AD-A115 491 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOO--ETC F/G 12/1
NONPARAMETRIC ESTIMATION OF DISTRIBUTION AND DENSITY FUNCTIONS --ETC(U)
MAY 82 J SWEEDER

UNCLASSIFIED AFIT/DS/MA/82-1 NL

1 OF 3

AD A
11549

1

UNITED STATES AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/DS/MA/82-1

1

NONPARAMETRIC ESTIMATION OF DISTRIBUTION AND DENSITY FUNCTIONS WITH APPLICATIONS

DISSERTATION

AFIT/DS/MA/82-1 James Sweeder
Captain USAF

Approved for public release; distribution unlimited

AFIT/DS/MA/82-1

NONPARAMETRIC ESTIMATION OF DISTRIBUTION AND DENSITY FUNCTIONS WITH APPLICATIONS

by

James Sweeder, B.S., M.S., M.B.A.

Captain USAF

Approved:

Albert H Moore May 20, 1982
Chairman

Brian W. Woodruff 20 May 1982

Joseph P. Cain 20 May 1982

20 May 1982

Accepted:

J S Przemieniecki 24 May 1982
Dean, School of Engineering

AFIT/DS/MA/82-1

NONPARAMETRIC ESTIMATION OF DISTRIBUTION AND DENSITY FUNCTIONS WITH APPLICATIONS

DISSERTATION

Presented to the Faculty of the School of Engineering
of the Air Force Institute of Technology
Air University
In Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

by

James Sweeder, B.S., M.S., M.B.A.
Captain USAF

May 1982

Approved for public release; distribution unlimited

Acknowledgements

This research was sponsored by the Air Force Wright Aeronautical Laboratories, Flight Dynamics Laboratory, Vehicle Synthesis Branch. I wish to thank the laboratory's management for their sponsorship and support. In particular, I especially want to thank Dr. Squire Brown for his invaluable assistance, consultation, and encouragement throughout the research project.

I am deeply indebted to the members of my advisory committee for their constant support, spirited interest, and helpful suggestions during all phases of the research. A special note of gratitude is due my committee chairman, Dr. Albert H. Moore. Through Dr. Moore's guidance, both my formal course work and this research effort were molded into a truly outstanding educational experience. His inspiration, unparalleled technical expertise, and valued friendship will long be remembered.

A brief note of thanks also to my student colleagues for their encouragement and to Mrs. Phyllis Reynolds for her superb assistance with the manuscript.

Finally and most importantly, to my wife, Cynthia, and to my daughter, Jennifer, I wish to express my deepest love and appreciation for their continued understanding and emotional support during the entire program.

Contents

List of Figures

Figure Page

List of Tables

Table Page

AFIT/DS/MA/82-1

Abstract

This report presents the theoretical development, evaluation, and applications of a new nonparametric family of continuous, differentiable, sample distribution functions. Given a random sample of independent, identically distributed, random variables, estimators are constructed which converge uniformly to the underlying distribution. A smoothing routine is proposed which preserves the distribution function properties of the estimators. Using mean integrated square error as a criterion, the new estimators are shown to compare favorably against the empirical distribution function. As density estimators, their derivatives are shown to be competitive with other continuous approximations. Numerous graphical examples are given. New goodness of fit tests for the normal and extreme value distributions are proposed based on the new estimators. Eight new goodness of fit statistics are developed. Extensive Monte Carlo studies are conducted to determine the critical values and powers for tests when the null hypothesis is completely specified and when the parameters of the null hypothesis are estimated. These tests were shown to be comparable with or superior to tests currently used. Forty-eight new estimators of the location parameter of a symmetric distribution are proposed based on the new models. For mild deviations from the normal distribution, some new estimators are shown to be superior to established robust estimators. Robust characteristics of the new estimators are discussed.

NONPARAMETRIC ESTIMATION OF DISTRIBUTION AND DENSITY FUNCTIONS WITH APPLICATIONS

I. Introduction

This dissertation develops and evaluates new nonparametric techniques for use in data analysis. A new family of nonparametric, continuous, differentiable sample distribution functions is proposed to model univariate random variables with continuous, unimodal densities. Much of the motivation for this research effort was the dominance of the empirical distribution function (EDF) as a basis for goodness of fit tests and robust estimation of parameters. This research presents a continuous, differentiable alternative to the EDF and its applications to statistical inference.

The EDF has long served as the mainstay for statistical inference. Only recently, as in a paper by Green and Hegazy, have other sample distribution functions even been considered as bases for goodness of fit tests (Ref 29). These alternatives are still classical step functions and are shown to generate powerful goodness of fit tests. The authors of the Princeton study on robust estimation of a location parameter, while using the EDF exclusively in their estimators, are careful to point out: "We ought not to close our eyes to other definitions of the empirical cumulative" (Ref 5:225). Their results, using the EDF, have given a large impetus to the search for robust estimators. Should not, then, a continuous, differentiable, alternative to the EDF offer the potential for improvement in goodness of fit testing and robust parameter estimation? This investigation shows that the new nonparametric family is a powerful tool for modeling univariate random variables, for goodness of fit tests and for robust estimation of the location parameter of a symmetric distribution.

Our analysis begins with the historical background of sample distribution functions given in Chapter II. Plotting positions for random samples and their relationship to sample distribution functions are discussed. Chapter III presents the theoretical development of the new family of nonparametric distribution functions. We demonstrate that the properties of a distribution function are preserved and discuss the conditions for uniform convergence. A routine is proposed to generate a smoother approximation for both the distribution and density functions. Six specific nonparametric models are generated from the new family and used for the remainder of the analysis. Three of these models are adaptive based on the estimated tail length of the underlying distribution from a random sample. Chapter IV examines the literature for techniques of distribution and density function estimation. A Monte Carlo analysis is then conducted to compare the distribution and density function estimates using mean integrated square error as the criterion. While not specifically designed as density function estimates, the new nonparametric models are shown to be competitive with or superior to two other continuous density function estimates. Several examples of the nonparametric estimates are graphically displayed. The chapter concludes with a discussion of a continuous nonparametric estimation of the hazard function which results from the differentiability of the distribution function estimate. Chapter V addresses the goodness of fit problem. After a brief historical survey, we propose eight new goodness of fit statistics. An extensive Monte Carlo analysis is conducted to determine the critical values for each test statistic for null distributions which are completely specified and when parameters are estimated. Two null distributions, the normal and the extreme value distributions, are considered. Subsequent Monte Carlo power studies show that the new tests are competitive with or superior to certain established goodness of fit tests. Chapter VI describes techniques for parameter estimation using the new models. Following a brief survey of location parameter estimation and robustness, we propose forty-eight new estimators of the location parameter of a symmetric distribution. The estimators are compared with the sample mean, sample median, and certain robust estimates proposed by Huber and Hampel. The comparisons are made using standardized empirical variances determined by Monte Carlo simulation, maximum and average relative deficiencies, and robust characteristics based on approximated influence curves over nine alternative symmetric distributions. For relatively mild deviations from the normal distribution, certain new nonparametric estimators are shown to have smaller deficiencies than the other estimators included in the study. The final chapter summarizes the major results of this research effort and also indicates potential applications of the new nonparametric models. We conclude with a discussion of areas for future research.

II. Background

Sample Distribution Functions (SDFs)

One of the initial steps in data analysis is the formulation of a sample cumulative distribution function. The most common of these is the empirical distribution function (EDF) whose properties are listed in Gibbons (Ref 27:73-75). Let $S_n(x)$ be the EDF.

$$S_n(x) = \begin{cases} 0 & x < X_{(1)} \\ i/n & X_{(i)} \leq x < X_{(i+1)} \quad i=1,\ldots,n-1 \\ 1 & x \geq X_{(n)} \end{cases}$$

It is easy to construct other sample distribution functions which are also step functions. Let $\{g_i\}$ $i=1,\ldots,n$ be a nondecreasing sequence of real numbers on $[0,1]$ with $g_n = 1$. Now define

$$G_n(x) = \begin{cases} 0 & x < X_{(1)} \\ g_i & X_{(i)} \leq x < X_{(i+1)} \quad i=1,\ldots,n-1 \\ 1 & x \geq X_{(n)} \end{cases}$$

Clearly $G_n(x)$ possesses all of the properties of a distribution function.

However, if we relax the property that $\lim_{x\to-\infty} G_n(x) = 0$ or $\lim_{x\to\infty} G_n(x) = 1$, we get improper sample distribution functions. An example is

$$G_n(x) = \begin{cases} 0 & x < X_{(1)} \\ i/(n+1) & X_{(i)} \leq x < X_{(i+1)} \quad i=1,\ldots,n-1 \\ n/(n+1) & x \geq X_{(n)} \end{cases}$$

It can be easily shown that the improper distribution function just defined has the same absolute convergence properties as the empirical distribution function. At this point, let us defer a discussion of the properties of either proper or improper distribution functions.

Several authors have considered specific alternatives to the empirical distribution function. In choosing a goodness of fit criterion, Pyke used the mean ranks as the basis for his modified empirical distribution function (Refs 10,68). Vogt also considered the mean ranks in his evaluation of maximal deviations from the EDF and his variant of the EDF (Ref 98). In a goodness of fit test for a completely specified continuous symmetric distribution, Schuster proposes an unbiased estimator $G_n(x)$ as the average of the EDF and another EDF based on reflecting the sample about the center of symmetry (Ref 82:1). He later considers the estimate of the distribution function when the center of symmetry is unknown. For a suitable choice of an estimator of the center of symmetry, it can be shown that the estimate formed by reflection about the estimated center of symmetry is asymptotically better than the EDF in specific cases (Ref 83). In testing for symmetry, Rothman and Woodroofe required their sample distribution function to be invariant under the transformation $x \to -x$. Thus, they used $2F_n^*(x) = S_n(x^+) + S_n(x^-)$ where S_n is the EDF (Ref 76). Hill and Rao generalized this sample distribution function in another article investigating the center of symmetry. They point out that the invariance property is preserved, if F_n^* is replaced by $F_n^{(a)}$ where $0 \leq a \leq 1$ and

$$F_n^{(a)}(x) = \begin{cases} aF_n(x^+) + (1-a)F_n(x^-) & x \leq 0 \\ (1-a)F_n(x^+) + aF_n(x^-) & x > 0 \end{cases}$$

for center of symmetry zero (Ref 36).

Forming continuous sample distribution functions is a simple task. Let $\{X_{(i)}\}$ $i=1,\ldots,n$ be an ordered sample. Choose a plotting rule for the $\{X_{(i)}\}$ to form the set of plotted values $\{G(X_{(i)})\}$ $i=1,\ldots,n$. A linear interpolation of the $G(X_{(i)})$ values for each interval $[X_{(i)}, X_{(i+1)}]$ gives a continuous function defined on $[X_{(1)}, X_{(n)}]$.• If $G(X_{(1)})=0$ and $G(X_{(n)})=1$, then the function is a proper distribution function. If not, we can construct extrapolation points $X_{(0)}$ and $X_{(n+1)}$ such that $G(X_{(0)})=0$ and $G(X_{(n+1)})=1$. Linear interpolation based on these extrapolated points again results in a continuous proper sample distribution function. Spline smoothing or exponential extrapolation for the $X_{(0)}$ and $X_{(n+1)}$ points are two other methods proposed by Andrews, et al., for forming alternatives to the EDF (Ref 5:224-225).

Whether we use a step function or a continuous one, the values of the sample distribution function at the observed data points can be used to estimate the underlying cumulative distribution function. The next section will examine several choices for these values, their use as plotting positions, and the relationship between plotting positions and sample distribution functions.

Plotting Positions

Used in graphical data analysis, plotting positions represent the estimated value of the underlying probability distribution function. As mentioned earlier, these plotting positions could be the values of some sample distribution functions at the observed data points.

As early as 1930, Hazen recognized that the values of the EDF were inappropriate for plotting annual flood data. He chose the midpoint of the jumps of the EDF as his plotting position (Ref 35). A limited survey comparing various choices of plotting positions was undertaken by Kimball (Ref 45). Some choices were based on specific underlying probability distributions. White proposes plotting positions for the Weibull distribution based on the expected value of reduced log-Weibull order statistics (Ref 107). For the normal distribution, Blom suggests plotting the ith order statistic at (i-.375)/(n+.25). He argues that this plotting rule

> . . . leads to a practically unbiased estimate of σ (the shape parameter) with a mean square deviation which is about the same as that of the unbiased best linear estimate.

He also states that Hazen's choice of plotting position for the normal ". . . leads to a biased estimate of σ with nearly minimum mean square deviation about σ" (Ref 7). While the previous discussion concerned some isolated plotting conventions, we now examine some basic systems of plotting positions.

<u>Rank Distributions</u>. Let $X_{(1)},\ldots,X_{(n)}$ be an ordered sample from an underlying probability distribution F(x). The distribution of $F(X_{(i)})$ i=1,...,n is the rank distribution. It can be shown that this distribution is a beta distribution for each i and is independent of the underlying distribution F, so long as F is differentiable (Refs 19, 44). A plotting position for the ith order statistic can be thought of as a point on the ith rank distribution. The question arises as to what point on the rank distribution should be used as a representative choice for $F(X_{(i)})$. Three measures of central tendency, the mean, median, and mode, are all contenders. $E(F(X_{(i)})) = i/(n+1)$, the mean rank, has the property that it divides [0,1] into n+1 equally probable intervals. The median rank, approximated by (i-.3)/(n+.4), can be used

as a better representative of skewed distributions, which most rank distributions are. For a unimodal distribution, the mode rank, (i-1)/(n-1), approximates the maximum of the probability density function of the rank distribution. Thus, the selection of a plotting position is equivalent to selecting a point from a beta distribution.

Blom's Formula. Plotting positions can also be derived from rather general expressions. Given choices of α and β such that $\alpha, \beta \leq 1$, a plotting position, G_i, can be defined as:

$$G_i = \frac{i-\alpha}{n-\alpha-\beta+1}$$

For specific choices of α and β, see reference 7. From the above formula, one can easily generate the same plotting positions in the rank distributions by judicious choices of α and β.

A slightly more general plotting position can be defined by

$$G_i = \frac{i+\alpha}{n+\beta} \quad \text{where } -1 \leq \alpha \leq \beta \leq 1$$

Once again, this formula allows for generation of common plotting positions by correct choices of α and β. Table II.1 summarizes some common plotting conventions.

TABLE II.1

PLOTTING POSITIONS OF THE ith ORDER STATISTIC

	Formula	Description
1.	i/n	value of the empirical distribution function
2.	$i/(n+1)$	mean rank
3.	$(i-1)/(n-1)$	mode rank
4.	$(i-.3)/(n+.4)$	median rank (approximation)
5.	$(i-.5)/n$	midpoint of the jump of the empirical distribution function
6.	$[n(2i-1)-1]/(n^2-1)$	average of the mean and mode ranks
7.	$(i-.375)/(n+.25)$	efficient approximation for the normal distribution
8.	$(i-\alpha)/(n-\alpha-\beta+1)$ $(\alpha,\beta\leq 1)$	Blom's general plotting position
9.	$(i+\alpha)/(n+\beta)$ $-1\leq\alpha\leq\beta\leq 1$	a more general plotting position

While the choice of plotting position is subject to the analyst's discretion, one must be aware of the problem of choosing plotting positions and generating a sample distribution function based on these positions. Once a plotting position is picked, any number of sample distribution functions can be constructed. However, given a specific plotting rule (midpoint of the jumps, limit from the right, etc.), a sample distribution step function uniquely determines the plotting positions.

III. New Nonparametric Sample Distribution Functions

Introduction

Having already seen the uses of various discrete plotting positions and their relationship to sample distribution step functions, we now propose a new family of approximations. The next section presents the theoretical development of a family of nonparametric, continuous, differentiable sample distribution functions. Properties of distribution functions are preserved and uniform convergence is demonstrated. A smoothing routine is selected which again preserves the distribution function properties. Three specific nonparametric models are developed by a detailed analysis of the stylized and random samples from selected members of the Generalized Exponential Power distribution. Finally, three adaptive nonparametric models were proposed based on using percentile ratios as a discriminant.

Theoretical Development

Consider a random sample $X_1,\ldots,X_n$ of size n from an unknown univariate, continuous, probability distribution function F. Let $X_{(1)},\ldots,X_{(n)}$ be the ordered sample. Now let $G_i = G(X_{(i)})$, i=1,...,n, be the plotting position for the ith order statistic based on some sample distribution function G.

Our goal is to estimate F by a nonparametric approach while preserving the following properties of the estimator, F_n:

1. F_n is differentiable
2. F_n is a distribution function
3. $F_n(X_{(i)}) = G_i$, i=1,...,n

Linear interpolation will, of course, satisfy conditions 2 and 3, but we require differentiability at the data points. What is needed is a family of nondecreasing curves on $[X_{(i)}, X_{(i+1)}]$ such that

$$\lim_{x \to X_i^-} F_n'(x) = \lim_{x \to X_i^+} F_n'(x) \text{ for each } i=1,\ldots,n$$

Arbitrarily, set the derivative equal to zero at each data point. Now, consider the midpoint of the interval $[X_{(i)}, X_{(i+1)}]$. Let

$$F_n\left(\frac{X_{(i)}+X_{(i+1)}}{2}\right) = \frac{G_i+G_{i+1}}{2}$$

Consider the function $-a\cos y$, which is monotonically increasing on the interval $[0, \pi]$ where a is a constant. Making the transformation

$$y = \pi\left(\frac{x-X_{(i)}}{X_{(i+1)}-X_{(i)}}\right)$$

yields

$$F_n(x) = \frac{G_i+G_{i+1}}{2} - a\cos\pi\left(\frac{x-X_{(i)}}{X_{(i+1)}-X_{(i)}}\right) \qquad (3.1)$$

Requiring $F_n(X_{(i)}) = G_i$ for each $i=1,\ldots,n$ gives

$$a = \frac{G_{i+1}-G_i}{2} .$$

Defining extrapolation points $X_{(0)}$ and $X_{(n+1)}$ such that $G_0 = 0$ and $G_{n+1} = 1$ completes the derivation. Thus, equation 3.1 becomes:

$$F_n(x)= \begin{cases} 0 & x<X_0 \\ G_i + \dfrac{G_{i+1}-G_i}{2}\left(1-\cos\pi\left(\dfrac{x-X_{(i)}}{X_{(i+1)}-X_{(i)}}\right)\right) & \\ & X_{(i)}\leq x<X_{(i+1)} \quad i=0,\ldots,n \\ 1 & x\geq X_{n+1} \end{cases} \qquad (3.2)$$

Differentiating, one immediately obtains an estimate of the probability density function.

$$f_n(x) = \begin{cases} \dfrac{\pi}{2}\left(\dfrac{G_{i+1}-G_i}{X_{(i+1)}-X_{(i)}}\right)\sin\pi\left(\dfrac{x-X_{(i)}}{X_{(i+1)}-X_{(i)}}\right) & \\ & X_{(i)}\leq x<X_{(i+1)}, \quad i=0,\ldots,n \\ 0 & \text{elsewhere} \end{cases} \qquad (3.3)$$

Clearly, the derived $F_n(x)$ satisfies the three properties required. However, the utility of such an estimate can certainly be questioned at this point.

Figures 3.1 and 3.2 show the estimates of the cumulative and density functions respectively for a random sample of size 20 from a normal distribution with zero mean and unit variance. The plotting positions chosen were the average of the mean and mode ranks. The extrapolation points $X_{(0)}$ and $X_{(n+1)}$ were chosen as: $X_{(0)} = 2X_{(1)} - X_{(2)}$ and $X_{(n+1)} = 2X_{(n)} - X_{(n-1)}$. The estimated CDF does approximate the true CDF in a continuous fashion, but provides the same inferences about the underlying population as the plotting positions themselves. The estimated PDF plot is analogous to a histogram with the intervals chosen to contain only one data point. Some shape of the underlying density can be inferred, especially with larger sample sizes, but any inference concerning the density shape or type is limited.

The basic undesirable property in the development thus far has been the zero derivative of the estimated cumulative distribution function at the data points. To avoid these zero derivatives, consider applying a variation of the jackknife. This technique was developed by Quenouille (Refs 70,71) as a means of reducing the bias of an estimator. In an abstract, Tukey proposes using the technique for robust interval estimation (Ref 96). An excellent survey and bibliography is given by Miller (Ref 58). More recent applications and extensions of

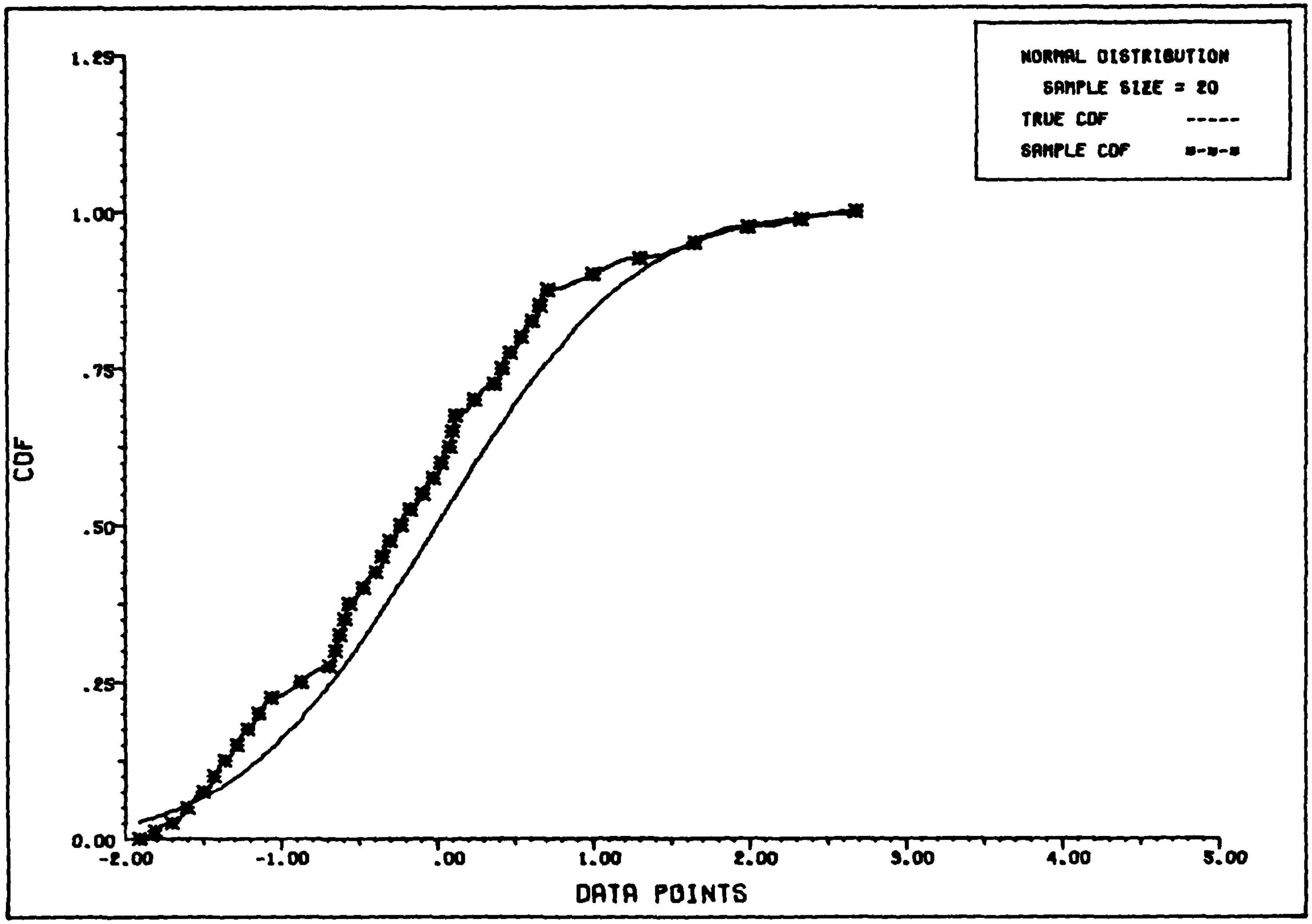

Figure 3.1. Sample CDF vs N(0,1)

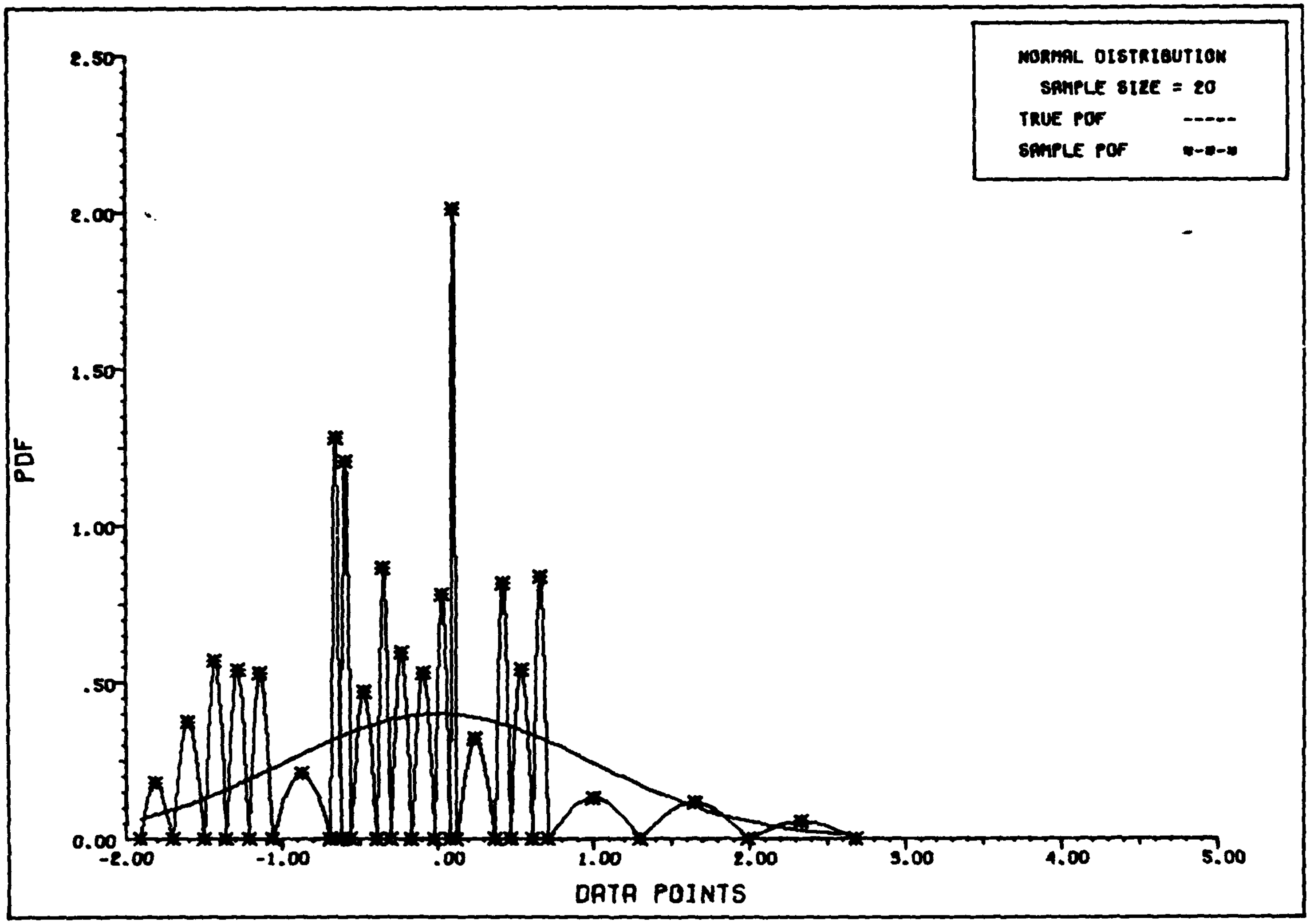

Figure 3.2. Sample PDF vs N(0,1)

the jackknife may be found in Gray, et al., and Cressie (Refs 15,28).

Analogous to Quenouille's development, let $X_{(1)},\ldots,X_{(n)}$ be an ordered sample. Choose $k \leq n/2$ to be the number of subsamples. Beginning at $X_{(1)}$ form the subsamples by assigning each successive order statistic to a new subsample until the k+1 order statistic is reached. Repeat this assignment process beginning with this order statistic, using the same ordering of subsamples, until all n order statistics are assigned.

Mathematically, if k is the number of subsamples, then $n=km+r$ where $m=[n/k]$ and $r=n$ modulo k. Now let ℓ index the subsamples, $\ell=1,\ldots,k$ and let $y_{(j,\ell)}$ be the jth element of subsample ℓ. Thus,

$$Y_{(j,\ell)} = X_{(\ell+k(j-1))}$$

where $j=1,\ldots,m$ if $\ell>r$

$j=1,\ldots,m+1$ if $\ell \leq r$

Clearly, there will be k ordered subsamples, r of which have size m+1 and k-r have size m.

Returning to the zero derivative problem, now that the subsamples are generated, consider the following estimate of the cumulative distribution function. Form k estimates, $SF_\ell(x)$, where $SF_\ell(x) = F_{n*}(x)$ for $\ell=1,\ldots,k$ and $F_{n*}(x)$ is the continuous, differentiable, sample distribution function defined in equation 3.2 and $n^* = \begin{cases} m & \text{if } \ell > r \\ m+1 & \ell \leq r \end{cases}$. The derivatives $SF'_\ell(x)$ are zero at each data point of the subsamples. Now simply average these estimates to form the sample cumulative function,

$$SF(x) = \frac{1}{k} \sum_{\ell=1}^{k} SF_\ell(x) \tag{3.4}$$

and sample density function

$$sf(x) = SF'(x) = \frac{1}{k} \sum_{\ell=1}^{k} SF'_\ell(x) \tag{3.5}$$

Note that each of the subsamples has its own extrapolated points, $Y_{(0,\ell)}$ and $Y_{(n^*+1,\ell)}$. Now let

$$X_{min} = \min_{\ell} \{Y_{(0,\ell)}\}$$

and $$X_{max} = \max_{\ell} \{Y_{(n^*+1,\ell)}\}.$$

Thus, the cumulative and density functions in equations 3.4 and 3.5 are formally defined as:

$$SF(x) = \begin{cases} 0 & x<X_{min} \\ \frac{1}{k} \sum_{\ell=1}^{k} SF_\ell(x) & X_{min} \leq x \leq X_{max} \\ 1 & x>X_{max} \end{cases} \tag{3.6}$$

$$sf(x) = \begin{cases} \frac{1}{k} \sum_{\ell=1}^{k} SF'_\ell(x) & X_{min} \leq x \leq X_{max} \\ 0 & \text{elsewhere} \end{cases} \tag{3.7}$$

Two important results occur by this averaging. First, while we required that $F_n(Y_{(j,\ell)}) = G_j$ for each data point in the subsample, $SF(Y_{(j,\ell)})$ is not necessarily equal to the $G_{(\ell+k(j-1))}$ for the entire sample. Thus, we are no longer tied to restricting our estimates to the plotting positions of the original sample. Second, while each $SF_\ell'(Y_{(j,\ell)}) = 0$, $SF'(Y_{(j,\ell)}) = 0$ only if there are at least k data points identically equal to $Y_{(j,\ell)}$. Since the assumed underlying distribution function is continuous, the probability of such an event is zero. Of course, in actual data sets, due to measurement accuracy, this event may occur. However, since it would require k occurrences in the same random sample to force a zero derivative, the limitation does not appear to be unreasonable. Figures 3.3 and 3.4 show the effect of averaging on the normal sample of size 20 considered previously. The number of subsamples, k, was chosen as four. Both the distribution and density functions are beginning to identify the shape of the underlying random variable.

Properties

Now that we have defined estimates for both the cumulative distribution and density functions by equations 3.6 and 3.7, we need to examine their properties. Specifically, we will consider the distribution function properties and uniform convergence.

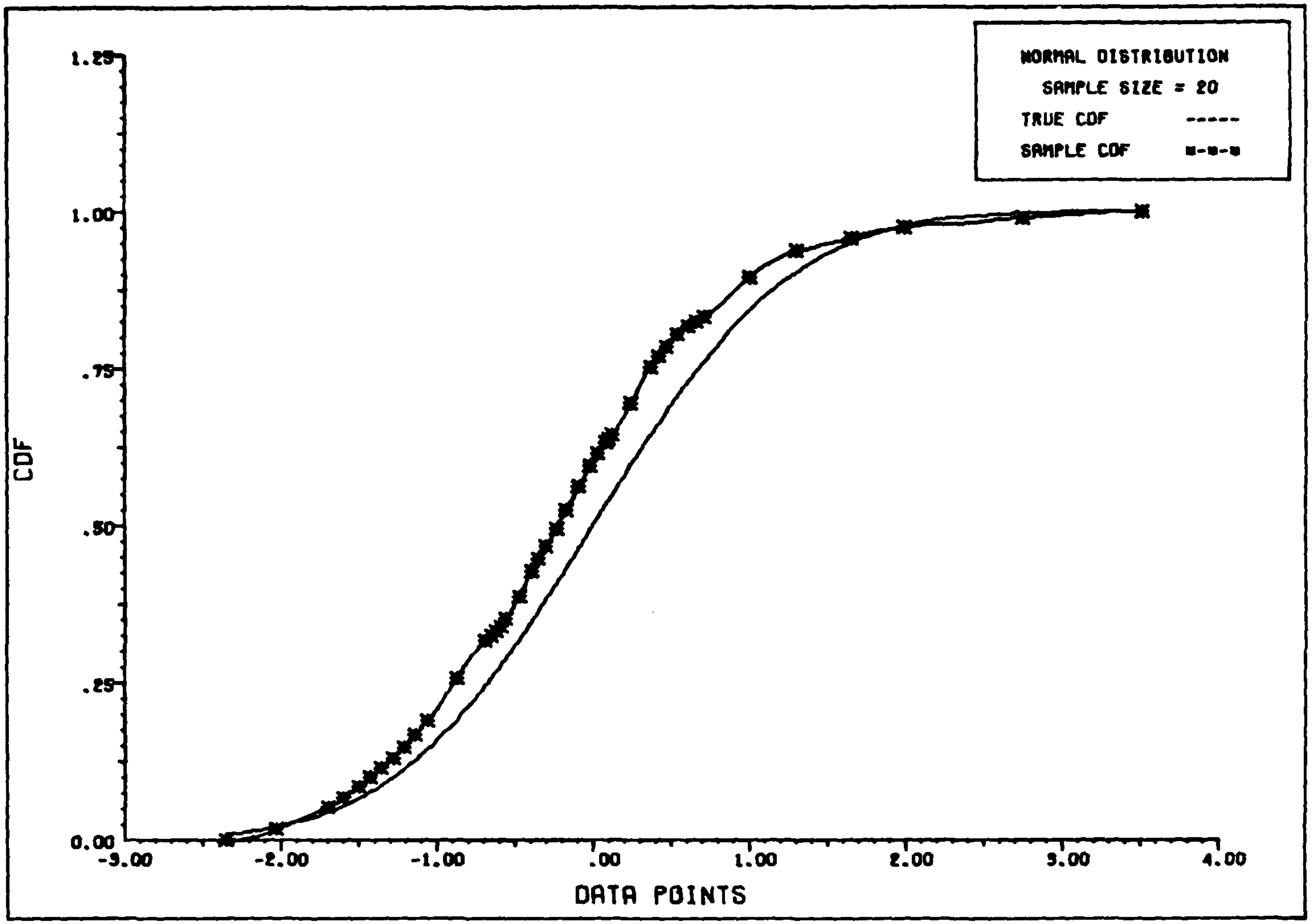

Figure 3.3. Sample CDF--4 Subsamples vs N(0,1)

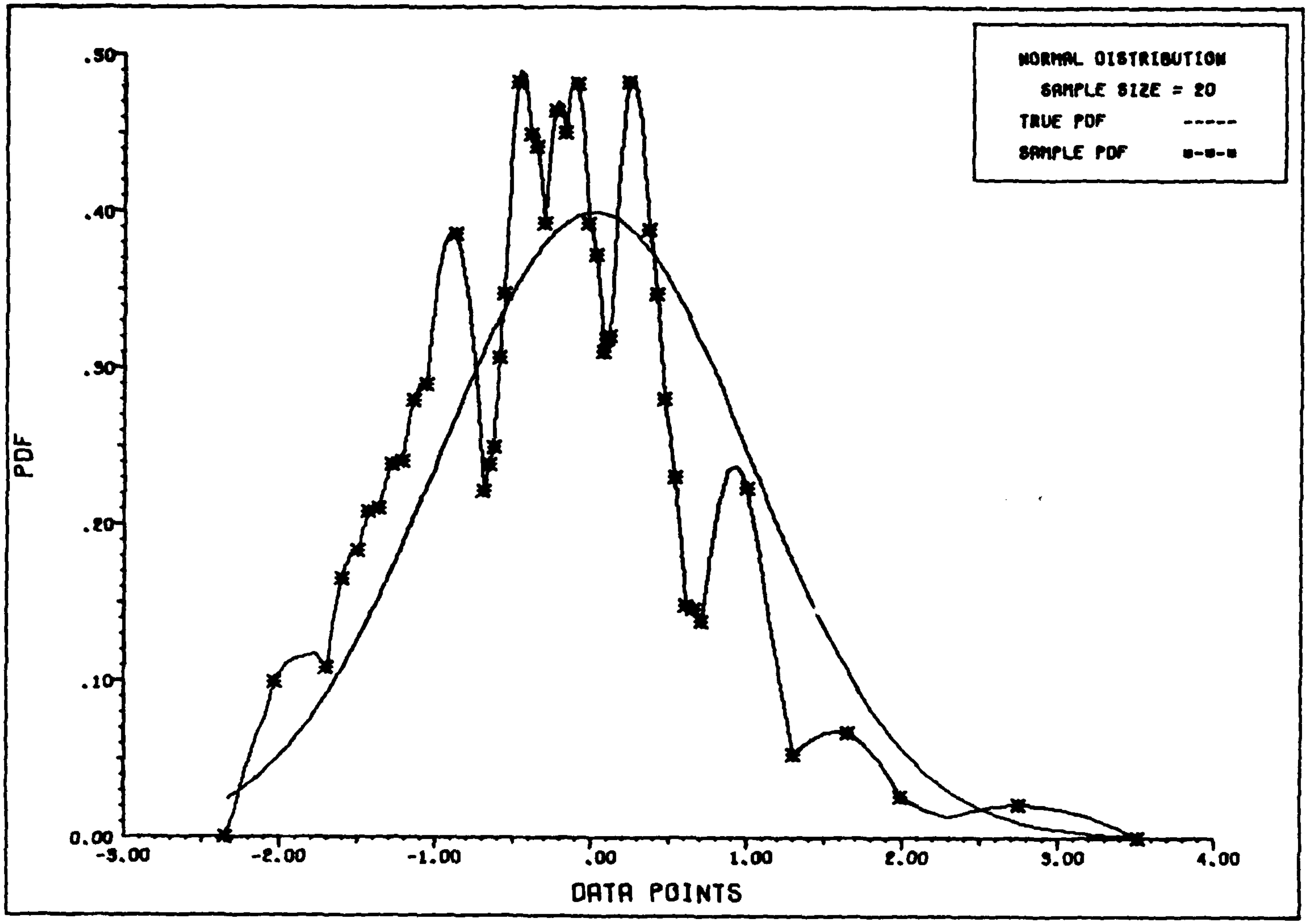

Figure 3.4. Sample PDF--4 Subsamples vs N(0,1)

Let R^1 be the real line, β the borel field on R^1 and P, a probability measure defined on β. The function F defined on (R^1, β, P) by $F(x) = P(\{t\varepsilon R^1: -\infty<t\leq x\})$ is the distribution function of P. Any standard probability text gives the properties of F (see references 13 and 49). F satisfies the following three properties:

1. F is nondecreasing
2. F is continuous from the right
3. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$

The function SF(x) defined in equation 3.6 clearly satisfies these properties. Further, since each $SF_\ell(x)$ is differentiable for each $x\varepsilon R^1$, SF(x) is also differentiable.

To examine the convergence of our estimator in equation 3.6, we begin by examining the convergence of step functions for subsamples.

Theorem 3.1. If $\overline{S}_{n*}$ is a sample distribution function based on a subsample of the form

$$\{Y_{(j,\ell)}\} \quad j=1,\ldots,n^*, \quad \ell=1,\ldots,k<\infty,$$

where $Y_{(j,\ell)} = X_{(\ell+k(j-1))}$

as defined in the previous section, and

$$n^* = \begin{cases} m & \text{if } \ell>r \\ m+1 & \text{if } \ell\leq r \end{cases}$$

then $\overline{S}_{n*}(x)$ converges uniformly to F(x) where

$$\overline{S}_{n*}(x) = \begin{cases} 0 & x < Y_{(1,\ell)} \\ j/n* & Y_{(j,\ell)} \leq x < Y_{(j+1,\ell)} \qquad j=1,...,n* \\ 1 & x \geq Y_{(n*,\ell)} \end{cases}$$

Proof. Without loss of generality, let F have a finite support [a, b] in R^1.

Let $\quad D = \sup\limits_{-\infty<x<\infty} |\overline{S}_{n*}(x) - F(x)| = |\frac{j}{n*} \cdot \frac{n}{i} S_n(x) - F(x)|$

where $\quad S_n(x)$ is the *EDF*.

Now $\quad D \leq \sup\limits_{-\infty<x<\infty} |S_n(x) - F(x)| + \left| \left(\frac{n*i-jn}{n*i}\right) S_n(x) \right|$

By construction, $n=km+r$, $i=\ell+k(j-1)$, $r<k$, and $\ell \leq k < \infty$. For simplicity, consider the case $n*=m$ ($n*=m+1$ is similar with slightly more algebra).

So, $\quad D \leq \sup\limits_{-\infty<x<\infty} |S_n(x)-F(x)| + \left| \left(\frac{m(\ell+k(j-1))-j(km+n)}{m(\ell+k(j-1))}\right) S_n(x) \right|$

$$\leq \sup_{-\infty<x<\infty} |S_n(x)-F(x)| + \left| \left(\frac{\frac{\ell}{j} - \frac{k}{j} - \frac{r}{m}}{\frac{\ell}{j} + k - \frac{k}{j}}\right) S_n(x) \right|$$

$$\lim_{n\to\infty} D \leq \lim_{n\to\infty} \left[D_n + \sup_{-\infty<x<\infty} \left| \left(\frac{\frac{\ell}{j} - \frac{k}{j} - \frac{r}{m}}{\frac{\ell}{j} + k - \frac{k}{j}}\right) S_n(x) \right| \right]$$

Case i: $x=a$

$n\to\infty$ implies $m\to\infty$, $j\to 1$, $S_n(x)\to 0$

Case ii: $x\varepsilon(a,b]$

$n\to\infty$ implies $m\to\infty$, $j\to\infty$

Since $\ell\leq k<\infty$ and $r<k<\infty$, and since $P[\lim_{n\to\infty} D_n=0]=1$ by Glivenko's Theorem (Ref 73:353), $P[\lim_{n\to\infty} D=0]=1$.

We now have established uniform convergence for sample distribution functions based on our constructed subsamples. Let us consider a general sample distribution function defined on these subsamples. We will continue to use $n^*=m$.

Theorem 3.2. $SF_\ell^-(x)$ converges uniformly to $F(x)$ where

$$SF_\ell^-(x) = \begin{cases} 0 & x<Y_{(1,\ell)} \\ (j+\alpha)/(m+\beta) & Y_{(j,\ell)}\leq x<Y_{(j+1,\ell)} \quad j=1,\ldots,m \\ 1 & x\geq Y_{(m+1,\ell)} \end{cases}$$

and $-1\leq\alpha\leq\beta\leq 1$, $Y_{(m+1,\ell)} = Y_{(m,\ell)} + \delta$

where $\delta\to 0$ as $m\to\infty$

Proof.

$$SF_\ell^-(x) = \begin{cases} 0 \cdot \overline{S}_m(x) & x<Y_{(1,\ell)} \\ \dfrac{j+\alpha}{m+\beta}\cdot\dfrac{m}{j}\,\overline{S}_m(x) & Y_{(j,\ell)}\leq x<Y_{(j+1,\ell)} \\ & \qquad j=1,\ldots,m \\ 1 \cdot \overline{S}_m(x) & x\geq Y_{(m+1,\ell)} \end{cases}$$

Now let $D_n^* = \sup\limits_{-\infty<x<\infty} \left| SF_\ell^-(x) - F(x) \right|$

$$\leq D_n + \sup_{-\infty<x<\infty} \left| \left(\frac{\frac{\beta}{m} - \frac{\alpha}{j}}{1 + \frac{\beta}{m}} \right) \bar{S}_m(x) \right|$$

Again, if x is an interior point or an end point the second term approaches zero as $n \to \infty$. Thus, by Theorem 3.1

$$P[\lim_{n\to\infty} D_n^* = 0] = 1$$

A slight modification of the hypothesis of Theorem 3.2 gives another family of estimators which converge uniformly to F(x). The proof of the following theorem is similar and thus omitted.

Theorem 3.3. $SF_\ell^+(x)$ converges uniformly to F(x) where

$$SF_\ell^+(x) = \begin{cases} 0 & x<Y_{(0,\ell)} \\ \frac{j+1+\alpha}{m+\beta} & Y_{(j,\ell)} \leq x < Y_{(j+1,\ell)} \quad j=0,1,\ldots,m-1 \\ 1 & x \geq Y_{(m,\ell)} \end{cases}$$

and $x \leq \alpha \leq \beta \leq 1$, $Y_{(0,\ell)} = Y_{(1,\ell)} - \delta$

where $\delta \to 0$ as $m \to \infty$.

We now have, by the previous two theorems, two families of sequences of estimators which converge uniformly to the underlying probability distribution function $F(x)$. Now consider $SF_\ell(x)$ as derived in the previous section and define $G_i = SF_\ell^-(Y_{j,\ell})$ for $j=0,1,\ldots,m+1$. Thus

$$G_{i+1} = SF_\ell^+(Y_{(j,\ell)}) \text{ for } j=0,1,\ldots,m$$

since $SF_\ell^-(Y_{(j,\ell)}) = SF_\ell^+(Y_{(j-1,\ell)})$.

We know by construction that

$$SF_\ell^-(x) \le SF_\ell(x) \le SF_\ell^+(x) \text{ for every } x.$$

This implies that

$$\lim_{n\to\infty} \sup_{-\infty<x<\infty} |SF^-(x) - F(x)|$$

$$\le \lim_{n\to\infty} \sup_{-\infty<x<\infty} |SF(x)-F(x)| \le \lim_{n\to\infty} \sup_{-\infty<x<\infty} |SF^+(x)-F(x)|$$

From Theorems 3.2 and 3.3, we can summarize with the following theorem.

Theorem 3.4. $SF_\ell(x)$ converges uniformly to $F(x)$ where

$$SF_\ell(x) = \begin{cases} 0 & x<Y_{(0,\ell)} \\ G_j + \frac{G_{j+1}-G_j}{2}\left(1-\cos\pi\left(\frac{x-Y_{(j,\ell)}}{Y_{(j+1,\ell)}-Y_{(j,\ell)}}\right)\right) & \begin{array}{l} Y_{(j,\ell)}\le x<Y_{(j+1,\ell)} \\ j=0,1,\ldots,m \end{array} \\ 1 & x\ge Y_{(m+1,\ell)} \end{cases}$$

and $G_j = G(Y_{(j,\ell)})$, $j=0,1,\ldots,m+1$

where

$$G(x) = \begin{cases} 0 & x<Y_{(1,\ell)} \\ (j+\alpha)/(m+\beta) & Y_{(j,\ell)}\le x<Y_{(j+1,\ell)} \quad j=1,\ldots,m \\ 1 & x\ge Y_{(m,\ell)} \end{cases}$$

for $-1\le\alpha\le\beta\le 1$

To prove our final result, we need a lemma.

Lemma 3.5. A finite convex combination of estimators which converge uniformly to $F(x)$ also converges uniformly to $F(x)$.

Proof. Let $\{T_{i,n}(x)\}$ $i=1,\ldots,k$ be a sequence of estimators converging uniformly to $F(x)$, i.e.,

$$P(\lim_{n\to\infty} \sup_{-\infty<x<\infty} |T_{i,n}(x) - F(x)|= 0) = 1 \text{ for } i=1,\ldots,k$$

and let $k<\infty$.

Now let $T_n(x) = \sum_{i=1}^{k} \alpha_i T_{i,n}(x)$

and $\sum_{i=1}^{k} \alpha_i = 1$

for $0\le\alpha_i\le 1$

$$\lim_{n\to\infty} \sup_{-\infty<x<\infty} |T_n(x) - F(x)|$$

$$= \lim_{n\to\infty} \sup_{-\infty<x<\infty} \left| \sum_{i=1}^{k} \alpha_i T_{i,n}(x) - \sum_{i=1}^{k} \alpha_i F(x) \right|$$

$$\leq \lim_{n\to\infty} \sup_{-\infty<x<\infty} \sum_{i=1}^{k} \alpha_i \left| T_{i,n}(x) - F(x) \right|$$

$$\leq \sum_{i=1}^{k} \alpha_i \lim_{n\to\infty} \sup_{-\infty<x<\infty} \left| T_{i,n}(x) - F(x) \right|$$

since $k<\infty$

Each term in the sum is zero by hypothesis. The uniform convergence of the finite convex combination follows immediately.

Applying the previous lemma to the function SF(x) as defined in equation 3.6, we can state the following theorem.

Theorem 3.6. SF(x) as defined in equation 3.6, converges uniformly to F(x).

At this point we have an estimator SF(x) of F(x) which is itself a continuous, differentiable distribution function and also converges uniformly. The same results, however, are not available for the derivative, sf(x). While it is true that sf(x) is continuous and differentiable almost everywhere, convergence properties will have to be inferred from the Monte Carlo analysis of Chapter IV.

Smoothing

Although the estimator family has been defined and the properties listed, a quick glance at Figures 3.3 and 3.4 indicates possible room for improvement. If we could dampen some of the sinusoidal activity in both the sample cumulative and sample density functions, our estimators should better approximate the underlying process. Two methods of such a smoothing were initially investigated: spline smoothing and a Fourier smoothing method.

Once SF(x) and sf(x) have been determined we can generate their values at each data point X_i to form the sets $\{SF(X_i)\}_{i=1,\ldots,n}$ and $\{sf(X_i)\}_{i=1,\ldots,n}$. At this point, however, note that we are not restricted to the original data set. We could choose a set $\{Z_j\}_{j=1,\ldots,m}$ and its corresponding sets $\{SF(Z_j)\}_{j=1,\ldots,m}$ and $\{sf(Z_j)\}_{j=1,\ldots,m}$ by an arbitrary rule, such as equally spaced points in the domain or inversion of SF(x) at some specified plotting positions. Thus m, the number of points used in smoothing, can be as large (or small) as we choose.

To apply spline smoothing (Ref 109) we can proceed in two directions: (1) independently smooth both the distribution and density functions, or (2) smooth only the distribution (density) function and analytically differentiate (integrate) to get the density (distribution) function. Proceeding in either of these directions opens the possibility of negative density values.

A second smoothing technique was hypothesized from the density and cumulative estimation work of Kronmal and Tarter (Refs 40, 48). Their investigation yielded estimates with impressive mean integrated square errors (MISEs). Analogous to the spline methods, we could use the Fourier approximation method of Kronmal and Tarter independently for the distribution and density functions or separately and derive the other. The same drawback occurs using the Fourier expansion as with splines--negative density values. Since our initial goal in this development was to preserve the distribution function properties of our estimators as well as add differentiability, it would be foolish at this point to abandon this aim in favor of the possible smoothing advantages of spline or Fourier expansions. Thus, both spline smoothing and the use of Fourier expansions were discarded.

The availability of both distribution and density function estimates at arbitrary points in the domain suggested an alternative approach. In a 1979 article, Efron (Ref 23) developed a "bootstrap method" related to the "double Monte Carlo" method proposed by Moore (Ref 59). Both methods estimate the distribution function based on sample data and then create a pseudosample by sampling from this estimated distribution. Rather than sampling from the estimated distribution, as these authors suggest, consider inverting the estimated distribution at specific points according to some rule. Specifically, solve $SF(Z_{(j)}) = G_j$ for $Z_{(j)}$, where $\{G_j\}_{j=1,\ldots,m}$ are predetermined plotting positions. The set $\{Z_{(j)}\}_{j=1,\ldots,m}$ is now a pseudosample based on some regular divisions, the plotting positions G_j, of SF(x). Having generated this pseudosample, now apply equations 3.6 and 3.7 to form new estimates of the distribution and density functions. Of course, this inversion process could be repeated and other estimates formed on the basis of new pseudosamples.

The previous derivation clearly preserves the distribution function properties of the estimators, as well as differentiability and continuity. By inverting SF(x) at the plotting positions G_j, we also preserve ordering and spacing information contained in the original sample, in contrast to the random sampling procedures of Moore and Efron. Although no formal proof of uniform convergence of this smooth distribution function estimator is presented, empirical evidence from graphical and Monte Carlo analysis of this estimator strongly suggests that uniform convergence is preserved. We will postpone a detailed analysis of these estimators to the results of Monte Carlo analyses of the next chapter.

Figures 3.5 through 3.9 give a graphical display of the smoothing technique proposed for our random sample of size 20 from the normal distribution. Figures 3.5 and 3.6 show the smoothed approximation and the true underlying standard normal distribution. Figures 3.7 and 3.8 compare the smoothed approximation to a normal distribution whose parameters are minimum variance unbiased estimates. Note the performance of the nonparametric model without the assumption of normality. Figure 3.9 compares the smoothed approximation to the empirical cumulative distribution function. Choices for the plotting positions, inversion points, and other variables have been made using methods discussed in the next section.

Choice of Variables for the Estimators

Since the approximation method and smoothing technique have been defined, we now seek to identify the variables needed to form our final estimators. The investigation will examine five sets of variables: (1) the number of subsamples for a given sample size; (2) plotting positions, $\{G_j\}_{j=1,\ldots,n^*}$ for each subsample; (3) extrapolation values, $Y_{(0)}$ and $Y_{(n^*+1)}$ for each subsample; (4) inversion points for the smoothing routine to generate the pseudosample; and (5) the number of inversions. Judicious choices of these sets of variables should give us an estimator with good approximating properties.

Due to the array of possible choices of the variables and their complex interaction in the estimators, it

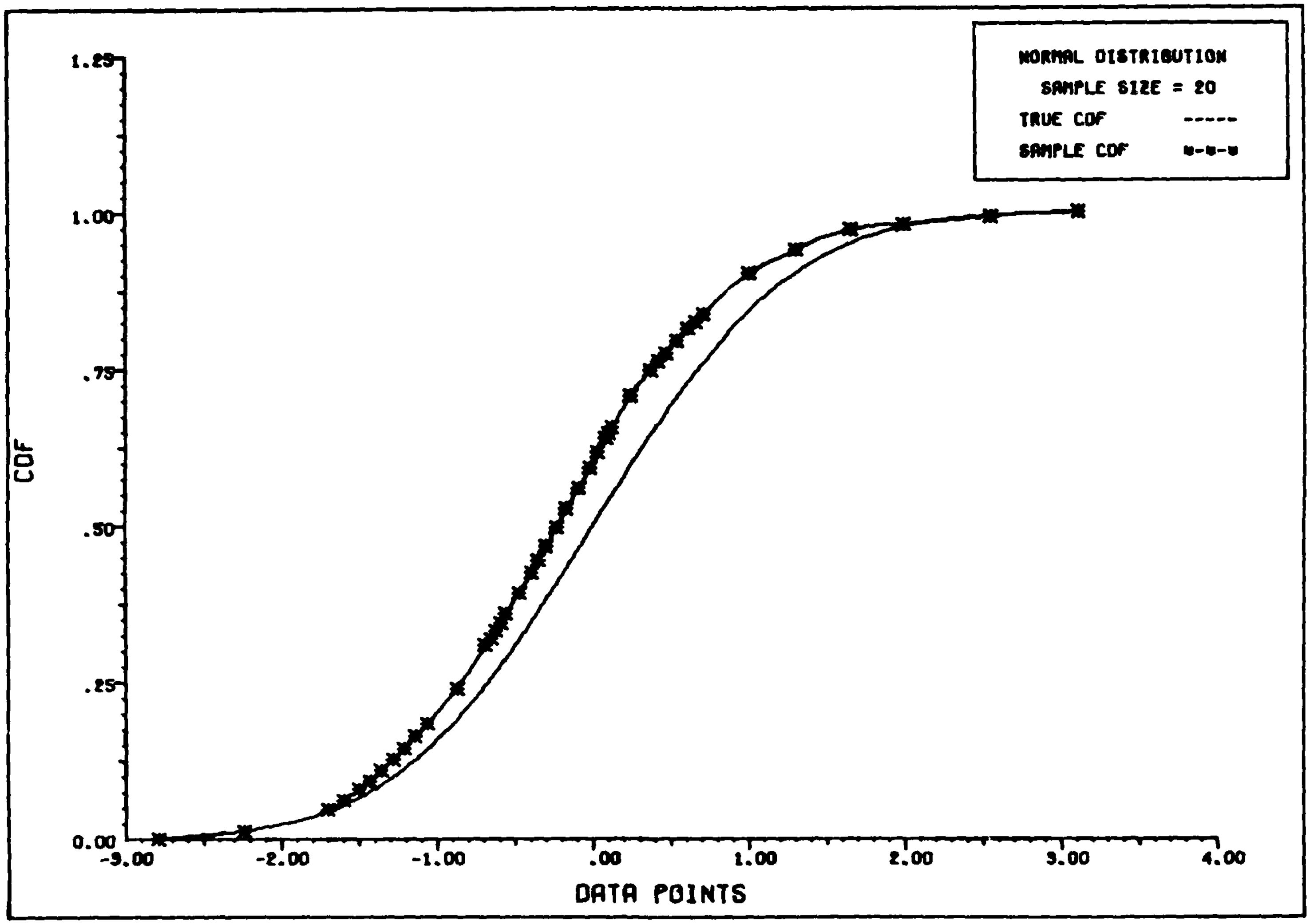

Figure 3.5. Sample CDF--Smoothed vs N(0,1)

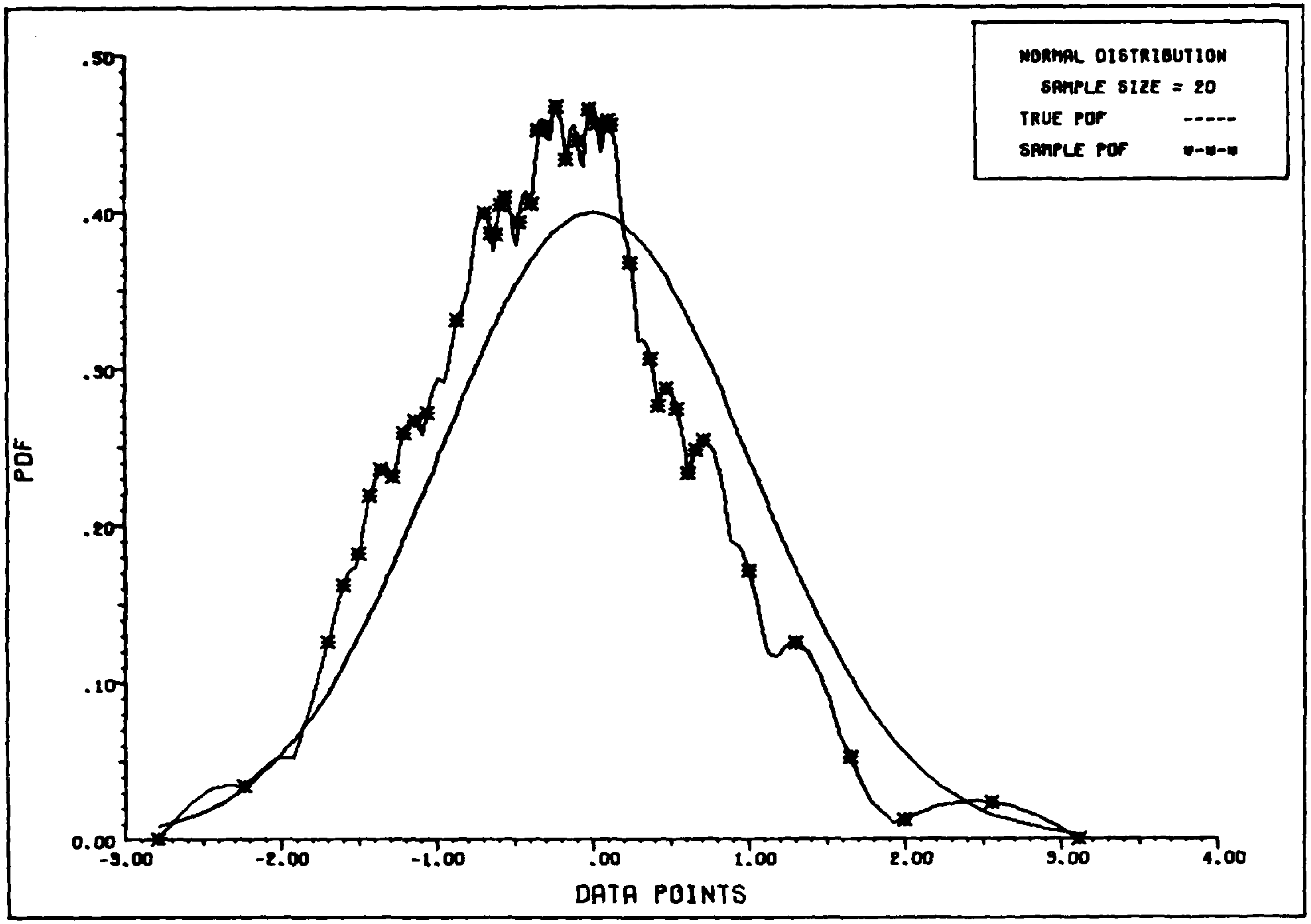

Figure 3.6. Sample PDF--Smoothed vs N(0,1)

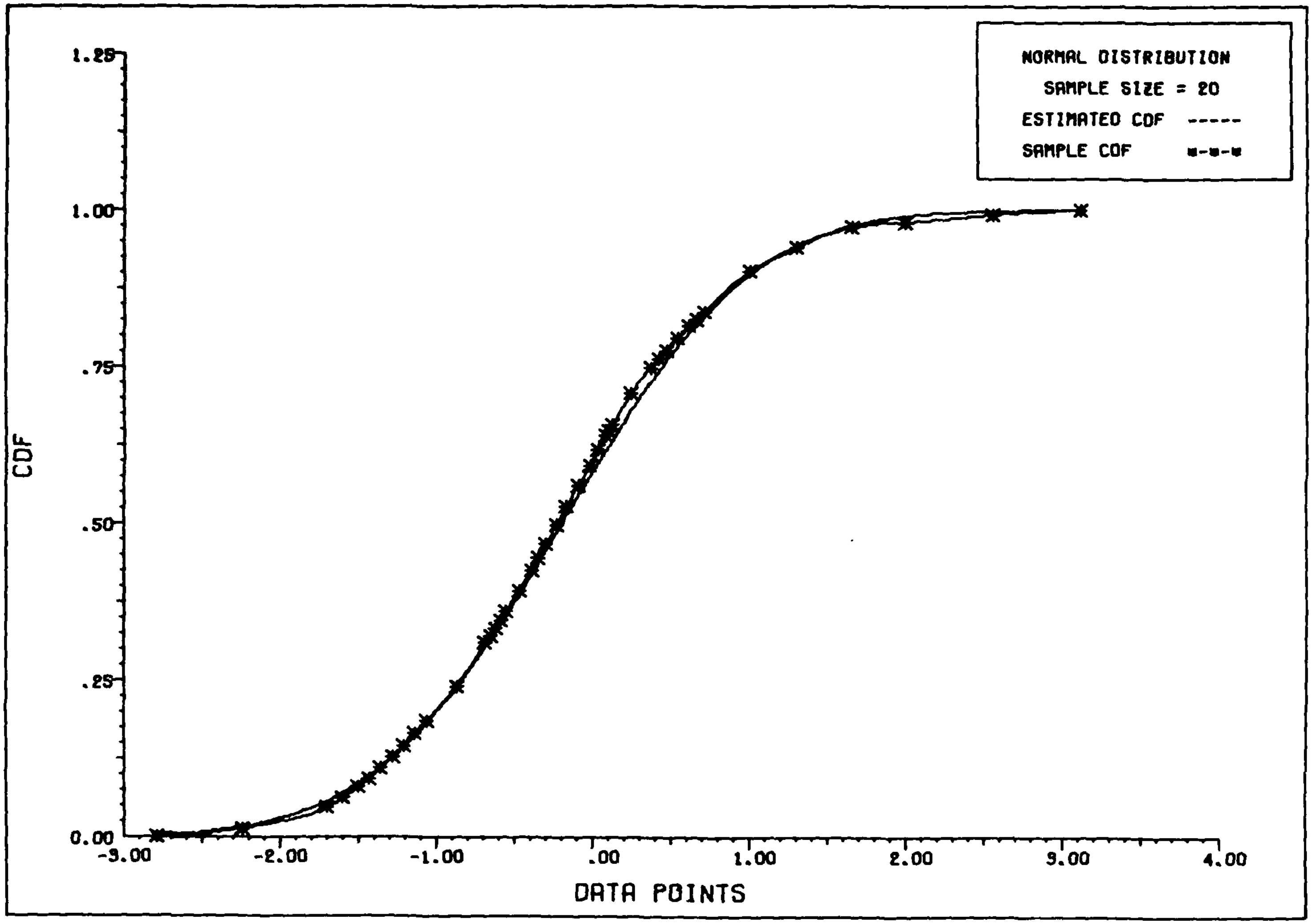

Figure 3.7. Sample CDF--Smoothed vs $N(\bar{X},S^2)$

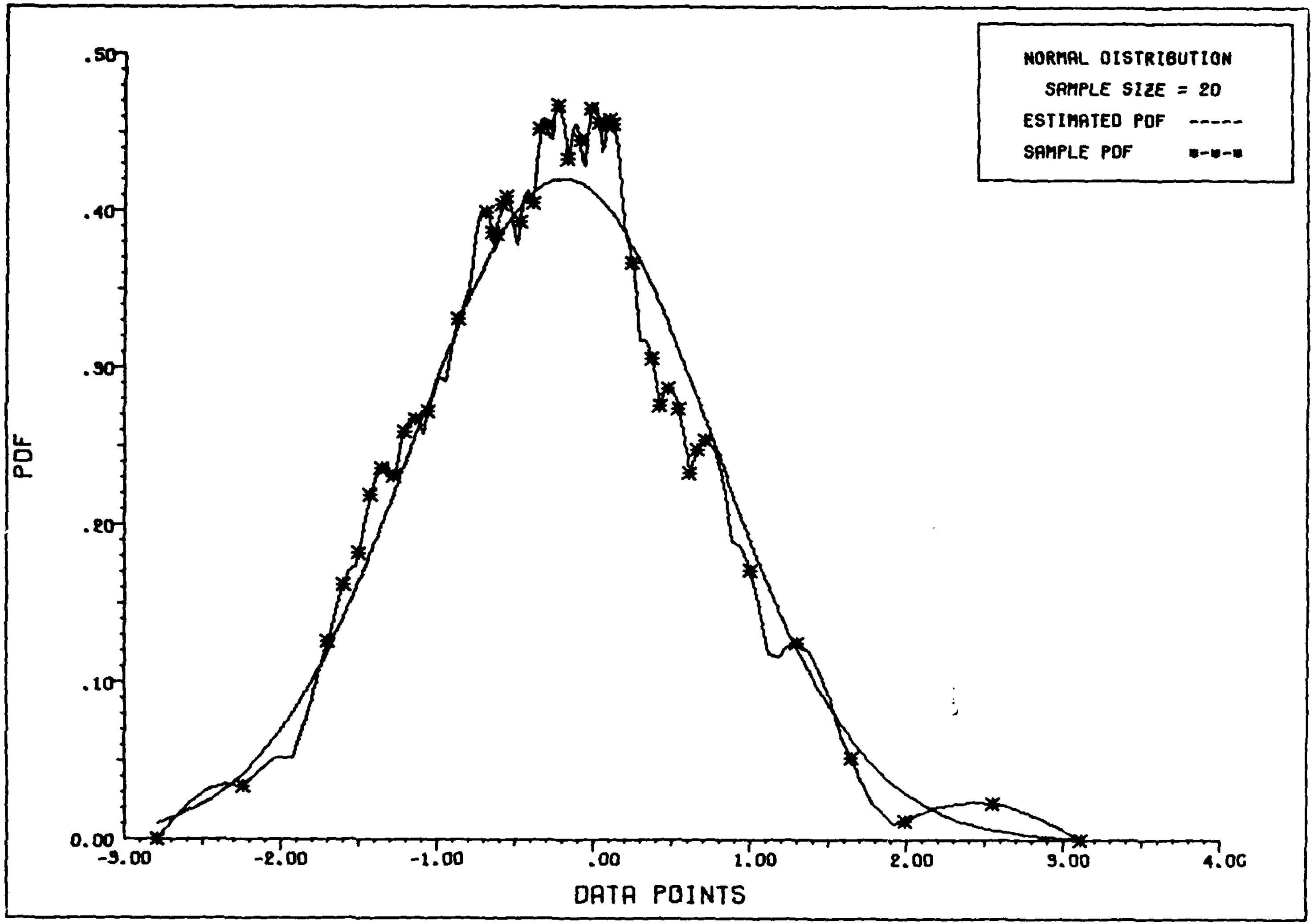

Figure 3.8. Sample PDF--Smoothed vs $N(\bar{X},S^2)$

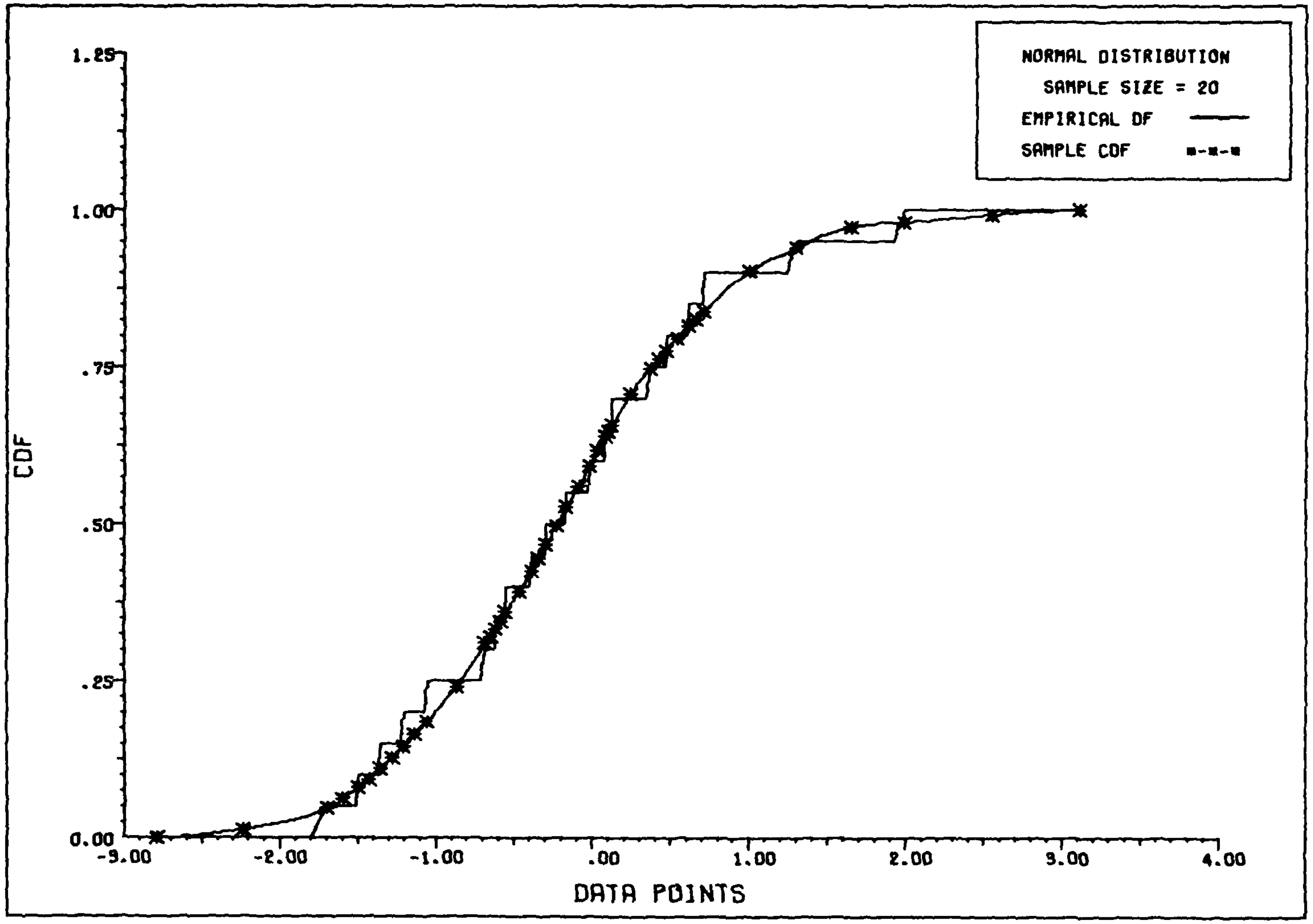

Figure 3.9. Sample CDF--Smoothed vs EDF

was necessary to restrict each set of variables to a manageable set of choices. We will rely on numerical and Monte Carlo analysis to determine the choices for our variables. No claim of optimality will be made, but we will attempt to justify our variable selections as reasonable for the situations considered. First, let us examine each set of variables and its restricted domain.

Number of Subsamples. Given an ordered sample of size n, let k be the number of subsamples generated via the method outlined earlier in this chapter. We require that $k \leq n/2$, for each subsample to contain at least two points, and also that k remains finite as n approaches infinity to satisfy the uniform convergence of the unsmoothed estimator of equation 3.6. For samples of size 100, k was initially chosen as an element of {5, 10, 15, 20}. Subsequent choices of the domain of k were made and will be identified at appropriate steps in the analysis.

Plotting Positions. Given each ordered subsample of size n*, a plotting position G_j, $j=1,\ldots,n^*$, is assigned to each order statistic. The following plotting positions were chosen from Table II.1:

1. Mean ranks

2. Median ranks

3. Midpoint of the jumps of the empirical distribution function

4. Average of the mean and mode ranks

5. Any of the above four plotting positions based on the entire sample, rather than each subsample. For example, each $Y_{(\ell,j)}$ has plotting position G_i, $i=1,\ldots,n$ associated with it where $Y_{(\ell,j)} = X_{(\ell+k(j-1))} = X_{(i)}$, the ith order statistic of the entire sample.

Extrapolation Values. For each subsample define $Y_{(0)} = Y_{(1)} - \Delta(Y_{(2)}-Y_{(1)})$ and $Y_{(n*+1)} = Y_{(n*)} + \Delta(Y_{(n*)} - Y_{(n*-1)})$ where Δ is the extrapolation value. The choices of Δ that were considered are:

1. 0, which puts a finite probability at each extreme order statistic of each subsample

2. 0.5

3. 1.0

4. 1.5

5. Choose Δ equal to the ratio $G_1/(G_2-G_1)$. This choice extrapolates the data points proportionately to their plotting positions. Since the plotting positions listed previously are symmetric, Δ is also equal to $(1-G_{n*})/(G_{n*}-G_{n*-1})$. Note that if plotting position 5 is used, then the extrapolation points are calculated only once based on the entire sample and then remain constant for each subsample.

Inversion Points. Once the subsamples are defined, we need a rule for inverting equation 3.6 to create a pseudosample. Our choices for inversion points are the first four plotting positions listed previously based on the entire sample. Thus the pseudosample $\{Z_i\}_{i=1,\ldots,N}$ is defined by $Z_i=SF^{-1}(G_i)$ where G_i is one of the four plotting conventions based on a sample of size N. Numerical calculations of $SF^{-1}(G_i)$ were accomplished via a Newton-Raphson method. Adjustments to the extreme points of the pseudosample were sometimes necessary. See Appendix 6 for a further discussion.

Number of Inversions. Since the inversion process can be repeated by creating another pseudosample, the number of repetitions needs to be determined. Due to the computational effort required and some preliminary investigation of repeated smoothing, a maximum of two inversions was considered practical. Estimators smoothed more than twice improved very little, if at all. Thus the number of inversions, I, was constrained to the set {0, 1, 2}.

Now that we have restricted our variables to manageable sets, let us now describe the procedure for selecting specific distribution function estimators by identifying particular choices of our variances. Our goal is to provide reasonable values for these variables in a limited situation in the hope of robustness over a wider class. To that end, let us consider only sample size 100 for the present. We also need a criterion for choice of the variables. A widely accepted criterion is mean integrated square error (MISE) (Refs 40, 48, 103, 104, 105). MISE = $E \int_{-\infty}^{\infty} [f(x)-\hat{f}(x)]^2 w(x)\,(dx)$, where f is the true function, $\hat{f}$ is the estimator, and w is the weight function. The integrated square error can be approximated numerically since our estimators are continuous. As a criterion, we will use an approximation to the integrated square error for both the distribution and density functions. For comparison purposes, other criteria were also used. These included Kolmogorov-Smirnov (K-S) distance, K-S integral and modified K-S integral distances, Cramer von Mises (CVM) and modified CVM integrals, Anderson-Darling (AD) and modified AD integrals and average square error (ASE). For a discussion of these criteria, see Appendix 1.

To numerically evaluate the variable choices, we also need to know the true underlying distribution. We chose three members of the Generalized Exponential Power Distribution family as our test distributions (see Appendix 2). The members chosen were the double exponential, normal, and uniform distributions. Although restricting ourselves to a symmetric family, the three members selected give three distinct measures of tail length, ranging from leptokurtic to mesokurtic to platykurtic. The density functions also possess unique central shapes--the double exponential being concave, the normal convex, and the uniform linear. As such, it was conjectured that estimators which performed well over this limited set of distributions would perform well over a much wider class.

The variable selection procedure, itself, consisted of two main steps: examination of "stylized" samples and examination of random samples. We shall deal with each in turn.

Stylized Samples. Given a sample size of 100, we generated a "stylized" sample by inverting each test distribution at the inversion points. We repeated the process for all four possible inversion values. Next, we calculated values for all of the distance criteria for the 400 combinations of the number of subsamples, plotting positions, extrapolation values and inversion points. The rationale at this stage is related to the underlying philosophy of Fisher consistency (Ref 73:281). Strict Fisher consistency requires that an estimator yield the true parameter when true proportions are realized in the sample. For our purposes, we require an estimator to be reasonably close to the true value when the input sample is stylized. Table III.1 summarizes the results of the stylized sample analysis. Four sets of variables were chosen for future consideration because of their "good" performance with respect to the modified CVM integral criterion. All three sets of variables which minimized the modified CVM integral for the distribution function

TABLE III.1

VARIABLE SETS BASED ON MODIFIED CVM INTEGRAL VALUES FOR THE DISTRIBUTION FUNCTION

	Distribution		
Variables[1]	Double Exponential	Normal	Uniform
(5,3,3,2)	$6.83\text{x}10^{-7}$	$3.78\text{x}10^{-7}$	$1.78\text{x}10^{-6}$
(5,4,3,2)	$3.28\text{x}10^{-7}$[2]	$6.19\text{x}10^{-7}$	$3.39\text{x}10^{-6}$
(5,5,3,2)	$6.91\text{x}10^{-7}$	$4.43\text{x}10^{-6}$	$1.13\text{x}10^{-9}$[2]
(5,4,5,3)	$1.32\text{x}10^{-6}$	$3.51\text{x}10^{-7}$[2]	$4.62\text{x}10^{-7}$

All entries listed are values of the modified Cramer von Mises integral of the distribution function.

Note 1: Variable sets are indexed based on their domains given earlier in this chapter. Terms correspond to (number of subsamples, plotting position, extrapolation value, inversion points).

Note 2: Minimum modified CVM integral value for that distribution.

were selected. The other set selected performed well for both the normal and double exponential distributions.

In examining the results of the stylized sample analysis, four observations were made. First, inversion points based on the median ranks outperformed the other choices. Second, plotting position 5 was clearly superior when the underlying distribution was uniform. This observation confirmed our intuition since all of the information in a sample from the uniform distribution is contained in the two extreme order statistics. Plotting position 5 uses an extrapolation scheme based on the entire sample and thus estimates the bounds of the distribution better than using extrapolated points based on the subsamples. Third, overall, the extrapolation values appeared arbitrary. Fourth, the number of subsamples determined in the "best" sets of variables seems low, probably due to the ideal spacings generated by the stylized samples. Based on these observations, we decided to fix the plotting positions, extrapolation values, and inversion points as determined by the four best variable sets. For these combinations, we now want to evaluate the functions on a limited number of random samples.

Random Samples. Given a fixed set of four combinations of plotting positions, extrapolation values, and inversion values as determined from the stylized samples, we now propose to determine choices for the number of subsamples and the number of inversions. Twenty-five random samples of size 100 from each of the test distributions were drawn and evaluated via averaged modified CVM integrals for both the distribution and density functions. Table III.2 lists the optimal choices of the sets of variables with respect to the CVM criteria. Based on the results of the random sample analysis, four conclusions were drawn: (1) there is no clear-cut optimal choice of variables across all three test distributions; (2) the optimal choice for the uniform performs poorly for the

TABLE III.2

OPTIMAL CHOICES FROM RANDOM SAMPLES

	Modified CVM Integral Values	
Variables[1]	Distribution Function	Density Function
1. Double Exponential		
A. (5,4,5,3,0)	$7.56\text{x}10^{-4}$[2]	$3.19\text{x}10^{-2}$
B. (15,4,3,2,2)	$7.80\text{x}10^{-4}$	$1.52\text{x}10^{-3}$[2]
2. Normal		
A. (25,4,3,2,1)	$1.27\text{x}10^{-3}$	$1.12\text{x}10^{-3}$[2]
B. (25,4,3,2,2)	$1.17\text{x}10^{-3}$[2]	$1.31\text{x}10^{-3}$
3. Uniform		
(25,5,3,2,2)	$5.00\text{x}10^{-4}$[2]	$1.22\text{x}10^{-3}$[2]

Note 1: Variables are listed in the same order as in Table III.1 with the last variable added being the number of inversions.

Note 2: Denotes minimum value for that criterion and distribution.

other two distributions; (3) plotting position 4, the average of the mean and mode ranks, outperformed plotting position 3, the midpoint of the jumps of the empirical distribution function, in every case; and (4) the inversion values at the median ranks outperformed the others in most cases. From these observations, we decided on forming three different models using the optimum, or nearly optimum, choices for each test distribution. Table III.3 summarizes the three models. Model 1 was developed from nearly optimum choices based on the double exponential distribution, Model 2 from the normal distribution, and Model 3 from the uniform distribution. These models were derived solely for sample size 100. Other random sample sizes were then investigated. Given random samples of size 20, 50, 175, and 250, we fixed all of the model parameters except for the number of subsamples. We also introduced a sixth pair of variables, N, the number of points to invert, and K, the number of subsamples used after an inversion. Based on twenty-five random samples from each sample size and using the modified CVM integral criterion, we developed nearly optimal selections of the number of subsamples, k, as well as N and K. Table III.4 gives the relationships between sample size and the number of subsamples for the three models based on their corresponding GEP distribution. These selections were denoted nearly optimal for two reasons. First, only a very few cases had N, the number of

TABLE III.3

NONPARAMETRIC MODELS 1, 2, AND 3

Model 1		
Number of subsamples	--	15
Plotting positions	--	average of mean and mode ranks
Extrapolation value	--	1.0
Inversion points	--	median ranks
Number of inversions	--	2
Model 2		
Number of subsamples	--	25
Plotting positions	--	average of mean and mode ranks
Extrapolation value	--	1.0
Inversion points	--	median ranks
Number of inversions	--	1
Model 3		
Number of subsamples	--	33
Plotting positions	--	median ranks of the entire sample
Extrapolation value	--	1.0
Inversion points	--	median ranks
Number of inversions	--	2

All models are valid for sample size 100 only.

TABLE III.4

NUMBER OF SUBSAMPLES VERSUS SAMPLE SIZE

Model	Sample Size (n)	Number of Subsamples (k)	Number of Inversion Points (N)	Number of Subsamples (K)
1	20	5	20	5
	50	10	50	10
	100	15	100	15
	175	30	100	15
	250	45	100	15
2	20	10	20	10
	50	25	50	25
	100	25	100	25
	175	35	100	25
	250	50	100	25
3	20	10	20	10
	50	25	50	25
	100	33	100	33
	175	80	100	33
	250	125	100	33

inversion points, greater than 100 as the optimal choice. The difference in the CVM criteria for the optimal choice and the value listed in Table III.4 was insignificant. For example, for sample size 50 using Model 3, the range of values for the modified CVM integral was [.00088, .00190] for the distribution function and [.00189, .00760] for the density function. The actual values chosen correspond to .00088 and .00190 for the distribution and density functions respectively. Thus, the decrease in the criteria did not justify the added computational effort to invert more than 100 points. The number of points in each pseudosample, N, was defined using the following algorithm:

$$N = \begin{cases} 20 & n \leq 20 \\ n & 20 < n < 100 \\ 100 & n \geq 100 \end{cases}$$

The number of subsamples for the pseudosample, K, was defined to be the corresponding k for n=N. Second, due to the high variability of such a small Monte Carlo sample size, we again opted for reasonable values which followed a generally regular trend.

The number of subsamples for sample sizes not listed in Table III.4 was arbitrarily determined by constructing step functions for each model such that the average number of points in each subsample followed a near linear interpolation through the k versus n points listed in the table. For sample sizes greater than 250, we use the value of k for n=250. This choice allows the models to exhibit the uniform convergence property shown earlier in this chapter since the number of subsamples stays finite. Figures 3.10, 3.11, and 3.12 show the plots of k versus n for the three models. Figure 3.13 shows the k-n relationship for model 2* developed in conjunction with an adaptive procedure discussed in the next section. Table III.5 shows the relationship of the average number of points in each subsample to the sample size for the three models.

Adaptive Approaches

Each of the three models generated in the previous section was based on stylized and random samples from a specific distribution. The variables for Models 1, 2, and 3 were chosen by comparison with the double exponential, normal, and uniform distributions respectively. While the models are strictly nonparametric and perform well given a specific underlying distribution, their performance for an unknown distribution is yet undetermined.

Since the three members of the GEP distribution represent vast differences in shapes and tail length, and since each nonparametric model proposed has been associated with a specific member of the GEP family, it became a natural extension to consider a nonparametric adaptive model using the three models already developed.

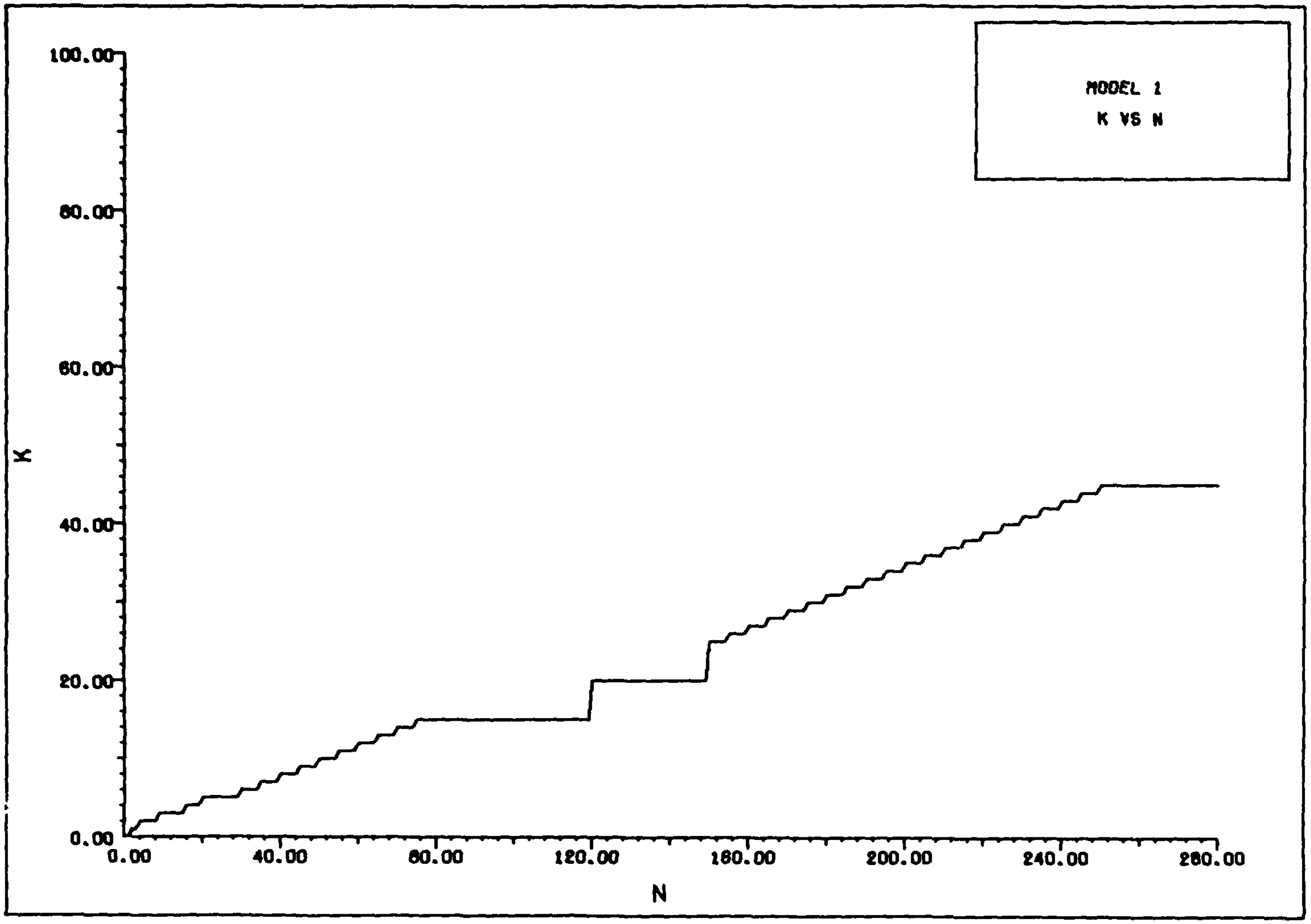

Figure 3.10. k vs n Plot--Model 1

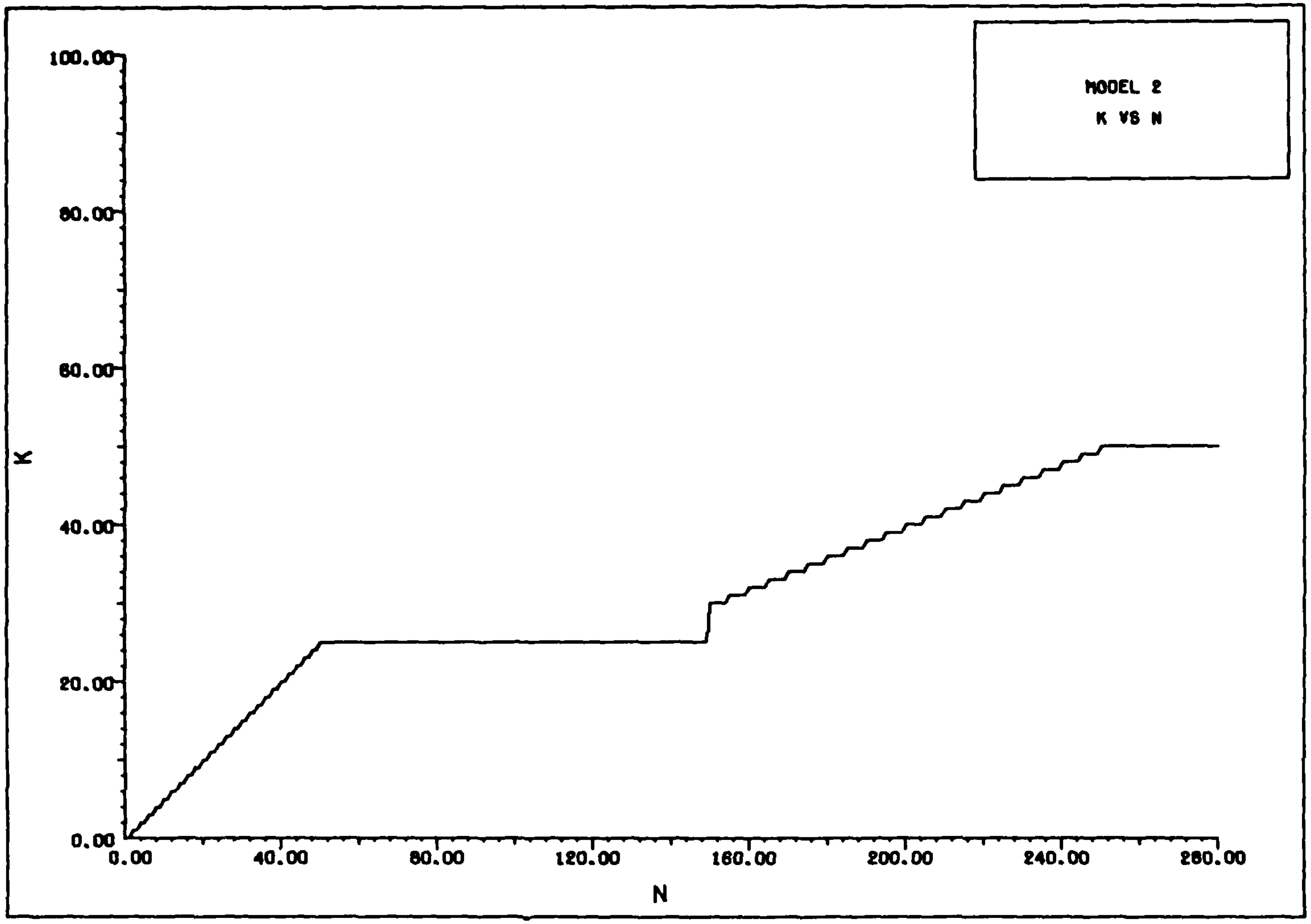

Figure 3.11. k vs n Plot--Model 2

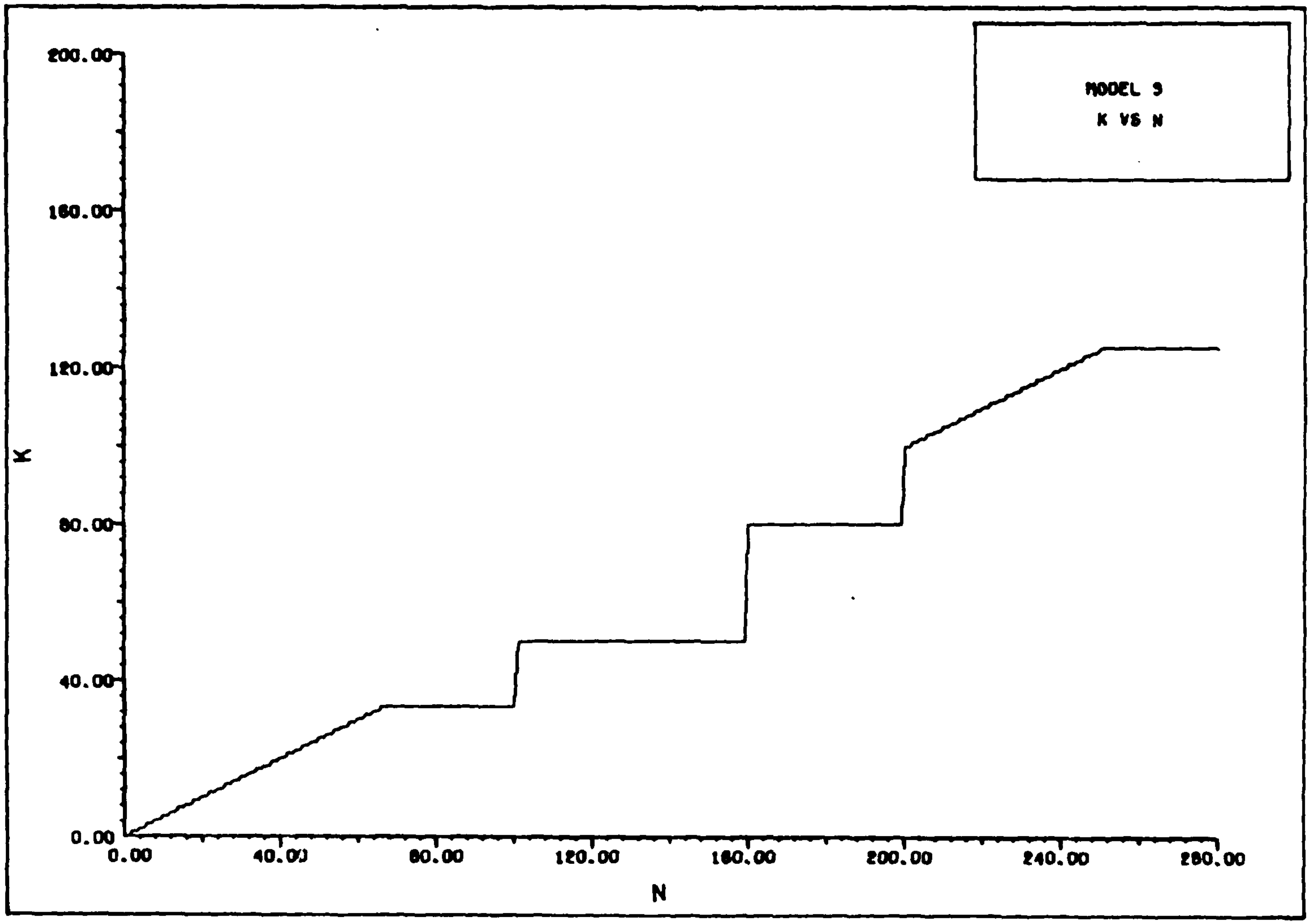

Figure 3.12. k vs n Plot--Model 3

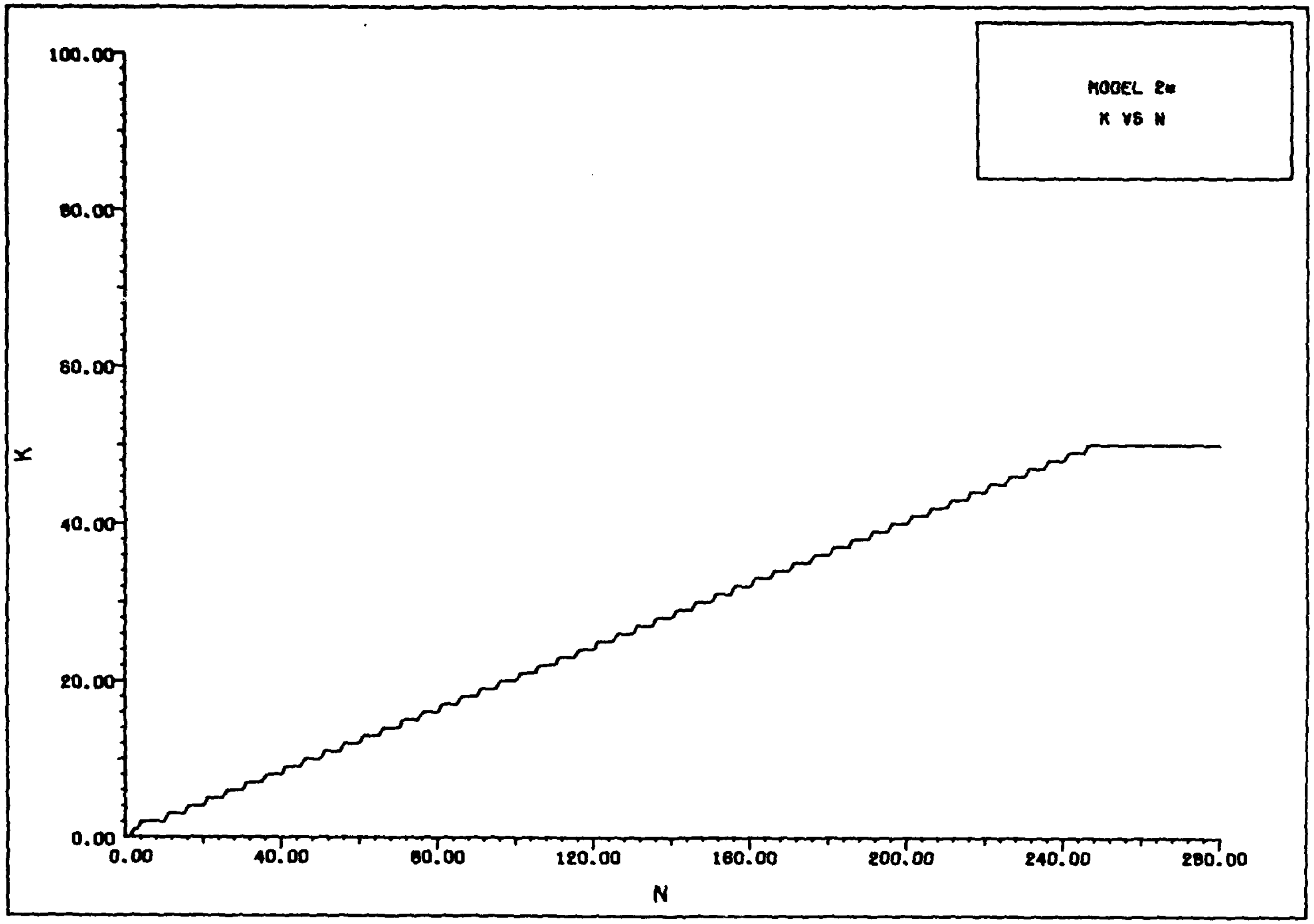

Figure 3.13. k vs n Plot--Model 2*

TABLE III.5

SELECTED VALUES OF k AND n FOR THE NONPARAMETRIC MODELS

Sample Size	Model 1		Model 2		Model 3		Model 2*	
(n)	k	n/k	k	n/k	k	n/k	k	n/k
5	2	2.5	2	2.5	2	2.5	2	2.5
10	3	3.33	5	2.0	5	2.0	2	5.0
15	3	5.0	7	2.14	7	2.14	3	5.0
20	5	4.0	10	2.0	10	2.0	4	5.0
25	5	5.0	12	2.08	12	2.08	5	5.0
50	10	5.0	25	2.0	25	2.0	10	5.0
75	15	5.0	25	3.0	33	2.27	15	5.0
100	15	6.67	25	4.0	33	3.33	20	5.0
150	25	6.0	30	5.0	50	3.0	30	5.0
200	35	5.71	40	5.0	100	2.0	40	5.0
250	45	5.56	50	5.0	125	2.0	50	5.0

To develop such a model, we need a discriminant. In the case of symmetric distributions, three discriminants based on tail length have been used: kurtosis, Hogg's Q statistic, and percentile ratios. Applications of the discriminants in parametric estimation problem can be found in Andrews, et al., Daniels, Harter, et al., Hogg, McNeese, and Moore, to name only a few (Refs 5, 17, 34, 38, 55, 60). For our purposes, we do not wish to restrict ourselves to modeling only symmetric populations. Both kurtosis and Hogg's Q statistic are not compatible with the asymmetric case. They tend to average the measures of both upper and lower tail length. However, it is possible to use percentile ratios as a discriminant for each tail individually. Thus, we can, heuristically at least, envision a model which could adequately portray a leptokurtic tail on one end and a platykurtic tail on the other.

Percentile Ratios. Let F be a continuous distribution function. Now define the lower and upper percentile ratios, PL and PU as follows:

$$PL = \frac{F^{-1}(.5) - F^{-1}(.025)}{F^{-1}(.5) - F^{-1}(.25)}$$

$$PU = \frac{F^{-1}(.975) - F^{-1}(.5)}{F^{-1}(.975) - F^{-1}(.75)}$$

By construction PL and PU are greater than or equal to unity. Table III.6 lists the lower and upper percentile ratios for some common distributions.

The next step was to examine the distributions of the percentile ratios themselves. We approximated these distributions by our nonparametric models. Monte Carlo samples of size 20, 50, 100, 175, 250, and 500 were drawn from each of the three GEP test distributions. The lower percentile ratio was then calculated. The process was repeated 100 times to get 100 values of PL for each sample size and test distribution. This is equivalent to 100 values of PU since the random samples were drawn from symmetric populations. We then used our nonparametric models to generate approximate distribution functions for PL (or PU) at each test distribution and sample size. Model 1 was used for the distribution of the percentile ratios computed from uniform and double exponential random samples. Model 2 was used for the distribution computed from normal random samples. Selection of these models was based on both graphical characteristics and the sample percentile ratios. At this point we imposed two constraints. First, since Model 3 tended to perform poorly if the true distribution was not uniform, we shall only use Model 3 when the sample strongly suggests a shape resembling the uniform. Let SPR be the sample percentile ratio, either lower or upper, and let PR_1 and PR_2 be the values of the

TABLE III.6

POPULATION PERCENTILE RATIOS

	Percentile Ratios	
Distribution	Lower	Upper
Normal	2.904	2.904
Uniform	1.900	1.900
Double Exponential	4.322	4.322
Triangular	2.651	2.651
Cauchy	12.706	12.706
Exponential	1.647	4.322
Weibull (2)	2.274	3.155
Weibull (3)	2.630	2.870
Beta (1, 2)	1.764	2.651
Beta (½, ½)	1.409	1.409
Largest Extreme Value	2.410	3.764

Shape parameters are given in parentheses. Triangular distribution has support [-2,2] Beta distribution has support [0,1]. All other distributions have been standardized with location parameter zero and scale parameter one.

percentile ratio where the adaptive procedure switches models. We set $P(SPR < PR_1 \mid \text{uniform distribution}) = .5$. Second, since both Models 1 and 2 perform reasonably well for both the double exponential and the normal distributions, set $P(SPR < PR_2 \mid \text{double exponential distribution}) = P(SPR > PR_2 \mid \text{normal distribution})$. Thus, we equate the probabilities of an incorrect choice. Based on these two constraints and our nonparametric distribution functions, we solved for PR_1 and PR_2 across all sample sizes considered. Values derived were PR_1=1.9 and PR_2=3.5. Table III.7 lists the approximate probabilities for the sample lower percentile ratio falling in any of the three intervals defined by PR_1 and PR_2 for the three underlying distributions and various sample sizes.

The construction of our nonparametric estimators allows the use of only one model for each sample considered. Having two different percentile ratios creates an ambiguity as to which model to finally choose. We resolved this dichotomy in two ways. First, Model 1 seemed to perform better when the underlying population was normal than Model 2 performed if the underlying population was double exponential. So, we chose Model 1 if both Models 1 and 2 are indicated. Actually, it turns out that the model number is its relative order of precedence. Second, we discovered that the uniform distribution could also be approximated well by using either Models 1 or 2 and

TABLE III.7

SELECTED PROBABILITIES--LOWER PERCENTILE RATIO (PL)

Sample	UNIFORM DISTRIBUTION		
Size	P(PL≤1.9)	P(1.9< PL< 3.5)	P(PL≥3.5)
20	.4326	.5025	.0649
50	.5178	.4738	.0084
100	.5541	.4428	.0031
175	.5085	.4915	0
250	.5544	.4456	0
500	.4881	.5119	0
Sample	NORMAL DISTRIBUTION		
Size	P(PL≤1.9)	P(1.9<PL<3.5)	P(PL≥3.5)
20	.0994	.5711	.3295
50	.0354	.7273	.2373
100	.0350	.7992	.1658
175	.0080	.8753	.1167
250	.0068	.9295	.0637
500	0	.9658	.0342
Sample	DOUBLE EXPONENTIAL DISTRIBUTION		
Size	P(PL≤1.9)	P(1.9<PL<3.5)	P(PL≥3.5)
20	.0592	.2715	.6693
50	.0231	.1851	.7918
100	.0026	.1594	.8380
175	.0012	.1222	.8766
250	.0013	.0972	.9015
500	0	.0375	.9625

forcing the extrapolated points for each subsample to be constants. These points are based on extrapolation from the entire sample.

From the previous three models and the fixed extrapolation point modification, Models 4 and 5 were developed. Model 4 uses the first three models depending on the values of the sample percentile ratios. Model 5 uses only Models 1 and 3.

In analyzing the relationship of k, the number of subsamples, and n, the sample size, it was evident from a graphical standpoint that the ratio of k/n determined how much detail the approximation possessed. So a choice of a nominal ratio of k/n seemed appealing. Since Models 1 and 2 performed reasonably well for double exponential and normal random samples, we postulated another model which is a compromise between the two in the sense of the k/n ratio. We chose the simple expression:

$$k = \begin{cases} \frac{n+4}{5} & n \leq 250 \\ 50 & n > 250 \end{cases}$$

Thus, for samples of size 250 or less, each subsample contains either 4 or 5 data points. Like Model 2, we kept the number of inversions at one. Denote this new model as Model 2* since, with the exception of the new choice of k, it uses the same variables as Model 2. An adaptive procedure, Model 6, was based on Models 2* and 3. A summary of all three adaptive models is given in Table III.8.

Summary

This chapter has traced the derivation of a nonparametric, continuous, differentiable, sample distribution function. First, we considered a simple scheme to extend plotting positions to a continuous, differentiable function. Then, we improved on our distribution and density estimators by the use of averaging functions based on subsamples, similar to the jackknife. Next we investigated the properties of uniform convergence and of distribution functions as they apply to our new estimators. Theorem 3.6 concludes the uniform convergence arguments. A smoothing routine, which again preserves the distribution function properties, was introduced. Next, a detailed analysis of stylized and random samples from representative members of the Generalized Exponential Power distribution resulted in selection of three initial nonparametric models. With the addition of the percentile ratios as discriminants of tail length, three adaptive models were then defined. Having completed the theoretical development of our six chosen models, our next goal is an evaluation and comparison of these techniques as estimators.

TABLE III.8

DECISION RULES FOR ADAPTIVE MODELS

Percentile Ratios		
Lower	Upper	Model 4
[1.0,1.9)	[1.0,1.9)	Model 3
[1.0,1.9)	[1.9,3.5]	Model 2--fixed $X_{(0)}$
[1.0,1.9)	(3,5,∞)	Model 1--fixed $X_{(0)}$
[1.9,3.5]	[1.0,1.9)	Model 2--fixed $X_{(n+1)}$
[1.9,3.5]	[1.9,3.5]	Model 2
[1.9,3.5]	(3.5,∞)	Model 1
(3.5,∞)	[1.0,1.9)	Model 1--fixed $X_{(n+1)}$
(3.5,∞)	[1.9,3.5]	Model 1
(3.5,∞)	(3.5,∞)	Model 1

Percentile Ratios		
Lower	Upper	Model 5
[1.0,1.9)	[1.0,1.9)	Model 3
[1.0,1.9)	[1.9,∞)	Model 1--fixed $X_{(0)}$
(1.9,∞)	[1.0,1.9)	Model 1--fixed $X_{(n+1)}$
(1.9,∞)	(1.9,∞)	Model 1

Percentile Ratios		
Lower	Upper	Model 6
[1.0,1.9)	[1.0,1.9)	Model 3
[1.0,1.9)	[1.9,∞)	Model 2*--fixed $X_{(0)}$
(1.9,∞)	[1.0,1.9)	Model 2*--fixed $X_{(n+1)}$
(1.9,∞)	(1.9,∞)	Model 2*

IV. Distribution and Density Function Estimation

Introduction

Having constructed six nonparametric models, we now propose to evaluate their performance and demonstrate their feasibility. We begin by surveying several other authors' estimates of the distribution function, both continuous estimates and step functions. Estimates of the density function are then examined. These include kernel estimates, orthogonal series estimates, delta sequences and a more recent entropy based estimate. The new nonparametric estimators are then compared on the basis of mean integrated square error of both density and distribution functions. Tables are given which list the results of Monte Carlo comparisons of the models over six distributions and six sample sizes. The results were compared with two other continuous density approximations. Convergence rates for the estimators are also approximated. Next some specific examples of the models are shown plotted for five different distributions. Finally the hazard function is estimated and plotted. As a tool, the hazard function, coupled with the density and distribution functions form a powerful discriminant of density types.

Historical Survey

Distribution Function Estimation. We have already examined some estimates of distribution functions in our discussion of sample distribution functions in Chapter II. Some were rather general, like Vogt's variant of the empirical distribution function, while others, like Schuster's, were concerned with reflecting points about the estimated location parameter of a symmetric distribution. The references in Chapter II describe rather simple step function approaches to estimating the distribution function.

Several other methods also merit discussion. While his estimate is still a step function, Turnbull developed an algorithm to calculate the maximum likelihood estimate $\hat{F}$ of an underlying distribution function F. He shows monotonic convergence of his algorithm to $\hat{F}$ and indicates an application to hypothesis testing, while considering data sets which are arbitrarily grouped, censored or truncated (Ref 97). For an average squared error loss function, Phadia showed that a step function estimator $\tilde{F}(t)$ is minimax.

$$\tilde{F}(t) = \frac{1}{2(m+1)} + \frac{1}{m(m+1)} \sum_{i=1}^{n} \delta_{X_i}(-\infty,t)$$

where $m = \sqrt{n}$ and δ_{X_i} is a measure on R^1 which assigns a unit mass to X_i. He further derived step function estimators which are best invariant and also best invariant confidence bands (Ref 67).

Continuous functions have also been developed. Smaga derives a smooth empirical distribution function in a manner similar to kernel estimates for a probability density (Ref 86). Orthogonal series estimators, based on trigonometric functions proposed by Kronmal and Tarter give a continuous approximation for the distribution function. Their Fourier series method produced impressive mean integrated square error values. A significant drawback to the method is the lack of distribution function properties of these estimators (Refs 40, 48).

While we are primarily concerned with nonparametric estimation, some rather general three or four parameter families of distributions can be used to approximate a distribution function. Recently, one such four parameter family was introduced by Ramberg, et al. Based on a generalization of Tukey's lambda function, this new distribution approximates a wide range of both symmetric and asymmetric populations (Ref 72).

In addition to the estimating methods presented both in this chapter and in Chapter II, the approaches to density estimation given in the next section provide the opportunity for further distribution function estimation. As we have seen, some authors attack the general problem of data modeling by investigating the distribution function.

We now consider those who chose a path of density function estimation.

Density Function Estimation. Oldest among the density function estimates is the histogram. Given a set of class intervals, the histogram is a maximum likelihood estimator. This dependence on internal selection, however, is a serious drawback. While the method of maximum likelihood has been a classical technique, recently the minimum distance method developed by Wolfowitz has inspired numerous articles, particularly in the sense of parametric estimation (Ref 108). Reiss proposes minimum distance estimators of unimodal densities. He proves consistency and gives a computational algorithm. Using the empirical distribution function and the Kolmogorov-Smirnov distance measures, Reiss' estimators are defined as constants between ordered sample data points. As such, the estimators are actually minimum distance histograms (Ref 74).

Since 1956, some significant continuous approximations have emerged. Much of the literature has been devoted to kernel estimators, first developed by Rosenblatt (Ref 75). Most of the important results are summarized in a recent book by Tapia and Thompson (Ref 94). Wegman and Davies discuss two recursive estimators closely related to kernel estimators. They also propose a sequential estimation procedure based on the recursive estimators (Ref 106).

Singh evaluates the mean square errors of a density estimator of the kernel type and its derivatives (Ref 85). Some further properties of kernel estimators are proposed by Schuster (Ref 81). Fourier inversion method of density estimation is proposed by Blum and Susarla. They show this estimator possesses mean square consistency and asymptotic normality (Ref 8).

Various estimation techniques based on orthogonal series expansions have also been developed. Kronmal and Tarter proposed estimators of both distribution and density functions using Fourier series. Expressions for the mean integrated square error are developed in terms of the variances of the Fourier coefficients. Both Schwartz and Walter evaluate the properties of a density estimator based on Hermite functions which are defined in terms of Hermite polynomials (Refs 84, 100). Watson proposes another orthogonal series estimator (Ref 102). Crain uses the set of normalized Legendre polynomials on [-1,1] as his orthogonal set. He incorporates both a restricted maximum likelihood approach and the information-theoretic distance defined by Kullback (Ref 14).

Watson and Ledbetter defined a density estimator as an average of square integrable functions. Expressions for these functions are derived based on a mean integrated square error criterion (Ref 103). Walter and Blum generalized many of the previously mentioned methods into one method based on "delta sequences," sequences of functions which converge to a generalized function δ. This delta sequence method includes kernel estimators, orthogonal series estimators, Fourier transform estimators and histograms (Ref 101). Convergence rates are also generalized from the results of Wahba (Ref 99).

Parzen has attempted to incorporate both parametric and nonparametric schemes in an approach to data modeling. He also introduces density quantile functions and a method of autoregressive density estimation (Ref 65).

Entropy approaches have also been suggested to estimate probability densities. MacQueen and Marschak discuss the rationale for using a maximum entropy approach to estimate Bayesian prior distributions (Ref 52). Miller, using the maximum entropy formalism given by Tribus (Ref 95), approximates a density function as a member of the exponential family of distributions, F. Miller's approximations are shown to be within computational accuracy when the underlying distribution is a member of F and accurate average values of the "information functions" are available (Ref 57).

Estimator Comparisons

Having examined previous distribution and density function estimators, we now wish to evaluate the new nonparametric estimators proposed in Chapter III. We begin by examining the criteria for comparison. Next we discuss the mechanics of the Monte Carlo study. Finally, we shall present the results and conclusions of the comparisons.

Criteria. To derive the various variables which make up our models, we previously used a modified CVM integral criterion. Here we will use this same criterion to evaluate the estimators. As mentioned in Appendix 1, this modified Cramer von Mises integral approximates the average square error and mean integrated square error (MISE) with weight function f.

If we restrict ourselves to the family of continuous distribution functions, F, which can be parameterized by location and scale parameters, we can show by construction that SF(x) belongs to F. Further, with respect to the distribution functions as the arguments, the modified KS integral, modified CVM integral and modified Anderson-Darling (AD) integral are all location and scale invariant. When the density functions are used in the arguments of these integrals, location invariance is preserved, but scale invariance is not. For example, let X be a random variable from a standard normal distribution. Now let Y = X/σ. Choose a random sample $\{X_i\}$ i=1,...,n and form $\{Y_i\}$ i=1,...,n. Now let $SF_X(x)$ and $sf_X(x)$ be the nonparametric approximations based on the sample $\{X_i\}$ i=1,...,n, and similarly for Y. Then

$$\int(f_Y(y)-sf_Y(y))^2\, dSF_Y(y) = \sigma^2 \int(f_X(x)-sf_X(x))^2 dSF_X(x).$$

Given the modified CVM integral value for a standardized distribution, we can compute the integral for another random variable with a different scale factor but the same distribution type.

Monte Carlo Mechanics. With our criteria defined we now generated random samples via the methods discussed in Appendix 3. Twenty-five samples of sizes 20, 50, 100, 175, 250 and 500 were drawn from each underlying distribution. These distributions included the double exponential, normal, uniform, triangular, Cauchy, and exponential. To keep a consistent comparison with other published results, the uniform and triangular distributions were defined on [0,1]. All other distribution functions had a zero location parameter and unit scale parameter. Each random sample was compared with nonparametric models 1 through 6. Values for both the MISE of the distribution function and density function were approximated by averaging the twenty-five modified CVM integrals. A standard error of each estimate was also calculated. As a numerical check, the average square errors were also calculated and were in close agreement with the modified CVM criterion.

Results. Tables IV.1 through IV.8 summarize the main results of the Monte Carlo study. Although a small Monte Carlo sample size was used, relative comparisons among the nonparametric models developed here can be made.

The same random samples were used to calculate the modified CVM integrals for each model. Tables which give approximate MISE also include the standard error of the estimate beneath each entry to give a measure of the Monte Carlo accuracy.

Table IV.1 shows a comparison among all six models using the approximate MISE of the distribution function for sample size 100. The last column lists the mean of the asymptotic distribution of the Cramer von Mises statistic, W^2, normalized by the sample size (Ref 4). This value is the MISE of the distribution function when the empirical distribution function is used as the estimator. Note that in all cases except for the Cauchy distribution, Models 1, 2 and the three adaptive models outperform the empirical distribution function in terms of MISE. Given an underlying uniform distribution, Model 3 is the clear choice. However, its poor performance for other distributions results from the fixed plotting positions based on the entire sample. The excellent performance of the adaptive models for the distributions considered is especially encouraging. These results indicate that, on the average, our nonparametric models are closer to the true distribution function than the empirical distribution function under the criterion of mean integrated square error.

TABLE IV.1

APPROXIMATE MISE--DISTRIBUTION FUNCTION--SAMPLE SIZE = 100

	Type of Estimate						
Distribution	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	$E(W^2)/n$
Double Exponential	.00080 *.00078*	.00092 *.00087*	.01642 *.00411*	.00080 *.00078*	.00080 *.00078*	.00085 *.00080*	.00167
Normal	.00136 *.00139*	.00121 *.00122*	.00663 *.00319*	.00131 *.00139*	.00136 *.00139*	.00126 *.00138*	.00167
Uniform	.00113 *.00079*	.00113 *.00074*	.00044 *.00040*	.00105 *.00074*	.00106 *.00080*	.00093 *.00076*	.00167
Triangular	.00110 *.00134*	.00099 *.00123*	.00267 *.00109*	.00099 *.00123*	.00110 *.00134*	.00103 *.00129*	.00167
Cauchy	.00192 *.00155*	.00243 *.00184*	.05176 *.01204*	.00192 *.00155*	.00192 *.00155*	.00205 *.00163*	.00167
Exponential	.00123 *.00080*	.00160 *.00101*	.01182 *.00468*	.00135 *.00093*	.00122 *.00082*	.00121 *.00085*	.00167

For the density functions, a direct comparison of our models with the estimators evaluated by Wegman was made. We chose only to repeat the two continuous density estimators tested, the naive estimator based on a uniform kernel and the trigonometric estimator of Kronmal and Tarter. For average square error values of histogram estimators, refer to Wegman (Ref 105). Table IV.2 gives the approximate MISE values for the density estimators. Note the competitive performance of our models of the density functions. No one estimator is clearly superior. Again the performance of the adaptive models is encouraging.

Remember that the motivation for the development of this new nonparametric family of estimators was based on modeling the distribution functions. The density estimators are merely analytic derivatives of these distribution functions. Since differentiation is an unbounded linear operator, one would suspect a large discrepancy between a differentiated estimate and one specifically designed to model the density function itself. The comparable performance of these new models against pure density estimators demonstrates their versatility.

It should also be noted that the trigonometric estimator introduced negative density values in samples from the normal, Cauchy and exponential distributions. Although the trigonometric density estimates do integrate to unity over their finite support, usually the interval

TABLE IV.2

APPROXIMATE MISE--DENSITY FUNCTION--SAMPLE SIZE = 100

	Type of Estimate							
Distribution	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Kernel[1]	Trigonometric[1]
Double Exponential	.00250	.00259	.02492	.00250	.00250	.00235	-	-
	.00152	*.00159*	*.00059*	*.00152*	*.00152*	*.00145*	-	-
Normal	.00228	.00156	.00942	.00206	.00228	.00184	.0012	.0012
	.00188	*.00136*	*.00151*	*.00178*	*.00188*	*.00161*	*.0010*	*.0012*
Uniform on [0,1]	.06845	.06481	.01387	.06438	.06268	.05964	.0439	.0297
	.01441	*.01157*	*.00438*	*.01989*	*.01881*	*.02459*	*.0187*	*.0480*
Triangular on [0,1]	.04486	.02806	.14131	.02806	.04486	.03488	.0322	.0439
	.04705	*.03001*	*.02621*	*.03001*	*.04705*	*.04031*	*.0177*	*.0319*
Cauchy	.00100	.00141	.00290	.00100	.00100	.00109	.0010	.0169
	.00058	*.00082*	*.00135*	*.00058*	*.00058*	*.00063*	*.0006*	*.0092*
Exponential	.03241	.03367	.01200	.02440	.02415	.02278	.0615	.0116
	.00571	*.00541*	*.00199*	*.00915*	*.00874*	*.00950*	*.0093*	*.0158*

Note 1: Values taken from Ref 105, Table II.

$[X_{(1)}, X_{(n)}]$, their utility is diminished by the negative values. Conversely, both the kernel estimator, when the kernel itself is chosen as a density function, and all of the new nonparametric models do possess all the properties of distribution functions.

The addition of the exponential distribution as an asymmetric example is significant. The performance of the adaptive models for both the distribution function and density function indicate that the new nonparametric approach also performs well over a very general class of probability distributions.

A further comparison of the density estimators was made for various sample sizes using the triangular distribution. Table IV.3 lists the values of the approximate MISE and the standard errors. The competitive nature of the new models, particularly the adaptive ones, is again evident. Tables IV.4 through IV.7 show the performance of Models 5 and 6 for various sample sizes and distributions. Both the MISEs for the distribution function and the density function are compared. Tables IV.4 and IV.6 include the mean of the asymptotic distribution of the normalized CVM statistic as a reference. These two models are significant in that they will form the bases for goodness of fit tests proposed in the next chapter.

TABLE IV.3

APPROXIMATE MISE--TRIANGULAR DISTRIBUTION ON [0,1]--DENSITY FUNCTION

	Type of Estimate						
Sample Size	Model 1	Model 2	Model 4	Model 5	Model 6	Kernel[1]	Trigonometric[1]
20	.11573 *.11236*	.05494 *.04106*	.10260 *.10890*	.11949 *.11946*	.19272 *.17059*	- -	- -
50	.06113 *.04551*	.03386 *.02552*	.04449 *.03342*	.06325 *.04769*	.07459 *.05308*	.0531 *.0380*	.0655 *.0680*
100	.04486 *.04705*	.02806 *.03001*	.02806 *.03001*	.04486 *.04705*	.03488 *.04031*	.0322 *.0177*	.0439 *.0319*
175	.03267 *.02348*	.02325 *.01819*	.02325 *.01819*	.03267 *.02348*	.02569 *.01933*	.0228 *.0110*	.0208 *.0143*
250	.02310 *.01765*	.01651 *.01206*	.01651 *.01206*	.02310 *.01765*	.01811 *.01365*	.0205 *.0130*	.0204 *.0174*
500	.01121 *.00511*	.00876 *.00332*	.00876 *.00332*	.01121 *.00511*	.00942 *.00459*	.0083 *.0030*	.0133 *.0074*

Note 1: Values taken from Ref 105, Table I.

TABLE IV.4

APPROXIMATE MISE--DISTRIBUTION FUNCTION--MODEL 5

	Distribution						
Sample Size	Double Exponential	Normal	Uniform	Triangular	Cauchy	Exponential	$E(W^2)/n$
20	.00608 *.00471*	.00770 *.00622*	.00487 *.00489*	.00521 *.00485*	.00915 *.00791*	.00742 *.00556*	.00833
50	.00219 *.00244*	.00318 *.00388*	.00196 *.00233*	.00262 *.00333*	.00425 *.00295*	.00239 *.00336*	.00333
100	.00080 *.00078*	.00136 *.00139*	.00106 *.00080*	.00110 *.00134*	.00192 *.00155*	.00122 *.00082*	.00167
175	.00074 *.00055*	.00080 *.00078*	.00078 *.00059*	.00074 *.00077*	.00128 *.00062*	.00107 *.00095*	.00095
250	.00064 *.00080*	.00054 *.00052*	.00081 *.00066*	.00053 *.00060*	.00106 *.00070*	.00097 *.00098*	.00067
500	.00027 *.00025*	.00024 *.00022*	.00042 *.00028*	.00027 *.00020*	.00101 *.00068*	.00049 *.00040*	.00033

AD-A115 491   AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOO--ETC F/G 12/1
NONPARAMETRIC ESTIMATION OF DISTRIBUTION AND DENSITY FUNCTIONS --ETC(U)
MAY 82 J SWEEDER
UNCLASSIFIED AFIT/DS/MA/82-1 NL

2 OF 3
AD A
115491

TABLE IV.5

APPROXIMATE MISE--DENSITY FUNCTION--MODEL 5

	Distribution					
Sample Size	Double Exponential	Normal	Uniform on [0,1]	Triangular on [0,1]	Cauchy	Exponential
20	.01548	.01153	.13181	.12949	.00282	.04370
	.01853	*.01644*	*.12691*	*.11946*	*.00132*	*.02701*
50	.00467	.00424	.07178	.06325	.00202	.02831
	.00307	*.00471*	*.04590*	*.04769*	*.00124*	*.01507*
100	.00250	.00228	.06268	.04486	.00100	.02415
	.00152	*.00188*	*.01881*	*.04705*	*.00058*	*.00874*
175	.00223	.00138	.04864	.03267	.00062	.01885
	.00202	*.00121*	*.02925*	*.02348*	*.00031*	*.01162*
250	.00164	.00084	.04595	.02310	.00049	.01758
	.00111	*.00061*	*.02690*	*.01765*	*.00027*	*.00756*
500	.00106	.00049	.03723	.01121	.00048	.01210
	.00079	*.00027*	*.01546*	*.00511*	*.00027*	*.00474*

TABLE IV.6

APPROXIMATE MISE--DISTRIBUTION FUNCTION--MODEL 6

	Distribution						
Sample Size	Double Exponential	Normal	Uniform	Triangular	Cauchy	Exponential	$E(W^2)/n$
20	.00609 *.00479*	.00763 *.00583*	.00480 *.00509*	.00521 *.00474*	.00724 *.00667*	.00622 *.00472*	.00833
50	.00211 *.00242*	.00328 *.00390*	.00193 *.00243*	.00270 *.00334*	.00373 *.00286*	.00209 *.00306*	.00333
100	.00085 *.00080*	.00126 *.00138*	.00093 *.00076*	.00103 *.00129*	.00205 *.00163*	.00121 *.00085*	.00167
175	.00080 *.00058*	.00081 *.00077*	.00068 *.00052*	.00068 *.00074*	.00133 *.00063*	.00103 *.00094*	.00095
250	.00070 *.00084*	.00052 *.00052*	.00069 *.00061*	.00050 *.00058*	.00120 *.00072*	.00088 *.00098*	.00067
500	.00029 *.00026*	.00023 *.00020*	.00036 *.00025*	.00026 *.00021*	.00091 *.00064*	.00042 *.00037*	.00033

TABLE IV.7

APPROXIMATE MISE--DENSITY FUNCTION--MODEL 6

	Distribution					
Sample Size	Double Exponential	Normal	Uniform on [0,1]	Triangular on [0,1]	Cauchy	Exponential
20	.02582 *.04060*	.01788 *.02496*	.15562 *.16928*	.19272 *.17059*	.00306 *.00185*	.06713 *.06474*
50	.00521 *.00316*	.00494 *.00525*	.06950 *.05329*	.07459 *.05308*	.00217 *.00189*	.02842 *.01757*
100	.00235 *.00145*	.00184 *.00161*	.05964 *.02459*	.03488 *.04031*	.00109 *.00063*	.02278 *.00950*
175	.00237 *.00184*	.00124 *.00103*	.03925 *.02264*	.02569 *.01933*	.00072 *.00037*	.01660 *.01247*
250	.00190 *.00123*	.00068 *.00051*	.03947 *.02313*	.01811 *.01365*	.00063 *.00032*	.01403 *.00799*
500	.00123 *.00086*	.00043 *.00020*	.03171 *.01316*	.00942 *.00459*	.00047 *.00029*	.00810 *.00447*

Based on the calculated criterion values, we derived empirical convergence rates for five of the models. Normalized to criterion values at sample size 50, Table IV.8 compares the empirical rates to convergence rates of order $n^{-.5}$, $n^{-.8}$, and n^{-1}. The distribution function models appear to converge at a rate near n^{-1}. This empirical result indicates that the smoothing process introduced in Chapter III does not appreciably affect the convergence of the estimators. Recall that the unsmoothed estimators displayed uniform convergence. Now, we have empirical evidence of the convergence of our distribution function models. The density function estimates appear to converge at a rate between $n^{-.5}$ and $n^{-.8}$. This rate is not as rapid as the theoretical convergence rate of the kernel estimate given by Rosenblatt or the approximate convergence rate for the trigonometric estimate given by Wegman (Refs 75 and 105). However, we have demonstrated empirical convergence of our density estimators, a property not analytically verifiable due to the differentiation operation. While the convergence rates appear somewhat slower, the previous tables show that the actual criterion values of our model estimators are very close to the methods currently available. Further, the use of nonparametric estimates for very large samples is a questionable procedure. Large samples are ideally suited to a parametric approach, since the amount of information available

TABLE IV.8

EMPIRICAL CONVERGENCE RATES

A. DISTRIBUTION FUNCTION						Rate	Rate	Rate
Sample Size	Model 1	Model 2	Model 4	Model 5	Model 6	$o(n^{-.5})$	$o(n^{-.8})$	$o(n^{-1})$
100	.4775	.4021	.4654	.4718	.4438	.7071	.5743	.5000
175	.3235	.2815	.3020	.3062	.2963	.5345	.3671	.2857
250	.2658	.2292	.2454	.2539	.2414	.4472	.2759	.2000
500	.1248	.1165	.1217	.1204	.1139	.3162	.1585	.1000
B. DENSITY FUNCTION								
100	.7244	.5625	.6992	.7009	.5717	.7071	.5743	.5000
175	.5867	.4736	.4877	.5148	.4117	.5345	.3671	.2857
250	.4912	.4117	.3938	.4114	.3396	.4472	.2759	.2000
500	.3362	.3291	.2884	.2843	.2375	.3162	.1585	.1000

Rates are normalized to sample size 50.

should provide model discrimination. Thus, all of the results of this analysis supports the use of the new nonparametric models for small and intermediate sample sizes. The results of investigations of samples of size 20 indicate that the strength of these models may lie in small sample analysis.

Graphical Comparisons

Much of the impetus for this research resulted from the ability to analyze many different random samples graphically. For criteria such as MISE, the accuracy of the approximations becomes obscured when dealing with such small quantities, at least for this author. MISE is also an average error, so a graphical approach may give more insight as to the influence that various portions of the density have on the mean value. For example, a graphical analysis showed that while the MISE of the density function for the exponential distribution using Model 3 was far superior, the poor estimation of tail values resulted in an extremely poor distribution function MISE. This observation calls to question the widely accepted use of MISE as a density function estimation criterion. Relying solely on MISE for the density function allows very poor estimators to appear quite good. Throughout this study, we have contended that density estimators should be compared with respect to criteria evaluation at their corresponding distribution functions as well as at the density function. A graphical examination is a simple way to expose these ill-conceived estimators.

To demonstrate the versatility of the new nonparametric estimators, we chose random samples of size 100 from the double exponential, uniform, triangular, Cauchy, and exponential distributions. The nonparametric model used in each case is the one with the smallest approximate MISE listed in Table IV.1. Figures 4.1 through 4.10 present the distribution function and density function approximations plotted against the true underlying processes. Table IV.9 lists the values of the approximate MISEs for the distribution and density functions for each random sample. Many other samples and distribution functions have been examined for different sample sizes. Other probability distributions analyzed included various beta distributions, including U shapes, Weibull distributions, gamma distributions, and extreme value distributions.

Hazard Function Estimation

The availability of a continuous density function estimator derived from a continuous, differentiable distribution function estimator automatically allows one to calculate a continuous hazard function estimator. The hazard function, defined by $h(x)=f(x)/(1-F(x))$, can be a powerful density function discriminant and is used

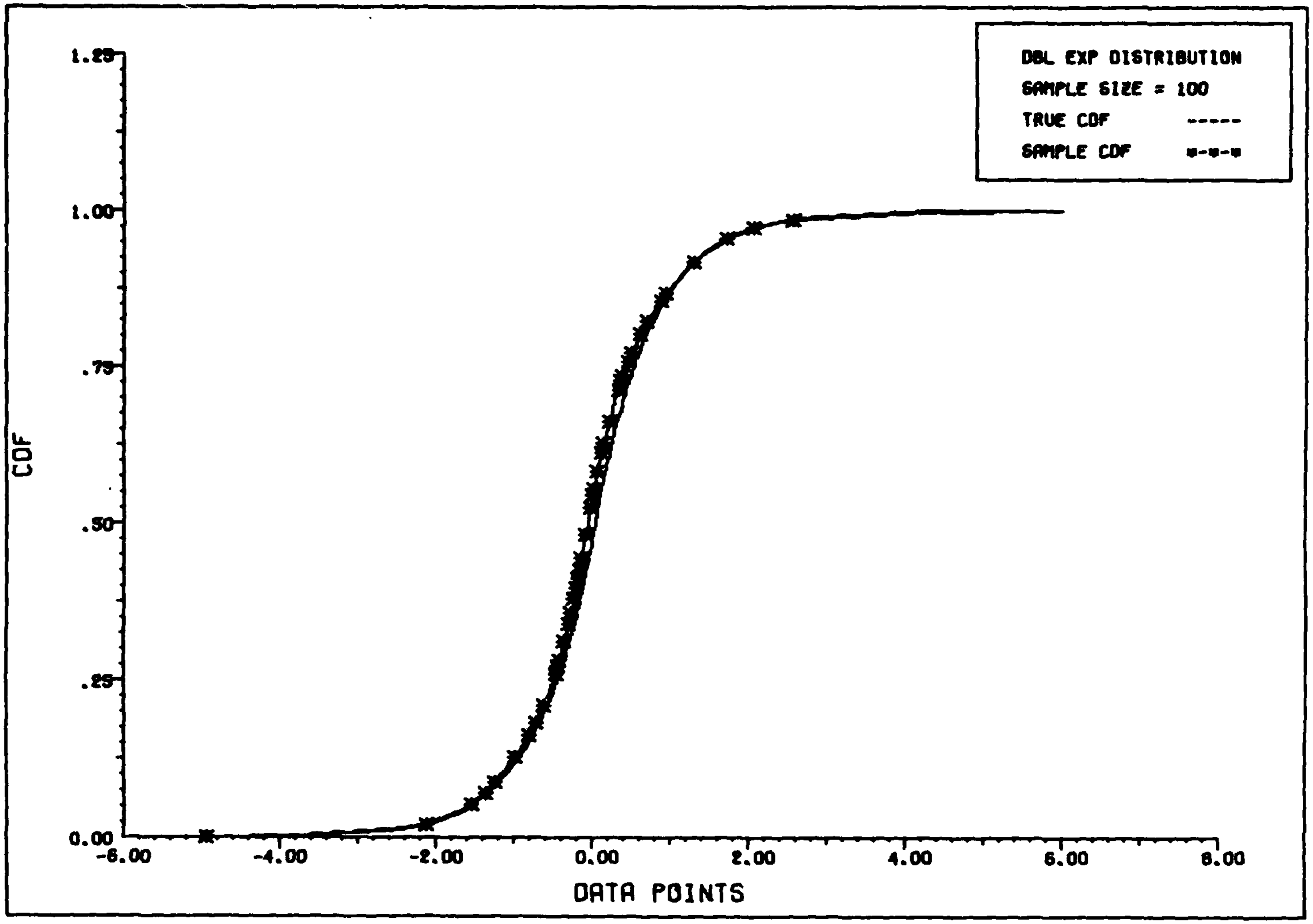

Figure 4.1. Double Exponential CDF vs Model 5

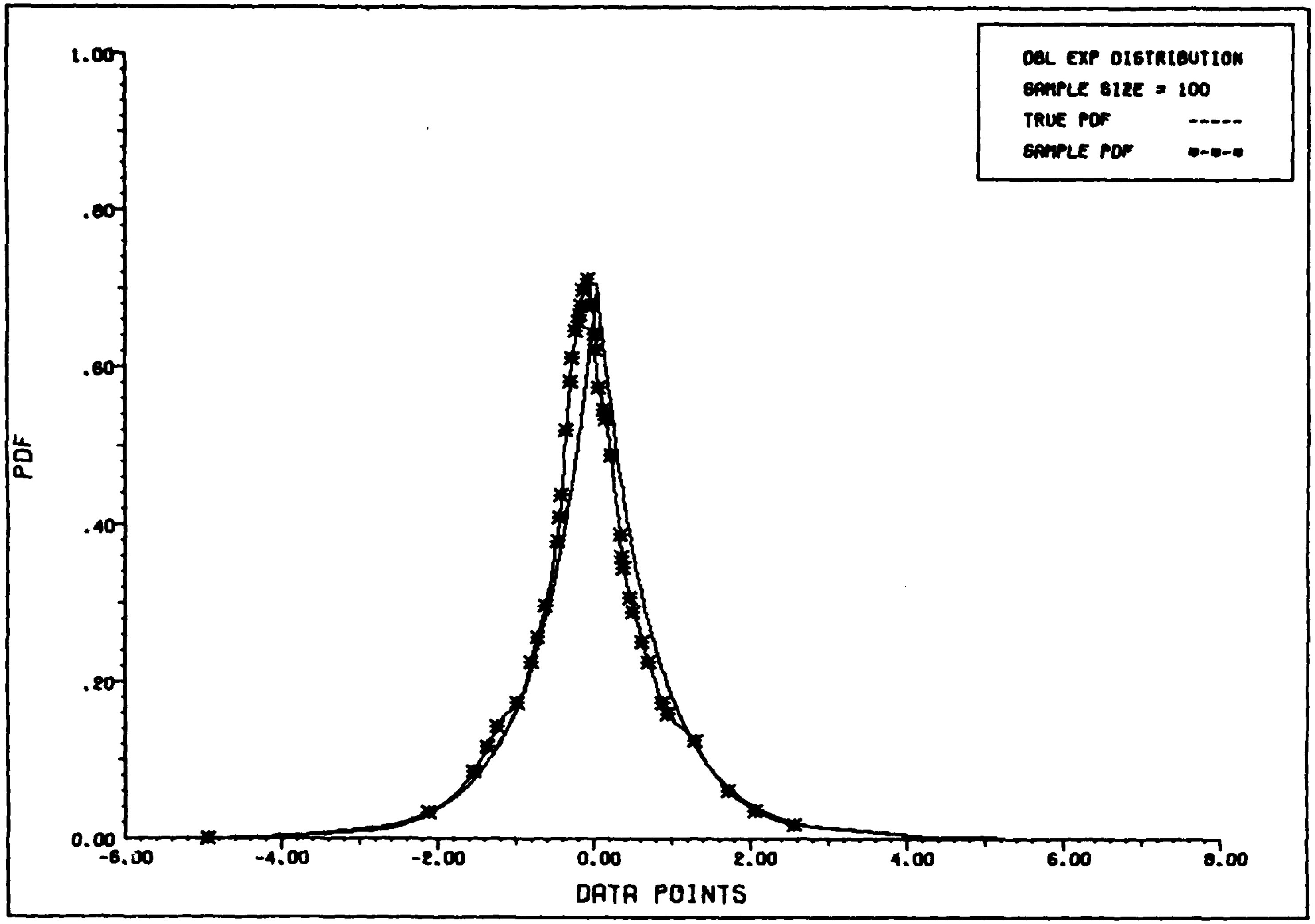

Figure 4.2. Double Exponential PDF vs Model 5

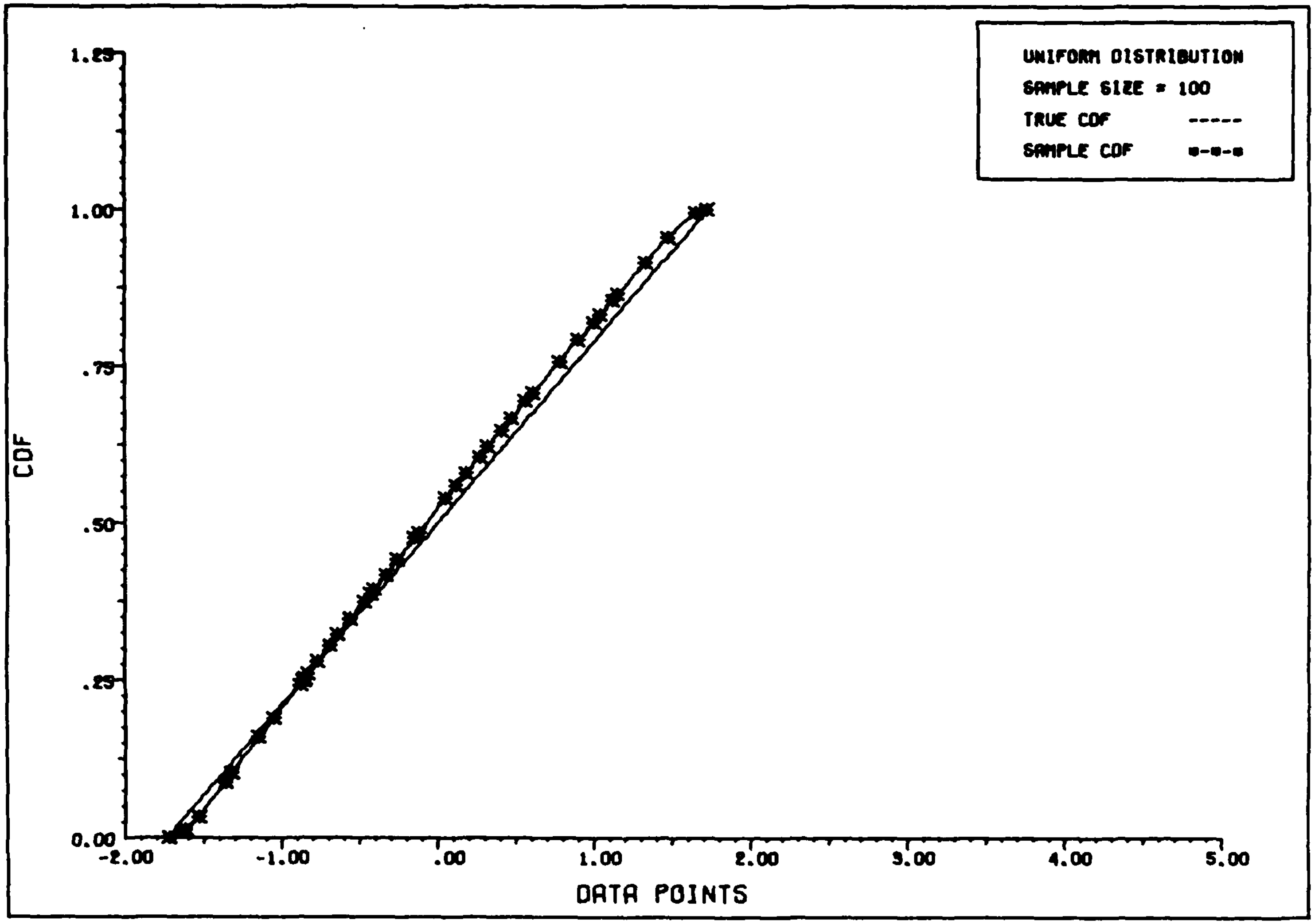

Figure 4.3. Uniform CDF vs Model 3

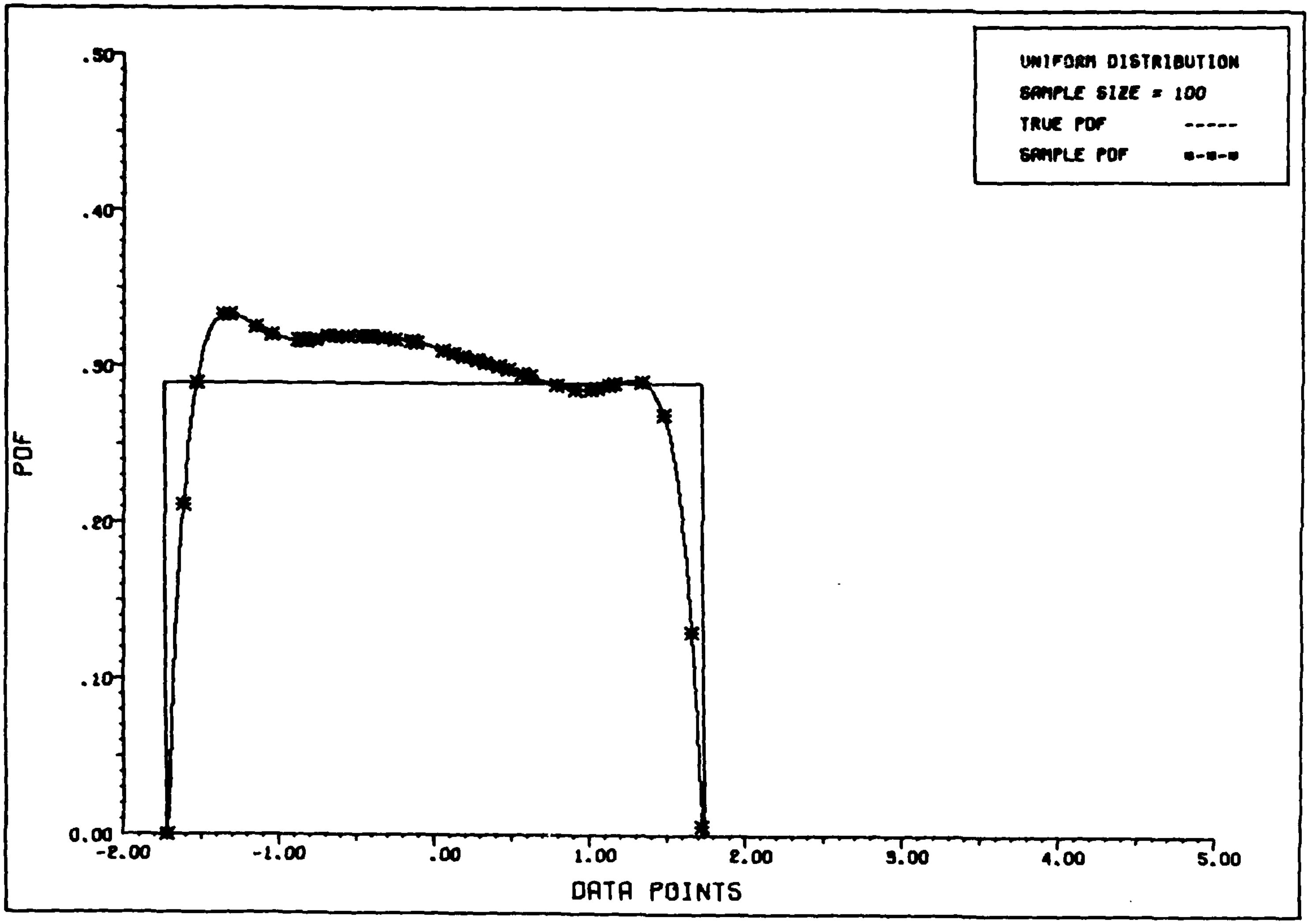

Figure 4.4. Uniform PDF vs Model 3

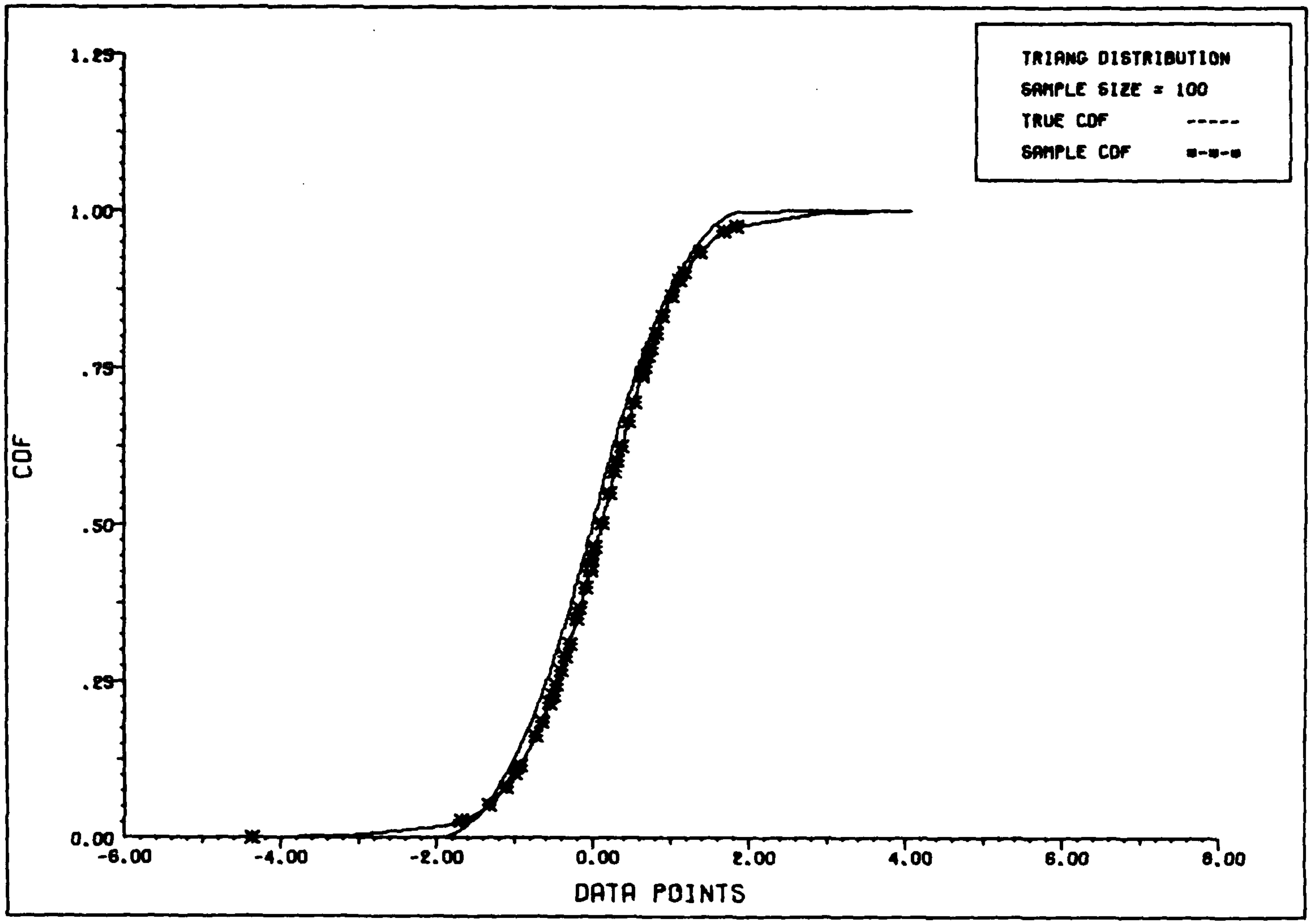

Figure 4.5. Triangular CDF vs Model 4

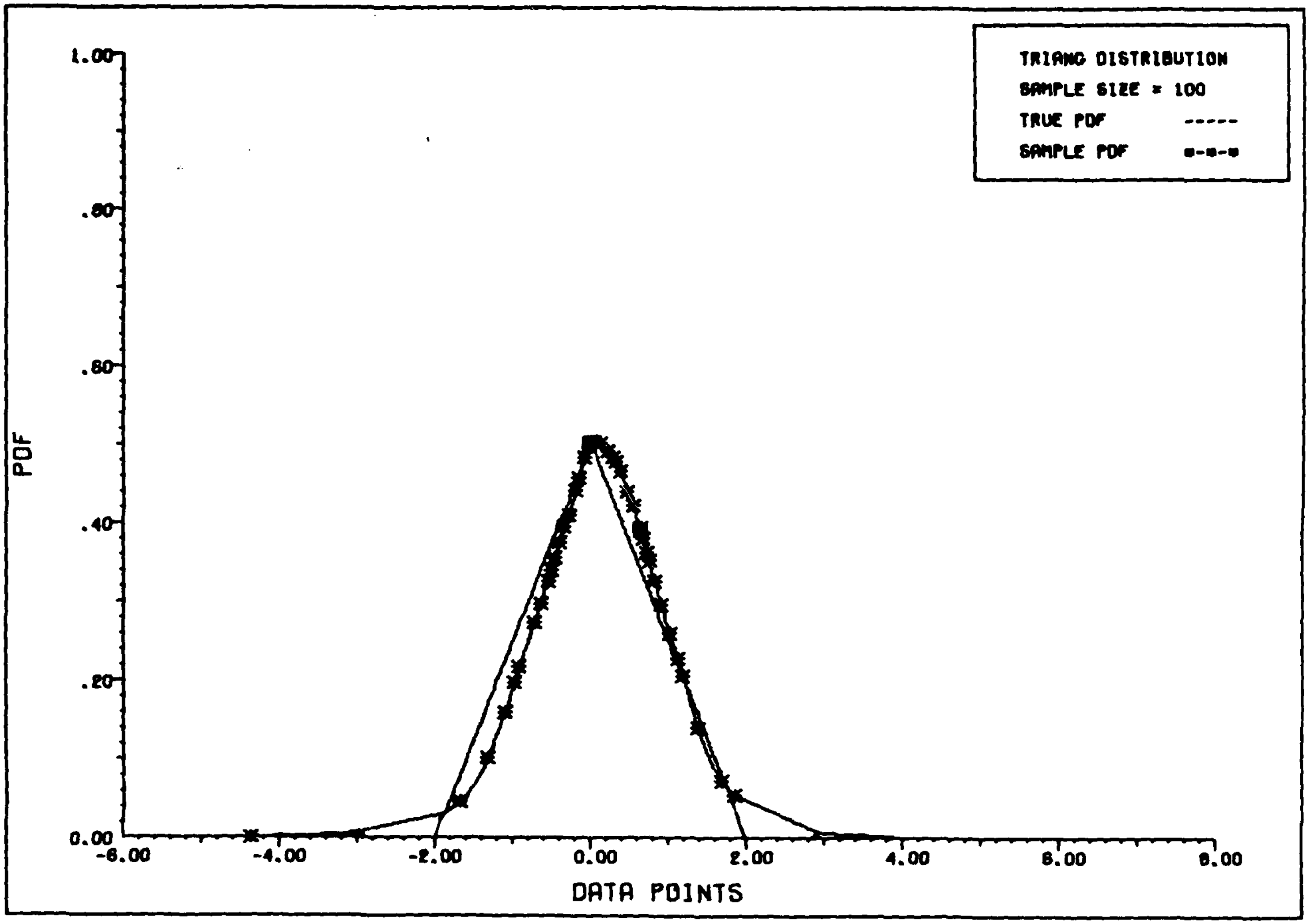

Figure 4.6. Triangular PDF vs Model 4

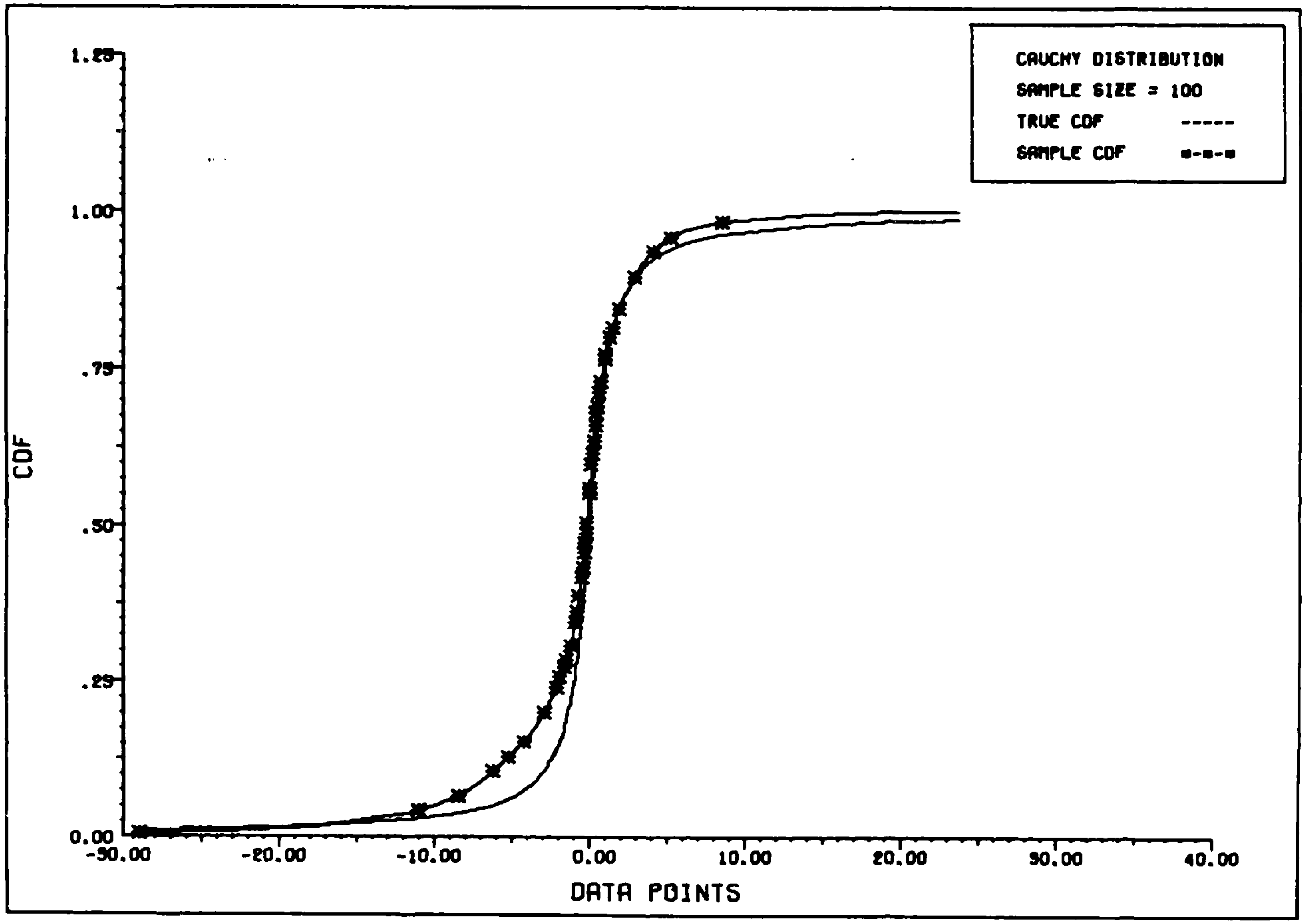

Figure 4.7. Cauchy CDF vs Model 5

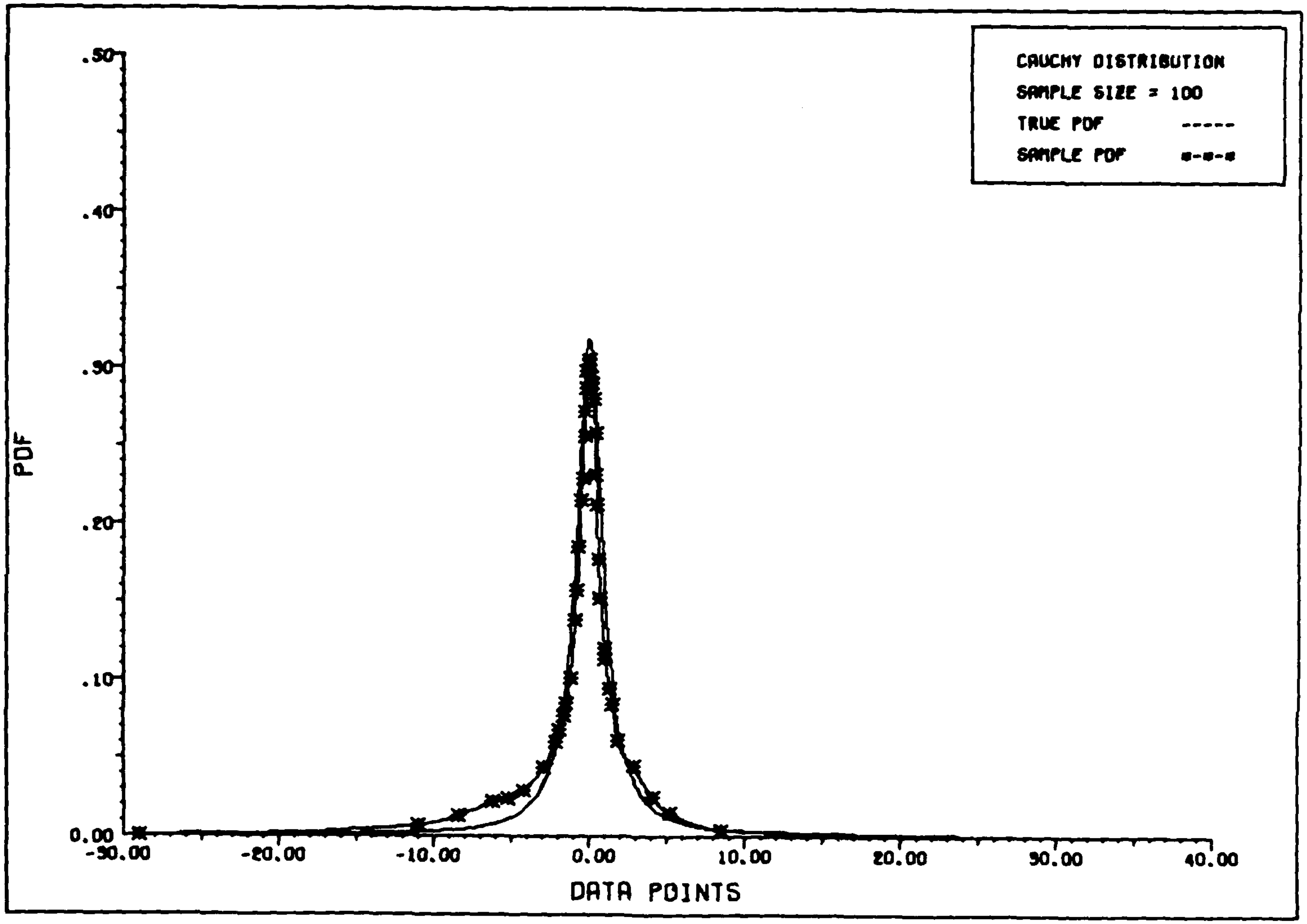

Figure 4.8. Cauchy PDF vs Model 5

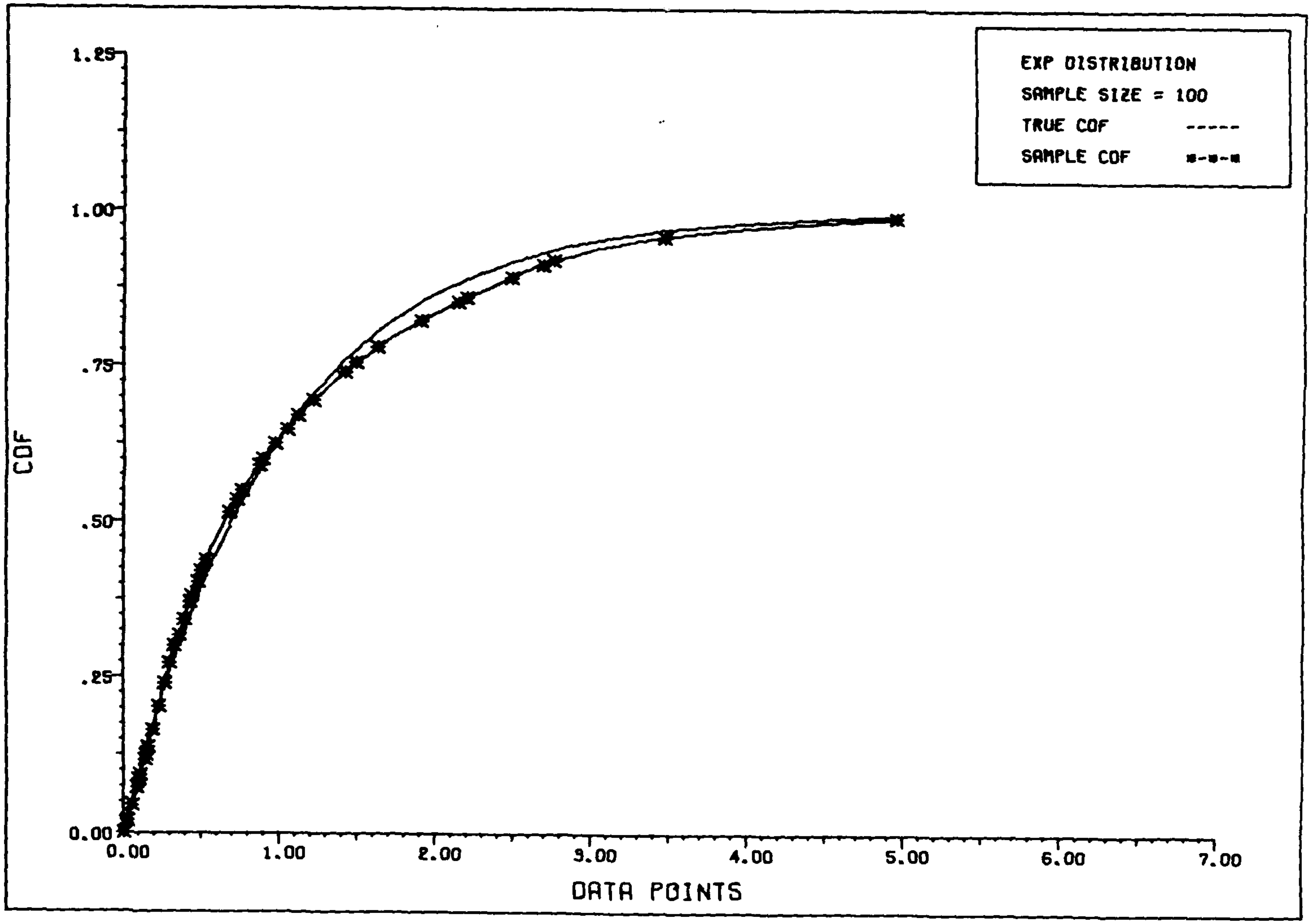

Figure 4.9. Exponential CDF vs Model 6

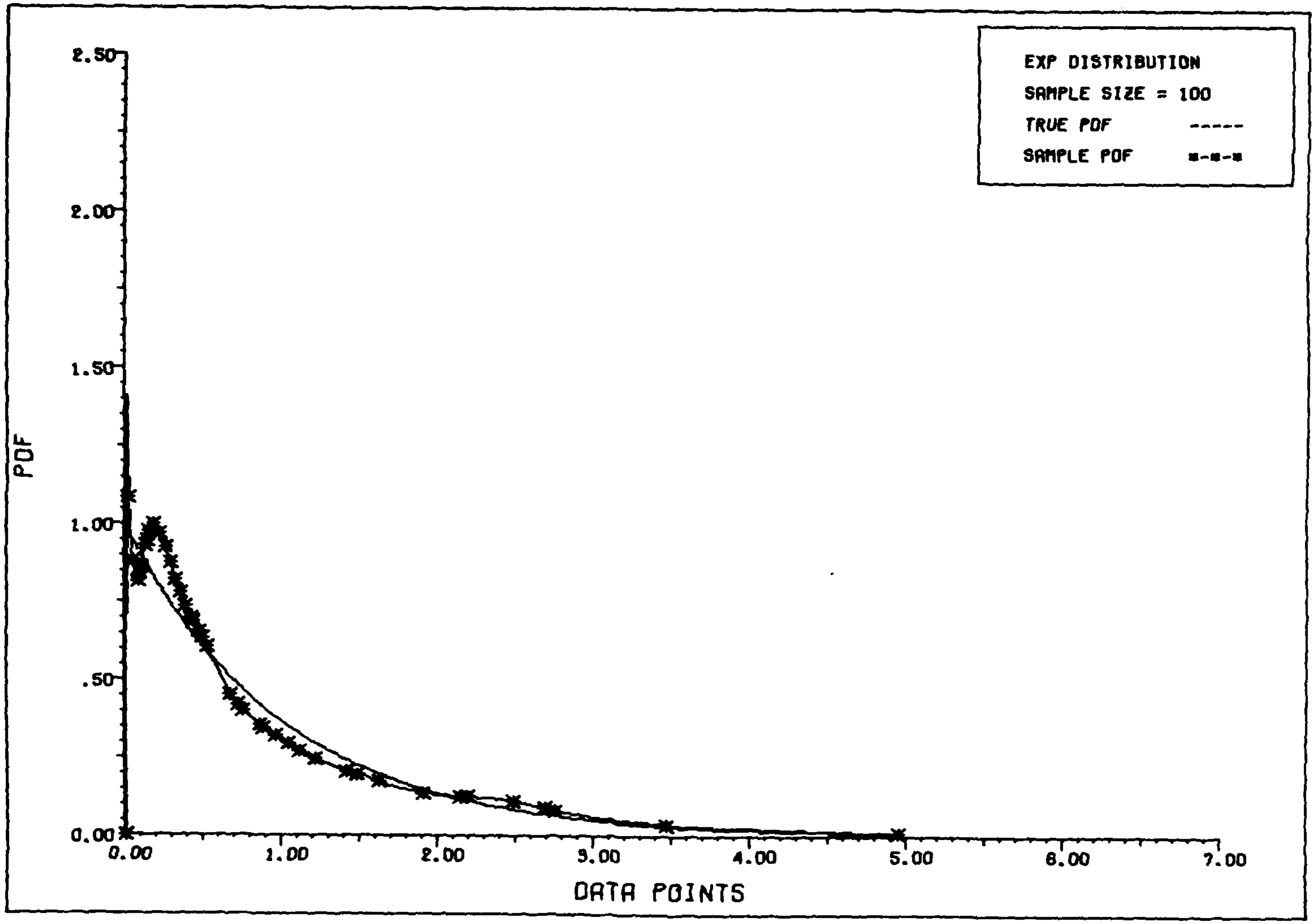

Figure 4.10. Exponential PDF vs Model 6

TABLE IV.9

APPROXIMATE MISE--RANDOM SAMPLES--SAMPLE SIZE 100

	MISE	
Distribution	Distribution Function	Density Function
Double Exponential	.00044	.00352
Uniform	.00054	.00125 (.01500)$^{(1)}$
Triangular	.00170	.00150 (.02403)$^{(1)}$
Cauchy	.00331	.00058
Exponential	.00031	.00786

Note 1: Density function MISE normalized to the interval [0,1].

extensively in reliability engineering and life testing. Early research in hazard analysis was done by Watson and Ledbetter, which prompted their later investigation of density estimation (Ref 103). An empirical approach to hazard function estimation can take the form of estimating the hazard function at the sample data points and fitting some least squares curve through the calculated points (Ref 44). Because of the necessity of using a differencing scheme to construct the density function estimate, the calculated hazard point estimates have magnified errors. The use of a continuous density approximation has a clear advantage.

Using the same models as the CDF and PDF plots, we constructed the hazard function estimates for the random samples plotted in the last section. Figures 4.11 through 4.15 show the estimators plotted versus the true population hazard function. The functions are only plotted between the first and last order statistic. Note the unique shape of each hazard function and the ability of the nonparametric estimator to follow the shape.

Armed with only the new nonparametric estimators and graphs of various distribution, density, and hazard functions, we now have a powerful tool for identifying the underlying distribution of the population from which a random sample is drawn.

Summary

We began our investigation into the utility of our new nonparametric estimators by surveying the literature for other distribution and density estimators. A Monte Carlo study was then described in which the new models were compared with established estimation schemes. The new estimators were very competitive in the mean integrated square error sense. Tables were developed showing the approximate MISE and standard error of the estimate. Based on these values, empirical convergence rates were indicated. We next discussed a graphical comparison of various random samples from five different

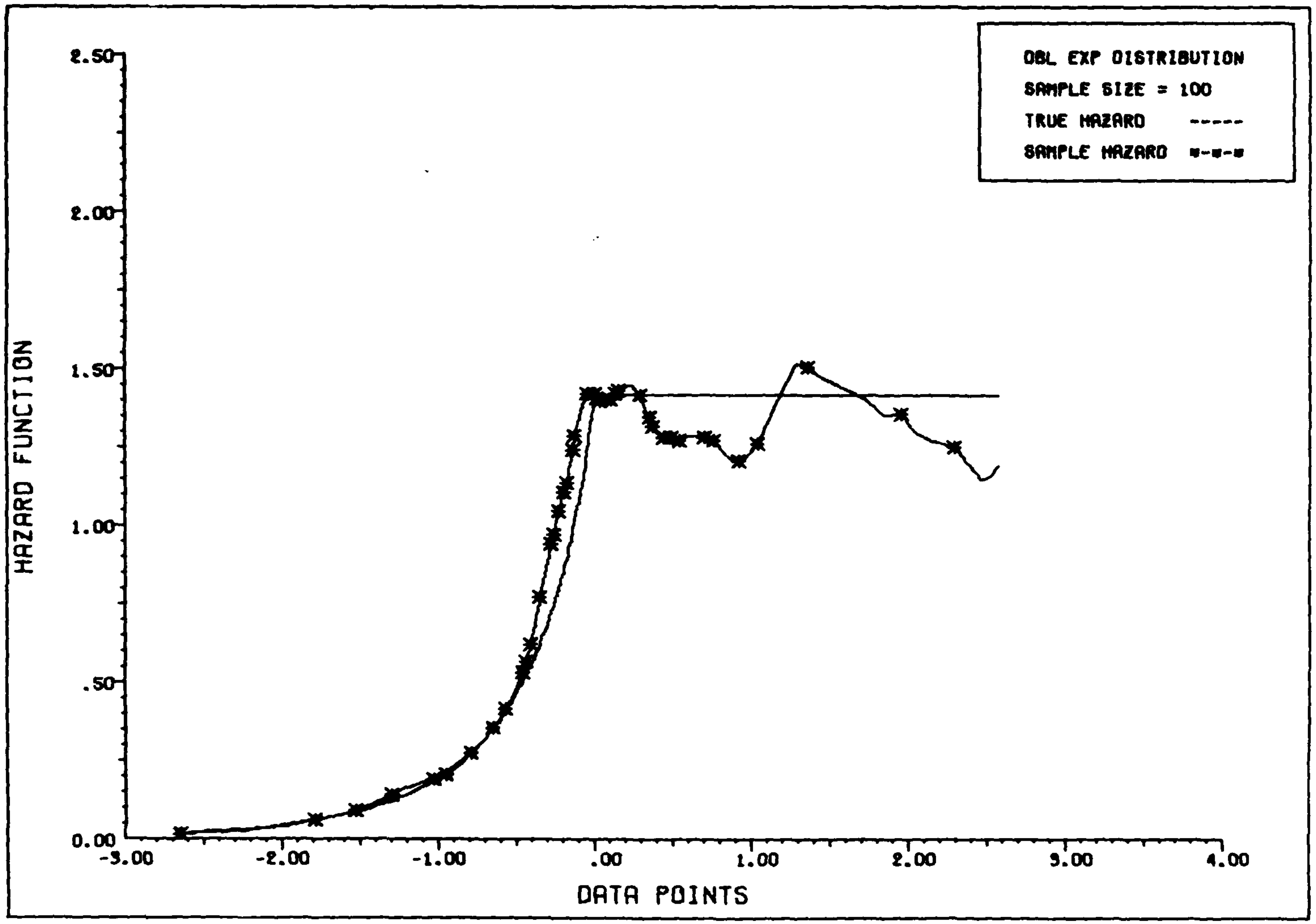

Figure 4.11. Double Exponential Hazard Function vs Model 5

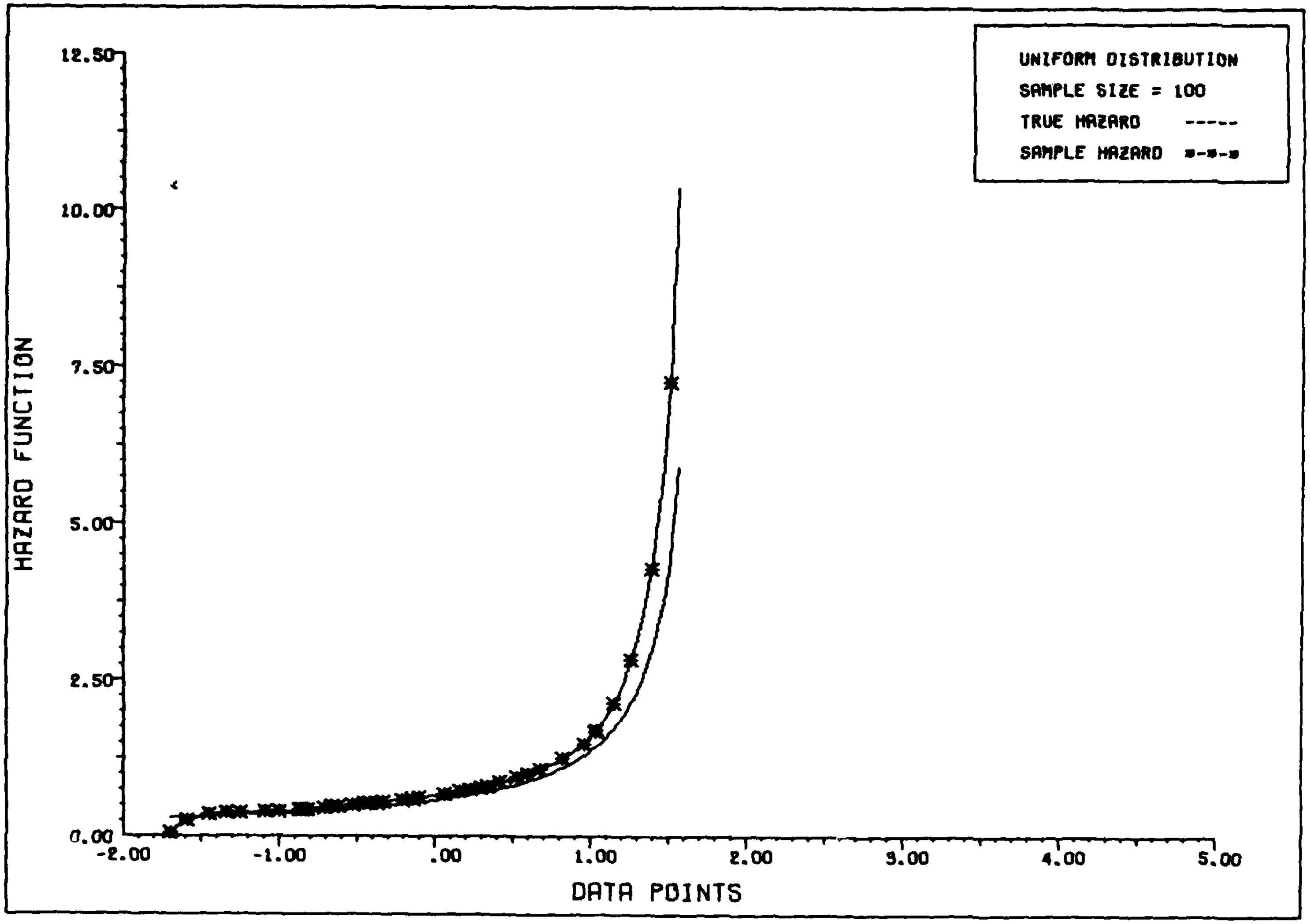

Figure 4.12. Uniform Hazard Function vs Model 3

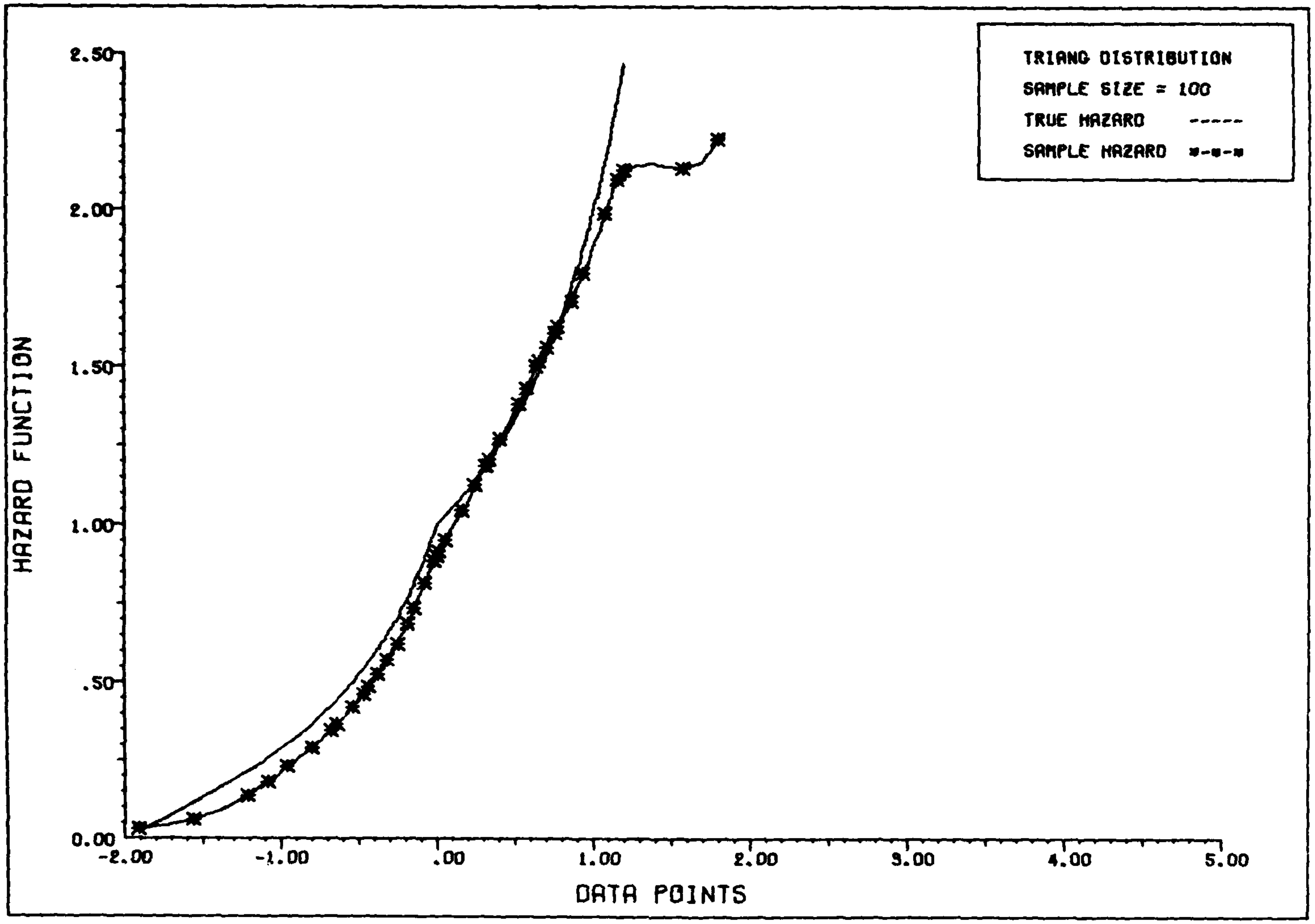

Figure 4.13. Triangular Hazard Function vs Model 4

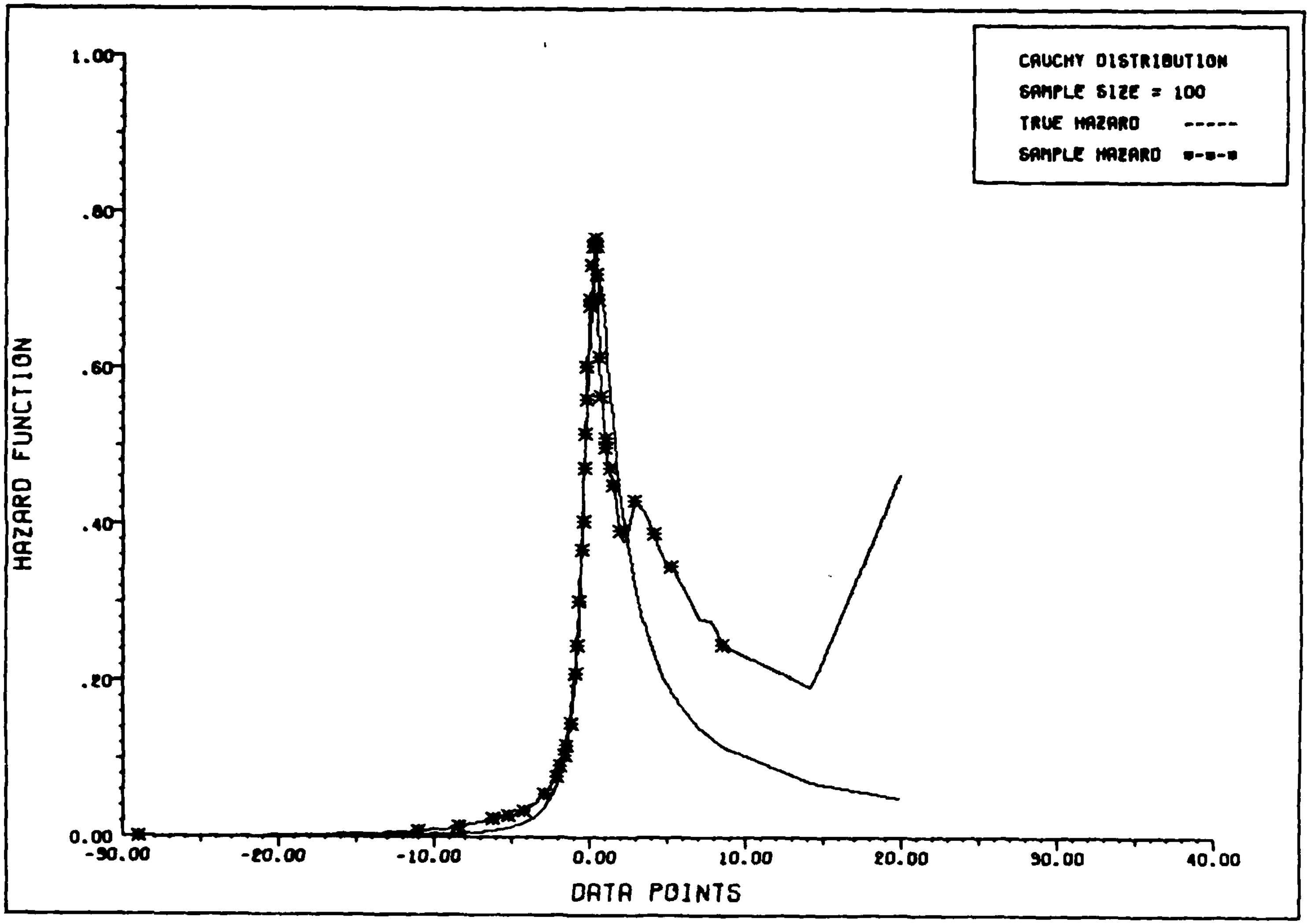

Figure 4.14. Cauchy Hazard Function vs Model 5

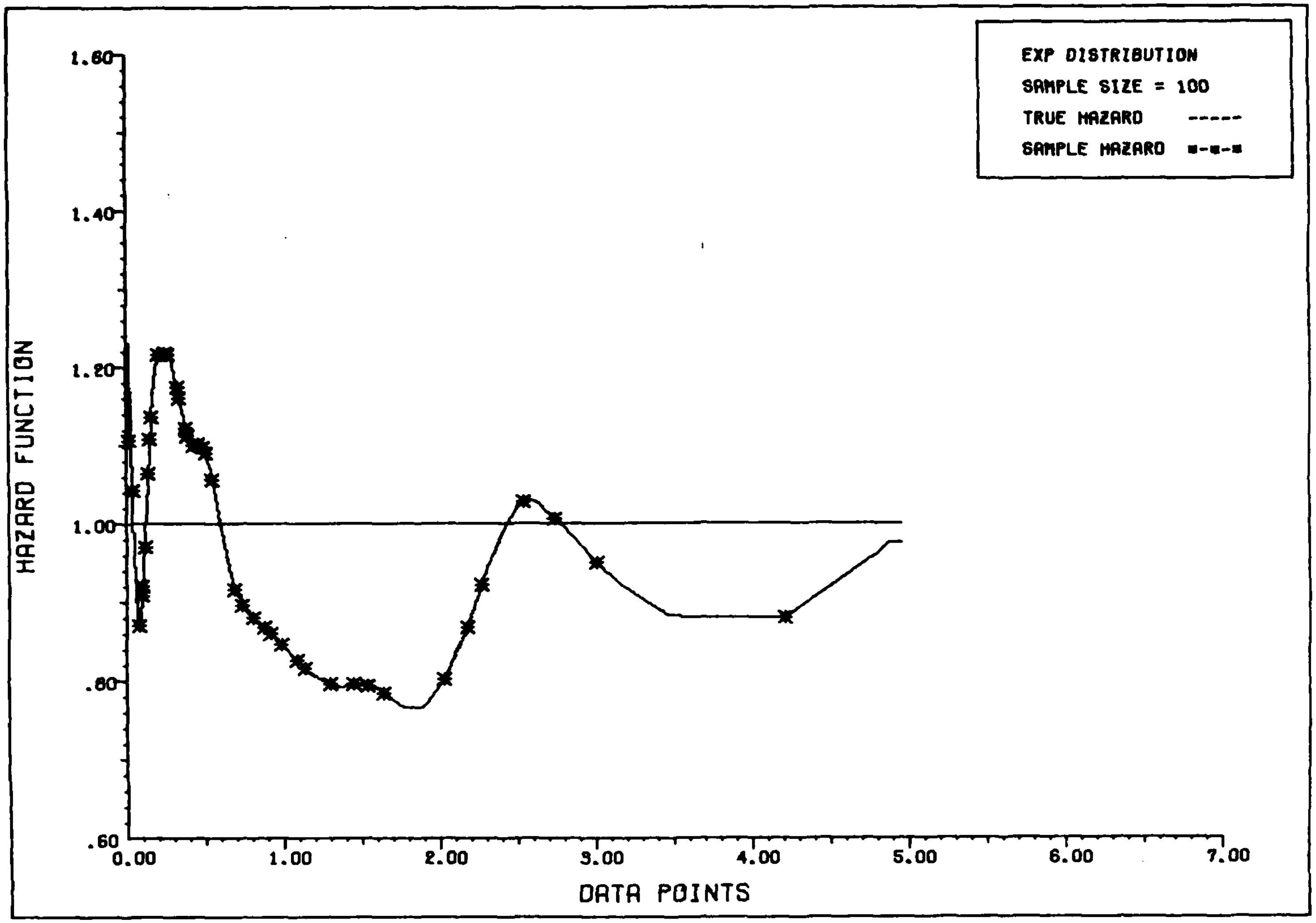

Figure 4.15. Exponential Hazard Function vs Model 6

distributions. We concluded with the development of an approximation to the hazard function, illustrated the hazard estimator for the five distributions, and argued for the simultaneous use of distribution, density, and hazard function graphs in solving problems in model discrimination.

We have demonstrated that our models are extremely competitive and closely approximate the true distribution function and density function. Their use as a population discriminant will be considered next in the development and evaluation of goodness of fit tests based on the new nonparametric estimators.

V. Goodness of Fit Tests

Introduction

Since the last chapter indicated that our models approximated the true underlying distribution with competitive precision, we will now use them as a basis for goodness of fit tests. We begin our discussion by a brief historical survey of goodness of fit tests. Next we introduce eight new test statistics based on two of the adaptive models and a sample distribution step function related to the median ranks. Then, we give the critical values of tests for the normal and extreme value distribution for both a completely specified null distribution and a null distribution whose parameters are estimated. Finally we present the results of power studies for both tests. Powers are also compared with some previously published methods.

Historical Survey

Goodness of fit test literature has not suffered from lack of attention. In our discussion, we are concerned with the goodness of fit problem in the context of life testing. Two important distributions used in life testing are the normal and the extreme value.

Forming the basis for goodness of fit tests is the selection of a test statistic. An excellent survey of distribution free statistics is given by Sahler (Ref 78). Consider now, some of the tests based in the statistics for the case of a completely specified null hypothesis. References in Sahler's survey give much of the historical background.

To avoid using extensive tables, Stephens proposed computational approximations for critical values of eleven common test statistics (Ref 88). Schuster uses a modified empirical distribution function to develop a test based on the Kolmogorov Smirnov statistic (Ref 82). Saniga and Miles evaluate some standard tests of normality against an alternative distribution which is a member of the asymmetric stable probability distribution family (Ref 80). Tests of symmetry have been proposed using the Cramer von Mises statistic and modified empirical distribution functions by Rothman and Woodroofe and Hill and Rao (Refs 36 , 76). For the Weibull distribution, or equivalently the extreme distribution value, Smith and Bain propose a goodness of fit test based on the correlation coefficient and evaluate both complete and censored samples in both the completely specified and composite hypothesis cases (Ref 87). Foutz attempts a more general approach to goodness of fit testing by using an empirical probability measure as a basis rather than the empirical distribution function (Ref 25). A novel approach of Dudwicz and van der Meulen uses entropy as the basis for a test of uniformity (Ref 20). Extensions to other distributions have not been published as yet.

While the aforementioned tests all use a completely specified null hypothesis, the work of David and Johnson shows that goodness of fit tests are independent of the true parameter values when invariant location and scale estimates are substituted and the test depends on the probability integral transform (Ref 18). This result opened the door for composite null hypothesis tests which estimate the parameters of the distribution by invariant estimators. Lilliefors pioneered the investigations of this type of developing tables for the KS statistic (Ref 50). Stephens conducted tests for uniformity, normality and exponentiality using modifications of the KS, CVM, AD, Kuiper and Watson statistics when the parameters were estimated (Ref 89). Green and Hegazy modify the KS, CVM, and AD tests by using other sample distribution functions as a basis for the test statistics. Their results show improvements in powers are possible when new sample distribution functions are used (Ref 29). Durbin proposes a generalized KS test when parameters are estimated and applies the result to tests of exponentiality and spacings (Ref 21). Durbin's results were based in part on the investigation of spacings done by Pyke (Ref 69). Pyke's work also motivated Mann, Scheuer and Fertig's development of two new statistics, L and S. They proposed tests based on these statistics for the two parameter Weibull or extreme values distribution (Ref 53). Littell, McClave, and Offen conducted power studies using the S statistic as well as four others for these same distributions (Ref 51). Stephens, following methods developed previously, computed critical values of modified CVM, AD and Watson statistics for tests of the extreme value distribution (Ref 90). A recent paper by Mihalko and Moore shows an application of a chi square test goodness of fit test to the two parameter Weibull when the parameters are estimated (Ref 56).

Test Procedures

The classical goodness of fit test can be stated as follows: from an observed random sample, $X_1,...,X_n$, test whether the sample comes from a population with distribution function F(x). Standard tests using EDF or modified EDF statistics are based on comparisons between F(x) and some sample distribution function. As we have generated new continuous, differentiable, sample distribution functions, we follow a similar approach to define our goodness of fit tests. Because of their outstanding performance using a mean integrated square error criterion over a wide range of distributions, we chose Models 5 and 6 to form the bases for our new tests.

Null Distributions and Situations Considered. One of the major applications of goodness of fit tests is in the area of life testing. For this reason, we chose two important and widely used failure distribution models, the normal and the extreme value distributions, for our null hypotheses.

The extreme value distribution considered in this entire analysis is the distribution of the largest value, whose cumulative distribution function is given by:

$$F(x) = \exp[-\exp\{-(\frac{x-\delta}{\sigma})\}]$$

where $-\infty<x<\infty$, $-\infty<\delta<\infty$, $\sigma>0$

Two specific hypotheses situations will also be considered. The first is the classical case of the null distribution, F(x), having all of its parameters completely specified. The second situation, and probably the more common one for the applied statistician, is the case where the functional form of the null distribution is hypothesized, but the parameters are estimated. Although both the normal and extreme value distributions are members of a two parameter family, we chose not to examine the situation where only one parameter is estimated and the other specified. We believe that the two situations considered here comprise the vast majority of cases encountered in actual practice.

The estimators used in the case of the normal distribution will be the uniformly minimum variance unbiased estimates, $\overline{X}$ and S. For the extreme value we will employ a Newton Raphson iteration technique to calculate the maximum likelihood estimators of the location and scale parameters.

Test Statistics. Eight new test statistics are proposed. The first set of these statistics is based on Models 5 and 6 and the modified distance measures listed in Appendix 1. Given the random sample, $X_1, \ldots, X_n$, let SF(x) be based on Model 5. Now define

$$D5 = \max_i \left| F(X_i) - SF(X_i) \right|$$

$$W5 = n\int_{-\infty}^{\infty} (SF(x)-F(x))^2 \, dSF(x)$$

$$A5 = n\int_{-\infty}^{\infty} (SF(x)-F(x))^2 [SF(x)(1-SF(x))] \, dSF(x)$$

Calculating SF(x) using Model 6 gives similar definitions for D6, W6, and A6. These first six test statistics are modifications of the classical KS, CVM and AD statistics.

Along the lines of the tests proposed by Green and Hegazy, we also propose two new test statistics based on a sample distribution step function (Ref 29). We wanted to use the median ranks in both a KS and CVM statistic, since, as plotting positions, they describe measures of central tendency for the mostly skewed rank distributions. The aim was to get the squared term in the summation for the CVM statistic to contain the difference between the hypothesized distribution function at that point and the median rank value. Working backwards, one sample distribution that will suffice is $F_n(x)$, where

$$F_n(x) = \begin{cases} .2\ /(n+.4) & x<X_{(1)} \\ (i+.2)/(n+.4) & X_{(i)}<x<X_{(i+1)} \quad i=1,\dots,n-1 \\ (n+.2)/(n+.4) & x>X_{(n)} \\ (i-.3)/(n+.4) & x=X_{(i)} \quad i=1,\dots,n \end{cases} \qquad (5.1)$$

Note that $F_n(X_i)$ is the midpoint of the jump from $F_n(X_i^-)$ to $F_n(X_i^+)$.

We now define two new statistics based on this $F_n(x)$.

$$\text{DMR} = \max_i \left| F(X_i) - \frac{i-.3}{n+.4} \right|$$

and $$\text{WMR} = \frac{n^2}{12(n+.4)^3} + \frac{n}{n+.4} \sum_{i=1}^{n} \left(F(X_i) - \frac{i-.3}{n+.4}\right)^2$$

Critical Values. Given the two distributions and two situations for the null hypothesis and the eight new goodness of fit statistics, we now generated critical values for each test statistic by the following method.

For fixed sample sizes of 10(10)50 we generated n ordered random variates from the null distribution (see Appendix 5 for a further discussion of random variate generation). We next calculated the approximate parameter estimates from the random sample. Finally, we calculated each of the eight new test statistics for this sample. The procedure was repeated 1000 times and values for each test statistic were ordered. Percentiles corresponding to alpha levels of .20, .15, .10, .05, .025, and .01 were determined. The entire process was then repeated five times and the critical values for each test statistic, at each sample size and alpha level were calculated by averaging the five corresponding percentiles. Appendix 3 gives the tables for the critical values for the normal and extreme value distributions, both when the null distribution is completely specified and when the parameters are estimated. Values are listed for five different sample sizes and six different alpha levels.

Tables V.1 and V.2 show the critical values across sample sizes and compares the eight new test statistic values with the classical values for the KS, CVM and AD statistics for a completely specified null hypothesis. Note the smaller values of the critical values for the new statistics (except A5 and A6 for sample size ≤ 30). This observation strengthens the claim made earlier that our new nonparametric model "better" approximates the true

TABLE V.1

COMPARISON OF CRITICAL VALUES FOR THE NORMAL DISTRIBUTION AT THE 5-PERCENT ALPHA LEVEL(1)

	Sample Size				
Statistic	10	20	30	40	50
$D^{(2)}$	.4094	.2941	.2418	.2102	.1884
D5	.3147	.2160	.1738	.1511	.1323
D6	.3108	.2228	.1765	.1543	.1349
DMR	.3509	.2687	.2211	.1963	.1748
$W^{2\ (2)}$	.5411	.5026	.4890	.4822	.4780
W5	.4513	.4267	.4067	.4101	.3998
W6	.4243	.4271	.4068	.4137	.4070
WMR	.4258	.4550	.4365	.4610	.4510
$A^{2\ (2)}$	2.492	2.492	2.492	2.492	2.492
A5	4.416	2.907	2.556	2.367	2.175
A6	4.013	2.837	2.563	2.388	2.218

Note 1: Null distribution is completely specified.

Note 2: Critical values calculated from formulae given by Stephens (Ref 89).

TABLE V.2

COMPARISON OF CRITICAL VALUES FOR THE EXTREME VALUE DISTRIBUTION AT THE 5-PERCENT ALPHA LEVEL[1]

	Sample Size				
Statistic	10	20	30	40	50
D[2]	.4094	.2941	.2418	.2102	.1884
D5	.3256	.2183	.1751	.1531	.1363
D6	.3205	.2111	.1764	.1542	.1376
DMR	.3536	.2661	.2221	.1953	.1769
W^2 [2]	.5411	.5026	.4890	.4822	.4780
W5	.4802	.4530	.4213	.4171	.4239
W6	.4444	.4363	.4128	.4152	.4242
WMR	.4284	.4491	.4317	.4473	.4537
A^2 [2]	2.492	2.492	2.492	2.492	2.492
A5	4.516	3.111	2.587	2.398	2.345
A6	4.104	3.014	2.572	2.367	2.343

Note 1: Null distribution is completely specified.

Note 2: Critical values calculated from formulae given by Stephens (Ref 89).

distribution than the EDF. "Better" is now in terms of KS, CVM and AD distance measures. Since each criterion for closeness of the true and approximated functions measures different qualities of the approximation, our distribution and density approximations of the last chapter gain more credibility.

While small critical values do indicate a high quality approximation, the real performance of a goodness of fit test is measured by its power.

Power Comparisons

Once the critical values were determined, we next evaluated the power of our new tests using various alternative distributions. Our first concern was the verification of our critical values for both distributions over all cases considered. Monte Carlo samples of size 1000 for the normal distribution and 2000 for the extreme value distribution were generated for each random sample size of 10(10)50. Tables V.3 and V.4 show the results of the critical value verifications at sample size 20 with the parameters of the null distributions estimated. All of the results indicated a good agreement between the alpha level and the power of the test using random samples generated by the null distribution. Thus, the critical values were empirically confirmed.

TABLE V.3

CRITICAL VALUE VERIFICATION FOR THE NORMAL DISTRIBUTION AT SAMPLE SIZE 20

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	201	156	105	53	26	14
D6	195	147	94	51	25	13
DMR	199	151	106	46	23	9
W5	202	156	102	52	24	14
W6	189	150	101	56	23	10
WMR	185	143	91	49	27	14
A5	201	155	108	51	24	14
A6	209	157	107	52	27	15

Entries represent the number of samples significant at the given alpha level for each test statistic calculated over a Monte Carlo sample of size 1000. The parameters of the null distribution were estimated.

TABLE V.4

CRITICAL VALUE VERIFICATION FOR THE EXTREME VALUE DISTRIBUTION AT SAMPLE SIZE 20

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	410	308	201	85	41	12
D6	395	282	188	94	35	10
DMR	410	328	228	111	52	15
W5	405	305	204	87	42	14
W6	399	310	202	89	43	10
WMR	389	296	209	107	51	13
A5	401	303	192	89	42	22
A6	405	311	192	92	42	15

Entries represent the number of samples significant at the given alpha level for each test statistic calculated over a Monte Carlo sample of size 2000. The parameters of the null distribution were estimated.

The general method followed in the power studies was to generate 1000 sets of random samples of size 10(10)50 for each alternative distribution. Then, the eight test statistics were calculated for each sample. The number of samples, for each sample size, which had test statistics that exceeded the critical values, was recorded. For a given alternate distribution, situation type, sample size, alpha level, and test statistic, the power of the test is the number of samples significant divided by 1000, the Monte Carlo size. Appendix 4 gives the results of some of the power studies for both null distributions, the normal and extreme value. The cases evaluated but not tabled include all of the results for alpha levels .20, .15, and .025. Several alternative distributions were not included in the tables but are discussed later in this chapter when each null distribution is examined. However, the tables do present the results for the most commonly used alpha levels and alternative distributions which provide variety and a basis for future comparisons.

Because of the similarity between Models 5 and 6, the correlation between the new test statistics should be rather high. To gain some insight into the correlations between all pairs of test statistics, over 1400 output matrices similar to Table V.5 were constructed for each null distribution, hypothesis situation, sample size, alpha level, and each alternative distribution. Each cell of

TABLE V.5

TYPICAL OUTPUT MATRIX OF POWER STUDIES

Null Distribution--Extreme Value, Parameters Estimated
Alternative Distribution--Normal
Sample Size--20
Alpha Level--.10

Statistic	D5	D6	DMR	W5	W6	WMR	A5	A6
D5	490							
D6	399	409						
DMR	225	221	252					
W5	468	391	221	491				
W6	416	376	218	417	419			
WMR	265	267	209	264	264	280		
A5	435	375	214	446	402	258	471	
A6	399	357	206	404	378	252	420	438

Entries represent the number of samples significant by both row and column statistics using a Monte Carlo sample of size 1000.

the matrix contains the number of samples significant by the corresponding row and column statistics. Diagonal terms were used to construct the power tables in Appendix 4.

Normal Distribution. Tables A4.1 through A4.6 in Appendix 4 list the results of the power study conducted for the normal distribution. We attempted to construct a meaningful alternative distribution when the null distribution parameters were completely specified. Sometimes the null distribution parameters were adjusted for simplicity. Eleven alternative distributions were considered.

For the double exponential, uniform, and Cauchy distributions, the location and scale parameters of the null and alternative distributions were zero and one respectively. For the exponential, gammas, and extreme value, the null distribution was modified to have the same mean and variance as the standard form of the alternative distribution. For example, the exponential distribution had a location parameter of zero and a scale parameter of one, while the normal distribution as the null distribution had location and scale parameters equal to one. The lambda distributions had zero mean and unit variance as did the corresponding normal as the null distribution. See Ramberg, et al., for a discussion of the four parameter lambda distribution (Ref 72).

Table V.6 lists selected results of the power study. Parameters for the null distribution have been estimated and only the results for an alpha level of .05 are shown. The powers for the three lambda distributions are included for comparison purposes. These three distributions are not included in the general tables of Appendix 4. To facilitate comparisons of our results with other published power studies, we included the classical KS, CVM, and AD statistics (listed as D, W_0 and A respectively) as well as two modified EDF statistics D_2 and A_{22}. D_2 is a summed KS distance between the hypothesized distribution and the EDF (summed over the data points). A_{22} is equal to n times the Anderson-Darling integral distance listed in Appendix 1 after $H_n(x)$ is substituted for SF(x) where

$$H_n(x) = (i+\tfrac{1}{2})/(n+1) \quad X_{(i)} \leq x < X_{(i+1)} \quad i=1,\ldots,n$$

See reference 29 for a further discussion of these two statistics. Note that these five test statistics used for comparison had powers calculated using different random samples than the ones used to calculate the powers for the eight new test statistics.

Several observations deserve mention. First, the tests based on Models 5 and 6 are superior in almost every instance to the tests based on median ranks. Second, for the gamma alternatives, it appears that D_2 and A_{22} have a

TABLE V.6

SELECTED POWER COMPARISONS FOR THE NORMAL DISTRIBUTION AT THE 5-PERCENT ALPHA LEVEL

Alternative Distribution	Sample Size	$D^{(1)}$	$D_2^{(2)}$	$W_0^{(2)}$	$A^{(2)}$	$A_{22}^{(2)}$	D5	D6	DMR	W5	W6	WMR	A5	A6
Double	20	220	260	248	262	239	289	282	207	319	285	254	201	169
Exponential	40	-	437	446	455	433	454	446	328	435	443	413	366	388
Uniform	20	120	149	134	173	200	159	83	88	26	32	131	265	263
	40	-	373	332	450	511	272	233	225	110	118	375	448	504
Cauchy	20	860	867	869	871	866	869	871	838	882	866	860	820	808
	40	-	992	991	990	992	991	992	980	990	990	992	988	989
Exponential	20	590	816	722	781	806	827	773	594	793	785	714	845	840
	40	-	986	969	988	991	984	983	914	980	978	967	991	991
Gamma-2	20	-	656	-	-	613	481	460	329	465	462	422	475	478
	40	-	894	-	-	905	800	779	613	807	803	733	845	848
Gamma-4	20	-	426	-	-	390	239	231	152	226	223	180	231	241
	40	-	635	-	-	616	507	479	351	512	502	435	553	541
Gamma-6	20	-	316	-	-	277	228	220	169	223	215	190	208	207
	40	-	498	-	-	472	329	306	223	317	310	257	323	332
Extreme	20	-	-	-	-	-	298	298	205	301	302	237	277	280
Value	40	-	-	-	-	-	534	499	362	534	578	441	523	521

TABLE V.6--Continued

Alternative Distribution	Sample Size	$D^{(1)}$	$D_2^{(2)}$	$W_0^{(2)}$	$A^{(2)}$	$A_{22}^{(2)}$	D5	D6	DMR	W5	W6	WMR	A5	A6
Lambda (0,5)(3)	20	-	-	-	-	-	172	143	109	167	150	110	120	98
	40	-	-	-	-	-	200	190	154	202	197	159	179	172
Lambda (0,9)(3)	20	-	-	-	-	-	253	233	179	274	239	187	195	171
	40	-	-	-	-	-	365	353	264	383	372	321	345	341
Lambda (1,4)(3)	20	-	-	-	-	-	356	338	252	344	346	308	331	340
	40	-	-	-	-	-	640	618	460	654	652	557	678	684

Note 1: Value for this statistic was taken from reference 89 , Table 5.

Note 2: Values for these statistics were taken from reference 29, Table 4.

Note 3: The lambda distribution is the four parameter distribution examined in reference . The distributions listed here all have zero location and unit scale parameters. Numbers in parentheses indicate the values of the skewness and kurtosis respectively of the distribution.

distinct advantage over the new tests. Again, however, caution is advised since the underlying random samples were different. Third, with the further exception of the uniform, the new tests based on Models 5 and 6 have very competitive powers.

Extreme Value Distribution. Tables A4.7 through A4.12 in Appendix 4 list the results of the power study conducted for the extreme value distribution. An attempt, as in the normal power study, was made to construct meaningful alternative distributions when the null distribution parameters were completely specified. Twelve alternative distributions were considered.

For the normal, uniform and double exponential distributions, the location and scale parameters were the mean and the square root of the variance of a standard extreme value distribution. The null distribution had zero location parameter and unit scale parameter. For the exponential, logistic and gamma distributions, location and scale parameters for both null and alternative distributions were set to zero and one respectively. As such, powers shown for the exponential appear quite high in the completely specified case. Power comparisons for the gamma distributions with shape parameters 2, 4 and 6 were made but are not listed in Appendix 4. Also not listed in Appendix 4 are the results of the power study for the four parameter lambda distribution with skewness equal to one and kurtosis equal to four. Random variables from chi square distributions with one degree and four degrees of freedom were also generated. Taking minus the natural logarithm of these random variables generates samples to compare against the extreme value distribution which are analogous to testing chi square random samples against a two parameter Weibull distribution. Although listed as χ^2 distributions, it should be noted that the actual comparison for the power determination was made between $-\ln(\chi^2)$ and the extreme value distribution.

Table V.7 lists selected results of the extreme value power study. Parameters for the null distributions have been estimated and only the results for an alpha level of .05 are shown. Parts of Table III of reference 51 are included to allow for comparisons to be made. However, again caution is advised since the random samples which generated both sets of powers were different. The values listed from reference 51 are rounded to compare with a Monte Carlo sample of size 1000. The D, W^2 and A^2 are the standard KS, CVM and AD test statistics. T is Smith and Bain's correlation statistic and S is Mann, Scheuer and Fertig's statistic. Both were referenced earlier in this chapter.

We note several trends. Again we detect the inferior performance of tests based on the median ranks

TABLE V.7

SELECTED POWER COMPARISONS FOR THE EXTREME VALUE DISTRIBUTION
AT THE 5-PERCENT ALPHA LEVEL

Alternative Distribution	Sample Size	$D^{(1)}$	$W^{2(1)}$	$A^{2(1)}$	$T^{(1)}$	$S^{(1)}$	D5	D6	DMR	W5	W6	WMR	A5	A6
Normal	10	91	98	84	87	175	161	145	98	159	138	98	159	140
	40	315	410	462	168	-	623	550	289	634	591	377	650	626
Uniform	10	-	-	-	-	-	175	157	106	107	109	131	217	220
	40	-	-	-	-	-	672	662	370	671	651	512	725	758
Double Exponential	10	188	217	201	199	252	261	254	187	274	242	202	236	197
	40	693	769	774	456	-	822	810	674	806	809	744	788	790
Cauchy	10	517	549	545	608	399	565	591	523	573	573	564	516	460
	40	975	983	984	1000	-	990	992	986	991	992	992	984	988
Logistic	10	120	141	131	127	203	214	189	108	220	187	117	210	179
	40	449	548	587	278	-	693	653	403	699	675	509	699	685
Exponential	10	-	-	-	-	-	79	117	155	76	143	188	106	152
	40	-	-	-	-	-	371	412	439	434	523	581	702	738
χ_1^2	10	90	89	95	69	31	40	64	67	48	65	77	49	62
	40	89	93	113	113	-	31	53	99	39	63	98	49	70
χ_4^2	10	55	57	48	44	78	82	72	54	81	70	46	82	83
	40	71	75	78	26	-	143	114	64	140	120	67	133	129

Note 1: Values for these statistics were taken from reference 51, Table III.

as compared to the corresponding tests using Models 5 and 6. Note that every test based on Models 5 and 6 is superior to all tests reported by Littell, McClave and Offen for the normal, double exponential, and logistic alternatives. Results for the uniform and exponential show the superiority of A5 and A6. Comparisons for the Cauchy indicate all test statistics are competitive. The χ^2 results exhibit a curious behavior. Like the T and S statistics, D5, W5 and A5 all show powers below the alpha level for some sample size. Thus, it appears that the statistics based on Model 5 are biased toward the χ_1^2 distribution. This same phenomena occurred in all eight test statistics when the alternative distribution was a gamma with shape parameter 4 and in the test statistics based on Models 5 and 6 when the alternative was the lambda distribution described earlier. These results indicate a bias of the test statistics toward the gamma and lambda distributions. Results of the χ_4^2 distribution were unexpected. For sample size 40, the new test statistics based on Models 5 and 6 show approximately 100 percent improvement in power over their corresponding classical test statistic.

With respect to the goodness of fit tests proposed for the extreme value distribution it should be noted that these are equivalent to tests for the two parameter Weibull distribution if the data are transformed into new random variables $Y_i = - \ln X_i$ where $\{X_i\}$ $i=1,\ldots,n$ is the sample to be compared with the Weibull.

Summary

The level of precision which we were able to attain in distribution and density function estimation laid the foundation for extending the application of our new nonparametric models into the goodness of fit arena. After a brief survey of the literature, we proposed eight new test statistics, six based on adaptive Models 5 and 6, and two of the modified EDF class. The generation of critical values and the Monte Carlo mechanics of the power studies was presented for goodness of fit tests for the normal and extreme value distributions. Appendices 3 and 4 contain much of the tabular results. What the power comparisons showed was that tests based on Models 5 and 6 were competitive when the null distribution was normal, and competitive, if not superior, when the null distribution was the extreme value. The magnitude of the improvement in power in the extreme value tests against normal, double exponential, and logistic alternatives strongly suggests that these new tests are superior over various alternatives. Tests for the two parameter Weibull are also possible since they are equivalent with tests for the extreme value distribution.

Thus far, we have been successful in distribution and density estimation, and goodness of fit testing.

The next chapter will venture into the realm of parametric estimation using our nonparametric distribution and density function models.

VI. Location Parameter Estimation for Symmetric Distributions

Introduction

Given a random sample of size n from a univariate continuous probability distribution, we have already generated nonparametric estimates of the distribution, density, and hazard functions as well as proposed new goodness of fit tests. Rather than a complete distribution estimate, one may wish to estimate only certain characteristics of the distribution. While the nonparametric procedure holds promise for estimating parameters from an assumed model in general, we now propose to examine one specific class of estimates, namely the estimates of the location parameter of a symmetric family of distributions. Our treatment begins with a literature overview of location estimates and a discussion of the concept of robustness. Many of the estimators identified were used in the celebrated Princeton robustness study (Ref 5). Because of the performance of the new nonparametric models in approximating underlying distributions, it was conjectured that estimators based on the models might exhibit some useful robust characteristics in the location problem. Based on some very elementary concepts of trimming and Winsorizing, we propose some 48 new estimators of the location parameter using these new models. Estimator evaluation is accomplished in terms of standardized empirical variances determined from a Monte Carlo analysis considering samples of size 20. Comparisons of estimators are made using relative deficiencies, both average and maximum, over subsets of nine alternate distributions. A large number of pairwise comparisons are graphically illustrated via deficiency plots. Finally, robustness characteristics are evaluated in the form of stylized sensitivity curves. The judicious use of the tables and figures of this chapter should allow an analyst to judge which estimator is appropriate for the alternative distributions he may expect. We include twelve other estimators for comparative purposes.

Historical Survey

Like goodness of fit tests, parameter estimation has not suffered from lack of attention in the literature. In this section we will briefly examine some recent studies which bear on the present investigation. We will limit our discussion to location parameter estimates of a symmetric distribution and considerations of robustness.

The concept of robustness is central to our investigation. Robustness, as defined by Hampel, simply means that small changes in the assumed underlying model should cause only a small change in the performance of an estimator (Ref 30). Excellent surveys of the development of robust techniques are given by Stigler, Hogg, and Huber (Refs 38, 42, 91, 93).

Computational formulae and applications for common robust estimates are given by Moore, Hogg and, to a limited extent, David (Refs 19, 39, 60). Some specific estimators deserve mention, particularly the "alphabet" estimators. Huber developed M-estimators, based on minimizing a function of the form $\sum_i \rho(X_i-T)$ where ρ is an arbitrary function. Specific choices of ρ result in the estimator T being the sample mean, sample median, or a maximum likelihood estimator (Ref 41). Hampel introduced a family of piecewise linear M-estimators (Ref 5). Given combinations of order statistics form a general class known as L-estimators. Besides trimmed and Winsorized means, this class includes estimators given by Alam, Harter, Gastwirth and others (Refs 2, 26, 33).

A recent article by Chan and Rhodin introduces asymptotically best linear estimates based on a finite number of symmetrically ranked order statistics. These estimates are shown to be more efficient than optimally trimmed or Winsorized means (Ref 12). Estimators based on rank tests, such as the Hodges-Lehmann estimator, belong to the class of R-estimators (Ref 37). More recently, a family of D-estimators was investigated by Parr (Ref 61). Originally proposed by Wolfowitz, a D-estimator minimizes some discrepancy (such as the CVM distance) between the empirical distribution function and an underlying parametric family (Ref 108). Parr and Schucany have shown that D-estimation is a competitive technique in estimating the location parameter of symmetric distributions by using the normal distribution as a projection model (Ref 63). D-estimation using a weighted CVM discrepancy is discussed by Parr and DeWit (Ref 64). Shaler states the conditions for existence and consistency of minimum discrepancy estimates (Ref 79). Beran proposes and evaluates minimum Hellinger distance estimators based on a discrepancy using a density function estimate and the underlying density function (Ref 6). The relationship between these types of estimates and goodness of fit tests is given by Easterling (Ref 22). For an exhaustive bibliography of minimum distance estimation, refer to Parr (Ref 62).

Various adaptive procedures have emerged. Hogg lists variations of estimators based on kurtosis, the statistic and percentile ratios (Ref 38). Harter proposed a variant of Hogg's estimator using certain maximum likelihood estimates and kurtosis as a discriminant (Ref 60). Optimal boundaries for various discriminants were determined by Rugg (Ref 77). Numerous other studies have been conducted using discriminants and generalized projection families such as the GEP distribution or the t distribution. Adaptive techniques incorporating both classical estimation procedures and minimum distance constraints have recently been investigated (Refs 3, 11, 16, 17, 24, 32, 34, 43, 55).

Perhaps the single most comprehensive study of estimates of the location parameter of a symmetric distribution was the Princeton study (Ref 5). While analyzing some 68 estimators, the authors are quick to point out that their study is not exhaustive. Stigler presents an interesting comparison of some of the estimators used in the Princeton study. He uses 24 original data sets from famous experiments conducted in the 18th and 19th century to determine the parallax of the sum, the mean density of the earth, and the velocity of light. Both his comments, while quite negative toward a large set of new robust estimators, and the comments of various discussants provide a refreshing discussion of the use of robust procedures (Ref 92).

Proposed New Estimators

The construction of the new nonparametric cumulative and density estimators implicitly gives us a technique for parameter estimation. This analysis only attempts to begin to explore the various procedures for estimating the parameters of an underlying distribution. We chose the family of symmetric distributions for two reasons. First, estimates of the location parameter can be constructed in very simple forms since the mean, median, and mode of the density are identical.

Second, comparisons with other estimates are readily available.

To form the estimators we use four of our nonparametric models--Models 2, 4, 5, and 6. The means and medians of the models comprise the first eight new estimators. The means were calculated using a modified Simpson's Rule integration routine and the medians were found by inverting the distribution function estimate using a Newton-Raphson technique. Estimators of this type are identified by Mean-Mn, Median-Mn, etc. where Mn denotes Model n, n=2,4,5,6.

Two other families of estimators were formed. Modified trimmed means were calculated by symmetrically trimming a percentage of observations from each end of the original ordered sample and then calculating the sample mean of the nonparametric density defined by the remaining data points and our models. Five different levels of trimming were used. The estimators are designated α percent T-Mn where α is the trimming proportion, α=5(5)25). Modified Winsorized means were calculated based on the density function determined by the entire original sample. To calculate the modified Winsorized means, let α be the amount (percentage) of Winsorizing. Calculate $SF^{-1}(\alpha)$ and $SF^{-1}(1-\alpha)$ where SF is the nonparametric estimator of the distribution function. Then, the modified Winsorized mean, $\hat{x}_{\alpha}$, is given by:

$$\hat{x}_\alpha = \int_{SF^{-1}(\alpha)}^{SF^{-1}(1-\alpha)} x dSF(x) + \alpha(SF^{-1}(\alpha) + SF^{-1}(1-\alpha))$$

What we have effectively done is to take the mean of a mixed distribution formed by truncating the nonparametric density at $SF^{-1}(\alpha)$ and $SF^{-1}(1-\alpha)$ and letting these two endpoints have a finite probability, namely α. This is analogous to the Winsorized mean where sample points are mapped back to the order statistics corresponding to the amount of Winsorizing. Modified Winsorized means are designated by α percent W-Mn where α is the amount of symmetric Winsorizing, α=5(5)25. This gives us a total of forty-eight new estimators proposed.

Estimator Evaluation

Using the Princeton study as a guide, we conducted a limited Monte Carlo analysis of three estimators. We generated 1000 Monte Carlo samples of size 20 from nine different distributions including the normal, double exponential, Cauchy and six contaminated normals. The normal, double exponential and Cauchy distributions all had a zero location parameter and a unit scale parameter. The contaminated normals consisted of ε percent observations from a normal with zero mean and a scale parameter of three and $(1-\varepsilon)$ percent observations from a standard normal. The contamination percentages used were 5, 10, 15, 25, 50, and 75. These distributions will be designated ε percent 3N where ε is the contamination percentage.

The distributions were grouped into classes of alternatives to the normal, using the same groupings as the Princeton study. The gentle, reasonable alternatives include the normal 5% 3N, 10% 3N, 15% 3N and 25% 3N. Gentle, unreasonable alternatives include 50% 3N and 75% 3N. Vigorous alternatives include the double exponential and the Cauchy. A fourth set of alternatives considered was the set of all distributions tested except the Cauchy. No specific short tailed distribution was tested in this portion of the study. The groupings relate to how the analyst views the practical world his data comes from. Using the normal distribution as a model of reality, the sampling mechanism and underlying process may allow for only mild departures from normality. In other cases, an analyst may want protection against a larger deviation in his underlying view of the world. By generating various sets of alternatives, we may infer the conditions under which certain estimators perform better.

For each random sample we calculated all 48 estimates. For comparison purposes, we also included the sample mean, sample median, and ten M-estimators, consisting of six Hubers and four Hampels. The Hubers includes H20, H17, H15, H12, H10, and H07, while the Hampels used were 25A, 21A, 17A, and 12A. For a complete definition of these estimators and their associated parameters, refer to the Princeton study (Ref 5). Results of this Monte Carlo study for the Hubers and Hampels are in excellent agreement with the variances given in that same study.

Table VI.1 gives the standardized empirical variances for all sixty estimators used. Table entries represent the mean square error of the estimate multiplied by the sample size. Even when actual variances are available, we used the empirical ones to compare estimators to keep relative rankings consistent. For example, the true variance of the sample mean is 1/n for an underlying normal population. Thus the table entry should be 1.000. We, however, will use our empirical variance entry of 0.990 for relative comparisons.

To synthesize this information into meaningful comparisons, we introduce the concept of deficiencies. The deficiency of an estimator is akin to Hogg's "insurance premium" of using a robust estimate. It is the penalty you pay if the distributional assumption, you chose not to make, is actually correct. Deficiencies are calculated as follows: Let T_{ij} be an estimator of type i over a set of test distributions indexed by j. Now let $T_{min,j}$ be the estimator with the smallest standardized empirical variance for distribution j.

TABLE VI.1

STANDARDIZED EMPIRICAL VARIANCES OF THE ESTIMATORS FOR SAMPLE SIZE 20

	Estimator	Normal	Double Exponential	Cauchy	5% 3N	10% 3N	15% 3N	25% 3N	50% 3N	75% 3N
1	Mean	.990	.975	4987.3	1.406	1.758	2.288	3.034	4.970	7.039
2	Median	1.432	.658	2.6	1.609	1.656	1.828	2.226	3.585	6.379
3	Mean-M2	.994	.905	2209.8	1.263	1.507	1.963	2.687	4.716	6.980
4	5%W-M2	1.002	.848	1156.8	1.250	1.450	1.852	2.503	4.376	6.702
5	10%W-M2	1.005	.830	60.0	1.227	1.397	1.771	2.395	4.289	6.709
6	15%W-M2	1.006	.826	35.9	1.222	1.379	1.774	2.351	4.256	6.711
7	20%W-M2	1.008	.822	23.2	1.223	1.377	1.739	2.334	4.219	6.677
8	25%W-M2	1.010	.814	17.7	1.224	1.374	1.733	2.321	4.167	6.617
9	5%T-M2	1.016	.812	17.0	1.179	1.292	1.601	2.219	4.255	6.776
10	10%T-M2	1.058	.753	6.9	1.198	1.261	1.490	1.946	3.822	6.597
11	15%T-M2	1.096	.704	4.5	1.235	1.281	1.484	1.861	3.490	6.404
12	20%T-M2	1.138	.669	3.5	1.289	1.320	1.508	1.862	3.273	6.239
13	25%T-M2	1.118	.644	3.0	1.333	1.370	1.540	1.894	3.174	6.128
14	Median-M2	1.015	.780	10.7	1.203	1.326	1.648	2.229	4.050	6.538
15	Mean-M4	1.022	1.043	6542.7	1.486	1.877	1.373	3.192	5.400	7.420
16	5%W-M4	1.006	.963	1920.0	1.340	1.623	2.097	2.876	5.020	7.161
17	10%W-M4	1.002	.908	62.1	1.266	1.479	1.914	2.656	4.744	6.993
18	15%W-M4	1.003	.863	35.8	1.235	1.408	1.801	2.485	4.500	6.836

TABLE VI.1--Continued

	Estimator	Normal	Double Exponential	Cauchy	5% 3N	10% 3N	15% 3N	25% 3N	50% 3N	75% 3N
19	20%W-M4	1.008	.820	23.0	1.220	1.357	1.708	2.330	4.256	6.688
20	25%W-M4	1.016	.777	14.8	1.213	1.317	1.624	2.181	4.009	6.548
21	5%T-M4	1.045	.885	23.0	1.215	1.338	1.715	2.384	4.657	7.102
22	10%T-M4	1.088	.822	8.3	1.234	1.307	1.573	2.121	4.217	7.071
23	15%T-M4	1.127	.758	5.0	1.230	1.318	1.539	2.001	3.743	6.894
24	20%T-M4	1.149	.692	3.9	1.280	1.324	1.560	1.918	3.450	6.644
25	25%T-M4	1.186	.674	3.3	1.321	1.364	1.533	1.905	3.273	6.394
26	Median-M4	1.067	.668	4.1	1.250	1.301	1.499	1.923	3.495	6.335
27	Mean-M5	1.002	1.053	6542.7	1.459	1.874	2.378	3.192	5.453	7.559
28	5%W-M5	.997	.976	1920.0	1.324	1.623	2.102	2.882	5.081	7.296
29	10%W-M5	.998	.916	62.1	1.254	1.479	1.918	2.658	4.777	7.074
30	15%W-M5	1.002	.865	35.8	1.225	1.407	1.802	2.482	4.504	6.865
31	20%W-M5	1.011	.816	23.0	1.210	1.355	1.706	2.322	4.230	6.663
32	25%W-M5	1.027	.768	14.8	1.205	1.315	1.620	2.168	3.952	6.473
33	5%T-M5	1.027	.888	23.0	1.189	1.332	1.672	2.362	4.652	7.188
34	10%T-M5	1.055	.820	8.2	1.203	1.280	1.542	2.052	4.158	6.995
35	15%T-M5	1.110	.748	4.9	1.217	1.293	1.506	1.913	3.659	1.767
36	20%T-M5	1.131	.696	3.8	1.268	1.314	1.525	1.880	3.407	6.530

TABLE VI.1--Continued

	Estimator	Normal	Double Exponential	Cauchy	5% 3N	10% 3N	15% 3N	25% 3N	50% 3N	75% 3N
37	25%T-M5	1.179	.663	3.2	1.329	1.356	1.529	1.897	3.232	6.322
38	Median-M5	1.118	.665	4.0	1.266	1.310	1.507	1.908	3.354	6.128
39	Mean-M6	1.000	1.035	2944.3	1.359	1.754	2.336	3.326	5.493	7.452
40	5%W-M6	.996	.957	866.0	1.278	1.558	2.049	2.893	5.056	7.209
41	10%W-M6	1.000	.897	45.5	1.224	1.425	1.849	2.597	4.719	7.023
42	15%W-M6	1.009	.837	25.0	1.204	1.352	1.715	2.360	4.368	6.794
43	20%W-M6	1.028	.779	13.6	1.194	1.302	1.603	2.156	4.013	6.580
44	25%W-M6	1.055	.731	7.8	1.204	1.279	1.528	2.008	3.723	6.410
45	5%T-M6	1.022	.882	18.6	1.191	1.323	1.679	2.385	4.663	7.125
46	10%T-M6	1.046	.808	8.3	1.196	1.278	1.535	2.056	4.156	6.923
47	15%T-M6	1.103	.740	5.0	1.218	1.285	1.502	1.914	3.659	6.700
48	20%T-M6	1.126	.693	3.8	1.264	1.308	1.517	1.871	3.388	6.491
49	25%T-M6	1.171	.662	3.2	1.319	1.350	1.517	1.883	3.234	6.275
50	Median-M6	1.214	.630	2.8	1.371	1.406	1.570	1.952	3.205	6.013
51	H20	1.006	.828	10.7	1.196	1.332	1.678	2.530	4.462	6.663
52	H17	1.018	.789	7.5	1.181	1.276	1.583	2.287	4.219	6.596
53	H15	1.034	.762	5.8	1.182	1.258	1.549	2.133	4.004	6.496
54	H12	1.073	.718	4.4	1.202	1.270	1.524	1.956	3.622	6.233
55	H10	1.109	.684	3.7	1.227	1.295	1.526	1.882	3.376	6.034

TABLE VI.1--<u>Continued</u>

	Estimator	Normal	Normal Exponential	Cauchy	5% 3N	10% 3N	15% 3N	25% 3N	50% 3N	75% 3N
56	H07	1.176	.634	2.9	1.290	1.360	1.567	1.848	3.185	5.738
57	25A	1.046	.745	3.6	1.167	1.267	1.565	2.111	3.946	6.496
58	21A	1.076	.721	3.3	1.183	1.274	1.546	1.987	3.755	6.428
59	17A	1.120	.688	2.9	1.219	1.302	1.540	1.890	3.502	6.252
60	12A	1.182	.658	2.6	1.273	1.361	1.575	1.849	3.307	5.991

Define efficiency $(T_{ij}) = \dfrac{\text{variance of } T_{\min,j}}{\text{variance of } T_{ij}}$.

Then, deficiency = 1 - efficiency. Naturally, one prefers deficiencies near zero.

For each set of alternatives we calculated two measures of deficiency, the maximum deficiency of an estimator for all distribution is the class and the average deficiency over the class. Again, depending on the sampling situation, one criterion may be more appropriate than another. An analyst faced with a large penalty for poor performance, would probably prefer the maximum relative efficiency criterion.

Tables VI.2 through VI.5 rank each of the 60 estimators with respect to both maximum relative and average relative deficiencies under each different set of alternative distributions. Notice in particular, the excellent performance of the new estimators under gentle, reasonable alternatives and under all alternatives except Cauchy (Tables VI.2 and VI.5). Of particular note is the fact that only one modified Winsorized mean is among the 20 leading estimators under either relative efficiency criterion for any set of alternatives. This estimator, 25%W-M6, is clearly the best of the modified Winsorized estimators that was proposed. Under gentle, reasonable alternatives, the modified trimmed mean, 10%T-M2, seems to perform "better" than the other estimators for either

TABLE VI.2

ESTIMATORS RANKED BY RELATIVE DEFICIENCIES UNDER GENTLE, REASONABLE ALTERNATIVES

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
1	10%T-M2	.063662	1	10%T-M2	.029250
2	Median-M4	.071697	2	15%T-M2	.035338
3	H12	.076936	3	H12	.039304
4	25%W-M6	.079710	4	15%T-M6	.042349
5	21A	.079712	5	21A	.043087
6	15%T-M2	.096405	6	25%W-M6	.043197
7	10%T-M5	.099457	7	Median-M4	.044031
8	10%T-M6	.101103	8	15%T-M5	.044905
9	15%T-M6	.102162	9	10%T-M6	.045542
10	H10	.106614	10	H10	.045879
11	15%T-M5	.107612	11	H15	.046170
12	Median-M5	.114106	12	25A	.047256
13	17A	.115875	13	10%T-M5	.048982
14	20%T-M6	.120092	14	17A	.050291
15	15%T-M4	.121191	15	20%T-M6	.053859
16	20%T-M5	.124137	16	Median-M5	.055790
17	25A	.124930	17	20%T-M5	.058061
18	10%T-M4	.128861	18	20%T-M2	.058942

TABLE VI.2--Continued

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
19	20%T-M2	.129572	19	5%T-M2	.060360
20	H15	.133671	20	H17	.061708
21	20%T-M4	.137705	21	20%W-M6	.062180
22	20%W-M6	.143129	22	15%T-M4	.066016
23	25%W-M5	.147845	23	25%W-M5	.068518
24	25%W-M4	.152944	24	25%W-M4	.069373
25	25%T-M6	.153950	25	20%T-M4	.072117
26	H07	.157750	26	10%T-M4	.073292
27	25%T-M5	.159765	27	Median-M2	.075153
28	12A	.162321	28	25%T-M6	.075465
29	25%T-M2	.163230	29	12A	.075954
30	25%T-M4	.164770	30	H07	.076328
31	5%T-M2	.167356	31	25%T-M5	.081813
32	Median-M2	.171057	32	25%T-M4	.084259
33	Median-M6	.184283	33	25%T-M2	.085990
34	H17	.192005	34	5%T-M5	.087989
35	25%W-M2	.203785	35	5%T-M6	.088216
36	20%W-M5	.204401	36	20%W-M5	.092347

TABLE VI.2--Continued

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
37	20%W-M4	.207193	37	15%W-M6	.094007
38	20%W-M2	.208556	38	20%W-M4	.094338
39	15%W-M2	.214205	39	H20	.096209
40	15%W-M6	.217156	40	25%W-M2	.099507
41	5%T-M5	.217902	41	20%W-M2	.100872
42	5%T-M4	.225132	42	5%T-M4	.102299
43	5%T-M6	.225225	43	15%W-M2	.102309
44	10%W-M2	.228627	44	Median-M6	.109341
45	15%W-M5	.255691	45	10%W-M2	.110717
46	15%W-M4	.256382	46	15%W-M5	.119410
47	5%W-M2	.261713	47	15%W-M4	.121415
48	H20	.269639	48	10%W-M6	.131827
49	10%W-M6	.288435	49	5%W-M2	.134161
50	10%W-M4	.304257	50	10%W-M5	.151421
51	10%W-M5	.304733	51	10%W-M4	.153471
52	Median	.308370	52	Mean-M2	.160306
53	Mean-M2	.312396	53	5%W-M6	.184499
54	5%W-M4	.357623	54	5%W-M5	.200633

TABLE VI.2--Continued

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
55	5%W-M5	.358778	55	5%W-M4	.203771
56	5%W-M6	.361396	56	Median	.236301
57	Mean	.391109	57	Mean	.239344
58	Mean-M4	.421147	58	Mean-M6	.248705
59	Mean-M5	.421247	59	Mean-M5	.267495
60	Mean-M6	.444439	60	Mean-M4	.274235

TABLE VI.3

ESTIMATORS RANKED BY RELATIVE DEFICIENCIES UNDER VIGOROUS ALTERNATIVES

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
1	12A	.042335	1	12A	.021168
2	Median	.042677	2	Median	.031582
3	Median-M6	.078546	3	Median-M6	.039273
4	17A	.114072	4	H07	.062906
5	H07	.119912	5	25%T-M2	.077423
6	25%T-M2	.133690	6	17A	.099370
7	25%T-M5	.093204	7	25%T-M5	.121496
8	25%T-M6	.202563	8	25%T-M6	.025646
9	25%T-M4	.206457	9	25%T-M4	.135768
10	21A	.208983	10	20%T-M2	.155183
11	20%T-M2	.252754	11	21A	.167309
12	25A	.288277	12	H10	.188831
13	H10	.298428	13	Median-M5	.206141
14	20%T-M6	.323308	14	20%T-M6	.206994
15	20%T-M5	.325717	15	20%T-M5	.210123
16	20%T-M4	.331943	16	20%T-M4	.210946
17	Median-M5	.359727	17	25A	.221070
18	Median-M4	.365213	18	Median-M4	.224578

TABLE VI.3--Continued

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
19	H12	.404083	19	15%T-M2	.262810
20	15%T-M2	.420788	20	H12	.263016
21	15%T-M5	.470095	21	15%T-M5	.313458
22	15%T-M4	.477791	22	15%T-M6	.314017
23	15%T-M6	.480075	23	15%T-M4	.323273
24	H15	.554150	24	H15	.363592
25	10%T-M2	.625332	25	10%T-M2	.394410
26	H17	.652505	26	25%W-M6	.403330
27	25%W-M6	.668621	27	H17	.426897
28	10%T-M5	.684490	28	10%T-M6	.453780
29	10%T-M4	.686053	29	10%T-M5	.457924
30	10%T-M6	.687374	30	10%T-M4	.459581
31	Median-M2	.757874	31	Median-M2	.474987
32	H20	.758589	32	H20	.498845
33	20%W-M6	.809498	33	20%W-M6	.500364
34	25%W-M5	.824678	34	25%W-M5	.502238
35	25%W-M4	.824851	35	25%W-M4	.507083
36	5%T-M2	.847017	36	5%T-M2	.535588

TABLE VI.3--Continued

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
37	25%W-M2	.853166	37	25%W-M2	.539632
38	5%T-M6	.860412	38	20%W-M5	.557652
39	5%-M5	.887383	39	20%W-M4	.559283
40	5%T-M4	.887480	40	20%W-M2	.560797
41	20%W-M5	.887495	41	15%W-M6	.571752
42	20%W-M4	.887505	42	5%T-M6	.572920
43	20%W-M2	.888242	43	15%W-M2	.582099
44	15%W-M6	.896433	44	5%T-M4	.587543
45	15%W-M4	.927574	45	5%T-M5	.588950
46	15%W-M5	.927596	46	15%W-M4	.598628
47	15%W-M2	.927713	47	10%W-M2	.598876
48	10%W-M6	.942957	48	15%W-M5	.599549
49	10%W-M2	.956743	49	10%W-M6	.620070
50	10%W-M4	.958217	50	5%W-M2	.627464
51	10%W-M5	.958231	51	10%W-M4	.632035
52	5%W-M6	.997006	52	10%W-M5	.634954
53	5%W-M2	.997758	53	Mean-M2	.651078
54	5%W-M4	.998649	54	5%W-M6	.669306

TABLE VI.3--Continued

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
55	5%W-M5	.998649	55	5%W-M4	.672136
56	Mean-M2	.998827	56	5%W-M5	.676320
57	Mean-M6	.999119	57	Mean	.676628
58	Mean	.999480	58	Mean-M6	.695156
59	Mean-M4	.999604	59	Mean-M4	.697769
60	Mean-M5	.999604	60	Mean-M5	.700585

TABLE VI.4

ESTIMATORS RANKED BY RELATIVE DEFICIENCES UNDER GENTLE, UNREASONABLE ALTERNATIVES

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
1	H07	.003600	1	H07	.001800
2	12A	.042172	2	Median-M6	.027790
3	Median-M6	.045769	3	25%T-M2	.031796
4	H10	.059802	4	12A	.041198
5	25%T-M2	.063592	5	25%T-M6	.052131
6	Median-M5	.063647	6	H10	.054423
7	20%T-M2	.080241	7	25%T-M5	.055147
8	25%T-M6	.085599	8	20%T-M2	.055289
9	25%T-M5	.092348	9	Median-M5	.058770
10	17A	.093692	10	25%T-M4	.066496
11	Median-M4	.094271	11	17A	.087968
12	25%T-M4	.102542	12	20%T-M6	.089674
13	15%T-M2	.103952	13	Median-M4	.093105
14	Median	.114627	14	20%T-M5	.094836
15	20%T-M6	.115959	15	15%T-M2	.097307
16	20%T-M5	.121250	16	H12	.101556
17	H12	.123723	17	Median	.107518
18	20%T-M4	.136371	18	20%T-M4	.108286

TABLE VI.4--Continued

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
19	15%T-M6	.143623	19	25%W-M6	.126130
20	25%W-M6	.147503	20	21A	.131015
21	15%T-M5	.152044	21	15%T-M6	.138129
22	21A	.154743	22	15%T-M5	.142377
23	15%T-M4	.167608	23	10%T-M2	.149892
24	10%T-M2	.169622	24	25%W-M5	.155229
25	25A	.195760	25	25A	.156245
26	25%W-M5	.196924	26	15%T-M4	.159848
27	H15	.207350	27	H15	.161987
28	25%W-M4	.208421	28	25%W-M4	.166075
29	20%W-M6	.209229	29	20%W-M6	.168608
30	Median-M2	.216419	30	Median-M2	.169390
31	10%T-M6	.236383	31	25%W-M2	.185606
32	10%T-M5	.236802	32	H17	.188904
33	25%W-M2	.238377	33	20%W-M2	.194159
34	10%T-M4	.247433	34	20%W-M5	.194225
35	20%W-M2	.247712	35	20%W-M4	.198185
36	H17	.247721	36	15%W-M2	.199644

TABLE VI.4--Continued

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
37	20%W-M5	.249634	37	10%W-M2	.202373
38	5%T-M2	.254083	38	5%T-M2	.203643
39	15%W-M2	.254311	39	10%T-M6	.203747
40	20%W-M4	.254361	40	10%T-M5	.208256
41	10%W-M2	.259999	41	5%W-M2	.209238
42	15%W-M6	.273357	42	H20	.213818
43	5%W-M2	.274677	43	15%W-M6	.214401
44	H20	.288800	44	10%T-M4	.217984
45	15%W-M4	.294757	45	15%W-M4	.227697
46	15%W-M5	.295437	46	15%W-M5	.229814
47	5%T-M5	.317806	47	Mean-M2	.252489
48	5%T-M4	.318534	48	10%W-M6	.255181
49	5%T-M6	.319448	49	10%W-M4	.255227
50	Mean-M2	.327106	50	5%T-M4	.255294
51	10%W-M6	.327421	51	5%T-M6	.257035
52	10%W-M4	.330970	52	5%T-M5	.259727
53	10%W-M5	.335609	53	10%W-M5	.262232
54	Mean	.361369	54	Mean	.273076

TABLE VI.4--Continued

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
55	5%W-M4	.367768	55	5%W-M4	.283218
56	5%W-M6	.372241	56	5%W-M6	.288132
57	5%W-M5	.375366	57	5%W-M5	.294440
58	Mean-M4	.412330	58	Mean-M4	.319500
59	Mean-M5	.417961	59	Mean-M6	.326108
60	Mean-M6	.422207	60	Mean-M5	.329434

TABLE VI.5

ESTIMATORS RANKED BY RELATIVE DEFICIENCIES UNDER ALL ALTERNATIVES EXCEPT CAUCHY

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
1	Median-M4	.094271	1	H07	.048893
2	15%T-M2	.104832	2	H10	.052185
3	H10	.106614	3	Median-M5	.056131
4	Median-M5	.114106	4	20%T-M2	.057863
5	17A	.115875	5	15%T-M2	.059517
6	20%T-M6	.120092	6	Median-M4	.061289
7	H12	.233723	7	12A	.063063
8	20%T-M5	.124137	8	17A	.064008
9	20%T-M2	.129572	9	25%T-M2	.064338
10	20%T-M4	.137705	10	H12	.065198
11	25%W-M6	.147503	11	25%T-M6	.066290
12	15%T-M6	.147959	12	20%T-M6	.067415
13	25%T-M6	.153950	13	25%T-M5	.071144
14	21A	.154743	14	20%T-M5	.071814
15	15%T-M5	.156820	15	Median-M6	.075286
16	H07	.157750	16	21A	.075388
17	25%T-M5	.159765	17	25%W-M6	.075786
18	12A	.162321	18	10%T-M2	.076190

TABLE VI.5--Continued

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
19	25%T-M2	.163230	19	25%T-M4	.077421
20	25%T-M4	.164770	20	15%T-M6	.079495
21	15%T-M4	.168755	21	15%T-M5	.083263
22	10%T-M2	.169622	22	20%T-M4	.083388
23	Median-M6	.184283	23	25A	.087829
24	25A	.195760	24	H15	.090982
25	25%W-M5	.196924	25	15%T-M4	.102316
26	H15	.207350	26	25%W-M5	.104106
27	25%W-M4	.208421	27	20%W-M6	.104918
28	20%W-M6	.209229	28	10%T-M6	.106924
29	Median-M2	.216419	29	25%W-M4	.108541
30	10%T-M6	.236383	30	H17	.110955
31	10%T-M5	.236802	31	10%T-M5	.111597
32	25%W-M2	.238377	32	Median-M2	.113330
33	10%T-M4	.247433	33	5%T-M2	.116656
34	20%W-M2	.247712	34	10%T-M4	.129442
35	H17	.247721	35	20%W-M5	.134749
36	20%W-M5	.249634	36	25%W-M2	.136855

TABLE VI.5--Continued

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
37	5%T-M2	.254083	37	20%W-M4	.137390
38	15%W-M2	.254311	38	20%W-M2	.140754
39	20%W-M4	.254361	39	15%W-M6	.143239
40	10%W-M2	.259999	40	15%W-M2	.143415
41	15%W-M6	.273357	41	H20	.143472
42	5%W-M2	.274677	42	10%W-M2	.149917
43	H20	.288800	43	5%T-M6	.155072
44	15%W-M4	.294757	44	5%T-M5	.156240
45	15%W-M5	.295437	45	5%T-M4	.163711
46	Median	.308370	46	15%W-M5	.166022
47	5%T-M5	.317806	47	15%W-M4	.166519
48	5%T-M4	.318534	48	5%W-M2	.168307
49	5%T-M6	.319448	49	Median	.179902
50	Mean-M2	.327106	50	10%W-M6	.183335
51	10%W-M6	.327421	51	10%W-M4	.197958
52	10%W-M4	.330970	52	10%W-M5	.199155
53	10%W-M5	.335609	63	Mean-M2	.201230
54	5%W-M4	.367768	54	5%W-M6	.230046

TABLE VI.5--Continued

Rank	Estimate	Maximum Relative Deficiency	Rank	Estimate	Average Relative Deficiency
55	5%W-M6	.372241	55	5%W-M4	.241364
56	5%W-M5	.375366	56	5%W-M5	.243254
57	Mean	.391109	57	Mean	.262081
58	Mean-M4	.421147	58	Mean-M6	.285867
59	Mean-M5	.421247	59	Mean-M5	.299738
60	Mean-M6	.444439	60	Mean-M4	.300764

deficiency criterion. For protection against vigorous alternatives Hampel's 12A seems to be the preferred choice.

As expected, no one estimator clearly surpassed the field. Depending on each sampling situation and the set of likely alternatives, the choice of an estimator is largely subject to analyst discretion.

Another comparison can be drawn between estimators or families of estimators. By plotting the deficiency of an estimator or a family of estimators under one alternative distribution versus another alternative, we get a graphical comparison of the relative performance of the estimators. Such deficiency plots, using the normal as one alternative in all cases, were constructed for the double exponential, Cauchy, and the contaminated normals. Figures 6.1 through 6.16 compare the deficiencies for the medians of some of the nonparametric models, the modified Winsorized estimator 25%W-M6, the family of Hubers, the family of Hampels, and the families of trimmed means for Models 2, 4, 5, and 6. For each specific alternative distribution, a set of two plots were generated for clarity. The first plot shows the comparison of the nonparametric medians and 25%W-M6 with the Hubers and Hampels. The medians on this plot are designated Mn where n is the model number. The second plot shows the comparison among the four families of trimmed means generated from Models 2, 4, 5, and 6. Each family is labeled by its corresponding

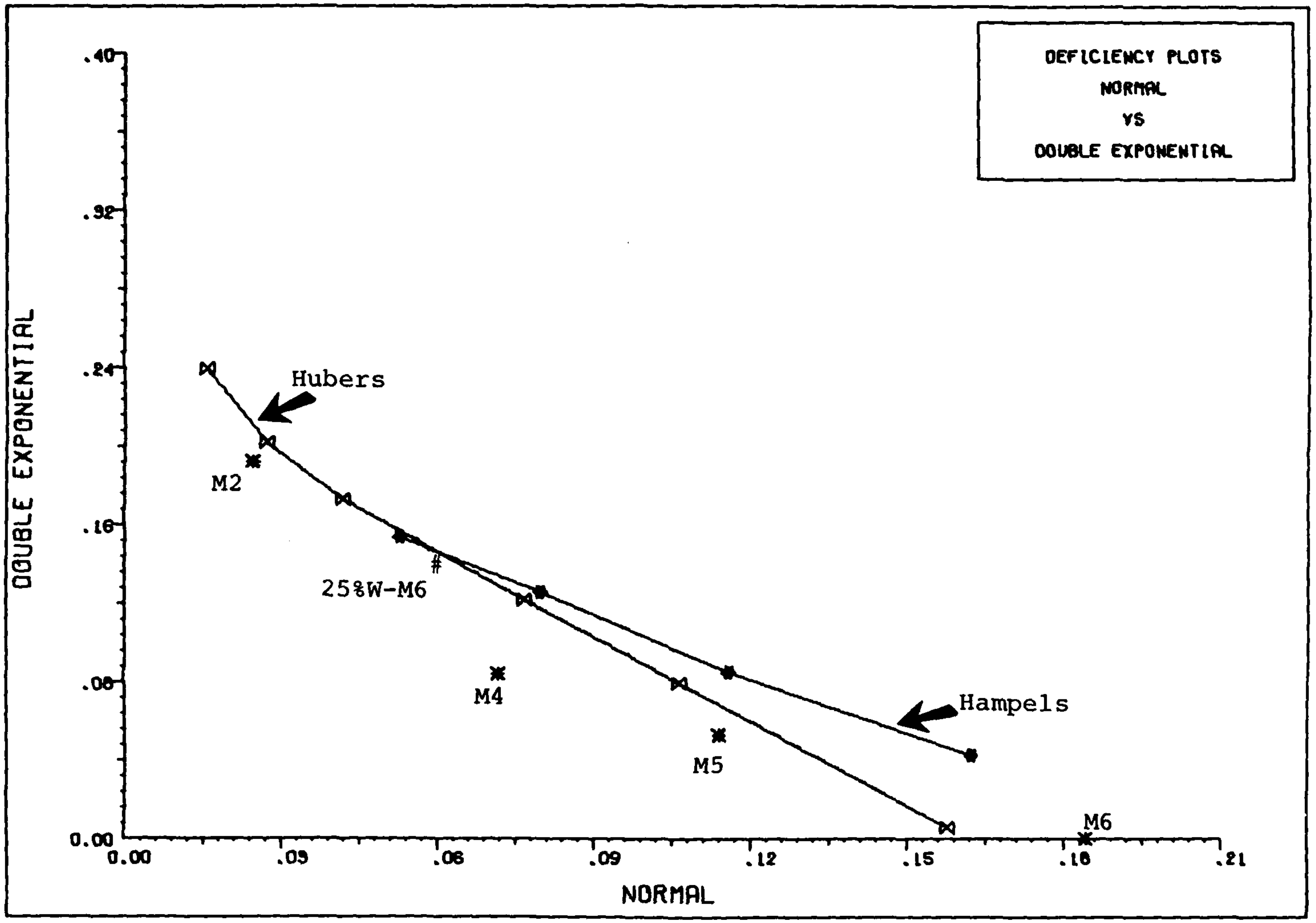

Figure 6.1. Deficiency Plot for Medians, 25%W-M6, Hubers and Hampels--Double Exponential vs Normal

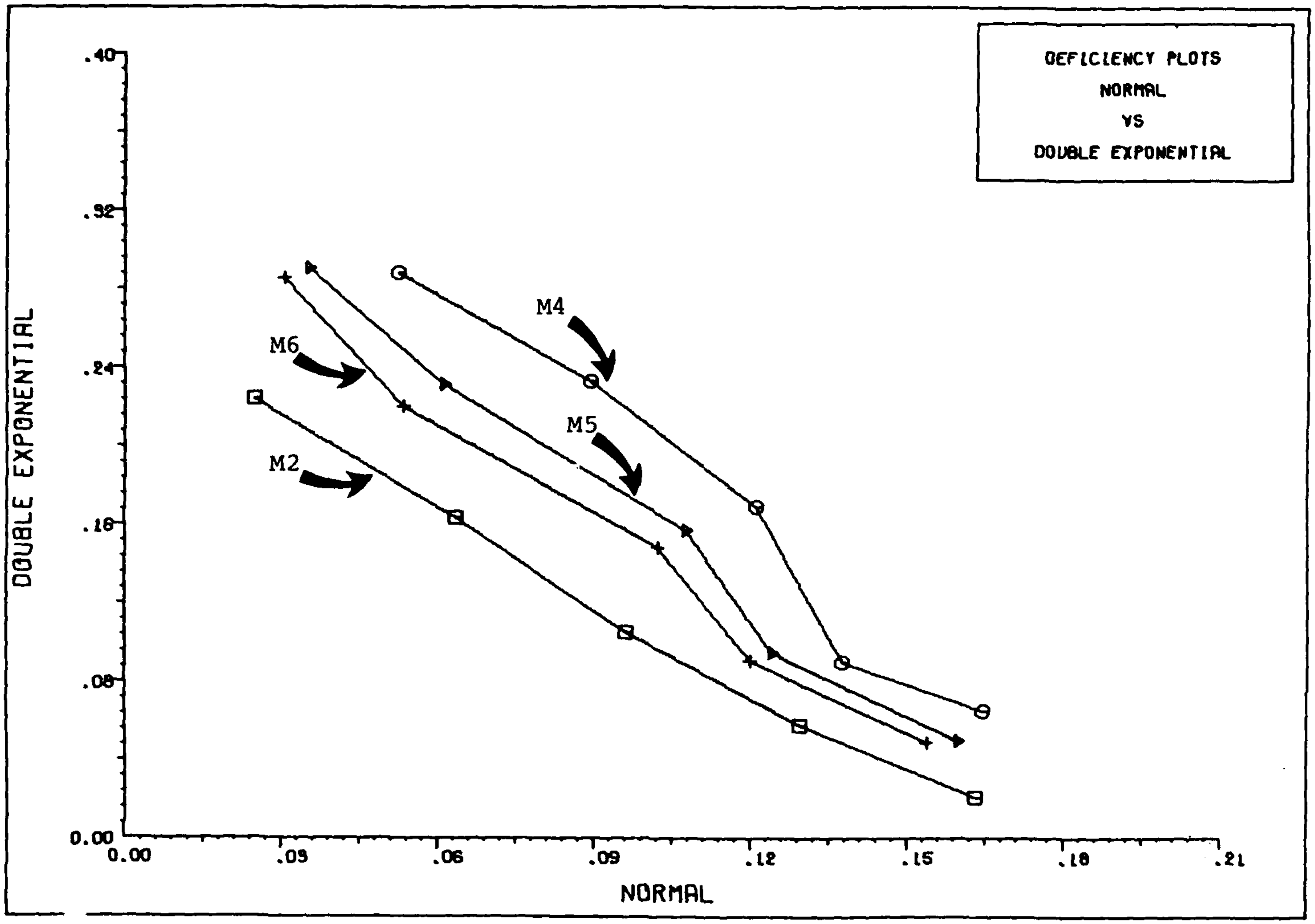

Figure 6.2. Deficiency Plot for Trimmed Means--
Double Exponential vs Normal

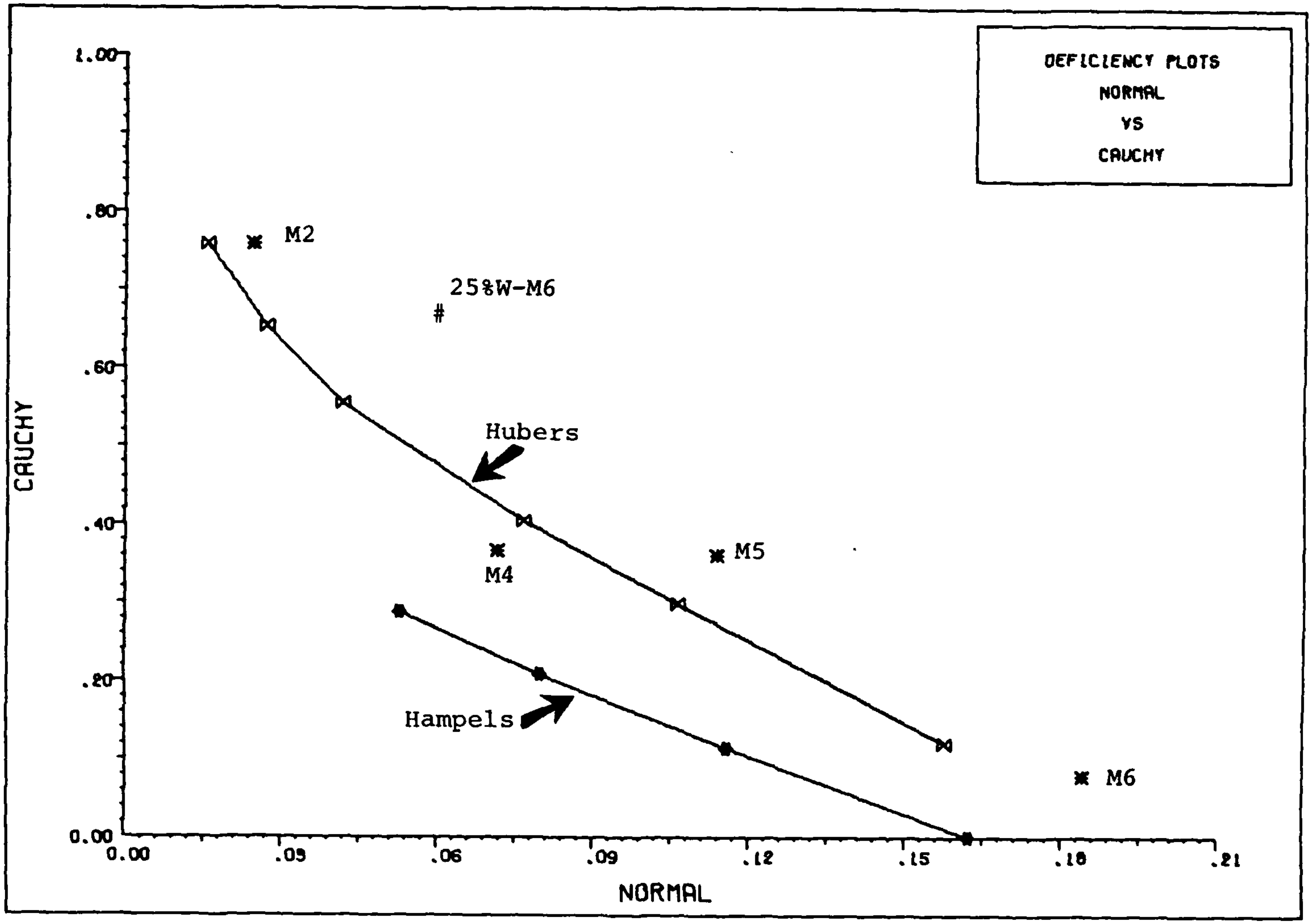

Figure 6.3. Deficiency Plot for Medians, 25%W-M6, Hubers and Hampels--Cauchy vs Normal

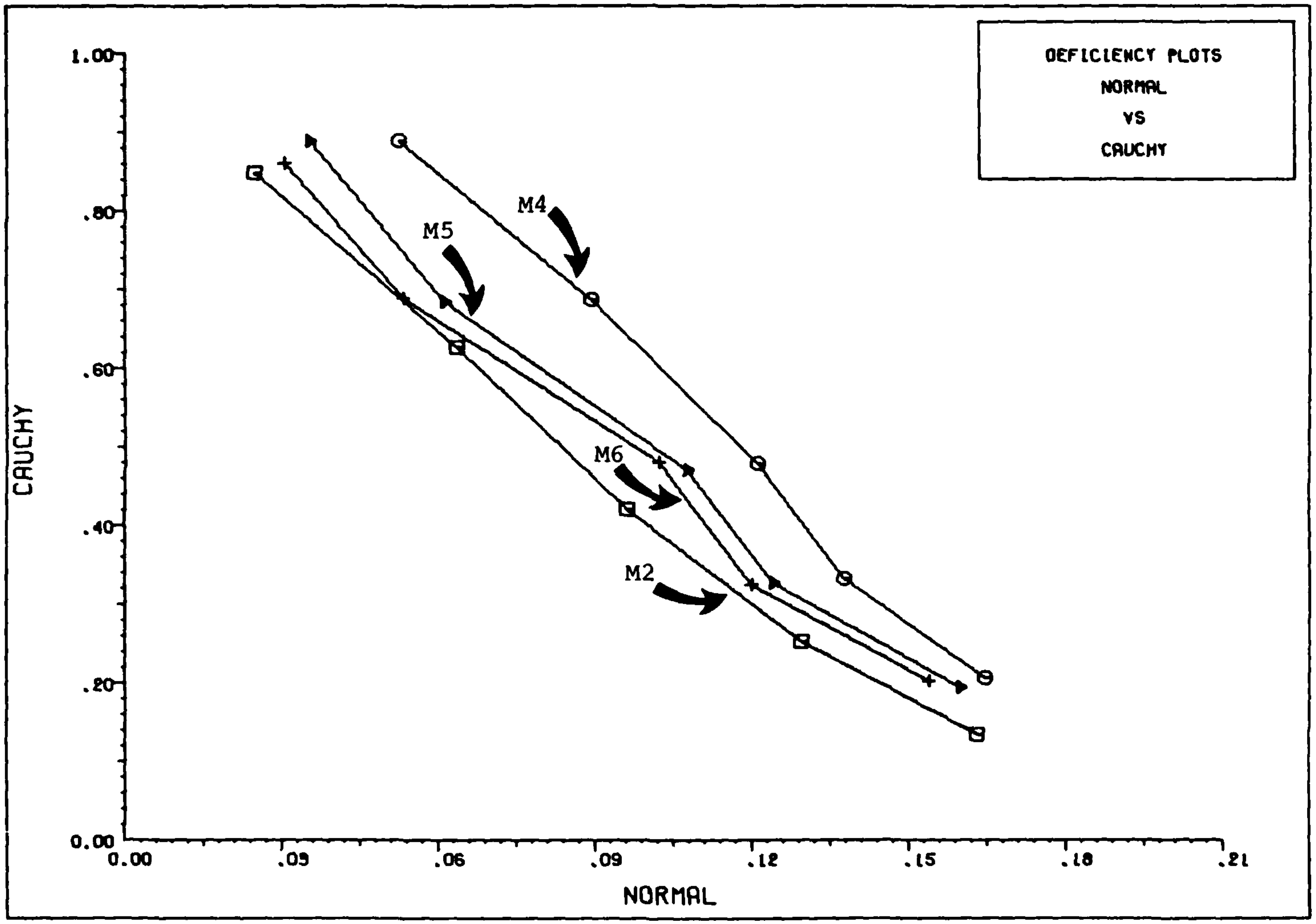

Figure 6.4. Deficiency Plot for Trimmed Means--
Cauchy vs Normal

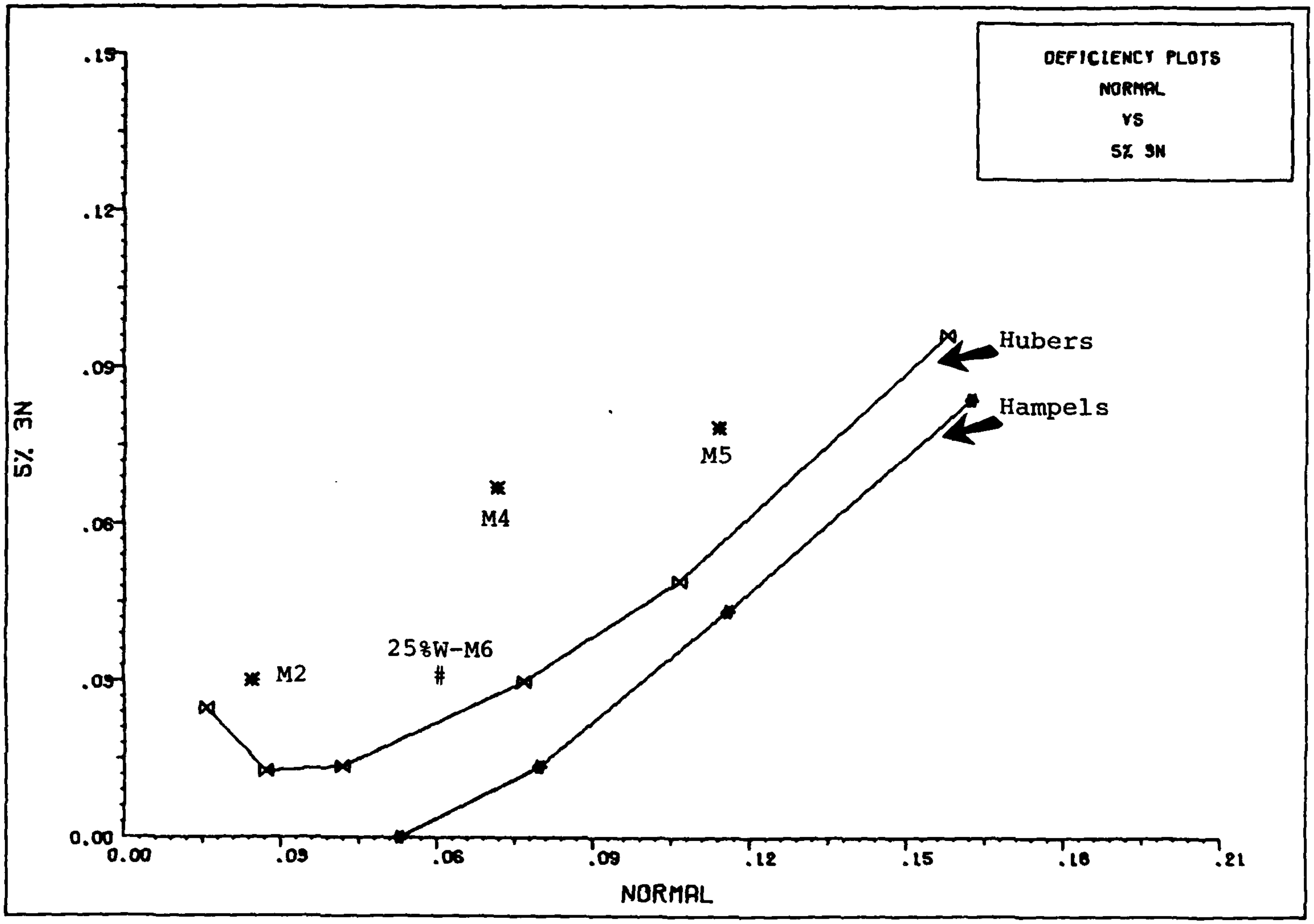

Figure 6.5. Deficiency Plot for Medians, 25%W-M6, Hubers and Hampels--5% 3N vs Normal

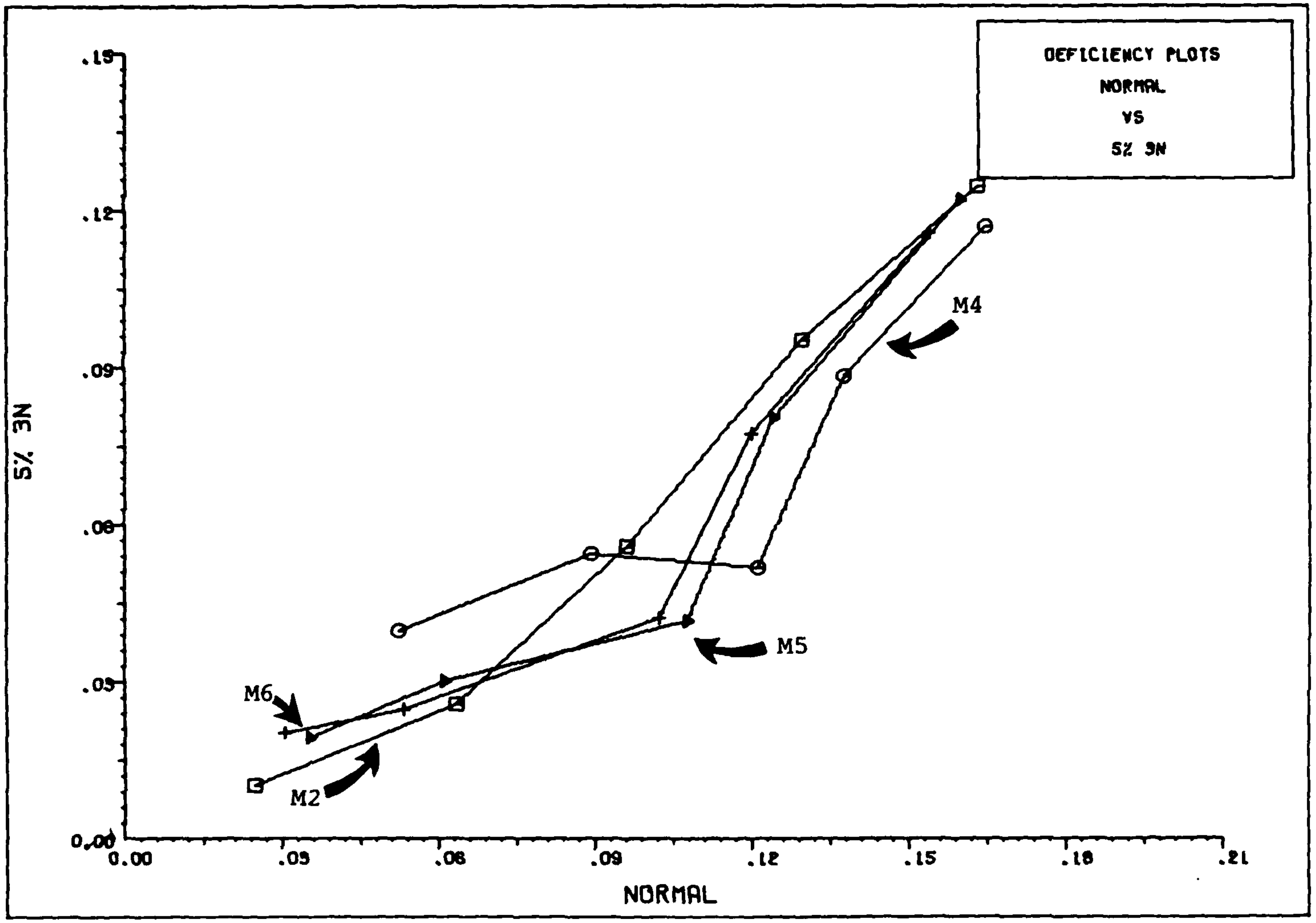

Figure 6.6. Deficiency Plot for Trimmed Means--
5% 3N vs Normal

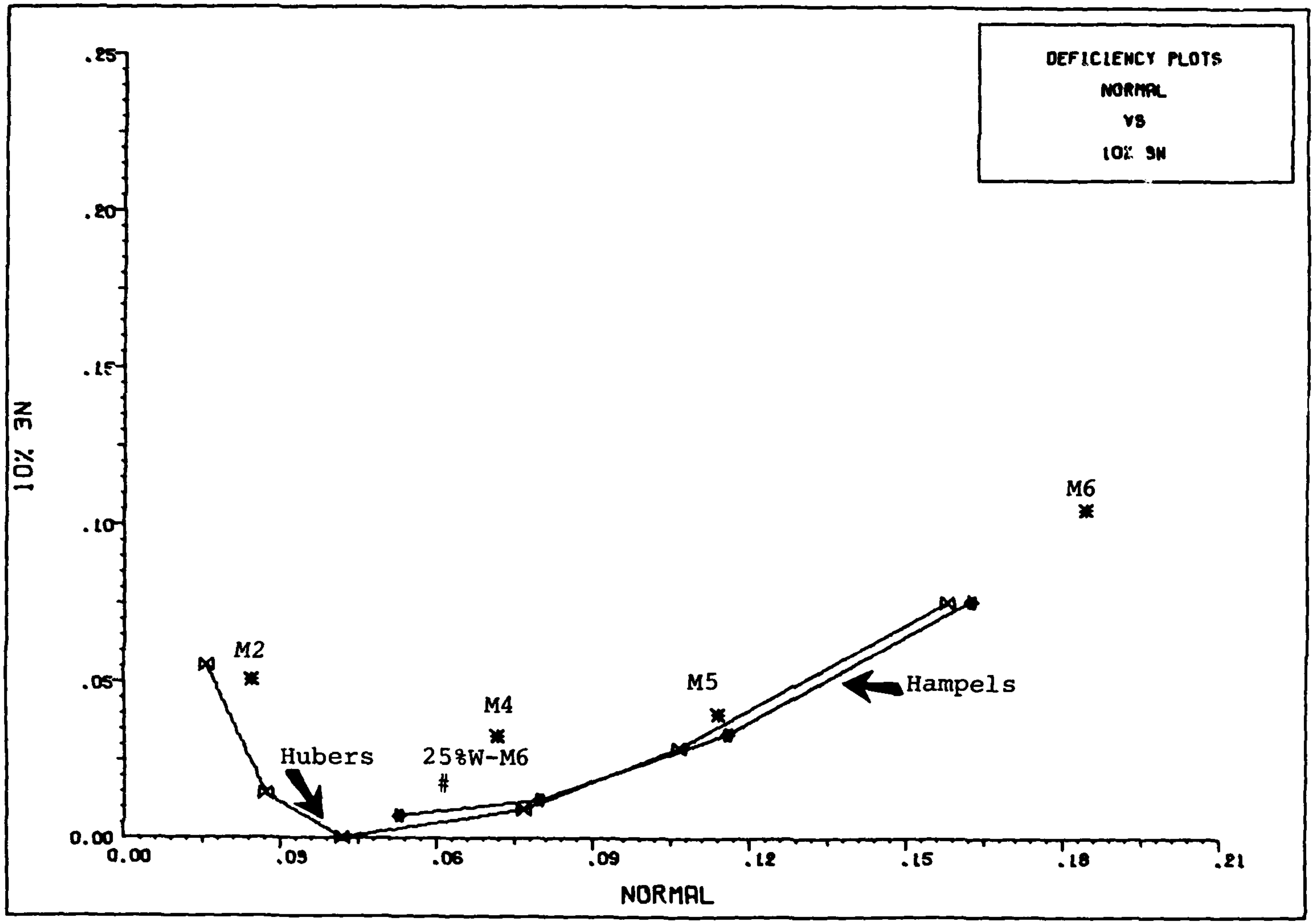

Figure 6.7. Deficiency Plot for Medians, 25%W-M6, Hubers and Hampels--10% 3N vs Normal

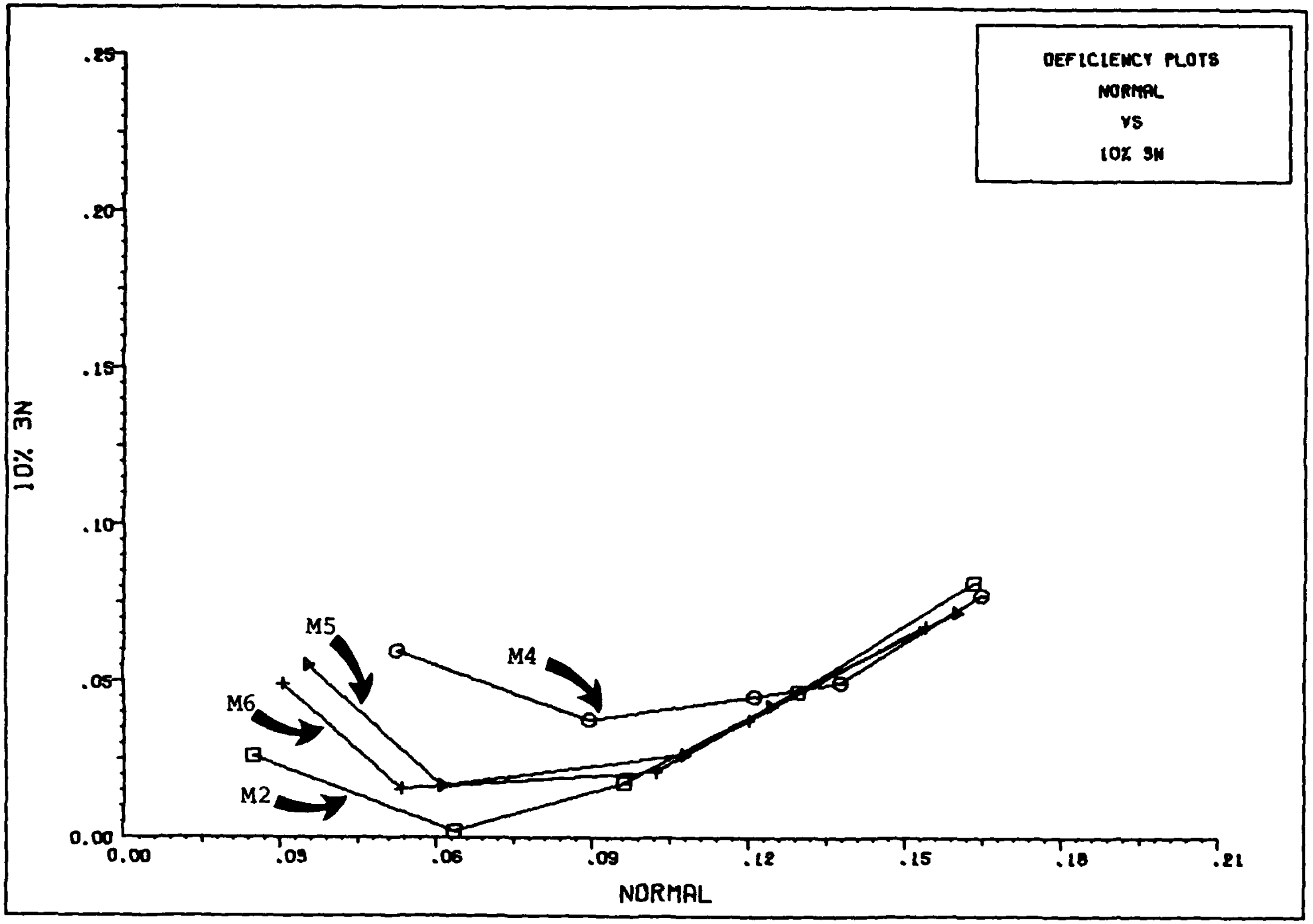

Figure 6.8. Deficiency Plot for Trimmed Means--10% 3N vs Normal

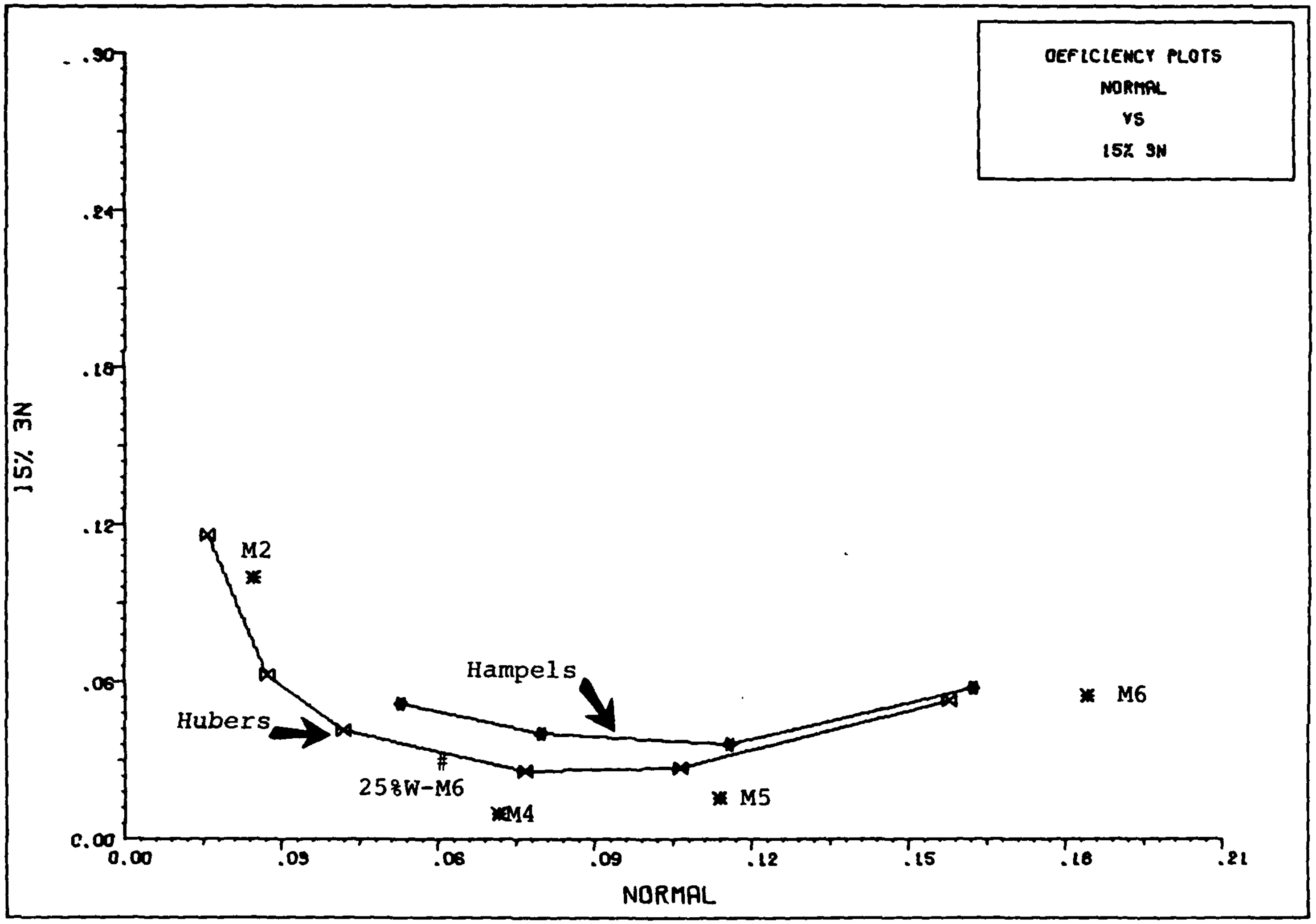

Figure 6.9. Deficiency Plot for Medians,25%W-M6, Hubers and Hampels--15% 3N vs Normal

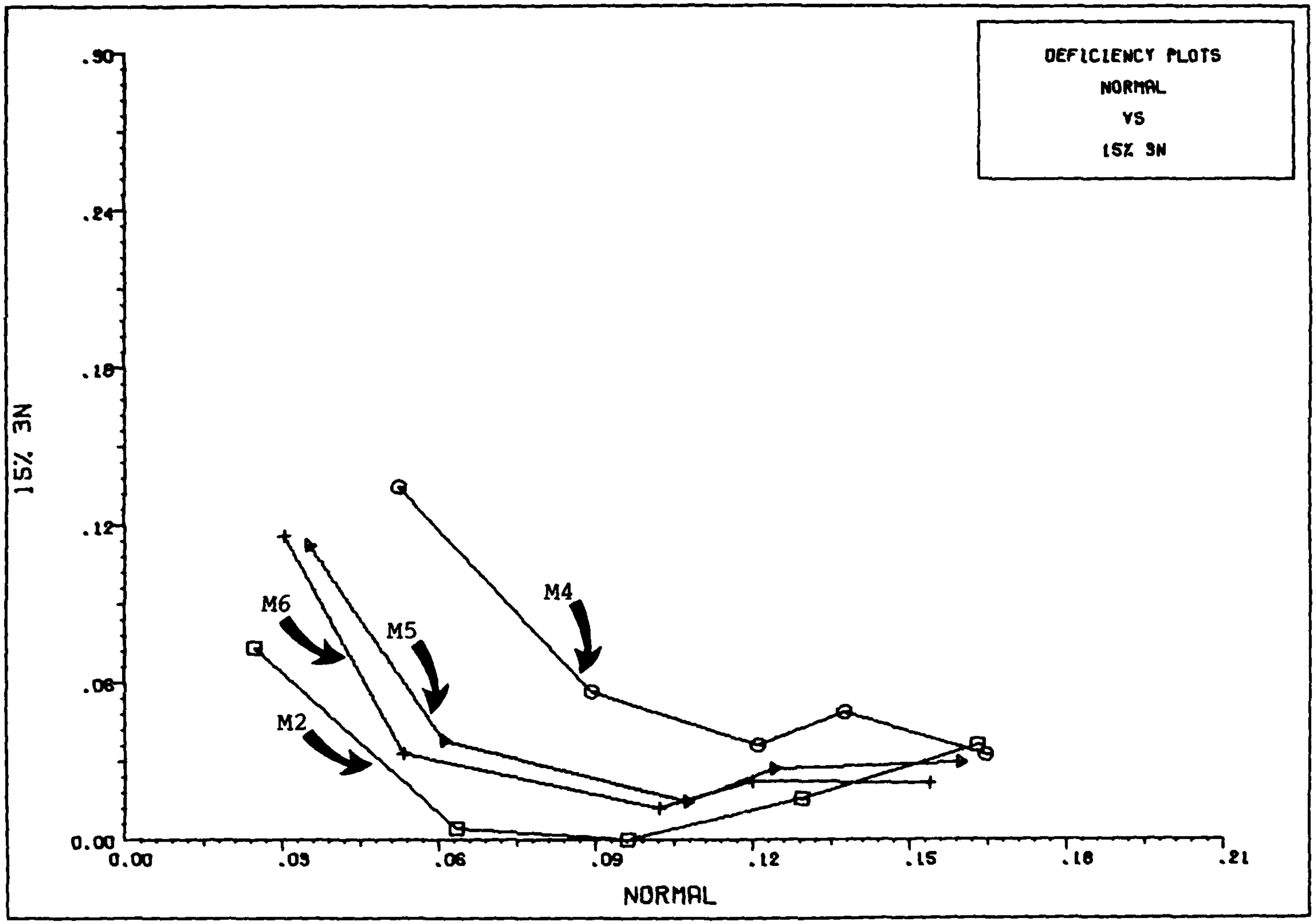

Figure 6.10. Deficiency Plot for Trimmed Means--
15% 3N vs Normal

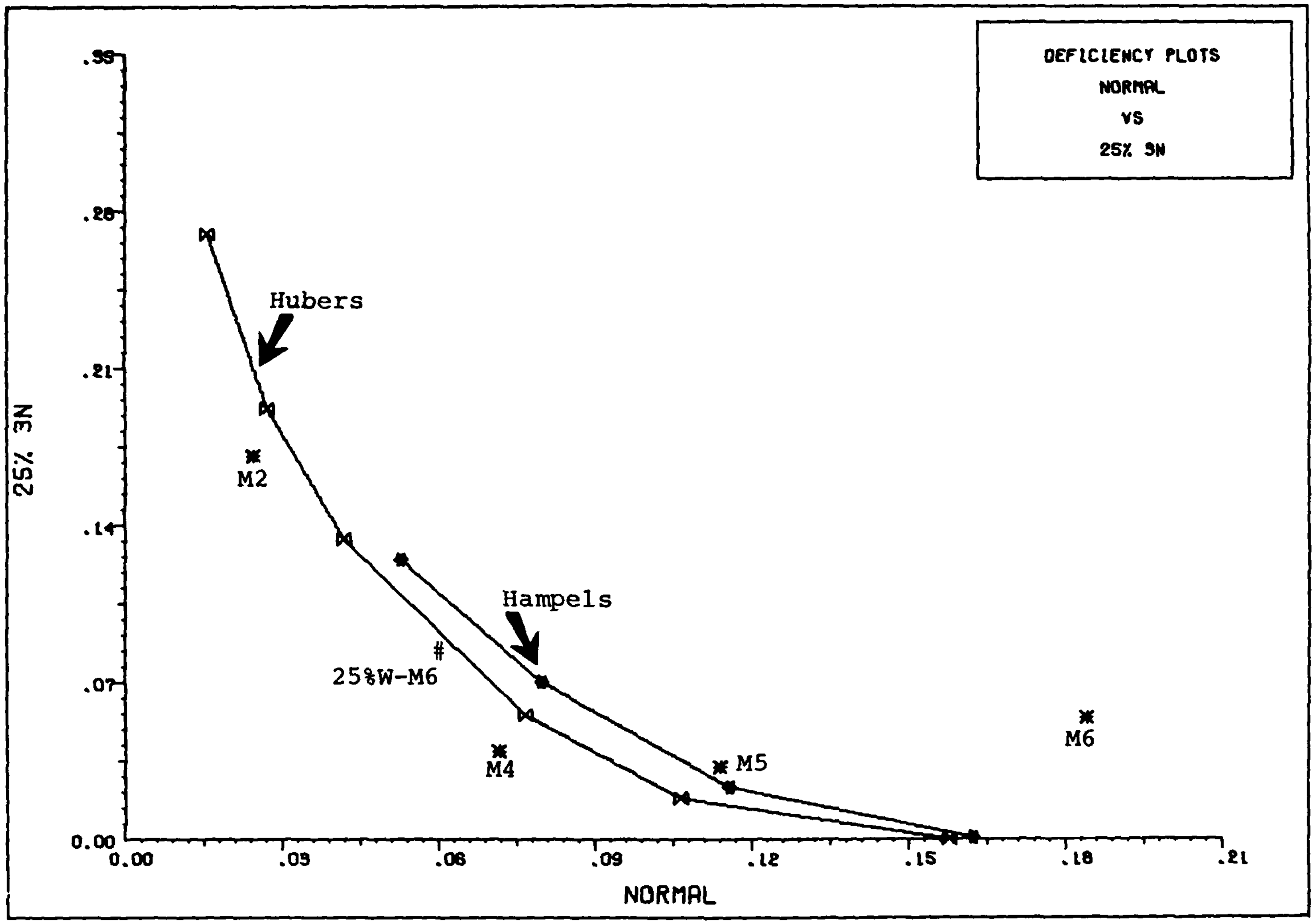

Figure 6.11. Deficiency Plot for Medians, 25%W-M6, Hubers, and Hampels--25% 3N vs Normal

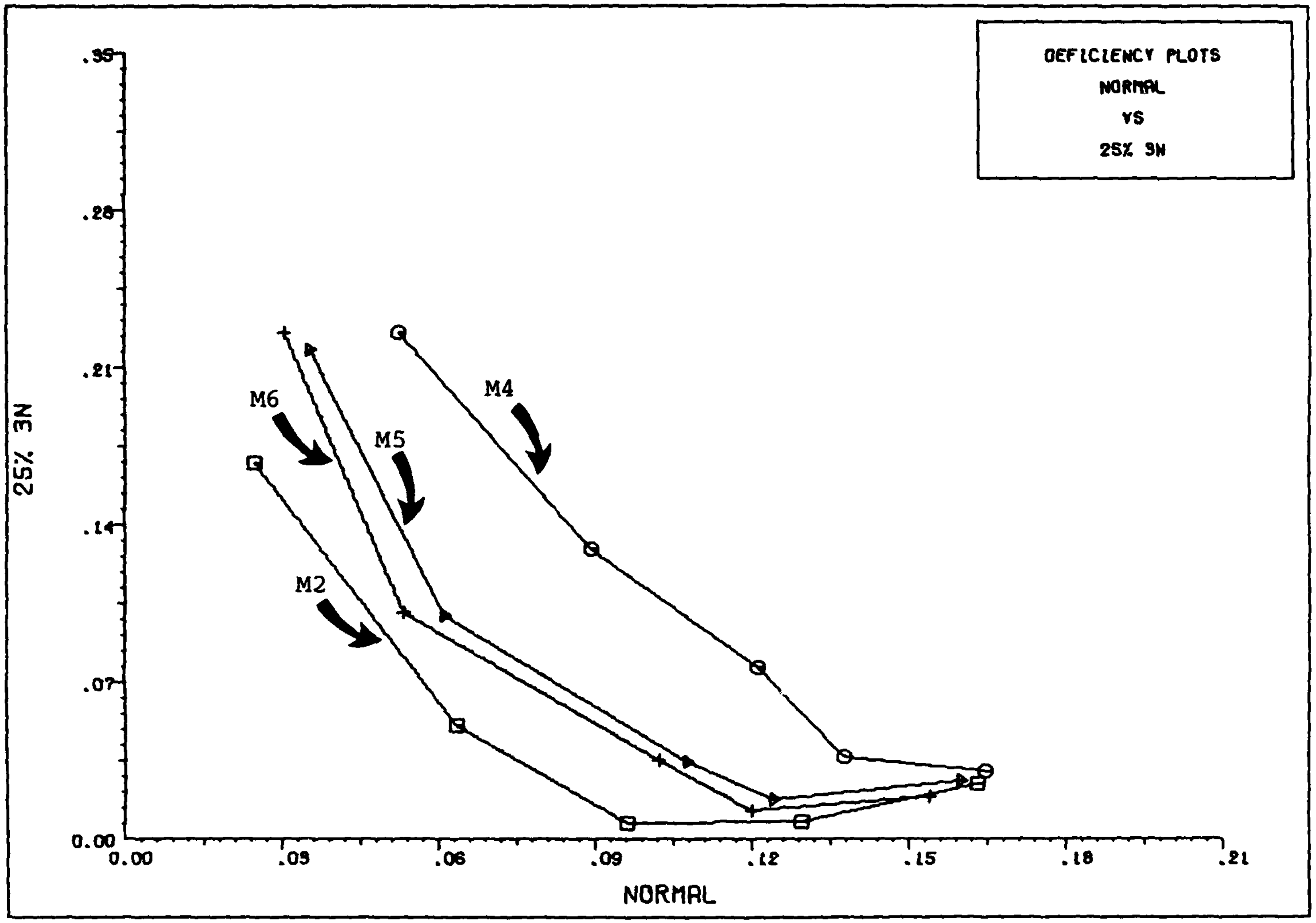

Figure 6.12. Deficiency Plot for Trimmed Means--
25% 3N vs Normal

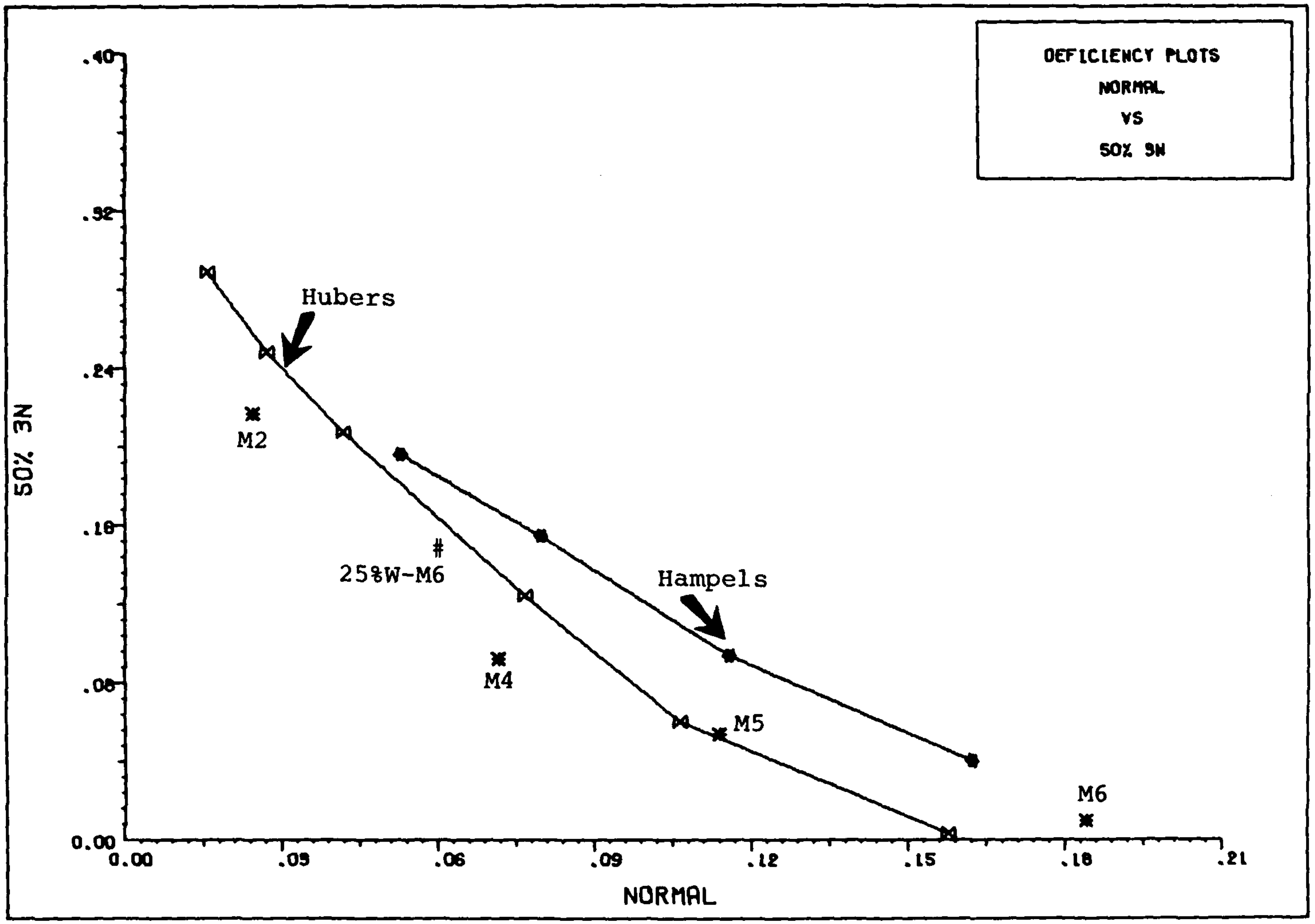

Figure 6.13. Deficiency Plot for Medians, 25%W-M6, Hubers and Hampels--50% 3N vs Normal

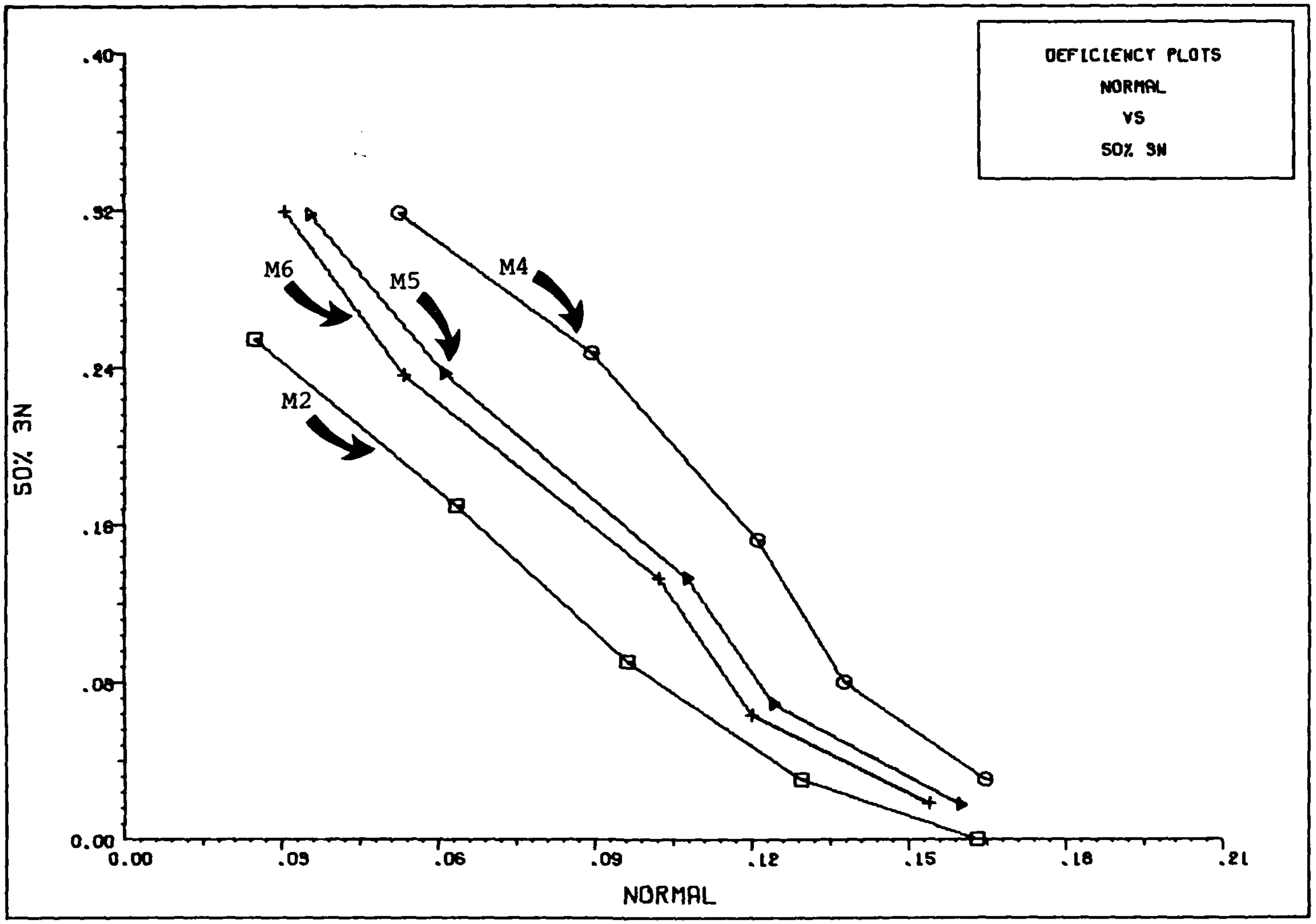

Figure 6.14. Deficiency Plot for Trimmed Means--
50% 3N vs Normal

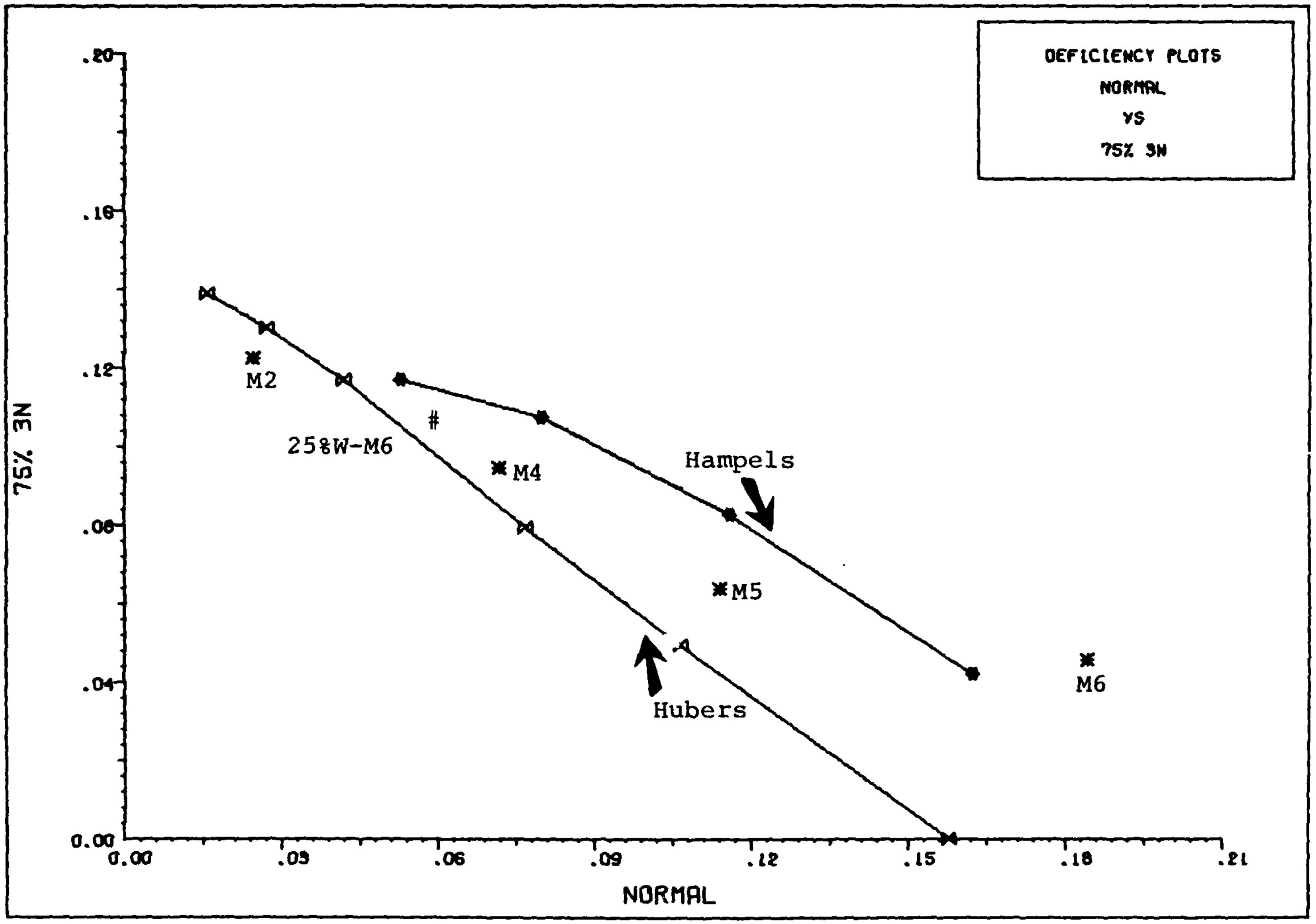

Figure 6.15. Deficiency Plot for Medians, 25%W-M6, Hubers and Hampels--75% 3N vs Normal

AD-A115 491 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOO--ETC F/G 12/1
NONPARAMETRIC ESTIMATION OF DISTRIBUTION AND DENSITY FUNCTIONS --ETC(U)
MAY 82 J SWEEDER
UNCLASSIFIED AFIT/DS/MA/82-1 NL

3 OF 3
AD A
115491

END
DATE
FILMED
07-82
DTIC

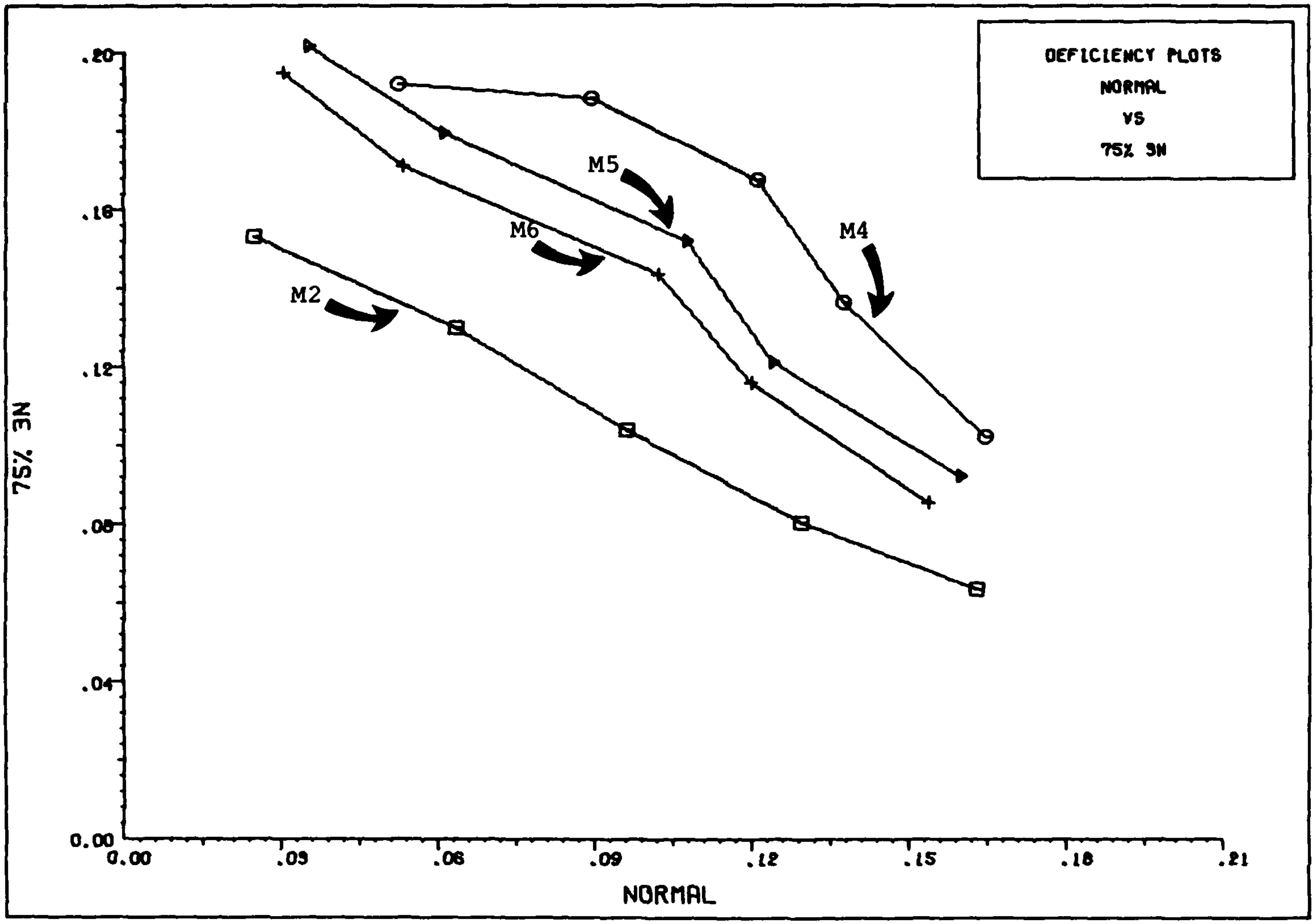

Figure 6.16. Deficiency Plot for Trimmed Means--
75% 3N vs Normal

model number. Since the modified Winsorized means as families and the means of the nonparametric models did not appear to be competitive estimators, we chose not to include their deficiency plots. We also chose to plot only the deficiency comparisons against a normal world. Based on the values in Table VI.1 other deficiency plots could be generated for any pair of alternative distributions.

As a final means of estimator evaluation, we use a tool developed by Hampel--the influence curve. Hampel describes the influence curve as ". . . essentially the first derivative of an estimator, viewed as a functional, at some distribution. . ." (Ref 31). We have chosen to approximate the influence curves for the finite sample case by the use of "stylized sensitivity curves," similar to the ones used in the Princeton study. These stylized sensitivity curves for sample size 20 were generated in the following manner. Let T(x) be a location parameter estimator. Generate a stylized sample from the normal distribution by inverting the standard normal distribution function at the median ranks for a sample size 19. To these 19 stylized order statistics add a 20th point at regular intervals across the real line. We chose 201 such data points at equally spaced intervals on [-3,3]. Calculate the estimator T(x) for each stylized sample of size 20. Plotting n T(x), where n=20, versus x, the added data point, gives us our estimated influence curve.

Figures VI.17 through VI.23 show the stylized sensitivity curves for some of the more competitive estimators determined by the relative efficiency criteria.

Viewing the stylized sensitivity curve as a derivative plot, we can determine how our estimators change with the addition of a new data point. Consider the curve for the median of Model 4 in Figure 6.17. The discontinuity at $x \simeq +2.4$ is due to the adaptive technique employed in the model. At that point, the percentile ratio dictated a model change. The other adaptive models were not similarly effected since the percentile ratios could not be low enough when using a stylized normal sample. Unlike the influence curve for the sample median which becomes constant only a very short distance from zero, the medians based on the nonparametric distribution models change slower as the added data point proceeds away from zero. The sample medium curves for Models 4 and 5 were still monotonically increasing in absolute value as data points were added further away from zero. The changes were very small at the ends of the interval considered, and were, however, decreasing in magnitude. The stylized sensitivity curve for Model 6 became constant for x values outside the interval $[X_{(3)}, X_{(17)}]$ where these order statistics are now based on the stylized sample of size 19. Curves for the modified trimmed means also become constant at some point away from zero, just as curves for simple

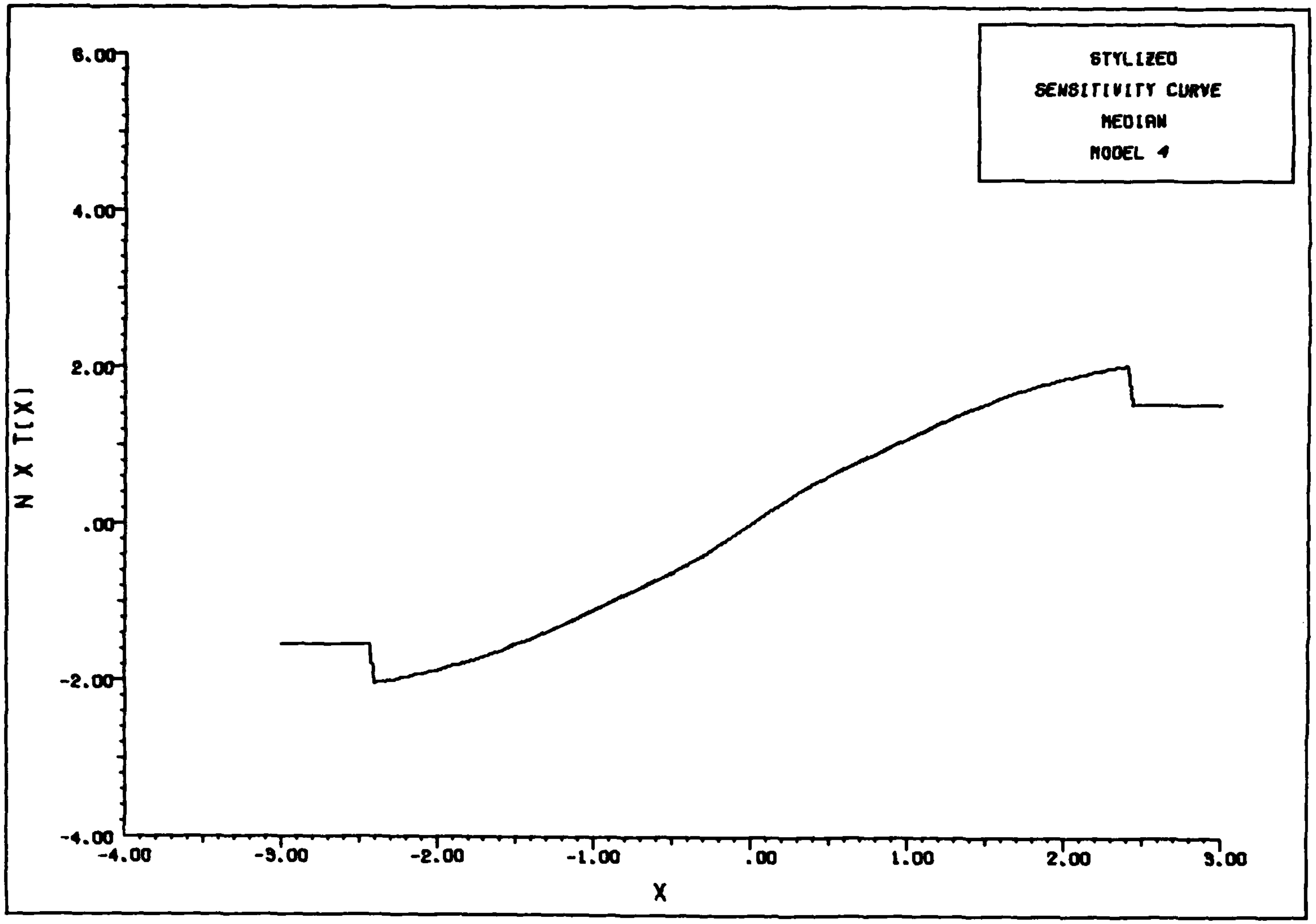

Figure 6.17. Stylized Sensitivity Curve for Median-M4

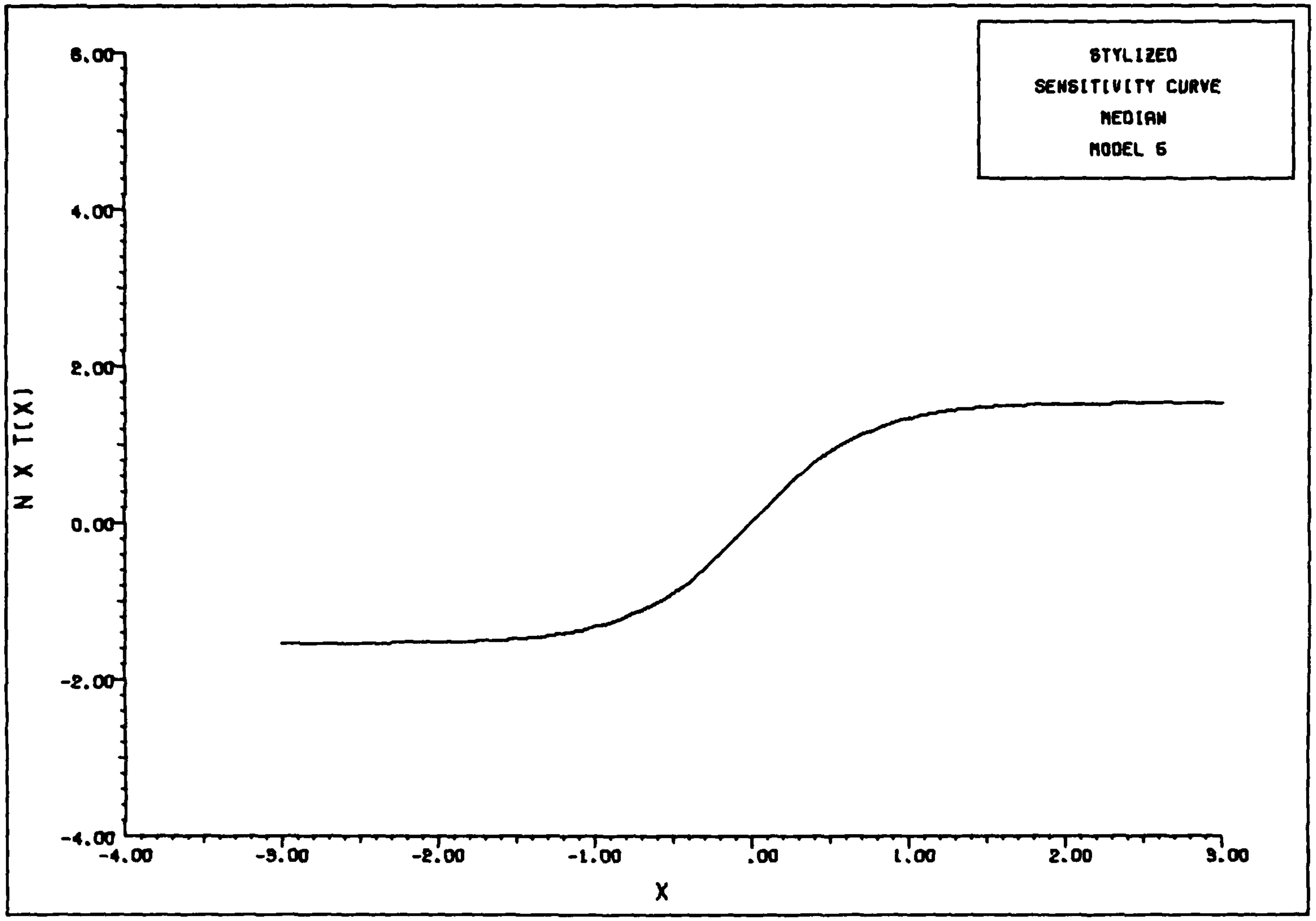

Figure 6.18. Stylized Sensitivity Curve for Median-M5

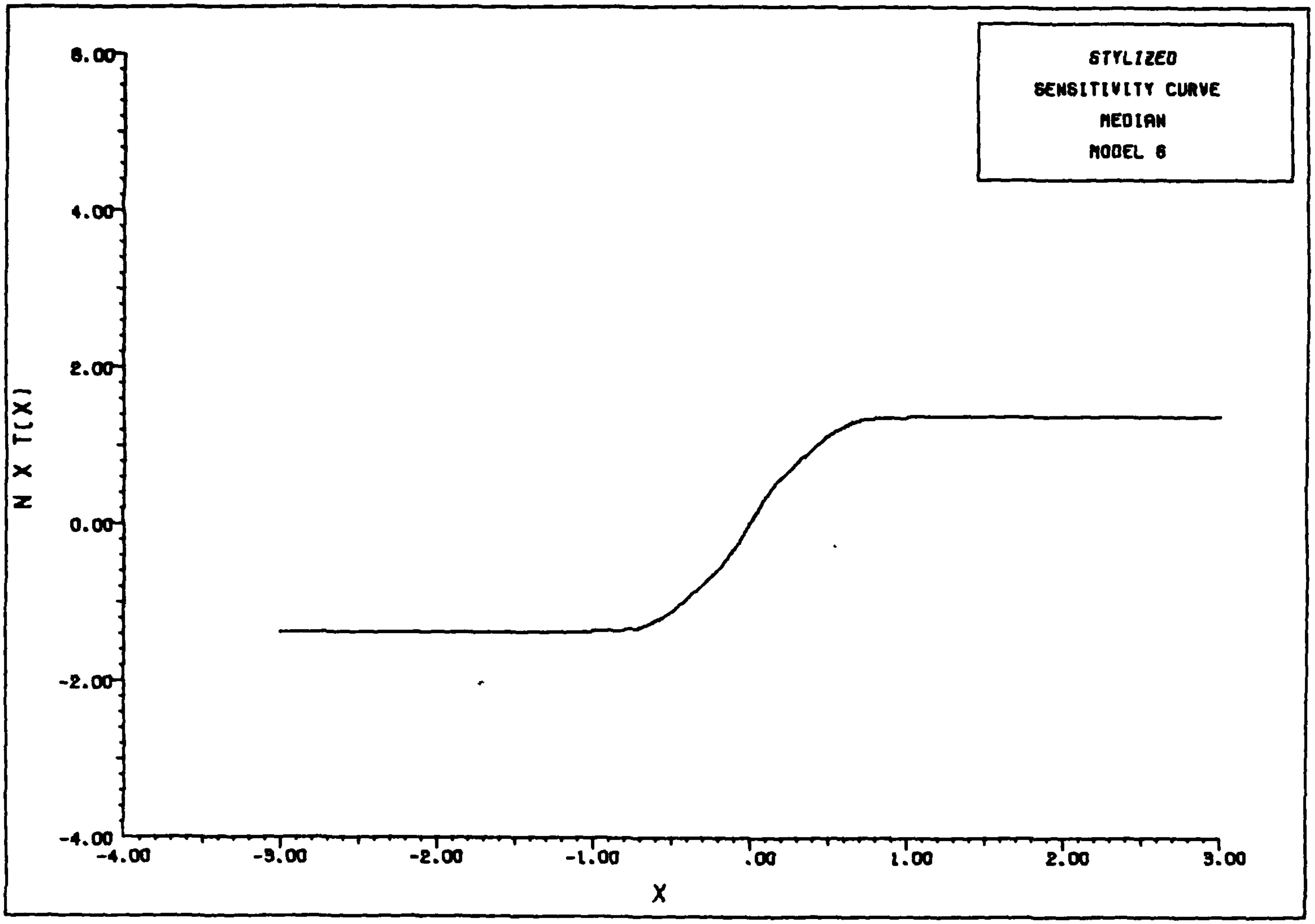

Figure 6.19. Stylized Sensitivity Curve for Median-M6

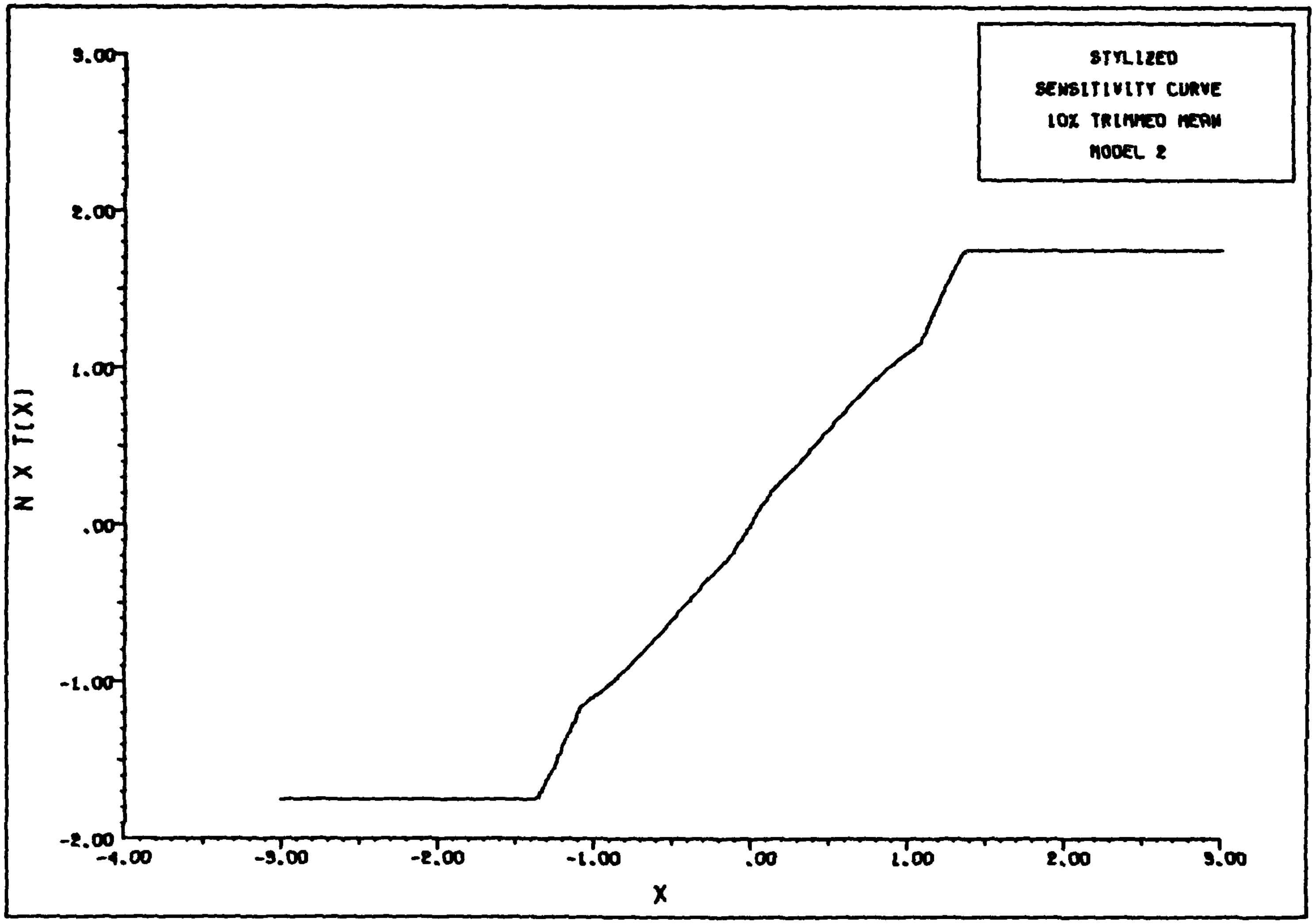

Figure 6.20. Stylized Sensitivity Curve for 10%T-M2

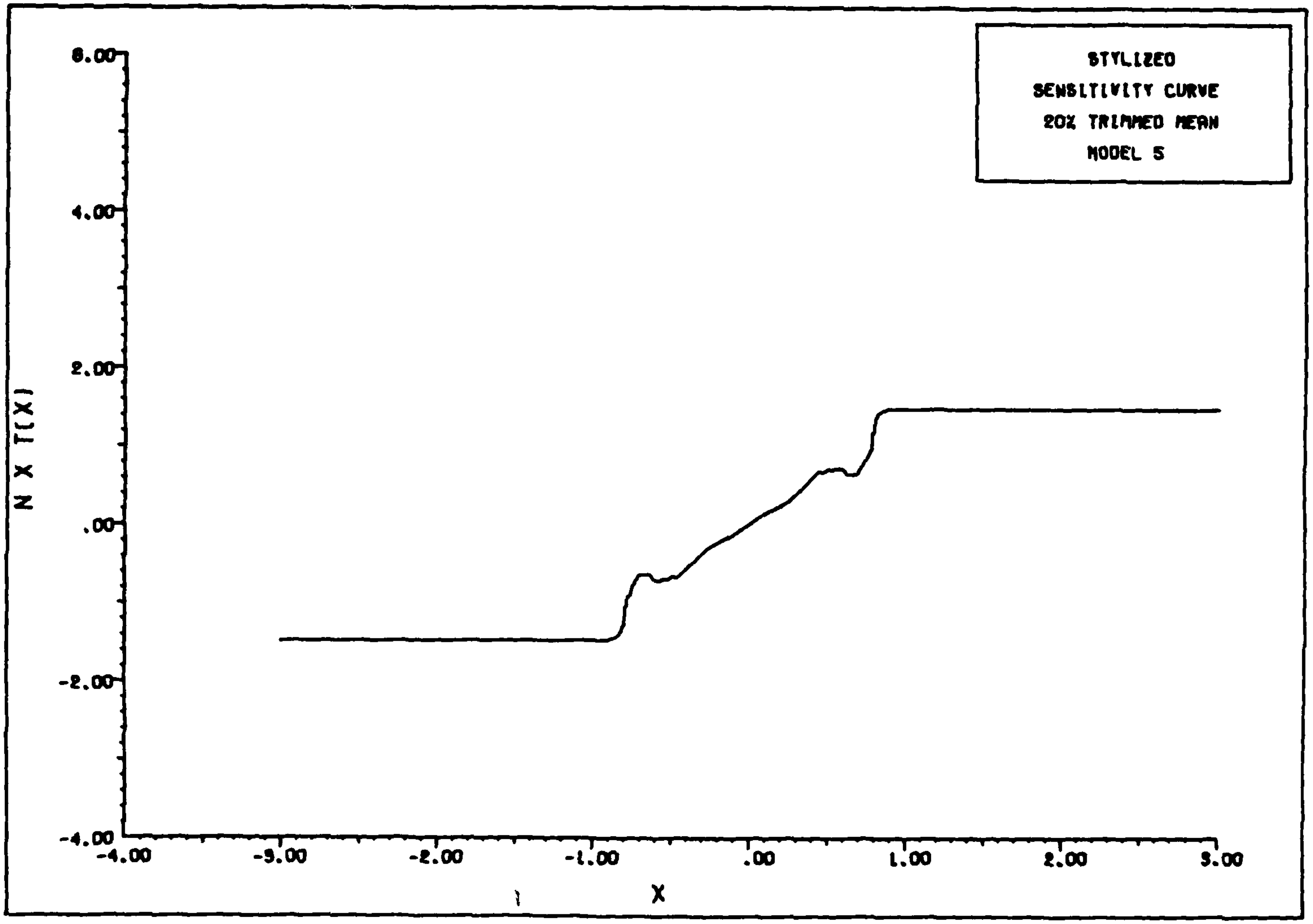

Figure 6.21. Stylized Sensitivity Curve for 20%T-M5

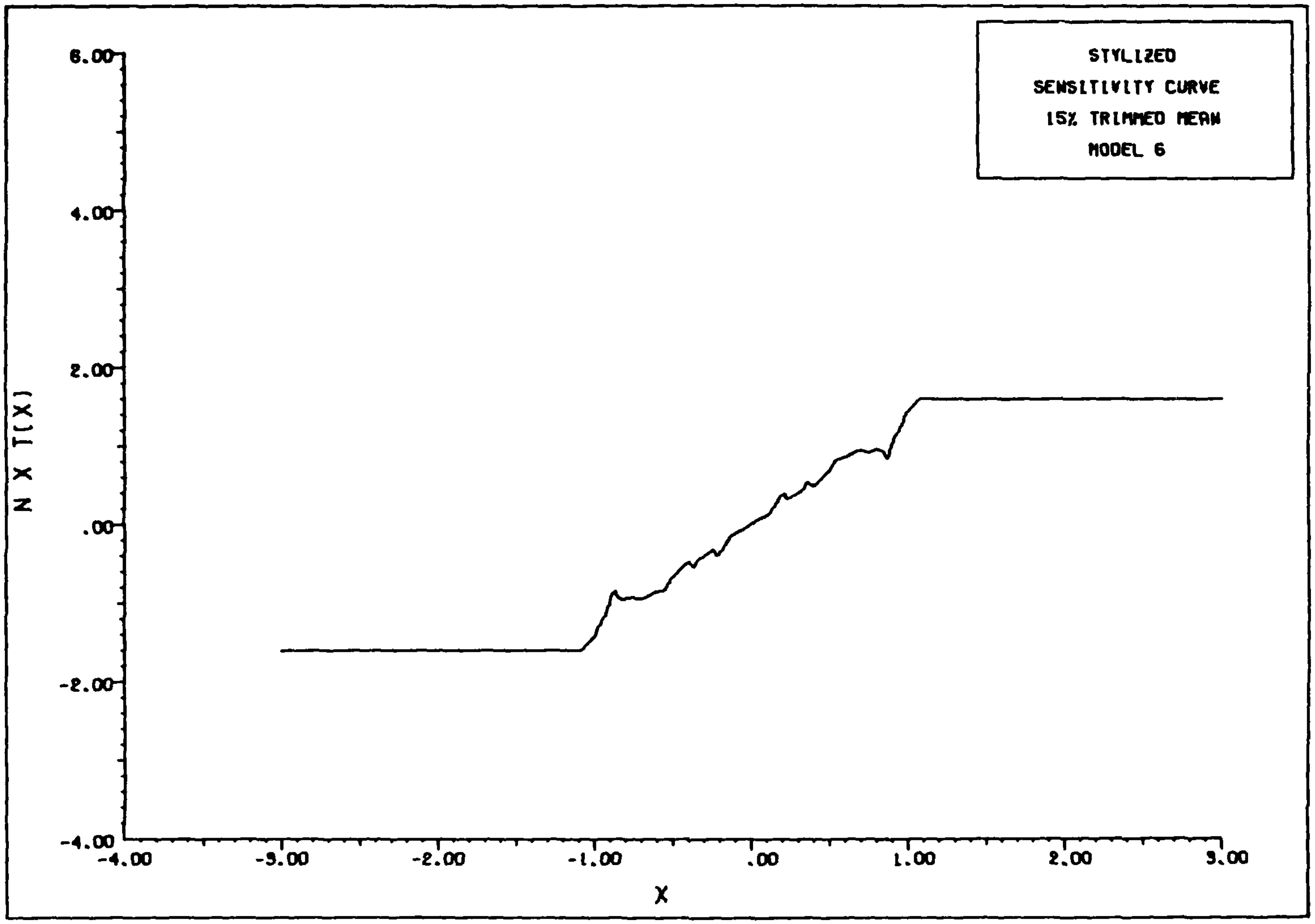

Figure 6.22. Stylized Sensitivity Curve for 15%T-M6

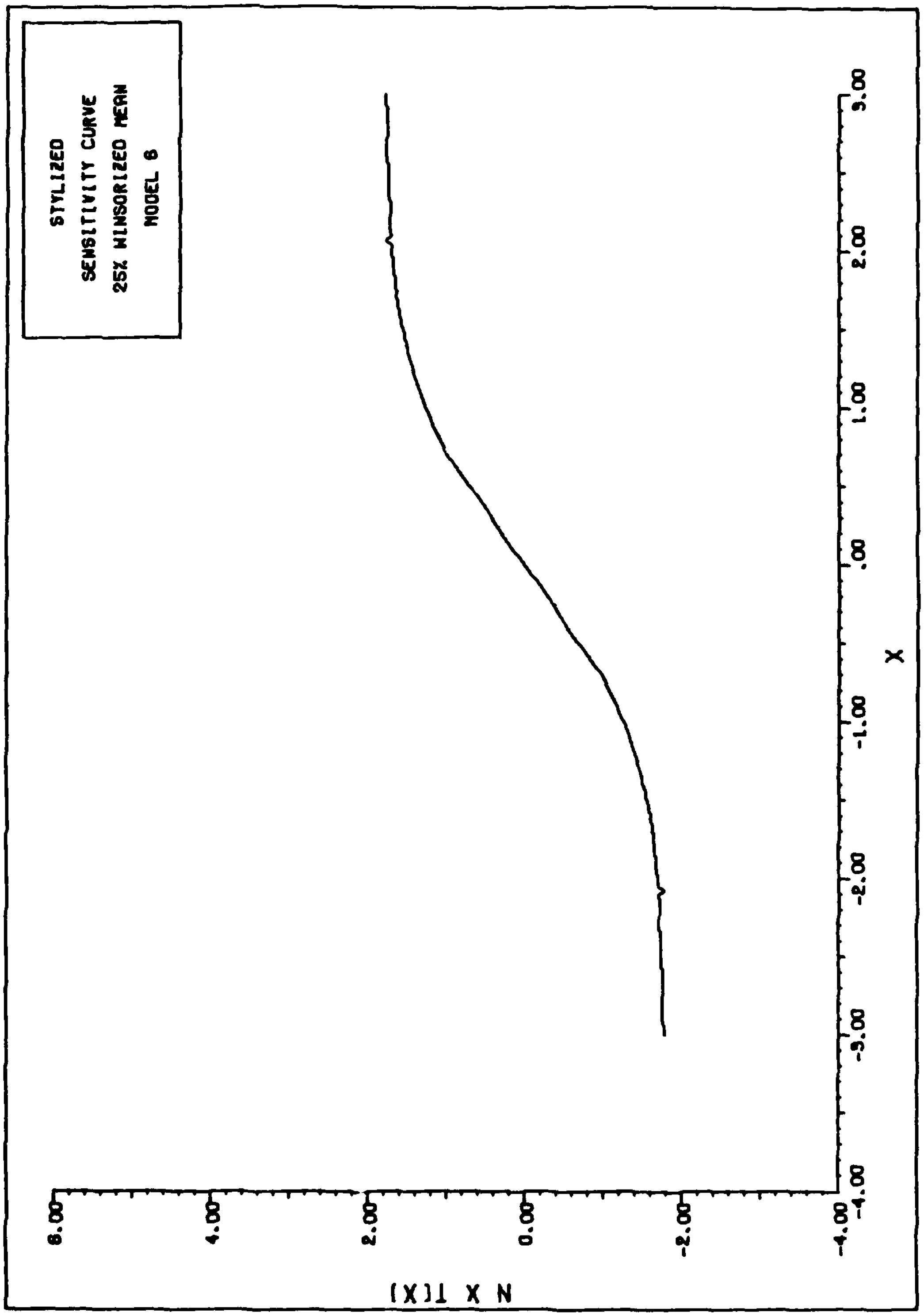

Figure 6.23. Stylized Sensitivity Curve for 25%W-M6

trimmed means do. This constant value of the sensitivity curve indicates that only the sign of the added data point is being noticed by the estimator. The actual value of the additional point could be at any point corresponding to the constant value of the curve. The "influence" on the estimator of two such points is thus identical. If an influence curve goes to zero, the estimator totally rejects the added data point. For our purposes, the value at which the influence curve initially becomes zero is termed the rejection point. Only the Hampels considered in this study have a finite rejection point. No nonparametric estimator proposed completely rejects outliers.

Returning to Figure 6.17, another type of "influence" can be seen. When the adaptive procedure comes into play, it lessens the effect on the estimator. Thus, a data point added to the sample at x=2.8 has a smaller effect on the median using Model 4 than a data point added at x=2.3.

The influence curve also allows for various other measures of robustness. One such measure is gross error sensitivity, the worst influence an outlier can cause. We approximate gross error sensitivity by the absolute value of the supremum of the stylized sensitivity curve. Of the new estimators proposed, the one with the smallest approximate gross error sensitivity was the median for Model 6, with a value of 1.37. When compared with the estimators evaluated by Hampel, only the sample median possesses a smaller gross error sensitivity at the standard normal distribution (Ref 31). For other measures of robustness, such as local shift sensitivity, asymptotic variance, and breakdown points, the reader is referred to Hampel's article.

Summary

This chapter has addressed one specific problem in parametric estimation, namely estimating the location parameter of a symmetric distribution. We began by reviewing some of the literature available concerning robustness aspects of the problem and various proposals for estimators. Besides M, L, R, and D estimators, adaptive techniques were also reviewed. Next we proposed some 48 new estimators based on the new nonparametric models. Model means and medians as well as modified trimmed and modified Winsorized means were defined. These 48 estimators were then evaluated along with the sample mean, sample median and estimators previously proposed by Huber and Hampel. A Monte Carlo analysis generated a standardized empirical variance for each estimator under nine alternative distributions. A relative deficiency comparison was then made over four classes of alternative distributions. Under mild deviations from the normal distribution, new nonparametric estimators possessed smaller average relative deficiency or smaller maximum relative deficiency than the Hubers or Hampels. Estimators and estimator families were further compared via deficiency plots using alternatives to the normal distribution. For some of the better estimators, approximate influence curves were presented. Robustness considerations using these stylized sensitivity curves showed that some of the new estimators are certainly competitive and robust.

VII. Summary, Applications, Limitations and Improvements

Summary

Motivated by the dominance of the empirical distribution function in practically every area of statistical inference, this research effort investigated an alternative to the EDF. After initially examining some other sample distribution functions and related plotting positions, we proposed a new nonparametric family of continuous, differentiable, sample distribution functions. We showed that members of this family possessed the properties of a distribution function and also converged uniformly to the underlying distribution. Six specific members of the family were chosen as models for the rest of the analysis. The new models were evaluated in three distinct areas--their ability to model probability distribution and density functions, their use as bases for goodness of fit tests, and their use in estimating the location parameter of symmetric distributions. We compared the distribution function estimates with the EDF using mean integrated square error as the criterion. A limited Monte Carlo analysis indicated that the new models were superior to the EDF for most of the distributions tested. The derivatives of the nonparametric distribution functions were also evaluated against specifically designed density estimates under the same error criterion. These new nonparametric models were shown to be competitive with or superior to other continuous density estimates. Eight new goodness of fit statistics were generated from the new models. An extensive Monte Carlo analysis confirmed that the new goodness of fit tests for the normal and extreme value distributions had comparable or greater power than the most powerful established tests. Forty-eight new estimators for the center of symmetric of a symmetric population were proposed based on the new models using modified trimmed and Winsorized means. For relatively mild variations of the normal distribution certain new nonparametric estimators were shown to have smaller standardized empirical variances than other robust estimators.

The overall performance of the six models tested has been impressive. Using the relatively simple concept of plotting positions and adding elementary properties of continuity and differentiability, we generated a very powerful tool for data analysis. Several applications of these models in problems of statistical inference are now suggested.

Applications

Given a random sample, our new nonparametric models can be used as representations of the distribution, density, and hazard functions of the underlying process without making any distributional assumption. The continuity of the functions allows for easy graphical depiction. Inferences about the underlying random variable can be made directly.

The new models can also serve as a discriminant for picking a parametric model. Having three continuous functions (distribution, density and hazard functions) to compare against selected parametric alternatives, one could choose a parametric model which had the same general characteristics as the nonparametric estimates. Initially, this could be done by graphical means, but goodness of fit criteria, using various distance measures, could provide a very powerful model discriminant. The modified distance measures of Appendix 1 allow for comparisons using different parametric models over the same finite support and the same probability measure.

Closely related to model discrimination is the problem of parametric estimation. Beginning with an assumed parametric family, parameter estimates are made using a modified distance measure. The parametric family is changed and the process repeated for each alternative family. The selection of the parametric model is then based on the smallest value of the distance criterion. The advantage of this technique is that both model discrimination and parametric estimation are performed simultaneously. A similar approach to the dual problem of model discrimination and parameter estimation was suggested by Borth, who used entropy as a criterion (Ref 9). Another proponent of this approach is Easterling who attacks parameter estimation problems by inverting goodness of fit tests (Ref 22). This is precisely what the above approach does with respect to the modified distance measures.

Another specific example of the use of the new nonparametric models is in the field of reliability. Due to high cost or destructive experiments, the reliability engineer is frequently faced with sparse data sets and the need for a tool of statistical inference. Our new models provide the capability of making reliability estimates from small data sets without the distribution assumptions usually made in reliability analysis. The goodness of fit test results for two widely used models in life testing, the normal and the extreme value, and the ability to estimate the hazard function by a continuous model indicate the applicability of the new nonparametric procedures to reliability problems. The continuity of the sample hazard function also creates the possibility of goodness of fit tests based on some distance measure between hazard functions. Tests using hazard functions have recently been proposed by Kochar (Refs 46, 47). While these tests are for the two sample problem, the new nonparametric models may provide a basis for a one sample test.

The new models also hold promise for use in simulation studies. Typically, Monte Carlo simulation is performed when the distribution of the dependent random variable is unknown. By taking a smaller Monte Carlo sample, the distribution of the dependent variable can be estimated nonparametrically. While no specific results are available to date, the potential benefits of reductions of Monte Carlo sample size warrant investigation. Such a technique could be used in large scale simulations such as cost analysis.

While all of the applications considered thus far dealt with complete random samples, the nonparametric techniques are also capable of modeling other types of data sets. Grouped data is easily handled, providing that the maximum number of data points in one group is at least as small as the number of subsamples used in the model. If not, small offset values can be introduced to insure that no subsample has two identical points. The generation of the nonparametric models from a grouped data set is identical to that of an ungrouped random sample. As such, we can get a continuous distribution function estimate and construct goodness of fit tests for grouped data in exactly the same manner as we constructed the tests in Chapter V.

Limitations

While extremely flexible, the new nonparametric models are subject to certain limitations. In the theoretical development, we arbitrarily set the derivative of the nonparametric distribution function equal to zero at each data point to insure differentiability. A consequence of this step is that $\lim_{x \to X_{min}} sf(x)$ and $\lim_{x \to X_{max}} sf(x)$ exist and are equal to zero. Obviously some density functions do not exhibit these same properties, for example, the uniform, the exponential or a U-shaped beta. All of the nonparametric estimates have density functions whose value is zero at the endpoints of their finite support. The fixed endpoint modifications introduced in the adaptive models attempt to minimize the effect of discontinuities of the underlying density functions. The nonparametric density estimates are continuous over R^1; in general, density functions are not.

Only unimodal densities were examined in the preceding chapters. A limited analysis was done on a bimodal distribution, the double triangular. The results indicated that, while bimodality may be inferred, the density estimate tended to attach unnecessary weight to the interval between the modes. A further analysis is necessary to determine the extent of this limitation.

Finally, the sinusoidal oscillation of the nonparametric estimates may be undesirable to some analysts. While not as smooth as the orthogonal series estimates, the new estimates do possess the distribution function properties lacking in the others. In all of the cases considered in this analysis, the smoothing procedure used tended to prevent radical motions in both the distribution and density functions.

Improvements

In examining our nonparametric models we chose only a representative few members of the family which showed good performance. We also limited ourselves to small sets of initial variables for the estimators. While we attempted to justify all of our choices are reasonable, we examined only a very small set of possible variables. The following are suggested as an initial list of possible improvements to the method. First, other variable sets for plotting positions, inversion points, etc., need to be explored. Their evaluation should still depend on a distance measure criterion, for both the distribution and density functions, perhaps some linear combination of both. Second, alternatives to the percentile ratios need to be considered as discriminants. Third, other functions besides the trigonometric ones need to be evaluated for forming the continuous, differentiable models. Some functions to consider are probability distribution functions, themselves; an analytic function with non-zero derivative at the endpoints which could be pieced together to form the sample distribution function would be ideal. Finally, modification of the technique to model censored samples would be an important contribution in reliability and life testing.

Our investigation of nonparametric, continuous, differentiable, sample distribution functions has covered a large area of statistical inference, from distribution and density estimation, to goodness of fit, to parameter estimation. Our models have shown some significant results, particularly at small sample sizes. Further refinements of techniques based on continuous sample distribution functions can further advance the field of statistical inference.

Bibliography

1. Abramowitz, Milton and Irene A. Stegun (editors). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. New York: Dover Publications, Inc., 1965.

2. Alam, Khurshed. Estimation of a Location Parameter. Technical Report N11. Arlington, Virginia: Office of Naval Research, August, 1971. (AD 736 164).

3. Almquist, Kenneth C. Adaptive Robust Estimation of Population Parameters Using Likelihood Ratio Techniques. MS Thesis, AFIT/GOR/MA/75D-1, Wright-Patterson Air Force Base, Ohio, Air Force Institute of Technology (1975).

4. Anderson, T. W. and D. A. Darling. "Asymptotic Theory of Certain 'Goodness of Fit' Criteria Based on Stochastic Processes," Annals of Mathematical Statistics, 23: 193-212 (1952).

5. Andrews, D. F., et al. Robust Estimates of Location: Survey and Advances. Princeton, New Jersey: Princeton University Press, 1972.

6. Beran, Rudolf. "Minimum Hellinger Distance Estimates for Parametric Models," Annals of Statistics, 5: 455-463 (1977).

7. Blom, Gunnar. Statistical Estimates and Transformed Beta-Variables. Stockholm: Almquist and Wiksells, 1958.

8. Blum, J. and V. Susarla. "A Fourier Inversion Method for the Estimation of a Density and its Derivatives," Journal of the Australian Mathematical Society (Series A), 23: 166-171 (1977).

9. Borth, David M. "A Total Entropy Criterion for the Dual Problem of Model Discrimination and Parameter Estimation," Journal of the Royal Statistical Society Series B--Methodological, 37: 77-87 (1975).

10. Brunk, H. D. "On the Range of the Difference Between Hypothetical Distribution Function and Pyke's Modified Empirical Distribution Function," Annals of Mathematical Statistics, 33: 525-532 (1962).

11. Caso, John. Robust Estimation Techniques for Location Parameter Estimation of Symmetric Distributions. MS Thesis, AFIT/GSA/MA/72-3, Wright-Patterson Air Force Base, Ohio, Air Force Institute of Technology (1972).

12. Chan, Lai K. and Lennart S. Rhodin. "Robust Estimation of Location Using Optimally Chosen Sample Quantiles," Technometrics, 22: 225-237 (1980).

13. Chung, Kai Lai. A Course in Probability Theory. New York: Academic Press, 1974.

14. Crain, Beadford R. "An Information Theoretic Approach to Approximating a Probability Distribution," SIAM Journal of Applied Mathematics, 32: 339-346 (March 1977).

15. Cressie, Noel. "Transformations and the Jackknife," Journal of the Royal Statistical Society Series B--Methodological, 43: 177-182 (1981).

16. Crowder, George E., Jr. Adaptive Estimation Based on a Family of Generalized Exponential Power Distributions. MS Thesis, AFIT/GOR/MA/77D-2, Wright-Patterson Air Force Base, Ohio, Air Force Institute of Technology (1977).

17. Daniels, Tony G. Robust Estimation of the Generalized t Distribution Using Minimum Distance Estimation. MS Thesis, AFIT/GOR/MA/80D-2, Wright-Patterson Air Force Base, Ohio, Air Force Institute of Technology (December 1980).

18. David, F. N. and N. L. Johnson. "The Probability Integral Transform when Parameters are Estimated from the Sample," Biometrika, 35: 182-190 (1948).

19. David, H. A. Order Statistics. New York: Wiley, 1970.

20. Dudewicz, Edward J. and Edward C. van der Mevlen. Entropy-Based Statistical Inference, I: Testing Hypotheses on Continuous Probability Densities, with Special Reference to Uniformity. Report No. 120. Leuven, Belgium: Department of Mathematics, Katholieke Universiteit Leuven, June 1979.

21. Durbin, J. "Kolmogorov-Smirnov Tests When Parameters are Estimated with Applications to Tests of Exponentiality and Tests on Spacings," Biometrika, 62: 5-22 (1975).

22. Easterling, Robert G. "Goodness of Fit and Parameter Estimation," Technometrics, 18: 1-9 (February 1976).

23. Efron, B. "Bootstrap Methods: Another Look at the Jackknife," Annals of Statistics, 7: 1-26 (1979).

24. Forth, Charles R. Robust Estimation Techniques for Population Parameters and Regression Coefficients. MS Thesis, AFIT/GSA/MA/74-1, Wright-Patterson Air Force Base, Ohio, Air Force Institute of Technology (1974).

25. Foutz, Robert V. "A Test for Goodness-of-Fit Based on an Empirical Probability Measure," Annals of Statistics, 8: 989-1001 (1980).

26. Gastwirth, J. "On Robust Procedures," JASA, 61: 929-948 (1966).

27. Gibbons, Jean D. Nonparametric Statistical Inference. New York: McGraw-Hill, 1971.

28. Gray, H. L., W. R. Schucany, and T. A. Watkins. "On the Generalized Jackknife and its Relation to Statistical Differentials," Biometrika, 62: 637-642 (1975).

29. Green, J. R. and Y. A. S. Hegazy. "Powerful Modified EDF Goodness-of-Fit Tests," JASA, 71: 204-209 (March 1976).

30. Hampel, Frank R. "A General Qualitative Definition of Robustness," Annals of Mathematical Statistics, 42: 1887-1896 (1971).

31. Hampel, Frank R. "The Influence Curve and its Role in Robust Estimation," JASA, 69: 383-393 (June 1974).

32. Harp, Tilford. Fully Adaptive Estimation of the Parameters of a t and Half t Distribution. MS Thesis, AFIT/GOR/MA/79-1, Wright-Patterson Air Force Base, Ohio, Air Force Institute of Technology (1979).

33. Harter, H. Leon. "The Use of Order Statistics in Estimation," Operations Research, 16: 783-798 (1968).

34. Harter, H. Leon, Albert H. Moore and Thomas F. Curry. "Adaptive Robust Estimation of Location and Scale Parameters of Symmetric Populations," Communications in Statistics--Theory and Methods, A8: 1473-1491 (1979).

35. Hazen, Allen. Flood Flows. New York: Wiley, 1930.

36. Hill, D. L. and P. V. Rao. "Tests of Symmetry Based on Cramer von Mises Statistics," Biometrika, 64: 489-494 (1977).

37. Hodges, J. L. and E. L. Lehmann. "Estimates of Location Based on Rank Tests," Annals of Mathematical Statistics, 34: 598-611 (1963).

38. Hogg, Robert V. "Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future Applications and Theory," JASA, 69: 909-927 (December 1974).

39. Hogg, Robert V. "An Introduction to Robust Estimation," in Robustness in Statistics, Robert L. Launer and Graham N. Wilkinson (editors). New York: Academic Press, 1979.

40. Holcomb, R. L., R. A. Kronmal, and M. E. Tartar. "A Description of New Computer Methods for Estimating the Population Density," Proceedings from the Association of Computing Machinery, 22. New York: Thompson Book Co., 511-519 (1967).

41. Huber, Peter J. "Robust Estimation of a Location Parameter," Annals of Mathematical Statistics, 35: 73-101 (1964).

42. Huber, Peter J. "The 1972 Wald Lecture: Robust Statistics: A Review," Annals of Mathematical Statistics, 43: 1041-1067 (1972).

43. Jorgenson, Loren W. Robust Estimation of Location and Scale Parameters. MS Thesis, AFIT/GSA/MA/73-2, Wright-Patterson Air Force Base, Ohio, Air Force Institute of Technology (1973).

44. Kapur, R. C. and L. R. Lamberson. Reliability in Engineering Design. New York: Wiley, 1977.

45. Kimball, B. F. "On the Choice of Plotting Positions on Probability Paper," JASA, 55: 546-560 (September 1960).

46. Kochar, Subhash C. "Distribution Free Comparison of Two Probability Distributions with Reference to Their Hazard Rates," Biometrika, 66: 437-442 (1979).

47. Kochar, Subhash C. "A New Distribution-Free Test for the Equality of Two Failure Rates," Biometrika, 68: 423-426 (1981).

48. Kronmal, R. A. and M. E. Tartar. "The Estimation of Probability Densities and Cumulatives by Fourier Series Methods," JASA, 63: 925-952 (1968).

49. Lamperti, John. Probability. New York: W. A. Benjamin, Inc., 1966.

50. Lilliefors, Hubert W. "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown," JASA, 399-402 (1967).

51. Littell, R. C., J. T. McClave, and W. W. Offen. "Goodness of Fit Tests for the Two Parameter Weibull Distribution," Communications in Statistics--Simula. Computa., B8: 257-269 (1979).

52. MacQueen, James and Jacob Marschak. "Partial Knowledge, Entropy and Estimation," Proceedings of the National Academy of Sciences, 3819-3824 (October 1975).

53. Mann, N. R., E. M. Scheuer, and K. W. Fertig. "A New Goodness of Fit Test for the Two-Parameter Weibull or Extreme Value Distribution with Unknown Parameters," Communications in Statistics, 2: 383-400 (1973).

54. McGrath, E. J. and D. C. Irving. Techniques for Efficient Monte Carlo Simulation. Volume II. Random Number Generation for Selected Probability Distributions. SAI Report SAI-72-590-LJ, Arlington, Virginia: Office of Naval Research, March 1973 (AD 762 722).

55. McNeese, Larry B. Adaptive Minimum Distance Estimation Techniques Based on a Family of Generalized Exponential Power Distributions. MS Thesis, AFIT/GOR/MA/80D, Wright-Patterson Air Force Base, Ohio, Air Force Institute of Technology (December 1980).

56. Mihalko, Daniel P. and David S. Moore. "Chi-Square Tests of Fit for Type II Censored Data," Annals of Statistics, 8: 625-644 (May 1980).

57. Miller, James E., Jr. Continuous Density Approximation on a Bounded Interval Using Information Theoretic Concepts. Ph.D. dissertation, AFIT/DS/MA/80-1, Wright-Patterson Air Force Base, Ohio, Air Force Institute of Technology, 1980.

58. Miller, Rupert G. "The Jackknife--a Review," Biometrika, 61: 1-15 (1974).

59. Moore, Albert H. "Extension of Monte Carlo Techniques for Obtaining System Reliability Confidence Limits from Component Test Data," Proceedings of National Aerospace Electronics Conference, 459-463 (May 1965).

60. Moore, Albert H. Robust Statistical Inference. Notes from a short course presented at the Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio, February 1981.

61. Parr, William C. Minimum Distance and Robust Estimation. Ph.D. dissertation, Dallas, Texas, Southern Methodist University, 1978.

62. Parr, William C. "Minimum Distance Estimation: A Bibliography," Unpublished Manuscript. Institute of Statistics, Texas A&M University, College Station, Texas, 1980.

63. Parr, William C. and William R. Schucany. "Minimum Distance and Robust Estimation," Dallas, Texas, Southern Methodist University, Department of Statistics, 1979 [to appear in JASA].

64. Parr, William C. and T. Dewet. "On Minimum CVM-Norm Parameter Estimation," Unpublished Manuscript. Institute of Statistics, Texas A&M University, College Station, Texas, and Department of Mathematical Statistics, Rhodes University, Grahamstown, South Africa, 1979.

65. Parzen, Emanuel. "Nonparametric Statistical Data Modeling," JASA, 74: 105-121 (March 1979).

66. Pennington, Ralph H. Introductory Computer Methods and Numerical Analysis. New York: Macmillan, 1965.

67. Phadia, Eswarlal G. On Estimation of a Cumulative Distribution Function. Ph.D. dissertation, Columbus, Ohio, Ohio State University, 1971.

68. Pyke, Ronald. "The Supremum and Infinum of the Poisson Process," Annals of Mathematical Statistics, 30: 568-576 (1959).

69. Pyke, Ronald. "Spacings," Journal of the Royal Statistical Society, Series B, 27: 395-436 (1965).

70. Quenouille, M. H. "Approximate Tests of Correlation in Time-Series," Journal of the Royal Statistical Society, Series B--Methodological, 11: 68-84 (1949).

71. Quenouille, M. H. "Notes on Bias in Estimation," Biometrika, 43: 353-360 (1956).

72. Ramberg, John S., Edward J. Dudewicz, Ranou R. Tadikamalla, and Edward F. Mykytka. "A Probability Distribution and its Uses in Fitting Data," Technometrics, 21: 201-214 (May 1979).

73. Rao, C. Radhakrishna. Linear Statistical Inference and Its Applications. New York: Wiley, 1965.

74. Reiss, R. D. "On Minimum Distance Estimators for Unimodal Densities," Metrika, 23: 7-14 (1976).

75. Rosenblatt, Murray. "Remarks on Some Nonparametric Estimates of a Density Function," Annals of Mathematical Statistics, 27: 832-837 (1956).

76. Rothman, E. D. and M. Woodroofe. "A Cramer-von Mises Type Statistic for Testing Symmetry," Annals of Mathematical Statistics, 43: 2035-2038 (1972).

77. Rugg, Bernard J. Adaptive Robust Estimation of Location and Scale Parameters Using Selected Discriminants. MS Thesis, AFIT/GOR/MA/74D-3, Wright-Patterson Air Force Base, Ohio, Air Force Institute of Technology (1974).

78. Sahler, W. "A Survey on Distribution-Free Statistics Based on Distances Between Distribution Functions," Metrika, 13: 149-169 (1968).

79. Sahler, W. "Estimation by Minimum Discrepancy Methods," Metrika, 15: 85-106 (1970).

80. Saniga, Erwin M. and James A. Miles. "Power of Some Standard Goodness-of-Fit Tests of Normality Against Asymmetric Stable Alternatives," JASA, 74: 861-865 (December 1979).

81. Schuster, Eugene F. "Estimation of a Probability Density Function and Its Derivatives," Annals of Mathematical Statistics, 40: 1187-1195 (1969).

82. Schuster, E. F. "On the Goodness-of-Fit Problem for Continuous Symmetric Distributions," JASA, 68: 713-715 (1973). Corrigenda JASA, 69: 288 (1974).

83. Schuster, Eugene F. "Estimating the Distribution Function of a Symmetric Distribution," Biometrika, 62: 631-636 (1975).

84. Schwartz, Stewart. "Estimation of a Probability Density by an Orthogonal Series," Annals of Mathematical Statistics, 38: 1261-1265 (1967).

85. Singh, R. S. "Mean Square Errors of Estimates of a Density and its Derivatives," Biometrika, 66: 177-180 (1979).

86. Smaga, Edward. "Smooth Empirical Distribution Function," Przeglad Statystyczny, 25.1: Warsaw, Poland (1978).

87. Smith, R. M. and L. J. Bain. "Correlation Type Goodness-of-Fit Statistics with Censored Samples," Communications in Statistics--Theory and Methods, A5: 119-132 (1976).

88. Stephens, M. A. "Use of Kolmogorov-Smirnov, Cramer von Mises and Relaxed Statistics Without Extensive Tables," Journal of the Royal Statistical Society, Series B, 32, No. 1: 115-122 (1970).

89. Stephens, M. A. "EDF Statistics for Goodness of Fit and Some Comparisons," JASA, 69: 730-737 (September 1974).

90. Stephens, M. A. "Goodness of Fit for the Extreme Value Distribution," Biometrika, 64: 583-588 (1977).

91. Stigler, Stephen M. Simon Newcomb, Percy Daniell, and the History of Robust Estimation, 1385-1920. Technical Report No. 319. Arlington, Virginia: Office of Naval Research, December 1972 (AD 757 026).

92. Stigler, Stephen M. "Do Robust Statistics Work with Real Data?" (With Discussants), Annals of Statistics, 5: 1055-1098 (1977).

93. Stigler, Stephen. "Studies in the History of Probability and Statistics XXXVIII--R. H. Smith, a Victorian Interest in Robustness," Biometrika, 67: 217-221 (1980).

94. Tapia, Richard A. and James R. Thompson. Nonparametric Probability Density Estimation. Baltimore: Johns Hopkins University Press, 1978.

95. Tribus, Myron. Rational Descriptions, Decisions, and Designs. New York: Pergamon Press, 1969.

96. Tukey, J. W. "Bias and Confidence in Not-Quite Large Samples," Annals of Mathematical Statistics, 29: 614 (1958).

97. Turnbull, Bruce W. "The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data," Journal of the Royal Statistical Society Series B--Methodological 38: 25 (1976).

98. Vogt, Herbert. "Concerning a Variant of the Empirical Distribution Function," Metrika, 25: 49-58 (1978).

99. Wahba, Grace. "Optimal Convergence Properties of Variable Knot, Kernel, and Orthogonal Series Methods for Density Estimation," Annals of Statistics, 3: 15-29 (1975).

100. Walter, Gilbert G. "Properties of Hermite Series Estimation of Probability Density," Annals of Statistics, 5: 1258-1264 (1977).

101. Walter, G. and J. Blum. "Probability Density Estimation Using Delta Sequences," Annals of Statistics, 7: 328-340 (1979).

102. Watson, G. S. "Density Estimation by Orthogonal Series," Annals of Mathematical Statistics, 40: 1496-1498 (1969).

103. Watson, G. S. and M. R. Leadbetter. "On the Estimation of the Probability Density I," Annals of Mathematical Statistics, 34: 480-491 (1963).

104. Wegman, Edward J. "Nonparametric Probability Density Estimation: I. A Summary of Available Methods," Technometrics, 14: 533-546 (1972).

105. Wegman, Edward J. "Nonparametric Probability Density Estimation: II. A Comparison of Density Estimation Methods," Journal of Statistical Computations and Simulation, 1: 225-245 (1972).

106. Wegman, Edward J. and H. I. Davies. "Remarks on Some Recursive Estimators of a Probability Density," Annals of Statistics, 7: 316-327 (1979).

107. White, John S. "The Moments of Log-Weibull Order Statistics," Technometrics, 11: 373-386 (1969).

108. Wolfowitz, J. "The Minimum Distance Method," Annals of Mathematical Statistics, 28: 75-88 (1957).

109. Wright, Ian W. Spline Methods in Statistics. Technical Report No. 77-1307. Bolling Air Force Base, D.C., Air Force Office of Scientific Research, 1977 (AD A049 197).

Appendix 1

Modified Distance Measures

A classical distance measure with respect to an integral criterion is given by:

$$\delta(F,G) = \int_{-\infty}^{\infty} (F(x)-G(x))^2 \ \psi(F(x)) \ dF(x)$$

where $\psi(F(x))$ is some preassigned weight function (Ref 78). For the Cramer von Mises distance, $G(x)$ is the empirical distribution function, $S_n(x)$, $\psi(F(x))=1$, and $F(x)$ is the postulated underlying model. Thus $\delta(F,S_n)$ is a CVM distance measure.

Given a measure, μF_n whose corresponding probability distribution function F_n is measurable, we can now consider an alternative distance measure, $\delta(F_n,F)$. Since $SF(x)$, as defined in equation 3.6, is continuous and differentiable, we can define:

$$\delta(SF,F) = \int_{X_{min}}^{X_{max}} (SF(x)-F(x))^2 \ \psi(SF(x)) \ dSF(x)$$

In the classical case, for $\psi(F(x))=1$, $\delta(F,G)$ is the integrated square error with a weight of f induced by the $dF(x)$ term. Using S_n as an approximation to F so that $dS_n(x)$ approximates $f(x)\,dx$ results in $\delta(F,G) \simeq$

$$\int_{-\infty}^{\infty} (F(x)-G(x))^2 \, dS_n(x),$$

which is the average square error between the distribution functions F and G (Ref 105). Since F is approximated by SF, we can also approximate the integrated square error $\delta(F,SF)$ by $\delta(SF,F)$, where $\psi(SF(x))=1$.

The following are some classical and modified distance measures used in the analysis where F is the underlying distribution function and SF is the continuous differentiable sample distribution function. Each distance measure is listed only with respect to closeness of the distribution functions F and SF. Substitution of f and sf for F and SF respectively in only the absolute value or squared terms gives the corresponding distance measure for the density functions. Note that the argument of both the weight function ψ and differentiation operator D is still the distribution function, not the density function.

1. Kolmogorov-Smirnov (KS) distance

$$\delta(F,SF) = \sup_{-\infty<x<\infty} |F(x) - SF(x)|$$

approximated by $\max_i |F(X_i)-SF(X_i)|$

2. KS integral distance

$$\delta(F,SF) = \int_{-\infty}^{\infty} |F(x)-SF(x)| \, dF(x)$$

3. Modified KS integral distance

$$\delta(SF,F) = \int_{-\infty}^{\infty} |SF(x)-F(x)| dSF(x)$$

4. Cramer von Mises (CVM) integral distance

$$\delta(F,SF) = \int_{-\infty}^{\infty} (F(x)-SF(x))^2 \; dF(x)$$

5. Modified CVM integral distance

$$\delta(SF,F) = \int_{-\infty}^{\infty} (SF(x)-F(x))^2 \; dSF(x)$$

6. Anderson Darling (AD) integral distance

$$\delta(F,SF) = \int_{-\infty}^{\infty} (F(x)-SF(x))^2/[F(x)(1-F(x)] \; dF(x)$$

7. Modified AD integral distance

$$\delta(SF,F) = \int_{-\infty}^{\infty} (SF(x)-F(x))^2/[(SF(x)(1-SF(x))] \; dSF(x)$$

8. Average square error

$$ASE = \frac{1}{n} \sum_{i=1}^{n} (F(X_i)-SF(X_i))^2$$

Appendix 2

Generalized Exponential Power (GEP) Distribution

The Generalized Exponential Power distribution is a three parameter family of symmetric distributions whose tail length ranges from extremely platykurtic to extremely leptokurtic (Ref 60). While, in general, the distribution function does not exist in closed form, the density function depends on μ, σ, and p, location, scale, and shape parameters respectively.

$$f(x;\mu,\sigma,p) = \frac{pg(p)}{2\Gamma(1/p)\sigma} \exp\left\{- \left[\frac{g(p)|x-\mu|}{\sigma}\right]^p\right\}$$

where

$$g(p) = \left[\frac{\Gamma(3/p)}{\Gamma(1/p)}\right]^{\frac{1}{2}}$$

and $-\infty<x<\infty$, $-\infty<\mu<\infty$, $0<\sigma<\infty$, $1\leq p\leq\infty$

For this distribution, $E(X)=\mu$ and $Var(X)=\sigma^2$.

Three special cases occur for specific choices of the shape parameter p:

1. p=1 reduces the GEP distribution to the Laplace or double exponential distribution.

2. p=2 reduces the GEP distribution to the normal distribution.

3. As $p \to \infty$, the GEP distribution approaches the uniform distribution. Although $p \to \infty$ is a limiting case, we include the uniform distribution to complete the family. To avoid the limit argument in discussions, we will consider $p=\infty$ to represent the uniform distribution.

Appendix 3

Critical Values

Tables A3.1 through A3.10 list the critical values of the eight new test statistics--D5, D6, DMR, W5, W6, WMR, A5, and A6. Two null hypothesis situations are considered: (1) the null distribution completely specified, and (2) the null distribution parameters estimated. For the normal distribution, the parameters were estimated using the uniformly minimum variance unbiased estimates $\overline{X}$ and S. For the extreme value distribution, the parameters were estimated using the maximum likelihood method. A Newton Raphson iteration scheme was employed. Critical values for the normal distribution are listed in Tables A3.1 through A3.5. Critical values for the extreme value distribution are listed in Tables A3.6 through A3.10. Values are given for sample sizes 10(10)50 and alpha levels .20, .15, .10, .05, .025, and .01.

TABLE A3.1

CRITICAL VALUES--NORMAL DISTRIBUTION--
SAMPLE SIZE 10

Null Distribution Completely Specified

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.2249	.2436	.2739	.3147	.3503	.3914
D6	.2238	.2439	.2712	.3108	.3487	.3903
DMR	.2656	.2846	.3114	.3509	.3853	.4192
W5	.2236	.2667	.3429	.4114	.5578	.7164
W6	.2090	.2549	.3178	.4243	.5218	.6767
WMR	.2239	.2622	.3240	.4258	.5106	.6509
A5	1.997	2.451	3.082	4.416	5.631	7.669
A6	1.812	2.193	2.806	4.013	5.370	7.306

Null Distribution Parameters Estimated

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.08559	.09379	.1045	.1202	.1342	.1519
D6	.0961	.1042	.1147	.1303	.1455	.1605
DMR	.1622	.1721	.1855	.2042	.2188	.2374
W5	.02626	.03120	.03801	.05103	.06626	.08648
W6	.02866	.03469	.04270	.05676	.06899	.09081
WMR	.07258	.07960	.09003	.1075	.1214	.1478
A5	.3596	.4414	.5551	.7616	1.024	1.312
A6	.3700	.4482	.5782	.7959	1.069	1.353

TABLE A3.2

CRITICAL VALUES--NORMAL DISTRIBUTION--
SAMPLE SIZE 20

Null Distribution Completely Specified

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.1521	.1666	.1885	.2160	.2354	.2685
D6	.1572	.1725	.1927	.2228	.2428	.2712
DMR	.2034	.2177	.2373	.2687	.2922	.3205
W5	.2018	.2491	.3199	.4267	.5299	.6916
W6	.2024	.2509	.3200	.4271	.5316	.6788
WMR	.2314	.2749	.3445	.4550	.5551	.6838
A5	1.447	1.755	2.183	2.907	3.791	5.325
A6	1.435	1.760	2.168	2.837	3.809	5.157

Null Distribution Parameters Estimated

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.05548	.06104	.06730	.07698	.08629	.09618
D6	.07071	.07698	.08498	.09649	.1083	.1204
DMR	.1335	.1409	.1510	.1646	.1754	.1921
W5	.02286	.02728	.03373	.04573	.05793	.07241
W6	.03240	.03866	.04739	.06295	.07948	.09941
WMR	.07858	.08654	.09843	.1212	.1396	.1662
A5	.2057	.2477	.3187	.4829	.6855	.9754
A6	.2656	.3250	.4123	.6126	.8104	1.112

TABLE A3.3

CRITICAL VALUES--NORMAL DISTRIBUTION--
SAMPLE SIZE 30

Null Distribution Completely Specified

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.1252	.1368	.1521	.1738	.1940	.2195
D6	.1281	.1390	.1540	.1765	.1962	.2232
DMR	.1717	.1835	.1992	.2211	.2407	.2661
W5	.1970	.2421	.3007	.4067	.5189	.6636
W6	.1982	.2428	.3015	.4068	.5243	.6624
WMR	.2365	.2757	.3371	.4365	.5554	.7058
A5	1.281	1.530	1.928	2.556	3.456	4.562
A6	1.277	1.534	1.903	2.563	3.396	4.517

Null Distribution Parameters Estimated

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.05076	.05525	.06136	.07162	.08047	.08940
D6	.05670	.06168	.06866	.07950	.08895	.09939
DMR	.1130	.1194	.1275	.1414	.1520	.1659
W5	.02544	.03011	.03764	.05045	.06426	.08333
W6	.03025	.03560	.04392	.05904	.07528	.09601
WMR	.07743	.08660	.09949	.1208	.1415	.1699
A5	.1816	.2198	.2747	.3948	.5619	.7823
A6	.2102	.2534	.3162	.4585	.6245	.8625

TABLE A3.4

CRITICAL VALUES--NORMAL DISTRIBUTION--
SAMPLE SIZE 40

Null Distribution Completely Specified

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.1066	.1162	.1289	.1511	.1709	.1916
D6	.1100	.1194	.1314	.1528	.1726	.1948
DMR	.1517	.1619	.1752	.1963	.2161	.2380
W5	.1957	.2234	.2915	.4101	.5133	.7017
W6	.1992	.2370	.2942	.4137	.5198	.7071
WMR	.2388	.2800	.3354	.4610	.5670	.7371
A5	1.159	1.390	1.723	2.367	3.176	4.183
A6	1.188	1.421	1.744	2.388	3.193	4.154

Null Distribution Parameters Estimated

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.04336	.04753	.05264	.06066	.06760	.07505
D6	.04942	.05352	.05936	.06798	.07591	.08455
DMR	.1016	.1075	.1134	.1239	.1346	.1456
W5	.02434	.02861	.03571	.04881	.06033	.07552
W6	.02997	.03510	.04333	.05841	.07208	.08959
WMR	.07907	.08729	.09978	.1211	.1433	.1654
A5	.1619	.1902	.2424	.3364	.4312	.5763
A6	.1964	.2309	.2864	.3942	.5003	.6480

TABLE A3.5

CRITICAL VALUES--NORMAL DISTRIBUTION--
SAMPLE SIZE 50

Null Distribution Completely Specified

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.09375	.1026	.1139	.1324	.1491	.1657
D6	.09685	.1054	.1167	.1349	.1516	.1692
DMR	.1363	.1456	.1583	.1748	.1926	.2129
W5	.1848	.2215	.2847	.3998	.4935	.6352
W6	.1903	.2287	.2921	.4070	.5046	.6440
WMR	.2325	.2740	.3305	.4510	.5541	.6931
A5	1.075	1.267	1.624	2.173	2.748	3.598
A6	1.112	1.319	1.659	2.218	2.784	3.619

Null Distribution Parameters Estimated

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.03915	.04272	.04772	.05455	.06073	.06843
D6	.04427	.04821	.05378	.06124	.06832	.07633
DMR	.09219	.09740	.1040	.1136	.1229	.1329
W5	.02435	.02934	.03571	.04717	.05780	.07374
W6	.03006	.03597	.04413	.05770	.07118	.08846
WMR	.07966	.08906	.1010	.1237	.1421	.1675
A5	.1620	.1911	.2335	.3120	.3920	.5080
A6	.1935	.2258	.2796	.3719	.4662	.5880

TABLE A3.6

CRITICAL VALUES--EXTREME VALUE DISTRIBUTION--
SAMPLE SIZE 10

Null Distribution Completely Specified

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.2318	.2534	.2808	.3256	.3656	.4104
D6	.2269	.2503	.2769	.3205	.3579	.4057
DMR	.2660	.2873	.3108	.3536	.3891	.4384
W5	.2401	.2868	.3559	.4802	.6194	.8060
W6	.2193	.2655	.3270	.4444	.5766	.7443
WMR	.2258	.2640	.3277	.4284	.5502	.7121
A5	2.060	2.578	3.269	4.516	6.049	8.173
A6	1.864	2.308	2.970	4.104	5.680	8.139

Null Distribution Parameters Estimated

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.08819	.09628	.1064	.1234	.1382	.1589
D6	.09683	.1052	.1162	.1316	.1446	.1646
DMR	.1646	.1739	.1867	.2069	.2247	.2471
W5	.03060	.03724	.04607	.06375	.08351	.1066
W6	.03277	.03936	.04948	.06446	.08231	.1068
WMR	.07576	.08359	.09478	.1124	.1320	.1544
A5	.3451	.4313	.5539	.7675	.9640	1.344
A6	.3586	.4367	.5500	.7644	1.010	1.340

TABLE A3.7

CRITICAL VALUES--EXTREME VALUE DISTRIBUTION--
SAMPLE SIZE 20

Null Distribution Completely Specified

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.1552	.1710	.1899	.2183	.2456	.2737
D6	.1585	.1733	.1911	.2211	.2489	.2760
DMR	.2048	.2183	.2356	.2661	.2911	.3183
W5	.2122	.2627	.3331	.4530	.5681	.7441
W6	.2061	.2516	.3201	.4363	.5523	.7129
WMR	.2336	.2722	.3316	.4491	.5514	.7138
A5	1.495	1.811	2.265	3.111	4.112	5.772
A6	1.465	1.767	2.202	3.014	4.056	5.731

Null Distribution Parameters Estimated

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.061170	.06642	.07342	.08512	.09431	.1078
D6	.06939	.07652	.08366	.09587	.1076	.1201
DMR	.1313	.1385	.1476	.1627	.1781	.1946
W5	.02757	.03302	.04118	.05543	.07108	.09624
W6	.03237	.03841	.04727	.06333	.08083	.1098
WMR	.07786	.08604	.09769	.1182	.1411	.1690
A5	.2189	.2724	.3623	.5310	.7965	1.185
A6	.2478	.3004	.3953	.5806	.8457	1.244

TABLE A3.8

CRITICAL VALUES--EXTREME VALUE DISTRIBUTION--
SAMPLE SIZE 30

Null Distribution Completely Specified

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.1245	.1360	.1512	.1751	.1958	.2205
D6	.1261	.1375	.1524	.1764	.1965	.2226
DMR	.1697	.1818	.1992	.2221	.2411	.2623
W5	.1988	.2383	.2968	.4213	.5244	.6631
W6	.1965	.2358	.2940	.4128	.5252	.6636
WMR	.2297	.2686	.3279	.4317	.5418	.6765
A5	1.279	1.523	1.909	2.587	3.339	4.461
A6	1.273	1.504	1.881	2.572	3.197	4.156

Null Distribution Parameters Estimated

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.05289	.05714	.06325	.07253	.08125	.09205
D6	.05660	.06117	.06748	.07660	.08682	.09707
DMR	.1120	.1178	.1252	.1385	.1494	.1625
W5	.02788	.03293	.04078	.05513	.07074	.09445
W6	.03094	.03655	.04480	.05850	.07518	.09842
WMR	.07716	.08507	.09728	.1194	.1419	.1678
A5	.1999	.2376	.2998	.4358	.5973	.8912
A6	.2175	.2562	.3186	.4448	.6115	.9352

TABLE A3.9

CRITICAL VALUES--EXTREME VALUE DISTRIBUTION--
SAMPLE SIZE 40

Null Distribution Completely Specified

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.1081	.1176	.1321	.1531	.1679	.1850
D6	.1098	.1206	.1348	.1542	.1693	.1869
DMR	.1507	.1623	.1762	.1953	.2124	.2309
W5	.1974	.2406	.2960	.4171	.5250	.6448
W6	.1969	.2398	.2957	.4152	.5133	.6365
WMR	.2331	.2735	.3401	.4477	.5476	.6613
A5	1.176	1.398	1.764	2.398	3.028	3.799
A6	1.186	1.414	1.754	2.367	3.022	3.826

Null Distribution Parameters Estimated

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.04923	.05265	.05720	.06406	.07202	.07997
D6	.05134	.05524	.06006	.06870	.07629	.08472
DMR	.1008	.1059	.1130	.1242	.1336	.1455
W5	.03104	.03627	.04378	.05671	.07083	.09323
W6	.03443	.03922	.04729	.06188	.07814	.09916
WMR	.08026	.08938	.1025	.1234	.1428	.1676
A5	.2109	.2503	.2995	.4034	.5309	.7445
A6	.2296	.2648	.3236	.4309	.5654	.7817

TABLE A3.10

CRITICAL VALUES--EXTREME VALUE DISTRIBUTION-- SAMPLE SIZE 50

Null Distribution Completely Specified

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.09797	.1067	.1181	.1363	.1530	.1727
D6	.09998	.1092	.1199	.1376	.1555	.1757
DMR	.1385	.1479	.1590	.1769	.1933	.2153
W5	.2042	.2425	.3032	.4239	.5267	.6965
W6	.2038	.2447	.3002	.4242	.5226	.6935
WMR	.2433	.2788	.3440	.4537	.5596	.7183
A5	1.173	1.403	1.733	2.345	2.978	3.813
A6	1.187	1.420	1.744	2.343	2.969	3.780

Null Distribution Parameters Estimated

	Alpha Level					
Statistic	.20	.15	.10	.05	.025	.01
D5	.04586	.04870	.05278	.05896	.06472	.07132
D6	.04669	.04976	.05413	.06058	.06797	.07452
DMR	.09065	.09508	.1014	.1110	.1185	.1295
W5	.03198	.03692	.04404	.05700	.06991	.08755
W6	.03388	.03984	.04734	.06140	.07556	.09243
WMR	.07911	.08751	.1006	.1185	.1392	.1647
A5	.2155	.2502	.2987	.3916	.4956	.6571
A6	.2271	.2637	.3173	.4142	.5208	.6863

Appendix 4

Power Comparisons

Tables A4.1 through A4.12 list the results of power comparisons made using the normal and extreme value distributions in the null hypothesis. Tables are listed by null distribution type (normal or extreme value), null hypothesis type (completely specified or parameters estimated) and alpha level (.10, .05, or .01). Each table includes eight distributions as alternative hypotheses and five different random sample sizes (four for the Cauchy). All entries represent the number of samples significant at the given alpha level from a Monte Carlo sample size of 1000 trials. Actual power of each test may be obtained by dividing each entry by 1000.

TABLE A4.1

POWER COMPARISONS FOR THE NORMAL DISTRIBUTION-- COMPLETELY SPECIFIED--ALPHA = .10

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Double Exponential	10	153	144	100	104	97	69	222	204
	20	160	159	125	126	110	94	224	199
	30	185	194	179	149	148	143	215	214
	40	193	207	200	145	149	164	204	212
	50	226	249	237	153	170	200	219	233
Uniform	10	91	103	157	90	99	135	74	77
	20	89	106	168	90	101	138	77	83
	30	104	116	215	105	109	169	114	122
	40	135	149	243	118	127	227	214	217
	50	139	158	294	120	131	266	300	313
Cauchy	10	400	356	282	365	322	277	144	134
	20	718	602	359	624	479	359	513	364
	30	907	843	456	791	704	491	807	736
	40	967	939	571	916	867	573	956	916
Exponential	10	199	212	212	186	187	192	243	250
	20	224	266	287	208	248	281	411	525
	30	414	429	401	363	388	420	891	935
	40	860	863	487	452	499	543	985	996
	50	967	972	634	620	648	685	995	1000

TABLE A4.1--Continued

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Gamma-2	10	143	145	135	125	130	127	155	150
	20	174	195	201	163	172	186	191	217
	30	223	230	236	195	204	228	259	290
	40	286	305	296	252	274	304	398	461
	50	362	387	358	292	318	388	519	592
Gamma-4	10	107	112	112	100	99	102	119	125
	20	110	123	121	107	111	117	116	120
	30	179	191	168	141	151	174	157	173
	40	199	205	202	158	174	208	174	194
	50	208	218	202	168	181	211	191	208
Gamma-6	10	106	123	119	99	107	107	109	100
	20	132	145	147	123	130	135	137	143
	30	160	160	168	145	149	159	135	139
	40	147	153	141	120	130	140	130	137
	50	178	183	171	155	160	180	171	180
Extreme Value	10	250	223	187	258	254	213	295	314
	20	446	404	324	471	425	363	593	596
	30	627	602	450	627	610	532	789	788
	40	754	735	593	764	751	686	901	899
	50	853	832	715	852	829	788	947	951

TABLE A4.2

POWER COMPARISONS FOR THE NORMAL DISTRIBUTION--
COMPLETELY SPECIFIED--ALPHA = .05

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Double Exponential	10	76	72	34	41	43	35	120	117
	20	93	78	54	59	55	37	150	143
	30	107	111	93	73	71	72	137	132
	40	96	101	97	59	57	57	115	117
	50	128	143	152	65	69	87	126	135
Uniform	10	58	60	87	60	63	73	33	30
	20	46	53	92	47	51	67	39	42
	30	69	79	128	62	69	110	66	70
	40	72	85	156	66	71	107	106	109
	50	82	90	201	60	68	128	160	160
Cauchy	10	272	226	158	231	183	169	75	70
	20	577	408	228	456	303	218	289	190
	30	806	705	322	634	514	334	622	500
	40	918	847	394	799	662	360	868	771
Exponential	10	127	130	140	110	116	114	139	133
	20	143	173	189	126	145	171	248	317
	30	271	290	291	229	254	296	640	736
	40	361	380	359	282	329	378	921	953
	50	856	870	484	409	455	531	978	994

TABLE A4.2--Continued

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Gamma-2	10	76	82	72	65	64	66	80	83
	20	110	116	117	95	102	109	130	138
	30	151	158	155	113	129	150	144	162
	40	193	208	202	143	160	190	214	266
	50	241	255	258	166	189	241	303	372
Gamma-4	10	55	59	62	57	53	55	67	62
	20	69	65	64	59	60	62	60	66
	30	88	97	116	73	76	90	90	90
	40	109	113	121	89	92	100	97	104
	50	131	140	140	94	105	128	95	115
Gamma-6	10	53	60	62	52	53	55	56	61
	20	56	57	59	46	44	48	44	44
	30	96	100	98	80	82	97	72	75
	40	77	82	83	60	67	75	56	65
	50	125	133	118	95	105	116	87	97
Extreme Value	10	141	139	91	156	144	120	148	169
	20	309	266	187	333	301	246	448	463
	30	459	431	297	470	441	384	663	647
	40	580	557	422	608	584	523	778	778
	50	739	716	556	732	710	650	878	884

TABLE A4.3

POWER COMPARISONS FOR THE NORMAL DISTRIBUTION--
COMPLETELY SPECIFIED--ALPHA = .01

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Double Exponential	10	13	12	6	8	7	4	26	30
	20	25	23	11	9	8	8	44	42
	30	24	21	17	17	18	13	34	32
	40	23	26	17	7	4	6	33	32
	50	28	29	32	13	13	12	34	34
Uniform	10	16	19	28	18	24	29	4	4
	20	11	13	19	12	13	18	10	11
	30	25	27	42	18	22	27	12	12
	40	23	26	44	14	17	26	11	13
	50	22	26	52	19	22	29	42	43
Cauchy	10	112	68	56	89	65	51	16	14
	20	319	197	87	161	100	86	41	30
	30	474	346	130	268	204	114	132	87
	40	690	516	148	371	250	110	372	224
Exponential	10	52	46	43	35	36	32	52	48
	20	56	73	69	42	50	58	71	73
	30	135	141	148	83	104	132	162	202
	40	149	167	179	98	115	160	347	472
	50	218	246	260	158	192	258	840	906

TABLE A4.3--Continued

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Gamma-2	10	23	20	22	15	14	14	23	19
	20	34	48	41	23	27	37	32	34
	30	41	44	51	23	28	39	32	35
	40	66	72	76	38	44	65	49	60
	50	98	111	106	47	63	97	73	107
Gamma-4	10	15	14	17	12	11	13	14	11
	20	17	19	20	11	15	18	10	11
	30	24	25	32	21	22	25	18	16
	40	35	40	35	24	25	31	19	24
	50	48	52	52	36	41	48	29	33
Gamma-6	10	12	12	16	7	7	9	13	13
	20	29	33	31	21	23	29	20	20
	30	28	29	30	24	25	27	21	19
	40	20	20	24	14	15	17	15	17
	50	40	40	37	25	26	32	23	25
Extreme Value	10	32	26	27	47	43	31	36	34
	20	108	91	49	122	109	91	151	160
	30	166	147	98	220	210	161	299	299
	40	281	241	176	314	284	245	464	469
	50	424	385	248	494	450	379	683	688

TABLE A4.4

POWER COMPARISONS FOR THE NORMAL DISTRIBUTION--
PARAMETERS ESTIMATED--ALPHA = .10

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Double Exponential	10	247	256	210	288	270	227	211	173
	20	396	385	298	429	397	334	351	295
	30	492	491	407	494	498	452	430	422
	40	577	567	445	563	569	516	517	525
	50	639	633	538	637	653	621	593	620
Uniform	10	190	163	147	84	90	189	251	263
	20	246	159	173	71	72	253	318	337
	30	330	288	288	160	154	389	444	490
	40	387	367	382	260	268	557	545	628
	50	444	453	444	413	390	629	660	722
Cauchy	10	651	664	616	702	677	647	604	531
	20	908	908	871	915	907	891	887	858
	30	973	973	960	972	972	969	968	970
	40	995	994	989	995	995	994	993	993
Exponential	10	581	578	441	540	577	524	573	608
	20	888	839	697	854	852	807	915	919
	30	968	966	857	966	966	937	985	986
	40	991	991	955	990	988	982	994	996
	50	999	999	989	999	999	997	1000	1000

TABLE A4.4--Continued

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Gamma-2	10	336	327	282	331	336	302	325	352
	20	598	583	446	583	601	543	609	639
	30	764	760	604	780	781	722	839	844
	40	882	858	727	886	877	828	903	905
	50	927	915	795	942	934	890	957	962
Gamma-4	10	238	225	191	216	230	207	222	230
	20	357	333	237	348	350	298	347	361
	30	512	491	387	511	513	450	541	544
	40	627	612	476	641	634	567	656	664
	50	677	656	536	715	700	617	728	734
Gamma-6	10	186	188	169	190	194	170	171	176
	20	317	311	241	314	318	283	308	304
	30	417	401	297	424	412	354	428	430
	40	457	431	311	469	460	366	459	463
	50	550	533	397	562	551	470	573	574
Extreme Value	10	246	246	200	257	263	231	240	226
	20	395	382	298	391	386	358	395	388
	30	515	499	382	526	524	461	529	517
	40	646	618	468	657	647	557	657	659
	50	711	689	530	729	717	642	731	728

TABLE A4.5

POWER COMPARISONS FOR THE NORMAL DISTRIBUTION--
PARAMETERS ESTIMATED--ALPHA = .05

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Double Exponential	10	161	178	132	206	184	163	135	110
	20	289	282	207	319	285	254	201	169
	30	377	382	289	387	385	350	311	302
	40	454	446	328	435	443	413	366	388
	50	512	526	410	522	537	503	474	499
Uniform	10	106	86	81	38	46	102	141	146
	20	159	83	88	26	32	131	265	263
	30	200	159	170	73	72	263	366	390
	40	272	233	225	110	118	375	448	504
	50	351	313	302	207	199	476	548	620
Cauchy	10	574	591	550	638	612	591	532	455
	20	869	871	838	882	866	860	820	808
	30	957	955	937	967	966	962	949	951
	40	991	992	980	990	990	992	988	989
Exponential	10	463	454	333	431	470	423	458	498
	20	827	773	595	793	785	714	845	840
	30	939	922	750	939	934	887	962	965
	40	984	983	914	980	978	967	991	991
	50	999	998	959	998	996	993	1000	1000

TABLE A4.5--Continued

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Gamma-2	10	237	234	183	244	240	206	240	240
	20	481	460	329	465	462	422	475	477
	30	665	636	460	693	691	604	746	759
	40	800	779	613	807	803	733	845	848
	50	881	861	689	891	881	835	923	928
Gamma-4	10	135	131	114	134	139	127	135	155
	20	239	231	152	226	223	180	231	241
	30	378	367	259	393	388	338	419	419
	40	507	479	351	512	502	435	533	541
	50	575	541	400	597	580	493	607	610
Gamma-6	10	109	115	95	115	120	98	101	101
	20	228	220	169	223	215	190	208	207
	30	292	286	185	315	314	244	313	304
	40	329	306	223	317	310	257	323	332
	50	416	401	278	447	426	354	437	435
Extreme Value	10	146	168	123	177	178	152	149	146
	20	298	298	205	301	302	237	277	280
	30	391	367	251	408	407	323	405	389
	40	534	499	362	534	518	441	523	521
	50	611	590	420	630	610	513	621	614

TABLE A4.6

POWER COMPARISONS FOR THE NORMAL DISTRIBUTION--
PARAMETERS ESTIMATED--ALPHA = .01

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Double Exponential	10	61	72	50	87	80	58	58	41
	20	131	132	92	150	131	116	55	50
	30	216	224	166	214	216	208	120	120
	40	250	250	179	254	263	253	169	192
	50	289	297	226	312	331	339	246	284
Uniform	10	19	17	25	6	13	29	29	26
	20	54	12	23	4	5	26	171	145
	30	69	34	47	13	15	80	279	282
	40	117	71	69	34	39	137	317	356
	50	148	94	91	51	47	230	392	451
Cauchy	10	442	468	443	489	471	474	400	337
	20	801	787	725	816	799	782	707	697
	30	924	930	896	926	929	929	886	887
	40	974	975	952	975	980	978	959	964
Exponential	10	246	239	165	215	268	224	239	282
	20	656	569	380	610	578	528	599	635
	30	835	797	530	826	816	730	887	892
	40	956	943	766	955	953	911	980	981
	50	988	985	844	989	986	961	995	997

TABLE A4.6--Continued

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Gamma-2	10	94	108	90	101	117	98	103	104
	20	285	251	163	285	280	230	231	252
	30	445	403	245	462	458	377	471	487
	40	630	583	373	662	647	543	716	724
	50	732	706	442	778	764	669	823	835
Gamma-4	10	41	47	42	40	50	42	43	38
	20	96	83	51	102	83	65	79	86
	30	203	183	112	204	201	160	178	186
	40	308	278	164	322	313	244	321	331
	50	361	339	193	402	390	312	410	417
Gamma-6	10	29	30	28	31	31	26	21	25
	20	91	85	58	95	93	67	63	73
	30	129	123	72	135	128	109	107	108
	40	150	131	83	157	152	124	158	164
	50	229	208	117	256	250	183	251	258
Extreme Value	10	51	61	38	61	64	54	46	47
	20	140	126	83	153	133	105	100	108
	30	209	200	117	224	217	173	193	192
	40	342	306	163	351	346	266	332	340
	50	393	369	238	437	418	335	425	433

TABLE A4.7

POWER COMPARISONS FOR THE EXTREME VALUE DISTRIBUTION--
COMPLETELY SPECIFIED--ALPHA = .10

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Normal	10	128	140	155	138	147	156	91	89
	20	168	181	189	149	164	185	111	112
	30	198	207	199	180	186	195	141	149
	40	244	239	232	209	215	221	170	174
	50	279	293	272	261	265	276	224	235
Double Exponential	10	145	142	125	124	117	116	155	154
	20	197	202	199	151	149	160	200	186
	30	286	300	270	223	230	246	233	239
	40	351	363	340	275	289	301	283	294
	50	407	416	406	311	342	365	315	335
Uniform	10	125	137	231	144	155	198	77	76
	20	143	169	291	134	153	261	105	114
	30	190	205	359	163	177	357	163	182
	40	264	276	437	218	231	431	339	358
	50	323	340	490	243	263	492	451	477
Cauchy	10	404	382	330	372	357	326	182	166
	20	732	676	473	627	558	514	533	458
	30	916	893	663	861	817	692	854	826
	40	965	946	773	940	909	816	964	943

TABLE A4.7--Continued

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Exponential	10	1000	1000	806	705	690	489	993	1000
	20	1000	1000	1000	1000	1000	1000	1000	1000
	30	1000	1000	1000	1000	1000	1000	1000	1000
	40	1000	1000	1000	1000	1000	1000	1000	1000
	50	1000	1000	1000	1000	1000	1000	1000	1000
Logistic	10	273	318	357	301	319	340	176	181
	20	539	560	547	463	496	545	366	397
	30	801	801	706	693	698	709	672	686
	40	926	908	822	817	820	797	839	839
	50	964	956	883	884	884	874	926	915
χ_1^2	10	297	302	322	329	313	304	169	169
	20	471	444	417	471	435	436	351	323
	30	676	667	587	656	650	617	599	594
	40	801	776	678	788	765	705	780	764
	50	883	866	764	861	852	795	883	860
χ_4^2	10	1000	1000	1000	999	1000	1000	989	991
	20	1000	1000	1000	1000	1000	1000	1000	1000
	30	1000	1000	1000	1000	1000	1000	1000	1000
	40	1000	1000	1000	1000	1000	1000	1000	1000
	50	1000	1000	1000	1000	1000	1000	1000	1000

TABLE A4.8

POWER COMPARISONS FOR THE EXTREME VALUE DISTRIBUTION--COMPLETELY SPECIFIED--ALPHA = .05

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Normal	10	63	65	84	67	69	74	47	50
	20	95	99	98	95	90	88	65	64
	30	125	123	124	97	105	119	68	72
	40	145	152	152	122	121	134	103	105
	50	181	189	175	140	145	154	112	114
Double Exponential	10	73	76	63	58	52	46	90	88
	20	115	115	102	85	88	93	115	113
	30	178	181	173	120	125	148	142	131
	40	231	252	238	153	164	192	161	181
	50	294	312	291	166	185	233	174	191
Uniform	10	71	87	128	77	87	109	46	48
	20	91	108	185	79	89	136	56	63
	30	114	119	231	89	95	203	80	85
	40	165	177	304	133	137	264	176	190
	50	192	211	326	129	141	306	266	293
Cauchy	10	276	246	218	250	222	218	113	105
	20	607	490	351	475	402	376	322	279
	30	840	792	524	707	647	574	684	619
	40	927	895	666	875	924	694	904	872

TABLE A4.8--Continued

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Exponential	10	822	834	318	380	372	248	908	960
	20	1000	1000	1000	991	988	924	1000	1000
	30	1000	1000	1000	1000	1000	1000	1000	1000
	40	1000	1000	1000	1000	1000	1000	1000	1000
	50	1000	1000	1000	1000	1000	1000	1000	1000
Logistic	10	177	205	223	212	220	249	109	113
	20	414	436	416	344	365	398	231	251
	30	672	677	598	541	563	590	490	497
	40	824	817	723	666	667	677	671	684
	50	919	906	820	781	775	786	827	828
χ_1^2	10	198	193	189	222	205	203	110	107
	20	353	330	293	348	315	316	240	202
	30	554	542	465	531	516	487	457	444
	40	691	663	562	662	634	579	646	613
	50	799	772	665	747	719	665	777	744
χ_4^2	10	995	995	997	999	999	1000	989	991
	20	1000	1000	1000	1000	1000	1000	1000	1000
	30	1000	1000	1000	1000	1000	1000	1000	1000
	40	1000	1000	1000	1000	1000	1000	1000	1000
	50	1000	1000	1000	1000	1000	1000	1000	1000

TABLE A4.9

POWER COMPARISONS FOR THE EXTREME VALUE DISTRIBUTION--COMPLETELY SPECIFIED--ALPHA = .01

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Normal	10	20	18	20	19	20	18	15	12
	20	33	39	33	24	26	27	16	13
	30	45	47	52	37	37	39	24	27
	40	67	69	56	43	49	55	21	23
	50	62	60	55	41	40	45	32	33
Double Exponential	10	15	12	8	6	6	6	21	21
	20	39	36	24	25	23	19	31	32
	30	59	59	60	36	34	38	35	39
	40	98	115	117	60	64	74	61	64
	50	106	110	115	50	54	74	57	64
Uniform	10	23	23	34	22	25	29	10	9
	20	29	32	56	25	28	34	11	10
	30	42	44	92	32	34	62	17	22
	40	79	82	126	58	62	93	47	48
	50	65	67	130	47	49	93	56	61
Cauchy	10	113	92	78	101	90	91	41	32
	20	356	251	179	222	192	185	76	63
	30	617	531	345	433	372	354	270	290
	40	834	757	463	642	581	501	637	544

TABLE A4.9--Continued

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Exponential	10	129	126	37	65	68	39	336	339
	20	1000	1000	1000	500	460	318	1000	1000
	30	1000	1000	1000	994	990	945	1000	1000
	40	1000	1000	1000	1000	1000	1000	1000	1000
	50	1000	1000	1000	1000	1000	1000	1000	1000
Logistic	10	66	71	90	85	90	92	25	21
	20	200	221	221	172	182	199	70	65
	30	396	409	389	324	328	348	218	243
	40	628	618	527	456	459	480	401	402
	50	752	728	619	538	540	563	545	549
χ_1^2	10	79	68	69	91	88	72	24	22
	20	189	164	149	171	146	127	55	43
	30	337	319	261	333	311	265	188	198
	40	482	457	344	437	406	362	357	319
	50	567	523	397	522	480	421	497	460
χ_4^2	10	971	974	974	986	990	992	925	881
	20	1000	1000	999	1000	1000	1000	1000	1000
	30	1000	1000	1000	1000	1000	1000	1000	1000
	40	1000	1000	1000	1000	1000	1000	1000	1000
	50	1000	1000	1000	1000	1000	1000	1000	1000

TABLE A4.10

POWER COMPARISONS FOR THE EXTREME VALUE DISTRIBUTION--PARAMETERS ESTIMATED--ALPHA = .10

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Normal	10	279	246	163	271	223	161	259	235
	20	490	409	252	491	419	280	471	438
	30	638	580	373	663	596	406	682	653
	40	772	684	420	739	700	488	758	744
	50	787	762	501	821	785	604	820	807
Double Exponential	10	363	349	263	372	327	277	338	286
	20	670	625	486	670	636	549	633	594
	30	806	783	641	794	782	710	776	765
	40	871	865	760	864	856	797	845	840
	50	924	917	846	911	910	882	902	905
Uniform	10	290	254	211	216	194	261	350	363
	20	537	488	319	476	410	414	557	575
	30	688	698	465	708	700	571	744	775
	40	765	774	531	776	775	663	807	841
	50	838	855	651	860	856	802	868	901
Cauchy	10	634	657	599	643	641	628	592	535
	20	900	906	876	897	908	883	866	875
	30	976	982	968	976	980	981	968	972
	40	992	994	991	992	993	993	992	993

TABLE A4.10--Continued

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Exponential	10	145	205	229	133	230	271	177	248
	20	163	319	378	149	380	458	296	447
	30	392	464	542	453	549	633	644	713
	40	547	585	583	571	656	692	801	832
	50	726	730	713	714	765	807	906	922
Logistic	10	308	284	182	321	279	203	317	266
	20	581	503	315	585	528	373	574	544
	30	707	673	480	718	688	546	719	700
	40	764	740	518	779	748	603	779	762
	50	850	827	641	861	843	730	860	853
X_1^2	10	90	99	142	93	112	143	99	114
	20	57	96	125	56	95	151	66	88
	30	85	112	150	90	120	164	92	116
	40	75	109	171	79	114	166	104	129
	50	94	103	174	88	127	186	120	144
X_4^2	10	148	128	112	140	130	108	151	133
	20	194	162	97	188	159	109	181	158
	30	206	186	123	200	189	142	205	189
	40	246	210	121	226	209	129	226	215
	50	218	203	135	229	212	139	224	214

TABLE A4.11

POWER COMPARISONS FOR THE EXTREME VALUE DISTRIBUTION--
PARAMETERS ESTIMATED--ALPHA = .05

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Normal	10	161	145	98	159	138	98	159	140
	20	327	297	159	349	308	189	342	327
	30	490	459	221	513	478	282	525	513
	40	623	550	289	634	591	377	650	626
	50	685	559	373	725	676	490	730	716
Double Exponential	10	261	254	187	274	242	202	236	197
	20	566	532	374	578	542	461	535	491
	30	734	712	551	727	719	634	700	693
	40	822	810	674	806	809	744	788	790
	50	889	889	784	874	879	842	859	864
Uniform	10	175	150	106	107	109	131	217	220
	20	394	337	202	302	256	274	453	453
	30	591	585	318	531	498	420	639	672
	40	672	662	370	671	651	512	725	758
	50	749	780	493	769	762	696	790	829
Cauchy	10	565	591	523	573	573	564	516	460
	20	859	874	824	854	872	865	814	826
	30	961	969	956	963	970	971	944	956
	40	990	992	986	991	992	991	984	988

TABLE A4.11--Continued

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Exponential	10	79	117	155	76	143	188	106	152
	20	97	221	250	81	248	324	172	294
	30	272	367	390	329	418	495	466	585
	40	371	412	439	434	523	581	702	738
	50	555	597	574	590	663	723	833	863
Logistic	10	214	189	108	220	187	117	210	179
	20	447	387	210	463	404	262	436	399
	30	628	593	365	633	610	451	626	618
	40	693	653	403	699	675	509	699	685
	50	787	769	521	792	776	633	796	789
X_1^2	10	40	64	67	48	65	77	49	62
	20	30	48	67	27	42	65	31	47
	30	39	65	78	46	65	83	41	58
	40	31	53	99	39	63	98	49	70
	50	40	56	102	41	71	120	54	76
X_4^2	10	82	72	54	81	70	46	82	83
	20	100	88	54	110	87	58	106	93
	30	129	121	58	130	125	69	129	118
	40	143	114	64	140	120	67	133	129
	50	152	147	79	148	137	77	141	131

TABLE A4.12

POWER COMPARISONS FOR THE EXTREME VALUE DISTRIBUTION--
PARAMETERS ESTIMATED--ALPHA = .01

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Normal	10	42	28	14	41	33	20	48	36
	20	126	128	48	139	118	84	121	119
	30	215	188	75	231	216	111	230	211
	40	356	323	127	357	339	206	366	359
	50	451	431	174	494	477	280	518	517
Double Exponential	10	120	113	70	129	115	97	119	94
	20	352	344	196	371	346	279	289	251
	30	565	544	365	569	569	473	486	471
	40	698	690	487	682	687	631	636	636
	50	799	794	611	777	792	742	749	766
Uniform	10	36	31	18	17	20	37	56	72
	20	146	115	40	72	56	68	232	232
	30	298	269	113	166	158	174	421	436
	40	451	410	155	291	270	283	522	559
	50	571	573	226	507	482	397	615	676
Cauchy	10	428	463	383	438	427	444	400	352
	20	765	795	722	761	787	784	659	680
	30	925	940	910	921	938	943	874	886
	40	972	982	961	970	979	985	950	960

TABLE A4.12--Continued

Alternative Distribution	Sample Size	D5	D6	DMR	W5	W6	WMR	A5	A6
Exponential	10	13	33	39	28	37	54	27	34
	20	31	84	92	25	71	125	47	67
	30	105	170	184	106	199	264	162	223
	40	147	215	228	176	267	336	388	456
	50	253	340	321	325	440	487	647	686
Logistic	10	77	68	36	76	64	49	80	64
	20	205	187	83	224	193	123	177	166
	30	394	375	174	401	395	259	379	362
	40	472	446	254	488	468	367	489	477
	50	615	584	325	639	622	467	626	628
χ_1^2	10	6	13	17	12	15	23	8	10
	20	8	8	14	6	4	12	9	9
	30	7	18	29	9	12	33	4	6
	40	5	10	22	4	9	35	9	13
	50	3	14	30	9	19	45	12	15
χ_4^2	10	18	15	8	19	17	9	17	16
	20	25	18	13	24	18	8	22	23
	30	31	27	17	33	30	17	33	28
	40	30	23	17	24	21	14	26	27
	50	43	36	16	44	39	18	37	39

Appendix 5

Computational Methods Used

This appendix describes various numerical methods used throughout this study. In particular, we will describe methods for random variate generation, numerical integration, and iterative solution for inverting the approximated distribution function. All calculations were performed using a CDC Cyber 74/750 system located at the Aeronautical Systems Division Computer Center, Wright-Patterson Air Force Base, Ohio.

Generating Random Variates

Depending on the underlying distribution, random variates were generated from two main sources. Uniform random variables were constructed using the multiplicative congruential generator described by McGrath and Irving (Ref 54). Random samples from the double exponential, exponential, triangular, and extreme value distributions were generated by applying the corresponding inverse probability integral transform to a set of uniform random variates. Random samples from the four parameter λ family of Rambert, et al., were generated by transforming uniform random variates using the percentile function $R(p) = \lambda_1 + [p^{\lambda_3} - (1-p)^{\lambda_4}]/\lambda_2$ where the λ_i, $i=1,\ldots,4$ are the parameters of the specific λ distribution, and p is a uniform random variate on [0,1] (Ref 72). Subroutines from the International Methematical and Statistical Libraries were used to generate random samples for the normal (using the polar method) Weibull, gamma, beta, and Cauchy distributions. If necessary, location and/or scale transformations were applied to adjust standard variates to specific underlying populations.

Numerical Integration

Two specific procedures used for evaluating the finite integral, $\int_a^b f(x)\,dx$, were Gaussian quadrature and Simpson's rule. Initially, in determining the variables for the nonparametric estimators, a sixteen point Gauss-Legendre quadrature scheme was used for the following integrands

1. $(F(x)-SF(x))^2\, sf(x)$

2. $(f(x)-sf(x))^2\, sf(x)$

Quadrature points and weights were taken from tables in reference 1, page 916. The interval of integration was the support of the nonparametric estimate $[X_{min}, X_{max}]$.

To evaluate the integrals used for comparisons of approximate mean integrated square error for both distribution and density functions and the integrals used in calculating the goodness of fit statistics, we used a modified Simpson's rule with error control (Ref 66). Given an ordered sample of size n and the two endpoints of the support of the nonparametric approximation, we constructed n+1 intervals of the form $[X_{(i)},X_{(i+1)}]$ i=0,...,n where $X_{(0)}=X_{min}$ and $X_{(n+1)}=X_{max}$. For each integrand, we used Simpson's rule on each interval. If the summed value of the approximation was not sufficiently close, we divided each interval in half and repeated the procedure. Integrands evaluated by this method included:

1. $(F(x)-SF(x))^2\, sf(x)$
2. $(f(x)-sf(x))^2\, sf(x)$
3. $(F(x)-SF(x))^2\, sf(x)/[SF(x)(1-SF(x))]$
4. $sf(x)$

A stopping criterion for integral convergence was selected based on the construction of our nonparametric density estimate. We know that $\int sf(x)\,dx = 1$ on $[X_{min},X_{max}]$. We also know that the underlying distribution function F and density function f are reasonably smooth. By using subintervals based on the data points, we should be able to detect any "spikes" in the integrands. Using this information, we used as the approximation to each integral, the value of the Simpson's rule calculations when $|sf(x)-1.0|\leq 0.01$. Since sf(x) is the "noisiest" contribution to the four integrands, approximating $\int sf(x)\,dx$ to a sufficient degree gives us a measure of confidence in the remaining integral approximations.

To see numerically how the choice of stopping criterion affected the other integrals, we generated twenty-five random samples of size 100 from the standard normal distribution. Then we calculated the modified CVM integrals for both the distribution and density functions as well as the integral of the density function approximation using all six nonparametric models. We used two different stopping criterion values, $|\int sf(x)\,dx - 1.0\,| \leq ERR$ where ERR = 0.01 or 0.001. Table A5.1 lists the average values of the integrals for the twenty-five samples. Each entry corresponds to a specific model approximation, integrand and choice of ERR. A comparison between the entries corresponding to ERR choices of 0.01 and 0.001 for each integrand shows that a tighter bound on the integral of the density approximation has a negligible effect. The convergence error criterion was then set at 0.01.

To evaluate the integrals associated with the location parameter estimates of Chapter VI, we again used a modified Simpson's rule. We divided the support into subintervals using the data points as before. However, since we only needed one integral evaluated, we chose a straightforward application of Simpson's rule with error control. The integral, $\int x\, sf(x)\,dx$, was said to converge when the change in the approximation was less than 0.1 percent.

TABLE A5.1

INTEGRAL COMPARISON BY MODEL AND STOPPING CRITERION

	INTEGRAND					
	$(F(x)-SF(x))^2 sf(x)$		$(f(x)-sf(x))^2 sf(x)$		$sf(x)$	
	ERR		ERR		ERR	
Model	0.01	0.001	0.01	0.001	0.01	0.001
1	.0014769	.0014758	.0025480	.0025453	1.0019406	1.0001775
2	.0013981	.0013968	.0017674	.0017654	1.0042248	1.0000453
3	.0014876	.0014863	.0022938	.0022909	1.0034149	1.0000980
4	.0066093	.0066093	.0098837	.0098837	0.9999472	0.9994720
5	.0014769	.0014758	.0025480	.0025453	1.0019406	1.0001775
6	.0014487	.0014475	.0021852	.0021828	1.0035479	0.9998360

Iterative Solution for Inverting the Approximated Distribution Function

To calculate the pseudosample points for the smoothing routine or to calculate any percentile, such as the median, we needed a method for inverting the sample distribution function. Since we can calculate the density function at any point a Newton Raphson iteration scheme was employed. The nth approximation $x^{(n)}$ was calculated as $x^{(n)}=x^{(n-1)} - SF(x^{(n-1)})/sf(x^{(n-1)})$. Convergence was defined when the absolute value of the difference between successive approximations was less than 10^{-5} (Ref 66).

Appendix 6

A Finite Support Modification to Insure Inclusion of All Original Data Points

For either an extremely leptokurtic or platykurtic distribution, the smoothing routine sometimes generated a pseudosample for which the support of the nonparametric distribution function did not contain the interval $[X_{(1)}, X_{(n)}]$ where $X_{(1)}$ and $X_{(n)}$ are the extreme order statistics of the original sample. To insure that the interval $[X_{min}, X_{max}]$, the support generated by the pseudosample, the following algorithm was added. If X_{min}, the lower endpoint of the finite support based on a pseudosample, is greater than $X_{(1)}$, the smallest order statistic of the original sample, replace the inversion point of the pseudosample determined by $FS^{-1}(G_1)$ by $X_{(1)}$, and similarly for X_{max} less than $X_{(n)}$. This modification uses the information that the distribution function is defined over at least the set $[X_{(1)}, X_{(n)}]$, and also only adds enough tail weight by adjusting the pseudosample to insure that the final support contains the original data points.

The above modification was used for all models except Model 3. Since Model 3 uses fixed $X_{(0)}$ and $X_{(n+1)}$ extrapolation points for all subsamples, we merely set

$X_{min} = X_{(0)}$ and/or $X_{max} = X_{(n+1)}$, where $X_{(0)}$ and $X_{(n+1)}$ were the extrapolation points based on the entire sample, whenever the interval $[X_{min}, X_{max}]$ did not contain $[X_{(1)}, X_{(n)}]$. This again insured that the final distribution function approximation was defined over a finite support which contained all of the data points.

Vita

James Sweeder was born on 23 November 1949 in Mount Carmel, Pennsylvania. He graduated from Our Lady of Lourdes Regional High School in Shamokin, Pennsylvania in 1967. Upon graduation from the United States Air Force Academy, he received both a Bachelor of Science degree in Mathematics and a commission in the United States Air Force in June 1971. In March 1972, he earned a Master of Science degree from Colorado State University, specializing in mathematics. He was then assigned to the Engineering Directorate of the Foreign Technology Division at Wright-Patterson AFB, Ohio as a mathematician and trajectory analyst until March 1975. He then served as a Minuteman III crew commander, instructor, evaluator, and senior evaluator for the 321st Strategic Missile Wing, Grand Forks AFB, North Dakota. While there, he received a Master of Business Administration degree from the University of North Dakota in December 1977. He entered the School of Engineering, Air Force Institute of Technology, in August 1979.

Permanent Address: 104 North Locust Street
Mount Carmel, Pennsylvania
17851

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFIT/DS/MA/82-1	2. GOVT ACCESSION NO. AD-A115491	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* NONPARAMETRIC ESTIMATION OF DISTRIBUTION AND DENSITY FUNCTIONS WITH APPLICATIONS		5. TYPE OF REPORT & PERIOD COVERED Ph.D. Dissertation 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) James Sweeder		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Air Force Institute of Technology (AFIT/EN) Wright-Patterson AFB, Ohio 45433		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Project 2404-01-79
11. CONTROLLING OFFICE NAME AND ADDRESS (AFWAL/FIMB) Air Force Wright Aeronautical Laboratories Flight Dynamics Laboratory Vehicle Synthesis Branch Wright-Patterson AFB, Ohio 45433		12. REPORT DATE May 1982 13. NUMBER OF PAGES 257
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

Approved for public release; IAW AFR 190-17

~~Frederic C. Lynch~~, Major, USAF
Director, Office of Public Affairs

LYNN E. WOLAVER
Dean for Research and
Professional Development

Lynn E Wolaver

28 MAY 1985

AIR FORCE INSTITUTE OF TECHNOLOGY (ATC)
WRIGHT-PATTERSON AFB, OH 45433

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Probability Distribution Function
Probability Density Estimation
Nonparametric Estimation
Goodness of Fit Tests
Location Parameter Estimation
Reliability
Hazard Function

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

This report presents the theoretical development, evaluation, and applications of a new nonparametric family of continuous, differentiable, sample distribution functions. Given a random sample of independent, identically distributed random variables, estimators are constructed which converge uniformly to the underlying distribution. A smoothing routine is proposed which preserves distribution function properties. Using mean integrated square error as a criterion, the new estimators are shown to compare favorably against the

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

empirical distribution function. As density estimators, their derivatives are shown to be competitive with other continuous approximations. Numerous graphical examples are given. New goodness of fit tests for the normal and extreme value distributions are proposed and eight new goodness of fit statistics are developed. Monte Carlo studies are conducted to determine the critical values and powers for tests when the null hypothesis is completely specified and when the parameters are estimated. These tests were shown to be comparable with or superior to tests currently used. Forty-eight new estimators of the location parameter of a symmetric distribution are proposed. For mild deviations from the normal distribution, some new estimators are shown to be superior to established robust estimators. Robust characteristics of the new estimators are discussed.