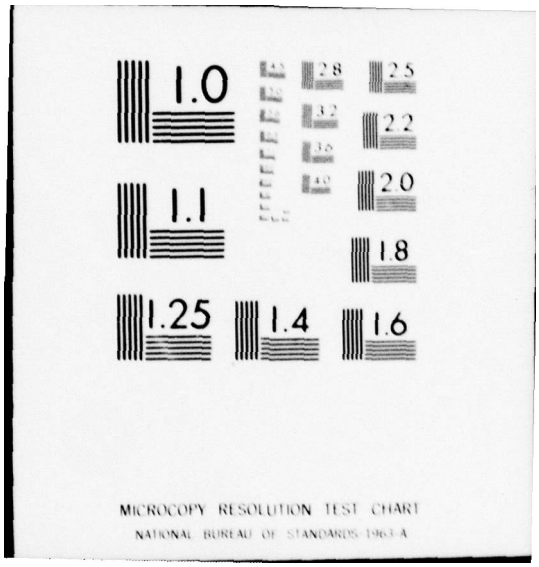
AD-A056 707 STATE UNIV OF NEW YORK AT BUFFALO AMHERST STATISTICA--ETC F/G 12/1
A DENSITY-QUANTILE FUNCTION PERSPECTIVE ON ROBUST ESTIMATION.(U)
MAR 78 E PARZEN DAAG29-76-G-0239

UNCLASSIFIED 60 ARO-13845.5-M NL

1 OF 1
AD
A056707

END
DATE
FILMED
9-78
DDC


1.0
1.1
1.25
1.4
1.6
2.8
2.5
2.2
2.0
1.8
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

State University of New York at Buffalo
Department of Computer Science
STATISTICAL SCIENCE
4230 Ridge Lea Road, Room A32, Amherst, New York 14228 Telephone: 716-831-1222

LEVEL II (18) ARO (19) 13845.5-M (12)

A DENSITY-QUANTILE FUNCTION PERSPECTIVE ON ROBUST ESTIMATION.*

by

(10) Emanuel Parzen

Statistical Science Division
State University of New York at Buffalo

An Invited Paper to be Published in the
Proceedings of the ARO Workshop in Robustness in Statistics

(9) GRANT TECHNICAL REPORT NO. ARO-4
STATISTICAL SCIENCE DIVISION REPORT NO. 60 (14)

(11) March 1978

(12) 84 p. (15) (6)

*Research supported by the Army Research Office (Grant DA AG29-76-0239).

A DENSITY-QUANTILE FUNCTION PERSPECTIVE ON ROBUST ESTIMATION*

by

Emanuel Parzen
Institute of Statistics
Texas A & M University

1. Introduction

Parametric statistical inference may be said to be concerned with statistical inference of idealized parameters from ideal data. Huber (1977), p. 1, writes: "The traditional approach to theoretical statistics was and is to optimize at an idealized parametric model."

Robust statistical inference may be said to be concerned with statistical inference of idealized parameters from semi-ideal data (by the use of methods which are insensitive against small deviations from the ideal assumptions). Huber (1977), p. 3, writes: the robust approach to theoretical statistics assumes "an idealized parametric model, but in addition one would like to make sure that methods work well not only at the model itself, but also in a neighborhood of it."

Exploratory data analysis may be said to be concerned with statistical inference from non-ideal data(often by seeking re-expressions (transformations) of the data that will make it more ideal). Exploratory data analysis helps pose the well-posed statistical questions to which classical parametric statistics provides answers.

---

*Research supported by the Army Research Office (Grant DA AG29-76-0239).

This paper provides an overview to a new general approach to statistical data analysis and parameter estimation which could be called the quantile function approach. The aims of descriptive statistics (to graphically summarize and display the data) are obtained by Quantile-Box plots of the sample quantile function. The aims of "goodness of fit" are obtained by fitting smooth quantile functions to the sample quantile function. The aims of parameter estimation, especially robust estimation of location and scale parameters, are attained by regression analysis of the sample quantile function. (The goal of a statistician in analyzing a batch of data $X_1, \ldots, X_n$ should be both "estimation of parameters" and "goodness of fit". By "goodness of fit" is meant fitting of the observed sample probabilities by a smooth probability law.)

Quantile functions are defined in section 2. Window estimators of location and scale parameters are defined in section 3; their equivalence to L-estimators is discussed in section 4. A conjectured expression is given in section 5 for the asymptotic variance of window estimators. New approaches being developed for non-parametric probability law modeling are mentioned in section 6; quantile box-plots are introduced in section 7. Section 8 discusses location and scale parameter estimation using trimmed samples. Robust regression is the subject of section 9. A new definition of statistics is proposed in section 10.

To carry out in practice robust estimation of location parameters this paper proposes computing means which adapt to the ends (by "ends"

one means the tail character of the distribution of the data). Three such methods are given in the paper:

(1) Iteratively reweighted estimators with weight function $w(x) = (1 + \frac{1}{m} x^2)^{-1}$ for suitable choices of m (section 3);

(2) Maximum likelihood estimation omitting extreme order statistics where the percentage of values omitted is determined from the goodness of fit of the corresponding smooth quantile functions (section 8);

(3) Adaptive L-estimation of location and scale parameters using autoregressive estimators of density-quantile functions (section 8).

A fourth method of robust location and scale parameter estimation is:

(4) Quantile box-plot diagnostics which indicate that mid-summaries and mid-scales are equal enough to provide naive estimators of location and scale (section 7).

2. Quantile Function

The quantile function $Q(u)$, $0 \le u \le 1$ of a random variable X is the inverse of its distribution function $F(x) = P(X \le x)$. The precise definition of Q is:

$$Q(u) = F^{-1}(u) = \inf \{x: F(x) \ge u\} .$$

Given a sample $X_1, \ldots, X_n$, we denote the sample distribution function by $\tilde{F}(x)$, $-\infty < x < \infty$; it is defined by

$$\tilde{F}(x) = \text{fraction of } X_1 \ldots, X_n \le x .$$

The Sample Quantile Function

$$\tilde{Q}(u) = \tilde{F}^{-1}(u) = \inf \{x: \tilde{F}(x) \ge u\}$$

can be computed explicitly in terms of the order statistics $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ (which are the values in the sample arranged in increasing order):

$$\tilde{Q}(u) = X_{(j)}, \quad \frac{j-1}{n} < u \leq \frac{j}{n} .$$

The foregoing definition of $\tilde{Q}(u)$ is a piecewise constant function. It is more convenient to define $\tilde{Q}(u)$ as a piecewise linear function. Divide the unit interval into $2n$ subintervals. For $u = (2j - 1)/2n$ define

$$\tilde{Q}\left(\frac{2j-1}{2n}\right) = X_{(j)}, \quad j = 1, 2, \ldots n .$$

For u in $\frac{2j-1}{2n} \leq u \leq \frac{2j+1}{n}$, $j = 1, 2, \ldots, n-1$,

define $\tilde{Q}(u)$ by linear interpolation; thus for u in this interval

$$\tilde{Q}(u) = n\left(u - \frac{2j-1}{2n}\right) X_{(j+1)} + n\left(\frac{2j+1}{2n} - u\right) X_{(j)}$$

In particular

$$\tilde{Q}\left(\frac{j}{2n}\right) = \frac{1}{2} X_{(j+1)} + \frac{1}{2} X_{(j)}$$

The population median is $Q(0.5)$. The sample median is $\tilde{Q}(0.5)$. Our definition of $\tilde{Q}(u)$ has the merit that $\tilde{Q}(0.5)$ is the usual definition of the sample median:

$$\tilde{Q}(0.5) = X_{(m+1)} \text{ if } n = 2m + 1 \text{ is odd,}$$
$$= \frac{1}{2}\left(X_{(m)} + X_{(m+1)}\right) \text{ if } n = 2m \text{ is even.}$$

The asymtotic distribution of $\tilde{Q}(u)$ satisfies: $\sqrt{n}\, fQ(u)\, \{\tilde{Q}(u) - Q(u)\}$ is asymptotically normal, with mean 0 and variance $u(1 - u)$, where $fQ(u)$ denotes the probability density function $f(x) = F'(x)$ evaluated at $x = Q(u)$; in symbols,

$$fQ(u) = f(Q(u)) \ .$$

We call $fQ(u)$ the density-quantile function.

Estimating the fQ-function is of interest for two reasons: as a way of estimating (1) the true probability density function $f(x)$, and (2) approximate confidence intervals for $Q(u)$ and especially for the true median $Q(0.5)$, since

$$\tilde{Q}(0.5) \pm \{\sqrt{n}\, fQ(0.5)\}^{-1}$$

is an approximate 95% confidence interval for the median $Q(0.5)$.

We call $q(u) = Q'(u)$ the quantile-density function. The identity

$$FQ(u) = u$$

implies the reciprocal relationship

$$fQ(u) \quad q(u) = 1 \ .$$

Thus we may write (using $\doteq$ to denote approximate equality)

$$\frac{1}{fQ(0.5)} = q(0.5) \doteq \frac{Q(0.75) - Q(0.25)}{0.75 - 0.25} = 2\{Q(0.75) - Q(0.25)\}$$

We define, for $0 \le p \le 0.5$,

$$R(p) = Q(1-p) - Q(p)$$

to be the p-range, and

$$\tilde{R}(p) = \tilde{Q}(1-p) - \tilde{Q}(p)$$

to be the sample p-range. When $p = 0.25$, we call $Q(0.75)$ and $Q(0.25)$ the quartiles,

$$R(0.25) = Q(0.75) - Q(0.25)$$

the quartile-range, and

$$\tilde{R}(0.25) = \tilde{Q}(0.75) - \tilde{Q}(0.25)$$

the sample quartile-range.

One can conclude that the median $Q(0.5)$ has a non-parametric estimator given by $\tilde{Q}(0.5)$, and an approximate 95% confidence interval given by

$$\tilde{Q}(0.5) \pm 2\tilde{R}(0.25)/\sqrt{n}$$

A use of a confidence interval of this kind for the median is discussed by McGill, Tukey, and Larsen (1978).

The aim of the foregoing discussion is to introduce the quantile function and illustrate how it is traditionally used to provide non-parametric measures of location (such as the median) and scale (such as the quartile range). Our aim is to use quantile functions to detect and describe ideal and non-ideal statistical models for data.

3. Location and Scale Estimation by Window Estimators

One of the points which this paper would like to make is that measures of location and scale of a data sample are interpretable only if they are probability based, in the sense that they are estimators of characteristics of the true quantile function of the random variable X.

We use μ and σ to denote measures of location and scale respectively. When μ and σ represent median and inter-quartile range, $\mu = Q(0.5)$ and $\sigma = Q(0.75) - Q(0.25)$. When μ and σ^2 represent mean and variance, they can be expressed in terms of Q by

$$\mu = \int_0^1 Q(u)du \quad , \quad \sigma^2 = \int_0^1 \{Q(u) - \mu\}^2 du \quad .$$

These formulas follow immediately from the basic fact that X is identically distributed as $Q(U)$ where U is uniformly distributed on the interval $[0,1]$.

When μ and σ^2 represent mean and variance, fully non-parametric estimators of μ and σ^2 are

$$\tilde{\mu} = \int_0^1 \tilde{Q}(u)\, du \quad , \quad \tilde{\sigma}^2 = \int_0^1 \{\tilde{Q}(u) - \tilde{\mu}\}^2\, du \quad ,$$

which are essentially the sample mean and the sample variance.

To efficiently estimate location and scale parameters μ and σ, it is customary to start with a model for the probability density function $f(x)$ of the form

$$f(x) = \frac{1}{\sigma} f_0\left(\frac{x - \mu}{\sigma}\right) \qquad (*)$$

where $f_0(x)$ is a known probability density function. Define $L(\mu, \sigma)$ to be $(1/n)$ times the log - likelihood of the sample $X_1, \ldots, X_n$; it is given by

$$L(\mu,\sigma) = -\log \sigma + \frac{1}{n} \sum_{i=1}^{n} \log f_0\left(\frac{X_i - \mu}{\sigma}\right)$$

One can express likelihood in terms of quantile functions:

$$L(\mu,\sigma) = -\log \sigma + \int_0^1 \log f_0\left(\frac{\tilde{Q}(u) - \mu}{\sigma}\right) du \quad .$$

The model (*) leads to a very simple formula for the true quantile function $Q(u)$ of the data:

$$Q(u) = \mu + \sigma Q_0(u)$$

where $Q_0(u)$ is a known quantile function corresponding to $f_0(x)$. For ease of writing we introduce the notation

$$\tilde{Q}_0(u) = \frac{\tilde{Q}(u) - \mu}{\sigma}$$

The maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}$ satisfy the log likelihood-derivative equations:

$$\frac{\partial}{\partial \mu} L(\hat{\mu}, \hat{\sigma}) = 0 , \quad \frac{\partial}{\partial \sigma} L(\hat{\mu}, \hat{\sigma}) = 0$$

To compactly write formulas for these derivatives, define

$$\psi(x) = \frac{-f_0'(x)}{f_0(x)} = -\frac{\partial}{\partial(x)} \log f_0(x) .$$

$$w(x) = \frac{1}{x} \psi(x)$$

Then $$\frac{\partial}{\partial \mu} L(\mu,\sigma) = \frac{1}{\sigma}\int_0^1 \psi(\tilde{Q}_0(u))\,du$$

$$= \frac{1}{\sigma^2}\int_0^1 \{\tilde{Q}(u) - \mu\}\, w(\tilde{Q}_0(u))\,du$$

$$\frac{\partial}{\partial \sigma} L(\mu,\sigma) = -\frac{1}{\sigma} + \frac{1}{\sigma^2}\int_0^1 \psi(\tilde{Q}_0(u))\,\{\tilde{Q}(u) - \mu\}du$$

$$= -\frac{1}{\sigma} + \frac{1}{\sigma^3}\int_0^1 w(\tilde{Q}(u))\,\{\tilde{Q}(u) - \mu\}^2\,du \quad .$$

In the normal case, $\psi(x) = x$, $w(x) = 1$ and $\hat{\mu}$ and $\hat{\sigma}^2$ are equal to the sample mean and variance respectively.

To obtain estimators $\hat{\mu}$ and $\hat{\sigma}$ without specifying $f_0(x)$, one introduces the concept of <u>iteratively reweighted estimators</u> of μ and σ^2 . Given estimators μ^* and σ^* define

$$\tilde{Q}_0^{\,*}(u) = \frac{\tilde{Q}(u) - \mu^*}{\sigma^*}$$

Then as "approximate" solutions of the log-likelihood derivative equations, one studies the estimators defined by

$$\hat{\mu} = \frac{\int_0^1 \tilde{Q}(u)\, w(\tilde{Q}_0^{\,*}(u))\,du}{\int_0^1 w(\tilde{Q}_0^{\,*}(u))\,du}$$

$$\hat{\sigma}^2 = \int_0^1 \{\tilde{Q}(u) - \hat{\mu}\}^2\, w(\tilde{Q}_0^{\,*}(u))\,du \quad .$$

These formulas for $\hat{\mu}$ and $\hat{\sigma}$ reduce to the sample mean and variance when one chooses $w(x) \equiv 1$.

Since we are concerned with forming estimators of location and scale which are satisfactory for long-tailed distributions it is natural to choose weight functions $w(x)$ corresponding to Students' t-distribution with m degrees of freedom,

$$f_0(x) = \frac{1}{\sqrt{m\pi}} \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})} \left(1 + \frac{x^2}{m}\right)^{-(m+1)/2} ,$$

for which

$$w(x) = -\frac{1}{x} (\log f_0(x))' = \frac{m+1}{m} \frac{1}{1 + \frac{1}{m} x^2}$$

We call this weight function a window, and we call $\hat{\mu}$ and $\hat{\sigma}$ window estimators.

To completely specify the window, one must specify a value for m (which we could call the "trimming width" of the window). The more normal the data is believed to be, the larger should m be chosen (say, $m = 25$) . The more Cauchy-distributed the data is believed to be, the closer to 1 should m be chosen (say, $m = 4$) . In practice, one might try both values of m , and compare the results. The constant m could also be estimated adaptively to yield "self-tuning" robust estimators of location and scale.

A window recommended by Tukey (see Mosteller and Tukey (1977), p. 205) is the bisquare window:

$$w_{\text{Bisquare}}(x) = (1 - (\frac{x}{c})^2)_+^2$$

where c is a suitably chosen constant. Tukey recommends that c be taken to be 6 or 4 when x is measured in units of σ. It seems likely that the choice of c should reflect one's beliefs about the long-tailed character of the data.

4. Weight Functions of L-Estimators

An L-estimator $\hat{\mu}$ of a location parameter is a linear combination of order statistics $X_{(1)} < \ldots < X_{(n)}$, which we write in the form

$$\hat{\mu} = \int_0^1 \tilde{Q}(u)\, W(u)\, du$$

for suitable weight function $W(u)$. Asymptotically efficient L-estimators of μ and σ in the model $Q(u) = \mu + \sigma Q_0(u)$, when f_0 is a symmetric density, are given by [see Parzen (1978), and summary in section 8]

$$\hat{\mu} = \int_0^1 \tilde{Q}(u)\, W_\mu(u)\, du \div \int_0^1 W_\mu(u)\, du$$

$$\hat{\sigma} = \int_0^1 \tilde{Q}(u)\, W_\sigma(u)\, du \div \int_0^1 W_\sigma(u)\, du$$

where $$W_\mu(u) = f_0Q_0(u)\, J_0'(u) = \frac{J_0'(u)}{Q_0'(u)} \quad \text{and}$$

$$W_\sigma(u) = J_0(u) + Q_0(u)\, W_\mu(u) \ .$$

$f_0Q_0(u)$ is the density-quantile function corresponding to Q_0 , and $J_0(u)$ is its score function defined by

$$J_0(u) = -(f_0Q_0)'(u) = \frac{-f_0'\, Q_0(u)}{f_0Q_0(u)} = \Psi(Q_0(u)) \ .$$

An L-estimator forms a weighted average of order statistics in which the weights depend on the ranks u . It is of interest to express the weights as a function of $Q_0(u)$, which is the size of the order statistics. One can derive such formulas starting from the general representation, given by Parzen (1978),

$$f_0Q_0(u) \sim (1-u)^{\alpha} \ , \quad Q_0(u) \sim (1-u)^{-(\alpha-1)}$$

where α , called the tail exponent, is assumed to satisfy $\alpha > 1$ (indicative of long tailed distributions). We write

$$W_\mu = J_0' f_0 Q_0 = -(f_0 Q_0)^2[(\log f_0 Q_0)'' + \{(\log f_0 Q_0)'\}^2]$$

$$W_\sigma = J_0 + Q_0 W = -(f_0 Q_0)(\log f_0 Q_0)' + Q W_\mu \quad .$$

Therefore

$$W_\mu(u) \sim (1-u)^{2(\alpha-1)}\, \alpha(1-\alpha) \sim \frac{1}{Q_0^2(u)}\, \alpha(1-\alpha)$$

$$W_\sigma(u) \sim (1-u)^{\alpha-1}\, \alpha(2-\alpha) \sim \frac{1}{Q_0(u)}\, \alpha(2-\alpha) \quad .$$

The main conclusion we desire to point out is that if one expresses $W_\mu(u)$ as a function w of $Q_0(u)$,

$$W_\mu(u) = w(Q_0(u)) \quad ,$$

then for long-tailed distributions, $w(x) \sim \frac{1}{x^2}$. By writing W_μ as a function of Q_0 , to an L-estimator one can form an equivalent iteratively reweighted estimator.

Given preliminary estimators μ^* and σ^* , form $\tilde{Q}_0^*(u) = \frac{\tilde{Q}(u) - \mu^*}{\sigma^*}$ and define

$$\hat{\mu} = \frac{\int_0^1 w(\tilde{Q}_0^*(u))\, \tilde{Q}(u)\, du}{\int_0^1 w(\tilde{Q}_0^*(u))\, du}$$

This estimator is a weighted average of $\tilde{Q}$ with weights a function only of the size of the standardized residuals $\tilde{Q}_0^*(u)$.

For Student's t-distribution with m degrees of freedom,

$$J_0(u) = \psi(Q_0(u)) = \frac{m+1}{m} \frac{Q_0(u)}{1 + \frac{1}{m} Q_0^2(u)}$$

Consequently

$$W_\mu(u) = w_\mu(Q_0(u)) , \quad W_\sigma(u) = w_\sigma(Q_0(u))$$

with $$w_\mu(x) = \frac{m+1}{m} \frac{1 - (x^2/m)}{[1 + (x^2/m)]^2} , \quad w_\sigma(x) = \frac{m+1}{m} \frac{2x}{[1 + (x^2/m)]}$$

These windows deserve further investigation. However they appear to support the recommendation that robust estimators of location and scale may be obtained from preliminary estimators μ^* and σ^* by the formulas (for a suitably chosen value of m)

$$\hat{\mu} = \frac{\int_0^1 \tilde{Q}(u) \{1 + \frac{1}{m} (\frac{\tilde{Q}(u) - \mu^*}{\sigma^*})^2\}^{-1} du}{\int_0^1 \{1 + \frac{1}{m} (\frac{\tilde{Q}(u) - \mu^*}{\sigma^*})^2\}^{-1} du}$$

$$\hat{\sigma}^2 = \int_0^1 \{\tilde{Q}(u) - \hat{\mu}\}^2 \{1 + \frac{1}{m} \left(\frac{\tilde{Q}(u) - \mu^*}{\sigma^*}\right)^2\}^{-1} du \left(\frac{m+1}{m}\right)$$

5. Variance and Influence Functions of Window Estimators

This section presents a conjectured formula for the asymptotic variance of a window estimator which is derived by representing it as an L-estimator

$$\hat{\mu} = \frac{\int_0^1 w(Q_0(u))\, \tilde{Q}(u)\, du}{\int_0^1 w(Q_0(u))\, du}$$

where $w(x) = (1 + \frac{1}{m} x^2)^{-1}$ The question of deriving the theory of $\hat{\mu}$ as an M-estimator is open for research; $\hat{\mu}$ is an M-estimator if it satisfies

$$\int_0^1 \psi\left(\frac{\tilde{Q}(u) - \hat{\mu}}{\sigma}\right) du = 0$$

for a suitable ψ function, here chosen to be

$$\psi(x) = \frac{x}{1 + (x^2/m)}$$

Under the assumption that the true quantile function is of the form $Q(u) = \mu + \sigma Q_0(u)$, and that $Q_0(1 - u) = -Q_0(u)$, signifying a symmetric distribution, we seek to find the variance V of the asymptotic distribution of $\sqrt{n}(\hat{\mu} - \mu)$, which is normal with zero mean and asymptotic variance V .

From the asymptotic distribution theory of L-estimators

$$V = \frac{\int_0^1 |V(u)|^2\, du}{\{\int_0^1 w Q_0(u)\, du\}^2}$$

where

$$V'(u) = w(Q_0(u))\, q(u) = \sigma\, w(Q_0(u))\, q_0(u)$$

$$V(u) = \sigma\, v(Q_0(u)) ,$$

defining

$$v(x) = \int^x w(y)\, dy = \sqrt{m}\, \tan^{-1}(x/\sqrt{m})$$

Further $v(x)$ is the influence function of the estimator (Huber (1977), p. 17).

Note that for fixed x, $v(x) \to x$ as $m \to \infty$. The formula for the variance of the robust estimator $\hat{\mu}$ may be written explicitly

$$\mathrm{Var}(\hat{\mu}) = \frac{\sigma^2}{n} \; \frac{\int_0^1 \{\sqrt{m}\, \tan^{-1}\, (Q_0(u)/\sqrt{m})\}^2\, du}{\{\int_0^1 (1 + \frac{1}{m}\, Q_0^{\,2}(u))^{-1}\, du\}^2}$$

and can clearly be regarded as a generalization of the traditional formula for the variance of the sample mean. It is derived under the assumption of a symmetric but possibly long-tailed distribution.

To estimate $\mathrm{Var}(\hat{\mu})$ in practice, one might replace σ by $\hat{\sigma}$ and $Q_0(u)$ by $(\tilde{Q}(u) - \hat{\mu})/\hat{\sigma}$ if a Quantile-Box plot of $\tilde{Q}(u) - \hat{\mu}$ indicates that it is symmetrically distributed about 0.

It should be noted that under the model $Q(u) = \mu + \sigma Q_0(u)$, with $Q_0(1 - u) = -Q_0(u)$, $\hat{\mu}$ estimates $\int_0^1 w(Q_0(u))\, Q(u)\, du \div \int_0^1 w(Q_0(u))\, du = \mu$ while $\hat{\sigma}^2$ estimates $\sigma^2 \int_0^1 w(Q_0(u))\, Q_0^2(u)\, du = \sigma^2 \int_0^1 Q_0^2(u)\, (1 + \frac{1}{m}\, Q_0^2(u))^2\, du$.

6. Non-parametric Probability Law Modeling

To interpret (as well as to form) location and scale parameters estimators from a data batch $X_1, \ldots, X_n$ one must model its probability law. This section briefly mentions some new approaches which are currently being developed for non-parametric probability law modeling (see Parzen (1978)). They all involve both graphical and numerical analysis of the sample quantile function $\tilde{Q}$ to find smoothing functions $\hat{Q}$.

Quantile Box-Plots are introduced in the next section.

Quantile Residual Brownian Bridge Test. To say that the true quantile function $Q(u)$ obeys the hypothesis $H_0 : Q(u) = \mu + \sigma Q_0(u)$ is to say that one can find values $\hat{\mu}$ and $\hat{\sigma}$ such that $\hat{Q}(u) = \hat{\mu} + \hat{\sigma} Q_0(u)$ fits $\tilde{Q}$. The fit of $\hat{Q}$ to $\tilde{Q}$ can be judged by displaying the quantile residuals

$$R(u) = f_0 Q_0(u) \{ \tilde{Q}(u) - \hat{Q}(u) \} , \quad 0 \le u \le 1$$

where $f_0 Q_0(u) = f_0(Q_0(u))$ is the density-quantile function corresponding to F_0. Under the null hypothesis $(\sqrt{n}/\sigma)\ R(u)$, $0 \le u \le 1$ is asymptotically distributed as a stochastic process $B(u)$, $0 \le u \le 1$ which is a modified Brownian Bridge process in the sense that its covariance kernel $E(B(u_1)B(u_2))$ is not $\min(u_1, u_2) - u_1 u_2$ but is modified due to the estimation of the parameters μ and σ. To test whether the sample path $R(u)$ looks like a sample path from a modified Brownian Bridge process one could use various functionals

whose asymptotic distribution is known from their role in the conventional theory of Goodness of Fit Tests. The sample process traditionally considered for goodness of fit tests is

$$\tilde{D}_0(u) = F_0((\tilde{Q}(u) - \hat{\mu})/\hat{\sigma})$$

To estimate σ (needed in the asymptotic distribution of $\hat{R}(u)$) one could use a non-parametric estimator such as

$$\tilde{\sigma}_0 = \int_0^1 f_0 Q_0(u) d\tilde{Q}(u) = \int_0^1 J_0(u)\, \tilde{Q}(u)\, du \quad .$$

To estimate $\hat{\mu}$ and $\hat{\sigma}$ needed to form $\hat{Q}(u)$, one could use quick and dirty estimators $\tilde{\mu}^*$ and $\tilde{\sigma}^*$ formed from Quantile Box-Plots, or one could use asymptotically efficient estimators formed from regression analysis of the continuous process $\tilde{Q}(u)$ (see section 8).

Cumulative Weighted Spacings Brownian Bridge Tests. To test whether the true quantile function $Q(u)$ is of the form $Q(u) = \mu + \sigma Q_0(u)$, one need not first estimate μ and σ. Instead, following Parzen (1978), form

$$\tilde{D}(u) = \frac{1}{\tilde{\sigma}_0} \int_0^u f_0 Q_0(t)\, d\tilde{Q}(t), \quad 0 \le u \le 1 ,$$

which is an estimator of

$$D(u) = \frac{1}{\sigma_0} \int_0^u f_0 Q_0(t)\, dQ(t) \quad 0 \le u \le 1 \quad \text{defining}$$

$\sigma_0 = \int_0^1 f_0 Q_0(u)\, dQ(u)$. Under the null hypothesis, $D(u) = u$, and it is

conjectured that $\sqrt{n}\,\{\tilde{D}(u) - u\}$, $0 \le u \le 1$ is asymptotically distributed as a Brownian Bridge Stochastic process.

By suitably choosing the null hypothesis H_0 and the standard density-quantile function f_0Q_0 , one can test the goodness of fit of any specified probability law (normal, exponential, Weibull, Cauchy, etc.) to the data.

Density-Quantile Function Autoregressive Estimation. Parzen (1978) discusses autoregressive estimators $\hat{d}(u)$ of

$$d(u) = D'(u) = \frac{1}{\sigma_0} \frac{f_0Q_0(u)}{fQ(u)}$$

which can be used to form estimators of $fQ(u)$.

The density quantile function fQ can be estimated also by forming autoregressive smoothers $\hat{D}_0(u)$ of

$$\tilde{D}_0(u) = F_0((\tilde{Q}(u) - \hat{\mu})/\hat{\sigma}),$$

with density

$$\tilde{d}_0(u) = f_0((\tilde{Q}(u) - \hat{\mu})/\hat{\sigma})\,\tilde{q}(u)/\hat{\sigma} .$$

The autoregressive density $\hat{d}_0(u) = \hat{D}_0'(u)$ is an estimator of

$$d_0(u) = \frac{f_0((Q(u) - \mu)/\sigma)}{fQ(u)\,\sigma}$$

7. Quantile Box-Plots Diagnostic Measures

Given a data batch $X_1, \ldots, X_n$, a successful approach to "display" of the data has been the box plot introduced by Tukey (1977). Five values from a set of data are conventionally used: the extremes, the upper and lower H-values (H is an abbreviation for hinges or quartiles), and the M-value (median). The basic configuration of the box-plot display is a vertical box of arbitrary width and length equal to the distance HH (defined as upper H-value minus lower H-value and called the H-spread). A solid line (called the M-line) is marked within the box at a distance MH above the lower end of the box (MH equals M minus lower H). Dashed lines are extended from the lower and upper ends of the box a distance equal to the distance of the extremes from the hinges. If one wants to indicate a confidence interval for the median, one might add a line perpendicular to the M-line at its midpoint, and of length $\pm$ HH/$\sqrt{n}$. The box-plot described should be called an H-Box Plot, because by replacing H-values by other types of values (called E-values and D-values) one can consider E-Box Plots and D-Box Plots.

The H-values are most conviently defined as $\tilde{Q}(0.25)$ and $\tilde{Q}(0.75)$, the 1/4 percentiles. The E-values are the 1/8 percentiles $\tilde{Q}(0.125)$ and $\tilde{Q}(0.875)$. The D-values are the 1/16 percentiles $\tilde{Q}(0.0625)$ and $\tilde{Q}(0.9375)$.

The mid-summaries of a data batch are

$$\tilde{\mu}(p) = \frac{1}{2}\{\tilde{Q}(1-p) + \tilde{Q}(p)\}, \qquad 0 \le p \le 0.5 .$$

Of particular interest are

$$\tilde{\mu}_M = \tilde{\mu}(0.5) \ , \quad \tilde{\mu}_H = \tilde{\mu}(0.25) \ , \quad \tilde{\mu}_E = \tilde{\mu}(0.125) \ ,$$

$$\tilde{\mu}_D = \tilde{\mu}(0.0625) \ .$$

When H_0: $Q(u) = \mu + \sigma\, Q_0(u)$ holds, and $Q_0(1-u) \equiv -Q_0(u)$, $\tilde{\mu}$ is an approximately unbiased estimator of μ .

The average of the extreme-values of the sample will be denoted $\tilde{\mu}(0)$. The closeness of $\tilde{\mu}(0)$ to the other $\tilde{\mu}$ values may indicate whether the data batch has a short tailed symmetric distribution such as the uniform.

The mid-spreads of a data batch are

$$\tilde{S}(p) = \tilde{Q}(1-p) - \tilde{Q}(p) \ , \qquad 0 \le p \le 0.5 \ .$$

Given a specified standardized quantile function Q_0, the mid-scales are defined by

$$\tilde{\sigma}(p) = \tilde{S}(p) \div S_0(p)$$

where $S_0(p) = Q_0(1-p) - Q_0(p)$ is the mid-spread of Q_0 . When H_0 holds, $\tilde{\sigma}(p)$ is an approximately unbiased estimator of σ . Of particular interest are

$$\tilde{\sigma}_H = \tilde{\sigma}(0.25) \ , \quad \tilde{\sigma}_E = \tilde{\sigma}(0.125) \ , \quad \tilde{\sigma}_D = \tilde{\sigma}(0.0625) \ .$$

Quick and dirty estimators of μ and σ are given by

$$\tilde{\mu}^* = \frac{1}{4}\{\tilde{\mu}_M + \tilde{\mu}_H + \tilde{\mu}_E + \tilde{\mu}_D\}$$

$$\tilde{\sigma}^* = \frac{1}{5}\{\tilde{\sigma}_H + 2\tilde{\sigma}_E + 2\tilde{\sigma}_D\}$$

Diagnostic tests for the validity of H_0 are obtained by testing for the equality of the various $\tilde{\mu}$ and $\tilde{\sigma}$ values. More quantitative diagnostic measures could be defined as follows:

$$\text{SKEW}(p) = \{\tilde{\mu}_M - \tilde{\mu}(p)\} \div \tilde{S}(p)$$

$$\text{TAIL}(p) = \log\{\tilde{S}(p) \div \tilde{S}(0.25)\}$$

$$\text{TAIL}_0(p) = \log\{S_0(p) \div S_0(0.25)\}$$

$$\text{TAIL}_\Phi(p) = \log\{\Phi^{-1}(p) \div \Phi^{-1}(0.25)\}$$

When SKEW(p) is not significantly different from zero, we consider the data batch to have a symmetric distribution.

When the data passes a SKEW test for symmetry, it is checked for normality by comparing TAIL(p) with $\text{TAIL}_\Phi(p)$; TAIL(p) significantly larger than $\text{TAIL}_\Phi(p)$ indicates a long-tail distribution, and TAIL(p) significantly shorter than $\text{TAIL}_\Phi(p)$ indicates either a short-tailed distribution (especially a uniform) or possibly a bimodal distribution.

A seven-number summary of a data batch is provided by its M, H, E and D values, which suffice to compute mid-summaries, mid-scales, SKEW, and TAIL measures. To find re-expressions (transformations) of the data which make it more normal, one needs only the seven-number summary of the re-expressed data batch which are easily found as re-expressions of the seven-number summary of the original data batch.

In addition to the analytical measures of the data, one should form a graphical display of the quantile function $\tilde{Q}(u)$ as a function on the unit interval $0 \leq u \leq 1$; the H, E, and D boxes are drawn superimposed.

Quantile-Box Plots enable the investigator to detect "non-ideal" aspects of data batches by testing the data for normality by tests which determine the directions in which data fails to be normal, such as (1) long-tailed distribution, (2) outliers, (3) bimodal distribution, (4) non-symmetric distribution.

To check for symmetry, inspect the shape of $\tilde{Q}(u)$ within the boxes, as well as compare mid-summaries and examine the SKEW diagnostic measures.

When the data passes the test for symmetry the question of whether it has a normal or long-tailed distribution is decided using the TAIL diagnostic measures. Small TAIL values may indicate bimodal distributions. Data sets with outliers may also yield small TAIL values.

If the graph $x = Q(u)$ has points with sharp rises ("infinite" slopes), then the probability density has a zero and will therefore have two (or more)

modes. If the points of sharp rise lie inside the H-Box we suspect the presence of several distinct populations generating the single data batch. If points of sharp rise lie outside the E-Box, we suspect outliers (values to be discarded for robust estimation).

A mode in the probability density function is indicated in the graph $x = \tilde{Q}(u)$ by a point of inflection (with "finite" slope). A horizontal segment in the graph is interpreted to mean a very large probability density there.

8. Location and Scale Parameter Estimation as Regression Analysis of Sample Quantile Process

One can consider estimators, denoted $\mu_{p,q}$ and $\sigma_{p,q}$, which use the sample quantile function $\tilde{Q}(u)$, $p \le u \le q$; this is equivalent to using a restricted set of order statistics $X_{(np)}, \ldots, X_{(nq)}$ or a trimmed sample. A compact derivation of such formulas is given by Parzen (1978) who gives the representation

$$\begin{bmatrix} \hat{\mu}_{p,q} \\ \hat{\sigma}_{p,q} \end{bmatrix} = \begin{bmatrix} I_{\mu\mu} & I_{\mu\sigma} \\ I_{\mu\sigma} & I_{\sigma\sigma} \end{bmatrix}^{-1} \begin{bmatrix} T_{\mu,p,q} \\ T_{\sigma,p,q} \end{bmatrix}$$

where

$$T_{\mu,p,q} = \int_p^q W_\mu(u)\,\tilde{Q}(u)\,du + \tilde{Q}(p)\,W_{\mu L}(p) + \tilde{Q}(q)\,W_{\mu R}(q)$$

$$T_{\sigma,p,q} = \int_p^q W_\sigma(u)\,\tilde{Q}(u)\,du + \tilde{Q}(p)\,W_{\sigma L}(p) + \tilde{Q}(q)\,W_{\sigma R}(q)$$

$$I_{\mu\mu} = \int_p^q W_\mu(u)\,du + W_{\mu L}(p) + W_{\mu R}(q)$$

$$I_{\mu\sigma} = \int_p^q W_\sigma(u)\,du + W_{\sigma L}(p) + W_{\sigma R}(q)$$

$$I_{\sigma\sigma} = \int_p^q W_\sigma(u)\,Q_0(u)\,du + W_{\sigma L}(p)\,Q_0(p) + W_{\sigma R}(q)\,Q_0(q)$$

The weight functions are expressed in terms of the density-quantile function $f_0Q_0(u) = f_0(Q_0(u))$ and the score function

$$J_0(u) = -(f_0Q_0)'(u) = \frac{-f_0'(F_0^{-1}(u))}{f_0(F_0^{-1}(u))} = \psi(Q_0(u)) .$$

$$W_{\mu}(u) = J_0'(u)\, f_0 Q_0(u)$$

$$W_{\sigma}(u) = J_0(u) + Q_0(u)\, W_{\mu}(u)$$

$$W_{\mu L}(p) = f_0 Q_0(p)\left[\frac{1}{p} f_0 Q_0(p) + J_0(p)\right]$$

$$W_{\mu R}(q) = f_0 Q_0(q)\left[\frac{1}{1-q} f_0 Q_0(p) - J_0(p)\right]$$

$$W_{\sigma L}(p) = Q_0(p)\, W_{\mu L}(p) - f_0 Q_0(p)$$

$$W_{\sigma R}(q) = Q_0(q)\, W_{\mu R}(p) + f_0 Q_0(q)$$

For normally distributed data,

$$f_0(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)x^2}, \quad F_0(x) = \Phi(x) = \int_{-\infty}^{x} \phi(y)\, dy ,$$

$$f_0 Q_0(u) = \phi\Phi^{-1}(u) = \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}\left|\Phi^{-1}(u)\right|^2 ,$$

$$J_0(u) = \Phi^{-1}(u) , \quad J_0'(u) = \{\phi\Phi^{-1}(u)\}^{-1} ,$$

$$W_{\mu}(u) = 1 , \quad W_{\sigma}(u) = 2\,\Phi^{-1}(u) .$$

$$W_{\mu L}(p) = \phi\Phi^{-1}(p)\left\{\frac{1}{p}\phi\Phi^{-1}(p) + \Phi^{-1}(p)\right\}$$

$$W_{\sigma R}(p) = \Phi^{-1}(p)\, W_{\mu L}(p) - \phi\Phi^{-1}(p) \quad .$$

When $q = 1 - p$, $I_{\mu\sigma} = 0$,

$$I_{\mu\mu} = 1 - 2p + 2W_{\mu L}(p)$$

$$I_{\sigma\sigma} = 2\int_p^q |\Phi^{-1}(u)|^2\, du + 2\Phi^{-1}(p)\, W_{\sigma L}(p)$$

The estimator

$$\hat{\mu}_{p,q} = \frac{\int_p^{1-p} \tilde{Q}(u)\, du + W_{\mu L}(p)\,\{\tilde{Q}(p) + \tilde{Q}(1-p)\}}{1 - 2p + 2W_{\mu L}(p)}$$

is similar to the Winsorized mean (with trimming proportion p).

Robust Maximum Likelihood Estimation of Mean and Variance of a Normal Distribution. We may be willing to assume that our data is more normal than longtailed, but the shape of the true distributions is deviating slightly from the assumed normal model due to "wrong" values in the data set. We propose the following exploratory data analysis for robust estimation of μ and σ from normal data with possible "outliers." We suggest the name "robust maximum likelihood estimators" for these estimators.

For selected values of p (at least $p = 0.05$, 0.25, and 0.45) , (1) compute the estimators $\hat{\mu}_{p,1-p}$ and $\hat{\sigma}_{p,1-p}$, and (2) plot the residuals.

$$\tilde{Q}(u) - \hat{Q}(u) = \tilde{Q}(u) - \hat{\mu}_{p,1-p} - \hat{\sigma}_{p,1-p}\Phi^{-1}(u) \quad ,$$

multiplied by $\phi\Phi^{-1}(u)$. Their values over the interval $p \le u \le 1 - p$ can be used to test the hypothesis H_0 . The residuals over the tail intervals $u \le p$ and $u \ge 1 - p$ can be used to test for the presence of "wrong values." One estimates μ and σ by those estimators corresponding to the lowest value of p for which one finds no "wrong values" over the tail intervals.

9. Robust Regression

Formulating the estimation of location and scale parameters as a problem of weighted regression of the sample quantile function, with weights a function of $Q_0(u)$, leads to the amazing conclusion that asymptotically efficient estimators of μ and σ are obtainable numerically by iterating ordinary regression calculations!

The iteratively reweighted estimators $\hat{\mu}$ and $\hat{\sigma}$ are also the solutions to the problem of estimating μ and σ in the following weighted least squares linear regression problem:

$$X_j = \mu + \epsilon_j$$

where ϵ_j are independent normal with mean 0 and variance satisfying

$$\mathrm{Var}(\epsilon_j) = \sigma^2 \frac{1}{w_j} \quad , \quad w_j = w(\epsilon_j^*) \quad , \quad \epsilon_j^* = \frac{X_j - \mu^*}{\sigma^*} \qquad (*)$$

A regression model can be written

$$Y_j = \beta_1 X_{1j} + \ldots + \beta_k X_{kj} + \epsilon_j \quad ,$$

where $\{\epsilon_j\}$ are independent random variables with quantile function known up to a parameter σ

$$Q_\epsilon(u) = \sigma Q_0(u)$$

To robustly estimate the coefficients $\beta_1, \ldots, \beta_k$, σ assume first $Q_0(u) = \Phi^{-1}(u)$, corresponding to normality, and by ordinary least squares linear regression obtain preliminary estimators $\beta_1^*, \ldots, \beta_k^*$; then form residuals

$$\epsilon_j^* = (Y_j - \beta_1^* X_{1j} + \ldots + \beta_k^* X_{kj}) \div \sigma_j^* \quad .$$

The next stage of estimators $\hat{\beta}_1, \ldots, \hat{\beta}_k$, $\hat{\sigma}^2$ are taken to be the least squares linear regression estimators under the assumption that ϵ_j have variances defined by (*) . This process is iterated to yield robust estimators (compare Huber (1977), p. 38, Algorithm W).

The long-tailed character of the residuals ϵ_j^* should also be examined, using Quantile-Box plots.

One might consider non-parametric non-linear regression of Y on $X_1, \ldots X_k$. A density-quantile approach to non-linear non-

parametric regression has been described by Parzen (1977), and is currently being investigated by Prof. J. P. Carmichael. It has been applied to time series analysis by Prof. M. Pagano. It provides means of checking whether robust estimation of variances and correlations is provided by robust estimation of linear regression coefficients.

10. Do we need a new definition of Statistics?

Can statistics be made a subject that provides intellectually exciting pastimes (for the young and the mature), is regarded as relevant by the creative scientist, and is appealing as a career to the mathematically talented?

An important step in achieving these desirable (and I believe attainable) goals is to alter the perception of the sample mean and the sample variance in elementary statistical instruction. Introductory Statistics is regarded by almost all college students (even by mathematically talented students) as a very dull subject. Perhaps one reason is that students enter the course knowing about a mean and a variance and leave the course knowing only about a mean and a variance. That statistics is in fact a live and vibrant discipline can be communicated to the student by emphasizing that there are many ways to estimate mean and variance, and more generally location and scale parameters.

I believe that the discipline of statistics can be made more "glamorous" if intellectually sound and demonstratively useful concepts of statistical data analysis and robust statistical inference are incorporated in introductory statistical instruction. The perspective which this paper proposes for interpreting robust statistical inference is equivalent to a proposal for the definition of statistics:

"Statistics is arithmetic done by the method of Lebesgue integration."

I realize this definition sounds unbelievable and may never sell to the introductory student. But at least statisticians should understand to what extent it is true. Perhaps it provides a basis for a new sect of statisticians.

Can we all agree that a basic problem of statistics is an arithmetical one: find the average $\bar{X}$ of a set of numbers $X_1, \ldots, X_n$? Even grade school students (in the U. S. A.) nowadays know the answer:

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) \quad .$$

In words: list the numbers, add them up, and divide by n . What should be realized is that the foregoing algorithm is the method of Riemann integration.

The method of Lebesgue integration finds $\bar{X}$ by first finding the distribution function $\tilde{F}(x)$ of the data, defined by $\tilde{F}(x)$ = fraction of $X_1, \ldots, X_n \leq x$, $-\infty < x < \infty$. Then $\bar{X}$ is found as the mean of this distribution function, defined by the integral

$$\bar{X} = \int_{-\infty}^{\infty} x \, d\tilde{F}(x) \quad .$$

To use an analogy to count a sack of coins, first arrange the coins in piles according to their value (pennies, nickels, dimes, quarters, and half-dollars), then count the number of coins in each pile, determine the value of each pile, and finally obtain $\bar{X}$ as the sum of the values of the piles, divided by n. The role of statistics is to find more accurate estimators of the true mean by fitting a smooth distribution function $\hat{F}(x)$ to $\tilde{F}(x)$.

Still more insight (and fidelity to the truth) is obtained by displaying the sample quantile function $\tilde{Q}(u) = \tilde{F}^{-1}(u) = \inf \{x: \tilde{F}(x) \geq u\}$, and fitting smooth quantile functions $\hat{Q}(u)$ to $\tilde{Q}(u)$. Then one computes the "sample average" by

$$\hat{\mu} = \int_0^1 \hat{Q}(u) \, du = \int_0^1 W_\mu(u) \, \tilde{Q}(u) \, du \div \int_0^1 W_\mu(u) \, du \quad .$$

In words, the "average" of a sample is a weighted average of the numbers in the sample arranged in increasing order, with the weight of a number depending on its rank. This is the essence of robust statistical data analysis; all the rest is commentary.

References

Huber, P. [1977]. Robust Statistical Procedures. Regional Conference Series in Applied Mathematics 27. SIAM: Philadelphia.

McGill, R., Tukey, J. W., Larsen, W. A. [1978]. "Variations of Box Plots," American Statistician, 32, 12-16.

Mosteller, F. and Tukey, J. W. [1977]. Data Analysis and Regression, Addison Wesley: Reading, Mass.

Parzen, E. [1977]. "Nonparametric Statistical Data Science (A Unified Approach Based on Density Estimation and Testing for "White Noise")." Technical Report No. 47, Statistical Science Division, State University of New York at Buffalo.

Parzen, E. [1978]. "Nonparametric Statistical Data Modeling," Journal of the American Statistical Association.

Tukey, J. W. [1977]. Exploratory Data Analysis. Addison Wesley: Reading, Mass.

Appendix

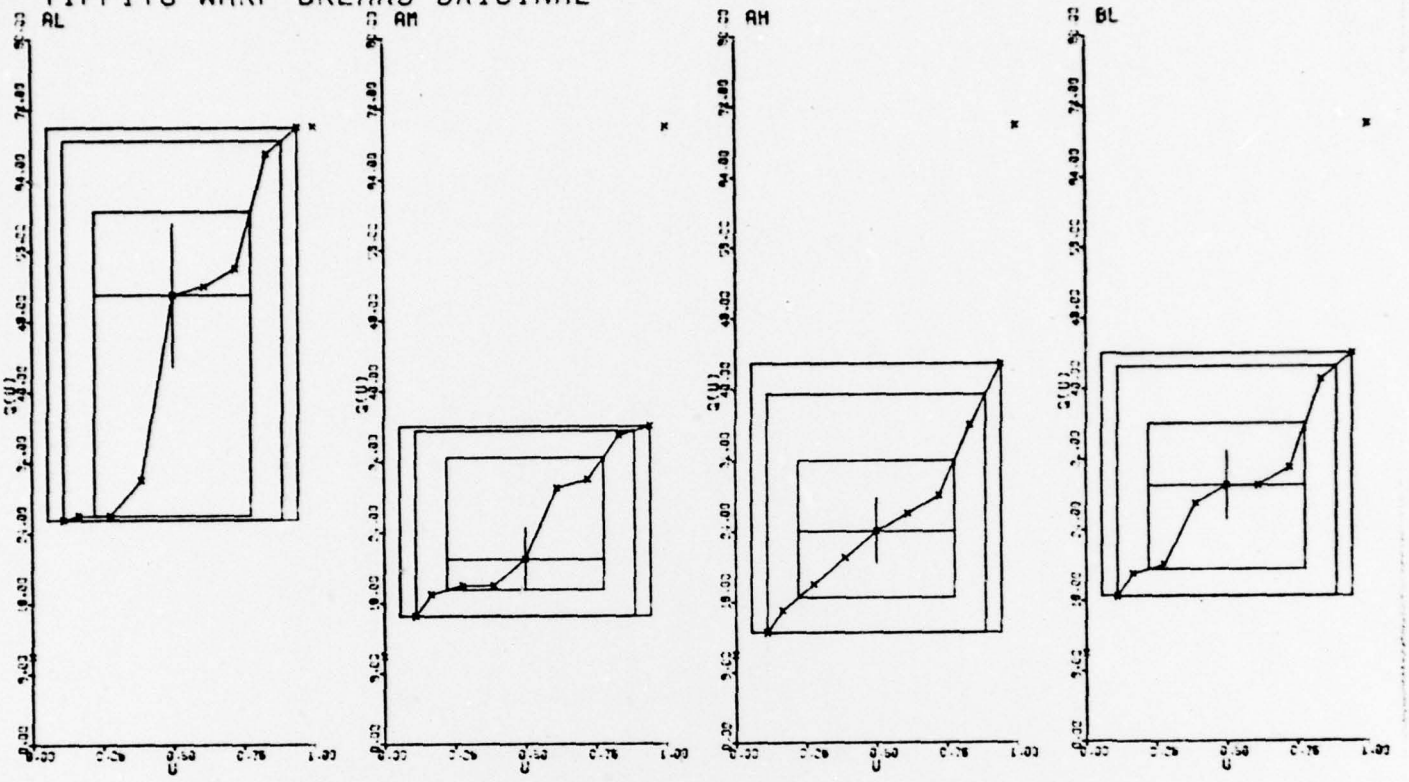
EXAMPLES OF QUANTILE-BOX PLOTS

TIPPETT'S WARP BREAK DATA (Compare box plots in McGill, Tukey, Larsen [1978]).

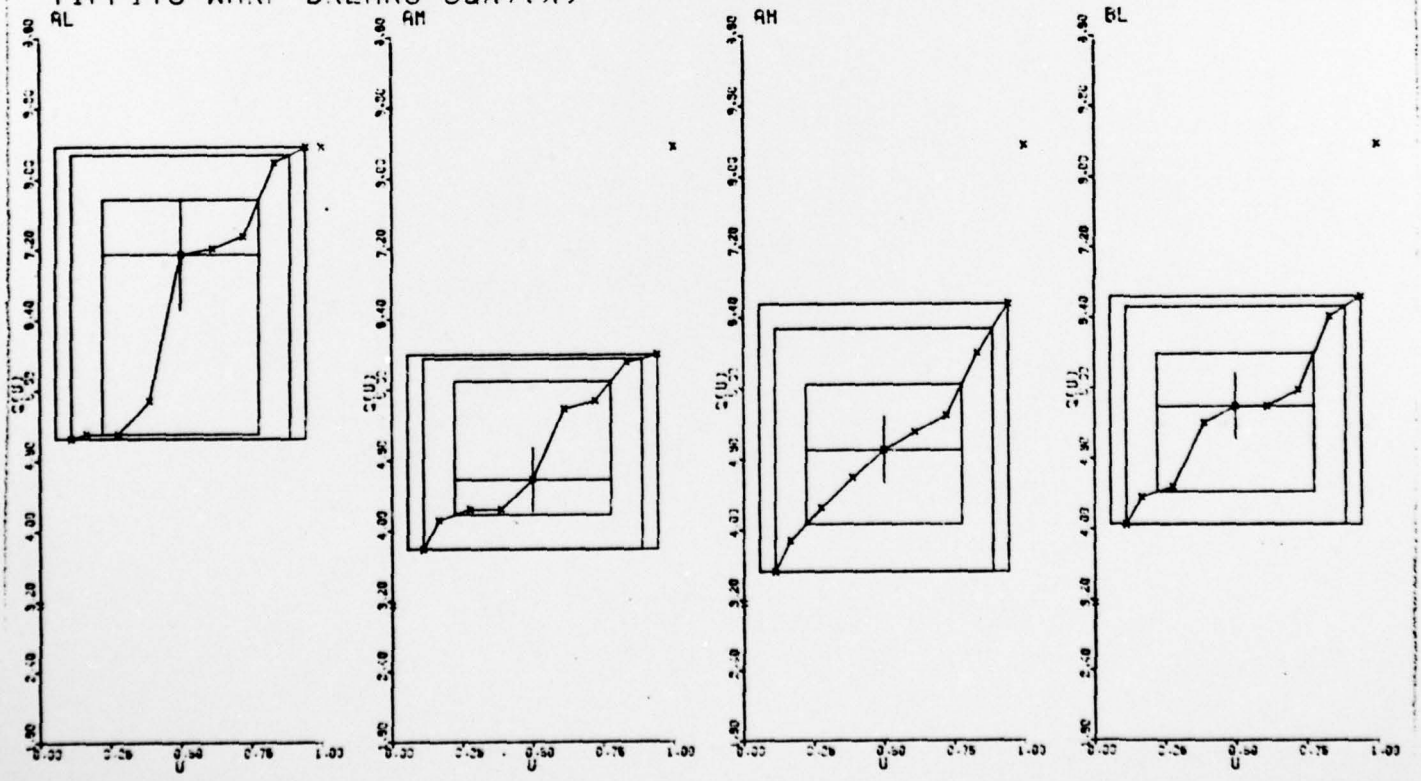
Fossil data from yellow Limestone formation of northwestern Jamaica (from Chernoff, H. (1973), "The Use of Faces to Represent Points in k-Dimensional Space Graphically," Journal of the American Statistical Association, 68, 361-368).

Variables 2 and 6 have zeroes in fQ (rises in Q). Variables 3 and 4 have proability masses (flat stetches in Q). Variables 1 and 5 are candidates for re-expression (logarithm for 1, square root for 5). Variables 2 and 5 suffice to classify the observations.

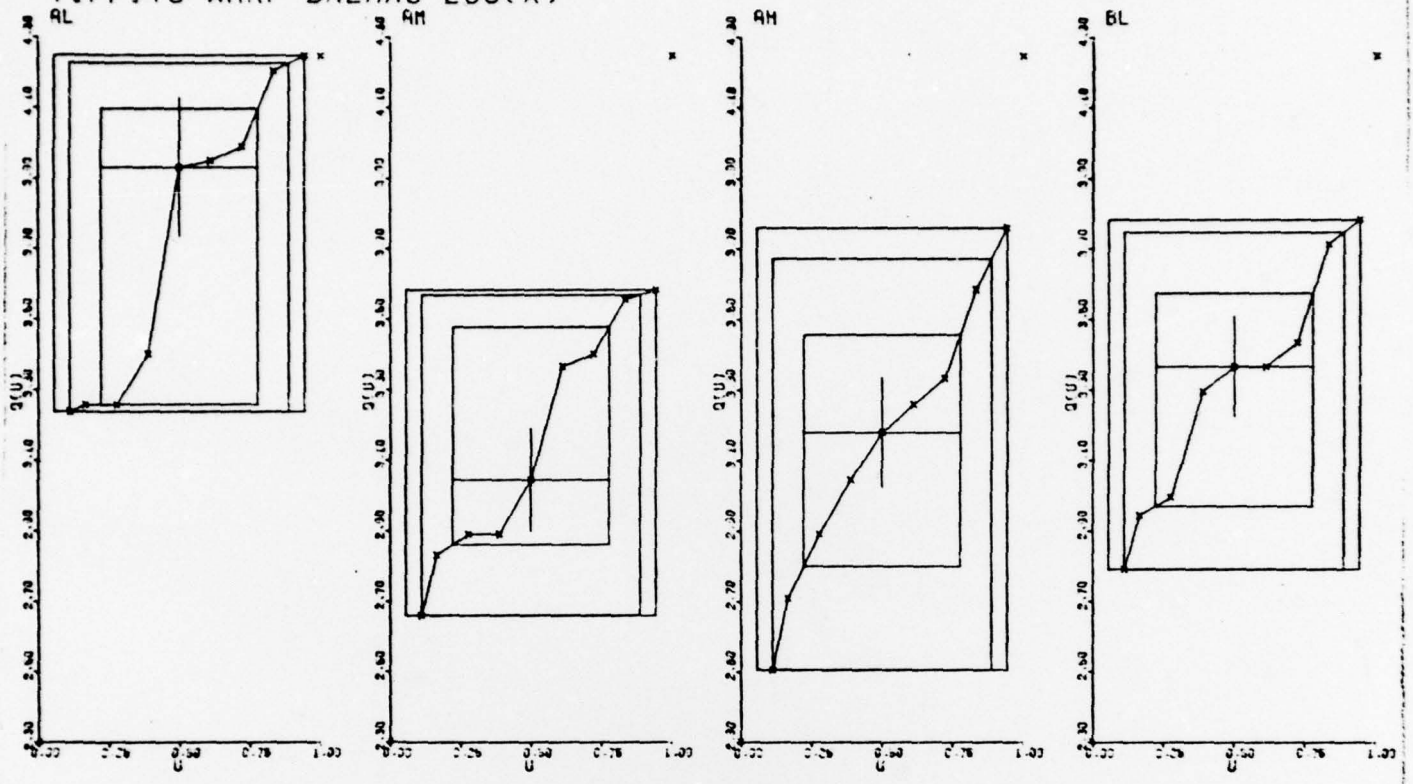
TIPPITS WARP BREAKS ORIGINAL



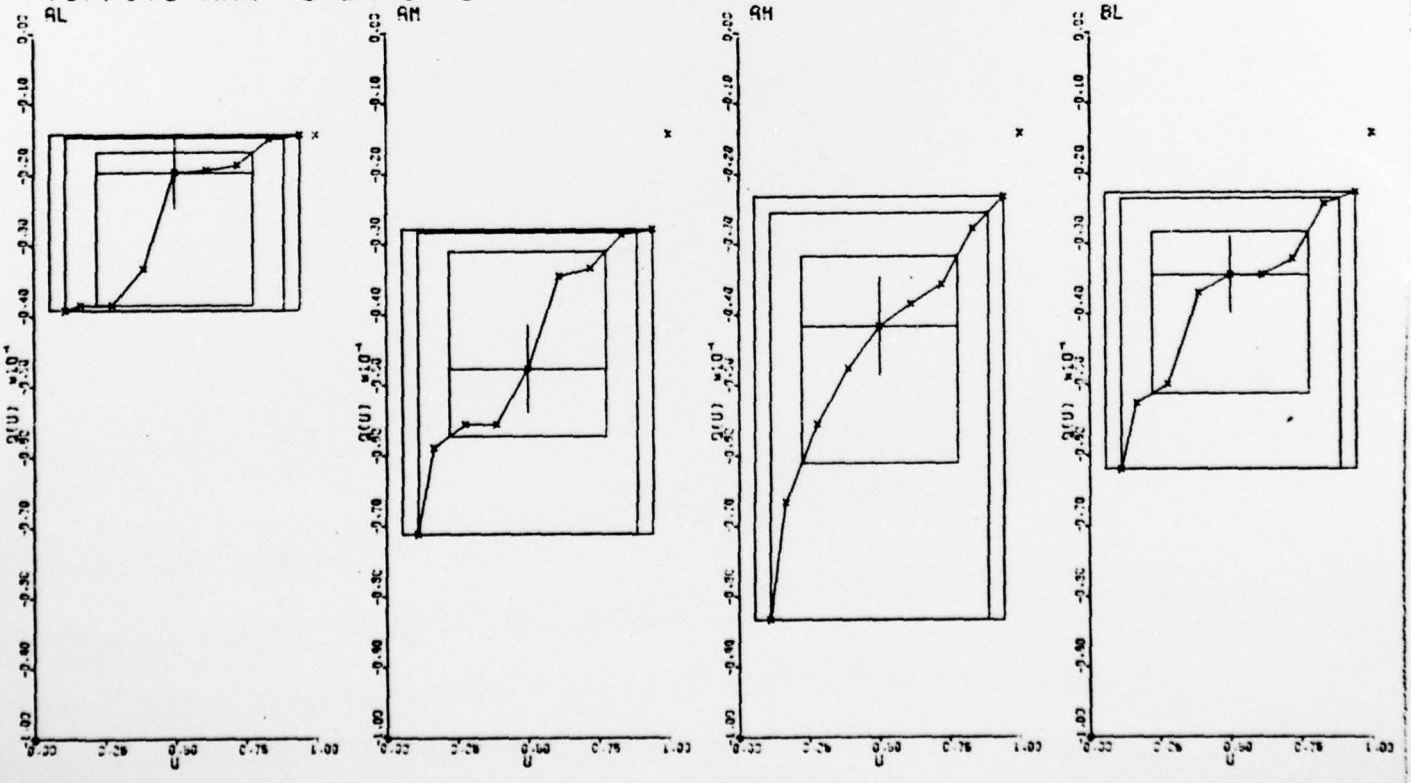
TIPPITS WARP BREAKS SQRT(X)



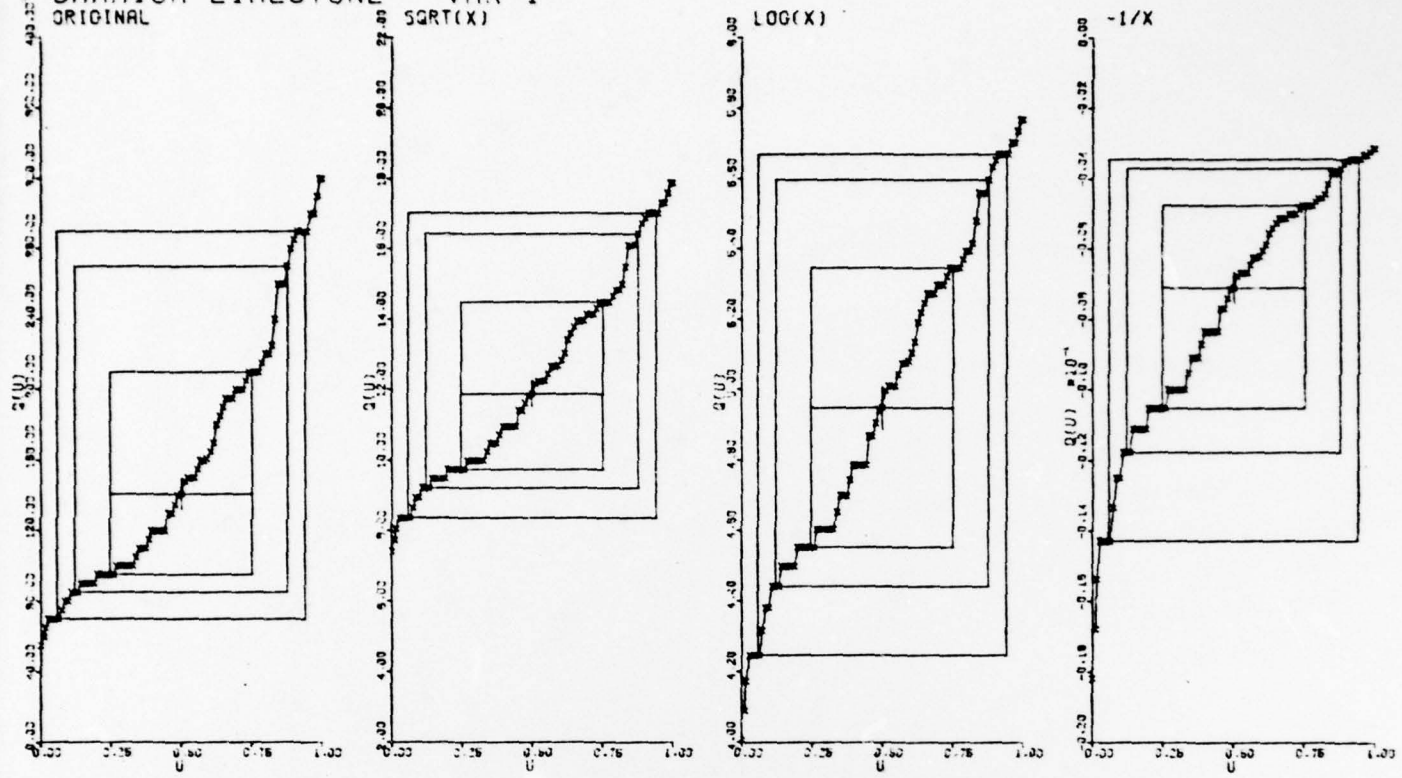
TIPPITS WARP BREAKS LOG(X)



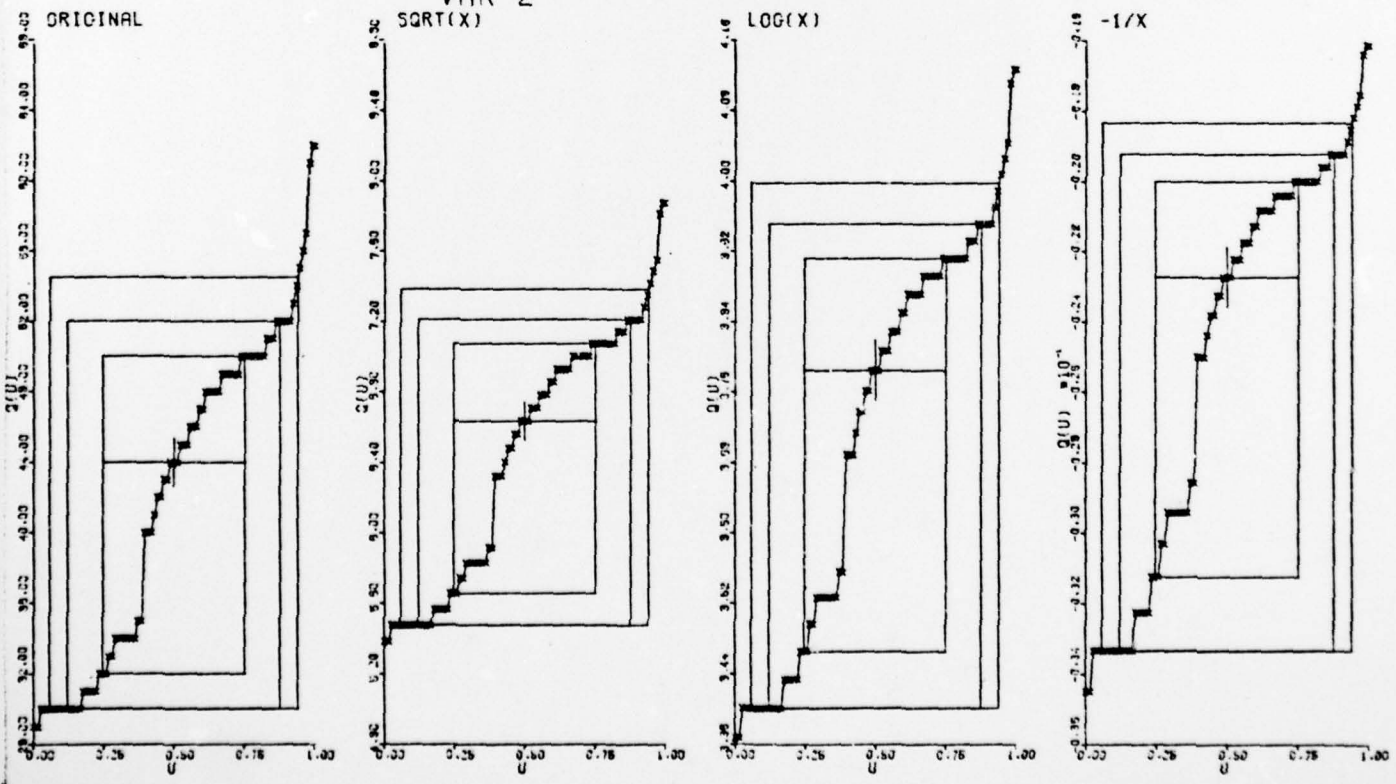
TIPPITS WARP BREAKS -1/X



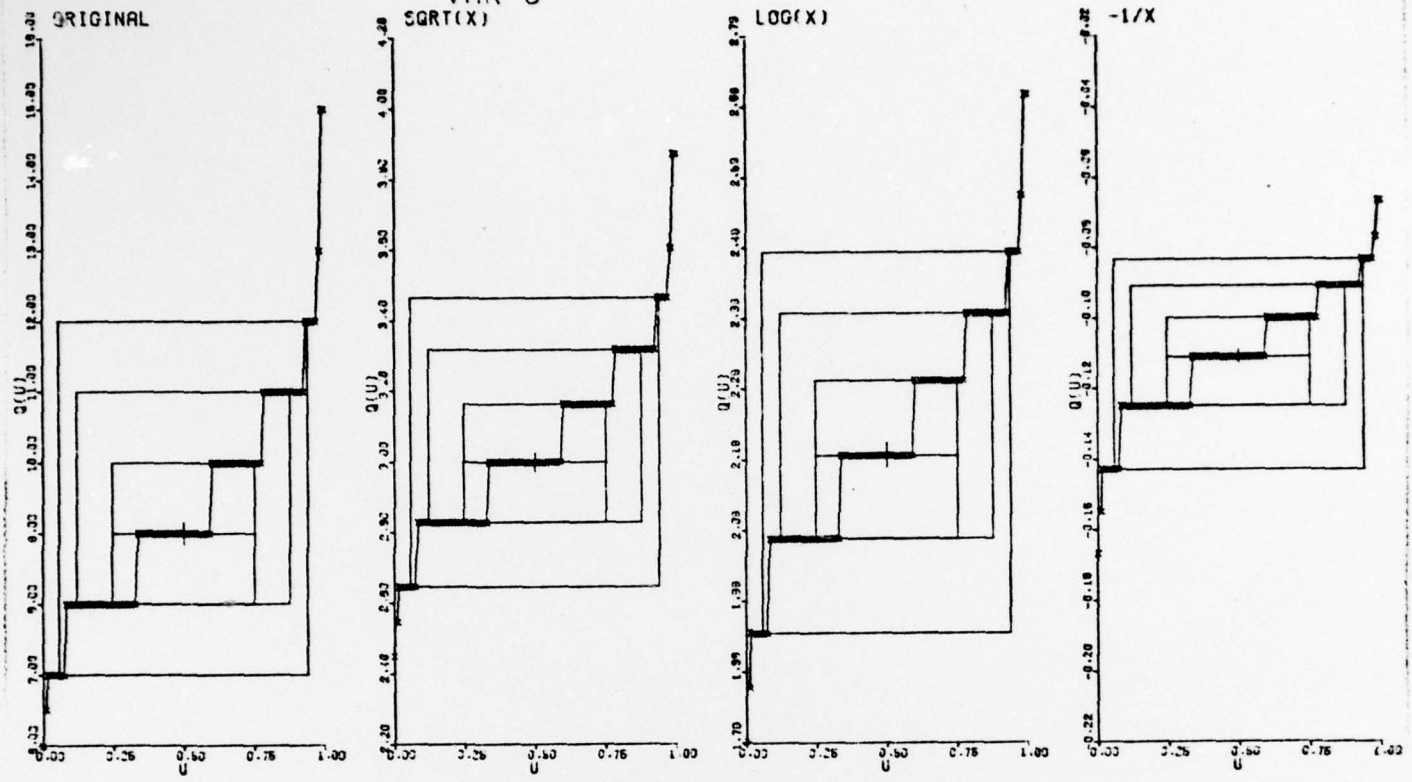
JAMAICA LIMESTONE VAR 1



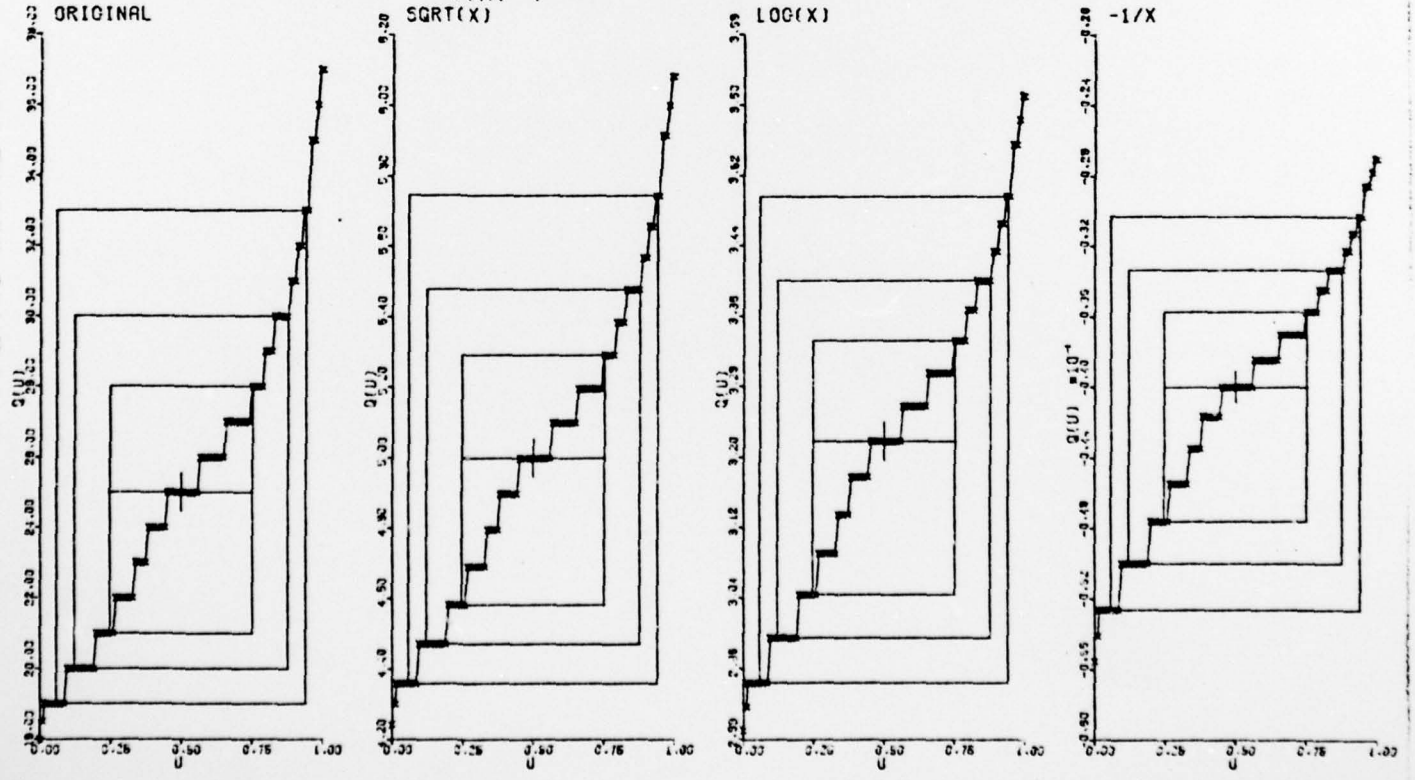
VAR 2

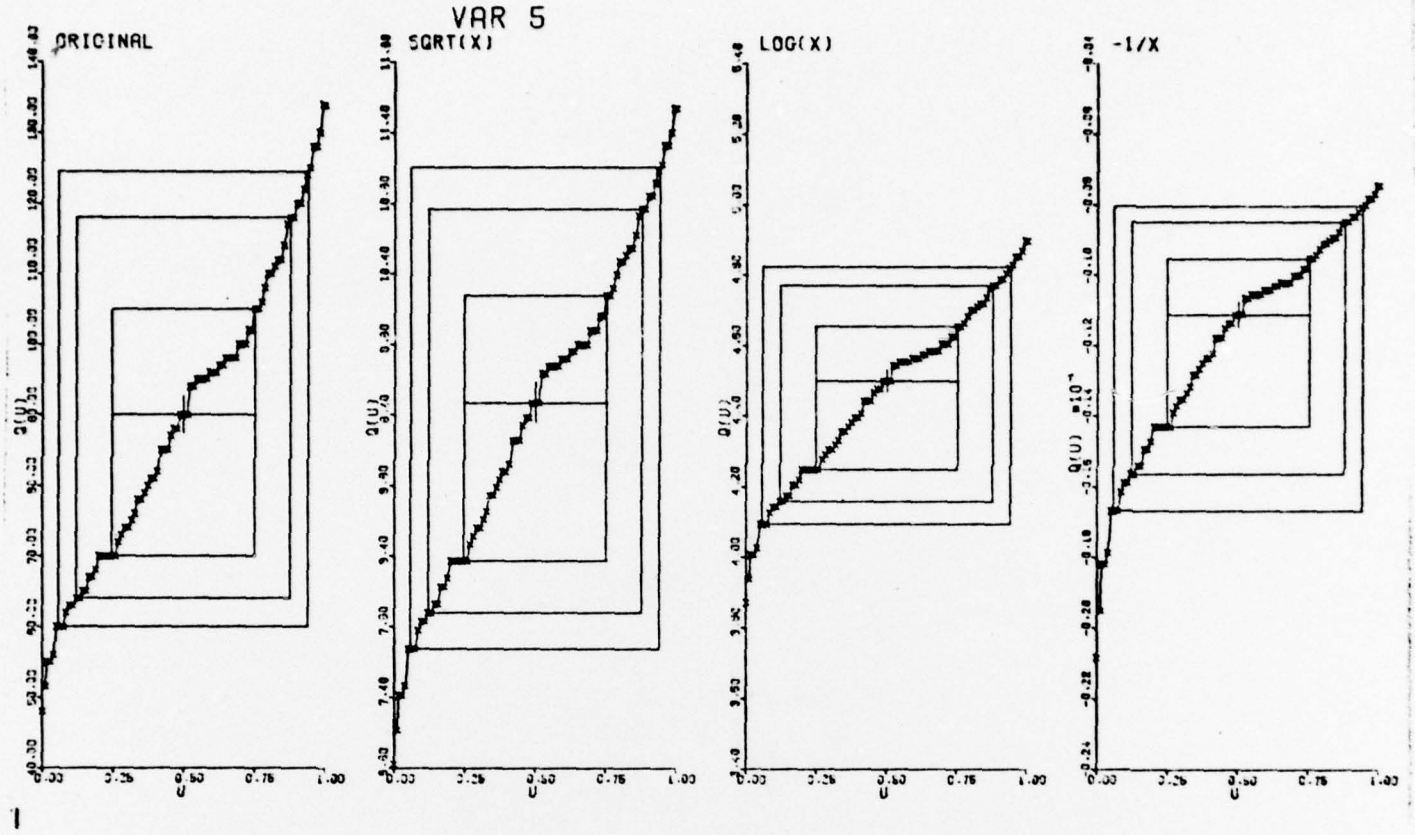


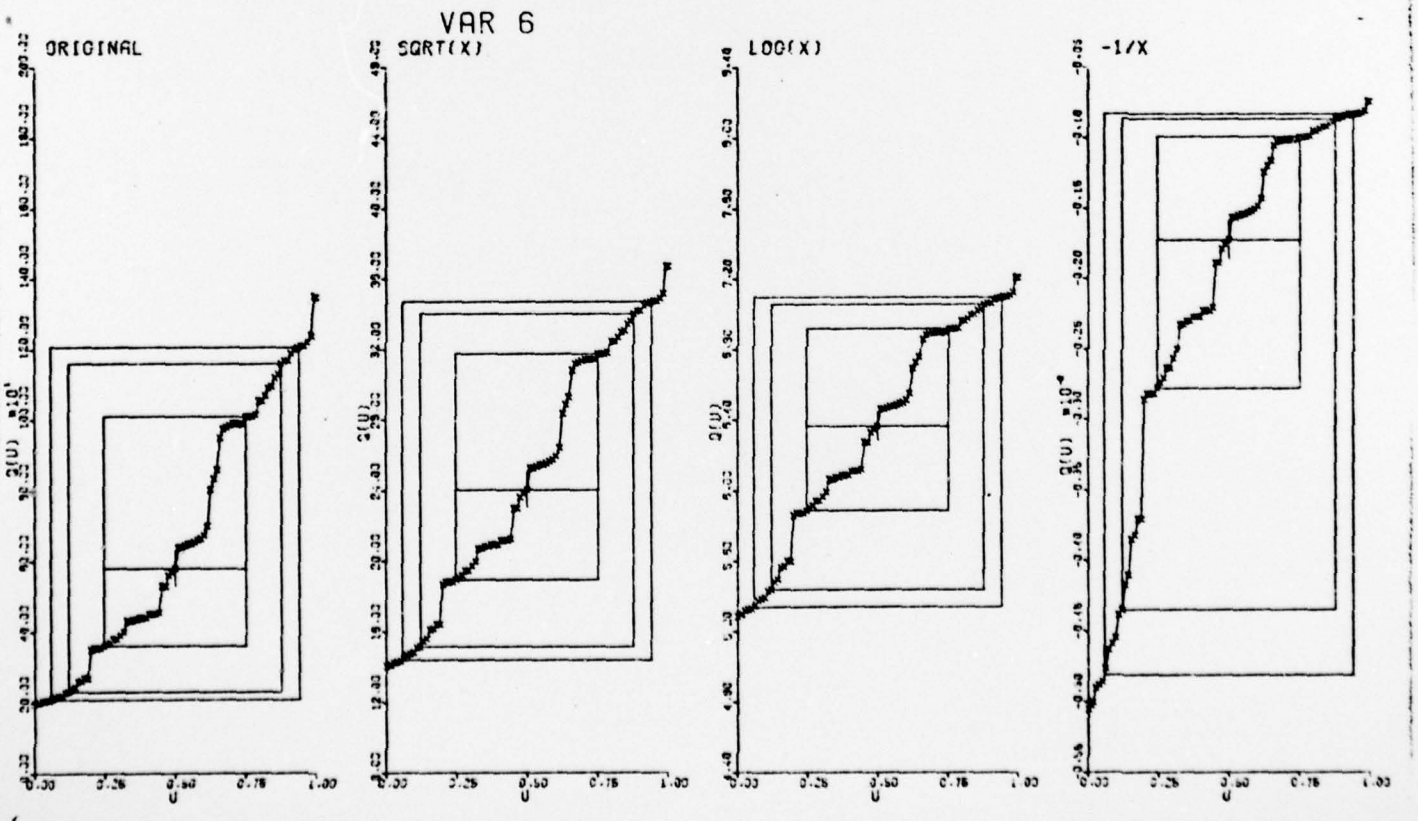
VAR 3



VAR 4




VAR 5
ORIGINAL
SQRT(X)
LOG(X)
-1/X
Q(U)
U


VAR 6
ORIGINAL
SQRT(X)
LOG(X)
-1/X
Q(U)
U

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ~~Technical Report No. ARO-4~~	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Density-Quantile Function Perspective on Robust Estimation		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Emanuel Parzen		8. CONTRACT OR GRANT NUMBER(s) DA AG29-76-G-0239
9. PERFORMING ORGANIZATION NAME AND ADDRESS Statistical Science Division State University of New York at Buffalo Amherst, New York 14226		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE March 1978
		13. NUMBER OF PAGES 35
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)

NA

18. SUPPLEMENTARY NOTES

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

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

Robust Statistics
Exploratory Data Analysis
Parameter Estimation
Quantile Functions
Box Plots
Robust Regression

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

This paper provides an overview to a new general approach to statistical data analysis and parameter estimation which could be called the quantile function approach. The aims of descriptive statistics (to graphically summarize and display the data) are obtained by Quantile-Box plots of the sample quantile function. The aims of "goodness of fit" are obtained by

(continued on page 2)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

fitting smooth quantile functions to the sample quantile function. The aims of parameter estimation, especially robust estimation of location and scale parameters, are attained by regression analysis of the sample quantile function. (The goal of a statistician in analyzing a batch of data $X_1, \ldots, X_n$ should be both "estimation of parameters" and "goodness of fit." By "goodness of fit" is meant fitting of the observed sample probabilities by a *smooth* probability law.)

Quantile functions are defined in Section 2. Window estimators of location and scale parameter estimation are defined in Section 3; their equivalence to L-estimators is discussed in Section 4. A conjectured expression is given in Section 5 for the asymptotic variance of window estimators. New approaches being developed for non-parametric probability law modeling are mentioned in Section 6; quantile box-plots are introduced in Section 7. Section 8 discusses location and scale parameter estimation using trimmed samples. Robust regression is the subject of Section 9. A new definition of statistics is proposed in Section 10.