Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Nonparametric Statistical Data Modeling

Emanuel Parzen [a]
[a] The Institute of Statistics, Texas A&M University, College Station, TX, 77843, USA
Published online: 05 Apr 2012.

To cite this article: Emanuel Parzen (1979) Nonparametric Statistical Data Modeling, Journal of the American Statistical Association, 74:365, 105-121

To link to this article: http://dx.doi.org/10.1080/01621459.1979.10481621

Nonparametric Statistical Data Modeling

EMANUEL PARZEN*

This article attempts to describe an approach to statistical data analysis which is simultaneously parametric and nonparametric. Given a random sample $X_1, \ldots, X_n$ of a random variable X, one would like (1) to test the parametric goodness-of-fit hypothesis H_0 that the true distribution function F is of the form $F(x) = F_0[(x - \mu)/\sigma]$, where F_0 is specified, and (2) when H_0 is not accepted, to estimate nonparametrically the true density-quantile function $fQ(u)$ and score function $J(u) = -(fQ)'(u)$. The article also introduces density-quantile functions, autoregressive density estimation, estimation of location and scale parameters by regression analysis of the sample quantile function, and quantile-box plots.

KEY WORDS: Quantile functions; Density-quantile functions; Quantile-box plots; Goodness of fit; Autoregressive density estimation; Tail of a distribution.

1. INTRODUCTION

This article aims to introduce new types of keys for exploratory data analysis (of continuous data) based on estimating the quantile function and density-quantile function. It appears that this approach leads to an exploratory data analysis which has a firm probability base. Consequently, the distinction between exploratory and confirmatory data analysis can be regarded as a distinction between confirmatory nonparametric statistical data analysis or modeling, and confirmatory *parametric* statistical data analysis.

The basic proposition of this article is that exploratory data analysis and conventional parametric statistical inference both should have as their aim the estimation of the quantile function $Q(u)$, $0 \leq u \leq 1$, of a random variable X of which the data $X_1, \ldots, X_n$ are independent (or dependent) observations. To estimate Q, one assumes a representation for it of the form:

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

which is equivalent to the classic location and scale parameter model for the probability density function $f(x) = (1/\sigma) f_0[(x - \mu)/\sigma]$. We call this representation hypothesis H_0. Four stages of this model can be distinguished:

1. *Parametric model:* Assume Q_0 is known. Then the aim is to estimate μ and σ, using either maximum likelihood estimation or optimal linear combinations of order statistics.
2. *Goodness of fit:* H_0 is tested for various specifications of Q_0 (corresponding to the familiar probability laws, such as normal, exponential, logistic, Weibull, Pareto, Cauchy, and so on).
3. *Robust parametric model:* Q_0 is specified by specifications which permit small deviations from an ideal model, such as "Q_0 symmetric and possibly long-tailed" or "Q_0 normal except for contamination by outliers."
4. *Nonparametric model:* Estimate Q_0, either by estimating the density quantile function $fQ(u) = f(Q(u))$, or through suitable plots of the sample quantile functions of transformations of the data.

The main aim of this article is to introduce a "density estimation" approach to goodness-of-fit tests which also yields estimations of Q. Given a specified hypothesis H_0: $Q(u) = \mu + \sigma Q_0(u)$, one can define a density $d(u)$, $0 \leq u \leq 1$, such that H_0 is equivalent to $d(u) \equiv 1$. Estimation of $d(u)$ provides a test of H_0 and also an estimator of the true fQ function when H_0 is rejected. Many density estimation methods are available; we believe the "autoregressive" method works best for small samples, and we describe it in detail.

Other contributions of this article are discussion of the properties of density-quantile functions (Section 9); the suggestion (outlined in Section 10) that the extensive literature on the "linear combination of order statistics" approach to estimation of location and scale parameters can be rigorously and compactly derived as a regression analysis of the sample quantile function $\tilde{Q}(u)$, $0 \leq u \leq 1$, regarded as a continuous-parameter time series; and quantile-box plots (Section 11), which are an extension of the box plots introduced by Tukey (1977). It should be acknowledged that many developments in this article are extensions of ideas of Tukey (1962, 1977), who has pioneered the development of statistical data analysis as a science.

The basic goal of this article is to contribute the development of a unified theory of statistical science (including statistical data analysis) in which various approaches (exploratory data analysis, parametric statistical inference, robust statistical inference, nonparametric statistical inference, goodness-of-fit procedures, and others) are used simultaneously.

* Emanuel Parzen is Distinguished Professor at the Institute of Statistics, Texas A&M University, College Station, TX 77843. This research was supported in part by the Army Research Office, Grant DA AG29-76-239.

This article was presented at a special JASA Invited Paper Session at the ASA Annual Meeting in San Diego, California, in August 1978. Some of the discussion following the presentation is included here, along with other comments subsequently received in writing.

2. QUANTILE FUNCTIONS AND DENSITY-QUANTILE FUNCTIONS

Given data $X_1, \ldots, X_n$, the aim of statistical data analysis is to find statistical regularities or patterns in the data (that is, to model its probability law). We call data analysis *probability-based* if it seeks to model the distribution function $F(x) = \Pr[X \leq x]$ and the probability density $f(x) = F'(x)$ under the assumption that the data are independent observations of a random variable X (with absolutely continuous distribution function F).

This article proposes to develop exploratory data analysis as confirmatory nonparametric statistical data modeling based on estimating the *quantile function* $Q(u)$, $0 \leq u \leq 1$, defined as follows: for a general distribution function $F(x)$ which is continuous from the right,

$$Q(u) = F^{-1}(u) = \inf\{x : F(x) \geq u\} \ . \tag{2.1}$$

It has the fundamental property that for $-\infty < x < \infty$ and $0 < u < 1$.

$$F(x) \geq u \quad \text{if and only if} \quad Q(u) \leq x \ .$$

Consequently, X is identically distributed as $Q(U)$, where U is uniformly distributed on $[0, 1]$, since

$$\Pr[Q(U) \leq x] = \Pr[U \leq F(x)] = F(x) \ .$$

We write this conclusion symbolically:

$$X \sim Q(U) \ . \tag{2.2}$$

When F is continuous, Q satisfies

$$Q(u) = \inf\{x : F(x) = u\} \ , \tag{2.3}$$
$$FQ(u) = u \ , \quad 0 \leq u \leq 1 \ .$$

Consequently, $F(X)$ is uniformly distributed on $[0, 1]$, since

$$\Pr[F(X) \geq u] = \Pr[X \geq Q(u)] = 1 - FQ(u) = 1 - u \ .$$

The notation $FQ(u)$ represents the composite function $F[Q(u)]$; we next define

$$fQ(u) = f(Q(u)) \ , \tag{2.4}$$

which we call the *density-quantile* function (and pronounce "eff-cue" function). We call

$$q(u) = Q'(u) \tag{2.5}$$

the *quantile-density* function. Tukey (1965) was among the first to give names to these important functions; he calls Q the "representing" function, and q the "sparsity" function.

Differentiating $FQ(u) = u$, we obtain

$$fQ(u)q(u) = 1 \ . \tag{2.6}$$

In words, fQ and q are reciprocals of each other (which justifies calling them by names which are the reverses of each other).

In the literature of nonparametric statistics, it is customary to define a function (see Hájek and Šidák 1967, p. 19):

$$J(u) = \frac{-f'(F^{-1}(u))}{f(F^{-1}(u))} = \frac{-f'Q(u)}{fQ(u)} \ . \tag{2.7}$$

We call $J(u)$ the *score* function of the probability density f; examples are given in the Table. Since $(fQ)'(u) = f'Q(u)q(u)$, it follows that

$$J(u) = -(fQ)'(u) \ ; \tag{2.8}$$

in words, the score function is the derivative of the density-quantile function. For purposes of estimation of the score function, definition (2.7) requires one to first estimate f, f', and F^{-1}; definition (2.8) requires one only to estimate $(fQ)'$.

From a knowledge of the quantile function $Q(u)$, $0 \leq u \leq 1$, of a random variable X, we can (1) simulate it (using 2.2), (2) compute expectations using the formula

$$E[g(X)] = E[gQ(U)] = \int_0^1 gQ(u)du \ , \tag{2.9}$$

and (3) compute measures of location and scale.

From (2.9), the mean μ and variance σ^2 are given by:

$$\mu = \int_0^1 Q(u)du \ , \quad \sigma^2 = \int_0^1 \{Q(u) - \mu\}^2 du \ . \tag{2.10}$$

Quantile Functions and Score Functions

Probability law	$Q(u)$	$J(u)$
Normal	$\Phi^{-1}(u)$	$\Phi^{-1}(u)$
Lognormal	$\exp\{\Phi^{-1}(u)\}$	$\exp\{-\Phi^{-1}(u)\}\{\Phi^{-1}(u) + 1\}$
Exponential	$\log(1-u)^{-1}$	1
Extreme Value	$\log\log(1-u)^{-1}$	$1 + \log(1-u)^{-1}$
Weibull	$\frac{1}{\beta}\{\log(1-u)^{-1}\}^{\beta}$	$\{\log(1-u)^{-1}\}^{-\beta}\{1 - \beta + \log(1-u)^{-1}\}$
Logistic	$\log\frac{u}{1-u}$	$2u - 1$
Double-Exponential	$\log 2u \ , \ u < \frac{1}{2}$ $-\log 2(1-u) \ , \ u > \frac{1}{2}$	$\text{sign}(2u-1)$
Cauchy	$\tan\pi(u - \frac{1}{2})$	$-\sin 2\pi u$
Pareto	$\frac{1}{\beta}(1-u)^{-\beta}$	$(1+\beta)(1-u)^{\beta}$

Another consequence of (2.9) is:

$$\int_0^1 g(fQ(u))du = \int_{-\infty}^{\infty} g(f(x))f(x)dx \ , \qquad (2.11)$$

whence

$$\int_0^1 fQ(u)du = \int_{-\infty}^{\infty} f^2(x)dx \ ,$$
$$\int_0^1 \log fQ(u)du = \int_{-\infty}^{\infty} f(x)\log f(x)dx \ . \qquad (2.12)$$

The right-hand integrals in (2.12) arise often in statistical theory, and we conjecture that it is because they are evaluations of the integrals of fQ and $\log fQ$.

Another important property of quantile functions is how they behave under transformations of random variables. Let F_X and F_Y denote the distribution functions of X and $Y = g(X)$. Similarly, let Q_X and Q_Y denote their quantile functions. If g is an increasing continuous function, then it is well-known that

$$F_Y(y) = F_X(g^{-1}(y)) \ . \qquad (2.13)$$

If, further, F_X is a strictly increasing continuous distribution, then:

$$Q_X F_X(x) = x \ , \qquad (2.14)$$

$$Q_Y(u) = g(Q_X(u)) \ , \qquad (2.15)$$

where (2.15) can be deduced from the fact that

$$F_Y(y) \geq u \text{ iff } F_X(g^{-1}(y)) \geq u \text{ iff } g^{-1}(y) \geq Q_X(u) \text{ iff } y \geq gQ_X(u) \ .$$

Two important consequences of (2.15) are:

If $Y = \mu + \sigma X$, $\sigma > 0$, then $Q_Y(u) = \mu + \sigma Q_X(u)$; if $Y = \log X$, then $Q_Y(u) = \log Q_X(u)$.

In general, if data $X_1, \ldots, X_n$ are assumed to be identically distributed as X, and if the quantile function of X can be transformed to the quantile function of a random variable Y by an increasing continuous transformation g, then the transformed data $g(X_1), \ldots, g(X_n)$ are identically distributed as Y.

3. TRANSFORMATIONS

To simulate a continuous random variable X, one starts with U which is uniformly distributed on (0, 1) and seeks an increasing function Ψ_1 such that $\Psi_1(U)$ and X are identically distributed; it is well-known that $\Psi_1(u) = Q(u)$, the quantile function of X. To estimate Q, we consider a more general problem. Let Y be a random variable with a specified quantile function Q_0. We seek to estimate from a random sample $X_1, \ldots, X_n$ of X, increasing functions Ψ and Ψ_1 such that:

$$\Psi(X) \sim Y \ , \quad \Psi_1(Y) \sim X \ , \qquad (3.1)$$

where $\sim$ means "identically distributed as." In particular, when an observed random variable X is not normal (or exponential), one seeks to find a transformation of data which is normal (or exponential).

The cumulative hazard function $H(x)$ in reliability theory, defined by $H(x) = -\log(1 - F_X(x))$, has the property that $H(X)$ is exponential with mean 1. Thus estimating $H(x)$ can be regarded as actually estimating a transformation to exponentiality.

We are thus led to consider the problem of estimating the transformation Ψ such that $\Psi(X)$ has a prescribed distribution function F_0; further, let Ψ_1 be the transformation such that $\Psi_1(Y) \sim X$. Using suitable axioms that Ψ and Ψ_1 be monotone functions, one could prove:

$$\Psi(x) = Q_0F(x) \ , \quad \Psi_1(y) = QF_0(y) \ . \qquad (3.2)$$

We *define* these to be the transformations desired, since clearly

$$Q_0F(X) \sim Y \ , \quad QF_0(Y) \sim X \ . \qquad (3.3)$$

To find Ψ and Ψ_1, we will find their derivatives

$$\psi(x) = \Psi'(x) \ , \quad \psi_1(y) = \Psi_1'(y) \ . \qquad (3.4)$$

The definitions $\Psi = Q_0F$ and $\Psi_1 = QF_0$ imply:

$$\psi(x) = q_0(F(x))f(x) \ , \qquad (3.5)$$
$$\psi_1(y) = q(F_0(y))f_0(y) \ .$$

Now let $x = Q(u)$ and $y = Q_0(u)$; we obtain:

$$\psi Q(u) = q_0(u)fQ(u) \ , \qquad (3.6)$$
$$\psi_1 Q_0(u) = q(u)f_0Q_0(u) \ .$$

One immediate conclusion is that ψQ and $\psi_1 Q_0$ are reciprocal functions; consequently, estimating their logarithms are equivalent problems. A second conclusion is that estimating ψQ and estimating fQ are equivalent problems since f_0Q_0 is a known function.

It turns out to be natural to estimate the density

$$d(u) = (1/\sigma_0)f_0Q_0(u)q(u) \ , \qquad (3.7)$$

where the normalizing constant σ_0 is defined by:

$$\sigma_0 = \int_0^1 f_0Q_0(u)q(u)du \ . \qquad (3.8)$$

Conditions for σ_0 to be finite are easily obtained from our general classification of fQ functions. We regard σ_0 as a scale parameter; its relationship to other measures of scale can be derived from the important formula (which follows by integration by parts):

$$\sigma_0 = \int_0^1 J_0(u)Q(u)du \ , \qquad (3.9)$$

assuming $f_0Q_0(u)Q(u) = 0$ for $u = 0, 1$.

We find it convenient to introduce the following terminology and definitions.

Definitions: $d(u)$ is the f_0Q_0 *transformation density* of X;

$$D(u) = \int_0^u d(t)dt \ , \quad 0 \leq u \leq 1 \ , \qquad (3.10)$$

is the f_0Q_0 *transformation distribution function* of X; and

$$\varphi(v) = \int_0^1 e^{2\pi i u v} d(u)du \ , \quad v = 0, \pm 1, \ldots \quad (3.11)$$

is the f_0Q_0 *transformation correlation function* of X.

A distribution function equal to $\sigma_0 D(u)$ has been extensively studied in reliability theory (see Barlow and Doksum 1972) under the notation

$$H_F^{-1}(u) = \int_0^{F^{-1}(u)} f_0[F_0^{-1}F(x)]dx \ , \quad (3.12)$$

which we write in our notation, letting $t = F(x)$, as:

$$H_F^{-1}(u) = \int_0^u f_0Q_0(t)q(t)dt \ . \quad (3.13)$$

What is novel in our approach is that we consider the density function and Fourier transform of this distribution function.

Recently, Barlow and Campo (1975) and Barlow and Proschan (1977) have studied the statistic

$$H_F^{-1}(u) = \int_0^{F^{-1}(u)} \{1 - F(x)\}dx \ , \quad (3.14)$$

which they call the "total time on test transform" of the distribution F, and use it to test for exponentiality. It is the same as our $\sigma_0 D(u)$ with $f_0Q_0 = 1 - u$, the density-quantile function of the exponential distribution.

4. SAMPLE QUANTILE FUNCTIONS

Given a sample $X_1, \ldots, X_n$ of a continuous random variable X, we denote the empirical distribution function (EDF) by:

$$\hat{F}(x) = \text{fraction of } X_1, \ldots, X_n \text{ that is } \leq x \ . \quad (4.1)$$

We shall give several definitions of the empirical quantile function (EQF) denoted by $\hat{Q}(u)$. The first definition is:

$$\hat{Q}(u) = \hat{F}^{-1}(u) = \inf\{x: \hat{F}(x) \geq u\} \ . \quad (4.2)$$

It is a piecewise constant function whose values are the order statistics $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$; more precisely,

$$\hat{Q}(u) = X_{(j)} \quad \text{for} \quad (j-1)/n < u \leq j/n \ , \quad j = 1, \ldots, n. \quad (4.3)$$

For $u = 0$, we define $\hat{Q}(0) = X_{(0)}$, where $X_{(0)}$ is taken to be either the sample minimum $X_{(1)}$ or a natural minimum when one is available (when X is nonnegative, one might take $X_{(0)} = 0$).

If one desires to form a smooth function from a step function, one should start with the smoothest reasonable definition (which is differentiable if possible). Consequently, a preferable definition of $\hat{Q}(u)$ might be the piecewise linear function:

$$\hat{Q}(u) = n\left(\frac{j}{n} - u\right) X_{(j-1)} + n\left(u - \frac{j-1}{n}\right) X_{(j)} \quad \text{for} \quad \frac{j-1}{n} \leq u \leq \frac{j}{n} \quad \text{and} \quad j = 1, \ldots, n \ . \quad (4.4)$$

Then $\hat{q}(u) = \hat{Q}'(u)$ is given by:

$$\hat{q}(u) = n(X_{(j)} - X_{(j-1)}) \quad \text{for} \quad (j-1)/n < u < j/n \quad \text{and} \quad j = 1, \ldots, n \ . \quad (4.5)$$

We call $n(X_{(j)} - X_{(j-1)})$, $j = 1, \ldots, n$, the *spacings* of the sample (see Pyke 1965, 1972). The most important fact about $\hat{q}(u)$ is that it is asymptotically exponentially distributed with mean $q(u)$. The sample spectral density of a stationary time series has an analogous property. Consequently there is an *isomorphism* between spacings and sample spectral densities; to any result about one, there is an analogous result about the other. The methods of proofs and exact hypotheses may need to be different for the two cases, but the statement of the conclusion is usually found to be the same.

Estimators which may behave better in small samples from symmetric densities can be obtained by adopting a shifted piecewise linear function as the definition of $\hat{Q}(u)$:

$$\hat{Q}(u) = n\left(\frac{2j+1}{n} - u\right) X_{(j)} + n\left(u - \frac{2j-1}{2n}\right) X_{(j+1)} \quad \text{for} \quad \frac{2j-1}{n} \leq u \leq \frac{2j+1}{2n} \quad \text{and} \quad j = 1, \ldots, n-1 \ . \quad (4.6)$$

We leave $\hat{Q}(u)$ undefined for $u < 1/2n$ or $u > 1 - (1/2n)$. The derivative of $\hat{Q}(u)$ is:

$$\hat{q}(u) = n(X_{(j+1)} - X_{(j)}) \ , \quad \frac{2j-1}{2n} < u < \frac{2j+1}{2n} \cdot \quad (4.7)$$

Finally, a Bayesian definition of $\hat{Q}(u)$ can be adopted, using the fractional order statistics process defined by Stigler (1977).

The asymptotic distribution of the quantile process $\hat{Q}(u)$, $0 \leq u \leq 1$, is usually studied in the literature for the first definition; the work most useful to us is that of Csörgő and Révész (1975, 1978) described in Sections 9 and 10 of this article (see also Shorack 1972). An open research problem is to show that this asymptotic distribution theory applies also to the other definitions of $\hat{Q}$ we have given.

The basic estimators we form in practice are linear functionals

$$T = \int_0^1 W(u)d\hat{Q}(u) \ .$$

For the first definition of $\tilde{Q}$,

$$T = \sum_{j=0}^{n-1} W\left(\frac{j}{n}\right)\{X_{(j+1)} - X_{(j)}\} \ . \tag{4.8}$$

For the second definition of $\tilde{Q}$,

$$T = \sum_{j=1}^{n} n(X_{(j)} - X_{(j-1)}) \int_{(j-1)/n}^{j/n} W(u)du \ . \tag{4.9}$$

We might evaluate the integral by a simple Simpson's rule approximation:

$$\int_{(j-1)/n}^{j/n} W(u)du = \tfrac{1}{6}\left\{W\left(\frac{j-1}{n}\right) + 4W\left(\frac{2j-1}{2n}\right) + W\left(\frac{j}{n}\right)\right\}. \tag{4.10}$$

For the third definition of $\tilde{Q}$,

$$T = \sum_{j=1}^{n-1} n(X_{(j+1)} - X_{(j)}) \int_{(2j-1)/2n}^{(2j+1)/2n} W(u)du \ . \tag{4.11}$$

When $W(u) = \exp\{2\pi iuv\} f_0 Q_0(u)$, we might approximate the last integral by:

$$f_0 Q_0(j/n) \exp\{2\pi iv(j/n)\} \sin(\pi v/n)/\pi v \ .$$

The distribution theory of linear functions of order statistics has an extensive literature (see Chernoff, Gastwirth, and Johns 1976, Moore 1968, Stigler 1974).

The use of the sample quantile function for purposes of statistical data analysis was pioneered by Wilk and Gnanadesikan (1968), who called it the "empirical cumulative distribution function." Their description of its advantages (p. 2) deserves to be quoted in full:

> For one-dimensional samples, the empirical cumulative distribution function (e.c.d.f.), i.e. a plot of the ith ordered value as ordinate against $(i - \frac{1}{2})/n$ as abscissa, provides an exhaustive representation of the data, under the following broad assumptions: (i) that the order of the observations is immaterial; (ii) that there is no classification of the observations, based on extraneous considerations, which one wishes to employ; and (iii) if the sample is non-random, then appropriate weights are specified. [J.W. Tukey points out that this is really the empirical inverse of the c.d.f. and suggests the term "empirical representing function" for it.]
>
> The use of the e.c.d.f. does not depend on any assumption of a parametric distributional specification. It may usefully describe data even when random sampling is not involved. Furthermore, the e.c.d.f. has additional advantages, including:
>
> (i) It is invariant under monotone transformation, in the sense of quantiles (but not, of course, in appearance).
>
> (ii) It lends itself to graphical representation.
>
> (iii) The complexity of the graph is essentially independent of the number of observations.
>
> (iv) It can be used directly and valuably in connexion with censored samples.
>
> (v) It is a robust carrier of information on location, spread and shape, and an effective indicator of peculiarities.
>
> (vi) It lends itself very well to condensation and to interpolation and smoothing.
>
> (vii) It does not involve the "grouping" difficulties that arise in using a histogram.
>
> (viii) It is directly associated with the body of statistical methods which may be referred to as probability plotting procedures, a discussion of which constitutes a major part of the sequel.

5. GOODNESS-OF-FIT TESTS

Given a random sample $X_1, \ldots, X_n$ of a continuous random variable X with distribution function $F(x)$, the simple goodness-of-fit problem is to test the hypothesis that $F(x) = F_0(x)$, where F_0 is a specified distribution function. The probability integral transformation $U = F_0(X)$ transforms the data $X_1, \ldots, X_n$ to new data $U_1, \ldots, U_n$ on the unit interval which one would like to test for a uniform distribution. When the hypothesis is

$$H_0: F(x) = F_0[(x - \mu)/\sigma] \ ,$$

where μ and σ are parameters to be estimated, it is customary to transform X_j to $\hat{U}_j = F_0[(X_j - \hat{\mu})/\hat{\sigma}]$, where $\hat{\mu}$ and $\hat{\sigma}$ are efficient estimators of μ and σ. It is further customary to base tests of the hypothesis on the deviations from the identity function u of the empirical distribution function of $\hat{U}_1, \ldots, \hat{U}_n$. We are proposing that one use instead the empirical quantile function which, under H_0, has the same theoretical properties.

We denote by $\tilde{D}_0(u)$, $0 \leq u \leq 1$, the empirical quantile function of $\hat{U}_1, \ldots, \hat{U}_n$; it can be expressed in terms of the sample quantile function $\tilde{Q}(u)$ of the original data $X_1, \ldots, X_n$ by

$$\tilde{D}_0(u) = F_0\left(\frac{\tilde{Q}(u) - \hat{\mu}}{\hat{\sigma}}\right). \tag{5.1}$$

Many test statistics could be considered for testing H_0; their theory is summarized by Durbin (1973).

In terms of $\tilde{D}_0(u)$ one can consider

$$\tilde{D}_0(0.5) \ , \quad \tilde{D}_0(0.75) - \tilde{D}_0(0.25) \ ,$$

$$\max_{0 \leq u \leq 1} |\tilde{D}_0(u) - u| \ , \quad \int_0^1 |\tilde{D}_0(u) - u|^2 du \ ,$$

$$\int_0^1 J_1(u)d\tilde{D}_0(u) \quad \text{for a specified } J_1(u) \ .$$

Alternatively, one can consider test statistics based on the Fourier transform

$$\tilde{\varphi}_0(v) = \int_0^1 \exp\{2\pi iuv\}d\tilde{D}_0(u) \ , \quad v = 0, \pm 1, \ldots, \tag{5.2}$$

and examine the sequence $|\tilde{\varphi}_0(v)|^2$, $v = 1, 2, \ldots$, or the quadratic form

$$\sum_{v \neq 0} k(v)|\tilde{\varphi}_0(v)|^2$$

for a specified $k(v)$. Similar statistics are considered in the conventional theory of goodness-of-fit tests under the name of "components." The density approach would consider

$$\tilde{d}_0(u) = \tilde{D}_0'(u) \ , \tag{5.3}$$

and statistics such as

$$\int_0^1 \log \tilde{d}_0(u)du \ , \quad \max_{0 \leq u \leq 1} \tilde{d}_0(u) \ .$$

It is beyond the scope of this article to make a comparative analysis of the merits of the various possible tests.

The distribution function $\tilde{D}_0(u)$ requires the estimation of μ and σ. By examining the formula for explicitly computing $\tilde{d}_0(u)$, one sees how to form a density $\tilde{d}(u)$ which can be used to test H_0 without first estimating μ and σ. Differentiating (5.1), one obtains:

$$\tilde{d}_0(u) = f_0\left(\frac{\tilde{Q}(u) - \hat{\mu}}{\hat{\sigma}}\right)\tilde{q}(u)\frac{1}{\hat{\sigma}}. \tag{5.4}$$

Consequently, define

$$\tilde{d}(u) = f_0Q_0(u)\tilde{q}(u)\frac{1}{\tilde{\sigma}_0}, \tag{5.5}$$

where

$$\tilde{\sigma}_0 = \int_0^1 f_0Q_0(u)\tilde{q}(u)du \ . \tag{5.6}$$

Note that (1) $\tilde{d}(u)$ can be regarded as a raw estimator of the density $d(u)$ defined by (3.7), and (2) H_0 is equivalent to the hypothesis that $d(u)$ equals the constant 1, since H_0 is equivalent to $f_0Q_0(u)q(u)$ equals a constant.

We call $\tilde{d}(u)$, $0 \leq u \leq 1$, the *weighted spacings* (or sample transformation-density function);

$$\tilde{D}(u) = \int_0^u \tilde{d}(t)dt \ , \quad 0 \leq u \leq 1 \ ,$$

the cumulative weighted spacings, or sample transformation-distribution function; and

$$\tilde{\varphi}(v) = \int_0^1 \exp\{2\pi iuv\}\tilde{d}(u)du \ , \quad v = 0, \pm 1, \ldots,$$

the sample transformation correlations.

The distribution theory of many of these statistics has already been studied in the literature. For a general $D(u)$, the almost sure convergence to 0 of

$$\max_{0\leq u\leq 1} |\tilde{D}(u) - D(u)|$$

was proved by Barlow and van Zwet (1970). The asymptotic distribution of

$$\int_0^1 \log \tilde{d}(u)du$$

is the same as that of the sample innovation variance, as given by Davis and Jones (1968) and Hannan and Nicholls (1977); compare Blumenthal (1968). The asymptotic distribution of $\tilde{\sigma}_0$ was found by Weiss (1963).

Under H_0, the asymptotic distribution of

$$\max_{0\leq u\leq 1} \tilde{d}(u)$$

is the same as the distribution in time series analysis (first found by Fisher 1929) of the maximum normalized periodogram ordinate of white noise.

An open research problem is the following.

Conjecture: Under H_0, the stochastic process $\sqrt{n}\{\tilde{D}(u) - u\}, 0 \leq u \leq 1$, is asymptotically distributed as a Brownian bridge process $B(u)$, $0 \leq u \leq 1$; this has been proved for $f_0Q_0(u) = 1 - u$, corresponding to the exponential distribution (Barlow, Richard E., personal communication, 1976). It would then follow that all statistics based on $\tilde{D}(u) - u$ have the same asymptotic distribution theory as the corresponding statistics based on $\tilde{F}(x) - x$, $0 \leq x \leq 1$, where $\tilde{F}(x)$ is the EDF of a random sample from a uniform distribution on $[0, 1]$ whose theory is summarized by Durbin (1973).

The foregoing framework includes as special cases many goodness-of-fit test statistics that are being proposed (for example, Andrews' test for normality (Gnandesikan 1977, p. 165) and tests for Weibull and extreme-value distributions introduced by Mann and Fertig 1975).

In Section 7 we propose goodness-of-fit tests based on determining the order of an autoregressive smoother of $\tilde{d}(u)$.

6. DENSITY-QUANTILE AUTOREGRESSIVE REPRESENTATIONS OF GENERALIZATIONS OF GOODNESS-OF-FIT HYPOTHESES

The concepts needed to state our new approach to statistical data analysis have now been defined. Given a random sample $X_1, \ldots, X_n$ of a random variable X, one would like to (1) test the goodness-of-fit hypothesis H_0 that the true df F is of the location-scale parameter form $F(x) = F_0((x - \mu)/\sigma)$, where μ and σ are parameters to be efficiently estimated, and F_0 is specified; (2) in particular, test normality, which corresponds to taking $F_0(x) = \Phi(x)$, the standard normal distribution function, with density-quantile function $f_0Q_0(u) = \varphi\Phi^{-1}(u)$; (3) when H_0 is not accepted, estimate the true eff-cue function $fQ(u)$ and score function $J(u) = -(fQ)'(u)$; (4) when the data are not normal or exponential, find a transformation of the data which transforms them to normality or exponentiality. All these aims can be accomplished by estimating

$$\sigma_0 = \int_0^1 f_0Q_0(u)q(u)du \ ,$$

the density

$$d(u) = (1/\sigma_0)f_0Q_0(u)q(u) \ , \quad 0 \leq u \leq 1 \ ,$$

its distribution function $D(u)$, and its Fourier transform $\varphi(v)$. It can be verified that H_0 is equivalent to *any one* of the following hypotheses:

$$\begin{aligned} Q(u) &= \mu + \sigma Q_0(u) \ ; \\ q(u) &= \sigma q_0(u) \ ; \\ fQ(u) &= (1/\sigma)f_0Q_0(u) \ ; \\ d(u) &= 1 \ ; \\ D(u) &= u \ ; \\ \varphi(v) &= 0 \quad \text{for} \quad v \neq 0 \ . \end{aligned} \tag{6.1}$$

When the density $d(u)$ is constant, it is called "white noise" in honor of an analogous situation in time series analysis. An approach to testing this hypothesis which

also provides an estimator of $d(u)$ when we do not believe it to be a constant is to represent it in a form called an autoregressive representation (since it is analogous to the spectral density of an autoregressive scheme in time series analysis). It is to be emphasized that knowledge of time series analysis is not needed to carry out or understand the concept of autoregressive representation of a density function.

Definition: A density $d(u)$, $0 \leq u \leq 1$, is said to be autoregressive of order m, or to have an *autoregressive representation of order m*, if it is of the form

$$d(u) = K_m|1 + \alpha_m(1)\exp\{2\pi iu\} + \ldots + \alpha_m(m)\exp\{2\pi ium\}|^{-2}, \quad (6.2)$$

where m is an integer called the *order* (whose determination is the most difficult estimation problem), K_m is a positive constant (corresponding to the finite memory m one-step-ahead mean squared prediction error), and $\alpha_m(1), \ldots, \alpha_m(m)$ are *complex-valued* coefficients satisfying the condition that

$$g_m(z) = 1 + \alpha_m(1)z + \ldots + \alpha_m(m)z^m$$

has all its roots outside the unit circle.

When $d(u) = f_0Q_0(u)/\sigma_0 fQ(u)$ is autoregressive of order m, one obtains a representation for fQ which generalizes the formula which holds in the location and scale parameter model:

$$fQ(u) = c_m|1 + \alpha_m(1)\exp\{2\pi iu\} + \ldots + \alpha_m(m)\exp\{2\pi ium\}|^2 f_0Q_0(u), \quad (6.3)$$

where

$$\frac{1}{c_m} = \int_0^1 |1 + \alpha_m(1)\exp\{2\pi iu\} + \ldots + \alpha_m(m)\exp\{2\pi ium\}|^2 f_0Q_0(u)q(u)du. \quad (6.4)$$

In fact we use low-order schemes to represent $d(u)$. We thus consider successively representations for $fQ(u)$ of the form:

$m = 0$: $fQ(u) = c_0 f_0Q_0(u)$;

$m = 1$: $fQ(u) = c_1|1 + \alpha_1(1)\exp\{2\pi iu\}|^2 f_0Q_0(u)$;

$m = 2$: $fQ(u) = c_2|1 + \alpha_2(1)\exp\{2\pi iu\} + \alpha_2(2)\exp\{2\pi iu2\}|^2 f_0Q_0(u)$;

and so on. It is clear that we have a sequence of representations for fQ which start with the hypothesis H_0 and ascend to the general representation

$$fQ(u) = f_0Q_0(u)c_\infty|1 + \alpha_\infty(1)\exp\{2\pi iu\} + \ldots + \alpha_\infty(m)\exp\{2\pi ium\} + \ldots|^2. \quad (6.5)$$

The infinite-order autoregressive representation (6.5) holds when conditions such as the following are true (see Geronimus 1960): first,

$$\frac{fQ(u)}{f_0Q_0(u)}, \quad \frac{f_0Q_0(u)}{fQ(u)}, \quad \log fQ(u), \quad \log f_0Q_0(u)$$

are all integrable over $0 \leq u \leq 1$; second, fQ and f_0Q_0 satisfy a smoothness condition such as differentiability. The speed of convergence of the approximations of order m to the infinite order case depends on the number of derivatives that exist, and is exponentially fast for infinitely differentiable functions.

Theorem 1: The coefficients of an autoregressive representation of order m for the f_0Q_0 transformation density $d(u)$ can be computed from a knowledge of the f_0Q_0 transformation correlations $\varphi(0)$, $\varphi(1)$, $\varphi(-1)$, $\ldots$, $\varphi(m)$, $\varphi(-m)$ up to lag m using the difference equation satisfied by $\varphi(v)$:

$$\varphi(-v) + \alpha_m(1)\varphi(1 - v) + \ldots + \alpha_m(m)\varphi(m - v) = 0, \quad v > 0; \quad (6.6)$$

$$\varphi(0) + \alpha_m(1)\varphi(1) + \ldots + \alpha_m(m)\varphi(m) = K_m. \quad (6.7)$$

Proof: Let z^* denote the complex conjugate of z. Since $d(u) = K_m\{g_m(\exp\{2\pi iu\})(g_m(\exp\{2\pi iu\}))^*\}^{-1}$, we can write:

$$\varphi(-v) + \alpha_m(1)\varphi(1 - v) + \ldots + \alpha_m(m)\varphi(m - v) = \int_0^1 \exp\{-2\pi iuv\}g_m(\exp\{2\pi iu\})d(u)du = \int_0^1 \exp\{-2\pi iuv\}K_m\{g_m(\exp\{2\pi iu\})\}^{*-1}du.$$

Now $\{g_m(\exp\{2\pi iu\})\}^*$ is a polynomial in $\exp\{-2\pi iu\}$ whose reciprocal has a convergent power series in positive powers of $\exp\{-2\pi iu\}$ (with constant term equal to 1) by virtue of the assumption on the location of the zeros of $g(z)$. Since

$$\int_0^1 \exp\{-2\pi iu(v + k)\}du = 0$$

for positive v and k, the above expression equals 0 for $v > 0$, and equals K_m for $v = 0$.

A symmetric density $f(x) = f(-x)$ satisfies

$$fQ(1 - u) = fQ(u), \quad J(1 - u) = -J(u), \quad Q(1 - u) = -Q(u). \quad (6.8)$$

We do not require $d(u)$ to be symmetric in the sense that $d(u) = d(1 - u)$. By permitting *complex-valued* coefficients $\alpha_m(j)$ in autoregressive representations, $d(u)$ can be any continuous function on $0 < u < 1$.

7. DENSITY-QUANTILE AUTOREGRESSIVE ESTIMATION

Given a sample $X_1, \ldots, X_n$ of a continuous random variable, rather than test a goodness-of-fit hypothesis by forming the sample functions $\tilde{d}(u)$, $\tilde{D}(u)$, and $\tilde{\varphi}(v)$, one can form a sequence of autoregressive densities of order m, $\hat{d}_m(u)$, $m = 0, 1, \ldots$, which are candidates for estimators of the true density $d(u)$. The sequence has the

property that $\hat{d}_0(u)$ is constant (identically equal to 1) and $\hat{d}_m(u)$ tends to $\tilde{d}(u)$ as m increases.

Let $g_m(z) = 1 + \alpha_m(1)z + \ldots + \alpha_m(m)z^m$, and

$$\hat{K}(g_m) = \int_0^1 |g_m(\exp\{2\pi iu\})|^2 \tilde{d}(u)du \ . \quad (7.1)$$

Let $\hat{\alpha}_m(1), \ldots, \hat{\alpha}_m(m)$ denote the values of $\alpha_m(1), \ldots, \alpha_m(m)$ minimizing $\hat{K}(g_m)$. We denote by $\hat{g}_m(z) = 1 + \hat{\alpha}_m(1)z + \ldots + \hat{\alpha}_m(m)z^m$ the minimizing polynomial; $\hat{K}(\hat{g}_m)$ is denoted by

$$\hat{K}_m = \int_0^1 |\hat{g}_m(\exp\{2\pi iu\})|^2 \tilde{d}(u)du \ . \quad (7.2)$$

The *autoregressive smoother of order m*, denoted by $\hat{d}_m(u)$, is defined to be

$$\hat{d}_m(u) = \hat{K}_m|1 + \hat{\alpha}_m(1)\exp\{2\pi iu\} + \ldots + \hat{\alpha}_m(m)\exp\{2\pi ium\}|^{-2} \quad (7.3)$$
$$= \hat{K}_m|\hat{g}_m(\exp\{2\pi iu\})|^{-2} \ .$$

Formulas for $\hat{\alpha}_m(j)$ and $\hat{K}_m$ are obtained by using the projection theorem in Hilbert space; $\hat{g}_m(z)$ satisfies the orthogonality conditions

$$\int_0^1 \hat{g}_m(\exp\{2\pi iu\})\exp\{-2\pi iuv\}\tilde{d}(u)du = 0 \ , \quad v = 1, \ldots, m \ , \quad (7.4)$$

which is equivalent to the *normal equations*

$$\bar{\varphi}(-v) + \hat{\alpha}_m(1)\bar{\varphi}(1 - v) + \ldots + \hat{\alpha}_m(m)\bar{\varphi}(m - v) = 0 \ , \quad v = 1, \ldots, m \ . \quad (7.5)$$

The orthogonality conditions (7.4) imply

$$\hat{K}_m = \int \hat{g}_m(\exp\{2\pi iu\})\tilde{d}(u)du$$
$$= 1 + \hat{\alpha}_m(1)\bar{\varphi}(1) + \ldots + \hat{\alpha}_m(m)\bar{\varphi}(m) \ . \quad (7.6)$$

The normal equations (7.5) are analogous to the so-called Yule-Walker equations considered in time series analysis, and can be solved recursively in m.

An estimator $\hat{d}_m(u)$ of $d(u)$ yields an estimator $\hat{f}Q_m(u)$ of fQ which is given explicitly by

$$\hat{f}Q_m(u) = \frac{|1 + \hat{\alpha}_m(1)e^{2\pi iu} + \ldots + \hat{\alpha}_m(m)e^{2\pi ium}|^2 f_0 Q_0(u)}{\int_0^1 |1 + \hat{\alpha}_m(1)e^{2\pi iu} + \ldots + \hat{\alpha}_m(m)e^{2\pi ium}|^2 f_0 Q_0(u)q(u)du} \ , \quad (7.7)$$

where $\hat{\alpha}_m(1), \ldots, \hat{\alpha}_m(m)$ are the solutions of the normal equations (7.5).

To compute $\hat{d}_m(u)$ one needs only $\bar{\varphi}(-m), \ldots, \bar{\varphi}(m)$. For purposes of statistical data analysis, the crucial question is the determination of a suitable value of the order m, which we denote by $\hat{m}$, such that $\hat{d}_{\hat{m}}(u)$ is an "optimal" estimator of $d(u)$.

The most satisfactory approach to order determination currently available seems to be a graphical approach for choosing the appropriate smooth estimator $\hat{d}_m(u)$ of $d(u)$, which judges visually how

$$\hat{D}_m(u) = \int_0^u \hat{d}_m(t)dt$$

fits $\tilde{D}(u)$. If it fits too well, one has over-smoothed, and the density $\hat{d}_m(u)$ will have spurious modes. One wants $\hat{D}_m(u)$ to follow $\tilde{D}(u)$, but not slavishly.

One would like a criterion function which is to be minimized to determine an "optimal" order $\hat{m}$ such that $\hat{d}_{\hat{m}}(u)$ is an "optimal" estimator of $d(u)$. Open for research is the criterion function CAT (criterion autoregressive transfer function) defined for $m = 1, 2, \ldots$ by

$$\mathrm{CAT}(m) = \frac{1}{n}\sum_{j=1}^{m} \hat{K}_j^{-1} - \hat{K}_m^{-1} \ , \quad (7.8)$$

which is analogous to criteria suggested by Parzen (1974, 1977a) for time series spectral estimation. At the present time we regard CAT as a test of H_0; when it chooses $\hat{m} = 0$, we regard it as confirmation that H_0 holds (especially when this hypothesis is accepted by tests based on $\tilde{d}$, $\tilde{D}$, and/or $|\bar{\varphi}|^2$).

For the small samples available in practice, one would choose very low orders m. But one feels comforted by a convergence theorem for large samples which envisages m growing to infinity as the sample size becomes infinite (although for very smooth densities, one expects to be able to choose m proportional to $\log n$). We give one example of a rigorous theorem concerning the consistency in probability of $\hat{d}_m(u)$ as an estimator of $d(u)$, adapted from the work of Carmichael (1976, 1978) on the autoregressive method for probability density estimation: Suppose that (1) $d(u)$, $d^{-1}(u)$, and $\log d(u)$ are integrable over $0 \le u \le 1$, and are continuous; (2) $d(u)$ is bounded above and below in the sense that $0 < d_L \le d(u) \le d_U < \infty$ a.e. in $[0, 1]$; (3) $d(u) = c(u)$ a.e. in $[0, 1]$, and c satisfies, for some $\alpha > 0.5$,

$$\sup_{|h| \le \delta} \int_0^1 |c(u + h) - c(u)|^2 du = O(\delta^{2\alpha}) \ . \quad (7.9)$$

Suppose that m is chosen as a function of the sample size n satisfying

$$\lim_{n\to\infty} m^3/n = 0 \ . \quad (7.10)$$

Then $\hat{d}_m(u)$ is a consistent in probability estimator of $d(u)$ uniformly in u; in symbols,

$$\lim_{n\to\infty} \sup_{0\le u\le 1} |\hat{d}_m(u) - d(u)| \to 0 \text{ in probability} \ .$$

To compare the autoregressive estimator $\hat{f}Q_m(u)$ defined by (7.7) with other possible estimators, one must realize that we are actually estimating the triple of functions fQ, q, and Q, and the basic aim is to form a smooth function $\hat{Q}$ which is an estimator of Q. One can distinguish three general approaches to forming estimators $\hat{Q}$, which we call (1) parametric, (2) nonparametric, and (3) nonparametric preflattened.

The parametric approach assumes a location and scale

parameter representation $Q(u) = \mu + \sigma Q_0(u)$, forms efficiently estimators $\hat{\mu}$ and $\hat{\sigma}$, and then takes $\hat{Q}(u) = \hat{\mu} + \hat{\sigma}Q_0(u)$ as the estimator of Q.

The nonparametric approach estimates Q at a point by averaging over the values of $\tilde{Q}(p)$ for p in a neighborhood of u. An estimator of this form is usually written as a kernel estimator

$$\hat{Q}(u) = \int_0^1 \tilde{Q}(p) \frac{1}{h} K\left(\frac{u-p}{h}\right) dp$$

for a suitable kernel K and bandwidth h. If one adopts the piecewise-linear definition of $\tilde{Q}$, one can differentiate this formula for $\hat{Q}$ to form a smooth estimator $\hat{q}$ of the quantile-density q:

$$\hat{q}(u) = \int_0^1 \tilde{q}(p) \frac{1}{h} K\left(\frac{u-p}{h}\right) dp \ .$$

Estimators of this form are in fact extensively studied in the literature of nonparametric density estimation (see Bofinger 1975, Moore and Yackel 1977). Another approach to fitting smooth curves $\hat{q}$ to the raw function $\tilde{q}$ is to use splines (see Wahba and Wold 1975).

The foregoing estimators of q will have good properties only at a fixed value of u; the consistency of estimation becomes worse the closer u is to 0 to 1, because $q(u)$ is in general a nonintegrable function. This problem can be overcome by multiplying $q(u)$ by a factor $f_0Q_0(u)$ which makes the product $f_0Q_0(u)q(u)$ an integrable function, which is not oscillating as much. When one smooths not $\tilde{q}(u)$ but $f_0Q_0(u)\tilde{q}(u)$, we call the approach *nonparametric preflattened smoothing.*

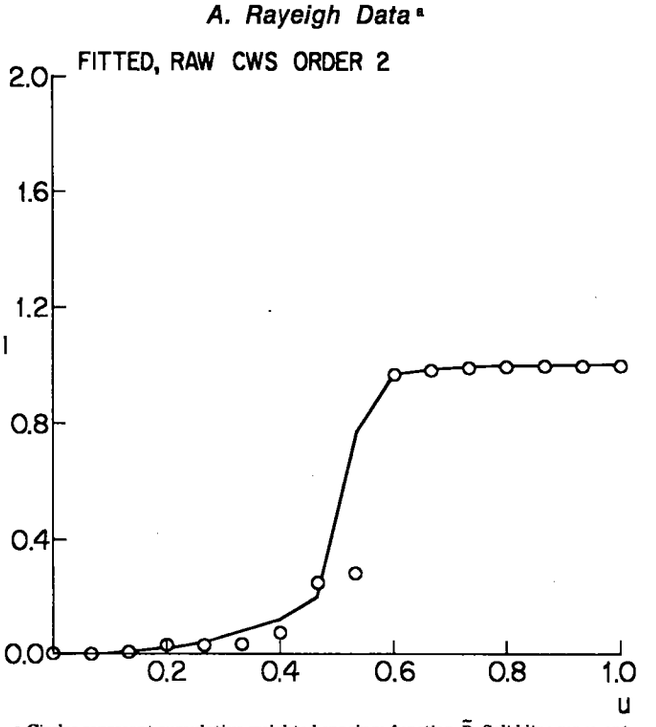
When one seeks to smooth $\tilde{d}(u) = f_0Q_0(u)\tilde{q}(u) \doteq \tilde{\sigma}_0$, the "kernel" approach would be to form estimators of the form:

$$\hat{d}(u) = \int_0^1 \tilde{d}(p) \frac{1}{h} K\left(\frac{u-p}{h}\right) dp \ .$$

The kernel approach is difficult to use in practice because of difficulties in optimally choosing h. The autoregressive approach to density estimation requires a choice of order m. It may be easier to compute (and therefore compare) autoregressive estimators corresponding to different values of m than to compute kernel estimators corresponding to different values of h.

8. COMPUTING ROUTINES AND EXAMPLES

A computer program which implements the data analysis approach described here has been developed by J.P. Carmichael and D. Tritchler. Given a sample $X_1, \ldots, X_n$, it (1) lists their order statistics; (2) plots the quantile function; (3) plots spacings. The f_0Q_0 functions of various familiar probability laws are available to be applied. For a specified f_0Q_0 function, the computer program (4) plots $\tilde{d}(u)$, the raw transformation-density function; (5) plots $\tilde{D}(u)$, the raw transformation-distribution function; (6) plots $|\tilde{\varphi}(v)|^2$, the square-modulus raw transformation-correlations. Next for $m = 1, 2, \ldots$, the autoregressive approximator $\hat{d}_m(u)$ is computed, and its distribution function $\hat{D}_m(u)$ is plotted superimposed on a graph of $\tilde{D}(u)$ to enable one to see how well $\hat{D}_m(u)$ fits $\tilde{D}(u)$. Criteria (especially CAT) to help determine the optimal order m of autoregressive approximation are tabulated. For each m, the density-quantile estimator $\hat{f}Q_m(u)$ corresponding to $\hat{d}_m(u)$ is plotted. In the absence of a rigorous procedure for determining the optimal order $\hat{m}$, we choose the smallest value of m for which $\hat{D}_m(u)$ "fits" $\tilde{D}(u)$.



[a] Circles represent cumulative weighted spacings function $\tilde{D}$. Solid line represents autoregressive estimator $\hat{D}_2$ of order 2.

Rayleigh example: Tukey (1977, p. 49) gives an example of data (Rayleigh's weights of a standard volume of "nitrogen" consisting of 15 measurements) which can be used to take a hard look at the advantage and disadvantages of graphical data analysis techniques. In 1893–1894, Rayleigh measured the densities of nitrogen produced by removing the oxygen from air and nitrogen produced by decomposition of a chemical compound; the discrepancies led him to investigate the composition of air chemically freed of oxygen. He discovered argon and was awarded the 1904 Nobel Prize in Physics.

Tukey discusses how to present the data so as to make it clear that it separates into two quite isolated subgroups, which one interprets as indicating that the single batch of weights might be two batches of weights (as in fact they are, one for "nitrogen" from air, the other for "nitrogen" from other sources).

The presence of two batches will be indicated by the shapes of the empirical quantile function (Figure E). A more rigorous indication would be the presence of two modes in the estimated density-quantile function. We usually estimate fQ taking as the base function f_0Q_0 the

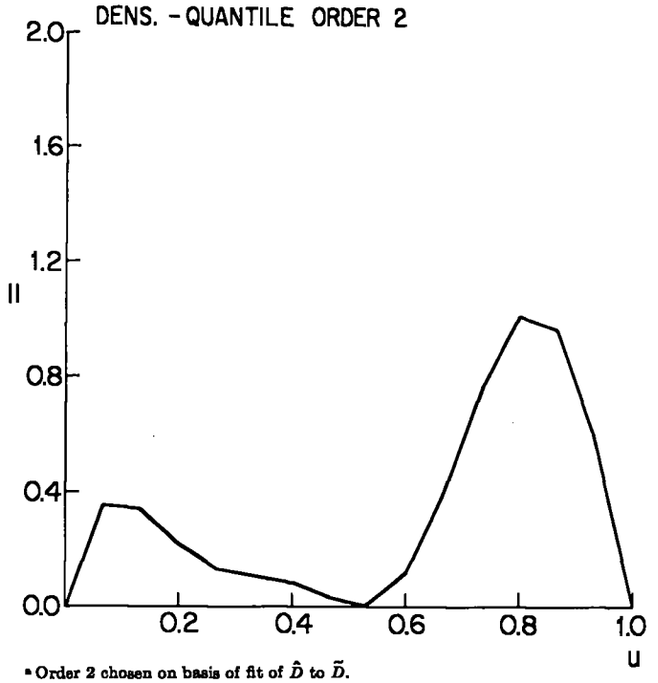
standard normal density, so that the procedure also provides a test of normality. The Rayleigh data is clearly nonnormal. We take order $m = 2$ as an optimal autoregressive approximation (on the criterion of the fit of $\hat{D}_2(u)$ to $\tilde{D}(u)$) and obtain the estimated density-quantile function whose plot appears in Figure A; it is bimodal.

The reader may find it interesting to compare the density-quantile function plot in Figure A with Tukey's (1977, p. 51) two batches box and whiskers plot. Our left-hand mode (representing "other than air" nitrogen measurements) is lower than the right-hand mode (representing "from air" nitrogen measurements), indicating that the left mode population is more variable than the right mode population.

Buffalo snowfall example: The 63 yearly values of snow precipitation in Buffalo (recorded to the nearest tenth of an inch) from 1910–1972 have been extensively analyzed by Carmichael (1976) and Thaler (1974) to illustrate and compare various probability density estimation techniques. It is interesting that Tukey (1977, p. 117) also suggests Buffalo snowfall as an example for analyses (and gives the data for 1918–1937).

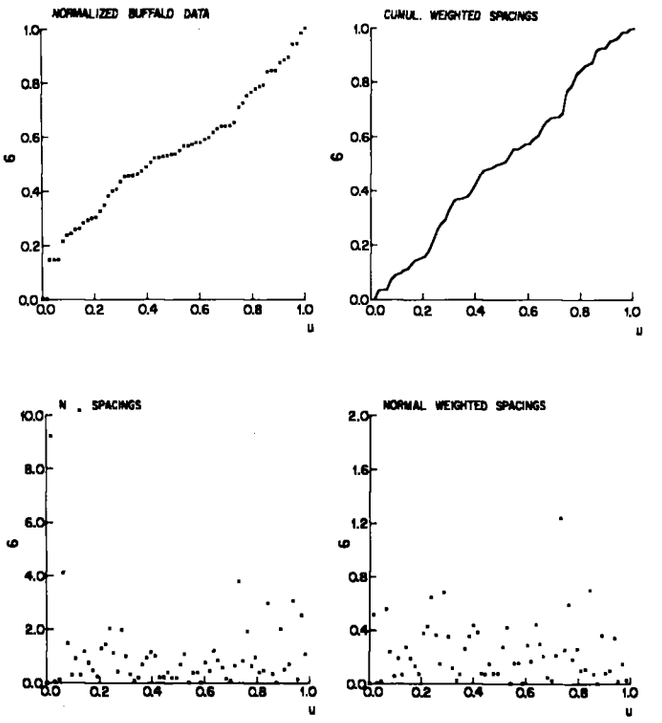
Conventional probability density estimation analyses of Buffalo snowfall indicate that it has either a unimodal or trimodal density, with the trimodal shape usually regarded as the more likely answer. Our density-quantile estimation procedure, with base f_0Q_0 taken to be the standard normal, also yields either a unimodal or trimodal density; the order 0 and order 1 autoregressive estimators $\hat{f}Q_1(u)$ are unimodal, and the order 2 autoregressive estimator $\hat{f}Q_2(u)$ is trimodal (see Figure B).

B. Rayleigh Data Autoregressive Estimator $\hat{f}Q$ of Density Quantile Function[a]



[a] Order 2 chosen on basis of fit of $\hat{D}$ to $\tilde{D}$.

C. Buffalo Snowfall Data[a]



[a] The sample quantile function $\tilde{Q}$ is in upper left graph, spacings or sample quantile-density function $\tilde{q}$ is in lower graph, normal weighted spacings $\tilde{d} = \varphi\Phi^{-1}\tilde{q}$ is in lower right graph, and cumulative weighted spacings $\tilde{D}$ is in upper right graph.

However, all our $\tilde{D}$- and $|\tilde{\varphi}|^2$-based diagnostic tests of the hypothesis H_0 that Buffalo snowfall is normal confirm that it is. Thus the trimodal density estimator often obtained in previous analyses seems to be incorrect. Figure D shows how $\tilde{D}$, the raw cumulative weighted spacings, is fitted by $\hat{D}$ corresponding to orders 1 and 2. Order 2 seems to overfit, and we would argue that order 1 illustrates an adequate fit which we regard as more likely to provide a model in accord with future observations.

The quantile-box plot of Buffalo snowfall, given in Figure E, also indicates that it is normal.

9. DENSITY-QUANTILE CLASSIFICATION OF PROBABILITY LAWS

An examination of the density-quantile functions $fQ(u)$ of familiar probability laws indicates that they can be classified according to their limiting behavior as u tends to 0 or 1. The behavior as $u \to 1$ can be described as

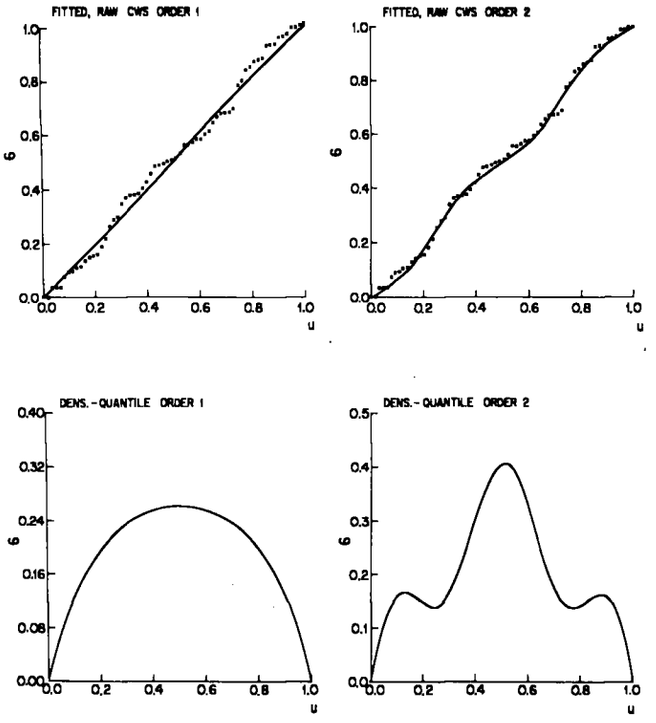
$$fQ(u) \sim (1 - u)^{\alpha}, \quad \alpha > 0,$$

where $g_1(u) \sim g_2(u)$ means $g_1(u) \div g_2(u)$ tends to a positive finite constant as $u \to 1$. When $\alpha = 1$, one *may* have a more refined description of the behavior:

$$fQ(u) \sim (1 - u)\left\{\log \frac{1}{1 - u}\right\}^{1-\beta}, \quad 0 \leq \beta \leq 1.$$

We call α the *tail-exponent* parameter and β the *shape*

D. Buffalo Snowfall Data[a]



[a] Upper left graph depicts $\tilde{D}$ by crosses and $\hat{D}_1$ by solid line; upper right graph depicts $\tilde{D}$ by crosses and $\hat{D}_2$ by solid line. Autoregressive estimators $\hat{f}Q_1$ and $\hat{f}Q_2$ of orders 1 and 2 appear in lower left and lower right graphs, respectively.

parameter of a distribution. A rigorous definition of the tail exponent is given at the end of the section.

The parameter ranges $\alpha < 1$, $\alpha = 1$, and $\alpha > 1$ correspond to the statistician's perception that probability laws have three types of tail behavior: (1) short tails or limited type; (2) medium tails or exponential type; (3) long tails or Cauchy type.

The names limited type, exponential type, and Cauchy type are used in the theory of extreme-value distributions to describe the type of distributions leading to the three types of extreme-value distributions (see Gumbel 1962).

The uniform distribution has $\alpha = 0$:

$$f(x) = 1\ , \quad 0 \le x \le 1\ ; \quad fQ(u) = 1\ , \quad 0 \le u \le 1\ .$$

An example of a short-tailed distribution is:

$$f(x) = c(1-x)^{c-1}\ , \quad 0 \le x \le 1\ ; \qquad fQ(u) = (1/\beta)(1-u)^{1-\beta}\ ,$$

where $c > 0$ and $\beta = 1/c$.

Examples of exponential type distributions are:

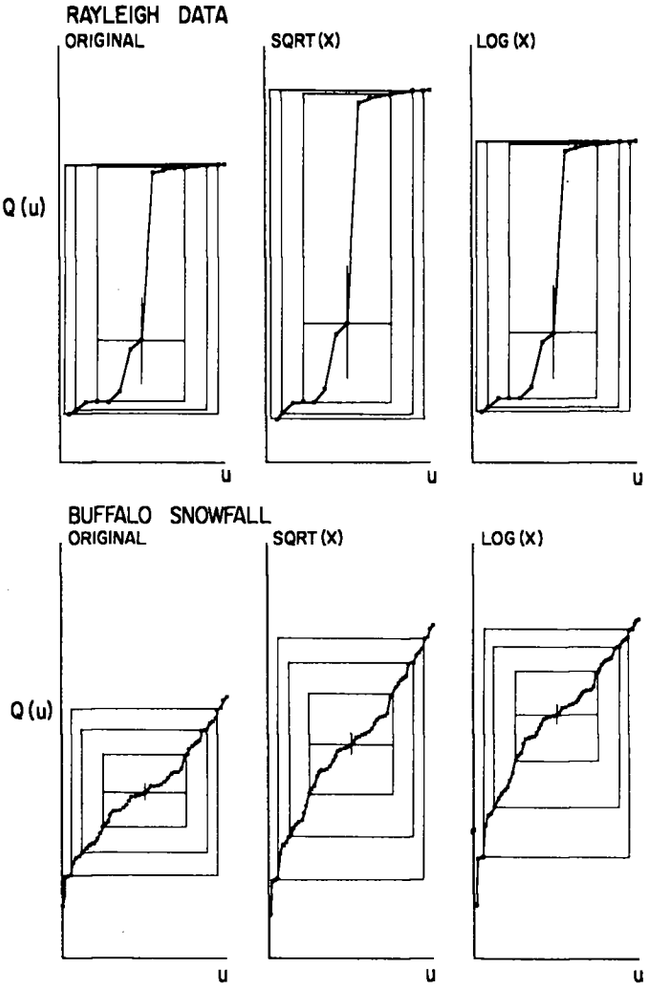
Exponential:

$$f(x) = e^{-x}\ , \quad x > 0\ ; \quad fQ(u) = 1 - u\ ;$$

Logistic:

$$f(x) = \frac{e^x}{(1+e^x)^2}\ , \quad -\infty < x < \infty\ ; \quad fQ(u) = u(1-u)\ ;$$

E. Quantile-Box Plots of Rayleigh Data and Buffalo Snowfall Data, Their Square Root and Logarithm[a]



[a] Buffalo snowfall is normal according to various diagnostic tests; its transformations are not normal. The Rayleigh data are bimodal, indicated by the sharp rise in the graph of its sample quantile function.

Weibull:

$$f(x) = cx^{c-1}e^{-x^c}\ , \quad x > 0\ ;$$

$$fQ(u) = \frac{1}{\beta}(1-u)\left\{\log\frac{1}{1-u}\right\}^{1-\beta}\ ; \quad c = 1/\beta > 0$$

Extreme value:

$$f(x) = e^x e^{-e^x}\ , \quad -\infty < x < \infty\ ;$$

$$fQ(u) = (1-u)\log\frac{1}{1-u}\ ;$$

Normal:

$$\phi(x) = \frac{1}{(2\pi)^{\frac{1}{2}}}e^{-\frac{1}{2}x^2}\ ;$$

$$\Phi(x) = \int_{-\infty}^{x}\phi(y)dy\ ;$$

$$fQ(u) = \frac{1}{(2\pi)^{\frac{1}{2}}}\exp -\tfrac{1}{2}|\Phi^{-1}(u)|^2$$

$$\sim (1-u)\left(2\log\frac{1}{1-u}\right)^{\frac{1}{2}}.$$

Note that in the β parameterization of exponential type distributions (those for which $\alpha = 1$), the values $\beta = 0$, 0.5, and 1 correspond to the extreme-value, normal, and exponential distributions, respectively. Note also that the β parameterization does not cover all exponential type distributions; in particular, it does not cover

Lognormal:

$$f(x) = (1/x)\phi(\log x) \; ; \quad fQ(u) = \phi\Phi^{-1}(u)e^{-\Phi^{-1}(u)} \; .$$

Examples of long-tailed distributions are:

Cauchy:

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}, \quad -\infty < x < \infty \; ;$$

$$fQ(u) = (1/\pi)\cos^2 \pi(u - (\tfrac{1}{2})) = (1/\pi)\sin^2 \pi u \sim (1-u)^2 \; ;$$

Reciprocal of a uniform:

$$f(x) = 1/x^2 \; , \quad x > 1 \; ; \quad fQ(u) = (1-u)^2 \; ;$$

Pareto $\beta > 0$:

$$f(x) = \{\beta x^{1+(1/\beta)}\}^{-1} \; , \; x > 1 \; ; \; fQ(u) = (1/\beta)(1-u)^{1+\beta} \; ;$$

Tukey $\lambda < 1$:

$$Q(u) = \frac{1}{\lambda}\{u^\lambda - (1-u)^\lambda\} \; ;$$

$$fQ(u) = \frac{(1-u)^\alpha}{1 + u^{-\alpha}(1-u)^\alpha}, \quad \alpha = 1 - \lambda \; .$$

The double-exponential distribution exemplifies another aspect of distributions which can be used to classify them—their differentiability.

Double-exponential:

$$f(x) = \tfrac{1}{2}e^{-|x|} \; ; \quad fQ(u) = u \quad \text{for} \quad u < 0.5$$
$$= 1 - u \quad \text{for} \quad u > 0.5 \; .$$

The nondifferentiability (at $x = 0$) of the double-exponential density makes the density-quantile function nondifferentiable at $u = 0.5$. Nondifferentiability of the density is equivalent to the characteristic function

$$\varphi(u) = \int_{-\infty}^{\infty} e^{iux} f(x)dx$$

decaying as $1/u^2$ as $u \to \infty$. Thus one can classify distributions according to the decay rate of (1) their densities and (2) their characteristic functions. The approach to statistical data analysis discussed in this article basically assumes that the densities we are considering are differentiable in order to obtain reasonable rates of consistency for our estimators.

Given data, the parameters we desire to estimate are location μ, scale σ, tail-exponent α, and (when $\alpha = 1$) shape β.

To efficiently estimate location and scale, one must know $f_0Q_0(u)$ or at least its tail exponent α. A formula given by Andrews (1973) for the tail area of a distribution suggests a fundamental formula for the limiting behavior of fQ functions as $u \to 1$, and also suggests a formula which might be used to rigorously define the tail exponent α of a distribution. Andrews's tail area approximation formula may be written:

$$1 - F(x) = \frac{f(x)}{g(x)}\frac{1}{\kappa - 1}\left[1 + \frac{1}{2}\left\{\frac{g'(x)}{g(x)} - \kappa\right\}\right],$$

defining $g(x) = f'(x)/f(x) = \{\log f(x)\}'$ and

$$\kappa = \lim_{x\to\infty} \frac{g'(x)}{g^2(x)}.$$

In this formula, let $u = F(x)$. Then $gQ(u) = -J(u)$, $g'Q(u) = (fQ)''(u)fQ(u)$, and

$$1 - u = \alpha\frac{fQ(u)}{J(u)}\left[1 + \frac{1}{2}\left\{\frac{fQ(u)(fQ)''(u)}{J^2(u)} - \kappa\right\}\right],$$

defining

$$\alpha = \frac{1}{1-\kappa}, \quad \kappa = \lim_{u\to 1}\frac{fQ(u)(fQ)''(u)}{J^2(u)}.$$

The ranges $\alpha < 1$, $\alpha = 1$, $\alpha > 1$ correspond to $\kappa < 0$, $\kappa = 0$, and $\kappa > 0$, respectively.

We are thus led to a rigorous definition of the tail exponent α:

$$\alpha = \lim_{u\to 1}\frac{(1-u)J(u)}{fQ(u)}.$$

This value of α satisfies approximately for u near 1:

$$-(\log fQ(u))' = \frac{J(u)}{fQ(u)} = \frac{\alpha}{1-u},$$

whence $\log fQ(u) = \alpha \log (1-u) + \text{constant}$, and

$$fQ(u) \sim (1-u)^\alpha \; ,$$

which is our intuitive definition of α.

One can state a general assumption describing the densities which we are considering.

A density f is called tail-monotone with exponent $\gamma > 0$ if (1) it is nondecreasing on an interval to the right of $a = \sup\{x: F(x) = 0\}$, and it is nonincreasing on an interval to the left of $b = \inf\{x: F(x) = 1\}$, where $-\infty \le a < b \le \infty$; and (2) $f(x) > 0$ on $a < x < b$, and

$$\sup_{a<x<b} F(x)(1-F(x))\frac{|f'(x)|}{f^2(x)} \le \gamma \; ,$$

or equivalently,

$$\sup_{0<u<1} u(1-u)\frac{|J(u)|}{fQ(u)} \le \gamma \; .$$

Tail-monotone densities are considered (without being so named) by Csörgő and Révész (1978), who demonstrate that they enjoy strong approximations of the quantile process; in Section 10 we apply this fact to estimation of location and scale parameters.

An example of a distribution function which does not satisfy (1) is:

$$1 - F(x) = \exp(-x - C\sin x) \; , \quad 0.5 < C < 1 \; .$$

Letting $x = Q(u)$, one obtains a relation for $Q(u)$:

$$-\log(1-u) = Q(u) + C \sin Q(u) ,$$

whence

$$1/(1-u) = q(u)\{1 + C \cos Q(u)\}$$

and

$$fQ(u) = (1-u)\{1 + C \cos Q(u)\} .$$

As $u \to 1$, $Q(u) \to \infty$, and $fQ(u)$ oscillates. The hazard quantile function

$$hQ(u) = \frac{fQ(u)}{1-u} = 1 + C \cos Q(u)$$

also oscillates.

10. ESTIMATION OF LOCATION AND SCALE PARAMETERS

The problem of estimation of location and scale parameters μ and σ arises under three sets of assumptions called (1) ideal model, (2) robust model, and (3) nonideal model.

The ideal model assumes that the true distribution function F of X may be represented in the form $F(x) = F_0[(x-\mu)/\sigma]$, where F_0 is a known distribution function, or equivalently the quantile function may be represented in the form $Q(u) = \mu + \sigma Q_0(u)$.

The nonideal model assumes F_0 is not known. It may be "estimated" from the data using a goodness-of-fit test for a hypothesis H_0: $Q = \mu + \sigma Q_0$, or by finding a function Ψ_1 such that $X \sim \Psi_1(Y)$, where Y has a specified distribution F_0. We have described an approach to goodness-of-fit tests which finds only the derivative $\psi_1(y) = \Psi_1'(y)$ and thus yields only a representation

$$\Psi_1(y) = \mu + \sigma \Psi_0(y) ,$$

where Ψ_0 is an indefinite integral of ψ_1. Then the quantile function Q of X has the representation:

$$Q(u) = \mu + \sigma \Psi_0 Q_0(u) .$$

The parameters μ and σ in this representation would be estimated in the same way one estimates any other pair of location and scale parameters.

The robust-ideal model assumes F_0 is symmetric but possibly long-tailed, or that F_0 is a contaminated normal distribution.

For each of these models, an important approach to obtaining computationally simple, asymptotically efficient estimators of location and scale parameters μ and σ is to use linear combinations of order statistics. This section outlines how the well-known results for such estimators can be compactly (and even rigorously) obtained by applying to the sample quantile process $\tilde{Q}(u)$, $0 \le u \le 1$, the theory of regression analysis on continuous-parameter time series from the reproducing kernel Hilbert space (RKHS) point of view given by Parzen (1961a,b, 1967).

A rigorous starting point is provided by the important theorems by Csörgõ and Révész (1978) on strong approximation of the quantile process.

Theorem 2: Let $X_1, \ldots, X_n$ be iid random variables with continuous df F and differentiable density f which is tail-monotone with exponent α (as defined at the end of Section 9). The quantile process $\tilde{Q}(u)$ is defined in terms of the order statistics $X_{(1)} < \ldots < X_{(n)}$, and $\tilde{Q}_U(u)$ denotes the quantile process of the uniformly distributed random variables $U_j = F(X_j)$. Let

$$R_n = \sup_{0<u<1} \sqrt{n}\,|fQ(u)\{\tilde{Q}(u) - Q(u)\} - \{\tilde{Q}_U(u) - u\}| .$$

Then almost surely

$$\begin{aligned} R_n &= O(n^{-\frac{1}{2}} \log\log n) && \text{if } \alpha < 1 , \\ &= O(n^{-\frac{1}{2}} (\log\log n)^2) && \text{if } \alpha = 1 , \\ &= O[n^{-\frac{1}{2}} (\log\log n)^{\alpha} (\log n)^{(1+\epsilon)(\alpha-1)}] && \text{if } \alpha > 1 , \end{aligned}$$

where $\epsilon > 0$ is arbitrary.

To state a theorem concerning the behavior of the uniform quantile process $\tilde{Q}_U$, recall the definition of a Brownian bridge $\{B(u),\ 0 \le u \le 1\}$; it is a zero-mean normal process with covariance kernel:

$$K_B(u_1, u_2) = \min(u_1, u_2) - u_1 u_2 .$$

Theorem 3: (Csörgõ and Révész 1975) A Brownian bridge $\{B_n(u),\ 0 \le u \le 1\}$ can be defined for each n such that almost surely

$$\sup_{0 \le u \le 1} |\sqrt{n}\{\tilde{Q}_U(u) - u\} - B_n(u)| = O(n^{-\frac{1}{2}} \log n) .$$

For purposes of statistical inference, we can interpret the foregoing results as follows: $\sqrt{n} fQ(u)\{\tilde{Q}(u) - Q(u)\}$ is distributed as a Brownian bridge $B(u)$. Under the representation $Q(u) = \mu + \sigma Q_0(u)$, we obtain

$$\sqrt{n}(1/\sigma) f_0 Q_0(u)\{\tilde{Q}(u) - \mu - \sigma Q_0(u)\} \sim B(u) .$$

Estimating μ and σ becomes a problem in regression analysis of continuous-parameter time series by writing

$$f_0 Q_0(u) \tilde{Q}(u) = \mu f_0 Q_0(u) + \sigma f_0 Q_0(u) Q_0(u) + \sigma_B B(u) ,$$

where

$$\sigma_B = \sigma/\sqrt{n}$$

is treated as a free parameter not constrained to be related to σ. Estimators of σ_B can be used to test the goodness of fit of the model.

Asymptotically efficient estimators of μ and σ can be formed, given a censored set of order statistics $X_{(np)}, \ldots, X_{(nq)}$, by time series regression analysis of the sample quantile function $\tilde{Q}(u)$ over a subinterval $p \le u \le q$ of $0 \le u \le 1$ (where we permit $p = 0$ or $q = 1$ as possible cases). To form the estimators $\hat{\mu}_{p,q}$ and $\hat{\sigma}_{p,q}$ based on these data, we need to compute the reproducing kernel inner product $\langle f, g \rangle_{p,q}$ of functions on the interval $p \le u \le q$ corresponding to the kernel $K_B(u_1, u_2)$. This RKHS consists of L_2 differentiable functions with inner

product

$$\langle f, g\rangle_{p,q} = \int_p^q f'(u)g'(u)du + \frac{1}{p} f(p)g(p) + \frac{1}{1-q} f(q)g(q) .$$

To verify this assertion, one need only verify the reproducing formula

$$\langle f, K_B(\cdot, t)\rangle_{p,q} = f(t) , \quad p \le t \le q .$$

Now

$$\begin{aligned} K_B(u, t) &= u(1-t) , \quad p \le u \le t \\ &= t(1-u) , \quad t \le u \le q \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial u} K_B(u, t) &= 1 - t , \quad p \le u \le t \\ &= -t , \quad t \le u \le q \end{aligned}$$

$$\begin{aligned} \int_p^q f'(u) \frac{\partial}{\partial u} K_B(u, t)du &= \int_p^t f'(u)(1-t)du \\ &\quad + \int_t^q f'(u)(-t)du \\ &= \{f(t) - f(p)\}(1-t) \\ &\quad + \{f(q) - f(t)\}(-t) \\ &= f(t) - (1-t)f(p) - tf(q) . \end{aligned}$$

Since $f(p)K_B(p, t) = f(p)p(1-t)$, and $f(q)K_B(q, t) = f(q)t(1-q)$, we have verified the reproducing property of our formula for inner product.

We conclude this section with formulas for optimal estimators of μ and σ in the model $Q = \mu + \sigma Q_0$, based on $\tilde{Q}(u)$, $p \le u \le q$. Define the information matrix

$$I(p, q) = \begin{bmatrix} I_{\mu\mu}(p, q) & I_{\mu\sigma}(p, q) \\ I_{\sigma\mu}(p, q) & I_{\sigma\sigma}(p, q) \end{bmatrix},$$

where

$$\begin{aligned} I_{\mu\mu}(p, q) &= \langle f_0Q_0, f_0Q_0\rangle_{p,q} , \\ I_{\mu\sigma}(p, q) &= I_{\sigma\mu}(p, q) = \langle f_0Q_0, Q_0(f_0Q_0)\rangle_{p,q} , \\ I_{\sigma\sigma}(p, q) &= \langle Q_0(f_0Q_0), Q_0(f_0Q_0)\rangle_{p,q} . \end{aligned}$$

Define

$$\begin{aligned} T_{n,\mu,p,q} &= \langle f_0Q_0, \tilde{Q}(f_0Q_0)\rangle_{p,q} , \\ T_{n,\sigma,p,q} &= \langle Q_0(f_0Q_0), \tilde{Q}(f_0Q_0)\rangle_{p,q} . \end{aligned}$$

Then the optimal estimators are given by

$$\begin{bmatrix} \hat{\mu}_{p,q} \\ \hat{\sigma}_{p,q} \end{bmatrix} = I^{-1}(p, q) \begin{bmatrix} T_{n,\mu,p,q} \\ T_{n,\sigma,p,q} \end{bmatrix},$$

with variance and covariance matrix

$$\begin{bmatrix} \operatorname{var}(\hat{\mu}_{p,q}) & \operatorname{cov}(\hat{\mu}_{p,q}, \hat{\sigma}_{p,q}) \\ \operatorname{cov}(\hat{\mu}_{p,q}, \hat{\sigma}_{p,q}) & \operatorname{var}(\hat{\sigma}_{p,q}) \end{bmatrix} = \sigma_B^2 I^{-1}(p, q) .$$

These estimators are similar to those given by Weiss and Wolfowitz (1970). Detailed implementation of these estimators is beyond the scope of this article.

11. QUANTILE-BOX PLOTS APPROACH TO EXPLORATORY DATA ANALYSIS

Autoregressive representations of the density-quantile function provide an approach to probability-based exploratory data analysis aimed at achieving the goals listed in Section 6. A "quick and dirty" approach to achieving the aims of probability-based exploratory data analysis is provided by quantile-box plots.

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

The quantile box plot of data consists of a graph of the sample quantile function $\tilde{Q}(u)$ as a function on the unit interval $0 \le u \le 1$ on which the H, E, and D boxes are drawn superimposed (see Figure E).

Quantile box plots enable the statistician to quickly classify a data batch into three basic types: (1) continuous and unimodal, which might be further classified as (a) normal, (b) symmetric but long-tailed, or (c) normal with outliers; (2) continuous but not unimodal; (3) discrete.

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 (to be defined).

When the data pass the test for symmetry, the question of whether they have a normal or long-tailed distribution is decided using the TAIL diagnostic measures (to be defined). Small TAIL values may indicate bimodal distributions. Data sets with outliers may also yield small TAIL values.

If the graph $x = \tilde{Q}(u)$ 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

D-box, we suspect outliers (values to be discarded for robust estimation).

Many horizontal segments in the graph $x = \tilde{Q}(u)$ are interpreted to mean probability masses, indicating a discrete random variable.

The *mid-summaries* of a data batch are:

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

Of particular interest are $\tilde{\mu}_M = \tilde{\mu}(0.5)$, $\tilde{\mu}_H = \tilde{\mu}(0.25)$, $\tilde{\mu}_E = \tilde{\mu}(0.125)$, and $\tilde{\mu}_D = \tilde{\mu}(0.0625)$.

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 by $\tilde{\mu}(0)$. 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:

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

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

Diagnostic tests of 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:

$$\begin{aligned}
\text{SKEW}(p) &= \{\tilde{\mu}_M - \tilde{\mu}(p)\} \div \tilde{S}(p) \ ; \\
\text{TAIL}(p) &= \log\{\tilde{S}(p) \div \tilde{S}(0.25)\} \ ; \\
\text{TAIL}_0(p) &= \log\{S_0(p) \div S_0(0.25)\} \ ; \\
\text{TAIL}\Phi(p) &= \log\{\Phi^{-1}(p) \div \Phi^{-1}(0.25)\} \ .
\end{aligned}$$

When SKEW(p) is not significantly different from zero, we consider the data batch to have a symmetric distribution. When the data pass a SKEW test for symmetry, they are checked for normality by comparing TAIL(p) with TAIL$\Phi(p)$; TAIL(p) significantly larger than TAIL$\Phi(p)$ indicates a long-tailed distribution, and TAIL(p) significantly shorter than 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 is easily found as re-expressions of the seven-number summary of the original data batch.

12. SOME OPEN RESEARCH PROBLEMS FOR EXTENSIONS

The approach described in this article formulates statistical estimation and testing problems as problems of density estimation and testing for white noise. This article discusses only the univariate one-sample case. Two-sample and multivariate (including nonparametric regression) problems can be treated similarly (see Parzen 1977b). This section describes some extensions of our results in the one-sample case whose theory and application are open for research.

Power Transformation to Normality: The transformation of a random variable X to a $N(\mu, \sigma^2)$ distribution is often assumed to be of one of the following forms for some appropriate choice of λ:

$$\begin{aligned}
\Psi(x) &= (1/\lambda)\{(x-\zeta)^\lambda - 1\} \ , \quad \lambda \neq 0 \ , \\
&= \log(x-\zeta) \ , \quad \lambda = 0 \ .
\end{aligned}$$

The derivative $\psi(x) = \Psi'(x)$ has a single formula:

$$\psi(x) = (x-\zeta)^{\lambda-1} \ .$$

The quantile function $Q(u)$ of X is then related to the standard normal quantile function $\Phi^{-1}(u)$ by

$$\begin{aligned}
\mu + \sigma\Phi^{-1}(u) &= (1/\lambda)\{(Q(u)-\zeta)^\lambda - 1\} \ , \quad \lambda \neq 0 \\
&= \log(Q(u)-\zeta) \ , \quad \lambda = 0 \ .
\end{aligned}$$

The density-quantile function of X satisfies

$$\log fQ(u) = -\log\sigma + \log\phi\Phi^{-1}(u) + (\lambda - 1)\log(Q(u)-\zeta) \ .$$

The problem is (1) to use these relations to estimate the parameters λ and ζ; and (2) to compare these estimators with the estimators of Box and Cox (1964).

Survival Data: Let $X_1, \ldots, X_n$ be a random sample from a single lifetime or survival distribution F with quantile function Q. However, one may fail to observe an X (called a "death") due to the previous occurrence of some other event Y (called a "loss") which has distribution H. The desired value X is censored on the right by Y, and one observes

$$Z = \min(X, Y) \ ,$$

with distribution function G satisfying

$$1 - G = (1-F)(1-H)$$

under suitable independence assumptions.

From the observed data $Z_1, \ldots, Z_n$, one can form an estimator $\tilde{F}$ of F introduced by Kaplan and Meier (1958). Its quantile function $\tilde{Q}$ is an estimator of Q. The asymptotic distribution theory of $\tilde{F}$ and $\tilde{Q}$ has been found by Breslow and Crowley (1974) and Sander (1975), respectively; the latter shows that $\sqrt{n}fQ(u)\{Q(u) - \tilde{Q}(u)\}$, $0 < u < 1$, converges in distribution (as a stochastic process) to a zero-mean Gaussian process with covariance kernel K given by

$$K(u_1, u_2) = (1-u_1)(1-u_2)\int_0^{\min(u_1,u_2)} dw(1-w)^{-2} \cdot \{1 - HQ(w)\}^{-1} \ .$$

When there is no censoring, $H = 0$ and $K(u_1, u_2)$

$= u_1(1 - u_2)$ for $u_1 < u_2$, the covariance kernel of the Brownian bridge.

The covariance kernel K has an integral representation which makes it easy to find its RKHS inner product. Thus one would have no difficulty extending the results of Section 10 to estimation of location and scale parameters from survival data.

Sampling the Quantile Process: Suppose that to compress the data, one seeks to reduce a sample of size n to k values, namely the order statistics $X_{(np_j)} \doteq \tilde{Q}(p_j)$ corresponding to specified percentiles $p_1, \ldots, p_k$. These percentiles can be chosen so that the optimal linear estimators $\hat{\mu}$ and $\hat{\sigma}$ that could be formed from them have variances which are a minimum over all choices of k points at which to sample $\tilde{Q}(u)$. Results of this kind could be deduced from the work of Sacks and Ylvisaker (1966) on designs of continuous-parameter time series regression problems; one would generalize and unify the extensive literature on this topic (see Hassanein 1977 for a good set of references).

[*Received January 1978. Revised May 1978.*]

REFERENCES

Andrews, D.F. (1973), "A General Method for the Approximation of Tail Areas," *Annals of Statistics*, 1, 367–372.

Barlow, Richard E., and Campo, Raphael (1975), "Total Time on Test Processes and Applications to Failure Data Analysis," in *Reliability and Fault Tree Analysis*, eds. Richard E. Barlow, Jerry B. Fussell, and Nozer D. Singpurwalla, Philadelphia: Society for Industrial and Applied Mathematics, 451–481.

———, and Doksum, Kjell A. (1972), "Isotonic Tests for Convex Orderings," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, I, 293–323.

———, and Proschan, F. (1977), "Asymptotic Theory of Total Time on Test Processes with Applications to Life Testing," in *Multivariate Analysis IV*, ed. Paruchuri R. Krishnaiah, Amsterdam: North-Holland and Publishing Co., 227–237.

———, and Van Zwet, Willem R. (1970), "Asymptotic Properties of Isotonic Estimators for the Generalized Failure Rate Function Part I: Strong Consistency," in *Nonparametric Techniques in Statistical Inference*, ed. Madan Lal Puri, Cambridge: Cambridge University Press, 159–173.

Blumenthal, Saul (1968), "Logarithms of Sample Spacings," *SIAM Journal on Applied Mathematics*, 16, 1184–1191.

Bofinger, Eve (1975), "Estimation of a Density Function Using Order Statistics," *Australian Journal of Statistics*, 17, 1–7.

Box, G.E.P., and Cox, D.R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Ser. B, 26, 211–252.

Breslow, N., and Crowley, J. (1974), "A Large Sample Study of the Life Table and Product Limit Estimates under Random Censorship," *Annals of Statistics*, 2, 437–453.

Carmichael, J.P. (1976), "The Autoregressive Method," unpublished Ph.D. thesis, Statistical Science Division, State University of New York at Buffalo.

——— (1978), "Consistency of the Autoregressive Method of Density Estimation," submitted for publication.

Chernoff, Herman, Gastwirth, Jospeh L., and Johns, M.V., Jr. (1967), "Asymptotic Distribution of Linear Combinations of Functions of Order Statistics with Applications to Estimation," *Annals of Mathematical Statistics*, 38, 52–72.

Csörgő, M., and Révész, P. (1975), "Some Notes on the Empirical Distribution Function and the Quantile Process," in *Limit Theorems of Probability*, Amsterdam: North-Holland Publishing Co., 59–71.

———, and Révész, P. (1978), "Strong Approximations of the Quantile Process," *Annals of Statistics*, 6, 882–894.

Davis, Herbert T., and Jones, Richard H. (1968), "Estimation of the Innovation Variance of a Stationary Time Series," *Journal of the American Statistical Association*, 63, 141–149.

Durbin, J. (1973), *Distribution Theory for Tests Based on the Sample Distribution Function*, Regional Conference Series in Applied Mathematics, 9, Philadelphia: Society for Industrial and Applied Mathematics.

Elderton, William Palin, and Johnson, Norman Lloyd (1968), *Systems of Frequency Curves*, Cambridge: Cambridge University Press.

Fisher, R.A. (1929), "Tests of Significance in Harmonic Analysis," *Proceedings of the Royal Society of London*, Ser. A., 125, 54–59.

Geronimus, Y.L. (1960), *Polynomials Orthogonal on a Circle and Interval* (translated from the Russian by D.E. Brown), New York: Pergamon Press.

Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations*, New York: John Wiley & Sons.

Gumbel, E.J. (1962), "Statistical Theory of Extreme Values (Main Results)," in *Contributions to Order Statistics*, eds. Ahmad E. Sarhan and Bernard G. Greenberg, New York: John Wiley & Sons, 56–93.

Hájek, Jaroslav, and Šidák, Zbyněk (1967), *Theory of Rank Tests*, New York: Academic Press.

Hannan, E.J., and Nicholls, D.F. (1977), "The Estimation of the Prediction Error Variance," *Journal of the American Statistical Association*, 72, 834–840.

Hassanein, K.M. (1977), "Simultaneous Estimation of the Location and Scale Parameters of the Gamma Distribution by Linear Functions of Order Statistics," *Scandinavian Actuarial Journal*, 60, 88–93.

Kaplan, E.L., and Meier, Paul (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481.

Mann, Nancy R., and Fertig, Kenneth W. (1975), "A Goodness-of-Fit Test for the Two Parameter vs. Three Parameter Weibull; Confidence Bounds for Threshold," *Technometrics*, 17, 237–245.

Moore, D.S. (1968), "An Elementary Proof of Asymptotic Normality of Linear Functions of Order Statistics." *Annals of Mathematical Statistics*, 39, 263–265.

———, and Yackel, James W. (1977), "Consistency Properties of Nearest Neighbor Density Function Estimators," *Annals of Statistics*, 5, 143–154.

Parzen, Emanuel (1961a), "An Approach to Time Series Analysis," *Annals of Mathematical Statistics*, 32, 951–989.

——— (1961b), "Regression Analysis of Continuous Parameter Time Series" *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, I, 469–489.

——— (1967), *Time Series Analysis Papers*, San Francisco: Holden-Day.

——— (1970), "Statistical Inference on Time Series by RKHS Methods," in *Proceedings of the Twelfth Biennial Seminar of the Canadian Mathematical Congress*, ed. Ronald Pyke, Montreal: Canadian Mathematical Congress, 1–37.

——— (1974), "Some Recent Advances in Time Series Modeling," *IEEE Transactions on Automatic Control*, AC-19, 723–730.

——— (1977a), "Multiple Time Series: Determining the Order of Approximating Autoregressive Schemes," in *Multivariate Analysis IV*, ed. Paruchuri R. Krishnaiah, Amsterdam: North-Holland Publishing Co., 283–296.

——— (1977b), "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.

Pyke, R. (1965), "Spacings," *Journal of the Royal Statistical Society*, Ser. B., 27, 395–449 (with discussion).

——— (1972), "Spacings Revisited," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, I, 417–427.

Sacks, Jerome, and Ylvisakar, Donald (1966), "Designs for Regression Problems with Correlated Errors," *Annals of Mathematical Statistics*, 37, 66–89.

Sander, J.M. (1975), "The Week Convergence of Quantiles of the Product-Limit Estimator," Technical Report No. 5, Department of Statistics, Stanford University.

Sarhan, Ahmad E., and Greenberg, Bernard G. (eds.) (1962), *Contributions to Order Statistics*, New York: John Wiley& Sons.

Shorack, Galen R. (1972), "Convergence of Quantile and Spacings Processes with Applications," *Annals of Mathematical Statistics*, 43, 1400–1411.

Stigler, Stephen M. (1974), "Linear Functions of Order Statistics with Smooth Weight Functions," *Annals of Statistics*, 2, 676–693.

——— (1977), "Fractional Order Statistics, with Applications," *Journal of the American Statistical Association*, 72, 544–550.

Thaler, H. (1974), "Nonparametric Probability Density Estimation and the Empirical Characteristic Function," unpublished Ph.D. thesis, Statistics Department, State University of New York at Buffalo.

Tukey, John W. (1962), "The Future of Data Analysis," *Annals of Mathematical Statistics*, 33, 1–67.

——— (1965), "Which Part of the Sample Contains the Information?," *Proceedings of the National Academy of Sciences*, 53, 127–134.

——— (1977), *Exploratory Data Analysis*, Reading Mass.: Addison-Wesley Publishing Co.

Wahba, G., and Wold, S. (1975), "Periodic Splines for Spectral Density Estimation: The Use of Cross Validation for Determining the Degree of Smoothing," *Communications in Statistics*, 4, 125–141.

Weiss, Lionel (1963), "On the Asymptotic Distribution of an Estimate of a Scale Parameter," *Naval Research Logistics Quarterly*, 10, 1–9.

———, and Wolfowitz, J. (1970), "Asymptotically Efficient Non-Parametric Estimators of Location and Scale Parameters," *Zeitschrift fur Wahrscheinlichkeits-Theorie und Verw. Gebiete*, 16, 134–150.

Wilk, M.B., and Gnanadesikan, R. (1968), "Probability Plotting Methods for the Analysis of Data," *Biometrika*, 55, 1–17.

Comment

JOHN W. TUKEY*

I am pleased to meet so many old friends, some under new names.

I congratulate our speaker on a great social feat, introducing medians and hinges to reproducing kernel Hilbert spaces.

More seriously, however, we have to thank the speaker for successfully weaving together important strands from:

1. the representing function and sparsity function view (e.g., Tukey 1965) of distributions (as complementary to the cumulative function and density function view);
2. the use of medians, hinges, eighths, etc. as simple tools in the examination, often graphical, of batches of data;
3. the autoregressive approach to time series analysis, where time-side calculations are linked with oscillation-frequency ideas.

One hopes that, in particular, this weaving together will make all three of these areas better known, both for themselves and jointly. I am sure that it would be regrettable, however, if any one of these areas came to be usually thought of only in connection with the other two.

I can foresee many discomforts as statisticians adapt to new ways of description and the necessary new words, whether they be those we have heard today or others. I can say, from experience, however, that the time and effort spent in learning the "representing approach"—the approach of regarding the observed values as coming to us via a function from some standard distribution, rather than as going away from us through a function, to a standard—is time well spent.

A few words of warning and interrelation are, I think, needed.

First, the score function $J(u)$ is Fisher's score for location expressed in terms of cumulation u rather than value y. There are many purposes for which this change is desirable, though there may well be significant exceptions. My warning is merely that you should recognize how the usage of the term "score function" differs.

Next, I must complain with all due politeness and vigor about the usage of the words "exploratory data analysis." As the putative originator of the phrase, I claim as large a right as anyone to say what this phrase has been intended to mean. In my view:

1. the usage in Parzen's article is not compatible with the general usage over the last few years;
2. if the usage in Parzen's article were to become general, it would be necessary to find a new phrase for the original meaning of "exploratory data analysis," something that I hope can be avoided.

My discomfort here is far from trivial.

In my view, as I indicated in a session yesterday, "exploratory data analysis" is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as for those we believe might be there. Except for emphasis on graphs, its tools are secondary to its purposes.

If we take the speaker literally, exploratory data analysis uses only one tool, looking at the pattern of observations in a single batch in such terms as median, hinges, etc. This is a profound misrepresentation. I doubt that more than ten percent of my book (Tukey 1977), for example, is devoted to such techniques. This is

* John W. Tukey is Donner Professor of Science and Professor, Department of Statistics, Princeton University, Princeton, NJ 08540, and Associate Executive Director-Research, Bell Laboratories, Murray Hill, NJ 07974. Research was supported by Army Research Office (Durham) Grant DAAG29-76-G-0298.