Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

DTIC FILE COPY

2

AD-A224 307

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

TITLE (and Subtitle)
Unification of Statistical Methods for
Continuous and Discrete Data

5. TYPE OF REPORT & PERIOD COVERED
Technical

6. PERFORMING ORG. REPORT NUMBER

AUTHOR(s)
Emanuel Parzen

8. CONTRACT OR GRANT NUMBER(s)
DAAL 03-90-G-0069

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Texas A&M University
Institute of Statistics
College Station, TX 77843

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS

12. REPORT DATE
May 1990

13. NUMBER OF PAGES
27

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

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

Bivariate

Entropy, ~~Comparison~~ density functions, Renyi information approximation, Chi-square Test information divergence, Multi-sample data analysis, Tests of homogeneity, ~~Bivariate~~ data analysis

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

This paper introduces notation and concepts which establish unity and analogues between various steps of statistical data analysis, estimation, and hypothesis testing by expressing them in terms of optimization and function approximation using information criteria to compare two distributions.

ARO 27574.1-MA

TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143

Department of STATISTICS
Statistical Interdisciplinary
Research Laboratory
E415EP@TAMVM1 BITNET

Emanuel Parzen
Distinguished Professor
Phone 409-845-3188
Fax 409-845-3144

UNIFICATION OF STATISTICAL METHODS FOR CONTINUOUS AND DISCRETE DATA

Emanuel Parzen

Department of Statistics

Texas A&M University

Technical Report No. #105

May 1990

Texas A&M Research Foundation

Project No. 6547

'Functional Statistical Data Analysis and Modeling'

Sponsored by the U. S. Army Research Office

Professor Emanuel Parzen, Principal Investigator

Approved for public release; distribution unlimited

UNIFICATION OF STATISTICAL METHODS FOR CONTINUOUS AND DISCRETE DATA

by Emanuel Parzen

Department of Statistics, Texas A&M University[1]

0. Introduction

This paper introduces notation and concepts which establish unity and analogues between various steps of statistical data analysis, estimation, and hypothesis testing by expressing them in terms of optimization and function approximation using information criteria to compare two distributions. The contents may be described as composed of two parts whose section titles are as follows.

Part I. Statistical Information Mathematics and Comparison Density Functions.

1. Traditional Entropy and Cross-Entropy
2. Comparison Density Functions
3. Renyi Information Approximation
4. Chi-square Information Divergence

Part II. Comparison Density Approach to Unity of Statistical Methods

5. One Sample Continuous Data Analysis
6. One Sample Discrete Data Analysis
7. Multi-sample Data Analysis and Tests of Homogeneity
8. Bivariate Data Analysis
9. Examples of One Sample and Multi-Sample Continuous Data Analysis

1. Traditional Entropy and Cross-Entropy

The (Kullback-Liebler) information divergence between two probability distributions F and G is defined by

$$I(F;G) = (-2)\int_{-\infty}^{\infty} \log\{g(x)/f(x)\}f(x)dx$$

[1] Research supported by the U.S. Army Research Office

when F and G are continuous with probability density functions $f(x)$ and $g(x)$; when F and G are discrete, with probability mass functions $p_F(x)$ and $p_G(x)$, information divergence is defined by

$$I(F;G) = (-2)\sum \log\{p_G(x)/p_F(x)\}p_F(x).$$

An information decomposition of information divergence is

$$I(F;G) = H(F;G) - H(F),$$

in terms of entropy $H(F)$ and cross-entropy $H(F;G)$; our definitions differ from usual definitions by a factor of 2:

$$H(F) = (-2)\int_{-\infty}^{\infty}\{\log f(x)\}f(x)dx,$$
$$H(F;G) = (-2)\int_{-\infty}^{\infty}\{\log g(x)\}f(x)dx.$$

2. Comparison Density Functions

Information divergence $I(F;G)$ is a concept that works for both multivariate and univariate distributions. This paper proposes that the univariate case is distinguished by the fact that we are able to relate $I(F;G)$ to the concept of comparison density $d(u;F,G)$ whose maximum entropy estimation provides significant extensions of information divergence.

Quantile domain concepts play a central role; $Q(u) = F^{-1}(u)$ is the quantile function. When F is continuous, we define the density quantile function $fQ(u) = f(Q(u))$, score function $J(u) = -(fQ(u))'$, and quantile density function

$$q(u) = 1/fQ(u) = Q'(u).$$

When F is discrete, we define $fQ(u) = p_F(Q(u))$, $q(u) = 1/fQ(u)$.

The comparison density $d(u;F,G)$ is defined as follows: when F and G are both continuous,

$$d(u;F,G) = g(F^{-1}(u))/f(F^{-1}(u));$$

when F and G are both discrete

$$d(u : F, G) = p_G(F^{-1}(u))/p_F(F^{-1}(u)).$$

In the continuous case $d(u; F, G)$ is the derivative of

$$D(u; F, G) = G(F^{-1}(u));$$

in the discrete case we define

$$D(u; F, G) = \int_0^u d(t; F, G)dt.$$

Let F denote the true distribution function of a continuous random variable Y. To test the goodness of fit hypothesis $H_0 : F = G$, one transforms to $W = G(Y)$ whose distribution function is $F(G^{-1}(u))$ and whose quantile function is $G(F^{-1}(u))$. The comparison density $d(u; F, G)$ and $d(u; G, F)$ are respectively the quantile density and the probability density of W.

3. Renyi Information Approximation

For a density $d(u)$, $0 < u < 1$, Renyi information (of index λ), denoted $IR_\lambda(d)$, is nonnegative and measures the divergence of $d(u)$ from uniform density $d_0(u) = 1$, $0 < u < 1$. It is defined:

$$IR_0(d) = 2\int_0^1 \{d(u)\log d(u)\}du;$$

$$IR_{-1}(d) = -2\int_0^1 \{\log d(u)\}du;$$

for $\lambda \neq 0$ or -1

$$IR_\lambda(d) = \{2/\lambda(1+\lambda)\}\log\int_0^1 \{d(u)\}^{1+\lambda}du.$$

To relate comparison density to information divergence we use the concept of Renyi information IR_λ which yields the important identity (and interpretation of $I(F; G)$!)

$$\begin{aligned} I(F; G) &= (-2)\int_0^1 \log d(u; F, G)du \\ &= IR_{-1}(d(u; F, G)) = IR_0(d(u; G, F)). \end{aligned}$$

Interchanging F and G: One can prove a basic identity:

$$IR_\lambda(d(u;F,G)) = IR_{-(1+\lambda)}(d(u;G,F))$$

Note $\lambda = -(1+\lambda)$ for $\lambda = -.5$. Hellinger information divergence is

$$IR_{-.5}(d) = -8\log\int_0^1 \{d(u)\}^{.5}du.$$

Minimizing $IR_\lambda(d)$ subject to constraints on d is equivalent, for $\lambda > 0$, to minimizing the L_p norm of d for $p = 1+\lambda$; we can apply the mathematical theory of this problem which is currently being developed (Chui, Deutsch, Ward (1990)). Note L_2 norm corresponds to $\lambda = 1$. The minimizing function $d\hat{}$ will satisfy $IR_\lambda(d\hat{}) \leq IR_\lambda(d)$.

Convergence Lemma. If $d_m(u)$ is a sequence of densities and $\lambda \geq 0$, $IR_\lambda(d_m(u))$ converges to 0 implies $\int_0^1 |d_m(u) - 1|du$ converges to zero.

Approximation Theory. To a density $d(u)$, $0 < u < 1$, approximating functions are defined by constraining (specifying) the inner product between $d(u)$ and a specified function $J(u)$, called a score function. We often assume that the integral over (0,1) of $J(u)$ is zero, and the integral of $J^2(u)$ is finite. A score function $J(u)$, $0 < u < 1$, is always defined to have the property that its inner product with $d(u)$, denoted

$$[J,d] = [J(u), d(u)] = \int_0^1 J(u)d(u)du,$$

is finite. The inner product is called a *component* or *linear detector*; its value is a measure of the difference between $d(u)$ and 1.

The question of which distributions to choose as F and G is often resolved by the following formula which evaluates the inner product between $J(u)$ and $d(u;F,G)$ as a moment with respect to G if $J(u) = \varphi(F^{-1}(u))$:

$$\left[\varphi\left(F^{-1}(u)\right), d(u;F,G)\right] = \int_{-\infty}^{\infty} \varphi(y)dG(y) = E_G[\varphi(Y)]$$

Often G is a raw sample distribution and F is a smooth distribution which is a model for G according to the hypothesis being tested.

Approximations in L_2 norm are based on a sequence $J_k(u)$, $k = 1,2,\ldots$, which is a complete orthonormal set of functions. Then if $d(u)$, $0 < u < 1$, is square integrable (equivalently, $IR_1(d)$ is finite) one can represent $d(u)$ as the limit of

$$d_m(u) = 1 + \sum_{k=1}^{m} [J_k, d] J_k(u), m = 1,2,\ldots.$$

When $\varphi_k(y)$, $k = 1,2,\ldots$, is complete orthonormal set for $L_2(F)$, $g(y)$ is approximated by

$$g_m(y) = f(y) \left\{ 1 + \sum_{k=1}^{m} E_G \left[\varphi_k(Y)\right] \varphi_k(y) \right\}$$

We call $d_m(u)$ a truncated orthogonal function (generalized Fourier) series.

An important general method of density approximation, called a weighted orthogonal function approximation, is to use suitable weights w_k to form approximations

$$d^*(u) = 1 + \sum_{k=1}^{\infty} w_k [J_k, d] J_k(u).$$

to $d(u)$. Often w_k depends on a "truncation point" m, and $w_k \to 1$ as $m \to \infty$.

We propose that non-parametric statistical inference and density estimation can be based on the same criterion functions used for parametric inference if one uses the minimum Renyi information approach to density estimation (which extends the maximum entropy approach); form functions $d_{\lambda,m}{}^{\wedge}(u)$ which minimize $IR_\lambda(d^{\wedge}(u))$ among all functions $d^{\wedge}(u)$ satisfying the constraints

$$[J_k, d^{\wedge}] = [J_k, d] \text{ for } k = 1,\ldots,m$$

where $J_k(u)$ are specified score functions. One expects $d_{\lambda,m}{}^{\wedge}(u)$ to converge to $d(u)$ as m tends to ∞, and $IR_\lambda(d_{\lambda,m}{}^{\wedge})$ to non-decreasingly converge to $IR_\lambda(d)$.

Quadratic Detectors. To test $H_0 : d(u) = 1$, $0 < u < 1$, many traditional goodness of fit test statistics (such as Cramer-von Mises and Anderson-Darling) can be expressed as quadratic detectors

$$\sum_{k=1}^{\infty} \{w_k \left[J_k, d\right]\}^2 = \int_0^1 \{d^*(u) - 1\}^2 \, du$$

$$= \int_0^1 \{d^*(u)\}^2 \, du - 1 = -1 + \exp IR_1(d^*).$$

We propose that these nonadaptive test statistics are only of historical interest since they are not as powerful as minimum Renyi information detectors $IR_\lambda(d_{\lambda,m}^{\wedge})$; in addition the latter provide unification of statistical methods.

Maximum entropy approximators correspond to $\lambda = 0$; $d_{0,m}^{\wedge}(u)$ satisfies an exponential model (whose parameters are denoted $\theta_1, \ldots, \theta_m$)

$$\log d_{0,m}^{\wedge}(u) = \sum_{k=1}^{m} \theta_k J_k(u) - \Psi(\theta_1, \ldots, \theta_m)$$

where Ψ is the integrating factor that guarantees that $d_{0,m}^{\wedge}(u)$, $0 < u < 1$, integrates to 1:

$$\Psi\left(\theta_1, \ldots, \theta_m\right) = \log \int \exp\left\{ \sum_{ji=1}^{m} \theta_k J_k(u) \right\} du$$

The approximating functions formed in practice are not computed from the true components $[J_k, d]$ but from raw estimators $[J_k, d^{\sim}]$ for suitable raw estimators $d^{\sim}(u)$. The approximating functions are interpreted as estimators of a true density. Methods proposed for unification and generalization of statistical methods use minimum Renyi information estimation techniques. Different applications of these methods differ mainly in how they define the raw density $d^{\sim}(u)$ which is the starting point of the data analysis.

4. Chi-square information divergence

In addition to Renyi information divergence (an extension of information statistics) one needs to use an extension of chi-square statistics which has been developed by Read and Cressie (1988). For $\lambda \neq 0, 1$, Chi-square divergence of index λ is defined for continuous F and G by

$$C_\lambda(F; G) = \int B_\lambda\left(\frac{g(y)}{f(y)}\right) f(y)dy$$

where

$$B_\lambda(d) = \frac{2}{(1+\lambda)}\left\{ d\left(\frac{d^\lambda - 1}{\lambda}\right) - d + 1 \right\}$$

$$B_0(d) = 2\left\{d \log d - d + 1\right\}$$

$$B_{-1}(d) = -2\left\{\log d - d + 1\right\}$$

Important properties of $B_\lambda(d)$ are:

$$B_\lambda(d) \geq 0, B_\lambda(1) = B'_\lambda(1) = 0,$$
$$B'_\lambda(d) = \frac{2}{\lambda}\left(d^\lambda - 1\right), B''_\lambda(d) = 2d^{\lambda-1}$$
$$B_1(d) = (d-1)^2$$
$$B_0(d) = 2(d \log d - d + 1)$$
$$B_{-.5}(d) = 4\left(d^{.5} - 1\right)^2$$
$$B_{-1}(d) = -2(\log d - d + 1)$$
$$B_{-2}(d) = d\left(d^{-1} - 1\right)^2$$

Renyi information of index λ is defined for continuous F and G: for $\lambda \neq 0$, 1

$$IR_\lambda\left(F;G\right) = \frac{2}{\lambda(1+\lambda)} \log \int \left\{\frac{g(y)}{f(y)}\right\}^{1+\lambda} f(y)dy$$
$$IR_0(F/G) = 2\int \left\{\frac{g(y)}{f(y)} \log \frac{g(y)}{f(y)}\right\} f(y)dy$$
$$IR_{-1}(F;G) = -2\int \left\{\log \frac{g(y)}{f(y)}\right\} f(y)dy$$

An analogous definition holds for discrete F and G.

The Renyi information and chi-square divergence measures are related:

$$IR_0(F;G) = C_0(F;G)$$
$$IR_{-1}(F;G) = C_{-1}(F;G)$$

For $\lambda \neq 0, 1$,

$$IR_\lambda(F;G) = \frac{2}{\lambda(1+\lambda)} \log \left\{1 + \left(\frac{\lambda(1+\lambda)}{2}\right) C_\lambda(F;G)\right\}$$

Interchange of F and G is provided by the Lemma:

$$C_\lambda(F;G) = C_{-(1+\lambda)}(G;F)$$
$$IR_\lambda(F;G) = IR_{-(1+\lambda)}(G;F)$$

For a density $d(u)$, $0 < u < 1$, define

$$C_\lambda(d) = \int_0^1 B_\lambda(d(u))du.$$

The comparison density again unifies the continuous and discrete cases. One can show that for univariate F and G

$$C_\lambda(F, G) = C_\lambda(d(u; F, G))$$

5. One Sample Continuous Data Analysis

We now apply statistical information mathematics to describe a unified approach to one sample continuous data analysis which uses optimization and approximation based on information criteria to develop methods which are simultaneously parametric, nonparametric, maximum entropy nonparametric, estimation, testing parametric hypotheses, and goodness of fit of parametric model. Let $Y_1, \ldots, Y_n$ be a random sample of a continuous random variable Y with true unknown distribution F and sample distribution $F^{\sim}$.

A parametric model $F(x;\theta)$ for F assumes that the true probability density function belongs to a parametric family $f(x;\theta)$ with distribution function $F(x;\theta)$. The maximum likelihood estimator $\theta^{\wedge}$ minimizes

$$I(F^{\sim}; F(\cdot;\theta)) = IR_{-1}(d(u; F^{\sim}, F(\cdot;\theta))).$$

To prove the proposition, we denote by $L(\theta)$ the twice average log likelihood function:

$$\begin{aligned} L(\theta) &= (2/n)\log f(Y_1, \ldots, Y_n; \theta) \\ &= 2E^{\sim}[\log f(Y;\theta)] \\ &= 2\int_{-\infty}^{\infty} \log f(y;\theta) dF^{\sim}(y). \end{aligned}$$

To maximize likelihood we express it as minus cross-entropy:

$$L(\theta) = -H(F^{\sim}; F(\cdot;\theta)).$$

Temporarily assuming away the fact that $\tilde{F}$ has only a symbolic density $\tilde{f}$, the maximum likelihood estimator $\hat{\theta}$ can be regarded as minimizing over θ

$$I(\tilde{F}; F(.;\theta)).$$

$\hat{\theta}$ may be interpreted as the parameter value θ for which the sample quantile function $F(\tilde{F}^{-1}(u);\theta)$ of the transformed variable $W_\theta = F(Y;\theta)$ is closest to uniform. Traditional goodness of fit statistics test how close to uniform is the sample distribution function $\tilde{F}(Q(u;\hat{\theta}))$ of $W_{\hat{\theta}}$ whose symbolic probability density is a raw estimator of $d(u; F(\cdot;\hat{\theta}), F)$.

Outline of statistical reasoning: We propose that the various steps of statistical reasoning compose 4 actions which are the goals of statistical science:

1. Make observations (step 0) and summarize by $\tilde{F}$; 2. Form expectations (steps 1 and 2) which is a parametric model for the observations expressed by $F(\cdot;\hat{\theta})$; 3. Compare observations and expectations (steps 3 and 4); 4. Revise model to fit observations (steps 5 and 6). The revised model is equivalent to a nonparametric estimator $\hat{F}$.

Step 0. *Observations.* The sample is summarized by its sample distribution function $\tilde{F}$ and its sample quantile function $\tilde{Q}$.

Step 1: *Parametric Model Specification.* Using diagnostic tools (such as the identification quantile function) identify a parametric family $F(x;\theta)$ such that for all θ

$$sup_u d(u; F(\cdot;\theta)), F) < \infty.$$

Step 2: Parameter estimation. Maximum likelihood estimator $\hat{\theta}$ can be obtained by minimizing

$$IR_{-1}(d(u; \tilde{F}, F(\cdot;\theta))$$

A parametric estimator of F is $\hat{F}(x) = F(x;\hat{\theta})$.

Step 2*: *Robust* parameter estimators $\hat{\theta}^{\lambda}$ can be obtained by minimizing

$$IR_\lambda\left(d\left(u; F^*, F(\cdot;\theta)\right)\right)$$

for a suitable smooth non-parametric distribution function estimator F^* and suitable values of λ, usually chosen in the interval $-1 \leq \lambda < 0$.

Step 3: Parametric hypothesis testing. To test a hypothesis H_0 about the parameter θ, let $\theta_{H_0}{}^{\wedge}$ denote the maximum likelihood estimator of θ under H_0; equivalent to likelihood ratio tests is the test statistic

$$IR_{-1}\left(d\left(u; F^{\sim}, F\left(\cdot; \theta_{H_0}{}^{\wedge}\right)\right)\right) - IR_{-1}\left(d\left(u; F^{\sim}, F\left(\cdot; \theta^{\wedge}\right)\right)\right)$$

Step 4: Goodness of fit test of $H_0 : F = F(\cdot; \theta^{\wedge})$ or equivalently $H_0 : d(u; F(\cdot; \theta^{\wedge}), F) = 0$. Test the significance of the difference from zero of

$$IR_0(d(u; F(\cdot; \theta^{\wedge}), F^{\sim}) = IR_{-1}(d(u; F^{\sim}, F(\cdot; \theta^{\wedge})).$$

Step 5: Maximum entropy goodness of fit tests and estimators $d_{0,m}{}^{\wedge}(u)$ of $d(u; F(\cdot; \theta^{\wedge}), F)$ are obtained by minimizing $I_0(d^{\wedge})$ among densities $d^{\wedge}(u)$ satisfying, for $k = 1, \ldots, m$ and specified score functions $J_k(u)$,

$$[J_k, d^{\wedge}] = [J_k, d^{\sim}]$$

defining $d^{\sim}(u) = d(u; F(\cdot; \theta^{\wedge}), F^{\sim})$. For m large enough $d_{0,m}{}^{\wedge}(u)$ equals $d^{\sim}(u)$ and $IR_0(d_{0,m}{}^{\wedge})$ increases to the test statistic of Step 4, $IR_0(d(u; F(\cdot; \theta^{\wedge}), F^{\sim}))$.

Step 6: Rejection simulation nonparametric estimation of F. Use an order determining criterion to determine an order $m^{\wedge}$ with the properties: if $m^{\wedge} = 0$, accept H_0; if one rejects H_0 use $d_{0,m^{\wedge}}{}^{\wedge}(u)$ as the density to be used in the rejection method of simulating a random sample from F. The combination of $F(\cdot; \theta^{\wedge})$ and $d_{0,m^{\wedge}}{}^{\wedge}(u)$ is regarded as an estimator $F^{\wedge}$.

We propose that order determining criteria should be regarded as providing density estimators which require further goodness of fit tests. We propose (as an open research problem) a method for testing if a smooth estimator $d^{\wedge}(u)$ adequately smooths a raw estimator $d^{\sim}$: test if the ratio $d^{\sim}(u)/d^{\wedge}(u)$ has as its best smoother a constant function.

6. One Sample Discrete Data Analysis

Step 1: Identify a parametric family of probability mass functions $p(x;\theta)$ to model the sample probability mass function $\tilde{p}(x)$.

Step 2: Parameter estimation. Maximum likelihood estimator $\hat{\theta}$ can be obtained by minimizing

$$IR_{-1}(d(u;\tilde{F},F(\cdot;\theta))) = (-2)\sum_{x} \log\{p(x;\theta)/\tilde{p}(x)\}\tilde{p}(x)$$

A parametric estimator of p is $\hat{p}(x) = p(x;\hat{\theta})$. Minimum chi-square estimation uses the modified chi-squared distance

$$IR_1(d(u;\tilde{F},F(\cdot;\theta))) = \sum_{x}\{(p(x;\theta)/\tilde{p}(x)) - 1\}^2\tilde{p}(x)$$

Step 3: Parametric hypthesis testing. To test a hypothesis H_0 about the parameter θ, let $\hat{\theta}_{H_0}$ denote the minimum-modified chi square estimator of θ under H_0; equivalent to likelihood ratio tests is the test statistic

$$IR_1(d(u;\tilde{F},F(\cdot;\hat{\theta}_{H_0}))) - IR_1(d(u;\tilde{F},F(\cdot;\hat{\theta})))$$

Step 4: Goodness of fit test of $H_0 : p = p(\cdot;\hat{\theta})$ or equivalently $H_0 : d(u;F(\cdot;\hat{\theta}),F) = 0$. Test the significance of the difference from zero of

$$IR_1(d(u;F(\cdot;\hat{\theta}),\tilde{F})) = IR_{-2}(d(u;\tilde{F},F(\cdot;\hat{\theta}))).$$

Step 5: Maximum entropy goodness of fit tests and estimators $\hat{d}_{0,m}(u)$ of $d(u;F(\cdot;\theta^*),F)$ are obtained by minimizing $IR_0(\hat{d})$ among densities $\hat{d}(u)$ satisfying, for $k = 1,\ldots,m$ and specified score functions $J_k(u)$,

$$[J_k,\hat{d}] = [J_k,\tilde{d}]$$

defining $\tilde{d}(u) = d(u;F(\cdot;\hat{\theta}),\tilde{F})$. For m large enough $\hat{d}_{0,m}(u)$ equals $\tilde{d}(u)$ and $IR_0(\hat{d}_{0,m})$ increases to a test statistic (alternative to that of Step 4) $IR_0(d(u;F(\cdot;\hat{\theta}),\tilde{F}))$.

Step 6 Rejection simulation nonparametric estimation of F. Use an order determining criterion to determine an order $m^{\wedge}$ with the properties: if $m^{\wedge} = 0$, accept H_0; if one rejects H_0 use $d_{0,m^{\wedge}}{}^{\wedge}(u)$ as the density to be used in the rejection method of simulating a random sample from F. The combination of $F(\cdot;\theta^{\wedge})$ and $d_{0,m^{\wedge}}{}^{\wedge}(u)$ is regarded as an estimator $F^{\wedge}$.

7. Multi-Sample Data Analysis and Tests of Homogeneity

Multi-sample data arises when one observes the values of a variable Y in several populations which can be regarded as indexed by a variable X. One can therefore regard multi-samples as independent observations of a bivariate random variable (X, Y). Conventional multi-sample statistical analysis is concerned with testing the hypothesis H_0 of homogeneity, which we express

$$H_0 : Y \text{ is independent of } X.$$

To formulate H_0 in terms of comparison density functions let us note that nonparametric statistics are based on replacing the response Y by its rank transform which in the population is

$$W = F_Y(Y)$$

The sample rank transform is

$$W^{\sim} = P_Y{}^{\sim}(Y)$$

where $P_Y{}^{\sim}(y)$ is sample mid-distribution function of Y defined by

$$P_Y{}^{\sim}(y) = F_Y{}^{\sim}(y) - .5p_Y{}^{\sim}(y),$$

in terms of the sample distribution function $F^{\sim}$ and sample probability mass function $p^{\sim}$.

The sample quantile function of the $W^{\sim}$ values which are rank transforms of Y values associated with a fixed value of X is an estimator of the conditional quantile function of W

$$Q_{W:X}(u) = F_{Y:X}(F_Y^{-1}(u)) = D(u; F_Y, F_{Y:X}).$$

The innovation of our approach is a new type of linear rank statistics which are estimators of the form, called sample components,

$$[J_k(u), d(u; F_Y{}^{\sim}, F_{Y:X}{}^{\sim})]$$

of population components

$$[J_k(u), d(u; F_Y, F_{Y:X})]$$

for suitable score functions $J_k(u)$.

Score functions $J(u) = \phi(F_Y^{-1}(u))$ satisfy

$$\left[J(u), d(u; F_Y, F_{Y|X})\right] = E_{Y|X}[\varphi(Y)] = E_{Y|X}[J(F_Y(Y))];$$

this component is the conditional mean given X of $J(W)$, where $W = F_Y(Y)$ is the rank transform of Y. A Wilcoxon statistic corresponds to $J(u) = (12)^{.5}(u - .5)$ whose sample component is equivalent to a rank-sum statistic. The conditional mean $E[Y|X]$ is a component with score function $J(u) = Q_Y(u)$.

The traditional approach to multiple sample tests of homogeneity is to test the significant difference from zero of the sample components. We propose that a comprehensive way to test the homogeneity hypothesis H_0 is to estimate the comparison density $d(u; F_Y, F_{Y|X})$ for each value of X, and various chi-square statistics

$$C_{X,\lambda} = C_\lambda(d(u; F_Y, F_{Y:X})).$$

8. Bivariate Data Analysis

Another approach to understanding the role of chi-squared measures of the difference of the comparison density from the uniform $d(u) = 1$, $0 < u < 1$, is to regard X and Y as random variables and express the homogeneity hypothesis H_0 as a hypothesis of independence:

$$H_0 : F_{Y|X}(y|x) = F_Y(y) \text{ for all } y \text{ and } x.$$

Bivariate data analysis can be unified by the *dependence density function* defined for $0 < u_1, u_2 < 1$ by

$$d_{X,Y}(u_1, u_2) = d(u_2; F_Y, F_{Y|X}(\cdot|Q_X(u_1))$$

Traditional maximum likelihood estimators (and EM algorithms) for response variables Y with covariates X can be based on the information measure of dependence, called mutual information, defined for continuous random variables X and Y by

$$\begin{aligned} I(Y|X) &= IR_{-1}(F_{X,Y}, F_X F_Y) = IR_{-1}(f_{X,Y}, f_X f_Y) \\ &= (-2)\int \log\{f_X(x) f_Y(y)/f_{X,Y}(x,y)\} f_{X,Y}(x,y)dxdy \end{aligned}$$

The fundamental relation usually used to study the information about Y in X measured by $I(Y|X)$ is

$$I(Y|X) = H(Y|X) - H(Y)$$

defining $H(Y|X) = E_X H(f_{Y|X})$, called conditional entropy of Y given X.

We obtain a fundamental relation expressing mutual information in terms of comparison density function of F_Y and $F_{Y|X}$ which measures how well f_Y models $Y|X$:

$$\begin{aligned} I(Y|X) &= E_X IR_0(d(u; F_Y, F_{Y|X})) \\ &= E_X C_0(d(u; F_Y, F_{Y|X})) \end{aligned}$$

This is proved by writing

$$\begin{aligned} I(Y|X) = 2\int & dx f_X(x) \\ & \int dy f_Y(y) \log\{f_{Y|X}(y|x)/f_Y(y)\}\{f_{Y|X}(y|x)/f_Y(y)\} \\ = 2E_X & \int_0^1 \log\{d(u; F_Y, F_{Y|X})\} d(u; F_Y, F_{Y|X})du \end{aligned}$$

Traditional chi-squared test statistics satisfy

$$\begin{aligned} C_\lambda(F_{X,Y}, F_X F_Y) &= E_X C_\lambda(d(u; F_{Y|X}, F_Y)) \\ &= E_X C_{-(1+\lambda)}(d(u; F_Y, F_{Y|X})) \end{aligned}$$

We define the chi-square divergence (of index λ) of Y given X to be

$$C_\lambda(Y|X) = E_X C_\lambda(d(u; F_Y, F_{Y|X}))$$

Traditional chi-square statistics for discrete data use $\lambda = 1$. Read and Cressie (1988) recommend $\lambda = 2/3$.

The notation is at hand to state our comparison density approach to multi-sample data analysis and tests of homogeneity:

Step 1: Form raw estimates for each value of X

$$d_X^{\sim}(u) = d(u; F_Y^{\sim}, F_{Y|X}^{\sim})$$

which is computed using the formula for comparison density function of sample discrete distributions.

Step 2: Form and test significance of difference from zero of various components

$$[J_k(u), d_X^{\sim}(u)]$$

for suitable score functions $J_k(u)$.

Step 3: Estimators of $d_X(u) = d(u; F_Y, F_{Y:X})$ by minimum Renyi information estimators

$$d_{X,\lambda,m}^{\wedge}(u)$$

subject to constraints

$$[J_k(u), d_X^{\wedge}(u)] = [J_k(u), d_X^{\sim}(u)]$$

Step 4: Smooth chi-squared tests of H_0 are based on smooth density estimators substituted in the population formulas

$$C_\lambda(Y|X) = E_X[C_{X,\lambda}],$$

$$C_{X,\lambda} = C_\lambda(d(u; F_Y, F_{Y|X}))$$

Further one can disaggregate $C_{X,\lambda}$ into statistics $C_{X,Y,\lambda}$ called "hanging Chi-squares". They are asymptotically distributed as Chi-Squared with 1 degree of freedom. If one rejects the hypothesis of homogeneity, the hanging Chi-squares help identify the sources of rejection.

This outline requires many details and examples to be understandable by statisticians not used to the point of view of statistical culture.

Contingency table data analysis. For $0 < p < 1$, define ODDS $(p) = p/(1-p)$. For r by c contingency table, total sample size N, one forms sample statistics

$$C_{\lambda}^{\sim}(Y|X) = \sum_{k=1}^{c} \{1 - p_X^{\sim}(k)\}\, C_{k,\lambda}^{\sim}$$

$$C_{k,\lambda}^{\sim} = \sum_{j=1}^{r} \{1 - p_Y^{\sim}(j)\}\, C_{j,k,\lambda}^{\sim}$$

$$C_{j,k,\lambda}^{\sim} = \text{ ODDS } (p_X^{\sim}(k)) \text{ ODDS } (p_Y^{\sim}(j)) B_{\lambda}\left(\frac{p_{Y|X}^{\sim}(j|k)}{p_Y^{\sim}(j)}\right)$$

Asymptotic distributions of test statistics:

$$(N-1)C_{\lambda}^{\sim}(Y|X) \text{ is Chi-square } ((r-1)(c-1))$$

$$(N-1)C_{k,\lambda}^{\sim} \text{ is Chi-square } (r-1)$$

$$(N-1)C_{j,k,\lambda}^{\sim} \text{ is Chi-square}$$

Multiple-Sample Goodness of Fit Tests. One can associate a weighted orthogonal series density estimator $d^*(u; F_Y, F_{Y|X})$ for each value of X, using suitable complete orthonormal functions $\varphi_j(u)$ and weights w_j.

$$C_{\lambda}^{\sim}(Y|X) = \sum_{k=1}^{c} \{1 - p_X^{\sim}(k)\}\, C_k^{\sim}$$

$$C_k^{\sim} = \text{ ODDS } (p_X^{\sim}(k)) \int_0^1 \left\{d^*(u; F_Y, F_{Y|X=k}) - 1\right\}^2 du$$

$$= \sum_{j=1}^{\infty} w_j^2 C_{j,k}^{\sim}$$

$$C_{j,k}^{\sim} = \text{ ODDS } \{p_X^{\sim}(k)\} \left[\varphi_j(u), d^{\sim}(u; F_Y, F_{Y|X=k})\right]^2$$

Cramer-von Mises Goodness of Fit Test;

$$w_j = 1/j\pi, \varphi_j(u) = 2^{.5}\cos(j\pi u).$$

$$\int_0^1 \{D(u) - u\}^2 \, du = \sum_{j=1}^{\infty} w_j^2 \, [\phi_j, d]^2$$

Anderson Darling Goodness of Fit Test:

$$w_j = \{1/j(j+1)\}^{.5}, \quad \varphi_j(u) = (2j+1)^{.5} \quad p_j(2u-1),$$

$$\int_0^1 \left\{ \{D(u) - u\}^2 / u(1-u) \right\} du = \sum_{j=1}^{\infty} w_j^2 \, [\varphi_j, d]^2 ;$$

$p_j(t)$ are Legendre polynomials on [-1,1].

Hermite Polynomial Goodness of Fit Test:

$$w_j = 1/j, \quad \varphi_j(u) = (j!)^{-5} H_j \left(\Phi^{-1}(u) \right);$$

$H_j(x)$ are Hermite polynomials.

9. Examples of One Sample and Multi-Sample Continuous Data Analysis:

National Bureau of Standards NB10 Measurements: Freedman, Pisani, Purves in their textbook on Statistics (p.94) report 100 measurements of the 10 gram check-weight NB10 made at the National Bureau of Standards. They report: "The normal curve does not fit at all well. The normal curve does fit the data with three outliers removed. The normal curve fitted to these measurements has an average of 404 micrograms below 10 grams, and a standard deviation of about 4 micrograms. But in a small percentage of cases, the measurements are quite a bit farther away from the average than the normal curve suggests. The overall standard deviation of 6 micrograms is a compromise between the standard deviation of the main part of the histogram (4 micrograms) and the three outliers, representing deviations of 18, -30, and 32 micrograms. In careful measurement work, a small percentage of outliers is expected. The only unusual aspect of the NB10 data is that the National Bureau of Standards reported its outliers; many investigators don't. Realistic performance parameters require the acceptance of all data that cannot be rejected for cause."

The NB10 data illustrates the statistical analysis strategy that we propose be routinely applied to data. Step 1. Specify a parametric probability model for the data (here the model is normal). Step 2. Estimate parameters of the model (here mean and standard deviation) to be 10 grams-404 micrograms and 6 micrograms respectively. Step 2*. Robust parameter estimation by Renyi information of index between 0 and 1 obtains as estimators of a normal model (fitted to the part of the data that can be well fitted by a normal model) the same mean and a standard deviation of 4 micrograms. Step 4: Goodness of fit test of normality by traditional tests. Step 5: Maximum entropy estimator of comparison density $d(u$; normal model, data) clearly indicates the nature of the data; a poor fit of normal model to data. Shape of $d\hat{}(u)$ in interior of interval (0,1) can be interpreted as expected curve if $d\hat{}(u)$ estimates

$$\begin{aligned}&d(u; N(0,(6)^2), N(0,(4)^2)) \\ &= \kappa \exp\left\{-.5\left(\kappa^2-1\right)\left(\Phi^{-1}(u)\right)^2\right\}, \kappa = 6/4.\end{aligned}$$

Peaks of $d\hat{}(u)$ at $u = 0, 1$ indicate longer tails than normal. In general, one must decide whether to consider these tails in $d\tilde{}(u)$ as outliers or as evidence that a longer tailed distribution than the normal should be used to model the data. In Figure 2 two graphs illustrate the comparison density estimation process: the raw estimator $d\tilde{}(u)$ superimposed on a smooth estimator $d\hat{}(u)$; the exponential model smooth estimator $d_{0,4}\hat{}$, the orthogonal polynomial estimator $d_{1,4}\hat{}$, and a naive step function estimator d^* representing increments of $D\tilde{}(u)$ on 8 equal subintervals. Diagnostic tools at step 1 which help identify probability models for the data are illustrated by a IQQ plot of the sample quantile function of the data versus the quantile function of a normal with density $f(x) = \exp(-\pi x^2)$. The informative quantile function of the sample is defined $QI\tilde{}(u) = \{Q\tilde{}(u) - Q\tilde{}(.5)\}/2\{Q\tilde{}(.75) - Q\tilde{}(.25)\}$.

Breaking Stress of Beam: Cheng and Stephens (1989) give a data set of breaking stress of 41 beam specimens cut from a single carbon block of graphite H590, and discuss goodness of fit .ests of the hypothesis that the data is normal. Let $F(\cdot;\theta\hat{})$ denote the normal distribution with maximum likelihood estimated value of θ. They show that Moran's statistic, which is equivalent to $IR_0(d(u;F(\cdot;\hat{\theta}),\tilde{F})$ "correctly" rejects the hypothesis that the sample is normal, in contrast to more traditional empirical distribution based statistics (such as Kolmogorov-Smirnov and Cramer-von Mises) which accept the hypothesis of normality for the sample tested. The comparison density estimation approach indicates the nature of the data; an excellent fit of normal model in interior of interval (0,1) but peaks at $u = 0, 1$ indicate outliers or long tails (clearly evident in stem and leaf table of the data). One conjectures that a symmetric extreme value distribution would be a more appropriate model. Figure 3 illustrates the comparison density estimation process for a normal model $F(\cdot;\hat{\theta})$. The graph of $D(u;F(\cdot;\hat{\theta}),\tilde{F})$ is graphically well fitted by a uniform distribution, and therefore passes traditional goodness of fit tests. The raw estimator $d(u;F(\cdot;\hat{\theta}),\tilde{F})$ is superimposed on a smooth estimator. The exponential model smooth estimator $\hat{d}(u)$ is superimposed on a step function estimator computed from increments of $D(u;F(\cdot,\hat{\theta}),\tilde{F})$ over 8 sub-intervals.

Cheng and Stephens Break Stress Data
(Stem and Leaf)

27	.55							
28								
29	.89							
30	.07	.65						
31	.23	.53	.53	.82				
32	.23	.28	.69	.98				
33	.28	.28	.74	.74	.86	.86	.86	
34	.15	.15	.15	.44	.62	.74	.74	
35	.03	.03	.32	.44	.61	.61	.73	.90
36	.20	.78						
37	.07	.36	.36	.36				
38								
39								
40	.28							

Multisample of ratio of assessed value to sale price of residential property: To illustrate the comparison density approach to testing multil-samples for homogeneity, we consider data analysed by Boos (1986) on ratio of assessed value to sale price of residential property in Fitchburg, Mass., 1979. The samples (denoted I, II, III, IV) represent dwellings in the categories single-family, two-family, three-family, four or more families. The sample sizes (54, 43, 31, 28) are proportions .346, .276, .199, .179 of the size 156 of the pooled sample. We interpret these proportions as $p_{X}{}^{\sim}(k)$, $k = 1, \ldots, 4$. We compute Legendre, cosine, Hermite components $\{C_{j,k}{}^{\sim}\}^{.5}$ up to order 4 of the 4 samples; they are asymptotically standard normal. We consider components greater than 2 (3) in absolute value to be significant (very significant).

Legendre, cosine, and Hermite components are very significant only for sample I, order 1 (-4.06, -4.22, -3.56 respectively). Legendre components are significant for sample IV, orders 1 and 2 (2.19, 2.31). Cosine components are significant for sample IV, orders 1 and 2 (2.36, 2.23) and sample III, order 1 (2.05). Hermite components are significant for sample IV, orders 2 and 3 (2.7 and -2.07).

Conclusions are that the four samples are not homogeneous (have the same distributions). Samples I and IV are significantly different from the pooled sample. Estimators of the comparison density provide a substantive conclusion; they show that sample I is more likely to have lower values than the pooled sample, and sample IV is more likely to have higher values, suggesting that one family homes are underassessed and four family homes are overassessed, while two and three family homes are fairly assessed.

When one compares components with traditional empirical distribution based tests one concludes that the insights are provided by the linear rank statistics of orthogonal polynomials rather than by portmanteau statistics of Cramer-von Mises or Anderson-Darling type. Comparison density functions, which compare each sample with the pooled sample, can provide the most substantive information.

REFERENCES

Alexander, William (1989) "Boundary kernel estimation of the two-sample comparison density function" Texas A&M Department of Statistics Ph.D. thesis.

Aly, E. A. A., M. Csorgo, and L. Horvath (1987) "P-P plots, rank processes, and Chernoff-Savage theorems" in *New Perspectives in Theoretical and Applied Statistics* (ed. M. L. Puri, J. P. Vilaplann, W. Wertz) New York: Wiley 135–156.

Boos, Dennis D. (1986) "Comparing k populations with linear rank statistics", *Journal of the American Statistican Association*, 81, 1018–1025.

Cheng, R. C. H. and Stephens, M. A. (1989) "A goodness of fit test using Moran's statistic with estimated parameters", *Biometrika*, 76, 385–392.

Chui, C. K., Deutsch, F., Ward, J. D. (1990) "Constrained best approximation in Hilbert space," *Constructive Approximation*, 6, 35–64.

Eubank, R. L., V. N. LaRiccia, R. B. Rosenstein (1987) "Test statistics derived as components of Pearson's Phi-squared distance measure", *Journal of the American Statistical Association*, 82, 816–825.

Freedman, D., Pisani, R., Purves, R. (1978) *Statistics*, New York: Norton.

Parzen, E. (1979) "Nonparametric statistical data modelling", *Journal of the American Statistical Association*, 74, 105–131.

Parzen, E. (1989) "Multi-sample functional statistical data analysis," in *Statistical Data Analysis and Inference*, (ed. Y. Dodge). Amsterdam: North Holland, pp. 71–84.

Shorack, Galen and John Wellner (1986) *Empirical Processes With Applications to Statistics*, New York: Wiley.

Figure 1

To understand the shapes of comparison density functions, graphs of $d(u;G,F)$ and $d(u;F,G)$ for two cases. Case 1: F normal (median 0, density at median 1), G Cauchy (median 0, density at median 1). Case 2: F normal (median 0, density at median 1), G symmetric extreme value (median 0, density at median 1).

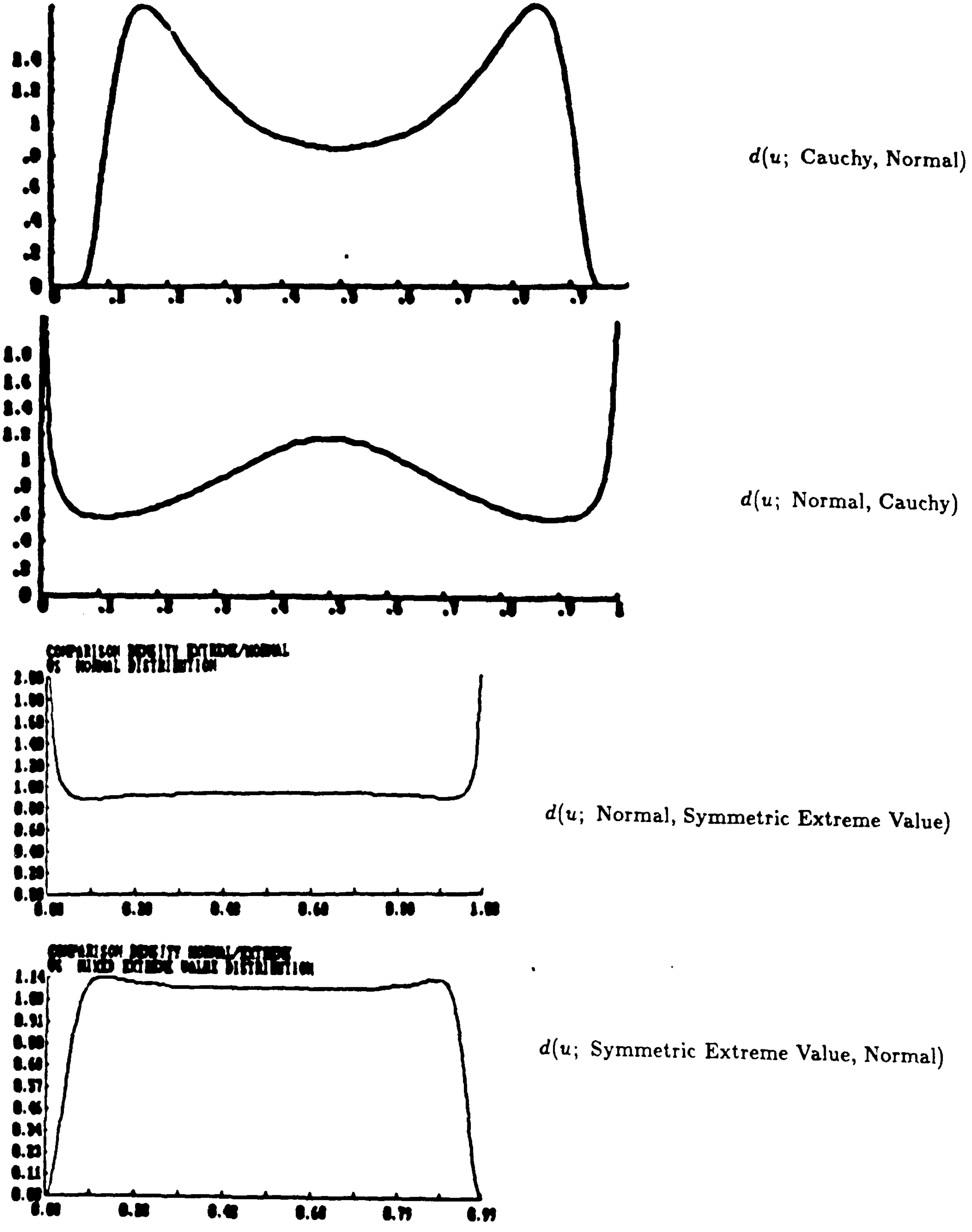
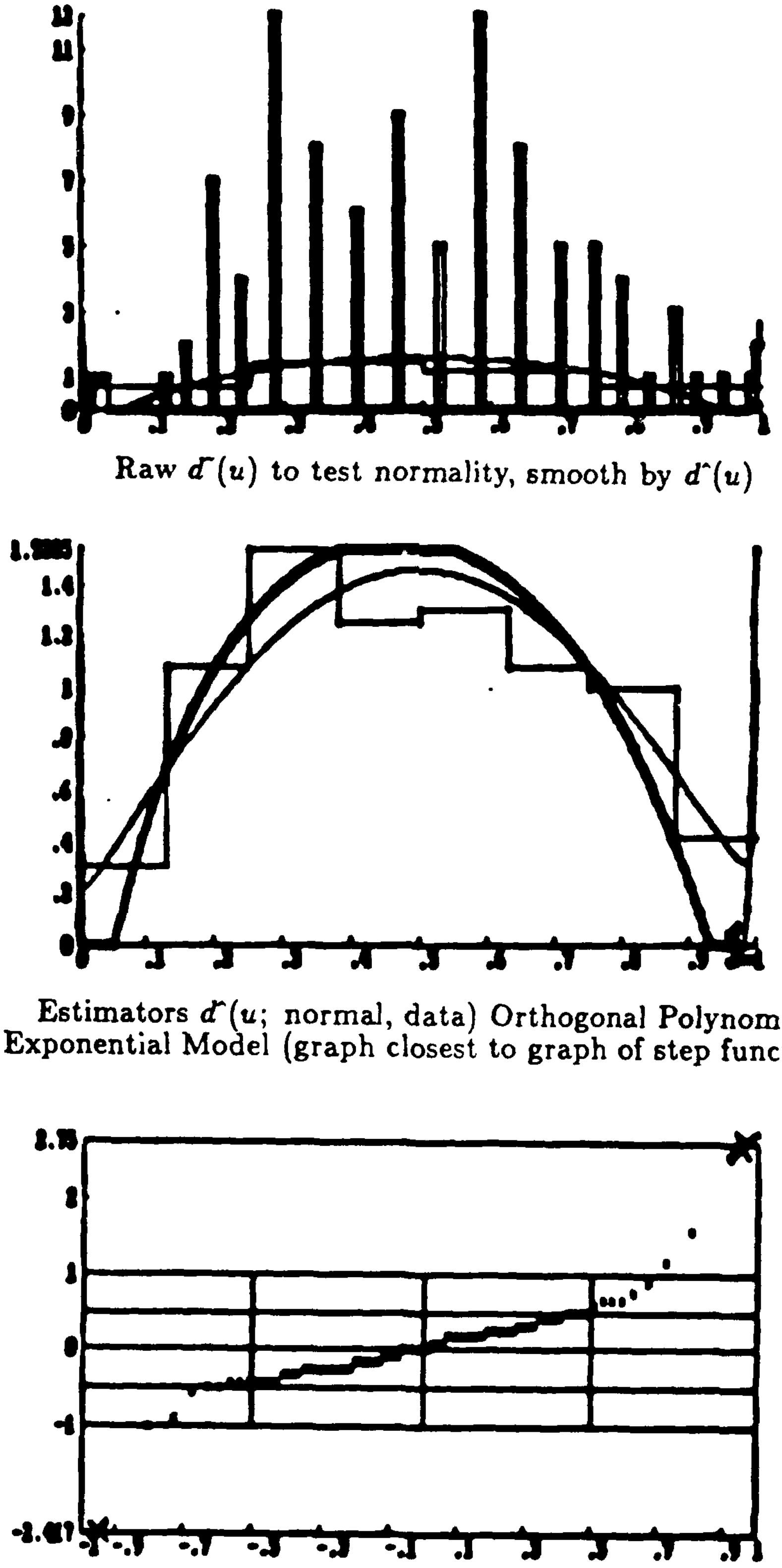


Figure 2
Test NB10 Measurements for Normality

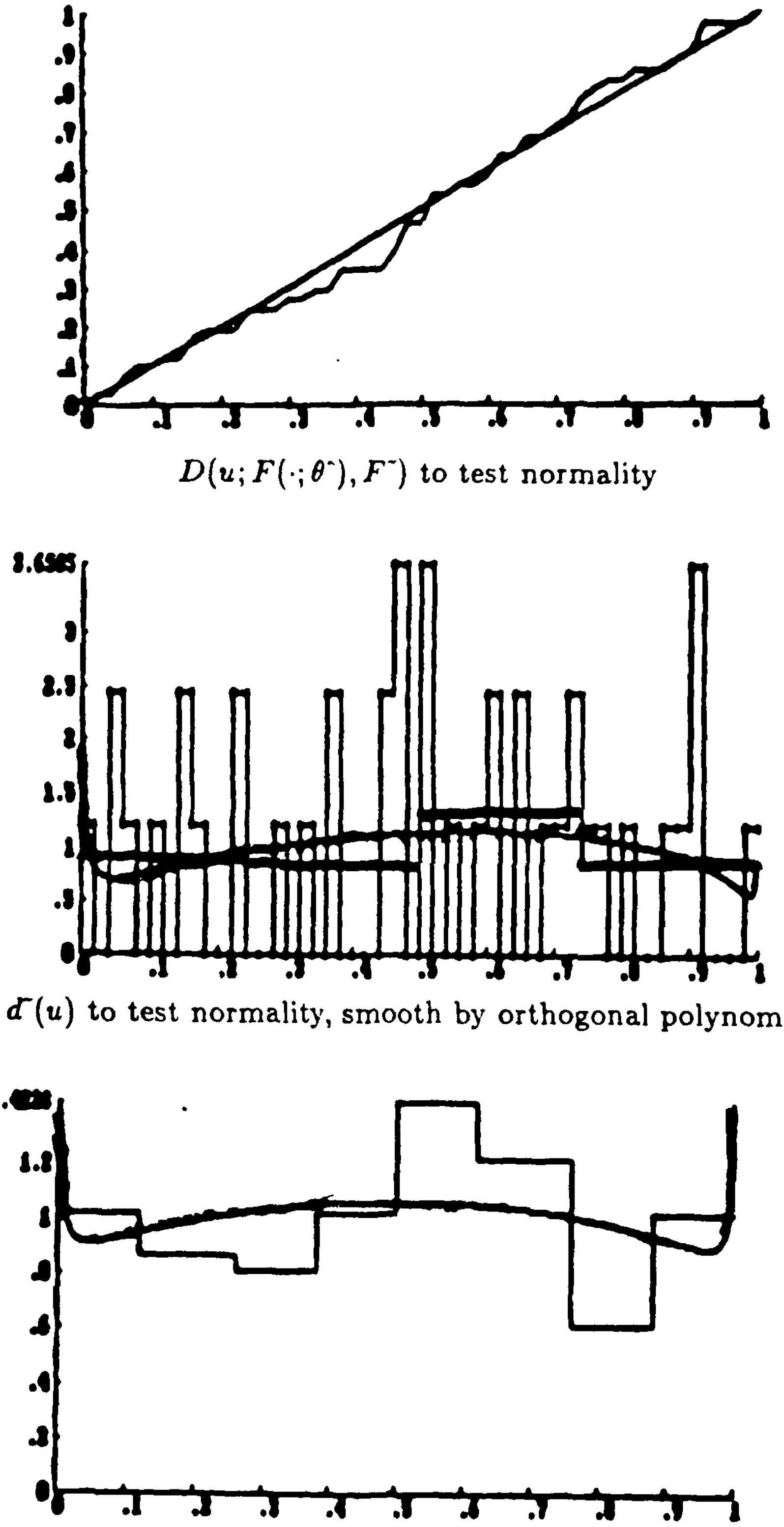


Raw $\tilde{d}(u)$ to test normality, smooth by $\hat{d}(u)$

Estimators $\tilde{d}(u;$ normal, data$)$ Orthogonal Polynomial; Exponential Model (graph closest to graph of step function):

IQQ Plot $(Q_0I, \tilde{Q}I)$ Q_0 Normal

Figure 3
Test Breaking Stress Measurements for Normality



$D(u;F(\cdot;\theta^{\wedge}),F^{\sim})$ to test normality

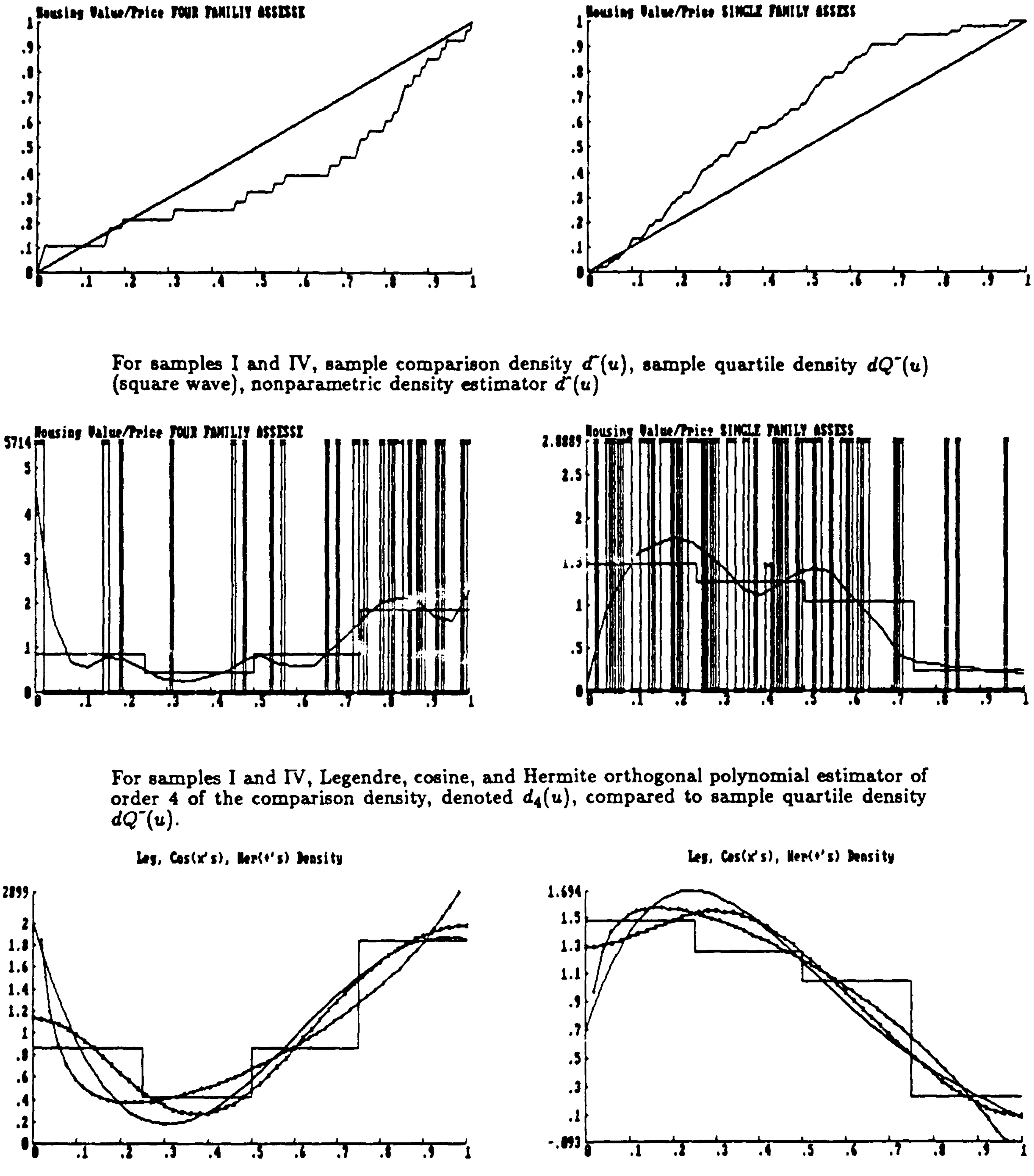
Raw $d^{\sim}(u)$ to test normality, smooth by orthogonal polynomial $d^{\wedge}(u)$

Estimation $d^{\wedge}(u$; normal, data) Exponential Model,
Step function

Figure 4

Ratio of assessed price to sale price of residential property

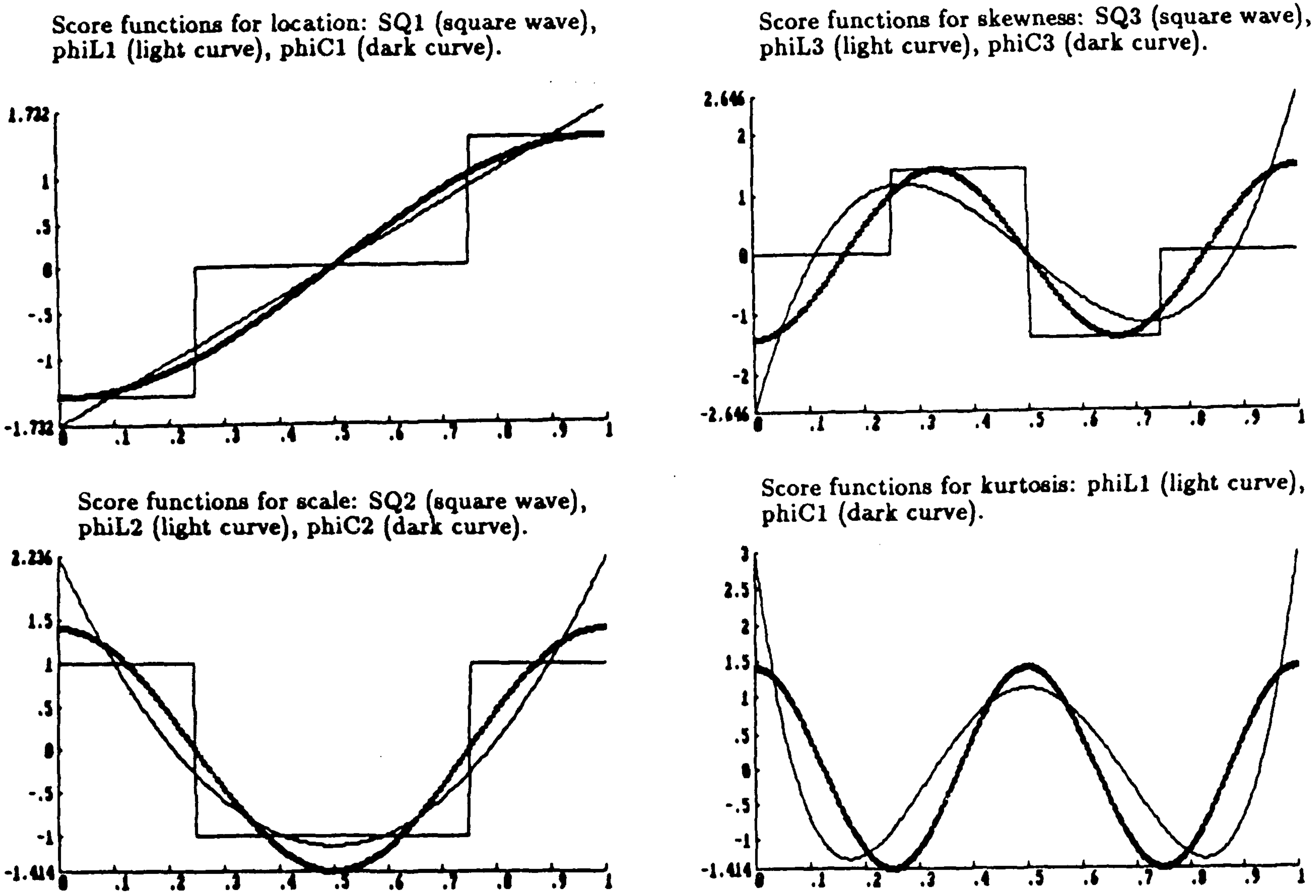
For samples I and IV, sample comparison distribution function $D^{\sim}(u)$



For samples I and IV, sample comparison density $d^{\sim}(u)$, sample quartile density $dQ^{\sim}(u)$ (square wave), nonparametric density estimator $\hat{d}^{\sim}(u)$

For samples I and IV, Legendre, cosine, and Hermite orthogonal polynomial estimator of order 4 of the comparison density, denoted $d_4(u)$, compared to sample quartile density $dQ^{\sim}(u)$.

Figure 5
Score Functions



Score functions for location: SQ1 (square wave), phiL1 (light curve), phiC1 (dark curve).

Score functions for skewness: SQ3 (square wave), phiL3 (light curve), phiC3 (dark curve).

Score functions for scale: SQ2 (square wave), phiL2 (light curve), phiC2 (dark curve).

Score functions for kurtosis: phiL1 (light curve), phiC1 (dark curve).

REFERENCES

Alexander, William (1989) "Boundary kernel estimation of the two-sample comparison density function" Texas A&M Department of Statistics Ph.D. thesis.

Aly, E. A. A., M. Csorgo, and L. Horvath (1987) "P-P plots, rank processes, and Chernoff-Savage theorems" in *New Perspectives in Theoretical and Applied Statistics* (ed. M. L. Puri, J. P. Vilaplann, W. Wertz) New York: Wiley 135–156.

Boos, Dennis D. (1986) "Comparing k populations with linear rank statistics", *Journal of the American Statistican Association*, 81, 1018–1025.

Cheng, R. C. H. and Stephens, M. A. (1989) "A goodness of fit test using Moran's statistic with estimated parameters", *Biometrika*, 76, 385–392.

Chui, C. K., Deutsch, F., Ward, J. D. (1990) "Constrained best approximation in Hilbert space," *Constructive Approximation*, 6, 35–64.

Eubank, R. L., V. N. LaRiccia, R. B. Rosenstein (1987) "Test statistics derived as components of Pearson's Phi-squared distance measure", *Journal of the American Statistical Association*, 82, 816–825.

Freedman, D., Pisani, R., Purves, R. (1978) *Statistics*, New York: Norton.

Parzen, E. (1979) "Nonparametric statistical data modelling", *Journal of the American Statistical Association*, 74, 105–131.

Parzen, E. (1989) "Multi-sample functional statistical data analysis," in *Statistical Data Analysis and Inference*, (ed. Y. Dodge). Amsterdam: North Holland, pp. 71–84.

Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness of Fit Statistics for Discrete Multivariate Data*, Springer Verlag, New York.

Renyi, A. (1961). "On measures of entropy and information." *Proc. 4th Berkeley Symp. Math. Statist. Probability, 1960*, 1, 547-461. University of California Press: Berkeley.

Shorack, Galen and John Wellner (1986) *Empirical Processes With Applications to Statistics*, New York: Wiley.