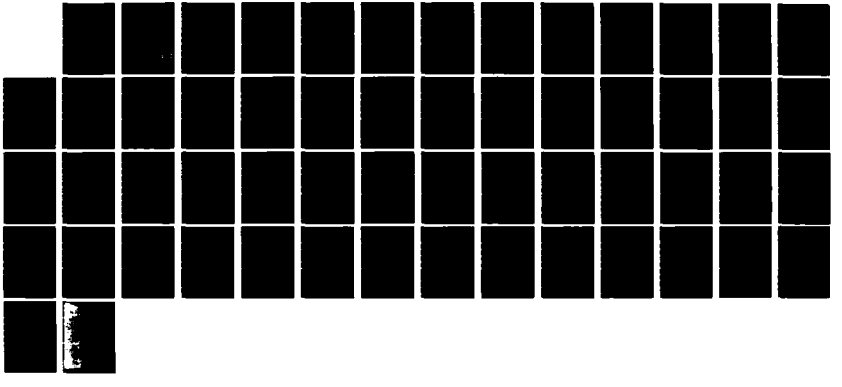
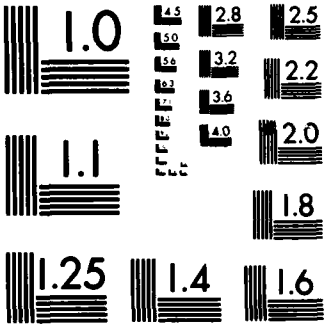
AD-A127 672 FUNSTAT QUANTILE APPROACH TO TWO SAMPLE STATISTICAL DATA ANALYSIS(U) TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS E PARZEN APR 83 TR-A-21 1/1

UNCLASSIFIED ARO-20140.1-MA DAAG29-83-K-0051 F/G 12/1 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

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

STATISTICS
Phone 409 - 845-3141

FUN.STAT QUANTILE APPROACH TO TWO SAMPLE STATISTICAL DATA ANALYSIS

by Emanuel Parzen

Institute of Statistics

Texas A&M University

Technical Report No. A-21

April 1983

Texas A&M Research Foundation
Project No. 4858

"Functional Statistical Data Analysis and Modeling"

Sponsored by the U. S. Army Research Office
Grant DAAG29-83-K-0051

Professor Emanuel Parzen, Principal Investigator

Approved for public release; distribution unlimited.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER A-21	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) FUN.STAT Quantile Approach to Two Sample Statistical Data Analysis		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Emanuel Parzen		8. CONTRACT OR GRANT NUMBER(s) DAAG29-83-K-0051
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 April 1983 13. NUMBER OF PAGES 55
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)

Two sample non-parametric tests, linear rank statistics, location-scale models, data analysis, convergence in distribution, density estimation, autoregressive density estimation, functional statistical inference, FUN.STAT, TWOSAM.

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

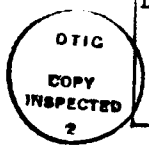
FUN.STAT is a name proposed by us to describe a synthesis of statistical reasoning which combines quantiles and quantile-densities, information and entropy, and functional statistical inference. This paper describes a FUN.STAT approach to the problem of statistical data analysis of two random samples, respectively representing two populations of interest. It is composed of four parts.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)



FUN.STAT QUANTILE APPROACH TO TWO SAMPLE STATISTICAL DATA ANALYSIS

by

Emanuel Parzen
Institute of Statistics
Texas A&M University

Abstract

FUN.STAT is a name proposed by us to describe a synthesis of statistical reasoning which combines quantiles and quantile-densities, information and entropy, and functional statistical inference. This paper describes a FUN.STAT approach to the problem of statistical data analysis of two random samples, respectively representing two populations of interest. It is composed of four parts. Part 1 describes how conventional approaches to two sample problems, including representations of linear rank statistics, are equivalent to functionals of a stochastic process $\tilde{D}(u)$. Part 2 motivates the autoregressive density estimation approach to the problem of functional statistical inference of $\tilde{D}(u)$ and states several conjectures concerning the properties of the density estimation approach. Part 3 outlines heuristic derivations of the asymptotic distribution theory of $\tilde{D}(u)$. Part 4 provides a summary and an example, using TWOSAM which is a computer program for autoregressive two sample statistical data analysis; it has been implemented as a Fortran program and as a SAS procedure.

Keywords: Two sample non-parametric tests, linear rank statistics, location-scale models, data analysis, convergence in distribution, density estimation, autoregressive density estimation, functional statistical inference, FUN.STAT, TWOSAM.

Research supported by U. S. Army Research Grant DAAG 29-83-K-0051.

FUN.STAT QUANTILE APPROACH TO

TWO SAMPLE STATISTICAL DATA ANALYSIS

Part 1. Two Sample Problems as Data Analysis of Stochastic Process $\tilde{D}(u)$, $0 \leq u \leq 1$

1.1. Univariate: one sample problem

The univariate: one sample problem of statistical data analysis considers a random sample $X_1, \ldots, X_n$ of a continuous random variable X and seeks to infer its probability law.

In the quantile approach to the study of the probability law of a random variable X, the functions to be estimated are [Parzen (1979)]:

distribution function $F(x) = \Pr[X \leq x]$, $-\infty < x < \infty$;

quantile function $Q(u) = F^{-1}(u) = \inf\{x: F(x) \geq u\}$, $0 \leq u \leq 1$;

probability density function $f(x) = F'(x)$, $-\infty < x < \infty$;

quantile-density $q(u) = Q'(u)$, $0 \leq u \leq 1$;

density-quantile $fQ(u) = f(F^{-1}(u)) = \{q(u)\}^{-1}$, $0 \leq u \leq 1$;

score-function $J(u) = -(fQ)'(u) = -f'F^{-1}(u)/fF^{-1}(u)$, $0 \leq u \leq 1$.

When F is continuous, $FF^{-1}(u) = u$. When F^{-1} is continuous, $F^{-1}F(x) = x$. We call X (or F) bi-continuous if both F and F^{-1} are continuous functions. When F is bi-continuous, then F^{-1} is a true inverse in the sense that $F(x) = u$ if and only if $F^{-1}(u) = x$.

To estimate a function [for example, $D(u)$, $0 \leq u \leq 1$] three types of estimators should be distinguished: fully parametric [denoted $\hat{\hat{D}}(u)$]; fully non-parametric [denoted $\tilde{D}(u)$]; and functional parametric [denoted $\hat{D}(u)$]. These types of estimators have the following characteristics: (1) fully non-parametric makes almost no assumptions about a model for the function; (2) fully parametric estimates the parameters of an assumed finite-parametric model for the function; and (3) functional parametric estimates the parameters of an approximating parametric model whose order is determined (selected) by the data.

Fully non-parametric estimators of $F(x)$ and $F^{-1}(u)$ [given a random sample $X_1,\ldots,X_n$ with order statistics $X_{(1)}<\ldots<X_{(n)}$] are defined by the sample distribution function $\tilde{F}(x)$ and the same quantile function $\tilde{F}^{-1}(u)$. We systematically use $\tilde{}$ to denote a sample function which is a raw (or fully non-parametric) estimator of a function defined on the ensemble or population.

Definition. Sample distribution $\tilde{F}$ and sample quantile $\tilde{F}^{-1}$. For a sample $X_1,\ldots,X_n$, $\tilde{F}$ and $\tilde{F}^{-1}$ are piecewise constant functions satisfying for $j=0,1,\ldots,n$

$$\tilde{F}(x) = \frac{j}{n} , \qquad X_{(j)} \leq x < X_{(j+1)} \quad ;$$

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

where $$X_{(0)} = -\infty, \; X_{(n+1)} = \infty.$$

A basic question of the univariate: one sample theory is the goodness of fit problem: to test the null hypothesis H_0: $F(x) = F_0(x)$, where $F_0(x)$ is a specified continuous distribution function. The mathematical statistician is concerned with finding the exact and asymptotic distributions under both null and alternative hypotheses of statistics such as

$$D_n = \sup_{-\infty<x<\infty} |\tilde{F}(x) - F_0(x)|$$

$$W_n^2 = \int_{-\infty}^{\infty} \{\tilde{F}(x) - F_0(x)\}^2 \, dF_0(x)$$

By a formal change of variable $x = Q_o(u) = F_o^{-1}(u)$ one can write

$$D_n = \sup_{0<u<1} |\tilde{F}F_o^{-1}(u)-u|$$

$$W_n^2 = \int_o^1 \{\tilde{F}F_o^{-1}(u) - u\}^2 \, du$$

One rigorously obtains these formulas by interpreting $\tilde{F}F_o^{-1}(u)$ as the sample distribution function $\tilde{F}_U(u)$ of the random variables $U_1 = F_o(X_1),\ldots,U_n = F_o(X_n)$. The null hypotheses H_o is equivalent to the hypothesis that $U = F_o(X)$ is uniform on $[0,1]$.

Alternative to testing the sample distribution function for uniformity, one could test for uniformity the sample quantile function $\tilde{F}_U^{-1}(u) = \tilde{Q}_U(u) = F_o\tilde{F}^{-1}(u)$.

A Brownian Bridge process $B(u)$, $0 \le u \le 1$ is a zero mean Gaussian process with covariance kernel

$$E[B(u_1)B(u_2)] = \min(u_1,u_2) - u_1u_2 \quad .$$

The asymptotic distribution of test statistics such as D_n and W_n^2 is based on the convergence theorems [as sample size n tends to ∞]: assuming H_o,

$$\tilde{B}_F(u) = \sqrt{n}\,\{\tilde{F}F_o^{-1}(u) - u\} \xrightarrow{D} B_F(u) \quad ,$$

$$\tilde{B}_{F^{-1}}(u) = \sqrt{n}\,\{F_o\tilde{F}^{-1}(u) - u\} \xrightarrow{D} - B_F(u)$$

where $\xrightarrow{D}$ denotes convergence in distribution of stochastic processes and $B_F(u)$ is a limiting process distributed as a Brownian Bridge process.

A quick and dirty definition of convergence in distribution of stochastic processes is as follows:

$$\{\tilde{B}_F(u),\ 0\le u\le 1\} \xrightarrow{D} \{B_F(u),\ 0\le u\le 1\}$$

if and only if for every bounded continuous functional $g(x(u),\ 0\le u\le 1)$ on a suitable metric space of functions $x(u)$, $0\le u\le 1$,

$$E[g(\tilde{B}_F(u),\ 0\le u\le 1)] \to E[g(B_F(u),\ 0\le u\le 1)].$$

Testing for uniformity a random sample $U_1,\ldots,U_n$ of points on the unit interval is a canonical problem of statistics in the sense that many other statistical problems can be transformed to this problem. One way to determine the appropriate transformation is by attempting to find the analogues of the test statistics D_n and W_n^2. To develop such analogues, one might compare computational formulas.

Computational formulas for D_n and W_n^2 [which are well known in the theory of goodness of fit tests, see Durbin (1973)] can be stated in terms of a general distribution function $\tilde{D}(u)$, $0\le u\le 1$ defined in terms of a set of specified constants U_j, where $0 = U_0 < U_1 < \ldots < U_n < U_{n+1} = 1$; define

$$\tilde{D}(u) = 0 \qquad 0 = U_0 \le u < U_1 \quad ;$$

$$= \frac{j}{n} \quad , \qquad U_j \le u < U_{j+1}$$

$$= 1 \quad , \qquad U_n \le u < U_{n+1} = 1 \quad .$$

Then

$$D_n = \max_{0\le u\le 1} |\tilde{D}(u)-u| = \max_{1\le j\le n} \left(\frac{j}{n} - U_j,\ U_j - \frac{j-1}{n}\right) \quad ;$$

$$W_n^2 = \int_0^1 \{\tilde{D}(u) - u\}^2\, du = \frac{1}{12n^2} + \frac{1}{n}\sum_{j=1}^{n} \left\{U_j - \frac{2j-1}{2n}\right\}^2$$

1.2. Univariate: two sample problem

The univariate: two sample problem of statistical data anlaysis considers random samples $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ respectively representing measurements on random variables X and Y with continuous distribution functions $F(x)$ and $G(x)$. We interpret X and Y as measurements of a physical variable in two different populations. The null hypothesis $H_0: F(x)=G(x)$ of equality of distributions is to be tested, if possible without specifying the distributional shapes. This paper assumes that F and G are bi-continuous (that is, F, G, F^{-1}, G^{-1} are continuous functions).

The random sample $X_1,\ldots,X_m$ has sample distribution $\tilde{F}$ and quantile $\tilde{F}^{-1}$. The random sample $Y_1,\ldots,Y_n$ has sample distribution $\tilde{G}$ and quantile $\tilde{G}^{-1}$. We denote by $\tilde{H}$ the sample distribution of the pooled sample $X_1,\ldots,X_m$, $Y_1,\ldots,Y_n$. It can be represented

$$\tilde{H}(x) = \lambda\tilde{F}(x) + (1-\lambda)\,\tilde{G}(x)$$

defining $N = m+n$ and $\lambda = m/N$. The limit of $\tilde{H}(x)$ is $H(x) = \lambda F(x) + (1-\lambda)\,G(x)$. We assume that as N tends to ∞, λ tends to a limit satisfying $0<\lambda<1$.

Techniques in the two-sample problem which are close counterparts of one sample techniques are the statistics

$$D_{mn} = \sup_{-\infty<x<\infty} |\tilde{F}(x) - \tilde{G}(x)| \quad ,$$

$$W^2_{mn} = \int_{-\infty}^{\infty} \{\tilde{F}(x) - \tilde{G}(x)\}^2 \, d\tilde{H}(x)$$

Durbin (1973) states computational formulas in terms of the ranks R_j, $j=1,\ldots,m$, in the pooled sample of the j-th largest observation in the X-sample. In our notation these formulas become

$$D_{mn} = \frac{1}{1-\lambda} \max_{j=1,\ldots,m} \left(\frac{j}{m} - \frac{R_j}{N} , \frac{R_j - 1}{N} - \frac{j-1}{m} \right)$$

$$W^2_{mn} = \frac{1}{(1-\lambda)^2} \frac{1}{m} \sum_{j=1}^{m} \left\{ \frac{R_j}{N} - \frac{2j-1}{2m} \right\}^2 + \frac{1}{12m^2} \frac{2m+n}{n}$$

By comparing these formulas with the general computational formulas in the one sample case one sees that the test statistics D_{mn} and W^2_{mn} are related to the statistics

$$\frac{1}{1-\lambda} \max_{0<u<1} |\tilde{D}(u)-u| , \frac{1}{(1-\lambda)^2} \int_0^1 \{\tilde{D}(u)-u\}^2 \, du$$

defining $\tilde{D}(u)$, $0 \leq u \leq 1$, as follows:

$$\tilde{D}(u) = 0 , \quad 0 \leq u < \frac{R_1}{N} ;$$

$$= \frac{j}{m} , \quad \frac{R_j}{N} \leq t < \frac{R_{j+1}}{N} ;$$

$$= 1 , \quad \frac{R_m}{N} \leq t < 1.$$

One aim of this paper is: (1) to relate the process $\tilde{D}(u)$ to the processes $\tilde{F}$ and $\tilde{H}^{-1}$, (2) to relate $\tilde{D}(u)$ to representations of linear rank statistics, (3) to relate $\tilde{D}(u)$ to quantiles, and (4) to use $\tilde{D}(u)$ graphically for a complete data analysis of the null hypothesis H_0.

1.3 Representations of $\tilde{D}(u)$

We have defined R_j to be the rank in the pooled sample of $X_{(j)}$, the j-th order statistic in the X sample. A more precise definition of rank is defined in terms of the sample distribution $\tilde{H}$:

$$R_j = N\tilde{H}(X_{(j)}). \qquad j=1,\ldots,m$$

Another insight into the definition of rank is provided by the formula, of later use,

$$\tilde{H}^{-1}(u) = X_{(j)} \quad , \frac{R_j-1}{N} < u \leq \frac{R_j}{N} \quad ;$$

note $H^{-1}(u)$ equals the k-th order statistic in the pooled sample for $\frac{k-1}{N} < u \leq \frac{k}{N}$.

The null hypothesis H_o: $F(x) = G(x)$ is often tested by means of linear rank statistics of the form

$$T_N = \frac{1}{m} \sum_{j=1}^{m} J(\frac{R_j}{N+1})$$

where $J(u)$ is a suitable weight function called a score function. Some frequently suggested score functions are listed in Table 1A.

Table 1A

Score functions for two-sample linear rank statistics

Test for Location Difference	Test for Scale Difference
$\Phi^{-1}(u)$ Normal scores test	$\lvert\Phi^{-1}(u)\rvert^2$
$u - 0.5$ Wilcoxon-Mann-Whitney Test	$(u - 0.5)^2$ Mood test
Sign $(u-0.5)$ Median test	Sign $\{\lvert u - 0.5 \rvert - 0.25\}$ Quantile test
	$\lvert u - 0.5 \rvert - 0.25$ Ansari-Bradley test

To study asymptotically the distribution of T_N various representations have been introduced. The celebrated Chernoff-Savage (1958) representation of a linear rank statistic T_N is

$$T_N = \int_{-\infty}^{\infty} J_N \left(\frac{N}{N+1} \tilde{H}(x)\right) \, d\tilde{F}(x)$$

The Pyke-Shorack (1968) representation is

$$T_N = \int_0^1 \tilde{\tilde{FH}}^{-1}(u) \, d\nu_N(u) = \sum_{i=1}^{N} \tilde{\tilde{FH}}^{-1}\left(\frac{i}{N}\right) \left\{\nu_N\left(\frac{i}{N}\right) - \nu_N\left(\frac{i-1}{N}\right)\right\}$$

for a suitable signed measure ν_N. Chernoff-Savage establish directly a limit theorem for T_N, while Pyke-Shorack derive the convergence properties of T_N from the convergence properties of the process $\tilde{\tilde{FH}}^{-1}(u)$, $0 \le u \le 1$ [see Pyke (1970)].

The functional statistical inference approach proposed in this paper is based on the proposition that $\tilde{\tilde{FH}}^{-1}$, the Pyke-Shorack two-sample process,

may be appropriate for theorem proving but is not directly useable for exploratory data analysis. The role of the score function J(u) is preserved, and functions whose graphs are suitable for data analysis are obtained, by using

$$\tilde{D}_1(u) = \tilde{H}\tilde{F}^{-1}(u), \text{ estimator of } D_1(u) = HF^{-1}(u) \quad ;$$

$$\tilde{D}(t) = \tilde{D}_1^{-1}(t), \text{ estimator of } D(t) = D_1^{-1}(t) = FH^{-1}(t).$$

Note that $\tilde{D}_1^{-1}(t)$ is a right continuous function which is the inverse of the left continuous function $\tilde{D}_1(u)$; it is defined by

$$\tilde{D}_1^{-1}(t) = \sup \{u: \tilde{D}_1(u) \leq t\}$$

Theorem 1B shows that $\tilde{D}(u)$ is computationally exactly the same as the process $\tilde{D}(u)$ in terms of which we approximately expressed D_{mn} and W_{mn}^2. Theorem 1A is further evidence that one can introduce a process $\tilde{D}(u)$ such that many conventional two sample statistics are functionals of this process. The univariate: two sample problem is thus transformed to a problem of statistical inference from a continuous parameter process $\tilde{D}(u)$, $0 \leq u \leq 1$. We call this a problem of functional statistical inference.

The branch of statistical theory which we call "functional (statistical) inference" (FUN.STAT) is a branch of "abstract inference" [Grenander (1981)].

Theorem 1A: Functional Representations of Linear Rank Statistic

$$T_N = \frac{1}{m} \sum_{j=1}^{m} J\left(\frac{R_J}{N+1}\right)$$

$$= \int_0^1 J\left(\frac{N}{N+1}\right) \tilde{D}_1(u)) \, du$$

$$= \int_0^1 J\left(\frac{N}{N+1} t\right) d\tilde{D}(t) \quad .$$

Proof: Define $J_N(u) = J\left(\frac{N}{N+1} u\right)$. In the Chernoff-Savage representation $T_N = \int_0^1 J_N(\tilde{H}(x)) \, d\tilde{F}(x)$ make the change of variable $u = \tilde{F}(x)$ to obtain $T_N = \int_0^1 J_N(\tilde{D}_1(u))du$. The change of variable $t = \tilde{D}_1(u)$ completes the proof.

Theorem 1B: Explicit Formulas for $\tilde{D}_1(u)$ and $\tilde{D}(t)$.

$\tilde{D}_1(u)$ $0 \leq u \leq 1$, is piecewise-constant, non-decreasing, left continuous, and satisfies

$$\tilde{D}_1(u) = \frac{R_j}{N} \, , \quad \frac{j-1}{m} < u \leq \frac{j}{m} \quad , \quad j=1,\ldots,m.$$

$\tilde{D}(t) = \tilde{D}_1^{-1}(t)$ is piecewise-constant, non-decreasing, right continuous, and satisfies $\tilde{D}(0) = 0$,

$$\tilde{D}\left(\frac{R_j}{N}\right) = \frac{j}{m} \quad , \quad j=1,\ldots,m \quad .$$

More precisely,

$$\tilde{D}\ (t) = 0 \quad , \quad 0 \leq t < \frac{R_1}{N} \quad ;$$

$$\tilde{D}\ (t) = \frac{j}{m} \quad , \quad \frac{R_j}{N} \leq t < \frac{R_{j+1}}{N} \quad , \qquad j=1,\ldots,m-1 \ ;$$

$$\tilde{D}\ (t) = 1, \qquad \frac{R_m}{N} \leq t < 1.$$

The Pyke-Shorack process $\tilde{F}\tilde{H}^{-1}(u)$ is given by

$$\tilde{F}\tilde{H}^{-1}(u) = \frac{j}{m} \quad , \qquad \frac{R_j-1}{N} < u \leq \frac{R_{j+1}-1}{N} \qquad ,$$

$$\tilde{F}\tilde{H}^{-1}(u) = 0 \quad , \qquad 0 < u \leq \frac{R_j-1}{N} \qquad ,$$

$$\tilde{F}\tilde{H}^{-1}(u) = 1 \quad , \qquad \frac{R_m-1}{N} < u \leq 1.$$

Example. Suppose $m=2$, $n=4$, $X_1=2$, $X_2=4$, $Y_1=1$, $Y_2=3$, $Y_3=5$, $Y_4=6$. Then $R_1=2$, $R_2=4$.

$$\tilde{D}(u) = \frac{2}{6} \ , \quad 0 < u \leq \frac{1}{2} \ ,$$

$$= \frac{4}{6} \ , \quad \frac{1}{2} < u \leq 1;$$

$$\tilde{D}\ (t) = 0 \ , \quad 0 \leq t < \frac{2}{6}$$

$$= \frac{1}{2} \qquad \frac{2}{6} \leq t < \frac{4}{6}$$

$$= 1 \qquad \frac{4}{6} \leq t < 1$$

$$\tilde{F}\tilde{H}^{-1}(u) = 0, \qquad 0 < u \leq \frac{1}{6}$$

$$= \frac{1}{2}, \qquad \frac{1}{6} < u \leq \frac{3}{6}$$

$$= 1, \qquad \frac{3}{6} < u \leq 1$$

The statistic

$$R_N = \frac{1}{m} \sum_{j=1}^{m} \frac{R_j}{N+1}$$

is the Wilcoxon statistic (up to a constant multiple); it corresponds to $J(t) = t$. The value of T_N is 3/7; it can be evaluated by the defining sum or by the representations

$$T_N = \int_0^1 \frac{6}{7} \tilde{D}(u)\, du = \frac{6}{7} \left[\frac{2}{6} \cdot \frac{1}{2} + \frac{4}{6} \cdot \frac{1}{2}\right] = \frac{3}{7}$$

$$T_N = \int_0^1 \frac{6}{7}\, t\, d\tilde{D}\ (t) = \frac{6}{7} \left[\frac{2}{6} \cdot \frac{1}{2} + \frac{4}{6} \cdot \frac{1}{2}\right] = \frac{3}{7}$$

The Pyke-Shorack representation would be evaluated

$$\frac{1}{2} \{\nu_N(\tfrac{3}{6}) - \nu_N(\tfrac{1}{6})\} + 1 \{\nu_N(1) - \nu_N(\tfrac{3}{6})\}$$

if one bothered to discover the values of the measure ν_N corresponding to $J(u)$.

Part 2. Functional Statistical Inference Approach to Two Sample Problem

2.1 Introduction

Part 1 has attempted to show that most conventional test statistics in the "univariate: two sample problem" are functionals of a stochastic process $\tilde{D}(u)$, $0 \leq u \leq 1$, and proposes that the problem of statistical inference should be posed as follows: what can we learn from a sample path of the stochastic process $\tilde{D}(u)$, $0 \leq u \leq 1$, assuming that it is the sum of a signal $D(u) = FH^{-1}(u)$ and a noise represented $C_D(u)/\sqrt{N}$:

$$\tilde{D}(u) = D(u) + \frac{1}{\sqrt{N}} C_D(u)$$

The covariance kernel of $C_D(u)$ in general is a function of the following unknown functions (which it is our goal to estimate)

$$D_F(u) = FH^{-1}(u), \qquad D_G(u) = GH^{-1}(u),$$

$$d_F(u) = D_F'(u), \qquad d_G(u) = D_G'(u).$$

Note that $D_F(u) = D(u)$ and $\lambda D_F(u) + (1-\lambda)D_G(u) = u$.

Part 3 outlines a heuristic proof of the following conjecture.

Conjecture 2A. Covariance Kernel of $C_D(u)$. $E[C_D(u)\, C_D(v)]$ equals, for $u<v$,

$$(1-\lambda)^2 [\lambda^{-1} d_G(u)\, d_G(v)\, D_F(u)(1-D_F(v))$$

$$+ (1-\lambda)^{-1} d_F(u)\, d_F(v)\, D_G(u)\, (1-D_G(v))]$$

Distribution of $C_D(u)$ under H_0 and local alternatives H_1. Under H_0: $F = G$, $D_F(u) = D_G(u) = u$, $d_F(u) = d_G(u) = 1$, $C_D(u)$ has covariance kernel (for $u<v$)

$$(1-\lambda)^2 [\lambda^{-1} u(1-v) + (1-\lambda)^{-1} u(1-v)] = (\frac{1-\lambda}{\lambda}) u (1-v)$$

which is the covariance kernel of $(\frac{1-\lambda}{\lambda})^{\frac{1}{2}}B(u)$. When H_0 is true, or under alternative hypotheses H_1 under which H_0 is only "gently" not true (as opposed to "violently" not true), then [$\stackrel{D}{=}$ denotes equal in distribution]

$$C_D(u) \stackrel{D}{=} (\frac{1-\lambda}{\lambda})^{\frac{1}{2}} B(u)$$

When the parameter in a statistical model is a function the statistical inference techniques used are called functional statistical inference. By introducing $\tilde{D}(u)$, $0 \leq u \leq 1$, the two sample problem has been formulated as a problem of functional statistical inference (abbreviated FUN.STAT) in which the parameters to be estimated or tested are the function $D(u) = FH^{-1}(u)$, its density $d(u) = D'(u)$, and its Fourier transform

$$\rho(v) = \int_0^1 e^{2\pi iuv} dD(u) = \int_0^1 e^{2\pi iuv} d(u)\, du \quad , v=0, \pm 1, \pm 2, \ldots .$$

The hypothesis H_0 is equivalent to

$$H_0: D(u) = u; \quad d(u) = 1; \quad \rho(v) = 0 \text{ for } v \neq 0 \quad .$$

To understand what we can learn from d(u), let us relate it to the underlying densities f(x) and g(x); the derivative of $D(u) = FH^{-1}(u)$ equals

$$d(u) = \frac{fH^{-1}(u)}{hH^{-1}(u)} \quad .$$

Consequently, the reciprocal $d^{-1}(u)$ satisfies

$$d^{-1}(u) = \frac{hH^{-1}(u)}{fH^{-1}(u)} = \frac{\lambda fH^{-1}(u) + (1-\lambda)\, gH^{-1}(u)}{fH^{-1}(u)}$$

$$= \lambda + (1-\lambda)\, \frac{gH^{-1}(u)}{fH^{-1}(u)} \quad .$$

Therefore: $d(u) \leq \lambda^{-1}$; d(u) tends to 0 if f(x) tends to 0; d(u) tends to λ^{-1} if g(x) tends to 0. By estimating $d^{-1}(u)$, one can <u>estimate the likelihood ratio</u> g(x)/f(x) without estimating g(x) and f(x) separately.

An estimator $\hat{d}(u)$ of d(u) generates an estimator of D(u) by

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

To form an estimator $\hat{d}(u)$ from $\tilde{D}(u)$ it is often convenient to introduce first a raw estimator of $\rho(v)$ denoted $\tilde{\rho}(v)$.

The <u>sample pseudo-correlations</u> are defined by, for $v = 0, \pm 1, \ldots,$

$$\tilde{\rho}(v) = \int_0^1 e^{2\pi iuv}\, d\tilde{D}(u)$$

$$= \frac{1}{m} \sum_{j=1}^{m} \exp\ 2\pi iv(R_j/N) \quad .$$

They obey the model (for alternative hypotheses close to H_o)

$$\tilde{\rho}(v) = \rho(v) + \left(\frac{1-\lambda}{N\lambda}\right)^{\frac{1}{2}} \eta(v) \quad , \; v=0,\pm 1,\ldots$$

defining

$$\eta(v) = \int_o^1 e^{2\pi i u v} \, dB(u) \quad , \; v=0,\pm 1,\ldots;$$

one can show that $\eta(v)$ is a sequence of independent N(0,1) random variables.

A Brownian Bridge B(u) can be represented [see Csörgö and Révész (1981)]

$$B(u) = \sum_{v \neq 0} \eta(v) \int_o^u e^{-2\pi i v t} dt$$

Under H_o, $|\tilde{\rho}(v)|^2$ is asymptotically distributed as a sequence of independent random variables such that $\frac{2N\lambda}{1-\lambda} |\tilde{\rho}(v)|^2$ is chi squared distributed with 2 degrees of freedom. A 95% significance level for this statistic is 6.

To test H_o one could examine if any values of $|\tilde{\rho}(v)|^2$, v=1,2,..., exceeds a threshold such as $3(1-\lambda)/N\lambda$.

Natural "portmanteau" test statistics for H_o are of the form

$\sum_{v>1} k_N(v) \, |\tilde{\rho}(v)|^2$ for a suitable weight function $k_N(v)$. The optimal choice of weights $k_N(v)$ depend on the alternative hypothesis against which one is testing H_o. If one makes an arbitrary choice of weights [such as $k_N(v) = 1/v$ for v>1], then there will always exist alternative hypotheses against which the test statistics has efficiency close to 0. If one always uses for a goodness of fit test the statistic

$$\sum_{v=1}^{20} |\tilde{\rho}(v)|^2$$

one will too often accept H_o when it is false but only a few values of $\rho(v)$ are significantly non-zero; if one always uses the statistic

$$\sum_{v=1}^{4} |\tilde{\rho}(v)|^2 \quad .$$

one will too often accept H_o when it is false because $\rho(v) = 0$ for $v=1,\ldots,4$ but is non-zero for $v>4$. To achieve an "optimal portmanteau" test statistic, one might consider

$$\sum_{v=1}^{M} |\tilde{\rho}(v)|^2$$

where the order M is determined by the data. Insight into how to choose M is provided by density estimation.

2.2 Density Estimation, Kernels, and Windows

Some insight into the problem of optimally choosing weights $k_N(v)$ can be obtained by examining the density estimation problem in which one seeks to optimally choose a test or estimator based on the data. Estimation of the density $d(u)$, $0 \le u \le 1$, can be based on its Fourier series representation:

$$d(u) = \sum_{v=-\infty}^{\infty} e^{-2\pi iuv} \rho(v) \quad .$$

A raw estimator $\tilde{\rho}(v)$ generates a symbolic raw estimator

$$\tilde{d}(u) = \sum_{v=-\infty}^{\infty} e^{-2\pi iuv} \tilde{\rho}(v) \quad .$$

The series defining $\tilde{d}(u)$ is symbolic because it does not converge. A natural class of estimators $\hat{d}(u)$ of $d(u)$ are of the form, called kernel estimators,

$$\hat{d}(u) = \sum_{v=-\infty}^{\infty} k_N(v)\, e^{-2\pi iuv}\, \tilde{\rho}(v) = \int_0^1 K_N(u-t)\, d\tilde{D}(t)$$

defining

$$K_N(t) = \sum_{v=-\infty}^{\infty} e^{-2\pi itv}\, k_N(v), \quad k_N(v) = \int_{-0.5}^{0.5} e^{2\pi itv}\, K_N(t)\, dt.$$

We call $k_n(v)$ a kernel, and $K_N(t)$ a window. The theoretical investigation of these estimators in the context of the two sample problem is still very open for research [see conjecture 2B below].

Example. The choice of weights

$$k_N(v) = \frac{\sin 2\pi vh}{2\pi vh} = \frac{1}{2h} \int_{-h}^{h} e^{2\pi itv}\, dt$$

may be shown to be equivalent to the estimator of $d(u)$ given by [for $h \le u \le 1-h$]

$$\hat{d}_h(u) = \frac{\tilde{D}(u+h) - \tilde{D}(u-h)}{2h}$$

which we call a gap or leap estimator, or a numerical derivative.

Example. The density estimator $\hat{d}(u)$ can be motivated as a Bayes estimator of $d(u)$ given the data $\tilde{\rho}(\cdot)$. Let

$$\hat{d}(u) = E[d(u)|\tilde{\rho}(\cdot)], \quad \hat{\rho}(v) = E[\rho(v)|\tilde{\rho}(\cdot)].$$

Then

$$\hat{d}(u) = \sum_{v=-\infty}^{\infty} e^{-2\pi iuv}\, \hat{\rho}(v) \quad .$$

The prior distribution of $\rho(\cdot)$ is that $\rho(v)$, $v=0,\pm1,\ldots$ is a sequence of independent zero mean normal random variables with variance $E|\rho(v)|^2 = \theta\, C(v)$, where θ is a scalar parameter and $C(v)$ is a known convergent sequence. Under local alternatives H_1, and conditional on the value of $\rho(\cdot)$, we consider $\tilde{\rho}(v)$, $v=1,2,\ldots$ to be independent with mean $\rho(v)$ and variance $C_N=(1-\lambda)/N\lambda$. Then $\hat{\rho}(v) = k(v)\tilde{\rho}(v)$, where

$$k(v) = \frac{\mathrm{Var}\,[\rho(v)]}{\mathrm{Var}[\tilde{\rho}(v)]} = \frac{\theta\, C(v)}{C_N + \theta C(v)}$$

$$= \left\{1 + \frac{C_N}{\theta\, C(v)}\right\}^{-1} \quad .$$

An important family of weights of this form is

$$k(v) = \{1 + (\frac{v}{M})^{2r}\}^{-1}$$

where one has to choose the exponent r and the truncation (or half-power) point M adaptively from the data. This choice of weights can also be motivated by formulating the density estimation problem as an optimization problem: choose $\hat{d}(u)$ to minimize

$$\int_0^1 |\tilde{d}(u) - \hat{d}(u)|^2 + p \int_0^1 |\hat{d}^{(r)}(u)|^2 \, du$$

where p is a penalty parameter to be specified by the researcher.

In the general context of functional statistical inference, when a kernel estimator is motivated by an optimization problem of the foregoing kind, we call it a non-parametric penalty estimator. A density estimator is called parametric select when it is a function of a finite number of parameters and the number is chosen by the sample. Autoregressive density estimators [described in section 2.4] are parametric select estimators.

Conjecture 2B. The asymptotic distribution of kernel density estimators can be developed from the theory [outlined in Part 3] of functionals $\int_0^1 J(u) \, d\tilde{D}(u)$ and the representation $\hat{d}(u) = \int_0^1 K_N(u-t) \, d\tilde{D}(t)$. Let $k(x)$ - $-\infty<x<\infty$, be a kernel generating function, and take $k_N(v)$ to be of the form

$$k_N(v) = k(\frac{v}{M})$$

We call M a bandwidth lag or truncation point or half-power point [depending on the standardization of $k(\cdot)$]; it is a function of N, and tends to ∞ as N tends to ∞. Then the asymptotic variance of $\hat{d}(u)$ is conjectured to satisfy

$$\frac{N}{M} \text{ Var } [\hat{d}(u)] = \overline{K^2} \ \frac{1-\lambda}{\lambda} \ d_F(u) \ d_G(u)$$

where

$$\overline{K^2} = \int_{-\infty}^{\infty} K^2(t) \, dt = \int_{-\infty}^{\infty} k^2(x) \, dx,$$

defining $K(t) = \int_{-\infty}^{\infty} e^{-2\pi i x t} \, k(x) \, dx$, $k(x) = \int_{-\infty}^{\infty} e^{2\pi i x t} \, K(t) \, dt$.

Example. The kernel generating function

$$k(x) = \frac{\sin 2\pi x}{2\pi x} \quad , \quad x \neq 0 \ ,$$

corresponds to the window generating function

$$K(t) = 0.5 \text{ for } |t| \leq 1$$

$$= 0 \qquad \text{otherwise} \quad .$$

One may verify that $\overline{K^2} = 0.5$; the weights

$$k_N(v) = k(\frac{v}{M}) = \frac{\sin 2\pi v/M}{2\pi v/M}$$

correspond to a leap estimator with M=1/h. The foregoing conjecture for the variance of a kernel estimator agrees with the formula in Part 3 for the differential variance of $\tilde{D}(u)$.

2.3 Parametric "Non-parametric" Tests based on Location Scale Parameter Models

A fully non-parametric estimator of $D(u) = FH^{-1}(u)$ is provided by $\tilde{D}(u)$. A functionally parametric estimator $\hat{D}(u)$ is provided by density estimators $d(u)$ based on the kernel method [section 2.2] or the autoregressive method [section 2.4]. A fully parametric estimator of $D(u)$ is based on estimating parameters in a finite parametric model.

A frequently adopted parametric model for the distribution functions F and G is the location-scale parameter model

$$F(x) = F_0\left(\frac{x-\mu_1}{\sigma_1}\right) , \quad G(x) = F_0\left(\frac{x-\mu_2}{\sigma_2}\right)$$

where $F_0(x)$ is a specified distribution function, and μ_1, σ_1, μ_2, σ_2 are unknown parameters. Equivalent parametric models for the quantile functions are

$$F^{-1}(u) = \mu_1 + \sigma_1 Q_0(u) \quad , \quad G^{-1}(u) = \mu_2 + \sigma_2 Q_0(u) \quad .$$

A model for $D_1(t) = HF^{-1}(t)$ is easily obtained:

$$D_1(t) = \lambda t + (1-\lambda) GF^{-1}(t)$$

$$= \lambda t + (1-\lambda) F_0\left(\frac{\mu_1 + \sigma_1 Q_0(t) - \mu_2}{\sigma_2}\right)$$

$$= \lambda t + (1-\lambda) F_0\left(Q_0(t) + \frac{\mu_1-\mu_2}{\sigma_2} + \left(\frac{\sigma_1}{\sigma_2} - 1\right) Q_0(t)\right)$$

$$= \lambda t + (1-\lambda)\, F_0\Big(Q_0(t) - \theta - \psi Q_0(t)\Big) \quad ,$$

defining parameters

$$-\theta = \frac{\mu_1 - \mu_2}{\sigma_2} \quad , \; -\psi = \frac{\sigma_1}{\sigma_2} - 1 \quad .$$

Alternative hypotheses H_1 which are local to H_0 correspond to assuming θ and ψ to be near zero; then one can employ a linear Taylor series expansion of $F_0(x)$ about $x = Q_0(t)$ to obtain

$$D_1(t) = t - (1-\lambda)\,\{\theta\, f_0 Q_0(t) + \psi\, Q_0(t)\, f_0 Q_0(t)\} \quad .$$

Our goal is an approximate parametric formula for $D(u)$.

Conjecture 2C. An approximation for $D(u)$, valid for θ and ψ near 0, is

$$D(u) = u + (1-\lambda)\,\theta\, f_0 Q_0(u) + (1-\lambda)\,\psi Q_0(u) f_0 Q_0(u)$$

A careful derivation of this approximation, and its consequences, is given by Prihoda (1981). By substituting a parametric formula for $D(u)$ in the model for $\tilde{D}(u)$ under local alternatives to H_0,

$$\tilde{D}(u) = D(u) + \left(\frac{1-\lambda}{N\lambda}\right)^{\frac{1}{2}} B(u),$$

one can obtain estimators $\hat{\theta}$ and $\hat{\psi}$ which provide parametric estimators of $D(u)$.

The model for $\tilde{D}(u)$ can be stated as a regression model in θ and ψ:

$$\left(\frac{-1}{1-\lambda}\right)\{\tilde{D}(u) - u\} = \theta\, f_0Q_0(u) + \psi\, Q_0(u)\, f_0Q_0(u) + \{N\lambda(1-\lambda)\}^{-\frac{1}{2}}B(u) \quad .$$

This representation is similar to a representation used by Parzen (1979) in the univariate: one sample case to form estimators of location and scale parameters in the model $F(x) = F_0(\frac{x-\mu}{\sigma})$:

$$f_0Q_0(u)\,\tilde{Q}(u) = \mu\, f_0Q_0(u) + \sigma\, Q_0(u)\, f_0Q_0(u) + \frac{\sigma}{\sqrt{N}}\, B(u) \quad .$$

Linear Rank statistics $\int_0^1 J(u)\, d\tilde{D}(u)$ arise naturally in expressions for estimators $\hat{\theta}$ and $\hat{\psi}$. The variance and covariance of optimal estimators $\hat{\theta}$ and $\hat{\psi}$, in the regular case, are given by inner products of $f_0Q_0(u)$ and $Q_0(u)\, f_0Q_0(u)$ in the Reproducing Kernel Hilbert Space (RKHS) corresponding to the process $\{N\lambda(1-\lambda)\}^{-\frac{1}{2}}\, B(u)$. One can use the data $\tilde{D}(u)-u$ over the full interval $0\leq u\leq 1$, or on a subinterval $0<p\leq u\leq q<1$, or at a discrete grid of values $u_1,\ldots,u_k$ in $(0,1)$.

Asymptotically efficient estimators $\hat{\theta}$ and $\hat{\psi}$ which are linear functionals in $\tilde{D}(u)$ are obtained by applying the theory of regression analysis of continuous parameter time series developed by Parzen (1961). Introduce the reproducing kernel Hilbert Space inner product between functions f(t) and

$g(t)$ on a subinterval $p \le t \le q$ of the unit interval corresponding to the covariance kernel $K(u_1,u_2) = \min(u_1,u_2) - u_1u_2$ of a Brownian Bridge process:

$$\langle f,g \rangle = \int_p^q f'(t)\, g'(t)\, dt + \frac{f(p)g(p)}{p} + \frac{f(q)g(q)}{1-q}$$

Digression. We find interesting an alternate expression:

$$\langle f,g \rangle = \int_0^1 \bar{f}'(t)\, \bar{g}'(t)\, dt$$

where

$$\bar{f}'(t) = f'(t) \qquad p<t<q$$

$$= \frac{1}{p} f(p), \qquad 0 \le t \le p$$

$$= \frac{1}{1-q} f(q), \qquad q \le t \le 1 \quad .$$

To form the inner product of $f(t)$ and $g(t)$ over $0 \le t \le 1$ we require $f(0) = f(1) = g(0) = g(1) = 0$; then

$$\langle f,g \rangle = \int_0^1 f'(t)g'(t)\, dt.$$

Note that $(f_oQ_o)'(u) = -J_o(u)$, $\{Q_o(u)\, f_oQ_o(u)\}' = 1 - J_o(u)Q_o(u)$. To form the estimators $\hat{\theta}$ and $\hat{\psi}$ we form: the information matrix

$$I = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} = \begin{bmatrix} \langle f_o Q_o, f_o Q_o \rangle & \langle f_o Q_o, Q_o \cdot f_o Q_o \rangle \\ \langle f_o Q_o, Q_o \cdot f_o Q_o \rangle & \langle Q_o \cdot f_o Q_o, Q_o \cdot f_o Q_o \rangle \end{bmatrix}$$

and the statistics

$$T_1 = \langle f_o Q_o, \{\tilde{D}(u)-u\} \rangle$$

$$T_2 = \langle Q_o \cdot f_o Q_o, \{\tilde{D}(u)-u\} \rangle \quad .$$

Then

$$\begin{bmatrix} \hat{\theta} \\ \hat{\psi} \end{bmatrix} = \left(\frac{-1}{1-\lambda}\right) I^{-1} \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} \quad .$$

In the symmetric case $f_o Q_o(u) = f_o Q_o(1-u)$, and $Q_o(u) = -Q_o(1-u)$. Then $I_{12} = 0$ when $q = 1-p$.

Let us explicitly evaluate the inner products when we use the interval $0 \leq u \leq 1$ and all the data $\tilde{D}(u)$, $0 \leq u \leq 1$. We must assume that $\int_0^1 J_o(u)\, du = \int_0^1 \{1-J_o(u)Q_o(u)\}\, du = 0$. Then

$$I_{11} = \int_0^1 |J_o(u)|^2 du, \quad I_{22} = \int_0^1 |1-J_o(u)Q_o(u)|^2\, du,$$

$$T_1 = \int_0^1 J_0(u)\, d\tilde{D}(u)$$

$$T_2 = \int_0^1 \{1 - J_0(u)\, Q_0(u)\}\, d\tilde{D}(u) \quad .$$

The covariance matrix of $\hat{\theta}$ and $\hat{\psi}$ is given in general by

$$\begin{bmatrix} \text{Var}\ (\hat{\theta}) & \text{Cov}\ (\hat{\theta},\hat{\psi}) \\ \text{Cov}\ (\hat{\theta},\hat{\psi}) & \text{Var}\ (\hat{\psi}) \end{bmatrix} = \frac{1}{N\lambda(1-\lambda)} \quad I^{-1}$$

To illustrate the meaning of the foregoing formula for variance, consider $\theta = (\mu_2-\mu_1)/\sigma_2$ in the normal case. Assume that $\sigma_1 = \sigma_2 = \sigma$. Let $\hat{\mu}_j$ be the sample means and $\hat{\sigma}^2$ the variance of the pooled sample. Then $(\hat{\mu}_1-\hat{\mu}_2)/\hat{\sigma}$ has asymptotic variance equal to $m^{-1} + n^{-1} = N/mn = \{N\lambda(1-\lambda)\}^{-1}\ I_{11}^{-1}$, since $J_0(u) = \Phi^{-1}(u)$ in the normal case.

To test H_0: F=G the analogue of conventional "non-parametric" test statistics (which could be called a parametric "non-parametric" test statistic for H_0) is the quadratic form [where * denotes transpose]

$$L = \begin{bmatrix} \hat{\theta} \\ \hat{\psi} \end{bmatrix}^* \begin{bmatrix} \text{Var}\ [\hat{\theta}] & \text{Cov}\ [\hat{\theta},\hat{\psi}] \\ \text{Cov}\ [\hat{\theta},\hat{\psi}] & \text{Var}\ [\hat{\psi}] \end{bmatrix}^{-1} \begin{bmatrix} \hat{\theta} \\ \hat{\psi} \end{bmatrix} ,$$

which, under the null hypothesis H_0, has a Chi-squared distribution with two degrees of freedom.

In terms of the test statistics T_1 and T_2 one can write

$$L = \frac{N\lambda}{1-\lambda} \begin{bmatrix} T_1 \\ T_2 \end{bmatrix}^* I^{-1} \begin{bmatrix} T_1 \\ T_2 \end{bmatrix}$$

Example: The logistic distribution has standard quantile function and score function

$$Q_0(u) = \log \frac{u}{1-u} \quad , \quad J_0(u) = 2u-1 \quad .$$

Therefore [see Eubank (1979)]

$$I_{11} = \int_0^1 (2u-1)^2 \, du = \frac{1}{3} \quad , \quad I_{22} = \frac{3 + \pi^2}{9} \quad .$$

A non-parametric test statistic for location [which is optimum for the logistic distribution] is

$$L_1 = \frac{N\lambda}{1-\lambda} \frac{|T_1|^2}{I_{11}} = \frac{12N\lambda}{(1-\lambda)} \left| \int_0^1 (u - \tfrac{1}{2}) \, d\tilde{D}(u) \right|^2$$

which is asymptotically equivalent to the Wilcoxon statistic $\sum_{j=1}^{m} R_j$. It is equivalent in a finite sample if we define $\tilde{D}(u)$ to be piecewise constant and equal to j/m at $u = R_j/(n+1)$; then

$$\int_0^1 u \, d\tilde{D}(u) = \frac{1}{m} \sum_{j=1}^{m} R_j/(N+1) \quad .$$

A non-parametric test statistic for scale [which is optimum for the logistic distribution] is

$$L_2 = \frac{N\lambda}{1-\lambda} \frac{|T_2|^2}{I_{22}}$$

$$= \frac{36\ N\lambda}{(3+\pi^2)(1-\lambda)} \left| \frac{1}{m} \sum_{j=1}^{m} \left(\frac{R_j}{N+1} - \frac{1}{2}\right) \log \left(\frac{R_j}{N+1-R_j}\right) \right|^2 .$$

This test may have been given first by Prihoda (1981).

Motivation of entropy or information measures as "portmanteau" test statistics. Parametric "non-parametric" tests of H_o given by L may be most powerful when one is testing alternative hypotheses which correspond to shifts in location and scale parameters. To obtain general "portmanteau" procedures, which do not require close specifications of the alternative hypothesis, let us re-express the statistic L in terms of the estimated density

$$\hat{d}(u) = 1 + (1-\lambda) \left[\hat{\theta}(f_o Q_o\ (u))' + \hat{\psi}\ (Q_o(u) f_o Q_o(u))'\right] .$$

One may verify that

$$\int_o^1 \left|\hat{d}(u) - 1\right|^2 du = \begin{bmatrix} T_1 \\ \\ T_2 \end{bmatrix}^* \quad I^{-1} \quad \begin{bmatrix} T_1 \\ \\ T_2 \end{bmatrix} = L\{(1-\lambda)/N\lambda\} .$$

One is led to conjecture that

$$\frac{N\lambda}{1-\lambda} \int_o^1 [\hat{d}(u)-1]^2\ du$$

is a test statistic for H_0 whose null distribution is chi-squared with degrees of freedom equal to the number of parameters in $\hat{d}(u)$. Next one is led to conjecture that the entropy or information measure

$$\frac{2N\lambda}{1-\lambda} \int_0^1 - \log \hat{d}(u)\, du$$

is a test statistic for H_0 whose null distribution is chi-squared with degrees of freedom equal to the number of parameters in $\hat{d}(u)$. We next introduce autoregressive estimators $\hat{d}(u)$ for which $\int_0^1 \log \hat{d}(u)\, du$ is evaluated as a parameter without integrating an estimated density.

2.4 Autoregressive estimation of d(u) and tests of H_0

The entropy H(f) of a probability density f(x), $-\infty<x<\infty$, is defined to be

$$H(f) = \int_{-\infty}^{\infty}\{-\log f(x)\} f(x)\, dx$$

$$= \int_0^1 \{-\log fF^{-1}(u)\}\, du.$$

For a density d(u), $0\leq u\leq 1$ we define

$$H(d) = \int_0^1 -\log d(u)\, du \geq 0$$

to be the entropy in the density-quantile sense. Density estimators $\hat{d}(u)$ whose entropy are easily evaluated are provided by the autoregressive method of density estimation.

Given $\tilde{\rho}(v)$, $v=0, \pm 1,\ldots,\pm m, \ldots$ one forms for $m=1,2,\ldots$

$$\hat{d}_m(u) = \hat{K}_m \left|1 + \hat{\alpha}_m(1)e^{2\pi iu} + \ldots + \hat{\alpha}_m(m)\, e^{2\pi ium}\right|^{-2}$$

where the complex-valued autoregressive coefficients $\hat{\alpha}_m(m)$ are computed by the Yule-Walker equations described below. Further

$$H(\hat{d}_m) = \int_0^1 -\log \hat{d}_m(u)\, du = -\log \hat{K}_m$$

is directly computed in terms of the parameter $\hat{K}_m$.

The Yule-Walker equations [which are solved to obtain $\hat{\alpha}_m(j)$, $j=1,\ldots, m$ and $\hat{K}_m$ from $\tilde{\rho}(v)$, $v=0, \pm 1, \ldots, \pm m$] are [see Parzen (1979)]

$$\sum_{j=0}^{m} \hat{\alpha}_m(j)\, \tilde{\rho}(j-k) = 0 , \quad k=1,\ldots,m,$$

$$\sum_{j=0}^{m} \hat{\alpha}_m(j)\, \tilde{\rho}(j) = \hat{K}_m$$

where $\hat{\alpha}_m(0) = 1$. They are solved using the recursive algorithm

$$\hat{\alpha}_m(m) = - \frac{1}{\hat{K}_{m-1}} \sum_{j=0}^{m-1} \hat{\alpha}_{m-1}(j)\, \tilde{\rho}(j-m)$$

and for $j=1,\ldots,m-1$

$$\hat{\alpha}_m(j) = \hat{\alpha}_{m-1}(j) + \hat{\alpha}_m(m)\, \hat{\alpha}^*_{m-1}(m-j)$$

where α^* is the complex conjugate of α. The autoregressive method of estimating densities was first implemented in a computer program, and its theory investigated, by Carmichael (1976, 1978).

A proposed diagnostic for determining the order m of an autoregressive estimator $\hat{d}_m(u)$ is the plot of

$$\bar{D}_m(u) = \int_0^u \frac{1}{\hat{d}(m)}\, d\tilde{D}(u) \quad .$$

An intuitive criterion for choosing an optimal order m is the smallest value of m for which $\bar{D}_m(u)$, $0 \leq u \leq 1$, is not significiantly different from $D_0(u) = u$, $0 \leq u \leq 1$.

This paper aims to raise the consciousness of statisticians about the FUN.STAT (functional statistical inference) approach to the two sample problem. At this time we can only state conjectures about the theorems that need to be proved [theoretically and/or by Monte Carlo calculations]. One theorem is about the large-sample properties of autoregressive density estimators; another theorem is about the use of estimates $-\log \hat{K}_m$ [of entropy or information measures] to test H_0 and to form order determining criteria for optimal autoregressive orders m. A noteworthy irony is that the orders m chosen in practice are small, and one might wonder about the relevance of a large sample consistency theorem.

Conjecture 2D. Formula for the asymptotic variance of $\hat{d}_m(u)$ as an estimator of d(u): As m tends to ∞ at a suitable rate [such as $m^3/N \to 0$ as $N \to \infty$] $\hat{d}_m(u)$ tends in probability to d(u) and its asymptotic variance satisfies

$$\frac{N}{m} \text{Var } [\hat{d}_m(u)] = 2\frac{1-\lambda}{\lambda} d_F(u)\, d_G(u) \quad .$$

(Note that m denotes the autoregressive order and not the size of the X sample.) This conjecture is based on the formulas for the variance of the kernel estimators conjectured in section 2.2, and the relations between the distributions of kernel and autoregressive estimators of the spectral density function of a stationary time series [conjectured by Parzen (1969), and confirmed by Berk (1974)].

Conjecture 2C. A "portmanteau" (alternative hypotheses unspecified) procedure for testing H_0 [which may have optimal properties] is of the form: accept H_0 if

$$- \log K_m \leq \frac{2m}{N} \frac{1-\lambda}{\lambda} \quad , \quad m=1,2, \ldots \quad .$$

It should be emphasized that further theoretical and Monte Carlo investigation is required to find the best multiple of the right hand side to use in practice [perhaps including a factor of log log m].

Conjecture 2F. A procedure for autoregressive density estimator order determination. If one rejects H_o because one of the inequalities in Conjecture 2E is violated, let $\hat{m}$ be the value of m minimizing a criterion of the form

$$AIC(m) = \log K_m + \frac{2m}{N} \frac{1-\lambda}{\lambda} \quad .$$

An estimator of d(u) is taken to be $\hat{d}_{\hat{m}}(u)$. Note that $AIC(\hat{m})<0$.

Conjecture 2G. Can one develop criteria for accepting H_o based on the values of $(\frac{2N\lambda}{1-\lambda})|\tilde{\rho}(v)|^2$, v=1,2,..., . Under H_o these statistics are asymptotically independent Chi-squared distributed with 2 degrees of freedom.

Part 3. Asymptotic Distributions of Stochastic Processes Arising in Two Sample Quantile Data Analysis

3.1 Introduction

The FUN.STAT approach to testing equality of two independent populations with bi-continuous distributions F and G respectively is based on:
(1) estimating parameters which are functions such as $D_1(u) = HF^{-1}(u)$, and $D(u) = FH^{-1}(u)$, and
(2) exploratory data analysis of the fully non-parametric estimators $\tilde{D}_1(u) = \tilde{H}\tilde{F}^{-1}(u)$ and $\tilde{D}(u) = \tilde{D}_1^{-1}(u)$.

This part discusses how to derive the asymptotic distribution of the stochastic processes $\tilde{D}_1(u)$, $0 \le u \le 1$, and $\tilde{D}(u)$, $0 \le u \le 1$. Our aim is to outline an operational calculus for intuitively deriving results concerning the distribution of empirical processes [such as $\tilde{D}(u)$], and for identifying stochastic processes $C_D(u)$ and $C_{D_1}(u)$ such that

$$\sqrt{N}\,\{\tilde{D}(u) - D(u)\} \xrightarrow{D} C_D(u), \quad \sqrt{N}\,\{D_1(u) - D_1(u)\} \xrightarrow{D} C_{D_1}(u) \quad ,$$

where $\xrightarrow{D}$ connotes convergence in distribution of stochastic processes. Our results are heuristic theorems, rather than rigorously proved theorems with carefully stated regularity assumptions.

Theorem 3A. Asymptotic distribution of a sample distribution function $\tilde{F}(x)$. $\tilde{F}(x)$ of a random sample $X_1,\ldots,X_m$ can be expressed in terms of $\tilde{F}_U(u)$ of $U_1 = F(X_1),\ldots,U_m = F(X_m)$ which are uniform on [0.1]. Note that $\tilde{F}F^{-1}(u) = \tilde{F}_U(u)$. One can show that there is a Brownian Bridge process $B(u)$, $0 \le u \le 1$, such that

$$m^{\frac{1}{2}}\{\tilde{F}_U(u) - u\} \xrightarrow{D} B(u)$$

We denote by $B_F(u)$ a Brownian Bridge process such that

$$\tilde{B}_F(u) = m^{\frac{1}{2}} \{\tilde{F}F^{-1}(u) - u\} \xrightarrow{D} B_F(u) \quad ,$$

Next define

$$\tilde{C}_F(x) = m^{\frac{1}{2}}\{\tilde{F}(x) - F(x)\} \xrightarrow{D} B_F(F(x))$$

The limiting process of $\tilde{F}(x)$ is denoted $C_F(x)$, where $C_F(x) = B_F(F(x))$.

Theorem 3B. Asymptotic distribution of sample quantile function $\tilde{F}^{-1}(u)$. $\tilde{F}_U^{-1}(u) = F\tilde{F}^{-1}(u)$ satisfies

$$\tilde{B}_{F^{-1}}(u) = \sqrt{m}\ \{F\tilde{F}^{-1}(u) - u\} \xrightarrow{D} - C_F\ (F^{-1}(u)) = -B_F(u)$$

$\tilde{F}^{-1}(u)$ under suitable conditions on $fF^{-1}(u)$ [see Csörgö and Révész (1981)] satisfies

$$\sqrt{m}\ \{\tilde{F}^{-1}(u) - F^{-1}(u)\} \xrightarrow{D} \frac{(-1)}{fF^{-1}(u)}\ B_F(u)$$

The limiting process of $\tilde{F}^{-1}(u)$ is denoted $C_{F^{-1}}(u)$. We note the basic relation

$$C_{F^{-1}}(u) = (-1)\ \{\frac{d}{du}\ F^{-1}(u)\}\ C_F(F^{-1}(u))$$

Proof: Write $F\tilde{F}^{-1}(u) - u = F\tilde{F}^{-1}(u) - \tilde{F}\tilde{F}^{-1}(u) + \tilde{F}\tilde{F}^{-1}(u) - u$. One may verify that

$$\sup_{0\le u\le 1} |\tilde{F}\tilde{\tilde{F}}^{-1}(u) - u| \le \frac{1}{m} \quad ,$$

$$\sqrt{m}\,\{\tilde{F}\tilde{\tilde{F}}^{-1}(u) - F\tilde{\tilde{F}}^{-1}(u)\} = \tilde{C}_F(\tilde{F}^{-1}(u)) \xrightarrow{D} C_F(F^{-1}(u)) = B_F(u) \quad .$$

The first conclusion may now be inferred. Next $m^{\frac{1}{2}}\{\tilde{F}^{-1}(u) - F^{-1}(u)\}$ equals

$$\{\frac{\tilde{F}^{-1}(u) - F^{-1}(u)}{F\tilde{F}^{-1}(u) - FF^{-1}(u)}\}\{m^{\frac{1}{2}}\,(F\tilde{F}^{-1}(u) - u)\}$$

The left bracket contains the reciprocal of a difference quotient which tends to $f(x)$ evaluated at $x = F^{-1}(u)$. The right bracket converges to $-B_F(u)$.

3.2 Conjectures in Distribution of $\tilde{D}(u)$

To apply Theorems 3A and 3B to the two-sample problem we first derive heuristically the asymptotic distribution of $\tilde{G}\tilde{\tilde{F}}^{-1}(u)$ as an estimator of $GF^{-1}(u)$:

$$\sqrt{N}\,\{\tilde{G}\tilde{\tilde{F}}^{-1}(u) - GF^{-1}(u)\}$$

$$= \sqrt{N}\,\{\tilde{G}\tilde{G}^{-1}(G\tilde{F}^{-1}(u)) - G\tilde{F}^{-1}(u) + G\tilde{F}^{-1}(u) - GF^{-1}(u)\}$$

$$= (1-\lambda)^{-\frac{1}{2}}\,\tilde{B}_G\,(G\tilde{F}^{-1}(u))$$

$$+ \lambda^{-\frac{1}{2}}\,\frac{GF^{-1}(F\tilde{F}^{-1}(u)) - GF^{-1}(u)}{F\tilde{F}^{-1}(u) - u}\;\tilde{B}_{F^{-1}}(u)$$

$$\xrightarrow{D} (1-\lambda)^{-\frac{1}{2}}\,B_G\,(GF^{-1}(u)) - \lambda^{-\frac{1}{2}}\{\frac{d}{du}\,GF^{-1}(u)\}\,B_F\,(u)$$

The asymptotic distribution of

$$\tilde{D}_1(u) = \tilde{H}\tilde{F}^{-1}(u) = \lambda\tilde{F}\tilde{F}^{-1}(u) + (1-\lambda)\,\tilde{G}\tilde{F}^{-1}(u)$$

as an estimator of $D_1(u) = HF^{-1}(u) = \lambda u + (1-\lambda)\,GF^{-1}(u)$ is described by the asymptotic distribution of $\tilde{H}\tilde{F}^{-1}(u) - HF^{-1}(u)$.

Conjecture 3C. $\sqrt{N}\,(\tilde{H}\tilde{F}^{-1}(u) - HF^{-1}(u))$ and

$\sqrt{N}(1-\lambda)\,(\tilde{G}\tilde{F}^{-1}(u) - GF^{-1}(u))$ converge in distribution to

$$C_{D_1}(u) = (1-\lambda)\,[(1-\lambda)^{-\frac{1}{2}}\,B_G(GF^{-1}(u)) - \lambda^{-\frac{1}{2}}\,\{\frac{gF^{-1}(u)}{fF^{-1}(u)}\}\,B_F(u)]$$

Let $d_1(u) = D_1'(u) = hF^{-1}(u)/fF^{-1}(u)$. Then $d_1(D_1^{-1}(u)) = hH^{-1}(u)/fH^{-1}(u)$. The asymptotic distribution of $\tilde{D}(u) = (\tilde{H}\tilde{F}^{-1})^{-1}(u)$ as an estimator of $D(u) = (HF^{-1})^{-1}(u) = FH^{-1}(u)$ is conjectured (using proofs similar to those used for sample quantile functions) to satisfy the following theorem.

Conjecture 3D. Asymptotic distribution of $\tilde{D}(u)$.

$$\sqrt{N}\,\{\tilde{D}(u) - D(u)\} \xrightarrow{D} C_D(u)$$

where $C_D(u) = \dfrac{(-1)}{d_1(D_1^{-1}(u))}\,C_{D_1}(D_1^{-1}(u))$ is explicitly given by

$$C_D(u) = (1-\lambda)\,[(1-\lambda)^{-\frac{1}{2}}\,\frac{fH^{-1}(u)}{hH^{-1}(u)}\,B_G(GH^{-1}(u)) - \lambda^{-\frac{1}{2}}\,\frac{gH^{-1}(u)}{hH^{-1}(u)}\,B_F(FH^{-1}(u))]\ .$$

The covariance kernel of $C_D(u)$, given in Conjecture 2A, shows that the complex-looking process $C_D(u)$ actually has much simplifying symmetry. It has been previously derived in Pyke-Shorack (1968) who show

$$\sqrt{N}\ \{\tilde{F}\tilde{H}^{-1}(u) - FH^{-1}(u)\} \xrightarrow{D} C_D(u) \quad .$$

An interesting question is whether the asymptotic distribution of $\tilde{D}(u)$ can be deduced from the Pyke-Shorack results using the fact [Theorem 1B] that $\tilde{D}(u) - \tilde{\tilde{F}}\tilde{\tilde{H}}^{-1}(u)$ equal 0 except for about m sub-intervals of length 1/N in which it equals 1/m.

Distribution of stochastic Stieltjes integrals and linear rank statistics. The process $\tilde{D}(u)$ has the important property that a linear rank statistic can be asymptotically represented as a stochastic Stieltjes integral

$$\int_0^1 J(u)\ d\tilde{D}(u)$$

for a suitable continuous score function J(u). Its limiting distribution can be described as follows:

$$\Delta(J) = \sqrt{N}\ \{\int_0^1 J(u)\ d\tilde{D}(u) - \int_0^1 J(u)\ dD(u)\}$$

is asymptotically normal with zero mean and covariance kernel $K_\Delta(J_1,J_2)$, heuristically representing

$$K_\Delta(J_1,J_2) = \text{Cov}\ [\Delta(J_1),\ \Delta(J_2)] \quad ,$$

and given by

$$K_\Delta(J_1,J_2) = \int_0^1 \int_0^1 J_1(u_1)\, J_2(u_2)\, dE[C_D(u_1)\, C_D(U_2)] \quad .$$

Explicitly

$$K_\Delta(J_1,J_2) = K_1(J_1,J_2) + K_3(J_1,J_2) - K_2(J_1,J_2)$$

where

$$K_1(J_1,J_2) = \frac{1-\lambda}{\lambda} \int_0^1 J_1(u)\, J_2(u)\, d_G(u)\, d_F(u)\, du \;;$$

$$K_2(J_1,J_2) = \frac{(1-\lambda)^2}{\lambda} \int_0^1 J_1(u)\, \{d_G(u)D_F(u)\}'\, du \int_0^1 J_2(u)\, \{d_G(u)D_F(u)\}'\, du$$

$$+ (1-\lambda)\int_0^1 J_1(u)\, \{d_F(u)D_G(u)\}'\, du \int_0^1 J_2(u)\{d_F(u)D_G(u)\}'\, du \;;$$

$$K_3(J_1,J_2) = (1-\lambda)^2 \int_0^1 \int_0^1 du_1\, du_2\, J_1(u_1)J_2(u_2)$$

$$[\{\lambda^{-1}\, d_G'(u_1)d_G'(u_2)D_F(\min(u_1,u_2)) + (1-\lambda)^{-1}d_F'(u_1)d_F'(u_2)D_G(\min(u_1,u_2)\}$$

$$+ e(u_1-u_2)\, \{\lambda^{-1}d_G'(u_1)d_G(u_2)d_F(u_2) + (1-\lambda)^{-1}d_F'(u_1)d_F(u_2)d_G(u_2)\}$$

$$+ e(u_2-u_1)\{\lambda^{-1}d_G'(u_2)d_G(u_1)d_F(u_1) + (1-\lambda)^{-1}d_F'(u_2)d_F(u_1)d_G(u_1)\}] \quad .$$

where $e(t) = 1$ or 0 as $t>0$ or $t<0$. Under the null hypothesis H_o

$$K_\Delta(J_1,J_2) = (\tfrac{1-\lambda}{\lambda})\{\int_0^1 J_1(u)J_2(u)\,du - \int_0^1 J_1(u)\,du \int_0^1 J_2(u)\,du\} .$$

One can obtain the asymptotic covariance of $\tilde{\rho}(v_1)$ and $\tilde{\rho}(v_2)$ by choosing $J_1(u) = e^{2\pi i u v_1}$, $J_2(u) = e^{-2\pi i u v_2}$.

3.3 Density Estimation and Differential Variance

Insight into the asymptotic variances of density estimators is provided by a formula for the variance of the fully non-parametric estimator of $d(u) = D'(u)$ given by the numerical derivative

$$\tilde{d}(u) = \frac{\tilde{D}(u+h) - \tilde{D}(u-h)}{2h} .$$

Conjecture 3E. A formula for the asymptotic variance of the numerical derivative $\tilde{d}(u)$ is

$$2hN \text{ Var } [\tilde{d}(u)] \doteq \frac{1-\lambda}{\lambda} d_F(u)\, d_G(u) .$$

The expression on the right hand side is called the differential variance of $\tilde{D}(u)$; it can be used to suggest conjectures concerning the asymptotic distributions of kernel and autoregressive density estimators [Conjectures 2B and 2D]. The form of the differential variance suggests that $\tilde{d}(u)$ has the distributional properties of a density-quantile estimator since an estimator of a probability density $d(u)$ has variance proportional to $d(u)$, while the variance of a density-quantile $d(u)$ has variance proportional to $d^2(u)$.

Conjecture 3F. A fully non-parametric estimator of $d_1(u) = D_1'(u) = \lambda + (1-\lambda)\{gF^{-1}(u)\ fF^{-1}(u)\}$ given by

$$\tilde{d}_1(u) = \frac{\tilde{D}_1(u+h) - (\tilde{D}_1(u-h)}{2h}$$

has asymptotic variance satisfying

$$2hN\ \mathrm{Var}[\tilde{d}_1'(u)] = \frac{1-\lambda}{\lambda}\{\frac{d}{du}\ GF^{-1}(u)\}\ \{\frac{d}{du}\ HF^{-1}(u)\} = \frac{1-\lambda}{\lambda}\ d_1(u)\{\frac{d_1(u)-\lambda}{1-\lambda}\}$$

Outline of a heuristic proof of Conjecture 3E. $2hN\ \mathrm{Var}\ [\tilde{d}(u)]$ approximately equals

$$\frac{1}{2h}\ E|C_D(u+h) - C_D(u-h)|^2$$

$$= \frac{(1-\lambda)^2}{2h}\ [(1-\lambda)^{-1}d_F^2(u)\ E|B_G(GH^{-1}(u+h)) - B_G(GH^{-1}(u-h))|^2$$

$$+ \lambda^{-1}\ d_G^2(u)\ E|B_F\ (FH^{-1}(u+h)) - B_F(FH^{-1}(u+h))|^2]$$

$$= \frac{(1-\lambda)^2}{\lambda(1-\lambda)}\ [\lambda\ d_F^2(u)\ d_G(u) + (1-\lambda)\ d_G^2(u)\ d_F(u)]$$

$$= \frac{1-\lambda}{\lambda}\ d_F(u)\ d_G(u)$$

since $\lambda d_F(u) + (1-\lambda)\ d_G(u) = 1$.

Covariance of linear functionals required to derive asymptotic variance of density estimators. A general theory of asymptotic distribution of density estimators can be developed by assuming that $K_\Delta(J_1,J_2)$ can be represented

$$K_\Delta(J_1,J_2) = \int_0^1 J_1(u)\, J_2(u)\, V_1(u)\, du$$

$$+ \int_0^1 \int_0^1 J_1(u)\, J_2(u)\, V_2(u_1,u_2)\, du_1 du_2$$

where $V_1(u)$ and $V_2(u_1,u_2)$ are integrable functions. We call $V_1(u)$ the differential variance; $V_2(u_1,u_2)$ vanishes in formulas for the asymptotic variance of kernel and autoregressive density estimators. Spectral averages of the spectral density of a stationary time series which is a linear process have the foregoing structure [see Parzen (1961), p. 982].

Part 4. Summary of Two Sample Quantile Data Analysis Using TWOSAM

To test the null hypothesis H_0: $F(x) = G(x)$ of equality of two populations, statisticians usually choose a test statistic (T_N, D_{mn}, W^2_{mn}, etc.), compute its value from the data, and test the significance of the computed value of the test statistic chosen. This paper shows that conventional test statistics can be represented as functionals of the process $\tilde{D}(u)$, $0 \leq u \leq 1$, and proposes an autoregressive density estimation approach to the data analysis of $\tilde{D}(u)$.

In addition to providing the applied statistician with the ability to analyze sample paths of continuous parameter stochastic processes [such as $\tilde{D}(u)$, $0 \leq u \leq 1$], this paper aims to stimulate the applied statistician to appreciate the basic probability theory of these stochastic processes.

A graphical (rather than an arithmetical) way to test H_0 is to plot $\tilde{D}(u)$, $0 \leq u \leq 1$, and examine whether its deviation from the uniform distribution $D_0(u) = u$ appears to be significantly different from the sample path of a Brownian Bridge with variance $(1-\lambda)/\lambda N$.

The proposed quantile data analysis approach to the univariate: two sample problem involves several stages.

Stage 1. Fully non-parametric analysis. Obtain for each of the two samples, and for the pooled sample, descriptive statistics and plots of the informative quantile function. Plot on one graph the quantile functions of the two samples. Plot $\tilde{D}(u)$.

Stage 2. Autoregressive analysis. Obtain: $|\tilde{\rho}(v)|^2$, square modulus of sample pseudo-correlations, for $v=1,\ldots,M$ where M is a specified maximum

order. Plot $\hat{d}_m(u)$ and $\bar{D}_m(u)$ for m=1,...,M. List values of $\hat{K}_m$, log $\hat{K}_m$, and AIC(m) for m=1,...,M. Obtain optimal order by AIC criterion.

The two sample non-parametric data modeling procedures described in this paper have been implemented in a computer program called TWOSAM. I would like to acknowledge the contributions to this program of the following colleagues during the course of their Ph.D. studies: Jean-Pierre Carmichael, Mike White, Tom Prihoda, Scott Anderson, Phil Spector, and Avi Harpaz (who deserves special thanks for the current version of the program).

To illustrate how the quantile approach to data analysis could be presented to students in an introductory statistics class, we consider a data set analyzed by Larsen and Marx (1981), p. 324.

An important problem of two sample data analysis arises in cases of disputed authorship. Were the 10 essays published in 1861 by "Quintus Curtius Snodgrass" actually written by Mark Twain? Let X and Y respectively denote the proportion of three-letter words in (eight) Twain essays and (ten) Snodgrass essays. For ease of writing, the sample values $X_1,...,X_8$, $Y_1,...,Y_{10}$ are multiplied by 1000 and 200 is subtracted. The samples then have order statistics:

X: 17, 17, 25, 29, 30, 35, 40, 62

Y: -4, 1, 2, 5, 7, 9, 10, 20, 23, 24.

A typical data analysis might include the following diagnostics.

I. An analysis of the two samples based on the t-test yields a t-value of 3.86 and rejects H_0 [at the .002 level]. That the distributions of X and

Y are very non-normal can be quickly examined by plotting the informative quantile functions of the samples. For a sample quantile function $\tilde{Q}(u)$ the informative quantile function IQ(u), which represent Q(u) normalized so that $Q(0.5) = 0$ and approximately $Q'(0.5)=1$, is estimated by

$$I\tilde{Q}(u) = \frac{\tilde{Q}(u) - \tilde{Q}(0.5)}{2\{\tilde{Q}(0.75) - \tilde{Q}(0.25)\}}$$

For a random sample $X_1,\ldots,X_m$, with order statistics $X_{(1)}<\ldots<X_{(m)}$, we define

$$\tilde{Q}(\tfrac{j}{m+1}) = X_{(j)}\ , \quad j=1,\ldots,m;$$

$\tilde{Q}(u)$ is defined by linear interpolation for other values of u. With this convention we obtain

u	1/9	2/9	3/9	4/9	5/9	6/9	7/9	8/9		
$I\tilde{Q}_X(u)$	-.38	-.38	-.14	-.02	.02	.17	.32	.98		
u	1/11	2/11	3/11	4/11	5/11	6/11	7/11	8/11	9/11	10/1
$I\tilde{Q}_Y(u)$	-.52	-.30	-.26	-.13	-.04	.04	.09	.52	.67	.70

These informative quantile functions indicate shorter-tailed distributions than the normal. A test based on the t-statistic might still be defended by those who believe that robustness justifies such procedures [this may be true only for distributions for which IQ(u) is not too asymmetric].

II. Conventional two-sample procedure. Apply a Wilcoxon rank sum test. Let R_j denote the ranks in the pooled sample of the X values.

R: 8, 9, 13, 14, 15, 16, 17, 18 .

One desires to test the significance (concerning equality of populations) of the rank sum equal to 110, or equivalently of the statistic

$$T = \frac{1}{m} \sum_{j=1}^{m} \frac{R_j}{N+1} \quad .$$

Note E[T] = 0.5. For the Mark Twain data, T = 110/152 = .7237. The variance of T is .0055; therefore $(T-E[T])/\sigma(T) = 3.02$. One concludes that the hypothesis H_0 that Twain wrote the Snodgrass papers is rejected [at the .001 level, using the normal approximation].

III. A graphical test can lead to a firm conclusion. An alternative to computing a statistic and determining its significance level is to plot $\tilde{D}(u)$, using the fact that it is a distribution function with jumps of size 1/m at the points R_j/N. $\tilde{D}(u)$ has the following values:

u	8/18	9/18	13/18	14/18	15/18	16/18	17/18	18/18
	.444	.5	.722	.778	.833	.889	.944	1.0
$\tilde{D}$ (u)	1/8	2/8	3/8	4/8	5/8	6/8	7/8	8/8
	.125	.25	.375	.5	.625	.75	.875	1.0

The graph of $\tilde{D}(u)$ is always below the uniform $D_0(u) = u$; we conclude that no reasonable test procedure would decide that Twain wrote the Snodgrass papers.

IV. Pseudo-correlations. The following table lists for the Mark Twain data the squared-modulus $|\tilde{\rho}(v)|^2$ of the pseudo-correlations of lags v=1,...,5:

v	1	2	3	4	5
$\lvert\tilde{\rho}(v)\rvert^2$	.2527	.1664	.0652	.1196	.0253

Since $2N\lambda/(1-\lambda) = 28.8$, the pseudo-correlation of lag 1 indicates that H_o should be rejected [at the .025 level; $28.8\lvert\tilde{\rho}(1)\rvert^2 = 7.3$].

V. Entropy and AIC. The following table lists for the Mark Twain data the entropy $-\log \hat{K}_m$, and order determining criterion AIC(m), for m=1,2,...,5:

m	1	2	3	4	5
$-\log \hat{K}_m$	.291	.696	1.669	1.980	2.740
AIC(m)	-.152	-.418	-1.252	-1.424	-2.046

One rejects H_o because AIC(m)<0 for some $m \geq 1$ (and indeed AIC is negative for all the values of m listed above). No optimal order $\hat{m}$ is chosen because AIC(m) does not achieve a relative minimum among the orders listed.

VI. Graphs of autoregressive density estimators $\hat{d}_m(u)$. When an order $\hat{m}$ is determined one considers $\hat{d}_{\hat{m}}(u)$ as an estimator of d(u). For the Mark Twain data, where the two samples are almost disjoint, no order is determined. The graphs of $\bar{D}_m(u)$ also indicate that a satisfactory estimator is not achieved among m=1,...,5. Since the sample sizes are so small here, one hesitates to consider larger values of m.

Actual graphs produced by TWOSAM are not included in this paper because the paper is too long and for the Snodgrass example the graphs are not actually needed to draw conclusions about H_o. Of course, one should study the graphs in order to discern information not contained in the numbers proposed as diagnostic measures.

Epilogue. What do we see as the future of the FUN.STAT Quantile approach to two-sample data analysis? It aims to provide statisticians with (1) new procedures which can detect differences in populations which are not diagnosed by conventional procedures, and (2) diagnostics of distributional shape which can enhance confidence in the use of conventional procedures. The theory of the new procedures is asymptotic, but they are practical to use in both very small and very large samples. The investigation of their properties, especially in small samples by Monte-Carlo methods, can be considered to provide many important research problems. We would like to emphasize our belief that it is unwise to rely on pure graphical data analysis based only on graphs which are not accompanied by diagnostic measures. FUN.STAT facilitates estimation of entropy and information measures which are particularly useful summary measures because they may provide comparisons between parametric and non-parametric analysis of a data set [see Parzen (1983)].

REFERENCES

Berk, K. N. (1974) Consistent autoregressive spectral estimates. Ann. Statist., 2, 489-502.

Carmichael, J. P. (1976) The Autoregressive Method: A Method of Approximating and Estimating Positive Functions. Ph.D. Thesis, Statistical Science, SUNY Buffalo

_____. (1978) Consistency of the Autoregressive Method of Density Estimation. Technical Report. Statistical Science, SUNY Buffalo.

_____ and Parzen, E. (1977) New Nonparametric Approach to the Two-Sample Problem. Technical Report. Statistical Science, SUNY Buffalo.

Chernoff, H. and Savage, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. Ann. Math Statist. 29, 972-94.

Csörgö, M. and Révész, P. (1981) Strong Approximations in Probability and Statistics, Academic Press: New York.

Eubank, R. L. (1979) A Density-Quantile Function Approach to Choosing Order Statistics for the Estimation of Location and Scale Parameters, Technical Report A-10, Texas A&M, Institute of Statistics.

Grenander, U. (1981) Abstract Inference, Wiley: New York.

Larsen, R. J. and Marx, M. L. (1981) An Introduction to Mathematical Statistics and its Applications, Prentice Hall: Englewood Cliffs, N.J.

Parzen, E. (1961) An approach to time series analysis Ann. Math. Statist., 32 (1961), 951-989.

_____. (1961) Regression analysis of continuous parameter time series Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I, Univ. California Press, Berkeley, Calif. 469-489.

_____. (1969) Multiple time series modeling Multivariate Analysis - II, edited by P. Krishnaiah, Academic Press: New York, 389-409.

_____. (1979) Nonparametric Statistical Data Modeling Journal of the American Statistical Association, (with discussion), 74, 105-131.

_____. (1980) Quantile Functions, Convergence in Quantile, and Extreme Value Distribution Theory, Technical Report B-3, Texas A&M, Institute of Statistics.

_____. (1982) Data Modeling Using Quantile and Density-Quantile Functions, Proceedings of 1980 Lisbon Academy of Sciences Symposium on Recent Advances in Statistics. Academic Press: New York. 23-52.

_____. (1983) Quantiles, Parametric-Select Density Estimation, and Bi-Information Parameter Estimators, Proceedings of the 14th Annual Symposium on the Interface of Computer Science and Statistics, New York: Springer Verlag.

_____. (1983) Entropy Interpretation of Goodness of Fit Tests, Technical Report B-8. Texas A&M, Institute of Statistics.

Prihoda, T. J. (1981) A Generalized Approach to the Two Sample Problem: The Quantile Appraoch, Technical Report B-5, Texas A&M, Institute of Statistics.

Pyke, R. (1970) Asymptotic Results for Rank Statistics in Nonparametric Techniques in Statistical Inference, ed. M. L. Puri, Cambridge: Cambridge University Press.

_____ and Shorack, G. (1968) Weak convergence of a two-sample empirical process and a new approach to Chernoff-Savage theorems. Ann. Math. Statist. 39, 755-71.

White, J. M. (1980) A Quantile Function Approach to the K-sample Quantile Regression Problem, Technical Report B-4, Texas A&M, Institute of Statistics.

Woodfield, T. J. (1982) Statistical Modeling of Bivariate Data, Technical Report B-7, Texas A&M, Institute of Statistics.

END

FILMED

5-83

DTIC