

June 17, 1998

**STATISTICAL METHODS MINING AND
NONPARAMETRIC QUANTILE DOMAIN DATA ANALYSIS**

by Emanuel Parzen
Department of Statistics
Texas A&M University

I am honored to attend the International Environmetrics Society (TIES) Ninth International Conference on Quantitative Methods for Environmental Sciences, July 3-6, 1998 in Australia. I am excited to speak at a Workshop to honor Ian MacNeill, who in his career has achieved an international reputation for excellence and leadership. Since he was my Ph.D. student at Stanford, I can claim him as a confirmation of my proverb that “to become a genius, the best way is to have students who are geniuses”.

ABSTRACT

This paper is a comprehensive review of some statistical data analysis methods whose theory I have been researching. Some aims of this paper are: (1.) to advocate the roles of quantile domain data analysis, using quantile functions and comparison density functions whose theory I have been exploring for 20 years (Parzen (1979), (1989), (1991), (1992), (1993)); (2.) unify conventional parametric and non-parametric statistics by representing them as score statistics (inner products or correlators of score functions and suitable density functions); (3.) show that expressing many test statistics as correlations generates new graphical diagnostic functions on the unit interval to find patterns in data. These concepts may appear

to be an *abstract* theory. I believe that they can be made computationally *concrete* and user-friendly. The only issue deciding their ultimate popularity are successful case studies which demonstrate to applied researchers that their presentation and interpretation can be made simple and powerful without having to be aware of the underlying theoretical and computational complexity. We propose the concept of “statistical methods mining” to describe this approach to applied research.

The mathematical statistical problems for which we discuss new methods are: one sample probability model identification (sections 3, 6) ; bivariate sample graphical representations alternative to scatter diagrams (section 8), cusums interpreted as correlations; applying two sample analysis to detect change and to discover homogeneous groups of data (sections 9–12); quantile limit theorems.

CONTENTS

0. Statistical methods mining and goals and coordinates of statistical data analysis
1. One sample notation
2. Quantile domain methods to measure location and scale
3. Box plots and sample shape identification quantile functions
4. Comparison distribution and comparison density functions
5. Change *PP* plots to compare discrete distributions
6. Quantile based rejection simulation using comparison density functions
7. Dependence comparison density functions for bivariate data
8. Quantile interpretation of correlations $R(X, Y)$, $R(X \leq Q_X(t), Y)$.
9. Two sample analysis and comparison density estimation
10. Two sample normal *t* tests, correlation $R(X = 1, Y)$
11. Two sample nonparametric Wilcoxon linear rank statistic, correlation $R(X = 1, F^{mid}(Y))$
12. Two sample omnibus nonparametric tests, maximum chi-square, accumulation chi-square, $R(X = Q_X(t), Y \leq Q_Y(u))$.
13. Quantile Limit Theorems.
14. Two sample comparison density function asymptotic distributions

0. Statistical methods mining and goals and coordinates of statistical data analysis

A main aim of this paper is to advocate “statistical methods mining” which I define to be the practice of data analysis by integrating (bridging, Brownian Bridging) the extensive range of classical and modern parametric, non-parametric, and function estimation statistical methods through frameworks for the diversity and history of statistical methods. The recent book “A History of Mathematical Statistics from 1750 to 1930” by Anders Hald (1998) demonstrates that statisticians have not done a satisfactory job of being aware of the contributions of past researchers (a goal of statistical methods mining). Statistical methods mining aims to benefit applied researchers by providing frameworks to learn history and the roles of various methods, to learn “why” and “when” as well as “how” and “what”.

To provide frameworks for the practice of statistics, and the roles of the discipline and profession of statistics, we propose three coordinates called: (1.) data model and notation, (2.) whole statistician, (3.) data inference.

The data model and notation coordinate formulates the problem in terms of the technical concepts, models, notation, and vocabulary of mathematical statistics. Hald (1998) demonstrates that notation has been slow to evolve but historically has been very important to developing a unifying discipline of statistics enabling technology transfer of statistical methods between different applied fields. David (1995) has collected a list of the “First (?) occurrence of common terms in mathematical statistics.” It is no joke that university professors explain to statis-

tics faculty why they could not recommend statistics courses to their students; “You will only confuse them because you use a different notation than we do.”

The whole statistician coordinate defines the goal of the proposed research, usually chosen from application (analyzing real data to answer real questions), theory (answering mathematical questions, forming estimators and studying their probability distribution under null and alternative hypotheses), and/or computation (developing and applying algorithms and software).

The data inference coordinate implements the goals defined in the whole statistician coordinate and encourages integrating conventional and modern estimation and testing procedures.

This paper discusses (1.) quantile domain of the data notation and models coordinate, (2.) theory and methods domain of the whole statistician coordinate, and (3.) some outlines of proofs and data analysis.

Practical statistical methods mining requires understanding about strategies for solving statistical problems and learning statistical methods (Parzen (1998)). We propose that both require a cycle of steps which one usually repeats (iterates) several times before reaching a satisfactory conclusion.

Our philosophy of data analysis is based on an attitude about the iterative process of data modeling. A model is equivalent to expectations (predictions) about what we observe. By comparing expectations (model) with reality (data) one can detect the *fit* of a model, develop diagnostics which help *revise* a model, and make *decisions* about the real models generating the observations in the data. We

summarize poetically our attitude about the common goals of science and statistics:
“The work of science (statistics) is to substitute facts (data) for appearances, and demonstrations (models) for impressions”.

1. One sample notation.

Coordinate 1 of statistical analysis of data X_1, \dots, X_n states data model and notation. Our first model assumes a random sample of a random variable X with unknown true distribution function

$$F(x) = P[X \leq x], \quad -\infty < x < \infty.$$

Many applications require (non-trivial) extensions to correlated data and censored data.

The standard assumptions of one sample statistical data analysis are tested by major fields of statistical theory whose analogies we believe should be learned (by learning what is a Brownian Bridge and what are its statistical applications).

1. Assumption of independent random variables is tested by stationary time series spectral analysis (test for trend or pattern in periodogram (squared Fourier transform), exponentially distributed function of frequency $w_j = j/n$).
2. Assumption of identically distributed random variables is tested by change-point analysis of cusums (test for trend or pattern in data plotted as a change density on unit interval $0 < \tau < 1$).
3. Assumption of normal (or other) parametric model for F is tested by goodness of fit and empirical processes theory (test for trend or pattern in change PP plot which compares model F^\wedge with data F^\sim).

The goal of statistical modeling is to estimate F , usually by identifying parametric models F_θ that approximately fit F (called model identification or selection). We believe that data analysis should be first parametric (fits by familiar probability laws), then nonparametric estimation of F by a process which revises a parametric model F^\wedge that fails to fit (section 6).

A distribution function F is called continuous with probability density function $f(x)$ if $F(x) = \int_{-\infty}^x f(y)dy$. An important (default) model is the Normal (μ, σ^2) parametric model

$$F(x) = F_{\mu,\sigma}(x) = \Phi\left(\frac{x - \mu}{\sigma}\right), \quad -\infty < x < \infty$$

where $\Phi(x)$ is the Normal (0,1) distribution function

$$\begin{aligned}\Phi(x) &= \int_{-\infty}^x \phi(y)dy, \\ \phi(x) &= (2\pi)^{-0.5} \exp(-x^2/2)\end{aligned}$$

Parametric estimation forms “efficient” (minimum information distance) estimators $\hat{\mu}$ and $\hat{\sigma}$ of the location parameter μ and scale parameter σ in a model

$$F(x) = F_0((x - \mu)/\sigma), \quad F^{-1}(u) = \mu + \sigma F_0^{-1}(u),$$

where $F_0(\cdot)$ is assumed known or required to be fitted (identified) from the data; the normal model assumes $F_0 = \Phi$. The parametric estimator F^\wedge for the true distribution F , assuming the parametric normal model, is

$$F^\wedge(x) = \Phi\left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right).$$

A *non-parametric estimator* of F is the sample distribution function

$$F^\wedge(x) = \text{fraction of sample } \leq x.$$

The important concept of mid- p value, due to Lancaster, can be represented by the sample mid-distribution function (Parzen (1996)) which we define, in terms of the sample probability mass function $p^{\sim}(x) = \text{fraction of sample equal to } x$, by

$$F^{\sim\text{mid}}(x) = F^{\sim}(x) - .5p^{\sim}(x).$$

An important tool is the sample rank transform

$$F^{\sim\text{mid}}(X(j; n)) = (j - .5)/n.$$

A sample (denoted X_1, \dots, X_n) has *order statistics* or *sample quantiles*, denoted $X(1, n) \leq \dots \leq X(n; n)$, which are the values in the sample arranged in increasing order. Sample quantiles provide powerful nonparametric and parametric methods of statistical analysis, data representation, and data compression of massive data sets because they can be applied to provide solutions to problems of simple (yet efficient) modeling of large amounts of data. Eisenberger and Posner (1965) describe how in the 1960's, at the Cal Tech Jet Propulsion Lab, they used sample quantiles to transmit and compress data from space probes, yet achieve efficient estimation of population parameters and tests of goodness of fit for very large samples.

When we observe bivariate data (X_j, Y_j) we should also report the data in order of increasing X values and denote them $(X(j; n), Y[j; n])$. We call $Y[j; n]$ the concomitant order statistics or co-order statistics.

Example: Let X be the level of estadiol in a fish and Y be the testosterone level in the same fish. We take fish samples from lake A (polluted) and lake B (unpolluted), and report the data:

Lake A (Polluted)

$X(j; n)$	23	30	37	38	53
$Y[j; n]$	24	7	6	22	8

Lake B (Unpolluted)

$X(j; n)$	15	16	19	27	29	36	64	72	85
$Y[j; n]$	33	54	60	12	47	75	20	53	100

One sees informally that levels Y of testosterone are much lower in lake A (polluted). Low testosterone levels reduce ability to reproduce. To assign a significance level (p -value) to this conclusion one must do a formal (contract) statistical analysis of the data.

An exploratory data analysis approach is to plot on the scatter diagram of each bivariate sample (see Figure 1) the sample conditional mean, defined (when all data points are distinct)

$$E^{\sim}[Y|X = x] = Y[j; n], \quad X^{\text{mid}}(j-1; n) < x < X^{\text{mid}}(j; n).$$

Mid-values of order statistics are defined (for $j = 1, 2, \dots, n-1$)

$$X^{\text{mid}}(j; n) = .5(X(j; n) + X((j+1); n));$$

$$X^{\text{mid}}(0; n) = X(1; n), \quad X^{\text{mid}}(n; n) = X(n; n).$$

Goodness of fit tests (Kolmogorov-Smirnov, Cramer-von Mises) of fit of a continuous model F^{\sim} to a discrete sample distribution F^{\sim} can be reviewed as combining statistics tests (section 5) of

$$F^{\sim}(X(j; n)) - F^{\sim\text{mid}}(X(j; n)) = U(j; n) - ((j - .5)/n),$$

called a mid-probability approximation, or

$$F^{\sim}(X^{\text{mid}}(j; n)) - F^{\sim}(X(j; n)) = U^{\text{mid}}(j; n) - (j/n),$$

called a continuity-correction approximation. These formulas (which assume no ties in data X_1, \dots, X_n) reflect two desirable ways of approximating discrete distributions (such as F^\sim) by continuous distributions (such as F^\wedge) which are more accurate than approximating $F^\sim(X(j; n))$ by $F^\wedge(X(j; n))$.

The variance (square of standard deviation) of the sample distribution F^\sim is denoted

$$\text{VAR}^\sim[X] = \frac{1}{n} \sum_{t=1}^n (X_t - \bar{X})^2$$

where $\bar{X} = E^\sim[X] = \frac{1}{n} \sum_{t=1}^n X_t$ is the sample mean. Statisticians need notation to clearly distinguish sample variance from the conventional definition, which I call “adjusted sample variance”:

$$S^2 = \text{VAR}^\sim_{\text{adj}}[X] = \frac{1}{n-1} \sum_{t=1}^n (X_t - \bar{X})^2$$

The two definitions of sample variance are calculated on every electronic calculator. Not using both definitions of sample variance makes many statistics textbooks very inelegant (for example, they define correlation coefficient with a divisor of $n-1$ and need complicated formulas for the pooled variance of two samples).

2. Quantile domain methods to measure location and scale

The quantile domain (of statistical inference, data analysis, and data modeling) is defined to be methods that use the inverse distribution or quantile function, denoted $Q(u) = F^{-1}(u)$, with precise definition as a left-continuous non-decreasing function:

$$Q(u) = \inf \{x : F(x) \geq u\}, \quad 0 < u < 1.$$

Note that a distribution function $F(x)$ is right-continuous.

When F is continuous with probability density f , $F(Q(u)) = u$ for all u and $f(Q(u))Q'(u) = 1$. We call $q(u) = Q'(u)$ the *quantile density* function; $fQ(u) = f(Q(u)) = 1/q(u)$ is called the *density quantile* function. This notation was introduced in Parzen (1979).

The sample quantile function $Q^\sim(u)$ is the inverse $F^{\sim-1}(u)$. In terms of the order statistics $X(j; n)$ one can give an explicit formula for the sample quantile function: for $j = 1, \dots, n$

$$Q^\sim(u) = X(j, n), \quad (j-1)/n < u \leq j/n.$$

This formula justifies calling order statistics sample quantiles (see Figures 2, 3, 4).

The sample quantile function is a “non-decreasing re-arrangement” of a function called a *change density* $\tilde{c}(\tau)$, $0 < \tau < 1$, of the time series X_j , which is the data arranged in order observed plotted as a function on the unit interval,

$$\tilde{c}(\tau) = X_j, \quad (j-1)/n < \tau < j/n.$$

When f is continuous and positive at $x = Q(u)$, as j tends to ∞ in such a way that $j/n \rightarrow u$, $0 < u < 1$, the asymptotic distribution of

$$\sqrt{n}(X(j; n) - Q(u)) \text{ is Normal}(0, u(1-u)(f(Q(u)))^{-2}).$$

Sheppard (1899) derived this asymptotic variance to study the errors of “method of quantiles” estimators (Hald (1998), p. 629). We outline a modern proof in section 13.

I regard this “sample quantile limit theorem” as an *existence theorem* for *statistical inference*. We do not need a parametric model for F to know a parametric model for the order statistics $X(j; n)$, thus avoiding the need to have some knowledge of (model for) the unknown distribution function F of X_j in order to infer it. Efficient estimators of parameters of model $Q(u) = \mu + \sigma Q_0(u)$, from uncensored and censored samples, can be formed from linear combinations of order statistics (using theory of regression analysis of continuous parameter time series Parzen (1979), Eubank (1988)). Applying this approach to censored survival data is a problem open for implementation (Parzen (1979)).

A continuous quantile $Q(p)$ can be interpreted as a parameter θ satisfying the estimating equation

$$F(\theta) - p = 0.$$

We call: $\theta = F^{-1}(p)$ an “inverse” parameter; $\theta = F(x) = E[I(X \leq x)]$ a “moment” parameter. If the distribution of a random variable X depends on a parameter θ , and if $T(X)$ is a statistic with mean $E_\theta[T(X)] = \tau$, we call τ a moment parameter and θ its inverse parameter. We regard θ and τ as inverse functions of each other, written $\tau = \tau(\theta)$, $\theta = \theta(\tau)$, $\tau = \theta^{-1}$, $\theta = \tau^{-1}$. The method of moments estimator of τ is $\hat{\tau} = \bar{T}$, the mean of $T(X_1), \dots, T(X_n)$, and the estimator of θ is $\hat{\theta} = \hat{\tau}^{-1} = \theta(\hat{\tau})$. The statistical properties of quantile estimators $Q^\sim(p)$ are prototypes for statistical properties of estimators $\hat{\theta}$ of parameters θ defined by estimating equations.

Important summary measures of a distribution and of a sample (originally proposed by Galton (1875), see Hald (1998), p. 602) are provided by the median

$MQ = Q2 = Q(.5)$, quartiles $Q1 = Q(.25)$ and $Q3 = Q(.75)$, and *quartile deviation*

$$DQ = 2\{Q(.75) - Q(.25)\}.$$

I believe that my definition of DQ is significant to applied statisticians because it states that the version of the inter-quartile range that is appropriate to measure scale is twice the inter-quartile range. Galton characterized scale or dispersion by the probable deviation, defined as half the interquartile range. We justify DQ as a numerical derivative of $Q(u)$ at $u = .5$ which crudely approximates another measure of scale (called the *density quantile deviation*)

$$DfQ = 1/fQ(.5) = 1/f(\text{median}) = q(.5) = Q'(.5).$$

Thus measures of location and scale are respectively provided by mean μ , MQ and standard deviation σ , DQ .

We recommend standardizations of probability distributions to make $f(\text{median}) = 1$ or $DQ = 1$. We call standardized quantile functions *shape identification* quantile function $QI(u)$. The standardized normal distribution satisfying median $MQ = 0$ and scale $f(\text{median}) = 1$ deserves more recognition, and we denote its distribution function $\Phi 1(y)$ and probability density function $\phi 1(y) = \exp(-\pi y^2)$. It is Normal $\{0, 1/2\pi\}$. Note $\Phi 1^{-1}(.75) = .675/2.5066 = .2693$ which we compare with $.25$; $\Phi 1^{-1}(.975) = .7819$; $\Phi 1^{-1}(.995) = 1.027$. Alternate definition $\Phi I(u)$ uses $DQ = 2.7$ as divisor, $\Phi I^{-1}(.75) = .25$, $\Phi I^{-1}(.995) = 2.575/2.7 = .9537$. We use $|QI(u)| > 1$ as a diagnosis of long tails or outliers.

In practice sample quantiles are computed by a variety of definitions. There is an extensive literature on nonparametric estimation of $Q(u)$ by smoothing $Q\tilde{(u)}$.

We recommend that the *functional* way to compute quantiles in practice is from a suitably defined continuous version $Q^{\sim\text{continuous}}(u)$ of the discrete sample quantile function $Q^{\sim}(u)$. Two possible continuous sample quantile functions are piecewise linear functions joining the points $(F^{\sim\text{mid}}(X(j; n)), X(j; n))$ or $(F^{\sim}(X(j; n)), X^{\text{mid}}(j; n))$. Using either of these definitions the median of a Poisson distribution with mean 2 is 1.84.

The sample median would be defined in practice as $Q^{\sim\text{continuous}}(.5)$. We conjecture that it is not important that this definition coincide with the traditional definition of sample median in terms of distinct order statistics $X_{(1)} < \dots < X_{(n)}$. When $n = 2m + 1$, an odd number, traditional sample median $= X_{(m+1)}$. When $n = 2m$, an even number, traditional sample median $= .5(X_{(m)} + X_{(m+1)}) = X_{(m)}^{\text{mid}}$.

3. Box plots and sample shape identification quantile functions

An important problem for applied statisticians is to model a sample quantile function $Q^{\sim}(u)$ by identification of probability distributions $Q_0(u)$, where $Q^{\sim}(u) = \hat{\mu} + \hat{\sigma}Q_0(u)$ fits $Q^{\sim}(u)$. We recommend Q_0 be standardized; $Q_0(.5) = 0$ and $Q_0'(.5) = 1$ or $2(Q_0(.75) - Q_0(.25)) = 1$. We discuss in section 6 models $Q^{\sim}(u)$ for which simulation is practical, so that we can create more data “like” that observed. To help identify suitable $Q_0(u)$, we propose plotting the *sample shape identification quantile function*, defined by (Parzen (1983))

$$QI^{\sim}(u) = (Q^{\sim}(u) - Q2^{\sim})/2(Q3^{\sim} - Q1^{\sim});$$

note $Q2^{\sim}$ is the sample median, and $Q1^{\sim}$ and $Q3^{\sim}$ are the sample first and third quartiles (possibly computed from a continuous version).

The identification of outliers, out-and-outliers, and tail behavior is a major goal of initial data analysis. Values u such that

$$|QI^{\sim}(u)| > 1$$

are diagnosed in our method as indicating outliers or long-tailed behavior. Note that Tukey's Box Plot diagnoses a value $QI^{\sim}(u)$ as an outlier if (in our notation)

$$QI^{\sim}(u) - QI^{\sim}(.75) > .75 \text{ or } QI^{\sim}(u) - QI^{\sim}(.25) < -.75$$

Note that $QI^{\sim}(.75) < .5$. Therefore $QI^{\sim}(u) > 1.25$ is an outlier.

Measures of skewness are: $fQ1 = -.25/QI(.25)$, $fQ2 = .25/QI(.75)$, $QI^{\sim}(.25) + .25$, $QI^{\sim}(.75) - .25$, $QI^{\sim}(.25) + QI^{\sim}(.75)$. These last three are zero for distributions symmetric about the median. If $fQ1 > 1$ we expect mode < median < mid-quartile < mean. If $fQ1 < 1$ we expect mean < midquartile < median < mode; midquartile $Q4 = .5(Q1 + Q3)$.

Measures of tail behavior are $QI^{\sim}(.05)$, $QI^{\sim}(.95)$; short tail if $QI^{\sim}(.95) < .5$, long tail if $QI^{\sim}(.95) > 1$, medium tail if $.5 < QI^{\sim}(.95) < 1$.

To learn how to use these plots one would have to study a portfolio of *population shape identification quantile functions* $QI(u)$ for various standard probability distributions. Figures 7,8 illustrate $QI^{\sim}(u)$ for samples from an exponential distribution of sizes 40 and 200. In an appendix we discuss simple examples of identification quantile analysis.

Example: Cushner-Peebles (1905) data, differences of excess hours of sleep under influence of two drugs, was analyzed by Student. Sample size $n = 10$. Order statistics $X(j; n)$ are: 0, 0.8, 1., 1.2, 1.3, 1.3, 1.4, 1.8, 2.4, 4.6.

Quantile diagnostics are $Q1 = 1$, $Q2 = 1.3$, $Q3 = 1.8$, $DQ = 2(Q3 - Q1) = 1.6$,

$$QI(.25) = (1 - 1.3)/1.6 = -.19, QI(.75) = (1.8 - 1.3)/1.6 = .3125,$$

$$QI(.05) = (0 - 1.3)/1.6 = -.81, QI(.95) = (4.6 - 1.3)/1.6 = 2.06.$$

The values which we can consider an outlier (those satisfying $|QI(u)| > 1$) are: 4.6.

Next we apply these diagnostics to the data set with 4.6 omitted: $MQ = 1.3$, $Q1 = .85$, $Q3 = 1.7$, $Q4 = 1.275$, $DQ = 1.7$, and diagnose a medium tailed symmetric distribution, since extreme QI values are $-.76$ and $.65$. Note $\Phi^{-1}(0.0556) = 1.6/2.7 = .59$, suggesting that we investigate an adequate fit of the sample quantile by the normal quantile.

A popular quick and dirty tool for statistical data analysis is the Box-Plot, made famous by Tukey. Our variation, called a “shape identification quantile box-plot”, plots $QI(u)$, $0 < u < 1$, with: (1) a dashed box over the interval $(.25, .75)$ representing the lower quartile, median, and upper quartile, used to diagnose symmetry and skewness; (2) an L shape over $(.16, .25)$ whose horizontal segment is at mean μ and vertical segment is drawn from $\mu - \sigma$ to μ ; (3) an L shape over $(.75, .84)$ whose horizontal segment is at mean μ and vertical segment is drawn from μ to $\mu + \sigma$, where σ is standard deviation, (4) dashed horizontal lines over $(0,1)$ at height 1 and -1, used to classify tail behavior as short tail, medium tail (medium-short, medium-medium, medium-long), or long tail.

Numerical definitions of tail behavior (and tail indices α_0 and α_1) can be given (Parzen (1979)) in terms of the behavior at $u = 0, 1$ of $fQ(u)$ and the score function

$J(u) = -(fQ(u))'$. We define

$$\alpha_0 = \lim_{u \rightarrow 0} \frac{-uJ(u)}{fQ(u)}, \quad \alpha_1 = \lim_{u \rightarrow 1} \frac{(1-u)J(u)}{fQ(u)}.$$

Short, medium, and long tail correspond to $\alpha < 1$, $\alpha = 1$, $\alpha > 1$. Current research on nonparametric estimation of these parameters can be searched under the name of “Hill’s estimators”.

If $1 - F(x) \sim x^{-\gamma}$ as $x \rightarrow \infty$, then $1 - u \sim Q(u)^{-\gamma}$ and $\alpha = 1 + (1/\gamma)$ as $u \rightarrow 1$. The tail index whose estimates have familiar properties is $1/\gamma$, not γ . Cauchy distribution has $\alpha = 2$, and mean is infinite. Variance is infinite if $\alpha > 1.5$.

Exponential distribution $1 - F(x) = e^{-x}$ has $1 - u = \exp(-Q(u))$, $fQ(u) = 1 - u$, $\alpha_1 = 1$, $\alpha_0 = 0$, $QI1 = -.1845$, $QI3 = .3155$.

Gamma distribution with parameter κ , with standard probability density

$$f_0(x) = x^{\kappa-1} e^{-x} / \Gamma(\kappa), x > 0,$$

has $f(x) \sim x^\kappa$ as $x \rightarrow 0$. The shape parameter to estimate tail behavior is $\beta = 1/\kappa$, since $\alpha_0 = 1 - \beta$ and $\alpha_1 = 1$. An analogous probability law is the Weibull distribution with parameter κ , with standard probability density

$$f_0(x) = \kappa x^{\kappa-1} e^{-x^\kappa}, x > 0,$$

$$Q_0(u) = (-\log(1 - u))^\beta,$$

$$f_0 Q_0(u) = (1/\beta)(1 - u)(-\log(1 - u))^{1-\beta}.$$

$$f_0 Q_0(u) / f_0 Q_0(.5) = 2(1 - u)(-\log(1 - u))^{1-\beta} / (\log 2)^{1-\beta}$$

4. Comparison distribution and comparison density functions.

The true distribution function F of a continuous random variable X can be characterized by the property that the transformed random variable $U = F(X)$ is Uniform(0,1). The usual approach to a goodness of fit test of a model F given a sample $X_j, j = 1, \dots, n$, is to transform the original data to $U_j = F(X_j)$ whose distribution is tested to be Uniform (0,1). We develop an approach which compares a model F and a true distribution function G which is estimated by the sample distribution F^{\sim} , using concepts of comparison distribution and comparison density functions.

Identifying a distribution function F for a random variable X , given a random sample X_1, \dots, X_n , requires measures of the distance between F and the sample distribution function F^{\sim} . To compare two probabilities p_1 and p_2 we recommend their ratio p_1/p_2 rather than their difference $p_1 - p_2$. Applying this philosophy to comparing two *continuous* distribution functions $F(x)$ and $G(x)$ we define

$$D(u; F, G) = G(F^{-1}(u)), \quad 0 < u < 1$$

called the comparison distribution function. Its density, called the comparison density function, satisfies

$$d(u; F, G) = D'(u; F, G) = \frac{g(F^{-1}(u))}{f(F^{-1}(u))}$$

We require $f(x) = 0$ implies $g(x) = 0$ in order for $d(u; F, G)$ to be well defined and integrate to 1. The notation for this condition is $G \ll F$.

When fitting a parametric family of distributions $F_{\theta}(x)$ one often specifies a null distribution $F_{\theta_0}(x)$; to study local alternative hypotheses important tools are

(Nikitin (1995))

$$D(u; \theta) = D(u; F_{\theta_0}, F_{\theta}), d(u; \theta) = d(u; F_{\theta_0}, F_{\theta}).$$

When F_{θ} is Normal($\theta, 1$) and $\theta_0 = 0$ one obtains

$$\log d(u; \theta) = \theta \Phi^{-1}(u) - .5\theta^2$$

which is a *linear* function of $\Phi^{-1}(u)$.

When F_{θ} is Normal($0, \theta^{-2}$), and $\theta_0 = 1$, one obtains

$$\log d(u; \theta) = \log \theta - .5 (\Phi^{-1}(u))^2 (\theta^2 - 1)$$

which is a *quadratic* function of $\Phi^{-1}(u)$.

These results help us interpret the various shapes that comparison density functions need to have for us to interpret them as indicating that the difference between two distributions F and G is a *difference in location* or a *difference in scale*. When a Normal model for a data set has correct location but too large an estimated standard deviation (due to symmetric outliers), the comparison density function of data and model will have a shape indicating a difference in scale.

5. Change PP plots to compare discrete distributions.

We are proud of our extension (Parzen (1993)) of the concepts of comparison distribution $D(u; F, G)$ and density $d(u; F, G)$ to F and G either continuous or discrete. Parzen (1997) calls such concepts *concrete statistics*, since they help unify methods for CONtinuous and disCRETE data.

For discrete distributions F and G with respective probability mass functions p_F and p_G , we define, assuming $p_G(x) > 0$ implies $p_F(x) > 0$ or $G \ll F$,

$$d(u; F, G) = \frac{p_G(F^{-1}(u))}{p_F(F^{-1}(u))},$$

$$D(u; F, G) = \int_0^u d(s; F, G) ds.$$

An important observation: $D(u, F, G) = G(F^{-1}(u))$ at values of u such that $F(F^{-1}(u)) = u$, called F -exact values of u . At other values of u $D(u; F, G)$ is defined by linear interpolation between its values at F exact values. The important concept of F exact values is a concept of concrete statistics, unification of discrete and continuous data analysis.

When F and G are discrete the graph of $D(u; F, G)$ is called a PP plot because it linearly connects the points $(0,0)$, $(1,1)$, $(F(x), G(x))$ for F – exact $u = F(x)$. We recommend in practice “change PP plots” which graph $D(u; F, G) - u$ by connecting the points $(F(x), G(x) - F(x))$.

A QQ plot, or quantile-quantile plot, of two distributions intuitively graphs the points $(F^{-1}(u), G^{-1}(u))$. A QQ plot is linear for a location scale model

$$G^{-1}(u) = \mu + \sigma F^{-1}(u).$$

One often interprets QQ plots in terms of how they differ from being linear. This method of testing for normality or exponentiality is widely used but seems to have little supporting theory. We recommend that the use of QQ plots should be supplemented by change PP plots and plots of two quantile functions (see Figures 2, 3, 4, 5, 6).

Goodness of fit tests for a discrete distribution (such as a binomial(k, p)) with outcomes $j = 0, 1, \dots, k$, compare a model of probabilities p_j , defined =probability observe outcome j , with empirical probability p_j^{\sim} , defined as observed relative frequencies $p_j^{\sim} = N_j/n$, where N_j is the number of observations of outcome j in n observations. The traditional chi-square statistic C can be represented (in terms of probabilities rather than counts N_j and np_j)

$$\begin{aligned} C &= n \sum_{j=0, \dots, k} (p_j^{\sim} - p_j)^2 / p_j \\ &= n \sum_{j=0, \dots, k} p_j ((p_j^{\sim} / p_j) - 1)^2 \\ &= n \int_0^1 (d(u; F, F^{\sim}) - 1)^2 du. \end{aligned}$$

One can use as a test statistic any measure of the distance of $d(u; F, F^{\sim})$ from 1.

Cumulative goodness of fit tests are defined in terms of cumulative probabilities

$$F_j = p_0 + p_1 + \dots + p_j, F_j^{\text{mid}} = F_j - .5p_j, F_j^{\sim} = p_0^{\sim} + p_1^{\sim} + \dots + p_j^{\sim},$$

$$CPP_j = F_j^{\sim} - F_j = D(F_j; F, F^{\sim}) - F_j,$$

and measure distance of function $\{CPP_j, j = 0, \dots, k - 1\}$ from zero. Statistic at an outcome value j to test the significance of the difference between the theoretical probability F_j and the empirical probability F_j^{\sim} is

$$Z_j = CPP_j / (F_j(1 - F_j))^{\cdot 5}, j = 0, 1, \dots, k - 1.$$

An analogue of a Wilcoxon type omnibus test statistic is $n^{\cdot 5}CPP^{\sim}$,

$$CPP^{\sim} = \sum_{j=0, \dots, k-1} p_j CPP_j.$$

An alternative statistic is the area under the continuous curve $CPP(u)$, given by

$$\begin{aligned} CPP^{\text{mid-}} &= \sum_{j=0, \dots, k} p_j CPP_j^{\text{mid}}, \\ CPP_j^{\text{mid}} &= .5(CPP_j + CPP_{j+1}). \end{aligned}$$

We define analogues of Cramer-von Mises W^2 , Watson U^2 , Anderson-Darling A^2 , Kolmogorov-Smirnov KS which are non-linear functionals of $D(u; F, F^\sim) - u$, $u = F_0, \dots, F_{k-1}$:

$$\begin{aligned}
W^2 &= 6n \sum_{j=0, \dots, k-1} p_j CPP_j^2 \\
&= 6n \sum_{j=0, \dots, k-1} p_j F_j (1 - F_j) Z_j^2, \\
U^2 &= n \sum_{j=0, \dots, k-1} p_j (CPP_j - CPP)^2, \\
A^2 &= n \sum_{j=0, \dots, k-1} p_j CPP_j^2 / F_j (1 - F_j) \\
&= n \sum_{j=0, \dots, k-1} p_j Z_j^2, \\
KS &= n^{.5} \max_{j=0, \dots, k-1} |CPP_j|.
\end{aligned}$$

Anderson-Darling variations are

$$\begin{aligned}
A_0^2 &= 2n \sum_{j=0, \dots, k-1} p_j (1 - F_j) Z_j^2, \\
A_1^2 &= 2n \sum_{j=0, \dots, k-1} p_j F_j Z_j^2,
\end{aligned}$$

Note $A^2 = .5(A_0^2 + A_1^2)$; therefore calculating A_0^2 and A_1^2 immediately yields third statistic A^2 . Kolmogorov Smirnov variations are

$$\begin{aligned}
KS_0 &= n^{.5} \max_{j=0, \dots, k-1} CPP_j, \\
KS_1 &= n^{.5} \min_{j=0, \dots, k-1} CPP_j, \\
KSZ &= n^{.5} \max_{j=0, \dots, k-1} |Z_j|.
\end{aligned}$$

Properties of similar statistics for goodness of fit of discrete distributions, and their components, are discussed by D. J. Best and J. C. W. Rayner (1997, Goodness of fit for the binomial distribution, *Australian Journal of Statistics*, 39, 355-364).

Components are linear functional statistics of the form

$$\begin{aligned}
T_\phi &= n^{.5} \sum_{j=0, \dots, k-1} p_j \tilde{\phi}(F^{-1}(F_j)) \\
T_J &= n^{.5} \int_0^1 d(u; F, F^\sim) J(u) du
\end{aligned}$$

where $\phi(j)$ and $J(u) = \phi(F^{-1}(u))$ are orthogonal polynomials called score functions.

A goodness of fit analysis consists of four phases:

1. Graphical analysis of the Change *PP* plot $D(u; F, F^\sim) - u$, $0 < u < 1$.
2. Nonlinear functionals such as W^2 and A^2 ,
3. Linear functionals T_J for suitable score functions $J(u)$,
4. Smoothing the sample comparison density function $d^\sim(u) = d(u; F, F^\sim)$ and testing if the optimal smoother is $d^\sim(u) = 1$ (which is equivalent to the null hypothesis that the model F is a good fit to the observations F^\sim). Smoothing $d^\sim(u)$ to form $\hat{d}^\sim(u)$ provides a nonparametric model for the true probability mass function $p(j)$.

6. Quantile based rejection simulation using comparison density functions

A fundamental property of quantile functions (Parzen (1979)) is the following theorem for monotone transformation $g(Y)$ of a random variable Y where g is non-decreasing and left-continuous:

$$Q_{g(Y)}(u) = g(Q_Y(u)), F_{g(Y)}(y) = F_Y(g^{-1}(y))$$

This theorem and our notation provide a quick proof that Y with quantile function $Q(u)$ has the same distribution as $Q(U)$ where U is Uniform[0, 1] (with quantile function $Q_U(u) = u$): $Q_{Q(U)}(u) = Q(Q_U(u)) = Q(u) = Q_Y(u)$.

A random sample Y_1, \dots, Y_n of Y can be simulated by $Y_j = Q(U_j)$ where U_1, \dots, U_n are a random sample from Uniform{0, 1}. A bootstrap sample Y_1, \dots, Y_N

can be simulated by $Y_j = Q^\sim(U_j)$ where $Q^\sim(u)$ is a sample quantile function formed from an original sample. Smooth bootstrap simulates $Y_j = Q^\wedge(U_j)$.

A *rejection* method simulates Y with distribution F from a sample from a distribution G . We formulate it in terms of the comparison density $d(u) = d(u; G, F)$, if we can assume for all u , $d(u) \leq c$ for a known constant c . Generate two independent Uniform (0,1) random variables U and W . The acceptance-rejection rule is: accept $Y = G^{-1}(U)$ as a sample from F if $W \leq d(U)/c$. The probability of acceptance is $1/c$. The probability that $U \leq G(y)$ and $W \leq d(U)/c$ equals $D(G(y))/c = F(y)/c$. The probability that our rejection simulated $Y \leq y$ equals $F(y)$. We say that this process simulates F by modeling its distance from G , a distribution which is easy to simulate.

We recommend to model the true distribution F of data by combining parametric modeling, goodness of fit, and density estimation: (1.) estimate a parametric model G^\wedge ; (2.) smooth the raw comparison density $d(u; G^\wedge, F^\wedge)$ and test whether $d(u) = 1$ is a good fit; (3.) thus arrive at nonparametric model f^\wedge which can be simulated by the rejection method from samples from G^\wedge . We call this 3-step process a “model” because it achieves a simulation goal of statistical data analysis: from the data observed infer an algorithm for generating additional data whose distribution is not significantly different from the originally observed data.

7. Dependence comparison density functions for bivariate data.

We use comparison distributions to provide new concepts to model the dependence of joint random variables X and Y which are either discrete or continuous

(Parzen (1992)). Our approach to modeling bivariate data (X, Y) compares the unconditional distribution F_Y with the conditional distribution of Y given $X = x$, measuring x as $Q_X(t)$, by a *dependence* comparison distribution

$$D(u; F_Y, F_{Y|X=Q_X(t)}), \quad 0 < u, t < 1.$$

The dependence comparison density functions can be shown to obey an *important identity* which is a form of Bayes rule: for $0 < u, t < 1$,

$$d(t, u) = d(u; F_Y, F_{Y|X=Q_X(t)}) = d(t; F_X, F_{X|Y=Q_Y(u)}).$$

The proof is immediate in the jointly continuous case, since then

$$d(t, u) = f_{X,Y}(Q_X(t), Q_Y(u)) / f_X Q_X(t) f_Y Q_Y(u).$$

The difficult case to prove is when X is discrete, Y is continuous.

Another important function is a *change* comparison distribution which compares F_Y with the conditional distribution of Y given $X \leq Q_X(\tau)$. It is defined for $0 < u, \tau < 1$ and denoted

$$D([0, \tau], u) = D(u; F_Y, F_{Y|X \leq Q_X(\tau)})$$

with change comparison density

$$d([0, \tau], u) = d(u; F_Y, F_{Y|X \leq Q_X(\tau)});$$

they compare two samples of Y values indexed by whether $X \leq Q_X(\tau)$ or $X > Q_X(\tau)$. A *Fundamental Lemma* of Concrete Statistics is: for τ X -exact and u

Y -exact

$$\tau d([0, \tau], u) = \int_0^\tau d(t, u) dt.$$

We believe this formula explains why estimators of the density on the left have the asymptotic distribution properties of probabilities (section 14).

From this lemma we obtain a “cusum” formula: for τ X -exact

$$\tau E[Y|X \leq Q_X(\tau)] = \int_0^\tau dt E[Y|X = Q_X(t)]$$

Proof (an exercise in quantile calculus; obvious for a sample):

$$\begin{aligned} \tau E[Y|X \leq Q_X(\tau)] &= \tau \int_0^1 du Q_Y(u) d(u; F_Y, F_{Y|X \leq Q_X(t)}) \\ &= \int_0^1 du Q_Y(u) \int_0^\tau dt d(t, u) \\ &= \int_0^\tau dt \int_0^1 du Q_Y(u) d(t, u) \\ &= \int_0^\tau dt E[Y|X = Q_X(t)] \end{aligned}$$

We can apply these algorithms when X is a deterministic rather than random variable. In particular X can be $1, \dots, n$, the index t of Y values Y_t , $t = 1, \dots, n$. Applications to changepoint detection are discussed in Parzen (1992), (1994). The goal of our research on theoretical statistics is to develop analogies between statistical methods which help unify their application. We propose analogies between testing for change and testing for dependence.

Statistics to test for change in a sequence X_t are

$$R(t, X_t),$$

cusum

$$C(\tau) = R(I(t \leq [n\tau], X_t))$$

scored rank cusum

$$C_{\text{Rank}}(\tau) = R(I(t \leq [n\tau], J(F^{\text{mid}}(X_t)))).$$

The cumulative spectral distribution of stationary time series analysis is the cusum process of set of score statistics for $j = 1, 2, \dots, [n/2]$

$$R^2(\cos 2\pi t j/n, X_t), R^2(\sin 2\pi t j/n, X_t);$$

sum is equivalent to periodogram at frequency $w_j = j/n$.

8. Quantile interpretation of correlations $R(X, Y), R(X \leq Q_X(t), Y)$.

The Quantile Domain provides many new correlations and functions for studying the relation between random variables X and Y . As an alternative to a scatter diagram of a sample of (X, Y) we propose graphing diagnostic functions which are suggested by a formula for correlation coefficient in the quantile domain. Define standardizations

$$S_X(X) = (X - E[X])/\sigma_X, S_Y(Y) = (Y - E[Y])/\sigma_Y.$$

Note the representation of conditional expectation

$$E[S_Y(Y)|X = Q_X(t)] = \int_0^1 S_Y(Q_Y(u))d(t, u)du.$$

Correlation coefficient $R(X, Y)$ can be represented using conditional expectations.

$$\begin{aligned} R(X, Y) &= E[S_X(X)S_Y(Y)] \\ &= E[S_X(X)E[S_Y(Y)|X]] \end{aligned}$$

In terms of quantile functions this abstract formula becomes a concrete formula

$$R(X, Y) = \int_0^1 S_X(Q_X(t))E[S_Y(Y)|X = Q_X(t)]dt$$

Thus a correlation coefficient is the correlator or inner product (integral of the product of two functions) of the following graphical or visualization diagnostic functions (illustrated in Figures 9, 10, 11):

$S_X(Q_X(t))$, the standardized quantile function of X ,

$E[S_Y(Y)|X = Q_X(t)]$, the change density of Y given X .

When X and Y are bivariate normal,

$$S_X(Q_X(t)) = \Phi^{-1}(t),$$

$$E[S_Y(Y)|X = Q_X(t)] = R(X, Y)\Phi^{-1}(t)$$

The change density is the function which one seeks to estimate by a nonparametric regression; we denote it $c_{S_Y(Y)}(t)$. Another important graphical diagnostic function is the “change process of Y given X ”, defined on $0 < t < 1$, which is a “cusum”,

$$C_{S_Y(Y)}(t) = tE[S_Y(Y)|X \leq Q_X(t)] = \int_0^t E[S_Y(Y)|X = Q_X(s)]ds.$$

From the formula (end of section) for Indicator Correlations we can show

$$R(X \leq Q_X(t), Y) = (t(1 - t))^{-.5}C_{S_Y(Y)}(t);$$

in words we can interpret a change process as a series of indicator correlations.

Diagnostics of change (dependence) are the omnibus statistics

$$\max_t R^2(X \leq Q_X(t), Y), \int_0^1 R^2(X \leq Q_X(t), Y)dt$$

as well as score statistics which test specified shapes. Theorems are known for the asymptotic distribution of the statistics under the null hypothesis of independence; see Csörgő and Horváth (1997).

In summary, to diagnose the change density we form its inner product with various score functions. Using the quantiles $S_X(Q_X(t))$ as the score function forms $R(X, Y)$. The indicator function of s for t fixed

$$\begin{aligned} I(s \leq t) &= 1 \text{ if } s < t \\ &= 0 \text{ if } s > t, \end{aligned}$$

forms the change process of Y given X and $R(X \leq Q_X(t), Y)$. Orthogonal polynomials, such as Legendre polynomials on the unit interval, are score function $K(s)$, which provide general non-parametric diagnostics,

$$\int_0^1 K(s)E[S_Y(Y)|X = Q_X(s)]ds;$$

we usually require

$$\int_0^1 K(s)ds = 0, \int_0^1 K^2(s)ds = 1.$$

Double score statistics are of the form

$$\int_0^1 ds K(s) \int_0^1 du J(u) d(s, u).$$

A conventional correlation is also a “double score” statistic of the joint dependence density $d(t, u)$:

$$R(X, Y) = \int_0^1 dt \int_0^1 du S_X(Q_X(t)) S_Y(Q_Y(u)) d(t, u).$$

Rank correlations correspond to score functions which are orthogonal polynomials.

The discrete-continuous data unification ability of correlations comes from their relations to conditional means when one of the variables being correlated is an indicator. Indicator random variables are defined, for any set B of real numbers, $I(X \text{ is in } B) = 1$ or 0 , as X is in B or X is not in B . From the general formula: for function $g(Y)$ and set B of real numbers

$$E[g(Y)|X \text{ is in } B] = E[g(Y)I(X \text{ is in } B)]/P[X \text{ is in } B]$$

one can show a *Fundamental Lemma*:

$$\begin{aligned} R(X \text{ is in } B, g(Y)) &= CORR[I(X \text{ is in } B), g(Y)] \\ &= (\text{odds } P[X \text{ is in } B])^{-5} E[S(g(Y))|X \text{ is in } B] \end{aligned}$$

where $S(g(Y)) = (g(Y) - \text{mean}(g(Y)))/\text{stdev}(g(Y))$ and odds $(p) = p/(1 - p)$.

9. Two sample analysis and comparison density estimation.

Parametric Student t -tests and nonparametric Wilcoxon tests of the homogeneity of two samples can be represented as a (bi-serial) correlation coefficient $R(X = 1, g(Y))$ between a function $g(Y)$ of the pooled sample of continuous response Y , and the indicator $I(X = 1)$ that an observation came from sample 1. Our quantile representation of correlation provides ways to “look at the data” by interpreting the correlation interpretation of a numerical test statistic

$$R(X = 1, g(Y)) = \int_0^1 du S(g(Q_Y(u))) E[S(I(X = 1))|Y = Q_Y(u)].$$

The first integral $S(g(Q_Y(u)))$ changes as we change g : it is a score function whose shape we are seeking to correlate with the second integrand which is the ultimate

function to be estimated and which we express, letting $\tau_1 = P(X = 1)$.

$$E[S(I(X = 1)|Y = Q_Y(u))] = (\text{odds } (\tau_1))^{.5} \left(\frac{P[X = 1|Y = Q_Y(u)]}{P(X = 1)} - 1 \right).$$

These quantile domain formulas lead us to a philosophy about the *ultimate* aim of statistical analysis of two samples: it is to estimate the comparison density

$$\begin{aligned} d(u; F_Y, F_{Y|X=1}) &= d(1; F_X, F_{X|Y=Q_Y(u)}) \\ &= \frac{P[X = 1|Y = Q_Y(u)]}{P[X = 1]} \\ &= \frac{f_{Y|X=1}(Q_Y(u))}{f_Y(Q_Y(u))} \end{aligned}$$

where F_Y is the distribution function in the pooled sample of the responses in samples one ($X = 1$) and two ($X = 2$).

Two sample problems are fundamental problems of applied statistics which compare treatment and control, or yesterday, today, and tomorrow. The comparison of two samples seeks to determine if there has been a change (the probability distributions of the two samples are different, a conclusion which we call non-stationary or heterogeneity) or if the difference between the two samples can be explained by fluctuations (a conclusion which we call not rejecting the null hypothesis that the two samples have the same probability distribution).

Applied statisticians routinely use to analyze two samples the two-sample t -test (several versions), the Wilcoxon rank sum test, and the two sample Anderson-Darling test. These important methods are badly and briefly discussed in most introductory statistics textbooks. We propose that thinking in the comparison density domain will provide us with a two-sample analysis strategy which unifies

for both continuous and discrete data the use of these usually separate statistical methods.

In mathematical statistics we formulate the two sample problem as the comparison of the continuous distributions F and G of variables X and Y respectively from INDEPENDENT samples X_1, \dots, X_m and Y_1, \dots, Y_n with respective sample distribution functions denoted $F_m(x)$ and $G_n(y)$. Let $X(j; m)$ and $Y(k; n)$ denote the respective order statistics of the samples (assumed to be distinct values for ease of exposition).

We call two sample analysis unpooled if it directly estimates the comparison distribution

$$D(u; F, G) = G(F^{-1}(u)), 0 < u < 1.$$

This requires assumptions (that are not always satisfied by F and G) that $D(0; F, G) = 0$, $D(1; F, G) = 1$. We call the discrete comparison distribution

$$D^{\sim}(u; F, G) = G_n(F_m^{-1}(u)), 0 < u < 1$$

an *unpooled* estimator of the comparison distribution $D(u; F, G)$.

When considering the asymptotic theory we assume that m/n tends to a limit which we represent in terms of

$$\tau_1 = \tau_m = m/(m+n), \tau_2 = \tau_n = n/(m+n),$$

the proportions of the total sample size $m+n$ in each sub-sample.

A *pooled* estimator of the comparison distribution of two continuous distributions (from independent samples of each) forms the pooled sample distribution

$$H_{m+n}(x) = \tau_m F_m(x) + \tau_n G_n(x)$$

which is regarded as an estimator of a pooled population distribution

$$H(x) = \tau_m F(x) + \tau_n G(x).$$

Assumptions of common support of F and G are not required in order to define the comparison distribution

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

and its continuous estimator (a triumph of our definition of the comparison distribution function of two discrete distributions)

$$D(u; H_{m+n}, F_m)$$

which linearly connects its values $F_m(H_{m+n}^{-1}(u))$ at H_{m+n} -exact u (the values $u = H_{m+n}(x)$ for some x , which are of the form $j/(m+n)$ when all values are distinct). Parzen (1999) argues that the asymptotic distribution theory of the pooled estimator follows immediately via the Quantile Limit Theorem of Doss and Gill (1992) from the easier asymptotic theory of the unpooled estimator.

An important note about notation: In terms of the notation at the beginning of this section H is equivalent to F_Y , F is equivalent to $F_{Y|X=1}$. Instead of X and Y representing the responses in the two samples, at the beginning of this section Y denotes the response variable, X represents the population being sampled, and the two sample problem is viewed as a paired sample (X, Y) problem.

We outline in sections 10 and 11 how conventional two sample t -statistic and Wilcoxon statistic can be expressed in terms of suitable score functions and the

comparison density function on the unit interval $0 < u < 1$

$$d(u) = d(u; H, F) = d(u; F_Y, F_{Y|X=1}).$$

A raw estimator of $d(u)$ is provided by (the piecewise constant function)

$$\tilde{d}(u) = d(u; H^{\sim}, F^{\sim}) = d(u; F^{\sim}_Y, F^{\sim}_{Y|X=1})$$

Important test statistics are score statistics or linear detectors or *linear rank statistics*

$$T(J) = \int_0^1 J(u) \tilde{d}(u) du.$$

The ultimate aim of two sample data analysis (to form a smooth estimator $\hat{d}(u)$, $0 < u < 1$) can be accomplished by a wide variety of methods for density estimation. A variety of norms, or entropy measures, of $\hat{d}(u)$ can be used to detect changes in the distributions of samples. We conjecture that our two sample data analysis methods apply equally to small and massive samples.

10. Two sample normal t tests, correlation $R(X = 1, Y)$

This section states the traditional two sample t test in terms of correlation.

Define: sample mean of sample j

$$\mu_j^{\sim} = E^{\sim}[Y|X = j];$$

sample mean of pooled sample,

$$\mu^{\sim} = \tau_1 \mu_1^{\sim} + \tau_2 \mu_2^{\sim};$$

sample variance of pooled sample

$$\sigma^2 = VAR[Y] = E[VAR[Y|X]] + VAR[E[Y|X]];$$

variance $\sigma_j^2 = VAR[Y|X = j]$ of sample j ;

$$\sigma_{ave}^2 = E[VAR[Y|X]] = \tau_1\sigma_1^2 + \tau_2\sigma_2^2, \text{ pooled variance.}$$

Verify

$$\mu_1 - \mu = \tau_2(\mu_2 - \mu)$$

$$\mu_2 - \mu = \tau_1(\mu_1 - \mu)$$

$$\begin{aligned} VAR[E[Y|X]] &= \tau_1(\mu_1 - \mu)^2 + \tau_2(\mu_2 - \mu)^2 \\ &= \tau_1\tau_2(\mu_1 - \mu_2)^2 \end{aligned}$$

Define (omitting \sim for ease of writing)

$$1 - R^2 = E[VAR[Y|X]]/VAR[Y] = \sigma_{ave}^2/\sigma^2,$$

$$R^2 = VAR[E[Y|X]]/VAR[Y] = \tau_1\tau_2(\mu_1 - \mu_2)^2/\sigma^2$$

The traditional t -statistic (to test homogeneity of two samples obeying respective parametric models $\text{Normal}(\mu_1, \sigma^2)$ and $\text{Normal}(\mu_2, \sigma^2)$) is $(n - 2)^{.5}T$, defining

$$\begin{aligned} T &= (\tau_1\tau_2)^{.5}(\mu_1 - \mu_2)/\sigma_{ave} \\ &= (\tau_1/\tau_2)^{.5}(\mu_1 - \mu)/\sigma_{ave}. \end{aligned}$$

Our favorite representation is

$$T = R/(1 - R^2)^{.5},$$

where R is a sample correlation coefficient; omitting $\tilde{}$ for ease of writing

$$\begin{aligned}
R(X = 1, Y) &= R(I(X = 1), Y) = E[S(I(X = 1))S(Y)] \\
&= E[((I(X = 1) - \tau_1)/(\tau_1(1 - \tau_1)))^{\cdot 5} E[S(Y)|X]] \\
&= \tau_1(1 - \tau_1)/(\tau_1(1 - \tau_1))^{\cdot 5}(\mu_1 - \mu)/\sigma \\
&\quad + \tau_2(1 - \tau_1)/(\tau_1(1 - \tau_1))^{\cdot 5}(\mu_2 - \mu)/\sigma \\
&= (\tau_1\tau_2)^{\cdot 5}(\mu_2 - \mu_1)/\sigma \\
&= (\tau_1/\tau_2)^{\cdot 5}(\mu_1 - \mu)/\sigma = R.
\end{aligned}$$

A statistic which can be represented as an asymptotically Normal $(0, 1/n)$ difference of two means we call a Z statistic. The most familiar example of a Z statistic is the Student t statistic to test mean of a one sample Normal.

To bridge parametric, nonparametric, and function estimation, we prepare to regard T as a score statistic by expressing

$$\begin{aligned}
(\mu_1\tilde{} - \mu\tilde{})/\sigma\tilde{} &= E\tilde{}[S(Y)|X = 1] \\
&= \int_0^1 J(s)d\tilde{}(s)ds,
\end{aligned}$$

where the score function

$$J(s) = S(Q_{Y\tilde{}}(s)) = (Q_{Y\tilde{}}(s) - \mu\tilde{})/\sigma\tilde{}.$$

The Wilcoxon statistics can be represented as a score statistic with score function

$$J(s) = (12)^{\cdot 5}(s - \cdot 5).$$

11. Two sample nonparametric Wilcoxon linear rank statistic, correlation $R(X = 1, F^{mid}(Y))$

The non-parametric two-sample Wilcoxon rank-sum test is

$$(1/n_1) \sum_{j=1}^{n_1} R_j = nE[F_Y^{\sim}(Y)|X = 1]$$

where R_j are ranks in pooled sample of observations in sample 1. To convert the ranks to a number between 0 and 1, one can form $R_j/(n + 1)$ or $(R_j - .5)/n$. We prefer the latter because it can be expressed in terms of the mid-rank transform $F_Y^{\sim mid}(Y)$ with mean .5 and variance (a remarkable formula!)

$$\begin{aligned} \sigma_{rank}^2 &= VAR[F_Y^{\sim mid}(Y)] \\ &= (1/12)(1 - E[|p_Y^{\sim}(Y)|^2]) \end{aligned}$$

The correlation coefficient $R(X = 1, F_Y^{\sim mid}(Y)) =$

$$(odds(\tau_1))^{.5} E[S(F_Y^{\sim mid}(Y))|X = 1],$$

which is a version of the Wilcoxon statistic which is asymptotically Normal(0, 1/n).

It can be represented as a score statistic with score function $J(s) = (12)^{.5}(s - .5)$.

12. Two sample omnibus nonparametric tests, maximum chi-square, accumulation chi-square, $R(X = Q_X(t), Y \leq Q_Y(u))$.

The conventional chi-square statistic is based on indicator correlations $R(X = x, Y = y)$; it is only appropriate when X is discrete and Y is discrete. For diagnosis of dependence in this case, and also in the case X discrete, Y continuous, a more powerful method is accumulation correlation coefficients

$$R(X = x, Y \leq y) = (odds p_X(x) odds F_Y(y))^{.5} ((F_{Y|X=x}(y)/F_Y(y)) - 1),$$

One can show that these are the basis of conventional goodness of fit tests for equality of two distributions when X and Y are expressed in terms of their percentiles; we define percentile accumulation correlations

$$R(X = Q_X(t), Y \leq Q_Y(u)) = (\text{odds } p_X(Q_X(t)))^{.5} (D(u; F_Y, F_{Y|X=Q_X(t)}) - u) / (u(1-u))^{.5};$$

they can be used to form a cumulative chi-squared statistic for each treatment t (more precisely treatment x with $t = F_X(x)$)

$$n \sum_{\text{exact } u} R^2(X = Q_X(t), Y \leq Q_Y(u)).$$

and a maximal chi-squared statistic

$$n \max_{\text{exact } u} R^2(X = Q_X(t), Y \leq Q_Y(u)).$$

13. Quantile Limit Theorems.

The order statistics or sample quantiles $X(j; n) = Q^\sim(j/n)$ have an asymptotic distribution when F is continuous with positive probability density f . We call this a Quantile Limit Theorem III, a theorem about the asymptotic distribution of Q^\sim which often is obtained from Quantile Limit Theorems I and II described below.

The sample quantile function and sample distribution function are always related by a fundamental representation (usually called Bahadur's representation), which I call Quantile Limit Theorem I: as $n \rightarrow \infty$

$$\sqrt{n}fQ(u)(Q^\sim(u) - Q(u)) + \sqrt{n}(F^\sim(Q(u)) - u) = R_n \rightarrow 0$$

This can be regarded as a special case of a general Quantile Limit Theorem given by Doss and Gill (1992) which we use in the two-sample problem to derive the

asymptotic distribution of pooled estimators from unpooled estimators (Parzen (1999)).

We call a Quantile Limit Theorem II, a theorem about the asymptotic distribution of F^\sim , described as a statement about $F^\sim(Q(u))$ which is an example of a sample comparison distribution function.

The sample quantile regarded as a function on the unit interval $0 < u < 1$ is a *dynamic statistic* since its probability distribution is a stochastic process:

$$B_n(u) = \sqrt{n}(F^\sim(Q(u)) - u), \quad 0 < u < 1 \rightarrow B(u), \quad 0 < u < 1.$$

We call this a Functional Quantile Limit Theorem II. When the observed sample consists of independent uncensored observations, $B(u)$, $0 < u < 1$, is a Brownian Bridge, a zero mean Gaussian process with covariance function

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

To test a null hypothesis that Q is the true quantile function we test if $B_n(u)$, $0 < u < 1$, behaves as a sample path of a Brownian Bridge which is also called “white noise”. Alternative hypotheses about Q can often be shown to imply that

$$\text{Data representation } B_n(u), \quad 0 < u < 1 = \text{signal} + \text{white noise}.$$

Typical (or generic) test statistics of the null hypothesis of white noise are: “linear detectors” which correlate the data function $B_n(u)$, $0 < u < 1$, with a score function representing the type of signal specified by the alternative hypothesis; “quadratic detectors” which are sums of squares of independent linear detectors;

“information theory detectors” which are entropy measures of comparison density functions. Learning the history and roles of this diversity of methods is the aim of statistical methods mining.

A modern proof of the asymptotic distribution of $X(j; n)$ uses the fact that it has the same distribution as $Q(U(j; n))$, where

$$U(j; n) = S(j; n + 1)/S(n + 1; n + 1), S(j; n + 1) = \frac{1}{n + 1}(Y_1 + \dots + Y_j);$$

Y_j are independent exponentially distributed random variables with mean 1. As $j/(n + 1)$ converges to u , $\sqrt{n + 1}(S(j; n + 1) - u)$, $0 < u < 1$, converges in distribution to $W(u)$, $0 < u < 1$, where $W(u)$ is a Wiener process (zero mean Gaussian process with covariance $E(W(u_1)W(u_2)) = \min(u_1, u_2)$). One can show $\sqrt{n + 1}(U(j; n) - u)$ has same asymptotic distribution as $\sqrt{n + 1}(S(j; n + 1) - uS(n + 1; n + 1))$ which converges to $B(u) = W(u) - uW(1)$, a Brownian Bridge. Finally, $\sqrt{n + 1}(X(j; n) - Q(u))$ has same asymptotic distribution as $Q'(u)B(u)$. We hope that this outline can motivate applied statisticians to want to learn the “large sample theory” required for a more rigorous proof.

14. Two sample comparison density function asymptotic distributions.

Estimates \hat{d} of traditional density functions d (such as kernel estimates of probability density functions, nearest neighbor probability density functions, and time series spectral density functions) have variances proportional to d or d^2 . For the pooled and unpooled comparison density functions that we have defined for two samples the asymptotic variances are different. We state without proof how the asymptotic variances of \hat{d} depends on the true d . We use the notation of section

9. We write the theorem in a form most convenient for interpretation of density estimation of

$$\begin{aligned}
 p(u; H, F) &= \tau_m d(u; H, F) \\
 &= \tau_m f(H^{-1}(u)) / (\tau_m f(H^{-1}(u)) + \tau_n g(H^{-1}(u))).
 \end{aligned}$$

One can interpret $p(u; H, F) = P[X = 1 | Y = Q(u)]$.

Theorem: For estimators $\hat{d}(u; H, F)$ and $\hat{d}(u; F, G)$ of comparison density estimators one can show that (writing \propto for “proportional to”)

$$\text{Var}[\hat{p}(u; H, F)] \propto p(u; H, F)(1 - p(u; H, F))$$

$$\text{Var}[(\tau_n/\tau_m)\hat{d}(u; F, G)] \propto (\tau_n/\tau_m)d(u; F, G) + ((\tau_n/\tau_m)d(u; F, G))^2$$

REFERENCES

- Aly, Emad-Eldin A. A., Csörgő, Miklós and Horváth, Lajos. (1987). $P - P$ plots, rank processes and Chernoff–Savage theorems. *New Perspectives in Theoretical and Applied Statistics* (ed. Madan L. Puri, Jose Perez Vilaplana and Wolfgang Wertz). pp. 135–156.
- Csörgő, M. and Révész, P. (1978). Strong Approximations of the Quantile Process, *Annals of Statistics*, **6**, 822–894.
- Csörgő, M. and Horvath, L. (1997). *Limit Theorems in Change-point Analysis*, Wiley: New York.
- Ćwik, Jan and Mielniczuk, Jan. (1993). Data-dependent bandwidth choice for a grade density kernel estimate. *Statist. Prob. Lett.* 16, 597–405.
- David, H. A. (1995). First (?) Occurrence of common terms in mathematical statistics. *Am. Statistician* 49, 121–133.
- Doss, Hani and Gill, Richard D. (1992). An elementary approach to weak convergence for quantile processes, with applications to censored survival data, *Journal of the American Statistical Association*, Vol. 87, No. 419, 869–877.
- Eisenberger, Isidore and Posner, Edward C. (1965). Systematic statistics used for data compression in space telemetry, *Journal of the American Statistical Association*, 60, 97–133.
- Eubank, R. L. (1988). Optimal grouping, spacing stratification, and piecewise

constant approximation, *SIAM Review*, 30:3, 404–420.

Eubank, R. L., LaRiccia, V. N. and Rosenstein, R. B. (1987). Test statistics derived as components of Pearson’s phi-squared distance measure. *J. Amer. Statist. Assoc.* 8, 816–25.

Hald, Anders. (1998). *A History of Mathematical Statistics from 1750 to 1930*. Wiley: New York.

Holmgren, E. C. (1995). The $P - P$ plot as a method of comparing treatment effects. *Journal of the American Statistical Association*, 90, 360–365.

Nikitin, Yakov. (1995). *Asymptotic efficiency of nonparametric tests*. Cambridge University Press, New York, NY. p. 174.

Ogden, Todd and Parzen, Emanuel. (1996). Changepoint Approach to Data Analytic Wavelet Thresholding, *Statistics and Computing*, 6, 93–99.

Ogden, Todd and Parzen, Emanuel. (1996). Data Dependent Wavelet Thresholding in Nonparametric Regression with Change Point Applications’, *Computational Statistics and Data Analysis*.

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

Parzen, Emanuel. (1979a). A density-quantile function perspective on robust estimation, *Robustness in Statistics*, ed. R. Launer and G. Wilkinson, Academic Press: New York. 237–258.

- Parzen, Emanuel. (1989). Multi-sample functional statistical data analysis, *Statistical Data Analysis and Inference Conference in Honor of C. R. Rao*, ed. Y. Dodge, Elsevier: Amsterdam. 71–84.
- Parzen, Emanuel. (1991). Unification of statistical methods for continuous and discrete data, *Proceedings Computer Science–Statistics INTERFACE '90*, ed. C. Page and R. LePage, Springer Verlag: New York, 235–242.
- Parzen, Emanuel. (1992). Comparison change analysis. *Nonparametric Statistics and Related Topics* (ed. A. K. Saleh), Elsevier: Amsterdam, 3–15.
- Parzen, Emanuel. (1993) Change *PP* plot and continuous sample quantile function, *Communications in Statistics*, 22, 3287–3304.
- Parzen, Emanuel. (1994). Comparison change analysis approach to changepoint estimation, *Applied Changepoint Analysis Symposium, Proceedings. Journal of Applied Statistical Science*, 1, 379–401.
- Parzen, Emanuel. (1994). Correlation Unification of Statistical Methods and Introductory Statistical Education, Technical Report, Texas A&M University, Department of Statistics.
- Parzen, Emanuel. (1998). Data Mining, Statistical Methods Mining, and History of Statistics, *Interface of Computing Science and Statistics*, ed. D. Scott. Interface Foundation of North America, Vol. 29, pages 365–374.
- Parzen, Emanuel. (1999). Statistical Methods Mining, Two Sample Data Analysis, Comparison Distributions, and Quantile Limit Theorems, *Asymptotic Methods*

in Probability and Statistics, ed. B. Szyszkowicz, Elsevier: Amsterdam.

References on Unpooled Comparison Distribution, for Censored Data:

Hsieh, F. (1995). The empirical process approach for semiparametric two-sample models with heterogeneous treatment effect. *J. R. Statist. Soc. B*, 57, 735–748.

Hsieh, F. (1996). A transformation model for two survival curves: An empirical process approach. *Biometrika*, 83, 3, 519–528.

Hsieh, F. and Turnbull, Bruce. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics* (1996) 24, 25–40.

Li, Gang, Tiward, Ram and Wells, Martin. (1996). Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers, *Journal of the American Statistical Association*, 91, 689–698.

Figure Captions

Figure 1. Sample Conditional Mean of Y (testosterone) given X (estodiol) for samples from a polluted lake A and an unpolluted lake B.

Figure 2. Time series plot of NB10 (National Bureau of Standards 10 gram standard weight).

Figure 3. Normal (0,1) quantile function and Sample quantile function of (NB10-mean 404.59)/standard deviation 6.43. Notice that quantile functions have different slopes at $u = .5$.

Figure 4. Normal (0,1) quantile function and Sample quantile functions of (NB10-mean 406.56)/standard deviation 4.75 of cleaned data set which omits outliers (here minimum and maximum of original data set). Notice that sample and normal quantile functions have equal slopes at $u = .5$.

Figure 5. Change PP plot $D(u) - u$ of original NB10 data set and Normal distribution with estimated parameters has cubic shape indicative of departure from fit due to difference in slopes.

Figure 6. Change PP plot $D(u) - u$ of cleaned NB10 data set and normal distribution with estimated parameters has shape indicative of Brownian Bridge sample path which accepts goodness of fit of two distributions.

Figure 7 and 8. Sample shape identification quantile functions of simulated samples

from an exponential distribution. Sample sizes $n = 40$ and $n = 200$.

Figure 9. Scatter plot of (X, Y) , data on Old Faithful geyser in Yellowstone National Park, X =duration of an eruption, Y =duration until next eruption

Figure 10. Alternative to scatter plot presentation of (X, Y) is the standardized quantile function of X and the change density defined as conditional mean of standardized Y , given $X = x$, as a function of percentiles t of values of X .

Figure 11. Conditional mean of standardized Y , given $X \leq x = Q_X(t)$, provides diagnostic measures of dependence of Y on X .