

**STATISTICAL METHODS MINING, TWO SAMPLE DATA
ANALYSIS, COMPARISON DISTRIBUTIONS, AND
QUANTILE LIMIT THEOREMS**

by Emanuel Parzen

Department of Statistics, Texas A&M University

*Dedicated to Miklos Csörgő to celebrate his 65th birthday in 1997
and 20 years of quantile research.*

0. Abstract.

In this paper I propose a map (coordinate system) of statistical methods whose aim is to provide a vision of specific methods without learning their details. I use this framework to motivate methods for modeling two samples, and testing their homogeneity, based on comparison distribution functions, quantile limit theorems, and comparison density functions. Theory is developed to be applicable to both uncensored and censored data. I propose that we need to help the application of methods of probability and statistics to the emerging field of “data mining”, which seeks to extract information from, and identify models for, data (possibly massive). I propose the name “statistical methods mining” to describe the process of applying (and describing the location of one’s research within) the virtual encyclopaedia of statistical knowledge. I propose that we publicize our research by using various frameworks, and maps, of the world of statistical knowledge. Thus in Ottawa the “Hungarian construction” should be called the “Canadian-Hungarian construction.”

1. 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 metrics to measure the distance between F and the sample distribution function F^\sim . To compare two probabilities p_1 and p_2 we recommend p_1/p_2 rather than $p_1 - p_2$. Applying this philosophy to comparing two continuous distribution functions $F(x)$ and $G(x)$ we define $D(u; F, G) = G(F^{-1}(u))$, $0 < u < 1$ 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.

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* of comparison density functions and when we can interpret them as indicating that the difference between two distributions F and G is a difference in location or a difference in scale.

Comparison distributions and density concepts can be used to compare two *discrete* distributions F and G with respective probability mass functions p_F and p_G . We define, assuming $p_F(x) > 0$ implies $p_G(x) > 0$,

$$d(u; F, G) = \frac{p_G(F^{-1}(u))}{p_F(F^{-1}(u))}, D(u; F, G) = \int_0^u d(s; F, G) ds.$$

One can show that $D(u, F, G) = G(F^{-1}(u))$ 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.

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 $u = F(x)$.

Change *PP* plot is $D(u; F, G) - u$.

2. Data Notation Coordinate 1 of Two Sample Data Analysis

To provide a map of classical and modern statistical methods we propose three coordinates called: data notation, whole statistician, data inference.

Two sample data analysis plays a central role in applied statistics to compare treatment and control, or to compare today with yesterday. Coordinate 1 Data Notation is usually expressed: Sample 1, X_1, \dots, X_m , true continuous distribution F ; Sample 2, Y_1, \dots, Y_n , true continuous distribution G . Inference problems are test hypothesis of homogeneity $H_0 : F = G$, and estimate comparison distribution $D(u; F, G) = G(F^{-1}(u))$, $0 < u < 1$.

We often prefer an alternative notation (which extends to comparing c samples):

Sample 1, Y_1, \dots, Y_{n_1} , true continuous distribution F_1 ;

Sample 2, $Y_{n_1+1}, \dots, Y_{n_1+n_2}$, true continuous distribution F_2 .

We find it useful to formally regard two sample data as paired observations (X, Y) where $X = 1$ or 2 denotes the population 1 or 2, and Y denotes the response. Intuitively we use the notation of conditional probability to write (for $j = 1, 2$) $F_j(y) = F_{Y|X=j}(y)$.

Let $n = n_1 + n_2$. Define $\tau_1 = n_1/n$, $\tau_2 = n_2/n$, the proportions of the pooled sample in each sample. The pooled sample $Y_1, \dots, Y_{n_1}, Y_{n_1+1}, \dots, Y_n$ can be regarded incorrectly but intuitively as a sample from the pooled distribution $F(y) = \tau_1 F_1(y) + \tau_2 F_2(y)$. We regard F intuitively as the unconditional distribution of Y . We also use $H(x) = F(x)$.

3. Whole Statistician Coordinate 2 of Two Sample Data Analysis

The whole statistician coordinate 2 is usually application, theory, or computation. Coordinate 2 is application if we are analyzing real data (for example, income of men and women). Our goals are to test the hypothesis of homogeneity of populations $H_0 : F_1 = F_2 = F$ and more generally to estimate comparison distributions such as $D(u; F_1, F_2) = F_2(F_1^{-1}(u))$, $0 < u < 1$. In this paper coordinate 2 is theory; we are concerned with forming estimators and studying their probability distribution.

4. Data Inference Coordinate 3 of Two Sample Data Analysis

Coordinate 3 Data Inference focuses on conventional and modern estimation and testing procedures. To test the hypothesis H_0 of equality of the distributions F_1 and F_2 , con-

ventional procedures are various two-sample t -statistics and a Wilcoxon non-parametric rank statistic. We argue that these do not extract information from the data as well as estimating a comparison distribution function, either $D(u; F_1, F_2) = F_2(F_1^{-1}(u))$, *unpooled* estimator or $D(u; F, F_1) = F_1(F^{-1}(u))$, *pooled* estimator . Modern research on comparing the distributions of two samples such as men's and women's incomes have tended to focus on estimating the unpooled estimator. We believe that it is not always definable when F_1 and F_2 do not have the same support. Therefore we recommend use of the pooled estimator and show how its asymptotic distribution theory is an application of the Quantile Limit Theorem.

5. Comparison Distribution Functions Asymptotic Distribution

To state the properties of comparison distribution functions in two sample data analysis, we use the traditional notation of two samples X_1, \dots, X_m and Y_1, \dots, Y_n assumed to be independent random samples of X and Y with continuous distribution functions F and G . Let F_m and G_n be their discrete sample distribution functions. The pooled sample $X_1, \dots, X_m, Y_1, \dots, Y_n$ has sample distribution $H_{m+n}(x) = \tau_m F_m(x) + \tau_n G_n(x)$, defining $\tau_m = m/(m+n)$, $\tau_n = n/(m+n)$. Denote their order statistics by $X(j; m)$ and $Y(k; n)$. One can represent the normalized rank of $X(j; m)$ in the pooled sample as $H_{m+n}(X(j; m))$.

Let $c = (m+n)/(m+n+1)$. A linear rank statistic can be defined

$$T^{\sim}(J) = \frac{1}{m} \sum_{j=1}^m J(cH_{m+n}(X(j; m))), \int_0^1 J(u)du = 0, \quad \int_0^1 J^2(u)du = 1.$$

$$T^{\sim}(J) = \int_{-\infty}^{\infty} J(cH_{m+n}(x))dF_m(x) = \int_0^1 J(cu)dF_m(H_{m+n}^{-1}(u)).$$

$T^{\sim}(J)$ is estimator of $T(J) = \int_0^1 J(u)dF(H^{-1}(u)) = \int_0^1 J(u)dD(u; H, F)$ which equals 0 under the null hypothesis that $H_0 : F = G = H$, $D(u; H, F) = u$. Statistical inference studies $T^{\sim}(J)$ for insight about $D^{\sim}(u; H, F) = F_m(H_{m+n}^{-1}(u))$ which is an estimator of $D(u; H, F) = F(H^{-1}(u))$. When studying the asymptotic distribution of D^{\sim} it appears

necessary to study the asymptotic distribution of an “inverse”

$$D^{\sim}(u; F, H) - u = H_{m+n}(F_m^{-1}(u)) - u = \tau_n(G_n(F_m^{-1}(u)) - u).$$

Comparing two distributions F and G seems natural to estimate $D(u; F, G) = G(F^{-1}(u))$ with natural estimator $D^{\sim}(u; F, G) = G_n(F_m^{-1}(u))$ which we call the “unpooled” estimator.

We claim that using “unpooled” comparison distributions is less general than estimating the “pooled” comparison distribution $D(u; H, F) = F(H^{-1}(u))$ for which we propose as an improved estimator $D^{\sim}(u; H, F) = D(u; H_{m+n}, F_m)$ which is improved since it is a continuous (rather than discrete) distribution.

We outline how the asymptotic distribution of $D^{\sim}(u; H, F) - D(u; H, F)$, $0 < u < 1$, can be derived (by invoking the general Quantile Limit Theorem of Doss and Gill (1992)) in terms of stochastic processes $B_G(u)$, $B_F(u)$ satisfying

$$m^{\cdot 5}(F_m(x) - F(x)) \rightarrow B_F(F(x)),$$

$$n^{\cdot 5}(G_n(x) - G(x)) \rightarrow B_G(G(x)).$$

$$m^{\cdot 5}(F_m^{-1}(u) - F^{-1}(u)) \rightarrow (-1/f(F^{-1}(u)))B_F(u)$$

In the IID case B_F and B_G are Brownian Bridges with covariance $E[B(u_1)B(u_2)] = \min(u_1, u_2) - u_1u_2$. Other formulas for B_F and B_G apply for censored data.

Pooled Comparison Distribution Limit Theorem: Under conditions specified in the references:

$$(m+n)^{\cdot 5}\tau_m(D(u; H_{m+n}, F_n) - D(u; H, F)), 0 < u < 1, \rightarrow$$

$$\tau_m d(u; H, F) B_G(\tau_n D(u; H, G))$$

$$- \tau_n d(u; H, G) B_F(\tau_m D(u; H, F)), 0 < u < 1.$$

We use symbolic notation $B_F(\tau u)$ to denote $\tau^{\cdot 5} B_F(u)$.

Conditions for this theorem in the IID uncensored case using the name “empirical rank process” are given by Ali, Csörgő, Horvath (1987).

We obtain this limit theorem for the pooled estimator by applying the Quantile Limit Theorem to the limit theorem for the unpooled estimator. 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 $p(u; H, F) = P[X = 1 | Y = Q(u)]$.

Idea of proof: Let $C_{un}(u)$, $0 < u < 1$, be the limit process of the unpooled estimator in the sense of convergence in distribution of stochastic processes:

$$(m+n)^{\cdot 5} (G_n(F_m^{-1}(u)) - G(F^{-1}(u))) \rightarrow C_{un}(u)$$

and therefore

$$(m+n)^{\cdot 5} (H_{m+n}(F_m^{-1}(u)) - H(F^{-1}(u))) \rightarrow \tau_n C_{un}(u).$$

Define

$$\begin{aligned} C_{D_1}(w) &= \tau_n C_{un}(w), \\ u &= D_1(w) = H(F^{-1}(w)), \\ w &= D(u) = D_1^{-1}(u) = F(H^{-1}(u)). \end{aligned}$$

Regarding $F_m(H_{m+n}^{-1}(w))$ as the quantile (inverse) of $H_{m+n}(F_m^{-1}(u))$, and $F(H^{-1}(w))$ as the quantile (inverse) of $H(F^{-1}(u))$, we obtain from the Quantile Limit Theorem that

$$(m+n)^{\cdot 5} (F_m(H_{m+n}^{-1}(u)) - F(H^{-1}(u))) \rightarrow C_D(u)$$

where

$$\begin{aligned} C_D(u) &= (-1/d_1(D_1^{-1}(u)) C_{D_1}(D_1^{-1}(u)) \\ &= (-f H^{-1}(u)/h H^{-1}(u)) C_{D_1}(F H^{-1}(u)) \\ &= -d(u; H, F) C_{D_1}(D(u; H, F)) \end{aligned}$$

The limit process $C_{un}(u)$ of the unpooled estimator can be represented

$$C_{un}(u) = \tau_n^{-\cdot 5} B_G(D(u; F, G)) - \tau_m^{-\cdot 5} d(u; F, G) B_F(u).$$

The above formula for $C_{un}(u)$ can be deduced from asymptotics which we write as equations:

$$\begin{aligned} G_n(F_m^{-1}(u)) - G(F_m^{-1}(u)) &= n^{-.5} B_G (G(F_m^{-1}(u))) = n^{-.5} B_G (D(u; F, G)), \\ G(F_m^{-1}(u)) - G(F^{-1}(u)) &= g(F^{-1}(u)) (F_m^{-1}(u) - F^{-1}(u)) \\ &= (g(F^{-1}(u))/f(F^{-1}(u))) f(F^{-1}(u)) (F_m^{-1}(u) - F^{-1}(u)) = d(u; F, G) m^{-.5} (-B_F(u)) \end{aligned}$$

Unpooled Comparison Distribution Limit Theorem: Proved for censored data in recent references listed. For interpretation of density estimation of

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

we write the asymptotic distribution of the unpooled estimator as follows under conditions specified in the references:

$$\begin{aligned} ((m+n)\tau_m)^{.5} (\tau_n/\tau_m) (G_n(F_m^{-1}(u)) - G(F^{-1}(u))) \\ \rightarrow B_G((\tau_n/\tau_m)D(u; F, G)) - (\tau_n/\tau_m)d(u; F, G)B_F(u). \end{aligned}$$

6. Comparison density functions asymptotic distribution.

Applying these results to 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”)

$$\begin{aligned} \text{Var} [\hat{p}(u; H, F)] &\propto p(u; H, F)(1 - p(u; H, F)) \\ \text{Var} [(\tau_n/\tau_M)\hat{d}(u; F, G)] &\propto (\tau_n/\tau_m)d(u; F, G) + ((\tau_n/\tau_m)d(u; F, G))^2 \end{aligned}$$

Note that for other kinds of density functions d the variance of an estimator \hat{d} is typically proportional to d or d^2 . A theorem on estimation of a comparison density is given by wik and Mielniczuk (1993) and Handcock and Janssen (1995).

References by Csrg and Parzen cited below celebrate 20 years of research on the theory and applications of quantiles.

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.
- Ćwik, Jan and Mielniczuk, Jan. (1993). Data-dependent bandwidth choice for a grade density kernel estimate. *Statist. Prob. Lett.* 16, 597–405.
- 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.
- 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. Soc.* 8, 816–25.
- Handcock, Mark S. and Janssen, Paul L. (1995). Statistical inference for the relative distribution and relative density.
- Nikitin, Yakov. (1995). *Asymptotic efficiency of nonparametric tests*. New York, NY: Cambridge University Press. p. 174.
- Parzen, Emanuel. (1979). Nonparametric Statistical Data Modeling *Journal of the American Statistical Association*, (with discussion). **74**, 105-131.
- 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, *Com-*

munications in Statistics, 22, 3287–3304.

Recent References on Unpooled Comparison Distribution, Especially for Censored Data:

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

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.