

QUANTILES, DATA MODELING, ALL STATISTICAL METHODS LEARNING

Emanuel Parzen, Texas A&M University

- 0) Introduction to almost baked ideas
 - *1) Population Quantile, Mid-Distribution Functions
 - *2) Quartile / Quantile Function
 - 3) Sample quantile continuous version
 - 4) Mid-quartile, quartile deviation
 - *5) SIEVE strategy of data analysis
 - *6) Stage S, Sample Summary Numbers
 - *7) $2Q$ plot, $2Q/Q$ plot
 - 8) Validation, Goodness of Fit, Comparison Distribution
 - *9) Validation dilemma, Components
 - *10) Design of score functions
 - *11) Stage E, Estimation non-parametric by Exponential Models
 - 12) Bivariate data, Scattergram / Conditional Quantiles
 - 13) Component correlations, multi-samples
 - 14) Two sample, multi-samples
 - 15) Multi Sample Multivariate Data
 - *16) Unified Theory of Component Correlations
 - *17) Cum–Cum Plots
 - 18) Appendix: Score functions, orthogonal polynomials, mid-distribution
- * To be presented at Florida State University Nonparametric Statistics Research Conference

0 Introduction to almost baked ideas

This paper outlines a unified framework for statistical methods and practice which completes the vision of nonparametric statistical modeling proposed in Parzen (1979). Ideas and notation outlined in this paper might be considered “almost baked” because they require an extensive book to provide detailed algorithms, examples, proofs, and exploration of the practical implications and applications of the formulas that we introduce.

The fundamental role of quantile data analysis in statistical science can be learned by asking what is the “Fundamental Theorem” of probability theory that makes statistical inference possible. It is often stated to be Law of Large Numbers for Sample Distribution Function $F^{\sim}(y)$; but we believe it is Central Limit Theorem and Extreme Value Theory for Sample Quantile Function $Q^{\sim}(u) = F^{\sim-1}(u)$ which says that one does not know the probability distribution of Y_1, \dots, Y_n but one does know asymptotically the probability distribution of their order statistics $Y(1, n) \leq \dots \leq Y(n, n)$.

We propose “all statistical methods learning,” a philosophy of thinking about statistical methods that

- (1) integrates statistical education, practice, scholarship, and research;
- (2) provide statisticians a framework to know about almost all (at least half) of statistical methods, a framework that makes esoteric expert knowledge exoteric common knowledge;
- (3) reduces the enormous gap that exists between “academically” known statistical theory and applied statistical methods, design introductory statistics courses that are explicit about teaching skills and understanding;
- (4) help students enjoy learning statistics because they are given clear advice to select appropriate methods and models by first classifying the type of data being analyzed: one sample univariate Y , bivariate (Y_1, Y_2) , multi-samples (Y, X) ;
- (5) practice the scientific method by learning ways of thinking that extend to the most advanced problems and that emphasize analogies between problems which link their solutions;
- (6) teach probability, conditional probability, and mid-distributions as the foundation for the collection and modeling of data (advocated by Lindley (2002)).

Our ideas began in 1976 in a study of the Total Time on Test Statistics which we formulated as testing for continuous distributions F_0 and F_1 the hypothesis $F_1 = F_0$ by testing $D(u; F_0, F_1) = F_1(F_0^{-1}(u)) = u$.

To compare continuous quantile functions $Q_0(s), Q_1(u)$ we propose solving for each u the value s satisfying

$$Q_1(u) = Q_0(s).$$

Applying F_1 to both sides we obtain

$$\begin{aligned} u &= D(s; F_0, F_1) \\ s &= D^{-1}(u; F_0, F_1) \\ Q_1(u) &= Q_0(D^{-1}(u; F_0, F_1)) \end{aligned}$$

The graph of $D(s; F_0, F_1)$ is estimated in practice by a *PP* plot, a concept which we extend to a Cum-Cum plot. Estimating the comparison density

$$d(s; F_0, F_1) = D'(s; F_0, F_1)$$

provide ultimate answers.

Bivariate data (Y, X) is analyzed by component correlations and by conditional comparison distributions

$$\begin{aligned} D(u; F_Y, F_{Y|X=Q_X}(t)), & \quad 0 \leq u, t \leq 1, \\ D(u; F_Y, F_{Y|X \leq Q_X}(t)). & \end{aligned}$$

A fundamental formula to estimate conditional quantiles is

$$Q_{Y|X=x}(u) = Q_Y(D^{-1}(u; F_Y, F_{Y|X=x}))$$

To implement this formula we estimate sequentially $d(s; F_Y, F_{Y|X=x})$ for each s as a function of x , and $D^{-1}(u; F_Y, F_{Y|X=x})$ as a function of u for fixed x .

1 Population Quantile, Mid-distribution Functions

Data analysis and modeling starts with Y_1, \dots, Y_n , a sample of random variable Y .

Probability model for a random variable Y described by several functions:

Distribution function $F(y) = F_Y(y) = P[Y \leq y]$, $-\infty < y < \infty$,

Probability mass function $p(y) = p_Y(y) = P[Y = y]$, $-\infty < y < \infty$,

Mid-distribution function

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

Quantile function $Q(u) = Q_Y(u) = F^{-1}(u)$.

Continuous F definition: solve $u = F(y)$ for $y = Q(u)$.

General definition: $Q(u) = \inf \{y : F(y) \geq u\}$.

When F is continuous $F(Q(u)) = u$, probability density $f(x) = F'(x)$ and *quantile density* $q(u) = Q'(u)$ satisfy

$$f(Q(u))q(u) = 1.$$

We call $f(Q(u))$ the *density quantile* function denoted $fQ(u)$. In words,

$$(\text{density quantile}) \times (\text{quantile density}) = 1.$$

A classification of tail behavior of probability distributions is obtained from a representation of $fQ(u)$, $u = 0$ and $u = 1$. There exist numbers α_0 and α_1 (called tail exponents) and slowly vary functions $L_0(u)$ and $L_1(1 - u)$: for u near 0

$$fQ(u) = u^{\alpha_0} L_0(u);$$

for u near 1

$$fQ(u) = (1 - u)^{\alpha_1} L_1(1 - u).$$

These representations provide elegant presentations of the theory of Extreme Value Distributions.

Tails of probability distributions are classified as short, medium, long:

Left Tail	α_0	Right Tail	α_1
Short	$\alpha_0 < 1$	Short	$\alpha_1 < 1$
Medium	$\alpha_0 = 1$	Medium	$\alpha_1 = 1$
Long	$\alpha_0 > 1$	Long	$\alpha_1 > 1$

2 Quartile / Quantile Function

Practical criteria for identifying measures of skewness and tail behavior can be defined in terms of a dimension-less function called the *quartile / quantile* function $Q/Q(u)$. Define

median	$Q(.5) = Q_2$
lower quartile	$Q(.25) = Q_1$
upper quartile	$Q(.75) = Q_3$
mid-quartile	$MQ = .5(Q_1 + Q_3)$
quartile deviation	$DQ = 2(Q_3 - Q_1)$

Define *quartile / quantile* function

$$Q/Q(u) = (Q(u) - MQ)/DQ$$

Verify $Q/Q(.25) = -.25$, $Q/Q(.75) = .25$. Values of u such that $|Q/Q(u)| > 1$ are called "Tukey outliers".

One can initially identify tail behavior from the value near 0 and 1 of $Q/Q(u)$:

Left Tail		
Long	Medium	Short
$Q/Q(u) < -1$	$-1 < Q/Q(u) < -.5$	$-.5 < Q/Q(u) < -.25$
Right Tail		
Short	Medium	Long
$.25 < Q(u) < .5$	$.5 < Q/Q(u) < 1$	$1 < Q/Q(u)$

3 Sample quantile continuous version

Sample versions of distribution and quantile functions are defined in terms of sample (empirical) probability: for set B of real numbers

$$P^\sim[B] = P^\sim[Y \text{ in } B] = \text{fraction of sample } Y_1, Y_2, \dots, Y_n \text{ in } B.$$

Sample distribution function $F^\sim(y) = P^\sim[Y \leq y]$.

Sample quantile $Q^\sim(u) = F^{\sim-1}(u)$

Order statistics $Y(1; n) \leq \dots \leq Y(n; n)$

$$Q^\sim(u) = Y(k; n), (k-1)/n < u \leq k/n$$

Distinct values in sample denoted $y_1 < \dots < y_r$

$$Q^\sim(u) = y_k, F^\sim(y_k) - p^\sim(y_k) < u \leq F^\sim(y_k)$$

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

Mid-distribution transform $F^{\sim\text{mid}}(Y)$ has mean and variance

$$E^\sim[F^{\sim\text{mid}}(Y)] = .5, \text{VAR}^\sim[F^{\sim\text{mid}}(Y)] = (1/12) \left(1 - \sum_{k=1}^r p^{\sim 3}(y_k) \right) = \sigma_{\text{mid}, Y}^2$$

which provide usual correction for ties in non-parametric statistics.

Sample quantile continuous version $Q^{\sim c}(u)$ is defined by join linearly

$$(F^{\sim\text{mid}}(y_k), y_k), k = 1, \dots, r$$

4 Mid-quartile, quartile deviation

To compute sample quartiles

$$Q_k = Q^{\sim c}(k/4), k = 1, 2, 3$$

determine $y_{j-1} < y_j$ so that

$$F^{\sim\text{mid}}(y_{j-1}) < k/4 \leq F^{\sim\text{mid}}(y_j).$$

Then compute Q_k by linear interpolation between y_{j-1} and y_j . We recommend computing sample summary:

median Q_2

quartiles Q_1, Q_3

mid-quartile $MQ = .5(Q_1 + Q_3)$

quartile deviation $DQ = 2(Q_3 - Q_1) = 2IQR$

interquartile range $IQR = Q_3 - Q_1$

histogram / quantile bin width $BW = IQR/2 = DQ/4$

histogram / quantile plot, with interval edge at MQ , is the universal solution to the problem of how to draw a histogram!

Example of sample diagnostics:

$$\begin{aligned}
Q_2 &= MQ, \text{ symmetric,} \\
Q_2 &< MQ, \text{ right skew (mean} > \text{median),} \\
Q_2 &> MQ, \text{ left skew.}
\end{aligned}$$

5 SIEVE Strategy of Data Analysis

Interactive process SIEVE is strategy for statistical data analysis and modeling:

- S: Sample distribution diagnosed by summary numbers of F^\sim , plots of Q^\sim^c , plots of Q/Q
- I: Identification of parametric probability models, parameters θ of F_θ ,
- E: Estimation of parameters θ^\wedge and model F^\wedge for true F
- V: Validation, goodness of fit of F^\wedge to F^\sim
- E: Estimation non-parametric of F^\wedge^E , exponential model.

There are important practical distinctions between E: estimation parametric and E: estimation non-parametric. Generalized linear models are parametric estimation, log linear models are non-parametric estimation.

6 Stage S, Sample Summary Numbers

First step in data analysis of univariate data is summary numbers of sample distribution F^\sim :

$$\begin{aligned}
\text{sample mean } \mu^\sim &= (1/n) \sum_{k=1}^n Y_k = \sum_{k=1}^r y_k p^\sim(y_k) \\
\text{sample variance } \sigma^{\sim 2} &= (1/n) \sum_{k=1}^n (Y_k - \mu^\sim)^2 = \sum_{k=1}^r (y_k - \mu^\sim)^2 p^\sim(y_k)
\end{aligned}$$

sample standard deviation σ^\sim

sample minimum MIN = $Q^\sim(0)$

sample maximum MAX = $Q^\sim(1)$

quartile lower $Q_1 = Q^\sim^c(.25)$

median $Q_2 = Q^\sim^c(.5)$

quartile upper $Q_3 = Q^\sim^c(.75)$

Mid-quartile MQ

Quartile deviation DQ

sample quartile / quantile function $Q/Q(u) = (Q^\sim^c(u) - MQ)/DQ$

skew index $Q/Q(.5) = (Q_2 - MQ)/DQ$

Left tail index $Q/QMIN = (MIN - MQ)/DQ$

$$Q/Q(.05) = (Q^\sim^c(.05) - MQ)/DQ$$

Right tail index $Q/QMAX = (MAX - MQ)/DQ$

$$Q/Q(.95) = (Q^\sim^c(.95) - MQ)/DQ$$

Note $Q/Q(u)$ is a sample quantile continuous version of “dimensionless” data

$$Y_k^Q = (Y_k - MQ)/DQ$$

Another form of dimensionless data is “normalized residuals”

$$Y_k^\nu = (Y_k - \mu^\sim) / \sigma^\sim$$

which have sample mean 0, sample variance 1.

A *Sample Mid-distribution Quantile Data Analysis*: McPherson (1989) reports responses from a sample of eight individuals $Y(j)$;

240, 194, 215, 194, 450, 240, 215, 215.

Our first step (in initial phase S) is to determine the distinct values y_j in the sample, and the sample mid-distribution $F^{\sim\text{mid}}(y_j)$. Form first order statistics $Y(j, n)$:

194, 194, 215, 215, 215, 240, 240, 450.

The continuous version of the sample quantile $Q^{\sim c}(u)$ is calculated from the table

y_j	Count	$F^{\sim\text{mid}}(y_j)$
194	2	.125
215	3	.4375
240	2	.75
450	1	.9375

Q_1 , the value at .25, is formed by interpolating between .125 and .4375; Q_2 , the value at .5, is formed by interpolating between .4375 and .75; Q_3 , the value at .75, is 240. Five number summary is

Min = 194	Min $Q/Q = -.367$, left tail short,
$Q_1 = 207.7$	$MQ = 221.35$
$Q_2 = 220$	$DQ = 74.6$
$Q_3 = 240$	Max $Q/Q = 3.065$, right tail outlier,
Max = 450	$Q/Q(.5) = -.018$, skewness almost zero,

The sample distribution is diagnosed as symmetric, left tail short, and right tail so far out as indicating bimodality (sampling from a distribution different from center of distribution).

McPherson reports the data are salaries of clerical employees in a company, except for 450 which is the salary of a manager.

7 2Q Plot, 2Q/Q plot

The SIEVE strategy of fitting probability models to data proposes as Stage S non-model analysis of the sample distribution function F^\sim . The results of Stage S are presented graphically by a plot of $Q^{\sim c}(u)$ and a plot of $Q/Q(u)$ from which we can learn tail behavior, skewness, normality, bimodality.

We call a $2Q$ plot the graphs of $Q^{\sim c}(u)$ and the quantile of Normal $(\mu^{\sim}, \sigma^{\sim 2})$ distribution $\mu^{\sim} + \sigma^{\sim} \Phi^{-1}(u)$ of a normal distribution after parameter estimation. We call a $2Q/Q$ plot graphs of $Q/Q(u)$ and the graph of the normal quartile / quantile function

$$\Phi^{-1}/\Phi^{-1}(u) = \Phi^{-1}(u)/2.7$$

since $\Phi(.675) = .75$, $\Phi(-.675) = .25$.

We plot $Q^{\sim c}(u)$ by joining linearly (where y_j are distinct values in the sample)

$$(F^{\sim \text{mid}}(y_j), y_j), j = 1, \dots, r;$$

we plot $Q/Q(u)$ by joining linearly

$$(F^{\sim \text{mid}}(y_j), (y_j - MQ)/DQ), j = 1, \dots, r.$$

8 Validation, Goodness of Fit, Comparison Distribution

Identification of parametric models $F_\theta(y)$ can be guided by the Q/Q plot. Estimation stage forms parameter estimators θ^\wedge and a model $F^\wedge(y) = F_{\theta^\wedge}(y)$.

Validation tests a null hypothesis $H_0 : \text{true } F = F_0$, for a specified F_0 , by testing (when F_0 is continuous)

$$H_0 : F_0(Y_1), \dots, F_0(Y_n) \text{ are Uniform } (0, 1) \text{ sample.}$$

In both the discrete and continuous cases we can graphically test $H_0 : F = F_0$ by a PP plot defined to join linearly the points listed; when F_0 is continuous

$$(0, 0), (F_0(y_j), F^{\sim \text{mid}}(y_j)), (1, 1);$$

when F_0 is discrete

$$(0, 0), (F_0^{\text{mid}}(y_j), F^{\sim \text{mid}}(y_j)), (1, 1)$$

or usually

$$(0, 0), (F_0(y_j), F^\sim(y_j)).$$

A PP plot can be regarded as a sample comparison distribution estimating the comparison distribution

$$D(u; F_0, F^\sim) = D(u; \text{model}, \text{data}).$$

Parameter estimation θ^\wedge can be defined as value of θ minimizing “distance” between uniform distribution $D_0(u) = u$ and

$$D(u; F^\sim, F_0) = D(u; \text{data}, \text{model}).$$

Important diagnostic plots that compare $F^\wedge(y)$ to $F^\sim(y)$ are the change *PP* plot

$$(F^\wedge(y_j), n \cdot^5 (F^{\sim\text{mid}}(y_j) - F^\wedge(y_j)))$$

and the Test *PP* plot

$$(F^\wedge(y_j), n \cdot^5 (F^{\sim\text{mid}}(y_j) - F^\wedge(y_j)) / (F^\wedge(y_j)(1 - F^\wedge(y_j)) \cdot^5)).$$

9 Validation dilemma, components

Statisticians do not yet have consensus about how to test goodness of fit. I call this the “statistical quandary or dilemma”: we have too many good answers to your problem and we do not know which one to recommend. We believe that “components” can be recommended as providing goodness of fit tests that work well for both discrete and continuous data.

The Chi-square test for discrete data is usually expressed

$$\text{Chi} = \sum \frac{(\text{observed counts} - \text{expected counts})^2}{\text{expected counts}}$$

We formulate this test as comparing empirical probabilities $p^\sim(y_j), j = 1, \dots, r$ and model probabilities $p^\wedge(y_j), j = 1, \dots, r$, by $\text{Chi} = n C$, defining

$$\begin{aligned} C &= \sum_{j=1}^r p^\wedge(y_j) \left(\frac{p^\sim(y_j)}{p^\wedge(y_j)} - 1 \right)^2 \\ &= \int_0^1 du \left(\frac{p^\sim(F^{\wedge-1}(u))}{p^\wedge(F^{\wedge-1}(u))} - 1 \right)^2 \end{aligned}$$

By theory of Hilbert space of functions on $[0, 1]$, we can choose (non-uniquely) orthogonal functions $\varphi_1(u), \dots, \varphi_{r-1}(u)$, compute coefficients (called components)

$$C_k = \int_0^1 du \frac{p^\sim(F^{\wedge-1}(u))}{p^\wedge(F^{\wedge-1}(u))} \varphi_k(u).$$

Approximately

$$C_k = \sum_{j=1}^r p^\wedge(y_j) \frac{p^\sim(y_j)}{p^\wedge(y_j)} \varphi_k(F^{\wedge\text{mid}}(y_j)),$$

and calculate the chi-square statistic C by

$$C = \sum_{k=1}^{r-1} C_k^2$$

We seek to design orthogonal functions of y similar to

$$\psi_k(y) = \varphi_k(F^{\wedge\text{mid}}(y))$$

and components

$$\begin{aligned} C_k &= \sum_j p^\sim(y_j) \psi_k(y_j) \\ &= E^\sim[\psi_k(Y)]. \end{aligned}$$

We call C_k “components”; they are sample means of score function $\psi_k(y)$ which satisfy

$$\begin{aligned} E^\wedge[\psi_k(Y)] &= \sum_y p^\wedge(y) \psi_k(y) = 0 \\ \text{VAR}^\wedge[\psi_k(Y)] &= \sum_y p^\wedge(y) \psi_k^2(y) = 1 \\ E^\wedge[\psi_{k_1}(Y) \psi_{k_2}(Y)] &= 0 \text{ if } k_1 \neq k_2. \end{aligned}$$

In a typical application,

$$\begin{aligned} nC^2 &= n(C_1^2 + C_2^2 + C_3^2 + C_4^2) = 6.6 \\ nC_1^2 &= 6.6 \text{ approximately.} \end{aligned}$$

The statistic nC^2 is tested for significance by regarding it as the value of a random variable which is Chi-square (4 degrees of freedom). The statistic nC_1^2 is tested for significance by regarding it as the value of a random variable which is Chi-square (1 degree of freedom). The component C_1 with observed value 6.6 has P -value .01, while Chi has P -value .15.

Using C_1 we interpret $p^\sim(y) - p^\wedge(y)$ as significantly different from zero in the direction of the score function $\psi_1(y)$.

10 Design of score functions

For goodness of fit of $F^\wedge(y)$ to $F^\sim(y)$ when true F is continuous, the Anderson Darling statistic is often recommended. Not so well known is the recommendation to use instead the components

$$C_k = E^\sim[\varphi_k(F^{\wedge\text{mid}}(Y))] = (1/n) \sum_{j=1}^n \varphi_k(F^{\wedge\text{mid}}(Y_j))$$

where $\varphi_k(u)$, $k = 1, 2, \dots$, are Legendre polynomials. Unification of methods for continuous and discrete data can be accomplished by constructing score functions for a discrete distribution

$$p(y), y = y_1, \dots, y_r.$$

Nonparametric statistical methods can be unified and extended by suitable score functions.

We recommend using two types of orthonormal score functions (details in appendix):

- (1) $g_k(y)$, based on original values;
- (2) $\psi_k(y)$, based on mid-distribution transformed values $F^{\text{mid}}(y_j)$.

For statistical data analysis of bivariate data (Y_1, Y_2) , two sample data, and multi-sample data, we will find it convenient to use as score functions $g_{k_1}(y_1)g_{k_2}(y_2)$ and $\psi_{k_1}(y_1)\psi_{k_2}(y_2)$, called “tensor products.”

11 Stage E, Estimation Non-parametric by Exponential Models

When we find significant components we use their score functions to form an exponential model estimator of the true $p(y)$ of the form

$$p^{\wedge E}(y) = p^{\wedge}(y) \exp(\theta_0 + \theta_1 \psi_1(y)).$$

This estimator has a maximum entropy interpretation; optimum θ^{\wedge} are those satisfying constraints

$$\sum_y (p^{\wedge E}(y) - p^{\sim}(y)) \psi_1(y) = 0.$$

Iterative Newton-Raphson estimation of θ_1 starts with an initial estimator

$$\theta_1^{\wedge(1)} = \sum_y p^{\sim}(y) \psi_1(y).$$

12 Bivariate data, Scattergram / conditional quantile

Data: $(Y_1(j), Y_2(j)), j = 1, \dots, n$, sample of bivariate random variables (Y_1, Y_2) .

A bivariate data analysis of (Y_1, Y_2) should be preceded by univariate data analyses of Y_1 and Y_2 . Questions investigated in bivariate data analysis are (1) bivariate normality, and (2) bivariate bimodality (two modes in the joint probability density $f_{Y_1, Y_2}(Y_1, Y_2)$). We recommend that one should check first for univariate bimodality because one usually seeks to identify bivariate bimodality that is not induced by univariate bimodality.

Bivariate independence: in terms of probability mass functions, Y_1 and Y_2 are independent if

$$\begin{aligned} p_{Y_1, Y_2}(Y_1, Y_2) &= p_{Y_1}(Y_1) p_{Y_2}(Y_2) \\ p_{Y_2|Y_1=y_1}(Y_2) &= p_{Y_2}(Y_2) \end{aligned}$$

For exploratory bivariate data analysis we recommend routine calculation of

- (1) scattergram / conditional quantile plot,
- (2) component correlations table.

Scattergram / conditional quantile plot graphs points $(Y_1(j), Y_2(j)), j = 1, \dots, n$ on (Y_1, Y_2) plane on which are drawn *vertical* lines at marginal quantiles of Y_1

$$Y_1 = Q_{1, Y_1}, Y_1 = Q_{2, Y_1}, Y_1 = Q_{3, Y_1}$$

and *horizontal* line at marginal quantiles of Y_2

$$Y_2 = Q_{1,Y_2}, Y_2 = Q_{2,Y_2}, Y_2 = Q_{3,Y_2}$$

These lines divide plane into 16 bins with marginal probabilities .25. The vertical lines divide plane into 4 quadrants. We plot in each quadrant conditional quantiles of set of Y_2 values corresponding to Y_1 values in the quadrant, which we denote

$$Q_{Y_2|Y_1 \text{ in quadrant}}(u), u = .25, .5, .75$$

For each $u = .25, .5, .75$ we call the linear join of the conditional quantile values the “conditional quantile of Y_2 as a function of Y_1 .” We can identify bivariate normality when the conditional quantiles are linear functions and quartile deviation of Y_2 as a function of Y_1 is constant.

A historically important example is the original bivariate data studied by Galton in his study of the heights of children and parents.

13 Component correlations, multi-samples

We change our notation for bivariate data to (Y, X) when X is an explanatory variable that can be a deterministic or random variable. Probabilities involving X are not population probability but are sample probability. We call data “multi-sample“ if it represents a response variable Y observed in c samples indexed by X . The data is recorded as bivariate data $(Y(j), X(j))$, where X has values $1, \dots, c$ representing the population being observed.

For multi-samples, conditional quantiles plot for $k = 1, \dots, c$ is a scattergram of the points

$$(F_X^{\sim \text{mid}}(k), Q_{Y|X=k}(u)), u = .25, .5, .75$$

On same graph plot horizontal lines at quartiles of unconditional distribution (pooled sample) of Y_1

$$y = Q_Y(u), u = .25, .5, .75$$

For bivariate data initial understanding (knowledge discovery) is provided by computing *conditional quantile* for Y given an interval $\{x : x_{k-1} < X \leq x_k\}$ for specified $x_0 < x_1 < \dots < x_c$. We are quantizing to transform the continuous variable X to a discrete variable.

For $j, k = 1, 2, \dots$ define *component correlation* of degrees (j, k) and score family g or ψ

$$\begin{aligned} C_{j,k}^g(Y, X) &= E^{\sim}[g_{j,Y}(Y)g_{k,X}(X)] \\ C_{j,k}^\psi(Y, X) &= E^{\sim}[\psi_{j,Y}(Y)\psi_{k,X}(X)] \end{aligned}$$

Pearson correlation is $C_{1,1}^g(X, Y) = \rho(X, Y)$. Spearman rank correlation is

$$C_{1,1}^\psi(Y, X) = \rho(F_Y^{\sim \text{mid}}(Y), F_X^{\sim \text{mid}}(X))$$

Simple linear regression, or linear prediction of degree 1, predicts Y given X by Y^μ , a function of X , given by

$$\frac{Y^\mu - \mu_Y}{\sigma_Y} = C_{1,1}^g(Y, X)g_{1,X}(X).$$

Linear prediction of degree 2 predicts Y given X by Y^μ , a function of X , given by

$$\frac{Y^\mu - \mu_Y}{\sigma_Y} = C_{1,1}^g(Y, X)g_{1,X}(X) + C_{1,2}^g(Y, X)g_{2,X}(X).$$

Open applied research problem: routinely compute and interpret $C_{1,2}^g(Y, X)$.

Formulas to compute and interpret component correlations (omitting \sim on probabilities for ease of writing)

$$\begin{aligned} C_{j,k}^g(Y, X) &= \sum_x p_X(x)\psi_{k,X}(x) \sum_y p_{Y|X=x}(y)\psi_{j,Y}(y) \\ &= E_X[\psi_{k,X}(X)E[\psi_{k,Y}(Y)|X=x]] \\ &= \int_0^1 du \psi_{k,X}(Q_X(u))E[\psi_{k,Y}(y)|X=Q_X(u)]. \end{aligned}$$

To graphically explore bivariate sample $(Y(j), X(j)), j = 1, \dots, n$ we plot:

(1) Scattergram Q/Q . Plot points

$$\left(Y^Q(j) = \frac{Y(j) - MQ_Y}{DQ_Y}, X^Q(j) = \frac{X(j) - MQ_X}{DQ_X} \right)$$

and plot horizontal and vertical lines at $-.25, 0, .25$ and circle at $(Q/Q_Y(.5), Q/Q_X(.5))$.

(2) Conditional mean components $E[\psi_{1,Y}(Y)|X=Q_X(u)]$ and $\psi_{1,X}(Q_X(u))$ on same graph, $0 \leq u \leq 1$.

(3) Conditional mean components $E[\psi_{1,Y}(Y)|X=Q_X(u)]$ and $\psi_{2,X}(Q_X(u))$ on same graph, $0 \leq u \leq 1$.

14 Two sample, multi-samples

We represent two samples as (Y, X) where $X = 1$ or 2 . We represent c multi-samples as (Y, X) , where $X = 1, 2, \dots, c$. We use the notation of conditional probability $p_{Y|X=k}(y)$ to represent the sample distribution of Y in the k -th sample. Hypotheses of homogeneity of distributions is expressed, for all k

$$F_{Y|X=k}(y) = F_Y(y) := \sum_{j=1}^r P[X=j]F_{Y|X=j}(y)$$

We call $F_Y(y)$ the unconditional or pooled distribution.

Conventional Wilcoxon statistic corrected for ties can be represented

$$(n-1)^5 R(X=1, F_Y^{\sim \text{mid}}(Y))$$

where

$$\begin{aligned}
R(X = 1, F_Y^{\sim\text{mid}}(Y)) &= \left(\frac{P[X = 1]}{1 - P[X = 1]} \right)^{.5} E^{\sim}[\psi_{1,Y}(Y)|X = 1] = C_{11}^{\psi}(Y, X). \\
E^{\sim}[\psi_{1,Y}(Y)|X = 1] &= \sum_y p_{Y|X=1}^{\sim}(y)(F_Y^{\sim\text{mid}}(y) - .5)/\sigma_{\text{mid},Y} \\
&= \sum_y p_Y^{\sim}(y) \frac{p_{Y|X=1}(y)}{p_Y^{\sim}(y)} (F_Y^{\sim\text{mid}}(y) - .5)/\sigma_{\text{mid},Y} \\
&= \int_0^1 du \, d(u; F_Y^{\sim}, F_{Y|X=1}^{\sim})(F_Y^{\sim\text{mid}}(Q_Y^{\sim}(u)) - .5)/\sigma_{\text{mid}}.
\end{aligned}$$

The sample comparison density $d(u; F_Y^{\sim}, F_{Y|X=1}^{\sim})$ is an estimator of true comparison density

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

Estimation of two sample comparison density is equivalent to logistic regression estimation of $P[X = 1|Y = y]$.

Conventional Kruskal-Wallis statistic corrected for ties can be represented

$$\begin{aligned}
(n - 1)R^2(F_Y^{\text{mid}}(Y)|X) &= \sum_x P^{\sim}[X = x](E^{\sim}[\psi_{1,Y}(Y)|X = x])^2 \\
&= \sum_x (1 - P[X = x])R^2(X = x, F_Y^{\text{mid}}(Y)).
\end{aligned}$$

The sum of squares defining the Kruskal-Wallis statistics can be represented as the square norm of a function and therefore as a sum of squares of components (Fourier coefficients)

$$\begin{aligned}
&\sum_x p_X^{\sim}(x)\psi_{k,X}(x)E^{\sim}[\psi_{1,Y}(Y)|X = x] \\
&= C_{1,k}^{\psi}(Y, X).
\end{aligned}$$

We conclude with an important finding: one can define statistics to test homogeneity of c samples of the form (for $m = 1, 2, \dots, c - 1$)

$$(n - 1) \sum_{k=1}^m |C_{1,k}^{\psi}(Y, X)|^2,$$

whose asymptotic distribution is chi-square with m degrees of freedom.

That component correlations are universally applicable tools of nonparametric statistical data modeling was first shown by Rayner and Best (2001); they propose the unifying

roles of component correlations in chi-square tests for independence of Y and X in contingency table analysis. The pioneer of these ideas was Henry Oliver Lancaster.

15 Multi-sample Multivariate Data

We propose that the general case of statistical data analysis is multi-sample multivariate data (or multi-way classification)

$$(Y_1, \dots, Y_m, X),$$

where X can be discrete ($X = 1, \dots, c$) or continuous.

Two important cases are:

two sample bivariate (Y_1, Y_2, X) where $X = 1, \dots, c$;

trivariate (Y_1, Y_2, Y_3) , three way classification.

Excellent discussions of multi-way classification discrete data analysis is given by Agresti (2002) in his book on categorical data analysis.

The goal of nonparametric estimation of $P[X|Y_1, \dots, Y_m]$ is approached by comparing conditional distribution $P[Y_1, \dots, Y_m|X]$ to unconditional distribution. Non-parametric solutions extending traditional methods are provided by Brunner, Dumhof, and Puri (2002).

For two sample bivariate data we propose constructing score functions $\psi_{k_i, Y_i}(y_i)$ from the marginal distribution of Y_i in the “pooled” sample. Our goal of estimating $P[X = k|Y_1, \dots, Y_m]/P[X = k]$ is approached by identifying multivariate score function

$$\psi_{k_1, Y_1}(y_1)\psi_{k_2, Y_2}(y_2)\psi_{k_3, Y_3}(y_3)$$

which are sufficient statistics for exponential model. As the first step in data analysis, we propose computing conditional component correlations (product moments)

$$C_{Y_1, Y_2, Y_3}(k_1, k_2, k_3|X = k) = E^{\sim}[\psi_{k_1, Y_1}(Y_1)\psi_{k_2, Y_2}(Y_2)\psi_{k_3, Y_3}(Y_3)|X = k]$$

Examples of this concept are given by Rayner and Best (2001), p. 147.

It should be noted that Hotelling’s bivariate two sample T -statistic is a tool for bivariate two sample data analysis. One can express it in terms of score function $g_{1, Y_i}(y)$. An analogous Wilcoxon statistic can be computed in terms of score function $\psi_{1, Y_i}(y)$.

16 Theory of Component Correlations

This section summarizes how to think about component correlations as providing diagnostic tools for many conventional statistical problems:

- (1) discrete data chi-square tests of independence and homogeneity,
- (2) discrete data log linear models,
- (3) non-parametric regression, prediction of Y by X ,
- (4) analysis of variance F tests

(5) non-parametric tests such as Wilcoxon and Kruskal-Wallis.

Important references are the pioneering research of Lancaster (1969) and canonical correlation extensions by Buja (1990).

Bivariate discrete data observes (Y, X) where Y has r ordered values y_1, \dots, y_r , X has c ordered values x_1, \dots, x_c . Chi-square test of a sample of size n is $\text{Chi} = n C$, defining in terms of sample probabilities $p_{Y,X}(y, x), p_Y(y), p_X(x)$,

$$\begin{aligned} C &= \sum_{y,x} (p_{Y,X}(y, x) - p_Y(y)p_X(x))^2 / p_X(x)p_Y(y) \\ &= \sum_{y,x} p_Y(y)p_X(x) \left(\frac{p_{Y,X}(y, x)}{p_Y(y)p_X(x)} - 1 \right)^2 \\ &= \sum_{y,x} (1 - p_Y(y))(1 - p_X(x)) R^2(Y = y, X = x) \end{aligned}$$

A “cumulant chi-square” statistic is defined in terms of $R^2(Y \leq y, X \leq x)$.

In terms of orthogonal polynomial score functions $g_{j,Y}(y), g_{k,X}(x)$ we can write C , which is the square norm of a function, as a sum of squares of “Fourier coefficients” $C_{j,k}(Y, X)$:

$$C = \sum_{j,k} C_{j,k}^2(Y, X)$$

where

$$\begin{aligned} C_{j,k}(Y, X) &= \sum_{x,y} p_Y(y)p_X(x) \frac{p_{Y,X}(y, x)}{p_Y(y)p_X(x)} g_{j,Y}(y)g_{k,X}(x) \\ &= E_{X,Y}[g_{j,Y}(Y)g_{k,X}(X)] \end{aligned}$$

The “Fourier coefficients” $C_{j,k}(Y, X)$ provide a representation of the raw joint probability mass function

$$\frac{p_{Y,X}(y, x)}{p_Y(y)p_X(x)} = 1 + \sum_{j,k} C_{j,k}(Y, X)g_{j,Y}(y)g_{k,X}(x)$$

A smooth estimator can be formed

$$p_{Y,X}^\wedge(y, x) = p_Y(y)p_X(x) \left\{ 1 + \sum_{\text{select}(j,k)} C_{j,k}(Y, X)g_{j,Y}(y)g_{k,X}(x) \right\}$$

Model selecting chooses significantly non-zero $C_{j,k}(Y, X)$.

We prefer a density estimator guaranteed to be positive given by an exponential model

$$p_{Y,X}^\wedge(y, x) = p_Y(y)p_X(x) \exp \left(\tau_0 + \sum_{\text{select}(j,k)} \tau_{j,k}g_{j,Y}(y)g_{k,X}(x) \right)$$

satisfying constraints (estimating equations) for all select (j, k)

$$\sum_{y,x} (p_{Y,X}^{\wedge}(y, x) - p_{Y,X}(y, x)) g_{j,Y}(y) g_{k,X}(x) = 0.$$

We unify density estimation and the problem of smooth regression of Y or X by writing a formula for conditional mean of $g_{j_0,Y}(Y)$ given X :

$$\begin{aligned} E[g_{j_0,Y}(Y)|X = x] &= \sum_y p_{Y|X=x}(y) g_{j_0,Y}(y) \\ &= \sum_y p_Y(y) \psi_{j_0,Y}(y) \left(1 + \sum_{\text{all } (j,k)} C_{j,k}(Y, X) g_{j,Y}(y) g_{k,X}(x) \right) \\ &= \sum_{\text{all } k} C_{j_0,k}(Y, X) g_{k,X}(x). \end{aligned}$$

Smooth estimator of conditional mean can be constructed

$$E^{\wedge}[g_{j_0,Y}(Y)|X = x] = \sum_{\text{select } k} C_{j_0,k}(Y, X) g_{k,X}(x).$$

This can be shown to be minimum mean square prediction of $g_{j_0,Y}(Y)$ given $g_{k,X}(X)$ for select k .

Simple linear regression is

$$E^{\wedge}[g_{1,Y}(Y)|X = x] = C_{1,1}(Y, X) g_{1,X}(x).$$

These formulas, derived for discrete data, **apply also for continuous data.**

17 Cum-Cum Plots

Multi-samples are represented as (Y, X) , $X = 1, \dots, c$. Sample distribution of Y in k -th sample is denoted $F_{Y|X=k}(y)$. Pooled distribution is “unconditional” distribution

$$F_Y(y) = \sum_k P[X = k] F_{Y|X=k}(y).$$

Let y_1, \dots, y_r denote ordered distinct values in pooled sample. A *PP* plot to test $F_{Y|X=k}(y) = F(y)$ joins linearly

$$(F_Y(y_j) = E_Y[I(Y \leq y_j)], F_{Y|X=k}(y_j) = E_{Y|X=k}[I(Y \leq y_j)]).$$

It estimates $F_{Y|X=k}(Q_Y(u))$ at u F_Y exact.

A *PP* plot can also be called a Cum-Cum plot, a concept which we can extend to cumulative sums of bivariate data, the linear join of

$$(E_X[I(X \leq x_k)], E_{Y,X}[\psi_{j,Y}(Y)I(X \leq x_k)])$$

where x_k denotes ordered distinct values of X . The cum-cum plot is a function of $0 \leq u \leq 1$ which estimates $u E[\psi_{j,Y}(Y)|X = Q_X(u)]$ for $u F_X$ -exact ($u = F_X(x)$ for some x). It provides tools for changepoint analysis, forming for $(Y_j, j), j = 1, \dots, n$, the linear join of

$$\left(\frac{k}{n}, \frac{1}{n} \sum_{j=1}^k (Y_j - \bar{Y}) \right) \quad k = 1, 2, \dots, n.$$

18 Appendix: Score functions, orthogonal polynomials, mid distribution

Suppose a discrete random variable takes c values y_j with probability $p(y_j)$. We often take $p(y) = p^\wedge(y)$ in a typical application. Define

$$\begin{aligned} \mu &= \sum_{j=1}^c y_j p(y_j) = E[Y], \\ \sigma^2 &= \sum_{j=1}^c (y_j - \mu)^2 p(y_j) = E[(Y - \mu)^2] \\ \beta_r &= \sum_{j=1}^c ((y_j - \mu)/\sigma)^r p(y_j) = E[((Y - \mu)/\sigma)^r] \end{aligned}$$

Following Best and Rayner (1997), p. 756, we define and compute orthogonal polynomials $g_k(y_j), k = 1, 2, \dots, c - 1$ as follows

$$\begin{aligned} g_0(y_j) &= 1 \\ g_1(y_j) &= (y_j - \mu)/\sigma \\ g_2(y_j) &= ((g_1^2(y_j) - 1) - \beta_3 g_1(y_j))/a_2, \\ a_2^2 &= \beta_4 - \beta_3^2 - 1 \end{aligned}$$

We derive (compute) orthogonal polynomials $g_3(y_j), \dots, g_{c-1}(y_j)$ by Gram-Schmidt orthonormalization, or by recurrence relations in Emerson (1968).

When $p(y)$ are Binomial probabilities, $y_j = j, j = 0, 1, \dots, c, g_k(j)$ can be defined in terms of Krawtchouk polynomials.

When $p(y)$ are uniform distribution on $1, \dots, c$ or on $(j - .5)/c, j = 1, \dots, c, g_k(y)$ are usual kinds of orthogonal polynomials and are the same as mid-distribution score functions for uniform distribution.

Non-parametric score functions are defined by transforming y_j to mid-distribution

transform

$$\begin{aligned}
 F^{\text{mid}}(y_j) &= F(y_j) - .5p(y_j) \\
 \mu_{\text{mid}} &= \sum p(y)F^{\text{mid}}(y_j) = .5 = E[F^{\text{mid}}(Y)] \\
 \sigma_{\text{mid}}^2 &= \sum p(y)(F^{\text{mid}}(y_j) - .5)^2 = \text{VAR}[F^{\text{mid}}(Y)] \\
 &= \left(\frac{1}{12}\right)\left(1 - \sum_j p^3(y_j)\right)
 \end{aligned}$$

The beautiful, not widely known, formula for σ_{mid}^2 has at least two ingenious proofs.

Many non-parametric methods (including corrections for ties) are of the forms

$$T^{\sim}(\psi_j) = E^{\sim}[\psi_j(y)] = \sum_y p^{\sim}(y)\psi_j(y) = \sum_y (p^{\sim}(y) - p^{\wedge}(y))\psi_j(y)$$

where $\psi_j(y)$ are mid-distribution score functions custom designed for model $p^{\wedge}(y)$ with mid-distribution $F^{\text{mid}}(y)$ by $\psi_0(y) = 1$,

$$\begin{aligned}
 \psi_1(y) &= (F^{\text{mid}}(y) - .5)/\sigma_{\text{mid}} \\
 \psi_2(y) &= ((\psi_1^2(y) - 1) - \beta_3\psi_1(y))/a_2, \\
 a_2^2 &= \beta_4 - \beta_3^2 - 1 \\
 \beta_r &= \sum_y p^{\wedge}(y)((F^{\text{mid}}(y) - .5)/\sigma_{\text{mid}})^r.
 \end{aligned}$$

References

- Agresti, A. (2002) *Categorical Data Analysis*. New York: Wiley.
- Best, C. J. and Rayner, J. C. W. (1997) “Goodness of fit for this binomial distribution,” *Aust. J. Statist.*, **39**(3), 355–364.
- Brunner, E., Dumhof, S., and Puri, M. L. (2002) “Weighted rank statistics in factorial designs with fixed effects,” *Statistica Neerlandica*, **56**(2), 179–194.
- Buja, A. (1990) “Remarks on functional canonical variates, alternating least squares methods, and ACE,” *Annals of Statistics*, **18**(3), 1032–1069.
- Emerson, P. L. (1968) “Numerical construction of orthogonal polynomials from a general recurring formula,” *Biometrics*, **24**, 695–701.
- Eubank, R. L. and LaRiccia, V. N. (1990) “Components of Pearson’s phi-squared distance measure for the k -sample problem,” *J. Amer. Stat. Assoc.*, **85**, 441–445.
- Lancaster, H. O. (1969) *The Chi-Squared Distribution*, New York: Wiley.
- Lindley, D. V. (2002) “Letter to the Editor: The experience of probability in data analysis,” *Teaching Statistics*, **24**(1), 22–23.

- McPherson, G. (1989) "The Scientists' View of Statistics—A Neglected Area," *J. Royal Statist. Soc. A*, **152**, 221–240.
- Parzen, E. (1979) "Nonparametric Statistical Data Modeling," *Journal of the American Statistical Association*, (with discussion), **74**, 105–131.
- Parzen, E. (1992) "Comparison Change Analysis," *Nonparametric Statistics and Related Topics* (ed. A.K. Saleh), Elsevier: Amsterdam, 3–15.
- Parzen, E. (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.
- Rayner, J. C. W. and Best, D. J. (1989) *Smooth Tests of Goodness of Fit*, Oxford University Press: New York.
- Rayner, J. C. W. and Best, D. J. (2001) *A Contingency Table Approach to Nonparametric Testing*, Chapman & Hall/CRC: Boca Raton.