September 2000

QUANTILE/QUARTILE PLOTS,CONDITIONAL QUANTILES,COMPARISON DISTRIBUTIONS

Emanuel Parzen
Texas A&M University, Department of Statistics

1. Histograms and Order Statistics

Histograms are traditionally plotted by statisticians to identify distributions that fit a sample (data set) $Y_1, \ldots, Y_n$.

Order statistics (sample values arranged in increasing order are denoted $Y(1;n) \leq \cdots \leq Y(n;n)$.

2. Sample Quantile Function

For identifying distributions that fit data we recommend the sample quantile function

$$Q^{\sim}(u) = F^{\sim^{-1}}(u) = Y(j;n), \quad (j-1)/n < u \leq j/n,$$

inverse of the sample distribution function

$$F^{\sim}(y) = E^{\sim}\left[I(Y \leq y)\right] = (1/n)\Sigma_{t=1}^{n}\, I(Y_t \leq y)$$

3. Population Ensemble Quantile Function

When we regard $Y_1, \ldots, Y_n$ as a random sample of Y, we define population distribution function $F(y)$, $-\infty < y < \infty$, and quantile function $Q(u)$, $0 \leq u \leq 1$,

$$F(y) = P\left[Y \leq y\right] = E\left[I(Y \leq y)\right]$$
$$Q(u) = F^{-1}(u) = \inf\left\{y : F(y) \geq u\right\}$$

4. Density Quantile, Quantile Density

If F is continuous, $F(Q(u)) = u$ for all u and (differentiating)

$$f(Q(u))\; Q'(u) = 1.$$

Quantile density $q(u) = Q'(u)$.
Density quantile $fQ(u) = f(Q(u))$.

Asymptotic distribution of sample quantiles which makes statistical inference possible is given by

$$\sqrt{n}\; fQ(u)(Q^{\sim}(u) - Q(u)) \to_d \; B(u)$$

where $B(u)$, $0 < u < 1$, is Brownian Bridge, zero mean Gaussian process with covariance $E\left[B(s)B(t)\right] = \min(s,t) - st$.

Analogous limit theorems can be proved for estimators of conditional quantile functions $Q_{Y|X=x}(u)$.

5.Tail Classification of Distributions

Tail classification of distribution functions can be described by exponents of regular variation α_0 and α_1:

$$fQ(u) = u^{\alpha_0}\ L_0(u), \qquad u \text{ near } 0,$$
$$fQ(u) = (1-u)^{\alpha_1}\ L_1(u), \quad u \text{ near } 0.$$

In terms of α we define

$$\begin{aligned} \alpha > 1,&\ \text{ long tail;} \\ \alpha = 1,&\ \text{ medium tail;} \\ 0 \leq \alpha = 1,&\ \text{ short tail;} \\ \alpha < 0,&\ \text{ infinitely short tail.} \end{aligned}$$

6.Continuous Sample Quantile

In practice we prefer continuous sample quantile

$$Q^{\sim c}((j-.5)/n) = Y(j;n),\ j = 1,\ldots,n,$$
$$Q^{\sim c}(0) = Y(1;n),\ Q^{\sim c}(1) = Y(n;n)$$

and defined by linear interpolation at other u values. The continuous quantile command in Splus is related to our definition by

$$Q^{\sim c}(p) = Q^{\sim c}_{\text{Splus}}(p + ((p-.5)/(n-1)))$$

Note $Q^{\sim c}_{\text{Splus}}((j-1)/(n-1)) = Y(j;n),\ j = 1,\ldots,n.$

7.Quantile Function of Transform

Assume $Y = g(W)$ where $g(w)$ is quantile-like (is non-decreasing and continuous from the left).

$$F_Y(y) = F_W(g^{-1}(y))$$
$$Q_Y(u) = g(Q_W(u))$$
$$Q_{Y|X=x}(u) = g(Q_{W|X=x}(u))$$

8.Distribution Transform

When F is continuous, $F_Y(Y)$ is Uniform $(0,1)$ since

$$Q_{F_Y(Y)}(u) = F_Y(Q_Y(u)) = u$$

We call $F_Y(Y)$ distribution transform or probability integral transform. More important is mid-distribution transform $F_Y^{\text{mid}}(Y)$, defining mid-distribution function

$$F_Y^{\text{mid}}(Y) = F_Y(Y) - .5p_Y(Y).$$

Randomized distribution is

$$F_Y^{\text{rand}}(y) = F_Y(y) - U(y)p_Y(y), \quad U(y) \text{ Uniform } (0,1).$$

9. Conditional Quantile Function

Bivariate data (X, Y) is often modeled by conditional mean $E[Y \mid X = x]$. To model conditional distribution

$$F_{Y|X=x}(y) = P[Y \leq y \mid X = x]$$

we recommend estimating conditional quantile function

$$Q_{Y|X=x}(u) = \inf \{y : F_{Y|X=x}(y) \geq u\}.$$

For jointly normal (X, Y) conditional distribution of Y given $X = x$ is normal,

$$Q_{Y|X=x}(u) = \mu_{Y|X=x} + \sigma_{Y|X=x}\Phi^{-1}(u).$$

10. Bayes Theorem Conditional Quantile Function

It is not true that $Q_Y(F_Y(y)) = y$ for all y, but for almost all values of the random variable Y

$$Q_Y(F_Y(Y)) = Y$$

Therefore by formula for quantile function of a tranform

$$Q_{Y|X=x}(u) = Q_Y(Q_{F_Y(Y)|X=x}(u))$$

We have reduced estimating the conditional quantile of Y to estimating the conditional quantile of the distribution transfrom $F_Y(Y)$ which we next interpret as a comparison distribution and estimate as a comparison density. One can apply this formula to Bayes estimation of a parameter θ from data X.

11. Comparison Distribution, Comparison Density

To compare two distributions F and G, a universal problem of statistical inference, we define concepts of comparison distribution $D(u; F, G)$, $0 \leq u \leq 1$, and comparison density

$$d(u; F, G) = D'(u; F, G).$$

When F and G are continuous, and $G << F$ ($f(y) = 0$ implies $g(y) = 0$),

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

When F and G are discrete, and $G << F$ (probability mass function $p_F(y) = 0$ implies $p_G(y) = 0$),

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

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

12. Exact u, PP plots

We call u F exact if u is in the range of F, $u = F(y)$ for some y. Then $Q(u) = \inf\{y : F(y) = u\}$, and $F(Q(u)) = u$. For u F exact

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

for other u, $D(u; F, G)$ is defined by linear interpolation between its value at exact u. Change comparison distribution

$$\text{Change } D(u; F, G) = G(F^{-1}(u)) - F(F^{-1}(u)).$$

Graph of D, called PP plot, connects linearly

$$(0,0),\ (F(y_j), G(y_j)),\ (1,1)$$

where y_j are jump points of F (assume F discrete).

13. Comparison Approach to Estimating Conditional Quantiles

Bayes' theorem for conditional quantiles is written

$$Q_{Y|X=x}(u) = Q_Y(Q_{F_Y(Y)|X=x}(u)) = Q_Y(s)$$

We propose to find

$$s = Q_{F_Y(Y)|X=x}(u)$$

by computing the function of s, $0 < s < 1$,

$$u = F_{F_Y(Y)|X=x}(s) = P\left[F_Y(Y) \le s \mid X = x\right]$$

When Y is continuous

$$u = F_{Y|X=x}(Q_Y(s)) = D(s; F_Y, F_{Y|X=x})$$

When Y is discrete and s is F_Y exact

$$F_Y(Y) \le s = F_Y(Q_Y(s)) \text{ if } fY \le Q_Y(s),$$

$$u = P\left[F_Y(Y) \le s \mid X = x\right] = P\left[Y \le Q_Y(s) \mid X = x\right] = D(s;\ F_Y,\ F_{Y|X=x})$$

14. Bayes Comparison Theorem for Conditional Quantiles

We can verify the following extension of Bayes theorem for conditional quantile functions:

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

Proof for Y discrete:
For discrete Y with ordered values $y_j, Q_Y(s) = y_j$ for s in interval $F_Y(y_{j-1}) < s \le F_Y(y_j)$.

For u in interval $F_{Y|X=x}(y_{j-1}) < u \le F_{Y|X=x}(y_j), s = D^{-1}(u; F_Y, F_{Y|X=x})$ varies linearly between $F(y_{j-1})$ and $F(y_j)$, and $Q_Y(s) = y_j$.

For interval $F_{Y|X=x}(y_{j-1}) < u \le F_{Y|X=x}(y_j), Q_{Y|X=x}(u) = y_j = Q_Y(s)$.

15. Formula for Conditional Mean From Conditional Quantile Function

Our computational process for computing conditional quantiles involves estimating in succession as a function of x

$$\begin{aligned} d(s; F_Y, F_{Y|X=x}), &\quad 0 < s < 1 \\ u = D(s; F_Y, F_{Y|X=x}), &\quad 0 < s < 1 \\ s = D(u; F_{Y|X=x}, F_Y), &\quad 0 < u < 1 \\ Q_{Y|X=x}(u) = Q_Y(s), &\quad 0 < u < 1 \end{aligned}$$

The relation between the above functions is illustrated by formulas for the conditional mean.

Quantile formulas for conditional mean. When Y continuous

$$\begin{aligned} E[Y \mid X = x] &= \int_{-\infty}^{\infty} y f_{Y|X=x}(y) dy \\ &= \int_{-\infty}^{\infty} y \frac{f_{Y|X=x}(y)}{f_Y(y)} dF_Y(y), y = Q_Y(s) \\ &= \int_0^1 Q_Y(s) \frac{f_{Y|X=x}(Q_Y(s))}{f_Y(Q_Y(s))} ds \\ &= \int_0^1 Q_Y(s) d(s; F_Y, F_{Y|X=x}) ds \\ &= \int_0^1 Q_Y(s) dD(s; F_Y, F_{Y|X=x}), u = D(s; F_Y, F_{Y|X=x}) \\ &= \int_0^1 Q_Y(D^{-1}(u; F_Y, F_{Y|X=x})) du \\ &= \int_0^1 Q_{Y|X=x}(u) du \end{aligned}$$

When Y discrete, similiar formulas start with

$$E[Y \mid X = x] = \sum_y y p_{Y|X=x}(y).$$

16. Logistic Regression Estimation of Conditional Comparison Density

Let $s_j = j/m, j = 0, 1, \ldots, m$ (often $m = 20$). Approximately

$$\begin{aligned} &d(s_j; F_Y, F_{Y|X=x}) \\ &= (s_j - s_{j-1})^{-1} P[Q_Y(s_{j-1}) < Y \le Q_Y(s_j) \mid X = x] \end{aligned}$$

which can be estimated for fixed s_j as a function of x by logistic regression (for which there exists many parametric and non-parametric methods).

17. Rejection Sampling Computation of Conditional Quantile of Distribution Transform

Combining these estimates as a function of x for fixed s_j one can form a piecewise constant estimate as a function of s for fixed x.

From the comparison density $d(s; F_Y, F_{Y|X=x}), 0 < s < 1$ one can estimate by rejection sampling the values of the comparison quantile

$$s = D^{-1}(u; F_Y, F_{Y|X=x}), 0 < u < 1.$$

18. Mid-Distribution Transform

A unifying role in non-parametric data analysis is played by the mid-distribution transform (equivalent to tied ranks of a sample)

$$\begin{aligned} F_Y^{\text{mid}}(Y) &= F_Y(Y) - .5p_Y(Y) \\ E\left[F_Y^{\text{mid}}(Y)\right] &= .5, \\ \text{VAR}\left[F_Y^{\text{mid}}(Y)\right] &= (1/12)(1 - E\left[p_Y^2(Y)\right]). \end{aligned}$$

The importance of $F_Y^{\text{mid}}(Y)$ in non-parametric statistical data analysis is illustrated by the formula for a score statistic to test $H_0 : F_{Y|X=x} = F_Y$:

$$T(J) = \int_0^1 J(u)d(u; F_Y, F_{Y|X=x})du.$$

When Y is discrete, with distinct values $y_1, \ldots, y_k$,

$$T(J) = \Sigma_{j=1}^k (p_{Y|X=x}(y_j)/p_Y(y_j)) \int_{F_Y(y_{j-1})}^{F_Y(y_j)} J(u)du.$$

Approximately

$$\begin{aligned} T(J) &= \Sigma_{j=1}^k p_{Y|X=x}(y_j)J(F_Y^{\text{mid}}(y_j)) \\ &= E\left[J(F_Y^{\text{mid}}(Y)) \mid X = x\right]. \end{aligned}$$

Linear rank statistics can be represented

$$T^{\sim}(J) = E^{\sim}\left[J(F_Y^{\sim\text{mid}}(Y)) \mid X = x\right].$$

19. Location Scale Quantile Models

The sample quantile $Q^{\sim}(u)$ is a non-parametric estimator of the population quantile $Q(u)$. A parametric estimator can be formed from a location-scale model

$$Q(u) = \mu + \sigma Q_0(u)$$

where $Q_0(u)$ is known.

Maximum likelihood estimators of μ and σ, denoted $\hat{\mu}$ and $\hat{\sigma}$, yield estimator $\hat{Q}(u) = \hat{\mu} + \hat{\sigma}Q_0(u)$.

Asymptotically efficient estimates, denoted $\mu_{n,L}$ and $\sigma_{n,L}$, can be formed as a linear functional of order statistics (sample quantile function); they can be computed by continuous parameter regression analysis from the asymptotic representation of the sample quantile as a linear regression

$$f_0Q_0(u)Q^{\sim}(u) = \mu f_0Q_0(u) + \sigma f_0Q_0(u)Q_0(u) + \sigma B(u).$$

20. Location Scale Models for Conditional Quantiles

$$Q_{Y|X=x}(u) = \mu_{Y|X=x} + \sigma_{Y|X=x}Q_0(u)$$

Estimators

$$\hat{Q}_{Y|X=x}(u) = \hat{\mu}_{Y|X=x} + \hat{\sigma}_{Y|X=x}Q_0(u)$$

can be formed in the same way as in the unconditional case from linear functions of our non-parametric estimators denoted $Q^{\sim}_{Y|X=x}(u)$, of the conditional quantile $Q_{Y|X=x}(u)$.

To compute these parametric estimators we assume $Q_0(u)$ known. We can compare several choices of $Q_0(u)$ by plotting $Q^{\sim}_{Y|X=x}(u)$ and $Q^{\wedge}_{Y|X=x}(u)$ on scatter diagrams.

Estimation of $\mu_{|X=x}$ is alternative to non-parametrically estimating $E\left[Y \mid X=x\right]$.

21. Confidence Intervals for $Q_Y(u)$ and Conditional Quantile $Q_{Y|X=x}(u)$

Confidence intervals for a parameter $Q(u)$ can be formed from the asymptotic distribution of $Q^{\sim}(u)$:

$$\sqrt{n}(Q^{\sim}(u) - Q(u)) \to_d B(u)/fQ(u)$$

This formula has a severe disadvantage , it requires estimation of $fQ(u)$.

From the values of the sample quantile functions $Q^{\sim}_{Y}(u)$ (Similarly for $Q^{\sim}_{Y|X=x}(u)$) one can obtain a confidence interval for $Q(u)$ using facts such as

$$\sqrt{n}(F^{\sim}_n(Q(u)) - u) \to_d B(u)$$

One can find functions $c_1(u)$ and $c_2(u)$ such that with probability greater than α, for all u,

$$u - (c_1(u)/\sqrt{n}) < F^{\sim}_n(Q(u)) < u + (c_2(u)/\sqrt{n}),$$
$$Q^{\sim}_n(u - (c_1(u)/\sqrt{n})) < Q(u) < Q^{\sim}_n(u + (c_2(u)/\sqrt{n}))$$

From parametric estimates of a location-scale model

$$\hat{Q}(u) = \hat{\mu} + \hat{\sigma}Q_0(u)$$

or

$$\hat{Q}(u) = \mu_{n,L} + \sigma_{n,L}Q_0(u)$$

one can derive a simultaneous confidence interval (Rosenkrantz (2000))

$$\hat{Q}(u) - c_1(u,n) \leq Q(u) \leq \hat{Q}(u) + c_2(u,n)$$

This location-scale model confidence interval is shorter than the non-parametric confidence intervals above.

22. Five Number Quantile Summary

Quantile function can "compress data" by a five number summary: values of $Q(u)$ at

$$u = .05, .25, .5, .75, .95$$

Median $Q(.5)$
Quartiles $Q(.25)$, $Q(.75)$
Mid Quartile $QM = .5(Q(.25) + Q(.75))$
Quartile deviation $QD = 2(Q(.75) - Q(.25))$
QD approximation to $Q(.5)$
Normal distribution $QD = 2.7\sigma$.

23. Quantile Quartile Function $Q/Q(u), 0 < u < 1$

To use quantile functions to identify distributions fitting data we propose

$$Q/Q(u) = (Q(u) - QM)/QD = Q_{Y^Q}(u),$$

defining transform of data

$$Y^Q = (Y - QM)/QD$$

Claim: plot of $y = Q/Q(u)$ contains all the insights of a box plot. Add to plot dotted lines

$$\begin{array}{ll} \text{horizontal} & y = -1, -.5, 0, .5, 1 \\ \text{vertical} & u = .05, .25, .5, .75, .95 \end{array}$$

Five number summary of distribution becomes

$$\begin{array}{rl} QM, & \text{location} \\ QD, & \text{scale} \\ Q/Q(.5), & \text{skewness} \\ Q/Q(.05), & \text{righttail} \\ Q/Q(.95), & \text{lefttail} \end{array}$$

Elegance of $Q/Q(u)$ is its universal values

$$Q/Q(.25) = -.25$$
$$Q/Q(.75) = .25$$

Tukey outliers correspond to $\mid Q/Q(u) \mid > 1$.
Tukey criteria for outlier value $Q(u)$ outside fences:

$$Q(u) > Q(.75) + 1.5(Q(.75) - Q(.25)), Q(u) - QM > QD,$$
$$Q(u) < Q(.25) - 1.5(Q(.75) - Q(.25)), Q(u) - QM < QD.$$

Diagnostics of left and right tail behavior:

$$\begin{array}{rll}
\text{Short}: & -5 < Q/Q(.05) & Q/Q(.95) < .5 \\
\text{Medium}: & -1 < Q/Q(.05) < -.5 & .5 < Q/Q(.95) < 1 \\
\text{Long}: & Q/Q(.05) < -1 & 1 < Q/Q(.95).
\end{array}$$

Diagnostics of skewness:

$$\begin{array}{lll}
Q/Q(.5) > 0, & \text{mean} < \text{median} < \text{mode}, & \text{left} - \text{skewed} \\
Q/Q(.5) < 0, & \text{mode} < \text{median} < \text{mean}, & \text{right} - \text{skewed}.
\end{array}$$

24. Conditional Quantile Five Number Summary

From the conditional quantile $Q_{Y|X=x}(u)$ at $u = .05, .25, .5, .75, .9$, we recommend five number summary:

location of conditional distribution

$$QM_{Y^Q|X=x} = (QM_{Y|X=x} - QM_Y)/QD_Y$$

scale of conditional distribution

$$QD_{Y^Q|X=x} = QD_{Y|X=x}/QD_Y$$

skewness of conditional distribution

$$Q/Q_{Y|X=x}(.5) = (Q_{Y|X=x}(.5) - QM_{Y|X=x})/QD_{Y|X=x}$$

right tail of conditional distribution

$$\begin{aligned} &Q/Q_{Y|X=x}(.95) = \\ &(Q_{Y|X=x}(.95) - QM)_{Y|X=x})/QD_{Y|X=x} \end{aligned}$$

$Q/Q_{Y|X=x}(.05)$ defined similarly.

25. Conditional Quantile Scatter Plots, Diagnostic Plots

We recommend plotting as a function of x for $u = .05, .25, .5, .75, .95$

$$y = Q_{Y^Q|X=x}(u)$$

on a scatterdiagram of (X^Q, Y^Q). Also plot $y = QM_{Y|X=x}$. On separate graphs, plot as function of x, $QD_{Y^Q|X=x}$, $Q/Q_{Y|X=x}(.05), Q/Q_{Y|X=x}(.95)$.

26. Unification of Conventional Statistical Methods

Conventional t test, Kruskal-Wallis test, goodness of fit) methods of two sample and multi-sample data analysis can be extended and unified by representing the data as (X, Y) and computing and graphing

conditional quantiles of Y given $X = x$,

conditional comparison quantiles of $F_Y^{\text{mid}}(Y)$,

conditional comparison distribution of $F_Y^{\text{mid}}(Y)$.

These are functions of u for each of the discrete x values of X. We plot summaries as a function of x plotted at $F_X^{\mathrm{mid}}(x)$.

To summarize data in many samples we represent (X_j, Y_j) where X_j denotes population (denoted $k = 1, \ldots, c$) from which response Y_j was observed. The values of Y for a fixed value X can be summarized by a conditional quantile function $Q_{Y|X}{}^{\sim}(u)$. The pooled sample of Y values is summarized by an unconditional quantile $Q_Y{}^{\sim}(u)$. As an alternative to running box plots to compare and summarize the different samples we propose conditional quantile/quartile plot which connects linearly points $(F_X^{mid}(k),\ Q/Q_{Y|X=k}(u)),\ k = 1, \ldots, c$. One plots this curve for $u = .05, .25, .5, .75, .95$. One also plots constant lines $Q/Q_Y(u),\ u = .05, .25, .5,\ .75, .95$.

References

Gilchrist, Warren (2000). *Statistical Modeling with Quantile Functions*, Chapman and Hall/CRC Press: Boca Raton.

Handcock, Mark and Martina Morris (1999). *Relative Distribution Methods in the Social Sciences*, Springer: New York.

Heckman, Nancy and R.H. Zamar (2000). "Comparing the Shapes of Regression Functions," *Biometrika*, 87, 135-144.

Parzen, Emanuel (1977). Discussion to Stone (1977).

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

Parzen, Emanuel (1989). "Multi-Sample Functional Statistical Data Analysis," *Statistical Data Analysis and Inference Conference in Honor of C.R. Rao,* (ed. Y. Dodge), Amsterdam: Elsevier, 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 (1996). "Concrete Statistics," *Statistics in Quality,* S. Ghosh, W. Schucany, W. Smith, Marcel Dekker: New York, 309-332.

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.

Rosenkrantz, Walter (2000). "Confidence Bands for Quantile Functions: A Parametric and Graphic Alternative for Testing Goodness of Fit," *The American Statistician,* 54, 185-190.

Stone, Charles (1977). "Consistent Nonparametric Regression," *Annals of Statistics,* 5, 595-645.