

QUANTILE PROBABILITY AND STATISTICAL DATA MODELING

by Emanuel Parzen
Texas A&M University

ABSTRACT

Quantile and conditional quantile statistical thinking, as I have innovated it in my research since 1976, is outlined in this comprehensive survey and introductory course in quantile data analysis. We propose that a (grand) unification of the theory and practice of statistical methods of data modeling may be possible by a quantile perspective. Our broad range of topics of univariate and bivariate probability and statistics are best summarized by the key words. Two fascinating practical examples are given involving positive mean and negative median investment returns and relation between radon concentration and cancer.

Key Words. Mid-distribution transform, Percent function, Percentile function, Quantile Function, Monotone Transform, Parameter inverse pivot quantile function, Confidence $Q - Q$ curve, Quantile/Quartile Function $Q/Q(u)$, Density quantile, Quantile density, Conditional quantile, Comparison Distribution, Comparison Density, Bayesian Inference using quantile simulation, bivariate dependence, component correlations.

0. Philosophy. Quantile and conditional quantile statistical methods are not widely practiced in introductory statistics courses. They were pioneered by Galton (1889) who computed medians and quartiles of conditional distributions of heights of sons given heights of parents, and discovered that they had constant scale and linear location. Galton thus pioneered regression, correlation, bivariate normal distributions, and conditional normal distributions. Many facts about quantiles have a long history and were known before 1900 (see Hald 1998)).

Quantile statistical thinking, as I have innovated it in my research since Parzen (1979), is outlined in this paper. My teaching philosophy has as its maxim: to learn more, learn more, and believe that learning a lot (answering all related questions) is easier than learning little (answering only the questions asked).

I teach that statistics (done the quantile way) can be simultaneously frequentist and Bayesian, confidence intervals and credible intervals, parametric and nonparametric, continuous and discrete data. Your first choice of data models is parametric; if they don't fit you provide nonparametric models for fitting and simulating the data. The practice of statistics, and the modeling (mining) of data, can be elegant and provide intellectual and sensual pleasure. Fitting distributions to data is an important industry in which

statisticians are not yet vendors. We believe that unifications of statistical methods can enable us to advertise “What is your question? Statisticians have answers!”

1. Probability Law of Random Variable Y . To describe the probability distribution of a random variable Y concepts include: distribution function $F(y) = P[Y \leq y]$, quantile function $Q(u) = F^{-1}(u)$, probability mass function $p(y) = P[Y = y]$, probability density function $f(y) = F'(y)$, and mid-distribution function $F^{mid}(y) = F(y) - .5p(y)$.

To denote the distinct concepts of $p(y)$ and $f(y)$, the same letter should not be used; using the same letter is detrimental to quantile domain and Bayesian reasoning. A discrete random variable can be described by $p(y)$ and a continuous can be described by $f(y)$.

Important examples of continuous distributions are standard exponential $f(y) = e^{-y}$, $F(y) = 1 - e^{-y}$, and standard normal $\phi(y), \Phi(y)$. Location-scale models for continuous random variables Y represent $Y = \mu + \sigma Y_0$ where Y_0 has standard distribution $F_0(y)$; then $F(y) = F_0((y - \mu)/\sigma)$. Normal (μ, σ) distribution has $F(y) = \Phi((y - u)/\sigma)$.

2. Mid-distribution Transform. The mid-distribution function concept $F^{mid}(y)$ is important for discrete distributions, especially sample distribution functions. When F is continuous $U = F(Y)$ is Uniform $(0,1)$. When F is discrete we use mid-distribution transform $W = F^{mid}(Y)$; it has mean $E(W) = .5$ and variance.

$$\text{VAR}(W) = (1/12)(1 - E[p^2(Y)])$$

I would appreciate information about published proofs of this elegant formula for $\text{VAR}(W)$; it is important for applications to data with ties (compare Heckman and Zamar (2000)).

3. Sample Distribution Function. A sample Y_1, \dots, Y_n has: a sample distribution function

$$F^\sim(y) = P^\sim[Y \leq y] = (1/n) \sum_{t=1}^n I(Y_t \leq y)$$

where $I(Y \leq y) = 1$ or 0 as $Y \leq y$ or $Y > y$; sample probability mass function $p^\sim(y) = P^\sim[Y = y]$; sample mid-distribution function $F^{\sim mid}(y) = F^\sim(y) - .5p^\sim(y)$. A continuous version $F^{\sim c}(y)$ of the discrete sample distribution $F^\sim(y)$ is defined below.

4. Percent Function. The distribution function can be denoted $u = F(y) = u(y)$ and called the percent function since $u(y)$ is the percent of the population whose values are less than or equal to y . Percent is similar to the p -value of the statistic T under a null hypothesis H_0 about the distribution of T .

5. Percentile Function. The percentile or quantile function is the inverse $y = Q(u) = F^{-1}(u) = y(u)$ of $u = F(y) = u(y)$. We call u the percent of y , and y the percentile of u .

To rigorously define $y = Q(u)$ suppose first that u is in the range of F ; there exists a value y such that $u = F(y)$. Define $y = Q(u)$ to be the smallest y such that $u = F(y)$ and $F(Q(u)) = u$. The general definition of quantile function: for $0 \leq u \leq 1$

$$Q(u) = F^{-1}(u) = \inf(y : F(y) \geq u)$$

The graph of $y = Q(u)$ is a rotation of the graph of $u = F(y)$. Experts on perception report that rotating a picture often helps us see patterns. Verify geometrically that

$$\int_{-\infty}^{\infty} |F_1(y) - F_2(y)| dy = \int_0^1 |Q_1(u) - Q_2(u)| du.$$

Quantile $y = Q(u)$ of standard exponential

$$u = F(y) = 1 - e^{-y}, \quad y = Q(u) = -\log(1 - u).$$

Quantile of standard Normal (0,1) is $\Phi^{-1}(u)$. Excellent approximation for ν large of quantile $Q_\nu(u)$ of Gamma(ν)/ ν is given by Wilson-Hillferty transformation:

$$Q_\nu(u) = \left(\left(1 - \frac{1}{9\nu}\right) + \frac{1}{3} \frac{1}{\nu^{.5}} \Phi^{-1}(u) \right)^3.$$

6. Quantile formula for mean, variance.

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} y dF(y) = \int_0^1 Q(u) du, \\ \text{VAR}(Y) &= \int_0^1 (Q(u) - E(Y))^2 du \end{aligned}$$

For a sample Y_1, \dots, Y_n the sample mean \bar{Y} should be computed NOT by

$$\bar{Y} = (1/n) \sum_{t=1}^n Y_t$$

but by

$$\bar{Y} = (1/n) \sum_{t=1}^n Y(j; n) = \int_0^1 Q^\sim(u) du$$

where $Y(1; n) \leq \dots \leq Y(n; n)$ are the order statistics of the sample and $Q^\sim(u)$ is the sample quantile function. Quantile thinking defines statistics as summation done by sorting (ranking) data before adding.

A mean can be a misleading summary of a distribution; one should always plot the quantile function to learn skewness and tails, and outliers (see Appendix for very practical example).

The sample mean $\bar{Y} = \mu^\sim = E^\sim[y]$, the mean of the sample distribution. The sample variance should be defined as the variance of the sample distribution, defined

$$\sigma^{\sim 2} = (1/n) \sum_{t=1}^n (Y_t - \bar{Y})^2.$$

We believe teaching statistics is made difficult by the popular definition of sample variance as

$$S^2 = \sum_{t=1}^n (Y_t - \bar{Y})^2 / (n - 1);$$

S^2 should be called the adjusted sample variance and accompanied by our general definition of sample variance.

7. Percentile Method of Simulation. Quantile function $Q(u)$ can be used to simulate Y from U which is Uniform (0,1) by $Y = Q(U)$; one can show

$$P[Q(U) \leq y] = P[U \leq F(y)] = F(y) = P[Y \leq y].$$

8. Credible intervals. A $1 - \alpha$ credible interval for Y can be obtained from

$$P[y(\alpha/2) = Q(\alpha/2) \leq Y \leq Q(1 - (\alpha/2)) = y(1 - (\alpha/2))] = 1 - \alpha.$$

Let θ be a parameter of a probability model for Y ; given a prior distribution one can compute the quantile function $Q(u)$ of the posterior distribution of θ given data. One can express Bayesian credible interval for θ , with credibility $1 - \alpha$;

$$P[\theta(\alpha/2) = Q(\alpha/2) \leq \theta \leq Q(1 - (\alpha/2)) = \theta(1 - (\alpha/2)) | \text{data}] = 1 - \alpha.$$

9. Confidence interval, parameter inverse pivot quantile function. Let θ be parameter of a probability model $f(y|\theta)$; regard θ as a constant to be estimated. Assume we can form a pivot $T^\sim(\theta)$ satisfying (1) it is a function of θ and the data, which is increasing in θ ; (2) its distribution when θ is the true parameter value is identical with the distribution of random variable T with quantile function $Q_T(u)$. Define $\theta(u)$, $0 < u < 1$, by

$$T^\sim(\theta(u)) = Q_T(u), \quad \theta(u) = T^{\sim-1}(Q_T(u)).$$

We call $\theta(u)$ parameter inverse pivot quantile function. It satisfies $F_T[T^\sim(\theta(Q_T(u)))] = u$.

Conventional confidence intervals and hypothesis tests can be expressed in terms of $\theta(u)$. A $1 - \alpha$ confidence interval for θ is

$$\theta(\alpha/2) \leq \theta \leq \theta(1 - (\alpha/2)),$$

because when θ is the true parameter value the set of samples for which

$$Q_T(\alpha/2) = T^\sim(\theta(\alpha/2)) \leq T^\sim(\theta) \leq T^\sim(\theta(1 - (\alpha/2))) = Q_T(1 - (\alpha/2))$$

has probability $1 - \alpha$. The rejection region $\theta_0 \leq \theta(\alpha)$ has probability $P[T^\sim(\theta_0) \leq Q_T(\alpha)] = \alpha$ under the hypothesis $H_0 : \theta = \theta_0$.

Our concept $\theta(u)$ should be compared with the concept $\hat{\theta}_\alpha$ defined in the bootstrap percentile method of confidence intervals (see Davison and Hinkley (1997), p. 193) as a random variable which is an end point of a confidence interval. They define $P[\theta < \hat{\theta}_\alpha] = \alpha$; the probability function P should be denoted P_θ to emphasize that it is calculated under the assumption that θ is the true parameter value. Our more rigorous definition of $\theta(u)$ writes the probability statement

$$P_\theta[T^\sim(\theta(u)) \leq T(\theta(u))] = P[T \leq Q_T(u)] = u.$$

The concept of parameter inverse pivot quantile $\theta(u)$ facilitates computing confidence intervals for several confidence levels between .5 and .99 in order to discover any asymmetry in the confidence interval about the point estimator of θ .

10. Quantile function of monotone transformations. A distribution function $F(y)$ is non-decreasing and continuous from the right. A quantile function is non-decreasing and continuous from the left. For $g(y)$ non-decreasing and continuous from the left we define $g^{-1}(z) = \sup\{y : g(y) \leq z\}$. A beautiful and powerful property of quantile functions is formula for quantile function of $g(Y)$:

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

11. Inverse properties of quantiles under inequalities. To prove the monotone transform theorem we use the fact that in general the inverse properties of quantile functions hold under inequalities: $F(Q(u)) \geq u$,

$$F(y) \geq u \text{ if and only if } y \geq Q(u).$$

Similarly for $g(y)$ non-decreasing and continuous from left

$$g(y) \leq t \text{ if and only if } y \leq g^{-1}(t).$$

The formula for $Q_{g(Y)}(u)$ follows from

$$F_{g(Y)}(t) = P[g(Y) \leq t] = P[Y \leq g^{-1}(t)] = F_Y(g^{-1}(t))$$

and equivalence of following inequalities:

$$F_{g(Y)}(t) \geq u, F_Y(g^{-1}(t)) \geq u, g^{-1}(t) \geq Q_Y(u), t \geq g(Q_Y(u)).$$

12. Sample quantile function. For theory we use sample quantile function defined by $y^\sim(u) = Q^\sim(u) = F^{\sim-1}(u)$; it is piecewise constant and can be expressed in terms of order statistics $Y(1; n) \leq \dots \leq Y(n; n)$ of sample:

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

We can think of sample percentile as a fractional order statistic

$$y^\sim(u) = Y([un]; n)$$

where $[un] = j$ if $j-1 < un \leq j$.

For practice we would like a definition of sample quantile whose sample median $y^\sim(.5)$ agrees with usual definition: if $n = 2m + 1$, $y^\sim(.5) = Y(m + 1; n)$; if $n = 2m$, $y^\sim(.5) = .5(Y(m; n) + Y(m + 1; n))$. A definition of sample quantile function which yields these formulas is continuous version sample quantile $Q^{\sim c}(u)$. If sample consists of distinct values define $Q^{\sim c}(u)$ as piecewise linear connecting values

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

Many computer programs (such as Splus and Excell) use ad hoc definitions

$$\begin{aligned} Q^{\sim c}((j-1)/(n-1)) &= Y(j; n), \\ Q^{\sim c}(j/(n+1)) &= Y(j; n), \\ Q^{\sim c}((j-a)/(n+1-2a)) &= Y(j; n) \end{aligned}$$

for some constant a . *Example:* 9, 10, 11, 21, 26, 48, 56, 60, 60, 99 has sample lower quartile 11 by definition $a = .5$ and 13.5 by definition $a = 1$.

Our definition extends to the case of ties in the sample. Denoting distinct values in sample by y_1, \dots, y_r , define Q^{\sim} as piecewise linear connecting

$$Q^{\sim c}(F^{\sim mid}(y_j)) = y_j$$

We consider $y^{\sim}(u) = Q^{\sim c}(u)$ to be a definition of fractional order statistic. A continuous version sample distribution $F^{\sim c}(y)$ is defined as piecewise linear connecting $F^{\sim c}(y_j) = F^{\sim mid}(y_j)$.

13. Confidence interval for quantile function. Let Y be continuous. The parameter $\theta = Q(p)$ can be defined from $F(y)$ by $F(\theta) - p = 0$. An estimator $\hat{\theta}$ of θ is defined to satisfy

$$F^{\sim c}(\hat{\theta}) - p = 0;$$

therefore $\hat{\theta} = Q^{\sim c}(p)$. A confidence interval for θ can be obtained by defining a pivot $T^{\sim}(\theta)$, a function of θ and data, by

$$T^{\sim}(\theta) = \frac{F^{\sim c}(\hat{\theta}) - p}{(p(1-p)/n)^{.5}} = Z$$

where Z is Normal (0,1); we are using the asymptotic distribution of $T^{\sim}(\theta)$ when θ is the true parameter value. The parameter inverse pivot quantile function $\theta(u), 0 \leq u \leq 1$, is defined to satisfy

$$T^{\sim}(\theta(u)) = Q_Z(u);$$

explicitly

$$Q^{\sim c}(p; u) = \theta(u) = Q^{\sim c}(p + (p(1-p)/n)^{.5}Q_Z(u))$$

We claim: (1) conventional large sample $1 - \alpha$ confidence interval for $\theta = Q(p)$ can be expressed

$$\theta(\alpha/2) \leq \theta \leq \theta(1 - (\alpha/2));$$

(2) a $1 - \alpha$ significant test of hypothesis $\theta = \theta_0$ is rejected if $\theta_0 \leq \theta(\alpha)$ or $\theta_0 \geq \theta(1 - \alpha)$ depending on whether the alternative hypothesis is $\theta_0 \leq \theta$ or $\theta \leq \theta_0$; (3) point estimation of θ is $\theta(.5) = Q^{\sim c}(p)$. For extensions see Rosenkrantz (2000).

14. Quartiles, Median, Quartile, Location, Scale: Important summary of a quantile function $Q(u)$ are quartiles $Q1 = Q(.25)$, $Q3 = Q(.75)$, and median $Q2 = Q(.5)$. Nonparametric measures of location are $Q2$ and mid-quartile

$$MQ = .5(Q1 + Q3).$$

Measure of scale is interquartile range $IQR = Q3 - Q1$. We prefer as measure of scale twice the interquartile range:

$$IQR2 = 2(Q3 - Q1)$$

Measure of skewness is $(Q2 - MQ)/IQR2$; its absolute value is bounded by .25.

General measures of scale have form $\int_0^1 J_0(u)Q(u)du$ for suitable score functions $J_0(u)$ such as $J_0(u) = \Phi^{-1}(u)$ or $J_0(u) = u - .5$.

Shapiro Wilks statistic to test normality of a random variable Y is a sample version of squared correlation

$$\rho^2(Q(u), \Phi^{-1}(u)) = \frac{[\int_0^1 \Phi^{-1}(u)Q(u)du]^2}{\int_0^1 [Q(u) - \int_0^1 Q(s)ds]^2 du}$$

As a test statistic we recommend $\log \rho^2$ because it is compared with zero, and is an entropy difference statistic since it is the difference of two estimators of $\log \sigma^2$.

15. Sample $Q - Q$ Plot. A sample $Q - Q$ plot compares a sample with a continuous quantile $Q_0(u)$ representing a model by plotting quantile functions

$$(Q_0(F^{\sim mid}(y_j), y_j) = (Q_0(u_j^{mid}), Q^{\sim c}(u_j^{mid})))$$

where $y_1 < \dots < y_r$ are distinct values in sample and $u_j^{mid} = F^{\sim mid}(y_j)$. We believe these widely used plots are difficult to interpret. It helps to align by making the functions equal at $u = .25$ and $u = .75$. This is accomplished by plotting quantile-quantile functions

$$(Q_0/Q_0(u_j^{mid}), Q^{\sim c}/Q^{\sim c}(u_j^{mid})).$$

An idea for research is the concept of “confidence $Q - Q$ curves” to compare a model Q_0 with sample quantile Q^{\sim} of data Y . Lower confidence $Q - Q$ curve joins linearly

$$(Q_o(F^{\sim mid}(y_j)), Q^\wedge(F_0^{\sim mid}(y_j); \alpha/2))$$

Upper confidence $Q - Q$ curve joins linearly

$$(Q_o(F^{\sim mid}(y_j)), Q^\wedge(F_0^{\sim mid}(y_j); 1 - (\alpha/2)))$$

A test if the model Q_0 fits data $Q^{\sim c}$: does a line exist between lower and upper confidence curves? If the graph of $y = g(x)$ fits between the confidence curves, we conclude $Y \stackrel{d}{=} g(X)$ since $Q_Y(u) = g(Q_0(u))$, where X has quantile $Q_0(u)$. Our goal is to identify transformations of the data to normality or exponential. For positive random variable Y , hazard function $H(Y)$ has property $H(Y)$ is exponential.

16. Quantile/Quantile Function $Q/Q(u)$: We define quantile/quantile function $Q/Q(u)$ of quantile $Q(u)$

$$Q/Q(u) = \frac{Q(u) - .5(Q(.25) + Q(.75))}{2(Q(.75) - Q(.25))}$$

Verify that $Q/Q(.25) = -.25, Q/Q(.75) = .25$.

If $Q/Q(u) > 1$ or $Q/Q(u) < -1$, we call u a Tukey outlier since the value $y = Q(u)$ lies outside the fences as defined by John Tukey in his pioneering work on exploratory data analysis. Measure of skewness is $Q/Q(.5)$. Measures of tail behavior are $Q/Q(.05)$, $Q/Q(.95)$. The distribution of stock market prices follows a power law (long tail) and is not Gaussian (medium tail).

Table: Quantile/quartile diagnostics of tail.

Left tail	Q/Q diagnostic
Short	$-.5 < Q/Q(.05) < -.25$
Medium	$-1 < Q/Q(.05) < -.5$
Long	$Q/Q(.05) < -1$
Right tail	Q/Q diagnostic
Short	$.25 < Q/Q(.95) < .5$
Medium	$.5 < Q/Q(.95) < 1$
Long	$1 < Q/Q(.95)$

17. Folio of Q/Q Plots and Data Modeling. For data analysis one plots the sample quantile/quartile function $Q^{\sim c}/Q^{\sim c}(u)$. From this normalized graph one can identify the shape of probability models to fit to data. To compare the fit of a location scale model $Q(u) = u + \sigma Q_0(u)$ one plots on the same graph the sample quantile/quartile function and $Q_0/Q_0(u)$.

From the sample quantile/quartile function one can diagnose symmetry and tail behavior of data, and identify standard distribution which might fit the data, and diagnose goodness of fit of models to the data. The study of a folio of Q/Q plots would enable a statistician to identify distributions to fit to data, and identify distributions (especially Normal) that do NOT fit the data. An example is studied in an appendix.

18. Density quantile and quantile density functions. If F is continuous, $F(Q(u)) = u$ for all u . Taking derivatives

$$f(Q(u))Q'(u) = 1$$

Define density quantile function $fQ(u) = f(Q(u))$, quantile density function $q(u) = Q'(u)$, score function

$$J(u) = -(fQ(u))' = \frac{-f'(Q(u))}{f(Q(u))}$$

In practice we assume representation near 0 and 1 as regularly varying functions:

$$\begin{aligned} fQ(u) &= u^{\alpha_0} L(u) \\ fQ(1-u) &= u^{\alpha_1} L(u) \end{aligned}$$

where $L(u)$ is a slowly varying or log-like function satisfying for fixed $y > 0$

$$L(yu)/L(u) \rightarrow 1 \text{ as } u \rightarrow 0$$

An example of a slowly varying function is $L(u) = (-\log u)^\beta$.

We call α_0 and α_1 tail exponents; they are used to classify tail behavior as short ($\alpha < 1$), medium ($\alpha = 1$), or long ($\alpha > 1$). Concept of tail behavior is widely used by

statisticians to describe non-normal distributions; tail exponents provide rigorous concepts of tail behavior needed to debate the statistical question: can the ends (tail) be used to justify the means?

19. Asymptotic distribution of sample quantiles. When Y is continuous $U = F(Y)$ is Uniform (0,1) and $Y = Q(U)$. The sample quantile of Y can be represented

$$Q_{\tilde{Y}}(u) = Q_Y(Q_{\tilde{U}}(u)).$$

By delta method of large sample theory

$$n^{.5}(Q_{\tilde{Y}}(u) - Q(u)) - q_Y(u)n^{.5}(Q_{\tilde{U}}(u) - u) \xrightarrow{P} 0.$$

One can show

$$n^{.5}(Q_{\tilde{U}}(u) - u) \xrightarrow{d} B(u)$$

where $B(u)$, $0 \leq u \leq 1$, is a Brownian Bridge, a zero mean Gaussian process with covariance kernel $E[B(u_1)B(u_2)] = \min(u_1, u_2) - u_1u_2$. One can conclude that

$$n^{.5}f_Y Q_Y(u)(Q_{\tilde{Y}}(u) - Q(u)) \xrightarrow{d} B(u).$$

The parameters μ and σ in a location scale model $Q(u) = \mu + \sigma Q_0(u)$, $f_Y Q_Y(u) = \frac{1}{\sigma} f_0 Q_0(u)$, then satisfy approximately a regression model

$$f_0 Q_0(u) Q_{\tilde{Y}}(u) = \mu f_0 Q_0(u) + \sigma f_0 Q_0(u) Q_0(u) + \frac{\sigma}{\sqrt{n}} B(u).$$

Using reproducing kernel Hilbert space theory of continuous parameter regression one can derive asymptotically efficient estimators μ^\wedge and σ^\wedge which are linear combinations of order statistics. One can also solve data compression problems of selecting a small number of values u_1, \dots, u_k such that $Q_{\tilde{Y}}(u_1), \dots, Q_{\tilde{Y}}(u_k)$ have as much information for estimation and modeling as the whole quantile function.

20. Conditional quantile function. When observing (X, Y) the mean and variance approach to statistical reasoning emphasizes conditional mean $E[Y|X = x]$ and conditional variance, which are mean and variance of conditional distribution

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

Conditional quantile is defined

$$Q_{Y|X=x}(u) = F_{Y|X=x}^{-1}(u).$$

We call this formula a brute force approach to calculating conditional quantile. An alternative can be developed using the fact that conditional probability has properties analogous to the properties of probability. Therefore for $g(y)$ non-decreasing and continuous from the left

$$Q_{g(Y)|X=x}(u) = g(Q_{Y|X=x}(u)).$$

One can show that $F(Q(u)) = u$ if u is in the range of F , $Q(F(y)) = y$ if y is in the range of Q . A random variable Y is in the range of Q with probability one. Therefore we have:

Theorem: Powerful representation: $Y = Q_Y(F_Y(Y))$ with probability one.

Note that Y is equal in distribution to $Q(U)$ where U is Uniform $(0,1)$. When Y is discrete $F(Y)$ is not uniform; still $Y = Q(F(Y))$. The representation of Y as a transform of $F(Y)$ yields:

Theorem: Conditional quartile representation

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

where $s = Q_{F(Y)|X=x}(u)$. To compute s we write

$$\begin{aligned} u &= F_{F(Y)|X=x}(s) = P[F(Y) \leq s | X = x] \\ &= P[Y \leq Q_Y(s) | X = x] = F_{Y|X=x}(Q_Y(s)). \end{aligned}$$

The relation between u and s is a special case of the concept of comparison distribution.

21. Comparison distribution PP plots. A fundamental problem of statistics is comparison of two distributions F and G , and testing hypothesis $H_0 : F(y) = G(y)$.

If we let $u = G(y)$, $y = G^{-1}(u)$ we can express the hypothesis

$$H_0 : F(G^{-1}(u)) = u.$$

We can write $H_0 : D(u; G, F) = u$ where $D(u; G, F)$ is the comparison distribution function whose definition is given for (1) F, G both continuous, (2) F, G both discrete, (3) F discrete (data), G continuous (model). A comparison distribution is called a relative distribution by Handcock and Morris (1999).

When F and G are both continuous with probability densities $f(y)$ and $g(y)$, we assume also $F \ll G$, defined $g(y) = 0$ implies $f(y) = 0$. Then

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

satisfies $D(0) = 0, D(1) = 1$. Comparison density is defined

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

When F, G are discrete with probability mass functions $p_F(y)$ and $p_G(y)$, we assume $p_G(y) = 0$ implies $p_F(y) = 0$ and define first comparison density function

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

Comparison distribution is defined

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

Verify that $D(u)$ is piecewise linear between its values at $u_j = G(y_j)$, where $y_1 < \dots < y_r$ are probability mass points of G , and

$$D(u_j) = F(G^{-1}(u_j)) = F(y_j).$$

The graph of $D(u)$ joins $(G(y_j), F(y_j))$ and is called a *PP* plot.

22. Comparison Density Rejection Simulation. The graph of $d(u)$ provides a rejection method of simulation which generates a sample Y_1, \dots, Y_n from F as an acceptable subset of a sample X_1, \dots, X_m from G . We assume a bound c , $d(u) \leq c$ for all u . Generate independent Uniform (0,1) U_1 and U_2 . If $U_2 \leq d(U_1)/c$, accept $X = G^{-1}(U_1)$ as an observed value of Y . Otherwise reject X . The probability of acceptance is $1/c$. To prove the acceptance-rejection rule, verify that the area under $d(u)$ from 0 to $G(y)$ equals $D(G(y)) = F(y)$. The probability that $U_1 \leq G(y)$ and $U_2 \leq d(U_1)/c$ has probability $F(y)/c$. The event $Y \leq y$ can be shown to have probability $F(y)$.

23. Bayesian Theorem for Posterior Distributions: Parametric statistical inference assumes a probability model depending on a parameter θ to be estimated. Bayesian inference assumes a prior distribution for the parameter θ which is a probability mass function $p(\theta)$ if θ is discrete, and is a probability density $f(\theta)$ if θ is continuous. The model for Y given θ is a probability mass function $p(Y|\theta)$ if Y is discrete, and a probability density function $f(Y|\theta)$ if Y is continuous.

The posterior distribution of θ given data Y is described by $p(\theta|Y)$ or $f(Y|\theta)$. To compute it we apply Bayes' theorem, which we state as a 2×2 table which generalizes the basic statement of Bayes' theorem for events A and B :

$$P[A|B]/P(A) = P[B|A]/P[B]$$

	Y discrete	Y continuous
θ discrete	$\frac{p(\theta Y)}{p(\theta)} = \frac{p(Y \theta)}{p(Y)}$	$\frac{p(\theta Y)}{p(\theta)} = \frac{f(Y \theta)}{f(Y)}$
θ continuous	$\frac{f(\theta Y)}{f(\theta)} = \frac{p(Y \theta)}{p(\theta)}$	$\frac{f(\theta Y)}{f(\theta)} = \frac{f(Y \theta)}{f(Y)}$

24. Bayesian Inference Using Quantile Simulation: The most informative way to compute the posterior distribution is by the posterior quantile function $Q_{\theta|Y}(u)$ using

$$\begin{aligned} Q_{\theta|Y}(u) &= Q_{\theta}(s) \\ s &= D^{-1}(u; F_{\theta}, F_{\theta|Y}) \\ u &= D(s; F_{\theta}, F_{\theta|Y}) \end{aligned}$$

One can simulate a sample from the posterior distribution using a sample from the prior distribution using rejection simulation and a formula for the comparison density $d(s; F_{\theta}, F_{\theta|Y})$.

When θ and Y are both continuous

$$\begin{aligned} d(s) &= d(s; F_\theta, F_{\theta|Y}) = f_{\theta|Y}(Q_\theta(s))/f_\theta(Q_\theta(s)) \\ &= f_{Y|\theta=Q_\theta(s)}(Y)/f_Y(Y) \end{aligned}$$

Monte Carlo simulation chooses independent Uniform (0,1) S and U ; accept $\theta = Q_\theta(S)$ if

$$\frac{d(S)}{\max_s d(s)} = \frac{f_{Y|\theta=Q_\theta(S)}(y)}{\max_\theta f_{Y|\theta}(Y)} \geq U$$

One compares the likelihood of Y under $\theta = Q_\theta(s)$ with maximum likelihood of Y .

25. Bivariate dependence density and component correlations. To model and measure dependence of bivariate data (Y, X) general tools are dependence density (or copula density)

$$d_{Y,X}(s, t) = d(s; F_Y, F_{Y|X=Q_X(t)})$$

and component correlations

$$C_{Y,X}(j, k) = \int_0^1 \int_0^1 ds dt d_{Y,X}(s, t) \phi_{Y,j}(s) \phi_{X,k}(t)$$

for suitable orthonormal score functions. Note

$$\begin{aligned} \int_0^1 ds d_{Y,X}(s, t) \phi_{Y,j}(s) &= E[\phi_{Y,j}(F_Y(Y)) | X = Q_X(t)] \\ C_{Y,X}(j, k) &= E[\phi_{Y,j}(F_Y(Y)) \phi_{X,k}(F_X(X))]. \end{aligned}$$

One way to construct orthonormal score functions is

$$\phi_{Y,j}(s) = g_j(F_Y^{-1}(s))$$

where $g_j(u)$ are orthonormal functions of y .

Empirical component correlations, estimated from data, are

$$C_{Y,X}(j, k) = E^\sim[\phi_{Y,j}(F_Y^{\sim mid}(Y)) \phi_{X,k}(F_X^{\sim mid}(X))].$$

To estimate $d_{Y,X}(s, t)$ we recommend logistic regression to estimate it as a function of t for s fixed. Apply it as a function of s for fixed t to compute conditional quantile $Q_{Y|X=Q_X(t)}(u)$, $0 < u < 1$, by rejection simulation from unconditional quantile $Q_Y(s)$, $0 < s < 1$.

26. Appendix: Investment strategy with positive mean gain, negative median gain. Investors should be aware that a stock market trading strategy can result in a positive mean gain, but negative gains for most investors. Each week an investor invests in an IPO (initial public offering) and sells after a week with gain 80% with probability

.5, loss 60% with probability .5. Let Y denote profit after two trades (two weeks) with an initial investment of \$10,000;

$$\begin{aligned} Y &= 22,400 \text{ if both trades gain;} \\ &= -2800 \text{ if one trade gains, one trade loses;} \\ &= -8400 \text{ if both trades lose} \end{aligned}$$

Probability mass function and mid-distribution of Y are:

y	$p(y)$	$F^{mid}(y)$
-8400	1/4	1/8
-2800	1/2	1/2
22400	1/4	7/8

Average gain $E(Y) = 2100$. Median $Q_2 = -2800$. In other words, the strategy is “winning” since mean is positive but actually losing since the median is negative.

Quartiles are found by interpolation. $Q_1 = 6533$, $Q_3 = 21000$, Quantile/quartile analysis $MQ = 7233.5$, $IQR_2 = 55066$, $(MIN - MQ)/IQR_2 = .284$, $(MAX - MQ)/IQR_2 = .275$. These diagnostics indicate very short tails which occurs when we have bimodality (two groups of small observed values and large observed values).

27. Appendix: Exploratory Data Analysis Comparison of Two Samples: Is high indoor radon concentration related to cancer of children in home? To study this question radon concentration is measured in two types of houses: houses in which a child diagnosed with cancer has been residing, and houses with no recorded cases of childhood cancer. Counts and distribution function in samples of homes (data from Devore (2004, p. 43), Example 1.20) are computed.

Table lists summary quantiles of the two samples. The conclusions of our data analysis are as follows:

Compare location (means, medians) of two samples: Cancer houses radon has greater location parameter than do non-cancer houses radon. What to do about an extreme observation of 210 in cancer houses which inflates mean?

Compare scale: Interquartile range (preferred to standard deviation) indicate variability of radon in non-cancer homes is greater than variability of radon in cancer homes.

Side by side boxplots: Radon non-cancer homes has skew distribution, radon in cancer homes is symmetric. Non-cancer radon variability is greater than cancer radon variability.

Identification of probability laws: Non-cancer homes diagnostics indicate fit by exponential distribution. Cancer homes indicate fit by normal distribution with outliers.

Comparison of two samples: Most general way to compare distributions of radon in cancer homes and non-cancer homes is to plot comparison distribution or PP plot of

$$(F_{\text{radon|cancer}}(y_j), F_{\text{radon|no cancer}}(y_j))$$

evaluated at values y_j obtained by pooling the values in each sample. Intuitively we consider $F_{\text{radon|cancer}}(y)$ to be conditional distribution $F_{Y|X=x}(y)$ of $Y =$ radon concentration given $X =$ type of homes, cancer or no cancer. We recommend as the most general method that one plot

$$(F_{\text{radon}}(y_j), F_{\text{radon|no cancer}}(y_j))$$

where $F_{\text{radon}}(y)$ is the distribution of radon in the pooled sample.

Quantile/quartile Q/Q plots. Figure 1 plots on one graph $Q/Q(u)$ for exponential and normal distributions and sample distribution of radon in non-cancer homes. Our speculation that exponential fits data is strengthened by this plot of the Q/Q curves. Figure 2 plots on one graph $Q/Q(u)$ for exponential and normal distribution and sample distribution of radon in cancer homes. This plot of the Q/Q curves support our speculation that normal with outliers fit data but also suggests that for better fit we consider as a model a Weibull distribution. The dots on the sample Q/Q curve represent the distinct values y_j^Q in the sample plotted at $u_j = F^{\sim mid}(y_j)$; we define $y_j^Q = (y_j - MQ)/IQR2$. These values are connected linearly to form sample $Q/Q(u)$. Note that sample Q/Q plots always have dots at $(.25, -.25)$, $(.75, .25)$ and $(.5, Q/Q(.5))$ which diagnose skewness. We do not usually plot the quantile function $Q(u)$ because information about shape comes from $Q/Q(u)$.

Comparison distribution plots. To test the hypothesis that the two sample (radon in non-cancer homes and radon in cancer homes) have the same distribution, general methods are PP plots of the two sample distribution functions which estimate the comparison distribution $D(u; F_{\text{cancer}}, F_{\text{no cancer}})$. Figure 3 plots this curve. Figure 4 plots an estimate of $D(u; F_{\text{pooled sample}}, F_{\text{no cancer}})$. Studying the two plots shows why we believe the second graph may be more useful as well as able to be plotted in general. Both graphs are plotted at the distinct values in the pooled sample (which in this example is 32).

Table. Numerical Summary and Diagnostics Radon Concentration in Cancer, No Cancer Homes

	Cancer Houses	No Cancer Houses
Sample size n	42	40
Number of distinct value	26	19
Sample mean \bar{Y}	22.8	19.2
Sample SD	31.7	17.0
S/\sqrt{n}	4.8	2.7
Sample MIN	3	3
Sample MAX	210	85
Next to MIN	5	5
Next to MAX	57	55
$Q1$	10.5	8
$Q2$	16	11
$Q3$	22	26.5
MQ	16.25	17.25
$IQR2$	23	37
$\frac{Q2-MQ}{IQR2}$	-.01	.17
Conclusion	Symmetric	Skew
Upper Fence= $MQ + IQR2$	39.25	54.5
Upper Outliers	45,57,210	55,55,85
$\frac{MIN-MQ}{IQR2}$	-.576	-.385
Conclusion	Normal with Outliers	Exponential

References

- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*, Cambridge University Press.
- DeVore, Jay. (2004). *Probability and Statistics* (sixth edition), Brooks/Cole: Belmont, California.
- Galton, F. (1889). *Natural Inheritance*, Macmillan: London.
- Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930*, Wiley: New York.
- 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 (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* (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 (1994). From comparison density to two sample data analysis, *The Frontiers of Statistical Modeling: An Informational Approach*, ed. H. Bozdogan, Kluwers: Amsterdam. 39–56.
- 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.

Figure 1.

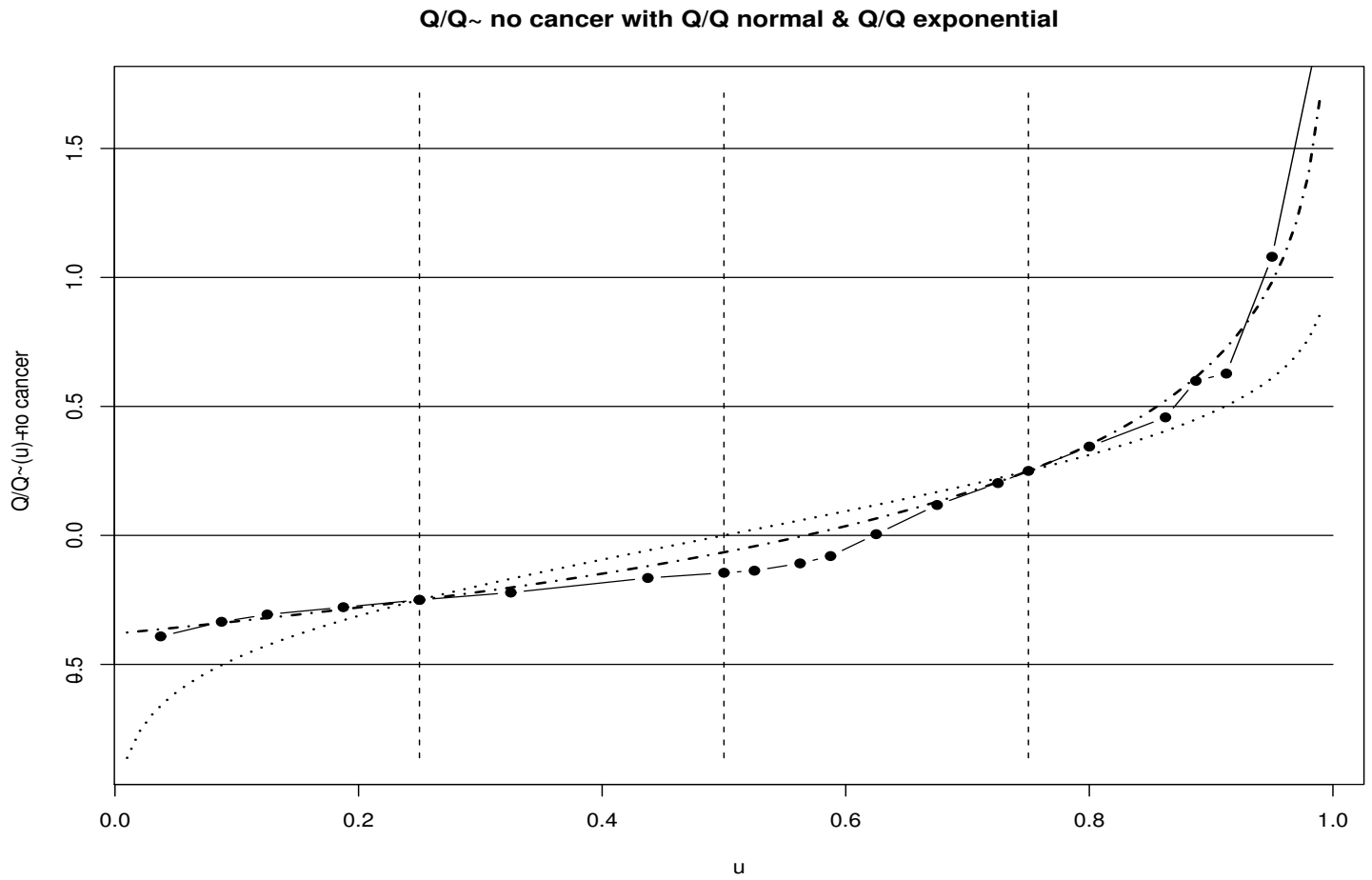


Figure 2.

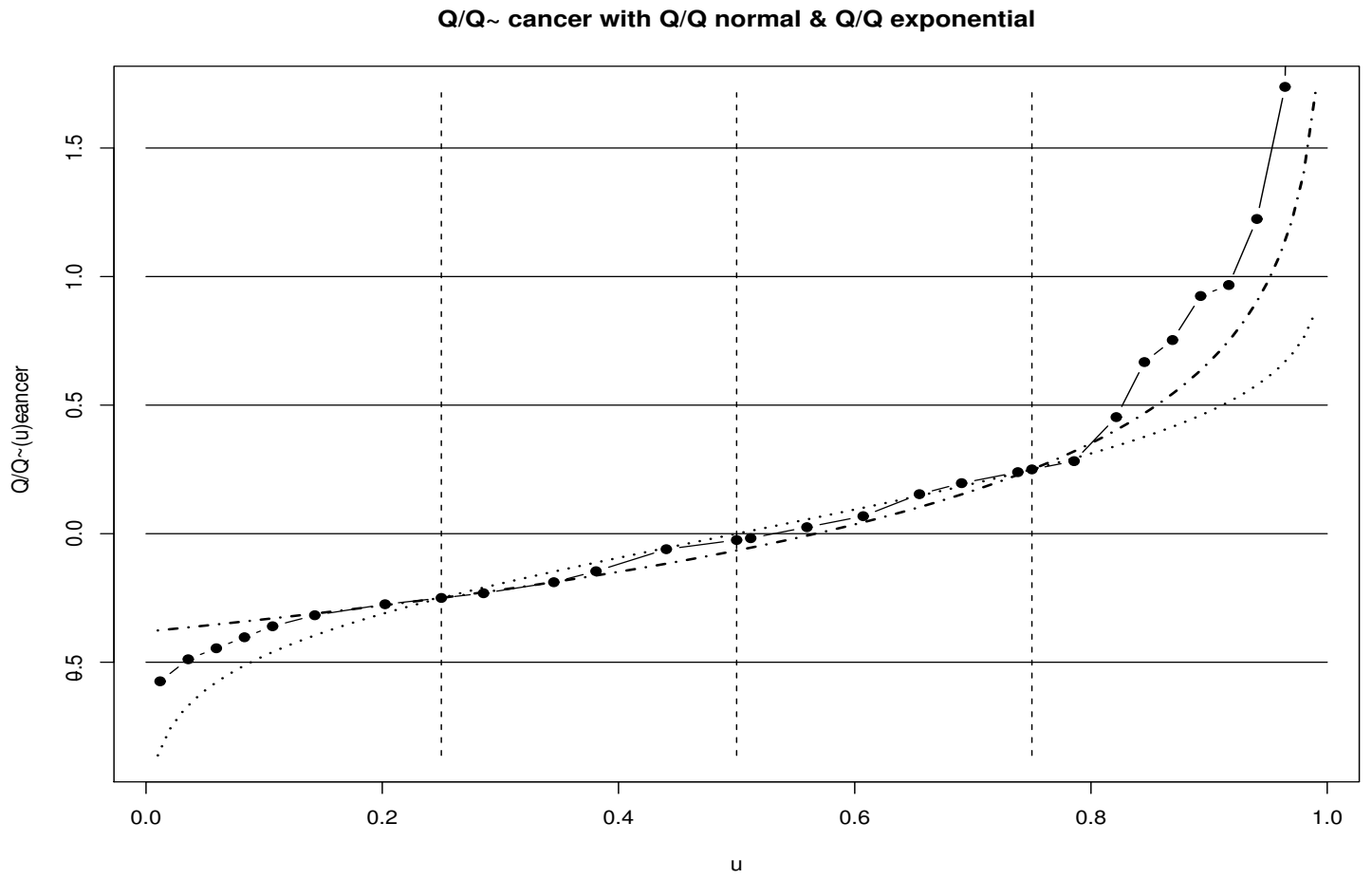


Figure 3.

$D_{\sim}(u, F_{\text{cancer}}, F_{\text{no cancer}})$

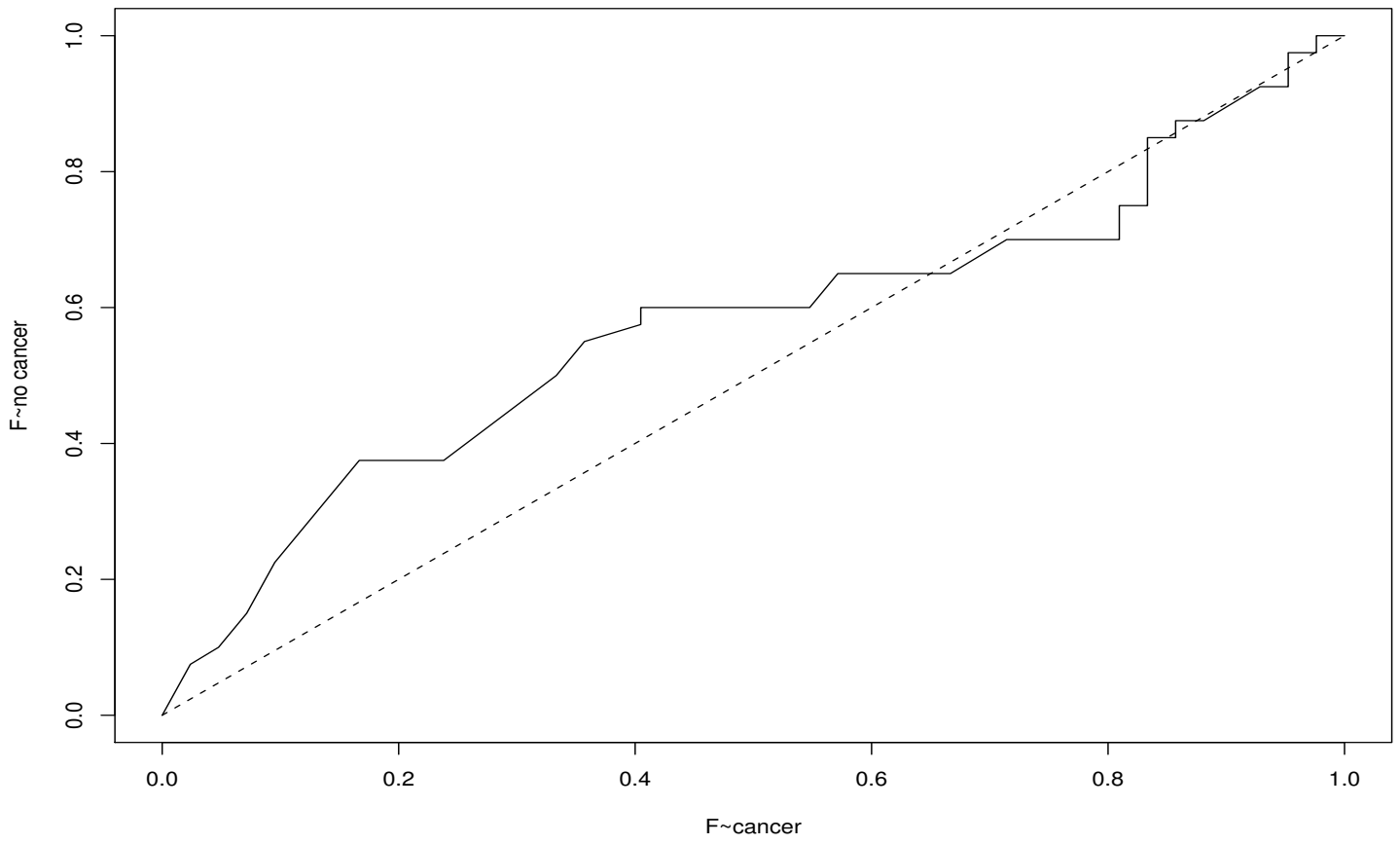


Figure 4.

$D_{\sim}(u, F \text{ pooled}, F \text{ no cancer})$

