QUANTILES

by

Randall L. Eubank

Technical Report No. SMU-DS-TR-185
Department of Statistics ONR Contract
May 1984

Research sponsored by the Office of Naval Research
Contract N00014-82-K-0207
Project NR 042-479

Department of Statistics
Southern Methodist University
Dallas, Texas 75275

Quantiles play a fundamental role in statistics although many times their use is disguised by notational and other artifices. They are the critical values we use in hypothesis testing and interval estimation and often are the characteristics of a distribution we wish most to estimate. Sample quantiles are utilized in numerous inferential settings and, recently, have received increased attention as a useful tool in data modeling.

Historically the use of sample quantiles in statistics dates back, at least, to Quetelet (1846) who considered the use of the semi-interquartile range as an estimator of the probable error for a distribution. Subsequent papers by Galton and Edgeworth (see eg. Galton (1889) and Edgeworth (1886, 1893) and references therein) discussed the use of other quantiles, such as the median, in various estimation settings. Sheppard (1899) and then Pearson (1920) studied the problem of optimal quantile selection for the estimation of the mean and standard deviation of the normal distribution by linear functions of subsets of the sample quantiles. Pearson's paper also contained most of the details involved in the derivation of the asymptotic distribution of a sample quantile. The large sample behaviour of a sample quantile was later investigated by Smirnoff (1935) who gave a rigorous derivation of its limiting distribution. Smirnoff's results were generalized in a landmark paper by Mosteller (1946) which, along with work

by Ogawa (1951), generated considerable interest in quantiles as estimation tools in location and scale parameter models. In more recent years quantiles have been utilized in a variety of problems of both classical and robust statistical inference and have played an important part in the work of Tukey (1977) and Parzen (1979a) on exploratory data analysis and nonparametric data modeling.

In this article the focus will be on the role of quantiles in various areas of statistics both as parameters of interest as well as means to other ends. We begin by defining the notion of population and sample quantiles.

Let F be a distribution function (d.f.) for a random variable X and define the associated quantile function (q.f.) by

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

Thus, for a fixed p in (0,1) the pth population quantile for X is Q(p). It follows from definition (1) that knowledge of Q is equivalent to knowledge of F. Further relationships between F and Q are

i) $FQ(u) \geq u$ with equality when F is continuous,

ii) $QF(x) \leq x$ with equality when F is continuous and strictly increasing, and

iii) $F(x) \geq u$ if and only if $Q(u) \leq x$.

Another important property of the q.f. which follows easily from iii) is that if U has a uniform distribution on [0,1] then

Q(U) and X have identical distributions. This fact provides one of the basic tools in many areas of statistical analysis. For example, in statistical simulation it has the consequences that a random sample from the uniform distribution may be used in conjunction with Q, to obtain a random sample from X.

The sample analog of Q is obtained by use of the empirical distribution function (e.d.f.). Let $X_{1:n}, X_{2:n}, \ldots, X_{n:n}$ denote the order statistics for a random sample of size n from a distribution F; then, the usual empirical estimator of F is

$$\tilde{F}(x) = \begin{cases} 0 \ , & x < X_{1:n} \ , \\ \frac{j}{n} \ , & X_{j:n} \leq x < X_{j+1:n}, \quad j = 1,\ldots,n-1 \ , \\ 1 \ , & x \geq X_{n:n} \ . \end{cases} \tag{2}$$

Replacing F with $\tilde{F}$ in (1) gives the sample or empirical quantile function (e.q.f.)

$$\tilde{Q}(u) = X_{j:n}, \ \frac{j-1}{n} < u \leq \frac{j}{n}, \ j = 1,\ldots,n. \tag{3}$$

Thus, the fundamental sample statistics $\tilde{Q}$, $\tilde{F}$ and the order statistics are all closely related. In fact, it is clear from (2) and (3) that knowledge of any one implies knowledge of the other two.

The previous discussions apply to both continuous and discrete random variables. However, in subsequent work it will be assumed that F is continuous and admits a density $f = F'$. In this case we also define the density-quantile function (d.q.f.) $fQ(u) = f(Q(u))$, $0 \leq u \leq 1$. Differentiation of $FQ(u) = u$ reveals that Q and fQ are related by

$Q'(u) = 1/fQ(u)$. Table 1 contains d.f.'s, q.f.'s, densities and d.q.f.'s for several common continuous distributions.

In the next two sections we study the properties of sample quantiles and their use in statistical inference. Subsequent sections then deal with the role of quantiles in exploratory data analysis as well as other areas of statistics.

Sample Quantiles: Asymptotic Properties and Nonparametric Inference

Sample quantiles provide nonparametric estimators of their population counterparts that are optimal in the sense that for any fixed p in (0,1) no other translation equivariate asymptotically median unbiased estimator is asymptotically more concentrated about $Q(p)$, than is $\tilde{Q}(p)$. This result is due to Pfanzagl (1975) who also shows that similar properties hold for tests about $Q(p)$ based on $\tilde{Q}(p)$.

There are several alternatives to $\tilde{Q}$ as defined in (3) that also have useful properties. This estimator duplicity stems, in part, from the discreteness of $\tilde{F}$ which entails that for $\tilde{F}(X_{j-1:n}) \leq p \leq \tilde{F}(X_{j:n})$ any value between $X_{j-1:n}$ and $X_{j:n}$ can, intuitively, act as the pth sample quantile. Thus, one could consider combining both $X_{j-1:n}$ and $X_{j:n}$ or, more generally, several order statistics in the neighborhood of $X_{j:n}$ to obtain an estimator of $Q(p)$. Such considerations have led to the usual definition for the sample median, which agrees

TABLE 1

DISTRIBUTION, QUANTILE, DENSITY AND DENSITY-QUANTILE FUNCTIONS FOR SELECTED PROBABILITY LAWS

PROBABILITY LAW	DISTRIBUTION FUNCTION	QUANTILE FUNCTION	DENSITY FUNCTION	DENSITY-QUANTILE FUNCTION
Normal	$\Phi(x) = \int_{-\infty}^{x}\phi(x)dx$	$\Phi^{-1}(u)$	$\phi(x)=(2\pi)^{-\frac{1}{2}}e^{-x^2/2}$	$\phi\Phi^{-1}(u)=(2\pi)^{-\frac{1}{2}}e^{-\lvert\Phi^{-1}(u)\rvert^2/2}$
Lognormal	$\Phi(\log x)$	$e^{\Phi^{-1}(u)}$	$\frac{1}{x}\phi(\log x)$	$\phi\Phi^{-1}(u)e^{-\Phi^{-1}(u)}$
Exponential	$1-e^{-x}$, $x>0$	$-\log(1-u)$	e^{-x}	$1-u$
Weibull	$1-e^{x^c}$, $c,x>0$	$c\{-\log(1-u)\}^{1-\frac{1}{c}}$	$cx^{c-1}e^{-x^c}$	$c(1-u)\{-\log(1-u)\}^{1-\frac{1}{c}}$
Extreme Value	$e^{-e^{-x}}$	$-\log\log(\frac{1}{u})$	$e^{-x}e^{-e^{-x}}$	$-u\log(u)$
Logistic	$\{1+e^{-x}\}^{-1}$	$\log(\frac{u}{1-u})$	$e^{-x}(1+e^{-x})^{-2}$	$u(1-u)$
Pareto	$1-(1+x)^{-\nu}$, $\nu,x>0$	$(1-u)^{-1/\nu}-1$	$\nu(1+x)^{-(\nu+1)}$	$\nu(1-u)^{1+\frac{1}{\nu}}$
Cauchy	$.5+\pi^{-1}\arctan(x)$	$\tan[\pi(u-.5)]$	$[\pi(1+x^2)]^{-1}$	$\pi^{-1}\sin^2(\pi u)$
Uniform	x, $0 \leq x \leq 1$	u	1	1
Reciprocal of a uniform	$1-\frac{1}{x+1}$, $x>0$	$(1-u)^{-1}-1$	$(x+1)^{-2}$	$(1-u)^2$
Double Exponential	$\begin{cases}\frac{1}{2}e^{x}, & x<0\\ 1-\frac{1}{2}e^{-x}, & x>0\end{cases}$	$\begin{cases}\log 2u, & u<.5\\ -\log 2(1-u), & u>.5\end{cases}$	$\frac{1}{2}e^{-\lvert x\rvert}$	$\begin{cases}u, & u<.5\\ 1-u, & u>.5\end{cases}$

with $\tilde{Q}(.5)$ only when n is odd, and have prompted several authors to propose linearized versions of $\tilde{Q}$ (see e.g. Parzen (1979a)). Estimators of Q(p) that utilize local smoothing of the order statistics near $X_{j:n}$ and appear to have good small sample properties have been suggested by Kaigh and Lachenbruch (1982), Kaigh (1983) and Harrell and Davis (1982). Reiss (1980) has considered the use of quasiquantiles and shown them to be superior to sample quantiles when compared on the basis of deficiency rather than efficiency.

For $0 < p < 1$ it is well known (c.f. Serfling (1980)) that, for fQ positive and continuous near p, $\tilde{Q}(p)$ is asymptotically normally distributed with mean Q(p) and variance $p(1-p)/nfQ(p)^2$. An extension of this result to k quantiles, for fixed $k \geq 1$, can be found in Mosteller (1946) and Walker (1968) with the case of k growing with n treated by Ikeda and Matsunawa (1972). Necessary and sufficient conditions for the existence of moments for sample quantiles and for the convergence of these moments to those of the limiting distribution are provided by Bickel (1967). For a discussion of the asymptotic properties of $\tilde{Q}(p)$ for certain types of dependent samples see Sen (1972) and Babu and Singh (1978).

It follows from the asymptotic distribution of $\tilde{Q}(p)$ that an asymptotic $100(1-\alpha)\%$ confidence interval for Q(p) is given by $\tilde{Q}(p) \pm \Phi^{-1}(\alpha/2)\sqrt{p(1-p)/nfQ(p)^2}$ which, unfortunately,

requires knowledge of fQ(p). This difficulty can be resolved by using, instead, the interval $(\tilde{Q}(k_1/n), \tilde{Q}(k_2/n))$ where k_1 and k_2 are integers chosen so that $k_1 \simeq np - \Phi^{-1}(\alpha/2)\sqrt{np(1-p)}$ and $k_2 \simeq np + \Phi^{-1}(\alpha/2)\sqrt{np(1-p)}$. This latter interval is asymptotically equivalent to the former but utilizes the asymptotic relationship between $\tilde{Q}(k_i/n)$ and $\tilde{Q}(p)$ to estimate fQ(p) (c.f. Serfling (1980, pg. 103)). An alternative, but similar, approach is given in Walker (1968). For an exact confidence interval based on order statistics see Wilks (1962, pg. 329). Interval estimates obtained by bootstrapping and jackknifing have been proposed by Harrel and Davis (1982) and Kaigh (1983).

In testing hypotheses about Q(p) the most widely known procedure is probably the quantile test. This test is based on the fact that if $H_0: Q(p) = Q_0(p)$ is true then the number of sample quantiles below or equal to $Q_0(p)$ will be binomial with parameters np and np(1-p). As a result, the binomial distribution may be utilized to obtain an exact test, or the normal approximation to the binomial for an approximate test, of H_0. The quantile test, as well as several other tests concerning the median, can be found in standard texts such as Conover (1971).

From a data modeling perspective what is of interest is not Q(p) for some particular p, but rather the entire function Q(·), as its knowledge is equivalent to knowing the data's underlying probability law. Thus, we now consider the construction

of nonparametric estimators, $\hat{Q}(\cdot)$, that are random functions or stochastic processes on (0,1) (this is the quantile domain analog of nonparametric probability distribution and density estimation). The natural estimator of $Q(\cdot)$ is $\tilde{Q}(\cdot)$ whose asymptotic distribution theory, when considered as a stochastic process, has been studied by Shorack (1972), Csörgő and Révész (1981), Csörgő (1983), Mason (1984) and others. From their work it follows that when fQ is positive and differentiable on [0,1] and satisfies certain other regularity conditions near 0 and 1, $\sqrt{n}\, fQ(u)\{\tilde{Q}(u)-Q(u)\}$ converges in distribution to a Brownian bridge process on (0,1), i.e., a zero mean normal process with covariance kernel $K(u,v) = u-uv$, $u \leq v$ (analogous results for a linearized version of $\tilde{Q}$ and for the case of randomly censored data can be found in Bickel (1967), Sander (1975) and Csörgő (1983)). Tests in this setting are of the goodness-of-fit variety. The asymptotic distribution of many classical statistics, such as $\sup_{0<u<1} f_0 Q_0(u)|\tilde{Q}(u)-Q_0(u)|$, are available under the null hypothesis $H_0: Q(\cdot) = Q_0(\cdot)$, for specified Q_0, from Csörgő and Révész (1981, pg. 171) and Csörgő (1983, Chap. 7). Consequently, such statistics can be utilized to conduct quantile based goodness-of-fit tests. Another goodness-of-fit procedure that can be naturally formulated in the quantile domain is the Shapiro-Wilk test for normality (see Csörgő and Révész (1981, pgs. 202-212) and Csörgő (1983)).

Several procedures are available for constructing smooth estimators of Q formed from suitably rich function classes. The Tukey lambda distribution and its generalizations as well as g-and-h distributions are examples of curves derived specifically for this purpose (see LAMBDA and g- AND- h DISTRIBUTIONS). Other techniques, developed by Parzen (1979a), utilize certain analogies with time series analysis to provide estimators for fQ and Q as well as goodness-of-fit tests.

Another important asymptotic result is the Bahadur representation for sample quantiles which describes the relationship between the $\tilde{F}$ and $\tilde{Q}$ processes. One statement of this result is that, for fQ positive and differentiable at p, with probability one

$$n^{1/2}(\tilde{Q}(p)-Q(p)) = \frac{n^{1/2}(p-\tilde{F}Q(p))}{fQ(p)} + O(n^{-1/4}(\log n)^{3/4}).$$

This has the immediate consequence that $n^{1/2}(\tilde{Q}(p)-Q(p))$ and $n^{1/2}(p-\tilde{F}Q(p))/fQ(p)$ have identical asymptotic distributions. The Bahadur representation may also be used to obtain a law of the iterated logarithm for sample quantiles, namely, with probability one

$$\overline{\lim_{n\to\infty}} \pm \frac{n^{1/2}[\tilde{Q}(p)-Q(p)]}{(2 \operatorname{loglog} n)^{1/2}} = \frac{[p(1-p)]^{1/2}}{fQ(p)} .$$

This problem was first considered by Bahadur (1966). More general results and references may be found in Kiefer (1970), Csörgő and Révész (1981) and Csörgő (1983). The case of

ϕ-mixing random variables is treated by Sen (1972) and Babu and Singh (1978).

Parameter Estimation

In this section we consider the use of quantile based estimators in parametric models of the form $F(x)=F_0(x;\underline{\theta})$, where F_0 is a known distributional form and $\underline{\theta}$ is a vector of unknown parameters. An important special case of this model, that we will focus on initially, is the location and scale parameter model where $F(x) = F_0(\frac{x-\mu}{\sigma})$ for μ and σ unknown location and scale parameters. In this instance it is readily seen that $Q(u) = \mu + \sigma Q_0(u)$, where Q_0 is the q.f. for F_0.

The problem of location parameter estimation for symmetric distributions has been a subject of extensive study. Several quick estimators of μ that are useful for data from symmetric distributions are based on symmetric quantile averages of the form $\tilde{\mu}(p) = [\tilde{Q}(p) + \tilde{Q}(1-p)]/2$, $0 < p \leq .5$. One example is the Tukey trimean $\{\tilde{\mu}(.5) + \tilde{\mu}(1/4)\}/2$ while another is the estimator suggested by Gastwrith (1966), $.4\tilde{\mu}(.5) + .6\tilde{\mu}(1/3)$, which was found to be nearly 80% as efficient (asymptotically) as the best estimators for the Cauchy, double exponential, logistic and normal distributions. For references and a general discussion of the robustness and efficiency properties of symmetric quantile averages see Brown (1981).

General classes of estimators for μ are also conveniently

(and usefully) formulated in the quantile domain. For example, if ψ is an odd function, an M-estimator of μ is a solution to $\int_0^1 \psi(\tilde{Q}(u)-\hat{\mu})du = 0$. Similarly, an R-estimator satisfies $\int_0^1 J[u-F(2\hat{\mu}-\tilde{Q}(u))]du = 0$, with J an odd function on [-1,1], whereas an L-estimator can be written explicitly as $\int_0^1 h(\tilde{Q}(u))dM(u)$ for some function h and some signed measure, M, on (0,1). See Huber (1981) and Fernholz (1983) for further discussion of these estimators.

Asymptotically efficient quantile based estimators of both μ and σ that are applicable to general F_0 (not necessarily symmetric) have been given by Parzen (1979a,b). Using results from the previous section it can be seen that, asymptotically, location and scale parameter estimation can be considered as a regression analysis problem for the quantile process via the model

$$f_0Q_0(u)\tilde{Q}(u)=\mu f_0Q_0(u)+\sigma f_0Q_0(u)Q_0(u)+\sigma_B B(u),\ u\in[0,1], \qquad (4)$$

where $\sigma_B = \sigma/\sqrt{n}$ and $B(\cdot)$ is a Brownian bridge process. Under appropriate restrictions on f_0Q_0 and the product $f_0Q_0 \cdot Q_0$, continuous time regression techniques can be utilized to obtain asymptotically efficient estimators

$$\begin{bmatrix} \hat{\mu} \\ \hat{\sigma} \end{bmatrix} = A^{-1} \begin{bmatrix} \int_0^1 W_\mu(u)\tilde{Q}(u)du \\ \int_0^1 W_\sigma(u)\tilde{Q}(u)du \end{bmatrix}, \qquad (5)$$

where A is the usual Fisher information matrix, for μ and σ with $W_\mu(u) = -(f_0Q_0)''(u)f_0Q_0(u)$ and $W_\sigma(u) = -[f_0Q_0(u)Q_0(u)]''f_0Q_0(u)$.

Many estimators based on quantiles and order statistics have strong ties to model (4) and the estimators in (5). For instance, L-estimation of location and scale (see L-STATISTICS) can be motivated from (5) through consideration of alternative weight functions in place of W_μ and W_σ. By using an analog of model (4) that holds for left and right censored data, estimators similar to those given by Weiss (1964) and Weiss and Wolfowitz (1970) can be obtained. Through sampling from model (4) at a set of $k < n$ points $U = \{u_1,\ldots,u_k\}$ which satisfy $0 < u_1 < \ldots < u_k < 1$, "observations" $f_0Q_0(u_i)\tilde{Q}(u_i)$ can be obtained that, asymptotically have means $\mu f_0Q_0(u_i) + \sigma f_0Q_0(u_i)Q_0(u_i)$, $i=1,\ldots,k$, and variance-covariance matrix consisting of the elements $\sigma_B^2 u_i(1-u_j)$, $i \leq j$, $i,j=1,\ldots,k$. Thus, generalized least squares may be utilized to obtain asymptotically best linear unbiased estimators of μ and σ. Since their derivation by Ogawa (1951) an extensive literature has developed on these latter estimators and the associated problem of optimal selection for the spacing, U(see OPTIMAL SPACING PROBLEMS).

The estimation of a particular quantile, Q(p) say, is often of interest in parametric settings such as the location and scale parameter model. As $Q(p) = \mu + \sigma Q_0(p)$, we see that to estimate Q(p) for this model it suffices to estimate μ and σ. This may be accomplished, for instance,

using the estimators in (5) or maximum likelihood estimators. Alternatives that have good asymptotic efficiency properties and provide computational savings by using appropriate subsets of the sample quantiles have been suggested by Kubat and Epstein (1980), Eubank (1981) and Koutrouvelis (1981). Estimators for extreme quantiles have been studied by Weissman (1978) and Boos (1984).

For the estimation of a parameter vector $\underline{\theta}$, not necessarily of the location/scale variety, LaRiccia (1982) has proposed a minimum quantile distance approach based on the distance measure

$$D(\underline{\theta}) = \int_0^1 W(u;\underline{\theta})[\tilde{Q}(u)-Q_0(u;\underline{\theta})]^2 du \ , \qquad (6)$$

where $W(u;\underline{\theta})$ is some specified weight function and $Q_0(u;\underline{\theta})$ is the quantile function for $F_0(x;\underline{\theta})$. Under certain restrictions, the estimator obtained by minimizing (6) as a function of $\underline{\theta}$ is asymptotically normal. An optimal weight function has also been provided for single parameter situations that, in the special case of location or scale parameter estimation, results in the estimator obtained by Parzen from model (4). Unlike minimum distance procedures based on F_n (c.f. Parr and Schucany (1980)) the robustness properties of minimum quantile distance estimators, such as those obtained from (6), have as yet to be extensively investigated. Nevertheless, this approach seems promising and is intuitively appealing since quantile based methods are closely related

to regression techniques (as exemplified by model (4)) and are directly related to various data oriented diagnostics such as Q-Q plots (see GRAPHICAL REPRESENTATION OF DATA). For the extension of these estimators to randomly censored data see Eubank and LaRiccia (1984).

Descriptive Statistics and Exploratory Data Analysis

Many of the diagnostic measures and tabular summaries utilized in descriptive and exploratory data analysis (EDA) can be conveniently formulated in terms of sample quantiles. For example, the 5, 7 and 9 number data summaries that are a basic tool in EDA are all, essentially, collections of symmetrically chosen sample quantiles. This point is illustrated by the 5-number summary proposed by Tukey (1977) which (for large n) is equivalent to the use of $\tilde{Q}(.5)$ (the median), $\tilde{Q}(.25)$ and $\tilde{Q}(.75)$ (the quartiles), and the extremes $\tilde{Q}(\frac{1}{n+1})$ and $\tilde{Q}(\frac{n}{n+1})$. Similarly, a 7-number summary suggested by Parzen (1979a) consists of the median and quartiles as well as the eighths, $\tilde{Q}(.125)$ and $\tilde{Q}(.875)$, and the sixteenths, $\tilde{Q}(.0625)$ and $\tilde{Q}(.9375)$. Such data summaries are frequently utilized to obtain a transformation which gives a data set an approximately symmetric or normal distribution (see Tukey (1977) and Emerson and Stoto (1982)). The transformation is first applied to the summary and, using various diagnostic measures, checked for the desired properties. If the diagnostics indicate the transformation is satisfactory,

it is then applied to the entire data set. Thus, in this case, models for the data are developed by modeling $\tilde{Q}$.

Symmetric quantile averages are frequently utilized with data summaries to provide measures of centrality as well as diagnostics. Familiar examples are the median, $\tilde{\mu}(.5)$, and the midrange, $\tilde{\mu}(1/n+1)$. Measures of spread are often constructed from the midspreads $\tilde{\sigma}(p) = \tilde{Q}(1-p)-\tilde{Q}(p)$, $0<p<.5$, as exemplified by the sample range, $\tilde{\sigma}(1/n+1)$ and the interquartile range $\tilde{\sigma}(.25)$. When the data is approximately normal $\tilde{\sigma}(p)/\{\Phi^{-1}(1-p)-\Phi^{-1}(p)\}$ provides an estimator of the population standard deviation, A special case of this is the pseudo-standard deviation, $\tilde{\sigma}(.25)/1.35$, discussed by Koopmans (1981, pg. 63). Various diagnostic measures based on $\tilde{\mu}(p)$ and $\tilde{\sigma}(p)$, including measures of skewness and tail length, may be found in Parzen (1979a, Section 11).

A useful graphical tool proposed by Parzen (1979a,b) is the quantile box plot. This plot is a graph of a linearized version of Q upon which p boxes have been superimposed with coordinates $(p,\tilde{Q}(p))$, $(p,\tilde{Q}(1-p))$, $(1-p,\tilde{Q}(p))$ and $(1-p,\tilde{Q}(1-p))$ for p=1/4, 1/8 and 1/16. A horizontal line is drawn across the quartile box (p=1/4) at the median $\tilde{Q}(.5)$ to aid in the visual assessment of symmetry. A vertical line with length $\tilde{\sigma}(.25)/\sqrt{n}$ is also frequently placed at the median to provide an approximate confidence interval for Q(.5) as well as an indication of the size of the data set from which the plot derives. When examining the plot one looks for sharp rises

(infinite slopes) which, when occuring inside the quartile box, indicate the presence of two or more modes and suggest the presence of outliers otherwise. Flat intervals (0 slopes) correspond to probability masses and, consequently, are indicative of discrete random variables.

An illustration of a quantile box plot is provided in Figure 1 for the Rayleigh data (Tukey (1977, p. 49)) that consists of 15 weights of standard volumes of nitrogen obtained from air and other sources. The sharp rise in $\tilde{Q}$ indicates possible bimodality. It was Rayleigh's recognition of this characteristic which led him to the discovery of argon.

Conditional Quantiles

When the relationship between two random variables, X and Y, is being studied it is frequently of interest to estimate the conditional quantiles of Y for a given value or values of X. An important special case of this problem occurs when X and Y satisfy a linear model. In this case one possible definition of an empirical quantile function has been given by Bassett and Koenker (1982). Procedures for inference about a conditional quantile, assuming a linear model, have been developed by Steinhorst and Bowden (1971) and Kabe (1976), under the assumption of normal errors. An alternative nonparametric approach has been suggested by Hogg (1975) that allows the error distribution

to depend on the independent variable. A parametric alternative to Hogg's procedure is provided by Griffiths and Willcox (1978). For the estimation of conditional quantiles, in general, consistent estimators have been derived by Stone (1977) under very mild conditions. An alternative is suggested by Parzen in the discussion of Stone's paper. In addition, if the conditional quantile function can be assumed monotone in X, strongly consistent estimators presented in Casady and Cryer (1976) can be utilized.

REFERENCES

Babu, J. G. and Singh, K. (1978). J. Multivariate Anal. 8, 532-549.

Bahadur, R. R. (1966). Ann. Math. Statist. 37, 577-580.

Bassett, G. and Koenker, R. (1982). J. Amer. Statist. Assoc. 77, 407-415.

Bickel, P. J. (1967). Proc. Fifth Berkeley Symp. Math. Statist. Prob. I, 575-591.

Boos, D. D. (1984). Technometrics 26, 33-39.

Brown, B. M. (1981). Biometrika 68, 235-242.

Casady, R. J. and Cryer, J. D. (1976). Ann. Statist. 4, 532-541.

Conover, W. J. (1971). Practical Nonparametric Statistics, New York: John Wiley.

Csörgő, M. (1983). Quantile Processes with Statistical Applications, Philadelphia, SIAM.

Csörgő, M. and Révész, P. (1981). Strong Approximations in Probability and Statistics, New York: Academic Press.

Edgeworth, F. Y. (1886). Philos. Mag. 22, 371-383.

Edgeworth, F. Y. (1893). Philos. Mag. 36, 98-111.

Emerson, J. D. and Stoto, M.A. (1982). J. Amer. Statist. Assoc. 77, 103-108.

Eubank, R. L. (1981). Ann. Statist. 9, 494-500.

Eubank, R. L. and LaRiccia, V. N. (1984). J. Multivariate Anal. 14, to appear.

Fernholz, L. T. (1983). von Mises Calculus for Differentiable Statistical Functionals, Lecture Notes in Statistics, No. 19, New York: Springer-Verlag.

Galton, F. (1889). Natural Inheritance, Macmillan and Co.: New York.

Gastwirth, J. L. (1966). J. Amer. Statist. Assoc. 61, 929-948.

Griffiths, D. and Willcox, M. (1978). J. Amer. Statist. Assoc. 73, 496-498.

Harrel, F. E. and Davis, D. E. (1982). Biometrika 69, 635-640.

Hogg, R. V.(1975). J. Amer. Statist. Assoc. 70, 56-59.

Huber, P. J. (1981). Robust Statistics, New York: John Wiley.

Ikeda, S. and Matsunawa, T. (1972). Ann. Inst. Statist. Math. 24, 33-52.

Kabe, D. G. (1976). J. Amer. Statist. Assoc. 71, 417-419.

Kaigh, W. D. (1983). Commun. Statist.-Theor. Meth. A12(21), 2427-2443.

Kaigh, W. D. and Lachenbruch, P. A. (1982). Commun. Statist.-Theor. Meth. A11(19), 2217-2238.

Kiefer, J. (1970). In: Proc. Conference on Nonparametric Techniques in Statistical Inference, Puri, M.L. (Ed.), 349-357.

Koopmans, L. H. (1981). An Introduction to Contemporary Statistics, Boston: Duxbury.

Koutrouvelis, I. A. (1981). Commun. Statist.-Theor. Meth. A10(2), 189-201.

Kubat, P. and Epstein, B. (1980). Technometrics 4, 575-581.

LaRiccia, V. N. (1982). Ann. Statist. 10, 621-624.

Mason, D. M. (1984). Ann. Prob. 12, 243-255.

Mosteller, F. (1946). Ann. Math. Statist. 17, 377-408.

Ogawa, J. (1951). Osaka Math. J. 3, 175-213.

Parr, W. C. and Schucany, W. R. (1980). J. Amer. Statist. Assoc. 75, 616-624.

Parzen, E. (1979a). J. Amer. Statist. Assoc. 74, 105-121.

Parzen, E. (1979b). In: Robustness in Statistics, Launer, R. and Wilkinson, G. (Ed.), 237-258. New York: Academic Press.

Pearson, K. (1920). Biometrika 13, 113-132.

Pfanzagl, J. (1975). In: Statistical Methods in Biometry, Ziegler, W. J. (Ed.) 111-126. Basel: Birkhauser Verlag.

Quetelet, A. (1846). Lettres à S.A.R. ie Duc Régnant de Saxe-Cobourg et Gotha, sur la Theorie des Probabilitiés Appliquée aux Sciences Morales et Politiques. M. Hayes, Bruxelles.

Reiss, R. D. (1980). Ann. Statist. 8, 87-105.

Sander, J. M. (1975). Tech. Rep. No. 11, Stanford Univ.

Sen, P. K. (1972). J. Multivariate Anal. 2, 77-95.

Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics, New York: John Wiley.

Sheppard, W. F. (1899). Philos. Trans. Roy. Soc. London Ser. A 192, 101-167.

Shorack, G. R. (1972). Ann. Math. Statist. 43, 1400-1411.

Smirnoff, N. V. (1935). Metron 12, 59-81.

Steinhorst, R. K. and Bowden, D. C. (1971). J. Amer. Statist. Assoc. 66, 851-854.

Stone, C. J. (1977). Ann. Statist. 5, 595-645.

Tukey, J. W. (1977). Exploratory Data Analysis, Addison-Wesley: Reading, Mass.

Walker, A. M. (1968). J. Roy. Statist. Soc. Ser. B. 30, 570-575.

Weiss, L. (1964). Naval Res. Logist. Quart. 11, 125-134.

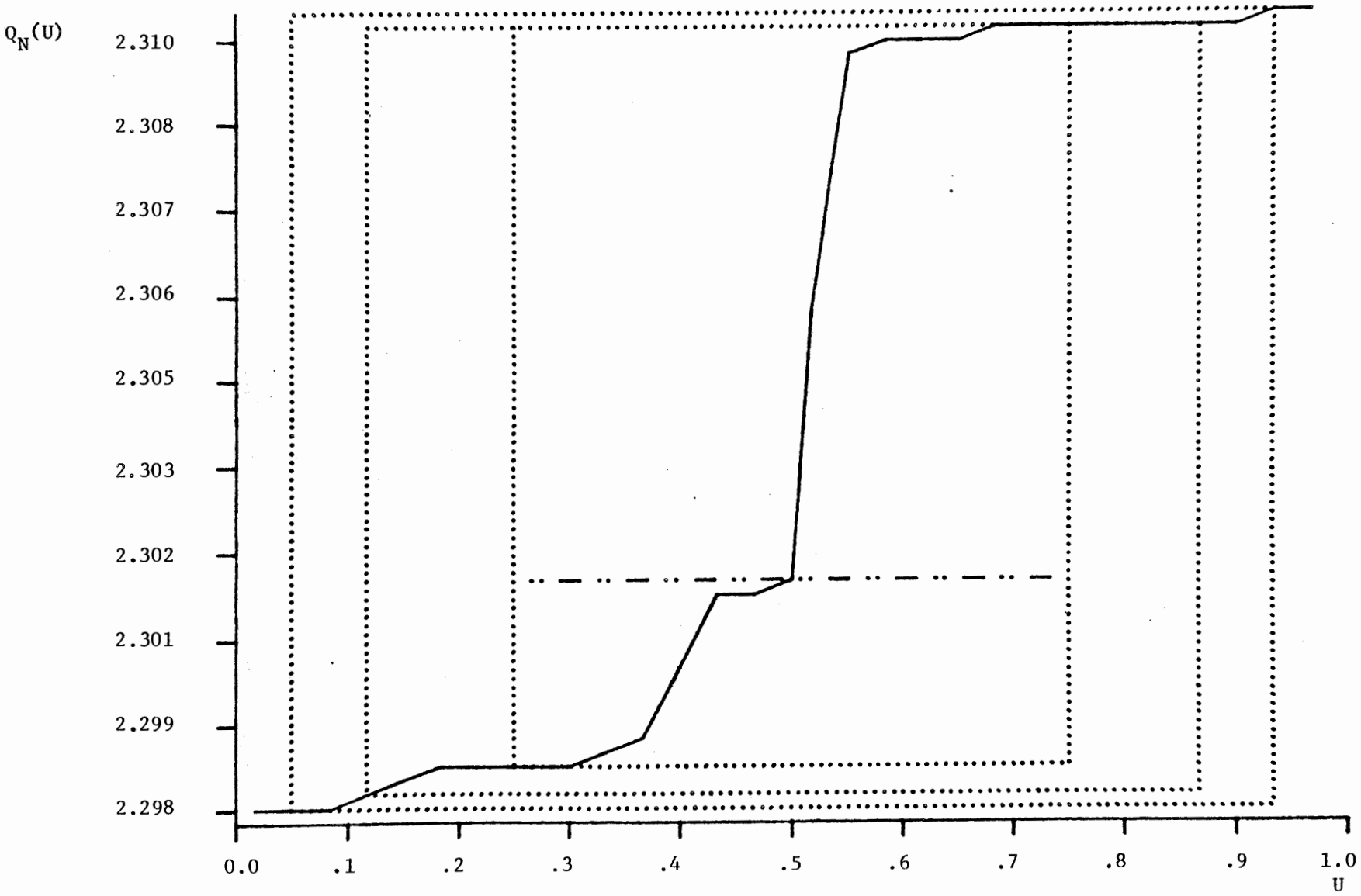
Weiss, L. and Wolfowitz, J. (1970). Z. Wahrsch. Verw. Gebiete 16, 134-150.

Weissman, I. (1978). J. Amer. Statist. Assoc. 73, 812-815.

Wilks, S. S. (1962). Mathematical Statistics, New York: John Wiley.

Research sponsored by Office of Naval Research contract N00014-82-K-0209.

FIGURE 1. QUANTILE BOX PLOT FOR THE RAYLEIGH DATA



UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER SMU-DS-TR185	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)*		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Randall L. Eubank		8. CONTRACT OR GRANT NUMBER(s) N00014-82-K-0209.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics Southern Methodist University Dallas, Texas 75275		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Arlington, VA. 22217		12. REPORT DATE May 1984
		13. NUMBER OF PAGES 31
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)*
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted for any putpose of The United States Government.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Conditional quantiles, density-quantile function, exploratory data analysis, L-estimator, location and scale parameters, M-estimators, minimum quantile distance, R-estimators, quantile function, sample quantiles

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

An overview is given of the role of quantiles in probability and statistics. The asymptotic properties of sample quantiles and their use in data modeling and nonparametric and parametric inference is considered.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*