Quantile Estimation by Orthogonal Polynomials

Bo Yang

A Thesis
in
The Department
of
Mathematics and Statistics

Presented in Partial Fulfilment of the Requirements
for the Degree of Master of Science at
Concordia University
Montreal, Quebec, Canada

January, 1994

ISBN 0-315-97662-4

Abstract

Quantile Estimation by Orthogonal Polynomials

Bo Yang

This thesis proposes a smooth alternative to the conventional sample quantile function as a nonparametric estimator of population quantile function. We view our functions as approximations of the sample quantile function rather than that of true quantile function. Thus, a kind of two-stage procedure is involved: the true quantile function is estimated by the sample quantile function and the sample quantile is approximated by an orthogonal polynomial. The proposed estimator has the same asymptotic distribution as the conventional sample quantile. Monte Carlo studies are conducted to compare the proposed estimator with the sample quantile.

Acknowledgements

This work has been supported by my professors and many other individuals. I wish to take this opportunity to thank them all. My special thanks go to Dr. Y.P.Chaubey for encouraging, being patient and always available for discussion, and to Dr. M.M.Bilinsky and Dr. J.Garrido for their helpful comments.

Contents

List of Tables

Chapter 1

Introduction

The estimation of population quantiles is of great interest when one is not prepared to assume a parametric form for the underlying distribution. In addition, quantiles often arise as the natural object to estimate when the underlying distribution is skewed. Assume that $F(x)$ is an absolutely continuous distribution function with corresponding probability density function $f(x)$. Define the quantile function $Q(p)$ to be the left continuous inverse of F given by $Q(p) = \inf\{x : F(x) \geq p\}$, $0 < p < 1$. The function $Q(p)$ has derivative given by $Q'(p) = \frac{1}{f[Q(p)]}$ when it exists and it is called the quantile-density function. For $0 < p < 1$ define ξ_p to be the pth quantile of F satisfying $\xi_p = Q(p)$. We assume throughout that $f(\xi_p) > 0$.

Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed with c.d.f. F, and let $X_{(1)} \leq X_{(2)} \cdots \leq X_{(n)}$ denote the corresponding order statistics. A traditional estimator of ξ_p is the pth sample quantile given by $\hat{Q}(p) = X_{([np]+1)}$, where $[np]$ denotes the integral part of np. The main drawback of sample quantiles is that they experience a substantial lack of efficiency, caused by the variability of individual order statistics.

Some alternatives to the conventional sample quantiles, as the estimators of the population quantiles, have been proposed by Parzen(1979, p. 113), Reiss(1980), Harrell and Davis (1982), Kaigh and Lachenbruch (1982), Kaigh (1983), Falk (1984), Yang (1985), Brewer (1986), Padgett(1986), and Sheather and Marron (1990).

Suppose that K is a density function symmetric about 0 and that $h \to 0$ as $n \to \infty$. Parzen (1979) proposed the kernel quantile estimator

$$PQ(p) = \sum_{i=1}^{n} [\int_{\frac{i-1}{n}}^{\frac{i}{n}} K_h(t-p)dt] X_{(i)} \tag{1.1}$$

where $K_h(u) = h^{-1}K(u/h)$. Clearly, $PQ(p)$ is a weighted average of the order statistics $X_{(i)}$, for which $\frac{i}{n}$ is close to p more heavily than those $X_{(i)}$ for which $\frac{i}{n}$ is far from p. Reiss (1980) used the above estimator to test hypotheses concerning the population quantiles. Yang's (1985) estimator is given by

$$YQ(p) = \sum_{i=1}^{n} [n^{-1} K_h(\frac{i}{n} - p)] X_{(i)} \tag{1.2}$$

In the same paper, Yang also established the asymptotic normality and mean squared consistency of $PQ(p)$ and $YQ(p)$. He showed that $PQ(p)$ and $YQ(p)$ are asymptotically equivalent in mean square and gave the mean square convergence rates of the two estimators. Falk (1984) investigated the asymptotic relative deficiency of the sample quantile with respect to $PQ(p)$. Padgett (1986) generalized the definition of $PQ(p)$ to right-censored data. Sheather and Marron (1990) gave several improved kernel quantile estimators as follows:

$$SM_1(p) = \sum_{i=1}^{n} [n^{-1} K_h(\frac{(i-\frac{1}{2})}{n} - p)] X_{(i)} \tag{1.3}$$

$$SM_2(p) = \sum_{i=1}^{n} [n^{-1} K_h(\frac{i}{(n+1)} - p)] X_{(i)} \tag{1.4}$$

$$SM_3(p) = \frac{\sum_{i=1}^{n} [K_h(\frac{i-\frac{1}{2}}{n} - p)] X_{(i)}}{\sum_{j=1}^{n} K_h(\frac{j-\frac{1}{2}}{n} - p)}. \tag{1.5}$$

They also established asymptotic equivalences between $YQ(p)$, $PQ(p)$, $SM_1(p)$, $SM_2(p)$, and $SM_3(p)$.

L- estimators with much different motivation are included in those proposed by Harrell and Davis (1982) and Kaigh and Lachenbruch (1982). Since $E(X_{[np]+1})$

converges to $Q(p) = F^{-1}(p)$, Harrell and Davis took an estimator of $E(X_{[np]+1})$ as the estimator of $Q(p)$ given by:

$$HD(p) = \frac{\int_0^1 F_n^{-1}(t)t^{(n+1)p-1}(1-t)^{(n+1)(1-p)-1}dt}{B[(n+1)p,(n+1)(1-p)]}. \tag{1.6}$$

where $B[(n+1)p,(n+1)(1-p)]$ is the normalizing constant of the beta distribution with parameters $(n+1)p$ and $(n+1)(1-p)$. The Harrell-Davis estimator can also be written as

$$HD(p) = \sum_{i=1}^{n} [\int_{\frac{i-1}{n}}^{\frac{i}{n}} \frac{\Gamma(n+1)}{\Gamma((n+1)p)\Gamma((n+1)q)} \times t^{(n+1)p-1}(1-t)^{(n+1)q-1}dt]X_{(i)} \tag{1.7}$$

where $q = 1-p$ [Maritz and Jarrett (1978)]. Although Harrell and Davis did not use such terminology, this is exactly the bootstrap estimator of $E(X_{([np]+1)})$. In this case, an exact calculation replaces the more common evaluation by simulated re-sampling (see Efron 1979, p.5). Seather and Marron (1990) demonstrated an asymptotic equivalence between $PQ(p)$ and $HD(p)$ for a particular value of the bandwidth h.

For a fixed integer k, let $Y_{(1)}, Y_{(2)}, \ldots, Y_{(k)}$ be the ordered statistics of a sub-sample of size k randomly selected without replacement from the complete sample $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. Then a sample quantile estimator of $Q(p)$ based on the sub-sample would be $Y_{([kp]+1)}$. Kaigh and Lachenbruch (1982) defined their alternative quantile estimator $KL(p)$ to be the average of all sub-sample quantile estimators from all $\binom{n}{k}$ sub-samples of size k. They showed that their estimator may be written as

$$KL(p) = \sum_{i=r}^{n+r-k} \frac{\binom{i-1}{r-1}\binom{n-i}{k-r}}{\binom{n}{k}} X_{(i)} \tag{1.8}$$

where $r = [p(k+1)]$. Sheather and Marron (1990) also established the asymptotic equivalence between $PQ(p)$ and $KL(p)$ where the bandwidth is a function of k. This relationship together with the optimal bandwidth theory [Sheather and Marron (1990), section 2] provided a theory for the choice of k that minimizes the asymptotic

MSE of $KL(p)$. See Kaigh (1988) for interesting generalizations of the ideas behind $KL(p)$.

Kaigh (1983) pointed out that $HD(p)$ is based on ideas related to the Kaigh-Lachenbruch estimator. The latter is based on sampling without replacement, whereas the former is based on sampling with replacement in the case $k = n$. A referee has pointed out one could thus generalize $HD(p)$ to allow arbitrary k, and this estimator as well as other generalizations were proposed and studied by Kaigh and Cheng (1988).

Brewer (1986) proposed an estimator of ξ_p, based on likelihood arguments. His estimator is given by

$$BQ(p) = \sum_{i=1}^{n} [\frac{1}{n} \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i+1)} p^{i-1}(1-p)^{n-i}] X_{(i)} \tag{1.9}$$

It was also shown to be asymptotically equivalent to $PQ(p)$ [see Sheather and Marron (1990)].

A kernel type estimator of $F(x)$ is

$$\hat{F}(x) = \int_{-\infty}^{x} (nh)^{-1} \sum_{i=1}^{n} K(\frac{X_i - t}{h}) dt \tag{1.10}$$

Then $\hat{Q}(p) = \inf\{x | \hat{F}(x) \geq p\}$ can be another estimator of $Q(p)$.

Our thesis is to propose and study another alternative to the sample quantile in an attempt to improve the precision of estimation of population quantiles and smooth the sample quantiles. The alternative estimator is a linear combination of order statistics and may be viewed as " quantile estimation by orthogonal polynomials" obtained by a Fourier expansion of the sample quantile over a series of orthogonal polynomials.

In general, the effective performance of nonparametric function estimators is critically dependent on the choice of a "smoothing parameter". If not enough smoothing is done, the estimate will be "rough", showing features which do not represent the function being estimated. On the other hand, if too much smoothing is done, important features of the curve may not show up since they are essentially smoothed away (Marron, 1986). In our estimator, the smoothing parameter is the number of the terms of orthogonal polynomials N. From the simulation studies, we can see that when $N = 10$, the estimator can achieve a very good approximation of the true quantile function.

In Chapter 2, the alternative estimator and some of their asymptotic properties are discussed. In Chapter 3, we employ Monte Carlo stimulation to compare the alternative estimator with the conventional sample quantile estimator. The computer program for the simulation studies is given in an appendix.

Chapter 2

Quantile Estimation by Orthogonal Polynomials

2.1 Approximation of Functions by Orthogonal Polynomials

Let $\{\phi_i, i \geq 0\}$ be a sequence of functions orthogonal with respect to a non-negative weight function w on B, usually R (real space) or $[0,1]$. Then ϕ_i should satisfy the following conditions:

$$\int_B w(x)\phi_i(x)\phi_j(x)dx = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

We will define (f,g) as $\int_B w(x)f(x)g(x)dx$. So

$$(\phi_i, \phi_j) = \int_B w(x)\phi_i(x)\phi_j(x)dx$$

Consider the problem of finding the best approximation to $f(x)$ generated by the orthogonal sequences $\phi_1(x)$, $\phi_2(x)$, ..., $\phi_N(x)$. With a simple computation, we can get the best approximation as follows:

$$\hat{f}(x) = \sum_{i=1}^{N} a_i \phi_i(x)$$

where $a_i = \int_B w(x) f(x) \phi_i(x) dx$.

In this section we generalize this special approximation problem by considering approximations generated by an infinite orthogonal polynomials.

Definition 1 *An infinite series of the form $\sum_{i=1}^{\infty} \phi_i(x)$ is said to converge to the function $\phi(x)$ if the sequence of partial sums $S_n(x) = \sum_{i=1}^{n} \phi_i(x)$ converges to $\phi(x)$.*

Lemma 1 *Let $\{\phi_i(x)\}$ be a sequence of orthogonal polynomials. A series of the form $\sum_{i=1}^{\infty} a_i \phi_i(x)$ converges to an function $\phi(x)$ if and only if $\sum_{i=1}^{\infty} |a_i|^2 < \infty$ and, in that case, we have $a_i = \int_B w(x) \phi(x) \phi_i(x)$.*

(See Luenberger (1969)).

Lemma 2 *(Bessel's Inequality) Let $f(x)$ be a function and suppose $\{\phi_i(x)\}$ is a sequence of orthogonal polynomials. Then*

$$\sum_{i=1}^{\infty} |(f, \phi_i)|^2 \leq ||f||^2$$

where $||f|| = \sqrt{(f, f)}$.

Since Bessel's inequality guarantees that $\sum_{i=1}^{\infty} |(f, \phi_i)|^2 < \infty$, Lemma 1 guarantees that the series $\sum_{i=1}^{\infty} (f, \phi_i) \phi_i$ converges to some function.

Definition 2 *An orthogonal system $\{\phi_i\}$ is called complete in $L_p(B)$ if for any function $f \in L_p(B)$, $\int_B f(x) \phi_i(x) dx = 0$, for all i, implies $f = 0$ almost everywhere.*

Example 1 Consider $L_2[-1, 1]$ consisting of square-integrable functions on the interval $[-1, 1]$. With Gram-Schmidt procedure, the independent functions $1, x, x^2, \ldots, x^n, \ldots$ produce a sequence of orthogonal polynomials $\{\phi_i\}_{i=1}^{\infty}$. It turns out that

$$\phi_i(x) = \sqrt{\frac{2n+1}{2}} P_i(x), \quad i = 1, 2, \ldots,$$

where the $P_i(x)$ are Legendre polynomials

$$P_i(x) = \frac{1}{2^i i!}\frac{d^i}{dx^i}\{(x^2-1)^i\}$$

We now show that this sequence is a complete sequence of orthogonal polynomials in $L_2[-1,1]$.

Assume that there exists an $f \in L_2[-1,1]$ orthogonal to each x^n, then we have

$$\int_{-1}^{1} x^n f(x)dx = 0 \quad \text{for } n = 0,1,2,\ldots.$$

The continuous function

$$F(x) = \int_{-1}^{x} f(t)dt$$

has $F(-1) = F(1) = 0$, and

$$\begin{aligned}\int_{-1}^{1} x^n F(x)dx &= \frac{x^{n+1}}{n+1}F(x)|_{-1}^{1} - \int_{-1}^{1}\frac{x^{n+1}}{n+1}f(x)dx \\ &= 0\end{aligned}$$

for $n = 0,1,2,\ldots$. Thus the continuous function F is also orthogonal to the polynomials.

Since F is continuous, the Weierstrass approximation theorem applies. Given $\epsilon > 0$, there is a polynomial

$$G(x) = \sum_{i=1}^{N} a_i x^i$$

such that

$$|F(x) - G(x)| < \epsilon$$

for all $x \in [-1,1]$. Therefore, since F is orthogonal to the polynomials,

$$\begin{aligned}\int_{-1}^{1}|F(x)|^2dx &= \int_{-1}^{1}F(x)[F(x)-G(x)]dx \\ &\le \epsilon\int_{-1}^{1}|F(x)|dx \\ &\le \epsilon\sqrt{2}\{\int_{-1}^{1}|F(x)|^2dx\}^{\frac{1}{2}}\end{aligned}$$

So we get

$$\int_{-1}^{1} |F(x)|^2 dx \leq 2\epsilon^2.$$

Since ϵ is arbitrary and F is continuous, we have that $F(x) = 0$, and therefore, $f(x) = 0$ almost everywhere. So the sequence of the orthogonal polynomials is complete.

Definition 3 *A complete system $\{\phi_i\}$ is called a basis in $L_p(B)$ if for every $f \in L_p(B)$ there is a unique convergent series expansion $\sum a_i \phi_i$.*

Definition 4 *The sequence $\{X_n\}$ is said to converge in $L_p(B)$ to X if and only if*

$$\lim_{n\to\infty} E[|X_n - X|^p] = 0$$

.

It is known that

Lemma 3 *A system is complete on $L_2(B)$ if and only if*

$$\int f^2 = \sum_{i=0}^{\infty} a_i^2, \quad \textit{for all } f \in L_2(B)$$

(See Devroye and Györfi (1984)). This is called Bessel's equality. In that case, we actually have convergence of the partial sums in $L_2(B)$:

Theorem 1 *If $f \in L_2(B)$, and the system is complete, then*

$$\int (f - \sum_{i=1}^{N} a_i \phi_i)^2 \to 0, \quad \textit{as } N \to \infty$$

(See Sansone (1977), p.23.)

2.2 General Properties

Let X_1, X_2, ..., X_n be a random sample drawn from a population with continuous density function f, and $\{\phi_i\}_{i=1}^{\infty}$ be a sequence of orthogonal polynomials on a set B, then $a_i = \int_B f(x)\phi_i(x)dx$ can be estimated without bias by

$$a_{ni} = \frac{1}{n}\sum_{j=1}^{n}\phi_i(X_j)$$

and $f(x)$ can be estimated by

$$f_n(x) = \sum_{i=0}^{N} a_{ni}\phi_i(x).$$

or

$$f_n(x) = \frac{1}{n}\sum_{j=1}^{n} K_N(x, X_j)$$

where

$$K_N(x, X_j) = \sum_{i=0}^{N}\phi_i(x)\phi_i(X_j).$$

The partial sum is defined as

$$S_m(f) = \sum_{i=0}^{m} a_i\phi_i$$

From Devroye and Györfi (1984), we have

Lemma 4 *Let f_n be an estimate of f by orthogonal polynomials with $m = m_n$ terms, and let f have the expansion*

$$\sum_{i=0}^{\infty} a_i\phi_i(x)$$

and $a_i = \int_B f(x)\phi_i(x)dx$. Assume also that all ϕ_i's are absolutely integrable on B. Then

$$\begin{aligned}
E(\int |f_n - f|) &\leq \int |S_m(f) - f| + E(\int |f_n - E(f_n)|) \\
&\leq \int |S_m(f) - f| + \int \sqrt{E((f_n - E(f_n))^2)} \\
&\leq \int |S_m(f) - f| + \frac{1}{\sqrt{n}} \int \sqrt{E(K^2(x, X_1))} \\
E(\int |f_n - f|) &\geq Max(\int |S_m(f) - f|, \frac{1}{2}\int E(|f_n - E(f_n)|)).
\end{aligned}$$

Lemma 5 *Let $f \in L_2(B)$, then the estimate f_n of f by orthogonal polynomials with m terms has the following expected L_2 error:*

$$
\begin{aligned}
E(\int (f_n - f)^2) &= \frac{1}{n}\sum_{i=0}^{m}(\int f\phi_i^2 - a_i^2) + \sum_{i=m+1}^{\infty} a_i^2 \\
&\leq \frac{1}{n}\sum_{i=0}^{m}\int f\phi_i^2 + \sum_{i=m+1}^{\infty} a_i^2 \\
&= \frac{1}{n}\int f(x)K_m(x,x)dx + \sum_{i=m+1}^{\infty} a_i^2.
\end{aligned}
$$

2.3 Asymptotic Properties of $\hat{Q}_N(p)$

We shall view our functions as approximation to the conventional quantile function, rather than the true quantile function. Thus, we need only assume that the sequence of orthogonal polynomials is complete in the class of step function on $(0,1)$.

Let $\hat{Q}$ be the traditional sample quantile function of an independent sample $\{X_j, 1 \leq j \leq n\}$ from an unknown distribution F. The generalized Fourier expansion of $\hat{Q}$, to "N terms", is given by

$$\hat{Q}_N(p) = \sum_{i=1}^{N} \hat{a}_i \phi_i(p).$$

and

$$\hat{Q}(p) = X_{([np]+1)}$$

where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are the order statistics of the X_i's.
Then we have

$$
\begin{aligned}
\hat{a}_i &= \int_0^1 \hat{Q}(p)\phi_i(p)w(p)dp \\
&= \int_0^1 X_{([np]+1)}\phi_i(p)w(p)dp
\end{aligned}
$$

$$= \int_0^{\frac{1}{n}} X_{(1)}\phi_i(p)w(p)dp + \int_{\frac{1}{n}}^{\frac{2}{n}} X_{(2)}\phi_i(p)w(p)dp + \cdots + \int_{\frac{n-1}{n}}^{1} X_{(n)}\phi_i(p)w(p)dp$$

$$= \sum_{j=1}^{n} X_{(j)} \int_{\frac{j-1}{n}}^{\frac{j}{n}} \phi_i(p)w(p)dp$$

So

$$\hat{Q}_N(p) = \sum_{i=1}^{N}\sum_{j=1}^{n} X_{(j)} \int_{\frac{j-1}{n}}^{\frac{j}{n}} \phi_i(x)w(x)dx\phi_i(p).$$

or

$$\hat{Q}_N(p) = \sum_{j=1}^{n} [\sum_{i=1}^{N} \int_{\frac{j-1}{n}}^{\frac{j}{n}} \phi_i(p)\phi_i(x)w(x)dx] X_{(j)}.$$

The estimator can also be re-expressed as

$$\hat{Q}_N(p) = \sum_{j=1}^{n} a_j X_{(j)}$$

where

$$a_j = \sum_{i=1}^{N} \int_{\frac{j-1}{n}}^{\frac{j}{n}} \phi_i(x)\phi_i(p)w(x)dx.$$

Thus, $\hat{Q}_N(p)$ is an L-statistic. Following David (1981, p.273), $\hat{Q}_N(p)$ is asymptotically normally distributed.

The efficacy of the orthogonal polynomial method rests on the simplicity of the choice of N, and on the ease with which the constants $\hat{a}_i$ can be calculated.

Lemma 6 *Let F be a twice differentiable distribution function, having finite support on R^1. Assume $\{\inf_{0\leq p\leq 1} f(F^{-1}(p))\} > 0$ and $\sup_{0\leq p\leq 1} |f'(F^{-1}(p))| < \infty$. Then*

$$sup_{0\leq p\leq 1}|\hat{Q}(p) - Q(p)| \to 0. \qquad a.e.$$

(See Csörgõ (1983)). From Chung(1974), we have

Lemma 7 *Convergence a.e. implies convergence in probability.*

i.e

$$\hat{Q}(p) - Q(p) \to 0 \quad \text{in probability} \quad \text{as } n \to \infty$$

Lemma 8 *If N/n[1] $\to \infty$ as $n \to \infty$, then*

$$E[\hat{Q}_N(p) - \hat{Q}(p)]^2 \to 0$$

This can be derived directly from Theorem 1. Thus we get

$$\hat{Q}_N(p) \to \hat{Q}(p) \quad \text{in} \quad L_2$$

.

Lemma 9 *If X_n converges to 0 in L_p, then it converges to 0 in probability.*

(See Chung (1974)). Therefore, we have

$$\hat{Q}_N(p) - \hat{Q}(p) \to 0 \quad \text{in probability as } n \to \infty \text{ (i.e. } N \to \infty)$$

Theorem 2 *If $N/n \to \infty$ as $n \to \infty$, then*

$$\hat{Q}_N(p) - Q(p) \to 0 \quad \textit{in probability as } n \to \infty$$

Proof. Since

$$|\hat{Q}_N(p) - Q(p)| \leq |\hat{Q}_N(p) - \hat{Q}(p)| + |\hat{Q}(p) - Q(p)|$$

Let

$$A = \{p : |\hat{Q}_N(p) - \hat{Q}(p)| \leq \frac{\epsilon}{2}\}$$
$$B = \{p : |\hat{Q}(p) - Q(p)| \leq \frac{\epsilon}{2}\}$$

and

$$C = \{p : |\hat{Q}_N(p) - Q(p)| \leq \epsilon\}$$

Since

$$A \cap B \subseteq C$$

[1] Actually, N can be often chosen much smaller than the sample size n.

We have

$$\bar{A} \cup \bar{B} \supseteq \bar{C}$$

where

$$\bar{A} = \{p : |\hat{Q}_N(p) - \hat{Q}(p)| > \frac{\epsilon}{2}\}$$

$$\bar{B} = \{p : |\hat{Q}(p) - Q(p)| > \frac{\epsilon}{2}\}$$

and

$$\bar{C} = \{p : |\hat{Q}_N(p) - Q(p)| > \epsilon\}$$

So we get

$$\mathcal{P}(\bar{C}) \leq \mathcal{P}(\bar{A}) + \mathcal{P}(\bar{B})$$

where $\mathcal{P}$ denotes the probability measure induced by F. With the assumption that

$$\hat{Q}_N(p) - \hat{Q}(p) \to 0 \quad \text{in probability as } n \to 0$$

$$\hat{Q}(p) - Q(p) \to 0 \quad \text{in probability as } n \to 0$$

i.e

$$\lim_{n\to\infty} \mathcal{P}(\bar{A}) = \lim_{N\to\infty} \mathcal{P}\{p : |\hat{Q}_N(p) - \hat{Q}(p)| \geq \frac{\epsilon}{2}\} = 0$$

$$\lim_{n\to\infty} \mathcal{P}(\bar{B}) = \lim_{n\to\infty} \mathcal{P}\{p : |\hat{Q}(p) - Q(p)| \geq \frac{\epsilon}{2}\} = 0$$

So

$$\lim_{n\to\infty} \mathcal{P}(\bar{C}) \leq \lim_{n\to\infty} \mathcal{P}(\bar{A}) + \lim_{n\to\infty} \mathcal{P}(\bar{B}) = 0$$

i.e.

$$\hat{Q}_N(p) - Q(p) \to 0 \quad \text{in probability as } n \to 0$$

Simulation results presented in the next chapter indicate the proposed quantile estimator is a promising alternative over the sample quantile function in terms of the smoothness and mean squared error.

Chapter 3

Simulation Studies

3.1 Simulation Studies

Monte Carlo simulation studies were conducted to compare the performance of the proposed estimators and the sample quantile for five families of distributions (Exponential, Normal, Gamma, Uniform, Triangular). Legendre polynomials were used to estimate the quantile function $\hat{Q}_N(p)$, because they provide a complete orthogonal basis.

Legendre polynomials of order n can be written as

$$P_n(t) = \frac{1}{2^n n!}\frac{d^n}{dt^n}[(t^2-1)^n].$$

The sequence is orthogonal and complete in $L^2[-1,1]$.

Furthermore, a total orthogonal sequence in the space $L^2[a,b]$ is L_n, where

$$L_n = \frac{1}{\|p_n\|}p_n, \qquad p_n(t) = P_n(s), \qquad s = 1 + 2\frac{t-a}{b-a}$$

The proof follows if we note that $a \le t \le b$ corresponds to $-1 \le s \le 1$ and the orthogonality is preserved under this linear transformation $t \longmapsto s$.

We thus have a total orthogonal sequence in $L^2[0,1]$ for the compact interval $[0,1]$.

With easy computation from the formulas above, we have

$$L_n(p) = \frac{(2n+1)^{\frac{1}{2}}}{n!}\frac{d^n}{dp^n}p^n(p-1)^n$$

where $n = 0, 1, 2, \ldots$. The first several terms are given as follows:

$$L_0(p) = 1$$

$$L_1(p) = \sqrt{3}(2p-1)$$

$$L_2(p) = \sqrt{5}(6p^2 - 6p + 1)$$

$$L_3(p) = \sqrt{7}(20p^3 - 30p^2 + 12p - 1)$$

And the weighted function for the Legendre Polynomials is $w(p) = 1$.

Because it is difficult to compute with this form in the computer, we use the recursive formula:

$$L_n(p) = (a_n p + b_n)L_{n-1}(p) - c_n L_{n-2}(p)$$

(see Luenberger (1969) P.75), $n = 2, 3, \ldots$, with

$$a_n = 2\sqrt{4 - \frac{1}{n^2}}$$

$$b_n = -\sqrt{4 - \frac{1}{n^2}}$$

$$c_n = \sqrt{\frac{2n+1}{2n-3}}\frac{n-1}{n}$$

The simulation was performed for five families of distributions that are commonly used. These distributions are shown in Table 1.

The mean squared errors of an estimators were used to measure efficiency, for example

$$MSE(\hat{Q}_N(p)) = E\{\hat{Q}_N(p) - Q(p)\}^2$$

$$MSE(\hat{Q}(p)) = E\{\hat{Q}(p) - Q(p)\}^2$$

These mean squared errors considered in Table 2 were estimated based on 10,000 independent Monte Carlo trials. The simulations were done on the Fisher workstation in the Department of Mathematics and Statistics of Concordia University. The exponential, normal, gamma, uniform, and triangle random numbers were generated by a SAS program on VAXII of Concordia University.

From the 10,000 samples, the estimated mean squared errors (MSE) of the estimators $\hat{Q}_N(p)$, $\hat{Q}(p)$ were computed, and the ratios of these estimated mean squared errors $MSE(\hat{Q}(p))/MSE(\hat{Q}_N(p))$ were calculated.

Table 3.1: Distributions Used in Simulations

Distribution	Density
Exponential	$f(x) = \beta exp(-\beta x), \quad x > 0$
Normal	$f(x) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} exp^{\frac{(x-\mu)^2}{2\sigma^2}}$
Gamma	$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} exp(-x/\beta)$
Uniform	$f(x) = \frac{1}{(b-a)} \quad a \leq x \leq b$
Triangular	$f(x) = \begin{cases} \frac{2x}{h} & if\ 0 \leq x \leq h \\ \frac{2(1-x)}{(1-h)} & if\ h \leq x \leq 1 \end{cases}$ where $0 < h < 1$.

Table 3.2: Ratio of the Mean Squared Error of $\hat{Q}(p)$ to that of $\hat{Q}_N(p)$ (n = 100, $N = 40$)

p	Exponential	Normal	Uniform	Gamma	Triangular
0.05	1.433735	1.317731	0.957361	0.768605	1.251294
0.10	1.255647	1.192206	0.995385	0.939847	1.167362
0.20	1.086199	1.143584	1.053124	0.990225	1.132148
0.30	1.072435	1.124017	1.114974	1.006874	1.115211
0.40	1.084842	1.067263	1.075361	1.013774	1.095464
0.50	1.087808	1.050122	0.978547	1.023494	1.063327
0.60	1.086553	0.959671	1.092713	0.991433	1.024368
0.70	1.050166	1.043096	1.078536	1.006323	0.991075
0.80	1.030233	1.096827	1.085785	1.041024	0.983681
0.90	1.120960	1.127729	1.248530	1.090807	0.962156
0.95	1.201535	1.071169	1.074269	1.113792	0.893721

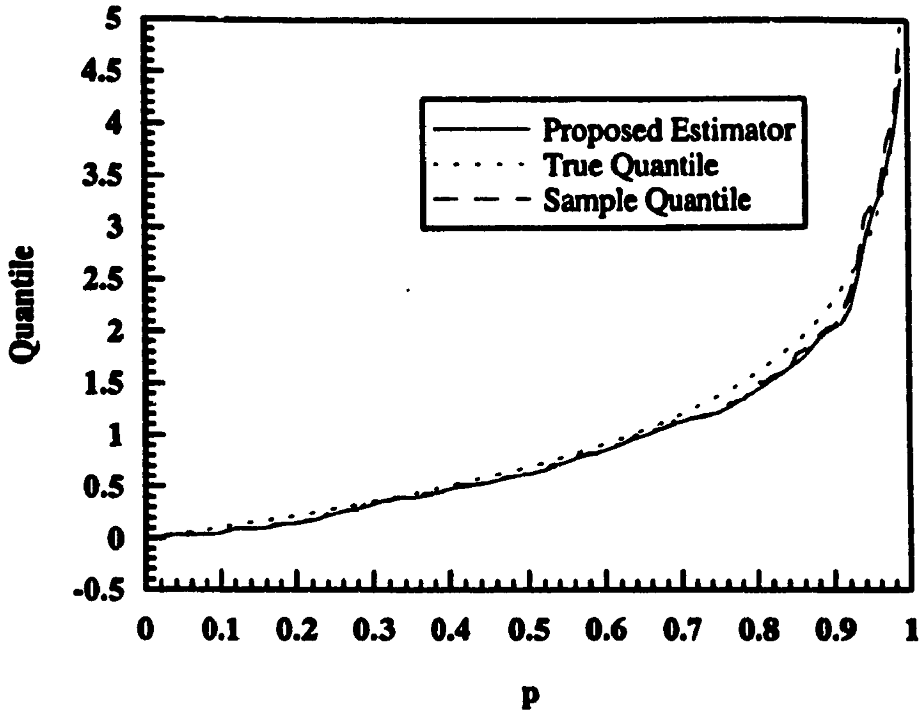

Figure 1. Plots of the Sample Quantile, Proposed Quantile Estimator and True Quantile of Exponential Distribution with n=100 and N=20.

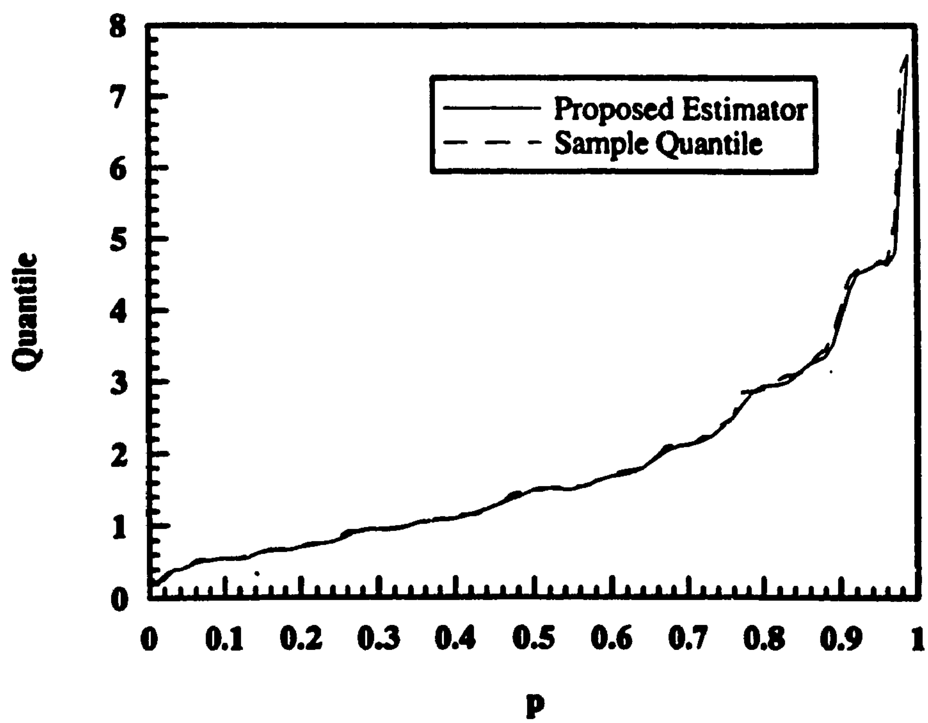

Figure 2. Plots of the Sample Quantile, Proposed Quantile Estimator and True Quantile of Gamma Distribution with n=100 and N=20.

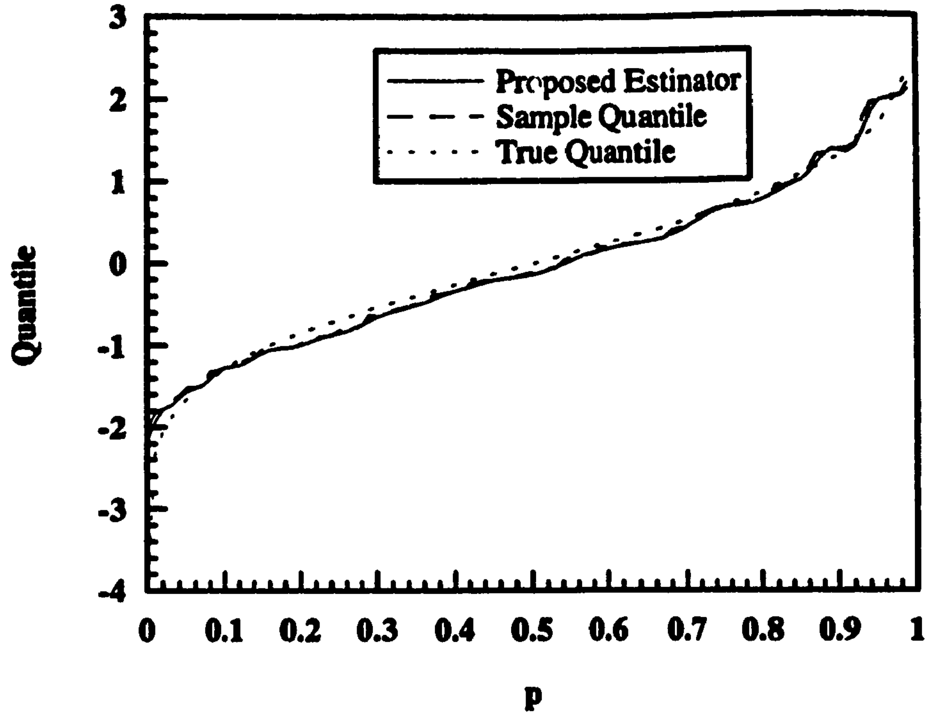

Figure 3. Plots of the Sample Quantile, Proposed Quantile Estimator and True Quantile of Normal Distribution with n=100 and N=20.

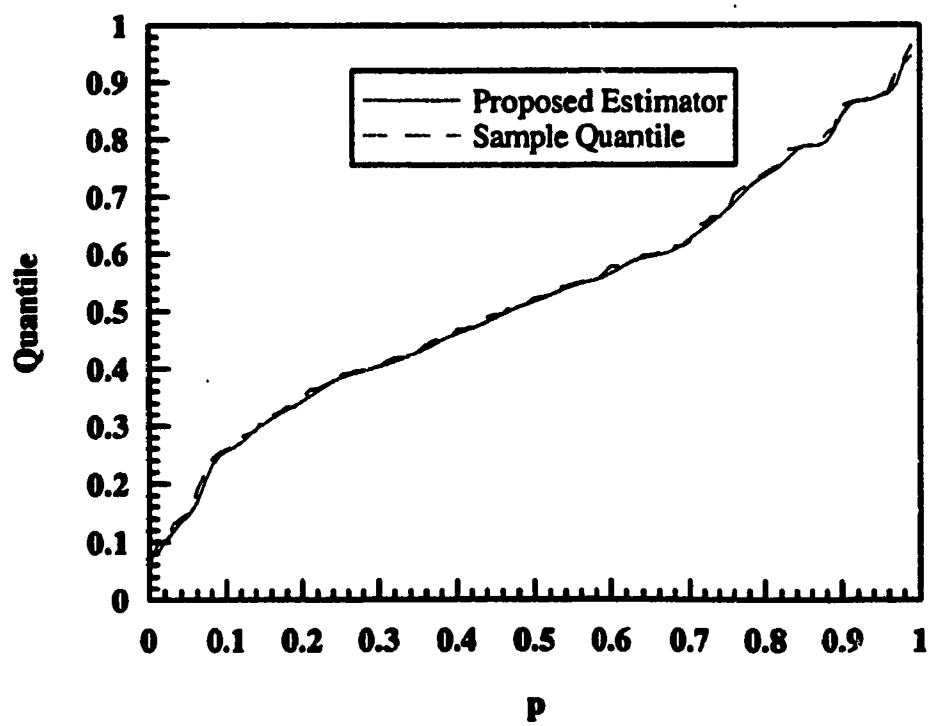

Figure 4. Plots of the Sample Quantile, Proposed Quantile Estimator and True Quantile of Tranigular Distribution with n=100 and N=20.

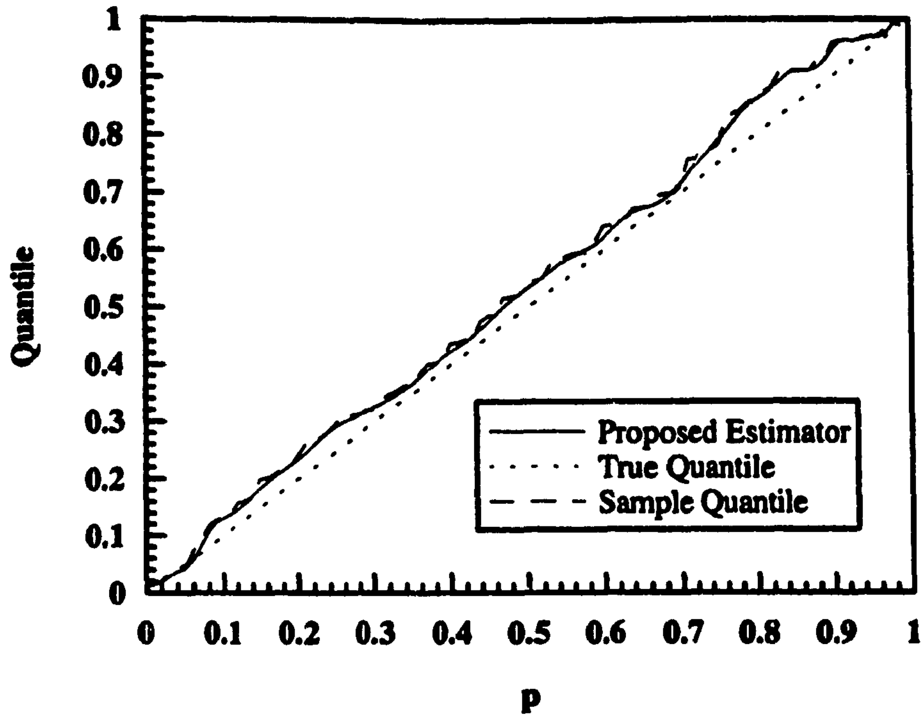

Figure 5. Plots of the Sample Quantile, Proposed Quantile Estimator and True Quantile of Uniform Distribution with n=100 and N=20.

3.2 Appendix Computer Program for Simulation Studies

This is a computer program written in C language. The random numbers are given by a program by SAS. From this program, we can get the estimation of the quantile estimation by orthogonal polynomial.

```
#include "head.h"
#define N 100
#define INT_N 500
float L[N];
float l(i,t)
int i;
float t;
{
  int j;L[0]=1.0;
  L[1]=sqrt(3.0)*(2.0*t-1.0);
  for(j=2;j<=i;j++)
    L[j]=((2.0*sqrt(4.0*j*j-1.0)/j)*t-sqrt(4.0*j*j-1.0)/j)*L[j-1]+
    (sqrt((2.0*j+1.0)/(2.0*j-3.0))*((1.0-j)/j))*L[j-2];
  return (L[i]);
}

main()
{
  int i,j,k,m,m1;
  float qhat,lamda,x[n],L1[N];
  FILE *fp,*gp;
  system("rm -f qhat.dat");
```

```
for(m1=0;m1<datanumber;m1++)
{
   if((fp=fopen("gensort.dat","r"))==NULL)
   {
      printf("Can't open file");
      exit();
   }
   for(m=0;m<m1;m++)
     for(i=0;i<n;i++)
       fscanf(fp,"%f",&x[0]);
   for(i=0;i<n;i++)
   {
     fscanf(fp,"%f",&x[i]);
   }
   fclose(fp);
   for(m=4;m<n;m+=5)
    if(m==4||m==n-6||m%10==9&&m<n-6&&m>4)
    {
      qhat=0.0;
      lamda=((float)m)/n;
      l(N-1,lamda);
      for(i=0;i<N;i++)L1[i]=L[i];
        for(j=0;j<n;j++)
           for(k=0;k<INT_N;k++)
           {
             l(N-1,(j*INT_N+(float)k)/(n*INT_N));
             for(i=0;i<N;i++)
               qhat=qhat+x[j]*((1.0/(n*INT_N))*L[i])*L1[i];
           }
```

```
        if((gp=fopen("qhat.dat", "a"))==NULL)
        {
           printf("Can't write file");
           exit();
        }
        fprintf(gp, "%f\n",qhat);
        fclose(gp);
     }
  }
}


#include "head.h"
main()
{
   FILE *fp,*gp;
   float f;
   int i,j;
   if((fp=fopen("gensort.dat","r"))==NULL)
   {
     printf("Can't open file");
     exit();
   }
   if((gp=fopen("gen2p.dat", "w"))==NULL)
   {
     printf("Can't write file");
     exit();
   }
   for(i=0;i<datanumber;i++)
```

```
for(j=0;j<n;j++)
{
  fscanf(fp,"%f",&f);
  if(j==5||j==n-5||j>5&&j<n-5&&j%10==0)
     fprintf(gp,"%f\n",f);
}
fclose(fp);fclose(gp);
}


#include "head.h"
main()
{
   float i,j,k,x[n];
   FILE *fp,*gp;
   int m,m1;
   if((fp=fopen("gen.dat","r"))!=NULL&&(gp=fopen("gensort.dat","w"))!=NULL)
   {
      for(m=0;m<datanumber;m++)
        {
          for(m1=0;m1<n;m1++)
          {
             fscanf(fp,"%*d %f",&x[m1]);
          }
          sort(x);
          for(m1=0;m1<n;m1++)
          {
              fprintf(gp,"%f\n",x[m1]);
          }
```

```
        }
    }
    fclose(fp);
    fclose(gp);
}

sort(x) float x[];
{
    float y;
    int i,j;
    for(i=0;i<n;i++)
      for(j=i+1;j<n;j++)
        if(x[i]>x[j])
        {
           y=x[i];
           x[i]=x[j];
           x[j]=y;
        }
}

#include "math.h"
#include "stdio.h"
#define n 100
main()
{
   int i;
   FILE *fp;
   if((fp=fopen("q.dat","w"))==NULL)
   {
```

```
        printf("Cannot write");
        exit();
    }
    for(i=5;i<n;i+=5)
      if(i==5||i==n-5||i>5&&i<n-5&&i%10==0)
         fprintf(fp,"%f\n", -log(1-i/100.0) );
    fclose(fp);
}
```

```
#include "head.h"
#define N 40
#define INT_N 100
main()
{
   int i,m,p;
   float qhatsum,f,f1,q[11];
   FILE *fp;
   if((fp=fopen("q.dat","r"))==NULL)
   {
       printf("Cannot read");
       exit();
   }
   for(i=0;i<=10;i++)
     fscanf(fp,"%f",&q[i]);
     fclose(fp);
   system("rm -f sum.dat");
   for(p=0;p<=10;p++)
   {
```

```
    qhatsum=0.0;
    if((fp=fopen("qhat.dat","r"))==NULL)
    {
       printf("Cannot read");
       exit();
    }
 for(m=0;m<datanumber;m++)
   for(i=0;i<=10;i++)
   {
      fscanf(fp,"%f",&f);
      if(i==p){qhatsum+=(f-q[i])*(f-q[i]);
   }
 }
 fclose(fp);
 if((fp=fopen("sum.dat","a"))==NULL)
 {
    printf("Cannot write");
    exit();
 }
 fprintf(fp,"%f\n",qhatsum/((float) datanumber));
 fclose(fp);
}
 for(p=0;p<=10;p++)
 {
   qhatsum=0.0;
   if((fp=fopen("gen2p.dat","r"))==NULL)
   {
      printf("Cannot read");
      exit();
```

```
    }
    for(m=0;m<datanumber;m++)
      for(i=0;i<=10;i++)
      {
        fscanf(fp,"%f",&f);
        if(i==p){qhatsum+=(f-q[i])*(f-q[i]);
      }
    }
    fclose(fp);
    if((fp=fopen("sum.dat","a"))==NULL)
    {
       printf("Cannot write");
       exit();
    }
    fprintf(fp, "%f\n", qhatsum/((float) datanumber));
    fclose(fp);
  }


#include "head.h"
#define N 40
#define INT_N 100
main()
{
  int i,m,p;
  float qhat[11],qbar[11];
  FILE *fp,*gp;
  if((fp=fopen("sum.dat","r"))==NULL)
  {
```

```
    printf("Cannot read");
    exit();
  }
  if((gp=fopen("quo.dat","w"))==NULL)
  {
    printf("Cannot write");
    exit();
  }
  for(i=0;i<11;i++)
    fscanf(fp,"%f",&qhat[i]);
  for(i=0;i<11;i++)
    fscanf(fp,"%f",&qbar[i]);
  fprintf(gp," P  \tqhat    \tqbar    \tqbar/qhat\n");
  for(i=0;i<11;i++)fprintf(gp, "0.%02d\t%f\t%f\t%f\n",
            i==0?5:i==10?95:10*i,qhat[i],qbar[i],qbar[i]/qhat[i]);
  fclose(fp);
  fclose(gp);
}
```

This is a computer program written in C language. The random numbers are given by a program in SAS. From this program, we can get the estimation of the quantile estimation by orthogonal polynomials.

```
#include "head.h"
#define N 100
#define INT_N 500
float L[N];
float l(i,t)
int i;
float t;
{
  int j;L[0]=1.0;
  L[1]=sqrt(3.0)*(2.0*t-1.0);
  for(j=2;j<=i;j++)
    L[j]=((2.0*sqrt(4.0*j*j-1.0)/j)*t-sqrt(4.0*j*j-1.0)/j)*L[j-1]+
    (sqrt((2.0*j+1.0)/(2.0*j-3.0))*((1.0-j)/j))*L[j-2];
  return (L[i]);
}

main()
{
  int i,j,k,m,m1;
  float qhat,landa,x[n],L1[N];
  FILE *fp,*gp;
  system("rm -f qhat.dat");
  for(m1=0;m1<datanumber;m1++)
  {
     if((fp=fopen("gensort.dat","r"))==NULL)
```

```
{
   printf("Can't open file");
   exit();
}
for(m=0;m<m1;m++)
  for(i=0;i<n;i++)
    fscanf(fp,"%f",&x[0]);
for(i=0;i<n;i++)
{
  fscanf(fp,"%f",&x[i]);
}
fclose(fp);
for(m=4;m<n;m+=5)
 if(m==4||m==n-6||m%10==9&&m<n-6&&m>4)
 {
   qhat=0.0;
   lamda=((float)m)/n;
   l(N-1,lamda);
   for(i=0;i<N;i++)Ll[i]=L[i];
     for(j=0;j<n;j++)
        for(k=0;k<INT_N;k++)
        {
          l(N-1,(j*INT_N+(float)k)/(n*INT_N));
          for(i=0;i<N;i++)
            qhat=qhat+x[j]*((1.0/(n*INT_N))*L[i])*Ll[i];
        }
   if((gp=fopen("qhat.dat", "a"))==NULL)
   {
      printf("Can't write file");
```

```
            exit();
        }
        fprintf(gp, "%f\n",qhat);
        fclose(gp);
    }
  }
}


#include "head.h"
main()
{
    FILE *fp,*gp;
    float f;
    int i,j;
    if((fp=fopen("gensort.dat","r"))==NULL)
    {
      printf("Can't open file");
      exit();
    }
    if((gp=fopen("gen2p.dat", "w"))==NULL)
    {
      printf("Can't write file");
      exit();
    }
    for(i=0;i<datanumber;i++)
      for(j=0;j<n;j++)
      {
        fscanf(fp,"%f",&f);
```

```
      if(j==5||j==n-5||j>5&&j<n-5&&j%10==0)
        fprintf(gp,"%f\n",f);
    }
  fclose(fp);fclose(gp);
}


#include "head.h"
main()
{
  float i,j,k,x[n];
  FILE *fp,*gp;
  int m,m1;
  if((fp=fopen("gen.dat","r"))!=NULL&&(gp=fopen("gensort.dat","w"))!=NULL)
  {
    for(m=0;m<datanumber;m++)
      {
       for(m1=0;m1<n;m1++)
       {
         fscanf(fp,"%*d %f",&x[m1]);
       }
       sort(x);
       for(m1=0;m1<n;m1++)
       {
          fprintf(gp,"%f\n",x[m1]);
       }
      }
  }
  fclose(fp);
```

```
   fclose(gp);
}

sort(x) float x[];
{
   float y;
   int i,j;
   for(i=0;i<n;i++)
     for(j=i+1;j<n;j++)
       if(x[i]>x[j])
       {
          y=x[i];
          x[i]=x[j];
          x[j]=y;
       }
}

#include "math.h"
#include "stdio.h"
#define n 100
main()
{
   int i;
   FILE *fp;
   if((fp=fopen("q.dat","w"))==NULL)
   {
       printf("Cannot write");
       exit();
   }
```

```
  for(i=5;i<n;i+=5)
    if(i==5||i==n-5||i>5&&i<n-5&&i%10==0)
      fprintf(fp,"%f\n", -log(1-i/100.0) );
  fclose(fp);
}


#include "head.h"
#define N 40
#define INT_N 100
main()
{
   int i,m,p;
   float qhatsum,f,f1,q[11];
   FILE *fp;
   if((fp=fopen("q.dat","r"))==NULL)
   {
      printf("Cannot read");
      exit();
   }
   for(i=0;i<=10;i++)
     fscanf(fp,"%f",&q[i]);
     fclose(fp);
   system("rm -f sum.dat");
   for(p=0;p<=10;p++)
   {
     qhatsum=0.0;
     if((fp=fopen("qhat.dat","r"))==NULL)
     {
```

```
            printf("Cannot read");
            exit();
        }
    for(m=0;m<datanumber;m++)
      for(i=0;i<=10;i++)
      {
         fscanf(fp,"%f",&f);
         if(i==p){qhatsum+=(f-q[i])*(f-q[i]);
      }
    }
    fclose(fp);
    if((fp=fopen("sum.dat","a"))==NULL)
    {
       printf("Cannot write");
       exit();
    }
    fprintf(fp,"%f\n",qhatsum/((float) datanumber));
    fclose(fp);
}
    for(p=0;p<=10;p++)
    {
      qhatsum=0.0;
      if((fp=fopen("gen2p.dat","r"))==NULL)
      {
         printf("Cannot read");
         exit();
      }
      for(m=0;m<datanumber;m++)
         for(i=0;i<=10;i++)
```

```
        {
          fscanf(fp,"%f",&f);
          if(i==p){qhatsum+=(f-q[i])*(f-q[i]);
        }
      }
      fclose(fp);
      if((fp=fopen("sum.dat","a"))==NULL)
      {
         printf("Cannot write");
         exit();
      }
      fprintf(fp, "%f\n", qhatsum/((float) datanumber));
      fclose(fp);
    }


#include "head.h"
#define N 40
#define INT_N 100
main()
{
  int i,m,p;
  float qhat[11],qbar[11];
  FILE *fp,*gp;
  if((fp=fopen("sum.dat","r"))==NULL)
  {
     printf("Cannot read");
     exit();
  }
```

```
  if((gp=fopen("quo.dat","w"))==NULL)
  {
     printf("Cannot write");
     exit();
  }
  for(i=0;i<11;i++)
     fscanf(fp,"%f",&qhat[i]);
  for(i=0;i<11;i++)
     fscanf(fp,"%f",&qbar[i]);
  fprintf(gp," P  \tqhat    \tqbar    \tqbar/qhat\n");
  for(i=0;i<11;i++)fprintf(gp, "0.%02d\t%f\t%f\t%f\n",
            i==0?5:i==10?95:10*i,qhat[i],qbar[i],qbar[i]/qhat[i]);
  fclose(fp);
  fclose(gp);
}
```

Bibliography

[1] Azzalini, A. (1981) *A Note on the Estimation of a Distribution Function and Quantiles by a Kernel Method. Biometrika*, 68, 326-328.

[2] Bahadur, R.R. (1966) *A Note on Quantiles in Large Samples. The Annals of Mathematical Statistics*, 37, 577-580.

[3] Bickel, P.J. and Doksum, K.A. (1977) *Mathematical Statistics: Basic Ideas and Selected Topics*, San Francisco: Holden-Day.

[4] Billingsley, P. (1979) *Probability and Measure*, New York: John Wiley.

[5] Brewer, K.R.W. (1986) *Likelihood Based Estimation of Quantiles and Density Estimation*, unpublished manuscript.

[6] Chaubey, Y. P. and Sen, P.K. *A Smooth Estimator Of The Survival Function.* Technical Report, Concordia University, Dept. of Mathematics and Statistics.

[7] Chung, K.L. (1974) *A Course in Probability Theory* (2nd ed.), New York: Academic Press.

[8] Csörgõ, M. (1983) *Quantile Processes With Statistical Application*, CBMS-NSF

[9] David, H.A. (1981) *Order Statistics* (2nd ed.), New York: John Wiley.

[10] Devroye, L. and Györfi, L. (1984) *Nonparametric Density Estimation: The L_1 View* New York: John Wiley

[11] Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7, 1-26.

[12] Falk, M. (1984) Relative Deficiency of Kernel Type Estimations of Quantiles. *The Annals of Statistics*, 12, 261-268.

[13] Hall, P. (1983) Orthogonal Series Distribution Function, with Application. *Journal of Royal Statistical Society. B*, 45, 81-88

[14] Hall, P.; Sheather, S.J. (1988) On the Distribution of a Studentized Quantile. *Journal of the Royal Statistical Society*, Ser.B, 50, 381-391.

[15] Harrell, F.E., and Davis, C.E. (1982) A New Distribution-Free Quantile Estimator. *Biometrika*, 69, 635-640.

[16] Kaigh, W.D. (1983) Quantile Interval Estimation. *Communications in Statistics, Part A - Theory and Methods*, 12, 2427-2443.

[17] Kaigh, W.D. (1988) O-Statistics and Their Applications. *Communications in Statistics, Part A - Theory and Methods*, 17, 2191-2210.

[18] Kaigh, W.D., and Cheng, C. (1988) Subsampling Quantile Estimators and Uniformity Criteria, unpublished manuscript.

[19] Kaigh, W.D. and Lachenbruch, P.A. (1982) A Generalized Quantile Estimator. *Communications in Statistics, Part A - Theory and Methods*, 11, 2217-2238.

[20] Luenberger, D.G. (1969) *Optimization by Vector Space Methods*. New York: John Wiley.

[21] Maritz, J.S. and Jarrett, R.G. (1978) A Note on Estimating the Variance of the Sample Median. *Journal of the American Statistical Association*, 73, 194-196.

[22] Müller, H. (1984) Smooth Optimum Kernel Estimators of Densities, Regression Curves and Modes. *The Annals of Statistics*, 12, 766-774.

[23] Padgett, W.J. (1986) *A Kernel-Type Estimator of a Quantile Function From Right-Censored Data. Journal of the American Statistical Association*, 81, 215-222.

[24] Padgett, W.J. and Thombs, L.A. (1986) *Smooth Nonparametric Quantile Estimation Under Censoring: Simulation and Bootstrap Methods. Communications in Statistics, Part B - Simulation and Computation*, 15, 1003-1025.

[25] Parzen, E. (1979) *Nonparametric Statistical Data Modeling. Journal of the American Statistical Association*, 74, 105-131.

[26] Pfanzagl, J. (1976) *Investigating the Quantile of an Unknown Distribution. Contributions to Applied Statistics, Experentia Supplementum*, 22, 111-126.

[27] Ramberg, J.S. Dudewicz, E.J., Tadikamalla, P.R., and Mykytka, E.F. (1979) *A Probability Distribution and Its Uses in Fitting Data. Technometrics*, 21, 201-214.

[28] Reiss, R. D. (1980) *Estimation of Quantiles in Certain Nonparametric Models. The Annals of Statistics*, 8, 87-105.

[29] Sansone, G. (1977) *Orthogonal Functions* Krieger, Huntington, New York

[30] Sheather, S. J. and Marron, J. S. (1990) *Kernel Quantile Estimators. Journal of the American Statistical Association*, 85, 410-416

[31] Stigler, S. M. (1974) *Linear Function of Order Statistics. The Annals of Statistics*, 2, 676-693.

[32] Yang, S.S. (1985) *A smooth Nonparametric Estimation of a Quantile Function. Journal of the American Statistical Association*, 80, 1004-1011

[33] Yashizawa, C.N., Sen, P.K., and Davis, C.E. (1985) *Asymptotic Equivalence of the Harrell-Davis Estimator and the Sample Median. Communications in Statistics, Part A - Theory and Methods*, 14, 2129-2136.