Electronic Journal of Statistics
Vol. 6 (2012) 664–671
ISSN: 1935-7524
DOI: 10.1214/12-EJS687

Asymptotic theorems for kernel U-quantiles

Stefan Ralescu

Queens College
City University of New York

Abstract: For a locally smooth statistical model, we investigate kernel U-quantiles estimators. Under suitable assumptions, we establish a strong Bahadur representation theorem, an invariance principle, and the asymptotic normality for randomly indexed sequences of observations.

AMS 2000 subject classifications: Primary 60K35.
Keywords and phrases: Bahadur representation, asymptotic normality, U-statistics, kernel estimator.

Received October 2011.

1. Introduction

Let $X_1, X_2, \ldots$ be independent random variables having common unknown distribution function (*df*) F. Let $h(x_1, \ldots, x_m)$ be a real-valued measurable function symmetric in its m arguments, and let

$$H(y) = P(h(X_1, \ldots, X_m) \leq y),\ y \in \Re$$

denote the distribution function of the random variable $h(X_1, \ldots, X_m)$. Since the df of the random variable $h(X_1, \ldots, X_m)$ is rarely known exactly, quantiles

$$H^{-1}(p) = \xi_p = \inf\{x\ :\ H(x) \geq p\},\ 0 < p < 1$$

and other features of the *df* H must be estimated from the data. The natural and most widely used estimator of the parameter ξ_p is given by U-quantile $H_n^{-1}(p)$ where for each $n \geq m$ and real y

$$H_n(y) = \frac{1}{n_{(m)}} \sum I(h(X_{i_1}, \ldots, X_{i_m}) \leq y)$$

the sum being taken over the $n_{(m)} = n(n-1)\cdots(n-m+1)$ m-tuples $(i_1, \ldots, i_m)$ of distinct elements from $\{1, 2, \ldots, n\}$ Note that when $h(x) = x$, the empirical *df* of U-statistic structure H_n reduces to the usual empirical *df* and $H_n^{-1}(p)$ becomes the usual pth sample quantile of F. For $m \geq 2$, a second choice of interest is $h(x_1, \ldots, x_m) = \frac{x_1 + x_2 + \cdots + x_m}{m}$ in which case the U-quantile $H_n^{-1}(\frac{1}{2})$ becomes the Hodges-Lehman estimator median $\left(m^{-1}(X_{i_1} + \cdots + X_{i_m})\right)$. Another interesting example corresponds to $h(x_1, x_2) = |x_1 - x_2|$ for which $H_n^{-1}(\frac{1}{2})$ provides an estimator for the spread measure $H^{-1}(\frac{1}{2})$, the median of the distribution of

$|X_1 - X_2|$, where X_1 and X_2 are independent with common distribution function F. The U-quantiles estimators have been investigated, among others, by Serfling [10], Choudhury and Serfling [2], Arcones [1], and Wendler [13]. A clear disadvantage of $H_n^{-1}(p)$ is its poor performance when H is smooth. Estimation of $H^{-1}(p)$ in smooth models plays a fundamental role in many statistical applications, especially in data-analytic and functional statistical methods (see Parzen [8]).

Studies have shown that a smoothed estimator $T_n(p)$ may be preferable to $H_n^{-1}(p)$. First, smoothing reduces the random variation in the data resulting in a more efficient estimator. Second, the "noise level" in the data is reduced by smoothing providing thus an estimator that better displays the interesting features of the *df*. Of the several alternative estimators that have been proposed, we consider the kernel U-quantile estimator

$$T_n(p) = \frac{1}{\alpha_n} \int_0^1 H_n^{-1}(t)\, k\left(\frac{p-t}{\alpha_n}\right) dt \tag{1.1}$$

where α_n is a specified sequence of positive constants (bandwidth) tending to zero and $k(x)$ is a known kernel function. In the case $h(x) = x$, this estimator has been proposed by Parzen [7] and has been studied by Falk [4, 5], Yang [14], Sheather and Marron [11], and Ralescu [9]. For the general case $T_n(p)$ has been investigated by Veraverbeke [12] who established its asymptotic normality. Using the kernel U-quantile estimator brings a clear improvement over the traditional U-quantile when H is differentiable. The size and order of the improvement is usually revealed when studying the Edgeworth expansion of $T_n(p)$ since using one or more terms beyond the normal approximation significantly improves the accuracy for small to moderate samples.

This paper studies further asymptotic properties of the kernel U-quantile estimator. In Section 2 we establish a strong Bahadur representation of $T_n(p)$. In particular, under regularity conditions on (α_n), the a.s. rate $O(n^{-\frac{3}{4}}(\log n)^{\frac{1}{4}})$ is obtained. In Section 3 we prove an invariance principle (functional CLT) for kernel U-quantiles, and in Section 4 we derive the asymptotic normality results for random samples sizes.

2. Asymptotic representation of $T_n(p)$

To study the strong asymptotic representation of $T_n(p)$, the following assumptions are needed:

- (A_1) H has a bounded second derivative in a neighborhood of ξ_p, such that $h(\xi_p) > 0$ where $h = H'$.
- (A_2) k is a density kernel with support included in $[-c, c]$, for some $c > 0$.
- (A_3) $\alpha_n = o(\epsilon_n)$ as $n \to \infty$, where $\epsilon_n \to 0$ in such a way that:

$$\liminf_{n\to\infty} \frac{n\epsilon_n^2}{\log n} > 0$$

The next result provides the Bahadur representation for the kernel U-quantiles.

Theorem 2.1. *If assumptions* (A_1)–(A_3) *are satisfied, then*

$$T_n(p) = \xi_p + \frac{p - H_n(\xi_p)}{h(\xi_p)} + O\left(\max\left(\alpha_n, \epsilon_n^2, \sqrt{\frac{\epsilon_n}{n}}\right)\right) \text{ a.s. as } n \to \infty \quad (2.1)$$

Proof. For any estimator δ_n of ξ_p, set:

$$R(\delta_n) = \delta_n - \xi_p - \frac{p - H_n(\xi_p)}{h(\xi_p)}$$

Let $W_{n,1} \le W_{n,2} \le \cdots \le W_{n,n_{(m)}}$ denote the generalized ordered statistics of the pseudo-sample $h(X_1, \ldots, X_{i_m})$ taken over $n_{(m)}$ m-tuples $(i_1, \ldots, i_m)$ of distinct elements from $\{1, 2, \ldots, n\}$. For real r, let $\lceil r \rceil$ denote the smallest integer greater than or equal to r. Introduce the function $k(i,t) = \lceil it \rceil$ for integer $i \ge 0$ and $0 < t < 1$.

On account of (A_2), for n sufficiently large

$$T_n(p) \in \left[H_n^{-1}(p - c\alpha_n), H_n^{-1}(p + c\alpha_n)\right]$$

Also, with $k_{ni} = k(n_{(m)}, p + (-1)^i c\alpha_n), i = 1, 2$, in view of (A_3), we have by Theorem 3.1 of Choudhuri and Serfling [2] that:

$$R(W_{n,k_{ni}}) = O\left(\max\left(\alpha_n, \epsilon_n^2, \sqrt{\frac{\epsilon_n}{n}}\right)\right) \text{ a.s. as } n \to \infty \quad (2.2)$$

Since for n sufficiently large:

$$R(T_n(p)) \in [R(W_{n,k_{n1}}), R(W_{n,k_{n2}})] \quad (2.3)$$

the conclusion (2.1) follows from (2.2) and (2.3). □

Remark 2.1. For U-quantiles, Choudhury and Serfling [2], and Dehling, Denker, and Philipp [3], obtained the rate $R(H_n^{-1}(p)) = O(n^{-\frac{3}{4}}(\log n)^{\frac{3}{4}})$ a.s. and $n \to \infty$. Arcones [1] proved the exact order

$$R(H_n^{-1}(p)) = O(n^{-\frac{3}{4}}(\log\log n)^{\frac{3}{4}}) \text{ a. s. as } n \to \infty$$

For the Kernel U-quantiles, if $\alpha_n = o(n^{-\frac{3}{4}}(\log n)^{\frac{1}{4}}))$ as $n \to \infty$, Theorem 2.1 gives the a.s. order $R(T_n(p)) = O\left(n^{-\frac{3}{4}}(\log n)^{\frac{1}{4}})\right)$.

Remark 2.2. As another illustration of the strong asymptotic representation (2.1), we deduce the law the iterated logarithm for the Kernel U-quantiles. More specifically, let $v_p^2 = Var[g_p(X_1)] > 0$ where

$$g_p(X_1) = E\{I(h(X_1, \ldots, X_m) \le \xi_p) | X_1\} - p$$

Suppose $\alpha_n = O\left(\sqrt{\frac{\epsilon_n}{n}}\right)$ and $n\epsilon_n^3 = O(1)$ as $n \to \infty$. If assumptions (A_1) and (A_2) hold, then a.s.

$$\limsup_{n\to\infty} \frac{\sqrt{n}(T_n(p) - \xi_p)}{\sqrt{2\log\log n}} = mv_p$$

3. The invariance principle for $T_n(p)$

Here we consider the Donsker type invariance principle for $T_n(p)$. The proof makes use of Theorem 2.1. Let

$$\begin{cases} Y_n(t) = 0 \text{ if } 0 \le t \le \frac{m-1}{n} \\ Y_n(t) = \frac{k(T_k(p)-\xi_p)h(\xi_p)}{mv_p\sqrt{n}} \text{ for } t = \frac{k}{n}, k = m, \dots, n \\ \text{and define } Y_n(t) \text{ by linear interpolation for the other } t \in [0,1] \end{cases} \tag{3.1}$$

We now prove that, for $n \to \infty$, the random function $Y_n(\cdot)$ converges weakly to a standard Brownian motion $W(\cdot)$ in the space $C[0,1]$ of all continuous functions on $[0,1]$ endowed with the uniform topology.

Theorem 3.1. *Let $Y_n(t)$ be given by (3.1). If assumptions (A_1)–(A_3) of Section 2 hold, and if $\alpha_n = o(n^{-\frac{1}{2}})$ as $n \to \infty$, then as $n \to \infty$*

$$Y_n(\cdot) \Longrightarrow W(\cdot) \tag{3.2}$$

Proof. Define the associated process $\{Y_n^*(t)\}_{0\le t\le 1}$ by

$$\begin{cases} Y_n^*(t) = 0 \text{ if } 0 \le t \le \frac{m-1}{n} \\ Y_n^*(\frac{k}{n}) = \frac{k(p-H_k(\xi_p))}{mv_p\sqrt{n}} \text{ for } k = m, \dots, n \\ \text{and for } t \in [\frac{k-1}{n}, \frac{k}{n}] \text{ with } k = m, \dots, n \\ Y_n^*(t) = Y_n^*(\frac{k-1}{n}) + n(t - \frac{k-1}{n})[Y_n^*(\frac{k}{n}) - Y_n^*(\frac{k-1}{n})] \end{cases} \tag{3.3}$$

Since for fixed y, $H_n(y)$ is a U-statistic, by the functional central limit theorem for U-statistics (Miller and Sen [6]) it follows that:

$$Y_n^*(t) \Longrightarrow W(\cdot) \text{ on } (C[0,1], \rho) \tag{3.4}$$

where ρ is the sup-norm in $C[0,1]$.

Therefore, to conclude the proof it suffices to show that for $n \to \infty$,

$$\rho(Y_n, Y_n^*) \xrightarrow{P} 0 \tag{3.5}$$

From Theorem 2.1, we have for $k \ge m$

$$Y_n\left(\frac{k}{n}\right) = \frac{k(p - H_k(\xi_p))}{mv_p\sqrt{n}} + \frac{kR_kh(\xi_p)}{mv_p\sqrt{n}} \tag{3.6}$$

where $R_k = O(\max(\alpha_k, \epsilon_k^2, \sqrt{\frac{\epsilon_k}{k}}))$ a.s. as $k \to \infty$ for any sequence ϵ_n satisfying assumption (A_3). Now, if $\sqrt{n}\alpha_n \to 0$, by taking $\epsilon_n = n^{-\frac{1}{2}+\alpha}$ with $0 < \alpha < \frac{1}{2}$, it is readily seen that $\frac{\alpha_n}{\epsilon_n} \to 0$, $\frac{n\epsilon_n^2}{\log n} \to \infty$, $\sqrt{n}\epsilon_n^2 \to 0$ and so $\sqrt{k}R_k \to 0$ a.s. as $k \to \infty$.

From (3.6), $\rho(Y_n, Y_n^*) = (\max_{m\le k\le n} k|R_k|)h(\xi_p)/mv_p\sqrt{n}$. For each n_0 with $m \le n_0 \le n$:

$$\frac{1}{\sqrt{n}} \max_{m\le k\le n} k|R_k| \le I(n_0, n) + II(n_0) \tag{3.7}$$

where

$$I(n_0, n) = \frac{n_0}{\sqrt{n}} \max_{m\le k\le n_0} |R_k| \text{ and } II(n_0) = \max_{k\ge n_0} \sqrt{k}|R_k|$$

First note that for n_0 fixed,

$$\lim_{n\to\infty} I(n_0, n) = 0 \text{ a.s.} \tag{3.8}$$

Also, since $\sqrt{n}R_n \to 0$ a.s. as $n \to \infty$, it follows that

$$II(n_0) \xrightarrow{P} 0 \text{ as } n_0 \to \infty \tag{3.9}$$

Combining (3.7)–(3.9) we conclude that (3.5) holds and the proof is complete. □

Remark 3.1. From Theorem 3.1 it is clear that the form of the asymptotic variance used by Veraverbeke (1987) is incorrect. In fact, our result shows that as $n \to \infty$

$$\frac{\sqrt{n}(T_n(p) - \xi_p)}{mv_p/h(\xi_p)} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

Remark 3.2. Theorem 3.1 will be used to prove the result presented in the next section. Further applications of the weak convergence

$$Y_n(\cdot) \Longrightarrow W(\cdot)$$

may be obtained as follows:

(a) Consider a sequence $\{r_k\}_{k\ge1}$ of positive real numbers such that $\lim_{k\to\infty} k^{-\frac{1}{2}} r_k = r$, $0 < r < \infty$. Let N_k denote the first time n such that $\sqrt{n}\,(T_n(p) - \xi_p)$ exceeds or reaches r_k. Let $G_k(x) = P\{N_k \le x\}$. Then, if $x_k > 0$ is a sequence that tends to infinity in such a way that $\lim_{k\to\infty} k^{-1}x_k = c > 0$, then on account of Theorem 3.1,

$$\lim_{k\to\infty} G_k(x_k) = \sqrt{\frac{2}{\pi}} \int_{th(\xi_p)\sqrt{c}/v_p}^{\infty} e^{-\frac{s^2}{2}} ds$$

(b) Under the assumptions of Theorem 3.1, the invariance principle implies that for $x > 0$:

$$\lim_{n\to\infty} P\left\{ \max_{m\le k\le n} \frac{k\,(T_k(p) - \xi_p)\,h(\xi_p)}{mv_p\sqrt{n}} > x \right\} = 2(1 - \Phi(x))$$

where Φ denotes the distribution function of the $N(0, 1)$ random variable.

4. Asymptotic normality for randomly indexed sequences of random variables

In many applied models statistical inference is based on a counting random sequence $\{N_k\}_{k\geq 1}$ of nonnegative integer valued random variables. For example, n might be the number of observations obtained within a fixed period of time. Applications connected to studies of randomly indexed samples appear often in queueing problems, insurance and liability applications. The situation is equally important in connection with stopping times arising in sequential tagging. Typically, in a sequential point or interval estimation problem, the sample size is not pre-determined and is itself an integer-valued random variable. For such stochastic sample sizes, the usual asymptotic normality results may require extra regularity conditions and a direct proof might be too involved.

Our next results establishes the asymptotic normality of $T_{N_k}(p)$ for random sample sizes.

Theorem 4.1. *Let $\{N_k\}_{k\geq 1}$ be a sequence of non-negative integer-valued random variables, and $\{n_k\}_{k\geq 1}$ a sequence of positive integers tending to ∞, such that*

$$\frac{N_k}{n_k} \xrightarrow{P} 1 \text{ as } k \to \infty \tag{4.1}$$

Then, under the conditions of Theorem 3.1, we have

$$\sqrt{N_k}\,(T_{N_k}(p) - \xi_p) \xrightarrow{\mathcal{D}} N\left(0, \left[\frac{mv_p}{h(\xi_p)}\right]^2\right) \tag{4.2}$$

Proof. Since $\sqrt{n_k}\,(T_{n_k}(p) - \xi_p)$ converges in distribution to a normal random variable with mean 0 and standard deviation $\frac{mv_p}{h(\xi_p)}$, it suffices to show that:

$$\sqrt{n_k}[T_{N_k}(p) - T_{n_k}(p)] \xrightarrow{P} 0 \text{ as } k \to \infty \tag{4.3}$$

To this end, let $0 < \delta < \frac{1}{2}$. Note that on $\left|\frac{N_k}{n_k} - 1\right| \leq \delta$ we have $N_k < 2n_k$, $\left|\frac{1}{N_k} - \frac{1}{n_k}\right| \leq \delta$ and the following estimate obtains:

$$\begin{aligned}\sqrt{n_k}\,|T_{N_k}(p) - T_{n_k}(p)| \leq a\delta &\left|Y_{2n_k}\left(\frac{1}{2}\right)\right| \\ &+ a\left|Y_{2n_k}\left(\frac{N_k}{2n_k}\right) - Y_{2n_k}\left(\frac{1}{2}\right)\right|\end{aligned} \tag{4.4}$$

with $a = \frac{2^{-\frac{1}{2}} mv_p}{h(\xi_p)}$.

For $\epsilon > 0$, we have the bound

$$P\left\{\sqrt{n_k}\,|T_{N_k}(p) - T_{n_k}(p)| \geq \epsilon\right\} \leq D_{1k} + D_{2k} \tag{4.5}$$

where

$$D_{1k} = P\left\{\left|\frac{N_k}{n_k} - 1\right| > \delta\right\} \text{ and}$$
$$D_{2k} = P\left\{\sqrt{n_k}\,|T_{N_k}(p) - T_{n_k}(p)| \geq \epsilon, \left|\frac{N_k}{n_k} - 1\right| \leq \delta\right\}$$

By assumption D_{1k} can be made arbitrarily small for sufficiently large k. To treat D_{2k}, in view of (4.4) we have:

$$D_{2k} \leq P\left\{\left|Y_{2n_k}\left(\frac{1}{2}\right)\right| \geq \frac{\epsilon}{2a\delta}\right\} + P\left\{\sup_{|t-s|\leq\delta} |Y_{2n_k}(t) - Y_{2n_k}(s)| \geq \frac{\epsilon}{2a}\right\} \quad (4.6)$$

Since $Y_{2n_k}(\frac{1}{2}) \xrightarrow{\mathcal{D}} N(0,2)$, as $k \to \infty$, there exists $\delta > 0$ such that the first term on the right hand side of (4.6) is less than or equal to $\frac{\epsilon}{3}$ for sufficiently large k. On the other hand, from Theorem 3.1, by the tightness property of $Y_{2n_k}(t)$, there exists $\delta > 0$ such that for all k sufficiently large, the second term of the right hand side of (4.6) is less than or equal to $\frac{\epsilon}{3}$.

Therefore, there exists $\delta > 0$, and $k_0 \geq 1$, such that for all $k \geq k_0$,

$$D_{1k} \leq \frac{\epsilon}{3} \text{ and } D_{2k} \leq \frac{2\epsilon}{3}$$

From these estimates and (4.5) we obtain (4.3). This completes the proof of the theorem. □

Remark 4.1. Theorem 4.1 may be used to study the sequential fixed-width confidence intervals for $\xi_p = H^{-1}(p)$ with given required accuracy. More precisely, by appropriately selecting the window-width, we can obtain random intervals I_n constructed from $T_n(p)$ with $length(I_n) \to 0$ with probability 1, as $n \to \infty$ such that for $d > 0$, if the random variable N_d is defined to be the first $n \geq 1$ for which $length(I_n) \leq 2d$, given $0 < \alpha < \frac{1}{2}$, we have:

(i) $\lim_{d\to 0} P\{\xi_p \in I_{N_d}\} = 1 - 2\alpha$ and
(ii) $N_d \approx cd^{-\frac{5}{2}}$ w. p. 1 as $d \to 0$

Details are omitted.

References

[1] Arcones, M. (1996). The Bahadur-Kiefer representation for U-quantiles. *Ann. Statist.* **24**, 1400–1422. MR1401857
[2] Choudhury, J. and R. Serfling (1988). Generalized order statistics, Bahadur representations and sequential nonparametric fixed-width confidence intervals. *J. Statist. Plann. Inference* **19**, 269–282. MR0955393

[3] Dehling, H., M. Denker and W. Philipp (1987). The almost sure invariance principle for the empirical process of U-statistic structure. *Annales de l'I.H.P.* **23**, 121–134. MR0891707
[4] Falk, M. (1984). Relative deficiency of kernel type estimators of quantiles. *Ann. Statistics.* **12**, 261–268. MR0733512
[5] Falk, M. (1985). Asymptotic normality of kernel quantile estimators. *Ann. Statistics.* **13**, 428–433. MR0773180
[6] Miller, R.G.,Jr. and P.K.Sen (1972). Weak convergence of U-statistics and von Mises' differentiable statistical functions. *Ann. Math. Statist.* **43**, 31–41. MR0300321
[7] Parzen, E. (1979). Nonparametric statistical data modeling. *J. Amer. Statist. Assoc.* **74**, 105–131. MR0529528
[8] Parzen, E. (1991). Unification of statistical methods for continuous and discrete data. *Proceedings of Computer Science and Statistics: Interface'90 (C. Page and R. La Page eds.)*, Springer Verlag, N. Y.
[9] Ralescu, S. S. (1995). Strong approximation theorems for integrated kernel quantiles. *Mathematical Methods of Statistics.* **4**, 201–215. MR1335155
[10] Serfling, R. (1984). Generalized L-, M- and R- estimates. *Ann. Statist.* **12**, 76–86. MR0733500
[11] Sheather, S. J. and J. S. Marron(1990). Kernel quantile estimators. *J. Amer. Statist. Assoc.* **85**, 410–416. MR1141741
[12] Veraverbeke, N. (1987). A Kernel-type estimator for generalized quantiles. *Statist. and Probab. Lett.* **5**, 175–180. MR0881191
[13] Wendler, M. (2011). Bahadur representation for U-quantiles of dependent data. *J. Multivariate Anal.* **102**, 1064–1079. MR2793876
[14] Yang, S. S. (1985). A smooth nonparametric estimator of a quantile function. *J. Amer. Statist. Assoc.* **80**, 1004–1011. MR0819607