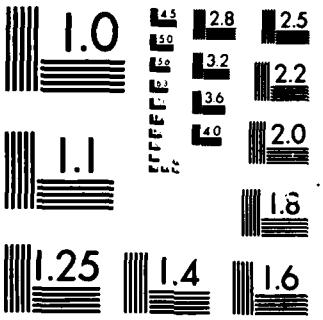
AD-A163 311 A BAHADUR REPRESENTATION FOR QUANTILES OF EMPIRICAL 1/1
DF'S OF GENERALIZED U.. (U) JOHNS HOPKINS UNIV BALTIMORE
MD DEPT OF MATHEMATICAL SCIENCES.. R J SERFLING NOV 85
UNCLASSIFIED TR-453 N00014-79-C-0801 F/G 12/1 NL



DEPARTMENT OF MATHEMATICAL SCIENCES
The Johns Hopkins University
Baltimore, Maryland 21218

A BAHADUR REPRESENTATION FOR QUANTILES OF EMPIRICAL DF'S OF GENERALIZED U-STATISTIC STRUCTURE

by

Robert J. Serfling

Technical Report No. 453
ONR Technical Report No. 85-6
November, 1985

Research supported by the U.S. Department of Navy under Office of Naval Research Contract No. N00014-79-C-0801.

ABSTRACT

A BAHADUR REPRESENTATION FOR QUANTILES OF
EMPIRICAL DF'S OF GENERALIZED U-STATISTIC STRUCTURE

A wide class of c-sample statistics can be represented conveniently as quantiles of a nonclassical empirical df having the structure of a generalized U-statistic. A key tool in studying classical quantiles has been the Bahadur representation. This paper provides a suitable extension of that tool to the generalized setting. As auxiliary lemmas, some new results for generalized U-statistics are developed. Also, some further results for empirical df's of generalized U-statistic structure and their corresponding quantiles are provided.

AMS 1980 *subject classifications*: Primary 60F15, Secondary 62E20

Key words and phrases: empirical distributions, quantiles, Bahadur representation, nonparametric estimation, probability inequalities, U-statistics

1. Introduction

Consider c independent collections of independent observations $\{X_1^{(1)},\ldots,X_{n_1}^{(1)}\},\ldots,\{X_1^{(c)},\ldots,X_{n_c}^{(c)}\}$ taken from df's $F^{(1)},\ldots,F^{(c)}$, respectively. (More generally, the X_i's may be random elements of an arbitrary space.) Let also a "kernel"

$$h(x_1^{(1)},\ldots,x_m^{(1)};\ldots;x_1^{(c)},\ldots,x_{m_c}^{(c)})$$

mapping $\mathbb{R}^{m_1+\ldots+m_c}$ to $\mathbb{R}$ be given, put $\underset{\sim}{F} = (F^{(1)},\ldots,F^{(c)})$, and denote by $H_{\underset{\sim}{F}}$ the df of $h(X_1^{(1)},\ldots,X_{m_1}^{(1)};\ldots;X_1^{(c)},\ldots,X_{m_c}^{(c)})$. An empirical df for estimation of $H_{\underset{\sim}{F}}$ is given, assuming $n_1 \geq m_1,\ldots,n_c \geq m_c$, by

$$(1.1)\quad H_{\underset{\sim}{n}}(y) = \Big[\prod_{j=1}^{c} (n_j)_{(m_j)}\Big]^{-1} \sum I\{h(X_{i_{11}}^{(1)},\ldots,X_{i_{1m_1}}^{(1)};\ldots;\ldots,X_{i_{cm_c}}^{(c)}) \leq y\}, y \in \mathbb{R}$$

where $\underset{\sim}{n} = (n_1,\ldots,n_c)$ and the sum is taken over all $(n_j)_{(m_j)} = n_j(n_j-1)\ldots(n_j-m_j+1)$ m_j-tuples $(i_{j1},\ldots,i_{jm_j})$ of distinct elements from $\{1,\ldots,n_j\}$, $1 \leq j \leq c$.

A wide class of parameters of F may be conveniently represented as $T(H_{\underset{\sim}{F}})$, for some choice of kernel h and for $T(\cdot)$ a suitable functional defined on df's. Such parameters may be estimated naturally by $T(H_{\underset{\sim}{n}})$. For the one-sample case (c=1), such an approach was introduced by Serfling (1984) and asymptotic normality results for $T(H_n)$ were established for $T(\cdot)$ an *L-functional*. Further such results were given by Janssen, Serfling and Veraverbeke (1984) for $T(\cdot)$ a more general type of L-functional.

Here we confine attention to the important special case of *quantile* L-functionals and thus to parameters of the form $\xi_p = H_F^{-1}(p)$ and their corresponding estimators $\hat{\xi}_{pn} = H_n^{-1}(p)$, $0 < p < 1$. For the case c=1 and the kernel $h(x) = x$, this reduces to estimation of the quantiles of F by the usual sample quantiles. For c=1 and the kernel $h(x_1,\ldots,x_m) = m^{-1}(x_1+\ldots+x_m)$, we have $\hat{\xi}_{\frac{1}{2},n}$ = median $\{m^{-1}(X_{i_1}+\ldots+X_{i_m}),\ 1 \le i_1 < \ldots < i_m \le n\}$, which is a *generalized Hodges-Lehmann location estimator*, the classical Hodges-Lehmann estimator corresponding to m=2 and the usual sample median corresponding to m=1. For the case c=1 and $h(x_1,x_2) = |x_1 - x_2|$, $\xi_{\frac{1}{2}}$ is a *spread* parameter discussed by Bickel and Lehmann (1979). For the case c=2, the kernel $h(x_1^{(1)};x_1^{(2)}) = x_1^{(2)} - x_1^{(1)}$ yields the classical two-sample *Hodges-Lehmann shift estimator* (Hodges and Lehmann, 1963), while kernels such as $(x_1^{(2)}+ x_2^{(2)}) - (x_1^{(2)}+ x_2^{(1)})$ yield competing estimators currently under investigation in the literature. Finally, as a general c-sample problem, let us consider *nonparametric analysis of variance*, where we have $F^{(j)}(x) = F(x-\mu_j)$, $1 \le j \le c$, for some unknown df F, and it is of interest to estimate parameters of the form $\theta = \sum_{j=1}^{c} d_j\mu_j$. For the case that θ is a *contrast* ($\sum_1^c d_j = 0$), Lehmann (1963) expressed θ in the form $\sum_{i=1}^{c}\sum_{j=1}^{c} a_{ij}(\mu_i - \mu_j)$ for appropriate constants a_{ij} and proposed the estimator $\hat{\theta}_L = \sum\sum a_{ij}\ \mathrm{med}\{X_k^{(i)} - X_\ell^{(j)}:\ 1 \le k \le n_i, 1 \le \ell < n_j\}$. An interesting literature has developed surrounding this approach, but heretofore the following very natural estimator has not received consideration:

$$(1.2) \qquad \hat{\theta} = \mathrm{median}\{\sum_{j=1}^{c} d_j X_{i_j}^{(j)}:\ 1 \le i_j \le n_j, 1 \le j \le c\}.$$

However, this estimator falls conveniently into the framework formulated above: with kernel $h(x_1^{(1)};x_1^{(2)};\ldots;x_1^{(c)}) = \sum_{j=1}^{c} d_j x_1^{(j)}$, we have $\theta = \xi_{\frac{1}{2}} = H_F^{-1}(\frac{1}{2})$ and $\hat{\theta} = \hat{\xi}_{\frac{1}{2},\underset{\sim}{n}} = H_{\underset{\sim}{n}}^{-1}(\frac{1}{2})$. Moreover, we need not require θ to be a contrast in order to formulate and handle this estimator. (A study of this estimator is currently in progress jointly with David Draper.)

In the case of the classical sample quantiles ($c = 1, h(x) = x$), a key role in obtaining properties of $\hat{\xi}_{pn}$ has been played by the so-called *Bahadur representation* whereby $\hat{\xi}_{pn}$ may be approximated by a sample mean within an error $O(n^{-3/4}(\log n)^{3/4})$ almost surely (see Bahadur (1966), Serfling (1980)). For the case $c=1$ and arbitrary kernel $h(x_1,\ldots,x_m)$, this has been extended by Choudhury and Serfling (1985) and applied to obtain nonparametric sequential fixed-width confidence interval procedures for estimation of parameters $\xi_p = H_F^{-1}(p)$. (This development also extended work of Geertsema (1970) which treated the cases $c=1, h(x) = x$ and $c=2, h(x_1,x_2) = \frac{1}{2}(x_1 + x_2)$.) For $m > 1$, the extended Bahadur representation approximates $\hat{\xi}_{pn}$ by a *U-statistic* (Hoeffding (1948); Serfling (1980)).

The present paper develops a Bahadur representation for $\xi_{p\underset{\sim}{n}}$ in the general c-sample case. For $c > 1$, the approximating term is a *generalized U-statistic*, which comes about because, for each fixed $y \in \mathbb{R}$, the random variable $H_{\underset{\sim}{n}}(y)$ given by (1.1) is itself a generalized U-statistic (Lehmann (1951); Serfling (1980)).

Our treatment requires some new basic results for arbitrary generalized U-statistics, which are presented in Section 2. Our

Bahadur representation theorem is developed in Section 3, along with an auxiliary lemma, of independent interest, giving an exponential probability inequality for a quantile of an empirical df of generalized U-statistic structure, i.e., for $\hat{\xi}_{p\underset{\sim}{n}}$. Some further results on the quantile process $\{H_{\underset{\sim}{n}}^{-1}(p) - H_{\underset{\sim}{F}}^{-1}(p), 0 < p < 1\}$ are provided in Section 4, where also we give some basic convergence results for $H_{\underset{\sim}{n}}(y)$ and $H_{\underset{\sim}{n}}^{-1}(p)$ as $\min(n_1,\ldots,n_c) \to \infty$.

2. A representation and some probability inequalities for generalized U-statistics

As noted in Section 1, the statistic defined by (1.1) is a special case of "generalized U-statistic," which is defined in general as follows. Given c samples and a kernel h as in Section 1, the corresponding generalized U-statistic is given by

$$(2.1) \qquad U_{\underset{\sim}{n}} = \left[\prod_{j=1}^{c} \binom{n_j}{m_j}\right]^{-1} \sum h(X_{i_{11}}^{(1)},\ldots,X_{i_{1m_1}}^{(1)};\ldots;\ldots,X_{i_{cm_c}}^{(c)}).$$

We can represent $U_{\underset{\sim}{n}}$ as an average of (dependent) averages of i.i.d. r.v.'s. Let $k_{\underset{\sim}{n}} = \min\{[n_1/m_1],\ldots,[n_c/m_c]\}$, where $[\cdot]$ denotes greatest integer part. Define the function

$$w(x_1^{(1)},\ldots,x_{n_1}^{(1)};\ldots;x_1^{(c)},\ldots,x_{n_c}^{(c)})$$

$$= k_{\underset{\sim}{n}}^{-1}[h(x_1^{(1)},\ldots,x_{m_1}^{(1)};\ldots;x_1^{(c)},\ldots,x_{m_c}^{(c)}) +$$

$$+ h(x^{(1)}_{m_1+1},\ldots,x^{(1)}_{2m_1};\ldots;x^{(c)}_{m_c+1},\ldots,x^{(c)}_{2m_c})$$

$$+ \ldots + h(x^{(1)}_{k_{\underset{\sim}{n}}m_1-m_1+1},\ldots,x^{(1)}_{k_{\underset{\sim}{n}}m_1};\ldots;\ldots,x^{(c)}_{k_{\underset{\sim}{n}}m_c})].$$

Let $\sum_p$ denote summation over all $\prod_{i=1}^{c}(n_i!)$ within-block permutations $(j_{11},\ldots,j_{1n_1};\ldots;j_{c1},\ldots,j_{cn_c})$ of $(1,\ldots,n_1;\ldots;1,\ldots,n_c)$ and $\sum$ denote summation as in (2.1). Then we easily have

$$k_{\underset{\sim}{n}} \sum_p w(x^{(1)}_{j_{11}},\ldots,x^{(1)}_{j_{1n_1}};\ldots;x^{(c)}_{j_{c1}},\ldots,x^{(c)}_{j_{cn_c}})$$

$$= k_{\underset{\sim}{n}}\left[\prod_{i=1}^{c}(n_i-m_i)!\right] \sum h(x^{(1)}_{i_{11}},\ldots,x^{(1)}_{i_{1m_1}};\ldots;\ldots,x^{(c)}_{i_{cm_c}})$$

and thus

$$(2.2)\quad U_{\underset{\sim}{n}} = \left[\prod_{i=1}^{c}(n_i!)\right]^{-1} \sum_p w(X^{(1)}_{j_{11}},\ldots,X^{(1)}_{j_{in_1}};\ldots;X^{(c)}_{j_{c1}},\ldots,X^{(c)}_{j_{cn_c}}).$$

This expresses $U_{\underset{\sim}{n}}$ as an average of $\prod_{i=1}^{c}(n_i!)$ terms, each of which is itself an average of $k_{\underset{\sim}{n}}$ i.i.d. random variables. This type of representation was introduced in the one-sample case by Hoeffding (1963) for the purpose of developing probability inequalities for U-statistics. We shall apply (2.2) in similar fashion.

LEMMA 2.1. *Let the kernel* h *have finite moment-generating function,*

$$\psi_h(s) = E_{\underset{\sim}{F}}\left\{e^{sh(X^{(1)}_{i_{11}},\ldots;\ldots;\ldots,X^{(c)}_{i_{cm_c}})}\right\} < \infty, 0 \le s \le s_0 \le \infty.$$

Then

$$(2.3) \qquad E_F\{e^{sU_n}\} \le \psi_h^{k_n}(s/k_n), \quad 0 \le s \le s_0 k_n.$$

PROOF. Applying the representation formula (2.2) and Jensen's inequality, we have

$$e^{sU_n} \le \Big[\prod_{i=1}^{c}(n_i!)\Big]^{-1} \sum_p \{\exp sw(X_{j_{11}}^{(1)},\ldots,X_{j_{1n_1}}^{(1)};\ldots;X_{j_{c1}}^{(1)},\ldots,X_{j_{cn_c}}^{(c)})\}.$$

Now taking expectations and applying the independence of the terms in any particular w-sum, we obtain (2.3). □

THEOREM 2.1. *Let the kernel* h *be bounded*: $a \le h \le b$. *Put* $\mu = E_F h$ *and* $\sigma^2 = \mathrm{Var}_F h$. *Then, for* $t > 0$ *and* $k \ge 1$,

$$(2.4) \qquad P\{U_n - \mu \ge t\} \le e^{-2k_n t^2/(b-a)^2}$$

and

$$(2.5) \qquad P\{U_n - \mu \ge t\} \le e^{-k_n t^2/2[\sigma^2 + (b-\mu)t/3]}.$$

PROOF. Follow exactly the proof of Theorem 5.6.1A of Serfling (1980). □

3. A probability inequality for quantiles and a Bahadur representation theorem

We first establish for $\hat{\xi}_{pn}$ an exponential probability inequality,

analogous to that for classical sample quantiles given by Theorem 2.3.2 of Serfling (1980).

THEOREM 3.1. *Let* $0 < p < 1$. *Suppose that* ξ_p *is the unique solution* y *of* $H_{\underset{\sim}{F}}(y-) \leq p \leq H_{\underset{\sim}{F}}(y)$. *Then, for every* $\varepsilon > 0$,

$$P\{|\hat{\xi}_{p\underset{\sim}{n}} - \xi_p| > \varepsilon\} \leq 2e^{-2k_{\underset{\sim}{n}}\delta_\varepsilon^2}, \tag{3.1}$$

where $\delta_\varepsilon = \min\{H_{\underset{\sim}{F}}(\xi_p+\varepsilon) - p, p - H_{\underset{\sim}{F}}(\xi_p-\varepsilon)\}$ *and*

$k_{\underset{\sim}{n}} = \min\{[n_1/m_1],\ldots,[n_c/m_c]\}$.

PROOF. Let $\varepsilon > 0$ and write

$$P\{|\hat{\xi}_{p\underset{\sim}{n}} - \xi_p| > \varepsilon\} = P\{\hat{\xi}_{p\underset{\sim}{n}} > \xi_p + \varepsilon\} + P\{\hat{\xi}_{p\underset{\sim}{n}} < \xi_p - \varepsilon\}.$$

Now

$$P\{\hat{\xi}_{p\underset{\sim}{n}} > \xi_p + \varepsilon\} = P\{p > H_{\underset{\sim}{n}}(\xi_p + \varepsilon)\}$$

$$= P\{\bar{H}_{\underset{\sim}{n}}(\xi_p + \varepsilon) - \bar{H}_{\underset{\sim}{F}}(\xi_p + \varepsilon) > \delta_{1\varepsilon}\},$$

where $\bar{H}_{\underset{\sim}{F}} = 1 - H_{\underset{\sim}{F}}$, $\bar{H}_{\underset{\sim}{n}} = 1 - H_{\underset{\sim}{n}}$, and $\delta_{1\varepsilon} = H_{\underset{\sim}{F}}(\xi_p + \varepsilon) - p$. Note that $\bar{H}_{\underset{\sim}{n}}$ is a generalized U-statistic based on the kernel $\mathbb{1}\{h(\cdot) > y\}$, so that by Theorem 2.1, formula (2.4), we have

$$P\{\hat{\xi}_{p\underset{\sim}{n}} > \xi_p + \varepsilon\} \leq e^{-2k_{\underset{\sim}{n}}\delta_{1\varepsilon}^2}.$$

Similarly we obtain, with $\delta_{2\varepsilon} = P - H_{\underset{\sim}{F}}(\xi_p - \varepsilon)$,

$$P\{\hat{\xi}_{p\underset{\sim}{n}} < \xi_p - \varepsilon\} \leq e^{-2k_{\underset{\sim}{n}}\delta_{2\varepsilon}^2}.$$

Thus (3.1) follows. □

The next result gives conditions under which $\hat{\xi}_{p\underset{\sim}{n}}$ lies within a suitably small neighborhood of ξ_p for all $\underset{\sim}{n}$ "sufficiently large", wp 1. The case c=1,h(x) = x, was given as Lemma 2.5.4B of Serfling (1980) and the case c=1,h(x) arbitrary, as Lemma 3.1 of Choudhury and Serfling (1985).

DEFINITION. An array $\{(n_1,\ldots,n_c)\} \subset \{1,2,\ldots\}^c$ satisfies *Condition A* if

$$\frac{\log \max(n_1,\ldots,n_c)}{\min(n_1,\ldots,n_c)} \to 0 \text{ as } \min(n_1,\ldots,n_c) \to \infty. \tag{3.2}$$

(This is trivially satisfied in the case c=1 and in general is not very restrictive.) The notation "$\min(n_1,\ldots,n_c) \overset{(A)}{\to} \infty$" shall denote restriction under Condition A.

LEMMA 3.1. *Let* $0 < p < 1$. *Suppose that* $H_{\underset{\sim}{F}}$ *is differentiable at* ξ_p, *with* $H'_{\underset{\sim}{F}}(\xi_p) = h_{\underset{\sim}{F}}(\xi_p) > 0$. *Then, under Condition A, with probability* 1

$$|\hat{\xi}_{p\underset{\sim}{n}} - \xi_p| \leq \frac{2(\log n_1 \ldots n_c)^{\frac{1}{2}}}{h_{\underset{\sim}{F}}(\xi_p) k_{\underset{\sim}{n}}^{\frac{1}{2}}}, \text{ for } \min(n_1,\ldots,n_c) \text{ sufficiently large.} \tag{3.3}$$

PROOF. Define δ_ε as in Theorem 3.1 and let ε be given by

$$\varepsilon_{\underset{\sim}{n}} = \frac{2(\log n_1 \ldots n_c)^{\frac{1}{2}}}{h_{\underset{\sim}{F}}(\xi_p) k_{\underset{\sim}{n}}^{\frac{1}{2}}} . \tag{3.4}$$

Then, by a routine argument (as in Serfling (1980), p. 96), we obtain

$$2k_{\underset{\sim}{n}} \delta^2_{\varepsilon_{\underset{\sim}{n}}} \geq 2 \log n_1 n_2 \ldots n_c, \text{ for } \varepsilon_{\underset{\sim}{n}} \text{ sufficiently small.}$$

Hence, by Theorem 3.1,

$$P\{|\hat{\xi}_{p\underset{\sim}{n}} - \xi_p| > \varepsilon_{\underset{\sim}{n}}\} \leq 2(n_1 \ldots n_c)^{-2}, \text{ for } \varepsilon_{\underset{\sim}{n}} \text{ sufficiently small.}$$

Now, by Condition A, it follows that $\varepsilon_{\underset{\sim}{n}} \to 0$ as $\min(n_1, \ldots, n_c) \to \infty$.

Thus

$$\sum_A P\{|\hat{\xi}_{p\underset{\sim}{n}} - \xi_p| > \varepsilon_{\underset{\sim}{n}}\} < \infty, \tag{3.5}$$

where $\sum_A$ denotes summation over $\underset{\sim}{n} = (n_1, \ldots, n_c)$ subject to Condition A. Thus (3.3) follows by the Borel-Cantelli lemma. □

The next lemma plays the key role in the proof of our Bahadur representation theorem. The case c=1,h(x) is due to Bahadur (1966) (also see Serfling (1980), Lemma 2.5.4E) and the case c=1,h arbitrary is covered by Lemma 3.2 of Choudhury and Serfling (1985).

LEMMA 3.2. *Let* $0 < p < 1$. *Suppose that* $H'_{\underset{\sim}{F}}$ *is bounded in a neighborhood of* ξ_p, *with* $H'_{\underset{\sim}{F}}(\xi_p) = h_{\underset{\sim}{F}}(\xi_p) > 0$. *Let* $\{a_{\underset{\sim}{n}}\}$ *be an array of positive constants satisfying*

$$a_{\underset{\sim}{n}} \sim c_0 k_{\underset{\sim}{n}}^{-\frac{1}{2}} (\log n_1 \ldots n_c)^{\frac{1}{2}}, \text{ as } \min(n_1,\ldots,n_c) \to \infty,$$

for some constant $c_0 > 0$. *Put*

$$D_{p\underset{\sim}{n}} = \sup_{|y| \le a_{\underset{\sim}{n}}} |[H_{\underset{\sim}{n}}(\xi_p + y) - H_{\underset{\sim}{n}}(\xi_p)] - [H_{\underset{\sim}{F}}(\xi_p + y) - H_{\underset{\sim}{F}}(\xi_p)]|.$$

Then with probability 1

$$(3.6) \qquad D_{p\underset{\sim}{n}} = O(k_{\underset{\sim}{n}}^{-3/4} (\log n_1,\ldots n_c)^{3/4}), \text{ as } \min(n_1,\ldots,n_c) \overset{(A)}{\to} \infty.$$

PROOF. Our approach is an adaptation of the proof of Lemma 2.5.4E of Serfling (1980), to which one may refer for details omitted here. First, let us note that Conditon A implies that $a_{\underset{\sim}{n}} \to 0$ as $\min(n_1,\ldots,n_c) \to \infty$.

Let $\{b_{\underset{\sim}{n}}\}$ be an array of positive integers such that $b_{\underset{\sim}{n}} \sim c_0 k_{\underset{\sim}{n}}^{\frac{1}{2}} (\log n_1 \ldots n_c)^{\frac{1}{2}}$ as $\min(n_1,\ldots,n_c) \overset{(A)}{\to} \infty$. For integers $r = -b_{\underset{\sim}{n}},\ldots,b_{\underset{\sim}{n}}$, put

$$\eta_{r,\underset{\sim}{n}} = \xi_p + a_{\underset{\sim}{n}} b_{\underset{\sim}{n}}^{-1} r,$$

$$\alpha_{r,n} = H_{\underset{\sim}{F}}(\eta_{r+1,\underset{\sim}{n}}) - H_{\underset{\sim}{F}}(\eta_{r,\underset{\sim}{n}}),$$

and

$$C_{\underset{\sim}{n}}(y) = [H_{\underset{\sim}{n}}(y) - H_{\underset{\sim}{n}}(\xi_p)] - [H_{\underset{\sim}{F}}(y) - H_{\underset{\sim}{F}}(\xi_p)].$$

By monotonicity of $H_{\underset{\sim}{n}}$ and $H_{\underset{\sim}{F}}$, we have

$$D_{pn} \leq K_n + \beta_n,$$

where

$$K_n = \max\{|C_n(\eta_{r,n})|: -b_n \leq r \leq b_n\}$$

and

$$\beta_n = \max\{\alpha_{r,n}: -b_n \leq r \leq b_n\}.$$

Since $\eta_{r+1,n} - \eta_{r,n} = a_n b_n^{-1} \sim k_n^{-3/4}$, we have by the Mean Value Theorem that

$$\beta_n \leq a_n b_n^{-1} \sup_{|y-\xi_p| \leq a_n} |h_F(y)| = O(k_n^{-3/4}), \tag{3.7}$$

as $\min(n_1,..,n_c) \overset{(A)}{\to} \infty$.

We now establish that with probability 1

$$K_n = O(k_n^{-3/4}(\log n_1 \ldots n_c)^{3/4}), \ \min(n_1,\ldots,n_c) \overset{(A)}{\to} \infty. \tag{3.8}$$

For this it suffices by the Borel-Cantelli Lemma to show that

$$\sum_A P\{K_n \geq \gamma_n\} < \infty, \tag{3.9}$$

where $\gamma_n = c_1 k_n^{-3/4}(\log n_1 \ldots n_c)^{3/4}$, with c_1 a positive constant to

be specified below. We will use

$$P\{K_{\underset{\sim}{n}} \geq \gamma_{\underset{\sim}{n}}\} \leq \sum_{r=-b_{\underset{\sim}{n}}}^{b_{\underset{\sim}{n}}} P\{|C_{\underset{\sim}{n}}(\eta_{r,\underset{\sim}{n}})| \geq \gamma_{\underset{\sim}{n}}\}.$$

Now $C_{\underset{\sim}{n}}(\eta_{r,\underset{\sim}{n}})$ is seen to be a generalized U-statistic based on a kernel having mean 0 and variance $P_{r,\underset{\sim}{n}}(1-P_{r,\underset{\sim}{n}})$, where $P_{r,\underset{\sim}{n}} = |H_{\underset{\sim}{F}}(\eta_{r,\underset{\sim}{n}}) - H_{\underset{\sim}{F}}(\xi_p)|$. Therefore, by Theorem 2.1, relation (2.5), we have

$$P\{|C_{\underset{\sim}{n}}(\eta_{r,\underset{\sim}{n}})| \geq \gamma_{\underset{\sim}{n}}\} \leq 2e^{-\theta_{r,\underset{\sim}{n}}}$$

where (by a simple analysis as in Serfling (1980), p. 99)

$$\theta_{r,\underset{\sim}{n}} \geq \frac{c_1^2}{8c_0 h_{\underset{\sim}{F}}(\xi_p)} (\log n_1 \ldots n_c),$$

Uniformly in $|r| \leq b_{\underset{\sim}{n}}$, for $\min(n_1,\ldots,n_c)$ sufficiently large. Thus, for c_1 chosen sufficiently large, we have

$$P\{|C_{\underset{\sim}{n}}(\eta_{r,\underset{\sim}{n}})| \geq \gamma_{\underset{\sim}{n}}\} \leq 2(n_1,\ldots n_c)^{-2}$$

uniformly in $|r| \leq b_{\underset{\sim}{n}}$, and hence

$$P\{K_{\underset{\sim}{n}} \geq \gamma_{\underset{\sim}{n}}\} \leq 6\, b_{\underset{\sim}{n}} (n_1 \ldots n_c)^{-2} = O((n_1 \ldots n_c)^{3/2}),$$

for $\min(n_1,\ldots,n_c)$ sufficiently large. Thus (3.9) follows and hence (3.8) is valid. Combining with (3.7), we have (3.6). □

We now are prepared to establish the following almost sure approximation of $\hat{\xi}_{p\underset{\sim}{n}}$ by a generalized U-statistic, extending Bahadur (1966). (Extension for the case c=1 is given in Choudhury and Serfling (1985).)

THEOREM 3.2. *Let* $0 < p < 1$. *Suppose that* $H_{\underset{\sim}{F}}$ *is twice differentiable at* ξ_p, *with* $H'_{\underset{\sim}{F}}(\xi_p) = h_{\underset{\sim}{F}}(\xi_p) > 0$. *Then*

$$\hat{\xi}_{p\underset{\sim}{n}} = \xi_p + \frac{p - H_{\underset{\sim}{n}}(\xi_p)}{h_{\underset{\sim}{F}}(\xi_p)} + R_{\underset{\sim}{n}} \tag{3.10}$$

where with probability 1

$$R_{\underset{\sim}{n}} = O(k_{\underset{\sim}{n}}^{-3/4}(\log n_1 \ldots n_c)^{3/4}), \quad \min(n_1, \ldots, n_c) \overset{(A)}{\to} \infty \tag{3.11}$$

PROOF. Under the conditions of the theorem, we may apply Lemma 3.1 to obtain

$$H_{\underset{\sim}{F}}(\hat{\xi}_{p\underset{\sim}{n}}) - H_{\underset{\sim}{F}}(\xi_p) = h_{\underset{\sim}{F}}(\xi_p)(\hat{\xi}_{p\underset{\sim}{n}} - \xi_p) + O(k_{\underset{\sim}{n}}^{-1} \log n_1 \ldots n_c)), \tag{3.12}$$

as $\min(n_1, \ldots, n_c) \overset{(A)}{\to} \infty$. By Lemma 3.2 and again appealing to Lemma 3.1, we may pass from (3.12) to: with probability 1

$$H_{\underset{\sim}{n}}(\hat{\xi}_{p\underset{\sim}{n}}) - H_{\underset{\sim}{n}}(\xi_p) = h_{\underset{\sim}{F}}(\xi_p)(\hat{\xi}_{p\underset{\sim}{n}} - \xi_p) + O(k_{\underset{\sim}{n}}^{-3/4}(\log n_1 \ldots n_c)^{3/4}), \tag{3.13}$$

as $\min(n_1, \ldots, n_c) \overset{(A)}{\to} \infty$. Finally, with probability 1 we have

$H_{\underset{\sim}{n}}(\xi_{p\underset{\sim}{n}}) = p + O([\prod_{j=1}^{c}(n_j)_{m_j}]^{-1}) = p + O(k_{\underset{\sim}{n}}^{-1})$, which with (3.13) yields (3.10) and (3.12). □

4. Furthur results

Here we provide a useful estimation of the maximum discrepancy between $\hat{\xi}_{p\underset{\sim}{n}}$ and ξ_p over an interval of p values and some basic convergence results for $H_{\underset{\sim}{n}}(y)$ and $\hat{\xi}_{p\underset{\sim}{n}}$.

THEOREM 4.1. *Let* $0 < t_0 < t_1 < 1$. *Suppose that* $H_{\underset{\sim}{F}}'(y) = h_{\underset{\sim}{F}}(y) > \Delta > 0, y \in (H_{\underset{\sim}{F}}^{-1}(t_0) - \varepsilon, H_{\underset{\sim}{F}}^{-1}(t_1) + \varepsilon)$, *for some* $\varepsilon > 0$. *Then with probability 1*

$$(4.1) \qquad \sup_{t_0<t<t_1} |H_{\underset{\sim}{n}}^{-1}(t) - H_{\underset{\sim}{F}}^{-1}(t)| = O(k_{\underset{\sim}{n}}^{-\frac{1}{2}}(\log n_1 \ldots n_c)^{\frac{1}{2}}), \min(n_1,\ldots,n_c) \overset{(A)}{\to} \infty.$$

PROOF. Let $a_{\underset{\sim}{n}} \sim c_0 k_{\underset{\sim}{n}}^{-\frac{1}{2}}(\log n_1 \ldots n_c)^{\frac{1}{2}}$, for some $c_0 > 0$, and let $\gamma_{\underset{\sim}{n}} \sim c_1 k_{\underset{\sim}{n}}^{-\frac{1}{2}}(\log n_1 \ldots n_c)^{\frac{1}{2}}$, with c_1 to be specified later. Partition (t_0, t_1) into $M = [2/a_{\underset{\sim}{n}}]$ subintervals each of length $\leq a_{\underset{\sim}{n}}$: $t_0 = s_0 < s_1 < \ldots < s_M = t_1$. Then we have

$$(4.2) \qquad \sup_{t_0<t<t_1} |H_{\underset{\sim}{n}}^{-1}(t) - H_{\underset{\sim}{F}}^{-1}(t)| \leq \max_{0\leq k\leq M} |H_{\underset{\sim}{n}}^{-1}(s_k) - H_{\underset{\sim}{F}}^{-1}(s_k)| + \Delta^{-1} a_{\underset{\sim}{n}}.$$

By Theorem 3.1 applied to $|H_{\underset{\sim}{n}}^{-1}(s_k) - H_{\underset{\sim}{F}}^{-1}(s_k)|$, we have

$$P\{|H_{\underset{\sim}{n}}^{-1}(s_k) - H_{\underset{\sim}{F}}^{-1}(s_k)| > \gamma_{\underset{\sim}{n}}\} \leq 2e^{-2k_{\underset{\sim}{n}}\delta_{\underset{\sim}{n}}^2},$$

with $\delta_{\underset{\sim}{n}}^2 = \max\{H_{\underset{\sim}{F}}(H_{\underset{\sim}{F}}^{-1}(s_k) + \gamma_{\underset{\sim}{n}}) - s_k, s_k - H_{\underset{\sim}{F}}(H_{\underset{\sim}{F}}^{-1}(s_k) - \gamma_{\underset{\sim}{n}})\}$.

Now, with $y = H_F^{-1}(s_k)$,

$$|H_F(y \pm \gamma_n) - H_F(y)| \geq \Delta \gamma_n$$

so that $\delta_n^2 \geq \Delta^2 \gamma_n^2$. Thus

$$\begin{aligned} P\{ \max_{0<k<M} |H_n^{-1}(s_k) - H_F^{-1}(s_k)| > \gamma_n \} \\ \leq \sum_{k=0}^{M} P\{|H_n^{-1}(s_k) - H_F^{-1}(s_k)| > \gamma_n\} \\ \leq 8a_n^{-1} e^{-2k_n \Delta^2 \gamma_n^2} \\ \leq 8a_n^{-1}(n_1 \ldots n_c)^{-\Delta^2 c_1^2} \\ = O((n_1 \ldots n_c)^{-2}), \end{aligned}$$

if we take c_1 sufficiently large. It follows by the Borel-Cantelli lemma that with probability 1

$$\max_{0<k<M} |H_n^{-1}(s_k) - H_F^{-1}(s_k)| = O(\gamma_n), \quad \min(n_1, \ldots, n_c) \overset{(A)}{\to} \infty,$$

whence, by (4.2), (4.1) follows. □

REMARK. For the case c=1, Theorem 4.1 yields with probability 1

$$\sup_{t_0<t<t_1} |H_n^{-1}(t) - H_F^{-1}(t)| = O(n^{-\frac{1}{2}}(\log n)^{\frac{1}{2}}), \quad n \to \infty. \tag{4.3}$$

This may be compared with

$$(4.4) \qquad \sup_{t_0<t<t_1} |H_n^{-1}(t) - H_F^{-1}(t)| = O_p(n^{-\frac{1}{2}}),$$

which is Lemma 4.2(b) of Janssen, Serfling and Veraverbeke (1984), given assuming in addition that h_F be Lipschitz continuous on the interval specified in Theorem 4.1. □

We now give some basic convergence results for $H_{\underset{\sim}{n}}(y)$ and $\hat{\xi}_{p\underset{\sim}{n}}$. Recalling that for each fixed y, $H_{\underset{\sim}{n}}(y)$ is a generalized U-statistic, we have with probability 1

$$(4.5) \qquad H_{\underset{\sim}{n}}(y) \to H_{\underset{\sim}{F}}(y), \ \min(n_1,\ldots,n_c) \to \infty,$$

by an SLLN for generalized U-statistics due to Sen(1977), and by Lehmann (1951) (or Serfling (1980)) we have

$$(4.6) \qquad [H_{\underset{\sim}{n}}(y) - H_{\underset{\sim}{F}}(y)]/\sigma_{\underset{\sim}{n}} \xrightarrow{d} N(0,1), \ \min(n_1,\ldots,n_c) \to \infty,$$

where $\sigma_{\underset{\sim}{n}}^2 = \sum_{j=1}^{c} m_j^2 \zeta_{1j}/n_j$ and $\zeta_{11},\ldots,\zeta_{1c}$ are certain positive parameters defined in terms of h and $\underset{\sim}{F}$. We note that (4.5) and (4.6) do not entail Condition A.

For $\hat{\xi}_{p\underset{\sim}{n}}$, matters are somewhat more complicated. One approach is to utilize our Bahadur representation (Theorem 3.2) to approximate $\hat{\xi}_{p\underset{\sim}{n}}$ by (a linear function of) the generalized U-statistic $H_{\underset{\sim}{n}}(\xi_p)$ and simply apply (4.5) and (4.6). For strong convergence, this approach immediately yields that with probability 1

$$(4.7) \qquad \hat{\xi}_{p\underset{\sim}{n}} \to \xi_p, \ \min(n_1,\ldots,n_c) \overset{(A)}{\to} \infty,$$

provided that H_F is twice differentiable at ξ_p with $h_F(\xi_p) > 0$. For asymptotic normality, we obtain that

$$(4.8) \qquad \frac{\hat{\xi}_{p\underset{\sim}{n}} - \xi_p}{\sigma_{\underset{\sim}{n}}/h_{\underset{\sim}{F}}(\xi_p))} \overset{d}{\to} N(0,1), \ \min(n_1,\ldots,n_c) \overset{(A)}{\to} \infty,$$

since trivially $\sigma_{\underset{\sim}{n}} k_{\underset{\sim}{n}}^{3/4} (\log n_1 \ldots n_c)^{-3/4} \to \infty$. However, both (4.7) and (4.8) entail Condition A and second-order differentiability conditions on $H_{\underset{\sim}{F}}$. By other methods avoiding the use of Theorem 3.2, one can obtain (4.7) and (4.8) under slightly milder assumptions. For practical purposes, however, the above results are generally satisfactory, so we shall not pursue these improvements here.

REFERENCES

Bahadur, R.R. (1966), "A note on quantiles in large samples," *Ann. Math. Statist.*, 37, 577-580.

Bickel, P.J. and Lehmann, E.L. (1979), "Descriptive statistics for nonparametric models. IV. Spread," in *Contributions to Statistics. Hájek Memorial Volume* (ed. by J. Jurečková), Academia, Prague, 33-40.

Choudhury, J. and Serfling, R.J. (1985), "Generalized order statistics, Bahadur representations, and sequential nonparametric fixed-width confidence intervals," Technical Report No. 445, Department of Mathematical Sciences, Johns Hopkins University, Baltmore, Maryland.

Geertsema, J.C. (1970), "Sequential confidence intervals based on rank tests," *Ann. Math. Statist.*, 41, 1016-1026.

Hodges, J.L., Jr. and Lehmann, E.L. (1963), "Estimates of location based on rank tests," *Ann. Math. Statist.*, 34, 598-611.

Hoeffding, W. (1948), "A class of statistics with asymptotically normal distribution," *Ann. Math. Statist.*, 19, 293-325.

Hoeffding, W. (1963), "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, 58, 13-30.

Janssen, P., Serfling, R. and Veraverbeke, N. (1984), "Asymptotic normality for a class of statistical functions and applications to measures of spread," *Ann. Statist.*, 12, 1369-1379.

Lehmann, E.L. (1951), "Consistency and unbiasedness of certain nonparametric tests," *Ann. Math. Statist.* 22, 165-179.

Lehmann, E.L. (1963), "Robust estimation in analysis of variance," *Ann. Math. Statist.*, 34, 957-966.

Sen, P.K. (1977), "Almost sure convergence of generalized U-statistics," *Ann. Prob.*, 5, 287-290.

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

Serfling, R.J. (1984), "Generalized L-, M- and R-statistics," *Ann. Statist.*, 12, 76-86.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT CATALOG NUMBER
ONR No. 85-6		

4. TITLE	5. TYPE OF REPORT & PERIOD COVERED
A Bahadur Representation for Quantiles of Empirical DF's of Generalized U-Statistic Structure	Technical Report
	6. PERFORMING ORGANIZATION REPORT NO. Technical Report No. 453

7. AUTHOR(s)	8. CONTRACT OR GRANT NUMBER(s)
Robert J. Serfling	ONR No. N00014-79-C-0801

9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Department of Mathematical Sciences The Johns Hopkins University Baltimore, Maryland 21218	

11. CONTROLLING OFFICE NAME & ADDRESS	12. REPORT DATE
Office of Naval Research Statistics and Probability Program Arlington, Virginia 22217	November, 1985
	13. NUMBER OF PAGES 19

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS (of this report) Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this report)

Approved for public release; distribution unlimited.

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

18. SUPPLEMENTARY NOTES

19. KEY WORDS

empirical distributions, quantiles, Bahadur representation, nonparametric estimation, probability inequalities, U-statistics

20. ABSTRACT

A wide class of c-sample statistics can be represented conveniently as quantiles of a nonclassical empirical df having the structure of a generalized U-statistic. A key tool in studying classical quantiles has been the Bahadur representation. This paper provides a suitable extension of that tool to the generalized setting. As auxiliary lemmas, some new results for generalized U-statistics are developed. Also, some further results for empirical df's of generalized U-statistic structure and their corresponding quantiles are provided.

END

FILMED

2-86

DTIC