AD-A086 372 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
CUMULATIVE PROCESSES: LINEAR COMBINATIONS OF ORDER STATISTICS A--ETC(U)
MAR 80 S LEURGANS DAAG29-75-C-0024
UNCLASSIFIED MRC-TSR-2056 NL

1 OF 1
AD A
086372

END
DATE
FILMED
8-80
DTIC

MRC Technical Summary Report #2056

CUMULATIVE PROCESSES: LINEAR COMBINATIONS OF ORDER STATISTICS AND PERCENTILES

Sue Leurgans

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

March 1980

(Received July 18, 1979)

Sponsored by

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, D.C. 20550

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

CUMULATIVE PROCESSES: LINEAR COMBINATIONS OF ORDER STATISTICS AND PERCENTILES

Sue Leurgans

Technical Summary Report #2056
March 1980

ABSTRACT

Let $\{X_{ni}, i \leq k(n), n \geq 1\}$ be a triangular array of row-wise independent random variables. If $S(X_{n1},\ldots,X_{nj})$ is a statistic based on $X_{n1},\ldots,X_{nj}$, a cumulative process is defined by $S_n(t) = S(X_{n1},\ldots,X_{n\,k(n)t})$. The asymptotic behavior of S_n is determined for S a percentile and for S a smoothly weighted linear combination of order statistics.

AMS (MOS) Subject Classifications - Primary: 60F05, 60G99
Secondary: 62E20, 62E30

Key Words - Invariance theorem, percentiles, linear combinations of order statistics

Work Unit Number 4 - Probability, Statistics, and Combinatorics

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024. This material is based upon work supported by the National Science Foundation under Grant Nos. MCS77-16974 and MCS78-09525.

SIGNIFICANCE AND EXPLANATION

Percentiles and linear combinations of order statistics are statistics which are sometimes preferred to averages because they can be less sensitive to the presence of a few wild observations. It is well known that for large samples, both percentiles and linear combinations of order statistics resemble averages in that the appropriately normalized statistic is approximately normally distributed, with parameters which depend on the underlying distributions.

This paper shows that percentiles and linear combinations of order statistics resemble averages in a stronger sense. It is well known that if a sequence of averages is plotted against the number of observations contributing to the average, and the resulting plot is rescaled appropriately, then for long sequences the picture will act like a realization of a Brownian motion path. This paper establishes that this is still true if the averages are replaced by percentiles or by linear combinations of order statistics.

Although this resemblance may appear to be an abstract probabilistic theorem, these results can be applied to estimation of monotone functions.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

CUMULATIVE PROCESSES: LINEAR COMBINATIONS OF ORDER STATISTICS AND PERCENTILES

Sue Leurgans

1. Introduction and Notation.

Let $\{X_{ni}, 1 \le i \le k(n), n \ge 1\}$ be a triangular array of row-wise independent random variables. Let $S(y_1,\ldots,y_k)$ denote a statistic based on $y_1,\ldots,y_k$. Define the cumulative processes S_n on $[0,1]$ by $S_n(t) = S(X_{n1},\ldots,X_{n,k(n)t})$. For example, if $S(y_1,\ldots,y_k) = y_1 + \ldots + y_k$, then S_n is the familiar cumulative sum processes. It is well known that if $\{\mu_{ni}, 1 \le i \le k(n), n \ge 1\}$ is the triangular array of expected values of the X_{ni}'s and M_n is the deterministic process defined by $M_n(t) = S(\mu_{n1},\ldots,\mu_{n,k(n)t})$, $k(n)^{1/2}(S_n - M_n)$ converges weakly to a Gaussian process which is determined by the variances of the X_{ni}'s. (This conclusion assumes only the existence of the first two moments and the validity of the Lindeberg condition.)

Since, in general, $S_n(t)$ is a statistic for t fixed, the study of the limiting behavior of the process S_n is the study of the limiting behavior of a sequence of statistics. It is natural to suppose that (under appropriate conditions) cumulative percentile processes and cumulative linear function of order statistic processes converge weakly to Gaussian processes. This paper focusses on these two processes under conditions which arise in monotone estimation.

Neither of these processes have certain convenient properties of the cumulative sum processes. The most convenient of these properties is that summing is a linear operation. If S is a linear function of order statistics or a percentile, $S(\underset{\sim}{y} + \underset{\sim}{z}) \ne S(\underset{\sim}{y}) + S(\underset{\sim}{z})$ for arbitrary $\underset{\sim}{y}$ and $\underset{\sim}{z}$. This causes the extension from the i.i.d. case to the independent, nonidentically distributed case to be more complicated for cumulative percentiles and cumulative linear functions of order statistics than for cumulative sums. This paper shows that under suitable conditions, these convenient properties are almost present in the sense that cumulative linear function of order

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024. This material is based upon work supported by the National Science Foundation under Grant Nos. MCS77-16974 and MCS78-09525.

statistic processes and cumulative percentile processes in $C[0,1]$ are asymptotically equivalent to sums of more tractable processes. For smoothly-weighted linear functions of order statistics, these processes are a non-random element of $C[0,1]$ and a cumulative sum process. For percentile processes, a non-random function and an empirical distribution function are used. The empirical distribution function is the cumulative sum obtained by formal substitution of delta-functions in the results for linear functions of order statistics. Unfortunately, this formal substitution does not meet the conditions of Section 2. Section 3 therefore consists of an independent proof for percentile processes. Section 4 compares the conditions imposed here with assumptions other authors have imposed in related problems.

Unless otherwise stated, $\{X_{ni}, 1 \leq i \leq k(n), n \geq 1\}$ will be a triangular array of random variables such that $\{X_{ni}, 1 \leq i \leq k(n)\}$ is (for every n) a set of independent random variables. F_{ni} will be the cumulative distribution function (CDF) of X_{ni}. $\tilde{F}_n$ is defined by $\tilde{F}_n(x) = \sum_{i=1}^{k(n)} F_{ni}(x)/k(n)$. The average CDF $\tilde{F}_n$ can be thought of as the CDF of a "randomly selected" member of $\{X_{ni}, 1 \leq i \leq k(n)\}$. This notation will be extended to $m \leq k(n)$ by defining $\tilde{F}_{n,m}(x) = \sum_{i=1}^{m} F_{ni}(x)/m$. For any set of k numbers, $x_1,\ldots,x_k$; $\{x_1,\ldots,x_k\}_{(m)}$ will denote the m^{th} order statistic of the set of numbers.

W will denote a standard Wiener process, usually in $C[0,1]$. Most of the processes constructed in this paper are to be thought of as members of $C[0,1]$, defined by linear interpolation between a finite set of points. This construction will be left implicit.

Integrals without limits are integrals over $[0,1]$.

Additional notation will be defined as needed. When quantity A is being defined and set equal to the expression B, this will be written either $A := B$ or $B =: A$.

2. *Smoothly Weighted Linear Combinations of Order Statistics*.

Let $X_{n,m;j}$ denote $\{X_{n1},\dots,X_{nm}\}_{(j)}$. Define $k(n)$ weighted sums of order statistics with weight function J by

$$S_{nm} = \sum_{j=1}^{m} J\left(\frac{j}{m+1}\right) X_{n,m;j} \qquad 1 \le m \le k(n)$$

Set $S_{n0} = 0$ and define a random function in $C[0,1]$ by $S_n(m/k(n)) = k(n))^{-1/2} S_{nm}$, $0 \le m \le k(n)$.

The first part of the theorem of this section will be proved under the five regularity assumptions, A1-A5.

A1: J is a real-valued function on the unit interval whose derivative J' exists everywhere on $[0,1]$ and satisfies a Hölder condition for $1/2 < \gamma \le 1$, that is, there is a constant K_L such that $|J'(u) - J'(v)| \le K_L |u - v|^{\gamma}$.

A2: The support of J is a compact subset of the interior of $[0,1]$.

A3: $\lim_{n\to\infty} (k(n))^{1/4} \max_{1\le j\le k(n)} \sup_{-\infty<x<\infty} |F_{nj}(x) - \tilde{F}_n(x)| = 0$.

A4: There exists an open set U *containing the support of* J *and a continuous* CDF F such that $\lim_{n\to\infty} \tilde{F}_n(x) = F(x)$, for all x such that $F(x) \in U$.

A5: $\{\tilde{F}_n, n \ge 1\}$ is a tight collection of cumulative distribution functions.

The second part of Theorem 2.1 imposes two more conditions: one a rate condition and the other limiting discontinuities.

A6: There exists an open set U containing the support of J such that $\overline{\lim}_{n\to\infty} \Delta_n (k(n))^{1/4} < \infty$, where $\Delta_n = \sup_{u\in U} |\tilde{F}_n^{-1}(u) - F^{-1}(u)|$.

A7: $\tilde{F}_n$ and F are strictly increasing on the support of J.

Note that conditions A3, A4, A5, A6 and A7 are trivial if all of the F_{nj} are the same continuous strictly increasing distribution function F.

The first two assumptions involve only the weight function J, and not the random variables. A1 is a smoothness condition for J', which forces J to be continuously differentiable. If $\gamma = 1$, the Hölder condition is a first-order Lipschitz condition.

Since J is a bounded function on a bounded interval it is no loss of generality to require $\gamma \leq 1$. The existence of J' everywhere ensures that the Mean Value Theorem can be applied straight-forwardly. A2 implies that the weighted sum of order statistics trims and guarantees the existence of certain integrals. The trimming implies that the arithmetic mean does not meet these assumptions. The smoothness conditions are not met by percentiles or by trimmed means. The next two assumptions involve the random variables only, and not the weight functions. A3 asserts that, as n gets large, the distributions of the random variables within the n^{th} row of the triangular array approach each other quickly enough that the nonidentical nature of the distributions within each row does not disturb the asymptotic behavior. A4 asserts that the mean distribution functions approach a limiting distribution function, except possibly in the tails. The limiting distribution determines the limiting variance. The tail condition A5 is used to show that the weights $J\left(\frac{j}{m+1}\right)$ can be replaced by $J\left(\frac{j}{m}\right)$. A5 does not require that the F_n converge in the tails, only that mass does not escape to infinity. The first part of Theorem 2.1 does not require any assumptions about the rate of convergence in A4. A6 is just such a condition. A7, which does not involve the tails of the mean CDF, is used to ignore ties.

<u>Theorem 2.1</u>. Under A1-A5, if $\{X_{n1},\ldots,X_{n\,k(n)}\}$ are mutually independent for every n, then

$$\frac{1}{\sigma}\,[S_n(\cdot) - D_n(\cdot) - C_n(\cdot)] \xrightarrow[n\to\infty]{w} W\ ,$$

where

$$\sigma^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} J(F(x))J(F(y))F(\min(x,y))(1-F(\max(x,y))dxdy$$

and D_n and C_n are defined by (1) and (2) respectively.

$$(1)\qquad D_n\left(\frac{m}{k(n)}\right) = \left(\frac{\sqrt{m}}{\sqrt{k(n)}}\right)\left[\int R_n U_{nm}(u)J'(V_{nm}(u))d\Gamma_{nm}(u) + \int R_n(u)J(u)dU_{nm}(u)\right]$$

$$(2)\qquad C_n\left(\frac{m}{k(n)}\right) = \left(\frac{m}{\sqrt{k(n)}}\right)\left[\int R_n(u)J(u)du + \int G(u)J(u)du + \int J(u)\left[u - \sum_{j=1}^{m}\frac{F^*_{nj}(u)}{m}\right]dG(U)\right]\ .$$

G, R_n, F^*_{nj}, Γ_{nm}, U_{nm} and V_{nm} are defined below. If A6 and A7 also hold, then

$$D_n \xrightarrow{P} 0 \ .$$

Sketch of proof of theorem:

The proof is based on rewriting the process S_n as the sum of ten explicit processes. Five of these processes will be grouped to form D_n and C_n. Four of the processes are shown to converge weakly to zero. The remaining process is shown to have a non-degenerate proper weak limit.

Since this proof is necessarily very intricate, most of the details are suppressed. The proof consists of five lemmas.

Throughout this section, G will be used for inverse distribution functions (assumed left-continuous, if there is any ambiguity): $G(u) = F^{-1}(u)$ and $G_n(u) = \tilde{F}_n^{-1}(u)$. Set $R_n(u) = G_n(u) - G(u)$. Thus A6 assumes that $(k(n))^{1/4} \sup_{u \in U} |R_n(u)|$ is bounded.

Transform the X_{ni}'s into $[0,1]$ by $Y_{ni} = \tilde{F}_n(X_{ni})$. Denoting the corresponding distribution functions $F^*_{ni} = F_{ni} \circ G_n$, it follows that $X_{ni} = G_n(Y_{ni})$ with probability one and that $Y_{nm;j} = \tilde{F}_n(X_{nm;j})$. Let Γ_{nm} be the empirical distribution function of $\{Y_{n1},\ldots,Y_{nm}\}$. In the course of this proof, the empirical distribution function Γ_{nm} will be compared with the uniform distribution function, and the difference will be rescaled to have a non-degenerate limiting distribution. Therefore set $U_{nm}(u) = \sqrt{m}\,(\Gamma_{nm}(u) - u)$. In this notation, off a null set, $S_{nm} = m \int J\left(\frac{m}{m+1} \Gamma_{n\,m}(u)\right) G_n(u)\, d\Gamma_{n\,m}(u)$. If G_n is continuous, the null set is empty. Since the failure of the equality on a null set will not affect the weak convergence, this possibility will not be dealt with explicitly in sequel.

Define the ten processes below by linear interpolation between the following points:

$$M_{jn}(0) = 0 \qquad 0 \le j \le 9$$

$$M_{0n}\left(\frac{m}{k(n)}\right) = \frac{m}{\sqrt{k(n)}} \int \left[J\left(\frac{m}{m+1}\,\Gamma_{nm}(u)\right) - J(\Gamma_{nm}(u))\right] G_n(u)\,d\Gamma_{nm}(u) \ ,$$

$$M_{1n}\left(\frac{m}{k(n)}\right) = \frac{\sqrt{m}}{\sqrt{k(n)}} \int U_{nm}(u)\,[J'(V_{nm}(u)) - J'(u)]G(u)\,d\Gamma_{nm}(u) \ .$$

(See Lemma 2.1 for V_{nm}.)

(3) $$M_{2n}\left(\frac{m}{k(n)}\right) = \frac{-1}{2\sqrt{k(n)}} \int U^2_{nm}(u)\,d[G(u)J'(u)] \ .$$

$$M_{3n}\left(\frac{m}{k(n)}\right) = \frac{1}{2m\sqrt{k(n)}} \sum_{j=1}^{m} G(Y_{nj})J'(Y_{nj}) \ .$$

$$M_{4n}\left(\frac{m}{k(n)}\right) = \frac{\sqrt{m}}{\sqrt{k(n)}} \int R_n(u)U_{nm}(u)J'(V_{nm}(u))\,d\Gamma_{nm}(u) \ .$$

(4) $$M_{5n}\left(\frac{m}{k(n)}\right) = \frac{\sqrt{m}}{\sqrt{k(n)}} \int R_n(u)J(u)\,dU_{nm}(u) \ \cdot$$

(5) $$M_{6n}\left(\frac{m}{k(n)}\right) = \frac{m}{\sqrt{k(n)}} \int R_n(u)J(u)\,du \ .$$

$$M_{7n}\left(\frac{m}{k(n)}\right) = \frac{m}{\sqrt{k(n)}} \left[\int J(u)\left[u - \sum_{j=1}^{m} \frac{F^*_{nj}(u)}{m}\right] dG(u)\right] \ .$$

$$M_{8n}\left(\frac{m}{k(n)}\right) = \frac{m}{\sqrt{k(n)}} \int G(u)J(u)\,du$$

$$M_{9n}\left(\frac{m}{k(n)}\right) = \sum_{j=1}^{m} \frac{Z_{nj}}{\sqrt{k(n)}} \ , \quad \text{where}$$

(6) $$Z_{nj} = \int J(u)\,[F^*_{nj}(u) - I_{\{Y_{nj} \le u\}}]\,dG(u) \ .$$

Note that all the component processes are piecewise linear with corners at the same coordinates. M_{0n} arises when the weights $J\left(\frac{i}{m+1}\right)$ are replaced by $J\left(\frac{i}{m}\right)$. M_{1n}, M_{2n} and M_{3n} are the processes defined in Guiahi (1975). The assumptions A1 and A2 imply that the total variation of GJ' is finite and therefore that the integral in (3) is finite. The integrals in (5) to (6) exist and are finite by A1 and A2. D_n and C_n are formed from M_{4n} through M_{8n}.

$$D_n(t) = M_{4n}(t) + M_{5n}(t), \quad 0 \le t \le 1 .$$

$$C_n(t) = M_{6n}(t) + M_{7n}(t) + M_{8n}(t), \quad 0 \le t \le 1 .$$

M_{8n} is the centering constant for asymptotic distributions of linear combinations of order statistics of independent, identically distributed random variables. M_{6n} is the correction necessary if the assumption that the random variables all have the same distribution is relaxed. (See Shorack (1973), Wesley (1977).) M_{7n} can be thought of as a further correction necessary when centering an entire process if the distributions F_{nj}, $1 \le j \le k(n)$, change systematically with j, for each n. Since M_{6n}, M_{7n} and M_{8n} involve non-random quantities only, C_n is a deterministic process.

If the average cumulative distribution functions F_n are all equal, R_n is identically zero. In that case, the processes M_{4n}, M_{5n} and M_{6n} are identically zero. If the F^*_{ni} are all standard uniform distributions, M_{7n} is zero. Therefore, if the random variables X_{nj} all have the same distribution, then M_{4n}, M_{5n}, M_{6n} and M_{7n} all vanish.

Lemma 2.1.

$$S_n(t) = \sum_{j=0}^{9} M_{jn}(t) .$$

The proof of this lemma involves many substitutions, and several integrations by parts. The mean value theorem is applied to J to define functions V_{nm} such that $J(\Gamma_{nm}(u)) = J(u) + (\Gamma_{nm}(u) - u)J'(V_{nm}(u))$ and $V_{nm}(u) = \theta\Gamma_{nm}(u) + (1 - \theta)u$, where u and θ are in $[0,1]$ and $0 \le m \le k(n)$. The details are similar to Moore (1967) and Guiahi (1975).

Lemma 2.2.

If A3 holds, then

a) $$\frac{1}{(k(n))^{1/4}} \max_{1 \le m \le k(n)} \sup_{0 \le u \le 1} |U_{nm}(u)| \xrightarrow[n\to\infty]{P} 0 .$$

If A6 also holds, then

b) $$\Delta_n \max_{0 \le m \le k(n)} \sup_{a_1 \le u \le a_2} |U_{nm}(u)| \xrightarrow[n\to\infty]{P} 0 .$$

The proof of this lemma is similar to that of Lemma 3.5 below. Once the same construction has been established the connection with a sequence of iid uniforms, the law of the iterated logarithm for Kolmogorov-Smirnov distances (see Csaki (1968)) and Lemma 3.2 are used to complete the proof.

Lemma 2.3.

$$(7) \qquad \max_{0\le m\le k(n)} |M_{jn}(m/k(n))| \xrightarrow[n\to\infty]{P} 0 \qquad 0 \le j \le 3 .$$

Proof: The convergence (7) will be established for each j in turn. In order to simplify the notation, we describe the proof of this lemma for $j = 0$ in the case in which the support of J is connected. In this case, A2 and A4 imply that a_1, a_2, b_1 and b_2 can be chosen such that $0 < a_1 < b_1 < b_2 < a_2 < 1$, (a_1, a_2) is an open set satisfying A4 and $[b_1, b_2]$ is the support of J. The trimming assumption and the Lipschitz condition for J (implied by A1) can be used to show tnat

$$\left|J\left(\frac{m}{m+1}\Gamma_{nm}(u)\right) - J(\Gamma_{nm}(u))\right|$$

$$\le |\Gamma_{nm}(u)|\frac{1}{(m+1)} K_L'[I_{\{u\in U\}} + I_{\{\Gamma_{nm}(a_1)>a_1+\varepsilon\}} + I_{\{\Gamma_{nm}(a_2)<a_2-\varepsilon\}}] .$$

The strong law of large numbers for Bernoulli random variables ensures the existence of a fixed $m_0(\varepsilon)$ such that for n sufficiently large,

$$P\{(I_{\{\Gamma_{nm}(a_1)>a_1+\varepsilon\}} + I_{\{\Gamma_{nm}(a_2)<a_2-\varepsilon\}})K_L' \ne 0 \text{ for some } m,\ m_0(\varepsilon) \le M \le k(n)\} < \varepsilon .$$

It follows that with probability at least $1 - \varepsilon$

$$|M_{0n}(t)| \le \frac{K_L'}{\sqrt{k(n)}} [\max_{1\le m\le m_0(\varepsilon)} \int |G_n(u)|d\Gamma_{nm}(u)] + \frac{K_L'}{\sqrt{k(n)}} \sup_{a_1<u<a_2} |G_n(u)| .$$

Because $m_0(\varepsilon)$ is fixed and A3 and A5 imply that the CDF's $\{F_{nj}\ 1 \le j \le k(n), n \ge 1\}$ are tight, the maximum of the $m_0(\varepsilon)$, $\{X_{n1}, \ldots, X_{n,m_0(\varepsilon)}\}$ is uniformly stochastically bounded in n and the first term above converges in probability to zero. A4 implies that $\sup\{G_n(u) : a_1 < u < a_2\}$ is bounded uniformly in n. Therefore with probability $1 - \varepsilon$, $|M_{0n}|$ is had by a random quantity which converges to zero in probability. Since ε can be chosen arbitrarily small, M_{0n} converges to zero weakly in $C[0,1]$.

By the Lipschitz Condition A1 and by the construction of V_{nm}

$$J'(V_{nm}(u)) - J'(u) \le K_L \;_{nm}(u) - u \ , \qquad 0 \le u \le 1 \ ,$$

where K_L is the Lipschitz constant. Substitution of this bound in M_{1n} gives

$$\left|M_{1n}\left(\frac{m}{k(n)}\right)\right| \le \frac{K_L m^{\frac{1-\gamma}{2}}}{\sqrt{k(n)}} \int_{b_1}^{b_2} |U_{nm}(u)|\,|\sqrt{m}\,(\Gamma_{nm}(u) - u)|^{\gamma}|G(u)|\,d\Gamma_{nm}(u)$$

$$\le \frac{K_L(k(n))^{\frac{1-\gamma}{2}}}{(k(n))^{1/2}} \sup_{0\le u\le 1} |U_{nm}(u)|^{1+\gamma} \sup_{b_1\le u\le b_2} |G(u)| \ .$$

The conclusion (7) for $j = 1$ follows immediately from Lemma 2.2, the continuous mapping theorem, and the fact that A1 forces $\gamma - 1$ to be nonpositive.

Similarly, the definition of M_{2n} (3) implies (8):

$$\left|M_{2n}\left(\frac{m}{k(n)}\right)\right| \le \frac{1}{\sqrt{k(n)}} \left[\max_{0\le m\le k(n)} \sup_{0\le u\le 1} U_{nm}^2(u)\right]\left[\frac{V(GJ')}{2}\right] \tag{8}$$

where V denotes total variation. Since A2 implies $V(GJ')$ is finite, Lemma 2.2 and the continuous mapping theorem imply (7) for $j = 2$. By definition of M_{3n},

$$\left|M_{3n}\left(\frac{m}{k(n)}\right)\right| \le \frac{1}{2\sqrt{k(n)}} \max_{0\le m\le k(n)} \sum_{i=1}^{m} \left|\frac{G(Y_{nj})J'(Y_{nj})}{m} I_{\{b_1\le Y_j\le b_2\}}\right| \ .$$

A1 implies that J' is a continuous function on a compact interval, hence a bounded function. A2 implies G is bounded on $[b_1, b_2]$. Therefore each mean inside the maximum, and the maximum itself, is bounded. Since $k(n)$ converges to infinity, (7) holds for $j = 3$. ∎

Lemma 2.4.

$$\frac{1}{\sigma} M_{9n} \xrightarrow[n\to\infty]{w} W \ .$$

Proof of Lemma 2.4:

Recall that M_{9n} is the normalized cumulative sum of $\{Z_{nj}, 1 \le j \le k(n)\}$. By Prohorov's generalization of Donsker's theorem (see Billingsley (1968)

(p. 77, pr. 10.1)), it suffices to show that the Z_{nj}'s satisfy Lindeberg's Condition and that the random functions τ_n converge weakly to the identity function on $[0,1]$, where $\tau_n(t) := \text{Var } M_{9n}(t)/\text{Var } M_{9n}(1)$. Since Z_{nj} depends only on X_{nj}, $\{Z_{n1},\dots,Z_{n\,k(n)}\}$ are mutually independent random variables. A2 implies that expectation and integration can be interchanged to give the following formulae:

$$EZ_{nj} = \int J(u)[F^*_{nj}(u) - EI_{\{Y_{nj}\le u\}}]dG(u) = 0 .$$

$$\sigma^2_{nj} := \text{Var } Z_{nj} = \iint J(u)J(v)[F^*_{nj}(u \wedge v) - F^*_{nj}(u)F^*_{nj}(v)]dG(u)dG(v) .$$

σ^2_{nj} is finite, because $|J(u)J(v)|$ dominates the integrand for all n and j and A2 implies that the dominating function is integrable with respect to $dG(u)dG(v)$. A3 and A4 imply $F^*_{nj}(u \wedge v) - F^*_{nj}(u)F^*_{nj}(v)$ converges to $u \wedge v - uv$ uniformly in j, for u and v in a neighborhood of the support of J. Therefore the Lebesque Dominated Convergence Theorem implies that σ^2_{nj} converges uniformly in j to $\sigma^2 := \iint J(u)J(v)[u \wedge v - uv]dG(u)dG(v)$. The uniformity of this convergence implies that $\sum_{j=1}^{t\,k(n)} \sigma^2_{nj}/k(n)$ converges uniformly to $\sigma^2 t$ and hence that $\tau_n(t)$ converges uniformly to t.

It remains only to verify the Lindeberg Condition for the Z_{nj}'s. Since σ^2_{nj} is of order $k(n)$, it suffices to show that the random variables Z^2_{nj} are uniformly integrable. Since both indicators and probabilities are bounded, the definition of Z_{nj} implies that

$$\int_{\{|Z_{nj}|\ge t\}} Z^2_{nj}dP \le 4P\{|Z_{nj}| \le t\} \iint |J(u)J(v)|dG(u)dG(v) .$$

Chebychev's inequality and the uniform convergence of σ^2_{nj} imply that for n sufficiently large $P\{|Z_{nj}| \ge t\} \le 2\sigma^2 t^{-2}$ uniformly in j. These two inequalities show that the Z^2_{nj} are uniformly integrable and hence that the Z_{nj} satisfy the Lindeberg Condition.

Lemma 2.5.

A1-A7 imply $D_n \xrightarrow[n\to\infty]{P} 0$.

Proof. Since $D_n = M_{4n} + M_{5n}$, it suffices to show M_{4n} and M_{5n} converge in probability to zero. The proof for M_{4n} follows that for M_{0n} in Lemma 2.3 and is omitted.

Expanding the definition of M_{5n} (4) and integrating by parts yields

$$(9) \qquad M_{5n}\left(\frac{m}{k(n)}\right) = \frac{m}{\sqrt{k(n)}}\left[-\sum_{j=1}^{m}\int_{(Y_{nj},1]}\frac{d(R_nJ(u))}{m} + \int ud(R_n\cdot J(u))\right].$$

A7 implies that F_n and F are strictly increasing and that G_n, G and R_n are continuous on the support of J. Therefore, R_nJ is a continuous function, and $(Y_{nj},1]$ can be replaced by $[Y_{nj},1]$ in (9). Collecting terms after this substitution and using A2 and A6 gives

$$\left|M_{5n}\left(\frac{m}{k(n)}\right)\right| \leq \sup_{a_1\leq u\leq a_2}|U_{nm}(u)| \sup_{b_1\leq u\leq b_2}|R_n(u)| \sup_{0\leq u\leq 1}|J(u)|$$

$$\leq \Delta_n \sup_{a_1\leq u\leq a_2}|U_{nm}(u)| \sup|J(u)| .$$

Lemma 2.2 can be extended to imply (in the presence of A6) that $\sup_{a_1\leq u\leq a_2}|U_{nm}(u)|$ converges to zero in probability. Therefore M_{5n} converges weakly to 0 in C[0,1], and the proof of Lemma 2.5 is complete.

3. Percentiles.

Let p be a number between 0 and 1. ξ_n will denote the p^{th} percentile of the average CDF $\tilde{F}_n$, that is, the number such that $\tilde{F}_n(\xi_n) = p$. $\xi_n(m/k(n))$ will denote the p^{th} percentile of $\bar{F}_{n,m}$. $\xi_n(m/k(n))$ will be the p^{th} sample percentile of $\{X_{nj}, 1 \le j \le m\}$ and $\hat{F}_{n,m}$ will be the empirical distribution function of the same set of random variables. Set

$$s_n^2(m/k(n)) = \sum_{i=1}^{m} F_{ni}(\xi_n(m/k(n))(1 - F_{ni}(\xi_n(m/k(n)))) .$$

When no ambiguity will result, $\hat{\xi}_n(1)$, $s_n(1)$ and $\hat{F}_{n,k(n)}$ will be referred to as $\hat{\xi}_n$, s_n and $\hat{F}_n$, respectively. The cumulative process discussed in this section is

$$W_n(t) := \frac{t\ k(n)(\hat{\xi}_n(t) - \xi_n(t))}{(k(n))^{1/2}} .$$

Define

$$V_n(t) := \frac{t\ k(n)}{(k(n))^{1/2}} \left\{\frac{p - \hat{F}_{n,k(n)t}(\xi_n(t))}{f_n(t)}\right\}$$

and $D_n(t) := W_n(t) - V_n(t)$. (See Assumption P2 below for the definition of f_n). Theorem 3.2 below asserts that D_n converges weakly to 0 in $C[0,1]$, and hence that W_n inherits the asymptotic behavior of V_n.

The first piece in the proof of Theorem 3.2 is Theorem 3.1, an extension to triangular arrays of Bahadur's approximation of quantiles. The proof uses Lemma 3.1, an inequality proved in Hoeffding (1956). Corollary 3.1 gives a reduction to i.i.d. sequences. Lemma 3.3 states the finite dimensional distributions of V_n, and Lemma 3.4 gives the tightness of V_n. The proof of Lemma 3.5 uses a convenient version of V_n and W_n. Corollary 3.1 to show that their difference D_n is tight. Theorem 3.2 gives the asymptotic behavior of W_n, the cumulative percentile process.

The following assumptions will be used in the first part of this section:

<u>P1</u>: $\{X_{ni}, 1 \le i \le k(n)\}$ are mutually independent random variables and $X_{ni} \sim F_{ni}$.

P2: $\tilde{F}_{n,k(n)t}$ is continuously differentiable at $\xi_n(t)$ with derivative $f_n(t)$ satisfying $0 < \inf_n f_n(t) \le \sup_n f_n(t) < \infty$.

P3: $\underline{\lim}_{n\to\infty} k(n)/\ln n = \infty$ and the sequence of constants a_n satisfies the two conditions below:

$$\lim_{n\to\infty} \frac{f_n^2(t)a_n^2k(n)}{\ln n} > 2p .$$

$$\overline{\lim}_{n\to\infty} a_n = 0$$

P4: $\underline{\lim}_{n\to\infty} k(n)(\ln n)^{-4} = \infty$

P5: The variances $s_n^2(t)$ satisfy

$$\lim_{n\to\infty} \frac{s_n^2(t)}{k(n)} =: \rho(t) > 0$$

$$\lim_{n\to\infty} \sup_{\substack{0\le t\le 1\\ 0\le s\le 1}} |\xi_n(s) - \xi_n(t)| = 0 .$$

P2 reduces to the usual assumption of a non-zero density in the i.i.d. case. P3 is primarily a pair of growth conditions on the constants a_n used in the approximation (Theorem 3.1). P2 and P4 imply that $a_n = \ln n/((k(n))^{1/2}f_n(t))$ satisfies P3 and that $O((a_n \ln n)^{1/2}) = o(1)$. P5 imposes some regularity conditions on F_{ni} as a function of i. The first condition of P5, although more restrictive then necessary, is appropriate for W_n as defined here. Additional notation will be introduced before Lemma 3.4 and will be used to give a simpler, more restrictive replacement for P5.

Theorem 3.1.

Under the assumptions P1, P2 and P3 for $t = 1$, with probability 1, $D_n(1) = O((a_n \ln n)^{1/2})$.

The conclusion of this theorem can be reexpressed as

$$(k(n))^{1/2}(\hat{\xi}_n - \xi_n) \overset{a.s.}{=} \frac{k(n)p - k(n)\hat{F}_n(\xi_n)}{f_n\sqrt{k(n)}} + O((a_n \ln n)^{1/2})$$

In the case of iid sequences $(F_{ni} = F,\ X_{ni} = X_i,\ k(n) = n,\ a_n = n^{-1/2}\ln n)$, this theorem is due to Bahadur (1966). The result was extended to m-dependent (possibly non-stationary) sequences by Sen (1968). Since the proof of Theorem 3.1 resembles their proofs, only the modifications of Bahadur's proof necessary for this extension will be given.

Lemma 3.1.

If $\{X_i, 1 \le i \le n\}$ are independent random variables taking the value 1 with probability p_i and the value 0 with probability $1 - p_i$, if $S_n := \sum_{i=1}^{n} X_i$, if $np = \sum^{n} p_i$ and if $a < np < b$, then $P\{a < s_n < b\}$ is minimized for $p_i \equiv p$. For a proof of this lemma, see Hoeffding (1956).

Sketch of proof of theorem:

Let b_n denote $(a_n k(n)/\ln n)^{1/2}$, I_n denote the interval $\xi_n \pm a_n$, I_{nr} denote the interval $[\xi_n + r\, a_n/b_n,\ \xi_n + (r+1)a_n/b_n]$ and U_{nr} the interval between ξ_n and $\xi_n + r\, a_n/b_n$. The first step of the proof is to show that, with probability 1,

$$\hat{F}_n(x) = \hat{F}_n(\xi_n) + \tilde{F}_n(x) - \tilde{F}_n(\xi_n) + O((a_n \ln n/k(n))^{1/2})$$

uniformly in I_n, or that

$$H_n = \sup_{x\in I_n} |\hat{F}_n(x) - \hat{F}_n(\xi_n) - (\tilde{F}_n(x) - \tilde{F}_n(\xi_n))| = O((a_n \ln n/k(n))^{1/2}) \ . \tag{10}$$

Simple algebra shows that

$$H_n \le \max_{|r|\le b_n} |\hat{F}_n(J_{nr}) - \tilde{F}_n(J_{nr})| + \max_{|r|\le b_n} |\tilde{F}_n(I_{nr})| \ .$$

The differentiability condition P2 implies the second term is $O(a_n/b_n) = O((a_n \ln n/k(n))^{1/2})$. Using Lemma 3.1, the probability that the first term exceeds any number γ_n can be bounded by the corresponding expression when $F_{ni} \equiv F_n$. This latter probability is itself bounded by a sum of $k(n)$ probabilities that a binomial random variable exceeds $\gamma_n k(n)$. The binomial probabilities are bounded as in Bahadur (1966) and the Borel-Cantelli Lemma is applied (with $\gamma_n = \gamma a_n/b_n$) to complete the proof of (10).

The second step is to show that if $q_n = p\,k(n) + o(k(n)a_n)$, then $\{X_{ni},\ 1 \le i \le k(n)\}_{(q_n)} \in I_n$ for all n sufficiently large with probability 1. This step follows from Lemma 3.1, Assumption P3 and the Borel-Cantelli Lemma. The remainder of the proof of Theorem 3.1 parallels Bahadur (1966).

Lemma 3.2.

If Y_i is a sequence of random variables and $\psi(n)$ a monotone deterministic sequence converging to infinity such that $\overline{\lim\limits_{n\to\infty}}\ Y_n/\psi(n) \le a < \infty$ with probability 1, then, with probability 1,

$$\overline{\lim_{n\to\infty}}\ \{\max_{m\le n} Y_m\}/\psi(n) \le a\ .$$

This lemma is based on a problem in Chung (1974) (p. 237, pr. 2). The proof is omitted.

Corollary 3.1.

If $\{X_i,\ i \ge 1\}$ is a sequence of independent, identically distributed random variables with cumulative distribution function F, if F has a continuous derivative f at its p^{th} percentile ξ, and if $k(n)$ satisfies P4, then, D_n converges weakly in $C[0,1]$ to the zero process and W_n converges weakly to $(p(1-p))^{1/2}W$.

Proof: Since W_n and V_n are based on a single sequence, defining $Y_m := m[(\hat{\xi}_m - \xi) - (p - \hat{F}_m(\xi))]$, it is easy to check that

$$\sup_{0\le t\le 1} |V_n(t) - W_n(t)| = \max_{1\le m\le k(n)} Y_m\ .$$

If a_n is set equal to $(\ln n)(k(n)^{1/2}f(\xi))^{-1}$, P4 implies P3. Theorem 1 therefore implies that with probability 1, $Y_{k(n)} = O(\ln n\ k(n)^{-1/4}) = o(1)$, or equivalently that $Y_{k(n)}(k(n))^{1/4}/\ln n = 0$, a.s. Lemma 2 implies that $\sup\limits_{0\le t\le 1} |W_n(t) - V_n(t)| = o_p(1)$. Thus $W_n - V_n$ converges weakly to the zero process. Because $\xi_n(t) \equiv \xi$, V_n is a normalized cumulative sum process $(V_n(t) = (k(n))^{-1} \sum\limits_{i=1}^{k(n)t} (p - I_{\{X_i \le \xi\}}))$, and Donsker's Theorem implies that V_n converges weakly to $(p(1-p))^{1/2}W$, a non-degenerate limit. Together with Stutsky's Theorem, this implies the same weak limit for W_n.

Lemma 3.3.

Under the assumptions P1, P2, P4 and P5, the finite-dimensional distribution functions of V_n converge to those of $W \circ \rho$. (See P5 for the definition of ρ.)

Proof: $V_n(t)$ can be rewritten as

$$\sum_{i=1}^{t\,k(n)} \frac{[p - I_{\{X_{ni} \leq \xi_n(t)\}}]}{k(n)^{1/2}} = \frac{s_n(t)}{k(n)^{1/2}} \left[\sum_{i=1}^{t\,k(n)} \frac{(p - I_{\{X_{ni} \leq \xi_n(t)\}})}{s_n(t)} \right]$$

The definition of $\xi_n(t)$ implies that $EV_n(t) = 0$. Assumption P5 implies that $s_n^2(t)$, the variance of $\sum_{i=1}^{t\,k(n)} (p - I_{\{X_{ni} \leq \xi_n(t)\}})$, becomes infinite with n. Since each summand is bounded, Lindeberg's condition for triangular arrays (see Billingsley (1968), p. 42) is satisfied trivially for all n sufficiently large, and the term in brackets converges in distribution to the standard normal distribution. P5 therefore implies that $V_n(t)$ converges weakly to a normal distribution with (positive) variance $\rho(t)$.

It remains only to show that $V_n(t_2) - V_n(t_1)$ is asymptotically independent of $V_n(t_1)$. An increment of V_n has the form

$$(11) \quad V_n(t_2) - V_n(t_1) = \frac{\sum_{i=t_1k(n)+1}^{t_2k(n)} p - I_{\{X_{ni} \leq \xi_n(t_2)\}}}{(k(n))^{1/2}} + \frac{\sum_{i=1}^{t_1k(n)} [I_{\{X_{ni} \leq \xi_n(t_2)\}} - I_{\{X_{ni} \leq \xi_n(t_1)\}}]}{(k(n))^{1/2}} .$$

The first term of (11) is independent of $V_n(t_1)$, by assumption P1. The second term is (up to a sign change) a normalized summation of $t_1k(n)$ independent Bernoulli random variables, with parameters $F_{ni}(a_n, b_n]$, where $a_n = \xi_n(t_1) \wedge \xi_n(t_2)$ and $b_n = \xi_n(t_1) \vee \xi_n(t_2)$. The variance of this normalized summation is

$$\sum_{i=1}^{t_1k(n)} \frac{F_{ni}(a_n,b_n)[1 - F_{ni}(a_n,b_n]]}{s_n^2} \leq \frac{k(n)t_1 F_{n\ t_1k(n)}(a_n,b_n]}{s_n^2}$$

$$\sim \frac{k(n)t_1 f_n(t_1)}{s_n^2} \left|\xi_n(t_1) - \xi_n(t_2)\right| \xrightarrow[n\to\infty]{} 0 .$$

The approximation follows from P2 an1 the convergence to zero is a consequence of P2 and P5. Since the variance of the second summation converges to zero, the summation converges weakly to zero, $V_n(t_2) - V_n(t_1)$ is asymptotically independent of $V_n(t_1)$. The extension to the joint distribution of $(W_n(t_1),\ldots,W_n(t_k))$ is routine. ■

Note that if the last condition of P5 is weakened, asymptotic independence does not obtain. If $\xi_n(t)$ exhibits other systematic behavior, a Brownian bridge component may result.

The notation which follows will be used for the rest of this section.

Define the stochastic majorant $\bar{F}_n$ and the stochastic minorant $\bar{F}'_n$ of $\{F_{nj}, 1 \le j \le k(n)\}$ by (12) and (13)

(12) $$\bar{F}_n(t) = \min_{1\le j\le k} F_{nj}(t)$$

(13) $$\bar{F}'_n(t) = \max_{1\le j\le k} F_{nj}(t)$$

Let $\bar{\xi}_n$ and $\bar{\xi}'_n$ be the corresponding percentiles defined by $\bar{F}_n(\bar{\xi}_n) = \bar{F}'_n(\bar{\xi}'_n) = p$. These definitions imply the following inequalities:

(14) $$\bar{F}_n^{-1}(u) \ge F_{ni}^{-1}(u) \ge \bar{F}'^{-1}_n(u) \quad 1 \le i \le k(n),\ 0 \le u \le 1$$

(15) $$\bar{\xi}_n \ge \xi_n(t) \ge \bar{\xi}'_n,\ 0 < t \le 1$$

A sufficient condition for assumption P5 with $\rho(t) \equiv p(1 - p)t$ can now be stated:

<u>P6</u>: There is a positive function H_1, a distribution function F and a sequence of positive numbers δ_n such that H and F have positive derivatives at ξ and the following inequalities hold:

(a) $\bar{F}_n^{-1}(u) \le F^{-1}(u) + \delta_n H(u)$

(b) $\bar{F}_n^{-1}(u) \ge F^{-1}(u) - \delta_n H(u)$

(c) $\varlimsup_{n\to\infty} \delta_n^2 k(n) < \infty$.

P6 is strong enough to imply that $\bar{F}_n$ and $\bar{F}'_n$ satisfy similar constraints in a neighborhood of ξ. In particular, P6 implies

$$(\bar{F}'_n(\bar{\xi}_n) - p) \vee (p - \bar{F}_n(\bar{\xi}'_n)) = O(\delta_n) . \tag{16}$$

One further assumption will also be used:

P7: The derivative of F at ξ, $f(\xi)$ satisfies

$$\lim_{n\to\infty} \sup_{0 \le t \le 1} |f_n(t) - f(\xi)| = 0 .$$

Let $\{U_j, 1 \le j < \infty\}$ be a sequence of independent uniform random variables. By the independence of the X_{nj}'s,

$$\{X_{nj}, 1 \le j \le k(n)\} \overset{d}{=} \{F^{-1}_{nj}(U_j), 1 \le j \le k(n)\} . \tag{17}$$

The symbol * will be used to denote the version of a process based on $\{F^{-1}_{nj}(U_j), 1 \le j \le k(n)\}$. For example,

$$V^*_n(t) = t\left\{ \sum_{i=1}^{t\,k(n)} (p - I_{\{F^{-1}_{ni}(U_i) \le \xi_n(t)\}}) \right\} (k(n)^{1/2}/f_n(t) .$$

By (17), $V^*_N \overset{d}{=} V_n$.

<u>Lemma 3.4.</u>

If P1 and P6 hold, the sequence of probability measures on $C[0,1]$ generated by the sequence of processes V_n is tight.

Proof: V^*_N reduces to

$$V^*_n(t) = \frac{1}{(k(n))^{1/2}} \sum_{i=1}^{t\,k(n)} [p - I_{\{U_i \le p\}} + I_{\{U_i \le p\}} - I_{\{U_i \le F_{ni}(\xi_n(t))\}}]$$

(The statement above will hold only almost surely if for some i, F_{ni} is flat at $\xi_n(t)$). Donsker's Theorem implies that

$$\sum_{i=1}^{t\,k(n)} \frac{(p - I_{\{U_i \le p\}})}{(k(n)p(1 - p))^{1/2}} \xrightarrow[n\to\infty]{w} W .$$

and hence that

$$\frac{1}{(k(n))^{1/2}} \sum_{i=1}^{t\,k(n)} (p - I_{\{U_i \le p\}}) \tag{18}$$

is tight. Set

$$(19) \qquad Y_{ni} = I_{\{U_i \le p\}} - I_{\{U_i \le F_{ni}(\xi_n(t))\}}$$

Since (18) is tight, V_n^* will be tight if

$$\sum_{i=1}^{t\,k(n)} \frac{Y_{ni}}{(k(n))^{1/2}}$$

is tight.

Y_{ni} is 1 with probability $(p - F_{ni}(\xi_n(t)))_+$, -1 with probability $(F_{ni}(\xi_n(t)) - p)_+$, and 0 with probability $1 - |p - F_{ni}(\xi_n(t))|$. The inequality (15) and the definitions (12) and (13) and the approximation (16) imply that there is a finite constant C such that

$$|Y_{ni}| \le I_{\{p-C\delta_n \le U_i \le p+C\delta_n\}} .$$

Therefore

$$\frac{\left|\sum_{i=1}^{k(n)t} Y_{ni}\right|}{(k(n))^{1/2}} \le \sum_{i=1}^{k(n)} \frac{I_{\{p-C\delta_n \le U_i \le p+C\delta_n\}}}{(k(n))^{1/2}} =: Z_n .$$

The variance of Z_n is less than $2k(n)C\delta_n/k(n) = 2C\delta_n$. Since P6 implies δ_n converges to zero, the variance of Z_n converges to zero, Z_n is tight, (19) is tight, and V_n is tight. ∎

Lemma 3.5.

Assumptions P1, P4, P6 and P7 imply that the sequence of probability measures on $C[0,1]$ generated by the sequence of processes D_n is tight.

Proof: Let $\bar{\xi}_N^*(m/k(n))$ denote the p^{th} percentile of $\{\bar{F}_n^{-1}(U_j),\ 1 \le j \le m\}$ and $\bar{F}^*_{n,m}$ the empirical CDF of the same set of random variables. The proof of Lemma 3.5 consists of showing that D_n^* is bounded above and below by the sum of three tight processes, and is therefore tight. Only the proof for the upper bound is given here, since the proof for the lower bound is essentially identical.

Define $D^*_{nm} = (k(n))^{1/2} D^*_n(m/k(n))/m$. This can be reexpressed as

$$D^*_{nm} = \hat{\xi}^*_n(m/k(n)) - \xi_n(m/k(n)) - \frac{p - \hat{F}^*_{nm}(\xi_n(m/k(n))}{f_n(m/k(n))}$$

Since (14) implies that $\bar{\xi}^*_N(m/k(n)) \geq \hat{\xi}^*(m/k(n))$, substitution in $D^*_{n,m}$ (and adding a complicated form of zero) yields

$$D^*_{nm} \leq [\bar{\xi}^*_n(m/k(n)) - \bar{\xi}_n - (\bar{F}_n(\bar{\xi}_n) - \bar{F}^*_{n,m}(\bar{\xi}_n))/f_n(m/k(n))] \tag{20}$$

$$+ [\bar{\xi}_n - \xi_n(m/k(n))] + [(\bar{F}_n(\bar{\xi}_n) - \bar{F}_{nm}(\xi_n(m/k(n))))/f_n(m/k(n))]$$

$$+ [(\hat{F}^*_{n,m}(\xi_n(m/k(n))) - \bar{F}^*_{m,n}(\bar{\xi}_n))/f_n(m/k(n))]$$

$$= \bar{D}^*_{nm} + [\bar{\xi}_n - \xi_n(m/k(n))] + [\hat{F}^*_{nm}(\xi_n(m/k(n))) - \bar{F}^*_{n,m}(\bar{\xi}_n)]/f_n \ , \tag{21}$$

where $\bar{D}^*_{nm}$ denotes the quantity inside the first set of square brackets on the RHS of (20). (The quantity inside the third set of square brackets on the RHS of (20) reduces to zero.) We shall show that each of the terms in (21) generates to a tight process when multiplied by $m(k(n))^{-1/2}$,

The process generated by the second term of (21) is uniformly bounded by $(k(n))^{1/2}(\bar{\xi}_n - \bar{\xi}'_n)$, which is itself bounded by $2(k(n))^{1/2}\delta_n H(u)$. The third condition of P6 ensures that this quantity is bounded, and hence that the second term of (21) generates a tight process.

The convergence of the other two processes will be inferred from the underlying sequence of uniform random variables. Set $Y_m := \{U_j, 1 \leq j \leq m\}_{(mp)}$. For the first process, note that the construction implies

$$\bar{\xi}^*_n(m/k(n)) - \bar{\xi}_n = \bar{F}^{-1}_n(Y_m) - \bar{F}^{-1}_n(p) \ . \tag{22}$$

Assumption P6 implies that the right hand side of (22) is less than $F^{-1}(Y_m) + \delta_n H(Y_m) - (F^{-1}(p) - \delta_n H(p))$. Therefore the process generated by $\bar{D}^*_{nm}$ satisfies the following inequality:

$$(23) \qquad \frac{m\,\bar{D}^*_{nm}}{(k(n))^{1/2}} \le \frac{m}{(k(n))^{1/2}} \left(F^{-1}(Y_m) - F^{-1}(p) - \sum_{i=1}^{m} \frac{(p - I_{\{F^{-1}(U_j) \le F^{-1}(p)\}})}{f(\xi)} \right)$$

$$(24) \qquad + \frac{m\,\delta_n}{(k(n))^{1/2}} (H(Y_m) + H(p))$$

$$(25) \qquad + \frac{1}{(k(n))^{1/2}} \left(\frac{1}{f_n(m/k(n))} - \frac{1}{f(\xi)} \right) \sum_{i=1}^{m} (p - I_{\{F^{-1}(U_j) \le F^{-1}(p)\}})$$

The first term of the bound is of the form of the process D_n of Corollary 3.1. Since the assumptions of Corollary 3.1 hold, this term converges weakly to zero and is consequently tight.

The process determined by (24) is uniformly bounded by $(k(n))^{1/2}\delta_n[\max\{|H(Y_m)| : m \le k(n)\} + |H(p)|]$. Since, *with probability 1*, Y_n *converges with* n to p, for every positive ε, there is an integer N_ε and a number $K_\varepsilon > \varepsilon$ such that $P\{\max\{|Y_n| : n > N_\varepsilon\} < \varepsilon/2\}$ and $P\{\max\{|Y_n| : n \le N_\varepsilon\} \ge K_\varepsilon\} < \varepsilon/2$, the sequence $\{\max[|H(Y(m))| : m \le k(n)], n \ge 1\}$ is tight. Assumption P6 implies that $m\delta_n(k(n))^{-1/2}$ is bounded for $m \le k(n)$ and therefore that the process (24) is tight.

The process generated by (25) contains a cumulative sum of i.i.d. Bernoulli random variables. Since this sum multiplied by $(k(n))^{-1/2}$ converges to a Gaussian process, and since Assumption P7 implies that $(1/f_n(m/k(n)) - 1/f(\xi))$ converges to zero, the process generated by (25) is also tight. Therefore the process corresponding to the left-hand side of (23) is tight.

For the third process, note that

$$(26) \qquad m|\hat{F}^*_{nm}(\xi_n(m/k(n))) - \bar{F}^*_{nm}(\bar{\xi}_n)|\,(k(n))^{-1/2} \le k(n)^{-1/2} \sum_{i=1}^{m} I_{\{\bar{F}_n(\xi_n) < U_i \le p\}} .$$

Therefore the process generated by the left-hand side of (26) is uniformly bounded by the random quantity

$$(27) \qquad k(n)^{-1/2} \sum_{i=1}^{k(n)} I_{\{\bar{F}_n(\xi_n) < U_i \le p\}} .$$

The sum is a binomial random variable with parameters $k(n)$ and $p - \bar{F}_n(\xi_n)$. Thus the variance of the bound (27) is less than $(p - \bar{F}_n(\xi_n))$, which converges to zero. Therefore the process generated by the third term of (21) is tight.

This completes the proof that the process D_n^* has an upper bound which is tight. The fact that D_n^* has a lower bound which is tight is shown in exactly the same manner. Therefcre D_n^* itself is a tight process, and since $D_n^* \overset{d}{=} D_n$, D_n is tight. ■

Theorem 3.2.

If P1, P2, P4, P6 and P7 hold, W_n converges weakly in $C[0,1]$ to $(p(1-p))^{1/2}W$.

Proof: Applying Theorem 3.1 to $\{X_{nj}, j \leq t_i k(n)\}$ for each fixed t_i in $[0,1]$ shows that the finite dimensional distribution functions of D_n are those of the zero process. By Lemma 3.5, D_n converges weakly to zero. Since P6 forces $\rho(t) = p(1-p)t$, Lemmas 3.3 and 3.4 imply that V_n converges weakly to $W \circ \rho = (p(1-p))^{1/2}W$. Since $W_n = V_n + D_n$, the theorem follows.

4. Discussion.

Many papers have considered cumulative processes based on i.i.d. sequences. These include Braun (1976) and Lai (1975) for rank test statistics, Miller and Sen (1972) for U-statistics and von Mises' differentiable statistical functions, and Guiahi (1975) and Ghosh and Sen (1976) for linear combinations of order statistics. All of these authors obtain limiting Gaussian processes. Lamperti (1964) showed that normalized cumulative maximum processes converge weakly to extremal processes. Welsch (1973) is the third in a series of papers extending Lamperti's conclusions to certain strong-mixing Gaussian sequences. However, none of these papers applies to non-trivial triangular arrays.

Some other papers are less closely related than their titles suggest. The strong quantile process approximation of Csorgo and Revesz (for example, Csorgo and Revesz (1978)) is primarily concerned with quantile processes as a function of p, rather than cumulative processes. Their method uses an embedding which is suitable for the i.i.d. case. Guiahi (1975) and Sen (1978) discuss a process based on a tail-sequence of linear combinations of order statistics from an i.i.d. sequence. Sen (1979) derives a Gaussian limit for a process based on $\{X_{(1)},\ldots,X_{(ntp)}\}$. This process differs from the process considered here in that the values of the X's themselves, rather than the values of their indices, determine which random variables are used to construct the process.

The weaker conclusion of asymptotic normality has been studied for percentiles and smooth linear combinations of order statistics in the case of independent, non-identically distributed random variables. The assumptions A1-A7 will therefore be compared with the assumptions of Stigler (1972) and Shorack (1972, 1973), as well as those of Guiahi. The assumptions P1-P7 will be compared with those of Sen (1968) and Weiss (1969).

Theorem 2.1 is an analog of Guiahi's Theorem 1. Guiahi derives his version with three regularity conditions and the assumption that $F_{ni} = F$, for all n and $1 \leq i \leq n$. A3, A4, A5, and A6 are all immediate in this case. The first regularity condition is that the first absolute moment of a random variable with cumulative distribution F be finite. There is no moment condition for F in A1-A7. The second condition is a smoothness condition for the weight function J. This condition is equivalent to A1, with $\gamma = 1$. A2 is not imposed. The third condition is that the total variation of the product of G and J' be finite. This condition is a consequence of A1 and A2. The proof here is similar to that of Guiahi, but the trimming condition A2 is imposed and tail conditions on F are weakened.

Stigler (1974) uses Hajek's projection theorem to derive the asymptotic normality of linear combinations of order statistics. Wesley (1977) discusses Stigler's results, gives a counterexample to the theorem as stated, and indicates a few corrections. (See also Stigler (1979)). Since the theorems in question concern the non-identically distributed case, the conditions are more complicated than Guiahi's. There are two conditions on the distributions: a tightness condition and a convergence condition. The tightness condition is the requirement that there exist a finite number M, a positive constant ε, and a cumulative distribution function H such that (28) and (29) hold.

$$(28) \qquad \begin{aligned} F_{nk}(y) &\le H(y) \qquad y \le -M \\ F_{nk}(y) &\ge H(y) \qquad y \ge M \end{aligned}$$

$$(29) \qquad \lim_{x\to\infty} x^{\varepsilon}(1 - H(x) + H(x)) = 0 \ .$$

The convergence condition is that (30) and (31) hold for almost all x and y.

$$(30) \qquad \lim_{n\to\infty} F_n(x) = F(x)$$

$$(31) \qquad \lim_{n\to\infty} \sum_{k=1}^{n} \frac{[F_{nk}(\min(x,y)) - F_{nk}(x)F_{nk}(y)]}{n} = K(x,y) \ .$$

Condition (30) asserts that the average cumulative distribution F_n converges weakly to F. The assumption (30) is stronger than A4, which is better suited to the trimming assumption in that the convergence is required only for x in a neighborhood of the support of the weight function J. A3 and A4 imply (30) for x and y in a neighborhood of the support of J, with $K(x,y) = F(\min(x,y)) - F(x)F(y)$.

Stigler places two conditions on the weight function J. The first condition is the same as A2, which requires that J trim. The second condition imposes some smoothness conditions on J. J is required to be bounded and to satisfy a Hölder condition for $\gamma > \frac{1}{2}$, except at possibly finitely many points, each of which is a null set for every F_{nk}^{-1}. This condition allows J to have a finite set of discontinuities as long as the corresponding quantiles are well-defined for all the distributions F_{nk}. This is a weaker condition than A1, which assumes J differentiable and imposes the Hölder condition on J'. The necessity of the condition that the discontinuities of J correspond to well-defined quantiles was demonstrated in Stigler (1973), where the limiting distribution of the trimmed mean was shown to be non-normal if the trimming fractions correspond to ill-defined percentiles.

The final stages of Stigler's proof are very similar to the proof given in this chapter. The asymptotic normality is obtained from a normalized sum of the random variables Z'_{nj}, where

$$Z'_{nj} = \int_{-\infty}^{\infty} [F_{nj}(y) - I_{\{X_{nj} \leq y\}}] J(F(y)) dy \ .$$

The random variables used in Section 2 are Z_{nj}, where

$$Z_{nj} = \int J(u) [F^*_{nj}(u) - I_{\{Y_{nj} \leq u\}}] dG(u) \ .$$

If the average cumulative distribution functions F_n are all equal to F and if F is strictly increasing, $Z'_{nj} = Z_{nj}$. Otherwise, the random variables are not necessarily equal.

Since the use of the projection theorem avoids the Mean Value Theorem, the differentiability conditions on J are unnecessary. In his proof that (in the notation of this chapter) $S_n(1)$ is asymptotically equivalent to a normalized sum of Z'_{nj}, Stigler avoids writing out the remainder term explicitly. However, his techniques cannot be extended to show that a corresponding remainder process converges in probability to zero in $C[0,1]$. The methods of Section 2 are also difficult to apply, because the remainder process (from Z'_{nj}) will not decompose conveniently into processes which can be treated individually.

Shorack has derived the asymptotic normality of linear combinations of order statistics from weak convergence of the empirical processes. His theorems apply to more general statistics than Stigler's, allowing finitely many percentiles to receive asymptotically non-negligible weight and including sums of (non-monotone) functions of order statistics. Shorack (1972) contains the basic proofs, but the conditions require that $F_{ni} = F_n$, $1 \le i \le n$. In a later paper (Shorack (1973)), a more general theorem is stated. The proof consists of a list of substitutions in the earlier proof. In the general case, one of two triples of assumptions is required. The first assumption of both triples is that there exist a finite M and a positive δ such that (32) and (33) hold for all u in the unit interval.

(32) $$|G(u)| \le M[u(1 - u)]^{\delta - \frac{1}{2}}$$

(33) $$|G_n(u)| \le M[u(1 - u)]^{\delta - \frac{1}{2}}$$

This condition is a fairly strict tightness condition. Shorack's theorems do not imply Stigler's results, since the Cauchy distribution satisfies (29), but not (32)]. The second basic assumption is that J be continuous, except possibly at ill-defined quantiles. The third condition is that (34) hold, with the same δ as above.

(34) $$\int [t(1 - t)]^{\frac{1-\delta}{2}} d|G_n - G|(t) .$$

The second assumption can be replaced by the condition that J be continuous and the third condition can be replaced by the convergence (35), for all u such that G is continuous at u.

(35) $$\lim_{n \to \infty} G_n(u) = G(u) .$$

Condition (35) implies that the quantiles of F_n converge to the quantiles of F, whenever the latter are well-defined. This condition is equivalent to the weak convergence (30).

Neither Shorack nor Stigler requires a condition resembling A3, which asserts that the F_{nk} approach each other quickly enough in Kolmogorov-Smirnov distance. Indeed, such a condition will not be necessary for asymptotic normality. Consider the situation in which, for every n, half of the F_{nk} equal H_1 and the rest equal H_2. The mean cumulative distribution function F_n will converge to H, the average of H_1 and H_2. The covariance function K will also be simply related to H_1 and H_2. Therefore, if H_1, H_2, and J are sufficiently regular, the weighted linear combination of order statistics will be asymptotically normal. However, a theorem like Theorem 2.1 cannot hold without additional conditions, because the process S_n is not a function of the order statistics of all n random variables, but also depends on the order statistics of $X_{n1},\ldots,X_{nm}$ for every $m \leq n$. The asymptotic behavior of the process will depend on which half of the F_{ni} are equal to H_1. The weak convergence of the process cannot be obtained without additional conditions, such as A3.

Sen (1968) extends Bahadur (1966) to m-dependent sequences. Weiss (1969) uses moment generating function techniques to study the joint asymptotic behavior of several sample percentiles based on a triangular array of independent, non-identically distributed random variables. We examine Sen's conditions for independence, Weiss's condition for a single percentile, and P1-P7 with $k(n) = n$.

The most obvious difference is that Weiss weakens the first condition of P5 (which is only slightly stronger than Sen's assumption that $\inf(s_n^2(1)/n)$ $0)$ to $\lim(n^{2/3}/s_n^2(1)) = 0$. The only other differences between Weiss and Sen are that Sen requires $\tilde{F}_n$ to be twice differentiable and Weiss imposes slightly more complicated bounds on the various density functions. Since neither Weiss nor Sen requires the percentiles ξ_n to converge, Assumption P6 is clearly stronger than the assumptions of Weiss and of Sen. However, P1-P7 avoid their requirement that each F_{ni} have a density. The results of Section 3 apply whenever the average distributions $\tilde{F}_{nm}$ are differentiable in a neighborhood of ξ_n.

Acknowledgements.

The second section of this paper is based on a chapter of the authors Ph.D. dissertation prepared in the Stanford Statistics Department under the guidance of Thomas W. Sager. The dissertation was commenced while the author was an NIH trainee and completed on an NSF Graduate Fellowship.

REFERENCES

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

Billingsley, P. (1968) *Convergence of Probability Measures*. Wiley.

Braun, H. (1976) Weak convergence of sequential linear rank statistics. *Ann. Statist. 4*: 554-575.

Chung, K-L. (1974) *A course in probability*. Second edition. Academic Press.

Csaki, E. (1968) An iterated logarithm law for semi-martingales and its application to the empirical distribution function. *Studia Sci. Math. Hungar 3*: 287-292.

Csörgö, M. and Revesz, P. (1978) Strong approximations of the quantile process. *Ann. Statist. 6*: 882-894.

Ghosh, M. and Sen, P. K. (1976) Asymptotic theory of sequential tests based on linear functions of order statistics. *Essays in Probability and Statistics*, Ogawa Volume (ed., S. Ikeda et.al.) 485-499.

Guiahi, F. (1975) Invariance Principle for Linear Combinations of Order Statistics. Technical Report, Dept. of Statistics, Stanford University.

Hoeffding, Wassily (1956) On the distribution of the number of successes in independent trials. *Ann. Math. Statist. 27*: 713-721.

Lai, T. L. (1975) On Chernoff-Savage statistics and sequential rank tests. *Ann. Statist. 3*: 825-845.

Lamperti, J. (1964) On extreme order statistics. *Ann. Math. Statist. 35*: 1726-1735.

Miller, R. G. and Sen, P. K. (1972) Weak convergence of U-statistics and von Mises' differentiable statistical functions. *Ann. Math. Statist. 43*: 31-41.

Moore, D. S. (1967) An elementary proof of asymptotic normality of linear functions of order statistics. *Ann. Math. Statist. 39*: 263-265.

Sen, P. K. (1968) Asymptotic normality of sample quantiles for m-dependent processes. *Ann. Math. Statist. 39*: 1724-1730.

Shorack, G. (1972) Functions of order statistics. Ann. Math. Statist. 43: 1400-1411.

Shorack, G. (1973) Convergence of reduced empirical and quantile processes with application to functions or order statistics in the non-iid case. Ann. Statist. 1: 146-152.

Stigler, S. M. (1973) The asymptotic distribution of the trimmed mean. Ann. Statist. 1: 472-477.

Stigler, S. M. (1974) Linear functions of order statistics with smooth weight functions. Ann. Statist. 2: 673-693.

Stigler, S. M. (1979) Correction to linear functions of order statistics with smooth weight functions. Ann. Statist. 7: 466.

Weiss, L. (1969) The asymptotic distribution of quantiles from mixed samples. Sankhya A 33: 313-318.

Welsch, R. (1973) A convergence theorem for extreme values from Gaussian sequences. Ann. Probab. 1: 398-404.

Wesley, R. A. (1977) Contributions to the theory of robust regression. Technical Report, Dept. of Statistics, Stanford University.

SL/ed

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
2056

2. GOVT ACCESSION NO.
AD-A086 372

3. RECIPIENT'S CATALOG NUMBER
9 Tech[illegible]

4. TITLE (and Subtitle)
CUMULATIVE PROCESSES: LINEAR COMBINATIONS OF ORDER STATISTICS AND PERCENTILES.

5. [redacted]RIOD COVERED
Summary Report, - no specific reporting period

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Sue Leurgans 14 MRC-TSR-[illegible]

8. CONTRACT OR GRANT NUMBER(s)
NSF-MCS78-09525
15 DAAG29-75-C-0024,
MCS77-16974.

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Mathematics Research Center, University of Wisconsin
610 Walnut Street
Madison, Wisconsin 53706

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Work Unit Number 4 - Probability, Statistics, and Combinatorics

11. CONTROLLING OFFICE NAME AND ADDRESS
(See Item 18 below)

12. REPORT DATE
March 1980

13. NUMBER OF PAGES
29

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
11 Mar 80 12 33

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

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

National Science Foundation
Washington, D.C. 20550

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Invariance theorem, percentiles, linear combinations of order statistics

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

Let $\{X_{ni}, i \leq k(n), n \geq 1\}$ be a triangular array of row-wise independent random variables. If $S(X_{n1},\ldots,X_{nj})$ is a statistic based on $X_{n1},\ldots,X_{nj}$, a cumulative process is defined by $S_n(t) = S(X_{n1},\ldots,X_{n\,k(n)t})$. The asymptotic behavior of S_n is determined for S a percentile and for S a smoothly weighted linear combination of order statistics.