The Annals of Probability
1978, Vol. 6, No. 2, 250–266

PROPERTIES OF THE EMPIRICAL DISTRIBUTION FUNCTION FOR INDEPENDENT NONIDENTICALLY DISTRIBUTED RANDOM VARIABLES[1]

BY MARTIEN C. A. VAN ZUIJLEN

University of Nijmegen

Some fundamental properties of the empirical distribution function are derived in the case of independent but not necessarily identically distributed random variables. The distribution functions of these random variables need not be continuous.

0. Introduction. Let k be a fixed positive integer and for each $N = 1, 2, \cdots$, let $X_{nN} = (X_{1nN}, X_{2nN}, \cdots, X_{knN})$, $n = 1, 2, \cdots, N$, be N mutually independent k-dimensional random vectors with joint distribution functions (df's)

$$F_{nN}(x_1, x_2, \cdots, x_k) = P(X_{1nN} \leqq x_1, X_{2nN} \leqq x_2, \cdots, X_{knN} \leqq x_k), \quad \text{for all} \quad -\infty < x_i < \infty, \quad i = 1, 2, \cdots, k, \tag{0.1}$$

and marginal df's $F_{1nN}, F_{2nN}, \cdots, F_{knN}$, i.e.,

$$F_{inN}(x) = P(X_{inN} \leqq x), \qquad \text{for all} \quad -\infty < x < \infty, \quad i = 1, 2, \cdots, k. \tag{0.2}$$

All random vectors are supposed to be defined on a single probability space $(\Omega, \mathscr{A}, P)$. For each N, moreover, let us define the joint empirical df $\mathbb{F}_N$ of $X_{1N}, X_{2N}, \cdots, X_{NN}$ by taking $N\mathbb{F}_N(x_1, x_2, \cdots, x_k)$ to be the number of elements in the set $\{X_{nN} : X_{1nN} \leqq x_1, X_{2nN} \leqq x_2, \cdots, X_{knN} \leqq x_k, n = 1, 2, \cdots, N\}$, for all $-\infty < x_i < \infty$, $i = 1, 2, \cdots, k$, and the averaged df's $\bar{F}_N$ and $\bar{F}_{iN}$, $i = 1, 2, \cdots, k$, as

$$\bar{F}_N(x_1, x_2, \cdots, x_k) = N^{-1} \sum_{n=1}^{N} F_{nN}(x_1, x_2, \cdots, x_k), \quad \text{for} \quad -\infty < x_i < \infty, \quad i = 1, 2, \cdots, k, \tag{0.3}$$

$$\bar{F}_{iN}(x) = N^{-1} \sum_{n=1}^{N} F_{inN}(x), \qquad \text{for} \quad -\infty < x < \infty. \tag{0.4}$$

We remark that $\bar{F}_N$ has all the properties of a k-variate df and that its marginal df's are the $\bar{F}_{iN}$, $i = 1, 2, \cdots, k$.

The classical theory on empirical df's deals with the case where the N random vectors $X_{1N}, X_{2N}, \cdots, X_{NN}$ are independent and identically distributed (i.i.d.). Our main purpose in this paper is to derive some fundamental properties of the empirical df in the non-i.i.d. case, where the N sample elements are assumed to be independent, but not necessarily identically distributed. In particular, we

Received December 10, 1975; revised March 11, 1977.

[1] Part of the preparation of this manuscript was done while the author was visiting the University of Oregon. Part of the research was done while the author was working at the Statistical Department of the Mathematical Centre in Amsterdam.

AMS 1970 *subject classifications.* Primary 60G17; Secondary 62G30.

Key words and phrases. Empirical distribution function, empirical process.

shall generalize results obtained by Govindarajulu, Le Cam and Raghavachari (1967), Ruymgaart, Shorack and van Zwet (1972), van Zuijlen (1976a), and van Zwet (van Zwet's theorem is published in Ruymgaart (1974)). It is rather striking that most of the theorems considered in the i.i.d. case remain valid in the non-i.i.d. case without any additional condition. Apart from the fact that the authors mentioned above derived these theorems in the i.i.d. case, they also assumed—with the exception of van Zwet—the underlying distribution functions to be continuous. It is our second aim in this paper to give a rigorous demonstration of the fact that, even in the non-i.i.d. case, most of the theorems considered also hold without this assumption.

Sections 1 and 2 deal with univariate and multivariate empirical df's in the case of continuous underlying df's. In Section 3 it will be shown that the continuity assumption is superfluous in almost all theorems derived.

The theorems are useful for proving asymptotic normality of rank statistics in a situation where the multivariate sample elements are allowed to have different df's and where the scores generating functions are allowed to tend to infinity near the boundary of the unit interval and to have a finite number of discontinuities of the first kind. The theorems may also be of interest in their own right.

The basic tools for our study are two related results of Hoeffding (1956), who showed that in a certain sense the non-i.i.d. case in not less favorable than the i.i.d. case, and a theorem of Billingsley (1968), page 94, on fluctuations of partial sums of random variables. We shall present these theorems without proofs.

Suppose that Z_n, $1 \leqq n \leqq N$, are independent random variables (rv's) with

$$P(Z_n = 1) = 1 - P(Z_n = 0) = p_n \,, \tag{0.5}$$

and suppose that

$$0 < \bar{p} = N^{-1} \textstyle\sum_{n=1}^{N} p_n < 1 \,. \tag{0.6}$$

THEOREM 0.1 (Hoeffding). *If*

$$f(k+2) - 2f(k+1) + f(k) > 0 \,, \qquad k = 0, 1, \cdots, N-2 \,, \tag{0.7}$$

then

$$E(f(\textstyle\sum_{n=1}^{N} Z_n)) \leqq \textstyle\sum_{k=0}^{N} f(k)\binom{N}{k}\bar{p}^k(1-\bar{p})^{N-k} \,, \tag{0.8}$$

where equality holds if and only if $p_1 = p_2 = \cdots = p_N = \bar{p}$.

THEOREM 0.2 (Hoeffding). *Let b and c be two integers such that*

$$0 \leqq b \leqq N\bar{p} \leqq c \leqq N \,.$$

Then

$$\textstyle\sum_{n=b}^{c} \binom{N}{n}\bar{p}^n(1-\bar{p})^{N-n} \leqq P(b \leqq \sum_{n=1}^{N} Z_n \leqq c) \leqq 1 \,.$$

Both bounds are attained. The lower bound is attained only if $p_1 = p_2 = \cdots = p_N = \bar{p}$ *unless* $b = 0$ *and* $c = N$.

Let $\xi_1, \cdots, \xi_m$ be random variables which need not be independent or identically distributed. Let $S_k = \xi_1 + \cdots + \xi_k$ ($S_0 = 0$), and put

$$M_m = \max_{0 \leqq k \leqq m} |S_k| \, . \tag{0.9}$$

THEOREM 0.3 (Billingsley). *Suppose that there exist $\gamma \geqq 0$, $\alpha > 1$, and nonnegative numbers $u_1, u_2, \cdots, u_m$ such that*

$$E(|S_j - S_i|^\gamma) \leqq (\textstyle\sum_{l=i+1}^{j} u_l)^\alpha \, , \qquad \text{for} \quad 0 \leqq i \leqq j \leqq m \, . \tag{0.10}$$

Then, there exists a positive number $K = K(\gamma, \alpha)$, only depending on γ and α such that for all $\lambda > 0$,

$$P(M_m \geqq \lambda) \leqq \frac{K}{\lambda^\gamma} (\textstyle\sum_{l=1}^{m} u_l)^\alpha \, . \tag{0.11}$$

1. Properties of the univariate empirical df in the case of continuous underlying distribution functions. In this section we shall deal with the case that $k = 1$, so that for $N = 1, 2, \cdots$, the univariate empirical df $\mathbb{F}_N$ is based on the N random variables $X_{1N}, X_{2N}, \cdots, X_{NN}$, with df's $F_{1N}, F_{2N}, \cdots, F_{NN}$ respectively. Moreover, for the time being we shall assume the underlying df's to be continuous. Before stating the theorems we first have to introduce some further notation.

We recall that the averaged univariate df $N^{-1} \sum_{n=1}^{N} F_{nN}$ is denoted by $\bar{F}_N$. For the set $X_{1N}, X_{2N}, \cdots, X_{NN}$, let us denote the order statistics by

$$X_{1:N} \leqq X_{2:N} \leqq \cdots \leqq X_{N:N} \, . \tag{1.1}$$

Let F be a df on $(-\infty, \infty)$, which is always taken to be right continuous. Define an inverse of this function by

$$F^{-1}(u) = \inf\{y : F(y) \geqq u\} \, , \qquad \text{for} \quad 0 < u \leqq 1 \, , \tag{1.2}$$

whereas $F^{-1}(0)$ is defined to be minus infinity. Here by way of exception a function is introduced which may assume an infinite value. According to (1.2), $F^{-1}(u)$ is nondecreasing, left continuous and satisfies $F(F^{-1}(u)) \geqq u$, for all $0 \leqq u \leqq 1$, with equality if and only if F is continuous. Furthermore it has the property that $F^{-1}(F(y)) \leqq y$, for all $y \in (-\infty, \infty)$, with equality if and only if F is strictly increasing.

We are now in a position to formulate the first theorem, which is immediate from the results obtained in van Zuijlen (1976a).

THEOREM 1.1. *For every $\beta \in (0, 1)$, every array of continuous underlying* df's $F_{1N}, F_{2N}, \cdots, F_{NN}$, $N = 1, 2, \cdots$ *and for every* $N = 1, 2, \cdots$, *we have*

$$P(\mathbb{F}_N(x) \leqq \beta^{-1}\bar{F}_N(x), \ \text{for} \ x \in (-\infty, \infty)) \geqq 1 - \tfrac{2}{3}\pi^2\beta(1-\beta)^{-4} \, , \tag{1.3}$$

$$P(\mathbb{F}_N(x) \geqq \beta\bar{F}_N(x), \ \text{for} \ x \in [X_{1:N}, \infty)) \geqq 1 - \tfrac{2}{3}\pi^2\beta^2(1-\beta)^{-4} \, , \tag{1.4}$$

$$\begin{aligned} P(\mathbb{F}_N(x) \geqq 1 - \beta^{-1}(1 - \bar{F}_N(x)), \ \text{for} \ x \in (-\infty, \infty)) \\ \geqq 1 - \tfrac{2}{3}\pi^2\beta(1-\beta)^{-4} \, , \end{aligned} \tag{1.5}$$

$$\begin{aligned} P(\mathbb{F}_N(x) \leqq 1 - \beta(1 - \bar{F}_N(x)), \ \text{for} \ x \in (-\infty, X_{N:N})) \\ \geqq 1 - \tfrac{2}{3}\pi^2\beta^2(1-\beta)^{-4} \, . \end{aligned} \tag{1.6}$$

PROOF. Immediate from the proof of the theorem in van Zuijlen (1976a) and the equality $\sum_{j=1}^{\infty} j^{-2} = \pi^2/6$.

REMARK 1.1. For $n = 1, 2, \cdots, N$ we introduce the rv's $\tilde{X}_{nN} \equiv \bar{F}_N(X_{nN})$. We denote by $\tilde{F}_{nN}$ the df of $\tilde{X}_{nN}$, and by $\tilde{\mathbb{F}}_N$ the empirical df based on $\tilde{X}_{1N}, \tilde{X}_{2N}, \cdots, \tilde{X}_{NN}$. Following Shorack (1973) we call $\tilde{\mathbb{F}}_N$ the reduced empirical df of $X_{1N}, X_{2N}, \cdots, X_{NN}$. Since the $F_{1N}, F_{2N}, \cdots, F_{NN}$ are assumed to be continuous and are clearly constant on any interval where $\bar{F}_N$ is constant, we have that the $\tilde{F}_{1N}, \tilde{F}_{2N}, \cdots, \tilde{F}_{NN}$ are continuous on [0, 1] and in view of the remark below (1.2) that

$$(1.7) \qquad \tilde{F}_{nN}(t) = F_{nN}(\bar{F}_N^{-1}(t)) \qquad \text{for} \quad t \in [0, 1], \quad n = 1, 2, \cdots, N,$$

and

$$(1.8) \qquad \bar{\tilde{F}}_N(t) \equiv N^{-1} \sum_{n=1}^{N} \tilde{F}_{nN}(t) = N^{-1} \sum_{n=1}^{N} F_{nN}(\bar{F}_N^{-1}(t)) = t, \qquad \text{for} \quad 0 \leqq t \leqq 1.$$

Next, let us prove four lemmas, which may be of independent interest and are used to derive generalizations to the non-i.i.d. case of results obtained in Govindarajulu, Le Cam and Raghavachari (1967), and Ruymgaart, Shorack and van Zwet (1972).

The first lemma supplies upper bounds for the central moments of $\sum_{n=1}^{N} Z_n$, where Z_n, $1 \leqq n \leqq N$, are independent Bernoulli (p_n) rv's defined in (0.5). We recall that $\bar{p} = N^{-1} \sum_{n=1}^{N} p_n$.

LEMMA 1.1. *For every $\alpha > \frac{1}{2}$, there exists $M_\alpha \in (0, \infty)$, such that for $N = 1, 2, \cdots$,*

$$(1.9) \qquad \begin{aligned} E|\textstyle\sum_{n=1}^{N} Z_n - N\bar{p}|^{2\alpha} &\leqq M_\alpha N\bar{p}, && \textit{for} \quad 0 \leqq \bar{p} \leqq N^{-1}, \\ &\leqq M_\alpha\{N\bar{p}(1-\bar{p})\}^\alpha, && \textit{for} \quad N^{-1} \leqq \bar{p} \leqq 1 - N^{-1}, \\ &\leqq M_\alpha N(1-\bar{p}), && \textit{for} \quad 1 - N^{-1} \leqq \bar{p} \leqq 1. \end{aligned}$$

PROOF. Since the assertion is trivial for $\bar{p} = 0$ or $\bar{p} = 1$ and $\alpha > \frac{1}{2}$, Theorem 0.1 ensures that it is sufficient to prove Lemma 1.1 in the case where $p_1 = p_2 = \cdots = p_N = \bar{p} \in (0, 1)$. First let us prove (1.9) for $N^{-1} \leqq \bar{p} \leqq 1 - N^{-1}$. Let $F_{N\bar{p}}(y)$ be the distribution function of $|\sum_{n=1}^{N} Z_n - N\bar{p}|(N\bar{p}(1-\bar{p}))^{-\frac{1}{2}}$. Then using an inequality due to S. N. Bernstein (see, e.g., Bahadur (1966), page 578), we have for $y > 0$ that

$$\begin{aligned} 1 - F_{N\bar{p}}(y) &= P(|\textstyle\sum_{n=1}^{N} Z_n - N\bar{p}| > y(N\bar{p}(1-\bar{p}))^{\frac{1}{2}}) \\ &\leqq 2 \exp\left(\dfrac{-y^2}{2 + 2y/3(N\bar{p}(1-\bar{p}))^{\frac{1}{2}}}\right). \end{aligned}$$

Moreover, for $y \geqq 1$ and $N^{-1} \leqq \bar{p} \leqq 1 - N^{-1}$, $N = 2, 3, \cdots$, we have

$$(N\bar{p}(1-\bar{p}))^{-\frac{1}{2}} \leqq (N/(N-1))^{\frac{1}{2}} \leqq 2^{\frac{1}{2}},$$

so that then

$$1 - F_{N\bar{p}}(y) \leqq 2 \exp\left(\frac{-y^2}{4y}\right) = 2 \exp\left(-\frac{y}{4}\right).$$

Hence, for $N^{-1} \leqq \bar{p} \leqq 1 - N^{-1}$, $N = 2, 3, \cdots$,

$$E\left|\frac{\sum Z_n - N\bar{p}}{(N\bar{p}(1-\bar{p}))^{\frac{1}{2}}}\right|^{2\alpha} = \int_0^\infty y^{2\alpha}\, dF_{N\bar{p}}(y) = 2\alpha \int_0^\infty y^{2\alpha-1}(1 - F_{N\bar{p}}(y))\, dy$$
$$\leqq 2\alpha \int_0^1 dy + 4\alpha \int_1^\infty y^{2\alpha-1} \exp\left(-\frac{y}{4}\right) dy\,,$$

so that (1.9) is proved for $N^{-1} \leqq \bar{p} \leqq 1 - N^{-1}$.

Let us next concentrate on the proof of (1.9) for $0 < \bar{p} \leqq N^{-1}$. For $k = 0, 1, \cdots, N$ we have $P(\sum_{n=1}^N Z_n = k) \leqq (N\bar{p})^k/k!$ and $k \leqq e^k$, so that for $\alpha > \frac{1}{2}$, $0 < \bar{p} \leqq N^{-1}$, $N = 1, 2, \cdots$,

$$E|\textstyle\sum_{n=1}^N Z_n - N\bar{p}|^{2\alpha} = \sum_{k=0}^N |k - N\bar{p}|^{2\alpha} P(\sum_{n=1}^N Z_n = k)$$
$$\leqq \sum_{k=0}^\infty |k - N\bar{p}|^{2\alpha} \frac{(N\bar{p})^k}{k!} \leqq N\bar{p}\left[1 + \sum_{k=1}^\infty \frac{k^{2\alpha}}{k!}\right]$$
$$\leqq N\bar{p}\left[1 + \sum_{k=1}^\infty \frac{e^{2\alpha k}}{k!}\right] = N\bar{p} \sum_{k=0}^\infty \frac{e^{2\alpha k}}{k!}$$
$$= N\bar{p}\exp(\exp(2\alpha))\,.$$

Relation (1.9) for $1 - N^{-1} \leqq \bar{p} \leqq 1$ follows from (1.9) for $0 \leqq \bar{p} \leqq N^{-1}$ by symmetry. □

With $\tilde{\mathbb{F}}_N$ as given in Remark 1.1, we define the reduced empirical process X_N by

$$X_N(t) = N^{\frac{1}{2}}(\tilde{\mathbb{F}}_N(t) - t)\,, \qquad \text{for} \quad 0 \leqq t \leqq 1\,. \tag{1.10}$$

From this definition of X_N and Lemma 1.1 we obtain:

LEMMA 1.2. *For every $\alpha > \frac{1}{2}$ there exists $M_\alpha \in (0, \infty)$ such that for every array of continuous* df's *$F_{1N}, F_{2N}, \cdots, F_{NN}$, $N = 1, 2, \cdots$, every $N = 1, 2, \cdots$ and every pair $s, t \in [0, 1]$,*

$$\begin{aligned} &E|X_N(t) - X_N(s)|^{2\alpha} \\ &\leqq M_\alpha N^{1-\alpha}|t-s|\,, && \text{if } 0 \leqq |t-s| \leqq N^{-1}\,, \\ &\leqq M_\alpha |t-s|^\alpha(1-|t-s|)^\alpha\,, && \text{if } N^{-1} \leqq |t-s| \leqq 1 - N^{-1}\,, \\ &\leqq M_\alpha N^{1-\alpha}(1-|t-s|)\,, && \text{if } 1 - N^{-1} \leqq |t-s| \leqq 1\,. \end{aligned} \tag{1.11}$$

PROOF. Let $\chi(S)$ denote the indicator function of a set S and let $\chi(S; s)$ denote the value of this function at the point s. Without loss of generality take $s < t$. Then,

$$\begin{aligned} E|X_N(t) - X_N(s)|^{2\alpha} &= N^{-\alpha}E|N\tilde{\mathbb{F}}_N(t) - N\tilde{\mathbb{F}}_N(s) - Nt + Ns|^{2\alpha} \\ &= N^{-\alpha}E|\textstyle\sum_{n=1}^N \chi((s, t]; \tilde{X}_{nN}) - N(t-s)|^{2\alpha} \\ &= N^{-\alpha}E|\textstyle\sum_{n=1}^N Z_n - N(t-s)|^{2\alpha}\,, \end{aligned} \tag{1.12}$$

where Z_n, $1 \leqq n \leqq N$, are independent Bernoulli (p_n) rv's, with (see (1.7))

$$p_n = \tilde{F}_{nN}(t) - \tilde{F}_{nN}(s)\,,$$

and hence $\bar{p} = t - s$. Relation (1.11) follows from (1.12) and Lemma 1.1. □

For $0 < \delta \leqq \frac{1}{2}$ we define the function q_δ as

$$q_\delta(t) = \{t(1-t)\}^{\frac{1}{2}-\delta}\,, \qquad \text{for} \quad 0 \leqq t \leqq 1\,. \tag{1.13}$$

Lemma 1.3, which will be derived from Lemma 1.2, tells us what happens with the upper bound in (1.11) if one replaces the process X_N by the process X_N/q_δ. Throughout this paper $\frac{0}{0}$ is defined to be zero.

LEMMA 1.3. *For every $\alpha > \frac{1}{2}$ there exists $\tilde{M}_\alpha \in (0, \infty)$ such that for every array of continuous* df's $F_{1N}, F_{2N}, \cdots, F_{NN}$, $N = 1, 2, \cdots$, *every* $N = 1, 2, \cdots$, *every pair $s, t \in [N^{-1}, 1 - N^{-1}] \cup \{0\} \cup \{1\}$ with $|t - s| \geqq N^{-1}$, and every $\delta \in (0, \frac{1}{2}]$,*

$$E\left|\frac{X_N(t)}{q_\delta(t)} - \frac{X_N(s)}{q_\delta(s)}\right|^{2\alpha} \leqq \tilde{M}_\alpha |t - s|^{2\alpha\delta}\,. \tag{1.14}$$

PROOF. Without loss of generality take $0 \leqq s < t \leqq \frac{1}{2}$. The c_r-inequality and Lemma 1.2 yield for $N^{-1} \leqq s < t \leqq \frac{1}{2}$, $t - s \geqq N^{-1}$,

$$\begin{aligned} E\left|\frac{X_N(t)}{q_\delta(t)} - \frac{X_N(s)}{q_\delta(s)}\right|^{2\alpha} &\leqq 2^{2\alpha-1}\left\{E\left|\frac{X_N(t) - X_N(s)}{q_\delta(t)}\right|^{2\alpha} + E\left|X_N(s)\left(\frac{1}{q_\delta(s)} - \frac{1}{q_\delta(t)}\right)\right|^{2\alpha}\right\} \\ &\leqq 2^{2\alpha-1}M_\alpha\left\{\frac{(t-s)^\alpha}{(t/2)^{2\alpha(\frac{1}{2}-\delta)}} + s^\alpha\left(\frac{1}{q_\delta(s)} - \frac{1}{q_\delta(t)}\right)^{2\alpha}\right\} \\ &\leqq 2^{2\alpha-1}M_\alpha\{2^\alpha(t-s)^{2\alpha\delta} + 2^\alpha(t-s)^{2\alpha\delta}\} \\ &= 2^{3\alpha}M_\alpha(t-s)^{2\alpha\delta}\,, \end{aligned}$$

because

$$s^{\frac{1}{2}}\left(\frac{1}{q_\delta(s)} - \frac{1}{q_\delta(t)}\right) \leqq 2^{\frac{1}{2}}(t-s)^\delta\,, \qquad \text{for} \quad 0 \leqq s < t \leqq \tfrac{1}{2}\,.$$

For $s = 0$, $t - s \geqq N^{-1}$ implies $t \geqq N^{-1}$ and although now $X_N(s) = 0$ the proof is still formally correct. □

LEMMA 1.4. *For every $\alpha > \frac{1}{2}$ there exist $M^* \in (0, \infty)$ and $M_\alpha{}^* \in (0, \infty)$ such that for every array of continuous* df's $F_{1N}, F_{2N}, \cdots, F_{NN}$, $N = 2, 3, \cdots$, *every* $N = 2, 3, \cdots$, *every $\delta \in (0, \frac{1}{2}]$ and every $c > 0$,*

$$\begin{aligned} &P\left(\sup_{|t-(k/N)| \leqq N^{-1}}\left|\frac{X_N(t)}{q_\delta(t)} - \frac{X_N(k/N)}{q_\delta(k/N)}\right| \geqq c\right) \\ &\qquad \leqq M_\alpha{}^*(cN^\delta)^{-2\alpha}\,, \qquad \textit{for} \quad k = 2, 3, \cdots, N-2\,, \\ &\qquad \leqq M^*(cN^\delta)^{-1}\,, \qquad \textit{for} \quad k = 1\,, \quad N-1\,, \end{aligned} \tag{1.15}$$

and

$$P\left(\sup_{|t-(k/N)| \leqq N^{-1}}\left|X_N(t) - X_N\left(\frac{k}{N}\right)\right| \geqq c\right) \leqq M_\alpha{}^*(cN^{\frac{1}{2}})^{-2\alpha}\,, \qquad \textit{for} \quad k = 1, 2, \cdots, N-1\,. \tag{1.16}$$

PROOF. We assume $k + 1 \leqq \frac{1}{2}N$; the proof for other values of k requires only minor modifications.

Suppose first that $2 \leqq k \leqq \frac{1}{2}N - 1$. Then

$$(1.17)\quad \begin{aligned} \sup_{|t-kN^{-1}|\leqq N^{-1}} &\left|\frac{X_N(t)}{q_\delta(t)} - \frac{X_N(kN^{-1})}{q_\delta(kN^{-1})}\right| \\ &\leqq \sup_{|t-kN^{-1}|\leqq N^{-1}} \left|\frac{X_N(t) - X_N(kN^{-1})}{q_\delta(t)}\right| \\ &\quad + |X_N(kN^{-1})| \left(\frac{1}{q_\delta((k-1)/N)} - \frac{1}{q_\delta(k/N)}\right) \\ &\leqq \frac{|X_N((k+1)/N) - X_N((k-1)/N)| + 4N^{-\frac{1}{2}}}{q_\delta((k-1)/N)} \\ &\quad + |X_N(kN^{-1})| \left(\frac{1}{q_\delta((k-1)/N)} - \frac{1}{q_\delta(k/N)}\right). \end{aligned}$$

Since $4\{N^{\frac{1}{2}}q_\delta((k-1)/N)\}^{-1} \leqq 2^{\frac{5}{2}}N^{-\delta}$, the reasoning in the proof of Lemma 1.3 shows that for $\alpha > \frac{1}{2}$,

$$(1.18)\quad E\left(\sup_{|t-kN^{-1}|\leqq N^{-1}} \left|\frac{X_N(t)}{q_\delta(t)} - \frac{X_N(kN^{-1})}{q_\delta(kN^{-1})}\right|\right)^{2\alpha} \leqq M_\alpha' N^{-2\alpha\delta}.$$

Application of Markov's inequality proves (1.15) for $2 \leqq k \leqq \frac{1}{2}N - 1$; taking $\delta = \frac{1}{2}$ we also obtain (1.16) for $2 \leqq k \leqq \frac{1}{2}N - 1$.

For $k = 1$ we note from Theorem 1.1 that for $0 < \beta \leqq \frac{1}{2}$,

$$P\left(\sup_{t\leqq N^{-1}} \frac{|X_N(t)|}{t} \geqq (\beta^{-1} - 1)N^{\frac{1}{2}}\right) \leqq 2^7\beta,$$

so that

$$P\left(\sup_{t\leqq N^{-1}} \frac{|X_N(t)|}{q_\delta(t)} \geqq 2^{\frac{1}{2}}(\beta^{-1} - 1)N^{-\delta}\right) \leqq 2^7\beta$$

and this proves (1.15) for $k = 1$ and $c \geqq 2^{\frac{1}{2}}N^{-\delta}$ and hence for all $c > 0$.

Finally we note that for $\alpha > \frac{1}{2}$,

$$(1.19)\quad E(\sup_{t\leqq N^{-1}} |X_N(t)|)^{2\alpha} \leqq E(|X_N(N^{-1})| + 2N^{-\frac{1}{2}})^{2\alpha} \leqq M_\alpha'' N^{-\alpha},$$

and the Markov inequality proves (1.16) for $k = 1$. □

Combination of Lemma 1.4 with Theorem 0.3 leads to the following two fundamental theorems:

THEOREM 1.2. *For every $\alpha > \frac{1}{2}$ there exists $\bar{M}_\alpha > 0$ such that for every array of continuous* df's $F_{1N}, F_{2N}, \cdots, F_{NN}$, $N = 1, 2, \cdots$, *every $N = 1, 2, \cdots$, every $0 \leqq a < b \leqq 1$ and every $c > 0$,*

$$(1.20)\quad \begin{aligned} P(\sup_{s,t\in[a,b]} &|X_N(t) - X_N(s)| \geqq c) \\ &\leqq \bar{M}_\alpha c^{-2\alpha}N^{1-\alpha}(b-a), \quad \text{if } b - a \leqq N^{-1}, \\ &\leqq \bar{M}_\alpha c^{-2\alpha}(b-a)^\alpha, \quad \text{if } b - a > N^{-1}. \end{aligned}$$

PROOF. If $b - a \leqq N^{-1}$, Lemma 1.2 and the c_r-inequality imply that for $\alpha > \frac{1}{2}$,

$$\begin{aligned} E(\sup_{s,t\in[a,b]} |X_N(s) - X_N(t)|)^{2\alpha} &\leqq E(|X_N(b) - X_N(a)| + 2N^{\frac{1}{2}}(b-a))^{2\alpha} \\ &\leqq 2^{2\alpha-1}(M_\alpha N^{1-\alpha}(b-a) + 2^{2\alpha}N^\alpha(b-a)^{2\alpha}) \\ &\leqq 2^{2\alpha-1}(M_\alpha + 2^{2\alpha})N^{1-\alpha}(b-a), \end{aligned}$$

and application of Markov's inequality proves the first part of the theorem. If $b - a > N^{-1}$, let k and $k + m$ be the smallest and largest integers in $[aN, bN]$, so that $m \leqq (b-a)N$. Define $S_i = X_N((k+i)/N) - X_N(k/N)$, $i = 0, 1, \cdots, m$. Then $S_0 = 0$ and from Lemma 1.2 we have

$$E|S_j - S_i|^{2\alpha} \leqq M_\alpha\left(\frac{j-i}{N}\right)^\alpha, \qquad \text{for} \quad 0 \leqq i \leqq j \leqq m.$$

It follows from Theorem 0.3 that for $\alpha > 1$,

$$\begin{aligned} P\left(\max_{0\leqq i\leqq m}\left|X_N\left(\frac{k+i}{N}\right) - X_N\left(\frac{k}{N}\right)\right| \geqq c\right) &\leqq M_\alpha' c^{-2\alpha}\left(\frac{m}{N}\right)^\alpha \\ &\leqq M_\alpha' c^{-2\alpha}(b-a)^\alpha. \end{aligned}$$

Combining this with the second part of Lemma 1.4 we find for $\alpha > 1$,

$$\begin{aligned} &P(\sup_{s,t\in[a,b]} |X_N(t) - X_N(s)| \geqq c) \\ &\quad\leqq 2P\left(\sup_{t\in[a,b]}\left|X_N(t) - X_N\left(\frac{k}{N}\right)\right| \geqq \frac{c}{2}\right) \\ &\quad\leqq 2M_\alpha^*\left(\frac{cN^{\frac{1}{2}}}{4}\right)^{-2\alpha}(m+1) + 2M_\alpha'\left(\frac{c}{4}\right)^{-2\alpha}(b-a)^\alpha \\ &\quad\leqq M_\alpha^* 2^{4\alpha+2}c^{-2\alpha}N^{-\alpha}(b-a)N + M_\alpha' 2^{4\alpha+1}c^{-2\alpha}(b-a)^\alpha \\ &\quad\leqq (M_\alpha^* 2^{4\alpha+2} + M_\alpha' 2^{4\alpha+1})c^{-2\alpha}(b-a)^\alpha. \end{aligned}$$

Since a probability is bounded by 1, the result remains true for $\alpha > \frac{1}{2}$ if we take $\bar{M}_\alpha \geqq 1$. □

THEOREM 1.3. *For every $\alpha > 0$ and every $\delta \in (0, \frac{1}{2}]$ there exist $\bar{M} > 0$ and $\bar{M}_{\alpha,\delta} > 0$ such that for every array of continuous* df's $F_{1N}, F_{2N}, \cdots, F_{NN}$, $N = 1, 2, \cdots$, *every $N = 1, 2, \cdots$ and every $c > 0$,*

$$P\left(\sup_{t\in[0,1]}\left|\frac{X_N(t)}{q_\delta(t)}\right| \geqq c\right) \leqq \bar{M}_{\alpha,\delta}c^{-2\alpha} + \bar{M}c^{-1}N^{-\delta}. \tag{1.21}$$

PROOF. Define $S_i = X_N(iN^{-1})/q_\delta(iN^{-1})$, for $i = 1, 2, \cdots, N$, $S_0 = 0$. Lemma 1.3 ensures that for $\alpha > \frac{1}{2}$,

$$E|S_j - S_i|^{2\alpha} \leqq \tilde{M}_\alpha\left(\frac{j-i}{N}\right)^{2\alpha\delta}, \qquad \text{for} \quad 0 \leqq i \leqq j \leqq N.$$

Theorem 0.3 implies, for $\alpha > (2\delta)^{-1}$,

$$P\left(\max_{0\leqq k\leqq N}\left|\frac{X_N(kN^{-1})}{q_\delta(kN^{-1})}\right| \geqq c\right) \leqq M_{\alpha,\delta}c^{-2\alpha}.$$

Application of Lemma 1.4 yields

$$P\left(\sup_{t\in[0,1]}\left|\frac{X_N(t)}{q_\delta(t)}\right| \geqq c\right) \leqq 2^{2\alpha}M_{\alpha,\delta}c^{-2\alpha} + 2^{2\alpha}NM_\alpha{}^*(cN^\delta)^{-2\alpha} + 4M^*(cN^\delta)^{-1},$$

which proves the theorem for $\alpha > (2\delta)^{-1}$ and hence for every $\alpha > 0$. □

The following corollary is immediate from Theorem 1.3:

COROLLARY 1.1. *For every* $\varepsilon > 0$ *and every* $\delta \in (0, \frac{1}{2}]$, *there exists* $M = M(\varepsilon, \delta)$, *such that for every array of continuous* df's $F_{1N}, F_{2N}, \cdots, F_{NN}$, $N = 1, 2, \cdots$, *and every* $N = 1, 2, \cdots$,

$$(1.22) \qquad P\left(\sup_{-\infty<x<\infty}\frac{N^{\frac{1}{2}}|\mathbb{F}_N(x) - \bar{F}_N(x)|}{q_\delta(\bar{F}_N(x))} \geqq M\right) \leqq \varepsilon .$$

PROOF. Since $\bar{F}_N$ is assumed to be continuous, we have

$$(1.23) \qquad \sup_{0\leqq t\leqq 1}\frac{N^{\frac{1}{2}}|\tilde{\mathbb{F}}_N(t) - t|}{q_\delta(t)} = \sup_{-\infty<x<\infty}\frac{N^{\frac{1}{2}}|\tilde{\mathbb{F}}_N(\bar{F}_N(x)) - \bar{F}_N(x)|}{q_\delta(\bar{F}_N(x))} .$$

Moreover, $\tilde{\mathbb{F}}_N \circ \bar{F}_N = \mathbb{F}_N$ with probability 1, so that (1.22) follows from (1.21) and (1.23). □

Corollary 1.1 is basic in the asymptotic theory of rank statistics in the case where the sample elements are allowed to have different df's. In particular this corollary can be used to counterbalance the growth of the scores generating functions near the boundary of the unit interval. In the i.i.d. case Corollary 1.1 is proved for the first time in Govindarajulu, Le Cam and Raghavachari (1967). Pyke and Shorack (1968) gave a simpler proof with the aid of the Poisson process and the Birnbaum–Marshall inequality. The result in the non-i.i.d. case for continuous underlying df's is already given in Sen (1970). However, it is clear from Shorack (1973) that the proof given by Sen is incorrect. The proof given here is different from the methods used by the authors mentioned above.

In order to formulate a corollary of Theorem 1.2 let us introduce for every positive integer m the function I_m on $[0, 1]$ defined by $I_m(1) = 1$ and

$$(1.24) \qquad I_m(u) = \frac{i-1}{m} \quad \text{for} \quad \frac{i-1}{m} \leqq u < \frac{i}{m}, \qquad i = 1, 2, \cdots, m .$$

COROLLARY 1.2. *For every* $\varepsilon > 0$ *and every* $c > 0$, *there exist* $N_0 = N_0(\varepsilon, c)$ *and* $m_0 = m_0(\varepsilon, c)$, *such that for every array of continuous* df's $F_{1N}, F_{2N}, \cdots, F_{NN}$, $N = 1, 2, \cdots$, *every* $N \geqq N_0$ *and every* $m \geqq m_0$,

$$(1.25) \qquad P(\sup_{0\leqq t\leqq 1}|X_N(I_m(t)) - X_N(t)| \geqq c) \leqq \varepsilon .$$

PROOF. Note that Theorem 1.2 implies that for every $\alpha > \frac{1}{2}$,

$$\begin{aligned} &P(\sup_{0\leqq t<1}|X_N(I_m(t)) - X_N(t)| \geqq c) \\ &\quad = P(\max_{k=1,2,\cdots,m}\sup_{(k-1)/m\leqq t<k/m}|X_N(I_m(t)) - X_N(t)| \geqq c) \\ &\quad \leqq \sum_{k=1}^m P\left(\sup_{(k-1)/m\leqq t<k/m}\left|X_N\left(\frac{k-1}{m}\right) - X_N(t)\right| \geqq c\right) \\ &\quad \leqq \bar{M}_\alpha c^{-2\alpha}\{\min(m, N)\}^{1-\alpha} . \end{aligned}$$

□

Corollary 1.2 is a generalization to the non-i.i.d. case of a theorem due to Ruymgaart, Shorack and van Zwet (1972). This result is especially useful in the asymptotic theory of rank statistics when one wants to replace certain integrals with respect to the measure induced by the empirical df by the corresponding integrals with respect to the measure induced by the averaged df.

A second consequence of Theorem 1.2 is Corollary 1.3. It is a stronger statement than Theorem 2.1 for $k = 1$.

COROLLARY 1.3. *For every* $\varepsilon > 0$ *there exists* $M = M(\varepsilon)$, *such that for every array of continuous* df's $F_{1N}, F_{2N}, \cdots, F_{NN}$, $N = 1, 2, \cdots$, *every* $N = 1, 2, \cdots$, *every* $0 \leqq a < b \leqq 1$,

$$P(\sup_{s,t\in[a,b]} |X_N(t) - X_N(s)| \geqq M(b-a)^{\frac{1}{2}}) \leqq \varepsilon . \tag{1.26}$$

PROOF. Apply Theorem 1.2 with $\alpha = 1$ and $c = M(b-a)^{\frac{1}{2}}$. □

The last theorem in this section is also of much help in the asymptotic theory of rank statistics in the non-i.i.d. case. For instance, it is useful when one wants to replace Theorem 1.1 and Corollary 1.1, which supply bounds for the empirical df $\mathbb{F}_N$, by similiar statements where bounds are given for the modified empirical df $\mathbb{F}_N^*$, defined as $\mathbb{F}_N^* = (N/(N+1))\mathbb{F}_N$.

THEOREM 1.4. *For* $N \in \{1, 2, \cdots\}$, *continuous* df's $F_{1N}, F_{2N}, \cdots, F_{NN}$ *and* $\alpha \in (0, N)$, *we have*

$$P(\bar{F}_N(X_{N:N}) \leqq 1 - \alpha/N) \leqq (1 - \alpha/N)^N \leqq e^{-\alpha} , \tag{1.27}$$

$$P(\bar{F}_N(X_{1:N}) \geqq \alpha/N) \leqq (1 - \alpha/N)^N \leqq e^{-\alpha} . \tag{1.28}$$

For α *restricted to the interval* $(0, 1)$, *we have, even if the sample elements are not independent*,

$$P(\bar{F}_N(X_{N:N}) \leqq 1 - \alpha/N) \geqq 1 - \alpha , \tag{1.29}$$

$$P(\bar{F}_N(X_{1:N}) \geqq \alpha/N) \geqq 1 - \alpha . \tag{1.30}$$

PROOF. Note that

$$\begin{aligned} P(\bar{F}_N(X_{N:N}) \leqq 1 - \alpha/N) &= P(X_{N:N} \leqq \bar{F}_N^{-1}(1 - \alpha/N)) \\ &= \prod_{n=1}^N F_{nN}(\bar{F}_N^{-1}(1 - \alpha/N)) . \end{aligned} \tag{1.31}$$

Hence, from the concavity of $\log y$ and Jensen's inequality we obtain

$$\begin{aligned} &\frac{1}{N} \sum_{n=1}^N \log F_{nN}(\bar{F}_N^{-1}(1 - \alpha/N)) \\ &= \frac{1}{N} \log \prod_{n=1}^N F_{nN}(\bar{F}_N^{-1}(1 - \alpha/N)) \\ &= \frac{1}{N} \log P(\bar{F}_N(X_{N:N}) \leqq 1 - \alpha/N) \leqq \log (1 - \alpha/N) , \end{aligned} \tag{1.32}$$

which proves (1.27). Relation (1.28) follows from application of (1.27) in the case of random variables $X'_{nN} \equiv -X_{nN}$, $n = 1, 2, \cdots, N$. In order to prove

(1.29) we remark that Bonferroni's inequality implies that

$$\begin{aligned} P(\bar{F}_N(X_{N:N}) \leqq 1 - \alpha/N) &= P(\bigcap_{n=1}^N [X_{nN} \leqq \bar{F}_N^{-1}(1 - \alpha/N)]) \\ &= 1 - P(\bigcup_{n=1}^N [X_{nN} > \bar{F}_N^{-1}(1 - \alpha/N)]) \\ &\geqq 1 - \sum_{n=1}^N P(X_{nN} > \bar{F}_N^{-1}(1 - \alpha/N)) \\ &= 1 - \sum_{n=1}^N (1 - P(X_{nN} \leqq \bar{F}_N^{-1}(1 - \alpha/N))) \\ &= 1 - \sum_{n=1}^N (1 - F_{nN}(\bar{F}_N^{-1}(1 - \alpha/N))) \\ &= 1 - N + N\bar{F}_N(\bar{F}_N^{-1}(1 - \alpha/N)) \\ &= 1 - N + N - \alpha = 1 - \alpha . \end{aligned}$$

Finally, (1.30) can be proved again from (1.29) with the aid of the rv's X'_{nN}. □

REMARK 1.2. The bounds derived in Theorem 1.4 are sharp in the sense that one can construct examples where these bounds are attained. If $F_{1N} = F_{2N} = \cdots = F_{NN} = \bar{F}_N$ then $P(\bar{F}_N(X_{N:N}) \leqq 1 - \alpha/N) = (1 - \alpha/N)^N$. Moreover, if F_{1N} is chosen such that $1 - F_{1N}(\bar{F}_N^{-1}(1 - \alpha/N)) = \alpha$ and $F_{2N}, F_{3N}, \cdots, F_{NN}$ such that $1 - F_{nN}(\bar{F}_N^{-1}(1 - \alpha/N)) = 0$ for $n = 2, 3, \cdots, N$, then

$$P(\bar{F}_N(X_{N:N}) \leqq 1 - \alpha/N) = 1 - \alpha .$$

REMARK 1.3. The continuity of the underlying df's is essential for the relations (1.28) and (1.29), as the following counterexample shows. Take $N = 2$, $\alpha = \frac{1}{2}$ and for $a < b$,

$$\begin{aligned} F_1(x) &= 0 \quad \text{for} \quad x < a & F_2(x) &= 0 \quad \text{for} \quad x < b \\ &= 1 \quad \text{elsewhere} , & &= 1 \quad \text{elsewhere} . \end{aligned}$$

We conclude this section by remarking that Mehra and Rao (1975) also used Billingsley's Theorem 0.3 fruitfully in their study of the one-dimensional empirical process, divided by certain q-functions, in the situation where the sample elements do have a common df, but where they are not necessarily independent.

2. A property of the multivariate empirical distribution function in the case of continuous underlying distribution functions.

In this section k is an arbitrary positive integer, so that for $N = 1, 2, \cdots$, the multivariate df $\mathbb{F}_N$ is based on the N random vectors $X_{nN} = (X_{1nN}, X_{2nN}, \cdots, X_{knN})$, $n = 1, 2, \cdots, N$, with df's $F_{1N}, F_{2N}, \cdots, F_{NN}$ respectively. Assuming for the moment again continuity of these underlying df's, we shall present a generalization of a slightly weaker version of a theorem due to van Zwet (Lemma 4.4 in Ruymgaart (1974)). See also Bahadur (1966). In fact van Zwet proved that, in the i.i.d. case, Theorem 2.1 below holds, without the factor $(\log (N + 1))^{\frac{1}{2}}$ in (2.1). We conjecture that one can dispense with this factor in the non-i.i.d. case too. This conjecture is clearly true for $k = 1$, where the theorem follows from Corollary 1.3. However, the present Theorem 2.1 is strong enough to handle problems connected with discontinuities in the scores generating functions of rank statistics (see, e.g., Ruymgaart (1974), van Zuijlen (1976b)).

By an abuse of notation we write $\mathbb{F}_N$ and $\bar{F}_N$ for the measure induced by the df's, thus $\mathbb{F}_N\{B\} = \int_B d\mathbb{F}_N$, $\bar{F}_N\{B\} = \int_B d\bar{F}_N$ for a Borel set B in $\mathbb{R}^k$. An interval in $\mathbb{R}^k$ is defined as the product set of k intervals, closed, open or half open, on the line.

THEOREM 2.1. *Let I be an interval in $\mathbb{R}^k$ and let $\mathscr{I} = \{I^* : I^*$ is an interval contained in $I\}$. For every $\varepsilon > 0$ and every positive integer k, there exists $M = M(\varepsilon, k)$, such that for every array of k-variate continuous* df's $F_{1N}, F_{2N}, \cdots, F_{NN}$, $N = 1, 2, \cdots$, *every interval I and every $N = 1, 2, \cdots$,*

$$(2.1)\qquad P\left(\sup_{I^* \in \mathscr{I}} |\mathbb{F}_N\{I^*\} - \bar{F}_N\{I^*\}| \leqq M\left(\frac{\log(N+1)\bar{F}_N\{I\}}{N}\right)^{\frac{1}{2}}\right) \geqq 1 - \varepsilon .$$

Before presenting the proof of this theorem, we shall prove a lemma which supplies an upper bound for $\sup_{I^* \in \mathscr{I}} |\mathbb{F}_N\{I^*\} - \bar{F}_N\{I^*\}|$ in terms of a maximum over a finite number of sets.

By $[a]$ we denote the largest integer in the number a.

LEMMA 2.1. *Let for $N = 1, 2, \cdots$ the k-dimensional* df's $F_{1N}, F_{2N}, \cdots, F_{NN}$ *be continuous and let I be an interval in $\mathbb{R}^k$ with $\bar{F}_N\{I\} > 0$, for $N = 1, 2, \cdots$. Define $\bar{F}_{iN}^{-1}(1 + a) = \infty$, for $a > 0$, where $\bar{F}_{iN} = N^{-1} \sum_{n=1}^N F_{inN}$, and let*

$$\mathscr{I} = \{I^* : I^* \text{ is an interval contained in } I\} ,$$

$$(2.2)\qquad \tilde{\mathscr{I}}_N = \Big\{\tilde{I}_N : \tilde{I}_N = I \cap \textstyle\prod_{i=1}^k \Big(\bar{F}_{iN}^{-1}\Big(\frac{n_{i1}}{N}\bar{F}_N\{I\}\Big), \bar{F}_{iN}^{-1}\Big(\frac{n_{i2}}{N}\bar{F}_N\{I\}\Big)\Big],$$

$$\text{for } k \text{ pairs of integers } (n_{i1}, n_{i2}), \text{ with } n_{i1} < n_{i2} \text{ and}$$

$$n_{ij} \in \Big\{0, 1, 2, \cdots, \Big[\frac{N}{\bar{F}_N\{I\}}\Big] + 1\Big\}, \text{ for } i = 1, 2, \cdots, k,$$

$$j = 1, 2\Big\} .$$

Then, for every $\omega \in \Omega$, $N = 1, 2, \cdots$, $k = 1, 2, \cdots$ we have

$$(2.3)\qquad \sup_{I^* \in \mathscr{I}} |\mathbb{F}_N\{I^*\} - \bar{F}_N\{I^*\}| \leqq \max_{\tilde{I}_N \in \tilde{\mathscr{I}}_N} |\mathbb{F}_N\{\tilde{I}_N\} - \bar{F}_N\{\tilde{I}_N\}| + 2kN^{-1}\bar{F}_N\{I\} .$$

PROOF. Let I^* be an arbitrary interval in I. Define

$$\bar{I}_N^* = \bigcap_{\tilde{I}_N \in \tilde{\mathscr{I}}_N; \tilde{I}_N \supset I^*} \tilde{I}_N , \qquad \underline{I}_N^* = \bigcup_{\tilde{I}_N \in \tilde{\mathscr{I}}_N; \tilde{I}_N \subset I^*} \tilde{I}_N .$$

Note that $\bar{I}_N^*$ and $\underline{I}_N^*$ are elements of $\tilde{\mathscr{I}}_N \cup \emptyset$ and that

$$(2.4)\qquad \bar{F}_N\{\bar{I}_N^*\} - \bar{F}_N\{\underline{I}_N^*\} \leqq 2kN^{-1}\bar{F}_N\{I\} .$$

If I^* is such that $\mathbb{F}_N\{I^*\} - \bar{F}_N\{I^*\} \geqq 0$, we have using (2.4)

$$\begin{aligned}|\mathbb{F}_N\{I^*\} - \bar{F}_N\{I^*\}| &= \mathbb{F}_N\{I^*\} - \bar{F}_N\{I^*\} \leqq \mathbb{F}_N\{\bar{I}_N^*\} - \bar{F}_N\{\underline{I}_N^*\} \\ &\leqq \mathbb{F}_N\{\bar{I}_N^*\} - \bar{F}_N\{\bar{I}_N^*\} + 2kN^{-1}\bar{F}_N\{I\} \\ &\leqq |\mathbb{F}_N\{\bar{I}_N^*\} - \bar{F}_N\{\bar{I}_N^*\}| + 2kN^{-1}\bar{F}_N\{I\} ,\end{aligned}$$

and if $\mathbb{F}_N\{I^*\} - \bar{F}_N\{I^*\} < 0$, we have

$$\begin{aligned}|\mathbb{F}_N\{I^*\} - \bar{F}_N\{I^*\}| &= \bar{F}_N\{I^*\} - \mathbb{F}_N\{I^*\} \leqq \bar{F}_N\{\bar{I}_N^*\} - \mathbb{F}_N\{\underline{I}_N^*\} \\ &\leqq \bar{F}_N\{\underline{I}_N^*\} - \mathbb{F}_N\{\underline{I}_N^*\} + 2kN^{-1}\bar{F}_N\{I\} \\ &\leqq |\mathbb{F}_N\{\underline{I}_N^*\} - \bar{F}_N\{\underline{I}_N^*\}| + 2kN^{-1}\bar{F}_N\{I\} . \qquad \square\end{aligned}$$

PROOF OF THEOREM 2.1. If $\bar{F}_N\{I\} = 0$, the theorem follows immediately. It proves to be convenient to consider the cases $0 < \bar{F}_N\{I\} \leqq 8 \log (N+1)/\varepsilon N$ and $\bar{F}_N\{I\} > 8 \log (N+1)/\varepsilon N$, for fixed $0 < \varepsilon < 1$, separately. Compare with Ruymgaart (1973), page 19.

First suppose that $0 < \bar{F}_N\{I\} \leqq 8(\varepsilon N)^{-1} \log (N+1)$, and choose $M = M_1(\varepsilon) = (2/\varepsilon)^{\frac{3}{2}}$. Then

$$M\left(\frac{\log (N+1)\bar{F}_N\{I\}}{N}\right)^{\frac{1}{2}} \geqq M\left(\frac{\varepsilon(\bar{F}_N\{I\})^2}{8}\right)^{\frac{1}{2}} \geqq \frac{\bar{F}_N\{I\}}{\varepsilon} . \tag{2.5}$$

Moreover, since

$$\sup_{I^* \in \mathscr{I}} |\mathbb{F}_N\{I^*\} - \bar{F}_N\{I^*\}| \leqq \max (\mathbb{F}_N\{I\}, \bar{F}_N\{I\}) , \tag{2.6}$$

we have from (2.5), (2.6) and Markov's inequality that the left-hand side of (2.1) is bounded below by

$$P(\max (\mathbb{F}_N\{I\}, \bar{F}_N\{I\}) \leqq \bar{F}_N\{I\}/\varepsilon) = P(\mathbb{F}_N\{I\} \leqq \bar{F}_N\{I\}/\varepsilon) \geqq 1 - \varepsilon .$$

Next we suppose that $\bar{F}_N\{I\} > 8(\varepsilon N)^{-1} \log (N+1)$. Application of Lemma 2.1 shows that for $M > M_2(k) = 4k(\log 2)^{-\frac{1}{2}}$ and $N = 1, 2, \cdots$, the left-hand side of (2.1) is bounded below by

$$\begin{aligned} &P\left(\max_{\tilde{I}_N \in \tilde{\mathscr{I}}_N} |\mathbb{F}_N\{\tilde{I}_N\} - \bar{F}_N\{\tilde{I}_N\}| \leqq M\left(\frac{\log (N+1)\bar{F}_N\{I\}}{N}\right)^{\frac{1}{2}} - 2k\frac{\bar{F}_N\{I\}}{N}\right) \\ &\geqq P\left(\max_{\tilde{I}_N \in \tilde{\mathscr{I}}_N} |\mathbb{F}_N\{\tilde{I}_N\} - \bar{F}_N\{\tilde{I}_N\}| \leqq \tfrac{1}{2}M\left(\frac{\log (N+1)\bar{F}_N\{I\}}{N}\right)^{\frac{1}{2}}\right) \\ &\geqq 1 - \sum_{\tilde{I}_N \in \tilde{\mathscr{I}}_N} P\left(|\mathbb{F}_N\{\tilde{I}_N\} - \bar{F}_N\{\tilde{I}_N\}| > \tfrac{1}{2}M\left(\frac{\log (N+1)\bar{F}_N\{I\}}{N}\right)^{\frac{1}{2}}\right) .\end{aligned} \tag{2.7}$$

Since $\frac{1}{2}M(N \log (N+1)\bar{F}_N\{I\})^{\frac{1}{2}} \geqq 1$ for $M \geqq M_3 = \frac{1}{2}2^{\frac{1}{2}}(\log 2)^{-\frac{1}{2}}$, Theorem 0.2 is applicable, so that we may assume that $N\mathbb{F}_N\{\tilde{I}_N\}$ in (2.7) is a binomial rv with parameters N and $\bar{F}_N\{\tilde{I}_N\}$.

With the aid of Bernstein's inequality (see, e.g., Bahadur (1966), page 578) we find, using $\max (\bar{F}_N\{\tilde{I}_N\}, 1 - \bar{F}_N\{\tilde{I}_N\}) \leqq 1$, that for $N = 1, 2, \cdots$, and $M > 0$,

$$\begin{aligned} &P\left(|\mathbb{F}_N\{\tilde{I}_N\} - \bar{F}_N\{\tilde{I}_N\}| \geqq \tfrac{1}{2}M\left(\frac{\log (N+1)\bar{F}_N\{I\}}{N}\right)^{\frac{1}{2}}\right) \\ &\leqq 2 \exp\left(-\frac{\frac{1}{4}M^2 N \log (N+1)\bar{F}_N\{I\}}{2N\bar{F}_N\{\tilde{I}_N\} + \frac{1}{3}M(N \log (N+1)F_N\{I\})^{\frac{1}{2}}}\right) .\end{aligned} \tag{2.8}$$

Moreover, since $\bar{F}_N\{I\} > 8(\varepsilon N)^{-1}\log(N+1) > 8N^{-1}\log(N+1)$ and $\bar{F}_N\{\tilde{I}_N\} \leqq \bar{F}_N\{I\}$, we obtain the following upper bound for (2.8):

$$(2.9) \qquad 2\exp\left(-\frac{\frac{3}{2}M^2 2^{\frac{1}{2}}\log(N+1)}{12.2^{\frac{1}{2}}+M}\right) \leqq 2\exp(-\tfrac{3}{4}M\log(N+1)),$$
$$\text{for} \quad M \geqq M_4 = 12.2^{\frac{1}{2}}.$$

Noting that the number of elements in $\tilde{\mathscr{I}}_N$ is bounded above by

$$\left(\frac{N}{\bar{F}_N\{I\}}+2\right)^{2k} \leqq \left(\frac{N^2\varepsilon}{8\log(N+1)}+2\right)^{2k} \leqq (5N^2)^{2k},$$

we obtain from (2.6)—(2.9) that for $M \geqq \max(M_1, M_2, M_3, M_4, 5\frac{1}{3}k)$, $N = 1, 2, \cdots$,

$$P\left(\sup_{I^*\in\mathscr{I}} |\mathbb{F}_N\{I^*\} - \bar{F}_N\{I^*\}| \leqq M\left(\frac{\log(N+1)\bar{F}_N\{I\}}{N}\right)^{\frac{1}{2}}\right)$$
$$(2.10) \qquad \geqq 1 - (5N^2)^{2k}2(N+1)^{-\frac{3}{4}M} \geqq 1 - 2.5^{2k}(N+1)^{4k-\frac{3}{4}M}$$
$$\geqq 1 - 2.5^{2k}.2^{4k-\frac{3}{4}M},$$

which completes the proof of the theorem. □

REMARK 2.1. If in Theorem 2.1 we take $I = \mathbb{R}^k$, we obtain the following result which is a kind of Glivenko–Cantelli theorem: for every $\varepsilon > 0$ and every positive integer k, there exists $M = M(\varepsilon, k)$ such that for every array of k-variate continuous df's $F_{1N}, F_{2N}, \cdots, F_{NN}$, $N = 1, 2, \cdots$, and every $N = 1, 2, \cdots$,

$$(2.11) \qquad P(\sup_{-\infty<x_1,x_2,\cdots,x_k<\infty} |\mathbb{F}_N(x_1, x_2, \cdots, x_k) - \bar{F}_N(x_1, x_2, \cdots, x_k)|$$
$$\leqq MN^{-\frac{1}{2}}(\log(N+1))^{\frac{1}{2}}) \geqq 1 - \varepsilon.$$

3. Discontinuous underlying distribution functions. In this section we shall establish a theorem which makes it clear that, without any additional condition, the most important results from the foregoing sections remain valid without the restriction of continuous underlying df's. For related results see, e.g., Behnen (1976) and Conover (1973).

An interval $I \subset \mathbb{R}^k$ is defined in the introduction of Section 2; the corresponding definition of the class of intervals $\mathscr{I}$ is given in Theorem 2.1. Given a set S, S^c will denote its complement, $\chi(S)$ its indicator function and $\chi(S; s)$ the value of this function at the point s, i.e.,

$$(3.1) \qquad \chi(S; s) = 1 \quad \text{for} \quad s \in S$$
$$= 0 \quad \text{for} \quad s \in S^c.$$

THEOREM 3.1. *Let k be a positive integer and let $\mathbb{F}_N$ be the empirical* df *based on N k-variate sample elements $X_{nN} = (X_{1nN}, X_{2nN}, \cdots, X_{knN})$, $n = 1, 2, \cdots, N$, where the X_{nN} are distributed independently according to given, possibly discontinuous* df's F_{nN}. *Let us denote for $i = 1, 2, \cdots, k$ by F_{inN} the* ith *marginal* df *of F_{nN}, let*

$\bar{F}_{iN} = N^{-1} \sum_{n=1}^{N} F_{inN}$, let $\{\xi_\nu^{(i)}, \nu = 1, 2, \cdots\}$ be the countable set of discontinuity points of $\bar{F}_{iN}$ and let $p_\nu^{(i)}$ be the height of the jump at $\xi_\nu^{(i)}$ of $\bar{F}_{iN}$. Finally let I be an interval in $\mathbb{R}^k$. There exist N k-variate random vectors $Y_{nN} = (Y_{1nN}, Y_{2nN}, \cdots, Y_{knN})$, $n = 1, 2, \cdots, N$, where the Y_{nN} are distributed independently according to continuous df's *G_{nN}, and an interval $\tilde{I} \subset \mathbb{R}^k$, such that*

$$(3.2) \qquad \bar{F}_N(x_1, x_2, \cdots, x_k) = \bar{G}_N(x_1 + \textstyle\sum_\nu p_\nu^{(1)}\chi([\xi_\nu^{(1)}, \infty); x_1), \cdots, x_k + \textstyle\sum_\nu p_\nu^{(k)}\chi([\xi_\nu^{(k)}, \infty); x_k))$$

and with probability one

$$(3.3) \qquad \mathbb{F}_N(x_1, x_2, \cdots, x_k) = \mathbb{G}_N(x_1 + \textstyle\sum_\nu p_\nu^{(1)}\chi([\xi_\nu^{(1)}, \infty); x_1), \cdots, x_k + \textstyle\sum_\nu p_\nu^{(k)}\chi([\xi_\nu^{(k)}, \infty); x_k))$$

and

$$(3.4) \qquad \sup_{I^* \in \mathscr{I}} \frac{|\mathbb{F}_N\{I^*\} - \bar{F}_N\{I^*\}|}{(\bar{F}_N\{I\})^{\frac{1}{2}}} \leqq \sup_{\tilde{I}^* \in \tilde{\mathscr{I}}} \frac{|\mathbb{G}_N\{\tilde{I}^*\} - \bar{G}_N\{\tilde{I}^*\}|}{(\bar{G}_N\{\tilde{I}\})^{\frac{1}{2}}},$$

where $\mathbb{G}_N$ denotes the empirical df *based on the Y_{nN}, $n = 1, 2, \cdots, N$ and $\bar{F}_N = N^{-1} \sum_{n=1}^{N} F_{nN}$, $\bar{G}_N = N^{-1} \sum_{n=1}^{N} G_{nN}$.*

Proof. Let $\{U_\nu^{(in)}, i = 1, 2, \cdots, k, n = 1, 2, \cdots, N, \nu = 1, 2, \cdots\}$ be a set of uniform (0, 1) distributed rv's, mutually independent and also independent of the random vectors X_{nN}, $n = 1, 2, \cdots, N$. Note that $\{\xi_\nu^{(i)}, \nu = 1, 2, \cdots\}$ contains the discontinuity points of each F_{inN}, $n = 1, 2, \cdots, N$.

Since $\sum_\nu p_\nu^{(i)} \leqq 1$ for $i = 1, 2, \cdots, k$, we can define for $n = 1, 2, \cdots, N$ the random vector $Y_{nN} = (Y_{1nN}, Y_{2nN}, \cdots, Y_{knN})$ as follows:

$$(3.5) \qquad Y_{inN} = X_{inN} + \textstyle\sum_\nu p_\nu^{(i)}\chi((\xi_\nu^{(i)}, \infty); X_{inN}) + \textstyle\sum_\nu p_\nu^{(i)} U_\nu^{(in)}\chi(\{\xi_\nu^{(i)}\}; X_{inN}),$$

for $i = 1, 2, \cdots, k$, so that X_{nN} is transformed stochastically to Y_{nN}. Let G_{nN} be the df of Y_{nN} and let $\mathbb{G}_N$ be the empirical df based on $Y_{1N}, Y_{2N}, \cdots, Y_{NN}$. It is clear that all the marginal df's of G_{nN} are continuous and hence G_{nN} is continuous. From definition (3.5) it is immediate that for $n = 1, 2, \cdots, N$ and $i = 1, 2, \cdots, k$,

$$(3.6) \qquad X_{inN} + \textstyle\sum_\nu p_\nu^{(i)}\chi((\xi_\nu^{(i)}, \infty); X_{inN}) \leqq Y_{inN} \leqq X_{inN} + \textstyle\sum_\nu p_\nu^{(i)}\chi([\xi_\nu^{(i)}, \infty); X_{inN}),$$

and hence

$$(3.7) \qquad [X_{inN} \leqq x_i] \Leftrightarrow [Y_{inN} \leqq x_i + \textstyle\sum_\nu p_\nu^{(i)}\chi([\xi_\nu^{(i)}, \infty); x_i)],$$

$$(3.8) \qquad [X_{inN} < x_i] \Leftrightarrow [Y_{inN} < x_i + \textstyle\sum_\nu p_\nu^{(i)}\chi((\xi_\nu^{(i)}, \infty); x_i)].$$

From (3.7) it is obvious that (with $\bar{G}_N = N^{-1} \sum_{n=1}^{N} G_{nN}$), the equalities (3.2) and (3.3) hold.

Next, let us construct from the given interval $I \subset \mathbb{R}^k$ an interval $\tilde{I} \subset \mathbb{R}^k$, such that (3.4) is satisfied. Therefore, we define for $i = 1, 2, \cdots, k$ the functions f_i

and g_i as follows:

(3.9) $f_i(x) = x + \sum_\nu p_\nu^{(i)}\chi([\xi_\nu^{(i)}, \infty); x)$, for $x \in (-\infty, \infty)$,

(3.10) $g_i(x) = x + \sum_\nu p_\nu^{(i)}\chi((\xi_\nu^{(i)}, \infty); x)$, for $x \in (-\infty, \infty)$.

Let $I = \prod_{i=1}^k I_i$ and let for $i = 1, 2, \cdots, k$, a_i and b_i be the endpoints of the interval $I_i \subset \mathbb{R}$, with $a_i \leqq b_i$. Let

(3.11)
$$\tilde{a}_i = g_i(a_i) \text{ for } a_i \in I_i, \qquad \text{and} \qquad \tilde{b}_i = g_i(b_i) \text{ for } b_i \in I_i^c,$$
$$= f_i(a_i) \text{ elsewhere}, \qquad\qquad = f_i(b_i) \text{ elsewhere}.$$

We define $\tilde{I} = \prod_{i=1}^k \tilde{I}_i$, wherefore $i = 1, 2, \cdots, k$, $\tilde{I}_i$ is the interval in $\mathbb{R}$ with the endpoints $\tilde{a}_i$ and $\tilde{b}_i$ and $\tilde{a}_i \in \tilde{I}_i$ iff $a_i \in I_i$, $\tilde{b}_i \in \tilde{I}_i$ iff $b_i \in I_i$. With the aid of (3.7) and (3.8) it can be verified that

(3.12) $$\bar{F}_N\{I\} = \bar{G}_N\{\tilde{I}\} \qquad \text{and} \qquad \mathbb{F}_N\{I\} = \mathbb{G}_N\{\tilde{I}\}.$$

Since analogously we can construct for every interval $I^* \subset \mathscr{I}$ an interval $\tilde{I}^* \subset \tilde{\mathscr{I}}$ satisfying (3.12) with $I = I^*$ and $\tilde{I} = \tilde{I}^*$, the proof is completed. □

COROLLARY 3.1. *Theorem* 1.1, *Corollary* 1.1, (1.27), (1.30) *and Theorem* 2.1 *also hold without the restriction to continuous underlying* df's.

PROOF. The assertion for Theorem 2.1 is immediate from (3.4). For $k = 1$, we denote by $Y_{1:N}$, $Y_{N:N}$ the first and last order statistic of the random variables $Y_{1N}, Y_{2N}, \cdots, Y_{NN}$, which are constructed in the proof of Theorem 3.1 (cf. (3.5)). In view of (3.7) we obtain

$$x \geqq X_{1:N} \Leftrightarrow x + \sum_\nu p_\nu^{(1)}\chi([\xi_\nu^{(1)}, \infty); x) \geqq Y_{1:N},$$
$$x < X_{N:N} \Leftrightarrow x + \sum_\nu p_\nu^{(1)}\chi([\xi_\nu^{(1)}, \infty); x) < Y_{N:N},$$

so that (3.2) and (3.3) imply that with probability one

(3.13) $$\sup_{x \geqq X_{1:N}} \frac{\bar{F}_N(x)}{\mathbb{F}_N(x)} \leqq \sup_{x \geqq Y_{1:N}} \frac{\bar{G}_N(x)}{\mathbb{G}_N(x)},$$

(3.14) $$\sup_{x < X_{N:N}} \frac{1 - \bar{F}_N(x)}{1 - \mathbb{F}_N(x)} \leqq \sup_{x < Y_{N:N}} \frac{1 - \bar{G}_N(x)}{1 - \mathbb{G}_N(x)},$$

(3.15) $$\sup_{-\infty < x < \infty} \frac{\mathbb{F}_N(x)}{\bar{F}_N(x)} \leqq \sup_{-\infty < x < \infty} \frac{\mathbb{G}_N(x)}{\bar{G}_N(x)},$$

(3.16) $$\sup_{-\infty < x < \infty} \frac{1 - \mathbb{F}_N(x)}{1 - \bar{F}_N(x)} \leqq \sup_{-\infty < x < \infty} \frac{1 - \mathbb{G}_N(x)}{1 - \bar{G}_N(x)},$$

(3.17) $$\sup_{-\infty < x < \infty} \frac{|\mathbb{F}_N(x) - \bar{F}_N(x)|}{q_\delta(\bar{F}_N(x))} \leqq \sup_{-\infty < x < \infty} \frac{|\mathbb{G}_N(x) - \bar{G}_N(x)|}{q_\delta(\bar{G}_N(x))}.$$

Moreover, with the aid of (3.2) one can show that

(3.18) $$\bar{F}_N(X_{N:N}) = \bar{G}_N(X_{N:N} + \sum_\nu p_\nu^{(1)}\chi([\xi_\nu^{(1)}, \infty); X_{N:N})) \geqq \bar{G}_N(Y_{N:N}),$$

(3.19) $$\bar{F}_N(X_{1:N}) = \bar{G}_N(X_{1:N} + \sum_\nu p_\nu^{(1)}\chi([\xi_\nu^{(1)}, \infty); X_{1:N})) \geqq \bar{G}_N(Y_{1:N}).$$

The proof can be completed from (3.13)—(3.19). □

REMARK 3.1. From Corollary 1.3, the proof of Corollary 1.1 and (3.4), it is immediate that for $k = 1$, Theorem 2.1 even holds without the factor $(\log (N + 1))^{\frac{1}{2}}$ in (2.1) and without the restriction to continuous underlying df's.

REMARK 3.2. Of course, as in the proof of Corollary 3.1, one can show that also the transformed versions (cf. (1.23)) of Theorem 1.2 and Theorem 1.3 remain valid without the restriction to continuous df's.

Acknowledgments. The author is indebted to Frits Ruymgaart for suggesting the problem, to Professor G. R. Shorack for his stimulating interest and to Professor W. R. van Zwet for his careful reading of the manuscript and his valuable comments, which resulted in considerable improvements of the original text.

REFERENCES

[1] BAHADUR, R. R. (1966). A note on quantiles in large samples. *Ann. Math. Statist.* **37** 557-580.

[2] BEHNEN, K. (1976). Asymptotic comparison of rank tests for the regression problem when ties are present. *Ann. Statist.* **4** 157-174.

[3] BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

[4] CONOVER, W. J. (1973). Rank tests for one sample, two samples, and k samples without the assumption of a continuous function. *Ann. Statist.* **1** 1105-1125.

[5] GOVINDARAJULU, Z., LE CAM, L. and RAGHAVACHARI, M. (1967). Generalization of theorems of Chernoff-Savage on asymptotic normality of nonparametric test statistics. *Proc. Fifth Berkeley Symp. Math Statist. Prob.* 609-638. Univ. of California Press.

[6] HOEFFDING, W. (1956). On the distribution of the number of successes in independent trials. *Ann. Math. Statist.* **27** 713-721.

[7] MEHRA, K. L. and RAO, M. S. (1975). Weak convergence of generalized empirical processes relative to d_q under strong mixing. *Ann. Probability* **3** 979-991.

[8] PYKE, R. and SHORACK, G. R. (1968). Weak convergence of a two-sample empirical process and a new approach to Chernoff-Savage theorems. *Ann. Math. Statist.* **39** 755-771.

[9] RUYMGAART, F. H. (1973). *Asymptotic Theory of Rank Tests for Independence.* Mathematical Centre Tracts 43, Amsterdam.

[10] RUYMGAART, F. H. (1974). Asymptotic normality of nonparametric tests for independence. *Ann. Statist.* **2** 892-910.

[11] RUYMGAART, F. H., SHORACK, G. R. and VAN ZWET, W. R. (1972). Asymptotic normality of nonparametric tests for independence. *Ann. Math. Statist.* **43** 1122-1135.

[12] SEN, P. K. (1970). On the distribution of one-sample rank order statistics. *Nonparametric Techniques in Statistics Information* (M. Puri, ed.) 53-72. Cambridge Univ. Press.

[13] SHORACK, G. R. (1970). A uniformly convergent empirical process. Tech. Report No. 20, Math. Dept., Univ. of Washington.

[14] SHORACK, G. R. (1972). Functions of order statistics. *Ann. Math. Statist.* **43** 412-427.

[15] SHORACK, G. R. (1973). Convergence of reduced empirical and quantile processes with application to functions of order statistics in the non-i.i.d. case. *Ann. Statist.* **1** 146-152.

[16] ZUIJLEN, M. C. A. VAN (1976a). Some properties of the empirical distribution function in the non-i.i.d. case. *Ann. Statist.* **4** 406-408.

[17] ZUIJLEN, M. C. A. VAN (1976b). *Empirical Distributions and Rank Statistics.* Mathematical Centre Tracts 79, Amsterdam.

MATHEMATISCH INSTITUUT DER KATHOLIEKE UNIVERSITEIT VAN NIJMEGEN
TOERNOOIVELD
NIJMEGEN
THE NETHERLANDS