The Annals of Probability
1992, Vol. 20, No. 2, 1086–1098

ALMOST SURE CONVERGENCE OF CERTAIN SLOWLY CHANGING SYMMETRIC ONE- AND MULTI-SAMPLE STATISTICS

BY N. HENZE AND B. VOIGT

Universität Karlsruhe

Let $X_j^{(i)}$, $i = 1, \dots, k$; $j \in \mathbf{N}$, be independent d-dimensional random vectors which are identically distributed for each fixed $i = 1, \dots, k$. We give a sufficient condition for almost sure convergence of a sequence $T_{n_1, \dots, n_k}$ of statistics based on $X_j^{(i)}$ $i = 1, \dots, k$; $j = 1, \dots, n_i$, which are symmetric functions of $X_1^{(i)}, \dots, X_{n_i}^{(i)}$ for each i and do not change too much when variables are added or deleted. A key auxiliary tool for proofs is the Efron–Stein inequality. Applications include strong limits for certain nearest neighbor graph statistics, runs and empty blocks.

1. Introduction. The Efron–Stein inequality [ESI, Efron and Stein (1981)], which essentially says that Tukey's jackknife estimate of variance is nonnegatively biased, has already had interesting applications in various fields [Hochbaum and Steele (1982), Steele (1981, 1982), Devroye (1987), Steele, Shepp and Eddy (1987)].

Alternative proofs, generalizations and analogues of the ESI were given by Karlin and Rinott (1982), Bhargava (1983), Vitale (1984), Rhee and Talagrand (1986), Steele (1986) and Vitale (1988).

It is the purpose of this paper to show how the ESI may be fruitfully applied to yield almost sure convergence of certain symmetric one- and multi-sample statistics with small fluctuation when variables are added or deleted. The main message is that in this case convergence of expectations implies almost sure convergence. For ease of reference we restate the ESI.

LEMMA 1.1 [Efron and Stein (1981)]. *Let $X_1, \dots, X_{n+1}$ be i.i.d. d-dimensional random vectors and $S(x_1, \dots, x_n)$ a real-valued symmetric statistic such that $E[S(X_1, \dots, X_n)^2] < \infty$. If $S_i = S(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_{n+1})$, $i = 1, \dots, n+1$, and $\bar{S} = (n+1)^{-1}\sum_{i=1}^{n+1} S_i$, we have*

$$\operatorname{Var}(S(X_1, \dots, X_n)) \leq E\left[\sum_{i=1}^{n+1} (S_i - \bar{S})^2\right].$$

2. Main result. Consider independent random vectors $X_j^{(i)}$, $i = 1, \dots, k$; $j \in \mathbf{N}$, in $\mathbf{R}^d$, where, for each i, $(X_j^{(i)})_{j \in \mathbf{N}}$ are identically distributed. For $(n_1, \dots, n_k) \in \mathbf{N}^k$, let $S_{n_1, \dots, n_k} = S_{n_1, \dots, n_k}(X_1^{(1)}, \dots, X_{n_1}^{(1)};\ X_1^{(2)}, \dots,$

Received September 1989; revised October 1990.

AMS 1980 *subject classifications*. Primary 60F15; secondary 62G10.

Key words and phrases. Almost sure convergence, Efron–Stein inequality, nearest neighbors, geometric probability, runs, empty blocks.

$X_{n_2}^{(2)}; \dots ; X_1^{(k)}, \dots, X_{n_k}^{(k)})$ be a real valued statistic which is symmetric in each group $X_1^{(i)}, \dots, X_{n_i}^{(i)}$, $i = 1, \dots, k$. Suppose that $E[S_{n_1,\dots,n_k}^2] < \infty$. In what follows, $n = n_1 + \cdots + n_k$ denotes the total sample size. For short, let $S_{\mathbf{n}} = S_{n_1,\dots,n_k}$.

LEMMA 2.1. *Assume that for each* $(n_1, \dots, n_k) \in \mathbf{N}^k$, *there is a positive constant* $d_{n_1,\dots,n_k}$ *such that for each* $i = 1, \dots, k$:

$$(2.1) \qquad |S_{n_1,\dots,n_k} - S_{n_1,\dots,n_{i-1},n_i+1,n_{i+1},\dots,n_k}| \le d_{n_1,\dots,n_k}, \qquad P\text{-}a.s.$$

Then

$$\operatorname{Var}\left(n^{-1} S_{n_1,\dots,n_k}\right) \le 2n^{-1} d_{n_1,\dots,n_k}^2.$$

PROOF. Letting $Y^{(i)} = (X_1^{(i)}, \dots, X_{n_i}^{(i)}; \dots ; X_1^{(k)}, \dots, X_{n_k}^{(k)})$, $i = 2, \dots, k$, we start with

$$\operatorname{Var}(S_{\mathbf{n}}) = E\left[\operatorname{Var}(S_{\mathbf{n}} | Y^{(2)})\right] + \operatorname{Var}\left(E\left[S_{\mathbf{n}} | Y^{(2)}\right]\right).$$

Put $S_{\mathbf{n}}^{(i)} = S_{\mathbf{n}}(X_1^{(1)}, \dots, X_{i-1}^{(1)}, X_{i+1}^{(1)}, \dots, X_{n_1+1}^{(1)}; Y^{(2)}]$ and

$$\overline{S}_{\mathbf{n}} = (n_1 + 1)^{-1} \sum_{i=1}^{n_1+1} S_{\mathbf{n}}^{(i)}.$$

By Lemma 1.1 and (2.1), we then have P-a.s.:

$$\begin{aligned}
\operatorname{Var}(S_{\mathbf{n}} | Y^{(2)}) &\le E\left[\sum_{i=1}^{n_1+1} \left(S_{\mathbf{n}}^{(i)} - \overline{S}_{\mathbf{n}}\right)^2 \middle| Y^{(2)}\right] \\
&\le \sum_{i=1}^{n_1+1} E\left[\left(S_{\mathbf{n}}^{(i)} - S_{n_1+1,n_2,\dots,n_k}\right)^2 \middle| Y^{(2)}\right] \\
&= (n_1 + 1) E\left[\left(S_{\mathbf{n}}^{(n_1+1)} - S_{n_1+1,n_2,\dots,n_k}\right)^2 \middle| Y^{(2)}\right] \\
&= (n_1 + 1) E\left[\left(S_{n_1,\dots,n_k} - S_{n_1+1,n_2,\dots,n_k}\right)^2 \middle| Y^{(2)}\right] \\
&\le 2n_1 d_{n_1,\dots,n_k}^2
\end{aligned}$$

and thus

$$\operatorname{Var}(S_{\mathbf{n}}) \le 2n_1 d_{n_1,\dots,n_k}^2 + \operatorname{Var}\left(E\left[S_{\mathbf{n}} | Y^{(2)}\right]\right).$$

Writing $g_{\mathbf{n}}^{(i)}(Y^{(i)}) = g_{n_1,\dots,n_k}^{(i)}(Y^{(i)}) = E[S_{\mathbf{n}} | Y^{(i)}]$ and applying (2.1) to the conditional expectation $g_{\mathbf{n}}^{(2)}(Y^{(2)})$, we obtain

$$\begin{aligned}
&\left(g_{n_1,\dots,n_k}^{(2)} - g_{n_1,\dots,n_{i-1},n_i+1,n_{i+1},\dots,n_k}^{2}\right)^2 \\
&\quad \le E\left[\left(S_{\mathbf{n}} - S_{n_1,\dots,n_{i-1},n_i+1,n_{i+1},\dots,n_k}\right)^2 \middle| Y^{(2)}\right] \\
&\quad \le d_{n_1,\dots,n_k}^2 \quad \text{a.s., } i = 2, \dots, k,
\end{aligned}$$

and proceeding as above it follows that

$$\begin{aligned}\operatorname{Var}\big(g_{\mathbf{n}}^{(2)}\big) &= E\big[\operatorname{Var}\big(g_{\mathbf{n}}^{(2)}|Y^{(3)}\big)\big] + \operatorname{Var}\big(E\big[g_{\mathbf{n}}^{(2)}|Y^{(3)}\big]\big)\\ &\le 2n_2 d^2_{n_1,\dots,n_k} + \operatorname{Var}\big(g_{\mathbf{n}}^{(3)}(Y^{(3)})\big).\end{aligned}$$

Iterating this reasoning for $i = 3,\dots,k-1$ and finally applying Lemma 1.1 to $g_{\mathbf{n}}^{(k)}(Y^{(k)})$ yields the assertion. □

LEMMA 2.2. *Let $(N_j)_{j\in\mathbf{N}}$ be a sequence of real-valued random variables such that $\lim_{j\to\infty} E[N_j] = b \in \mathbf{R}$ exists. If*

$$\sum_{j=1}^{\infty} P\big(\big|N_j - E[N_j]\big| > \varepsilon\big) < \infty$$

for each $\varepsilon > 0$, we have $\lim_{j\to\infty} N_j = b$, P-a.s..

PROOF. Use the Borel–Cantelli lemma and the triangle inequality. □

We now state our main result.

THEOREM 2.3. *In addition to the conditions stated at the beginning of this section, assume the following:*

(a) *There is a positive constant c with $|n^{-1}S_{\mathbf{n}}| \le c$, P-a.s.*

(b) *There are positive constants $K, \alpha_1,\dots,\alpha_k$ with $\alpha_1 + \cdots + \alpha_k > k-2$ and a sequence $(d_{n_1,\dots,n_k})_{n_1,\dots,n_k\in\mathbf{N}}$ of positive real numbers such that*

$$|S_{n_1,\dots,n_k} - S_{n_1,\dots,n_{i-1},n_i+1,n_{i+1},\dots,n_k}| \le d_{n_1,\dots,n_k}, \qquad i = 1,\dots,k,\ P\text{-a.s.},$$

where

$$d_{n_1,\dots,n_k} \le K\big(n_1^{1-\alpha_1} n_2^{1-\alpha_2} \cdots n_k^{1-\alpha_k}\big)^{1/4k}.$$

Let $(n_1,\dots,n_k) = (n_1(j),\dots,n_k(j))_{j\in\mathbf{N}}$ be a fixed sequence in $\mathbf{N}^k$ such that $\lim_{j\to\infty} n_i(j) = \infty$ $(i = 1,\dots,k)$ and

$$\tau_i = \lim_{j\to\infty} n_i(j)\big(n_1(j) + \cdots + n_k(j)\big)^{-1} > 0, \qquad i = 1,\dots,k,$$

exists (for $k = 1$ set $\tau_1 = 1$). If for some constant b,

$$\lim_{j\to\infty} E\Big[\big(n_1(j) + \cdots + n_k(j)\big)^{-1} S_{\mathbf{n}}\Big] = b,$$

we have

$$\lim_{j\to\infty} \big(n_1(j) + \cdots + n_k(j)\big)^{-1} S_{\mathbf{n}} = b, \qquad P\text{-a.s.}$$

PROOF. From condition (b) and Lemma 2.1, we have

$$\operatorname{Var}\big(n^{-1}S_{\mathbf{n}}\big) \le 2n^{-1} d^2_{n_1,\dots,n_k}. \tag{2.2}$$

Let $(n_1(j),\dots,n_k(j))_{j\in\mathbf{N}}$ be a sequence in $\mathbf{N}^k$ with the properties stated above. Given $j \in \mathbf{N}$ and $i \in \{1, \dots, k\}$ choose $a_i(j) \in \mathbf{N}$ such that $a_i(j)^p \le n_i(j) < (a_i(j)+1)^p$, where $p = 2k$. It is easily seen that for some positive constant M, we have

$$\frac{a_{i_1}(j)}{a_{i_2}(j)} \ge M, \qquad 1 \le i_1, i_2 \le k;\ j \ge 1. \tag{2.3}$$

Without loss of generality assume that

$$(a_1(j_1), \dots, a_k(j_1)) \ne (a_1(j_2), \dots, a_k(j_2)) \quad \text{if } j_1 \ne j_2. \tag{2.4}$$

Let $\varepsilon > 0$ be fixed, and let

$$N_j = \left(a_1(j)^p + \cdots + a_k(j)^p\right)^{-1} S_{a_1(j)^p, \dots, a_k(j)^p}, \qquad d(j) = d_{a_1(j)^p, \dots, a_k(j)^p}.$$

It then follows from (2.2), (2.3), (2.4) and condition (b) that there is a positive constant L such that

$$\begin{aligned}
&\sum_{j=1}^{\infty} P\left(\left|N_j - E[N_j]\right| > \varepsilon\right)\\
&\quad \le \varepsilon^{-2} \sum_{j=1}^{\infty} \operatorname{Var}(N_j)\\
&\quad \le 2\varepsilon^{-2} \sum_{j=1}^{\infty} d(j)^2 \left[a_1(j)^p + \cdots + a_k(j)^p\right]^{-1}\\
&\quad = 2\varepsilon^{-2} \sum_{j=1}^{\infty} d(j)^2 \left[\frac{a_1(j)^2 \cdots a_1(j)^2}{a_1(j)^2 \cdots a_k(j)^2} + \cdots + \frac{a_k(j)^2 \cdots a_k(j)^2}{a_1(j)^2 \cdots a_k(j)^2}\right]^{-1}\\
&\qquad \times \left(\prod_{i=1}^{k} a_i(j)^2\right)^{-1}\\
&\quad \le L \sum_{j=1}^{\infty} d(j)^2 \left[a_1(j)^2 \cdots a_k(j)^2\right]^{-1}\\
&\quad \le L \sum_{i_1=1}^{\infty} \cdots \sum_{i_k=1}^{\infty} \left(d_{i_1^p, \dots, i_k^p}\right)^2 \left[i_1^2 \cdots i_k^2\right]^{-1}\\
&\quad \le LK^2 \sum_{i_1=1}^{\infty} \cdots \sum_{i_k=1}^{\infty} \left(i_1^{p(1-\alpha_1)} \cdots i_k^{p(1-\alpha_k)}\right)^{1/p} \left[i_1^2 \cdots i_k^2\right]^{-1}\\
&\quad = LK^2 \sum_{i_1=1}^{\infty} \cdots \sum_{i_k=1}^{\infty} \left[i_1^{1+\alpha_1} \cdots i_k^{1+\alpha_k}\right]^{-1}\\
&\quad < \infty.
\end{aligned}$$

From Lemma 2.2, we deduce that

$$\lim_{j \to \infty} \left(a_1(j)^p + \cdots + a_k(j)^p\right)^{-1} S_{a_1(j)^p, \dots, a_k(j)^p} = b, \qquad P\text{-a.s.} \tag{2.5}$$

Consider now the following interpolation argument:

Let $j \in \mathbf{N}$ be fixed and let $(n_1, \ldots, n_k) = (n_1(j), \ldots, n_k(j))$. There is exactly one vector $(a_1, \ldots, a_k) = (a_1(j), \ldots, a_k(j)) \in \{(a_1(i), \ldots, a_k(i)) \colon i \in \mathbf{N}\}$ such that

$$\begin{aligned} a_i^p \le n_i < (a_i + 1)^p &= \sum_{j=0}^{p} \binom{p}{j} a_i^{p-j} \\ &\le a_i^p + (2^p - 1) a_i^{p-1}, \qquad i = 1, \ldots, k. \end{aligned}$$

It follows that

$$0 \le n_i - a_i^p < (2^p - 1) a_i^{p-1} \le (2^p - 1) n_i^{(p-1)/p}, \qquad i = 1, \ldots, k, \tag{2.6}$$

and thus for sufficiently large j,

$$a_i^{-p} < \left(n_i - (2^p - 1) n_i^{(p-1)/p}\right)^{-1}, \qquad i = 1, \ldots, k. \tag{2.7}$$

Letting $T(i_1, \ldots, i_k) = (i_1 + \cdots + i_k)^{-1} S_{i_1, \ldots, i_k}$, $(i_1, \ldots, i_k) \in \mathbf{N}^k$, and using the triangle inequality, we have

$$\begin{aligned} &|(n_1 + \cdots + n_k)^{-1} S_{n_1, \ldots, n_k} - (a_1^p + \cdots + a_k^p)^{-1} S_{a_1^p, \ldots, a_k^p}| \\ &\quad = |T(n_1, \ldots, n_k) - T(a_1^p, \ldots, a_k^p)| \\ &\quad \le \sum_{i_1 = a_1^p}^{n_1 - 1} |T(i_1 + 1, n_2, \ldots, n_k) - T(i_1, n_2, \ldots, n_k)| \\ &\qquad + \sum_{i_2 = a_2^p}^{n_2 - 1} |T(a_1^p, i_2 + 1, n_3, \ldots, n_k) - T(a_1^p, i_2, n_3, \ldots, n_k)| \\ &\qquad + \cdots + \sum_{i_k = a_k^p}^{n_k - 1} |T(a_1^p, a_2^p, \ldots, a_{k-1}^p, i_k + 1) - T(a_1^p, \ldots, a_{k-1}^p, i_k)|. \end{aligned}$$

It will be seen that each of the k sums (depending on j) in this upper estimate tends to zero as $j \to \infty$. Since the reasoning is the same for each sum, only the first sum is considered. From conditions (a) and (b), it follows that P-a.s.,

$$\begin{aligned} &|T(i_1 + 1, n_2, \ldots, n_k) - T(i_1, n_2, \ldots, n_k)| \\ &\quad = \frac{\begin{array}{r} |(i_1 + n_2 + \cdots + n_k) S_{i_1 + 1, n_2, \ldots, n_k} \\ - (i_1 + 1 + n_2 + \cdots + n_k) S_{i_1, n_2, \ldots, n_k}| \end{array}}{(i_1 + 1 + n_2 + \cdots + n_k)(i_1 + n_2 + \cdots + n_k)} \\ &\quad \le \frac{|S_{i_1 + 1, n_2, \ldots, n_k} - S_{i_1, n_2, \ldots, n_k}|}{i_1 + 1 + n_2 + \cdots + n_k} \\ &\qquad + \frac{|S_{i_1, n_2, \ldots, n_k}|}{(i_1 + 1 + n_2 + \cdots + n_k)(i_1 + n_2 + \cdots + n_k)} \\ &\quad \le \frac{1}{i_1 + 1 + n_2 + \cdots + n_k} \left(d_{i_1, n_2, \ldots, n_k} + c\right). \end{aligned}$$

Observe that by (2.6),

$$\sum_{i_1=a_1^p}^{n_1-1} \frac{c}{i_1+1+n_2+\cdots+n_k} \le c \sum_{i_1=a_1^p}^{n_1-1} a_1^{-p} = c a_1^{-p}(n_1 - a_1^p)$$
$$< c a_1^{-1}(2^p - 1),$$

where the last term tends to zero as $j \to \infty$. Furthermore, invoking (2.6) and (2.7) and putting $\alpha = \alpha_1 + \cdots + \alpha_k$, straightforward algebra yields

$$\sum_{i_1=a_1^p}^{n_1-1} \frac{d_{i_1, n_2, \dots, n_k}}{i_1+1+n_2+\cdots+n_k}$$
$$\le (2^p - 1) K \left[\left(\frac{n_2}{n_1} \right)^{1-\alpha_2} \cdots \left(\frac{n_k}{n_1} \right)^{1-\alpha_k} \right]^{1/4k}$$
$$\times \frac{n_1^{(k-2-\alpha)/4k}}{\left(1 - (2^p - 1) n_1^{-1/p}\right)^{(4k-1+\alpha_1)/4k}}.$$

Since by assumption $\alpha > k - 2$, we see that the last term tends to zero as $j \to \infty$. Summarizing, we have the following: For each $\varepsilon > 0$, there is a $j_0 \in \mathbf{N}$ such that for each $j \ge j_0$:

$$|T(n_1(j), \dots, n_k(j)) - T(a_1(j)^p, \dots, a_k(j)^p)| < \varepsilon, \qquad P\text{-a.s.}$$

In view of (2.5) the proof of Theorem 2.3 is complete. □

3. Applications.

3.1. *One-sample nearest neighbor statistics.* Consider a sequence $X_1, X_2, \dots$ of i.i.d. random vectors (points) in $\mathbf{R}^d$, $d \ge 1$, with a.e. continuous Lebesgue density $f(\cdot)$, and let $\|\cdot\|$ be an arbitrary norm on $\mathbf{R}^d$. For $i = 1, \dots, n$ and $r = 1, \dots, n-1$, let $N_n^{(r)}(X_i)$ denote the rth-nearest neighbor of X_i among the points $\{X_j: 1 \le j \le n; j \ne i\}$ with respect to $\|\cdot\|$. Note that $N_n^{(r)}$ depends on all X_i, $i = 1, \dots, n$. Obviously, ties may be neglected since their occurrence is an event of probability 0. In what follows, $I\{A\}$ denotes the indicator of an event A. The random variable

$$n^{-1} R_n^{(l,r)} = n^{-1} \sum_{i=1}^{n} I\{X_i = N_n^{(l)}(N_n^{(r)}(X_i))\}$$

is the fraction of points $X_1, \dots, X_n$ which are the lth-nearest neighbor to their own rth-nearest neighbor. It has been studied by various authors [Clark and Evans (1955), Clark (1955), Dacey (1969), Schwarz and Tversky (1980), Cox (1981), Pickard (1982), Henze (1986, 1987)], usually under the ideal model of events within a d-dimensional homogeneous Poisson process.

To state a strong limit theorem for $n^{-1} R_n^{(l,r)}$, let λ denote d-dimensional Lebesgue measure and write μ for $(d-1)$-dimensional Hausdorff measure (surface area) normalized such that $\mu\{x \in \mathbf{R}^d\colon \|x\| = 1\} = 1$. Generically, $S(x, \rho)$ is the open $\|\cdot\|$-sphere with radius ρ centered at x, and $\mathbf{0} = (0, \dots, 0)$ is shorthand for the origin in $\mathbf{R}^d$. For u with $\|u\| = 1$, let

$$p(u) = \frac{\lambda[S(\mathbf{0},1) \cap S(u,1)]}{\lambda[S(\mathbf{0},1)]}, \qquad q(u) = \frac{\lambda[S(\mathbf{0},1)]}{\lambda[S(\mathbf{0},1) \cup S(u,1)]}.$$

Observe that $q(u) = (2 - p(u))^{-1}$. We finally write

$$\mathbf{b}(m,j,p) = \binom{m}{j} p^j (1-p)^{m-j}, \qquad \mathbf{w}(m,j,p) = \binom{m-1+j}{m-1} p^m (1-p)^j$$

for the probabilities of the binomial and negative binomial distribution, respectively.

THEOREM 3.1. *We have*

$$\lim_{n\to\infty} n^{-1} R_n^{(l,r)} = t_r(l), \qquad P\text{-}a.s.,$$

where

$$t_r(l) = \int_{\|u\|=1} \sum_{j=0}^{\kappa} \mathbf{b}(r-1, j, p(u)) \mathbf{w}(r, l-1-j, q(u)) \mu(du)$$

and $\kappa = \min(r-1, l-1)$.

PROOF. Clearly $R_n^{(l,r)}$ is a symmetric function of $X_1, \dots, X_n$ satisfying $|n^{-1} R_n^{(l,r)}| \le 1$ a.s. It was shown in Henze [(1987), Theorem 1.1] that $\lim E[n^{-1} R_n^{(l,r)}] = t_r(l)$. From Corollary S1 of Bickel and Breiman (1983), which may be easily generalized to rth nearest neighbors, we deduce that there is a universal positive constant Δ_r depending only on r and $\|\cdot\|$ such that, for any set $z_1, \dots, z_n$ of n distinct points in $\mathbf{R}^d$, z_1 can be the rth nearest neighbor for at most Δ_r other points. This entails

$$|R_{n+1}^{(l,r)} - R_n^{(l,r)}| \le \Delta, \qquad P\text{-a.s.},$$

for a constant Δ depending only on r, l and $\|\cdot\|$, so that the assertion follows immediately from Theorem 2.3. □

Another interesting problem concerning nearest neighbors is the fact that, although each point X_i has a unique nearest neighbor, it is not necessarily the nearest neighbor of precisely one other point. The problem of finding the probability that a random point is the nearest neighbor of precisely s other points is of interest in various fields [Tversky and Rinott (1983), Maloney (1983)] and has been investigated in the situation of a homogeneous d-dimensional Poisson process [Roberts (1969), Newman, Rinott and Tversky (1983), Newman and Rinott (1985)]. Let

$$n^{-1}T_n^{(s)} = n^{-1}\sum_{j=1}^{n} I\left\{\sum_{\substack{i=1\\ i\neq j}}^{n} I\{X_j = N_n^{(1)}(X_i)\} = s\right\}$$

be the fraction of random points $X_1, \dots, X_n$ that are the nearest neighbor of precisely s other points.

THEOREM 3.2. *We have*

$$\lim_{n\to\infty} n^{-1}T_n^{(s)} = \mathbf{p}(s), \qquad P\text{-}a.s.,$$

where

$$\mathbf{p}(s) = \frac{1}{s!}\sum_{\nu=0}^{\infty}\frac{1}{\nu!}(-1)^{\nu}\delta_{s+\nu}, \qquad s \geq 0,$$

$$\delta_r = \int \underset{\Gamma_r}{\cdots} \int \exp\left[-\lambda\left(\bigcup_{i=1}^{r} S(x_i, |x_i|)\right)\right] dx_1 \cdots dx_r,$$

$$\Gamma_r = \left\{(x_1, \dots, x_r) \in [\mathbf{R}^d]^r : |x_j| < \min_{1\leq\nu\leq r;\, \nu\neq j} |x_j - x_\nu|, 1 \leq j \leq r\right\}.$$

REMARK. Observe that $\Gamma_r = \varnothing$ and thus $\delta_r = 0$ for sufficiently large r.

PROOF. Henze [(1987), Theorem 1.4] showed that $\lim_{n\to\infty} E[n^{-1}T_n^{(s)}] = \mathbf{p}(s)$. The assertion now follows from Theorem 2.3 by analogy with the reasoning given in the proof of Theorem 3.1. □

In connection with a nonparametric multivariate two-sample test [Henze (1988), see also Schilling (1986)] the nearest neighbor graph statistic

$$C_n^{(r)} = (nr)^{-1}\sum_{j=1}^{n}\left(D_{n,j}^{(r)} - r\right)^2$$

is of interest. Here

$$D_{n,j}^{(r)} = \sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{\nu=1}^{r} I\{X_j = N_n^{(\nu)}(X_i)\}$$

is the number of points $X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_n$ for which X_j is one of the rth nearest neighbors. In terms of graph theory, $D_{n,j}^{(r)}$ is the indegree of vertex X_j in the union of the nearest, second nearest, ..., rth nearest neighbor graph of $X_1, \dots, X_n$. In this way, $C_n^{(r)}$ may be regarded as an empirical variance of indegrees.

THEOREM 3.3. *We have*

$$\lim_{n\to\infty} C_n^{(r)} = 1 - r + r^{-1} \sum_{l,s=1}^{r} c(l,s), \qquad P\text{-}a.s., \tag{3.1}$$

where

$$c(l,s) = \sum_{\alpha,\beta=0}^{1} \sum_{\nu=0}^{\bar{\nu}} \frac{1}{\nu!\delta!\eta!} \int\int_{S_{\alpha,\beta}} \lambda(S_1 \cap S_2)^{\nu} \lambda(S_1 \setminus S_2)^{\delta} \lambda(S_2 \setminus S_1)^{\eta}$$
$$\times \exp[-\lambda(S_1 \cup S_2)]\, du_1\, du_2,$$
$$\bar{\nu} = \min(l + \alpha - 2, s + \beta - 2), \qquad \delta = l - \nu + \alpha - 2,$$
$$\eta = s - \nu + \beta - 2,$$
$$S_{\alpha,\beta} = \left\{(u_1, u_2) \in [\mathbf{R}^d]^2 : \mathbf{0} \in S(u_1, |u_1 - u_2|)^{\alpha} \cap S(u_2, |u_1 - u_2|)^{\beta}\right\},$$
$$S_j = S\left(u_j, |u_j|\right) \quad \textit{for } j = 1, 2 \textit{ and } A^1 = A,\ A^0 = A^c \textit{ for a set } A.$$

PROOF. Some algebra and symmetry give

$$E\left[C_n^{(r)}\right] = 1 - r + r^{-1} \sum_{l,s=1}^{r} (n-1)(n-2)$$
$$\times P\left(X_3 = N_n^{(l)}(X_1), X_3 = N_n^{(s)}(X_1)\right).$$

By straightforward but tedious calculations along the lines of Henze [(1987), Section 3] it can be shown that the expectation of $C_n^{(r)}$ converges to the right-hand side of (3.1). Corollary S1 of Bickel and Breiman (1983) implies that the conditions of Theorem 2.3 are fulfilled for $S_n := nC_n^{(r)}$. Since $C_n^{(r)}$ is a symmetric function of $X_1, \dots, X_n$, the assertion follows. □

3.2. *Nearest neighbor comparisons for two samples.* Consider two sequences (samples) $X_1, X_2, \dots, X_{n_1}, \dots; Y_1, Y_2, \dots, Y_{n_2}, \dots$ of independent d-dimensional random vectors (points), where $X_1, X_2, \dots$ $(Y_1, Y_2, \dots)$ are identically distributed according to a Lebesgue density $f(\cdot)$ $(g(\cdot))$ which is assumed to be continuous a.e. Let

$$Z_j = \begin{cases} X_i, & \text{if } 1 \le i \le n_1, \\ Y_{i-n_1}, & \text{if } n_1 + 1 \le i \le n_1 + n_2, \end{cases}$$

and put $n = n_1 + n_2$. Define $N_n^{(r)}(Z_j)$ to be the rth nearest neighbor of Z_j among $Z_1, \dots, Z_{j-1}, Z_{j+1}, \dots, Z_n$, and let

$$I_j(r) = I\left\{Z_j \text{ and } N_n^{(r)}(Z_j) \text{ belong to the same sample}\right\}.$$

Then

$$T_{n_1,n_2}^{(r)} = \sum_{j=1}^{n} \sum_{\nu=1}^{r} I_j(\nu)$$

is the number of all νth nearest neighbor comparisons ($\nu = 1, \dots, r$) in which points and their neighbors belong to the same sample. $T^{(r)}_{n_1,n_2}$ may be used as a statistic for testing the hypothesis H_0: $f = g$ a.e against general alternatives [Schilling (1986), Henze (1988)].

THEOREM 3.4. *As $n_1, n_2 \to \infty$ such that $n_1/(n_1 + n_2) \to \tau$, $0 < \tau < 1$, we have*

$$\lim(nr)^{-1}T^{(r)}_{n_1,n_2} = D(f,g,\tau), \qquad P\text{-}a.s.,$$

where

$$D(f,g,\tau) = \int \frac{\tau^2 f(x)^2 + (1-\tau)^2 g(x)^2}{\tau f(x) + (1-\tau)g(x)}\,dx. \tag{3.2}$$

PROOF. The proof of $\lim E[(nr)^{-1}T^{(r)}_{n_1,n_2}] = D(f,g,\tau)$ is given for the case $r = 1$ in Henze (1988). The general case $r > 1$ follows similarly. Obviously, $T^{(r)}_{n_1,n_2}$ is a symmetric function within each of the two samples. Since $|(nr)^{-1}T^{(r)}_{n_1,n_2}| \le 1$ and

$$|T^{(r)}_{n_1,n_2} - T^{(r)}_{n_1,n_2+1}| \le \Delta, \qquad |T^{(r)}_{n_1,n_2} - T^{(r)}_{n_1+1,n_2}| \le \Delta$$

for a constant Δ depending only on r and the chosen norm $\|\cdot\|$ [use again Corollary S1 of Bickel and Breiman (1983)], Theorem 2.3 yields the assertion. □

REMARK. Theorem 3.4 may be generalized to the case of k independent samples. If $f_j(\cdot)$ denotes the density of points from the jth sample of size n_j and $T^{(r)}_{n_1,\dots,n_k}$ stands for the number of all νth nearest neighbor type coincidences ($\nu = 1,\dots,r$), we have, as $n_j \to \infty$ with $n_j/(n_1 + \cdots + n_k) \to \tau_j$, $0 < \tau_j < 1$, $j = 1,\dots,k$:

$$\lim[(n_1 + \cdots + n_k)r]^{-1}T^{(r)}_{n_1,\dots,n_k} = \int \frac{\sum_{j=1}^k \tau_j^2 f_j(x)^2}{\sum_{j=1}^k \tau_j f_j(x)}\,dx, \qquad P\text{-a.s.}$$

3.3. *Runs and empty blocks.* Let $X_1, X_2, \dots, X_{n_1}, \dots$; $Y_1, Y_2, \dots, Y_{n_2}, \dots$ be two samples of independent real-valued random variables, where $X_1, X_2, \dots (Y_1, Y_2, \dots)$ are identically distributed according to a Lebesgue density $f(\cdot)$ ($g(\cdot)$) which is assumed to be continuous a.e. Let R_{n_1,n_2} denote the total number of runs (sequences of maximal length within the same sample) when the pooled sample $X_1,\dots,X_{n_1},Y_1,\dots,Y_{n_2}$ is arranged in ascending order [Wald and Wolfwitz (1940)]. As above, let $n = n_1 + n_2$. The following strong law of large numbers for R_{n_1,n_2} seems to be new.

THEOREM 3.5. *As $n_1, n_2 \to \infty$ such that $n_1/(n_1 + n_2) \to \tau$, $0 < \tau < 1$, we have*

$$\lim n^{-1}R_{n_1,n_2} = 1 - D(f,g,\tau), \qquad P\text{-}a.s.,$$

with $D(f,g,\tau)$ given in (3.2).

PROOF. Let

$$R_{ix}^{(1)} = \min\{X_j - X_i : j = 1, \dots, n_1, X_j > X_i\},$$

$$R_{iy}^{(1)} = \min\{Y_j - X_i : j = 1, \dots, n_2, Y_j > X_i\},$$

$$R_{ix}^{(2)} = \min\{X_j - Y_i : j = 1, \dots, n_1, X_j > Y_i\},$$

$$R_{iy}^{(2)} = \min\{Y_j - Y_i : j = 1, \dots, n_2, Y_j > Y_i\},$$

with the convention $\min \varnothing = \infty$. Then

$$R_{n_1,n_2} = 1 + \sum_{i=1}^{n_1} I\{R_{iy}^{(1)} < R_{ix}^{(1)}\} + \sum_{i=1}^{n_2} I\{R_{ix}^{(2)} < R_{iy}^{(2)}\}$$

which, in other words, is one plus the number of points in the pooled sample whose nearest neighbor *to the right* is of different sample type. By conditioning on X_1, respectively, Y_1 and arguing along the lines of Henze [(1988), Theorem 4.1] we have

$$\lim P\left(R_{iy}^{(1)} < R_{ix}^{(1)}\right) = \int \frac{(1-\tau)g(x)}{\tau f(x) + (1-\tau)g(x)} f(x)\,dx,$$

$$\lim P\left(R_{ix}^{(2)} < R_{iy}^{(2)}\right) = \int \frac{\tau f(x)}{\tau f(x) + (1-\tau)g(x)} g(x)\,dx,$$

which entails

$$\begin{aligned}\lim E\left[n^{-1}R_{n_1,n_2}\right] &= 2\tau(1-\tau)\int \frac{f(x)g(x)}{\tau f(x) + (1-\tau)g(x)}\,dx \\ &= 1 - D(f, g, \tau).\end{aligned}$$

Since R_{n_1,n_2} is a symmetric function within each sample satisfying $|n^{-1}R_{n_1,n_2}| \le 1$ and $|R_{n_1,n_2} - R_{n_1+1,n_2}| \le 2$, $|R_{n_1,n_2} - R_{n_1,n_2+1}| \le 2$, the assertion follows from Theorem 2.3. □

Observe that $D(f, g, \tau) \ge \tau^2 + (1-\tau)^2$ with equality if, and only if, $f = g$ a.e. Consequently, Theorem 3.5 yields a simple consistency proof of the run test under weak restrictions on the densities f and g.

Let, in the situation stated at the beginning of 3.3, $X_{(1)} < \cdots < X_{(n_1)}$ be the ordered X-sample and let $B_1 = (-\infty, X_{(1)}]$, $B_j = (X_{(j-1)}, X_{(j)}]$, $j = 2, \dots, n_1$, $B_{n_1+1} = (X_{(n_1)}, \infty)$ the blocks generated by $X_1, \dots, X_{n_1}$. Then

$$E_{n_1,n_2} = \sum_{i=1}^{n_1+1} I\left\{\bigcap_{j=1}^{n_2} \{Y_j \notin B_i\}\right\}$$

is the number of empty X-blocks which may be used to test the hypothesis $f = g$ a.e. [Wilks (1962)]. Also the following strong limit law for E_{n_1,n_2} seems to be new.

THEOREM 3.6. *As $n_1, n_2 \to \infty$ such that $n_1/(n_1 + n_2) \to \tau$, $0 < \tau < 1$, we have*

$$\lim n^{-1} E_{n_1, n_2} = \int \frac{\tau^2 f(x)^2}{\tau f(x) + (1 - \tau) g(x)}\, dx, \qquad P\text{-}a.s. \tag{3.3}$$

PROOF. Observe that, with the notation of the proof of Theorem 3.5,

$$\left| E_{n_1, n_2} - \sum_{i=1}^{n_1} I\{R_{iy}^{(1)} > R_{ix}^{(1)}\} \right| \le 2.$$

Since the right-hand side of (3.3) is the almost sure limit of $n^{-1}\sum_{i=1}^{n_1} I\{R_{iy}^{(1)} > R_{ix}^{(1)}\}$ (see the proof of Theorem 3.5), we are done. □

Acknowledgment. The authors would like to thank an Associate Editor for his careful reading of the manuscript.

REFERENCES

BHARGAVA, R. P. (1983). A property of the jackknife estimation of the variance when more than one observation is omitted. *Sankhyā Ser. A* **45** 112–119.

BICKEL, P. J. and BREIMAN, L. (1983). Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *Ann. Probab.* **11** 185–214.

CLARK, P. J. (1955). Grouping in spatial distributions. *Science* **123** 373–374.

CLARK, P. J. and EVANS, F. C. (1955). On some aspects of spatial pattern in biological populations. *Science* **121** 397–398.

COX, T. F. (1981). Reflexive nearest neighbours. *Biometrics* **37** 367–369.

DACEY, M. F. (1969). Proportion of reflexive nth order neighbors in spatial distribution. *Geographical Analysis* **1** 385–388.

DEVROYE, L. (1987). An application of the Efron–Stein inequality in density estimation. *Ann. Statist.* **15** 1317–1320.

EFRON, B. and STEIN, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9** 586–596.

HENZE, N. (1986). On the probability that a random point is the jth nearest neighbor to its own kth nearest neighbor. *J. Appl. Probab.* **23** 221–226.

HENZE, N. (1987). On the fraction of random points with specified nearest neighbour interrelations and 'degree of attraction'. *Adv. in Appl. Probab.* **19** 873–895.

HENZE, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist.* **16** 772–783.

HOCHBAUM, D. and STEELE, J. M. (1982). Steinhaus's geometric location problem for random samples in the plane. *Adv. in Appl. Probab.* **14** 56–67.

KARLIN, S. and RINOTT, Y. (1982). Application of ANOVA type decompositions of conditional variance statistics including jackknife estimates. *Ann. Statist.* **10** 485–501.

MALONEY, L. T. (1983). Nearest neighbor analysis of point processes: Simulations and evaluations. *J. Math. Psychology* **27** 251–260.

NEWMAN, C. M. and RINOTT, Y. (1985). Nearest neighbors and Voronoi regions in high-dimensional point processes with various distance functions. *Adv. in Appl. Probab.* **17** 794–809.

NEWMAN, C. M., RINOTT, Y. and TVERSKY, A. (1983). Nearest neighbors and Voronoi regions in certain point processes. *Adv. in Appl. Probab.* **15** 726–751.

PICKARD, D. K. (1982). Isolated nearest neighbors. *J. Appl. Probab.* **19** 444–449.

RHEE, W. T. and TALAGRAND, M. (1986). Martingale inequalities and the jackknife estimate of variance. *Statist. Probab. Lett.* **4** 5–6.

ROBERTS, F. D. K. (1969). Nearest neighbours in a Poisson ensemble. *Biometrika* **56** 401–406.

SCHILLING, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* **81** 799–806.

SCHWARZ, G. and TVERSKY, A. (1980). On the reciprocity of proximity relations. *J. Math. Psychology* **22** 157–175.

STEELE, J. M. (1981). Complete convergence of short paths and Karp's algorithm for the tsp. *Math. Oper. Res.* **6** 374–378.

STEELE, J. M. (1982). Optimal triangulation of random samples in the plane. *Ann. Probab.* **10** 548–553.

STEELE, J. M. (1986). An Efron–Stein inequality for nonsymmetric statistics. *Ann. Statist.* **14** 753–758.

STEELE, J. M., SHEPP, L. A. and EDDY, W. F. (1987). On the number of leaves of a euclidean minimal spanning tree. *J. Appl. Probab.* **24** 809–826.

TVERSKY, A. and RINOTT, Y. (1983). Nearest neighbor analysis of point processes: Applications to multidimensional scaling. *J. Math. Psychology* **27** 235–250.

VITALE, R. A. (1984). An expansion for symmetric statistics and the Efron–Stein inequality. In *Inequalities in Statistics and Probability* (Y. L. Tong, ed.). IMS *Lecture Notes–Monograph Ser.* **5** 112–114.

VITALE, R. A. (1988). A differential version of the Efron–Stein inequality: Bounding the variance of a function of an infinitely divisible variable. *Statist. Probab. Letters* **7** 105–112.

WALD, A. and WOLFOWITZ, J. (1940). On a test whether two samples are from the same population. *Ann. Math. Statist.* **11** 147–162.

WILKS, S. S. (1962). *Mathematical Statistics*. Wiley, New York.

INSTITUT FÜR MATHEMATISCHE STOCHASTIK
UNIVERSITÄT KARLSRUHE
ENGLERSTR. 2, D-7500 KARLSRUHE 1
GERMANY