stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK SD 106/74 MAY

F.H. RUYMGAART & M.C.A. VAN ZUYLEN

RANK STATISTICS FOR INDEPENDENCE AND THE k-SAMPLE PROBLEM

Preliminary report

2e boerhaavestraat 49 amsterdam

1. INTRODUCTION

Rank statistics that are essentially intended to provide tests for the hypothesis of independence on the basis of a random sample from one single bivariate distribution function (df) may, formally, be computed in the same way when the sample elements originate from k possibly different bivariate dfs (throughout k will denote a fixed but arbitrary positive integer). In such a case we ignore the fact that we are dealing with k samples rather than one and compute the rank statistics on the basis of the combined sample.

Large absolute values of these statistics, however, need no longer give evidence for positive or negative dependence between the coordinates of each or any of the sample elements. Large absolute values may e.g. also occur when the coordinates of each sample element are independent but the k bivariate dfs are all different and strongly concentrated along some monotone curve in the plane.

We shall not dwell on the practical merits of rank statistics for independence computed under these non-standard conditions but concentrate on some points concerning the (asymptotic) distribution of the statistics. In the first place we shall prove that these rank statistics have a normal distribution in the limit when the samples are drawn from k fixed arbitrary bivariate dfs. For $k=1$ results of this kind may be found in e.g. [1], [5], [8], [9] and [10]. Secondly we shall point out how our first result for k samples from bivariate dfs entails as a corollary the asymptotic distribution of rank statistics, used in the univariate k-sample problem, when the k samples are drawn from k fixed arbitrary univariate dfs. In the latter case the asymptotic distribution of rank statistics for the univariate k-sample problem has already been obtained in e.g. [2] (for $k=2$), [3], [4], [5] and [6]. Therefore our main objective is to display the relationship between rank statistics for independence as computed for the combination of the k bivariate samples and rank statistics for the univariate k-sample problem.

To be more precise let us suppose that for $i = 1, \dots, k$ the two-dimensional random vector $(X^{(i)}, Y^{(i)})$ has bivariate df $H_{(i)}(x,y) = \Pr(X^{(i)} \leq x, Y^{(i)} \leq y)$ and marginal dfs $F_{(i)}(x) = \Pr(X^{(i)} \leq x)$ and $G_{(i)}(y) = \Pr(Y^{(i)} \leq y)$ for $x, y \in (-\infty, \infty)$. Let us also suppose that for each k-tuple of positive integers $N(1), \dots, N(k)$ we are given k mutually independent random samples

$$(1.1) \qquad (X_1^{(i)}, Y_1^{(i)}), \dots, (X_{N(i)}^{(i)}, Y_{N(i)}^{(i)}) \qquad \text{for } i = 1, \dots, k,$$

where the i-th sample, having upper index i, has been drawn from the bivariate df $H_{(i)}$. All samples are supposed to be defined on a single probability space $(\Omega, \mathcal{A}, P)$. Given any finite set D the number of elements in D will be denoted by $\# D$. For the i-th sample we define the bivariate empirical df $H_{N(i)}^{(i)}$ by $N(i) H_{N(i)}^{(i)}(x,y) = \#\{(X_n^{(i)}, Y_n^{(i)}) : X_n^{(i)} \leq x, Y_n^{(i)} \leq y, n = 1, \dots, N(i)\}$ and the marginal empirical dfs $F_{N(i)}^{(i)}$ and $G_{N(i)}^{(i)}$ by $N(i) F_{N(i)}^{(i)}(x) = \#\{X_n^{(i)} : X_n^{(i)} \leq x, n = 1, \dots, N(i)\}$ respectively $N(i) G_{N(i)}^{(i)}(y) = \#\{Y_n^{(i)} \leq y, n = 1, \dots, N(i)\}$, where $x, y \in (-\infty, \infty)$. The rank $R_{nN(i)}^{(i)}$ of $X_n^{(i)}$ and $Q_{nN(i)}^{(i)}$ of $Y_n^{(i)}$ is defined as $\#\{X_m^{(i)} : X_m^{(i)} \leq X_n^{(i)}, m = 1, \dots, N(i)\}$ respectively $\#\{Y_m^{(i)} : Y_m^{(i)} \leq Y_n^{(i)}, m = 1, \dots, N(i)\}$, for $n = 1, \dots, N(i)$. Finally the ordered set of first and second coordinates will be denoted by $X_{1:N(i)}^{(i)} \leq \dots \leq X_{N(i):N(i)}^{(i)}$ respectively $Y_{1:N(i)}^{(i)} \leq \dots \leq Y_{N(i):N(i)}^{(i)}$. We have the relations

$$R_{nN(i)}^{(i)} = N(i) F_{N(i)}^{(i)}(X_n^{(i)}), \quad Q_{nN(i)}^{(i)} = N(i) G_{N(i)}^{(i)}(Y_{N(i)}^{(i)}).$$

Let us put $N = \sum_{i=1}^{k} N(i)$, $\nu_i = \nu_i(N) = N(i)/N$ for $i = 1, \dots, k$ and $\nu = (\nu_1, \dots, \nu_k)$. The combined sample consists of all N random vectors given in (1.1). We shall denote an arbitrary ordering of the random vectors contained in the combined sample by

$$(1.2) \qquad (X_1, Y_1), \dots, (X_N, Y_N).$$

We define $\overline{H}_{(\nu)} = \sum_{i=1}^{k} \nu_i H_{(i)}$, $\overline{F}_{(\nu)} = \sum_{i=1}^{k} \nu_i F_{(i)}$ and $G_{(\nu)} = \sum_{i=1}^{k} \nu_i G_{(i)}$. Let us observe that $\overline{H}_{(\nu)}$ has all the properties of a bivariate df and that its marginal dfs are $\overline{F}_{(\nu)}$ and $\overline{G}_{(\nu)}$. For the combined sample we define the bivariate empirical of H_N by $NH_N(x,y) = \#\{(X_n, Y_n): X_n \leq x, Y_n \leq y, n=1,\dots,N\}$ and the marginal empirical dfs F_N and G_N by $NF_N(x) = \#\{X_n : X_n \leq x, n=1,\dots,N\}$ respectively $NG_N(y) = \#\{Y_n : Y_n \leq y, n=1,\dots,N\}$, where $x, y \in (-\infty,\infty)$. The rank R_{nN} of X_n and Q_{nN} of Y_n is defined as $\#\{X_m : X_m \leq X_n, m=1,\dots,N\}$ respectively $\#\{Y_m : Y_m \leq Y_n, m=1,\dots,N\}$ for $n=1,\dots,N$. The ordered set of first and second coordinates will be denoted by $X_{1:N} \leq \dots \leq X_{N:N}$ respectively $Y_{1:N} \leq \dots \leq Y_{N:N}$. We have the relations $H_N = \sum_{i=1}^k \nu_i H_{N(i)}^{(i)}$, $F_N = \sum_{i=1}^k \nu_i F_{N(i)}^{(i)}$, $G_N = \sum_{i=1}^k \nu_i G_{N(i)}^{(i)}$ and

$$R_{nN} = N F_N(X_n), \quad Q_{nN} = N G_N(Y_n).$$

The rank statistics mentioned in the first paragraph that we are interested in are of the type

$$N^{-1} \sum_{n=1}^N a_N(R_{nN})\, b_N(Q_{nN}), \tag{1.3}$$

where, for each $N=1,2,\dots$, $a_N(n)$ and $b_N(n)$ are real numbers, called scores, for $n=1,\dots,N$. We shall exclusively be concerned with the case where the scores are generated by sufficiently smooth functions J and K defined on $(0,1)$, referred to as scores generating functions, according to

$$a_N(n) = J(n/(N+1)), \quad b_N(n) = K(n/(N+1)). \tag{1.4}$$

Introducing the modified empirical dfs

$$F_{N(i)}^{(i)*} = [N/(N+1)]\, F_{N(i)}^{(i)}, \quad G_{N(i)}^{(i)*} = [N/(N+1)]\, G_{N(i)}^{(i)},$$

$$F_N^* = [N/(N+1)]\, F_N, \quad G_N^* = [N/(N+1)]\, G_N.$$

we may write the statistic (1.3) in the convenient form

$$T_N = \iint_{x,y\in(-\infty,\infty)} J(F_N^*(x))\, K(G_N^*(y))\, dH_N(x,y) = \iint J(F_N^*)\, K(G_N^*)\, dH_N \tag{1.5}$$

In contrast to [1], [5], [8], [9] and [10] we shall allow the scores generating function to depend on N or, more generally, on the vector ν of relative sample sizes, so that we should write $J_{(\nu)}$ and $K_{(\nu)}$ rather than J and K. In order not to overdo the notation we shall, however, adhere to the latter notation and suppress subscript. This dependence on ν, which is essential for our present purposes, does not affect the root of the regularity conditions and will not bring about additional technical complications.

Statistics of the kind given in (1.5) are generally used to detect those deviations from the null hypothesis that

$$(1.6) \qquad \text{both } H_{(i)} = F_{(i)} \times G_{(i)} \text{ for } i=1,\dots,k \text{ and } H_{(1)} = \dots = H_{(k)},$$

for which the first condition in (1.6) is not fulfilled. Indeed T_N will in general assume large absolute values in the case where all $H_{(i)}$ are equal but not of product type, more in particular in the case where X and Y are positively or negatively dependent. However, the following example makes clear that large absolute values of T_N may also occur when all $H_{(i)}$ are of product type, and hence $X^{(i)}$ and $Y^{(i)}$ independent for $i=1,\dots,k$, but not equal. Let

$$(1.7) \qquad \mathcal{U}_{(i)}(z) \qquad \text{for } z \in (-\infty, \infty) \text{ and } i=1,\dots,k$$

denote the df of the uniform distribution over $(i-1, i)$ and define

$$H_{(i)}(x,y) = \mathcal{U}_{(i)}(x)\,\mathcal{U}_{(i)}(y) \qquad \text{for } x,y \in (-\infty,\infty) \text{ and } i=1,\dots,k.$$

In this case T_N will assume large absolute values because the supports of the dfs are concentrated along the main diagonal in the plane and the supports are pairwise disjunct.

Our main theorem, Theorem 2.1, asserts the asymptotic normality of statistics of the type (1.5) for a fixed arbitrary k-tuple of underlying bivariate dfs. In Theorem 2.2 we present the asymptotic normality of linear combinations of statistics that play a basic role in the univariate k-sample problem for a fixed k-tuple of underlying univariate dfs. The result of Theorem 2.2 is well known. However, we shall give a new proof of Theorem 2.2 by pointing out that it is an almost immediate modification of the proof of Theorem 2.1.

Let us conclude this section by setting forth this relationship and choose

(1.8) $\qquad H_{(i)}(x,y) = F_{(i)}(x)\,\mathcal{U}_{(i)}(y)$ for $(x,y)\in(-\infty,\infty)\times(-\infty,\infty)$ and $i=1,\dots,k$.

Here $F_{(i)}$ is an arbitrary univariate df and $\mathcal{U}_{(i)}$ is defined in (1.7). Let us introduce the modified relative sample sizes

$$\nu_i^* = [N/(N+1)]\,\nu_i,$$

and observe that for the special choice of the $H_{(i)}$ in (1.8) we have

(1.9) $\qquad P(\bigcap_{i=1}^{k} \bigcap_{n=1}^{N(i)} \{Q_{nN}/(N+1) \in (\sum_{j=1}^{i-1} \nu_i^*, \sum_{j=1}^{i} \nu_i^*]\}) = 1.$

Let us finally introduce the functions

(1.10) $\qquad \chi_i^*(t) = 1 \quad$ for $t \in (\sum_{j=1}^{i-1} \nu_i^*, \sum_{j=1}^{i} \nu_i^*]$,

$\qquad\qquad = 0$ elsewhere,

(1.11) $\qquad \chi_\lambda^*(t) = \sum_{i=1}^{k} \lambda_i\, \chi_i^*(t) \qquad$ for $t\in(0,1)$.

The dependence on N (through ν) of these functions is obvious.

Test statistics for the univariate k-sample problem are based on statistics of the type

(1.12) $\qquad N^{-1} \sum_{n:\, X_n = X_m^{(i)} \text{ for some } m} a_N\,(R_{nN}) \quad$ for $i=1,\dots,k$,

where $a_N(n)$ are scores derived from a scores generating function J as in (1.4). Joint asymptotic normality of the k rank statistics in (1.12) is equivalent to asymptotic normality of each linear combination

(1.13) $\qquad N^{-1} \sum_{i=1}^{k} \lambda_i \sum_{n:\, X_n = X_m^{(i)} \text{ for some } m} a_N\,(R_{nN})$.

It is our purpose to prove asymptotic normality of statistics of type (1.13) for a vector $\lambda = (\lambda_1,\dots,\lambda_k)$ of fixed arbitrary constants.

This is achieved in Theorem 2.2 by writing the statistics (1.13) in the form (1.5). Let us introduce

$$(1.14)\qquad T_N^{(i)} = \iint J(F_N^*)\,\chi_i^*(G_N^*)\,dH_N ,$$

$$(1.15)\qquad T_{\lambda N} = \sum_{i=1}^{k} \lambda_i T_N^{(i)} = \iint J(F_N^*)\,\chi_\lambda^*(G_N^*)\,dH_N .$$

The statistics in (1.15) arise as special cases from those in (1.5) by choosing $K = \chi_\lambda^*$. Provided we choose the $H_{(i)}$ according to (1.8) it is an immediate consequence of (1.9) and (1.11) that the statistic $T_{\lambda N}$ in (1.15) equals, with probability 1, the statistic in (1.13). We summarize this result in the next theorem.

<u>THEOREM 1.1.</u> Provided we choose $H_{(1)}, \ldots, H_{(k)}$ according to (1.8) we have

$$(1.16)\qquad P(\{ N^{-1}\sum_{i=1}^{k} \lambda_i \sum_{n:\, X_n = X_m^{(i)} \text{ for some } m} a_N(R_{nN}) = T_{\lambda N} \}) = 1 .$$

Here $\lambda = (\lambda_1, \ldots, \lambda_k)$ is a vector of fixed arbitrary constants and $T_{\lambda N}$ is defined in (1.15).

As a special case let us take $k = 2$ and $\lambda_1 = 1, \lambda_2 = 0$. For this λ we have $T_{\lambda N} = T_N^{(1)}$ and (with probability 1)

$$(1.17)\qquad N^{-1}\sum_{n:\, X_n = X_m^{(1)} \text{ for some } m} a_N(R_{nN}) = T_N^{(1)}$$

These rank statistics are of the common type in the univariate two-sample problem for testing the null hypothesis that

$$F_{(1)} = F_{(2)} .$$

Taking in this case, moreover,

$$J(s) = 2s - 1 \qquad \text{for } s \in (0,1),$$

we obtain Wilcoxon's two-sample rank statistic.

It should be noted that there is an alternative way to cover asymptotic normality of rank statistics for independence and both univariate and bivariate k-sample problems in one single theorem. Let us take $k=2$ and consider

$$\tilde{T}_N^{(1)} = \iint J(F_N^*)\,K(G_N^*)\,d\,H_{N(1)}^{(1)}. \qquad (1.16)$$

Statistics of the form (1.16) include the following special cases.

1. For J and K appropriately chosen, $\tilde{T}_N^{(1)}$ is a test statistic, suitable for bivariate 2-sample problems.

2. For $K=1$ on $(0,1)$, $\tilde{T}_N^{(1)}$ reduces to the usual test statistic for the univariate 2-sample problem.

3. Let us suppose that $H_{(1)} = H_{(2)}$ and that $N(1) = N$, $N(2) = 0$. Then $\tilde{T}_N^{(1)}$ is the usual test statistic for the independence problem.

It is most likely that asymptotic normality of the statistics $\tilde{T}_N^{(1)}$ can be performed along essentially the same lines as that of the statistics T_N. Nevertheless we shall not deal with this approach here because it seems impossible to include simple linear rank statistics in this way. In Section 4 we conjecture that this will be possible by generalizing the approach started in Theorem 1.1. Moreover, usually statistics for bivariate k-sample problems are combinations of statistics for univariate k-sample problems, so that the set-up chosen here might be the most general after all.

The material is organized as follows. In Section 2 we give the main result on asymptotic normality, Theorem 2.1, and its corollary Theorem 2.2. The proof of these theorems is deferred to the appendix. A straightforward generalization to the multivariate case is considered in Section 3. In Section 4, finally, we conjecture two generalizations of Theorem 2.1 leading to asymptotic normality of rank statistics for univariate symmetry and of simple linear rank statistics in the case of univariate underlying dfs.

2. STATEMENT OF THE THEOREMS

Before presenting the theorems let us introduce some more notation and conventions, to be used throughout the present and the subsequent sections. The symbol M will be employed as a generic constant, not depending on $H_{(i)}$, $F_{(i)}$, $G_{(i)}$ for $i=1,\dots,k$ and also independent of N. For any univariate df Ψ (which is tacitly assumed to be taken right-continuous) we define an inverse Ψ^{-1} on $(0,1)$ by

$$\Psi^{-1}(u) = \inf\{x : \Psi(x) \geq u\} \quad \text{for } u \in (0,1). \tag{2.1}$$

By this definition Ψ^{-1} is left-continuous on $(0,1)$.

For convenience, we shall use only q- and reproducing u-shaped functions (for a definition see [13, Appendix] of a special but common type based on the function

$$R(u) = [u(1-u)]^{-1} \quad \text{for } u \in (0,1). \tag{2.2}$$

We shall also use the notation

$$\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} \exp(-w^2/2)\,dw.$$

Let us fix two positive integers l and m and two sets of points $\mathcal{S}_1 = \{s_1,\dots,s_l\} \subset (0,1)$ and $\mathcal{S}_2 = \{t_1,\dots,t_m\} \subset (0,1)$. Let, moreover, $\gamma > 0$ be a fixed constant chosen such that for each $0 < \zeta \leq \gamma$ we have

$$\mathcal{O}_{1,\varsigma} = \cup_{i=1}^{\ell} (s_i - \varsigma, s_i + \varsigma) \subset (0,1),$$
(2.3)
$$\mathcal{O}_{2,\varsigma} = \cup_{j=1}^{m} (t_j - \varsigma, t_j + \varsigma) \subset (0,1).$$

For each $N = 1, 2, \ldots$ let be given two sets of points

$$\mathcal{S}_{1N} = \{s_{1N}, \ldots, s_{\ell N}\} \subset \mathcal{O}_{1,\gamma/2},$$
(2.4)
$$\mathcal{S}_{2N} = \{t_{1N}, \ldots, t_{mN}\} \subset \mathcal{O}_{2,\gamma/2}.$$

At these and only these points the scores generating functions are allowed to have discontinuities of the first kind. For $N = 1, 2, \ldots$ let us define

$$\Delta_{1iN} = J(s_{iN}+) - J(s_{iN}-), \quad \Delta_{1N} = \Sigma_{i=1}^{\ell} |\Delta_{1iN}|,$$
(2.5)
$$\Delta_{2jN} = K(t_{jN}+) - K(t_{jN}-), \quad \Delta_{2N} = \Sigma_{j=1}^{m} |\Delta_{2jN}|.$$

The underlying dfs $H_{(1)}, \ldots, H_{(k)}$ will always be restricted to the class

$$\mathcal{H} = \{H : H \text{ is a bivariate df, continuous on } (-\infty,\infty) \times (-\infty,\infty)\},$$

so that with probability 1 the ranks of the first respectively second coordinates are all different, as well for each of the k samples separately as for the combined sample. The price for discontinuities in the scores generating function is a kind of local differentiability condition on the transformation $\bar{H}_{(\nu)}(\bar{F}_{(\nu)}^{-1}, \bar{G}_{(\nu)}^{-1})$ of $\bar{H}_{(\nu)}$ to the unit square. We shall say that a density $\bar{h}_{(\nu)}$ (with respect to Lebesgue measure in the unit square) exists on the Borel set $B_0 \subset (0,1) \times (0,1)$ if, for each Borel set $B \subset B_0$,

$$\iint_{(s,t)\in B} d\bar{H}_{(\nu)}(\bar{F}_{(\nu)}^{-1}(s), \bar{G}_{(\nu)}^{-1}(t)) = \iint_{(s,t)\in B} \bar{h}_{(\nu)}(s,t)\, ds\, dt.$$

To describe the properties of $\bar{h}_{(\nu)}$ that we need in our theorems let us introduce a fixed set $\mathcal{N}$ in the open unit interval with Lebesgue-measure zero, i.e.

$$\mathcal{N} \subset (0,1) \quad , \quad \int_{s \in \mathcal{N}} ds = 0 .$$

We are now ready to formulate the conditions for the theorems and start with those needed for Theorem 2.1 which deals with the statistics T_N given in (1.5).

C1. SCORES GENERATING FUNCTIONS

C1.1. For $N=1,2,\ldots$ the scores generating function J is continuously differentiable on the set $(0,1)-\mathcal{D}_{1N}$ and on this set we have, for some $\alpha \geq 0$

$$|J^{(i)}(\cdot)| \leq M\,[R(\cdot)]^{\alpha+i} \qquad \text{for } i=0,1.$$

C1.2. For $N=1,2,\ldots$ the scores generating function K is continuously differentiable on the set $(0,1)-\mathcal{D}_{2N}$ and on this set we have, for some $\beta \geq 0$,

$$|K^{(i)}(\cdot)| \leq M\,[R(\cdot)]^{\beta+i} \qquad \text{for } i=0,1.$$

C2. UNDERLYING DISTRIBUTION FUNCTIONS

C2.1. For $i=1,\ldots,k$ we have $H_{(i)} \in \mathcal{H}$.

C2.2. For $N=1,2,\ldots$ the density $\bar{h}_{(\nu)}$ exists on $\mathcal{O}_{1,\gamma} \times (0,1)$, where $\bar{h}_{(\nu)}(\cdot, t)$ is continuous on $\mathcal{O}_{1,\gamma}$ for all $t \in (0,1)-\mathcal{N}$ and, for some $b \geq 0$,

$$|\bar{h}_{(\nu)}(s,t)| \leq M\,[R(t)]^{b} \quad \text{for } (s,t) \in \mathcal{O}_{1,\gamma} \times (0,1).$$

C2.3. For $N=1,2,\ldots$ the density $\bar{h}_{(\nu)}$ exists on $(0,1) \times \mathcal{O}_{2,\gamma}$, where $\bar{h}_{(\nu)}(s,\cdot)$ is continuous on $\mathcal{O}_{2,\gamma}$ for all $s \in (0,1)-\mathcal{N}$ and, for some $a \geq 0$,

$$\bar{h}_{(\nu)}(s,t) \leq M[R(s)]^{a} \qquad \text{for } (s,t) \in (0,1) \times \mathcal{O}_{2,\gamma}.$$

For the standardization of T_N we shall use the quantities

(2.6) $$\mu_N = \iint J(\bar{\bar{F}}_{(\nu)})\, K(\bar{G}_{(\nu)})\, d\bar{H}_{(\nu)},$$

(2.7) $$\sigma_N^2 = \mathrm{Var}\,(\Sigma_{i=0}^{4}\, A_{iN}),$$

where the A_{iN} are defined in (A.4).

THEOREM 2.1. Suppose that C1.1 and C1.2 hold for fixed α, β with $\alpha+\beta < 1/2$ and that C2.1 is satisfied for the fixed underlying dfs $H_{(1)}, \ldots, H_{(k)}$. Moreover, C2.2 holds for fixed b with $\beta + b < 1$ if $\limsup_{N\to\infty} \Delta_{1N} > 0$, and C2.3 holds for fixed a with $\alpha + a < 1$ if $\limsup_{N\to\infty} \Delta_{2N} > 0$. Let us finally assume that $0 < \liminf_{N\to\infty} \nu_i(N) \leq \limsup_{N\to\infty} \nu_i(N) < 1$ for $i = 1, \ldots, k$.

Then the quantities μ_N and σ_N^2, defined in (2.6) and (2.7), are finite for $N = 1, 2, \ldots$ and

(2.8) $$\lim_{N\to\infty} \sup_{z\in(-\infty,\infty)} |P(\{N^{1/2}(T_N - \mu_N)/\sigma_N \leq z\}) - \Phi(z)| = 0,$$

provided $\liminf_{N\to\infty} \sigma_N^2 > 0$. Here T_N is defined in (1.5).

In the next theorem we deal with the statistics T_{2N} under the additional condition that the $H_{(i)}$ satisfy (1.8). This assumption has the following impact on C2.1 and C2.2.

LEMMA 2.1. Suppose that $H_{(i)} = F_{(i)} \times U_{(i)}$ for $i = 1, \ldots, k$. Then C2.1 is equivalent to the condition that $F_{(i)}$ is continuous on $(-\infty, \infty)$ for $i = 1, \ldots, k$, and C2.2 is equivalent to the condition that $F_{(i)}(\bar{\bar{F}}_{(\nu)}^{-1})$ has a continuous density on $\mathcal{O}_{1,\gamma}$.

PROOF. The first assertion is immediate. As far as the second is concerned let us first observe that we have

(2.9) $\bar{G}_{(\nu)}(y) = \sum_{j=1}^{i-1} \nu_j + \nu_i y \qquad$ for $y \in (i-1, i)$.

because $G_{(i)} = \mathcal{U}_{(i)}$, so that

(2.10) $d\, G_{(i)}(\bar{G}_{(\nu)}^{-1}(t))/dt = 0 \qquad$ for $t \notin (\sum_{j=1}^{i-1} \nu_j, \sum_{j=1}^{i} \nu_j)$,

$= 1/\nu_i \quad$ for $t \in (\sum_{j=1}^{i-1} \nu_j, \sum_{j=1}^{i} \nu_j)$.

Consequently we have under the present conditions that

(2.11) $\partial \bar{H}_{(\nu)}(\bar{F}_{(\nu)}^{-1}(s), \bar{G}_{(\nu)}^{-1}(t))/\partial t = F_{(i)}(\bar{F}_{(\nu)}^{-1}(s))$ for $(s,t) \in (0,1) \times (\sum_{j=1}^{i-1} \nu_j, \sum_{j=1}^{i} \nu_j)$

from which the second assertion of the lemma follows at once. $\square$

For the standardization of $T_{\lambda N}$ we shall use the quantities

(2.12) $\mu_{\lambda N} = \sum_{i=1}^{k} \lambda_i \nu_i \int J(\bar{F}_{(\nu)})\, dF_{(i)}$,

(2.13) $\sigma^2_{\lambda N} = \mathrm{Var}(\sum_{i=0}^{2} A_{i\lambda N})$,

where the $A_{i\lambda N}$ are defined in (A.27)-(A.28) (see also (A.25)).

THEOREM 2.2. Suppose that C1.1 holds for fixed α with $\alpha < 1/2$. Let for the fixed underlying dfs $H_{(i)}$, satisfying (1.8) so that $H_{(i)} = F_{(i)} \times \mathcal{U}_{(i)}$ for $i = 1, \ldots, k$, C2.1 be satisfied. Moreover, they satisfy C2.2 if $\limsup_{N\to\infty} \Delta_{1N} > 0$. Let us choose $m = k-1$ and assume that $\sum_{i=1}^{j} \nu_i(N) \in \mathcal{O}_{2,\gamma/2}$ for $j = 1, \ldots, k-1$ and $N = 1, 2, \ldots$

Then the quantities $\mu_{\lambda N}$ and $\sigma^2_{\lambda N}$, defined in (2.11) and (2.12) are finite for $N = 1, 2, \ldots$ and

(2.14) $\lim_{N\to\infty} \sup_{x \in (-\infty,\infty)} |P(\{N^{1/2}(T_{\lambda N} - \mu_{\lambda N})/\sigma_{\lambda N} \le x\}) - \Phi(x)| = 0$,

provided $\liminf_{N\to\infty} \sigma^2_{\lambda N} > 0$. Here $\lambda = (\lambda_1, \ldots, \lambda_k)$ is a fixed vector and $T_{\lambda N}$ is defined in (1.15).

The proof of Theorem 2.2 will be reduced to that of Theorem 2.1, using decomposition (A.4) as a common starting point. In the proof of the theorems we shall frequently use Hölder's inequality in the form

$$(2.15) \qquad \iint |\varphi(\bar{F}_{(\nu)})\psi(\bar{G}_{(\nu)})| d\bar{H}_{(\nu)} \leq \left[\int_0^1 |\varphi(s)|^{\xi} ds\right]^{1/\xi} \left[\int_0^1 |\psi(t)|^{\eta}\right]^{1/\eta},$$

where φ and ψ are functions on $(0,1)$ such that the integrals in (2.10) exist and where ξ and η satisfy $\xi > 1$, $\eta > 1$ and $\xi^{-1} + \eta^{-1} = 1$. The finiteness of an absolute moment of order larger than 2 of T_N is guaranteed by the inequality (2.10), provided the orders of magnitude of J and K near 0 and 1, governed by α respectively β, are balanced by the requirement that $\alpha + \beta < 1/2$. This explains that for $T_{\lambda N}$, where $K = \chi_\lambda^*$ and hence C1.2 is satisfied with $\beta = 0$, we have on J the weak condition that $\alpha < 1/2$.

We also see that discontinuities of the scores generating functions are compensated by a smooth behavior of the transformed df $\bar{H}_{(\nu)}(\bar{F}_{(\nu)}^{-1}, \bar{G}_{(\nu)}^{-1})$ on strips along the lines where these discontinuities occur. Under the conditions of Theorem 2.2 (with $H_{(\cdot)}$ as in (1.8)) however, C2.3 is only satisfied in the case where $F_{(1)} = \ldots = F_{(k)}$. For this reason the terms A_{4N} and B_{4N} in (A.4) cannot be properly defined, since χ_λ^* is a simple step function. Moreover, $A_{5N} = 0$ because $K^{(1)} = \chi_\lambda^{*(1)} = 0$ where defined on $(0,1)$. Therefore we use the modification (A.25) of the decomposition (A.4) for the proof of Theorem 2.2.

3. RANK STATISTICS FOR MULTIVARIATE ANALYSIS

More generally, let us suppose that for fixed arbitrary p the sample elements are p-dimensional random vectors originating from k fixed arbitrary p-variate dfs. For each $q=1,\dots,p$ let us poole the q-th coordinate of the k samples and denote the ranks of these pooled coordinates by $R_{q1N},\dots,R_{qNN}$. Statistics of type (1.3) with R_{nN} and Q_{nN} replaced by R_{qnN} and R_{rnN} respectively will be denoted by T_{qrN} for $1 \leq q < r \leq p$. Statistics of type (1.11) with R_{nN} replaced by R_{qnN} and where the summation extends only over those n for which R_{qnN} is the rank of the q-th coordinate of an element in the i-th sample will be denoted by $T_{qN}^{(i)}$ for $i=1,\dots,k$ and $q=1,\dots,p$. For a survey of statistics used in several multivariate problems and the limiting distributions of these statistics we refer to [5].

For testing various kinds of dependences in the underlying p-variate dfs the test statistic is in general derived from the $(p-1)p/2$-dimensional random vector with components

$$T_{qrN} \quad \text{for } 1 \leq q < r \leq p. \tag{3.1}$$

Asymptotic multivariate normality of this vector is equivalent to asymptotic univariate normality of all linear combinations

$$\sum_{1 \leq q < r \leq p} \lambda_{qr} T_{qrN}. \tag{3.2}$$

Using the expansion (A.4) for each of the T_{qrN} separately, we can show asymptotic normality of the statistics (3.2) in a way quite similar to that for $p=2$ given in the appendix.

Test statistics for testing the equality of the underlying p-variate dfs are in general derived from the kp-dimensional random vector with components

$$T_{qN}^{(i)} \quad \text{for } i=1,\dots,k \text{ and } q=1,\dots,p. \tag{3.3}$$

Let us write $T_{\lambda qN} = \sum_{i=1}^{k} \lambda_i T_{qN}^{(i)}$, analogously to (1.12). Showing multivariate asymptotic normality of the vector (3.3) is equivalent to showing that, for all vectors $\lambda = (\lambda_1, \ldots, \lambda_k)$, all linear combinations

$$(3.4) \qquad \sum_{q=1}^{p} \rho_q T_{q\lambda N}$$

have a univariate normal distribution in the limit. With the aid of Theorem 1.1 we may use expansion (A.28) for each $T_{q\lambda N}$ separately. Hence we can show asymptotic normality of the statistics (3.4) in a manner quite similar to that for $p=1$, given in the appendix.

4. RANK STATISTICS FOR SYMMETRY AND SIMPLE LINEAR RANK STATISTICS

We conjecture that Theorem 2.1, appropriately modified, continues to hold in either of the following two cases:

1. The sample sizes $N(1), \ldots, N(k)$ are random variables tending to ∞ in a suitable sense.
2. The bivariate df of (X_n, Y_n) in (1.2), $H_{(n,N)}$ say, is continuous but otherwise arbitrary. In particular these dfs need no longer be members of the set $\{H_{(1)}, \ldots, H_{(k)}\}$ of k given dfs introduced in Section 1.

Moreover, we conjecture that the proof of the corresponding extensions of Theorem 2.2 is again only a minor modification of the proof of the extension of Theorem 2.1.

The first extension is useful to derive asymptotic normality of one-sample rank statistics for testing the hypothesis of symmetry for fixed arbitrary underlying univariate dfs. Results of this kind have already been given in e.g. [5] and [7]. The second extension leads to a general theorem on asymptotic normality of simple linear rank statistics (which include statistics of the kind $T_{\lambda N}$ in (1.12)) for fixed sets of underlying univariate dfs. Results of the latter kind have already been obtained in [3] and [4]. In part A and B of the present section we show how these rank statistics are related to those of type (1.5).

A. RANK STATISTICS FOR SYMMETRY

For completeness let us start with the relation between rank statistics for symmetry and two-sample rank statistics, which we copy from [7]. Let $Z_1, \ldots, Z_N$ be mutually independent identically distributed random variables with continuous univariate df Ψ, and suppose we wish to test whether Ψ is symmetric about 0. Let $X_1^{(1)}, \ldots, X_{N(1)}^{(1)}$ represent the sample consisting of all $|Z|$'s with $Z > 0$ and $X_1^{(2)}, \ldots, X_{N(2)}^{(2)}$ the sample of all $|Z|$'s with $Z < 0$. For $x \geq 0$ each $X^{(1)}$ has df $F_{(1)}(x) = (\Psi(x) - \Psi(0))/(1-\Psi(0))$

and each $X^{(2)}$ has df $F_{(2)}(x) = (\Psi(0) - \Psi(-x))/\Psi(0)$. Even if N is not a random variable, $N(1)$ and $N(2)$ do be random variables satisfying $P(\{N(1)+N(2)=N\})=1$. In view of the extension we have in mind, N is allowed to be a random variable too (as is also the case in [7]). One-sample rank statistics for testing symmetry can be presented as two-sample rank statistics $T_N^{(1)}$ defined in (1.11) (with random $N(1)$). Like in the case of Theorem 1.1 we can again show that $T_N^{(1)}$ satisfies (1.15), by choosing $H_{(i)}$ as in (1.8) with $F_{(i)}$ as given above ($i=1,2$).

8. SIMPLE LINEAR RANK STATISTICS

Let $X_{1N}, \ldots, X_{NN}$ have continuous univariate dfs $F_{(1,N)}, \ldots, F_{(N,N)}$ respectively. Denote the ranks of the X's by $R_{1N}, \ldots, R_{NN}$, as usual. Simple linear rank statistics are of the form

$$(4.1) \qquad S_N = N^{-1} \sum_{n=1}^{N} c_N(n)\, a_N(R_{nN}).$$

Let us (in contradistinction to [3] and [4]) assume that the regression constants $c_N(n)$ are generated by a function K on $(0,1)$ according to

$$(4.2) \qquad c_N(n) = K(n/(N+1)),$$

for $N=1,2,\ldots$ and $n=1,\ldots,N$. Like in the case of Theorem 1.1 we have

$$(4.3) \qquad P(\{S_N = \iint J(F_N^*)\, K(G_N^*)\, dH_N\}) = 1,$$

provided we choose, analogously to (1.8), $H_{(n,N)}(x,y) = F_{(n,N)}(x)\, U_{(n)}(y)$ for $x, y \in (-\infty, \infty)$.

In order to prove asymptotic normality of S_N it is likely that we shall again have to balance the rates of growth near 0 and 1 of J (generating the scores) and K (generating the regression constants).

APPENDIX: PROOF OF THEOREMS 2.1 AND 2.2

Without loss of generality we shall the scores generating functions J and K suppose to be right-continuous. Conditions C1.1 and C1.2 imply that J and K can be decomposed as

(A.1) $$J = J_c + J_d \quad , \quad K = K_c + K_d ,$$

where J_c, K_c are continuous throughout $(0,1)$ and J_d, K_d are simple right-continuous step functions having discontinuities of the first kind only at the elements of $\mathcal{D}_{1N}$, $\mathcal{D}_{2N}$ respectively (see (2.4)). The dependence on N of the functions J_c, J_d, K_c and K_d is again suppressed in the notation.

It will occasionally be convenient to use the function

(A.2) $$\begin{aligned} e(x) &= 0 \qquad \text{for } x \in (-\infty, 0) , \\ &= 1 \qquad \text{for } x \in [0, \infty) . \end{aligned}$$

With the aid of this function we can explicitly write

(A.3) $$\begin{aligned} J_d(s) &= \Sigma_{i=1}^{\ell} \; \Delta_{1iN} \; e(s - s_{iN}) \qquad \text{for } s \in (0,1), \\ K_d(t) &= \Sigma_{j=1}^{m} \; \Delta_{2jN} \; e(t - t_{jN}) \qquad \text{for } t \in (0,1), \end{aligned}$$

see (2.5).

Starting with the proof of Theorem 2.1 we shall first give the decomposition

(A.4) $$N^{1/2}(T_N - \mu_N) = \Sigma_{i=0}^{4} \; A_{iN} + \Sigma_{i=1}^{6} \; B_{iN} ,$$

which is of essential importance in the proof of both theorems. Here T_N is defined in (1.5), μ_N in (2.6), and

$$A_{0N} = N^{1/2} \iint J(\bar{F}_{(\nu)}) K(\bar{G}_{(\nu)}) \, d(H_N - \bar{H}_{(\nu)}) ,$$

$$A_{1N} = N^{1/2} \iint (F_N - \bar{F}_{(\nu)}) J^{(1)}(\bar{F}_{(\nu)}) K(\bar{G}_{(\nu)})\, d\bar{H}_{(\nu)},$$

$$A_{2N} = N^{1/2} \iint (G_N - \bar{G}_{(\nu)}) J(\bar{F}_{(\nu)}) K^{(1)}(\bar{G}_{(\nu)})\, d\bar{H}_{(\nu)},$$

$$A_{3N} = N^{1/2} \sum_{i=1}^{\ell} [F_N(\bar{F}_{(\nu)}^{-1}(s_{iN})) - s_{iN}][\Delta_{1iN} \int_0^1 K(t)\, \bar{h}_{(\nu)}(s_{iN}, t)\, dt],$$

$$A_{4N} = N^{1/2} \sum_{j=1}^{m} [G_N(\bar{G}_{(\nu)}^{-1}(t_{jN})) - t_{jN}][\Delta_{2jN} \int_0^1 J(s)\, \bar{h}_{(\nu)}(s, t_{jN})\, ds],$$

$$B_{1N} = N^{1/2} \iint [J_c(F_N^*) - J_c(\bar{F}_{(\nu)})] K(\bar{G}_{(\nu)})\, dH_N - A_{1N},$$

$$B_{2N} = N^{1/2} \iint J(\bar{F}_{(\nu)}) [K_c(G_N^*) - K_c(\bar{G}_{(\nu)})]\, dH_N - A_{2N},$$

$$B_{3N} = N^{1/2} \iint [J_d(F_N^*) - J_d(\bar{F}_{(\nu)})] K(\bar{G}_{(\nu)})\, dH_N - A_{3N},$$

$$B_{4N} = N^{1/2} \iint J(\bar{F}_{(\nu)}) [K_d(G_N^*) - K_d(\bar{G}_{(\nu)})]\, dH_N - A_{4N},$$

$$B_{5N} = N^{1/2} \iint [J_c(F_N^*) - J_c(\bar{F}_{(\nu)})][K(G_N^*) - K(\bar{G}_{(\nu)})]\, dH_N,$$

$$B_{6N} = N^{1/2} \iint [J_d(F_N^*) - J_d(\bar{F}_{(\nu)})][K(G_N^*) - K(\bar{G}_{(\nu)})]\, dH_N.$$

The functions $\bar{F}_{(\nu)}^{-1}$ and $\bar{G}_{(\nu)}^{-1}$ are defined according to (2.1).

Let us start with the asymptotic normality of the A-terms, deferring the asymptotic negligibility of the B-terms to the second part of the section. The limiting distribution of the A-terms may be derived in much the same way as in [8] and [10]. Again we appeal to some version of the central limit theorem by first writing

$$\text{(A.5)} \qquad \sum_{i=0}^{4} A_{iN} = N^{-1/2} \sum_{n=1}^{N} Z_{nN},$$

where $Z_{nN} = \sum_{i=0}^{2} A_{inN} + \sum_{i=1}^{2} A'_{inN}$ and

$$A_{0nN} = J(\bar{F}_{(\nu)}(X_n))\, K(\bar{G}_{(\nu)}(Y_n)) - \mu_N,$$

$$A_{1nN} = \iint [c(\bar{F}_{(\nu)} - \bar{F}_{(\nu)}(X_n)) - \bar{F}_{(\nu)}] J^{(1)}(\bar{F}_{(\nu)}) K(\bar{G}_{(\nu)}) d\bar{H}_{(\nu)},$$

$$A_{2nN} = \iint [c(\bar{G}_{(\nu)} - \bar{G}_{(\nu)}(Y_n)) - \bar{G}_{(\nu)}] J(\bar{F}_{(\nu)}) K^{(1)}(\bar{G}_{(\nu)}) d\bar{H}_{(\nu)},$$

$$A_{3nN} = \sum_{i=1}^{\ell} [c(s_{iN} - \bar{F}_{(\nu)}(X_n)) - s_{iN}][\Delta_{1iN} \int_0^1 K(t) \bar{h}_{(\nu)}(s_{iN}, t) dt]$$

$$A_{4nN} = \sum_{j=1}^{m} [c(t_{jN} - \bar{G}_{(\nu)}(Y_n)) - t_{jN}][\Delta_{2jN} \int_0^1 J(s) \bar{h}_{(\nu)}(s, t_{jN}) ds]$$

The rvs Z_{nN} depend on $(X_1, Y_1), \ldots, (X_N, Y_N)$ through (X_n, Y_n) only and, consequently, are mutually independent. Let us for the moment assume that there exists a $\delta > 0$ such that

(A.6) $$\sup_{N=1,2,\ldots} N^{-1} \sum_{n=1}^{N} \mathcal{E}(|Z_{nN}|^{2+\delta}) < \infty,$$

or equivalently that

(A.7) $$\sup_{N=1,2,\ldots} N^{-1} \sum_{n=1}^{N} \mathcal{E}(|A_{inN}|^{2+\delta}) < \infty, \text{ for } i = 0, 1, \ldots, 4.$$

It is easy to verify that

$$\sum_{n=1}^{N} \mathcal{E}(Z_{nN}) = 0 \quad , \quad \sigma_N^2 = N^{-1} \sum_{n=1}^{N} \operatorname{Var}(Z_{nN}),$$

where σ_N^2 is defined in (2.7). Because one of the conditions of the theorem is that $\liminf_{N\to\infty} \sigma_N^2 > 0$, we may apply the central limit theorem in a special form which is essentially due to Esseen (see also e.g. [11, formula (3.33)]. This entails

$$\lim_{N\to\infty} \sup_{-\infty < z < \infty} |P(\{[N\sigma_N^2]^{-1/2} \sum_{n=1}^{N} Z_{nN} \leq z\}) - \Phi(z)| = 0.$$

Let us next turn to the proof of (A.7), and note that by C1.1, C1.2 and the definition of $\bar{H}_{(\nu)}$ we have the bound

$$(A.8) \qquad N^{-1}\sum_{n=1}^{N} \mathcal{E}\left(|A_{0nN}|^{2+\delta}\right) \leq M \iint [R(\bar{F}_{(\nu)})]^{\alpha(2+\delta)}[R(\bar{G}_{(\nu)})]^{\beta(2+\delta)}\, d\bar{H}_{(\nu)}$$

for all $N=1,2,\ldots$ Because for any $\delta>0$ the function $c(z)$, defined in (A.2), satisfies

$$|c(u-v)-u| \leq M\, R^{1/2-\delta}(v)\, R^{-1/2+\delta}(u) \quad \text{for } u,v\in(0,1),$$

we deduce in the same way from C1.1 and C1.2 that

$$(A.9) \qquad N^{-1}\sum_{n=1}^{N} \mathcal{E}\left(|A_{1nN}|^{2+\delta}\right) \leq M \int_0^1 [R(s)]^{(1/2-\delta)(2+\delta)}\, ds \;\times$$

$$\iint [R(\bar{F}_{(\nu)})]^{\alpha+1/2+\delta}[R(\bar{G}_{(\nu)})]^{\beta}\, d\bar{H}_{(\nu)},$$

for all $N=1,2,\ldots$ A similar bound can be found for $N^{-1}\sum_{n=1}^{N}\mathcal{E}\left(|A_{2nN}|^{\delta}\right)$. Let us note that the right-hand sides in (A.8) and (A.9) do not depend on N.

We shall use throughout that $\alpha,\beta>0$ and $\alpha+\beta<1/2$. Application of (2.10) with $\xi=(\alpha+\beta)/\alpha$ and $\eta=(\alpha+\beta)/\beta$ yields that the integral on the right of (A.8) is bounded by

$$\int_0^1 [R(u)]^{(\alpha+\beta)(2+\delta)}\, du < \infty,$$

provided $\delta(>0)$ is chosen sufficiently small to ensure that $(\alpha+\beta)(2+\delta)<1$. The first integral on the right of (A.9) is finite since $(1/2-\delta)(2+\delta)<1$ for any $\delta>0$. To the second integral on the right of (A.9) we apply (2.10) with $\xi=(\alpha+1/2+2\delta)^{-1}$ and $\eta=(1/2-\alpha-2\delta)^{-1}$ (it is assumed that $\delta\ (>0)$ satisfies $2\delta<1/2-\alpha$ so that $\xi>1$, $\eta>1$) and obtain the bound

$$\left\{\int_0^1 [R(s)]^{(\alpha+1/2+\delta)/(\alpha+1/2+2\delta)} ds\right\}^{\alpha+1/2+2\delta} \times$$

$$\left\{\int_0^1 [R(t)]^{\beta/(1/2-\alpha-2\delta)} dt\right\}^{1/2-\alpha-2\delta} < \infty,$$

provided $0 < 2\delta < 1/2 - \alpha - \beta$. From these observations we may conclude that there exists a number $\delta > 0$ for which (A.7) is satisfied for $i = 0, 1, 2$.

Using once more C1.1 and C1.2 we even find that

$$(A.10) \qquad |A_{3nN}| \leq M \int_0^1 R^{\beta+b}(t)\, dt,$$

for all $N = 1, 2, \ldots$ and $n = 1, \ldots, N$. The integral on the right in (A.10) is a finite constant because $\beta + b < 1$. A similar observation holds concerning the $|A_{4nN}|$. Hence for any $\delta > 0$, (A.7) holds for $i = 3, 4$. This completes the proof of (A.7) and consequently the proof of the asymptotic normality of the A-terms.

As far as the asymptotic negligibility of the B-terms is concerned, let us note that the present B-terms can be obtained from the B- and C-terms in [8 , formula (3.2)] , by consistently replacing F, G, H by $\bar{F}_{(\nu)}, \bar{G}_{(\nu)}, \bar{H}_{(\nu)}$ and $\alpha_i, s_i, \beta_j, t_j$ by $\Delta_{1iN}, s_{iN}, \Delta_{2jN}, t_{jN}$ ($i = 1, \ldots, l$; $j = 1, \ldots, m$; $N = 1, 2, \ldots$). This correspondence is not merely a notational one. Inspecting the proofs for the asymptotic negligibility of the remainder terms in [8 , Section 5] we see that these proofs would carry over to the present B-terms without essential changes, provided only the basic lemmas in [8 , Section 4] remain valid under the modifications mentioned above. In order to avoid needless repetition of arguments we shall content ourselves with a review of the lemmas in the form required for application in the present case.

By symmetry we need only consider B_{1N}, B_{3N}, B_{5N} and B_{6N}. In the sequel we shall use, without explicit reference, that

$$0 < \liminf_{N\to\infty} N(i)/N \leq \limsup_{N\to\infty} N(i)/N < 1 \text{ for } i = 1, \ldots, k,$$

and that

$$X_{1:N} = \min_{i=1,\dots,k}\{X^{(i)}_{1:N(i)}\}, \quad X_{N:N} = \max_{i=1,\dots,k}\{X^{(i)}_{N(i):N(i)}\}.$$

We shall start with the modified versions of the lemmas specific for the terms B_{1N} and B_{5N} which are based on the continuous component J_c of J. The idea of the proofs is quite similar to that of [10, Lemmas 6.1 and 6.2].

<u>LEMMA A.1.</u> For each $0<\delta\leq 1/2$ we have, as $N\to\infty$, uniformly for $H_{(1)},\dots,H_{(k)}\in\mathcal{H}$

(A.11) $$\sup_{[X_{1:N},X_{N:N}]} |U^*_N(\bar{F}_{(\nu)}) - U_N(\bar{F}_{(\nu)})|[R(\bar{F}_{(\nu)})]^{1/2-\delta} = o_p(1),$$

(A.12) $$\sup_{[X_{1:N},X_{N:N}]} |U^*_N(\bar{F}_{(\nu)})|[R(\bar{F}_{(\nu)})]^{1/2-\delta} = O_p(1).$$

<u>PROOF:</u> For arbitrary $\varepsilon>0$ there exists a $\beta=\beta_\varepsilon$ with $0<\beta<1$ such that

(A.13) $$P(\cap_{i=1}^k\{\beta/N \leq F_{(i)}(X^{(i)}_{1:N(i)}) \leq F_{(i)}(X^{(i)}_{N(i):N(i)}) \leq 1-\beta/N\}) \geq 1-\varepsilon,$$

uniformly in all continuous $F_{(1)},\dots,F_{(k)}$ and for all $N=1,2,\dots$ Let us note that

$$[N(i)/N]s \leq \bar{F}_{(\nu)}(F^{-1}_{(i)}(s)) \leq 1-[N(i)/N](1-s) \quad \text{for } s\in[0,1].$$

Hence, application of the non-decreasing function $\bar{F}_{(\nu)}(F^{-1}_{(i)})$ to the quantities within the brackets in (A.13) yields that for some $\tilde{\beta}=\tilde{\beta}_\varepsilon$ with $0<\tilde{\beta}<1$ we have

(A.14) $$P(\{\tilde{\beta}/N \leq \bar{F}_{(\nu)}(X_{1:N}) \leq \bar{F}_{(\nu)}(X_{N:N}) \leq 1-\tilde{\beta}/N\}) \geq 1-\varepsilon,$$

uniformly in all continuous $F_{(1)},\dots,F_{(k)}$ and for all $N=1,2,\dots$, which proves (A.11).

For (A.12) we need [11, formula (2.4)], which is an extension of [6, Lemma 2.2] together with (A.11). □

<u>LEMMA A.2.</u> For each $\xi>0$ we have, as $N\to\infty$, uniformly for $H_{(1)},\dots,H_{(k)}\in\mathcal{H}$,

(A.15) $$\sup_{[X_{1:N},X_{N:N}]} [R(F^*_N)/R(\bar{F}_{(\nu)})]^{\xi} = O_p(1).$$

PROOF: For an arbitrary $\varepsilon>0$ there exists a $\tilde\beta=\tilde\beta_\varepsilon$ with $0<\tilde\beta<1$ such that

$$P(\cap_{i=1}^{k}\{F_{N(i)}^{(i)*}(X_{N(i):N(i)}^{(i)})=N/(N+1)\le 1-\tilde\beta(1-F_{(i)}(X_{N(i):N(i)}^{(i)}\})\ge 1-\varepsilon,$$

uniformly in all continuous $F_{(1)},\ldots,F_{(k)}$. Jointly with [13, Lemma A.3] it follows that there exists a $\beta=\beta_\varepsilon$ with $0<\beta<1$ such that

$$P(\cap_{i=1}^{k}\{\beta F_{(i)}\le F_{N(i)}^{(i)*}\le 1-\beta(1-F_{(i)})\text{ on }[X_{1:N(i)}^{(i)},X_{N(i):N(i)}^{(i)}]\})\ge 1-\varepsilon,$$

uniformly in all $F_{(1)},\ldots,F_{(k)}$ and for all $N=1,2,\ldots$ Let us define

$$\begin{aligned}\Phi_{N(i)}^{(i)*}&=F_{N(i)}^{(i)*}&&\text{on }[X_{1:N(i)}^{(i)},X_{N(i):N(i)}^{(i)}],\\&=N/[(N+1)N(i)]&&\text{on }(-\infty,X_{1:N(i)}^{(i)})\quad\text{for }i=1,\ldots,k,\end{aligned}$$

so that $P(\cap_{i=1}^{k}\{\beta F_{(i)}\le\Phi_{N(i)}^{(i)*}\le 1-\beta(1-F_{(i)})\text{ on }(-\infty,\infty)\})\ge 1-\varepsilon$, uniformly in all $F_{(1)},\ldots,F_{(k)}$ and for all $N=1,2,\ldots$ Moreover, let us introduce

$$\Phi_N^{*}=\Sigma_{i=1}^{k}\,\nu_i\,\Phi_{N(i)}^{(i)}\qquad\text{for }i=1,\ldots,k,$$

which consequently has the property that

$$\text{(A.16)}\qquad P(\{\beta\bar F_{(\nu)}\le\Phi_N^{*}\le 1-\beta(1-\bar F_{(\nu)})\text{ on }(-\infty,\infty)\})\ge 1-\varepsilon,$$

uniformly in all $F_{(1)},\ldots,F_{(k)}$ and for all $N=1,2,\ldots$

The definition of Φ_N^{*} entails that $\Phi_N^{*}\le F_N^{*}+[(1/\nu_1)+\cdots+(1/\nu_k)]/(N+1)\le F_N^{*}+k$ on $(-\infty,\infty)$, which in turn implies the inequality

$$\text{(A.17)}\qquad F_N^{*}\ge\Phi_N^{*}/(1+k)\qquad\text{on }[X_{1:N},X_{N:N}]$$

We also have

$$\text{(A.18)}\qquad F_N^{*}\le\Phi_N^{*}\qquad\text{on }(-\infty,\infty).$$

Combining (A.16) - (A.18) it follows that

$$(A.19)\quad P(\{[\beta/(1+k)]\overline{F}_{(v)} \le F_N^* \le 1-[\beta/(1+k)](1-\overline{F}_{(v)}) \text{ on } [X_{1:N}, X_{N:N}]\}) \ge 1-\varepsilon$$

uniformly in all continuous $F_{(1)}, \ldots, F_{(k)}$ and for all $N = 1, 2, \ldots$

With the aid of (3.14) the proof of the lemma may be concluded by utilizing the reproducing u-shaped character of the function R^{ξ} (see [13, Definition A3 ; C, proof of Lemmas 4.2 and 6.1). $\square$

The next two lemmas, specific for the terms B_{3N} and B_{6N} based on the discontinuous component J_d of J, are modified versions of [8, Lemmas 4.3 and 4.4] The content of the present Lemma A.3 is only a minor extension of a lemma, due to Van Zwet (see [8, Lemma 4.4). To write the content of this lemma in a convenient form let us denote $\iint_B dH_{(i)} = H_{(i)}\{B\}$, $\iint_B d\overline{H}_{(v)} = \overline{H}_{(v)}\{B\}$ and $\iint_B dH_N = H_N\{B\}$ for any Borel set B in the plane. By an interval in the plane the product set of two intervals on the line will be understood. The proof of the present Lemma A.4 is based on an idea by Shirahata (see [12, Remark 2 in Section 3]) Given any set A, A^c will denote its complement, $\chi(A)$ its indicator function and $\chi(A;a)$ the value of this function at the point a.

LEMMA A.3 (Van Zwet): Let $I_1, I_2, \ldots$ be a sequence of intervals in the plane and let $\mathcal{J}_N = \{I_N^* : I_N^* \text{ is an interval contained in } I_N\}$, $N = 1, 2, \ldots$ Then

$$(A.20)\quad \sup_{I_N^* \in \mathcal{J}_N} |H_N\{I_N^*\} - \overline{H}_{(v)}\{I_N^*\}| = O_p([\overline{H}_{(v)}\{I_N\}/N]^{1/2}),$$

as $N \to \infty$, uniformly in all sequences of intervals $I_1, I_2, \ldots$ and all bivariate distribution functions $H_{(1)}, \ldots, H_{(k)}$ (continuous or not).

PROOF: The validity of the lemma for $k=1$ entails, for arbitrary $\varepsilon > 0$, the existence of a constant $M = M_\varepsilon$ such that the sets

$$\Omega_{iN} = \{\omega : \sup_{I_N^* \in \mathcal{J}_N} |H_{N(i)}^{(i)}\{I_N^*\} - H_{(i)}\{I_N^*\}| \le M[H_{(i)}\{I_N\}/N]^{1/2}\}$$

satisfy $P(\bigcap_{i=1}^k \Omega_{iN}) \geq 1-\varepsilon$, for all N and all $H_{(1)}, \dots, H_{(k)}$ (continuous or not).
Application of Jensen's inequality shows that

$$M \Sigma_{i=1}^k \nu_i \left[H_{(i)}\{I_N\}/N(i)\right]^{1/2} = M \Sigma_{i=1}^k \left[\nu_i H_{(i)}\{I_N\}\right]^{1/2} N^{-1/2}$$

$$\leq k^{1/2} M \left[\bar{H}_{(\nu)}\{I_N\}/N\right]^{1/2},$$

so that by the triangle inequality

$$P(\{\sup_{I_N^* \in \mathcal{I}_N} |H_N\{I_N^*\} - \bar{H}_{(\nu)}\{I_N^*\}| \leq k^{1/2} M [\bar{H}_{(\nu)}\{I_N\}/N]^{1/2}\})$$

$$\geq P(\bigcap_{i=1}^k \Omega_{iN}) \geq 1-\varepsilon,$$

for all $N = 1, 2, \dots$ and all $H_{(1)}, \dots, H_{(k)}$ (continuous or not). □

LEMMA A.4 (Shirahata): Let φ, ψ be a pair of finite positive measurable functions on $(0,1)$ with $\int_0^1 \varphi(t)dt < \infty$ and $\int_0^1 \varphi(t)\psi(t)dt < \infty$. Let $\mathcal{T} \subset (-\infty, \infty)$ be some index set and θ_0 an accumulation point (possibly $-\infty$ or $+\infty$) of $\mathcal{T}$. For each $\theta \in \mathcal{T}$ let φ_θ be a measurable function on $(0,1)$ such that $|\varphi_\theta| \leq \varphi$ on $(0,1)$. Suppose that $\varphi_\theta \to 0$ as $\theta \to \theta_0$, almost everywhere on $(0,1)$ with respect to Lebesgue measure.
Then we have for $i = 1, \dots, k$, as $\theta \to \theta_0$ and $N \to \infty$,

$$(A.21) \quad N^{1/2} \int\int |[c(F_N^* - s_{iN}) - c(\bar{F}_{(\nu)} - s_{iN})] \varphi_\theta(\bar{G}_{(\nu)})| d\bar{H}_{(\nu)} = o_p(1),$$

$$(A.22) \quad N^{1/2} \int\int |[c(F_N^* - s_{iN}) - c(\bar{F}_{(\nu)} - s_{iN})] \varphi_\theta(\bar{G}_{(\nu)})| dH_N = o_p(1),$$

uniformly for $H_{(1)}, \dots, H_{(k)} \in \mathcal{H}$ satisfying that for all $(s,t) \in O_1 \times (0,1)$ the density $h_{(\nu)}$ exists as a continuous function bounded by $\psi(t)$.

PROOF: Let us first note that $|F_N^* - \bar{F}_{(\nu)}| \leq \Sigma_{i=1}^k [N(i)/(N+1)] \times |F_{N(i)}^{(i)} - [(N+1)/N] F_{(i)}|$ on $(-\infty, \infty)$, so that $\sup_{(-\infty,\infty)} |F_N^* - \bar{F}_{(\nu)}| = O_p(N^{-1/2})$ uniformly for all continuous $F_{(1)}, \dots, F_{(k)}$. Hence, given an arbitrary $\varepsilon > 0$ there exists a constant $M = M_\varepsilon$ such that the set

$$\Omega_N = \Omega_{N\varepsilon} = \{\omega : \sup_{(-\infty,\infty)} |F_N^* - F_{(\nu)}| \leq M N^{-1/2}\}$$

has probability $P(\Omega_N) \geq 1-\varepsilon$, for all continuous $F_{(1)},\dots, F_{(k)}$ and all $N=1,2,\dots$ Let us prove the lemma for an arbitrary but fixed index i and put for brevity

$$S_N = S_{NM} = [s_{iN} - MN^{-1/2}, s_{iN} + MN^{-1/2}].$$

For arbitrary fixed $\zeta > 0$ we have

$$P(\{N^{1/2}\iint |[c(F_N^* - s_{iN}) - c(\bar{F}_{(\nu)} - s_{iN})]\varphi_\theta(\bar{G}_{(\nu)})| d\bar{H}_{(\nu)} \leq \zeta\})$$

$$\geq P(\{\chi(\Omega_N) N^{1/2}\iint \chi(S_N; \bar{F}_{(\nu)})|\varphi_\theta(\bar{G}_{(\nu)})| d\bar{H}_{(\nu)} \leq \zeta\}) - P(\Omega_N^c)$$

$$\geq P(\{N^{1/2}[2MN^{-1/2}]\int_0^1 |\varphi_\theta(t)|\psi(t)dt \leq \zeta\}) - \varepsilon \geq 1-2\varepsilon,$$

provided N is chosen sufficiently large (this choice is independent of the k-tuple $H_{(1)},\dots, H_{(k)}$ as long as it satisfies the conditions of the lemma). The last transition follows from the dominated convergence theorem. The numbers $\zeta > 0$ and $\varepsilon > 0$ being arbitrary, this shows (A.21).

By a similar reasoning we see that for arbitrary but fixed $\zeta > 0$,

$$P(\{N^{1/2}\iint |[c(F_N^* - s_{iN}) - c(\bar{F}_{(\nu)} - s_{iN})]\varphi_\theta(\bar{G}_{(\nu)})| dH_N \leq \zeta\})$$

$$\geq P(\{\chi(\Omega_N) N^{-1/2}\sum_{n=1}^N \chi(S_N; \bar{F}_{(\nu)}(X_n))|\varphi_\theta(\bar{G}_{(\nu)}(Y_n))| \leq \zeta\}) - P(\Omega_N^c)$$

$$\geq 1 - \zeta^{-1} N^{1/2}[2MN^{-1/2}]\int_0^1 |\varphi_\theta(t)|\psi(t)dt - \varepsilon \geq 1-2\varepsilon,$$

provided N is chosen sufficiently large (this choice is again independent of the k-tuple $H_{(1)},\dots, H_{(k)}$ as long as it satisfies the conditions of the lemma). The last but one transition follows from Markov's inequality and the last transition from the dominated convergence theorem. The numbers $\zeta > 0$ and $\varepsilon > 0$ being arbitrary this shows (A.22). $\square$

Let us finally turn to the proof of Theorem 2.2, trying to reduce it as much as possible to that of Theorem 2.1. In this case $H_{(i)} = F_{(i)} \times U_{(i)}$ (see (1.8)) $K = \chi_\lambda^*$ (see (1.11)), $m = k-1$ and $t_{jN} = \nu_1^* + \dots + \nu_j^*$ for $j = 1, \dots, k-1$.

For direct application of Theorem 2.1, however, we would also need C2.3 because $\Delta_{2N} = |\lambda_2 - \lambda_1| + \dots + |\lambda_k - \lambda_{k-1}| > 0$ (unless $\lambda_1 = \dots = \lambda_k$). It is apparent from (2.11) that C2.3 is trivially satisfied in the case where $F_{(1)} = \dots = F_{(k)}$ but also that it will not be fulfilled in general. Therefore we cannot decompose $T_{\lambda N}$ as in (A.4). In particular A_{4N} and B_{4N}, arising from a decomposition of

$$\text{(A.23)} \qquad N^{1/2} \iint J(\bar{F}_{(\nu)}) [\chi_\lambda^*(G_N^*) - \chi_\lambda(\bar{G}_{(\nu)})] \, dH_N,$$

cannot be defined. In the present simpler case, however, no such decomposition is desired and it will be even convenient to add B_{5N} and B_{6N} to (A.23). Instead of A_{4N}, B_{4N}, B_{5N} and B_{6N} we thus arrive at a single new term which can simply be shown to be of second order.

An even simpler decomposition of $T_{\lambda N}$ is obtained in the following way. Let us introduce the functions

$$\chi_i(t) = 1 \qquad \text{for } t \in (\Sigma_{j=1}^{i-1} \nu_j, \Sigma_{j=1}^{i} \nu_j],$$
$$= 0 \qquad \text{elsewhere},$$

(A.24) $$\chi_\lambda(t) = \Sigma_{i=1}^{k} \lambda_i \chi_i(t) \qquad \text{for } t \in (0,1).$$

Because $\chi_\lambda^*(G_N^*) = \chi_\lambda(G_N)$ it follows that

$$T_{\lambda N} = \iint J(F_N^*) \chi_\lambda^*(G_N^*) dH_N = \iint J(F_N^*) \chi_\lambda(G_N) dH_N.$$

The decomposition of $T_{\lambda N}$ on which we shall finally base our proof of Theorem 2.2 is the one formally obtained from (A.4) by consistently replacing K by χ_λ and G_N^* by G_N and by combining A_{4N}, B_{4N}, B_{5N} and B_{6N} in one single new term, $C_{\lambda N}$ say. Because $\chi_\lambda^{(1)} = 0$ on $(0,1) - \mathcal{D}_{\lambda N}$ the terms A_{3N} and B_{3N} vanish in the decomposition, which can be written

(A.25) $$N^{1/2}(T_{\lambda N} - \mu_{\lambda N}) = \Sigma_{i=0}^{2} A_{i\lambda N} + \Sigma_{i=1}^{2} B_{i\lambda N} + C_{\lambda N}.$$

Here $\mu_{\lambda N}$ is defined in (2.12), $\sigma_{\lambda N}^2$ in (2.13) and

$$A_{0\lambda N} = N^{1/2} \iint J(\bar{F}_{(\nu)}) \chi_\lambda(\bar{G}_{(\nu)}) d(H_N - \bar{H}_{(\nu)}),$$

$$A_{1\lambda N} = N^{1/2} \iint (F_N - \bar{F}_{(\nu)}) J'(\bar{F}_{(\nu)}) \chi_\lambda(\bar{G}_{(\nu)}) d\bar{H}_{(\nu)},$$

$$A_{2\lambda N} = N^{1/2} \Sigma_{i=1}^{\ell} [F_N(\bar{F}_{(\nu)}^{-1}(s_{iN})) - s_{iN}][\alpha_{iN} \int_0^1 \chi_\lambda(t) \bar{h}_{(\nu)}(s_{iN}, t) dt]$$

$$B_{1\lambda N} = N^{1/2} \iint [J_c(F_N^*) - J_c(\bar{F}_{(\nu)})] \chi_\lambda(\bar{G}_{(\nu)}) dH_N - A_{1\lambda N},$$

$$B_{2\lambda N} = N^{1/2} \iint [J_d(F_N^*) - J_d(\bar{F}_{(\nu)})] \chi_\lambda(\bar{G}_{(\nu)}) dH_N - A_{2\lambda N},$$

$$C_{\lambda N} = N^{1/2} \iint [J(\bar{F}_{(\nu)}) + J(F_N^*)][\chi_\lambda(G_N) - \chi_\lambda(\bar{G}_{(\nu)})]\, dH_N .$$

The A_λ- terms in (A.28) are special cases of some of the A- terms occurring in (A.4) and may be dealt with in the same way. The B_λ- terms in (A.25) are special cases of slightly modified versions of some of the B- terms in (A.4) and can be treated in essentially the same way. Only $C_{\lambda N}$ is new. Using property (2.9) of $\bar{G}_{(\nu)}$, the similar property

$$\text{(A.26)} \qquad G_N(y) = \sum_{j=1}^{i-1} \nu_j + \nu_i \qquad \text{for } y \in (i-1, i),$$

of G_N and the definition of χ_λ in (A.24) it is immediate that

$$C_{\lambda N} = 0 .$$

By means of (2.9) and (A.26) the A_λ - terms can be written as

$$\text{(A.27)} \qquad A_{0\lambda N} = N^{1/2} \sum_{i=1}^{k} \nu_i \lambda_i \int J(\bar{F}_{(\nu)})\, d(F_{N(i)}^{(i)} - F_{(i)}),$$

$$\text{(A.28)} \qquad A_{1\lambda N} = N^{1/2} \sum_{i=1}^{k} \nu_i \lambda_i \int (F_N - \bar{F}_{(\nu)})\, J^{(1)}(\bar{F}_{(\nu)})\, dF_{(i)},$$

$$\text{(A.29)} \qquad A_{2\lambda N} = N^{1/2} \sum_{i=1}^{\ell} [F_N(\bar{F}_{(\nu)}^{-1}(s_{iN})) - s_{iN}][\alpha \sum_{j=1}^{k} \lambda_j \nu_j f_{(j,\nu)}(s_{iN})]$$

From these expressions it is clear that the A_λ - terms do in fact not depend on the auxiliary random components related to the Y's. For $k=2$ the terms $A_{0\lambda N}$ and $A_{1\lambda N}$ correspond to the random variables given in [2, (4.5) and (4.6)]. (In [2] the random variable $A_{2\lambda N}$ is not present since there $\alpha_{1N} = \ldots = \alpha_{\ell N} = 0$ for all N.) The expressions for $\mu_{\lambda N}$ and $\sigma^2_{\lambda N}$ can best be derived from the latter formulas. In the same way the B_λ - terms can be seen to have a simpler form which no longer contains the auxiliary random components related to the Y's.

REFERENCES

[1] BHUCHONGKUL, S. (1964). A class of nonparametric tests for independence in bivariate populations. Ann. Math. Statist. 35, 138-149.

[2] CHERNOFF, H. and SAVAGE, I.R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. Ann. Math. Statist. 29, 972-994.

[3] DUPAČ, V. and HÁJEK, J. (1969). Asymptotic normality of simple linear rank statistics under alternatives II. Ann. Math. Statist. 40, 1992-2017.

[4] HÁJEK, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. Ann. Math. Statist. 39, 325-346.

[5] PURI, M.L. and SEN, P.K. (1971). Nonparametric Methods in Multivariate Analysis. Wiley, New York.

[6] 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.

[7] PYKE, R. and SHORACK, G.R. (1968). Weak convergence and a Chernoff-Savage theorem for random sample sizes. Ann. Math. Statist. 39, 1675-1685.

[8] RUYMGAART, F.H. (1972). Asymptotic normality of nonparametric tests for independence. To appear in Ann. Statist..

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

[10] 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.

[11] SEN, P.K. (1970). On the distribution of the one-sample rank order statistic. Nonparametric Techniques in Statistical Inference, Puri ed., 53-73. University Press, Cambridge.

[12] SHIRAHATA, S. (1974). Locally most powerful rank tests for independence with censored data. To appear in Ann. Statist..

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