Statistica Neerlandica (2005) Vol. 59, nr. 3, pp. 313–323

Generalized two-sample *U*-statistics for clustered data

Mei-Ling Ting Lee*

Channing Laboratory, Brigham and Women's Hospital, Boston, MA,
Department of Medicine, Harvard Medical School, Boston, MA and
Biostatistics Department, Harvard School of Public Health, Boston, MA

Herold G. Dehling†

Fakultät für Mathematik, Ruhr-Universität Bochum,
44780 Bochum, Germany

In this paper we investigate two-sample *U*-statistics in the case of clusters of repeated measurements observed on individuals from independent populations. The observations on the *i*-th individual in the first population are denoted by $X_1^{(i)}, \dots, X_{J_i}^{(i)}$, $1 \le i \le m$, and those on the *k*-th individual in the second population are denoted by $Y_1^{(k)}, \dots, Y_{L_k}^{(k)}$, $1 \le k \le n$. Given the kernel $\phi(x, y)$, we define the generalized two-sample *U*-statistic by

$$U_{m,n} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{k=1}^{n} \sum_{j=1}^{J_i} \sum_{l=1}^{L_k} \phi(X_j^{(i)}, Y_l^{(k)}).$$

We derive the asymptotic distribution of $U_{m,n}$ for large sample sizes. As an application we study the generalized Mann–Whitney–Wilcoxon rank sum test for clustered data.

Key Words and Phrases: clustered data, generalized *U*-statistics, asymptotic normality, Mann–Whitney–Wilcoxon rank sum test.

1 Introduction

One of the most popular tests in non-parametric statistics is Wilcoxon's two-sample rank test. We are given two independent samples

$$X_1, \dots, X_m$$
$$Y_1, \dots, Y_n.$$

The observations in the first sample, X_i, $1 \le i \le m$ are independent identically distributed with distribution function $F(x) = P(X_i \le x)$. The same holds for the observations in the second sample, Y_j, $1 \le j \le n$, whose distribution function is $G(y) = P(Y_j \le y)$. We consider the problem of testing the hypothesis that $F = G$

*Lee's work was supported in part by NIH grant EY12269-04.
†herold.dehling@rub.de

© VVS, 2005. Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

against the alternative that G is stochastically larger than F, i.e. that $G(x) \leq F(x)$ for all x and that $G(x) < F(x)$ for at least one x. The Mann–Whitney–Wilcoxon test statistic is then given by

$$W = \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n}\left(1_{\{X_i<Y_k\}} + \frac{1}{2}1_{\{X_i=Y_k\}}\right),$$

rejecting the hypothesis for large values of W.

The test statistic W is a special case of a two-sample U-statistic. Given a kernel function $h : \mathbb{R}^2 \to \mathbb{R}$ we define the U-statistic

$$U = \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n} h(X_i, Y_k).$$

Hence W is a U-statistic with kernel $h(x, y) = 1_{\{x<y\}} + \frac{1}{2}1_{\{x=y\}}$.

In a medical context, the two-sample experimental design arises in the context of comparison of two treatments. We have a total of $n + m$ patients of which m receive medication A and n receive medication B. The X_i, $1 \leq i \leq m$, are then the observations taken on patients in the first group and the Y_k, $1 \leq k \leq n$, are the observations taken on the patients in the second group.

Clustered data

Often, we face a situation where we are given repeated measurements taken from the same individual, e.g. measurements taken from two eyes of the same individual, blood pressure taken from left and right arms. In this case, the clustered observations are vectors

$$\mathbf{X}^{(i)} = (X_j^{(i)})_{1\leq j\leq J_i}$$
$$\mathbf{Y}^{(k)} = (Y_l^{(k)})_{1\leq l\leq L_k}$$

where $X_1^{(i)},\ldots,X_{J_i}^{(i)}$ denote the measurements taken on the i-th individual (cluster) in the first sample and similarly $Y_1^{(k)},\ldots,Y_{L_k}^{(k)}$ for the second sample. Concerning the distribution, we make the following assumptions

1. The vectors $\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(m)}$, $\mathbf{Y}^{(1)},\ldots,\mathbf{Y}^{(n)}$ from m individuals in the first sample and n individuals in the second sample are independent.
2. The components $X_1^{(i)},\ldots, X_{J_i}^{(i)}$ of the same vector are exchangeable, i.e. for any permutation $\sigma(1),\ldots,\sigma(J_i)$, the vector $(X_{\sigma(1)}^{(i)},\ldots,X_{\sigma(J_i)}^{(i)})$ has the same distribution as $(X_1^{(i)},\ldots,X_{J_i}^{(i)})$. The same assumption holds for the Y-observations.
3. The one-dimensional marginal distributions of all components in the $\mathbf{X}$-vectors are identical and denoted by F and similarly for the $\mathbf{Y}$-vectors where the one-dimensional marginal distributions are denoted by G.

Given a square integrable kernel $\phi(x, y)$ we define the clustered kernel

$$\Phi_{i,k}(\mathbf{X}^{(i)}, \mathbf{Y}^{(k)}) := \sum_{j=1}^{J_i}\sum_{l=1}^{L_k}\phi(X_j^{(i)}, Y_l^{(k)})$$

© VVS, 2005

and the generalized U-statistic of clustered data by

$$\begin{aligned}U_{m,n} &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n}\Phi_{i,k}(\mathbf{X}^{(i)},\mathbf{Y}^{(k)})\\ &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n}\sum_{j=1}^{J_i}\sum_{l=1}^{L_k}\phi(X_j^{(i)},Y_l^{(k)}).\end{aligned} \tag{1}$$

This U-statistic has been introduced and applied in the works of Obuchowski (1997), Rosner and Grove (1999) and Lee and Rosner (2001). It is the main goal of this paper to establish asymptotic normality of $U_{m,n}$ and to derive expressions for the mean and the variance. We will show how general U-statistics theory, originally developed for non-clustered data, can be adapted to the case of clustered data.

In order to present our main results, we introduce the Hoeffding decomposition of the kernel $\phi(x, y)$,

$$\phi(x,y) = \theta + \phi_1(x) + \phi_2(y) + \psi(x,y), \tag{2}$$

where the terms on the right hand side are defined by

$$\theta = \mathrm{E}(\phi(X_j^{(i)},Y_l^{(k)})) = \int\int \phi(x,y)\mathrm{d}F(x)\mathrm{d}G(y) \tag{3}$$

$$\phi_1(x) = \mathrm{E}(\phi(x,Y_l^{(k)})) - \theta = \int \phi(x,y)\mathrm{d}G(y) - \theta \tag{4}$$

$$\phi_2(y) = \mathrm{E}(\phi(X_j^{(i)},y)) - \theta = \int \phi(x,y)\mathrm{d}F(x) - \theta \tag{5}$$

$$\psi(x,y) = \phi(x,y) - \phi_1(x) - \phi_2(y) - \theta. \tag{6}$$

The validity of (2) follows directly from the definition of $\psi(x, y)$ in (6).

Theorem 1. *Under the above assumptions, the generalized two-sample U-statistic for clustered data has mean $\mu_{m,n}$ and variance $\sigma_{m,n}^2$ given by*

$$\begin{aligned}\mu_{m,n} &= \bar{J}_m\bar{L}_n\theta\\ \sigma_{m,n}^2 &= \frac{1}{m}(\bar{L}_n)^2\Big(\bar{J}_m\mathrm{Var}(\phi_1(X_1^{(1)})) + \bar{J}_m^{(2)}\mathrm{Cov}(\phi_1(X_1^{(1)}),\phi_1(X_2^{(1)}))\Big)\\ &\quad + \frac{1}{n}\bar{J}_m^2\Big(\bar{L}_n\mathrm{Var}(\phi_2(Y_1^{(1)})) + \bar{L}_n^{(2)}\mathrm{Cov}(\phi_2(Y_1^{(1)}),\phi_2(Y_2^{(1)}))\Big)\\ &\quad + \frac{1}{mn}\bar{J}_m\bar{L}_n\mathrm{Var}(\psi(X_1^{(1)},Y_1^{(1)}))\\ &\quad + \frac{1}{mn}\bar{J}_m^{(2)}\bar{L}_n\mathrm{Cov}(\psi(X_1^{(1)},Y_1^{(1)}),\psi(X_2^{(1)},Y_1^{(1)}))\\ &\quad + \frac{1}{mn}\bar{J}_m\bar{L}_n^{(2)}\mathrm{Cov}(\psi(X_1^{(1)},Y_1^{(1)}),\psi(X_1^{(1)},Y_2^{(1)}))\\ &\quad + \frac{1}{mn}\bar{J}_m^{(2)}\bar{L}_n^{(2)}\mathrm{Cov}(\psi(X_1,Y_1^{(1)}),\psi(X_2^{(1)},Y_2^{(1)})),\end{aligned}$$

© VVS, 2005

where the four quantities $\bar{J}_m, \bar{J}_m^{(2)}, \bar{L}_n, \bar{L}_n^{(2)}$ *are defined by* $\bar{J}_m := \frac{1}{m}\sum_{i=1}^m J_i$, $\bar{J}_m^{(2)} := \frac{1}{m}\sum_{i=1}^m (J_i^2 - J_i)$, $\bar{L}_n := \frac{1}{n}\sum_{k=1}^n L_k$ *and* $\bar{L}_n^{(2)} := \frac{1}{n}\sum_{k=1}^n (L_k^2 - L_k)$, *respectively.*

As m, n become large, the U-statistic $U_{m,n}$ becomes asymptotically normal. Moreover, in the expression for the variance of $U_{m,n}$, the first two terms dominate. Thus asymptotically the variance becomes

$$\tau_{m,n}^2 := \frac{1}{m}(\bar{L}_n)^2\left(\bar{J}_m \mathrm{Var}(\phi_1(X_1^{(1)})) + \bar{J}_m^{(2)}\mathrm{Cov}(\phi_1(X_1^{(1)}), \phi_1(X_2^{(1)}))\right)$$
$$+\frac{1}{n}\bar{J}_m^2\left(\bar{L}_n \mathrm{Var}(\phi_2(Y_1^{(1)})) + \bar{L}_n^{(2)}\mathrm{Cov}(\phi_2(Y_1^{(1)}, \phi_2(Y_2^{(1)})))\right).$$

THEOREM 2. *Suppose that the cluster sizes* J_i *and* L_k *are bounded, i.e. that* $\sup_i J_i < \infty$ *and* $\sup_k L_k < \infty$. *Then, under the above assumptions, the generalized U-statistic* $U_{m,n}$ *is asymptotically normal with mean* $\mu_{m,n}$ *and variance* $\tau_{m,n}^2$, *i.e.*

$$\frac{U_{m,n} - \mu_{m,n}}{\tau_{m,n}} \to N(0, 1)$$

in distribution, as $\min(m, n) \to \infty$.

The rest of this paper is organized as follows: in section 3 we provide the proofs of Theorem 1 and Theorem 2. In section 2 we apply the general theory to the generalized Mann–Whitney–Wilcoxon test for clustered data.

2 Generalized Mann–Whitney–Wilcoxon rank sum test

The generalized two-sample Mann–Whitney–Wilcoxon rank sum for clustered data is given by

$$W = \frac{1}{mn}\sum_{i=1}^m\sum_{k=1}^n\sum_{j=1}^{J_i}\sum_{l=1}^{L_k}\left(1_{\{X_j^{(i)}<Y_l^{(k)}\}} + \frac{1}{2}1_{\{X_j^{(i)}=Y_l^{(k)}\}}\right). \tag{7}$$

Observe that W is a generalized U-statistic with kernel $\phi(x, y) = 1_{\{x<y\}} + \frac{1}{2}1_{\{x=y\}}$. ROSNER and GROVE (1999) have proposed W as a test statistic for testing the hypothesis $H : F = G$ against the alternative that G is stochastically larger than F, i.e. that $1 - G(x) \geq 1 - F(x)$ for all $x \in \mathbb{R}$ and $1 - G(x) > 1 - F(x)$ for some x.

Rosner and Grove obtained an explicit formula for the variance of W under the null hypothesis. They also conducted a simulation study indicating asymptotic normality of W for large sample sizes. We will show in this section that asymptotic normality can be obtained as a consequence of Theorem 2.

For the kernel

$$\phi(x,y) = 1_{\{x<y\}} + \frac{1}{2}1_{\{x=y\}}$$

the parameters and functions entering in the formula for the mean and the variance of the associated generalized U-statistic can be calculated explicitly. We obtain

© VVS, 2005

$$\begin{aligned}\theta &= P(X_1^{(1)} < Y_1^{(1)}) + \frac{1}{2}P(X_1^{(1)} = Y_1^{(1)})\\ \phi_1(x) &= P(x < Y_1^{(1)}) + \frac{1}{2}P(Y_1^{(1)} = x) - \theta\\ \phi_2(y) &= P(X_1^{(1)} < y) + \frac{1}{2}P(X_1^{(1)} = y) - \theta.\end{aligned}$$

The parameters entering in the formula for the asymptotic variance of the Mann–Whitney–Wilcoxon rank sum are then

$$\begin{aligned}v_x &= \mathrm{Var}(\phi_1(X_1^{(1)}))\\ v_y &= \mathrm{Var}(\phi_2(Y_1^{(1)}))\\ r_x &= \mathrm{Cov}(\phi_1(X_1^{(1)}), \phi_1(X_2^{(1)}))\\ r_y &= \mathrm{Cov}(\phi_2(Y_1^{(1)}), \phi_2(Y_2^{(1)}))\end{aligned}$$

THEOREM 3. *Suppose that the cluster sizes J_i and L_k are bounded. Then the generalized Mann–Whitney–Wilcoxon rank sum W is asymptotically normal with mean $\mu_{m,n} = \bar{J}_m\bar{L}_n(P(X_1^{(1)} < Y_1^{(1)}) + \frac{1}{2}P(X_1^{(1)} = Y_1^{(1)}))$ and variance*

$$\tau_{m,n}^2 = \frac{1}{m}(\bar{L}_n)^2\left(\bar{J}_m v_x + \bar{J}_m^{(2)} r_x\right) + \frac{1}{n}(\bar{J}_m)^2\left(\bar{L}_n v_y + \bar{L}_n^{(2)} r_y\right).$$

Asymptotic distribution in the continuous case

The above expressions for mean and variance of W can be considerably simplified when the marginal distributions F and G are continuous. In this case, we may replace the kernel $1_{\{x<y\}} + 1_{\{x=y\}}$ by $\phi(x, y) = 1_{\{x<y\}}$. We thus obtain

$$\begin{aligned}\theta &= P(X_1^{(1)} < Y_1^{(1)})\\ \phi_1(x) &= P(x < Y_1^{(1)}) - \theta = 1 - G(x) - \theta\\ \phi_2(y) &= P(X_1^{(1)} < y) - \theta = F(y) - \theta.\end{aligned}$$

For the parameters v_x, v_y, r_x, r_y, we obtain the following expressions:

$$\begin{aligned}v_x &= \mathrm{Var}(G(X_1^{(1)})) = P(X_1^{(1)} \geq Y_1^{(1)}, X_1^{(1)} \geq Y_1^{(2)}) - (P(X_1^{(1)} \geq Y_1^{(1)}))^2\\ v_y &= \mathrm{Var}(F(Y_1^{(1)})) = P(X_1^{(1)} \geq Y_1^{(1)}, X_1^{(2)} \geq Y_1^{(1)}) - (P(X_1^{(1)} \geq Y_1^{(1)}))^2\\ r_x &= \mathrm{Cov}(G(X_1^{(1)}), G(X_2^{(1)})) = P(X_1^{(1)} \geq Y_1^{(1)}, X_2^{(1)} \geq Y_1^{(2)}) - (P(X_1^{(1)} \geq Y_1^{(1)}))^2\\ r_y &= \mathrm{Cov}(F(Y_1^{(1)}), F(Y_2^{(1)})) = P(X_1^{(1)} \geq Y_1^{(1)}, X_1^{(2)} \geq Y_2^{(1)}) - (P(X_1^{(1)} \geq Y_1^{(1)}))^2.\end{aligned}$$

Note that the expressions for r_x involves two different observations $X_1^{(1)}$, $X_2^{(1)}$ of the same subject from the X-population and two observations $Y_1^{(1)}$, $Y_1^{(2)}$ from two

© VVS, 2005

different subjects of the Y-population. In the same way, the expression for r_y involves two observations $Y_1^{(1)}$, $Y_2^{(1)}$ of the same subject in the Y-population and two observations $X_1^{(1)}$, $X_1^{(2)}$ from different subjects in the X-population.

In the special case when there is no clustering, i.e. when $J_i = L_k = 1$ for all i and k, the formula for the asymptotic variance $\tau_{m,n}^2$ reduces to the well-known formula for the asymptotic variance of the ordinary Mann–Whitney–Wilcoxon rank sum, i.e.

$$\tau_{m,n}^2 = \frac{1}{m} v_x + \frac{1}{n} v_y = \frac{1}{m} \mathrm{Var}(G(X)) + \frac{1}{n} \mathrm{Var}(F(Y)).$$

Asymptotic distribution under the null hypothesis, continuous case
Under the null hypothesis $F = G$, both $G(X)$ and $F(Y)$ will have a uniform distribution on the unit interval [0, 1]. Thus we obtain $v_x = v_y = 1/12$ and

$$r_x = P(X_1^{(1)} \geq Y_1^{(1)}, X_2^{(1)} \geq Y_1^{(2)}) - \frac{1}{4} \tag{8}$$

$$r_y = P(X_1^{(1)} \leq Y_1^{(1)}, X_1^{(2)} \leq Y_2^{(1)}) - \frac{1}{4}. \tag{9}$$

In the special case of no clustering we obtain the following formula for the asymptotic variance of W,

$$\tau_{m,n}^2 = \frac{1}{m} v_x + \frac{1}{n} v_y = \frac{1}{12} \frac{m+n}{mn}.$$

Note that $\tau_{m,n}^2$ is close to the exact variance of W under the null hypothesis,

$$\mathrm{Var}(W) = \frac{1}{12} \frac{m+n+1}{mn}.$$

The parameters r_x and r_y can be estimated from the data by their sample analogues. As estimator for r_x we take the average number of times that $X_j^{(i)} \geq Y_l^{(k)}$ and $X_{j'}^{(i)} \geq Y_{l'}^{(k')}$ where $1 \leq i \leq m$, $1 \leq k \neq k' \leq n$, $1 \leq j \neq j' \leq J_i$, $1 \leq l, l' \leq L_k$. The number of such indices is

$$\begin{aligned}\left(\sum_{i=1}^{m} J_i(J_i - 1)\right)\left(\sum_{1 \leq k \neq k' \leq n} L_k L_k'\right) &= m\bar{J}_m^{(2)}\left(\left(\sum_{k=1}^{n} L_k\right)^2 - \sum_{k=1}^{n} L_k^2\right) \\ &= m\bar{J}_m^{(2)}\left(n^2(\bar{L}_n)^2 - n\bar{L}_n^{(2)} + n\bar{L}_n\right)\end{aligned}$$

so that we obtain the following estimator for r_x:

$$\hat{r}_x = \frac{\sum_{i=1}^{m} \sum_{1 \leq k \neq k' \leq n} \sum_{1 \leq j \neq j' \leq J_i} \sum_{l=1}^{L_k} \sum_{l'=1}^{L_{k'}} 1_{\{X_j^{(i)} \geq Y_l^{(k)}, X_{j'}^{(i)} \geq Y_{l'}^{(k')}\}}}{m\bar{J}_m^{(2)}\left(n^2(\bar{L}_n)^2 - n\bar{L}_n^{(2)} + n\bar{L}_n\right)} - \frac{1}{4}.$$

3 Proofs of main theorems

In this section, we will present the proofs of Theorems 1 and 2. These proofs require some auxiliary results which we will formulate first and then prove in the next section. The main tool in the proofs is the Hoeffding decomposition for generalized two-sample U-statistics. From the decomposition 2 of the kernel $\phi(x, y)$, we obtain the following decomposition of $U_{m,n}$,

$$\begin{aligned} U_{m,n} &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n}\sum_{j=1}^{J_i}\sum_{l=1}^{L_k}\phi(X_j^{(i)}, Y_l^{(k)}) \\ &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n}\sum_{j=1}^{J_i}\sum_{l=1}^{L_k}\Big(\theta + \phi_1(X_j^{(i)}) + \phi_2(Y_l^{(k)}) + \psi(X_j^{(i)}, Y_l^{(k)})\Big) \\ &= \bar{J}_m\bar{L}_n\theta + \bar{L}_n\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{J_i}\phi_1(X_j^{(i)}) + \bar{J}_m\frac{1}{n}\sum_{k=1}^{n}\sum_{l=1}^{L_k}\phi_2(Y_l^{(k)}) \\ &\quad + \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n}\Psi_{i,k}(\mathbf{X}^{(i)}, \mathbf{Y}^{(k)}), \end{aligned} \tag{10}$$

where $\Psi_{i,k}$ is defined by

$$\Psi_{i,k}(\mathbf{X}^{(i)}, \mathbf{Y}^{(k)}) := \sum_{j=1}^{J_i}\sum_{l=1}^{L_k}\psi(X_j^{(i)}, Y_l^{(k)}).$$

Proposition 1. *The three sums on the right hand side of* 10*, i.e.* $\sum_{i=1}^{m}\sum_{j=1}^{J_i}\phi_1(X_j^{(i)})$, $\sum_{k=1}^{n}\sum_{l=1}^{L_k}\phi_2(Y_l^{(k)})$ *and* $\sum_{i=1}^{m}\sum_{k=1}^{n}\Psi_{i,k}(\mathbf{X}^{(i)}, \mathbf{Y}^{(k)})$*, are pairwise uncorrelated.*

As a consequence of Proposition 1, we can compute the variance of $U_{m,n}$ as the sum of the variances of the terms on the right hand side of 10. In what follows, we shall determine each of these variances separately.

Proposition 2. *The variances of the linear terms in the Hoeffding decomposition are given by*

$$\mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{J_i}\phi_1(X_j^{(i)})\right) = \frac{1}{m}\Big(\bar{J}_m\mathrm{Var}(\phi_1(X_1^{(1)})) + \bar{J}_m^{(2)}\mathrm{Cov}(\phi_1(X_1^{(1)}), \phi_1(X_2^{(1)}))\Big) \tag{11}$$

$$\mathrm{Var}\left(\frac{1}{n}\sum_{k=1}^{n}\sum_{l=1}^{L_k}\phi_2(Y_l^{(k)})\right) = \frac{1}{n}\Big(\bar{L}_n\mathrm{Var}(\phi_2(Y_1^{(1)})) + \bar{L}_n^{(2)}\mathrm{Cov}(\phi_2(Y_1^{(1)}), \phi_2(Y_2^{(1)}))\Big). \tag{12}$$

© VVS, 2005

PROPOSITION 3. *The variance of the second order term in the Hoeffding decomposition is given by*

$$\mathrm{Var}\left(\frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n}\Psi_{i,k}(\mathbf{X}^{(i)},\mathbf{Y}^{(k)})\right)=\frac{1}{m^2n^2}\sum_{i=1}^{m}\sum_{k=1}^{n}\mathrm{Var}(\Psi_{i,k}(\mathbf{X}^{(i)},\mathbf{Y}^{(k)})), \tag{13}$$

where

$$\begin{aligned}\mathrm{Var}\Big(\Psi_{i,k}(\mathbf{X}^{(i)},\mathbf{Y}^{(k)})\Big) &= J_iL_k\mathrm{Var}\Big(\psi(X_1^{(1)},Y_1^{(1)})\Big)\\ &+J_i(L_k^2-L_k)\mathrm{Cov}\Big(\psi(X_1^{(1)},Y_1^{(1)}),\psi(X_1^{(1)},Y_2^{(1)})\Big)\\ &+(J_i^2-J_i)L_k\mathrm{Cov}\Big(\psi(X_1^{(1)},Y_1^{(1)}),\psi(X_2^{(1)},Y_1^{(1)})\Big)\\ &+(J_i^2-J_i)(L_k^2-L_k)\mathrm{Cov}\Big(\psi(X_1^{(1)},Y_1^{(1)}),\psi(X_2^{(1)},Y_2^{(1)})\Big).\end{aligned} \tag{14}$$

PROOF OF THEOREM 1. The expression for the mean $\mu_{m,n}$ follows directly from linearity of the mean and the fact that $E(\phi(X_j^{(i)},\ Y_l^{(k)}))\ =\ \theta$. The expression for the variance is a direct consequence of Proposition 1, Proposition 2 and Proposition 3. □

LEMMA 1. *Let* $\xi^{(i)}\ =\ (\xi_1^{(i)},\ldots,\ \xi_{J_i}^{(i)})$*,* $i=1,2,\ldots$ *be independent vectors whose coordinates* $\xi_1^{(i)},\ldots,\ \xi_{J_i}^{(i)}$ *are exchangeable for any fixed i. Assume moreover that* $\sup_i J_i<\infty$, $\inf_i \mathrm{Var}(\sum_{j=1}^{J_i}\xi_j^{(i)})\ >\ 0$ *and that all pairs* $(\xi_j^{(i)},\ \xi_{j'}^{(i)}), j\neq j'$*, have the same distribution. Then*

$$S_n:=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\xi_j^{(i)}$$

is asymptotically normally distributed with mean $\mu_n\ =\ \bar{J}_n\mathrm{E}(\xi_1^{(1)})$ *and variance*

$$s_n^2=\frac{1}{n}\left(\bar{J}_n\mathrm{Var}\Big(\xi_1^{(1)}\Big)+\bar{J}_n^{(2)}\mathrm{Cov}\Big(\xi_1^{(1)},\xi_2^{(1)}\Big)\right).$$

PROOF OF THEOREM 2. Applying the Hoeffding decomposition for generalized U-statistics (10), we get

$$\begin{aligned}U_{m,n}-\bar{J}_m\bar{L}_n\theta &= \bar{L}_n\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{J_i}\phi_1(X_j^{(i)})+\bar{J}_m\frac{1}{n}\sum_{k=1}^{n}\sum_{l=1}^{L_k}\phi_2(Y_l^{(k)})\\ &+\frac{1}{nm}\sum_{i=1}^{m}\sum_{k=1}^{n}\Psi_{i,k}(\mathbf{X}^{(i)},\mathbf{Y}^{(k)})\end{aligned}$$

Of the three terms on the right hand side, the third is of lower order than the first two and may thus be neglected. The first two terms are independent and we can apply the Central Limit Theorem of Lemma 1 to each of them. Together this yields Theorem 2. □

© VVS, 2005

4 Appendix: proofs of auxiliary results

We first observe that the functions ϕ_1, ϕ_2 and ψ appearing in the Hoeffding decomposition (2) of $\phi(x, y)$ satisfy the following identities:

$$E(\phi_1(X_j^{(i)})) = E(\phi_2(Y_l^{(k)})) = 0, \tag{15}$$

$$E(\psi(x, Y_l^{(k)})) = E(\psi(X_j^{(i)}, y)) = 0, \tag{16}$$

for all x, y. The identities (15) follow directly from Fubini's theorem and independence of $X_j^{(i)}$ and $Y_l^{(k)}$. For a proof of the second identity, we observe that $E(\phi(x, Y_l^{(k)})) = \phi_1(x) + \theta$ and $E(\phi(X_j^{(i)}, y)) = \phi_2(y) + \theta$, by definition of ϕ_1 and ϕ_2, respectively. We thus obtain

$$\begin{aligned} E\Big(\psi(x, Y_l^{(k)})\Big) &= E\Big(\phi(x, Y_l^{(k)}) - \phi_1(x) - \phi_2(Y_l^{(k)}) - \theta\Big) \\ &= \phi_1(x) + \theta - \phi_1(x) - E(\phi_2(Y_l^{(k)})) - \theta = 0. \end{aligned}$$

Similarly, one can show that $E(\psi(X_i^{(j)}, y)) = 0$.

Proof of Proposition 1. It suffices to show pairwise uncorrelatedness of the summands occuring in the last three terms of (10), i.e. of $\phi_1(X_j^{(i)})$, $\phi_2(Y_2^{(k)})$ and $\psi(X_{j'}^{(i')}, Y_{l'}^{(k')})$. The first two terms are independent by assumption, and hence it remains to show that $\phi_1(X_j^{(i)})$ and $\psi(X_{j'}^{(i')}, Y_{l'}^{(k')})$ are uncorrelated (and the same for $\phi_2(Y_2^{(k)})$ and $\psi(X_{j'}^{(i')}, Y_{l'}^{(k')})$). We obtain by Fubini's theorem and independence of the X- and Y-observations

$$E\Big(\phi_1(X_j^{(i)})\psi(X_{j'}^{(i')}, Y_{l'}^{(k')})\Big) = E\Big(E\Big(\phi_1(X_j^{(i)})\psi(X_{j'}^{(i')}, Y_{l'}^{(k')})|X_j^{(l)}, X_{j'}^{(l')}\Big)\Big) = 0,$$

where we have used (16) in the last step. □

Proof of Proposition 2. By independence of the vectors $\mathbf{X}^{(i)}$, $1 \le i \le m$, we have

$$\mathrm{Var}\left(\sum_{i=1}^{m}\sum_{j=1}^{J_i}\phi_1(X_j^{(i)})\right) = \sum_{i=1}^{m}\mathrm{Var}\left(\sum_{j=1}^{J_i}\phi_1(X_j^{(i)})\right).$$

The variance inside the sum can again be expressed as

$$\begin{aligned} \mathrm{Var}\left(\sum_{j=1}^{J_i}\phi_1(X_j^{(i)})\right) &= \sum_{j=1}^{J_i}\mathrm{Var}(\phi_1(X_j^{(i)})) + \sum_{1\le j_1\ne j_2\le J_i}\mathrm{Cov}(X_{j_1}^{(i)}, X_{j_2}^{(i)}) \\ &= J_i\mathrm{Var}(\phi_1(X_1^{(i)})) + J_i(J_i - 1)\mathrm{Cov}(X_1^{(i)}, X_2^{(i)}), \end{aligned}$$

where in the final step we have used exchangeability of the observations $X_1^{(i)}, \ldots, X_{J_i}^{(i)}$. This proves the first identity, and the same proof also yields the second identity. □

Proof of Proposition 3. For a proof of (13) it suffices to show that $\psi(X_j^{(i)}, Y_l^{(k)})$ and $\psi(X_{j'}^{(i')}, Y_{l'}^{(k')})$ are uncorrelated whenever $i \neq i'$ or $k \neq k'$. If the latter holds, i.e. if $k \neq k'$, we get using (16)

$$\begin{aligned} & E\Big(\psi(X_j^{(i)}, Y_l^{(k)})\psi(X_{j'}^{(i')}, Y_{l'}^{(k')})\Big) \\ & \quad = E\Big(E\Big(\psi(X_j^{(i)}, Y_l^{(k)})\psi(X_{j'}^{(i')}, Y_{l'}^{(k')})|X_j^{(i)}, X_{j'}^{(i')}, Y_{l'}^{(k')}\Big)\Big) \\ & \quad = E\Big(\psi(X_{j'}^{(i')}, Y_{l'}^{(k')})E\Big(\psi(X_j^{(i)}, Y_l^{(k)})|X_j^{(i)}, X_{j'}^{(i')}, Y_{l'}^{(k')}\Big)\Big) = 0, \end{aligned}$$

i.e. $\psi(X_j^{(i)}, Y_l^{(k)})$ and $\psi(X_{j'}^{(i')}, Y_{l'}^{(k')})$ are uncorrelated. In order to prove (14), we note first that

$$\begin{aligned} \mathrm{Var}(\Psi_{i,k}(\mathbf{X}^{(i)}, \mathbf{Y}^{(k)})) &= \mathrm{Var}\left(\sum_{j=1}^{J_i}\sum_{l=1}^{L_k}\psi(X_j^{(i)}, Y_l^{(k)})\right) \\ &= \sum_{j=1}^{J_i}\sum_{l=1}^{L_k}\sum_{j'=1}^{J_i}\sum_{l'=1}^{L_k}\mathrm{Cov}(\psi(X_j^{(i)}, Y_l^{(k)}), \psi(X_{j'}^{(i)}, Y_{l'}^{(k)}). \end{aligned}$$

Concerning the covariances occuring on the r.h.s. of this identity, we have to distinguish four different cases, namely (i) $j = j'$, $l = l'$, (ii) $j = j'$, $l \neq l'$, (iii) $j \neq j'$, $l = l'$ and (iv) $j \neq j'$, $l \neq l'$. There are respectively J_{iLk}, $J_i(L_k^2 - L_k)$, $(J_i^2 - J_i)L_k$ and $(J_i^2 - J_i)(L_k^2 - L_k)$ such index quadruples and we get the following values for the covariance:

$$\begin{aligned} \mathrm{Cov}(\psi(X_j^{(i)}, Y_l^{(k)}), \psi(X_j^{(i)}, Y_l^{(k)})) &= \mathrm{Var}(\psi(X_1^{(1)}, Y_1^{(1)})) \\ \mathrm{Cov}(\psi(X_j^{(i)}, Y_l^{(k)}), \psi(X_j^{(i)}, Y_{l'}^{(k)})) &= \mathrm{Cov}(\psi(X_1^{(1)}, Y_1^{(1)}), \psi(X_1^{(1)}, Y_2^{(1)})) \\ \mathrm{Cov}(\psi(X_j^{(i)}, Y_l^{(k)}), \psi(X_{j'}^{(i)}, Y_l^{(k)})) &= \mathrm{Cov}(\psi(X_1^{(1)}, Y_1^{(1)}), \psi(X_2^{(1)}, Y_1^{(1)})) \\ \mathrm{Cov}(\psi(X_j^{(i)}, Y_l^{(k)}), \psi(X_{j'}^{(i)}, Y_{l'}^{(k)})) &= \mathrm{Cov}(\psi(X_1^{(1)}, Y_1^{(1)}), \psi(X_2^{(1)}, Y_2^{(1)})). \end{aligned}$$

Putting everything together, we obtain (14). □

Proof of Lemma 1. Without loss of generality we may assume that $E\xi_j^{(i)} = 0$. We define the block sums

$$U_i := \sum_{j=1}^{J_i}\xi_j^{(i)},$$

and note that $(U_i)_{i\geq 1}$ is a sequence of independent random variables, satisfying

$$\mathrm{Var}(U_i) = J_i\mathrm{Var}(\xi_1^{(1)}) + J_i(J_i - 1)\mathrm{Cov}(\xi_1^{(1)}, \xi_2^{(1)}).$$

Observe that $S_n = \frac{1}{n}\sum_{i=1}^n U_i$ and that

$$\operatorname{Var}\left(\sum_{i=1}^{n} U_i\right) = \sum_{i=1}^{n} \operatorname{Var}(U_i) = n^2 s_n^2,$$

by definition of s_n^2. Thus

$$\frac{S_n}{s_n} = \sum_{i=1}^{n} \frac{U_i}{n s_n}$$

is a sum of independent random variables whose variances add to 1. In order to prove asymptotic normality, it suffices to establish the Lindeberg condition. We have

$$\begin{aligned} L_n(\epsilon) &= \sum_{i=1}^{n} E\left(\frac{U_i}{n s_n}\right)^2 1_{\{|U_i| \geq \epsilon n s_n\}} \\ &= \frac{1}{n^2 s_n^2} \sum_{i=1}^{n} E(U_i^2) 1_{\{|U_i| \geq \epsilon n s_n\}} \end{aligned}$$

By assumption, $n^2 s_n^2 = \sum_{i=1}^{n} \operatorname{Var}(U_i) \geq n\sigma^2$ where $\sigma^2 := \inf_i \operatorname{Var}(U_i)$. Moreover, by the Cauchy–Schwarz inequality we have $U_i^2 \leq J_i \sum_{j=1}^{J_i} (\xi_1^{(i)})^2$ and thus

$$L_n(\epsilon) \leq \frac{1}{n\sigma^2} \sum_{i=1}^{n} J_i^2 E\left(\xi_1^{(i)}\right)^2 1_{\{|U_i| \geq \epsilon n s_n\}}.$$

The family of random variables $(\xi_1^{(i)})^2$ is uniformly integrable. Moreover,

$$P(|U_i| \geq \epsilon n s_n) \leq \frac{1}{\epsilon^2} \frac{\operatorname{Var}(U_i)}{(n s_n)^2} \leq \frac{J_i^2 E(\xi_1^{(1)})^2}{n\sigma^2} \rightarrow 0,$$

by Chebychev's inequality. Thus $L_n(\epsilon) \rightarrow 0$, establishing the Lindeberg condition. □

References

Lee, M.-L. T. and B. Rosner (2001), The average for areas under correlated ROC curves: a nonparametric approach based on generalized two-sample Wilcoxon statistics, *Applied Statistics* **50**, 337–344.

Obuchowski, N. A. (1997), Nonparametric analysis of clustered ROC curve data, *Biometrics* **53**, 567–578.

Rosner, B. and D. Grove (1999), Use of the Mann–Whitney U-test for clustered Data, *Statistics in Medicine* **18**, 1387–1400.

Received: December 2004. Revised: January 2005.

© VVS, 2005