Записки научных
семинаров ПОМИ
Том 260, 1999 г.

P. L. Conti and Ya. Yu. Nikitin

RATES OF CONVERGENCE FOR A CLASS OF RANK TESTS FOR INDEPENDENCE

Abstract. Almost optimal rate of convergence result is obtained for a large class of rank statistics for testing independence including the Gini's and Spearman's rank correlation coefficients as well as the Spearman's footrule.

1. Introduction

Let $\{(X_i, Y_i);\ i \geq 1\}$ be a sequence of independent and identically distributed (i.i.d.) random variables (r.v.'s) taking values in $\mathbb{R}^2$, and defined on some probability space $(\Omega, \mathcal{F}, P)$; $(X_1, Y_1), \dots, (X_n, Y_n)$ is a sample of size n. Denote by $H(x, y) = P(X_1 \leq x, Y_1 \leq y)$ the joint distribution function (d.f.) of (X_1, Y_1), and by $F(x) = H(x, +\infty)$ and $G(y) = H(+\infty, y)$ the marginal d.f.'s of X_1 and Y_1, respectively. In the sequel we suppose that H, F, and G are continuous.

Testing for independence between the components of the bidimensional r.v. (X_1, Y_1) using a rank test is one of the important and interesting problems of nonparametric statistics. Classical rank statistics for this problem are Kendall's τ [12], Spearman's rank correlation coefficient and footrule [17, 18], Gini's cograduation coefficient [10]. Other test statistics and related results can be found in Daniels [8], Bhuchongkul [1], Ruymgaart [15]; see also the book of Puri and Sen [14] with numerous bibliographical references.

In this paper we focus our attention on a new class of rank statistics recently introduced by Cifarelli, Conti and Regazzini [5]. Let R_i be the rank of X_i among X's, and S_i the rank of the corresponding Y_i among Y's. Furthermore, let $g : [0,1] \to \mathbb{R}_+$ be a continuous increasing and convex function, such that $g(0) = 0$. The test statistics considered in [5] take the general form (up to an inessential scale factor)

$$\gamma_{g,n} = \frac{1}{n} \sum_{i=1}^{n} \left\{ g\left(n^{-1} |R_i + S_i - (n+1)|\right) - g\left(n^{-1} |R_i - S_i|\right) \right\}. \qquad (1)$$

Research supported by RFBR, grant No. 99-01-00724 .

Note that when $g(x) = x$, the statistic (1) reduces (up to a scale factor) to Gini's cograduation index; when $g(x) = x^2$, (1) turns out to be the Spearman's rank correlation coefficient (again up to a scale factor).

It was proved in [5] that, under suitable conditions,

$$\sqrt{n}(\gamma_{g,n} - \gamma_g) \xrightarrow{d} N(0, \sigma_g^2),$$

where

$$\gamma_g = \int\limits_{\mathbb{R}^2} \{g(|F(x) + G(y) - 1|) - g(|F(x) - G(y)|\} \; dH(x, y)$$

and σ_g^2 is the corresponding asymptotic variance.

In this note we go a step further by obtaining the rate of convergence for the statistic (1). In principle, this problem was already studied by Conti [6]. However, the result obtained there was based on very restrictive regularity conditions, and did not cover many interesting cases. For instance, no rate of convergence was obtained for Gini's cograduation coefficient, which is of interest in view of its optimality properties (see [5]).

Rates of convergence for rank statistics were studied by several authors. A good review is in the paper by Hallin and Puri [11], which contains also extensive bibliographic references. However, the general class considered in that paper does not contain statistics of the form (1). An estimate for the rate of convergence was obtained by Bolthausen and Götze [3] (see also [2]) for a general class of multivariate statistics that contains (1) as a special case. But their results cover only the case $H(x, y) = F(x)G(y)$. In other words, the result is restricted to the case of independence of X_i's and Y_i's.

The main goal of the present paper is to derive a rate of convergence for statistics of the form (1) by imposing conditions which are considerably less stringent than in [6] and without assuming independence between X_i's and Y_i's. As a by-product, we will obtain a "nearly optimal" rate of convergence for the Gini's statistic.

Our results may be used also when calculating the Pitman efficiency of tests of independence based on γ_g. Local Bahadur efficiency of such tests and conditions of their local optimality were studied recently in Conti and Nikitin [7].

2. Rate of convergence: basic results

The main result of the present section is Theorem 1 which is based on following conditions (i), (ii). Such conditions can be slightly relaxed, see the remark below.

(i) The derivative $g'(x)$ exists for every $x \in (0,1)$, it is continuous on $(0,1)$ and is such that $\lim_{x\to 0^+} g'(x)$ and $\lim_{x\to 1^-} g'(x)$ both exist and are finite.

(ii) The function $g'(x)$ satisfies a Hölder condition of order $\alpha \in (0,1]$:

$$|g'(x+u) - g'(x)| \le C_g |u|^\alpha$$

for every $x \in (0,1)$, $x + u \in (0,1)$, C_g being a positive constant.

Remark 1. If $\lim_{x\to 0^+} g'(x) = 0$, then the function $f(u,v) = g(|u-v|)$ is differentiable on the unit square; this is the case covered in [6]. However, for many interesting cases, e. g., Gini's cograduation coefficient, this hypothesis is not fullfilled.

We start here from an idea developed in [6], i.e., to approximate $\gamma_{g,n}$ by a (nondegenerate) U-statistic of degree 2:

$$U_{g,n} = \binom{n}{2}^{-1} \sum_{i<j} \Psi(X_i, Y_i, X_j, Y_j). \tag{2}$$

The kernel $\Psi(\cdot\,;\cdot\,)$ is essentially defined as in [6]. Now we will give its exact form. Put $f(x,y) = g(|x-y|)$, and define:

$$f_{10}(x,y) = \frac{\partial f}{\partial x} = g'(|x-y|)\,\mathrm{sign}\,(x-y);$$

$$f_{01}(x,y) = \frac{\partial f}{\partial y} = -f_{10}(x,y).$$

Consider the Heaviside function $u(x) = 1_{\{x\ge 0\}}$, and introduce the functions:

$$\begin{aligned} v_C(x_i,y_i;x_j,y_j) =& f(x_i,1-y_i) + (u(x_i-x_j) - x_i) f_{10}(x_i,1-y_i) \\ &- (u(y_i-y_j) - y_i) f_{01}(x_i,1-y_i), \\ k_C(x_i,y_i;x_j,y_j) =& \{v_C(x_i,y_i;x_j,y_j) + v_C(x_j,y_j;x_i,y_i)\}/2. \end{aligned}$$

The U-statistic $U_{C,n}$ is defined as

$$U_{C,n} = \binom{n}{2}^{-1} \sum_{i<j} k_C(X_i,Y_i;X_j,Y_j).$$

Similarly, consider the functions:

$$\begin{aligned} v_D(x_i,y_i;x_j,y_j) &= f(x_i,y_i) + (u(x_i-x_j) - x_i) f_{10}(x_i,y_i) \\ &\quad + (u(y_i-y_j) - y_i) f_{01}(x_i,y_i) \\ k_D(x_i,y_i;x_j,y_j) &= \{v_D(x_i,y_i;x_j,y_j) + v_D(x_j,y_j;x_i,y_i)\}/2. \end{aligned}$$

The U-statistic $U_{D,n}$ is then defined as

$$U_{D,n} = \binom{n}{2}^{-1} \sum_{i<j} k_D(X_i, Y_i; X_j, Y_j).$$

Finally, the U-statistic $U_{g,n}$ from (2) possesses kernel Ψ of the form:

$$\Psi(X_i, Y_i; X_j, Y_j) = k_C(X_i, Y_i; X_j, Y_j) - k_D(X_i, Y_i; X_j, Y_j).$$

It was shown in [6] that $\sqrt{n}(\gamma_{g,n} - U_{g,n})$ tends in probability to zero as $n \to \infty$.

Lemma 1 refines this result. Without loss of generality, the marginal d.f.'s F and G can be assumed uniform in the interval $(0,1)$.

Lemma 1. *Assume that the d.f. $H(x,y)$ has a bounded density $h(x,y)$ and let $\sup_{x,y} h(x,y) \leq K$. Then, under conditions* (i), (ii)*:*

$$P(\sqrt{n}|U_{g,n} - \gamma_{g,n}| > C^* K \epsilon_n) \leq$$
$$\leq Q\left(\exp\left\{-2(n/\log n)^{\alpha}\ \epsilon_n^2\right\} + n^{-2} + \exp\left\{-K^2 n^{1-\alpha} (\log n)^{\alpha}\right\}\right) \quad (3)$$

for every sequence $\{\epsilon_n\}_{n\geq 1}$ of positive numbers, Q and C^* being positive constants independent of K.

Proof. See Appendix.

Now we are able to prove our main result. Define

$$\mu(h) = E\{g(|X_1 + Y_1 - 1|) - g(|X_1 - Y_1|)\},$$
$$\sigma^2(h) = \lim_{n\to\infty} \mathrm{Var}\,(U_{g,n})$$

Theorem 1. *Let $\mathcal{H}$ be a class of densities $h(x,y)$ with uniform marginal densities such that $\sup_{h\in\mathcal{H}} \sup_{x,y} h(x,y)$ is finite. Furthermore, assume that conditions* (i), (ii) *are fulfilled. Then:*

$$\sup_{h\in\mathcal{H}} \sup_{y\in\mathbb{R}} |P((\gamma_{g,n} - \mu(h))/\sigma(h)) - \Phi(y)| = O(n^{-\alpha/2}(\log n)^{(1+\alpha)/2}).$$

Proof. Using a standard argument (see, e.g., [16], pp. 228–229), we have

$$\sup_{y\in\mathbb{R}} |P((\gamma_{g,n} - \mu(h))/\sigma(h)) < y) - \Phi(y)| \leq$$

$$\leq 3 \sup_{y\in\mathbb{R}} |P((U_{g,n} - \mu(h))/\sigma(h)) < y) - \Phi(y)| + \frac{2}{\pi}\epsilon_n + P(|U_{g,n} - \gamma_{g,n}| > \epsilon_n)$$

for every sequence $\{\epsilon_n\}_{n\geq 1}$ of positive real numbers. The result follows from the Berry-Esséen bound for U-statistics from [4] and from Lemma 1 by taking $K \geq \sup_{h\in\mathcal{H}} \sup_{x,y} h(x,y)$ and $\epsilon_n = C^* K n^{-\alpha/2}(\log n)^{(1+\alpha)/2}$. $\square$

Remark 2. Conditions (i), (ii) can be slightly weakened. For instance, Theorem 1 still holds if g' is bounded and piecewise continuous, provided that left and right limits exist at every discontinuity point, and Hölder condition takes place at every interval of continuity.

Remark 3. Gini's cograduation coefficient satisfies condition (ii) with $\alpha = 1$. It follows immediately from Theorem 1 that the rate of convergence in this case is $O(n^{-1/2}\log n)$. The same holds also for Spearman's footrule, which is essentially "a half" of Gini's coefficient.

3. Appendix

Proof of Lemma 1. Consider $f_{10}(x,y)$ and $f_{01}(x,y)$ as defined above. Following [6], it is possible to conclude that with probability 1

$$\gamma_{g,n} = C_n - D_n + O(1/n),$$

where

$$C_n = \int\limits_{[0,1]^2} g(|F_n(x) + G_n(y) - 1|)\, dH_n(x,y),$$
$$D_n = \int\limits_{[0,1]^2} g(|F_n(x) - G_n(y)|)\, d\, H_n(x,y),$$

and where F_n, G_n, and H_n are the marginal and the joint empirical d.f.'s.

Consider the two U-statistics $U_{C,n}$, $U_{D,n}$ defined as above. The term $R_{C,n} = \sqrt{n}(C_n - U_{C,n})$ satisfies the following inequality [6]:

$$|R_{C,n}| \leq \sqrt{n} D_{1,n} A_n + \sqrt{n} D_{2,n} B_n,$$

where

$$
\begin{aligned}
A_n &= \int\limits_{[0,1]^2} \phi_n(x, 1-y)\, dH_n(x, y), \\
B_n &= \int\limits_{[0,1]^2} \phi_n(x, y)\, dH_n(x, y), \\
D_{1,n} &= \sup_{x \in [0,1]} |F_n(x) - x|, \\
D_{2,n} &= \sup_{y \in [0,1]} |G_n(y) - y|, \\
\Delta_x &= \{u : |u| \le D_{1,n}, (x+u) \in [0,1]\}, \\
\Delta_y &= \{v : |v| \le D_{1,n}, (y+v) \in [0,1]\}, \\
\phi_n(x, 1-y) &= \sup_{(u,v) \in \Delta_x \times \Delta_y} |f_{10}(x+u, 1-y-v) - f_{10}(x, 1-y)|, \\
\phi_n(x, y) &= \sup_{(u,v) \in \Delta_x \times \Delta_y} |f_{10}(x+u, y+v) - f_{10}(x, y)|.
\end{aligned}
$$

Take now a sequence $\{\epsilon_n\}_{n \ge 1}$ of positive reals decreasing to zero, split the unit square into three regions

$$
\begin{aligned}
T_1 &= \{(x, y) : 0 \le x < 1 - 2\epsilon_n, 0 \le y \le x\}, \\
T_2 &= \{(x, y) : 0 \le y < 1 - 2\epsilon_n, 0 \le x \le y\}, \\
T_3 &= [0,1]^2 \setminus \left(T_1 \bigcup T_2\right)
\end{aligned}
$$

and define A_{in}'s, $i = 1, 2, 3$, exactly as A_n, except for the integration regions that now are T_i's. Hence $A_n = A_{1n} + A_{2n} + A_{3n}$.

Suppose now that $(x, y) \in T_1$. If $|u| < \epsilon_n$ and $|v| < \epsilon_n$, then we have

$$
\operatorname{sign}(x + y - 1) = \operatorname{sign}(x + y + u + v - 1),
$$

and

$$
|f_{10}(x+u, 1-y-v) - f_{10}(x, 1-y)| =
$$

$$
= |g'(1 - x - y - u - v) - g'(1 - x - y)| \le C_g(|u|^\alpha + |v|^\alpha).
$$

Hence, it is possible to write:

$$
A_{1n} \le C_g(D_{1,n}^\alpha + D_{2,n}^\alpha)
$$

whenever $D_{1,n} < \epsilon_n$ and $D_{2,n} < \epsilon_n$. Using DKW (Dvoretzky–Kiefer–Wolfowitz) inequality [9, 13] we get from this the chain of inequalities

$$\begin{aligned} P(A_{1n} > \epsilon_n) &\leq P(D_{1,n}^{\alpha} + D_{2,n}^{\alpha} > C_g^{-1}\epsilon_n) + 2P(D_{1,n} \geq \epsilon_n) \\ &\leq 2P\left(D_{1,n}^{\alpha} > (2C_g)^{-1}\epsilon_n)\right) + 2P(D_{1,n} \geq \epsilon_n) \\ &\leq 8(\exp\{-2n(2C_g)^{-2/\alpha}\epsilon_n^{2/\alpha}\} + \exp\{-2n\epsilon_n^2\}). \end{aligned}$$

The same arguments hold also for $P(A_{2n} > \epsilon_n)$. Furthermore, observing that $\phi_n(x,y)$ is smaller than $2C_g$ for every $(x,y) \in [0,1]^2$, it is easy to see that

$$A_{3n} \leq 2C_g \int\limits_{T_3} dH_n(x,y) \leq 8KC_g\epsilon_n + 2C_g|P_n(T_3) - P_H(T_3)|,$$

P_H being the probability measure corresponding to the density h, and P_n the corresponding empirical measure. Hence we obtain the inequality

$$P(A_{3n} > 10KC_g\epsilon_n) \leq 2\exp\{-nK^2\epsilon_n^2\}.$$

Taking into account all these estimates, it is easy to see that, for appropriate constants C_A and Q_A, independent of K, we can write

$$P(A_n > KC_A\epsilon_n) \leq Q_A \exp\left\{-2n\epsilon_n^{2/\alpha}\right\} + \exp\left\{-nK^2\epsilon_n^2\right\}.$$

From this and DKW inequality we finally get

$$\begin{aligned} &P(\sqrt{n}D_{1,n}A_n > KC_A\epsilon_n) \\ =&P(\sqrt{n}D_{1,n}A_n > KC_A\epsilon_n,\ n^{\frac{1-\alpha}{2}}(\log n)^{\alpha/2}D_{1,n} > \epsilon_n) \\ +&P(\sqrt{n}D_{1,n}A_n > KC_A\epsilon_n,\ n^{\frac{1-\alpha}{2}}(\log n)^{\alpha/2}D_{1,n} \leq \epsilon_n) \\ \leq&P\left(n^{\frac{1-\alpha}{2}}(\log n)^{\alpha/2}D_{1,n} > \epsilon_n\right) + P((n/\log n)^{\alpha/2}A_n > KC_A) \\ \leq&Q_A^*\left(\exp\left\{-2(n/\log n)^{\alpha}\epsilon_n^2\right\} + n^{-2} + \exp\left\{-K^2n^{1-\alpha}(\log n)^{\alpha}\right\}\right) \end{aligned}$$

with Q_A^* independent of K.

In the same way it can be shown that:

$$P(\sqrt{n}D_{1,n}A_n > KC_B\epsilon_n) \leq$$

$$\leq Q_B^*(\exp\{-2(n/\log n)^{\alpha}\epsilon_n^2\} + n^{-2} + \exp\{-K^2n^{1-\alpha}(\log n)^{\alpha}\})$$

with Q^*_B, C_B independent of K. It follows from these inequalities that

$$\begin{aligned}
&P(|R_{C,n}| > K(C_A + C_B)\epsilon_n)\\
\le&2(Q^*_A + Q^*_B)(\exp\{-2(n/\log n)^\alpha \epsilon_n^2\} + n^{-2})\\
+&2(Q^*_A + Q^*_B)(\exp\{-K^2 n^{1-\alpha}(\log n)^\alpha\}).
\end{aligned}$$

Finally, by applying the same technique to the term $R_{D,n} = \sqrt{n}(D_n - U_{D,n})$, we get the conclusion of lemma. □

References

1. S. Bhuchongkul, *A class of nonparametric tests for independence in bivariate populations.* — Ann. Math. Statist. **35** (1964), 138–149.
2. F. Bolthausen, *An estimate of the remainder in a combinatorial central limit theorem.* — Z. Wahrsch. Verw. Gebiete **66** (1984), 379–386.
3. E. Bolthausen and F. Götze, *The rate of convergence for multivariate sampling statistics.* — Ann. Statist. **21,** No. 4 (1993), 1692–1710.
4. H. Callaert and P. Janssen, *The Berry-Esséen theorem for U-statistics.* — Ann. Statist. **6** (1978), 417–421.
5. D. M. Cifarelli, P. L. Conti, and E. Regazzini, *On the asymptotic distribution of a general measure of monotone dependence.* — Ann. Statist. **25** (1996), 1386–1399.
6. P. L. Conti, *Asymptotic inference on a general measure of monotone dependence.* — J. Ital. Statist. Soc. **3** (1994), 213–241.
7. P. L. Conti and Ya. Yu. Nikitin, *Asymptotic efficiency of independence tests based on Gini's rank association coefficient, Spearman's footrule and their generalizations.* — Commun. in Statistics. Theory and Methods, 1999. (to appear).
8. H. E. Daniels, *The relations between measures of correlations in the universe of sample permutations.* — Biometrika **33** (1944), 129–135.
9. A. Dvoretzky, J. Kiefer, and J. Wolfowitz, *Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator.* — Ann. Math. Statist. **27** (1956), 642–669.
10. C. Gini, *L'ammontare e la composizione della ricchezza delle nazioni.* F.lli Bocca, Torino (1914).
11. M. Hallin and M. L. Puri, *Some asymptotic results for a broad class of nonparametric statistics.* — J. of Stat. Plann. and Infer. **32** (1992), 165–196.
12. M. G. Kendall, *Rank Correlation Methods.* (4th ed.), Hafner, New York (1970).
13. P. Massart, *The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality.* — Ann. Prob. **18** (1990), 1269–1283.
14. M. L. Puri and P. K. Sen, *Nonparametric Methods in Multivariate Analysis.* Wiley, New York (1971).

15. F. H. Ruymgaart, *Asymptotic normality of nonparametric tests for independence.* — Ann. Statist. **2** (1974), 892–910.
16. R. J. Serfling, *Approximation Theorems of Mathematical Statistics.* Wiley, New York (1980).
17. C. Spearman, *The proof and measurement of association between two things.* — Am. J. Psychol. **15** (1904), 72–101.
18. C. Spearman, *"Footrule" for measuring correlation.* — Brit. J. Psychol. **2** (1906), 89–108.

Department of Statistical Sciences
University of Bologna, Italy

Санкт-Петербургский
государственный
университет

Поступило 17 февраля 1999 г.