Periodica Mathematica Hungarica Vol. 8 (1), (1977), pp. 53—55

A REMARK ON THE APPROXIMATION OF THE SAMPLE DF IN THE MULTIDIMENSIONAL CASE

by

G. TUSNÁDY (Budapest)

Let $X_1, X_2, \ldots, X_n$ be i.i.d.r.v-s on the k-dimensional unit cube with

$$\mathsf{P}(X_1 < \mathbf{t}) = \lambda(\mathbf{t}) \quad \text{if} \quad \mathbf{0} \leqq \mathbf{t} \leqq \mathbf{1}, \tag{1}$$

where $\mathbf{t} = (t_1, t_2, \ldots, t_k)$, and $\mathbf{1} = (1, 1, \ldots, 1)$ are k-dimensional vectors, $\lambda(\mathbf{t}) = \prod_{i=1}^{k} t_i$ is the k-dimensional volume of the rectangle determined by the origin and the point $\mathbf{t}$, and for two vectors $\mathbf{a}$ and $\mathbf{b}$ the inequality $\mathbf{a} < \mathbf{b}$ means that each coordinate of $\mathbf{a}$ is less than the corresponding coordinate of $\mathbf{b}$. The empirical distribution function $F_n(\mathbf{t})$ based on the sample $X_1, X_2, \ldots, X_n$ is the function

$$F_n(\mathbf{t}) = \frac{1}{n} \sum_{i: X_i < \mathbf{t}} 1 \qquad \text{if} \qquad \mathbf{0} \leqq \mathbf{t} \leqq \mathbf{1}, \tag{2}$$

and the k-dimensional Brownian bridge $B(\mathbf{t})$ is defined by

$$B(\mathbf{t}) = W(\mathbf{t}) - \lambda(\mathbf{t}) W(\mathbf{1}) \quad \text{if} \quad \mathbf{0} \leqq \mathbf{t} \leqq \mathbf{1}, \tag{3}$$

where $W(\mathbf{t})$ is a k-dimensional Wiener process, i.e., $W(\mathbf{t})$ is a Gaussian process with independent increments, variance equal to the k-dimensional volume. In this remark we investigate the approximation of $F_n(\mathbf{t})$ by $B(\mathbf{t})$ in the case $k = 2$.

The one-dimensional case was investigated in [2], where we proved that there is a version of F_n and B such that

$$\mathsf{P}\left(\sup_{\mathbf{0} \leq \mathbf{t} \leq \mathbf{1}} | n(F_n(\mathbf{t}) - \lambda(\mathbf{t})) - n^{\frac{1}{2}} B(\mathbf{t}) | > C \log n + x\right) < K e^{-\lambda x} \tag{4}$$

olds for all x, where C, K, λ are positive absolute constants (Theorem 3). nvestigating the approximation of the whole sequence $\{F_n, n = 1, 2, \ldots\}$ we

AMS (MOS) subject classifications (1970). Primary 60B10; Secondary 62G10.
Key words and phrases. Invariance, embedding, testing multidimensional distributions.

got a theorem of two-dimensional character (Theorem 4). As we shall see, the result presented here is strongly connected with this latter theorem.

THEOREM. *In the case $k = 2$ for any n there is a version of F_n and B such that*

$$\mathsf{P}\left(\sup_{\mathbf{0}\leq\mathbf{t}\leq\mathbf{1}} |n(F_n(\mathbf{t}) - \lambda(\mathbf{t})) - n^{\frac{1}{2}}B(\mathbf{t})| > (C\log n + x)\log n\right) < Ke^{-\lambda x} \tag{5}$$

holds for all x, where C, K, λ are positive absolute constants.

PROOF. The proof is based on the following version of Theorem 5 of [2]:

LEMMA. *In the case $k = 1$ for any n there are measurable functions*

$$e_i(w, f) \qquad i = 1, 2, \ldots, n \tag{6}$$

defined on the product of the spaces $C(0, 1)$ and $D(0, 1)$ such that for any independent pair of the one-dimensional Wiener process W and empirical DF F_n the random variables

$$\varepsilon_i = e_i(W, F_n) \qquad\qquad i = 1, 2, \ldots, n \tag{7}$$

are i.i.d.r.v-s with distribution

$$\mathsf{P}(\varepsilon_i = 0) = \mathsf{P}(\varepsilon_i = 1) = \frac{1}{2}, \tag{8}$$

the set of random variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ is independent of F_n, and

$$\mathsf{P}\left(\sup_{\mathbf{0}\leq\mathbf{t}\leq\mathbf{1}}\left|\frac{n}{2}(\tilde{F}(\mathbf{t}) - \lambda(\mathbf{t})) - \left(\frac{n}{2}\right)^{\frac{1}{2}}W(\mathbf{t})\right| > (C\log n + x)\right) < Ke^{-\lambda x} \tag{9}$$

holds for all x, where C, K, λ are positive absolute constants, and $\tilde{F}_n$ is defined by

$$\tilde{F}_n(\mathbf{t}) = \frac{2}{n}\sum_{i=1}^{nF_n(\mathbf{t})} \varepsilon_i\,. \tag{10}$$

The proof of this lemma is similar to the proof of Theorem 5 of [2]. We shall use the conditional quantile transformation, and the dyadic scheme in the following way. Suppose $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are arbitrary i.i.d.r.v-s of distribution (8), and the set $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\}$ is independent of F_n. Let m and

$$\mathbf{0} = \mathbf{t}_0 < \mathbf{t}_1 < \ldots < \mathbf{t}_m = \mathbf{1}$$

be arbitrary and let $\tilde{F}_n$ defined by (10). The conditional distribution of $\frac{n}{2}\tilde{F}_n(\mathbf{t}_i)$ under the condition that $\{F_n(\mathbf{t}), \mathbf{0} \leq \mathbf{t} \leq \mathbf{1}\}$ and

$$\{\tilde{F}_n(\mathbf{t}_0), \tilde{F}_n(\mathbf{t}_1), \ldots, \tilde{F}_n(\mathbf{t}_{i-1}), \tilde{F}_n(\mathbf{t}_{i+1}), \ldots, \tilde{F}_n(\mathbf{t}_m)\}$$

are given is a hypergeometric distribution with parameters depending on the condition, and the conditional distribution of $\tilde{F}_n(\mathbf{1})$ given $\{F_n(\mathbf{t}), \mathbf{0} \leq \mathbf{t} \leq \mathbf{1}\}$ is a binomial distribution with parameters $\left(n, \frac{1}{2}\right)$. Hence we can transform the appropriate parts of $W(\mathbf{t})$ step by step to $\tilde{F}_n(\mathbf{1})$, $\tilde{F}_n\left(\frac{\mathbf{1}}{\mathbf{2}}\right)$, $\tilde{F}_n\left(\frac{\mathbf{1}}{\mathbf{4}}\right)$, $\tilde{F}_n\left(\frac{\mathbf{3}}{\mathbf{4}}\right)$ using in each step the conditional distribution of the new variable on the condition that $\{F_n(\mathbf{t}), \mathbf{0} \leq \mathbf{t} \leq \mathbf{1}\}$ and the just defined $\tilde{F}_n(\mathbf{t}_i)$-s are given. This is the same construction as the construction of the proof of Theorem 5, the only difference is that here $\tilde{F}_n(\mathbf{1})$ is also random variable. Hence the further details of the proof are omitted.

The theorem follows from the lemma in the same way as Theorem 4 follows from Theorem 5 in [2]. Hence its proof is also omitted.

REMARK 1. In the case $k = 1$ we proved in [2] that our result is the best possible in the sense that there are positive absolute constants A, B such that for any n and any version of F_n, B

$$\mathsf{P}\left(\sup_{\mathbf{0} \leq \mathbf{t} \leq \mathbf{1}} |n(F_n(\mathbf{t}) - \lambda(\mathbf{t})) - n^{\frac{1}{2}} B(\mathbf{t})| > A \log n\right) \geq B.$$

In the case $k = 2$ the situation is different: we do not know whether for a given $\varepsilon > 0$ are there positive constants A, B such that

$$\mathsf{P}\left(\sup_{\mathbf{0} \leq \mathbf{t} \leq \mathbf{1}} |n(F_n(\mathbf{t}) - \lambda(\mathbf{t})) - n^{\frac{1}{2}} B(\mathbf{t})| > A (\log n)^{1+\varepsilon}\right) \geq B. \tag{11}$$

REMARK 2. In case $k > 2$ the best known available results are given by CSÖRGŐ and RÉVÉSZ [1]. They give an approximation of order $n^{\frac{k-1}{2k}}$. We do not know whether (11) is true for any $k > 1$ or not.

REFERENCES

[1] M. CSÖRGŐ and P. RÉVÉSZ, A new method to prove Strassen type laws of invariance principle. I, II, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **31** (1974/75), 255–259 and 261–269. *MR* **51** # 11605a

[2] J. KOMLÓS, P. MAJOR and G. TUSNÁDY, An approximation of partial sums of independent RV's, and the sample DF. I, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **32** (1975), 111–131. *MR* **51** # 11605b

(Received October 15, 1975)

MTA MATEMATIKAI KUTATÓ INTÉZETE
H–1053 BUDAPEST
REÁLTANODA U. 13–15.
HUNGARY