A Berry-Esseen bound for U-statistics in the non-i.i.d. case

I.B. Alberink

June 30, 1998

Abstract

Let $X_1, \ldots, X_n$ be independent, not necessarily identically distributed random variables. An optimal Berry-Esseen bound is derived for U-statistics of order 2, that is, statistics of the form $T = \sum_{1 \le i < j \le n} g_{ij}(X_i, X_j)$, where the g_{ij} are measurable functions such that $\mathbb{E}|g_{ij}(X_i, X_j)| < \infty$. An application is given concerning Wilcoxon's rank-sum test.

Key words and phrases: U-statistics, Berry-Esseen bound, non-i.i.d. sampling, rate of convergence, Wilcoxon's rank-sum test.
A.M.S. classification (1991): 62E20.

1 Introduction and result

Let $X_1, \ldots, X_n$ be independent, not necessarily identically distributed random variables, taking values in an arbitrary measurable space $(\mathcal{X}, \mathcal{B})$. We suppose that $n \ge 2$ and restrict ourselves to so-called U-statistics of order 2, that is,

$$T \;=\; T(X_1, \ldots, X_n) \;=\; \sum_{1 \le i < j \le n} g_{ij}(X_i, X_j), \tag{1}$$

where the g_{ij} are certain measurable functions on $\mathcal{X}^2$. Our aim is to prove a Berry-Esseen type theorem for such statistics.

If the expectation values $\mathbb{E}|g_{ij}(X_i, X_j)|$ are finite, we can introduce

$$\begin{aligned} T_j(X_j) &:= \mathbb{E}(T|X_j) - \mathbb{E}T \qquad (2) \\ &= \sum_{i=1}^{j-1} \{\mathbb{E}(g_{ij}(X_i, X_j)|X_j) - \mathbb{E}g_{ij}(X_i, X_j)\} + \sum_{i=j+1}^{n} \{\mathbb{E}(g_{ji}(X_j, X_i)|X_j) - \mathbb{E}g_{ji}(X_j, X_i)\}, \end{aligned}$$

for $1 \le j \le n$, and

$$\begin{aligned} T_{ij}(X_i, X_j) \;&:=\; \mathbb{E}(T|X_i, X_j) - \mathbb{E}(T|X_i) - \mathbb{E}(T|X_j) + \mathbb{E}T \qquad (3) \\ &=\; g_{ij}(X_i, X_j) - \mathbb{E}(g_{ij}(X_i, X_j)|X_i) - \mathbb{E}(g_{ij}(X_i, X_j)|X_j) + \mathbb{E}g_{ij}(X_i, X_j), \end{aligned}$$

for $1 \leq i < j \leq n$. We abbreviate the $T_j(X_j)$ and $T_{ij}(X_i, X_j)$ by T_j and T_{ij} respectively. Using (2) and (3) we obtain the so-called Hoeffding decomposition of T (see, for example, Van Zwet (1984) or Bentkus, Götze, Van Zwet (1997)), given by

$$T \;=\; \mathbb{E}T + \sum_{j=1}^{n} T_j + \sum_{1 \leq i < j \leq n} T_{ij} \;=:\; \mathbb{E}T + L + U. \tag{4}$$

It consists of mutually uncorrelated random variables such that $\mathbb{E}T_i = \mathbb{E}T_{jk} = \mathbb{E}T_i T_{jk} = \mathbb{E}T_j T_k = 0$, for all i and $j < k$. Hence, in the case that $\mathbb{E}T^2 < \infty$ the sum in (4) gives us an orthogonal decomposition of T in the corresponding L^2-space.

We set

$$s_j^2 := \mathbb{E}T_j^2, \quad s^2 := \mathrm{var}(L) = \textstyle\sum_{j=1}^n s_j^2, \quad \beta_j := \mathbb{E}\,|T_j|^3, \quad \beta := \textstyle\sum_{j=1}^n \beta_j$$

and

$$\Delta^2 := \mathrm{var}(U) = \textstyle\sum_{1 \leq i < j \leq n} \mathbb{E}T_{ij}^2 .$$

Let Φ denote the standard normal distribution function and let

$$D_n \;:=\; \sup_x |\mathbb{P}(T \leq x) - \Phi(x)|.$$

Our result reads as follows:

Theorem 1. *Suppose that $\mathbb{E}T = 0$ and $s^2 = 1$. Then*

$$D_n \;\leq\; 17\,(\beta + \Delta). \tag{5}$$

We take a look at the special case in which $X, X_1, \ldots, X_n$ is an i.i.d. sample and the g_{ij} equal some symmetric function g. Then, in case $\mathbb{E}T = 0$,

$$T_j = (n-1)\,\mathbb{E}(g(X, X_j)|X_j), \quad T_{ij} = g(X_i, X_j) - \mathbb{E}(g(X_i, X)|X_i) - \mathbb{E}(g(X, X_j)|X_j),$$

and setting

$$\tilde{\beta} := \mathbb{E}\,|n^{1/2}\,T_1|^3 \quad \text{and} \quad \tilde{\Delta}^2 := \mathbb{E}\,|n^{3/2}\,T_{12}|^2$$

it is easily seen that $\beta = n^{-1/2}\tilde{\beta}$ and $\Delta^2 \leq \frac{1}{2}n^{-1}\,\tilde{\Delta}^2$. Van Zwet (1984) in this case derives, under the conditions of Theorem 1, the Berry-Esseen bound

$$D_n \leq C\,n^{-1/2}\,(\tilde{\beta} + \tilde{\Delta}^2), \tag{6}$$

with unspecified absolute constant C. Theorem 1 itself leads to

$$D_n \leq 17\,n^{-1/2}\,(\tilde{\beta} + 2^{-\frac{1}{2}}\tilde{\Delta}),$$