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

I.B. Alberink

June 30, 1998

Abstract

Let $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ 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,

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

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

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

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

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

for $1 \le i < j \le n$. 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

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

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

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

Let Φ denote the standard normal distribution function and let

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

Our result reads as follows:

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

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

We take a look at the special case in which $X, X_1, \dots, X_n$ 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 $\Delta^2 \le \frac{1}{2}n^{-1}\,\tilde{\Delta}^2$. Van Zwet (1984) in this case derives, under the conditions of Theorem 1, the Berry-Esseen bound

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

with unspecified absolute constant C. Theorem 1 itself leads to

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

in fact a somewhat better result.

Friedrich (1989) considers the more general case of an independent, not identically distributed sample, for which under the conditions of Theorem 1 the following bound arises:

$$D_n \leq c(p)(\beta + n^{5/2}\,\gamma_1\,\gamma_{3,p}^p + n^2\,\gamma_1^2\,\gamma_{3,\frac{3}{2}}), \tag{7}$$

where

$$p \in [\tfrac{3}{2}, 2), \quad \gamma_1 := \max_{1\leq j\leq n} \beta_j^{1/3} \quad \text{and} \quad \gamma_{3,p}^p := \max_{1\leq i<j\leq n} \mathbb{E}\,|T_{ij}|^p,$$

and $c(p)$ is again an unspecified constant, depending only on p. Under appropriate moment conditions this will give the desired order of $n^{-1/2}$. Notice that in the non-i.i.d. case this result is less general and precise than (5) since it contains maxima of moments.

The conditions in Theorem 1 are there merely for ease of exposition. The one thing that we really need in order to get to a Berry-Esseen type theorem for T is that its linear part has positive and finite variance, that is to say, that $0 < s^2 < \infty$. If $s^2 = 0$ the statistic is degenerate and the normal distribution will not approximate the distribution of T. In case $0 < s^2 < \infty$ we may apply Theorem 1 to the random variable $(T - \mathbb{E}T)/s$ and see that

$$\sup_x \left| \mathbb{P}\left(\frac{T - \mathbb{E}T}{s} \leq x\right) - \Phi(x)\right| \;\leq\; 17\left(\frac{\beta}{s^3} + \frac{\Delta}{s}\right). \tag{8}$$

Moreover,

$$\sup_x \left| \mathbb{P}\left(\frac{T - \mathbb{E}T}{\sigma} \leq x\right) - \Phi(x)\right| \;\leq\; 17.1\left(\frac{\beta}{\sigma^3} + \frac{\Delta}{\sigma}\right), \tag{9}$$

where $\sigma^2 := \mathrm{var}(T) = s^2 + \Delta^2$ and $0 < s^2 \leq \sigma^2 < \infty$. We will not prove that (8) implies (9): in essence this proof is already given in the Proof of Theorem 1.1 in Bentkus, Götze, Van Zwet (1997), Section 5.

2 Proof of the Theorem

Without loss of generality we may suppose that

$$n \geq 3 \quad \text{and} \quad \beta, \Delta \leq 17^{-1}.$$

Indeed, in case $n = 2$ we have that $\beta = \beta_1 + \beta_2 \geq s_1^3 + s_2^3 = s_1^3 + (1 - s_1^2)^{3/2} \geq 1/\sqrt{2}$ so $D_2 \leq 1 \leq \sqrt{2}\,\beta$. In case $\beta \geq 17^{-1}$ or $\Delta \geq 17^{-1}$ clearly $D_n \leq 1 \leq 17(\beta + \Delta)$.

We proceed like Van Zwet (1984), using in addition a very useful technique taken from Bentkus, Bloznelis, Götze (1996), so-called 'randomization'. The calculations needed in order to get precise constants are usually rather easy and therefore omitted.

Let $e\{s\} := \exp\{is\}$, for $s \in \mathbb{R}$, and let Z be a standard normally distributed random variable which is independent of the sample. We will basically use Esseen's Smoothing Lemma (see Feller (1971), section XVI.3), telling us that

$$D_n \;\le\; \frac{1}{\pi}\int_{-H}^{H} \left|\mathbb{E}\, e\{tT\} - \mathbb{E}\, e\{tZ\}\right| / |t|\, dt + \frac{24}{\pi\sqrt{2\pi}}\, H^{-1}$$

for any $H > 0$. Taking $H := \min\{\beta^{-1}, \Delta^{-1}\}$ it is easy to see that

$$D_n \;\le\; 2\pi^{-1} \int_0^H \left|\mathbb{E}\, e\{tT\} - \mathbb{E}\, e\{tZ\}\right| / t\, dt + 3.1\, H^{-1}, \tag{10}$$

which will be our point of departure. The plan of attack is now as follows. Let $h := \sqrt{H}$. Then

$$\int_0^H \left|\mathbb{E}\, e\{tT\} - \mathbb{E}\, e\{tZ\}\right| / t\, dt \le \varepsilon_1 + \varepsilon_2 + \varepsilon_3, \tag{11}$$

with

$$\begin{aligned}
\varepsilon_1 &:= \int_0^h |\mathbb{E}\, e\{tL\} - \mathbb{E}\, e\{tZ\}|/t\, dt,\\
\varepsilon_2 &:= \int_0^h |\mathbb{E}\, e\{tT\} - \mathbb{E}\, e\{tL\}|/t\, dt,\\
\varepsilon_3 &:= \int_h^H |\mathbb{E}\, e\{tT\} - \mathbb{E}\, e\{tZ\}|/t\, dt.
\end{aligned}$$

By (10) and (11), in order to prove Theorem 1 it will suffice to show that

$$\varepsilon_1 \le 1.5\,\beta, \quad \varepsilon_2 \le 3.6\,\Delta \quad \text{and} \quad \varepsilon_3 \le 0.6\,\beta + 16.1\,\Delta. \tag{12}$$

We are going to prove that this is indeed the case.

2.1 Estimation of ε_1

The first inequality in (12) can be proved fairly easily by copying the proof of Theorem 2 in Feller (1971), Section XVI.5: the main difference is that one has to take $T = \beta^{-1/2}$ instead of $T = \frac{8}{9}\beta^{-1}$.

2.2 Estimation of ε_2

The fact that $\varepsilon_2 \le 3.6\,\Delta$ we will prove rather classically, using Taylor expansions. First we need the following lemma:

Lemma 2. *Let $0 \le t \le h$ and suppose that $n \ge 3$ and $\beta \le 17^{-1}$. Then for all $1 \le i < j \le n$ we have: $|\mathbb{E}\, e\{t\sum_{k\,:\,k\ne i,j} T_k\}| \le \exp\{-c_1 t^2\}$, with $c_1 = 0.26$.*

Proof of Lemma 2. Here and in the following we will make use of Taylor expansions of functions around 0. As to them we will use the following scheme. Let $j \in \{0, 1, 2, \dots\}$: for any real or complex-valued function f on $\mathbb{R}$, possessing a continuous $(j+1)^{\text{th}}$ derivative on some open interval containing both 0 and x,

$$f(x) \;=\; \sum_{l=0}^{j} f^{(l)}(0)\,\frac{x^l}{l!} + r_{j+1}, \tag{13}$$

where

$$r_{j+1} \;:=\; \int_0^1 (1-s)^j\, f^{(j+1)}(sx)\,\frac{x^{j+1}}{j!}\,ds.$$

Introducing a random variable τ which is uniformly distributed on $(0,1)$, we may also write r_{j+1} as

$$r_{j+1} \;=\; \mathbb{E}(1-\tau)^j\, f^{(j+1)}(\tau x)\,\frac{x^{j+1}}{j!}.$$

Looking at the Taylor expansion of $e\{tT_k\}$ around 0, we see that

$$\mathbb{E}\,e\{tT_k\} \;=\; 1 - \tfrac{1}{2}s_k^2\,t^2 + \tfrac{1}{2}(it)^3\,\mathbb{E}(1-\tau)^2 e\{t\tau T_k\}T_k^3, \tag{14}$$

where τ is uniformly distributed on $(0,1)$ and independent of all other random variables. It follows from this that

$$|\mathbb{E}\,e\{tT_k\}| \;\le\; \exp\{-\tfrac{1}{2}t^2(s_k^2 - 2^{-\frac{1}{2}}\beta_k\beta^{-1/2})\}. \tag{15}$$

Indeed, in case $s_k^2 \ge 2\beta$ we have that $s_k^2 \le \beta_k^{2/3} \le \beta_k^{2/3}\, s_k(2\beta)^{-1/2} \le 2^{-\frac{1}{2}}\beta_k\beta^{-1/2}$ in which case (15) is true because $|\mathbb{E}\,e\{tT_k\}| \le 1$. Otherwise we have that $1 - \frac{1}{2}s_k^2\,t^2 \ge 0$ and (15) follows easily from (14), combined with the fact that $1 - y \le \exp\{-y\}$ for all $y \in \mathbb{R}$.

It is furthermore clear that $s_j^2 \le \beta_j^{2/3} \le \beta^{2/3} \le 17^{-2/3}$ (all j), so that

$$\begin{aligned}|\mathbb{E}\,e\{t\textstyle\sum_{k\,:\,k\neq i,j} T_k\}| \;&=\; \textstyle\prod_{k\,:\,k\neq i,j}|\mathbb{E}\,e\{tT_k\}| \;\le\; \exp\{-\tfrac{1}{2}t^2\textstyle\sum_{k\,:\,k\neq i,j}(s_k^2 - 2^{-\frac{1}{2}}\beta_k\beta^{-1/2})\}\\ &\le\; \exp\{-\tfrac{1}{2}t^2\,(1 - s_i^2 - s_j^2 - 2^{-\frac{1}{2}}\beta^{1/2})\} \;\le\; \exp\{-c_1t^2\}\,:\end{aligned}$$

this completes the proof. □

Expanding $e\{tU\}$ around 0 (see (13)) and using Lemma 2 (with τ and c_1 as in its proof) we see that

$$\begin{aligned}\varepsilon_2 \;&=\; \textstyle\int_0^h |\mathbb{E}\,e\{tL\}\,\{e\{tU\} - e\{0\}\}|/t\,dt\\ &\le\; \textstyle\int_0^h |\mathbb{E}\,e\{tL\}U|\,dt + \int_0^h t\,|\mathbb{E}\,e\{tL\}(1-\tau)e\{t\tau U\}\,U^2|\,dt\\ &\le\; \textstyle\int_0^h \sum_{i<j}|\mathbb{E}\,e\{tL\}T_{ij}|\,dt + \int_0^h \tfrac{1}{2}t\,\mathbb{E}\,U^2\,dt\\ &=\; \textstyle\int_0^h \sum_{i<j}|\mathbb{E}\,e\{t\sum_{k\,:\,k\neq i,j}T_k\}||\mathbb{E}\,e\{t(T_i+T_j)\}T_{ij}|\,dt + \tfrac{1}{4}h^2\,\Delta^2\\ &\le\; \tfrac{1}{4}\Delta + \textstyle\int_0^h \exp\{-c_1t^2\}\,\sum_{i<j}|\mathbb{E}\,e\{t(T_i+T_j)\}T_{ij}|\,dt.\end{aligned}$$

Expanding $e\{tT_k\}$ as $1 + it\,\mathbb{E}_\tau e\{t\tau T_k\}\,T_k$ (by $\mathbb{E}_\tau S$ we mean the conditional expectation of S given all random variables except τ) we see that $|\mathbb{E}\,e\{t(T_i+T_j)\}T_{ij}| \le t^2\,\mathbb{E}\,|T_iT_jT_{ij}|$ so that, using Hölder's inequality as well as the Cauchy-Schwarz inequality, we see that

$$\begin{aligned}\textstyle\sum_{i<j} |\mathbb{E}\,e\{t(T_i+T_j)\}T_{ij}| &\le t^2 \textstyle\sum_{i<j}\{\mathbb{E}\,|T_iT_j|^2\}^{1/2}\{\mathbb{E}\,T_{ij}^2\}^{1/2}\\ &\le t^2\left(\textstyle\sum_{i<j}\mathbb{E}\,|T_iT_j|^2\right)^{1/2}\left(\textstyle\sum_{i<j}\mathbb{E}\,T_{ij}^2\right)^{1/2}.\end{aligned}$$

Of course $\sum_{i<j}\mathbb{E}\,|T_iT_j|^2 \le \left(\sum_j \mathbb{E}\,T_j^2\right)^2 = 1$, so

$$\varepsilon_2 \;\le\; \tfrac{1}{4}\Delta + \left(\textstyle\int_0^\infty t^2 \,\exp\{-c_1t^2\}\,dt\right)\Delta \;\le\; 3.6\,\Delta,$$

which proves (12) for ε_2.

2.3 Estimation of ε_3

To obtain the proper bound for ε_3 we have to be more careful.

Van Zwet (1984) in the i.i.d. case treats this integral by splitting up the sample into two parts, $X_1,\dots,X_m$ versus $X_{m+1},\dots,X_n$, where m is a function of t. In this way T is split up as $L_1+L_2+U_1+U_2$, where

$$L_1 = \textstyle\sum_{j=1}^m T_j, \quad L_2 = \sum_{j=m+1}^n T_j, \quad U_2 = \sum_{m<i<j} T_{ij}$$

and $U_1 = T - (L_1+L_2+U_2)$: now U_1 is removed by a Taylor expansion and L_1 is independent of (L_2, U_2), which leads to the result which is needed.

In our case this approach is not satisfactory as in a sense it privileges the first m random variables over the remaining $n-m$. This problem is overcome by means of the above-mentioned randomization. Take any fixed $t \in [h, H]$: now let $\overline{X}_1,\dots,\overline{X}_n$ be an independent copy of $X_1,\dots,X_n$, and let $\alpha_1,\dots,\alpha_n$ be an independent sample from a Bernoulli distributed population with

$$\mathbb{P}(\alpha_j = 1) = 1 - \mathbb{P}(\alpha_j = 0) = m(t) := 12\,t^{-2}\log t$$

(note that always $m(t) \in [0,1]$), independent of all other random variables and of each other.

Instead of the sample $X_1,\dots,X_n$ we may take a so-called 'mixed' sample, by which we mean the original one with some of the X_j replaced by independent copies $\overline{X}_j$, and apply our statistic T to the new sample. As an example we look at the case in which only the first random variable is exchanged, i.e., at the sample $\overline{X}_1, X_2,\dots,X_n$. Now we get $T = T(\overline{X}_1, X_2,\dots,X_n)$, which has a Hoeffding decomposition of form

$$\hat{T}_1(\overline{X}_1) + \textstyle\sum_{j=2}^n \hat{T}_j(X_j) + \sum_{j=2}^n \hat{T}_{1j}(\overline{X}_1, X_j) + \sum_{2\le i<j\le n}\hat{T}_{ij}(X_i,X_j),$$

with the $\hat{T}_j$ and the $\hat{T}_{ij}$ defined via (2) and (3) (using the mixed sample). Though it is not clear beforehand that the random variables

$$\hat{T}_1(\overline{X}_1),\ \hat{T}_2(X_2),\ \dots,\ \hat{T}_n(X_n),\ \hat{T}_{12}(\overline{X}_1, X_2),\ \dots,\ \hat{T}_{n-1,n}(X_{n-1},X_n)$$

are identical in distribution to the original $T_1(X_1), \ldots, T_{n-1,n}(X_{n-1}, X_n)$, elementary considerations on the conditional expectations involved show that indeed they are. Because of this we can speak of $T_1(\overline{X}_1)$ or $T_{14}(\overline{X}_1, X_4)$ instead of $\hat{T}_1(\overline{X}_1)$ or $\hat{T}_{14}(\overline{X}_1, X_4)$ (in the case of a mixed sample containing both $\overline{X}_1$ and X_4).

The idea is now the following. Let

$$\tilde{T} \;=\; \tilde{T}(t) \;:=\; \tilde{L} + \tilde{U}, \tag{16}$$

with

$$\tilde{L} \;=\; \tilde{L}(t) \;=\; L_1 + L_2 \;:=\; \sum_{j=1}^{n} \alpha_j \, T_j(X_j) + \sum_{j=1}^{n} (1-\alpha_j)\, T_j(\overline{X}_j),$$

and $\tilde{U} = \tilde{U}(t) = U_1 + U_2 + U_3$ with

$$U_1 := \textstyle\sum_{i<j} \alpha_i \alpha_j \, T_{ij}(X_i, X_j), \quad U_2 := \sum_{i \neq j} \alpha_i (1-\alpha_j)\, T_{ij}(X_i, \overline{X}_j) \quad \text{and}$$
$$U_3 := \textstyle\sum_{i<j} (1-\alpha_i)(1-\alpha_j)\, T_{ij}(\overline{X}_i, \overline{X}_j)$$

(where, e.g., by $T_{43}(X_4, \overline{X}_3)$ we actually mean $T_{34}(\overline{X}_3, X_4)$). Then we have that

$$T \;\stackrel{d}{=}\; \tilde{T},$$

where by $\stackrel{d}{=}$ we mean equality in distribution, as this is so for each fixed outcome of $\alpha_1, \ldots, \alpha_n$. Hence $\mathbb{E}\, e\{t\tilde{T}\} = \mathbb{E}\, e\{tT\}$, and we may in the following consider $\mathbb{E}\, e\{t\tilde{T}\}$ instead of $\mathbb{E}\, e\{tT\}$.

Now we turn to

$$\varepsilon_3 = \textstyle\int_h^H |\mathbb{E}\, e\{t\tilde{T}\} - \mathbb{E}\, e\{tZ\}|/t \, dt.$$

First we note that

$$\textstyle\int_h^H |\mathbb{E}\, e\{tZ\}|/t \, dt \;=\; \int_h^H t^{-1} \exp\{-\frac{1}{2}t^2\} \, dt \;\leq\; \frac{1}{2} \cdot 10^{-3} \, (\beta + \Delta),$$

so we may instead turn to the estimation of $\int_h^H |\mathbb{E}\, e\{t\tilde{T}\}|/t \, dt$. Using (13), we first remove U_1 by noticing that

$$\begin{aligned} &\textstyle\int_h^H |\mathbb{E}\, e\{t(\tilde{L} + \tilde{U})\} - \mathbb{E}\, e\{t(\tilde{L} + \tilde{U} - U_1)\}|/t \, dt \\ &\qquad \leq\; \textstyle\int_h^H \mathbb{E}|U_1| \, dt \;\leq\; \int_h^H \{m(t)^2 \Delta^2\}^{1/2} \, dt \;\leq\; 7.8\, \Delta, \end{aligned}$$

so that we are only in need of an upper bound for

$$\delta := \textstyle\int_h^H |\mathbb{E}\, e\{t(L_1 + L_2 + U_2 + U_3)\}|/t \, dt.$$

Denoting the sequences $\overline{X}_1, \ldots, \overline{X}_n$ and $\alpha_1, \ldots, \alpha_n$ by $\overline{X}$ and α, we have:

$$\delta \leq \textstyle\int_h^H \mathbb{E}|\mathbb{E}\,(e\{t(L_1 + L_2 + U_2 + U_3)\} \,|\, \overline{X}, \alpha)|/t \, dt = \int_h^H \mathbb{E}|\mathbb{E}\,(e\{t(L_1 + U_2)\} \,|\, \overline{X}, \alpha)|/t \, dt.$$

Now set

$$\begin{aligned}\delta_1 &:= \int_h^H \mathbb{E}\,|\mathbb{E}\,(e\{t(L_1+U_2)\} - e\{tL_1\}\,|\,\overline{X},\alpha)|/t\,dt,\\ \delta_2 &:= \int_h^H \mathbb{E}\,|\mathbb{E}\,(e\{tL_1\}\,|\,\overline{X},\alpha)|/t\,dt \;=\; \int_h^H \mathbb{E}\,|\mathbb{E}\,(e\{tL_1\}\,|\,\alpha)|/t\,dt.\end{aligned}$$

We will verify that

$$\delta \;\le\; \delta_1+\delta_2 \;\le\; \{2\cdot 2.3\,\Delta + 3.1\,\Delta\} + \tfrac{1}{2}\,(\beta+\Delta), \tag{17}$$

which will complete the proof of (12).

As to δ_2 we need the following lemma:

Lemma 3. *Let $h \le t \le H$ and suppose that $\beta \le 17^{-1}$: then $\mathbb{E}\,|\mathbb{E}\,(e\{tL_1\}\,|\,\alpha)|^2 \le t^{-4}$.*

Proof of Lemma 3. Let $\hat{X}_1,\dots,\hat{X}_n$ be a second independent copy of $X_1,\dots,X_n$. Write $\hat{L}_1 := \sum_{j=1}^n \alpha_j T_j(\hat{X}_j)$, and denote the $T_k(X_k)$ and $T_k(\hat{X}_k)$ by T_k and $\hat{T}_k$ respectively. Then we have that

$$\mathbb{E}\,|\mathbb{E}\,(e\{tL_1\}\,|\,\alpha)|^2 = \mathbb{E}\,\mathbb{E}\,(e\{t(L_1-\hat{L}_1)\}\,|\,\alpha) = \mathbb{E}\,e\{t(L_1-\hat{L}_1)\}, \tag{18}$$

since $\hat{L}_1$ is conditionally independent of L_1.

By taking $\alpha_k(T_k-\hat{T}_k)$ instead of T_k in (14) we see that

$$\begin{aligned}\mathbb{E}\,e\{t\alpha_k(T_k-\hat{T}_k)\} \;&=\; 1-\tfrac{1}{2}t^2\,\mathbb{E}\,(\alpha_k(T_k-\hat{T}_k))^2\\ &\quad +\tfrac{1}{2}(it)^3\,\mathbb{E}\,(1-\tau)^2 e\{t\tau\alpha_k(T_k-\hat{T}_k)\}(\alpha_k(T_k-\hat{T}_k))^3,\end{aligned}$$

where τ again is uniformly distributed on $(0,1)$ and independent of all other random variables. As a result

$$\mathbb{E}\,e\{t\alpha_k(T_k-\hat{T}_k)\} \;=\; 1-\tfrac{1}{2}m(t)t^2\,2s_k^2 + R_k \tag{19}$$

with $|R_k| \le \frac{1}{2}t^3\,\mathbb{E}\,(1-\tau)^2\alpha_k\,|T_k-\hat{T}_k|^3 \le \frac{2}{3}m(t)t^2\beta^{-1}\beta_k$, where we used the fact that

$$\begin{aligned}\mathbb{E}\,|T_k-\hat{T}_k|^3 \;&\le\; \mathbb{E}\,(|T_k|+|\hat{T}_k|)(T_k^2-2T_k\hat{T}_k+\hat{T}_k^2) \;=\; \mathbb{E}\,(|T_k|+|\hat{T}_k|)(T_k^2+\hat{T}_k^2)\\ &=\; 2\,\mathbb{E}\,|T_k|(T_k^2+\hat{T}_k^2) \;=\; 2\,\{\beta_k + s_k^2\,\mathbb{E}\,|T_k|\} \;\le\; 4\beta_k.\end{aligned}$$

In case $1 \ge \frac{1}{2}m(t)t^2\,2s_k^2 = 12\,s_k^2\,\log t$ for all t, as in the Proof of Lemma 2 (19) leads to

$$|\mathbb{E}\,e\{t\alpha_k(T_k-\hat{T}_k)\}| \;\le\; \exp\{-m(t)t^2(s_k^2-\tfrac{2}{3}\beta_k\beta^{-1})\}. \tag{20}$$

This is for example the case if $1 \ge 12\,s_k^2\log\beta^{-1}$: if not then $1 < 12\,s_k^2\log\beta^{-1}$ and $s_k^2 \le s_k^2\,\{(12\log\beta^{-1})^{1/2}s_k\} \le \sqrt{12}\,\beta_k\,\log^{1/2}\beta^{-1} \le \frac{2}{3}\beta_k\beta^{-1}$ and (20) is trivially true.

As a consequence we always have the inequality (20), so that

$$\begin{aligned}|\mathbb{E}\,e\{t(L_1-\hat{L}_1)\}| \;&=\; \textstyle\prod_{k=1}^n |\mathbb{E}\,e\{t\alpha_k(T_k-\hat{T}_k)\}|\\ &\le\; \exp\{-m(t)t^2\textstyle\sum_{k=1}^n(s_k^2-\frac{2}{3}\beta_k\beta^{-1})\} \;=\; t^{-4},\end{aligned}$$

which, together with (18), finishes the proof. $\square$

With the help of Lemma 3 it is clear that

$$\delta_2 \;\le\; \int_h^H \{\mathbb{E}\,|\mathbb{E}\,(e\{tL_1\}\,|\,\alpha)|^2\}^{1/2}/t\,dt \;\le\; \tfrac{1}{2}h^{-2} \;\le\; \tfrac{1}{2}(\beta+\Delta).$$

On the other hand we have that

$$\delta_1 \;\le\; \delta_3+\delta_4 \;:=\; \int_h^H \mathbb{E}\,|\mathbb{E}\,(e\{tL_1\}U_2\,|\,\overline{X},\alpha)|\,dt + \tfrac{1}{2}\int_h^H t\,\mathbb{E}U_2^2\,dt.$$

As to δ_4, since $\mathbb{E}U_2^2 = 2m(t)(1-m(t))\Delta^2 \le 2\Delta^2\,m(t)$ we see that

$$\delta_4 \;\le\; 12\,\Delta^2\int_h^H t^{-1}\log t\,dt \;\le\; 3.1\,\Delta.$$

With regard to δ_3 we need one final randomization: let $\xi_1,\ldots,\xi_n$ be an independent sample from a Bernoulli distributed population, independent of all other random variables, with $\mathbb{P}(\xi_j=0)=\mathbb{P}(\xi_j=1)=\frac{1}{2}$ for all j. Now as concerns $\mathbb{E}\,|\mathbb{E}\,(e\{tL_1\}U_2\,|\,\overline{X},\alpha)|$ we may replace every X_j by $\xi_jX_j+(1-\xi_j)\tilde{X}_j$, where $\tilde{X}_1,\ldots,\tilde{X}_n$ is a third independent copy of $X_1,\ldots,X_n$. In this way we come to:

$$\begin{aligned}\delta_3 \;&=\; \int_h^H \mathbb{E}\,|\mathbb{E}\,(e\{t(L_{11}+L_{12})\}(U_{21}+U_{22})\,|\,\overline{X},\alpha)|\,dt\\ &\le\; \int_h^H \mathbb{E}\,|\mathbb{E}\,(e\{t(L_{11}+L_{12})\}(U_{21}+U_{22})\,|\,\overline{X},\alpha,\xi)|\,dt,\end{aligned} \tag{21}$$

with

$$\begin{aligned}&L_{11} := \textstyle\sum_{j=1}^n \alpha_j\xi_j\,T_j(X_j), \quad L_{12} := \sum_{j=1}^n \alpha_j(1-\xi_j)\,T_j(\tilde{X}_j),\\ &U_{21} := \textstyle\sum_{j=1}^n \alpha_j\xi_j\sum_{l\ne j}(1-\alpha_l)\,T_{jl}(X_j,\overline{X}_l) \quad \text{and}\\ &U_{22} := \textstyle\sum_{j=1}^n \alpha_j(1-\xi_j)\sum_{l\ne j}(1-\alpha_l)\,T_{jl}(\tilde{X}_j,\overline{X}_l).\end{aligned}$$

The inequality in (21) holds because in general $\mathbb{E}\,|\mathbb{E}\,(V\,|\,V_1,V_2)| \le \mathbb{E}\,|\mathbb{E}\,(V\,|\,V_1,V_2,V_3)|$. Of course we then have that

$$\delta_3 \;\le\; \sum_{j=1}^2\int_h^H \mathbb{E}\,|\mathbb{E}\,(e\{t(L_{11}+L_{12})\}U_{2j}\,|\,\overline{X},\alpha,\xi)|\,dt \;=:\; \delta_5+\delta_6.$$

Here

$$\begin{aligned}\delta_5 \;&=\; \int_h^H \mathbb{E}\,|\mathbb{E}\,(e\{tL_{11}\}U_{21}\,|\,\overline{X},\alpha,\xi)\mathbb{E}\,(e\{tL_{12}\}\,|\,\overline{X},\alpha,\xi)|\,dt\\ &\le\; \int_h^H \mathbb{E}\;\mathbb{E}\,(|U_{21}|\,|\,\overline{X},\alpha,\xi)\,|\mathbb{E}\,(e\{tL_{12}\}\,|\,\alpha,\xi)|\,dt.\end{aligned}$$

Hölder's inequality and (18) lead us to:

$$\begin{aligned}\delta_5 \;&\le\; \int_h^H \{\mathbb{E}\mathbb{E}^{\,2}(|U_{21}|\,|\,\overline{X},\alpha,\xi)\}^{1/2}\{\mathbb{E}\,|\mathbb{E}\,(e\{tL_{12}\}\,|\,\alpha,\xi)|^2\}^{1/2}\,dt\\ &\le\; \int_h^H \{\mathbb{E}U_{21}^2\}^{1/2}\{\mathbb{E}\,|\mathbb{E}\,(e\{tL_{12}\}\,|\,\alpha,\xi)|^2\}^{1/2}\,dt,\end{aligned}$$

where $\mathbb{E}U_{21}^2 = 2\cdot\frac{1}{2}\cdot m(t)(1-m(t))\,\Delta^2 \le m(t)\,\Delta^2$. As in the proof of Lemma 3 we see that

$$\mathbb{E}\,|\mathbb{E}\,(e\{tL_{12}\}\,|\,\alpha,\xi)|^2 \;\le\; t^{-2}$$

(the only difference being in fact that all α_k are replaced by $\alpha_k(1-\xi_k)$, which leads to an additional factor of $\frac{1}{2}$ in the exponent). We come to

$$\delta_5 \;\leq\; \int_h^H m^{1/2}(t)\,\Delta\, t^{-1}\, dt \;\leq\; 2.3\,\Delta.$$

In the same way we get exactly the same bound for δ_6. Hence it is clear that $\delta_1 \leq 2\cdot 2.3\,\Delta + 3.1\,\Delta$, which concludes the proof of (17). This in turn concludes the proof of Theorem 1.

3 Application: Wilcoxon's rank-sum test

We conclude the paper with an application of Theorem 1 to Wilcoxon's rank-sum test.

Let $Y, Y_1, \ldots, Y_m$ and $Z, Z_1, \ldots, Z_n$ be independent i.i.d. samples, with distribution functions F_Y and F_Z respectively. We assume that these two functions are continuous and we are going to test the null hypothesis that they are equal. We do this as follows: let R_j denote the rank number of Y_j among $Y_1, \ldots, Y_m, Z_1, \ldots, Z_n$, that is, let

$$R_j \;:=\; \sum_{i=1}^m I\{Y_j \geq Y_i\} + \sum_{i=1}^n I\{Y_j \geq Z_i\}.$$

Note that ties (equal outcomes) occur with probability 0 as we take the distribution functions to be continuous. Wilcoxon's rank-sum statistic is defined by:

$$T \;:=\; \sum_{j=1}^m R_j. \tag{22}$$

It is well-known (see, e.g., Gibbons, Chakraborti (1992) or Pestman (1998)) that for relatively large m and n the distribution of the statistic (22) under the null hypothesis may be approximated by a normal one. Thus T gives a way to test whether $F_Y = F_Z$, which in that case will be referred to as F. We talk about Wilcoxon's rank-sum test, which is equivalent to the also well-known Mann-Whitney test.

Corollary 4. *Let T be Wilcoxon's rank-sum statistic, as defined in (22). Under the null hypothesis we have the following result:*

$$\sup_x \left| \mathbb{P}\left(\frac{T - \frac{1}{2}m(m+n+1)}{\sqrt{\frac{1}{12}mn(m+n)}} \leq x \right) - \Phi(x) \right| \;\leq\; \frac{17}{\sqrt{m+n}} \left\{ \tfrac{3}{4}\sqrt{3}\,\frac{m^2+n^2}{\sqrt{mn}\,(m+n)} + 1 \right\}.$$

Proof of Corollary 4. Let $(X_1, \ldots, X_{m+n}) := (Y_1, \ldots, Y_m, Z_1, \ldots, Z_n)$: under the null hypothesis this constitutes an i.i.d. sample. Therefore we may write T in the form (1), by just taking

$$g_{ij}(x,y) \;:=\; I\{i \leq m < j\} \left(I\{x \geq y\} + \frac{m+1}{2n} \right).$$

Setting $\xi_j := \mathbb{E}(I\{Y_j \geq Z\} \mid Y_j) - \frac{1}{2}$, for $j = 1, \ldots, m$, and $\eta_k := \mathbb{E}(I\{Y \geq Z_k\} \mid Z_k) - \frac{1}{2}$, for $k = 1, \ldots, n$, we see that

$$T_j \;=\; I\{1 \leq j \leq m\}\, n\, \xi_j + I\{m < j \leq m+n\}\, m\, \eta_{j-m}$$

for all j. Furthermore, for all $1 \leq i < j \leq n$ we have that

$$T_{ij} = I\{i \leq m < j\} \left(I\{Y_i \geq Z_{j-m}\} - \mathbb{E}(I\{Y_i \geq Z\} \mid Y_i) - \mathbb{E}(I\{Y \geq Z_{j-m}\} \mid Z_{j-m}) + \tfrac{1}{2} \right).$$

Now we can determine s^2, β and Δ^2. First we note that

$$\mathbb{E}\, |\xi_1|^2 \;=\; \mathbb{E}\, \left|\mathbb{E}(I\{F(Y_1) \geq F(Z)\} \mid Y_1) - \tfrac{1}{2}\right|^2 \;=\; \mathbb{E}\, \left|F(Y_1) - \tfrac{1}{2}\right|^2 \;=\; \tfrac{1}{12},$$

since $F(Z)$ and $F(Y_1)$ are independent and uniformly distributed on $(0,1)$. Of course then also $\mathbb{E}\,|\xi_j|^2 = \frac{1}{12}$ for all j. In the same way it is seen that always $\mathbb{E}\,|\eta_k|^2 = \frac{1}{12}$ and $\mathbb{E}\,|\xi_j|^3 = \mathbb{E}\,|\eta_k|^3 = \mathbb{E}\,|F(Y_1) - \frac{1}{2}|^3 = \frac{1}{32}$, and that, for all $1 \leq i \leq m < j \leq n$, we have:

$$\mathbb{E}\, \left|I\{Y_i \geq Z_{j-m}\} - \mathbb{E}(I\{Y_i \geq Z\} \mid Y_i) - \mathbb{E}(I\{Y \geq Z_{j-m}\} \mid Z_{j-m}) + \tfrac{1}{2}\right|^2 \;=\; \tfrac{1}{12}.$$

It turns out that $\mathbb{E}T = \frac{1}{2}m(m+n+1)$,

$$s^2 = \tfrac{1}{12}mn(m+n), \quad \beta = \tfrac{1}{32}mn(m^2+n^2) \quad \text{and} \quad \Delta^2 = \tfrac{1}{12}mn.$$

Corollary 4 now is an easy consequence of Theorem 1. $\square$

In case $\min\{m,n\}/\max\{m,n\} \geq c$ for some $c > 0$ and all m and n, Corollary 4 gives us a Berry-Esseen bound. In the case that, for example, always $m = 1$, the bound is useless, as was to be expected. Note that Van Zwet's result (6) is not applicable here, since Wilcoxon's rank-sum statistic is non-symmetric. Friedrich's result (7) *is* applicable but gives a result which is more complicated.

References

[1] V. Bentkus, M. Bloznelis, F. Götze, A Berry-Esséen Bound for Student's Statistic in the Non-I.I.D. Case, 1996, *Journal of Theoretical Probability 9, p.765-796.*

[2] V. Bentkus, F. Götze and W.R. van Zwet, An Edgeworth expansion for symmetric statistics, 1997, *Ann. Statist. 25, p.851-896.*

[3] W. Feller, *An Introduction to Probability Theory and Its Applications. Vol. II*, 2nd Ed., Wiley, 1971, New York.

[4] K.O. Friedrich, A Berry-Esseen bound for functions of independent random variables, 1989, *Ann. Statist. 17, p.170-183.*

[5] J.D. Gibbons, S. Chakraborti, *Nonparametric statistical inference*, 3rd Ed., Marcel Dekker, Inc., 1992, New York.

[6] R. Helmers, W.R. van Zwet, The Berry-Esseen bound for U-statistics, *Statistical Decision Theory and Related Topics III*, Vol. 1 (S.S. Gupta and J.O. Berger, eds.), p.497-512, Academic Press, 1982, New York.

[7] W.R. Pestman, *Mathematical Statistics, An Introduction*, Walter de Gruyter, 1998, Berlin.

[8] W.R. van Zwet, A Berry-Esseen bound for symmetric statistics, 1984, *Z. Wahrsch. verw. Gebiete 66, p.425-440.*

Ivo Alberink
Department of Mathematics
University of Nijmegen
Postbus 9010
6500 GL, Nijmegen
The Netherlands
email: alberink@sci.kun.nl