stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE STATISTIEK SW 91/82 DECEMBER
(DEPARTMENT OF MATHEMATICAL STATISTICS)

R. HELMERS

THE BERRY-ESSEEN BOUND FOR STUDENTIZED U-STATISTICS

Preprint

kruislaan 413 1098 SJ amsterdam

Printed at the Mathematical Centre, 413 Kruislaan, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

1980 Mathematics Subject classification: Primary 60 F 05

Secondary 62 E 20

The Berry-Esseen bound for studentized U-statistics*)

by

R. Helmers

ABSTRACT

Callaert and Veraverbeke (1981) recently obtained a Berry-Esseen type bound of order $n^{-\frac{1}{2}}$ for Studentized non degenerate U-statistics of degree two. The condition these authors need to obtain this order bound is the finiteness of the 4.5th absolute moment of the kernel h. In this note it is shown that this assumption can be weakened to that of a finite $(4+\varepsilon)$th absolute moment of the kernel h, for some $\varepsilon > 0$. Our proof resembles part of Helmers and van Zwet (1982) where an analogous result is obtained for the Student t-statistic. The present note extends this to Studentized U-statistics.

KEY WORDS & PHRASES : *Berry-Esseen bound, Studentized U-statistic, Student t-statistic, jackknifing, rate of convergence*

*) This report will be submitted for publication elsewhere.

Let $X_1, X_2, \ldots, X_n$, $n\geq 2$ be i.i.d. random variables with common distribution function F. Let h(x,y) be a realvalued function, symmetric in its arguments, and with $Eh(X_1,X_2)= \nu$. Define a U-statistic by

$$(1) \qquad U_n = \binom{n}{2}^{-1} \sum_{1\leq i<j\leq n} h(X_i, X_j)$$

and suppose that $g(X_1) = E[h(X_1, X_2)- \nu|X_1]$ has a positive variance σ_g^2.
Let

$$S_n^2 = 4(n-1)(n-2)^{-2} \sum_{i=1}^n [(n-1)^{-1} \sum_{\substack{j=1\\ j\neq i}}^{n} h(X_i, X_j)-U_n]^2$$

and note that $n^{-1} S_n^2$ is the jackknife estimator of the variance of U_n; i.e. S_n^2 is the sample variance of the "pseudo-values" $nU_n-(n-1)U_{n-1}^i$, where

$$U_{n-1}^i = \binom{n-1}{2}^{-1} \sum_{\substack{1\leq j<k\leq n\\ j\neq i,\, k\neq i}} h(X_j, X_k),$$

for $i=1, 2, \ldots, n$.

THEOREM. *If* $E|h(X_1, X_2)|^{4+\varepsilon}<\infty$, *for some* $\varepsilon>0$, *and* $\sigma_g^2 > 0$ *then, for* $n\to\infty$

$$(2) \quad \sup_x |P(\{n^{\frac{1}{2}}S_n^{-1} (U_n-\nu) \leq x\}) - \Phi(x)|= O(n^{-\frac{1}{2}})$$

Callaert and Veraverbeke (1981) proved the theorem for the special case $\varepsilon=\frac{1}{2}$. The purpose of this note is to show that the theorem is also valid in its present form. Our proof will rely heavily on the proof given by Callaert and Veraverbeke. However, to deal with the part of their proof which required the full force of their 4.5th absolute moment assumption we will modify their proof and employ the following lemma to obtain a sharper result.

LEMMA *Let*

$$(3) \quad V_n = \binom{n}{2}^{-1} \sum_{1\leq i<j\leq n} h_n (X_i, X_j)$$

be a U-*statistic with a varying kernel* h_n *of the form*

$$(4) \quad h_n = \alpha+ n^{-1}\beta$$

where α *and* β *are symmetric functions of its two arguments with* $E\alpha(X_1,X_2)=\nu$ *and* $E\beta(X_1, X_2)=0$. *Suppose that* $\gamma(X_1)=E[\alpha(X_1, X_2)-\nu|X_1]$ *has a positive variance* σ_γ^2. If $E|\gamma(X_1)|^3 < \infty$ *and, for some* $\eta > 0$,

$$(5) \quad E|\alpha(X_1, X_2)|^{\frac{5}{3} + \eta}< \infty, \quad E|\beta(X_1, X_2)|^{1 + \eta}< \infty$$

then, for $n \to \infty$

(6) $\sup_x |P(\{\tau_n^{-1}(V_n^n - \nu) \le x\}) - \Phi(x)| = O(n^{-\frac{1}{2}})$

where $\tau_n^2 = 4\, n^{-1}\sigma_\gamma^2$

PROOF. The lemma is a simple consequence of theorem 4.1 of [2]. □

PROOF OF THE THEOREM. As in [1] we write

(7) $$\frac{n^{\frac{1}{2}}(U_n-\nu)}{S_n} = \frac{n^{\frac{1}{2}}(U_n-\nu)}{2\sigma_g}\, 2\,\sigma_g\, S_n^{-1}$$

and establish a stochastic expansion for $2\sigma_g\, S_n^{-1}$. Using nothing more than the finiteness of $E|h(X_1, X_2)|^{4+\varepsilon}$, for some $\varepsilon > 0$, it is proved in [1] that

(8) $$2\,\sigma_g\, S_n^{-1} = 1 - \frac{1}{8}\,\sigma_g^{-2}\, n^{-1}\, \Sigma_{i=1}^n f(X_i) + R_n$$

where the function f is given by

(9) $$f(x) = 4(g^2(x) - \sigma_g^2) + 8\int_{-\infty}^{\infty} g(y)\,(h(x,y)-g(x)-g(y))\, dF(y)$$

for real x and R_n is a remainder term which is of order $n^{-\frac{1}{2}}(\ln n)^{-1}$, except on a set with probability $O(n^{-\frac{1}{2}})$, as $n \to \infty$.

It follows directly from (7) and (8) (cf [1], page 197) that

(10) $$P(\{|n^{\frac{1}{2}}(U_n-\nu)\, R_n| \ge 2\sigma_g\, n^{-\frac{1}{2}}\})$$
$$\le P(\{|R_n| \ge n^{-\frac{1}{2}}(\ln n)^{-1}\}) + P(\{|n^{\frac{1}{2}}(U_n-\nu)| \ge 2\sigma_g \ln n\}) = O(n^{-\frac{1}{2}})$$

where we have applied the lemma (with α=h and β=0) to obtain the orderbound in the last line. As in [1], (7), (8) and (10) together imply that it suffices now to establish a Berry-Esseen bound for

(11) $$W_n = 2^{-1}\,\sigma_g^{-1}\, n^{\frac{1}{2}}(U_n-\nu)\,(1 - \tfrac{1}{8}\,\sigma_g^{-2} n^{-1} \sum_{i=1}^n f(X_i))$$

instead of obtaining such a bound for $n^{\frac{1}{2}} S_n^{-1}(U_n-\nu)$.

By slightly modifying the decomposition of W_n employed in [1] we write

(12) $$W_n = W_{n1} + W_{n2}$$

where $2\sigma_g n^{-\frac{1}{2}} W_{n1} + \nu$ is a U-statistic with varying kernel h_n of the form V_n (cf(3)) with $h_n = \alpha + n^{-1}\beta$ where α and β are given by

(13) $$\alpha(x,y) = h(x,y) - \frac{1}{8}\,\sigma_g^{-2}\,(g(x)\, f(y) + g(y)\, f(x))$$

and

(14) $$\beta(x,y) = -\frac{1}{8}\,\sigma_g^{-2}\,((h(x,y)-\nu)(f(x)+f(y)) - 2(g(x)\, f(y) + g(y) f(x)) - 2\mu)$$

with $\mu = \int_{-\infty} g(x) f(x)\, d\,F(x)$ and where W_{n2} is a remainder term satisfying

$E\, W_{n2} = O(n^{-\frac{1}{2}})$ and

(15) $P(\{|W_{n2}-EW_{n2}| \geq n^{-\frac{1}{2}}\}) = O(n^{-\frac{1}{2}})$

We note in passing that W_{n1} and W_{n2} are precisely equal to the terms

$\frac{n^{\frac{1}{2}}}{2\sigma_g} U_n^* + Z_{n1} - EZ_{n1} + Z_{n2}$ and $EZ_{n1} + Z_{n3}$ in [1] which together form the decomposition of W_n employed in that paper. The orderbound (15) was proved in [1] requiring $\sigma_g^2 > 0$ and the finiteness of $Eh^4(X_1, X_2)$. Thus W_{n2} is also of negligible order of magnitude under our present assumptions. It remains to consider W_{n1}. The statistic $2\sigma_g\, n^{-\frac{1}{2}} W_{n1} + \nu$ is a U-statistic of the form V_n (cf (3)) with varying kernel $h_n = \alpha + n^{-1}\beta$ where α and β are given by (13) and (14) and satisfy the requirements $E\alpha(X_1, X_2)=\nu$ and $E\beta(X_1, X_2) = 0$. It follows that, if the assumptions of the lemma are satisfied, we have the Berry-Esseen bound

(16) $\sup_x |P(\{W_{n1} \leq x\}) - \Phi(x)| = O(n^{-\frac{1}{2}})$.

To check the assumptions needed for (16) we note first that in this case $\gamma\,(X_1)=E\,[\alpha(X_1, X_2)- \nu\,|X_1] = E[h(X_1, X_2) - \nu\;|X_1] = g(X_1)$ and an application of Jensen's inequality for conditional expectations yields $E|g(X_1)|^3 \leq E|h(X_1, X_2)-\nu|^3 < \infty$, so that the assumptions $\sigma_\gamma^2 > 0$ and $E|\gamma(X_1)|^3 < \infty$ of the lemma are clearly satisfied. Secondly we verify assumption (5) of the lemma. By the independence of X_1 and X_2, the c_r-inequality, and the relations (13) and (14) we see that it suffices to show that the $(\frac{5}{3} + \eta)$th absolute moments of $h(X_1, X_2)$, $g(X_1)$ and $f(X_1)$ and the $(1+\eta)$ th absolute moment of $h(X_1, X_2)$. $f(X_1)$ are all finite, for some $\eta > 0$. In view of the remark following (16) we need only to consider the last two of these moments. Application of Schwarz inequality, the c_r-inequality and relation (9) easily leads to the requirements $E(g(X_1))^{4+4\eta} < \infty$, $E(h(X_1, X_2))^{2+2\eta} < \infty$. Jensen's inequality for conditional expectations can be applied once more to find that we only need $Eh(X_1,X_2)^{4+4\eta} < \infty$ to guarantee this. As $\eta > 0$ is arbitrary, the proof of (16) is now complete. Combining (16) with (15), the remark preceeding (15) and the argument leading to (11) completes the proof of the theorem. □

REMARKS

1. The idea behind the present modification of the proof given by Callaert & Veraverbeke (1981) is that by applying the Berry-Esseen bound (6) to W_{n1} we implicitly use rather delicate characteristic functions methods, whereas in Callaert & Veraverbeke (1981) crude momentbounds are employed to deal with part of W_{n1}. As a consequence it is possible to relax their 4.5th absolute momentassumption - which Callaert & Veraverbeke (1981) really need only in their treatment of the W_{n1}-term - to that of a finite $(4+\varepsilon)$th absolute moment for the kernel h, for some $\varepsilon > 0$.
2. If we take $h(x,y) = \frac{1}{2}(x+y)$ the statistic $n^{\frac{1}{2}}S_n^{-1}(U_n-\nu)$ reduces to the one-sample Student t-statistic. For this very special case the theorem was proved in Helmers and van Zwet (1982) in a similar fashion. Note, however, that in this case W_{n1} simplifies, whereas W_{n2} becomes even non random so that relation (15) is superfluous. The theorem yields the rate $n^{-\frac{1}{2}}$ for the accuracy of the normal approximation for Student's t, provided $0<E|X_1|^{4+\varepsilon}<\infty$, for some $\varepsilon > 0$, whereas Callaert and Veraverbeke (1981) need a finite and positive 4.5th absolute moment for F to prove this.

REFERENCES

[1] H. Callaert and N. Veraverbeke (1981), The order of the normal approximation for a Studentized U-statistic, *Ann. Statist.* vol.9, no.1, 194-200.

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