The Annals of Statistics
1978, Vol. 6, No. 2, 417–421

THE BERRY–ESSEEN THEOREM FOR U-STATISTICS

BY HERMAN CALLAERT AND PAUL JANSSEN

Limburgs Universitair Centrum, Belgium

Assuming only the existence of the third absolute moment we prove that $\sup_x |P(\sigma_n^{-1}U_n \leqq x) - \Phi(x)| \leqq C\nu_3\sigma_g^{-3}n^{-\frac{1}{2}}$ where U_n is a U-statistic. This concludes a series of investigations on the Berry–Esseen theorem for U-statistics by Grams and Serfling, Bickel, and Chan and Wierman.

1. Introduction. Let $X_1, X_2, \cdots, X_n$, $n \geqq 2$, be i.i.d. random variables with common distribution function F. Define a U-statistic by $U_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} h(X_i, X_j)$ where h is a symmetric function of two variables with $Eh(X_1, X_2) = 0$ and such that $g(X_1) = E(h(X_1, X_2) \mid X_1)$ has a positive variance σ_g^2. It then follows from Hoeffding (1948) that the distribution function (df) of $\sigma_n^{-1}U_n$ converges for $n \to \infty$ to the standard normal df Φ under the sole condition of the existence of $Eh^2(X_1, X_2)$. A study of the rate of this convergence started in 1973 with a paper by Grams and Serfling showing that $\sup_x |P(\sigma_n^{-1}U_n \leqq x) - \Phi(x)|$ is of the order $O(n^{-r/(2r+1)})$, $n \to \infty$, when $Eh^{2r} < \infty$, leading to $O(n^{-\frac{1}{2}+\varepsilon})$, $\varepsilon > 0$, when h has finite moments of all orders. An order bound of exactly $O(n^{-\frac{1}{2}})$ was found by Bickel (1974) assuming U-statistics with bounded kernels h. Chan and Wierman (1977) succeeded in weakening considerably the assumptions of the previous theorems obtaining the order bound of $O(n^{-\frac{1}{2}})$ when the fourth moment exists and $O(n^{-\frac{1}{2}} \log^{\frac{1}{2}} n)$ for kernels h with finite third absolute moment. We now prove that $O(n^{-\frac{1}{2}})$ can be attained requiring only the existence of the third absolute moment which is a natural assumption for a Berry–Esseen theorem. With only some easy additional computations it will be shown that $\sup_x |P(\sigma_n^{-1}U_n \leqq x) - \Phi(x)| \leqq C\nu_3\sigma_g^{-3}n^{-\frac{1}{2}}$ for all $n \geqq 2$, indicating a nice analogy with the classical Berry–Esseen theorem.

The notation of this paper mainly adheres to that of Chan and Wierman and also the technique used in the first part of the proof is based on their work.

2. The Berry–Esseen bound. Let $\hat{U}_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} (g(X_i) + g(X_j))$ be the projection of U_n. Note that $E\hat{U}_n = 0$ and put $\hat{\sigma}_n^2 = E\hat{U}_n^2$ which equals $4\sigma_g^2 n^{-1}$. Denote $(U_n - \hat{U}_n)\hat{\sigma}_n^{-1}$ by Δ_n and split it up into two parts Δ_n' and Δ_n'' such that $\Delta_n' = \binom{n}{2}^{-1}\hat{\sigma}_n^{-1} \sum_{1 \leq i < j \leq c_n} Y_{ij}$ with $Y_{ij} = h(X_i, X_j) - g(X_i) - g(X_j)$. The quantity c_n will be determined in the course of the proof and plays an essential role in obtaining various order bounds for several terms to be estimated.

THEOREM. *Let $U_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} h(X_i, X_j)$ be a U-statistic such that $g(X_1) = E(h(X_1, X_2) \mid X_1)$ has positive variance σ_g^2. If $\nu_3 = E|h(X_1, X_2)|^3 < \infty$ then there exists*

Received March 1977; revised June 1977.
AMS 1970 *subject classification.* Primary 60F05.
Key words and phrases. Berry–Esseen bound, U-statistics.

an absolute constant C such that for all $n \geqq 2$

$$\sup_x |P(\sigma_n^{-1}U_n \leqq x) - \Phi(x)| \leqq C\nu_3\sigma_g^{-3}n^{-\frac{1}{2}} . \tag{1}$$

PROOF. First note that

$$\sigma_g^2 = Eg^2(X_1) = Eh(X_1, X_2)h(X_1, X_3) \leqq Eh^2(X_1, X_2) \leqq \nu_3^{\frac{2}{3}} < \infty ,$$

so that $\nu_3\sigma_g^{-3} \geqq 1$. Also $E|g(X_1)|^3 \leqq \nu_3$. Let $S_n = \hat{\sigma}_n^{-1}\hat{U}_n = n^{-\frac{1}{2}}\sigma_g^{-1}\sum_{i=1}^n g(X_i)$ which is a standardized sum of i.i.d. random variables with finite third absolute moment. Note that $\hat{\sigma}_n^{-1}U_n = (S_n + \Delta_n') + \Delta_n''$ and write

$$\begin{aligned}\sup_x |P(\hat{\sigma}_n^{-1}U_n \leqq x) - \Phi(x)| \\ \leqq \sup_x |P(S_n + \Delta_n' \leqq x) - \Phi(x)| + P(|\Delta_n''| \geqq a_n) + O(a_n) .\end{aligned} \tag{2}$$

Then, with φ_X the characteristic function of X,

$$\begin{aligned}\int_0^{\varepsilon n^{\frac{1}{2}}} t^{-1}|e^{-t^2/2} - \varphi_{S_n+\Delta_n'}(t)|\,dt \\ \leqq \int_0^{\varepsilon n^{\frac{1}{2}}} t^{-1}|e^{-t^2/2} - \varphi_{S_n}(t)|\,dt + \int_0^{\varepsilon n^{\frac{1}{2}}} t^{-1}|\varphi_{S_n}(t) - \varphi_{S_n+\Delta_n'}(t)|\,dt .\end{aligned} \tag{3}$$

From the proof of the classical Berry–Esseen theorem it follows that

$$\int_0^{\varepsilon n^{\frac{1}{2}}} t^{-1}|e^{-t^2/2} - \varphi_{S_n}(t)|\,dt \leqq C_1 E|g(X_1)|^3\sigma_g^{-3}n^{-\frac{1}{2}} \leqq C_1\nu_3\sigma_g^{-3}n^{-\frac{1}{2}} \tag{4}$$

for $\varepsilon = \sigma_g^3/E|g(X_1)|^3$. Throughout the C_k are absolute constants.

We now start the estimation of the last integral in (3). Writing η for the characteristic function $\varphi_{g(X_1)}$ we have $|\eta(\vartheta)| \leqq \exp(-\frac{1}{3}\vartheta^2\sigma_g^2)$ for $|\vartheta| \leqq \varepsilon\sigma_g^{-1}$ and ε as above. Also note that $E(f(X_i)Y_{ij}) = 0$ for any bounded Borel-measurable function f. Then $|\varphi_{S_n}(t) - \varphi_{S_n+\Delta_n'}(t)| = |Ee^{itS_n}(1 - e^{it\Delta_n'})| \leqq |Ee^{itS_n}it\Delta_n'| + \frac{1}{2}t^2E\Delta_n'^2$. Now, for $0 \leqq t \leqq \varepsilon n^{\frac{1}{2}}$ and $n \geqq 4$,

$$\begin{aligned}|Ee^{itS_n}\Delta_n'| &= \tfrac{1}{2}\sigma_g^{-1}n^{\frac{1}{2}}\binom{n}{2}^{-1}\Big|\sum_{1\leqq i<j\leqq c_n} E\Big(\exp\Big[itn^{-\frac{1}{2}}\sigma_g^{-1}\sum_{k\neq i,j} g(X_k)\Big]\Big) \\ &\qquad \times E(\exp[itn^{-\frac{1}{2}}\sigma_g^{-1}[g(X_i) + g(X_j)]]Y_{ij})\Big| \\ &= \tfrac{1}{2}\sigma_g^{-1}n^{\frac{1}{2}}\binom{n}{2}^{-1}|\eta(n^{-\frac{1}{2}}\sigma_g^{-1}t)|^{n-2} \\ &\qquad \times \Big|\sum_{1\leqq i<j\leqq c_n} E(\exp[itn^{-\frac{1}{2}}\sigma_g^{-1}g(X_i)] - 1)(\exp[itn^{-\frac{1}{2}}\sigma_g^{-1}g(X_j)] - 1)Y_{ij}\Big| \\ &\leqq \tfrac{1}{2}\sigma_g^{-3}n^{-\frac{1}{2}}t^2e^{-t^2/6}E|g(X_i)g(X_j)Y_{ij}| \\ &\leqq \tfrac{3}{2}\nu_3\sigma_g^{-3}n^{-\frac{1}{2}}t^2e^{-t^2/6} .\end{aligned}$$

Hence, for $d_n \leqq \varepsilon n^{\frac{1}{2}}$,

$$\begin{aligned}\int_0^{d_n} t^{-1}|Ee^{itS_n}it\Delta_n'|\,dt &\leqq \tfrac{3}{2}\nu_3\sigma_g^{-3}n^{-\frac{1}{2}}\int_0^{d_n} t^2e^{-t^2/6}\,dt \\ &\leqq \tfrac{9}{4}(6\pi)^{\frac{1}{2}}\nu_3\sigma_g^{-3}n^{-\frac{1}{2}} .\end{aligned} \tag{5}$$

Remark that the last estimate, which is also valid for $n = 2, 3$ as may be seen by direct computations, is independent of the choice of d_n. Also c_n does not play any particular role at this moment. Further, since $E\Delta_n'^2 \leqq Eh^2(X_1, X_2)\sigma_g^{-2}n^{-1} \leqq \nu_3\sigma_g^{-3}n^{-1}$, we have

$$\tfrac{1}{2}E\Delta_n'^2\int_0^{d_n} t\,dt \leqq \tfrac{1}{4}\nu_3\sigma_g^{-3}n^{-1}d_n^2 \tag{6}$$

which for $d_n = n^{\frac{1}{4}}$ yields the upper bound $\frac{1}{4}\nu_3\sigma_g^{-3}n^{-\frac{1}{2}}$. The estimates (5) and (6)

are sufficient for all n such that $\varepsilon n^{\frac{1}{2}} \leqq n^{\frac{1}{2}}$. If $d_n < \varepsilon n^{\frac{1}{2}}$ we write

$$\begin{aligned}|Ee^{itS_n}(1 - e^{it\Delta_n'})| &\\ &\leqq |E\exp[itn^{-\frac{1}{2}}\sigma_g^{-1}\textstyle\sum_{k>c_n} g(X_k)]|E|\exp[itn^{-\frac{1}{2}}\sigma_g^{-1}\textstyle\sum_{k\leqq c_n} g(X_k)](1 - e^{it\Delta_n'})| \\ &\leqq tE|\Delta_n'|\,|\eta(n^{-\frac{1}{2}}\sigma_g^{-1}t)|^{n-c_n} \\ &\leqq t\nu_3\sigma_g^{-3}n^{-\frac{1}{2}}\exp(-(n - c_n)n^{-1}t^2/3)\,.\end{aligned}$$

Hence

$$\int_{d_n}^{\varepsilon n^{\frac{1}{2}}} t^{-1}|\varphi_{S_n}(t) - \varphi_{S_n+\Delta_n'}(t)|\,dt \leqq \nu_3\sigma_g^{-3}n^{-\frac{1}{2}}\int_{d_n}^{\varepsilon n^{\frac{1}{2}}}\exp(-(n - c_n)n^{-1}t^2/3)\,dt\,. \tag{7}$$

Here the flexibility of a choice for d_n and c_n already becomes clear. As an example we could take $d_n = n^{\frac{1}{2}}$ and $c_n = [n - 3n^{\frac{1}{2}}\log n]$ which is, in the Chan and Wierman paper, a crucial choice for obtaining the overall order bound of $O(n^{-\frac{1}{2}}\log^{\frac{1}{2}} n)$. The estimate in (7) then becomes $C_2\nu_3\sigma_g^{-3}n^{-\frac{1}{2}}$ but it is easily seen that many other d_n and c_n provide an analogous bound. From the classical Berry–Esseen argument, together with (4), (5), (6), (7) and a suitable choice of d_n and c_n we obtain

$$\sup_x |P(S_n + \Delta_n' \leqq x) - \Phi(x)| \leqq C_3\nu_3\sigma_g^{-3}n^{-\frac{1}{2}}$$

yielding the desired result for the first term on the right-hand side of (2). For an estimate of the other terms in (2) we use the following

LEMMA. *With the notations and the assumptions as in the above theorem one has*

$$E|\Delta_n''|^3 \leqq C_4\nu_3\sigma_g^{-3}(n - c_n)^{\frac{3}{2}}n^{-3} \qquad \text{for } n = 2, 3, \cdots. \tag{8}$$

We postpone the proof of the lemma to the end of this paper.

It now follows from the Markov inequality that

$$P(|\Delta_n''| \geqq a_n) \leqq C_4\nu_3\sigma_g^{-3}n^{-3}(n - c_n)^{\frac{3}{2}}a_n^{-3} \tag{9}$$

and again we have a lot of freedom in choosing a_n and c_n. One systematic way consists in taking

$$a_n = [n^{-3}(n - c_n)^{\frac{3}{2}}]^{\frac{1}{3}} \tag{10}$$

and then choosing c_n such that $a_n \leqq C_5 n^{-\frac{1}{2}}$. This yields

$$P(|\Delta_n''| \geqq \nu_3\sigma_g^{-3}a_n) \leqq P(|\Delta_n''| \geqq a_n) \leqq C_4\nu_3\sigma_g^{-3}a_n \tag{11}$$

which is sufficient for obtaining the estimate $C_6\nu_3\sigma_g^{-3}n^{-\frac{1}{2}}$ for the last two terms in (2). We finally note that $\sigma_n^2 = \binom{n}{2}^{-1}(Eh^2(X_1, X_2) + 2(n - 2)\sigma_g^2)$, and hence for $n \geqq 2$

$$\left|\frac{\sigma_n}{\hat{\sigma}_n} - 1\right| \leqq \left|\frac{\sigma_n^2}{\hat{\sigma}_n^2} - 1\right| \leqq \frac{1}{n - 1}\left(1 + \frac{Eh^2(X_1, X_2)}{2\sigma_g^2}\right) \leqq 3\nu_3\sigma_g^{-3}n^{-1}$$

completing the proof of the theorem.

3. Some remarks on the quantities a_n, c_n, and d_n. If one only wants the order bound $O(n^{-\frac{1}{2}})$ in the Berry–Esseen theorem, the above proof could be somewhat

simplified by making in advance a suitable choice of a_n, c_n and d_n and then, without further discussion, showing that (1) holds. We preferred not to follow this way for the following reason. If one is interested in some numerical value for the constant C in (1) the possibility of altering a_n, c_n and d_n within some suitable ranges may provide easier and sharper numerical bounds on the various C_k appearing in the proof. In fact, within our framework, one is able to distinguish between dominant terms yielding $O(n^{-\frac{1}{2}})$ and terms which are $o(n^{-\frac{1}{2}})$ for $n \to \infty$.

To make this explicit we determine the ranges for the quantities under consideration. From (7) and (10) it follows that $n - c_n$ may vary from $O(n^{\frac{1}{2}} \log n)$ to $O(n^{\frac{3}{4}})$ while d_n is at most $O(n^{\frac{1}{2}})$ by (6). Taking $n - c_n = n^{\frac{3}{4}}$ in (7) we find $O(n^{\frac{1}{4}}(\log n)^{\frac{1}{2}})$ as a lower bound for d_n. Finally a_n is related to c_n by (10). It now becomes clear that apart from the estimate in (4) which obviously cannot be smaller than $O(n^{-\frac{1}{2}})$, the only dominant term appears in (5). All other estimates relevant in the proof of (1) can be made $o(n^{-\frac{1}{2}})$ by staying away from the endpoints of the ranges indicated above. This makes it possible to write (1) as

$$\sup_x |P(\sigma_n^{-1} U_n \leqq x) - \Phi(x)| \leqq C_1' \nu_3 \sigma_g^{-3} n^{-\frac{1}{2}} + C_2' o(n^{-\frac{1}{2}})$$

where C_1' is considerably smaller than C_2'.

PROOF OF THE LEMMA. Define ξ_j by

$$\binom{n}{2} \hat{\sigma}_n \Delta_n'' = \sum_{j=c_n+1}^{n} \sum_{i=1}^{j-1} Y_{ij} = \sum_{j=c_n+1}^{n} \xi_j \,. \tag{12}$$

We have $\xi_1 = 0$ and $E(\xi_{j+1} \mid \xi_1, \cdots, \xi_j) = 0$ a.s. for $j = 1, 2, \cdots$. Hence the ξ_j are martingale summands and, by optional skipping, $V_k = \sum_{j=c_n+1}^{c_n+k} \xi_j$ forms a martingale, $k = 1, 2, \cdots, n - c_n$. Applying a theorem of Dharmadhikari, Fabian and Jogdeo (1968) we get for $k = n - c_n$

$$E|V_{n-c_n}|^3 \leqq 2^{12}(n - c_n)^{\frac{3}{2}} \max_{c_n+1 \leqq j \leqq n} E|\xi_j|^3 \,. \tag{13}$$

But for fixed $j \geqq c_n + 1$, $W_k = \sum_{i=1}^{k} Y_{ij}$, $k = 1, 2, \cdots, j - 1$, is also a martingale and the same argument yields for $j = 2, 3, \cdots$

$$E|\xi_j|^3 = E|W_{j-1}|^3 \leqq 2^{12}(j - 1)^{\frac{3}{2}} \max_{1 \leqq i \leqq j-1} E|Y_{ij}|^3 \leqq 2^{12} 3^3 (j - 1)^{\frac{3}{2}} \nu_3 \,. \tag{14}$$

Then, from (12), (13) and (14)

$$E|\binom{n}{2} \hat{\sigma}_n \Delta_n''|^3 \leqq 2^{24} 3^3 (n - c_n)^{\frac{3}{2}} (n - 1)^{\frac{3}{2}} \nu_3$$

and hence

$$E|\Delta_n''|^3 \leqq 2^{24} 3^4 (n - c_n)^{\frac{3}{2}} n^{-3} \nu_3 \sigma_g^{-3} \qquad \text{for } n = 2, 3, \cdots,$$

which proves the lemma.

4. The c-sample case. Although we proved the Berry–Esseen theorem only for a one-sample U-statistic of order two, the result remains valid for the general case of multisample U-statistics of arbitrary order provided that the minimum sample size tends to infinity. The proof is based on the same ideas as used in this paper but becomes computationally more involved and will not be given

here. For more details we refer to the Ph. D. thesis to be completed by Paul Janssen.

NOTE. In the proof of the theorem the quantity Δ_n'' has been separated from the U-statistic and handled by the Markov inequality. An alternative procedure consists in writing $\Delta_n = \Delta_{n1} + \Delta_{n2}$, then making a Taylor-expansion of $Ee^{it\Delta_{n2}}$ and using an independence argument. This method will be displayed in a forthcoming paper on the Edgeworth expansion for U-statistics, co-authored by N. Veraverbeke and the authors.

Acknowledgment. The authors thank Professor W. R. van Zwet for the suggestion of showing the dependence of the order bound on F and h explicitly.

REFERENCES

[1] BICKEL ,P. J. (1974). Edgeworth expansions in nonparametric statistics. *Ann. Statist.* **2** 1–20.

[2] CALLAERT, H., JANSSEN, P. and VERAVERBEKE, N. (1977). Edgeworth expansion for U-statistics. Unpublished manuscript.

[3] CHAN, Y. and WIERMAN, J. (1977). On the Berry–Esseen theorem for U-statistics. *Ann. Probability* **5** 136–139.

[4] DHARMADHIKARI, S. W., FABIAN, V. and JOGDEO, K. (1968). Bounds on the moments of martingales. *Ann. Math. Statist.* **39** 1719–1723.

[5] GRAMS, W. F. and SERFLING, R. J. (1973). Convergence rates for U-statistics and related statistics. *Ann. Statist.* **1** 153–160.

[6] HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** 293–325.

DEPARTMENT OF MATHEMATICS
LIMBURGS UNIVERSITAIR CENTRUM
B-3610 DIEPENBEEK, BELGIUM