The Annals of Probability
1977, Vol. 5, No. 1, 136–139

ON THE BERRY–ESSEEN THEOREM FOR U-STATISTICS

By Y.-K. Chan[1] and John Wierman

University of Washington

Assuming the existence of fourth moment only, we prove the Berry–Esseen theorem for U-statistics. Assuming the third absolute moment, we obtain the order bound $O(n^{-\frac{1}{2}}\log^{\frac{1}{2}} n)$. This improves earlier results of Bickel, and Grams and Serfling.

Let $X_1, \cdots, X_n$ be independent random variables with the same distribution function F. Let h be a symmetric function of two variables such that $h(X_1, X_2)$ has mean 0, and such that $E(h(X_1, X_2) \mid X_1)$ has a positive variance. Introduce, as in [3], the U-statistic $U \equiv U_n = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} h(X_i, X_j)$. The subscript n for U_n and those random variables defined later will be suppressed throughout this paper. Let σ^2 denote the variance of U. Let Φ denote the standard normal distribution. We prove the following theorem.

Theorem. *If $h(X_1, X_2)$ has finite third absolute moment, then* $\sup_x |P(\sigma^{-1}U \leqq x) - \Phi(x)| = O(n^{-\frac{1}{2}}\log^{\frac{1}{2}} n)$ *as* $n \to \infty$. *If $h(X_1, X_2)$ has finite fourth moment, then* $\sup_x |P(\sigma^{-1}U \leqq x) - \Phi(x)| = O(n^{-\frac{1}{2}})$.

Bickel [1] obtained the order bound $O(n^{-\frac{1}{2}})$ with the assumption that h is bounded. Grams and Serfling [2] have the order bound $O(n^{-\frac{1}{2}+\varepsilon})$ ($\varepsilon > 0$ arbitrary) assuming the existence of all moments for $h(X_1, X_2)$. The key step in the present proof is the consideration of $S + \Delta'$ below. The rest is a combination of the techniques used in [1] and [2].

First, assume that $h(X_1, X_2)$ has finite third moment. Introduce the function $g(x) = \int h(x, y)\, dF(y)$. Thus $E(h(X_1, X_2) \mid X_1) = g(X_1)$, and therefore $g(X_1)$ has mean 0 and positive variance σ_g^2. Let $\hat{U} \equiv 2n^{-1} \sum_{i=1}^{n} g(X_i) = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (g(X_i) + g(X_j))$, the projection of U. Then $\hat{U}$ also has a positive variance $\hat{\sigma}^2$. Simple computations show

$$\hat{\sigma}^2 = 4n^{-1}\sigma_g^2 \qquad \text{and}$$

$$\sigma^2 = \binom{n}{2}^{-2}\left\{\binom{n}{2} Eh^2(X_1, X_2) + n(n-1)(n-2)\sigma_g^2\right\}.$$

For each n let $I = [n - 3n^{\frac{1}{2}} \log n]$. Define the random variables

$$S \equiv \hat{U}/\hat{\sigma} = 2n^{-1}\hat{\sigma}^{-1} \sum_{i=1}^{n} g(X_i),$$

$$Y_{ij} \equiv h(X_i, X_j) - g(X_i) - g(X_j), \qquad (1 \leqq i < j \leqq n)$$

$$\Delta \equiv (U - \hat{U})/\hat{\sigma} = \binom{n}{2}^{-1}\hat{\sigma}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Y_{ij},$$

$$\Delta' \equiv \binom{n}{2}^{-1}\hat{\sigma}^{-1} \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} Y_{ij},$$

$$\Delta'' \equiv \Delta - \Delta' = \binom{n}{2}^{-1}\hat{\sigma}^{-1} \sum_{i=1}^{n-1} \sum_{j=(I \vee i)+1}^{n} Y_{ij}.$$

Received December 1, 1975.

[1] Research partially supported by NSF Grant MPS75-09084.

AMS 1970 *subject classification*. Primary 60F07.

Key words and phrases. Berry–Esseen bounds, U-statistics.

Note that Y_{ij} $(1 \leqq i < j \leqq n)$ are uncorrelated and have mean 0. Let φ_X stand for the characteristic function for a random variable X. From the ordinary Berry–Esseen theorem, we have

$$\int_0^{\varepsilon_1 n^{\frac{1}{2}}} t^{-1}|e^{-t^2/2} - \varphi_S(t)|\, dt = O(n^{-\frac{1}{2}})$$

where ε_1 is a positive constant independent of n. We will obtain a similar order bound for $\varphi_{S+\Delta'}$. Write η for the characteristic function $\varphi_{g(X_1)}$. There exists $\varepsilon \in (0, \varepsilon_1)$ so small that $|\eta(\theta)| \leqq 1 - \theta^2\sigma_g{}^2/3 \leqq \exp(-\theta^2\sigma_g{}^2/3)$ if $|\theta| \leqq \varepsilon/\sigma_g$. In the following assume n is so large that $n^{\frac{1}{4}} > \varepsilon^{-1}$ and $n - 2 > n/2$. We estimate

$$\begin{aligned}
&\int_0^{n^{\frac{1}{4}}} t^{-1}|\varphi_S(t) - \varphi_{S+\Delta'}(t)|\, dt \\
&\quad = \int_0^{n^{\frac{1}{4}}} t^{-1}|Ee^{itS}(1 - e^{it\Delta'})|\, dt \\
&\quad \leqq \int_0^{n^{\frac{1}{4}}} [t^{-1}|Ee^{itS}it\Delta'| + t^{-1}E|1 + it\Delta' - e^{it\Delta'}|]\, dt \\
&\quad \leqq \binom{n}{2}^{-1}\hat{\sigma}^{-1} \textstyle\sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \int_0^{n^{\frac{1}{4}}} |Ee^{itS}Y_{ij}|\, dt + \int_0^{n^{\frac{1}{4}}} t^{-1}E(t^2\Delta'^2)\, dt \\
&\quad \leqq \hat{\sigma}^{-1} \int_0^{n^{\frac{1}{4}}} |\eta^{n-2}(2n^{-1}\hat{\sigma}^{-1}t)| \times |Ee^{2in^{-1}\hat{\sigma}^{-1}t(g(X_1)+g(X_2))}Y_{12}|\, dt + n^{\frac{1}{2}}E\Delta'^2 \\
&\quad \leqq \hat{\sigma}^{-1} \int_0^{n^{\frac{1}{4}}} \{\exp(n-2)(-4n^{-2}\hat{\sigma}^{-2}t^2\sigma_g{}^2/3)\} \\
&\qquad\quad \times |Ee^{2in^{-1}\hat{\sigma}^{-1}t(g(X_1)+g(X_2))}Y_{12}|\, dt + O(n^{-\frac{1}{2}}) \\
&\quad \leqq \hat{\sigma}^{-1} \int_0^{n^{\frac{1}{4}}} e^{-t^2/6}|Ee^{2in^{-1}\hat{\sigma}^{-1}t(g(X_1)+g(X_2))}Y_{12}|\, dt + O(n^{-\frac{1}{2}}) \\
&\quad = \hat{\sigma}^{-1} \int_0^{n^{\frac{1}{4}}} e^{-t^2/6}|E[e^{2in^{-1}\hat{\sigma}^{-1}t(g(X_1)+g(X_2))} - 1 \\
&\qquad\quad - 2in^{-1}\hat{\sigma}^{-1}t(g(X_1) + g(X_2))]Y_{12}|\, dt + O(n^{-\frac{1}{2}}) \\
&\quad \leqq \hat{\sigma}^{-1} \int_0^{n^{\frac{1}{4}}} e^{-t^2/6}E|4n^{-2}\hat{\sigma}^{-2}t^2(g(X_1) + g(X_2))^2Y_{12}|\, dt + O(n^{-\frac{1}{2}}) \\
&\quad \leqq 4n^{-2}\hat{\sigma}^{-3}E|(g(X_1) + g(X_2))^2Y_{12}| \int_0^\infty e^{-t^2/6}t^2\, dt + O(n^{-\frac{1}{2}}) \\
&\quad = O(n^{-\frac{1}{2}})\,.
\end{aligned}$$

(In the second equality above we used the fact that $g(X_1) + g(X_2)$ and Y_{12} are uncorrelated.) On the other hand

$$\begin{aligned}
&\int_{n^{\frac{1}{4}}}^{\varepsilon n^{\frac{1}{2}}} t^{-1}|\varphi_S(t) - \varphi_{S+\Delta'}(t)|\, dt \\
&\quad = \int_{n^{\frac{1}{4}}}^{\varepsilon n^{\frac{1}{2}}} t^{-1}|Ee^{itS}(1 - e^{it\Delta'})|\, dt \\
&\quad = \int_{n^{\frac{1}{4}}}^{\varepsilon n^{\frac{1}{2}}} t^{-1}|Ee^{2in^{-1}\hat{\sigma}^{-1}t(g(X_{I+1})+\cdots+g(X_n))}Ee^{2in^{-1}\hat{\sigma}^{-1}t(g(X_1)+\cdots+g(X_I))}(1 - e^{it\Delta'})|\, dt \\
&\quad \leqq \int_{n^{\frac{1}{4}}}^{\varepsilon n^{\frac{1}{2}}} t^{-1}|\eta^{n-I}(2n^{-1}\hat{\sigma}^{-1}t)| \times 2\, dt \\
&\quad \leqq \int_{n^{\frac{1}{4}}}^{\varepsilon n^{2}} 2t^{-1}\exp[(n - I)(-4n^{-2}\hat{\sigma}^{-2}t^2\sigma_g{}^2/3)]\, dt \\
&\quad \leqq \int_{n^{\frac{1}{4}}}^{\varepsilon n^{\frac{1}{2}}} 2t^{-1}\exp[(3n^{\frac{1}{2}}\log n)(-4n^{-2}\hat{\sigma}^{-2}t^2\sigma_g{}^2/3)]\, dt \\
&\quad = \int_{n^{\frac{1}{4}}}^{\varepsilon n^{\frac{1}{2}}} 2t^{-1}n^{-1}\, dt \\
&\quad = O(n^{-\frac{1}{2}})\,.
\end{aligned}$$

Combining, we have $\int_0^{\varepsilon n^{\frac{1}{2}}} t^{-1}|e^{-t^2/2} - \varphi_{S+\Delta'}(t)|\, dt = O(n^{-\frac{1}{2}})$, and the usual Berry–Esseen argument yields

$$\sup_x |P(S + \Delta' \leqq x) - \Phi(x)| = O(n^{-\frac{1}{2}})\,.$$

Therefore, for any choice of the constants a_n, an elementary calculation gives

$$(*) \qquad \sup_x |P(S + \Delta \leqq x) - \Phi(x)| \leqq O(n^{-\frac{1}{2}}) + P(|\Delta - \Delta'| \geqq a_n) + O(a_n)\,.$$

If we let $a_n = n^{-\frac{1}{2}} \log^{\frac{1}{2}} n$, we have

$$\begin{aligned} P(|\Delta - \Delta'| \geqq a_n) &= P(|\Delta''| \geqq a_n) \leqq a_n^{-2} E\Delta_n''^2 \\ &= a_n^{-2}\tbinom{n}{2}^{-2}\hat{\sigma}^{-2} \textstyle\sum_{i=1}^{n-1} \sum_{j=(I\vee i)+1}^{n} EY_{ij}^2 \\ &\leqq a_n^{-2}\tbinom{n}{2}^{-2}\hat{\sigma}^{-2}(n \times 3n^{\frac{1}{2}} \log n) EY_{12}^2 \\ &= O(n^{-\frac{1}{2}} \log^{\frac{1}{2}} n) . \end{aligned}$$

Combining, we see that

$$\begin{aligned} \sup_x |P(U/\hat{\sigma} \leqq x) - \Phi(x)| &= \sup_x |P(S + \Delta \leqq x) - \Phi(x)| \\ &= O(n^{-\frac{1}{2}} \log^{\frac{1}{2}} n) . \end{aligned}$$

This, together with the observation that $|1 - \hat{\sigma}/\sigma| = O(n^{-1})$, implies

$$\sup_x |P(U/\sigma \leqq x) - \Phi(x)| = O(n^{-\frac{1}{2}} \log^{\frac{1}{2}} n) .$$

The first assertion of the theorem is proved.

Now assume that $h(X_1, X_2)$, and therefore Y_{12}, has finite fourth moment. Then, if we let $a_n = n^{-\frac{1}{2}}$, we have

$$\begin{aligned} P(|\Delta - \Delta'| \geqq a_n) &= P(|\Delta''| \geqq a_n) \leqq a_n^{-4} E\Delta_n''^4 \\ &= n^2\tbinom{n}{2}^{-4}\hat{\sigma}^{-4} \textstyle\sum\sum\sum\sum EY_{i_1j_1}Y_{i_2j_2}Y_{i_3j_3}Y_{i_4j_4} , \end{aligned}$$

where in the summation the subscript i_k $(k = 1, 2, 3, 4)$ runs through $1, \cdots, n-1$, the subscript j_k runs through $(I \vee i_k) + 1, \cdots, n$ $(k = 1, 2, 3, 4)$. Many terms in the summation vanish. More precisely, since Y_{ij} is a function only of X_i and X_j, we see that a summand in the above sum vanishes unless every subscript occurs at least twice. Consider a nonzero summand. Suppose that it contains four distinct subscripts $\alpha, \beta, \gamma, \delta$. At least two of $\alpha, \beta, \gamma, \delta$ must be each equal to some of j_1, j_2, j_3, j_4. Hence there are most $\binom{4}{2}n^2(n - I)^2$ way to select the quadruple $(\alpha, \beta, \gamma, \delta)$. Suppose next that a nonzero summand contains 3 distinct subscripts α, β, γ. At least one of α, β, γ must each be equal to some j_1, j_2, j_3, j_4. Hence there are at most $\binom{3}{1}n^2(n - I)$ ways to select the triple (α, β, γ). Evidently there are at most $2^8n^2 + n$ summands which contain fewer than three distinct subscripts. Combining, we see that the number of nonzero terms in the above sum is bounded by a constant multiple of $n^2(n - I)^2$. Consequently, with $a_n = n^{-\frac{1}{2}}$, we have

$$\begin{aligned} P(|\Delta - \Delta'| \geqq a_n) &\leqq Cn^2\tbinom{n}{2}^{-4}\hat{\sigma}^{-4}n^2(n - I)^2 \\ &\leqq Cn^2\tbinom{n}{2}^{-4}4^{-2}n^2\sigma_g^{-4}n^2(1 + 3n^{\frac{1}{2}} \log n)^2 \\ &= O(n^{-\frac{1}{2}}) . \end{aligned}$$

(Here C is some constant dependent on $\varphi_{h(X_1,X_2)}$ but not on n.) This and (*) yields

$$\sup_x |P(S + \Delta \leqq x) - \Phi(x)| = O(n^{-\frac{1}{2}}) .$$

As $|1 - \hat{\sigma}/\sigma| = O(n^{-1})$, this again implies

$$\sup_x |P(U/\sigma \leqq x) - \Phi(x)| = O(n^{-\frac{1}{2}}) .$$

Although stated in terms of a one-sample U-statistic of order two, the theorem of this paper may be extended to the general case of multisample U-statistics of arbitrary order, provided only that the minimum sample size tends to infinity.

Consider a c-sample U-statistic with sample sizes $n_1, n_2, n_3, \cdots, n_c$, and corresponding blocks of $m_1, m_2, \cdots, m_c$ arguments in the kernel. Letting n denote the minimum sample size, define $I = [n - 3n^{\frac{1}{2}} \log n]$. The difference $\Delta = (U - \hat{U})/\hat{\sigma}$ is partitioned into Δ' and $\Delta'' = \Delta - \Delta'$ with Δ' having the form

$$\Delta' = [\prod_{i=1}^{c} \binom{n_i}{m_i}]^{-1} \hat{\sigma}^{-1} \sum Y_{i_{11} i_{12} \cdots i_{cm_c}},$$

where the sum is over all combinations $i_{11}, i_{12}, \cdots, i_{cm_c}$ for which all indices from the smallest sample are less than or equal to I.

The Berry–Esseen theorem for independent nonidentically distributed random variables is applied to $S = \hat{U}/\hat{\sigma}$. $S + \Delta'$ is then handled by Fourier analysis, using the fact that Δ' is independent of the last $n - I$ observations from the smallest sample. The counting arguments used for the number of terms of nonzero expectation in Δ'^2, Δ''^2, and Δ''^4 are more involved than those presented. Moment bounds found then handle Δ'' by the methods of Grams and Serfling, to obtain the rate of convergence for $U/\hat{\sigma}$. The equation allowing replacement of $\hat{\sigma}$ by σ is obtained from Hájek's projection lemma and the moment bound of Grams and Serfling, which combine to show $|\sigma^2 - \hat{\sigma}^2| = O(n^{-2})$. For details, see [4].

REFERENCES

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

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

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

[4] WIERMAN, J. (1975). First passage percolation, a Berry–Esseen theorem for U-statistics, and optimal stopping. Ph. D. thesis, Univ. of Washington.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF WASHINGTON
SEATLE, WASHINGTON 98195

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MINNESOTA 55455