King Fahd University of Petroleum & Minerals

DEPARTMENT OF MATHEMATICAL SCIENCES

Technical Report Series

TR 040

May 1982

On the Order of Normal Approximation of a Studentized U-Statistic

Ibrahim A. Ahmad

DHAHRAN 31261 ● SAUDI ARABIA ● www.kfupm.edu.sa/math/ ● E-mail: mathdept@kfupm.edu.sa

ON THE ORDER OF NORMAL APPROXIMATION OF A STUDENTIZED U-STATISTIC*

By

Ibrahim A. Ahmad
Department of Mathematical Sciences
University of Petroleum and Minerals
Dhahran, Saudi Arabia

Abstract

A recent result of Callaert and Veraverbeke (1981, Ann. Statist., 9, 194-200) is proved by a different and simpler technique that results in weakening the required moment condition. A random version of this result is also given providing a studentized analogue of a result of Ahmad (1980, Ann. Statist., 8, 1395-1398).

AMS 1980 Subject Classification: Primary: 60F05 Secondary: 62E20.

Keywords and phrases: Order of normal approximation, studentized U-statistics, random indices, Galton-Watson process.

*The author is indebted to H. Callert and N. Veraverbeke for allowing him to see their paper prior to publication. This research is supported by a grant from the University of Petroleum & Minerals.

1. Introduction. Let $\{X_n\}_{n=2}^{\infty}$ be a sequence of i.i.d. random variables with common distribution function F. Let $h(x, y)$ be a real-valued symmetric function of two variables such that $Eh(X_1, X_2) = \theta$. Define a U-statistics by:

$$(1.1) \qquad U_n = \binom{n}{2}^{-1} \Sigma_{1 \le i < j \le n} h(X_i, X_j),$$

and assume that $g(X_1) = E[h(X_1, X_2)|X_1]$ has a positive variance σ_g^2.

Hoeffding (1948) proved that the asymptotic distribution of $(U_n - \theta) / (\text{Var } U_n)^{\frac{1}{2}}$ is standard normal while more recently, Callaert and Janssen (1978) found out that if $E|h(X_1, X_2)|^3 < \infty$, then the order in the normal approximation is $O(n^{-\frac{1}{2}})$. On the other hand, Berk (1966) and independently Hoeffding (1961) proved that U_n converges to θ with probability one as $n \to \infty$. Rates of Convergence in this almost sure law as well as other results have been investigated in Ahmad (1981).

Next, if $\{N_n\}$ is a sequence of integer-valued random variables independent of the X_n's such that $EN_n = m_n$ and $\text{Var } N_n = m_n^{(2)}$, then a random-sample size U-statistic is defined by:

$$U_{N_n} = \binom{N_n}{2}^{-1} \Sigma_{1 \le i < j \le N_n} h(X_i, X_j).$$

The order of normal approximation of U_{N_n} has been given in Ahmad (1980) and is of the order $m_n^{-\frac{1}{2}} + \sqrt{m_n^{(2)}}/m_n + (m_n^{(2)}/m_n)^{\frac{1}{2}}$, see (1.5) of Ahmad (1980). This approximation finds application, among other possibilities, in deriving approximate bounds for the super critical Gulton-Watson branching process.

The purpose of the present note is two-fold, first is to give a simpler proof of the result of Callaert and Veraverbeke (1981) concerning

a studentized version of the U-statistic, viz., $\sqrt{n}\,(U_n - \theta)/S_n$, where

$$(1.3) \qquad S_n^2 = 4(n-1)\,(n-2)^{-2} \sum_{i=1}^{n} [V_n(X_i) - U_n]^2,$$

where $V_n(X_i) = (n-1)^{-1} \sum_{j \neq i}^{n} h(X_i, X_j)$. Our method of proof while keeps the basic idea of Callaert and Veraverbeke about decomposing S_n^2 intact it applies a simple device that not only facilitates the proof but also reduces the moments condition. Secondly, a random-indices version of this studentized U-statistic is studied and the order of normal approximation is derived for $\sqrt{N_n}(U_{N_n} - \theta)/S_{N_n}$ and the result is then noted how it could be used to obtain order of approximation in the super-critical Galton-Watson processes with unknown variance.

2. Main results.

THEOREM 1. *If* $Eh^4(X_1, X_2) < \infty$ *and if* $\sigma_g^2 > 0$, *then as* $n \to \infty$,

$$(2.1) \qquad \sup_x \left| P[\sqrt{n}(U_n - \theta) \leq xS_n] - \Phi(x) \right| = O(n^{-\frac{1}{2}}),$$

where

$$(2.2) \qquad S_n^2 = 4(n-1)\,(n-2)^{-2} \sum_{i=1}^{n} [V_n(X_i) - U_n]^2,$$

with $V_n(X_i) = (n-1)^{-1} \sum_{j \neq i} h(X_i, X_j)$.

NOTE: Callart and Veraverbeke (1981) proved the above result assuming that $E|h(X_1, X_2)|^{9/2} < \infty$, thus our Theorem 1 is an improvement.

PROOF. Instead of decomposing S_n^{-1} as in Callaert and Veraverbeke (1981) we proceed using the following simple and well-known device (cf. Lemma 1 of Michel and Pfanzagl (1971)): If $\{\xi_n\}$ and $\{\eta_n\}$ are two sequences of random variables, then for any sequence of positive real numbers $\{a_n\}$,

$$(2.3)\qquad \sup|P[\xi_n \le x\eta_n] - \Phi(x)| \le \sup_x|P[\xi_n \le x] - \Phi(x)| + P[|\eta_n - 1| \ge a_n] + O(a_n).$$

Applying this device we have

$$(2.4)\qquad \sup_x|P[\sqrt{n}\,(U_n - \theta) \le xS_n] - \Phi(x)|$$

$$\le \sup_x|P[\sqrt{n}(U_n - \theta) \le x(2\sigma_g)] - \Phi(x)| + P[|\frac{S_n}{2\sigma_g} - 1| \ge a_n] + O(a_n).$$

But it follows that the first term in the right-hand-side of (2.4) is $O(n^{-\frac{1}{2}})$, cf. Callaert and Janssen (1978). Thus we need only to deal with the other two terms. Let us choose $a_n = n^{-\frac{1}{2}}$, thus the last term is also $O(n^{-\frac{1}{2}})$. All that remains is to deal with the middle term. But

$$(2.5)\qquad P[|\frac{S_n}{2\sigma_g} - 1| \ge a_n] \le P[|\frac{S_n^2}{4\sigma_g^2} - 1| \ge a_n] = P[|S_n^2 - 4\sigma_g^2| \ge \epsilon_n],$$

where $\epsilon_n = 4\sigma_g^2\, a_n$. Using the decomposition of Callaert and Veraverbeke (1981) of $S_n^2 - 4\sigma_g^2$ we have

$$(2.6)\quad S_n^2 - 4\sigma_g^2 = n^{-1}\Sigma_{i=1}^n [4(g^2(X_i) - \sigma_g^2) + 8\tilde{g}(X_i)] - 4\binom{n}{2}^{-1}\Sigma_{i<j}\,[(g(X_i) + g(X_j))\cdot$$

$$\psi(X_i, X_j) - \tilde{g}(X_i) - \tilde{g}(X_j)] - 8n^{-1}\Sigma_{i=1}^n [g(X_i)\,\binom{n-1}{2}^{-1}\,\Sigma_{k<m}^{(i)}\,\psi(X_k, X_m)]$$

$$+4(n-2)^{-1}\Sigma_{i=1}^n [\binom{n-1}{2}^{-1}\Sigma_{k<m}^{(i)}\psi(X_i, X_k)\,\psi(X_i, X_m) - 4n(n-1)\,(n-2)^{-2}\cdot$$

$$[\binom{n}{2}^{-1}\Sigma_{i<j}\psi(X_i, X_j)]^2 + 4n(n-2)^{-2}\,\binom{n}{2}^{-1}\,\Sigma_{i<j}\,\psi^2(X_i, X_j)$$

$$= T_n + \sum_{\ell=1}^{6} R_{\ell n}, \quad \text{say.}$$

where $\psi(X_1, X_2) = h(X_1, X_2) - \theta - g(X_1) - g(X_2)$ is the orthogonal complement of $h(X_1, X_2)$, and $\tilde{g}(x) = \int g(y)\, h(x, y)\, dF(y)$. Hence

$$(2.7)\qquad P[|S_n^2 - 4\sigma_g^2| \ge \epsilon_n] \le P[|T_n| \ge \epsilon_n/7] + \sum_{\ell=1}^{6} P[|R_{\ell n}| \ge \epsilon_n/7],$$

where $\varepsilon_n = 4\sigma_g^2 a_n = 4\sigma_g^2 n^{-\frac{1}{2}}$. But in Callaert and Veraverbeke (1981) they show that if $0 < Eh^4(X_1, X_2) < \infty$, then $E(R_{\ell n}^2) = O(n^{-2})$, $\ell = 1, \ldots, 6$. Then for all $n \geq 1$, $\ell = 1, \ldots, 6$,

$$(2.8) \qquad P[|R_{\ell n}| \geq \tfrac{4}{7}\sigma_g^2 n^{-\frac{1}{2}}] \leq E(R_{\ell n})^2/(\tfrac{4}{7}\sigma_g^2 n^{-\frac{1}{2}}) = O(n^{-1}).$$

All that is left out is to evaluate $P[|T_n| \geq \varepsilon_n/7]$. First, note that by the c_r-inequality (Loeve(1963) p. 155),

$$(2.9) \qquad \mathrm{Var}\,[(g^2(X_1) - \sigma_g^2) + \tilde{g}(X_1)] \leq 2\,[\mathrm{Var}(g^2(X_1) - \sigma_g^2) + \mathrm{Var}\,\tilde{g}(X_1)].$$

$$\leq 4(E\,g^4(X_1) + \sigma_g^4) + E\,\tilde{g}^2(X_1).$$

But from Lemma 1 of Callaert and Veraverbeke (1981) if $Eh^4(X_1, X_2) < \infty$, then $Eg^4(X_1) < \infty$ and since obviously also $E\tilde{g}^2(X_1) < \infty$, we see that

$$(2.10) \qquad \mathrm{Var}\,[(g^2(X_1) - \sigma_g^2) + \tilde{g}(X_1)] < \infty.$$

Hence, as in Callaert and Veraverbeke (1981) we take n large enough to consider $P[T_n^2 > \varepsilon_n/7]$.

$$(2.11) \qquad P[|T_n|^2 \geq \varepsilon_n/7] = P[|T_n|^2 \geq 4\sigma_g^2/7n^{\frac{1}{2}}]$$

$$\leq \frac{n^{\frac{1}{2}}\,\mathrm{Var}(T_n)}{(4\sigma_g^2/7)} = n^{-\frac{1}{2}}\,\frac{\mathrm{Var}[(g^2(X_1)-\sigma_g^2)+\tilde{g}(X_1)]}{4\sigma_g^2/7} = O(n^{-\frac{1}{2}}).$$

This concludes the proof. QED.

Next, we give a random-sample size version of Theorem 1 under the set-up outlined in the introduction. This gives a studentized version of a Theorem of Ahmad (1980).

THEOREM 2. Under the conditions of Theorem 1 and assuming that

$\{N_n\}$ and $\{X_n\}$ are independent, then as $n \to \infty$,

$$(2.12) \qquad \sup_x \left| P[\sqrt{N_n}(U_{Nn} - \theta) \leq x\, S_{Nn}] - \Phi(x) \right| = O(m_n^{-\frac{1}{2}} + \sqrt{m_n^{(2)}}/m_n + (\sqrt{m_n^{(2)}}/m_n)^{\frac{1}{2}}),$$

where $EN_n = m_n$, $\mathrm{Var}(N_n) = m_n^{(2)}$ are such that $m_n \to \infty$ as $n \to \infty$.

PROOF. Since $\{N_n\}$ and $\{X_n\}$ are independent and putting $p_{n,k} = P[N_n = k]$, we have,

$$(2.13) \qquad \sup_x \left| P[\sqrt{N_n}(U_{N_n} - \theta) \leq x S_{N_n}] - \Phi(x) \right|$$

$$\leq \Sigma_{k=1}^{\infty} p_{n,k} \sup_x \left| P[\sqrt{k}\,(U_k - \theta) \leq S_k\, x] - \Phi(x) \right|$$

$$\leq \Sigma_{k=1}^{\infty} p_{n,k} \sup_x \left| P[\sqrt{k}\,(U_k - \theta) \leq 2\sigma_g x] - \Phi(x) \right|$$

$$+ \Sigma_{k=1}^{\infty} p_{n,k}\, P\left[\left|\frac{S_k}{2\sigma_g} - 1\right| \geq a_n\right]$$

$$+ \Sigma_{k=1}^{\infty} p_{n,k}\, a_k = J_{1n} + J_{2n} + J_{3n}, \quad \text{say}$$

Now, it follows from a Theorem of Ahmad (1980) that

$$(2.14) \qquad J_{1n} = O(m_n^{-\frac{1}{2}} + \sqrt{m_n^{(2)}}/m_n + (\sqrt{m_n^{(2)}}/m_n)^{\frac{1}{2}}), \quad \text{as} \quad n \to \infty.$$

Next choosing $a_k = k^{-\frac{1}{2}}$ gives,

$$(2.15) \qquad J_{3n} = \Sigma_{k=1}^{\infty} p_{k,n}\, a_k^{-\frac{1}{2}} = \Sigma_{k:\,|m_n - k| \leq m_n/2}\, p_{n,k}\, k^{-\frac{1}{2}} + \Sigma_{k=|m_n - k| > m_n/2}\, p_{n,k}\, k^{-\frac{1}{2}}$$

$$\leq (\tfrac{m_n}{2})^{-\frac{1}{2}} + \Sigma_{k=|m_n - k| > m_n/2}\, p_{n,k}\, \frac{2|k - m_n|}{k^{\frac{1}{2}}\, m_n} \leq (\tfrac{m_n}{2})^{-3/4} + \frac{c}{m_n} E|N_n - m_n|$$

$$\leq C\{m_n^{-\frac{1}{2}} + \sqrt{m_n^{(2)}}/m_n\}.$$

Finally we evaluate J_{2n}.

$$(2.16)\qquad J_{2n} = \Sigma_{k:|k-m_n|\le m_n/2}\, p_{n,k}\, P\left[\left|\frac{S_k}{2\sigma_g} - 1\right| \ge a_k\right] + \Sigma_{k:|k-m_n|>m_n/2}\, P\left[\left|\frac{S_k}{2\sigma_g} - 1\right| \ge a_k\right]$$

But as before the second term in the right-hand-side of (2.16) is less than or equal to $\sqrt{m_n^{(2)}}/m_n$. Let us evaluate the first term. For sufficiently large n, and using Theorem 2.1

$$(2.16)\qquad \Sigma_{k:|k-m_n|\le m_n/2}\, p_{n,k}\, P\left[\left|\frac{S_k}{2\sigma_g} - 1\right| \ge a_k\right]$$

$$\le \Sigma_{k:|k-m_n|\le m_n/2}\, p_{n,k}\, P[|T_k| \ge \epsilon_k/7] + \sum_{\ell=1}^{6} \Sigma_{k:|k-m_n|\le m/2} P[|R_{\ell n}| \ge \epsilon_k/7]$$

$$\le C\{\Sigma_{k:|k-m_n|\le m_n/2}\, p_{n,k}\, k^{-\frac{1}{2}} + \Sigma_{\ell=1}^{6}\, k:|k-m_n|\le m_n/2\, p_{n,k}\, k^{-\frac{1}{2}}$$

$$\le C\{(m_n/2)^{-\frac{1}{2}} + (m_n/2)^{-\frac{1}{2}}\} \le C\{m_n^{-\frac{1}{2}}\}.$$

Hence $J_{2n} = O(m_n^{-\frac{1}{2}} + \sqrt{m_n^{(2)}}/m_n)$. Collecting the bounds of J_{1n}, J_{2n}, and J_{3n} we reach the desired conclusion. QED.

As an application of the above theorem let us consider a super-critical Galton-Watson process by first considering a sequence $\{X_n\}$ of i.i.d. random variables. Let z_0 be a fixed number and define the stochastic process: $Z_0 = z_0$, $Z_1 = X_1 + \ldots + X_{Z_0}$, ..., $Z_2 = X_{Z_0+1} + \ldots + X_{Z_0+Z_1}$, ..., $Z_n = X_{Z_0+\ldots+Z_{n-2}+1} + \ldots + X_{Z_0+\ldots+Z_{n-1}}$. Assume that $EZ_1 = m > 0$. Let Z_n^* denote the random variable Z_n under the probability measure conditional on $Z_n > 0$. Since $m^n Z_n^{-\frac{1}{2}} (W - W_n)$ (where $W_n = m^{-n} Z_n$ and W is the a.s. limit of W_n as $n \to \infty$) has the same distribution as that of

$(Z_n^*)^{\frac{1}{2}}\ [X_1 + \ldots + X_{Z_n^*}]$, then $\mathrm{Var}\ X_1 = \sigma^2/(m^2 - m)$. In many cases this quantity is unknown and thus we can estimate $\mathrm{Var}\ X_1$ by $S^2_{Z_n^*} = \frac{1}{Z_n^*} \Sigma_{i=1}^{Z_n^*} (X_i - \bar{X}_{Z_n^*})^2$, with $\bar{X}_{Z_n^*} = \frac{1}{Z_n^*} \Sigma_{i=1}^{Z_n^*} X_i$. Thus we can apply Theorem 2.2 above if we assume that $E|Z_n|^4 < \infty$ to conclude that for sufficiently large n,

$$(2.17) \qquad \sup_x \left| P[\Sigma_{i=1}^{Z_n^*} X_i \leq x S_{Z_n^*} \sqrt{Z_n^*}] - \Phi(x) \right| = O(m^{-\frac{1}{2}} + \sqrt{m_2}/m + (\sqrt{m^{(2)}}/m)^{\frac{1}{2}}).$$

cf. (3.3) of Ahmad (1980).

REFERENCES

AHMAD, I.A. (1981). Some asymptotic properties of U-statistics. Scand. J. Statist., 8, 175-182.

AHMAD, I.A. (1980). On the Berry-Esseen theorem for random U-statistics. Ann. Statist., 8, 1395-1398

BERK, R.H. (1966). Limiting behavior of posterior distributions when the model is incorrect. Ann. Math. Statist., 37, 457-462.

CALLAERT, H. and VERAVERBEKE, N. (1981). The order of the normal approximation for a studentized U-statistic. Ann. Statist., 9, 194-200.

CALLAERT, H. and JANSSEN, P. (1978). The Berry-Esseen theorem for U-statistics. Ann. Statist., 6, 417-421.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distributions. Ann. Math. Statist., 19, 293-325.

HOEFFDING, W. (1961). The strong law of large numbers for U-statistics. Institute of Statistics Memo Series No. 302, University of North Carolina, Chapel Hill, N.C.

LOEVE, M. (1963). Probability Theory. Von Nostrand.

MICHEL, R. and PFANZAGL, J. (1971). The accurary of the normal approximation for the minimum contrast estimates. Z. Wharscheinlichkeitstheorie und verw. Geb., 18, 73-84.