ON THE ASYMPTOTIC DISTRIBUTION OF CERTAIN INCOMPLETE U-STATISTICS

by

Alan J. Lee

University of Auckland
and
University of North Carolina at Chapel Hill

Abstract

We consider incomplete U-statistics based on the arithmetic mean of m quantities g_i where g is the kernel of the complete U-statistic evaluated at a randomly chosen subsample of a sample of size n. The asymptotic distribution of the incomplete U-statistic is obtained in terms of that of the complete statistic, and the asymptotic relative efficiency of the incomplete statistic discussed. Comparisons are made with other incomplete U-statistics.

Some Key Words: Incomplete U-statistic, asymptotic distribution, asymptotic relative efficiency.

*The work of this author was partially supported by the Air Force Office of Scientific Research under Grant AFOSR-75-2796.

§1. Introduction. Let $X_1,\ldots,X_n$ be independent and identically distributed random variables with distribution function F, and let

$$\theta = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1,\ldots,x_\nu) \prod_{i=1}^{\nu} F(dx_i) \tag{1}$$

where the kernel g is symmetric in its ν arguments. As is well known, an unbiased estimator of θ is furnished by the U-statistic

$$U_n = \binom{n}{\nu}^{-1} \sum g(X_{i_1},\ldots,X_{i_\nu}) \tag{2}$$

where the sum ranges over all $\binom{n}{\nu}$ ν-subsets of the X_i. Such estimators may require excessive computation when ν and n are not small. The dependence among the terms in (2) suggests that an incomplete statistic of the form

$$U_m^* = \frac{1}{m} \sum g(X_{i_1},\ldots,X_{i_\nu}) \ , \tag{3}$$

where the sum ranges over $m < \binom{n}{\nu}$ ν-subsets chosen systematically or randomly, may estimate θ almost as well as (2). Such incomplete U-statistics have been studied by Blom [1] and Brown and Kildea [2]. In the present paper we consider the case where the m subsets in (3) are chosen at random with replacement from the $\binom{n}{\nu}$ possible subsets and derive the asymptotic distribution of the resulting (normalized) incomplete U-statistic. We discuss the efficiency of such statistics compared to the corresponding complete U-statistic and also compared to the balanced incomplete U-statistic proposed by Brown and Kildea.

We first summarize some results on complete U-statistics due to Hoeffding [3].

Let $\Phi_c(x_1,\ldots,x_c) = E(g(x_1,\ldots,x_c,X_{c+1},\ldots,X_\nu))$ and define $\zeta_c = \mathrm{Var}\ \Phi_c(X_1,\ldots,X_c)$ for $c = 1,2,\ldots,\nu$; $\zeta_0 = 0$. If $\zeta_k = 0$ for $k \le d$ and $\zeta_{d+1} \neq 0$, U_n is said to be stationary of order d at F. In this case $n^{\frac{d+1}{2}}\left(U_n-\theta\right)$ has a nondegenerate limit distribution, and $n^{d+1}\ \mathrm{Var}\ U_n$ converges to $\binom{\nu}{d+1}^2 (d+1)!\ \zeta_{d+1}$. See e.g. Puri and Sen [6], p. 54.

§2. Efficiency of the incomplete U-statistic. Suppose now the incomplete U-statistic U_m^* is constructed by selecting m subsets at random with replacement. Then (Blom [1]) U_m^* is unbiased, and its variance is given by $\mathrm{Var}\ U_m^* = \frac{\zeta_\nu}{m} + \left(1-\frac{1}{m}\right)\mathrm{Var}\ U_n$. Other relations between the second moments of U_n and U_m^* are $\mathrm{Var}\ U_m^* - \mathrm{Var}\ U_n = E\left|U_m^* - U_n\right|^2$ and $\mathrm{Cov}(U_n,U_m^*) = \mathrm{Var}\ U_n$. The relative efficiency of U_m^* compared with U_n is thus $\mathrm{Var}\ U_n \Big/ \left(\frac{\zeta_\nu}{m} + \left(1-\frac{1}{m}\right)\mathrm{Var}\ U_n\right)$ which converges as n,m tend to infinity to

$$\frac{(d+1)!\ \binom{\nu}{d+1}^2 \rho_{d+1}}{\alpha + (d+1)!\ \binom{\nu}{d+1}^2 \rho_{d+1}}$$

where $\alpha = \lim n^{d+1}m^{-1}$ and $\rho_{d+1} = \frac{\zeta_{d+1}}{\zeta_\nu}$. Note that since $0 < \frac{\zeta_{d+1}}{d+1} \le \frac{\zeta_\nu}{\nu}$ (Hoeffding [3]), $0 < \rho_{d+1} \le \frac{d+1}{\nu}$. Thus the asymptotic relative efficiency of U_m^* is zero, unity or between these two values according as α is ∞, 0 or a finite positive number. In the first case the asymptotic distribution of $\sqrt{m}(U_m^*-\theta)$ is normal with zero mean and variance ζ_ν, as proved below. In the second case, the asymptotic distribution of $n^{\frac{d+1}{2}}\left(U_m^*-\theta\right)$ coincides with that of $n^{\frac{d+1}{2}}\left(U_n-\theta\right)$, since

$$E\left|n^{\frac{d+1}{2}}\left(U_m^*-\theta\right) - n^{\frac{d+1}{2}}\left(U_n-\theta\right)\right|^2 = n^{d+1}E\left|U_m^* - U_n\right|^2$$

$$= n^{d+1}(\text{Var } U_m^* - \text{Var } U_n)$$

$$= n^{d+1}\left(\frac{\zeta_\nu}{m} - \frac{\text{Var } U_n}{m}\right)$$

which converges to zero.

The case $0 < \alpha < \infty$ is treated in §4 below.

§3. Comparison with Brown and Kildea's incomplete U-statistic. In [2] Brown and Kildea propose an incomplete U-statistic for the case $d=0, \nu=2$ based on systematic selection of Kn pairs X_i, X_j in such a way that each r.v. appears in exactly 2K pairs and each pair shares an r.v. with 2(2K-1) other pairs. The asymptotic efficiency of this statistic, relative to the complete U-statistic, is $2K\rho_1(\frac{1}{2}+(2K-1)\rho_1)^{-1}$, which is superior to that of U_m^* which is $K\rho_1(1+K\rho_1)^{-1}$ when $m=Kn, d=0, \nu=2$. Brown and Kildea's statistic, which has an asymptotically normal distribution, is thus to be preferred over U_m^* in this case. However, the efficiency of U_m^* improves as ν and d increase, provided that ρ_{d+1} is not too small. The construction of "balanced" incomplete U-statistics poses difficult combinatorial problems for high ν and d, and the resulting asymptotics are also not simple, so the use of statistics of type U_n^* may be regarded as a satisfactory albeit nonoptimal alternative.

§4. Asymptotic distribution of U_m^* when $0 < \alpha \leq \infty$. Let $\Psi_n(t)$ be the characteristic function of $n^{\frac{d+1}{2}}\left(U_n-\theta\right)$ and $\Psi(t)$ the c.f. of the limit

distribution. The c.f. of $\sqrt{m}(U_n-\theta)$ is thus $\Psi_n\left(m^{\frac{1}{2}}n^{\frac{-(d+1)}{2}}t\right)$, which converges to $\Psi(t\alpha^{-\frac{1}{2}})$ or 1 according as α is finite or infinite. Using these facts, we can now prove our asymptotic result.

Theorem. Suppose $\lim_{n,m\to\infty} n^{d+1}m^{-1} = \alpha$ where $0 < \alpha \leq \infty$. Then $\sqrt{m}(U_m^*-\theta)$ converges in distribution to a distribution with characteristic function

$$\Phi(t) = \begin{cases} e^{-\zeta_\nu t^2/2}\,\Psi(t\alpha^{-\frac{1}{2}}) & 0 < \alpha < \infty, \\ e^{-\zeta_\nu t^2/2} & \alpha = \infty. \end{cases}$$

Proof. Given a sample $x_1,\ldots,x_n$, let g_i, $i = 1,\ldots N$, be the values of the kernel $g(x_{i_1},\ldots,x_{i_\nu})$ where $x_{i_1},\ldots,x_{i_\nu}$ range over the $N = \binom{n}{\nu}$ subsets of $\{x_1,\ldots,x_n\}$. The conditional distribution of U_m^* given $\underset{\sim}{X} = (x_1,\ldots,x_n)$ is that of $\frac{1}{m}\sum_{j=1}^{m} Y_j$ where the Y_j are independent and identically distributed random variables with distribution $Pr[Y_j = g_i] = 1/N$, which has mean $\frac{1}{N}\sum_{i=1}^{N} g_i = U_N$ and variance $\sigma_n^2 = \frac{1}{N}\sum_{i=1}^{N}(g_i-U_n)^2$. Let $\phi_n(t)$ be the conditional characteristic function of the r.v. $\sqrt{m}\left(\frac{1}{m}\sum_{j=1}^{m} Y_j-U_n\right)/\sigma_n$ so that $\phi_n(t)$ converges to $e^{-t^2/2}$ by the central limit theorem. It follows that

$$E[\exp(it\sqrt{m}(U_m^*-\theta))\,|\,\underset{\sim}{x}] = E\left[e^{it\sqrt{m}\left(\sum_{j=1}^{m} Y_j/m-\theta\right)}\right]$$

$$= \phi_n(t\sigma_n)e^{i\sqrt{m}(U_n-\theta)t}$$

and so

$$E[\exp(it\sqrt{m}(U_m^*-\theta))] = E\left[\phi_n(t\sigma_n)e^{i\sqrt{m}(U_n-\theta)t}\right].$$

Consider

$$\left|E\left[\phi_n(t\sigma_n)e^{i\sqrt{m}(U_n-\theta)t}\right] - e^{-t^2\zeta_\nu}\,\Psi(t\alpha^{-\frac{1}{2}})\right|$$

$$\leq \left|E\left[\phi_n(t\sigma_n)e^{i\sqrt{m}(U_n-\theta)t}\right] - \phi_n(t\zeta_\nu^{\frac{1}{2}})\Psi_n(t\alpha_n^{-\frac{1}{2}})\right|$$

$$+ \left|\phi_n(t\zeta_\nu^{\frac{1}{2}})\Psi_n(t\alpha_n^{-1}) - e^{-\frac{1}{2}t\zeta_\nu}\,\Psi(t\alpha^{-\frac{1}{2}})\right|$$

where $\alpha_n = m^{-1}n^{d+1}$. The second term obviously converges to 0, while the first is less than

$$E\left|\phi_n(t\sigma_n)e^{i\sqrt{m}(U_n-\theta)t} - \phi_n(t\zeta_\nu^{\frac{1}{2}})e^{i\sqrt{m}(U_n-\theta)t}\right|$$

$$\leq \int_{-\infty}^{\infty}\left|\phi_n(tx) - \phi_n(t\zeta_\nu^{\frac{1}{2}})\right|dH_n(x) \qquad (4)$$

where H_n is the distribution function of the r.v. σ_n. Now $\sigma_n^2 = \frac{1}{N}\sum g^2(X_{i_1},\ldots,X_{i_\nu}) - U_N^2$ and $\frac{1}{N}\sum g^2(X_{i_1},\ldots,X_{i_\nu})$ is a U-statistic and so converges in probability to $E(g^2(X_1,\ldots,X_\nu)) = \zeta_\nu + \theta^2$. Similarly U_n converges in probability to θ and so σ_n^2 converges in probability to ζ_ν which implies that $\sigma_n \xrightarrow{P} \zeta_\nu^{\frac{1}{2}}$. Now write the integral in (4) above as

$$\int_{|x-\zeta_\nu^{\frac{1}{2}}|\leq\varepsilon}\left|\phi_n(tx) - \phi_n(t\zeta_\nu^{\frac{1}{2}})\right|H_n(dx) + \int_{|x-\zeta_\nu^{\frac{1}{2}}|>\varepsilon}\left|\phi_n(tx) - \phi_n(t\zeta_\nu^{\frac{1}{2}})\right|H_n(dx). \qquad (5)$$

The second integral in (5) is less than $2\Pr[|\sigma_n - \zeta_\nu^{\frac{1}{2}}| > \varepsilon]$ while the first

is less than $\sup_{|x-\zeta_\nu^{1/2}|\le\varepsilon} |\phi_n(tx) - \phi_n(t\zeta_\nu^{1/2})|$ so (5) is less than

$$\sup_{|x-\zeta^{1/2}|\le\varepsilon} |\phi_n(tx) - \phi(tx)| + \sup_{|x-\zeta^{1/2}|\le\varepsilon} |\phi(tx) - \phi(t\zeta)| + |\phi_n(t\zeta) - \phi(t\zeta)| + 2\Pr[|\sigma_n - \zeta_\nu^{1/2}| > \varepsilon] \tag{6}$$

where $\phi(t) = \lim \phi_n(t) = e^{-t^2/2}$. Now ϕ_n converges uniformly to ϕ on compact sets so by (4), (5) and (6)

$$\limsup E|\phi_n(t\sigma_n) - \phi_n(t\zeta_\nu^{1/2})| \le \sup_{|x-\zeta^{1/2}|\le\varepsilon} |\phi(tx) - \phi(t\zeta_\nu^{1/2})|$$

which by the uniform continuity of characteristic functions can be made arbitrarily small by suitable choice of ε. Thus $\lim E|\phi_n(t\sigma_n) - \phi_n(t\zeta_\nu^{1/2})| = 0$, and the proof is complete. Note that if $d = 0$ the distribution of $n^{1/2}(U_n-\theta)$ is normal with mean $\nu^2\zeta_1$ and so the asymptotic distribution of $\sqrt{m}(V_m-\theta)$ is normal with mean zero and variance $\zeta_\nu+\nu\zeta_1\alpha^{-1}$. When $d = 1$, $n(U_n-\theta)$ has an asymptotic distribution equal to that of $\sum \lambda_j(Z_j^2-1)$ for iid $N(0,1)Z_j$, where the λ_j are the eigenvalues of a certain integral operator (see [4] for details). The limit distribution of $\sqrt{m}(U_m^*-\theta)$ is thus that of $X + \sum \lambda_j'(Z_j^2-1)$ where X has an $N(0,\zeta_\nu)$ distribution independent of the Z_j, and $\lambda_j' = \alpha^{1/2}\lambda_j$. Finally, when $\alpha = \infty$, the asymptotic distribution of $\sqrt{m}(U_m^*-\theta)$ is $N(0,\zeta_\nu)$ irrespective of d. □

§5. U-statistics based on sampling without replacement. The above results, with slight modifications, remain true when we sample without replacement in the construction of the incomplete U-statistic.

Let U'_m be such a U-statistic, then

$$\text{Var } U'_m = E(E(U'_m-\theta^2)\,|\,\underset{\sim}{X}))$$

$$= E\left(\frac{\sigma_m^2}{m}\left(\frac{N-m}{N-1}\right)+(U_n-\theta)^2\right)$$

where $\sigma_m^2 = \frac{1}{m}\sum_{i=1}^{m}(g_i-U_n)^2$. Now

$$E(\sigma_m^2) = E\left(\frac{1}{m}\sum_{i=1}^{m}(g_i-\theta)^2\right) + E(U_n-\theta)^2$$

$$= \zeta_\nu - \text{Var } U_n$$

so $\text{Var } U'_m = \frac{\zeta_\nu}{m}\left(\frac{N-m}{N-1}\right) + \left(1-\frac{1}{m}\frac{(N-m)}{(N-1)}\right)\text{Var } U_n$, which corresponds to Blom's result for the with replacement case. The theorem of section 4 remains true, with only slight modifications needed in the proof. From Madow [5], Corollary 1, we note that conditional on X, $\Pr\left[\frac{\sqrt{m}(U'_m-U_n)}{\sigma_n(1-(m-1)/(N-1))^{\frac{1}{2}}} \le x\right]$ converges to $\Psi(x)$ as $m,n \to \infty$ provided σ_n does not converge to zero, $(m-1)/(N-1) < 1 - \varepsilon$ for all sufficiently large m,n and the g_i are uniformly bounded.

Assuming the latter (without loss of generality, see [2]) the proof of the theorem of section 4 applies almost without change.

References

[1] Blom, G. (1976). Some properties of incomplete U-statistics. *Biometrika, 63*, pp. 573-580.

[2] Brown, B.M. and Kildea, D.G. (1978). Reduced U-statistics and the Hodges-Lehmann estimator. *Ann. Statist., 6*, pp. 828-835.

[3] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist., 19*, pp. 293-325.

[4] Lee, A.J. (1979). On the asymptotic distribution of U-statistics. *Institute of Statistics Mimeo Series #1255*, University of North Carolina at Chapel Hill.

[5] Madow, W.G. (1948). On the limiting distributions of estimates based on samples from finite universes. *Ann. Math. Statist., 19*, pp. 535-545.

[6] Puri, M. and Sen, P.K. (1971). *Non Parametric Methods in Multivariate Analysis*. Wiley, New York.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* On the Asymptotic Distribution of Certain Incomplete U-Statistics		5. TYPE OF REPORT & PERIOD COVERED TECHNICAL 6. PERFORMING ORG. REPORT NUMBER Mimeo Series No. 1259
7. AUTHOR(s) Alan J. Lee		8. CONTRACT OR GRANT NUMBER(s) Grant AFOSR-75-2796
9. PERFORMING ORGANIZATION NAME AND ADDRESS		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research Bolling AFB Washington, DC 20332		12. REPORT DATE December 1979 13. NUMBER OF PAGES 9
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release--distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Incomplete U-statistic, asymptotic distribution, asymptotic relative efficiency

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

We consider incomplete U-statistics based on the arithmetic mean of m quantities g_i where g is the kernel of the complete U-statistic evaluated at a randomly chosen subsample of a sample of size n. The asymptotic distribution of the incomplete U-statistic is obtained in terms of that of the complete statistic, and the asymptotic relative efficiency of the incomplete statistic discussed. Comparisons are made with other incomplete U-statistics.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*