ON INCOMPLETE U-STATISTICS HAVING MINIMUM VARIANCE

by

Alan J. Lee

University of Auckland
and
University of North Carolina at Chapel Hill

Summary

Let U be an incomplete U-statistic of order k -- that is to say an arithmetic mean of m quantities $g(X_{i_1},\ldots,X_{i_k})$ where $X_1,\ldots,X_n$ is a random sample from some distribution, g is a function symmetric in its k arguments, and the sum is taken over m k-subsets $(i_1,\ldots,i_k)$ of $\{1,\ldots,n\}$. The problem of how to choose the m subsets to make the variance of U a minimum is discussed. Some results on the asymptotic properties of U are given.

Key Words and Phrases: incomplete U-statistic, design, balanced incomplete blocks, consistent estimation.

The work of this author was supported by the Air Force Office of Scientific Research under Grant AFOSR-80-0080.

1. Introduction.

Let $X_1,\ldots,X_n$ be a random sample from a distribution with d.f. F, where F is a member of a class $\mathcal{F}$ of dfs. Let θ be a parameter such that

$$\theta = \theta(F) = E[g(X_1,\ldots,X_k)] \text{ for all } F \in \mathcal{F} \tag{1}$$

where g is symmetric in its k arguments. In this case the functional $\theta(F) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1,\ldots,x_k) \prod_{i=1}^{k} dF(x_i)$ is said to be regular, and under certain conditions on $\mathcal{F}$, the minimum variance unbiased estimator of $\theta(F)$ is the complete U-statistic

$$U^c = \frac{1}{\binom{n}{k}} \sum g(X_{i_1},\ldots,X_{i_k}) \tag{2}$$

where the sum is taken over all $\binom{n}{k}$ k-subsets of $\{1,\ldots,n\}$.

The amount of computation required to calculate (2) may be excessive if n and k are large. Because of the high degree of dependence between the terms of (2), it appears that discarding some of the terms involved in the U-statistic will not appreciably inflate the variance. We are thus led to consider incomplete U-statistics of the form

$$U_m = \frac{1}{m} \sum g(X_{i_1},\ldots,X_{i_k}) \tag{3}$$

where the sum now extends over m k-subsets chosen in some manner. Such incomplete U-statistics have been studied by Blom (1976), Brown and Kildea (1978), and Lee (1979).

In the present paper we address the question of how to choose the sets defining (3) to make the variance a minimum. We also discuss asymptotic properties of sequences of optimal incomplete U-statistics.

Since the variance of an incomplete U-statistic is always greater than that of the corresponding complete statistic, one is trading off efficiency against simple computation. However, in some cases it is possible to achieve considerable savings in computation coupled with a negligible or zero decrease in asymptotic relative efficiency. We return to this point in §4.

2. The variance of an incomplete U-statistic.

Following Hoeffding (1948), define functions $g_c(x_1,\ldots,x_c) = E[g(x_1,\ldots,x_c, X_{c+1},\ldots,X_k)]$ for $c = 1,2,\ldots,k$ and $\sigma_c^2 = \mathrm{Var}(g_c(X_1,\ldots,X_c))$. Let S_j, $j = 1,\ldots,m$, be m k-subsets of $\{1,2,\ldots,n\}$ and let $g_j = g(X_{i_1},\ldots,X_{i_k})$ where $S_j = \{i_1,\ldots,i_k\}$. The U-statistic based on the S_j is $U_m = \frac{1}{m}\sum_{j=1}^{m} g_j$ and we call the sets S_j the design of the U-statistic. The design is conveniently described by the $n\times m$ incidence matrix N whose i-j element is 1 if $i \in S_j$ and zero otherwise.

From Blom (1976) the variance of U_m is given by

$$\mathrm{Var}\ U_m = \frac{1}{m^2}\sum_{c=1}^{k} f_c\sigma_c^2 \tag{4}$$

where f_c is the number of pairs of sets $S_j, S_{j'}$ having exactly c elements in common. f_c is the number of elements of N'N equal to c. If we assume the sets S_j to be distinct, then $f_k = m$. We now develop an alternative expression for the variance. Let $a(i_1,\ldots,i_\ell)$ be the number of sets that contain the set $\{i_1,\ldots,i_\ell\}$, so that

$$a(i_1,\ldots,i_\ell) = \sum_{j=1}^{m} n_{i_1 j}\cdots n_{i_\ell j}$$

where $N = (n_{ij})$. Define for $\nu = 1,\ldots,k$

$$A_\ell = \sum_{1\le i_1<\ldots<i_\ell\le n} a^2(i_1,\ldots,i_\ell)$$

$$B_\ell = \sum_{1\le i_1<\ldots<i_\ell\le n} a(i_1,\ldots,i_\ell) \ .$$

Define the Stirling numbers $S_\mu^{(\nu)}$ and $S_\nu^{(\mu)}$ of the first and second kind by

$$x^\nu = \sum_{\mu=1}^{\nu} S_\nu^{(\mu)}\, x(x-1)\ldots(x-\mu+1)$$

and

$$x(x-1)\ldots(x-\mu+1) = \sum_{\nu=1}^{\mu} S_\mu^{(\nu)}\, x^\mu \ .$$

Then

$$\sum_{i_1=1}^{n}\cdots\sum_{i_\nu=1}^{n}\left(\sum_{j=1}^{m} n_{i_1 j}\cdots n_{i_\nu j}\right)^2 = \sum_{\mu=1}^{\nu} \mu!\, S_\nu^{(\mu)}\, A_\mu \tag{5}$$

$$\sum_{i_1=1}^{n}\cdots\sum_{i_\nu=1}^{n}\sum_{j=1}^{m} n_{i_1 j}\cdots n_{i_\nu j} = \sum_{\mu=1}^{\nu} \mu!\, S_\nu^{(\mu)}\, B_\mu \ . \tag{6}$$

But the left-hand side of (5) and (6) are respectively equal to

$$\sum_{j=1}^{m} \sum_{j'=1}^{m} \left(\sum_{i=1}^{n} n_{ij} n_{ij'} \right)^{\nu} = \sum_{c=1}^{k} c^{\nu} f_c$$

and

$$\sum_{j=1}^{m} \left(\sum_{i=1}^{n} n_{ij} \right)^{\nu} = mk^{\nu}$$

so

$$\sum_{c=1}^{k} c^{\nu} f_c = \sum_{\mu=1}^{\nu} \mu! \, S_{\nu}^{(\mu)} A_{\mu} \tag{7}$$

and

$$mk^{\nu} = \sum_{\mu=1}^{\nu} \mu! \, S_{\nu}^{(\mu)} A_{\mu} \, . \tag{8}$$

Multiplying each side of (7) and (8) by $S_{\ell}^{(\nu)}$, summing over ν from 1 to ℓ and using the identity

$$\sum_{\nu=\mu}^{\ell} S_{\nu}^{(\mu)} S_{\ell}^{(\nu)} = \delta_{\mu\ell}$$

yields from (7)

$$\ell! \, A_{\ell} = \sum_{c=1}^{k} f_c \left\{ \sum_{\nu=1}^{\ell} S_{\ell}^{(\nu)} c^{\nu} \right\}$$

$$= \sum_{c=1}^{k} f_c \, c(c-1)\ldots(c-\ell+1)$$

and so

$$A_\ell = \sum_{c=\ell}^{k} f_c \binom{c}{\ell} . \tag{9}$$

Similarly from (8)

$$\ell!\, B_\ell = m \sum_{\nu=1}^{\ell} S_\ell^\nu k^\nu = m\, k(k-1)\ldots(k-\ell+1)$$

and so

$$B_\ell = m\binom{k}{\ell} . \tag{10}$$

Finally multiplying (9) by $(-1)^{\ell-\nu}\binom{\ell}{\nu}$, summing over ℓ from ν to k and using the identity $\sum_{\ell=\nu}^{c} (-1)^{\ell-\nu}\binom{\ell}{\nu}\binom{c}{\ell} = \binom{c}{\nu}\delta_{c\nu}$ we obtain

$$f_\nu = \sum_{\ell=\nu}^{k} (-1)^{\ell-\nu}\binom{\ell}{\nu} A_\ell \tag{11}$$

so by (4)

$$\text{Var } U_m = \frac{1}{m^2}\left[\sum_{c=1}^{k} \sigma_c^2 \sum_{\ell=\nu}^{k} (-1)^{\ell-\nu}\binom{\ell}{\nu} A_\ell\right]$$

$$= \frac{1}{m^2}\left[\sum_{\ell=1}^{k} d_\ell A_\ell\right]$$

where $d_\ell = \sum_{c=1}^{\ell} (-1)^{\ell-c}\binom{\ell}{c}\sigma_c^2$.

The quantities d_ℓ are non-negative by Lemma 5.1 of Hoeffding (1948), so Var U_m is minimized by designs having all A_ℓ a minimum.

In general, finding designs which simultaneously minimize the A_ℓ is a difficult problem in 0-1 programming. However, in certain situations we may be able to recognize minimum A_ℓ by simple arguments. In particular, note that A_ℓ is the sum of squares of $\binom{n}{\ell}$ non-negative integers $a(i_1,\ldots,i_\ell)$ whose sum is $B_\ell = m\binom{k}{\ell}$. If a design exists for which the $a(i_1,\ldots,i_\ell)$ are all equal, then A_ℓ will be a minimum. A necessary condition for this is that $m\binom{k}{\ell}/\binom{n}{\ell}$ is a positive integer.

Alternatively, if $m\binom{k}{\ell} < \binom{n}{\ell}$ and a design exists for which the $a(i_1,\ldots,i_\ell)$ are all zeroes and ones, then $m\binom{k}{\ell}$ are unity and the rest zero, and so $A_\ell = B_\ell$. Clearly for any design $A_\ell \geq B_\ell$, so A_ℓ is minimized if the $a(i_1,\ldots,i_\ell)$ are zeroes and ones. Moreover if for any $\nu \geq 2$, $a(i_1,\ldots,i_\nu)$ is equal to zero or one for all $i_1,\ldots,i_\nu$, then A_ℓ attains its minimum for $\ell \geq \nu$.

To see this, note that if $a(i_1,\ldots,i_\nu)$ is zero or one for all subsets $i_1,\ldots,i_\nu$ then necessarily $a(i_1,\ldots,i_\ell)$ is zero or one for $\ell \geq \nu$ and so $A_\ell = B_\ell$ and so attains a minimum.

We now turn to the description of designs for which the A_ℓ have one of the two properties above.

3. Designs having minimum variance.

3.1. *Designs based on tactical configurations.* A class of designs for which the A_ℓ are sums of squares of equal terms are those based on tactical configurations. A tactical configuration (t.c.) $C(k,\ell,\delta,n)$ is a system of m k-subsets of $\{1,\ldots,n\}$ such that every ℓ-subset of $\{1,\ldots,n\}$

is contained in exactly δ of the m subsets of the system. A necessary condition for the existence of a t.c. is that the quantities

$$\delta_h = \delta\binom{n-h}{\ell-h}/\binom{k-h}{\ell-h}$$

are integers for $h = 0,1,\ldots,\ell-1$. If such a t.c. exists it is also a $C(k,h,\delta_h,n)$ t.c. for $h = 1,2,\ldots,\ell-1$. The number m of sets is given by

$$m = \delta\binom{n}{\ell}/\binom{k}{\ell} .$$

These facts may be found for example in Raghavarao (1971).

Suppose now that a $C(k, k-1, \delta, n)$ tactical configuration exists. Then the incomplete U-statistic based on this t.c. has minimum variance, for by definition every ℓ-subset of $\{1,\ldots,n\}$ is contained in δ_ℓ subsets S_j of the system, and so each A_ℓ is a sum of squares of equal terms and so is a minimum.

Special cases.

a) $k = 2$. A $C(2,1,\delta,n)$ configuration is simply a system for which each element i is contained in the same number δ of sets S_j. Designs (for any k) having this property are termed balanced by Blom. The balanced designs for $k = 2$ have been considered by Brown and Kildea (1978), who prove the asymptotic normality of U-statistics based on them.

b) $k = 3$. A $C(3,2,\delta,n)$ t.c. is just a balanced incomplete block design for n varieties with $m = \delta\binom{n}{2}/3$ blocks of 3 plots each. Each pair of indices (varieties) occurs in $\lambda = \delta$ sets (blocks), and the design is balanced in the sense of Blom with each index appearing in

$r = \delta(n-1)/2$ of the sets S_j. The variance of the incomplete U-statistic is given by

$$\text{Var } U_m = \frac{1}{m}(3\sigma_1^2(r-2\lambda+1)+3\sigma_2^2(\lambda-1)+\sigma_3^2) \ .$$

A sequence of designs having $m = 0\left(\frac{n^2}{6}\right)$ are those based on Steiner triple systems. See e.g. Raghavarao (1971) p. 86. These are BIBD's with parameters $k = 3$ and

$$n = 6t + 3 \qquad m = (3t+1)(2t+1) \qquad r = 3t + 1 \qquad \lambda = 1$$

or

$$n = 6t + 1 \qquad m = t(6t+1) \qquad r = 3t \qquad \lambda = 1$$

and exist for every positive integer t. The variances of U-statistics based on them are $(9t\sigma_1^2+\sigma_3^2)/(3t+1)(2t+1)$ and $(3(3t-1)\sigma_1^2+\sigma_3^2)/t(6t+1)$ respectively.

c) $k = 4$. $C(4,3,1,n)$ configurations are called quadruple systems and have $m = \binom{n}{3}/4$ subsets. They exist (Hanani (1960)) when $n = 2$ or $4 \pmod 6$. Little is known about t.c.'s for which $k > 4$.

3.2. *Other designs based on tactical configurations.* Suppose that a $C(k,\nu,1,n)$ configuration exists for $\nu \geq 2$. Then any ν-set is contained in at most 1 of the S_j and so A_ℓ is minimized for $\ell \geq \nu$. Since the design is a $C(k,\ell,\delta_\ell,n)$ t.c. for $1 \leq \ell < \nu$ the A_ℓ are also minimized for $\ell < \nu$ and so the U-statistic has minimum variance. In particular BIBD designs for arbitrary k,n,m are of this type if λ (the number of blocks containing any pair of varieties) is unity.

3.3. *Designs based on partially balanced incomplete blocks.* We consider designs based on two associate classes with n varieties, k plots per block and m blocks. Let $n_1, n_2, \lambda_1, \lambda_2$ be the parameters of the association scheme, so that every variety has n_1 first associates and n_2 second associates; any pair of varieties that are i^{th} associates occurs in λ_i blocks, i = 1,2. Suppose $\lambda_1 = 1$ and $\lambda_2 = 0$ (or $\lambda_1 = 0$ and $\lambda_2 = 1$); any pair of varieties occurs in at most one block and so A_2 is a minimum. Since the PBIB is balanced in the Blom sense, A_1 is also a minimum. Since the $a(i_1, i_2)$ are zeroes and ones, A_ℓ is also a minimum for $\ell \geq 2$. Thus the U-statistic based on a PBIB with two associate classes has minimum variance if $\lambda_1 = 1$ and $\lambda_2 = 0$.

3.4. *Designs based on cyclic permutations.* We now consider a class of balanced designs, with $m = nK$ for some integer K, that have the property that the off-diagonal elements of the matrix NN' are zeroes and ones. Since A_2 is the sum of squares of the upper off-diagonal elements of NN', such designs correspond to minimum variance U-statistics. We first consider the case $K = 1$. Then $m = n$ and $r = k$, and for any balanced design N is square with row and column sums equal to k. It follows (see e.g. Raghavarao (1971), p. 107) that N can be written

$$N = \sum_{\ell=1}^{k} P_\ell$$

where P_ℓ is a permutation matrix, i.e. one whose i-j element is of the form $\delta_{i,p(j)}$ for some permutation p of the integers $\{1,\ldots,n\}$. Conversely, any set of k permutations $p_1,\ldots,p_k$ will generate a balanced design, provided $p_{\ell_1}(j) \neq p_{\ell_2}(j)$ for distinct ℓ_1, ℓ_2 and $j = 1,\ldots,n$.

Now let q be the cyclic permutation $q(j) = (j+1) \pmod n$, and for integers α_ℓ, $0 \le \alpha_\ell < n$, set $p_\ell = q^{\alpha_\ell}$. Let P_ℓ be the permutation matrix corresponding to p_ℓ, and set $N = \sum_{\ell=1}^{k} P_\ell$. We note that, provided the α_ℓ are distinct, N is the incidence matrix of a balanced design. Also $NN' = \sum_{\ell_1=1}^{k} \sum_{\ell_2=1}^{k} P_{\ell_1} P'_{\ell_2}$ where $P_{\ell_1} P'_{\ell_2}$ is the permutation matrix corresponding to the permutation $q^{\alpha_{\ell_2} - \alpha_{\ell_1}}$.

Now call the elements of an $n \times n$ matrix for which $i - j = c \pmod n$ the c-diagonal of the matrix. The matrix $P_{\ell_1} P'_{\ell_2}$ has ones on its $\alpha_{\ell_2} - \alpha_{\ell_1}$ diagonal and zero elsewhere. Thus, provided the quantities $\alpha_{\ell_2} - \alpha_{\ell_1}$ for $\ell_1, \ell_2 = 1, \ldots, k$ are distinct (mod n) when they are not zero, NN' has its off-diagonal elements zero or one, and hence N is the incidence matrix of a minimum variance U-statistic. The table below gives suitable values for the α_ℓ for different k values.

TABLE 1

Values of α_ℓ for different k values. The range of n yielding minimum variance U-statistics appears in parentheses.

	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 4$	$\ell = 5$	
k = 2	0	1				$(n \ge 3)$
k = 3	0	1	3			$(n \ge 7)$
k = 4	0	1	4	6		$(n \ge 13)$
k = 5	0	1	4	9	11	$(n \ge 23)$
k = 5	0	1	4	14	16	$(n = 21)$

For $K > 1$, it may be possible to construct suitable incidence matrices of the form $N = [N_1 \vdots N_2 \vdots \ldots \vdots N_K]$ where each N_ℓ, $\ell = 1,\ldots,K$, is of the above form. Then $NN' = N_1N_1' + \ldots + N_KN_K'$ will have its off-diagonal elements zero or one if the matrices $N_\ell N_\ell'$ have their non-zero diagonals all distinct. For example if $k = 3$ and $K = 2$ we can choose N_1 based on α_ℓ values $0,1,3$ and N_2 based on $0,4,9$. The resulting N has distinct non-zero diagonals for $n \geq 19$ and so is a design for a minimum variance U-statistic.

4. Asymptotic results.

First we comment on the asymptotic relative efficiency of the designs above. Let U_{m_n} be a sequence of U-statistics based on $n \times m_n$ incidence matrices N_n which have the property that $A_2 = m\binom{k}{2}$ or equivalently that the off-diagonal elements of N_nN_n' are zeroes and ones and are balanced. Then, dropping subscripts and assuming $\sigma_1^2 > 0$,

$$\text{Var } U_m = \frac{1}{m^2}(f_1\sigma_1^2 + m\sigma_k^2)$$

where

$$f_1 = \sum_{\ell=1}^{k} (-1)^{\ell-\nu} \binom{\ell}{\nu} A_\ell = \sum_{\ell=2}^{k} (-1)^{\ell-\nu} \binom{\ell}{\nu} m\binom{k}{\ell} + rmk$$

since

$$A_1 = rmk\,,\ A_\nu = m\binom{k}{\nu}\,, \qquad \nu \geq 2\,.$$

Thus

$$f_1 = m \sum_{\ell=1}^{k} (-1)^{\ell-\nu} \binom{\ell}{\nu}\binom{k}{\nu} + rmk - mk$$

$$= mk(r-1)$$

so

$$\text{Var } U_m = \frac{1}{m}(k(r-1)\sigma_1^2 + \sigma_k^2) \ .$$

The variance of the complete U-statistic is

$$\text{Var } U_n^c = \frac{1}{n}k^2\sigma_1^2 + 0\left(\frac{1}{n}\right)$$

so the asymptotic relative efficiency of the complete versus the incomplete statistic is

$$\lim_{n\to\infty} \frac{\frac{1}{n}k^2\sigma_1^2 + 0\left(\frac{1}{n}\right)}{\frac{1}{m}(k(r-1)\sigma_1^2 + \sigma_k^2)} = \lim_{n\to\infty} \frac{k^2\sigma_1^2}{k^2\sigma_1^2\left(\frac{r-1}{r}\right) + \frac{k\sigma_k^2}{r}} \ .$$

Suppose that $\frac{m}{n} \to \infty$ so that $r \to \infty$. Then the ARE is unity, and if $\frac{m}{n} \to b$, say, then $r \to bk$ and the ARE is $bk\sigma_1^2/(k(bk-1)\sigma_1^2 + \sigma_k^2) < 1$. For example, a sequence of designs having the former property are those for $k = 3$ based on the Steiner triple systems, and a sequence having the latter property are those for $k=3$, $m=n$ based on cyclic permutations with $b = 1$ and an ARE of $k^2\sigma_1^2/(k(k-1)\sigma_1^2 + \sigma_k^2)$.

For results on asymptotic normality, once again consider a sequence of designs based on incidence matrices N_n of the type described above.

We distinguish two cases: (i) $r \to \infty$; (ii) $r \to bk$ with b finite. If $r \to \infty$, then $n \text{ Var } U_m = \frac{n}{m}(k(r-1)\sigma_1^2 + \sigma_k^2) = \frac{k^2(r-1)}{r}\sigma_1^2 + \frac{k\sigma_k^2}{r}$, and so $\lim n \text{ Var } U_m = k^2\sigma_1^2 = \lim_n \text{Var } U_n^c$ where U_n^c is the corresponding complete statistic. It follows from §5 of Blom (1976) that the asymptotic distributions of $\sqrt{n}(U_n^c - \theta)$ and $\sqrt{n}(U_{m_n} - \theta)$ coincide and are $N(0, k^2\sigma_1^2)$. When $r \to kb < \infty$, then $m_n \text{ Var } U_{m_n}$ converges to $k(bk-1)\sigma_1^2 + \sigma_k^2$ and a trivial extension of the argument of Brown and Kildea (1978), Theorem 1, shows that $\sqrt{m}(U_{m_n} - \theta)$ converges in distribution to $N(0, k(bk-1)\sigma_1^2 + \sigma_k^2)$.

To construct confidence intervals for θ, we need a consistent estimator of $k^2\sigma_1^2$ or $k(bk-1)\sigma_1^2 + \sigma_k^2$ as appropriate. Define for $i = 1,\ldots,n$

$$W_i = mU_m - (m-r)U_{(i)}$$

where $U_{(i)} = \frac{1}{m-r}\sum_{(i)} g_j$ and the sum $\sum_{(i)}$ is taken over all sets S_j not containing i. Then if $\underset{\sim}{W}' = (W_1,\ldots,W_n)$ and $\underset{\sim}{g} = (g_1,\ldots,g_m)$, $\underset{\sim}{W} = N\underset{\sim}{g}$. Consistent estimators of $k^2\sigma_1^2$ and $k(kb-1)\sigma_1^2 + \sigma_k^2$ may be based on the statistic $\sum(W_i - \overline{W})^2$ where $\overline{W} = \frac{1}{n}\sum_{i=1}^{n} W_i$. Similar statistics are employed in the complete U-statistic case for consistent estimation of $k^2\sigma_1^2$. Consider

$$\sum(W_i - \overline{W})^2 = \underset{\sim}{W}'(I_n - \frac{1}{n}J_n)\underset{\sim}{W}$$

$$= \underset{\sim}{g}(N'N - \frac{k^2}{n}J_m)\underset{\sim}{g}$$

where I_n and J_n are square identity matrices and matrices of 1's respectively. The covariance matrix of the vector $\underset{\sim}{g}$, π say, has elements

$\pi_{ij} = \sigma_c^2$ where c is the i-j element of N'N. Thus

$$E\left(\sum_{i=1}^{n}(W_i-\overline{W})^2\right) = \text{trace } \Pi\left(N'N-\frac{k^2}{n}J_m\right) + \theta^2 \underset{\sim}{1}'\left(N'N-\frac{k^2}{n}J_m\right)\underset{\sim}{1}$$

$$= (f_1\sigma_1^2 + kf_k\sigma_k^2) - \frac{k^2}{n}(f_1\sigma_1^2 + f_k\sigma_k^2) \qquad (12)$$

$$= mk(r-1)\left(1-\frac{k^2}{n}\right)\sigma_1^2 + mk\left(1-\frac{k}{n}\right)\sigma_k^2$$

Now consider the case $r \to \infty$. The estimator $\hat{\beta}_\infty = \frac{k}{mr}\sum_{i=1}^{n}(W_i-\overline{W})^2$ is an asymptotically unbiased estimator of $k^2\sigma_1^2$, since $\lim_n E\left(\frac{k}{mr}\sum(W_i-\overline{W})^2\right) = \lim\left(k^2\frac{(r-1)}{r}\sigma_1^2 + \frac{k^2}{r}\sigma_k^2\right) = k^2\sigma_1^2$. It is also consistent; to prove this (under the assumption $E|g(X_1,\ldots,X_k)|^4 < \infty$) write

$$\frac{k}{mr}\sum_{i=1}^{n}(W_i-\overline{W})^2 = \frac{k}{mr}\underset{\sim}{g}N'N\underset{\sim}{g} - \frac{k^3}{nmr}\left(\sum_{j=1}^{m}g_j\right)^2$$

$$= \frac{k}{mr}\sum_{(1)}g_jg_{j'} + \frac{k^2}{rm}\sum_{j=1}^{m}g_j^2 - k^2U_m^2 \qquad (13)$$

where $\sum_{(1)}$ denotes the sum over all sets S_j having one element in common. Since U_m converges to θ in probability, the last term in (13) converges to $-k^2\theta^2$ in probability. Also $\frac{1}{m}\sum_{j=1}^{m}g_j^2$ is an incomplete U-statistic based on the kernel g^2 and so converges in probability to $E(g^2) = \sigma_k^2 + \theta^2$. Thus the second term converges in probability to zero.

For the first term, consider

$$\mathrm{Var}\left(\frac{k}{mr}\sum_{(1)} g_j g_{j'}\right) = \frac{k^2}{m^2 r^2}\sum_{j,j',\ell,\ell'} \mathrm{cov}(g_j g_{j'}, g_\ell g_{\ell'})$$

where the sum is taken over all j,j',ℓ,ℓ' such that the sets $S_j \cap S_{j'}$ and $S_\ell \cap S_{\ell'}$ each have exactly one element. If $(S_j \cup S_{j'}) \cap (S_\ell \cup S_{\ell'}) = \phi$ the corresponding term in the sum is zero, since $g_j g_{j'}$ and $g_\ell g_{\ell'}$ are independent in this case. The number of terms for which this is not the case is less than $m.rk.(2k-1)r.kr = 0(mr^3)$ and so $\mathrm{Var}\left(\frac{k}{mr}\sum_{(1)} g_j g_{j'}\right) = \frac{k^2}{m^2 r^2} 0(mr^3) = \frac{k^3}{n} 0(1)$ converges to zero. Since $E\left(\frac{k}{mr}\sum_{(1)} g_j g_{j'}\right) = \frac{kf_1}{mr}(\sigma_1^2 + \theta^2) = k^2\frac{(r-1)}{r}(\sigma_1 + \theta^2)$ it follows that the first term of (13) converges in probability to $k^2(\sigma_1^2 + \theta^2)$ and so (13) converges in probability to $k^2\sigma_1^2$. Thus when $r \to \infty$, $\hat{\beta} = \frac{k}{mr}\sum_{i=1}^{n}(W_i - \overline{W})^2$ is a consistent estimator of $\beta = k^2\sigma_1^2$. It follows that when $r \to \infty$ $\sqrt{\hat{\beta}_\infty^{-1}}\, n \left(U_m - \theta\right)$ has asymptotically a normal distribution with mean zero and variance unity.

For the case $r \to kb < \infty$, the estimate must be modified. Consider $\frac{1}{m}\sum (W_i - \overline{W})^2$; from (12) we obtain $\lim_n E\,\frac{1}{m}\sum (W_i - \overline{W})^2 = k(kb-1)\sigma_1^2 + k\sigma_k^2$ so $\frac{1}{m}\sum (W_i - \overline{W})^2$ is not an asymptotically unbiased estimate of $k(bk-1)\sigma_1^2 + \sigma_k^2$. Using a decomposition similar to (13) we obtain

$$\frac{1}{m}\sum (W_i - \overline{W})^2 = \frac{1}{m}\sum_{(1)} g_j g_{j'} + \frac{k}{m}\sum_{j=1}^{m} g_j^2 - rkU_m^2 ,$$

which using the arguments above converges in probability to

$$k(kb-1)(\sigma_1^2 + \theta^2) + k(\sigma_k^2 + \theta^2) - k^2 b\theta^2 = k(kb-1)\sigma^2 + k\sigma_k^2 .$$

To adjust the estimate, consider

$$\frac{1}{m-1} \sum_{j=1}^{m} (g_j - \bar{g})^2 = \frac{1}{m-1} \underset{\sim}{g}'(I_m - \frac{1}{m} J_m)\underset{\sim}{g} .$$

We have

$$E\left[\frac{1}{(m-1)} \sum_{j=1}^{m} (g_j - \bar{g})^2\right] = \frac{1}{m-1} \text{ trace } \Pi(I_m - \frac{1}{m} J_m)$$

$$= \frac{1}{m-1}(m\sigma_k^2 - \frac{1}{m}(f_1\sigma_1^2 + f_k\sigma_k^2))$$

$$= \sigma_k^2 - \frac{k(r-1)}{m-1}\sigma_1^2$$

and $\frac{1}{m-1} \sum_{j=1}^{m} (g_j - \bar{g})^2 = \frac{1}{m-1} \sum_{j=1}^{m} g_j^2 - \frac{1}{m-1} U_m^2$ converges in probability to σ_k^2. Thus $\frac{1}{m-1} \sum (g_j - \bar{g})^2$ is an asymptotically unbiased and consistent estimator of σ_k^2. It follows from the above that

$$\hat{\beta} = \frac{k}{mr} \sum_{i=1}^{n} (W_i - \bar{W})^2 - \frac{k-1}{m-1} \sum_{j=1}^{m} (g_j - \bar{g})^2$$

is an asymptotically unbiased and consistent estimator of $k(kb-1)\sigma_1^2 + \sigma_k^2$ and so $\sqrt{\hat{\beta}^{-1}}\, n \left(U_m - \theta\right)$ is asymptotically N(0,1).

References

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

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

Hanani, H. (1960). On quadruple systems. *Can. J. Math. 12*, 145-157.

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

Lee, A.J. (1979). On the asymptotic distribution of certain incomplete U-statistics. *Institute of Statistics Mimeo Series #1259*, University of North Carolina at Chapel Hill.

Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Experimental Design*. Wiley: New York.

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 Incomplete U-Statistics Having Minimum Variance	5. TYPE OF REPORT & PERIOD COVERED: TECHNICAL
	6. PERFORMING ORG. REPORT NUMBER: Mimeo Series No. 1283
7. AUTHOR(s): Alan J. Lee	8. CONTRACT OR GRANT NUMBER(s): Grant AFOSR-80-0080
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 Air Force Base, Washington, DC 20332	12. REPORT DATE: April 1980
	13. NUMBER OF PAGES: 18
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, design, balanced incomplete blocks, consistent estimation

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

Let U be an incomplete U-statistic of order k -- that is to say an arithmetic mean of m quantities $g(X_{i_1},\ldots,X_{i_k})$ where $X_1,\ldots,X_n$ is a random sample from some distribution, g is a function symmetric in its k arguments, and the sum is taken over m k-subsets $(i_1,\ldots,i_k)$ of $\{1,\ldots,n\}$. The problem

20. of how to choose the m subsets to make the variance of U a minimum is discussed. Some results on the asymptotic properties of U are given.