SMALL SAMPLE VARIANCE ESTIMATORS FOR U-STATISTICS

William R. Schucany

Daniel M. Bankson

Department of Statistical Science
Southern Methodist University, Dallas, Texas 75275

SMU/DS/TR-206

February 1988

SMALL SAMPLE VARIANCE ESTIMATORS FOR U-STATISTICS

William R. Schucany and Daniel M. Bankson*

Department of Statistical Science
Southern Methodist University
Dallas, Texas 75275

ABSTRACT

New unbiased estimators are presented for the dominant and lower-order terms of the variance expansion for U-statistics. In small samples these provide important corrections to the usual estimate of asymptotic standard error which is based on the leading term in the expansion. The new estimators for the first term cannot be recommended. The ordinary jackknife estimator is found to be more effective than the direct estimates of the separate terms.

Key Words: Approximate standardization, Bootstrap, Jackknife, Second order, Studentized quantity

* Research supported in part by ONR Contract N00014-85-K-0340

1. Introduction

Useful expansions for the variance of a U-statistic appear in the seminal paper by Hoeffding (1948). These are presented and illustrated in current texts, such as Randles and Wolfe (1979), Chapter 3 and Serfling (1980), Chapter 4. Under certain regularity conditions the large sample normality of this class of unbiased estimators makes these statistics obvious candidates for approximate tests and confidence intervals. A simple expression for the dominant term of the variance expansion and a related consistent estimator of the asymptotic variance given by Sen (1960) has been used in particular applications to produce studentized quantities. In large samples these have worked well. For U-statistics of degree two Callaert and Veraverbeke (1981) and more recently Helmers (1985) give conditions under which the studentized statistic has the same rate of convergence to normality as the exactly standardized quantity. These authors studentize with the jackknife estimator of the asymptotic variance. In the context of asymptotics certain other consistent estimators of the variance will yield the same desirable results. Obviously, however, in small samples one or the other of these may be preferred on the basis of its bias, mean square error or some other property related to the expected performance of some function of the estimator.

Consider the special case of a U-statistic of degree two. This choice is simply for definiteness and to ease the notational burden. The issues and approaches developed extend to U-statistics of degree greater than two. Let $X_1, X_2, \ldots, X_n$ be iid random variables ($n \geq 2$) with common distribution

function F. Also let $h(x,y)$ be a real-valued function, symmetric in its arguments with $E_F[h(X_1, X_2)] = \gamma$. The function h is usually called the kernel and the corresponding U-statistic of degree 2 is

$$U_n = \binom{n}{2}^{-1} \sum_{i<j}\sum h(X_i, X_j) \quad .$$

If we assume in addition that $E[h^2(X_1,X_2)] < \infty$ and $h_1(X_1) = E[h(X_1,X_2) \mid X_1]$ has a positive variance, $\zeta_1 > 0$, then it is known that the distribution of $(U_n - \gamma)/[\mathrm{Var}(U_n)]^{1/2}$ converges to the standard normal as $n \to \infty$. The general expression for the variance reduces to

$$\sigma_n^2 = \mathrm{Var}\ [U_n] = \frac{2(n-2)}{\binom{n}{2}} \zeta_1 + \frac{1}{\binom{n}{2}} \zeta_2 \ , \tag{1.1}$$

where $\zeta_1 = \mathrm{Cov}[h(X_1, X_2), h(X_1, X_3)]$ and $\zeta_2 = \mathrm{Var}[h(X_1, X_2)]$.

Clearly $n\sigma_n^2 \to 4\zeta_1 = 4\ \mathrm{Var}[h_1(X)]$ and Sen (1960) established the consistency of an estimator based on sample-based components that parallel the conditional expectations, $h_1(X_i)$. His estimator employs statistics

$$\hat{h}_1(X_i) = \frac{1}{n-1} \sum_{\substack{j=1 \\ j\neq i}}^{n} h(X_i,X_j) \quad , \quad i=1, \ldots, n$$

and then the sample variance of these to estimate ζ_1. The consistent estimator of the asymptotic variance due to Sen (1960) is

$$4\hat{\zeta}_1 = \frac{4}{n-1} \sum_{i=1}^{n} [h_1(X_i) - U_n]^2 \ . \tag{1.2}$$

A relationship between this estimator of ζ_1, generalized to U-statistics of degree m, and the jackknife estimator of $\mathrm{Var}[U_n]$ in that setting appears in Sen (1977). In that article the sample variance of the pseudo-values, S_J^2, is related to $\hat{\zeta}_1$ by

$$S_J^2 = m^2 \frac{(n-1)^2}{(n-m)^2} \hat{\zeta}_1 \quad . \tag{1.3}$$

With m = 2 this is the estimator used by Callaert and Veraverbeke (1981). They comment on the desirable feature that this expression reduces to the usual s^2 when the U-statistic degenerates to degree one. In the same discussion they mention that while S_J^2 is biased, it is nonnegative and that all alternatives that they found having less bias could be negative with positive probability.

The issues of bias and the choice between S_J^2 and $4\hat{\zeta}_1$ are of no practical importance in large samples. However when n=10, there is a substantial difference and furthermore in such small samples one would rely on estimates of the asymptotic variance with some hesitation. Still, there are situations in which the $\mathrm{Var}[U_n]$ is desired even though the sample size is too small for one to invoke large-sample normality comfortably. For example, there may be numerous replications of the small experiment and it may be reasonable to average the U-statistics and pool the separate variances. It is just such an application that motivates the present investigation of estimators of ζ_2.

2. An Unbiased Estimator of the Variance of the Kernel

In employing a U-statistic of degree three to test for symmetry in small samples of size n=20 and 30, Randles et al (1980) found that the approximately distribution-free studentized quantity did not yield the desired levels unless all 3 terms of $Var[U_n]$ were included. A similar finding about the nonneglible contribution from ζ_2 to the variance of Kendall's τ has been reported by Samara and Randles (1988). In this specific case they make use of the fact that $\zeta_2 = 1 - \tau^2$ and substitute the underlying statistic to produce an estimator of ζ_2.

In a recent dissertation in the Department of Religion at Duke University, Vinson (1984) reported the results of an experiment in which a random sample of individuals (n = 10) participated in a physical simulation constructed to shed some light on the controversy over whether the gospels written by Matthew and Luke were written independently but used Mark as their common source document. A basic measurement used by some biblical scholars is the proportion of the words in a passage that are different from the source document yet coincide in two later writers' versions of that passage. Consequently, using the 10 subjects who each independently edit a common source paragraph, the statistic that summarizes their pairwise rate of these so-called "minor agreements" is a U-statistic of degree 2. Even though the definition of the kernel suggests the possible use of some classical parametric models for counting variables, the natural tendency to handle text in phrases (clusters of random length) renders these inadequate. The thrust of this lack of a tractable distribution for

$h(X_i, X_j)$ is that a nonparametric approach is needed if one wishes to estimate $\zeta_2 = \mathrm{Var}[h(X_1, X_2)] = E[h^2(X_1, X_2)] - \gamma^2$.

For the remaining sections assume that $n \geq 4$. Using a straightforward U-statistic approach to this problem, define

$$\tilde{\zeta}_2 = \frac{1}{\binom{n}{2}} \sum_{i<j}\sum h^2(X_i,X_j) - \frac{1}{\binom{n}{4}} \sum_{i<j<k<\ell}\sum\sum\sum h_c^* (X_i,X_j,X_k,X_\ell), \qquad (2.1)$$

where h_c^* is a symmetrized unbiased kernel for γ^2 based on cross products, namely

$$h_c^*(X_i, X_j, X_k, X_\ell) = \frac{1}{3} [h_{ij}h_{k\ell} + h_{ik}h_{j\ell} + h_{i\ell}h_{jk}] ,$$

where h_{rs} is abbreviated notation for $h(X_r,X_s)$. It is easy to see that $\tilde{\zeta}_2$ is unbiased for ζ_2.

Rewriting (2.1) in standard form for a U-statistic of degree four yields

$$\tilde{\zeta}_2 = \frac{1}{\binom{n}{4}} \sum_{i<j<k<\ell}\sum\sum\sum [h_s^*(X_i,X_j,X_k,X_\ell) - h_c^*(X_i,X_j,X_k,X_\ell)] ,$$

where h_s^* is a symmetric kernel for the expected square,

$$h_s^*(X_i,X_j,X_k,X_\ell) = \frac{1}{6} [h_{ij}^2 + h_{ik}^2 + h_{i\ell}^2 + h_{jk}^2 + h_{j\ell}^2 + h_{k\ell}^2] .$$

It follows easily by rearranging the terms of h_s^* and h_c^* that

$$\tilde{\zeta}_2 = \frac{1}{\binom{n}{4}} \sum_{i<j<k<\ell}\sum\sum\sum k^*(X_i,X_j,X_k,X_\ell) , \tag{2.2}$$

where

$$k^*(X_i,X_j,X_k,X_\ell) = \frac{1}{6}\left[(h_{ij} - h_{k\ell})^2 + (h_{ik} - h_{j\ell})^2 + (h_{i\ell} - h_{jk})^2\right] .$$

It is obvious from this represention that $\tilde{\zeta}_2 \geq 0$ with probability 1. In this form it is also apparent that this is a natural extension of the familiar kernel of degree two for S^2. In other words, in place of $\frac{1}{2}(X_i-X_j)^2$ the essential ingredient of k^* is $\frac{1}{2}[h(X_i,X_j)-h(X_k,X_\ell)]^2$.

For the "minor-agreements" application there were 10 independent replications of the experimental editing of a common source paragraph by the 10 subjects. Using (2.2) the estimated magnitudes of the second terms in (1.1) were on the order of 20% to 40% of the first term. The 10 estimates of $\mathrm{Var}[U_n]$ are to be pooled to estimate the within-paragraph contribution to the variance of the average rate. The ultimate goal being to produce a reliable estimate of standard error to studentize the difference between this rate and rates obtained from other experimental settings. Even though n=10 may be somewhat small for the normal approximation to be adequate for any single paragraph, the averaging of 10 such U-statistics should produce adequately near-normal behavior. Hence there is a need to include the asymptotically negligible term in (1.1) at the same time that the asymptotic distribution theory is to be used. Consequently, in the interest of the conservatism of any subsequent test of

the null hypothesis of no difference in rates, the proposed estimate of $\mathrm{Var}[U_n]$ is

$$VU = \frac{2(n-2)}{\binom{n}{2}} \hat{\zeta}_1 + \frac{1}{\binom{n}{2}} \tilde{\zeta}_2 \ . \tag{2.3}$$

It seems clear that VU should be superior to the usual consistent estimator of the asymptotic variance, which will be denoted by $VA = 4\hat{\zeta}_1/n$.

From (1.2) and (1.3) it can be seen that the jackknife variance estimator, $VJ = S_J^2/n$, is larger than the first term in (2.3) by a factor of $(n-1)^3/(n-2)^3$. At n=10 this is $(9/8)^3 = 1.424$. This 42% inflation is not unreasonable in that VJ does represent an estimate of $\mathrm{Var}[U_n]$. Consequently, in this small sample setting VJ may well be competitive with VU.

Since VU does represent a greater computational burden, a simulation was performed to examine the issue of which is to be preferred. The U-statistic of degree two that is the subject of the Monte Carlo experimentation is

$$S^2 = \frac{1}{\binom{n}{2}} \sum_{i<j} \sum \frac{1}{2} (X_i - X_j)^2 \tag{2.4}$$

The design and various analyses of the simulation are reported in the next section.

Before we examine the Monte Carlo results, a fourth alternative approach should be considered. Following an approach similar to that for the construction of $\tilde{\zeta}_2$, one could define an unbiased estimator of $\zeta_1 = E[h_{12}h_{13}] - \gamma^2$. Consider

$$\tilde{\zeta}_1 = \frac{1}{\binom{n}{3}} \sum\sum\sum_{i<j<k} h_o^*(X_i, X_j, X_k) - \frac{1}{\binom{n}{4}} \sum\sum\sum\sum_{i<j<k<\ell} h_c^*(X_i, X_j, X_k, X_\ell),$$

where h_o^* is the symmetric kernel for terms with one overlapping argument,

$$h_o^*(X_i, X_j, X_k) = \frac{1}{3}[h_{ij}h_{ik} + h_{ij}h_{jk} + h_{ik}h_{jk}]$$

and h_c^* is as in (2.1). This is the natural U-statistic for ζ_1 mentioned by Callaert and Veraverbeke (1981). It is reasonably easy to construct examples for which $\tilde{\zeta}_1 < 0$. It is even more disturbing that the same such example can produce a negative estimate for σ_n^2 when $\tilde{\zeta}_1$ and $\tilde{\zeta}_2$ are substituted in (1.1). In 10,000 simulations with $U_n = s^2$, n=10 and standard normal data these outcomes occurred several times. For this reason (coupled with the fact that this unbiased estimator of σ_n^2 also had a larger MSE than VU) the estimator $\tilde{\zeta}_1$ cannot be recommended and is not included in the comparisons in Section 3.

Finally, one might consider applying the bootstrap to this variance estimation problem. This setting provides an illustration of problems for which the resampling approach is unnecessary. In other words, one can evaluate directly the functionals $\mu(F) = \gamma$ and $\sigma^2(F)$ at the empirical

distribution function, F_n. However, note that the bootstrap estimate of γ is not identical to U_n, but simply

$$\gamma = \int\int h(x,y)dF_n(x)\ dF_n(y) = \frac{1}{n^2} \sum_i \sum_j h(X_i,\ X_j)\ .$$

This is what Serfling (1980) calls a V-statistic and demonstrates its asymptotic equivalence to U_n. Some care should be taken with evaluation of the kernel whenever $i = j$. Technically this is not permissible. However, one might simply define $h(x,x) \equiv 0$. This occurs automatically for s^2 but not for other U-statistics such as the Wilcoxon Signed-rank, where $h_{ij} = I(X_i + X_j > 0) = 1$ if $X_i + X_j$ is positive and zero otherwise. Even if such technicalities are handled, it does not appear that anything fundamentally new results from the consideration of the bootstrap estimator, $\sigma^2(F_n)$. This leaves open the possibility however, that more sophisticated bootstrapping (or iterated bootstrapping) involving percentiles of studentized U-statistics may yield competitive approximate confidence intervals.

Before examining the results of some small sample experiments it should be noted that the techniques of this section can produce straightforward extensions for U-statistics of degree greater than 2. For example for $m = 3$ the unbiased estimator $\tilde{\zeta}_3$ would be similar to (2.2). The major alteration being that k^* would be of degree 6. The more important correction term would involve ζ_2. However, it is not clear that the additional computational effort will produce better intervals and tests, even though there may be an improvement in the traditional performance criteria for the point estimators of $Var(U_n)$.

3. Small Sample Efficiency and Validity

The numerical results in this section were produced by a Fortran program using IMSL routines on an IBM 3081-D at Southern Methodist University. The primary experiment involves samples of size n=10 from a standard normal for which there were 10,000 Monte Carlo repetitions. The U-statistic of interest is the sample variance (2.4). For this specific case the desired quantity is $\sigma_n^2 = \mathrm{Var}(S^2) = 2/9$. The three estimators examined are: VA, Sen's asymptotic variance estimator; VU a new estimator (2.3) having a direct unbiased estimate of the second order term; and VJ, the jackknife.

Each of the 3 estimators were evaluated for each of the 10,000 replications. By taking this natural blocking of the experiment into account the sensitivities of the various comparisons among the competing estimators is better than the usual summary tables of means and standard errors would suggest. To capitalize on this matching condition it is necessary to accumulate magnitudes and directions of the errors in each estimator as well as pairwise differences in their squares. These and other performance criteria from 1000 subsets of 10 pooled samples were also analyzed for other sample sizes and distributions, but only the n=10, normal case is reported in detail.

The usual simple comparison of the variance estimators indicates that 1) the bias in VA is in the anticipated undesirable direction and 2) there

may be some advantage to VU, since its expectation is closer to the true value without a substantial penalty in mean square error (MSE).

Estimator	MSE	Bias
VA	.056	-.026
VU	.065	-.004
VJ	.090	.025

The difference between the MSEs of VU and VJ is statistically significant. The average of 10,000 paired differences in squared errors differs from zero by approximately 17 standard errors. However, for any single random sample of size 10 the estimated standard deviation of this difference is approximately .15. Therefore, as always, one may wish to look beyond the squared error criterion.

With regard to the familiar notion of Pitman closeness, the advantage appears to switch from VU to VJ. The estimated $P[\,|VU - \sigma_n^2| > |VJ - \sigma_n^2|\,]$ = .60. Most of these cases of VJ closer than VU occur when $VU < VJ < \sigma_n^2$ (about 58 out of 60). The second most frequent occurence is $\sigma_n^2 < VU < VJ$ (about 32%). Indeed, $VU > VJ$ only about 6% of the time, which would be consistent with a more conservative performance of approximate tests based on VJ.

Perhaps the most relevant criterion for comparison of these variance estimators is the actual significance level of the associated test that might typically be employed. A conventional approximate test of H_0: $\gamma = \gamma_0$ vs. H_1: $\gamma \neq \gamma_0$ is to compare the magnitude of $Z = (U_n -$

$\gamma_0)/(\text{est. Var})^{\frac{1}{2}}$ to 1.96. Let ZA, ZU and ZJ represent such statistics corresponding to VA, VU and VJ, respectively. As the figures below demonstrate, it is seldom advisable to rely on large sample normality at n=10. The actual significance levels may be in the 16 - 20% range rather than close to the nominal .05. However, the situation is considerably better for the application discussed in the previous section. When the U-statistics and their variance estimates were pooled across 10 independent subgroups each of size n=10 the actual levels are noticeably smaller. In the table below error rates for two independent sets of 10,000 are summarized for each situation. The maximum standard error for the unpooled rates is < .004. For the pooled case there are 1000 realizations and at the nominal .05 rate the standard errors are approximately .007.

Test	Unpooled		Pooled	
ZA	.189	.192	.076	.088
ZU	.172	.177	.065	.072
ZJ	.163	.166	.056	.051

Combining the 2000 realizations of the pooled test, the level for ZJ is not significantly great than .05 and the other two are. An exact version of McNemar's test for correlated proportions [Lehmann (1975), p. 268-9] confirms the significance of the difference between ZU and ZJ. There were 30 instances in which the two tests produced different conclusions and all 30 were that ZU erroneously rejected and ZJ did not.

For the same U-statistic but an exponential population rather than a normal the outcomes were similar. Although VU had significantly smaller MSE than VJ, VJ was Pitman closer about 77% of the time and the levels the

test based on pooled samples were significantly different favoring the jackknife. As one might expect, in the presence of greater skewness the ultimate error rates were not as close to the nominal as in the normal case.

For the Wilcoxon signed-rank statistic the fundamental U-statistic estimates $P(X_i + X_j > 0)$. For distributions that are symmetric about 0, $\gamma = \frac{1}{2}$, and the same grouping and sample sizes were examined in this setting with normal data. Here VU had the smallest MSE of the three. In contrast to the previous two examples, VU was Pitman closer than VJ in about 65% of the samples. Nevertheless, consistent with the earlier cases of ZJ was the superior approximate test. The observed error rates for the pooled versions of ZA, ZU and ZJ were. .092, .076 and .054, respectively.

For examination of the bootstrap estimators of $\mathrm{Var}[U_n]$, the associated V-statistics and an unbiased modification see Lee (1985). In that article it may be seen that these two estimators and VJ have variances that are equal up to $O(n^{-3})$. However, the three differ with respect to bias. In a small simulation study reported there, one sees that while VJ may not be the preferred point estimator of $\mathrm{Var}[U_n]$, it yields a studentized quantity with better validity of approximate confidence intervals for Kendall's τ.

For certain specific U-statistic problems there will undoubtedly be special approaches that may be better than the general approach of studentizing $U_n - \gamma$ with $VJ^{\frac{1}{2}}$. For example much has been written about the effectiveness of approximate symmetrizing or variance stabilizing transformations, such as $\log(s^2)$. However, in the present study we have

elected not to examine such proposals because they are not generally available. The conclusion from the studies reported here is that the new unbiased estimator, $\tilde{\zeta}_2$, in (2.2) provides a worthwhile improvement over VA. At the same time it must be noted that the ordinary jackknife estimate, VJ, is easier to compute and leads to superior approximate tests.

REFERENCES

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

Helmers, R. (1985). The Berry-Esseen bound for studentized U-statistics, Cand. J. Statist., 13, 79-82.

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

Lee, A.J. (1985). On estimating the variance of a U-statistic, Commun. Statist. 14, 289-301.

Lehmann, E.L. (1975). Nonparametrics: Statitical Methods Based on Ranks, Holden-Day, San Francisco.

Randles, R.H., Fligner, M.A., Policello, G.E., and Wolfe, D.A. (1980). An asymptotically distribution-free test for symmetry versus asymmetry, J. Amer. Statist. Assoc., 75, 168-172.

Randles, R.H. and Wolfe, D.A. (1979). Introduction to the Theory of Nonparametric Statistics, Wiley, New York.

Samara, B. and Randles, R.H. (1988). A test for correlation based on Kendall's tau, Commun. Statist., 17, (to appear).

Sen, P.K. (1960). On some convergence properties of U-statistics, Cal. Statist. Assoc. Bull., 10, 1-18.

Sen, P.K. (1977). Some invariance principles relating to jackknifing and their role in sequential analysis, Ann. Statist., 5, 316-329.

Serfling, R.J. (1980). Approximation Theorems in Mathematical Statistics, Wiley, New York.

Vinson, R.B. (1984). The Significance of the Minor Agreements as an Argument Against the Two-Document Hypothesis, unpublished Ph.D. dissertation, Department of Religion, Duke University.