# PROBABILITY SAMPLE U-STATISTICS: THEORY AND APPLICATIONS FOR COMPLEX SAMPLE DESIGNS

Ralph E. Folsom, Research Triangle Institute

## 1. Introduction

This paper summarizes research presented in my 1984 dissertation. The dissertation (Folsom, 1984) was devoted to the development of probability sample U-statistics theory. Applications of the theory were made to solve variance and variance component estimation problems for complex sample designs. In the domain of classical statistics where sampling from probability distributions is assumed, U-statistics theory has played an important role in the area of robust nonparametric inference. Variance components estimation has been one classical area where U-statistics theory has been applied to good advantage. By extending U-statistics theory into the realm of complex probability samples, the unbiased Yates-Grundy-Sen variance estimator and associated variance component estimators are identified as degree 2 probability sample U-statistics. Considering the central role that variance and variance component estimates play in probability sample design and inference, the associated U-statistics theory provides a valuable new research and analysis tool for survey statistics.

## 2. Unbiased Covariance Estimators for U-Statistics

In the initial development, the need for new unbiased covariance estimators for U-statistics is recognized. Due to the deeply stratified nature of the typical probability sample, stratum specific sample sizes will often be too small to justify existing large sample variance approximations. For a stratified sample with $n_h = 4$ independent normal selections per stratum, the variance of the sample variance estimator is underestimated by 33 percent using Sen's large sample variance approximation. The standard jackknife variance estimator for this example results in a 51 percent overestimate. The unbiased U-statistic covariance estimator developed in Chapter 2 of the dissertation is analogous to the Yates-Grundy-Sen (YGS) variance estimator for degree 1 Horvitz-Thompson statistics. As in the degree 1 case, the YGS variance estimator for a degree m U-statistic requires computation of degree 2m joint sample inclusion probabilities. For the special case of with replacement sampling an unbiased U-statistic variance estimator is developed that requires no computation of higher order variance weights. In the minimum sample size case (n = 4) for degree m = 2 variance estimability, the with replacement variance estimator has a simple form reminiscent of the squared difference estimator for paired with replacement selections. With $f_a$ depicting a degree 2 kernel, the with replacement variance estimator for the n = 4 special case is computed as the simple average of three terms of the form

$$(f_a - \bar{f})(\bar{f} - f_{s-a}) \qquad (1)$$

where $\bar{f}$ is the U-statistic estimator and $f_{s-a}$ is the degree 2 kernel based on the complementary half-sample s-a. Note that for n = 4 and m = 2, there are $\binom{n}{2} = 6$ degree 2 (half-sample) kernels and 3 half-sample/complement pairs.

## 3. Multistage Sample Results

In Chapter 3 of the dissertation, the single stage design results of Chapter 2 were extended to multistage probability samples. For primary sample designs involving nonreplacement or minimum replacement selections, an extension of Durbin's theorem for unbiased degree m = 1 multistage variance estimation was required. Given the development of parallel notation, the associated degree m version of Durbin's theorem proves to be directly analogous to the general degree m = 1 result. For the special case of with replacement primary samples, the variance estimator based on a copy of the single stage form in equation (1) is unbiased; that is, with the estimated psu-level kernel $\hat{f}_a$ based on subsequent stages of sampling replacing $f_a$ in (1), then equation (1) provides an unbiased vari-ance estimator for the multistage U-statistic $\bar{f}_{II}$. As in the degree m = 1 case, it is shown that for with replacement primary selections this copy of the first stage variance form

is unbiased without further correction for within psu vari-ance contributions.

4. Nonlinear Functions of U-Statistics

Chapter 4 of the dissertation explores the issue of variance estimation for nonlinear func-tions of probability sample U-statistics. For a general class of probability sample esigns including with replacement and minimum replacement selections as defined by Chromy (1979), the Taylor series or delta method variance approximation is shown to be asymptotically unbiased. As in the degree m = 1 case, it is shown that the delta method approximation can be formed by replacing the sample kernels $f_a$ in the linear function variance formula by a linearized kernel, say $\ell_a$, where

$$\ell_a \equiv \hat{Q} \underset{\sim}{f}_a \qquad (2)$$

and $\hat{Q}$ is the row vector of estimated first order partial derivatives of the non-linear function with respect to the vector of U-statistics. The elements of the column vector $\underset{\sim}{f}_a$ in equation (2) are the kernels $f_{at}$ associated with the t-th constituent U-statistic, say $U_t$. The jackknife and balanced repeated replication (BRR) sample reuse methods for nonlinear variance approximation were also explored. The existing versions of these sample reuse methods do not allow one to properly account for the effects of non-replacement or minimum replacement sampling when the sampling fractions vary from stratum to stratum. Furthermore, for unequal probability sample designs only the with replacement variance approximation can be reproduced when these sample reuse methods are applied to linear statistics. The BRR method has also been restricted in the past to designs with an equal number of sampling units per stratum. To remove these design restrictions and simultaneously extend sample reuse methods to general probability sample U-statistics, a new class of pairwise jackknife and BRR variance estimators was developed. If (ab) labels distinct pairs of degree m kernels, the YGS variance estimator for a single stage U-statistic $\bar{f}$ has the form

$$\mathrm{var}(\bar{f})_{YGS} = \sum_{a \varepsilon s_m} \sum_{b>a} w_{ab} (\tilde{F}_a - \tilde{F}_b)^2 \qquad (3)$$

where $\tilde{F}_a \equiv \binom{n}{m}^{-1} f_a$ is the sample size scaled degree m kernel. If the number of kernel pairs with a > b is denoted by $R = \binom{n}{m}[\binom{n}{m} - 1]/2$, then the pairwise BRR estimator is based on an orthogonal design matrix with +1 and -1 elements of dimension greater than R. Letting $\Lambda$ depict a selection of R contrast columns from the smallest such orthogonal design matrix with dimension A greater than R, the new BRR replicates can be depicted by

$$\underset{\sim}{f}_{brr} \equiv \bar{f} J_A + \Lambda W^{\frac{1}{2}} \underset{\sim}{e} \qquad (4)$$

where $\underset{\sim}{e}$ is an R x 1 column vector with elements

$$e(ab) \equiv (\tilde{F}_a - \tilde{F}_b).$$

For the special case considered here, we assume that the $w_{ab}$ variance weights are all positive. This assumption is satisfied for the degree m = 1 case by any number of nonreplacement unequal probability selection schemes. With this assumption, $W^{\frac{1}{2}}$ can be identified as an R x R diagonal matrix with $(w_{ab})^{\frac{1}{2}}$ on the diagonal. The vector $J_A$ in equation (4) denotes an A x 1 column vector of 1's, and $\Lambda$ is an A x R matrix whose columns are orthogonal contrasts. Since the columns of $\Lambda$ are contrast vectors, (+1's and -1's that sum to zero), it is clear that the A replicate statistics in $\underset{\sim}{f}_{brr}$ (A x 1) average to the full sample U-statistic $\bar{f}$. This follows from the fact that

$$\begin{aligned} \bar{f}_{brr} &\equiv (J_A^T \underset{\sim}{f}_{brr} \div A) \\ &= \bar{f} + J_A^T \Lambda W^{\frac{1}{2}} \underset{\sim}{e} \div A \\ &= \bar{f} + 0. \end{aligned}$$

The simple mean square among these new BRR kernels is

$$\begin{aligned} \mathrm{var}(\bar{f})_{brr} &= (\underset{\sim}{f}_{brr} - \bar{f} J_A)^T (\underset{\sim}{f}_{brr} - \bar{f} J_A) \div A \\ &= \underset{\sim}{e}^T W^{\frac{1}{2}} \Lambda^T \Lambda W^{\frac{1}{2}} \underset{\sim}{e} \div A \qquad (5) \end{aligned}$$

The fact that the columns of $\Lambda$ are orthogonal contrast vectors of +1's and -1's leads to the result

$$\text{var}(\bar{f})_{brr} = \underset{\sim}{e}^T W \underset{\sim}{e}$$
$$= \sum_{(ab)} w_{ab}\, e(ab)^2$$
$$\equiv \text{var}(\bar{f})_{YGS} \qquad (6)$$

The results in equations (5) and (6) confirm that the new BRR replicate equations in (4) lead to a sample reuse variance form that reproduces the unbiased YGS variance estimator for linear func-tions of U-statistics. Note that this result is true for any n. One can also produce the complementary set of BRR replicate estimators $\underset{\sim}{f}^*_{brr}$ by subtracting the residual vector $\Lambda\, W^{\frac{1}{2}}\, \underset{\sim}{e}$ from $\bar{f}\, J_A$. These complementary BRR replicates can be used to form $\text{var}(\bar{f})^*_{brr}$ which is averaged with equation (6). Alternately, one can form the BRR difference variance estimator

$$\text{var}(\bar{f})^D_{brr} = (\underset{\sim}{f}_{brr} - \underset{\sim}{f}^*_{brr})^T (\underset{\sim}{f}_{brr} - \underset{\sim}{f}^*_{brr}) \div 4A \qquad (7)$$

It is easy to see that equation (7) is also equivalent to the YGS estimator. The BRR estimators in (5) and (7) are easily shown to duplicate the delta YGS variance estimator for linear functions of U-statistics. The equivalence of the new BRR approximation and the delta-YGS estimator extends to quadratic functions of U-statistics when the BRR-D estimator in equation (7) is employed.

5. U-Statistics Applications

Three areas of application were illustrated in Chapter 5 of the dissertation. The first application demonstrates the utility of probability sample U-statistics theory as a research tool for exploring the small sample properties, bias and mean squared error, of a new class of YGS ratio estimators. To improve the stability of YGS variance estimators, the following ratio adpatation was proposed for the degree m = 1 case

$$\text{var}(\hat{Y}_+)_{YGS\text{-}R} = [1 - n \sum_{k=1}^{N} \phi_k^2]\, S^2_{\Omega y} / n \qquad (8)$$

where $\phi_k = \pi_k/n$ is the single draw selection probability for population unit k, and

$$S^2_{\Omega y} = \sum_{i=1}^{n} \sum_{j>i} \Omega_{ij}\, (y_i - y_j)^2 / 2$$

is the weighted average of the degree 2 variance kernels $(y_i - y_j)^2/2$ with $y_i = (Y_i/\phi_i)$ depicting the single draw variate for sampling unit i. The weights $\Omega_{ij}$ for averaging these kernels are proportional to the YGS variance weights $w_{ij} = [(\pi_i\, \pi_j \div \pi_{ij}) - 1]$; that is,

$$\Omega_{ij} \equiv w_{ij} \div [\sum_{i=1}^{n} \sum_{j>i} w_{ij}] .$$

The term in square brackets in equation (8) is proportional to the expected value of the YGS variance weight sum in the denominator of $S^2_{\Omega y}$. The U-statistics theory developed in the dissertation provides a mechanism for approximating the bias and mean squared error of equation (8) and the variance of the competing YGS variance estimator. Such an analytical evaluation of the small sample properties of these variance estimators would be an efficient alternative to Monte-Carlo methods.

The second area of application explored treats the problem of robust confidence interval estimation for probability sample statistics. An approximation for the degrees of freedom associated with the sample t statistic

$$t = (\hat{\theta} - \theta) \div [\text{var}(\hat{\theta})]^{\frac{1}{2}}$$

is developed. The approach proposed is to use U-statistics theory to estimate the variance of the observed noncentral t statistic

$$t_* = \hat{\theta} \div [\text{var}(\hat{\theta})]^{\frac{1}{2}} \qquad (9)$$

Having obtained the estimator $\text{var}(t_*)$, this statistic is equated to the variance function of the noncentral t. The square of $t_*$ is similarly equated to $[E(t_*)]^2 + \text{Var}(t_*)$. The resulting two equations are then solved for the noncentrality parameter $\delta \equiv \theta \div [\text{Var}(\theta)]^{\frac{1}{2}}$ and the desired degrees of freedom parameter.

The final U-statistics application yields a variance and covariance component model for a vector of subpopulation proportions $\hat{\underset{\sim}{P}}$ from a complex two stage sample design. The design has a constant second stage sample size ($\bar{m}$) selected from each primary unit. The observed sample size for domain-d across all psus is denoted by

$m_d$. The following design effect estimator is developed for the covariance matrix of $\hat{\underset{\sim}{P}}$

$$\text{cov}(\hat{\underset{\sim}{P}}) = \hat{V}_M \, [I + \text{cvw}^2] \, \hat{Q} \, [I + \bar{m} \, \hat{\alpha} + (\bar{m} - 1) \, \hat{\Omega}] \tag{10}$$

where $\hat{V}_M$ estimates the simple random sampling diagonal covariance matrix with diagonal elements

$$\hat{V}_M(d) = \hat{P}_d \, (1 - \hat{P}_d) \, / \, m_d.$$

The second quantity in square brackets is the unequal weighting design effect with $\text{cvw}^2$ denoting a diagonal matrix with diagonal elements representing the squared coefficent of variation for the sample weights associated with domain d members. The third multiplicative design factor in (10) is the diagonal matrix $\hat{Q}$ with elements

$$\hat{Q}_d \equiv \{ \, [\hat{PW}_d \div \hat{P}_d] + [(1 - \hat{PW}_d) \div (1 - \hat{P}_d)] - 1 \, \}$$

where $\hat{PW}_d$ is a version of the domain d p-value computed as a weighted average of the one-zero Y variate using squared sample weights $W^2$. This quantity measures the effect of optimum specification of the unequal selection probabilities. These $\hat{Q}_d$ quantities are less than one and therefore effect variance reduction when domain d members belonging to the rarest Y variate response level are oversampled. That is, when $P_d < .5$ then the parameter $Q_d < 1$ when domain d members with $Y = 1$ are overrepresented.

The matrix $\hat{\alpha}$ measures the design effects of primary unit stratification and nonreplacement psu selection. Since both of these factors reduce variance, quadratic forms in $\hat{\alpha}$ are expected to be negative. The $\hat{\Omega}$ matrix combines the effect of within psu clustering, second stage stratification and nonreplacement sampling. Quadratic forms in $\hat{\Omega}$ will be positive when variance inflating effects of clustering dominate the variance reducing effects of second stage stratification and nonreplacement selection. When the psus are large and heterogeneous an effectively stratified second stage sample can dominate the effect of psu clustering so that quadratic forms in $\hat{\Omega}$ are negative. Simple matrix analogs of analysis of variance type mean squares were developed for estimating $\alpha$ and $\Omega$.

For the two stage unequal probability design considered, the $\hat{\alpha}$ and $\hat{\Omega}$ estimators are such that the design effect form in (10) is equivalent to the asymptotically unbiased delta method variance estimator based on the YGS estimation formula. To improve the stability of subpopulation variance and covariance estimates, it was proposed that the composite design effect component matrices $\hat{\alpha}$ and $\hat{\Omega}$ be pooled across multiple binary response variables Y. The U-statistics theory provides a formal mechanism for testing the equivalence of matrices $\hat{\Omega}_t$ associated with different outcomes $Y_t$.

6. Recommendations for Related Research

Turning to recommendations for related research, the applications illustrated in section 5 suggest a number of empirical investigations to test the performance of proposed methodologies. The new pairwise BRR sample reuse methodology as applied to degree 2 statistics also suggests a robust confidence interval strategy that should be explored. The degree 2 version of this methodology can be applied to the delta variance based sample t statistic to produce replicate t values of the form

$$t_r = (\hat{\theta}_r - \hat{\theta}) \div [\text{var}_\Delta(\hat{\theta})_r]^{\frac{1}{2}}.$$

While it might seem more natural to employ a BRR variance estimator in the replicate t statistics $t_r$, the methodology proposed requires that the full sample t statistic have the form of a nonlinear function of U-statistics. Unlike $\text{var}_\Delta(\hat{\theta})$, the full sample delta variance estimator, the full sample BRR variance approximations $\text{var}_{BRR}(\hat{\theta})$ are not in the form of nonlinear U-statistic functions.

Letting $t_{\alpha/2}$ and $t_{(1-\alpha/2)}$ denote the lower and upper $\alpha/2$ percentage points of the empirical cdf derived from the $t_r$ values, the robust nonsymmetric $(1 - \alpha)$ level confidence interval is

$$\hat{\theta}_U = \hat{\theta} - t_{\alpha/2} \, [\text{var}_\Delta(\hat{\theta})]^{\frac{1}{2}}$$

and

$$\hat{\theta}_L = \hat{\theta} - t_{(1-\alpha/2)} \, [\text{var}_\Delta(\hat{\theta})]^{\frac{1}{2}} .$$

By using the degree 2 version of the pairwise BRR algorithm, this BRR-bootstrap method produces an empirical cdf for t with a variance that properly accounts for the degree 2 nature of $\text{var}_{\Delta}(\hat{\theta})$. Rao and Wu (1984) have proposed a degree 1 bootstrap algorithm for unequal probability samples that bases replicate statistics $\hat{\theta}_r$ on with replacement samples from the n(n-1) ordered sampling unit pairs (ij). Specifically, each of Rao and Wu's bootstrap replicate statistics requires a with replacement sample $\not\not\not_2(r)$ of m = n(n-1) pairs. With $W_{ij}$ denoting the YGS variance weights for a linear Horvitz-Thompson total estimator $\hat{Y}_{\sim HT}$, the associated replicate replicate totals have the form

$$\hat{Y}_{\sim r} = \hat{Y}_{\sim HT} + \sum_{(ij)\varepsilon \not\not\not_2(r)} W_{ij}^{\frac{1}{2}} (\pi_i^{-1} Y_{\sim i} - \pi_j^{-1} Y_{\sim j}) .$$

The empirical cdf of the replicate statistics $\hat{\theta}_r = F(\hat{Y}_{\sim r})$ is then used to define bootstrap intervals.

A sample simulation study is needed to contrast the actual confidence level and expected half-width of these alternative bootstrap intervals with standard symmetric t intervals. A symetric t interval based on the robust df estimate proposed in Section 5 would be an interesting competitor.

In addition to the empirical studies suggested here, a U-statistics central limit theorem for nonreplacement and minimum replacement designs should be developed. The asymptotic framework utilized by Madow (1945), Hájek (1964), and Fuller (1975) can be employed to develop the desired result. This framework assumes a sequence of populations of size $N_t$ and associated sample sizes $n_t$ with the property that as $n_t \to \infty$ the corresponding sampling fractions $f_t \equiv (n_t \div N_t)$ converge to $f < 1$. Rick Williams, an RTI colleague, is currently working with Professor P. K. Sen to develop the required central limit theory.

## 7. Dissertation Reference List


Arvesen, J. N. (1969). Jackknifing U-statistics. Annals of Mathematical Statistics 40: 2076-2100.

Bayless, D. L. and J. N. Rao. (1970). An empirical study of stabilities of estimators and variance estimators in unequal probability sampling (n = 3 or 4). Journal of the American Statistical Association 65: 1645-1667.

Brewer, K. W. R. (1963). A model of systematic sampling with unequal probabilities. Australian Journal of Statistics 5: 5-13.

Chromy, J. R. (1979). Sequential sample selection methods. American Statistical Association 1979 Proceedings of the Section on Survey Research Methods: 401-406.

Cramér, H. (1946). Mathematical Methods of Statistics. Princeton University Press: 254.

Durbin, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. Journal of the Royal Statistical Society B15: 262-269.

Durbin, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. Applied Statistics 16: 152-164.

Fellegi, I. P. (1963). Sampling with varying probabilities without replacement: Rotating and non-rotating samples. Journal of the American Statistical Association 58: 183-201.

Folsom, R. E., D. L. Bayless, and B. V. Shah. (1971). Jackknifing for variance components in complex sample survey designs. American Statistical Association, 1971 Proceedings of the Social Statistics Section: 36-39.

Folsom, Ralph E., Jr. (1980). U-statistics estimation of variance components for unequal probability samples with nonadditive interviewer and respondent errors. American Statistical Association 1980 Proceedings of the Section on Survey Research Methods: 137-142.

Folsom, R. E., R. L. Williams. (1981). Design Effects and the Analysis of Survey Data, Research Triangle Institute, Research Triangle Park, North Carolina.

Folsom, R. E., (1984). Probability Sample U-Statistics: Theory and Applications for Complex Sample Designs. Institute of Statistics Mimeo Series No. 1464, Chapel Hill, N.C.

Fuller, W. A. (1975). Regression analysis for sample survey. Sankya, ser. C. 37: 117-132.

Gray, G. B. (1975). Components of variance model in mutli-stage stratified sample. Survey Methodology 1: 27-43.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. Annals of Mathematical Statistics 35: 1491-1523.

Hansen, M. H., and W. N. Hurwitz. (1943). On the theory of sampling from finite populations. Annals of Mathematical Statistics 14: 33-362.

Hartley, H. O. and J. N. K. Rao. (1962). Sampling with unequal probabilities and without replacement. Annals of Mathematical Statistics 33: 350-374.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. Annals of Mathematical Statistics 19: 293-325.

Horvitz, D. G., and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47: 663-685.

Isaki, Cary T. (1983). Variance estimation using auxiliary information. Journal of the American Statistical Association 78: 117-123.

Madow, W. G. (1945). On the limiting distributions of estimates based on samples from finite universes. Annals of Mathematical Statistics 19: 535-545.

Quade, Dana. (1967). Nonparametric partial correlation. Institute of Statistics Mimeo Series No. 526, Chapel Hill, N.C.

Rao, J. N. K., A. J. Scott. (1981). The analysis of categorical data from complex sample surveys: Chi-squared test of goodness of fit and independence in two-way tables. Journal of the American Statistical Association 76: 221-230.

Rao, J. N. K. and C. F. J. Wu. (1983). Inference from stratified samples: Second order analysis of three methods for nonlinear statistics. Journal of the American Statistical Association, 80: 620-630.

Rao, J. N. K. and C. F. J. Wu. (1984). Bootstrap inference for sample surveys. American Statistical Association 1984 Proceedings of the Section on Survey Research Methods: 106-112.