U-STATISTICS IN UNEQUAL PROBABILITY SAMPLES

Rick L. Williams, Research Triangle Institute and P. K. Sen, UNC-Chapel Hill
Rick L. Williams, P.O. Box 12194, Research Triangle Park, NC 27709

1. INTRODUCTION

U-statistics are an important class of estimators. They arise as a generalization of the sample mean, or of forming an average. The large sample invariance properties of this class are well understood for independent and identically distributed sequences of observations. The research described below extended some of these results to unequal probability without replacement sampling from a finite population.

First, consider a sequence $\{X_i;\ i\geq 1\}$ of independent and identically distributed (i.i.d.) random variables each having a distribution function (d.f.) F. Let $\mathbf{F}$ be the space of all d.f.'s belonging to some specified class. A homogeneous statistical function of degree m (≥ 1) is given by

$$\theta = \int \cdots \int g(x_1,\ldots,x_m)dF(x_1)\ldots dF(x_m)$$

$$= E_F[g(X_1,\ldots,X_m)] \quad \text{for all } F \text{ in } \mathbf{F},$$

where $\mathbf{F} = \{F : |\theta(F)| < \infty\}$ and g is a Borel measurable function called a kernel. As is usual in this situation, assume without loss of generality that g is a symmetric function of its m arguments. If there exists a symmetric kernel g of degree m for which the above holds, then θ is termed an estimatable parameter.

Hoeffding (1948) introduced an unbiased estimator, called a U-statistic, of θ. Suppose that $n \geq m$, then U_n is an unbiased estimator of θ given by

$$U_n = \binom{n}{m}^{-1} \sum_{1\leq i_1 < \cdots < i_m \leq n} g(X_{i_1},\ldots,X_{i_m})$$

Hoeffding introducted the projection method to establish the large sample distribution of U_n. This is the same basic approach adopted for use with unequal probability sampling.

The projection of a U-statistic

$$\hat{U}_n = \sum_{i=1}^{n} E[U_n|X_i] - (n-1)\theta \ .$$

Hoeffding demonstrated that the random variables

$$Y_n = \sqrt{n}\ (\hat{U}_n - \theta) \quad \text{and} \quad Z_n = \sqrt{n}\ (U_n - \theta)$$

have the same limiting distribution by showing that Y_n and Z_n are equivalent in quadratic mean as n trends to infinite.

The advantage of considering $\hat{U}_n$ is that it is a simple degree one summation rather than a degree m statistic. When the original observations are independent and identically distributed, $\hat{U}_n$ is just the sum of n i.i.d. random variables. Thus, the well developed theory for the mean or total of i.i.d. random variables extends to the degree m statistic U_n.

Because of the shortness of space, a very brief summary of the theoretically results are given next. A detailed presentation of a simulation study is then presented.

The case of equal probability sampling without replacement from a finite population has been well covered by Nandi and Sen (1963), Sen (1960) and Sen (1972). It was shown that, for simple random sampling without replacement from a finite population, there is convergence to a Brownian bridge process.

Folsom (1984) extended the class of U-statistics to unequal probability samples. He defined the U-statistic population parameter and its estimator. In 1988, Williams determined the project of a U-statistic for unequal probability samples selected with less than full replacement. He then demonstrated that a U-statistic and its projection were equivalent in quadratic mean when sampling without replacement. This was done under the assumption of "uniform asymptotic negligibility." In addition, it was assumed that certain ratios of products of the expected sample inclusion frequencies, with the same sampling units in both the numerator and denominator, approach one. With this result, central limit theorems for linear statistics from unequal probability samples can be extended to U-statistics.

2. NUMERICAL SIMULATION

The results described above demonstrate that the distribution of a U-statistic estimated from an unequal probability sample converges to a normal law as the sequence of samples and populations become infinitely large. In practice, we are allowed only one such population and sample. The advisability of using the normal distribution as a basis for inference in this situation is assessed in the numerical simulation presented here. The simulation proceeds by selecting 1,000 independent samples using Sampford's method from a population of U.S. counties. Two different U-statistics are estimated from each sample and the empirical distribution of the 1,000 estimates for each statistic is compared with a normal distribution.

The two U-statistics used in this simulation -- Kendall's rank correlation and the with replacement variance component -- differ in that one has a narrow constricted range, while the other may take on any finite nonnegative value. The rank correlation is more likely to benefit from nearly equal probabilities of selection, while the variance component is more suited to unequal selection probabilities proportional to the size of the observation. Three sets of size measures, used to generate the selection probabilities, are considered. The first set covers a very wide range and is related to the size of the observations. The other two sets of size measures cover progressively smaller ranges.

2.1 The Data

The data for this simulation are taken from the 1986 Area Resource File (ARF) obtained from

the Health Resources and Services Administration of the U.S. Public Health Service (1987). The ARF contains one record for each of the 3,080 counties in the U.S., a subset of which will serve as the study population, with the counties being the sampling units. The 1980 U.S. Census population of each county, the 0.57 root of the population and the natural logarithm of the population are used as the size measures for selecting the samples. The 1984 count of the number of short-term general hospitals and short-term general hospital beds in the county are the analysis variables. The study population was restricted to the 2,000 smallest counties, as measured by their 1980 population count, to eliminate the possibility of large self-representing (selection probability greater than one) units and to control the cost of selecting the samples.

Even though only the 2,000 smallest counties are included, they are still very diverse in total population size. They range in size from 91 persons to 35,376 with a ratio of largest to smallest of 389. The other two size measures are used to generate samples from a less diverse set of probabilities. The 0.57 root transformation was chosen to provide a set of size measures that had an approximate 30 to 1 range of sizes, while the log transformation provides a set of sizes measures with a less than 2.5 to 1 range. These two transformations of the size measures were mainly chosen to restrict the variability of the selection probabilities and not to determine an optimum size measure.

2.2 Design of the Simulation

Six sets of 1,000 independently replicated samples were drawn from the 2,000 study counties using the rejective version of Sampford's method (1967). The six sets result from using each of the three size measures with sample sizes of 50 and 100. We attempted to consider sample sizes greater than 100, but, as the sample size increases, the rejective version of Sampford's method almost always rejects a candidate with replacement sample because of at least one duplicated unit in the sample.

For each of the 6,000 samples, two U-statistics were estimated. The first U-statistic considered was Kendall's (1938) rank correlation, often called τ_a. A general description of this statistic is given in Kendall (1970) or in Conover (1980). To formulate this as a U-statistic, consider two bivariate observations -- (X_1,Y_1) and (X_2,Y_2). The kernel for τ_a is

$$g\left[\left(X_1,Y_1\right),\ \left(X_2,Y_2\right)\right] = \binom{2000}{2}^{-1} \operatorname{sgn}\left(X_1 - X_2\right) \operatorname{sgn}\left(Y_1 - Y_2\right)$$

with

$$\operatorname{sgn}(z) = \begin{cases} -1 & \text{if } z < 0 \\ 0 & \text{if } z = 0 \\ 1 & \text{if } z > 0 \end{cases} .$$

This kernel will be averaged over all pairs of observations to find the rank correlation between the number of hospitals and the number of hospital beds in a county.

The second U-statistic used in this simulation is the with replacement variance component given by

$$\sigma^2_{wr} = \sum_{a=1}^{N} \phi(a)\left[y_a/\phi(a) - Y_+\right]^2$$

where Y_+ is the population total and $\phi(a)$ is the relative size measure of the a-th county. This statistic is useful in designing new surveys and is discussed by Folsom (1984). Divided by the sample size, this is the variance of the estimated total under with replacement sampling. This statistic was chosen because it was likely to benefit from probability proportional to size sampling if the analysis variable, Y, is proportional to the size measure. The with replacement variance component will be found for the number of hospital beds in a county. A kernel for this statistic is

$$g(a,b) = \phi(a)\ \phi(b)\left[y_a/\phi(a) - y_b/\phi(b)\right]^2 .$$

In the following, the relative with replacement variance component will be reported. This is obtained by dividing by the squared total, Y_+^2. This does not effect the results, but simplifies the presentation.

2.3 Findings

Table 1 displays the mean and standard error over all of the estimates for the rank correlation and relative with replacement variance component. Notice that, in every case, the mean estimate is very close to the true value (also in the table). Next, note that, for the rank correlation, the standard error of the mean estimate decreases as the size measure is changed from the total population size, to the root population size, to the log of the population size. The converse is true for the relative with replacement variance component. In the latter case, the relative standard error is generally smallest for the total population size measure, intermediate for the root population size measure, and largest for the log population size measure. This fits with the general intuition since the kernel for the rank correlation only takes on the values -1, 0, and 1, which will be more "proportional" to the less diverse transformed size measures. On the other hand, since the number of hospital beds in a county should grow with the population of the county, we expect that the with replacement variance component should enjoy the benefit of proportional to size sample selection.

Next, as a first attempt to assess how "normal" the distribution of the estimates is, the skewness and kurtosis coefficients of each set of 1,000 estimates is presented in Table 2. Recall that skewness measures how symmetric the distribution is with the normal distribution having a skewness of 0. Kurtosis measures how "heavy" the tails of the distribution are with the normal distribution having a kurtosis of 3. For Kendall's rank correlation with the log population size measure, the skewness and kurtosis of the sample estimates are very close

the values for a normal distribution for both sample sizes of 50 and 100. They move progressively further from the normal distribution values as the size measure is changed to the root of population size to the total population size. This again fits with the preliminary expectations. Turning to the with replacement variance component estimates, we see that all of the coefficients differ markedly from the normal distribution values. However, as expected, they are closer to the normal values for the total population size measure than for the other size measures.

A graphical depiction of the frequency distribution of the estimates is given by the histograms in Figures 1 and 2. These data have been standardized by subtracting the true value of the population statistic from each estimate and then dividing by the empirical standard error of the 1,000 independent estimates. The first histogram, for Kendall's rank correlation, is fairly symmetric. Contrast this with the next histogram, for the with replacement variance component, which show a highly skewed distribution with a long right hand tail. The two cases shown here are illustrative of the histograms for the other cases.

A more specific graphical comparison with the normal distribution is given in Figures 3 and 4. Here normal quantile plots compare the standardized empirical values with the quantiles of the standard normal distribution. Each plot contains a 45^0 line for reference. If the observed estimates were truly normal, they would produce a straight line with slope σ and intercept μ. Since the empirical values have been standardized, they will correspond to the reference line if they approximate a normal distribution.

Figure 1 contains the quantile plot for the rank correlation. Notice that the empirical distribution conforms quite well to the standard normal. When the total population size measure was used (not shown here), the graphs were less similar. This observation is further reenforced by the Kolmogorov-Simirnov (K-S) tests given in Table 3. The K-S test statistics were compared with the two-sided asymptotic critical values of levels 0.200, 0.100, 0.050, 0.025, and 0.010 to determine the range of the significance level of each test. This shows that the empirical distribution of the rank correlation is significantly different from that of a normal distribution when using total population as a measure of size (P-value $<$ 0.01). However, when using the other size measures, the empirical distribution differs significantly, at the 0.05 level, from the normal distribution only for samples of size 50 with the root population size measure.

Turning to the with replacement variance component, Figure 2 shows very clearly that the empirical distribution differs markedly from a normal distribution for this statistic. This is further demonstrated by the K-S tests shown in Table 4. In every case, the test rejects (P-value $<$ 0.01) that the estimates are from a normal distribution.

The final numerical findings of this study are given in Table 5 where the empirical tail and confidence interval coverage probabilities determined for each set of 1,000 estimates are shown. These were obtained by comparing the standardized empirical values with the 0.050, the 0.025, and the 0.010 upper and lower quantiles of the standard normal distribution to find the proportion of times that the empirical values were in the tails of the distribution.

For Kendall's rank correlation, the tail probabilities are very near their nominal levels for both sample sizes and are fairly symmetric when using the log population size measure. With the root population size measure, the probabilities are also well behaved. They are skewed toward the upper tail, but the overall confidence interval is near the proper nominal levels. Sampling using the total population size measure yields the most highly skewed intervals. However, even these intervals are reasonable for the total coverage probabilities of the intervals.

Moving on to the with replacement variance component, the situation is different. As noted when considering the frequency histograms in Figure 2, the distribution of the empirical values is very skewed which results in none of the standardized estimates falling into the lower tail. However, the total probability of the confidence interval when using the total population size measure is near the correct nominal level. When sampling with the other two size measures, the situation is much worse. Again, none of the estimates fall in the lower tail and the empirical level of the confidence interval differs substantially from the nominal level.

2.4 Conclusions

The results of this simulation study are mixed. The validity of the large sample theory developed by Williams (1988) is evident from the results for Kendall's rank correlation. It is clear that the large sample distribution of estimates for this statistic under Sampford's method is approaching that of a normal distribution. The rate of convergence appears to be related to the variability of the selection probabilities. Less variable selection probabilities seem to enhance the rate of convergence. The potential advantages of probability proportional to size sampling are mitigated by the restricted range of the kernel. However, even with highly variable selection probabilities, the normal distribution provides an adequate approximation for sample sizes of 50 to 100.

On the other hand, a warning is sounded by the results for the with replacement variance component. For this statistic, their is some evidence that the large sample distribution of the estimates is moving toward a normal distribution. However, for sample sizes of 50 to 100 under Sampford's method, the distribution of the estimates is still highly skewed with a long right hand tail. The slower rate of convergence for this statistic is not completely surprising since variances tend to generate a chi-square distribution with a long right hand tail. Such a distribution is slow to approach normality in large samples. It was hoped that unequal probability selection proportional to total population size would hasten the approach to normality when compared to the more

restricted range size measures. While there is some slight evidence that this is the case, the effect was far from substantial. The empirical distribution of the estimates deviated greatly from that of a normal distribution.

REFERENCES

Cochran, W. G. (1977), Sampling Techniques, Third Edition, John Wiley and Sons, New York, New York.

Conover, W. J. (1980), Practical Nonparametric Statistics, Second Edition, John Wiley and Sons, New York, New York.

Folsom, R. E. (1984), "Probability Sample U-Statistics: Theory and Applications for Complex Sample Designs," Ph.D. dissertation, University of North Carolina, Chapel Hill, North Carolina.

Hoeffding, W. (1948), "A Class of Statistics with Asymptotically Normal Distribution," Annals of Mathematical Statistics Vol. 19, 293-325.

Kendall, M. G. (1938), "A New Measure of Rank Correlation," Biometrika, Vol. 30, 81-93.

Kendall, M. G. (1970), Rank Correlation Methods, Fourth Edition, Griffin, London.

Nandi, H. K. and Sen, P. K. (1963), "On the Properties of U-Statistics When the Observations are not Independent. Part Two: Unbiased Estimation of the Parameters of a Finite Population," Calcutta Statistical Association Bulletin, Vol. 12, 125-148.

Sampford, M. R. (1967), "On Sampling Without Replacement with Unequal Probabilities of Selection," Biometrika, Vol. 54, 499-513.

Sen, P. K. (1960), "On Some Convergence Properties of U-Statistics," Calcutta Statistical Association Bulletin, Vol. 10, 1-18.

Sen, P. K. (1972), "Finite Population Sampling and Weak Convergence to a Brownian Bridge," Sankhya A, Vol. 34, 85-90.

U. S. Public Health Service (1987), The Area Resource File (ARF) System, U. S. Department of Health and Human Services, Public Health Service, Health Resources and Services Administration, Bureau of Health Professions, Office of Data Analysis and Management, Washington, D. C., ODAM Report No. 4-87, NTIS Accession No. HRP-0907022.

Williams, R. L. (1988), "Large Sample Theory for U-Statistics in Unequal Probability Samples," Ph.D. dissertation, Unversity of North Carolina, Chapel Hill, North Carolina.

Table 1 Mean and Standard Error from 1,000 Replicated Samples of Sizes 50 and 100 by Type of Size Measure

Size Measure / Sample Size	Kendall's Rank Correlation			Relative With Replacement Variance Component		
	True Value	Mean	Standard Error	True Value	Mean	Standard Error
Total Population	0.234			0.692		
50		0.239	0.087		0.699	0.447
100		0.242	0.069		0.690	0.303
Root Population	0.234			0.680		
50		0.242	0.058		0.670	0.421
100		0.240	0.038		0.688	0.321
Log Population	0.234			0.943		
50		0.240	0.034		0.933	0.668
100		0.240	0.023		0.958	0.499

Table 2 Skewness and Kurtosis from 1,000 Replicated Samples of Sizes 50 and 100 by Type of Size Measure

Size Measure / Sample Size	Kendall's Rank Correlation		Relative With Replacement Variance Component	
	Skewness	Kurtosis	Skewness	Kurtosis
Total Population				
50	1.30	6.25	2.00	6.87
100	1.25	6.38	1.32	4.41
Root Population				
50	0.70	4.69	3.08	13.40
100	0.16	3.17	2.02	6.84
Log Population				
50	-0.41	2.95	4.05	22.18
100	-0.47	3.55	2.66	10.31

Table 3 Kolmogorow-Smirnov (K-S) Test of Normality from 1,000 Replicated Samples of Sizes 50 and 100 by Type of Size Measure

Size Measure / Sample Size	Kendall's Rank Correlation		Relative With Replacement Variance Component	
	K-S Statistic	P-Value Range	K-S Statistic	P-Value Range
Total Population				
50	0.084	<.010	0.204	<.01
100	0.067	<.010	0.167	<.01
Root Population				
50	0.047	.025 - .050	0.229	<.01
100	0.024	>.200	0.198	<.01
Log Population				
50	0.038	.100 - .200	0.217	<.01
100	0.037	.100 - .200	0.215	<.01

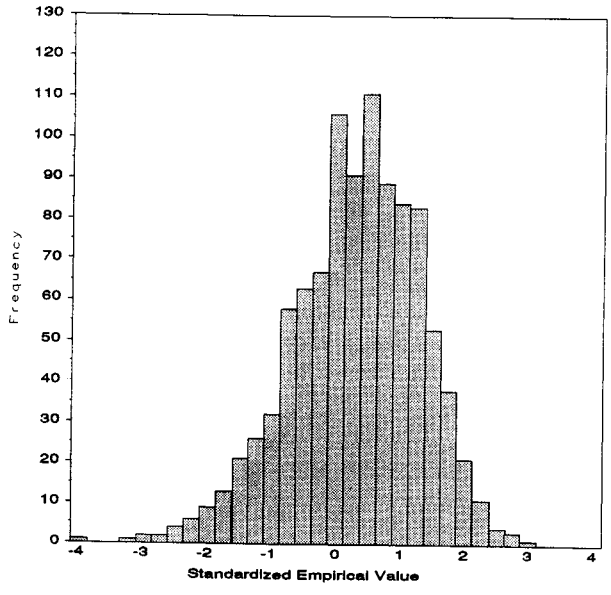

FIGURE 1 FREQUENCY DISTRIBUTION FOR KENDALL'S RANK CORRELATION: SAMPLES OF SIZE 100, PPS TO LOG POPULATION SIZE

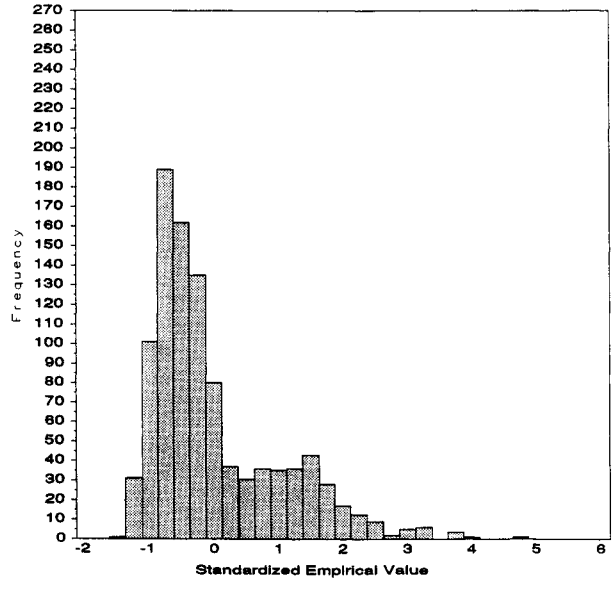

FIGURE 2 FREQUENCY DISTRIBUTION FOR WITH REPLACEMENT VARIANCE COMPONENT: SAMPLES OF SIZE 100, PPS TO POPULATION SIZE

Table 4 Empirical Tail Probabilities from 1,000 Replicated Samples of Sizes 50 and 100 by Type of Size Measure

Size Measure Nominal Tail Level	n=50 Lower Tail	n=50 Upper Tail	n=50 Both Tails	n=100 Lower Tail	n=100 Upper Tail	n=100 Both Tails
Kendall's Rank Correlation						
Total Population						
0.050	0.013	0.063	0.076	0.005	0.070	0.075
0.025	0.004	0.047	0.051	0.000	0.052	0.052
0.010	0.000	0.031	0.031	0.000	0.032	0.032
Root Population						
0.050	0.017	0.057	0.074	0.029	0.072	0.101
0.025	0.006	0.038	0.044	0.014	0.040	0.054
0.010	0.002	0.029	0.031	0.003	0.015	0.018
Log Population						
0.050	0.045	0.056	0.101	0.036	0.075	0.111
0.025	0.024	0.019	0.043	0.020	0.028	0.048
0.010	0.011	0.006	0.017	0.011	0.011	0.022
With Replacement Variance Component						
Total Population						
0.050	0.000	0.106	0.106	0.000	0.080	0.080
0.025	0.000	0.092	0.092	0.000	0.047	0.047
0.010	0.000	0.052	0.052	0.000	0.029	0.029
Root Population						
0.050	0.000	0.061	0.061	0.000	0.086	0.086
0.025	0.000	0.047	0.047	0.000	0.082	0.082
0.010	0.000	0.040	0.040	0.000	0.074	0.074
Log Population						
0.050	0.000	0.038	0.038	0.000	0.059	0.059
0.025	0.000	0.027	0.027	0.000	0.056	0.056
0.010	0.000	0.025	0.025	0.000	0.056	0.056

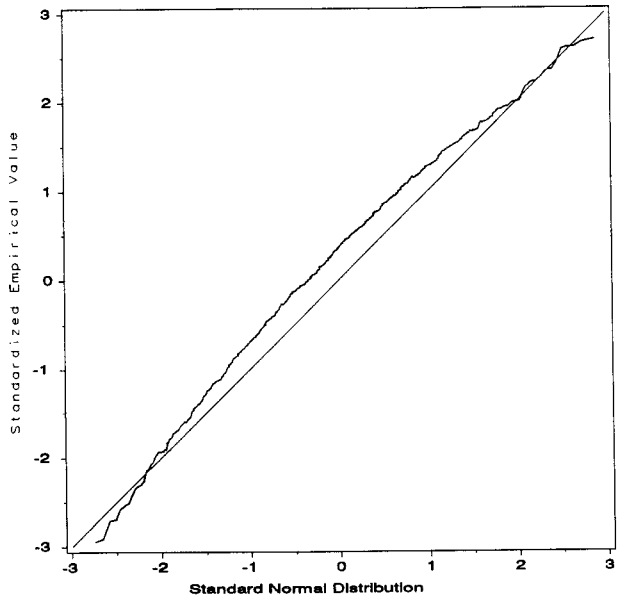

FIGURE 3 NORMAL QUANTILE PLOT FOR KENDALL'S RANK CORRELATION: SAMPLES OF SIZE 100, PPS TO LOG POPULATION SIZE

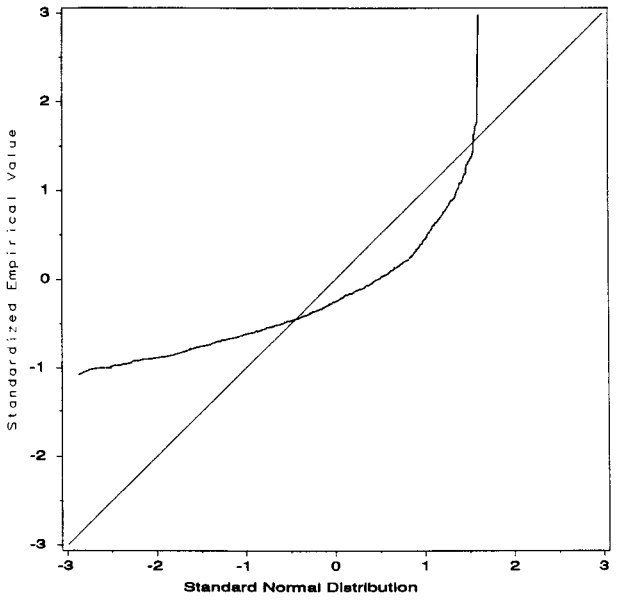

FIGURE 4 NORMAL QUANTILE PLOT FOR WITH REPLACMENT VARIANCE COMPONENT: SAMPLES OF SIZE 100, PPS TO POPULATION SIZE