ESTIMATING THE VARIANCE OF PERCENTILES USING REPLICATE WEIGHTS

John W. Rogers
Westat, 1650 Research Blvd., Rockville, MD 20850

KEY WORDS: Percentiles, Replicate weights, Variance, Sample surveys

1. Introduction

Different programs use different formulas for estimating percentiles for data from a random sample (for examples, see Hyndman and Fan, 1996). Estimating percentiles for survey data based on a complex probability sample design requires selection of a formula for calculating the percentiles. The choice can affect the bias of the percentile estimate. Variances for survey statistics can be calculated using several different approaches. The most common approaches use replicate weights and Taylor series linearization.

When using replicate weights, there are several alternative approaches for estimating the variance of percentiles but no generally accepted method (Wolter, 1985). The replicate weight approach calculates the variance of a percentile from percentile estimates using the full sample weight and each replicate weight. An approach proposed by Woodruff, calculates the confidence interval (CI) and variance from the sample cumulative distribution function (CDF) and the variance of *p*, the proportion of the population less than the full sample percentile estimate.

The sample weights can often be constructed in different ways, however, they must be consistent with the sample design and analysis needs. Jackknife replicate weights are often used and are the only replicate weights considered in this paper.

This research was initiated to test the hypothesis that rounding or binning the data can, at least in some circumstances, improve the performance of confidence intervals when using jackknife replicate weights. This paper reports on simulations to evaluate alternate approaches for calculating percentiles and the variance and confidence intervals (CIs) for the percentile estimates using jackknife replicate weights. One unexpected result of the research was quantification of estimation bias associated with rounding or binning the data.

The formulas used in this paper for calculating a percentile of a variable assume the variable:

1) Is continuous and real valued such that interpolated values could be possible population values, and
2) Has a distribution that is reasonably smooth, such that interpolation between observed values provides a reasonable estimate of intermediate percentiles.

The simulations assume a stratified sample of clusters. Within clusters the population values are assumed to be random around means that vary among strata and clusters.

Section 2 discusses the hypothesis to be tested. Before testing the hypothesis it was necessary to select a formula for calculating percentiles. Section 3 discusses simulations to evaluate different formulas. Section 4 discusses the simulations to test the hypothesis. Finally, Section 5 provides a discussion of the results.

2. Hypothesis

For a simple random sample with n observations, jackknife variance estimates of a statistic can be calculated by:

1) Estimating the statistic of interest, $\hat{\theta}$,
2) Sequentially deleting each observation *i* and calculating the statistic of interest $\hat{\theta}_i$, and
3) Estimating the variance of $\hat{\theta}$ by:

$$\sum_{i=1}^{n} \frac{n-1}{n} \left(\hat{\theta} - \hat{\theta}_i\right)^2 \qquad (1)$$

This estimate of variance is generally considered to have n-1 degrees of freedom.

However, unweighted Jackknife estimates for percentiles are known to be inconsistent (Efron, 1982). Using generally accepted methods for calculating medians, when the sample size is even, the median estimate is the average of the middle two observations. Each jackknife estimate differs from the full sample estimate by half the difference between the observations closest to the full sample estimate. As a result, the jackknife estimate of the variance is a function of the local behavior of the sample CDF at the median. Similar conclusions apply for other percentiles and when the sample size is odd.

The simulations reported here were initiated to test the general hypotheses that smoothing the sample CDF will make the variance less dependent on the local behavior of the CDF and improve CI coverage. One approach to smoothing the sample CDF is to bin the data and use the count and mean (or mid-point) within each bin to define the sample CDF. For this research, bin boundaries were equally spaced along the measurement scale.

Rounding of the data is a special type of binning that corresponds to using equal length bins and using the midpoint of the bin as the value for each observation in the bin. Rounding or binning of survey data may occur because the data collection process only records a fixed number of significant digits or because the data are coded into ranges. If the data are binned as part of the analysis procedure, the bins and the binned data value can be specified by the analyst.

3. Estimating Percentiles

Hyndman and Fan (1996) provide a summary of commonly used formulas used for calculating percentiles. The various formulas fall into two general categories, discontinuous and piecewise continuous functions to approximate the population quantile function. The usual sample quantile function is a discontinuous step function. Use of a discontinuous quantile function has disadvantages when using jackknife replicate weights. In particular, the estimated variance of a percentile may be zero, particularly if there are tied values in the data. The performance of the jackknife percentile variance is improved by interpolating between the steps of the sample quantile function. For this paper, only piecewise continuous functions with linear interpolation are considered.

The following formula (presented by Hyndman and Fan) defines plotting points. Linear interpolation between the plotting points is used to define the piecewise continuous sample quantile function. The location of the plotting points depends on two parameters, α and β. Assume a continuous underlying variable and a piecewise-linear sample quantile function p = f(x) defined by plotting points (x_k, p_k), where x_k is the k^{th} ordered observation,

$$p_k = \frac{k-\alpha}{n+1-\alpha-\beta}, \text{ and} \qquad (2)$$

α and β are constants between 0 and 1. The percentile corresponding to the desired percentage P is obtained by interpolation between neighboring plotting points.

Different software programs use different α and β values. For symmetric distributions, Hyndman and Fan argue that α and β should be equal. For random samples they recommended setting α and β between 1/3 and 3/8. Unless α and β equal 1.0, there will be percentages less than p_1 and greater than p_n for which percentiles cannot be calculated. Some software uses the maximum or minimum for these values. Here these percentiles were set to missing.

Hyndman and Fan implicitly assume that there are no ties and do not discuss weighted data. Equation (3) is proposed here as a reasonable analog of equation (2) for weighted data and data with tied values.

$$x_k = k^{th} \text{ ordered unique value, } k = 1 \ldots n$$

$$S_k = \sum_{j=1}^{k} W_j, \; W_k = \text{weight for observations equal to } x_k \qquad (3)$$

$$p_k = \frac{S_k - W_k\alpha}{S_n + W_k(1-\alpha-\beta)}$$

The simulations for this paper use formula (3) to estimate percentiles. As with formula (2), there may be upper and lower percentages for which the percentiles cannot be calculated. Binning the data tends to increase the sum of weights for the lowest and highest bins and therefore increase the range of percentages for which percentiles cannot be calculated.

This paper makes a distinction between using interpolation between plotting points to provide to obtain a piecewise continuous sample quantile function and rounding or binning to smooth the quantile function by reducing the number of unique values from which the quantile function is calculated.

3.1 Simulations for Selecting α and β

Simulations using different values of α and β were used to determine how these values affect the bias of the percentile estimates. These simulations used the following parameters:

- Sample design: random sample from an infinite population (i.e., no differences between cluster and strata means).
- Data distribution: standard normal (N(0,1)) or a uniform (U(0,1)).
- Sample sizes:16, 64, 256, 1024
- Bins based on rounding, rounding unit: .4, 0.2, 0.1, 0.05, 0.025, 0.01, and 0.0001 for N(0,1) and 0.2, 0.1, 0.05, 0.025, 0.01, and 0.0001 for U(0,1). Note, for the uniform distribution the bins were defined as even divisions between 0 and 1,
- Alpha: 0.1, 0.3, 0.5, 0.7 for N(0,1) and 0.0, 0.2, 0.4, 0.6 for U(0,1).
- Beta = Alpha.
- Simulations: 14,400 for the three smallest rounding units and 1600 for other simulations

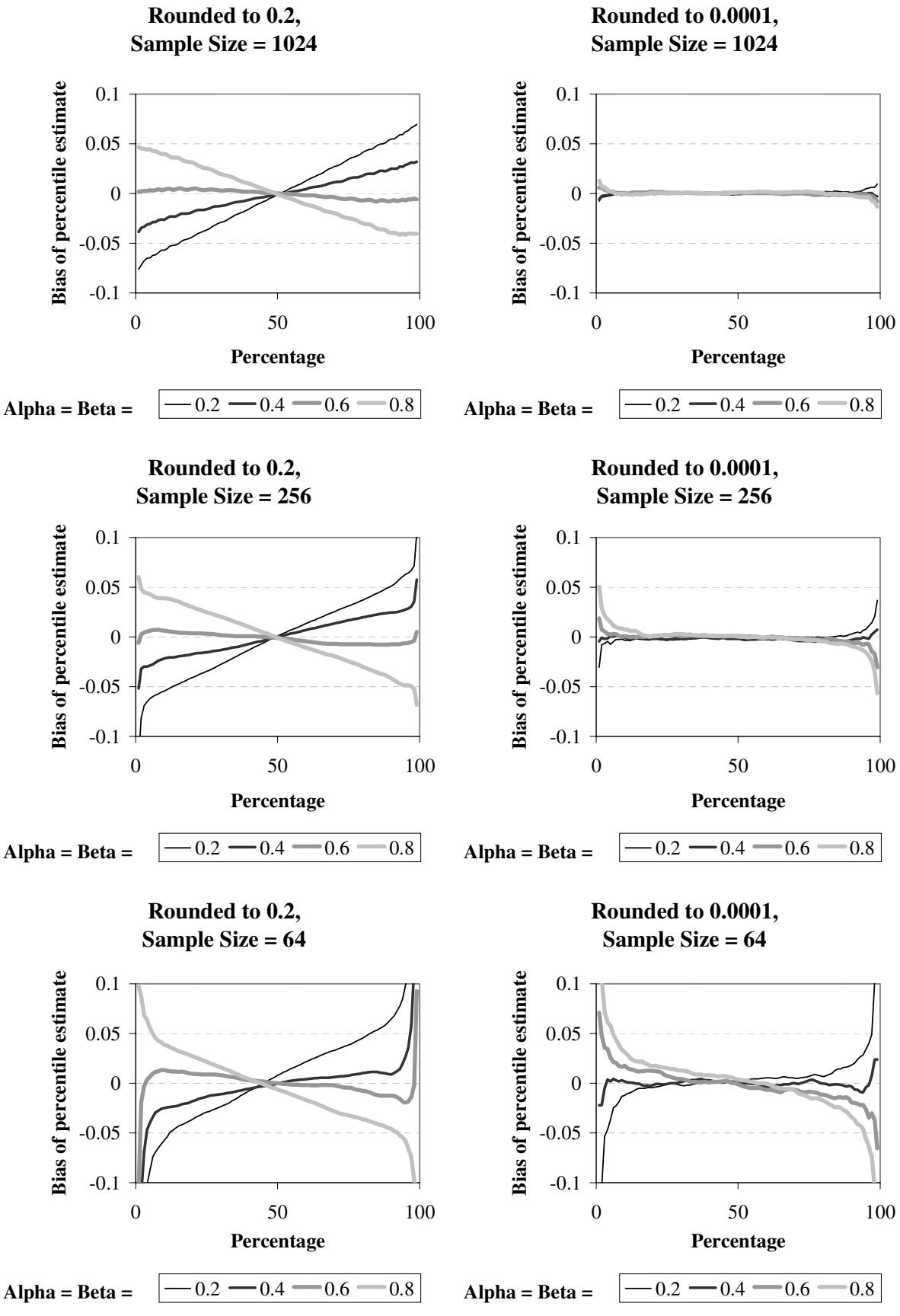
For each simulation, the bias was calculated for all percentages from 1 to 99.

Figure 1 shows example results for a random sample 64, 256, or 1024 normally distributed observations rounded to units of 0.2 or 0.0001. Except for extreme percentages, the bias appears to be a linear function of the percentage. Similar patterns were found using both normal and uniform distributions.

3.2 Choosing α, β to minimize Bias

For most plots in Figure 1, for alpha equal to 0.2 or 0.4 estimates of upper percentiles are greater than the true value and estimates of lower percentiles are less than the true value, on average. The reverse is true for alpha of 0.6 and 0.8. Interpolating between the curves suggests that an alpha around 0.5 will minimize the bias for most percentages (i.e., the slope versus percentage will be zero). The alpha for which the slope of bias versus percentage (for percentages between 20 and 80) is zero was calculated (using interpolation) for all simulated conditions and will be referred to as the optimal alpha.

Figure 1 Bias versus α and β, Random Sample from a Normal Distribution



The optimal alpha depends on the distribution. Figure 2 shows the optimal alpha for a uniform distribution versus rounding unit and sample size. The optimal alpha for rounded data increases from near zero to an asymptote at about 0.5 as the rounding unit increases and increases as the sample size increases.

Figure 2 Optimal alpha for a uniform distribution

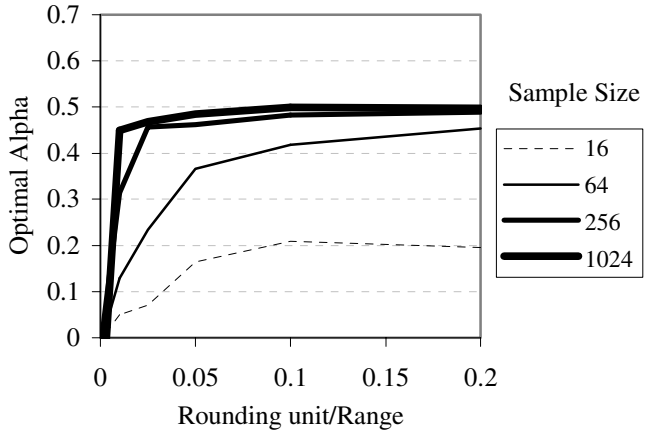
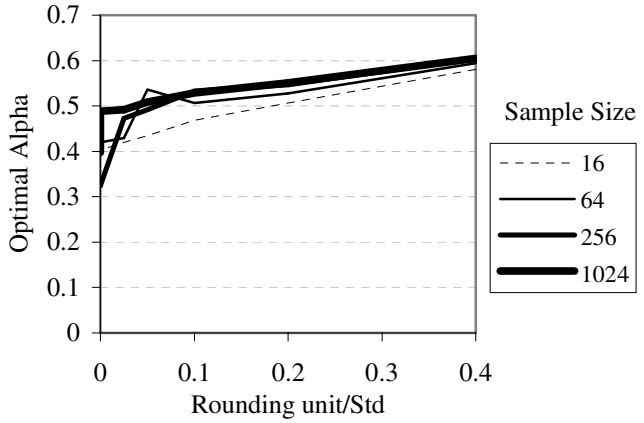


Figure 3 shows the optimal alpha for a normal distribution versus rounding unit and sample size. The optimal alpha for rounded data increases from near 0.4 to about 0.6 as the rounding unit increases from essentially no rounding to a rounding unit of 0.4. The optimal alpha increases slightly with increasing sample size.

Figure 3 Optimal alpha for a normal distribution



The bias is more sensitive to the choice of alpha for larger rounding units. Based on an empirical fit to the data, the bias for percentages between 20% and 80% is approximately:

$$Bias = -.020(Pct - 50)(\alpha - Optimal\,\alpha)R \qquad (4)$$

where Pct = the percentage for the desired percentile, α is the alpha value use to estimate the percentile, $Optimal\,\alpha$ is the optimal alpha for minimizing bias and R is the rounding unit. The constant (.020) is the essentially the same for both the standard normal and standard uniform distribution. The constant is essentially the same for larger sample sizes (256 or greater in these simulations) but differs somewhat for smaller sample sizes. For uniform and normal distributions with other standard deviations, the constant in equation (4) would need to be scaled appropriately.

If one value of alpha is to be used to all calculations, I recommend choosing alpha for data with significant rounding, for which the potential bias can be largest. For data from a uniform or normal distribution, setting alpha between 0.5 and 0.6 will perform relatively well for all but small sample sizes. For small sample sizes, the bias will be small relative to the standard error.

The calculation of percentiles uses linear interpolation between plotting points. The CDF for the normal distribution is non-linear. Therefore linear interpolation provides only an approximation to the CDF. For the larger rounding units, the error due to using linear interpolation to approximate a non-linear function can be seen in the simulation results. Since sample CDFs can be transformed to be linear, at least approximately, I recommend using alpha equal to 0.5 unless there is significant rounding and adequate information for specifying another value for alpha. Using 0.5, users would get better estimates if data were transformed to have an approximately uniform distribution. For the subsequent simulations, alpha and beta are set to 0.5.

4. Estimating Variance and Confidence Intervals

This research used two methods for estimating the variance and confidence intervals for percentiles, Woodruff (see Sarndal, Swensson, and Wretman, 1992), and replicate weights. The replicate weight approach calculates the percentile estimate for weighted subsets of the data represented by the replicate weights. The variance and confidence intervals are calculated from the variance of the replicate estimates. The Woodruff method calculates a confidence interval for the percentage of the sample below the estimated percentile. That confidence interval is then transformed to the to the measurement scale using the inverse of the sample quantile function. These two methods are described below in mathematical terms. The following sections describe the simulations to evaluate the hypothesis that rounding or binning the data can improve the performance of the confidence intervals for percentiles.

The steps for calculating the Woodruff CI are:

1. For desired percentage P, Calculate $x_P = f^{-1}(P)$
2. Calculate p, the population percentage less than x_P, and its confidence interval, $[p_L, p_U]$ (using replicate weights and normal assumptions)
3. The CI around x_P is:
$$\left[f^{-1}(P-(p-p_L)),\, f^{-1}(P+(p_U-p))\right]$$
4. Woodruff CI cannot be calculated if CI limits are beyond the range of the quantile function

The steps for calculating the replicate weight CI are:

1. For desired percentage P, Calculate $x_P = f^{-1}(P)$
2. For each replicate weight, r = 1…k, calculate the estimated quantile function f_r and
3. $x_{Pr} = f_r^{-1}(P)$
4. Calculate the confidence interval on x_P using the replicate estimates and normal assumptions, see equation (1)
5. The replicate weight CI cannot be calculated if any of the full sample or replicate weight estimates are beyond the range of the quantile function

4.1 Simulations Using Clusters of Random Data

Simulations were performed to assess the effect of rounding and clustering on the coverage of percentile CIs. For these simulations there were no differences between cluster means. The simulations used the following parameters:

- Sample design: clustered samples of random data
- Sample size of 64 or 1024
- Cluster sizes of 4, 8, and 16 with a sample size of 64 and 4, 16, 64, and 256 with a sample size of 1024
- Normal and uniform data distributions
- Rounding units of 0.2, 0.1, 0.05, 0.01, and 0.000001

For each simulation the CI coverage was calculated for all percentages from 1 to 99.

Overall, rounding increases CI coverage. The simulation results suggest that, for a fixed sample size, larger clusters are associated with closer-to-nominal CI coverage, particularly for smaller sample sizes. However, additional simulations using other sample and cluster sizes would be needed to be confident of those conclusions. Rounding had more effect on coverage than cluster size. Figure 4 shows example results for rounding units of 0.1 and 0.000001 and two sample/cluster size combinations.

The percentage of confidence intervals that could be calculated drops off as the percentage for the percentile approaches 0% or 100%. For these extreme percentages the plots show the coverage for those confidence intervals that could be calculated. If no confidence intervals could be calculated, the plots show the coverage for the nearest percentage for which the confidence interval coverage could be calculated.

With essentially no rounding, the coverage of the Woodruff confidence intervals is close to the nominal level of 95%. With rounding the Woodruff coverage is greater than 95% and the confidence intervals are conservative.

The jackknife replicate weight CI coverage is generally less than the nominal level without rounding and close to the nominal level with rounding and/or larger clusters

Additional CI coverage results include:

- The replicate weight CI can be calculated for a wider range of percentages than the Woodruff CI. The range decreases with fewer/wider bins.
- For variance estimates, the effective Df using jackknife replicate weights is less than using the Woodruff method, which is less than the number of clusters.
- Estimation of extreme percentiles is problematic because:
  - Calculation of estimates may not be possible
  - Bias is often larger for extreme percentiles
    - The optimal α may be different for extreme percentiles than for intermediate percentiles
  - CI coverage can diverge from the nominal coverage

Figure 5 shows the average CI coverage versus sample size and rounding unit for a uniform distribution using Woodruff and replicate weights methods. CI coverage is the average of all CIs that could be calculated.

4.2 Simulations with Complex Samples

Limited simulations were performed using complex sample designs with stratified samples of clustered data. These simulations used the following parameters:

- Distribution: Normal distribution ($N(\mu,1)$)
- 2 strata, 8 clusters of 16 observations per strata (sample size of 256)
- Mean (μ) differed among clusters within strata and between strata :
  - Std. of cluster means: 0.0 or 1.0
  - Difference between strata means, 0 or 1.0
- Number of bins, 20, 100, infinite (i.e., no binning). The mean within each bin was used for the data value.

As before, for each simulation the CI coverage was calculated for all percentages from 1 to 99.

Figure 4 CI Coverage, Random Sample from a Uniform Distribution, α = 0.5

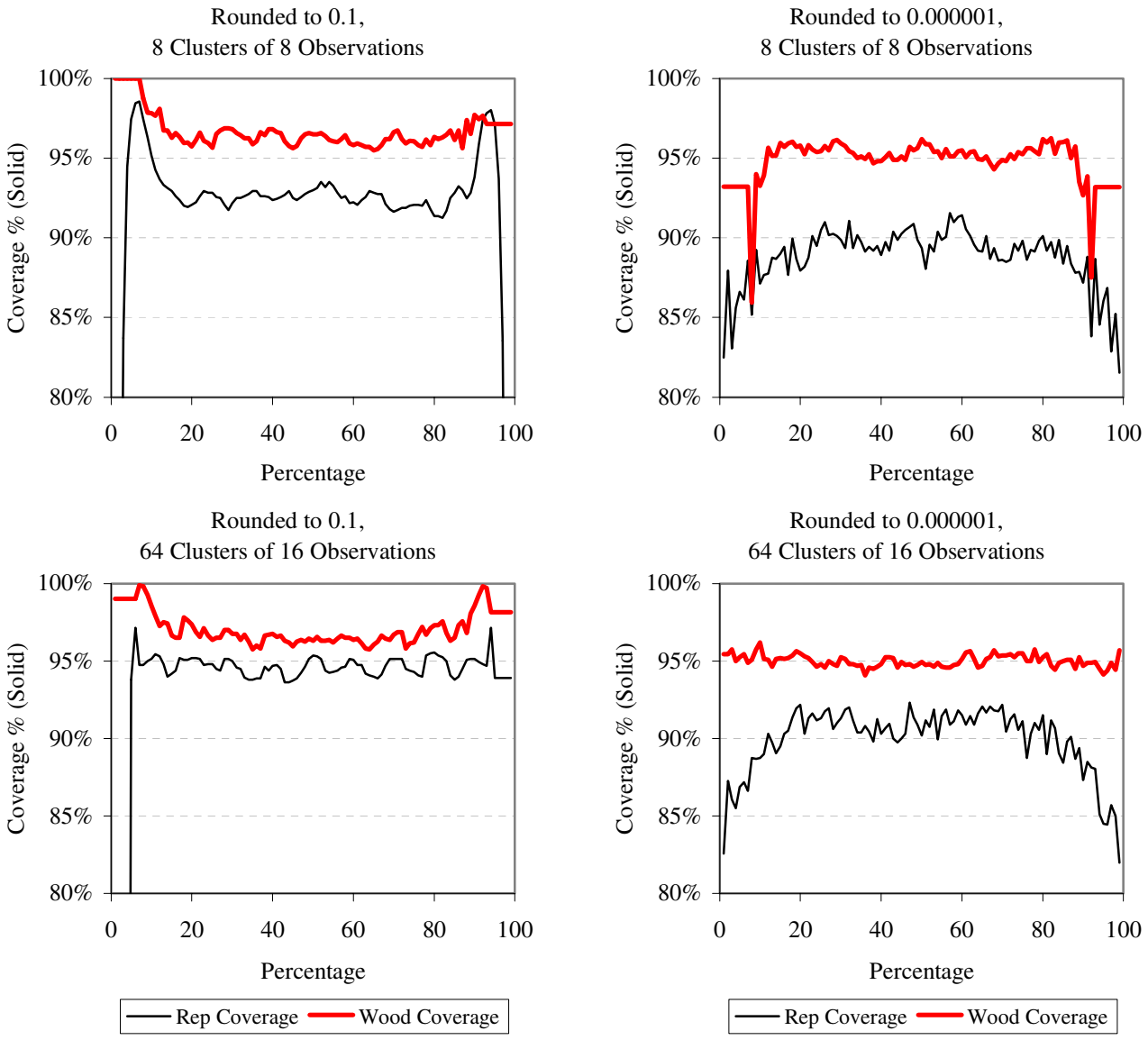
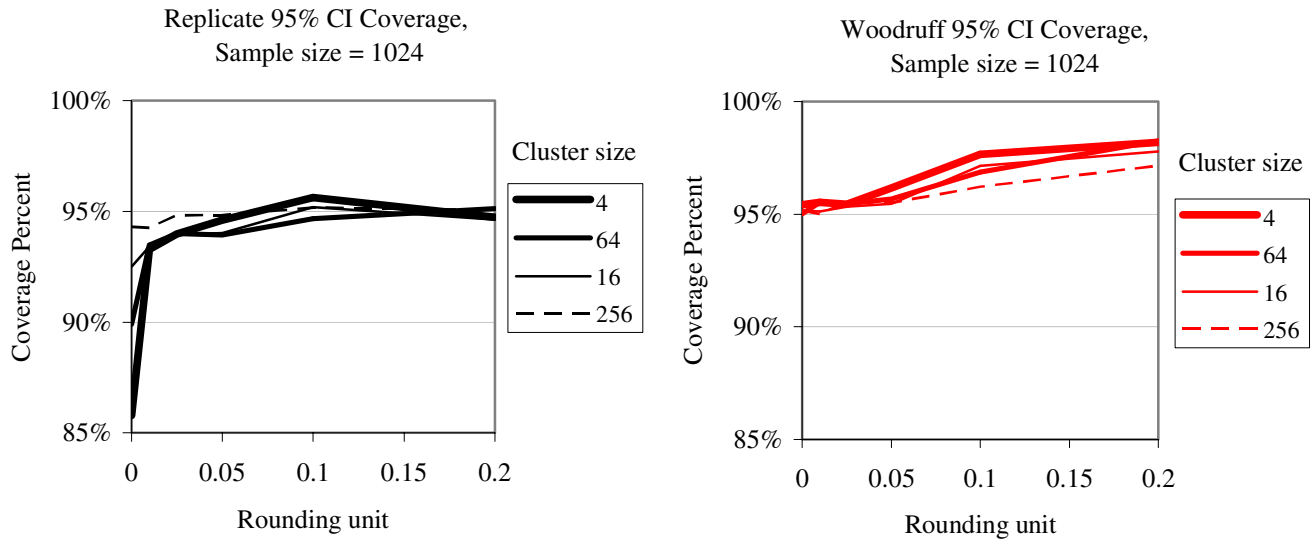


Figure 5 Average CI Coverage versus rounding unit, Replicate and Woodruff methods, Uniform Distribution, α = 0.5



Based on the limited set of complex sample simulations, the results are similar to those reported earlier (binning increases CI coverage) except that binning the data is less effective at increasing coverage when there are differences between clusters.

Figure 6 shows the simulation results for no binning or 20 bins for simulations with differences between strata, between clusters, and differences between both strata and clusters.

5. Discussion

Perhaps the most important result from this research is the extent to which the bias of the estimated percentile depends on the choice of the formula for estimating the percentile, represented here by the choice of alpha and beta. Based on the simulations I recommend setting $\alpha = \beta = 0.5$, unless adequate data are available for selecting other values.

This research looks at the effect of binning or rounding of data on the bias and coverage of 95% confidence intervals. Overall, binning increases coverage of the CIs. Without binning Woodruff CIs have coverage close to the nominal level. With binning, coverage of Woodruff confidence intervals is generally higher (more conservative) than the nominal level. When using Woodruff confidence intervals, binning is not recommended. If the data have already been binned or rounded as part of the data collection/processing process, Woodruff CIs appear to be conservative. Binning decreases the range of percentages for which CIs can be calculated. For the Woodruff method, this limitation may be unacceptable.

Binning can improve the coverage of replicate weight confidence intervals. Most of the improvement appears to be achieved with only slight rounding or use of many bins. However, further improvement is associated with more rounding or use of fewer bins. Rounding the data reduces the range of percentages for which CIs can be calculated while improving the coverage. Even with rounding, percentile CIs can be calculated for most percentages when using replicate jackknife weights.

Without rounding, the Woodruff method is preferred for calculating confidence intervals. For more extreme percentiles, if the Woodruff CI cannot be calculated, the replicate weight interval is a reasonable choice. Some rounding or binning may improve the estimate. However, both of the tested approaches perform poorly for estimating the most extreme percentiles.

Limited simulations suggest that binning has relatively little effect when there are differences between clusters. Many surveys collect binned data by dividing the response range into bins and asking the respondent to pick the relevant bin. Also, many survey designs randomly select clusters for which differences among clusters are expected. In this situation, the replicate jackknife CIs are expected to have CI coverage that is lower than but close to the nominal level. Woodruff CIs are expected to be conservative.

Prior research has shown that jackknife replicate CIs have coverage closer to the nominal level with designs with larger clusters. (Kovar, Rao, and Wu, 1988) Although this research is consistent with those findings, the effect of cluster size was found to be small in the limited number of simulations with complex sample designs.

The following are limitations of this research:

- Binning the data, in effect, smoothes the quantile function. Other smoothers may have more desirable results.
- Coverage of the jackknife replicate weight CI might be improved by using fewer degrees of freedom when calculating the CI
- Although the results for complex samples are consistent with other simulations, additional work is needed.
- Simulations of functions of percentiles (such as inter-quartile range) have not been performed.

References

Hyndman R.J., and Fan Y. Sample Quantiles in Statistical Packages, *The American Statistician,* November 1996, Vol 50. No 4, pp 361-365.

Efron B., (1982) *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia PA: SIAM

Kovar. J.G., Rao, J.N.K., and Wu. C.F.J. Bootstrap and Other Methods to Measure Errors in Survey Estimates, *Canadian Journal of Statistics*, Vol. 16 Supplement, pp 25-46.

Sarndal, C., Swensson, B., and Wretman, J. (1992) Model *Assisted Survey Sampling*. New York: Springer-verlag.

Wolter, K.M. (1985) *Introduction to variance estimation*. New York: Springer-Verlag

Figure 6 Complex Samples, Normal Distribution, Sample Size = 256, α = 0.5