Empirical Study on the Second-stage Sample Size

Yan Liu, Mary Batcher, Ryan Petska and Amy Luo
Yan Liu, Ernst & Young LLP, 1225 Connecticut Ave., NW, Washington, DC 20036

Abstract

In a typical research setting, two-stage stratified sampling is typically done in situations where both the populations and the samples are large. But in the case of an audit setting, where business records are sampled and reviewed, sampling is typically done on relatively small populations and samples. For this setting, there are two common methods[1] used for variance estimation; the classical design-based approach or a resampling approach. The classical design-based approach directly incorporates the second-stage sample size into the variance formula, while the typical resampling approach does not explicitly express the second-stage sample size but it is implied in the variance formula.

It is known that as the second-stage sample size increases, the overall variance decreases; but how large of a second-stage sample size is 'large enough?' In this paper, we will investigate the impact the second-stage sample size has on the overall estimation in different estimation approaches in the two-stage stratified, audit sampling setting.

Key words: Jackknife; Ratio Estimation; Stratified Sampling; Two-Stage Sampling.

1. Background

Much of the research for multi-stage stratified sampling is limited to both large populations and large samples. But in the audit practice, business sampling is typically done on relatively small samples due to time and cost constraints. For example: in two-stage business sampling the first-stage could be a location. If a large first-stage sample is taken, we may need to travel to many different places in order to pull the necessary records. This can be very costly and timely. At the second-stage, the cost of pulling out and reviewing a single record could also be costly. So, we want to minimize the sample, at both stages, as much as possible.

In a typical business sampling situation, there exists a list which consists of a relatively small number of entities, and a corresponding list, for each entity, that contains a large number of business records. In other words, the first-stage population is small and the second-stage populations are large. The quantity to be estimated may be, for example, the amount subject to sales tax, the amount deductible from income tax, or an amount that is in error. The estimates for these quantities have a lower bound of zero but can take on large positive values, sometimes millions of dollars. In addition, there are always requirements to minimize the impact of the sampling on company operations and to keep the sample size as small as possible, while still achieving good precision. If an entity is selected, that entity will then provide their list of invoices with corresponding dollar amounts. For those entities not in the sample, only the total invoice amounts at the first-stage are available from the financial report.

In this type of problem, a two-stage stratified sample design is often used. The classical design-based approach does give us a closed form of variance estimation, but the formula is very complicated for two-stage stratified sampling and gets even more difficult for additional stages. On the other hand, resampling approaches are fairly straightforward and easy to implement in multistage sample designs. For the classical design-based approach, the second-stage sample size is explicitly expressed in the variance formula, while the contribution to variance from the second-stage sample size is implied in the variance calculation, but not explicitly expressed, for resampling approaches. In the classical design-based approach, the second-stage sample size can be calculated using assumptions about the costs and on the ratio of the variance components of the two stages (Lohr, p.156). Though, this may become extremely difficult for sample designs with more than two stages. The second-stage sample size cannot be calculated from a variance formula for resampling approaches. In general, the statistical properties of the variance estimators for resampling approaches are limited to simulation or empirical studies (Sarndal, p.419).

In this paper, we compare two methods of variance estimation – the closed form of the design-based approach and the Jackknife; one of the most common resampling methods. Specifically, we will look at the impact the second-stage sample size has on the overall variance estimation using simulations.

[1] The model-based approach is a good choice if there is a good model fit, see Valliant, Dorfman and Royall (2000). In this paper, we only intend to compare the two design-based methods.

2. Simulated Data - Typical Auditing Situation

Our hypothetical, typical population will consist of 31 entities (PSUs) and within each entity there will exist hundreds to thousands of invoices (SSU). For each invoice, there is an invoice amount (x) which is known and a qualified amount (y) which could be anywhere between zero to the full invoice amount. Our goal is to estimate the total qualified amount in the population. The total invoice amounts for each of the population entities is known and their distribution is highly skewed. The distribution of invoice amounts within entities is also very skewed. Figure 1 shows the distribution of the total invoice amounts for all 31 entities. Figure 2 shows an example of a distribution of the invoice amounts within a single entity. Figure 3 is the scatterplot of the total qualified amount against the total invoice amount for the 31 entities in the first-stage population. Figure 4 is the scatterplot of the qualified amounts against the invoice amounts for invoices within a single entity. The plot in Figure 4 shows the typical relationship between the qualified amount and the invoice amount at the SSU level. Due to the distributions of the design variable x (the invoice amount) being skewed at both, the first and second-stages, a sample design that is stratified at both stages[2] and sampled without replacement is appropriate. We expect some changes in the qualified percentages across entities, but the changes may not be substantial, as shown in Figure 3. Therefore, the combined ratio estimation method is used.

Given the known values of the invoice amounts (x), the qualified amount (y) is simulated using

$$y = \begin{cases} \beta x + u(1-\beta)x, & \text{with probability } \beta \\ \beta x - u\beta x, & \text{with probability } (1-\beta) \end{cases} \qquad (3.1)$$

where u is a random number from *Uniform* $(0,1)$. For each entity, a value is assigned to β that can be viewed as the approximate qualified percent per entity. For the 20 entities in our population with a relatively small total invoice amount, β is randomly assigned a value of 0.5 or 0.6; for the ten entities with a medium to large total invoice amount, β is randomly assigned 0.6 or 0.7; and for the one largest entity, β is set to be 0.7.

The simulated values of y are scattered around the line βx within the range of $(0, x)$. Figure 4 gives the scatterplot of simulated y against x at the SSU level for a single entity. Figure 3 gives the scatterplot of simulated y against x at the PSU level for all entities.

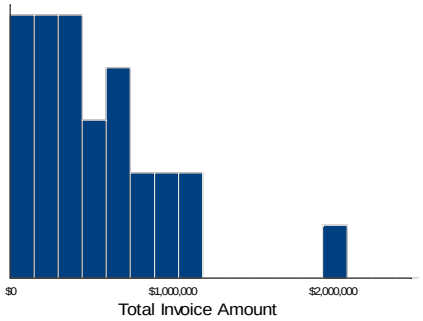

Figure 1. Frequency Distribution of Total Invoice Amount Per Entity (PSU)

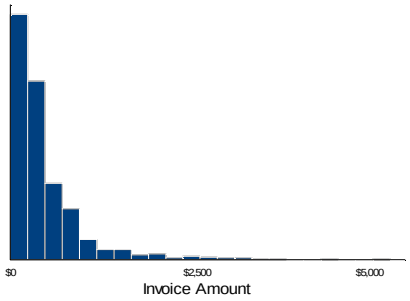

Figure 2. Typical Frequency Distribution of Invoice Amount (SSU)

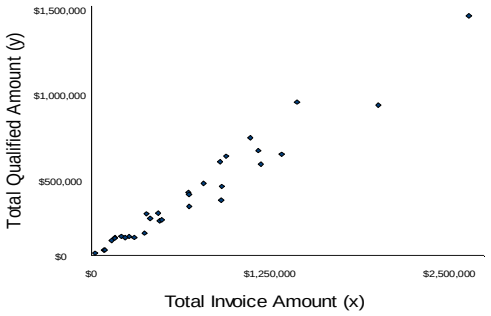

Figure 3. Relationship of Qualified Amount (y) and Invoice Amount (x), PSU Level

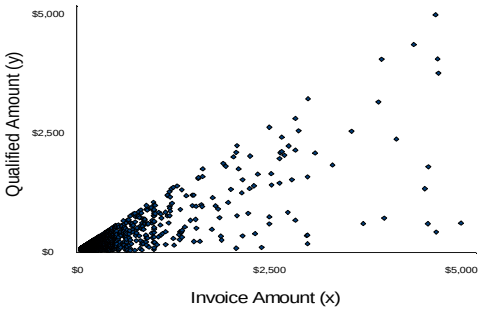

Figure 4. Typical Relationship of Qualified Amount (y) and Invoice Amount (x), SSU Level

[2] The other method to consider using is Probability Proportionate to Size (PPS) sampling at the first stage and stratified sampling at the second-stage. This method was not explored in this paper.

Table 1. The Population Summary by Stratum

PSU Stratum	Stratum Definition: Total Invoice Amount Per Entity (PSU)		Number of PSUs	Total Invoice Amount	Number of SSUs	Number of SSU Strata Per PSU
	Minimum	Maximum				
1	20,557	624,770	20	5,673,319	28,241	2
2	717,236	1,598,935	10	9,479,716	43,584	3
Certainty	2,103,016	2,103,016	1	2,103,016	13,340	3
Total			**31**	**17,256,051**	**85,165**	

3. Sample Design on the Simulated Data.

Some notations are defined in the following.

- The population is stratified at both stages.
- The strata are called PSU strata at the first-stage and SSU strata at the second-stage.
- The population of N PSU units (entities) is divided into L strata; $h = 1, 2, \cdots, L$.
- Within each stratum h, there are N_h PSU units; $i = 1, 2, \cdots, N_h$.
- Within the i^{th} PSU of stratum h, the SSU units (invoices), are divided into L_{hi} strata; $k = 1, 2, \cdots, L_{hi}$.
- Within stratum k of PSU (h, i), there are M_{hik} elementary units.
- From the M_{hik} SSU units of cell (h, i, k), m_{hik} elementary units are randomly selected.
- At the SSU level, x_{hikj} is the known invoice amount and y_{hikj} is the qualified amount; $j = 1, 2, \cdots, M_{hik}$.
- X is the total invoice amount of the population PSU units

The 31 entities (PSU) are stratified into three strata by the total invoice amount per entity, as shown in Table 1.

For each PSU, the invoices are also stratified into two strata if the PSU falls within stratum 1 or three strata if the PSU falls within stratum 2 or in the certainty stratum, based upon the total invoice amount. The SSU stratum boundaries are created independently within each entity using the Delaneous-Hodges method for stratification.

For each sample, nine PSUs are sampled – one is taken with certainty and four are randomly selected from each of the random strata. Then, for each of the sampled PSUs, a number of SSUs are randomly selected from each SSU stratum. To compare different second-stage sample sizes, four scenarios are used, as summarized in Table 2. There are 23 SSU strata from a sample of nine PSUs. In scenario 1, ten SSUs are randomly selected from each SSU stratum and the total number of sampled SSUs is 230. Similarly, there are 690 and 1,380 sampled SSUs if the SSU sample size per SSU stratum is 30, as in scenario 2, or 60, as in scenario 3. In scenario 4, all SSUs are taken for each sampled PSU. At this point, the sample design becomes a one-stage design and serves as a benchmark for our comparisons. The number of sampled SSUs in scenario 4 depends on the selected PSUs, and averaged approximately 36,421 SSUs per sample.

Table 2. Scenarios of Second-Stage Sample Size

Scenario	Sample Size Per SSU Stratum m_{hik}	Total SSU Sample Size $\sum_{h,i,k} m_{hik}$
1	10	230
2	30	690
3	60	1,380
4	Full	Average 36,421

The above simulation process is repeated 1,000 times with a different seed for sample selection every time.

4. Estimation Formula

To achieve our simulation results, we use a two-stage design stratified at both stages. The qualified amount is estimated using the combined ratio estimator and the variance is estimated by using both the closed form and the Jackknife methods.

Point Estimator. The combined ratio estimator for a two-stage stratified sample design is

$$\hat{Y}_{Rc} = X \frac{\hat{Y}_{st}}{\hat{X}_{st}}, \tag{4.1}$$

where

$$\hat{Y}_{st} = \sum_{h=1}^{L} \hat{Y}_h = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} \hat{Y}_{hi} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} \sum_{k=1}^{L_{hi}} M_{hik} \bar{y}_{hik}$$

and $\bar{y}_{hik} = \frac{1}{m_{hik}} \sum_{j=1}^{m_{hik}} y_{hikj}$.

$\hat{X}_{st}$ is defined similarly.

Variance Estimator. In this paper, we compare two methods of variance estimation – the closed form of the design-based approach and the Jackknife; a common resampling method.

Estimated Variance - Closed Form. There is a closed form for the variance of $\hat{Y}_{Rc}$. We refer to Cochran (1977) and elaborate the standard variance formula as follows:

$$V(\hat{Y}_{Rc}) = \sum_{h=1}^{L} \left[\frac{N_h^{\,2}}{n_h}(1-f_{h1}) \frac{\sum_{i=1}^{N_h}(D_{hi}-\bar{D}_h)^2}{N_h-1} + \frac{N_h}{n_h} \sum_{i=1}^{N_h} \sum_{k=1}^{L_{hi}} \frac{M_{hik}^{\,2}(1-f_{h2ik}) S_{dh2ik}^{\,2}}{m_{hik}} \right] \quad (4.2)$$

where

$f_{h1} = \frac{n_h}{N_h}$ (subscript 1 means the first-stage)

$f_{h2ik} = \frac{m_{hik}}{M_{hik}}$ (subscript 2 means the second-stage)

$D_{hi} = Y_{hi} - RX_{hi}$

$\bar{D}_h = \frac{\sum_{i=1}^{N_h} D_{hi}}{N_h}$

$S_{h2ik}^{\,2} = \frac{1}{M_{hik}-1} \sum_{j=1}^{M_{hik}} \left[\left(y_{hikj} - Rx_{hikj}\right) - \left(\bar{Y}_{hik} - R\bar{X}_{hik}\right) \right]^2$

The estimated variance is

$$v(\hat{Y}_{Rc}) = \sum_{h=1}^{L} \left[\frac{N_h^{\,2}}{n_h}(1-f_{h1}) \frac{\sum_{i=1}^{n_h}(\hat{D}_{hi}-\hat{\bar{D}}_h)^2}{n_h-1} + \frac{N_h}{n_h} \sum_{i=1}^{n_h} \sum_{k=1}^{L_{hi}} \frac{M_{hik}^{\,2}(1-f_{h2ik}) s_{d'h2ik}^{\,2}}{m_{hik}} \right] \quad (4.3)$$

where

$\hat{D}_{hi} = \hat{Y}_{hi} - \hat{R}\hat{X}_{hi}$

(note $\hat{Y}_{hi} = \sum_{k=1}^{L_{hi}} M_{hik}\bar{y}_{hik}$ and $\hat{X}_{hi} = \sum_{k=1}^{L_{hi}} M_{hik}\bar{x}_{hik}$)

$\hat{\bar{D}}_h = \frac{\sum_{i=1}^{n_h} \hat{D}_{hi}}{n_h}$

and

$s_{d'h2ik}^{\,2} = \frac{1}{m_{hik}-1} \sum_{j=1}^{m_{hik}} \left[\left(y_{hikj} - \hat{R}x_{hikj}\right) - \left(\bar{y}_{hik} - \hat{R}\bar{x}_{hik}\right) \right]^2$.

Estimated Variance – Jackknife. As one of the resampling methods, the Jackknife is flexible and simple to implement in complex sample designs. We refer to Wolter (1985) for the Jackknife method where $\hat{\theta}$ is the estimate from the full sample, i.e. $\hat{\theta} = X\frac{\hat{Y}_{st}}{\hat{X}_{st}}$, the same as (4.1). Let $\hat{\theta}_{(hi)}$ denote the estimator of the same functional form as $\hat{\theta}$ obtained after deleting the i^{th} PSU in the h^{th} stratum from the sample.

Define the "pseudovalue" $\hat{\theta}_{hi}$ as

$\hat{\theta}_{hi} = (Lw_h + 1)\hat{\theta} - Lw_h\hat{\theta}_{(hi)}$

where

$w_h = (n_h - 1)(1 - n_h / N_h)$.

Note that if the dropped PSU is in a certainty stratum, the pseudovalue $\hat{\theta}_{hi}$ is the same as the value of $\hat{\theta}$ calculated from the full sample, i.e., $\hat{\theta}_{hi} = \hat{\theta}$, since $w_h = (n_h - 1)(1 - n_h / N_h) = 0$.

The Jackknife estimator of θ is defined by

$$\hat{\theta}^1 = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \hat{\theta}_{hi} / Ln_h \quad (4.4)$$

One version of the Jackknife variance estimator is defined as

$$v_J(\hat{\theta}) = \sum_{h=1}^{L} \frac{w_h}{n_h} \sum_{i=1}^{n_h} (\hat{\theta}_{(hi)} - \hat{\theta}_{(h.)})^2 , \quad (4.5)$$

where

$\hat{\theta}_{(h.)} = \sum_{i=1}^{n_h} \hat{\theta}_{(hi)} / n_h$.

$v_J(\hat{\theta})$ is approximately unbiased for both $Var(\hat{\theta})$ in (4.1) and $Var(\hat{\theta}^1)$ in (4.5).

5. Two Issues in the Calculations

In our setting, there are two issues in the standard estimation of variance. The first involves how to treat the certainty PSU when estimating the variance using the Jackknife approach. In the Jackknife approach, the second-stage variance of the certainty PSU is often ignored. But in our business sampling case, the second-stage variance of the certainty PSU can be relatively large and have a significant influence on the overall variance, and therefore should be included. In order to do so, this portion of the variance must be calculated separately. The variance of the certainty PSU can be calculated using either a closed form, if possible or, a resampling method. At this point, this becomes fairly easy because it is simply a one-stage design. This analogue applies to sample designs with more than two stages where the units used at the first-stage of subsampling are the basis for the formation of replicates in order to calculate the variance of the PSU (Wolter, p. 31).

The second issue deals with the calculation of the degrees of freedom. Typically for larger samples and populations, a 'rule of thumb' method is used to calculate the degrees of freedom; which is the number of sampled PSUs minus the number of PSU strata. A better estimate of the degrees of freedom here is to use the Satterthwaite adjustment[3]. We calculate the Satterthwaite degrees of freedom by assuming the usual assumptions of normality and independence.

$$DF = \frac{2E(v(\hat{\theta}))^2}{V(v(\hat{\theta}))} = \frac{\left(\sum_{h=1}^{2} v_h(\hat{\theta})\right)^2}{\sum_{h=1}^{2} \frac{(v_h(\hat{\theta}))^2}{n_h - 1}} \qquad (5.1)$$

where $v_h(\hat{\theta})$ is the estimated variance of the h^{th} PSU stratum. The normal assumption may not hold, but the Satterthwaite approximation still seems to work well in our simulation.

6. Simulation Results

For each scenario in Table 2, we drew 1,000 samples and calculated the estimated qualified amount and its corresponding estimated variance using both the closed form and Jackknife methods.

Bias Comparison. The relative bias, $\frac{\hat{Y} - Y}{Y}$, is calculated for each of the 1,000 samples. Here Y is the true qualified amount and $\hat{Y}$ is the estimate. The 1,000 samples are first arranged in the order of increasing values of the relative bias and then grouped into ten sets of the 100 samples. The first group consists of the 100 samples with the largest relative negative biases, and the last group contains the 100 samples with the largest relative positive biases. The average bias for each of these ten groups is then calculated and compared across each scenario and estimation method. In general, the biases of the Jackknife estimates are slightly larger than those of the closed form estimates for each scenario, but overall the two methods perform very similar in terms of bias. The significant bias differences occur across different scenarios or SSU sample sizes. Therefore, Figure 5 only presents the bias comparison of four scenarios from the closed form calculation. Compared to the benchmark of the full SSU sample, the scenario of 10 units per SSU stratum has a significantly larger average bias; about 1.5 times that of the benchmark for the largest negative and positive bias groups. The scenarios consisting of 30 and 60 units per SSU stratum are much closer to the benchmark in terms of the average relative bias.

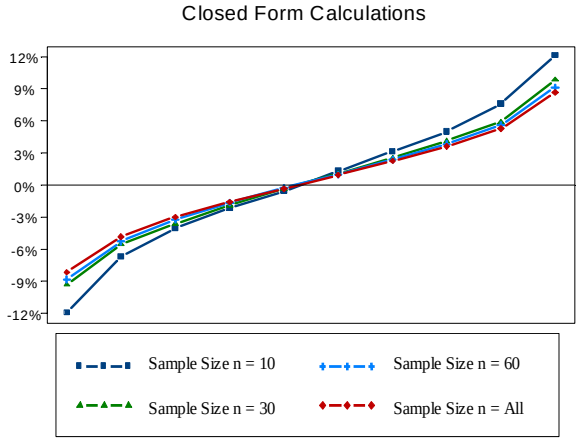

Figure 5. Relative Bias Comparison of Four Scenarios

Relative Precision Comparison. Another way of measuring the closeness between the estimated qualified amount $\hat{Y}$ and the true qualified amount Y is to use the relative width of the confidence interval or relative precision, defined as $\frac{t(df)\sqrt{v(\hat{Y})}}{Y}$. The df is 6, by 'rule of thumb,' which is used in a large sample setting. In the small sample situation, the Satterthwaite approximation method (5.1) is used to have better, or a more conservative, coverage. The relative precision at the 90 percent confidence level was calculated for each of the 1,000 samples. Following the same methodology as in our bias comparisons, the 1,000 samples were first arranged in order of increasing values of the relative precision and then grouped into ten sets of 100 samples. The first group consists of the

[3] Additional references can be found in Rust and Rao (1996).

Table 3. Average Relative Precision by Group for Different SSU Sample Sizes and Different Variance Estimation Methods

Group	$m_h = 10$		$m_h = 30$		$m_h = 60$		$m_h =$ Full	
	Closed	Jackknife	Closed	Jackknife	Closed	Jackknife	Closed	Jackknife
1	7.9%	7.0%	6.5%	6.0%	6.1%	5.9%	5.9%	5.9%
2	9.7%	8.9%	8.1%	7.8%	7.8%	7.8%	7.6%	7.8%
3	10.7%	10.1%	9.0%	8.7%	8.7%	8.6%	8.4%	8.5%
4	11.6%	11.0%	9.7%	9.4%	9.3%	9.3%	8.9%	9.0%
5	12.4%	11.8%	10.3%	10.1%	9.8%	9.8%	9.4%	9.5%
6	13.3%	12.6%	11.0%	10.8%	10.4%	10.4%	9.8%	10.0%
7	14.2%	13.5%	11.6%	11.6%	11.0%	11.0%	10.3%	10.5%
8	15.3%	14.7%	12.4%	12.4%	11.6%	11.6%	10.9%	11.0%
9	17.0%	16.4%	13.3%	13.2%	12.4%	12.5%	11.5%	11.6%
10	20.7%	19.8%	15.6%	15.5%	14.0%	14.3%	12.4%	12.8%

100 samples with the smallest relative precision, and the last group contains the 100 samples whose relative precision levels are the largest. Then, for each of the ten groups, we calculated the average relative precision $\frac{1}{100}\sum_{i=1}^{100}\frac{t(df)\sqrt{v(\hat{Y})}}{Y}$. Table 3 displays the relative precision by group for both the closed form and the Jackknife methods. As shown in Table 3, the relative precision decreases as the SSU sample size increases. The settings of $m_h = 30$ and $m_h = 60$ produce relative precisions that are very close to the full SSU sample size setting. But the total SSU sample sizes for the settings of $m_h = 30$ and $m_h = 60$ differ very much from that of full SSU setting. As shown in Table 2, the sample sizes for the settings of $m_h = 30$ and $m_h = 60$ are 690 units and 1,380 units respectively while the full SSU sample size is 36,421 units on the average.

Coverage Rate Comparison. The coverage rate is a measure closely related to relative precision. Table 4 gives the coverage rate, the proportion of confidence intervals that contain the true population total *Y*, for our simulation results calculated for a 90 percent confidence interval.

Table 4. Coverage Rate for Different SSU Sample Sizes and Different Variance Estimation Methods

Estimation Method	SSU Stratum Sample Size m_h			
	10	30	60	Full
Closed	89.6%	88.6%	89.0%	90.3%
Jackknife	87.7%	88.0%	89.1%	90.6%

As shown in Table 4, the different estimation methods and different SSU sample sizes result in minor differences in the coverage rate. The Jackknife results were calculated by using both the point estimate and variance estimate calculated from the Jackknife method. In practice, it is often the case that the point estimate is calculated by use of a closed form calculation and the variance of the estimate is calculated by using the Jackknife. This combined use basically causes no change in Table 4.

Effect of Adjustments Based on the Two Issues. The coverage rates shown in Table 4 are calculated using the Satterthwaite adjustment for degrees of freedom along with the additional variance adjustment of the certainty PSU for the Jackknife variance calculation. To see the overall impact of these adjustments individually, Table 5 presents the coverage rates calculated with and without the Satterthwaite adjustment to the degrees of freedom, and with and without accounting for the second-stage variance for the certainty PSU. The table shows that the Satterthwaite adjustment improves the coverage rate for both the closed form and the Jackknife methods. It also shows that variance adjustment of certainty PSU improves the coverage rate for the Jackknife estimation.

Table 5. Coverage Rate for Different Settings

Estimation Method	Degrees of Freedom	Variance Adjustment	SSU Stratum Sample Size			
			10	30	60	Full
Closed Form	6		88.4%	87.6%	88.3%	88.6%
Closed Form	Satterthwaite		89.6%	88.6%	89.0%	90.3%
Jackknife	6	No	85.6%	85.9%	87.8%	88.9%
Jackknife	Satterthwaite	No	87.0%	87.7%	88.8%	90.6%
Jackknife	6	Yes	86.3%	86.2%	88.0%	88.9%
Jackknife	Satterthwaite	Yes	87.7%	88.0%	89.1%	90.6%

7. Conclusion

Simulations were also performed on population data that has less variation at both stages. The outcome of this analysis gave similar results to those presented in this paper.

Through the results of our analysis, we can conclude that:

- Both the closed from and the Jackknife estimation methods perform similarly, especially as the SSU sample sizes get larger. The variance estimation of the closed form is very complicated for a two-stage stratified sampling and gets even more difficult for additional stages. On the other hand, the Jackknife formula for variance estimation is a more straightforward for a multi-stage sample design. Therefore, a Jackknife estimation method seems to be a good choice for a multi-stage sample design

- The choice of a secondary sample size depends on the population distribution itself. In this type of setting, sampling somewhere between 30 and 60 units per SSU stratum seems to provide us with reasonable estimates needed for our business sampling situation. In other words, less than 4 percent of the SSU units need to be reviewed, compared to a 100 percent review, in order to achieve reasonable estimates.

- The Satterthwaite approximation for degrees of freedom definitely helps improve the coverage rate and should be used.

- The second-stage variance for the certainty PSU should be counted in the Jackknife variance estimation in order to get a more accurate variance estimate.

8. References

1. Cochran, W.G. (1977). *Sampling Technique*, 3rd ed. New York: Wiley.
2. Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press.
3. Rust, K.F & Rao, J.N.K. (1996). Variance Estimation for Complex Surveys Using Replication Techniques. *Statistical Methods in Medical Research*, 5: 283-310
4. Särndal , C.E., Swensson, B. & Wretman, J. (1991). *Model-Assisted Survey Sampling*. New York: Springer-Verlag.
5. Valliant, R., Dorfman, A. H. & Royall, R. M. (2000), *Finite Population Sampling and Inference, a Prediction Theory*, New York: Wiley
6. Wolter, Kirk (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag