# PPS VS STRATIFIED SAMPLING IN MODERN AUDITS

Wendy Rotz, Eric Falk, Ryan Petska, Jinhee Yang,
Ernst & Young LLP, 1225 Connecticut Ave. NW, Washington DC 20036

**Key words:** pps, stratified sampling, rare events

**Introduction:** Audit situations often involve sampling both to detect and quantify rare events. Today's audits are used to estimate Medicare overpayments, amounts owed to a state in unclaimed funds, or to estimate tax credits and deductions. All involve sampling from a record file where the item of interest may occur in less than twenty percent of the records, perhaps even less than ten percent. A dollar estimate with a reasonable confidence interval is desired and, due to regulator oversight, textbook methodologies are preferred.

Typically, stratified random sampling is used but probability proportionate to size (pps) sampling is a long known alternative. It is more commonly used for rare event detection and to place a, sometimes rather extreme, upper bound on an estimated error amount. However, if there are enough errors contained in the data, pps sampling can also be used to create a two-sided confidence interval based on the normal distribution. According to Roberts[1], a general rule of thumb is there must be at least 20 occurrences of the event before a normal distribution assumption holds in pps. This somewhat limits how "rare" the event can be when sample size budget constraints are a consideration.

Yet confidence intervals based on stratified estimates are also imperfect because the variates under consideration are typically far from normally distributed, with the majority of their values zero and a small number of non-zero values. However, Roberts gives two conditions when the estimated values are approximately normal: 1) the error rates are at least five percent and the dollar size of the errors is relatively small, and 2) the error rates are at least 30 percent but the dollar size of the errors is moderate or large.[2]

This paper explores pps versus stratified random sampling for various distributions of errors found in common modern audit settings. Standard estimators, such as mean per unit (MPU), ratio, and regression estimators are considered from stratified samples and compared to pps samples of equal sizes. Three populations of varying size were created. Within each population three samples of different sizes were selected. Four types of error distributions were considered and each type was considered with three different error rates.

**Population:** The population design variable, *X*, was simulated with a gamma distribution because we found this most closely approximated the typical populations we are finding in our audit settings. For the sake of simplicity, we limited our simulation analyses to three population sizes: 5,000, 50,000, and 150,000 for a small, medium, and large population. The populations are nested.

**Designs:** Three sample sizes were considered: 150, 300, and 600. For each sampling population, a stratified random sample was designed with strata boundaries based on *X* from the medium size population of 50,000. The cumulative square root of the frequency method was used to define the strata boundaries. The same stratum definitions for the medium population were used for the small and large population, adjusting optimum allocation to account for the different size certainty strata in the three populations.

For the stratified samples, a minimum sample size of 20 was set for non-certainty strata, realizing that in a few of the simulations some strata may have little or no errors. This may happen in practice (usually when a higher error rate was expected during the design stage than occurred in the actual sample). We did not employ advanced techniques to weight these strata.[3] Part of our study was to determine how detrimental this is to estimation when we only consider standard approaches. We anticipated that the estimates may be less stable and we would underestimate the variance when the error rate is low.

[1] Roberts, D. M., *Statistical Auditing,* American Institute of Certified Public Accountants, Inc, New York, 1978, p. 117

[2] Roberts, D. M., *op.cit.*, p. 104

[3] Liu, Y., Batcher, M., and Rotz, W., "Application of the Hypergeometric Distribution in a Special Case of Rare Events", 2001 Proceedings of the American Statistical Association, Section on Survey Methodology

The pps samples were drawn with replacement using a common approach of applying a skip interval to the cumulative dollar amount size. Some large records were selected multiple times, but there were not many and since a skip interval was used, we could determine the large units would usually only be selected once or twice, never more than 3 times. Therefore, in accordance with Kish[4] the pps sample was unstratified. An area for future study is to determine when and the degree to which estimation might be improved with a stratified pps approach.

**Estimation Variables:** All of the variables considered for estimation are from mixed distributions and are based on the types of variables we see in practice. In addition, all are based on the assumption there is a constant probability, *p*, that the dependent variable, y, will be an "error" but the distribution of the dollar values in error are different for each variable. Another area for future study is to allow *p* to vary loosely related to the size of the *X*. Three constant levels of *p* (*5%*, *10%*, and *20%*) were considered for four different types of error distributions making a total of twelve estimation variables studied.

The four types of error distributions were again based upon findings in audit settings: Type A) an "All or Nothing" variable where the error amounts are equal to the design variable, *X*; Type B) an "Almost Flat" amount where the errors vary uniformly about a constant; Type C) a "Linear" amount where the errors are a linear combination of *X*; and Type D) a "Mixed" amount where the errors are frequently equal to *X*, but can vary from zero up to *X*. For the sake of simplicity, no negative errors (understatements) were considered. Type A and Type D are the most common types of variables we have encountered. Type B and Type C are somewhat extreme cases of atypical situations found in practice. Since the Type B errors are unrelated to *X*, it is expected that both stratified and pps samples would have poor estimates. See Figure 1 for a visual description of the types of error distributions used in this study.

The dependent variables were simulated as follows:

Let $I_i = 1$ if the ith record is in "error" and let $I_i=0$ otherwise, where $i=1,2,\ldots N$ and $\boldsymbol{N}$ is the population size. Then p is the probability that $I_i= 1$. Let $u_{1i}$ and $u_{2i}$ be iid random variables uniformly distributed between zero and one. Let $e_i$ be iid $N(0,1)$. For the Type D variables, let $I_{2i} =1$ if the error amount is $x_i$, and $I_{2i} =0$ otherwise. Furthermore, let $P(I_{2i} =1)=80\%$. Then the mathematical expressions of our four distributions are:

Type A) All or Nothing: $Y_i = I_i X_i$ ;
Type B) Almost Flat**:** $Y_i = I_i(40 +20u_{1i})$
Type C) Linear: $Y_i = I_i (400 +0.1 X_i +200e_i)$ and
Type D) Mixed: $Y_i = I_i (I_{2i} X_i +(1\text{-}I_{2i})u_{2i}X_i)$ .

Note that all of the estimation variables have mixed distributions. The first three are a mixture of two distributions: 1) a large group of zeros when $I_i$ =0, and 2) a smaller set of non-zero values when $I_i$ =1. The forth type is a mixture of three distributions: 1) a large group of zeros, 2) a small set of values equal to x, and 3) an even smaller set of values varying between zero and x. For each of the four types of distributions, there are three sets of variables for the three levels of *p*. Figure 1 shows the four error distributions used in this study. Figure 2 illustrates the data structure being used within each of the three population sizes.

**Simulation**

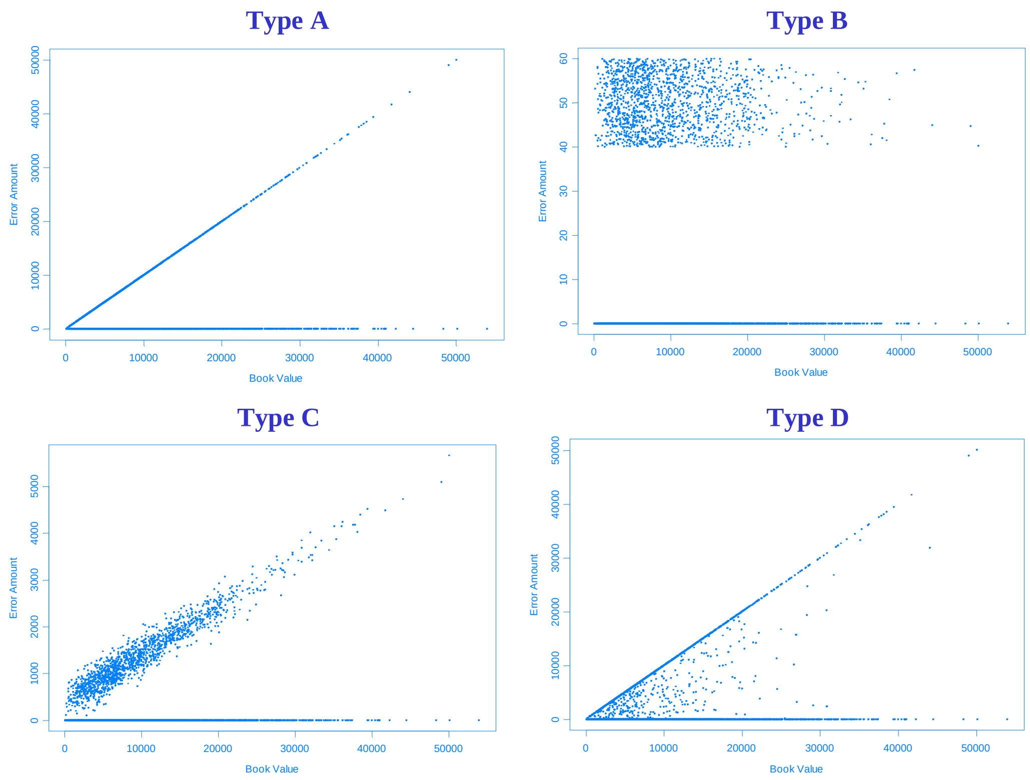
One thousand simulations were run under pps and stratified sample designs for each of the three populations and three sample sizes, making a total of 1,000 · 2 · 3 ·3 = 18,000 simulations. Each simulation included selecting a sample size from a sampling population and producing estimates. Estimates of the total error amount were made for the twelve variables studied to make sets of one thousand estimates for each variable under each design for each population and sample size. Each of the four types of error distributions had 27 scenarios tested (three levels each of error rates, population sizes and sample sizes to make 3 · 3 · 3 = 27 scenarios. Stratified estimates were calculated using MPU, ratio, and regression estimators. Estimates for the pps sample were based on standard methodology.[5]

We found that all three stratified estimators behaved about the same in this study, so unless specified otherwise in the results below, they are referred to collectively as the stratified estimates.

[4] Kish, L. *Survey Sampling* John Wiley and Sons, Inc. New York pp. 245-246

[5] Roberts, D. M. *Statistical Auditing* American Institute of Certified Public Accountants, Inc New York 1978, p. 117

Figure 1. Four Types of Error Distributions

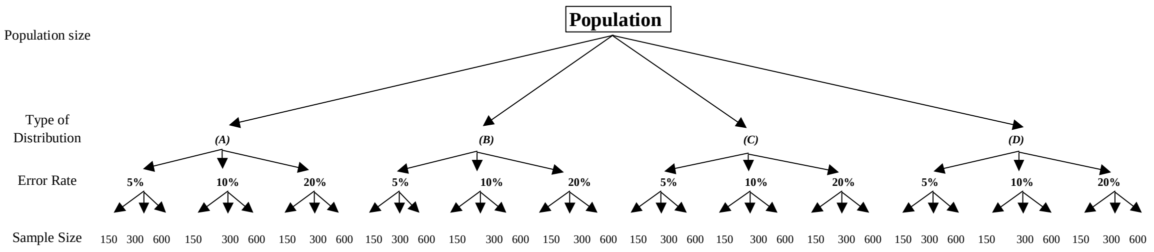

Type A

Type B

Type C

Type D

Figure 2. Data Structure



**Comparison:** The Mean Square Errors (MSE) and percent of samples containing the true population value in their confidence intervals are compared across all of the estimates as well as the distribution of the estimates. In addition, the confidence interval width, for each estimator type, was compared. The distribution of the amounts selected for the pps samples were also compared to the stratified sample. We noted the average number of pps sample selections that fell in each stratum to determine how different the pps selections are from a stratified random sample.

**MSE Results:** The pps estimates had marginally smaller MSEs than the stratified estimates for the Type A and Type D distributions. However, the PPS MSEs were substantially larger for the Type B and C variables. As expected, the MSEs for all estimators were generally smaller for larger sample sizes and higher sampling fractions (smaller population sizes). The MSEs were also smaller for smaller error rates and therefore smaller total errors.

In addition, we found the pps method underestimated the true value about 60 to 70 percent of the time with Type B and C variables. There was a tendency for the stratified estimates to underestimate the true value.

We found in most scenarios, stratified methods underestimated the true value about 50 to 55 percent of the 1,000 simulations. This observation prevailed, even for larger sample sizes and larger error rates, although the percentages were closer to 50 percent.

**Width of Confidence Intervals:** The median width of the confidence interval for pps was shown to be consistently smaller than the width of the confidence interval for the stratified estimators. As expected, the median confidence interval width was smaller for larger sample sizes. The median confidence interval width increased as the error percentage increased.

**PPS 90% Confidence Interval Coverage:** It was expected that roughly 90 percent of the confidence intervals would contain the true value of the parameter estimated. However, both pps and stratified intervals rarely demonstrated this in the hypothesis tests of $p \geq .90$ with $\alpha=.05$ and n=1,000. The normality of the sampling distribution and estimate of the variance are a factor.

In each of the twenty-seven pps scenarios for type B estimates the one thousand pps "90%" confidence intervals contained the true value in only 60 to 70 percent of the simulations (larger coverage with larger sample sizes). Poor performance was expected in Type B estimates because these had nearly constant errors unrelated to *X*. However, the pps Type C "90%" confidence intervals only covered the true value 70 to 80 percent of the time and these variables had linearly related errors.

The pps coverage was better for Type A and D variables. The percent coverage was close to 90 percent and even well exceeded it in many instances with the small population of 5,000 records.

**Stratified 90% Confidence Interval Coverage:** With a few exceptions noted below, the coverage of the stratified "90%" confidence intervals was roughly 85 to 90 percent, occasionally exceeding 90 percent by a small amount. The only percentage below 80 percent occurred for the worst-case scenario, which was the smallest sample of 150 from the largest population size of 150,000 with the smallest error rate of 5 percent on the most problematic variable Type B. In this scenario, all three estimators: MPU, ratio, and regression only had about 72 percent coverage. Also all three estimators typically had about 80 to 87 coverage for the Type B and C variables with an error rate of 5 percent. However, coverage in the stratified confidence intervals out-performed pps coverage in all Type B and C scenarios. Although the stratified confidence intervals generally had coverage just under 90 percent, they were consistently closer to 90 percent than pps.
The causes of under-coverage are confounded. There are normality considerations discussed in more detail further below. We used Satterthwaite's approximation to the degrees of freedom to somewhat account for the complexity of the design. However this approximation is imperfect because it assumes normality of $Y_i$, not just $\hat{Y}$ and as noted the $Y_i$ values are far from normally distributed. In most of the scenarios, even those with small sample sizes, the t-value was not much more than 1.7 compared to the normal value of 1.645 (at the 90% confidence level), so the use of Satterthwaite's approximation only had a small influence. Widening the confidence intervals according to Satterthwaite's approximation was more conservative than using 1.645; however, it may not have been conservative enough.

The estimate of the variance is another issue. Recall that sparse stratum errors were expected and raised the concern of under-estimating the variance. Since actual population variances and covariances were known in the simulations, the variance of the estimates was calculated using known population strata data. These true variances were compared to the variances estimated from the sample. It was found that we were more likely to under-estimate the variance than over-estimate it. However, the under estimation of the variance was only mild for Type A and D variables; the mean square error of the estimated variance divided by the square of the variance was about 20 percent or less when *p=5%*, under 10 percent when *p=10%* and under 5 percent when *p=20%.* For Type B and C variables, it was often more than 40 percent when *p=5%*, but usually under 20 percent when *p=10%* and under 5 percent when *p=20%.* Therefore, as the error rate is going up we are doing a better job of estimating the variance.

Despite the confounding causes of under-coverage of the confidence intervals, namely normality, possible insufficiency of Satterthwaite approximation, and underestimating variances, the overall compounded effect is usually only a slight under-coverage.

**Normality of Sampling Distributions:** To partially assess causes for undercoverage of the 90 percent confidence intervals, the sampling distributions of the pps and stratified estimates were tested for normality using the Kolmogorov-Smirnov test statistic calculated in SAS with *n=1,000* and *α=.05*. In the *108* pps tests, twenty-seven scenarios for four types of variables, only 5 of the 108 tests did not reject a hypothesis of normality. In light of so many failures, before writing off these five cases as Type II errors, we note that they were all from the Type D distribution, and occurred for the medium and large population and the medium and large error rate.

The results seem to indicate that the normality assumption of pps estimates needs further investigation. Clearly, many of our scenarios had more than 20 errors, yet whether the sample size was 150 or 600 and whether the error rate was 5 percent or 20 percent almost all scenarios and all types of errors failed a normality test for the distribution of the pps estimate. In spite of these failures, however, pps did exhibit good confidence interval coverage for Type A and D variables.

For the stratified estimates, about half the scenarios failed the normality test. The normality assumption was more likely to hold for Type A and D variables and when p=20%. However, there were no clear patterns to form a general rule as to when the sampling distributions were normally distributed. Again, despite the normality test failures, the confidence intervals only slightly undercovered in most scenarios.

**Distributions of the PPS Sample Selection:** Both pps and stratified samples select larger records with higher probabilities. We were curious just how different the pps sample distribution was in comparison to the stratified samples. We found that for smaller populations, the pps sample tends to select more large values than the stratified sample. As the population increases, the distribution of pps sample selections look more like those of a simple random sample, while the stratified sample remains highly concentrated in the tails.

**Conclusion:** We did not observe overwhelming evidence to convert from stratified to pps approaches. Although pps may have some merits with Type A and D distributions, stratified approaches performed more reliably for all types of estimates and would be preferred if the type of error distribution is unknown.

**Next Steps:** The analysis may be taken further by considering stratification with pps and/or incorporating error rates that vary in relation to *X*. Replicated variance estimates, although not a standard textbook method, are becoming more common, may improve upon the stratified estimate confidence intervals, and may be a viable solution to the textbook closed formulas with assumptions that do not quite apply in these audit settings.

**Acknowledgements:** The authors would like to thank the following people for their thoughts, aid, and efforts in completing this study: Kellie Wills, Ru Sun, Archana Joshee, Kevin Messina, Carla Inclan, David McGrath, Susan Garille-Higgins, John Matson, Greg Densmore, and Taras Smetaniouk.

**References:**

1. Cochran, W. G. Sampl*ing Techniques*, pp. 250-259 & 294-299, 3rd ed. New York, NY.: John Wiley & Sons, Inc., 1977.
2. Guy, D. M.; Carmichael, D.R. and Whittington, O. R. *Audit Sampling An Introduction*, pp.29, 34 & 184-198, 4th ed. New York, NY: John Wiley & Sons, Inc. 1998.
3. Hansen, M.H.; Hurwitz, W. N. and Madow, W. G., *Sample Survey Methods and Theory*, pp.341-348, 362, 391, 522 & 605, Vol. I. New York, NY: John Wiley & Sons, Inc. 1993.
4. Hansen, M.H.; Hurwitz, W. N. and Madow, W. G., *Sample Survey Methods and Theory,* pp. 62 & 213, Vol. II. New York, NY: John Wiley & Sons, Inc. 1993.
5. Kish, L., *Survey Sampling*, pp. 220-247. New York, NY: John Wiley & Sons, Inc. 1965.
6. Lohr, S. L., *Sampling: Design and Analysis*, pp. 190 & 211-212. Pacific Grove, CA: Brooks/Cole Publishing Company 1999.
7. Liu, Y., Batcher, M., and Rotz, W. "Application of the Hypergeometric Distribution in a Special Case of Rare Events", 2001 Proceedings of the American Statistical Association, Section on Survey Methodology
8. Neter, J. and Loebecke, J. *Behavior of Major Statistical Estimators in Sampling Accounting Populations an Empirical Study* American Institute of Certified Public Accountants Inc. New York
9. Roberts, D. M., *Statistical Auditing*, pp. 63-64 & 214-216. New York, NY: American Institute of Certified Public Accountants, Inc. 1978.
10. Valliant, R.; Dorfman, A. H. and Royall, R. M., *Finite Population Sampling and Inference A Prediction Approach*, pp. 74-77. New York, NY: John Wiley & Sons, Inc. 2000.
11. Wilburn, A.J., *Practical Statistical Sampling For Auditors*, pp. 112-117 & 122-125. New York, NY: Marcel Dekker, Inc. 1984.
12. Statistical Modeling and Analysis in Auditing, Panel on Nonstandard Mixtures of Distributions, Statistical Science, 1989, Vol. 4, No. 1, 2-33.