Efficient Sampling Design and Estimation in Audit Data (II)

Yan Liu[1], Mary Batcher[2], Fritz Scheuren[1]
NORC, 1350 Connecticut Ave., NW, Suite 500, Washington, DC 20036[1]
Ernst & Young LLP, 1225 Connecticut Ave., NW, Washington, DC 20036[2]

Abstract

For the special audit data where the qualified invoice amount is somewhere between zero and the full invoice amount, we have discussed in Part I of this paper(Liu, Batcher and Scheuren, 2005) about stratum boundaries, sample size calculation and sample size allocations. In design-based approach, the properties of the two estimation methods - Mean Per Unit estimation and Ratio estimation are well known. Here, we compare the two estimation methods from a different approach. We compare their bias and variance using the realization process. We also perform simulations to compare the balance of number of strata and stratum sample size under different settings.

Key words: mixture distributions; audit sampling; stratified sampling; Mean Per Unit (MPU) estimation; Ratio Estimation; realization process.

1. Introduction

One typical audit situation is that there exists a list of invoice with a known invoice amount. The distribution of the invoice amounts is highly skewed, as shown in Figure 1. For each invoice, there is a qualified amount associated with it, whether qualified for taxable amount or for tax credit. In part I of this paper, we discussed two types of populations that we often face in auditing. One is the special population where invoices are divided into two categories according to whether or not invoices are qualified, called *Population One*. The other population type arises when some invoices have a qualified amount between zero and the full invoice amount, called *Population Two*. Figure 2 and Figure 3 show the scatterplot of the qualified amount against the invoice amount for these two populations.

Assume that the population parameter to be estimated is the total qualified amount. The typical estimation methods used in this type of audit data are Mean Per Unit (MPU) method and Ratio method (or combined ratio method in stratified sampling design).

The properties of these two methods are well known when design-based approach is taken. It is known, for example, that unconditionally, the MPU estimate is unbiased while the ratio estimate is biased. It is also known that the variance estimator of MPU estimate is exact and unbiased while the variance estimator of the ratio estimate is approximate and biased.

Figure 1. Typical Frequency Distribution of Invoice Amount (x)

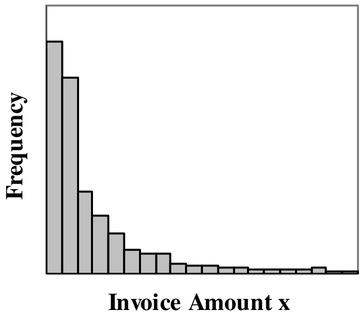


Figure 2. Population One

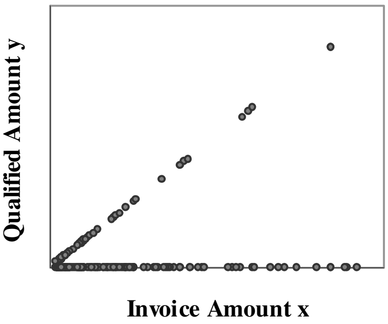
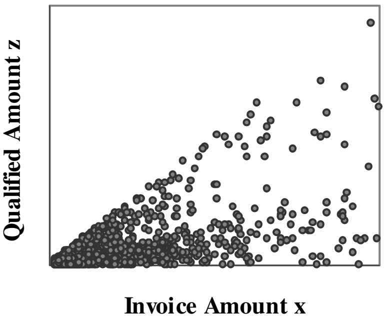


Figure 3. Population Two



For the stratified sampling with small stratum sample size, it is often thought that the MPU estimator is "safer" than the ratio estimator. By "safer", we mean that it has better properties in terms of bias and stability of variance estimation. In this paper, we look at theses properties from a different approach using a "realization process" that generates the audit populations.

2. Realization Process of Population One

To characterize the population distribution, we assume that qualified invoices and non-qualified invoices are randomly distributed among the N population units. Let x_i be the known invoice amount for invoice i and y_i be the unknown qualified amount for invoice i. According to Roberts (1978), the N population units in Population One (Figure 2.) may be characterized as a realization of the following process:

$$y_i = \begin{cases} x_i, & \text{with probability } p \\ 0, & \text{with probability } (1\text{-}p) \end{cases}, \quad i = 1, 2, \cdots, N \tag{2.1}$$

The properties of this process in terms of averages over all possible realizations are denoted as E_p. Some important results for MPU estimator and ratio estimator are given by Roberts (1978).

For the MPU estimator, it can be shown that

$$\hat{Y} = N\bar{y}, \tag{2.2}$$

and for ratio estimator, that

$$\hat{Y}_R = \hat{R}X \tag{2.3}$$

where

$$\hat{R} = \frac{\bar{y}}{\bar{x}}, \quad \bar{y} = \frac{1}{n}\sum_i^n y_i \text{ and } \bar{x} = \frac{1}{n}\sum_i^n x_i .$$

The corresponding variances are

$$V(\hat{Y}) = N(N/n-1)S_y^{\ 2}, \tag{2.4}$$

$$V(\hat{Y}_R) = N(N/n-1)S_d^{\ 2}, \tag{2.5}$$

where $S_y^{\ 2}$ is the variance of y and $S_d^{\ 2}$ is the variance of $d_i = y_i - Rx_i$.

Under the realization process (2.1), Roberts (1978) proves that

$$E_p(S_y^{\ 2}) \approx p\left(S_x^{\ 2} + (1-p)\bar{X}^2\right), \tag{2.6}$$

and

$$E_p(S_d^{\ 2}) \approx p(1-p)\left(S_x^{\ 2} + \bar{X}^2\right), \tag{2.7}$$

when the population size, N, is reasonably large. Combining equations (2.6) and (2.7), we have

$$E_p\left(S_y^{\ 2}\right) = E_p\left(S_d^{\ 2}\right) + p^2 S_x^{\ 2}. \tag{2.8}$$

3. Realization Process of Population Two

As described in Part I of this paper, we also developed a realization process for Population Two. Let z_i (in order to distinguish from y_i in Population One) be the unknown qualified amount of invoice i. There could be many scenarios for Population Two. One scenario is that points of z_i are randomly scattered around the line px_i. This can be characterized as a realization of the following process:

$$z_i = \begin{cases} px_i + u(1-p)x_i, & \text{with probability } p \\ px_i - upx_i, & \text{with probability } (1\text{-}p) \end{cases}, \quad i = 1, 2, \cdots, N \tag{3.1}$$

where u is a random number from *Uniform* $(0,1)$.

The MPU estimator and ratio estimator for the total qualified amount are:

$$\hat{Z} = N\bar{z}, \tag{3.2}$$

$$\hat{Z}_R = \hat{R}Z \tag{3.3}$$

where

$$\hat{R} = \frac{\bar{z}}{\bar{x}}, \quad \bar{z} = \frac{1}{n}\sum_i^n z_i \text{ and } \bar{x} = \frac{1}{n}\sum_i^n x_i .$$

The corresponding variances are

$$V(\hat{Z}) = N(N/n-1)S_z^{\ 2}, \tag{3.4}$$

$$V(\hat{Z}_R) = N(N/n-1)S_{d(z)}^{\ 2}, \tag{3.5}$$

where $S_z^{\ 2}$ is the variance of z and $S_{d(z)}^{\ 2}$ is the variance of $d_i(z) = z_i - Rx_i$.

Under the realization process (3.1), it can be shown that

$$E_p\left(S_z^{\ 2}\right) = E_p\left(S_{d(z)}^{\ 2}\right) + p^2 S_x^{\ 2}. \tag{3.6}$$

The proof of (3.6) is given in the following.

Since $p = E(\hat{R})$, under model (3.1), we have

$$d_i(z) = z_i - Rx_i \approx z_i - px_i$$

Taking the variance on both sides of the equation:

$$z_i - px_i \approx d_i(z),$$

we have

$$S_z^2 + p^2 S_x^2 - 2pCov(z,x) = S_{d(z)}^2. \qquad (3.7)$$

From the realization process (3.1), it is immediate that

$$E_p(zx) = pS_x^2 + p\overline{X}^2.$$

Therefore,

$$\begin{aligned} Cov_p(z,x) &= E_p(zx) - E_p(z)E_p(x) \\ &= pS_x^2. \qquad (3.8) \end{aligned}$$

Substituting equation (3.8) into equation (3.7), we obtain equation (3.6).

Since from Part I of this paper (Liu, Batcher and Scheuren, 2005) it can be shown that

$$E_p(S_{d(z)}^2) = \frac{1}{3} E_p(S_d^2). \qquad (3.9)$$

The variances of the two estimators under the realization process (3.1) can be obtained from equations (2.7), (3.4), (3.5), (3.6) and (3.9).

4. Bias Comparison - MPU Versus Ratio Estimator

Population One. Estimators from the simple random samples in Population One are known to be such that unconditionally, the MPU estimate (2.2) is unbiased while the ratio estimate (2.3) is biased. Conditionally, for the given sample ($x_1, x_{2,\cdots,} x_n$), under realization process characterized by equation (3.1) of Population One, the MPU estimate is conditionally biased while the ratio estimate is unbiased. The conditional biases of $\hat{Y}$ and $\hat{Y}_R$ under the Population One realization process are

$$E_p(\hat{Y} - Y) = Np(\overline{x} - \overline{X}) \qquad (4.1)$$

$$E_p(\hat{Y}_R - Y) = \frac{E_p(\overline{y})}{\overline{x}} X - Y = 0. \qquad (4.2)$$

The MPU estimator would underestimate the population total if $\overline{x} < \overline{X}$; and overestimate if $\overline{x} > \overline{X}$. The audit populations here are typically very skewed to the right and it is often the case that $\overline{x} < \overline{X}$. To keep the bias small, a balanced sample is desired. That is, a sample that satisfies $\overline{x} \approx \overline{X}$.

Population Two. In a way that parallels our Population One results, it can be shown that this same conclusion holds under Population Two, as characterized by equation (3.1). The conditional biases in this second case are:

$$E_p(\hat{Z} - Z) = Np(\overline{x} - \overline{X}) \qquad (4.3)$$

$$E_p(\hat{Z}_R - Z) = 0 \qquad (4.4)$$

Achieving Balance. Stratification on invoice amount x is often used, which can help the sample to be better balanced. Stratum boundaries and sample size calculation are discussed in part I of this paper. The conditional biases under stratification for Population One are:

$$\begin{aligned} E_p(\hat{Y} - Y) &= p\sum_h N_h(\overline{x}_h - \overline{X}_h) \\ &= p(\hat{X}_{st} - X). \qquad (4.5) \end{aligned}$$

$$\begin{aligned} E_p(\hat{Y}_{CR} - Y) &= \frac{E_p(\overline{y}_{st})}{\overline{x}_{st}} X_{st} - Y \\ &= \frac{p\overline{x}_{st}}{\overline{x}_{st}} - Y = 0. \qquad (4.6) \end{aligned}$$

Similarly, for Population Two,

$$E_p(\hat{Z} - Z) = p(\hat{X}_{st} - X). \qquad (4.7)$$

$$E_p(\hat{Z}_{CR} - Z) = 0. \qquad (4.8)$$

5. Variance Comparison – MPU Versus Ratio estimator

The variance of the ratio estimator is smaller than the mean per unit estimator for both a simple random sample design and a stratified sampling design and under both realization processes (2.1) and (3.1). The differences in the variances of the two estimators are given below.

From equations (2.4) and (2.5), the expected variances of the two estimators under simple random sample and realization process (2.1) of Population One are

$$E_p\left(V(\hat{Y})\right) = N(N/n - 1)E_p\left(S_y^2\right), \qquad (5.1)$$

$$E_p\left(V(\hat{Y}_R)\right) = N(N/n - 1)E_p\left(S_d^2\right). \qquad (5.2)$$

Using equations (2.8), (5.1) and (5.2), the variance difference of two estimators is

$$\begin{aligned} E_p\left(V(\hat{Y})\right) - E_p\left(V(\hat{Y}_R)\right) &= N(N/n - 1)p^2 S_x^2 \qquad (5.3) \\ &> 0 \end{aligned}$$

For the stratified sample design, similar conclusions hold:

$$E_p\left(V(\hat{Y}_{st})\right)-E_p\left(V(\hat{Y}_{CR})\right)$$
$$=p^2\sum_h N_h(N_h/n_h-1)S_{hx}{}^2 \qquad (5.4)$$
$$>0$$

Similarly, under the realization process (3.1) of Population Two, we have

$$E_p\left(V(\hat{Z})\right)=N(N/n-1)E_p\left(S_z{}^2\right), \qquad (5.5)$$

$$E_p\left(V(\hat{Z}_R)\right)=N(N/n-1)E_p\left(S_{d(z)}{}^2\right). \qquad (5.6)$$

Using equations (3.6), (5.5) and (5.6), we get the following:

$$E_p\left(V(\hat{Z})\right)-E_p\left(V(\hat{Z}_R)\right)$$
$$=N(N/n-1)p^2S_x{}^2. \qquad (5.7)$$

$$E_p\left(V(\hat{Z}_{st})\right)-E_p\left(V(\hat{Z}_{CR})\right)$$
$$=p^2\sum_h N_h(N_h/n_h-1)S_{hx}{}^2. \qquad (5.8)$$

Note that (5.3), (5.4), (5.7) and (5.8) show that the variance differences of two estimators are the same under both realization processes. But actual variances under the two realization processes are different, which is shown by equation (3.9).

6. Comparison of Sample Estimates Using Simulations

Theoretically, to summarize sections 2 through 5, we have compared MPU estimator and ratio estimator for the special audit populations. From the realization pint of view, we have shown that

- The MPU estimator is biased and the ratio estimator is unbiased,
- The degree of bias depends on the how close $\bar{x}$ is to $\bar{X}$,
- The variance of ratio estimator is smaller than the variance of MPU. estimator

This assumes that the populations follow the realization process models (2.1) or (3.1) exactly.

In practice, real populations do not follow models exactly and we need to estimate the variance using sample data. Since the stratified sample design in often used, an important issue is the balance between the number of strata and the stratum sample size. For example, for a fixed sample size of 90 units, we can use the setting of 9 strata with 10 units per stratum or the setting of 3 strata with 30 units per stratum. A deep stratified sample may result in smaller bias because of $\bar{x}$ closer to $\bar{X}$.

On the other hand, it is intuitive that the stability of the estimated variance needs a larger stratum sample size. In the following simulations, we will look at the bias and the variance estimates of the two estimation methods from sample data under different settings. For this exercise, the typical design-based variance estimates for MPU estimator and for combined ratio estimator are used, see Cochran (1977).

The simulation population includes 3,865 invoices. Four variables are generated as described in Table 1. For a fixed sample size of 90 invoices, two design settings are used, as described in Table 2.

Table 1. Four Simulated Variables

Variable Name	Population Model	Value of p
y_1	Equation (2.1)	0.3
y_2	Equation (2.1)	0.7
z_1	Equation (3.1)	0.3
z_2	Equation (3.1)	0.7

Table 2. Two Design Settings

Total sample size	Number of Strata	Number of Invoices Per Stratum
90	3	30
90	9	10

The stratum boundaries are set up using the results from Part I of this paper:

$$X_h\sqrt{CV_{hx}{}^2+1}=C\,,\ h=1,2,\cdots,L\,. \qquad (6.1)$$

where CV_{hx} is the coefficient of variation of x for stratum h and C is a constant.

For each setting in Table 2, we drew 2,000 samples and calculated the estimated qualified amount and its corresponding estimated variance using both MPU method and combined ratio method

Bias Comparisons. The relative biases $\text{Bias}(y_1)$, $\text{Bias}(y_2)$, $\text{Bias}(z_1)$ and $\text{Bias}(z_2)$ are calculated for each of the 2,000 samples. Here $Bias(y_1)=\hat{Y}_1/Y_1-1$ and so forth. For each variable, the 2,000 bias values are sorted in increasing order; the ordered values of the four variables are merged together in order. To compare the two estimation methods – MPU estimator and ratio estimator, we look at the linear regression of

MPU bias on the ratio bias through the origin for each variable and each sample design setting. The following Table 3 gives the regression coefficient.

Table 3. The Regression Coefficient of MPU Bias on Ratio Bias

Variable	Setting of 3 Strata	Setting of 9 Strata
y_1	1.003	1.026
y_2	1.019	1.132
z_1	1.008	1.077
z_2	1.068	1.365

The value of R^2 is larger that 0.99 for all regressions in Table 3. It shows that ratio estimator has a smaller bias, in general, for both design settings.

Further, we look at the four scenarios of two design settings and two estimation methods. We take the ratio method and 9 strata setting as the benchmark and regress the bias of the each scenario on it. Table 4 gives the coefficient of linear regression through the origin. All the R^2 values are greater than 0.99.

Table 4. The Regression Coefficient of Bias on Bench Mark Bias

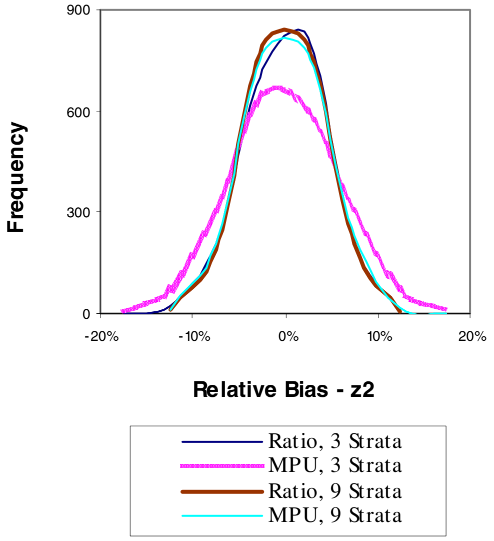
Variable	Ratio 9 Strata	MPU 9 Strata	Ratio 3 Strata	MPU 3 Strata
y_1	1.000	1.003	1.059	1.087
y_2	1.000	1.019	1.039	1.178
z_1	1.000	1.008	1.063	1.146
z_2	1.000	1.068	1.055	1.447

Table 4 indicates that the ratio method in the 9 strata setting has the smallest bias in general; the scenario of MPU and 9 strata and the scenario of Ratio estimation and 3 strata are close; and the MPU estimation and 3 strata is the worst scenario. Table 4 also suggests that if MPU estimator is used, deep stratification can reduce the bias, especially when the ratio type relationship between qualified amount and invoice amount (that is, variable z_2) is stronger.

Note that the simulation comparisons here are in terms of repeated samples. Bias properties by equations (4.5) – (4.8) are in terms of repeated generations of qualified amount y_1 (y_2, z_1 and z_2) by the model for a given sample. Under both scenarios, a better balanced sample and the ratio method reduce bias.

The graph for variable z_2 is given in the following Figure 4. For other variables, the bias distributions are not much different.

Figure 4. Distribution of Bias



Relative Precision Comparisons. Another way of measuring the closeness between the estimated qualified amount and the true qualified amount is to use the relative width of the confidence interval or relative precision, defined as $\frac{t_\alpha(df)\sqrt{v(\hat{Y}_1)}}{Y_1}$ for y_1; with similar definitions for the other variables. The degrees of freedom (df) is 87 for the 3 strata setting and df =81 for the 9 strata setting. 90% confidence level is used. The same comparison technique in bias comparison is used here.

For each variable, the 2,000 relative precision values are sorted in increasing order; the ordered values of the four variables are merged together in order. First, we compare the two estimation methods – MPU estimator and ratio estimator. For each variable and each sample design setting, we fit a linear regression of relative precision of MPU estimate on the relative precision of ratio estimate through origin. The following Table 5 gives the regression coefficient.

Table 5. Coefficient of Regression of MPU Relative Precision on Ratio Relative Precision at 90% C.L.

Variable	Setting of 3 Strata	Setting of 9 Strata
y_1	1.034	1.003
y_2	1.146	1.023
z_1	1.084	1.010
z_2	1.386	1.071

The value of R^2 is larger that 0.99 for all regressions in Table 5. The table shows that ratio estimator has a smaller value of relative precision or a better precision in both design settings.

Next, we look at the four scenarios from two design settings and two estimation methods. We take the ratio method and 9 strata setting as the bench mark and fit the linear regression through origin for the relative precision of each scenario. Table 6 gives the coefficient of linear regression through origin. All the R^2 values are greater than 0.99.

Table 6 indicates that the ratio method in the 9 strata setting has a better precision in general; the scenario of MPU and 9 strata and the scenario of ratio estimation and 3 strata are close; the MPU estimation with 3 strata design is the worst under these scenarios.

Table 6. The Regression Coefficient of Relative Precision on Benchmark Relative Precision at 90% C.L.

Variable	Ratio 9 Strata	MPU 9 Strata	Ratio 3 Strata	MPU 3 Strata
y_1	1.000	1.003	1.064	1.100
y_2	1.000	1.023	1.059	1.213
z_1	1.000	1.010	1.062	1.151
z_2	1.000	1.071	1.062	1.473

Coverage Rate Comparison. The coverage rate is a measure closely related to relative precision. Table 7 gives the coverage rate, the proportion of simulated samples whose confidence intervals contain the true population value at a 90 percent confidence level.

Table 7. Coverage Rate at 90% C.L.

Variable	Ratio 9 Strata	MPU 9 Strata	Ratio 3 Strata	MPU 3 Strata
y_1	89.8%	89.2%	90.0%	89.7%
y_2	89.8%	89.5%	90.3%	91.1%
z_1	89.9%	90.1%	89.5%	89.4%
z_2	90.0%	89.4%	89.6%	90.7%

As shown in Table 7, coverage rates are similar for each variable, each design setting and each estimation method. They are all close to their nominal level of 90%. Therefore, the performance comparison depends on the bias and relative precision.

7. References

Cochran, W. G. (1977), *Sampling Techniques*, New York: John Wiley & Sons

IRS REV PROC.

Liu, Y., Batcher, M. & Scheuren F. (2005). Efficient Sampling Design in Audit Data. *Journal of Data Science, Vol 3, No. 3.*

Roberts, D.M. (1978), *Statistical Auditing*, New York: American Institute of Certified Public Accountants, Inc.