Efficient Sample Design in Special Ratio Type Data

Yan Liu, Mary Batcher and Fritz Scheuren
Yan Liu, Ernst & Young LLP, 1225 Connecticut Ave., NW, Washington, DC 20036

Key words: Mixture Distributions; Audit Sampling; Stratified Sampling; Neyman Allocation; Ratio Estimation

Abstract

Auditors are often faced with reviewing a sample drawn from the special populations where ratio type estimation is appropriate. One is the special population where invoices are divided into two categories according to whether or not invoices are qualified. In other words, the qualified amount follows a nonstandard mixture distribution in which the qualified amount is either zero with a certain probability or the same as the known invoice amount with a certain probability. The other is the population where some invoices are partially qualified. In other words, some invoices have a qualified amount between zero and the full invoice amount. Under these populations, the typical sample design is the stratified random sample design and the estimation method is a ratio type method. We focus on efficient sample design and provide some guidelines in setting up stratum boundaries, calculating sample size and allocating sample size optimally across strata.

1. Introduction

There are two types of population 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. In other words, the qualified amount is either zero or the same as the known invoice amount depending on which category the invoice falls in. This type of populations is called *Population One*. Figure 1 shows the scatterplot of the qualified amount against the invoice amount for population one.

The other type is the population where some invoices have a qualified amount between zero and the full invoice amount. This type of population is called *Population Two*. Figure 2 shows the scatterplot of the qualified amount against the invoice amount for population two.

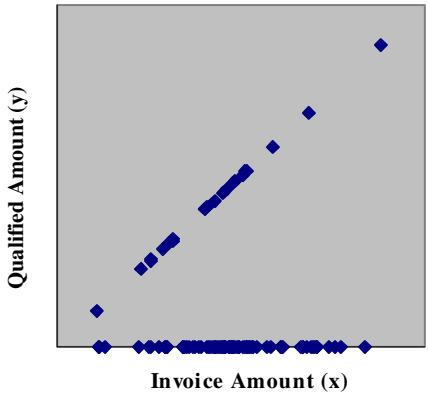

Figure 1. Population One

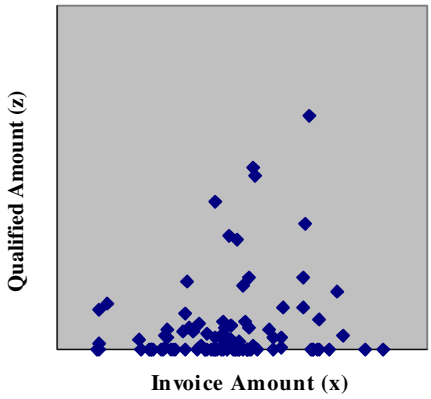

Figure 2. Population Two

For these two populations, the typical sample design is a stratified random sample design using the known invoice amount as the stratifying variable. In this paper, we assume all the 'outliers' are taken with certainty.

We first summarize the characteristics of population one. Suppose that the population includes N invoices and each has a known invoice amount. The invoices are divided into two classes – qualified class C and non-qualified class $\tilde{C}$. If an invoice is in class C, then the qualified amount is equal to its invoice amount; otherwise the qualified amount is zero. In this paper, we assume that the percentage of invoices in one class is in a reasonable range. If the

percentage of invoices in one class is extreme, either very small or very large; a hypergeometric estimation method is recommended (Liu, Batcher and Rotz, 2001).

Further, 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 may be characterized as a realization of the following process:

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

The properties of this process in terms of averages over all possible realizations, denoted as E_p, lead to some useful applications. We first outline these properties summarized by Roberts.

The population parameter to be estimated is the ratio:

$$R = \sum_{i=1}^{N} y_i \Big/ \sum_{i=1}^{N} x_i \tag{2}$$

The corresponding sample estimate under simple random sample is:

$$\hat{R} = \frac{\bar{y}}{\bar{x}}, \tag{3}$$

where $\bar{y} = \frac{1}{n}\sum_{i}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$.

The variance of $\hat{R}$ is:

$$V(\hat{R}) = \frac{(1-f)}{n\bar{X}^2} S_d{}^2, \tag{4}$$

where $S_d{}^2$ is the variance of $d_i = y_i - Rx_i$ and $S_d{}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - Rx_i)^2$.

Under the realization process of population units described in equation (1),

$$E_p(R) = p \tag{5}$$

and

$$E_p(S_d{}^2) \approx p(1-p)\left(S_x{}^2 + \bar{X}^2\right) \tag{6}$$

when the population size, N, is reasonably large.

We now expand the above properties to population two where some invoices are partially qualified. Suppose the average ratio is still $E_p(R) = p$; there are many scenarios of the relationship between the qualified amount, denoted as z_i (in order to distinguish it from y_i in population one), and the invoice amount x_i. One scenario is that points of z_i are randomly scattered around the line px_i. So the N population units can be characterized as a realization of the following process:

$$z_i = \begin{cases} px_i + u(1-p)x_i, & \text{with probability } p \\ px_i - upx_i, & \text{with probability } (1-p) \end{cases},$$

$$i = 1, 2, \cdots, N \tag{7}$$

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

Under the realization process of population units described in equation (7), we still have $E_p(R) = p$ for the ratio $R = \sum_{i=1}^{N} z_i \Big/ \sum_{i=1}^{N} x_i$. Corresponding to formula (4), the variance of $\hat{R} = \frac{\bar{z}}{\bar{x}}$ is $V(\hat{R}) = \frac{(1-f)}{n\bar{X}^2} S_{d(z)}{}^2$. Now, $S_{d(z)}{}^2$ is the variance of $d_i(z) = z_i - Rx_i$. Rewrite $d_i(z)$ as

$$d_i(z) = z_i - Rx_i \approx z_i - px_i$$
$$= \begin{cases} u(1-p)x_i, & \text{with probability } p \\ -upx_i, & \text{with probability } (1-p) \end{cases},$$

$$i = 1, 2, \cdots, N \tag{8}$$

Rewrite d_i in population one as:

$$d_i = y_i - Rx_i \approx y_i - px_i$$
$$= \begin{cases} (1-p)x_i, & \text{with probability } p \\ -px_i, & \text{with probability } (1-p) \end{cases},$$

$$i = 1, 2, \cdots, N \tag{9}$$

Comparing equations (8) and (9), we have

$d_i(z) = ud_i$.

Therefore,

$$E_p(S_{d(z)}{}^2) = \frac{E_p(S_d{}^2)}{3} \quad (10)$$

Note that most scenarios of population two fall between the process characterized in equations (1) and the process characterized in equation (7). Therefore, we may expect the value of $S_{d(z)}{}^2$ to lie between $S_d{}^2/3$ and $S_d{}^2$ for most population two scenarios.

Determination of Stratum Boundaries

At the design stage, we only have knowledge about the invoice amount x. In order to determine stratum boundaries and sample size, we use equation (6) as the approximation of $S_d{}^2$. Given the number of strata L, stratum boundaries under Neyman optimum allocation can be determined such that $N_h S_{hd}$ ($h = 1, 2, \cdots, L$) is about the same for all strata (see Cochran, 1977). That is,

$$N_h \sqrt{p_h(1-p_h)\left(S_{hX}{}^2 + \bar{X}_h{}^2\right)} = C,$$

$$h = 1, 2, \cdots, L \quad (11)$$

where C is some constant.

If we are comfortable with the assumption that all the qualified invoices are evenly distributed in the population, p_h is about the same across all the strata. We can therefore use the known $\left(S_{hX}{}^2 + \bar{X}_h{}^2\right)$. Equation (11) is reduced to:

$$N_h \sqrt{S_{hX}{}^2 + \bar{X}_h{}^2} = C, \quad h = 1, 2, \cdots, L \quad (12)$$

Rewrite equation (12) as:

$$X_h \sqrt{CV_{hx}{}^2 + 1} = C, \quad h = 1, 2, \cdots, L \quad (13)$$

where CV_{hx} is the coefficient of variation of x for stratum h.

Equation (13) leads to an important application of setting up stratum boundaries. First, it should be easy to set up stratum boundaries under Neyman allocation using equation (13). Further note that $CV_{hx}{}^2$ is much smaller than 1 in many accounting applications. Therefore, equation (13) can be approximated by $X_h = C$, $h = 1, 2, \cdots, L$. In other words, the equal invoice amount per stratum gives us the approximate stratum boundaries under Neyman allocation. To be more accurate, we may first set up stratum boundaries based on the equal invoice amount; and then adjust the boundaries based on the coefficient of variation per stratum.

The above guidelines of optimum stratum boundaries also apply to population two, which is supported by equation (10). Note that there are many scenarios for population two and equation (10) is one of them. The stratum boundaries under Neyman allocation may vary for different scenarios, but the equal invoice amount criterion can provide a useful approximation for other scenarios as long as the assumption holds. That is, the qualified invoices are randomly scattered in the population.

If the qualified percentage tends to increase or decrease as the invoice amount increases or decreases, we may incorporate information about different qualified percentages in different strata into equation (13). That is, we can set up stratum boundaries using:

$$X_h \sqrt{CV_{hx}{}^2 + 1} \sqrt{p_h(1-p_h)} = C, \quad h = 1, 2, \cdots, L \quad (14)$$

Sample Size Determination and Allocation

The above stratum boundary criterion yields equal stratum sample sizes for all strata. The sample size formula for population one is:

$$n_1 = \frac{t^2 L \sum N_h{}^2 S_{hd}{}^2}{A^2 + t^2 \sum N_h S_{hd}{}^2} \quad (15)$$

where t is the t-value corresponding to the confidence level and A is the desired absolute precision.

For the model of equation (7), the sample size is:

$$n_2 = \frac{t^2 L \sum N_h{}^2 S_{hd(z)}{}^2}{B^2 + t^2 \sum N_h S_{hd(z)}{}^2} \qquad (16)$$

where B is the desired absolute precision.

Since $S_{hd(z)}{}^2 \approx \frac{S_{hd}{}^2}{3}$ by equation (10), we have:

$$n_2 = \frac{t^2 L \sum N_h{}^2 S_{hd}{}^2}{3B^2 + t^2 \sum N_h S_{hd}{}^2} \qquad (17)$$

Compare equations (15) and (17), the same sample size leads to $B = 0.58A$. In other words, the same sample size can give a better precision for population two than for population one. For the assumed qualified percent p, the sample size to calculated to achieve a certain precision under population one is a conservative estimate of the sample size needed to achieve the same precision for some unknown scenario of population two. We should caution that it maybe too conservative sometimes. As in the above analysis, the sample size calculated under population one can give a 42% shorter margin of error for the scenario described in equation (7).

Simulation

The simulation population includes 3,231 invoices after removing the largest invoices with certainty. Figure 3 gives the histogram based on invoice amount -- the design variable x.

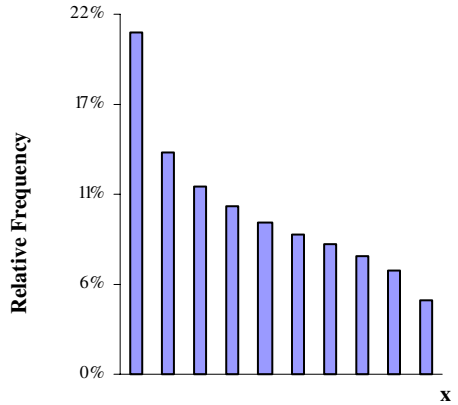

Figure 3. Histogram of the Simulation Population

The population is divided into five strata with equal stratum amounts on x. The population summary is presented in Table 1.

Table 1. Simulation Population Summary by Stratum

	Range of x				
h	**Min**	**Max**	N_h	X_h	CV_{hx}
1	10,260	46,920	1,112	37,446,730	29%
2	46,960	59,771	702	37,484,099	7%
3	59,857	70,040	576	37,448,657	5%
4	70,078	84,950	491	37,500,290	6%
5	84,951	193,405	350	37,525,292	23%

Variable y is created based on the equation (1) to represent population one and variable z is created based on equation (7) to represent one of the scenarios in population two. $p = 0.2$ is used in creating variables y and z.

The Neyman allocations across strata based on different variables are given in Table 2.

Table 2. Neyman Allocation Comparison

	Based on Simulated Variables		Based on the Known Variable x	
h	(a) d_y	(b) d_z	(c) Equation (6)	(d) x
1	19.6%	20.4%	20.5%	41.8%
2	20.5%	19.0%	19.8%	10.0%
3	20.1%	19.4%	19.7%	6.5%
4	19.6%	20.6%	19.8%	8.1%
5	20.1%	20.7%	20.3%	33.6%

The sample size allocation across strata would be best determined by the variable of interest, y or z. In ratio type estimation, the Neyman allocation percentages are calculated for variable y by $N_h S_{hd} \Big/ \sum_h N_h S_{hd}$, where $S_{hd}{}^2$ is the variance of $d = y - Rx$ and $R = \sum_{i=1}^{N} y_i \Big/ \sum_{i=1}^{N} x_i$. The results are given in column (a). Column (b) gives the Neyman allocation percentages based on variable z. These percentages are calculated using $N_h S_{hd(z)} \Big/ \sum_h N_h S_{hd(z)}$, where $S_{hd(z)}{}^2$ is the variance of $d(z) = z - Rx$ and $R = \sum_{i=1}^{N} z_i \Big/ \sum_{i=1}^{N} x_i$. The

numbers in columns (a) and (b) are very close stratum by stratum. This indicates that stratum boundaries under Neyman allocation are about the same whether they are determined by population one type of data (y) or by the type of data of population two (z). In practice, the values of variable y are unknown at the design stage. Fortunately, S_{hd} is well approximated by $\sqrt{p(1-p)(S_{hx}^2 + \overline{X}_h^2)}$. Therefore, Neyman allocation percentages can be actually calculated by

$$Neyman\ Allocation = \frac{N_h\sqrt{S_{hx}^2 + \overline{X}_h^2}}{\sum_h N_h\sqrt{S_{hx}^2 + \overline{X}_h^2}} \qquad (18)$$

The above formula (18) involves only the known values of variable x. The results are shown in column (3). Comparing the numbers in column (3) to those in column (1), there are only minor differences. Therefore, we can make Neyman allocation at the design stage without knowing the variables of interest. As a comparison, the Neyman allocation percentages using $N_h S_{hx} \Big/ \sum_h N_h S_{hx}$ are also presented in column (4). The numbers in column (4) are quite different from those in the other three columns. This indicates that the Neyman allocation based on the variance of the design variable x alone is very inefficient. It under-allocates for certain strata and over-allocates for other strata by a large degree. In summary, Neyman allocations can be calculated using equation (18) for both population one and population two.

The allocation percentages across strata in column (1) are very close, which indicates an equal sample size across strata is appropriate. This confirms our earlier finding that stratum boundaries under Neyman allocation are approximated by setting an equal invoice amount per stratum.

The above simulation is based on $p = 0.2$. Other simulations using $p = 0.5$ and $p = 0.8$ lead to the same conclusion.

Using formula (15), the sample sizes in order to reach a relative precision of 10% at 90% confidence level are given in Table 3.

Table 3. Sample Size Comparison

Assumed p	Using Simulated y	Using Simulated z	Using Roberts Formula
0.2	788	329	797
0.5	247	88	253
0.8	70	21	68

The sample sizes using Roberts' formula are obtained by substituting equation (6) into sample size formula (15). As shown in Table 3, Roberts gives sample sizes very close to those obtained using simulated variable y. The simulated variable z achieves the same relative precision with smaller sample sizes. For many situations in practice, the variable of interest is between y and z. Therefore, Roberts gives somewhat conservative sample sizes for these situations. As the values of p increase, the sample sizes decrease. However, even though the overall sample size needed to achieve desired precision levels may be very small, the stratum sample size should not be allowed to become too small in order to reduce bias and stabilize the variance estimation.

Conclusion

For our special ratio type data, assuming the qualified amounts are randomly spread throughout the population, the stratum boundaries under Neyman allocation can be obtained approximately by setting up equal stratum amounts on the design variable x. The stratum boundaries can be modified by considering the coefficient of variation of x per stratum using equation (13). More modification can be made using equation (14) if there is prior knowledge about different values of p for different strata.

The sample size calculated from the Roberts formula tends to be conservative in practice for many scenarios of population two.

Future Work

We plan to analyze the effectiveness of different numbers of strata and the stratum sample size. For example, for a fixed sample size of 100 units, we may compare the setting of 4 strata with 25 units per stratum and the setting of 2 strata with 50 units per stratum.

We also plan to explore the relationship between the value of p and the gains achieved using ratio estimation.

Reference

Cochran W.G. (1977). *Sampling Technique*, 3rd ed. New York: Wiley.

Liu, Y, Batcher, M & Rotz, W (2001), Application of the Hypergeometric Distribution
In a Special Case of Rare Events, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

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