Journal of Data Science **3**(2005), 213-222

Efficient Sampling Design in Audit Data

Yan Liu[1], Mary Batcher[1] and Fritz Scheuren[2]
[1]*Ernst & Young and* [2]*University of Chicago*

Abstract: Auditors are often faced with reviewing a sample drawn from special populations. 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. For these settings, the typical sample design is stratified random, with the estimation method employing a ratio type method. This paper focuses on efficient sample design for this setting and provides some guidelines in setting up stratum boundaries, calculating sample size and allocating sample size optimally across strata.

Key words: Audit sampling, Neyman allocation, ratio estimation, stratified sampling.

1. Introduction

Much of traditional sampling theory was developed in the household survey context. Sampling business records presents very different challenges and often requires different solutions. Most commonly, the quantity to be estimated is financial. It may be, for example, the amount subject to sales tax, the amount deductible from income tax, or the amount that is in error in the business records. The sampling unit is frequently invoices. 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. Whether we are reviewing for the traditional audit purpose of identifying and quantifying errors in business records or for determining taxable amounts, we can generally classify our sampling as audit sampling, where we are beginning with a recorded amount and making some quantitative determination about that original amount.

There are 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. In other words, the qualified amount is either zero or the same as the known invoice amount, depending on which category the invoice falls into. This type of populations is called *Population One*. Figure 1 (left half) shows the scatterplot of the qualified amount against the invoice amount for population one. The other population type arises when some invoices have a qualified amount between zero and the full invoice amount. This type of population is called *Population Two*. Figure 1 (right half) shows the scatterplot of the qualified amount against the invoice amount for population two.

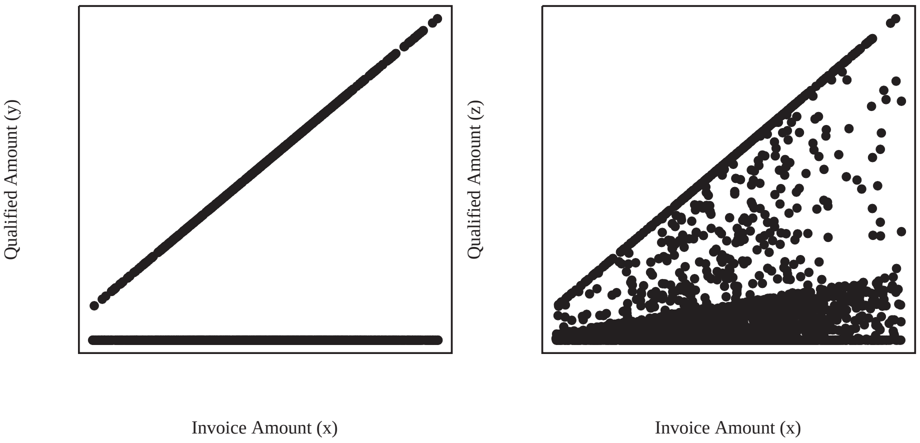

Figure 1: Population one (left part) and population two (right part)

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 the cases with the largest recorded amounts (or potential 'outliers') are taken with certainty.

We first summarize the characteristics of population one. Suppose that the population includes 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). Here, however, we will assume a binomial model applies.

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 and 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:

$$\begin{aligned} y_i &= x_i, && \text{with probability} p \\ &= 0, && \text{with probability} (1-p) \end{aligned} \tag{1.1}$$

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 (1978). The population parameter to be estimated is the ratio:

$$R = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} \tag{1.2}$$

The corresponding sample estimate under simple random sample is:

$$\hat{R} = \frac{\bar{y}}{\bar{x}} \tag{1.3}$$

where $\bar{y} = n^{-1} \sum_i^n y_i$ and $\bar{x} = n^{-1} \sum_i^n x_i$. The variance of $\hat{R}$, for large n, is approximately:

$$V(\hat{R}) = \frac{1-f}{n\bar{X}^2} S_d^2 \tag{1.4}$$

where S_d^2 is the variance of $d_i = y_i - Rx_i$ and

$$S_d^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - Rx_i)^2 = \frac{1}{N-1} \sum_{i=1}^{N} d_i^2 \tag{1.5}$$

Under the realization process of population units described in equation (1.1),

$$E_p R = p \tag{1.6}$$

and

$$E_p(S_d^2) \approx p(1-p)(S_x^2 + \bar{X}^2) \tag{1.7}$$

when the population size, N, is also reasonably large.

We now expand the above properties to population two where some invoices are partially qualified. In order to relate population two to population one and make use of the results from population one, we assume the same average ratio for population two, i.e., $E_p(R) = p$. There should be 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 are randomly scattered around the line . So the population units can be characterized as a realization of the following process:

$$\begin{aligned} z_i &= px_i + u(1-p)x_i, && \text{with probability } p \\ &= px_i - upx_i, && \text{with probability } (1-p), \end{aligned} \tag{1.8}$$

$i = 1, 2, \ldots, N$, where u is a random number from *Uniform*(0, 1). Under the realization process of population units described in equation (1.8), we still have $E_p(R) = p$ for the ratio $R = \sum_{i=1}^{N} z_i / \sum_{i=1}^{N} x_i$. Corresponding to formula (1.4), the approximate variance of $\hat{R} = \bar{z}/\bar{x}$ is $V(R) = \frac{1-f}{n\bar{X}^2} S^2_{d(z)}$. Now, $S^2_{d(z)}$ is the variance of $d_i(z) = z_i - Rx_i$. Rewrite $d_i(z)$ as

$$\begin{aligned} d_i(z) = z_i - Rx_i \approx z_i - px_i &= u(1-p)x_i, && \text{with probability } p \\ &= -upx_i, && \text{with probability } (1-p), \end{aligned} \tag{1.9}$$

$i = 1, 2, \ldots, N$.

Rewrite d_i in population one as:

$$\begin{aligned} d_i = y_i - Rx_i \approx y_i - px_i &= (1-p)x_i, && \text{with probability } p \\ &= -px_i, && \text{with probability } y(1-p), \end{aligned} \tag{1.10}$$

$i = 1, 2, \ldots, N$. Comparing equations (1.9) and (1.10), we have $d_i = ud_i$.

Now from equation (1.5), $S_d^2 = (N-1)^{-1} \sum_{i=1}^{L} d_i^2$. Therefore, $S^2_{d(u)} = (N-1)^{-1} \sum_{i=1}^{N} (ud_i)^2 = u^2 S_d^2$. $E_p(S^2_{d(z)}) = E_p(S_d^2) = E_p(u^2)E_p(S_d^2)$, since u and d are independent. $E_p(u^2) = 1/3$, since $u \sum$ *Uniform*(0, 1). Therefore,

$$E_p(S^2_{d(z)}) = \frac{E_p(S_d^2)}{3} \tag{1.11}$$

Note that most scenarios of population two fall between the process characterized in equations (1.1) and the process characterized in equation (1.8). Therefore, we may expect the value of $S^2_{d(z)}$ to lie between $S_d^2/3$ and S_d^2 for most scenarios of population two.

2. Determination of Stratum Boundaries

At the design stage, we only have knowledge about the invoice amount. In practice, the Dalenius-Hodges method (Cochran 1977, pp. 127-131 and Särndla, *et al.* 1991, pp. 463-464) is often used to set up stratum boundaries based on the values of x. Then sample size is allocated by the Neyman rule (Cochran 1977, Chapter 5), based on knowledge of x. This works well only if the correlation between x and y is strong, say a correlation coefficient of 0.9 or more. This is often not the case in practice. Therefore, for our special ratio type data, we develop a new method to determine stratum boundaries and sample size allocation using the special relationship between x and y. Specifically, we use equation (1.7) as the approximation of S_d^2.

Given the number of strata and the same sample size per stratum, stratum boundaries under Neyman optimum allocation can be determined such that $N_h S_{hd}$ $(h = 1, 2 \ldots, L))$ is about the same for all strata. That is,

$$N_h \sqrt{p_h(1 - p_h)(S_{hx}^2 + \bar{X}_h^2)} = C, \quad h = 1, 2, \ldots, L \tag{2.1}$$

where C is a 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 $(S_{hx}^2 + \bar{X}_h^2)$. Equation (2.1) is reduced to:

$$N_h \sqrt{S_{hx}^2 + \bar{X}_h^2} = C, \quad h = 1, 2, \ldots, L \tag{2.2}$$

Now we can rewrite equation (2.2) as:

$$X_h \sqrt{CV_{hx}^2 + 1} = C, \quad h = 1, 2, \ldots, L \tag{2.3}$$

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

Equation (2.3) 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 (2.3). Further note that CV_{hx}^2 is much smaller than 1 in many accounting applications. Therefore, equation (2.3) can be approximated by $X_h = C, h = 1, 2, \ldots, L$ if the distribution of invoice amount x is not highly skewed. In other words, since X_h is the total value of the invoices in stratum h, then what is being said is that setting equal the total invoice amount per stratum gives us the approximate stratum boundaries for the same sample size per stratum 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 described by equation (1.8), which is supported by equation (1.11). Note that there are many scenarios for population two and equation (1.8) is one of them. The stratum boundaries for the same sample size per stratum 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 that the CV's are small 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 (2.3). That is, we can set up stratum boundaries using:

$$X_h\sqrt{CV_{hx}^2+1}\sqrt{p_h(1-p_h)}=C, \qquad h=1,2,\ldots,L \tag{2.4}$$

3. 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:

$$n_1=\frac{t^2L\sum N_h^2S_{hd}^2}{A^2+t^2\sum N_h s_{hd}^2} \tag{3.1}$$

where t is the t-value corresponding to the confidence level and A is the desired absolute precision or margin of error. For population two, that is described by the model of equation (1.8), the sample size is:

$$n_2=\frac{t^2L\sum N_h^2S_{hd(z)}^2}{B^2+t^2\sum N_hS_{hd(z)}^2} \tag{3.2}$$

where B is the desired absolute precision. Since $S_{hd(z)}^2\approx S_{hd}^2/3$ by equation (1.11), we have:

$$n_2=\frac{t^2L\sum n_h^2S_{hd}^2}{3B^2+t^2\sum N_hS_{hd}^2} \tag{3.3}$$

Compare equations (3.1) and (3.3), 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 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 (1.8).

4. Simulation

The simulation population includes 3,231 invoices after removing the largest invoices with certainty. Figure 2 gives the histogram based on invoice amount — the design variable x.

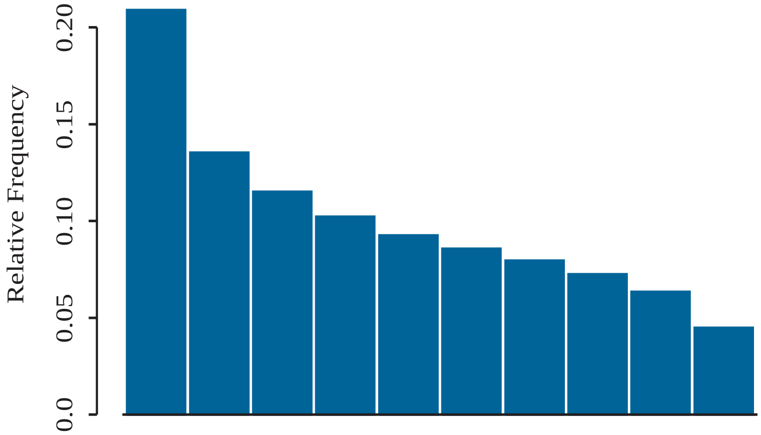

Figure 2: Histogram of the simulated population

The population is divided into five strata with equal stratum total dollar 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 (1.1) to represent population one and variable z is created based on equation (1.8) 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.

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 $N_h S_{hd} / \sum_h N_h S_{hd}$, where S_{hd}^2 is the variance of $d = y - Rx$ and $R = \sum_{i=1}^{N} y_i / \sum i = 1^N x_i$. The results are given in column (a). Column (b) gives the Neyman allocation percentages based on variable z. These percentages are calculated using $N_h S_{hd(z)} / \sum_h N_h S_{hd(z)}$, where $S_{hd(z)}^2$ is the variance of $d(z) = z - Rx$ and $R = \sum_{i=1}^{N} z_i / \sum_{i=1}^{N} x_i$. The numbers

Table 2: Neyman Allocation Comparison

	Based on Simulated Variables of Interest		Based on the Known Covariate Variable x	
	(a)	(b)	(c)	(d)
h	d_y	d_x	Equation (1.7)	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%

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 would have been 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 $\sqrt{p(1-p)(S_{hx}^2+\bar{X}_h^2)}$. Therefore, Neyman allocation percentages can be actually calculated by

$$\text{Neyman Allocation} = \frac{N_h\sqrt{S_{hx}^2+\bar{X}_h^2}}{\sum_h N_h\sqrt{S_{hx}^2+\bar{X}_h^2}} \tag{4.1}$$

The above formula (4.1) involves only the known values of variable x. The results are shown in column (c) of Table 2. Comparing the numbers in column (c) to those in column (a), there are only minor differences. Therefore, we can achieve Neyman allocation regarding to the variable of interest (y or z) at the design stage without knowing the variable of interest. As a comparison, the Neyman allocation percentages regarding to the design variable x using $N_hS_{hx}/\sum_h N_hS_{hx}$ are also presented in column (d). The numbers in column (d) 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 s trata by a large degree. In summary, Neyman allocations can be calculated using equation (4.1) for both population one and population two.

The allocation percentages across strata in column (a) are very close, which indicates an equal sample size across strata is appropriate. This confirms our earlier finding that stratum boundaries by equation (2.4) are well approximated by setting an equal invoice amount per stratum if the distribution of x is not highly skewed and qualifying percentage p_h is about the same across all the strata.

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 (3.1), 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 Simu. y	Using Simu. z	Using Roberts' Formula
0.2	788	329	797
0.5	247	88	253
0.8	70	21	68

The sample sizes using Roberts (1978)'s formula are obtained by substituting equation (1.7) into sample size formula (3.1). As shown in Table 3, Roberts (1978) gives sample sizes very close to those obtained using the 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 (1978) 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.

5. Conclusion

For our special ratio type data, assuming the qualified amounts are randomly spread throughout the population, the stratum boundaries with equal stratum sample size under Neyman allocation can be obtained approximately by setting up equal total stratum amounts on the design variable x. The stratum boundaries can, then, be modified by considering the coefficient of variation of x per stratum, using equation (2.3). Even more modification can be made using equation (2.4) if there is prior knowledge about different values of p for different strata. The sample size calculated from the Roberts (1978) formula tends to be conservative in practice for many scenarios of population two.

6. 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.

References

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

Liu, Y, Batcher, M and 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.* American Institute of Certified Public Accountants, Inc.

Särndal, C. E., Swensson B. and Wretman J. (1991). *Model-assisted Survey Sampling.* Springer Verlag.

Received April 3, 2004; accepted July 5, 2004.

Yan Liu
Ernst & Young LLP
1225 Connecticut Ave., NW
Washington, DC 20036, USA
Yan.Liu@ey.com

Mary Batcher
Ernst & Young LLP
1225 Connecticut Ave., NW
Washington, DC 20036, USA
Mary.Batcher@ey.com

Fritz Scheuren
NORC, University of Chicago
1402 Ruffner Road
Alexandria, VA 22302, USA
Scheuren@aol.com