RANDOM SUM ESTIMATORS

AND THEIR EFFICIENCY

Yurii B. Shvetsov
Department of Mathematical Sciences
Montana State University

January 15, 2004

A writing project submitted in partial fulfillment
of the requirements for the degree

of

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Yurii B. Shvetsov

This writing project has been read by the writing project director and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and is ready for submission to the Statistics Faculty.

1/15/2004

Date

John J Borkowski

John J. Borkowski
Writing Project Director

Abstract

In this paper, we review the random sum estimators of the mean. First, we describe certain consulting problems that motivate this study. Then, we compute the expectation and variance of a random sum estimator, and investigate the conditions under which this estimator is more efficient than the simple random sample estimator. Finally, a simulation is conducted to analyze the confidence intervals based on the random sum estimation and to test the results of the relative efficiency study.

1 Introduction

Random sums of random variables have been studied in the theory of stochastic processes and stochastic modelling for quite some time. Due to their usefulness in this area of statistics, numerous results on random sums have been obtained through the last 30 years. Such results range from purely theoretical, that concern a large class of random sums ([4], [5], [6]), to theorems on convergence or distribution of random sums of identically distributed (independent or not) random variables that follow a certain distribution ([8], [9], [12]).

Random sums have also made their way into classic texts. Feller [3] invokes them as examples on several occasions and states the Central Limit Theorem for random sums. Mood, Graybill, and Boes [7] offer a brief discussion of random sums, with sketchy derivations of their expectation and variance. Finally, Taylor and Karlin [10] present a rather extensive treatment of random sums, including a detailed discussion of moments and distribution of random sums, and several motivating examples. In particular, this texts illustrates the differences in applying the Central Limit Theorem, depending on whether the phenomenon under consideration is modeled by a random sum or not.

Little attention, however, has been paid to the theory of point estimation for random sums in the survey sampling literature. This paper addresses this situation.

In simple terms, a random sum is a sum of a random number of random variables. The number of the terms N in the sum, as well as the individual terms, can have various distributions. In stochastic theory, N is often assumed to follow Poisson or geometric distribution. In the discussion that follows, N has a hypergeometric distribution, and the terms in the sum are identically distributed random variables, not necessarily independent. We show that random sum approach can be useful in populations that have a mixture distribution with a significant fraction of zero values and a nonzero component that follows a continuous distribution.

The idea of utilizing random sums in point estimation arises from a number of statistical consulting problems that exhibit mixture distributions. We shall now describe two such problems from the practice of Prof. Borkowski [1].

1. **Medicaid overpayment study** . Medicaid providers receive payments based on claims submitted for reimbursement. When a provider submits a claim and receives payment for items not supported by Medicaid, then this provider

receives an overpayment. Audits are conducted by Medicaid to estimate the total amount of overpayment for every provider. Because the number of claims submitted by one provider is large, it is unrealistic to audit all claims. Thus, a random sample is taken from the population of claims, and the total provider overpayment is estimated based on the overpayment in the sample.

It is not unusual for a provider to submit about 50,000 claims in a 12- to 16-month period. From the population of claims from a particular provider, a sample of at least 200 claims is taken. Confidence interval estimates are based on data from the sample, and are used to determine how much health care providers must reimburse Medicaid for the overpayment they received.

Because not all claims contain an overpayment, there is a large fraction of zeroes (no overpayment) in the sample. So the distribution of overpayments is a mixture of a zero part and a positive overpayment part that we assume to be continuously distributed.

2. **Assessment of highway maintenance activities**. The study was conducted to assess the quality of highways in the state of Montana. A sample of 1/10 mile highway segments was taken, and on every segment, 23 characteristics were measured for deficiencies, such as the proportion of striping that does not meet standards and the number of feet of fencing not up to standards.

   For every characteristic measured, if the entire 1/10 mile segment meets standards, a value of 0 is assigned. Therefore, a large fraction of zero values is present in the sample, and each characteristic follows the aforementioned mixture distribution.

We mention yet another practical problem with a mixture distribution of the same type. The author had a conversation with a physicist who conducted the following experiment. A thin sheet of material is bombarded with alpha particles. On the opposite side of the sheet, a detector records the energy level of every particle that it captures. Clearly, not all alpha particles make it through the sheet, because some are reflected. In that case, a value of zero is assigned. Thus a significant number of zeroes in the recorded data.

This paper is focused on the analysis of random sum estimators of the mean in finite populations that exhibit a mixture distribution with a significant fraction of zeroes. After stating the necessary definitions and assumptions in Section 2, we introduce the random sum estimator in Section 3 and discuss its most basic properties. Then, in Section 4, we obtain the formulas for variance estimates used in the construction of confidence intervals. Based on these formulas, we perform a relative efficiency study and a simulation to test the findings of the study. This is described in Section 5.

In the subsequent discussion, we shall adopt the terminology from the Medicaid study.

2 Preliminaries

Consider a finite population of units ("claims"), and let N denote the size of the population. We are interested in measuring the characteristic Y ("overpayment") in every unit. The value of Y in the i-th unit will be denoted by y_i. Throughout the paper, we assume that the population has a significant portion of units with $y_i = 0$ ("no-overpayment part"), and that for the rest of the population, $y_i > 0$ ("overpayment part"). For the Medicaid study, the zero portion was typically near 40%. Furthermore, we assume that Y is continuously distributed on the overpayment part of the population.

Let N_1 denote the number of units in the overpayment part of the population. To simplify the indexing, without loss of generality we assume that the first N_1 units belong to the overpayment part. Thus we have

$$\begin{aligned} y_i >& \ 0 \quad \text{for} \quad i &= 1, \ldots, N_1 \\ y_i =& \ 0 \quad \text{for} \quad i &= N_1 + 1, \ldots, N \end{aligned} \tag{1}$$

Define q to be the fraction of the population with overpayments:

$$q = \frac{N_1}{N} \,.$$

Further, μ will denote the mean of Y over the entire population, and μ_1 the mean of Y across the overpayment part:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} y_i \,, \qquad \mu_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} y_i \,. \tag{2}$$

Similarly, σ^2 denotes the variance of Y over the entire population, and σ_1^2 the variance of Y across the overpayment part. In addition, τ denotes the total overpayment in the population:

$$\tau = \sum_{i=1}^{N} y_i \,.$$

We now state some elementary but useful identities.

Lemma 1. *With the assumptions stated above, the following are true:*

(a) $\mu = q\mu_1$;

(b) $\tau = N\mu = N_1\mu_1 = qN\mu_1$.

In addition, by performing simple algebraic transformations one can easily show the following.

Lemma 2. $\sigma^2 = q\sigma_1^2 + q(1-q)\mu_1^2$.

In the next section we shall use the definition of a random sum of random variables, which we now state [10].

Definition 3. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent and identically distributed random variables, and let K be a discrete random variable, independent of $\xi_1, \xi_2, \ldots$ and having the probability mass function $p_K(n) = \Pr\{K = n\}$ for $n = 0, 1, \ldots$. Define the random sum X by

$$X = \begin{cases} 0 & : \text{ if } K = 0, \\ \xi_1 + \cdots + \xi_K & : \text{ if } K > 0. \end{cases} \tag{3}$$

We save space by abbreviating (3) to simply write $X = \xi_1 + \cdots + \xi_K$, understanding that $X = 0$ whenever $K = 0$.

Using the notation given in this definition, we state an important proposition [10].

Proposition 4. *Suppose that ξ_k and K have finite moments:*

$$\begin{aligned} E[\xi_k] &= \mu, & \operatorname{Var}[\xi_k] &= \sigma^2, \\ E[K] &= \nu, & \operatorname{Var}[K] &= \tau^2, \end{aligned}$$

then the moments of X are given by

$$E[X] = \mu\nu, \qquad \operatorname{Var}[X] = \nu\sigma^2 + \mu^2\tau^2\,. \tag{4}$$

We note that Proposition 4 is valid if the terms of the random sum are independent and identically distributed. However in our case, as we shall see in the next section, the independence assumption is not satisfied, and the formula for the variance of a random sum will have to be adjusted.

3 Random sum estimators

Suppose that a random sample of size n is taken from the population of size N that has N_1 units with $y_i > 0$ (overpayment) and $N - N_1$ units with $y_i = 0$ (no overpayment). Denote by n_1 the number of units in the sample with overpayment. Without loss of generality, we assume that the units in the sample are ordered so that the first n_1 units form the overpayment part of the sample.

The goal of taking the sample is to estimate the mean overpayment in the population, i.e. μ. The sample mean is typically used for that purpose:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,.$$

$\bar{y}$ is an unbiased estimator of μ, and due to the Central Limit Theorem, it is approximately normal for large sample size n. However, for smaller n, $\bar{y}$ is very poorly approximated by the normal distribution, because the population has a significant fraction of zeroes. To account for this mixture of zero and nonzero parts, we suggest a random sum estimator of the following form:

$$\hat{\mu}_R = \frac{1}{n}\sum_{i=1}^{n_1} y_i\,, \tag{5}$$

where n_1 has a hypergeometric distribution:

$$n_1 \sim \text{Hyp}\ (N, N_1, n)\ . \tag{6}$$

Note that $\sum_{i=1}^{n_1} y_i = \sum_{i=1}^{n} y_i$ because $\sum_{i=n_1+1}^{n} y_i = 0$, and hence, $\hat{\mu}_R = \bar{y}$. Therefore, the two estimators, $\bar{y}$ and $\hat{\mu}_R$, always yield the same estimated value of μ. In other words, it is the same quantity, which is looked at from two different viewpoints. However, their estimated variances, and consequently the confidence intervals based on these two estimators, are different.

We shall now show that $\hat{\mu}_R$ is an unbiased estimator of μ. Note first that since in $\hat{\mu}_R$, y_i always belongs to the overpayment part, then

$$E[y_i] = \mu_1\ , \quad i = 1, 2, \ldots, n_1\ .$$

Now, applying Proposition 4 and Lemma 1, we obtain

$$E[\hat{\mu}_R] = \frac{1}{n}\, E\Big[\sum_{i=1}^{n_1} y_i\Big] = \frac{1}{n}\, E[n_1]\, E[y_i] = \frac{1}{n}\, \frac{nN_1}{N}\, \mu_1 = q\mu_1 = \mu\ .$$

Thus $\hat{\mu}_R$ is unbiased.

We conclude this section by stating the Central Limit Theorem for random sums, as it appears in Feller [3].

Theorem 5 (Central Limit Theorem for Random Sums). *Let $X_K = \xi_1 + \cdots + \xi_K$, where ξ_i and K are mutually independent random variables. Suppose that ξ_i have a common distribution F with zero expectation and variance* 1. *Further, let $N_1, N_2, \ldots$ be positive integer-valued random variables such that*

$$n^{-1}\, N_n \xrightarrow{p} 1\ . \tag{7}$$

Then the distribution of $X_{N_n}/\sqrt{n}$ tends to $\mathfrak{N}$.

As noted in [3], $X_{N_n}/\sqrt{n}$ is not normalized to unit variance. In addition, Theorem 5 applies to cases with $E[N_n] = \infty$ and even when expectations exist, (7) does not imply that $n^{-1}\, E[N_n] \longrightarrow 1$.

4 Variance formulas

Recall that the variance of the sample mean in finite populations is given by

$$\text{Var}[\bar{y}] = \frac{N-n}{Nn}\, \sigma^2\ , \tag{8}$$

where $\frac{N-n}{N}$ is the finite population correction. Thus the estimated variance of the sample mean is

$$\widehat{\text{Var}}[\bar{y}] = \frac{N-n}{Nn}\, s^2\ . \tag{9}$$

We shall now obtain a formula for the variance of the random sum estimator $\hat{\mu}_R$.

Proposition 6. *The variance of the random sum estimator $\hat{\mu}_R$ of the population mean μ, is given by the following:*

$$\mathrm{Var}[\hat{\mu}_R] = \frac{q}{n}\,\sigma_1^2\left(1-\frac{n-1}{N-1}\right) \;+\; \frac{\mu_1^2}{n}\,q(1-q)\,\frac{N-n}{N-1}\,. \tag{10}$$

PROOF. Let $M = \min\{n, N_1\}$ and denote $h(k)$ the hypergeometric probability mass function: $h(k) = \mathrm{P}(n_1 = k)$. By formulas of conditional probability:

$$\begin{aligned}
\mathrm{Var}[\hat{\mu}_R] &= E\left[\Big(\frac{1}{n}\sum_{i=1}^{n_1} y_i\Big)^2\right] - \mu^2 = \sum_{k=0}^{M} E\left[\Big(\frac{1}{n}\sum_{i=1}^{n_1} y_i\Big)^2 \Big| n_1 = k\right]\cdot \mathrm{P}(n_1 = k) \;-\; \mu^2 \\
&= \frac{1}{n^2}\sum_{k=0}^{M} E\left[\sum_{i=1}^{k} y_i^2 \;+ \sum_{i\neq j} y_i y_j\right] h(k) \;-\; \mu^2 \\
&= \frac{1}{n^2}\sum_{k=0}^{M} \Big(k\,E[y_i^2] \;+ k(k-1)\,E[y_i y_j]\Big)\,h(k) \;-\; \mu^2 \\
&= \frac{1}{n^2}\sum_{k=0}^{M} \Big(k(\sigma_1^2+\mu_1^2) \;+ k(k-1)\big(\mathrm{Cov}(y_i,y_j)+\mu_1^2\big)\Big)\,h(k) \;-\; \mu^2 \\
&= \frac{1}{n^2}\,(\sigma_1^2+\mu_1^2)\sum_{k=0}^{M} k h(k) \;+ \frac{1}{n^2}\,\mu_1^2 \sum_{k=0}^{M} k(k-1)h(k) \;-\; \mu^2 \\
&\quad + \frac{1}{n^2}\sum_{k=0}^{M} k(k-1)\,\mathrm{Cov}(y_i,y_j)\,h(k)\,,
\end{aligned} \tag{11}$$

because $E[y_i^2] = \sigma_1^2 + \mu_1^2$ and $E[y_i y_j] = \mathrm{Cov}\,(y_i, y_j) \;+ \mu_1^2$. Now the first three terms in (11) correspond to the variance of $\hat{\mu}_R$ for the case when the summands in the random sum are independent, i.e. $\mathrm{Cov}[y_i, y_j] = 0$. So by Proposition 4, their sum is equal to

$$\sigma_1^2\,E[n_1] + \mu_1^2\,\mathrm{Var}[n_1] = \frac{q}{n}\,\sigma_1^2 \;+\; \frac{\mu_1^2}{n}\,q(1-q)\,\frac{N-n}{N-1}\,. \tag{12}$$

However, because a sample without replacement is taken from a finite population, the items in the sample are not independent, and

$$\mathrm{Cov}[y_i, y_j] = -\frac{\sigma_1^2}{N_1 - 1}\,,$$

if both y_i and y_j belong to the overpayment part of the sample. Thus, the last term in (11) is non-vanishing and is equal to

$$\begin{aligned}
&-\frac{\sigma_1^2}{(N_1-1)n^2}\sum_{k=0}^{M} k(k-1)h(k) \\
&= -\frac{\sigma_1^2}{(N_1-1)n^2}\Big(\mathrm{Var}[n_1] + \left(E[n_1]^2\right) - E[n_1]\Big) \\
&= -\frac{\sigma_1^2}{(N_1-1)n^2}\left(\frac{nN_1}{N}\,\frac{(N-N_1)(N-n)}{N(N-1)} + \frac{n^2N_1^2}{N^2} - \frac{nN_1}{N}\right) \\
&= -\frac{q}{n}\,\sigma_1^2\,\frac{n-1}{N-1}\,.
\end{aligned} \tag{13}$$

Now equation (10) follows from (11), (12) and (13). □

Analogous to (9), the estimated variance of the random sum estimator can be computed as follows:

$$\widehat{\mathrm{Var}}[\hat{\mu}_R] = \frac{\hat{q}}{n}\,s_1^2\left(1-\frac{n-1}{N-1}\right) + \frac{\bar{y}_1^2}{n}\,\hat{q}(1-\hat{q})\,\frac{N-n}{N-1}\,, \tag{14}$$

where $\hat{q} = n_1/n$.

By the Central Limit Theorem for finite populations, the sample mean is approximately normal for sufficiently large n, so the confidence interval for μ with the significance level α, based on the sample mean, is

$$\left[\bar{y} - t_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}[\bar{y}]}\,, \bar{y} - t_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}[\bar{y}]}\,\right],$$

where $\widehat{\mathrm{Var}}[\bar{y}]$ is given by (8). For the same reason, the confidence interval based on $\hat{\mu}_R$ is

$$\left[\hat{\mu}_R - t_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}[\hat{\mu}_R]}\,, \hat{\mu}_R - t_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}[\hat{\mu}_R]}\,\right],$$

where $\widehat{\mathrm{Var}}[\hat{\mu}_R]$ is given by (10).

We note that for relatively small n, the condition of approximate normality may not hold for either the sample mean or the random sum estimator. Therefore, this condition, under which we can use the t-distribution in the above formulas, needs verification. Apart from a purely theoretical verification, which is beyond the scope of this paper, we recommend performing bootstrapping to study the distribution of estimators in question.

5 Comparison of estimators

We are interested in analyzing the efficiency of the random sum estimator, compared to the sample mean. More specifically, we would like to determine under what conditions on the population the random sum estimator yields a smaller estimated variance and, as a consequence, a narrower confidence interval. The following parameters will be considered:

1) Population size N,

2) Fraction of nonzeroes q,

3) Mean of the overpayment part of the population μ_1,

4) Coefficient of variation of the overpayment part $k_1 = \dfrac{\sigma_1}{\mu_1}$,

5) Sample size n.

To measure efficiency, we use the following efficiency index:

$$\mathrm{Eff} = \sqrt{\frac{\mathrm{Var}(\hat{\mu}_R)}{\mathrm{Var}(\;\bar{y}\;)}}\,. \tag{15}$$

This index reflects the relative width of the confidence intervals for μ constructed on $\hat{\mu}_R$ and $\bar{y}$. We shall try to identify the combinations of parameters that yield $\mathrm{Eff} \le 0.99$, that is, for which there is a visible improvement that results from using the random sum estimator.

5.1 Relative efficiency study

Using formulas (8) and (10) and keeping in mind Lemma 2, we can rewrite (15) as follows:

$$\mathrm{Eff} = \sqrt{\frac{N}{N-1}}\,, \tag{16}$$

which is almost indistinguishable from 1 if the population size N is large. However, Lemma 2 does not hold for the estimated variances s^2 and s_1^2, so the empirical, or observed, efficiency will differ from the theoretical value (16). To make this adjustment, we first reformulate Lemma 2 for the estimated variances.

Lemma 7. *The estimated variances of the overall sample mean $\bar{y}$ and the overpayment sample mean $\bar{y}_1$ satisfy the following equation:*

$$s^2 = \frac{\hat{q}\,n-1}{n-1}\,s_1^2 + \frac{n}{n-1}\,(1-\hat{q})\,\hat{q}\,\bar{y}_1^2\,.$$

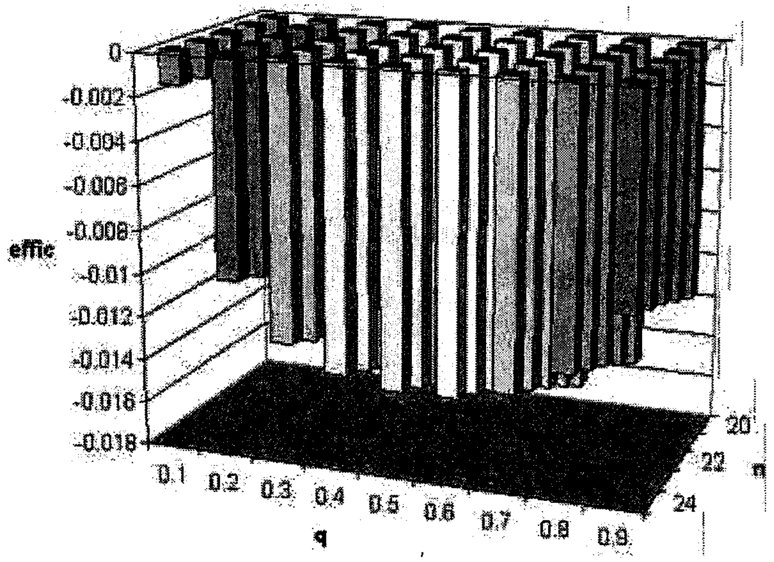
We then analyze the efficiency index of the form

$$\mathrm{Eff} = \sqrt{\frac{\widehat{\mathrm{Var}}(\hat{\mu}_R)}{\widehat{\mathrm{Var}}(\;\bar{y}\;)}}\,, \tag{17}$$

where $\widehat{\mathrm{Var}}(\hat{\mu}_R)$ and $\widehat{\mathrm{Var}}(\bar{y})$ are given by (9) and (14). Such analysis was performed in Microsoft Excel®. A typical graph from the Excel analysis is shown in Figure 1, where the logarithm of Eff is measured along the vertical axis.

The following conclusions were reached based on the Excel analysis.

Figure 1: A typical graph of log(Eff) for fixed N, μ_1, k_1.



1. $\hat{\mu}_R$ is better than $\bar{y}$ when the sample size n is small. For example, if the population size $N = 1000$, then $n = 20$ results in Eff ~ 0.98, and $n = 50$ results in Eff ~ 0.996.

2. Population size N and the true mean μ have no effect on the efficiency of $\hat{\mu}_R$, provided N is sufficiently large.

3. If the coefficient of variation is small, i.e. $k_1 \leq 0.1$, efficiency is nearly constant across the values of $q = 0.1 \ldots 0.9$.

4. If $k_1 > 0.1$, curvature is present with respect to q, with efficiency best at $q = 0.5$ and worse at $q = 0.1$ than at $q = 0.9$.

5. If $k_1 \geq 0.32$, then $\hat{\mu}_R$ is not more efficient than $\bar{y}$ for values of q close to 0.

6. If $k_1 \geq 0.8$, then $\hat{\mu}_R$ is not more efficient than $\bar{y}$ for any value of q. In particular, the efficiency index is close to 1 for $q > 0.5$, and greater than 1 for $q < 0.5$.

5.2 Simulation study

The conclusions of the previous section were based on purely theoretical values of the variances and the efficiency index. Moreover, no conclusion was made about the coverage rate of the confidence intervals based on the two estimators. Due to these facts, a simulation study was conducted in order to test whether the above conclusions hold for the sample statistics, and to compare the coverage rates of the confidence intervals built on $\hat{\mu}_R$ and $\bar{y}$.

Since prior evidence shows that the true mean μ has no effect on the efficiency index, it was fixed throughout the simulation at the level $\mu = 10$. For the rest of the factors, a factorial design was adopted with 4 levels of k_1 and 3 levels of all other factors. Thus 108 combinations of factors were used. The factor levels are listed in Table 1.

Table 1: Simulation design

Factor	Levels			
N	100	550	1000	
k_1	0.1	0.3	0.5	0.7
q	0.1	0.5	0.9	
n	10	20	40	

For every combination of N, k_1 and q, a simulated population with these factor levels is generated in SAS®. Each population is then read into MATLAB®, where for every sample size, 1000 samples are taken, and the 95% confidence intervals based on $\hat{\mu}_R$ and $\bar{y}$ constructed for every sample. Then we compare the length of the confidence intervals and check their coverage rate. The results are averaged across the 1000 samples, and written to an output file. The SAS and MATLAB programs used in the simulation are listed in Appendices A.1 and A.2.

The following three tables summarize the results of the simulation. The results are listed separately for every level of the sample size.

Table 2: Simulation results for sample size $n = 10$

N	k	q	Coverage Rate SRS	Coverage Rate RS	$\lvert CI_{\text{SRS}} \rvert$	$\lvert CI_{\text{RS}} \rvert$	Efficiency
100	0.1	0.1	99.7	99.7	6.2128	5.9466	0.95715
		0.5	92.6	91.2	7.0381	6.7242	0.95539
		0.9	69.2	69.2	3.6994	3.5614	0.96270
	0.3	0.1	99.9	99.9	5.6629	5.6132	0.99122
		0.5	95.5	95.0	7.3421	7.1128	0.96877
		0.9	94.8	94.7	5.5521	5.4593	0.98329
	0.5	0.1	99.7	99.7	6.0210	6.0882	1.01116
		0.5	94.9	94.7	8.2038	8.0256	0.97827
		0.9	95.5	95.5	7.2381	7.1735	0.99107
	0.7	0.1	99.9	99.9	9.4351	9.5969	1.01715
		0.5	93.1	93.4	9.4817	9.4375	0.99533
		0.9	94.4	94.6	8.6615	8.6287	0.99621
550	0.1	0.1	99.7	99.7	6.2313	5.9288	0.95145
		0.5	94.4	92.6	7.1952	6.8457	0.95142
		0.9	73.1	73.1	3.9988	3.8363	0.95936
	0.3	0.1	99.5	99.2	6.5693	6.3870	0.97225
		0.5	97.1	96.5	7.7628	7.4891	0.96475
		0.9	91.3	91.1	5.8948	5.7760	0.97985
	0.5	0.1	99.8	99.8	7.0727	6.9270	0.97940
		0.5	93.2	93.4	8.1005	7.9550	0.98204
		0.9	95.7	95.6	7.6616	7.5763	0.98887
	0.7	0.1	99.5	99.6	7.1085	7.2033	1.01334
		0.5	95.2	95.2	9.3138	9.1998	0.98776
		0.9	95.3	95.3	9.2141	9.1383	0.99177
1000	0.1	0.1	99.4	99.4	6.2843	5.9789	0.95139
		0.5	94.5	92.0	7.2069	6.8551	0.95119
		0.9	72.5	72.4	3.9278	3.7670	0.95907
	0.3	0.1	99.6	99.5	6.3441	6.1494	0.96931
		0.5	95.6	95.1	7.8016	7.5316	0.96539
		0.9	94.1	94.1	5.7533	5.6288	0.97836
	0.5	0.1	99.7	99.5	6.6483	6.6389	0.99859
		0.5	94.8	94.4	8.4185	8.2400	0.97879
		0.9	94.5	94.4	7.4134	7.3259	0.98819
	0.7	0.1	99.7	99.7	7.0257	7.1383	1.01603
		0.5	93.1	93.1	9.7527	9.6397	0.98841
		0.9	93.9	93.8	9.1870	9.1202	0.99273

Table 3: Simulation results for sample size $n = 20$

N	k	q	Coverage Rate SRS	Coverage Rate RS	$\lvert CI_{\text{SRS}}\rvert$	$\lvert CI_{\text{RS}}\rvert$	Efficiency
100	0.1	0.1	99.5	99.2	3.0706	3.0159	0.98218
		0.5	94.1	93.6	4.3464	4.2616	0.98048
		0.9	90.6	90.5	2.4995	2.4560	0.98257
	0.3	0.1	99.6	99.5	2.8356	2.8727	1.01306
		0.5	94.7	94.4	4.5283	4.4699	0.98709
		0.9	94.3	94.3	3.4641	3.4425	0.99378
	0.5	0.1	98.9	98.9	2.9723	3.0502	1.02621
		0.5	94.8	94.7	5.0878	5.0464	0.99187
		0.9	94.9	94.9	4.4929	4.4833	0.99787
	0.7	0.1	99.2	99.2	4.6453	4.7932	1.03184
		0.5	93.7	93.7	5.9362	5.9360	0.99997
		0.9	94.1	94.2	5.4200	5.4223	1.00043
550	0.1	0.1	99.7	99.7	3.2181	3.1441	0.97700
		0.5	95.4	95.2	4.6812	4.5711	0.97648
		0.9	88.5	88.5	2.7760	2.7179	0.97907
	0.3	0.1	99.8	99.6	3.4278	3.4066	0.99382
		0.5	96.1	96.0	4.9901	4.9049	0.98292
		0.9	95.0	95.0	3.9269	3.8873	0.98990
	0.5	0.1	99.6	99.5	3.7109	3.7053	0.99849
		0.5	94.7	94.3	5.3042	5.2594	0.99155
		0.9	93.8	93.7	5.0071	4.9804	0.99467
	0.7	0.1	96.3	96.6	3.6925	3.7865	1.02545
		0.5	93.6	93.6	6.1058	6.0729	0.99461
		0.9	94.7	94.4	5.9725	5.9506	0.99633
1000	0.1	0.1	99.9	99.9	3.2760	3.2003	0.97689
		0.5	94.1	93.7	4.6835	4.5719	0.97617
		0.9	88.4	88.4	2.7963	2.7362	0.97851
	0.3	0.1	99.9	99.9	3.2279	3.2033	0.99237
		0.5	94.8	94.3	5.0825	4.9957	0.98292
		0.9	94.3	94.1	3.7781	3.7371	0.98916
	0.5	0.1	98.2	98.3	3.3781	3.4339	1.01652
		0.5	94.9	94.7	5.5353	5.4769	0.98943
		0.9	96.1	96.1	4.9586	4.9298	0.99419
	0.7	0.1	97.1	97.7	3.5666	3.6620	1.02674
		0.5	95.1	94.9	6.5016	6.4630	0.99406
		0.9	95.6	95.4	6.0913	6.0694	0.99640

Table 4: Simulation results for sample size $n = 40$

N	k	q	Coverage Rate SRS	Coverage Rate RS	$\lvert CI_{\text{SRS}}\rvert$	$\lvert CI_{\text{RS}}\rvert$	Efficiency
100	0.1	0.1	95.6	95.5	1.5738	1.5646	0.99414
		0.5	94.6	94.4	2.5713	2.5529	0.99283
		0.9	94.3	94.2	1.5397	1.5299	0.99359
	0.3	0.1	94.2	94.6	1.4665	1.4867	1.01377
		0.5	94.6	94.6	2.6861	2.6758	0.99617
		0.9	95.1	95.1	2.0600	2.0586	0.99932
	0.5	0.1	92.7	92.9	1.5486	1.5844	1.02312
		0.5	94.9	94.9	3.0160	3.0111	0.99839
		0.9	94.9	94.9	2.6671	2.6708	1.00139
	0.7	0.1	90.6	91.0	2.4430	2.5117	1.02815
		0.5	95.1	95.1	3.5318	3.5409	1.00258
		0.9	95.0	95.0	3.2167	3.2258	1.00281
550	0.1	0.1	99.1	99.0	1.8761	1.8558	0.98918
		0.5	95.6	95.3	3.1387	3.1034	0.98875
		0.9	92.4	92.3	1.9110	1.8915	0.98979
	0.3	0.1	95.4	95.5	1.9783	1.9784	1.00003
		0.5	94.2	94.1	3.3498	3.3229	0.99195
		0.9	94.3	94.2	2.6264	2.6146	0.99550
	0.5	0.1	94.1	94.3	2.1107	2.1198	1.00433
		0.5	95.2	95.2	3.5854	3.5722	0.99630
		0.9	94.4	94.4	3.3918	3.3845	0.99784
	0.7	0.1	91.5	92.4	2.1964	2.2444	1.02185
		0.5	92.9	92.9	4.0695	4.0605	0.99779
		0.9	95.3	95.3	4.0378	4.0325	0.99869
1000	0.1	0.1	98.9	98.8	1.9039	1.8831	0.98908
		0.5	93.5	93.5	3.1714	3.1346	0.98841
		0.9	92.4	92.2	1.9298	1.9093	0.98938
	0.3	0.1	96.0	95.9	1.9230	1.9219	0.99942
		0.5	94.3	94.2	3.4328	3.4046	0.99180
		0.9	95.0	94.8	2.5772	2.5636	0.99474
	0.5	0.1	92.6	92.8	2.0674	2.0950	1.01336
		0.5	94.5	94.2	3.7533	3.7344	0.99499
		0.9	95.6	95.6	3.3272	3.3183	0.99732
	0.7	0.1	92.3	92.7	2.1691	2.2211	1.02398
		0.5	94.5	94.6	4.3976	4.3866	0.99748
		0.9	94.4	94.4	4.1531	4.1467	0.99847

The simulation confirmed earlier conclusions about the efficiency of the random sum estimator $\hat{\mu}_R$. In other words, the conclusions we reached for the theoretical formulas also hold for the sample statistics.

We may also notice that the random sum estimator is the most efficient compared to the sample mean, when the coefficient of variation $k_1 = 0.1$. In addition, as the population size increases from 100 to 1000, the efficiency improves slightly (from 0.957 to 0.951). As sample size increases from 10 to 40, the efficiency index increases to 1, meaning that $\hat{\mu}_R$ does not present any advantages at $n = 40$.

In the best possible scenario, the random sum estimator offers a 5% improvement.

The simulation also offers new information about the coverage rates. As Tables 2, 3, and 4 indicate, the coverage rate is generally close to the nominal value of 95%. Thus, the simulation provides evidence in support of our assumption that both estimators are nearly normal for smaller sample size. This evidence in no way constitutes a definitive proof of the assumption. It can be completely proved (or disproved) only by further research.

Several more observations on coverage rates are worth mentioning. First, the coverage rates are close to 99% when $q = 0.1$, that is, when there are relatively few positive values in the population. This suggests a better approximation by the normal distribution, which results in the coverage rate significantly above the nominal level.

Second, when $k_1 = 0.1, \quad q = 0.9$, we observe a very low coverage rate for both estimators. For $n = 10$, it goes as low as 69%. This fact is troublesome and should be analyzed further.

Third, when $q = 0.5$, we notice the largest discrepancy between the coverage rates of $\hat{\mu}_R$ and $\bar{y}$, with the rate for the random sum estimator more than 1% lower. This phenomenon is most pronounced when $k_1 = 0.1$. We conclude that although $\hat{\mu}_R$ is most efficient when $q = 0.5$, this efficiency is largely achieved through reduction in the coverage rate. This may be acceptable in some situations, while undesirable in others.

6 Conclusions

Our research has shown that random sum estimators of the mean in finite populations should be considered for practical use in limited circumstances. This includes situations when only a small sample can be taken from the population, and when the positive part of the population has relatively small variability. That, of course, implies that the population must be a mixture of zero and positive parts.

In such situations, the random sum estimator offers a 5% reduction in the length of the confidence interval, with the coverage rate close to the nominal level. In the Medicaid example, where the lower end of the confidence interval is used to bill health care providers for overpayment, a 5% reduction in length could mean thousands of dollars more being returned to Medicaid.

However, the use of random sum estimators can be much harder to justify in the court than the standard sample mean, and the option of random sum estimation should be carefully evaluated in all individual situations.

Another possibility is to make use of the additional information that is often available. In the Medicaid study, the total payment for every claim is readily accessible. If we assume that the amount of overpayment is directly proportional to the total payment, then the ratio between the two could be estimated. Approaching the problem from the random sum viewpoint, we arrive at random sum ratio estimators, which is a topic of separate research [2]. Preliminary results of that research indicate that the improvement that results from using random sum ratio estimators is much more substantial than what we see here.

References

[1] J. Borkowski, *Private communication*, 2003.

[2] J. Borkowski and Y. Shvetsov, *Random sum ratio estimators in finite populations*, in progress.

[3] W. Feller, *An introduction to probability theory and its applications, Volume II*, 2nd edition, John Wiley and Sons, New York, 1971.

[4] T. Jiang, C. Su, and Q. H. Tang, *Limit theorems for the random sum of partial sums on independent, identically distributed random variables* (in Chinese), J. Univ. Sci. Technol. China **31** (2001), 394-399.

[5] V. Yu. Korolev, *On the convergence of distributions of random sums of independent random variables to stable laws*, Theory Probab. Appl. **42** (1997), 695-696.

[6] V. M. Kruglov, *Weak compactness of random sums of independent random variables*, Theory Probab. Appl. **43** (1999), 203-220.

[7] A. M. Mood, F. A. Graybill, and D. C. Boes, *Introduction to the theory of statistics*, 3rd edition, McGraw-Hill, New York, 1974.

[8] Z. Rychlik and T. Walczyński, *Convergence in law of random sums with nonrandom centering*, J. Math. Sci (New York) **106** (2001), 2860-2864.

[9] D. O. Selivanova, *Estimates for the rate of convergence in some limit theorems for geometric random sums*, Moscow Univ. Comput. Math. Cybernet. **1995**, no. 2, 27-31.

[10] H. M. Taylor and S. Karlin, *An introduction to stochastic modeling*, Academic Press, Orlando, 1984.

[11] S. K. Thompson, *Sampling*, 2nd edition, John Wiley and Sons, New York, 2002.

[12] P. Vellaisamy and B. Chaudhuri, *Poisson and compound Poisson approximations for random sums of random variables*, J. Appl. Probab. **33** (1996), 127-137.

Appendices

A1. SAS Program

```
dm 'LOG; clear; OUT; clear;';
options nodate nonumber ls=100 ps=3000;


data  popul (drop= i N q ky1 N1 muy1 sy1 seed1 u tailprob);

  N    = 1000;
  q    =  0.9;
  muy1 =   10;
  ky1  =  0.7;

  N1 = N*q;
  y=0; u=0;
  sy1 = muy1*ky1;
  tailprob = probnorm(-muy1/sy1);
  seed1 = round(ranuni(0)*1000000);
  retain seed1 ;

  do i = 1 to N1;
     call ranuni(seed1,u);
     y = probit(u*(1-tailprob) + tailprob);
     y = muy1 + y*sy1;
     output;
  end;

  do i = N1+1 to N;
     y = 0;
     output;
  end;


proc print data = popul;
   ID y;


run;
quit;
```

A2. MATLAB Program

```
% Simulation of random sum estimator of the mean
% Final version


K = 1000;                       % Number of runs for simulation
n = 10;                         % Sample size
trials = 36;                    % Number of trials for every sample size

fileout = fopen('results_n10.dat','w');
fprintf(fileout, 'Trial  CovRate_S  Cov Rate_R  CI Length_S');
fprintf(fileout, 'CI Length_R Efficiency\n\n');

tvalue = tinv(.975,n-1);


for trial = 1:trials

  fname = [ 'popul_' num2str(trial,'%2d') '.dat' ];
  Y = load(fname);
  [N,a] = size(Y); if a~=1
  error('Incorrect data for the population');
end

ymean = mean(Y(:));             % true mean of population
clevel_s = 0;                   % true confidence level of SRS CI
clevel_r = 0;                   % true confidence level of RS CI
avlen_s = 0;                    % average length of CI
avlen_r = 0;

i = 1; while (i <= K)
  J = sample_wor(n,N);     % take a random sample from population
  sample = Y(J);

  n1 = 0;                  % determine the dimension of overpayment part
  while (n1<n)&(sample(n1+1))
     n1 = n1 + 1;
  end

  if n1>1

   % compute statistics for SRS and RS estimators
   ybar_s = mean(sample(:));
   ybar_r = mean(sample(1:n1));
```

```
    s = samplevar(sample(:));
    s1 = samplevar(sample(1:n1));
    q = n1/n;

    var_s = (N-n)/(N*n)*s;
    var_r = q/n*s1*(1-(n-1)/(N-1)) + ybar_r^2 /n *(1-q)*q*(N-n)/(N-1);

    cil_s = ybar_s - tvalue*sqrt(var_s);
    ciu_s = ybar_s + tvalue*sqrt(var_s);
    cil_r = q*ybar_r - tvalue*sqrt(var_r);
    ciu_r = q*ybar_r + tvalue*sqrt(var_r);

    len_s = ciu_s - cil_s;
    len_r = ciu_r - cil_r;
    avlen_s = avlen_s + len_s;
    avlen_r = avlen_r + len_r;

    capt_s = (cil_s < ymean)&(ymean < ciu_s);
    capt_r = (cil_r < ymean)&(ymean < ciu_r);

    clevel_s = clevel_s + capt_s;
    clevel_r = clevel_r + capt_r;

    i = i+1;
  end

end

clevel_s = clevel_s/K*100;
clevel_r = clevel_r/K*100;

avlen_s = avlen_s/K;
avlen_r = avlen_r/K;
rlen = avlen_r/avlen_s;

fprintf(fileout, '%12.4f  & %12.4f  & %12.4f  & %12.4f  & %6.5f \\\\ \n',
                  clevel_s,clevel_r,avlen_s,avlen_r,rlen);
trial

end

status = fclose(fileout);
```

```
function Sample = samplewor(ss,ps)
%
% A function for selecting a sample without replacement
%
%   ss = sample size
%   ps = population size
%
clear Sample
r = [1:ps]';
out = [];
for j=1:ss;
  kk = unidrnd(length(r));
  out = [out;r(kk)];
  r(kk) = [];
end;
Sample = sort(out);
return
```

```
function svar = samplevar(X)
% SAMPLEVAR compute sample variance s^2 of X

clear svar;
[n,a] = size(X);
if n==1
   n = a;
elseif a==1
else
   error('Incorrect dimension');
end

svar = sum((X-mean(X)).^2)/(n-1);
```