Hindawi
International Journal of Mathematics and Mathematical Sciences
Volume 2023, Article ID 3064736, 5 pages
https://doi.org/10.1155/2023/3064736

Research Article

Estimation of Finite Population Mean under Probability-Proportional-to-Size Sampling in the Presence of Extreme Values

Richard Ayinzoya[1] **and Dioggban Jakperik**[2]

[1]*Department of Statistics and Actuarial Science, School of Mathematical Sciences, C. K. Tedam University of Technology and Applied Sciences, Navrongo, Ghana*
[2]*Department of Biometry, School of Mathematical Sciences, C. K. Tedam University of Technology and Applied Sciences, Navrongo, Ghana*

Correspondence should be addressed to Dioggban Jakperik; jdioggban@cktutas.edu.gh

Received 7 April 2023; Revised 6 June 2023; Accepted 21 June 2023; Published 27 June 2023

Academic Editor: Bikash Koli Dey

This article developed an estimator for finite population mean under probability-proportional-to-size sampling in the presence of extreme values. Theoretical properties such as bias, variance, and consistency are derived. Monte Carlo simulations were performed to assess the consistency and efficiency of the proposed estimator. It is found that the proposed estimator is more efficient than the competing estimators for all values of c between 0 and 1. The gain in precision of the proposed estimator is much higher than that of its competitors for small values of c. Empirical applications of the proposed estimator are illustrated using three real data sets, and the results revealed that the proposed estimator performed better than the conventional and Sarndal (1972) estimators.

1. Introduction

Over the years, attempts have been made by many researchers to enhance estimates of population parameters such as mean, total, and median with optimum statistical properties [1]. For instance, recent studies [2, 3] and many others proposed estimators for various parameters with the aim of finding estimates which describe data with extreme values in a good manner. However, the mean estimate is either under- or overestimated in situations where extreme values are present in the study variable such as expenditure, taxes, income, production, and consumption. These extreme values introduce significant bias since they increase the estimation variance. Conventional estimation methods are unable to provide realistic and precise estimates in such cases. Specialized techniques based on nonparametric, semiparametric, and biased reduction densities are employed to increase precision of such estimates [4, 5]. Alternative approaches to developing estimators for finite population means is the use of regression-based estimators, see for instance [6]. Though significant gains in precision can be obtained with these methods, they are generally computationally laborious and time consuming, especially in large samples [7, 8]. These, among others, called for consideration of alternative methods that fairly are user friendly without sacrificing precision.

To overcome this challenge, Sarndal [9] proposed an unbiased estimator for a finite population mean in the presence of extreme values under simple random sampling. The authors in other works [2, 10, 11] proposed an improved ratio-type estimator for the estimation of finite population mean when there exist minimum or maximum values. Moreover, a ratio, product, and regression type estimators for the estimation of finite population mean when there exist extreme values were proposed [1, 10, 12]. Other procedures have been proposed in recent studies aimed at increasing precision of mean estimates when variability in the study population is high [1, 6, 13].

Although these approaches have achieved significant improvement in the precision of population parameters, the gain in precision and computational efficiencies still leave much to be desired. This study therefore seeks to develop an efficient estimator for the estimation of the finite population mean in the presence of extreme values.

The rest of the paper is organized as follows. Section 2 presents literature review on existing mean estimators; Section 3 contains the proposed estimator and the derivations of its theoretical properties. In Section 4, the comparisons of the theoretical properties of the proposed estimator with the competing estimators are carried out. The simulation and empirical studies are contained in Sections 5 and 6, respectively, whilst conclusion is in Section 7.

2. Review of Existing Mean Estimators

Consider selecting a random sample of size n from a population of size N and the probability of selection p_i associated with the size of the primary units, an unbiased estimator of population mean and variance under probability-proportional-to-size sampling scheme are given as follows:

$$\overline{y}_z = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{Np_i}\right), \tag{1}$$

and

$$V(\overline{y}_z) = \frac{\sigma_z^2}{n}, \tag{2}$$

where

$$\sigma_{y_z}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i}{Np_i} - \overline{Y}\right)^2, \tag{3}$$

and y is the study variable, Y_i is the value of the study variable y for the i^{th} population unit, $i= 1,\ldots,N$, p_i is the selection probability of i^{th} unit in the population at any given draw. y_z is the define variate given by $y_z = (Y_i/Np_i)$. $i = 1,2,\ldots,N$. Without loss of generality, $\overline{y}_z = \overline{y}$.

To avoid overestimating or underestimating the population mean when observations in actual data contain unexpected large or small values, Sarndal [9] suggested an unbiased estimator given by the following equation:

$$\overline{y}_s = \begin{cases} \overline{y} + c, & \text{if sample contain } y_{\min} \text{ but not } y_{\max}, \\ \overline{y} - c, & \text{if sample contain } y_{\max} \text{ but not } y_{\min}, \\ \overline{y}, & \text{for all other sample contain,} \end{cases} \tag{4}$$

with

$$V(\overline{y}_s) = \lambda s_y^2 - \frac{2\lambda \mathrm{nc}}{N-1}(y_{\max} - y_{\min} - \mathrm{nc}), \tag{5}$$

where $\lambda = (1/n)$, c is a constant with

$$c_{\mathrm{opt}} = \frac{(y_{\max} - y_{\min})}{2n}, \qquad V(\overline{y}) = \lambda s_y^2, \tag{6}$$

is the conventional variance.

The minimum mean variance is given by the following equation:

$$V(\overline{y}_s) = \lambda s_y^2 - \frac{\lambda(y_{\max} - y_{\min})^2}{2(N-1)}. \tag{7}$$

The major drawback of this estimator is its slow convergence for small values of c leading to reduced precision of mean estimates. To address this challenge, Ahmad and Shabbir [1] proposed a product ratio estimator using an auxiliary variable. This led to a complex estimator without significant gain in precision. The square root transformation of c provides the needed stability of the variance and hence improves precision remarkably without much computational efforts.

3. Proposed Estimator

The proposed estimator is a modification of the Sarndal estimator [9] for finite population mean. An estimator of finite population mean when extreme values are present under probability proportional to size sampling scheme is proposed.

Let $y_i, i = 1,2,\ldots,n$ be independent and identically distributed random samples with mean $\overline{y}$ under probability proportional to size sampling scheme and c, a non-negative constant. The proposed estimator is formulated as follows:

$$\overline{y}_{\mathrm{pp}} = \begin{cases} \overline{y} + \sqrt{c}\,, & \text{if sample contain } y_{\min} \text{ but not } y_{\max}, \\ \overline{y} - \sqrt{c}\,, & \text{if sample contain } y_{\max} \text{ but not } y_{\min}, \\ \overline{y}, & \text{for all other samples.} \end{cases} \tag{8}$$

The bias of the proposed estimator is

$$\begin{aligned} \text{Bias} &= E(\overline{y}_{\mathrm{pp}}) - \overline{Y}, \\ \Rightarrow B(\overline{y}_{\mathrm{pp}}) &= \left(\sum_{s_1}(\overline{y} + \sqrt{c}\,) + \sum_{s_2}\left(\overline{y} - \sqrt{c} + \sum_{s_3}(\overline{y})\right)p_i - \overline{Y}\right. \\ B(\overline{y}_{\mathrm{pp}}) &= \sum_{s_i \in s}\overline{y}p_i - \overline{Y}, \end{aligned} \tag{9}$$

but

$$\sum_{s_i \in s}\overline{y}p_i = \overline{Y} \qquad \Rightarrow B(\overline{y}_{pp}) = 0. \tag{10}$$

The variance of the proposed estimator is given by

$$V(\overline{y}_{pp}) = \lambda s_y^2 - \frac{2\lambda n\sqrt{c}}{N-1}(y_{\max} - y_{\min} - n\sqrt{c}\,), \tag{11}$$

where $\lambda = (1/n)$, c is a constant and the conventional variance is

$$V(\overline{y}) = \lambda s_y^2. \tag{12}$$

Consistency of the proposed estimator is a limiting function of the bias, thus,

$$\begin{aligned} &\lim_{n\to\infty}\left(B\left(\overline{y}_{\mathrm{pp}}\right)\right)_n = \lim_{n\to\infty}\left(\overline{y}_{\mathrm{pp}}\right) = 0,\\ &\lim_{n\to\infty}\left(B\left(\overline{y}_{\mathrm{pp}}\right)\right)_n = 0. \end{aligned} \tag{13}$$

is trivial.

4. Comparison of Estimators

In this section, the proposed estimator is compared with the conventional and Sarndal [9] estimators under the probability-proportional-to-size sampling scheme.

4.1. Condition (i). From equations (2) and (11),

$$V(\overline{y}) - V\left(\overline{y}_{pp}\right) = \frac{2\lambda n\sqrt{c}}{N-1}\left(Y_{\max} - Y_{\min} - n\sqrt{c}\,\right) > 0. \tag{14}$$

Suppose $(y_{\max} - y_{\min} - n\sqrt{c}\,) = 0$, and $\sqrt{c}$ is positive. Consequently,

$$0 < \sqrt{c} < \frac{Y_{\max} - Y_{\min}}{n}. \tag{15}$$

4.2. Condition (ii). From equations (5) and (11),

$$\begin{aligned} V\left(\overline{y}_s\right) - V\left(\overline{y}_{\mathrm{pp}}\right) &= \frac{2\lambda \mathrm{nc}}{N-1}\left(Y_{\max} - Y_{\min} - \mathrm{nc}\right) - \frac{2\lambda n\sqrt{c}}{N-1}\left(Y_{\max} - Y_{\min} - n\sqrt{c}\,\right) > 0,\\ &\Rightarrow c\left(Y_{\max} - Y_{\min} - nc\right) - \sqrt{c}\left(Y_{\max} - Y_{\min} - n\sqrt{c}\,\right) > 0. \end{aligned} \tag{16}$$

Suppose $(Y_{\max} - Y_{\min} - nc) - \sqrt{c}\,(Y_{\max} - Y_{\min} - n\sqrt{c}\,) = 0$, and $\sqrt{c}$ is positive.

Consequently,

$$0 < \sqrt{c} < \frac{Y_{\max} - Y_{\min}}{n}. \tag{17}$$

Thus, the proposed estimator performs better than the conventional and the Sarndal estimator [9] when conditions (i) and (ii) are satisfied.

5. Simulation

Monte Carlo simulations were performed for samples of size $n = 50, 100, 150, 200, 250, 300$ for 500 replications under the probability-proportional-to-size sampling scheme for a finite population of size 5000 units with an extremely minimum value of 10 and maximum value of 4900. The variance of the proposed estimator and the Sarndal estimator [9] were determined for different sample sizes of n and values of c. The variance of the conventional estimator and $(Y_{\max} - Y_{\min})$ were assumed to be constant for each sample size, n. The variance of the proposed estimator and Sarndal estimator [9] were computed at $c = 0.01, 0.03, 0.05, 0.07, 0.1, 0.3, 0.5, 0.7, 1.0, 1.3, 1.5, 1.7, 2.0$. Tables 1–6 show the variance of the estimator for each n value at different values of c, respectively. As seen in Table 1, the variance of the proposed estimator is smaller than that of Sarndal estimator [9] for all c values between 0 and 1. Furthermore, as the sample size increases, lower variances are observed for both estimators. This suggests that the mean approaches the population mean, hence demonstrates consistency of the proposed estimator.

The variances of estimators for sample size 100 are presented in Table 2.

For sample size 150, the results are indicated in Table 3.

The variance of estimators for sample size 200 is shown in Table 4.

The variance of estimators at sample size 250 is shown in Table 5.

Finally, results of the variance of estimator for sample size 300 are indicated in Table 6.

6. Empirical Applications

To determine the performance of the proposed estimator relative to some existing estimators, three data sets from three different populations were used. Two data sets were obtained from literature [14, 15], and the third data set was extracted from Ghana Living Standard Survey Round 7 data [16]. The estimates for these populations are given in the following.

Population 1 (see [15])

Y: area under wheat crop in 1964.

$N = 34$, $n = 12$, $Y_{\max} = 634$, $Y_{\max} = 6$, $S_y^2 = 22564.56$, $\overline{Y} = 199.441$.

Population 2 (see [14])

Y: population size in 1930 (in 1000).

$N = 49$, $n = 20$, $Y_{\max} = 634$, $Y_{\max} = 46$, $S_y^2 = 15158.83$, $\overline{Y} = 127.7959$.

Population 3 (see [16])

Y: total amount on house expenses.

Table 1: Variance of estimators for n = 50.

c	Estimators		
	$V(\bar{y})$	$V(\bar{y}_s)$	$V(\bar{y}_{pp})$
0.01	42807.3100	856.1266	855.9508
0.03	42807.3100	856.0875	855.8079
0.05	42807.3100	856.0484	855.7097
0.07	42807.3100	856.0094	855.6300
0.1	42807.3100	855.9508	855.5295
0.3	42807.3100	855.5611	855.0806
0.5	42807.3100	855.1730	854.7728
0.7	42807.3100	854.7865	854.5234
1.0	42807.3100	854.2098	854.2098
1.3	42807.3100	853.6367	853.9416
1.5	42807.3100	853.2566	853.7801
1.7	42807.3100	852.8781	853.6294
2.0	42807.3100	852.3134	853.4195

Table 2: Variance of estimators for n = 100.

c	Estimators		
	$V(\bar{y})$	$V(\bar{y}_s)$	$V(\bar{y}_{\text{pp}})$
0.01	18535.1300	185.3317	185.1561
0.03	18535.1300	185.2926	185.0136
0.05	18535.1300	185.2536	184.9158
0.07	18535.1300	185.2145	184.8365
0.1	18535.1300	185.1561	184.7366
0.3	18535.1300	184.7680	184.2917
0.5	18535.1300	184.3831	183.9879
0.7	18535.1300	184.0014	183.7425
1.0	18535.1300	183.4349	183.4349
1.3	18535.1300	182.8756	183.1727
1.5	18535.1300	182.5067	183.0152
1.7	18535.1300	182.1411	182.8685
2.0	18535.1300	181.5985	182.6646

Table 3: Variance of estimators for n = 150.

c	Estimators		
	$V(\bar{y})$	$V(\bar{y}_s)$	$V(\bar{y}_{\text{pp}})$
0.01	14289.9900	95.2470	95.0716
0.03	14289.9900	95.2080	94.9295
0.05	14289.9900	95.1689	94.8321
0.07	14289.9900	95.1300	94.7532
0.1	14289.9900	95.0716	94.6539
0.3	14289.9900	94.6851	94.2130
0.5	14289.9900	94.3034	93.9132
0.7	14289.9900	93.9265	93.6718
1.0	14289.9900	93.3702	93.3702
1.3	14289.9900	92.8247	93.1140
1.5	14289.9900	92.4670	92.9605
1.7	14289.9900	92.1142	92.8178
2.0	14289.9900	91.5939	92.6199

Table 4: Variance of estimators for n = 200.

c	Estimators		
	$V(\bar{y})$	$V(\bar{y}_s)$	$V(\bar{y}_{\text{pp}})$
0.01	11399.6700	56.9788	56.8035
0.03	11399.6700	56.9397	56.6619
0.05	11399.6700	56.9007	56.5649
0.07	11399.6700	56.8618	56.4863
0.1	11399.6700	56.8035	56.3877
0.3	11399.6700	56.4186	55.9508
0.5	11399.6700	56.0402	55.6550
0.7	11399.6700	55.6681	55.4175
1.0	11399.6700	55.1220	55.1220
1.3	11399.6700	54.5903	54.8717
1.5	11399.6700	54.2438	54.7223
1.7	11399.6700	53.9037	54.5836
2.0	11399.6700	53.4056	54.3916

Table 5: Variance of estimators for n = 250.

c	Estimators		
	$V(\bar{y})$	$V(\bar{y}_s)$	$V(\bar{y}_{\text{pp}})$
0.01	8872.7680	35.4715	35.2964
0.03	8872.7680	35.4325	35.1552
0.05	8872.7680	35.3935	35.0586
0.07	8872.7680	35.3546	34.9805
0.1	8872.7680	35.2964	34.8824
0.3	8872.7680	34.9132	34.4495
0.5	8872.7680	34.5379	34.1577
0.7	8872.7680	34.1706	33.9243
1.0	8872.7680	33.6347	33.6347
1.3	8872.7680	33.1168	33.3905
1.5	8872.7680	32.7815	33.2450
1.7	8872.7680	32.4543	33.1103
2.0	8872.7680	31.9784	32.9244

Table 6: Variance of estimators for n = 300.

c	Estimators		
	$V(\bar{y})$	$V(\bar{y}_s)$	$V(\bar{y}_{\text{pp}})$
0.01	7307.527	24.3389	24.1640
0.03	7307.527	24.2998	24.0232
0.05	7307.527	24.2609	23.9270
0.07	7307.527	24.2221	23.8492
0.1	7307.527	24.1640	23.7518
0.3	7307.527	23.7823	23.3229
0.5	7307.527	23.4102	23.0351
0.7	7307.527	23.0478	22.8056
1.0	7307.527	22.5221	22.5221
1.3	7307.527	22.0180	22.2838
1.5	7307.527	21.6939	22.1424
1.7	7307.527	21.3794	22.0116
2.0	7307.527	20.9257	21.8317

$N = 9594$, $n = 500$, $Y_{\max} = 73247.86$, $Y_{\max} = 10$, $S_y^2 = 3345353$, and $\overline{Y} = 951.5913$.

Table 7 shows the variance associated with each of the estimators in different populations. It is observed that the variance of the proposed estimator is smaller in each population compared with the conventional and Sarndal [9] estimators. The proposed estimator is a better estimator of mean than existing ones, especially for large sample sizes [7, 8].

Table 7: Variance of estimators for different populations.

Estimators	Population 1	Population 2	Population 3
$V(\overline{y})$	1880.3800	757.9415	6690.7060
$V(\overline{y}_s)$	137.8448	25.8554	12.6449
$V(\overline{y}_{\mathrm{pp}})$	130.1441	20.9896	12.3552

Table 8: PRE of estimators for different populations.

Estimators	Population 1	Population 2	Population 3
$\overline{y}$	100.0000	100.0000	100.0000
$\overline{y}_s$	7.3307	3.4113	0.1890
$\overline{y}_{\mathrm{pp}}$	6.9212	2.7693	0.1847

The following expression is used for efficiency comparison:

$$\mathrm{PRE} = \frac{V(\overline{y}_i)}{V(\overline{y})} \times 100, i = s, pp. \tag{18}$$

The percent relative efficiencies are summarized in Table 8.

Clearly, the proposed estimator is consistently better than its competitors in both simulation and applications, especially when the value of c is less than unity.

7. Conclusion

A new estimator for a finite population mean under the probability proportional to size sampling in the presence of extreme values is proposed. Theoretical properties such as bias and variance were derived. Empirical studies on real life data and simulation studies were performed, and the proposed estimator was compared with existing estimators. Empirical results confirmed the proposed estimator to have smaller variance than the conventional and Sarndal [9] estimators. The proposed mean estimator was found to be better and more efficient than the existing estimators for small values of c.

Data Availability

All the data used in this study are published data and hence publicly available.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] S. Ahmad and J. Shabbir, "Use of extreme values to estimate finite population mean under pps sampling scheme," *Journal of Reliability and Statistical Studies*, vol. 11, no. 2, pp. 99–112, 2018.

[2] U. Daraz, J. Shabbir, and H. Khan, "Estimation of finite population mean by using minimum and maximum values in stratified random sampling," *Journal of Modern Applied Statistical Methods*, vol. 17, no. 1, pp. 1–20, 2018.

[3] H. P. Singh, A. Mishra, and S. K. Pal, "Improved estimator of population total in PPS sampling," *Communications in Statistics-Theory and Methods*, vol. 47, no. 4, pp. 912–934, 2018.

[4] D. Jakperik, R. O. Otieno, and G. O. Orwa, "A semi-parametric multiplicative bias reduction density with a parametric start," *Advances and Applications in Statistics*, vol. 53, no. 6, pp. 715–729, 2018.

[5] D. Jakperik, R. O. Otieno, and G. O. Orwa, "Variance estimation for poverty indicators via linearization technique," *Journal of Computations and Modeling*, vol. 9, no. 1, pp. 1–9, 2019.

[6] U. Shahzad, N. H. Al-Noor, M. Hanif, I. Sajjad, and M. Muhammad Anas, "Imputation based mean estimators in case of missing data utilizing robust regression and variance-covariance matrices," vol. 51, no. 8, pp. 4276–4295, 2022.

[7] J. Dioggban, "A note non nonparametric regression modeling using a density function," *African Journal of Applied Statistics*, vol. 7, no. 2, pp. 993–1000, 2020.

[8] D. Jakperik, R. O. Otieno, and J. Okungu, "Estimating a finite population total using a density function," *Eurasian Journal of Mathematics*, vol. 4, no. 1, pp. 32–37, 2021.

[9] C.-E. Särndal and C. E. Sarndal, "Sample survey theory vs. general statistical theory: estimation of the population mean," *International Statistical Review/Revue Internationale de Statistique*, vol. 40, no. 1, pp. 1–12, 1972.

[10] A. Iftikhar, H. Shi, S. Hussain et al., "Estimation of finite population mean in presence of maximum and minimum values under systematic sampling scheme," *AIMS Mathematics*, vol. 7, no. 6, pp. 9825–9834, 2022.

[11] M. Khan and J. Shabbir, "Some improved ratio, product, and regression estimators of finite population mean when using minimum and maximum values," *The Scientific World Journal*, vol. 2013, pp. 1–7, 2013.

[12] S. Al-Marzouki, C. Chesneau, S. Akhtar et al., "Estimation of finite population mean under PPS in presence of maximum and minimum values," *AIMS Mathematics*, vol. 6, no. 5, pp. 5397–5409, 2021.

[13] S. Ali, M. Khan, and J. Shabbir, "Using extreme values and fractional raw moments for mean estimation in stratified sampling," *Hacettepe Journal of Mathematics and Statistics*, vol. 47, no. 2, pp. 383–402, 2018.

[14] W. G. Cochran, *Sampling Techniques*, John Wiley & Sons, Hoboken, NY, USA, 1977.

[15] M. N. Murthy, *Sampling Theory and Methods*, Statistical Publishing Society, Barrackpore, India, 1967.

[16] Ghana Statistical Service, "Living standards survey Round 7 (GLSS 6)," *Poverty Profile in Ghana (2016-2017)*, Ghana Statistical Service, Accra, Ghana, 2017.

[17] A. Y. Al-Hossain and M. Khan, "Efficiency of ratio, product, and regression estimators under maximum and minimum values, using two auxiliary variables," *Journal of Applied Mathematics*, vol. 2014, Article ID 693782, 2014.