Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 2, 2025 (New Series), pp 323–329
http://www.ssca.org.in/journal

Prabhu-Ajgaonkar's 1967 Result Revisited

Bikas K. Sinha[1] **and Manisha Pal**[2]

[1]*Retired Professor, Stat-Math Unit, Indian Statistical Institute, Kolkata, India*
[2]*Department of Statistics, St. Xavier's University, Kolkata, India*

Received: 13 March 2025; Revised: 30 March 2025; Accepted: 02 April 2025

Abstract

This short communication revisits the work of Ajgaonkar (1967), wherein the author claimed that the precision of an estimator of population total (or population mean) decreases on increasing the sample size, if the estimator under consideration is an average function, and the population elements are drawn with varying probabilities of selection at each draw. Prabhu-Ajgaonkar (1967) justified this claim through some examples. However, this paper disproves the claim so made!

Key words: Finite population; Unequal probability sampling; Unbiased estimation of population total; Prabhu-Ajgaonkar's claim; Lanke's estimator.

1. Introduction

In survey sampling theory, it is a common belief that an increase in sample size will result in increase of precision of a linear unbiased estimator of the population total (or population mean) under a given sampling scheme. However, Prabhu-Ajgaonkar (1967) made a claim that looked counter-intuitive. He stated that when the population units are drawn with varying probabilities of selection at each draw and the estimator/parameter under consideration is an average function, the statement does not hold good. Examples were presented, whereby he compared unbiased estimators based on sample sizes 1 and 2 to emphasize his point. His study did not raise any question/concern.

Meanwhile, Lanke (1975) published his Ph.D. thesis which received much attention of the researchers in the area of statistical inference, particularly in finite population inference. For any given unbiased estimator (linear or non-linear) of a population parameter, Lanke (1975) defined an improved unbiased estimator by Rao-Blackwellization. The improved estimator is now frequently termed as "Lanke's Estimator". By virtue of the result, we have that in comparison to any homogeneous linear unbiased estimator of a finite population total based on a given sampling scheme, the Lanke's estimator is superior!

In this short communication, we revisit Prabhu-Ajgaonkar's claim. We consider an example of finite population cited in Prabhu-Ajgaonkar (1967), and show that under the same

Corresponding Author: Bikas K. Sinha
Email: bikasksinha118@gmail.com

sampling design with sample size $n = 2$, the Lanke's unbiased estimator of the population total is not only more precise than the unbiased estimator considered by Prabhu-Ajgaonkar, but also better than the unique unbiased estimator for $n = 1$. Further, we establish the superiority of Lanke's estimator for any general finite population size(N).

2. Illustration as in Prabhu-Ajgaonkar (1967)

To illustrate his claim, Prabhu-Ajgaonkar (1967) used the three populations furnished by Yates and Grundy (1953), each having $N = 4$ units with unequal probabilities of selection in a draw. For sample size $n = 1$, sampling is done with unequal probabilities, while for $n = 2$, sampling is done without replacement, and the first unit is drawn with unequal probabilities while in the second draw the remaining units are assumed to have equal probabilities of selection.

We consider the first population, for which the values (Y_i) of the study character Y and the selection probabilities (p_i) of the units are as given in Table 1.

Table 1: Population of size $N = 4$

Unit i	1	2	3	4	Total
Y_i	0.5	0.2	0.1	0.2	7.0
p_i	0.1	0.2	0.3	0.4	1.0

The population total is $T(Y) = 7.0$.

For a sample of size $n = 1$, the unique unbiased estimator $\hat{T}_1(Y)$ of $T(Y)$ gives the estimate

$$e(s_i) = \frac{Y_i}{p_i}, \quad \text{for sample } s_i = \{i\}, \quad i = 1, 2, 3, 4.$$

It is readily verified that $E[\hat{T}_1(Y)] = T(Y) = 7.0$, and $\text{Var}[\hat{T}_1(Y)] = 1.00$.

Next, the sampling design is extended to the case of $n = 2$ by drawing exactly one of the remaining 3 units at random. In effect, therefore, one is dealing with the extended design specified in Table 2 below.

Table 2: Extended sampling design for $n = 2$

Sample $s(i, j)$	(1,2)	(1,3)	(1,4)	(2,3)	(2,4)	(3,4)
$p^*(s(i, j))$	$\frac{0.3}{3}$	$\frac{0.4}{3}$	$\frac{0.5}{3}$	$\frac{0.5}{3}$	$\frac{0.6}{3}$	$\frac{0.7}{3}$

In the above table, $p^*(s(i, j)) = \dfrac{p_i + p_j}{3}$, for $i, j = 1, 2, 3, 4$ with $i < j$.

The unbiased estimator $\hat{T}_2(Y)$ considered by Prabhu-Ajgaonkar (1967) for $n = 2$ is

the Horvitz-Thompson estimator (HTE). It is given by

$$\hat{T}_2(Y|s(i,j)) = \left(\frac{Y_i}{\pi_i} + \frac{Y_j}{\pi_j}\right),$$

where π_i denote the inclusion probabilities of the population units.

The computation of the estimates $e[s(i,j)]$ of the population total are indicated in Table 3 below.

Table 3: HTE of the population total for $n = 2$

Sample $s(i,j)$	(1,2)	(1,3)	(1,4)
$e[s(i,j)]$	$3\left(\frac{0.5}{1.2} + \frac{1.2}{1.4}\right)$	$3\left(\frac{0.5}{1.2} + \frac{2.1}{1.6}\right)$	$3\left(\frac{0.5}{1.2} + \frac{3.2}{1.8}\right)$
Sample $s(i,j)$	(2,3)	(2,4)	(3,4)
$e[s(i,j)]$	$3\left(\frac{1.2}{1.4} + \frac{2.1}{1.6}\right)$	$3\left(\frac{1.2}{1.4} + \frac{3.2}{1.8}\right)$	$3\left(\frac{2.1}{1.6} + \frac{3.2}{1.8}\right)$

Thus,

$$E[\hat{T}_2(Y)] = 7.00 = T(Y), \quad \text{and} \quad \text{Var}[\hat{T}_2(Y)] = 2.8833 > 1 = \text{Var}[\hat{T}_1(Y)].$$

Note: Prabhu-Ajgaonkar (1967) wrongly reported the variance as 6.32 – it was, of course, a computational error. This was his contention to claim that sample size $n = 1$ fares better than sample size $n = 2$. However, the above estimator is far from being Lanke's estimator!

3. Lanke's estimator

Lanke's unbiased estimator $\hat{T}_L(Y)$ of the population total, based on the extended sampling design given in Table 2, gives the sample estimates of the population total as:

$$e^*[s(i,j)] = \frac{Y_i + Y_j}{p_i + p_j}, \quad i,j = 1,2,3,4,\ i < j.$$

We note that:

1. $E[\hat{T}_L(Y)] = 7.00$
2. $E[\hat{T}_L^2(Y)] = 49.3629 \ \Rightarrow \ \text{Var}[\hat{T}_L(Y)] = 49.3629 - 7.00^2 = 0.3629$

which implies:

$$\text{Var}[\hat{T}_L(Y)] = 0.3629 < 1 = \text{Var}[\hat{T}_1(Y)].$$

Thus, it transpires that Lanke's estimator for sample size $n = 2$ fares better than the unique unbiased estimator for $n = 1$, as it should.

Remark 1: It may be noted that **Lanke's estimator is different from the usual Horvitz-Thompson Estimator (HTE)**, which was taken by Prabhu-Ajgaonkar (1967) in his study. The HTE uses inclusion probabilities in the denominator, while Lanke's estimator uses the sum of selection probabilities of the sample.

4. Variance inequality for any finite population

In this section, the variance inequality

$$\mathrm{Var}(\hat{T}_L(Y)) \leq \mathrm{Var}(\hat{T}_1(Y))$$

is established, where $\hat{T}_1(Y)$ and $\hat{T}_L(Y)$ denote, respectively, the unbiased estimators of the population total $T(Y)$ based on sample sizes $n = 1$ and $n = 2$. Here, $\hat{T}_1(Y)$ is the Horvitz–Thompson estimator (HTE), and $\hat{T}_L(Y)$ is Lanke's estimator under the sampling design indicated in Prabhu-Ajgaonkar (1967).

Let Y_i, $i = 1, 2, \ldots, N$ denote the values of the study variable, and p_i the corresponding probabilities of selection for each unit. Then:

$$\mathrm{Var}(\hat{T}_1(Y)) = \sum_{i=1}^{N} \frac{Y_i^2}{p_i} - [T(Y)]^2 .$$

$$\mathrm{Var}(\hat{T}_L(Y)) = \sum_{i<j} \frac{(Y_i + Y_j)^2}{\frac{p_i+p_j}{2(N-1)}} - [T(Y)]^2$$

A sufficient condition for

$$\mathrm{Var}(\hat{T}_1(Y)) \geq \mathrm{Var}(\hat{T}_L(Y))$$

is that for all pairs (i, j) with $i < j$:

$$(N-1)\left[\frac{Y_i^2}{p_i} + \frac{Y_j^2}{p_j}\right] \geq \frac{(Y_i + Y_j)^2}{p_i + p_j}$$

This inequality is equivalent to:

$$(N-2)\left[\frac{Y_i^2}{p_i} + \frac{Y_j^2}{p_j}\right] + \frac{p_j Y_i^2}{p_i(p_i+p_j)} + \frac{p_i Y_j^2}{p_j(p_i+p_j)} - \frac{2Y_iY_j}{p_i+p_j} \geq 0$$

which simplifies to:

$$(N-2)\left[\frac{Y_i^2}{p_i} + \frac{Y_j^2}{p_j}\right] + \frac{1}{p_ip_j(p_i+p_j)}(p_jY_i - p_iY_j)^2 \geq 0 \quad \text{for all } i < j.$$

Since the left-hand side is a sum of non-negative quantities and $N \geq 2$, this inequality always holds. Hence, we conclude

$$\mathrm{Var}(\hat{T}_1(Y)) \geq \mathrm{Var}(\hat{T}_L(Y)).$$

Thus, the claim made by Prabhu-Ajgaonkar (1967) is not tenable.

Remark 2: It may be mentioned here that Pal and Sinha (2024) devoted a chapter to the use of additional resources for the improvement of an estimator. This area of research has not yet been duly exposed to the researchers in the broad area of statistical inference.

5. Estimation of $\mathrm{Var}(\hat{T}_L(Y))$

We proceed to unbiasedly estimate $\mathrm{Var}(\hat{T}_L(Y))$ in the general set-up described in Section 4. Since

$$\mathrm{Var}(\hat{T}_L(Y)) = E\left[\hat{T}_L(Y)^2\right] - T^2(Y),$$

it suffices to find an unbiased estimator of $T^2(Y)$.

Now,

$$T^2(Y) = \left(\sum_{i=1}^{N} Y_i\right)^2 = \sum_{i=1}^{N} Y_i^2 + \sum_{\substack{i,j=1 \\ i\neq j}}^{N} Y_i Y_j.$$

Note that the cross-product term can be written as

$$\sum_{\substack{i,j=1 \\ i\neq j}}^{N} Y_i Y_j = \sum_{\substack{i,j=1 \\ i\neq j}}^{N} \frac{Y_i^2 + Y_j^2}{2(N-1)} + \sum_{\substack{i,j=1 \\ i\neq j}}^{N} Y_i Y_j,$$

which simplifies the expression.

Consider the following estimator of $T^2(Y)$ based on the sample $s(i,j)$:

$$\hat{T}^2(Y) \mid_{s(i,j)} = \frac{1}{p_i + p_j}\left[Y_i^2 + Y_j^2 + 2(N-1)Y_i Y_j\right].$$

It is easily verified that this is an unbiased estimator of $T^2(Y)$.

Hence, an unbiased estimate of $\mathrm{Var}(\hat{T}_L(Y))$ for the sample $s(i,j)$ is:

$$\widehat{\mathrm{Var}}(\hat{T}_L(Y)) \mid_{s(i,j)} = \left(\frac{Y_i + Y_j}{p_i + p_j}\right)^2 - \frac{1}{p_i + p_j}\left[Y_i^2 + Y_j^2 + 2(N-1)Y_i Y_j\right].$$

This can also be expressed as a quadratic form:

$$\widehat{\mathrm{Var}}(\hat{T}_L(Y)) \mid_{s(i,j)} = \frac{1}{(p_i + p_j)^2}\begin{pmatrix} Y_i & Y_j \end{pmatrix} A(i,j) \begin{pmatrix} Y_i \\ Y_j \end{pmatrix},$$

where

$$A(i,j) = \begin{pmatrix} 1-(p_i+p_j) & \{1-(p_i+p_j)\} - (p_i+p_j)(N-2) \\ \{1-(p_i+p_j)\} - (p_i+p_j)(N-2) & 1-(p_i+p_j) \end{pmatrix}.$$

Clearly, $\widehat{\mathrm{Var}}(\hat{T}_L(Y)) \mid_{s(i,j)} \geq 0$ provided $A(i,j)$ is positive semi-definite. This is ensured when
(i). $1-(p_i+p_j) \geq 0$, and
(ii). $\det A(i,j) \geq 0$.

Condition (ii) is equivalent to:

$$\frac{1-(p_i+p_j)}{p_i+p_j} \geq \frac{N-2}{2}, \quad \text{for } N > 2. \tag{2}$$

5.1. Numerical illustration

For the example in Section 2 with $N = 4$, and

$$(Y_1, p_1) = (0.5, 0.1), \quad (Y_2, p_2) = (1.2, 0.2), \quad (Y_3, p_3) = (2.1, 0.3), \quad (Y_4, p_4) = (3.2, 0.4),$$

the unbiased estimates of $\mathrm{Var}(\hat{T}_L(Y))$ are shown in Table 4.

Table 4: Unbiased estimates of $\mathrm{Var}(\hat{T}_L(Y))$ for various pairs (i,j)

Sample $s(i,j)$	(1,2)	(1,3)	(1,4)	(2,3)	(2,4)	(3,4)
$p^*(s(i,j))$	$\frac{0.3}{3}$	$\frac{0.4}{3}$	$\frac{0.5}{3}$	$\frac{0.5}{3}$	$\frac{0.6}{3}$	$\frac{0.7}{3}$
$\widehat{\mathrm{Var}}(\hat{T}_L(Y))$	14.478	14.850	14.580	1.620	−4.089	−21.202

Remark 3. The estimates for samples $(2,4)$ and $(3,4)$ are negative since the left-hand side of condition (2) is

$$\frac{1-0.6}{0.6} = \frac{2}{3}, \quad \text{and} \quad \frac{1-0.7}{0.7} = \frac{3}{7},$$

which are both less than the right-hand side value 1 (since $N = 4$ implies $(N-2)/2 = 1$). Hence, matrix $A(i,j)$ is not positive semi-definite in those cases, leading to negative variance estimates.

Acknowledgement

The authors gratefully acknowledge the anonymous reviewer for the insightful comments and constructive suggestions, which significantly improved the clarity and presentation of the manuscript.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Lanke, J. (1975). *Some Contributions to the Theory of Survey Sampling*. Unpublished Doctoral Thesis, Lund University, Sweden.

Pal, M. and Sinha, B. K. (2024). *Selected Topics in Statistical Inference: Theory and Applications*. Springer.

Prabhu-Ajgaonkar, S. (1967). The effect of increasing sample size on the precision of an estimator. *American Statistician*, **21**, 26–28.

Yates, F. and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society. Series B*, **15**, 253–261.