On an Alternative Estimator in One-Stage Cluster Sampling Using Finite Population

Lukman Abiodun Nafiu
Department of Mathematics and Statistics
Federal University of Technology
P. M. B. 65, Minna, Nigeria.

Abstract

This paper investigates the use of a one-stage cluster sampling design in estimating the population total. We focus on a special design where certain number of visits is being considered for estimating the population size and a weighted factor of N_i/n_i^2 is introduced. In this study,we proposed a new estimator and compared it with some of the existing estimators in a one-stage sampling design. Eight (8) data sets were used to justify this paper and computation was done with software developed in Microsoft Visual C++ programming language. For all the populations considered, the bias and variance of our proposed estimator are the least among all estimators compared. All the estimated population totals are also found to fall within the computed confidence intervals for α = 5%. The coefficients of variations obtained for the estimated population totals using both illustrated and life data show that our newly proposed estimator has the least coefficient of variation. Therefore, our newly proposed estimator ($\hat{Y}_{1NPE}$) is recommended when considering a one-stage cluster sampling design.

Keywords: Sampling, cluster, one-stage, design, estimator, bias, variance and finite population.

1. Introduction

In a census, each unit (such as person, household or local government area) is enumerated, whereas in a sample survey, only a sample of units is enumerated and information provided by the sample is used to make estimates relating to all units (Kish, 1967). In one-stage cluster sampling, the estimate varies due to different samples of primary units yielding different estimates. Cochran (1977) opines that subsampling has a great variety of applications. Fink (2002) compares one-stage cluster sampling with simple random sampling and observes that one-stage cluster sampling is better in terms of efficiency. Kalton (1983) gives the reason for one-stage sampling as administrative convenience. Okafor (2002) states that one-stage sampling makes fieldwork and supervision relatively easy.

2. Aim and Objectives

The aim of this paper is to propose a new estimator for a one-stage cluster sampling design and objectives to be achieved include:

(i) investigating some of the existing estimators used in one-stage cluster sampling design and compare them in terms of efficiency and administrative convenience.

(ii) developing new estimator that is more efficient and precise than already existing estimators in one-stage cluster sampling design.

(iii) comparing these estimators (conventional and newly proposed) using eight (8) data sets.

3. Data Used

There are eight (8) categories of data used in this paper. The first four (4) data sets were obtained from Horvitz and Thompson (1952), Raj (1972), Cochran (1977) and Okafor (2002) respectively. The second four (4) data sets used are of secondary type and were collected from Niger State Ministry of Health, Minna, Niger state, Nigeria (2007) and National Bureau of Statistics (2007). We constructed a sampling frame from all diabetic patients with chronic eye disease (Glaucoma and Retinopathy) in the twenty-five (25) Local Government Areas of the state between years 2005 and 2008 as found in Nafiu (2012).

4. Methods and Materials

4.1 Proposed One-Stage Cluster Sampling Scheme

Let Y denote the population value of a variable of interest and y denote the sample value of individuals involved in the variable of interest (number of diabetic patients). Given n_i number of sample value of individuals for the visits made within N_i number of population value of a variable of interest for all days availabe, in line with Thompson (1992), we now propose our estimator in one-stage cluster sampling as:

$$\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i \qquad (1)$$

and

$$\bar{\bar{y}} = \sum_{i=1}^{n} \sum_{j=1}^{n_i} y_{ij}/n_i \,.\, N_i/n_i \qquad (2)$$

Therefore,

$$\hat{Y}_{1NPE} = \frac{N}{n}\sum_{i=1}^{n} \sum_{j=1}^{n_i} y_{ij}/n_i \,.\, N_i/n_i \qquad (3)$$

Equation (3) can be written as;

$$\hat{Y}_{1NPE} = \frac{1}{\gamma}\sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{N_i}{n_i^2} y_{ij} \qquad (4)$$

where $\gamma = \frac{n}{N}$ is the known sampling fraction and y_{ij} denotes the number of individuals in the sample.

Theorem 1: $\hat{Y}_{1NPE}$ is unbiased for the population total Y
Proof:
We show that

$$\begin{aligned}
E\left(\hat{Y}_{1NPE}\right) &= E\left(\frac{1}{\gamma}\sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{N_i}{n_i^2} y_{ij}\right) \\
&= E\left(\frac{N}{n}\sum_{i=1}^{n} \sum_{j=1}^{n_i} (N_i/n_i^2) y_{ij}\right) \\
&= E\left(\frac{N}{n}\sum_{i=1}^{n} (N_i/n_i^2) \sum_{j=1}^{n_i} y_{ij}\right) \\
&= NE\left(\sum_{i=1}^{n} (N_i/n_i)\, \bar{y}_i\right) \\
&= Y \qquad (5)
\end{aligned}$$

This shows that $\hat{Y}_{1NPE}$ is unbiased.
Hence;

$$\begin{aligned}
V\left(\hat{Y}_{1NPE}\right) &= E\{\hat{Y}_{1NPE} - E\left(\hat{Y}_{1NPE}\right)\}^2 \\
&= E\{(\hat{Y}_{1NPE} - E\left(\hat{Y}_{1NPE}\right))(\hat{Y}_{1NPE} - E\left(\hat{Y}_{1NPE}\right))\} \\
&= E\left(\hat{Y}_{1NPE}^2\right) - \{E\left(\hat{Y}_{1NPE}\right)\}^2 \\
&= E\{\left(\frac{1}{\gamma}\sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{N_i}{n_i^2} y_{ij})^2\right\} - \{E\left(\frac{1}{\gamma}\sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{N_i}{n_i^2} y_{ij}\right)\}^2 \\
&= \frac{1}{\gamma^2}\left(\sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(\frac{N_i}{n_i^2})^2 - \frac{N_i}{n_i^2}\right) \left(E\left(y_{ij}^2\right) - \left(E\left(y_{ij}\right)\right)^2\right) \\
&= \frac{1}{\gamma^2}\sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(\frac{N_i^2}{n_i^4} - \frac{N_i}{n_i^2}\right) V\left(y_{ij}\right) \\
&= \frac{1}{\gamma^2}\sum_{i=1}^{n} \left(\frac{N_i^2}{n_i^4} - \frac{N_i}{n_i^2}\right) \sigma_i^2
\end{aligned}$$

$$V\left(\hat{Y}_{1NPE}\right) = \frac{1}{\gamma^2}\sum_{i=1}^{n} \left(\frac{N_i^2}{n_i^4} - \frac{N_i}{n_i^2}\right) \sigma_i^2 \qquad (6)$$

Hence, an unbiased estimator of $V\left(\hat{Y}_{1NPE}\right)$ is:

$$\hat{V}\left(\hat{Y}_{1NPE}\right) = \frac{N^2}{n^2}\sum_{i=1}^{n} \left(\frac{N_i^2}{n_i^4} - \frac{N_i}{n_i^2}\right) s_i^2$$

$$\hat{V}\left(\hat{Y}_{1NPE}\right) = \frac{N^2}{n^2}\sum_{i=1}^{n} \frac{N_i(N_i - n_i^2)}{n_i^4}) s_i^2 \qquad (7)$$

where $s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i} (y_{ij} - \hat{Y}_{1NPE})^2$

Theorem 2: $\hat{V}(\hat{Y}_{1NPE})$ is unbiased for $V(\hat{Y}_{1NPE})$
Proof:
We note that

$$E\{\hat{V}(\hat{Y}_{1NPE})\} = E\{\frac{1}{\gamma^2}\sum_{i=1}^{n}((\frac{N_i}{n_i^2})^2 - \frac{N_i}{n_i^2})s_i^2\}$$

$$= \frac{1}{\gamma^2}E\{\sum_{i=1}^{n}(\frac{N_i}{n_i^2})^2 - \frac{N_i}{n_i^2})s_i^2\}$$

$$= \frac{1}{\gamma^2}[E\{\sum_{i=1}^{n}(\frac{N_i}{n_i^2})^2\, s_i^2\} - E\{\sum_{i=1}^{n}\frac{N_i}{n_i^2}s_i^2\}]$$

$$= \frac{N^2}{n^2}[E\{\sum_{i=1}^{n}(\frac{N_i}{n_i^2})^2\, s_i^2\} - E\{\sum_{i=1}^{n}\frac{N_i}{n_i^2}s_i^2\}]$$

$$= E\{\frac{N^2}{n^2}\sum_{i=1}^{n}(\frac{N_i}{n_i^2})^2\, s_i^2\} - E\{\frac{N^2}{n^2}\sum_{i=1}^{n}\frac{N_i}{n_i^2}s_i^2\}$$

$$= E(\hat{Y}_{1NPE}^2) - (E(\hat{Y}_{1NPE}))^2 + (E(\hat{Y}_{1NPE}))^2 - (E(\hat{Y}_{1NPE}))^2$$

$$= E(\hat{Y}_{1NPE} - E(\hat{Y}_{1NPE}))^2$$

$$= V(\hat{Y}_{1NPE})$$

That is;

$$E\{\hat{V}(\hat{Y}_{1NPE})\} = V(\hat{Y}_{1NPE}) \qquad (8)$$

Hence, $\hat{V}(\hat{Y}_{1NPE})$ is an unbiased sample estimator of the proposed estimator ($\hat{Y}_{1NPE}$) in one-stage cluster sampling design.

5. Results

The estimated population totals computed with the aid of software developed (Microsoft Visual C^{++}) are given in table 1 for the illustrated data and in table 2 for the life data. The biases are presented in tables 3 and 4 for illustrated data and life data respectively. The estimated variances computed using the software developed are given in table 5 for illustrated data and in table 6 for life data respectively. Tables 7 and 8 give the standard errors of the estimated population totals using a one-stage cluster sampling design for illustrated data and life data respectively. The values for confidence intervals for estimated population totals are presented in table 9 for illustrated data and in table 10 for life data. Coefficient of Variations for the estimated population totals using one-stage sampling schemes are given in table 11 for illustrated data and in table 12 for life data.

6. Discussion of Results

Table 3 gives the biases of the estimated population totals for illustrated data for our own estimator as 16, 264, 2 and 143 for cases I – IV respectively while table 4 gives those of the four life data sets as 129, 149, 128 and 122 respectively. This implies that our own estimator has the least biases using both data sets. Table 5 shows the variances obtained using illustrated data for our own estimator as 3052.6168, 30401182.1107, 1.3703 and 148715.0244 for cases I – IV respectively while table 6 shows those of life data sets as 11257.1327, 12008.3612, 12202.6286 and 13101.9827 respectively meaning that our own estimator has the least variances using both data sets.

Table 7 shows the obtained standard errors for the estimated population totals using illustrated data for our own estimator as 55.2505, 5513.7267, 1.1706 and 1219.5061 for cases I – IV respectively while table 8 shows those of life data sets as 106.0996, 109.5827, 110.4655 and 114.4640 respectively meaning that our own estimator has the least standard errors using both data sets.

The confidence intervals of the estimated populations for illustrated data in table 1 are given in table 9 and for life data in table 2 are given in table 10 showing that all the estimated population totals fall within the computed intervals as expected. For our own estimator, table 11 gives the coefficients of variations for the estimated population totals using illustrated data as 13.12%, 5.57%, 3.00% and 8.8% for cases I – IV respectively while table 12 gives those of life data sets as 0.40%, 0.42%, 0.41% and 0.40% respectively which means that our newly proposed one-stage cluster estimator has the least coefficient of variation.

Conclusion

The estimates presented in table 5 for the illustrated data and in table 6 for the life data indicate that substantial reductions in the variances were obtained through the use of newly proposed estimators without forfeiting an unbiased estimate of the sampling variances. We observed from these tables that irrespective of the data considered, the variances of newly proposed estimators are always less than those of already existing estimators in one-stage cluster sampling designs. Tables 7 and 8 which give standard errors in the illustrated data and the life data respectively reveal similar results.

Recommendations

When a complete list of sampling units (frame) from which to draw our sample does not exist and it is uneconomical to obtain information from a sample of elements of the population scattered all over the area, the newly proposed estimator ($\hat{Y}_{1NPE}$) is preferred to the already existing estimators considered in this study. It is therefore recommended to be used in one-stage cluster sampling design.

References

Cochran, W.G. (1977). *Sampling Techniques*. Third Edition. New York: John Wiley and Sons.

Fink, A. (2002). *How To Sample In Surveys.* Thousand Oaks, C.A.: Sage Publications.

Horvitz, D. G. and Thompson, D. J. (1952). "A Generalization of Sampling without Replacement from a Finite Universe". Journal of American Statistical Association. 47: 663-685.

Kalton, G. (1983). *Introduction to Survey Sampling.* Thousand Oaks, C.A.: Sage Publications.

Kish, L. (1967). *Survey Sampling*. New York: John Wiley and Sons.

Nafiu, L. A. (2012). "An Alternative Estimation Method for Multistage Cluster Sampling in Finite Population". Unpulished Ph.D Thesis. University of Ilorin, Nigeria.

National Bureau of Statistics (2007). *Directory of Health Establishments in Nigeria.* Nigeria: Abuja Printing Press.

Niger State (2007). *Niger State Statistical Year Book*. Nigeria: Niger Press Printing & Publishing.

Okafor, F. (2002). *Sample Survey Theory with Applications.* Nigeria: Afro-Orbis Publications.

Raj, D. (1972). *The Design of Sample Surveys.* New York, USA: McGraw-Hill, Inc.

Thompson, S. K. (1992). *Sampling*. New York: John Wiley and Sons.

Table 1: Estimated Population Totals for Illustrated Data

Estimator	Case I	Case II	Case III	Case IV
$\hat{Y}_{HH}$	401	99,113	28	15,097
$\hat{Y}_{HHG}$	434	139,919	25	13,916
$\hat{Y}_{HT}$	387	127,315	35	14,653
$\hat{Y}_{RHC}$	460	109,336	43	15,673
$\hat{Y}_{C}$	425	99,391	33	15,001
$\hat{Y}_{T}$	399	100,637	41	14,849
$\hat{Y}_{O}$	486	135,186	48	14,171
$\hat{Y}_{1NPE}$	421	98,966	39	13,855

Table 2: Estimated Population Totals for Life Data

Estimator	Population 1	Population 2	Population 3	Population 4
$\hat{Y}_{HH}$	28,393	29,105	29,247	29,472
$\hat{Y}_{HHG}$	24,204	26,428	26,551	27,096
$\hat{Y}_{HT}$	25,804	29,002	29,031	29,501
$\hat{Y}_{RHC}$	26,043	27,309	27,609	29,094
$\hat{Y}_{C}$	24,214	28,610	28,791	28,851
$\hat{Y}_{T}$	27,096	27,451	28,142	28,612
$\hat{Y}_{O}$	25,621	27,301	27,451	27,777
$\hat{Y}_{1NPE}$	24,639	25,010	26,551	28,407

Table 3: Biases of Estimated Population Totals for Illustrated Data

Estimator	Case I	Case II	Case III	Case IV
$\hat{Y}_{HH}$	51	724	13	642
$\hat{Y}_{HHG}$	43	533	15	625
$\hat{Y}_{HT}$	48	615	8	654
$\hat{Y}_{RHC}$	46	717	6	673
$\hat{Y}_{C}$	31	675	6	577
$\hat{Y}_{T}$	56	564	12	590
$\hat{Y}_{O}$	61	494	7	685
$\hat{Y}_{1NPE}$	16	264	2	143

Table 4: Biases of Estimated Population Totals for Life Data

Estimator	Population 1	Population 2	Population 3	Population 4
$\hat{Y}_{HH}$	266	202	154	231
$\hat{Y}_{HHG}$	166	205	178	263
$\hat{Y}_{HT}$	208	231	184	196
$\hat{Y}_{RHC}$	217	233	177	236
$\hat{Y}_{C}$	151	175	191	273
$\hat{Y}_{T}$	222	223	143	218
$\hat{Y}_{O}$	187	206	182	250
$\hat{Y}_{1NPE}$	129	149	128	122

Table 5: Variances of the Estimated Population Totals for Illustrated Data

Estimator	Case I	Case II	Case III	Case IV
$\hat{V}(\hat{Y}_{HH})$	6,493.4234	54,200,329.5000	3.7960	1,848,102.4684
$\hat{V}(\hat{Y}_{HHG})$	5,534.8856	48,528,190.1696	3.6303	1,798,200.5201
$\hat{V}(\hat{Y}_{HT})$	4,942.5637	43,617,400.2350	2.0357	1,709,593.1507
$\hat{V}(\hat{Y}_{RHC})$	4,732.7848	36,285,763.3158	2.0026	1,694,104.4845
$\hat{V}(\hat{Y}_{C})$	4,649.8003	35,382,000.6394	2.0004	1,691,831.3946
$\hat{V}(\hat{Y}_{T})$	3,508.7040	32,000,281.9013	1.8656	1,684,551.6747
$\hat{V}(\hat{Y}_{O})$	3,298.4906	31,280,9105494	1.4413	1,662,588.4542
$\hat{V}(\hat{Y}_{1NPE})$	3,052.6168	30,401,182.1107	1.3703	1,487,195.0244

Table 6: Variances of the Estimated Population Totals for Life Data

Estimator	Population 1	Population 2	Population 3	Population 4
$\hat{V}(\hat{Y}_{HH})$	21,561.4967	20, 538.4533	18,963.0372	18,699.3613
$\hat{V}(\hat{Y}_{HHG})$	19,401.0964	20,106.6515	17,456.3615	18,642.0138
$\hat{V}(\hat{Y}_{HT})$	16,567.0363	18,057.2492	14,189.0924	16,438.0924
$\hat{V}(\hat{Y}_{RHC})$	14,425.2555	14,553.3278	14,071.5123	15,303.6736
$\hat{V}(\hat{Y}_{C})$	12,328.5246	14,237.9614	13,504.2438	15,110.4435
$\hat{V}(\hat{Y}_{T})$	12,036.2412	14,125.0305	13,002.6232	14,900.6139
$\hat{V}(\hat{Y}_{O})$	11,812.0825	12,972.2387	12,753.9342	13,688.3221
$\hat{V}(\hat{Y}_{1NPE})$	11,257.1327	12,008.3612	12,202.6286	13,101.9827

Table 7: Standard Errors for Estimated Population Total for Illustrated Data

Estimator	Case I	Case II	Case III	Case IV
$\hat{Y}_{HH}$	80.5818	7,362.0873	1.9483	1,359.4493
$\hat{Y}_{HHG}$	74.3968	6,966.2178	1.9053	1,340.9700
$\hat{Y}_{HT}$	70.3034	6,604.3471	1.4268	1,307.5141
$\hat{Y}_{RHC}$	68.7952	6,023.7665	1.4151	1,301.5777
$\hat{Y}_{C}$	68.1894	5,948.2771	1.4144	1,300.7042
$\hat{Y}_{T}$	59.2343	5,656.8792	1.3659	1,297.9028
$\hat{Y}_{O}$	57.4325	5,592.9340	1.2005	1,289.4140
$\hat{Y}_{1NPE}$	55.2505	5,513.7267	1.1706	1,219.5061

Table 8: Standard Errors for Estimated Population Total for Life Data

Estimator	Population 1	Population 2	Population 3	Population 4
$\hat{Y}_{HH}$	146.8383	143.3124	137.7063	136.7456
$\hat{Y}_{HHG}$	139.2878	141.7979	132.1225	136.5358
$\hat{Y}_{HT}$	128.7130	134.3773	119.1180	128.2111
$\hat{Y}_{RHC}$	120.1052	120.6372	118.6234	123.7080
$\hat{Y}_{C}$	111.0339	119.3229	116.2078	122.9245
$\hat{Y}_{T}$	109.7098	118.8488	114.0290	122.0681
$\hat{Y}_{O}$	108.6834	113.8957	112.9333	116.9971
$\hat{Y}_{1NPE}$	106.0996	109.5827	110.4655	114.4640

Table 9: Confidence Intervals of Estimated Population Totals for Illustrated Data

Estimator	Case I	Case II	Case III	Case IV
$\hat{Y}_{HH}$	(243,559)	(84643,113543)	(24,31)	(12432,17762)
$\hat{Y}_{HHG}$	(288,580)	(126265,153573)	(21,29)	(11288,16544)
$\hat{Y}_{HT}$	(249,525)	(114370,140260)	(32,38)	(9630,19676)
$\hat{Y}_{RHC}$	(325,595)	(97529,121143)	(40,46)	(13122,18224)
$\hat{Y}_{C}$	(291,559)	(87732,111050)	(30,36)	(12452,17550)
$\hat{Y}_{T}$	(282,516)	(94980,106294)	(38,44)	(12305,17393)
$\hat{Y}_{O}$	(373,599)	(124224,146148)	(46,50)	(11644,16698)
$\hat{Y}_{1NPE}$	(313,529)	(88159,109723)	(37,41)	(11465,16245)

Table 10: Confidence Intervals of Estimated Population Totals for Life Data

Estimator	Population 1	Population 2	Population 3	Population 4
$\hat{Y}_{HH}$	(28110,28680)	(28820,29390)	(28840,29370)	(29200,29740)
$\hat{Y}_{HHG}$	(23930,24480)	(26150,26710)	(26170,26690)	(26830,27360)
$\hat{Y}_{HT}$	(25550,26060)	(28740,29270)	(28800,29260)	(29250,29750)
$\hat{Y}_{RHC}$	(25810,26280)	(27070,27550)	(27380,27840)	(28850,27340)
$\hat{Y}_{C}$	(24000,24430)	(28340,28540)	(28560,29020)	(28610,29090)
$\hat{Y}_{T}$	(26880,27310)	(29220,29680)	(27920,28370)	(28370,28850)
$\hat{Y}_{O}$	(25410,25830)	(27070,27520)	(27230,27670)	(27550,28010)
$\hat{Y}_{1NPE}$	(24430,24850)	(24800,25220)	(26330,26770)	(28180,28630)

Table 11: Coefficients of Variation for Illustrated Data

Estimator	Case I	Case II	Case III	Case IV
$\hat{Y}_{HH}$	20.10%	7.43%	6.96%	9.00%
$\hat{Y}_{HHG}$	17.14%	4.98%	7.62%	9.62%
$\hat{Y}_{HT}$	18.17%	5.19%	4.08%	8.92%
$\hat{Y}_{RHC}$	14.96%	5.51%	3.29%	8.30%
$\hat{Y}_{C}$	16.04%	5.98%	4.29%	8.67%
$\hat{Y}_{T}$	14.85%	5.62%	3.33%	8.74%
$\hat{Y}_{O}$	11.82%	4.14%	2.50%	9.10%
$\hat{Y}_{1NPE}$	13.12%	5.57%	3.00%	8.80%

Table 12: Coefficients of Variation for Life Data

Estimator	Population 1	Population 2	Population 3	Population 4
$\hat{Y}_{HH}$	0.52%	0.49%	0.47%	0.46%
$\hat{Y}_{HHG}$	0.58%	0.54%	0.50%	0.50%
$\hat{Y}_{HT}$	0.50%	0.48%	0.41%	0.43%
$\hat{Y}_{RHC}$	0.46%	0.44%	0.43%	0.43%
$\hat{Y}_{C}$	0.46%	0.43%	0.41%	0.43%
$\hat{Y}_{T}$	0.43%	0.43%	0.41%	0.43%
$\hat{Y}_{O}$	0.43%	0.44%	0.42%	0.42%
$\hat{Y}_{1NPE}$	0.40%	0.42%	0.41%	0.40%