On the Coverage Properties of the Ratio Based Estimator in Presence of Non Response Error

Charles Wanyingi Nderitu, Herbert Imboga, Samuel Mwangi Gathuka

Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email address:
nderitucharles90@gmail.com (C. W. Nderitu), imbogaherbert@jkuat.ac.ke (H. Imboga), sammgathuka@gmail.com (S. M. Gathuka)

Abstract: Sample surveys are taken with the assumption that all the sampled elements will respond. However, this is not always the case. Sometimes missing values occur in the survey data due to some reasons. In cases of such missing values, any inference from the data will survey from a non-response error. Therefore, the researcher needed to put all measures in place to prevent the occurrence of the missing values in the data. However, this is not easily achieved. The non-response may occur even after all measures to prevent it have been put in place. Therefore, there is a need to correct the error if it so happens. The current paper seeks to improve the Hansel and Hurwitz (1946) estimator using poststratification. The proposed estimator can be as well be improved. Therefore, the current study proposes an improvement of the Hansel and Hurwitz (1946) estimator using the median of the auxiliary variable. The efficiency of the new proposed estimator is checked using the confidence interval length. Which is the on-coverage property of the estimator. On to the recommendation a band with that will reduce the variance in case of high non-response rate is thus suggested for further studies. Beside we suggest further studies on how both variances and bias will be minimized without any of them being minimized in expense of the other.

Keywords: Confidence Interval, Variance, No-Response, Mean

1. Introduction

According to Hartley, H. O. (2019) [8] the idea of the sample survey exists under the assumption that the sample is a representation of the study population. Mainly, [14] samples are selected based on the probability schemes where each element is assumed to have an equal likelihood of being selected [4]. However, a non-response error occurs. This diminishes the inferences from such surveys. As mentioned by Nayak, M. S. (2019) [11] non-response error may occur under the following circumstances; When the respondent's response deviates from the concept implied by the researcher, non-response can also result from failure to measure the same element. Besides, [12] adds that the error may arise from human error the researcher mail fails to collect some data, further, due to the unwillingness of the respondent to respond to all or some of the questions in the survey [13].

As a matter of mentioned fact, the current study applies the poststratification on the Hansel and Hurwitz (1946) estimator and compares the new model performance in terms of the coverage properties. According to [9] the closer the coverage rate is to the true population rate the more efficient the estimator is. This means the estimator yields the narrowest confidence interval length. Other researchers who have discuses on coverage properties include; [1-3, 5-7, 10, 16].

The present study considers a study variable Y whose population is stratified into two. The response stratum and the non-response stratum. The study variable Y is assumed to have a population size of N with N_1 being the size of the response stratum and $N_2 = N - N_1$, being the size of the non-response stratum.

Considering a simple random sample without replacement of sample size n. we shall have n_1 being the sample size of the response stratum and $n_2 = n - n_1$ being the sample size of the non-response stratum. In the initial attempt of the non-response post ratification correction, Hansen and Hurwitz suggest a resampling scheme where the second sample of size r is taken from the sample non-response such that, $k = \frac{n_2}{r}$ For $k \geq 1$. suppose $\tilde{y}_r$ and $\tilde{y}_{nr}^*$ be the sample mean for the y character based on the n_1 and r units, respectively.

Hansen and Hurwitz (1946) proposed an ubiased estimator defined by;

$$\tilde{y}^* = w_1 \tilde{y}_r + w_2 \tilde{y}_{nr}^* \tag{1}$$

Where, $w_i = \frac{n_i}{n}$

The variance of the estimator $\tilde{y}^*$ is defined as $V(\tilde{y}^*) = \left(\frac{1}{n} - \frac{1}{N}\right) s_y^2 + \frac{(k-1)N_2}{N} \frac{s_{\{y2\}}^2}{n}$.

Where s_y^2 denotes the population means square for the character y in N_1 response units and the population mean square for the character in the N_2 non-response units is denoted by $s_{\{y2\}}^2$.

2. Proposed Estimator

2.1. Model Formulation

The current study applies to reweight poststratification to the Hansen and Hurwitz estimator in equation (1). The suggested estimator is defined as

$$\tilde{y}_r^* = \tilde{y}^* \left(\frac{\bar{X} + M}{\bar{\tilde{x}} + M}\right) \tag{2}$$

Where m is the median of the auxiliary variable.

In order to derive the asymptotic properties of the estimator the equation (2) was expanding up to the second order approximation under the following assumptions.

$$\tilde{y}^* = \bar{Y}(1 + e_1),$$

$$e_2 = \left(\frac{\bar{\tilde{x}}}{\bar{X}} - 1\right), E(e_1) = E(e_2) = 0, C_x^2 = \frac{S_x^2}{\bar{X}^2} C_y^2 = \frac{S_y^2}{\bar{Y}^2}$$

$$E(e_1^2) = \frac{1-f}{n} \cdot C_y^2 + \frac{w_2(k-1)}{n} C_{\{y2\}}^2$$

Where,

$$E(e_2^2) = \frac{V(\bar{\tilde{x}})}{\bar{X}^2} = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{S_x^2}{\bar{X}^2}$$

$$E(e_1 e_2) = \frac{cov(\tilde{y}^*, \bar{\tilde{x}})}{\bar{X}\bar{Y}} = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{S_{xy}}{\bar{X}\bar{Y}}$$

$$S_{xy} = \frac{1}{N-1} \Sigma_{i=1}^N (x_i - \bar{X})(y_i - \bar{Y})$$

$$S_x = \frac{1}{N-1} \Sigma_{i=1}^N (x_i - \bar{X})^2$$

$$S_y = \frac{1}{N-1} \Sigma_{i=1}^N (y_i - \bar{Y})^2.$$

Thus, expressing the equation (2) in terms of the errors and expanding up to the first-order approximation we get,

$$\tilde{y}_r^* = \bar{Y}\left[1 - \Psi_p e_2 + \left(\Psi_p e_2\right)^2 + e_1 - \Psi_p e_1 e_2\right] \tag{3}$$

2.2. Derivation of the Model's Variance

The variance of the estimator $\hat{\theta}$ concerning the unknown population parameter θ is defined as;

$$Var(\hat{\theta}) = E_\theta\left[\left(\hat{\theta} - \theta\right)^2\right]$$

The variance of the proposed estimator $\tilde{y}_r^*$ is defined as; Given

$$Var(\hat{\theta}) = E_\theta\left[\left(\hat{\theta} - \theta\right)^2\right]$$

The variance of the proposed estimator is thus defined as;

$$Var(\tilde{y}_r^*) = E_\theta\left[\left(\tilde{y}_r^* - \bar{Y}\right)^2\right]$$

Using the equation. we get,

$$Var(\tilde{y}_r^*) = E_\theta\left[\bar{Y}\left[1 - \Psi_p e_2 + \left(\Psi_p e_2\right)^2 + e_1 - \Psi_p e_1 e_2\right] - \bar{Y}\right]^2$$

$$= \bar{Y}^2 E_\theta\left[-\Psi_p e_2 + \left(\Psi_p e_2\right)^2 + e_1 - \Psi_p e_1 e_2\right]^2$$

Squaring up to first-degree approximation and taking the expectation, the variance becomes.

$$\bar{Y}^2 \left(\frac{1}{n} - \frac{1}{N}\right)\left[\left(\Psi_p\right)^2 C_x^2 - 2\Psi_p \rho C_x C_y + C_y^2\right] \tag{4}$$

Based on equation (4) the confidence interval length of the standard error of the estimate is defined as

$$SE = \sqrt{\frac{Var(\tilde{y}_r^*)}{n}}$$

Thus the confidence interval length is defined as

$$CIL = \left\{(\tilde{y}_r^*) + Z_{\left\{1-\frac{\alpha}{2}\right\}} * SE\right\} - \left\{(\tilde{y}_r^*) - Z_{\left\{1-\frac{\alpha}{2}\right\}} * SE\right\}$$

Which is equal to

$$CIL = 2 * \left\{Z_{\left\{1-\frac{\alpha}{2}\right\}} * SE\right\}$$

3. Efficiency Comparison

In this section, we illustrate the variance expression of the proposed estimator together with some existing estimators. The illustration is shown in table 1.

Table 1. *A comparison of the estimators' Variance.*

Estimator	Variance
Proposed model	$\bar{Y}^2 \left(\frac{1}{n} - \frac{1}{N}\right)\left[\left(\Psi_p\right)^2 C_x^2 - 2\Psi_p \rho C_x C_y + C_y^2\right]$
$\hat{\bar{Y}}_r = \frac{\tilde{y}}{\bar{x}}\bar{X} = r * \bar{X}$, Classical ratio	$\bar{Y}^2 \left(\frac{1}{n} - \frac{1}{N}\right)\left[C_x^2 - 2\rho C_x C_y + C_y^2\right]$
$\bar{y}^* = w_1\bar{y}_1 + w_2\bar{y}_{h2}$. Hansen and Hurwitz (1946)	$\bar{Y}^2 \left(\frac{1}{n} - \frac{1}{N}\right)[C_x^2] + \frac{\frac{(k-1)N_2}{Nn} * S_{y2}^2}{\bar{Y}^2}$
$T_R = \bar{y}^* \frac{\bar{X}}{\bar{x}}$ Rao 1986	$\bar{Y}^2 \left(\frac{1}{n} - \frac{1}{N}\right)[C_x^2 - \rho C_x C_y]$
$\bar{y}_{er}^* = \bar{y}^* \exp\left[\frac{\bar{X} - \bar{x}}{\bar{x} + \bar{X}}\right]$ The Singh and Kumar	$\bar{Y}^2 \left(\frac{1}{n} - \frac{1}{N}\right)\left[\frac{C_x^2}{4} - 2\rho C_x C_y + C_y^2\right]$

4. Data Analysis

In this section, we present the estimated confidence interval length from two simulated populations. The estimation of the confidence interval length was done at 95% under various percentages on the non-response error and values of k.

4.1. The Review of the Singh and Kumar Estimator for the Confidence Interval Length

Singh et al [15], proposed a ratio-exponential based estimator of finite population mean in presence of the non-response error. The proposed estimator was found to result in the narrowest confidence interval compared to the existing one, however as the non-response rate increases the confidence

length increases. The study concludes by suggesting other parameters such as median to be tested for their performance as far as confidence interval length is concerned.

4.2. Estimates of the Population Mean

We estimated the population means for the two populations under the (0%, 25%, 50%, and 90%) the non-response. Further, we considered various k values (k=1, 1.5, 2 and 3). Considering the first population, *see Table 2,* all the estimators approximated almost the same value of the mean under 0% non-response rate and k=1. However, as the non-response rate increases and the k value changes the estimates changes as well. Noticeably, the change in the new model was not that pronounced compared to the other existing.

Table 2. *The estimate of the population means for the first population by the non-response rate and k values.*

K	Non-response percentage	Hansel &Hurwitz	New	Rao	Ratio	Sign & Kumar
	0	1.9610	1.9610	1.9682	1.9682	1.9646
1	25	1.9721	1.9610	1.9706	1.9566	1.9705
	50	1.9981	1.9728	1.99762	1.9532	1.9828
	90	2.0045	1.9815	1.9999	1.9632	1.9965
	0	1.9609	1.9609	1.9660	1.9605	1.9635
1.5	25	1.9774	1.9628	1.9740	1.9706	1.9740
	50	1.9877	1.9791	1.9886	1.9825	1.9851
	90	2.0334	2.0100	2.0180	2.0456	2.023
	0	1.9609	1.9609	1.9607	1.9609	1.9607
2	25	1.9684	1.9609	1.9700	1.9591	1.9668
	50	1.9786	1.9652	1.9719	1.9562	1.9752
	90	2.0138	1.9923	2.0286	1.9523	2.0212
	0	1.9609	1.9607	1.9604	1.9609	1.9607
3	25	1.9832	1.9801	1.9949	1.9795	1.9891
	50	1.9895	1.9804	1.9970	1.9899	1.9964
	90	2.0444	2.0347	2.0399	2.0598	2.0387

Considering the second population, a similar observation was exhibited. However, at a severe non-response rate (90%) and k=3 the estimated values of the mean in all the estimators varies greatly and were big compared to other instances.

Table 3. *The estimate of the population means for the second population by the non-response rate and k values.*

k	Non-Response percentage	Hansel &Hurwitz	New	Rao	Ratio	Sign & Kumar
	0	1.9687	1.9686	1.9698	1.9678	1.9689
1	25	1.9983	1.9796	1.9999	1.9985	1.9897
	50	2.0057	2.0003	2.0056	2.0068	2.0053
	90	2.0084	2.0032	2.0076	2.0089	2.0062
	0	1.9686	1.9686	1.9691	1.9601	1.9683
1.5	25	1.9691	1.9679	1.9689	1.9697	1.9680
	50	2.0062	1.9894	2.0041	2.0064	2.0032
	90	2.0328	2.0010	2.0054	2.0364	2.0034
	0	1.9686	1.9681	1.9745	1.9745	1.9716
2	25	1.9987	1.9776	1.9892	2.0158	1.9816
	50	2.0275	2.0079	2.0100	2.0358	2.0086
	90	2.0378	2.0205	2.0332	2.0458	2.0315
	0	1.9686	1.9671	1.9695	1.9655	1.9671
3	25	1.9897	1.9772	1.9798	1.9921	1.9772
	50	2.0034	1.9893	2.0014	2.0039	1.9993
	90	2.0237	2.0015	2.0167	2.0297	2.0124

The results in Table 4 show that the narrowest confidence interval length was from the proposed estimator. It can be furthest noted that the interval length increases with an increase in the non-response rate. Besides the length would be big when k=1.5 and smallest

when k=1. Considering the second population, *see Table 5,* similarly, the proposed estimator yields the narrowest confidence interval length. However, this condition is true except for the severe percentage of the non-response rate.

Table 4. *The 95% confidence interval length for the respective estimates of the population mean for the first population.*

k	Non-response rate	Hansel &Hurwitz	New	Rao	Ratio	Sign&Kumar
	0	0.000542	0.000441	0.000553	0.000615	0.000477
1	25	0.000651	0.000563	0.000614	0.00066	0.000598
	50	0.000747	0.000572	0.000675	0.00087	0.000603
	90	0.000801	0.000615	0.000712	0.000945	0.000642
	0	0.000608	0.000601	0.000602	0.000601	0.000601
1.5	25	0.000667	0.00061	0.00067	0.000737	0.000641
	50	0.000746	0.000642	0.000739	0.000777	0.00067
	90	0.000774	0.000649	0.000745	0.000863	0.000685
	0	0.000624	0.000456	0.000615	0.000701	0.000601
2	25	0.000643	0.000552	0.000637	0.000741	0.000613
	50	0.000803	0.000636	0.000771	0.000853	0.000665
	90	0.00084	0.000685	0.000814	0.000893	0.00072
	0	0.00054	0.000459	0.000517	0.000602	0.000478
3	25	0.000683	0.000602	0.000648	0.000744	0.000632
	50	0.000784	0.000658	0.000771	0.000818	0.000706
	90	0.000812	0.000753	0.000789	0.00089	0.000768

Table 5. *The 95% confidence interval length for the respective estimates of the population mean for the second population.*

	Row Labels	Hansel &Hurwitz	New	Rao	Ratio	Sign&Kumar
	0	0.00042	0.000338	0.000466	0.000466	0.000432
1	25	0.000431	0.000358	0.000486	0.000486	0.000435
	50	0.002413	0.001384	0.00202	0.002772	0.001621
	90	0.002632	0.002126	0.002352	0.002786	0.001792
	0	0.000433	0.000361	0.000486	0.000492	0.000435
1.5	25	0.001347	0.001162	0.00136	0.001367	0.001377
	50	0.002528	0.001391	0.002156	0.002972	0.001678
	90	0.002635	0.001948	0.002384	0.003374	0.001794
	0	0.000414	0.000338	0.000466	0.000608	0.000435
2	25	0.001085	0.000382	0.000803	0.001307	0.001058
	50	0.002398	0.001377	0.001996	0.002738	0.001572
	90	0.002614	0.002126	0.002346	0.003077	0.001754
	0	0.000961	0.000188	0.000917	0.000492	0.000435
3	25	0.001347	0.001162	0.00136	0.001367	0.001444
	50	0.002566	0.001618	0.002366	0.002758	0.00238
	90	0.002642	0.002922	0.002385	0.003371	0.002649

5. Conclusion

It was found that the Hansen and Hurwitz estimator was unbiased, however the estimator suffers from the increased variance. Thus, the confidence interval of the Hansel and Hurwitz estimator was large. Therefore, the current study incorporated the median of the auxiliary information to reduce the variance of the proposed estimator. The results shows that the variance was reduced competitively to the other existing estimators. However, there exist cases under severe non response variance when the proposed estimator does not yield the narrowest confidence interval length.

In conclusion the study recommends a bandwidth that would as well reduce the variance in high non-response rates. Further, the study recommends a further study of the ratio exponential form of the current estimator and be assessed on the coverage properties.

References

[1] Calonico, S. C. (2018). Coverage error optimal confidence intervals for local polynomial regression. *arXiv preprint arXiv: 1808.01398.*

[2] Calonico, S. C. (2018). On the effect of bias estimation on coverage accuracy in nonparametric inference. pp. 113 (522), 767-779.

[3] Calonico, S. C. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs.. *The Econometrics Journal*, 23 (2), 192-210.

[4] Frank, B. &. (2021). Comparison of Variance Estimators for Systematic Environmental Sample Surveys: Considerations for Post-Stratified Estimation. *Forests*, 12 (6), 772.

[5] Gardasevic, J. (2019). Standing Height and its Estimation Utilizing Tibia Length Measurements in Adolescents from Western Region in Kosovo. *International Journal of Morphology,* 37 (1).

[6] Gardasevic, J. M. (2019). RELATIONSHIP BETWEEN TIBIA LENGTH MEASUREMENTS AND STANDING HEIGHT.. *Anthropologie*, (1962-), 57 (3), 263-270.

[7] Goulet-Pelletier, J. C. (2018). A review of effect sizes and their confidence intervals, Part I: The Cohen's family.. *The Quantitative Methods for Psychology*, 14 (4), 242-265.

[8] Hartley, H. O. (2019). A new estimation theory for sample surveys. *Biometrika,* 55 (3), 547-557.

[9] Herbert, I. O. (2018). Incorporation Of The Jackknifing Procedure Into The Three-Stage Cluster Sampling Design In The Estimation Of Finite Population Totals.

[10] Lee, D. S. (2021). Valid t-ratio Inference for IV (No. w29124). *National Bureau of Economic Research.*

[11] Nayak, M. S. (2019). "Strengths and weaknesses of online surveys.". *technology*, 6 (2019): 7.

[12] Nderitu, C. W. (2022). Estimation of Finite Population Mean Using Ratio Estimator Based on Known Median of Auxiliary Variable in the Presence of Non-Response.. *American Journal of Theoretical and Applied Statistics,* 11 (2), 75-82.

[13] Singh, G. N. (2021). Enhanced estimation of the population distribution function in the presence of non-response. *Ain Shams Engineering Journal,* 12 (3), 3109-3119.

[14] Singh, P. S. (2018). Effect of measurement error and non-response on the estimation of the population mean.. *Investigación Operacional*, 39 (1), 108-120.

[15] Singh, R. K. (2009). Estimation of mean in presence of non-response using an exponential estimator. *Infinite Study.*

[16] Yang, L. D. (2020). Estimation of incubation period and serial interval of COVID-19: analysis of 178 cases and 131 transmission chains in Hubei province, China.. *Epidemiology & Infection*, 148.