ISSN:0976-4933
Journal of Progressive Science
Vol. 02, No.01, pp 09-16 (2011)

Estimation of harmonic mean using AUXILIARY INFORMATION under double sampling scheme

Sheela Misra and S.K. Yadav[1]
Department of Statistics, Lucknow University, Lucknow (India)
Email- misra_sheela@yahoo.com

Abstract

Estimation of population harmonic mean, using auxiliary information under double sampling scheme using ratio type estimator. Its bias and mean square error (MSE) are found to the first order of approximation. An optimum subclass of estimators is also obtained and a comparative study with the conventional estimator is made. It has further been shown that estimation of parametric values involved in the optimum subclass does not reduce the efficiency of the proposed estimator. An empirical example showing the increased efficiency of proposed estimator over conventional estimator is also included as an illustration.

Key Words- Harmonic mean, Auxiliary information, Estimator, Bias, Mean square error, order of approximation, Bounds, Finite Population Correction, efficiency.

1. Introduction

The use of auxiliary information can increase the precision of an estimator when study variable is highly correlated with auxiliary variable x. There are certain situations when harmonic mean is more useful like in averaging rates and ratios than other measures of averages. For example, it is the most appropriate average where unit of observation (such as per day, per hour, per unit, per worker etc.) remains the same and the act being performed, such as covering distance, is constant. The estimators for the population harmonic mean under simple random sampling have been proposed by Gupta (2008) using auxiliary variables when mean $\overline{X}$ of auxiliary variable x is known. We have proposed estimators of population harmonic mean of the main variable y under study when $\overline{X}$ is not known under double sampling scheme.

Let there be N unit in the population and let (Y_i , X_i) , i=1,2,….,N be the values of observation for the i^{th} unit of the population according to the study variable y and the auxiliary variable x respectively. Let a simple random sample of size n' from this population is taken without replacement having sample

[1] Dr.Subhash Kumar Yadav, Department of Mathematics and Statistics, **Dr RML Avadh University, Faizabad (India)** Email- drskystats@gmail.com

values(y_i, x_i), i= 1, 2…, n' assuming without any loss of generality that first n' units have been selected in the sample from N units of the population to estimate $\overline{X}$. Further let a simple random sample of size n out of n' is drawn without replacement to estimate $\overline{X}$. We further assume that no value is zero and negative.

The population harmonic mean, the parameter of interest, is given by $H_m = \dfrac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{Y_i}}$ (1.1)

The following estimators h_0, h_1 *and* h_1^* of harmonic mean H_m, assuming deviation about mean to be small as compared with mean, with and without using auxiliary information under simple random sampling were proposed by Gupta. $h_0 = \overline{y}\left(1-\alpha\dfrac{s_y^2}{\overline{y}^2}\right)$ (1.2)

And for negatively correlated variables y and x we may take $h_1 = \overline{y}\left(1-\alpha\dfrac{s_y^2}{\overline{y}^2}\right)\left(\dfrac{\overline{x}}{\overline{X}}\right)$, (1.3)

whereas for positively correlated variables y and x we may take $h_1^* = \overline{y}\left[1-\alpha\dfrac{s_y^2}{\overline{y}^2}\right]\left(\dfrac{\overline{X}}{\overline{x}}\right)$, (1.4)

where $\alpha = \dfrac{n-1}{n}$ or a suitably chosen scalar.

A generalized class of estimators h_c of harmonic mean may be taken $h_c = \overline{y}\left[1-\alpha\dfrac{s_y^2}{\overline{y}^2}\right]\left(\dfrac{\overline{x}}{\overline{X}}\right)^{\lambda}$ (1.5)

where α and λ are the characterizing scalar to be determined by minimizing mean square error $MSE(h_c)$.

All above estimators are for estimating harmonic mean H_m when $\overline{X}$ is known. But in practice it has been found that it is hardly known. For the cases when $\overline{X}$ is not known, double sampling scheme is used. Under this sampling scheme we proposed the following estimators for the harmonic mean of the population characteristics under study using auxiliary information- $h_{1d} = \overline{y}\left(1-\alpha\dfrac{s_y^2}{\overline{y}^2}\right)\left(\dfrac{\overline{x}}{\overline{x}'}\right)$ (1.6)

When y and x are negatively correlated variables. $h_{1d}^* = \overline{y}\left(1-\alpha\dfrac{s_y^2}{\overline{y}^2}\right)\left(\dfrac{\overline{x}'}{\overline{x}}\right)$ (1.7)

When y and x are positively correlated variables $h_{cd} = \bar{y}\left(1-\alpha\frac{s_y^2}{\bar{y}^2}\right)\left(\frac{\bar{x}}{\bar{x}'}\right)^{\lambda}$, (1.8)

where α and λ are the characterizing scalars to be chosen suitably and determined by minimizing mean square error $MSE(h_{cd})$.

Let us define,

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_i,\ \bar{Y}=\frac{1}{N}\sum_{i=1}^{N} Y_i,\ \sigma_Y^2=\frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^2,\ S_Y^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^2,\ C_Y^2=\frac{S_Y^2}{\bar{Y}^2},\ C_X^2=\frac{S_X^2}{\bar{X}^2}$$

$$s_y^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2,\ s_x^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2,\ \mu_{rs}=\frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^r\left(X_i-\bar{X}\right)^s,\ \beta_2=\frac{\mu_{40}}{\mu_{20}^2}$$

2. BIAS OF THE ESTIMATOR h_{1d}

Let, $e_0=\frac{\bar{y}-\bar{Y}}{\bar{Y}} \Rightarrow \bar{y}=\bar{Y}(1+e_0)$, $e_1=\frac{s_y^2-S_y^2}{S_y^2} \Rightarrow s_y^2=S_y^2(1+e_1)$

$$e_2=\frac{\bar{x}-\bar{X}}{\bar{X}} \Rightarrow \bar{x}=\bar{X}(1+e_2),\ e_2'=\frac{\bar{x}'-\bar{X}}{\bar{X}} \Rightarrow \bar{x}'=\bar{X}(1+e_2') \tag{2.1}$$

So that $E(e_0)=E(e_1)=E(e_2)=E(e_2')=0$

Assuming population size N to large be enough in comparison to sample size n, we may ignore finite population correction factor (f.p.c) and get,

$$E\left(e_0^2\right)=\frac{C_Y^2}{n}\qquad E\left(e_1^2\right)=\frac{2}{n-1}\left\{1+\frac{n-1}{2n}(\beta_2-3)\right\},\ E\left(e_2^2\right)=\frac{C_x^2}{n},$$

$$E\left(e_2'^2\right)=\frac{C_x^2}{n'}\ E\left(e_0e_1\right)=\frac{\mu_3}{n\bar{Y}S_y^2},\ E\left(e_2'^2\right)=\frac{C_x^2}{n'},\ E\left(e_0e_2\right)=\frac{C_{yx}}{n},$$

(2.2)

$$E\left(e_0e_2'\right)=\frac{C_{yx}}{n'},\ E\left(e_1e_2\right)=\frac{\mu_{21}}{n\bar{X}S_y^2},\ E\left(e_1e_2'\right)=\frac{\mu_{21}}{n'\bar{X}S_y^2}$$

From equation (1.7) we have $h_{1d} = \bar{y}\left(1-\alpha\frac{s_y^2}{\bar{y}^2}\right)\left(\frac{\bar{x}}{\bar{x}'}\right)$

Substituting values of $\bar{y}, s_y^2, \bar{x}' \, and \, \bar{x}$ from (2.1) and (2.2) in above equation we get

$$h_{1d} = \bar{Y}(1+e_0)\left(1-\alpha\frac{S_y^2}{\bar{Y}^2}\frac{(1+e_1)}{(1+e_0)^2}\right)\left(\frac{\bar{X}(1+e_2)}{\bar{X}'(1+e_2^{'})}\right) \quad (2.3)$$

$$h_{1d} = \left[\bar{Y}(1+e_0) - \alpha\bar{Y}C_y^2(1+e_1)(1+e_0)^{-1}\right](1+e_2)(1+e_2^{'})^{-1}$$

Expanding $(1+e_0)^{-1}$ and put $\alpha\bar{Y}C_y^2 = k$ in (2.3), we get,

$$h_{1d} = \left[\bar{Y}(1+e_0) - k(1+e_1)(1-e_0+e_0^2+..)\right](1+e_2)(1-e_2^{'}+e_2^{'2}+..)$$

$$h_{1d} = \left[\bar{Y}(1+e_0) - k(1+e_1)(1-e_0+e_0^2+..)\right](1+e_2-e_2^{'}-e_2e_2^{'}+e_2^{'2}+..)$$

$$h_{1d} = \bar{Y} - k + \bar{Y}\left(e_0+e_2-e_2^{'}+e_0e_2-e_0e_2^{'}-e_2e_2^{'}+e_2^{'2}+..\right) - k\left(e_1+e_2\right.$$
$$\left.-e_0-e_2^{'}+e_1e_2-e_1e_2^{'}-e_0e_1-e_0e_2-e_0e_2^{'}+e_0^2+...\right)$$

Put $\bar{Y} - k = H$ in above we get

$$h_{1d} - H = \bar{Y}\left(e_0+e_2-e_2^{'}+e_0e_2-e_0e_2^{'}-e_2e_2^{'}+e_2^{'2}+..\right) - k\left(e_1+e_2-e_0\right.$$
$$\left.-e_2^{'}+e_1e_2-e_1e_2^{'}-e_0e_1-e_0e_2-e_0e_2^{'}+e_0^2+...\right) \quad (2.4)$$

Taking expectation both the side we get bias (h_{1d}) up to $O\left(\frac{1}{n}\right)$,

$$Bias(h_{1d}) = E(h_{1d} - H)$$

$$= \bar{Y}E(e_0e_2) - kE(e_1e_2) + kE(e_0e_1) - \bar{Y}E(e_0e_2^{'}) + kE(e_1e_2^{'}) + kE(e_1e_2)$$
$$+kE(e_0e_2) - kE(e_0^2) + \bar{Y}E(e_2^{'2}) - kE(e_0e_2^{'}) + kE(e_0^2)$$

$$= (\bar{Y}+k)E(e_0e_2) - (\bar{Y}+k)E(e_0e_2^{'}) + kE(e_0e_1) - kE(e_1e_2) - kE(e_0^2). \quad (2.5)$$

Now putting $\overline{Y} + k = k_0$ in the equation (2.5) and using equation (2.2), we get,

$$Bias(h_{1d}) = k_0\left(\frac{1}{n} - \frac{1}{n'}\right)C_{yx} + \frac{k\mu_{30}}{n\overline{Y}S_y^2} - k\left(\frac{1}{n} - \frac{1}{n'}\right)\frac{\mu_{21}}{\overline{X}S_y^2} - k\frac{C_y^2}{n}$$

$$= \left(\frac{1}{n} - \frac{1}{n'}\right)\left\{k_0 C_{yx} - k\frac{\mu_{21}}{\overline{X}S_y^2}\right\} + \frac{k}{n}\left\{\frac{\mu_{30}}{\overline{Y}S_y^2} - C_y^2\right\} = 0 \quad (2.6)$$

If $k_0 C_{yx} = k\frac{\mu_{21}}{\overline{X}S_y^2}$ and $\frac{\mu_{30}}{\overline{Y}S_y^2} = C_y^2$, $\Rightarrow \rho = \frac{k\overline{Y}\mu_{21}C_y}{k_0\overline{X}\mu_{30}C_x}$ (2.7)

Under above condition h_{1d} becomes unbiased estimator of H.

3. THE MEAN SQUARE ERROR OF ESTIMATOR h_{1d}

The mean square error of the proposed estimator h_{1d} is given by squaring equation (2.4) and taking expectation up to $O\left(\frac{1}{n}\right)$ we have, $MSE(h_{1d}) = E(h_{1d} - H)^2$ (3.1)

$$= E\left\{\overline{Y}(e_0 + e_2 - e_2') - k(e_1 + e_2 - e_0 - e_2')\right\}^2 = E\left\{(\overline{Y} + k)e_0 - ke_1 - (\overline{Y} - k)e_2' + (\overline{Y} - k)e_2\right\}^2$$

Now putting $\overline{Y} + k = k_0$ and $\overline{Y} - k = H$ in the above equation we have,

$$MSE(h_{1d}) = k_o^2 E(e_0^2) + H^2\left\{E(e_2^2) + E(e_2'^2) - 2E(e_2 e_2')\right\} + k^2 E(e_1^2)$$

$$- 2kk_0 E(e_0 e_1) + 2Hk\left\{E(e_1 e_2') - E(e_1 e_2)\right\} - 2k_0 HE\left\{e_0 e_2' - e_0 e_2\right\}$$

From equation (2.2) we get,

$$MSE(h_{1d}) = k_0^2\frac{C_y^2}{n} + \left(\frac{1}{n} - \frac{1}{n'}\right)H^2 C_x^2 - 2k_0 H\left(\frac{1}{n} - \frac{1}{n'}\right)C_{yx} + k^2\frac{2}{n-1}\left\{1 + \frac{n-1}{n}(\beta_2 - 3)\right\}$$

$$- \frac{2k}{S_y^2}\left[\frac{k_0\mu_{30}}{n\overline{Y}} - \left(\frac{1}{n} - \frac{1}{n'}\right)\frac{H\mu_{21}}{\overline{X}}\right] \quad (3.2)$$

4.UPPER BOUND FOR MSE h_{1d}

Form thesis, we have $k \in [0, \overline{Y}]$ and $k_0 \in [\overline{Y}, 2\overline{Y}]$. Also since it is well know that $0 \le H.M \le A.M$ therefore $H \in [0, \overline{Y}]$, Substituting these limit of k, k_0 and H we get from equation (3.2)

$$MSE(h_{1d}) \le 4\overline{Y}^2 \frac{C_y^2}{n} + (\frac{1}{n} - \frac{1}{n'})\overline{Y}^2 C_x^2 + 4\overline{Y}^2 (\frac{1}{n} - \frac{1}{n'})C_{yx}$$

$$+ \overline{Y}^2 \frac{2}{n-1}\left\{1 + \frac{n-1}{n}(\beta_2 - 3)\right\} - \frac{2\overline{Y}^2}{S_y^2}\left[\frac{2\mu_{30}}{n\overline{Y}} + (\frac{1}{n} - \frac{1}{n'})\frac{\mu_{21}}{\overline{X}}\right] \quad (4.1)$$

For positively skewed distribution $\mu_{30} > 0$, $\rho > 0$ and if $\mu_{21} < 0$ we have,

$$MSE(h_{1d}) \le \overline{Y}^2 \left[4\frac{C_y^2}{n} + (\frac{1}{n} - \frac{1}{n'})C_x^2 + 4(\frac{1}{n} - \frac{1}{n'})C_{yx} + \frac{2}{n-1}\left\{1 + \frac{n-1}{2n}(\beta_2 - 3)\right\}\right] \quad (4.2)$$

If distribution is mesokurtic, $\beta_2=3$, then

$$MSE(h_{1d}) \le \overline{Y}^2 \left[4\frac{C_y^2}{n} + (\frac{1}{n} - \frac{1}{n'})\{C_x^2 + 4C_{yx}\} + \frac{2}{n-1}\right] \quad (4.3)$$

Further we may obtain from equation, $k \le \overline{Y} \Rightarrow \overline{Y}\alpha C_y^2 \le \overline{Y} \Rightarrow C_y^2 \le \frac{1}{\alpha} = \frac{n}{n-1}$

$$MSE(h_{1d}) \le \overline{Y}^2 \left[\frac{4}{n-1} + (\frac{1}{n} - \frac{1}{n'})\{C_x^2 + 4C_{yx}\} + \frac{2}{n-1}\right]$$

$$\le \overline{Y}^2 \left[\frac{6}{n-1} + (\frac{1}{n} - \frac{1}{n'})\{C_x^2 + 4C_{yx}\}\right], \quad (4.4)$$

C_x^2, being a stable quantity, may be known from previous experience, pilot survey or literature and if $\rho_{yx} > 0$ hence may be taken as constant say C_0 then, $MSE(h_{1d}) \le \overline{Y}^2 \left[\frac{6}{n-1} + C_0\right]$, (4.5)

which is the upper bound of $MSE(g_{1d})$.

EMPIRICAL EXAMPLE

For demonstration purpose let us consider following example where $\mu_{30} > 0$ & $\rho > 0$ i.e., x and y are positively correlated.

From the data dealing with weight (y) in kg and height (x) in c.m in a study of N=277 children between age group of 3 to 36 months, the required value of population parameter are calculated. Further to study the property of proposed estimator, two random sample of size 30 each, was taken and required sample values calculated.

$\overline{Y} = 6.587726$, $\overline{X} = 68.23105$, $S_y^2 = 6.691442$, $S_x^2 = 156.989902$, $S_{yx} = 26.029657$,

$C_y^2 = 0.154187$, $C_x^2 = 0.033721$, $C_{yx} = 0.0579096$, $\mu_{30} = 8.28540357$, $\mu_{21} = 26.7871327$,

$\mu_{40} = 124.976079$, $\mu_{20} = 6.667286$, $\beta_2 = 2.811439303$, $H_m = 5.547495256$,

For the sample of size n'= 100

$\bar{x} = 67.18$, $\bar{y} = 6.412$, $s_y^2 = 5.9616727$, $s_x^2 = 144.02788$, $s_{yx} = 23.24584$,

$c_y^2 = 0.1450176$, $c_x^2 = 0.0319129$, $\hat{\mu}_{30} = 5.418344$, $\hat{\mu}_{21} = 30.4639$, $\hat{\mu}_{40} = 96.43139$,

$\hat{\beta}_2 = 2.76829082$, $\hat{\mu}_{20} = 5.902056$

And for n= 30

$\bar{x} = 66.7$, $\bar{y} = 6.293333$, $s_y^2 = 5.3634034$, $s_x^2 = 115.66552$, $s_{yx} = 17.498$ $c_y^2 = 0.1354188$,

$c_x^2 = 0.0259987$, $\hat{\mu}_{30} = 1.479231$, $\hat{\mu}_{21} = 19.59672$, $\hat{\mu}_{40} = 58.70783$ $\hat{\beta}_2 = 2.1840459$,

$\hat{\mu}_{20} = 5.184622$, $h_0 = 5.5913371, h_1 = 5.88133794278$, $h_{1d} = 5.9237441805$

$MSE(h_0) = 0.255740$, $MSE(h_1) = 0.1487723$ and $MSE(h_{1d}) = 0.1508628$.

$M\hat{S}E(h_0) = 0.22140725566$, $M\hat{S}E(h_1) = 0.142199960$ and $M\hat{S}E(h_{1d}) = 0.16834540922$.

Relative efficiency of h_{1d} over h_0 as follows

$$Efficiency = \frac{MSE(h_0)}{MSE(h_{1d})} \times 100 = 169.51846\%$$

Relative efficiency of h_{1d} over h_0 based on estimated MSE of h_0 and h_{1d} is

$$Efficiency = \frac{M\hat{S}E(h_0)}{M\hat{S}E(h_{1d})} \times 100 = 131.51963\%$$

Results and Conclusion

It is clear that values of the estimators $h_0 = 5.5913371$, $h_1 = 5.88133794278$ $h_{1d} =$ 5.9237441805 are very close to exact and approximated population harmonic means $H_m = 5.547495256$, and $H = 5.547595414$ respectively having very small mean square errors given by $MSE(h_0) = 0.2557403$, $MSE(h_1) = 0.1487723$ & $MSE(h_{1d}) = 0.1508628$ and estimated mean square errors given by $M\hat{S}E(h_0) = 0.221407256$, $M\hat{S}E(h_1) = 0.14219996$ & $M\hat{S}E(h_{1d}) = 0.16834541$. Relative and estimated relative efficiency of h_{1d} which utilizes auxiliary information under double sampling scheme is 169.51846% and 131.51963% respectively over estimator h_0 which does not utilizes auxiliary information in the example under consideration. Thus we see that if the proper information on auxiliary variable is not available i.e, $\overline{X}$ is not known, double sampling technique is used to estimate it. We observe here that the MSE of the estimator h_{1d} of population Harmonic mean under double sampling scheme is not much increased by the MSE of the estimator h_1 under simple random sampling. Empirically it is only reduced by 0.0020905.So we can say it is very close to the estimator under simple random sampling where information on auxiliary variable is completely available.

References

1. Dalabehera, M. and Sahoo, L.N. (2000). An unbiased estimator in two-phase sampling using two auxiliary variables. *Journal of Indian Society of Agril. Statistics*, 53(2):134-140.
2. Gupta, R.K. (2008). On estimation of finite population mean using auxiliary variables. Ph. D. thesis submitted and accepted by Lucknow University.
3. Misra, S. (1992). On estimation in sampling theory. Ph. D. thesis submitted and accepted by Lucknow University.
4. Misra, S., Maneesha and Singh R. Karan (2004). Double Sampling Ratio Estimator in the presence Measurement Error (2004), *Lucknow Journal of Science*, 1(2): 21-27.
5. Misra S. and Singh R. Karan (2005): A Class of Estimators for Estimating Population Mean Using Auxiliary Information on Moments about Zero, *Lucknow Journal of Science*, 2(1): 29-39.
6. Sahoo, J. and Sahoo, L.N. (1993): A class of estimators in two-phase sampling using two auxiliary variables. *Journal of Indian Statistical Association*, 31:107-114.
7. Sahoo, L.N., Senapati, S.C. and Singh, G.N. (2005) An alternative class of estimators in two-stage sampling with two auxiliary variables. *Journal of the Indian Statistical Association*, 43 (2):147-156.
8. Shabbir, J. and Gupta. S. (2007). On estimating the finite population mean with known population proportion of an auxiliary variable. *Pak. J. Statist*. 23(1):1-9.
9. Singh, G. (1984). On use of auxiliary information in sampling theory. Ph.D. thesis submitted and accepted by Lucknow University.
10. Singh, R. Karan (1982). Generalized double sampling estimators for the ratio and product of population parameters. *Journal of the Indian Statistical Association*, 20: 39-49.
11. Sukhatme, B.V. (1962). Some ratio-type estimators in two-phase sampling. *Journal of the Indian Statistical Association,* 57:628-632.

Received on 05.09.2010 Accepted on 27.12.2010