Generalized Estimators of Population Median using Auxiliary Information

H.S. Jhajj[1] and H. K. Bhangu[2]

Abstract

For estimating the median of the population, we have proposed two estimators using linear transformations using the information on median of the auxiliary variable. The expressions for biases, mean square errors and their minimum values have been obtained. It has been shown that proposed estimators are always efficient than the ratio estimator and equally efficient to the other estimators derived from a different approach respectively defined by Kuk and Mak (1989). The comparison of estimators among the proposed estimators with respect to their biases has also been done.The results have been illustrated by carrying out the simulation study.

Keywords: Median estimation, Auxiliary variable, Mean squared errors, Bias, Simple random sampling, Population Median, Sample Median.

1. Introduction

In survey sampling, statisticians have given more attention to the estimation of population mean, total, variance etc. but median is regarded as a more appropriate measure of location than mean when the distribution of variables such as income, expenditure etc is highly skewed. In such situations, it is necessary to estimate median. First of all some statisticians such as Gross(1980), Sedransk and Meyer(1978), Smith and Sedransk(1983) have considered the problem of estimating the median by dealing exclusively with variable under study Y only.

1. Prof. & Head, Department of Statistics, Punjabi University, Patiala-147002, India. Email: drhsjhajj@yahoo.co.in

2. Department of Community Medicine, SGRDIMSAR, Amritsar-143501, India. Email: harpreet3182@gmail.com

Kuk and Mak (1989) are the first to introduce the estimation of median of study variable Y by using information of the values on the auxiliary variable X highly correlated with Y for the units in the sample and its known median M_X for the whole population. Later problem of estimation of median was discussed by various authors such as Chambers and Dunstan(1986), Rao et al.(1990), Mak and Kuk(1993), Rueda et al.(2001),Arcos et al.(2005), Garcia and Cebrian(2001), Meeden(1995), and Singh, S. et al(2007).

Using known value of population median M_X of the auxiliary variable X, Kuk and Mak (1989) suggested an estimator for the population median M_Y of study variable Y under simple random sampling similar to ratio estimator of its population mean as

$$\hat{M}_{YR} = \hat{M}_Y \frac{M_X}{\hat{M}_X} \qquad (1.1)$$

where, $\hat{M}_Y$ and $\hat{M}_X$ are the estimators of M_Y and M_X respectively based on a simple random sample of size n drawn from the population.

Let Y_i and X_i denote the values on the i^{th} unit of the population i = 1, 2, 3, …, N for the study variable Y and auxiliary variable X respectively and corresponding small letters denote the values in the sample.

Suppose that $Y_{(1)}, Y_{(2)},\ldots,Y_{(n)}$ are the values of Y on the sample units in ascending order. Further, let t be an integer such that $Y_{(t)} \le M_Y \le Y_{(t+1)}$ and let p = t/n be the proportion of Y values in the sample that are less than or equal to the median value M_Y, an unknown population parameter. If $\hat{p}$ is a predictor of p, the sample median $\hat{M}_Y$ can be written in terms of quantiles as $\hat{Q}_Y(\hat{p})$, where $\hat{p} = 0.5$.

Kuk and Mak (1989) define a matrix of proportions (P_{ij}, (i,j=1,2)) of units in the population as

	$X \leq M_X$	$X > M_X$	Total
$Y \leq M_Y$	P_{11}	P_{12}	$P_{1.}$
$Y > M_Y$	P_{21}	P_{22}	$P_{2.}$
Total	$P_{.1}$	$P_{.2}$	1

Where for instance, P_{11} denotes the proportion of the units in the population with $Y \leq M_Y$ and $X \leq M_X$. In practice, the P_{ij} are usually unknown but can be estimated by p_{ij} based on a similar cross-classification of the sample. Thus, p_{11}, for instance, represents the proportion of units in the sample with $Y \leq M_Y$ and $X \leq M_X$. For estimating the population median M_Y of study variable Y, Kuk and Mak(1989) has also proposed two other estimators, position estimator $\widehat{M}_{YP}$ and stratification estimator $\widehat{M}_{YS}$ respectively derived from a different approach.

$$\widehat{M}_{YP} = \hat{Q}_Y(\hat{p}_1) \tag{1.2}$$

where, $\hat{p}_1 = 2/n\{n_X p_{11} + (n - n_X)(\frac{1}{2} - p_{11})\}$

where, n_X be the number of units in the sample with $X \leq M_X$.

$$\widehat{M}_{YS} = \inf\{y: \tilde{F}_Y(y) > 1/2\} \tag{1.3}$$

where, $\tilde{F}_Y(y) \cong \frac{1}{2}\{\tilde{F}_{Y1}(y) + \tilde{F}_{Y2}(y)\}$ and for any value of y, let $\tilde{F}_{Y1}(y)$ be the proportion among those units in the sample with $X \leq M_X$ that have Y values less than or equal to y.

Similarly, $\tilde{F}_{Y2}(y)$ is the proportion among those with X > M_X.

Defining

$$e_0 = \frac{\hat{M}_Y}{M_Y} - 1 , \qquad e_1 = \frac{\hat{M}_X}{M_X} - 1$$

such that E(e_k) ≅ 0 and | e_k | < 1 for k = 0,1

Using results of Kuk and Mak (1989) up to the first order of approximation, we have

$$E(e_0^2) = (1-f)(4n)^{-1}\,[M_Y f_Y(M_Y)]^{-2} \tag{1.4}$$

$$E(e_1^2) = (1-f)(4n)^{-1}[M_X f_X(M_X)]^{-2} \tag{1.5}$$

$$E(e_0 e_1)=(1-f)(4n)^{-1}[4P_{11}(X,Y)-1][M_X M_Y f_X(M_X) f_Y(M_Y)]^{-1} \tag{1.6}$$

where it is being assumed that as N→∞, the distribution of the bivariate variable (X,Y) approaches to a continuous distribution with marginal densities $f_X(x)$ and $f_Y(y)$ for X and Y respectively. This assumption holds in particular under a superpopulation model framework, treating the values of (X,Y) in the population as a realization of N independent observations from a continuous distribution. We also assume that $f_X(x)$ and $f_Y(y)$ are positive.

2. The Proposed Estimators and their results

When the median M_X of the auxiliary variable X is known, we propose following estimators of population median using linear transformation under the simple random sampling design as

$$\hat{M}_{H1} = \frac{\hat{M}_Y}{\hat{M}_X}\left[\hat{M}_X + \alpha\left(M_X - \hat{M}_X\right)\right] \tag{2.1}$$

$$\hat{M}_{H2} = \frac{\hat{M}_Y}{M_X}\left[M_X + v\left(\hat{M}_X' - M_X\right)\right] \tag{2.2}$$

where $\hat{M}_X' = \frac{NM_X - n\hat{M}_X}{N-n}$

where $0 \leq \alpha \leq 1$, $0 \leq v \leq 1$

Assuming that sample size is large enough such that terms involving e_i's more than second degree are negligible in the expansions of estimators $\widehat{M}_{H1} and\ \widehat{M}_{H2}$ in terms of e_i's while obtaining biases and mean squared errors.

Using results (1.2) – (1.4), the biases and MSE's of $\widehat{M}_{H1} and\ \widehat{M}_{H2}$, up to first order of approximation are

Bias $(\widehat{M}_{H1}) = M_Y\ \alpha(1-f)(4n)^{-1}[\ \{M_X f_X(M_X)\}^{-2}$

$$-\{4P_{11}(X,Y)-1\}\{M_X M_Y f_X(M_X) f_Y(M_Y)\}^{-1}\ ] \quad (2.3)$$

Bias $(\widehat{M}_{H2}) = -M_Y\ \frac{nv}{(N-n)}(1-f)(4n)^{-1}\{4P_{11}(X,Y)-1\}\{M_X M_Y f_X(M_X) f_Y(M_Y)\}^{-1}$ (2.4)

MSE $(\widehat{M}_{H1}) = (1-f)(4n)^{-1}[\{\ f_Y(M_Y)\}^{-2} + \alpha\left(\frac{M_Y}{M_X}\right)^2\ \{f_X(M_X)\}^{-2}(\alpha - 2C)\ ]$ (2.5)

where C = $\frac{[4P_{11}(X,Y)-1]M_X f_X(M_X)}{M_Y f_Y(M_Y)}$ (2.6)

From (2.5), we note that MSE of $\widehat{M}_{H1}$decreases with the decrease in the value of α provided

C $\leq \alpha/2$.

Similarly, up to the first order of approximation, we get

MSE$(\widehat{M}_{H2}) = (1-f)(4n)^{-1}[\ \{f_Y(M_Y)\}^{-2} + \theta\left(\frac{M_Y}{M_X}\right)^2\ \{f_X(M_X)\}^{-2} v(\theta v - 2C)\ ]$ (2.7)

where $\theta = \frac{n}{N-n}$

In (2.7), we note note that MSE of $\widehat{M}_{H2}$ decreases with decrease in the value of v provided

C $\leq \theta v/2$.

MSE ($\widehat{M}_{H1}$) minimizes for

$$\alpha = \frac{[4P_{11}(X,Y)-1]M_X f_X(M_X)}{M_Y f_Y(M_Y)} = C \tag{2.8}$$

and its minimum value is given by

$$MSE_{min}\ (\widehat{M}_{H1}) = (1-f)(4n)^{-1}\{f_Y(M_Y)\}^{-2}\{1-(4P_{11}(X,Y)-1)^2\} \tag{2.9}$$

Bias of optimum estimator $\widehat{M}_{H1}$ is given by

$$\text{Bias}\ (\widehat{M}_{H1}) = (1-f)(4n)^{-1}\{4P_{11}(X,Y)-1\}[\{M_X f_X(M_X) f_Y(M_Y)\}^{-1} - \{4P_{11}(X,Y)-1\}M_Y{}^{-1}\{f_Y(M_Y)\}^{-2}] \tag{2.10}$$

Similarly, MSE($\widehat{M}_{H2}$) minimizes for

$$v = \frac{(N-n)[4P_{11}(X,Y)-1]M_X f_X(M_X)}{nM_Y f_Y(M_Y)} = \frac{(N-n)}{n}C \tag{2.11}$$

and its minimum value is given by

$$MSE_{min}\ (\widehat{M}_{H2}) = (1-f)(4n)^{-1}\{f_Y(M_Y)\}^{-2}[1-\{4P_{11}(X,Y)-1\}^2] = MSE_{min}\ (\widehat{M}_{H1}) \tag{2.12}$$

and bias of optimum estimator of $\widehat{M}_{H2}$ is given by

$$\text{Bias}\ (\widehat{M}_{H2}) = -(1-f)(4n)^{-1}\{4P_{11}(X,Y)-1\}^2 M_Y{}^{-1}\{f_Y(M_Y)\}^{-2} \tag{2.13}$$

Using the expressions (2.10) and (2.13), we have

$$\frac{|\text{Bias}\ (\widehat{M}_{H1})|}{|\text{Bias}(\widehat{M}_{H2})|} = \left|\frac{M_Y f_Y(M_Y)}{[4P_{11}(X,Y)-1]M_X f_X(M_X)} - 1\right| \tag{2.14}$$

Expression (2.14) shows that bias of $(\widehat{M}_{H2})$ is smaller than the bias of $(\widehat{M}_{H1})$

$$\text{if} \qquad \rho_c < \frac{1}{2}\frac{M_Y f_Y(M_Y)}{M_X f_X(M_X)} \tag{2.15}$$

where $\rho_c = [4P_{11}(X,Y) - 1]$,the correlation coefficient between the variables X and Y,goes from -1 to 1 as $P_{11}(X,Y)$ increases from 0 to 1/2 which implies that $\rho_c$is negative and positive for $P_{11}(X,Y)$ belongs to $[0 \ \frac{1}{4})$ and $(\frac{1}{4} \ \frac{1}{2}]$ respectively.

Note : We have seen that value of C remains fairly stable in repeated survey. So the value of C may often be more or less known on the basis of previous data, past experience, a pilot survey or otherwise, more information about the range of possible values of C may be available in practical situations.

Using the additional knowledge of C in addition to known value of population median M_X of auxiliary variable X, we can construct from (2.1) and (2.2) efficient estimators of population median M_Y of study variate Y.

3. Comparison

To compare the proposed estimators with $\widehat{M}_{YR}$ given by Kuk and Mak(1989) and usual sample median $\widehat{M}_Y$, we first write the expressions of MSEs of estimators $\widehat{M}_{YR}$ and $\widehat{M}_Y$of population median up to the first order of approximation as

$$\text{MSE}\,(\widehat{M}_{YR}) = (1-f)(4n)^{-1}\,[\,\{f_Y(M_Y)\}^{-2} + \left(\frac{M_Y}{M_X}\right)^2 \{f_X(M_X)\}^{-2}$$

$$-2\{4P_{11}(X,Y) - 1\}(\frac{M_Y}{M_X})\{f_X(M_X)f_Y(M_Y)\}^{-1}] \tag{3.1}$$

MSE $(\widehat{M}_Y) = (1-f)(4n)^{-1}\{f_Y(M_Y)\}^{-2}$ (3.2)

Using (2.9) & (3.1), we have

MSE $(\widehat{M}_{YR})$ - MSE $(\widehat{M}_{H1}) = [\frac{M_Y}{M_X}\{f_X(M_X)\}^{-1} - \{f_Y(M_Y)\}^{-1}\{4P_{11}(X,Y)-1\}]^2$

$\geq$ 0, which is always true. (3.3)

Similarly, using (2.9) & (3.2), we have

MSE $(\widehat{M}_Y)$ - MSE $(\widehat{M}_{H1}) = \{4P_{11}(X,Y)-1\}^2$

$\geq$ 0, which is always true. (3.4)

From (3.3) and (3.4), we note that the estimator $\widehat{M}_{H1}$ is always efficient than the estimator $\widehat{M}_{YR}$ defined by Kuk and Mak (1989) and usual sample median $\widehat{M}_Y$ but it is equally efficient to the other two estimators proposed by KUK and Mak (1989).

4. Numerical Illustration

To obtain the rough idea about the efficiencies of proposed estimators over the existing ones, simulation study has been carried out using R software in which we drew 10,00,000 repeated samples from a bivariate normal population for different correlation coefficient values with different samples sizes having Medians : M_Y = 4, M_X = 3, Means: μ_X = 4, μ_Y = 3 and Standard deviations : σ_Y = 3, σ_X = 1. Numerical values of results are given in table 4.1

Table 4.1 Biases of different estimators

Correlation coefficient (ρ_c)	Sample size(n)	Bias					
		$\widehat{M}_{H1}$	$\widehat{M}_{H2}$	$\widehat{M}_{Y}$	$\widehat{M}_{YR}$	$\widehat{M}_{YP}$	$\widehat{M}_{YS}$
0.3	3	0.049	0.045	1.89	0.154	0.151	0.152
	5	0.032	0.029	1.15	0.094	0.093	0.093
	7	0.022	0.019	0.84	0.066	0.058	0.057
	9	0.018	0.015	0.67	0.053	0.049	0.049
0.5	3	0.055	0.051	1.89	0.097	0.088	0.087
	5	0.037	0.032	1.15	0.062	0.058	0.057
	7	0.026	0.023	0.84	0.044	0.036	0.034
	9	0.022	0.018	0.67	0.037	0.030	0.030
0.7	3	0.025	0.022	1.89	0.029	0.028	0.027
	5	0.021	0.019	1.15	0.026	0.026	0.026
	7	0.017	0.013	0.84	0.021	0.019	0.017
	9	0.015	0.012	0.67	0.017	0.016	0.014
0.9	3	0.058	0.046	1.89	0.059	0.812	0.811
	5	0.036	0.01	1.15	0.029	0.486	0.484
	7	0.021	0.011	0.84	0.017	0.355	0.358
	9	0.014	0.005	0.67	0.011	0.273	0.269

Table 4.2 Comparison of efficiencies of estimators

Correlation coefficient (ρ_c)	Sample size(n)	MSE				Relative efficiencies		
		$\widehat{M}_Y$	$\widehat{M}_{YR}$	$\widehat{M}_{H1}$	$\widehat{M}_{H2}$	$\widehat{M}_Y$	$\widehat{M}_{YR}$	$\widehat{M}_{H1}$
0.3	3	1.89	3.516	1.88	1.88	100	53.75	100.5
	5	1.15	2.124	1.13	1.13	100	54.14	101.8
	7	0.84	1.491	0.82	0.82	100	56.33	102.4
	9	0.67	1.122	0.65	0.65	100	59.82	103.1
0.5	3	1.89	2.114	1.71	1.71	100	89.40	110.5
	5	1.15	1.258	1.01	1.01	100	91.41	113.8
	7	0.84	0.901	0.72	0.72	100	93.22	116.7
	9	0.67	0.698	0.56	0.56	100	95.98	119.6
0.7	3	1.89	1.457	1.39	1.39	100	129.7	136.0
	5	1.15	0.881	0.82	0.82	100	130.5	140.2
	7	0.84	0.639	0.59	0.59	100	131.4	142.4
	9	0.67	0.505	0.46	0.46	100	132.7	145.7
0.9	3	1.89	0.835	0.81	0.81	100	226.4	233.3
	5	1.15	0.501	0.48	0.48	100	229.5	239.6
	7	0.84	0.365	0.35	0.35	100	230.1	240.0
	9	0.67	0.287	0.27	0.27	100	233.4	248.2

From Table 4.1, we see that proposed estimators have lower bias than all the three estimators proposed by Kuk and Mak(1989) and usual sample median estimator. Also the bias of estimator $\widehat{M}_{H2}$ is lower than

estimator $\widehat{M}_{H1}$. From table 4.2, it is clear that proposed estimators always have higher efficiency than the ratio estimator defined by Kuk and Mak(1989) and usual sample median estimator.

5. Conclusion

The theoretical study shows that proposed estimators are always more efficient than ratio estimator defined by Kuk and Mak (1989) as well as usual sample median estimator for all the situations. It has also been shown that both the estimators $\widehat{M}_{H1}$ and $\widehat{M}_{H2}$ are equally efficient but in spite of exact bias of $\widehat{M}_{H2}$ as compared to the bias of $\widehat{M}_{H1}$ taken up to first order of approximation is smaller than $\widehat{M}_{H1}$. Biases of the proposed estimators are less than the estimators proposed by Kuk and Mak(1989).It is also shown that efficient estimators can be constructed by choosing the values of α and v in the proposed estimators corresponding to given values of C or range of C. Numerical results given in table 4.2 by using simulation also show that the proposed estimators are always efficient than ratio estimator defined by Kuk and Mak (1989) as well as usual sample median estimator and table 4.1 shows that bias of proposed estimators is less than the estimators proposed by Kuk and Mak(1989) and usual sample median estimator.

References

[1] A.Arcos, M. Rueda, M.D. Martinz, S. Gonzalez, and Y. Roman, "Incorporating the auxiliary information available in variance estimation", *Applied Mathematics and Computation* , vol. 160, pp.387-399, 2005.

[2] R.L.Chambers, R. Dunstan, "Estimating distribution functions from survey data", *Biometrik*a, vol.73, pp.597 -604,1986.

[3] S.T. Gross, "Median estimation in sample surveys". *Proc. Surv. Res. Meth. Sect. Amer. Statist. Ass.*, pp. 181-184, 1980.

[4] Y.C.A. Kuk and T.K. Mak, "Median estimation in the presence of auxiliary information", *J.R. Statist*. Soc. B, vol. 2, pp.261-269, 1989.

[5] T.K. Mak and A.Y.C. Kuk , "A new method for estimating finite population quantiles using auxiliary information", *The Canadian Journal of Statistics* vol. 25, pp.29-38, 1993.

[6] G. Meeden , "Median estimation using auxiliary information". *Survey Methodology*, vol. 21, pp.71-77,1980.

[7] J.N.K. Rao, J.G. Kovar, H.J. Mantel , "On estimating distribution functions and quantiles from survey data using auxiliary information". *Biometrika* vol. 77, pp. 365-375, 1990.

[8] M. Rueda, and A. Arcos, "On estimating the median from survey data using multiple auxiliary information". *Metrika*, vol. 54, pp.59-76, 2001.

[9] J. Sedransk and J. Meyer, "Confidence intervals for the quantiles of a finite population: simple random sampling and stratified simple random sampling".*J.R.Statist.* Soc. B, vol. 40, pp.239 -252, 1978.

[10] B.W. Silverman, *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall, 1986.

[11] S. Singh, P.S. Housila and L.N. Upadhyaya,. "Chain ratio and regression type estimators for median estimation in survey sampling". *Staistical Papers*, no.1, vol. 48, pp. 23-46, 2007.

[12] P. Smith and J. Sedransk, "Lower bounds for confidence coefficients for confidence intervals for finite population quantiles". Commun. Statist. Theor. Meth., vol.12, pp.1329- 1344,1983.