*American Journal of Applied Mathematics and Statistics, 2014, Vol. 2, No. 3, 163-167*
Available online at http://pubs.sciepub.com/ajams/2/3/12

DOI:10.12691/ajams-2-3-12

# Estimation of Population Total in the Presence of Missing Values Using a Modified Murthy's Estimator and the Weight Adjustment Technique

**Oyoo David Odhiambo, Christopher Ouma Onyango***

Department of Statistics and Actuarial Science, Kenyatta University-Kenya
*Corresponding author: chrisouma2004@yahoo.co.uk

*Received March 27, 2014; Revised May 08, 2014; Accepted May 22, 2014*

**Abstract** Use of Murthy's method in estimation of population parameters, such as population totals, population means, and population variances has been limited to surveys where survey data values are complete. This study applies weight adjustment technique to estimate a population total under simple random sampling without replacement. The asymptotic properties show that the estimated population total is sufficient for the true population total. The proposed estimator is obtained by symmetrizing Murthy's estimator.

***Keywords:*** *Murthy's estimator, missing values, weight adjustment*

**Cite This Article:** Oyoo David Odhiambo, and Christopher Ouma Onyango, "Estimation of Population Total in the Presence of Missing Values Using a Modified Murthy's Estimator and the Weight Adjustment Technique." *American Journal of Applied Mathematics and Statistics*, vol. 2, no. 3 (2014): 163-167. doi: 10.12691/ajams-2-3-12.

## 1. Introduction

In sample surveys, modeling an optimal estimator that best estimates finite population total has been of interest to modern statisticians (Ouma et al., 2010). Estimation methods for some population parameters include, among others, ratio estimation, Horvitzand Thompson estimation, and Yates and Grundy estimation. From these studies, various estimators of population have been obtained. In this paper, we have obtained a new estimator by symmetrizing Murthy's estimator. We have then estimated finite population total in the presence of missing data using the derived estimator. As a way of correcting the 'missingness' of data, weight adjustment method has been used.

### 1.1 Background of the Problem

In sample surveys, completeness of observed datais one factor that influences inferences made on results of a study. Daroga and Chaudhary (2002) explained that missing data distort validity and reliability of a study. Consequently, variousmethods of correcting missing data have been proposed in sample surveys. Some of the methods include: imputation techniques,partial deletion andresampling (Brewer, 2002, Broemeling,2009). Singh and Solanki (2012) later not only supported Broemeling'sproposal **(**2009), but also observed that previous studies have not extensively used samples with missing data. This research has, therefore, filled this gap by using a sample with missing values. Singh and Solanki (2012) further observed that previous studies have only focused on ordered sampling procedures. However, not all sets of data are ordered. In filling this gap, this study utilizes Murthy's estimation method, which involves unordered sampling procedures (Murthy, 1957).

## 2. Murthy's Estimation

Murthy's estimator has been used for constructing unbiased estimators of population totals and/or mean from a sample of fixed size. Let $\hat{Y}_M$ be an estimator of population parameter $\theta$ based on the ordered sample ($s_i$), Murthy's estimator for population total is given by

$$\hat{Y}_M = \frac{\sum_i^n P(s/i) y_i}{P(s)}$$

Where,
*P(s/i)* = conditional probability of getting the set of units that was drawn, given that the *i-th* unit was drawn first.
*P(s)* = unconditional probability of getting the set of units that was drawn
Consider a random selection of three population units *i*, *j*, and *k* are randomly selected from a population of size *N* with the corresponding selection probabilities be $z_i$, $z_j/(1-z_i)$, and $z_k/(1-z_i-z_j)$.

Then we can show that Murthy's estimator, $\hat{Y}_M$ is unbiased for the population total *Y* and its variance for n = 2 is given by

$$Var\left(\hat{Y}_M\right) = \sum_s P(s) \hat{Y}_M{}^2 - Y^2$$

$$= \sum_i^N \sum_{j>i}^N \frac{z_i z_j (2 - z_i - z_j)}{(1 - z_i)(1 - z_j)} \hat{Y}_M{}^2 - Y^2$$

Which can be rearranged as follows

$$Var\left(\hat{Y}_M\right) = \sum_{i}^{N}\sum_{j>i}^{N} \frac{z_i z_j (1 - z_i - z_j)}{2 - z_i - z_j}\left(\frac{y_i}{z_i} - \frac{y_j}{z_j}\right)^2$$

# 3. Proposed Estimator

The proposed estimator is given by

$$t_w = \sum_{c=1}^{k}\sum_{i\in\Phi_c} w_{ci} y_{ci}$$

Where $w_{ci}$ = weight adjustment of i[th] unit in group c and $w_{ci}$ an be expressed as

$w_{ci} = \frac{N_c}{m_c}$, where $N_c$ = population size in group c, $m_c$ = number of units with complete data

## 3.1. Derivation of the Proposed Estimator

By assuming any two population units $y_i$ and $y_j$ and the corresponding selection probabilities $p_i$ and $p_i$, Shahbaz (2004) modified Murthy's estimator as

$$t_M = \frac{1}{2}\left[\frac{y_i}{p_i} + \frac{y_j}{p_j}\right]$$

And Shahbaz and Ayesha (2008) symmetrized the partitioned estimator as $T_1$ and $T_2$ given by

$$T_1 = y_1 + \frac{y_2}{p_2}(1 - p_2)$$

and

$$T_2 = y_2 + \frac{y_1}{p_1}(1 - p_1)\left[k - \frac{p_2}{1 - p_2}\right],$$

where $k = \sum_{i=1}^{N}\frac{p_i}{1 - p_i}$

Suppose the symmetrization is such that $T_2 = y_2 + \frac{y_1}{p_1}(1 - p_1)$, then define $T^*_m$ as

$$T^*_m = \frac{1}{2}(T_1 + T_2) = \frac{1}{2}\left[\frac{y_1}{p_1} + \frac{y_2}{p_2}\right] \quad (1)$$

Equation (1) is only for selecting 2 units. Suppose we consider n units, we get $T^*$ given by

$$T^* = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} \quad (2)$$

Since the study involves estimating finite population total in the presence of missing data, we apply weighting adjustment to correct the "missingness" of responses. We proceed as follows;

For any population of size N, as $n \to N, p_i \to 1 \forall i$, then $p_i \approx p_j$ and $n/N \to 1$. That is, for large n, the inclusion probabilities are asymptotically equal and $p_i \to 1/N$ (Cochran 1977)

Using the results for large n and asymptotic value of $p_i$, equation (2) reduces to

$$\sum_{i=1}^{n} y_i = T_w \quad (3)$$

Equations (3) and the proposed estimator are similar if the weighting constant $w_{ci} = 1$ and $c = 1$. Our task is therefore to determine the value of $w_{ci}$.

Consider the set $U\{1,2,3,\ldots..N\}$ and $S = \{1,2,3,\ldots..n\}$ be a set chosen from U.

Define a population of size N as $U_y = \{y_i / i \in U\}$ and a sample of size n as $S_y = \{y_i / i \in U\}$.

Let the respective population and sample totals be

$$t_p = \sum_{i=1}^{N} y_i \quad and \quad t_s = \sum_{1\in S} y_i$$

And the corresponding population and sample means are given by

$$\bar{Y} = \frac{t_p}{N} \quad and \quad \bar{y} = \frac{t_s}{n}$$

Since $\bar{y}$ is unbiased for $\bar{Y}$ it follows that $N\bar{Y} = E(N\bar{y})$ and hence the estimator of population is

$$\hat{t}_p = N/n\, t_s = \sum_{i\in S} w_i y_i \quad \text{where} \quad \text{w}_\text{i} = N/n.$$

## 3.2. Weighting Adjustment

Suppose the population can be classified to form k groups based on auxiliary information $X_i (i = 1,2,\ldots\ldots N)$. Using the definition of S above, let us partition S as

$$S = \bigcup_{c=1}^{k} S_c sothat\, U = S \bigcup S'$$

Using the k classes, there exists partitions $U_1$, $U_2$, ……… , $U_k$ such that $S_c \subset U_c$, $\forall c = 1,2,\ldots k$.

Let $\Phi_c$ be the set containing identified numbers of responding units in class c (i.e with no missing information).

$$\Rightarrow \Phi_c \subset S_c, c = 1,2,\ldots\ldots k.$$

Let the sizes of $U_c$, $S_c$, and $\Phi_c$ are $N_c$, $n_c$, and $m_c$ respectively, then by letting $m_c > 1$, we have

$$m = \sum_{c=1}^{k} m_c, n = \sum_{c=1}^{k} n_c, N = \sum_{c=1}^{k} N_c$$

Consider any class c (c = 1, 2, …..k), $m_c$ is used to represent $n_c$. This implies that each of the mc units has a weight of $n_c/m_c$.

Let $y_{ci}$ be a study observation with an identification number i in class c. If we define

$$t_{sc} = \sum_{i \in S_c} y_{ci}, \text{ then } t_p = \sum_{c=1}^{k} t_c \text{ where } t_c = \sum_{i \in U_c} y_{ci}$$

And from equation (4), $t_c$ can be estimated by $\frac{N_c}{n_c} t_{sc}$.

That is,

$$\hat{t}_c = \frac{N_c}{n_c} t_{sc}. \quad (5)$$

Then, for known $N_c$,

$$\hat{t}_c = \frac{N_c}{n_c} \hat{t}_{sc} = \frac{N_c n_c}{n_c m_c} t_{sc}^{*} = \frac{N_c}{m_c} \sum_{i \in \Phi_c} y_{ci} \quad (6)$$

Equation (6) implies that a sample of size $m_c$ is used to represent a population of size $N_c$. The overall adjusted estimator can thus be written as

$$\hat{t}_p = t_w = \sum_{c=1}^{k} \sum_{i \in \Phi_c} w_{ci} y_{ci} \quad (7)$$

Where $w_{ci} = \frac{N_c}{m_c}$. And $w_{ci}$ can be expressed as $w_{ci}$ = $w_1.w_2$, where $w_1 = \frac{N_c}{n_c}$ is the base weight in class c and $w_2 = \frac{n_c}{m_c}$ is the non-response adjusted weight in class c.

# 4. Properties of the Proposed Estimator

## 4.1. Unbiasedness

Define a vector $r' = (m_c, n_c, N_c)'$ so that $R' = (m_1, m_2, \ldots m_k, n_1, n_2 \ldots n_k, N_1, N_2 \ldots N_K)'$

Now,

$$E\left(t_w / R\right) = E\left\{\sum_{c=1}^{k} \sum_{i \in \Phi} \frac{N_c}{m_c} y_{ci} / R\right\}$$

$$= \sum_{c=1}^{k} E\left(t_c / R\right) = \sum_{c=1}^{k} t_c = t_p.$$

Hence the estimator is unbiased.

## 4.2. Variance of the Proposed Estimator

Since the nature of sampling makes the entire sampling procedure analogous to Simple Random Sampling (SRS). Suppose we consider one of the classes and use a sample of size $m_c$ to estimate parameters in a population of size $N_c$, we can apply the procedures in SRS to derive this variance.

Since $\bar{y}$ is unbiased for $\bar{Y}$ it follows that $Var\left(N_c \bar{Y}\right) = Var\left(N_c \bar{y}\right) = N^2{}_c Var(\bar{y})$

Recall: $Var\left(\bar{y}\right) = E\left(\bar{y}^2\right) - [\bar{y}]^2 = E\left(\bar{y}^2\right) - \bar{y}^2$ (Cochran, 1977)

Now

$$E\left(\bar{y}^2\right) = E\left[\frac{1}{n} \sum_{1}^{n} y_i\right]^2$$

$$= \frac{1}{n^2} E\left(\sum_{1}^{n} y_i\right)^2 = \frac{1}{n^2} E\left(\sum_{1}^{n} y_i\right)^2$$

$$= \frac{1}{n^2} E\left\{\sum_{1}^{n} y_i^2 + \sum_{i}^{n} \sum_{j}^{n} y_i y_j\right\}$$

Define $a_i = \begin{cases} 1 & \text{if } i-th \text{ unit is in the sample} \\ 0, & \text{otherwise} \end{cases}$

In SRS, $P(a_i = 1) = n / N$

$$P\left(a_i a_j = 1\right) = P\left(a_i = 1 and a_j = 1\right)$$

And

$$= P(a_i = 1).P\left(a_j = 1 / a_i = 1\right) = \frac{n}{N} . \frac{n-1}{N-1}$$

Hence

$$E\left(\bar{y}^2\right) = \frac{1}{n^2} E\left\{\sum_{1}^{N} a_i\ y_j\ ^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j\ y_i\ y_j\right\}$$

$$= \frac{1}{n^2} \left\{\sum_{1}^{N} y_i^2 E\left(a_i\right) + \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j\ E\left(a_i a_j\right)\right\}$$

$$= \frac{1}{n^2} \left[\frac{n}{N} \sum_{1}^{N} y_i^2 + \frac{n(n-1)}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j\right]$$

which on simplification gives,

$$= \frac{N-n}{nN(N-1)} \sum_{1}^{N} y_i^2 + \frac{N^2(n-1)}{Nn(N-1)} \bar{y}^2$$

and this simplifies to,

$$Var\left(\bar{y}\right) = \frac{N-n}{nN(N-1)} \sum_{1}^{N} y_i^2 + \frac{\left(N^2 n - N^2\right)}{Nn(N-1)} \bar{y}^2 - \bar{y}^2$$

$$= \frac{N-n}{nN(N-1)} \sum_{1}^{N} y_i^2 + \frac{N(N-n)}{nN(N-1)} \bar{Y}^2$$

$$= \frac{N-n}{nN} S^2 \qquad \text{(Cochran, 1977)}$$

Hence, $Var\left(\bar{y}\right) = \frac{N_c - m_c}{N_c m_c} S_c^2$ where, $S_c^2 = \frac{1}{N_c - 1}\left[\sum_{1}^{N_c} Y_i^2 - N\bar{Y}_c^2\right]$. Thus,

$$N_c^2 Var\left(\bar{y}\right) = N_c^2 . \frac{N_c - m_c}{N_c m_c} S_c^2 \quad (8)$$

Now,

$$Var\left(t_w / R\right) = \sum_{c=1}^{k} Var\left(t_c / \mathrm{E}\right) = \sum_{c=1}^{k} N_c^2 . \frac{N_c - m_c}{N_c m_c} S_c^2$$

But in SRS, sample variance ( $s_c^2$ ) is unbiased for population variance ( $S_c^2$ ). Where

$$s_c^{\,2} = \frac{1}{m_c - 1}\left[\sum_{1}^{m_c} y_i^{\,2} - m_c \bar{y}_c^{\,2}\right]$$

Therefore, overall variance of the estimator is

$$Var\left(t_w / R\right) = \sum_{c=1}^{k} N_c^{\,2} \cdot \frac{N_c - m_c}{N_c m_c} s_c^{\,2} \qquad (9)$$

### 4.3. Consistency of the Proposed Estimator

Consider the proposed estimator $t_w$ and finite population total $t_p$. A sequence of point estimators $t_w^* = \left(y_{c1}, y_{c2}, \ldots\ldots. y_{cm_c}\right) \forall c$, is said to be weakly consistent for $t_p$ if $t_w^*$ converges in probability to $t_p$.

That is,

$$\lim_{m_c \to \infty} P\left\{\left|t_w - t_p\right| > \varepsilon\right\} = 0, \quad \textit{for every } \varepsilon > 0$$

Proof: By Chebychev's inequality, for every $\varepsilon > 0$.

$$P\left\{\left|t_w - t_p\right| > \varepsilon\right\} \le \frac{Var\left(t_w / R\right)}{\varepsilon^2} = \frac{1}{\varepsilon^2} \sum_{c=1}^{k} N_c^{\,2} \cdot \frac{N_c - m_c}{N_c m_c} s_c^{\,2}$$

Taking limits as $m_c \to \infty$, the right hand side $\to 0$.

Hence, $t_w \xrightarrow{P} t_p$, which is the necessary and sufficient condition for consistency.

### 4.4. Bias of the Proposed Estimator

From equation (8), we assume that $N_c$ ($\forall c$) is known. Suppose that $N_c$ is not known, we need to estimate $N_c$ and consequently a new $t_w^*$. Suppose the classification is such that the subpopulation ratio $n_c / N_c$ is equal to $n / N$. That is, sampling distribution of $n_c / N_c$ is centered on $n / N$.

$$\Rightarrow E\left(n_c / N_c\right) = n / N \Rightarrow N_c = N . n_c / n \qquad (10)$$

Replacing equation (10) in equation (7), we have

$$t_w^* = \sum_{c=1}^{k} \sum_{i \in \Phi_c} N / n \cdot \frac{n_c}{m_c} y_{ci} \qquad (11)$$

And consequently $Var\left(t_w^* / \mathrm{E}\right)$ becomes

$$Var\left(t_w^* / R\right) = \sum_{c=1}^{k} \left(N \frac{n_c}{n}\right)^2 \cdot \frac{N_c - m_c}{N_c m_c} s_c^{\,2} \qquad (12)$$

We can thus obtain Bias ($t_w^*$) instead of Bias ($t_w$)

Bias $\left(t_w^*\right) = E\left[t_w^* - t_p\right] = E\left[t_w^*\right] - t_p$, since $t_p$ is constant. (Cochran, 1977)

$$E\left[t_w^*\right] = E\left[\sum_{c=1}^{k} \sum_{i \in \Phi_c} N / n \cdot \frac{n_c}{m_c} y_{ci}\right] = N / n \sum_{c=1}^{k} E\left[\sum_{i \in \Phi_c} \frac{n_c}{m_c} y_{ci}\right] \qquad (13)$$

But from previous workings,

$$\hat{t}_{sc} = \frac{n_c}{m_c} \sum_{i \in \Phi_c} y_{ci}$$

$$\Rightarrow E\left[t_w^*\right] = \sum_{c=i}^{k} N / n \, \hat{t}_{sc} \qquad (14)$$

From equations (6) and (7)

$$t_p = \sum_{c=i}^{k} N_c / n_c \, \hat{t}_{sc} \qquad (15)$$

Substituting (14) and (15) in equation (13) and simplifying, we obtain

$$Bias\left(t_w^*\right) = \sum_{c=1}^{k} \left(N / n - N_c / n_c\right) \hat{t}_{sc} \qquad (16)$$

Clearly, Bias ($t_w^*$) vanishes if $N / n = N_c / n_c$.

### 4.5. Expected Mean Squared Error (MSE) of the Proposed Estimator

$$MSE\left(t_w^*\right) = E\left[t_w^* - t_p\right]^2$$

$$= E\left[t_w^* + \mathrm{E}\left(t_w^*\right) - \mathrm{E}(t_w^*) - t_p\right]^2$$

(Daroga and Chaudhary, 2002)

$$= E\left[t_w^* - E(t_w^*)\right]^2 + \left[E\left(t_w^*\right) - t_p\right]^2$$

$$= Var\left(t_w^*\right) + \left[Bias(t_w^*)\right]^2$$

Where,

$$Var\left(t_w^* / R\right) = \sum_{c=1}^{k} \left(N \frac{n_c}{n}\right)^2 \cdot \frac{N_c - m_c}{N_c m_c} s_c^{\,2}$$

$$\textit{and} \quad Bias(t_w^*) = \sum_{c=1}^{k} \left(N / n - N_c / n_c\right) \hat{t}_{sc}.$$

## References


[1] Brick, J.M. and Kalton, G. (1996) Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.

[2] Broemeling, D. L. (2009). *Bayesian Methods for Measures of Agreement (Chapman & Hall/CRC Biostatistics Series)*. Chapman and Hall/CRC Press.

[3] Cochran, W. G. (1977). *Sampling Techniques*. 3rd Edition. New York, John Wiley.

[4] Chang, C. and Ferry, B. (2012). Weighting Methods in Survey Sampling. *Section on Survey Research Methods-JSM*, 4768-4782.

[5] Daroga, S. and Chaudhary, F. (2002). *Theory and Analysis of Sample Survey Designs*. New Delhi: New Age International (P) Limited Publishers.

[6] Murthy, M. N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhya*, 18, 379-390.

[7] Ouma, C., Odhiambo, R. and Orwa, G. (2010). Bootstrapping in Model-Based Estimation of a Finite Population Total Under Two-Stage Cluster Sampling With Unequal Cluster Sizes. *Annals of Statistics*, July Issue, 171-184.

[8] Salehi, M. and Seber, G. A. (2002). Theory & Methods: A New Proof of Murthy's Estimator which Applies to Sequential Sampling. *Australian & AMP New Zealand Journal of Statistics*, 43(3), 281-286.

[9] Shahbaz, Q. M., and Ayesha, S. (2008). A new symmetrized estimator of population total in unequal probability sampling. Journal of Statistics, 13(1), 20-25.

[10] Singh, H. P. and Solanki, R. S. (2012). An alternative procedure for estimating the population mean in simple random sampling. *Pakistan Journal of Statistics and Operation Research*, 8(2), N 1816-2711.