THE EGYPTIAN STATISTICAL JOURNAL
ISSR, CARIO, UNIV., VOL.(42), NO.2, 1998

Estimating Finite Population Parameters Using Stratified Sample in The Presence of The Auxiliary Information

by: Mounira A. Hussein **

Introduction:

Stratification is a common technique to increase the precision of the finite population estimators .It is suggested when it is possible to divide a heterogenous population into homogenous subpopulations. It is useful when the data of known precision are wanted or when sampling proplems differ markedly in different parts of the population (Cochran ,1977). Auxiliary information is often used for increasing the precision of the estimators of the population parameters .The resulting methods of estimation ,which make use of the auxiliary information ,under certain conditions give more efficient estimates of the population parameters (Hussein, 1988). When the auxiliary variable has a high positive correlation with the dependent variable, two ratio estimators were developed for stratified sample, the separate and the combined estimators.Their mathematical formula can be presented respectively as shown below.

$$\hat{Y}_{RS} = \sum_{h=1}^{L} \frac{\bar{y}_h}{\bar{x}_h} X_h \quad \text{and} \quad \hat{Y}_{RC} = \frac{\bar{y}_{st}}{\bar{x}_{st}} X$$

where $\bar{y}_{st} = \sum_{h=1}^{L} N_h \bar{y}_h$ and $\bar{x}_{st} = \sum_{h=1}^{L} N_h \bar{x}_h$

When the auxiliary variable has a moderate positive correlation with the dependent variable, product type of estimator can be presented respectively as shown below:

---

* Mounira A. Hussein is an Associate Professor , College of Commerce, Menoufia University.

$$\hat{Y}_{as} = \sum_{h=1}^{L} \frac{\hat{X}^*_h}{X_h} \hat{Y}_h \quad and \quad \hat{Y}_{ac} = \hat{X}_{st}^* \frac{\hat{Y}_{st}}{X}$$

$$where \quad \hat{X}^* = (N_h X_h - n_h \hat{X}_h) / (N_h - n_h)$$

$$\hat{X}_{st}^* = \sum_{h=1}^{L} \hat{X}_h^* \quad and \quad \hat{Y}_{st} = \sum_{h=1}^{L} \hat{Y}_h$$

When the auxiliary variable has a high negative correlation with the dependent variable, a product type of estimator for stratified sample can be presented as a straight forward extension of the well-known product type of estimator for simple random sample (Cochran 1977, Scheaffer, Mendenhall and Ott 1983). This estimate can be presented as follows:

$$\hat{Y}_{PS} = \sum_{h=1}^{L} \frac{\bar{x}_h}{X_h} \bar{y}_h$$

When we have a moderate negative correlation with the dependent variable ,two other alternative product estimators were proposed (Srivenkataramana and Bandyopadhyay 1980, Hussein 1988). Their mathematical formula can be presented respectively as follows:

$$\hat{Z}_{as} = \sum_{h=1}^{L} \frac{X_h}{\hat{X}^*_h} \hat{Y}_h \quad and \quad \hat{Z}_{ac} = \frac{X}{\hat{X}^*_{st}} \hat{Y}_{st}$$

When the auxiliary variable has a low correlation, the stratified sample mean is the best estimator which can be presented as follows:

$$\hat{Y}_{st} = \sum_{h=1}^{L} N_h \bar{y}_h$$

The mean square error of the total estimator using the mean per element,

ratio and product type estimators for each stratum were proposed in Hussein (1995) as follows:

Estimator	Mean Square Error
$\hat{Y}_{st}$	$\sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 C_{yyh}$
$\hat{Y}_{Rs}$	$\sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \{C_{yyh}-2C_{yxh}+C_{xxh}\}$
$\hat{Y}_{as}$	$\sum_{h=1}^{L} \frac{1-f_h}{n_h} Y_h^2 \{C_{yyh}-2g_h C_{yxh}+g_h^2 C_{xxh}\}$

In practical situations the correlation between the dependent variable and the auxiliary variable differ from one stratum to another (Scheaffer, Mendenhall, and Ott 1983). It can be high for some strata, moderate in some strata and low in others. So it is important to develop a general estimator that deals with any practical situation . Also since some of the variables have symmetric distributions and others have skewed distribution it is quite important to develop this estimator for the median not only for the mean and the total.

Median is often regarded as a more appropriate measure of allocation than the mean with a highly skewed distribution (Kuk and Mak 1989). Most of the research on median estimation in stratified random sample deals exclusively with the survey variable of interest alone and does not make explicit use of auxiliary variables in the construction of the estimators (Kuk and Mak 1989).

Notation and Definitions:

The following notations are used in this paper :

1- $\bar{x}_h$ and $\bar{y}_h$ are the sample means of the auxiliary variable X and the dependent variable Y of the stratum h.

2- N_h and $n_h < N_h$ are the population and the sample size in stratum h .

3- Y_h is the population total of the characteristic under study in stratum h.

4- L : number of strata.

5- X : total population of the auxiliary variable.

6- $\hat{Y}_h = N_h \bar{y}_h$ and $\hat{x}_h = N_h \bar{x}_h$ are unbiased estimators for Y_h and x_h based on a sample of size n_h .

7- C_{yyh}, C_{xxh}, and C_{yxh} are the relative variance for the characteristic under study, relative variance for the auxiliary variable and their relative covariance respectively.

Study Objectives:

This paper is organized to accomplish the following objectives :

1-Developing the mathematical formula of two generalized estimators of the total population from stratified sample which deal with different levels of correlation between the auxiliary variable and the dependent variable.The first for positive correlation and the other for negative correlation.

2-Constructing the mathematical formula of their mean square errors and comparing their mean square errors with other estimators cited in the introduction.

3- Constructing two estimators for the median to deal with highly skewed distributions

General Estimator of The Total Population:

A- The auxiliary variable is positively correlated with the dependent variable:

Let

$$H_h(x) = \frac{\hat{X}_h}{X_h} = \frac{\bar{x}_h}{\bar{X}_h}$$

Since

$$\hat{Y}_{RS} = \sum_{h=1}^{L} \frac{\bar{y}_h}{\bar{x}_h} X_h = \sum_{h=1}^{L} (N_h \bar{y}_h) \frac{\bar{X}_h}{\bar{x}_h}$$

Then

$$\hat{Y}_{RS} = \sum_{h=1}^{L} \frac{\hat{Y}_h}{H_h(x)}$$

Let $$g_h = \frac{n_h}{N_h - n_h}$$

Since

$$\hat{Y}_{as} = \sum_{h=1}^{L} \frac{(N_h X_h - n_h \hat{X}_h)}{(N_h - n_h) X_h} \hat{Y}_h = \sum_{h=1}^{L} \{\frac{N_h}{(N_h - n_h)} \hat{Y}_h - \frac{n_h}{(N_h - n_h)} \frac{\hat{X}_h}{X_h} \hat{Y}_h\}$$

Then $$\hat{Y}_{as} = \sum_{h=1}^{l} \hat{Y}_h [1 + g_h - g_h H_h(x)]$$

Suppose that the mean per element is the best estimator (it has the least mean square error according to the criteria cited in (Hussein,1995)) for the first stratum, and the ratio estimator is the best estimator for the second stratum, and the alternative product estimator is the best estimator for the third stratum. The general estimator and its mean square error can be prsented as follows:

$$\hat{Y}_s = [\hat{Y}_1 \quad \hat{Y}_2 \quad \hat{Y}_3] \begin{Bmatrix} 1 \\ \frac{1}{H_2(x)} \\ 1 + g_3 - g_3 H_3(x) \end{Bmatrix}$$

$$M(\hat{Y}_s) = [1 \ 1 \ 1] \begin{Bmatrix} 1 & 1 & 1 \\ 0 & -2 & -2g_3 \\ 0 & 1 & g_3^2 \end{Bmatrix} \Theta \begin{Bmatrix} c_{yy1} & c_{yy2} & c_{yy3} \\ c_{yx1} & c_{yx2} & c_{yx3} \\ c_{xx1} & c_{xx2} & c_{xx3} \end{Bmatrix} * \begin{Bmatrix} k_1 \\ k_2 \\ k_3 \end{Bmatrix}$$

$$M(\hat{Y}_s) = \underset{\sim}{1}' G \Theta C \underset{\sim}{k} \quad \text{where } k_h = \frac{1 - f_h}{n_h} \hat{Y}_h$$

B-The auxiliary variable is negatively correlated with the dependent variable:

Suppose that the mean per element is the best estimator for the first stratum, and the product estimator is th best estimator for the second stratum, and the alternative product estimator is the best estimator for the third stratum. The general estimator $\hat{Z}s$ and its mean square error can be prsented as follows:

$$Since\ \hat{Z}_{as} = \sum_{h=1}^{l} \hat{Y}_h \Big/ \frac{X_h}{\hat{Xh}} \quad where\ \hat{Xh} = \frac{(N_h X_h - n_h \hat{Xh})}{(N_h - n_h)}$$

$$Therefore,\ \hat{Zas} \sum_{h=1}^{l} \hat{Y}_h (1+g_h - g_h H_h(x))^{-1}$$

$$Since \quad \hat{Y}_{ps} = \sum_{h=1}^{L} \hat{Y}_h \frac{\hat{Xh}}{Xh}$$

$$Therfore, \quad \hat{Y}_{ps} = \sum_{h=1}^{L} \hat{Y}_h H_h(x)$$

$$\textbf{Thus,} \quad \hat{Z}_s = [\hat{Y}_1 \quad \hat{Y}_2 \quad \hat{Y}_3] \begin{Bmatrix} 1 \\ H_2(x) \\ \dfrac{1}{1+g_3 - g_3 H_3(x)} \end{Bmatrix}$$

$$M(\hat{Z}_s) = [1\ 1\ 1] \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 2g_3 \\ 0 & 1 & g_3^2 \end{bmatrix} \Theta \begin{bmatrix} c_{yy1} & c_{yy2} & c_{yy3} \\ c_{yx1} & c_{yx2} & c_{yx3} \\ c_{xx1} & c_{xx2} & c_{xx3} \end{bmatrix} * \begin{Bmatrix} k_1 \\ k_2 \\ k_3 \end{Bmatrix}$$

$$M(\hat{Z}_s) = 1' G_1 \Theta C' k$$

Where C is the coefficient of variation matrix.

G is the matrix that can be determined after applying the criteria (Hussein ,1995) of the best estimator for positively correlated variable.

G_1 is the matrix that can be determined after applying the criteria of the best estimator for negatively correlated variable.

Median Estimators For Stratified Random Sample:

In this paper two estimators are proposed for the finite population median in the presence of an auxiliary variable .These estimators are applicable in situation where only a grouped frequency distribution of the auxiliary variable is known.These estimators can be expressed as follows:

A- ***The Position Estimator:***

h		$x \le M_{xh}$	$x \ge M_{xh}$
1	$Y \le M_{y1}$	P_{111}	P_{121}
	$Y > M_{y1}$	P_{211}	P_{221}
2	$Y \le M_{y2}$	P_{112}	P_{122}
	$Y > M_{y2}$	P_{212}	P_{222}
		$P_{.11}$	$P_{.21}$
L	$Y \le M_{yL}$	P_{11L}	P_{12L}
	$Y > M_{yL}$	P_{21L}	P_{22L}
		$P_{.1L}$	$P_{.2L}$

where for instance P_{11h} denotes the proportion of units in the population in the stratum h with $X_h \le M_{xh}$ and $Y_h \le M_{yh}$.

Let n_{xh} be the number of units in the sample with $X_h \le M_{xh}$. then if P_{ijh} are known we can estimate the proportion of the Y's in the srata h of the sample that are less than or equal to M_{yh} which is denoted by P_h as suggested for simple random sample by Kuk and Mak (1989) as follows:

$$\hat{P}_h = n_h^{-1}\{n_{xh} P_{11h}/P_{.1h} + (n_h - n_{xh}) P_{12h}/P_{.2h}\}$$

Since we are dealing with the stratified sample, the appropriate estimator is:

$$\hat{P}_{st} = \sum_{h=1}^{L} \frac{N_h \hat{P}_h}{N}$$

Thus My is approximately the sample Pth quantile $\hat{Q}_Y(P_{st})$. Therefore, we propose an estimator of My which is given by: $\hat{M}_{YP} = \hat{Q}_Y(P_{st})$

This estimator is called a position estimator.

B- The Stratified Estimator

The cross-classification used in constructing $\hat{M}_{YP}$ motivates another way of estimating My as follows:

Let $\tilde{F}_{Y1h}(Y)$ be the proportion among those units in the sample with $X_h \le M_{xh}$ that have Y_h values less than or equal to Y.

Similarly, $\tilde{F}_{Y2h}(Y)$ is the proportion among those units with $X_h > M_{xh}$. Then

$\tilde{F}_Y(Y)$ can be estimated as suggested for the simple random sample by Kuk and Mak (1989) as follows:

$$\tilde{F}_{Yh}(Y) = N_h^{-1} N_{xh} \tilde{F}_{Y1h}(y) + N_h^{-1}(N_h - N_{xh}) \tilde{F}_{Y2h}(y)$$

where N_{xh} is the number of units in the population with $X_h \le M_{xh}$. The estimate which is appropriate for stratified random sample is:

$$\tilde{F}_Y(Y) = \sum_{h=1}^{L} \frac{N_h}{N} \tilde{F}_{Yh}(y)$$

Thus, an estimator of My is given by:

$$\hat{M}_{Ys} = \inf\{Y : \tilde{F}_Y(y) \ge \frac{1}{2}\}$$

This estimator is called a stratified estimator.

Asymptotic Distributions of Estimators:

To derive the asymptotic distribution of $\hat{M}_{yp}$, we will use the same equation for simple random sample used by Kuk and Mak (1989) after modifying it for stratified sample as follows:

$$\hat{M}_{yp} - M_y = \left\{f_y(M_y)\right\}^{-1}\left\{\hat{F}_y(\hat{M}_{yp}) - F_y(M_y)\right\} + O_p(n^{-\frac{1}{2}})$$

$$\hat{M}_{yp} - M_y = \left\{f_y(M_y)\right\}^{-1}\left\{\hat{P}_{st} - P_{st}\right\} + O_p(n^{-1})$$

Then $$E\left\{\hat{M}_{yp} - M_y\right\}^2 = \left\{f_y(M_y)\right\}^{-2} E\left\{\hat{P}_{st} - P_{st}\right\}^2$$

Where $$E\left\{\hat{P}_{st} - P_{st}\right\}^2 = \sum_{h=1}^{L}\left\{\frac{N_h}{N}\right\}^2 E\left\{\hat{P}_h - P_h\right\}^2$$

and $$E\left\{\hat{P}_h - P_h\right\}^2 = 2(1-f_h)P_{1|h}(1-2P_{1|h})n_h^{-1}$$

Then

$$E\left\{\hat{M}_{yp} - M_y\right\}^2 = \left\{f_y(M_y)\right\}^{-2}\sum_{h=1}^{L}2\left\{\frac{N_h}{N}\right\}^2 \times (1-f_h)P_{1|h}(1-2P_{1|h})n_h^{-1}$$

Hence the position estimator is asymptotically normal with mean M_y and variance:

$$V(\hat{M}_{yp}) = \left\{f_y(M_y)\right\}^{-2}\sum_{h=1}^{L}2\left\{\frac{N_h}{N}\right\}^2 (1-f_h)P_{1|h}(1-2P_{1|h})n_h^{-1}$$

To derive the asymptotic variance of the stratified estimator, we will use the same equation used by Kuk and Mak (1989) after modifying it for stratified sample as follows:

$$E\left\{\hat{M}_{ys} - M_y\right\}^2 = \left\{f_y(M_y)\right\}^{-2}\sum_{h=1}^{L}\left\{\frac{N_h}{N}\right\}^2 Var\{\tilde{F}_{yh}(M_y)\}$$

$$E\left\{\hat{M}_{ys} - M_y\right\}^2 = \left\{f_y(M_y)\right\}^{-2}\sum_{h=1}^{L}2\left\{\frac{N_h}{N}\right\}^2 \otimes (1-f_h)P_{1|h}(1-2P_{1|h})n_h^{-1}$$

Numerical Examples:

In the following, two numerical examples from sampling literature are used to illustrate the gain in precision for the proposed estimators. In the first example, we are choosing among the ratio esimator, product type of estimator and the mean per element estimator for each stratum according to criteria cited in Hussein (1995).

Then, we are calculating the generalized estimator for stratified sample, which has the least mean square error among stratified ratio estimator, stratified product type estimator and stratified sample mean. In the second example, we have highly skewe population where the median is more appropriate measure of location than the mean Here, we are comparing among the median estimator in simple random sample and the two proposed estimators in stratified sample. As expected, stratification and using the auxiliary information increased the precision of the two proposed estimators.

Example (1):

Manufacturing companies for specific industry wish to estimate the total numbe of man-hours lost because of sickness.The population consists of two companies .Company A has 1000 employees while company B has 1500 employees. Simple random sample of 10 observations from each company is chosen . Number of man-hour lost in current year(y)and number of man hour lost in previous year (x) are collected for each sample unit. It is assumed that number of man-hour lost in previous year is 16300 for company A and 12800 for company B. This example is used by (Scheaffer, Mendenhall,and Ott,1983) for stratified ratio estimator. Applying the criteria for best estimator to each stratum (Hussein, 1995), We conclude:

For stratum 1:

$$0.01 = g_1 < \frac{2C_{yx1}}{C_{xx1}} = 1.9385 > 1 + g_1 = 1.01$$

For stratum 2:

$$0.0067 = g_2 < \frac{2C_{yx2}}{C_{xx2}} = 0.667 < 1 + g_2 = 1.006$$

So, the ratio estimator is the best estimator in Stratum 1, and the alternative produ - type estimator is the best estimator in Stratum 2. Therefore, the best estimator fo the total population can be expressed as follows:

$$\hat{Y}_s = \left\{\hat{Y}_1 \quad \hat{Y}_2\right\}\left\{\begin{array}{c} 1 \\ \hline H_1(x) \\ 1+g_2-g_2H_2(x) \end{array}\right\} = 24935.13$$

Table (1) shows the gain in precision from using the general estimator in stratified sample compared with other estimators.

Table (1)

Gain In precision from using the general Estimator in stratified sample

The Sample Design	Estimator	Estimate	√ MSE	Relative Efficiency
Stratified Sample	$\bar{Y}$	26500	34.43	100.00
Stratified Sample	$\bar{Y}_{as}$	26489.1	34.024	100.5
Stratified Sample	$\bar{Y}_{rs}$	26629.6	27.74	124.12
Stratified Sample	$\bar{Y}_s$	24935.1	25.21	136.6

Example (2):

Numbers of inhabitants in thousands of 64 large cities in the United states were obtained in 1920 (x) and 1930 (y). The cities are arranged in to strata the first containing the 16 large cities and the second the remaining 48 cities. Table (1) shows the population characteristics of the survey variable. The values of the skewness measure shows that the survey variable has a skewed distribution and the median is more appropriate measure of location than the mean. This example was used by Cochran (1977) to illustrate the gain in precision from the stratified sample. Table (2) illustrates the gain in precision from using the position and stratified median estimators.

Table (2)

population characteristics of the dependent variable

Number of observations	*64*	*16*	*48*
Mean	*305.75*	*629.375*	*197.875*
Median	*253.00*	*575.500*	*169.500*
Standard Deviation	*229.016*	*232.041*	*74.706*
Interquartile Range	*222.5*	*346*	*136.5*
Skeweness	*1.904*	*1.257*	*.449*
Correlation	*.939*	*.764*	*.909*
P_{11}	*.4687*	*.4375*	*.4583*

Table (3)

Gain In precision from using the position and stratified median estimators

The Sample Design	Estimator	Estimate	Standard Error	Relative Efficiency
Simple Random Sample	The Mean	278.78	23.66	100.00
Simple Random Sample	The Median	205.00	2.02	1173.6
Stratified Sample	Position Estimator	204.63	0.047	50340.4
Stratified Sample	Stratified Estimator	212.04	0.047	50340.4

References:

1-***Bandyopadhyay, S.*** (1980), Improved Ratio and Product Estimators, Sankhya, 2, series c, pt.1 and 2, pp.45-49.

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

3-***Chang D.*** S. (1986), The Asymptotic Distribution of Multivariate Product Estimator, Chinese Journal of Mathematics, Vol. 14, No. 3.

4-***Hussein, M.A.*** (1986), On Increasing The Precision of Finite Population Estimators Using Auxiliary Variable ,Unpublished Ph.D. Thesis, University of Iowa, USA.

5- ***Hussein, M. A.*** (1992), Alternative Estimators for Stratified Random Sampling. The Egyptian Statistical Journal, ISSR, Cairo University,Vol. 35, No. 2, 1992.

6- ***Hussein, M. A.*** (1995), An Estimator in Cluster Sample Combined with Stratification. The Egyptian Statistical Journal, ISSR,Cairo University,Vol. 46, No. 2, 1995.

7-***Ouyang, Z., Srivastava, J.N.and Schreuder, H.T.*** (1992), A general Ratio Estimator and Its Application in Model Based Inference, Ann. Inst. Statist. Math. Vol. 45, No. 1, 113-127.

8-***Sekkappan, R.*** M. (1986), Estimation in Sampling from Finite Populations

under the General Linear Regression Model, Journal of Indian statistical Association ,Vol. 24, 91-98.

9- ***Scheaffer, R. L. , Mendenhall , W. and Ott , L.*** *(1983) . Elementary Survey Sampling . Wadsworth , Inc.*

10-***Srivenkataramana, T.***(1980) , Dual to The Ratio Estimator in Sample Surveys, Biometrika, 67,1, pp. 193-204.