The Sample Mean Under Stratified Random Sampling

Arturo Y. Pacificador, Jr.[1]

ABSTRACT

This paper presents some results on the design-based properties of the simple arithmetic mean from a stratified random sample in the estimation of the population mean. Such estimator is often used when the stratified sample is treated as a simple random sample. Such estimator is generally biased.

KEYWORDS: Stratified random sampling, sample mean, design-based estimator.

1. Introduction

Consider a finite and identifiable population of N units. Further suppose that this population can be characterized by some parametric function such as the population mean whose value is assumed to be unknown and is the object of inference. Survey sampling theory in its classical formulation is concerned with the choice of an appropriate sampling strategy for purposes of estimating unknown parametric functions (Chaurhuri & Vos, 1988). By sampling strategy, we refer to the specification of (1) the sampling design (p) - a rule or function p defined over the set S, the set of possible samples, such that it satisfies

(i) $p(s) \geq 0$ *for all* $s \in S$; *and,*

(ii) $\sum_{\forall s} p(s) = 1$

and; (2) the specification of the estimator (t). The performance of a strategy $H=(p,t)$ is assessed in terms of its two characteristics namely: $E_p(t) = \sum_{\forall s} t(s,y)p(s)$, the design-expectation; and $M_p(t) = E_p(t-T)^2 = \sum_{\forall s} (t(s,y) - T)^2 p(s)$, the design-mean squared error of t. The quantity $B_p(t) = E_p(t-T)$ is called the design-bias of t. Among all strategies, the preferred ones are those with controlled magnitudes of the bias and mean squared error of t.

One such sampling design is *stratified sampling* and is described as follows (Cochran, 1977):

(a) First, the population of N units are divided into subpopulations called *strata* of N_1, N_2, ..., N_L units, respectively.

[1] Associate Professor and Director Institute of Mathematical Sciences and Physics, UP Los Baños

(b) These subpopulations are nonoverlapping and together they comprise the whole population such that

$$N_1 + N_2 + \ldots + N_L = N$$

(it is assumed in here that N_h is known)

(c) When the strata have been determined, a sample is drawn from each without any restriction of sampling design used (i.e. a different sampling design may be used in the selection of units in each stratum). However, for this paper, we assume that simple random sampling is used in each stratum. Further, the sample selection are made independently in the different strata.

(d) The sample sizes within strata are denoted by n_1, n_2, . . . , n_L, respectively. It is assumed in here that the sample sizes are predetermined and fixed prior to the selection of units in each stratum.

This paper presents the design-based properties of the sample mean under stratified random sampling.

2. Notations

Basically, the notations adopted by Cochran (1977) will be used all throughout the paper. Let the suffix *h* denote the stratum and *i* the unit within a stratum. In particular let

N_h	total number of units in the hth stratum
n_h	sample size in the hth stratum
L	number of strata formed
$W_h = \frac{N_h}{N}$	stratum weight
$\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi}$	true mean for the hth stratum
$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$	sample mean for the hth stratum
$S_h^2 = \frac{\sum_{i=1}^{N_h}(y_{hi} - \bar{Y}_h)^2}{N_h - 1}$	true variance for the hth stratum
$\bar{Y} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h} y_{hi} = \sum_{h=1}^{L} W_h \bar{Y}_h$	true mean for the entire population and is the object of estimation in this paper.

3. Estimation of the Mean

Under stratified (simple) random sampling, an unbiased estimator of the population mean denoted by $\bar{y}_{st}$, , is defined as

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{1}$$

and its variance (assuming that the N_h's are large) is

$$Var(\bar{y}_{st}) = \sum W_h^2 \frac{S_h^2}{n_h} \tag{2}$$

Details of derivations and proofs are shown in Cochran (1977) pp. 91-93.

Another estimator of the population mean can be defined as the simple arithmetic mean of the entire stratified sample. That is,

$$\begin{aligned} \bar{y} &= \frac{1}{n}\sum_{h=1}^{L}\sum_{i=1}^{n_h} y_{hi} \\ &= \frac{1}{n}\sum_{h=1}^{L} n_h \bar{y}_h \qquad (3) \\ &= \sum_{h=1}^{L} w_h \bar{y}_h \end{aligned}$$

where $w_h = n_h / n$

Such estimator is at times employed in practice especially in situations where a stratified sample is treated as a simple random sample.

4. Design Based Properties of $\bar{y}$.

A. Unbiasedness

The design-based expectation of $\bar{y}$, denoted by $E_p(\bar{y})$, is given by

$$E_p(\bar{y}) = E_p\left(\sum_{h=1}^{L} w_h \bar{y}_h\right) = \sum_{h=1}^{L} E_p(w_h \bar{y}_h) \tag{4}$$

Assuming that the sample size in each stratum , n_h, is predetermined prior to the actual selection of samples and SRS is used in the selection of samples in each stratum, then

$$E_p(\bar{y}) = \sum_{h=1}^{L} w_h \bar{Y}_h \qquad (5)$$

A closer look into (5) shows that:

(a) Unless $w_h = W_h \quad \forall h$, $\bar{y}$ is generally a biased estimator of $\bar{Y}$.

(b) If the stratum means are equal (say $\bar{Y}_h = \bar{Y}_1 \quad \forall h$), then

$$E_p(\bar{y}) = \sum_{h=1}^{L} w_h \bar{Y}_1 = \bar{Y}_1 \quad since \quad \sum_{h=1}^{L} w_h = 1 \qquad (6)$$

and,

$$\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_1 = \bar{Y}_1 \quad since \quad \sum_{h=l}^{L} W_h = 1 \qquad (7)$$

Thus, from the relations given by (6) and (7), $\bar{y}$ is an unbiased estimator of $\bar{Y}$ if the stratum means are equal.

As an additional note, the case of equal stratum means may be realized whenever stratification is not too effective.

Generally, $\bar{y}$ is generally a biased estimator of $\bar{Y}$. While it is desirable to work with unbiased estimators, it must be noted however that biased estimators are not necessarily useless and under certain situations, biased estimators may even be more superior than unbiased estimators (e.g. ratio estimators) specially if the magnitude of the bias is negligible and the estimator is precise leading to an accurate estimator. The magnitude of the bias of $\bar{y}$ as an estimator of $\bar{Y}$ can be measured by $B_p(\bar{y})$ and is defined as

$$\begin{aligned} B_p(\bar{y}) &= E_p(\bar{y}) - \bar{Y} \\ &= \sum_{h=1}^{L} w_h \bar{Y}_h - \sum_{h=1}^{L} W_h \bar{Y}_h \qquad (8) \\ &= \sum_{h=1}^{L} (w_h - W_h) \bar{Y}_h \end{aligned}$$

Note that (8) is a function of (a) the difference between w_h and W_h and (b) the true mean of the h<u>th</u> stratum. The expression also shows that it is independent of the overall sample size *n*. It would be small if (a) the difference between w_h and W_h is small for all stratum; or, (b) said difference is small for stratum having large means. No generalizations however can be made as to the direction of the bias of $\bar{y}$.

1. $B_p(\bar{y})$ under equal allocation.

If the sample size n is to be allocated equally for each stratum, then

$$n_h = \frac{n}{L} \Rightarrow w_h = \frac{1}{L} \tag{9}$$

Thus, from (9), (8) can be expressed as

$$B_p(\bar{y}) = \sum_{h=1}^{L} (\frac{1}{L} - W_h)\bar{Y}_h \tag{10}$$

(10) would be equal to zero (i.e., $\bar{y}$ is an unbiased estimator of $\bar{Y}$) whenever ($1/L$) equals the stratum weight W_h. in all strata. This would be possible if the stratum weights are the same (in which case it is equal to (1/L)) or the stratum sizes are equal. That is

$$\begin{aligned} &If \quad W_h = W_1 \quad (say) \qquad \forall h \\ &then \\ &\sum_{h=1}^{L} W_h = \sum_{h=1}^{L} W_1 = 1 \\ &\qquad = W_1 L = 1 \qquad \Rightarrow \qquad W_h = W_1 = \frac{1}{L} \end{aligned}$$

Hence, under equal allocation, $\bar{y}$ is unbiased for $\bar{Y}$ if the stratum sizes are equal.

2. $B_p(\bar{y})$ *under proportional allocation*

Under proportional allocation, the sample size allocated for each stratum is determined by the following relation

$$w_h = W_h \quad \Rightarrow \quad n_h = nW_h \tag{11}$$

Obviously, under this allocation scheme, $\bar{y}$ is unbiased for $\bar{Y}$ and likewise, $\bar{y} = \bar{y}_{st}$.

3. $B_p(\bar{y})$ *under optimum (Neyman's) allocation*

Under this allocation scheme which is due to Neyman (1934) and Tschuprow (1923), the sample size to be allocated in the hth stratum is determined by minimizing the variance of the mean, $\bar{y}_{st}$, given a linear cost function (Cochran,1977). Assuming that the cost of obtaining measurements for each unit is the same for all stratum, then the sample size for each stratum is given by

$$n_h = n\frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} \qquad (12)$$

Note that under this allocation scheme, if the stratum variances, S_h^2, are equal, then (12) is equivalent to proportional allocation. Thus, under this scheme and with equal stratum variances $\bar{y}$ is an unbiased estimator of $\overline{Y}$.

Table 1 presents a summary of the unbiasedness property of $\bar{y}$ under three allocation schemes considered and under certain conditions.

Table 1: Summary of unbiasedness property of $\bar{y}$ under three allocation schemes and three conditions considered.

	CONDITIONS		
ALLOCATION SCHEME	EQUAL STRATUM SIZES	EQUAL STRATUM MEANS	EQUAL STRATUM VARIANCES
Equal	**Unbiased**	**Unbiased**	**No generalizations**
Proportional	**Unbiased**	**Unbiased**	**Unbiased**
Optimum	**No generalizations**	**No generalizations**	**Unbiased**

B. Variance and Mean Squared Error (MSE)

The variance and mean squared error (MSE) are two measures that assesses the precision and accuracy of estimators in general. Together with the bias, the standard error (positive square root of the variance), an assessment of whether the bias is negligible or not can be made. For instance, if the bias of an estimator is less than one-tenth of its standard error then it is considered negligible (almost an unbiased estimator) (Cochran,1977).

The design-based (or true) variance of $\bar{y}$, denoted by $Var_p(\bar{y})$, is (assuming large N_h)

$$Var_p(\bar{y}) = Var_p(\sum_{h=1}^{L} w_h \bar{y}_h)$$

$$= \sum_{h=1}^{L} w_h^2 \, Var_p(\bar{y}_h) \qquad (13)$$

$$= \sum_{h=1}^{L} w_h^2 \frac{S_h^2}{n_h} \qquad (13)$$

From (13), it can be noted that $\bar{y}$ becomes more precise whenever the sample size n increases. In addition, (13) may be minimized if the sample size in a stratum with large $Var_p(\bar{y}_h)$ is chosen so as to make w_h as small as possible. The difference (or ratio) between (13) and (2) may be used as an indicator of the precision (or loss of it) of $\bar{y}$ as compared to $\bar{y}_{st}$. The said difference is given as

$$Var_p(\bar{y}) - Var_p(\bar{y}_{st}) = \sum_{h=1}^{L}(w_h^2 - W_h^2)\frac{S_h^2}{n_h} \tag{14}$$

Note that from (14), if $w_h = W_h$ (in all strata) then the estimator $\bar{y}$ is as precise as $\bar{y}_{st}$. This condition is realized under proportionate stratified sampling in which case $\bar{y} = \bar{y}_{st}$. Outside of this condition, the magnitude and direction of said difference is dependent on the difference between w_h^2 and W_h^2 as well as the variance of the sample mean in the hth stratum. Note that the difference between w_h^2 and W_h^2 in each stratum may either be positive or negative and is very important in the determination of the magnitude as well as the direction of the overall difference. For instance if the sum of the negative terms in the difference is larger than the sum of the positive terms, then $\bar{y}$ is more precise than $\bar{y}_{st}$ and hence as far as precision is concerned, nothing was lost when stratification was employed. Such a scenario is very possible.

However, since $\bar{y}$ is generally a biased estimator of $\bar{Y}$, a more appropriate measure in comparing $\bar{y}$ with $\bar{y}_{st}$ would be the mean squared error (MSE) [5]. The MSE of $\bar{y}$, is defined as

$$\begin{aligned} M_p(\bar{y}) &= Var_p(\bar{y}) + \{Bias_p(\bar{y})\}^2 \\ &= \sum_{h=1}^{L} w_h^2\frac{S_h^2}{n_h} + \left\{\sum_{h=1}^{L}(w_h - W_h)\bar{Y}_h\right\}^2 \end{aligned} \tag{15}$$

Note that from (15), as the sample size increases, the first component of (15) (variance term) decreases however the second component (bias term) will remain unchanged unless the increase in the sample would result into a change in w_h however this would not guarantee that the bias term would decrease (it may in fact increase). Comparing (15) with (2) (which is also the MSE of $\bar{y}_{st}$ being an unbiased estimator of $\bar{Y}$), the following difference function can be defined

$$MSE_p(\bar{y}) - MSE_p(\bar{y}_{st}) = \sum_{h=1}^{L}(w_h^2 - W_h^2)\frac{S_h^2}{n_h} + \left\{\sum_{h=1}^{L}(w_h - W_h)\bar{Y}_h\right\}^2 \tag{16}$$

From (16), it can be noted that even if $\bar{y}$ is more precise than $\bar{y}_{st}$ (the first term being less than zero), it is possible that $\bar{y}$ will be less accurate than $\bar{y}_{st}$, specially if the bias term (2nd component) cannot be ignored and may even be larger than first term. Thus while no precision may be lost (even gaining some) when stratification is employed, there might be some loss in accuracy as a result of ignoring stratification.

5. Illustration

The results derived is illustrated for the case of an actual population. Because of data confidentiality, the actual population used in this study will not be described in detail. Only descriptive information will be presented to characterize the population used for this study. The entire population was divided into three strata using the Mahalonobis' rule of equalization of stratum totals. The resulting descriptive measures (mean, CV) as a result of stratification and for the entire population is are presented in Table 2.

Table 2: Descriptive measures of the actual population as a result of stratification.

STRATUM	WEIGHT	MEAN	C.V.(%)
1	0.54	4,654	55.1
2	0.27	12,964	36.0
3	0.19	40,487	52.9
Entire Pop'n	1.00	13,679	120.2

It can be noted from Table 2 that the resulting stratification can be considered efficient since it was able to increase the variability between strata (as seen from the comparison between stratum means) and reduce the variation within strata. Thus, it is expected that stratification in this case would improve the efficiency of the estimates.

For this study, all possible one-decimal values for the sample weights such that the sum of these weights equals one were considered. The relative bias (bias/true mean) (%) denoted by rel. bias, the ratio of the bias and standard error of $\bar{y}$ denoted by |b|/s, and the ratio of the MSE of $\bar{y}$ with the Var of $\bar{y}_{st}$ denoted by mse/vyst, were computed under different combinations of the "sample" weight and for varying sample sizes *n=50, 100, and 150* . The results are presented in Table 3.

Note that the results in Table 3 shows that when stratification is efficient as in this case, practically ignoring stratification leads to a great loss in accuracy and even made more pronounced when the sample size is increased. More illustrations were made for the case of 2 strata by Pacificador (1995).

Table 3: Example for actual population.

							n=50		n=100		n=150	
w(1)	w(2)	w(3)	W(1)	W(2)	W(3)	rel bias	|b|/s	mse/vyst	|b|/s	mse/vyst	|b|/s	mse/vyst
0.1	0.1	0.8	0.54	0.27	0.19	149.66	7.53	383.25	10.65	759.86	13.04	1136.46
0.1	0.2	0.7	0.54	0.27	0.19	129.54	6.94	316.59	9.81	626.74	12.02	936.88
0.1	0.3	0.6	0.54	0.27	0.19	109.42	6.30	221.36	8.91	437.29	10.91	653.21
0.1	0.4	0.5	0.54	0.27	0.19	89.30	5.59	137.43	7.91	270.60	9.68	403.77
0.1	0.5	0.4	0.54	0.27	0.19	69.18	4.79	73.65	6.78	144.23	8.30	214.81
0.1	0.6	0.3	0.54	0.27	0.19	49.06	3.86	31.38	5.46	60.78	6.68	90.17
0.1	0.7	0.2	0.54	0.27	0.19	28.94	2.70	8.60	3.82	16.17	4.67	23.73
0.1	0.8	0.1	0.54	0.27	0.19	8.82	1.07	0.74	1.51	1.13	1.85	1.52
0.2	0.1	0.7	0.54	0.27	0.19	123.47	6.63	297.86	9.37	589.10	11.48	880.34
0.2	0.2	0.6	0.54	0.27	0.19	103.35	5.96	228.68	8.43	451.10	10.33	673.53
0.2	0.3	0.5	0.54	0.27	0.19	83.23	5.23	140.69	7.39	276.42	9.05	412.14
0.2	0.4	0.4	0.54	0.27	0.19	63.11	4.39	71.74	6.21	139.94	7.60	208.14
0.2	0.5	0.3	0.54	0.27	0.19	42.98	3.40	27.82	4.80	53.42	5.88	79.03
0.2	0.6	0.2	0.54	0.27	0.19	22.86	2.15	6.31	3.04	11.49	3.72	16.68
0.2	0.7	0.1	0.54	0.27	0.19	2.74	0.34	0.39	0.47	0.43	0.58	0.47
0.3	0.1	0.6	0.54	0.27	0.19	97.27	5.63	183.81	7.96	362.00	9.75	540.19
0.3	0.2	0.5	0.54	0.27	0.19	77.15	4.86	122.99	6.87	240.98	8.42	358.97
0.3	0.3	0.4	0.54	0.27	0.19	57.03	3.98	61.32	5.63	119.00	6.90	176.68
0.3	0.4	0.3	0.54	0.27	0.19	36.91	2.93	21.85	4.15	41.43	5.08	61.00
0.3	0.5	0.2	0.54	0.27	0.19	16.79	1.59	4.02	2.25	6.89	2.75	9.77
0.3	0.6	0.1	0.54	0.27	0.19	-3.33	0.41	0.41	0.58	0.47	0.72	0.53
0.4	0.1	0.5	0.54	0.27	0.19	71.08	4.49	92.63	6.35	180.89	7.78	269.15
0.4	0.2	0.4	0.54	0.27	0.19	50.95	3.57	48.69	5.05	93.83	6.19	138.98
0.4	0.3	0.3	0.54	0.27	0.19	30.83	2.46	15.98	3.48	29.70	4.26	43.42
0.4	0.4	0.2	0.54	0.27	0.19	10.71	1.02	2.32	1.44	3.50	1.77	4.68
0.4	0.5	0.1	0.54	0.27	0.19	-9.41	1.18	0.83	1.67	1.31	2.05	1.79
0.5	0.1	0.4	0.54	0.27	0.19	44.88	3.16	34.10	4.47	65.10	5.47	96.09
0.5	0.2	0.3	0.54	0.27	0.19	24.76	1.99	10.80	2.81	19.43	3.44	28.05
0.5	0.3	0.2	0.54	0.27	0.19	4.64	0.45	1.33	0.63	1.55	0.77	1.77
0.5	0.4	0.1	0.54	0.27	0.19	-15.48	1.97	1.64	2.78	2.94	3.41	4.25
0.6	0.1	0.3	0.54	0.27	0.19	18.68	1.51	6.37	2.13	10.79	2.61	15.22
0.6	0.2	0.2	0.54	0.27	0.19	-1.44	0.14	1.09	0.20	1.11	0.24	1.14
0.6	0.3	0.1	0.54	0.27	0.19	-21.56	2.78	2.84	3.93	5.36	4.81	7.88
0.7	0.1	0.2	0.54	0.27	0.19	-7.51	0.73	1.50	1.04	2.03	1.27	2.55
0.7	0.2	0.1	0.54	0.27	0.19	-27.63	3.61	4.40	5.10	8.48	6.25	12.57
0.8	0.1	0.1	0.54	0.27	0.19	-33.71	4.47	6.12	6.32	11.94	7.73	17.77

6. Summary and Conclusion

The simple arithmetic mean as estimator of the population mean under stratified sampling is generally biased and its bias is dependent on the differences between w_h and W_h and stratum means except when samples are allocated proportionately across strata. Particular allocation schemes likewise provide situations under which the sample mean is unbiased such as: (1) equal stratum means or equal stratum sizes under equal allocation; and, (2) equal stratum variances under optimum allocation.

these information is to be used for planning purposes, then one may not afford ignoring stratification.

The study also shows that before stratification is to be ignored, some exploratory efforts must be done to ensure that ignoring it may not unnecessarily lead to a loss in accuracy.

It would be interesting to look into the effects of ignoring other sampling designs commonly used such as systematic sampling and multi-stage sampling both for descriptive and analytic uses of survey data. In relation to stratified sampling, the study has not yet dealt with the problem of variance estimation and it is hoped that future improvements of this study will include such area.

7. References

Cochran, W.G. 1977. *Sampling Techniques*. 3rd ed., John Wiley & Sons, USA.

Chaudhuri,A. And J.W.E.Vos. 1988. *Unified Theory and Strategies of Survey Sampling. North Holland Series in Statistics and Probability.*

Pacificador,A.Y.Jr. 1995. *Effects of Ignoring Sampling Design: The Stratified Random Sampling Case.* UPLB Faculty, Staff, and Students Professorial Chair Paper in Statistics. UP Los Baños, College, Laguna.

8. Acknowledgments

The author would like to thank the enlightening comments and views of the referees.