The Effect of Stratified Sampling with proportional Allocation on Inference about Population Mean

Temesgen Zewotir*

A proper analysis of survey data requires that sampling design be taken into account, when conclusions are wanted about finite population. However, many computer programs for standard statistical analysis implicitly assumes simple random sampling. Accordingly, this study undertakes to examine the effect of proportional stratified sampling on the common type of simple statistical analysis. The result of the study indicates that the usual use of standard procedures can lead to erroneous variance estimate and therefore invalid conclusions.

Introduction

The ultimate goal of sampling is statistical inference. That is, estimation of the characteristics of the population and making decisions on the characteristics of the population. Estimates that are unbiased are a desired features of sampling plan. On the other hand, a plan that gives a small bias is not ruled out of consideration if it has other attractive features. To mention the simplest case, ratio estimate under simple random sampling is a popular estimate which is biased with appealing precision, (see Cochran, 1977; Barnett, 1991). Smaller variance (or higher precision) is an attractive feature of an estimator.

In social, business, economic and political studies stratified random sampling technique is more widely used than simple random sampling technique. The reason is that stratified sampling increases precision, ensures adequate representation and creates administrative convenience (see Hansen, Hurwitz and Madow, 1953; Chaudhuri, 1992; and Tryfus, 1996). Particularly, proportional allocation of the sample is the most widely used stratified sampling technique. Because, it requires only the knowledge of stratum size; but besides the stratum size Neyman allocation and optimum allocation require the knowledge of stratum variance and stratum cost (Godfrey, Roshwalb and Wright, 1984).

* Mr. Temesgen Zwetoir is a Lecturer in the Department of Statistics and Demography at the National University of Lesotho.

Since the stratified sample is selected randomly, albeit in different fashion from simple random sample, the methods of population characteristics estimation are also different. The statistical properties of estimators, for the same population parameter, obtained from stratified and simple random sampling techniques may not be identical; that is, in terms of bias, precision, cost, simplicity, and so on. The theories related to this have been discussed in the literatures (Hansen, Hurwitz and Madow, 1953; Kish 1965; Cochran, 1977; Sarndal, Swensson and Wretman, 1992; and Tryfos, 1996).

In most practical cases, however, the data collected with stratified sampling technique is analysed as if the data were from simple random sampling design. The primary reason is that the methods of analysis in simple random sampling is popular and simple. The second reason is that statistical analysis are readily available and easily invoked in many commercial statistical package; and most statistical packages provide testes of hypothesis and estimations about various parameters with the assumption of simple random sampling technique. In other words the analysis will be done as if the sample comes from simple random sample where samples came from simple stratified random sampling technique.

The objective of this paper is, therefore, to assess the loss or gain, if any, in statistical inference about population mean if the sampling technique is considered as simple random sampling when the actual sampling technique was stratified sampling with proportional allocation.

Basic Formulations

In stratified sampling the population of size N is first divided into non-overlapping L strata of size N_1, N_2, . . ., N_L, respectively. Such that an independent simple random sample of size n_1, n_2,..., n_L (where $n_h \geq 2$, $h = 1, 2, \ldots L$), respectively, will be taken from each stratum as if the stratum were a population in its own right. So that the various samples are lumped together to form a single sample in the population. That is, $n = \Sigma n_h$ and $N = \Sigma N_h$. In proportional stratified sampling the size of the sample from stratum h, n_h, is in the same proportion to the sample size, n, as N_h is to N, that is, $n_h = (N_h/N)n = W_h n$.

Since the most practical approach of sampling is sampling without replacement all considerations in this paper are sampling without replacement. The

estimator of the population mean is then:

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

$$\text{where,} \qquad \bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$$

$$var(\hat{\mu}_{st}) = \frac{(1-f)}{n} \sum_{h=1}^{L} W_h \sigma_h^2 \tag{1b}$$

where f=n/N

$$\sigma_h^2 = \frac{\sum_{i=1}^{N_h} (Y_{hi} - \mu_h)^2}{N_h - 1}$$

$$\mu_h = \frac{\sum_{i=1}^{N_h} Y_{hi}}{N_h}$$

In practice, σ_h^2 is unknown and, therfore, the unbiased sample estimate is

$$var(\hat{\mu}_{st}) = \frac{(1-f)}{n} \sum_{h=1}^{L} W_h s_h^2 \tag{1c}$$

$$\text{where,} \qquad s_h^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1}$$

The $(1-\alpha)100\%$ confidence interval for μ under the assumption the distribution of (1a) is normally distributed and (1c) well determined is given by

$$\hat{\mu}_{st} \pm t_{\frac{\alpha}{2}}(n_e)\, se(\hat{\mu}_{st}) \tag{2}$$

As Cochran (1977) noted the exact distribution of (1c) along with the formula for effective degrees of freedom, n_e, is considered complicated and complex. Most sampling texts use Z instead of t with the assumption the sample size in each stratum is large. It is Cochran (1977) who noted the effective degrees of freedom and student's t-distribution.

Under the assumption that y_{hi} are normal the simplified derivation is as follows:

For mean square, MS, obtained from a normal distribution, and corresponding degrees of freedom, df, the following holds true.

$$\frac{df.MS}{E(MS)} \sim \chi^2(df) \quad \text{whereupon} \quad var\left(\frac{df.MS}{E(MS)}\right) = 2df \tag{3}$$

for details refer Searle, Casella and McCulloch (1992).

Let

$$X = var(\hat{\mu}_{st}), \qquad g_h = \frac{1-f}{n} W_h$$

clearly, X is a mean square derived from normal distribution. Therefore,

$$\frac{n_e \cdot X}{E(X)} \sim \chi^2(n_e) \quad hence \quad var\left(\frac{n_e \cdot X}{E(X)}\right) = 2n_e \tag{4}$$

$$E(X) = \sum_{h=1}^{L} g_h \sigma_h^2 \tag{5}$$

$$var(X) = \sum_{h=1}^{L} g_h^2 var(s_h^2) = \sum_{h=1}^{L} g_h^2 \frac{\sigma_h^4}{n_h - 1} \tag{6}$$

Applying (4) in (5) and (6) we get

$$n_e = \frac{[\sum_{h=1}^{L} g_h \sigma_h^2]^2}{\sum_{h=1}^{L} g_h^2 \frac{\sigma_h^4}{n_h - 1}} \tag{7}$$

The sample estimate is

$$n_e = \frac{[\sum_{h=1}^{L} g_h s_h^2]^2}{\sum_{h=1}^{L} g_h^2 \frac{s_h^4}{n_h - 1}} \tag{8}$$

Furthermore, it can be shown that

$$\min_{h=1}^{L}(n_h - 1) \leq n_e \leq \sum_{h=1}^{L} (n_h - 1) \tag{9}$$

Likewise to test a hypothesis that

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0 \tag{10}$$

the test statistic is

$$t = \frac{\hat{\mu}_{st} - \mu_0}{se(\hat{\mu}_{st})} \tag{11}$$

which is distributed as student's t-distribution with degrees of freedom n_e.

If a simple random sample of size n was drawn from the same population, the unbiased estimator of μ would be

$$\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{12a}$$

and

$$var(\hat{\mu}) = \frac{(1-f)}{n}\sigma^2 \tag{12b}$$

$$where, \qquad \sigma^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N-1} \tag{12c}$$

The unbiased sample estimate of (12b) is

$$var(\hat{\mu}) = \frac{(1-f)}{n} s^2 \tag{13a}$$

$$where, \qquad s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} \tag{13b}$$

Consequently the 100(1-α)% confidence interval for μ and test statistic for (10) becomes

$$\hat{\mu} \pm t_{\frac{\alpha}{2}}(n-1)\, se(\hat{\mu}) \tag{14}$$

and

$$t = \frac{\hat{\mu} - \mu_0}{se(\hat{\mu})} \sim t(n-1) \tag{15}$$

respectively.

Comparison

If the sampling technique is stratified sampling and the analysis is done as if the data came from simple random sampling, we use (14) instead of (10) and (15) instead of (11). Now the comparison is between (1a) and (12a), t value with degrees of freedom n_e and (n-1), and (1b) and (12b) consequently between (1c) and (13a). Clearly under proportional allocation (12a) is equal to (1a). In fact, rounding $n_h = W_h n$ to the nearest integer may not allow us to get results as mathematically justified. And for a moderate sample size in each stratum, the difference between the theoretical t value at degrees of freedom n_e and (n-1) is not remarkable for fixed small α, (refer student's t-distribution table). Therefore the remaining comparison is between the variances.

From (12c) it easy to derive

$$\sigma^2 = \sum_{h=1}^{L} \frac{N_h - 1}{N-1} \sigma_h^2 + \sum_{h=1}^{L} \frac{N_h}{N-1} (\mu_h - \mu)^2 \tag{16}$$

If N_i and N are large (as likely in practice)

$$\sigma^2 = \sum_{h=1}^{L} W_h \sigma_h^2 + \sum_{h=1}^{L} W_h (\mu_h - \mu)^2 \qquad (17)$$

Hence, from (17), (12b) and (1b)

$$var(\hat{\mu}) = var(\hat{\mu}_{st}) + \frac{1-f}{n} \sum_{h=1}^{L} W_h (\mu_h - \mu)^2 \qquad (18)$$

$var(\hat{\mu}_{st}) = var(\hat{\mu})$ only when all $\mu_h = \mu$, that is when all strata means are equal. In practice stratification is conducted towards homogeneity within stratum and heterogeneity between strata. It is therefore rather a rare case that μ_h will be equal to μ.

From (18), the estimated variance based on the simple random sampling assumption overestimates the actual variance from stratified sampling with proportional allocation. Hence the resulting confidence intervals in (14) are in error. In fact, the confidence interval calculated from the simple random sample (overestimated variance) will have a true confidence level at least equal to the $100(1-\alpha)\%$ aimed at the stratified sampling with proportional allocation. The confidence interval is then conservative. With regard to hypothesis testing, the calculated t value in (15) underestimates the actual t value in (11). Consequently, it leads to accepting the null hypothesis when it is actually false.

Application

We are interested in estimating the average 1981 military expenditure, of 124 countries. Using CO124 population data available in Sarndal, Swensson and Wretman (1992, pp 662-665, Appendix D). This population consists of 124 countries divided into 6 strata: 1 Africa, 2 = Asia (non-Soviet), 3 = Europe(non-Soviet part), 4 = North and Central America, 5 = Oceania and USSR, 6 = South America. To estimate μ, the mean 1981 military expenditure (in millions of US dollars), the total sample size considered is n = 50. The

sample size in each stratum is determined by proportional allocation, after which a random sample (without replacement) is drawn in each stratum. The summary statistics is presented in Table 1.

The statistical analysis when one considers the sample as stratified sampling with proportional allocation (stpa) and when one assumes as a simple random sample (srs) drawn from 124 countries is given in Table 2.

Table 1: Summary Statistics

Stratum	W_h	n_h	Stratum sample variance (s_h^2)	Stratum sample mean
Africa	0.306	15	113239.5	242.933
Asia	0.266	13	36526549.0	2984.231
Europe	0.194	10	88113701	7320.80
N & C America	0.123	6	2907132	747.33
Oceania and USSR	0.032	2	4851613	1950.5
South America	0.089	4	58008.92	299.25

Table 2: Results in the actual and assumed sample design

Design	Estimate	Variance	df	95 % Confidence Interval
Actual (stpa)	2459.406	325833.70	19	(1264.67, 3654.14)
Assumed (srs)	2504.58	32709392.7	49	(-8988.60, 13997.76)

Clearly to test the hypothesis stated earlier, at $\alpha=0.05$, we reject H_0 if μ_0 is less than the lower limit or greater than the upper limit in the confidence limits given in Table 2, otherwise we accept H_0 . Apparently the confidence interval in the assumed design is conservative and hence misleading. The deceptiveness of the assumed sampling design in hypothesis testing about the mean 1981 military expenditure in the 124 countries is also striking.

Conclusion

Inference about the population mean is widely used statistical analysis. In order to have a valid result, however, the analyses should be done in line with the survey design. Under a frequently used stratified sampling, which is proportional allocation, if the statistical analysis is done as if the data came from the simple random sampling the confidence interval will be conservative; and in hypothesis testing, type II error is maximized. That is, erroneously accepting the null hypothesis when it is false is highly tenable.

References

Barnet, V. Sample Survey Principles and Methods, 2nd edn. (London: Oxford University Press, 1991).

Chaudhuri, A. Survey Sampling: Theory and Methods. (New York: Marcel Dekker, 1992).

Cochran, W.G. Sampling Techniques, 3rd edn. (New york: John Wiley, 1977).

Godfrey, J. Roshwalb, A. and Wright, R.L. Model-based stratification in inventory cost estimation. Journal of Business and Economic Statistics 2, 1-9, 1984.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. Sample Survey Methods and Theory, Vol. I and II. (New York: John Wiley, 1953).

Kish, L. Survey Sampling. (New York: John Wiley, 1965).

Sarndal, C.E., Swensson,B. and Wretman, J. Model Assisted Survey Sampling. Springer-Verlag, 1992.

Searle, S.R., Casella, G. and McCulloch, C.E. Variance Components. (New York: John Wiley, 1992).

Tryofos, P. Sampling Methods for Applied Research: Text and cases. (New York: John Wiley, 1996).

Notes to contributors

Contribution to the LSSR from academics and others with specialist knowledge in various fields of the Social Sciences are welcome and should be submitted in English.

1. The length of articles should be 8000 words and must be accompanied by an abstract of not more than 300 words.

2. Articles should be typed in double spacing on A4 paper. Two hard copies must be submitted together with a diskette preferably on Wordperfect 5.1, 6.0 or 6.1.

3. Broad divisions in the text must be indicated by clear headings and sub-headings where appropriate.

4. Maps, diagrams and graphs should be camera-ready and submitted separately.

5. References to books and articles should be identified in the text by the surname author, year of publication and page reference, placed in parenthesis e.g. (Ake 1996:61). Only the year of publication and page are indicated in a case where the author is mentioned in the sentence.

6. If the same author is referred to more than once on books or articles published in the same year, the references are distinguished sequentially e.g. (Ake 1996a; 1966b etc.).

7. Quotations of more than 40 words should be indented and single spaced. Shorter quotations must be indicated by double quotation marks.

8. Endnotes should be used as additional explanatory material to a point referred to in the text. Footnotes are not allowed.

9. Bibliographic references should be placed at the end of the article in alphabetical order. For books indicate author's surname and initials, full title of the book (bold), publisher, place of publication, date of publication e.g. Ake, C. Democracy and Department in Africa, (The Brookings Institution: Washington, 1996). For journal articles Provide author's surname and initials, full title of article, journal, (bold) volume, number and date e.g. Leys, C., "The Crisis in 'Development Theory'", New Political Economy, Vol. 1 No. 1 March, 1996.

10. All articles published in LSSR shall be refereed. Alterations may be made by the referees and editorial board.

11. Data access and estimation procedures: Author(s) should report, describe, and/or reference the complete estimation procedure used to derive the results presented in the manual. This should include data documentation and model specification used but not presented in the manuscript. Author(s) are expected to make available, at cost, all data used to other researchers for replication purposes for a period of 5 years.

12. Author(s) identification: To protect their anonymity during the review process, Author(s) should not identify themselves within the manuscript. Attach a separate page including names of author(s), biographical information as well as title of the manuscript. Include only the title on the first page of the text.

13. Table: Place each table on a separate page; double-space all material omitting vertical rules. LSSR prefers tables created without the use of word-processor table format tools.

14. Figures: Place figures, charts, and graphs at the end of the manuscript each on a separate unnumbered page. Computer generated graphics are preferred. All figures should be camera-ready.

15. Hard & Disk copies: Four hard copies of the manuscript should be send to the editor. In addition, the manuscript should be submitted on 3.5 inch disk in WordPerfect 4.2-6.1 formats.