IMPLICATIONS OF OPTIMUM AND APPROXIMATELY OPTIMUM STRATIFICATION

Dallas W. Anderson, Leslie Kish, Richard G. Cornell

The University of Michigan

The principal reason for stratification in the design of sample surveys is to reduce the variance of sample estimates. The factors that influence the reduction of variance include: choice of stratification variables, number of strata, determination of stratum boundaries, and allocation of the sample.

The purpose of this paper is to quantify gains in precision due to the choice of stratum boundaries. We make the common and somewhat restrictive assumptions that there is only one estimation variable Y and only one stratification variable X. Furthermore, we assume that Y and X have a bivariate normal distribution with correlation ρ between Y and X.

In the context of stratified random sampling, gains in precision will be quantified by comparing the variance of the stratified mean with the variance of the unstratified mean. Two types of sample allocation will be considered: Neyman allocation and proportional allocation.

For each type of sample allocation, the immediate problem is how to choose stratum boundaries. The results of this study will be of greater value if the boundaries are determined in some unique manner, and optimum boundaries seem to be the most reasonable standard. Optimum boundaries minimize the variance of a particular estimator for a given type of sample allocation.

There is an extensive literature on optimum stratification, but unfortunately there are no rules for choosing optimum boundaries that are both explicit and practical. Typically minimal equations are derived which the optimum boundaries must satisfy, but these systems of equations are not easily solved. The thrust of much of the work in the area of optimum stratification has been toward the development of relatively simple rules for determining approximately optimum stratum boundaries.

One of the more widely known rules is the cum. $\sqrt{f}$ rule of Dalenius and Hodges [5]. This rule is based on the set of equations:

$$\int_{-\infty}^{x'_h} \sqrt{f(x)}dx = (h/L) \int_{-\infty}^{\infty} \sqrt{f(x)}dx, \qquad (1)$$

$$h=1,\ldots,L-1,$$

where f is the p.d.f. of the stratification variable and L is the number of strata. The roots $x'_1,\ldots,x'_{L-1}$ are the approximately optimum stratum boundaries.

In the last few years several writers have proposed cum. cube root rules. For example, see Singh [13] or Thomsen [14]. In the setting dealt with in this paper, the cube root rules of Singh and Thomsen coincide and give approximately optimum stratum boundaries $x''_1,\ldots,x''_{L-1}$ which are the roots of the set of equations:

$$\int_{-\infty}^{x''_h} \sqrt[3]{f(x)}dx = (h/L) \int_{-\infty}^{\infty} \sqrt[3]{f(x)}dx, \qquad (2)$$

$$h=1,\ldots,L-1.$$

We will compare boundaries determined by the cum. $\sqrt{f}$ rule and by the cum. $\sqrt[3]{f}$ rule to optimum stratum boundaries obtained by solving minimal equations using iterative techniques. One method of comparison will be to examine the various sets of boundaries and to note gross differences. The other method will be to calculate a ratio of variances which will measure the relative loss of precision resulting from use of approximately optimum boundaries rather than optimum boundaries.

The general theoretical framework implicit in this discussion is found in Singh and Sukhatme [12], [13]. Techniques for solving minimal equations are given in Murthy [9] and Sethi [11].

Applying the theory and methods mentioned above, we obtained optimum stratum boundaries in the case of Neyman allocation. Also obtained were the approximately optimum stratum boundaries given by the cum. $\sqrt{f}$ rule and by the cum. $\sqrt[3]{f}$ rule. Table 1 allows an easy comparison of some selected sets of boundaries.

In Table 1 the stratification variable is assumed to have a standard normal distribution; ie. $X \sim N(0,1)$. There is no loss of generality because the transformation $x_h = \mu_2+\sigma_2 x_h^s$ converts standardized stratum boundaries $\{x_h^s\}$ found in the table to optimum and approximately optimum stratum boundaries for a stratification variable distributed $N(\mu_2,\sigma_2^2)$.

In the case where the number of strata exceeds two, the optimum points of stratification depend on $|\rho|$, as well as the value of L. These points are symmetric about zero, and for fixed L, they decrease in absolute value as $|\rho|$ increases. Greater decreases are associated with higher values of $|\rho|$. Also the amount of decrease is related to the distance from the mean. Those points which are farthest from the mean show the greatest decrease.

The square root rule (1) and the cube root rule (2) yield approximately optimum stratum boundaries which are symmetric about zero. One

TABLE 1

Comparison of Optimum Points of Stratification* for Neyman Allocation with Approximately Optimum Points Determined by the Cum. $\sqrt{f}$ Rule and the Cum. $\sqrt[3]{f}$ Rule

Number of Strata	$\|\rho\|$	Optimum Points of Stratification	Approximately Optimum Points of Stratification Cum. $\sqrt{f}$ Rule	Approximately Optimum Points of Stratification Cum. $\sqrt[3]{f}$ Rule
3	0.25	0.61	0.61	0.75
	0.95	0.58		
	0.99	0.56		
4	0.25	0.00, 0.98	0.00, 0.96	0.00, 1.17
	0.95	0.00, 0.93		
	0.99	0.00, 0.90		
5	0.25	0.38, 1.24	0.36, 1.19	0.44, 1.46
	0.95	0.37, 1.19		
	0.99	0.35, 1.14		
6	0.25	0.00, 0.66, 1.45	0.00, 0.61, 1.37	0.00, 0.75, 1.68
	0.95	0.00, 0.63, 1.39		
	0.99	0.00, 0.60, 1.33		
7	0.25	0.28, 0.87, 1.61	0.26, 0.80, 1.51	0.31, 0.98, 1.85
	0.95	0.27, 0.84, 1.55		
	0.99	0.26, 0.80, 1.48		
8	0.25	0.00, 0.50, 1.05, 1.75	0.00, 0.45, 0.96, 1.63	0.00, 0.55, 1.17, 1.99
	0.95	0.00, 0.49, 1.02, 1.69		
	0.99	0.00, 0.46, 0.97, 1.61		
9	0.25	0.22, 0.68, 1.20, 1.86	0.20, 0.61, 1.08, 1.73	0.24, 0.75, 1.33, 2.11
	0.95	0.22, 0.66, 1.16, 1.81		
	0.99	0.21, 0.63, 1.11, 1.72		
10	0.25	0.00, 0.40, 0.83, 1.32, 1.97	0.00, 0.36, 0.74, 1.19, 1.81	0.00, 0.44, 0.91, 1.46, 2.22
	0.95	0.00, 0.39, 0.81, 1.29, 1.91		
	0.99	0.00, 0.38, 0.77, 1.23, 1.82		

*The stratification variable is distributed N(0,1). Use the transformation $x_h = \mu_2 + \sigma_2 x_h^s$ to convert standardized stratum boundaries $\{x_h^s\}$ found in the table to optimum and approximately optimum stratum boundaries $\{x_h\}$ for a stratification variable distributed $N(\mu_2, \sigma_2^2)$. The stratum boundaries are symmetric about the mean and only nonnegative values have been included in this table.

TABLE 2

$V_{cum.\ \sqrt{f}}/V_{min}$ and $V_{cum.\ \sqrt[3]{f}}/V_{min}$ under Neyman Allocation

$\|\rho\|$	Number of Strata		
	3	5	10
0.10	1.00000*	1.00000	1.00001
	1.00009**	1.00005	1.00001
0.50	1.00000	1.00010	1.00025
	1.00291	1.00160	1.00034
0.95	1.00143	1.00012	1.00255
	1.04275	1.03703	1.01107
0.99	1.00536	1.00292	1.00148
	1.07203	1.09408	1.05147

*$V_{cum.\ \sqrt{f}}/V_{min}$

**$V_{cum.\ \sqrt[3]{f}}/V_{min}$

can verify that $x_h'' = \sqrt{1.5}\, x_h'$ when the stratification variable X is standard normal. Note that the square root rule performs noticeably better than the cube root rule for large values of $|\rho|$.

Let $V_{\min}$, $V_{\text{cum.}\sqrt{f}}$, and $V_{\text{cum.}\sqrt[3]{f}}$ denote the variance of the stratified mean when optimum stratum boundaries are used, when the cum. $\sqrt{f}$ rule is used, and when the cum. $\sqrt[3]{f}$ rule is used, respectively. The ratios $V_{\text{cum.}\sqrt{f}}/V_{\min}$ and $V_{\text{cum.}\sqrt[3]{f}}/V_{\min}$ quantify the loss of precision incurred when using approximately optimum stratum boundaries rather than optimum boundaries. Table 2 gives values of these ratios under Neyman allocation for L=3,5,10 and $|\rho|$=0.10,0.50,0.95,0.99.

Table 2 indicates that the Dalenius and Hodges cum. $\sqrt{f}$ rule performs so well, even when ρ is moderate, that for all practical purposes this method gives optimum stratum boundaries. Losses in precision due to the use of this rule vary with L and ρ but are usually less than 0.5%. Table 2 also indicates that for large values of $|\rho|$ the use of the cum. $\sqrt[3]{f}$ rule leads to losses in precision which are not negligible. For example, when L=5 and $|\rho|$=0.99 there is a 9.4% loss in precision under Neyman allocation using the cum. $\sqrt[3]{f}$ rule but only a 0.3% loss using the cum. $\sqrt{f}$ rule.

For the purpose of discussing gains due to stratification under Neyman allocation consider Table 3. Let Var($\bar{y}$) denote the variance of the unstratified mean. Table 3 was constructed by calculating [Var($\bar{y}$)-$V_{\min}$]/Var($\bar{y}$) for L=2,...,10 and $|\rho|$=0.00,0.05,...,0.95,0.96,0.97,0.98,0.99. These values are interpreted as the proportion the variance is reduced when using a stratified mean, rather than an unstratified mean, to estimate the population mean. These are relative decreases, valid for any choice of the mean and variance of the marginal variates Y and X.

Many values of L and $|\rho|$ were included in Table 3 because it is felt that this table has practical value. Given that the estimation variable Y and the stratification variable X are approximately normal and a reasonable guess is available for ρ, an investigator can look at the appropriate line in Table 3 and find estimates of the gains in precision which can be achieved under Neyman allocation by stratifying on the variable X. The gains increase to ρ^2 as L increases, but there is a point of diminishing returns. With a knowledge of the cost of additional strata and the worth of increased precision, the investigator can choose a value of L (between 2 and 10) which is a compromise between the conflicting desires of minimizing cost and maximizing precision.

Table 3 indicates that the correlation between the estimation variable Y and the stratification variable X is an important consideration. Large gains due to stratification come from stratifying on a variable which is highly correlated

TABLE 3

[Var($\bar{y}$)-$V_{\min}$]/Var($\bar{y}$) under Neyman Allocation

$\lvert\rho\rvert$	Number of Strata								
	2	3	4	5	6	7	8	9	10
0.00	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.05	0.00159	0.00202	0.00221	0.00230	0.00236	0.00239	0.00241	0.00243	0.00244
0.10	0.00637	0.00810	0.00883	0.00920	0.00942	0.00956	0.00965	0.00972	0.00977
0.15	0.01432	0.01822	0.01986	0.02070	0.02120	0.02151	0.02172	0.02187	0.02198
0.20	0.02546	0.03239	0.03530	0.03680	0.03768	0.03824	0.03862	0.03889	0.03908
0.25	0.03979	0.05062	0.05516	0.05751	0.05888	0.05975	0.06034	0.06076	0.06107
0.30	0.05730	0.07289	0.07943	0.08281	0.08479	0.08604	0.08689	0.08749	0.08794
0.35	0.07799	0.09922	0.10812	0.11272	0.11540	0.11711	0.11827	0.11909	0.11969
0.40	0.10186	0.12960	0.14123	0.14723	0.15074	0.15297	0.15448	0.15555	0.15633
0.45	0.12892	0.16404	0.17875	0.18634	0.19078	0.19360	0.19551	0.19687	0.19786
0.50	0.15915	0.20254	0.22070	0.23006	0.23554	0.23902	0.24138	0.24305	0.24428
0.55	0.19258	0.24511	0.26707	0.27839	0.28501	0.28923	0.29208	0.29409	0.29558
0.60	0.22918	0.29174	0.31787	0.33134	0.33921	0.34422	0.34760	0.35000	0.35177
0.65	0.26897	0.34245	0.37311	0.38890	0.39812	0.40400	0.40797	0.41078	0.41285
0.70	0.31194	0.39726	0.43279	0.45109	0.46177	0.46857	0.47317	0.47642	0.47882
0.75	0.35810	0.45618	0.49695	0.51792	0.53016	0.53794	0.54321	0.54694	0.54968
0.80	0.40744	0.51926	0.56562	0.58942	0.60331	0.61214	0.61811	0.62234	0.62545
0.85	0.45996	0.58657	0.63886	0.66565	0.68126	0.69118	0.69789	0.70264	0.70613
0.90	0.51566	0.65825	0.71684	0.74673	0.76412	0.77515	0.78261	0.78789	0.79177
0.95	0.57455	0.73471	0.80002	0.83310	0.85224	0.86434	0.87250	0.87828	0.88252
0.96	0.58671	0.75066	0.81740	0.85112	0.87059	0.88289	0.89117	0.89703	0.90133
0.97	0.59900	0.76687	0.83508	0.86945	0.88925	0.90173	0.91012	0.91605	0.92040
0.98	0.61141	0.78338	0.85313	0.88818	0.90829	0.92093	0.92941	0.93539	0.93977
0.99	0.62395	0.80023	0.87166	0.90742	0.92787	0.94066	0.94921	0.95521	0.95959

with the variable of interest. For example, to achieve a 50% increase in precision, $|\rho|$ would have to exceed seven-tenths.

In practice gains of stratification are often most important for sampling cluster means, rather than for elements. Both Y and X will then be cluster means whose distributions tend toward normal distributions by the central limit theorem. Thus, our results for the bivariate normal have a great deal of relevance. Furthermore, due to "aggregation" high values of ρ between Y and X are not uncommon.

Investigating gains in precision for proportional allocation proved to be somewhat easier than for Neyman allocation because the variance of the stratified mean $\text{Var}(\bar{y}_{st})$ could be expressed in a simple relationship with $\text{Var}(\bar{y})$:

$$\text{Var}(\bar{y}_{st}) = \text{Var}(\bar{y})(1-c\rho^2), \qquad (3)$$

where c is a function of L and the stratum boundaries $\{x_h\}$.

The following conclusions were found to hold for proportional allocation:

1. The optimum stratum boundaries do not depend on ρ, as was the case with Neyman allocation. These boundaries can be found in Table 4 of Sethi [11].

2. The cum. $\sqrt{f}$ rule leads to only negligible losses of precision.

3. The cum. $\sqrt[3]{f}$ rule performs better under proportional allocation than under Neyman allocation, but the loss of precision exceeds that which would be incurred using the cum. $\sqrt{f}$ rule.

4. Under the normality assumption and with optimum or nearly optimum stratum boundaries, proportional allocation does almost as well as Neyman allocation at reducing variance. This result follows from the fact that these conditions lead to stratum variances which are approximately equal.

Further discussion and additional tables for both Neyman and proportional allocation are given in the yet unpublished paper by Anderson, Kish, and Cornell [1]. Copies of this paper are available upon request.

REFERENCES

[1] Anderson, D.W., Kish, L. & Cornell, R.G. "Quantifying gains from stratification for optimum and approximately optimum strata using a bivariate normal model," Tech. Report 4 (1975), Department of Biostatistics, The University of Michigan, Ann Arbor, Michigan.

[2] Cochran, W.G. Sampling Techniques, 2nd ed. New York: John Wiley and Sons, Inc., 1963.

[3] Cochran, W.G. "Comparisons of methods for determining stratum boundaries," Bulletin of the International Statistical Institute, 38 (2), Tokyo (1961), pp. 345-58.

[4] Dalenius, T. "The problem of optimum stratification," Skandinavisk Aktuarietidskrift, 33 (1950), pp. 203-13.

[5] Dalenius, T. and Hodges, J.L., Jr. "Minimum variance stratification," Journal of the American Statistical Association, 54 (1959), pp. 88-101.

[6] Hess, I., Sethi, V.K. & Balakrishnan, T.R. "Stratification: A practical investigation," Journal of the American Statistical Association, 61 (1966), pp. 74-90.

[7] Kish, L. Survey Sampling. New York: John Wiley and Sons, Inc., 1965.

[8] Kpedekpo, G.M.K. "Recent advances on some aspects of stratified sample design. A review of the literature," Metrika, 20 (1973), pp. 54-64.

[9] Murthy, M.N. Sampling Theory and Methods. Calcutta: Statistical Publishing Society, 1967.

[10] Serfling, R.J. "Approximately optimal stratification," Journal of the American Statistical Association, 63 (1968), pp. 1298-1309.

[11] Sethi, V.K. "A note on optimum stratification of populations for estimating the population means," Australian Journal of Statistics, 5 (1963), pp. 20-33.

[12] Singh, R. and Sukhatme, B.V. "Optimum stratification," Annals of the Institute of Statistical Mathematics, 21 (1969), pp. 515-28.

[13] Singh, R. "Approximately optimum stratification on the auxillary variable," Journal of the American Statistical Association, 66 (1971), pp. 829-33.

[14] Thomsen, I. "A comparison of approximately optimal stratification given proportional allocation with other methods of stratification and allocation," Metrika, to appear in 1975.