A COMPARISON OF APPROXIMATE STRATIFICATION TECHNIQUES UNDER AN AREA SAMPLING FRAME, TWO EMPIRICAL STUDIES

Robert D. Tortora, USDA

Nicholas J. Ciancio, USDA

Introduction

Recently work has been done comparing various approximate optimum stratification techniques. Hess, et.al., [6] and Cochran, [2] compared various techniques for actual populations. In [2], Cochran was concerned with eight different populations ranging from income per tax return, population of US cities, resources of commercial banks, number of farms per area sampling unit, and proportion of gross bank loans. In [6], the stratification and primary estimation variable was size of hospital. A brief discussion examines other estimation variables with high correlations (> .9) with the stratification variable, Kish, et.al., [1] compared various stratification techniques for a specified bivariate population where the stratification is carried out on an auxiliary variable X and estimation is made for a variable Y. Kpedeko, [7] in a review of the literature on stratification techniques calls for further empirical studies to evaluate some of these methods for different types of data.

This paper compares five approximate optimum stratification techniques when an auxiliary variable is used for stratification and when one is interested in estimating crop acreage and livestock totals. The stratification techniques are 1) cum$\sqrt{f}$, 2) Durbin, 3) Ekman, 4) Sethi, and 5) Equal Aggregate Output (EAO). The Statistical Reporting Services' (SRS) area frame is used in two States to make the comparisons. In the area frame the land area is classified (stratified) according to land use in order to achieve homogeneity within strata. The sampling unit is a segment, which is a piece of land with boundaries delineated on a map. Every parcel of land within a segment is accounted for during a survey.

The stratification variable for the area frame is the percent of land under cultivation. For each segment this is defined as the total cropland in acres in the segment divided by the total acres in the segment times 100. The crops acreage variables of interest are the three most important income-producing crops for US farmers, vis., corn, wheat, and soybeans. Similarly, the important livestock variables are cattle and hogs and these variables will also be studied.

The data used in this study are from the 1975 June Enumerative Survey for Ohio and Kansas. The segments are from the agricultural strata and the population sizes are N=252 and N=435 segments in Ohio and Kansas, respectively. Even though the above five commodities are the most important within Ohio and Kansas, these States differ demographically and geographically. Kansas is a more homogeneous farm state with more area under irrigation. The average size of farm in Kansas is larger (616 acres vs. 150 acres). Ohio has more farms (117,000 vs. 81,000) and less land in farms (17.5×10^6 acres vs 50×10^6 acres). The segment size is Kansas ranges from 1 to 4 square miles while in Ohio the segment size is ½ to 1 square miles.

Optimum allocation for fixed sample size is used to determine sample sizes in the strata. This technique is the one used by SRS.

The comparisons are made for 2, 3, 4 or 5 strata. Currently SRS uses four strata with stratum boundary values 15%, 50%, and 75%. The variances for the approximate techniques will be compared to the variance under the current SRS technique. The total number of strata is held to 5 for two reasons. First, to strain the stratification techniques, which depend more or less on the assumption that the number of strata L is reasonably large, so that within a stratum the frequency function can be assumed to be rectangular. Secondly, for practical purposes of frame construction, it is very difficult to efficiently divide the area frame into a large number of strata.

The Approximate Methods

Let $X_0, X_1, \ldots, X_L$ be the stratum boundaries, the strata being numbered 1, 2, ..., L. let S_h be the standard deviation is stratum h and $W_h = N_h/N$ be the ratio of the number of sampling units in stratum h to the total number in the population. The usual estimate of the population total is

$$y_{st} = N\Sigma W_h \bar{y}_h$$

where $\bar{y}_h$ is the sample mean in statum h. Its variance is

$$V(y_{st}) = N^2 \Sigma W_h^2 S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right).$$

For a fixed total sample size of n, $V(y_{st})$ is minimized by taking $n_n = nN_h S_h / \Sigma N_h S_h$. The minimum variance is

$$(1)\quad V_{min}(y_{st}) = \frac{1}{n}(\Sigma N_h S_h)^2,$$

ignoring the fpc. Equation (1) becomes the basis for further caculations.

A discussion of the actual implementation of the approximate stratification techniques can be found in [2], [6] or [7].

The Study Variables

Table 1 gives the shapes of the stratification variables in the two States. It is simply the percentage of the total number of segments lying within each tenth of the range of the percent of land under cultivation.

Table 1: Percentage of segments falling into successive tenths of the stratification variable.

	Ohio	Kansas
0-10	8.73	10.11
10-20	7.14	3.91
20-30	5.56	5.06
30-40	5.16	4.60
40-50	9.52	7.82
50-60	7.52	8.74
60-70	9.52	10.57
70-80	14.68	12.64
80-90	15.87	15.40
90-100	16.28	21.15

The distribution in Ohio is closest to a rectangular distribution, while in Kansas the distribution is close to being two-tailed. Cum$\sqrt{f}$, Durbin and Ekman compute approximately the same stratum boundary values, as do EAO and Sethi. Also all techniques seem to be relatively insensitive to the distribution of the stratification variable although in Kansas the values are higher than in Ohio.

Table 2 and 3 give the shape of the frequency distribution of the variables to be estimated.

Table 2: Percentage of segments having crop acreage falling into given classes by State, (Ohio/Kansas).

acres	corn	wheat	soybeans
0-10	17.46/76.78	34.92/13.79	31.75/80.46
10-50	26.59/ 8.04	38.10/ 9.66	19.05/ 8.50
50-100	36.11/ 5/98	22.62/12.18	24.60/ 6.21
100-250	19.44/ 4.83	4.36/30/11	23.81/ 4.60
> 250	0.00/ 4.37	0.00/34.26	0.79/ 0.03

Table 3: Percentage of segments having livestock numbers falling into given classes by States, (Ohio/Kansas).

Number of livestock	cattle	hogs
0	69.65/85.29	28.97/31.49
1 - 50	21.43/ 8.28	40.48/25.98
51 - 100	3.97/ 2.53	18.65/22.30
101 - 500	4.36/ 3.45	11.90/19.54
> 500	1.59/ 0.45	0.00/ 0.69

The tables show that in Ohio corn and wheat appear unimodal while soybeans appear bimodal. In Kansas corn and soybeans are skewed to the left while wheat is skewed to the right. The general shape of the distribution for hogs are the same. For cattle, the distributions are skewed to the left with the distribution in Ohio being fatter.

Finally Table 4 lists the estimated correlation coefficients between the stratification variable and the variables of interest. Note that none of the correlations are near the correlations in [6] and thus should strain the techniques.

Table 4: Estimated correlation coefficients between the stratification variable and the variables of interest, (Ohio/Kansas).

	corn	wheat	soybeans
Percent of land under cultivation	.554 / .240	.614 / .609	.652 / .028
	cattle	hogs	
percent of land under cultivation	.118 / .077	-.216 / -.356	

The correlation coefficients between the crops and the auxiliary variable are consistent in Ohio while in Kansas wheat has the highest correlation. In both states the magnitude of the correlations between livestock and percent cultivated are small.

Comparison of the Rules

Equation (1) is used as the basis for the comparisons for each technique. The variance under the approximate stratification technique is compared to the variance obtained under the present stratification used by SRS. The results are presented in Table 5.

The separation of boundary values (cum$\sqrt{f}$, Durbin, and Ekman vs. EAO and Sethi) is reflected in Table 5. As the number of strata increases the differences between the techniques does not change as much for the crop variables, i.e., for those with higher correlation coefficient with the stratification variable and where the stratification variable is more nearly rectangular (Ohio). Whenever EAO or Sethi have a smaller ratio for 2 strata than either cum$\sqrt{f}$, Durbin, or Ekman they retain their smaller ratio as the number of strata increases. For the negatively correlated variables in both states cum$\sqrt{f}$, Durbin and Ekman perform much better than the other techniques. For the highest positively correlated variables (soybeans in Ohio and wheat in Kansas) cum$\sqrt{f}$, Durbin, and Ekman perform well. To get a feel for the performance of the techniques across all variables of interest a plot of the technique(s) (with smallest ratios from Table 5 vs. the correlation coefficients (r) is given in Figure 1. For the range of correlation coefficients we see that cum$\sqrt{f}$, Durbin and Ekman perform well for negatively correlated variables as well as for moderately correlated values (r > .5). For small positive values of r all techniques seem to perform on a par.

Table 5: Variance under the approximate stratification technique divided by the variance under the current SRS stratification (Ohio/Kansas).

technique	strata	corn	wheat	soybeans	cattle	hogs
cum$\sqrt{f}$	2	3.434 / 2.499	3.229 / 2.963	3.165 / 3.012	3.224 / 2.341	3.528 / 3.374
	3	1.334 / 1.133	1.221 / 1.316	1.261 / 1.282	1.090 / 1.149	1.546 / 1.470
	4	0.781 / 0.643	0.676 / 0.719	0.699 / 0.703	0.707 / 0.603	0.882 / 0.658
	5	0.471 / 0.394	0.434 / 0.460	0.421 / 0.442	0.437 / 0.342	0.773 / 0.461

Table 5: (Con't)

technique	strata	corn	wheat	soybeans	cattle	hogs
Durbin	2	3.595 2.499	3.959 2.963	3.334 3.012	3.239 2.341	3.514 3.374
	3	1.440 1.046	1.357 1.236	1.418 1.266	1.260 0.989	1.622 1.398
	4	0.806 0.643	0.724 0.719	0.721 0.703	0.718 0.603	0.861 0.658
	5	0.483 0.377	0.452 0.439	0.433 0.420	0.446 0.285	0.538 0.444
Ekman	2	3.909 2.499	3.229 2.963	3.932 3.012	3.024 2.341	3.527 3.374
	3	1.369 1.133	1.254 1.316	1.297 1.282	1.210 1.149	1.566 1.470
	4	0.753 0.714	0.656 0.798	0.638 0.750	0.650 0.720	0.865 0.889
	5	0.498 0.440	0.464 0.481	0.425 0.428	0.467 0.424	0.554 0.429
EAO	2	4.281 2.316	4.055 3.776	3.981 3.253	3.139 1.601	5.187 4.783
	3	1.955 0.818	1.639 1.604	1.642 1.451	1.541 0.803	2.494 2.404
	4	1.045 0.450	0.942 0.912	0.934 0.836	0.844 0.343	1.650 1.433
	5	0.723 0.250	0.548 0.632	0.497 0.522	0.560 0.200	1.099 1.034
Sethi	2	4.281 1.963	4.055 3.320	3.981 3.480	3.139 1.635	5.013 4.396
	3	1.915 0.896	1.610 1.556	1.439 1.698	1.511 0.783	2.425 2.165
	4	1.048 0.497	0.925 0.847	0.945 0.937	0.842 0.339	1.696 1.284
	5	0.758 0.260	0.590 0.604	0.529 0.583	0.648 0.195	1.230 0.936

From Figure 1 we see cum$\sqrt{f}$ performs best 18 times, but Durbin is best 14 times, and Ekman is best 9 times. Cum$\sqrt{f}$ peforms well over the range of r, Durbin does well with smaller values of r, and Ekman does well with larger values of r. Figure 2 presents a graphic display of the worst of the best for the regions were the three techniques cum$\sqrt{f}$, Durbin, and Ekman perform well. Figure 2 presents the technique with the largest ratio from Table 6. The trends exhibited here indicate that cum$\sqrt{f}$ can give larger variances for negative r, Durbin for moderate r, and Ekman across the range of r.

Dalenius [3] suggested the approximation $V_L/V_{L-1} = (L-1)^2/L^2$ (for rectangular distributions) to quantify the gains caused by stratification. For L = 2, 3, 4, and 5 we get from the formula 0.250, 0.444, 0.562 and 0.640 respectivly. Table 6 presents the average gain by crop or livestock for each technique.

In general, the average gain is slightly more than that estimated by Dalenius. There is less gain in precision for the variables than with lower correlation (livestock in Ohio) with the stratification variable than with those with higher correlation (crops in Ohio). Comparing gains against the distribution of the stratifying variable we see that there are more gains (25 vs 15) for the unimodal distribution (Ohio). Defining any gain exceeding $(L-1)^2/L^2$ as significant, we see that for the unimodal distribution there are more significant gains (22 vs. 14). Finally, examining significant gains by correlation and by technique we see that cum$\sqrt{f}$ does best for the higher correlations (crop in Ohio and Kansas), Durbin performs about the same for crop and livestock and the remaining three produce more significant gains for the lower correlation (livestock in Ohio and Kansas).

Table 6: The average gain V_L/V_{L-1} for crops and livestock by straitification technique (Ohio/Kansas)

		crops	livestock
cum f	2	.188 .231	.248 .220
	3	.388 .440	.388 .464
	4	.546 .554	.610 .486
	5	.616 .627	.747 .634
Durbin	2	.202 .231	.248 .220
	3	.401 .419	.425 .418
	4	.541 .584	.550 .540
	5	.608 .598	.623 .574
Ekman	2	.176 .231	.287 .220
	3	.470 .441	.422 .464
	4	.522 .607	.545 .616
	5	.679 .597	.679 .536
EAO	2	.263 .256	.158 .226
	3	.424 .422	.496 .502
	4	.559 .508	.606 .512
	5	.602 .624	.662 .653

Table 7: (Con't)

		crops	livestock
Sethi	2	.236 / .228	.292 / .216
	3	.402 / .457	.482 / .486
	4	.593 / .558	.628 / .512
	5	.640 / .620	.747 / .652

Summary

This study compared five approximate techniques for stratification in an agricultural setting. The comparisons were based on area sampling units from two States. The stratification variable (percent of land under cultivation) was different from the variables to be estimated (corn, wheat, soybeans, cattle and hogs).

The rules divided themselves into two groups based on stratum boundary values, cum$\sqrt{f}$, Durbin, Ekman, and Equal Aggregate Output, Sethi. Comparisons were based on variances obtained using the current SRS stratification. Cum$\sqrt{f}$, Durbin, and Ekman performed well for variables either with negative correlations or moderate positive correlations with the stratification variable. All five rules were comparable for small positive correlations.

Using Dalenius' approximation, $(L-1)^2/L^2$, for gains due to increasing the number of strata it was found that the most significant gains were produced when the stratification variable was unimodal. Ekman, Equal Aggregate Output, and Sethi had more significant gains for variables not highly correlated with the stratification variable, gains and cum$\sqrt{f}$ produced more significant gaims with the higher correlated variables. It was found that the approximation $(L-1)^2/L^2$ was an overestimate of the gains due to increasing the number of strata (concurring with the results in [2]).

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, Department of Biostatistics, The University of Michigan, (1975), Ann Arbor, Michigan 48104

[2] Cochran, W.G. "Comparison of Methods for Determining Stratum Boundaries", Bulletin of the International Statistical Institute, 38(2), Tokyo (1961), pp. 345-358.

[3] Dalenius, T. Sampling in Sweden, Chapter 8, Almquist and Wiksell, Stockholm (1957).

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

[5] Hansen, M.H., Hurwitz, W.N., Madow, W.G. Sample Survey Methods and Theory, John Wiley and Sons, Inc. 1953.

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

[7] Kpedekpo, G.M.K. "Recent Advances on Some Aspects of Stratified Sample Design. A Review of the Literature", Metrika, 20 (1073), pp. 54-64.

[8] Mahalanobis, P.C. "Some Aspects of the Design of Sample Surveys. Sankhya 12 (1952) pp. 1-7.

number of strata										
5	E	D	D	S	C	EAO	C	D	C	C
4	CD	D	CD	S	E	EAO	E	CD	E	E
3	D	C	D	EAO	C	EAO	C	D	C	C
2	CDE	D	CDE	EAO	EAOS	S	C	CDE	CE	C
		-.356	-.216	.028	.077	.118	.240	.554	.609	.614	.652	r

Figure 1: Best technique (smallest ratio from Table 6) vs. Correlation coefficient (ordered by increasing magnitude). C=cum$\sqrt{f}$, D=Durbin, E+Ekman, EAO=EAO, S=Sethi.

number of strata						
5	C	C	E	E	E	D
4	E	C	D	E	D	D
3	CE	D	D	CE	D	D
2	CDE	C	E	CDE	D	E
		-.356	-.216	.554	.609	.614	.652	r

Figure 2: Worst technique (largest ratio from Table 6) vs. correlation coefficient over ranges where cum$\sqrt{f}$, Durbin and Ekman perform well.