Approximate Methods to Obtain the Optimum Stratum Boundaries: A Comparative Study

Ghadah A. Alsakkal[1] & Mowafaq Muhammed Al Kassab[2]

[1]Department of Computer Engineering, Tishk International University, Erbil, Iraq
[2]Department of Mathematics Education, Tishk International University, Erbil, Iraq
Correspondence: Ghadah A. Alsakkal, Tishk International University, Erbil, Iraq.
Email: ghada.alsakkal@tiu.edu.iq

Doi: 10.23918/eajse.v7i1p197

Abstract: In this article, we shall present an approximately optimal method for constructing stratum boundary points when the sample is allocated proportionally. The method is based on an equal partitioning of the cumulative $f^{6/7}$, where f is the distribution of the stratification variable. We show that in many practical situations, this technique compares favorably with approximately optimal stratification and allocation methods of previously suggested.

Keywords: Proportional Allocation, Stratification, Optimum Strata Boundaries, Cumulative Frequency Distribution, Efficiency

1. Approximately Optimal Stratification with Proportional Allocation

Denote A(Y) = $\int_{-\infty}^{\infty}\left(f(y)\right)^{\frac{6}{7}}dy$

We confine our attention to the finite interval [a, b], outside of which the probability density function $f(y)$ may be assumed to be zero with negligible error. Let $y_1 < y_2 < \cdots < y_{l-1}$ be the boundary points defining a construction of L strata within the interval [a, b] setting $y_0 = a$, and $y_l = b$.

Denote

$$A_h\,(Y) = \int_{Y_{h-1}}^{Y_h}\left(f(y)\right)^{\frac{6}{7}}dy$$

$$\sigma_h^2 = \frac{1}{w_h}\int_{Y_{h-1}}^{Y_h} y^2\, f(y)\, dy - \mu^2$$

Were

$$\mu_h = \frac{1}{w_h}\int_{Y_{h-1}}^{Y_h} y\, f(y)\, dy$$

$$W_h = \int_{Y_{h-1}}^{Y_h} f(y)\, dy$$

Assume that $f(y)$ be approximated within the hth stratum by its mean value μ_h. Then the weight, variance, and Ah(Y) of the hth stratum approximately are:

Received: April 26, 2021
Accepted: June 22, 2021
Alsakkal, G.A., & Al Kassab, M.M. (2021). Approximate Methods to Obtain the Optimum Stratum Boundaries: A Comparative Study. *Eurasian Journal of Science & Engineering*, 7(1), 197-204.

$$W_h = \mu_h\,(y_h - y_{h-1}) \quad [1]$$

$$\sigma_h^2 = (y_h - y_{h-1})^2 / 12 \quad [2]$$

and

$$A_h\,(Y) = \mu_h^{6/7}\,(y_h - y_{h-1}) \quad [3]$$

On ignoring the finite correction factors, Var $(\bar{y}_{st})$ under proportional allocation is given by

$$\text{Var}\,(\bar{y}_{st}) = \frac{1}{n}\sum_{h=1}^{L} w_h\ \sigma_h^2 \quad [4]$$

Substituting (2.1), (2.2), and (2.3) into (2.4), we get:

$$\text{Var}\,(\bar{y}_{st}) = \frac{1}{12n}\sum_{h=1}^{L} A_h^{7/6}\,(Y)\,(y_h - y_{h-1})^{76} \quad [5]$$

Since $\sum_{h=1}^{L} A_h(Y) = A(Y)$ is independent of the choice of boundary points, (2.5) is minimum when Ah(Y) is constant for all h, i.e.

$A_h\,(Y) = A(Y)/L$, in this case

$$\text{Var}\,(\bar{y}_{st}) = A^{7/6}\,(Y) / (12nL^2) \quad [6]$$

If the stratification is done by means of auxiliary variable X, and the regression of Y on X is linear, that is

$$Y_{hi} = \alpha + \beta X_{hi} + U_{hi} \quad [7]$$

Where Uhi are independent of each other and of Xhi and E(Uhi)=0, var (Uhi)=σ^2 . Dalenius and Hodges (1959) give equations for intermediate stratum boundaries on the X scale which make Var $(\bar{y}_{st})$ minimum for proportional allocation. The solution consists of applying rule (1.1) to X. We apply, here, the cum f6/7 rule given above to X. (4) gives

$$\text{Var}\,(\bar{y}_{st}) = \frac{1}{n}\sum_{h=1}^{L} W_h(X)\,[\,\beta^2\,\sigma_h^2\,(X) + \sigma^2 \quad [8]$$

Since $\beta 2 = \rho^2\ \frac{\sigma^2(Y)}{\sigma^2(X)}$ and σ^2 = (1-p2) σ^2 (Y), where ρ is the correlation coefficient between X and Y. (8) becomes:

$$\text{Var}\,(\bar{y}_{st}) = \frac{1}{n}\,\sigma^2\,(Y)\,[\ \frac{\rho^2\,A^{\frac{7}{6}}\,(X)}{12\,L^2\ \sigma^2\,(X)} + (1\text{-}\rho^2)\ \ ] = \frac{1}{n}\,\sigma^2\,(Y)\,[\ \frac{M^2\,(X)\rho^2}{L^2} + (1\text{-}\rho^2)]$$

Where

$$M^2\,(X) = \frac{A^{\frac{7}{6}}\,(X)}{12\,\sigma^2(X)}$$

2. Approximate Methods

As mentioned, the five suggested methods proposed, along with their researchers, are:

Method 1: The following method of stratification is studied by Thomson (1976). First the cumulative $f^{1/3}$ is formed, and then the $f^{1/3}$ scale is partitioned into equal intervals. The variance of the stratified mean Var $(\bar{y}_{1st})$ using this stratification and allocation method is given by:

$$\text{Var}\ (\bar{y}_{1st})) = \frac{H^3\ (Y)}{12\ nL^2} \qquad [9]$$

Where H(Y) = $\int_{-\infty}^{\infty} f^{1/3}\ (y)\ dy$

Method 2: This method of stratification is studied and recommended in several books and articles, as proposed by Cochran (1961, 1963), Dalenius (1957), Ekman (1959), Hess et al. (1966), Kish (1965), and Serfling (1968). First, the cumulative f1/2 is formed, and then the f1/2 scale is partitioned into equal intervals. The allocation consists of taking equally as many observations from each stratum. An approximation to the variance the stratified mean $\bar{y}_{2st}$, using this stratification and allocation method, is given by Serfling (1968):

$$\text{Var}\ (\bar{y}_{2st}) = \text{K}^2\ (\text{Y})/\ 12\text{nL}^2 \qquad [10]$$

Where K(Y) = $\int_{-\infty}^{\infty} f^{1/2}(y)\ dy$

Method 3: The following method of stratification is studied by Wasan (2017). First the cumulative $f^{3/5}$ is formed, and then the $f^{3/5}$ scale is partitioned into equal intervals. The variance of the stratified mean $\bar{y}_{3st}$ using this stratification and allocation method is given by:

$$\text{Var}\ (\bar{y}_{3st})\ = \frac{Z^{5/3}\ (Y)}{12\ nL^2} \qquad [11]$$

Where Z(Y) = $\int_{-\infty}^{\infty} f^{3/5}\ (y)\ dy$

Method 4: Next, is the stratification method researched by Al-kassab (1993), at the start, the cumulative $f^{2/3}$ is formed, and then the $f^{2/3}$ scale is partitioned into equal intervals. The variance of the stratified mean $\bar{y}_{4st}$ using this stratification and allocation method is given by:

$$\text{Var}\ (\bar{y}_{4st}) = \frac{M^{3/2}\ (Y)}{12\ nL^2} \qquad [12]$$

Where M(Y) = $\int_{-\infty}^{\infty} f^{2/3}\ (y)\ dy$

Method 5: Last but not least, this method of stratification studied by Al-kassab and Aldaghestani(1997). First the cumulative $f^{5/6}$ is formed, and then the $f^{5/6}$ scale is partitioned into equal intervals. The variance of the stratified mean $\bar{y}_{5st}$ using this stratification and allocation method is given by:

$$\text{Var}\ (\bar{y}_{5st}) = \frac{C^{6/5}\ (Y)}{12\ nL^2} \qquad [13]$$

When C(Y) = $\int_{-\infty}^{\infty} f^{5/6}\ (y)\ dy$

3. Comparison of Methods

A comparison will be done between our suggested method and the other five methods to see the relative efficiencies of these methods (*Wackerly, and et al. 2008)*. Starting off the comparison with method 1, the efficiency of method 2 relative to method 1, from (3.2) and (3.1) is:

$$\text{eff (H, K)} = \frac{\text{Var}\,(\bar{y}_{2}st)}{\text{Var}\,(\bar{y}_{1}st)} = \frac{K^2\,(Y)}{H^3\,(Y)} \quad , \qquad [14]$$

and from (11) and (9) it follows that:

$$\text{eff (H, Z)} = \frac{\text{Var}\,(\bar{y}_{3}st)}{\text{Var}\,(\bar{y}_{1}st)} = \frac{Z^{5/3}\,(Y)}{H^3\,(Y)} \quad , \qquad [15]$$

and from (12) and (9) it follows that:

$$\text{eff (H, M)} = \frac{\text{Var}\,(\bar{y}_{4}st)}{\text{Var}\,(\bar{y}_{1}st)} = \frac{M^{3/2}\,(Y)}{H^3\,(Y)} \quad , \qquad [16]$$

and from (13) and (9) it follows that:

$$\text{eff (H, C)} = \frac{\text{Var}\,(\bar{y}_{5}st)}{\text{Var}\,(\bar{y}_{1}st)} = \frac{C^{6/5}\,(Y)}{H^3\,(Y)} \quad , \qquad [17]$$

and from (6) and (9) it follows that:

$$\text{eff(H, A)} = \frac{\text{Var}\,(\bar{y}st)}{\text{Var}\,(\bar{y}_{1}st)} = \frac{A^{7/6}\,(Y)}{H^3\,(Y)} \quad , \qquad [18]$$

Second comparison with method 2, from (3.4) and (3.3) follows that:

$$\text{eff (K, Z)} = \frac{\text{Var}\,(\bar{y}_{3}st)}{\text{Var}\,(\bar{y}_{2}st)} = \frac{Z^{5/3}\,(Y)}{K^2\,(Y)} \qquad [19]$$

from (19) and (11), it follows that:

$$\text{eff (K, M)} = \frac{\text{Var}\,(\bar{y}_{4}st)}{\text{Var}\,(\bar{y}_{2}st)} = \frac{M^{3/2}\,(Y)}{K^2\,(Y)} \qquad [20]$$

from (20) and (11), it follows that:

$$\text{eff (K, C)} = \frac{\text{Var}\,(\bar{y}_{5}st)}{\text{Var}\,(\bar{y}_{2}st)} = \frac{C^{6/5}\,(Y)}{K^2\,(Y)} \qquad [21]$$

and from (2.6) and (11), it follows that:

$$\text{eff (K, A)} = \frac{\text{Var}\,(\bar{y}st)}{\text{Var}\,(\bar{y}_{2}st)} = \frac{A^{7/6}\,(Y)}{K^2\,(Y)} \qquad [22]$$

Third comparison with method 3, from (12) and (11) follows that:

$$\text{eff (Z, M)} = \frac{\text{Var}\,(\bar{y}_{4}st)}{\text{Var}\,(\bar{y}_{3}st)} = \frac{M^{3/2}\,(Y)}{Z^{5/3}\,(Y)} \qquad [23]$$

from (13) and (11), it follows that:

$$\text{eff (Z, C)} = \frac{\text{Var}\,(\bar{y}_{5}st)}{\text{Var}\,(\bar{y}_{3}st)} = \frac{C^{6/5}\,(Y)}{Z^{5/3}\,(Y)} \qquad [24]$$

and from (6) and (11), it follows that:

$$\text{eff (Z, A)} = \frac{\text{Var}\,(\bar{y}st)}{\text{Var}\,(\bar{y}_{3}st)} = \frac{A^{7/6}\,(Y)}{Z^{5/3}\,(Y)} \qquad [25]$$

Fourth comparison with method 4, from (3.5) and (3.4) follows that:

$$\text{eff (M, C)} = \frac{\text{Var}\,(\bar{y}_5 st)}{\text{Var}\,(\bar{y}_4 st)} = \frac{C^{6/5}\,(Y)}{M^{3/2}\,(Y)} \qquad [26]$$

and from (6) and (12), it follows that:

$$\text{eff (M, D)} = \frac{\text{Var}\,(\bar{y}_{st})}{\text{Var}\,(\bar{y}_4 st)} = \frac{D^{7/6}\,(Y)}{M^{3/2}\,(Y)} \qquad [27]$$

Finally, we will compare method 5 with the suggested method, equation (6):

$$\text{eff (C, A)} = \frac{\text{Var}\,(\bar{y}_{st})}{\text{Var}\,(\bar{y}_5 st)} = \frac{A^{7/6}\,(Y)}{C^{6/5}\,(Y)} \qquad [28]$$

Notice that all the ratios are independent of the number of strata L, apart from the fact that the approximations become more accurate as the number of strata increases.

4. Comparing the Various Approaches Numerically

This section compares the suggested method with the five previous methods numerically and shows how the suggested one is effective in determining the optimum stratum boundaries. For this purpose, we have considered the hospital data, obtained from the National Center for Health Statistics Hospital Discharge Survey (Valliant and et al. 2000), which have been grouped into 20 corresponding classes. In Table 1, the class frequencies are given in column 2 while their cumulative roots are given in the other columns.

Table 1: Frequency distribution and cumulative roots for all methods

Classes	f_i	$Cum\ \sqrt[3]{f}$	$Cum\ \sqrt{f}$	$Cum\ f^{3/5}$	$Cum\ f^{2/3}$	$Cum\ f^{5/6}$	$Cum\ f^{6/7}$
10-58	53	3.756	7.280	10.828	14.109	27.346	30.057
58-108	50	7.440	14.351	21.284	27.681	53.396	58.650
108-156	48	11.074	21.279	31.488	40.889	78.575	86.260
156-205	29	14.146	26.664	39.029	50.328	95.120	104.186
205-254	30	17.254	32.141	46.725	59.983	112.139	122.641
254-303	34	20.493	37.972	55.021	70.478	131.029	143.185
303-352	28	23.530	43.264	62.405	79.699	147.097	160.580
352-400	27	26.530	48.460	69.630	88.699	162.686	177.441
400-449	13	28.881	52.065	74.290	94.228	171.163	186.453
449-498	17	31.452	56.189	79.763	100.839	181.765	197.794
498-547	18	34.073	60.431	85.428	107.708	192.884	209.705
547-596	15	36.539	64.304	90.506	113.790	202.435	219.893
596-644	7	38.452	66.950	93.720	117.449	207.497	225.194
644-693	5	40.162	69.186	96.346	120.373	211.320	229.167
693-742	3	41.605	70.918	98.279	122.453	213.818	231.732
742-791	3	43.047	72.650	100.213	124.533	216.316	234.296
791-840	5	44.757	74.886	102.839	127.457	220.140	238.269
840-888	2	46.017	76.300	104.355	129.045	221.922	240.080
888-937	4	47.604	78.300	106.652	131.565	225.097	243.362
937-986	2	48.864	79.715	108.168	133.152	226.878	245.173

Applying equations (9), (10), (11), (12), (13), and (6) respectively, n Var ($\bar{y}_{st}$) for all the approximate methods will be given in Table 2.

Table 2: gives the values of n Var ($\bar{y}_{st}$) for all the approximation methods for different number of strata L = 2,3,4,5

Number of strata	$Cum\ \sqrt[3]{f}$	$Cum\ \sqrt{f}$	$Cum\ f^{3/5}$	$Cum\ f^{2/3}$	$Cum\ f^{4/5}$	$Cum\ f^{5/6}$	$Cum\ f^{6/7}$
2	2430.668	132.385	51.159	32.010	16.028	13.987	12.778
3	1080.297	58.838	22.737	14.226	7.123	6.216	5.679
4	607.667	33.096	12.790	8.002	4.007	3.497	3.195
5	388.907	21.182	8.185	5.122	2.564	2.238	2.045

Notice from Table 2 that as the number of strata increase, n Var ($\bar{y}_{st}$) decrease. The $Cum\ f^{6/7}$ method also has minimum value in comparison with the other methods. In order to find the efficiency of these methods and to pinpoint the most efficient method, the equations of section four must be put into use respectively; the relative efficiencies in Table 3 are:

Table 3: Relative efficiencies of the approximate methods

Method	Relative Efficiency				
$H^3(Y)$ = 116672.108					
$K^2(Y)$ = 6354.481	eff (H, K) =0.0545				
$Z^{\frac{5}{3}}(Y)$ = 2455.630	eff (H, Z) =0.0210	eff (K, Z) =0.3864			
$M^{\frac{3}{2}}(Y)$ = 1530.461	eff (H, M) =0.0132	eff (K, M) =0.2418	eff (Z, M) =0.6257		
$C^{\frac{6}{5}}$(Y) = 671.353	eff (H, C) =0.0058	eff (K, C) =0.1057	eff (Z, C) =0.2734	eff (M, C) =0.4369	
$A^{\frac{7}{6}}(Y)$ = 613.364	eff (H, A) =0.0053	eff (K, A) =0.0965	eff (Z, A) =0.2498	eff (M, A) =0.3992	eff (C, A) =0.9136

Table 3 shows that the variance of the $Cum\ \sqrt{f}$ is approximately 5.5% of the variance the $Cum\ \sqrt[3]{f}$, and the variance of the $Cum\ f^{3/5}$ is approximately 2.1% of the variance the $Cum\ \sqrt[3]{f}$, …, and the variance of the $Cum\ f^{6/7}$is approximately 0.5% of the variance the $Cum\ \sqrt[3]{f}$. Thevariance of the $Cum\ f^{3/5}$is approximately 38.6% of the variance the $Cum\ \sqrt{f}$,, and the variance of the $Cum\ f^{6/7}$is approximately 9.7% of the variance the $Cum\ \sqrt{f}$. The variance of the $Cum\ f^{2/3}$is approximately 62.6% of the variance the $Cum\ f^{3/5}$, …, and the variance of the

$Cum\ f^{6/7}$is approximately 25% of the variance the $Cum\ f^{3/5}$. The variance of the $Cum\ f^{5/6}$is approximately 43.7% of the variance the$Cum\ f^{2/3}$, and the variance of the $Cum\ f^{6/7}$is approximately 39.9% of the variance the$Cum\ f^{2/3}$. Finally, the variance of the $Cum\ f^{6/7}$is approximately 91.4% of the variance the$Cum\ f^{5/6}$.

5. Conclusion

To sum it up, the $Cum\ f^{6/7}$ has proved to be the most efficient out of the presented approximate methods. All the ratios of the variances of the stratified mean are independent of the number of strata L, apart from the fact that these approximations become more accurate as the number of strata increases. Putting the differences between the formulas and the numerical data into consideration, a conclusion can be drawn that affirms the suggested method achieves the least variance out of all presented methods, therefore making it the most efficient and accurate method.

References

Al-kassab M.M.T. and Aldaghestani, T.H. (1997). Using the Linear Model to Find Stratum Boundaries Based on the Proportional Allocation: A comparative Study. *Mutah Journal for Research and Studies, 1294)*, 195-209.

Al-kassab M.M.T. (1993). Approximately Optimal Stratification Using Proportional Allocation. *Tanmeat Alrafideen Journal, 41*.

Cochran, W. (1963). *Sampling techniques*. Wiley, N.Y.

Cochran, W. (1961). Comparison of methods for determining stratum boundaries. *Bull. Int. Statist. Inst.* 3pt, (2), 345-358.

Dalenius, T. (1957). *Sampling in Sweden*. Almqvist & Wiksell, Stockholm.

Dalenius, T. and Hodges, J. L. (1959). Minimum variance stratification. *J. Amer Statist. Ass.*, 54, 88-101.

Ekman, G. (1959). An approximation useful in univariate stratification. *J. Amer. Statist. Ass.*, 30, 219-229.

Hess, I. Sethi, V. K. and Balakrishnan, T. R. (1966) Stratification: A practical investigation. *J. Amer. Statist. Ass*. 613 74-90.

Kish L. (1965). *Survey Sampling.* New York 1965.

Serfling, R. J. (1968). Approximately optimal stratification. *J. Amer. Statist. Ass*. 63, 1298-1309.

Singh, R. (1971). Approximately optimal stratification on the auxiliary variable. *J. Amer. Statist. Ass.* 66, 829-833.

Thomson, I. (1976). Comparison of approximately optimal stratification given proportional allocation with other methods of stratification and allocation. *Metrika, Band 23*, Seite 15-25, 1976.

Valliant, R., Dorfman, A.H., & Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York.

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical Statistics with Applications* (Seventh ed.). Belmont.

Wasan, A.A. (2017). *Approximate optimal strata based on proportional allocation*. Postgraduate Diploma Thesis. University of Mosul.