Liet. matem. rink, **47***, spec. nr.,* 2007, 369–374

Stratification of populations with skewed distribution[*]

Dalius PUMPUTIS (VPU, MII)
e-mail: dpumputis@yahoo.co.uk

Abstract. The problem of efficient stratification in the case of skewed population is considered. Four stratification methods are examined. A new adjusted geometric stratification method is introduced. This method is compared by simulation with the Dalenius-Hodges cumulative root frequency method, the geometric method proposed by Gunning and Horgan [2], and the power method offered by Plikusas in [6]. The simulation results show that in most cases considered the power method is the most efficient one.

Keywords: stratification; skewed population; power method; adjusted geometric method.

1. Introduction

Survey statisticians are always concerned in selecting the best sample design which gives more accurate estimates of the population parameters of interest. One of the classical and still efficient sample designs is a stratified sample design: the survey population is divided into several non-overlapping parts (strata), the sample is drawn from each part independently, then according to the selection method, the population parameters are estimated on the basis of the sample drawn. In survey practice the most popular stratification method is *the cumulative root frequency stratification method* considered by Dalenius and Hodges [1, 4]. It is useful even nowadays. A review of different stratification methods for the skewed populations is considered in [3]. In the following section, we formulate a stratification problem in general. The new stratification method is presented in Section 3.

2. Stratification problem

Consider a finite population $\mathcal{U} = \{u_1, u_2, \dots, u_N\}$ of N elements. Let y be a study variable defined on the population $\mathcal{U}$ and taking values $\{y_1, y_2, \dots, y_N\}$. Let us consider a *stratified simple random sample* obtained by partitioning the population into non-overlapping groups, called strata, and then selecting a simple random sample from each stratum. Suppose that the number of strata H is fixed and known. Denote by U_h the stratum h, by s, $s \subset \mathcal{U}$, a stratified random sample set, drawn from the population $\mathcal{U}$, and by s_h a simple random sample selected from the stratum h.

Using the proper stratification strategy, we can get estimators of the population parameters of interest which provide more precise estimates at a lower survey cost. The aim of a survey statistician is to decide how to select the best stratification algorithm

[*] The research is supported by the Grant of Lithuanian science foundation, T-07149.

in order to maximize the precision of considered estimators, i.e, to minimize variance, MSE or the coefficient of variation (cv) of estimators.

The classical stratification problem is formulated by choosing the population mean as a parameter of interest and minimizing the variance of its estimator:

$$\hat{\mu} = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_h.$$

Here $\bar{y}_h$ is the sample mean in stratum h, N_h is the number of elements in stratum h, and the product $N_h\bar{y}_h$ is a well known Horvitz-Thompson estimator of the stratum h total.

Stratification procedure deals with several issues. How to choose the stratification variable? How can the strata boundaries be determined? How many strata should there be? How large sample should be selected? How to allocate the sample to the strata defined?

We suppose the number of strata H and the sample size n to be chosen, and consider the second issue assuming that the sample is distributed according to the Neyman optimal allocation [5].

Let the variable y be known and its values be arranged in an ascending order. Denote by k_0 and k_H the smallest and largest values of y respectively. The problem is to find intermediate stratum boundaries $k_1, k_2, \dots, k_{H-1}$ such that $var(\hat{\mu})$ be minimal. An assumption that the variable y is known is unrealistic, therefore we will use auxiliary variable x for stratification. This auxiliary variable x should be well correlated with the study variable y. The principle remains the same: the values of variable x are arranged in an ascending order and we are looking for the stratum boundaries which minimize variance of the mean estimator $var(\hat{\mu}_x)$ for the variable x.

Tore Dalenius has showed that stratum boundaries with the above-mentioned property exist and satisfy the following equations:

$$\frac{(k_h - \mu_h)^2 + S_h^2}{S_h} = \frac{(k_h - \mu_{h+1})^2 + S_{h+1}^2}{S_{h+1}}, \quad h = 1, 2, \dots, H-1, \tag{1}$$

where S_h, μ_h are the standard deviation and mean of the stratum h. There are $H-1$ equation, moreover, both S_h and μ_h depend on k_h. Thus, we have complicated iterative equations. Some additional problems arise:

a) how to select the first approximation of the solution k_h, $h = 1, \dots, H-1$;
b) whether the iteration procedure converge.

3. Some stratification methods

I. *The cumulative root frequency method.* Denote by $f(x)$ a continuous density of the auxiliary variable x. Assuming that the distribution of x in each stratum is approximately uniform, Dalenius and Hodges [1, 4] have showed that the minimum variance of the population mean estimator is approximately achieved when the strata boundaries $k_h^{(f)}$ are chosen so that

$$\int_{k_0^{(f)}}^{k_1^{(f)}} \sqrt{f(x)}\,\mathrm{d}x = \int_{k_1^{(f)}}^{k_2^{(f)}} \sqrt{f(x)}\,\mathrm{d}x = \cdots = \int_{k_{H-1}^{(f)}}^{k_H^{(f)}} \sqrt{f(x)}\,\mathrm{d}x.$$

If the distribution of the variable x is discrete, then $f(x)$ is the frequency function of x. So, the rule is to choose stratum boundaries $k_h^{(f)}$ so that the following totals

$$\sum_{l \in U_h} \sqrt{f(x_l)}$$

be approximately the same.

II. *Geometric method.* An interesting method is presented by Gunning and Horgan [2]. They have proposed a new algorithm for construction of stratum boundaries, based on an observation that, with near optimum boundaries, the coefficient of variation of the stratification variable x is the same in all strata:

$$\frac{S_1}{\mu_1} = \frac{S_2}{\mu_2} = \cdots = \frac{S_H}{\mu_H}.$$

Assuming that the distribution of the variable x within each stratum is uniform, the following expression for the approximately optimum stratum boundaries has been obtained:

$$k_h^{(g)} = k_0^{(g)} r^h, \quad r = \left(\frac{k_H^{(g)}}{k_0^{(g)}}\right)^{1/H}, \quad h = 0, 1, \ldots, H.$$

So, the stratum boundaries are terms of a geometric progression. This method is called *Geometric method* and it is proposed for skewed populations.

III. *Power method.* A simple and efficient method is proposed in Plikusas [6]. The boundaries $k_h^{(p)}$ are chosen so that the totals

$$\sum_{l \in U_h} x_l^{\alpha} = \text{const}, \quad h = 1, 2, \ldots, H$$

be approximately the same. Unfortunately this method does not have a theoretical reasoning so far, and its advantages can be shown by simulation. Many experiments show that the parameter α should be in the range from 0.5 to 0.7. There is a hypothesis that the parameter α depends on the exponential distribution parameter λ.

IV. *Adjusted geometric method.* Using the same idea of Gunning and Horgan [2] to equalize the coefficients of variation of each stratum and assuming that the distribution within each stratum is exponential, we get iterative equations for defining the strata boundaries:

Table 1. Comparison of stratification methods. Number of strata $H = 5$, Sample size $n = 50$.

Population	Stratification					Stratum		
skewness	method	cv		1	2	3	4	5
3.13	Cum$\sqrt{f}$	0.1592	$k_h^{(f)}$	75709	118149.8	173649.1	255265.9	
			N_h	83	72	57	47	41
			n_h	9	8	8	11	14
	Geometric	0.1785	$k_h^{(g)}$	57826.9	91532.2	144883.2	229330.8	
			N_h	51	62	79	56	52
			n_h	3	5	11	12	19
	Power	0.1501	$k_h^{(p)}$	81624	122303	180321	263000	
			N_h	96	69	55	44	36
			n_h	12	8	9	10	11
	Adjusted	0.1755	$k_h^{(adj)}$	58129.7	92350.2	146364.2	231123.6	
	geometric		N_h	51	65	77	56	51
			n_h	3	6	10	13	18
4.98	Cum$\sqrt{f}$	0.1635	$k_h^{(f)}$	181	354.7	571.9	832.4	
			N_h	78	67	66	51	38
			n_h	7	6	7	7	22
	Geometric	0.2272	$k_h^{(g)}$	22.9	71.4	223.2	697.3	
			N_h	9	22	68	141	60
			n_h	0	0	4	24	22
	Power	0.1599	$k_h^{(p)}$	261.2	464.3	649.1	872.3	
			N_h	117	62	48	40	33
			n_h	15	7	4	5	19
	Adjusted	0.1877	$k_h^{(adj)}$	25.6	88.1	288.4	849.7	
	geometric		N_h	11	25	89	140	35
			n_h	0	1	6	28	15
5.95	Cum$\sqrt{f}$	0.1332	$k_h^{(f)}$	176771.2	883576	1767081.9	5654508.1	
			N_h	151	84	31	19	15
			n_h	4	8	3	11	24
	Geometric	0.2288	$k_h^{(g)}$	842.4	10138.1	122006.8	1468292.4	
			N_h	1	30	95	136	38
			n_h	0	0	1	11	38
	Power	0.1231	$k_h^{(p)}$	329353	1036455	3948983	9051276	
			N_h	189	56	31	14	10
			n_h	8	6	11	11	14
	Adjusted	0.1808	$k_h^{(adj)}$	1182	21194.1	375856.7	3758593.2	
	geometric		N_h	4	49	147	75	25
			n_h	0	0	4	13	32
10.91	Cum$\sqrt{f}$	0.1730	$k_h^{(f)}$	6.8	15.6	24.4	39	
			N_h	79	100	61	37	23
			n_h	4	6	4	4	32
	Geometric	0.0693	$k_h^{(g)}$	3.1	9.7	30.2	94.1	
			N_h	54	39	169	37	1
			n_h	1	2	32	15	0
	Power	0.1968	$k_h^{(p)}$	11	16	22	37	
			N_h	115	64	50	43	28
			n_h	9	2	2	4	32
	Adjusted	0.0645	$k_h^{(adj)}$	3.5	12.2	40	116.1	
	geometric		N_h	54	95	128	22	1
			n_h	2	9	30	9	0

$$k_h^{(adj)} = \frac{I_1(h)I_2(h+1)k_{h+1}^{(adj)} + I_1(h+1)I_2(h)k_{h-1}^{(adj)}}{I_1(h)I_2(h+1) + I_1(h+1)I_2(h)},$$

where

$$I_1(h) = \int_{k_{h-1}^{(adj)}}^{k_h^{(adj)}} t\mathrm{e}^{\lambda t}\,\mathrm{d}t, \quad I_2(h) = \int_{k_{h-1}^{(adj)}}^{k_h^{(adj)}} \mathrm{e}^{\lambda t}\,\mathrm{d}t.$$

Let us compare the described stratification methods by simulation.

4. Simulation study

We compare all the mentioned stratification methods considering four real populations of size 300 having a skewed distribution which is close to exponential. The sample size $n = 50$ is distributed into five strata, using Neyman's optimal allocation. The known variable x is used for the stratification and the results are presented for the study variable y which is highly correlated ($\rho \approx 0.9$) with the variable x.

$m = 1000$ samples s_j are drawn. The strata boundaries and the coefficient of variation of the estimate of μ_y are calculated for each method. The simulation results for some skewed populations are presented in Table 1.

For the most skewed populations the power method is the best one. The geometric method is simple, but precision is lowest in the most cases considered. It can be observed, for example, in the case of the first population.

The coefficient of skewness for the second and third populations is higher, but the efficiency of all methods remains almost the same. Moreover, there appear more significant differences between the power method and the others.

It should be mentioned, that for very skewed populations the power method is not best. This situation illustrates the fourth population with the highest coefficient of skewness. The adjusted geometric method is preferable in this case.

The simulation was also performed for populations with a normal distribution. Then the cumulative root frequency method is most suitable, however differences in efficiency of stratification methods are minimal.

References

1. W.G. Cochran, *Sampling Techiques*, New York. John Wiley and Sons (1977).
2. P. Gunning, J.-M. Horgan, A new algorithm for the construction of stratum boundaries in skewed populations, *Survey Methodology*, **30**(2), 159–166 (2004).
3. J.-M. Horgan, Stratification of skewed populations: a review, *International Statistical Review*, **74**(1), 67–76 (2006).
4. T. Dalenius, J.-L. Hodges, Jr., Minimum variance stratification, *Journal of the American Statistical Association*, **54**(285), 88–101 (1959).
5. A. Plikusas, D. Krapavickaitė, *Imčių teorijos pagrindai*, Vilnius, Technika (2005).
6. A. Plikusas, Sampling methods in Lithuanian official statistics, Design and parameter estimation problems, in: *Probability Theory and Mathematical Statistics, Proceedings of the Seventh Vilnius Conference*, Uthrecht, VSP (1999).

REZIUMĖ

D. Pumputis. Asimetrinių populiacijų sluoksniavimas

Straipsnyje nagrinėjamas populiacijų sluoksniavimo uždavinys, kai tyrimo kintamojo skirstinys yra asimetrinis. Pasiūlytas naujas – pataisytasis geometrinis sluoksniavimo metodas. Šis metodas modeliuojant lyginamas su trimis kitais žinomais metodais: kvadratinės šaknies iš skirstinio dažnio, geometriniu ir laipsninio sluoksniavimo metodu. Modeliavimo rezultatai rodo, kad vidutiniškai asimetrinėms populiacijoms geriausiai tinka laipsninio sluoksniavimo metodas, o ypač asimetrinėms populiacijoms geriausias yra pataisytasis geometrinis sluoksniavimas.