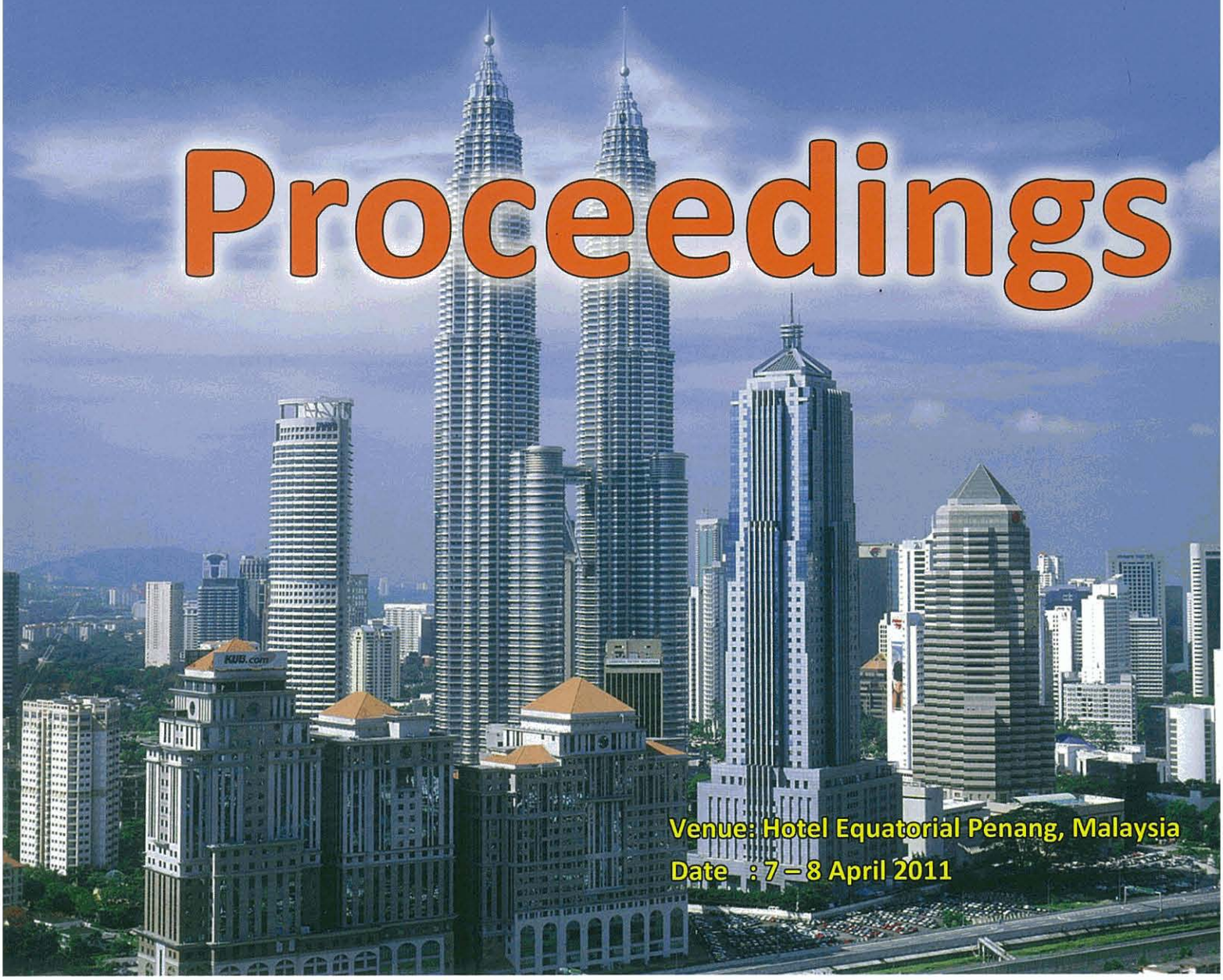
ORS 2011

Annual International Conference on

Operations Research and Statistics

Proceedings



Venue: Hotel Equatorial Penang, Malaysia

Date : 7 – 8 April 2011

Published and Organized by Global Science and Technology Forum (GSTF)

www.globalstf.org

Determining Optimum Strata Boundaries for Skewed Population with Log-normal Distribution

M.G.M. Khan and Dinesh Rao
School of Computing, Information & Mathematical Sciences
The University of the South Pacific, Suva, Fiji

A.H. Ansari and M. J. Ahsan
Department of Statistics and Operations Research
Aligarh Muslim University, Aligarh, India

Abstract: The method of choosing the best boundaries that make strata internally homogeneous as far as possible is known as optimum stratification. To achieve this, the strata should be constructed in such a way that the strata variances for the characteristic under study be as small as possible. If the frequency distribution of the study variable is known, the Optimum Strata Boundaries (OSB) could be obtained by cutting the range of the distribution at suitable points. In this paper the problem of finding the OSB for a skewed population with standard Log-normal distribution is studied. The problem is then redefined as the problem of determining Optimum Strata Widths (OSW) and is formulated as a Mathematical Programming Problem (MPP) that seeks minimization of the variance of the estimated population parameter under Neyman allocation subject to the constraint that sum of the widths of all the strata is equal to the total range of the distribution. The formulated MPP turns out to be a multistage decision problem that can be approached by dynamic programming technique. A numerical example is presented to illustrate the application and computational details of the proposed method. The results are compared with the Dalenius and Hodge's cum $\sqrt{f}$ method, which reveals that the proposed technique is more efficient and also useful for a skewed population when the other methods may fail to obtain OSB.

1. Introduction

When a single characteristic is under study and its frequency distribution is known, one could determine the strata boundaries. The problem was first considered by Dalenius(1950). Subsequently many authors attempted in this direction. Most of the authors obtained the calculus equations for the strata boundaries, which are ill adapted to practical computations. Khan et al. (2002) formulated this problem as a MPP and developed a solution procedure using dynamic programming. They worked out OSB to the populations having uniform and right triangular distributions.

In this paper the authors extend the dynamic programming approach to determine the OSB for a skewed population with standard Log-normal distribution under Neyman allocation. Section 2 provides the details formulation of the problem of finding OSW as an MPP. The solution procedure using dynamic programming technique to solve the MPP is then discussed in Section 3 and the computational details of the solution procedure are illustrated with a numerical example in Section 4. Finally, in section 5, an investigation is carried out to compare the results obtained by the proposed technique and the cum $\sqrt{f}$ method of Dalenius and Hodge's (1959).

2. Formulation of the Problem

Let the population be stratified into L strata and the estimation of the population mean of study variable x is of interest. Let $f(x)$ denotes frequency function and x_0 and x_L be the smallest and largest values of x respectively. Then the problem of determining the strata boundaries is to cut up the range, $x_L - x_0 = d$ (say), at intermediate points $x_1 \le x_2 \le,...,\le x_{L-1}$ such that the variance of the stratified sample mean $\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h$ under Neyman allocation given by $V(\bar{x}_{st}) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h \sigma_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h \sigma_h^2$ is minimum. If the finite population correction (f.p.c) is ignored, minimize $V(\bar{x}_{st})$ is equivalent to minimize

$$\sum_{h=1}^{L} W_h \sigma_h \qquad (1)$$

where, W_h and σ_h are obtained in terms of boundary points by

$$W_h = \int_{x_{h-1}}^{x_h} f(x)dx \qquad \text{and}$$

$$\sigma_h^2 = \frac{1}{W_h}\int_{x_{h-1}}^{x_h} x^2 f(x)dx - \mu_h^2, \qquad \text{where}$$

$$\mu_h = \frac{1}{W_h}\int_{x_{h-1}}^{x_h} x f(x)dx. \qquad (2)$$

If $y_h = x_h - x_{h-1}$ denotes the width of the h th stratum, obviously $\sum_{h=1}^{L} y_h = d$. Then, (1) is expressed as a function of y_h; $h = 1,2,...,L$ only and the problem of determining OSB is reduced to the problem of determining OSW. Let $f_h(y_h) = W_h \sigma_h$, thus the above problem can be stated as the following MPP:

$$\text{Minimize} \quad \sum_{h=1}^{L} f_h(y_h) \qquad \text{subject to} \quad \sum_{h=1}^{L} y_h = d \text{ and}$$

$$y_h \ge 0;\ h = 1,2,\ldots,L.$$

Let x follows standard Log-normal distribution with the parameter $\sigma > 0$. Then

Proceedings of the Annual International Conference on Operations Research and Statistics (ORS 2011)

ISBN: 978-981-08-8407-9

$$f(x)=\frac{\exp\left[-\left(\left(\ln(x)\right)^2/\left(2\sigma^2\right)\right)\right]}{x\sigma\sqrt{2\pi}};\quad x>0. \qquad (3)$$

Using (2) and (3) the problem of determining OSW may be expressed as an MPP:

$$\text{Min}\ \sum_{h=1}^{L}\left\{Sqrt\left\{\left[-\frac{1}{2}\exp\left(2\sigma^2\right)\left(erf\left(\frac{2\sigma^2-\ln\left(y_h+x_{h-1}\right)}{\sigma\sqrt{2}}\right)-erf\left(\frac{2\sigma^2-\ln\left(x_{h-1}\right)}{\sigma\sqrt{2}}\right)\right)\right]\left[\frac{1}{2}\left(erf\left(\frac{\ln\left(y_h+x_{h-1}\right)}{\sigma\sqrt{2}}\right)-erf\left(\frac{\ln\left(x_{h-1}\right)}{\sigma\sqrt{2}}\right)\right)\right]-\left[\frac{1}{2}\exp\left(\frac{\sigma^2}{2}\right)\left(erf\left(\frac{\sigma^2-\ln\left(y_h+x_{h-1}\right)}{\sigma\sqrt{2}}\right)-erf\left(\frac{\sigma^2-\ln\left(x_{h-1}\right)}{\sigma\sqrt{2}}\right)\right)\right]^2\right\}\right\},$$

subject to $\sum_{h=1}^{L}y_h=d$, and $y_h\geq 0;\ h=1,2,\ldots,L.$

3. The Solution Using Dynamic Programming

Let $f(k,d_k)$ denotes the minimum value of the objective function of (4) at a stage $k(<L)$ when $d_k<d$ is the total available width for division into k strata. Using dynamic programming technique to the MPP (4), the recurrence relations are:

$$f(1,d_1)=Sqrt\left\{\left[-\frac{1}{2}\exp\left(2\sigma^2\right)\left(erf\left(\frac{2\sigma^2-\ln(d_1+x_0)}{\sigma\sqrt{2}}\right)-erf\left(\frac{2\sigma^2-\ln(x_0)}{\sigma\sqrt{2}}\right)\right)\right]\left[\frac{1}{2}\left(erf\left(\frac{\ln(d_1+x_0)}{\sigma\sqrt{2}}\right)-erf\left(\frac{\ln(x_0)}{\sigma\sqrt{2}}\right)\right)\right]-\left[\frac{1}{2}\exp\left(\frac{\sigma^2}{2}\right)\left(erf\left(\frac{\sigma^2-\ln(d_1+x_0)}{\sigma\sqrt{2}}\right)-erf\left(\frac{\sigma^2-\ln(x_0)}{\sigma\sqrt{2}}\right)\right)\right]^2\right\}$$

(5) at $y_1^*=d_1$, for $k=1$.

$$f(k,d_k)=\min_{0\leq y_k\leq d_k}\left\{Sqrt\left\{\left[-\frac{1}{2}\exp(2\sigma^2)\left(erf\left(\frac{2\sigma^2-\ln(d_k+x_0)}{\sigma\sqrt{2}}\right)-erf\left(\frac{2\sigma^2-\ln(d_k-y_k+x_0)}{\sigma\sqrt{2}}\right)\right)\right]\left[\frac{1}{2}\left(erf\left(\frac{\ln(d_k+x_0)}{\sigma\sqrt{2}}\right)-erf\left(\frac{\ln(d_k-y_k+x_0)}{\sigma\sqrt{2}}\right)\right)\right]-\left[\frac{1}{2}\exp\left(\frac{\sigma^2}{2}\right)\left(erf\left(\frac{\sigma^2-\ln(d_k+x_0)}{\sigma\sqrt{2}}\right)-erf\left(\frac{\sigma^2-\ln(d_k-y_k+x_0)}{\sigma\sqrt{2}}\right)\right)\right]^2\right\}+f\left(k-1,d_k-y_k\right)\right\},$$

(6)

for $k\geq 2$.

4. A Numerical Example

Assume that x follows the standard Log-normal distribution with $\sigma=1$, in the interval $[0.00001,13.00001]$, that is $x_0=0.00001$ and $x_L=13.00001$. The recurrence relations (5) and (6) are solved by executing a computer program developed for the solution procedure. Table 1 gives the optimum strata boundary points (x_h^*) and the values of objective function $\sum_{h=1}^{L}W_h\sigma_h$ for $L=2, 3, 4, 5$ and 6.

Table 1. OSB for a standard Log-normal study variable.

No. of strata L	Optimum Strata boundaries: $x_h^*=x_{h-1}^*+y_h^*$	Variance $\sum_{h=1}^{L}W_h\sigma_h$
2	2.23653	0.8569355124
3	1.30860, 3.65945	0.5773613579
4	0.95460, 2.20738, 4.74155	0.4358095763
5	0.76590, 1.60922, 2.97289, 5.59430	0.3501356776
6	0.64768, 1.28199, 2.19156, 3.63412, 6.28459	0.2926636591

5. Discussion

A numerical investigation is carried out below to compare the proposed technique to the commonly used Dalenius and Hodge's cum $\sqrt{f}$ method. For this purpose, we generate data of size $N=1000$ for a standard Log-normal population with $\sigma=1$ and the density function $f(x)=\left(1/x\sqrt{2\pi}\right)\exp\left[-\left(\ln(x)\right)^2/2\right]$, which have been grouped into 15 equal classes. In Table 2 the class frequencies are given in column 2 while their cumulative roots are given in column 3. For this example the smallest and the largest values of x are $x_0=0.025$ and $x_L=32.4$ respectively. Therefore, the range of the distribution $d=32.375$.

Table 2. Frequency distribution of x and cum $\sqrt{f(x)}$.

Class	Frequency $f(x)$	Cum $\sqrt{f(x)}$
0.025-2.18	7915	89.0
2.18-4.34	1390	126.2
4.34- 6.5	391	146.0
6.5-8.66	155	158.5
8.66-10.8	62	166.3
10.8-13	34	172.2
13-15.1	19	176.5
15.1-17.3	8	179.4
17.3-19.5	9	182.4
19.5-21.6	5	184.6
21.6-23.8	4	186.6
23.8-25.9	3	188.3
25.9-28.1	2	189.7
28.1-30.3	2	191.2
30.3-32.4	1	192.2

For this distribution the OSB are determined by using both cum $\sqrt{f}$ and dynamic programming methods. For each $L=2,3$ and 4 the variance $\sum_{h=1}^{L}W_h\sigma_h$ is calculated, which is used to compare the efficiency of the two methods. The results of this investigation are given in Table 3.

Table 3. Relative efficiency of dynamic programing method

L	Cum $\sqrt{f}$ method		DP method		Relative Efficiency
	OSB	$\sum_{h=1}^{L}W_h\sigma_h$	OSB	$\sum_{h=1}^{L}W_h\sigma_h$	
2	4.34	1.13876	2.60509	1.04443	109.03172
3	2.18 6.5	0.75022	1.47386 4.54269	0.70562	106.32069
4	2.18 4.34 6.5	0.65746	1.06071 2.59123 6.19841	0.53353	123.22831

From the last column of table it can be seen that the dynamic programming method is more efficient for optimum stratification for all $L=2,3$ and 4. It has also been seen that, if L is large (more than 4), the cum $\sqrt{f}$ method fails to determine the OSB as the population is skewed, whereas, the proposed dynamic approach is useful for a skewed population with any number of strata.

6. References

[1] Dalenius, T. (1950). The problem of optimum stratification-II. Skand. Aktuartidskr, 33, 203-213.

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

[3] Khan, E.A., Khan, M.G.M., and Ahsan, M.J. (2002). Optimum stratification: A mathematical programming approach. Culcutta Statistical Association Bulletin, 52 (special), 205-208.

ORS 2011 PARTNER UNIVERSITIES

GSTF PARTNER UNIVERSITIES

ISBN 978-981-08-8407-9