Cut-off sampling for right skewed long tail distribution

Sang Eun Lee*,
Kyonggi University, Republic of Korea, sanglee62@kgu.ac.kr

Key-Il Shin
Hankuk University of Foreign Studies, Republic of Korea, keyshin@hufs.ac.kr

Abstract

Cut-off sampling method has been widely used for business survey which has generally a right skewed population with a long tail. The modified cut-off sampling Hidiroglou (1987) suggested is the method which determines the cut-off point by minimizing the sample size. In this paper we suggest a new cut-off point determines by using the underlying distribution which are log-normal and gamma distribution. Small Monte-Carlo simulation studies are performed to confirm the theoretical results.

Key words : Truncated log-normal distribution, truncated gamma distribution, cut-off point, k-th moments.

1. Introduction

One of popular sampling designs for right skewed long tail distributions is the cut-off sampling. It is known that many business survey variables follow right skewed long tail distribution. The modified cut-off sampling suggested by Hidiroglou (1986) is a special case of stratified sampling which makes population into two parts : a take-all stratum and a take-some strata. The take-some stratum is a stratum taking some samples from the stratum and take-all stratum takes all elements as samples in that stratum.

A crucial point to design the cut-off sampling is the determination of the cut-off point which separates take-all and take-some stratum. Hidiroglou (1986) suggested an algorithm of determining the minimum sample size n and the cut-off point t with a given precision. However in most recent business surveys as well as household surveys, non-response rates are getting higher. These non-responses produce non-sampling error which should be controlled for reducing total error.

Especially in modified cut-off sampling, frequently non-responses occur in take-all stratum. The usual methods to solve this problem are weight adjustment and imputation. The weight adjustment method is widely used for handling non-responses. However this method may be properly applied to non-responses occurred in take-some stratum.

On the other hand the imputation method can be used for the both strata. However this method usually needs extra auxiliary variables for achieving desired precision. So for the case of non-existence of extra information from auxiliary variables in take-all stratum this method may fail to achieve the desired precision. Hence the reduction of sample size assigned to take-all stratum may fundamentally solve the problem. That is, small size of take-all stratum may reduce the non-sampling error stemmed from non-responses. Lee and Shin (2011) and Lee (2011) studied about this topic.

In this paper, with given sample size n by the modified cut-off method, we re-calculate the new cut-off point, t_{OP} , that makes take-all stratum size smaller for some cases.

In section 2, we investigate the new cut-off point for log-normal and gamma distribution with given sample size n. In section 3, compare the new cut-off point and

the usual cut-off point suggested by Hidiroglou (1986) using simulated data. Real data analysis is performed in section 4 and summary and conclusion are in section 5.

2. A new cut-off point determination

The cut-off sampling is applied for highly skewed data. So we assume that the underlying survey data distribution follows log-normal distribution or gamma distribution. In this section we determined the new cut-off point by minimizing the variance of total estimator with predetermined sample size n and given population size N.

2.1 Log-normal distribution

First we consider log-normal distribution. The log-normal pdf is given by

$$f(x) = \frac{exp(-\mu + \sigma^2/2)}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{\log(x) - (\mu - \sigma^2)}{\sigma}\right)^2\right] \quad (1)$$

Then we can easily obtained the n-th moment given by

$$E(X^n) = exp\left(n\mu + \frac{1}{2}n^2\sigma^2\right) \quad (2)$$

and so we have $E(X) = exp\left(\mu + \frac{1}{2}\sigma^2\right)$and $Var(X) = exp(2\mu + \sigma^2) \times (exp(\sigma^2) - 1)$.

Now given the cut-off point or truncated point x_0, denote $b_0 = \frac{logx_0 - \mu}{\sigma}$ then the n-th moment of truncated log-normal distribution is obtained by

$$E(X^n|x_0) = exp(n\mu + \frac{1}{2}n^2\sigma^2) \times \frac{\Phi(b_0 - n\sigma)}{\Phi(b_0)}$$

where $\Phi(\cdot)$ is the cumulative normal function. Therefore we can easily obtained the truncated mean and variance using followings.

$$E(X|x_0) = exp(\mu + \frac{1}{2}\sigma^2) \times \frac{\Phi(b_0 - \sigma)}{\Phi(b_0)}$$
$$E(X^2|x_0) = exp(2\mu + 2\sigma^2) \times \frac{\Phi(b_0 - 2\sigma)}{\Phi(b_0)} \quad (3)$$

Now turn back to modified cut-off sampling. Let's denote the population size, N, sample size, n, and cut-off point t which is a function of truncated point x_0. For take-some stratum, we assume that the sample mean is unbiased. Then MSE of estimated total, $\hat{t}_t$, for cut-off sampling is the same as variance. So we have

$$MSE = Var(\hat{t}_t) = \frac{(N-t)(N-n)}{(n-t)} S^2_{[N-t]} \quad (4)$$

Here t is the corresponding take-all stratum size to x_0. Since we want to minimize MSE or variance by t to obtain minimizing point t_{op}, we need to make $S^2_{[N-t]}$ a function of t.

Now from (3), we have

$$S^2_{[n-t]} = exp(2\mu + 2\sigma^2) \times \frac{\Phi(b_0 - 2\sigma)}{\Phi(b_0)} - \left(exp(\mu + \frac{1}{2}\sigma^2) \times \frac{\Phi(b_0 - \sigma)}{\Phi(b_0)}\right)^2$$

where $b_0 = \Phi^{-1}(1 - \frac{t}{N})$. Therefore we have

$$S^2_{[n-t]} = exp(2\mu + 2\sigma^2) \times \frac{\Phi\left(\Phi^{-1}\left(1-\frac{t}{N}\right)-2\sigma\right)}{\left(1-\frac{t}{N}\right)} - \left(exp\left(\mu + \frac{1}{2}\sigma^2\right) \times \frac{\Phi\left(\Phi^{-1}\left(1-\frac{t}{N}\right)-\sigma\right)}{\left(1-\frac{t}{N}\right)}\right)^2 \quad (5)$$

Recall that $b_0 = \frac{log x_0 - \mu}{\sigma}$ and $b_0 = \Phi^{-1}\left(1 - \frac{t}{N}\right)$. Hence we can easily see that t is a function of x_0. Then we can find t_{op} by minimizing (4) with (5). Even though the analytic solution is not easy, but we can easily obtain the solution numerically. Following figure 1. shows $Var(\hat{t}_t)$ with respect to t as an example.

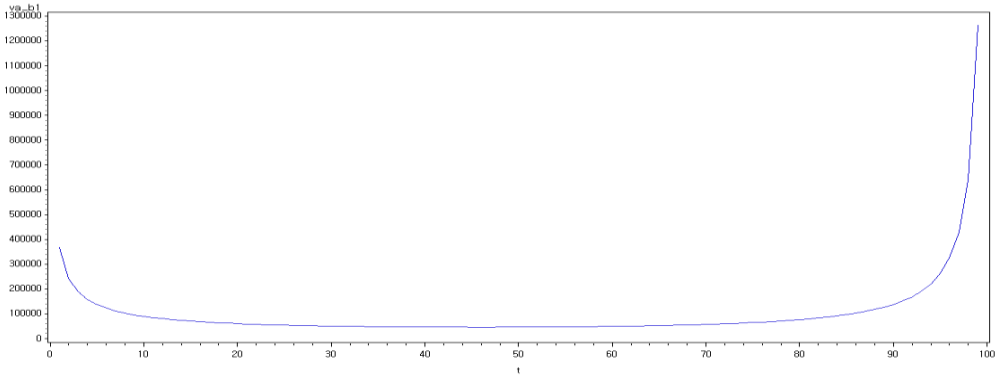

Figure 1. The variance of total estimator by the size of take-all stratum for log-normal distribution. ($\mu = 10,\ \sigma = 2,\ N = 5{,}000,\ n = 100$)

2.2 Gamma distribution

In this section we consider gamma distribution,$\Gamma(\alpha, \beta)$ and gamma pdf is given by

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} exp(-x/\beta)$$

Also the pdf of the truncated gamma distribution is defined by following

$$f(x|\alpha, \beta, x_0) = x^{\alpha-1} exp(-x/\beta)/I(\alpha, \beta, x_0),\ \text{for}\ 0 < x < x_0 \quad (6)$$

where $I(\alpha, \beta, x_0) = \int_0^{x_0} x^{\alpha-1}\, exp(-x/\beta)\, dx$ and $\alpha > 0,\ \beta > 0$.

Then simply we have

$E(X|x_0) = \int_0^{x_0} x^\alpha\, exp(-x/\beta)\, dx / I(\alpha, \beta, x_0) = I(\alpha + 1, \beta, x_0)/I(\alpha, \beta, x_0)$ and
$E(X^2|x_0) = I(\alpha + 2, \beta, x_0)/I(\alpha, \beta, x_0)$.

Therefore we have

$$S^2_{[N-t]} = \frac{I(\alpha + 2, \beta, x_0)}{I(\alpha, \beta, x_0)} - \left[\frac{I(\alpha + 1, \beta, x_0)}{I(\alpha, \beta, x_0)}\right]^2$$

$$= \frac{\Gamma(\alpha+2)\beta^{\alpha+2}CDF - \Gamma(\alpha+2,\beta,x_0)}{\Gamma(\alpha)\beta^{\alpha}CDF - \Gamma(\alpha,\beta,x_0)} - \left[\frac{\Gamma(\alpha+1)\beta^{\alpha+1}CDF - \Gamma(\alpha+1,\beta,x_0)}{\Gamma(\alpha)\beta^{\alpha}CDF - \Gamma(\alpha,\beta,x_0)}\right]^2$$

$$= \frac{\alpha(\alpha+1)\beta^2 CDF - \Gamma(\alpha+2,\beta,x_0)}{CDF - \Gamma(\alpha,\beta,x_0)} - \alpha^2\beta^2\left[\frac{CDF - \Gamma(\alpha+1,\beta,x_0)}{CDF - \Gamma(\alpha,\beta,x_0)}\right]^2 \quad (7)$$

Now since $I(\alpha,\beta,x_0) = \int_0^{x_0} x^{\alpha-1}\, exp(-x/\beta)\, dx$, we have

$$x_0 = I^{-1}\left(1 - \frac{t}{N}\middle|\alpha,\beta\right) \quad (8)$$

Now plugging the (8) and (7) into equation (4) and minimizing equation (4) by t then t_{op} is obtained by numerically. The figure 2. shows $Var(\hat{t}_t)$ with respect to t.

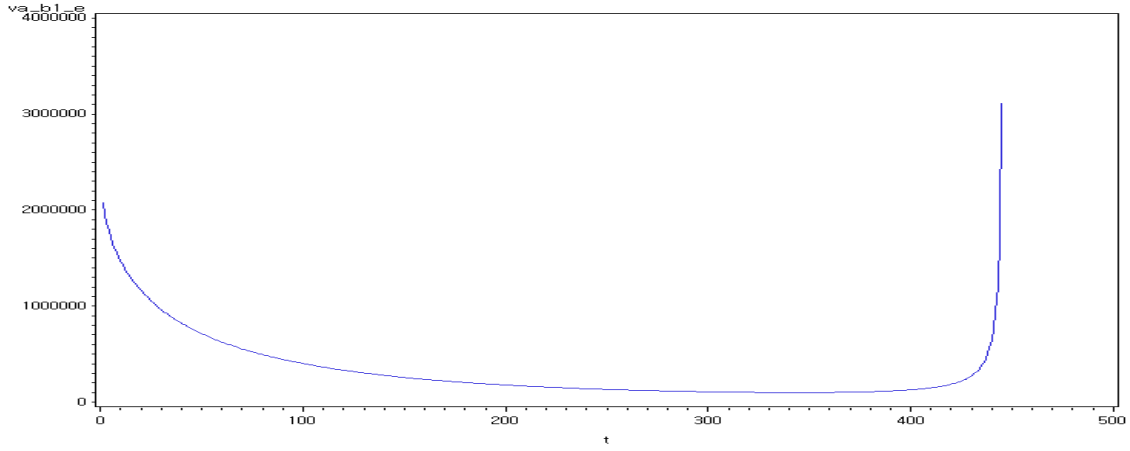

Figure 2. The variance of total estimator by the size of take-all stratum for Gamma distribution ($\alpha = 0.05,\ \beta = 30,\ N = 5{,}000,\ n = 430,\ e = 0.05$)

3. Comparison of cut-off points

The cut-off point suggested by Hidiroglou (1986) and the new cut-off point are compared with simulated data with log-normal and gamma distribution. For given $N = 5{,}000$ and μ, σ^2 log-normally distributed data is generated. Also with $N = 5{,}000$ and α and β, we generate gamma distributed data. Also with expected error, $e = 0.03, 0.05$, the sample size n is predetermined by using Hidiroglou (1986) method.

3.1 Log-normal distribution

For given $N = 5{,}000$, $\mu = 10$ and various σ from 0.5 to 5, log-normally distributed data is generated. The table 3.1 shows the sample size, n and cut-off point, t_H by using Hidiroglou method. Also the new cut-off point, t_{OP}, obtained by developed in this paper is also tabulated.

Table 3.1. Optimal cut-off point for Log-normal distribution

e	σ	n	t_H	t_{OP}
0.03	0.5	1447	565	818
	1.0	1411	698	789
	1.5	1304	708	708
	2.0	1189	675	623
	2.5	1079	637	546
	3.0	977	585	477
	4.0	799	500	365
	5.0	655	413	281
0.05	0.5	809	201	371
	1.0	863	330	404
	1.5	830	373	384
	2.0	774	376	350
	2.5	712	368	314
	3.0	652	353	279
	4.0	540	304	219
	5.0	444	259	172

In table 3.1, as σ increases, n goes up and down. Also t_H and t_{OP} goes up and down as σ increases. However t_{OP} goes down faster than t_H. So beyond some values of σ, we have $t_{OP} \leqq t_H$. That means for large values of σ, we can reduce the size of take-all stratum.

3.2 Gamma distribution

With $N = 5{,}000$, $\beta = 30$ and various α from 0.001 to 1, we generate gamma distributed data. With $e = 0.03$, 0.05 and using Hidiroglou method, we calculate the sample size n and cut-off point t_H. Also we calculate t_{OP} suggested cut-off point in this study. Since the take-all stratum size does not depend on β, we do not consider various β in simulation. Table 3.2 shows the results.

Table 3.2 Optimal cut-off point for gamma distribution

e	α	n	t_H	t_{OP}
0.03	0.001	28	25	25
	0.005	96	82	84
	0.01	181	156	158
	0.03	402	324	335
	0.5	1441	873	902
	1	1592	793	865
0.05	0.001	26	24	23
	0.005	82	71	71
	0.01	154	125	131
	0.03	331	268	266
	0.5	959	450	485
	1	958	299	372

From table 3.2, the t_H is always smaller than t_{OP}. Hence instead of using t_{OP}, t_H may reduce the non-sampling error by reducing the size of take-all stratum.

4. Real data analysis

The Korea briquette consumption survey data in 2012 is used for analysis. Since this data do not follow log-normal distribution, we can not use the results developed in section 2. However we can directly calculate the equation (4) using numerical method. For this we calculate cut-off point, t_H , suggested by Hidiroglou and sample size n. And also calculating the equation (4) with size n, numerically we obtain the new cut-off point t_{OP}. The results are in Table 4.1.

Table 4.1 Optimal cut-off point for Korea briquette consumption survey

e	n	t_H	t_{OP}
0.01	1196	1034	970
0.02	983	752	619
0.03	819	546	466
0.05	589	305	257

From table 4.1, the new cut-off point, t_{OP} is smaller than t_H. This result shows that for some data, we can reduce the size of take all stratum by minimizing the equation(4) directly.

5. Summary and conclusion

This new cut-off point may reduce the number of non-responses and the non-sampling by reducing the size of take-all stratum itself.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education science and Technology(2012R1A1A2003919).

References

1. Hidiroglou, M. A., (1986). The construction of a self-representing stratum of large units in survey design, The American Statistician, Vol. 4, No. 1, 27-31.
2. Lee, S. E. Shin, K.-I., (2011), Alternative determination of cut-off point based on MSE, Proceedings of ISI 2011, Dublin.
3. Lee, S. E., (2011), The cut-off point based on MSE in modified cut-off sampling, Journal of the Korean Official Statistics, Vol. 16, No. 1, 82-94.