Truncated Triangular Distribution for Multiplicative Noise and Domain Estimation

Jay J. Kim [1], Dong M. Jeong [2]

[1] National Center for Health Statistics, Centers for Disease Control and Prevention, 3311 Toledo Road, Hyattsville, MD 20782

[2] Korea National Statistical Office, Building 13F, National Credit Union, 949 Dunsan-dong, Seo-gu, Daejeon, 302-120, Republic of Korea

Abstract

The truncated triangular distribution has been used for masking microdata. The random variable following the truncated triangular distribution can serve as a multiplicative noise factor for the masking. The most desirable candidate distribution is the symmetric one which is centered at and truncated symmetrically about 1. This is because the multiplicative noise factor of 1 or very close to 1 does not protect confidentiality at all. The probability density function of the truncated triangular distribution has been developed by Kim [4] and applied to the 2006 Korean Householder Income and Expenditure Survey (HIES) data. Formulas for the domain estimation for the data masked by the multiplicative noise mentioned above are developed. In this paper, we will show domain estimation formulas and some results of the application of the truncated triangular distribution on the HIES data.

Key Words: Masking, confidentiality, truncation, noise generation

1. Introduction

Since around 1980, the U.S. Energy Information Administration (EIA) has been using multiplicative noise for masking the number of heating and cooling days in an area, etc, in their public use micro data file from the Residential Energy Consumption Survey. EIA uses noise which follows the truncated normal distribution [Hwang, (2)]. Evans, et al [1] proposed the use of multiplicative noise to mask economic data. They considered noise which follows distributions such as normal and truncated normal distributions. Kim and Winkler [4] considered multiplicative noise which follows the truncated normal distribution. The U.S. Bureau of the Census uses a truncated triangular distribution for masking the Commodity Flow Survey data. Kim [5] developed the probability density function (pdf) of the truncated triangular distribution and showed that the estimate from the data masked by the distribution is unbiased if the triangular distribution is symmetric about 1 and truncated symmetrically about 1. In this paper, we will review noise that follows the truncated triangular distribution, develop formulas for domain estimation, and report the results of applying noise to the HIES data.

Multiplicative noise has the following form:

$$y_i = x_i e_i ,\ i = 1,\ 2, \ldots\ldots , n ,$$

where y_i is the masked variable for the i^{th} unit such as person, household, establishment, etc., x_i is the corresponding un-masked variable and e_i (>0) is the noise.

Since noise is generated independently of the original data, x_i and e_i are independent.

$$E(y) = E(x)E(e) .$$

If $E(e)=1$ is used, then $E(y)=E(x)$.

Also

$$V(y)=V(x)V(e)+\mu_e^2V(x) + \mu_x^2V(e)\,, \tag{1}$$

where $\mu_x = E(x)$ and $\mu_e = E(e)$.

Thus

$$V(x) = \frac{V(y) - \mu_x^2V(e)}{V(e)\ + \mu_e^2}\,. \tag{2}$$

Letting $\mu_y = E(y)$ and imposing the condition that $\mu_y = \mu_x$ and $\mu_e = 1$, we can express equation (2) as

$$V(x) = \frac{V(y) - \mu_y^2V(e)}{V(e)\ + 1}\,. \tag{3}$$

If the data disseminating agency provides data users with the variance of the noise variable used, users can estimate the variance of the original data using the above formula. In the process, estimates of $V(y)$ and μ_y from the released data are substituted into equation (3).

2 Truncated Triangular Distribution

2.1 Triangular Distribution

A truncated triangular distribution is a modified form of a triangular distribution, and thus we first consider a triangular distribution which is shown in Figure 1. The triangular distribution is very useful. It can be used for approximating the normal, gamma and beta distributions. The triangular distribution is analytically easier to handle than the normal distribution.

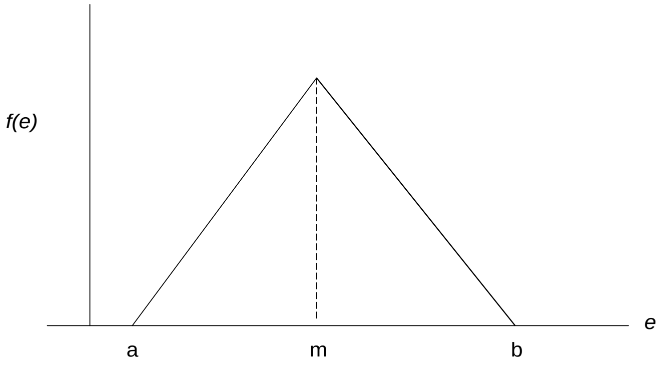

Figure 1. Triangular Distribution

The triangular distribution of a random variable e as shown above has the following form.

$$f(e) = \begin{cases} \frac{2}{b-a}\frac{e-a}{m-a}, & a \le e < m \\ \frac{2}{b-a}\frac{b-e}{b-m}, & m \le e < b \end{cases} \quad (4)$$

Note that *m* is the mode of this distribution. If the distribution is symmetric about *m*, *m* is also the population mean and median of *e*. If the distribution is symmetric about m, $m-a=b-m$ or $a+b=2m$, then equation (4) reduces to

$$f(e) = \begin{cases} \frac{e-a}{(m-a)^2}, & a \le e < m \\ \frac{b-e}{(b-m)^2}, & m \le e < b. \end{cases} \quad (5)$$

The expected value of e is

$$E(e) = \frac{m+a+b}{3}. \quad (6)$$

Note that the expected value of *e* is a simple mean of the minimum, maximum and mode of *e*. Suppose the triangular distribution is symmetric about *m*, then equation (6) reduces,

$$E(e) = m.$$

The variance of e is

$$V(e) = \frac{b^2 - ab + a^2 + m^2 - m(a+b)}{18}. \quad (7)$$

If the triangular distribution is symmetric about *m*, then equation (7) reduces to

$$V(e) = \frac{(b-m)^2}{6}. \quad (8)$$

2.2 Truncated Triangular Distribution

When triangular distribution-based noise is used, one must avoid using the number close to one (1) for noise, because multiplying by a number very close to 1 does not change the original value that much, and thus the original value does not get any protection. In addition, the probability density for *e* is the greatest, when *e* is near 1. This implies that the largest number of units do not get protection. Hence, it has been suggested [Evans, et al (1)] to truncate the mid-section, or the section near 1 of the triangular distribution. The truncated triangular distribution has the following shape.

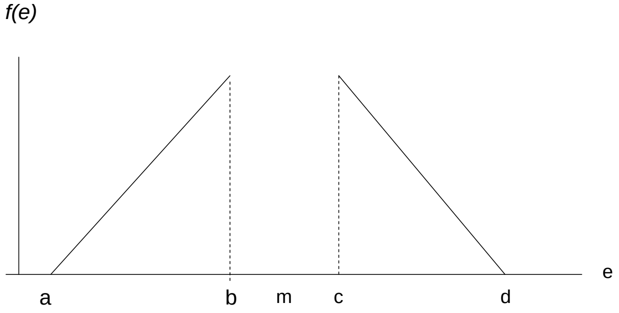

Figure 2. Truncated Triangular Distribution

Suppose the distribution is truncated at b and c, $c > b$, as shown in Figure 2. In this case, the pdf has the following form [Kim, (5)]:

$$f(e) = \begin{cases} \dfrac{2(d-m)}{(b-a)^2(d-m)+(d-c)^2(m-a)}(e-a), & a \le e < b \\ \dfrac{2(m-a)}{(b-a)^2(d-m)+(d-c)^2(m-a)}(d-e), & c \le e < d\,. \end{cases} \tag{9}$$

If the triangular distribution is symmetric about *m* and the truncation is also symmetric about *m*, then the formula for the pdf reduces to

$$f(e) = \begin{cases} \dfrac{e-a}{(d-c)^2}, & a \le e < b \\ \dfrac{d-e}{(d-c)^2}, & c \le e < d\,. \end{cases} \tag{10}$$

The expected value of e without assuming symmetry of the distribution and truncation is

$$E(e) = \frac{(d-m)(b-a)^2(2b+a) \; + (m-a)(d-c)^2(2c+d)}{3[(b-a)^2(d-m)+(d-c)^2(m-a)]} \tag{11}$$

Suppose again the triangular distribution is symmetric about *m* and the mid-section of the distribution is truncated symmetrically about *m*. Then

$$E(e) = m\,. \tag{12}$$

Thus, the mean of the symmetric triangular distribution, whose mid-section is symmetrically truncated, is the same as that of the un-truncated triangular distribution. Consequently, the data masked by noise following the symmetrically truncated symmetric triangular distribution with *m*=1 will provide an unbiased mean of the original data.

The variance of e is

$$V(e)=\frac{1}{18\left[(b-a)^2(d-m)+(d-c)^2(m-a)\right]^2}\left\{(b-a)^6(d-m)^2+(d-c)^6(m-a)^2\right.$$

$$\left.+(b-a)^2(d-m)(d-c)^2(m-a)\left[(3b+a)^2+(3c+d)^2+2\left(a^2+d^2\right)-4(2b+a)(2c+d)\right]\right\}. \qquad (13)$$

For the symmetric triangular distribution with symmetric truncation about *m*, we have

$$V(e)=m^2-\frac{16mc+8md-2(d^2+2dc+3c^2)}{12}. \qquad (14)$$

3 Domain Estimation

Let a superscript s indicate domain s. Then for the domain s,

$$y_i^s=x_i^s e_i,\ i=1,\ 2,\ .\ .\ .\ .\ ,n^s. \qquad (15)$$

Note in the above that n^s is the domain size. Since we do not generate different noise for each domain separately, we do not have a superscript *s* for e_i in equation (15).

Since noise is generated independently of the original unmasked variable,

$$E(y^s)=E(x^s)E(e). \qquad (16)$$

Hence,

$$E(x^s)=\frac{E(y^s)}{E(e)}. \qquad (17)$$

Using $E(e)=1$,

$$E(y^s)=E(x^s). \qquad (18)$$

From equation (1), we have for a domain *s*,

$$V(y^s)=V(x^s)V(e)+\mu_e^2V(x^s)+(\mu_x^s)^2V(e). \qquad (19)$$

In the above, $\mu_x^s=E(x^s)$. With $\mu_e=E(e)=1$ and $\mu_y^s=E(y^s)=\mu_x^s$,

$$V(x^s)=\frac{V(y^s)-(\mu_y^s)^2V(e)}{V(e)\ +1} \qquad (20)$$

Using equation (20), data users can estimate the variance of a domain.

4 Application of the Masking Scheme to a Survey Data

We applied the truncated triangular distributed noise approach to the Korea National Statistical Office (KNSO)'s Household Income and Expenditures Survey (HIES) data [Jeong, (3)]. For generating noise using the truncated triangular distribution, we tried four different sets of parameters, that is, four different combinations of minimum noise, maximum noise and lower and upper truncation points for noise as shown in Table 1 below. Each combination corresponds to a method in Table 1 below.

Table 1. Parameters for Truncated Triangular Distribution for Generating Noise (*e*)

Method	Minimum Noise	Lower Truncation Point	Mode	Upper Truncation Point	Maximum Noise
I	0.6	0.99	1.0	1.01	1.4
II	0.6	0.90	1.0	1.10	1.4
III	0.4	0.99	1.0	1.01	1.6
IV	0.4	0.90	1.0	1.10	1.6

The ranges of noise for Methods I and II are the same. Similarly, the ranges for Methods III and IV are the same. However, the ranges of noise for Methods I and II are narrower than those for Methods III and IV. The truncation regions for Methods I and III are the same. Similarly, the truncation regions for Methods II and IV are the same. Note the width of the truncation regions for Methods I and III are narrower than those for Methods II and IV. In summary, Method I is treated most favorably in terms of the narrower range and truncation region for noise. On the other hand, Method IV is treated least favorably.

After generating noise, we calculated the mean, minimum and maximum for each of the four noise datasets. The parameter and observed values of noise are shown in Table 2.

Table 2. Parameter and Observed Values for the Generated Noise (*e*)

Method		Mean	Minimum	Maximum
I	Parameter	1.0000	0.6000	1.4000
	Observed	1.0003	0.6041	1.3916
II	Parameter	1.0000	0.6000	1.4000
	Observed	0.9994	0.6041	1.3964
III	Parameter	1.0000	0.4000	1.4000
	Observed	1.0004	0.4062	1.5874
IV	Parameter	1.0000	0.4000	1.4000
	Observed	0.9987	0.4062	1.5947

In Table 2, most observed values are close to their corresponding parameter values. However, the observed maximum values for Methods III and IV are more than 10 percent away from their parameter values. This may be because their noise distributions can take values farther from 1.

Table 3. Means of Wages and Food Costs Data Masked by Truncated Triangular Distributed Noise (*e*) by Methods and Geographic Domains

	Area	Original Data	Method I	Method II	Method III	Method IV
Wages	Nation	1,967,254	1,966,919	1,968,925	1,966,025	1,962,908
	Seoul	2,133,225	2,123,645	2,141,393	2,114,724	2,124,118
	Others	1,941,942	1,943,018	1,942,622	1,943,348	1,938,323
Food Costs	Nation	175,216	175,264	175,142	175,441	175,161
	Seoul	194,296	194,334	194,500	194,360	194,115
	Others	172,460	172,509	172,346	172,708	172,423

Table 3 shows the means of the data masked by truncated triangular distributed noise along with that of the original data. To help compare the methods, we calculated an absolute relative difference between means for the original and masked data using the original mean as the norm. The results in percent are shown in Table 4.

Table 4. Absolute Relative Difference (in Percent) between Means for Original and Masked Data – Truncated Triangular Distributed Noise

	Area	Method I	Method II	Method III	Method IV
Wages	Nation	0.017	0.085	0.062	0.221
	Seoul	0.449	0.383	0.867	0.427
	Other	0.055	0.035	0.072	0.186
Food Costs	Nation	0.027	0.042	0.128	0.031
	Seoul	0.020	0.105	0.033	0.093
	Other	0.028	0.066	0.144	0.021

The absolute relative differences in Table 4 are ranked from the smallest to the largest values among four methods, as shown in Table 5.

Table 5. Ranks of Four Methods for Means – Truncated Triangular Distributed Noise

	Area	Method I	Method II	Method III	Method IV
Wages	Nation	1	4	3	2
	Seoul	3	1	4	2
	Other	2	1	3	4
Food Costs	Nation	1	3	4	2
	Seoul	1	4	2	3
	Other	2	3	4	1

In general, the differences for Method I are the smallest or close to the smallest among all methods. Note that Method I got the most favorable treatment. One exception is the mean household wages for Seoul. The relative difference for Seoul is much higher than those for the other areas and for all averages of food costs. The difference for Seoul mean household wages turns out to be relatively large disregarding the methods.

The standard deviations for the masked and original data are shown in Table 6 below.

Table 6. Standard Deviations of Data Masked by Truncated Triangular Distributed Noise by Methods and Geographic Domains

	Area	Original Data	Method I	Method II	Method III	Method IV
Wages	Nation	1,238,689	1,238,159	1,284,962	1,298,934	1,389,932
	Seoul	1,371,615	1,345,896	1,431,407	1,394,682	1,481,191
	Others	1,215,169	1,219,069	1,259,080	1,282,207	1,297,701
Food Costs	Nation	149,618	154,834	154,659	164,666	163,828
	Seoul	169,632	173,319	164,678	186,269	176,604
	Others	146,297	146,087	147,318	155,452	156,104

To gain a better insight into differences among the methods, absolute relative differences were computed and are shown in Table 6.

Table 7. Absolute Relative Difference between Standard Deviations for Original and Masked Data – Truncated Triangular Distributed Noise

	Area	Method I	Method II	Method III	Method IV
Wages	Nation	0.043	3.736	4.864	12.210
	Seoul	1.875	4.359	1.682	7.989
	Others	0.321	3.614	5.517	6.792
Food Costs	Nation	3.486	3.369	10.058	9.498
	Seoul	2.174	2.920	9.808	4.110
	Others	0.144	0.698	6.258	6.703

The absolute relative differences for standard deviations are much higher than those for means. Generally, differences for Method I are smaller than the others.

The absolute relative differences in Table 7 are ranked from the smallest to the largest values among four methods, as shown in Table 8.

Table 8. Ranks of Four Methods for Standard Deviations – Truncated Triangular Distributed Noise

	Area	Method I	Method II	Method III	Method IV
Wages	Nation	1	2	3	4
	Seoul	2	3	1	4
	Others	1	2	3	4
Food Costs	Nation	2	1	4	3
	Seoul	1	2	4	3
	Others	1	2	3	4

Table 8 shows the ranks for Method I are 1 except for 2 cases, where they are 2. When Method I's rank is 2, it is close second.

5 Concluding Remarks

For masking the Commodity Flow Survey data, the U.S. Bureau of the Census uses multiplicative noise that follows the truncated triangular distribution. Kim (5) developed the probability density function (pdf) for the truncated triangular distribution. In this paper, we developed formulas for domain estimation.

The truncated triangular distributed noise was applied to the 2006 Korean Household Income and Expenditures Survey data. Depending on the width of the distribution and the truncation region used to generate noise, we can have different sets of noise. For this study, we generated four sets of noise, which are labeled "methods I - IV." The most favorable treatment was given to Method I: a narrower distribution and truncation region. Method IV was given the worst treatment: a wider distribution and truncation region. We masked the data using the noise and calculated means and standard deviations for each of the four masked datasets. As could be conjectured, Method I generally produces mean and variance closest to those of the original data and Method IV farthest from them. The estimates of the mean based on Method I are generally excellent. The estimates of the standard deviations are not as precise as those of means. Note that domain estimation formulas were used to estimate the standard deviations for Seoul and Other areas.

6 References

1. Evans, T., Zayatz, L., and Slanta, J. (1998) Using Noise for Disclosure Limitation of Establishment Tabular Data, Journal of Official Statistics, Vol. 14, No. 4, 537-551.
2. Hwang, J.T. (1986) Multiplicative Errors-in-Variables Models with Applications to Recent Data Released by the U.S. Department of Energy, Journal of the American Statistical Association, Vol. 81, No. 395, 680-688.
3. Jeong, D.M. (2008) Schemes for Masking the Household Income and Expenditures Survey Data, Internal Memorandum. Korea National Statistics Office.
4. Kim, J.J. and Winkler, W. E. (2001) Multiplicative Noise for Masking Continuous Data, Proceedings of the Survey Methods Research Section, American Statistical Association, CD Rom.
5. Kim, J.J. (2007), Application of Truncated Triangular and Trapezoidal Distributions for Developing Multiplicative Noise, Proceedings of the Survey Methods Research Section, American Statistical Association, CD Rom.

The findings and conclusions in this paper are those of the author and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.