Application of the Truncated Triangular and the Trapezoidal Distributions

Jay J. Kim
ORM, NCHS, 3311 Toledo rd, Hyattsville, MD, 20782

Abstract

It has been suggested that the truncated triangular distribution be used for masking microdata. The random variable which follows the truncated triangular distribution serves as a multiplicative noise factor. The natural candidate distribution is the one which is centered at 1, but truncated symmetrically around 1, because it is not desirable to use some values which are too close to 1 due to confidentiality concerns. Instead of truncating the middle section of the distribution, we can assign low probabilities for that part of the distribution. This procedure will render some variations of the trapezoidal distribution. Two other approaches are to use the upside-down triangular distribution or parabloa for the middle section of the distribution. In this paper, we will derive the probability density function for each of the four distributions and investigate their properties.

KEY WORDS: confidentiality, masking, truncated triangular distribution

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 and marginal rates for utility, such as electricity and gas in their public use micro data file from the Residential Energy Consumption Survey. Evans, et al (1998) proposed use of multiplicative noise to mask economic data. They considered noise which follows distributions such as normal and truncated normal distributions. Kim and Winkler (2002, 2003) considered a noise distribution following the truncated normal distribution. Massell and Russell (2006) proposed to use a truncated triangular distribution for masking economic data. In this paper, we will consider noise that follows the truncated triangular, modified trapezoidal, double triangular and parabola-filled truncated triangular distributions. As the above probability density functions are not available at this moment, we will derive them and investigate some of their properties in this paper.

Multiplicative noise has the following form:

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

where y_i is the masked variable for the i^{th} unit such as person, household, establishment or governmental unit, x_i is the corresponding un-masked variable and e_i is the noise.

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

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

Usually *E(e)= 1* is used, and thus

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

Also

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

Thus

$$V(x) = \frac{V(y) - \mu_x^2 V(e)}{V(e) + \mu_e^2} \qquad (1)$$

Since the data disseminating agencies provide the mean and variance of the noise distribution to the data users, users can estimate the variance of characteristics from the masked data using the formula in equation (1).

2. Noise Distributions.

2.1 Triangular Distribution

Since a truncated triangular distribution is a modified form of a triangular distribution, we consider a triangular distribution first.

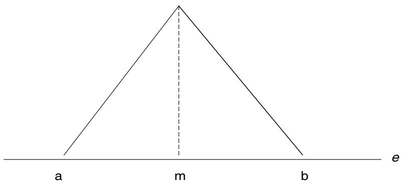

Figure 1. Triangular Distribution

Triangular distribution of *e,* as shown above has the following form.

$$f(e)=\begin{cases}\dfrac{2}{b-a}\dfrac{e-a}{m-a}, & a\le e<m\\[2ex] \dfrac{2}{b-a}\dfrac{b-e}{b-m}, & m\le e<b\end{cases}\qquad(2)$$

Note that *m* is the mode of this distribution. If it is a symmetric distribution around *m*, *m* is also the mean and median of the variable *e*. If the distribution is symmetric around *m*, i.e., $m-a=b-m$ then equation (2) reduces to

$$f(e)=\begin{cases}\dfrac{e-a}{(m-a)^2}, & a\le e<m\\[2ex] \dfrac{b-e}{(b-m)^2}, & m\le e<b\end{cases}\qquad(3)$$

The expected value of *e* is

$$E(e)=\frac{m+a+b}{3}\qquad(4)$$

Suppose the triangular distribution is symmetric around *m*, i.e., $m-a=b-m$ or $a+b=2m$, then the above expectation reduces to *m*.

The variance of *e* is

$$V(e)=\frac{b^2-ab+a^2+m^2-m(a+b)}{18}.\qquad(5)$$

If $m-a=b-m$ or $a+b=2m$, then the above variance reduces to

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

2.2 Truncated Triangular Distribution

When noise following a triangular distribution is used, using 1 for *m* is not desirable. This is because multiplication by 1 does not change the original value at all. That is, if the original value is multiplied by 1 for a unit, the unit does not get any protection from disclosure. Even worse, the probability of noise being 1 is the highest in this type of distribution. This implies that the largest number of units do not get protection. Thus Massell and Russell suggested using noise that follows the triangular distribution whose middle portion, or the part near 1, is truncated. The truncated triangular distribution mentioned above is shown in Figure 2.

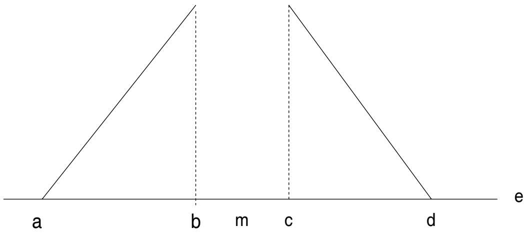

Figure 2. Truncated Triangular Distribution

Suppose the distribution is truncated at b and c, $c>b$, as shown above, which are around *m.* One of the vertices, denoted by b in equation (2), is now denoted by *d.* In this case, the probability density function (pdf) of *e* has the following form:

$$f(e)=\begin{cases}\dfrac{2(e-a)}{k(m-a)(d-a)}, & a\le e<m\\[2ex] \dfrac{2(d-e)}{k(d-m)(d-a)}, & m\le e<d\end{cases}\qquad(7)$$

where *k* needs be determined.

Now

$$\int_a^b\frac{2(e-a)}{k(m-a)(d-a)}de+\int_c^d\frac{2(d-e)}{k(d-m)(d-a)}de$$

$$= \frac{(b-a)^2}{k(m-a)(d-a)} + \frac{(d-c)^2}{k(d-m)(d-a)} \ . \qquad (8)$$

In order for the function in equation (7) to become a pdf, equation (8) needs to be 1. Thus

$$k = \frac{(d-m)(b-a)^2 + (m-a)(d-c)^2}{(d-a)(m-a)(d-m)}$$

Using the above k, the truncated triangular distribution of e, truncated at b and c is,

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

If the triangular distribution is symmetric around m and the truncation is also symmetric around m, then equation (9) can be simplified as follows:

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

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

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

$$+ \frac{(m-a)(d-c)^2(2c+d)}{3[(b-a)^2(d-m)+(d-c)^2(m-a)]} \qquad (11)$$

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

$$E(e)=m \qquad (12)$$

Thus, the mean of the symmetrically truncated triangular distribution, where the mid-section of the distribution 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.

Example 1.

If a = 0.5, d = 1.5, m = 1, b = 0.75 and c = 1.25, then $E(e)$ without truncation is

$$E(e) = \frac{1+0.75+1.25}{3} = 1.$$

The expected value of e after the truncation is

$E(e) = 1$, since m = 1.

As shown in the formulas, they are the same.

The variance of e of the truncated triangular distribution is,

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

$$\Big\{(b-a)^6(d-m)^2+(d-c)^6(m-a)^2$$

$$+(b-a)^2(d-m)(d-c)^2(m-a)\big[(3b+a)^2$$

$$+(3c+d)^2+2\left(a^2+d^2\right)-4(2b+a)(2c+d)\big]\Big\} \ (13)$$

If the truncation is symmetric around m and the triangular distribution is also symmetric around m, then

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

Example 2. Using the data in Example 1, $V(e)$ without truncation is .04167, but after truncation, it is .1146. The variance of e after truncation is 2.75 times that of the variable e without truncation.

The variance of the truncated triangular distribution is greater, since the values around the mean, which have the highest frequencies, are not there anymore.

2.3 Trapezoidal Distribution

If we connect the top two points in the truncated part of the truncated triangular distribution, we have a trapezoidal distribution. Instead of not allowing noise (e) to have any positive probability between b and c, we can allow some small (equal) probability. In that case, we have a modified trapezoidal distribution, which is shown in Figure 4. To derive the probability density function for the modified trapezoidal distribution, we start with the trapezoidal distribution in Figure 3.

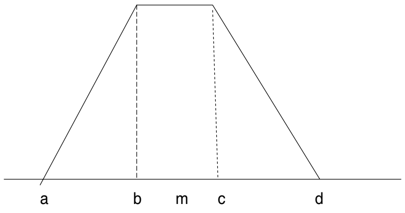

Figure 3. Trapezoidal Distribution

Let the trapezoidal distribution have the following form.

$$f(e)=\begin{cases}\dfrac{2k_1}{(d-a)(b-a)}(e-a), & a\le e<b\\[2ex] \dfrac{2k_2}{d-a}, & b\le e<c\\[2ex] \dfrac{2k_3}{(d-a)(d-c)}(d-e), & c\le e<d\end{cases}\qquad(15)$$

Depending on values of k's, the height of top line changes.

If $f(b)=f(c),\ k_1=k_2=k_3=\dfrac{d-a}{d+c-b-a}$,

then the distribution has the following form:

$$f(e)=\begin{cases}\dfrac{2}{d+c-b-a}\ \dfrac{e-a}{b-a}, & a\le e<b\\ \dfrac{2}{d+c-b-a}, & b\le e<c\\ \dfrac{2}{d+c-b-a}\ \dfrac{d-e}{d-c}, & c\le e<d\end{cases}\qquad(16)$$

The above is the same form as given by van Dorp and Kotz.

Suppose the trapezoidal distribution is symmetric around m, then $E(x)=m$.

2.4 Modified Trapezoidal Distribution

When numbers around 1 are multiplied for masking records, it is better to avoid using 1 or numbers too close to 1. This is the reason why the truncated normal and triangular distributions are considered as candidates for a noise distribution, where truncation is done very near to 1. An alternative approach to the truncated distribution is using a distribution which allows a small probability, instead of a zero probability, in the mid-section of the noise distribution. The shape of the noise distribution then will have a variation of the trapezoidal distribution (see Figure 4 below), where there is no connectivity in the two highest points and the mid-section is dropped down.

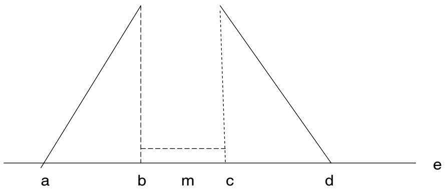

Figure 4. Modified Trapezoidal Distribution

Suppose $k_1=k_3=k$ and $k_2=\dfrac{k}{q}$. Then

$$f(e)=\begin{cases}\dfrac{2k}{d-a}\ \dfrac{e-a}{b-a}, & a\le e<b\\ \dfrac{2k}{(d-a)q}, & b\le e<c\\ \dfrac{2k}{d-a}\ \dfrac{d-e}{d-c}, & c\le e<d\end{cases}\qquad(17)$$

$$\int_\Omega f(e)de=\frac{2k}{(d-a)(b-a)}\int_a^b(e-a)de\ +$$

$$\frac{2k}{q(d-a)}\int_b^c 1de+\frac{2k}{(d-a)(d-c)}\int_c^d(d-e)de\quad(18)$$

The equation above needs to sum up to 1 in order for the function in equation (17) to be a probability density function.

Equation (18) becomes

$$\frac{k[q(d-c+b-a)+2(b-c)]}{(d-a)q} \qquad (19)$$

Hence,

$$k=\frac{(d-a)q}{q(d-c+b-a)+2(b-c)} \qquad (20)$$

Plugging k in the equation (17), we get

$$f(e)=\begin{cases}\dfrac{2q}{q(d-c+b-a)+2(c-b)}\dfrac{e-a}{b-a}, \\ \qquad for\ \ a\le e<b \\ \dfrac{2}{q(d-c+b-a)+2(c-b)}, \\ \qquad for\ b\le e<c \\ \dfrac{2q}{q(d-c+b-a)+2(c-b)}\dfrac{d-e}{d-c}, \\ \qquad for\ \ c\le e<d \end{cases}$$

(21)

The height of the middle section of the distribution depends on the value of q. If a small probability in the mid-section is desired, then a large q should be used. If a smaller q is used, the probability will go up.

The expected value of this distribution is the same as the regular trapezoidal distribution.

2.5 Double Triangular Distribution

The middle truncated section of the triangular distribution can be filled with an upside-down triangular distribution. In that case, the middle triangle should be narrower than the outer triangle before truncation. Here we have a double triangular distribution as shown in Figure 5. Note that b and c can be farther away from m than the truncated triangular distribution, because we don't want to have high probabilities of e near m.

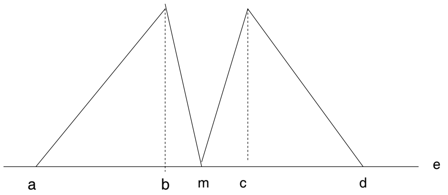

Figure 5. Double Triangular Distribution

The "double" triangular distribution shown above, assuming two triangles are symmetric around m, is

$$f(e)=\begin{cases}\dfrac{1}{(m-a)(b-a)}(e-a), & a\le e<b \\ \dfrac{1}{(m-a)(m-b)}(m-e), & b\le e<m \\ \dfrac{1}{(d-m)(c-m)}(e-m), & m\le e<c \\ \dfrac{1}{(d-m)(d-c)}(d-e), & c\le e<d \end{cases} \qquad (22)$$

$$E(e)=\frac{1}{m-a}\left[\frac{-ab-a^2}{6}+\frac{m^2+mb}{6}\right]$$

$$+\frac{1}{d-m}\left[\frac{-mc-m^2}{6}+\frac{d^2+dc}{6}\right] \qquad (23)$$

Suppose $m - b = c - m,\ m - a = d - m$. Equation (23) can be simplified as follows.

$$E(e)=\frac{2m+(a+d)+(b+c)}{6}=m \qquad (24)$$

2.5 Parabola-Filled Truncated Triangular Distribution

The truncated part of the triangular distribution can be filled with a parabola (see Figure 6). It even can be filled with a u-shaped parabola. The latter can be more desirable in the sense that the probabilities of e near m would be very small in the u-shaped distribution.

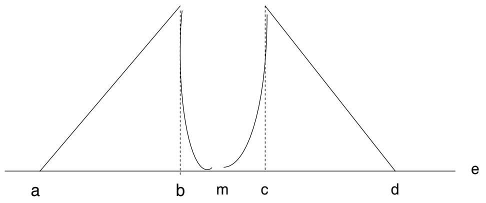

Figure 6. Parabola-Filled Truncated Triangular Distribution

Let

$$f(e)=\begin{cases}\dfrac{1}{(m-a)(b-a)}(e-a), & a\le e<b\\[2ex] k(x-m)^2, & b\le e<c\\ \dfrac{1}{(d-m)(d-c)}(d-e), & c\le e<d\end{cases}\qquad(25)$$

To determine k in equation (25), we set the following to 1.

$$\int_\Omega f(e)de=\int_a^b\frac{1}{(m-a)(b-a)}(e-a)\,de+\int_b^c k(e-m)^2\,de$$

$$+\int_c^d\frac{1}{(d-m)(d-c)}(d-e)de\ .$$

The above is,

$$\frac{b-a}{2(m-a)}+\frac{d-c}{2(d-m)}+k[\frac{c^3-b^3}{3}$$

$$-m(c^2-b^2)+m^2(c-b)]$$

Assuming a symmetric distribution, i.e., $m - a = d - m$, we have

$$\frac{b-a+d-c}{2(d-m)}+k[\frac{c^3-b^3}{3}-m(c^2-b^2)+m^2(c-b)]=1$$

Thus

$$k=\frac{3}{2(d-m)[c^2+bc+b^2-3m(c+b)+3m^2]}$$

Hence

$$f(e)=\begin{cases}\dfrac{1}{(m-a)(b-a)}(e-a), & a\le e<b\\[2ex] \dfrac{3(e-m)^2}{2(d-m)[c^2+bc+b^2-3m(c+b)+3m^2]}, & \\ \qquad\qquad for\quad b\le e<c & \\ \dfrac{1}{(d-m)(d-c)}(d-e), & c\le e<d\end{cases}$$

$$(26)$$

The expected value of e is,

$$E(e)=\frac{1}{m-a}\left(\frac{2b^2-ba-a^2}{6}\right)$$

$$+\frac{1}{d-m}\left(\frac{d^2+dc-2c^2}{6}\right)$$

$$+\frac{(c-b)\Big[3(c+b)(c^2+b^2)-8m(c^2+cb+b^2)}{8(d-m)[c^2+bc+b^2-3m(c+b)+3m^2]}$$

$$+\frac{6m^2(c+b)\Big]}{8(d-m)[c^2+bc+b^2-3m(c+b)+3m^2]}$$

Assuming a symmetric distribution for the parabola and triangular distribution, the above expected value reduces to m.

3. Concluding Remarks

It has been suggested to use noise following a truncated triangular distribution for masking sensitive variables using a multiplicative noise scheme. However, the formula for its density function has not been available. This paper developed the formula. Furthermore, three other candidates for the noise distribution, which are modifications of the truncated triangular distribution, are also developed here. If the triangular distribution is symmetric and the truncation is also symmetric around m, or if the triangular distribution is symmetric and the rectangle, triangle or parabola filling the truncated portion is also symmetric around m, then the expected value of e is m. If m = 1 is used, the expected value of the masked data will be the same as that of the unmasked data. In other words, the sample mean of the masked data will be an unbiased estimate of the unmasked data.

The data disseminating agencies will provide the data users with the mean and variance of the noise being used for masking. The data users can use the information to estimate the sample variance from the masked data. Note that users cannot estimate the mean and variance of the noise from the masked data.

In further research, the domain estimation formula needs to be developed for this type of masking. These distributions need to be tried with real data and compared to find out which one performs best.

4. References

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.

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, 690-688.

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.

Kim, J..J. and Winkler, W. E. (2001) Multiplicative Noise for Masking Continuous Data, Research Report Series, Statistics #2003-01, Statistical Research Division, U.S. Bureau of the Census.

Massell, P. and Russell, N. (2006) Protecting Confidentiality of Commodity Flow Survey Tabular Data by Adding Noise to Underlying Microdata, presented at a Washington Statistical Society seminar, October 24, 2006.

van Dorp, J.R. and Kotz, S (2003) Generalized Trapezoidal Distributions, Metrika, Vol. 58, Issue 1, 85-97.

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.