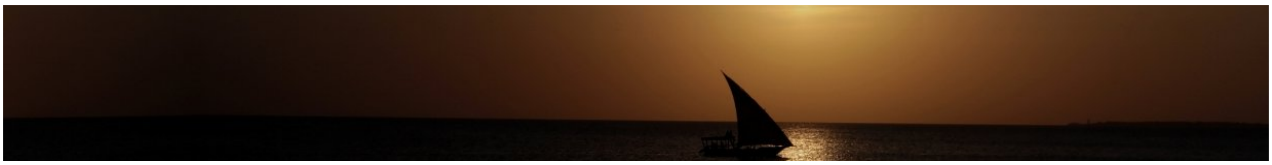

NavIncerta Library

Truncation of Lognormal Distributions

Domain:	Probabilistic analysis
Author:	Henk Krijnen
Reviewed by:	Thijs Koeling
Version:	1.0
Date:	November 5, 2017

NavIncerta
Address: Oude Delft 71F
2611BC Delft
The Netherlands
Email: info@navincerta.com
Phone: +31654385214

1 Why would we be interested in the truncation topic?

In this paper we will take a look at the mathematics of chopping up or truncating a lognormal distribution. There are two applications:

- Suppose a variable is lognormally distributed and there is some reason to subdivide the total range in several intervals or segments. What are the expectations of these segments? What weights should be applied in for example a decision tree to properly represent the various segments by the discrete expectation values?
- Suppose we would like to ignore the tail of a lognormal distribution as one takes the view that a real variable in question cannot assume very large values to infinity. What are the moments of a lognormal distribution with the tail cut off?

2 General double-sided truncation

The formula for the probability density function of the lognormal distribution is:

$$f(X) = \frac{1}{X\sigma\sqrt{2\pi}} e^{-\frac{(lnX-\mu)^2}{2\sigma^2}} \tag{1}$$

In this general case X ranges from 0 to infinity, whilst μ and σ denote the mean and standard deviation of the underlying normal distribution of $ln(X)$.

To derive the formulas for the moments (i.e. mean, variance and skewness) of truncated or discretized distributions, we will provide the derivation in the most general form. This is inspired by a derivation available in the discussion section of Wikipedia on the lognormal distribution, but presented here in a more generalized and complete version.

Suppose we would like to assess the characteristics of the distribution of a variable which is based on a lognormal X but bounded by A to the left and B to the right.

The n^{th} (raw) moment of this distribution is defined as:

$$M_n = \frac{\int_A^B X^n \frac{1}{X\sigma\sqrt{2\pi}} e^{-\frac{(lnX-\mu)^2}{2\sigma^2}} dX}{F_X(B) - F_X(A)} \tag{2}$$

F_X is the cumulative probability function of the lognormal variable X. The chance that X lies between A and B is thus $F_X(B) - F_X(A) = P(A, B)$. We need to divide by $P(A, B)$ in order to normalize the calculation of the moments. To clarify, note that when the moments are calculated for regular, untruncated distributions P equals 1. We now define a new variable y as follows:

$$y = \frac{lnX - \mu}{\sigma}$$
$$X = e^{\sigma y + \mu}$$
$$dX = \sigma e^{\sigma y + \mu} dy$$

Hence, we get:

$$P(A,B) \times M_n = \int_{\frac{lnA-\mu}{\sigma}}^{\frac{lnB-\mu}{\sigma}} \frac{e^{n(\sigma y+\mu)}}{e^{\sigma y+\mu}\sigma\sqrt{2\pi}} e^{-\frac{y^2}{2}} \sigma e^{\sigma y+\mu} dy$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\frac{lnA-\mu}{\sigma}}^{\frac{lnB-\mu}{\sigma}} e^{n(\sigma y+\mu)} e^{-\frac{y^2}{2}} dy$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\frac{lnA-\mu}{\sigma}}^{\frac{lnB-\mu}{\sigma}} e^{n\sigma y+n\mu-\frac{y^2}{2}} dy$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\frac{lnA-\mu}{\sigma}}^{\frac{lnB-\mu}{\sigma}} e^{-\frac{1}{2}(y-n\sigma)^2+\frac{1}{2}n^2\sigma^2+n\mu} dy$$

$$= \frac{e^{\frac{1}{2}n^2\sigma^2+n\mu}}{\sqrt{2\pi}} \int_{\frac{lnA-\mu}{\sigma}}^{\frac{lnB-\mu}{\sigma}} e^{-\frac{1}{2}(y-n\sigma)^2} dy$$

We now need to apply one more substitution: $v = y - n\sigma$ and $dv = dy$:

$$M_n = \frac{1}{P(A,B)} \times e^{\frac{1}{2}n^2\sigma^2+n\mu} \times \frac{1}{\sqrt{2\pi}} \int_{\frac{lnA-\mu}{\sigma}-n\sigma}^{\frac{lnB-\mu}{\sigma}-n\sigma} e^{-\frac{1}{2}v^2} dv$$

We recognize the integral as the cumulative distribution function of the normal distribution. The probability $P(A,B)$ can be obtained by integrating the lognormal density function from A to B, which is equivalent to integrating the normal density function from $\frac{lnA-\mu}{\sigma}$ to $\frac{lnB-\mu}{\sigma}$. If we work this out and use the appropriate symbol Φ[1]:

$$M_n = e^{\frac{1}{2}n^2\sigma^2+n\mu} \times \frac{\Phi(\frac{lnB-(\mu+n\sigma^2)}{\sigma}) - \Phi(\frac{lnA-(\mu+n\sigma^2)}{\sigma})}{\Phi(\frac{lnB-\mu}{\sigma}) - \Phi(\frac{lnA-\mu}{\sigma})} \quad (3)$$

Clearly this equation will lead to formulas that can easily be coded in a spreadsheet (by using the NORMDIST(x,0,1,1) function for Φ(x)).

3 Chopping up a lognormal

An application of the partial moment concept is discretization. This is about representing a lognormal distribution by a set of discrete values. To each of these values a weight is assigned. Suppose we would like to subdivide the distribution as shown in several parts for which we can define the bounding values A_1, A_2, A_3...

[1] Cumulative distribution function of the standard normal distribution $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt$

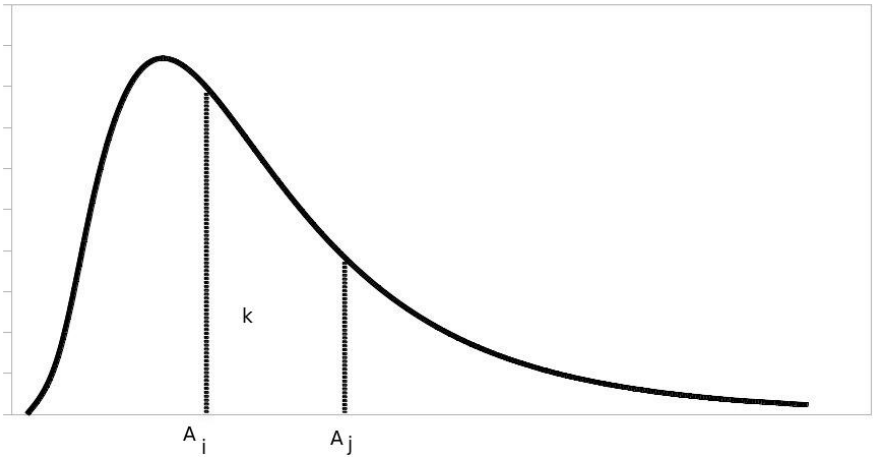

Figure 1: Partial Expectation

To find the expectation of each sector we take formula (3) and apply it only for the first moment to the sector bounded by A_i to the left and A_j (see Figure 1) to the right:

$$\tilde{X}_k = E[X|A_i < X < A_j] = e^{\frac{1}{2}\sigma^2+\mu} \times \frac{\Phi(\frac{lnA_j-(\mu+\sigma^2)}{\sigma}) - \Phi(\frac{lnA_i-(\mu+\sigma^2)}{\sigma})}{\Phi(\frac{lnA_j-\mu}{\sigma}) - \Phi(\frac{lnA_i-\mu}{\sigma})} \tag{4}$$

$$= \tilde{X} \times \left[\frac{\Phi(\frac{lnA_j-(\mu+\sigma^2)}{\sigma}) - \Phi(\frac{lnA_i-(\mu+\sigma^2)}{\sigma})}{\Phi(\frac{lnA_j-\mu}{\sigma}) - \Phi(\frac{lnA_i-\mu}{\sigma})} \right] \tag{5}$$

The weight is simply obtained by taking the denominator of above equation:

$$P_k = P(A_i, A_j) = \Phi\left(\frac{lnA_j - \mu}{\sigma}\right) - \Phi\left(\frac{lnA_i - \mu}{\sigma}\right) \tag{6}$$

Suppose we have chopped up the lognormal distribution in n parts, $k = 1, ..n$. The following should now apply:

$$\tilde{X} = \sum_{k=1}^{n} P_k \times \tilde{X_k} \tag{7}$$

with P_k calculated by (6) and $\tilde{X}_k$ by (5)

Example application

Suppose the price of some product has been modelled using a lognormal distribution. Suppose furthermore that if the price exceeds a certain threshold value, a different contractual condition will apply than if the price is lower than the threshold. You want to show the two cases as two branches in a decision tree. What weights and values should be used?

The product price has been modelled by the following assumptions:

- Low (P_{90}) = 23
- Mid (P_{50}) = 46
- High (P_{10}) = 92

The threshold is 50. So there are two cases: price < 50 and price > 50.

From the percentile values we can easily calculate the characteristics of the lognormal distribution (see also the paper 'The Lognormal Distribution' in the NavIncerta Library).

The mean equals: 53.24528

The parameters of the underlying normal distribution:

μ = 3.82864

σ = 0.54087

The mean for the case ' price < 50 ' can be calculated by (5) as follows:

Mean [case price < 50] = $53.24528 \times \frac{\Phi\left(\frac{ln(50)-(3.82864+0.54087^2)}{0.54087}\right)-0}{\Phi\left(\frac{ln(50)-3.82864}{0.54087}\right)-0} = 33.2$

Likewise, the mean for the case ' price > 50 ' can be calculated by (5) as follows:

Mean [case price > 50] = $53.24528 \times \frac{1-\Phi\left(\frac{ln(50)-(3.82864+0.54087^2)}{0.54087}\right)}{1-\Phi\left(\frac{ln(50)-3.82864}{0.54087}\right)} = 79.0$

The chance that price < 50 follows from: $\Phi\left(\frac{ln(50)-3.82864}{0.54087}\right)$ = 0.56

The chance that price > 50, then of course, equals 0.44

Thus we can develop a decision tree with one branch based on a case at a price of 33.2 with weight 0.56 and the second branch based at a price of 79.0 with weight 0.44.

Similar calculations follow in situations where the lognormal is subdivided in multiple segments, although it will be advised to build a spreadsheet rather than performing the calculations by hand as in the above example.

4 Right-sided truncation

One objective of the truncation concept is to cut off the tail of the lognormal distribution. After all, many physical quantities cannot assume an unlimited numeric value and tails towards infinity do not make sense. Normally such tails do not present a problem as the associated probabilities are extremely small. However, if the distribution has significant skewness then the tail can have undesirable effects.Therefore, where needed a truncation of the distribution can be applied to

eliminate these problems. For example, we could specify that the lognormal distribution is truncated at 1%-point of the cumulative descending distribution to avoid these problems. This does not for practical purposes change the probability model but ensures it is stable also at higher skews. Of course, if we are sure that the distributions to be modelled will not exhibit more than moderate skews there is no need to implement the calculations derived in this section.

First we must realize that the expressions (2) and (3) apply to the raw moments of the lognormal distribution. We generally work with the central moments. For the first moment (the expectation) there is no difference but for the variance and skewness we need some additional manipulations. Secondly, for this purpose we do not need to truncate the lower end (A_i in above articulations).

If $A \to 0$ then $\Phi(\frac{lnA-\mu}{\sigma}) \to 0$.

We therefore write $X|0 < X < B$ as the variable to focus on. We will abbreviate with $X|B$ or X_B.

For the first moment (expectation) (3) becomes:

$$E[X|B] = e^{\frac{1}{2}\sigma^2+\mu} \times \frac{\Phi(\frac{lnB-(\mu+\sigma^2)}{\sigma})}{\Phi(\frac{lnB-\mu}{\sigma})} \tag{8}$$

Or:

$$E[X|B] = E[X] \times \frac{\Phi(\frac{lnB-(\mu+\sigma^2)}{\sigma})}{\Phi(\frac{lnB-\mu}{\sigma})} \tag{9}$$

For the second moment, the variance, we need a little more involved approach. Remember, the formula (3) applies to raw moments, whilst the variance is equivalent to the second central moment. We can relate them using (3) as follows:

$$VAR[X|B] = E[X^2|B] - E[X|B]^2 \tag{10}$$

$$= e^{2\mu+2\sigma^2} \times \frac{\Phi(\frac{lnB-(\mu+2\sigma^2)}{\sigma})}{\Phi(\frac{lnB-\mu}{\sigma})} - \left[e^{\mu+\frac{1}{2}\sigma^2} \times \frac{\Phi(\frac{lnB-(\mu+\sigma^2)}{\sigma})}{\Phi(\frac{lnB-\mu}{\sigma})}\right]^2 \tag{11}$$

$$= e^{2\mu+2\sigma^2} \times \frac{\Phi(\frac{lnB-(\mu+2\sigma^2)}{\sigma})}{\Phi(\frac{lnB-\mu}{\sigma})} - \left[E[X] \times \frac{\Phi(\frac{lnB-(\mu+\sigma^2)}{\sigma})}{\Phi(\frac{lnB-\mu}{\sigma})}\right]^2 \tag{12}$$

Now remember that $e^{\mu} = median = X_{P_{50}} = P_{50}$ and $e^{\sigma^2} = \left(\frac{\tilde{X}}{P_{50}}\right)^2$, also with $E[X] = \tilde{X}$. With this, we can write:

$$VAR[X|B] = \frac{\tilde{X}^4}{{P_{50}}^2} \times \frac{\Phi(\frac{lnB-(\mu+2\sigma^2)}{\sigma})}{\Phi(\frac{lnB-\mu}{\sigma})} - \left[\tilde{X} \times \frac{\Phi(\frac{lnB-(\mu+\sigma^2)}{\sigma})}{\Phi(\frac{lnB-\mu}{\sigma})}\right]^2 \tag{13}$$

We now would like to apply a right-sided cutoff α. For example, if $\alpha = 1\%$ this means that we ignore the last 1% of the tail of the distribution. We then must choose a value for B accordingly. Then (9) and (12) become:

$$E[X|B] = \tilde{X} \times \frac{\Phi(\frac{lnB-(\mu+\sigma^2)}{\sigma})}{1-\alpha} \tag{14}$$

$$VAR[X|B] = \frac{\tilde{X}^4}{{P_{50}}^2} \times \frac{\Phi(\frac{lnB-(\mu+2\sigma^2)}{\sigma})}{1-\alpha} - \left[\tilde{X} \times \frac{\Phi(\frac{lnB-(\mu+\sigma^2)}{\sigma})}{1-\alpha}\right]^2 \tag{15}$$

The third central moment is given by:

$$TM[X|B] = E\left[[X|B] - E[X|B]\right]^3 \tag{16}$$

We work this out, with $\tilde{X_B} = E[X|B]$:

$$TM[X|B] = E[X^3|B] + 3{\tilde{X_B}}^2 E[X|B] - 3\tilde{X_B}E[X^2|B] - {\tilde{X_B}}^3 \tag{17}$$

$$= E[X^3|B] - 3\tilde{X_B}E[X^2|B] + 2{\tilde{X_B}}^3 \tag{18}$$

In this equation, the first term is actually the third raw moment, which we can calculate with (3) as follows:

$$E[X^3|B] = e^{\frac{1}{2}9\sigma^2+3\mu} \times \frac{\Phi(\frac{lnB-(\mu+3\sigma^2)}{\sigma})}{1-\alpha} \tag{19}$$

$$E[X^3|B] = \frac{\tilde{X}^9}{P_{50}^6} \times \frac{\Phi(\frac{lnB-(\mu+3\sigma^2)}{\sigma})}{1-\alpha} \tag{20}$$

Likewise, the second term will be:

$$-3\tilde{X_B}e^{2\sigma^2+2\mu} \times \frac{\Phi(\frac{lnB-(\mu+2\sigma^2)}{\sigma})}{1-\alpha} \tag{21}$$

$$-3\tilde{X_B}\frac{\tilde{X}^4}{{P_{50}}^2} \times \frac{\Phi(\frac{lnB-(\mu+2\sigma^2)}{\sigma})}{1-\alpha} \tag{22}$$

$\tilde{X_B}$ can be obtained from (14)

Summarizing, the first three moments of a truncated lognormal distribution can be calculated using the expressions as detailed below. Lognormal variable X is truncated at value B; $P(X > B) = \alpha$, μ and σ are the expectation and standard deviation of $ln(X)$, $X|B$ is the stochastic variable that results from truncating X at B, $\tilde{X} = E[X]$, $\tilde{X_B} = E[X|B]$.

The expectation:

$$\boxed{\tilde{X_B} = E[X|B] = \tilde{X} \times \frac{\Phi(\frac{lnB-(\mu+\sigma^2)}{\sigma})}{1-\alpha}}$$

The variance:

$$\boxed{VAR[X|B] = \frac{\tilde{X}^4}{{P_{50}}^2} \times \frac{\Phi(\frac{lnB-(\mu+2\sigma^2)}{\sigma})}{1-\alpha} - \left[\tilde{X} \times \frac{\Phi(\frac{lnB-(\mu+\sigma^2)}{\sigma})}{1-\alpha}\right]^2}$$

The third moment:

$$\boxed{TM[X|B] = \frac{\tilde{X}^9}{P_{50}^6} \times \frac{\Phi(\frac{lnB-(\mu+3\sigma^2)}{\sigma})}{1-\alpha} - 3\tilde{X_B}\frac{\tilde{X}^4}{{P_{50}}^2} \times \frac{\Phi(\frac{lnB-(\mu+2\sigma^2)}{\sigma})}{1-\alpha} + 2{\tilde{X_B}}^3}$$

5 Application: the moments of a truncated lognormal

We now aim to truncate the tail of a lognormal distribution and cut off the last 1%. Can we develop some approximations for the moments, as the formulas above are a little complicated?

As usual, we work from the percentiles, as low-mid-high descriptions of distributions are most commonly used in practice. Let's define, as in other NavIncerta papers, the asymmetry ratio as:

$$Asymmetry\ Ratio = AR = \frac{P_{10} - P_{50}}{P_{50} - P_{90}} \tag{23}$$

We now define a series of lognormal distributions with increasing skew (or increasing asymmetry ratio).

The P_{90} (low value) is set at 1.

The P_{50} (mid value or median) is set at 2.

The P_{10} (high value) varies from 3.1 with an increment of 0.5 to 11.6.

We now calculate the characteristics of the (shifted) lognormal distribution that fit the pre-defined percentiles. Then we apply a 1% cut off of the tail and calculate the moments of the truncated distribution using the formulas given above.

The results are shown in the following table:

Percentile definition lognormal				Moments full lognormal			Swanson	Truncated Distribtution		
P90	P50	P10	AR	Mean	SD	Skew	Mean	Mean	SD	Skew
1	2	3.1	1.1	2.0305	0.8215	0.22383	2.0300	2.00638	0.78906	0.05211
1	2	3.6	1.6	2.1855	1.0822	1.19292	2.1800	2.14223	0.99375	0.75418
1	2	4.1	2.1	2.3483	1.4244	2.14429	2.3300	2.27925	1.23893	1.25232
1	2	4.6	2.6	2.5207	1.8502	3.22794	2.4800	2.41889	1.51471	1.64702
1	2	5.1	3.1	2.7035	2.3679	4.54115	2.6300	2.56159	1.81708	1.97603
1	2	5.6	3.6	2.8971	2.9887	6.17673	2.7800	2.70749	2.14382	2.25846
1	2	6.1	4.1	3.1021	3.7255	8.23703	2.9300	2.85661	2.49350	2.50573
1	2	6.6	4.6	3.3186	4.5925	10.84072	3.0800	3.00893	2.86506	2.72540
1	2	7.1	5.1	3.5470	5.6054	14.12760	3.2300	3.16443	3.25767	2.92276
1	2	7.6	5.6	3.7875	6.7813	18.26287	3.3800	3.32305	3.67064	3.10171
1	2	8.1	6.1	4.0405	8.1384	23.44143	3.5300	3.48473	4.10341	3.26521
1	2	8.6	6.6	4.3062	9.6967	29.89243	3.6800	3.64944	4.55546	3.41556
1	2	9.1	7.1	4.5849	11.4777	37.88420	3.8300	3.81711	5.02635	3.55461
1	2	9.6	7.6	4.8768	13.5041	47.72954	3.9800	3.98769	5.51569	3.68382
1	2	10.1	8.1	5.1822	15.8006	59.79161	4.1300	4.16115	6.02311	3.80442
1	2	10.6	8.6	5.5013	18.3935	74.49027	4.2800	4.33742	6.54828	3.91740
1	2	11.1	9.1	5.8345	21.3109	92.30908	4.4300	4.51647	7.09091	4.02361
1	2	11.6	9.6	6.1819	24.5826	113.80288	4.5800	4.69826	7.65072	4.12377

Figure 2: Table of lognormal distributions and associated 1% truncated lognormal distributions

In the table, also the 'Swanson Mean' is shown. This mean is calculated with the following formula.

$$Swanson\ Mean = 0.3 \times P_{90} + 0.4 \times P_{50} + 0.3 \times P_{10} \tag{24}$$

This Swanson Mean is being applied widely to approximate the mean of a lognormal distribution. We observe that whereas it is inaccurate for a full lognormal distribution, especially at higher skews, it is a reasonable approximation for a 1% truncated lognormal.

Therefore we recommend to continue to use the Swanson Mean for practical reasons, provided that a lognormal distribution with a 1% cutoff applied is considered a reasonable representation of the variable in question.

However, if we were to use Swanson based equations for calculating the variance and skew, gross errors would develop at higher skews.

We can however arrive at approximation formulas for the second and the third moment as follows.

For the standard deviation we fit a second order polynomial to the column SD of the truncated lognormal as a function of the Asymmetry Ratio (AR). We then get for the standard deviation (SD_t) of the truncated distribution:

$$SD_t = (0.042AR^2 + 0.368AR + 0.3) \times (P50 - P90) \tag{25}$$

For calculating the skewness we fit a logarithmic function to the column Skew of the truncated lognormal as a function of the Asymmetry Ratio (AR). We then get for the skew (γ_t) of the truncated distribution:

$$\gamma_t = 1.885 \times ln(AR) - 0.143 \tag{26}$$

The rationale for the use of these formulas, which are valid up to AR = 10, is as follows. For quite a number of variables there is some logic to model these with a lognormal distribution. In practice it can happen that percentile values (low-mid-high) are chosen such that a high Asymmetry Ratio results. If subsequently a lognormal is fitted, the standard deviation and skewness become very large (also inspect the table above). To avoid this, one could choose to standardize on the above approximation formulas. They in effect 'protect' the analyst against using (too) highly skewed distributions.

The justification can be found in the consideration that, although variables may be considered to be generated by multiplicative processes (volumes, costs, revenues) and hence a lognormal would be an appropriate choice, in reality such variables will not assume values that are part of the tail to infinity. Hence, chopping off 1% (although an arbitrary number) is a reasonable thing to do.