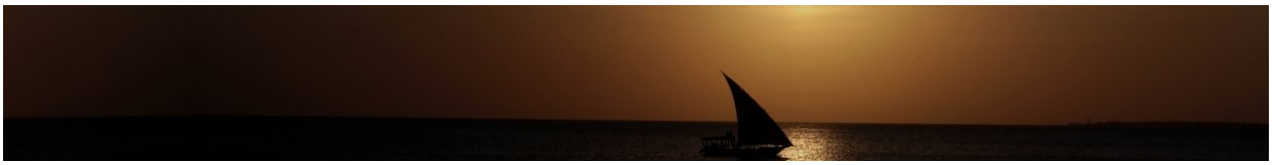

NavIncerta Library

A new method for discretizing continuous distributions

Domain:	Probabilistic analysis
Author:	Henk Krijnen
Reviewed by:	Thijs Koeling
Version:	1.0
Date:	August 25, 2017

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

1 Why is this relevant?

In probabilistic analysis uncertainties in variables are often described by continuous probability distributions. In decision trees such uncertainties may need to be represented by several discrete cases, spanning the uncertainty range. The question is often what weights should be assigned to the branches of the tree.

There are several rules being used. The most common is Swanson's rule. Assuming the three discrete values are chosen at P_{90} (low), P_{50} (mid) and P_{10} (high) percentiles, then the rule prescribes that weights of 0.3, 0.4 and 0.3 respectively are to be used. We will discuss later where this rule comes from.

The question we discuss is whether there is a more generic approach that allows establishment of the weights of a few discrete **arbitrary** values that are supposed to adequately characterize the continuous distribution.

The approach described in this paper allows calculation of such weights based on the following inputs: the mean and standard deviation of the distribution to be discretized, and three values considered to be representative of the distribution, i.e. roughly spanning the uncertainty range (a low, mid and high).

To ensure that these three values with their weighting properly represent the distribution, we require that 1) the mean of the discrete distribution (i.e. the weighted average of the three values chosen) is the same as the mean of the continuous distribution, but ALSO that the standard deviation of the discrete version of the distribution is the same as the original (meaning that the range represented by the three values is reasonable).

Any continuous distribution with mean $\tilde{X}$ and standard deviation Σ can be represented by three values L, M and H (low, mid, high) ensuring that mean and standard deviation are preserved, by using following weights:

$$W_M = \frac{\tilde{X}^2 + \Sigma^2 + LH - (L+H)\tilde{X}}{(M-H)(M-L)} \tag{1}$$

$$W_L = \frac{\tilde{X}^2 + \Sigma^2 + MH - (M+H)\tilde{X}}{(L-H)(L-M)} \tag{2}$$

$$W_H = \frac{\tilde{X}^2 + \Sigma^2 + LM - (L+M)\tilde{X}}{(H-M)(H-L)} \tag{3}$$

Whilst in this way we are preserving the mean and standard deviation (i.e. the first and second moments), these formulas do not cater for preserving the skew (third moment). If required, this can be accomplished by a simple goal seek procedure in which we keep two of the chosen values fixed, for example L and M, and subsequently goal seek H by applying a constraint through the formula for the skew for the discrete distribution.

A practical procedure for application of the concept presented would be:

- Examine the continuous distribution to be discretized and extract mean, standard deviation and skew;
- Choose two values, for example the low and the mid. Note that it is not needed to know at what percentile values these cases are positioned;
- Apply the goal seek procedure to find the third value and the three weights, using the above formulas as well as the input characteristics of the continuous distribution (mean, standard deviation and skew), with a constraint on the skew;

- If necessary, iterate if the weights turn out to be odd (< 0 or > 1), or if the third (calculated) value appears unrealistic.
- The cases can now be used in a decision tree with weights that exactly preserve the mean, width and skew of the distribution.

More details are provided in the text below.

If the skew is not very important, the goal seek part can be eliminated and the three weights can be found directly using the formulas above.

A **practical application** of this procedure could for example be for the following situation. Suppose you have done work on evaluating three cases that conformed to the P_{90}, P_{50} and P_{10} of a variable. They will feature in a decision tree. Suppose the thinking around this variable evolves and its probability distribution is updated. You now do not need to develop three new cases, but you can use the existing cases. You can apply the above procedure to find corresponding updated weights for the decision tree (unless the change of the distribution is very large).

A special case is a **symmetric distribution**, for which the following applies (with $\Delta = H - M = M - L$):

$$W_L = W_H = \frac{1}{2}\frac{\Sigma^2}{\Delta^2} \tag{4}$$

$$W_M = 1 - \frac{\Sigma^2}{\Delta^2} \tag{5}$$

If the symmetric distribution is normal, then **Swanson's rule** is valid: $W_L = W_H = 0.3, W_M = 0.4$.

Reference is made to (Bickel et al, 2011) for a more elaborate discussion of various discretization methods (which however does not contain the above generic approach).

2 Discretization by matching mean and variance

Suppose we have a continuous distribution representing the uncertainty of some variable. Let's assume we have available the expectation of the distribution and its standard deviation. We would like to replace this continuous distribution by three discrete values that roughly characterize the position and shape of the continuous distribution, for example to be applied in a decision tree. What weights should be used?

The approach to this problem that can be followed is to match the mean and standard deviation (i.e. match the first two moments). The discrete distribution replacing the continuous distribution with three values (low, mid and high) must have the same mean and the same standard deviation.

So we have chosen three values L, M, H (low, mid, high) with which the entire distribution is to be represented. We hence would like to know the weights W_L, W_M, W_H. The expectation of the

original distribution is $\tilde{X}$ and the standard deviation is Σ. If we want the discrete distribution to properly represent the original distribution, we will require:

$$W_L \times L + W_M \times M + W_H \times H = \tilde{X} \tag{6}$$
$$W_L \times (L - \tilde{X})^2 + W_M \times (M - \tilde{X})^2 + W_H \times (H - \tilde{X})^2 = \Sigma^2 \tag{7}$$
$$W_L + W_M + W_H = 1 \tag{8}$$

(6) expresses that the weighted average of the three values should match the expectation of the original distribution, whilst (7) implies that the variance of the discrete distribution matches too. Equation (8) obviously provides the requirement that the sum of the weights equals 1.

We thus have a set of three equations with three unknowns: W_L, W_M, W_H. We multiply (8) by H and subtract (6). We also multiply by L and again subtract (6).

$$W_L \times (H - L) + W_M \times (H - M) = H - \tilde{X} \tag{9}$$
$$W_L = \frac{H - \tilde{X} - W_M(H - M)}{H - L} \tag{10}$$
$$W_M \times (L - M) + W_H \times (L - H) = L - \tilde{X} \tag{11}$$
$$W_H = \frac{L - \tilde{X} - W_M(L - M)}{L - H} \tag{12}$$

We now insert ((10)) and ((12)) in ((7)). This results in:

$$\Sigma^2 = \frac{H - \tilde{X} - W_M(H - M)}{H - L} \times (L - \tilde{X})^2 + W_M \times (M - \tilde{X})^2 +$$
$$+ \frac{L - \tilde{X} - W_M(L - M)}{L - H} \times (H - \tilde{X})^2$$

$$\Sigma^2 = -W_M \left(\frac{H - M}{H - L}(L - \tilde{X})^2 - (M - \tilde{X})^2 + \frac{L - M}{L - H}(H - \tilde{X})^2 \right) +$$
$$+ \frac{H - \tilde{X}}{H - L}(L - \tilde{X})^2 + \frac{L - \tilde{X}}{L - H}(H - \tilde{X})^2$$

$$\Sigma^2 - \frac{H - \tilde{X}}{H - L}(L - \tilde{X})^2 - \frac{L - \tilde{X}}{L - H}(H - \tilde{X})^2 =$$
$$= -W_M \left(\frac{(H - M)(L - \tilde{X})^2 - (L - M)(H - \tilde{X})^2}{H - L} - (M - \tilde{X})^2 \right)$$

$$\Sigma^2 - \frac{(H-\tilde{X})(L-\tilde{X})^2 - (L-\tilde{X})(H-\tilde{X})^2}{H-L}$$
$$= -W_M \left(\frac{(H-M)(L-\tilde{X})^2 - (L-M)(H-\tilde{X})^2}{H-L} - (M-\tilde{X})^2 \right)$$

$$\Sigma^2 - \frac{(H-\tilde{X})(L^2-2L\tilde{X}+\tilde{X}^2) - (L-\tilde{X})(H^2-2H\tilde{X}+\tilde{X}^2)}{H-L}$$
$$= -W_M \left(\frac{(H-M)(L^2-2L\tilde{X}+\tilde{X}^2) - (L-M)(H^2-2H\tilde{X}+\tilde{X}^2)}{H-L} - (M-\tilde{X})^2 \right)$$

Evaluating the long terms in the last equation:

$$\begin{aligned}
&(H-\tilde{X})(L^2-2L\tilde{X}+\tilde{X}^2) - (L-\tilde{X})(H^2-2H\tilde{X}+\tilde{X}^2) = \\
&HL^2 - 2HL\tilde{X} + H\tilde{X}^2 - \tilde{X}L^2 + 2L\tilde{X}^2 - \tilde{X}^3 + \\
&- LH^2 + 2LH\tilde{X} - L\tilde{X}^2 + \tilde{X}H^2 - 2H\tilde{X}^2 + \tilde{X}^3 = \\
&= -HL(H-L) + \tilde{X}^2(H-L) + (H-L)\tilde{X}(H+L) - 2\tilde{X}^2(H-L) \\
&= -HL(H-L) - \tilde{X}^2(H-L) + (H-L)\tilde{X}(H+L)
\end{aligned}$$

Similarly:

$$\begin{aligned}
&(H-M)(L^2-2L\tilde{X}+\tilde{X}^2) - (L-M)(H^2-2H\tilde{X}+\tilde{X}^2) \\
&= HL^2 - 2HL\tilde{X} + H\tilde{X}^2 - ML^2 + 2ML\tilde{X} - M\tilde{X}^2 + \\
&- LH^2 + 2LH\tilde{X} - L\tilde{X}^2 + MH^2 - 2MH\tilde{X} + M\tilde{X}^2 = \\
&= -HL(H-L) + (H-L)\tilde{X}^2 + M(H+L)(H-L) - 2M\tilde{X}(H-L)
\end{aligned}$$

Now the equation becomes:

$$\Sigma^2 + HL - \tilde{X}(H+L) + \tilde{X}^2 = -W_M(-HL + \tilde{X}^2 + M(H+L) - 2M\tilde{X} - M^2 + 2M\tilde{X} - \tilde{X}^2)$$
$$W_M = \frac{\tilde{X}^2 + \Sigma^2 + HL - \tilde{X}H - \tilde{X}L}{M^2 + HL - MH - ML}$$

We can re-write for W_M:

$$W_M = \frac{\tilde{X}^2 + \Sigma^2 + LH - (L+H)\tilde{X}}{(M-H)(M-L)} \tag{13}$$

If we follow the same procedure for W_L and W_H, or following considerations of symmetry :

$$W_L = \frac{\tilde{X}^2 + \Sigma^2 + MH - (M+H)\tilde{X}}{(L-H)(L-M)} \tag{14}$$

$$W_H = \frac{\tilde{X}^2 + \Sigma^2 + LM - (L+M)\tilde{X}}{(H-M)(H-L)} \tag{15}$$

This works for any distribution (symmetric or skewed) and for any combination of mean, standard deviation and low, mid, high numbers. In situations of high skews or when the low, mid, high numbers are not representative of the distribution to be described, the weights will assume odd numbers, including negatives.

3 Dealing with the skewness: goal seek

Not included in the above concept is the preservation of the skewness of the original continuous distribution. Conceptually this can be done by requiring, in addition to (6), (7) and (8), that:

$$\gamma = \frac{W_L \times (L - \tilde{X})^3 + W_M \times (M - \tilde{X})^3 + W_H \times (H - \tilde{X})^3}{\Sigma^3} \tag{16}$$

In this expression γ is the skewness of the continuous distribution.

To accomplish this in a practical way, a goal seek procedure can be implemented as follows. In a spreadsheet:

- Create three cells for entering mean, standard deviation and skewness. Enter the relevant values.
- Create three cells for entering low, mid and high. Enter two of the three values.
- Code the equations (13),(14) and (15).
- Code equation (16).
- Create a goal seek with formula cell containing (16), the target value being the skewness of the continuous distribution and the variable cell being the unfilled value of the low, mid, high set.

4 Symmetric distributions

Suppose that the distribution to be discretized is symmetric. We choose L and H such that $H - M = M - L$. Then, of course, $M = \tilde{X}$ and $(L - \tilde{X})^2 = (H - \tilde{X})^2$. The weights for the high and low must be equal: $W_L = W_H$.

Let's look again at ((6)), ((7)) and ((8)). Let's denote $H - M = M - L = \Delta$. Equation ((7)) then becomes:

$$W_L \Delta^2 + W_H \Delta^2 = \Sigma^2 \tag{17}$$

$$W_L = W_H = \frac{1}{2} \frac{\Sigma^2}{\Delta^2} \tag{18}$$

$$W_M = 1 - \frac{\Sigma^2}{\Delta^2} \tag{19}$$

5 Swanson's rule

The so called Swanson rule is much used to discretize lognormal distributions. Simply put, if we have a low, mid and high of a lognormal distribution the mean of the distribution can be calculated as follows:

$$Mean \approx 0.3 \times Low + 0.4 \times Mid + 0.3 \times High \tag{20}$$

Bickel has shown that this rule is not valid for lognormal distributions, but for normal distributions. For lognormals the rule approximates the mean, but only at low to moderate skews.

If the continuous distribution is normal and the Low and High are at the 10%-percentiles, then: $\Delta = \Phi(10\%) \times \Sigma = 1.281552 \times \Sigma$. Hence we get:

$$W_L = W_H = \frac{1}{2} \frac{\Sigma^2}{(1.281552 \times \Sigma)^2} = 0.3044 \tag{21}$$

$$W_M = 1 - W_L - W_H = 0.3911 \tag{22}$$

These are the weights (rounded to 0.3 and 0.4) that are used in the Swanson rule. This perspective shows that indeed the Swanson rule applies to normal distributions and preserves the first and second moments.

This is considerably more straightforward than the derivation in (Hurst et al, 2000).

6 Application and examples

Examples, using (13),(14), and (15):

Mean	**St.Dev.**	**Low**	**Mid**	**High**	W_L	W_M	W_H
100	40	20	100	150	0.15	0.60	0.25
100	50	20	100	150	0.24	0.38	0.38
100	60	20	100	150	0.35	0.10	0.55
100	70	20	100	150	0.47	-0.23	0.75

As we can see, in this arbitrary example, the weight for the mid value turns negative when the standard deviation equals 70. Of course, this is not intuitive. Although universally applicable, the concept is not universally usable as it is impossible to explain why certain values would get a negative weight, although this would be theoretically correct. However, it is just a matter of choosing a different set of values representing the distribution, for example in the last case (replacing a high of 150 by 200):

Mean	**St.Dev.**	**Low**	**Mid**	**High**	W_L	W_M	W_H
100	70	20	100	200	0.34	0.39	0.27

If the distribution is skewed, and we also know the skewness, then we can add the procedure as highlighted in section 3. Let's say, using the last case, we pick a low and a mid and let the high be found using the goal seek. Examples for two different values of the skewness:

Mean	**St.Dev.**	**Skew**	**Low**	**Mid**	W_L	W_M	W_H	**High (calc.)**
100	70	0.5	20	100	0.31	0.47	0.22	215
100	70	1.7	20	100	0.22	0.69	0.09	299

In case the mean, standard deviation and skewness of the continuous distribution are not available but instead percentile values have been estimated, for example P_{90} (low), P_{50} (mid) and P_{10} (high) then first, using these values, the characteristics of the continuous distribution must be derived. This will depend on the type of distribution. For the lognormal distribution such formulas can be found in the article 'The lognormal distribution' (also in the NavIncerta Library).

Let's work out an example. We have a lognormal distribution with estimated $P_{90} = 15$ (low), $P_{50} = 30$ (mid) and $P_{10} = 60$ (high). The characteristics can be calculated to be:

Mean: 34.7252

Standard deviation: 20.2427

Skewness: 1.9469

Mean	**St.Dev.**	**Skew**	**Low**	**Mid**	W_L	W_M	W_H	**High (calc.)**
34.7252	20.2427	1.9	15	30	0.15	0.72	0.12	86.75

Hence, we would need to replace the original high value of 60 by 86.75 (and use the weights as calculated) for the discrete distribution having the same mean, standard deviation and skewness as the original continuous lognormal distribution.

If we are not concerned about preserving the skewness, and wish to stick to the original set of three discrete values, then we can calculate the weights to be applied:

Mean	**St.Dev.**	**Low**	**Mid**	**High**	W_L	W_M	W_H
34.7	20.2	15	30	60	0.43	0.20	0.37

Another example: the triangular distribution.

If we have a triangular distribution with min = 10, mode = 20 and max = 35, then we calculate the distribution characteristics to be:

Mean: 21.67

Standard deviation: 5.137

Skewness: 0.19125

If we were to use the same set of values (min, mode, max) to represent the triangular distribution, then following would be the weights to be applied:

Mean	**St.Dev.**	**Skew**	**Low**	**Mid**	W_L	W_M	W_H	**High (calc.)**
21.67	5.137	0.19125	10	20	0.07	0.70	0.23	29.91

Again, if we are not concerned about preserving the skewness and wish to stick to the original set of three discrete values, then we calculate the following weights:

Mean	**St.Dev.**	**Low**	**Mid**	**High**	W_L	W_M	W_H
21.67	5.137	10	20	35	0.02	0.86	0.12

7 Conclusion

Using the approach proposed in this paper the weights for any set of three values representing a discrete equivalent of a continuous distribution can easily be calculated enforcing that the discrete distribution has the same mean and standard deviation as the original continuous distribution.

By adding a simple goal seek procedure also the skewness can be preserved, although this requires fixing two of the three discrete values and finding the third value.

The procedure applies to any type of continuous distribution and the discrete values chosen to represent it may or may not be positioned at known percentile values.

It is assumed that in general there is no need to preserve the moments higher than the third (for example the kurtosis, related to the 4th moment). Should this be required, however, reference is made to (Bickel et al, 2011) containing solutions for several distributions on the basis of the Gaussian quadrature approach. The calculation procedures then become more restrictive and complicated.

8 References

Discretization, Simulation, and Swanson's (Inaccurate) Mean, J. Eric Bickel, SPE, and Larry W. Lake, SPE, The University of Texas at Austin, John Lehman, Strategic Decisions Group, SPE Economics and Management, 2011.

Swanson's 30-40-30 rule, A. Hurst, G. C. Brown, and R. I. Swanson, AAPG Bulletin, v. 84, no. 12 (December 2000), pp. 1883–1891.