The BEDA (Belief-Enhanced Decision Analysis) Method

A Decision-Analytic Method by Prof. Mats Danielson, Dept. of Computer and Systems Sciences, Stockholm University, Forum 100, SE-164 40 Kista, Sweden, mad@dsv.su.se

Summary

The DELTA Method is a distribution-free decision analysis method for the handling and evaluation of decision and risk trees (Danielson, 1997). It has thereafter in 2001-2002 been extended from probabilistic decision situations also to cover decisions under multiple criteria. Decision alternatives are evaluated by so-called contractions of the intervals combined with several complementary evaluation rules. The advantage of a distribution-free approach is the generality and freedom from assumptions that it allows. However, a disadvantage is the unintuitive interpretation of the results of a contraction. In order to alleviate that problem, an additional analysis method is introduced in this report, based on a belief mass interpretation of the output intervals from DELTA. Each input and output interval consists of a lower bound, an upper bound, and a focal point. These three points are interpreted as parameters for belief distributions (Dirichlet distributions for probabilities and criteria weights, triangle distributions for values).

Decision Analysis Background

This section is built on (Danielson, 1997), which describes the DELTA Method for interval decision analysis that was later generalised to multi-level trees (the original text handles only single-level trees, but the generalisation is straightforward and does not introduce any new concepts). Decisions under risk (probabilistic decisions) are often given a tree representation. This is the reading of the tree as a sequence of events leading up to the final consequences, the end nodes. As an example, consider the tree in Figure 1, a screenshot from the software *Decide*IT, built on the representation and evaluation algorithms in (Danielson, 1997) by the company Preference AB (www.preference.nu).

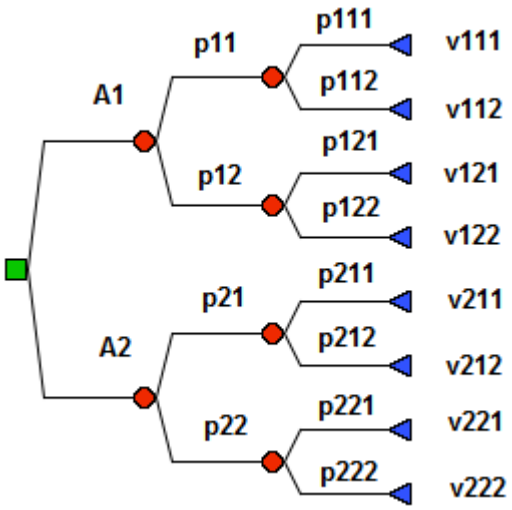

Figure 1 *A two-level decision tree from DecideIT*

A decision tree consists of a root node, representing a decision, a set of intermediary (event) nodes, representing some kind of uncertainty about which event will eventually occur, and consequence nodes, representing possible final outcomes. Usually, probability distributions are assigned in the form of weights in the probability nodes as measures of the uncertainties involved. The informal semantics are simply that given that an alternative A_i is chosen, there is a probability p_{ij} that an event occurs. This event can either be a consequence with a value v_{ijk} assigned to it or another event. Usually, the maximisation of the expected value is used as an evaluation rule. For instance, in Figure 1 above, the expected value of alternative A_i is

$$E(A_i) = \sum_{j=1}^{2} p_{ij} \sum_{k=1}^{2} p_{ijk} v_{ijk}$$

In case of precise probability and utility assessments, this is straightforwardly evaluated. However, when the probabilities and utilities are imprecise, several complications appear, including the non-uniqueness of the expected value of an alternative (leading to the need to find upper and lower bounds). The first step in obtaining a solution is generalising the decision tree structure.

Representation

Let a **decision frame** represent a tree decision problem. This is convenient for presentational purposes. The idea with such a frame is to collect all information necessary for the model in one structure. One of the building blocks of a decision frame is a graph.

A **graph** is a structure $\langle I,N,E\rangle$, where I is an index set, N is a set $\{n_i\}$, $i\in I$, of nodes, and E is a set $\{(n_i,n_j)\}$, $i,j\in I$, $i\neq j$, of edges (node pairs). A **tree** is a connected graph without cycles.

An **r-tree** (rooted tree) is a tree $\langle I,N,E,r\rangle$ where exactly one node n_r has the property $\neg\exists\ k : (n_k,n_r)\in E$. n_r is called the root of the tree. The set N is partitioned into two subsets of **leaf nodes** (N^L) and **intermediate nodes** (N^I). $n_i \in N^I$ *iff* $\exists\ k : (n_i,n_k)\in E$. Since $N^L = N \setminus N^I$, $n_i\in N^L$ **iff** $\neg\exists\ k : (n_i,n_k)\in E$. The index set I is partitioned accordingly: an index $i\in I^I$ *iff* $n_i\in N^I$ and an index $i\in I^L$ **iff** $n_i\in N^L$. An intermediate node $n_i\in N^I$ has children indices $C_i = \{j : (n_i,n_j)\in E\}$.

Then, all the rooted trees representing alternatives are joined together into a decision frame. In the sequel, the notation is used that the *n* children of a node x_i are denoted, $x_{i1}, x_{i2},\ldots,x_{in}$ and the m children of the node x_{ij} are denoted $x_{ij1}, x_{ij2},\ldots,x_{ijm}$, etc.

Decision-maker **statements** of probability and value are translated into **constraints** (inequalities) in order to be entered into the decision problem. Range statements (i.e. intervals) translate into **range constraints**, inequalities involving a single variable. A reasonable interpretation of such statements is that the estimate is not outside of the given interval. For a value scale [a,b], there is a default range constraint $v_{ij}\in[a,b]$ for each value variable. Likewise, there is a default range constraint $p_{ij}\in[0,1]$ for each probability variable (although, in practice, the normalisation

takes care of this). Comparative statements compare the probabilities of two consequences occurring with one another, such as "*the events C_1 and C_2 are equally probable*" or "*the event C_3 is more likely to occur than C_4*". Those statements are translated into **comparative constraints**, inequalities involving more than one variable. The term **interval constraints** is used for the kinds of constraints above. A collection of interval constraints concerning the same set of variables is called a constraint set, and it forms the basis for the representation of decision situation statements.

Given an index set I and a set of variables $\{x_i\}_{i\in I}$, a **constraint set** in $\{x_i\}_{i\in I}$ is a set of interval constraints in $\{x_i\}_{i\in I}$.

Initially, it is important to determine whether the elements in a constraint set are at all compatible with each other. This is the question of whether a constraint set has a solution, i.e. if there exists any vector of real numbers that can be assigned to the variables.

Given an index set I and a set of variables $\{x_i\}_{i\in I}$, a constraint set X in $\{x_i\}_{i\in I}$ is **consistent** *iff* the system of weak inequalities in X has a solution. Otherwise, the constraint set is **inconsistent**. A constraint Z is **consistent with** a constraint set X **iff** the constraint set $\{Z\} \cup X$ is consistent. The collection of all consistent instances of a constraint set X is called the solution set to X.

Given an index set I and a consistent constraint set X in $\{x_i\}_{i\in I}$ and a function *f*, the **maximum** is ${}^{X}\max(f(x)) =_{def} \sup(a \mid \{f(x) > a\} \cup X \text{ is consistent})$. In a similar way, the **minimum** is ${}^{X}\min(f(x)) =_{def} \inf(a \mid \{f(x) < a\} \cup X \text{ is consistent})$.

Given an index set I, a consistent constraint set X in $\{x_i\}_{i\in I}$ and a function *f*, ${}^{X}\mathbf{argmax}(f(x))$ is a solution vector that is a solution to ${}^{X}\max(f(x))$, and ${}^{X}\mathbf{argmin}(f(x))$ is a solution vector that is a solution to ${}^{X}\min(f(x))$.

Note that argmax and argmin need not be unique. The feasible box (i.e., the set of feasible variable assignments) can be calculated if the constraint set is consistent. The feasible box is a concept that in each dimension signals which parts are infeasible within the constraint set. Intuitively, the feasible box represents a conservative extension of the solution set of a set of constraints.

Given an index set I and a consistent constraint set X in $\{x_i\}_{i\in I}$, the set of optimum pairs $\{\langle {}^{X}\min(x_i), {}^{X}\max(x_i)\rangle\}_{i\in I}$ is the **feasible box** of the set and is denoted $\langle {}^{X}\min(x_i), {}^{X}\max(x_i)\rangle_I$.

This feasible box represents upper and lower probabilities if X consists of probabilities and upper and lower values if X consists of values. For convexity reasons, the entire interval between those extremal points is feasible. Using this concept, an application program can display to the user which statements are incompatible or which parts of intervals are incompatible with the rest of the statement set. Hence, at all times, an application program can maintain a consistent model of the user's problem in collaboration with the user.

Probability and Value Constraint Sets

There are two types of constraint sets (*c-sets*), probability c-sets and value c-sets. The smallest c-set unit is the event node c-set, which collects all probability statements made regarding a specific event node in an r-tree.

Given an r-tree $T = \langle I,N,E,r\rangle$ and an event node n_i, consider the set C_i of disjoint and exhaustive consequences of the event (children nodes), user event statements in $\{p_j\}_{j\in Ci}$, and a discrete, finite probability mass function $\Pi:n_j\rightarrow[0,1]$ over C_i. Let p_j denote the function value $\Pi(n_j)$. Π obeys the standard probability axioms, and thus $p_j\in[0,1]$ and $\Sigma_j\, p_j = 1$ are default constraints. Then the **event node c-set** P_i is derived from the set of user range and comparative statements with the following content.

- A feasible box $\langle a_k,b_k\rangle$, $k\in C_i$, which represents the user and default range constraints $\forall k \in C_i : p_k \in [0,1]$.
- All user comparative constraints.
- The normalisation constraint $\sum_{k\in C_i} p_k = 1$.

Thus, the c-set transforms statements into linear constraints while maintaining the same meaning. A c-set is more convenient to handle than a pure set of statements. An event node c-set is characterising a set of discrete probability distributions. The next aggregation level is that of a probability c-set, which collects together all probability statements belonging to all nodes in the same tree.

Given an r-tree $T = \langle I,N,E,r\rangle$ with all event nodes n_i, $i\in I^I$. Then the **probability c-set** P is all event c-sets P_j combined, i.e. feasible boxes, normalisations, and user comparative statements.

Requirements similar to those for probability variables are found for value variables. There are apparent similarities and differences between probability and value statements. The normalisation ($\Sigma_k\, p_{ik} = 1$) requires the probability variables of an intermediate node to sum to one. No such constraint exists for the value variables. Further, the value scale endpoints can be arbitrarily selected and need not be [0,1] as in the probability case.

Given an r-tree $T = \langle I,N,E,r\rangle$, consider the set N^L of leaf nodes. Then a **value c-set** is derived from the set of user range and comparative statements. The user statements, together with the default statements $\forall k \in I^L : v_k \in [0,1]$, form the c-set constraints in the following way.

- A hull $\langle a_k,b_k\rangle$, $k\in I^L$, which represents the user and default range constraints.
- All user comparative constraints.

Similar to probability c-sets, a value c-set is characterising a set of value functions. The statements are transformed into a set of linear constraints. Using the above concepts of constraint and c-set, a decision situation is modelled by a decision frame. To begin with, each alternative is represented by a tree frame.

Given a decision alternative, statements are made about the probabilities of the events as well as the values of the consequences. A **tree frame** is a structure $\langle T,P,V\rangle$ containing the following representation of the alternative:

- A rooted tree $T = \langle I,N,E,r\rangle$ with index set partitions I^I and I^L, and, for each $i \in I^I$, the child index set C_i.
- A probability c-set P in variables $\{p_i\}$, $i \in I\backslash\{r\}$, representing all probability statements in the form of a feasible box and constraints.
- A value c-set V in variables $\{v_i\}$, $i \in I^L$, representing all value statements in the form of a feasible box and constraints.

All alternatives are modelled in the same structure. This structure (the decision frame) fully represents the entire decision problem, and all evaluations are made relative to it. The probability and value c-sets, together with structural information, constitute the decision frame.

Given a probabilistic decision situation with m alternatives, a **decision frame** is a structure $\langle m,F\rangle$, $F = \{F_i\}$ for $i \in \{1,...,m\}$, where $F_i = \langle T_i,P_i,V_i\rangle$ is a tree frame for alternative A_i. Thus, the decision frame contains , for each alternative, a decision tree structure and a tree frame.

Evaluation Algorithms

Now that the representation structure is defined, the next item is algorithms for computing upper and lower bounds for the expected value in the tree, i.e. optimisation of sums of products derived from the tree structure. The primary evaluation rule is based on the expected value. Since neither probabilities nor values are fixed numbers, evaluating the expected value yields multilinear objective functions (with bilinear functions as a special case for one-level trees). Evaluate the expected value of an alternative given a decision frame $\langle m, \{\langle T_i,P_i,V_i\rangle\}\rangle$, i.e.

$$EV(A_i) = \sum_{i_1=1}^{n_{i_0}} p_{ii_1} \sum_{i_2=1}^{n_{i_1}} p_{ii_1i_2} \cdots \sum_{i_{m-1}=1}^{n_{i_{m-2}}} p_{ii_1i_2\cdots i_{m-2}i_{m-1}} \sum_{i_m=1}^{n_{i_{m-1}}} p_{ii_1i_2\cdots i_{m-2}i_{m-1}i_m} v_{ii_1i_2\cdots i_{m-2}i_{m-1}i_m},$$

where $p_{\ldots i_j \ldots}$, $j \in \{1,\ldots,m\}$ denote probabilities in P_i and $v_{\ldots i_j \ldots 1}$ denote values in V_i. Optimisation of such non-linear expressions subject to linear constraints (the probability and value constraint sets) are described in (Danielson, 1997).

The contraction is a generalised sensitivity analysis to be carried out in an arbitrary number of dimensions. In non-trivial decision situations, when an information frame contains numerically imprecise information, the different principles suggested above are often too weak to yield a conclusive result. Often, a far too crowded set of candidates is received. One way to proceed could be to determine the stability of the relation between the consequence sets under consideration. A natural way to investigate this is to consider values near the boundaries of the intervals as being less reliable than more central values due to interval statements being deliberately imprecise. This is takne into account by measuring the dominated regions indirectly using the concept of contraction.

The principle of contraction is motivated by the difficulties of performing simultaneous sensitivity analysis in several dimensions at the same time. It can be hard to gain a real understanding of the solutions to large decision problems using only one-dimensional analyses since different combinations of dimensions can be critical to the evaluation results. Investigating all possible such combinations would lead to a procedure of high complexity in the number of cases to investigate. Using contractions, this difficulty is circumvented. The contraction avoids the complexity inherent in combinatorial analyses. However, it is still possible to study the stability of a result by gaining a better understanding of how important the interval boundary points are. By co-varying the contractions of an arbitrary set of intervals, it is possible to gain much better insight into the influence of the structure of the information frame on the solutions. Both the set of intervals under investigation and the scale of individual contractions can be controlled. Consequently, a contraction can be regarded as a focus parameter that zooms in on central subintervals of the full statement intervals.

X is a base with the variables $x_1,\ldots,x_n$, $\pi \in [0,1]$ is a real number, and $\{\pi_i \in [0,1] : i = 1,\ldots,n\}$ is a set of real numbers. $[a_i, b_i]$ is the interval corresponding to the variable x_i in the solution set of the base, and $\bar{k} = (k_1,\ldots,k_n)$ is a consistent point in X. A ***π*-contraction** of X is to add the interval statements $\{x_i \in [a_i+\pi\cdot\pi_i\cdot(k_i-a_i), b_i-\pi\cdot\pi_i\cdot(b_i-k_i)] : i = 1,\ldots,n\}$ to the base X. $\bar{k}$ is called the **contraction point** (or focal point).

By varying π from 0 to 1, the intervals are decreased proportionally using the gain factors in the π_i-set, thereby facilitating the study of co-variation among the variables. This is a form of sensitivity analysis, which is described in more detail in (Danielson, 1997). It is implemented in the tool *Decide*IT, from which Figure 2 is taken. It shows the comparison of expected values between two alternatives. In the figure, the contraction progresses from left (0%) to right (100%), where the original intervals are reduced to single numbers at the contraction point $\bar{k}$.

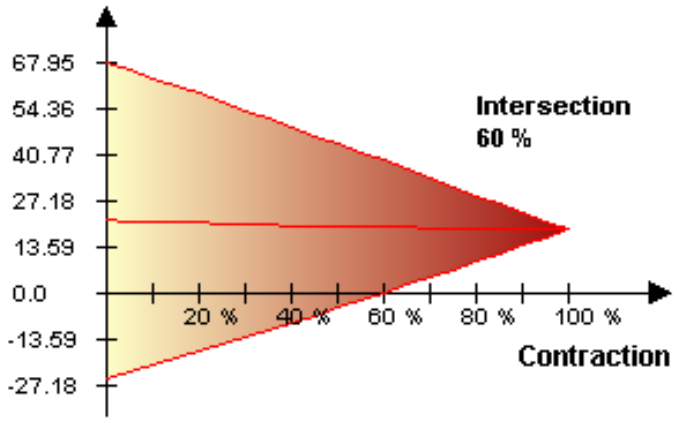

Figure 2 *Contraction graph in DecideIT*

At 60% contraction, the intervals have shrunk to the extent that there is no longer any consistent assignment of variables that make the lower alternative a feasible choice. This represents the

state-of-the-art in interval decision analysis at the time of writing (2012). However, the company that commissioned this research task had a desire for a sensitivity analysis method that keeps the complete intervals in the analysis process. This research question led to the BEDA method.

The BEDA Method

A key observation in the DELTA method is that the belief in points closer to the endpoints of the intervals is lower than the belief in more central points. This is the reason for the contraction procedure above. The same observation underlies the BEDA method, but it is effectuated differently – by assigning explicit distributions of belief on the intervals. The distributions used for expressing beliefs are well-known distributions from statistics: the Dirichlet distribution for probabilities (since they need to sum to one following Kolmogorov's axiom system) and the triangle and uniform distributions for utilities/values, the choice depending on whether there are two or three points defining an interval. The properties of both Dirichlet and triangle distributions are well described in (Kotz & van Dorp, 2004). In this document, the evaluation of such distributions will be discussed.

Begin by revisiting the expression for the expected value:

$$\mathrm{EV}(A_i) = \sum_{i_1=1}^{n_{i_0}} p_{ii_1} \sum_{i_2=1}^{n_{i_1}} p_{ii_1i_2} \cdots \sum_{i_{m-1}=1}^{n_{i_{m-2}}} p_{ii_1i_2\cdots i_{m-2}i_{m-1}} \sum_{i_m=1}^{n_{i_{m-1}}} p_{ii_1i_2\cdots i_{m-2}i_{m-1}i_m} v_{ii_1i_2\cdots i_{m-2}i_{m-1}i_m 1},$$

To evaluate this expression, and thus arrive at an analysis of the decision situation, employ calculation methods for the two operators addition and multiplication. The addition operator is handled by ordinary convolution, i.e. if h is the distribution over a sum $z = x + y$ whose components have distributions $f(x)$ and $g(y)$, then $h(z)$ is

$$h(z) = \frac{d}{dz}\int_0^z f(x)g(z-x)dx.$$

The multiplication operator is treated analogously. Using the same assumptions as above, if h is the distribution over a product $z = x \cdot y$, $h(z)$ is found by letting

$$H(z) = \iint_{\Gamma_x} f(x)g(y)dxdy = \int_0^1 \int_0^{z/x} f(x)g(y)dxdy = \int_z^1 f(x)G(z/x)dx$$

where G is a primitive function to g, $\Gamma_z = \{(x,y) \mid x \cdot y \leq z\}$, and $0 \leq z \leq 1$. Then $h(z)$ is the corresponding density function

$$h(z) = \frac{d}{dz}\int_z^1 f(x)G(z/x)dx = \int_z^1 \frac{f(x)g(z/x)}{x}dx.$$

In theory, the products are calculated and the abovementioned convolution of two densities then effectuates the summations of the products. This combination of operators computes the distribution over the expected utility. In practice, however, these calculations are very complicated for a decision-analytic tool to carry out, especially when additional requirements are added,

such as asymmetry in the input distributions and truncated distributions due to the input intervals being narrower than the default [0, 1] range assumed in the standard theory.

The evaluation method in BEDA is based on the principle of going concern (PGC). It is the same PGC observation that enables the use of probability theory as a risk calculus. The probability of an event occurring is the proportion of times it occurs if the event is repeated an infinite number of times. In using probabilities for modelling real-life events, the approximation is used that the probability best represents the risk involved. For this approximation to be reasonable, several events need to take place for the real-world outcomes to cancel out in the sense that they, on average, tend to the probability. This is the assumption of going concern, and the approximation is viable in most decision situations, which is why probability calculus is accepted for use in this way. The same PGC reasoning applied to distributions involves the central limit theorem and law of large numbers in statistics. This leads to the well-founded approximation that the total distribution of expected value over a large number of decision situations will tend to the normal distribution. Using this approximation, the evaluation in the BEDA method amounts to finding parameters for a suitable approximately normal distribution. Two factors slightly complicate matters. **i**) The input distributions are seldom symmetric in the sense that their mean values are not midway between the lower and upper boundaries of the intervals. And even if they were, the multiplication operator's non-linearity still yields an asymmetric result. **ii**) The lower and upper bounds themselves introduce truncations into the resulting distributions, leading to non-standard outcomes.

This eventually turns the BEDA evaluation into a moment calculus using the NEMO (net moment) technique. NEMO includes all moments that have noticeable impact on the end result and excludes those that have negligible impact to save computation time. This entails that no moments higher than three are considered for the calculations.

Skew-normal Distribution

The skew-normal distribution is a continuous probability distribution that generalises the normal distribution to allow for non-zero skewness.

Let $\phi(x)$ denote the standard normal probability density function

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

with the cumulative distribution function given by

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\, dt = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$$

where erf(·) is the error function. Then the probability density function of the skew-normal distribution is given by

$$f(x) = 2\phi(x)\Phi(\alpha x)$$

where α is the shape parameter, φ is the standard normal density, and Φ its cumulative distribution function. To add location and scale parameters, transform

$$x \to \frac{x-\xi}{\omega}\ .$$

When $\alpha = 0$ the result is the standard normal distribution. When $\alpha = 1$ it models the distribution of the maximum of two independent standard normal variates. When the shape parameter's absolute value increases, the distribution's skewness increases. The limit as $|\alpha| \to \infty$ results in the folded normal distribution or half-normal distribution. The distribution is right-skewed if $\alpha > 0$ and left-skewed if $\alpha < 0$. When α changes its sign, the density is reflected about $x = 0$. The skew-normal density function with location ξ, scale ω, and shape parameter α is

$$f(x) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\left(\frac{x-\xi}{\omega}\right)\right).$$

The probability density function is

$$\frac{1}{\omega\pi}e^{-\frac{(x-\xi)^2}{2\omega^2}}\int_{-\infty}^{\alpha\left(\frac{x-\xi}{\omega}\right)}e^{-\frac{t^2}{2}}\,dt$$

The cumulative distribution function is

$$\Phi\left(\frac{x-\xi}{\omega}\right) - 2T\left(\frac{x-\xi}{\omega}, \alpha\right)$$

where $T(h, a)$ is Owen's T-function. $T(h, a)$ is defined by

$$T(h,a) = \frac{1}{2\pi}\int_0^a \frac{e^{-\frac{1}{2}h^2(1+x^2)}}{1+x^2}dx \quad (-\infty < h, a < +\infty)\,.$$

$T(h, a)$ gives the probability of the event ($X > h$ and $0 < Y < a \cdot X$) where X and Y are independent standard normal random variables.

Mean of the skew-normal distribution:

$$\xi + \omega\delta\sqrt{\frac{2}{\pi}}$$

where

$$\delta = \frac{\alpha}{\sqrt{1+\alpha^2}}$$

Variance of the skew-normal distribution:

$$\omega^2\left(1 - \frac{2\delta^2}{\pi}\right)$$

Truncated Distributions

The truncated (skew-)normal distribution is the probability distribution of a (skew-)normally distributed random variable whose value is either bounded from below, from above, or both. Without loss of generality, suppose that

$$X \sim N(\mu, \sigma^2)$$

has a normal distribution and lies within the interval

$$X \in (a, b),\ -\infty \le a < b \le \infty.$$

Then X conditional on $a < X < b$ has a truncated normal distribution. Its probability density function, f, for $a \le x \le b$, is given by

$$f(x; \mu, \sigma, a, b) = \frac{\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

and by $f = 0$ otherwise.

Here, $\phi(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\xi^2\right)$ is the probability density function of the standard normal distribution and $\Phi(\cdot)$ is its cumulative distribution function.

Two sided truncation:

$$\mathrm{E}(X \mid a < X < b) = \mu + \frac{\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}\sigma$$

$$\mathrm{Var}(X \mid a < X < b) = \sigma^2 \left[1 + \frac{\frac{a-\mu}{\sigma}\phi\left(\frac{a-\mu}{\sigma}\right) - \frac{b-\mu}{\sigma}\phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} - \left(\frac{\phi\left(\frac{a-\mu}{\sigma}\right) - \phi\left(\frac{b-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}\right)^2\right]$$

One sided truncation (upper tail):

$$\mathrm{E}(X \mid X > a) = \mu + \sigma\lambda(\alpha)$$

$$\mathrm{Var}(X \mid X > a) = \sigma^2[1 - \delta(\alpha)],$$

where

$$\alpha = (a - \mu)/\sigma,\ \lambda(\alpha) = \phi(\alpha)/[1 - \Phi(\alpha)]$$

and

$$\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha].$$

One sided truncation (lower tail):

$$\mathrm{E}(X \mid X < b) = \mu - \sigma\frac{\phi(\beta)}{\Phi(\beta)}$$

$$\operatorname{Var}(X \mid X < b) = \sigma^2 \left[1 - \beta \frac{\phi(\beta)}{\Phi(\beta)} - \left(\frac{\phi(\beta)}{\Phi(\beta)}\right)^2\right],$$

where

$$\beta = (b - \mu)/\sigma.$$

Moments

In BEDA, the NEMO (net moment) calculus determines the output distributions. The n^{th} moment of a real-valued continuous function $f(x)$ of a real variable about a value c is

$$\mu'_n = \int_{-\infty}^{\infty} (x - c)^n \, f(x) \, dx.$$

The function $f(x)$ is a probability density function. The n^{th} moment about zero of a probability density function $f(x)$ is the expected value of X^n and is called the *raw moment* ($c = 0$). The moments about its mean are called *central moments* ($c = \mu$ where μ is the mean value). These describe the shape of the function independently of translation.

If f is a probability density function, then the value of the integral above is called the n^{th} moment of the probability distribution. If F is a cumulative probability distribution function of any probability distribution, then the n^{th} moment of the probability distribution is given by the Riemann-Stieltjes integral

$$\mu'_n = \mathrm{E}(X^n) = \int_{-\infty}^{\infty} x^n \, dF(x)$$

where X is a random variable that has this cumulative distribution F and E[ ] is the expectation operator.

Central Moments

The k^{th} central moment of a real-valued random variable X is $\mu_k = \mathrm{E}[(X - \mathrm{E}[X])^k]$ where E is the expectation operator. For a continuous probability distribution with probability density function $f(x)$, the central moment about the mean μ is

$$\mu_k = \mathrm{E}\left[(X - \mathrm{E}[X])^k\right] = \int_{-\infty}^{+\infty} (x - \mu)^k f(x) \, dx.$$

The n^{th} central moment is translation-invariant for all $n > 0$, i.e. for any random variable X and any constant c

$$\mu_n(X + c) = \mu_n(X).$$

For all n, the n^{th} central moment is homogeneous of degree n

$$\mu_n(cX) = c^n \mu_n(X).$$

The additivity property for independent random variables X and Y is

$$\mu_n(X+Y) = \mu_n(X) + \mu_n(Y) \text{ provided } n \leq 3.$$

The corresponding formulas for the fourth and fifth central moments are

$$\mu_4(X+Y) = \mu_4(X) + \mu_4(Y) + 4\,\mu_3(X)\,\mu_Y + 4\,\mu_X\,\mu_3(Y) + 6\,\sigma^2_X\,\sigma^2_Y$$
$$\mu_5(X+Y) = \mu_5(X) + \mu_5(Y) + 5\,\mu_4(X)\,\mu_Y + 5\,\mu_X\,\mu_4(Y) + 10\,\mu_3(X)\,\sigma^2_Y + 10\,\mu_3(Y)\,\sigma^2_X.$$

The lower central moments have the following interpretations:

- The first central moment μ_1 is zero.
- The second central moment μ_2 is the variance, denoted σ^2, where σ is the standard deviation.
- The third central moment μ_3 is denoted $\underset{\sim}{y}$ and is used to define the third standardised moment γ_1 which is called skewness.
- The fourth central moment μ_4 is denoted $\underset{\sim}{b}$ and is used to define the fourth standardised moment γ_2 which is called kurtosis.
- The fifth central moment μ_5 is used to define the fifth standardised moment which is called hyperskewness.

The *standardised moment* is the central moment divided by the standard deviation of the same power as its number, i.e. μ_n / σ^n for the n^{th} moment.

Moment	Raw moment	Central moment	Standardised moment
1	μ or E[X] (mean)	0	0
2	μ_2' or E[X^2]	μ_2 or σ^2 (variance)	1
3	μ_3' or E[X^3]	μ_3 or $\underset{\sim}{y}$	γ_1 (skewness)
4	μ_4' or E[X^4]	μ_4 or $\underset{\sim}{b}$	γ_2 (non-excess kurtosis)
5	μ_5' or E[X^5]	μ_5	ζ_1 (hyperskewness)

***Table 1**: Raw, central, and standardised moments*

The relations between the central moments μ_n and raw moments μ_n' are

$$\mu_2 = -\mu_1'^2 + \mu_2'$$
$$\mu_3 = 2\,\mu_1'^3 - 3\,\mu_1'\,\mu_2' + \mu_3'$$
$$\mu_4 = -3\,\mu_1'^4 + 6\,\mu_1'^2\,\mu_2' - 4\,\mu_1'\,\mu_3' + \mu_4'$$
$$\mu_5 = 4\,\mu_1'^5 - 10\,\mu_1'^3\,\mu_2' + 10\,\mu_1'^2\,\mu_3' - 5\,\mu_1'\,\mu_4' + \mu_5'$$

and vice versa

$$\mu_2' = \mu_2 + \mu_1'^2$$
$$\mu_3' = \mu_3 + 3\,\mu_2\,\mu_1' + \mu_1'^3$$
$$\mu_4' = \mu_4 + 4\,\mu_3\,\mu_1' + 6\,\mu_2\,\mu_1'^2 + \mu_1'^4$$
$$\mu_5' = \mu_5 + 5\,\mu_4\,\mu_1' + 10\,\mu_3\,\mu_1'^2 + 10\,\mu_2\,\mu_1'^3 + \mu_1'^5.$$

Variance

If a random variable X has the expected value $\mu = \mathrm{E}[X]$, then the variance of X is

$$\operatorname{Var}(X) = \mathrm{E}\left[(X - \mu)^2\right].$$

It can be expanded as follows:

$$\begin{aligned}
\operatorname{Var}(X) &= \mathrm{E}\left[(X - \mu)^2\right] \\
&= \mathrm{E}\left[X^2 - 2\mu X + \mu^2\right] \\
&= \mathrm{E}\left[X^2\right] - 2\mu\ \mathrm{E}[X] + \mu^2 \\
&= \mathrm{E}\left[X^2\right] - 2\mu^2 + \mu^2 \\
&= \mathrm{E}\left[X^2\right] - \mu^2 \\
&= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2.
\end{aligned}$$

If the random variable X is continuous with probability density function $f(x)$, then the variance equals the second central moment given by

$$\operatorname{Var}(X) = \int (x - \mu)^2 f(x)\, dx\,,$$

where μ is the expected value

$$\mu = \int x\, f(x)\, dx\,,$$

and where the integrals are definite integrals taken for x over the range of X.

The variance is invariant with respect to changes in the location or scale parameter.

$$\begin{aligned}
&\operatorname{Var}(X + a) = \operatorname{Var}(X). \\
&\operatorname{Var}(aX) = a^2 \operatorname{Var}(X).
\end{aligned}$$

The variance of a sum of two random variables is given by

$$\operatorname{Var}(aX + bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2ab\ \operatorname{Cov}(X, Y),$$

and in general, for the sum of N random variables, the variance is

$$\operatorname{Var}\left(\sum_{i=1}^{N} X_i\right) = \sum_{i,j=1}^{N} \operatorname{Cov}(X_i, X_j) = \sum_{i=1}^{N} \operatorname{Var}(X_i) + \sum_{i \neq j} \operatorname{Cov}(X_i, X_j).$$

The variance of a finite sum of uncorrelated (not necessarily independent) random variables is equal to the sum of their variances. This stems from the above identity and the fact that for uncorrelated variables the covariance is zero.

$$\operatorname{Cov}(X_i, X_j) = 0 \ (i \neq j) \Rightarrow \operatorname{Var}\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} \operatorname{Var}(X_i).$$

These results lead to the variance of a linear combination as

$$\begin{aligned}
\operatorname{Var}\left(\sum_{i=1}^{N} a_i X_i\right) &= \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j \operatorname{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{N} a_i^2 \operatorname{Var}(X_i) + \sum_{i \neq j} a_i a_j \operatorname{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{N} a_i^2 \operatorname{Var}(X_i) + 2 \sum_{i<j} a_i a_j \operatorname{Cov}(X_i, X_j).
\end{aligned}$$

If all variables have the same variance σ^2 then, since division by n is a linear transformation, this implies that the variance of their mean is

$$\operatorname{Var}\left(\overline{X}\right) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}\left(X_i\right) = \frac{\sigma^2}{n}.$$

If the variables are correlated, then the variance of their sum is the sum of their covariances.

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \operatorname{Cov}(X_i, X_j).$$

The scaling property plus the covariance property $\operatorname{Cov}(aX, bY) = ab \cdot \operatorname{Cov}(X, Y)$ jointly imply that

$$\operatorname{Var}(aX + bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) + 2ab \operatorname{Cov}(X, Y).$$

The expression above can be extended to a weighted sum of multiple variables.

$$\operatorname{Var}\left(\sum_{i} a_i X_i\right) = \sum_{i} a_i^2 \operatorname{Var}(X_i) + 2 \sum_{i} \sum_{j>i} a_i a_j \operatorname{Cov}(X_i, X_j)$$

If X and Y are two random variables and the variance of X exists, then

$$\operatorname{Var}(X) = \operatorname{Var}(\operatorname{E}(X|Y)) + \operatorname{E}(\operatorname{Var}(X|Y)).$$

If two variables X and Y are independent, the variance of their product is

$$\operatorname{Var}(XY) = [E(X)]^2 \operatorname{Var}(Y) + [E(Y)]^2 \operatorname{Var}(X) + \operatorname{Var}(X) \operatorname{Var}(Y).$$

The variance of the product of two random variables X and Y in general is

$$\mathrm{Var}(XY) = \mathrm{E}(X^2Y^2) - [\mathrm{E}(XY)]^2$$

where the expressions are expanded as

$$\mathrm{E}(X^2Y^2) = \mathrm{Cov}(X^2,Y^2) + \mathrm{E}(X^2)\,\mathrm{E}(Y^2)$$

and

$$[\mathrm{E}(XY)]^2 = [\mathrm{Cov}(X,Y) + \mathrm{E}(X)\,\mathrm{E}(Y)]^2.$$

Thus, Var(*XY*) of two possibly dependent random variables *X* and *Y* can be expressed as

$$\mathrm{Var}(XY) = \mathrm{Cov}(X^2,Y^2) + [\mathrm{Var}(X) + [\mathrm{E}(X)]^2]\cdot[\mathrm{Var}(Y) + \mathrm{E}[(Y)]^2] - [\mathrm{Cov}(X,Y) + \mathrm{E}(X)\,\mathrm{E}(Y)]^2.$$

If the random variables *X* and *Y* are independent, i.e. $\mathrm{Cov}(X^2,Y^2) = \mathrm{Cov}(X,Y) = 0$, the expression reduces to

$$[\mathrm{Var}(X) + [\mathrm{E}(X)]^2] \cdot [\mathrm{Var}(Y) + [\mathrm{E}(Y)]^2] - [\mathrm{E}(X)\,\mathrm{E}(Y)]^2$$

and the two $[E(X)\,E(Y)]^2$ terms cancel out yielding

$$\mathrm{Var}(X)\,\mathrm{Var}(Y) + \mathrm{Var}(X)\,[\mathrm{E}(Y)]^2 + \mathrm{Var}(Y)\,[\mathrm{E}(X)]^2$$

which leads back to the expression for independent products above.

These additive and multiplicative properties of the second central moment yield expressions for the variance in decision and risk trees. This variance is then matched to the B-normal distribution that is derived from the skew-normal distribution below and calculated by NEMO.

Covariance

The covariance between two jointly distributed real-valued random variables *X* and *Y* with finite second moments is defined as the expected product of their deviations from their individual expected values

$$\mathrm{cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right]$$

where E[*X*] is the expected value of *X*. By using the linearity of expectations, this can be simplified to

$$\begin{aligned}\mathrm{cov}(X,Y) &= \mathrm{E}[(X - \mathrm{E}[X])\,(Y - \mathrm{E}[Y])] \\ &= \mathrm{E}[XY - X\,\mathrm{E}[Y] - \mathrm{E}[X]Y + \mathrm{E}[X]\,\mathrm{E}[Y]] \\ &= \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] - \mathrm{E}[X]\,\mathrm{E}[Y] + \mathrm{E}[X]\,\mathrm{E}[Y] \\ &= \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y].\end{aligned}$$

For several random variables having the same parent node, there are a larger number of covariances. These are more easily collected in matrix form. For random vectors $\mathbf{X} \in \mathrm{R}^m$ and $\mathbf{Y} \in \mathrm{R}^n$, the m × n covariance matrix is

$$\begin{aligned}\operatorname{cov}(\mathbf{X},\mathbf{Y}) &= \operatorname{E}\left[(\mathbf{X}-\operatorname{E}[\mathbf{X}])(\mathbf{Y}-\operatorname{E}[\mathbf{Y}])^{\mathrm{T}}\right] \\ &= \operatorname{E}\left[\mathbf{X}\mathbf{Y}^{\mathrm{T}}\right] - \operatorname{E}[\mathbf{X}]\operatorname{E}[\mathbf{Y}]^{\mathrm{T}},\end{aligned}$$

where $\mathbf{Y}^{\mathbf{T}}$ is the transpose of $\mathbf{Y}$. The (i, j)-th element of this matrix is equal to the covariance $\operatorname{cov}(X_i, Y_j)$ between the i-th scalar component of $\mathbf{X}$ and the j-th scalar component of $\mathbf{Y}$. By symmetry, $\operatorname{cov}(Y, X)$ is the transpose of $\operatorname{cov}(X, Y)$. Using a matrix form simplifies the handling of covariances.

Skewness

The skewness γ_1 of a random variable X is the third standardised moment

$$\gamma_1 = \operatorname{E}\left[\left(\tfrac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\operatorname{E}\left[(X-\mu)^3\right]}{(\operatorname{E}\left[(X-\mu)^2\right])^{3/2}}$$

where μ_3 is the third central moment, σ is the standard deviation, and E[ ] is the expectation operator.

The formula expressing skewness in terms of the raw moment $\operatorname{E}[X^3]$ is derived as follows:

$$\begin{aligned}\gamma_1 &= \operatorname{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \\ &= \frac{\operatorname{E}[X^3] - 3\mu\operatorname{E}[X^2] + 3\mu^2\operatorname{E}[X] - \mu^3}{\sigma^3} \\ &= \frac{\operatorname{E}[X^3] - 3\mu(\operatorname{E}[X^2] - \mu\operatorname{E}[X]) - \mu^3}{\sigma^3} \\ &= \frac{\operatorname{E}[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}\ .\end{aligned}$$

If two variables X and Y with third central moments $\underset{\sim}{\mu}_X$ and $\underset{\sim}{\mu}_Y$ are independent, it follows from the additivity property that the third central moment $\underset{\sim}{\mu}_{X+Y}$ of their sum $X+Y$ is

$$\underset{\sim}{\mu}_{X+Y} = \underset{\sim}{\mu}_X + \underset{\sim}{\mu}_Y.$$

The third central moment $\underset{\sim}{\mu}_{XY}$ of their product XY is

$$\begin{aligned}\underset{\sim}{\mu}_{XY} &= \operatorname{E}[(X-\mu_X)^3\,(Y-\mu_Y)^3] \\ &= \operatorname{E}[X^3Y^3 - 3\,\mu_X\,\mu_Y\,X^2Y^2 + 3\,\mu^2_X\,\mu^2_Y\,XY - \mu^3_X\,\mu^3_Y] \\ &= \operatorname{E}[X^3Y^3] - 3\,\mu_X\,\mu_Y\operatorname{E}[X^2Y^2] + 3\,\mu^2_X\,\mu^2_Y\operatorname{E}[XY] - \mu^3_X\,\mu^3_Y \\ &= \operatorname{E}[X^3]\operatorname{E}[Y^3] - 3\,\mu_X\,\mu_Y\operatorname{E}[X^2]\operatorname{E}[Y^2] + 2\,\mu^3_X\,\mu^3_Y\end{aligned}$$

in terms of the raw moments $\operatorname{E}[X^3]$, $\operatorname{E}[Y^3]$, $\operatorname{E}[X^2]$ and $\operatorname{E}[Y^2]$.

Since

$$\mathrm{E}[X^2] = \sigma^2_X + \mathrm{E}^2[X]$$

and

$$\mathrm{E}[X^3] = \underset{\sim}{\mu}_X + 3\,\mu_X\,\sigma^2_X + \mu^3_X$$

this yields

$$\begin{aligned}\underset{\sim}{\mu}_{XY} &= \mathrm{E}[X^3]\,\mathrm{E}[Y^3] - 3\,\mu_X\,\mu_Y\,\mathrm{E}[X^2]\,\mathrm{E}[Y^2] + 2\,\mu^3_X\,\mu^3_Y \\ &= (\underset{\sim}{\mu}_X + 3\,\mu_X\,\sigma^2_X + \mu^3_X)\,(\underset{\sim}{\mu}_Y + 3\,\mu_Y\,\sigma^2_Y + \mu^3_Y) \\ &\quad - 3\,\mu_X\,\mu_Y\,(\sigma^2_X + \mathrm{E}^2[X])\,(\sigma^2_Y + \mathrm{E}^2[Y]) + 2\,\mu^3_X\,\mu^3_Y \\ &= \underset{\sim}{\mu}_X\,\underset{\sim}{\mu}_Y + 3\,\underset{\sim}{\mu}_X\,\mu_Y\,\sigma^2_Y + 6\,\mu_X\,\sigma^2_X\,\mu_Y\,\sigma^2_Y \\ &\quad + 3\,\mu_X\,\sigma^2_X\,\underset{\sim}{\mu}_Y + \underset{\sim}{\mu}_X\,\mu^3_Y + \mu^3_X\,\underset{\sim}{\mu}_Y\end{aligned}$$

in terms of the central moments $\underset{\sim}{\mu}_X$, $\underset{\sim}{\mu}_Y$, σ^2_X and σ^2_Y.

Since the general skewness is $\gamma_1 = \underset{\sim}{\mu}_X / \sigma^3_X$, the additive and multiplicative properties of the third central moment yield expressions for the skewness in decision and risk trees. This skewness is then matched to the B-normal distribution skewness that is later derived from the skew-normal distribution.

Skewness of sums (and differences)

For two random variables X and Y, the skewness of the sum $X+Y$ is

$$S_{X+Y} = \frac{1}{\sigma^3_{X+Y}}\left[\sigma^3_X S_X + 3\sigma^2_X\sigma_Y S(X,X,Y) + 3\sigma_X\sigma^2_Y S(X,Y,Y) + \sigma^3_Y S_Y\right]$$

where S_X is the skewness of X, $S()$ is the coskewness of X and Y, and σ_X is the standard deviation of X. It then follows that the sum of two random variables can be skewed ($S_{X+Y} \neq 0$) even if both random variables are completely symmetric by themselves ($S_X = 0$ and $S_Y = 0$).
For three random variables X, Y, and Z, the coskewness $S(X, Y, Z)$ is defined as

$$S(X,Y,Z) = \frac{\mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])(Z-\mathrm{E}[Z])]}{\sigma_X\sigma_Y\sigma_Z}$$

where $\mathrm{E}[X]$ is the expected value of X. Ordinary skewness is a special case of coskewness where the three random variables are identical. This is analogous to the variance $\mathrm{Var}(X)$ of a random variable being the same as the covariance with itself, i.e. $\mathrm{Cov}(X, X)$. Thus, the skewness $S_X = S(X, X, X)$ of a random variable X can be expressed as

$$S(X,X,X) = \frac{\mathrm{E}\left[(X-\mathrm{E}[X])^3\right]}{\sigma^3_X}$$

For each $S()$-term in the formula for S_{X+Y}, the corresponding expression can be derived. For example, let $R(X, Y, Y)$ be $\sigma_X\sigma_Y\sigma_Y \cdot S(X, Y, Y)$. Then $R(X, Y, Y)$ can be expanded as

$$\begin{aligned}R(X,Y,Y) &= \mathrm{E}[(X-\mu_X)\,(Y-\mu_Y)\,(Y-\mu_Y)] \\ &= \mathrm{E}[XY^2 - \mu_X Y^2 - 2\mu_Y XY + 2\mu_X\mu_Y Y + \mu_Y{}^2 X - \mu_X\mu_Y{}^2].\end{aligned}$$

Since $E[XY^2] = E[X]E[Y^2] + Cov(X, Y^2)$ and $E[XY] = E[X]E[Y] + Cov(X, Y)$, and by the linearity of the expectation operator $E[\ ]$, $R(X, Y, Y)$ reduces to a combination of covariances.

$$\begin{aligned}
R(X, Y, Y) &= E[XY^2 - \mu_X Y^2 - 2\mu_Y XY + 2\mu_X\mu_Y Y + \mu_Y^2 X - \mu_X\mu_Y^2] \\
&= E[XY^2] - \mu_X E[Y^2] - 2\mu_Y E[XY] + 2\mu_X\mu_Y^2 + \mu_Y^2 E[X] - \mu_X\mu_Y^2 \\
&= E[XY^2] - \mu_X E[Y^2] - 2\mu_Y E[XY] + 2\mu_X\mu_Y^2 + \mu_Y^2\mu_X - \mu_X\mu_Y^2 \\
&= E[X]E[Y^2] + Cov(X, Y^2) - \mu_X E[Y^2] - 2\mu_Y(E[X]E[Y] + Cov(X, Y)) + 2\mu_X\mu_Y^2 \\
&= \mu_X E[Y^2] + Cov(X, Y^2) - \mu_X E[Y^2] - 2\mu_Y(\mu_X\mu_Y + Cov(X, Y)) + 2\mu_X\mu_Y^2 \\
&= Cov(X, Y^2) - 2\mu_Y Cov(X, Y).
\end{aligned}$$

Using the above formula and similar expressions for the other $R(\ )$-terms, S_{X+Y} can be computed as a sum of covariances, allowing skewness to be propagated and calculated. However, simulations indicate that the dependence effect on skewness (or third central moment) of $X+Y$ or $X-Y$ is negligible and not worth the effort. Thus, it is not included in the NEMO calculus.

Higher moments

The fourth standardized moment is defined as

$$\beta_2 = \frac{\mu_4}{\sigma^4}$$

where μ_4 is the fourth central moment and σ is the standard deviation. The basic definition of kurtosis is

$$Kurt(X) = E[X^4] - 3\ E^2[X^2].$$

Kurtosis is more commonly defined as the fourth cumulant divided by the square of the second cumulant

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\sigma^4} - 3$$

which is also known as excess kurtosis. This formula hints at a kinship between the second and fourth moments. However, simulations show that moments higher than three are of little influence in the BEDA method and are thus not included in the NEMO calculus.

The B-normal Distribution

The B-normal distribution method uses a skew-normal distribution defined using Owen's T and adapted to belief use. It expresses the resulting distribution of outcomes of events in the context of belief evaluations. The joint distribution of all modelled consequence pairs is approximately skew-normally distributed, with the approximation improving as the number of consequences increase. The parameters of the distribution are determined by the NEMO calculus where moments are used to represent properties of the output distributions. The adaptions of skew-normality to belief use (B-normality) consist of:

- location and scale parameters to match the expected value and variance with the normal distribution while maintaining the same skewness
- interpolated truncation toward the orthogonal hull
- handling of large skew ($\gamma_1 > 0.955$), where skew-normality does not hold, by successive limiting

- interpretation of decision and risk trees as PGC situations

The skewness γ_1 of the B-normal distribution depends only on the shape parameter and is symmetric about the origin:

$$\gamma_1 = \frac{4-\pi}{2} \cdot \frac{\left(\delta\sqrt{2/\pi}\right)^3}{(1-2\delta^2/\pi)^{3/2}}$$

The limit of the skewness is finite and depends on the sign of α. The maximum skewness ($\gamma_1 \approx 0.995272$) is obtained by setting $\delta = 1$. The skewness resulting from the B-normal method can be larger, though. For example, it is at most $\gamma_1 = 1.4$ for the product XY where X is uniformly distributed and Y is triangular with $a = c = 0$ and $b > 0$. Then unmoderated skew-normality cannot be used since α cannot be determined from

$$\delta = \frac{\alpha}{\sqrt{1+\alpha^2}}.$$

The requirement of maintaining the skewness (alignment) yields

$$|\delta| = \sqrt{\frac{\pi}{2} \cdot \frac{|\hat{\gamma}_1|^{\frac{2}{3}}}{|\hat{\gamma}_1|^{\frac{2}{3}} + ((4-\pi)/2)^{\frac{2}{3}}}}$$

where $\hat{\gamma}_1$ is the B-skewness from the approximation. The sign of δ is the same as the sign of $\hat{\gamma}_1$. The shape parameter of the approximation is then

$$\hat{\alpha} = \delta/\sqrt{1-\delta^2}.$$

It can be seen that the B-normal distribution is limited by $|\delta| < 1$ since $|\alpha| \to \infty$ as $|\delta| \to 1$. In reality, it is limited by $|\delta| < 0.995037$, yielding a maximal $|\alpha|$ of 10, above which the skewness is limited. Thus, the practical skewness limit is $\gamma_1 \approx 0.955557$. For $|\alpha| > 10$, the skew-normal distribution tends to a half-normal distribution which is not appropriate in belief modelling.

B-normal Alignment

To employ the B-normal method, the skewed distribution must be aligned to have the same variance and expected value as its unskewed counterpart plus displaying the correct shape (skew). The alignment (matching) of the B-normal distribution is done in three steps:

1. Obtain the correct shape α from the skew γ_1 by solving for δ in the formula in the previous section and substituting δ by α, also from the previous section.
2. Once the shape α is determined, this changes the variance of the B-normal distribution. To bring it back to the desired variance σ^2, use the formula for the variance from the skew-normal section and solve for ω^2.
3. Now, since the shape and variance are determined, the expected value of the distribution is changed. To bring it back to the desired expected value E(X), use the formula for the mean from the skew-normal section and solve for ξ.

Once the parameters α, ω^2 and ξ have been obtained, the B-normal distribution is parametrically determined and ready to use. The process is easiest shown with an example. The skew-normal

parameters location, scale, and shape correspond to the mean, variance, and skewness of a normal distribution. The mean of the skew-normal distribution is

$$E(X) = \xi + \omega\delta\sqrt{\frac{2}{\pi}}.$$

With parameters for location ξ = 0, scale ω = 1, and shape α = 0, a standard normal distribution (with mean 0 and standard deviation 1) is obtained:

$$E(X) = 0 + 1\frac{0}{\sqrt{1+0^2}}\sqrt{\frac{2}{\pi}} = 0$$

Start with fixating the desired α, say α = 5. Then δ will change and the mean will also change and become positive. To keep the mean at 0 with a positive α, other parameters, e.g. ω, must be decreased. The variance is

$$Var(X) = \omega^2\left(1 - \frac{2\delta^2}{\pi}\right).$$

Obtain the value of ω by the variance formula:

$$1 = \omega^2\left(1 - \frac{2\frac{5}{\sqrt{1+5^2}}^2}{\pi}\right) = 0.3878656\ \omega^2.$$

Then ω should be ≈ 1.605681 (negative or positive). The mean becomes

$$0 = \xi \pm 1.605681 \cdot 0.9805807\sqrt{\frac{2}{\pi}} = \xi \pm 1.256269.$$

Thus, the following parameters yield the intended distribution: ξ ≈ ± 1.256269, ω ≈ ± 1.605681 (the opposite sign of the location) and α = 5. These are the parameters for the B-normal counterpart of a normal distribution with mean 0 and variance 1, but with a skewness introduced through α.

The B-normal CDF

The cumulative distribution function (CDF) describes the probability that a real-valued random variable *X* with a given probability distribution will be found at a value less than or equal to *x*.

For every real number *x*, the cumulative distribution function of a real-valued random variable *X* is given by

$$F_X(x) = \mathrm{P}(X \leq x),$$

where the right-hand side represents the probability that the random variable *X* takes on a value less than or equal to *x*. The probability that *X* lies in the interval (*a*, *b*], where $a < b$, is therefore

$$\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a).$$

where (*a*, *b*] is a semi-closed interval.

The CDF of a continuous random variable *X* can be defined in terms of its probability density function *f* as follows:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt.$$

Every cumulative distribution function *F* is (not necessarily strictly) monotone, non-decreasing, and right-continuous. Furthermore,

$$\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to +\infty} F(x) = 1.$$

If the CDF *F* is absolutely continuous, then there exists a Lebesgue-integrable function *f*(*x*) such that

$$F(b) - F(a) = \mathrm{P}(a < X \leq b) = \int_a^b f(x)\,dx$$

for all real numbers *a* and *b*. The function *f* is equal to the derivative of *F* almost everywhere, and is the probability density function of the distribution of *X*.

In the use of the B-normal distribution, the PGC principle leads to the B-normal approximation of the total (joint) multivariate distribution a set of decisions. In the approximation, the tails are cut off for values outside the orthogonal hull of the combined random variables and the cumulative distribution function for the tails is interpolated between the cut-off points and the tangents of the B-normal CDF curve instead of a steep truncation as in a truncated skew-normal distribution.

Individual Risk Distributions

When not much is known about the underlying distribution of an outcome, it is reasonable to use a two-point distribution for modelling requiring only upper and lower bounds. But if the modal outcome is also known or could be reasonably estimated, then the probability of the outcome can be better represented by a three-point distribution. The triangular distribution, along with the Beta and Erlang distributions, is therefore widely used in project management models (such as PERT) to model events which take place within an interval defined by a minimum and maximum value.

Simulations have shown that the triangular distribution yields results similar to Beta-PERT in general and to Erlang-PERT in particular. Therefore, the triangular distribution can be seen as a very good representative of the class of three-point distributions. Thus, the class of two-point

distributions is represented by a uniform distribution and a triangular distribution represents the class of three-point distributions. Together, they cover a very wide range of modelling needs.

Uniform Distribution

The probability density function of the continuous uniform distribution is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

The values of $f(x)$ at the two boundaries a and b are unimportant because they do not alter the values of the integrals of $f(x)\,dx$ over any interval, nor of $x \cdot f(x)\,dx$ or any higher moment.

In terms of mean μ and variance σ^2, the probability density is

$$f(x) = \begin{cases} \frac{1}{2\sigma\sqrt{3}} & \text{for } -\sigma\sqrt{3} \le x - \mu \le \sigma\sqrt{3} \\ 0 & \text{otherwise} \end{cases}$$

The cumulative distribution function is

$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \le x < b \\ 1 & \text{for } x \ge b \end{cases}$$

The first two moments of the distribution are

$$E(X) = \frac{1}{2}(a+b)$$

and

$$V(X) = \frac{1}{12}(b-a)^2$$

Due to the shape of the distribution, the skewness is zero. Solving for parameters a and b, given known first and second moments E(X) and V(X), yields

$$a = E(X) - \sqrt{3V(X)}$$
$$b = E(X) + \sqrt{3V(X)}$$

Triangular Distribution

The triangular distribution is used as a representative for a class of distributions (three-point distributions) commonly used in business management. It is used as a subjective description of a population for which there is only limited sample data, and especially in cases where the relationship between variables is known but data is scarce (possibly because of the high cost of

collection). It is based on best estimates of the minimum and maximum as well as the modal value.

The triangular distribution is a continuous probability distribution with lower limit *a*, upper limit *b*, and mode *c*, where $a \le c \le b$ and $a < b$. The probability density function is given by

$$f(x|a,b,c) = \begin{cases} 0 & \text{for } x < a, \\ \frac{2(x-a)}{(b-a)(c-a)} & \text{for } a \le x \le c, \\ \frac{2(b-x)}{(b-a)(b-c)} & \text{for } c < x \le b, \\ 0 & \text{for } b < x, \end{cases}$$

The cumulative distribution function is

$$\begin{cases} 0 & \text{for } x < a, \\ \frac{(x-a)^2}{(b-a)(c-a)} & \text{for } a \le x \le c, \\ 1 - \frac{(b-x)^2}{(b-a)(b-c)} & \text{for } c < x \le b, \\ 1 & \text{for } b < x. \end{cases}$$

Median:

$$\begin{cases} a + \frac{\sqrt{(b-a)(c-a)}}{\sqrt{2}} & \text{for } c \ge \frac{a+b}{2}, \\ b - \frac{\sqrt{(b-a)(b-c)}}{\sqrt{2}} & \text{for } c \le \frac{a+b}{2}. \end{cases}$$

Mean:

$$\frac{a+b+c}{3}$$

Variance:

$$\frac{a^2+b^2+c^2-ab-ac-bc}{18}$$

Skewness ($\mu_3 / \sigma^{3/2}$):

$$\frac{\sqrt{2}(a+b-2c)(2a-b-c)(a-2b+c)}{5(a^2+b^2+c^2-ab-ac-bc)^{\frac{3}{2}}}$$

Defining $t = b - a$ (range) and $q = (c - a)/t$ (relative mode within range), the first raw moment and the following four central moments can be calculated as

$\mu = a + t \cdot (q + 1)/3$
$\mu_2 = t^2 \cdot (1 - q + q^2)/18$
$\mu_3 = t^3 \cdot (2 - 3q - 3q^2 + 2q^3)/270$

$\mu_4 = t^4 \cdot (1 - q + q^2)^2 / 135$
$\mu_5 = 2t^5 \cdot (2 - 3q - 3q^2 + 2q^3) \cdot (1 - q + q^2) / 1701$

which simplifies the algorithms. Note that $18 = 2 \cdot 3^2$, $135 = 5 \cdot 3^3$, $270 = 2 \cdot 5 \cdot 3^3$ and $1701 = 7 \cdot 3^5$ if the denominators are split into prime factors.

Note 1: Note that $\mu_5 = 40\ \mu_2\ \mu_3 / 7$. The similarities between the 3rd and 5th central moments indicate why the 5th standardised moment is called hyperskewness. The hyperskewness of a random variable is its ordinary skewness moderated by its variance. If $\mu_3 = 0$ then $\mu_5 = 0$. Since $\mu_2 \geq 0$ it follows that μ_3 and μ_5 have the same sign. For a given skew, high hyperskewness corresponds to heavy tails and small movement of mode, while low hyperskewness corresponds to larger changes in shoulders.

Note 2: Further note that $\mu_4 = 12\ (\mu_2)^2 / 5$. These two notes indicate why moments higher than three are not used for the determination of B-normality but rather used for checking the fit since they add no new information. Their information is already contained in the lower moments.

Compared with most other three-point distributions, the triangular distribution considers the entire interval [*a*, *b*] to a larger extent and is thus a better and more consistent companion to the two-point uniform distribution in the B-normal method. The statements by the decision-maker are interpreted in a more similar way, which adds consistency to the process. The first raw moment of many three-point distributions can be written $\mu_1(\lambda) = (a + b + \lambda c) / (\lambda + 2)$ with the triangular distribution having $\lambda = 1$ as parameter.[1] Beta-PERT usually has $\lambda = 4$ and Erlang-PERT has $\lambda = 3$ but various other parameter values, including non-integers, have been suggested to represent the user input. In practise, a higher λ value tends to underestimate the uncertainty. The underestimation is then amplified through multiplication. The triangular distribution is less centre-weighted and thus less prone to underestimation, even if the differences are not large. However, there are no compelling reasons to use any three-point distribution other than triangular in belief modelling.

Dirichlet Distribution

The Dirichlet distribution is a family of continuous multivariate probability distributions parameterised by a vector $\boldsymbol{\alpha}$ of positive numbers. It is a multivariate generalisation of the beta distribution. The Dirichlet distribution of order $K \geq 2$ with parameters $\alpha_1,\ldots,\alpha_K > 0$ has a probability density function given by

$$f(x_1, \ldots, x_{K-1}; \alpha_1, \ldots, \alpha_K) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

where

$$x_K = 1 - \sum_{i=1}^{K-1} x_i$$

and bounded by

$$x_1, \ldots, x_{K-1} \geq 0$$
$$x_1 + \cdots + x_{K-1} \leq 1$$

[1] The companion two-point uniform distribution has $\lambda = 0$.

Thus, $x_1,\ldots,x_K$ belongs to a standard simplex. The normalising constant in the density function is the multivariate beta function, which can be expressed in terms of the gamma function

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}, \qquad \boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_K).$$

Moments

Let X be a vector of Dirichlet distributed random variables

$$X = (X_1,\ldots,X_K) \sim \mathrm{Dir}(\alpha)$$

and let the first $K-1$ components be distributed according to the function above. Then the last component is given by

$$X_K = 1 - \sum_{i=1}^{K-1} X_i.$$

Thus, there is a loss of one degree of freedom (DoF). Let the parameter α_0 be

$$\alpha_0 = \sum_{i=1}^{K} \alpha_i$$

Then the expected value and variance are

$$\mathrm{E}[X_i] = \frac{\alpha_i}{\alpha_0},$$
$$\mathrm{Var}[X_i] = \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}$$

Since there is a strong dependency when a DoF is lost, the covariance must also be considered in the mass calculations. It can be seen that the covariance is separable, and the subcovariance calculation technique mentioned in section 2 can be employed.

$$\mathrm{Cov}[X_i, X_j] = \frac{-\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)}$$

More generally, moments of Dirichlet-distributed random variables can be expressed as

$$\mathrm{E}\left[\prod_{i=1}^{K} x_i^{\beta_i}\right] = \frac{B(\boldsymbol{\alpha}+\boldsymbol{\beta})}{B(\boldsymbol{\alpha})} = \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\Gamma\left[\sum_{i=1}^{K}(\alpha_i+\beta_i)\right]} \times \prod_{i=1}^{K}\frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)}$$

Mode

The mode of the distribution is the vector $(x_1, ..., x_K)$ with

$$x_i = \frac{\alpha_i - 1}{\alpha_0 - K}, \qquad \alpha_i > 1$$

Marginal distributions

The marginal distributions of Dirichlet are beta distributions

$X_i \sim \mathrm{Beta}(\alpha_i, \alpha_0 - \alpha_i)$

which is why they are important in a software implementation and therefore discussed next.

Beta Distribution

The beta distribution is a family of continuous probability distributions defined on [0, 1] and parameterised by two shape parameters, α and β, that control the shape of the distribution. The probability density function of the beta distribution, for $0 \le x \le 1$ and shape parameters $\alpha, \beta > 0$, is a power function of the variable x and its reflection $(1 - x)$.

$$\begin{aligned} f(x;\alpha,\beta) &= \text{constant} \cdot x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1} \\ &= \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} \end{aligned}$$

where $\mathbf{\Gamma}(\,)$ is the gamma function. The beta function $\mathbf{B}(\,)$ acts as a normalisation constant to ensure that the total probability integrates to 1.

Mode and expected value

The mode of a beta distributed random variable X with $\alpha, \beta > 1$ is the most likely value of the distribution (corresponding to the peak in the pdf) and is given by

$$\frac{\alpha - 1}{\alpha + \beta - 2}$$

The expected value μ of a beta distributed random variable X with two parameters α and β is a function of the ratio β / α of these parameters.

$$\begin{aligned} \mu = \mathrm{E}[X] &= \int_0^1 x f(x;\alpha,\beta)\,dx \\ &= \int_0^1 x\,\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\,dx \\ &= \frac{\alpha}{\alpha+\beta} \\ &= \frac{1}{1+\frac{\beta}{\alpha}} \end{aligned}$$

Variance

The variance of a beta distributed random variable X with parameters α and β is

$$\operatorname{var}(X) = \mathrm{E}[(X-\mu)^2] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

The deviation around the mean for the distribution is

$$\mathrm{E}[|X - E[X]|] = \frac{2\alpha^\alpha \beta^\beta}{\mathrm{B}(\alpha,\beta)(\alpha+\beta)^{\alpha+\beta+1}}$$

It can be approximated as

$$\frac{\text{mean abs. dev. from mean}}{\text{standard deviation}} = \frac{\mathrm{E}[|X - E[X]|]}{\sqrt{\operatorname{var}(X)}} \approx \sqrt{\frac{2}{\pi}}\left(1 + \frac{7}{12(\alpha+\beta)} - \frac{1}{12\alpha} - \frac{1}{12\beta}\right), \text{ if } \alpha,\beta > 1.$$

At the limit $\alpha \to \infty$, $\beta \to \infty$, the ratio of the mean absolute deviation to the standard deviation (for the beta distribution) becomes equal to the ratio of the same measures for the normal distribution, i.e. $(2/\pi)^{½}$. This is the same ratio that appears in the B-normal distribution.

Skewness

The skewness of the distribution in terms of the shape parameters α and β is

$$\gamma_1 = \frac{\mathrm{E}[(X-\mu)^3]}{(\operatorname{var}(X))^{3/2}} = \frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$$

For more straightforward computations, the skewness can instead be expressed in terms of the variance var and the mean μ. This simplifies the computations.

$$\gamma_1 = \frac{\mathrm{E}[(X-\mu)^3]}{(\operatorname{var}(X))^{3/2}} = \frac{2(1-2\mu)\sqrt{\operatorname{var}}}{\mu(1-\mu)+\operatorname{var}} \text{ if } \operatorname{var} < \mu(1-\mu)$$

This concludes the description of the BEDA method and the implementation phase ensues.

Summary

Interval decision analysis often results in overlaps, meaning that even though one alternative (A_1) has a higher expected value than another (A_2) and should be preferred according to the principle of maximising the expected value, the best possible variable assignments for alternative A_2 is higher than the worst possible variable assignments for alternative A_2. This common situation makes it impossible to discard A_2 based on reasons of admissibility. In order not to end the analysis inconclusively, the DELTA method (Danielson, 1997) employs the concept of contraction, in which the intervals are decreased proportionally until all possible variable assignments yield A_1 having higher expected value than A_2, i.e. making A_1 admissible. The contraction is measured as a percentage such that the original intervals represent 0% and the intervals decreased to singular points represent 100% contraction. The lower the contraction level

required to reach admissibility, the more stable the analysis result is. In other words, the contraction is a sensitivity analysis of the result. But it has the disadvantage that the contracted intervals must be accepted by the decision-maker to represent the original decision situation. While a low contraction rate easily makes this possible, higher contraction rates might constitute a different problem formulation than the original one.

To alleviate this disadvantage, the BEDA method is introduced in this report. It measures the belief in each of the input intervals and, based on that information, delivers a measure of belief in the output, i.e. in the expected values of the different alternatives analysed. This is done while taking the complete input intervals into account, as opposed to the DELTA method that zooms in on central sub-intervals.

To sum up, the BEDA method consists of three steps:
1. Modelling the belief in the input intervals by assigning belief distributions to them. The distributions can be, as in this report, Dirichlet/Beta distributions for probabilities and criteria weights and triangle or uniform distributions for utilities or values, but other distributions can be used if desired.
2. Through the PGC principle, and thus using the B-normal distribution for the analysis output (the expected values), determine the appropriate parameters by employing the NEMO calculus. The appropriateness of the selected parameters has been verified through extensive simulations.
3. Calculate the resulting belief in each alternative by a B-normal calculation. The beliefs can be compared by using the difference concepts (delta and gamma) from the DELTA method, resulting in belief levels for sets or pairs of alternatives.

Acknowledgement

The work in this report is research commissioned by [redacted]. The results in the report are made publically available after an embargo period, while the specific algorithms and computer code are intended only for internal use by [redacted].

Version 2 120909: Formula on page 18 corrected.

Sources

Azzalini, A. (1985). "A class of distributions which includes the normal ones", *Scandinavian Journal of Statistics*, **12**, pp.171–178.

Danielson, M. (1997). *Computational Decision Analysis*, PhD thesis, Royal Institute of Technology, Stockholm, Sweden.

Kendall, M.G. & Stuart, A. (1969). *The Advanced Theory of Statistics, Volume 1: Distribution Theory*, 3rd edition, Griffin.

Kotz, S. & van Dorp, J.R. (1999). "A novel extension to the triangular distribution and its parameter estimation", *The Statistician*, **51**, pp.63-79.

Kotz, S. & van Dorp, J.R. (2004). *Beyond Beta, Other Continuous Families of Distributions with Bounded Support and Applications*, World Scientific Press, Singapore.

Loeve, M. (1977). "Probability Theory", *Graduate Texts in Mathematics*, Volume 45, 4th edition, Springer-Verlag.

O'Hagan, A. & Leonhard, T. (1976). "Bayes estimation subject to uncertainty about parameter constraints", *Biometrika*, **63**, pp.201–202.

Owen, D.B. (1956). "Tables for computing bivariate normal probabilities", *Ann. Math. Statist.*, **27**, pp.1075–1090.