Analytic Uncertainty Modeling

Henry Neimeier
The MITRE Corporation
7525 Colshire Drive
McLean, Virginia, 22102, USA

Abstract

The analytic uncertainty modeling technique is useful whenever sensitivity analysis is important. It provides the entire resulting probability distribution instead of a single uncertain point estimate of the mean. Both analytic development costs, and computer execution costs are far less than in discrete event simulation. The price paid is some lack in modeling flexibility.

Discrete simulation requires multiple long simulation runs to obtain a statistically significant point estimate. The different result values from multiple runs with identical parameter values but different random number seeds, are averaged to obtain the point estimate of the mean result value. Conversely, the analytic solution gives the entire resulting probability distribution with minimal calculation. The analytic solution also considerably simplifies sensitivity analysis. A single analytic run is done for each input parameter setting. Discrete event simulation requires multiple runs for each input parameter, to obtain a statistically significant mean result.

In functional economic analysis we are interested in the relative future costs of alternative systems. There are uncertainties in process performance, resource requirements, cost estimates, investment required, workload, interest and inflation rates. There is also uncertainty in the future projection of these elements. Analytic uncertainty modeling provides a simple way of calculating output measure uncertainty from model input parameter uncertainties.

Analytic Uncertainty Modeling

Overview

The beta distribution analytic technique outlined in this paper is useful whenever sensitivity analysis is important. It provides the entire resulting probability distribution vice a single uncertain point estimate of the mean. Both analytic development costs, and computer execution costs are far less than in discrete event simulation. The price paid is some lack in modeling flexibility.

With the appropriate choice of parameter values, the beta distribution closely fits all the classical probability distributions. The sums and products of beta variates are also approximately beta distributed. This paper documents the error in beta approximation. The beta distribution can be fit based on the minimum, mean, maximum, and standard deviation statistics. In a complex results calculation all that is required is to keep track of these statistics as the calculation proceeds. At any point in a calculation, the probability distribution of the result, can be derived by fitting a beta distribution based on the four statistics.

Discrete Event Simulation Times

Discrete simulation requires multiple long simulation runs to obtain a statistically significant point estimate. The different result values from multiple runs with identical parameter values but different random number seeds, are averaged to obtain the point estimate of the mean result value. Conversely, the analytic solution gives the entire resulting probability distribution with minimal calculation. The analytic solution also considerably simplifies sensitivity analysis. A single analytic run is done for each parameter setting vice multiple runs for a statistically significant result. The simulation time (T) required to be 95 percent confident in a relative error (e) is approximated by the following equation for open GI/G/1 (general independent inter arrival times, general service times, 1 server) queuing networks:

$$T = 8\,t\,(C_a^2 + C_s^2)\,Z^2 / (r^2(1-r^2)\,e^2)$$

Where:

- T = simulation time for a specified relative error
- t = service time
- C_a^2 = square coefficient of variation in inter arrival time (variance in inter arrival time divided by the mean inter arrival time squared)
- C_s^2 = square coefficient of variation in service times
- Z = unit normal deviate (Z=2 for 95 percent confidence
- r = utilization (service time divided by inter arrival time)
- e = tolerated relative error

Figure 1 is a semi-log plot of simulation time required for 95 percent confidence in a specified relative error as a function of utilization. It represents the exponential inter arrival and service time case (Ca=Cs=1). Note that at high utilization and low relative errors extremely long simulation times are required. To achieve 5 percent relative error in the mean on an 80 percent utilized queuing network requires one million service times.

In functional economic analysis we are interested in the relative future costs of alternative systems. There are uncertainties in process performance, resource requirements, cost estimates, investment required, workload, interest and inflation rates. There is also uncertainty in the future projection of these elements. Thus there is uncertainty in the discounted present value cost distribution for each alternative system. A plot of cumulative probability versus cost, aids the decision process. The entire range in cost distribution is of interest. Figure 2 shows the expected

number of simulation events required to obtain an event in the tail of the result distribution when using discrete event simulation. The equation plotted is:

$$E = 1 / P^{C}$$ Where:

E = expected number of simulation events

P = distribution tail probability

C = uncertain model components

The lower the tail probability, and the more components in the model, the more events are required. For example, an average of one million simulation events are required in a six component model to simultaneously be in the 10 percent tail of all component distributions. To simultaneously be in the 1 percent tail requires an average of one trillion simulation events. In the limit it requires an infinite number of simulation events to capture the entire range of results. Thus discrete event simulation is not practical if one is interested in the entire result distribution in other than very small models with few components. If the minimum and maximum distribution values are not needed then discrete event simulation is practical. However, even in this case the model development, execution, and sensitivity analysis costs are higher.

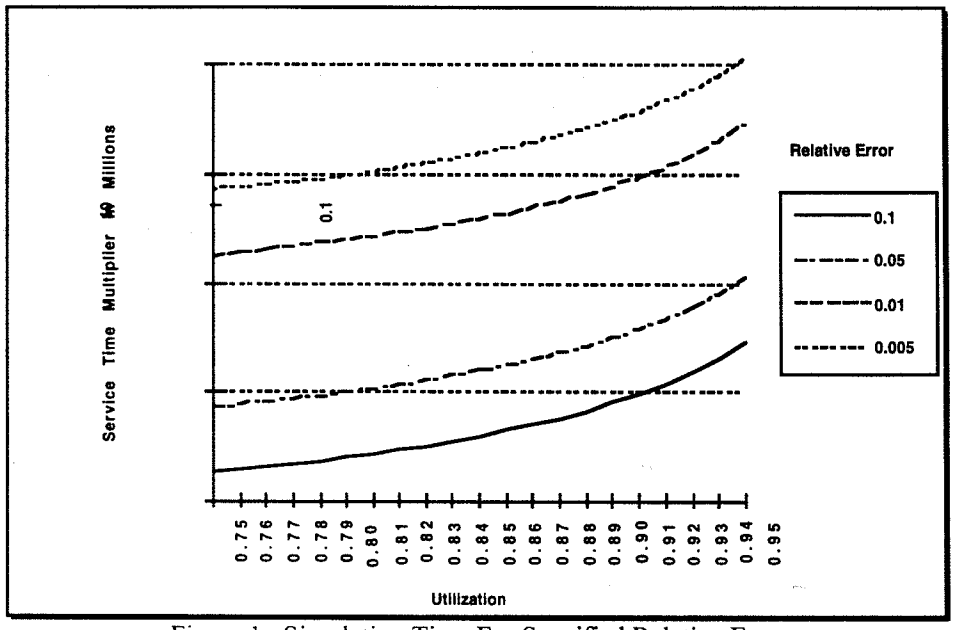


Figure 1. Simulation Time For Specified Relative Error

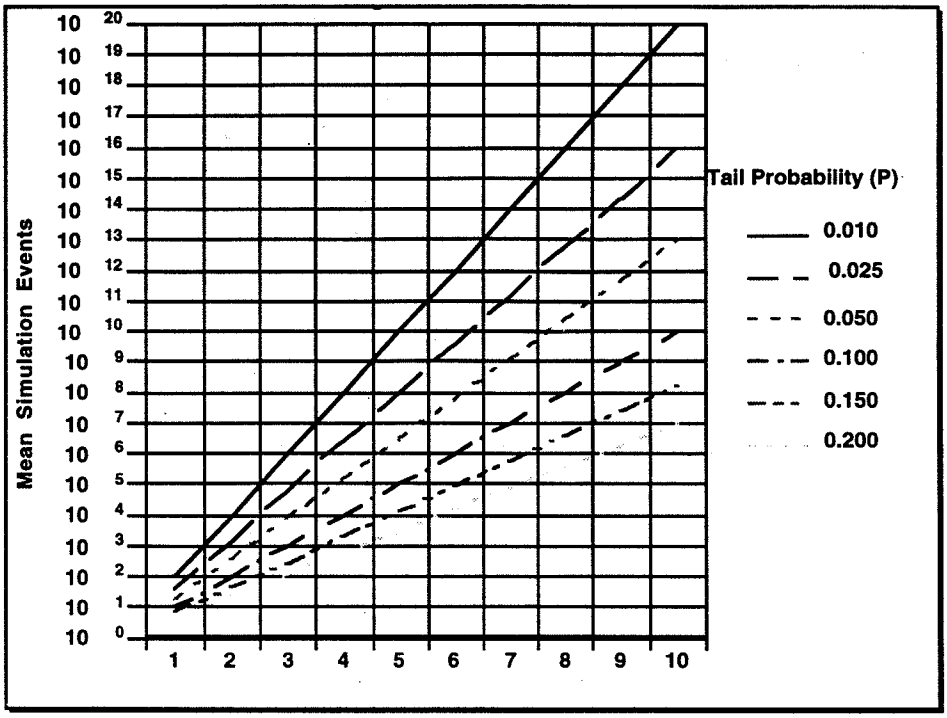


Figure 2. Simulation Events For A Specified Tail Probability Versus Model Components ($1/P^C$)

Beta distribution

The beta density distribution is bounded by high (max) and low (min) values.

$$P = C \{(x\text{-min})/(\text{max-min})\}^{a-1} \{1-(x\text{-min})/(\text{max-min})\}^{b-1}$$

Where:

P is the probability density

C normalizes the area under the distribution to unity

$$(C = 1/\int_0^1 x^{a-1} (1-x)^{b-1} dx)$$

min is the minimum variable value

max is the maximum variable value

If a random variable is bounded then its distribution function is uniquely determined by its moments (reference 3, p 126). The beta distribution can be fit based on the minimum, maximum, mean and variance. The fit parameters a and b are based on a range variable r and a skewness variable s, and are defined in standard terms as:

$$r = \text{variance} / (\text{max} - \text{min})^2$$
$$s = (\text{mean} - \text{min}) / (\text{max} - \text{min})$$
$$a = s^2 (1 - s) / r - s$$
$$b = s (1 - s) / r - 1 - a$$

If (a-1)(b-1)>0 then the beta distribution has a mode at

$$\text{min} + (\text{max} - \text{min}) (a - 1) / (a + b - 2)$$

Figure 3 plots the beta distribution a and b parameter values as a function of the r and s parameters. The ratio of standard deviation to range (square root of r) selects the appropriate curve.

Separate curves are presented for a and b parameters at a selected set of "r" values. The symbol legend is at the right. The symbol for the a parameter curve is a darkened version of the b parameter symbol. The "s" skewness parameter ((mean-minimum) /(maximum-minimum)) is plotted along the abscissa. Select the appropriate value and read the a and b values off the ordinate. Note that the smaller the standard deviation the larger the a and b parameters. For the mean at the midpoint between minimum and maximum values (s=.5) we have a symmetric distribution. The a and b parameters are equal in this symmetric case.

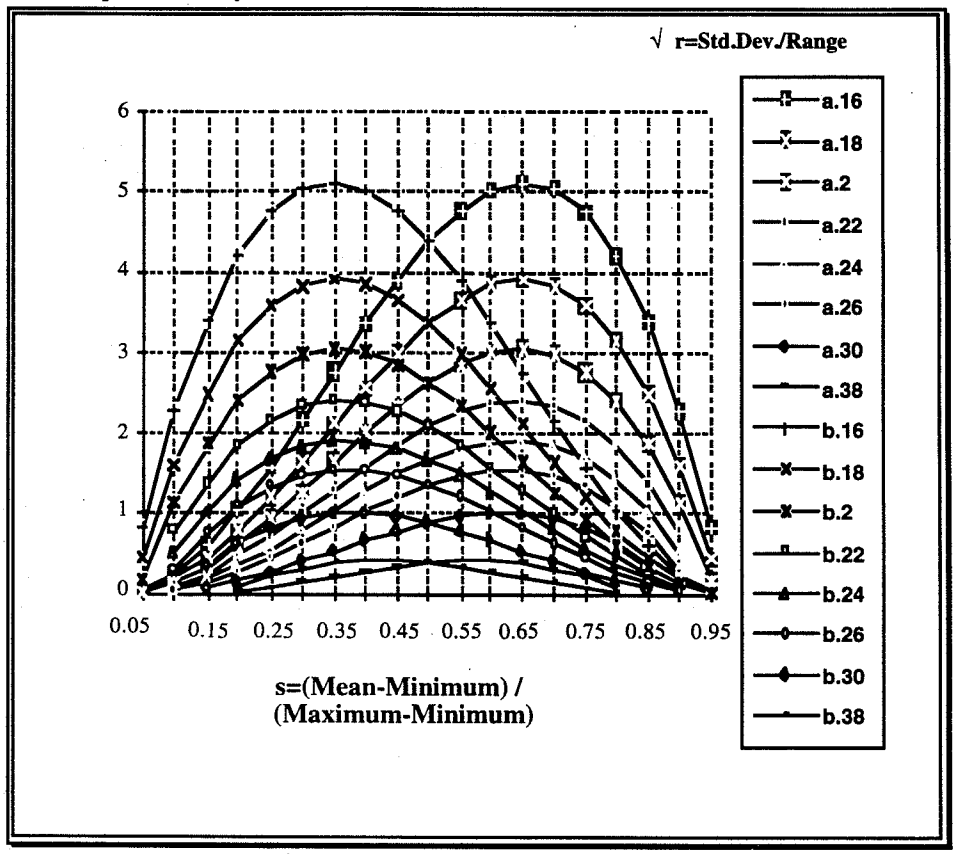


Figure 3. Beta Distribution a and b Parameters

Figure 4 shows examples of different beta distribution shapes. For the mean midway between the minimum and maximum values, the distribution is symmetric with equal a and b parameter values. Unity a and b parameters yield a uniform distribution (1,1). Values less than 1 lead to a "U" shaped distribution (.5,.5). With a and b equal 2 (2,2) the distribution has a parabolic shape. At higher a and b values (6,6; 12,12) the distribution has a shape similar to a normal distribution. The lower the standard deviation relative to the range, the higher the a and b parameter values, and the more peaked the distribution shape. If a is greater than b then the distribution has a negative skew (9,2). Conversely, if a is less than b the distribution has a positive skew. Triangular distributions (1,3) and left (.5,3) and right "J" shaped distribution shapes are also possible.

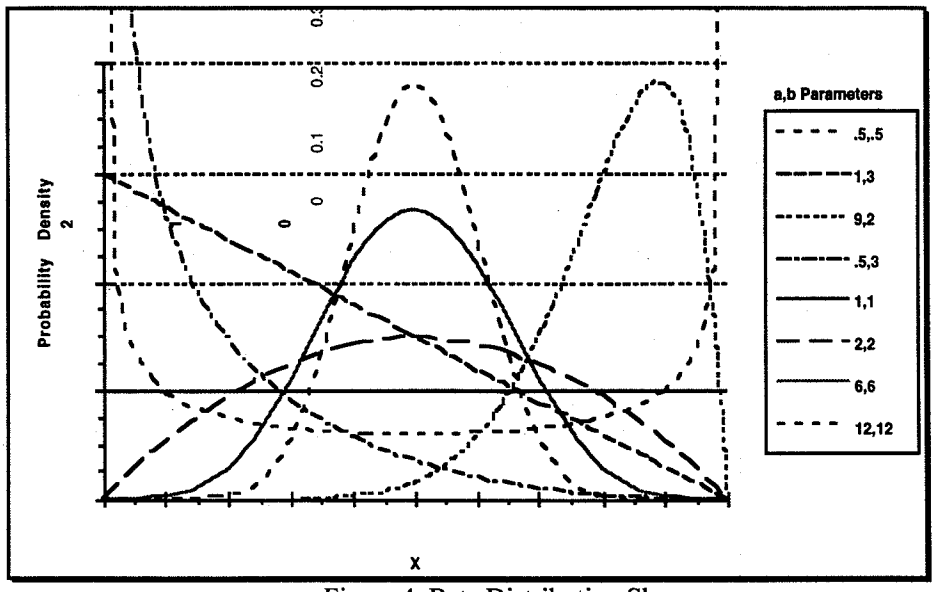


Figure 4. Beta Distribution Shapes

Spanning the classical distribution space

Figure 5 is a skew (÷b1) kurtosis (b2) plot of the popular classical continuous statistical distributions. The beta distribution covers most of the classical distribution area except the log normal line. In the case of the log normal line a log transformation is performed before the beta distribution is fit. The beta distribution shapes are shown on the chart regions. Beta distributions can fit uniform, triangular, J shaped, U shaped, exponential, Erlang, hyper-exponential, gamma, and normal distributions.

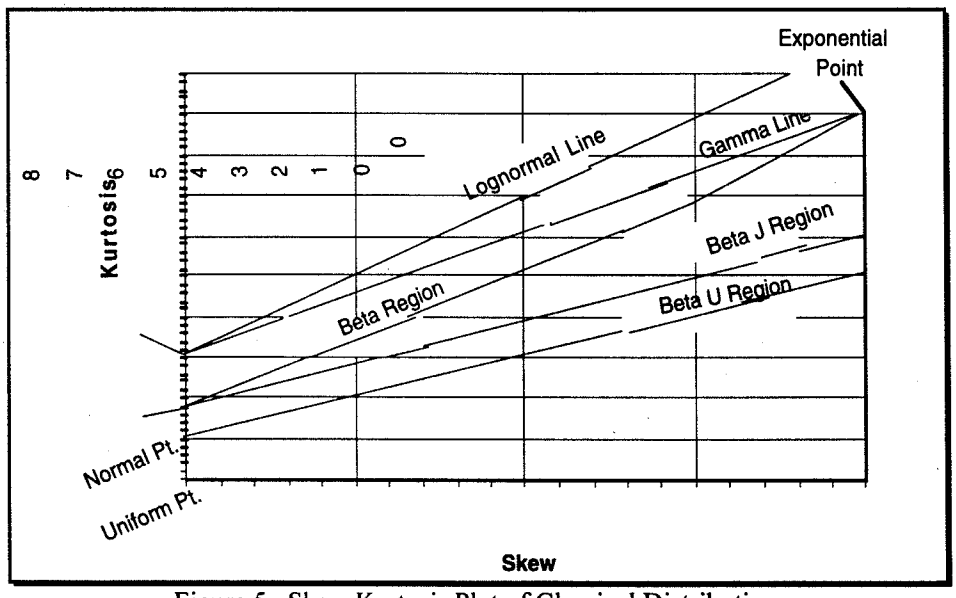


Figure 5. Skew Kurtosis Plot of Classical Distributions

Beta approximation errors

The beta distribution closely fits all the classical distributions. The sums and products of beta variates are also closely approximated by a beta distribution. The resulting beta distribution can easily be calculated. One does not have to resort to computationally intensive convolution for the sums of

random variates. All that is required is to calculate the minimum, maximum, mean and standard deviation statistics. The following paragraphs calculate the error in using the beta distribution to fit sums or products of uniform random variables. Then the error in approximating the normal distribution is calculated. Finally the mean and variance formulas for different types of calculations are presented.

Sums and products of uniform random variables

The uniform distribution is fit exactly by the beta distribution (a=b=1). The sum of two uniform variates is a triangular distribution with a discontinuity at the mode. In the case of discontinuities, the error in beta approximation is minimal. The cumulative probability distribution function of the sum of n independent uniform random variables between 0 and 1 is :

$$Pr[Sn \pounds x] = S\ (-1)^j\ (n!\ /\ (j!\ (n-j)!\ ))\ (x-j)^n\ /\ n!$$

Where:

x is a random variate between 0 and n

n is the number of uniform (0-1) random variates that are summed

Pr[Sn] is the cumulative probability of the sum of n independent uniform random variates

The summation (S) is over all integers $j < x$

The minimum, maximum, mean, and variance of the distribution of the sum distribution were used to fit a beta distribution. The difference between the fit beta distribution and the actual sum distribution was calculated over the entire range of distribution values. Figure 6 gives the results for sums and products of 2,3, or 4 uniform distributions.

The maximal error for sums ranges from 1.2 percent for 2 variates down to less than 1/2 percent for the sum of four uniform variates. The error rates in the tails of the distribution are far less. This is the area we are most interested in when making confidence statements.

Since the product of variates tend to be log normally distributed (outside the skew kurtosis beta region) the beta approximation for the product is not as good as the sum. Log transformation of the data before fitting would give a closer fit, but this would increase the calculation complexity.

Using the uniform distribution as an example, the cumulative distribution function of the product of two uniform distributions is:

$$Pr[x \pounds p] = p - p\ \ln(p)$$

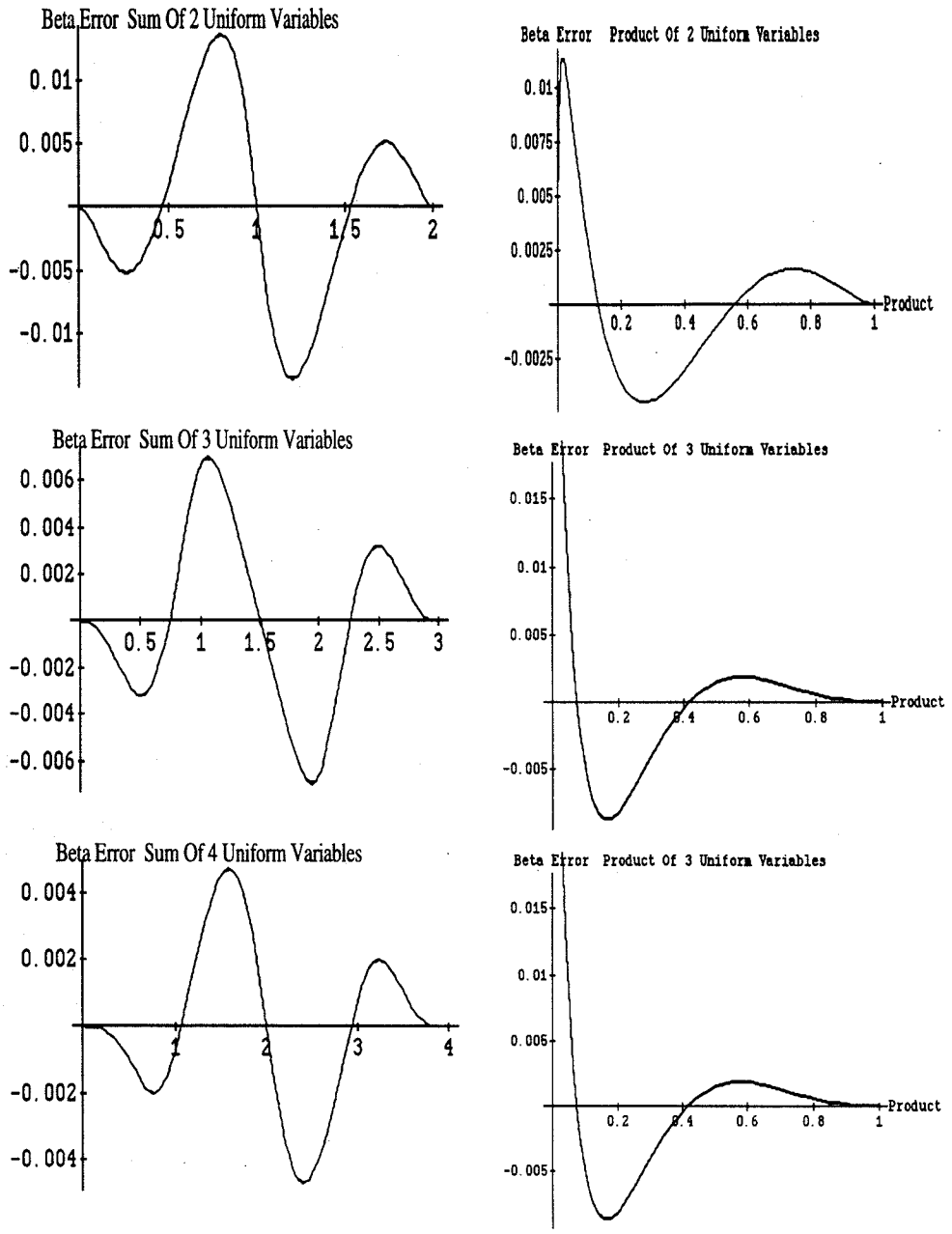


Figure 6. Beta Error Functions

The cumulative distribution function of the product of three random variates is:

$$Pr[x \pounds p] = p - p \ln(p) + p \ln(p)^2 / 2$$

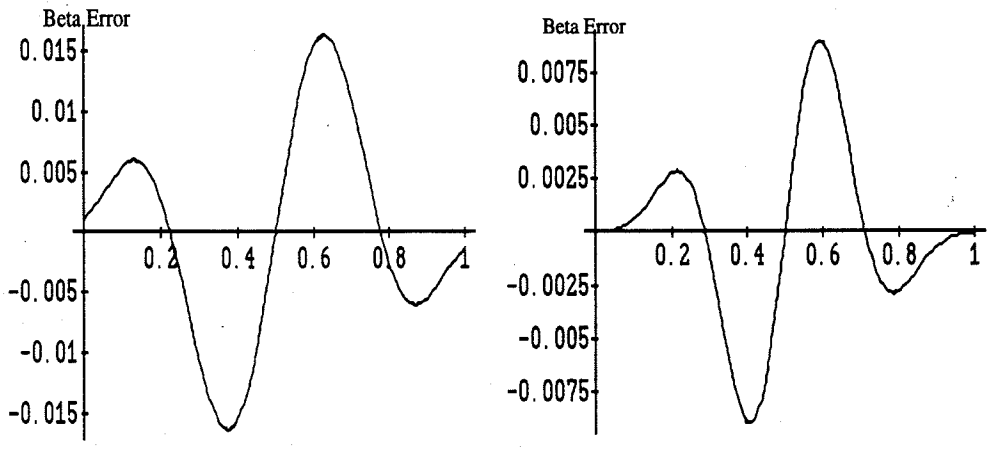
The cumulative distribution function of the product of four random variates is:

$$Pr[x \pounds p] = p - p \ln(p) + p \ln(p)^2 / 2 - p \ln(p)^3 / 6$$

The error increases as the number of terms in the product increases. However the error in the upper tail is very reasonable (<1/2 percent). Kotlarski and Jambunathan have investigated general conditions under which products of independent variables have a beta distribution. Springer uses Mellin transforms to calculate the products and quotients of beta random variables.

Beta approximation to normal distribution

The sum of independent variates approaches a normal distribution as more variates are summed (central limit theorem) The symmetric normal distribution tails extend from minus infinity to plus infinity. However the area in the tails beyond 3 or 4 standard deviations from the mean is minimal (0.27 percent for 3 standard deviations, 0.006 percent for 4 standard deviations). Figure 7 shows the error in approximating the normal distribution with a Beta distribution.



Range = 6 Standard Deviations (a = b = 4)

Range = 8 Standard Deviations (a = b =7 .5)

Figure 7. Beta Errors In Approximating Normal Distributions

The maximal error is 1.5 percent when the beta range is set to 6 standard deviations (+-3) or 0.75 percent when the beta range is set to 8 standard deviations (+-4). Note that again the error in the high interest tails of the distribution is far less.

Mean and variance statistics

In practice any calculation with beta variates usually yields beta variates. Thus if one encodes parameter uncertainty with a beta distribution, the results of a calculation with many uncertain parameters will also be beta distributed. This allows analytic solution for the resulting distribution without resorting to discrete simulation. This greatly reduces the calculation requirement and simplifies parametric sensitivity analysis.

In a process cost or performance calculation, operations must be performed on uncertain parameter values to determine expected results and uncertainty in results. Operations include: addition, subtraction, multiplication, power, polynomial function, general function. Using the Beta distribution we must keep track of the calculation minimum, maximum, mean (u) and variance (s^2) values. The variance of the sum or difference of two independent parameters is just the sum of their component variances. If the parameters are correlated the following equation is used:

$$s^2_{1+2} = s^2_1 + s^2_2 + 2\,s_{12}; \qquad s^2_{1-2} = s^2_1 + s^2_2 - 2\,s_{12}$$

Where s_{12} is the covariance of parameters 1 and 2. The variance of the product of two uncertain independent parameters 1 and 2 is given by the following formula:

$$s^2_{12} = s^2_1\, s^2_2 + u_1^{\,2}\, s^2_2 + u_2^{\,2}\, s^2_1$$

If parameter a is a constant with value a (minimum = mean = maximum = a, and variance = 0) then the variance of the product of parameters 1 and 2 is:

$$s^2_{12} = a^2\, s^2_2$$

In the case of division, the maximum result is the maximum quotient divided by the minimum divisor. Conversely the minimum result is the minimum quotient divided by the maximum divisor. The variance of the result and the corrected mean are obtained from:

$$s_{1/2} = s_1^2 / s_2^2 + u_1^2 s_1^2 + 1/(u_2^2 s_2^2)$$
$$u_{1/2} = (u_1/u_2)(1+(s_2/u_2)^2)$$

If y is a linear function of n input variables ($y=a_0+a_1x_1+a_2x_2+...+a_nx_n$) then the variance of y is given by:

$$\sum_{i=1}^{n} (\partial f/\partial x_i)^2 s^2_i + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} (\partial f/\partial x_i)(\partial f/\partial x_j) s_{ij}^2$$

If the input variables are independent the second term is ignored. If y is not a linear function of the input variables the above equation is only an approximation. In the case of a general single variable function (y=f(x)) with known derivatives, the variance of the result is approximated by the following:

$$(\partial y/\partial x)^2 s_x^2 + 1/2(\partial^2 y/\partial x^2)^2 s_x^4 + (\partial^3 y/\partial x^3)(\partial y/\partial x) s_x^4 + ...$$

Usually only the first term provides a reasonable approximation. Additional terms can be found in the Tukey report. Approximations for functions of several variables (z=f(x,y,...)) are derived based on the multidimensional Taylor series. Note that the mean y (y=f(x)) is not equal to the result of substituting the mean x value into the function, if the function is non-linear. The shift in mean value is given by:

$$1/2 (\partial^2 y/\partial x^2) s_x^2 + 1/8 (\partial^4 y/\partial x^4) s_x^4 + ...$$

For a general function, a reasonable first-cut approximation is to calculate the output for mean, and mean plus standard deviation inputs. The squared difference of these outputs is an approximation of the output variance.

Conclusions

In functional economic analysis we are interested in the relative future costs of alternative systems. There are uncertainties in process performance, cost estimates, investment requirements, and workload. Thus there is uncertainty in the discounted present value cost distribution for each alternative system. This paper presented an analytic technique to calculate the result distribution for alternatives from estimates of component uncertainties. The technique has wide application It considerably simplifies sensitivity analysis. It should be considered when the probability distribution of a result is desired rather than a single point estimate of the mean. Both analytic development costs, and computer execution costs are far less than in discrete event simulation. The price paid is some lack in modeling flexibility.

References

Whitt, W. 1989.*Planning Queueing Simulations,* Management Science, Vol 35, No.11, November 1989.

Johnson, N., Kotz,S. 1970. *Continuous Univariate Distributions-2,* John Wiley & Sons, New York.

Wilks, S. 1962 *Mathematical Statistics,* New York, John Wiley & Sons,

New York.
Tukey,J. *The Propagation of Errors, Fluctuations, and Tolerances Basic Generalized Formulas,* AD155082, Princeton University, Princeton,
New Jersey.
Koltarski,I.1962 *On Groups of N Independent Random Variables Whose Product Follows The Beta Distribution,* Warsaw, COLLOQUIUM MATHEMATICUM, vol IX, FASC.2,pp325-332,Warsaw, Poland.
Jambunathan, M.1954.*Some Properties Of Beta And Gamma Distributions,* The Annals of Mathematical Statistics, Vol.25,pp 401-405,1954.
Springer,M.1979. *The Algebra of Random Variables,* John Wiley & Sons, New York.
Seiler,F. 1987. *Error Propagation for Large Errors,* Risk Analysis, Vol.7, No.4,pp 509-518.