DISTRIBUTION SELECTION

INTRODUCTION

When quantifying the effects of uncertainty on a system there is a dilemma that many face: the selection of most appropriate distribution for each of the input variables. This article will cover a collection of principles and various methods that can be used when deciding upon the most suitable distribution for an input variable. Some attention will be given to the application of these methods; however, it is ultimately up to you to determine the best application for each situation that you encounter.

ENTROPY

Entropy was first used as a scientific term when it was used to describe the way in which the quality of energy in a system will tend to decrease. Since then its use has moved into other fields, including probability. Within the context of probability entropy is used a measure of the uncertainty regarding the value or state of a system. The less certain we are of the state or value a system will take, the greater the entropy of that system.

The expression for entropy H is as shown in the formula below.

$$H = -\sum_{i=1}^{n} p_i \ln p_i$$

Where:
n is the number of states or values possible
p_i is the probability of the i^{th} possible state occurring

If there were only one possible state then the probability of that state would have a value of unity, and the above expression would be equal to zero: no entropy or uncertainty. When you wish to select a distribution for an input variable you should maximize the entropy given the knowledge you have. By doing this, you can ensure that you do not select a distribution that implicitly places unconfirmed constraints upon the random variability (or uncertainty). Thus, by selecting a distribution that maximizes the entropy you select an unbiased distribution that reflects the knowledge you actually have. How is such a distribution identified?

The above definition of entropy can be applied to the case where p is replaced by a continuous Probability Density Function (PDF) $f(x)$. The general form of $f(x)$ that maximizes the entropy can then be found. The general form is as shown below.

$$f(x) = e^{-\lambda_0 - \sum_{i=1}^{n} \lambda_i}$$

Where:
n is the number moments that are known
λ_0 and λ_i are constant to be solved for

With the above expression for the PDF and n moments known, you can develop an equation for a distribution that will capture the information you have while still maximizing uncertainty. To actually do this can be difficult; the solution needs to be found numerically due to the nature of the function. However, others have investigated this and reported on cases where the maximum entropy distribution is also a standard distribution. Some of these cases are summarized in the table below. This is based on the information found in the PhD thesis *Charcaterisation of the variability of design parameters* by Maxine Nelson. The thesis is held at the Swinburne University of Technology Library, and might be difficult for you to get a hold of. Therefore, if you would like more information on developing maximum entropy distributions you might like to take a look at *Maximum Entropy Models in Science and Engineering* by Jagat Narain Kapur or *Rational descriptions decisions and designs* by Myron Tribus.

Information	**Maximum Entropy Distribution**
Upper and lower limits	Uniform
Mean and standard deviation	Normal
Mean, standard deviation and limits	Normal: truncated
Mean and limits	Exponential: truncated & shifted

The above table, and the maximum entropy method, is essentially based on the information that can be extracted form the data that you have at hand. If you have data then this approach is ideal. However, you may not always have data, and you will need to use other sources of information when deciding upon the best distribution for an input variable.

THE CENTRAL LIMIT THEOREM

The central limit theorem deals with the distribution that is produced when distributions are added or multiplied together. The most commonly noted form is the additive form. The additive central limit theorem can be stated as follows:

'The sum of a large number of independent but not necessarily identically distributed random variables is approximately normal provided that no one random variable contributes appreciably to the sum; that is, no term dominates the others.' From *The probability tutoring book* by Carol Ash, 1992.

An extension to the central limit theorem is that for multiplication. The multiplicative form of the central limit theorem says that the distribution of a product of a series of random variables tends toward a Lognormal distribution.

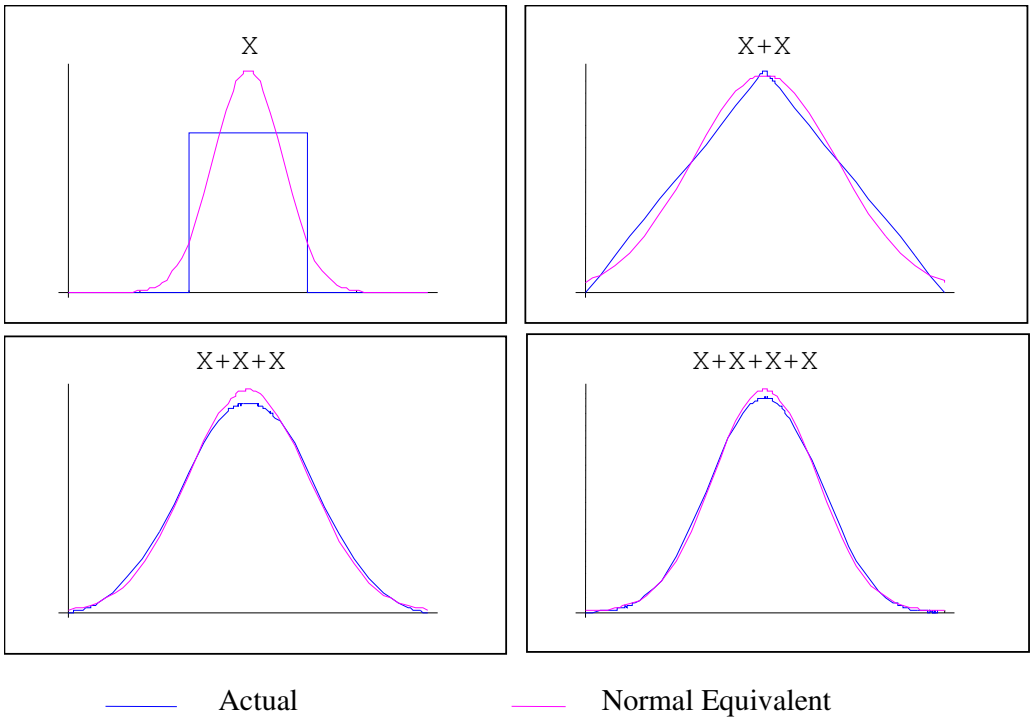
Consider an input variable or parameter that is associated with a phenomenon that is additive or multiplicative in nature. That variable or parameter will most likely have a distribution that closely resembles a Normal or Lognormal distribution respectively. Therefore, if you can determine that the phenomenon associated with a random variable or parameter is either multiplicative or additive, then you are able to select a distribution that will likely be a good approximation of the actual distribution for that variable or parameter.

Examples:

> The profit of a business or a company division is the result of the addition of the various sources of income minus the various costs. If each source of income and cost is independently random, then the central limit theorem would suggest that the profit will be closely approximated by a Normal distribution. This is regardless of the distribution that each income source or cost has. Note that subtraction is simply the addition of a negative number (or distribution).
>
> Many natural phenomena can be described by the multiplication of various variables. For example, the wear exhibited by a surface exposed to rubbing is proportional to the pressure of contact, multiplied by the displacement, multiplied by the wear coefficient, divided by the hardness.

If each of these were independently random, then according to the central limit theorem the evident wear would have a distribution approximating a Lognormal distribution.

As stated above, when independent random variables are added together, the resulting distribution tends toward a Normal distribution. Also, multiplication results in a tendency toward a Lognormal distribution. Still a question remains: how strong is the tendency, or how quickly are the effects evident? The figure below shows the resultant distribution from the addition of independent, identical uniformly distributed random variables. The Uniform distribution was used because of its significant difference from the Normal distribution. For each case, the resultant distribution (in pink) is shown with the equivalent (same mean and standard deviation) Normal distribution (in blue). The number of Uniform distributions added together is indicated by the expression above each plot.



The figure above demonstrates that the sum of two uniformly distributed variables is a Triangular distribution and that the Normal equivalent is similar. When three uniformly distributed variables are added, the difference from the

Normal equivalent is further reduced; and when four are added together, the difference is difficult to distinguish by sight.

There is considerable difference between a Normal distribution and a Uniform distribution. Further, most distributions encountered in practice tend to have some kind of a hump in the middle, with a tail at either end. This would suggest that typical distributions in the real world are closer to a Normal than a Uniform distribution. This would in turn mean that adding real life random variables together would result in a Normal like distribution with fewer additions than shown in the above figure.

The effectiveness of the central limit theorem and the tendency for most distribution to be more Normal like than Uniform like allow us to draw a conclusion:

> The addition of a small number of random variables can be well represented by a Normal distribution, and the multiplication of a small number of random variables can be well represented by a Lognormal distribution.

Therefore, the central limit theorem can be applied to many situations, even if the number of contributing random variables cannot be determined. What is important is that you can determine if the underling phenomenon is multiplicative or additive.

DIMENSIONAL CONSTRAINTS

Model constraints place restrictions on the form that a model can take. If we understand these constraints we are able to gain insight into how random variability will propagate through a system. By understanding how random variability will be propagated we can ascertain how much information we must specify for the distribution of the input variables. In this section we will consider a particular constraint: dimensional homogeneity.

Dimensional homogeneity demands that the dimensions (or units) on one side of an expression be the same as the dimensions on the other side. For example, the speed of a car (in miles per hours) is found by taking the distance traveled (in miles) and dividing that by the time taken (in hours). It would not make sense to

subtract the time from the distance to get speed. All models must satisfy the constraint of dimensional homogeneity, and thus their form is also constrained.

To explain this, we will use a simple example: a clutch. When considering a clutch we are interested in the torque that it can transmit. The torque transmitted (in Newton metres Nm) by a new clutch can be calculated using the formula below.

$$T = \frac{2Ff\left(r_o^{\,3} - r_i^{\,3}\right)}{3\left(r_o^{\,2} - r_i^{\,2}\right)}$$

Where:
F is clamping force (in Newtons N)
f is the coefficient of friction (friction can be treated as having no dimensions)
r_o is the outer radius (in meters m)
r_i is the inner radius (in meters m)

Dimensional homogeneity demands that the dimensions on one side of the equal sign must be the same as on the other. In this case the dimensions on the left hand side are Newton meters (Nm); therefore, the units on the right hand side must also be equal to Newton meters. Without going into too much detail on the topic of dimensional analysis, the requirement of the dimensions on both being equal means we can arrange an equation so that it consists of dimensional groups. To demonstrate, the equation for the torque above can be algebraically manipulated into the following non-dimensionalized equation/model.

$$T = \frac{2}{3} \times F\ r_o\ f \times \frac{\left(1 - \left(\frac{r_i}{r_o}\right)^3\right)}{\left(1 - \left(\frac{r_i}{r_o}\right)^2\right)}$$

From inspection, the first group, r_i/r_o, is dimensionless and the other group, $F\ r_o f$, has the dimensions of Newton meters (just like the torque), and the equation is dimensionally balanced. These two groups are a collection of variables that are multiplied together, and from the central limit theorem as a group they will likely be well represented by a Lognormal distribution. Because we know that each group will be well represented by a Lognormal distribution it doesn't really matter what

distribution is allocated to each input variable as long as we have a good measure of the average and the variability: the mean and the standard deviation. Therefore, we can choose any distribution for the input variables as long as we get the mean and the standard deviation right and still be confident that we will be able to predict a representative distribution for the output. Or can we? This might seem too simple, and in fact it is.

In the above equation we cannot assume that there is independence between the two groups of variables (r_i / r_o and $F\ r_o f,$); they are both a function of r_o. This means that if we are going to take advantage of the central limit theorem, we will need to account for this interdependency. Experience has found that the full distribution must be found for the 'shared' variables in a non-dimensionalized model, if the maximum accuracy in the output model is desired. However, it has also been found that changes to the distribution of the shared variable often have a minor effect (frequently none) upon the output distribution.

From the above it can be concluded that for models that produce relatively large groupings of numbers after being non-dimensionalized the distribution for each input variable is often unimportant and only the shared variables require extra attention. This can reduce your workload when defining the input distribution considerably. However, you will need to become familiar with dimensional techniques to do this. Still, it is worth doing this because of the time that you can save when selecting distributions for the input variables. If you would like to improve you skills in this area, an excellent text on the topic of dimensional analysis is *Applied Dimensional Analysis and Modeling* by Thomas Szirtes. A question remains: what other constraints might we take advantage of if a model does not produce large dimensional groups?

While all models can be manipulated so that they are made up of functions of dimensional groups, sometimes the groups themselves are many in number and each is small in size (perhaps only 2 constituent variables at most). An example would be the cash balance at the end of a period in a business model. Everything will have the units of dollars, and non-dimensionalization will have little effect on the form of the equation. This would limit the effects of the central limit theorem, and you might feel that you need to ascertain the most representative distribution for each input variable.

However, such models usually allow for consideration that can still enable you to determine those variables that require little effort when defining their distributions.

Models that do not allow for the formation of large dimensional groups after non-dimensionalization often exhibit a similar quality. Along with the cash balance example in the previous paragraph, most business financial models, tolerance stacks and conservation equations (mass and energy for example) do not lend themselves to non-dimensionalization. The reason for this is that all of these situations deal with variables with same units, and produce models that are the summation of those variables. Models that are essentially a series of sums do not lend themselves to non-dimensionalization, but we can still take advantage of the central limit theorem. Because these models are a summation of random variables, it would be expected that the output would be well represented by a Normal distribution regardless of the actual distribution of each input variable. Therefore, in such cases you would really only need to be certain of the mean and the standard deviation for each input variable.

In this section we have seen that by considering the form of our model and the central limit theorem we are able to ascertain how much information we must accurately specify when defining the distribution for an input variable. For models that lend themselves to non-dimensionalization, we really only need to specify the complete distribution for those input variable that are shared between dimensional groups. For other variables we need only specify the mean and the standard deviation, and any distribution type can be chosen. For those models that do not lend themselves to non-dimensionalization, we should determine if these models are additive. If so, then we know the output will be close to a Normal distribution, and we only need to specify the mean and the standard deviation for each input variable; any distribution type can then be chosen.

A note for those lacking dimensional analysis skills

> It might be that you do not have skills in dimensional analysis and you do not have the time to develop them. An alternative to developing and applying these skills is to change the distribution for each input variable from one type to another that is significantly different and observing the result. If the change causes a negligible change in your predictions then the

variable is not a shared one, and the distribution type is unimportant. If there is a difference, then the variable should be treated as a shared one, and the full distribution should be found.

ESTIMATING MOMENTS

The two most important properties of any distribution are the mean and variance (or standard deviation). Before you can start to model the variability of a system output, you need to understand the nature of the variability for each input variable. A good place to start is the tolerance that is usually specified for the variable of interest or the maximum and minimum values that you think are possible. This is the focus of this section.

Typically, the tolerance range is assumed to be equal to about 6 standard deviations when the Normal distribution is being used. This means that 99.7% of the time we expect the value of the variable will be within the tolerance range. This is a rough and ready approach and at times we may expect there to be a different proportion within range. The following are some other possible settings:

Set the standard deviation equal to 1/2 of the tolerance range if you are expecting 68% to be within tolerance.
Set the standard deviation equal to 1/4 of the tolerance range if you are expecting 95% to be within tolerance.
Set the standard deviation set equal to 1/6 of the tolerance range if you are expecting 99.7% to be within tolerance.

Other researchers have extended upon this idea to include the amount of data you have or the type of manufacturer/service provider you are using. Haugen says in his paper *Modern Statistical Materials Selection* that estimates should be based on the range of values measured from collected data and recommends the following rules:

From about 4 samples set the standard deviation equal to half the tolerance range. Set the mean to the mid point between the maximum and minimum
From about 25 samples set the standard deviation equal to 1/4 of the tolerance range. As above for the mean.
From about 500 samples set the standard deviation equal to 1/6 of the tolerance range. As above for the mean.

Shooman, in his book *Probabilistic reliability – An Engineering Approach*, uses the reputation of the manufacturer (or service provider) to estimate the standard deviation and uses the following rules:

> If the manufacturer is little known or inexperienced or if you are in early development, set the standard deviation equal to 1/2 the tolerance range.
> If the manufacturer is military, reputable or experienced set the standard deviation equal to 1/6 of the tolerance range.

The above is only a collection of general rules to provide you with guidance if you need it. There may very well be times when you have reason to believe that the moments should be determined in a different manner.

CLOSING

We have seen from the above that we can:

- Choose a distribution that is least bias given our understanding of the situation by utilizing maximum entropy distributions
- Use the central limit theorem to determine what distribution can be expected when we know the basic nature of the respective underlying phenomena
- Use dimensional constraints to determine which input variables require the most comprehensive determination of their distributions
- Use a collection of heuristics to determine the most appropriate mean and standard deviation from sample data.

However, while these concepts help us determine appropriate distribution with greater ease, they do not actually determine them for us. They require us to apply our understanding of each situation. Therefore, it must always be remembered that it is ultimately up to us to choose the input distribution for each input variable. The above concepts are only tools, which we must be certain we have used correctly.