Random Numbers and Simulation

To develop an understanding of the set of possible outcomes for a particular scenario, we can set up and observe the actual events of interest or we can use a computer to model or simulate the events. Simulation has several advantages. First, a computer can be used to try out many thousands of different possibilities in a relatively short period of time. Second, it is often expensive, impractical, or even hazardous to actually try out a representative range of all the possibilities.

Simulation modeling involves the generation of theoretically random events based on one or more simultaneous, and possibly interacting, probability distributions. Each event is generated in a very methodical and structured fashion. Everything follows a set of rules. Because of these characteristics, TK is an excellent tool for building simulation into your models, whether you are simply experimenting with adding some random "noise" to a variable or modeling a complete phenomenon.

The Statistics section of the TK Library contains distribution functions for computing probabilities. TK also includes built-in functions for generating random numbers from these distributions.

Below are descriptions of the random number generators included in TK.

RAND

The RAND function generates a random number between 0 and 1 from a uniform distribution. It includes an option to specify a different minimum and maximum value, RAND(min,max), which is the same as the expression RAND()*(max-min) + min.

RANDBERN

You may wish to simulate a situation in which a dichotomous variable is involved. This variable may only take on two values, each with some specified probability which is independent of the previous or the next event. This is known as a Bernoulli process. You can use this to simulate situations where things are reported as either successful or failed, on or off, open or closed, etc., with certain probabilities.

The RANDBERN function generates a Bernoulli number – 0 or 1 -- which may be linked with some physical or economic process. RANDBERN requires a single input – the probability of the event occurring. For example, you might have a rule such as

```
flow1 = RANDBERN(0.2)*valve1(D,T,density,viscosity)
```

which would simulate the operation of a valve in an open or closed (RANDBERN = 1 or 0) position.

Example:

Suppose that on your drive to work every day, you pass through a particular intersection where the light is green for you only 20% of the time. Simulate your next 50 trips through the intersection.

Use the List Fill Command and choose Bernoulli distribution. Enter trips as the list name, 0.2 as the probability and 50 for the number of values. TK generates 50 numbers in the list trips . The 0's indicate that you stopped, 1's indicate that you had a green light.

RANDBETA

The standard beta distribution (with limits of 0 and 1) is commonly used to model variation in the proportion or percentage of a quantity occurring in different samples, such as the proportion of a 24-hour day that a machine is operational or the proportion of a certain element in a chemical compound. The beta distribution is easily extended over an arbitrary domain to model a broader range of situations.

The beta distribution is defined by the formulas

$$y = \frac{x - c}{d - c}$$

$$p = \frac{\Gamma(a + b)}{\Gamma(a) \cdot \Gamma(b)} \cdot y^{(a - 1)} \cdot (1 - y)^{(b - 1)}$$

The first equation simply rescales x to be between 0 and 1, with c and d defined as the minimum and maximum possible values of the independent variable x. Variables a and b are the shape parameters. If both a>1 and b>1, the probability is 0 at both x=c and x=d and rises to a unique maximum between c and d. If both a<1 and b<1, the probability decreases from infinity at x = c and x = d to a unique minimum. If a>1 while b<1 the probability increases from 0 to infinity as x increases, and the reverse occurs if a<1 and b>1. For a=b=1, the distribution is uniform from c to d. The mean and variance are given by the equations

$$\text{mean} = \frac{a}{a + b}$$

$$\text{variance} = \frac{a \cdot b}{(a + b)^{2} \cdot (a + b + 1)}$$

The RANDBETA function requires four inputs – the shape parameters a and b, and the limits c and d.

Example:

A 20 in. bar is clamped in a fixed position at each end. The probability of the bar snapping at any point along the bar is given by a beta distribution with parameters a = b = 3. Simulate 1000 snaps.

Use the List Fill Command and choose the Beta Distribution with the indicated parameters. The resulting values will be the locations of the 1000 snaps.

RANDBETA can also be used for generating values from Snedecor's F distribution, which is used extensively in the statistical testing of variances. Generate a random beta value using x = RANDBETA(n1/2,n2/2,0,1). Then F = (n2/n1)*(x/(1-x)).

Values from the Student's t distribution can also be generated using RANDBETA, since t=sqrt(F), assuming n1=1. And because a t distribution is symmetric, the sign must be randomly applied. Here is the required formula for producing t, assuming x = RANDBETA(1,n2/2,0,1)

t = (RANDBERN(0.5)*-1)*SQRT(n2*x/(1-x))

RANDBIN

You may wish to simulate a situation in which n independent Bernoulli trials, each with identical probability p, are grouped together to form a single Binomial trial. The value of a single binomial random number, then, is the sum of results of the individual Bernoulli trials. The RANDBIN function generates a random binomial value. The mean of a binomial distribution is defined as n*p and the variance is n*p*(1-p).

The RANDBIN function requires that you provide it with two inputs -- the probability of success on any given trial and the number of trials per event. For example, you might have a rule such as Y = RANDBIN(0.35,4), where Y would be the number of successes in 4 attempts for a situation where the probability of success on any given trial is 35%.

Example:

Suppose you are simulating giving a 10 item test to each of 100 subjects. Items are either answered correctly or incorrectly so they may be considered as Bernoulli trials. Each subject's test score is their number of correct answers. If we assume that each item is equally difficult for each subject (each item has an equal probability of correct solution), then the distribution of 20 total scores follows a binomial distribution. Let's assume that each item has a 70% chance of being answered correctly.

Use the List Fill Command, choosing the Binomial Distribution. Enter "scores" as the name of the list to store the values. Enter 0.7 as the probability and 10 as the number of trials per event. Ask for 100 values to be generated. The resulting list will contain simulated test scores for each of the 100 students.

RANDEXP

The RANDEXP function generates values from an exponential distribution. Exponential distributions are commonly used for modeling the times at which objects or people arrive at a particular place for some sort of service. One input is required -- the mean of the distribution. For an exponential distribution, the mean and standard deviation have the same value.

Example:

Suppose you wanted to simulate the response time at a certain on-line computer terminal (the elapsed time between the end of a user's inquiry and the beginning of the system's response to that inquiry) and you assume that this time has an exponential distribution with a mean of 5 seconds.

Use the List Fill Command and choose an Exponential Distribution with mean 5 and ask for 1000 values in the list named response. The resulting list contains 1000 exponentially distributed numbers.

RANDGEO

Suppose that a group of independent events, each of which results in a success with probability p, are continually performed until a success occurs. The number of events necessary has a geometric distribution. The RANDGEO function generates random values from a geometric distribution. One input is required when using the RANDGEO function – the probability of success on any event. The mean of a geometric distribution is the inverse of this probability, 1/p. The variance is defined as (1-p)/p^2.

Example:

Suppose a factory uses a machine which is made up of 8 identical components running in series, each with an identical probability of breaking down on any given day. If one component breaks down, the whole machine is down to repair the component. Assuming that any single component breaks down with probability .015 on any given day, simulate 100 days of activity for the machine.

Use the List Fill Command and select Geometric Distribution. Enter X as the list name, 0.015 as the probability, and ask for 100 values. TK generates the 100 list elements representing the 100 days. Each time a value of 9 or greater appears in the list, your simulated machine was ok for that day.

RANDGAMMA

The gamma distribution yields a wide variety of skewed shapes, depending on the values of the two parameters. The equation for the gamma probability distribution is

$$\text{if } x => 0 \ \text{ then } p = \frac{x^{(a-1)} - e^{\left[\frac{x}{b}\right]}}{b^{a} \bullet \Gamma(a)}$$

The mean is defined as a*b and the variance is a*b^2. The minimum value is 0. The distribution can be shifted by adding the desired minimum value.

Example:

Suppose that the reaction time of an individual to a certain stimulus has a gamma distribution with parameters a = 2 and b = 2. The expression RANDGAMMA(2,2) will generate a random reaction time.

RANDHGEO

The hypergeometric distribution is based on experiments satisfying the following conditions:

1. The finite population to be sampled consists of N individuals, objects, or elements.
2. Each object can be characterized as a success or failure and there are k successes in the population.
3. A sample of n objects is drawn in such a way that each subset of size n is equally likely to be chosen. That is, the n objects are drawn as a set, without replacement.

The random variable of interest, X, is the number of successes in the sample. The mean of the distribution is m = n*k/N. The standard deviation is sd = m*(1-k/N)*(N-1)/(n-1).

Example:

A standard deck of playing cards (jokers removed) includes 13 cards from each suit. Generate a random number representing the number of hearts dealt to an individual in a five-card hand. The TK expression would be

```
RANDHGEO(52,5,13)
```

RANDNBIN

The negative binomial distribution is based on experiments satisfying the following conditions:

1. An experiment consists of a sequence of independent trials, with each trial resulting in a success or failure.
2. The probability of success, p, is constant from trial to trial.
3. The experiment continues until a total of r successes have been observed, where r is a specified positive value. That is, r may be fractional.

The random variable of interest, X, is the number of failures which precede the rth success. The mean of the distribution is (1-p)*r/p. The variance is (1-p)*r/(p^2). RANDNBIN requires two inputs, p and r.

Example:

A pediatrician wishes to recruit couple, each of whom is expecting their first child, to participate in a new natural childbirth regimen. Assuming that the probability is 0.2 that a couple agrees to participate, generate a random number indicating the number of couples asked before five are found who agree to participate. The TK expression is

```
RANDNBIN(0.2,5)
```

RANDNORM

This function is very useful for generating random noise assumed to be distributed normally. It is also useful for generating random values for many other physically measured properties. RANDNORM requires two inputs -- the mean and standard deviation of the distribution. For example, if you have a variable t representing the thickness of a part, and you assume that t is normally distributed with mean 0.125 and standard deviation 0.002, you can simulate the thickness using the rule

```
t = RANDNORM(0.125,0.002)
```

RANDNORM can also be accessed through the List Fill Command.

RANDPOIS

The Poisson Distribution is commonly used for modeling events which occur in series over a specified period of time. The distribution values are the number of objects in the system at any given time. The mean of the distribution is required. The mean and standard deviation are equal for this distribution.

Example:

Suppose you are interested in simulating a traffic intersection controlled by stop signs in all four directions. During a period of the day between 6:30 and 8:00 a.m. northbound traffic is backed up at the intersection with an average of 9 vehicles waiting at any given time. Generate 100 simulated observations of the number of vehicles waiting in the northbound lane.

Use the List Fill Command with a Poisson Distribution. Use 9 as the mean value and “vn” as the name of the list. Ask for 100 values. The resulting vn list contains the simulated number of vehicles observed waiting in the northbound lane on 100 consecutive days.

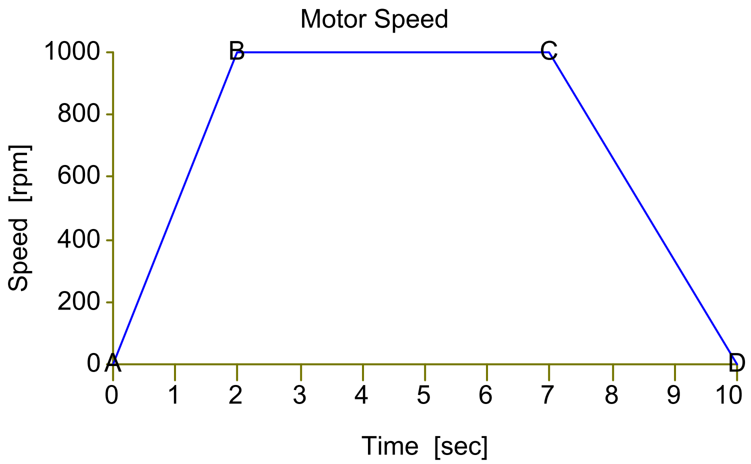
RANDTRAP

There are many instances where a probability distribution is dictated by a mechanical process. In some of these cases, a trapezoid can be used. Other geometric shapes which are often used include a rectangle (representing a uniform distribution) and a triangle (see RANDTRI below). A trapezoid can be used to simulate situations where a velocity ramps up, levels off, then slows down. A trapezoid might also be used in modeling manufacturing tolerances.

The inputs to the RANDTRAP function are the four x-axis points of the trapezoid. They must be in sequence. That is, the projection of the “top” of the trapezoid must lie within the “bottom”.

Example:

Suppose you need to simulate the operating speed of a motor in a manufacturing process. At any random point in time, the speed of the motor is defined by the following trapezoid.



If the motor operates for 10 seconds and then idles for 10 seconds, the following TK rule can be used to simulate the operation of the motor.

X = RANDBERN(0.5)*RANDTRAP(0,2,7,10)

The RANDBERN function represents the fact that the motor is operating half of the time. The RANDTRAP function returns a value between 0 and 10.

RANDTRI
See RANDTRAP for details on generating random numbers from geometric shapes. The concept is the same. The three inputs required by RANDTRI are the x-axis coordinates of the triangle.

RANDWEIB
The Weibull distribution can be used for a variety of applications but is most frequently applied in the study of reliability. Weibull distributions can take on a variety of shapes as the three parameters vary. Here is the equation for the cumulative probability.

$$\text{if } x > c \ \text{ then } CDF = 1 - e^{\left[-a \bullet (x-c)^{b}\right]} \ \text{ else } CDF = 0$$

The mean and variance are defined by the equations below.

$$\text{mean} = c + a^{\left[\frac{-1}{b}\right]} \bullet \Gamma\left[1 + \frac{1}{b}\right]$$

$$\text{var} = a^{\left[\frac{-2}{b}\right]} \bullet \Gamma\left[1 + \frac{2}{b}\right] - (\text{mean} - c)^{2}$$

The RANDWEIB function allows you to simulate Weibull values. It requires three inputs – the a, b, and c parameters of the distribution. The a parameter is the scale parameter which stretches or compresses the distribution in the x direction. The b parameter interacts with the a parameter in determining the height and shape. The c parameter is simply the lower limit of the distribution. Note the use of the built-in GAMMA function in the evaluation of the mean and variance.

If a value of c is assumed, it is possible to use TK to solve for a and b, given values for the mean and variance. Iteration is required to solve the two simultaneous equations above.

Example:
The lifetime X (in hundreds of hours) of a certain type of light bulb is known to have a Weibull distribution with parameters a = 2, b = 3, c = 0. Simulate the life of 10000 of these light bulbs.

Use the List Fill Command and select the Weibull Distribution. Input the list name as bulbs and enter the a, b, and c parameters along with 10000 for the list size. The resulting list contains the simulated lifespans for 10000 bulbs.

Estimating Distribution Parameters

In the previous section we saw how to generate random numbers from various distributions. In each case, we assumed that we already knew the parameters defining the shapes of the distributions. If you have simulated inputs for several variables in a TK model, it may not be clear as to what distributions the resulting outputs will have. This section describes techniques which can be used to estimate the parameters given sample data or simulation results.

Moment Matching

Given a list of values, compute the mean and variance of the sample. For any particular distribution, we know the formulas relating the mean and variance to the unknown parameters. These formulas can be solved using the sample data information as inputs.

Example:
An experiment was conducted measuring the efficiency of a rocket engine using a new fuel mixture. A sample of 50 of the engines were tested and their performance was

measured as the number of seconds before the fuel ran out. The resulting 50 data points had a sample mean of 2.35 and a variance of 0.62.

From previous experience, you consider that the data may have come from a gamma distribution. We know that a gamma distribution has two parameters, a and b, which define the shape. We also know that a and b are related to the mean and variance by the formulas

mean = a*b   variance = a*b^2

or equivalently,

b = variance/mean and a = mean/b

Given the sample mean and variance, we can now estimate that b=.46 and a=2.94 . This allows us to use the gammaPDF and gammaCDF functions from the TK Library to compute probabilities for any specified values or to determine the value corresponding with a particular probability.

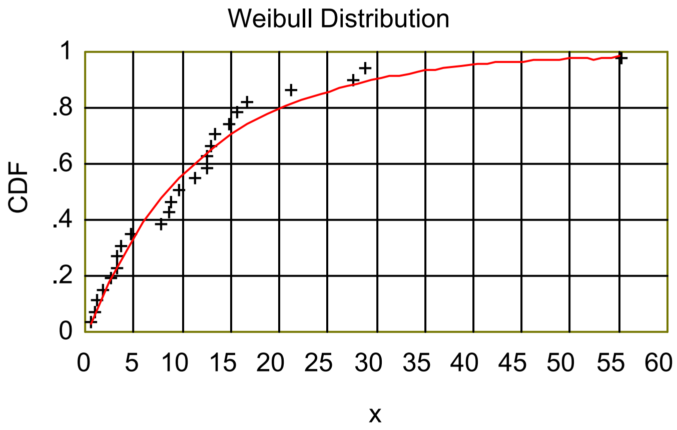
The moment matching technique can be used for all the distributions discussed earlier. The simplest case is the exponential distribution which requires only a single moment, the mean. The Weibull distribution requires that you provide your own estimate of the minimum possible value. The other two parameters can then be estimated by solving two simultaneous equations involving the mean and variance. TK is well equipped to help.

Least-Squares Estimation

The Statistics section of the TK Library includes several tools in the Curve Fitting folder that can be used to evaluate the results of simulation studies. Each output list can be processed individually. The simplest approach is to launch a second TK and load the library tool. Then copy the output list from the first TK into the list sheet of the second TK.

Here are the results from a sample run of the Weibull fitting tool.

St	Input	Name	Output	Unit	Comment
					Analysis of Weibull Distributions
	'sample	data			Input listname of data values
					Summary Statistics
		n	25		Sample size
		MEAN	12.02		Mean
		VAR	141.948333		Variance
		SD	11.9142072		Standard Deviation
					Weibull Distribution
					CDF = 1 - exp(-a*(x-c)^b) (x>c)
		a	.097622929		Estimated a parameter
		b	.93603907		Estimated b parameter
		c	.330036435		Estimated minimum value (c parameter)
		r	.992116046		correlation coefficient
		test	5.68863436		goodness of fit test statistic
		p	.458957023		chisquare probability (6 d.f.)



Note that the Method of Moments estimates for a and b would be 0.09031 and 0.98123 based on the values of the mean and variance. Those are similar to the least squares estimates reported above.

See the simulation example in the Engineering and Science section of the TK Library.