Quick Review Guide to Probability and @RISK

Basic glossary ("frequentist" philosophy)

Event: a subset of the possible outcomes of your "experiment".

Probability of an event: if you were to average over a very large number of replicas of the experiment, the fraction of the time that the experiment outcome is in the event.

Random variable: a number whose value depends on the outcome of the experiment.

Expected value of a random variable: if you were to average over a very large number of replicas of the experiment, the average value you would observe for the random variable. Also called the "mean" or average. Usually denoted E[X], where X is the random variable.

Independence: when knowing about one event or random variable doesn't give you any extra information about another event or random variable. For example, if I take two coins out of my pocket, flip one, and get heads, that doesn't give you any information about whether the second one will come up heads.

Variance and standard deviation: measures of how much a random variable varies; technically, the variance of X is $E[X^2] - E[X]^2$, and the standard deviation is the square root of the variance.

Common types of random variables you can generate with @RISK

Binomial distribution: RISKBINOMIAL(*n*, *p*)

Interpretation: Number of "heads" counted in n flips of a coin, with each flip having a chance p of being "heads." The possible values are 0, 1, ... , p. The expected value is np.

Use: When each of a known number n of items has an independent chance p of having some property. We wish to count the subset having the property. Also, if you want a variable with a chance p of being 1 and a chance $1 - p$ of being 0, you can use RISKBINOMIAL(1, p).

Arbitrary "discrete" distribution: RISKDISCRETE(*list1*, *list2*)

Interpretation: Will produce each value in *list1* with probability equal to the corresponding element of *list2*.

Use: Generating random values from a table giving values and probabilities. The two lists must have the same length. The values in *list2* should add up to 1. If, not @RISK scales them proportionally so they do.

Exponential distribution: RISKEXPON(*r*)

Interpretation/Use: The time between successive arrivals of customers at a service facility such as a store, telephone call center, etc. The average time between arrivals is 1/*r*.

Normal distribution: RISKNORMAL(*m*, *s*)

Interpretation: The classic "bell curve" with mean value *m* and standard deviation *s*.

Use: Random variables that are sums or averages of large numbers of other, independent random variables; see the discussion of the central limit theorem below.

Poisson distribution: RISKPOISSON(*m*)

Interpretation/Use: Number of customers arriving at a facility in a given span of time, when the average number is *m*. The idea is that you have a very large pool of customers, each of whom has a very small chance of requesting service in a given time period. Mathematically, the distribution is obtained by taking the limit of Binomial(N, m/N) as N goes to infinity. The value returned can, in theory, be any nonnegative integer. However, the average value *m* need not be an integer.

Uniform distribution: RISKUNIFORM(*m*, *M*)

Interpretation: A random quantity equally likely to take a value the computer can represent between *m* and *M*, specifically any x with $m \leq x < M$.

The central limit theorem

Then central limit theorem is the foundation of most of the statistics that you've learned in other courses. Omitting the technical details, it states that if you add up (or average) a "large" number N of independent random variables, then, no matter what the distributions of the individual variables look like, the distribution of the sum (or average) will be very close to the classic "normal", bell-curve distribution. Because of calculus-related technicalities, I have left "large" and "close" fuzzy here. Suffice it to say that for the purposes of most simulations, $N = 50$ is probably already "large".

The main use of the central limit theorem in simulation is as follows. Suppose we have a large number of independent random variables to simulate (such as the water usage of each of 200 homeowners in a certain neighborhood between the hours of 5 and 6 PM on a Tuesday in May). Suppose also that the only way that these variables affect the outcome

of the simulation is through their sum. Then we can save a lot of work by observing that central limit theorem says that the sum must essentially have a normal distribution. So, we can just simulate the value of the sum, rather than each of its individual components, resulting in a much simpler model. Thus, if we know the mean and standard deviation of the sum, we can (in @RISK) just use RISKNORMAL to simulate the value of the sum.

To get the mean of the sum, we just use the basic fact that if you add a bunch of random variables, then the expected value of the sum is the sum of their expected values. So, if we have N random variables with mean a, then the sum has expected value Na. The standard deviation of the sum is a bit harder. We use the fact that if random variables are independent, their *variances* (the squares of their standard deviations) add. From that, one can derive that if you add N independent random variables, each with standard deviation b, the sum will have standard deviation $N^{1/2}b$ ($N^{1/2}$ mean the square root of N).

Thus, if we add a "large" number N of independent random variables, each with mean a and standard deviation b, the sum will be well-approximated by Normal(Na, $N^{1/2}b$).

If we need an average instead of a sum, the situation is very similar, except that you have to divide through by N. Thus, the average will be distributed like Normal(a, $b/N^{1/2}$) - this fact is the key to deriving simple confidence intervals and sample sizes in statistics.

Note that just because a random variable has a standard deviation, it doesn't have to be normal. Essentially all random variables have standard deviations. It's just that the standard deviation is part of the commonly accepted way to distinguish members of the family of normal variables.

@RISK mechanics

Specifying simulation outputs: RISKOUTPUT()

If you want @RISK to keep statistical output information on some quantity *value* computed by your model, just add RISKOUTPUT(*name*) to the corresponding formula. The name argument is optional and specifies what @RISK should call the output in its report. If *name* is omitted, @RISK uses the cell coordinates of the RISKOUTPUT formula.

For example, putting the formula = RISKOUTPUT(“profit”) + B23 - B22 into cell B24 instructs @RISK that the quantity B23 - B22 is an output variable named "profit". B23 - B22 would also appear as the value of cell B24.

Specifying scenarios: RISKSIMTABLE(*list, name*)

You use this function to specify a decision variable for which you want @RISK to try various different values. The possible values are in the argument *list*, which is a block of cells. The *name* argument is optional, and specifies a name for the parameter in the output report; if you omit it, @RISK will use the cell coordinates.