Method 3

Constructing an empirical distribution from data

Situation

You have a set of random and representative observations of a single model variable, for example the number of children in American families (we'll look at a joint distribution for two or more variables at the end of this section), and you have enough observations to feel that the range and approximate random pattern has been captured. You want to use the data to construct a distribution directly.

Technique

Since the data already captures the pattern, one can simply use the empirical distribution of the data rather than fitting it to a parametric distribution. The main thing to keep on mind when using an empirical distribution is that extrapolating beyond the observed data can be difficult and subjective. Below, we outline three options to construct an empirical distribution:

1. Discrete Uniform: uses only the list of observed values

2. Cumulative: creates a cumulative distribution, and therefore allows values between those observed, and possibly values beyond the observed range

3. Histogram: similar to a cumulative distribution, but can be more efficient with large datasets.

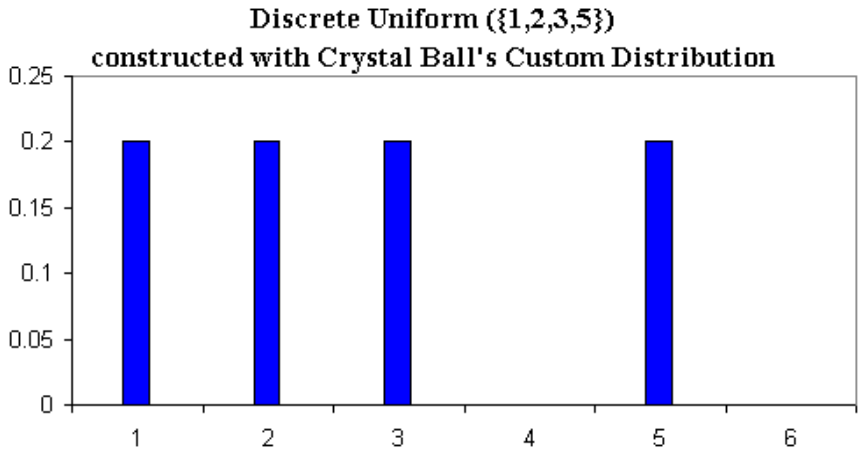
Option 1: A Discrete Uniform distribution

How to construct a discrete uniform distribution varies for different simulation software packages. Model Empirical Distributions provides an example.

The links to the Discrete Uniform software specific models are provided here:

Model Empirical Distributions (sheet Discrete) provides an example.

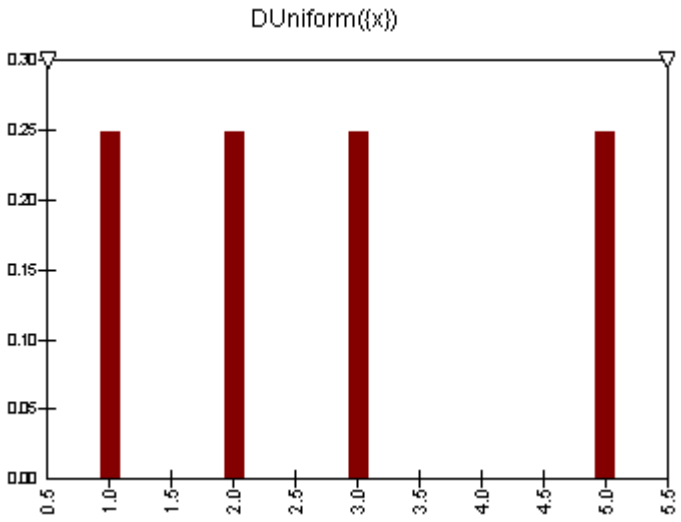
With Crystal Ball you can construct a Discrete Uniform distribution using Crystal Ball's Custom Distribution. The Discrete Uniform takes one parameter: a list of values. It then randomly picks any one of those values with equal probability (sampling with replacement). Thus, for example, a Discrete Uniform with data ({1,2,3,5}) will generate, with each iteration, one of the four values 1, 2, 3 or 5 (each value has during each iteration a probability of being picked of 25%). The figure below shows what the probability distribution looks like.



Let's imagine that we have our data in a column of Cells B1:B50. By simply using Crystal Ball's Custom Distribution to generate a Discrete Uniform distribution, we will generate a distribution that replicates the pattern of the observed data. You can use the Discrete Uniform distribution for both discrete and continuous data providing you have sufficient observations.

Model Empirical Distributions (sheet DUniform) provides an example.

@RISK offers a Discrete Uniform distribution that takes one parameter: a list of values. It then randomly picks any one of those values with equal probability (sampling with replacement). Thus, for example, =RiskDUniform({1,2,3,5}) will generate, with each iteration, one of the four values 1, 2, 3 or 5 (each value has during each iteration a probability of being picked of 25%). The figure below shows what the probability distribution looks like.



Let's imagine that we have our data in an array of Cells called 'Observations'. By simply writing =RiskDUniform(Observations) we will generate a distribution that replicates the pattern of the observed data. You can use the DUniform distribution for both discrete and continuous data providing you have sufficient observations.

Option 2: A Cumulative distribution

If your data are continuous you also have the option of using a Cumulative distribution.

Our best guess of the cumulative probability of a data point in a set of observations turns out to be r/(n+1) where r is the rank of the data point within the data set and n is the number of observations. Thus, when choosing this option, one needs to:

- Rank the observations in ascending order
- In the column to the left of the observations, calculate the rank of the data: write a column of values 1, 2, … n
- In the column immediately to the right of the data, calculate the cumulative probability F(x) = rank/(n+1)
- Use the data and F(x) columns as inputs to the distribution

Note that the minimum and maximum values of x only have any effect on the very first and last interpolating lines to create the Cumulative distribution, and so the distribution is less and less sensitive to the values chosen as more data are used in its construction.

Model Empirical Distributions provides an example.

The links to the Cumulative Distribution software specific models are provided here:

Model  Empirical Distributions (sheet Cumul) provides an example.

This model is generated using Crystal Ball's Custom Distribution. If you specify that the data are cumulative, Crystal Ball then constructs an empirical cumulative distribution by straight-line interpolation between the points defined on the curve.

It should be noted that for constructing this model in Crystal Ball, one should use the data and F(x) columns (two neighboring columns) as inputs to the Custom distribution, and specify in the "Define assumption" menu of the custom distribution that it is cumulative data.

Model  Empirical Distributions (sheet Cumul) provides an example.

This is a distribution that @RISK offers that takes four parameters: a minimum, a maximum, a list of values, and a list of cumulative probabilities associated with those values. From these parameters, it then constructs an empirical cumulative distribution by straight-line interpolation between the points defined on the curve.

It should be noted that for constructing this model in @Risk, one should use the data and F(x) columns as inputs to the RiskCumul distribution, together with subjective estimates of what the minimum and maximum values might be.

Option 3: A histogram distribution

Sometimes (admittedly, not as often as we'd like) we have enormous amounts of random observations that we would like to construct a distribution from (for example, the generated values from another simulation). The Discrete Uniform and Cumulative options described above start to get a bit slow at that point, and model the variable in unnecessarily fine details. A more practical approach now is to create a histogram of the data and use that instead.

Model Empirical Distributions provides an example.

The links to the Histogram Distribution software specific models are provided here:

Model  Empirical Distributions (sheet Histogram) provides an example.

The array function FREQUENCY( ) in Excel will analyze a data set and say how many lie within any number of contiguous bin ranges. The Crystal Ball distribution Histogram distribution, also constructed by using Crystal Ball's Custom Distribution, needs three neighboring columns: the first has minimum possible value, the second the maximum possible value, and the third the bin frequencies (or probabilities), which is just the FREQUENCY( ) array.

Model  Empirical Distributions (sheet Histogrm) provides an example.

The array function FREQUENCY( ) in Excel will analyze a data set and say how many lie within any number of contiguous bin ranges. The @RISK distribution Histogrm has three parameters: the minimum possible value, the maximum possible value, and an array of bin frequencies (or probabilities), which is just the FREQUENCY( ) array.

Creating an empirical joint distribution for two or more variables

For data that are collected in sets (pairs, triplets, etc), there may be correlation patterns inherent in the observations, and that we would like to maintain while fitting empirical distributions to data. An example is data of people's weight and height, where there is clearly some relationship between them.

Model Empirical Distributions provides an example.

The links to the Empirical Joint Distribution software specific models are provided here:

Model  Empirical Distributions (sheet Joint) provides an example.

A combination of using Crystal Ball's Discrete Uniform distribution with an Excel VLOOKUP() or OFFSET( ) function allows us to do this easily.

Model  Empirical Distributions (sheet Joint) provides an example.

A combination of using an IntUniform distribution with an Excel VLOOKUP() or OFFSET( ) function allows us to do this easily.