Using TCAce™ Information about Uncertainty To Support Decision Making

A Concise Tutorial

Sylvatica
Analysis and Software Benefiting Industry and Environment: Industrial Ecology, Pollution Prevention, Life Cycle Assessment
www.sylvatica.com *207-676-7640*

Using TCAce™ Information about Uncertainty To Support Decision Making

A Concise Tutorial

Contents

Sylvatica thanks the Dow Chemical Company for sponsoring the development of this tutorial, and Duane Koch of Dow Chemical for his leadership in applying and strengthening TCAce and the AIChE Total Cost Assessment method.

Sylvatica

Analysis and Software Benefiting Industry and Environment: Industrial Ecology, Pollution Prevention, Life Cycle Assessment

Introduction

One of the major reasons to use TCAce when evaluating the potential cost consequences of projects or alternatives is its treatment of uncertainty. Specifically, TCAce allows you to test the consequences of all sorts of uncertain events and their uncertain costs upon the possible results of your decisions. But not only does TCAce process lots of input uncertainties; it also produces lots of information about the outcome uncertainties. This document is intended to help you make productive use of this information in your decision making.

In particular, this document is designed to help you get the most insight possible from TCAce to support decisions. It helps you go beyond simple measures such as the "expected value" to consider what *ranges* of cost outcomes are possible, what your confidence can be that the actual costs will lie between given bounds, what the worst-case and best-case possibilities are, etc. It should even help you understand the significance of the expected value a little better.

This document is organized as follows. First, we introduce the basic statistical concepts you need in order to make the most of TCAce outputs. Next, we very briefly describe what Monte Carlo analysis is and how it works. We then describe how to use the Analytica portion of TCAce in order to view the different types of information about uncertainty which we have described. Finally, we illustrate the concepts and the use of the software on a series of examples.

Basic Concepts of Uncertainty and Probability

First, we introduce the following concepts:

- Mean (also known as the average, and the expected value)
- Mode
- Median
- Probability density
- Cumulative probability
- Percentiles (also known as fractiles and probability bands)
- Minimum and maximum values

The figures on this page are examples of the ***probability density*** and the ***cumulative probability*** plots provided in the Analytica portion of TCAce. They serve to illustrate and differentiate the three important concepts of ***mode, median***, and ***mean***. These curves pertain to the same uncertain variable. They are two different ways of looking at that variable – at the probabilities of different values which that uncertain variable might take.[1]

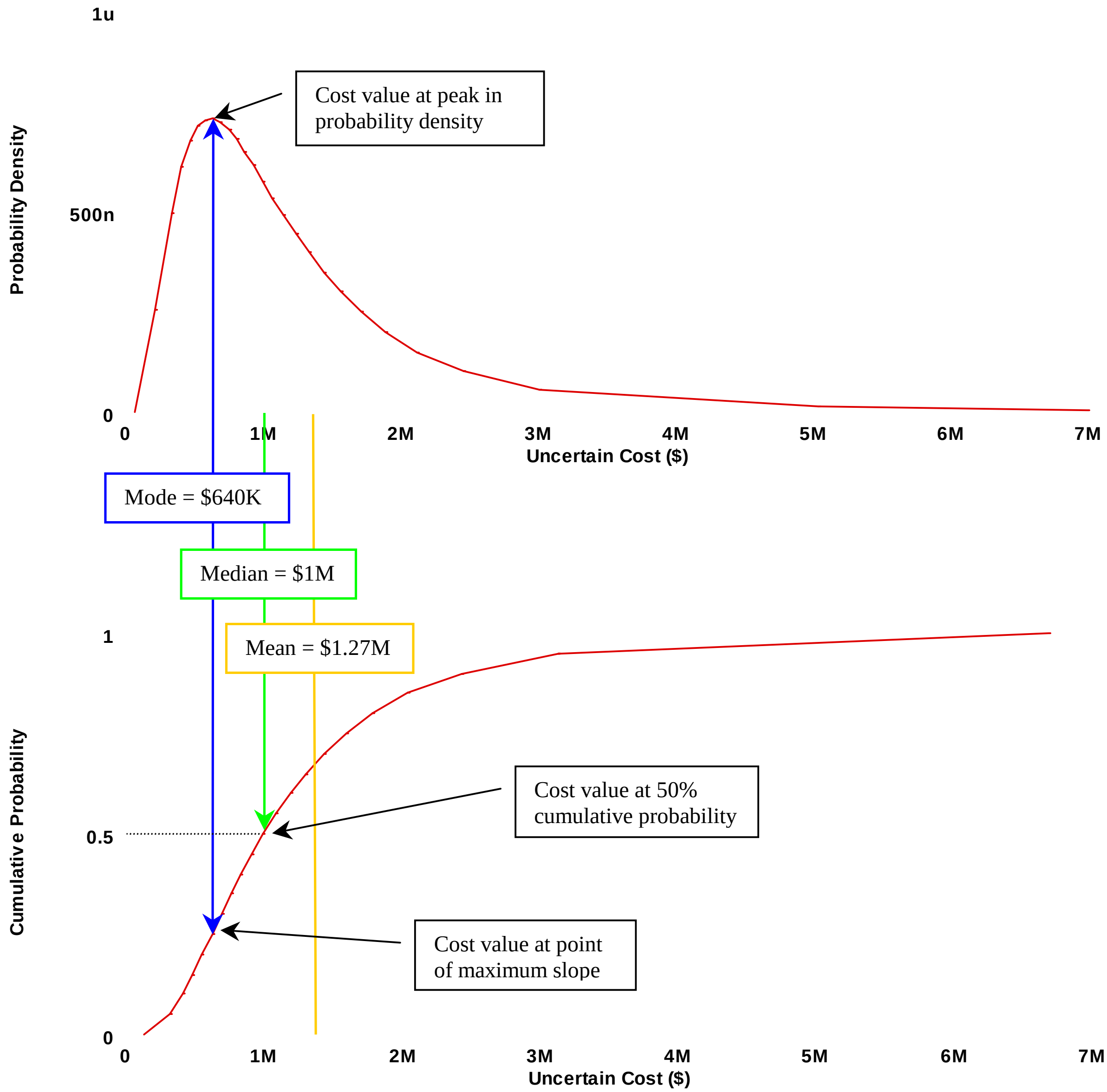


Figure 1: Mode, Median, & Mean in the Probability Density & Cumulative Probability Curves

[1] The figures pertain to a lognormal distribution with median of 1.0 (million dollars) and a geometric standard deviation of 2.0.

Concept 1: Relative likelihood and the probability density curve.

The relative height of the probability density curve at any cost point along the horizontal axis indicates the relative probability of the uncertain cost being that particular amount. Thus, the single *most likely* value for cost in Figure 1 is $640K.[2] This most likely value is termed the ***mode,*** and is identifiable graphically as the peak of the probability density curve. It can also be identified graphically on the cumulative probability curve as the point of maximum slope. What is the mode useful for? Well, if you are gambling, and you get a reward for being right (or a penalty for being wrong), but the penalty for being wrong has nothing to do with the size of your prediction error, then you should use the mode as your prediction. It has a higher chance of being right than any other estimate. (This is *not* the same thing as saying it has the lowest expected prediction error, as we will see.)

Concept 2: Area under the probability density curve

An important concept is that ***the probability of the cost being within a certain range is given by the area under the probability density curve over that range***.[3] Thus, the probability that the cost is between $2M and $3M is given by the area under the density curve within this cost region. (Note that by definition the total area under the entire density curve is 1.0, so the area under a portion of the curve equals that portion's share of the total area.) This concept helps us understand the meaning of the *median*, as well as the usefulness of the cumulative probability curve.

The ***median*** is the value for which there is an equal chance that the actual cost is either above it or below it. In the figures on the previous page, the median cost is $1M. Thus, there is a 50% chance that the actual cost will be *above* $1M, and a 50% chance that the actual cost will be *below* $1M. Graphically, the area under the density curve to the left of the median is equal to the area under the curve to the right of the median.

Now, ***the cumulative probability curve is calculated as the cumulative area under the probability density curve***, moving from left to right – that is, from 0 cost to higher and higher costs. Concept 2 told us that the probability of the cost falling within a given range is given by the area under the under the density curve for that region. This leads to concept 3.

Concept 3: The cumulative probability curve

The height of the cumulative probability curve for a given cost value gives the probability that the actual cost is equal or less than that cost value. *The cumulative probability curve rises from 0 (at the minimum possible cost) to 1 (at the maximum*

[2] Strictly speaking, for a continuous probability density function, any single precise cost value would have essentially zero probability, since there are a nearly infinite number of possible costs between the maximum and minimum values. However, if we consider the probabilities of costs being in the neighborhood of a finite set of possible costs (e.g., rounded to thousand dollar increments), then each cost neighborhood has a nonzero probability whose relative magnitude is indicated by the relative height of the probability density curve in that cost neighborhood.

[3] In calculus terms, the probability is given by the integral of the probability density function over that interval.

possible cost). It equals 50% (or 0.5) at exactly the median cost. Note that this must be true, since there is a 50% probability that the actual cost will be at or below the median.

The cumulative probability curve is also a helpful graphical way to understand the meaning of confidence intervals, also known as fractiles, or percentiles, or probability bands. Probability bands are illustrated in Figure 2. The 10th probability band or percentile, for example, is the value which the actual cost has only a 10% chance of being below. The 90th percentile is the value which the actual cost has a 90% chance of being below (or a 10% chance of being above). Thus, the actual cost has an 80% chance of being somewhere between the 10th and 90th percentiles – that is, between $412K and $2.427M. We can refer to this range as the 80% confidence interval for the actual cost, centered about the median.

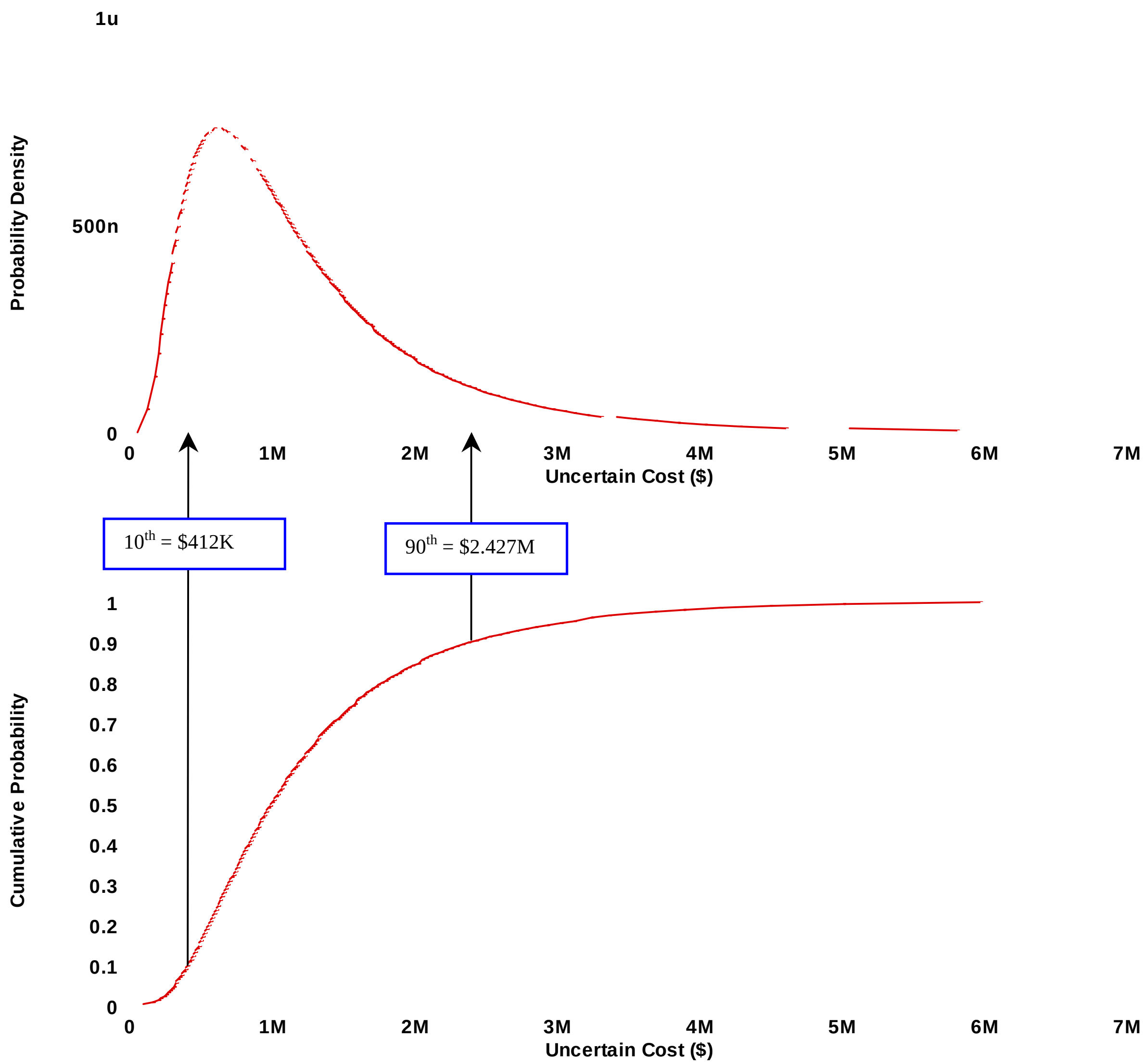


Figure 2: 10th and 90th Percentiles in the Probability Density & Cumulative Probability Curves

We haven't mentioned the ***minimum*** or ***maximum*** values yet. These are commonly understood concepts that really require little explanation. Perhaps the most important point to emphasize here is that the estimated maximum and minimum values are still only estimates, not guarantees. And, they are only as good as the input information which you provide to TCAce about the possible costs stemming from scenario occurrence. We will look a bit further at minimum and maximum cost estimates in the third section of this document.

Finally, what about the ***mean?*** The mean value is also known as the arithmetic average, or simply the "average" value. It is also referred to as the "expected value." We tend to think of the mean as also referring to the "most likely" value (the mode), but this only true if the density curve is symmetric. (The density curve in Figures 1 and 2 is not symmetric, but instead is referred to as "positively skewed.") So what is the physical meaning of the mean value?

Mean = Average = Expected Value

For one thing, ***the mean is the cost estimate which has the lowest expected error*** (strictly speaking, the mean is the "least squares error estimate") for the actual value. By using the mean as our "best guess" for the actual, we will minimize the expected absolute magnitude of the error between our estimate and whatever the actual turns out to be.

Also, if we have a cost which is expected to recur, then the best estimate of the total sum of these separate incidents is given by the mean times the number of occurrences. For example, let's say we have a cost whose mean is $1,000, which will occur annually for 10 years. The minimum-expected-error estimate for the total is 10* $1K = $10K.

A Brief Introduction to Monte Carlo Analysis

Monte Carlo Analysis refers to a family of simulation methods for quantifying the influence of uncertain assumptions (or "inputs" to mathematical models) upon conclusions that depend on them ("outputs" or model results).

In a nutshell, Monte Carlo analysis involves three steps.

1) First, the uncertainty in inputs is described using probability distributions, which express the likelihood that each input could take on certain possible values.
2) Second, a possible value is randomly selected for each of the uncertain inputs, and the results of the model are computed using a full set of selected values, one for each uncertain variable in the model. The results are saved in memory as a single "run" or "iteration" or "sample." In TCAce simulations include the dimension of time, and so an "iteration" or "run" is an entire possible future, an entire and self-consistent set of possible annual outcomes for each scenario and its potential costs, from year 0 through the final year.
3) Finally, after performing this randomly-select-and-compute procedure many (for example, 100) times, the (100) different simulated outcomes ("iterations" or "runs") are evaluated to provide a picture of the ranges of possible outcomes and their relative likelihood.

Controlling How the Monte Carlo Simulation is Performed

TCAce's probabilistic simulation model has been programmed using the Analytica modeling environment, so it provides you with extensive control over how the Monte Carlo simulation is performed. For example, you can set the number of iterations or runs (the "sample size"). You can also choose the algorithm for Analytica to use in selecting possible values for each of the uncertain inputs for each run. Finally, you can control what sorts of output statistics are computed to describe the results.

To learn more about the options outlined below and their influence upon results, select "Help Topics" from the "Help" menu, click on the "Index" tab, and then type "Uncertainty Setup Dialog Box" in the search field and hit Return.

To access the Monte Carlo controls, select "Uncertainty Options..." from the "Result" menu. You can directly specify the sample size in the resulting dialog box.

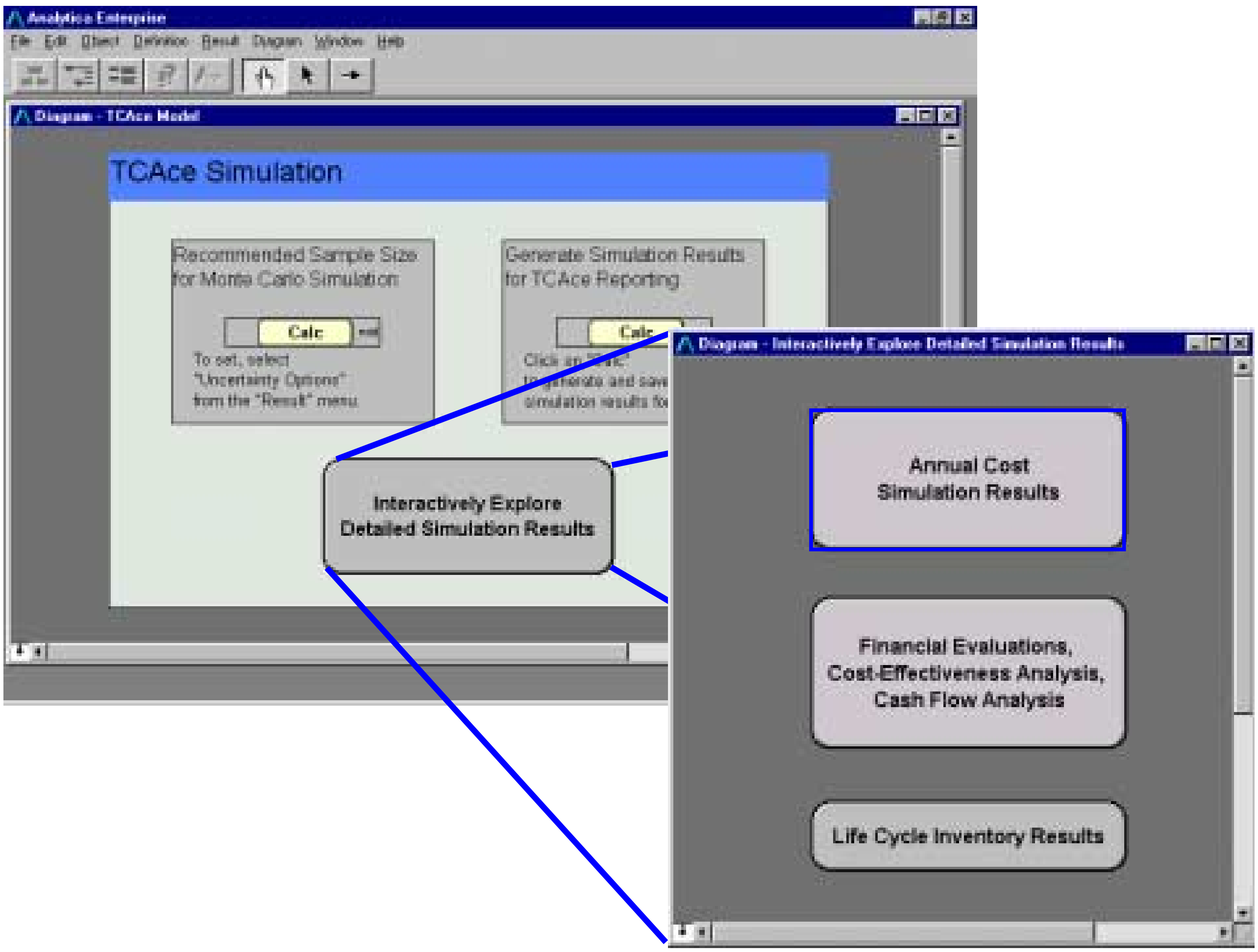
The "Analysis Option:" menu within this dialog box gives you the ability to select:

- The sample size, sampling method, randomization method, and the random seed;
- The statistics to use in reporting probabilistic results;
- The probability bands (percentiles) to show in results tables and graphs;
- The methods (and resolution) used for drawing probability density and cumulative probability density values.

Viewing Information About Uncertainty in TCAce

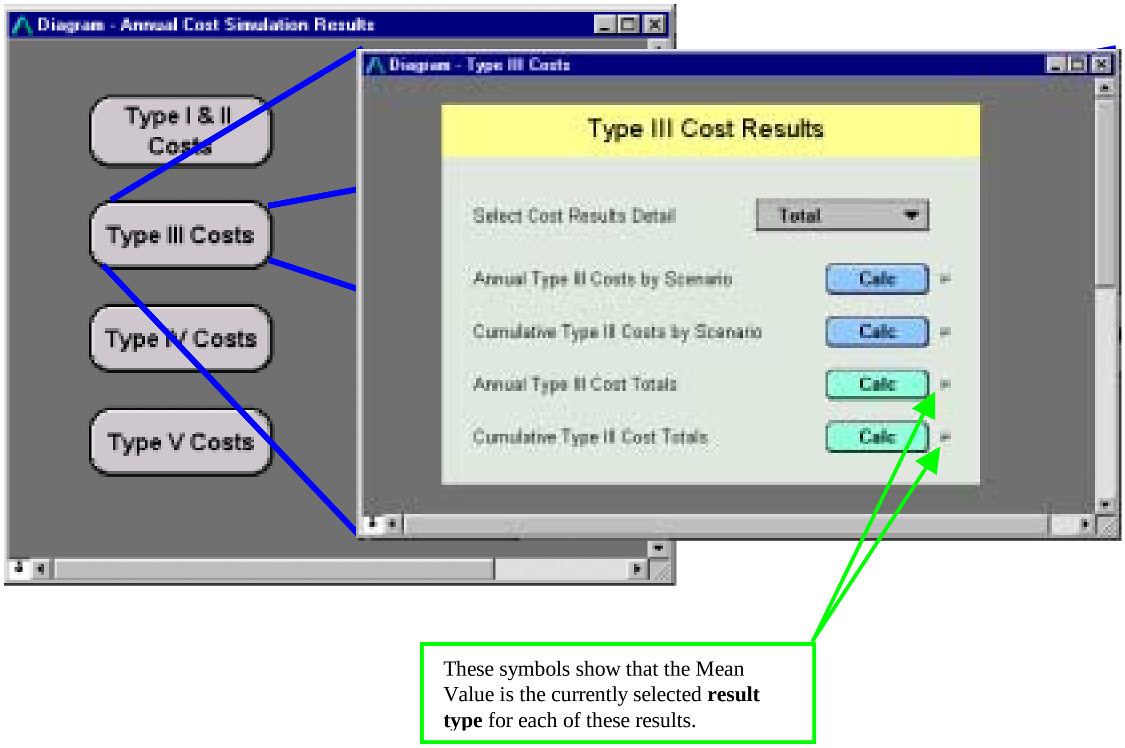
As we have seen, there are several different ways to look at an uncertain result. Here we review how you access these different ways within the Analytica portion of TCAce.

To explore information about uncertainty for decision support in TCAce, double-click on "Interactively Explore Detailed Simulation Results" within the main window. You then can chose among Annual Cost Simulation Results, Financial Evaluations, and Life Cycle Inventory Results (see the figure below).



Let's choose Annual Cost Simulation Results for the present example; double-click on it. You will see the window displayed on the next page. From there, double-click on "Type III Costs."

Select a result to investigate by clicking on its "Calc" button.



A **results window** will open up, displaying the pre-selected result type, in either graphical or tabular format. You use the controls within the results window to change the *type of result* you are investigating, the *slice* of the data which you are observing, and the *format* for displaying it (graphical or tabular).

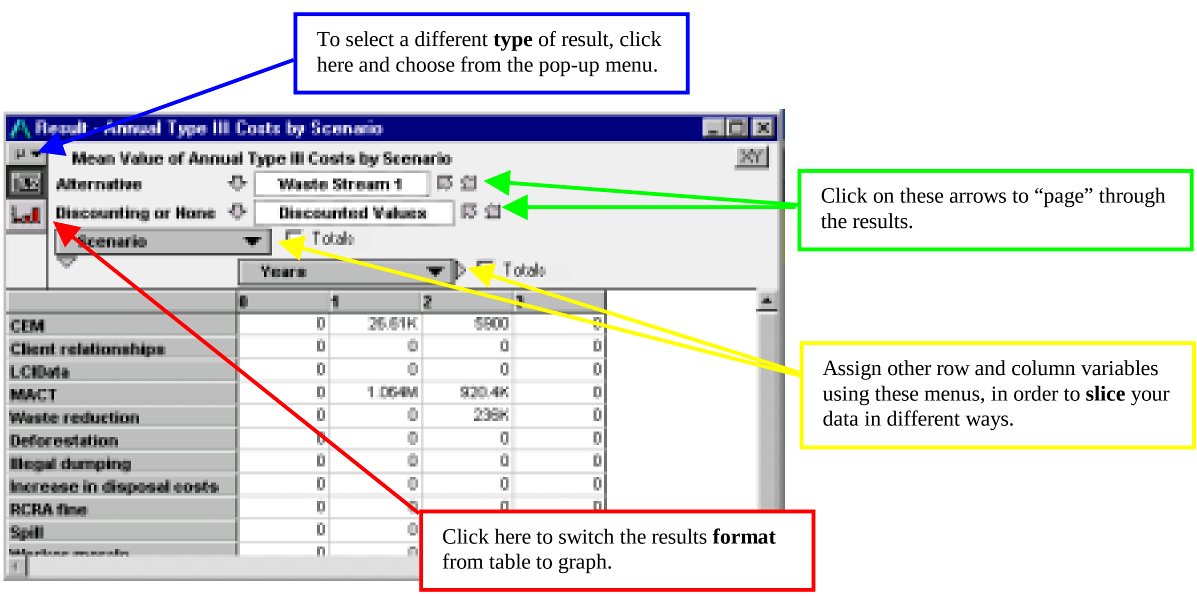


Figure 3: A **Results Window**

Using the result type menu, you can choose among the following options, each of which we have examined earlier in this document:

- "**Mid**" results are calculated without Monte Carlo analysis, simply by choosing the median value for each uncertain variable and computing the result. They are the only non-probabilistic result type;
- The **mean** or expected value of the result;
- **Statistics** for the result, including min, median, mean, max, and standard deviation;
- **Probability bands**, which we have also referred to as percentiles or fractiles;
- The **probability density** curve
- The **cumulative probability**; and
- The **sample** of results for each individual "run" or "iteration" of the Monte Carlo simulation.

Discussion of Some Sample Results

Finally, let's consider some hypothetical cost results for some example scenarios. Each of these examples is modeled in the file "Tutorial.ana". You could also try creating each simple scenario using TCAce. Recall that:

1) scenarios have specified probabilities of occurring;
2) scenarios are allowed to occur either once only (non-recurrable) or multiple times (recurrable);
3) *if* they occur, scenarios lead to one or more costs ("cost drivers") for which we may specify either best-guess values, a range of values with uniform probability, or a triangular distribution with min, max, and most likely (mode) value.

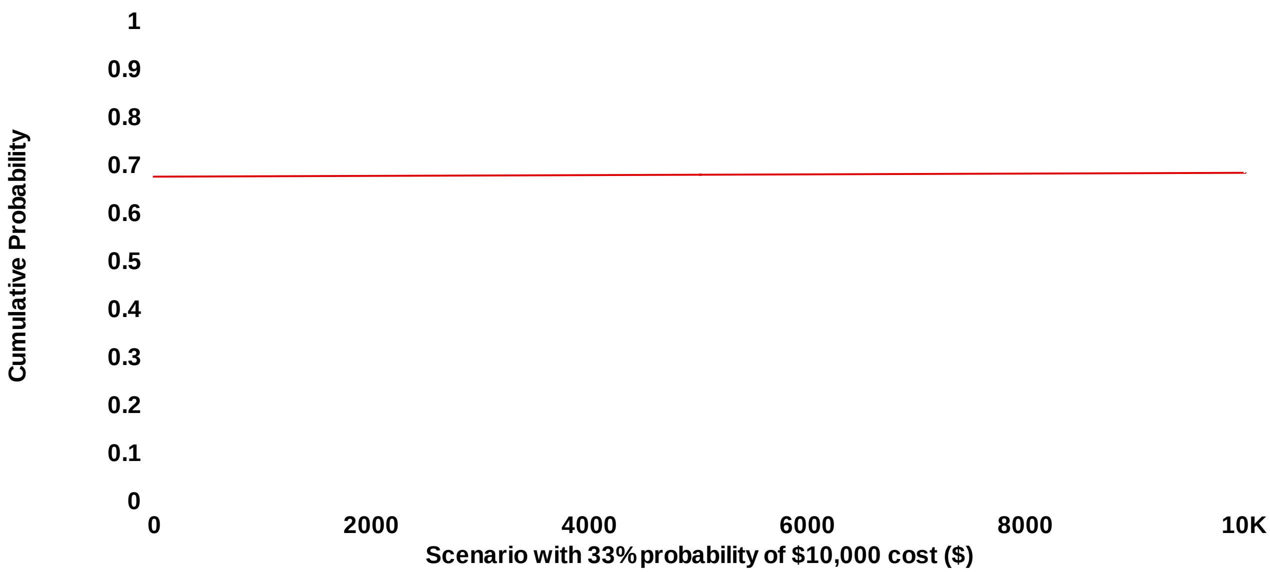
<u>Example 1: Single Possible Scenario with a Certain Cost, No Time Dimension</u>
First, let's imagine a scenario which has a 33% chance of occurring, and which – if it does occur – would lead to a single best-guess cost of $10,000. This is essentially like rolling the dice. Roll a 1 or a 2: we incur the cost of $10,000, roll a 3-6: we incur no cost at all. This outcome is called a "discrete" random variable, because it can take on only a discrete or finite set of possible values – either $0 or $10K, in this case. What will be the statistics, the probability density, and the cumulative probability for this scenario's cost?

Well, there is a 67% chance of zero cost, and a 33% chance of a $10K cost. What is the median cost? (Recall, the median cost is that cost which there is a 50% chance of exceeding.) It must be zero. Thus, the "mid" value for this cost is zero as well.

Let's try looking at the cumulative probability curve for this simple scenario to clarify things; it is shown below.

What is the minimum possible value of this cost? Zero.
What it the maximum possible value: $10K.
Thus, the cumulative probability curve will start at 0 probability of a cost below $0, and rise to a 100% probability of a cost at or below $10K. What will it do in between?

We can adopt the "frequentist" point of view of probability to make things easier to conceptualize. Imagine that we performed the experiment 100 times; i.e., we rolled the dice 100 times, or we had a way of considering 100 different, equally likely, possible future outcomes of this scenario. We measure the costs of each possible outcome, represent it as a bar lying on its side, and stack these cost outcome bars on top of each other, ranked from lowest cost to highest cost. The resulting graph would look like the cumulative probability curve shown below.



Sixty-seven of the 100 trials will have zero cost. Then, the remaining 33 trials will have a $10K cost. Thus, the 50th trial, in rank order by ascending cost, *which is the median cost,* is zero. So is the 10th trial, which is the 10th percentile. The 90th percentile is $10K.

What is the **mode**, or the most likely outcome? It, too, is $0. If you have to bet, bet on zero. But what should our guess be if we want to minimize the expected sum of our prediction errors, over the set of possible futures? The answer is the **mean** cost, which equals $3300.

Notice something: The mean cost, which is our minimum-expected-error guess, is $3300, which is not even a possible outcome! We will never be exactly correct with this guess.

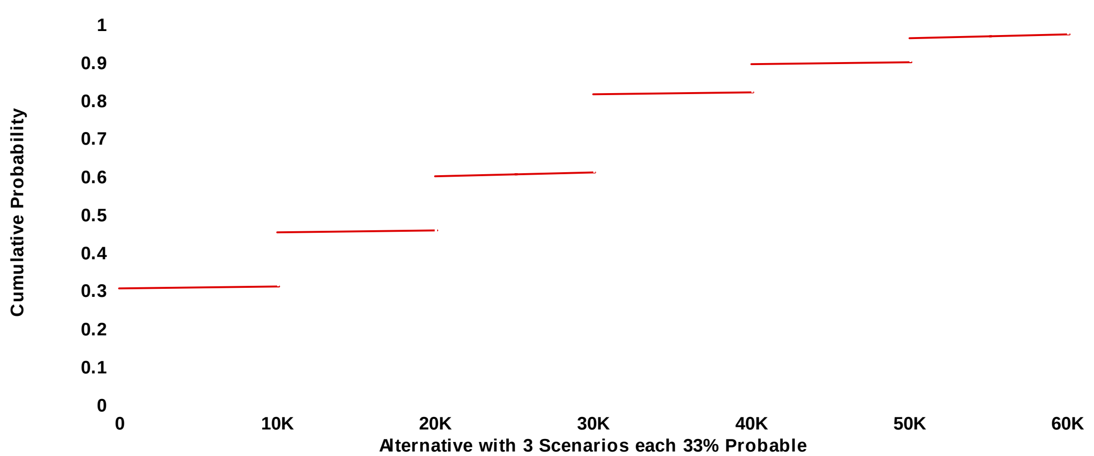
67% of the time the mean is $3300 too high, and the remaining 33% of the time it is $6700 too low. Yet, no other guess will yield a smaller average prediction error if we run the experiment many times. No other guess has a smaller expected error.

<u>Example 2: Multiple Possible Scenarios, Each with Certain Costs, No Time Dimension</u>

Now consider an alternative that has three different scenarios of the sort in example 1: three scenarios, each with a 33% chance of occurring. The scenarios are *independent*, meaning that whether one occurs or not will have no bearing on whether either of the other two occurs. The first scenario would lead to a cost of $10K. The second, if it happens, would lead to a cost of $20K. The third would yield a cost of $30K. What are the possible results, and their probabilities?

First, let's list each of the possible outcomes, or permutations.

Outcome	Cost	Probability of Outcome	Cumulative probability
None occurs	→ $0	0.67*0.67*0.67 = 0.3	0.3
#1 occurs	→ $10K	0.33*0.67*0.67 = 0.148	0.448
#2 occurs	→ $20K	0.67*0.33*0.67 = 0.148	0.596
#3 occurs	→ $30K	0.67*0.67*0.33 = 0.148	0.744
#1 & #2 occur	→ $30K	0.33*0.33*0.67 = 0.073	0.817
#1 & #3 occur	→ $40K	0.33*0.67*0.33 = 0.073	0.89
#2 & #3 occur	→ $50K	0.67*0.33*0.33 = 0.073	0.963
All occur	→ $60K	0.33*0.33*0.33 = 0.036	1.0



What is the single most likely cost value outcome, the mode? $0.
What is the median cost, the cost at 50% cumulative probability? $20K
What are the 10th and 90th percentiles, respectively? $0 and $50
What is the mean? (this is not observable from the plot above) $19.8K

You may have noticed that the "mid" value computed for the alternative is $0, even though the median cost is $20K. Why is this so? Because the mid is calculated using the median values for each of the inputs. The median for each scenario separately is $0, so the mid value for the composite (sum) of the scenarios is also $0.

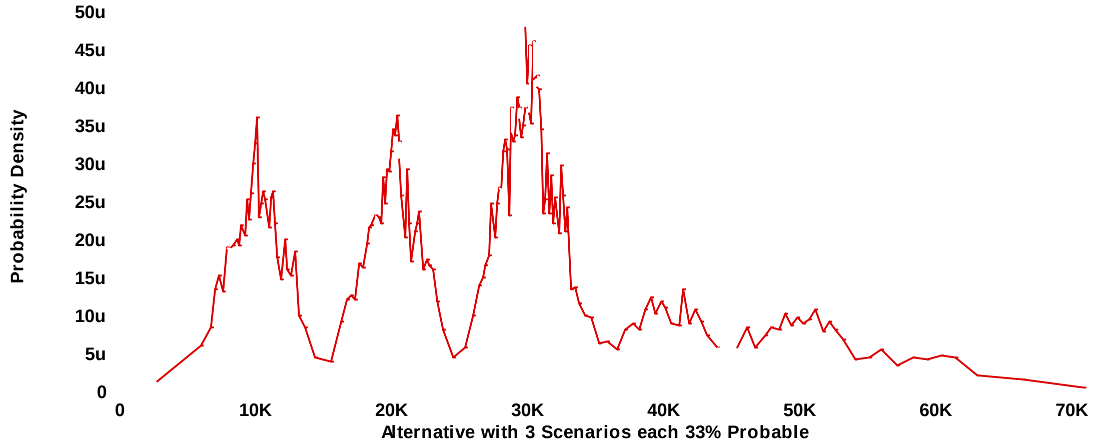
Example 3: Multiple Possible Scenarios, Each with Uncertain Costs, No Time Dimension

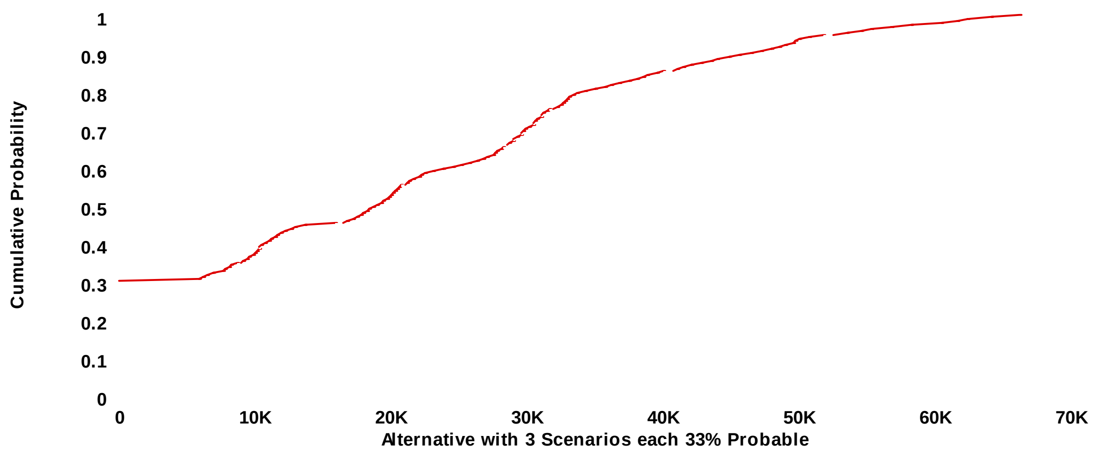
Finally, consider modifying the input information about example 2 as follows. Instead of certain costs of $10K, $20K, and $30K respectively, let's say that these costs are uncertain. Their most likely values are still given as above, but each could be up to $5K different from its expected value.

We model this uncertainty with triangular distributions for each cost:

Scenario 1: Triangular with min = $5K, mode = $10K, max = $15K.
Scenario 2: Triangular with min = $15K, mode = $20K, max = $25K.
Scenario 3: Triangular with min = $25K, mode = $30K, max = $35K.

The probability density and cumulative probability plots for this alternative are shown below. In contrast with the other examples, the cumulative probability plot is now smooth, and the probability density plot is intelligible. This is due to the "continuous" uncertainty introduced by using distributions to model uncertainty in the scenario costs.





We can observe three noticeable modes, or peaks, within the probability density plot, at the costs of $10K, $20K, and $30K. A the left hand side of this plot there is a very tall spike (mostly out of view) for the 30% probability of a cost value of zero.