Basic Overview: Monte Carlo Simulation

11/28/2016

Monte Carlo Simulation in Data Science

- Monte Carlo has numerous uses, but the main one is **to quantify and explore risks**--or different scenarios.
- While you can study your past data and business practices to learn from your successes and mistakes, a Monte Carlo simulation gives you a possible look into the future, and helps test your understanding of the situation today.
- Monte Carlo methods are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results, i.e. by running simulations many times in succession in order to calculate those same probabilities with machine learning. Monte Carlo simulations are often the precursor to building machine learning algorithms for specific classes of problems.

What is Monte Carlo simulation?

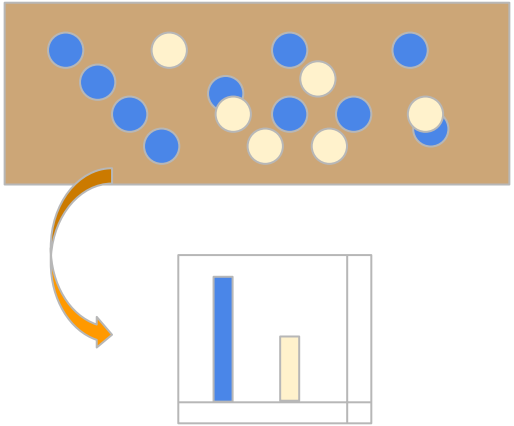
- Risk analysis *vs* What-if Analysis *vs* Scenario Analysis *vs* Sensitivity Analysis….
  - Monte Carlo Simulation tests thousands of tiny variations in scenarios

- Monte Carlo simulation provides the decision-maker with **a range of possible outcomes and the probabilities they will occur** for any choice of action.
  - It shows the extreme possibilities—the outcomes of going for broke and for the most conservative decision—along with all possible consequences for middle-of-the-road decisions.

A Monte Carlo simulation is simply a way to understand how inputs into a system might affect the outputs.

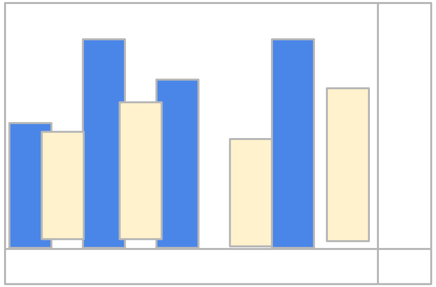
A simple example to illustrate...

Let's say you have a box of coins that you pour out onto a table. Once they all drop, you count tails and heads and come up with a ratio that you plot on a graph.

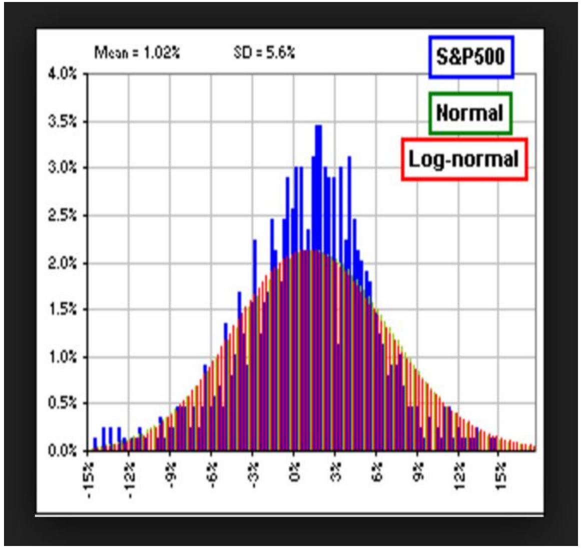


Then you do this same experiment 1000 times to see what the results and the graph look like. This process uses a Monte Carlo Method.

⇒ * 1000 (or whatever n trials you want)



Now, if you programmed a computer to model this same process and you ran the model 1000 times or more, graphing the results, you would have a Monte Carlo simulation.



How Monte Carlo simulation works...

- Its core idea is to use random samples of parameters or inputs to explore the behavior of a complex process.
    - To use this form of risk analysis you'll need numerical values related to the process, and basic analytical tools.
    - The simulation then begins building models of possible results by substituting a range of values—**a probability distribution**—for any factor that has inherent uncertainty.
    - It then calculates results over and over, each time using a different set of random values from the probability functions.

- Depending upon the number of uncertainties and the ranges specified for them, a Monte Carlo simulation could involve thousands or tens of thousands of recalculations before it is complete. **Monte Carlo simulation produces distributions of possible outcome values.**

How Monte Carlo simulation works…cont'd

- By using probability distributions, variables can have different probabilities of different outcomes occurring.
- During a Monte Carlo simulation, values are sampled at random from the input probability distributions.
    - Each set of samples is called an iteration, and the resulting outcome from that sample is recorded.
    - Monte Carlo simulation does this hundreds or thousands of times, and the result is a probability distribution of possible outcomes.
    - In this way, Monte Carlo simulation provides a much more comprehensive view of what may happen. It tells you not only what could happen, but how likely it is to happen.

An example ….

Let's say you want to plan for staffing a new customer service call center. You have estimates of fixed costs, variable costs per rep, call-handling times, and some past data on call volume.

- You start with a basic mathematical model that calculates costs based on your average handling time and expected call volume. (Note: This can be done easily in a spreadsheet such as Excel or use a statistical model….such as a Bayesian Regression model)

- Next, you ask "which factors are uncertain?" – such as call-handling time, and call volume per hour in this example. For each one, you replace a fixed number with an "uncertain variable" that can take on a range of possible values.

- Then, you run a series of "Monte Carlo trials" (think of spreadsheet what-if scenarios) where you plug in randomly chosen values for each uncertain variable (within its own range).

- **Getting an accurate range of possible values for each uncertain variable (a probability distribution) is a key step in creating a risk analysis model.** Instead of simply using parameters dictated by market experts or other outside influences, you can use past data and distribution fitting to help you choose.

Another example...

https://www.planacademy.com/monte-carlo-101-monte-carlo-schedule-simulations-work/

What Are The Benefits of Monte Carlo Simulations?

Monte Carlo simulation provides a number of advantages over deterministic, or “single-point estimate” analysis:

- *Probabilistic Results.* Results show not only what could happen, but how likely each outcome is.
- *Graphical Results.* Because of the data a Monte Carlo simulation generates, it’s easy to create graphs of different outcomes and their chances of occurrence. This is important for communicating findings to other stakeholders.
- *Sensitivity Analysis.* With just a few cases, deterministic analysis makes it difficult to see which variables impact the outcome the most. In Monte Carlo simulation, it’s easy to see which inputs had the biggest effect on bottom-line results.
- *Scenario Analysis:* In deterministic models, it’s very difficult to model different combinations of values for different inputs to see the effects of truly different scenarios. Using Monte Carlo simulation, analysts can see exactly which inputs had which values together when certain outcomes occurred. This is invaluable for pursuing further analysis.
- *Correlation of Inputs.* In Monte Carlo simulation, it’s possible to model interdependent relationships between input variables. It’s important for accuracy to represent how, in reality, when some factors goes up, others go up or down accordingly.

Quick check

EXERCISES

- Is it okay to use a normal distribution for the Monte Carlo simulation? Explain your answer.
- What happens to the error when the normal distribution is used?

EXERCISES

- What happens when you run the simulation over and over for the same n? Is the result what you expect?

Common probability distributions include:

Normal – Or “bell curve.” The user simply defines the mean or expected value and a standard deviation to describe the variation about the mean. Values in the middle near the mean are most likely to occur. It is symmetric and describes many natural phenomena such as people’s heights. Examples of variables described by normal distributions include inflation rates and energy prices.

Lognormal – Values are positively skewed, not symmetric like a normal distribution. It is used to represent values that don’t go below zero but have unlimited positive potential. Examples of variables described by lognormal distributions include real estate property values, stock prices, and oil reserves.

Uniform – All values have an equal chance of occurring, and the user simply defines the minimum and maximum. Examples of variables that could be uniformly distributed include manufacturing costs or future sales revenues for a new product.

Triangular – The user defines the minimum, most likely, and maximum values. Values around the most likely are more likely to occur. Variables that could be described by a triangular distribution include past sales history per unit of time and inventory levels.

PERT- The user defines the minimum, most likely, and maximum values, just like the triangular distribution. Values around the most likely are more likely to occur. However values between the most likely and extremes are more likely to occur than the triangular; that is, the extremes are not as emphasized. An example of the use of a PERT distribution is to describe the duration of a task in a project management model.

Discrete – The user defines specific values that may occur and the likelihood of each. An example might be the results of a lawsuit: 20% chance of positive verdict, 30% change of negative verdict, 40% chance of settlement, and 10% chance of mistrial.