Monte Carlo Simulations – A Case Study

Monte Carlo Simulations is a method to estimate outcomes of processes in which analytical solution is impossible. As can be understood from the name *Monte Carlo*, the technique heavily depends on randomness.

This document discusses the technique while it is being applied to a real-world data set. The data set consists of the body weight measurements of an individual, over the course of 44 days. The intention of the analysis is to have an idea about what is most likely to happen in the upcoming 30 days.

The code and data for the experiments can be reached at:

- https://github.com/bgunyel/monte_carlo

Günyel

The below figure shows the body weight measurements of the individual over the past 44 days.

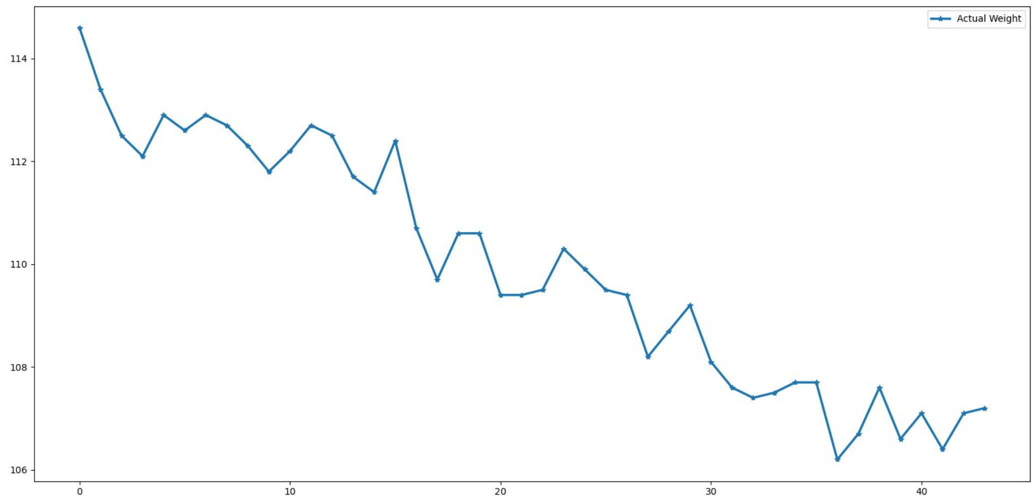

Illustration 1 - Actual body weight data for 44 days

As can be seen in Illustration 1, the body weight measurements of the individual started at 114.6 kg and ended at 107.2 kg.

The main idea of Monte Carlo application in this specific scenario is to draw random numbers for the **daily weight changes** for each of the next 30 days. Hence, body weight of the individual will be calculated for each of the next 30 days. The main assumption in this analysis is that daily weight changes are **independent and identically distributed** (we can test whether this assumption is true by utilizing time-series techniques, but we skip this test for keeping the analysis simple).

The immediate question arises: “how are we going to draw random numbers?” We will use the *Empirical Distribution* of the daily weight changes of the past 44 days. In other words, the simulated future daily weight changes will have the same characteristic with the actual past daily weight changes. By utilizing the *Inverse Transform Technique*, we generate random daily weight changes for the next 30 days and calculate the body weight of the individual. More interested readers might have a look at the following resources:

- https://en.wikipedia.org/wiki/Cumulative_distribution_function
- https://en.wikipedia.org/wiki/Empirical_distribution_function
- https://www.cyut.edu.tw/~hchorng/downdata/1st/SS8_Random%20Variate.pdf

Günyel

Some examples of the randomly generated simulation outcomes can be viewed in Illustration 2.

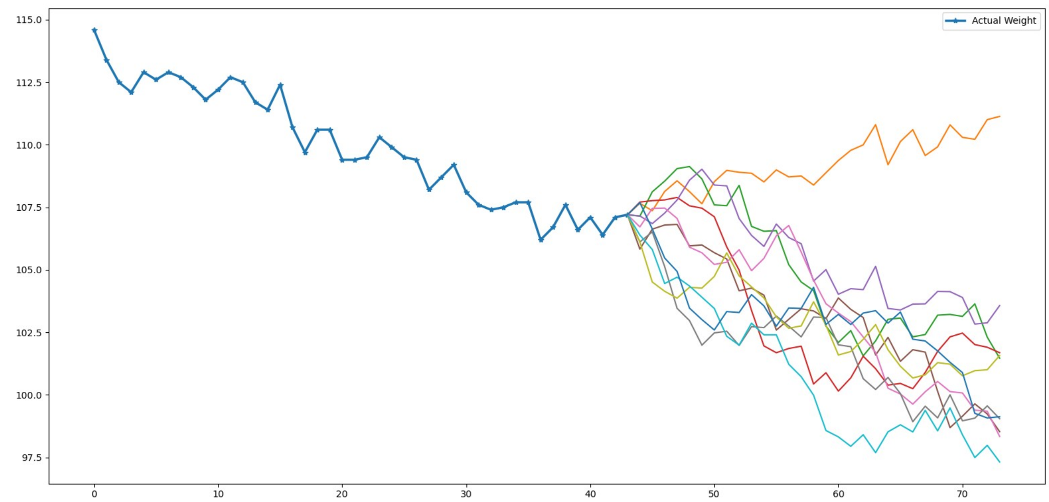

Illustration 2 - Some random simulation results

As can be seen in Illustration 2, most of the scenarios give consistent results with the main trend in the actual body weight data. However, as can be seen in the figure, there can be some outlier simulation results (the orange curve at the top), as well. In order for these outlier outcomes not to hurt the overall computation, the technique is applied many times and statistical significance is obtained. Please have a look at the following resource if you are unfamiliar with *Confidence Interval* concept:

- https://en.wikipedia.org/wiki/Confidence_interval

Illustration 3 shows the median simulation result (orange curve) and the 0.95 Confidence Interval (via red and green curves) obtained after running the simulation 10 times.

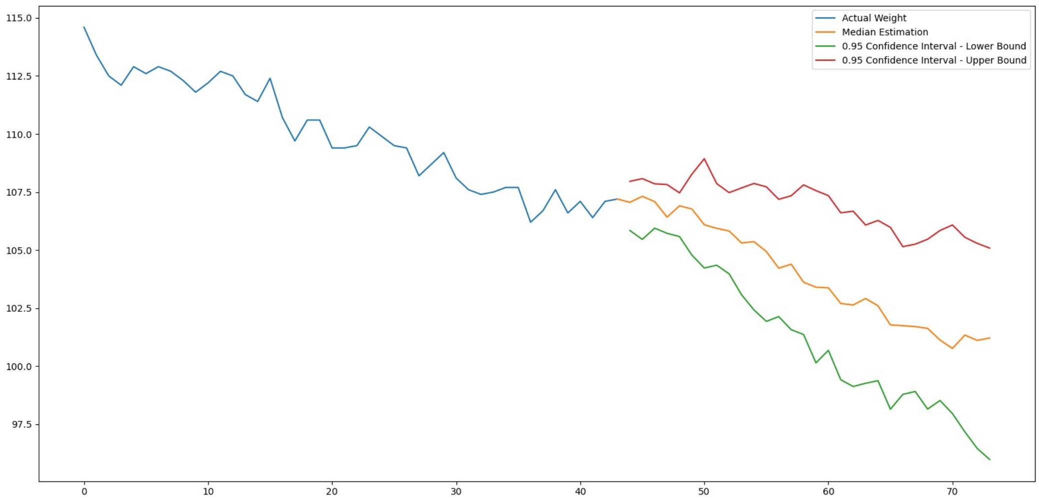

Illustration 3 - Simulation results with 10 repetitions

You can observe that the median output (orange curve) is quite noisy since 10 repetitions may not be sufficient to have statistical significance. Also, please observe that the 0.95 Confidence Interval has an increasing span between lower bound and upper bound as the simulation day advances in time. This makes sense because, any error made in simulating the body weight in day t affects the simulation of days t+1, t+2, t+3, … To be more precise, the daily weight changes are independent from each other but the body weight at day t is calculated by using the body weight at day t-1. On the other hand, the body weight at day t-1 is calculated by using the body weight at day t-2, and so on.

We can see the simulation results obtained with 100 repetitions and 1000 repetitions in Illustration 4 and Illustration 5, respectively with the corresponding 0.95 Confidence Intervals.

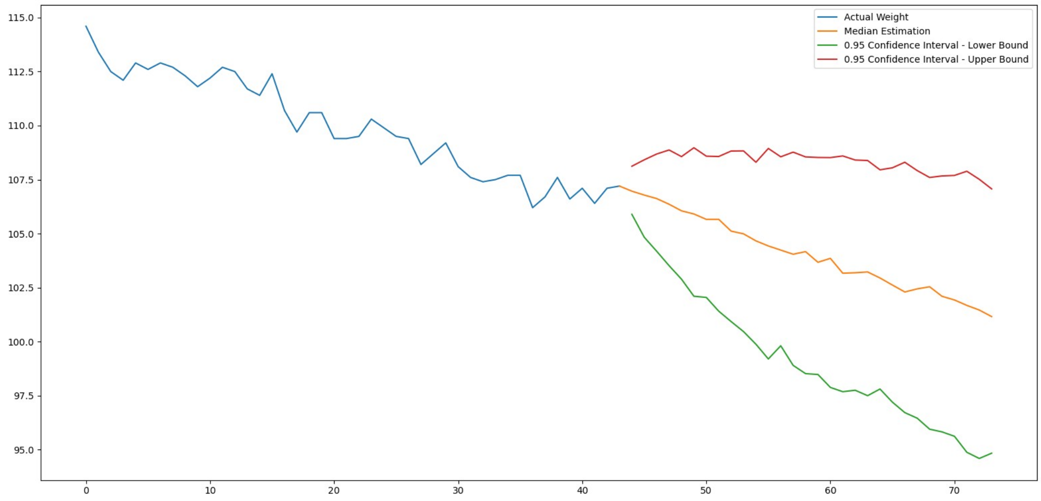

Illustration 4 - Simulation results with 100 repetitions

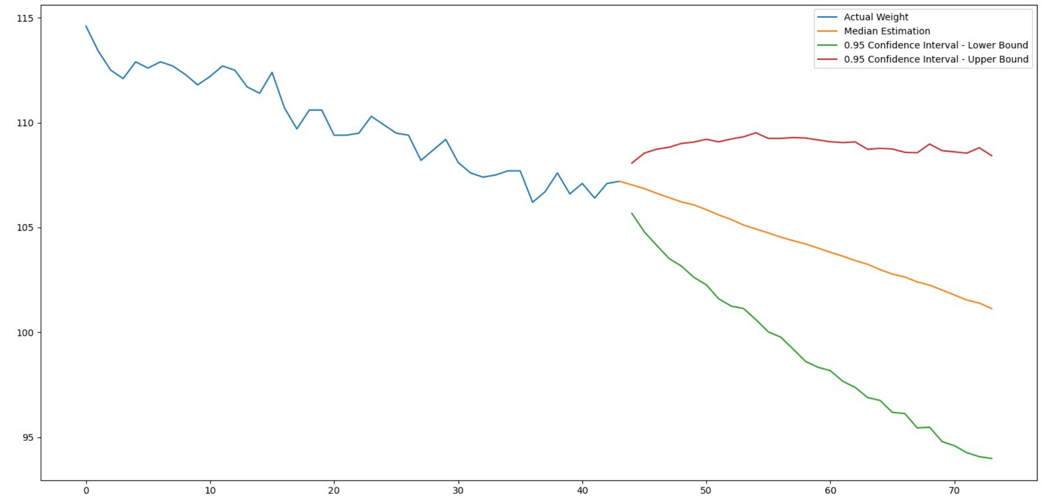

Illustration 5 - Simulation results with 1000 repetitions

Some conclusions can be drawn with an overall comparison of the simulation results in Illustration 3, Illustration 4, and Illustration 5. As the number of repetitions increases from 10 to 100 and to 1000, the median simulation result becomes smoother and becomes better at representing the overall trend of the body weight change. The same smoothing impact can be observed on the upper and lower bound curves of the 0.95 Confidence Interval. This can be explained by the *Law of Large Numbers*. Simply stated, the average of the results obtained from a number of trials becomes closer to the expected value of the process as more trials are performed [https://en.wikipedia.org/wiki/Law_of_large_numbers].

On the other hand, as the number of repetitions increases, the span of the 0.95 Confidence Interval increases, which is the direct consequence of the application of randomness more. Variance reduction techniques can be utilized to decrease the span of the 0.95 Confidence Interval but this is out of the scope of this document.

In conclusion, Monte Carlo Simulations is a theoretically well-established technique which is widely used in:

- situations to predict/simulate a not-yet-happened scenario
- situations in which analytical solution to the problem is unknown (e.g. computing some very advanced integrals in quantum physics)

The overall recipe for applying the Monte Carlo Simulations technique could be written as:

- Analyze the scenario and identify the elements which include randomness
- Identify the distribution of the random elements (either fitting a theoretical distribution or deducing the empirical distribution)
- Draw random numbers for the random elements according to the distribution you derived.
- Repeat the scenario many times to obtain statistically significant results.

Bertan Günyel