A Review of Basic Monte Carlo Methods

Julian Haft

May 9, 2014

Introduction

One of the most powerful techniques in statistical analysis developed in this past century is undoubtedly that of Monte Carlo simulation. First conceived by scientists working on the nuclear bomb in the forties, the theory of Monte Carlo simulation emerged simultaneously with the first computers capable of producing such simulations. In the 70 some odd years following they have become ubiquitous in such diverse fields as finance, chemistry, physics, biology and economics.

This paper serves to elucidate the theoretical foundations of the Monte Carlo methods, while demonstrating some basic applications.

Motivation

Frequently the properties of a statistical model cannot easily be developed through statistical theory alone, due to complexity or gaps in understanding, necessitating the use of empirical methods. How do we empirically study a theoretical model? Well we have to find an actual instance of it that is conducive to study. However this is usually an impossible proposition. So we use Monte Carlo methods to generate data indistinguishable from data collected from an actual phenomena that adhered to the specifications of our model.

Monte Carlo methods are thus uniquely suited to empirically studying the properties of theoretical models and this is how they're most often used (Though there are a plethora of novel applications like the decryption of ciphers and computer learning models.).

Heuristic explanation

Monte Carlo methods are based in a rather simple analogy between probability and volume. The idea is encapsulated in the following method for calculating the volume of the unit circle. Let X and Y be i.i.d. $U_{[-1,1]}$. Generate a sample of size N from the pair (X, Y). Eliminate all (x_i, y_i), in the sample, not satisfying $x_i^2 + y_i^2 \leq 1$. Take the ratio of the number of remaining values over the total number of values calculated. That ratio should be approximately the ratio of

area the unit disk and the box $[-1,1]^2$ and it is also an approximation of the probability that an observed value of (X,Y) lies within the unit disk.

The above procedure relies heavily on the idea that we can simulate uniform random variables or equivalently that we can simulate $U_{[0,1]}$ (As all uniform distributions can be represented as scalings and translations of $U_{[0,1]}$.). Unsurprisingly we can in fact simulate $U_{[0,1]}$ perfectly according to any reasonable set of tests for randomness. General Monte Carlo methods are primarily predicated on the possibility of picking countably many random variables from a known probability distribution. Surprisingly this boils down to being able to simulate $U_{[0,1]}$, as almost any distribution can be simulated from a uniform distribution.

In the following section we will outline the methods for simulating random variables from a uniform distribution beginning with the simplistic inverse probability transformation and ending with the extraordinarily powerful accept-reject method.

Simulating random variables

It will become apparent in the exposition of methods that almost all Monte Carlo methods require us to be able simulate indefinitely large samples from arbitrarily distributed random variable. In this section we demonstrate that we can in fact satisfy this requirement for almost any conceivable distribution.

The first method exploits a connection between cumulative distribution functions and the uniform distribution on the unit interval.

Inverse Transform

Let X be a continuous random variable with pdf f and cdf F. Then we have:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Let $U \sim U_{[0,1]}$ then, assuming F has an inverse, we have

$$P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x).$$

So $F^{-1}(U) \sim X$. For the general case, where we don't know if F has an inverse, we use right continuity and monotonicity of a continuous cdf to define an analytic inverse:

$$F^{-1}(x) = \inf\{t | F(t) \geq x\,, 0 < t < 1\}.$$

Clearly $F^{-1}(U) \sim X$ for all continuous X. However, since F^{-1} may not have a closed form this method is often useless.

Example

As an example of this method consider $X \sim \beta(\frac{1}{2}, \frac{1}{2})$, a special case of the β distribution where:

$$F(x) = \frac{2}{\pi} \arcsin(\sqrt{x}), \text{ so } F^{-1}(x) = \sin^2\left(\frac{\pi x}{2}\right).$$

We drew 10^4 samples from $U_{[0,1]}$ and applied F^{-1} to each of them and then plotted a histogram of our data with forty bars and superimposed the pdf of X on top of it:

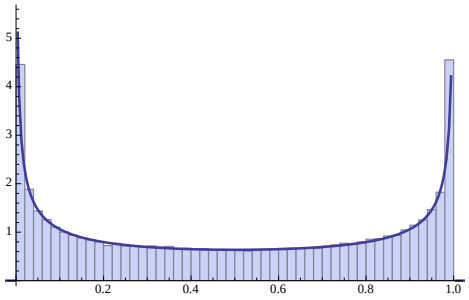

Figure 1: Simulation results

The treated data clearly has the desired distribution. However we cannot use this method to simulate a general β-distribution, as its cdf generally doesn't permit a closed form and therefore doesn't permit a closed form inverse. Thus we need additional methods before we can be satisfied that the Monte Carlo methods provided rely on reasonable assumptions.

Accept-Reject

The following method is far more general and has a much more intuitive core approach, however it takes a bit of effort to expose this intuition. The following procedure is not the actual method, but it is that the accept-reject method approximates.

Suppose we wish to simulate a random variable X with known pdf f. Let (X, U) be a the pairing of X with a random variable U that satisfies $U|X = x \sim U_{[0,f(x)]}$, so:

$$\begin{aligned} P(X = x \wedge U = u) &= P(U = u|X = x)P(X = x) \\ &= \left(\frac{1}{f(x)}\right) f(x) \\ &= 1 \end{aligned}$$

Thus:

$$P(X = x) = \int_0^{f(x)} P(U = u \wedge X = x)du$$

The above integral gives the marginal density of X, so:

$$P(X = x) = f(x).$$

Now this result is completely unexciting since in order to get it we would have to be able to simulate X. Nevertheless consider what samples from (X, U) actually are, namely points under the probability distribution of X. The idea of the accept reject method is generate samples from $(Y, U_{[0,1]})$. Where Y is a random variable with a pdf g with a support that contains the support of f and which differs from f by at most a multiplicative factor M. is a random variable following the uniform unit distribution. From this sample we then throw out all pairs (Y_i, U_i) for which U_i is not less than $f(Y_i)$ by eliminating all pairs where:

$$U_i \leq \frac{f(Y)}{Mg(Y)}.$$

Though we inevitably will have thrown out some pairs under the pdf of X, all the remaining pairs are all under the graph of f, thus the marginal probability of Y for the accepted pairs is X distributed. Thus the set of accepted Ys in the sample are indistinguishable from a sample from X.

The mathematical justification for this method can be found in the following calculation which proves that $P(Y \leq x|U \text{ is accepted}) = P(X \leq x)$ (Since the proof is quite arduous and not terribly enlightening, unless you're interested in exploiting the relationship between joint and conditional probabilities, it may be skipped.)

$$P\left(Y \leq y | U \leq \frac{f(Y)}{Mg(Y)}\right) = \frac{P\left(Y \leq y,\, U \leq \frac{f(Y)}{Mg(Y)}\right)}{P\left(U \leq \frac{f(Y)}{Mg(Y)}\right)}$$

First we solve for the denominator.

$$\begin{aligned} P\left(U \leq \frac{f(Y)}{Mg(Y)}\right) &= \int_{-\infty}^{\infty} P\left(U \leq \frac{f(Y)}{Mg(Y)} | Y = y\right) P(Y = y) dy \\ &= \int_{-\infty}^{\infty} \frac{f(y)}{Mg(y)} g(y) dy \\ &= \frac{1}{M} \end{aligned}$$

Next we calculate

$$\begin{aligned} P\left(U \leq \frac{f(Y)}{Mg(Y)} | Y \leq y\right) &= \frac{P\left(U \leq \frac{f(Y)}{Mg(Y)}, Y \leq y\right)}{P(Y \leq y)} \\ &= \frac{1}{G(y)} \int_{-\infty}^{y} P\left(U \leq \frac{f(Y)}{Mg(Y)} | Y = t\right) P(Y = t) dt \\ &= \frac{1}{G(y)} \int_{-\infty}^{y} \frac{f(t)}{Mg(t)} g(t) dt \\ &= \frac{F(y)}{MG(y)} \end{aligned}$$

Using this we evaluate the numerator:

$$\begin{aligned} P\left(Y \leq y, U \leq \frac{f(Y)}{Mg(Y)}\right) &= P\left(U \leq \frac{f(Y)}{Mg(Y)} | Y \leq y\right) P(Y \leq y) \\ &= \frac{F(y)}{MG(y)} G(y) \\ &= \frac{F(y)}{M} \end{aligned}$$

So

$$P\left(Y \leq y | U \leq \frac{f(Y)}{Mg(Y)}\right) = F(y)$$

In the above calculation we incidentally calculated the probability that a pair is accepted, here we reproduce it more explicitly.

$$\begin{aligned} P\left(U \leq \frac{f(Y)}{Mg(Y)}\right) &= \int_{-\infty}^{\infty} P\left(U \leq \frac{f(Y)}{Mg(Y)} | Y = y\right) P(Y = y) dy \\ &= \int_{-\infty}^{\infty} \frac{f(y)}{Mg(y)} g(y) dy \\ &= \frac{1}{M} \end{aligned}$$

Limitations

The accept-reject method is far more generally applicable than the inverse transform as it only requires us knowing f and being able to identify a probability distribution Ythat is appropriate for use in the algorithm. Unfortunately identifying such a Y can be difficult and if M is large then generating a sample of approximately size N from X can be computationally intensive as we would have to generate a sample of size $M \cdot N$ from (Y, U).

Example

As an example suppose we want to simulate $X \sim \beta(2,3)$, let f be the pdf of X. First we calculate $M = \max_{x\in[0,1]} f(x) = \frac{16}{9}$. Then we take $Y \sim U_{[0,1]}$ and $U_{[0,1]}$ and generate 10^5 pairs (Y,U), we expect to keep 56,250 after filtering out the ones which did not satisfy $U \leq \frac{9f(Y)}{16}$. In the simulation I ran $56,464$ were accepted. The plotted histogram below shows that the accepted values are distributed $\beta(2,3)$

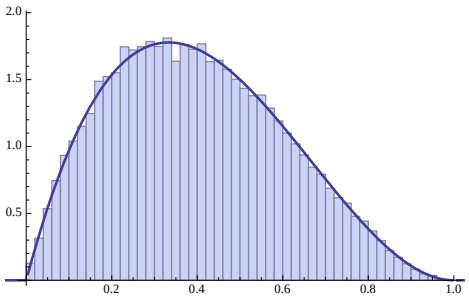

Figure 2: Simulation results

Discrete RVs

Suppose X is a discrete random variable whose range contains a countably infinite or finite number of values, $x_1, x_2, \ldots, x_k, \ldots$. Suppose f is the pdf of X and let $F(x) = P(X \leq x) = \sum_{t\leq x} f(t)$. We can simulate X by sampling from $U \sim U_{[0,1]}$ and for each U calculating a corresponding Y equal to the least x_j such that $u \leq F(x_j)$. We clearly have $Y \sim X$.

We have demonstrated three methods of generating random variables from other random variables. Our claim is that the above methods allow us to simulate a random variable with almost any distribution.

Monte Carlo Integration

Let X be a continuous RV with pdf f that takes values in χ. Let g be a function on χ. One of the central techniques in Monte Carlo simulations is the approximation of the following class of integrals:

$$E_f[g(X)] = \int_\chi f(x)g(x)dx.$$

The method is as follows: generate a sample $(X_1, \ldots, X_n)$ from X and calculate the following sum

$$\hat{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

The strong law of large numbers ensures that our estimator with converge to $E_f[g(X)]$ as n approaches infinity. The variance of $\hat{g}_n$ can be expressed as:

$$\text{var}(\hat{g}_n) = \frac{1}{n}\int_{\chi} \left(g(x) - E_f[g(X)]\right)^2 f(x)dx.$$

Of course we can use Monte Carlo integration to estimate $\text{var}(\hat{g}_n)$ by calculating:

$$\sigma_n = \frac{1}{n^2}\sum_{i=1}^{n}(g(X_i) - \hat{g}_n)^2.$$

Let 1_C is the characteristic function of the segment of the unit disk in the first quadrant of $\mathbb{R}^2$. The method we presented in the introduction to approximate the area of this segment is equivalent to using the Monte Carlo integral to approximate $E_f[1_C(X)]$, where f is the pdf of $X \sim (U_{[0,1]}, U_{[0,1]})$.

The utility of these estimators lies in their general applicability. However there are various conditions that can make the method very computationally expensive. We go over one in the following example.

Example

Let $X \sim N(0,1)$, let $1_{|X|>2.6}$ be the indicator function for the inequality in the subscript, and let f be the pdf of X.

Note that the expected value of $E[1_{|X|>2.6}] = \int_{-\infty}^{\infty} 1_{|X|>2.6}(x)f(x)dx$ is equivalent to $1 - \int_{-2.6}^{2.6} f(x)dx$ or the probability that $|X| > 2.6$. The exact numerical value of the integral is 0.009322.

Generating a sample of size 10^5 from X 950 values in the support of $g = 1_{|X|>2.6}$ giving

$$\hat{g}_{10^5} = \frac{950}{10^5} = 0.0095.$$

This is a rather poor estimate for such a gigantic sample. Intuitively this is because the only values of interest to us are extremely rare values of X and therefore we have to simulate X many times to get one such value.

Importance sampling

The shortcoming illustrated in the example above can be overcome by using a method called importance sampling. Let X be a random variable with pdf f which takes values in χ and let Y be another random variable with pdf q that takes values in Ω, let h be a function defined on χ and suppose that f, q, and h satisfy the following condition

$$\text{support}(f) \cap \text{support}(h) \subseteq \text{support}(q).$$

Then we can change the underlying distribution by instead estimating the integral on the right hand side of the following equality,

$$E_f[h(X)] = \int_{\chi} h(x)\frac{f(x)}{q(x)}q(x) = E_q\left[h(X)\frac{f(X)}{q(X)}\right]$$

Thus given a sample $(Y_1, \ldots, Y_n)$ from Y we can use the following estimator, for the integral $E_f[h(X)]$,

$$v = \frac{1}{n} \sum_{i=1}^{n} \frac{f(Y_i)}{q(Y_i)} h(Y_i).$$

In order to see the utility of the method lets return to the problem we examined in last example.

Example

Suppose $X \sim N(0,1)$ and $h = 1_{2.6<X}$. Let Y be exponential distribution truncated at 2.6 and let g be the pdf of Y, so $g(x) = \exp(-(x-2.6))$. Now let's generate a sample of size 10^4 from Y, which we can do by generating $Z \sim \epsilon(1)$ and setting $Y = Z + 2.6$. We then use the sample to evaluate:

$$E_f[1_{X>2.6}(X)] \approx \frac{1}{n} \sum_{i=1}^{n} \frac{f(Y_i)}{g(Y_i)}.$$

Using this procedure I calculated $E_f[1_{X>2.6}](X)] \approx 0.0046815$ which is quite close to the actual value of 0.00466119. Using the symmetry of the normal distribution we get $E_f[1_{|X|>2.6}] \approx 0.009362$. Thus using importance sampling we've produced a much better estimate using a tenth of the sample size.

Limitations

Despite its benefits importance sampling has many limitations. While the estimator using any compatible g will converge in the limit to the desired integral, some choices g may so drastically diminish the efficiency of the estimator to the point that for any N an estimator computed from a sample of size N may deviate from the actual value by an unacceptable margin. An importance sampling estimator derived from a g for which f/g is unbounded will suffer from this problem due to having an unbounded variance when sampled for different N because there will be values of Y for which the associated weight is arbitrarily large.

In order to get an acceptable estimate it is therefore necessary to choose a g for which there exists and m such that $mf \geq g$. Thus the criterion for choosing an acceptable g is essentially the same as for choosing a g under the accept reject method.

The central limit theorem tells us that the variations of $(\hat{h}_n^q - E_f[h(X)])$ are asymptotically Gaussian if f/q is bounded.[2]

Conclusion

There are several additional central methods that fall under the heading of Monte Carlo methods, however, they require mathematical knowledge well beyond the prerequisites for this class. One of these methods, optimization, can

be briefly summarized as calculating the maximum of a bounded function h on a bounded domain by taking a sample from the uniform on the domain and taking the max value in the image of the sample under h. However, this is a terribly slow and inefficient method and gains in efficiency require an understanding of multivariable calculus. Thus the technique of optimization could not be satisfactorily developed in this paper. The results presented above only represent a small subset of the techniques used in Monte Carlo experiments, but should provide an adequate understanding of how most basic experiments work and the central idea of the Monte Carlo simulation.

References

[1] Paul Glasserman. *Monte Carlo methods in financial engineering.* Springer Science + Business Media, New York, 2004.

[2] W.S. Kendall, J.M. Marin, and C.P. Robert. Confidence bands for brownian motion and applications to monte carlo simulation. *Statistics and Computing*, 17(1):1–10, March 2007.

[3] Christian P. Robert and George Casella. *Introducing Monte Carlo methods with R.* Use R! Springer, New York, 2010.

[4] Christian P. Robert and George Casella. *Monte Carlo statistical methods.* Springer, New York, 2nd edition, 2010. Christian P. Robert, George Casella. ill. ; 24 cm. Springer texts in statistics Éd. antérieure: 2004. Bibliogr.: p. [591]-622.