1. MONTE-CARLO METHODS

Consider the following problems of practical interest that one often encounters.

(1) Find the probability of an event. For example, it could be that X, Y are independent $N(0,1)$ random variables and we need an approximate numerical value of the probability that $X\cos(Y+X^2) - \tan(X) > 1$.

(2) Find the integral of a function. For example, we may want to find a numerical approximation for $\int_0^1 \frac{\cos x}{1+x^4} dx$.

In this section, we describe how we can solve these questions provided we know how to simulate random numbers (which is why we spent time learning how to do that!). The justification for the method comes from the weak law of large numbers (WLLN), which will show that if take simulate our random numbers infinitely many times, then we can get the exact answer. A better justification is from Chebyshev's inequality, which will show us the extent and probability of our error if we sample a large, but finitely many random numbers.

The problem of integration: Let $\varphi : [0,1] \to \mathbb{R}$ be a continuous function. We would like to compute $I = \int_0^1 \varphi(x)dx$. Most often we cannot compute the integral explicitly and for an approximate value we resort to numerical methods. Here is an idea to use random numbers.

Let $U_1, U_2, \ldots, U_n$ be i.i.d. Unif$[0,1]$ random variables and let $X_1 = \varphi(U_1), \ldots, X_n = \varphi(U_n)$. Then, X_k are i.i.d. random variables with common mean and variance

$$\mu = \int_0^1 \varphi(x)dx = I, \qquad \sigma^2 = \text{Var}(X_1) = \int_0^1 (\varphi(x) - I)^2 dx.$$

This gives the following method of finding I. Fix a large number N (appropriately) and pick N uniform random numbers U_k for $1 \le k \le N$. Then, define

$$\hat{I}_N := \frac{1}{N}\sum_{k=1}^{N} \varphi(U_k).$$

Present $\hat{I}_N$ as an approximate value of I.

In what sense is this an approximation of I, and why? Indeed, by the WLLN we have $\mathbf{P}\{|\hat{I}_N - I| \ge \delta\} \to 0$, and hence we expect $\hat{I}_N$ to be close to I. How large should N be? For this, we fix two numbers $\epsilon = 0.01$ and $\delta = 0.001$ (you may change the numbers). By Chebyshev's inequality, observe that $\mathbf{P}\{|\hat{I}_N - I| \ge \delta\} \le \sigma^2/N\delta^2$.

First find N so that $\sigma^2/N\delta^2 < \epsilon$, i.e., $N = \lceil \frac{\sigma^2}{\delta^2} \rceil$. Then, the random variable $\hat{I}_N$ is within δ of I with probability greater than $1-\epsilon$. This is a probabilistic method, hence there is a possibility of large error, but with a small probability. Observe that N grows proportional to *square* of $1/\delta$. To increase the accuracy by 10, you must increase the number of samples by a factor of 100.

One last point. To find N we need σ^2 which involves computing another integral involving φ which we do not know how to compute! Here we do not need the exact value of the integral. For example, if our function satisfies $-M \leq \varphi(x) \leq M$ for all $x \in [0,1]$, then also $-M \leq I \leq M$ and hence $(\varphi(x) - I)^2 \leq 4M^2$. This means that $\sigma^2 \leq 4M^2$. Therefore, if we take $N = \lceil \frac{4M^2}{\delta^2} \rceil$, then the value of N is larger than required for the desired accuracy. We can work with this N. Note that the dependence of N on δ does not change.

Exercise 1. We know that $\int_0^1 \frac{1}{1+x^2} dx = \frac{\pi}{4}$. Based on this, devise a method to find an approximate value of π. Use the R software to implement your method, and see how many sample you need to get an approximation to 1, 2 and 3 decimal places consistently (consistently means with a large enough probability, say 0.9).

Exercise 2. Devise a method to approximate e and π (there are many possible integrals).

Check the files 'e_estimation.R' and 'pi_estimation.R'.

This method can be used to evaluate integrals over any interval. For instance, how would you find $\int_a^b \varphi(t)dt$, $\int_0^\infty \varphi(t)e^{-t}dt$ or $\int_{-\infty}^\infty \varphi(t)e^{-t^2}dt$, where φ is a function on the appropriate interval? It can also be used to evaluate multiple integrals (and consequently to find the areas and volumes of sets). The only condition is that it should be possible to evaluate the given function φ at a point x on the computer. To illustrate, consider the problem of finding the area of a region $\{(x,y) : 0 \leq x, y, \leq 1,\ 2x^3y^2 \geq 1,\ x^2 + 2y^2 \leq 2.3\}$. It is complicated to work with such regions analytically, but given a point (x,y), it is easy to check on a computer whether all the constraints given are satisfied.

As a last remark, how do Monte-Carlo methods compare with the usual numerical methods? In the latter, usually a number N and a set of points $x_1, \ldots, x_N$ are fixed along with some weights $w_1, \ldots, w_N$ that sum to 1. Then, one presents $\tilde{I}_N := \sum_{k=1}^N w_k \varphi(x_k)$ as the approximate value of I. Lagrange's method, Gauss quadrature, etc. are of this type. Under certain assumptions on φ, the accuracy of these integrals can be like $1/N$ as opposed to $1/\sqrt{N}$ in Monte-Carlo. But, when those assumptions are not satisfied, $\tilde{I}_N$ can be way off I. One may regard this as a game of strategy as follows.

I present a function φ (say bounded between -1 and 1) and you are expected to give an approximation to φ. Quadrature methods do a good job generically, but if I knew the procedure you use, then I can give a function for which your result is entirely wrong (for example, I pick a function φ which vanishes at each of the quadrature points!). However, with Monte-Carlo methods, even if I know the procedure, there is no way to prevent you from getting an approximation of accuracy $1/\sqrt{N}$. This is because neither of us know where the points U_k will fall!

Estimating probabilities: We now turn to the problem of estimating probabilities. Let us take the same example that was given earlier, of estimating the probability of the event E that $X\cos(Y + X^2) - \tan(X) > 1$, where X, Y are i.i.d. $N(0,1)$ random variables.

In principle, this can be done by finding the distribution of $X\cos(Y + X^2) - \tan(X)$ (using change of variables), but in practice that is difficult to execute (the required integrals are hard to calculate explicitly). A versatile method is to use the idea that $\mathbf{P}(E)$ is the proportion of times the event E occurs in a long series of independent trials (in other words, the law of large numbers!). This is how we do that.

Sample $X_1, Y_1, X_2, Y_2, \ldots, X_n, Y_n$ independently on a computer using some random number generator. For each $1 \le k \le n$, calculate the value of $Z_k := X_k \cos(Y_k + X_k^2) - \tan(X_k)$ and set $\xi_k = 1$ if $Z_k > 1$ and $\xi_k = 0$ if $Z_k \le 1$. Then, ξ_k are i.i.d. Ber(p) random variables with $p = \mathbf{P}\{E\}$. Therefore, by the law of large numbers, $\hat{p}_n := \frac{1}{n}(\xi_1 + \cdots + \xi_n)$ is close to p for large n. Hence, we present $\hat{p}_n$ as an approximation to p.

To make a quantitative statement, observe that $\mathbf{E}[\hat{p}_n] = p$ and $\text{Var}(\hat{p}_n) = \frac{1}{n}\text{Var}(\xi_1) = \frac{p(1-p)}{n}$. Since $p(1-p) \le \frac{1}{4}$ (true for all $0 < p < 1$), using Chebyshev's inequality, we see that

$$\mathbf{P}\left\{|\hat{p}_n - p| \ge \delta\right\} \le \frac{1}{4n\delta^2}.$$

What n to choose? We first decide the accuracy δ that we want. Since our estimate is based on random numbers, there is always a chance of making a mistake, for example, if all ξ_ks turn out to be 0, then our estimate $\hat{p}_n = 0$. In other words, there is no guarantee in this method that we will get $\hat{p}_n$ within δ of p. Hence, we also fix a probability of error, say ϵ. Then, what we want is

$$\frac{1}{4\delta^2 n} \le \epsilon \quad \text{or, equivalently} \quad n \ge \frac{1}{4\delta^2\epsilon}.$$

For example, if we set $\delta = 0.05$ and $\epsilon = 0.01$, then we get $n > 10000$. Observe the dependence of n on δ and ϵ. If we halve δ, then n must increase 4 times, while if we halve ϵ, we only need an increase in n by a factor of 2.

The versatility of the method is that any and every probability one cares about can be estimated this way. The only thing needed is that we should be able to simulate the event in reasonable time. The situations where one can explicitly calculate probabilities by analytical methods are far and few.

Example 3. Let us return to the birthday paradox. Analytical calculation shows that the probability of the event of having at least two people with the same birthday in a group of 23 people is 0.5073. Here, are the estimates we got from simulations.

A single simulation consists of drawing 23 random numbers from $1, 2, \ldots, 365$, and checking to see if there are two equal numbers among them. We did this 50 times and took the proportion of times we had a coincident birthday. The proportion turned out to be 0.52.

Then, we repeated the whole procedure 10 times. Each time we get a different estimate. They turned out to be $0.50, 0.52, 0.52, 0.46, 0.52, 0.50, 0.48, 0.44, 0.46, 0.56$.

Exercise 4. For $n = 20, 40, \ldots, 100$, find the probability that a random group of n people will have three coincident birthdays. Can you find a value of n such that this probability is close to $\frac{1}{2}$?

2. Convergence in Distribution

Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables with corresponding sequence of distribution functions (DFs) as $\{F_n\}_{n\geq 1}$. Define

$$F(x) = \lim_{n\to\infty} F_n(x) \; \forall \; x \in \mathbb{R}.$$

Let us look at the following examples.

Example 5. Let $X_n \sim U(-n, n)$ for $n = 1, 2, \ldots$ Then, the DF of X_n is

$$F_n(x) = \begin{cases} 0, & \text{if } x < -n \\ \frac{x+n}{2n}, & \text{if } -n \leq x < n, \\ 1, & \text{if } x \geq n. \end{cases}$$

Clearly, $F(x) = \lim_{n\to\infty} F_n(x) = \frac{1}{2} \; \forall \; x \in \mathbb{R}$ and F is *not* a DF. (Why?)

Example 6. Let $X_n \sim N\left(0, \frac{1}{n}\right)$ for $n = 1, 2, \ldots,$ and let X be a random variable degenerate at 0 (i.e., $\mathbf{P}(\{X = 0\}) = 1$). Then, for $x \in \mathbb{R}$,

$$F(x) = \mathbf{P}(\{X \leq x\}) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{if } x \geq 0. \end{cases}$$

$$\begin{aligned} F_n(x) &= \mathbf{P}\left(\{X_n \leq x\}\right) \\ &= \Phi(\sqrt{n}x), \text{ since } \sqrt{n}X_n \sim N(0,1) \\ \rightarrow F(x) &= \begin{cases} 0, & \text{if } x < 0 \\ \frac{1}{2}, & \text{if } x = 0 \\ 1, & \text{if } x > 0, \text{ as } n \to \infty. \end{cases} \end{aligned}$$

Clearly, F is *not* a DF. (Why?)

The above examples illustrate that a sequence $\{F_n\}_{n\geq 1}$ may converge at all points, but the limiting function $F(x) = \lim_{n\to\infty} F_n(x)$ for $x \in \mathbb{R}$ may *not* be a DF. To fix ideas, consider the *simple* example.

Example 7. Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables with $\mathbf{P}\left(\left\{X_n = \frac{1}{n}\right\}\right) = 1$ for $n = 1, 2, \ldots$ Then, the DF of X_n is

$$F_n(x) = \begin{cases} 0, & \text{if } x < \frac{1}{n}, \\ 1, & \text{if } x \geq \frac{1}{n}. \end{cases}$$

Clearly,

$$F(x) = \lim_{n\to\infty} F_n(x) = \begin{cases} 0, & \text{if } x \leq 0 \\ 1, & \text{if } x > 0. \end{cases}$$

So, F is not right continuous at $x = 0$, and hence *not* a DF!

However, F can be easily converted to a distribution function

$$F^*(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1, & \text{if } x \geq 0, \end{cases}$$

by changing its value at the point 0 (the point of discontinuity of F).

Since $\mathbf{P}\left(\left\{X_n = \frac{1}{n}\right\}\right) = 1$ for $n = 1, 2, \ldots$ and $\lim_{n\to\infty} \frac{1}{n} = 0$, a natural approximation of F_n seems to be the DF of the random variable degenerate at 0 (i.e., $\mathbf{P}(X = 0) = 1$). Note that F^* is the DF of this random variable, which is degenerate at 0. The above discussion suggests that it is too restrictive to require

$$\lim_{n\to\infty} F_n(x) = F^*(x) \; \forall \; x \in \mathbb{R},$$

and exceptions should be permitted at the points of discontinuities of F^*.

Definition 8. The sequence $\{X_n\}_{n\geq 1}$ is said to converge in distribution to X (written as $X \xrightarrow{D} X$) as $n \to \infty$ if

$$\lim_{n\to\infty} F_n(x) = F(x) \; \forall \; x \in C_F,$$

where C_F is the set of continuity points of F.

Remark 9. Some points:

- The set C_F is dense in $\mathbb{R}$.
- If F (the limiting DF) is continuous (i.e., $C_F = \mathbb{R}$), then $\lim_{n\to\infty} F_n(x) = F(x) \; \forall \; x \in \mathbb{R}$.
- Let f and f_n be the pmfs (with *integer* support set) of X and X_n, respectively, for $n = 1, 2, \ldots$ Then,

$$f_n(x) \to f(x) \; \forall \; x \implies X_n \xrightarrow{D} X \text{ as } n \to \infty.$$

- Let f and f_n be the pdfs of X and X_n, respectively, for $n = 1, 2, \ldots$ Then,

$$f_n(x) \to f(x) \; \forall \; x \implies X_n \xrightarrow{D} X \text{ as } n \to \infty.$$

Continuity theorem:
Suppose that there exists $a > 0$ such that the mgfs M and M_n of X and X_n, respectively, are finite on $(-a, a)$ for all $n = 1, 2, \ldots$ If $\lim_{n\to\infty} M_n(t) = M(t) \; \forall \; t \in (-a, a)$, then $X_n \xrightarrow{D} X$ as $n \to \infty$.

Remarks:

(1) If $X_n \xrightarrow{D} X$, then $h(X_n) \xrightarrow{D} h(X)$ for any continuous function h.

(2) $X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$, but the *reverse implication is not true* in general.

Example 10. Consider the $U(0,1)$ distribution. Define $X_n = \mathbf{1}_{[0,1/2+1/n]}$ for $n = 1, 2, \ldots$ and $X = \mathbf{1}_{[1/2,1]}$. Indeed,

$$F_n(t) = \begin{cases} 0 & t < 0 \\ 1/2 - 1/n & 0 \le t < 1 \\ 1 & t \ge 1, \end{cases}$$

converges pointwise to

$$F_X(t) = \begin{cases} 0 & t < 0 \\ 1/2 & 0 \le t < 1 \\ 1 & t \ge 1. \end{cases}$$

On the other hand, $|X_n - X| = \mathbf{1}_{[0,1/2]\cup[1/2+1/n,1]}$ (Check!). So, $\mathbf{P}\{|X_n - X| > 1/2\} = 1/2 + (1 - 1/2 - 1/n) = 1 - 1/n \to 1$ as $n \to \infty$, and there is no convergence in probability!

(3) $X_n \xrightarrow{P} c$ if and only if $X_n \xrightarrow{D} c$ for $c \in \mathbb{R}$.

Slutsky's Theorem: If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} c \in \mathbb{R}$, then $X_n + Y_n \xrightarrow{D} X + c$ and $X_n Y_n \xrightarrow{D} cX$.

3. CENTRAL LIMIT THEOREM

Let $X_1, X_2, \ldots$ be i.i.d. random variables with expectation μ and variance σ^2. We saw that $\bar{X}_n$ has mean μ and standard deviation $\sigma/\sqrt{n}$.

WLLN roughly means that $\bar{X}_n$ is close to μ, within a few multiples of $\sigma/\sqrt{n}$ (as shown by Chebyshev's inequality). Now, we look at $\bar{X}_n$ with a finer microscope. In other words, we ask for the probability that $\bar{X}_n$ is within the tiny interval $\left[\mu + \frac{a}{\sqrt{n}}, \mu + \frac{b}{\sqrt{n}}\right]$ for any $a < b$. The answer turns out to be surprising and remarkable!

Central limit theorem: Let $X_1, X_2, \ldots$ be i.i.d. random variables with expectation μ and variance σ^2. We assume that $0 < \sigma^2 < \infty$. Define $Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$, and let Z be a $N(0,1)$ random variable. Then,

$$Z_n \xrightarrow{D} Z \text{ as } n \to \infty.$$

Proof: Follows from the continuity theorem of mgfs, but we will not discuss it here.

Let Φ be the cdf of $N(0,1)$. For any $a < b$, we have

$$\mathbf{P}\left\{\mu + a\frac{\sigma}{\sqrt{n}} \le \bar{X}_n \le \mu + b\frac{\sigma}{\sqrt{n}}\right\} \to \Phi(b) - \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt \text{ as } n \to \infty.$$

What is remarkable about this? The end result does not depend on the distribution of X_is at all! Only the mean and variance of the distribution were used! As this is one of the most important theorems in all of probability theory, we restate it in several forms, all equivalent to the above.

Restatements of central limit theorem: Let $S_n = X_1 + \cdots + X_n$. Let Z be a $N(0,1)$ random variable. Then, of course $\mathbf{P}\{a < Z < b\} = \Phi(b) - \Phi(a)$.

(1) $\mathbf{P}\{a < \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \le b\} \to \Phi(b) - \Phi(a) = \mathbf{P}\{a < Z < b\}$ as $n \to \infty$. Put another way, this says that for large n, the random variable $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ has $N(0,1)$ distribution, approximately.

(2) Yet another way to say the same is that S_n has approximately normal distribution with mean $n\mu$ and variance $n\sigma^2$. That is,

$$\mathbf{P}\left\{a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\right\} \to \mathbf{P}\{a < Z < b\} \text{ as } n \to \infty.$$

The central limit theorem is deep, surprising and useful. The following example gives a hint as to why.

Example 11. Let $U_1, \ldots, U_n$ be i.i.d. Uniform($[-1, 1]$) random variables. Let $S_n = U_1 + \cdots + U_n$, let $\bar{U}_n = S_n/n$ (sample mean) and let $Y_n = S_n/\sqrt{n}$. Consider the problem of finding the distribution

of any of these. Since they are got from each other by scaling, finding the distribution of one is the same as finding that of any other. For uniform $[-1, 1]$, we know that $\mu = 0$ and $\sigma^2 = 1/3$. Hence, CLT tells us that

$$\mathbf{P}\left\{\frac{a}{\sqrt{3}} < Y_n < \frac{b}{\sqrt{3}}\right\} \to \Phi(b) - \Phi(a) \text{ as } n \to \infty.$$

Equivalently, $\mathbf{P}\{a < Y_n < b\} \to \Phi(b\sqrt{3}) - \Phi(a\sqrt{3})$ as $n \to \infty$. For large n (practically, $n = 50$ is large enough) we may use this limit as a good aproximation to the probability we want.

Why is this surprising? The way to find the distribution of Y_n would be this. One can find the density of $S_n = U_1 + \cdots + U_n$ (in principle! the actual integration may be intractable!). Then we can find the density of Y_n by another change of variable (in one dimension). Having got the density of Y_n, we integrate it from a to b to get $\mathbf{P}\{a < Y_n < b\}$. This is clearly a daunting task (if you don't feel so, just try it for $n = 5$).

The CLT cuts short all this and directly gives an approximate answer! And what is even more surprising is that the original distribution does not matter - we only need to know the mean and variance of the original distribution!

Check the file 'CLT.R'.