

Probabilities and Statistics - Lecture 10

Olariu E. Florentin
May, 2017

1 Computer simulation and Monte Carlo (MC) Methods

Estimating expectation with MC method

Estimating lengths, areas and volumes

Estimating areas of regions with unknown boundaries

Monte Carlo integration

Estimating probabilities with MC method

2 Bibliography

- The process of generating random values from a density is called *simulation* (some call it *Monte Carlo simulation*).
Statistics and Data with R by Y. Cohen, J. Y. Cohen
- *Monte Carlo method* is any method which solves a problem by generating suitable random numbers and observing that fraction of the numbers obeying some property or properties. The method is useful for obtaining numerical solutions to problems which are too complicated to solve analytically.
mathworld.wolfram.com
- A value of a random variable (or a value from a density) is called a *quantile* or a *variate*.

- A *Monte Carlo method* generates many (sometimes millions) of such variates associated with a probability distribution and the process is called the simulation of that distribution.
- The simulation is used for finding the expectation, or the variance of a distribution, or another associated parameter.
- The so-called simulation depends on the "quality" of the variates. The most used random numbers are variates from standard uniform continuous distribution, $U(0, 1)$, or from a discrete uniform distribution, U_n .
- Almost every programming language has such a *random numbers generator*, but these generators give only pseudorandom or quasirandom numbers (uniform variates).
- The most widely used pseudorandom number generator (PRNG) is Mersenne-Twister (the default in R).

Estimating expectation with MC method

- Suppose that we have a random variable X and we want to estimate its expectation $\mu = \mathbb{E}[X]$.
- We first generate a Monte Carlo sequence of random variables (the variates) X_1, X_2, \dots, X_N independent and identically distributed (i. i. d.) with X . An unbiased estimator of μ is

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N},$$

because $\mathbb{E}[\bar{X}] = \mu$. If $\text{Var}[X] = \sigma^2$, then

$$\text{Var}[\bar{X}] = \frac{\sum_{i=1}^N \text{Var}[X_i]}{N^2} = \frac{\sigma^2}{N}.$$

Estimating expectation with MC method - Example

- **Example.** A retailer sells a perishable commodity and each day he places an order for 100 units. Each unit that is sold gives a profit of 55 cents and units not sold at the end of the day are discarded at a loss of 40 cents per unit. The demand, X , on any given day is uniformly distributed on $[80, 140]$. Estimate the expected profit.

- **Solution.** If P is the profit, then

$$P = \begin{cases} 55, & \text{if } X \geq 100 \\ 0.55X - 0.4(100 - X), & \text{if } X < 100 \end{cases}$$

- We generate N values for X , and compute P_1, P_2, \dots, P_N , then take the average (the sample mean).
- For five independent samples with $N = 10000$ we get

51.7796 51.82632 51.87036 51.84095 51.88509

Estimating expectation with MC method - Example

- The exact value of the expected profit is

$$\int_{80}^{100} \frac{0.95x - 40}{60} dx + \int_{100}^{140} \frac{55}{60} dx = 51.83333$$

- **Example.** A very powerful computer is used by 250 independent subscribers. Each day, each subscriber, independently, uses the computer with probability 0.3. The number of tasks sent by each active user has Geometric distribution with parameter 0.15, and each task takes a $\Gamma(10, 3)$ distributed computer time (in minutes). Tasks are processed consecutively. Estimate the expectation of the total requested time.
- **Solution.** The total requested time $T = T_1 + \dots + T_X$ consists of times T_i requested by X active users. The number of active users X is $Binomial(250, 0.3)$.

Estimating expectation with MC method - Example

- Each of the active users sends a *Geometric*(0.15) number of tasks Y_i . Thus, each $T_i = T_{i,1} + \dots + T_{i,Y_i}$, where $T_{i,j}$ is $\Gamma(10, 3)$.

- Three independent estimations give the following probabilities

1494.901 1492.228 1489.696

- All these values are just above the 24-hours period (1440 minutes).

- Example.** Two web servers deliver the same pages to web clients. The time to process a HTTP request has an exponential distribution with $\lambda_1 = 0.03\text{ms}^{-1}$ for the first server and $\lambda_2 = 0.04\text{ms}^{-1}$ for the second. The roundtrip latency contains also the time the request and the response travel through Internet which has an exponential distribution with $\lambda = 1\text{ms}^{-1}$.

Estimating expectation with MC method - Example

- **Example cont'd.** It is known that a client is directed to the first server with probability 0.4, and to the second server with probability 0.6. Estimate the average time a client has to wait until it receives a response to its request.
- **Solution.** A simulation (or a run) for this problem consists in generating first a standard uniform variate U , then depending on its value generating an exponential variate with $\lambda = 0.03$ or 0.04 ; the result is added to an exponential variate with $\lambda = 1$:

$$T = X + \begin{cases} Y, & \text{if } U < 0.4 \\ Z, & \text{if } U \geq 0.4 \end{cases},$$

where $U : U(0,1)$, $X : Exp(1)$, $Y : Exp(0.03)$, $Z : Exp(0.04)$.

From $N = 10000$ of runs we get 29.48822 ms.

- Let U be a standard uniform variable; U belongs to a set $A \subseteq [0, 1]$ with probability

$$P(U \in A) = \int_A 1 \, du = \text{length of } A.$$

- Let $X = \chi_A$ be the indicator (characteristic) function of bset A and X_1, X_2, \dots, X_N a sequence of random variables i. d. with X .

$$X(u) = \chi_A(u) = \begin{cases} 1, & u \in A \\ 0, & \text{otherwise} \end{cases}$$

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N},$$

- The sequence (X_i) can be obtained by generating a sequence of independent standard uniform variables U_1, U_2, \dots, U_N , and $X_i = \chi_A(U_i)$.
- The length of A is approximatively \bar{X} which is the proportion of U_i that fall into A .
- Let $A \subseteq [a, b]$; if U is the uniform variable on $[a, b]$, then

$$P(U \in A) = \int_A 1 \, du = (b - a) \int_A \frac{1}{b - a} \, du = \text{length of } A.$$

- We generate a sequence of independent uniform variables on $[a, b]$: U_1, U_2, \dots, U_N ($X_i = \chi_A(U_i)$). The length of A is approximatively $(b - a) \cdot \bar{X}$ which is the proportion of U_i that fall into A times $(b - a)$.
- Of course computing lengths doesn't represent a real problem; however the method can be used to estimate areas and volumes.

- Let B be a two-dimensional set that lies in $[0, 1] \times [0, 1]$; two independent standard uniform variables have a joint density

$$f_{U,V}(u, v) = \begin{cases} 1, & (u, v) \in B \\ 0, & \text{otherwise} \end{cases}$$

- Now, the area of B is

$$P((U, V) \in B) = \iint_B 1 \, dudv.$$

- We give an algorithm for estimating the area of $B \subseteq [0, 1]^2$:

- 1 Generate an even number of independent standard uniform variables: $U_1, \dots, U_N, V_1, \dots, V_N$;
- 2 Let N_B be the number of pairs (U_i, V_i) that belongs to B .
- 3 Estimate the area of B by N_B/N .

- Let B be a two-dimensional set that lies in $[a, b] \times [a, b]$; two independent uniform variables on $[a, b]$ have a joint density

$$f_{U,V}(u, v) = \begin{cases} 1/(b-a)^2, & (u, v) \in B \\ 0, & \text{otherwise} \end{cases}$$

- Now, the area of B is

$$P((U, V) \in B) = \iint_B 1 \, dudv = (b-a)^2 \iint_B \frac{1}{(b-a)^2} \, dudv.$$

- An algorithm for estimating the area of $B \subseteq [a, b]^2$ is:

- 1 Generate an even number of independent uniform variables on $[a, b]$: $U_1, \dots, U_N, V_1, \dots, V_N$;
- 2 Let N_B be the number of pairs (U_i, V_i) that belongs to B .
- 3 Estimate the area of B by $(b-a)^2 \cdot N_B/N$.

- **Example 1.** Let B be the unit disc in the real plane:

$$B = \{(u, v) : u^2 + v^2 \leq 1\} \subseteq [-1, 1]^2.$$

- We generate $N = 10000$ random independent uniform values (or variates) on $[-1, 1]$ (in R we use `runif(1, -1, 1)` 10000 times or `runif(10000, -1, 1)`).
- We get an estimate of 3.1368 for the area of this disc which is $\pi = 3.14159$.

- **Example 2.** Let B be an ellipse ($a = 4, b = 3$):

$$B = \{(u, v) : u^2/a^2 + v^2/b^2 \leq 1\} \subseteq [-4, 4] \times [-3, 3] \subseteq [-4, 4]^2.$$

- We generate $N = 10000$ random independent pairs of uniform values (or variates) on $[-4, 4]$.
- We get an estimate of 37.4528 for the area of this ellipse which is $\pi ab = 12\pi = 37.69911$.

- Let C be a three-dimensional set that lies in $[a, b] \times [a, b] \times [a, b]$; three independent uniform variables on $[a, b]$ have a joint density

$$f_{U,V,W}(u, v, w) = \begin{cases} 1/(b-a)^3, & (u, v, w) \in C \\ 0, & \text{otherwise} \end{cases}.$$

- Now, the volume of C is

$$P((U, V, W) \in C) = \iiint_C 1 \, dudvdw = (b-a)^3 \iiint_C \frac{dudvdw}{(b-a)^3}$$

- An algorithm for estimating the volume of $C \subseteq [a, b]^3$ is:

- Generate a number multiple of 3 of independent uniform variables on $[a, b]$: $U_1, \dots, U_N, V_1, \dots, V_N, W_1, \dots, W_N$.
- Let N_C be the number of triplets (U_i, V_i, W_i) that belongs to C .
- Estimate the volume of C by $(b-a)^3 \cdot N_C/N$.

- Let us estimate the volume of the unit ball¹:

$$C = \{(u, v, w) : u^2 + v^2 + w^2 \leq 1\} \subseteq [-1, 1]^3.$$

- We generate $N = 10000$ random independent triplets of uniform values (or variates) on $[-1, 1]$, and we get an estimate of 4.184 for the volume of this ball which is $4\pi/3 = 4.18879$.
- Second we generate $N = 50000$ random independent triplets of uniform values (or variates) on $[-1, 1]$, and we get an estimate of 4.18816 for the volume of the unit sphere.
- As the number of dimensions increases we need more variates to get good approximations for our parameter.
- This is the *curse of dimensionality* which become visible when working in high-dimensional spaces.

¹Usually by sphere we understand only the boundary of the given set.

Estimating volumes - Example

- Let estimate the volume of the 8-dimensional unit ball (which has the volume equal with $\pi^4/24 = 4.058712$):

$$C = \left\{ (u_1, \dots, u_8) : \sum_{i=1}^8 u_i^2 \leq 1 \right\} \subseteq [-1, 1]^8.$$

- The following table contains the estimates for various length of the sequences for five different MC simulations:

| run | $N = 1000$ | $N = 20000$ | $N = 50000$ | $N = 100000$ |
|----------------|------------|-------------|-------------|--------------|
| 1. | 2.816 | 3.3920 | 4.11136 | 3.99872 |
| 2. | 4.096 | 4.1600 | 4.01408 | 3.98592 |
| 3. | 3.584 | 4.3776 | 4.06528 | 4.04992 |
| 4. | 3.328 | 4.0704 | 4.2496 | 4.13440 |
| 5. | 4.864 | 3.6480 | 4.22912 | 4.00896 |
| average | 3.7376 | 3.9296 | 4.133888 | 4.035584 |
| absolute error | 0.321112 | 0.129112 | 0.075176 | 0.023128 |

Estimating areas of regions with unknown boundaries

- In order to approximate areas or volumes by MC methods, knowing exact boundaries is not necessary.
- To apply one of the above algorithms it is sufficient to know how to determine if a given point belongs to the involved set (for which we measure area, volume etc).
- Thus, it is not required that the sample region has a rectangular shape; with different scales along the axes, random points may be generated on a rectangle or even a more complicated set.
- One way to generate random point in a region of arbitrary shape is to draw a larger rectangular shape set around it and generate uniformly distributed coordinates until the corresponding point belongs to the region.

Estimating areas of regions with unknown boundaries - Example

- **Example.** An emergency is reported at a nuclear power plant. It is necessary to assess the size of the region exposed to the radioactivity. Boundary of the region cannot be determined, however, the level of radioactivity can be measured at any given location.
- **Solution.** A rectangle of 10×8 km is chosen around the exposed area. Pairs of uniform random numbers (U_i, V_i) are generated (that is random points in the covering rectangle).
- The radioactivity is measured at all the obtained random locations. The area is then estimated as the proportion of measurements above the normal level multiplied by the area of the sampling rectangle.
- Suppose that radioactivity is measured at 50 random sites, and it is found above the normal levels at 18 locations. The exposed area is then estimated as $\frac{18}{50} \cdot 80 \text{ km}^2 = 28.8 \text{ km}^2$.

- A length, an area or a volume can be viewed as a definite integral from a certain function.
- Hence we can extend the Monte Carlo method to definite integrals. Suppose we want to integrate a certain function h from a to b :

$$H = \int_a^b h(u) du.$$

- We can approximate this integral by averaging samples of h at random uniform variates within the interval $[a, b]$.
- If U_1, U_2, \dots, U_n are independent uniform random variables on $[a, b]$ (for which the density function is $1/(b-a)$ on this interval and 0 outside), the Monte Carlo estimator for F is

$$F_N = \frac{b-a}{N} \sum_{i=1}^N h(U_i).$$

- This is because, for an uniform random variable, U , on $[a, b]$, the expectation of $f(U)$ is

$$\mathbb{E}[h(U)] = \int_a^b h(u)f(u) du,$$

where f is the density of the uniform distribution on $[a, b]$.

- Hence

$$\mathbb{E}[h(U)] = \int_a^b h(u) \frac{1}{b-a} du,$$

and

$$H = \int_a^b h(u) du = (b-a)\mathbb{E}[h(U)].$$

- By using the Monte Carlo estimation of the above expectation we get

$$H \approx \frac{b-a}{N} \sum_{i=1}^N h(U_i) = F_N,$$

for a sequence of independent uniform random variables on $[a, b]$ (II.)

- From the (Strong) Law of Large Numbers $P\left(\lim_{N \rightarrow \infty} F_N = H\right) = 1$; the variance of this estimator is

$$\text{Var}[F_N] = \frac{(b-a)^2}{12N} = \mathcal{O}(1/N),$$

as the variance of the uniform distribution on $[a, b]$ is $(b-a)^2/12$.

- Since the standard deviation is a measure of spread, the former relation can be read like this: we must quadruple the number of samples in order to reduce the error (the standard deviation) by half.

Monte Carlo integration - Example

- Let's try to estimate the following (improper) integral:

$$\int_0^{\infty} e^{-u^2/2} du,$$

(we already know that $\int_0^{\infty} e^{-u^2/2} du = \sqrt{\pi/2} = 1.253314$).

- First observe that $\lim_{a \rightarrow \infty} \int_0^a e^{-u^2/2} du = \int_0^{\infty} e^{-u^2/2} du$, hence

for large enough value of a we have $\int_0^{\infty} e^{-u^2/2} du \approx \int_0^a e^{-u^2/2} du$.

Let us choose $a = 10$.

- For different values of N we get the following averages after 30 independent estimates.

| | $N = 1000$ | $N = 10000$ | $N = 20000$ | $N = 50000$ |
|----------|------------|-------------|-------------|-------------|
| average | 1.247216 | 1.259898 | 1.250592 | 1.251562 |
| st. dev. | 0.08749 | 0.02256 | 0.01898 | 0.01045 |

- The former definite integral can be written like this:

$$H = \frac{1}{b-a} \int_a^b (b-a)h(u) du = \mathbb{E}[(b-a)h(U)],$$

where U has an uniform continuous distribution on $[a, b]$.

- Following the following procedure we can use *any continuous distribution* in place of an uniform one.

- Let X be a random continuous distribution with density f such that $f(u) > 0$, for all $u \in [a, b]$, and $f(u) = 0$ for every $u \notin [a, b]$.

- We can write

$$H = \int_a^b h(x) dx = \int_a^b \frac{h(x)}{f(x)} f(x) dx = \mathbb{E} \left[\frac{h(X)}{f(X)} \right].$$

- Thus we can estimate H by choosing N variates of X (X_1, \dots, X_N) and computing the following average:

$$H \approx \frac{1}{N} \sum_{i=1}^N \frac{h(X_i)}{f(X_i)}.$$

- The above method is not limited to a finite interval $[a, b]$. We can approximate in this way improper (but convergent) integrals.

- We can estimate over an interval $(a, b) \subseteq \mathbb{R}$ the only requirement is that the support of f , i. e., $\text{supp}(f) = \{x \in \mathbb{R} : f(x) \neq 0\}$ to include (a, b) .

Improved Monte Carlo integration - Example

- For example, by choosing f to be standard normal density we can perform Monte Carlo integration from $-\infty$ to ∞ or, if we choose f to be the exponential density, we can perform MC integration from 0 to ∞ .
- Let's estimate again

$$\int_0^{\infty} e^{-u^2/2} du,$$

this time using the exponential density with $\lambda = 1$ (not an approximation using the upper limit of integration).

| | $N = 1000$ | $N = 10000$ | $N = 20000$ | $N = 50000$ |
|----------|------------|-------------|-------------|-------------|
| average | 1.254416 | 1.254476 | 1.253978 | 1.253035 |
| st. dev. | 0.01454 | 0.00349 | 0.00313 | 0.00176 |

Estimating probabilities with MC method

- Estimating a probability is one of the most typical applications of Monte Carlo method.
- Let X be a real random variable and $A \subseteq \mathbb{R}$; the probability $p = P(X \in A)$ is estimated by

$$\hat{p}_N = \frac{\#\{X_i \in A\}}{N}.$$

- Obviously the number of $X_1, X_2, \dots, X_n \in A$ is a discrete random variable having a binomial distribution ($B(N, p)$).
- The expectation and the variance of \hat{p}_N are

$$\mathbb{E}[\hat{p}_N] = \frac{Np}{N} = p \text{ and}$$

$$\text{Var}[\hat{p}_N] = \frac{Np(1-p)}{N^2} = \frac{p(1-p)}{N}.$$

Accuracy of estimating probabilities with MC method

- How accurate is this method of approximating p by \hat{p}_N (which is an unbiased estimator)?
- Using the normal approximation of the binomial distribution,

$$\frac{N\hat{p} - Np}{\sqrt{Np(1-p)}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/N}} : N(0, 1).$$

- Therefore,

$$\begin{aligned} P(|\hat{p} - p| > \epsilon) &= P\left(\frac{|\hat{p} - p|}{\sqrt{p(1-p)/N}} > \frac{\epsilon}{\sqrt{p(1-p)/N}}\right) \approx \\ &\approx 2\Phi\left(-\frac{\epsilon}{\sqrt{p(1-p)/N}}\right) = 2 \cdot \text{pnorm}\left(-\frac{\epsilon}{\sqrt{p(1-p)/N}}\right), \end{aligned}$$

where $\Phi(\cdot)$ is the distribution function of a standard normal random variable (in R $\Phi(z) = \text{pnorm}(z)$).

Accuracy of estimating probabilities with MC method

- How we design a Monte Carlo study that attains the desired accuracy?
- That is, for given ϵ and $0 < \alpha < 1$, how large must be N such that

$$P(|\hat{p} - p| > \epsilon) \leq \alpha ?$$

- The main obstacle is that the value of p is unknown (otherwise the estimate doesn't make any sense).
- We have two possibilities to estimate the quantity $p(1-p)$:

① First, we can use a "guess" (preliminar estimate) of p , if available.

② Second, we can use an upper bound of the mentioned quantity

$$p(1-p) \leq 1/4, \forall p \in [0, 1].$$

Accuracy of estimating probabilities with MC method

- In the first case, if p^* is the "guess", then we have to solve the inequality

$$2\Phi\left(-\frac{\epsilon}{\sqrt{p^*(1-p^*)/N}}\right) \leq \alpha.$$

- Let $z_a = \Phi^{-1}(a) = \text{qnorm}(a)$, where $a \in (0, 1)$. The above inequality becomes

$$-\frac{\epsilon}{\sqrt{p^*(1-p^*)/N}} \leq z_{\frac{\alpha}{2}} \text{ or } \sqrt{p^*(1-p^*)/N} \leq -\frac{\epsilon}{z_{\frac{\alpha}{2}}}.$$

(Note that, for $a < 1/2$, we have $z_a < 0$.)

- We get a lower bound for N :

$$N \geq p^*(1-p^*) \left(\frac{z_{\frac{\alpha}{2}}}{\epsilon}\right)^2.$$

Accuracy of estimating probabilities with MC method

- In the second case, if we don't have a "guess", then

$$N \geq \frac{1}{4} \left(\frac{z_{\frac{\alpha}{2}}}{\epsilon} \right)^2 = \left(\frac{z_{\frac{\alpha}{2}}}{2\epsilon} \right)^2.$$

Estimating probabilities with MC method - Example

- **Example.** A very powerfull server is used by 250 independent subscribers. Each day, each subscriber, independently, uses the server with probability 0.3. The number of tasks sent by each active user has Geometric distribution with parameter 0.15, and each task takes a $\Gamma(10, 3)$ server time (in minutes). Tasks are processed consecutively. What is the probability that the total requested time is less than 24 hours? Estimate this probability, attaining the margin of error ± 0.01 with probability 0.99.

- **Solution.** The total requested time $T = T_1 + \dots + T_X$ consists of times T_i requested by X active users. The number of active users X is $Binomial(250, 0.3)$.

Estimating probabilities with MC method - Example

- Each of the active users sends a *Geometric*(0.15) number of tasks Y_i . Thus, each $T_i = T_{i,1} + \dots + T_{i,Y_i}$, where $T_{i,j}$ is $\Gamma(10, 3)$.
- We cannot "guess" an estimate of the probability of interest, i. e., $P(T < 24)$. In order to attain the required accuracy ($\alpha = 0.001$, $\epsilon = 0.001$) we need

$$N \geq \frac{1}{4} \left(\frac{z_{\frac{\alpha}{2}}}{\epsilon} \right)^2 = \frac{1}{4} \left(\frac{z_{0.005}}{0.01} \right)^2 = \frac{1}{4} \left(\frac{-2.57529}{0.01} \right)^2 = 16587.24,$$

as $z_{0.005} = qnorm(0.005) = -2.57529$.







- Thus we need $N = 16588$ simulations (which is big enough in order to use the above normal approximation).

Estimating probabilities with MC method - Example

- Three independent estimations give the following probabilities

0.4262117 0.4202435 0.4259103

- The probability we got is not so small, it seems quite probable that all tasks will be completed in a single day.

-  Baron, M., *Probability and Statistics for Computer Scientist*, Chapman & Hall/CRC Press, 2013 or the electronic edition <https://ww2.ii.uj.edu.pl/~z1099839/naukowe/RP/rps-michael-byron.pdf>
-  Johnson, J. L., *Probability and Statistics for Computer Science*, Wiley Interscience, 2008.
-  Lipschutz, S., *Theory and Problems of Probability*, Schaum's Outline Series, McGraw-Hill, 1965.
-  Ross, S. M., *A First Course in Probability*, Prentice Hall, 5th edition, 1998.
-  Shao, J., *Mathematical Statistics*, Springer Verlag, 1998.
-  Stone, C. J., *A Course in Probability and Statistics*, Duxbury Press, 1996.