Advanced Practical 2022/2023
Operations Research Case

Lecture: Monte Carlo Algorithm Basic Simulation Principles

Guanlian Xiao

Department Operations Analytics
Vrije Universiteit Amsterdam

June 2023

Basic Simulation Principles

The Monte Carlo Algorithm

Simulation Model

- The basis of simulation is the static simulation model.
- Also called Monte Carlo simulation.
- A probability model is determined by random variables (or process) $\boldsymbol{X} = (X_1, X_2, \ldots)$ (*input variables*).
- The output of the model is the variable $Y = h(\boldsymbol{X})$ for some response function h
- Y is called *output variable*.
- The goal is to compute the *performance measure*

$$J \doteq \mathbb{E}[Y] = \mathbb{E}[h(\boldsymbol{X})] = \int_{\mathcal{X}} h(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x},$$

where $f(\boldsymbol{x})$ is the pdf of the random vector $\boldsymbol{X}$;

SLLN

When $Y_1, Y_2, \ldots$ are iid replications of Y then

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \stackrel{\text{a.s.}}{\to} \mathbb{E}[Y] \quad (n \to \infty).$$

Note, this is applied to $Y = h(X)$. Thus, it holds that if $X_1, X_2, \ldots$ are iid, then

$$\frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{X}_i) \stackrel{\text{a.s.}}{\to} \mathbb{E}[h(\boldsymbol{X})] \quad (n \to \infty).$$

Estimators

- Consider finitely many iid output replications $Y_1, Y_2, \ldots, Y_n$.
- This is called a *sample* with sample size n.
- Their average is called *sample average estimator*:

$$\overline{Y}_n \doteq \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

- Properties:
  –> $\mathbb{E}\left[\overline{Y}_n\right] = J$; i.e., $\overline{Y}_n$ is an *unbiased estimator* of the performance measure J.

  –> For large n, according to SLLN, $\overline{Y}_n$ is not so random; i.e., it is almost a constant function; i.e., $\overline{Y}_n \approx J$ for almost all samples.

Statistics

Let $\sigma^2 \doteq \mathrm{Var}(Y)$ be the variance of the output variable.

CLT

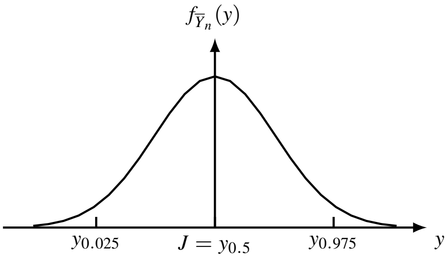
When $Y_1, Y_2, \ldots$ are iid replications of Y, and $\overline{Y}_n$ is the sample average estimator based on n samples; then

$$\sqrt{n}\big(\overline{Y}_n - J\big) \xrightarrow{\mathcal{D}} \mathsf{N}(0, \sigma^2).$$

Normal Distribution

Interpretation:

$$\overline{Y}_n \overset{\mathcal{D}}{\approx} J + \mathsf{N}(0, \sigma^2/n) \overset{\mathcal{D}}{\sim} \mathsf{N}(J, \sigma^2/n).$$



Confidence Interval

- Suppose that σ^2 is known.
- The CLT is the basis for reporting simulation output.
- Your estimator is $\overline{Y}_n$ computed as the average of n iid output data by simulation.
- The standard deviation of the estimator is called *standard error*:

$$\mathrm{SE}[\overline{Y}_n] \doteq \sqrt{\mathrm{Var}(\overline{Y}_n)} = \sigma/\sqrt{n}.$$

- Let $\alpha \in (0, 1)$ be the significance level, typically $\alpha = 5\%$.
- Let z_p be the p-th quantile of the standard normal distribution; i.e. $\Phi(z_p) = \mathbb{P}(Z \leq z_p) = p$.
- Use $1 - \Phi(1 - z_p) = \Phi(z_p)$.
- Then a $100(1 - \alpha)$*% confidence interval* is

$$\left(\overline{Y}_n - z_{1-\alpha/2}\mathrm{SE}[\overline{Y}_n],\ \overline{Y}_n + z_{1-\alpha/2}\mathrm{SE}[\overline{Y}_n]\right)$$

- Interpretation: when you would repeat the experiment of estimating J using sample n, and when you would calculate the associated confidence intervals, then about $100(1 - \alpha)\%$ of these confidence intervals would contain J.

Unknown Standard Error

- Almost always the standard error of the estimator is unknown.
- Equivalently, the variance $\sigma^2 = \mathrm{Var}[Y]$ of the output variable is unknown.
- This variance can be estimated by the *sample variance*

$$S^2 \doteq \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \overline{Y}_n\right)^2.$$

- Hence, replace in your computations $\mathrm{SE}[\overline{Y}_n]$ by its estimate

$$\widehat{\mathrm{SE}} = S/\sqrt{n}.$$

Properties

Properties of the sample variance.

(i). Unbiased estimator: $\mathbb{E}[S^2] = \sigma^2$.

(ii). Strongly consistent estimator: $S^2 \xrightarrow{\mathbb{P}} \sigma^2$; i.e.,

$$\lim_{n\to\infty} \mathbb{P}(|S^2 - \sigma^2| > \epsilon) = 0 \quad \forall \epsilon > 0.$$

(iii).

$$\frac{\overline{Y}_n - J}{\sigma/\sqrt{n}} \overset{\mathcal{D}}{\sim} \mathrm{N}(0,1) \quad \Rightarrow \quad \frac{\overline{Y}_n - J}{S/\sqrt{n}} \overset{\mathcal{D}}{\sim} t_{n-1},$$

the (Student) t distribution with $n-1$ degrees of freedom.

(iv). Asymptotic normality:

$$\sqrt{n}\left(\overline{Y}_n - J\right)/S \xrightarrow{\mathcal{D}} \mathrm{N}(0,1).$$

Normal vs Student t

- According to property (iii), we should use the Student t distribution in stead of the normal distribution when constructing confidence intervals:

$$\left(\overline{Y}_n - t_{n-1,1-\alpha/2}\, S/\sqrt{n},\ \overline{Y}_n + t_{n-1,1-\alpha/2}\, S/\sqrt{n}\right),$$

  where $t_{n-1,q}$ is the q-th quantile for the t-distribution with $n-1$ degrees of freedom.

- According to property (iv), there is not much difference between the critical points for large sample size n.

α	0.15	0.1	0.05	0.025
$t_{9,1-\alpha/2}$	1.574	1.833	2.262	2.685
$t_{49,1-\alpha/2}$	1.462	1.677	2.010	2.312
$t_{99,1-\alpha/2}$	1.451	1.660	1.984	2.276
$t_{249,1-\alpha/2}$	1.443	1.651	1.970	2.255
$t_{999,1-\alpha/2}$	1.441	1.646	1.962	2.245
$z_{1-\alpha/2}$	1.440	1.645	1.960	2.241

The Monte-Carlo Algorithm

Summary. Suppose we wish to compute a performance measure $J = \mathbb{E}[Y]$ where the output variable $Y = h(\boldsymbol{X})$ can be calculated as a function h of a sequence of random input variables $\boldsymbol{X}$. Suppose $\sigma^2 = \mathrm{Var}(Y)$ also unknown.

Monte Carlo algorithm

1. Repeat for $i = 1, \ldots, n$:
   - *(i).* Generate $\boldsymbol{X}_i$ independently of previous runs.
   - *(ii).* Compute output $Y_i = h(\boldsymbol{X}_i)$.
2. Compute sample average estimator for J: $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
3. Compute sample variance estimator for σ^2: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2$.
4. Report estimate, confidence interval, and/or (estimated) standard error.

Estimating a Probability

Suppose that it is asked to compute the probability that the output Y has value in a set A.

Clearly

$$J = \mathbb{P}(Y \in A) = \mathbb{E}[I\{Y \in A\}],$$

with $I\{\cdot\}$ the indicator function.

In other words, it is again an expected value to estimate.

Exercise: the sample variance estimator is simplified to

$$S^2 = \frac{n}{n-1}\overline{Y}_n\left(1 - \overline{Y}_n\right).$$

One Sample versus Sample of Samples

The usual approach is a single sample of size n as in the Monte Carlo algorithm of the previous slide.

Some people prefer k samples, each of length m.

Let's compare by setting $n = k \times m$.

The i-th sample ($i = 1, \ldots, k$) has outputs $Y_j^{(i)}, j = 1, \ldots, m$ with sample average $\overline{Y}_m^{(i)}$.

The overal sample average estimator becomes:

$$\overline{Y}_{km} \doteq \frac{1}{k}\sum_{i=1}^{k}\overline{Y}_m^{(i)} = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{m}\sum_{j=1}^{m}Y_j^{(i)} = \frac{1}{k \times m}\sum_{i=1}^{k}\sum_{j=1}^{m}Y_j^{(i)} = \frac{1}{n}\sum_{t=1}^{n}Y_t = \overline{Y}_n.$$

Conclude: for the estimated value it does not matter.

Next, what about the variances?

One versus Many (cont'd)

When all Y_t's are i.i.d., the (theoretical) variances are

$$\mathrm{Var}[\overline{Y}_{km}] = \mathrm{Var}\Big[\frac{1}{k \times m}\sum_{i=1}^{k}\sum_{j=1}^{m} Y_j^{(i)}\Big] = \mathrm{Var}\big[\frac{1}{n}\sum_{t=1}^{n} Y_t\big] = \frac{1}{n}\mathrm{Var}[Y] = \mathrm{Var}[\overline{Y}_n].$$

Both estimators have the same variance.

What about their estimates by the sample variance estimators?

One versus Many (cont'd)

The sample variance of $\overline{Y}_m^{(1)}, \ldots \overline{Y}_m^{(k)}$ is

$$\widetilde{S}_k^2 \doteq \frac{1}{k-1} \sum_{i=1}^{k} \left(\overline{Y}_m^{(i)} - \overline{Y}_{km}\right)^2.$$

Hence $\mathrm{Var}[\overline{Y}_{km}]$ is estimated by $\widetilde{S}_k^2/k$.

Exercise to work out,

$$\frac{1}{k}\widetilde{S}_k^2 = \frac{1}{k}\,\frac{1}{k-1} \sum_{i=1}^{k} \left(\overline{Y}_m^{(i)} - \overline{Y}_{km}\right)^2 = \frac{1}{k-1}\left(\frac{1}{k} \sum_{i=1}^{k} \left(\overline{Y}_m^{(i)}\right)^2 - \overline{Y}_n^2\right),$$

versus (single sample of size n)

$$\frac{1}{n}S_n^2 = \frac{1}{n}\,\frac{1}{n-1} \sum_{t=1}^{n} \left(Y_t - \overline{Y}_n\right)^2 = \frac{1}{n-1}\left(\frac{1}{n} \sum_{t=1}^{n} Y_t^2 - \overline{Y}_n^2\right).$$

Experiment by yourself.

Illustrative Example

The Problem

- Three points are chosen randomly from the unit square.
- Compute the expected area of the triangle that they form.

Simulation Model

- The input of the system are three IID $X_1, X_2, X_3 \in [0, 1]^2$.
- The output $Y = h(X_1, X_2, X_3)$ is the area of the triangle they form.
- There are several explicit formulas. We choose Heron's formula:
- Define

$$
\begin{aligned}
L_1 &= \text{distance between } X_1 \text{ and } X_2; \\
L_2 &= \text{distance between } X_1 \text{ and } X_3; \\
L_3 &= \text{distance between } X_2 \text{ and } X_3; \\
S &= \frac{1}{2}(L_1 + L_2 + L_3).
\end{aligned}
$$

- Then

$$
Y = \sqrt{S(S - L_1)(S - L_2)(S - L_3)}.
$$

Running the Simulation Model

It is straightforward to program the simulatioin model. We get output (for sample size 1000):

```
sample size 1000
estimation 0.0781065
standard error 0.00210456
95% confidence interval ( 0.0739815 , 0.0822314 )
relative width 10.5623 %
```

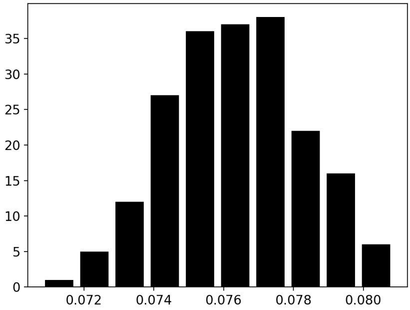
To get smaller confidence interval, for instance about 5% relative width,

```
sample size 5000
estimation 0.0772075
95% confidence interval ( 0.0753025 , 0.0791125 )
relative width 4.93474 %
```

Verify Normality

Do 200 experiments of size 1000 and test normality.

- Histogram of the 200 simulated data:



- Compute moments (e.g. skewness and kurtosis) and compare with of the theoretical values.
- Run a test for normality (Jarque-Bera, Shapiro-Wilk, ...).