Lecture 4 - Monte Carlo methods

Computational Statistics and Applications

Vu Quoc Hoang (vqhoang@fit.hcmus.edu.vn)
Tran Thi Thao Nhi (thaonhitt2005@gmail.com)
FIT - HCMUS

February 5, 2023

Agenda

1. The beginning
2. Monte Carlo method
3. Monte Carlo estimates
4. Variance reduction methods
5. Applications to statistical inference

Agenda

1. **The beginning**
2. Monte Carlo method
3. Monte Carlo estimates
4. Variance reduction methods
5. Applications to statistical inference

The beginning

Question. Compute the integral

$$I = \int_0^\infty x^{0.9} e^{-x} dx.$$

Answer. $I = \Gamma(0.9 + 1) = \Gamma(1.9)$, where Γ is gamma function defined by

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx.$$

Such value has no **closed-form formula** but can be calculated by **numerical analysis**

$$I \approx 0.9618.$$

However, these numerical analysis methods face difficulties in high dimensional spaces. The example above illustrates how we use random numbers to overcome the difficulties, which is called **Monte Carlo method**.

The beginning - The 1^{st} solution

We can rewrite I as

$$I = \int_0^\infty x^{0.9} e^{-x} dx = \int_{-\infty}^\infty x^{0.9} e^{-x} \mathbb{I}_{[0,\infty)}(x) dx = \int_{-\infty}^\infty x^{0.9} f(x) dx$$

where $f(x) = e^{-x}\mathbb{I}_{[0,\infty)}(x)$ is the probability density function of distribution $\text{Exp}(\lambda = 1)$. Thus,

$$I = E_{X\sim \text{Exp}(1)}\left(X^{0.9}\right) = E_f\left(X^{0.9}\right).$$

From **random sample** $X_1, X_2, ..., X_N \overset{\text{iid}}{\sim} \text{Exp}(1)$, we can **estimate** I by **constructing the statistic** $\hat{I} = \frac{1}{N}\sum_{i=1}^N X_i^{0.9}$.
Since $\hat{I}$ is an **unbiased estimator** of I, which means $E(\hat{I}) = I$, where **standard error** is given by

$$\sigma(\hat{I}) = \sqrt{\text{Var}(\hat{I})} = \sqrt{\frac{\text{Var}_f(X^{0.9})}{N}} = \frac{\sigma_f(X^{0.9})}{\sqrt{N}}.$$

The beginning - The 1^{st} solution (cont.)

Based on methods to generate random samples for $Exp(1)$ distribution discussed in the preceding lesson ($U \sim \mathcal{U}(0,1)$, then $X = -\ln U \sim \text{Exp}(1)$). We randomly generate $N = 10$ specific numbers $X_1, ..., X_{10}$ with

$$\hat{I} = \frac{1}{N} \sum_{i=1}^{N} X_i^{0.9} = 1.1692.$$

With these samples, we estimate $\sigma_f(X^{0.9})$ by the statistic

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (X_i^{0.9} - \hat{I})^2}.$$

Thus, the standard deviation (standard error) is estimated by $\hat{\sigma}(\hat{I}) = \frac{s}{\sqrt{N}} = 0.4118$. Since $\sigma(\hat{I}) \propto \frac{1}{\sqrt{N}}$, to get the 0.001 of standard error, there is a need of $N \approx 1.7 \times 10^6$ random numbers $\mathcal{U}(0,1)$.

The beginning - The 2^{nd} solution

We rewrite I as

$$I = \int_0^\infty \left(\frac{1}{x^{0.1}}\right) xe^{-x}dx = \int_{-\infty}^\infty \left(\frac{1}{x^{0.1}}\right) xe^{-x}\mathbb{I}_{[0,\infty)}(x)dx = \int_{-\infty}^\infty \left(\frac{1}{x^{0.1}}\right) g(x)dx$$

where $g(x) = xe^{-x}\mathbb{I}_{[0,\infty)}(x)$ is the probability density function of Erlang$(k = 2, \lambda = 1)$ distribution. Then,

$$I = E_{X\sim\text{Erlang}(2,1)}\left(\frac{1}{X^{0.1}}\right) = E_g\left(\frac{1}{X^{0.1}}\right).$$

Since $X = X_1 + X_2 \sim \text{Erlang}(2, 1)$ if $X_1, X_2 \overset{\text{iid}}{\sim} \text{Exp}(1)$, we randomly generate samples for $X \sim \text{Erlang}(2, 1)$ by sampling $U_1, U_2 \sim \mathcal{U}(0, 1)$ before returning $X = -\ln U_1 U_2$.
With $N = 10$ random numbers in $\mathcal{U}(0, 1)$, we have $\hat{I} = \frac{1}{N/2}\sum_{i=1}^{N/2}\frac{1}{X_i^{0.1}} = 0.9408$ with the standard error $\hat{\sigma}(\hat{I}) = 0.0375$. To get the 0.001 of standard error, we need $N \approx 14000$ random numbers on $\mathcal{U}(0, 1)$.

Agenda

1. The beginning
2. **Monte Carlo method**
3. Monte Carlo estimates
4. Variance reduction methods
5. Applications to statistical inference

Monte Carlo method

Monte Carlo method

- Generate a large set of samples from the statistical model simulated on a computer.
- Learn about the behavior of the model by studying the computer-generated set of samples instead of the model itself.

Monte Carlo is, sort of, a method where "randomness is used to solve problems".

Example

- The expected value of a random variable can be approximated by generating a large number of samples of the random variable and then considering the average value.
- The probability of an event can be approximated by generating a large number of samples and then considering the proportion of samples where the event occurs.
- The quality of a method for statistical inference can be assessed by repeatedly generating synthetic data with a known distribution and then analysing how well the inference method recovers the known properties of the underlying distribution from the synthetic data sets.

Monte Carlo method (cont.)

Many interesting questions can be reduced to computing the expectations of the forms $E(f(X))$ where X is a random "object" from the system under consideration and f is a real-valued function, determining some quantity of interest in the system. There are three different methods to compute the expectation

1. We can find the answer analytically. For example, if the distribution of X has a density function φ, we have

$$E(f(X)) = \int f(x)\varphi(x)dx.$$

   This method only works if we can know φ and solve the resulting integral.
2. We can try to use numerical integration to get an approximation of the value of the integral. This method often works well when X is in a low-dimensional space Xp x bng các phng pháp s.
3. We can use Monte Carlo estimation. This technique is based on the **strong law of the large numbers**: if $(X_j)_{j\in\mathbb{N}}$ is a sequence of iid random variables with the same distribution as X, then

Monte Carlo method - Example 1

For $X \sim \mathcal{N}(0,1)$, we generate iid $X_1, ..., X_N$ with the same distribution as X. Then, we estimate

$$E(X) \approx \frac{1}{N} \sum_{j=1}^{N} X_j = \overline{X}$$

$$E(\sin(X)^2) \approx \frac{1}{N} \sum_{j=1}^{N} \sin(X_j)^2$$

With $N = 10000$, after generating the specific samples, we have

$$E(X) \approx 0.0020, \quad E(\sin(X)^2) \approx 0.4292.$$

Monte Carlo method - Example 2

Let X be a random variable, we have $P(X \in A) = E(\mathbb{I}_A(X))$. Thus, we can estimate

$$P(X \in A) = E(\mathbb{I}_A(X)) \approx \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}_A(X_j).$$

Example. For $X \sim \mathcal{N}(0,1), a = -1$, we generate iid $X_1, ..., X_N$ with the same distribution as X. Then, we estimate

$$P(X \leq a) \approx \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}_{(-\infty,a]}(X_j).$$

With $N = 10000$, after generating the specific samples, we have

$$P(X \leq a) \approx 0.8406.$$

Thus,

$$P(X \leq a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = \Phi(a) = 0.8413.$$

Monte Carlo method - Example 3

Let $X \sim \mathcal{U}(a,b)$ where the density function is $\varphi(x) = \frac{1}{b-a}\mathbb{I}_{[a,b]}$, then we estimate

$$\int_a^b f(x)dx = (b-a)\int_{-\infty}^{\infty} f(x)\varphi(x)dx = (b-a)E(f(x)) \approx \frac{(b-a)}{N}\sum_{j=1}^{N} f(X_j).$$

Ví d. We can generate $X_1, ..., X_N \overset{\text{iid}}{\sim} \mathcal{U}(0, 2\pi)$ and then use the approximation

$$\int_0^{2\pi} e^{\cos x}dx \approx \frac{2\pi}{N}\sum_{j=1}^{N} e^{\cos X_j}.$$

With $N = 10000$, after generating the specific samples, we have

$$\int_0^{2\pi} e^{\cos x}dx \approx 8.0082.$$

Monte Carlo method - Example 4

Consider a simple Bayesian inference problem, where we want to make inference about $X \sim \text{Exp}(1)$ using a single observation y of $Y \sim \mathcal{N}(0, X)$. To solve this problem, we have to find the posterior distribution $X|Y = y$, which is the conditional distribution of X given the observation $Y = y$.

First, the prior probability of X, $X \sim \text{Exp}(1)$, has the density function

$$p_X(x) = e^{-x}\mathbb{I}_{[0,\infty)}(x).$$

The conditional distribution of $Y \sim \mathcal{N}(0, X)$, given $X = x$ has the conditional density function

$$p_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi x}} e^{-y^2/(2x)}.$$

Then, using Bayes' rule, the posterior probability $X|Y = y$ has the density function

$$p_{X|Y=y}(x) = \frac{p_{Y|X=x}(y)p_X(x)}{p_Y(y)} = \frac{p_{Y|X=x}(y)p_X(x)}{\int p_{Y|X=u}(y)p_X(u)du} \propto p_{Y|X=x}(y)p_X(x)$$

Monte Carlo method - Example 4 (cont.)

Let

$$f(x) = \frac{1}{\sqrt{x}} e^{-y^2/(2x)-x} \mathbb{I}_{[0,\infty)}(x) \propto p_{Y|X=x}(y) p_X(x) \propto p_{X|Y=y}(x),$$

we can use the rejection sampling algorithm to generate samples from the posterior distribution of $X|Y = y$. Note that, the rejection sampling algorithm can still be applied when the normalising constant Z_f is unknown.

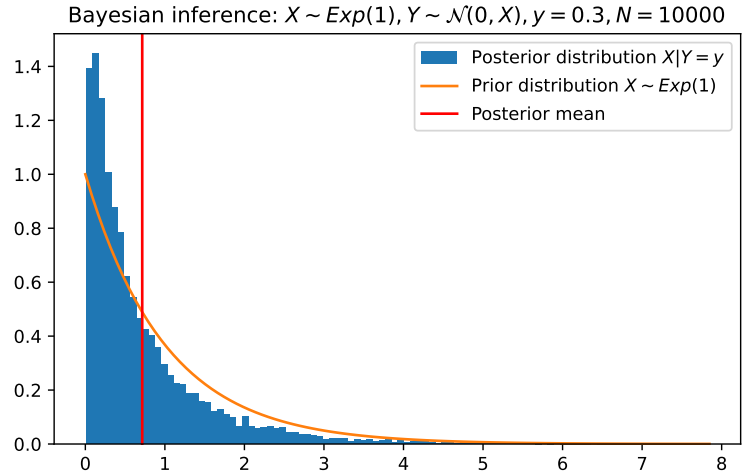
In detail, we try to use an Exp(1)-distribution for the proposals in envelope rejection sampling algorithm, and a constant

$$c = \frac{1}{|y|} e^{-1/2},$$

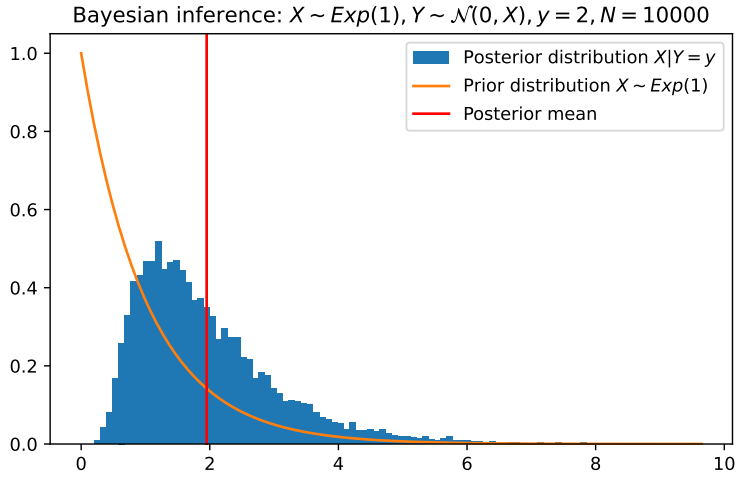
we generate iid $X_1, ..., X_N$ from the posterior distribution $X|Y = y$, draw a histogram and estimate

$$E(X|Y = y) \approx \frac{1}{N} \sum_{j=1}^{N} X_j = \bar{X}, \quad Var(X|Y = y) \approx \frac{1}{N-1} \sum_{j=1}^{N} \left(X_j - \bar{X}\right)^2.$$

Monte Carlo method - Example 4 (cont.)



Monte Carlo method - Example 4 (cont.)



Agenda

1. The beginning
2. Monte Carlo method
3. **Monte Carlo estimates**
4. Variance reduction methods
5. Applications to statistical inference

Monte Carlo estimates

Let X be a random variable and f be a real-valued function, **Monte Carlo estimate** for $E(f(X))$ is given by

$$Z_N^{MC} = \frac{1}{N}\sum_{j=1}^{N} f(X_j)$$

where $X_1, ..., X_N$ are iid with the same distribution as X.
Note: Since the estimate Z_N^{MC} is constructed from random samples X_j, it is a random quantity itself.

Monte Carlo estimates (cont.)

MCS algorithm. (Monte Carlo eStimate)
Input:

- distribution of X,
- a real-valued function f,
- $N \in \mathbb{N}$.

Output: an estimate Z_N^{MC} for $E(f(X))$.

1: $s \leftarrow 0$
2: **for** $j = 1, 2, ..., N$ **do**
3: generate X_j, with the same distribution as X has
4: $s \leftarrow s + f(X_j)$
5: **end for**
6: **return** s/N

Monte Carlo error

Let $\hat{\theta} = \hat{\theta}(X)$ be an estimator for a parameter θ, some definitions are given as follows:

- **Bias** of the estimator is given by

$$\text{bias}(\hat{\theta}) = E_\theta(\hat{\theta}(X) - \theta) = E_\theta(\hat{\theta}(X)) - \theta,$$

- **Standard error**

$$\text{se}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})} = \sqrt{E_\theta\left((\hat{\theta}(X) - E_\theta(\hat{\theta}(X)))^2\right)},$$

- **Mean squared error - MSE**

$$\text{MSE}(\hat{\theta}) = E_\theta\left((\hat{\theta}(X) - \theta)^2\right),$$

- **Root-mean-square error - RMSE**

$$\text{RMSE}(\hat{\theta}) = \sqrt{\text{MSE}(\hat{\theta})}.$$

Proposition. $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{bias}(\hat{\theta})^2 = \text{se}(\hat{\theta})^2 + \text{bias}(\hat{\theta})^2$.

Monte Carlo error (cont.)

Since the estimate Z_N^{MC} is random, the Monte Carlo $Z_N^{MC} - E(f(X))$ is also random. To quantify the magnitude of this random error, we use the concepts of bias and mean squared error from statistics.

Proposition. Monte Carlo estimate Z_N^{MC} for $E(f(X))$

$$Z_N^{MC} = \frac{1}{N} \sum_{j=1}^{N} f(X_j)$$

has

$$\text{bias}(Z_N^{MC}) = 0,$$

and

$$\text{MSE}(Z_N^{MC}) = \text{Var}(Z_N^{MC}) = \frac{1}{N}\text{Var}(f(X)).$$

Monte Carlo error - Example

Back to the preceding example, given $X \sim \mathcal{N}(0,1)$, the Monte Carlo estimate for $E(\sin(X)^2)$ is given by

$$Z_N^{MC} = \frac{1}{N}\sum_{j=1}^{N} \sin(X_j)^2$$

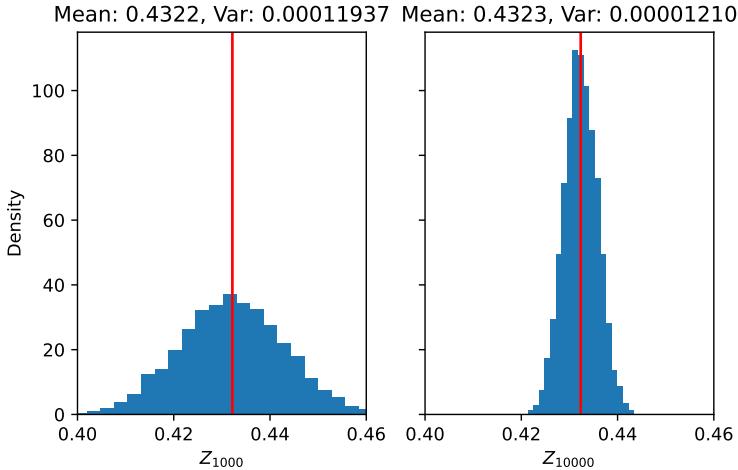
where $X_1, X_2, ..., X_N \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.

This proposition shows that Z_N^{MC} is the unbiased estimate for $E(\sin(X)^2)$, where $\text{bias}(Z_N^{MC}) = 0$, and

$$\text{MSE}(Z_N^{MC}) = \text{Var}(Z_N^{MC}) = \frac{\text{Var}(f(X))}{N} \propto \frac{1}{N},$$

$$\text{RMSE}(Z_N^{MC}) = \sqrt{\text{MSE}(Z_N^{MC})} = \frac{\sigma(f(X))}{\sqrt{N}} \propto \frac{1}{\sqrt{N}}.$$

Monte Carlo error - Example (cont.)



Choice of sample size

If the value of $\mathrm{Var}(f(X))$ is unknown, we can estimate

$$\mathrm{MSE}(Z_N^{MC}) = \frac{\mathrm{Var}(f(X))}{N} \approx \frac{\hat{\sigma}^2}{N},$$

where

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{j=1}^{N} \left(f(X_j) - Z_N^{MC} \right)^2$$

is the sample variance of the generated values $f(X_1), f(X_2), ..., f(X_N)$.

To determine an efficient sample size N (the standard error gets bigger for the smaller sample sizes, while the larger sample sizes incur higher costs), we can run with an "appropriate" N, estimate the error, then modify the sample size accordingly ($\mathrm{RMSE}(Z_N^{MC}) \propto \frac{1}{\sqrt{N}}$) and execute another run. (See the beginning example)

Choice of sample size (cont.)

If the sample variance $Var(f(X))$ or its upper bound is known, to achieve error $\text{MSE}(Z_N^{MC}) \leq \epsilon^2$, the sample size N must satisfy

$$N \geq \frac{Var(f(X))}{\epsilon^2}.$$

Example 1. Assume $Var(f(X)) = 1$. To estimate $E(f(X))$ so that the error satisfies $\text{MSE} \leq \epsilon^2 = 0.01^2$, we can use a Monte Carlo estimate with

$$N \geq \frac{Var(f(X))}{\epsilon^2} = \frac{1}{0.01^2} = 10000.$$

Choice of sample size (cont.)

Example 2. Let X be a real-valued random variable and $A \subseteq \mathbb{R}$. As we have seen in the preceding example, we can estimate $p = P(X \in A)$ by

$$Z_N^{MC} = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}_A(X_j).$$

The variance of the Monte Carlo samples is

$$Var(\mathbb{I}_A(X)) = E(\mathbb{I}_A(X)^2) - E(\mathbb{I}_A(X))^2 = p - p^2 = p(1-p).$$

Thus, we can achieve MSE $\leq \epsilon^2$ by choosing

$$N \geq \frac{p(1-p)}{\epsilon^2}.$$

This bound depends on the unknown probability p but, since $p(1-p) \leq 1/4$, $\forall p \in [0,1]$ choosing

$$N \geq \frac{1}{4\epsilon^2}.$$

Refined error bounds

The **central limit theorem** can be used to obtain refined bounds for the Monte Carlo error.

Proposition. Let $\alpha \in (0,1)$, and $q_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ where Φ is the CDF of the standard normal distribution $\mathcal{N}(0,1)$. Let $\sigma^2 = Var(f(X))$ and

$$N \geq \frac{q_\alpha^2 \sigma^2}{\epsilon^2}$$

Then, the Monte Carlo estimate Z_N^{MC} for $E(f(X))$ (approximately for large N)

$$P\left(\left|Z_N^{MC} - E(f(X))\right| \leq \epsilon\right) \geq 1 - \alpha.$$

An alternative way to express the result of the proposition above is to replace the point estimator Z_N^{MC} by a confidence interval

$$P\left(E(f(X)) \in \left[Z_N^{MC} - \frac{\sigma q_\alpha}{\sqrt{N}}, Z_N^{MC} + \frac{\sigma q_\alpha}{\sqrt{N}}\right]\right) \geq 1 - \alpha.$$

Refined error bounds (cont.)

As a special case, for $\alpha = 5\%$, we have $q_{0.05} = \Phi^{-1}(0.975) \approx 1.96$, thus we need

$$N \geq \frac{1.96^2 \sigma^2}{\epsilon^2}$$

in order to have an absolute error of at most ϵ with at least 95% probability. We see that approximately 4 times as many samples are required as for the condition $\text{RMSE}(Z_N^{MC}) \leq \epsilon$.

Example. Assume $Var(f(X)) = 1$, to estimate Z_N^{MC} for $E(f(X))$ so that the $\left|Z_N^{MC} - E(f(X))\right|$ is at most $\epsilon = 0.01$ with probability at least $1 - \alpha = 95\%$, we can use the samples with

$$N \geq \frac{1.96^2 \, Var(f(X))}{\epsilon^2} = \frac{1.96^2}{(0.01)^2} = 38416.$$

Refined error bounds (cont.)

If the standard deviation $\sigma^2 = Var(f(X))$ is unknown, it can be replaced by

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{j=1}^{N} \left(f(X_j) - Z_N^{MC} \right)^2$$

which is samples variance of $f(X_1), f(X_2), ..., f(X_N)$ to estimate σ^2.

Accordingly, $q_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) = P(\mathcal{N}(0,1) \geq \alpha/2)$ should be replaced by the corresponding quantile $q_\alpha^{N-1} = P(\text{Student}(N-1) \geq \alpha/2)$ where Student$(N-1)$ is Student's t-distribution with $N-1$ degrees of freedom. However, when N is quite large, $q_\alpha \approx q_\alpha^{N-1}$.

Example. In the beginning example, if we use 10 random numbers $\mathcal{U}(0,1)$, confidence interval 95% for $I = \int_0^\infty x^{0.9} e^{-x} dx$ is $[0.8917, 1.3096]$ and $[0.8367, 1.0450]$ according to the first and second solution, respectively. Furthermore, if using 14000 random numbers, the confidence interval will be $[0.9594, 0.9632]$.

Agenda

1. The beginning
2. Monte Carlo method
3. Monte Carlo estimates
4. **Variance reduction methods**
5. Applications to statistical inference

Introduction

As we have seen, the Monte Carlo estimate Z_N^{MC} for $E(f(X))$ has the mean square error

$$\text{MSE}(Z_N^{MC}) = \text{Var}(Z_N^{MC}) = \frac{\text{Var}(f(X))}{N}.$$

To improve the efficiency of the estimate, we find a way to reduce the variance ($\text{Var}(Z_N^{MC})$).

Importance sampling

Assume that X is a random variable with density function ϕ, that f is a real-valued function, and ψ is another probability density function with $\psi(x) > 0$ whenever $f(x)\phi(x) > 0$. We have

$$E_{X\sim\phi}(f(X)) = \int f(x)\phi(x)dx = \int \frac{f(x)\phi(x)}{\psi(x)}\psi(x)dx = E_{Y\sim\psi}\left(\frac{f(Y)\phi(Y)}{\psi(Y)}\right).$$

Then, **importance sampling estimate** for $E(f(X))$ is given by

$$Z_N^{IS} = \frac{1}{N}\sum_{j=1}^{N}\frac{f(Y_j)\phi(Y_j)}{\psi(Y_j)},$$

where $Y_1, Y_2, ..., Y_N$ are iid with density function ψ.

Importance sampling (cont.)

IS algorithm. (Importance Sampling)
Input:

- a real-valued function f,
- the density function ϕ of X,
- an density function ψ,
- $N \in \mathbb{N}$.

Output: an estimate Z_N^{IS} for $E(f(X))$.

1: $s \leftarrow 0$
2: **for** $j = 1, 2, ..., N$ **do**
3: generate $Y_j \sim \psi$
4: $s \leftarrow s + f(Y_j)\phi(Y_j)/\psi(Y_j)$
5: **end for**
6: **return** s/N

Importance sampling (cont.)

Proposition. The importance sampling estimate $Z_N^{IS} = \frac{1}{N}\sum_{j=1}^{N} \frac{f(Y_j)\phi(Y_j)}{\psi(Y_j)}$ for $E(f(X))$ has bias$(Z_N^{IS}) = 0$ and

$$\begin{aligned}\mathrm{MSE}(Z_N^{IS}) &= \frac{1}{N}\mathrm{Var}\left(\frac{f(Y)\phi(Y)}{\psi(Y)}\right) \\ &= \frac{1}{N}\left(\mathrm{Var}(f(X)) - E\left(f(X)^2\left(1-\frac{\phi(X)}{\psi(X)}\right)\right)\right).\end{aligned}$$

The importance sampling method is efficient if both of the following criteria are satisfied:

- The samples $Y_1, Y_2, ..., Y_N$ can be generated efficiently from ψ,
- $\mathrm{Var}\left(f(Y)\phi(Y)/\psi(Y)\right)$ is small or, equivalently, the constant $c_\psi = E\left(f(X)^2\left(1-\frac{\phi(X)}{\psi(X)}\right)\right)$ is large. Thus, we will get good efficiency if we choose ψ to be approximately proportional to the function $f\phi$ or even if we choose a distribution such that ψ is big whenever $|f|$ is big. (This is the reason why this method is called importance sampling!)

Importance sampling - Example

Let X be a real-valued random variable and $A \subset \mathbb{R}$, then the importance sampling estimate for $P(X \in A) = E(\mathbb{I}_A(X))$ is given by

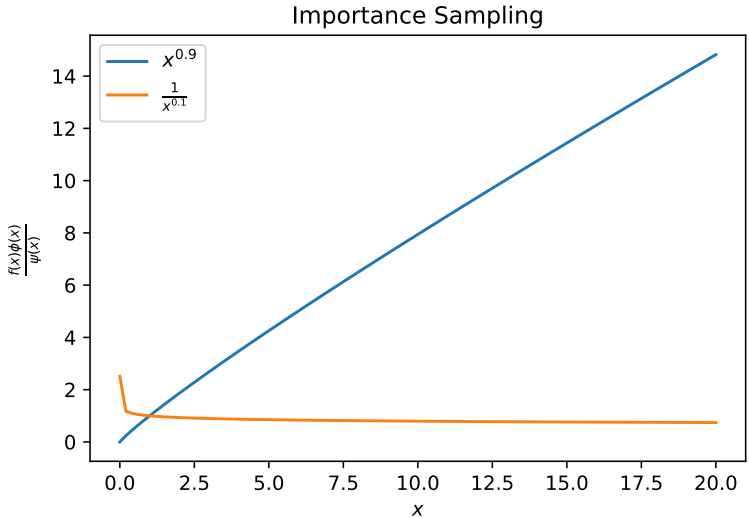
$$Z_N^{IS} = \frac{1}{N} \sum_{j=1}^{N} \frac{\mathbb{I}_A(Y_j)\phi(Y_j)}{\psi(Y_j)},$$

where ψ is a probability density function, which satisfies $\psi(x) > 0$ for all points $x \in A$ with $\phi(x) > 0$, and $Y_1, Y_2, ..., Y_N$ are a sequence of iid random variables with density function ψ. Then,

$$\begin{aligned} \text{MSE}(Z_N^{IS}) &= \frac{1}{N}\text{Var}(\mathbb{I}_A(X)) - \frac{1}{N}E\left(\mathbb{I}_A(X)\left(1 - \frac{\phi(X)}{\psi(X)}\right)\right) \\ &= \text{MSE}(Z_N^{MC}) - \frac{1}{N}E\left(\mathbb{I}_A(X)\left(1 - \frac{\phi(X)}{\psi(X)}\right)\right). \end{aligned}$$

Thus, $\text{MSE}(Z_N^{IS}) < \text{MSE}(Z_N^{MC})$ if we can choose the density ψ such that $\psi > \phi$ on the set A.

Importance sampling - The beginning example



Antithetic variables

Assume that X, X' are identically distributed random variables, which are unnecessarily independent. Then we have

$$E\left(\frac{f(X)+f(X')}{2}\right) = \frac{E(f(X))+E(f(X'))}{2} = E(f(X)),$$

$$\mathrm{Var}\left(\frac{f(X)+f(X')}{2}\right) = \frac{1}{2}\mathrm{Var}(f(X)) + \frac{1}{2}\mathrm{Cov}(f(X), f(X')).$$

Antithetic variables estimate for $E(f(X))$ with sample size $N \in 2\mathbb{N}$ is given by

$$Z_N^{AV} = \frac{1}{N}\sum_{k=1}^{N/2}\left(f(X_k)+f(X_k')\right),$$

where (X_k, X_k') are iid copies of (X, X'), which are called **antithetic pair**.

Antithetic variables (cont.)

AV algorithm. (antithetic variables)
Input:

- a real-valued function f,
- $N \in \mathbb{N}$ even.

Output: the estimate Z_N^{AV} for $E(f(X))$.

1: $s \leftarrow 0$
2: **for** $k = 1, 2, ..., N/2$ **do**
3: generate $(X_k, X_k') \sim (X, X')$
4: $s \leftarrow s + f(X_k) + f(X_k')$
5: **end for**
6: **return** s/N

Antithetic variables (cont.)

Proposition. Let X, X' be 2 random variables with the same distribution and let $\rho = \mathrm{Cor}(f(X), f(X'))$, then the antithetic variables estimate $Z_N^{AV} = \frac{1}{N}\sum_{k=1}^{N/2}(f(X_k) + f(X_k'))$ for $E(f(X))$ satisties

$$bias(Z_N^{AV}) = 0,$$

and

$$\mathrm{MSE}(Z_N^{AV}) = \frac{1}{N}\mathrm{Var}(f(X))(1+\rho).$$

To conclude, the antithetic variables method is effective if and only if $\rho < 0$. Equivalently, we need to construct the pairs X, X' of samples such that both values have the correct distribution but, at the same time, $f(X), f(X')$ are negatively correlated.

Antithetic variables (cont.)

A first idea for construction antithetic pairs, if the distribution of X is symmetric, is to use $X' = -X$, then in many cases we have $\mathrm{Cor}(f(X), f(X')) < 0$.

Example 1. Let $X \sim \mathcal{N}(0,1)$, and consider the problem of estimating the probability

$$p = P(X \in [1,3]) = E(\mathbb{I}_{[1,3]}(X)).$$

Since the distribution of X is symmetric, we can try to use (X, X') with $X' = -X$. For this choice we find

$$\rho = \mathrm{Cor}(\mathbb{I}_{[1,3]}(X), \mathbb{I}_{[1,3]}(-X)) = -\frac{p}{1-p}.$$

Thus, by the result of the above proposition, we have

$$\mathrm{MSE}(Z_N^{AV}) = (1+\rho)\mathrm{MSE}(Z_N^{MC}) = \left(1 - \frac{p}{1-p}\right)\mathrm{MSE}(Z_N^{MC}).$$

Since $p \approx 0.16$, $\rho \approx -0.19$, thus $\mathrm{MSE}(Z_N^{AV}) \approx 81\%\mathrm{MSE}(Z_N^{MC})$.

Antithetic variables (cont.)

Another method for generating antithetic pairs: if F is the distribution function of X whose inverse can be applied easily, we can use $X = F^{-1}(U), X' = F^{-1}(1-U)$ where $U \sim \mathcal{U}(0,1)$, then

- X, X' have the same distribution since $U, 1-U$ are $\mathcal{U}(0,1)$,
- $\mathrm{Cor}(X, X') \leq 0$ since F^{-1} monotonically (increasing). Moreover, if f monotonically (inscreasing or decreasing), $\mathrm{Cor}(f(X), f(X')) \leq 0$ as we can be seen in the following proposition.

Proposition. Let $g : \mathbb{R} \to \mathbb{R}$ be monotonically (increasing or decreasing) and $U \sim \mathcal{U}(0,1)$, then

$$\mathrm{Cor}(g(U), g(1-U)) \leq 0.$$

Example 2. In the beginning example, to compute the integral

$$I = \int_0^\infty x^{0.9} e^{-x} dx = \int_{-\infty}^\infty x^{0.9} e^{-x} \mathbb{I}_{[0,\infty)}(x) dx = E_{X \sim \mathrm{Exp}(1)} \left(X^{0.9} \right).$$

Antithetic variables (cont.)

We can use the Monte Carlo estimate

$$Z_N^{MC} = \frac{1}{N}\sum_{i=1}^{N}(-\ln U_i)^{0.9}$$

where $U_1, U_2, ..., U_N \overset{\text{iid}}{\sim} \mathcal{U}(0,1)$. Vì $x^{0.9}$ is monotonically (increasing), thus we can use antithetic variables estimate

$$Z_N^{AV} = \frac{1}{N}\sum_{i=1}^{N/2}\left((-\ln(U_i))^{0.9} + (-\ln(1-U_i))^{0.9}\right).$$

The correlation $\rho = \text{Cor}(f(X), f(X')) = \text{Cor}\left((-\ln(U))^{0.9}, (-\ln(1-U))^{0.9}\right)$ is hard to compute but easy to estimate from the samples $\hat{\rho} = -0.7114$. Thus, $\text{MSE}(Z_N^{AV}) \approx 29\%\text{MSE}(Z_N^{MC})$.

Control variates

Let f, g be real-valued functions where $g \approx f$, then we have

$$E(f(X)) = E\left(f(X) - g(X)\right) + E(g(X)).$$

If $E(g(X))$ can be computed, then we can use our knowledge of $E(f(X) - g(X))$ to assist with the estimation of $E(f(X))$. Since $g \approx f$, $\mathrm{Var}(f(X) - g(X)) < \mathrm{Var}(f(X))$.

Let g be a function such that $E(g(X))$ is known, **control variates estimate** for $E(f(X))$ is given by

$$Z_N^{CV} = \frac{1}{N} \sum_{j=1}^{N} \left(f(X_j) - g(X_j)\right) + E(g(X)),$$

where $X_1, X_2, ..., X_N$ are iid copies of X. The random variable $g(X)$ is called **control variate** for $f(X)$.

Control variates (cont.)

CV algorithm. (control variates)
Input:

- a real-valued function f,
- a function $g \approx f$ such that $E(g(X))$ is known,
- $N \in \mathbb{N}$ even.

Output: a estimate Z_N^{CV} for $E(f(X))$.

1: $s \leftarrow 0$
2: **for** $j = 1, 2, ..., N$ **do**
3: generate $X_j \sim X$
4: $s \leftarrow s + f(X_j) - g(X_j)$
5: **end for**
6: **return** $s/N + E(g(X))$

Control variates (cont.)

Proposition. The control variates estimate $Z_N^{CV} = \frac{1}{N}\sum_{j=1}^{N}(f(X_j) - g(X_j)) + E(g(X))$ satisfies

$$\text{bias}(Z_N^{CV}) = 0,$$

and

$$\text{MSE}(Z_N^{CV}) = \frac{1}{N}\text{Var}\left(f(X) - g(X)\right).$$

Thus, the control variates method is effective if

- We can find a simpler function $g \approx f$ such that $E(g(X))$ is analytically computable,
- $f(X) - g(X)$ has smaller variance than $f(X)$.

Control variates (cont.)

The control variates method described above is a special case of a more general method. Using a correlated control variate Y, every random variable Z can be transformed into a new random variable $\tilde{Z}$ with the same mean but smaller variance. This technique is described in the following proposition.

Proposition. Let Z be a random variable with $E(Z) = \mu$ and $\mathrm{Var}(Z) = \sigma^2$, random variable Y has $\mathrm{Cor}(Y, Z) = \rho$. We define

$$\tilde{Z} = Z - \frac{\mathrm{Cov}(Y, Z)}{\mathrm{Var}(Y)}(Y - E(Y))$$

then the random variable $\tilde{Z}$ satisfies $E(\tilde{Z}) = \mu$ and

$$\mathrm{Var}(\tilde{Z}) = (1 - \rho^2)\sigma^2 \leq \sigma^2.$$

Control variates - Example

To compute an integral $I = \int_0^1 e^{x^2} dx$, which is hard to compute, we consider another "similar" but simpler integral

$$J = \int_0^1 e^x dx = e^x \Big|_{x=0}^{x=1} = e^1 - e^0 = e - 1.$$

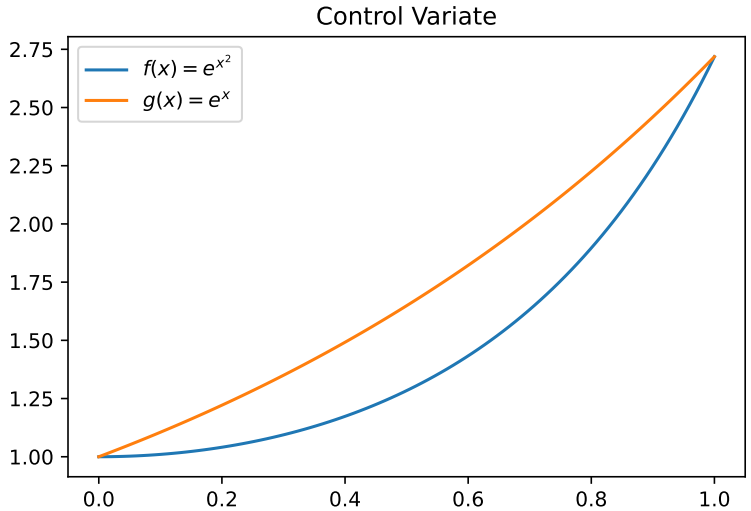
With $f(x) = e^{x^2}, g(x) = e^x, X \sim \mathcal{U}(0,1)$, we have $I = E(f(X)), J = E(g(X))$.
Generating $X_1, X_2, ..., X_N \overset{\text{iid}}{\sim} \mathcal{U}(0,1)$, then the Monte Carlo estimate for I is $Z_N^{MC} = \frac{1}{N}\sum_{j=1}^N f(X_j)$ and the control variates estimate for I is $Z_N^{CV} = \frac{1}{N}\sum_{j=1}^N (f(X_j) - g(X_j)) + e - 1$.

With $N = 10000$, after generating the specific samples, we have

$$Z_N^{MC} \approx 1.4680, \quad Z_N^{CV} \approx 1.4617$$

and $\text{MSE}(Z_N^{CV}) \approx 6\%\text{MSE}(Z_N^{MC})$.

Control variates - Example (cont.)



Agenda

1. The beginning
2. Monte Carlo method
3. Monte Carlo estimates
4. Variance reduction methods
5. **Applications to statistical inference**

Introduction

Statistical inference problems

- We have observed data $x = (x_1, ..., x_n)$.
- We consider a family $(P_\theta)_{\theta \in \Theta}$ of probability distributions, where θ is the parameter vector of the models and Θ is the set of all possible parameter values.
- We assume that the observed data are a sample of a random variable $X = (X_1, ..., X_n) \sim P_\theta$ for an unknown parameter value $\theta \in \Theta$.
- The aim here is to decide which statistical model P_θ the observed data x could have been generated by.

Point estimators

- A **point estimator** for the parameter θ is any function of the random sample X with values in Θ, typically denoted by $\hat{\theta} = \hat{\theta}(X) = \hat{\theta}(X_1, ..., X_n)$.
- **Bias** of an estimator $\hat{\theta} = \hat{\theta}(X)$ for a parameter θ is given by

$$\text{bias}_\theta(\hat{\theta}) = E\left(\hat{\theta}(X)\right) - \theta, \forall \theta \in \Theta.$$

- For given value of θ, the Monte Carlo estimate for the bias is given by

$$\widehat{\text{bias}_\theta}(\hat{\theta}) = \frac{1}{N}\sum_{j=1}^{N} \hat{\theta}(X^{(j)}) - \theta,$$

  where the samples $X^{(j)} = (X_1^{(j)}, ..., X_n^{(j)})$ are iid copies of X.
- *Note:* n is the size of samples, while N is the number of samples used in the Monte Carlo estimate; $\hat{\theta}$ is an estimator for θ while $\widehat{\text{bias}_\theta}(\hat{\theta})$ is an estimator for the bias of an estimator $\hat{\theta}$.

Point estimators - Example

Let $\rho \in [-1, 1]$ and $X, \eta \sim \mathcal{N}(0, 1)$, and define

$$Y = \rho X + \sqrt{1 - \rho^2}\eta.$$

Then

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\rho}{\sqrt{1 \times 1}} = \rho.$$

The correlation can be estimated by the **sample correlation**

$$\hat{\rho}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and (X_i, Y_i) is a sequence of iid copies of (X, Y).

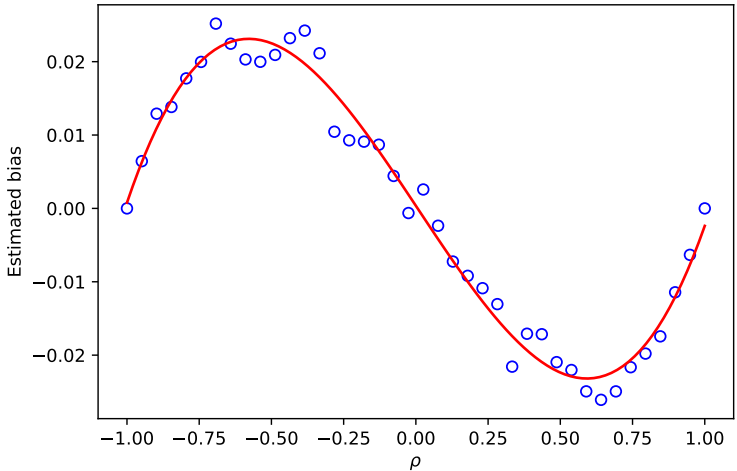
Point estimators - Example (cont.)

For given ρ, the Monte Carlo estimator $\widehat{\text{bias}_\rho}(\hat{\rho})$ can be computed using the following steps:

1. $S \leftarrow 0$
2. **for** $j = 1, 2, ..., N$ **do**
3. generate $X_1^{(j)}, ..., X_n^{(j)} \sim \mathcal{N}(0, 1)$
4. generate $\eta_1^{(j)}, ..., \eta_n^{(j)} \sim \mathcal{N}(0, 1)$
5. let $Y_i^{(j)} = \rho X_i^{(j)} + \sqrt{1 - \rho^2}\eta_i^{(j)}, i = 1, 2, ..., n$
6. compute $\hat{\rho}^{(j)} = \hat{\rho}(X^{(j)}, Y^{(j)})$ which is the sample correlation
7. $S \leftarrow S + \hat{\rho}^{(j)}$
5. **end for**
6. return $S/N - \rho$

The estimate $\widehat{\text{bias}_\rho}(\hat{\rho})$ can be run for different values of $\rho \in [-1, 1]$ to get the dependence of the bias on the parameter ρ.

Point estimators - Example (cont.)



Point estimators (cont.)

Standard error of an estimator $\hat{\theta} = \hat{\theta}(X)$ for parameter θ is given by

$$\mathsf{se}_\theta(\hat{\theta}) = \sigma_\theta(\hat{\theta}(X)) = \sqrt{\mathsf{Var}_\theta(\hat{\theta}(X))}, \forall \theta \in \Theta.$$

For given θ, Monte Carlo estimate for standard error is defined by

$$\widehat{\mathsf{se}_\theta}(\hat{\theta}) = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}\left(\hat{\theta}(X^{(j)}) - \bar{\hat{\theta}}\right)^2},$$

where

$$\bar{\hat{\theta}} = \frac{1}{N}\sum_{j=1}^{N}\hat{\theta}(X^{(j)})$$

and $X^{(j)}$ are iid copies of X.

Confidence intervals

A **confidence interval** with confidence coefficient $1-\alpha$ for a parameter θ is a random interval $[U, V] \subset \mathbb{R}$ where $U = U(X), V = V(X)$ are functions of the random sample $X = (X_1, ..., X_n)$, such that

$$P_\theta(\theta \in [U(X), V(X)]) \geq 1 - \alpha, \forall \theta \in \Theta.$$

In many cases, a confidence interval for a parameter θ can be constructed by considering a point estimator $\hat{\theta} = \hat{\theta}(X)$ for θ as follows:

$$P_\theta(\theta \in [\hat{\theta} - \epsilon, \hat{\theta} + \epsilon]) \geq 1 - \alpha,$$

where $\epsilon > 0$ is an appropriate value choosen from the distribution of $\hat{\theta} - \theta$.

Confidence intervals - Example

Let $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with a known variance σ^2, construct the confidence interval for the unknown mean μ.

The typically point estimator for μ is

$$\hat{\mu} = \hat{\mu}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

where

$$\hat{\mu} - \mu \sim \mathcal{N}(0, \sigma^2/n).$$

Then, if choosing $\epsilon = \frac{\sigma q_\alpha}{\sqrt{n}}$ where $q_\alpha = \Phi^{-1}(1 - \alpha/2)$, we have a confidence interval $1 - \alpha$ for μ which is

$$I(X) = \left[\bar{X} - \frac{\sigma q_\alpha}{\sqrt{n}}, \bar{X} + \frac{\sigma q_\alpha}{\sqrt{n}}\right].$$

Confidence intervals - Example (cont.)

If the variance σ^2 is not known, we can use a confidence interval

$$I(X) = \left[\bar{X} - \frac{\hat{\sigma} q_\alpha^{n-1}}{\sqrt{n}}, \bar{X} + \frac{\hat{\sigma} q_\alpha^{n-1}}{\sqrt{n}}\right],$$

where

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

and q_α^{n-1} is the $1 - \alpha/2$-quantile of Student's t-distribution with $n - 1$ degrees of freedom.

Confidence intervals (cont.)

If X_i are not normally distributed and n is small, the confidence intervals are no longer accurate. Monte Carlo estimates can be used, both to construct and access the confidence intervals, to approximate the expectation as

$$P_\theta(\theta \in [U, V]) \approx \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}_{[U(X^{(j)}), V(X^{(j)})]}(\theta),$$

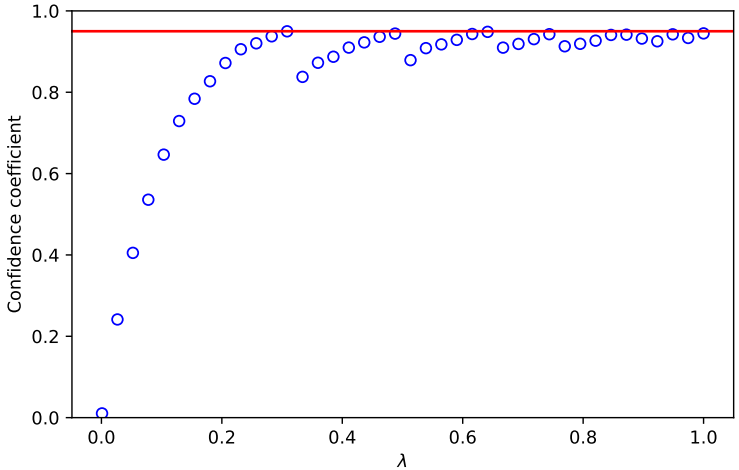
where the vectors $X^{(j)} = (X_1^{(j)}, ..., X_n^{(j)})$ are iid copies of $X = (X_1, ..., X_n)$.

Confidence intervals - Example

Let $X_1, ..., X_n$ be independent and Poisson-distributed with parameter λ, a Monte Carlo estimate for the confidence interval $P_\lambda(U \leq \lambda \leq V)$ in the preceding example for given λ, can be obtained by the following algorithm:

1. $k \leftarrow 0$
2. **for** $j = 1, 2, ..., N$ **do**
3. generate $X_1, ..., X_n \sim Poisson(\lambda)$
4. $\hat{\mu} \leftarrow \sum_{i=1}^{n} X_i / n$
5. $\hat{\sigma} \leftarrow \sqrt{\sum_{i=1}^{n} (X_i - \hat{\mu})^2 / (n-1)}$
6. $U \leftarrow \hat{\mu} - p_{n,\alpha} \hat{\sigma} / \sqrt{n}$
7. $V \leftarrow \hat{\mu} + p_{n,\alpha} \hat{\sigma} / \sqrt{n}$
8. **if** $U \leq \lambda \leq V$ **then**
9. $k \leftarrow k + 1$
10. **end if**
11. **end for**
12. return k/N

Confidence intervals (cont.)



Hypothesis tests

A **statistical hypothesis test** of size $\alpha \in (0,1)$ for the hypothesis $H_0 = \{\theta \in \Theta_0\}$ ($\Theta_0 \subset \Theta$) is given by a function $T = T(X)$ of the random sample $X = (X_1, ..., X_n)$ together with a set C, such that

$$P_\theta(T(X) \in C) \leq \alpha, \forall \theta \in \Theta_0.$$

The test **reject** the hypothesis H_0 if and only if $T(X) \in C$. T is called the **test statistic** and C is called the **critical region** of the test.
A statistical test can fail in 2 different ways

- **Type I error**: $\theta \in \Theta_0$ but $T(X) \in C$ (H_0 is wrongly rejected despite being true).
- **Type II error**: $\theta \notin \Theta_0$ but $T(X) \notin C$ (H_0 is wrongly not rejected despite being wrong).

Hypothesis tests - Example

Skewness

$$\gamma = E\left(\left(\frac{X-\mu}{\sigma}\right)^3\right) = \frac{E\left((X-\mu)^3\right)}{\sigma^3}$$

of a random variable X with mean μ and standard deviation σ can be estimated by

$$\hat{\gamma}_n = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{3/2}},$$

where $X_1, ..., X_n$ are iid copies of X and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\gamma = 0$ and

$$\sqrt{\frac{n}{\sigma}}\hat{\gamma}_n \longrightarrow \mathcal{N}(0,1)$$

as $n \to \infty$.

Hypothesis tests - Example (cont.)

Assume that we want to construct a test for the null hypothesis

$$H_0 : X \sim \mathcal{N}(., \sigma^2).$$

For large n, we can use the test statistic

$$T = \sqrt{n/\sigma}|\hat{\gamma}_n|$$

and the critical region

$$C = (1.96, \infty) \subset \mathbb{R}$$

to construct a test of size $\alpha = 5\%$. We reject H_0 if and only if

$$|\hat{\gamma}_n| \geq 1.96\sqrt{\sigma/n}.$$

Hypothesis tests - Example (cont.)

One problem with the test constructed in the preceding example is that the convergence of the distribution of $\sqrt{n/\sigma}\hat{\gamma}_n$ to $\mathcal{N}(0,1)$ is very slow. For small or moderate n, the probability of wrongly rejecting H_0 (type I error) may be bigger than α.

We can use Monte Carlo method to estimate the probability of type I errors of statistical tests as follows:

- For $j = 1, 2, ..., N$, generate samples $(X_1^{(j)}, ..., X_n^{(j)})$ according to the distribution given by the H_0.
- Compute $T^{(j)} = T(X_1^{(j)}, ..., X_n^{(j)})$ for $j = 1, 2, ..., N$.
- Check for which percentage of samples H_0 is (wrongly) rejected:

$$P(T \in C) \approx \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}_C(T^{(j)}).$$

References

Chapter 3. Jochen Voss. *An Introduction to Statistical Computing - A Simulation-based Approach*. John Wiley & Sons, 2014.

Chapter 1, 5. J. S. Dagpunar. *Simulation and Monte Carlo - With applications in finance and MCMC*. John Wiley & Sons, 2007.