Bài 4 - Phương pháp Monte Carlo (Monte Carlo methods)

Thống kê máy tính và ứng dụng (Computational Statistics and Applications)

Vũ Quốc Hoàng (vqhoang@fit.hcmus.edu.vn)
FIT - HCMUS

Ngày 5 tháng 2 năm 2023

Nội dung

1. Minh họa mở đầu
2. Phương pháp Monte Carlo
3. Ước lượng Monte Carlo
4. Các phương pháp giảm phương sai
5. Một số ứng dụng cho suy diễn thống kê

Nội dung

Minh họa mở đầu

Yêu cầu. Tính tích phân

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

Trả lời. $I = \Gamma(0.9 + 1) = \Gamma(1.9)$, với Γ là hàm gamma, được định nghĩa

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

Giá trị này không có **công thức đóng** (closed-form formula) để tính nhưng có thể được xấp xỉ bằng **phân tích số** (numerical analysis)

$$I \approx 0.9618.$$

Tuy nhiên, các phương pháp số gặp khó khăn khi tính các tích phân nhiều chiều. Minh họa này dùng các số ngẫu nhiên để tính, gọi là **phương pháp Monte Carlo** (Monte Carlo method), mà có thể giúp vượt qua khó khăn trên.

Minh họa mở đầu - Phương án 1

Nhận xét I có thể được viết lại là

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

với $f(x) = e^{-x}\mathbb{I}_{[0,\infty)}(x)$ là hàm mật độ xác suất của phân phối $\text{Exp}(\lambda = 1)$. Như vậy

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

Từ **mẫu ngẫu nhiên** (random sample) $X_1, X_2, ..., X_N \overset{\text{iid}}{\sim} \text{Exp}(1)$, ta có thể **ước lượng** (estimate) I bằng **thống kê** $\hat{I} = \frac{1}{N}\sum_{i=1}^{N} X_i^{0.9}$.
Ta đã biết $\hat{I}$ là một **ước lượng không chệch** (unbiased estimator) của I, nghĩa là $E(\hat{I}) = I$, với **sai số chuẩn** (standard error)

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

Minh họa mở đầu - Phương án 1 (tt)

Dùng phương pháp sinh số ngẫu nhiên cho phân phối $Exp(1)$ từ bài trước ($U \sim \mathcal{U}(0,1)$ thì $X = -\ln U \sim \mathrm{Exp}(1)$), ta sinh $N = 10$ số ngẫu nhiên $X_1, \ldots, X_{10}$ cụ thể và có

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

Cũng từ mẫu này ta ước lượng $\sigma_f(X^{0.9})$ bằng thống kê

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

Từ đó ta có ước lương cho sai số chuẩn $\hat{\sigma}(\hat{I}) = \frac{s}{\sqrt{N}} = 0.4118$.

Vì $\sigma(\hat{I}) \propto \frac{1}{\sqrt{N}}$ nên để có sai số chuẩn 0.001 ta cần $N \approx 1.7 \times 10^6$ số ngẫu nhiên $\mathcal{U}(0,1)$.

Minh họa mở đầu - Phương án 2

Ta cũng nhận xét, I có thể được viết lại là

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

với $g(x) = xe^{-x}\mathbb{I}_{[0,\infty)}(x)$ là hàm mật độ xác suất của phân phối Erlang$(k = 2, \lambda = 1)$.
Như vậy

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

Vì $X = X_1 + X_2 \sim \text{Erlang}(2,1)$ nếu $X_1, X_2 \overset{\text{iid}}{\sim} \text{Exp}(1)$ nên ta có thể sinh số ngẫu nhiên cho $X \sim \text{Erlang}(2,1)$ bằng cách sinh $U_1, U_2 \sim \mathcal{U}(0,1)$ và trả về $X = -\ln U_1 U_2$.
Cũng dùng $N = 10$ số ngẫu nhiên $\mathcal{U}(0,1)$ ta có $\hat{I} = \frac{1}{N/2}\sum_{i=1}^{N/2}\frac{1}{X_i^{0.1}} = 0.9408$ với sai số chuẩn $\hat{\sigma}(\hat{I}) = 0.0375$. Để có sai số chuẩn 0.001 ta cần $N \approx 14000$ số ngẫu nhiên $\mathcal{U}(0,1)$.

Nội dung

Phương pháp Monte Carlo

Phương pháp Monte Carlo (Monte Carlo method)

- Sinh mẫu từ mô hình thống kê bằng mô phỏng trên máy tính.
- Dùng các tính chất thống kê của mẫu để nghiên cứu mô hình.

Nói nôm na, Monte Carlo là phương pháp "giải bài toán bằng ngẫu nhiên".

Ví dụ

- Kỳ vọng của một biến ngẫu nhiên có thể được xấp xỉ bằng trung bình của một lượng lớn các mẫu của biến ngẫu nhiên.
- Xác suất của một biến cố có thể được xấp xỉ bằng tần suất biến cố xảy ra trong một lượng lớn lần lặp lại thí nghiệm.
- Chất lượng của một phương pháp phân tích có thể được đánh giá bằng cách sinh nhiều dữ liệu từ một mô hình được chọn và đối chiếu kết quả phân tích được từ phương pháp với các tính chất đã biết của mô hình.

Phương pháp Monte Carlo (tt)

Rất nhiều bài toán được đưa về việc tính $E(f(X))$ với X là "đối tượng" ngẫu nhiên mô tả hệ thống cần nghiên cứu và f là một hàm giá trị thực mô tả đại lượng quan tâm từ hệ thống. Giá trị trên thường được tính bằng 3 cách

1. Tính toán chính xác bằng giải tích. Chẳng hạn, nếu phân phối của X có hàm mật độ là φ, ta có
$$E(f(X)) = \int f(x)\varphi(x)dx.$$
Phương pháp này gặp khó khăn khi không biết φ hoặc khó tính tích phân.
2. Xấp xỉ bằng các phương pháp số. Phương pháp này gặp khó khi không gian của X phức tạp (chẳng hạn, có số chiều lớn).
3. Ước lượng bằng Monte Carlo. Phương pháp này dựa trên **dạng mạnh của luật số lớn** (strong law of large numbers): nếu $(X_j)_{j\in\mathbb{N}}$ là một dãy iid như X thì
$$E(f(X)) = \lim_{N\to\infty} \frac{1}{N}\sum_{j=1}^{N} f(X_j) \text{ (với xác suất 1).}$$

Phương pháp Monte Carlo - Ví dụ 1

Với $X \sim \mathcal{N}(0,1)$, sinh $X_1, ..., X_N$ iid như X, ta có thể ước lượng

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

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

Với $N = 10000$, sinh mẫu cụ thể ta có

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

Phương pháp Monte Carlo - Ví dụ 2

Cho biến ngẫu nhiên X, ta có $P(X \in A) = E(\mathbb{I}_A(X))$. Do đó, ta có ước lượng

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

Ví dụ. Với $X \sim \mathcal{N}(0,1), a = -1$, sinh $X_1, ..., X_N$ iid như X, ta có thể ước lượng

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

Với $N = 10000$, sinh mẫu cụ thể ta có

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

Đối chiếu với giá trị

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

Phương pháp Monte Carlo - Ví dụ 3

Cho $X \sim \mathcal{U}(a, b)$ với hàm mật độ xác suất $\varphi(x) = \frac{1}{b-a}\mathbb{I}_{[a,b]}$, ta có ước lượng

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

Ví dụ. Sinh $X_1, ..., X_N \overset{\text{iid}}{\sim} \mathcal{U}(0, 2\pi)$ ta có thể ước lượng

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

Với $N = 10000$, sinh mẫu cụ thể ta có

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

Phương pháp Monte Carlo - Ví dụ 4

Xét một bài toán suy diễn Bayes đơn giản: ta muốn suy diễn về $X \sim \text{Exp}(1)$ bằng cách sử dụng một quan sát y từ $Y \sim \mathcal{N}(0, X)$. Để giải bài toán này, ta cần tìm phân phối hậu nghiệm $X|Y = y$, là phân phối có điều kiện của X khi biết $Y = y$.

Trước hết, phân phối tiên nghiệm của X, $X \sim \text{Exp}(1)$, có hàm mật độ

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

Phân phối có điều kiện của $Y \sim \mathcal{N}(0, X)$ khi biết $X = x$ có hàm mật độ

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

Từ đó, dùng công thức Bayes, phân phối hậu nghiệm $X|Y = y$ có hàm mật độ

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

Phương pháp Monte Carlo - Ví dụ 4 (tt)

Đặt

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

ta có thể dùng phương pháp lấy mẫu loại bỏ để sinh mẫu từ phân phối hậu nghiệm $X|Y = y$. Lưu ý, phương pháp lấy mẫu loại bỏ không cần f phải được chuẩn hóa.

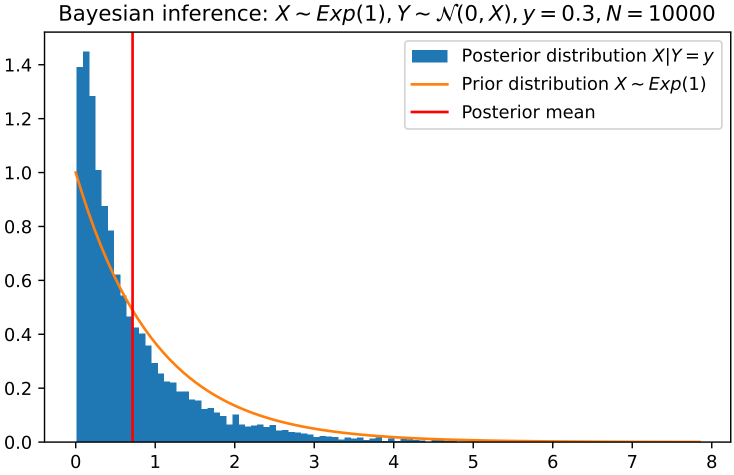
Cụ thể, dùng thuật toán lấy mẫu loại bỏ theo khuôn với phân phối đề cử là Exp(1) và hằng số

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

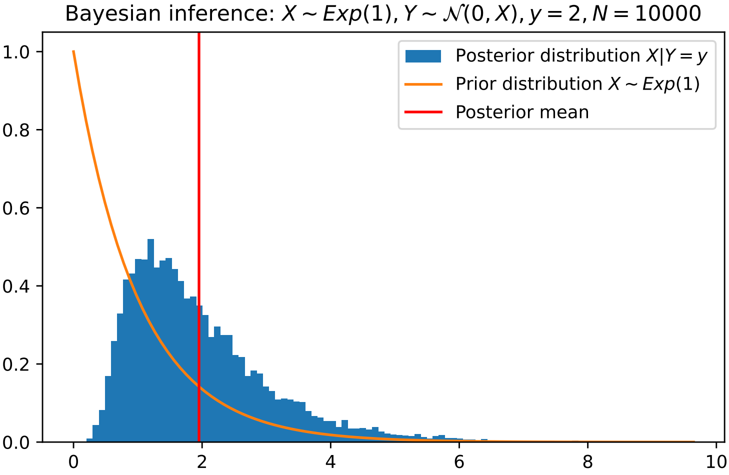
ta sinh mẫu $X_1, ..., X_N$ iid theo phân phối hậu nghiệm $X|Y = y$, vẽ histogram và tính

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

Phương pháp Monte Carlo - Ví dụ 4 (tt)



Phương pháp Monte Carlo - Ví dụ 4 (tt)



Nội dung

Ước lượng Monte Carlo

Cho biến ngẫu nhiên X và hàm giá trị thực f, **ước lượng Monte Carlo** (Monte Carlo estimate) cho $E(f(X))$ được định nghĩa là

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

với $X_1, ..., X_N$ iid như X.
Lưu ý: các X_j ngẫu nhiên nên ước lượng Z_N^{MC} là biến ngẫu nhiên.

Ước lượng Monte Carlo (tt)

Thuật toán MCS. (Monte Carlo estimate)

Input:

- phân phối của X,
- hàm f giá trị thực,
- $N \in \mathbb{N}$.

Output: ước lượng Z_N^{MC} cho $E(f(X))$.

$s \leftarrow 0$
for $j = 1, 2, ..., N$ **do**
    sinh X_j cùng phân phối với X
    $s \leftarrow s + f(X_j)$
end for
return s/N

Sai số Monte Carlo

Với $\hat{\theta} = \hat{\theta}(X)$ là một ước lượng cho tham số θ, ta định nghĩa

- **Độ chệch** (bias) của ước lượng là

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

- **Sai số chuẩn** (standard error)

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

- **Trung bình bình phương sai số** (mean squared error - MSE)

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

- **Căn trung bình bình phương sai số** (root-mean-square error - RMSE)

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

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

Sai số Monte Carlo (tt)

Vì Z_N^{MC} là ngẫu nhiên nên sai số Monte Carlo $Z_N^{MC} - E(f(X))$ cũng ngẫu nhiên. Để đánh giá sai số Monte Carlo, ta sử dụng các khái niệm trên từ thống kê.

Mệnh đề. Ước lượng Monte Carlo Z_N^{MC} cho $E(f(X))$

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

có

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

và

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

Sai số Monte Carlo - Ví dụ

Trở lại ví dụ trước, cho $X \sim \mathcal{N}(0,1)$, ước lượng Monte Carlo cho $E(\sin(X)^2)$ là

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

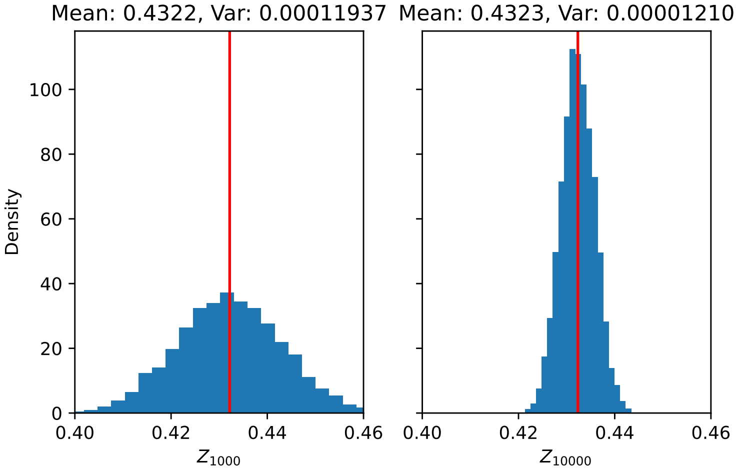
với $X_1, X_2, ..., X_N \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$.

Mệnh đề trên cho biết Z_N^{MC} là ước lượng không chệch cho $E(\sin(X)^2)$, tức là $\text{bias}(Z_N^{MC}) = 0$, và

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

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

Sai số Monte Carlo - Ví dụ (tt)



Lựa chọn cỡ mẫu

Nếu không biết $\text{Var}(f(X))$, ta có thể ước lượng

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

với

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

là phương sai mẫu của $f(X_1), f(X_2), ..., f(X_N)$.

Để lựa chọn cỡ mẫu N phù hợp (N nhỏ thì sai số lớn, còn N lớn thì chi phí tính toán lớn), ta có thể chạy thử với N "vừa phải", ước lượng sai số, từ đó điều chỉnh cỡ mẫu cho phù hợp ($\text{RMSE}(Z_N^{MC}) \propto \frac{1}{\sqrt{N}}$) và chạy lại. (Xem minh họa mở đầu.)

Lựa chọn cỡ mẫu (tt)

Nếu biết $Var(f(X))$ hoặc biết chặn trên của nó thì để đạt được sai số $\text{MSE}(Z_N^{MC}) \leq \epsilon^2$, N phải thỏa

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

Ví dụ 1. Giả sử $Var(f(X)) = 1$, để ước lượng $E(f(X))$ với $\text{MSE} \leq \epsilon^2 = 0.01^2$, ta có thể sử dụng ước lượng Monte Carlo với

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

Lựa chọn cỡ mẫu (tt)

Ví dụ 2. Cho X là một biến ngẫu nhiên có giá trị thực và $A \subseteq \mathbb{R}$. Từ ví dụ trước, ta có thể ước lượng $p = P(X \in A)$ bằng

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

Ta có

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

Do đó, ta có thể đạt được MSE $\leq \epsilon^2$ bằng cách chọn

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

Mặc dù không biết p nhưng $p(1-p) \leq 1/4$, $\forall p \in [0,1]$ nên ta có thể chọn

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

Đánh giá chi tiết sai số

Từ **định lý giới hạn trung tâm** (central limit theorem), ta có mệnh đề sau, giúp đánh giá chi tiết sai số Monte Carlo.

Mệnh đề. Cho $\alpha \in (0,1)$, đặt $q_\alpha = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ với Φ là CDF của phân phối $\mathcal{N}(0,1)$, đặt $\sigma^2 = Var(f(X))$ thì với

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

ước lượng Monte Carlo Z_N^{MC} cho $E(f(X))$ thỏa (khi N đủ lớn)

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

Cách khác để mô tả kết quả của Mệnh đề trên là thay thế ước lượng điểm Z_N^{MC} bằng khoảng tin cậy

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

Đánh giá chi tiết sai số (tt)

Với $\alpha = 5\%$, từ Mệnh đề trên ta có $q_{0.05} = \Phi^{-1}(0.975) \approx 1.96$, do đó ta cần

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

để có sai số tuyệt đối không quá ϵ với xác suất ít nhất là 95%. Cỡ mẫu này gấp gần 4 lần cỡ mẫu để có $\mathrm{RMSE}(Z_N^{MC}) \leq \epsilon$.

Ví dụ. Giả sử $Var(f(X)) = 1$, để ước lượng Z_N^{MC} cho $E(f(X))$ có sai số tuyệt đối $\left|Z_N^{MC} - E(f(X))\right|$ không quá $\epsilon = 0.01$ với xác suất ít nhất là $1 - \alpha = 95\%$, ta có thể dùng cỡ mẫu

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

Đánh giá chi tiết sai số (tt)

Nếu không biết $\sigma^2 = Var(f(X))$ thì ta có thể dùng

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

là phương sai mẫu của $f(X_1), f(X_2), ..., f(X_N)$ để ước lượng cho σ^2.

Tương ứng, $q_\alpha = \Phi^{-1} \left(1 - \frac{\alpha}{2}\right) = P(\mathcal{N}(0,1) \geq \alpha/2)$ nên được thay thế bằng $q_\alpha^{N-1} = P(\text{Student}(N-1) \geq \alpha/2)$ với Student$(N-1)$ là phân phối Student có $N-1$ bậc tự do. Tuy nhiên, khi N khá lớn thì $q_\alpha \approx q_\alpha^{N-1}$.

Ví dụ. Trong minh họa mở đầu, dùng 10 số ngẫu nhiên $\mathcal{U}(0,1)$, khoảng tin cậy 95% cho $I = \int_0^\infty x^{0.9} e^{-x} dx$ theo Phương án 1 là $[0.8917, 1.3096]$ còn theo Phương án 2 là $[0.8367, 1.0450]$. Hơn nữa, nếu dùng 14000 số thì khoảng tin cậy là $[0.9594, 0.9632]$.

Nội dung

Giới thiệu

Như ta đã thấy, ước lượng Monte Carlo Z_N^{MC} cho $E(f(X))$ có trung bình bình phương sai số

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

Để tăng tính hiệu quả của ước lượng, ta tìm cách giảm phương sai ($\mathrm{Var}(Z_N^{MC})$).

Lấy mẫu quan trọng

Cho X là biến ngẫu nhiên với hàm mật độ ϕ, f là hàm giá trị thực, ψ là một hàm mật độ với $\psi(x) > 0$ khi $f(x)\phi(x) > 0$, ta có

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

Từ đó, **ước lượng lấy mẫu quan trọng** (importance sampling estimate) cho $E(f(X))$ được định nghĩa là

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

trong đó $Y_1, Y_2, ..., Y_N$ iid với hàm mật độ ψ.

Lấy mẫu quan trọng (tt)

Thuật toán IS. (Importance Sampling)

Input:

- hàm f giá trị thực,
- hàm mật độ ϕ của X,
- hàm mật độ ψ,
- $N \in \mathbb{N}$.

Output: ước lượng Z_N^{IS} cho $E(f(X))$.

$s \leftarrow 0$
for $j = 1, 2, ..., N$ **do**
    sinh $Y_j \sim \psi$
    $s \leftarrow s + f(Y_j)\phi(Y_j)/\psi(Y_j)$
end for
return s/N

Lấy mẫu quan trọng(tt)

Mệnh đề. Ước lượng lấy mẫu quan trọng $Z_N^{IS} = \frac{1}{N}\sum_{j=1}^{N} \frac{f(Y_j)\phi(Y_j)}{\psi(Y_j)}$ cho $E(f(X))$ có $\text{bias}(Z_N^{IS}) = 0$ và

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

Như vậy, phương pháp lấy mẫu quan trọng cho hiệu quả khi

- $Y_1, Y_2, ..., Y_N$ có thể được sinh một cách hiệu quả từ hàm mật độ ψ,
- $\text{Var}\left(f(Y)\phi(Y)/\psi(Y)\right)$ nhỏ, tức là, hằng số $c_\psi = E\left(f(X)^2\left(1 - \frac{\phi(X)}{\psi(X)}\right)\right)$ lớn. Ta có được điều này khi ψ cùng dạng với $f\phi$ hay ψ lớn khi $|f|$ lớn. (Đây là lí do phương pháp này có tên là lấy mẫu quan trọng!)

Lấy mẫu quan trọng - Ví dụ

Cho X là một biến ngẫu nhiên nhận giá trị thực và $A \subset \mathbb{R}$, ước lượng lấy mẫu quan trọng cho $P(X \in A) = E(\mathbb{I}_A(X))$ là

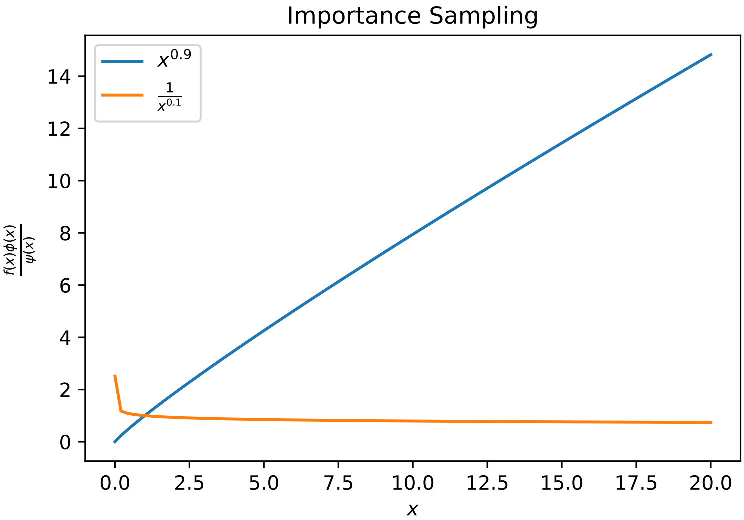
$$Z_N^{IS} = \frac{1}{N} \sum_{j=1}^{N} \frac{\mathbb{I}_A(Y_j)\phi(Y_j)}{\psi(Y_j)},$$

trong đó, ψ là một hàm mật độ xác suất thỏa $\psi(x) > 0$ với mọi $x \in A$ mà $\phi(x) > 0$, và $Y_1, Y_2, ..., Y_N$ iid với hàm mật độ ψ. Ta có

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

Như vậy, $\text{MSE}(Z_N^{IS}) < \text{MSE}(Z_N^{MC})$ khi ta chọn được $\psi > \phi$ trên A.

Lấy mẫu quan trọng - Xem lại minh họa mở đầu



Biến đối nghịch

Cho X, X' cùng phân phối nhưng không nhất thiết độc lập, ta có

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

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

Ước lượng biến đối nghịch (antithetic variables estimate) cho $E(f(X))$ với cỡ mẫu $N \in 2\mathbb{N}$ được định nghĩa là

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

trong đó các (X_k, X_k') iid như (X, X') và được gọi là các **cặp biến đối nghịch** (antithetic pair).

Biến đổi nghịch (tt)

Thuật toán AV. (antithetic variables)
Input:

- hàm f giá trị thực,
- $N \in \mathbb{N}$ chẵn.

Output: ước lượng Z_N^{AV} cho $E(f(X))$.

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

Biến đổi nghịch (tt)

Mệnh đề. Cho X, X' là 2 biến ngẫu nhiên cùng phân phối với $\rho = \text{Cor}(f(X), f(X'))$, ước lượng biến đổi nghịch $Z_N^{AV} = \frac{1}{N}\sum_{k=1}^{N/2}(f(X_k) + f(X_k'))$ cho $E(f(X))$ có

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

và

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

Như vậy, phương pháp biến đổi nghịch cho hiệu quả khi $\rho < 0$. Tức là, ta cần X, X' cùng phân phối và $f(X), f(X')$ có tương quan nghịch.

Biến đối nghịch (tt)

Ý tưởng đầu tiên giúp xây dựng các cặp biến đối nghịch là: nếu X có phân phối đối xứng thì ta có thể dùng $X' = -X$, khi đó ta thường có $\text{Cor}(f(X), f(X')) < 0$.

Ví dụ 1. Cho $X \sim \mathcal{N}(0, 1)$, cần ước lượng xác suất

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

Vì phân phối của X đối xứng nên ta có thể thử dùng cặp (X, X') với $X' = -X$. Với lựa chọn này, ta có

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

Do đó, từ Mệnh đề trên, ta có

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

Vì $p \approx 0.16$ nên $\rho \approx -0.19$, do đó $\text{MSE}(Z_N^{AV}) \approx 81\%\text{MSE}(Z_N^{MC})$.

Biến đối nghịch (tt)

Một ý tưởng khác giúp xây dựng các cặp biến đối nghịch là: nếu X có hàm phân phối F dễ tìm hàm ngược thì ta có thể dùng $X = F^{-1}(U), X' = F^{-1}(1-U)$ với $U \sim \mathcal{U}(0,1)$, khi đó

- X, X' cùng phân phối do $U, 1-U$ cùng phân phối $\mathcal{U}(0,1)$,
- $\text{Cor}(X, X') \leq 0$ vì F^{-1} đơn điệu (tăng). Hơn nữa, nếu f đơn điệu (tăng hoặc giảm) thì ta có $\text{Cor}(f(X), f(X')) \leq 0$ như mệnh đề sau cho thấy.

Mệnh đề. Nếu $g : \mathbb{R} \to \mathbb{R}$ là hàm đơn điệu (tăng hoặc giảm) và $U \sim \mathcal{U}(0,1)$ thì

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

Ví dụ 2. Trong ví dụ mở đầu, ta muốn tính tích phân

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

Biến đổi nghịch - Ví dụ (tt)

Ta có thể dùng ước lượng Monte Carlo

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

với $U_1, U_2, ..., U_N \overset{\text{iid}}{\sim} \mathcal{U}(0,1)$. Vì $x^{0.9}$ là hàm đơn điệu (tăng) nên ta có thể thử dùng ước lượng biến đổi nghịch

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

Hệ số tương quan $\rho = \text{Cor}(f(X), f(X')) = \text{Cor}\left((-\ln(U))^{0.9}, (-\ln(1-U))^{0.9}\right)$ khó tính nhưng dễ dàng ước lượng từ mẫu $\hat{\rho} = -0.7114$. Do đó $\text{MSE}(Z_N^{AV}) \approx 29\%\text{MSE}(Z_N^{MC})$.

Biến kiểm soát

Cho f, g là các hàm giá trị thực với $g \approx f$, ta có

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

Nếu ta có thể tính $E(g(X))$ thì ta có thể ước lượng $E(f(X))$ qua ước lượng của $E(f(X) - g(X))$. Vì $g \approx f$ nên $\mathrm{Var}(f(X) - g(X)) < \mathrm{Var}(f(X))$.

Cho g là hàm với $E(g(X))$ đã biết, **ước lượng biến kiểm soát** (control variates estimate) cho $E(f(X))$ được định nghĩa là

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

trong đó $X_1, X_2, ..., X_N$ iid như X. Biến ngẫu nhiên $g(X)$ được gọi là **biến kiểm soát** (control variate) cho $f(X)$.

Biến kiểm soát (tt)

Thuật toán CV. (control variates)
Input:

- hàm f giá trị thực,
- hàm $g \approx f$ với $E(g(X))$ đã biết,
- $N \in \mathbb{N}$ chẵn.

Output: ước lượng Z_N^{CV} cho $E(f(X))$.

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

Biến kiểm soát (tt)

Mệnh đề. Ước lượng biến kiểm soát $Z_N^{CV} = \frac{1}{N}\sum_{j=1}^{N}(f(X_j) - g(X_j)) + E(g(X))$ có

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

và

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

Như vậy, phương pháp biến kiểm soát cho hiệu quả khi

- Tìm được hàm $g \approx f$ sao cho $E(g(X))$ dễ tính,
- $f(X) - g(X)$ có phương sai nhỏ hơn phương sai của $f(X)$.

Biến kiểm soát (tt)

Phương pháp biến kiểm soát mô tả ở trên là một trường hợp đặc biệt của một phương pháp tổng quát hơn. Cụ thể, nếu dùng biến kiểm soát có tương quan Y thì mọi biến ngẫu nhiên Z đều có thể được biến đổi thành biến ngẫu nhiên mới $\tilde{Z}$ có cùng kì vọng nhưng phương sai nhỏ hơn như Mệnh đề sau cho thấy.

Mệnh đề. Cho biến ngẫu nhiên Z với $E(Z) = \mu$ và $\text{Var}(Z) = \sigma^2$, biến ngẫu nhiên Y có $\text{Cor}(Y, Z) = \rho$, định nghĩa

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

thì biến ngẫu nhiên $\tilde{Z}$ có $E(\tilde{Z}) = \mu$ và

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

Biến kiểm soát - Ví dụ

Ta muốn tính tích phân $I = \int_0^1 e^{x^2} dx$. Tích phân này khó tính nhưng có một tích phân "gần giống" như vậy lại dễ tính hơn

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

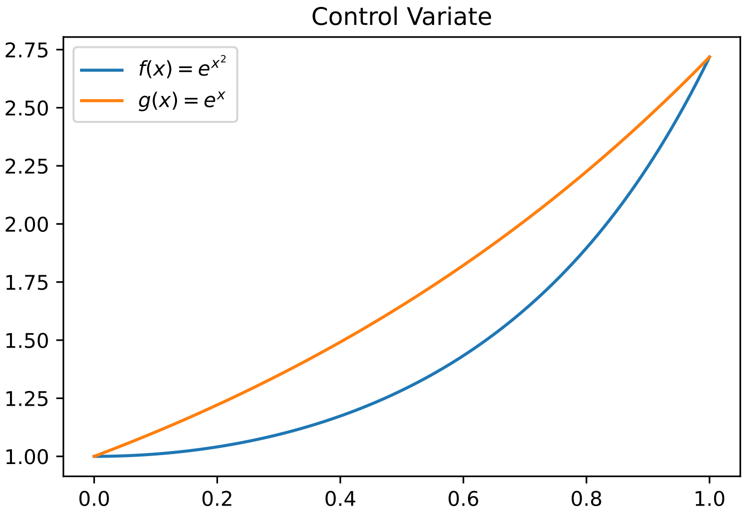
Với $f(x) = e^{x^2}, g(x) = e^x, X \sim \mathcal{U}(0,1)$, ta có $I = E(f(X)), J = E(g(X))$. Sinh $X_1, X_2, ..., X_N \overset{\text{iid}}{\sim} \mathcal{U}(0,1)$, ta có ước lượng Monte Carlo cho I là $Z_N^{MC} = \frac{1}{N}\sum_{j=1}^N f(X_j)$ và ước lượng biến kiểm soát cho I là $Z_N^{CV} = \frac{1}{N}\sum_{j=1}^N (f(X_j) - g(X_j)) + e - 1$.

Với $N = 10000$, sinh mẫu cụ thể ta có

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

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

Biến kiểm soát - Ví dụ (tt)



Nội dung

Giới thiệu

Bài toán suy diễn thống kê

- Ta có dữ liệu quan sát $x = (x_1, ..., x_n)$.
- Ta xem xét họ $(P_\theta)_{\theta \in \Theta}$ các phân phối xác suất, trong đó θ là vector tham số của mô hình và Θ là không gian tham số.
- Ta giả sử dữ liệu quan sát là một mẫu của biến ngẫu nhiên $X = (X_1, ..., X_n) \sim P_\theta$ với giá trị tham số $\theta \in \Theta$ chưa biết.
- Ta xác định mô hình P_θ mà dữ liệu x được sinh ra từ đó.

Ước lượng điểm

- Một **ước lượng điểm** (point estimator) cho tham số θ là hàm bất kì của mẫu ngẫu nhiên X với giá trị trong Θ, kí hiệu $\hat{\theta} = \hat{\theta}(X) = \hat{\theta}(X_1, ..., X_n)$.
- **Độ chệch** (bias) của ước lượng $\hat{\theta} = \hat{\theta}(X)$ cho tham số θ được định nghĩa là

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

- Với mỗi θ, ước lượng Monte Carlo cho độ chệch là

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

  trong đó các $X^{(j)} = (X_1^{(j)}, ..., X_n^{(j)})$ iid như X.
- *Lưu ý:* n là kích thước mẫu dữ liệu còn N là số mẫu dùng trong ước lượng Monte Carlo; $\hat{\theta}$ là một ước lượng cho θ còn $\widehat{\text{bias}_\theta}(\hat{\theta})$ là một ước lượng cho độ chệch của ước lượng $\hat{\theta}$.

Ước lượng điểm - Ví dụ

Cho $\rho \in [-1,1]$ và $X, \eta \sim \mathcal{N}(0,1)$, định nghĩa

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

Khi đó

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

Hệ số tương quan này có thể được ước lượng bằng **hệ số tương quan mẫu** (sample correlation)

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

trong đó $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ và các (X_i, Y_i) iid như (X, Y).

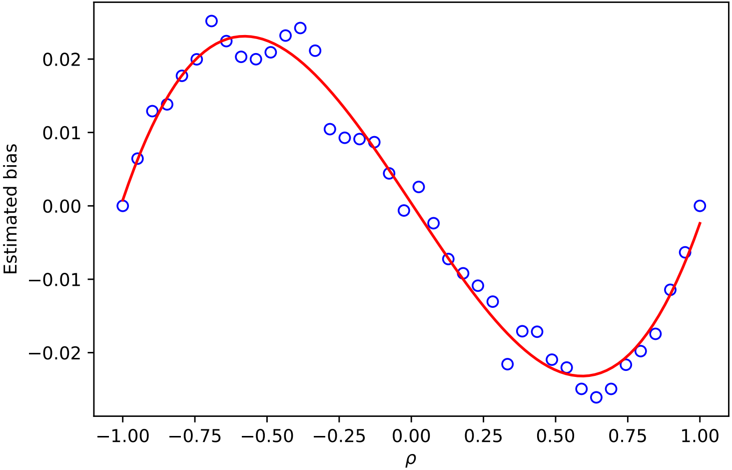
Ước lượng điểm - Ví dụ (tt)

Với mỗi ρ, ước lượng Monte Carlo $\widehat{\text{bias}_\rho}(\hat{\rho})$ có thể được tính như sau

1. $S \leftarrow 0$
2. **for** $j = 1, 2, ..., N$ **do**
3. sinh $X_1^{(j)}, ..., X_n^{(j)} \sim \mathcal{N}(0, 1)$
4. sinh $\eta_1^{(j)}, ..., \eta_n^{(j)} \sim \mathcal{N}(0, 1)$
5. đặt $Y_i^{(j)} = \rho X_i^{(j)} + \sqrt{1 - \rho^2}\eta_i^{(j)}, i = 1, 2, ..., n$
6. tính $\hat{\rho}^{(j)} = \hat{\rho}(X^{(j)}, Y^{(j)})$ là hệ số tương quan mẫu
7. $S \leftarrow S + \hat{\rho}^{(j)}$
5. **end for**
6. trả về $S/N - \rho$

Ước lượng $\widehat{\text{bias}_\rho}(\hat{\rho})$ có thể được tính cho các giá trị khác nhau của $\rho \in [-1, 1]$ để thấy sự phụ thuộc của độ chệch vào tham số ρ.

Ước lượng điểm - Ví dụ (tt)



Ước lượng điểm (tt)

Sai số chuẩn (standard error) của ước lượng $\hat{\theta} = \hat{\theta}(X)$ cho tham số θ được định nghĩa là

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

Với mỗi θ, ước lượng Monte Carlo cho sai số chuẩn là

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

trong đó

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

và các $X^{(j)}$ iid như X.

Khoảng tin cậy

Một **khoảng tin cậy** (confidence interval) với độ tin cậy $1-\alpha$ cho tham số θ là một khoảng ngẫu nhiên $[U, V] \subset \mathbb{R}$ với $U = U(X), V = V(X)$ là các hàm của mẫu ngẫu nhiên $X = (X_1, ..., X_n)$ sao cho

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

Trong nhiều trường hợp, khoảng tin cậy cho tham số θ có thể được xây dựng dựa trên ước lượng điểm $\hat{\theta} = \hat{\theta}(X)$ cho θ theo dạng

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

trong đó giá trị $\epsilon > 0$ được chọn phù hợp từ phân phối của $\hat{\theta}-\theta$.

Khoảng tin cậy - Ví dụ

Cho $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ với phương sai σ^2 đã biết, xây dựng khoảng tin cậy cho tham số chưa biết μ.

Ước lượng điểm hay dùng cho μ là

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

với

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

Từ đó, nếu chọn $\epsilon = \frac{\sigma q_\alpha}{\sqrt{n}}$ với $q_\alpha = \Phi^{-1}(1-\alpha/2)$ thì ta có một khoảng tin cậy $1-\alpha$ cho μ là

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

Khoảng tin cậy - Ví dụ (tt)

Nếu phương sai σ^2 chưa biết, ta có thể dùng khoảng tin cậy

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

trong đó

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

và q_α^{n-1} là phân vị mức $1 - \alpha/2$ của phân phối Student với $n - 1$ bậc tự do.

Khoảng tin cậy (tt)

Trường hợp các X_i không có phân phối chuẩn và n nhỏ thì các khoảng tin cậy trên không còn chính xác nữa. Ta có thể dùng các ước lượng Monte Carlo để xây dựng hoặc đánh giá các khoảng tin cậy dựa trên xấp xỉ

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

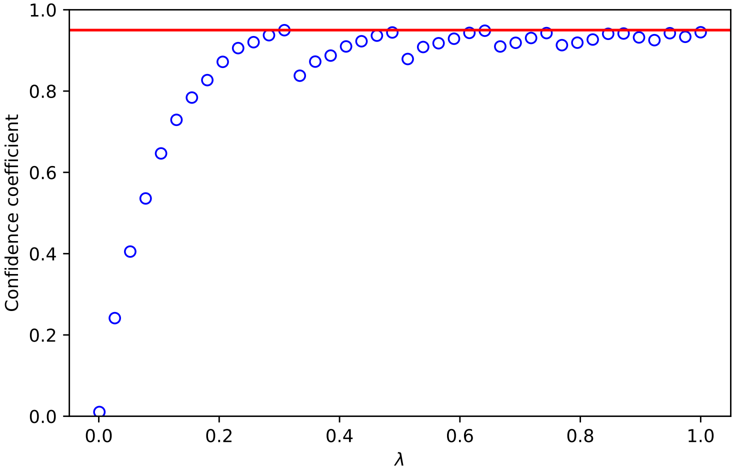
trong đó các $X^{(j)} = (X_1^{(j)}, ..., X_n^{(j)})$ iid như $X = (X_1, ..., X_n)$.

Khoảng tin cậy - Ví dụ

Cho $X_1, \ldots, X_n$ độc lập và cùng phân phối Poisson với tham số λ, ta đánh giá khoảng tin cậy $P_\lambda(U \leq \lambda \leq V)$ ở trên với mỗi λ được cho bằng ước lượng Monte Carlo như sau

1. $k \leftarrow 0$
2. **for** $j = 1, 2, \ldots, N$ **do**
3. sinh $X_1, \ldots, X_n \sim Poisson(\lambda)$
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. trả về k/N

Khoảng tin cậy (tt)



Kiểm định giả thuyết

Một **kiểm định giả thuyết thống kê** (statistical hypothesis test) cỡ $\alpha \in (0, 1)$ cho giả thuyết $H_0 = \{\theta \in \Theta_0\}$ ($\Theta_0 \subset \Theta$) được cho bởi một hàm $T = T(X)$ của mẫu ngẫu nhiên $X = (X_1, ..., X_n)$ và một tập C sao cho

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

Kiểm định **bác bỏ** (reject) giả thuyết H_0 khi và chỉ khi $T(X) \in C$. T được gọi là **thống kê kiểm định** (test statistic) và C được gọi là **miền bác bỏ** (critical region) của kiểm định.

Một kiểm định có thể sai lầm theo 2 cách

- **Sai lầm loại I** (type I error): $\theta \in \Theta_0$ nhưng $T(X) \in C$ (H_0 bị bác bỏ nhầm).
- **Sai lầm loại II** (type II error): $\theta \notin \Theta_0$ nhưng $T(X) \notin C$ (H_0 không được bác bỏ đúng).

Kiểm định giả thuyết - Ví dụ

Độ lệch (skewness)

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

của biến ngẫu nhiên X với kỳ vọng μ và độ lệch chuẩn σ có thể được ước lượng bởi

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

trong đó $X_1, ..., X_n$ iid như X và $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.
Nếu $X \sim \mathcal{N}(\mu, \sigma^2)$ thì $\gamma = 0$ và

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

khi $n \to \infty$.

Kiểm định giả thuyết - Ví dụ (tt)

Giả sử ta muốn xây dựng một kiểm định cho giả thuyết

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

Với n lớn, ta có thể dùng thống kê kiểm định

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

với miền bác bỏ

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

để xây dựng một kiểm định cỡ $\alpha = 5\%$. Ta bác bỏ H_0 khi và chỉ khi

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

Kiểm định giả thuyết - Ví dụ (tt)

Một vấn đề với kiểm định trên là sự hội tụ theo phân phối của $\sqrt{n/\sigma}\hat{\gamma}_n$ về $\mathcal{N}(0,1)$ là rất chậm. Với n không đủ lớn, xác suất bác bỏ nhầm H_0 (sai lầm loại I) có thể lớn hơn α.

Ta có thể dùng ước lượng Monte Carlo để ước lượng xác suất sai lầm loại I của kiểm định như sau:

- Với $j = 1, 2, ..., N$, sinh $(X_1^{(j)}, ..., X_n^{(j)})$ theo phân phối được cho bởi giả thuyết H_0.
- Tính $T^{(j)} = T(X_1^{(j)}, ..., X_n^{(j)})$ với mỗi $j = 1, 2, ..., N$.
- Tính tỉ lệ số lần H_0 bị bác bỏ (nhầm):

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

Tài liệu tham khảo

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.