Bài giảng R, số 5
-Tích phân và phương pháp Monte Carlo với R-

TS.Tô Đức Khánh

10/03/2025

1 Tích phân cơ bản

Xét các bài toán tích phân cơ bản

$$I = \int_a^b f(x)\,\mathrm{d}x$$

Ta có 3 phương pháp cơ bản để xấp xỉ tích phân này như sau:

- Phương pháp Riemann
$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=0}^{n-1}(x_{i+1}-x_i)f\left(x_i\right).$$

- Phương pháp Trapezoidal (hình thang)
$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=0}^{n-1}\frac{x_{i+1}-x_i}{2}\left(f\left(x_i\right)+f\left(x_{i+1}\right)\right).$$

- Phương pháp Simpson
$$\int_a^b f(x)\,\mathrm{d}x \approx \sum_{i=0}^{n-1}\frac{x_{i+1}-x_i}{6}\left[f(x_i)+4f\left(\frac{x_i+x_{i+1}}{2}\right)+f(x_{i+1})\right].$$

Ta dễ dàng thi hành các đoạn công thức tính của các phương pháp trong `R`.

Thực hành 1.1: Viết các hàm tính tích phân theo 3 phương pháp được liệt kê ở trên.

Thực hành 1.2: Áp dụng các hàm trên để tính các tích phân sau:

(a) $\displaystyle\int_0^\pi x\sin(x)\,\mathrm{d}x$

(b) $\displaystyle\int_0^1 x^2\exp(-x)\,\mathrm{d}x$

(c) $\displaystyle\int_0^1 \exp(2x)\sin(3x)\,\mathrm{d}x$

So sánh kết quả này với kết quả thu được bởi việc sử dụng hàm `integrate()`.

Bài tập

Bài tập 1.1: Xét tích phân

$$I = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x$$

với μ và σ là hai tham số cho trước.

(a) Sử dụng phép đổi biến $u = \dfrac{1}{1+\exp(-x)}$, xác định dạng mới của tích phân theo u.

(b) Áp dụng các hàm tính tích phân để xác định giá trị của tích phân trong các trường hợp: (i) $a = 0, \mu = 2, \sigma = 1.2$; (ii) $a = -0.6, \mu = 0, \sigma = 0.5$; (iii) $a = 2, \mu = 1.3, \sigma = 2$.

So sánh kết quả này với kết quả thu được bởi việc sử dụng hàm `integrate()` và `pnorm()`.

Bài tập 1.2: Sử dụng các thuật toán trên, hãy viết hàm tính giá trị của hàm phân phối xác suất tích lũy (cdf) của biến ngẫu nhiên có phân phối Cauchy với hàm mật độ xác suất được cho bởi:

$$f(x) = \frac{1}{\theta\pi\left(1 + ((x-\alpha)/\theta)^2\right)},$$

với $-\infty < x < +\infty$. So sánh kết quả với kết quả thu được bởi việc sử dụng hàm `pcauchy()`.

2 Phương pháp Tích phân Monte Carlo

2.1 Phương pháp mô phỏng giá trị ngẫu nhiên

Trong `R`, các biến ngẫu nhiên thông dụng đều có các hàm mô phỏng giá trị ngẫu nhiên:

- phân phối đều $\mathcal{U}(a, b)$, `runif(n, min, max)`;
- phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$, `rnorm(n, mean, sd)`;
- phân phối Cauchy $\mathcal{C}(\alpha, \beta)$, `rcauchy(n, location, scale)`, với `location` là α và `scale` là β;
- phân phối mũ $\mathrm{Exp}(\lambda)$, `rexp(n, rate)`.
- phân phối Gamma $\mathcal{G}(\alpha, \beta)$, `rgamma(n, shape, rate)` , với `shape` là α và `rate` là β.

Bên cạnh đó, ta có thể phát sinh dữ liệu ngẫu nhiên bằng các phương pháp

- Đổi biến dựa trên biến phân phối đều trên (0,1), tức là $\mathcal{U}(0, 1)$.
- Dựa trên hàm ngược của hàm phân phối xác suất tích lũy
- Rejection Sampling

Đối với đổi biến, ta có bảng tổng hợp các phép biến đổi cho các phân phối như sau

- Phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$: tạo $U_1 \sim \mathcal{U}(0, 1)$ và $U_2 \sim \mathcal{U}(0, 1)$, với U_1 và U_2 độc lập nhau; xác định hai kết quả bởi $X_1 = \mu + \sigma\sqrt{-2\log(U_1)}\cos(2\pi U_2)$ và $X_2 = \mu + \sigma\sqrt{-2\log(U_1)}\sin(2\pi U_2)$.
- Phân phối log-normal $\mathcal{LN}(\mu, \sigma^2)$: tạo $Y \sim \mathcal{N}(\mu, \sigma^2)$; xác định kết quả $X = \exp(Y)$.
- Phân phối Cauchy $\mathcal{C}(\alpha, \beta)$: tạo $U \sim \mathcal{U}(0, 1)$, xác định kết quả $X = \alpha + \beta\tan\{\pi(U - 0.5)\}$.
- Phân phối mũ $\mathrm{Exp}(\lambda)$: tạo $U \sim \mathcal{U}(0, 1)$, xác định kết quả $X = \log(U)/\lambda$.
- Phân phối Chi-square χ^2_ν: tạo 1 dãy gồm ν biến ngẫu nhiên $Y_1 \ldots, Y_\nu$ tuân theo phân phối chuẩn $\mathcal{N}(0, 1)$; xác định kết quả $X = \sum_{i=1}^{\nu} Y_i$.
- Phân phối t-student t_ν: tạo biến $Y \sim \mathcal{N}(0, 1)$, $Z \sim \chi^2_\nu$, xác định kết quả $X = Y/\sqrt{Z/\nu}$.

Đối với phương pháp dùng hàm ngược, ta thực hiện các bước sau:

1. xác định hàm ngược $F_X^{-1}(\cdot)$;
2. ta tạo ngẫu nhiên $U \sim \mathcal{U}(0,1)$;
3. xác định X thông qua $F_X^{-1}(U)$.

Trong trường hợp, ta không có công thức tường minh của $F_X^{-1}(\cdot)$, thì việc xác định X sẽ thực hiện thông qua phép nội suy tuyến tính (linear interpolation), như sau.

1. xác định vùng quan sát khả dĩ cho hàm mật độ $f_X(\cdot)$;
2. tạo một lưới $x_1, \ldots, x_m$ trên vùng quan sát khả dĩ;
3. tính $u_i = F_X(x_i)$;
4. tạo ngẫu nhiên $U \sim \mathcal{U}(0,1)$
5. xác định X theo đường tuyến tính nội suy giữa hai điểm liền kề $u_i \leq U \leq u_j$

$$X = \frac{u_j - U}{u_j - u_i} x_i + \frac{U - u_i}{u_j - u_i} x_j.$$

Thực hành 2.1.1: Áp dụng phương pháp đổi biến để mô phỏng dữ liệu theo các phân phối sau:

- Phân phối chuẩn $\mathcal{N}(2, 3^2)$.
- Phân phối Chi-square χ_5^2.
- Phân phối t-student t_4.

Sử dụng Quantile-Quantile plot được tạo bởi `qqplot()` và `qqline()` để kiểm chứng kết quả.

Thực hành 2.1.2: Áp dụng phương pháp dùng hàm ngược, mô phỏng dữ liệu theo các phân phối mũ $\text{Exp}(\lambda)$ với $\lambda = 3$. Sử dụng Quantile-Quantile plot được tạo bởi `qqplot()` và `qqline()` để kiểm chứng kết quả.

Bài tập

Bài tập 2.1.1: Xét biến ngẫu nhiên có phân phối Weibull với hàm xác suất được cho bởi

$$f(x; \theta, \alpha) = \frac{\alpha}{\theta} \left(\frac{x}{\theta}\right)^{\alpha-1} \exp\left\{-\left(\frac{x}{\theta}\right)^{\alpha}\right\},$$

với $x > 0$ và $\theta, \alpha > 0$. Ta tính được

$$F(x; \theta, \alpha) = 1 - \exp\left\{-\left(\frac{x}{\theta}\right)^{\alpha}\right\}.$$

Áp dụng phương pháp dùng hàm ngược, mô phỏng dữ liệu theo phân phối Weibull với, $\theta = 2$, $\alpha = 2$.

Bài tập 2.1.2: Áp dụng phương pháp dùng hàm ngược và nội suy tuyến tính hãy tạo ngẫu nhiên giá trị của biến ngẫu nhiên tuân theo phân phối chuẩn $\mathcal{N}(2, 1)$. Chú ý, để xác định vùng khả dĩ cho lưới, ta có thể áp dụng quy tắc $k\sigma$ của phân phối chuẩn.

Bài tập 2.1.3: Áp dụng phương pháp dùng hàm ngược và nội suy tuyến tính hãy tạo ngẫu nhiên giá trị của biến ngẫu nhiên tuân theo phân phối có hàm mật độ được cho bởi

$$f(x; \mu) = \mu^2 x \exp(-\mu x),$$

với $x \geq 0$ và $\mu > 0$. Xét 02 trường hợp $\mu = 1/2$ và $\mu = 3$. Chú ý, để xác định vùng khả dĩ cho lưới, ta có thể giải phương trình $F_X(x) = 0.99999$ tìm x (tức là x là giới hạn trên của vùng có xác suất quan sát được giá trị của X là 99.999%). Trong trường hợp này, phương trình có thể giải bằng phương pháp lặp Newton-Raphson.

2.2 Tích phân Monte Carlo

Xét tích phân

$$I = \int_{\mathbb{R}^p} g(\boldsymbol{x}) f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

Xấp xỉ Monte Carlo

$$\widehat{I} = \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{X}_i)$$

trong đó, mẫu ngẫu nhiên $\mathbf{X}_1, \ldots, \mathbf{X}_n$ được tạo từ phân phối với hàm mật độ $f(\boldsymbol{x})$.

Ví dụ 2.2.1: Xét tích phân

$$I = \int\limits_{-\infty}^{+\infty} (3 - x^2) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \, \mathrm{d}x.$$

Ta tách được $g(x) = 3 - x^2$, và

$$f(x) = frac1\sqrt{2\pi} \exp\left(-\frac{x^2}{2}\right)$$

là hàm mật độ của phân phối chuẩn $\mathcal{N}(0,1)$. Theo đó, ta sẽ phát sinh mẫu ngẫu nhiên gồm n quan sát tuân theo phân phối chuẩn $\mathcal{N}(0,1)$. Và áp dụng công thức xấp xỉ tích phân Monte Carlo

$$\widehat{I} = \frac{1}{n} \sum_{i=1}^{n} (3 - X_i^2),$$

Trong `R`, ta định nghĩa

```
I_est <- function(data){
  return(mean(3 - data^2))
}
```

Với $n = 15$, ta thực hiện các bước như sau:

```
n_sim <- 15
x_sim <- rnorm(n = n_sim, mean = 0, sd = 1)
I_est(data = x_sim)
```

```
## [1] 1.9628
```

Để kiểm chứng độ chính xác của ước lượng tích phân Monte Carlo, ta thực hiện một quy trình mô phỏng Monte Carlo, với 1000 lần lặp, trong mỗi lần lặp ta phát sinh 1 mẫu ngẫu nhiên có cỡ 15 và tính ước lượng tích phân Monte Carlo.

```
out_sim <- sapply(1:1000, function(i, n_sim){
  x_sim <- rnorm(n = n_sim, mean = 0, sd = 1)
  return(I_est(data = x_sim))
}, n_sim = 15)
```

Tính trung bình và phương sai, ta thu được.

```
mean(out_sim)
```

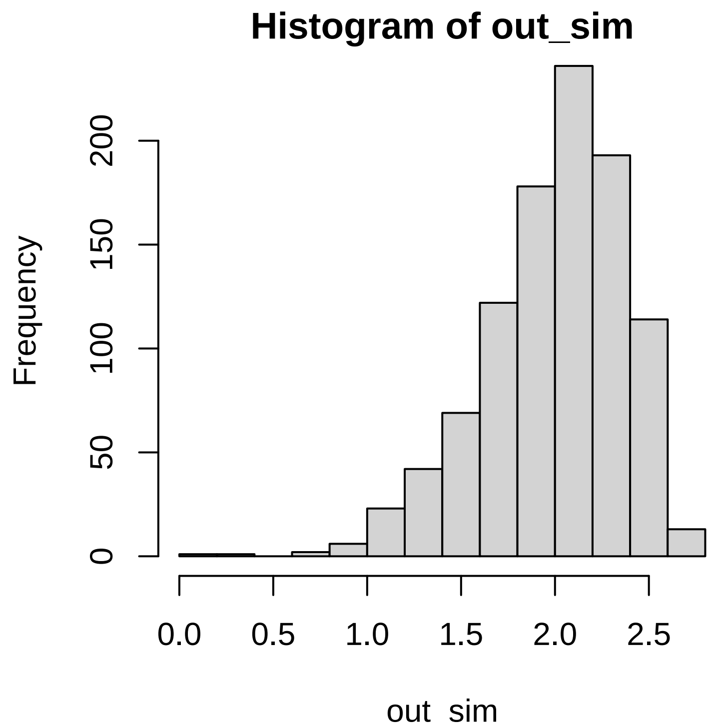
```
## [1] 1.994098
```

```
var(out_sim)
```

```
## [1] 0.1388133
```

Histogram cho thấy sự xấp xỉ phân phối chuẩn là chưa đạt được với cỡ mẫu $n = 15$.

```
hist(out_sim)
```



Thực hành 2.2.1: Tăng cỡ mẫu n và lập lại quá trình tính trong Ví dụ 2.2.1. Vẽ histogram và nhận xét.

Ví dụ 2.2.2: Xét bài toán xấp xỉ số π theo tích phân bội

$$\pi = 4\int_0^1\int_0^1 \mathrm{I}(x_1^2 + x_2^2 \leq 1)\frac{1}{4}\,\mathrm{d}x_1\mathrm{d}x_2.$$

Tích phân bội này có thể phân tách thành tích phân của hàm $\mathrm{I}(X_1^2 + X_2^2 \leq 1)$, trong đó, X_1 và X_2 là hai biến ngẫu nhiên độc lập theo phân phối đều $\mathcal{U}(-1, 1)$ với hàm mật độ xác suất đồng thời là

$$f(x_1, x_2) = f(x_1)f(x_2) = \frac{1}{4}.$$

Xấp xỉ Monte Carlo cho π là

$$\pi \approx 4\,\frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(X_{1i}^2 + X_{2i}^2 \leq 1).$$

Trong R, ta định nghĩa

```
pi_est <- function(x1, x2) {
  n <- length(x1)
  return(4*mean(x1^2 + x2^2 <= 1))
}
```

Với $n = 15$, ta thực hiện các bước như sau:

```r
n_sim <- 15
X1 <- runif(n = n_sim, min = -1, max = 1)
X2 <- runif(n = n_sim, min = -1, max = 1)
pi_est(X1, X2)
```

```
## [1] 3.466667
```

Kết quả thu được là chưa đủ gần với $\pi \approx 3.141593$. Thực hiện mô phỏng Monte Carlo với 1000 lần lặp để kiểm chứng độ chính xác của xấp xỉ này.

```r
out_pi_sim <- sapply(1:1000, function(i, n_sim){
  X1 <- runif(n = n_sim, min = -1, max = 1)
  X2 <- runif(n = n_sim, min = -1, max = 1)
  return(pi_est(X1, X2))
}, n_sim = 15)
```

Tính trung bình và phương sai, ta thu được.

```r
mean(out_pi_sim)
```

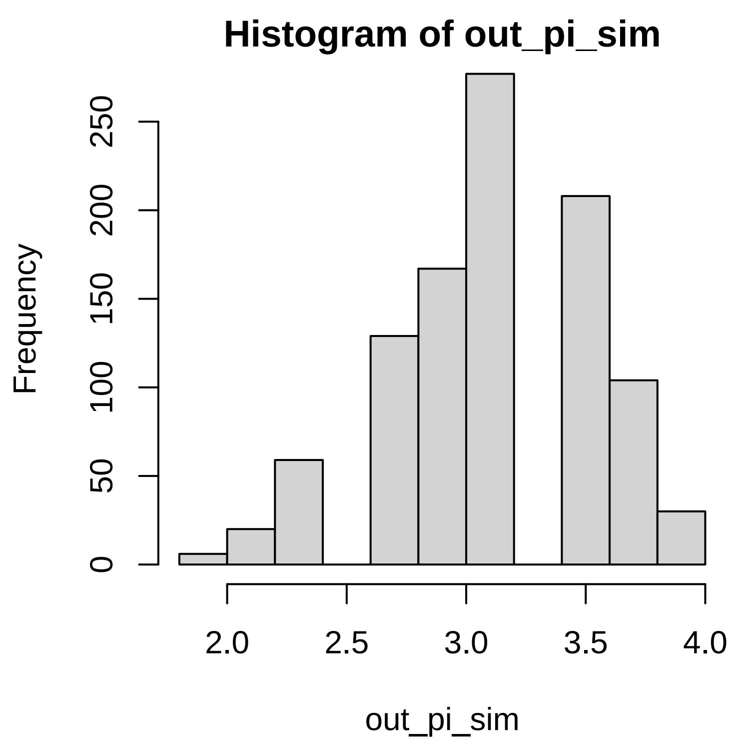
```
## [1] 3.145067
```

```r
var(out_pi_sim)
```

```
## [1] 0.1804873
```

Histogram cho thấy sự xấp xỉ phân phối chuẩn là chưa đạt được với cỡ mẫu $n = 15$.

```r
hist(out_pi_sim)
```



Thực hành 2.2.2: Tăng cỡ mẫu n và lặp lại quá trình tính trong Ví dụ 2.2.2. Vẽ histogram và nhận xét.

Bài tập

Bài tập 2.2.1: Áp dụng phương pháp tích phân Monte Carlo tính các tích phân sau:

- $I = \int_0^1 x^2 \mathrm{d}x.$
- $I = \int_0^\pi \sqrt{x^3 + \sqrt{x}} - x^2 \sin(4x) \, \mathrm{d}x.$
- $I = \int_0^1 \int_{-2}^2 x^2 \cos(xy) \, \mathrm{d}x\mathrm{d}y.$
- $I = \int_0^\infty \frac{3}{4} x^4 \exp(-x^3/4) \, \mathrm{d}x.$

2.3 Importance sampling

Xét tích phân:

$$I = \int_{\mathbb{R}^p} g(\boldsymbol{x}) f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

Đối với importance sampling, xấp xỉ Monte Carlo cho tích phân này được tính bởi:

$$\widehat{I}_{IS} = \frac{1}{n} \sum_{i=1}^n g(\mathbf{X}_i) \frac{f(\mathbf{X}_{\boldsymbol{i}})}{h(\mathbf{X}_{\boldsymbol{i}})} = \frac{1}{n} \sum_{i=1}^n g(\mathbf{X}_i) w(\mathbf{X}_i),$$

với bộ mẫu ngẫu nhiên $\mathbf{X}_1, \ldots, \mathbf{X}_n$ độc lập, được tạo ngẫu nhiên từ hàm mật độ $h_{\mathbf{X}}(\cdot)$, với $\mathbf{X}_i = (X_{1i}, \ldots, X_{ni})$.

Về mặt kỹ thuật, ta cần chọn $h(\cdot)$ thỏa các tiêu chí sau:

- $h(x)$ là hàm mật độ xác suất trong vùng xác định của X;
- $\frac{f(x)}{h(x)} = c < +\infty$;
- $h(x)$ có đuôi nặng hơn $f(x)$ (tức là giá trị của $h(x)$ tại các điểm x nhỏ hoặc lớn, là không quá nhỏ so với $f(x)$).

Ví dụ 2.3.1: Xét tích phân sau

$$I = \int_0^1 \frac{\exp(-x)}{1 + x^2} \, \mathrm{d}x.$$

Dễ thấy rằng, tích phân này tương đương với

$$I = \int_0^1 \frac{\exp(-x)}{1 + x^2} 1 \, \mathrm{d}x = \mathbb{E}_X \left(\frac{\exp(-X)}{1 + X^2} \right),$$

với X là biến ngẫu nhiên tuân theo phân phối đều $\mathcal{U}(0, 1)$. Do đó, ta có thể áp dụng xấp xỉ Monte Carlo thông thường.

- tạo bộ biến ngẫu nhiên $X_1, \ldots, X_n$ tuân theo phân phối đều $\mathcal{U}(0, 1)$
- tính xấp xỉ

$$\widehat{I}_1 = \frac{1}{n} \sum_{i=1}^n \frac{\exp(-X_i)}{(1 + X_i^2)}.$$

Tuy nhiên, phương sai của ước lượng $\widehat{I}_1$ có thể lớn. Ta sẽ áp dụng thuật toán importance sampling.

Trước tiên, ta nhận xét rằng

$$I = \int_0^1 \frac{\exp(-x)}{1+x^2}\frac{h(x)}{h(x)}\,\mathrm{d}x = \mathbb{E}_h\left(\frac{\exp(-X)}{1+X^2}\frac{1}{h(X)}\right),$$

với $h(x)$ là một hàm mật độ của X xác định trên $(0,1)$. Để xác định $h(x)$, ta phân thành hai trường hợp sau:

1. $g(x) = \dfrac{1}{1+x^2}$, $f(x) = \exp(-x)$: dễ thấy $f(x)$ có dạng là hàm mật độ xác suất của phân phối mũ $\mathrm{Exp}(1)$ (có vùng xác định là $(0,+\infty)$). Theo như tiêu chí chọn $h(x)$, ta có thể chọn $h(x) \propto f(x)$. Tuy nhiên, $h(x) = \exp(-x)$ không phải là hàm mật độ xác suất trên $(0,1)$, do tích phân toàn miền của $h(x)$ không bằng 1:
$$\int_0^1 h(x)\,\mathrm{d}x = \int_0^1 \exp(-x)\,\mathrm{d}x = 1-\exp(-1).$$
Do đó, ta có thể chọn $h(x) = \dfrac{\exp(-x)}{1-\exp(-1)}$ với $x \in (0,1)$. Khi đó, $\dfrac{f(x)}{h(x)} = 1-\exp(-1) \approx 0.63212$ với mọi $x \in (0,1)$, tỷ lệ này là phù hợp với lý thuyết.

2. $g(x) = \pi\exp(-x)$, $f(x) = \dfrac{1}{\pi(1+x^2)}$: dễ thấy $f(x)$ có dạng là hàm mật độ xác suất của phân phối Cauchy $\mathcal{C}(0,1)$ (có vùng xác định là $(-\infty,+\infty)$). Theo như tiêu chí chọn $h(x)$, ta có thể chọn $h(x) \propto f(x)$. Tuy nhiên, $h(x) = \dfrac{1}{\pi(1+x^2)}$ không phải là hàm mật độ xác suất trên $(0,1)$, do tích phân toàn miền của $h(x)$ không bằng 1:
$$\int_0^1 h(x)\,\mathrm{d}x = \int_0^1 \frac{1}{\pi(1+x^2)}\,\mathrm{d}x = \frac{1}{4}.$$
Do đó, ta có thể chọn $h(x) = 4\dfrac{1}{\pi(1+x^2)}$ với $x \in (0,1)$. Khi đó, $\dfrac{f(x)}{h(x)} = 0.25$ với mọi $x \in (0,1)$, tỷ lệ này là phù hợp với lý thuyết.

Như vậy, ta có hai lựa chọn cho $h(x)$ và có hai xấp xỉ tương ứng

- $h(x) = \dfrac{\exp(-x)}{1-\exp(-1)}$, xấp xỉ Monte Carlo importance sampling
$$\widehat{I}_2 = \frac{1}{n}\sum_{i=1}^n \frac{1-\exp(-1)}{1+X_i^2},$$
với $X_1,\ldots,X_n$ là mẫu ngẫu nhiên tạo từ phân phối có hàm mật độ xác suất $h(x) = \dfrac{\exp(-x)}{1-\exp(-1)}$.

- $h(x) = 4\dfrac{1}{\pi(1+x^2)}$, xấp xỉ Monte Carlo importance sampling
$$\widehat{I}_3 = \frac{1}{n}\sum_{i=1}^n \frac{\pi\exp(-X_i)}{4},$$
với $X_1,\ldots,X_n$ là mẫu ngẫu nhiên tạo từ phân phối có hàm mật độ xác suất $h(x) = 4\dfrac{1}{\pi(1+x^2)}$.

Để tạo ngẫu nhiên theo các phân phối này, ta dùng phương pháp hàm ngược. Cụ thể như sau:

- với $h(x) = \dfrac{\exp(-x)}{1-\exp(-1)}$, hàm phân phối tích lũy là
$$F_1(x) = \int_0^x \frac{\exp(-s)}{1-\exp(-1)}\,\mathrm{d}s = \frac{1-\exp(-x)}{1-\exp(-1)},$$

với $x \in (0, 1)$. Từ đây, suy ra công thức xác định X theo hàm ngược

$$X = F_1^{-1}(U) = -\log\left(1 - U(1 - \exp(-1))\right),$$

với $U \sim \mathcal{U}(0, 1)$.

- với $h(x) = 4\dfrac{1}{\pi(1 + x^2)}$, hàm phân phối tích lũy là

$$F_2(x) = \int_0^x \frac{4}{\pi(1 + s^2)}\,\mathrm{d}s = \frac{4}{\pi}\tan^{-1}(x),$$

với $x \in (0, 1)$. Từ đây, suy ra công thức xác định X theo hàm ngược

$$X = F_1^{-1}(U) = \tan\left(\frac{U\pi}{4}\right),$$

với $U \sim \mathcal{U}(0, 1)$.

Ta tiến hành thi hành thuật toán trong `R` như sau:

```
n_sim <- 50
u_sim <- runif(n_sim)
```

```
x_2 <- -log(1 - u_sim*(1 - exp(-1)))
I2_est <- (1 - exp(-1)) * mean(1/(1 + x_2^2))
I2_est
```

```
## [1] 0.5303541
```

```
x_3 <- tan(u_sim*pi/4)
I3_est <- pi * mean(exp(-x_3))/4
I3_est
```

```
## [1] 0.5418026
```

Thực hiện mô phỏng Monte Carlo 5000 lần để kiểm tra độ hiệu quả của các ước lượng.

```
out_sim_is <- sapply(1:5000, function(i, nsim) {
  u_sim <- runif(n_sim)
  ## standard Monte Carlo
  x_1 <- u_sim
  I1_est <- mean(exp(-x_1)/(1 + x_1^2))
  ## IS 1
  x_2 <- -log(1 - u_sim*(1 - exp(-1)))
  I2_est <- (1 - exp(-1)) * mean(1/(1 + x_2^2))
  ## IS 2
  x_3 <- tan(u_sim*pi/4)
  I3_est <- pi * mean(exp(-x_3))/4
  res <- c(I1_est = I1_est, I2_est = I2_est, I3_est = I3_est)
  return(res)
}, nsim = 50)
```

Trung bình cho thấy các kết quả khá gần nhau, và gần với giá trị chính xác 0.5247971.

```
rowMeans(out_sim_is)
```

```
##    I1_est    I2_est    I3_est
## 0.5253496 0.5249918 0.5251186
```

Độ lệch chuẩn của các ước lượng cho thấy các ước lượng Monte Carlo Importance Sampling cho độ lệch chuẩn (hoặc phương sai) thấp hơn ước lượng Monte Carlo thông thường `I1_est`.

```r
apply(out_sim_is, 1, sd)
```

```
##     I1_est     I2_est     I3_est
## 0.03467524 0.01385255 0.02004829
```

Ngoài ra, ta cũng thấy rằng, độ lệch chuẩn của phương pháp IS thứ 2 là cao hơn so với phương pháp IS thứ 1.

Thực hành 2.3.1 Xét bài toán suy luận Bayes cho trung bình của phân phối location-scale t-student.

$$f(x;\mu,\sigma,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{(x-\mu)^2}{\sigma\nu}\right)^{-(\nu+1)/2},$$

trong đó, $-\infty < \mu < +\infty$ là tham số cho location (vị trí của vùng dữ liệu), $\sigma > 0$ là tham số cho scale (biên độ rộng của phân phối), và $\nu > 0$ là bậc tự do.

Ta xét trường hợp, $\sigma = 1$ và $\nu = 3$. Khi đó

$$f(x;\mu) = \frac{\Gamma(2)}{\sqrt{3\pi}\Gamma\left(\frac{3}{2}\right)}\left(1+\frac{(x-\mu)^2}{3}\right)^{-2}.$$

Giả sử ta có n quan sát $x_1,\ldots,x_n$ tuân theo phân phối này. Với non-informative prior (tiên nghiệm không có thông tin), $\pi(\theta) = 1$, ta tìm được phân phối hậu nghiệm

$$\pi(\mu;x_1,\ldots,x_n) = \frac{\prod\limits_{i=1}^{n} f(x_i|\mu)}{\int\limits_{-\infty}^{+\infty}\prod\limits_{i=1}^{n} f(x_i|\mu)\,\mathrm{d}\mu} = \frac{\prod\limits_{i=1}^{n}\left(3+(x_i-\mu)^2\right)^{-2}}{\int\limits_{-\infty}^{+\infty}\prod\limits_{i=1}^{n}\left(3+(x_i-\mu)^2\right)^{-2}\,\mathrm{d}\mu}$$

Ta tính được kỳ vọng hậu nghiệm cho μ là

$$\mathbb{E}(\mu|x_1,\ldots,x_n) = \frac{\int\limits_{-\infty}^{+\infty}\mu\prod\limits_{i=1}^{n} f(x_i|\mu)\,\mathrm{d}\mu}{\int\limits_{-\infty}^{+\infty}\prod\limits_{i=1}^{n} f(x_i|\mu)\,\mathrm{d}\mu} = \frac{\int\limits_{-\infty}^{+\infty}\mu\prod\limits_{i=1}^{n}\left(3+(x_i-\mu)^2\right)^{-2}\,\mathrm{d}\mu}{\int\limits_{-\infty}^{+\infty}\prod\limits_{i=1}^{n}\left(3+(x_i-\mu)^2\right)^{-2}\,\mathrm{d}\mu}$$

Hãy áp dụng phương pháp Importance sampling để tính xấp xỉ tích phân này. Chú ý, μ là biến ngẫu nhiên.

Bài tập

Bài tập 2.3.1: Áp dụng phương pháp Monte Carlo importance sampling để tính tích phân sau

$$\int_1^{+\infty}\frac{x^2}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x$$

Bài tập 2.3.2: Xét bài toán tính $\sigma^2 = \mathbb{E}(X^2)$ với X là biến ngẫu nhiên có hàm mật độ xác suất tỷ lệ với hàm $h(x) = \exp(-|x|^3/3)$, với $x \in (-\infty,\infty)$.

(a) Sử dụng phương pháp Monte Carlo importance sampling tính xấp xỉ σ^2.

(b) Một ước lượng khác là sử dụng biến thể của tích phân Riemann, cụ thể

$$\sigma_R^2 = \frac{\sum\limits_{i=1}^{n-1} X_{[i]}^2\left(X_{[i+1]}-X_{[i]}\right)h(X_{[i]})}{\sum\limits_{i=1}^{n-1}\left(X_{[i+1]}-X_{[i]}\right)h(X_{[i]})},$$

trong đó, $X_{[1]} \leq X_{[2]} \ldots \leq X_{[n]}$ là mẫu đã được sắp xếp tăng dần liên quan đến giá trị của biến ngẫu nhiên $X_1, X_2, \ldots, X_n$. Hãy áp dụng phương pháp này.

(c) Thực hiện mô phỏng Monte Carlo để so sánh sự hiệu quả của hai phương pháp trong câu (a) và (b).