R 语言-5

郑泽靖　zzjstat2023@163.com

北京师范大学统计学院

2025 年 11 月 20 日

蒙特卡罗方法简介

定义

蒙特卡罗 (Monte Carlo) 方法是一种以概率统计理论为指导的数值计算方法。它使用随机数（或伪随机数）来解决计算问题。

- **核心思想**：当所求解的问题是某随机事件出现的概率，或者是某随机变量的数学期望时，通过“实验”的方法得出频率，以此逼近概率或期望。
- **应用场景**：数值积分, 物理模拟等。

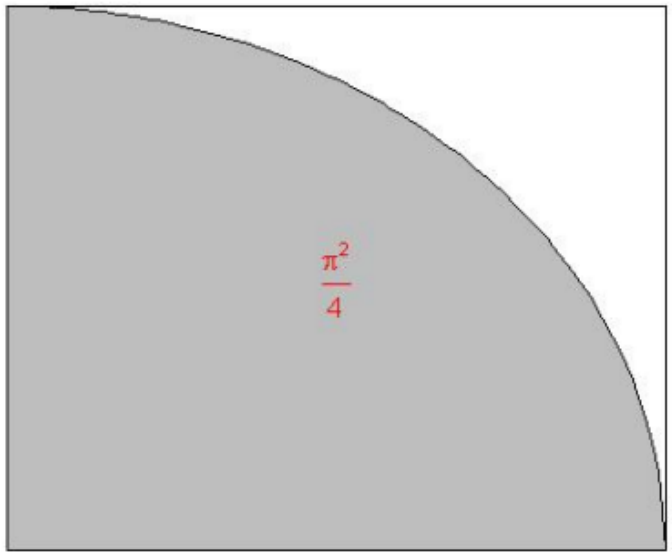
经典案例：计算圆周率 π

考虑单位正方形 S 和其内切圆 C（第一象限）：

$$\begin{cases} 0 \le X \le 1 \\ 0 \le Y \le 1 \end{cases}$$

若 $X, Y \sim U(0,1)$ 独立同分布，则点落入圆内的概率为：

$$P(X^2 + Y^2 \le 1) = \frac{\text{圆面积}/4}{\text{正方形面积}} = \frac{\pi/4}{1} = \frac{\pi}{4}$$



R 代码实现：计算 π

基于大数定律，我们可以估计：

$$\pi \approx 4 \times \frac{\#\{X_i^2 + Y_i^2 \leq 1\}}{n}$$

```r
p_est <- function(n) {
    x <- runif(n); y <- runif(n)
    k <- (x^2 + y^2 <= 1)
    return(4 * sum(k) / n)
}

num <- c(1e3, 1e4, 1e5, 1e6) # 样本量
results <- c()
for (N in num) {
    results <- c(results, p_est(N))
}

data.frame(N=num, Est=results,
           Error=abs(results - pi))
```

理论基石：大数定律

Kolmogorov 强大数定律

若 $X_1, X_2, \ldots, X_n$ 是独立同分布 (i.i.d.) 的随机变量且 $\mathrm{E}|X| < \infty$，则：

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} X_k = \mathrm{E}(X)\right) = 1$$

这意味着：只要样本量 n 足够大，样本均值将几乎处处收敛于理论期望。

应用 1：估计期望 $\mathrm{E}(X)$

以 $X \sim U(0,1)$ 为例，理论值 $\mathrm{E}(X) = 0.5$。

```r
mean_unif <- function(n) {
    x <- runif(n, 0, 1)
    mean(x)
}
num <- c(10, 100, 1000, 10000, 100000)
results <- c()
for (N in num) {
    results <- c(results, mean_unif(N))
}
print(results)
```

随着 n 增加，样本均值迅速逼近 0.5。

应用 2：估计概率 $\mathrm{P}(A)$

概率可以看作是示性函数 $I_A(X)$ 的期望：

$$\mathrm{P}(A) = \mathrm{E}[I_A(X)] \approx \frac{1}{n}\sum_{i=1}^{n} I_A(X_i) = \text{频率}$$

以 $X \sim B(10, 0.5)$ 为例，估计 $P(X = 4)$：

```r
p_binom <- function(n) {
    x <- rbinom(n, size = 10, prob = 0.5)
    mean(x == 4) # 使用 mean() 计算逻辑向量的比例
}
results <- c()
for (N in num) {
    results <- c(results, p_binom(N))
}
```

理论值：‘dbinom(4, 10, 0.5)‘ ≈ 0.205。

应用 3：估计复杂函数的期望

假设 $X_1, X_2 \sim N(0,1)$，估计 $\theta = E|X_1 - X_2|$。

此处解析解较复杂，但蒙特卡罗模拟非常简单：

```r
m <- 100000
x1 <- rnorm(m); x2 <- rnorm(m) # 向量化操作效率更高
g_val <- abs(x1 - x2)
est <- mean(g_val)

cat("估计值:", est, "\n")
cat("理论值:", 2/sqrt(pi))
```

理论值 ≈ 1.128379。这种方法在处理多维复杂分布时优势巨大。

蒙特卡罗估计方法小结

我们可以通过生成随机样本来估计以下各类统计量：

- **期望值** $\mathrm{E}[X]$：直接计算样本均值 $\bar{x}$。
- **概率值** $\mathrm{P}(A)$：计算事件发生的频率（示性函数的均值）。
- **分布函数** $F(x)$：计算 $X_i \leq x$ 的经验比例（Glivenko-Cantelli 定理保证了收敛性）。
- **积分** $\int g(x)\,dx$：转化为期望形式进行估计。

蒙特卡罗积分：原理

目标：计算定积分 $\theta = \int_a^b g(x)\,dx$。

方法：引入在 (a, b) 上的均匀分布随机变量 $X \sim U(a, b)$，其概率密度函数为 $f(x) = \frac{1}{b-a}$。

我们可以将积分改写为期望的形式：

$$\theta = \int_a^b g(x)\,dx = (b-a)\int_a^b g(x)\,\underbrace{\frac{1}{b-a}}_{f(x)}\,dx = (b-a)\mathrm{E}[g(X)]$$

估计量：

$$\hat{\theta} = \frac{b-a}{n}\sum_{i=1}^{n} g(X_i), \quad X_i \sim U(a, b)$$

积分示例 1：区间 $[0, 1]$

计算 $\theta = \int_0^1 e^{-x} dx$。此时 $b - a = 1$。

```r
m <- 10000
x <- runif(m)              # 默认 min=0, max=1
theta_hat <- mean(exp(-x))

print(paste("MC估计:", round(theta_hat, 5)))
print(paste("真实值:", round(1 - exp(-1), 5)))
```

积分示例 2：一般区间 $[2, 4]$

计算 $\theta = \int_2^4 e^{-x} dx$。注意这里要乘以区间长度 $(4 - 2) = 2$。

```r
m <- 10000
x <- runif(m, min = 2, max = 4)
# 公式: (b-a) * mean(g(x))
theta_hat <- 2 * mean(exp(-x))

print(paste("MC估计:", round(theta_hat, 5)))
print(paste("真实值:", round(exp(-2) - exp(-4), 5)))
```

误差分析：中心极限定理 (CLT)

蒙特卡罗估计的误差服从什么分布?

Lindeberg-Lévy CLT

若 $X_1, \ldots, X_n$ 独立同分布，期望 μ，方差 $\sigma^2 < \infty$，则：

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

或者标准化形式：

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0, 1)$$

这告诉我们，蒙特卡罗估计的误差收敛速度为 $O(1/\sqrt{n})$。

CLT 验证：可视化

验证：当 n 较大时，标准化样本均值近似标准正态分布。以 $X \sim U(0,1)$ 为例，$\mu = 0.5, \sigma^2 = 1/12$。

```
verify_clt <- function(m, n) {
    # m: 模拟次数, n: 每次的样本量
    z_scores <- numeric(m)
    true_sd <- sqrt(1/12)
    for (i in 1:m) {
        x <- runif(n)
        # 构造标准化变量
        z_scores[i] <- (mean(x)-0.5)/(true_sd/sqrt(n))
    }
    return(z_scores)
}
data <- verify_clt(m=2000, n=100)
```

CLT 验证：绘图结果

接上页代码，绘制直方图与理论曲线对比：

```r
# 绘制频率直方图
hist(data, prob = TRUE, breaks = 40,
     main = "Standardized Mean vs N(0,1)",
     col = "lightblue", border = "white")

# 添加标准正态分布密度曲线
curve(dnorm(x), add = TRUE, col = "red", lwd = 2)
```

- 蓝色直方图代表模拟数据的分布。
- 红色曲线代表标准正态分布 $N(0,1)$。
- 两者高度重合说明了 CLT 的有效性。

练习题 1：定积分计算

利用蒙特卡罗方法计算如下定积分的近似值，并与真实值进行比较：

$$I = \int_1^5 \ln(x)\, dx$$

提示：

- 积分区间为 $[1, 5]$，长度为 4。
- 真实值计算公式：$\int \ln(x) dx = x \ln x - x$。
- 真实值结果约为 $5 \ln 5 - 5 - (1 \ln 1 - 1) \approx 4.047$。

练习题 2：多维随机变量概率估计

设 X, Y, Z 相互独立且均服从 $U(0,1)$ 分布。请编写 R 代码估计以下事件发生的概率：

$$P(X+Y+Z \leq 1)$$

思考：

- 理论上，这是三维单位立方体中被平面 $x+y+z=1$ 截出的四面体的体积。
- 理论值为 $\frac{1}{3!} = \frac{1}{6} \approx 0.16667$。

Thanks!