STAT3003: 统计算法基础

蒙特卡洛方法（3）

曾靖

中国科学技术大学管理学院

大纲

大纲

1. 方差缩减

本节内容来自于《统计计算使用 *R*》第 *5* 章，《统计方法》*3.3* 小节

1 蒙特卡罗方法的方差缩减

1.1 增大样本数

- 对于目标参数 $\theta = E(h(X))$，生成 m 个与 X 同分布的独立样本 $X_1, \ldots, X_m$，并计算 $h(X_i)$。Monte Carlo 估计量为

$$\boxed{\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} h(X_i).}$$

其方差为

$$\boxed{\mathrm{Var}(\hat{\theta}) = \frac{1}{m}\mathrm{Var}(h(X)).}$$

- 最简单的方差缩减方法即为增大试验次数 m。

$$\boxed{\frac{se(\hat{\theta}_{\text{new}})}{se(\hat{\theta}_{\text{old}})} = \frac{1}{100} \quad \Longrightarrow \quad \frac{m_{\text{new}}}{m_{\text{old}}} = 100^2 = 10{,}000.}$$

1.2 对偶变量法：一元情况

- 考虑关于一元随机变量 X 的参数 $\theta = E(h(X))$，这里 $X \in \mathbb{R}$。令 $x_1, \ldots, x_m$ 为与 X 同分布的 i.i.d. 样本，则 θ 的 Monte
- 考虑两个随机变量 $X_1, X_2 \sim X$，则平均值 $(h(X_1) + h(X_2))/2$ 的方差为

$$\begin{aligned}&\mathrm{Var}\left(\frac{h(X_1)+h(X_2)}{2}\right)\\ =&\frac{1}{4}(\mathrm{Var}(h(X_1)) + \mathrm{Var}(h(X_2)) + 2\mathrm{Cov}(h(X_1), h(X_2))).\end{aligned}$$

- 若 $h(X_1)$ 和 $h(X_2)$ 具有负相关性，则平均值 $\frac{h(X_1)+h(X_2)}{2}$ 可以被降低。

Theorem 1. 设 $g(x)$ 为单调函数，$U \sim \text{Unif}(0,1)$，则 $\text{Cov}(g(U), g(1-U)) \leq 0$.

- 记随机变量 X 的 CDF 为 $F(x)$, 生成 $U \sim \text{Unif}(0,1)$，则

$$F^{-1}(U), F^{-1}(1-U) \sim X$$

- 根据 Theorem 1，如果 h 为单调函数，则

$$h(F^{-1}(U)), h(F^{-1}(1-U)) \sim h(X)$$

且两者具有负相关性。

- 因此，只要生成一个样本 U，我们就可以构造出两个与 $h(X)$ 同分布且负相关的样本。

更一般地，我们有如下定理：

Theorem 2. 设 $f(x), g(x)$ 同单调增或同单调减, 则对于任何随机变量 X 我们有 $\mathrm{Cov}(f(X), g(X)) \geq 0$.

例 1.1 积分估计

Question: 利用一元对偶变量法估计标准正态分布函数

$$\boxed{\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.}$$

- 令 $\theta = E[xe^{-(xU)^2/2}]$，这里 $U \sim \text{Unif}(0,1)$。
- 如果 $x > 0, \Phi(x) = 0.5 + \theta/\sqrt{2\pi}$；如果 $x < 0$, $\Phi(x) = 1 - \Phi(-x)$。
- 因此只需用对偶变量法估计参数 θ。

```r
MC.Phi <- function(x, m = 10000, antithetic = TRUE){
  u <- runif(m/2)
  if (!antithetic) v <- runif(m/2)
  else v <- 1 - u
  u <- c(u, v)
  cdf <- numeric(length(x))
  for (i in 1:length(x)) {
    g <- x[i] * exp(-(u * x[i])^2 / 2)
    cdf[i] <- mean(g) / sqrt(2 * pi) + 0.5
    }
  cdf
}
```

```r
x <- seq(.1, 2.5, length=5)
Phi <- pnorm(x)
MC1 <- MC.Phi(x, antithetic = FALSE) # 非对偶法
MC2 <- MC.Phi(x) # 对偶法
print(round(rbind(x, MC1, MC2, Phi), 5))
```

```
##         [,1]    [,2]    [,3]    [,4]    [,5]
## x    0.10000 0.70000 1.30000 1.90000 2.50000
## MC1 0.53983 0.75795 0.90293 0.97116 0.99439
## MC2 0.53983 0.75804 0.90323 0.97131 0.99378
## Phi 0.53983 0.75804 0.90320 0.97128 0.99379
```

Question: 比较非对偶估计和对偶估计的方差。

```
m <- 1000
MC1 <- MC2 <- numeric(m)
x <- 1.95
for (i in 1:m) {
  MC1[i] <- MC.Phi(x, m = 1000, antithetic = FALSE)
  MC2[i] <- MC.Phi(x, m = 1000)
}
print(var(MC1))
```

```
## [1] 5.186911e-05
```

```
print(var(MC2))
```

```
## [1] 2.040312e-07
```

```
print(var(MC2)/var(MC1))
```

```
## [1] 0.003933578
```

1.2 对偶变量法：多元情况

考虑多元随机向量 $(X_1, \dots, X_n)$ 的参数 $\theta = E(h(X_1, \dots, X_n))$。

Definition 1：如果对于所有 $j = 1, \dots, n$，我们有 $x_j \le y_j$，则记 $(x_1, \dots, x_n) \le (y_1, \dots, y_n)$。

Definition 2：如果对于 $(x_1, \dots, x_n) \le (y_1, \dots, y_n)$ 我们有 $g(x_1, \dots, x_n) \le g(y_1, \dots, y_n)$，则称 n 元函数 $g = g(x_1, \dots, x_n)$ 称为单调增。

注：类似可定义单调减的 n 元函数。如果 n 元函数是单调增的或单调减的，则称该函数为单调的。

Theorem 3. 设 f, g 为关于 $(x_1, \dots, x_n)$ 的同单调递增或递减函数, 对任何多元随机变量 $(X_1, \dots, X_n)$ 且 X_j 之间相互独立，我们有 $\text{Cov}(f(X_1, \dots, X_n), g(X_1, \dots, X_n)) \geq 0$.

根据 Theorem 3，取 $f(x_1, \dots, x_n) = -g(1 - x_1, \dots, 1 - x_n)$, 我们有如下推论。

Corollary 1. 设 $g(x_1, \dots, x_n)$ 为关于 $(x_1, \dots, x_n)$ 的单调函数, $U_i \sim \text{Unif}(0, 1)$, $i = 1, \dots, n$ 且 U_i 间相互独立，则 $\text{Cov}(g(U_1, \dots, U_n), g(1 - U_1, \dots, 1 - U_n)) \leq 0$.

- 记每个随机变量 X_i 的 CDF 为 $F_i(x), i = 1, \dots, n$，如果 $h(x_1, \dots, x_n)$ 为单调函数，则 $h(F_1^{-1}(u_1), \dots, F_n^{-1}(u_n))$ 也是单调函数。
- 生成 $U_i \sim \text{Unif}(0, 1), i = 1, \dots, n$ 且 U_i 间相互独立。根据 Corollary 1，如果 $h(x_1, \dots, x_n)$ 为单调函数，随机变量

$$h(F_1^{-1}(U_1), \dots, F_n^{-1}(U_n)) \sim h(X_1, \dots, X_n),$$
$$h(F_1^{-1}(1 - U_1), \dots, F_n^{-1}(1 - U_n)) \sim h(X_1, \dots, X_n),$$

并且两者存在负相关关系。

- 对于目标参数 $\theta = E(h(X_1, \ldots, X_n))$，假设 h 是单调的，利用对偶变量法得到 Monte Carlo 估计
  - 独立生成 n 个均匀分布变量 $U_i \sim \text{Unif}(0,1), i = 1, \ldots, n$;
  - 构造随机变量 $Y_j = h(F_1^{-1}(U_1), \ldots, F_n^{-1}(U_n))$;
  - 构造对偶变量 $Y_j^+ = h(F_1^{-1}(1-U_1), \ldots F_n^{-1}(1-U_n))$，则 $Y_j, Y_j^+ \sim h(X_1, \ldots, X_n)$ 且 Y_j, Y_j^+ 负相关;
  - 重复上述步骤 $m/2$ 次，构造对偶 Monte Carlo 估计量

$$\hat{\theta} = \frac{1}{m} \sum_{j=1}^{m/2} (Y_j + Y_j^+).$$

结论：

- 对偶 Monte Carlo 估计量仍是 θ 的无偏估计
- 只要保证 $U_1, \ldots, U_n$ 相互独立，并且 h 单调，根据 Corollary 1，对偶 Monte Carlo 估计量的方差会得到缩减。
- 相比于仅单 Monte Carlo 估计 $\hat{\theta} = \frac{1}{m}\sum_{j=1}^{m} Y_j$，对偶 Monte Carlo 估计只需要重复 $m/2$ 次模拟实验，就可以得到更有效率（方差更小）的估计量。

特例：

- 如果 $X_i, i = 1, \ldots, n,$ 服从正态分布，我们不需要借助均匀分布变量，利用正态分布的对称性直接构造对偶变量。
- 假设 $X_i \sim N(\mu_i, \sigma^2)$，令 $Z_i = 2\mu_i - X_i$，则 $X_i \sim Z_i$。
- 如果 h 是单调函数。则可证明 $Y_j = h(X_1, \ldots, X_n)$ 与 $Y_j^+ = h(Z_1, \ldots, Z_n)$ 同分布于 $h(X_1, \ldots, X_n)$ 且负相关。
- 对偶 Monte Carlo 估计量为

$$\hat{\theta} = \frac{1}{m} \sum_{j=1}^{m/2} (Y_j + Y_j^+).$$

例 1.2 样本均值超过阈值的概率

Question: 设 $X_i \overset{iid}{\sim} N(0,1)$，$a>0$，$\bar{X}=n^{-1}\sum_{i=1}^n X_i$，估计 $\theta=P(\bar{X}>a)$。

- $h(x_1,\ldots,x_n)=I\left(n^{-1}\sum_{i=1}^n x_i>a\right)$ 对每个分量单调递增。
- 每次生成 $Z_{1j},\ldots,Z_{nj} \overset{iid}{\sim} N(0,1)$，并与其相反数配对。令

$$\bar{Z}_j=n^{-1}\sum_{i=1}^n Z_{ij},\quad Y_j=I(\bar{Z}_j>a),\quad Y_j^+=I(-\bar{Z}_j>a).$$

- 重复 $m/2$ 次，对偶估计量为

$$\hat{\theta}=\frac{1}{m}\sum_{j=1}^{m/2}(Y_j+Y_j^+).$$

1.3 控制变量法

- 考虑参数 $\theta = E(h(X))$，并设 $\mu = E(g(X))$ 已知。对任意常数 c，构造

$$Y = h(X) + c\{g(X) - \mu\}.$$

因为 $E(Y) = \theta$，所以 Y 是 θ 的无偏估计量；称 $g(X)$ 为 **控制变量**。

- 构造变量 Y 的方差为

$$\mathrm{Var}(Y) = \mathrm{Var}(h(X)) + c^2\mathrm{Var}(g(X)) + 2c\,\mathrm{Cov}(h(X), g(X)).$$

- 若 $X_1, \ldots, X_m$ 是 X 的独立样本，令 $Y_i = h(X_i) + c\{g(X_i) - \mu\}$，则

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} Y_i = \frac{1}{m}\sum_{i=1}^{m} \left[h(X_i) + c\{g(X_i) - \mu\}\right].$$

- 取

$$c^* = -\frac{\mathrm{Cov}(h(X), g(X))}{\mathrm{Var}(g(X))},$$

可使 Y 的方差最小:

$$\mathrm{Var}(Y) = \mathrm{Var}(h(X)) - \frac{\mathrm{Cov}^2(h(X), g(X))}{\mathrm{Var}(g(X))}.$$

- 相对于直接用 $h(X)$ 估计 θ，方差缩减比为

$$\begin{aligned}\frac{\mathrm{Var}(h(X)) - \mathrm{Var}(Y)}{\mathrm{Var}(h(X))} &= \frac{\mathrm{Cov}^2(h(X), g(X))}{\mathrm{Var}(h(X))\mathrm{Var}(g(X))} \\ &= \mathrm{Corr}^2(h(X), g(X)).\end{aligned}$$

- $h(X)$ 和 $g(X)$ 的相关性越强，控制变量带来的方差缩减通常越大。
- 使用时需知道 $\mu = E(g(X))$；求最优系数还需 $\text{Cov}(h(X), g(X))$ 和 $\text{Var}(g(X))$，若这些量不能直接计算，通常需要额外的 Monte Carlo 估计。
- 因此，应选择与 $h(X)$ 相关性较强，且 $E(g(X))$、$\text{Cov}(h(X), g(X))$ 和 $\text{Var}(g(X))$ 易于计算的控制变量 $g(X)$。

例 1.3

Question: 利用控制变量方法估计积分

$$\theta = \int_0^1 e^u du.$$

- 令 $U \sim \text{Unif}(0,1)$，则该问题等价于估计 $\theta = E(h(U)) = E(e^U)$，其中 $h(x) = e^x$。
- 取控制变量 $g(U) = U$，则 $\mu = E(g(U)) = 0.5$。并且

$$\begin{aligned} &\text{Var}(g(U)) = 1/12, \\ &\text{Cov}(h(U), g(U)) = 1 - (1/2)(e-1). \end{aligned}$$

- 因而最优系数为

$$c^* = -\frac{\text{Cov}(h(U), g(U))}{\text{Var}(g(U))} = -12 + 6(e-1).$$

- 构造变量为

$$Y = h(U) + c^*\{g(U) - \mu\} = e^U + c^*(U - 0.5).$$

```r
m <- 10000
c_opt <- -12 + 6 * (exp(1) - 1)
U <- runif(m)
T1 <- exp(U)
T2 <- exp(U) + c_opt * (U - 1/2) ## 控制变量方法
mean(T1)
```

```
## [1] 1.710791
```

```r
mean(T2)
```

```
## [1] 1.719018
```

```r
(var(T1) - var(T2)) / var(T1)
```

```
## [1] 0.9838001
```

1.4 重要抽样法

- 对于积分 $\theta = \int_A g(x)dx$ 的简单 Monte Carlo 估计如下：

$$\hat{\theta} = \frac{|A|}{m} \sum_{i=1}^{m} g(u_i), \quad u_i \sim \text{Unif}(A),$$

- 该方法会有如下三点问题：
  1. 这种方法无法应用到无限区间 A 上；
  2. 当存在很多 $g(u) \approx 0$ 的区域时，会造成很多样本的浪费；
  3. 当函数 $g(u)$ 在区间 A 上不平稳时，$g(u)$ 的方差会很大。

1.4 重要抽样法

- 假设随机变量 X 的密度函数为 $f(x)$，且支撑在 A 上。我们把问题转化为估计

$$\theta = \int_A g(x)dx = \int_A \frac{g(x)}{f(x)} f(x)dx = E[g(X)/f(X)].$$

其中，函数 $f(x)$ 被称为重要函数。

- 重要抽样估计量为

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} \frac{g(x_i)}{f(x_i)}, \quad x_1, \ldots, x_m \sim X.$$

这里，每个样本的权重为 $1/\{mf(x_i)\}$。

注：简单 *Monte Carlo* 估计中，从 $\text{Unif}(A)$ 抽样，每个样本的权重均为 $|A|/m$。

1.4 重要抽样法：直观解释

重要抽样法选取与 $|g(u)|$ 形状接近的密度 $f(u)$，以 $1/f(u)$ 作权重，这样

- 在 $|g(x)|$ 比较小的地方权重 $1/f(x)$ 会比较大；
- 在 $|g(x)|$ 比较大的地方权重 $1/f(x)$ 会比较小。

从而有如下效果：

1. 减少接近零的无效抽样：在 $|g(u)| \approx 0$ 的区域，$f(u)$ 也较小，抽到这些区域的概率降低。
2. 稳定加权后的取值：在 $|g(u)|$ 大的地方，权重 $1/f(u)$ 小；在 $|g(u)|$ 小的地方，权重大，使 $g(u)/f(u)$ 的波动减小，从而降低估计方差。

例 1.4

Question: 利用不同的重要函数估计积分

$$\int_0^1 \frac{e^{-x}}{1+x^2}dx.$$

考虑如下重要函数:

1. (简单抽要) $f_1(x) = 1, 0 < x < 1$;
2. (指数分布) $f_2(x) = e^{-x}, 0 < x < \infty$;
3. (Cauchy 分布) $f_3(x) = \frac{1}{\pi(1+x^2)}, -\infty < x < \infty$;
4. $f_4(x) = e^{-x}(1-e^{-1})^{-1}, 0 < x < 1$;
5. $f_5(x) = \frac{4}{\pi(1+x^2)}, 0 < x < 1$.

```r
m <- 10000
theta.hat <- se <- numeric(5)
g <- function(x) {
  exp(-x - log(1+x^2)) * (x > 0) * (x < 1)
}

x <- runif(m) #using f1
fg <- g(x)
theta.hat[1] <- mean(fg)
se[1] <- sd(fg) / sqrt(m)

x <- rexp(m, 1) #using f2
fg <- g(x) / exp(-x)
theta.hat[2] <- mean(fg)
se[2] <- sd(fg) / sqrt(m)
```

```r
x <- rcauchy(m) #using f3
i <- c(which(x > 1), which(x < 0))
x[i] <- 2 #to catch overflow errors in g(x)
fg <- g(x) / dcauchy(x)
theta.hat[3] <- mean(fg)
se[3] <- sd(fg) / sqrt(m)

u <- runif(m) #f4, inverse transform method
x <- - log(1 - u * (1 - exp(-1)))
fg <- g(x) / (exp(-x) / (1 - exp(-1)))
theta.hat[4] <- mean(fg)
se[4] <- sd(fg) / sqrt(m)

u <- runif(m) #f5, inverse transform method
x <- tan(pi * u / 4)
fg <- g(x) / (4 / ((1 + x^2) * pi))
theta.hat[5] <- mean(fg)
se[5] <- sd(fg) / sqrt(m)
```

```r
options(width = 60)
print(rbind(theta.hat, se), digits = 6)
```

```
##                  [,1]       [,2]       [,3]         [,4]
## theta.hat 0.52801489 0.52492070 0.51536584 0.523486060
## se        0.00245172 0.00418894 0.00944808 0.000967642
##                  [,5]
## theta.hat 0.52290135
## se        0.00141042
```

结论：

- 对于函数 $f_2(x)$ 和 $f_3(x)$，由于支撑集远比 $[0,1]$ 大，从而会导致大量样本 $g(x_i)/f(x_i)I([0,1])$ 为 0，进而导致方差 $\mathrm{Var}[g(X)/f(X)]$ 过大。
- 函数 $f_4(x)$ 和 $f_5(x)$ 则由于和函数 $g(x) = \frac{e^{-x}}{1+x^2}$ 分别共享了分子与分母，因此形状更为接近，从而缩减了方差 $\mathrm{Var}[g(X)/f(X)]$。

1.5 分层抽样法

我们总是可以把一个积分拆分成在不同不相交区间上的积分之和。以积分

$$\theta = \int_0^1 \frac{e^{-x}}{1+x^2} dx$$

为例，该积分可以写成如下形式

$$\begin{aligned}\theta =& \frac{1}{4}\left(4\int_0^{1/4} \frac{e^{-x}}{1+x^2} dx + 4\int_{1/4}^{1/2} \frac{e^{-x}}{1+x^2} dx\right.\\ &\left. +4\int_{1/2}^{3/4} \frac{e^{-x}}{1+x^2} dx + 4\int_{3/4}^{1} \frac{e^{-x}}{1+x^2} dx\right)\\ =& \frac{1}{4}(\theta_1+\theta_2+\theta_3+\theta_4).\end{aligned}$$

- 如果分层抽样来分别估计 $\theta_1, \theta_2, \theta_3, \theta_4$，则

$$\begin{aligned}
\theta_1 &= E(g(U^{(1)})), \quad U^{(1)} \sim \text{Unif}(0, 1/4), \\
\theta_2 &= E(g(U^{(2)})), \quad U^{(2)} \sim \text{Unif}(1/4, 1/2), \\
\theta_3 &= E(g(U^{(3)})), \quad U^{(3)} \sim \text{Unif}(1/2, 3/4), \\
\theta_4 &= E(g(U^{(4)})), \quad U^{(4)} \sim \text{Unif}(3/4, 1).
\end{aligned}$$

对于每层，生成 $m' = m/4$ 个样本：

$$\begin{aligned}
U_i^{(1)} \sim \text{Unif}(0, 1/4),\ U_i^{(2)} \sim \text{Unif}(1/4, 1/2), \\
U_i^{(3)} \sim \text{Unif}(1/2, 3/4),\ U_i^{(4)} \sim \text{Unif}(3/4, 1),
\end{aligned}$$

这里 $i = 1, \ldots, m'$.

- 相应的 Monte Carlo 估计量为

$$\hat{\theta}_1 = \frac{1}{m'}\sum_{i=1}^{m'} g(U_i^{(1)}), \quad \hat{\theta}_2 = \frac{1}{m'}\sum_{i=1}^{m'} g(U_i^{(2)}),$$

$$\hat{\theta}_3 = \frac{1}{m'}\sum_{i=1}^{m'} g(U_i^{(3)}), \quad \hat{\theta}_4 = \frac{1}{m'}\sum_{i=1}^{m'} g(U_i^{(4)}).$$

- 通过分层抽样得到的 θ 的估计量为 $\hat{\theta}' = \frac{1}{4}\sum_{k=1}^{4} \hat{\theta}_k$.

- 如果直接估计 θ，则 $\theta = E(g(U))$，这里 $U \sim \text{Unif}(0,1)$，$g(x) = e^{-x}/(1+x^2)$。则 Monte Carlo 估计量为

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} g(U_i), \quad U_i \sim \text{Unif}(0,1), \quad i = 1, \ldots, m.$$

- 比较两种 Monte Carlo 估计量的期望和方差
  - 期望:

$$E(\hat{\theta}) = \theta, \quad E(\hat{\theta}') = \frac{1}{4}\sum_{k=1}^{4} E(\hat{\theta}^{(k)}) = \frac{1}{4}\sum_{k=1}^{4} \theta_k = \theta.$$

  - 方差:

$$\mathrm{Var}(\hat{\theta}) = \frac{1}{m}\mathrm{Var}(g(U_i)), \quad \mathrm{Var}(\hat{\theta}') = \frac{1}{16}\sum_{k=1}^{4} \mathrm{Var}(\hat{\theta}_k),$$

由于对于每个 k,

$$\mathrm{Var}(\hat{\theta}_k) = \frac{4}{m}\mathrm{Var}(g(U_i^{(k)})),$$

则

$$\mathrm{Var}(\hat{\theta}') = \frac{1}{4m}\sum_{k=1}^{4} \mathrm{Var}(g(U_i^{(k)})).$$

- 由于我们进行了分层抽样，相比于 U_i，$U_i^{(k)}$ 是在更小的区间 (层) 上进行抽样，则 $\mathrm{Var}(g(U_i^{(k)})) \leq \mathrm{Var}(g(U_i))$，因此

$$\mathrm{Var}(\hat{\theta}') \leq \frac{1}{4m} \sum_{k=1}^{4} \mathrm{Var}(g(U_i)) = \mathrm{Var}(\hat{\theta}).$$

- 通过这个直观的例子，我们可以看出通过将积分分割成多个区间上的小积分，并进行分层抽样对每个小积分做 Monte Carlo 估计，可以降低 Monte Carlo 估计量的方差。

```r
n <- 500 # sample size
m <- 1000 # number of Monte Carlo
T2 <- numeric(4)
estimates <- matrix(0, m, 2)
g <- function(x) {exp(-x - log(1+x^2))}
for (i in 1:m) {
  estimates[i, 1] <- mean(g(runif(n)))
  T2[1] <- mean(g(runif(n/4, 0, .25)))
  T2[2] <- mean(g(runif(n/4, .25, .5)))
  T2[3] <- mean(g(runif(n/4, .5, .75)))
  T2[4] <- mean(g(runif(n/4, .75, 1)))
  estimates[i, 2] <- mean(T2)
}
apply(estimates, 2, mean)
```

```
## [1] 0.5249950 0.5247453
```

```r
apply(estimates, 2, var)
```

```
## [1] 1.249068e-04 8.032615e-06
```

解释

- 本质上分层抽样就是把参数 $\theta = E(g(X))$ 写成 $\theta = E\{E(g(X) \mid I)\}$，再进行两步 Monte Carlo 估计。
  - 第一步：在每一层 $(I = k)$ 中，给出 $E(g(X) \mid I = k)$ 的 Monte Carlo 估计；
  - 第二步：将 $E(g(X) \mid I)$ 记为 $h(I)$，则 $\theta = E(h(I))$，给出 $E(h(I))$ 的 Monte Carlo 估计。

分层抽样的方差

Theorem 4. 设 $P(I=k)=1/K$，$k=1,\dots,K$，考虑参数

$$\theta = E(g(X)), \qquad \theta_k = E(g(X)\mid I=k), \qquad \theta = \frac{1}{K}\sum_{k=1}^{K}\theta_k.$$

简单 Monte Carlo 抽样：$X_i \overset{\text{i.i.d.}}{\sim} X$，$i=1,\dots,m$。估计量为

$$\hat{\theta}^m = \frac{1}{m}\sum_{i=1}^{m} g(X_i).$$

分层抽样：$X_i^{(k)} \overset{\text{i.i.d.}}{\sim} X\mid I=k$，$i=1,\dots,m_k$。各层独立抽样，每层样本量 $m_k = m/K$。估计量为

$$\hat{\theta}_k = \frac{1}{m_k}\sum_{i=1}^{m_k} g(X_i^{(k)}), \qquad \hat{\theta}^s = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k,$$

若 $E[g(X)^2]<\infty$，则两个估计量均无偏，且

$$\mathrm{Var}(\hat{\theta}^s) \le \mathrm{Var}(\hat{\theta}^m).$$

- 回顾混合模型的知识，我们可以将均匀分布变量 $U \sim \text{Unif}(0,1)$ 写成一个混合模型, 即

$$f_U(u) = \frac{1}{4}\sum_{k=1}^{4} f_k(u), \qquad f_k(u) = 4\, I_{((k-1)/4,\, k/4)}(u).$$

其中 f_k 是第 k 层的均匀分布密度。记潜变量 I 为在 $\{1,2,3,4\}$ 上的均匀分布，指示 U 具体服从哪一层的分布，即如果 $I = k$，则 $U \sim U^{(k)}$。根据 U 的混合模型，我们知道

$$U \mid (I=1) \sim \text{Unif}(0,1/4), \quad U \mid (I=2) \sim \text{Unif}(1/4,1/2),$$
$$U \mid (I=3) \sim \text{Unif}(1/2,3/4), \quad U \mid (I=4) \sim \text{Unif}(3/4,1).$$

- 根据 Theorem 4 可知 $\hat{\theta}'$ 的方差比 $\hat{\theta}$ 小。