

STATYSTYCZNE MODELOWANIE DANYCH BIOLOGICZNYCH

WYKŁAD 10



Plan wykładu

1. Całkowanie przy użyciu metod Monte Carlo
2. Redukcja wariancji
3. Pakiet R

Całkowanie przy użyciu metod Monte Carlo

Niech $g(x)$ będzie funkcją, dla której chcemy wyznaczyć całkę $\int_a^b g(x)dx$ (zakładamy, że taka całka istnieje). Niech X będzie zmienną losową z funkcją gęstości $f(x)$, wtedy

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

Gdy dostępna jest próba losowa z rozkładu zmiennej losowej X to nieobciążonym estymatorem $E[g(X)]$ jest średnia z próby.

Rozważmy problem estymacji $\theta = \int_0^1 g(x)dx$. Jeśli $X_1, \dots, X_m$ jest próbą losową z rozkładu jednostajnego na odcinku $(0, 1)$ wtedy

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

zbiega do $E[g(X)] = \theta$ z prawdopodobieństwem 1 (na podstawie Mocnego Prawa Wielkich Liczb). Estymator Monte Carlo dla $\int_0^1 g(x)dx = \overline{g_m(X)}$.

Wyznaczyć estymator Monte Carlo całki

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

oraz porównać jego wartość z wartością rzeczywistą.

```
1 m<-10000
2 x<-runif(m)
3 theta.hat<-mean(exp(-x))
4 print(theta.hat)
5 [1] 0.6318515
6 print(1-exp(-1))
7 [1] 0.6321206
```

Całkowanie MC

Aby wyznaczyć całkę $\int_a^b g(t)dt$ należy dokonać zamiany zmiennych tak, aby granice całkowania były od 0 do 1. Transformacja liniowa ma postać

$$y = \frac{t - a}{b - a} \quad \text{i} \quad dy = \frac{1}{b - a} dt.$$

Po zamianie zmiennych otrzymujemy całkę

$$\int_a^b g(t)dt = \int_0^1 g(y(b - a) + a)(b - a)dy.$$

Alternatywnie możemy zastąpić rozkład jednostajny na odcinku $(0, 1)$ przez rozkład jednostajny na odcinku (a, b) i wyznaczyć całkę

$$\int_a^b g(t)dt = (b - a) \int_a^b g(t) \frac{1}{b - a} dt.$$

Wyznaczyć estymator Monte Carlo całki

$$\theta = \int_2^4 e^{-x} dx$$

oraz porównać jego wartość z wartością rzeczywistą.

```
1 m<-10000
2 x<-runif(m,min=2,max=4)
3 theta.hat<-mean(exp(-x))*2
4 print(theta.hat)
5 [1] 0.1173496
6 print(exp(-2)-exp(-4))
7 [1] 0.1170196
```


Ogólna metoda całkowania przy użyciu metody Monte Carlo

1. Wygenerować niezależne zmienne losowe $X_1, \dots, X_m$ z rozkładu jednostajnego na odcinku (a, b) .
2. Wyznaczyć $\overline{g(X)} = \frac{1}{m}g(X_i)$.
3. $\hat{\theta} = (b - a)\overline{g(X)}$.

Będziemy chcieli użyć metod Monte Carlo do estymacji dystrybuanty dla standardowego rozkładu normalnego

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

Warto zauważyć, że nie możemy bezpośrednio użyć algorytmu przedstawionego powyżej, ponieważ dolna granica całkowania wynosi $-\infty$. Jednakże możemy podzielić nasz problem na dwa przypadki: (i) $x < 0$ oraz (ii) $x \geq 0$ oraz użyć założenie o symetrii rozkładu normalnego do wyznaczenia całki dla $x < 0$.

Powyższy problem możemy rozwiązać poprzez generowanie zmiennych losowych z rozkładu jednostajnego na odcinku $(0, x)$, ale będzie to od nas wymagało zmiany parametru x dla każdej różnej wartości dystrybuanty. Łatwiej będzie użyć algorytmu, który zakłada losowanie liczb z rozkładu jednostajnego na odcinku $(0, 1)$ dla każdej wartości dystrybuanty. W tym celu musimy dokonać zamiany zmiennych

$$y = \frac{t}{x} \quad \text{i} \quad dt = x dy.$$

Wtedy mamy

$$\theta = \int_0^1 x e^{-(xy)^2/2} dy.$$

Estymatorem całki jest

$$\hat{\theta} = \overline{g_m(u)} = \frac{1}{m} \sum_{i=1}^m x e^{-(u_i x)^2/2}.$$

Estymatory dystrybuanty:

- $0.5 + \hat{\theta}/\sqrt{2\pi}$ dla $x > 0$,
- $\Phi(x) = 1 - \Phi(-x)$ dla $x < 0$.

```
1 x<-seq(.1,2.5,length=10)
2 m<-10000
3 u<-runif(m)
4 cdf=numeric(length(x))
5 for (i in 1:length(x)) {
6   g<-x[i]*exp(-(u*x[i])^2/2)
7   cdf[i]<-mean(g)/sqrt(2*pi)+.5
8 }
9
10 Phi<-pnorm(x)
11 print(round(rbind(x,cdf,Phi),3))
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
x	0.10	0.367	0.633	0.900	1.167	1.433	1.700	1.967	2.233	2.500
cdf	0.54	0.643	0.737	0.817	0.880	0.926	0.959	0.980	0.992	1.000
Phi	0.54	0.643	0.737	0.816	0.878	0.924	0.955	0.975	0.987	0.994

Niech $I(\cdot)$ będzie indykátorem oraz $Z \sim \mathcal{N}(0,1)$. Wtedy dla dowolnej stałej x mamy

$$E[I(Z \leq x)] = P(Z \leq x) = \Phi(x).$$

$\Phi(x)$ jest dystrybuantą standardowego rozkładu normalnego.

Wygeneruj próbę losową $z_1, \dots, z_m$ ze standardowego rozkładu normalnego. Wtedy średnia z próby

$$\hat{\Phi}(x) = \frac{1}{m} \sum_{i=1}^m I(z_i \leq x)$$

zbiega z prawdopodobieństwem 1 do swojej wartości oczekiwanej $E[I(Z \leq x)] = P(Z \leq x) = \Phi(x)$

Całkowanie MC

```
1 x<-seq(.1,2.5,length=10)
2 m<-10000
3 z<-rnorm(m)
4 dim(x)<-length(x)
5 p<-apply(x,MARGIN=1, FUN=function(x,z) {mean(z<x)}, z=z)
6
7 Phi<-pnorm(x)
8 print(round(rbind(x,p,Phi),3))
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
x	0.100	0.367	0.633	0.900	1.167	1.433	1.700	1.967	2.233	2.500
p	0.548	0.649	0.740	0.817	0.881	0.927	0.957	0.975	0.987	0.994
Phi	0.540	0.643	0.737	0.816	0.878	0.924	0.955	0.975	0.987	0.994

Wariancja $\hat{\theta}$ może być estymowana przy pomocy wzoru

$$\hat{\sigma}^2 = \frac{1}{m} \sum_{i=1}^m [g(x_i) - \overline{g(x)}]^2.$$

Wyznacz wariancję estymatora dystrybuanty standardowego rozkładu normalnego w punkcie 2 oraz wyznacz dla niego 95% przedział ufności.

```
1 x<-2
2 m<-10000
3 z<-rnorm(m)
4 dim(x)<-length(x)
5 g<-(z<rep(x,m))
6 v<-mean((g-mean(g))^2)/m
7 cdf<-mean(g)
8
9 c(cdf,v)
10 [1] 9.771000e-01 2.237559e-06
11
12 c(cdf-1.96*sqrt(v),cdf+1.96*sqrt(v))
13 [1] 0.9741681 0.9800319
14
15 pnorm(2)
16 [1] 0.9772499
```


Redukcja wariancji

Redukcja wariancji

Jeżeli $\hat{\theta}_1$ i $\hat{\theta}_2$ są estymatorami parametru θ oraz $Var(\hat{\theta}_2) < Var(\hat{\theta}_1)$ wtedy procent redukcji wariancji uzyskany poprzez użycie estymatora $\hat{\theta}_2$ zamiast $\hat{\theta}_1$ wyraża się wzorem

$$100 \left(\frac{Var(\hat{\theta}_1) - Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)} \right).$$

Redukcja wariancji

Rozważmy średnie dwóch zmiennych U_1 i U_2 pochodzących z tego samego rozkładu. Jeżeli zmienne losowe U_1 i U_2 są niezależne to

$$\text{Var}\left(\frac{U_1 + U_2}{2}\right) = \frac{1}{4} (\text{Var}(U_1) + \text{Var}(U_2)).$$

W ogólnej wersji mamy

$$\text{Var}\left(\frac{U_1 + U_2}{2}\right) = \frac{1}{4} (\text{Var}(U_1) + \text{Var}(U_2) + 2\text{Cov}(U_1, U_2)).$$

Z powyższego faktu wynika, że wariancja średniej $\text{Var}\left(\frac{U_1 + U_2}{2}\right)$ jest mniejsza jeżeli zmienne losowe U_1 i U_2 są ujemnie skorelowane niż jeżeli są niezależne.

Twierdzenie 1

Jeżeli zmienne losowe $X_1, \dots, X_n$ są niezależne oraz funkcje f i g są funkcjami rosnącymi wtedy

$$E[f(X)g(X)] \geq E[f(X)]E[g(X)]$$

Twierdzenie 2

Jeżeli $g = g(X_1, \dots, X_n)$ jest funkcją monotoniczną, wtedy zmienne

$$Y = g(F_X^{-1}(U_1), \dots, F_X^{-1}(U_n))$$

oraz

$$Y' = g(F_X^{-1}(1 - U_1), \dots, F_X^{-1}(1 - U_n))$$

są skorelowane ujemnie.

Redukcja wariancji

Metoda ta jest łatwa do zastosowania. Jeżeli potrzebujemy m replikacji Monte Carlo, należy utworzyć $\frac{m}{2}$ replikacji

$$Y_j = g(F_X^{-1}(U_1^{(j)}), \dots, F_X^{-1}(U_n^{(j)}))$$

oraz $\frac{m}{2}$ replikacji

$$Y'_j = g(F_X^{-1}(1 - U_1^{(j)}), \dots, F_X^{-1}(1 - U_n^{(j)}))$$

gdzie $U_i^{(j)}$ są niezależnymi zmiennymi losowymi z rozkładu jednostajnego na odcinku $(0, 1)$, $i = 1, \dots, n$ oraz $j = 1, \dots, m/2$. Wtedy nasz estymator ma postać

$$\hat{\theta} = \frac{1}{m} \{Y_1 + Y'_1 + Y_2 + Y'_2 + \dots + Y_{m/2} + Y'_{m/2}\} = \frac{2}{m} \sum_{j=1}^{m/2} \left(\frac{Y_j + Y'_j}{2} \right).$$

Redukcja wariancji

Chcemy ponownie wyestymować dystrybuantę standardowego rozkładu normalnego przy pomocy tradycyjnej metody oraz nowej metody przedstawionej na poprzednim slajdzie. Naszym celem jest parametr

$$\theta = E_U[xe^{-(xU)^2/2}],$$

gdzie U jest zmienną losową z rozkładu jednostajnego na odcinku $(0,1)$. Aby wykonać zadanie należy wyznaczyć próbę losową $u_1, \dots, u_{m/2}$ z rozkładu jednostajnego na odcinku $(0,1)$ i policzyć

$$Y_j = xe^{-(u_j x)^2/2}, \quad j = 1, \dots, m/2$$

oraz

$$Y'_j = xe^{-((1-u_j)x)^2/2}, \quad j = 1, \dots, m/2.$$

Średnia z próby wynosi

$$\hat{\theta} = \frac{2}{m} \sum_{j=1}^{m/2} \left(\frac{xe^{-(u_j x)^2/2} + xe^{-((1-u_j)x)^2/2}}{2} \right).$$

Redukcja wariancji

```
1 MC.Phi=function(x,R=10000,antithetic=TRUE) {  
2   u=runif(R/2)  
3   if (!antithetic) v=runif(R/2) else  
4     v=1-u  
5   u=c(u,v)  
6   cdf=numeric(length(x))  
7   for (i in 1:length(x)) {  
8     g=x[i]*exp(-(u*x[i])^2/2)  
9     cdf[i]=mean(g)/sqrt(2*pi)+0.5  
10  }  
11  cdf  
12 }
```

Redukcja wariancji

```
1 m<-1000
2 MC1<-MC2<-numeric(m)
3 x<-1.95
4 for (i in 1:m) {
5   MC1[i]=MC.Phi(x,R=1000,anti=FALSE)
6   MC2[i]=MC.Phi(x,R=1000)
7 }
8
9 print(sd(MC1))
10 [1] 0.006874616
11 print(sd(MC2))
12 [1] 0.0004392972
13 print((var(MC1)-var(MC2))/var(MC1))
14 [1] 0.9959166
```


Redukcja wariancji

```
1 x=seq(.1,2.5,length=5)
2 Phi=pnorm(x)
3 set.seed(123)
4 MC1=MC.Phi(x,anti=FALSE)
5 set.seed(123)
6 MC2=MC.Phi(x)
7 print(round(rbind(x,MC1,MC2,Phi),5))
```

Redukcja wariancji

	[,1]	[,2]	[,3]	[,4]	[,5]
x	0.10000	0.70000	1.30000	1.90000	2.50000
MC1	0.53983	0.75825	0.90418	0.97311	0.99594
MC2	0.53983	0.75805	0.90325	0.97132	0.99370
Phi	0.53983	0.75804	0.90320	0.97128	0.99379

Redukcja wariancji

```
1 m<-10000
2 a<- -12+6*(exp(1)-1)
3 U<-runif(m)
4 T1<-exp(U)
5 T2<-exp(U)+a*(U-1/2)
6
7 mean(T1)
8 [1] 1.717834
9 mean(T2)
10 [1] 1.718229
11 (var(T1)-var(T2))/var(T1)
12 [1] 0.9838606
```

Wykład opracowany na podstawie książki:

- Maria L. Rizzo 'Statistical Computing with R'