서울대학교 통계학과 **이재용 교수**

07강

베이지안 통계학

몬테카를로 방법

한국방송통신대학교
Korea National Open University

목차

01 난수발생

역함수방법

F는 역함수가 존재하는 누적분포함수라 하자.

$U \ \sim \ U(0,1)$이라면, $F^{-1}(U) \ \sim \ F$이다.

* 예) 지수분포

$$U \sim Exp(1) \Leftrightarrow X \overset{d}{=} -\log(1-U), U \sim U(0,1)$$

$$\Leftrightarrow X \overset{d}{=} -\log U, U \sim U(0,1)$$

합격불합격방법(acceptance-rejection method)

목적과 상황

* 밀도함수 f에서 난수를 발생시키고자 한다.
f에서 직접 난수를 발생시키는 것은 어렵지만,
아래의 조건을 만족하는 밀도함수 g에서 난수를 발생시키는 것은 쉽다고 하자.

$$f(x) \le Mg(x),\ x : f(x) > 0.$$

여기서 $M > 0$은 상수이다.

알고리듬

1. $X \sim g$와 $U \sim U(0, 1)$를 발생시킨다.
2. $U \le f(X)/Mg(X)$이면, $Y = X$이라 하고, 그렇지 않으면 1로 돌아간다.

예. 삼각분포에서 확률변수의

▫ 다음의 확률밀도함수에서 확률변수를 생성하는 문제를 고려해 보자.

$$f(x) = \begin{cases} x, & 0 \le x \le 1, \\ 2 - x, & 1 \le x \le 2 \end{cases}$$

$$g(x) = 1,\ 0 \le x \le 2$$

라 정의하면

$$\sup_{x \in [0,2]} \frac{f(x)}{g(x)} = 1$$

임을 알 수 있다.

▫ 알고리듬

- ⋆ 단계 1. $X \sim U(0,2), U \sim U(0,1)$을 생성한다.
- ⋆ 단계 2. 합격비율 $f(X)$를 계산한다.
- ⋆ 단계 3. $U \le f(X)$ 이면, $Y = X$ 를 반환하고, 그렇지 않으면 단계 1 로 돌아간다.

02 몬테카를로 방법

목적

$\boldsymbol{X} \sim \boldsymbol{g}(\boldsymbol{x})$**일 때, 적분값**

$$\boldsymbol{B} = \mathbb{E}[\boldsymbol{f}(\boldsymbol{X})] = \int \boldsymbol{f}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})\boldsymbol{dx},$$

을 구하고자 한다.

알고리듬

$X_1, \ldots, X_n \overset{iid}{\sim} g$ **이라 하자. 다음과 같이**

$$\hat{B} \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

를 계산하고, $\hat{B}$**를** B**를 추정하는데 사용한다.**

추정오차

$\hat{\boldsymbol{B}}$**의 추정오차** $= \widehat{\boldsymbol{SE}}(\hat{\boldsymbol{B}}) = \sqrt{\frac{\mathrm{v}}{\boldsymbol{n}}}.$

여기서,

$$\mathrm{v} = \frac{n}{n-1}\sum_{i=1}^{n}\left(f(X_i) - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\right)^2$$

이다.

예. π 의 추정

결과

$U_1, U_2 \overset{i.i.d}{\sim} U(0,1)$ 이면,

$$\mathbb{P}(u_1^2 + u_2^2 < 1) = \frac{\pi}{4}$$

이다.

알고리듬

$(U_{i1}, U_{i2}) \overset{i.i.d}{\sim} U(0,1) \times U(0,1), i = 1,\cdots,n,$ 을 추출한다.
이를 이용해 $P(u_1^2 + u_2^2 < 1)$를 다음의 추정량으로

$$\hat{P}_n = \frac{1}{n}\#\left(u_{i1}^2 + u_{i2}^2 < 1\right).$$

π 는

$$\hat{\pi} = 4\hat{P}_n$$

으로 추정한다. π의 95% 신뢰구간은

$$4\hat{P}_n \pm 8.84\sqrt{\frac{\hat{P}_n(1-\hat{P}_n)}{n}}$$ 이다.

예. 정규모형과 코시 사전분포

▫ 문제

$X|\theta \sim N(\theta, 1)$**이고,**

사전분포가 $\theta \sim Cauchy(0, 1)$**일 때,**

사후분포의 평균을 구해보자.

예. 정규모형과 코시 사전분포.

▫ 풀이

사후분포의 밀도함수는

$$\pi(\theta|x) \propto \frac{1}{1+\theta^2} e^{-\frac{1}{2}(x-\theta)^2}$$

이다. 따라서, 사후분포의 평균은

$$E(\theta|x)=\frac{\int \frac{1}{1+\theta^2} e^{-\frac{1}{2}(\theta-x)^2} d\theta}{\int \frac{1}{1+\theta^2} e^{-\frac{1}{2}(\theta-x)^2} d\theta}$$

이다. 위 적분의 결과는 수식으로 주어지지 않는다.

* 알고리듬

$\theta_1, \theta_2, \cdots, \theta_m \overset{i.i.d}{\sim} N(x,1)$를 추출한다.

그리고 분자와 분모의 적분을 각각 추정해서 사후분포의 평균을 구한다. 추정량은

$$\hat{\theta}^m \approx \frac{\sum_{i=1}^m \frac{\theta_i}{1+\theta_i^2}}{\sum_{i=1}^m \frac{1}{1+\theta_i^2}}$$

와 같이 주어진다.

중요도 추출(importance sampling)

▫ 문제

$X \sim g(x)$이고,

$$B = \int f(x)g(x)dx$$

를 구하려고 한다.
그런데, $g(x)$에서 표본을 추출하는 것은 어렵고,

$\pi(x)$ 에서 표본을 추출하는 것은 쉽다고 하자.

▫ 알고리듬

1. $X_1, \ldots, X_n \overset{iid}{\sim} \pi(x)$ 를 추출한다.
2. 가중치 $w_i = \frac{g(x_i)}{\pi(x_i)}$, i=1, . . . , n,를 계산한다.
3. B를 다음의 두 값으로 추정한다.

$$\hat{B}_1 = \frac{1}{n}\sum_{i=1}^{n} w_i f(x_i)$$

$$\hat{B}_2 = \frac{\sum_{i=1}^{n} w_i f(x_i)}{\sum_{i=1}^{n} w_i}$$

중요도 추출(importance sampling)

- $\hat{B}_1$은 B의 불편추정량이지만, $\hat{B}_2$은 불편추정량은 아니다. 하지만 편이는 크지 않다.
- $\hat{B}_1$를 구하기 위해서는 가중치를 정확히 알아야 한다. 반면에 $\hat{B}_2$를 구하기 위해서는 가중치가 모르는 상수의 곱으로 나타나도 상관없다.

  즉 $\hat{B}_1$을 구하기 위해서는 $g(x)$를 정확하게 알아야 하지만, $\hat{B}_2$를 구하기 위해서는

  $$g(x) = \text{모르는 상수} \times \text{아는 함수형태}$$

  이어도 상관없다.
- *중요도추출은* $g(x)$ 에서 표본을 추출하기 어려울 때 사용할 수 있다.

예. 삼각분포의 평균

$$f(x) = \begin{cases} x, & 0 \le x \le 1, \\ 2 - x, & 1 \le x \le 2 \end{cases}$$

라 할 때,

$$\mathrm{B} = \int_0^2 x f(x) dx$$

를 중요도추출을 구하는 방법에 대해 알아보자.

예. 삼각분포의 평균

▫ 풀이

$\pi(x) = \frac{1}{2} I(0 \le x \le 2)$를

$U(0,2)$의 밀도함수로 정의하자

$$I = \int_0^2 xf(x)dx = \int_0^2 x\frac{f(x)}{\pi(x)}\pi(x)dx$$

임을 이용하여 다음의 알고리듬을

구성할 수 있다.

* **알고리듬**

단계1. $X_1, \cdots, X_n \overset{i.i.d.}{\sim} U(0,2)$를 **생성한다**.

단계2. **가중치** $W_i = \frac{f(X_i)}{\pi(X_i)}$**=2** $f(X_i)$**, i=1,2,...,n,를 계산한다.**

단계3. **다음 두 가지 중 하나의 식을 이용해 추정값을 계산한다.**

단계1..1 $\hat{I}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i w_i$ **를 계산한다**.

단계2..2 $\hat{I}_2 = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$ **를 계산한다**.

03 실습

합격불합격 방법을 이용한 π 의 계산

알고리듬

$(U_{i1}, U_{i2}) \overset{i.i.d}{\sim} U(0,1), i = 1,\cdots,n,$ 를 추출한다.

이를 이용해 $P(u_1^2 + u_2^2 < 1)$를

다음의 추정량으로 추정한다.

$$\hat{p}_n = \frac{1}{n}\#\left(u_{i1}^2 + u_{i2}^2 < 1\right).$$

π는

$$\hat{\pi} = 4\hat{P}_n$$

으로 추정한다. π의 95% 신뢰구간은

$$4\hat{p}_n \pm 8.84\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$$

이다.

* 코드

```
n = 1000
u1 = runif(n)
u2 = runif(n)
phat = mean( u1^2 + u2^2 < 1)
pihat = 4*phat; pihat
cihw = 8.84*sqrt(phat*(1-phat)/n)
ci = c(pihat - cihw, pihat + cihw);ci
```

참고문헌

- Hoff, Peter D. A first course in Bayesian statistical methods의 4장. Springer Science & Business Media, 2009.

수고하셨습니다.

감사합니다.