STAT3006:TUTORIAL3

Sampling methods:

1. Inverse Method.
2. Accept-Reject method.

RANDOM SAMPLES

- In statistics, we use data to extract information, but how is the data produced?

- The most basic problem: how to get a sequence of (uniform) random numbers?
  - Pseudorandom number generator.
  - Why called pseudo? If you know the seed (X_0), you know any following number in the sequence.
  - Examples:
    - linear congruential generator (often has poor performance): $X_{n+1} = (aX_n + c) \ mod \ m$.
    - Mersenne Twister (most used).
  - Why called random? The sequence $\{X_0, X_1, \dots, X_n\}$ can pass numerous statistical tests for randomness (e.g. Diehard tests, Kolmogorov-Smirnov test).
  - Throughout the tutorial 3, we assume we can sample from uniform distribution on [0,1]

INVERSE METHOD

- We want to sample from $F(x)$, where $F(x)$ is a distribution function.
- Define the generalized inverse function of F, $F^{-}(y)$, as $\inf\{x: F(x) \geq y\}$.
- Inverse method:
  - Draw a sample U from unif[0,1];
  - Let X be $F^{-}(U)$.
- Some requirements:
  - F is a univariate function
  - F^{-} has a closed form or F^{-} can be easily calculated numerically.

INVERSE METHOD

- A discrete case: drawing samples from Binomial(n, p)
- $p(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x}$
- $F(x|n,p) = \sum_{t \leq x} \binom{n}{t} p^t (1-p)^{n-t}$
- Inverse method:
  - U from unif[0,1]
  - If $\mathrm{F}(\mathrm{x}|\mathrm{n},\mathrm{p}) < U \leq F(x+1|n,p)$, let $\mathrm{X} = \mathrm{x} + 1$.

INVERSE METHOD

```
#samples from binomial distribution
m <- 2000 # sample number
n <- 8 # trial number
p <- 0.4 # success probability
x <- 0:n
prob_mass <- factorial(n) / (factorial(n-x) * factorial(x)) * p^x * (1-p)^(n-x) #probability mass
prob_cumsum <- cumsum(prob_mass) #cumulative sum of porbability masses
u_vec <- runif(m) #uniform samples
x_vec <- NULL #binomial samples

for(i in 1:m){
        x_vec <- c(x_vec, min( which(prob_cumsum >= u_vec[i]) ) - 1)
}

hist(x_vec, breaks = seq(-0.5, 8.5, by = 1), freq = FALSE) #histogram
```

INVERSE METHOD

- A continuous case: drawing samples from a truncated normal distribution N(0,1)I(x>2.5)
- An intuitive approach is fist drawing samples from N(0,1) and then only selecting samples greater than 2.5.
- Is it valid? Yes, but not effective.
- Inverse method (what is F^{-}):
  - U from unif[0,1]
  - Let $X = F^{-}(U)$
  - $F^{-}(u) = \mathbf{\Phi}^{-1}(\left(1 - \mathbf{\Phi}(2.5)\right)U + \mathbf{\Phi}(2.5))$

INVERSE METHOD

- #samples from truncated normal distribution
- m <- 2000 #sample number
- trun_point <- 2.5
- u_vec <- runif(m) #uniform samples
- const <- pnorm(2.5)
- x_vec <- qnorm((1-const)*u_vec + const)

- hist(x_vec, breaks = 20, freq = FALSE) #histrogram

ACCEPT-REJECT METHOD

- When F and the inverse of F is not tractable.
- We are interested in sampling from a pdf $f(x)$.
- There exists a pdf $g(x), s.t. f(x) < Mg(x)$.
- Sampling from $g(x)$ is easy.
- Then we can apply accept-reject method:
  - Step1: draw $Y \sim g(x), U \sim unif[0,1]$
  - Step2: if $U < \frac{f(Y)}{Mg(Y)}$, let $X = Y$.
  - The acceptance rate is $1/M$. The smaller the M is, the more effective the accept-reject method is.

ACCEPT-REJECT METHOD

- Example: we want to sample from a truncated normal distribution $N(0,1)I(x > 2.5)$.
- How to determine the $g(x)$?
- What is $f(x)/Mg(x)$?
- $g(x)$ is a shifted exponential pdf.

ACCEPT-REJECT METHOD

- #samples from truncated normal distribution
- m <- 2000
- i <- 0
- M <- exp(-2.5^2/2)/(2.5*sqrt(2*pi)*(1-pnorm(2.5)))
- ratio <- function(y){sqrt(2*pi)*exp(-y^2/2 + 2.5*(y-2.5) + 2.5^2/2)}
- x_vec <- NULL
- while(i <= m){
- y <- rexp(1, 2.5) + 2.5
- u <- runif(1)
- if(u <= ratio(y)){
- x_vec <- c(x_vec, y)
- i <- i+1
- }
- }
- hist(x_vec, breaks = 20, freq = FALSE) #histrogram