Computational Statistics Homework 1

吴嘉骜 21307130203

2023 年 9 月 24 日

1 ex 3.5

A discrete random variable X has probability mass function

x	0	1	2	3	4
$p(x)$	0.1	0.2	0.2	0.2	0.3

Use the inverse transform method to generate a random sample of size 1000 from the distribution of X. Construct a relative frequency table and compare the empirical with the theoretical probabilities. Repeat using the R `sample` function.

Solution:

First, we use the inverse transform method to generate the required random sample. We first calculate the cumulative distribution function of X:

$$F(x) = \begin{cases} 0, & x < 0 \\ 0.1, & 0 \le x < 1 \\ 0.3, & 1 \le x < 2 \\ 0.5, & 2 \le x < 3 \\ 0.7, & 3 \le x < 4 \\ 1, & x \ge 4 \end{cases}$$

Then we generate u from $Unif(0,1)$ 1000 times, and use the inverse of F to transform the sample to the required distribution. If $u \in [0, 0.1]$, then $x = 0$; if $u \in (0.1, 0.3]$, then $x = 1$; if $u \in (0.3, 0.5]$, then $x = 2$; if $u \in (0.5, 0.7]$, then $x = 3$; if $u \in (0.7, 1]$, then $x = 4$. The R code and result are as follows:

```
> # Inverse transform method
> n <- 1000
> p <- c(0.1, 0.2, 0.2, 0.2, 0.3)
> p_cdf <- cumsum(p)
> x <- rep(n,0)
> for(i in 1:n){
+   u <- runif(1)
+   x[i] <- sum(u>p_cdf)
+ }
> rbind(table(x)/n,p)
```

```
        0     1     2     3     4
     0.096 0.198 0.214 0.194 0.298
p    0.100 0.200 0.200 0.200 0.300
```

We can see that the empirical probabilities are very close to the theoretical probabilities.

Then we use the R `sample` function to generate the required random sample. The R code and result are as follows:

```
> # Sample function
> n <- 1000
> p <- c(0.1, 0.2, 0.2, 0.2, 0.3)
> y <- sample(0:4, size=n, prob=p, replace=T)
> rbind(table(y) / n, p)
        0     1     2     3     4
     0.098 0.205 0.197 0.193 0.307
p    0.100 0.200 0.200 0.200 0.300
```

2 ex 3.6

Prove that the accepted variates generated by the acceptance-rejection sampling algorithm are a random sample from the target density f_X.

Proof:

Let X be the random variable generated by the Acceptance-Rejection sampling algorithm, and Y be the auxiliary random variable from density g_Y. We assume that there exits a constant c such that $f_X(t) \leq cg_Y(t)$ for all t where $f_X(t) > 0$.

Let U be a random variable from $Unif(0,1)$, and A be the set of accepted values y.

Given $Y = y$, the probability that y is accepted is

$$P(A|Y=y) = P(U < \frac{f_X(y)}{cg_Y(y)}|Y=y) = \frac{f_X(y)}{cg_Y(y)}.$$

Then we have

$$P(A) = \int_{-\infty}^{\infty} P(A|Y=y)g_Y(y)dy = \int_{-\infty}^{\infty} \frac{f_X(y)}{cg_Y(y)}g_Y(y)dy = \frac{1}{c}.$$

By Bayes' theorem, we have

$$g_{Y|A}(y|A) = \frac{P(A|Y=y)g_Y(y)}{P(A)} = \frac{f_X(y)}{cg_Y(y)} \cdot g_Y(y) \cdot \frac{1}{1/c} = f_X(y).$$

Therefore, the accepted variates generated by the Acceptance-Rejection sampling algorithm are a random sample from the target density f_X.

3 ex 3.9

The rescaled Epanechnikov kernel is a symmetric density function

$$f_e(x) = \frac{3}{4}(1-x^2), \quad |x| \leq 1.$$

Devroye and Györfi give the following algorithm for simulation from this distribution. Generate iid $U_1, U_2, U_3 \sim$ Uniform$(-1, 1)$. If $|U_3| \geq |U_2|$ and $|U_3| \geq |U_1|$, deliver U_2; otherwise deliver U_3. Write a function to generate random variates from f_e, and construct the histogram density estimate of a large simulated random sample.

Solution:

The R code is as follows:

```
# Rescaled Epanechnikov kernel
rREpan <- function(n){
    u1 <- runif(n, -1, 1)
    u2 <- runif(n, -1, 1)
    u3 <- runif(n, -1, 1)
    x <- rep(n, 0)
    for(i in 1:n){
    if(abs(u3[i]) >= abs(u2[i]) && abs(u3[i]) >= abs(u1[i]))
        x[i] <- u2[i]
    else
        x[i] <- u3[i]
    }
    return(x)
}

x <- rREpan(10000)
hist(x,prob=T)
y <- seq(-1, 1, 0.001)
fe <- 0.75 * (1 - y^2)
lines(y,fe)
```

The histogram density estimate of a 10000-size random sample together with the theoretical density is as follows:

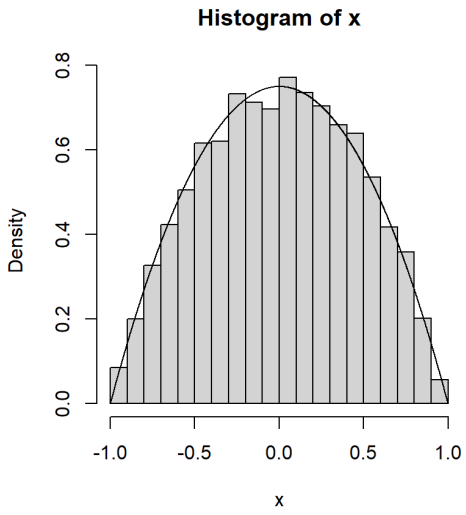

Figure 1: Histogram of a Rescaled Epanechnikov kernel sample

4 ex 3.10

Prove that the algorithm given in Exercise 3.9 generates variates from the density f_e.

Proof:

The algorithm equvalently generates iid $V_1, V_2, V_3 \sim$ Uniform(0, 1), and choose the first or second order statistic (not the largest) V with equal probability $\frac{1}{2}$, then let $X = \pm V$ with probability $\frac{1}{2}$ respectively. Notice that V is a mixture of $V_{(1)}$ and $V_{(2)}$, where $V_{(1)}$ and $V_{(2)}$ are the first and second order statistics of V_1, V_2, V_3. So the cumulative distribution function of V is $F_V(v) = \frac{1}{2}F_{V_{(1)}}(v) + \frac{1}{2}F_{V_{(2)}}(v)$.

As is known, the pdf of $V_{(k)}$ is $f_{V_{(k)}}(v) = k\binom{n}{k}v^{k-1}(1-v)^{n-k}$, $v \in (0, 1)$.

So the pdf of $V_{(1)}$ is $f_{V_{(1)}}(v) = 3(1-v)^2$, and the pdf of $V_{(2)}$ is $f_{V_{(2)}}(v) = 6v(1-v)$, and the pdf of V is $f_V(v) = \frac{1}{2}f_{V_{(1)}}(v) + \frac{1}{2}f_{V_{(2)}}(v) = \frac{3}{2}(1-v^2)$.

Therefore, the pdf of X is $f_X(x) = \frac{1}{2}f_V(x) = \frac{3}{4}(1-x^2)$, which is the density f_e.

5 ex 3.12

Simulate a continuous Exponential-Gamma mixture. Suppose that the rate parameter Λ has Gamma(r, β) distribution and Y has Exp(Λ) distribution. That is, $(Y|\Lambda = \lambda) \sim f_Y(y|\lambda) = \lambda e^{-\lambda y}$. Generate 1000 random observations from this mixture with $r = 4$ and $\beta = 2$.

Solution:

The R code is as follows:

```
# Exponential-Gamma mixture
n <- 1000
r <- 4
beta <- 2
lambda <- rgamma(n,r,beta)
x <- rexp(n,lambda)
```

6 ex 3.13

It can be shown that the mixture in Exercise 3.12 has a Pareto distribution with cdf

$$F(y) = 1 - \left(\frac{\beta}{\beta + y}\right)^r, \quad y \geq 0.$$

(This is an alternative parameterization of the Pareto cdf given in Exercise 3.3.) Generate 1000 random observations from the mixture with $r = 4$ and $\beta = 2$. Compare the empirical and theoretical (Pareto) distributions by graphing the density histogram of the sample and superimposing the Pareto density curve.

Solution:

The probability density function of Pareto distribution is

$$f(y) = \frac{r\beta^r}{(y+\beta)^{r+1}}, \quad y \geq 0.$$

The R code is as follows:

```r
# Exponential-Gamma mixture and Pareto dist
n <- 1000
r <- 4
beta <- 2
lambda <- rgamma(n,r,beta)
x <- rexp(n,lambda)
hist(x,prob=T)
y <- sort(x)
f_pareto=r * beta^r / (beta + y)^(r + 1)
lines(y,f_pareto)
```

The histogram density estimate of a 1000-size random sample together with the theoretical density is as follows:

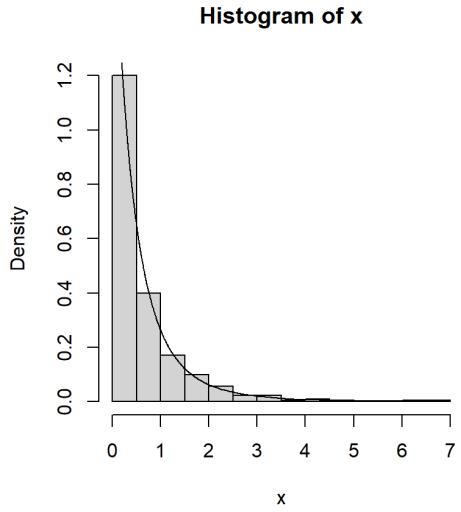

Figure 2: Histogram of an Exponential-Gamma mixture sample

From the figure we can see that the histogram density estimate is close to the theoretical density.

7 ex 3.14

Generate 200 random observations from the 3-dimensional multivariate normal distribution having mean vector $\mu = (0, 1, 2)$ and covariance matrix

$$\Sigma = \begin{bmatrix} 1.0 & -0.5 & 0.5 \\ -0.5 & 1.0 & -0.5 \\ 0.5 & -0.5 & 1.0 \end{bmatrix}.$$

using the Choleski factorization method. Use the R `pairs` plot to graph an array of scatter plots for each pair of variables. For each pair of variables, (visually) check that the location and correlation approximately agree with the theoretical parameters of the corresponding bivariate normal distribution.

Solution:

The R code and result are as follows:

```
> rmvn_Chol <- function(n,mu,sigma){
    +   d <- length(mu)
    +   CT <- chol(sigma)
    +   Z <- matrix(rnorm(n*d),n,d)
    +   X <- Z%*%CT + matrix(mu,n,d,byrow=T)
    +   return(X)
    + }
> sigma <- matrix(c(1, -0.5, 0.5, -0.5, 1, -0.5, 0.5, -0.5, 1), 3, 3)
> mu <- c(0, 1, 2)
> x <- rmvn_Chol(200, mu, sigma)
> colMeans(x)
[1] 0.03386865 0.96643718 2.07497880
> cor(x)
           [,1]       [,2]       [,3]
[1,]  1.0000000 -0.5018619  0.4559275
[2,] -0.5018619  1.0000000 -0.5306230
[3,]  0.4559275 -0.5306230  1.0000000
> pairs(x)
```

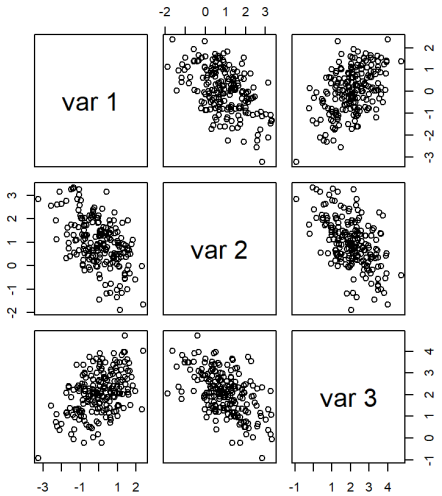

Figure 3: Pairs plot of a 3-dimensional multivariate normal distribution

From the sample mean and sample correlation we can see that the generated sample is close to the specified parameters.

From the `pairs` plot we can see that the centers of the three variables are approximately $(0, 1, 2)$, and the correlations approximately agree with Σ.