《统计算法基础》第二次作业

姓名：王凯栋　　学号：PB20071441　　日期：2023/3/24

目录

一. 实验目的

- 了解 Copula Model 生成随机数的方式
- 了解随机模拟在数值计算中的具体方法
- 结合程序理解统计算法 MCMC 的使用方式

二. 实验过程

- 完成 PPT 关于 Copula Model 若干问题
- 完成统计计算若干题目 (推导及代码)
- 完成统计计算使用 R 若干代码问题

三. 实验内容

Copula Model

Question 1

对于多元 t 分布，$X \sim T_v(\mu,\Sigma;p)$，说明其 Copula 完全由矩阵 $M = diag(\Sigma)^{-1/2}\Sigma diag(\Sigma)^{-1/2}$ 确定。

解：

首先，对于多元 t 分布 $X \sim T_v(\mu,\Sigma;p)$，它的密度函数可以表示为：

$$f(x) = \frac{\Gamma(\frac{v+p}{2})}{\Gamma(\frac{p}{2})\sqrt{det(\pi v\Sigma)}}(1+1/v(x-\mu)^T\Sigma^{-1}(x-\mu))^{-\frac{v+p}{2}}$$

设多元随机变量 X 边缘分布为 $F_i(x)$，联合密度函数为 $F(x_1,x_2,\cdots,x_p)$, 则 X 的 Copula 函数为：

$$C(u_1,u_2,\cdots,u_p) = F(F_1^{-1}(u_1),F_2^{-1}(u_2),\cdots,F_p^{-1}(u_p))$$

其中 F_i^{-1} 为分量 X_i 的逆分布函数，u_i 为第 i 个分量边缘分布函数的取值，所以只要求出 F_i^{-1}，我们便可以得到 $Copula$ 的具体形状。

$F_i \sim \sigma_{ii}^{1/2}t_v$,

由于 σt 的概率密度函数查得为：

$$f(t) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(v/2)\sqrt{\pi v\sigma_{ii}}(1+\frac{(t-\mu)^2}{v\sigma_{ii}})^{\frac{v+1}{2}}}$$

所以 F_i 对应的分布函数为

$$F_i(x) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(v/2)\sqrt{\pi v\sigma_{ii}}}\int_{-\infty}^{x}(1+\frac{(t-\mu)^2}{v\sigma_{ii}})^{-\frac{v+1}{2}}dt$$

再由

$$F^{-1}(x) = inf\{x : F(X) \geq x\}$$

所以

$$F_i^{-1}(u_i) = \mu_i + \sigma_{ii}^{1/2}t_v^{-1}(u_i)$$

带入上式，得

$$
\begin{aligned}
C(u_1,u_2,\cdots,u_p) =& F(\mu_1+\sigma_{11}^{1/2}t_v^{-1}(u_1),\mu_2+\sigma_{22}^{1/2}t_v^{-1}(u_2),\cdots,\mu_p+\sigma_{pp}^{1/2}t_v^{-1}(u_p)) \\
=& \int_{-\infty}^{x} f(F_1^{-1}(u_1),F_2^{-1}(u_2),\cdots,F_p^{-1}(u_p))|diag(\Sigma)|^{1/2}du
\end{aligned}
$$

其中 $|diag(\Sigma)|$ 为求导时得到。(因为 F 作用得多元变量可以表示为 $(\mathbf{t_v^{-1}}(\mathbf{u}))diag(\Sigma)^{1/2}$，所以在做变换时就多出来了这一项)

所以研究 $f(F_1^{-1}(u_1),F_2^{-1}(u_2),\cdots,F_p^{-1}(u_p))$ 即可，

$f(x)=\frac{\Gamma(\frac{v+p}{2})}{\Gamma(\frac{p}{2})\sqrt{det(\pi v\Sigma)}}(1+1/v(x-\mu)^T\Sigma^{-1}(x-\mu))^{-\frac{v+p}{2}}$ 中，自变量为 $F_i^{-1}(u_i)=\mu_i+\sigma_{ii}^{1/2}t_v^{-1}(u_i)$, 所以有

$$
\begin{aligned}
&(x-\mu)^T\Sigma^{-1}(x-\mu) \\
=&(t_v^{-1}(u_1),t_v^{-1}(u_2),\cdots,t_v^{-1}(u_p))^T diag(\Sigma^{1/2})\Sigma^{-1}diag(\Sigma^{1/2})(t_v^{-1}(u_1),t_v^{-1}(u_2),\cdots,t_v^{-1}(u_p)) \\
=&(\mathbf{t_v^{-1}}(\mathbf{u}))^T M^{-1}(\mathbf{t_v^{-1}}(\mathbf{u}))
\end{aligned}
$$

$$
\frac{|diag(\Sigma)|^{1/2}}{\sqrt{det(\Sigma)}}=\sqrt{det(M)^{-1}}
$$

所以

$$
C(u_1,u_2,\cdots,u_p)
$$

中，与 Σ 有关得部分均已经转化为 M 来表示，所以其 Copula 完全由矩阵 $M=diag(\Sigma)^{-1/2}\Sigma diag(\Sigma)^{-1/2}$ 与 v 决定。

Question 2

利用 t Copula model 生成满足多元 t 分布 $T_v(0,\Sigma;2)$ 的随机变量 X, 其中 $v=3,\sigma_{ij}=0.5^{|i-j|},i,j=1,2$。另一方面，利用 $X=Y/\sqrt{u/v}$,$Y\sim N(0,\Sigma),u\sim\chi_v^2$ 的方式生成随机变量 X. 通过可视化的方式来说明两种方法生成的样本大致吻合。

统计计算使用 R

P135/例 5.13 在例 5.10 中我们通过重要函数 $f_3(x)=e^{-x}/(1-e^{-1}),0<x<1$ 得到了最好的结果。通过 10000 次重复实验我们得到了估计值 $\hat{\theta}=$

0.5257801 和标准误差 0.0970314. 现在我们把区间 (0,1) 分成 5 个子区间 $(j/5,(j+1)/5), j=0,1,\cdots,4$

在第 j 个子区间上根据密度

$$\frac{5e^{-x}}{1-e^{-x}}, \frac{j-1}{5} < x < \frac{j}{5}$$

生成随机变量。实现过程留作练习。

5.6 在例 5.7 中，通过控制变量法计算了

$$\theta = \int_0^1 e^x dx$$

的蒙特卡罗积分。现在考虑对偶变量法。计算 $Cov(e^U, e^{1-U})$ 和 $Var(e^U + e^{1-U})$, 其中 $U \sim Uniform(0,1)$.(和简单蒙特卡罗方法比较) 使用对偶变量法方差缩减百分比能达到多少？

```
set.seed(123)
m <- 1000000
u <- runif(m,0,1)
y <- 1-u
cov <- cov(exp(u),exp(y))
var <- var(exp(u)+exp(y))
cat(" 协方差为:",cov,"\n")
```

```
## 协方差为: -0.2344374
```

```
cat(" 方差为: ",var,"\n")
```

```
## 方差为:  0.01565713
```

下面对比方差缩减的百分比

```
MC_Phi<- function(R=10000,anti=TRUE){
  u<-runif(R/2)
  if(!anti) v<-runif(R/2)else
    v<-1-u
  u<-c(u,v)
  g<-exp(u)
  theta <-mean(g)
  return(theta)
}

m<- 1000
MC1<-MC2<-numeric(m)
for(i in 1:m){
  MC1[i]=MC_Phi(R=10000,anti=FALSE)
  MC2[i]=MC_Phi(R=10000,anti=TRUE)
}
cat(" 普通方法得到的方差为",var(MC1),"\n")
```

```
## 普通方法得到的方差为 2.290312e-05
```

```
cat(" 对偶变量法得到的方差为",var(MC2),"\n")
```

```
## 对偶变量法得到的方差为 7.358485e-07
```

```
rate=(var(MC1)-var(MC2))/var(MC1)
print(rate*100)
```

```
## [1] 96.78713
```

可见，方差缩减了约 96.7%。

5.14 使用重要抽样法得到

$$\int_1^{+\infty} \frac{x^2}{\sqrt{2\pi}} e^{-x^2/2} dx$$

的蒙特卡罗估计。

解：令重要函数为

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}I(x>1)$$

$$g(x)=\frac{x^2}{\sqrt{2\pi}}e^{-x^2/2}I(x>1)$$

所以 $\frac{g(x)}{f(x)}=x^2I(x>1)$

```
set.seed(12)
m<-10000
g<-function(x){
  x^2/sqrt(2*pi)*exp(-x^2/2)*(x>1)
}
x<-rnorm(m)
fg<- x^2*(x>1)
theta_hat<-mean(fg)
print(theta_hat)
```

```
## [1] 0.4032721
```

得到的估计 $\hat{I}=0.4032721$

统计计算

3.5 设 $X\sim N(0,1)$, 则 $\theta=P(X>4.5)=3.398\times10^{-6}$。

(1) 如果直接生成 N 个 X 的随机数，用 $X_i>4.5$ 的比例估计 $P(X>4.5)$, 平均多少个样本点中才能有一个样本点满足 $X_i>4.5$?

解：设 Y 为满足 $X_i>4.5$ 的样本个数，则 $Y\sim Geo(3.398\times10^{-6})$

所以 $EY=1/p=1/(3.398\times10^{-6})\approx29429$

所以平均 29429 个样本点中才能有一个样本点满足足 $X_i>4.5$。

```
print(1/(3.398*10e-6))
```

```
## [1] 29429.08
```

(2) 取 V 为指数分布 $Exp(1)$，令 $W = V + 4.5$, 用 W 的样本进行重要抽样估计 θ, 取样本点个数 $N = 1000$, 求估计值并估计误差的大小。

解：

$$p(v) = e^{-v} I(v > 0)$$

做 $W = V + 4.5$ 变换, 于是得到 W 的概率密度函数为

$$p(w) = e^{-w+4.5} I(w > 4.5)$$

$$\theta = \int_{4.5}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

```
set.seed(123)
N=1000
v <- rexp(N,1)
w=v+4.5
g<-function(x){
  1/(sqrt(2*pi))*exp(-x^2/2)
}
fg<-g(w)/exp(-w+4.5)
theta <- mean(fg)
print(theta)
```

```
## [1] 3.280402e-06
```

得到的估计为 3.28×10^{-6}, 比较接近 3.398×10^{-6}

下面来估计误差的大小

```r
m <- 10000
Theta <- numeric(m)
for(i in 1:m){
  v<- rexp(N,1)
  w=v+4.5
  fg <- g(w)/exp(-w+4.5)
  Theta[i] <- mean(fg)
}
sd(Theta)
```

```
## [1] 1.411106e-07
```

```r
c(theta-1.96*sd(Theta),theta+1.96*sd(Theta))
```

```
## [1] 3.003825e-06 3.556978e-06
```

```r
cat(" 误差大约为",1.96*sd(Theta),"\n")
```

```
## 误差大约为 2.765767e-07
```

3.7 设 $\{U_i, i=1,2,\cdots\}$ 为独立同 $U(0,1)$ 分布的随机变量序列。令 M 为序列中第一个比前一个值小的元素的符号，即

$$M = min\{m : U_1 \le U_2 \le \cdots \le U_{m-1}, U_{m-1} > U_m, m \ge 2\}$$

(1) 证明 $P(M>n) = \frac{1}{n!}, n \ge 2$

解：

$$P(M>n) = P(U_1 \le U_2 \le \cdots \le U_n)$$

因为 $(U_1, U_2, \cdots, U_n)$ 之间的独立同分布性知道，其从大到小排列共有 $n!$ 中排列方式 $P(U_1 \le U_2 \le \cdots \le U_n)$ 恰为其中一种，所以有

$$P(M>n) = P(U_1 \le U_2 \le \cdots \le U_n) = \frac{1}{n!}$$

(2) 用概率论中的恒等式 $EM = \sum_{n=0}^{\infty} P(M > n)$, 证明 $EM = e$.

证明：

$$EM = \sum_{n=0}^{\infty} P(M > n) = \sum_{n=0}^{\infty} \frac{1}{n!} = e$$

上式得自于 Taylor 级数的展开。

(3) 生成 M 的 N 个独立抽样，用平均值 $\bar{M}$ 估计 e.

```
MC_M<-function(N=10000){
  count <-0
  M <- numeric(N)
  while(count<N){
    u<- runif(1,0,1)
    v<- runif(1,0,1)
    sig=2
    while(u<=v){
      u<-v
      v<- runif(1,0,1)
      sig=sig+1
    }
    count=count+1
    M[count]<- sig
  }
  return(mean(M))
}

e_hat = MC_M(N=10000)
print(e_hat)
```

```
## [1] 2.706
```

可见，得到的估计与 e 十分接近

(4) 估计 $\bar{M}$ 的标准差，给出 e 的近似 95 置信区间

```
m <- 1000
sdM = numeric(m)
for(i in 1:m){
  sdM[i] <- MC_M(N=10000)
}
cat("bar{M} 的标准差为",sd(sdM),"\n")
```

```
## bar{M}的标准差为 0.008703005
```

```
cat("e 的置信区间为 [",e_hat-1.96*sd(sdM),",",e_hat+1.96*sd(sdM),"]","\n")
```

```
## e的置信区间为[ 2.688942 , 2.723058 ]
```

3.9 用随机模拟法计算二重积分 $\int_0^1\int_0^1 e^{(x+y)^2}dydx$, 用对立变量法改善精度。

解：$g(x,y)=e^{(x+y)^2}$ 是关于 (x,y) 的单调增函数

```
MC_I<- function(R=10000,anti=TRUE){
  x<-runif(R/2,0,1)
  y<- runif(R/2,0,1)
  if(!anti){
    x1<-runif(R/2,0,1)
    y1<- runif(R/2,0,1)
  }else{
    x1 <- 1-x
    y1 <- 1-y
  }
  x=c(x,x1)
  y=c(y,y1)
  g<-exp((x+y)^2)
  I_hat <-mean(g)
  return(I_hat)
}
```

```r
m<- 1000
MC1<-MC2<-numeric(m)
for(i in 1:m){
  MC1[i]=MC_I(R=10000,anti=FALSE)
  MC2[i]=MC_I(R=10000,anti=TRUE)
}

cat(" 普通方法得到的估计为:",MC_I(R=10000,anti = FALSE),"\n")
```

```
## 普通方法得到的估计为: 4.997806
```

```r
cat(" 对偶变量法法得到的估计为:",MC_I(R=10000,anti = TRUE),"\n")
```

```
## 对偶变量法法得到的估计为: 4.940414
```

```r
rate <- (var(MC1)-var(MC2))/var(MC1)
cat(" 使用对偶变量法精度提升了",rate*100,"%")
```

```
## 使用对偶变量法精度提升了 37.57776 %
```