Questão 3

Mário/Rodney

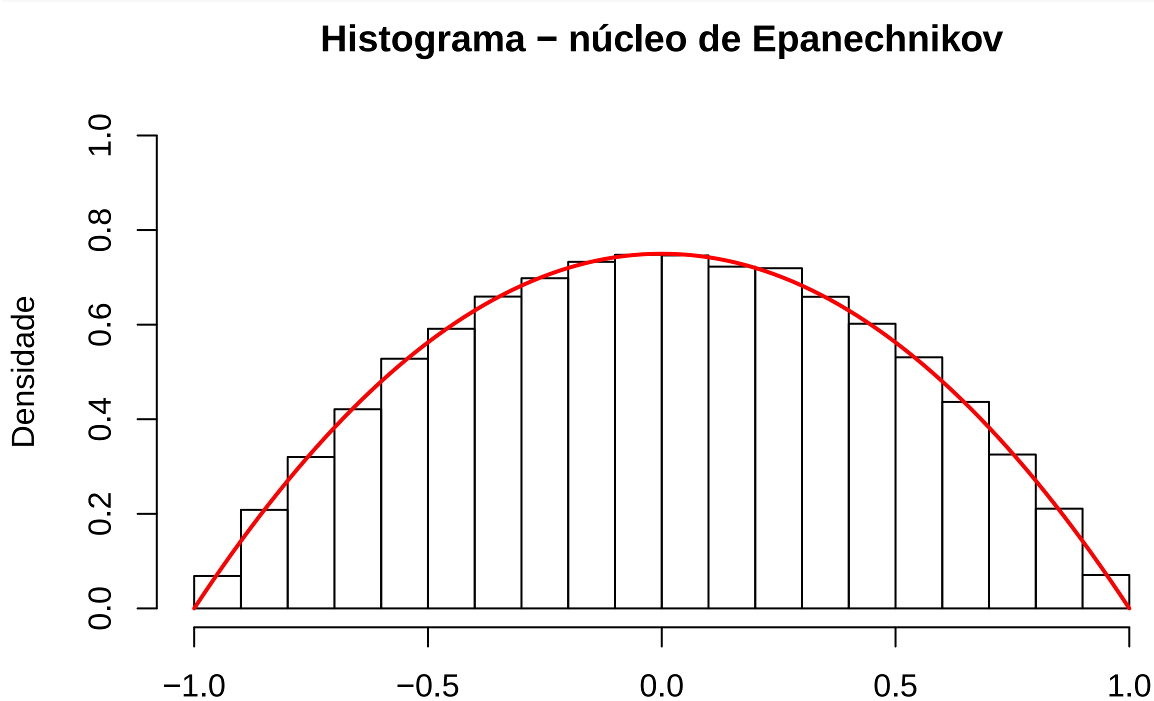
April 6, 2017

Questão 3:

No código a seguir, simulamos uma amostra da variável Epanechnikov usando o algoritmo descrito na questão e, em seguida, construimos um histograma das variáveis geradas.

```
# densidade Epanechnikov
d_epach = function(x) 3*(1 - x^2)/4
# função para gerar variáveis aleatórias com distribuição Epanechnikov
r_epach = function(n){
  v = rep(0,n)
  for(i in 1:n){
    u1 = runif(1,-1,1); u2 = runif(1,-1,1); u3 = runif(1,-1,1);
    if((abs(u3)>abs(u2))&&(abs(u3)>abs(u1))){
      v[i] = u2
    }else{
      v[i] = u3
    }
  }
  return(v)
}

# histograma de uma amostra da Epanechnikov e a densidade verdadeira
hist(r_epach(100000),prob=TRUE,main="Histograma - núcleo de Epanechnikov",
     ylim=c(0,1),ylab="Densidade",xlab="")
curve(d_epach, xlim=c(-1,1),add=TRUE,col=2,lwd=2)
```



Comparando o histograma com a densidade Epanechnikov, vemos que os valores simulados estão bastante próximos da densidade teórica.

Nessa parte vamos provar que o algoritmo de fato gera valores da densida Epanechnikov. Sabemos que $U_i \sim U(-1,1)$, $i \in \{1,2,3\}$, daí a função de distribuição de U_i é

$$F_U(u) = \frac{u+1}{2},$$

então, para $Y_i = |U_i|$ temos que

$$F_{Y_i}(y) = P(|U_i| \leq y) = P(-y \leq U_i \leq y) = F_{U_i}(y) - F_{U_i}(-y) = \left(\frac{y+1}{2}\right) - \left(\frac{-y+1}{2}\right) = y,$$

ou seja, $Y_i \sim U(0,1)$, $i \in \{1,2,3\}$.

Sendo X a variável aleatória gerada de acordo com o algoritmo descrito na questão, temos que

$$X = \begin{cases} U_2 & \text{se } |U_3| > max\{|U_1|,|U_2|\} \\ U_3 & \text{se } |U_3| \leq max\{|U_1|,|U_2|\} \end{cases},$$

e como U_2 e U_3 são simétricas em torno de zero, pelo algoritmo temos que X também o será. Logo, $P(X<0) = 1/2$. Considerando então $0 < x < 1$, temos que

$$\begin{aligned} P(|X| \leq x) &= P(|U_3| > max\{|U_1|,|U_2|\}, |U_2| \leq x) + P(|U_3| > max\{|U_1|,|U_2|\}, |U_3| \leq x) \\ &= P(Y_3 > max\{Y_1,Y_2\}, Y_2 \leq x) + P(Y_3 > max\{Y_1,Y_2\}, Y_3 \leq x). \end{aligned}$$

Sabemos que a densidade de Y_i é $f_{Y_i}(y) = 1$ para $0 < y < 1$. Assim, calculando a primeira probabilidade temos que

$$\begin{aligned} P(Y_3 > max\{Y_1,Y_2\}, Y_2 \leq x) &= P(Y_1 \leq Y_2 \leq Y_3, Y_2 \leq x) + P(Y_2 \leq Y_2 \leq Y_3, Y_2 \leq x) \\ &= \int_0^x \int_0^{y_2} \int_{y_2}^1 dy_3 dy_1 dy_2 + \int_0^x \int_{y_2}^1 \int_{y_1}^1 dy_3 dy_1 dy_2 \\ &= \int_0^x \int_0^{y_2} (1-y_2) dy_1 dy_2 + \int_0^x \int_{y_2}^1 (1-y_1) dy_1 dy_2 \\ &= \int_0^x (1-y_2) y_2 dy_1 dy_2 + \int_0^x \left[(y_1 - \frac{y_1^2}{2})\Big|_{y_2}^1 \right] dy_2 \\ &= \left(\frac{y_2^2}{2} - \frac{y_2^3}{3}\right)\Big|_0^x + \int_0^x \left[1 - \frac{1}{2} - y_2 + \frac{y_2^2}{2}\right] dy_2 \\ &= \frac{x^2}{2} - \frac{x^3}{3} + \left[\frac{y_2}{2} - \frac{y_2^2}{2} + \frac{y_2^3}{6}\right]\Big|_0^x \\ &= \frac{x^2}{2} - \frac{x^3}{3} + \frac{x}{2} - \frac{x^2}{2} + \frac{x^3}{6} = \frac{x}{2} - \frac{x^3}{6} \end{aligned}$$

Na segunda probabilidade, temos que

$$\begin{aligned} P(Y_3 > max\{Y_1,Y_2\}, Y_3 \leq x) &= P(Y_3 \leq Y_2 \leq Y_1, Y_3 \leq x) + P(Y_2 \leq Y_3 \leq Y_1, Y_3 \leq x) \\ &\quad + P(Y_3 \leq Y_1 \leq Y_2, Y_3 \leq x) + P(Y_1 \leq Y_3 \leq Y_2, Y_3 \leq x) \\ &= 2[P(Y_3 \leq Y_2 \leq Y_1, Y_3 \leq x) + P(Y_2 \leq Y_3 \leq Y_1, Y_3 \leq x)], \end{aligned}$$

onde na primeira igualdade listamos as possíveis formas que o evento $[Y_3 > max\{Y_1, Y_2\}, Y_3 \leq x]$ pode ocorrer e na segunda igualdade usamos o fato de Y_1 e Y_2 serem identicamente distribuídas. Daí

$$\begin{aligned}
P(Y_3 > max\{Y_1, Y_2\}, Y_3 \leq x) &= 2\left[\int_0^x \int_{y_3}^1 \int_{y_2}^1 dy_1 dy_2 dy_3 + \int_0^x \int_0^{y_3} \int_{y_3}^1 dy_1 dy_2 dy_3\right] \\
&= 2\left[\int_0^x \int_{y_3}^1 (1-y_2) dy_2 dy_3 + \int_0^x \int_0^{y_3} (1-y_3) dy_2 dy_3\right] \\
&= 2\left[\int_0^x \left[(y_2 - \frac{y_2^2}{2})\Big|_{y_3}^1 \right] dy_3 + \int_0^x (1-y_3) y_3 dy_3\right] \\
&= 2\left[\int_0^x \left[\frac{1}{2} - y_3 + \frac{y_3^2}{2}\right] dy_3 + \left(\frac{y_3^2}{2} - \frac{y_3^3}{3}\right)\Big|_0^x\right] \\
&= 2\left[\left[\frac{y_3}{2} - \frac{y_3^2}{2} + \frac{y_3^3}{6}\right]\Big|_0^x + \frac{x^2}{2} - \frac{x^3}{3}\right] \\
&= 2\left[\frac{x}{2} - \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^2}{2} - \frac{x^3}{3}\right] = x - \frac{x^3}{3}
\end{aligned}$$

Portanto,

$$P(|X| \leq x) = \frac{x}{2} - \frac{x^3}{6} + x - \frac{x^3}{3} = \frac{3x}{2} - \frac{3x^3}{6} = \frac{3x - x^3}{2},$$

e pela simetria de X, temos que

$$P(X \leq x) = P(X \leq 0) + P(0 < X \leq x) = \frac{1}{2} + \frac{P(|X| \leq x)}{2} = \frac{1}{2} + \frac{3x - x^3}{4},$$

onde $0 < x < 1$. Novamente usando o fato da distribuição de X ser simétrica, temos para $0 < x < 1$ que

$$F_X(-x) = 1 - F_X(x) = 1 - \left[\frac{1}{2} + \frac{3x - x^3}{4}\right] = \frac{1}{2} - \frac{(3x - x^3)}{4} = \frac{1}{2} + \frac{3(-x) - (-x)^3}{4},$$

ou seja, a função de distribuição de X é $F_X(x) = 1/2 + (3x - x^3)/4$, para $x \in [-1, 1]$, e derivando $F_X(x)$ em relação a x obtemos a função densidade de X, que é

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

assim, vemos que X tem distribuição Epanechnikov, e o método de geração está correto.

Questão 10:

A questão pede que estimemos a esperança de X^2 de uma distribuição normal padrão definida no intervalo $[1, \infty)$ pelo método da amostragem por importância. A esperança a ser estimada é descrita a seguir:

$$E(X^2) = \int_1^\infty \frac{x^2}{\sqrt{2\pi}} exp\left(\frac{-x^2}{2}\right)$$

Uma distribuição interessante para ser utilizada é a distribuição normal padrão truncada em 1, porque possui o mesmo suporte que a densidade que temos $[0, \infty)$ e, além disso, a razão entre ambas gera uma expressão estável. A estabilidade se dá pela semelhança entre a distribuição normal que a questão oferece e a distribuição normal truncada.

A razão referida anteriormente pode ser definida, com alguma simplificação de notação, da seguinte maneira:

$$R = \frac{g(x)}{f(x)},$$

onde $g(x)$ denota a distribuição original, no caso a distribuição normal padrão pré multiplicada por x^2 e $f(X)$ a função de importância. É importante observar que a variância do estimador depende de R, logo é aconselhável obter razões R mais estáveis.

A implementação do algoritmo de importância segue abaixo:

```
MCNormTruncX2 <- function(n,lower=1,upper=Inf,media=0,dp=1){
# verificando se o pacote truncnorm precisa ser instalado
if("truncnorm" %in% rownames(installed.packages())==FALSE)
{install.packages("truncnorm");library(truncnorm)}
  else library(truncnorm)

rtn      <- rtruncnorm(n=n,a=lower,b=upper,mean=media,sd=dp)
int1     <- ((rtn^2)*dnorm(x=rtn,mean=media,sd=dp))/dtruncnorm(x=rtn,
        a=lower,b=upper,mean=media,sd=dp)
impMC1   <- mean(int1)
varinf1  <- var(int1)/n

result   <- list(thetahat = impMC1,
                 varhat   = varinf1)
return(result)
}
cbind(rbind(MCNormTruncX2(n=100),MCNormTruncX2(n=1e4),MCNormTruncX2(n=1e6)),c(100,1e4,1e6))
```

```
##      thetahat  varhat
## [1,] 0.4012294 0.0005732426 100
## [2,] 0.3993562 6.550068e-06 10000
## [3,] 0.4007475 6.854187e-08 1e+06
```

Note que os valores de truncagem estão pré definidos, mas nada impede que os valores de *lower* e *upper* sejam substituidos. O mesmo ocorre com os valores da média e do desvio-padrão. Note também que os resultados indicam que o estimador é consistente, pois à medida que se eleva n, a variância do estimador reduz considerávelmente.