Universidad ORT Uruguay
Facultad de Ingeniería

Probabilidad y Estadística

Obligatorio

Diciembre 2020

Documentación

Docente:
Mateus Castelli

Alumno:
Matias Hernández 169236

Índice

Comandos

dpois(x, lambda):

Es la probabilidad de que ocurran ‘x’ sucesos en un período cuando el valor esperado de eventos es lambda.

Ejemplo: ¿Cuál es la probabilidad de hacer entre 2 y 4 ventas en la semana, si el promedio de ventas es de 3 por semana?

```
dpois(x = 2, lambda = 3) +
dpois(x = 3, lambda = 3) +
dpois(x = 4, lambda = 4)
## [1] 0.6434504
# Valor esperado de ventas = lambda = 3
# Varianza = lambda = 3
```

dhyper(x, m, n, k):

Es el número de éxitos en una muestra de tamaño k (sin reposición) cuando la población tiene M éxitos y N no-éxitos.

Ejemplo: ¿Cuál es la probabilidad de seleccionar x=14 manzanas rojas de una muestra que contiene k=20 manzanas rojas desde una urna que tiene m=70 manzanas rojas y n=30 manzanas verdes?

```
# probabilidad
dhyper(x = 14, m = 70, n = 30, k = 20)
## [1] 0.2140911
# Valor esperado de manzanas rojas
k * m / (m + n)
## [1] 14
# Varianza
k * m / (m + n) * (m + n - k) / (m + n) * n / (m + n - 1)
## [1] 3.393939
```

dbinom(x, n, p):

La distribución binomial es una distribución de probabilidad discreta. Describe el resultado de n ensayos independientes en un experimento. Se supone que cada ensayo tiene solo dos resultados, éxito o fracaso. Si la probabilidad de un ensayo exitoso es 'p', entonces la probabilidad de tener 'x' resultados exitosos en un experimento de 'n' ensayos independientes es la siguiente:

$f(x) = \binom{n}{x} p^x (1-p)^{(n-x)}$ donde x = 0, 1, 2, …, n.

Y la misma está representada por la función dbinom(r, n, p).

Ejemplo: Suponga que hay doce preguntas de opción múltiple en una prueba de clase de inglés. Cada pregunta tiene cinco posibles respuestas, y solo una de ellas es correcta. Encuentre la probabilidad de tener cuatro o menos respuestas correctas si un estudiante intenta responder todas las preguntas al azar.

Solución: Dado que solo una de cada cinco respuestas posibles es correcta, la probabilidad de responder una pregunta correctamente al azar es 1/5 = 0.2. Podemos encontrar la probabilidad de tener exactamente 4 respuestas correctas mediante intentos aleatorios de la siguiente manera.

```
dbinom(x=4, n=12, p=0.2)
## [1] 0.1329
```

Para encontrar la probabilidad de tener cuatro o menos respuestas correctas mediante intentos aleatorios, se aplica la función dbinom con x = 0 , … , 4.

```
dbinom(x=0, n=12, p=0.2) +
dbinom(x=1, n=12, p=0.2) +
dbinom(x=2, n=12, p=0.2) +
dbinom(x=3, n=12, p=0.2) +
dbinom(x=4, n=12, p=0.2)
## [1] 0.9274
```

dgeom(x, p, log=FALSE):

dgeom determina la probabilidad de que 'x' fallas ocurran antes del primer éxito cuando la probabilidad del suceso es 'p'.

Ejemplo: Un entrevistador aleatoreamente selecciona personas en la calle hasta que encuentra alguien que haya asistido a un juego de fútbol de la última temporada. Cuál es la probabilidad de que el entrevistado encuentre x=3 personas que no hayan asistido a un juego de la última temporada antes de encontrar a una persona que sí haya asistido, cuando la probabilidad de que una persona atienda a un juego de la última temporada es p=0.20.

Solución:

```
dgeom(x = 3, p = 0.20)
## [1] 0.1024
```

barplot:

Crea gráficos de barras, donde la altura es un vector o una matriz.

- Si la altura es un vector, los valores determinan las alturas de las barras en el gráfico.
- Si la altura es una matriz y la opción beside=FALSO, entonces cada barra del gráfico corresponde a una columna de altura y los vectores de la columna dan las alturas de las "subbarras" apiladas.
- Si la altura es una matriz y beside=VERDADERO, entonces los valores en cada columna se yuxtaponen en lugar de apilar.

Incluya la opción names.arg=(vector de caracteres) para etiquetar las barras.

La opción horiz=TRUE para crear un diagrama de barras horizontal.

c:

Combina valores dentro de un vector o lista.

punif(q, min=0, max=1, lower.tail=TRUE, log.p=FALSE):

Retorna la función de distribución de una variable con distribución Normal.

- q: vector de cuantiles.
- min, max: Determina los límites de la distribución.
- Log.p: Si es TRUE, las probabilidades p son dadas como log(p).
- lower.tail: Si es TRUE, las probabilidades son P[X <= x], sino serían P[X > x].

plot:

Dado un conjunto de puntos para el eje de las x y otro conjunto de puntos para el eje de las y, plot se encarga de graficar dichos puntos.

Existen otros parámetros que pueden ser usados para personalizar las gráficas, pero la versión simple, la cual sólo grafica es:

```
plot(x, y)
```

Un ejemplo más completo, con otros posibles parámetros sería:

```
plot(
  x,
  sin(x),
  main="Overlaying Graphs",
  ylab="",
  type="l",
  col="blue")
  lines(x,cos(x), col="red")
  legend("topleft",
  c("sin(x)","cos(x)"),
  fill=c("blue","red")
)
```

pnorm:

Calcula la función de distribución acumulada de una variable aleatoria X con distribución Normal.

```
pnorm(x, mean=50, sd=20)
```

- ‘x’: Debería ser el número tal que $P(X \leq x)$
- ‘mean’: Es el valor medio de la distribución.
- ‘sd’: Es la desviación estándar de la distribución.

rnorm:

Genera múltiples variables con distribución normal.

```
rnorm(x, mean=50, sd=20)
```

- ‘x’: Es la cantidad de variables aleatorias que se van a generar de forma aleatoria.
- ‘mean’: Es el valor medio de la distribución.
- ‘sd’: Es la desviación estándar de la distribución.

dnorm:

Determina la densidad de probabilidad de una variable aleatoria con distribución normal.

```
dnorm(x, mean=50, sd=20)
```

- ‘x’: Es una secuencia de valores que determinarán los límites de la distribución.
- ‘mean’: Es el valor medio de la distribución.
- ‘sd’: Es la desviación estándar de la distribución.

curve:

Dibuja una curva correspondiente a una función sobre el intervalo [from, to]

Puede graficar también una expresión en la variable xname.

curve(expr, from, to, n, add, type, xname, xlab, ylab, log, xlim, …)

- 'expr' es el nombre de una función, llamada o expresión escrita como una función de x, la cual será evaluada a un objeto del mismo largo que x.
- 'from, to' es el rango sobre el cual la función será graficada.
- 'n' es el número de valores de x que serán evaluados.
- 'add' si es TRUE, se agregan a un gráfico existente. Si NA crea un nuevo gráfico tomando por defecto los límites y la escala del eje x del gráfico previo.
- 'type' es el tipo de gráfico.
- 'xname' da el nombre para ser usado en el eje x.
- 'xlab, ylab, log'
- 'xlim'

hist:

Para crear un histograma, se utilizará la función siguiente:

hist(v, main, xlab, xlim, ylim, breaks, col, border)

- 'v' es un vector que contiene valores numéricos utilizados en el histograma.
- 'main' indica el título del gráfico.
- 'xlab' se usa para dar una descripción al eje x.
- 'xlim' se usa para especificar un rango de valores en el eje x.
- 'ylim' se usa para especificar un rango de valores en el eje y.
- 'breaks' se usa para mencionar el ancho de cada barra.
- 'col' se usa para configurar el color de las barras.
- 'border' se usa para configurar el color de los bordes de cada barra.

matrix(elements, byrow, nrow):

El primer argumento es la colección de elementos que R organizará en las filas y columnas de la matriz.

El argumento 'byrow' indica que la matriz está llena de filas. Si queremos que la matriz se llene con las columnas, simplemente se coloca 'byrow=FALSE'.

El tercer argumento 'nrow' indica cuantas filas debe tener la matriz.

dchisq:

Determina la densidad de probabilidad de una variable aleatoria con distribución chi cuadrado.

```
dchisq(4:8, df = 7)
```

Por ejemplo, en el ejemplo de arriba, se calcula la densidad de probabilidad para los valores enteros desde el 4 hasta el 8 de una curva chi cuadrado con df = 7.

Ejercicio 1

Parte 1)

El número de canchas de fútbol alquiladas en el club Andresito sigue una distribución de Poisson con un promedio de 10 canchas alquiladas por día.

A. **¿Cuál es la probabilidad de que en un día cualquiera se alquilen mínimo 7 y máximo 11 canchas de fútbol?**

Sea la V.A.D $X / X = "Cantidad\ de\ canchas\ alquiladas\ en\ el\ día"$.

Como el promedio de alquiler de canchas es de 10 por día $\Rightarrow$ $E(X) = \lambda = 10$.

X respeta la distribución de Poisson con parámetro $\lambda = 10 \Rightarrow X \sim Poi(\lambda = 10)$.

La probabilidad de que en un día cualquiera se alquilen mínimo 7 y máximo 11 canchas de fútbol es representado por la función de probabilidad siguiente:

$$P(7 \leq X \leq 11) = P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10) + P(X = 11).$$

Se realiza el cálculo en R:

```
dpois(x = 7, lambda = 10) +
dpois(x = 8, lambda = 10) +
dpois(x = 9, lambda = 10) +
dpois(x = 10, lambda = 10) +
dpois(x = 11, lambda = 10)
# [1] 0.5666347
```

$\Rightarrow \boxed{P(7 \leq X \leq 11) = 0.5666347}$

B. ¿Cuál es la probabilidad de que en tres días seguidos se alquilen 36 canchas de fútbol?

Sea la V.A.D. $Y/Y =$ "$Cantidad\ de\ canchas\ alquiladas\ en\ 3\ días\ seguidos$".

Como se alquilan en promedio, 10 canchas por día ⇒ en 3 días se alquilarán en promedio 30 canchas ⇒ $E(Y) = 30$.

Y respeta la distribución de Poisson con parámetro $\lambda = 30 \Rightarrow Y \sim Poi(\lambda = 30)$.

La probabilidad de que en tres días seguidos se alquilen 36 canchas de fútbol es determinado por la función de probabilidad siguiente: $P(Y = 36)$

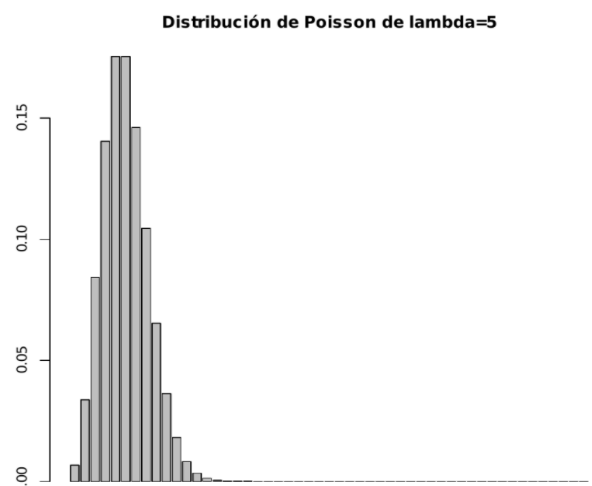
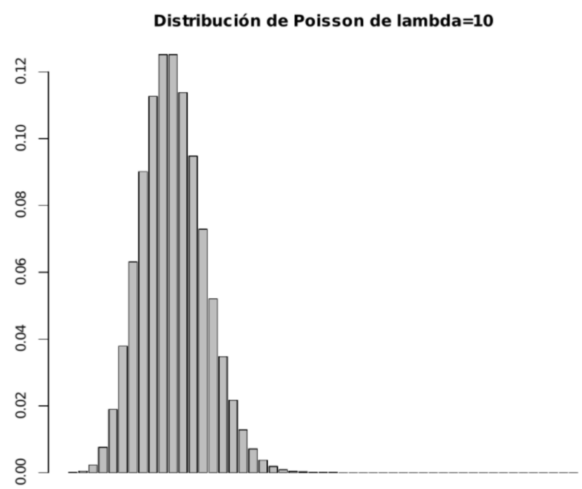
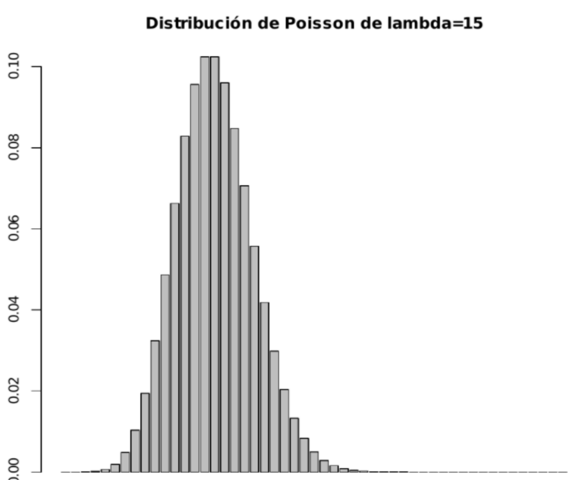
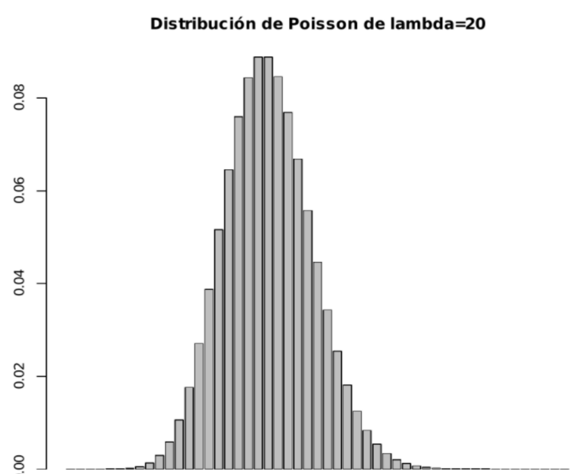
Se realiza el cálculo en R:

```
dpois(x = 36, lambda = 30)
# [1] 0.03775683
```

⇒ $\boxed{P(Y = 36) = 0.03775683}$

Parte 2)

Representar los diagramas de barra de una distribución Poisson con los valores de λ= 5, 10, 15, 20. Los valores de ocurrencias “x” tomando un rango de 0 a 50.

Caso $\lambda = 5$: 	Caso $\lambda = 10$: 
Caso $\lambda = 15$: 	Caso $\lambda = 20$: 

Ejercicio 2)

En una antigua facultad de Ingeniería se sabe que la probabilidad de aprobar el examen final de Probabilidad y Estadística es del 52%. Lucas se va a presentar dicho examen. ¿Cuál es la probabilidad de que lo apruebe en la tercera oportunidad?

Sea la V.A.D. $X/X = "Aprobar\ el\ examen\ en\ el\ intento"$.

$X \sim Ber(p = 0.52)$ o $X \sim Bin(n = 1, p = 0.52)$ con $R_X = \{0, 1\}$

```
P_X1 <- dbinom(1, size=1, prob=0.52)
# [1] 0.52
```

$P(X = 1) = 0.52$

Como se debe calcular considerando múltiples intentos:

Sea la V.A.D. $Y/Y = "Cantidad\ de\ veces\ que\ se\ aprueba\ el\ examen"$.

$Y \sim Bin(n = 3, p = 0.52)$ con $R_Y = \{0, 1, 2, 3, \dots\}$

Entonces, para 3 intentos independientes con probabilidad de éxito de 0.52 cada uno, se tiene una probabilidad puntual de:

$P(X = j) = C_j^n * p^j * (1-p)^{n-j}$ con $j = 1$

```
P_Y1 <- dbinom(1, size = 3, prob = 0.52)
# [1] 0.359424
```

$P(Y = 1) = 0.359424$

Como esta es la probabilidad de que se apruebe el examen 1 vez en 3 intentos, entonces se debe dividir entre 3 (esto es posible ya que los intentos son independientes), porque sólo me interesa conocer cuando se aprueba en la tercera oportunidad.

Sea la V.A.D. $Z/Z =$ "Intento en el que se aprueba el examen".

Entonces $P(Z = 3) = \frac{P(X=1)}{3}$

```
P_Z3 <- P_Y/3
# [1] 0.119808
```

$$\boxed{P(Z = 3) = 0.119808}$$

Ejercicio 3)

En cierto punto de información de Atlántida el tiempo dedicado a orientar al público sigue una distribución exponencial con un tiempo medio de 2.5 minutos.
¿Cuál es la probabilidad de que el tiempo de orientación al público sea menos de 4 minutos?

Sea la V.A.D. $X/X = "Tiempo\ de\ orientación\ al\ público"$.

E(X) = 2.5 min. (tiempo medio)

$$E(X) = \frac{1}{\lambda} = 2.5 \Rightarrow \lambda = \frac{1}{2.5} = 0.4$$

$$X \sim Exp(\lambda = 0.4)$$

Como el tiempo de orientación al público debe ser menos de 4 minutos, entonces:

$$P(X < 4) = P(X \leq 4) = F_X(4)$$

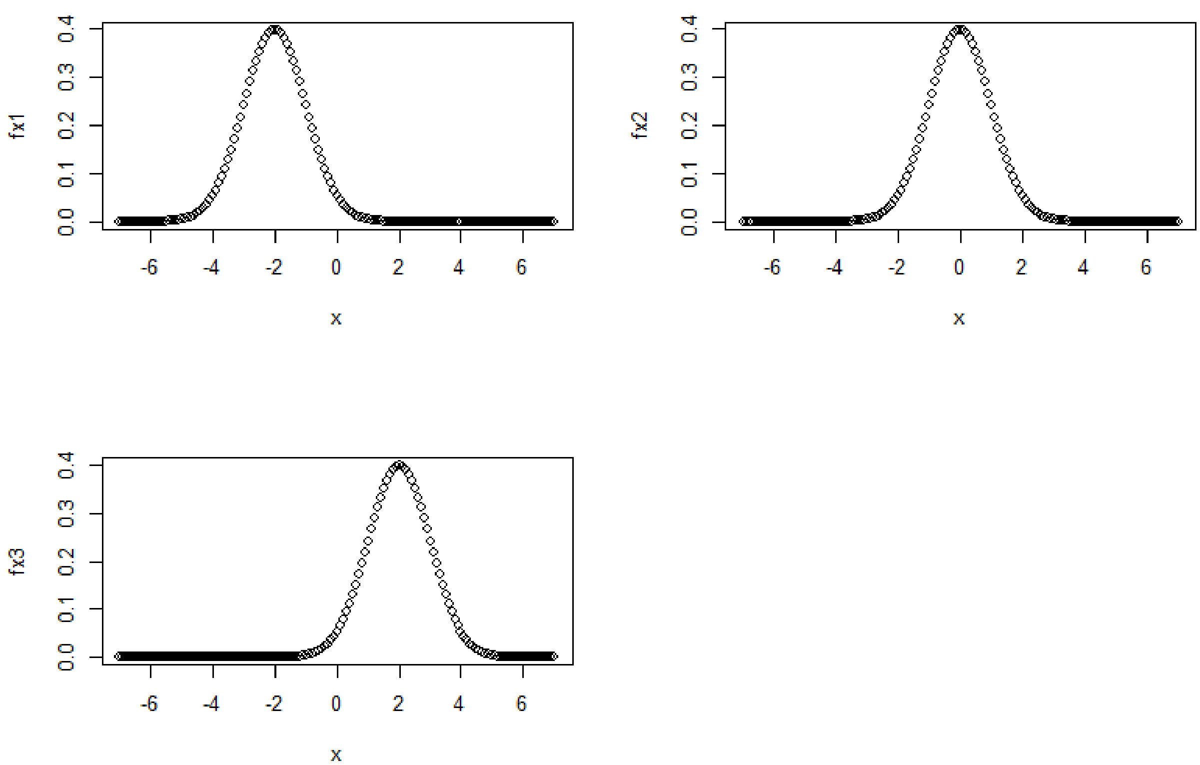
```
F_X4 <- pexp(4, rate=0.40)
# [1] 0.7981035
```

$$\boxed{P(X < 4) = 0.7981}$$

Ejercicio 4)

Graficar la función de densidad de probabilidad de una distribución normal desviación estándar σ=1 variando con los valores de μ=-2, 0, 2.

```
x <- seq(-7, 7, by = .1)

fx1 <- dnorm(x, mean = -2, sd = 1)
fx2 <- dnorm(x, mean = 0, sd = 1)
fx3 <- dnorm(x, mean = 2, sd = 1)

par(mfrow = c(2, 2))

plot(x, fx1)
plot(x, fx2)
plot(x, fx3)
```



Ejercicio 5)

a) Simular una muestra de 5 variables aleatorias independientes normales con media 5 y desvío estándar 1.
Calcular los estadísticos $\bar{X}$ y S^2.

Se obtienen 5 V.A. Ind. Normales con media 5 y desvío estándar 1:

```
muestra <- rnorm(5, mean = 5, sd = 1)
# [1] 5.375674 3.673414 5.391707 3.975449 5.526854
```

$$X_1 = 5.375674$$

$$X_2 = 3.673414$$

$$X_3 = 5.391707$$

$$X_4 = 3.975449$$

$$X_5 = 5.526854$$

Se procede a calcular el estadístico del promedio $\overline{X_5} = \frac{X_1+X_2+X_3+X_4+X_5}{5}$:

```
EstProm <- sum(muestra)/5
# [1] 4.788619
```

$$\boxed{\overline{X_5} = 4.788619}$$

Se procede a calcular el estadístico Cuasi Varianza $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$ con n=5:

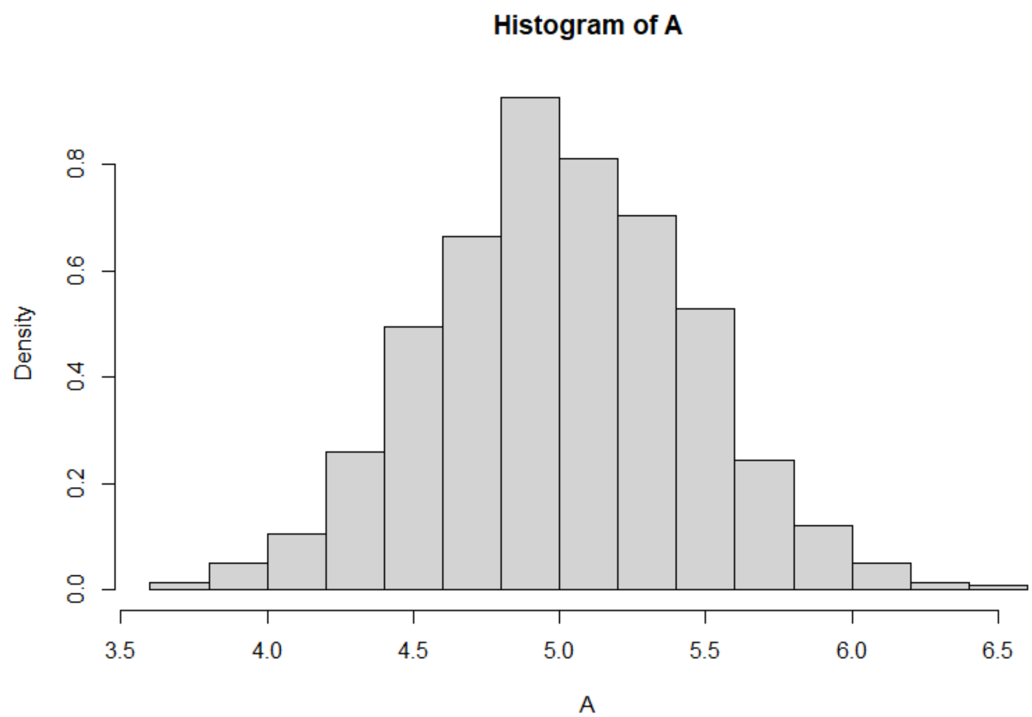
```
n = 5
mses <- sapply(muestra, function(x) (x - EstProm)^2)
EstCuaVar <- (1/(n-1))*sum(mses)
# [1] 0.7895663
```

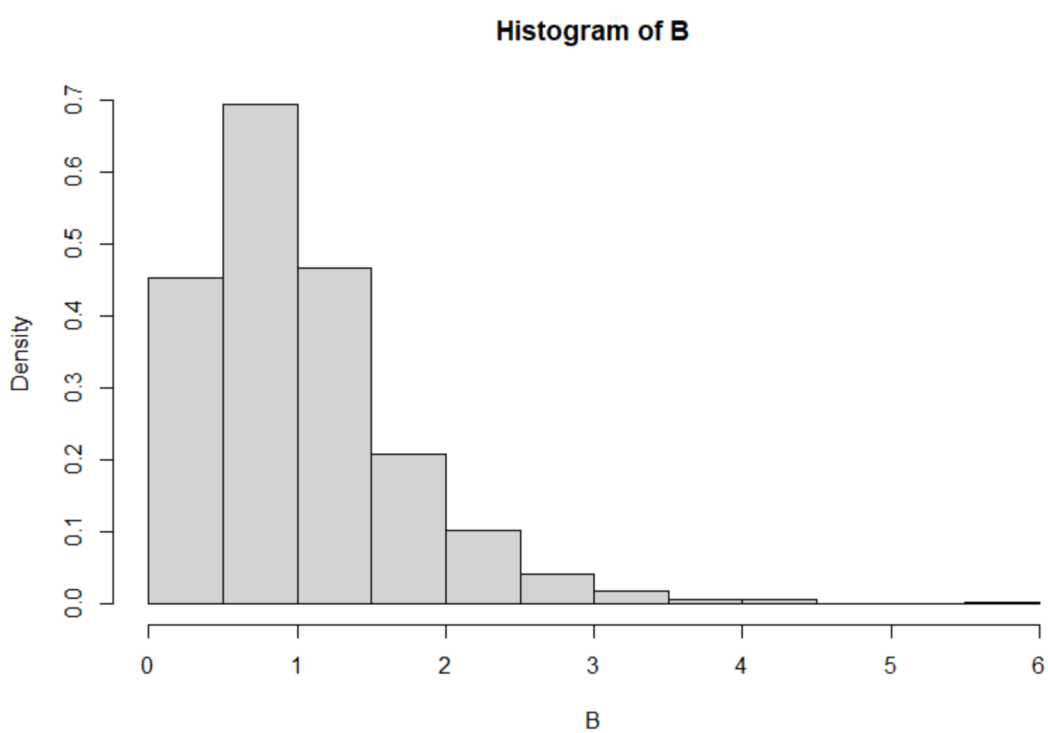
$$\boxed{S^2 = 0.7895663}$$

b) Mediante un comando "for", repetir la parte a) 1000 veces, almacenando los cálculos por separado de X y S2 en variables que se llamarán A y B. Calcular estadísticos descriptivos.

```
dataLength <- 1000
A <- vector(mode = "list", length = dataLength)
B <- vector(mode = "list", length = dataLength)

for (i in 1:dataLength)
{
  n = 5

  muestra <- rnorm(n, mean = 5, sd = 1)

  EstProm <- sum(muestra)/n
  A[i] <- EstProm


  mses <- sapply(muestra, function(x) (x - EstProm)^2)
  EstCuaVar <- (1/(n-1))*sum(mses)
  B[i] <- EstCuaVar
}

A <- gsub(",", "", A)
A <- as.numeric(A)
hist(A, freq = FALSE)
summary(A)

B <- gsub(",", "", B)
B <- as.numeric(B)
hist(B, freq = FALSE)
summary(B)
```

Histogramas:

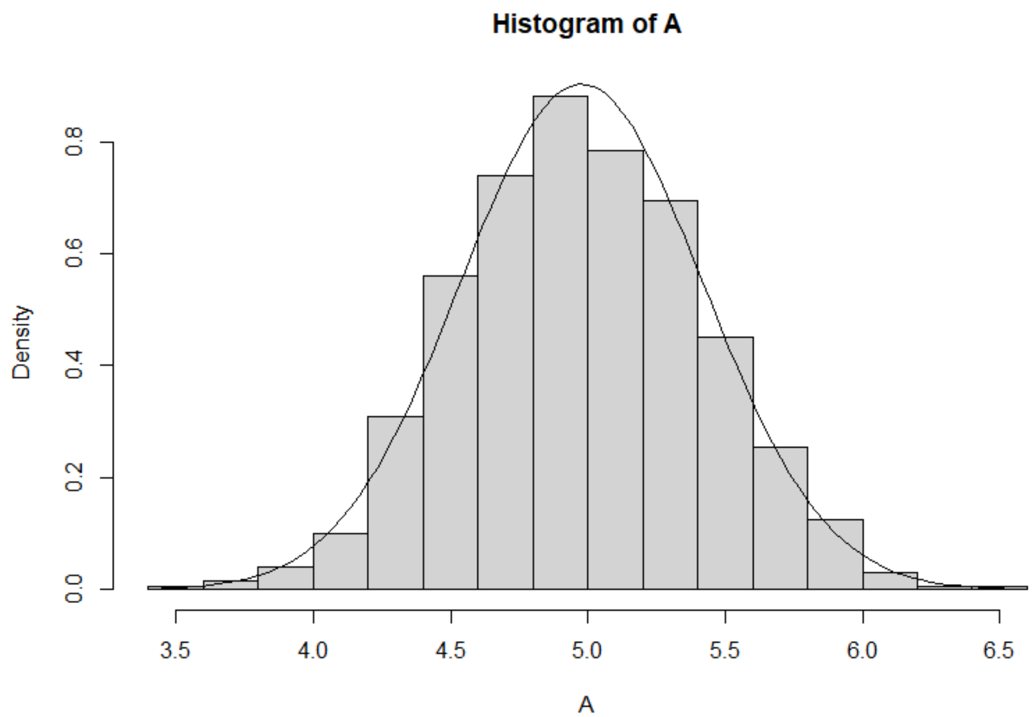


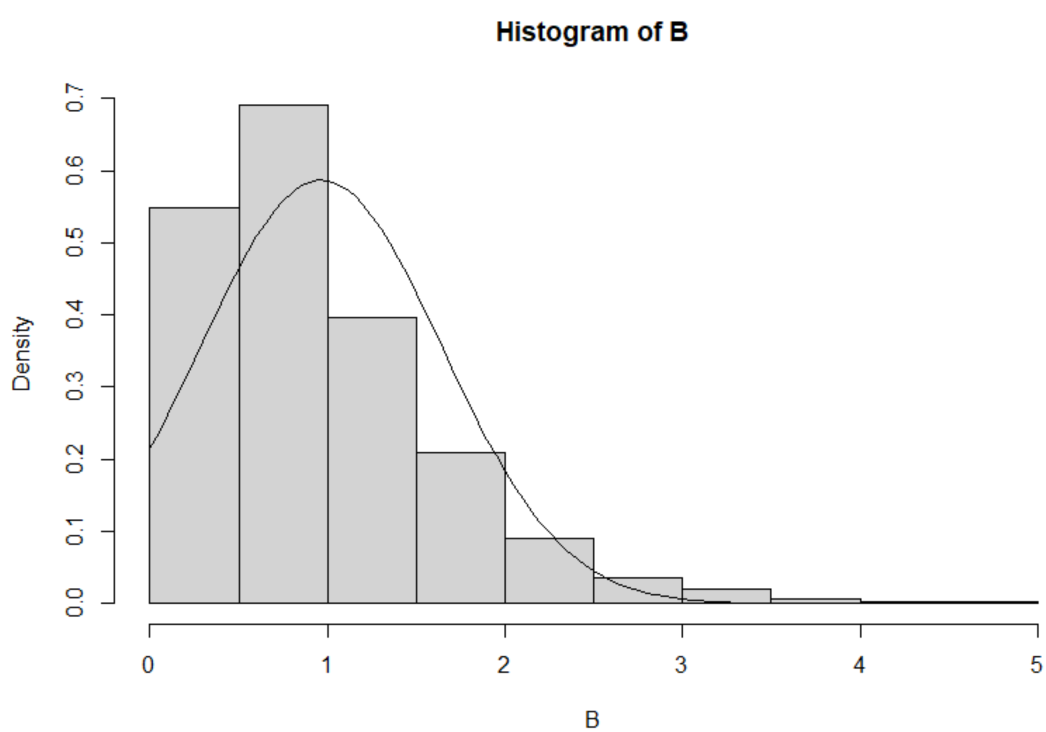


c) Realizar un histograma de las variables A y B, graficando densidad en el eje y. Comparar con el gráfico de la función de densidad de una variable normal.

```
dataLength <- 1000
A <- vector(mode = "list", length = dataLength)
B <- vector(mode = "list", length = dataLength)

for (i in 1:dataLength)
{
  n = 5

  muestra <- rnorm(n, mean = 5, sd = 1)

  EstProm <- sum(muestra)/n
  A[i] <- EstProm



  mses <- sapply(muestra, function(x) (x - EstProm)^2)
  EstCuaVar <- (1/(n-1))*sum(mses)
  B[i] <- EstCuaVar
}

A <- gsub(",", "", A)
A <- as.numeric(A)
hist(A, freq = FALSE)
summary(A)
curve(dnorm(x, mean(A), sd(A)), add=T)

B <- gsub(",", "", B)
B <- as.numeric(B)
hist(B, freq = FALSE)
summary(B)
curve(dnorm(x, mean(B), sd(B)), add=T)
```

Gráficos:


Histogram of A
Density
0.0
0.2
0.4
0.6
0.8
3.5
4.0
4.5
5.0
5.5
6.0
6.5
A


Histogram of B
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
2
3
4
5
B

¿Ajustan los datos de las variables A y B a una distribución gaussiana? Justifique su respuesta aplicando algún test de normalidad (explíquelo):

Test de Shapiro Wilks:

Sirve para estimar si una variable tiene una distribución normal o no.

Plantea la hipótesis nula que una muestra proviene de una distribución normal. Elegimos un nivel de significancia y tenemos una hipótesis alternativa que sostiene que la distribución no es normal.

$$H_0: X \sim N(\mu, \vartheta^2)$$

$$H_1: X \nsim N(\mu, \vartheta^2)$$

El test Shapiro-Wilks intenta rechazar la hipotesis nula a nuestro nivel de significanza.

Seleccionando como nivel de significanza de 0.15 y realizando el test en R:

```
> shapiro.test(A)

        Shapiro-Wilk normality test

data:  A
W = 0.99749, p-value = 0.1272
```

```
> shapiro.test(B)

        Shapiro-Wilk normality test

data:  B
W = 0.90603, p-value < 2.2e-16
```

Se puede ver que se obtiene como nivel de significanza, 0.1272 para A y un número bastante inferior a 0.15 para B, por lo tanto:

Se puede determinar que se rechaza la hipótesis nula y que se ajustan los datos a una distribución gaussiana.

Ejercicio 6)

Los siguientes comandos permiten estudiar una Chi cuadrado con cierta cantidad(K) de grados de libertad:

A<-matrix(0,1000,5)
B<-matrix(0,1000,1)
for(i in 1:1000) for(j in 1:5) {A[i,j]=rnorm(1,mean=5,sd=1)}
for(i in 1:1000) {B[i]=4*sd(A[i,])^2}
hist(B,freq=FALSE)

Utilice el commando curve y dchisq dados a continuación para obtener x de forma experimental:

```
A<-matrix(0,1000,5)
B<-matrix(0,1000,1)
for(i in 1:1000) for(j in 1:5) {A[i,j]=rnorm(1,mean=5,sd=1)}
for(i in 1:1000) {B[i]=4*sd(A[i,])^2}
hist(B,freq=FALSE)


curve(dchisq(x,4),0.20,add=T)
```

Para K=4, se puede ver que se obtiene una grafica similar a la indicada en la letra.