Probabilidad y Estadística Matemática

Morfín Chávez Guillermo

1.

Escriba un programa que genere k números aleatorios con distribución uniforme en el intervalo $(0,1)$, y que calcule el promedio muestral y la varianza muestral. Verifique, ejecutando varias veces el programa, que aunque en cada simulación se generan distintos números aleatorios (cuide que cada vez que se corra, que varié la semilla del generador de números aleatorios), que siempre sucederá que la media muestral se acerca a $1/2$ y que la varianza muestral se acerca a $1/12$. Ejecute el programa para diversos valores ascendentes de k, y observe lo que ocurre a medida que estos aumentan. Comente.

```
#Sea
n=100
x=runif(n)
mean(x)
#Otenemos
[1] 0.542648
var(x)
#Obtenemos
[1] 0.09711465
#Para valores ascendentes de k
prom=c()
con=1
for(k in c(10,100,1000,10000,100000))
{
  x=runif(k)
  prom[con]=mean(x)
  con=con+1
}
prom
# Uno de los resultados al ejecutar
[1] 0.5981123 0.4198007 0.5216803 0.4966859 0.4999518
variance=c()
con=1
for(k in c(10,100,1000,10000,100000))
{
  x=runif(k)
  variance[con]=var(x)
  con=con+1
}
variance
# Uno de los resultados al ejecutar
[2] 0.06024582 0.07074343 0.08573964 0.08328064 0.08363378
```

Debe notarse que al ejecutar para valores ascendentes de k el valor se aproxima cada vez más al valor exacto de la Esperanza y de la Varianza (1/2 para la primera y 1/12 para la segunda); observese [1] y [2] en el código. De donde se verifica la ley de los grandes números, i.e. la convergencia cuando $k \to \infty$.

2.

Mediante simulación (Genere números aleatorios, haga el histograma, etc.) pruebe y comente, que:

2.1. Si $Z \sim N(0,1)$ y $Y \sim \chi^2_5$, entonces $\frac{Z}{\sqrt{Y/5}} \sim t_5$

```
 t_5=c()
for(i in 1:1000)
{
  t_5[i]=rnorm(1)/sqrt(rchisq(1,5)/5)
}
hist(t_5, freq = FALSE,  breaks = 50, xlab = "Distribucion",
xlab = "Densidad", ylim = c(0,0.5), main = "Demostracion_por_simulacion")
curve(dt(x,5), add = TRUE)
```

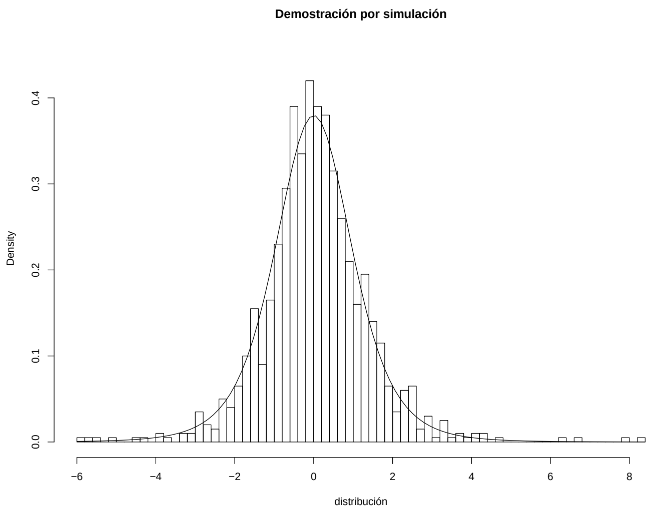

Figura 1: Histograma para la variable aleatoria y curva de t_5

Al generar el histograma se puede observar una gran similitud con la distribución para una variable aleatoria t-Student de parámetro 5. Utilizando el argumento "breaks"podemos generar más clases para visualizar el resultado más detalladamente.

2.2. Si $X_1, ..., X_{50} \overset{ind}{\sim} N(\mu = 2, \sigma = 2)$, entonces $\frac{\sqrt{50}(\overline{X}_{50}-2)}{\sqrt{S^2_{50-1}}} \sim t_{49}$

```
 t_49=c()
for(i in 1:1000)
{
  muestra <- rnorm(50,2,2)
  var <- var(muestra)
  mean <- mean(muestra)
  t_49[i] <- ((sqrt(50))*(mean-2))/(sqrt(var))
}
hist(t_49,prob=T, , xlab = "distribucion",
main = "Demostracion␣por␣simulacion_para_t_49")
curve(dt(x,49),add=T)
```

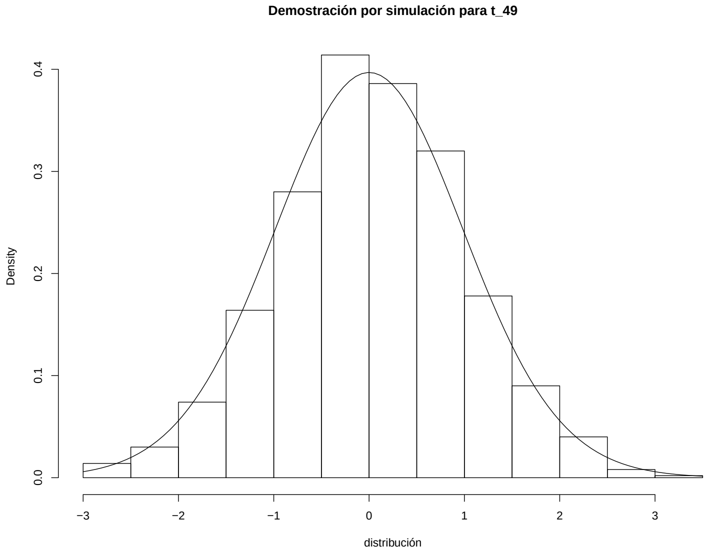

Figura 2: Histograma para la variable aleatoria y curva de t_49

Notese que al iterar un número considerable de veces el histograma tienda a tener la forma de la distribución t-Student de parámetro 49. Razonamos que las variables aleatorias se distribuyen de la manera esperada.

2.3. Si $X \sim \chi^2_2$ y $Y \sim \chi^2_4$, entonces $\frac{X/2}{Y/4} \sim F_{2,4}$

```
distr=c()
for(i in 1:1000)
{
  distr[i]=(rchisq(1,2)/2)/(rchisq(1,4)/4)
}
hist(distr, freq = FALSE, breaks = 50, xlab = "Distribucion",
ylim = c(0,0.62), main = "Demostracion_por_simulacion")
curve(df(x,4,2), add = TRUE)
```

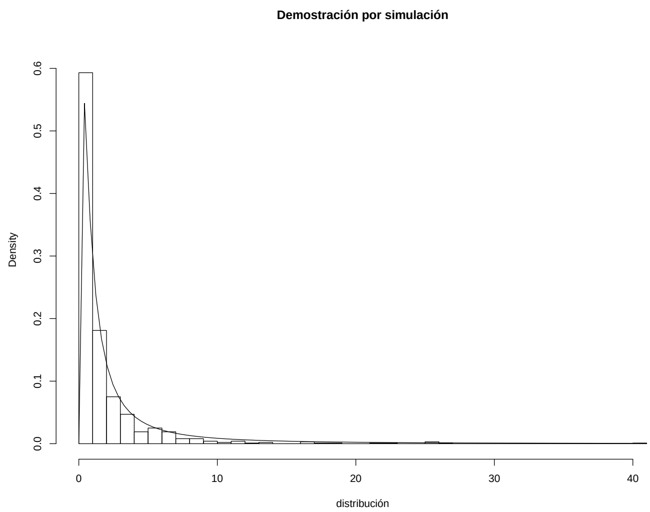

Figura 3: Histograma para la variable aleatoria y curva de $F_{2,4}$

Al visualizar la correlación entre el histograma y la curva de la distribución Fisher de 2 y 4 grados de libertad, es posible percatarse que las variables aleatorias que generan el histograma se distribuyen como $F_{2,4}$.

2.4. Sean $X_1, X_2, ...X_{20}$ y $Y_1, Y_2, ..., Y_{30}$ dos muestras aleatorias independientes de dos poblaciones normales, con 20 observaciones de la población 1 y 30 observaciones de la población 2, si las varianzas de las poblaciones 1 y 2 son σ_1^2 y σ_2^2, respectivamente, entonces:

$$Q = \frac{S^2_{20-1}\sigma_2^2}{S^2_{30-1}\sigma_1^2}$$

tiene una distribución F con 19 y 29 grados de libertad

```
 Q=c()
for(i in 1:1000)
{
  muestra_1 <- rnorm(20,0,2)
  muestra_2 <- rnorm(30,2,4)
  var_1 <- var(muestra_1)
  var_2 <- var(muestra_2)
  Q[i] <- ((var_1)*(4^2))/((var_2)*(2^2))
}
hist(Q,prob=T, ylim = c(0,1),xlab = "Distribuci n",
main = "Demostracion␣por␣simulacion_para_F,19,29")
curve(df(x,19,29),add=T)
```

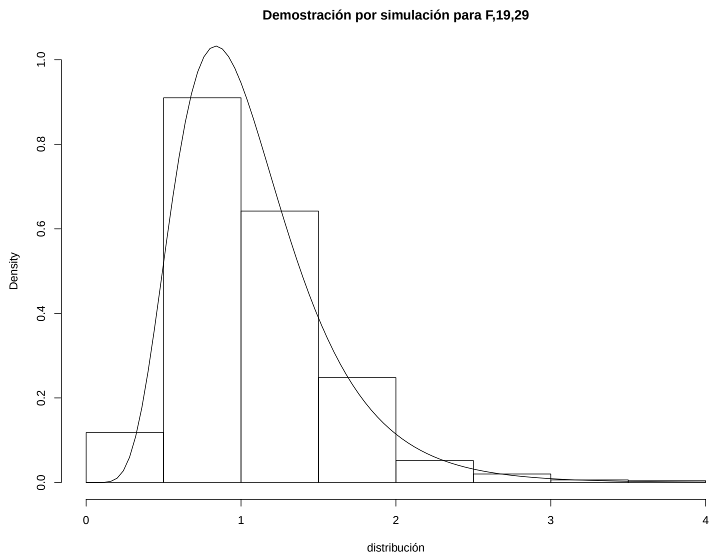

Figura 4: Histograma para la variable aleatoria y curva de $F_{19,29}$

Al igual que en los casos anteriores la simulación y el histograma nos permiten mostrar como al iterar un número considerable de veces la variable aleatoria en cuestión, podemos observar como esta se ajusta (al observar el histograma) a una distribución Fisher de 19 y 29 grados de libertad.

3.

En el siguiente ejercicio realice los cálculos a mano y verifíquelos usando R. Considere el conjunto de datos:
0 0 0 1 0 1 0 2 1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 1 1 1 0 2 0 1 0 0 0 0 0 0 0 0 1 1 0 2 1 0 0 0
0 0 3 1 0 0 1 1 0 3 0 1 1 1 2 0 0 0 1 1 0 1 2 0 0 0 0 0 0 1 0 2 1 1 0 1 0 0 1 0 0 0 0 0 0 0 1

3.1. Calcule el promedio y la mediana.

```
  datos<- scan("datos.txt") #Aqui se guarda la lista de datos
  mean(datos)
  > 0.46
 table(datos) #Esto es para observar la frecuencia de los valores
 > datos
```

```
 0  1  2  3
64 28  6  2
  median(datos)
 > 0
```

La suma de todos los datos es

$$\sum_{datos} = 46,$$

por otro lado tenemos que hay $n = 100$ datos. Así la media es $\frac{46}{100} = 0.46$.
Es posible observar que el 0 es la mediana, dado que los dos valores que se encuentran justo a la mitad de los datos ordenados son 0.

3.2. Calcular la varianza muestral y desviación estándar muestral.

```
var(datos)
> 0.4933333
sqrt(var(datos))
> 0.7023769
```

Tenemos que la varianza muestral está dada por

$$s^2 = \frac{1}{99}\sum_{i=1}^{n}(x_i - \bar{x})$$

$$\frac{64}{99}(-0.46)^2 + \frac{28}{99}(0.54)^2 + \frac{6}{99}(1.54)^2 + \frac{2}{99}(2.54)^2 = 0.4933333$$

Así, para la desviación estándar muestral

$$s = \sqrt{0.4933333} = 0.7023769$$

3.3. ¿Que distribución podría proponer para las variables aleatorias de las cuales se generaron los datos?

Proponemos la siguiente distribución hipergeométrica de parámetros $m = 9$, $n = 91$, $k = 5$ y simulamos 100 valores para comparar con las variables aleatorias dadas.

```
n=100
rhyper(n,9,91,5)
show(x)
  >0 0 0 0 1 0 0 0 1 2 1 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 1 0 1 0 0 2 0 0 0
  0 1 1 2 1 0 1 1 0 0 0 0 0 0 0 2 1 1 1 0 1 0 0 0 0 0 0 1 0 0 1 1 0 1 1 0
  0 1 2 1 0 1 0 1 0 0 0 0 0 1 1 3 0 1 0 0 0 0 1 0 1 0 1 0 0
table(x)#Para comparar con el conjunto de datos dado
>x
 0  1  2  3
60 34  5  1
```

3.4. Construya un histograma. Utilice $k = \sqrt{100} = 10$ clases. Llene el cuadro y grafique

Hicimos uso del paquete 'fdth' (Frecuancy Distribution Table, histograms and Poligons) para generar el histograma y la tabla de frecuencias.

```
dist <- fdt(datos, k=10)
dist #para visualizar la tabla de distribucion de frecuencia completa
plot(dist, type="rfh") #Para generar el histograma y comparar en 3.5
```

Intervalo	Frecuencia Relativa
[0,0.3)	0.64
[0.3,0.61)	0.00
[0.61,0.91)	0.00
[0.91,1.2)	0.28
[1.2,1.5)	0.00
[1.5,1.8)	0.00
[1.8,2.1)	0.06
[2.1,2.4)	0.00
[2.4,2.7)	0.00
[2.7,3]	0.02

Cuadro 1: Tabla de distribución de frecuencia relativa.

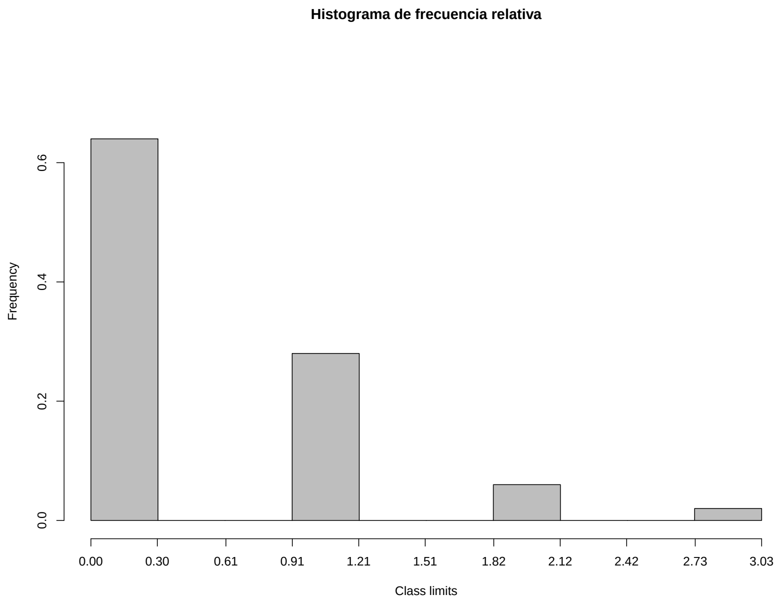

Figura 5: Histogramas de las variables aleatorias

3.5. Haga una figura de la distribución del inciso 3.3. y compare con el histograma del inciso 3.4.

```
 #Para generar la comparacion
p=dhyper (0:3 ,9 ,91 ,5)
plot (0:3,p,type="h",xlim=c(0,3),ylim=c(0,0.7),
col="blue",ylab="probabilidad", xlab = "Distribucion")
par(mfrow=c(1,2))
plot(dist , type="rfh")
```

En general podemos hacer la comparación porque hemos hecho una indución a partir de las medias de tendencia y dispersión de los datos, además, con base en el conocimiento de las distintas distribuciones, podemos interpretar que distribución se ajusta a los datos de la mejor manera.

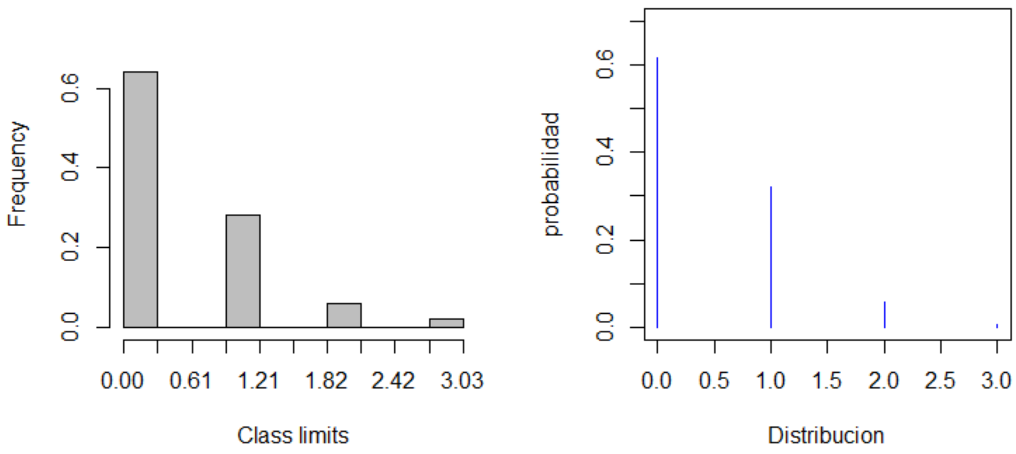

Figura 6: Histogramas de las variables aleatorias y la distribución de probabilidad

En la figura 6 tenemos en el lado izquierdo el histograma para las frecuencias del conjunto de datos, en el lado derecho podemos observar el histograma de la distribución de variables aleatorias que se indica en el inciso 3.3. Nótese que comparando ambas figuras resulta que el ajuste de la distribución hipergeométrica genera variables aleatorias muy similares a los datos que tenemos inicialmente, así, mediante la observación se ha mostrado que los datos se ajustan a la distribución hipergeométrica con parámetros (9,91,5).