Probability and Stadistics 1

Sonia Castro

12/1/2021

```r
setwd("/Users/sonia/Desktop")
library(MASS)
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```r
library(EnvStats)
```

```
##
## Attaching package: 'EnvStats'

## The following object is masked from 'package:MASS':
##
##     boxcox

## The following objects are masked from 'package:stats':
##
##     predict, predict.lm

## The following object is masked from 'package:base':
##
##     print.default
```

```r
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:EnvStats':
##
##     qqPlot
```

```r
library(tables)
library(RcmdrMisc)
```

```
## Loading required package: sandwich
```

```
library(multcomp)
```

```
## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

Exercici 1

Genereu 100 dades d'una distribució Normal amb paràmetres $\mu_1 = 2,3$ i $\sigma_1^2 = 0.08$. A aquesta mostra l'anomenarem *mostra 1*

```
set.seed(1)
mostra1 <- rnorm(100, 2.3, sqrt(0.08))
```

- **Estimeu l'esperança de la distribuó puntualment i per mitjà d'un interval de confiança. Feu-ho a mà i amb l'R.**

Estimem l'esperança sabent que el seu estimador puntual és la mitjana i per tant:

$$\hat{\mu} = \bar{x} = \frac{1}{N}\sum_{i=0}^{100} x_i$$

que en el nostre cas dona:

```
mean <-0
for (i in 1:100){
  mean <- mean + mostra1[i]
}
mean<-mean/100
mean
```

```
## [1] 2.330798
```

Ara ho fem amb la comanda "mean()" de R:

```
mean(mostra1)
```

```
## [1] 2.330798
```

Estimem l'esperança per mitjà d'in interval de confiança. Sabem que $\frac{\bar{x}-\mu}{S/\sqrt{n}} \sim t-Student_{n-1}$ i volem fer un interval de confiança de $1-\alpha$. Per tant $P(t_{-\alpha/2} \leq \frac{\bar{x}-\mu}{S/\sqrt{n}} \leq t_{\alpha/2}) = 1-\alpha$ i aïllant trobem $P(\bar{x}-t_{\alpha/2}S/\sqrt{n} \leq \mu \leq \bar{x}+t_{\alpha/2}S/\sqrt{n}) = 1-\alpha$

L'interval de confiança de μ és:

$$\bar{x} \pm t_{\alpha/2}S/\sqrt{n}$$

Ho càlculem amb $\alpha = 0.05$:

```
n=100

mean(mostra1)+ qt(0.025,n-1) * (sd(mostra1)/sqrt(n))
```

```
## [1] 2.280389
```

```
mean(mostra1)- qt(0.025,n-1) * (sd(mostra1)/sqrt(n))
```

```
## [1] 2.381207
```

Y ara amb R:

```
t.test(mostra1, mu = 2.3)
```

```
##
##  One Sample t-test
##
## data:  mostra1
## t = 1.2123, df = 99, p-value = 0.2283
## alternative hypothesis: true mean is not equal to 2.3
## 95 percent confidence interval:
##  2.280389 2.381207
## sample estimates:
## mean of x
##  2.330798
```

- **Estimeu la variància de la distribució puntualment i per mitjà d'un interval de confiança. Feu-ho a mà i amb l'R. L'estimador de la variància surt més proper del límit inferior o del superior de l'interval?. Justifiqueu perquè.**

Estimem la variància puntualment sabent que $\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_x-\bar{x})^2$

```
n=100
s = 0
for(i in 1:100){
  s = s + (mostra1[i] - mean(mostra1))**2
}

v <- 1/(n-1) * s
v
```

```
## [1] 0.06454097
```

I amb la comanda "var()" de R:

```
var(mostra1)
```

```
## [1] 0.06454097
```

Estimem per mitjà d'un intèrval de confiança. Sabem que $(n-1)S^2/\sigma^2 \sim X^2_{n-1}$ i com volem un interval amb confiança $1-\alpha$, $P(X^2_{\alpha/2,n-1} \leq (n-1)S^2/\sigma^2 \leq X^2_{1-\alpha/2,n-1}) = 1-\alpha$ i aïllant obtenim $P(\frac{(n-1)S^2}{X^2_{1-\alpha/2,n-1}} \geq \sigma^2 \geq \frac{(n-1)S^2}{X^2_{\alpha/2,n-1}}) = 1-\alpha$. Per tant el IC de σ^2 és:

$$\Big[\frac{(n-1)S^2}{X^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{X^2_{1-\alpha/2,n-1}}\Big]$$

Ho calculem:

```
((n-1)*var(mostra1))/qchisq(0.975,n-1,lower.tail=TRUE)
```

```
## [1] 0.04975437
```

```
((n-1)*var(mostra1))/qchisq(0.025,n-1,lower.tail=TRUE)
```

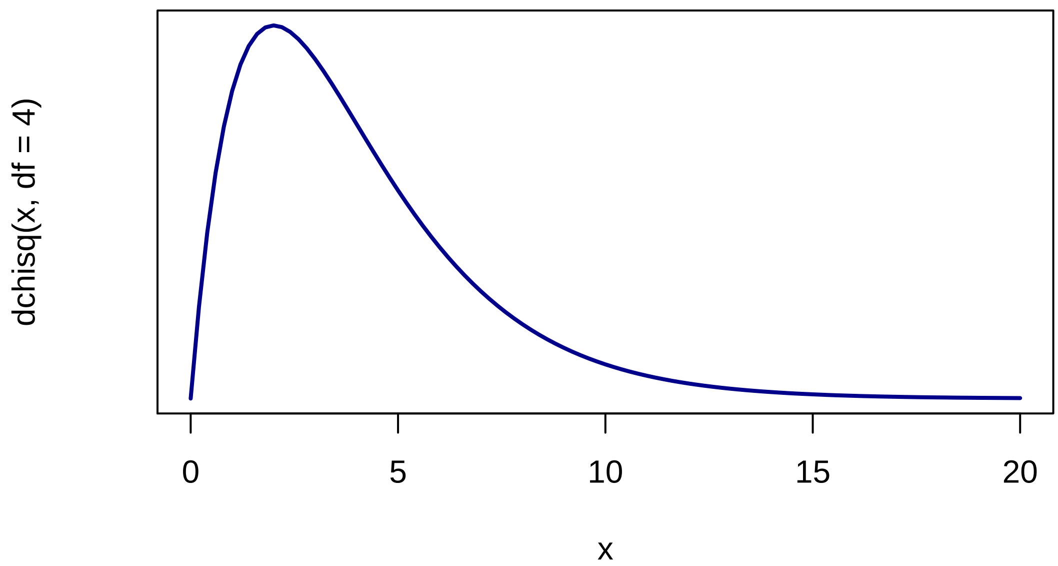
```
## [1] 0.08709735
```

I amb la comanda R:

```
varTest(mostra1, sigma.squared = 0.08, alternative = "two.sided")
```

```
##
##  Chi-Squared Test on Variance
##
## data:  mostra1
## Chi-Squared = 79.869, df = 99, p-value = 0.1584
## alternative hypothesis: true variance is not equal to 0.08
## 95 percent confidence interval:
##  0.04975437 0.08709735
## sample estimates:
##   variance
## 0.06454097
```

```
layout(matrix(1), widths = lcm(15), heights = lcm(10))
curve(dchisq(x, df=4), xlim=c(0,20),
      col="darkblue", lwd=2, yaxt="n")
```

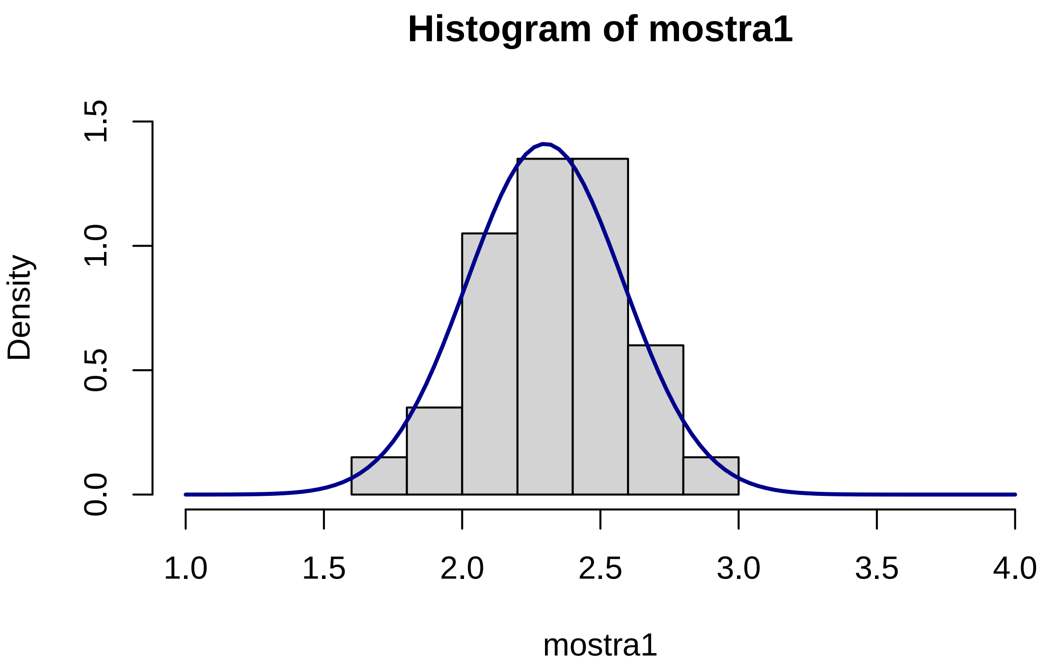


L'estimador de la variància queda més proper al límit inferior, perquè la distribució chi quadrat no és simètrica i si observes la seva funció de densitat acumula més probabilitat a la meitat esquerra que a la meitat dreta.

Per entendre-ho millor per exemple observem la funció densitat d'una chi quadrat de df=4, si crees un CI desde el quantil d'ordre 0.025 fins al quantil d'ordre 0.975, hi ha molta més probabilitat acumulada a la part esquerra del CI que a la dreta i per tant més probabilitat que el valor estimat estigui més aprop del límit esquerre.

- **Feu l'histograma associat a les dades i mostreu-lo juntament amb el gràfic la densitat d'una Normal.**

```
layout(matrix(1), widths = lcm(15), heights = lcm(10))
hist(mostra1,prob = TRUE,ylim=c(0, 1.5), xlim=c(1,4))
curve(dnorm(x, mean = 2.3, sd=sqrt(0.08)),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")
```



- **Porteu a terme el test d'hipòtesi $H_0 : \mu = 2.3$ vs $H_1 : \mu \neq 2.3$. Què concloïu?**

Portem a terme el test d'hipòtesi amb la comanda R "t.test".

```
t.test(mostra1, mu=2.3)
```

```
##
##  One Sample t-test
##
## data:  mostra1
## t = 1.2123, df = 99, p-value = 0.2283
## alternative hypothesis: true mean is not equal to 2.3
## 95 percent confidence interval:
##  2.280389 2.381207
## sample estimates:
## mean of x
##  2.330798
```

Com que el p-value $> \alpha = 0.05$, no rebutgem la hipòtesi nula. Concluim que les dades tenen prou força com per no rebutjar que el valor esperat és 2.3.

- **Porteu a terme el test d'hipòtesi $H_0 : \mu = 2.5$ vs $H_1 : \mu < 2.5$. Què concloïu?**

```
t.test(mostra1, mu=2.5, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  mostra1
## t = -6.6602, df = 99, p-value = 7.737e-10
## alternative hypothesis: true mean is less than 2.5
## 95 percent confidence interval:
##     -Inf 2.37298
## sample estimates:
## mean of x
##  2.330798
```

Com que el p-value $< \alpha = 0.05$,rebutgem la hipòtesi nula $\mu = 2.5$. Concluim que les dades tenen prou força com per a dir que el valor esperat és menor que 2.5.

- **Definiu una variable de la qual la mostra anterior en pugui ser representativa. Expliqueu amb un paràgraf les conclusions de l'anàlisi realitzat en base a la variable que heu definit.**

La variable és el sou mitjà d'un més de persones de 30 a 50 anys a un barri d'Itàlia en milers de euros, amb un valor esperat de 2,33 (2330 euros). Amb una variància petita, ja que la majoria es troben aproximadament entre 1.400 y 3300.

Després de l'anàlisis realitzat concluim que les dades tenen prou força com per no rebutjar que el valor esperat del sou de persones de 30 a 50 anys en aquest barri és estadísticament igual a 2300 euros. I per rebutjar que el valor esperat del sou sigui major o igual a 2500 euros.

Exercici 2

Genereu 200 dades d'una Normal amb espeprança $\mu = 1.9$ i $\sigma_2^2 = 0.12$. A aquesta mostra l'anomenarem *mostra 2*.

```
mostra2 <- rnorm(200, 1.9, sqrt(0.12))
```

- **Calculeu analíticament l'expressió general d'un interval de confiança (IC) per a la diferència dels valors esperats de dues poblacions Normals (s'han de veure tots els passos).**

Anomenarem $X \sim N(\mu_x, \sigma)$ a la distribució normal de la mostra1 i $Y \sim N(\mu_x, \sigma)$ a la distribució normal de la mostra2 (independents). Suposarem variàncies desconegudes iguals. Com que hem d'estimar la diferència dels valors esperats, escollim el paràmetre $\theta = \mu_x - \mu_y$. El seu estimador és $\hat{\theta} = \bar{X} - \bar{Y}$ que és no esbiaixat, ja que $E(\hat{\theta}) = E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_x - \mu_y = \theta$.

Tenim, $\bar{X} \sim N(\mu_x, \sigma/n_1)$ i $\bar{Y} \sim N(\mu_y, \sigma/n_2)$. Com que la variància és la mateixa l'estimem tenint en compte totes les dades amb

$$Sp^2 = \frac{(n-1)S_x^2 + (m-1)S_Y^2}{n_1 + n_2 - 2} = \frac{\sum_1^{n_1}(X_i - \bar{X})^2 + \sum_1^{n_2}(Y_i - \bar{Y})^2}{n_1 + n_2 - 2}$$

Sabem que $\frac{X_i - \bar{X}}{\sigma} \sim N(0,1)$ per a tota i i que $\sum_1^m \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{m-1}$.

Per tant $\frac{\sum_1^{n_1}(X_i - \bar{X})^2 + \sum_1^{n_2}(Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$, és a dir, $\frac{(n_1+n_2-2)Sp^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$. (1)

Com que $Var(\bar{X}-\bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \sigma^2/n_1 + \sigma^2/n_2$, tenim $\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{\sigma^2/n_1+\sigma^2/n_2}} \sim N(0,1)$. (2)

Sabem que si Z es distribueix com una N(0,1), i Y com una χ^2_n llavors $\frac{Z}{\sqrt{Y/n}} \sim t_n$. Per tant tenint en compte (1) i (2):

$\frac{\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{\sigma^2/n_1+\sigma^2/n_2}}}{\sqrt{\frac{(n_1+n_2-2)Sp^2}{\sigma^2(n_1+n_2-2)}}} \sim t_{n_1+n_2-2}$ i simplificant tenim que $\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{Sp\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$

Ara sabem que $P(-t_{\alpha/2,n_1+n_2-2} \leq \frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{Sp\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leq t_{\alpha/2,n_1+n_2-2}) = 1-\alpha$ i aïllant ens queda

$$(\bar{X}-\bar{Y}) - t_{\alpha/2,n_1+n_2-2}Sp\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq \mu_x - \mu_y \leq (\bar{X}-\bar{Y}) + t_{\alpha/2,n_1+n_2-2}Sp\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

És a dir el CI és:

$$(\bar{X}-\bar{Y}) \pm t_{\alpha/2,n_1+n_2-2}Sp\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

- **Calculeu a partir de els vostres dues mostres i del que heu calculat en l'apartat anteior, l'IC per a $\mu_1 - \mu_2$.**

```
sp <- sqrt((99 * (sd(mostra1)**2) + 199 *(sd(mostra2)**2))/298)


mean(mostra1)- mean(mostra2) + (qt(0.025, 298)* sp * sqrt(1/100 + 1/200))
```

```
## [1] 0.3557008
```

```
mean(mostra1)- mean(mostra2) - (qt(0.025, 298)* sp * sqrt(1/100 + 1/200))
```

```
## [1] 0.5087131
```

```
t.test(mostra1, mostra2,var.equal = TRUE )
```

```
##
##  Two Sample t-test
##
## data:  mostra1 and mostra2
## t = 11.118, df = 298, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3557008 0.5087131
## sample estimates:
## mean of x mean of y
##  2.330798  1.898591
```

- **En base a les vostres dues mostres, feu el test de comparació de variàncies per tal de veure si les variàncies son iguals o diferents estadísticament. Feu el test per valors crítics i per p-valors i comproveu que el resultat és el mateix.**

Proposem les següents hipòtesis:

$$\begin{cases} H_0: & \frac{\sigma_1^2}{\sigma_2^2} = 1 \\ H_1: & \frac{\sigma_1^2}{\sigma_2^2} \neq 1 \end{cases}$$

Primer de tot fem el test d'hipòtesi per valors crítics:

```
ratiosd<-(sd(mostra1))^2/(sd(mostra2))^2

ratiosd/qf(0.025,99,199)
```

```
## [1] 0.7730149
```

```
ratiosd/qf(1-0.025,99,199)
```

```
## [1] 0.3896731
```

Com podem observar, l'1 no es troba dins de l'interval definit. Per tant, podem rebutjar la hipòtesi nul · la, és a dir, les variàncies són diferents.

A continuació fem el test per p-valors:

```
pvalor<-2*(pf(ratiosd,99,199))
pvalor
```

```
## [1] 0.0008234913
```

Observem que el p-valor $< \alpha = 0.05$, per tant, rebutgem la hipòtesi nul · la
H_0.

Comprovem els nostres resultats.

```
var.test(mostra1,mostra2, alternative="two.side", conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  mostra1 and mostra2
## F = 0.5434, num df = 99, denom df = 199, p-value = 0.0008235
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3896731 0.7730149
## sample estimates:
## ratio of variances
##          0.5434005
```

- **En base als resultats de l'apartat anterior, compareu els valors esperats de les dues mostres. Quin resultat obteniu?**

En aquest cas l'estadístic és

$$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{s_x^2}{n_1}+\frac{S_y^2}{n_2}}} \sim N(0,1)$$

I el CI és:

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2}\sqrt{\frac{s_x^2}{n_1} + \frac{S_y^2}{n_2}}$$

```r
mean(mostra1) - mean(mostra2) + (qnorm(0.025)*sqrt(var(mostra1)/100 + var(mostra2)/200))
```

```
## [1] 0.3632098
```

```r
mean(mostra1) - mean(mostra2) - (qnorm(0.025)*sqrt(var(mostra1)/100 + var(mostra2)/200))
```

```
## [1] 0.5012041
```

```r
t.test(mostra1, mostra2, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  mostra1 and mostra2
## t = 12.277, df = 256.83, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.3628831 0.5015308
## sample estimates:
## mean of x mean of y
##  2.330798  1.898591
```

Com que el p-value $\leq \alpha = 0.05$,rebutgem la hipòtesi nula $\mu_x = \mu_y$. Concluim que les dades tenen prou força com per a rebutjar que el valor esperat de les dues mostres és igual.

- **Definiu una variable en la qual al mostra 2 en pugui ser representativa, i que tingui sentit ser comparada amb la variable de l'exercici anterior. En base a aquestes variables expliqueu mitjançant un paràgraf els resultats que heu obtingut.**

La variable és el sou mitjà en un mes de persones de 16 a 30 anys al mateix barri d'Itàlia que la mostra1 en milers d'euros, amb un valor esperat de 1898 euros (menor que en adults amb més edat). Això es podria donar perquè a que a aquestes edats la situació laboral es més variada donat que molts estudien i que els sous en general tendeixen a ser més baixos.

Després de l'anàlisi concluim que rebutgem que les variàncies siguin estadísticament iguals (en aquest cas els sous entre els joves varien mes) i que els valors esperats de sou en els dos grups d'edats no són estadísticament iguals.

Exercici 3

- **De la definició formal de les distribucions t-d'Student i Fisher, deduïu analíticament quina és la distribució del quadrat d'una variable amb distribució t-d'Student.**

Sabem que sigui $Z \sim N(0,1)$ i $Y \sim \chi^2_n$,

$$\frac{Z}{Y/\sqrt{n}} \sim t_n$$

També sabem que siguin $Z_i \sim N(0,1)$ independents,

$$Z_1^2 + ... + Z_n^2 \sim \chi^2_n$$

i que siguin $Y_1 \sim \chi^2_{n_1}$ i $Y_2 \sim \chi^2_{n_2}$ independents,

$$\frac{Y_1/n_1}{Y_2/n_2} \sim F_{n_1,n_2}$$

Sigui $Z \sim N(0,1)$ i $Y \sim \chi^2_n$, una t-d'Student al quadrat seria $\frac{Z^2}{\sqrt{Y/n}^2} = \frac{Z^2}{Y/n}$.

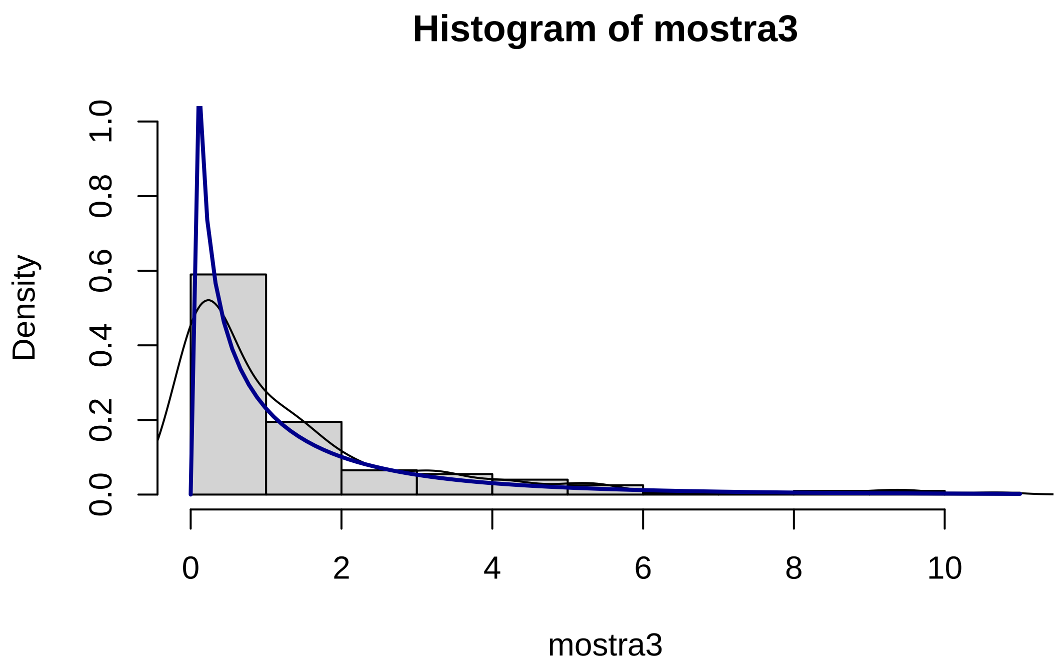
On $Z^2 \sim \chi^2_1$ i per tant tenint en compte la definició d'una distribució de Fisher:

$$\frac{Z^2/1}{Y/n} = \frac{Z^2}{Y/n} \sim F_{1,n}$$

- **Genereu 200 dades d'una t-d'Student amb 10 graus de llibertat. Calculeu el quadrat dels valors obtinguts i comproveu que la mostra resultant prové de la distribució que heu especificat en l'apartat anterior. Il · lustreu-ho gràficament.**

```
set.seed(2)
mostra3 <- rt(200, 10)
for (i in 1:200){
  mostra3[i] = mostra3[i] * mostra3[i]
}
```

```
layout(matrix(1), widths = lcm(15), heights = lcm(10))
hist(mostra3,prob = TRUE,ylim=c(0,1))
lines(density(mostra3), xlim=c(0.5,20))
curve(df(x,1,10), col="darkblue", lwd=2,add=TRUE)
```



Exercici 4

Per a les dades obtingudes en l'Exercici 1, i suposant que le valor esperat és desconegut, porteu a terme el test de Raó de Versemblança per tal de contrastar si les dades provenen del model Normal especificant en la hipòtesi nul · la o del model Normal especificat en l'alternativa:

$H_0 : \sigma^2 = 0.1$ vs $\sigma^2 desconeguda$

Com que no coneixem μ, l'estimem amb $\bar{X}$. Primer calculem l'estimador màxim versemblant de σ^2:

$$L(\sigma^2, \mu; x) = \prod_1^n f(x_i; \sigma^2, \mu) = \prod_1^n \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\bar{x})^2}{2\sigma^2}} = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{\frac{-\sum_1^n (x_i-\bar{x})^2}{2\sigma^2}}$$

Fem el logaritme:

$$l(\sigma^2, \mu; x) = log(\frac{1}{(\sqrt{2\pi}\sigma)^n} e^{\frac{-\sum_1^n (x_i-\bar{x})^2}{2\sigma^2}}) = \frac{-n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_1^n (x_i - \bar{X})^2$$

Derivem respecte σ^2 i ens queda $\frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_1^n (x_i - \bar{X})^2$ i al igualar amb 0 ens queda $\sigma^2 = \frac{\sum_1^n (x_i-\bar{X})^2}{n}$.

L'estadístic per aquest test és $-2log\left(\frac{max\hat{\theta}\in\theta_0 L(\hat{\theta},X)}{max\hat{\theta}\in\theta L(\hat{\theta},X)}\right) \sim \chi^2_{dif.paràmetres}$, que en el nostre cas és:

```
set.seed(1)

n<-100
s<- 0
```

```r
for(i in 1:100){
  s<- s +(mostra1[i] - mean(mostra1))**2
}


#estimador màxima versemblança variancia
estv <-s/n

#L fent servir el valor de la hipòtesi nula
L0 <- (1/((2*pi*0.1)**(n/2))) * (exp(-s/(2*0.1)))
#L fent servur l'estimador a partir de la nostra mostra
L <- (1/((2*pi*estv)**(n/2))) * (exp(-s/(2*estv)))

#creem l'estadistic de prova
quo <-L0/L
estad <- -2 * log(quo)
estad
```

```
## [1] 8.687592
```

```r
#rebutgem H0 quan l'estadístic és més gran que:
qchisq(0.95,1)
```

```
## [1] 3.841459
```

Com que l'estadístic és més gran que $\chi^2_{0.95,1}$, rebutgem la hipòtesi nula. Concluim que les dades tenen prou força com per a dir que la variància de la mostra és desconeguda.