

A2

Carlos A. García

March 30, 2019

Càrrega del fitxer

Carregueu l'arxiu de dades en R. Independentment del fitxer que vàreu obtenir en l'activitat 1, useu el fitxer "rawData_clean.csv". Un cop carregat el fitxer, valideu que els tipus de dades són els correctes. Si no és així, feu les conversions de tipus oportunes.

Primer establim el directori de feina

```
setwd("D:/Users/cagarcia/uoc/M2.954 - Estadística avançada/A2")
```

Llegim el fitxer proporcionat a la pràctica

```
satisfaccioLaboral <- read.csv2("rawData_clean.csv", header = TRUE, sep = ",", dec = ".")
attach(satisfaccioLaboral)
```

```
summary(satisfaccioLaboral)
```

```
##              city      sex  educ_level job_type  happiness
## Barcelona      :640    F:960    N:480      C :960    Min.    : 1.900
## Madrid          :640    M:960    P:480      PC:960    1st Qu.: 4.900
## Santiago de Compostela:640      S:480      Median : 5.900
##                               U:480      Mean   : 5.883
##                               3rd Qu.: 6.800
##                               Max.    :10.000
##      age      seniority      sick_leave      sick_leave_b
## Min.    :18.00  Min.    : 0.00  Min.    : 0.000  Min.    :0.0000
## 1st Qu.:33.00  1st Qu.:15.00  1st Qu.: 0.000  1st Qu.:0.0000
## Median :40.00  Median :20.00  Median : 0.000  Median :0.0000
## Mean   :40.33  Mean   :19.03  Mean   : 6.667  Mean   :0.2464
## 3rd Qu.:47.00  3rd Qu.:24.00  3rd Qu.: 0.000  3rd Qu.:0.0000
## Max.   :67.00  Max.   :35.00  Max.   :78.000  Max.   :1.0000
##  work_hours  Cho_initial  Cho_final
## Min.    :26.20  Min.    :0.95  Min.    :0.990
## 1st Qu.:35.50  1st Qu.:1.15  1st Qu.:1.250
## Median :38.20  Median :1.25  Median :1.350
## Mean   :38.29  Mean   :1.20  Mean   :1.351
## 3rd Qu.:40.90  3rd Qu.:1.25  3rd Qu.:1.450
## Max.   :88.00  Max.   :1.45  Max.   :1.680
```

Validem que els tipus de dades (classes) són correctes

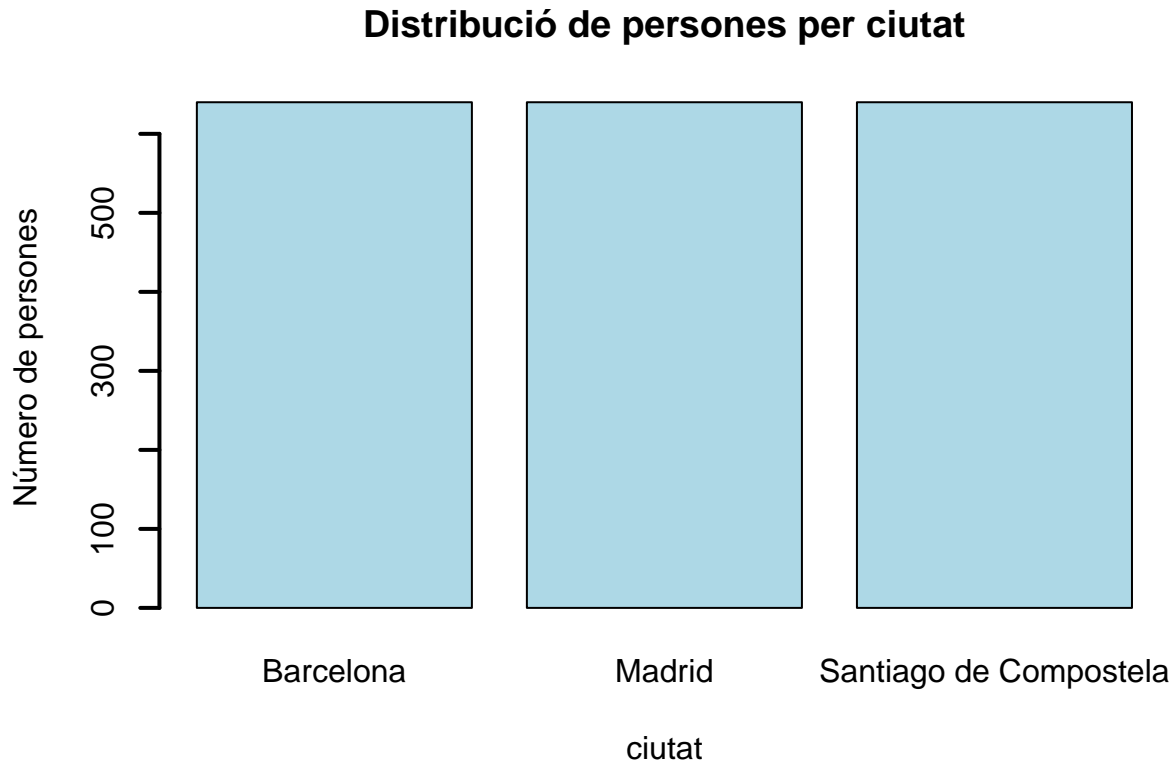
```
sapply(satisfaccioLaboral, class)
```

```
##      city      sex  educ_level  job_type  happiness
## "factor" "factor" "factor"    "factor" "numeric"
##      age  seniority  sick_leave  sick_leave_b  work_hours
## "integer" "integer" "integer"  "integer" "numeric"
## Cho_initial  Cho_final
```

```
##      "numeric"      "numeric"
```

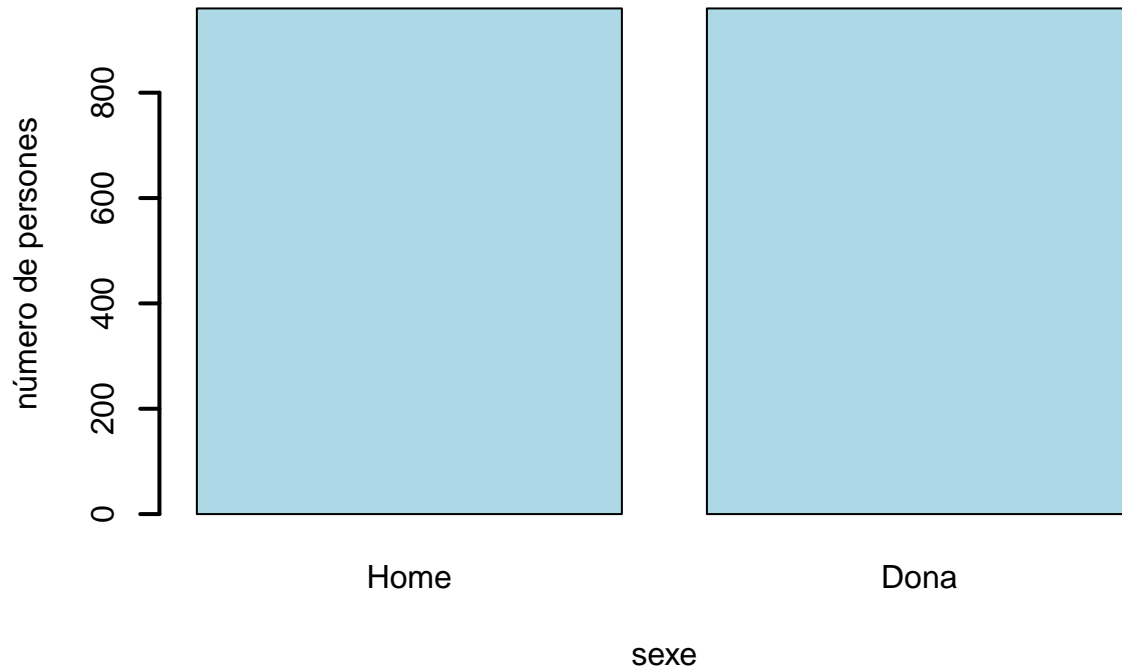
Gràfic de totes les variables:

```
plot(city, main="Distribució de persones per ciutat", xlab="ciutat",  
      ylab="Número de persones", col="#ADD8E6", lwd = 2)
```



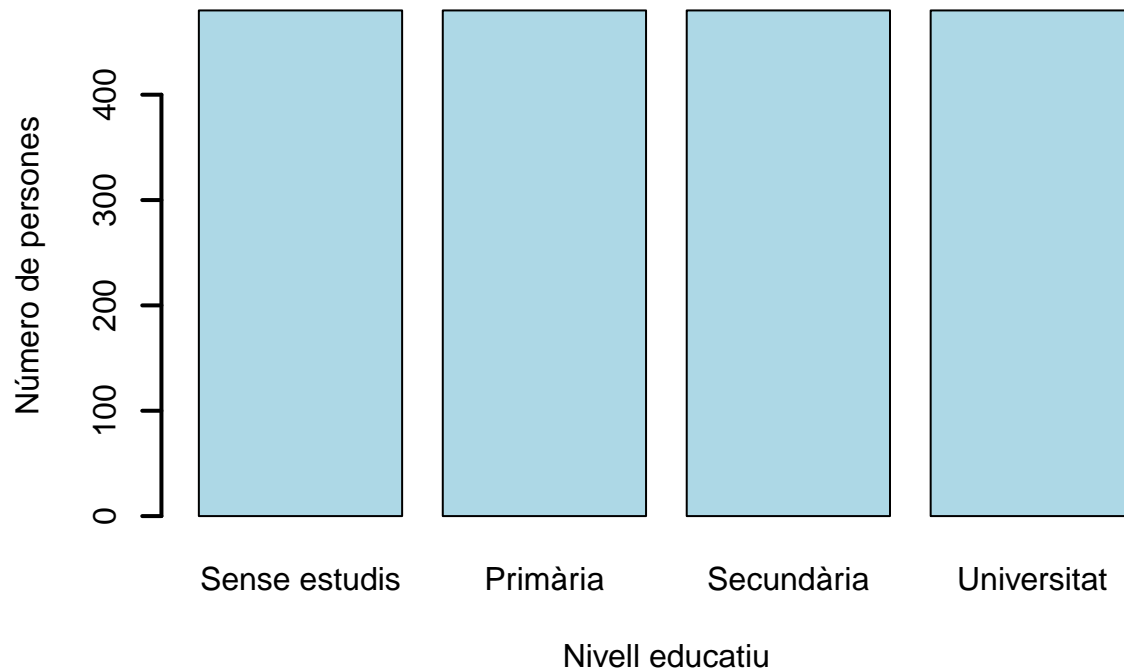
```
sexDesc <- factor(sex, levels=c("M","F"), labels=c("Home","Dona"))  
plot(sexDesc, main="Distribució de persones per sexe", xlab="sexe",  
      ylab="número de persones", col="#ADD8E6", lwd = 2)
```

Distribució de persones per sexe



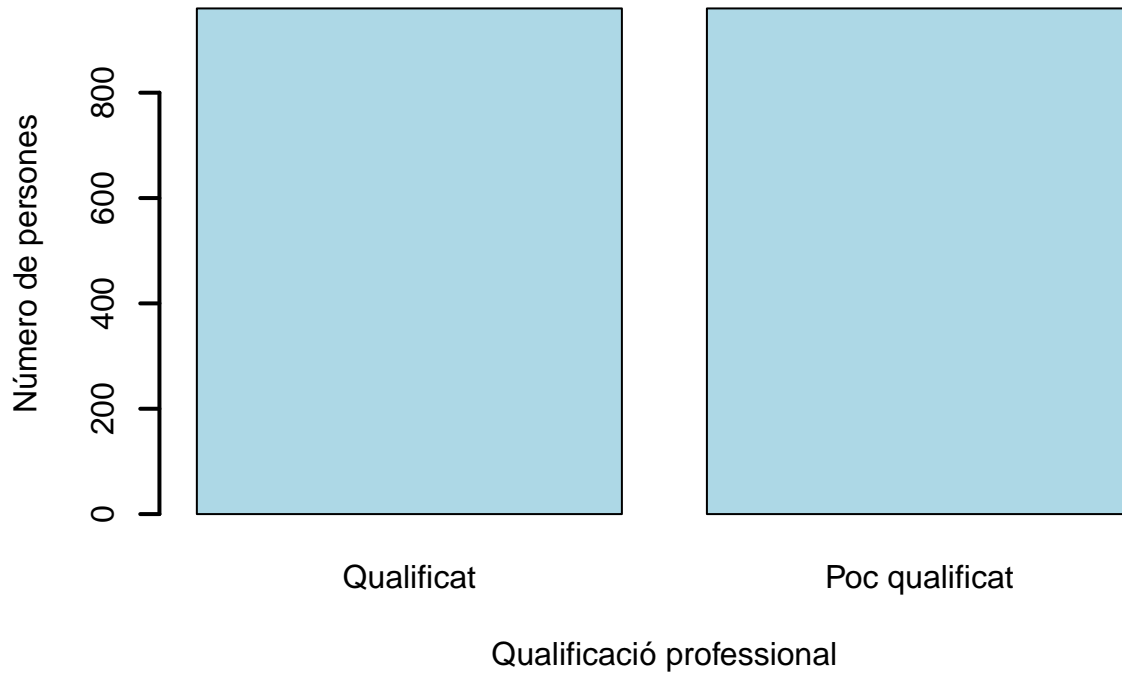
```
educDesc <- factor(educ_level, levels=c("N","P","S","U"),  
                  labels=c("Sense estudis", "Primària", "Secundària", "Universitat"))  
plot(educDesc, main="Distribució de persones per nivell d'educació",  
     xlab="Nivell educatiu", ylab="Número de persones", col="#ADD8E6", lwd = 2)
```

Distribució de persones per nivell d'educació



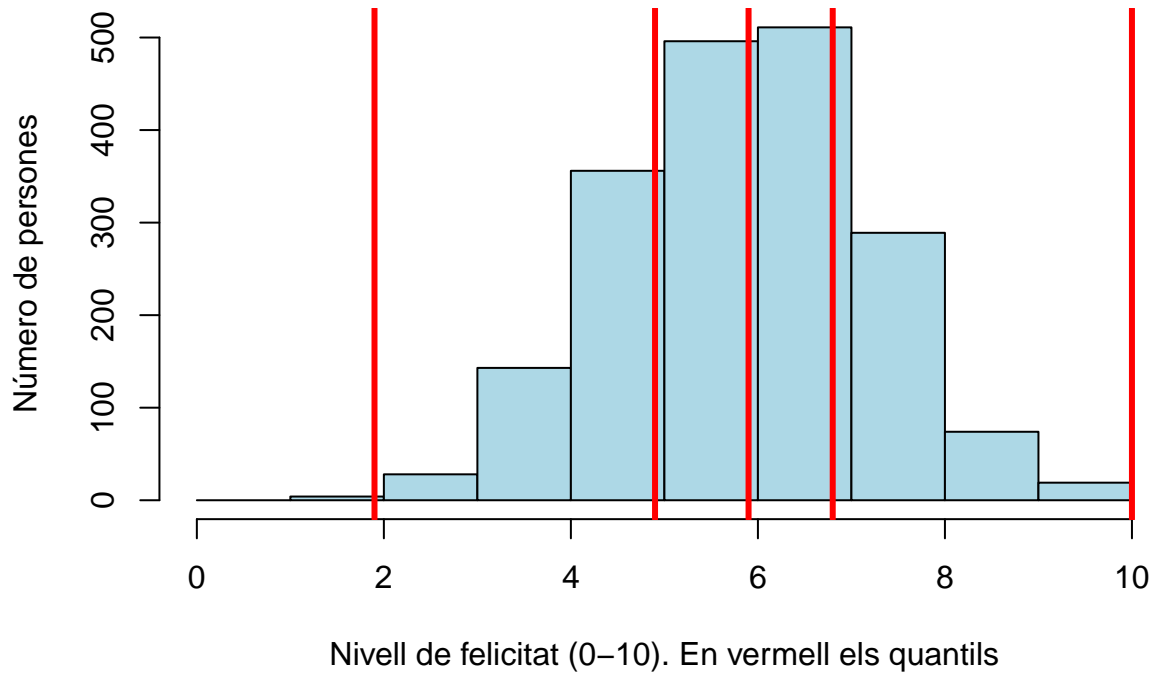
```
jobTypeDesc <- factor(job_type, levels=c("C","PC"), labels=c("Qualificat", "Poc qualificat"))  
plot(jobTypeDesc, main="Distribució de persones per qualificació professional",  
      xlab="Qualificació professional", ylab="Número de persones", col="#ADD8E6", lwd = 2)
```

Distribució de persones per qualificació professional

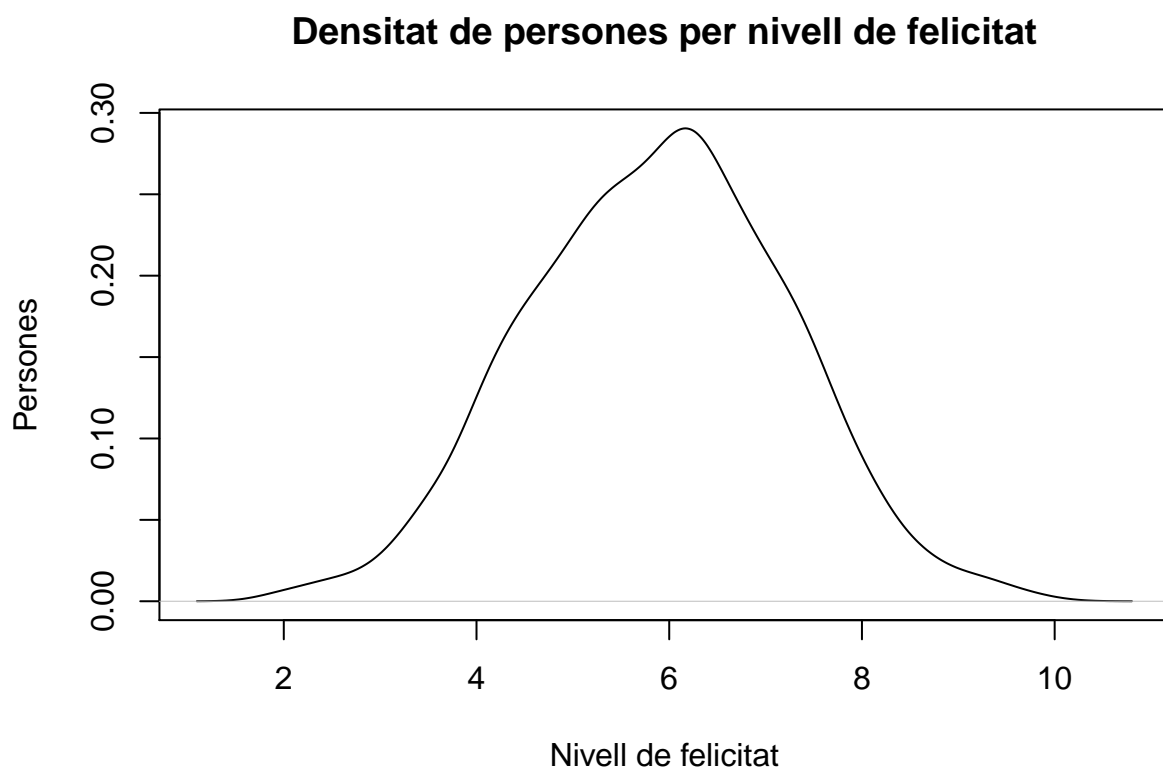


```
hist(happiness, breaks=c(0:10), main="Distribució de persones per nivell de felicitat",  
     xlab="Nivell de felicitat (0-10). En vermell els quantils",  
     ylab="Número de persones", col="#ADD8E6")  
abline(v=quantile(happiness), col="red", lwd=3)
```

Distribució de persones per nivell de felicitat

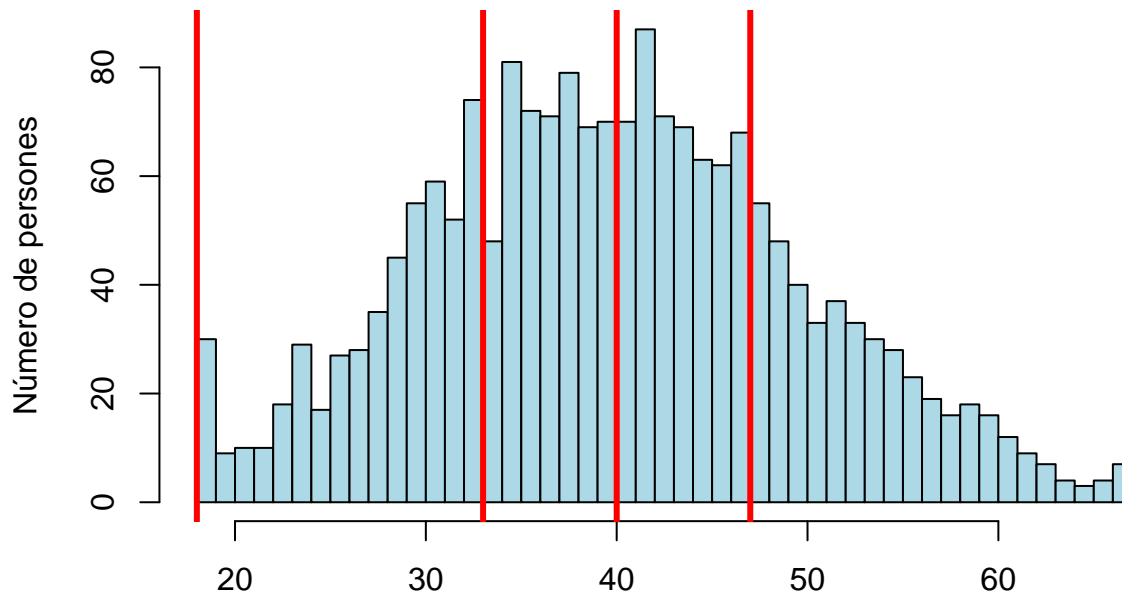


```
den.happiness <- density(happiness)
plot(den.happiness, col="black", main="Densitat de persones per nivell de felicitat",
     xlab="Nivell de felicitat", ylab="Persones")
```



```
hist(age, breaks=c(min(age):max(age)), main="Distribució de persones per edat",  
      xlab="Edat de les persones de la mostra. En vermell els quantils", ylab="Número de persones", col="red",  
      abline(v=quantile(age), col="red", lwd=3))
```

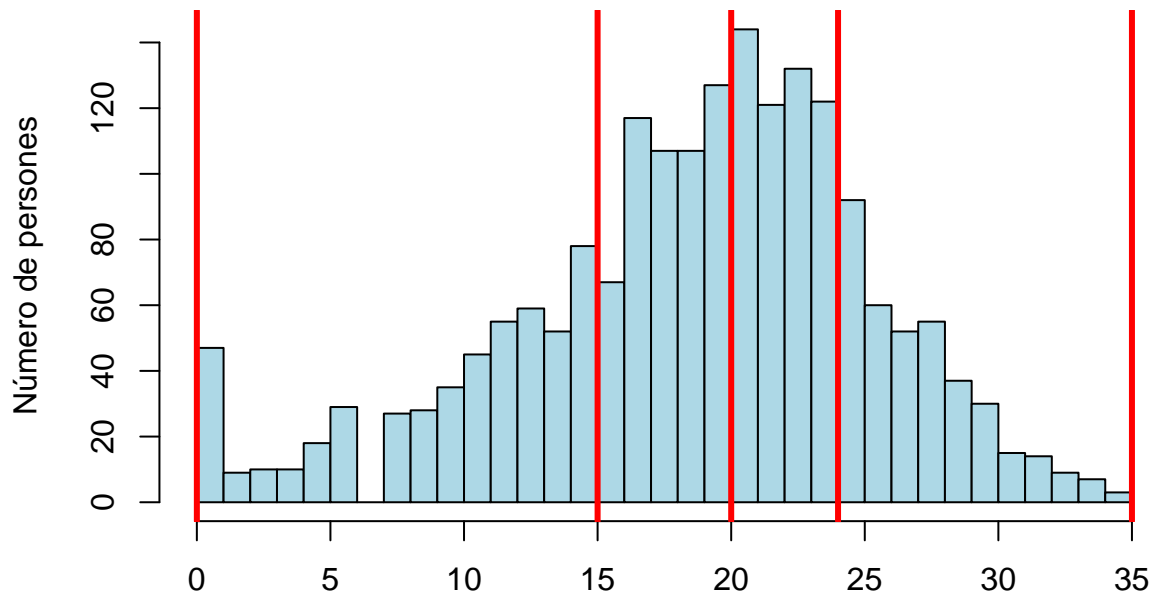
Distribució de persones per edat



Edat de les persones de la mostra. En vermell els quantils

```
hist(seniority, breaks=c(min(seniority):max(seniority)), main="Distribució de persones per anys d'exper",  
     xlab="Experiència de les persones de la mostra. En vermell els quantils", ylab="Número de persones",  
     abline(v=quantile(seniority), col="red", lwd=3))
```

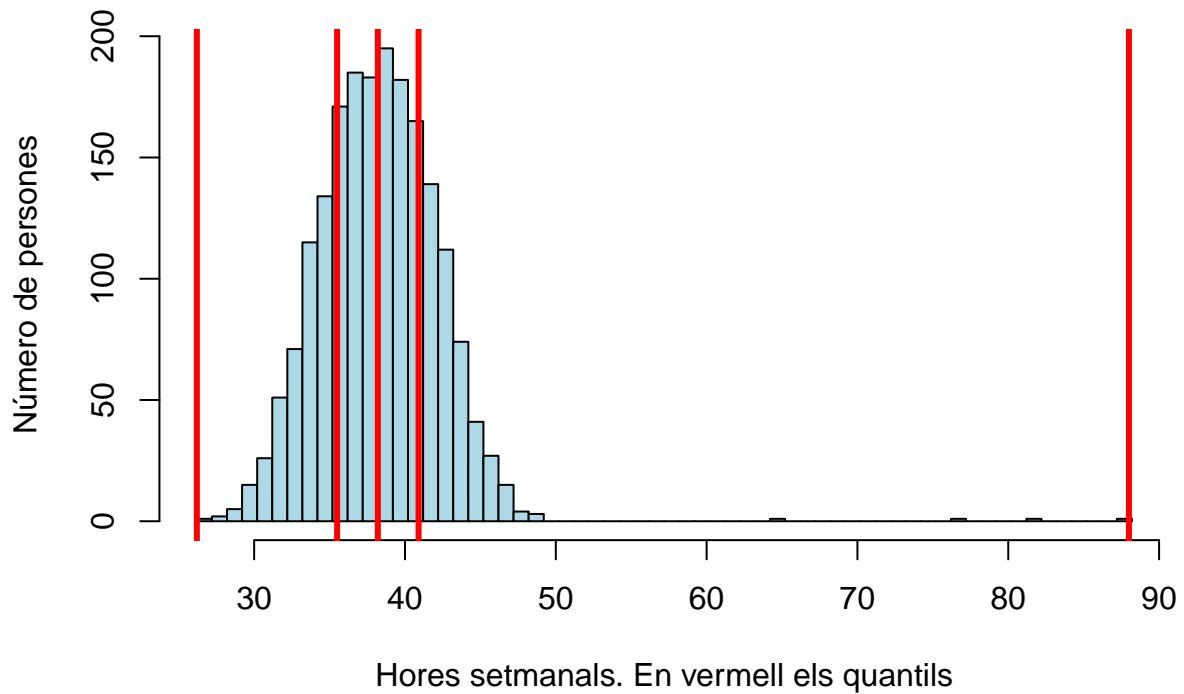

Distribució de persones per anys d'experiència



Experiència de les persones de la mostra. En vermell els quantils

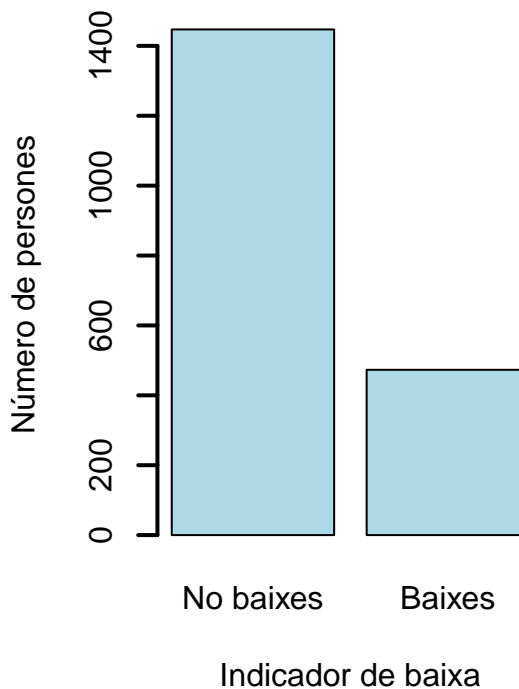
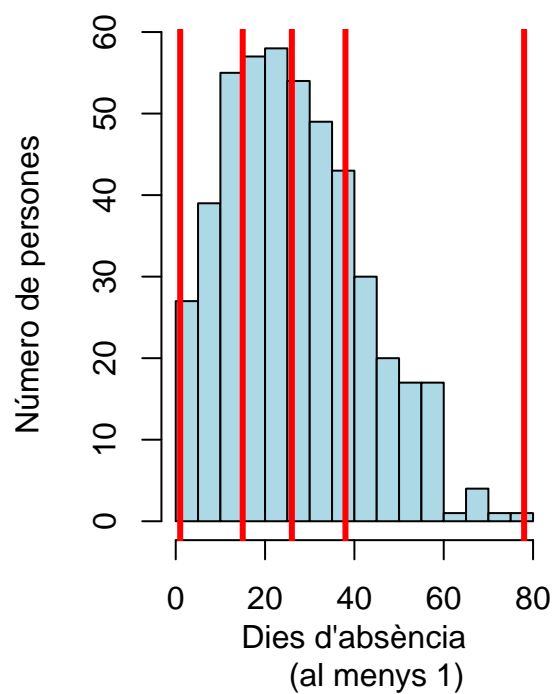
```
max_work_hours <- max(work_hours)+1
hist(work_hours, breaks=c(min(work_hours):max_work_hours), main="Distribució de persones per hores de f",
      xlab="Hores setmanals. En vermell els quantils", ylab="Número de persones", col="#ADD8E6")
abline(v=quantile(work_hours), col="red", lwd=3)
```

Distribució de persones per hores de feina setmanals

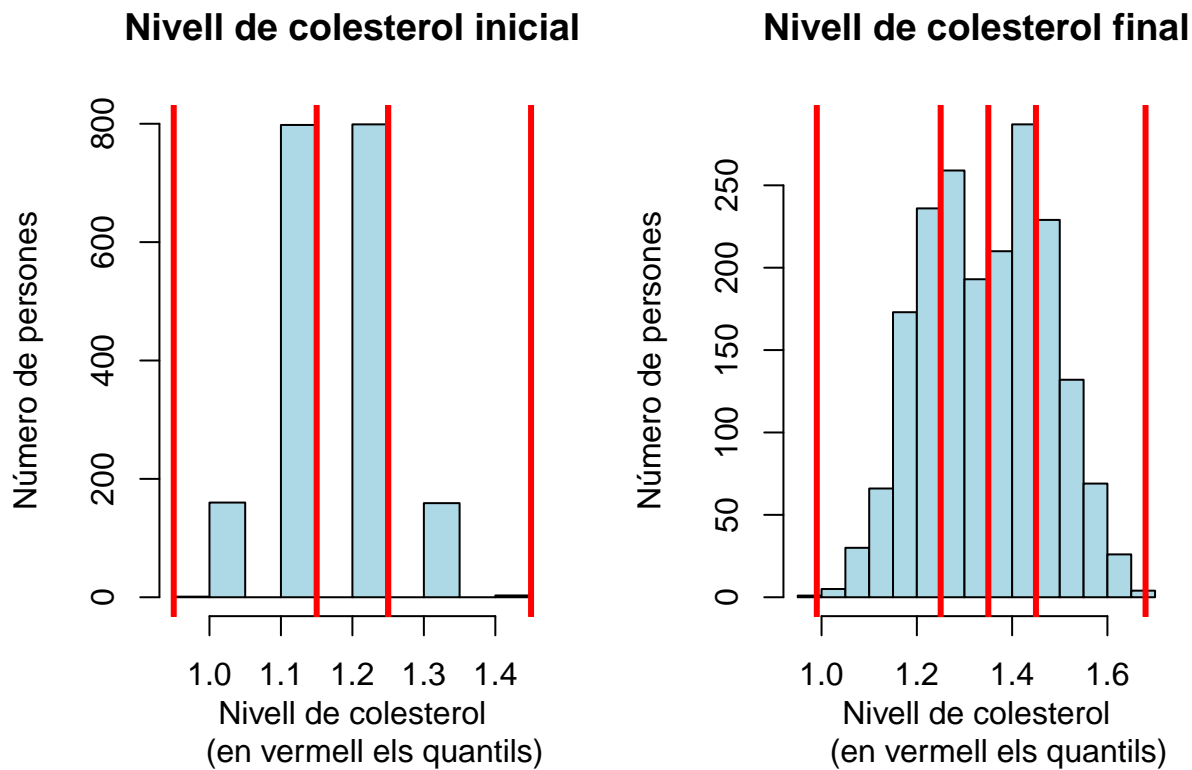


```
par(mfrow=c(1,2))
max_sick_leave <- max(sick_leave) + 5
hist(sick_leave[sick_leave>0], breaks=seq(from=0, to=max_sick_leave, by=5), main="Persones per dies d'absència",
      xlab="Dies d'absència", ylab="Número de persones", col="#ADD8E6")
abline(v=quantile(sick_leave[sick_leave>0]), col="red", lwd=3)
sick_leave_bDesc <- factor(sick_leave_b, levels=c("0","1"), labels=c("No baixes", "Baixes"))
plot(sick_leave_bDesc, main="", xlab="Indicador de baixa", ylab="Número de persones", col="#ADD8E6", lwd = 2)
```

Persones per dies d'absència



```
hist(Cho_initial, main="Nivell de colesterol inicial",
     xlab="Nivell de colesterol
       (en vermell els quantils)", ylab="Número de persones", col="#ADD8E6")
abline(v=quantile(Cho_initial), col="red", lwd=3)
hist(Cho_final, main="Nivell de colesterol final",
     xlab="Nivell de colesterol
       (en vermell els quantils)", ylab="Número de persones", col="#ADD8E6")
abline(v=quantile(Cho_final), col="red", lwd=3)
```



```
dev.off()
```

```
## null device
##          1
```

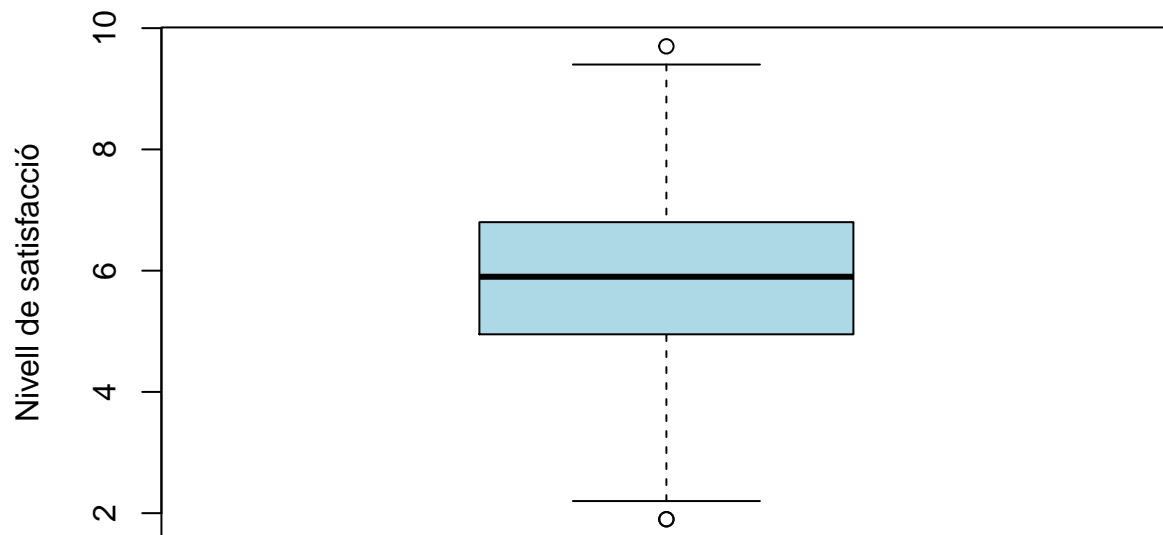
Satisfacció en el treball en relació al sexe

Boxplot

Càlcul de happiness en funció del sexe Primer desglosem l'informació en funció del sexe

```
satisfaccioLaboral.dones <- satisfaccioLaboral[sex=="F",]
satisfaccioLaboral.homes <- satisfaccioLaboral[sex=="M",]
boxplot(satisfaccioLaboral.dones$happiness, main="Satisfacció laboral (dones)",
        xlab="", ylab="Nivell de satisfacció", col="#ADD8E6")
```

Satisfacció laboral (dones)



```
boxplot.stats(satisfaccioLaboral.dones$happiness)$out
```

```
## [1] 9.7 1.9 1.9
```

```
summary(satisfaccioLaboral.dones$happiness)
```

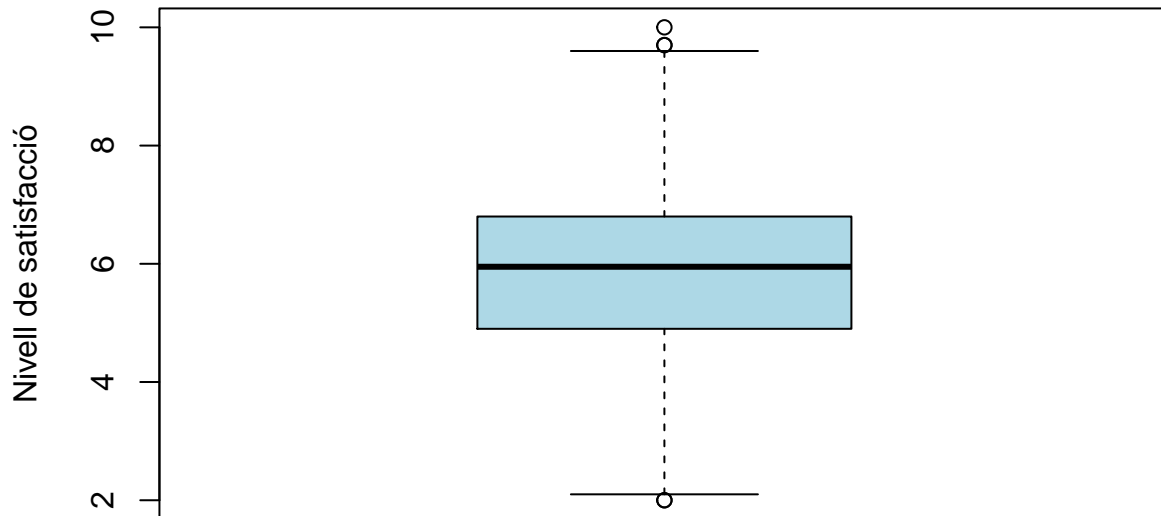
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.900  4.975   5.900   5.884   6.800   9.700
```

```
sd(satisfaccioLaboral.dones$happiness)
```

```
## [1] 1.351982
```

```
boxplot(satisfaccioLaboral.homes$happiness, main="Satisfacció laboral (homes)",
        xlab="", ylab="Nivell de satisfacció", col="#ADD8E6")
```

Satisfacció laboral (homes)



```
boxplot.stats(satisfaccioLaboral.homes$happiness)$out
```

```
## [1] 10.0  9.7  2.0  2.0  9.7
```

```
summary(satisfaccioLaboral.homes$happiness)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      2.000  4.900  5.950  5.883  6.800  10.000
```

```
sd(satisfaccioLaboral.homes$happiness)
```

```
## [1] 1.343949
```

A nivell visual no hi ha una gran diferència per sexe. Els interquantils són pràcticament els mateixos. Sí que tenim més valors outliers per adult en el cas d'homes (3 vs 1) i més per baix en cas de dones (2 vs 1).

Calculem els tres intervals (happiness de la mostra total, happiness de les dones i happiness dels homes):

```
margError <- qt(0.05, df=(nrow(satisfaccioLaboral)-1),
               lower.tail=FALSE)*sd(happiness)/sqrt(nrow(satisfaccioLaboral))
```

```
intEsq <- mean(happiness)-margError
```

```
intDer <- mean(happiness)+margError
```

```
print(c(intEsq, intDer))
```

```
## [1] 5.832461 5.933685
```

```
margError <- qt(0.05, df=(nrow(satisfaccioLaboral)-1), lower.tail=FALSE)*
               sd(satisfaccioLaboral.dones$happiness)/sqrt(nrow(satisfaccioLaboral))
```

```
intEsq <- mean(satisfaccioLaboral.dones$happiness)-margError
```

```
intDer <- mean(satisfaccioLaboral.dones$happiness)+margError
```

```
print(c(intEsq, intDer))
```

```
## [1] 5.832870 5.934422
```

```
margError <- qt(0.05, df=(nrow(satisfaccioLaboral)-1), lower.tail=FALSE)*  
             sd(satisfaccioLaboral.homes$happiness)/sqrt(nrow(satisfaccioLaboral))  
intEsq <- mean(satisfaccioLaboral.homes$happiness)-margError  
intDer <- mean(satisfaccioLaboral.homes$happiness)+margError  
print(c(intEsq, intDer))
```

```
## [1] 5.832026 5.932974
```

L'interval de confiança ens dona un 90% de probabilitats de que la mitjana poblacional es trobi en el rang indicat.

Escriure la hipòtesi nul · la i alternativa

Si mirem les dades anteriors, podem apreciar que els valors estadístics descriptius són molt similars: la mitja i la mediana són pràcticament idèntiques. La major diferència la trobem a la variança.

Analitzarem dues possibles hipòtesis nul · les (i en conseqüència dues alternatives) per intentar veure si el nivell de happiness és diferent en homes que en dones:

1. Hipòtesi nul · la: Les mitjanes són diferents. Alternativa: són iguals i, per tant, indistingibles.
2. Hipòtesi nul · la: Les variàncies són diferents. Alternativa: són iguals i, per tant, indistingibles.

Mètode

En el nostre cas, no coneixem ni la mitjana ni la variança poblacional. Només coneixem la mitjana i la variança de la mostra. El nostre objectiu és comprovar si el nivell de felicitat (happiness) depèn del sexe de la persona. Això ho podem validar comparant les mitjanes i variàncies mostrals. En aquest cas, per a contrastar les mitjanes, un mètode adequat és el t de Student. El mateix és aplicable al contrast de variàncies.

Calcular l'estadístic de contrast, el valor crític i el valor p

Per lo que podem veure, hem de rebutjar ambdues hipòtesis nul·les. El rang de la mitjana inclou el valor 0 i el de la variança l'1.

```
var.test(happiness ~ sex, alternative='two.sided', conf.level=.95, var.equal=FALSE,  
         data=satisfaccioLaboral)
```

```
##
```

```
## F test to compare two variances
```

```
##
```

```
## data: happiness by sex
```

```
## F = 1.012, num df = 959, denom df = 959, p-value = 0.8536
```

```
## alternative hypothesis: true ratio of variances is not equal to 1
```

```
## 95 percent confidence interval:
```

```
## 0.8915994 1.1486351
```

```
## sample estimates:
```

```
## ratio of variances
```

```
## 1.011989
```

```
t.test(happiness~sex, alternative='two.sided', conf.level=.95, var.equal=FALSE,  
       data=satisfaccioLaboral)
```

```
##
```

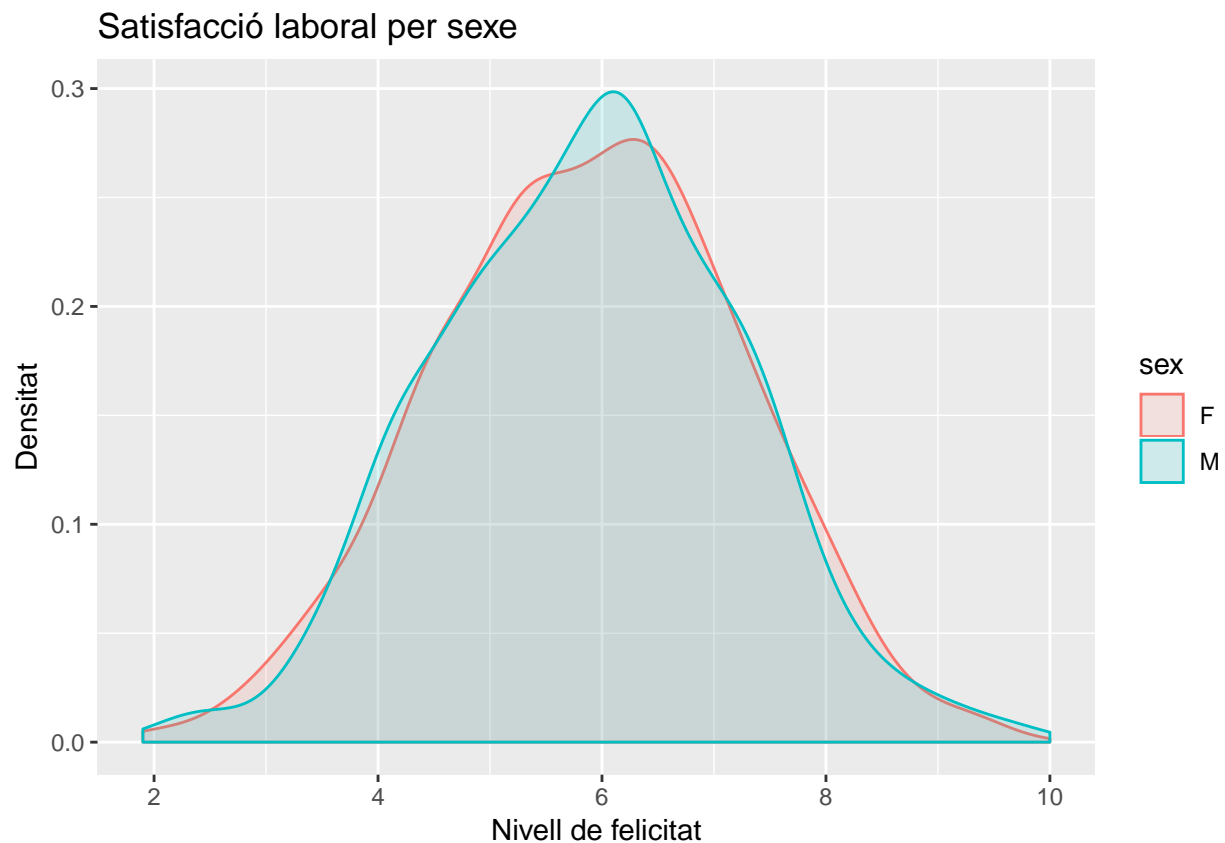
```
## Welch Two Sample t-test
```

```
##
## data: happiness by sex
## t = 0.018623, df = 1917.9, p-value = 0.9851
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.1195195  0.1218111
## sample estimates:
## mean in group F mean in group M
##      5.883646      5.882500
```

Com es pot veure adalt, els respectius valors de p són $p\text{-value} = 0.8536$ i $p\text{-value} = 0.9851$.

Gràficament, podem veure que les funcions de densitat se semblen:

```
library(ggplot2)
distinctSex <- c("M", "F")
ds <- subset(satisfaccioLaboral, sex %in% distinctSex)
p <- ggplot(data=ds, aes(happiness, group=sex, colour=sex, fill=sex))
p <- p + ggtitle("Satisfacció laboral per sexe") +
  xlab("Nivell de felicitat") + ylab("Densitat") +
  theme(legend.position="right")+
  geom_density(alpha=0.15)
p
```



Si considerem ambdues distribucions per separat (dones i homes), podem calcular els valors crítics de la mitja amb la distribució Khi-quadrat i calcular l'interval de confiança:


```
qchisq(c(0.975,0.025), df=nrow(satisfaccioLaboral), lower.tail=TRUE)

## [1] 2043.338 1800.451

((nrow(satisfaccioLaboral) - 1)*mean(satisfaccioLaboral.dones$happiness)) /
qchisq(c(0.975,0.025), df=nrow(satisfaccioLaboral), lower.tail=TRUE)

## [1] 5.525625 6.271050

((nrow(satisfaccioLaboral) - 1)*mean(satisfaccioLaboral.homes$happiness)) /
qchisq(c(0.975,0.025), df=nrow(satisfaccioLaboral), lower.tail=TRUE)

## [1] 5.524548 6.269829
```

Interpretar el resultat

Segons la mostra de dades, podem afirmar que no hi ha diferència de happiness depenent del sexe de la persona. Les mostres són molt similars i els indicadors no ens permeten afirmar que hi ha cap diferència entre ambdues mostres (dones i homes).

Test no paramètric

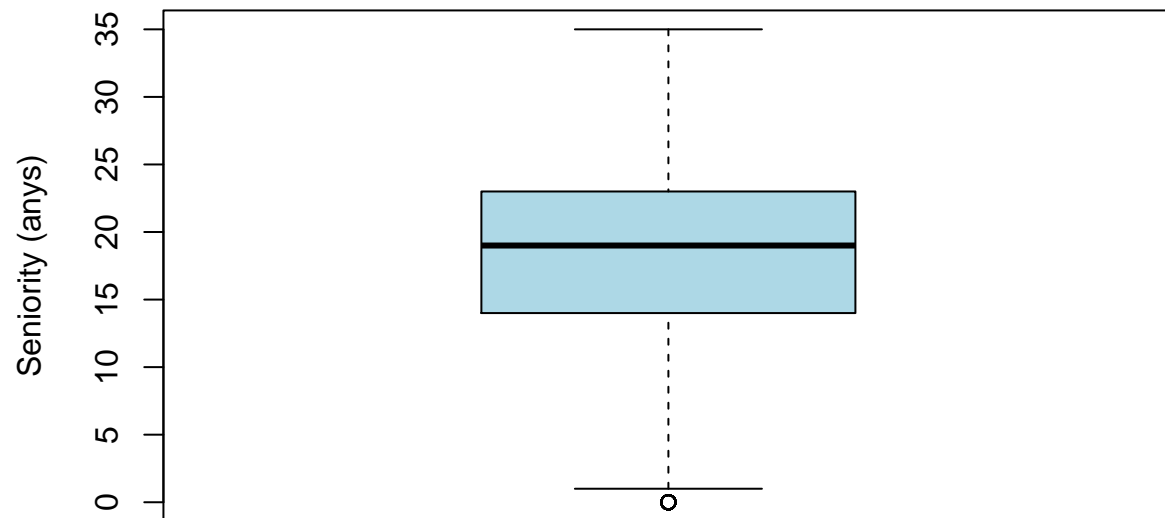
Calcular si hi ha diferències de seniority entre els treballadors qualificats (C) i poc qualificats(PC)

Hipòtesi nul · la i alternativa

Mirem si realment hi ha diferències entre els boxplot de les dues mostres (C i PC). Això ens donarà una pista de com de disperses són les dades.

```
satisfaccioLaboral.seniorityC <- satisfaccioLaboral[job_type=="C",]
satisfaccioLaboral.seniorityPC <- satisfaccioLaboral[job_type=="PC",]
boxplot(satisfaccioLaboral.seniorityC$seniority, main="Seniority (nivell C)",
        xlab="", ylab="Seniority (anys)", col="#ADD8E6")
```

Seniority (nivell C)

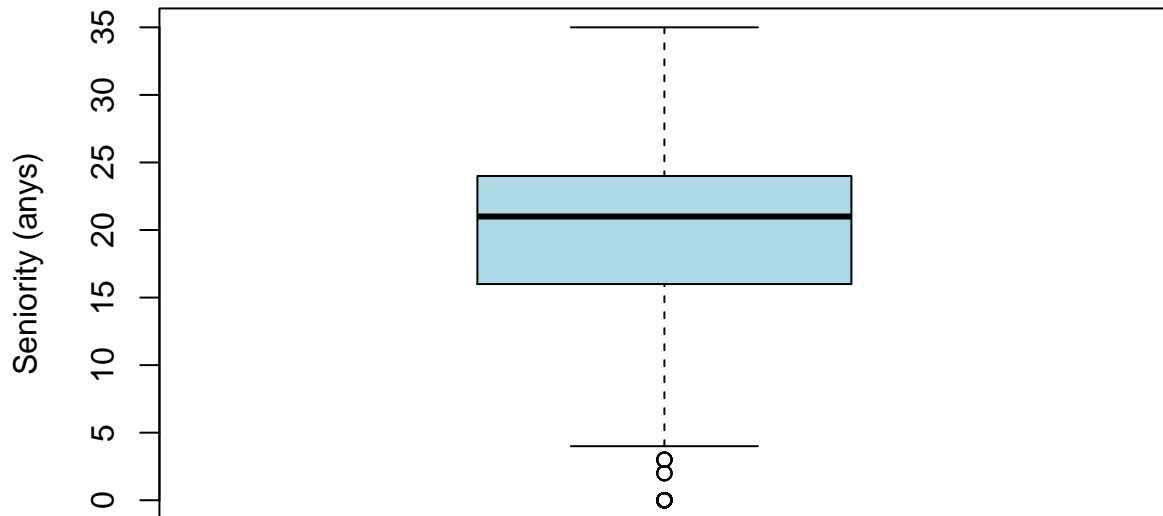


```
boxplot.stats(satisfaccioLaboral.seniorityC$seniority)$out
```

```
## [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
boxplot(satisfaccioLaboral.seniorityPC$seniority, main="Seniority (nivell PC)",
        xlab="", ylab="Seniority (anys)", col="#ADD8E6")
```

Seniority (nivell PC)



```
boxplot.stats(satisfaccioLaboral.seniorityPC$seniority)$out
```

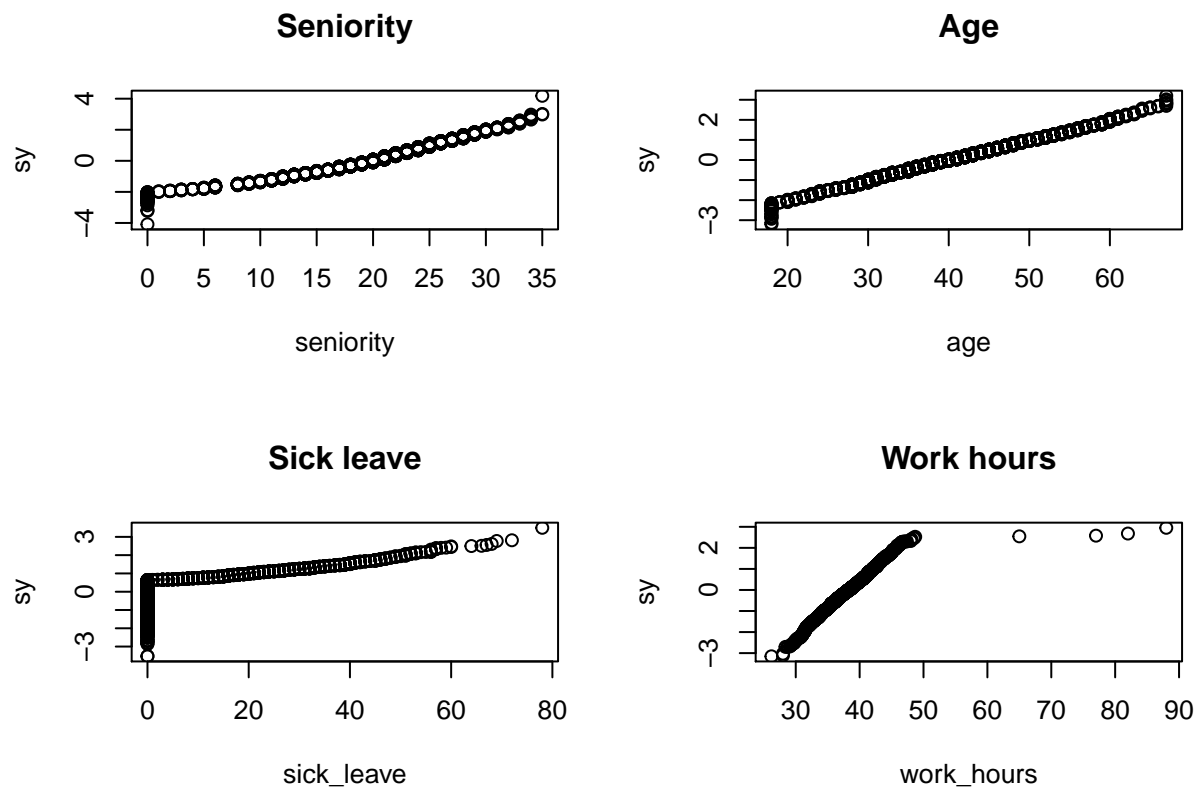
```
## [1] 3 3 0 2 0 0 3 0 0 0 3 0 0 0 2 0 0 2 0 0 0 0 0 0 0 0
```

1. Hipòtesi nul·la: Les mitjanes són diferents. Alternativa: són iguals i, per tant, indistingibles.

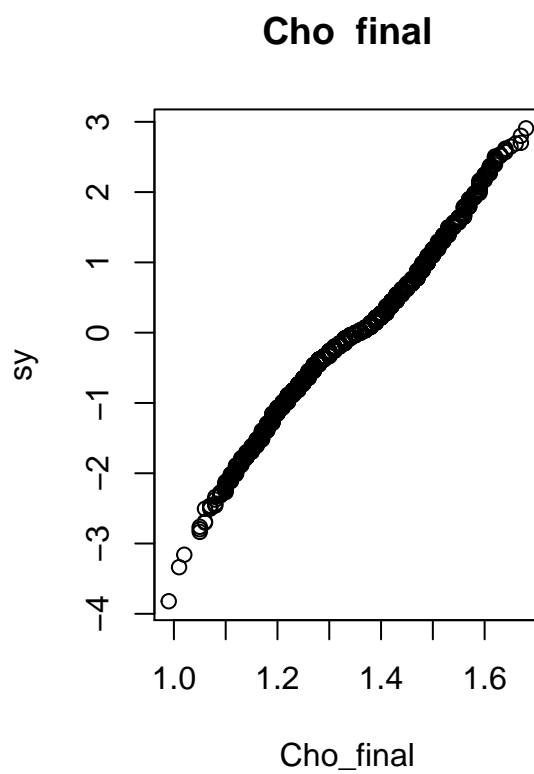
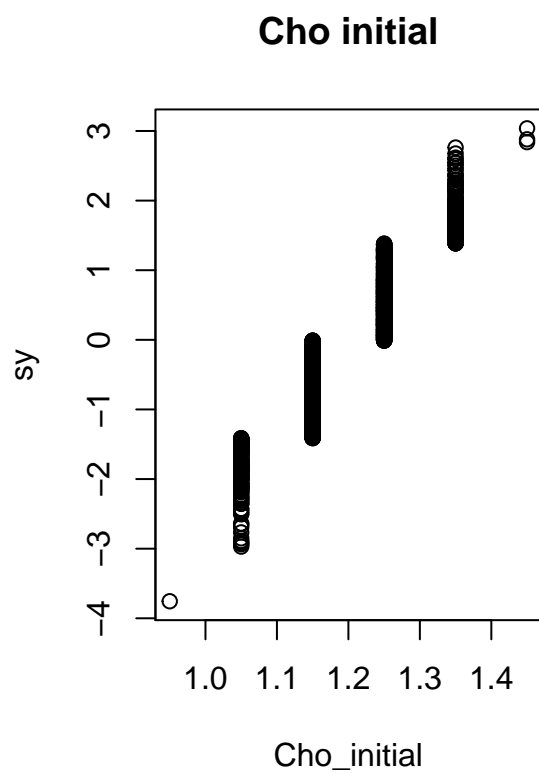
Assumpció de normalitat

Tests de normalitat de les variables numèriques

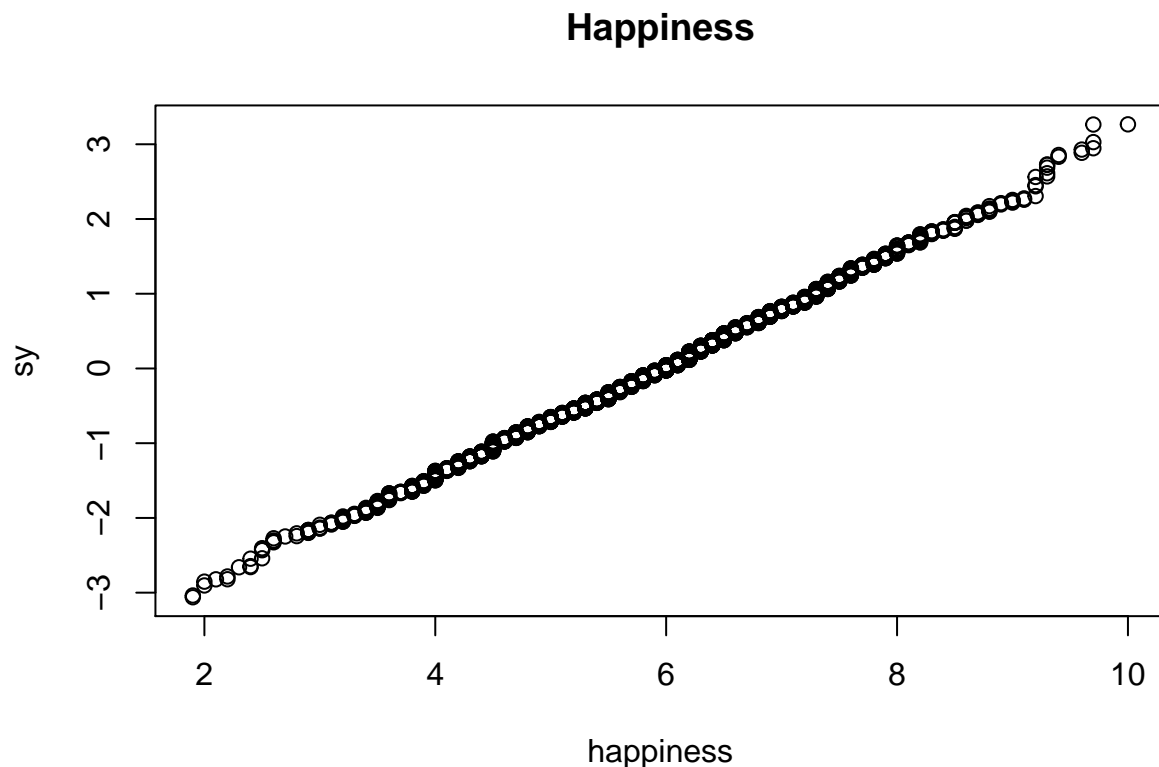
```
par(mfrow=c(2,2))
qqplot(seniority, rnorm(nrow(satisfaccioLaboral)), main="Seniority", ylab = NULL)
qqplot(age, rnorm(nrow(satisfaccioLaboral)), main="Age", ylab = NULL)
qqplot(sick_leave, rnorm(nrow(satisfaccioLaboral)), main="Sick leave", ylab = NULL)
qqplot(work_hours, rnorm(nrow(satisfaccioLaboral)), main="Work hours", ylab = NULL)
```



```
par(mfrow=c(1,2))
qqplot(Cho_initial, rnorm(nrow(satisfaccioLaboral)), main="Cho initial", ylab = NULL)
qqplot(Cho_final, rnorm(nrow(satisfaccioLaboral)), main="Cho final", ylab = NULL)
```



```
qqplot(happiness, rnorm(nrow(satisfaccioLaboral)), main="Happiness", ylab = NULL)
```



Com es pot veure visualment, la variable que més segueix la distribució normal és Happiness. Recordem que:

1. Si la línia és recta, llavors la variable segueix totalment la distribució normal
2. Si la línia fa curva, llavors les dades poden estar esbiaixades (podem veure, per exemple, una lleu corba a la variable seniority)
3. Si hi ha valors que no segueixen la línia, llavors aquests no segueixen la distribució normal. És el cas de seniority, age, sick leave, inici i final de happiness, ...

Ho podem validar executant el test de Shapiro-Wilk per a totes les variables. Si el valor és proper a 1, llavors s'accepta l'hipòtesi de que la variable segueix una distribució normal. Com es pot veure a continuació, happiness o age segueixen molt fortament la distribució normal.

```
shapiro.test(seniority)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  seniority
## W = 0.96881, p-value < 2.2e-16
```

```
shapiro.test(age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  age
## W = 0.9946, p-value = 1.887e-06
```

```
shapiro.test(sick_leave)

##
##  Shapiro-Wilk normality test
##
## data:  sick_leave
## W = 0.55292, p-value < 2.2e-16

shapiro.test(work_hours)

##
##  Shapiro-Wilk normality test
##
## data:  work_hours
## W = 0.89675, p-value < 2.2e-16

shapiro.test(Cho_initial)

##
##  Shapiro-Wilk normality test
##
## data:  Cho_initial
## W = 0.85874, p-value < 2.2e-16

shapiro.test(Cho_final)

##
##  Shapiro-Wilk normality test
##
## data:  Cho_final
## W = 0.98553, p-value = 5.142e-13

shapiro.test(happiness)

##
##  Shapiro-Wilk normality test
##
## data:  happiness
## W = 0.99829, p-value = 0.04441
```

Test U de Mann-Whitney

Hipòtesi alternativa (H_a): Els valors de seniority depenen del tipus de treball (qualificat o no qualificat).
 Hipòtesi nul·la (H_0): Els valors de seniority no depenen del tipus de treball. Nivell de significació. Per a tot valor de probabilitat igual o menor que 0.05, s'accepta H_a y se rebutja H_0 .

$\alpha = 0.05$

```
wilcox.test(satisfaccioLaboral.seniorityC$seniority,
            satisfaccioLaboral.seniorityPC$seniority, correct=FALSE)

##
##  Wilcoxon rank sum test
##
## data:  satisfaccioLaboral.seniorityC$seniority and satisfaccioLaboral.seniorityPC$seniority
## W = 411070, p-value = 4.141e-05
## alternative hypothesis: true location shift is not equal to 0
```

Podem afirmar que la hipòtesi alternativa és correcta; hi ha relació entre el job type i el seniority, ja que el resultat és inferior a α .

Càlculs manuals

Recordem que per a una mostra suficientment gran, la mostra produïda pel test es distribueix de forma normal, on:

$$Z = \frac{U - \bar{U}}{\sigma_u}$$

On:

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - \sum R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - \sum R_2$$

```
n_1 <- nrow(satisfaccioLaboral.seniorityC)
n_2 <- nrow(satisfaccioLaboral.seniorityPC)
satisfaccioLaboral$seniority.rank <- rank(satisfaccioLaboral$seniority, na.last = TRUE,
                                           ties.method = "average")
R_1 <- sum(satisfaccioLaboral$seniority.rank[satisfaccioLaboral$job_type == 'PC'])
R_2 <- sum(satisfaccioLaboral$seniority.rank[satisfaccioLaboral$job_type == 'C'])
R_1
```

```
## [1] 971810.5
```

```
R_2
```

```
## [1] 872349.5
```

Calculem els Valors de U_1 i U_2 :

```
U_1 <- n_1 * n_2 + (n_1 * (n_1 + 1) / 2) - R_1
U_2 <- n_1 * n_2 + (n_2 * (n_2 + 1) / 2) - R_2
U_1
```

```
## [1] 411069.5
```

```
U_2
```

```
## [1] 510530.5
```

La fórmula simplificada de σ_u és:

$$\sigma_u = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

```
sigma_u <- sqrt((n_1 * n_2 * (n_1 + n_2 + 1)) / 12)
sigma_u
```

```
## [1] 12146.31
```

Calculem \bar{U} com:

$$\bar{U} = \frac{n_1 n_2}{2}$$

```
Mitja_U <- (n_1 * n_2) / 2
Mitja_U
```

```
## [1] 460800
```


I finalment calculem la distribució (com es pot veure, el valor absolut de z és el mateix tant si feim servir U_1 com U_2):

$$Z = \frac{U - \bar{U}}{\sigma_u}$$

```
z <- (U_1-Mitja_U)/sigma_u
z
```

```
## [1] -4.094289
```

```
z2 <- (U_2-Mitja_U)/sigma_u
z2
```

```
## [1] 4.094289
```

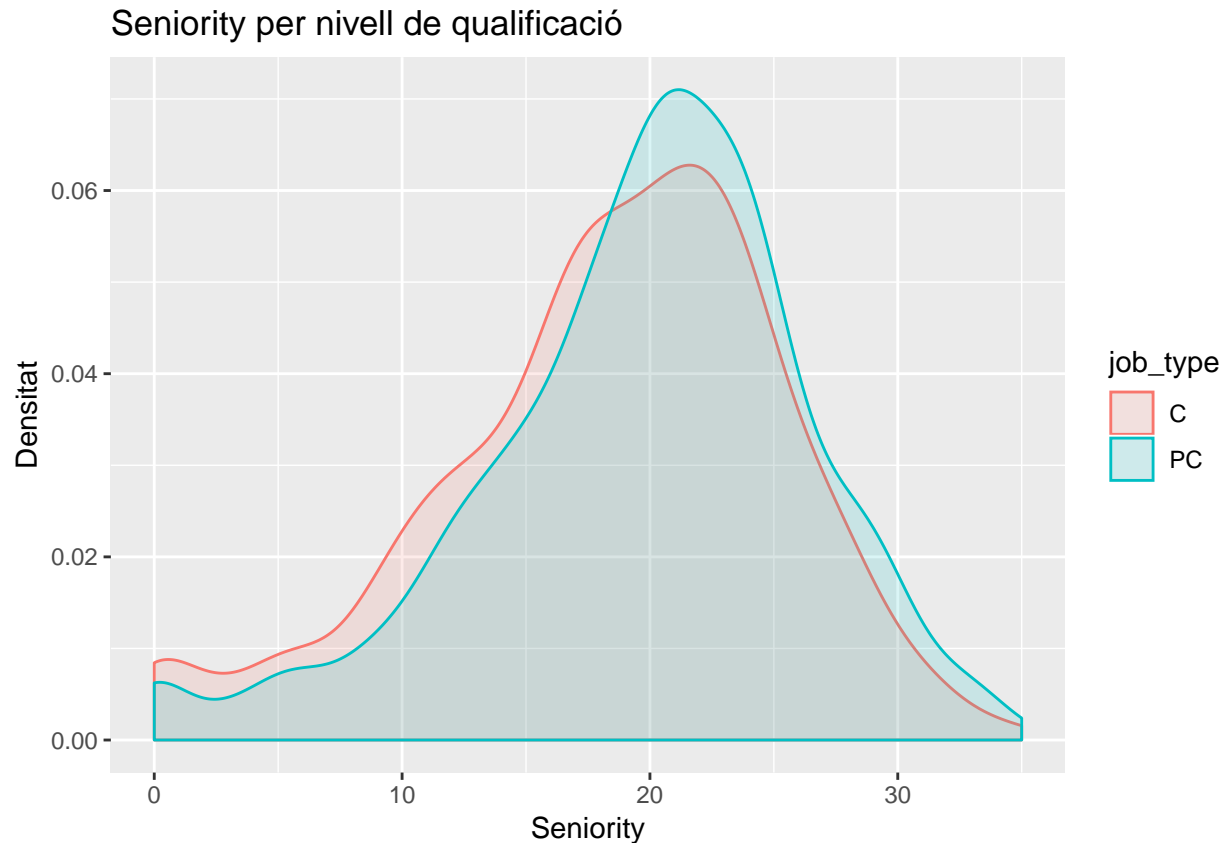
Si calculem l'àrea associada a aquesta z, podem veure que efectivament es troba a la zona de rebuig de l'hipòtesi nul·la:

```
pnorm(z)
```

```
## [1] 2.117326e-05
```

Gràficament es pot veure que les gràfiques són semblants:

```
distinctQua <- c("C", "PC")
ds <- subset(satisfaccioLaboral, job_type %in% distinctQua)
p <- ggplot(data=ds, aes(seniority, group=job_type, colour=job_type, fill=job_type))
p <- p + ggtitle("Seniority per nivell de qualificació") +
  xlab("Seniority") + ylab("Densitat") +
  theme(legend.position="right")+
  geom_density(alpha=0.15)
p
```



Interpretació

El test U de Mann-Whitney ens permet entendre si dues mostres tenen la mateixa distribució. Si la mostra és prou gran, podem afirmar que, aplicant el test, els resultats s'acosten a una distribució Normal. Això ens permet definir fàcilment les àrees d'acceptació i de rebuig. En el nostre cas, ha estat de rebuig de l'hipòtesi nul · la i acceptació de l'alterna.

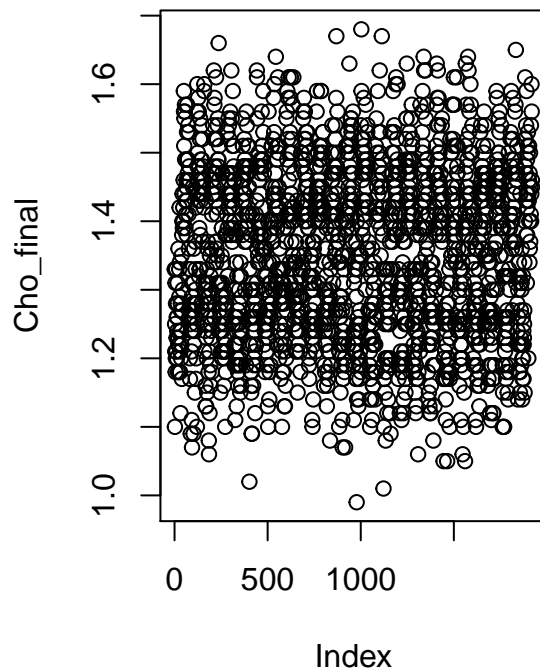
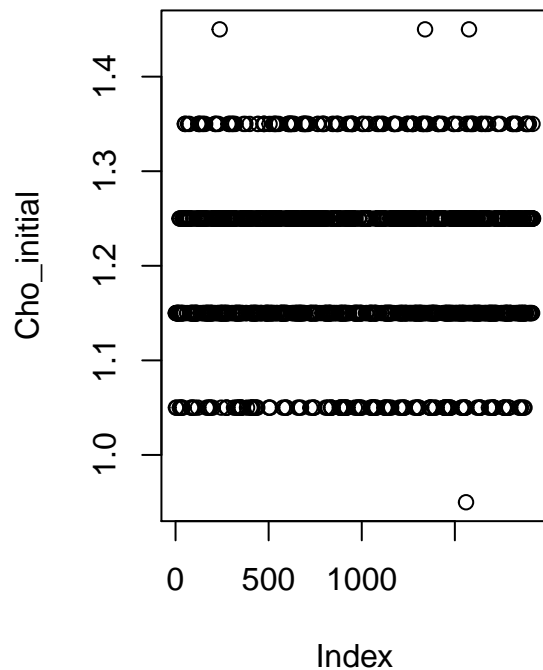
Reflexió sobre tests paramètrics i no paramètrics

Les distribuïcions conegudes (com evidentment és la Normal) tenen un sèrie de característiques i propietats que faciliten el seu tractament. Per exemple, tenim perfectament definit el càlcul de les àrees d'acceptació o rebuig. Una prova és la funció `pnorm`. No necessita informació de la distribució per a calcular l'àrea associada. En definitiva: els càlculs són més senzills, més intuïtius i més coneguts. A més, les comparacions entre distribuïcions són més senzilles si són conegudes.

Test sobre el nivell de colesterol

Com ja hem vist, les dues distribuïcions són molt diferents. De fet, abans ja hem vist que `Cho_final` s'acosta molt a una distribució normal mentre que `cho_initial` té molt poca varietat de valors:

```
par(mfrow=c(1,2))
plot(Cho_initial)
plot(Cho_final)
```



A primer cop d'ull, si restem al colesterol final el colesterol inicial, podem apreciar que la mitja és major que 0. A més, visualment es pot veure que el 0 és fora de l'interval de confiança de la mitja. La primera impressió és que sí sembla que el colesterol final és major que el colesterol inicial.

```
cho_dif <- Cho_final - Cho_initial
summary(cho_dif)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.0600  0.1000   0.1500   0.1511  0.2000   0.3600
```

```
S_dif <- sd(cho_dif)
S_dif
```

```
## [1] 0.07112589
```

```
M_dif <- mean(cho_dif)
M_dif
```

```
## [1] 0.1511146
```

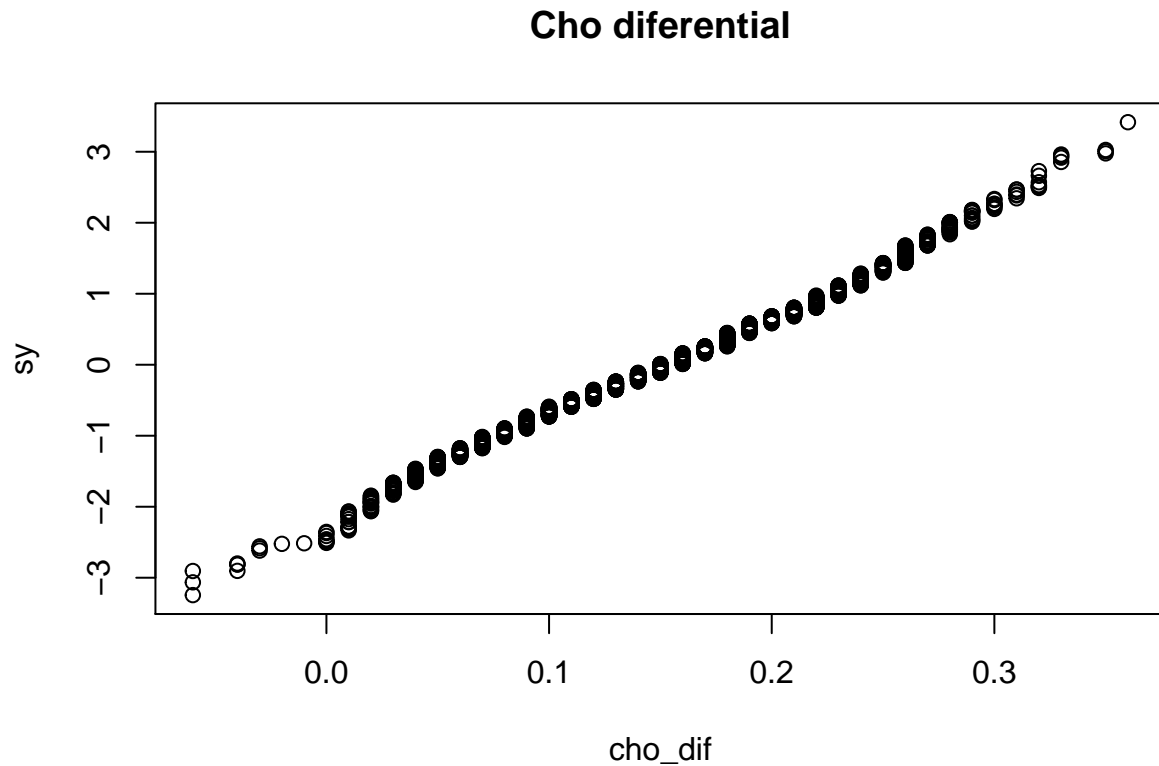
```
boxplot.stats(cho_dif)
```

```
## $stats
## [1] -0.04  0.10  0.15  0.20  0.33
##
## $n
## [1] 1920
##
## $conf
## [1] 0.1463942 0.1536058
```

```
##
## $out
## [1]  0.35 -0.06 -0.06 -0.06  0.35  0.35  0.36
```

Comprovem si la resta entre el colesterol final i l'inicial s'adequa a la distribució Normal. Com podem veure, és apropiat però no arriba a ajustar-se a una distribució Normal.

```
qqplot(cho_dif, rnorm(nrow(satisfaccioLaboral)), main="Cho diferencial", ylab = NULL)
```



```
shapiro.test(cho_dif)
```

```
##
## Shapiro-Wilk normality test
##
## data:  cho_dif
## W = 0.99422, p-value = 8.166e-07
```

Utilitzarem, doncs, la distribució t de Student per a calcular l'interval de confiança:

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

Per a un interval de confiança de 0.1 i 5 graus de llibertat, obtenim els valors de t de Student:

```
qt(c(.05, .95), df=5)
```

```
## [1] -2.015048  2.015048
```

Calculem l'interval de confiança:

```
M_dif + (qt(0.05, df=5) * (S_dif/sqrt(nrow(satisfaccioLaboral))))
```

```
## [1] 0.1478437
```

```
M_dif - (qt(0.05, df=5) * (S_dif/sqrt(nrow(satisfaccioLaboral))))
```

```
## [1] 0.1543854
```

Com es pot veure, l'interval és clarament damunt del 0. Això vol dir que la mostra cho_final és més alta que cho_inicial i que pertanyen a distribucions distintes.

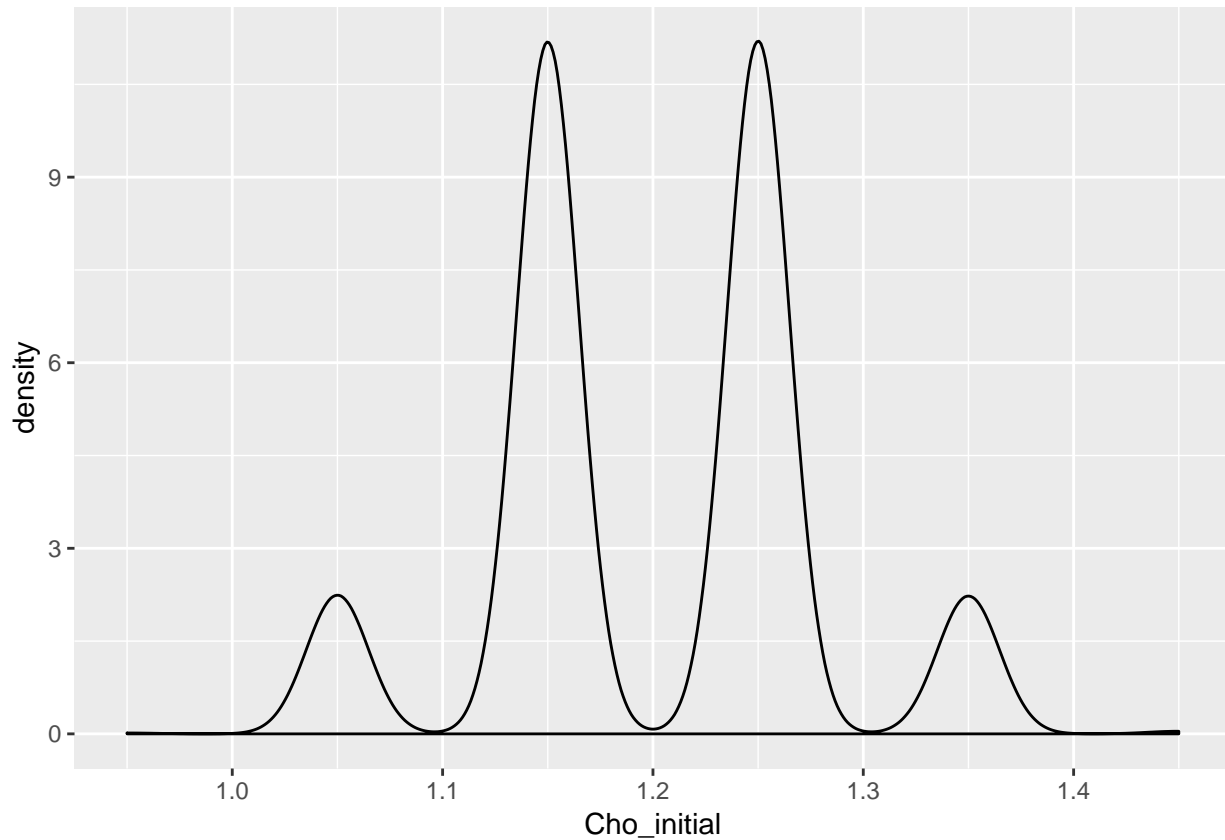
Hipòtesi nul · la i alternativa

H_0 : La distribució de la mostra del colesterol final és superior a la del colesterol inicial

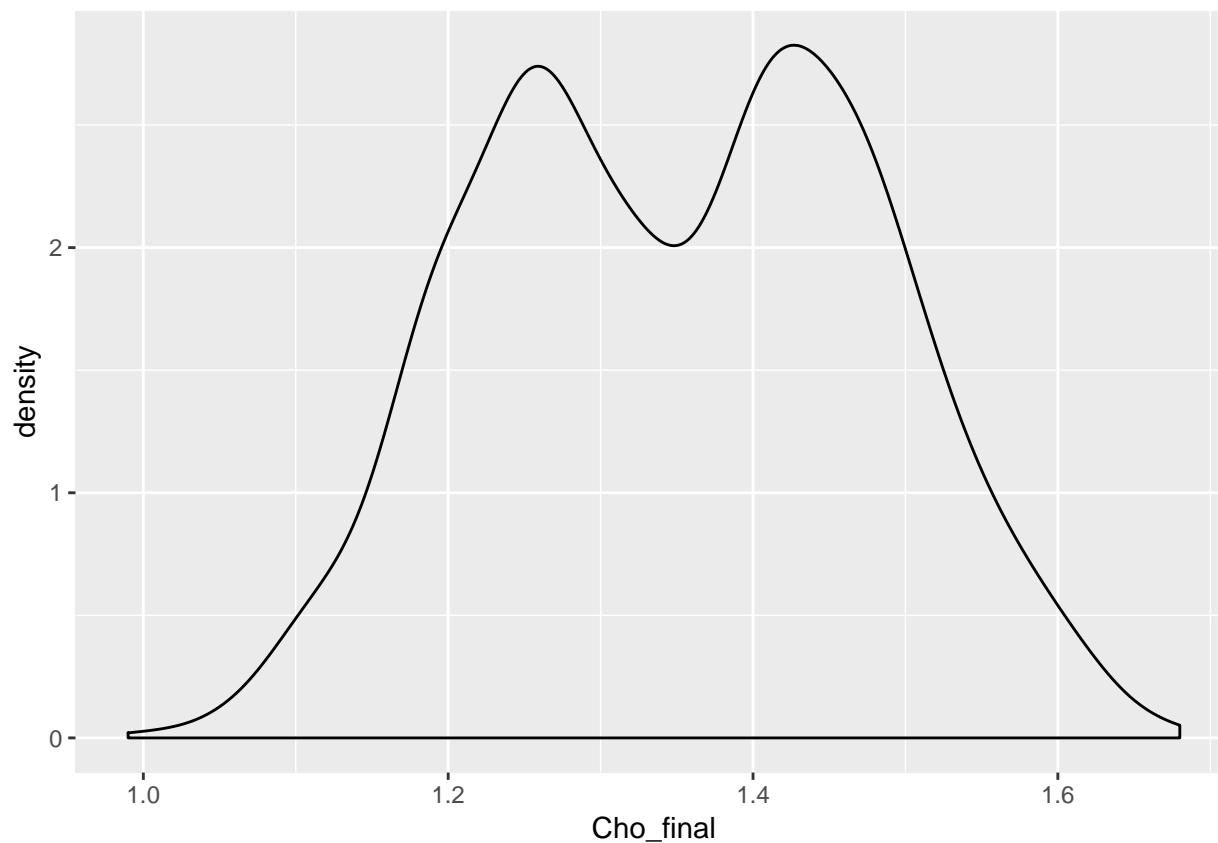
H_a : Les diferències són degudes a l'atzar

Gràficament, es pot veure que les funcions de distribució no se semblen:

```
ggplot(satisfaccioLaboral, aes(Cho_initial)) +  
  geom_density(alpha=0.15)
```



```
ggplot(satisfaccioLaboral, aes(Cho_final)) +  
  geom_density(alpha=0.15)
```



Mètode

El ja aplicat. Com es pot veure, els càlculs s'han aplicat sobre la resta de colesterol final - colesterol inicial. Està clar que si el valor és positiu és perquè $\text{cho_final} > \text{cho_inicial}$. Si és 0, són iguals. Si és negatiu, $\text{cho_final} < \text{cho_inicial}$. El primer que hem fet és veure si aquesta resta segueix una distribució Normal. No és així per poc. Apliquem el test t de Student, que ens permet determinar l'interval de confiança per a una població amb mitja i variància poblacional desconeguda. L'interval NO inclou el 0 i és positiu. Això vol dir que $\text{cho_final} > \text{cho_inicial}$.

Càlculs

Els ja fets.

Interpretació

S'accepta l'hipòtesi nul·la. El nivell de colesterol s'ha incrementat entre mostres.