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

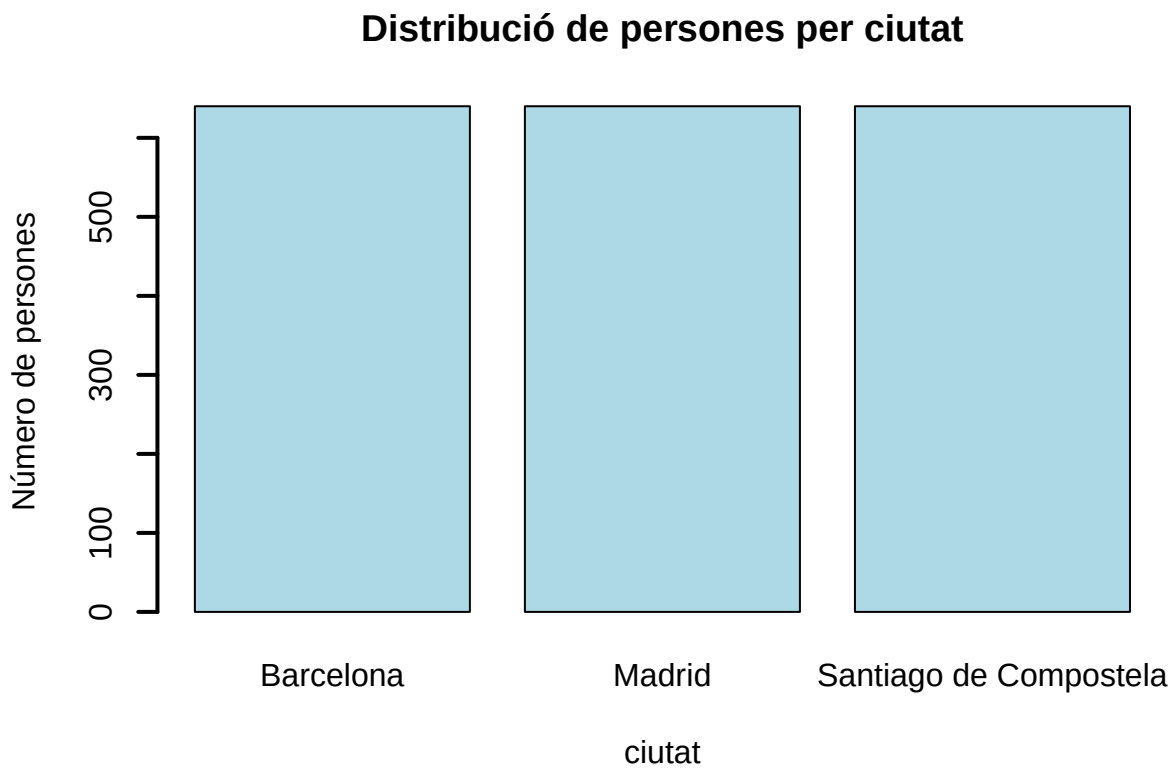
```
sapply(satisfaccioLaboral, class)
```

```
##         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
```

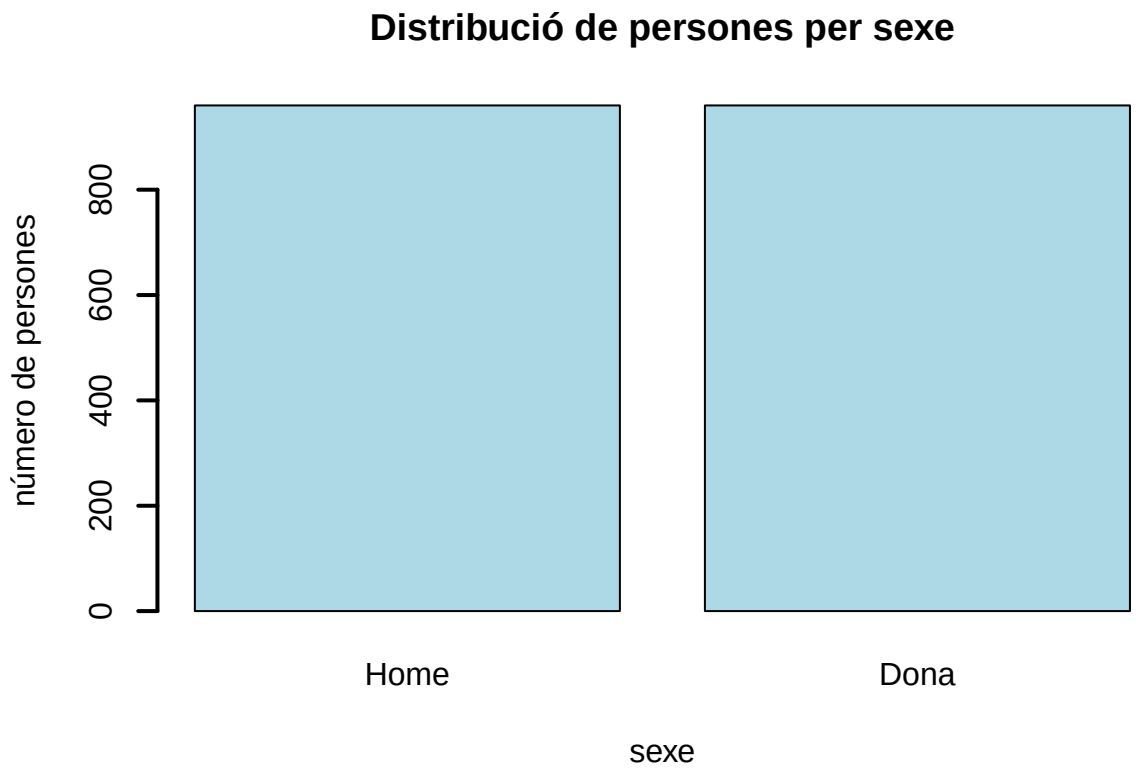
```
##    "numeric"    "numeric"
```

Gràfic de totes les variables:

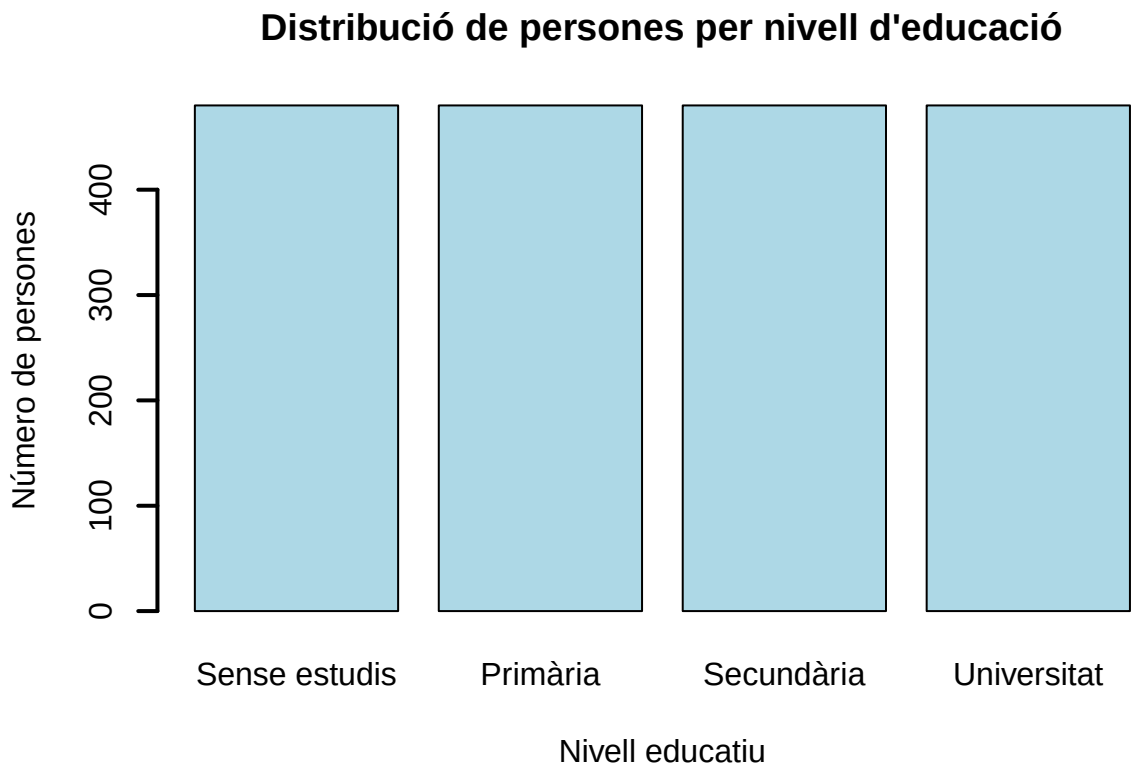
```
plot(city, main="Distribució de persones per ciutat", xlab="ciutat",
     ylab="Número de persones", col="#ADD8E6", lwd = 2)
```



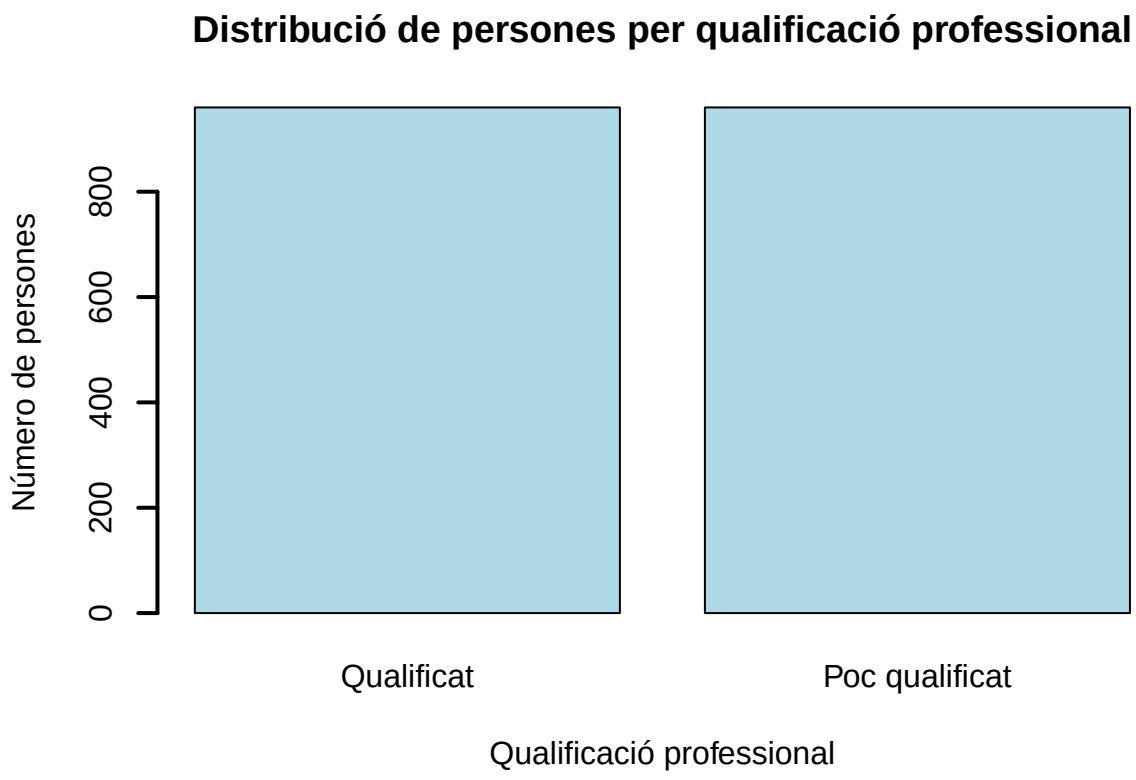

```
sexDesc <- factor(sex, levels=c("M","F"), labels=c("Home","Dona"))
plot(sexDesc, main="Distribució de persones per sexe", xlab="sexe",
     ylab="número de persones", col="#ADD8E6", lwd = 2)
```



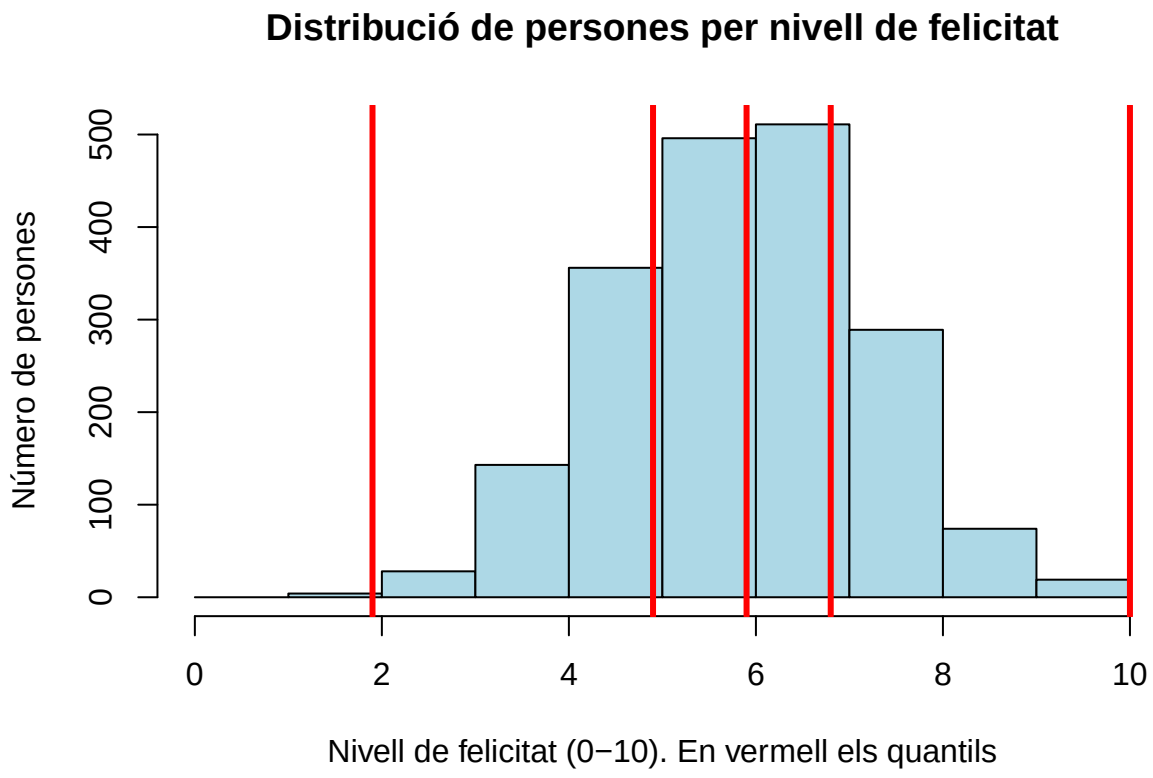

```
educDesc <- factor(educ_level, levels=c("N","P","S","U"), labels=c("Sense estudis","Primària", "Secundà
plot(educDesc, main="Distribució de persones per nivell d'educació",
     xlab="Nivell educatiu", ylab="Número de persones", col="#ADD8E6", lwd = 2)
```



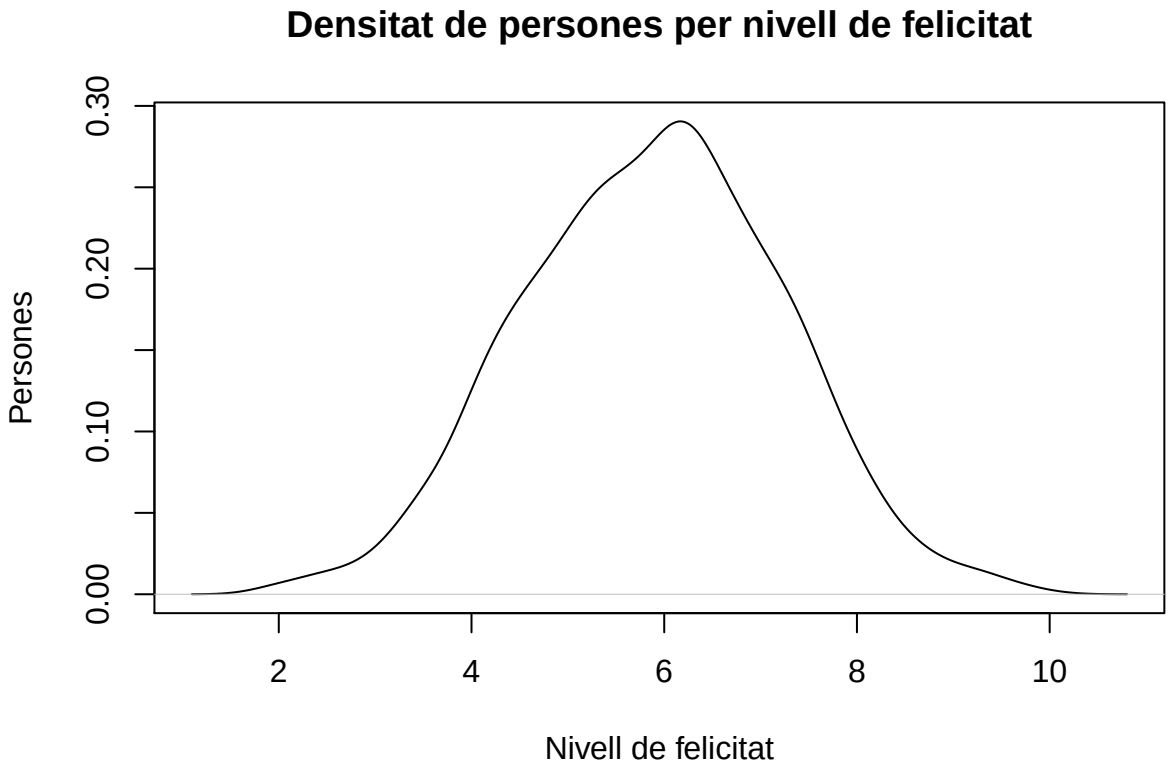

```
jobTypeDesc <- factor(job_type, levels=c("C","PC"), labels=c("Qualificat", "Poc qualificat"))
plot(jobTypeDesc, main="Distribució de persones per qualificació professional",
     xlab="Qualificació professional", ylab="Número de persones", col="#ADD8E6", lwd = 2)
```



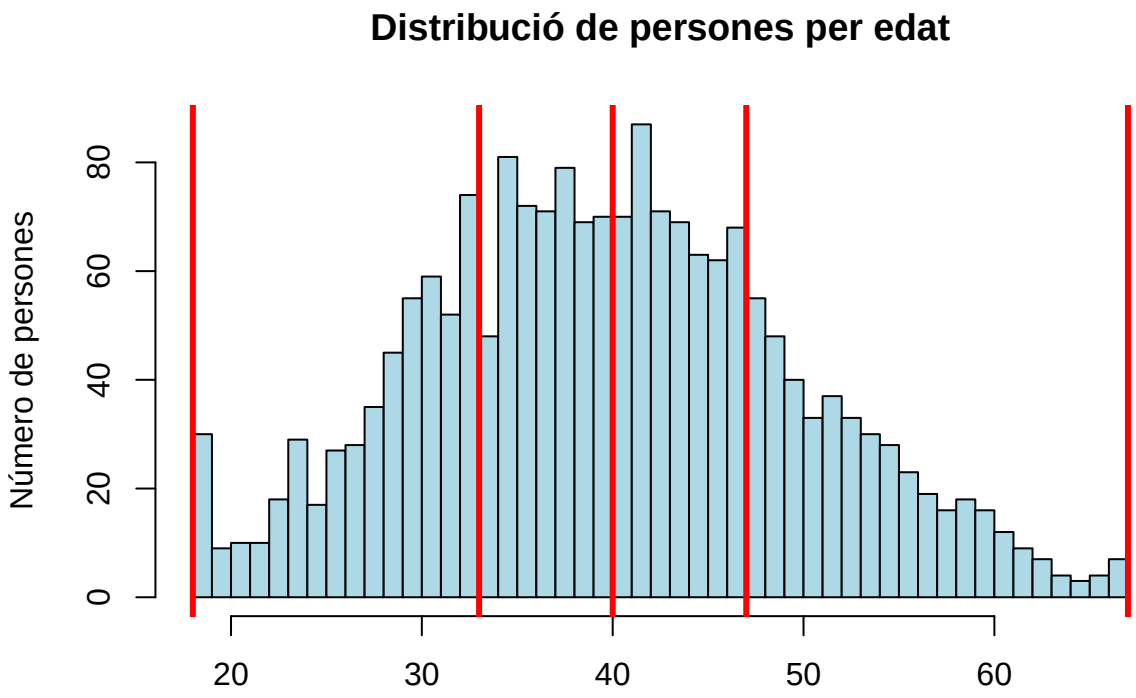

```
hist(happiness, breaks=c(0:10), main="Distribució de persones per nivell de felicitat",
     xlab="Nivell de felicitat (0-10). En vermell els quantils", ylab="Número de persones", col="#ADD8E
abline(v=quantile(happiness), col="red", lwd=3)
```



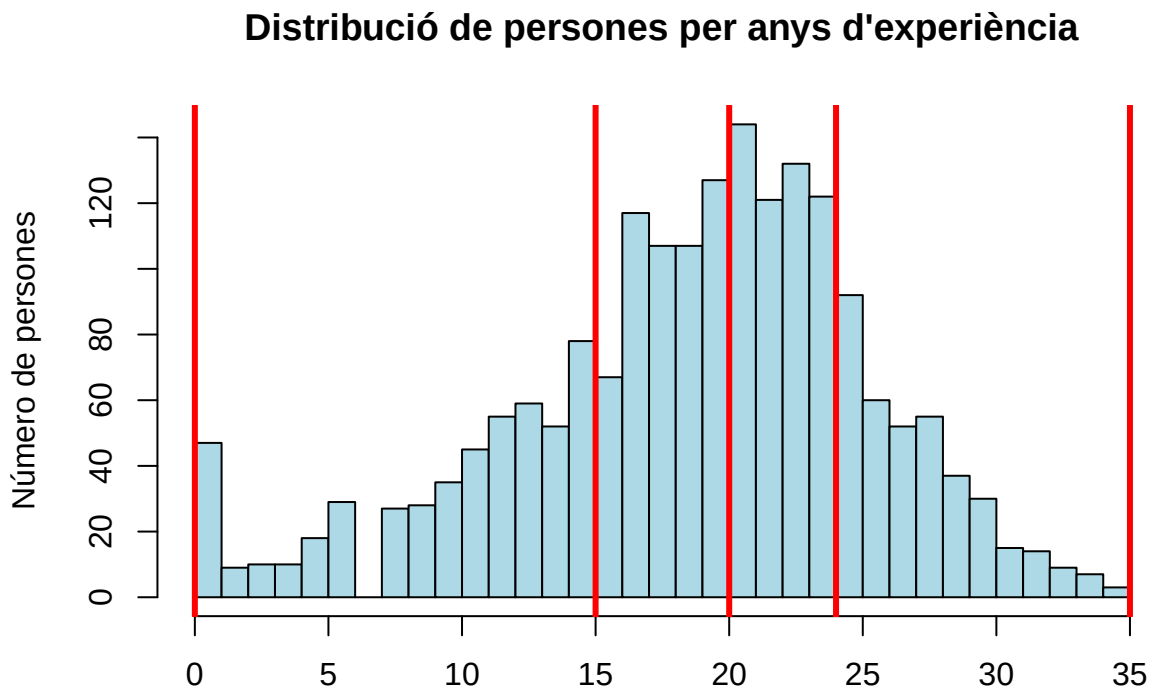

```
den.happiness <- density(happiness)
plot(den.happiness, col="black", main="Densitat de persones per nivell de felicitat",
     xlab="Nivell de felicitat", ylab="Persones")
```




```
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=
abline(v=quantile(age), col="red", lwd=3)
```

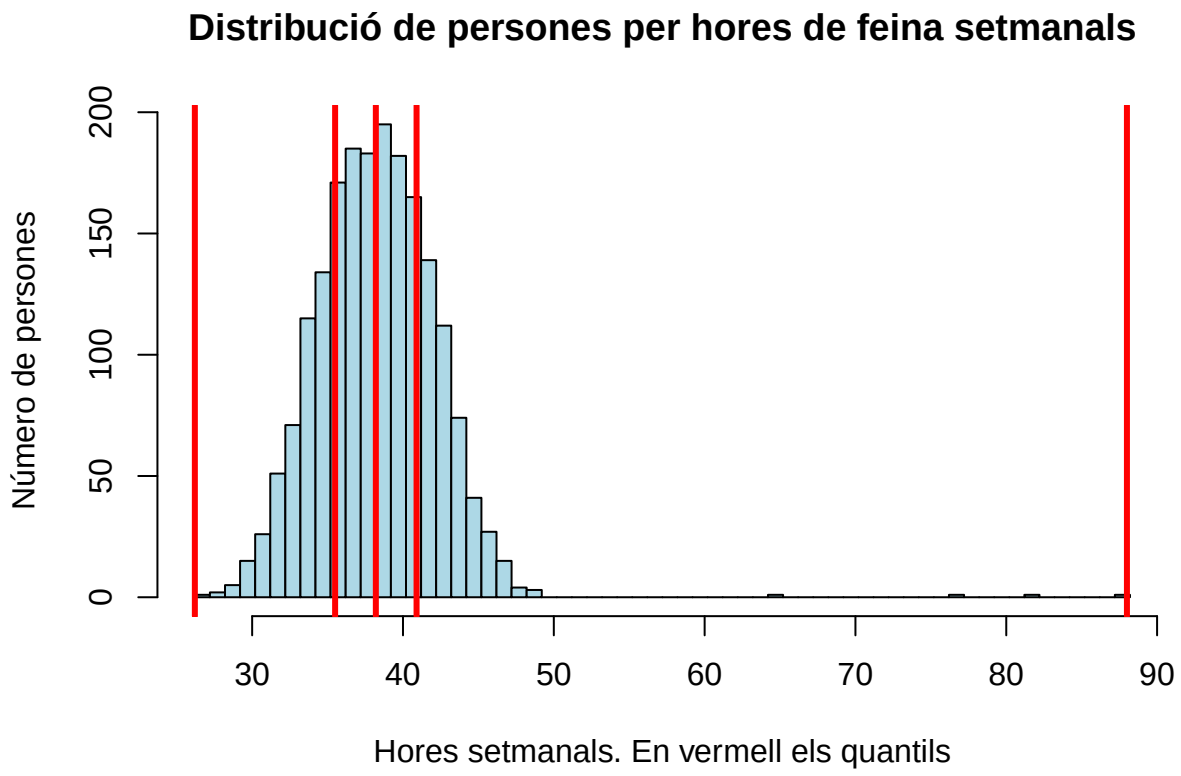



```
hist(seniority, breaks=c(min(seniority):max(seniority)), main="Distribució de persones per anys d'experi
     xlab="Experiència de les persones de la mostra. En vermell els quantils", ylab="Número de persones"
abline(v=quantile(seniority), col="red", lwd=3)
```

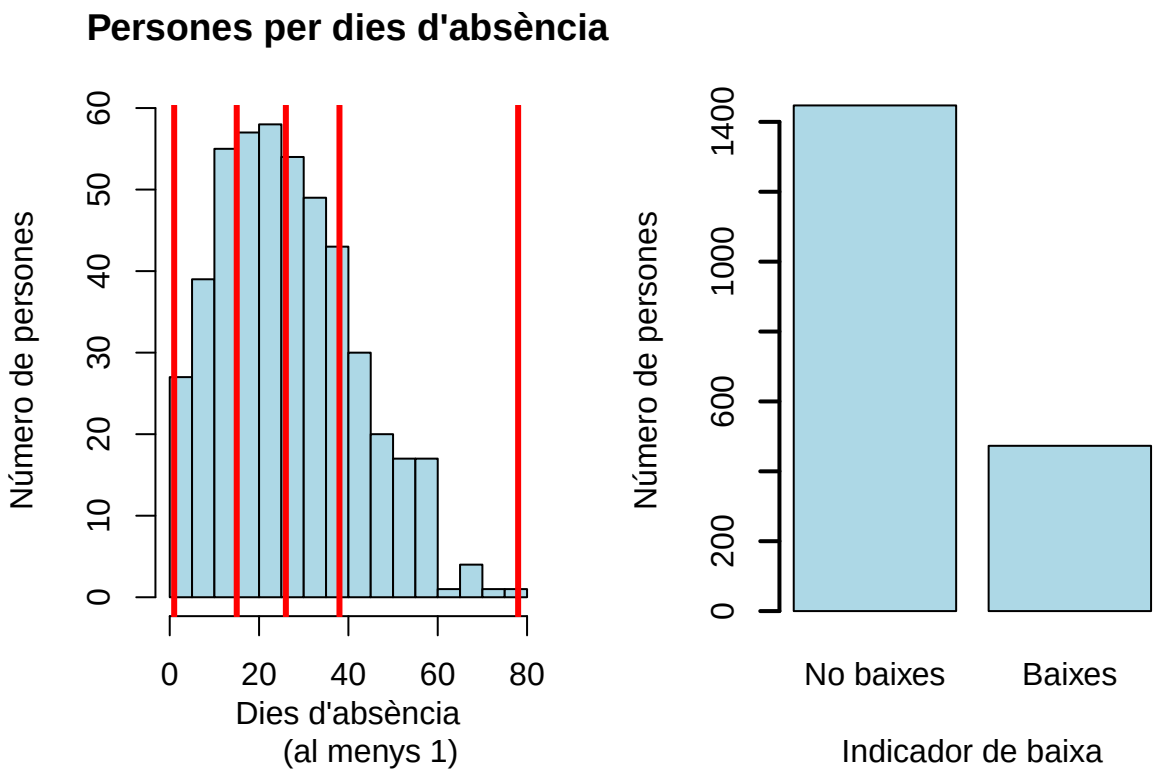


Experiència de les persones de la mostra. En vermell els quantils

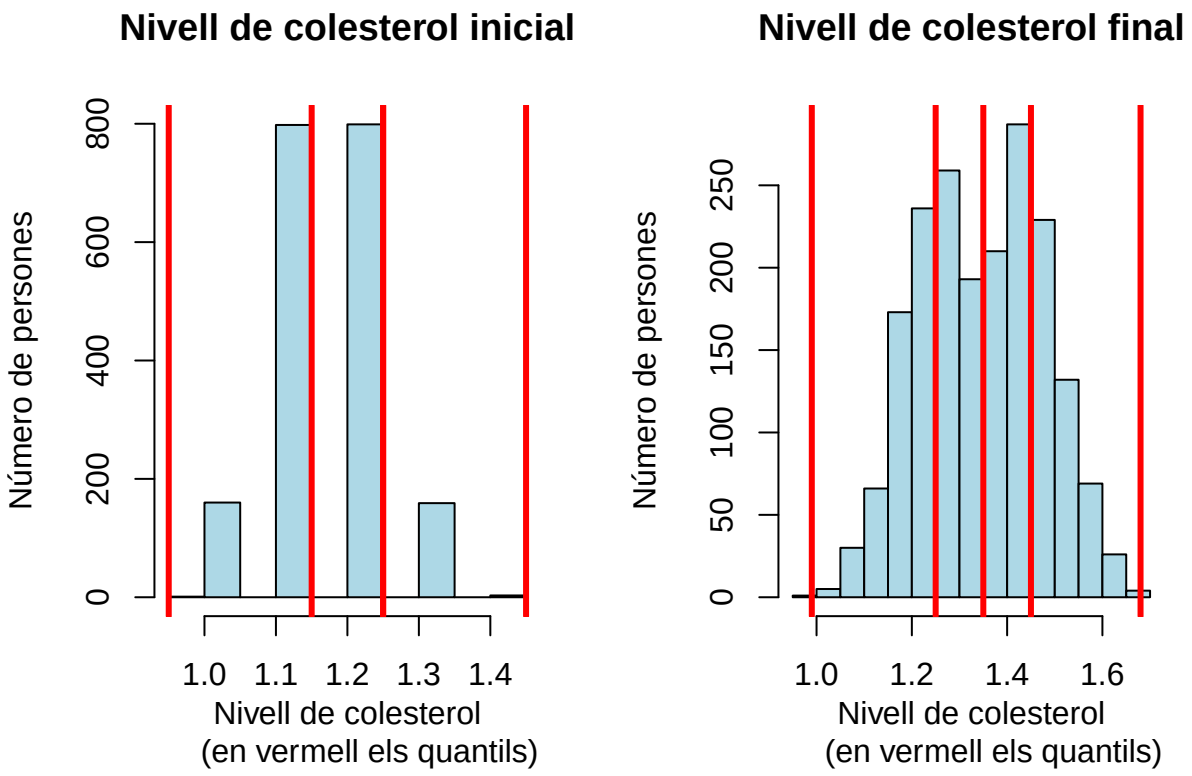
```r
max_work_hours <- max(work_hours)+1
hist(work_hours, breaks=c(min(work_hours):max_work_hours), main="Distribució de persones per hores de f
     xlab="Hores setmanals. En vermell els quantils", ylab="Número de persones", col="#ADD8E6")
abline(v=quantile(work_hours), col="red", lwd=3)
```




```
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'al
     xlab="Dies d'absència
     (al menys 1)", 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)
```




```
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()
```

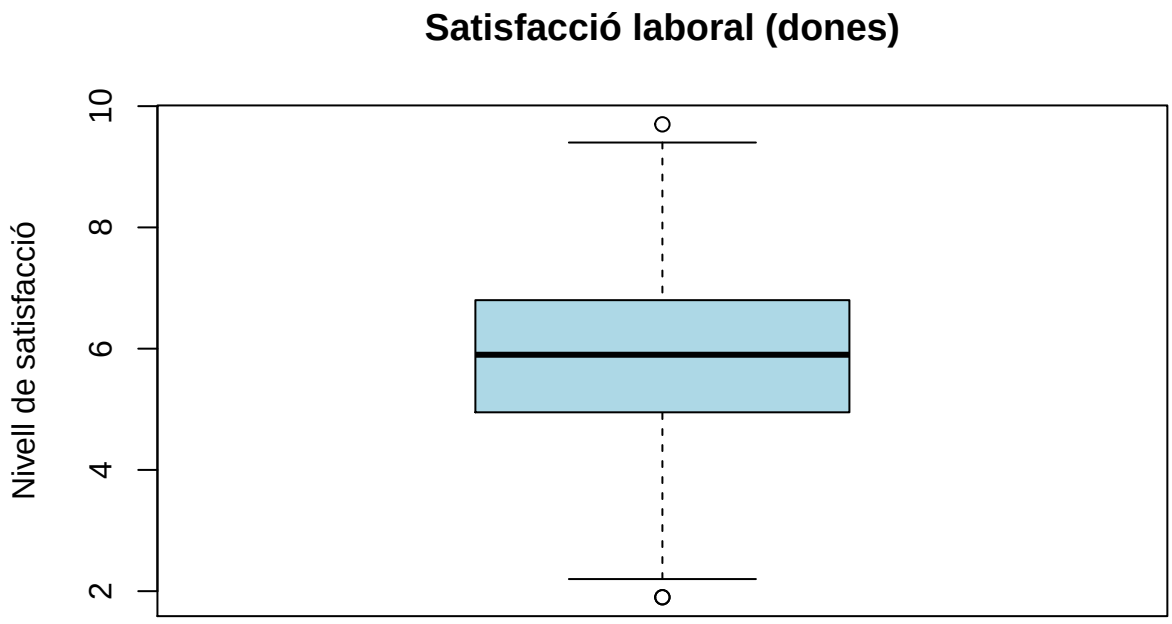
```
## null device
##           1
```

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")
```




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

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

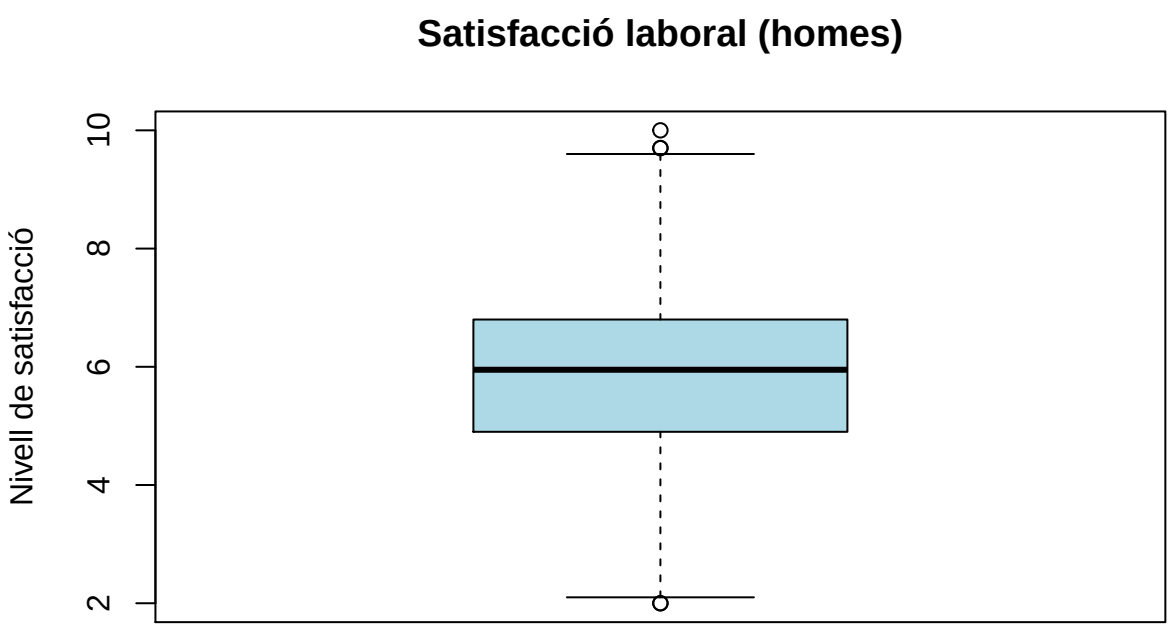
```
summary(satisfaccioLaboral.dones$happiness)
```

```
##    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")
```




```
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 adalt en el cas d'homes (3 vs 1) i més per abaix 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
```

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

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

```r
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 dóna 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, indistinguibles.
2. Hipòtesi nul · la: Les variances són diferents. Alternativa: són iguals i, per tant, indistinguibles.

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 clar: intentar veure si el nivell de felicitat (happiness) depèn del sexe de la persona. Això ho podrem comprovar comparant les mitjanes i variances mostrals. En aquest cas, per a contrastar les mitjanes, un mètode adequat és el t de Student. El mateix és aplicable al contrast de variances.

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

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

```r
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
```

```r
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-value = 0.8536 i p-value = 0.9851.

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

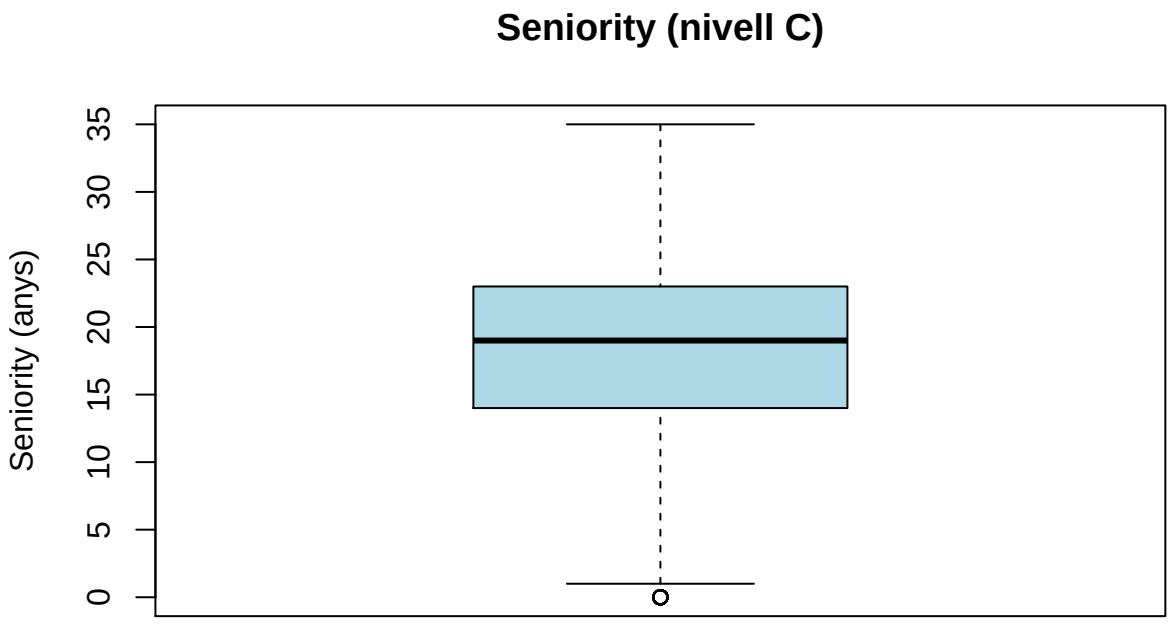
Segons la mostra de dades, podem afirmar que no hi ha diferència de happiness depenen del sexe de la persona. Les mostres són molt similars i els indicadors no ens permeten afirmar que hi ha cap diferència entre ambdues mostres (dones i homes).

Test no paramètric

Hipòtesi nul · la i alternativa

Primer mirem si realment hi ha diferències entre els boxplot de les dues mostres (C i PC)

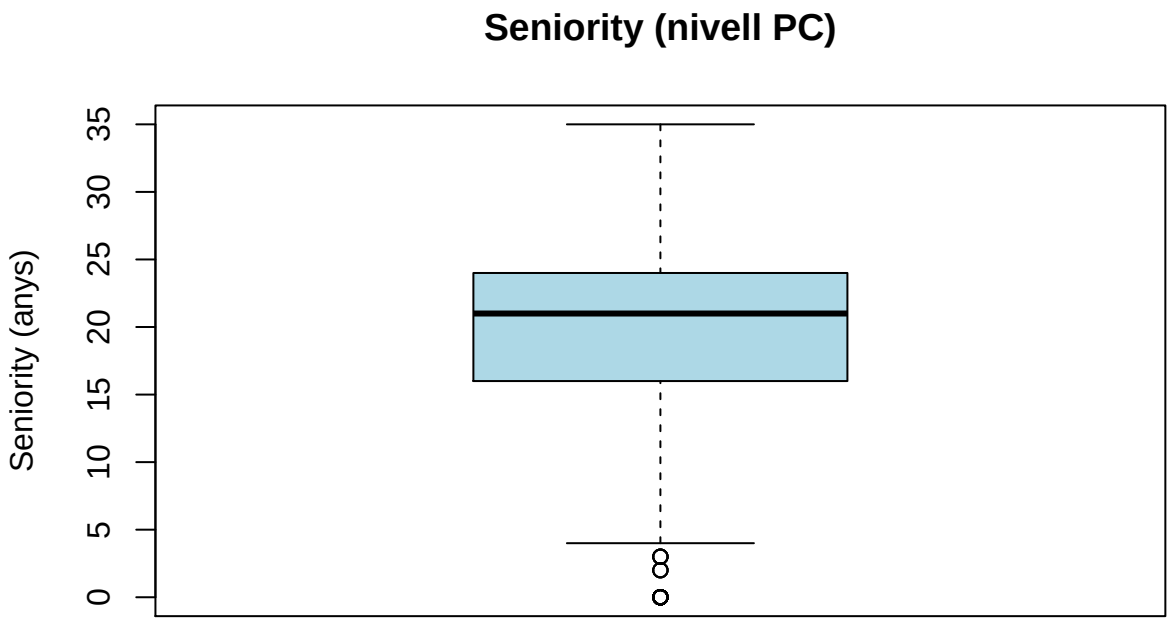
```
satisfaccioLaboral.seniorityC <- satisfaccioLaboral[job_type=="C",]
satisfaccioLaboral.seniorityPC <- satisfaccioLaboral[job_type=="PC",]
boxplot(satisfaccioLaboral.seniorityC$seniority, main="Seniority (nivell C)",
   xlab="", ylab="Seniority (anys)", col="#ADD8E6")
```




```
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 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")
```




```
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, indistinguibles.

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

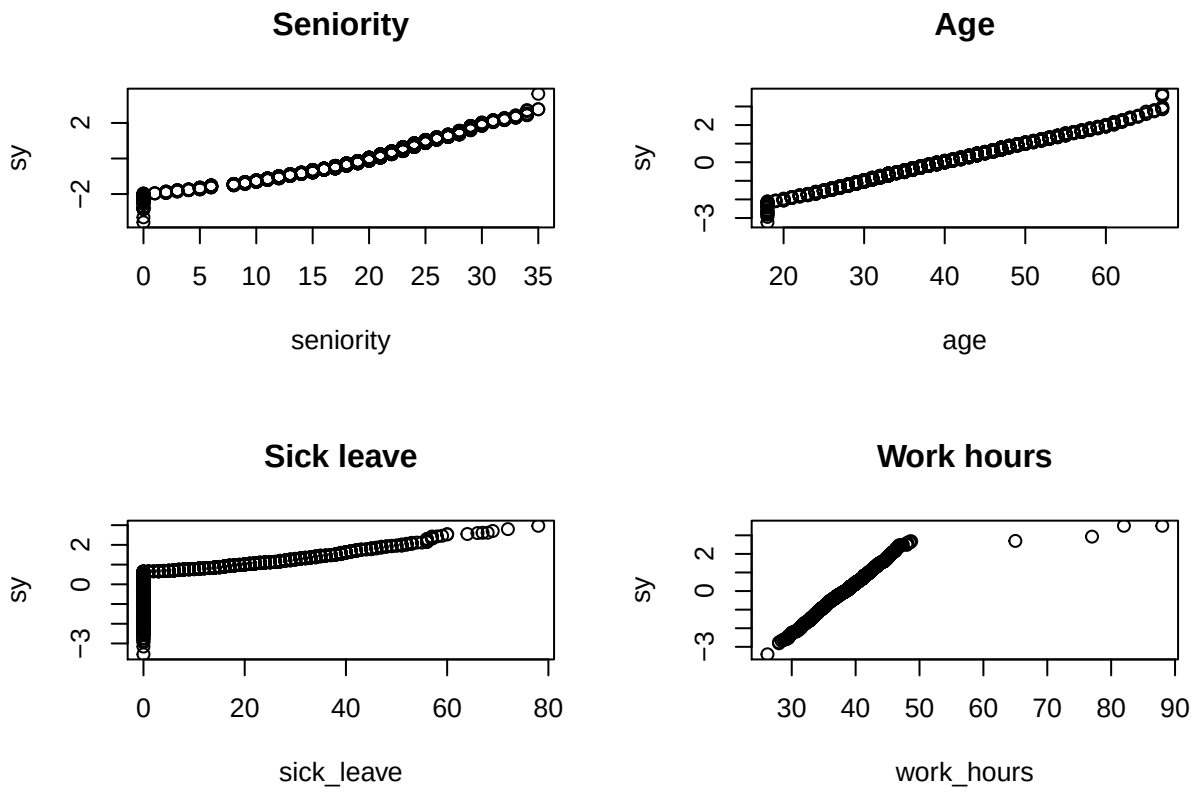
```
##
##  Wilcoxon rank sum test
##
## data:  satisfaccioLaboral.seniorityC$seniority and satisfaccioLaboral.seniorityPC$seniority
## W = 411070, p-value = 4.141e-05
## alternative hypothesis: true location shift is not equal to 0
```

Amb un número tan proper a 0 com p-value = 4.141e-05 podem afirmar que la hipòtesi nul · la és correcta.

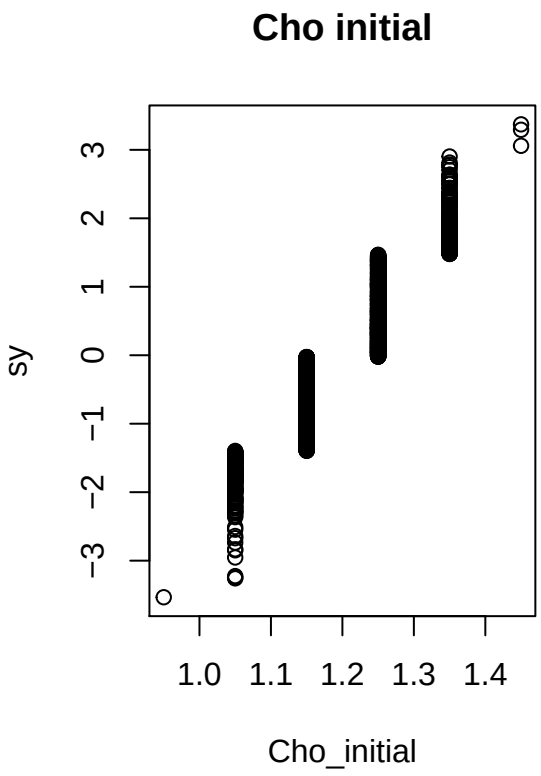
Assumpció de normalitat

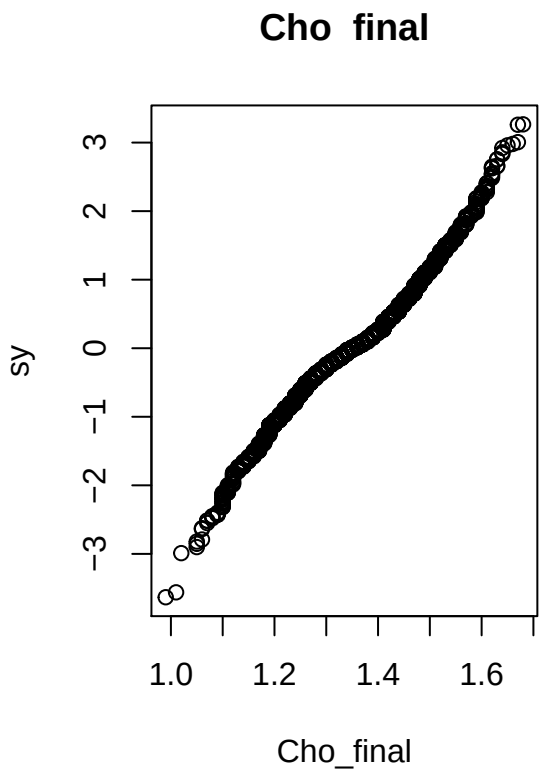
Tests de normalitat de les variables numèriques

```
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)
```

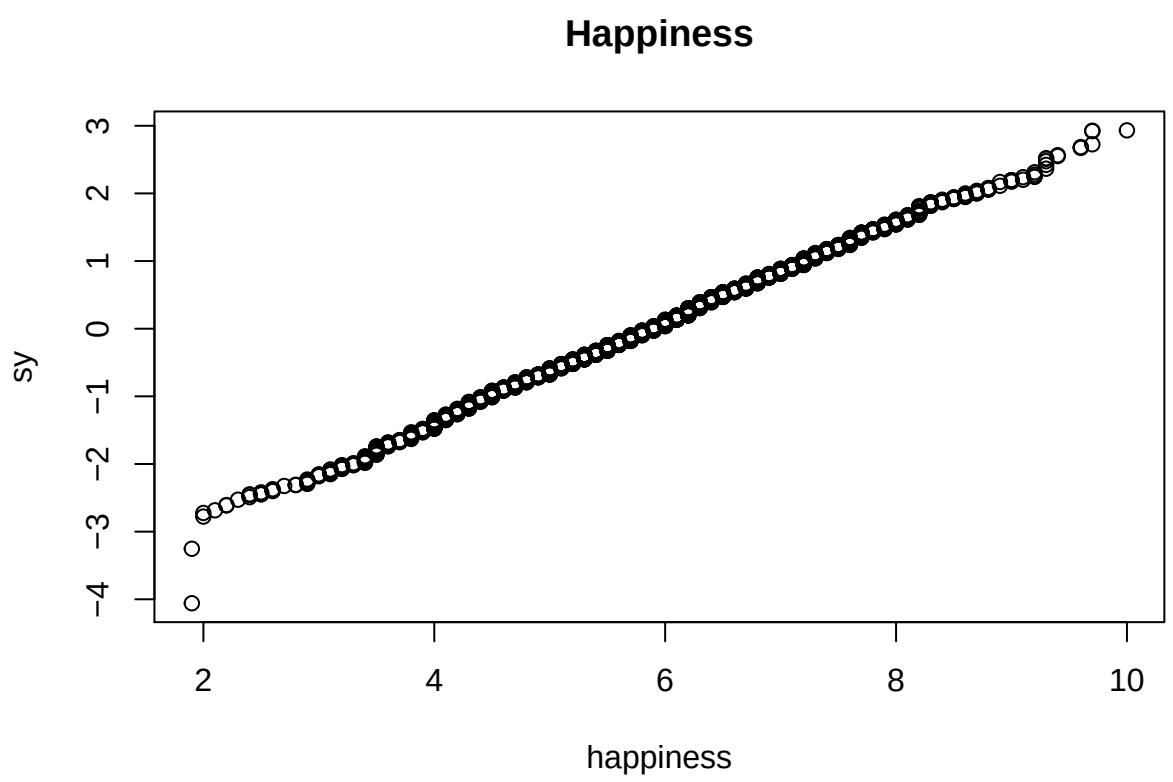



```r
par(mfrow=c(1,2))
qqplot(Cho_initial, rnorm(nrow(satisfaccioLaboral)), main="Cho initial", ylab = NULL)
qqplot(Cho_final, rnorm(nrow(satisfaccioLaboral)), main="Cho  final", ylab = NULL)
```






```
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 corva 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
```

```r
shapiro.test(sick_leave)
```

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

```r
shapiro.test(work_hours)
```

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

```r
shapiro.test(Cho_initial)
```

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

```r
shapiro.test(Cho_final)
```

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

```r
shapiro.test(happiness)
```

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