A2: Analítica Descriptiva i Inferencial

Proposta de solució

Semestre 2018.2

Contents

1 Introducció

Les dades que s'analitzen en aquesta activitat corresponen a l'anàlisi de satisfacció laboral treballat en l'activitat 1. L'objectiu és investigar la satisfacció laboral de la població en relació a les altres variables.

L'arxiu es denomina *rawData_clean.csv*, i conté 1920 registres i 12 variables. Aquestes variables són: city, sex, educ_level, job_type, happiness, age, seniority, sick_leave, sick_leave_b, work_hours, Cho_initial, Cho_final

Les dades recollides per a cada treballador són:

1. La ciutat del lloc de treball
2. El sexe del treballador

3. El nivell d'estuis del treballador (1: sense estudis, 2: estudis primaris, 3: educació secundària o educació professional, 4: universitaris)
4. El tipus de treball (C: qualificat, PC: poc qualificat)
5. La satisfacció laboral del treballador (compresa entre 0 i 10)
6. L'edat (anys)
7. L'antiguitat (en anys)
8. Els dies de baixa en el darrer any laboral
9. Si el treballador va estar de baixa o no en el darrer any (variable binària)
10. Les hores que treballa a la setmana en promig
11. El nivell de colesterol en la penúltima revisió mèdica (mmol/L)
12. El nivell de colesterol en la darrera revisió mèdica (mmol/L).

Nota important a tenir en compte per a lliurar l'activitat:

- És necessari lliurar el fitxer Rmd i el fitxer de sortida (PDF o html). El fitxer de sortida ha d'incloure el codi i el resultat de la seva execució (pas a pas). S'ha de respectar la numeració dels apartats de l'enunciat.
- No realitzeu llistat dels conjunts de dades, donat que aquests poden ocupar varies pàgines. Si voleu comprovar l'efecte d'una instrucció sobre les dades, podeu usar la funció head que mostra les 10 primeres files del conjunt de dades.

2 Analítica descriptiva

2.1 Càrrega del fitxer

Carregueu l'arxiu de dades en R. Independentment del fitxer que vàreu obtenir en l'activitat 1, useu el fitxer "rawDataClean.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.

```
df <- read.csv( file="rawData_clean.csv")
dim( df )
```

```
## [1] 1920   12
```

```
str(df)
```

```
## 'data.frame':    1920 obs. of  12 variables:
##  $ city        : Factor w/ 3 levels "Barcelona","Madrid",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex         : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ educ_level  : Factor w/ 4 levels "N","P","S","U": 1 1 1 1 1 1 1 1 1 1 ...
##  $ job_type    : Factor w/ 2 levels "C","PC": 1 1 1 1 1 1 1 1 1 1 ...
##  $ happiness   : num  7.2 5.9 7 6.3 8.9 7.2 6.1 9.2 7.6 5.7 ...
##  $ age         : int  45 36 34 37 35 29 26 44 23 46 ...
##  $ seniority   : int  23 18 16 18 17 11 8 23 5 22 ...
##  $ sick_leave  : int  0 0 0 0 0 6 0 0 24 36 ...
##  $ sick_leave_b: int  0 0 0 0 0 1 0 0 1 1 ...
##  $ work_hours  : num  35.4 41.3 36.6 40.5 36 ...
##  $ Cho_initial : num  1.15 1.05 1.15 1.15 1.15 1.05 1.15 1.15 1.15 1.15 ...
##  $ Cho_final   : num  1.33 1.1 1.25 1.21 1.31 1.18 1.2 1.28 1.33 1.21 ...
```

2.2 Anàlisi descriptiva visual

Representeu de manera gràfica les variables del conjunt de dades, de manera que es pugui copsar de manera visual i resumida els valors i distribucions de les variables del conjunt de dades. Escolliu els gràfics més apropiats en cada cas.

```
library(ggplot2)

#grouping data to show pie charts
library(dplyr)
```

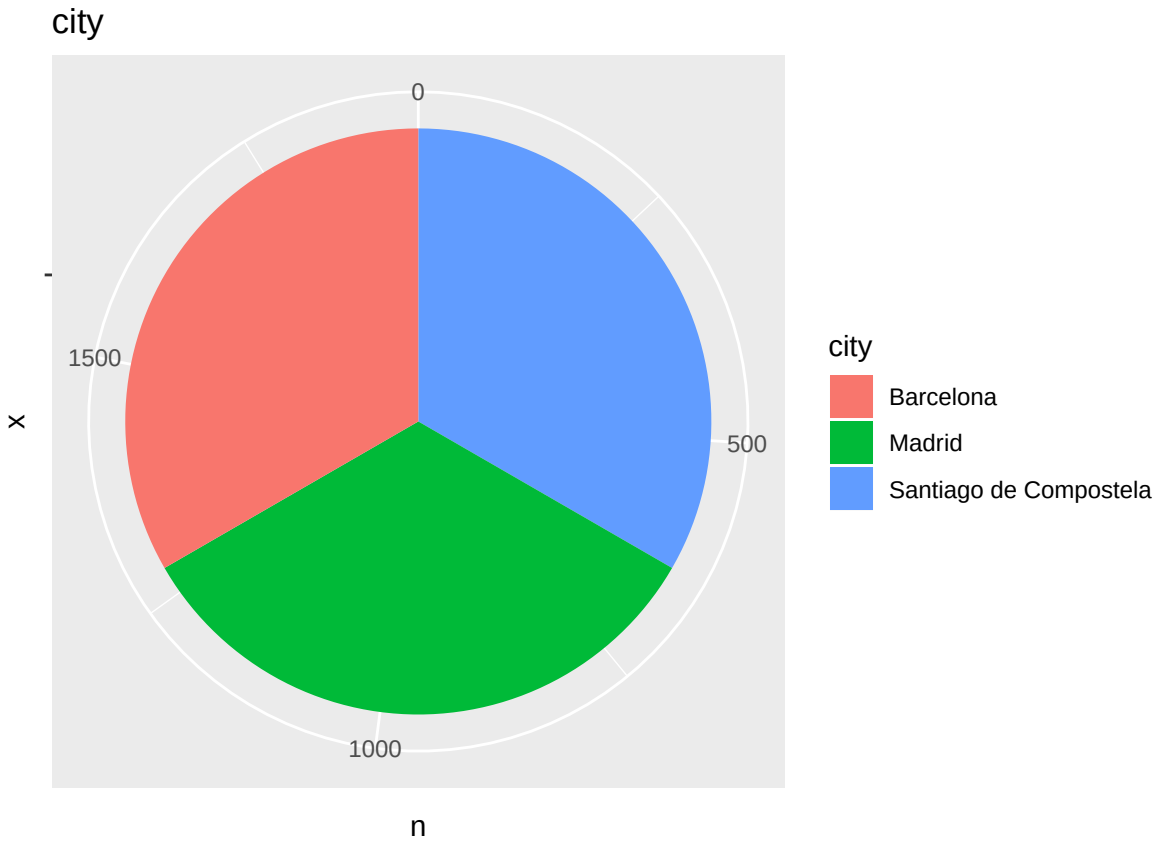
```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:kableExtra':
##
##     group_rows
```

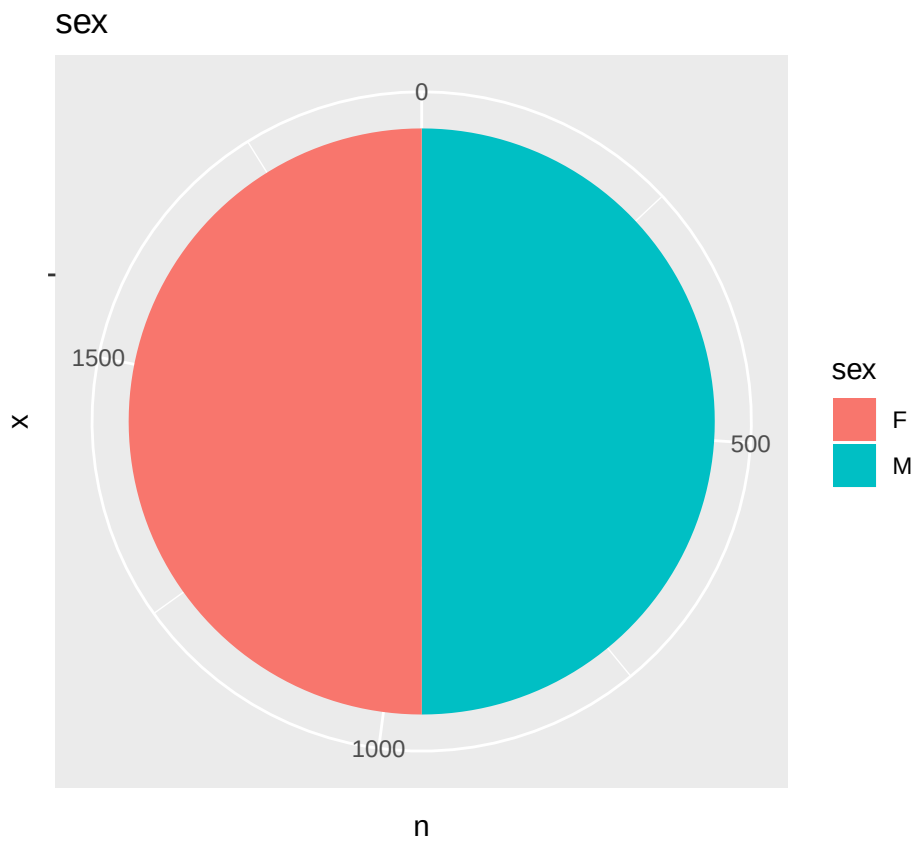
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

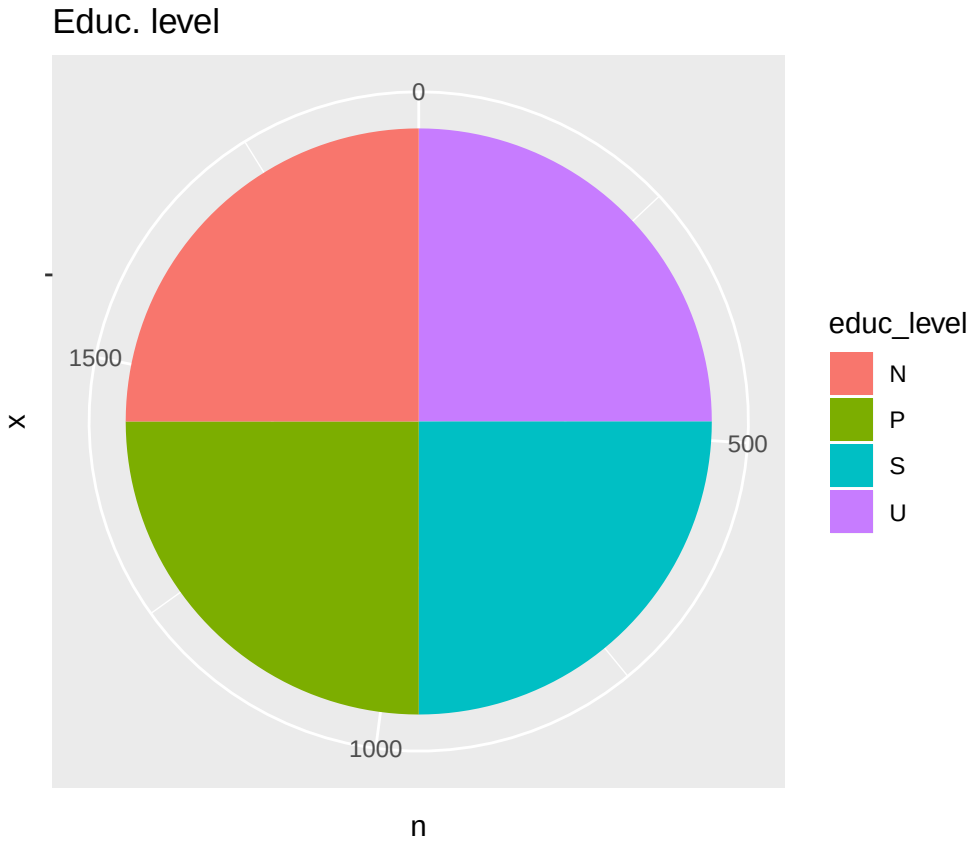
```
citysum <- summarize( group_by(df, city), n=length(city))
ggplot( citysum, aes(x="", y=n, fill=city)) +
  geom_bar(width = 1, stat = "identity") +
    coord_polar("y", start=0) + ggtitle("city")
```



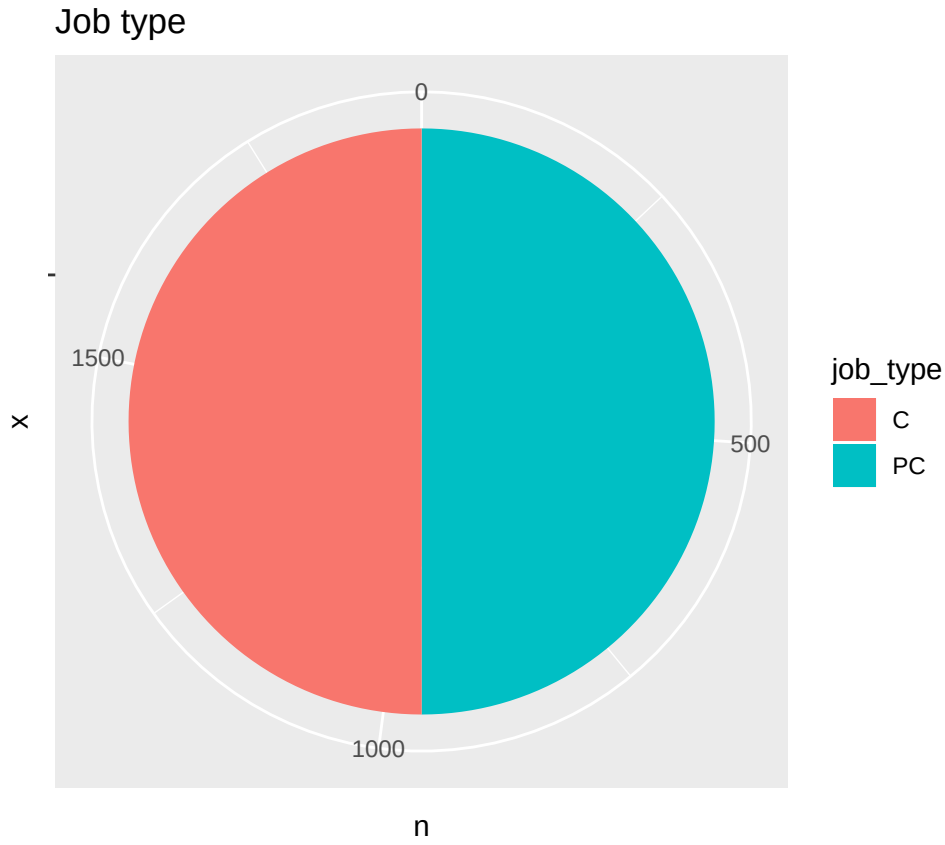

```
sxsum <- summarize( group_by(df, sex), n=length(sex))
ggplot( sxsum, aes(x="", y=n, fill=sex)) +
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) + ggtitle("sex")
```



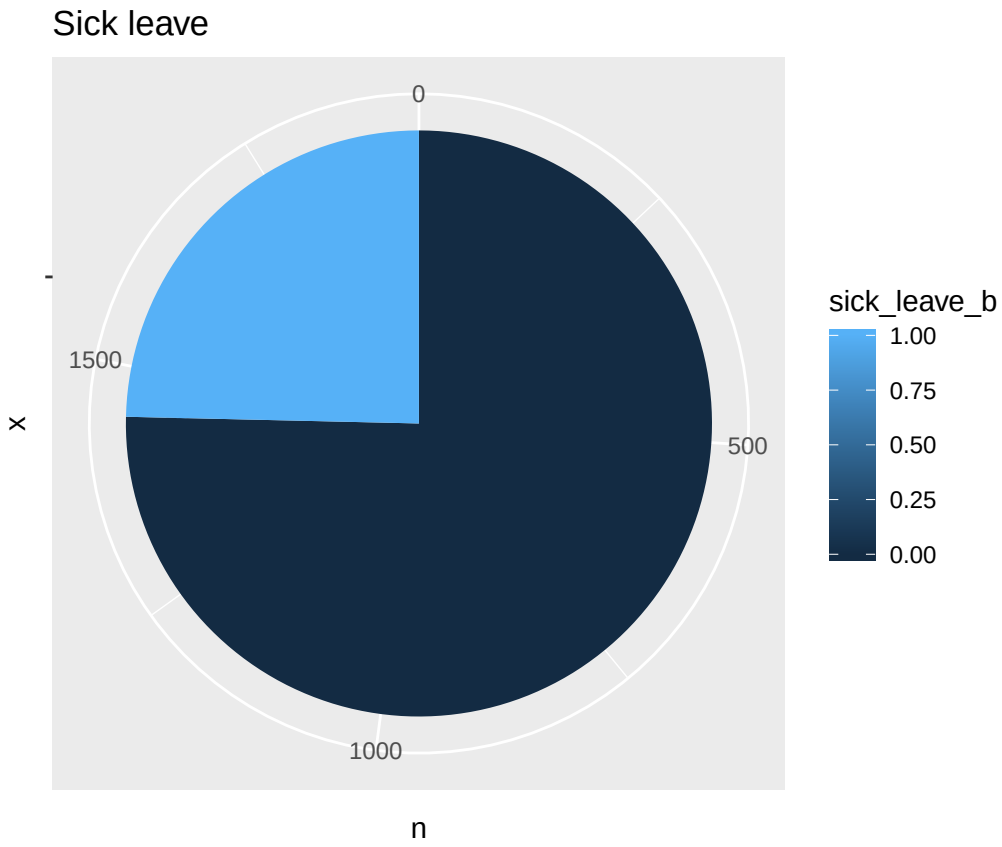

```
edsum <- summarize( group_by(df, educ_level), n=length(educ_level))
ggplot(edsum, aes(x="", y=n, fill=educ_level))+
  geom_bar(width = 1, stat = "identity") +
  coord_polar("y", start=0) + ggtitle("Educ. level")
```




```
jtsum <- summarize( group_by(df, job_type), n=length(job_type))
ggplot(jtsum, aes(x="", y=n, fill=job_type))+
  geom_bar(width = 1, stat = "identity") + ggtitle("Job type")+
  coord_polar("y", start=0)
```




```
slsum <- summarize( group_by(df, sick_leave_b), n=length(sick_leave_b))
ggplot(slsum, aes(x="", y=n, fill=sick_leave_b))+
  geom_bar(width = 1, stat = "identity") + ggtitle("Sick leave")+
  coord_polar("y", start=0)
```

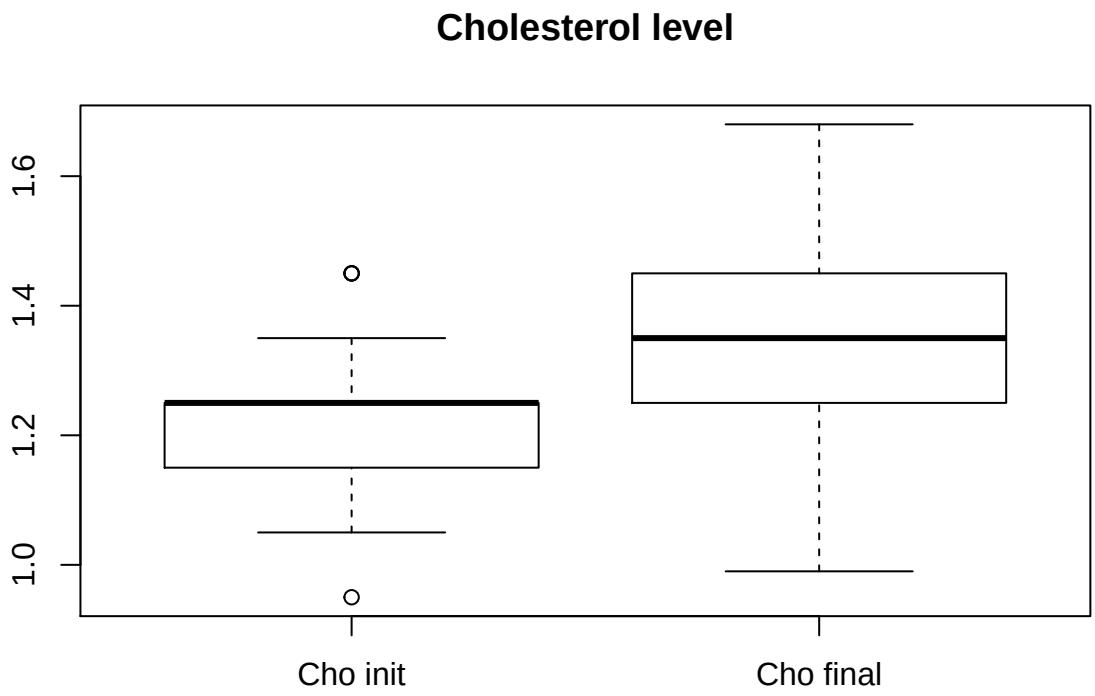



```
#---
par(mfrow=c(1,5))

boxplot(df$happiness, main="Happiness")
boxplot(df$age, main="Age")
boxplot(df$seniority, main="Seniority")
boxplot(df$work_hours, main="Work hours")
boxplot(df$sick_leave, main="Sick leave")
```


```
par(mfrow=c(1,1))

boxplot( df$Cho_initial, df$Cho_final, names=c("Cho init", "Cho final"), main="Cholesterol level" )
```



3 Estadística inferencial

3.1 Intervals de confiança

Representeu, si no ho heu fet en la secció anterior, la distribució de valors de la variable happiness en un histograma. Interpreteu el gràfic.

```
hist(df$happiness)
```

Interpretació: El valor de happiness segueix una distribució normal amb valors entre 1 i 10, centrat en la mitjana de 6.

A continuació, calculeu l'interval de confiança del 90% de la satisfacció laboral (happiness) dels treballadors. També, expliqueu quina és la interpretació d'interval de confiança de la variable happiness.

Nota: S'han de realitzar els càlculs manualment. No es poden usar funcions d'R que calculin directament l'interval de confiança com t.test o similar. Sí que podeu usar funcions com qnorm, pnorm, qt i pt.

```
#Càlcul INTERVAL DE CONFIANÇA
n<-nrow( df )
alpha<-1-0.90
#Error típic
errorTipic <- sd(df$happiness) / sqrt( n )
errorTipic
```

```
## [1] 0.03075509
```

```
#Valor z
z<-qnorm( 1-alpha/2 )
z
```

```
## [1] 1.644854
```

```
#Marge d'error
error<- z * errorTipic
error
```

```
## [1] 0.05058762
```

```r
#Interval
c( mean(df$happiness) - error, mean(df$happiness) + error )
```

```
## [1] 5.832485 5.933661
```

```r
#Comprovació amb t Student.
#No dona exactament igual perquè s'ha assumit distribució normal (anteriorment).
t.test( df$happiness, conf.level=0.90 )
```

```
##
##  One Sample t-test
##
## data:  df$happiness
## t = 191.29, df = 1919, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  5.832461 5.933685
## sample estimates:
## mean of x
##  5.883073
```

Interpretació: L'interval de confiança obtingut de la variable happiness és un estimat de l'interval que pot contenir el valor real del paràmetre happiness promig de la població. El nivell de confiança del 90% representa que el 90% percentatge d'intervals que es podrien construir a partir d'infinites mostres de la població contindrien el valor real del happiness promig.

3.2 Satisfacció en el treball en relació al sexe

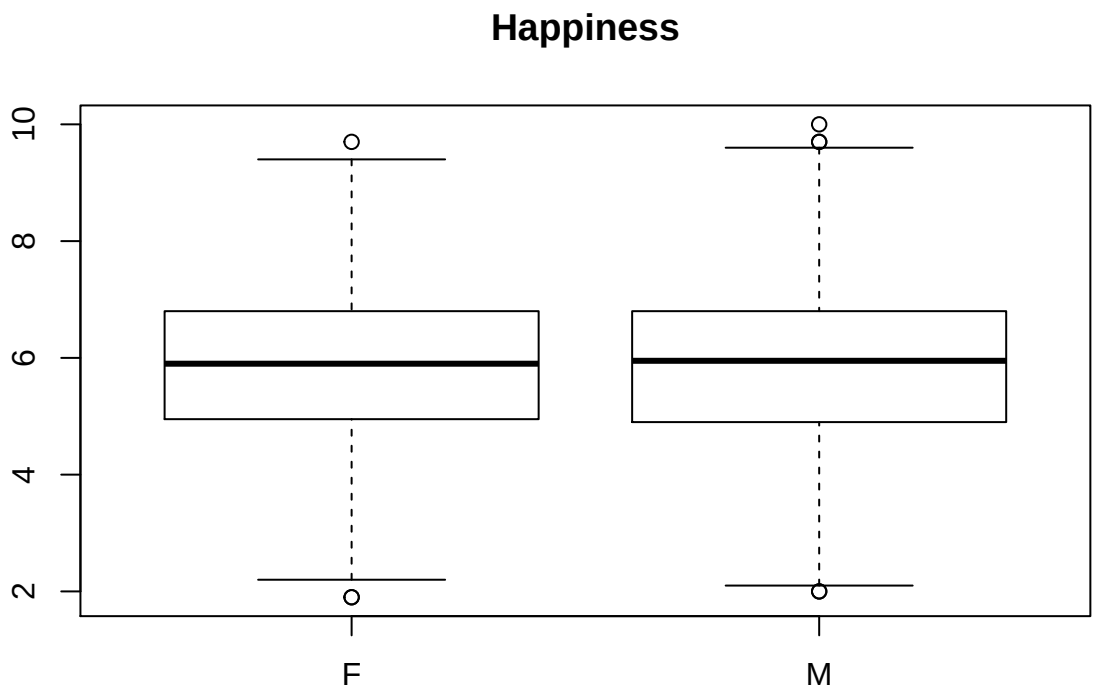
Ens preguntem si existeixen diferències significatives en el grau de felicitat (happiness) dels homes en relació a les dones. Per a respondre a aquesta pregunta, seguiu els passos que es detallen a continuació.

Nota: S'han de realitzar tots els càlculs manualment. No es poden usar funcions R que calculin directament el contrast com t.test o similar. Sí que podeu usar funcions com: **qnorm**, **pnorm**, **qt** i **pt**.

3.2.1 Boxplot

Mostreu la distribució de happiness d'homes i dones per separat en un boxplot. Expliqueu si considereu que hi ha diferències apreciables visualment.

```r
boxplot( happiness~sex, df, main="Happiness")
```



Interpretació: No s'aprecien diferències gaire notables entre la felicitat de dones i homes en el treball.

3.2.2 Escriure la hipòtesi nul · la i alternativa

Escriviu la hipòtesi nul · la i alternativa tenint en compte que es vol testejar si el nivell de satisfacció (happiness) és diferent en homes i dones.

$H0 : \mu_F = \mu_M$

$H1 : \mu_F \neq \mu_M$

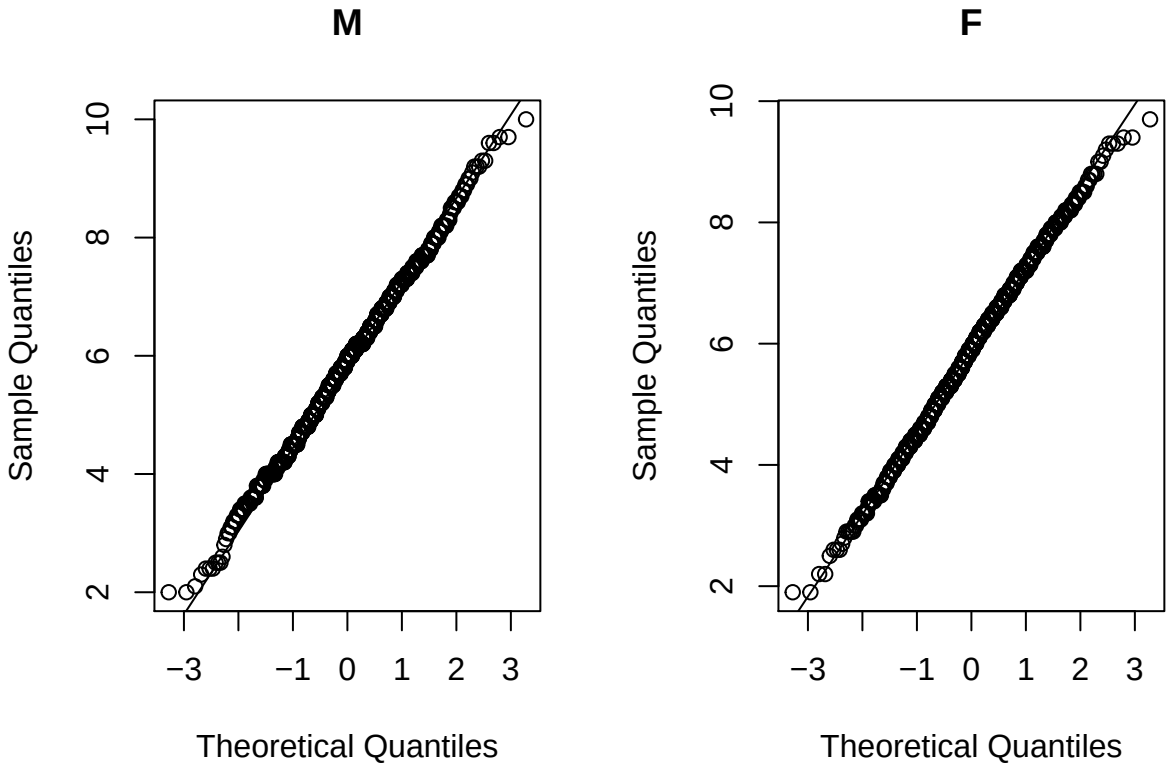
3.2.3 Mètode

Indiqueu quin és el mètode més apropiat per a fer aquesta anàlisi, en funció de les característiques de la mostra i l'objectiu de l'anàlisi.

```
#Comprovem si es compleix normalitat a les mostres
#hist(df$happiness[df$sex=="M"], breaks=15, main="Men's Happiness")
#hist(df$happiness[df$sex=="F"], breaks=15, main="Women's Happiness")

par(mfrow=c(1,2))
qqnorm(df$happiness[df$sex=="M"], main="M")
qqline(df$happiness[df$sex=="M"])

qqnorm(df$happiness[df$sex=="F"], main="F")
qqline(df$happiness[df$sex=="F"])
```




```
par(mfrow=c(1,1))

#Test de normalitat
#H0: la mostra segueix una distribució normal
shapiro.test(df$happiness[df$sex=="M"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$happiness[df$sex == "M"]
## W = 0.99759, p-value = 0.1716
```

```
shapiro.test(df$happiness[df$sex=="F"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$happiness[df$sex == "F"]
## W = 0.99779, p-value = 0.2316
```

Si el valor p del test Shapiro Wilk és superior a 0.05, implica que la distribució de les dades no és significativament diferent de la distribució normal. És a dir, podem assumir normalitat.

Segons el test Shapiro Wilk i el plot Q-Q, podem assumir normalitat. Per tant, apliquem un test de dues mostres independents per a la diferència de les mitjanes. Apliquem el cas de dades normals, amb variança desconeguda. El test és bilateral (dues cues).

Per a decidir si apliquem variàncies iguals o diferents, apliquem el test F. Podeu trobar la informació a: http://www.sthda.com/english/wiki/f-test-compare-two-variances-in-r

```r
#Test d'igualtat de variances
#H0: F (rati de variàncies) = 1
#H0 : F diferent d'1
var.test(df$happiness[df$sex=="M"], df$happiness[df$sex=="F"], alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  df$happiness[df$sex == "M"] and df$happiness[df$sex == "F"]
## F = 0.98815, 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:
##  0.8705985 1.1215799
## sample estimates:
## ratio of variances
##          0.9881527
```

El resultat del test F és que podem assumir igualtat de variàncies. Per tant, apliquem test t de dues mostres independents per a la diferència de mitjanes, variàncies desconegudes i iguals. El test és bilateral.

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

Realitzeu els càlculs amb un nivell de confiança del 95%.

```r
#
F <- df[df$sex=='F',]
nF<-nrow( F )
meanF<-mean(F$happiness)
sdF <- sd( F$happiness)
#
M <- df[df$sex=='M',]
nM<-nrow( M )
meanM<-mean(M$happiness)
sdM <- sd( M$happiness)
#

s <-sqrt( ((nF-1)*sdF^2 + (nM-1)*sdM^2 )/(nF+nM-2) )
Sb <- s*sqrt(1/nF + 1/nM)
t <- (meanF-meanM)/ Sb
t
```

```
## [1] 0.0186235
```

```r
alfa <- (1-0.95)
tcritical <- qt( alfa/2, df=nF+nM-2, lower.tail=FALSE )

#two-sided
pvalue<-pt( abs(t), df=nF+nM-2, lower.tail=FALSE )*2

#Print info
info<-data.frame(nF,meanF,sdF,nM,meanM,sdM,t,tcritical,pvalue)
info %>%   kable() %>%   kable_styling()
```

nF	meanF	sdF	nM	meanM	sdM	t	tcritical	pvalue
960	5.883646	1.351982	960	5.8825	1.343949	0.0186235	1.961202	0.9851434

```r
#Comprovació t-test:
t.test( F$happiness, M$happiness, conf.level=0.95, paired=FALSE,
        var.equal=TRUE, alternative="two.sided" )
```

```
##
##  Two Sample t-test
##
## data:  F$happiness and M$happiness
## t = 0.018623, df = 1918, 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 of x mean of y
##  5.883646  5.882500
```

3.2.5 Interpretar el resultat

El valor de p és 0.9851434 i no és inferior a alfa=0.05 i per tant, no podem rebutjar la hipòtesi nul · la de que el nivell de satisfacció laboral és el mateix en les dones que en els homes. De manera anàloga, el valor t observat no és superior al valor crític. I per tant, estem a la zona d'acceptació de la hipòtesi nul.la.

Així doncs, tal com mostrava el boxplot, no s'observen diferències significatives entre homes i dones en quant a satisfacció laboral.

3.3 Test no paramètric

Quan les assumpcions de normalitat no es compleixen, cal aplicar tests no paramètrics. El **test de suma de rangs de Wilcoxon**, també anomenat **prova U de Mann-Whitney** és l'equivalent no paramètric al test t per a dues mostres independents. En aquest apartat us demanem que apliqueu aquest test. Es realitzarà per a contrastar si hi ha diferències en l'antiguitat (seniority) en funció del tipus de treball (job_type).

Podeu consultar:

- https://www.r-bloggers.com/wilcoxon-mann-whitney-rank-sum-test-or-test-u/

3.3.1 Hipòtesi nul · la i alternativa

Escriviu la hipòtesi nul · la i l'alternativa. Recordeu que la pregunta és si hi ha diferències significatives en seniority segons el tipus de treball.

$H0 : \mu_C = \mu_{PC}$

$H1 : \mu_C \neq \mu_{PC}$

3.3.2 Assumpció de normalitat

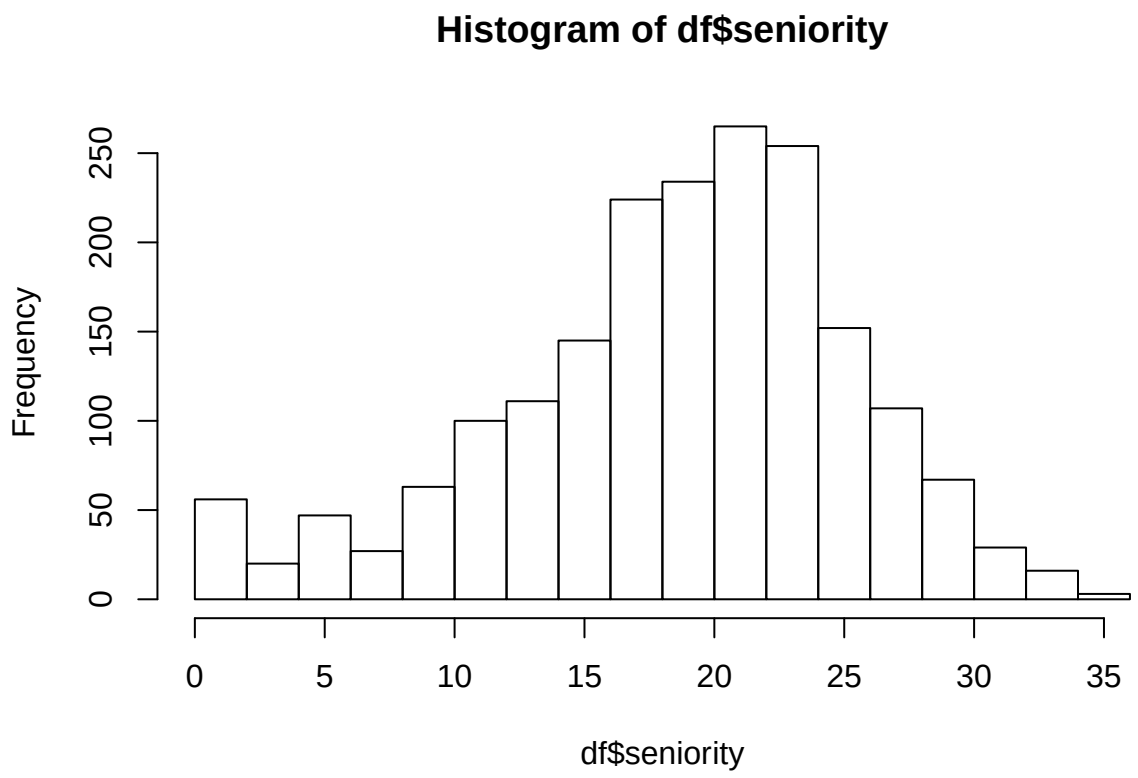
Comproveu si es compleix l'assumpció de normalitat en les dades. Per a fer-ho, podeu mostrar els gràfics **Q-Q plots** i aplicar a continuació el test Shapiro-Wilk, interpretant els resultats.

Per a realitzar aquest apartat, podeu consultar la informació següent:

- Quantile-Quantile Plots: https://www.rdocumentation.org/packages/stats/versions/3.5.3/topics/qqnorm

- Interpretació dels Q-Q plots: https://data.library.virginia.edu/understanding-q-q-plots/
- Test Shapiro-Wilk: https://es.wikipedia.org/wiki/Test_de_Shapiro
- Funció shapiro.test d'R: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/shapiro.test.html

```
hist(df$seniority,breaks=20)
```

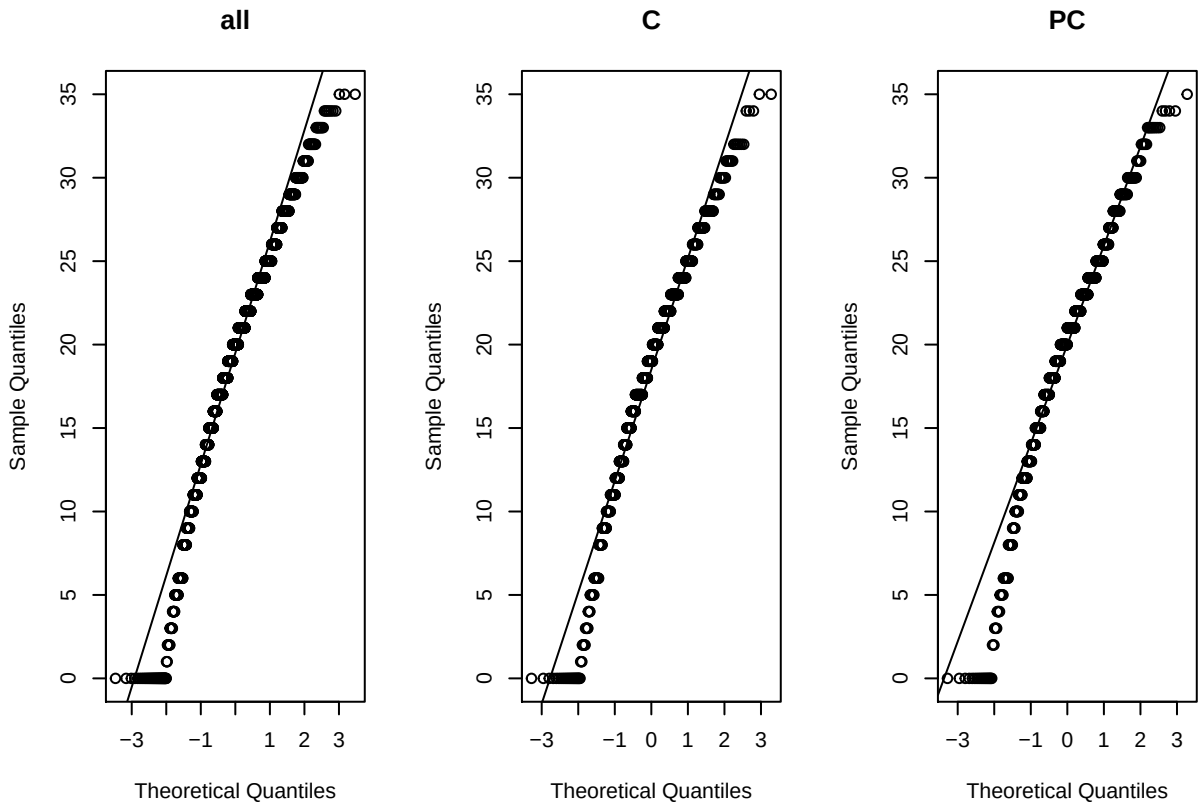



```
df_t1 <- df[df$job_type=="C",]
df_t2 <- df[df$job_type=="PC",]

par(mfrow=c(1,3))
qqnorm( df$seniority,main="all")
qqline(df$seniority)


#t1
qqnorm( df_t1$seniority,main="C")
qqline(df_t1$seniority)

#t2
qqnorm( df_t2$seniority,main="PC")
qqline(df_t2$seniority)
```




```
par(mfrow=c(1,1))

#
shapiro.test(df$seniority)
```

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

```
shapiro.test(df_t1$seniority)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_t1$seniority
## W = 0.9701, p-value = 3.775e-13
```

```
shapiro.test(df_t2$seniority)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_t2$seniority
## W = 0.96711, p-value = 6.565e-14
```

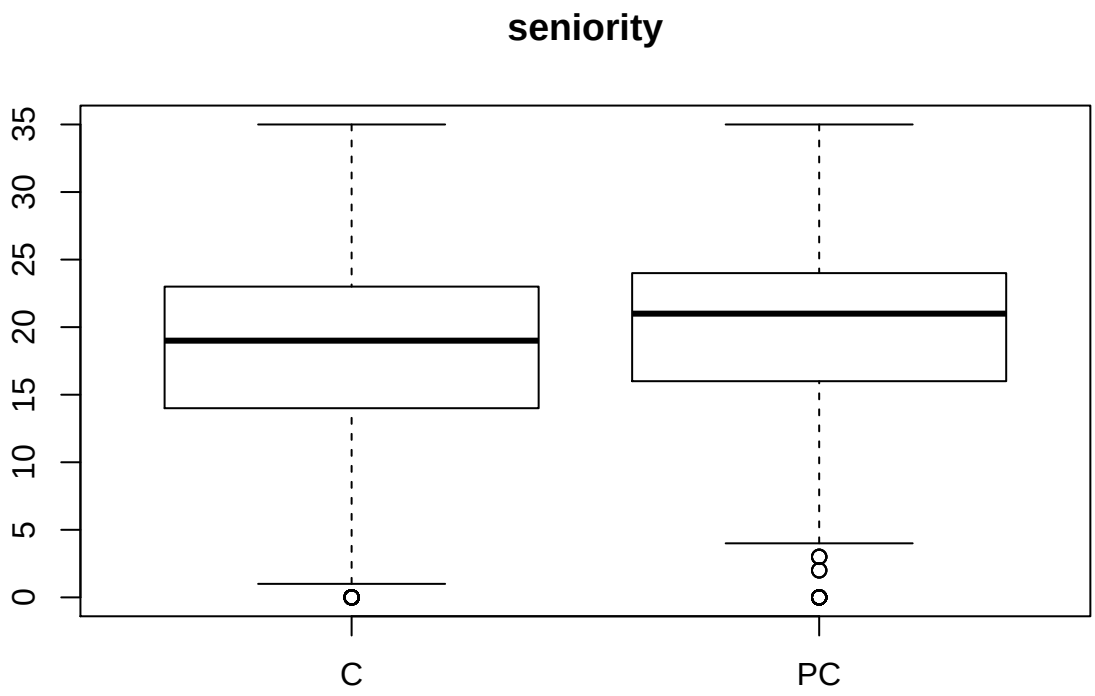
Interpretació: tant els gràfics Q-Q com el test de Shapiro proven que no podem assumir normalitat en les

dades.

3.3.3 Test U de Mann-Whitney

Apliqueu el test U de Mann-Whitney per a testejar si hi ha diferències significatives en l'antiguitat en funció del tipus de treball. Podeu usar la funció **wilcox.test**.

```
boxplot( df$seniority[df$job_type=="C"], df$seniority[df$job_type=="PC"],
         names=c("C","PC"), main="seniority")
```




```
wilcox.test(df$seniority[df$job_type=="C"], df$seniority[df$job_type=="PC"],
              alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  df$seniority[df$job_type == "C"] and df$seniority[df$job_type == "PC"]
## W = 411070, p-value = 4.141e-05
## alternative hypothesis: true location shift is not equal to 0
```

3.3.4 Càlculs manuals

Per tal d'aprofundir en les operacions del test de suma de rangs de Wilcoxon, realitzeu els càlculs de la suma de ranks del test U de Mann-Whitney i realitzeu el contrast. Podeu consultar l'explicació pas a pas de les fórmules a: https://www.monografias.com/trabajos106/prueba-u-mann-whitney-dos-muestras-independientes/prueba-u-mann-whitney-dos-muestras-independientes.shtml

```r
typeQ  <- df[df$job_type=="C",]
typePQ <- df[df$job_type=="PC",]
nQ <- nrow( typeQ )
nPQ <- nrow( typePQ )

ranks <- rank(c(typeQ$seniority, typePQ$seniority))
length(ranks)
```

```
## [1] 1920
```

```r
sumrank.Q = sum( ranks[1:nQ] )
sumrank.Q
```

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

```r
U1 <- nQ*nPQ + (nQ*(nQ+1))/2 - sumrank.Q

sumrank.PQ = sum( ranks[(nQ+1):length(ranks)])
sumrank.PQ
```

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

```r
U2 = nQ*nPQ + (nPQ*(nPQ+1))/2 - sumrank.PQ

c(U1,U2)
```

```
## [1] 510530.5 411069.5
```

```r
U<-min(U1,U2)
U
```

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

```r
#Contrast
mu=nQ*nPQ/2
Su <- sqrt( (nQ*nPQ*(nQ+nPQ+1))/12 )
z<-(U-mu)/Su
z
```

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

```r
pvalue<-pnorm(z,lower.tail=TRUE)*2  #multipliquem X2 per ser bilateral
pvalue
```

```
## [1] 4.234652e-05
```

```r
#Print info
info<-data.frame(U1,U2,U,Su,z,pvalue)
info %>%   kable() %>%   kable_styling()
```

U1	U2	U	Su	z	pvalue
510530.5	411069.5	411069.5	12146.31	-4.094289	4.23e-05

```r
##Considerant el nombre d'empats i fent el càlcul precís
CalculEmpats <- function( ranks ){
  ranks.sorted <- sort( ranks )
  unique.ranks <- sort( unique( ranks ) )
  Lii<-0
```

```
  for (rnk in (unique.ranks)){
    print(rnk)
    idx <- grep( as.character(rnk), as.character(ranks.sorted) )
    length(idx)
    Lii<- c( Lii, length(idx) )
  }
  return (Lii)
}

Lis <- CalculEmpats( ranks )
```

```
## [1] 22.5
## [1] 46
## [1] 52
## [1] 61.5
## [1] 71.5
## [1] 85.5
## [1] 109
## [1] 137
## [1] 164.5
## [1] 196
## [1] 236
## [1] 286
## [1] 343
## [1] 398.5
## [1] 463.5
## [1] 536
## [1] 628
## [1] 740
## [1] 847
## [1] 964
## [1] 1099.5
## [1] 1232
## [1] 1358.5
## [1] 1485.5
## [1] 1592.5
## [1] 1668.5
## [1] 1724.5
## [1] 1778
## [1] 1824
## [1] 1857.5
## [1] 1880
## [1] 1894.5
## [1] 1906
## [1] 1914
## [1] 1919
```

```
Lis
```

```
##  [1]   0  44  81   9  10  10 140 173  27  28  35  45  55  59  52  78  67
## [18] 117 107 107 127 144 121 132 122  92  60  52  55  37  30  15  14   9
## [35]   7   3
```

```r
sum(Lis)
```

```
## [1] 2264
```

```r
N <-nQ+nPQ
sum.Lis <- sum( Lis^3 - Lis ) / 12
Factor_w_Li <- ((N^3-N)/12 - sum.Lis )

Su <- sqrt( (nQ*nPQ)/ (N*(N-1)) * (Factor_w_Li) )
Su
```

```
## [1] 12123.67
```

```r
z<-(U-mu)/Su
z
```

```
## [1] -4.101935
```

```r
pvalue<-pnorm(z,lower.tail=TRUE)*2  #multipliquem X2 per ser bilateral
pvalue
```

```
## [1] 4.097085e-05
```

```r
#Hi ha poca diferència amb la correcció pel nombre d'empats
info<-data.frame(U1,U2,U,Su,z,pvalue)
info %>%   kable() %>%   kable_styling()
```

U1	U2	U	Su	z	pvalue
510530.5	411069.5	411069.5	12123.67	-4.101935	4.1e-05

3.3.5 Interpretació

Interpreteu el resultat del test.

El valor p del test de suma de rangs de Wilcoxon és inferior a 0.05. Per tant, podem rebutjar la hipòtesi nul · la. Així doncs, concloem que hi ha diferències en l'antiguitat en funció del tipus de treball.

3.3.6 Reflexió sobre tests paramètrics i no paramètrics

Si els tests no paramètrics no realitzen assumpcions sobre la normalitat de les dades, i per tant, són menys restrictius en la seva aplicabilitat, per què no s'usen els tests no paramètrics com a primera opció (o opció per defecte)?

Els tests no paramètrics tenen menys potència, és a dir, menys capacitat de detectar diferències quan aquestes existeixen.Per tant, sempre que es compleixin les assumpcions d'un test paramètric, és aconsellable usar-lo.

3.4 Test sobre el nivell de colesterol

A continuació ens preguntem si el nivell de colesterol dels treballadors ha augmentat en l'última revisió en relació a la penúltima revisió. Apliqueu un test adient per a realitzar aquest contrast.

Nota: S'han de realitzar els càlculs manualment. No es poden usar funcions d'R que calculin directament l'interval de confiança com t.test o similar. Sí que podeu usar funcions com qnorm, pnorm, qt i pt.

3.4.1 Hipòtesi nul · la i alternativa

Escriviu la hipòtesi nul · la i alternativa.

$H0 : \mu_f = \mu_i$

$H1 : \mu_f > \mu_i$

Aplicant un test aparellat, és equivalent a:

$H0 : \mu_{dif} = 0$

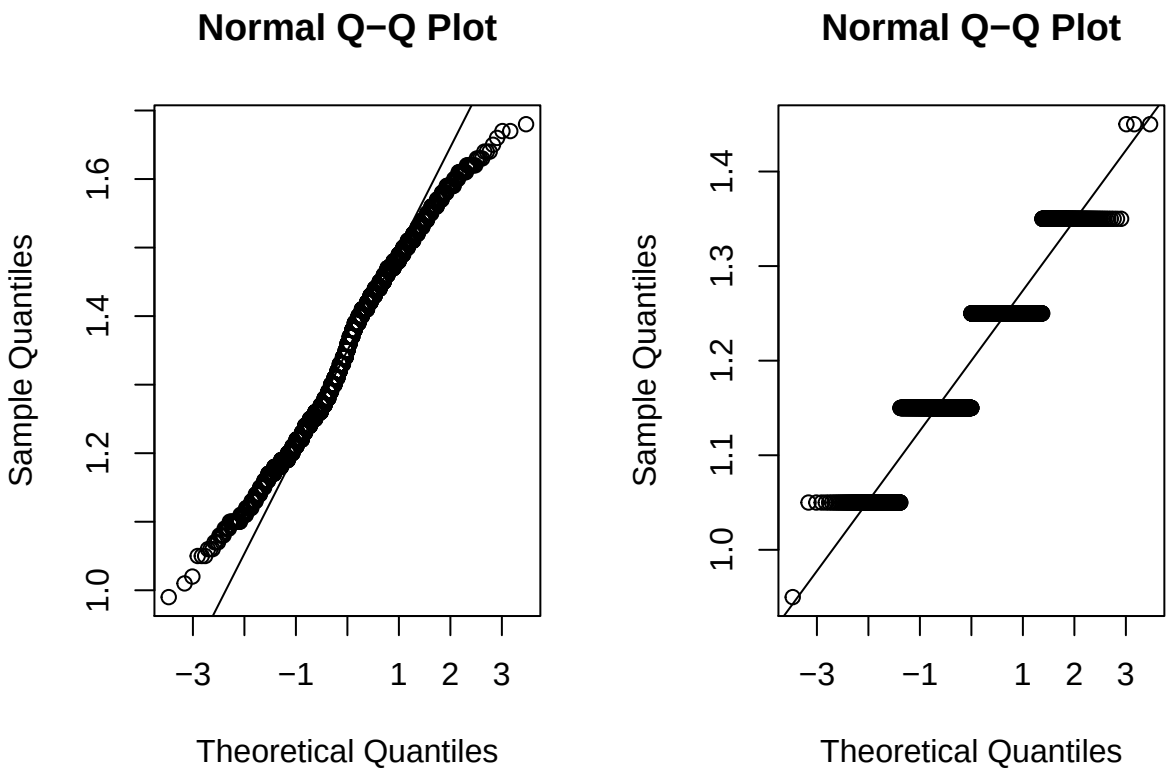
$H1 : \mu_{dif} > 0$

on dif=$Cho_f - Cho_i$.

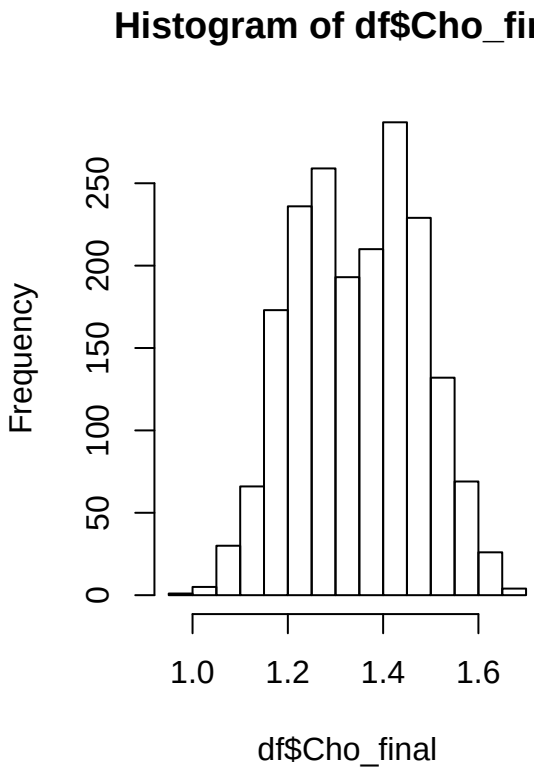
3.4.2 Mètode

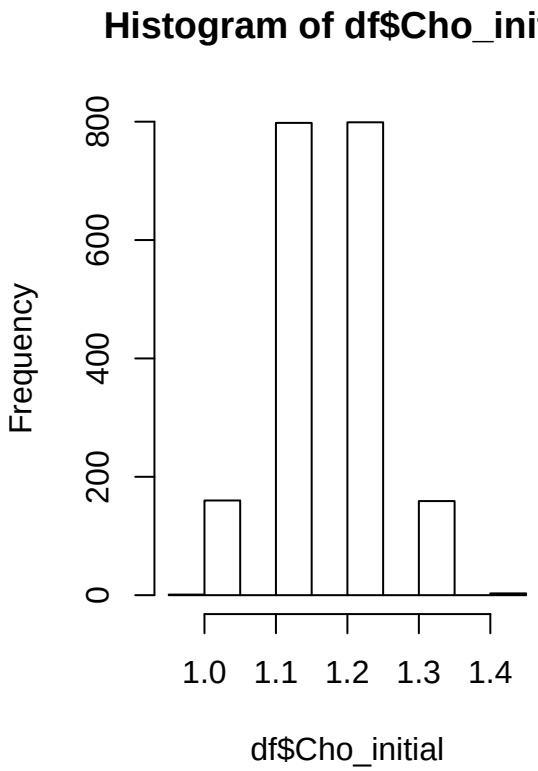
Escriviu el mètode que usareu per a realitzar el contrast i les assumpcions sobre les dades que es realitzen per a aplicar el mètode. Apliqueu si s'escau tets per a validar les assumpcions sobre les dades.

```
#Test de normalitat de les dades
 par(mfrow=c(1,2))
 qqnorm(df$Cho_final)
 qqline(df$Cho_final)
 qqnorm(df$Cho_initial)
 qqline(df$Cho_initial)
```



```r
hist(df$Cho_final)
hist(df$Cho_initial)
```






```r
par(mfrow=c(1,1))

shapiro.test( df$Cho_final )
```

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

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

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

No es compleixen les assumpcions de normalitat sobre les dades. Però com treballem amb una mida de mostres elevada n=1920, podem assumir que es compleix el teorema del límit central. Amb aquesta assumpció, apliquem un test t de mostres aparellades. Es considera també vàlid els que heu triat un test no paramètric com el test de rangs i signes de Wilcoxon, que és el test equivalent a la suma de rangs de Wilcoxon (test U Mann-Whitney) per a mostres aparellades.

3.4.3 Càlculs

Realitzeu tots els càlculs que s'escaiguin: estadístic de contrast, valor crític, valor p...

Apliquem el test de mostres aparellades.

```
dif <- df$Cho_final - df$Cho_initial

n<-length(dif)
mean = mean(dif)
mu0<-0

t<- (mean(dif)-mu0) / (sd(dif)/sqrt(n))
tcritical <- qt( 1-0.95, df=n-1, lower.tail=FALSE)
pvalue <- pt( t, df=n-1, lower.tail=FALSE)

info<-data.frame(n,mean,t,tcritical,pvalue)
info %>% kable() %>% kable_styling()
```

n	mean	t	tcritical	pvalue
1920	0.1511146	93.09562	1.645648	0

```
#Validació
t.test( dif, mu=0, alternative="greater" )
```

```
##
##  One Sample t-test
##
## data:  dif
## t = 93.096, df = 1919, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  0.1484433       Inf
## sample estimates:
## mean of x
## 0.1511146
```

```
#amb test de Wilcoxon
wilcox.test( df$Cho_final, df$Cho_initial, paired=TRUE, alternative="greater")
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  df$Cho_final and df$Cho_initial
## V = 1827700, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

3.4.4 Interpretació

Interpreteu el resultat.

El valor p és 0. Per tant, podem rebutjar la hipòtesi nul · la de que no hi ha diferències en el colesterol entre l'última prova i la penúltima, a favor de la hipòtesi alternativa. El test de Wilcoxon també permet rebutjat la hipòtesi nul · la.

4 Reflexió final

Elaboreu unes conclusions finals sobre l'anàlisi realitzat. Podeu parlar dels resultats obtinguts en aquesta anàlisi així com una reflexió més general sobre la idoneïtat dels tests i els tipus de tests que hem emprat. No cal excedir la mitja pàgina.

En aquesta activitat s'han treballat els tests d'hipòtesis de dues mostres. A continuació, es revisen els conceptes treballats:

Tests de dues mostres independents o aparellats El tests de dues mostres poden ser aparellats, com és el cas dels nivells de colesterol, en què el mateix individu és mesurat en dos moments diferents de temps. Les mostres independents (no aparellades) provenen de dues mostres diferents, com pot ser el cas de la satisfacció laboral de les dones i dels homes, o l'antiguitat dels empleats amb treball qualificat o poc qualificat.

Tests paramètrics i no paramètrics Els tests paramètrics realitzen assumpcions sobre les dades com és l'assumpció de normalitat. Si les dades segueixen una distribució normal, o es pot aplicar el teorema del límit central, es poden aplicar tests paramètrics. En cas contrari, cal aplicar tests no paramètrics. Per exemple, el test de Wilcoxon, per a mostres aparellades, o el test de suma de rangs de Wilcoxon (test U de Mann-Whitney) per a dues mostres independents.

En aquesta activitat, s'ha aplicat el test de suma de rangs de Wilcoxon a nivell il · lustratiu, però es podria haver assumit l'aplicació del teorema del TLC doncs la mida de la mostra és superior a 30.

Tests bilaterals o unilaterals Cal tenir en compte que si la hipòtesi alternativa testeja si una dada és superior o inferior a l'altra, el test s'ha de formular com a test unilateral (per la dreta o per l'esquerra, respectivament). Si la hipòtesi alternativa testeja la diferència, llavors el test és bilateral. Els càlculs del valor p canvien segons si el test és bilateral o unilateral.

Resultats observats Pel que fa als resultats de l'anàlisi, no hi ha diferències en la satisfacció laboral entre homes i dones. Es troben diferències en l'antiguitat en funció de si el treball és qualificat o poc qualificat. Caldria averiguar en quin sentit es produeix aquesta diferència. Per últim, el colesterol dels treballadors ha augmentat de manera significativa en el darrer any. Caldrà revisar el menjar del restaurant de l'empresa.

5 Puntuacions dels apartats

- Apartat 2 (10%)
- Apartat 3.1 (10%)
- Apartat 3.2 (20%)
- Apartat 3.3 (20%)
- Apartat 3.4 (20%)
- Apartat 4 (10%)
- Qualitat de l'informe dinàmic (10%)