# PAC4_Franco_Granell_Àlex

Àlex Franco Granell

Gener de 2023

## 1. Preporcessament

Carregue el fitxer proporcionat a l'AV i consulte el tipus de dades de cada atribut.

```
## 'data.frame':    253 obs. of  4 variables:
##  $ AE    : chr  "1.871878" "1.91312" "2.58114" "2.17827" ...
##  $ Tipo  : chr  "NF" "NF" "NF" "NF" ...
##  $ genero: chr  "M" "F" "M" "F" ...
##  $ edad  : int  54 60 40 55 59 63 62 62 26 48 ...
```

S'hi pot observar que els tipus de les dades és incorrecte per la capacitat pulmonar, per tant s'hauria de corregir. A més aprofitaré per convertir les variables de **Tipo** i **genero** com a factors.

A més, a través d'una inspecció visual de les dades he detectat que hi ha valors de **AE** amb comes en lloc de punts, i que els valors de **Tipo** no són únics. Ho podem veure amb:

```
# Valors no únics:
unique(dfBase$Tipo)
```

```
##  [1] "NF"     "FP"     "NI"     "FL"     "FM  "   "  FM  " "FM"     "fm"
##  [9] "FI"     "fi"
```

```
# Valors numèrics erronis:
dfBase$AE[58]
```

```
## [1] "1,990184"
```

Ara modificaré les dades per aconseguir els valors requerits i finalment comprovaré els valors nuls del dataset.

```
# Canvie Tipo:
dfBase$Tipo <- trimws(dfBase$Tipo)
dfBase$Tipo <- toupper(dfBase$Tipo)

# Canvie AE:
dfBase$AE <- gsub(",", ".", dfBase$AE)

# Canvie el tipus de les dades:
fumadors <- dfBase %>% transmute(c.pulmonar = as.numeric(AE),
                                 tipo = as.factor(Tipo),
```

```
                                genero = as.factor(genero),
                                edad = edad)

# Comprove que totes les dades són correctes:
unique(fumadors$tipo)
```

```
## [1] NF FP NI FL FM FI
## Levels: FI FL FM FP NF NI
```

```
unique(fumadors$genero)
```

```
## [1] M F
## Levels: F M
```

```
# Finalment comprove els valors nuls del dataset i elimine l'anterior:
remove(dfBase)
colSums(is.na(fumadors))
```

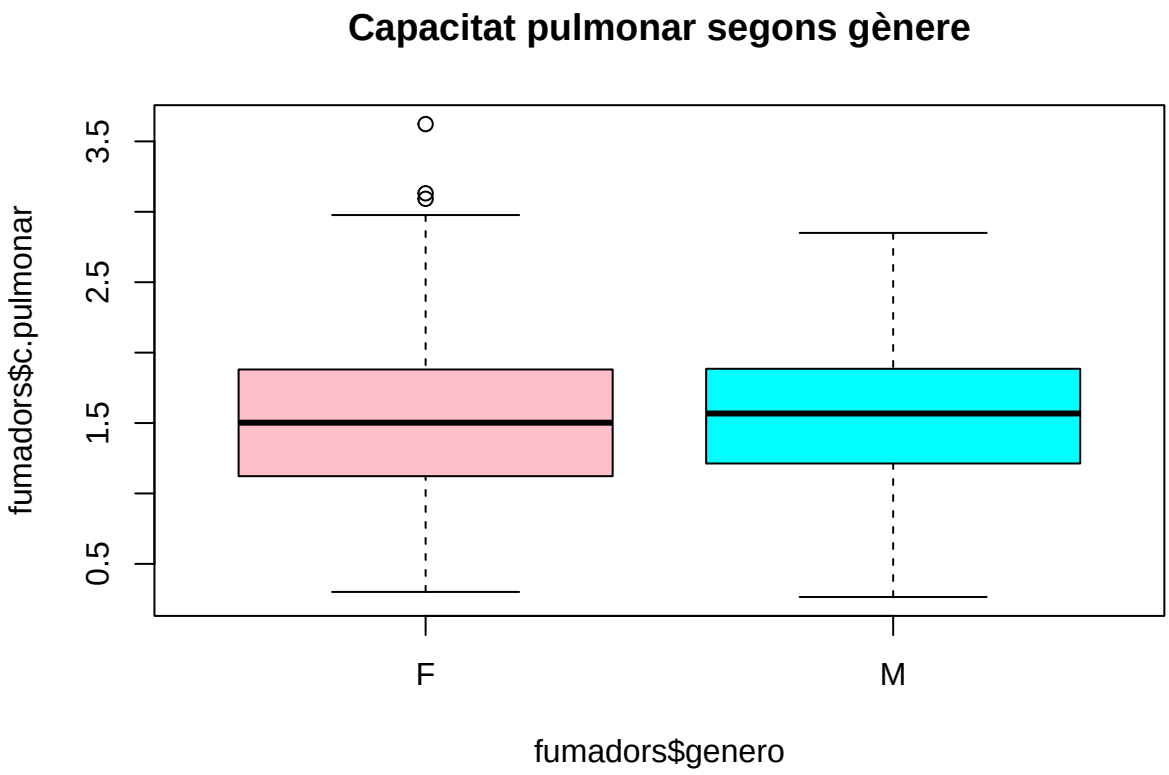
```
## c.pulmonar       tipo     genero       edad
##          0          0          0          0
```

Ara totes les dades són correctes, i a més s'ha observat que no hi ha nuls al dataset. Per tant no es tractaran aquests. Mostre les distribucions resultants:

# 2. Anàlisi descriptiva de la mostra

## 2.1 Capacitat pulmonar i gènere

Observe la relació entre la capaciatat pulmonar i el gènere:



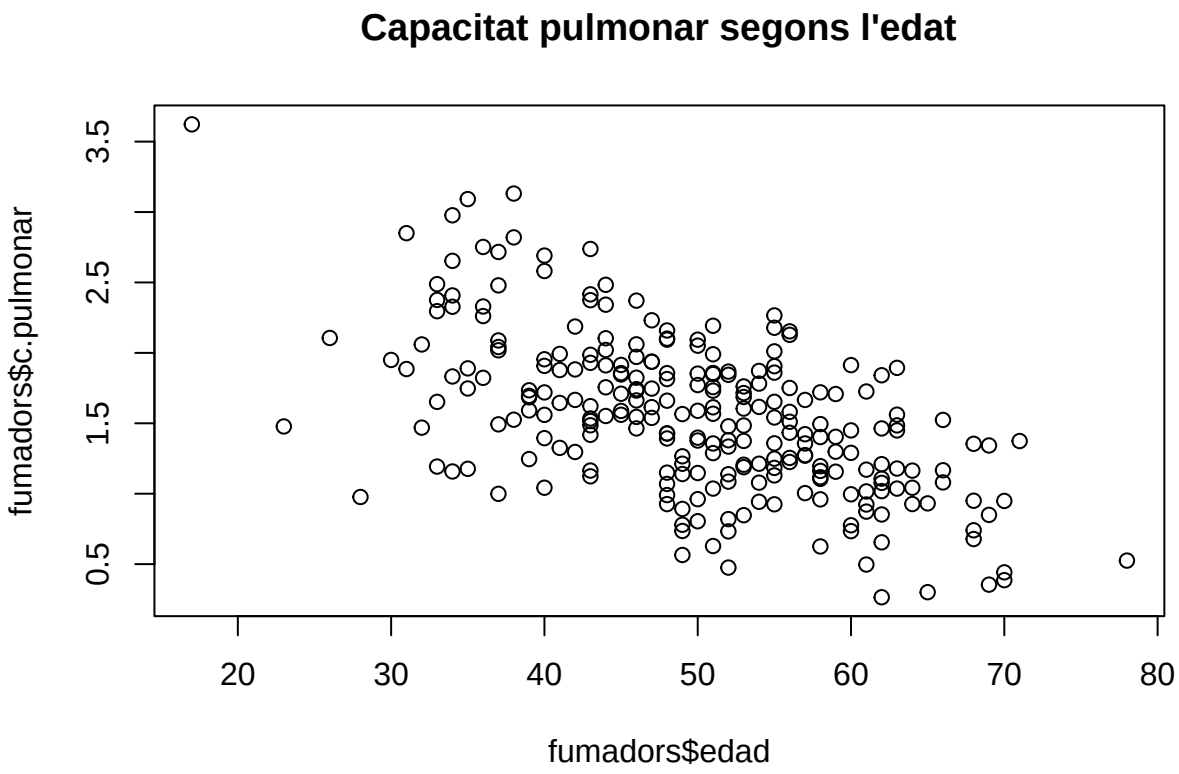

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3008  1.1312  1.5026  1.5233  1.8792  3.6226
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2649  1.2129  1.5677  1.5838  1.8853  2.8499
```

La distribució de les dades sembla prou semblant per ambdós gèneres, amb un lleuger major nombre de dones amb major capacitat pulmonar. S'observa també que hi ha més dones amb pijor capacitat pulmonar (al Q1). Passe ara a revisar la capacitat pulmonar segons l'edat:

## 2.2 Capacitat pulmonar i edat

```
plot(fumadors$edad,fumadors$c.pulmonar,main = "Capacitat pulmonar segons l'edat")
```




Pel que s'hi pot veure, la capacitat pulmonar decreix a mesura que s'augmenta l'edat. Passe ara a revisar els tipus:

### 2.3 Capacitat pulmonar segons tipus de fumadors

```
# Nombre de tius de fumador:
summary(fumadors$tipo)
```

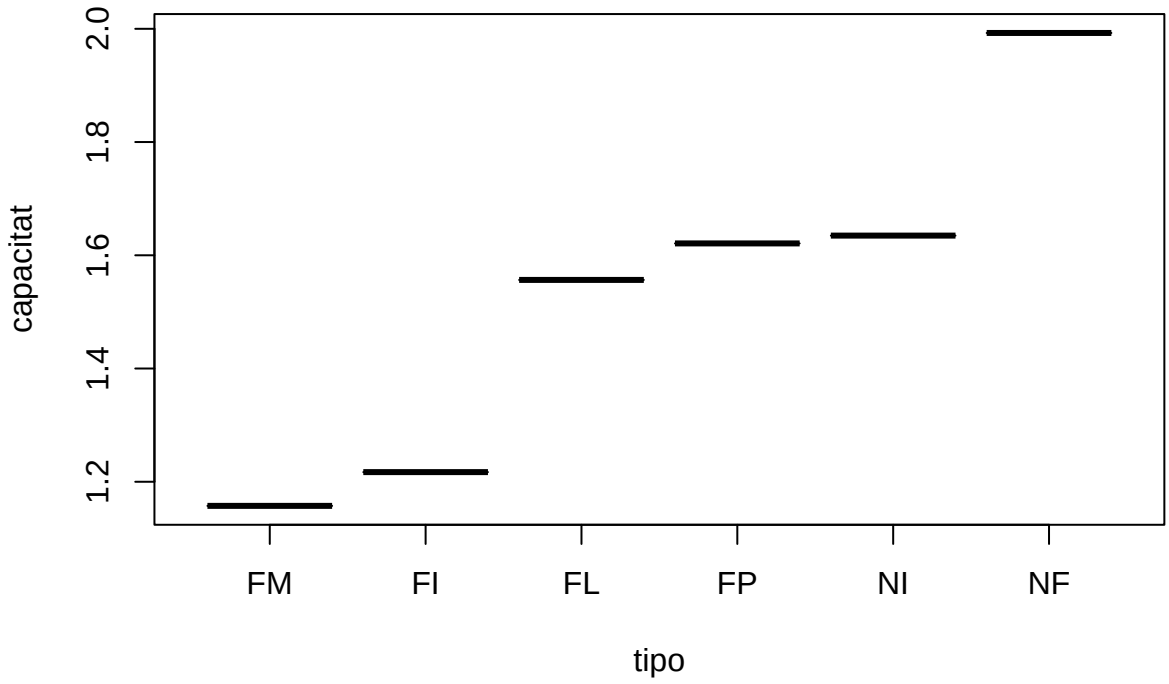
```
## FI FL FM FP NF NI
## 41 41 39 40 50 42
```

```
# Mitjana de cada tipus:
tipus <-fumadors %>%
  group_by(tipo) %>%
  summarise(capacitat = mean(c.pulmonar))

# Reordenem les dades:
tipus <- tipus[order(tipus$capacitat),]
tipus$tipo <- factor(tipus$tipo, levels = tipus$tipo[order(tipus$capacitat)])

# Mostre les mitjanes per tipus:
as.data.frame(tipus)
```

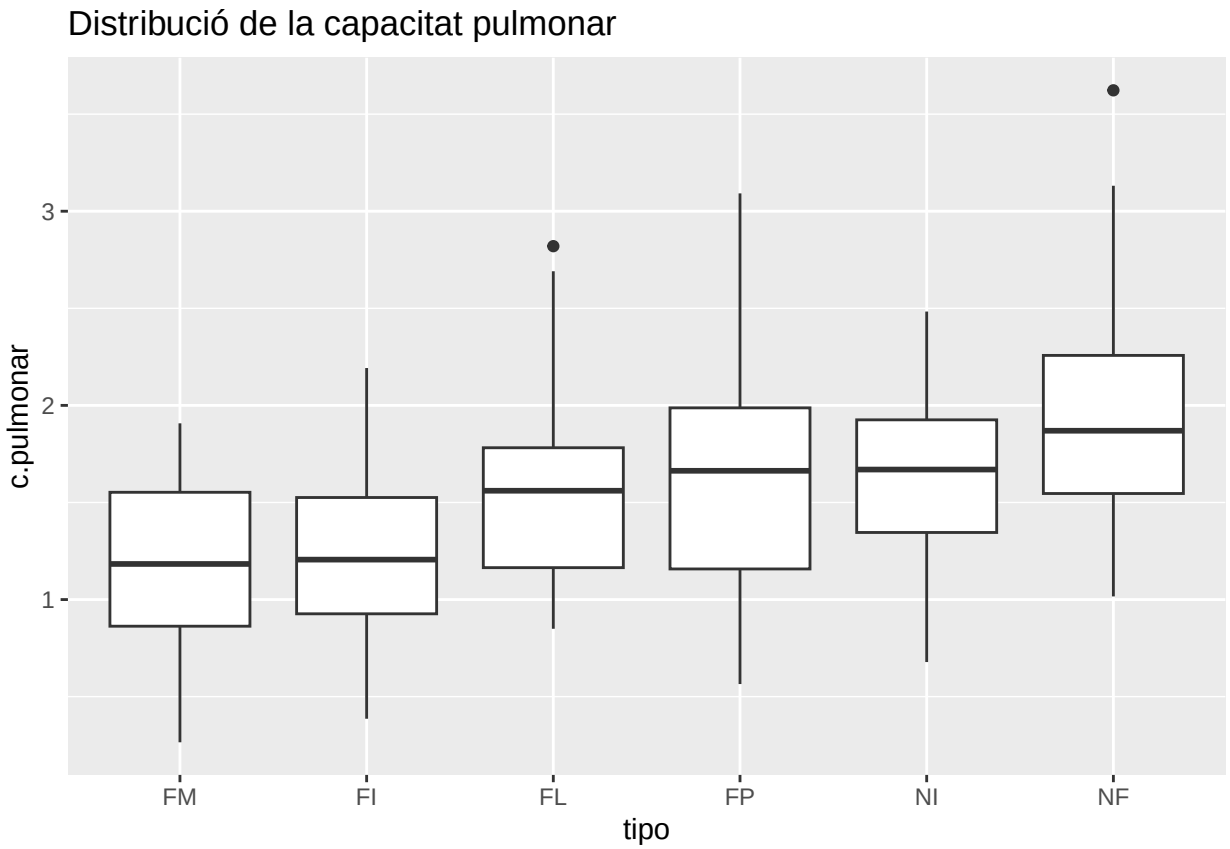
```
##   tipo capacitat
```

```
## 1   FM  1.157442
## 2   FI  1.217035
## 3   FL  1.556476
## 4   FP  1.620952
## 5   NI  1.634737
## 6   NF  1.992625
```

```
plot(tipus)
```




```
# Faig el boxplot per cada capacitat pulmonar:
fumadors$tipo <- factor(fumadors$tipo, levels = tipus$tipo[order(tipus$capacitat)])
ggplot(fumadors,aes(tipo,c.pulmonar))+geom_boxplot()+ggtitle("Distribució de la capacitat pulmonar")
```




Les conclusions que s'hi poden extraure de tot l'anàlisi és que: per a un nombre semblant dels tipus de fumadors (recompte de la mostra semblant) s'hi veu que aquelles persones que no fumen, no inhalen el fum o son fumadors passius tenen millor capacitat pulmonar que els fumadors de tots els tipus. Destaca que els fumadors moderats presenten pijor capacitat pulmonar que els fumadors intensius i que les mitjanes dels fumadors lleugers, els passius i els que no inhalen són quasi iguals.

Concretament s'hi veu al boxplot que el que pijor capacitat pulmonar tenen són els fumadors moderats, seguits pels intensius, que els segueixen els fumadors lleugers (amb nombres del Q3-Max relativament elevats), seguits pels fumadors pasius (amb una gran dispersió de les dades: observar que és el plot amb més distància Min-Max), seguits pels fumadors que no inhalen el fum (amb valors de dispersió semblant als fumadors lleugers), i finalment els no fumadors, que presenten les millors dades de capacitat pulmonar.

La conclusió que es podria extraure d'aquestes dades és que si no fumes tens millor capacitat pulmonar en general, les persones que no inhalen el fum generalment tenen millor capacitat pulmonar (relació directa entre ingerir fum i pijor capacitat pulmonar), l'etiqueta de fumador passiu és molt vaga i resulta en una gran dispersió de les dades, i finalment que hi ha gent que fuma de manera intensiva i te millor capacitat pulmonar que la gent que fuma moderadament (possible adaptació?).

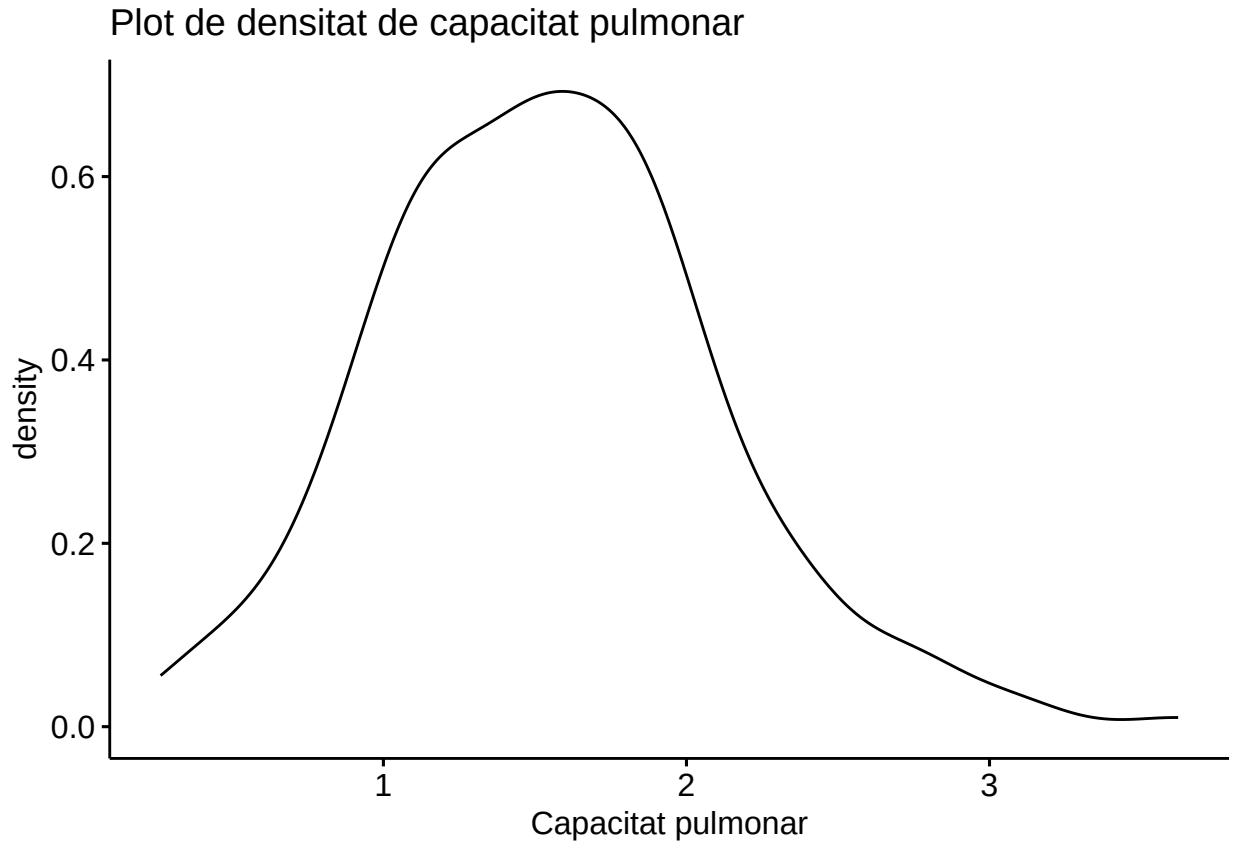
## 3. Interval de confiança de la capacitat pulmonar

Abans de calcular l'interval de confiança de la capacitat pulmonar, comprovaré que les dades compleixen la normalitat:

```
# Test de normalitat
shapiro.test(fumadors$c.pulmonar)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fumadors$c.pulmonar
## W = 0.98869, p-value = 0.04484
```

```
ggdensity(fumadors$c.pulmonar,
          main = "Plot de densitat de capacitat pulmonar",
          xlab = "Capacitat pulmonar")
```




El test de Shapiro-Wilk ens ha donat un p-valor menor al nivell de significància, 0.05, el qual ens faria descartar la normalitat de les dades. Així mateix, observant la distribució de la capacitat pulmonar, s'hi pot veure que segueix prou una distribució normal amb una lleugera cua a la dreta. Aquesta estaria causada per les dades de la gent que no és fumadora o és més jove (amb més capacitat pulmonar) i s'ha de recordar que en aquest dataset estem mesurant la capacitat pulmonar a persones fumadores, per això les dades s'han desplaçat cap a una menor capacitat pulmonar.

Tenint això present, i acollint-me al Teorema central del límit, concloure que les dades segueixen una distribució normal. Es pot argumentar que aquestes dades no segueixen formalment una distribució normal perquè s'hi mesclen dades de gent fumadora, amb gent que no fuma, però que per la població la capacitat pulmonar seguirà una distribució normal.

Nota: no he volgut fer una transformació per no complicar la interpretació dels resultats. Amb una transformació d'arrel quadrada es soluciona la normalitat de la capacitat pulmonar de la mostra.

Asumint que la capacitat pulmonar és normal, ho serà també per als homes i les dones. Així mateix, es confirma amb:

```
shapiro.test(homes$c.pulmonar)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  homes$c.pulmonar
## W = 0.98857, p-value = 0.4874
```

```
shapiro.test(dones$c.pulmonar)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dones$c.pulmonar
## W = 0.97595, p-value = 0.01227
```

Ambdos p-valors confirmen la normalitat de les dades. Ara calcularé l'interval de confiança de la mitjana pulmonar per ambdós gèneres, a una confiança del 95%:

```
#Interval de confiança al 95%
t.test(homes$c.pulmonar)$"conf.int"
```

```
## [1] 1.482301 1.685224
## attr(,"conf.level")
## [1] 0.95
```

```
t.test(dones$c.pulmonar)$"conf.int"
```

```
## [1] 1.428428 1.618132
## attr(,"conf.level")
## [1] 0.95
```

Podem veure que l'interval per als homes és de [1.48, 1.68] i el de les dones és de [1.43, 1.62]. Segons aquestes dades, i per l'observat a l'apartat anterior, s'hi podria intuir que la capacitat pulmonar mitjana de les dones és lleugerament menor al dels homes. No queda ben clar perquè ben bé podrien ser mitjanes iguals. Ho estudiarem en més en detall al apartat següent.

# 4. Diferències de capacitat pulmonar segons sexe

## 4.1 Hipòtesi

A partir de les conclusions extretes a l'apartat anterior, on s'observava que la mitjana de les dones podria ser menor a la dels homes, formule les hipotesi següents:

**H0:** y(cp.homes) = y(cp.dones)

**H1:** y(cp.homes) != y(cp.dones)

De ser certa l'hipotesi alternativa, es podria considerar que la mitjana de les dones serà menor a la dels homes segons el que hem observat a l'apartat anterior.

## 4.2 Contrast

El test que aplicaré serà el d'un contrast d'hipòtesi de dues mostres amb mitjana i variància desconegudes. Com he comprovat a l'apartat anterior, es compleix la norma de normalitat en ambdós gèneres. Ara comprovaré la igualtat de variàncies:

```
var.test(homes$c.pulmonar,dones$c.pulmonar)
```

```
##
##  F test to compare two variances
##
## data:  homes$c.pulmonar and dones$c.pulmonar
## F = 0.86133, num df = 108, denom df = 143, p-value = 0.4152
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6066144 1.2339167
## sample estimates:
## ratio of variances
##           0.861326
```

Amb el p-valor major al nivell de significació, el test mostra que les variàncies són iguals. Per tant serà un test de dues mostres amb mitjanes i variàncies desconegudes però iguals. El test serà bilateral.

## 4.3 Càlculs

Mostre ara els càlculs necessaris per fer el contrast:

```
# Guarde el test:
ttest.bilateral <- function( x1, x2, CL=95, var.equal=FALSE ){
  mean1<-mean(x1); n1<-length(x1); sd1<-sd(x1)
  mean2<-mean(x2); n2<-length(x2); sd2<-sd(x2)
  alfa <- (1-CL/100)

#variances iguals
  if (var.equal){
    S <- sqrt( ( (n1-1)*sd1^2 + (n2-1)*sd2^2 ) / (n1+n2-2) )
    t <- (mean1-mean2) / (S * sqrt(1/n1+1/n2) )
    df <- n1+n2-2
```

```
  }
  else{
    #variances diferents
    Sb <- sqrt( sd1^2/n1 + sd2^2/n2 )
    denom <- ( (sd1^2/n1)^2/(n1-1) + (sd2^2/n2)^2/(n2-1))
    df <- ( (sd1^2/n1 + sd2^2/n2)^2 ) / denom
    t<- (mean1-mean2) / Sb #valor observat
  }

  tcritical <- qt( alfa/2, df, lower.tail=FALSE )

  pvalue<-pt( abs(t), df, lower.tail=FALSE )*2

  #Guarde el resultat
  info<-c(mean1, mean2, t,tcritical,pvalue,df)
  names(info)<-c("mean1", "mean2", "t","tcritical", "pvalue", "df")
  return (info)
}

# Faig el càlcul:
test_1 <- ttest.bilateral(homes$c.pulmonar,dones$c.pulmonar, CL = 95, var.equal = TRUE)
test_1
```

```
##       mean1       mean2           t   tcritical      pvalue          df
##   1.5837624   1.5232801   0.8531624   1.9694602   0.3943827 251.0000000
```

### 4.4 Interpretació

A través dels càlculs exposats anteriorment, podem veure que el valor observat cau dintre del rang d'acceptació de la hipòtesi nul · la. A més, el p-valor és major a la significació escollida. És a dir, com 0.85 està dintre de l'interval [-1.96, 1.96] i el p-valor és $0.39 > 0.05$. Per tant, no podem rebutjar l'hipotesi nul · la i hem d'acceptar que la mitjana de la capacitat pulmonar és la mateixa tant per a homes com per a dones.

Podriem dir que la diferència observada a la mostra es deu a un nombre major de dones fumadores (o que fumen amb més intensitat) que d'homes fumadors.

## 5. Diferències en la capacitat pulmonar entre Fumadors i no Fumadors

### 5.1 Hipòtesi

En aquest apartat volem comprovar si realment fumar afecta a la capacitat pulmonar. Per tot el que hem estudiat a apartats anteriors, es pot intuïr que sí que influeix, concretament que la fa empijorar. Per tant les hipòtesis per aquest test seran:

**H0:** y(cp.fuma) >= y(cp.no.fuma)

**H1:** y(cp.fuma) < y(cp.no.fuma)

## 5.2 Contrast

Agruparé ara les dades:

```
# Agrupe les dades:
fuma <- fumadors[fumadors$tipo %in% c("NI","FL","FM","FI"),]
no.fuma <- fumadors[fumadors$tipo %in% c("NF","FP"),]
```

Ara, a priori, totes les dades haurien de seguir complint les propietats de normalitat i igualtat de variàncies perquè seguim parlant de capacitat pulmonar. Igualment, les calcularé per demostrar-ho:

```
shapiro.test(fuma$c.pulmonar)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fuma$c.pulmonar
## W = 0.99487, p-value = 0.8442
```

```
shapiro.test(no.fuma$c.pulmonar)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  no.fuma$c.pulmonar
## W = 0.97594, p-value = 0.09325
```

```
var.test(fuma$c.pulmonar,no.fuma$c.pulmonar)
```

```
##
##  F test to compare two variances
##
## data:  fuma$c.pulmonar and no.fuma$c.pulmonar
## F = 0.79901, num df = 162, denom df = 89, p-value = 0.2187
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5477312 1.1426311
## sample estimates:
## ratio of variances
##          0.7990148
```

Es pot comprovar que les propietats comentades segueixen mantenint-se. Normalitat amb els p-valors majors a 0.05, i el p-valor del test de variàncies també major a 0.05.

Per tant, llevant-nos això del damunt, ara aplicaré un contrast de hipòtesis per a dues mostres amb mitjana desconeguda i variància desconeguda però igual. El test serà unilateral per l'esquerra.

### 5.3 Preparació de les dades

Ja les he preparades, mostre els seus valors únics:

```
unique(fuma$tipo)
```

```
## [1] NI FL FM FI
## Levels: FM FI FL FP NI NF
```

```
unique(no.fuma$tipo)
```

```
## [1] NF FP
## Levels: FM FI FL FP NI NF
```

### 5.4 Càlculs

```
ttest.unilateral <- function( x1, x2, CL=95, alternative="less", var.equal=FALSE ){
  mean1<-mean(x1); n1<-length(x1); sd1<-sd(x1)
  mean2<-mean(x2); n2<-length(x2); sd2<-sd(x2)
  alfa <- (1-CL/100)
  #variances iguals
  if (var.equal){
    S <- sqrt( ( (n1-1)*sd1^2 + (n2-1)*sd2^2 ) / (n1+n2-2) )
    t <- (mean1-mean2) / (S * sqrt(1/n1+1/n2) )
    df <- n1+n2-2
  }else{
  #variances diferents
  Sb <- sqrt( sd1^2/n1 + sd2^2/n2 )
  denom <- ( (sd1^2/n1)^2/(n1-1) + (sd2^2/n2)^2/(n2-1))
  df <- ( (sd1^2/n1 + sd2^2/n2)^2 ) / denom
  t<- (mean1-mean2) / Sb #valor observat
  }
  #less
  if (alternative=="less"){
  tcritical <- qt( alfa, df, lower.tail=TRUE )
  pvalue<-pt( t, df, lower.tail=TRUE )
  }else{ #greater
  tcritical <- qt( alfa, df, lower.tail=FALSE )
  pvalue<-pt( t, df, lower.tail=FALSE )
  }

  #Guarde el resultat
  info<-c(mean1, mean2, t,tcritical,pvalue,df)
  names(info)<-c("mean1", "mean2", "t","tcritical", "pvalue", "df")
  return (info)
}

# Aplique el test
test_2 <- ttest.unilateral(fuma$c.pulmonar,no.fuma$c.pulmonar,CL = 95, var.equal = TRUE)
test_2
```

```
##         mean1         mean2             t     tcritical        pvalue
##  1.395786e+00  1.827437e+00 -6.329761e+00 -1.650947e+00  5.613478e-10
##            df
##  2.510000e+02
```

El càlcul en auest cas ens mostra el següent: el valor observat cau fora de l'interval d'accpetació i el p-valor és menor al nivell de significació. És a dir, -6.33 està fora de [-1.65,INF] i el p-valor = 5.6e-10 $< 0.05$. Per tant cal rebutjar l'hipotesi nul $\cdot$ la i acceptar l'aternativa. És a dir, els fumadors presenten una capacitat pulmonar menor a les persones no fumadores i els que són fumadors passius.

# 6. Anàlisi de regressió lineal

Ara procuraré de crear un model de regressió lineal multiple a partir de totes les dades. La variable explicada serà la capacitat pulmonar i com es demana a l'enunciat, no excloure cap variable explicativa. Per tant passe directament a fer el càlcul.

## 6.1 Càlcul

Construisc el model:

```
model_c.pulmonar <- lm(c.pulmonar ~ tipo + genero + edad,
                   data=fumadors)
summary(model_c.pulmonar)
```

```
##
## Call:
## lm(formula = c.pulmonar ~ tipo + genero + edad, data = fumadors)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05421 -0.25126 -0.00321  0.23288  1.03947
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.787768   0.136568  20.413  < 2e-16 ***
## tipoFI      -0.046357   0.082133  -0.564 0.572986
## tipoFL       0.292102   0.082363   3.547 0.000468 ***
## tipoFP       0.347985   0.082807   4.202 3.70e-05 ***
## tipoNI       0.377166   0.081602   4.622 6.15e-06 ***
## tipoNF       0.735450   0.078432   9.377  < 2e-16 ***
## generoM     -0.002321   0.047033  -0.049 0.960680
## edad        -0.030951   0.002276 -13.601  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3655 on 245 degrees of freedom
## Multiple R-squared:  0.583,  Adjusted R-squared:  0.5711
## F-statistic: 48.94 on 7 and 245 DF,  p-value: < 2.2e-16
```

## 6.2 Interpretació

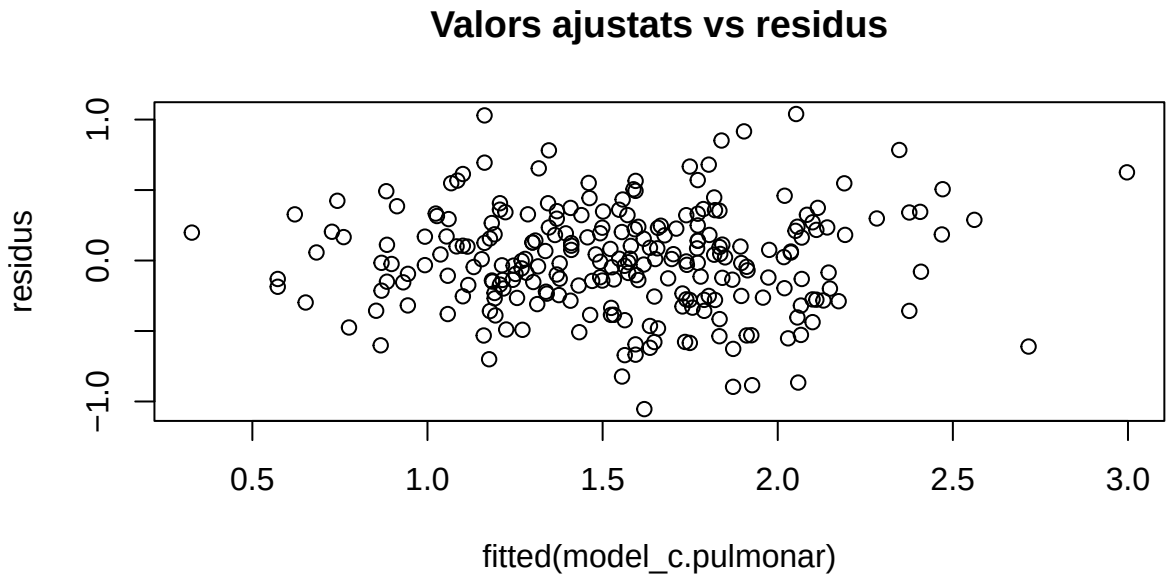
Pel que s'hi pot veure al model aconseguit anteriorment, les variables més explicatives (amb més potència estadística al valor t) són les de **edad** i els diferents **tipus** de fumadors. Concretament s'hi veu que les persones no fumadores determinen prou més que la resta la capacitat pulmonar, seguidament per les persones que no inhalen el fum, els fumadors passius, i finalment els fumadors lleugers.

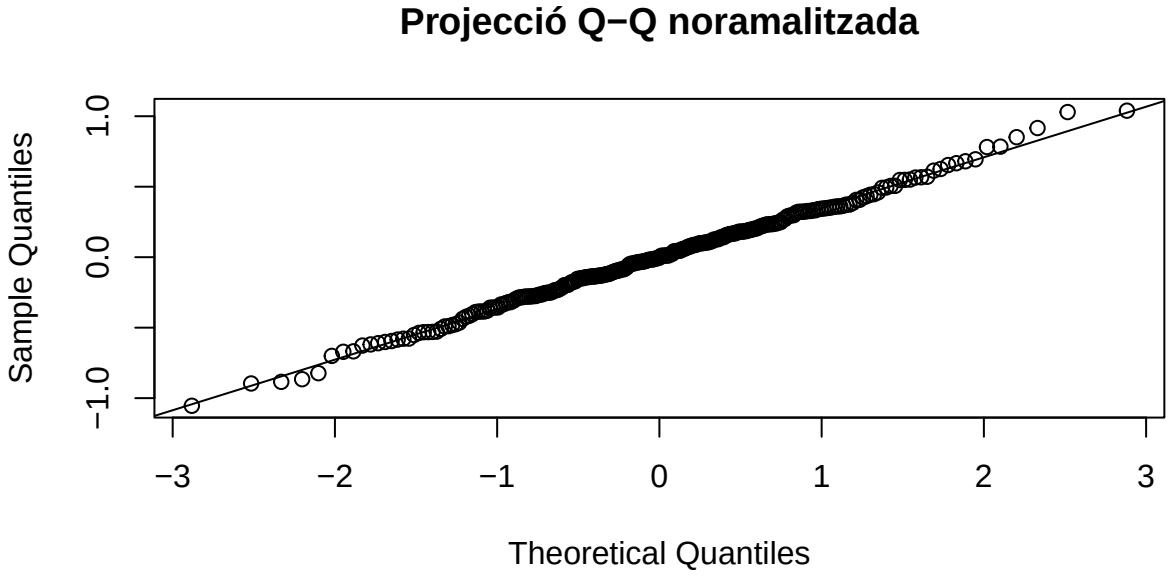
Les variables de fumadors intensius i el gènere de les persones són les variables que menys potència estadística tenen en comparació a la resta. De més contribució al model a menys seria: "edad, tipoNF, tipoNI, tipoFP, tipoFL, tipoFI, generoM" o sintetitzant "edad, tipo, genero".

### 6.3 Bondat de l'ajust

Pel que podem veure als resultats del model, el valor del RSE és menut [0.36], el qual ens explica que hi hauria pocs errors al nostre model. A més el valor de l'R quadrat ens dona un valor prou positiu, del 57%, el qual significa que tenim una bona proporció de les dades explicades a través d'aquest model. A més, l'estadístic F el tenim moderadament elevat.

Observaré ara la distribució dels residus i el gràfic Q-Q per acabar d'analitzar la qualitat de l'ajust:




En aquests darrers gràfics veiem que els residus estan prou espaiats i no s'hi veu una distribució de les dades clara: bona senyal. A més s'hi veu que les dades s'ajusten molt bé a la línia dels quantils teòrics, per tant els residus segueixen una distribució normal.

Amb tot, podem concloure que el model és bo per predir les dades.

## 6.4 Predicció

Ara provaré a calcular la capacitat pulmonar de persones de 30 anys fins als 80, per a cada tipus de fumador. Primer cree el dataframe i el mostre:

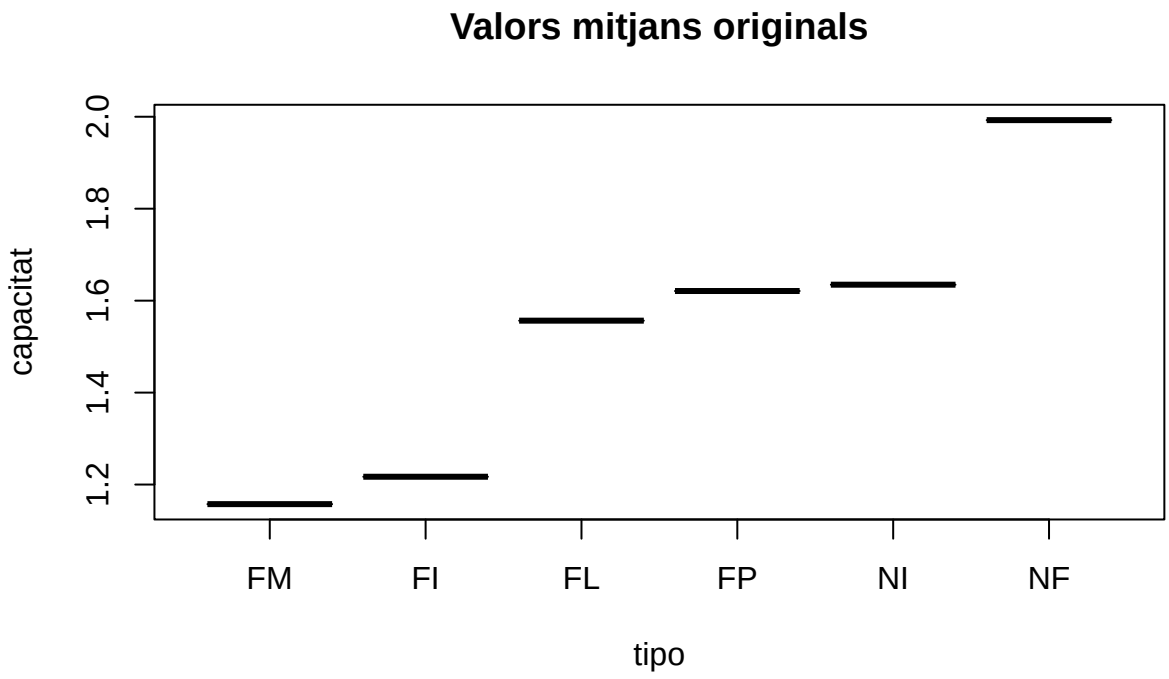
```
## 'data.frame':    306 obs. of  4 variables:
##  $ c.pulmonar: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ tipo      : Factor w/ 6 levels "FM","FI","FL",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ genero    : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ edad      : num  30 31 32 33 34 35 36 37 38 39 ...
```

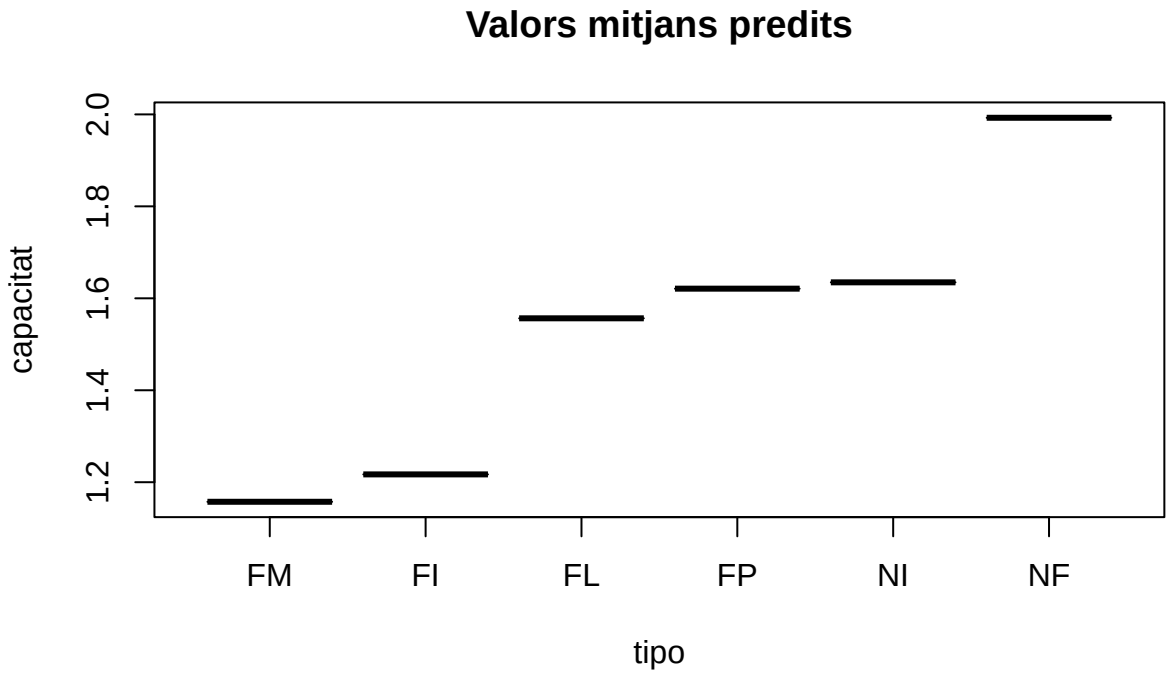
Una vegada tenim les dades preparades, aplique la predicció a totes elles:

```
for (row in c(1:nrow(prova_model))) {
  valor <- data.frame(tipo = prova_model[row,2],
                      genero = prova_model[row,3],
                      edad = prova_model[row,4])
  prova_model[row,1] <- predict(model_c.pulmonar, newdata = valor)
}

# Mostre un resum de la taula:
head(prova_model)
```

```
##   c.pulmonar tipo genero edad
## 1   1.856905   FM      M   30
## 2   1.825953   FM      M   31
## 3   1.795002   FM      M   32
## 4   1.764050   FM      M   33
## 5   1.733099   FM      M   34
## 6   1.702148   FM      M   35
```

Per comparar-ho amb els valors anteriors, cree la taula de mitjanes i mostre el gràfic anterior i l'actual:




Pel que s'hi pot veure, observant les mitjanes de cada tipus, la predicció ha funcionat correctament per les dades sugerides.

# 7. ANOVA unifactorial

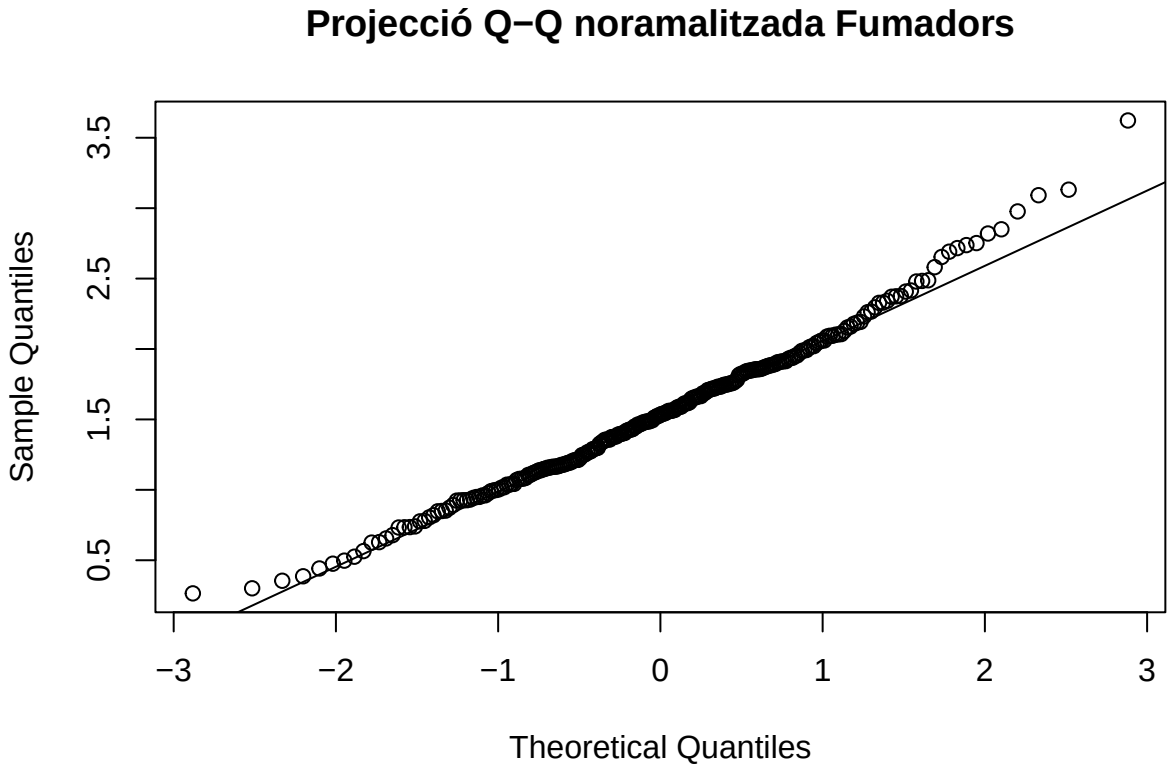
En aquest apartat estudiarem la variabilitat de la capacitat pulmonar pot explicar-se pel factor tipus de fumador. Concretament volem saber:

- Si existeixen diferències en la capacitat pulmonar dels diversos tipus de fumadors/no fumadors.
- Si existeixen diferències, entre quins grups hi han aquestes diferències.

## 7.1 Normalitat

Comprove primerament que les dades segueixen una distribució normal. Aquest test ja l'havia aplicat abans, però el torne a mostrar amb la seua representació al gràfic Q-Q:

```
##
##  Shapiro-Wilk normality test
##
## data:  fumadors$c.pulmonar
## W = 0.98869, p-value = 0.04484
```




Com s'ha comentat en apartats anteriors, el test estadístic aplicat ens dona un p-valor menor al nivell de significància, el qual ens faria rebutjar la hipòtesi nul · la i acceptar que les dades no segueixen una distribució normal. També, com he comentat anteriorment, considere que aquestes dades poden estar un poc esbiaixades per la mescla de persones fumadores/no fumadores i acollint-me al Teorema Central del Límit considere que les dades seran normals.

A més, a través del gràfic de densitat de l'apartat anterior esmentat, a més del gràfic Q-Q que mostre actualment, s'hi veu que les dades s'aproximen molt a les d'una distribució normal (Q-Q a la diagonal) amb una lleugera desviació per les dades de la gent que presenta molt bona capacitat pulmonar (aquesta assumpció es veia millor al gràfic de densitat).

Per tant, considere que les dades són normals.

## 7.2 Homocedasticitat Homogeneitat de variàncies

Ara en aquest apartat compararé la variància dels diferents tipus per veure si presenten homocedasticitat i s'hi pot aplicar l'anova sense problemes:

```
# Test de bartlett:
bartlett.test(c.pulmonar ~ tipo, data = fumadors)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  c.pulmonar by tipo
## Bartlett's K-squared = 3.2658, df = 5, p-value = 0.6591
```

A través del test de Barlett veiem que el p-valor és major al nivell de significància i per tant mantenim la hipòtesi nul · la i acceptem que les variàncies són iguals.

## 7.3 Hipòtesi nul · la i alternativa

Una vegada hem esclarit que les dades són normals i ténen variàncies iguals, plantege les hipòtesis de l'anova (represente els factors amb T, hi ha sis tipus de fumador):

**H0:** T1 = T2 = T3 = T4 = T5 = T6 = 0

**H1:** Ti != Tj per a algun i != j

## 7.4 Càlcul ANOVA

En aquest punt, calculare l'ANOVA i la mostre:

```
# Calcule l'anova:
fum.aov <- aov(c.pulmonar ~ tipo, data = fumadors)

# La mostre:
summary(fum.aov)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## tipo          5  20.86   4.171   17.88 4.03e-15 ***
## Residuals   247  57.63   0.233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### 7.5 Interpretació

Podem comprovar que com el p-valor és menor al nivell de significància, 4.03e-15 $< 0.05$, hi ha una diferència significativa entre els tipus de fumadors. Ens ve a dir que segons el tipus de fumador, hi ha una gran variància entre les mitjanes de cadascun.

Si relacionem aquestes conclusions, amb el gràfic de boxplot de l'apartat 2.3, podem ovservar que efectivament les dades estaven distribuïdes amb prou diferències (al igual que les seues mitjanes). Podem asumir també pel resultat del contrast d'hipotesis dels fumadors i els no fumadors que les diferències radiquen principalment entre les persones que fumen i les que no. És en l'apartat 2.3 en el que s'ha detallat com s'havia d'interpretar cada tipus.

### 7.6 Aprofundint en ANOVA

### 7.7 Força de la relació

## 8. Comparacions múltiples

Com que a l'apartat anterior hem rebutjat l'hipòtesi nul · la que ens feia acceptar la diferència entre els factors, ara aplicaré un test de comparació multiple per veure quines són aquestes diferències entre els tipus.

### 8.1 Test pairwise

Aplique el test pairwise sense cap correcció i posteriorment l'analitze:

```
pairwise.t.test(fumadors$c.pulmonar,fumadors$tipo, p.adjust.method = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fumadors$c.pulmonar and fumadors$tipo
##
##    FM      FI      FL      FP      NI
## FI 0.58175 -       -       -       -
## FL 0.00027 0.00165 -       -       -
## FP 2.9e-05 0.00021 0.54864 -       -
## NI 1.3e-05 0.00011 0.46122 0.89733 -
## NF 2.6e-14 5.4e-13 2.6e-05 0.00035 0.00048
##
## P value adjustment method: none
```

El test de pairwise ens indica quines són les diferències entre les mitjanes de cada factor. A través dels p-valors de cada parell podem saber quines parelles presenten diferències a les seues mitjanes. Quan el p-valor és menor a 0.05, ens indica que les parelles presenten mitjanes diferents. Per tant del càlcul anterior s'hi pot extraure:

Les variables de FM i FI estan aparellades, com també ho estan FL i FP, FL i NI, i FP i NI. És a dir, les dades que presenten mitjanes semblants són: els fumadors moderats i intensius (suposem que poca capacitat pulmonar), fumadors lleugers i passius (c.p. moderada), fumadors lleugers i que no ingereixen fum (c.p. moderada), i finalment fumadors passius i que no ingereixen fum (c.p. moderada). Les conclusions sobre la c.p les he inferit en base als resultats de el boxplot de l'apartat 2.3 i de les seues mitjanes observades.

Pel que fa a les variables que presenten diferències entre les seues mitjanes tenim: fumadors intensius amb tota la resta que no són els moderats (pijor c.p); fumadors lleugers amb els moderats i intensius (tenen millor cp que ambdós fumadors) i amb els no fumadors (tenen pijor c.p); fumadors passius amb els moderats i intensius (millor c.p) i amb els no fumadors (pijor c.p.); fumadors que no ingereixen amb fumadors moderats i intensius (millor c.p.) i els no fumadors (pijor c.p.); i finalment s'observa que els no fumadors presenten diferències amb tots els grups (tenen la millor c.p).

Aquest test vindria a confirmar les conclusions que s'han anat extraient a tota la pac.

## 8.2 Correcció de Bonferroni

Per fer correctament una comparació múltiple de diversos factors és més correcte aplicar una correcció per fer-la més precisa. Aplique la correcció de Bonferroni i analitze els resultats:

```
pairwise.t.test(fumadors$c.pulmonar,fumadors$tipo, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fumadors$c.pulmonar and fumadors$tipo
##
##    FM      FI      FL      FP      NI
## FI 1.00000 -       -       -       -
## FL 0.00409 0.02477 -       -       -
## FP 0.00043 0.00315 1.00000 -       -
## NI 0.00020 0.00160 1.00000 1.00000 -
## NF 4.0e-13 8.1e-12 0.00039 0.00522 0.00717
##
## P value adjustment method: bonferroni
```

Una vegada calculada, veiem que els resultats són més exactes pel que fa a les dades aparellades, però no hi ha diferències per als factors que mostraven diferències entre mitjanes (p-valor menor a 0.05). Per tant el resultat d'aquest test, amb la correcció corresponent, s'hauria d'interpretar com el resultat de l'anterior.

# 9. ANOVA multifactorial

Ara, a l'anàlisi de les variàncies, afegirem també el factor de gènere per veure si s'observen diferències notables amb l'anàlisi anterior.

## 9.1 Anàlisi visual

Primerament, però, faré un estudi visual de les dades per veure si existeixen efectes principals o d'interacció entre gènere i tipus de fumador. Agrupe les dades i les mostre com a taula per gènere:

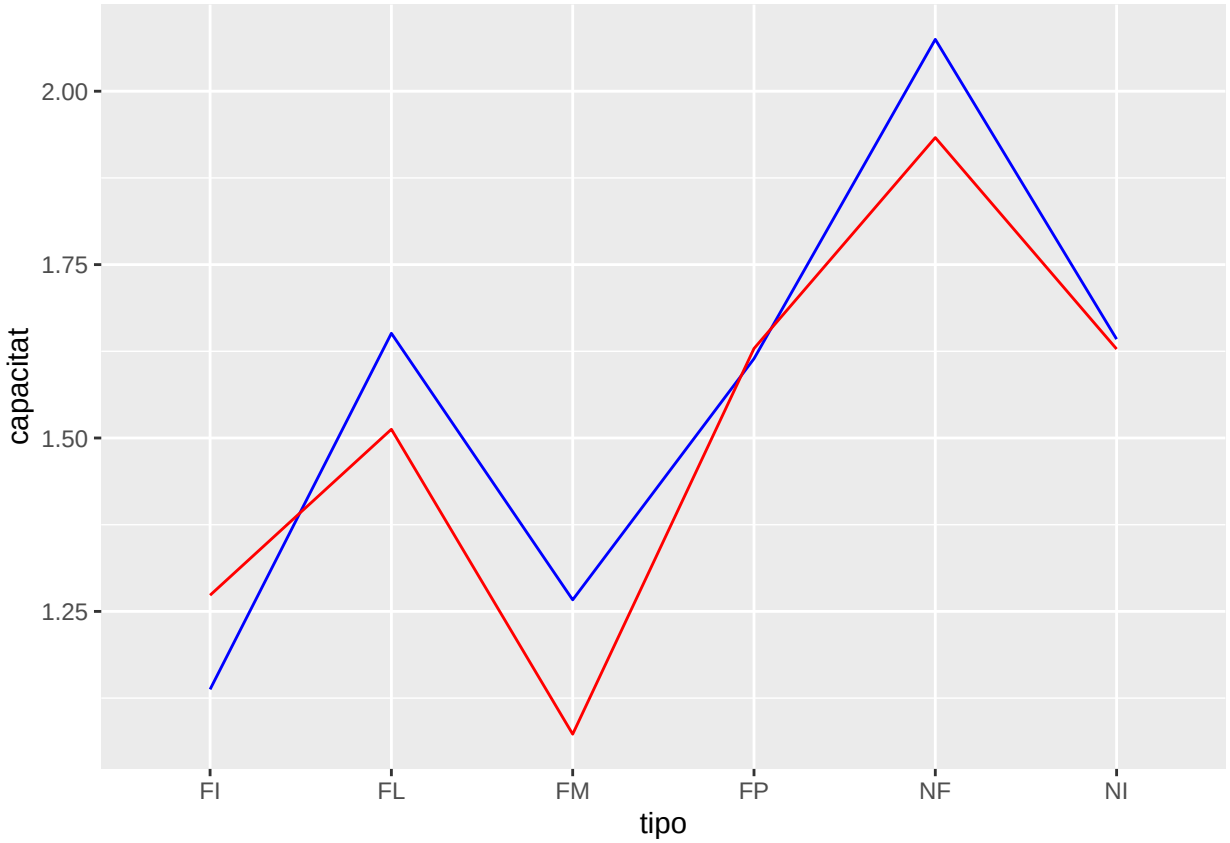
```
# Aprofite els datasets 'homes' i 'dones' anteriors:

# Mitjana de cada tipus H:
tipusH <- homes %>%
  group_by(tipo) %>%
  summarise(capacitat = mean(c.pulmonar))

# Mitjana de cada tipus D:
tipusD <- dones %>%
  group_by(tipo) %>%
  summarise(capacitat = mean(c.pulmonar))

#Mostre Homes
tipusH
```

```
## # A tibble: 6 x 2
##   tipo  capacitat
##   <fct>     <dbl>
## 1 FI         1.14
## 2 FL         1.65
## 3 FM         1.27
## 4 FP         1.61
## 5 NF         2.07
## 6 NI         1.64
```

```
#Mostre Dones
tipusD
```

```
## # A tibble: 6 x 2
##   tipo  capacitat
##   <fct>     <dbl>
## 1 FI         1.27
## 2 FL         1.51
## 3 FM         1.07
## 4 FP         1.63
## 5 NF         1.93
## 6 NI         1.63
```

Ara que veiem les dades, mostre la seua representació gràfica per observar l'interacció:




Interpretació dels resultats: Com s'hi pot veure al gràfic anterior, sí que existeix interacció. Veiem que les dones fumadores intensives tenen millor c.p, i per al tipus de fumadors passius també passa així. Altrament, a la resta de tipus els homes sempre tenen millor capacitat pulmonar. Com les rectes no son paraleles s'hi observa la interacció de les dades.

## 9.2 ANOVA multifactorial

Calcularé ara l'anàlisi de la variància de la capacitat pulmonar per als factors de gènere i tipus, amb la seua interacció. Les hipotesis serien les mateixes que a l'exercici 7, però ara amb dos factors.

```
# Calcule l'anova:
fum.aov2 <- aov(c.pulmonar ~ tipo + genero + tipo:genero, data = fumadors)

# La mostre:
summary(fum.aov2)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## tipo          5  20.86   4.171  17.739 5.81e-15 ***
## genero        1   0.20   0.197   0.838    0.361
## tipo:genero   5   0.76   0.153   0.650    0.661
## Residuals   241  56.67   0.235
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## 9.3 Interpretació

Observant els resultats del càlcul anterior s'hi observa que les diferències entre les mitjanes radiquen exclusivament segons els tipus de fumadors. Ho podem saber pels p-valors recollits al test: el p-valor de tipo és el únic significatiu que ens fa descartar l'hipotesi nul · la de igualtat de variàncies, mentre que tant gereno, com la seua interacció amt tipo, son majors del nivell de significació escollit i per tant es manté l'hipotesi nul · la per a aquests.

És a dir, com ja haviem investigat amb el contrast d'hipotesi sobre el gènere, tant el gènere com la interacció d'aquest amb el tipus no són rellevants per explicar la variància entre les mitjanes de la capacitat pulmonar. El factor determinant és el tipus. Més detalladament, els tipus analitzats a l'apartat 8.

# 10. Resum tècnic

| N | Pregunta | Resultat | Conclusió |
|---|---|---|---|
| 1. | Hi ha inconsistències? | Es canvien els tipus de Tipo, genero a factor. Es corregeixen les comes de AE. Es renombren les variables AE i Tipo | S'han preparat les dades per fer l'anàlisi |
| 2.1 | Quines diferències hi ha entre cp i gènere? | Es mostra el gràfic | Hi ha poques diferències de gènere, amb una lleu desviació cap a una pijor capacitat pulmonar de les dones |
| 2.2 | Quines diferències hi ha entre cp i edat? | Es mostra el gràfic | A mesura que augmenta l'edat empijora la capacitat pulmonar |
| 2.3 | Quines diferències hi ha entre cp i els tipus de fumador? | Es mostra el gràfic | Hi ha diferències notables entre els fumadors i no fumadors. A més consum, pijor cp. |
| 3. | Interval de confiança de la capacitat pulmonar homes i dones | [1.48,1.68] per als homes i [1.43,1.62] per a les dones | Hi ha una lleugera diferència als intervals, però no sembla molt significativa |
| 4. | Hi ha diferències en la cp de les dones i els homes? | Contrast d'hipotesis amb test de dues mostres amb mitjanes i variancies desconegudes però iguals. Bilateral. Guardat a la funció ttest.bilateral() | No s'observen diferències significatives per a la població |
| 5. | Hi ha diferències en la cp dels fumadors vs els no fumadors? | Contrast d'hipotesis amb test de dues mostres amb mitjanes i variancies desconegudes però iguals. Unilateral per l'esquerra. Guardat a la funció ttest.unilateral() | La mitjana poblacional de la capacitat pulmonar dels fumadors és menor a la dels no fumadors |
| 6. | Trobar un model de regressió lineal múltiple que predisca la capacitat pulmonar per als valors passats | El model s'ha guardat a la variable model_c.pulmonar | El model s'ha generat correctament i s'ha provat amb 300 registres predint la cp de manera correcta. A la predicció de prova s'han aconseguit els resultats esperats |
| 7 | Fes l'ANOVA unifactorial de cp per a tipo | Es guarda l'anova a la variable fum.aov | Es descarta l'hipotesi nul · la i s'accepta que hi ha diferències entre els factors de tipus |
| 8. | Comparacions multiples per esbrinar les diferències entre els tipus | S'aplica el pairwise.test sense i amb correcció de Bonferroni. Resultats equivalents | Els factors aparellats que presenten més variació a les mitjanes són, altra vegada, principalment tres grups. Els no fumadors es diferencien de tots. Els fumadors lleugers, moderats, i que no ingereixen. I finalment els fumadors moderats i intensius |

| N | Pregunta | Resultat | Conclusió |
|---|---|---|---|
| 9. | ANOVA multifactorial per a genero i tipo | Es guarda l'anova a la variable fum.aov2 | S'observa graficament una interacció entre les variables de genero i tipo, però al fer l'anova s'hi veu que sols presenta diferències a les mitjanes el factor de tipo. És a dir, el genere no fa variar les mitjanes significativament |

# 11. Resum executiu

En aquesta PAC he estudiat un dataset amb dades sobre fumadors de diversos tipus. Ho he fet amb l'objectiu d'esbrinar de quines maneres afecta el consum de tabac a la capacitat pulmonar dels consumidors. Les conclusions que s'han extret són:

- No hi ha diferències entre els fumadors homes i dones. En gènere no és un factor rellevant per a l'estudi de la capacitat pulmonar en relació al consum de tabac.
- A mesura que s'augmenta l'edat es va disminuint en capacitat pulmonar.
- El tipus de consumidor importen molt per esbrinar la pèrdua de capacitat pulmonar produïda pel tabac. Concretament, s'ha trobat que les persones que no fumen, o son fumadors passius, presenten una major capacitat pulmonar a qualsevol edat que les persones que son fumadores (de qualsevol tipus). Entrant en més detall, s'han observal principalment tres grups diferents pel que fa a la capacitat pulmonar:
  - Els no fumadors són els que millor capacitat pulmonar presenten.
  - Els fumadors passius, fumdors que no ingereixen el fum, o els fumadors lleugers (d'un a vint cigarrets al dia, durant >=20 anys) presenten reduccions en les seues capacitats pulmonars. Aquesta capacitat pulmonar varia prou segons altres factors (com l'edat, o el propi tipus), però es sap que és menor que la dels no fumadors i semblant per al grup. Cal destacar que els fumadors passius també veuen reduïda la seua capacitat pulmonar.
  - Els fumadors moderats (11-39 cigarrets al dia, durant >= 20 anys) i els fumadors intensius (>= 40 cigarrets al dia, durant >= 20 anys) presenten greus reduccions de la capacitat pulmonar. Amb poca diferència relativa entre ambdós grups.