Estadística Avançada - Activitat 4

Solució

Semestre 2022.1

Índex

Introducció

En una recerca mèdica es va estudiar la capacitat pulmonar dels fumadors i no fumadors. Es van recollir dades d'una mostra de la població fumadora, no fumadora i fumadors passius. A cada persona es va realitzar un test de capacitat pulmonar consistent a avaluar la quantitat d'aire expulsat (AE).

La mostra de n individus es va categoritzar en 6 tipus:

- No fumadors (NF)
- Fumadors passius (FP)
- Fumadors que no inhalen (NI): persones que fumen però no inhalen el fum.
- Fumadors lleugers (FL): persones que fumen i inhalen d'un a 10 cigarrets al dia durant 20 anys o més.
- Fumadors moderats (FM): persones que fumen i inhalen entre 11 i 39 cigarrets per dia durant 20 anys o més.
- Fumadors intensius (FI): persones que fumen i inhalen 40 cigarrets o més durant 20 anys o més.

En aquesta activitat s'analitzarà si la capacitat pulmonar està influïda pel tipus de fumador. Per a això, s'apliquessin diferents tipus d'anàlisis, revisant el contrast d'hipòtesis de dues mostres, vists en l'activitat A2, i després realitzant anàlisis més complexes com ANOVA.

Notes importants a tenir en compte per al lliurament de 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 d'incloure un índex o taula de continguts. I s'ha de respectar la numeració dels apartats de l'enunciat.
- No realitzeu llistats dels conjunts de dades, ja que aquests poden ocupar diverses pàgines. Si voleu comprovar l'efecte d'una instrucció sobre un conjunt de dades podeu usar la funció **head** i **tail** que mostren les primeres o últimes files del conjunt de dades.

---

1 Preprocessament

Carregar el fitxer de dades "Fumadores.csv". Consulteu els tipus de dades de les variables i si és necessari, apliqueu les transformacions apropiades. Esbrinar possibles inconsistències en els valors de Tipus, AE, gènere i edat. En cas que existeixin inconsistència, corregiu-les.

```
filename="Fumadores.csv"
 data <- read.csv( filename, sep=";")
 head(data)
```

```
##         AE Tipo genero edad
## 1 1.871878   NF      M   54
## 2  1.91312   NF      F   60
## 3  2.58114   NF      M   40
## 4  2.17827   NF      F   55
```

```
## 5 1.707732   NF      F   59
## 6 1.561215   NF      F   63
```

```
sapply( data, class)
```

```
##          AE        Tipo      genero        edad
## "character" "character" "character"   "integer"
```

```
summary(data)
```

```
##       AE                Tipo              genero               edad
##  Length:253         Length:253         Length:253         Min.   :17.00
##  Class :character   Class :character   Class :character   1st Qu.:43.00
##  Mode  :character   Mode  :character   Mode  :character   Median :50.00
##                                                           Mean   :49.76
##                                                           3rd Qu.:57.00
##                                                           Max.   :78.00
```

```
str( data )
```

```
## '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 ...
```

```
#Revisem Tipo
unique( data$Tipo )
```

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

```
data[ data$Tipo=="fi", ]$Tipo <- "FI"
data[ data$Tipo=="fm", ]$Tipo <- "FM"
data$Tipo<-trimws( data$Tipo )
data$Tipo <- as.factor( data$Tipo )
levels( data$Tipo )
```

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

```
#Revisem gènere
unique(data$genero)
```

```
## [1] "M" "F"
```

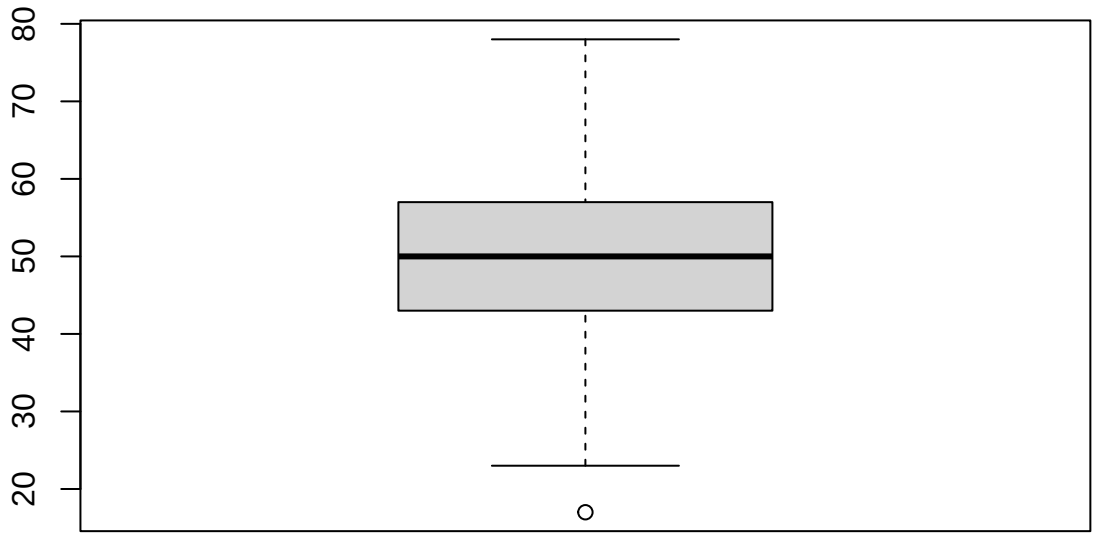
```
#Revisem AE. Coma y punto decimal
data$AE[ grep(",", data$AE) ]
```

```
##  [1] "1,885287" "1,990184" "2,09365"  "1,70995"  "1,25422"  "1,58875"
##  [7] "1,644625" "1,004136" "1,581052" "1,665934" "0,942632" "1,58774"
## [13] "1,085856" "0,44163"  "1,714654"
```

```
data$AE<-as.numeric( gsub( ",", "\\.", data$AE))

#Revisem edat
class(data$edad)
```

```
## [1] "integer"
```

```
boxplot(data$edad)
```




```
#Valors extrems en campo edat
data$edad[data$edad<20]
```

```
## [1] 17
```

```
#Possibles inconsistències entri edat y tipus de fumador
data[data$edad<33 & data$Tipo=="FL",]
```

```
##          AE Tipo genero edad
## 162 1.94971   FL      M   30
```

```
data[data$edad<33 & data$Tipo=="FI",]
```

```
##           AE Tipo genero edad
## 230 0.976464   FI      M   28
## 236 1.469072   FI      M   32
## 242 1.477476   FI      F   23
```

```
data[data$edad<33 & data$Tipo=="FM",]
```

```
## [1] AE     Tipo   genero edad
## <0 rows> (or 0-length row.names)
```

Preprocés realitzat:

- S'ha normalitzat el format de número de AE, corregint la coma decimal pel punt decimal.
- Es normalitza el format del Tipus de fumador.
- Es troben incosistencias entre edat i Tipus de fumador (edats inferiors a 30 anys amb fumadors lleugers i intensius que fumen durant més de 20 anys). Per falta d'informació, aquests registres es deixen intactes.

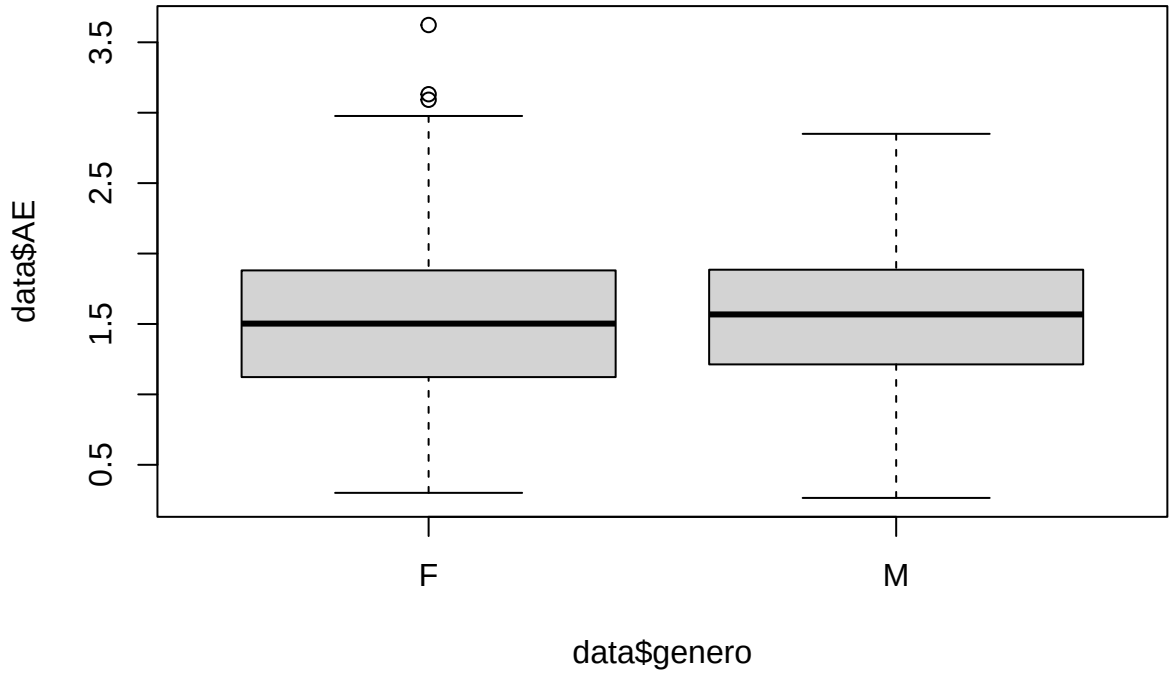
Però es podrien eliminar o realitzar una imputació en el valor d'edat o tipus de fumador. Per falta d'informació, es deixa tal com està.

2 Anàlisi descriptiva de la mostra

2.1 Capacitat pulmonar i gènere

Mostreu la capacitat pulmonar en relació al gènere. S'observen diferències?

```
boxplot( data$AE ~ data$genero )
```

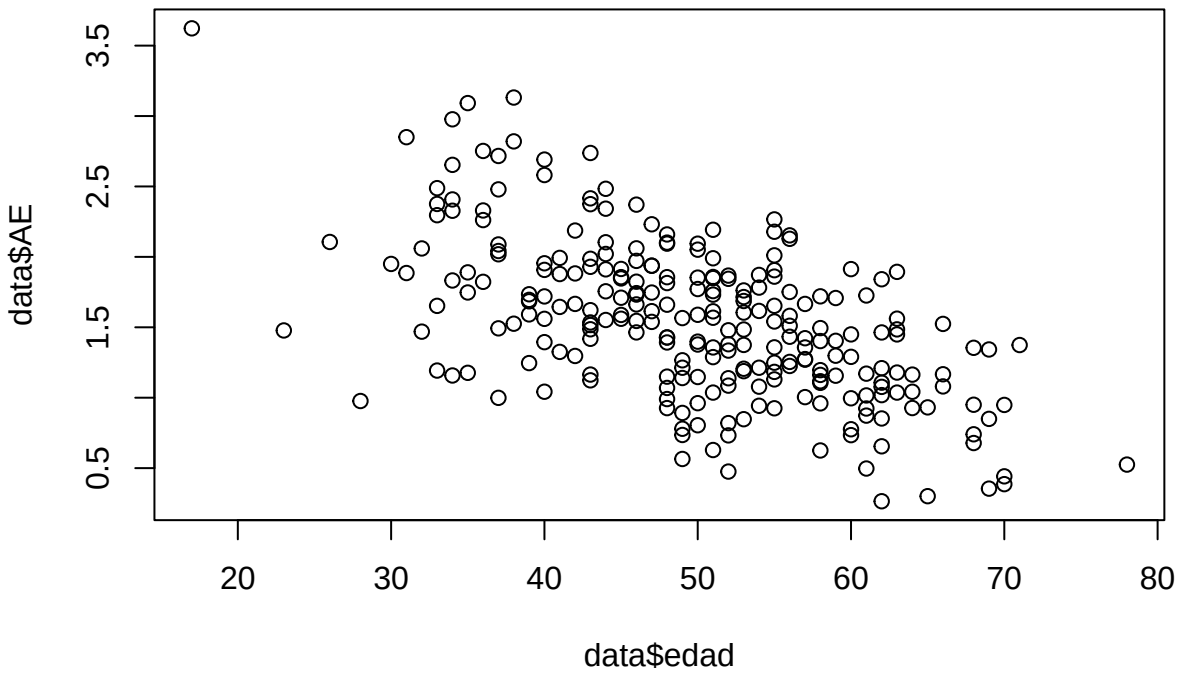


Pràcticament no s'observen diferències entre capacitat pulmonar i gènere.

2.2 Capacitat pulmonar i edat

Mostreu la relació entre capacitat pulmonar i edat usant un gràfic de dispersió. Interpreteu.

```
plot( data$edad, data$AE )
```



S'observa una tendència a la baixa en la capacitat pulmonar a mesura que augmenta l'edat.

2.3 Tipus de fumadors i capacitat pulmonar

Mostreu el nombre de persones en cada tipus de fumador i la mitjana de AE de cada tipus de fumador. Mostreu un gràfic que visualitzi aquesta mitjana. Es recomana que el gràfic estigui ordenat de menys a més AE.

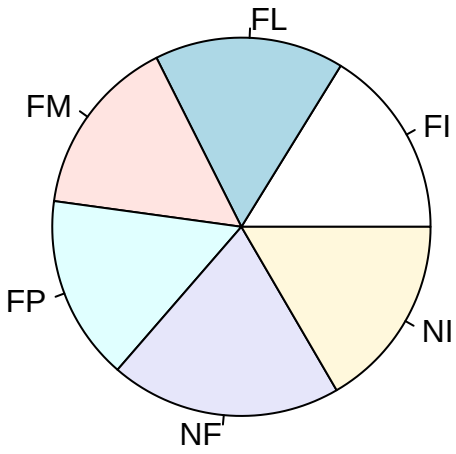
Després, s'ha de representar un boxplot on es mostri la distribució de AE per cada tipus de fumador. Interpreteu els resultats.

Nota: Per a calcular la mitjana o altres variables per a cada tipus de fumador, podeu usar les funcions **summarize** i **group_by** de la llibreria **dplyr** que us seran de gran utilitat. Per a realitzar la visualització de les dades, podeu usar la funció **ggplot** de la llibreria **ggplot2**.

```
#Nombre de persones per cada tipus de fumador
table( data$Tipo )
```

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

```
pie(table(data$Tipo))
```




```
#Estadístiques de cada grup
DS <- summarize( group_by(data, Tipo), AEmedia=mean(AE), n=length(AE),
                 sd=sd(AE), edadmedia=mean(edad),
                 fem=length(genero[genero=="F"]),
                 male=length(genero[genero=="M"]))
DS
```

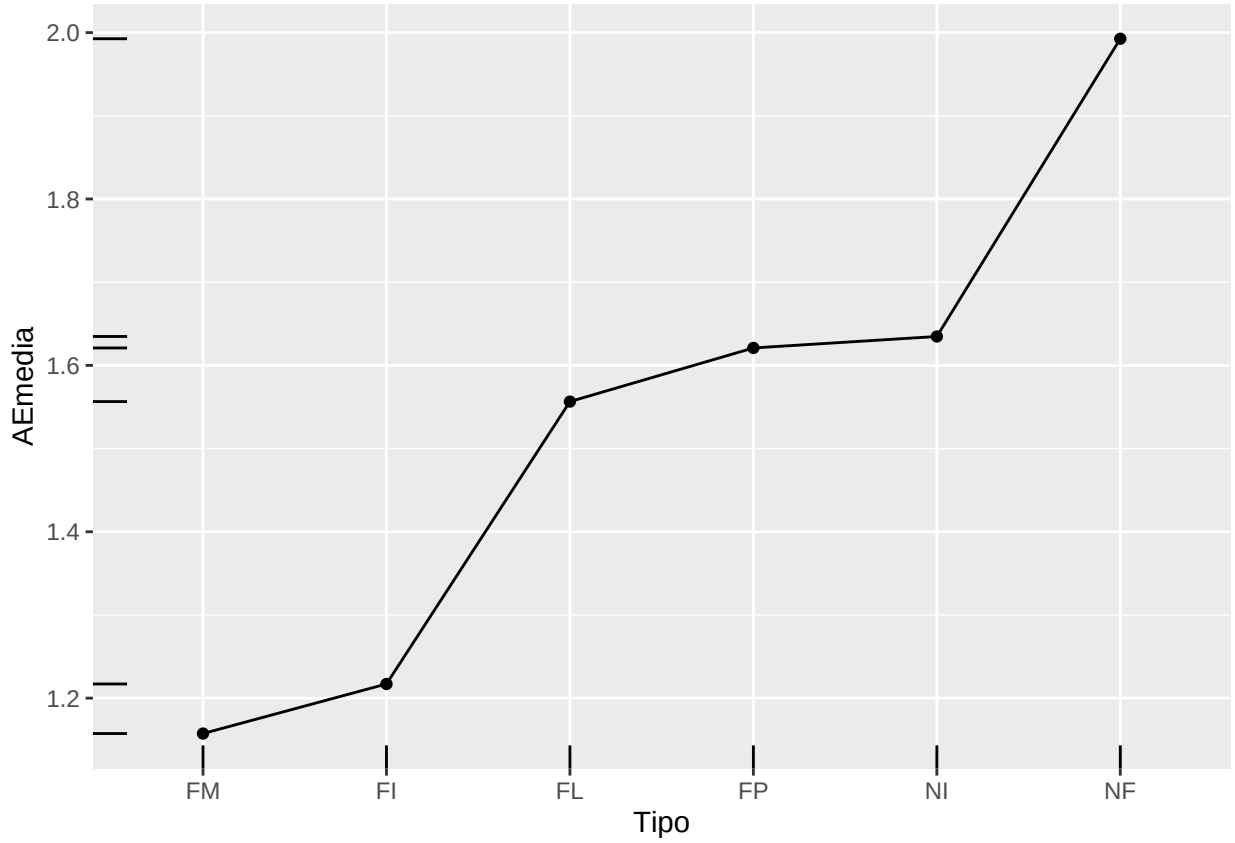
```
## # A tibble: 6 x 7
##   Tipo  AEmedia     n    sd edadmedia   fem  male
##   <fct>   <dbl> <int> <dbl>     <dbl> <int> <int>
## 1 FI       1.22    41 0.465      49.2    24    17
## 2 FL       1.56    41 0.484      49.2    28    13
## 3 FM       1.16    39 0.421      52.6    22    17
## 4 FP       1.62    40 0.518      48.9    18    22
## 5 NF       1.99    50 0.536      49.4    29    21
## 6 NI       1.63    42 0.451      49.4    23    19
```

```
#Preparem el dataset per a mostrar un gràfic de la mitjana, ordenat segons mitjana.
DS$Tipo <- factor( DS$Tipo, levels=DS$Tipo[order(DS$AEmedia)])
library(ggplot2)
```

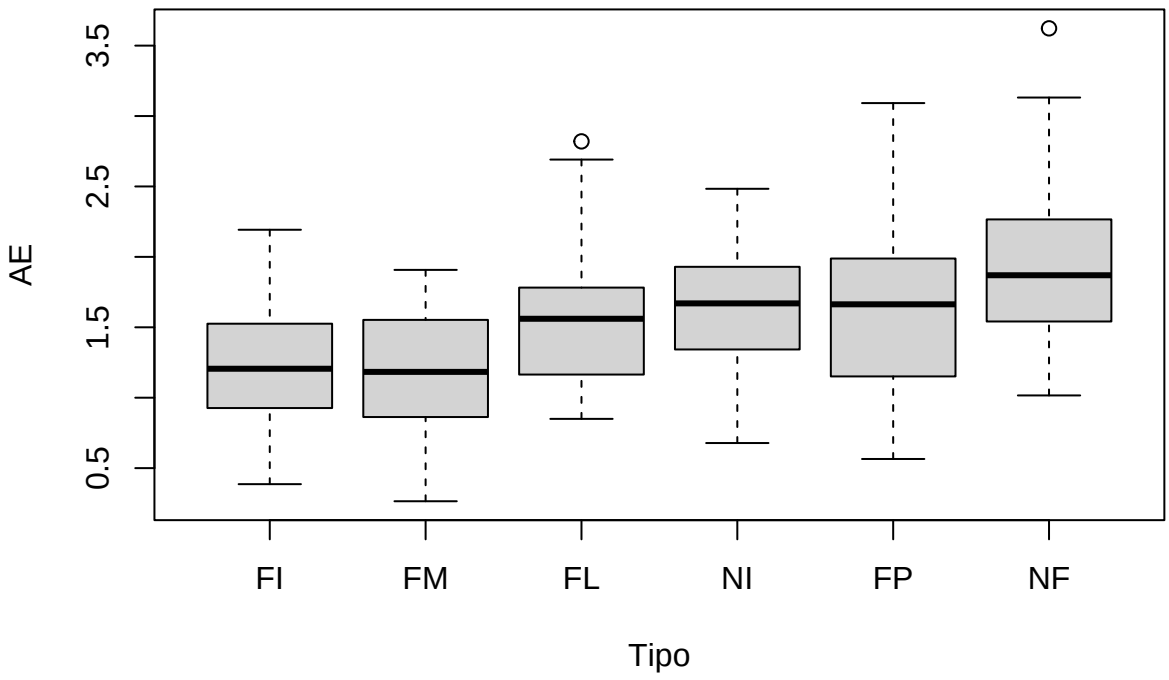
```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
ggplot(DS, aes(x=Tipo, y=AEmedia, group=1)) +
  geom_point() + geom_line() + geom_rug()
```




```
#Ordenem segons tipus de fumador
data$Tipo <- factor( data$Tipo, levels=c("FI","FM","FL","NI","FP","NF"))
boxplot( AE~Tipo, data)
```



La mostra conté aproximadament el mateix nombre de persones per cada tipus de fumador. Quant a la capacitat pulmonar s'observen diferències entre els tipus de fumador, sent el tipus "no fumador" el que té major capacitat pulmonar i els fumadors intensius i moderats els que tenen menor capacitat pulmonar.

---

3 Interval de confiança de la capacitat pulmonar

Calculeu l'interval de confiança al 95% de la capacitat pulmonar de les dones i homes per separat. Abans d'aplicar el càlcul, reviseu si es compleixen les assumpcions d'aplicació de l'interval de confiança. Interpreteu els resultats. A partir d'aquests càlculs, s'observen diferències significatives en la capacitat pulmonar de dones i homes?

Resposta: Com la mostra és superior a 30, podem assumir que la mitjana de AE segueix una distribució normal, segons el teorema del límit central.

```
n<-length( data$AE )
n
```

```
## [1] 253
```

```
my.IC <- function(x){
  alfa <- 0.05
  error.estandar <- sd( x ) / sqrt(n)
  z <- qnorm( 0.025, lower.tail=FALSE )
  margen.error <- z* error.estandar
  media<- mean(x)
  IC.inf <- media - margen.error
```

```
  IC.sup <- media + margen.error
  return (c(IC.inf, IC.sup))
}

IC.F<-my.IC( data[data$genero=="M",]$AE ); IC.F
```

```
## [1] 1.517912 1.649613
```

```
IC.M<-my.IC( data[data$genero=="F",]$AE ); IC.M
```

```
## [1] 1.452326 1.594234
```

Com veiem, els intervals de confiança estan solapats i per tant, no podem afirmar que existeixin diferències en la capacitat pulmonar entre homes i dones, com ja havíem observat visualment en la secció anterior.

---

4 Diferències en capacitat pulmonar entre dones i homes

Apliqueu un contrast d'hipòtesi per a avaluar si existeixen diferències significatives entre la capacitat pulmonar de dones i homes. Seguiu els passos que s'indiquen a continuació.

Nota: Realitzeu el càlcul manualment sense usar les funcions `t.test` o equivalents. Podeu usar `qnorm`, `qt`, `pnorm`, `pt`,.

4.1 Hipòtesi

Escriure la hipòtesi nul · la i alternativa.

$H_0 : \mu_F = \mu_M$

$H_1 : \mu_F \neq \mu_M$

4.2 Contrast

Expliqueu quin tipus de contrast aplicareu i per què. Si és necessari, valideu les assumpcions del test.

Resposta: apliquem un contrast de dues mostres independents sobre la mitjana. Comprovem si podem assumir homocedasticitat.

```
var.test( data[data$genero=="M",]$AE , data[data$genero=="F",]$AE )
```

```
##
##  F test to compare two variances
##
## data:  data[data$genero == "M", ]$AE and data[data$genero == "F", ]$AE
## 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
```

El resultat del test no mostra diferències significatives entre variàncies. Per tant, aplicarem un test de dues mostres independents sobre la mitjana amb variàncies desconegudes iguals. El test és bilateral.

4.3 Càlculs

Apliqueu els càlculs del contrast. Mostreu el valor observat, el valor de contrast i el valor p.

```
my.ttest <- function( x1, x2, CL=95, alternative="two.sided" ){
  mean1<-mean(x1); n1<-length(x1); sd1<-sd(x1)
  mean2<-mean(x2); n2<-length(x2); sd2<-sd(x2)
  alfa <- (1-CL/100)

  #variançies iguals
    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
    lt<-FALSE

  if (alternative=="two.sided"){
      tcritical <- qt( alfa/2, df, lower.tail=FALSE )    #two sided
      pvalue<-pt( abs(t), df, lower.tail=FALSE )*2       #two sided
  }
  else{
      lt <- ifelse(alternative=="less", TRUE, FALSE)
      tcritical <- qt( alfa, df, lower.tail=lt )
      pvalue<-pt( t, df, lower.tail=lt )
  }

  #Guardem el resultat en un named vector
  info<-c(mean1, mean2, t,tcritical,pvalue,df)
  names(info)<-c("mean1", "mean2", "t","tcritical", "pvalue", "df")
  return (info)
}

tAE.FM<-my.ttest( data[data$genero=="M",]$AE , data[data$genero=="F",]$AE, alternative="two.sided")
tAE.FM
```

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

```
#Comprovació:
t.test( data[data$genero=="M",]$AE , data[data$genero=="F",]$AE, alternative="two.sided", var.equal=TRU
```

```
##
##  Two Sample t-test
##
## data:  data[data$genero == "M", ]$AE and data[data$genero == "F", ]$AE
## t = 0.85316, df = 251, p-value = 0.3944
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.07913636  0.20010078
## sample estimates:
## mean of x mean of y
##  1.583762  1.523280
```

4.4 Interpretació

Interpreteu els resultats i compareu les conclusions amb els intervals de confiança calculats anteriorment.

Resposta: No podem rebutjar la hipòtesi nul · la. Per tant, no existeixen diferències significatives en la

capacitat pulmonar entre homes i dones amb un nivell de confiança del 95%. Aquesta conclusió és consistent amb el càlcul dels intervals de confiança realitzat anteriorment.

---

5 Diferències en la capacitat pulmonar entre Fumadors i No Fumadors

Podem afirmar que la capacitat pulmonar dels fumadors és inferior a la de no fumadors? Incloeu dins de la categoria de no fumadors els fumadors passius. Seguiu els passos que s'indiquen a continuació.

Nota: Realitzar el càlcul manualment sense usar les funcionis `t.test` o equivalents. Podeu usar `qnorm`, `qt`, `pnorm`, `pt`,.

5.1 Hipòtesi

Escriure la hipòtesi nul · la i alternativa.

$H_0 : \mu_{FUM} = \mu_{NFUM}$

$H_1 : \mu_{FUM} < \mu_{NFUM}$

5.2 Contrast

Expliqueu quin tipus de contrast aplicareu i per què. Si és necessari, valideu les assumpcions del test.

Resposta: apliquem un contrast de dues mostres independents sobre la mitjana. Comprovem si podem assumir homocedasticitat.

```
Fum <- data[data$Tipo!="NF" & data$Tipo!="FP", ]
NFum<- data[data$Tipo=="NF" | data$Tipo=="FP", ]

var.test( Fum$AE, NFum$AE )
```

```
##
##  F test to compare two variances
##
## data:  Fum$AE and NFum$AE
## 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
```

El resultat del test no mostra diferències significatives entre variàncies. Per tant, aplicarem un test de dues mostres independents sobre la mitjana amb variàncies desconegudes iguals. El test és bilateral.

Resposta: Contrast de mitjanes de dues mostres independents. Assumim distribució normal i cas de variància poblacional desconegudes iguals. Apliquem un contrast unilateral.

5.3 Preparació de les dades

```
#S'han creat anteriorment les mostres Fum i NFum
n1 <- nrow( Fum )
n2 <- nrow( NFum )
```

```
n1; n2
```

```
## [1] 163
```

```
## [1] 90
```

5.4 Càlculs

Apliqueu els càlculs del contrast. Mostreu el valor observat, el valor de contrast i el valor p.

```
#Usem la funció my.ttest
tAE.FNF<-my.ttest( Fum$AE, NFum$AE, alternative="less" ); tAE.FNF
```

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

```
#Comprovació amb t.test
t.test( Fum$AE, NFum$AE, alternative="less", var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Fum$AE and NFum$AE
## t = -6.3298, df = 251, p-value = 5.613e-10
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.3190665
## sample estimates:
## mean of x mean of y
##  1.395786  1.827437
```

5.5 Interpretació

Interpreteu el resultat del contrast.

Atès que el valor p és menor que $\alpha = 0.05$ rebutgem la hipòtesi nul · la a favor de la hipòtesi alternativa, segons la qual la mitjana de la capacitat pulmonar dels fumadors és inferior a la dels no fumadors amb un nivell de confiança del 95%.

---

6 Anàlisi de regressió lineal

Realitzem una anàlisi de regressió lineal per a investigar la relació entre la variable capacitat pulmonar (AE) i la resta de variables (tipus, edat i gènere). Construïu i interpreteu el model.

6.1 Càlcul

Construïu el model de regressió lineal.

```
  mylm <- lm( AE ~ ., data)
  summary(mylm)
```

```
##
## Call:
```

```
## lm(formula = AE ~ ., data = data)
##
## 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)  2.741411   0.128797  21.285  < 2e-16 ***
## TipoFM       0.046357   0.082133   0.564    0.573
## TipoFL       0.338459   0.080850   4.186 3.96e-05 ***
## TipoNI       0.423523   0.080259   5.277 2.89e-07 ***
## TipoFP       0.394342   0.081470   4.840 2.30e-06 ***
## TipoNF       0.781808   0.077004  10.153  < 2e-16 ***
## generoM     -0.002321   0.047033  -0.049    0.961
## 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

Interpreteu el model i la contribució de cada variable explicativa sobre la variable AE.

Resposta: La variable edat influeix en la capacitat pulmonar amb un coeficient negatiu, és a dir, que a mesura que augmenta l'edat disminueix la capacitat pulmonar. La variable Tipus de Fumador és significativa. La categoria de referència és FI (intensiu). Hi ha diferències significatives en AE entre tots els tipus de fumadors excepte el moderat, en relació amb el fumador intensiu. Finalment, el gènere no influeix a la capacitat pulmonar. Els resultats són consistents amb els tests realitzats anteriorment sobre gènere, edat i tipus de fumador, i l'anàlisi visual que s'ha mostrat en relació amb els tipus de fumador.

6.3 Bondat d'ajust

Avalueu la qualitat del model.

Resposta: El model explica el 58,3% de la variabilitat en la capacitat pulmonar. Probablement hi ha altres variables que influeixen en la capacitat pulmonar i que no estan incloses al model, com la realització d'exercici físic o si la persona viu en un entorn amb alta contaminació.

6.4 Predicció

Realitzeu una predicció de la capacitat pulmonar per a cada tipus de fumador des dels 30 anys d'edat fins als 80 anys d'edat (podeu assumir gènere home). Mostreu una taula amb els resultats. Mostreu també visualment la simulació.

```
rango.edad <- seq(30,80,1)
N<-length(rango.edad); N
```

```
## [1] 51
```

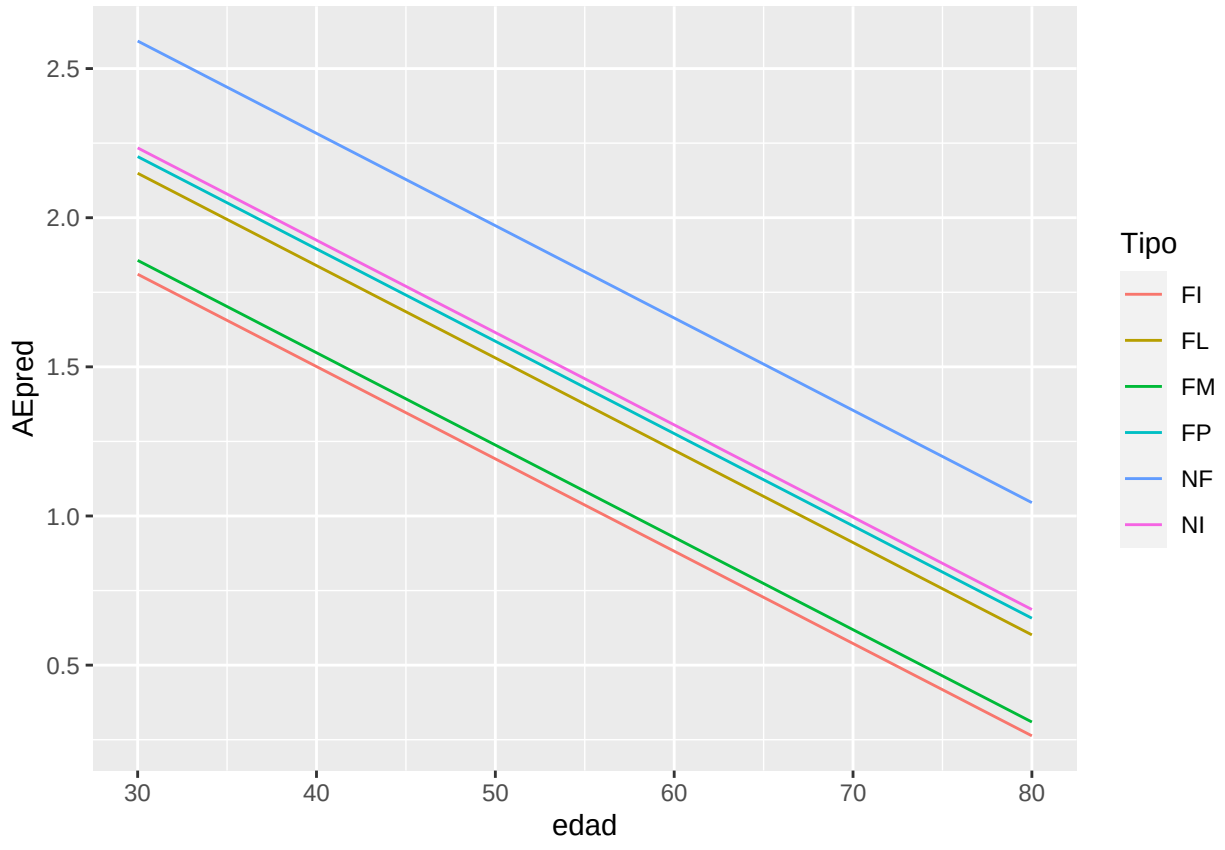
```
tipo <- c("NF", "FP", "NI", "FL", "FM", "FI")
rango.tipo<-sort( rep( tipo, N) )

sim <- data.frame( Tipo=rango.tipo, genero="M", edad=rango.edad ); head(sim)
```

```
##   Tipo genero edad
## 1   FI      M   30
## 2   FI      M   31
## 3   FI      M   32
## 4   FI      M   33
## 5   FI      M   34
## 6   FI      M   35
```

```
  sim$AEpred <- predict( mylm, sim)
  head(sim)
```

```
##   Tipo genero edad   AEpred
## 1   FI      M   30 1.810547
## 2   FI      M   31 1.779596
## 3   FI      M   32 1.748645
## 4   FI      M   33 1.717693
## 5   FI      M   34 1.686742
## 6   FI      M   35 1.655790
```

```
#sim$Tipo <- factor( sim$Tipo, levels=DS$Tipo[order(DS$AEmedia)])
library(ggplot2)
ggplot(sim, aes(x=edad, y=AEpred, group=Tipo,color=Tipo)) +
  geom_line()
```



7 ANOVA unifactorial

A continuació es realitzarà una anàlisi de variància, on es desitja comparar la capacitat pulmonar entre els sis tipus de fumadors/no fumadors classificats prèviament. L'anàlisi de variància consisteix a avaluar si la variabilitat d'una variable dependent pot explicar-se a partir d'una o diverses variables independents, denominades factors. En el cas que ens ocupa, ens interessa avaluar si la variabilitat de la variable AE pot explicar-se pel factor tipus de fumador.

Hi ha dues preguntes bàsiques a respondre:

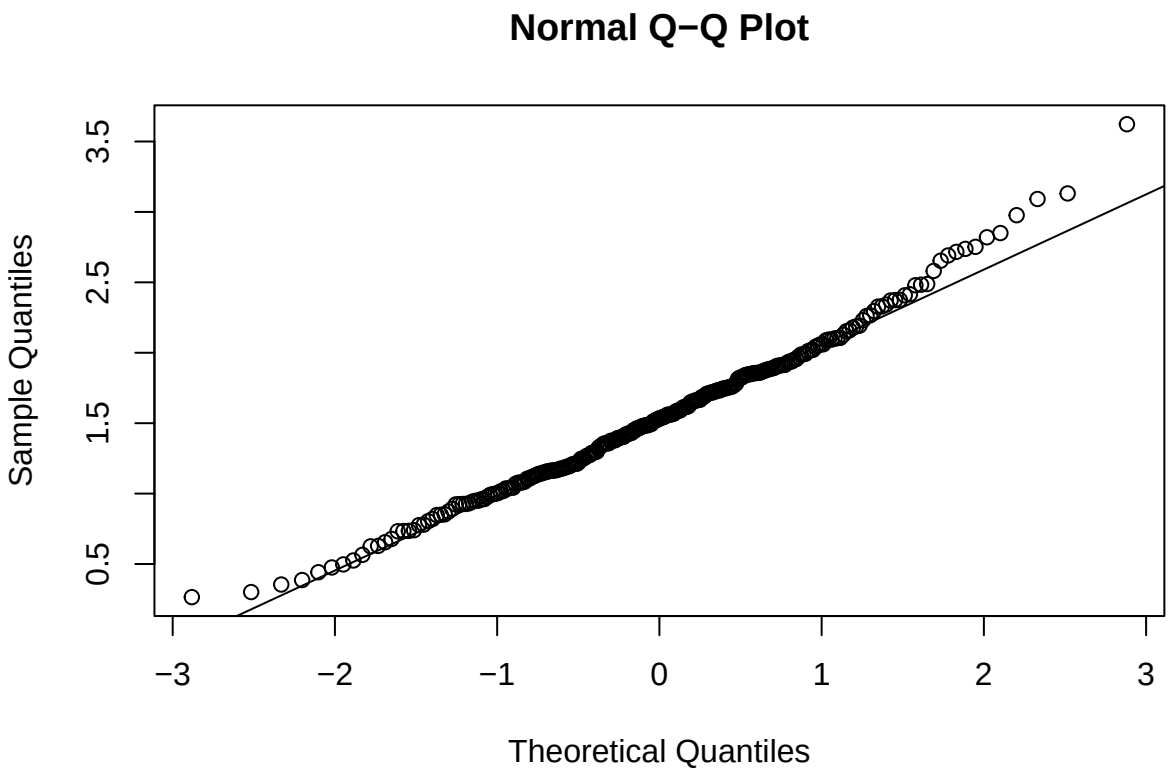
- Existeixen diferències entre la capacitat pulmonar (AE) entre els diferents tipus de fumadors/no fumadors?
- Si existeixen diferències, entre quins grups estan aquestes diferències?

7.1 Normalidad

Evaluar si el conjunto de datos cumple las condiciones de aplicación de ANOVA. Seguid los pasos que se indican a continuación. Mostrad visualmente si existe normalidad en los datos y también aplicar un test de normalidad.

Nota: Podeu usar el gràfic Normal Q-Q i el test Shapiro-Wilk, per a avaluar la normalitat dels residus.

```
qqnorm(data$AE)
qqline(data$AE)
```




```
#H0: la mostra (de mida n) segueix una distribució normal
#es rebutja  H0 si p value < alfa
```

```
#Si se aplica Shapiro (en tota la mostra)
ST <- shapiro.test(data$AE)
ST
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$AE
## W = 0.98869, p-value = 0.04484
```

```
class(ST)
```

```
## [1] "htest"
```

```
pvalue<-ST[[2]]   #pvalue
pvalue
```

```
## [1] 0.04483706
```

Interpretació: Al test de Shapiro-Wilk, si $Pr(D) \leq \alpha$ es rebutjaria la hipòtesi nul · la de normalitat en les dades.

El valor p del test de Shapiro ha donat 0.0448371. Per tant, es rebutjaria la hipòtesi nul · la de normalitat, encara que aquesta desviació respecte a la normalitat no és gaire pronunciada.

La condició de normalitat s'ha de complir per a cada grup. Per això, cal aplicar la prova de normalitat a cada grup (tipus de fumador). També es valora que es representi el gràfic per a cada tipus de fumador.

```
#"The distribution of Y within each group is normally distributed."  It's the same thing as Y|X and in

DS <- summarize( group_by(data, Tipo), n=length(AE), p.shapiro=shapiro.test(AE)[[2]])
DS
```

```
## # A tibble: 6 x 3
##   Tipo      n p.shapiro
##   <fct> <int>     <dbl>
## 1 FI       41    0.607
## 2 FM       39    0.234
## 3 FL       41    0.0415
## 4 NI       42    0.783
## 5 FP       40    0.404
## 6 NF       50    0.0364
```

El test de Shapiro-Wilk presenta en tots els grups (excepte un) valors p superiors a 0.05. Estrictament, no es podria rebutjar la hipòtesi nul.la de normalitat, encara que com la desviació és poc pronunciada, seguim amb la aplicació de ANOVA paramètric.

7.2 Homocedasticitat: Homogeneïtat de variàncies

Altra de les condicions d'aplicació de ANOVA és la igualtat de variàncies (homocedasticitat). Aplicar un test per a validar si els grups presenten igual variància. Apliqueu el test adequat i interpreteu el resultat.

Nota: podeu fer servir tests com el de Levene o Bartlett test.

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma_6^2$

H_1 : Almenys hi ha diferències entre dos grups: $\sigma_i^2 \neq \sigma_j^2$

```
#Levene Test
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
LT <- leveneTest(AE ~Tipo, data)
LT
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   5  0.4241 0.8317
##       247
```

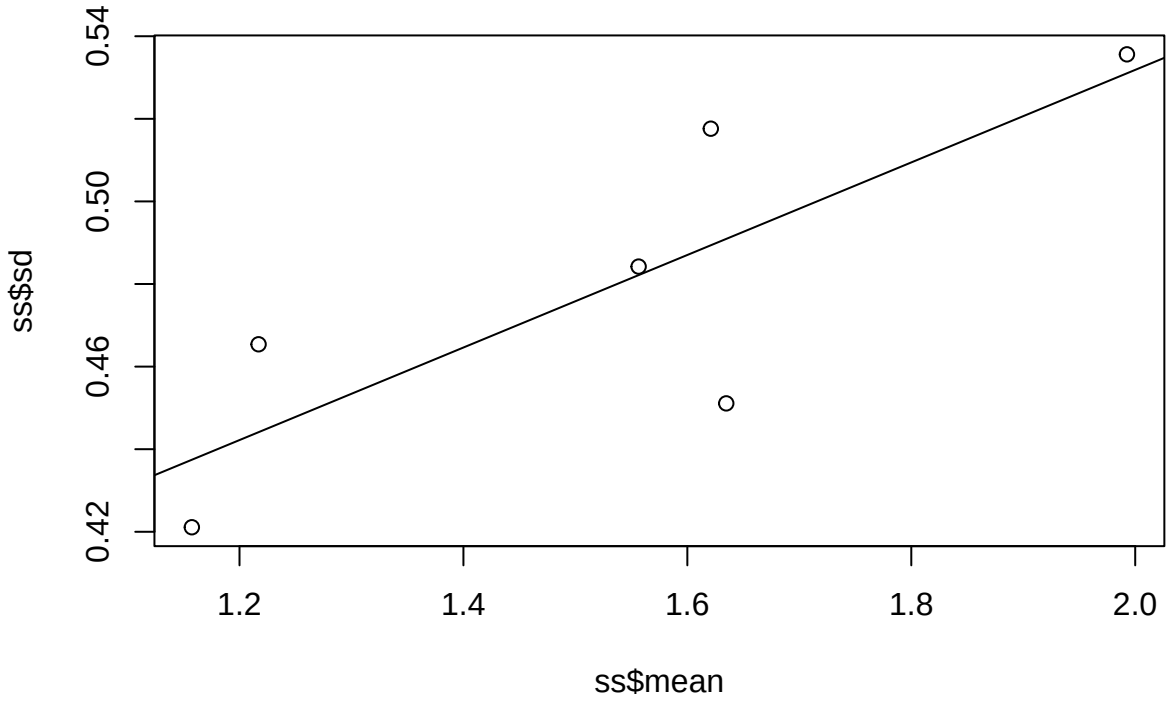
```
LT$`F value`[1]
```

```
## [1] 0.4241176
```

```
pvalue<-LT$`Pr(>F)`[1]; pvalue
```

```
## [1] 0.8316893
```

```
#Gràfic de dispersió per Tipo. y=dispersió, x=mitjana del grup
ss <- summarize( group_by(data, Tipo), sd=sd(AE), mean=mean(AE))
reg<-lm(sd ~ mean, data = ss)
plot( ss$mean, ss$sd )
abline(reg)
```




```
# Bartlett Test of Homogeneity of Variances
bartlett.test(AE~Tipo, data)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  AE by Tipo
## Bartlett's K-squared = 3.2658, df = 5, p-value = 0.6591
```

```
# Figner-Killeen Test of Homogeneity of Variances. No paramétrico
fligner.test(AE~Tipo, data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  AE by Tipo
## Fligner-Killeen:med chi-squared = 1.6636, df = 5, p-value = 0.8935
```

Interpretació del test de Levene: El valor del test de Levene és $Pr(F)$=0.8316893. Per a un nivell de significació $\alpha = 0,05$, $Pr(F) \geq \alpha$. Per tant, no es rebutja la hipòtesi nul · la d'igualtat de variàncies. Es compleix la condició d'homoscedasticitat. Altres tests d'homogeneïtat de variàncies com el test Barlett o Fligner-Killeen obtenen resultats anàlegs.

7.3 Hipòtesi nul · la i alternativa

Independentment dels resultats sobre la normalitat u homocedasticitat de les dades, prosseguirem amb l'aplicació de l'anàlisi de variància. Concretament, s'aplicarà ANOVA d'un factor (one-way ANOVA o independent samples ANOVA) per a investigar si existeixen diferències en el nivell d'aire expulsat (AE) entre

els diferents tipus de fumadors. Escriviu la hipòtesi nul · la i alternativa.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

H_1 : Almenys hi ha diferències entre dos grups: $\mu_i \neq \mu_j$

7.4 Càlcul ANOVA

Podeu usar la funció **aov**.

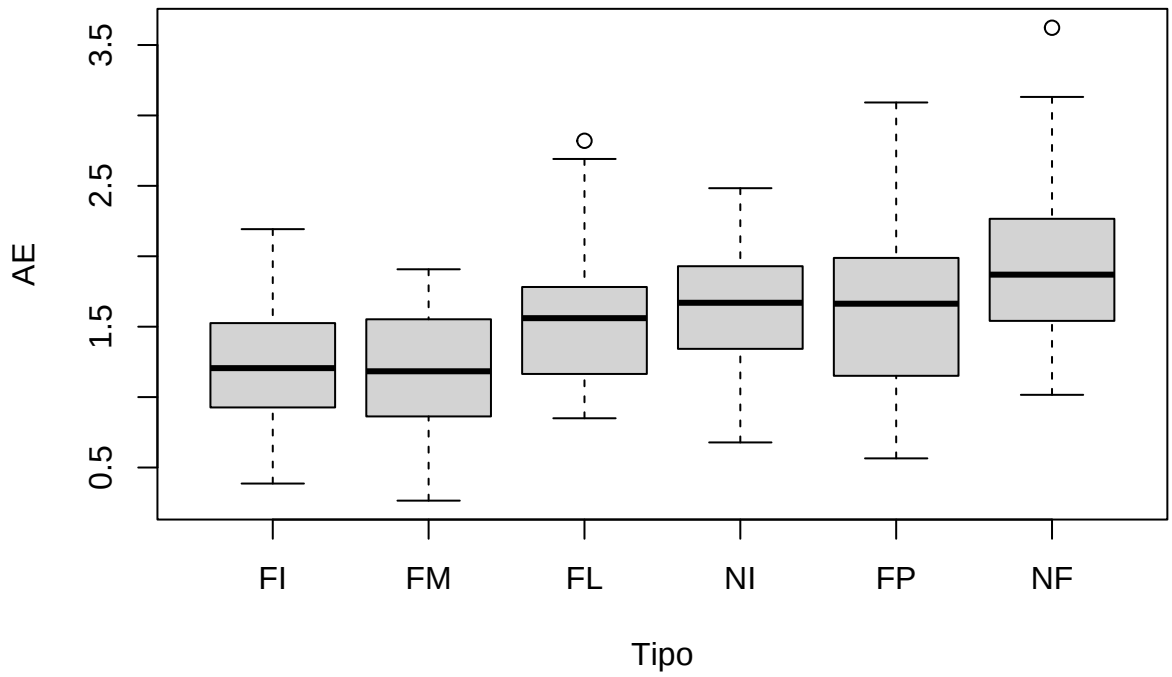
```
#Cálculo one-way ANOVA
my.aov <- aov( AE~Tipo, data )
sum.aov <- summary( my.aov)
sum.aov
```

```
##              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

Interpreteu els resultats de la prova ANOVA i relacionar-los amb el resultat gràfic del boxplot mostrat en l'apartat 2.3.

```
data$Tipo <- factor( data$Tipo, levels=c("FI","FM","FL","NI","FP","NF"))
boxplot( AE~Tipo, data)
```



L'estadístic F=17.8774417.

El valor p és $4.0257861 \times 10^{-15}$. Per tant, podem rebutjar la hipòtesi nul · la que les diferències entre els grups siguin iguals. Aquest resultat s'observa visualment al diagrama de caixes (boxplot), on s'observen diferències entre les mitjanes dels grups.

7.6 Aprofundint en ANOVA

A partir dels resultats del model retornat per **aov**, identificar les variables SST (Total Sum of Squares), SSW (Within Sum of Squares), SSB (Between Sum of Squares) i els graus de llibertat. A partir d'aquests valors, calcular manualment el valor F, el valor crític (a un nivell de confiança del 95%), i el valor p. Interpreteu els resultats i expliqueu el significat de les variables SST, SSW i SSB.

```
#Càlculs
SSB <- sum.aov[[1]]$`Sum Sq`[1]
SSW<- sum.aov[[1]]$`Sum Sq`[2]
SST<-SSB + SSW
k<-length( levels(data$Tipo))
n<-length(data$AE)
F <- ( SSB / (k-1)) / ( SSW / (n-k))

#observed statistic
F<- (SSB/(k-1))/(SSW/(n-k))
F
```

```
## [1] 17.87744
```

```
#critical value
f.critical <- qf( 0.05, df1=k-1, df2=n-k, lower.tail=FALSE )
f.critical
```

```
## [1] 2.250576
```

```
#p value
p.value <- pf( F, df1=k-1, df2=n-k, lower.tail=FALSE)
p.value
```

```
## [1] 4.025786e-15
```

Com es pot observar, el càlcul de F es realitza a partir de la variància entre grups que és SSB/(k-1), on SSB=20.855837 i (k-1)=5. El denominador és la variància dins dels grups i correspon a SSW/(n-k), on SSW=57.6300772, i (n-k)=247. El còmput de F dóna 17.8774417, el qual coincideix amb el resultat del model anova calculat. El càlcul del valor p s'ha fet amb la funció `pf` a partir de l'estadística F i els graus de llibertat (k-1) i (n-k), respectivament.

7.7 Força de la relació

Calculeu la força de la relació i interpreteu el resultat.

```
fuerza <- SSB / SST
fuerza
```

```
## [1] 0.2657271
```

Interpretació:

La força de la relació representa en quina mesura el coneixement del grup de pertinença determina el valor a la variable dependent. Segons el resultat els grups a què pertany una persona explica el 26.5727133 % de la variabilitat en la capacitat pulmonar.

8 Comparacions múltiples

Independentment del resultat obtingut en l'apartat anterior, realitzem un test de comparació múltiple entre els grups. Aquest test s'aplica quan el test ANOVA retorna rebutjar la hipòtesi nul · la d'igualtat de mitjanes. Per tant, procedirem com si el test ANOVA hagués donat com a resultat el rebuig de la hipòtesi nul · la.

8.1 Test pairwise

Calculeu les comparacions entre grups sense cap mena de correcció. Podeu usar la funció **pairwise.t.test**. Interpreteu els resultats.

```
pairwise.t.test(data$AE, data$Tipo, p.adj = "none")
```

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

Interpretació:

- NF (no fumador): presenta diferències significatives amb tots els grups.
- FP (fumador passiu): Té capacitat pulmonar significativament diferent del No fumador, Fumador intensiu i Fumador moderat. Equivalent a FL (Fumador lleuger) ia NI (no inhala).
- NI (fumador no inhala): diferències significatives amb NF (no fumador), FI (fumador intensiu) i FM (fumador moderat). Equivalent a FP (passiu), FL (lleuger).
- FL (fumador lleuger): té AE significativament diferent del FI (intensiu) i FM (moderat). Equivalent a NI (no inhala) i FP (passiu). També és significativament diferent del NF (no fumador).
- FI (intensiu) i FM (moderat) són equivalents entre si.

8.2 Correcció de Bonferroni

Apliqueu la correcció de Bonferroni en la comparació múltiple. Interpreteu el resultat i contrasteu el resultat amb l'obtingut en el test de comparacions múltiples sense correcció.

```
library(DescTools)
```

```
##
## Attaching package: 'DescTools'

## The following object is masked from 'package:car':
##
##     Recode

## The following objects are masked from 'package:psych':
##
##     AUC, ICC, SD
```

```
  PostHocTest( my.aov, method="bonferroni")
```

```
##
##   Posthoc multiple comparisons of means : Bonferroni
##     95% family-wise confidence level
##
## $Tipo
##              diff      lwr.ci    upr.ci    pval
## FM-FI -0.05959277 -0.37983497 0.2606494 1.00000
## FL-FI  0.33944056  0.02322673 0.6556544 0.02477 *
## NI-FI  0.41770160  0.10337562 0.7320276 0.00160 **
## FP-FI  0.40391730  0.08573327 0.7221013 0.00315 **
## NF-FI  0.77558970  0.47394093 1.0772385 8.1e-12 ***
## FL-FM  0.39903333  0.07879113 0.7192755 0.00409 **
## NI-FM  0.47729437  0.15891614 0.7956726 0.00020 ***
## FP-FM  0.46351007  0.14132231 0.7856978 0.00043 ***
## NF-FM  0.83518247  0.52931345 1.1410515 4.0e-13 ***
## NI-FL  0.07826103 -0.23606494 0.3925870 1.00000
## FP-FL  0.06447674 -0.25370729 0.3826608 1.00000
## NF-FL  0.43614914  0.13450037 0.7377979 0.00039 ***
## FP-NI -0.01378430 -0.33009223 0.3025236 1.00000
## NF-NI  0.35788811  0.05821894 0.6575573 0.00717 **
## NF-FP  0.37167240  0.06795894 0.6753859 0.00522 **
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
  #pairwise.t.test(data$AE, data$Tipo, p.adj = "bonferroni")
```

Interpretació: Hi ha diferències significatives entre:

- NF (no fumador) i la resta de tipus.
- FI (intensiu) presenta diferències amb tots els tipus, excepte el FM (moderat).
- FM (moderat) presenta diferències amb tots els tipus, excepte el FI (intensiu).
- FL (lleuger) té diferències amb FP (passiu) i alhora amb FM (moderat) i FI (intensiu).
- NI presenta diferències amb FI (intensiu) i FM (moderat), a més de les diferències amb NF.
- FP (passiu) només presenta diferències amb FI (intensiu) i FM (moderat), a més de la diferència amb NF.

Es pot veure que detecta menys diferències. És un test més conservador.

9 ANOVA multifactorial

En una segona fase de la recerca es evalua l'efecte del gènere com a variable independent, a més de l'efecte del tipus de fumador, sobre la variable AE.

9.1 Anàlisi visual

Es realitzarà un primer estudi visual per a determinar si existeixen efectes principals o hi ha efectes d'interacció entre gènere i tipus de fumador. Per a això, seguir els passos que s'indiquen a continuació:

1. Agrupeu el conjunt de dades per tipus de fumador i gènere i calculeu la mitjana de AE en cada grup. Podeu usar les instruccions **group_by** i **summarise** de la llibreria **dplyr** per a realitzar aquest procés.

Mostreu el conjunt de dades en forma de taula, on es mostri la mitjana de cada grup segons el gènere i tipus de fumador.

2. Mostrar en un gràfic el valor de AE mitjà per a cada tipus de fumador i gènere. Podeu realitzar aquest tipus de gràfic usant la funció **ggplot** de la llibreria **ggplot2**.

3. Interpreteu el resultat sobre si existeixen només efectes principals o existeix interacció. Si existeix interacció, expliqueu com s'observa i quins efectes produeix aquesta interacció.

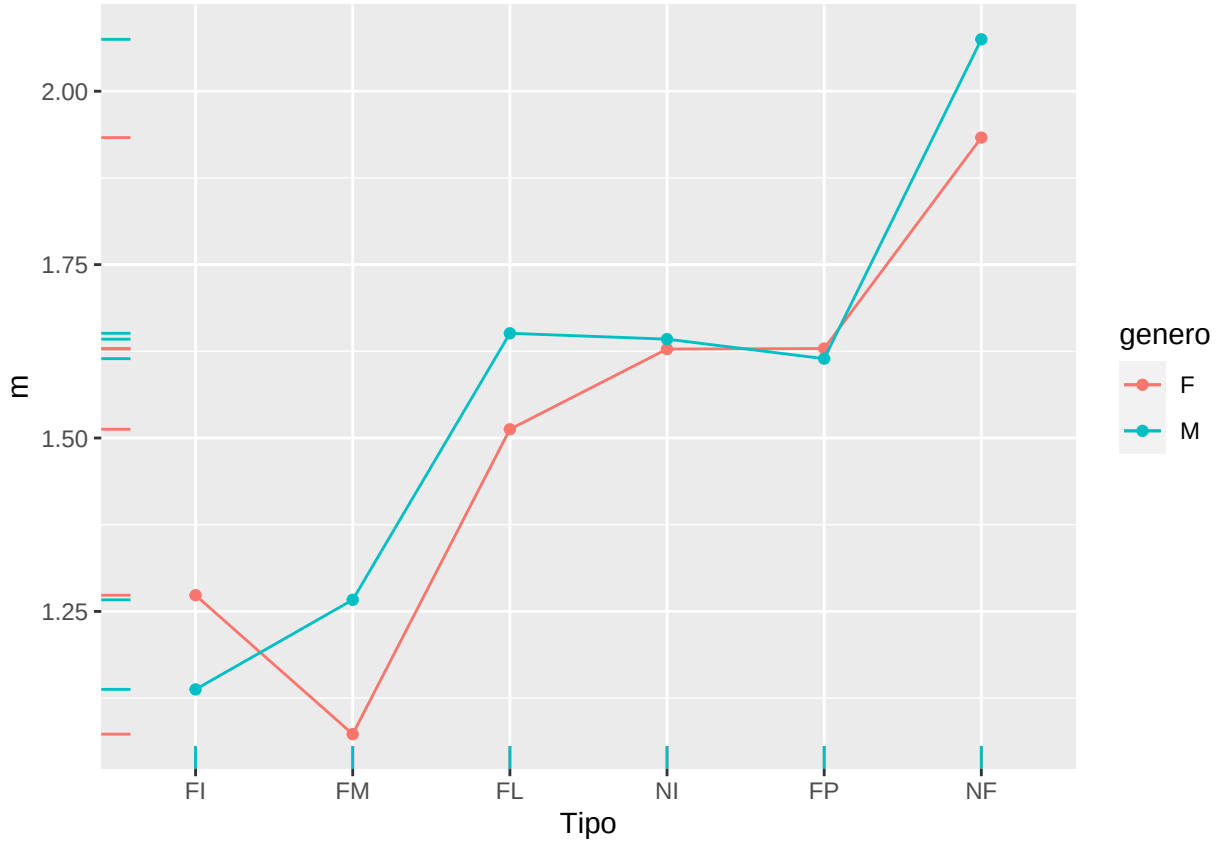
```
#data$Tipo <- factor( df$Tipo, levels=c("FI","FM","FL","NI","FP","NF"))

#S'agrupa el dataset per tipo i género i es calcula la mitjana per cada grup.
data %>% group_by(Tipo, genero) -> DS2
DS3 <- summarise( DS2, m=mean(AE))
```

```
## `summarise()` has grouped output by 'Tipo'. You can override using the
## `.groups` argument.
```

```
DS3
```

```
## # A tibble: 12 x 3
## # Groups:   Tipo [6]
##    Tipo  genero     m
##    <fct> <chr>  <dbl>
##  1 FI    F       1.27
##  2 FI    M       1.14
##  3 FM    F       1.07
##  4 FM    M       1.27
##  5 FL    F       1.51
##  6 FL    M       1.65
##  7 NI    F       1.63
##  8 NI    M       1.64
##  9 FP    F       1.63
## 10 FP    M       1.61
## 11 NF    F       1.93
## 12 NF    M       2.07
```

```
library(ggplot2)
ggplot(DS3, aes(x=Tipo, y=m, group=genero, color=genero)) +
  geom_point() + geom_line() + geom_rug()
```



Interpretació: Segons el gràfic mostrat, hi ha efectes principals de la variable Tipus i de la variable gènere. Pel que fa a la variable gènere, s'observa que la capacitat pulmonar de les dones és lleugerament inferior a la dels homes. La possible interacció entre Tipus i gènere no és gaire visible. Cal estudiar amb el càlcul d'anova.

9.2 ANOVA multifactorial

Calculeu ANOVA multifactorial per a avaluar si la variable dependent AE es pot explicar a partir de les variables independents gènere i tipus de fumador. Incloeu l'efecte de la interacció només si s'ha observat dita interacció a l'anàlisi visual de l'apartat anterior. Interpretar el resultat.

```
my.aov2 <- aov( AE~Tipo + genero + genero*Tipo, data )
my.aov2
```

```
## Call:
##    aov(formula = AE ~ Tipo + genero + genero * Tipo, data = data)
##
## Terms:
##                       Tipo   genero Tipo:genero Residuals
## Sum of Squares  20.85584  0.19699     0.76465  56.66844
## Deg. of Freedom        5        1           5       241
##
## Residual standard error: 0.4849111
## Estimated effects may be unbalanced
```

```
sum.aov2<-summary( my.aov2 ); sum.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
```

Interpretació: S'observa que la variabilitat d'AE s'explica fonamentalment pel tipus de fumador (p<2 10e-6). No s'observen efecte de la variable gènere ni tampoc no s'observa interacció significativa entre gènere i tipus de fumador.

10 Resum tècnic

Realitzeu una taula amb el resum tècnic de les preguntes de recerca plantejades al llarg d'aquesta activitat.

N	Pregunta	Resultat	Conclusió
P1	IC AE dones 95% IC AE homes 95%	1.5179115, 1.6496132 1.452326, 1.5942343	Els intervals es solapen. No existeixen diferències al 95% NC
P2	Contrast AE F vs M	t=0.8531624 p=0.3943827	No existeixen diferències en AE entre homes i dones al 95%
P3	Contrast AE Fum vs NoF	t=-6.3297609 p=$5.6134782 \times 10^{-10}$	Existeixen diferències en AE entre fumadors i no fumadors al 95%
P4	Anàlisi de regresió	R2=0.5830461	Variables independents significatives: edat, tipus de fumador
P5	ANOVA unifactorial	F=17.8774417 p=$4.0257861 \times 10^{-15}$	hi ha diferències significatives en AE segon tipus de fumador.
P5	ANOVA multifactorial	F=17.7391745(Tipo) p=5.809109×10^{-15} (Tipo)	Efecte principal de tipus de fumador Sense efecte en gènere ni interacció.

11 Resum executiu

Escriviu un resum executiu com si haguéssiu que comunicar a una audiència no tècnica. Per exemple, podria ser un equip de gestors o decisors, als quals se'ls ha d'informar sobre les conseqüències de fumar sobre la capacitat pulmonar, perquè puguin prendre les decisions necessàries.

S'ha realitzat un estudi de la capacitat pulmonar d'una població de fumadors en comparació amb no fumadors. La població de fumadors s'ha classificat en Fumador Intensiu, Moderat, Lleuger i No Inhala, segons els hàbits de consum de cigarrets i anys de fumador. La població de no fumadors s'ha categoritzat com a No fumador i Fumador passiu.

En general, s'ha observat que la capacitat pulmonar disminueix amb l'edat en tots els grups. En canvi, no s'observen diferències significatives en la capacitat pulmonar segons el gènere amb un nivell de confiança del 95%. S'observen diferències significatives molt notables en la capacitat pulmonar entre els diferents tipus de fumador. Concretament, fumador intensiu i moderat tenen capacitat pulmonar equivalent. El fumador lleuger té capacitat pulmonar equivalent al fumador que no inhala. I el fumador passiu té capacitat pulmonar equivalent a un fumador lleuger o que no inhala.

Així mateix, s'ha desenvolupat un model de predicció amb què podem fer una estimació de la capacitat pulmonar d'una persona a partir del tipus de fumador i edat. L'estimació és aproximada, ja que tan sols és capaç d'explicar el 58% de la variabilitat de la capacitat pulmonar a la població. No obstant això, es pot utilitzar com a model de simulació.

Puntuació dels apartats

- Pregunta 1: 10%

- Pregunta 2: 10%
- Pregunta 3: 10%
- Preguntes 4,5: 10%
- Pregunta 6: 10%
- Pregunta 7: 10%
- Pregunta 8: 10%
- Pregunta 9: 10%
- Pregunta 10: 10%
- Pregunta 11: 10%