Estadística Avanzada - Actividad 4

Propuesta de solución

Semestre 2022.1

Índice

Introducción

En una investigación médica se estudió la capacidad pulmonar de los fumadores y no fumadores. Se recogieron datos de una muestra de la población fumadora, no fumadora y fumadores pasivos. A cada persona se realizó un test de capacidad pulmonar consistente en evaluar la cantidad de aire expulsado (AE). La muestra de n individuos se categorizó en 6 tipos:

- No fumadores (NF)
- Fumadores pasivos (FP)
- Fumadores que no inhalan (NI): personas que fuman pero no inhalan el humo.
- Fumadores ligeros (FL): personas que fuman e inhalan de uno a 10 cigarrillos al día durante 20 años o más.
- Fumadores moderados (FM): personas que fuman e inhalan entre 11 y 39 cigarrillos por día durante 20 años o más.
- Fumadores intensivos (FI): personas que fuman e inhalan 40 cigarrillos o más durante 20 años o más.

En esta actividad se analizará si la capacidad pulmonar está influida por el tipo de fumador. Para ello, se aplicaran distintos tipos de análisis, revisando los contrastes de hipótesis de dos muestras, vistos en la actividad A2, y luego realizando análisis más complejos como ANOVA.

Notas importantes a tener en cuenta para la entrega de la actividad:

- Es necesario entregar el fichero Rmd y el fichero de salida (PDF o html). El fichero de salida debe incluir el código y el resultado de su ejecución (paso a paso). Se debe incluir un índice o tabla de contenidos. Y se debe respetar la numeración de los apartados del enunciado.
- No realizar listados de los conjuntos de datos, puesto que estos pueden ocupar varias páginas. Si queréis comprobar el efecto de una instrucción sobre un conjunto de datos podéis usar la función **head** o **tail** que muestran las primeras o últimas filas del conjunto de datos.

---

1 Preprocesado

Cargar el fichero de datos "Fumadores.csv". Consultar los tipos de datos de las variables y si es necesario, aplicar las transformaciones apropiadas. Averiguar posibles inconsistencias en los valores de Tipo, AE, género y edad. En caso de que existan inconsitencias, corregidlas.

```
data <- read.csv( "Fumadores.csv", 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 ...
```

```
#Revisamos 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"
```

```
#Revisamos género
unique(data$genero)
```

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

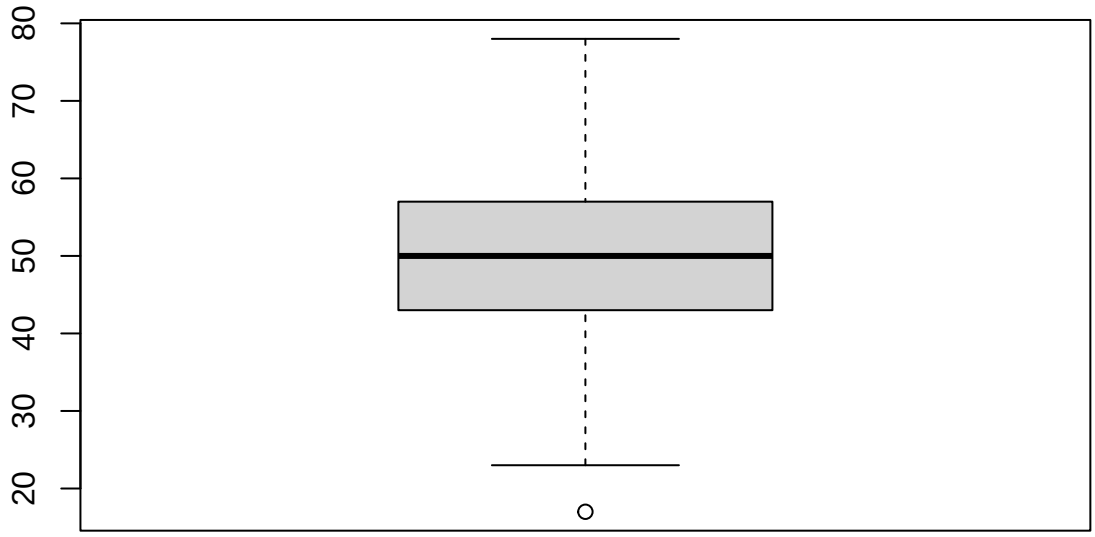
```
#Revisamos AE. Coma y punto decimal
data$AE[ grep(",", data$AE) ]
```

```
##  [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))

#Revisamos edad
class(data$edad)
```

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

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




```
#Valores extremos en campo edad
data$edad[data$edad<20]
```

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

```
#Posibles inconsistencias entre edad y tipo 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)
```

Preproceso realizado:

- Se ha normalizado el formato de número de AE, corrigiendo la coma decimal por el punto decimal.
- Se normaliza el formato del Tipo de fumador.

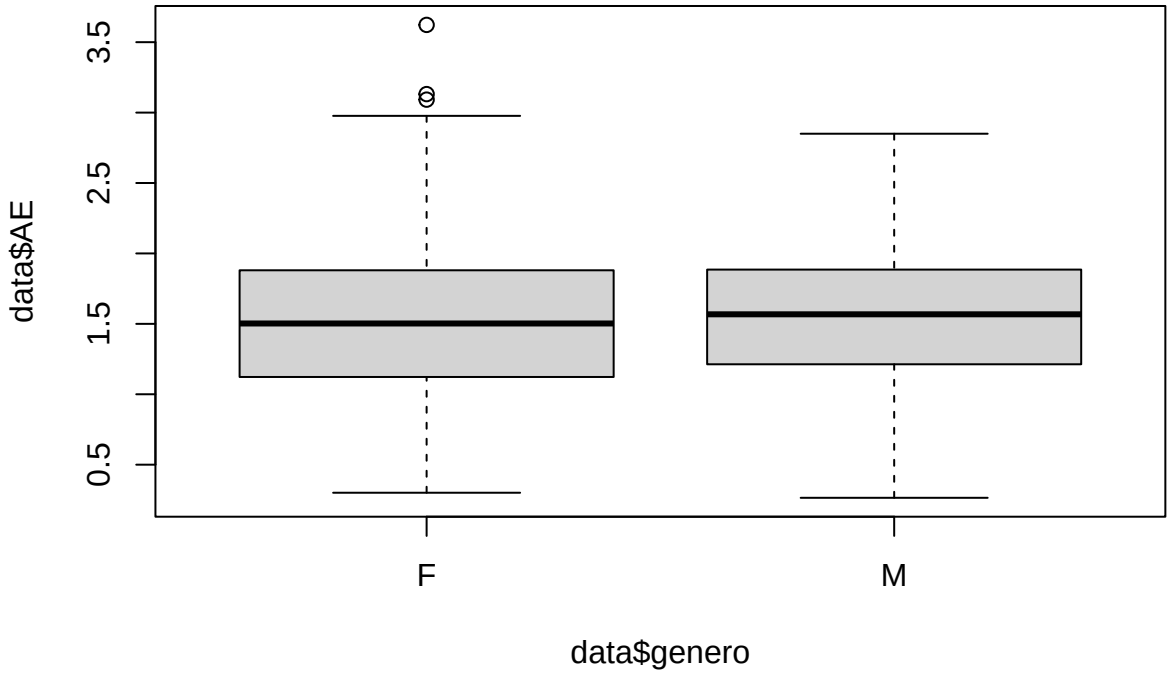
- Se encuentran incosistencias entre edad y Tipo de fumador (edades inferiores a 30 años con fumadores ligeros e intensivos que fuman más de 20 años). Por falta de información, estos registros se dejan intactos. Pero se podrían eliminar o realizar una imputación en el valor de edad o tipo de fumador.

2 Análisis descriptivo de la muestra

2.1 Capacidad pulmonar y género

Mostrar la capacidad pulmonar en relación al género. ¿Se observan diferencias?

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

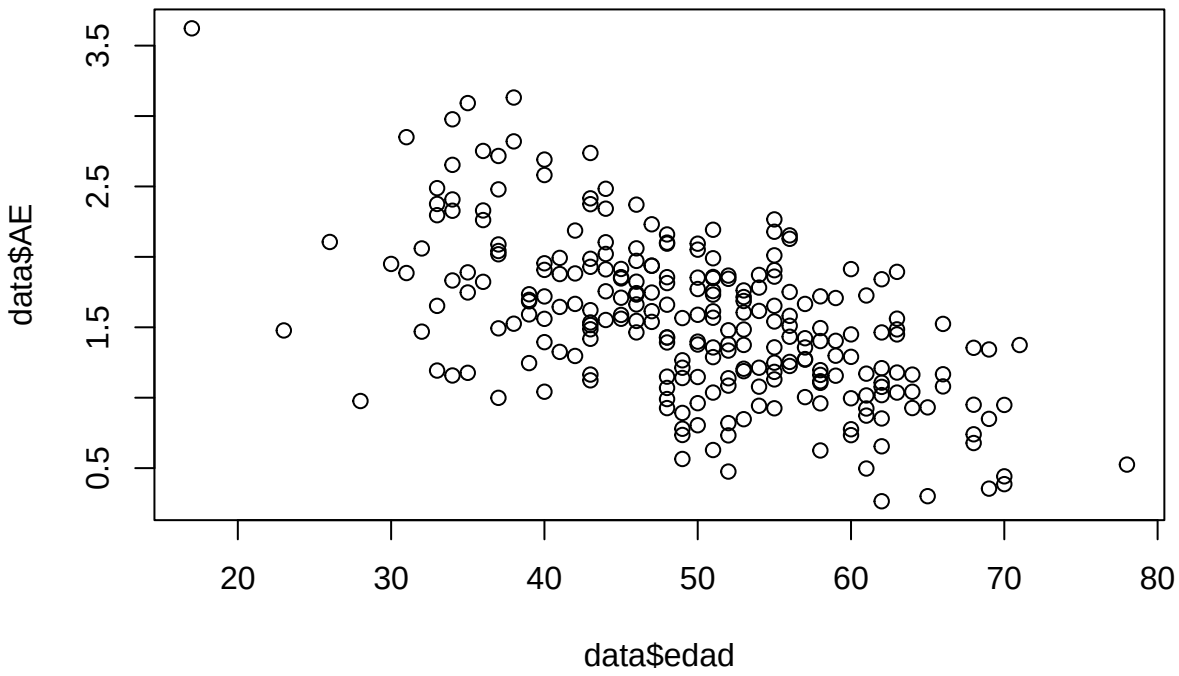


Prácticamente no se observan diferencias entre capacidad pulmonar y género.

2.2 Capacidad pulmonar y edad

Mostrar la relación entre capacidad pulmonar y edad usando un gráfico de dispersión. Interpretar.

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



Se observa una tendencia a la baja en la capacidad pulmonar a medida que aumenta la edad.

2.3 Tipos de fumadores y capacidad pulmonar

Mostrar el número de personas en cada tipo de fumador y la media de AE de cada tipo de fumador. Mostrad un gráfico que visualice esta media. Se recomienda que el gráfico esté ordenado de menos a más AE.

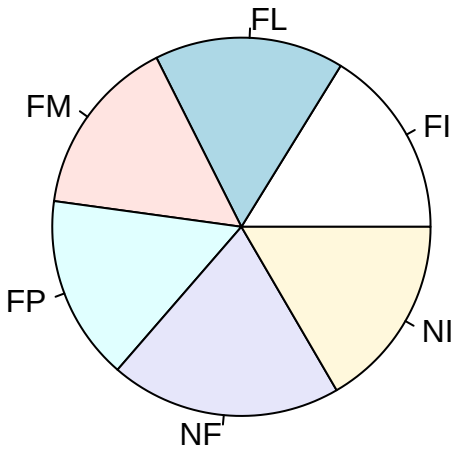
Luego, se debe representar un boxplot donde se muestre la distribución de AE por cada tipo de fumador. Interpretar los resultados.

Nota: Para calcular la media o otras variables para cada tipo de fumador, podéis usar las funciones **summarize** y **group_by** de la libreria **dplyr** que os serán de gran utilidad. Para realizar la visualización de los datos, podéis usar la función **ggplot** de la librería **ggplot2**.

```
#Número de personas por cada tipo de fumador
table( data$Tipo )
```

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

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




```
#Estadísticas de cada grupo
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
```

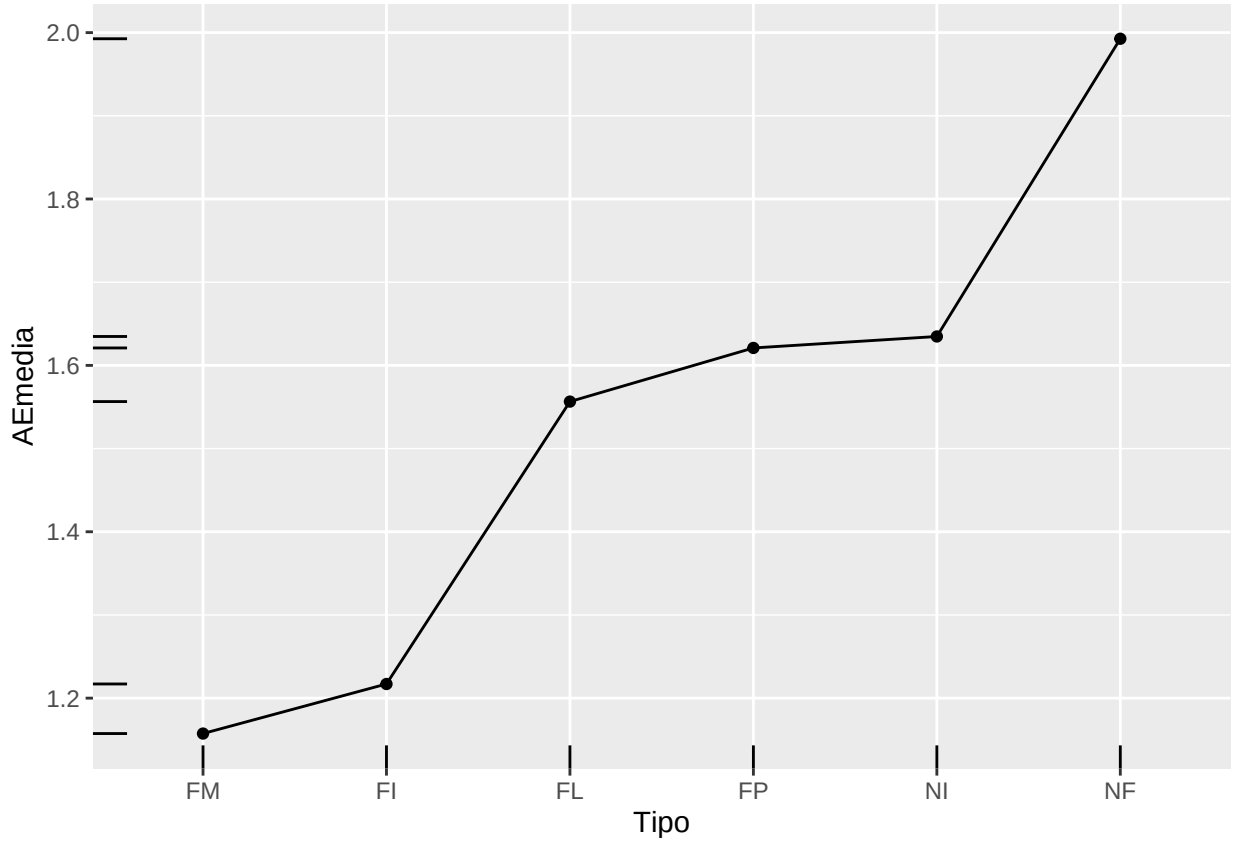
```
## # A tibble: 6 x 7
##   Tipo  AEmedia     n    sd edadmedia   fem  male
##   <fct>   <dbl> <int> <dbl>     <dbl> <int> <int>
## 1 FI       1.22    41 0.465      49.2    24    17
## 2 FL       1.56    41 0.484      49.2    28    13
## 3 FM       1.16    39 0.421      52.6    22    17
## 4 FP       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
```

```
#Preparamos el dataset para mostrar un gráfico de la media, ordenado según media.
DS$Tipo <- factor( DS$Tipo, levels=DS$Tipo[order(DS$AEmedia)])
library(ggplot2)
```

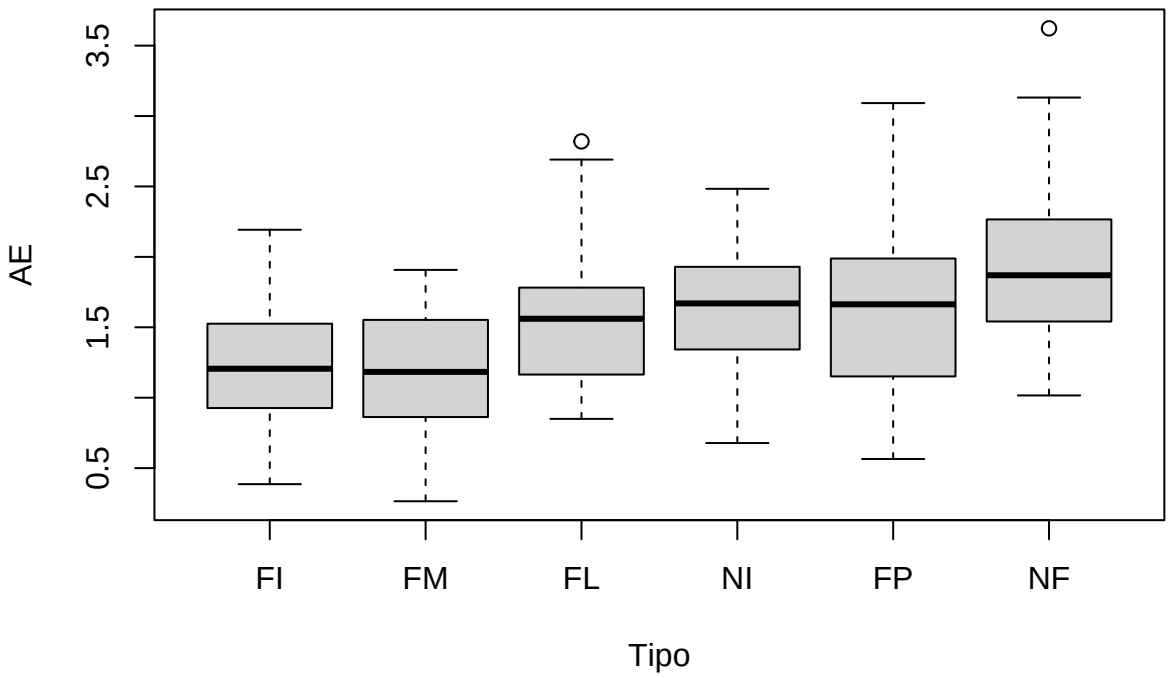
```
##
## Attaching package: 'ggplot2'
```

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

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




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



La muestra contiene aproximadamente el mismo número de personas por cada tipo de fumador. En cuanto a la capacidad pulmonar se observan diferencias entre los tipos de fumador, siendo el tipo "no fumador" el que tiene mayor capacidad pulmonar y los fumadores intensivos y moderados los que tienen menor capacidad pulmonar.

---

3 Intervalo de confianza de la capacidad pulmonar

Calcular el intervalo de confianza al 95% de la capacidad pulmonar de las mujeres y hombres por separado. Antes de aplicar el cálculo, revisar si se cumplen las asunciones de aplicación del intervalo de confianza. Interpretar los resultados. A partir de estos cálculos, ¿se observan diferencias significativas en la capacidad pulmonar de mujeres y hombres?

Nota: Realizar el cálculo manualmente sin usar las funciones `t.test` o equivalentes. Podéis usar `qnorm`, `qt`, `pnorm`, `pt`, ...

Respuesta: Como la muestra es superior a 30, podemos asumir que la media de AE sigue una distribución normal, según el teorema del límite 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
```

Como vemos, los intervalos de confianza están solapados y por tanto, no podemos afirmar que existan diferencias en la capacidad pulmonar entre hombres y mujeres, como ya habíamos observado visualmente en la sección anterior.

---

4 Diferencias en capacidad pulmonar entre mujeres y hombres

Aplicar un contraste de hipótesis para evaluar si existen diferencias significativas entre la capacidad pulmonar de mujeres y hombres. Seguid los pasos que se indican a continuación.

Nota: Realizar el cálculo manualmente sin usar las funciones `t.test` o equivalentes. Podéis usar `qnorm`, `qt`, `pnorm`, `pt`, ...

4.1 Hipótesis

Escribir la hipótesis nula y alternativa.

$H_0 : \mu_F = \mu_M$

$H_1 : \mu_F \neq \mu_M$

4.2 Contraste

Explicad qué tipo de contraste aplicaréis y por qué. Si es necesario, validad las asunciones del test.

Respuesta: aplicamos un contraste de dos muestras independientes sobre la media. Comprobamos si podemos asumir homocedasticidad.

```
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 resultado del test no muestra diferencias significativas entre varianzas. Por tanto, aplicaremos un test de dos muestras independientes sobre la media con varianzas desconocidas iguales. El test es bilateral.

4.3 Cálculos

Aplicad los cálculos del contraste. Mostrar el valor observado, el valor de contraste y 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)

  #varianzas iguales
    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 )
  }

  #Guardamos el resultado 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
```

```
#Comprobación:
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 Interpretación

Interpretad los resultados y comparad las conclusiones con los intervalos de confianza calculados anteriormente.

Respuesta: No podemos rechazar la hipótesis nula. Por tanto, no existen diferencias significativas en la capacidad pulmonar entre hombres y mujeres con un nivel de confianza del 95%. Esta conclusión es consistente con el cálculo de los intervalos de confianza realizado anteriormente.

---

5 Diferencias en la capacidad pulmonar entre Fumadores y No Fumadores

¿Podemos afirmar que la capacidad pulmonar de los fumadores es inferior a la de no fumadores? Incluid dentro de la categoría de no fumadores los fumadores pasivos. Seguid los pasos que se indican a continuación.

Nota: Realizar el cálculo manualmente sin usar las funciones `t.test` o equivalentes. Podéis usar `qnorm`, `qt`, `pnorm`, `pt`, ...

5.1 Hipótesis

Escribir la hipótesis nula y alternativa.

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

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

5.2 Contraste

Explicad qué tipo de contraste aplicaréis y por qué. Si es necesario, validad las asunciones del test.

Respuesta: aplicamos un contraste de dos muestras independientes sobre la media. Comprobamos si podemos asumir homocedasticidad.

```
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 resultado del test no muestra diferencias significativas entre varianzas.

Respuesta: Contraste de medias de dos muestras independientes. Asumimos distribución normal y caso de varianza poblacional desconocidas iguales. Aplicamos un contraste unilateral.

5.3 Preparar los datos para realizar el contraste

```
#Se han creado anteriormente las muestras Fum y NFum
n1 <- nrow( Fum )
n2 <- nrow( NFum )

n1; n2
```

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

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

5.4 Cálculos

Aplicad los cálculos del contraste. Mostrar el valor observado, el valor de contraste y el valor p.

```
#Usamos la función 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
```

```
#Comprobación  con 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 Interpretar el resultado del contraste

Dado que el valor p es menor que $\alpha = 0.05$ rechazamos la hipótesis nula a favor de la hipótesis alternativa, según la cual la media de la capacidad pulmonar de los fumadores es inferior a la de los no fumadores con un nivel de confianza del 95%.

---

6 Análisis de regresión lineal

Realizamos un análisis de regresión lineal para investigar la relación entre la variable capacidad pulmonar (AE) y el resto de variables (tipo, edad y género). Construid e interpretad el modelo.

6.1 Cálculo

```
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 Interpretación

Interpretar el modelo y la contribución de cada variable explicativa sobre la variable AE.

Respuesta: La variable edad influye en la capacidad pulmonar con un coeficiente negativo, es decir, que a medida que aumenta la edad disminuye la capacidad pulmonar. La variable Tipo de Fumador es significativa. La categoría de referencia es FI (intensivo). Existen diferencias significativas en AE entre todos los tipos de fumadores excepto el moderado, en relación al fumador intensivo. Finalmente, el género no influye en la capacidad pulmonar. Los resultados son consistentes con los contrastes realizados anteriormente sobre género, edad y tipo de fumador, y el análisis visual que se ha mostrado en relación a los tipos de fumador.

6.3 Bondad de ajuste

Evaluar la calidad del modelo.

Respuesta: El modelo explica el 58.3% de la variabilidad en la capacidad pulmonar. Probablemte hay otras variables que influyen en la capacidad pulmonar y que no están incluidas en el modelo, como la realización de ejercicio físico o si la persona vive en un entorno con alta contaminación.

6.4 Predicción

Realizad una predicción de la capacidad pulmonar para cada tipo de fumador desde los 30 años de edad hasta los 80 años de edad (podéis asumir género hombre). Mostrad una tabla con los resultados. Mostrad también visualmente la simulación.

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

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

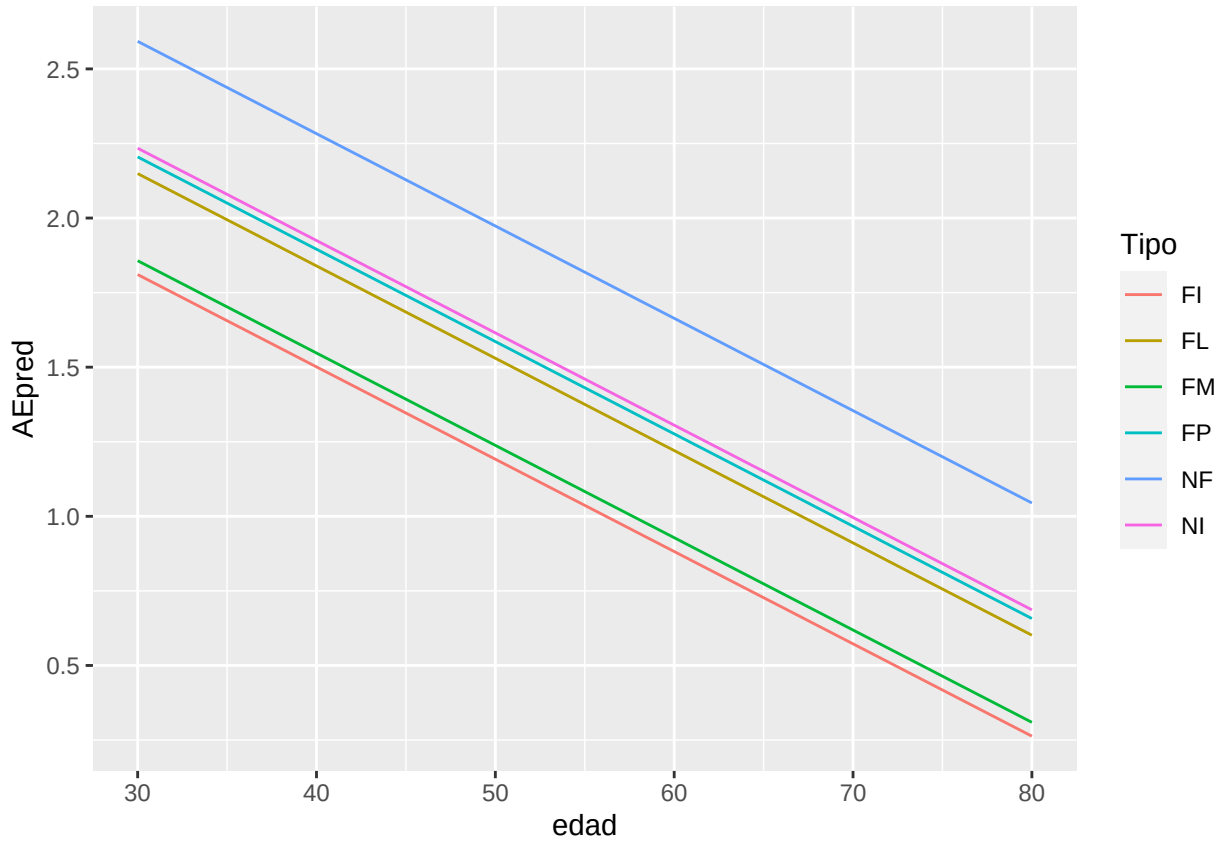
```
  tipo <- c("NF", "FP", "NI", "FL", "FM", "FI")
  rango.tipo<-sort( rep( tipo, N) )

  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

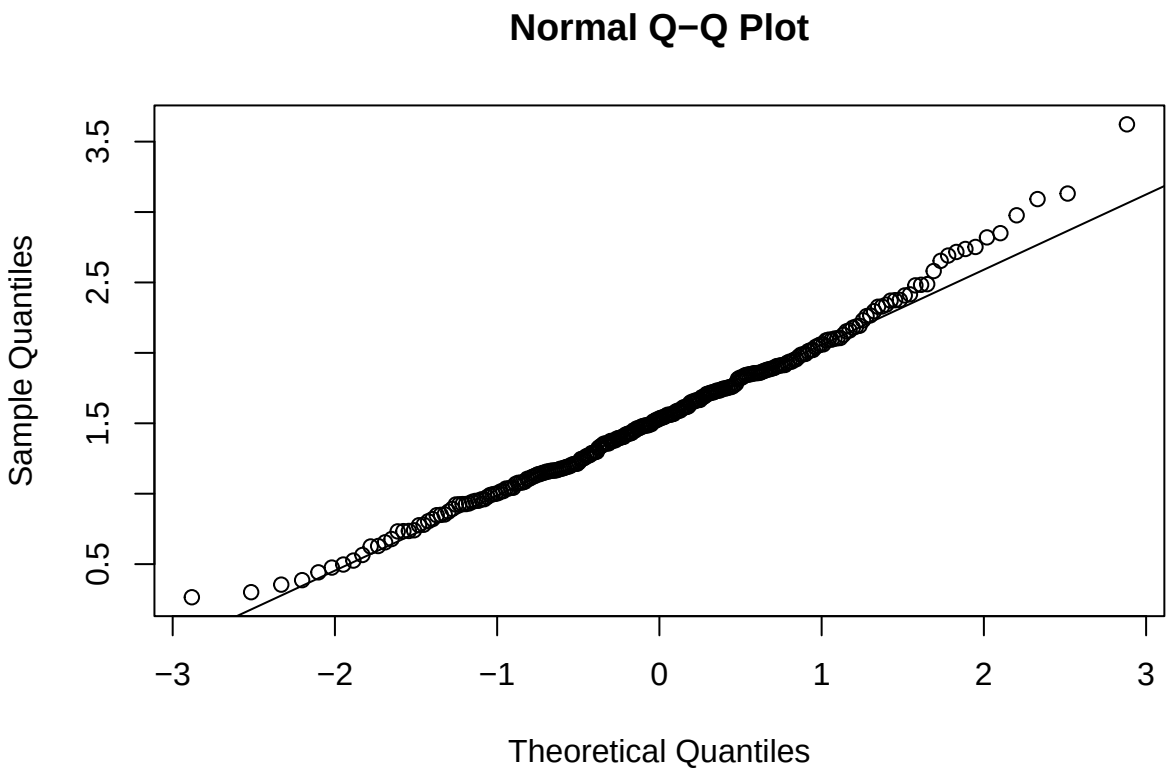
A continuación se realizará un análisis de varianza, donde se desea comparar la capacidad pulmonar entre los seis tipos de fumadores/no fumadores clasificados previamente. El análisis de varianza consiste en evaluar si la variabilidad de una variable dependiente puede explicarse a partir de una o varias variables independientes, denominadas factores. En el caso que nos ocupa, nos interesa evaluar si la variabilidad de la variable AE puede explicarse por el factor tipo de fumador. Hay dos preguntas básicas a responder:

- ¿Existen diferencias entre la capacidad pulmonar (AE) entre los distintos tipos de fumadores/no fumadores?
- Si existen diferencias, ¿entre qué grupos están estas diferencias?

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.

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




```
#H0: la muestra (de tamaño n) sigue una distribución normal
#Se rechaza  H0 si p value < alfa

#Si se aplica Shapiro (en toda la muestra)
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
```

Interpretación: En el test de Shapiro-Wilk, si $Pr(D) \leq \alpha$ se rechazaría la hipótesis nula de normalidad en los datos.

El valor p del test de Shapiro ha dado 0.0448371. Por tanto, se rechazaría la hipotesis nula de normalidad, aunque esta desviación respecto la normalidad no es muy pronunciada.

La condición de normalidad se debe cumplir para cada grupo. Por ello, se debe aplicar la prueba de normalidad a cada grupo (tipo de fumador). También se valora que se represente el plot para cada tipo de fumador.

```r
#"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 arroja en todos los grupos (excepto uno) valores p superiores a 0.05. Estrictamente, no se podría rechazar la hipótesis nula de normalidad, aunque como la desviación es poco pronunciada, seguimos con la aplicación de ANOVA paramétrico.

7.2 Homocedasticidad: Homogeneidad de varianzas

Otra de las condiciones de aplicación de ANOVA es la igualdad de varianzas (homocedasticidad). Aplicar un test para validar si los grupos presentan igual varianza. Aplicad el test adecuado e interpretar el resultado.

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

H_1 : Al menos existen diferencias entre dos grupos: $\sigma_i^2 \neq \sigma_j^2$

```r
#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
```

```r
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
```

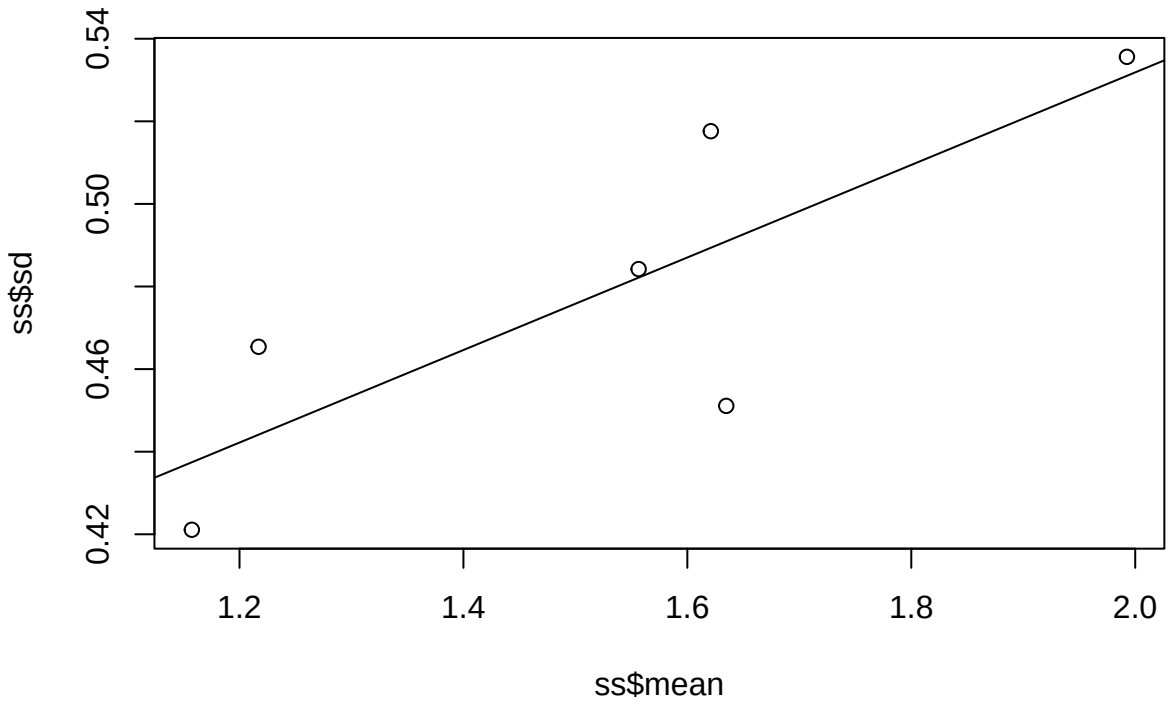
```r
LT$`F value`[1]
```

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

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

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

```
#Gráfico de dispersión por tipo. y=dispersión, x=media del grupo
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
```

Interpretación del test de Levene: El valor del test de Levene es $Pr(F)$=0.8316893. Para un nivel de significación $\alpha = 0,05$, $Pr(F) \geq \alpha$. Por tanto, no se rechaza la hipótesis nula de igualdad de varianzas. Se cumple la condición de homocedasticidad. Otros tests de homogeneidad de varianzas como el test Barlett o Fligner-Killeen obtienen resultados análogos.

7.3 Hipótesis nula y alternativa

Independientemente de los resultados sobre la normalidad e homocedasticidad de los datos, proseguiremos con la aplicación del análisis de varianza. Concretamente, se aplicará ANOVA de un factor (one-way ANOVA o independent samples ANOVA) para investigar si existen diferencias en el nivel de aire expulsado (AE) entre los distintos tipos de fumadores. Escribid la hipótesis nula y alternativa.

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

H_1 : Al menos dos grupos son distintos: $\mu_i \neq \mu_j$

7.4 Cálculo ANOVA

Podéis usar la función **aov**.

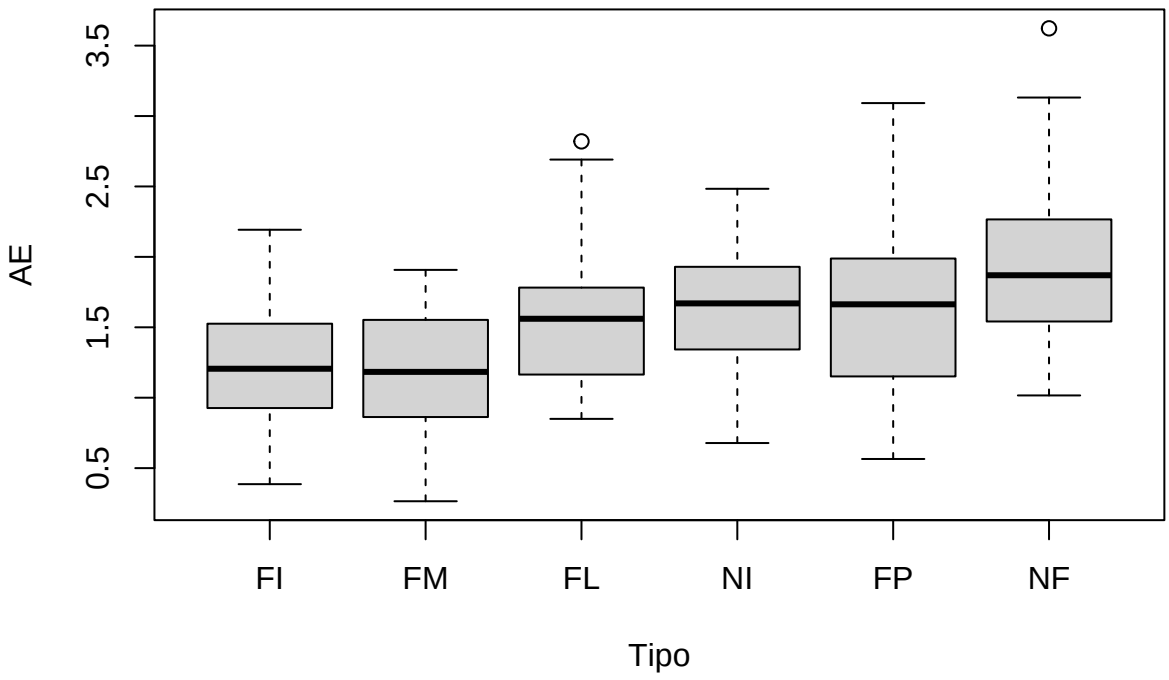
```
#Cálculo one-way ANOVA
my.aov <- aov( AE~Tipo, data )
sum.aov <- summary( my.aov)
sum.aov
```

```
##              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 Interpretación

Interpretar los resultados de la prueba ANOVA y relacionarlos con el resultado gráfico del boxplot mostrado en el apartado 2.3.

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



El estadístico F=17.8774417.

El valor p es $4.0257861 \times 10^{-15}$. Por tanto, podemos rechazar la hipótesis nula de que las diferencias entre los grupos sean iguales. Este resultado se observa visualmente en el diagrama de cajas (boxplot), donde se observan diferencias entre las medias de los grupos.

7.6 Profundizando en ANOVA

A partir de los resultados del modelo devuelto por **aov**, identificar las variables SST (Total Sum of Squares), SSW (Within Sum of Squares), SSB (Between Sum of Squares) y los grados de libertad. A partir de estos valores, calcular manualmente el valor F, el valor crítico (a un nivel de confianza del 95%), y el valor p. Interpretar los resultados y explicar el significado de las variables SST, SSW y SSB.

```
#Cálculos
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
```

Como se puede observar, el cálculo de F se realiza a partir de la varianza entre grupos que es SSB/(k-1), donde SSB=20.855837 y (k-1)=5. El denominador es la varianza dentro de los grupos y corresponde a SSW/(n-k), donde SSW=57.6300772, y (n-k)=247. El cómputo de F da 17.8774417, el cual coincide con el resultado del modelo anova calculado. El cálculo del valor p se ha realizado con la función `pf` a partir del estadístico F y los grados de libertad (k-1) y (n-k) respectivamente.

7.7 Fuerza de la relación

Calcular la fuerza de la relación e interpretar el resultado.

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

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

Interpretación: La fuerza de la relación representa en qué medida el conocimiento del grupo de pertenencia determina el valor en la variable dependiente. Según el resultado los grupos al que pertenece una persona explica el 26.5727133 % de la variabilidad en la capacidad pulmonar.

8 Comparaciones múltiples

Independientemente del resultado obtenido en el apartado anterior, realizamos un test de comparación múltiple entre los grupos. Este test se aplica cuando el test ANOVA devuelve rechazar la hipótesis nula de igualdad de medias. Por tanto, procederemos como si el test ANOVA hubiera dado como resultado el rechazo de la hipótesis nula.

8.1 Test pairwise

Calcular las comparaciones entre grupos sin ningún tipo de corrección. Podéis usar la función **pairwise.t.test**. Interpretar los resultados.

```
  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
```

Interpretación:

- NF (no fumador): presenta diferencias significativas con todos los grupos.
- FP (fumador pasivo): Tiene capacidad pulmonar significativamente distinta del No fumador, Fumador intensivo y Fumador moderado. Equivalente a FL (Fumador ligero) y a NI (no inhala).
- NI (fumador no inhala): diferencias significativas con NF (no fumador), FI (fumador intensivo) y FM (fumador moderado). Equivalente a FP (pasivo), FL (ligero).
- FL (fumador ligero): tiene AE significativamente diferente del FI (intensivo) y FM (moderado). Equivalente a NI (no inhala) y FP (pasivo). También es significativamente diferente del NF (no fumador).
- FI (intensivo) y FM (moderado) son equivalentes entre si.

8.2 Corrección de Bonferroni

Aplicar la corrección de Bonferroni en la comparación múltiple. Interpretar el resultado y contrastar el resultado con el obtenido en el test de comparaciones múltiples sin corrección.

```
  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")
```

Interpretación: Hay diferencias significativas entre:

- NF (no fumador) y el resto de tipos.
- FI (intensivo) presenta diferencias con todos los tipos, excepto el FM (moderado).
- FM (moderado) presenta diferencias con todos los tipos, excepto el FI (intensivo).
- FL (ligero) tiene diferencias con FP (pasivo) y a la vez con FM (moderado) e FI (intensivo).
- NI presenta diferencias con FI (intensivo) y FM (moderado), además de las diferencias con NF.
- FP (pasivo) solo presenta diferencias con FI (intensivo) y FM (moderado), además de la diferencia con NF.

Se puede ver que detecta menos diferencias. Es un test más conservador.

9 ANOVA multifactorial

En una segunda fase de la investigación se evalua el efecto del género como variable independiente, además del efecto del tipo, sobre la variable AE.

9.1 Análisis visual

Se realizará un primer estudio visual para determinar si existen efectos principales o hay efectos de interacción entre género y tipo de fumador. Para ello, seguir los pasos que se indican a continuación:

1. Agrupar el conjunto de datos por tipo de fumador y género y calcular la media de AE en cada grupo. Podéis usar las instrucciones **group_by** y **summarise** de la librería **dplyr** para realizar este proceso. Mostrar el conjunto de datos en forma de tabla, donde se muestre la media de cada grupo según el género y tipo de fumador.
2. Mostrar en un plot el valor de AE medio para cada tipo de fumador y género. Podéis realizar este tipo de gráfico usando la función **ggplot** de la librería **ggplot2**.
3. Interpretar el resultado sobre si existen sólo efectos principales o existe interacción. Si existe interacción, explicar cómo se observa y qué efectos produce esta interacción.

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

  #Se agrupa el dataset por tipo y género y se calcula la media para cada grupo.
  data %>% group_by(Tipo, genero) -> DS2
  DS3 <- summarise( DS2, m=mean(AE))
```

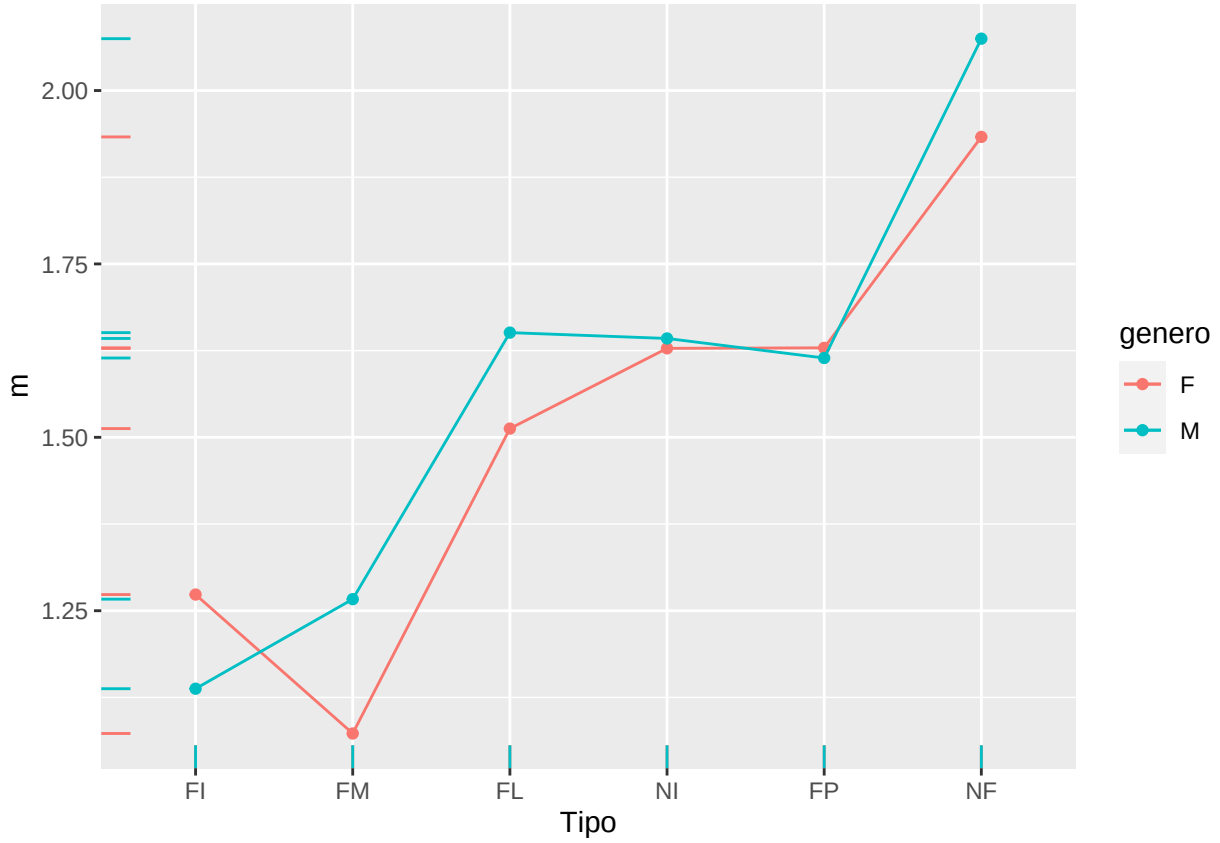
```
## `summarise()` has grouped output by 'Tipo'. You can override using the
## `.groups` argument.
```

```
  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()
```



Interpretación: Según el gráfico mostrado, hay efectos principales de la variable Tipo y de la variable género. En relación a la variable género, se observa que la capacidad pulmonar de las mujeres es ligeramente inferior a la de los hombres. La posible interacción entre Tipo y género no es muy visible. Se deberá estudiar con el cálculo de anova.

9.2 ANOVA multifactorial

Calcular ANOVA multifactorial para evaluar si la variable dependiente AE se puede explicar a partir de las variables independientes género y tipo de fumador. Incluid el efecto de la interacción sólo si se ha observado dicha interacción en el análisis visual del apartado anterior. Interpretad el resultado.

```
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
```

Interpertación: Se observa que la variabilidad de AE se explica fundamentalmente por el tipo de fumador (p<2 10e-6). No se observan efecto de la variable género ni tampoco se observa interacción significativa entre género y tipo de fumador.

10 Resumen técnico

Realizad una tabla con el resumen técnico de las preguntas de investigación planteadas a lo largo de esta actividad.

N	Pregunta	Resultado	Conclusión
P1	IC AE mujeres 95%	1.5179115, 1.6496132	Los intervalos se solapan.
	IC AE hombres 95%	1.452326, 1.5942343	No existen diferencias al 95% NC
P2	Contraste	t=0.8531624	No existen diferencias en AE
	AE F vs M	p=0.3943827	entre hombres y mujeres al 95%
P3	Contraste	t=-6.3297609	Existen diferencias en AE
	AE Fum vs NoF	p=$5.6134782 \times 10^{-10}$	entre fumadores y no fumadores al 95%
P4	Análisis de regresión	R2=0.5830461	Variables independientes significativas:
			edad, tipo de fumador
P5	ANOVA unifactorial	F=17.8774417	Hay diferencias significativas
		p=$4.0257861 \times 10^{-15}$	en AE según tipo de fumador.
P5	ANOVA multifactorial	F=17.7391745(Tipo)	Efecto principal de tipo de fumador
		p=5.809109×10^{-15} (Tipo)	Sin efecto en género ni interacción.

11 Resumen ejecutivo

Escribid un resumen ejecutivo como si tuvieráis que comunicar a una audiencia no técnica. Por ejemplo, podría ser un equipo de gestores o decisores, a los cuales se les debe informar sobre las consecuencias de fumar sobre la capacidad pulmonar, para que puedan tomar las decisiones necesarias.

Se ha realizado un estudio de la capacidad pulmonar de una población de fumadores en comparación con no fumadores. La población de fumadores se ha clasificado en Fumador Intensivo, Moderado, Ligero y No Inhala, según los hábitos de consumo de cigarrillos y años de fumador. La población de no fumadores se ha categorizado como No fumador y Fumador pasivo.

En general se ha observado que la capacidad pulmonar disminuye con la edad en todos los grupos. En cambio, no se observan diferencias significativas en la capacidad pulmonar según el género con un nivel de confianza

del 95%. Se observan diferencias significativas muy notables en la capacidad pulmonar entre los distintos tipos de fumador. Concretamente, fumador intensivo y moderado tienen capacidad pulmonar equivalente. El fumador ligero tiene capacidad pulmonar equivalente al fumador que no inhala. Y el fumador pasivo tiene capacidad pulmonar equivalente a un fumador ligero o que no inhala.

Asismismo, se ha desarrollado un modelo de predicción con el que podemos realizar una estimación de la capacidad pulmonar de una persona a partir del tipo de fumador y edad. La estimación es aproximada, puesto que tan solo es capaz de explicar el 58% de la variabilidad de la capacidad pulmonar en la población. Sin embargo puede usarse como modelo de simulación.

12 Puntuación de la actividad

- Pregunta 1: 10%
- Pregunta 2: 10%
- Pregunta 3: 10%
- Preguntas 4,5: 10%
- Pregunta 6: 10%
- Pregunta 7: 10%
- Pregunta 8: 10%
- Pregunta 9: 10%
- Pregunta 10: 10%
- Pregunta 11: 10%