Actividad 4: Análisis de varianza y repaso del curso

Lukaz Martin Doehne

2023-01-23

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.

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=";", stringsAsFactors=TRUE, fileEncoding = "UTF-8")
str(data)
```

```
## 'data.frame':    253 obs. of  4 variables:
##  $ AE    : Factor w/ 253 levels "0,44163","0,942632",..: 190 198 242 224 160 143 122 180 220 217 ...
##  $ Tipo  : Factor w/ 10 levels "  FM  ","fi",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ genero: Factor w/ 2 levels "F","M": 2 1 2 1 1 1 2 2 2 1 ...
##  $ edad  : int  54 60 40 55 59 63 62 62 26 48 ...
```

Reemplazamos comas(,) por puntos(.) en la variable *AE*

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

Convertimos variable AE en númerica

```
data$AE<-as.numeric(data$AE)
```

Unificamos valores de variable *Tipo* a FM, FI, FL, FP, NF o NI

```
data$Tipo<-gsub(" ", "", data$Tipo) # Quitamos espacios en blanco
data$Tipo<-as.factor(toupper(data$Tipo)) # Convertimos todo en mayúsculas y factorizamos
```

Convertimos la variable *edad* a númerica

```
data$edad<-as.numeric(data$edad)
```

No se encuentran valores nulos en el dataset

```
sum(is.na(data))
```

```
## [1] 0
```

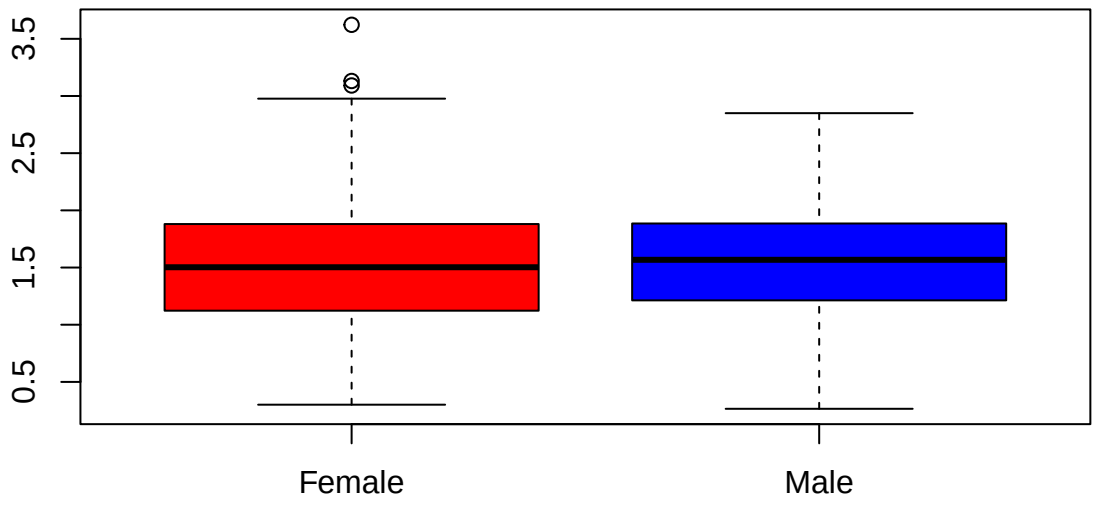
2 Análisis descriptivo de la muestra

2.1 Capacidad pulmonar y género

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

Para diferenciar visualmente he optado por un boxplot

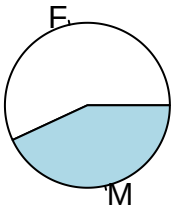
```
boxplot(data$AE[data$genero=="F"],data$AE[data$genero=="M"], names =c("Female","Male"), col=c("red","blu
```



A simple vista parece que el genero no influye significativamente en la capacidad pulmonar. Tenemos una media de 1.52 para mujeres y 1.58 para hombres (AE).Por lo que el de la mujer es levemente inferior. Además, podemos observar que para las mujeres hay más outliers.

También observamos que el dataset está medianamente equilibrado. Por lo que damos la interpretación cómo válida.

```
pie(table(data$genero))
```

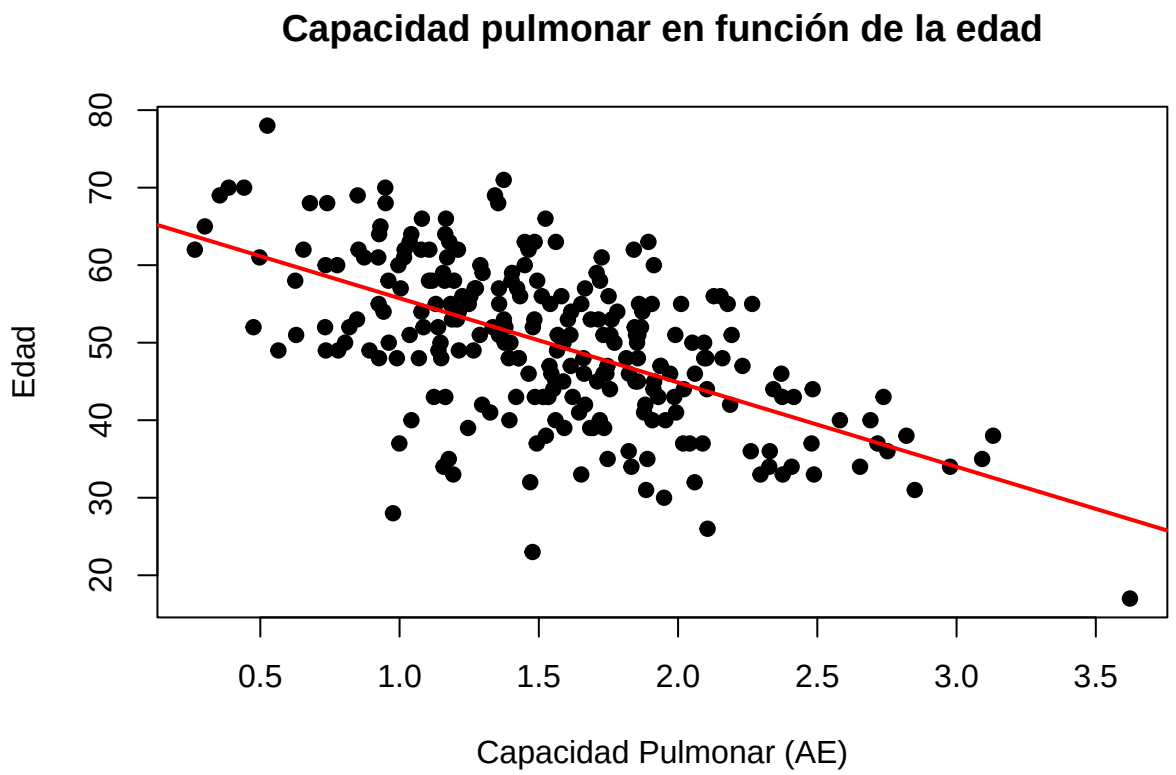


- 57% Mujeres
- 43% Hombres

2.2 Capacidad pulmonar y edad

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

```
attach(data)
plot(AE, edad, main="Capacidad pulmonar en función de la edad",
     xlab="Capacidad Pulmonar (AE)", ylab="Edad", pch=19)
abline(lm(edad~AE), col="red", lwd=2)
```



Cómo podemos observar en la gráfica los puntos se encuentran repartidos pero con una clara tendencia que visualizamos con la línea roja: a mayor edad, menor capacidad pulmonar.

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.

Número de personas en cada tipo de fumador

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

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

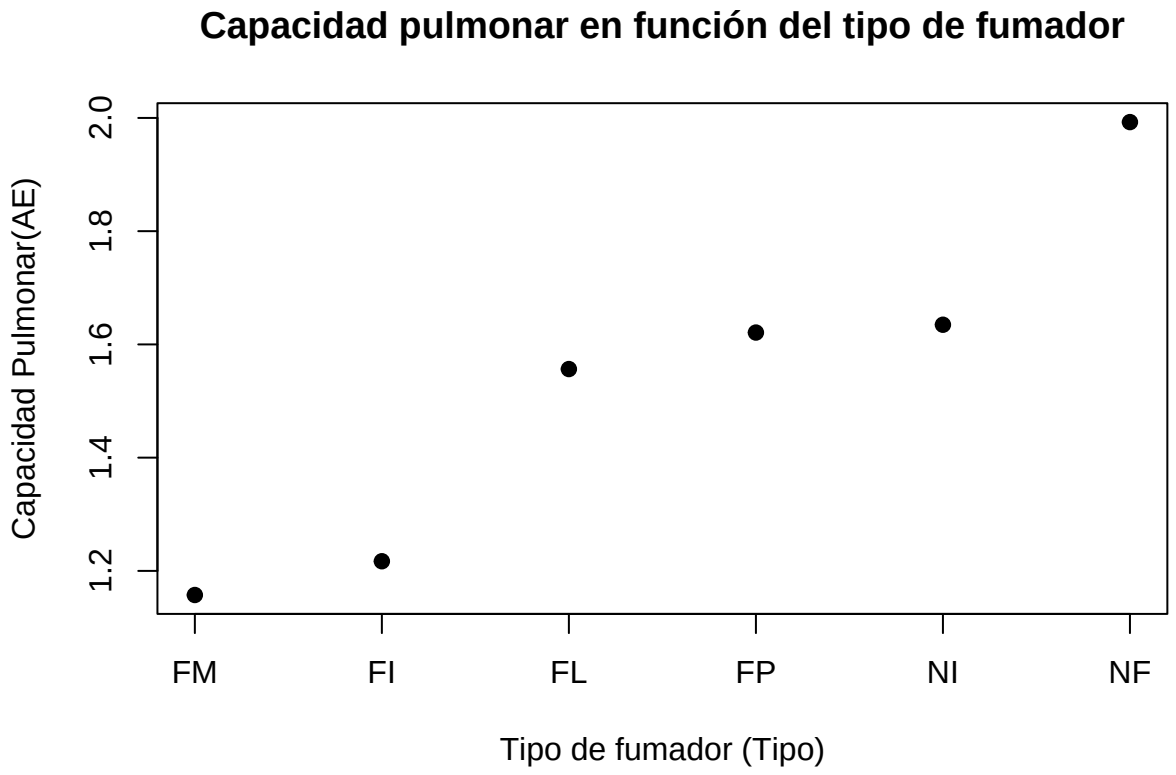
Media AE de cada tipo

```
mean_AE<-sapply(names(table(data$Tipo)), function(x) mean(data$AE[data$Tipo==x]))
mean_AE
```

```
##       FI       FL       FM       FP       NF       NI
## 1.217035 1.556476 1.157442 1.620952 1.992625 1.634737
```

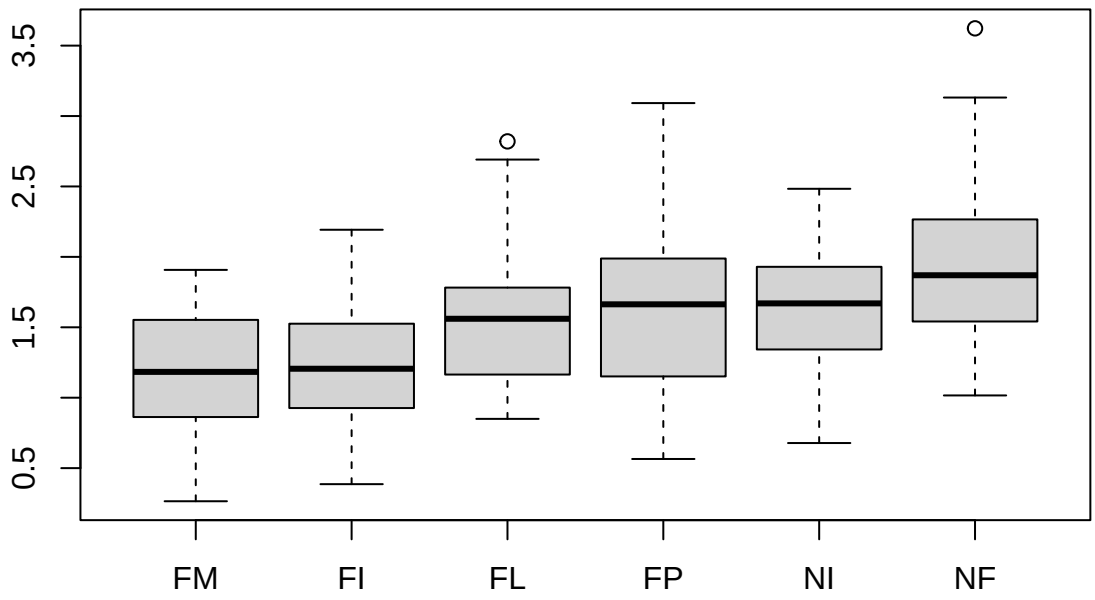
Gráfico para visualizar media

```
plot(c(1,2,3,4,5,6),as.numeric(mean_AE[c(3,1,2,4,6,5)]), main="Capacidad pulmonar en función del tipo d
     xlab="Tipo de fumador (Tipo)", ylab="Capacidad Pulmonar(AE)", xaxt="n",pch=19)
axis(1, at=1:6, labels=names(mean_AE[c(3,1,2,4,6,5)]))
```



Boxplot distribución *AE* según tipo de fumador

```
boxplot(data$AE[data$Tipo=="FM"],data$AE[data$Tipo=="FI"],data$AE[data$Tipo=="FL"],
data$AE[data$Tipo=="FP"],data$AE[data$Tipo=="NI"],data$AE[data$Tipo=="NF"],
names =c("FM","FI","FL","FP","NI","NF"))
```



Observamos que la cantidad de registros según Tipo es bastante equilibrada. Esto indica que a la hora de recoger los datos probablemente se busco hacer las pruebas a diferentes Tipos, ya que si las muestras fuesen aleatorias el porcentaje de No Fumadores sería mayor. Pero esto nos es útil para validar las conclusiones que podamos extraer.

Además podemos ver que la media de capacidad pulmonar disminuye cuánto más fumes (por norma general).

- Fumadores moderados (FM): 1.16
- Fumadores intensivos (FI): 1.22
- Fumadores ligeros (FL): 1.56
- Fumadores pasivos (FP): 1.62
- Fumadores que no inhalan (NI): 1.63
- No fumadores (NF): 1.99

También podemos observar un outlier en los No Fumadores. Con una capacidad pulmonar de 3.62 y de 17 años, probablemente se trate de un atleta.

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.

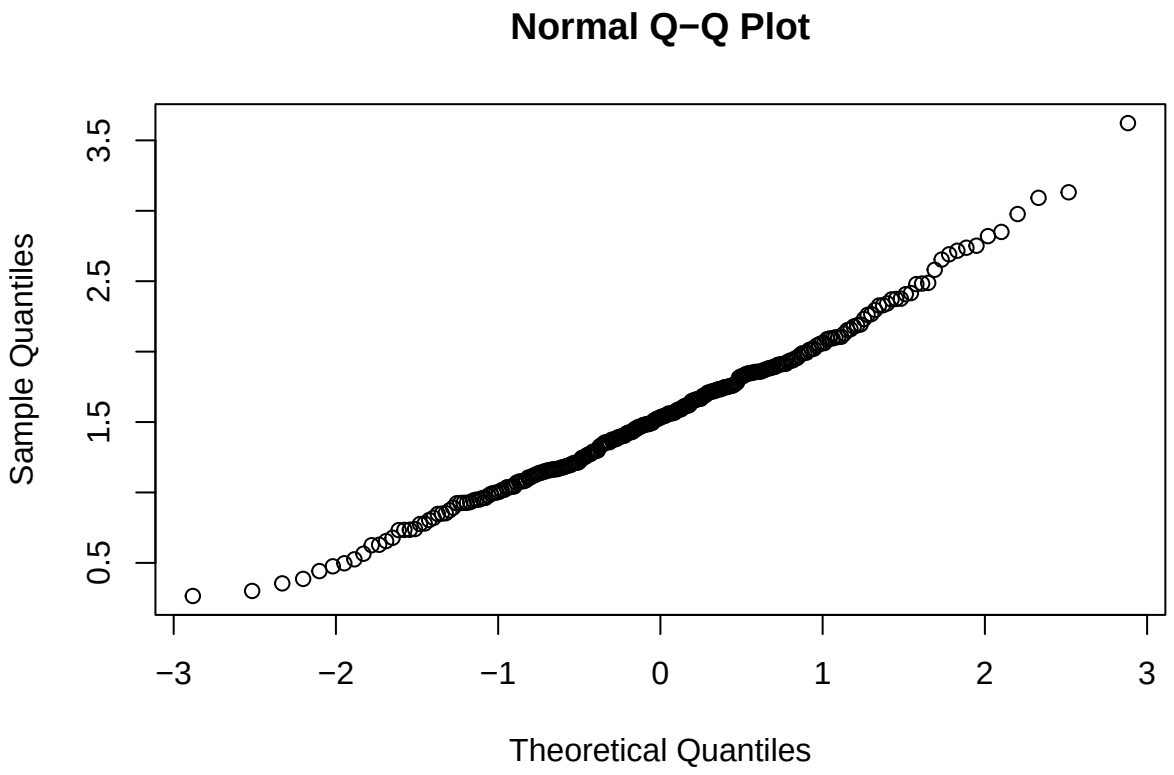
Para que la muestra sea fiable ha de ser representativa. (Sin sesgos y de un tamaño adecuado) Para averiguar si la muestra de la población es normal haremos uso de el **teorema del límite central** (TLC)

Este teorema establece que el contraste de hipótesis sobre la media de una muestra se aproxima a una distribución normal aunque la población original no siga una distribución normal, siempre que el tamaño de la muestra n sea suficientemente grande. Por suficientemente grande, se suele considerar superior a 30 elementos, n $>$ 30.

```
nrow(data)
```

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

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



Por el teorema del límite central, podemos asumir normalidad, puesto que tenemos una muestra de tamaño grande N=253 y en el plot parece que AE sigue una distribución normal.

Recupero la función creada el la A2 para el intervalo de confianza.

```
IC <- function( x, NC ){
alpha<-1-NC
sd<-sd(x)
N<-length(x)
mean<-mean(x)
z<-qnorm(alpha/2, lower.tail=FALSE)
L<-mean-(z*(sd/sqrt(N)))
U<-mean+(z*(sd/sqrt(N)))
return (c(L,U))
}
```

```
IC(data$AE[data$genero=="F"], 0.95)
```

```
## [1] 1.429231 1.617330
```

```
IC(data$AE[data$genero=="M"], 0.95)
```

```
## [1] 1.483438 1.684087
```

Podemos observar que los intervalos de confianza para las mujeres es de (1.429, 1.617) y el de hombres (1.483, 1.684) por lo que el valor de los hombres es levemente superior. El intervalo de confianza indica que el 95% de las muestras se encontrarían en este rango (con un número elevado de muestras).

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.

4.1 Hipótesis

Escribir la hipótesis nula y alternativa.

La pregunta que queremos responder es: **¿La capacidad pulmonar media de las mujeres es diferentes a la de los hombres?**

- H0: mean_female = mean_male
- H1: mean_female != mean_male

4.2 Contraste

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

Por el teorema del límite central, asumimos normalidad. Ya que tenemos una muestra de un tamaño grande y aplicamos el test de la media.

El test a aplicar es **un test de dos muestras sobre la media con varianzas desconocidas**. Por lo que hemos de comparar las varianzas de las dos muestras.

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

```
##
##  F test to compare two variances
##
## data:  data_female and data_male
## F = 1.161, num df = 143, denom df = 108, p-value = 0.4152
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8104275 1.6484936
## sample estimates:
## ratio of variances
##           1.161001
```

El resultado del test es un valor p>0.05.Por tanto, asumimos igualdad de varianzas.En consecuencia, el test es 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.

```
CM<-function(x, y, NC=0.95, var.equal=TRUE, alternative="two.sided"){
  sd.x <- sd(x)
  N.x<-length(x)
  mean.x<-mean(x)
  sd.y <- sd(y)
  N.y <- length(y)
  mean.y <- mean(y)
  alpha <- 1-NC
  if(var.equal){ # Varianzas desconocidas iguales (pg. 42 apuntes)
    S<-sqrt((((N.x-1)*(sd.x^2))+((N.y-1)*(sd.y^2))) / (N.x+N.y-2))
    denom<-S*sqrt((1/N.x)+(1/N.y))
    t<-(mean.x-mean.y)/denom
    gl<-N.x+N.y-2 # Grados de libertad
  }
  else{ # Varianzas desconocidas diferentes
    denom<-((((sd.x^2)/N.x)^2)/(N.x-1))+((((sd.y^2)/N.y)^2)/(N.y-1))
    gl<-((((sd.x^2)/N.x)+((sd.y^2)/N.y))^2)/denom
    t<-(mean.x-mean.y)/sqrt(((sd.x^2)/N.x)+((sd.y^2)/N.y))
  }
  if (alternative=="two.sided"){ # se reparte izquierda y derecha de la curva
    crit_value<-qt(1-(alpha/2),gl)
    pvalue<-pt(abs(t), gl, lower.tail=FALSE )*2 # por eso multiplicamos por 2 p
  }
  else if (alternative == "greater"){
    crit_value<-qt(1-alpha, gl)
    pvalue<-pt(t, gl, lower.tail=FALSE )
  }
  else if (alternative == "less"){
    crit_value <-qt(alpha, gl, lower.tail=TRUE )
    pvalue<-pt(t, gl, lower.tail=TRUE )
  }
  return (c(t,crit_value,pvalue,gl))
}
```

Haremos uso de esta función en el siguiente apartado también.

```
cm_genero<-CM(data$AE[data$genero=="F"], data$AE[data$genero=="M"],var.equal=TRUE, alternative="two.sid
cm_genero
```

```
## [1]  -0.8531624   1.9694602   0.3943827 251.0000000
```

4.4 Interpretación

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

```
data.frame(T=c(cm_genero[1]),CRIT_VALUE=c(cm_genero[2]), PVALUE=c(cm_genero[3]),GL=c(cm_genero[4]))
```

```
##            T CRIT_VALUE    PVALUE  GL
## 1 -0.8531624    1.96946 0.3943827 251
```

El valor crítico para un nivel de confianza del 95% es 1.96946.

El valor T es -0.8531624. Por lo tanto **aceptamos la hipótesis nula**, ya que se encuentra en el rango de valores aceptados (-1.96,1.96).

Queda confirmado por el pvalue de 0.3943827 siendo mayor que el nivel de significación alpha=0.05.

Al aceptar la hipótesis nula asumimos que las mujeres tienen en promedio una capacidad pulmonar igual a la de los hombres.

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.

5.1 Hipótesis

Escribir la hipótesis nula y alternativa.

La pregunta que queremos responder es: **¿La capacidad pulmonar media de los fumadores es inferior a la de los no fumadores?**

- H0: mean_fumador = mean_no_fumador
- H1: mean_fumador < mean_no_fumador

5.2 Contraste

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

Por el teorema del límite central, asumimos normalidad. Ya que tenemos una muestra de un tamaño grande y aplicamos el test de la media.

El test a aplicar es **un test de dos muestras sobre la media con varianzas desconocidas**. Por lo que hemos de comparar las varianzas de las dos muestras.

```
data_fumador<-data$AE[data$Tipo=="FI" | data$Tipo=="FL" | data$Tipo=="FM"]
data_no_fumador<-data$AE[data$Tipo=="NF" | data$Tipo=="NI" | data$Tipo=="FP"]
var.test(data_fumador,data_no_fumador)
```

```
##
##  F test to compare two variances
##
## data:  data_fumador and data_no_fumador
## F = 0.84225, num df = 120, denom df = 131, p-value = 0.3399
## alternative hypothesis: true ratio of variances is not equal to 1
```

```
## 95 percent confidence interval:
##  0.5932431 1.1990678
## sample estimates:
## ratio of variances
##          0.8422516
```

El resultado del test es un valor p>0.05.Por tanto, asumimos igualdad de varianzas.En consecuencia, el test es de dos muestras independientes sobre la media con varianzas desconocidas iguales. El test es unilateral por la izquierda.

5.3 Preparación de los datos

Preparad las muestras. Una de ellas contiene los valores de AE de los fumadores y la otra, los valores de AE de los no fumadores y fumadores pasivos.

Ya hemos las hemos separado en el paso anterior.

- Los fumadores bajo la variable **data_fumador**
- Los no fumadores bajo la variable **data_no_fumador**

5.4 Cálculos

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

```
cm_tipo<-CM(data_fumador, data_no_fumador,var.equal=TRUE, alternative="less")
cm_tipo
```

```
## [1] -7.049383e+00 -1.650947e+00  8.663692e-12  2.510000e+02
```

5.5 Interpretación

Interpretar el resultado del contraste

```
data.frame(T=c(cm_tipo[1]),CRIT_VALUE=c(cm_tipo[2]), PVALUE=c(cm_tipo[3]),GL=c(cm_tipo[4]))
```

```
##           T CRIT_VALUE       PVALUE  GL
## 1 -7.049383  -1.650947 8.663692e-12 251
```

El valor crítico para un nivel de confianza del 95% es -1.650947.

El valor T es -7.049383. Por lo que se sobrepasa del valor crítico y **rechazamos la hipótesis nula**, ya que se encuentra fuera del rango de valores aceptados (-1.65,1.65).

Queda confirmado por el pvalue de 8.663692e-12 siendo mucho menor que el nivel de significación alpha=0.05.

Al rechazar la hipótesis nula concluimos que la capacidad pulmonar de los fumadores es inferior a la de los no fumadores.

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, siguiendo los pasos que se especifican a continuación.

6.1 Cálculo

Calculad el modelo de regresión lineal. Podéis usar la función lm.

```
modelo1 <- lm(AE~Tipo + edad + genero, data)
```

6.2 Interpretación

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

```
summary(modelo1)
```

```
##
## Call:
## lm(formula = AE ~ Tipo + edad + genero, 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 ***
## TipoFL       0.338459   0.080850   4.186 3.96e-05 ***
## TipoFM       0.046357   0.082133   0.564    0.573
## TipoFP       0.394342   0.081470   4.840 2.30e-06 ***
## TipoNF       0.781808   0.077004  10.153  < 2e-16 ***
## TipoNI       0.423523   0.080259   5.277 2.89e-07 ***
## edad        -0.030951   0.002276 -13.601  < 2e-16 ***
## generoM     -0.002321   0.047033  -0.049    0.961
## ---
## 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
```

El valor de Adjusted R-squared es de 0.571. Es decir, que el modelo explica un 57% de la varianza de la capacidad pulmonar de los registros. Lo que indica una capacidad explicativa moderadamente buena. El valor de p-value es de 2.2e-16 por lo que significativamnete menor que 0.05, indicando que el conjunto de variables explicicativas contribuyen significativamente en la capacidad pulmonar.

La variable *edad* es significativamente correlacionada negativamente con la variable *AE*. Es decir, que a mayor edad, menor capacidad pulmonar. También destacan los tipos No-Fumador(NF) y Fumador que no inhala(NI) que se encuentran en relación positiva a la variable *AE*.

6.3 Bondad de ajuste

Evaluad la calidad del modelo.

Podemos evaluar la calidad utilizando la función de *anova()*

```r
anova(modelo1, test="Chisq")
```

```
## Analysis of Variance Table
##
## Response: AE
##            Df Sum Sq Mean Sq  F value Pr(>F)
## Tipo        5 20.856  4.1712  31.2280 <2e-16 ***
## edad        1 24.905 24.9047 186.4526 <2e-16 ***
## genero      1  0.000  0.0003   0.0024 0.9607
## Residuals 245 32.725  0.1336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observamos que todas las variables Tipo y edad son muy significativas, pero la variable genero no demasiado.

También podemos aplicar la suma de cuadrados de los residuos de Pearson para evaluar la eficiencia, en conjunto con el test Chi-cuadrado.

```r
sum(residuals(modelo1,type="pearson")^2)
```

```
## [1] 32.72501
```

```r
1-pchisq(sum(residuals(modelo1,type="pearson")^2),1)
```

```
## [1] 1.061627e-08
```

Apreciamos que el modelo sigue siendo significativo ya que el resultado es menor a la suma de cuadrados de los residuos de Pearson

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.

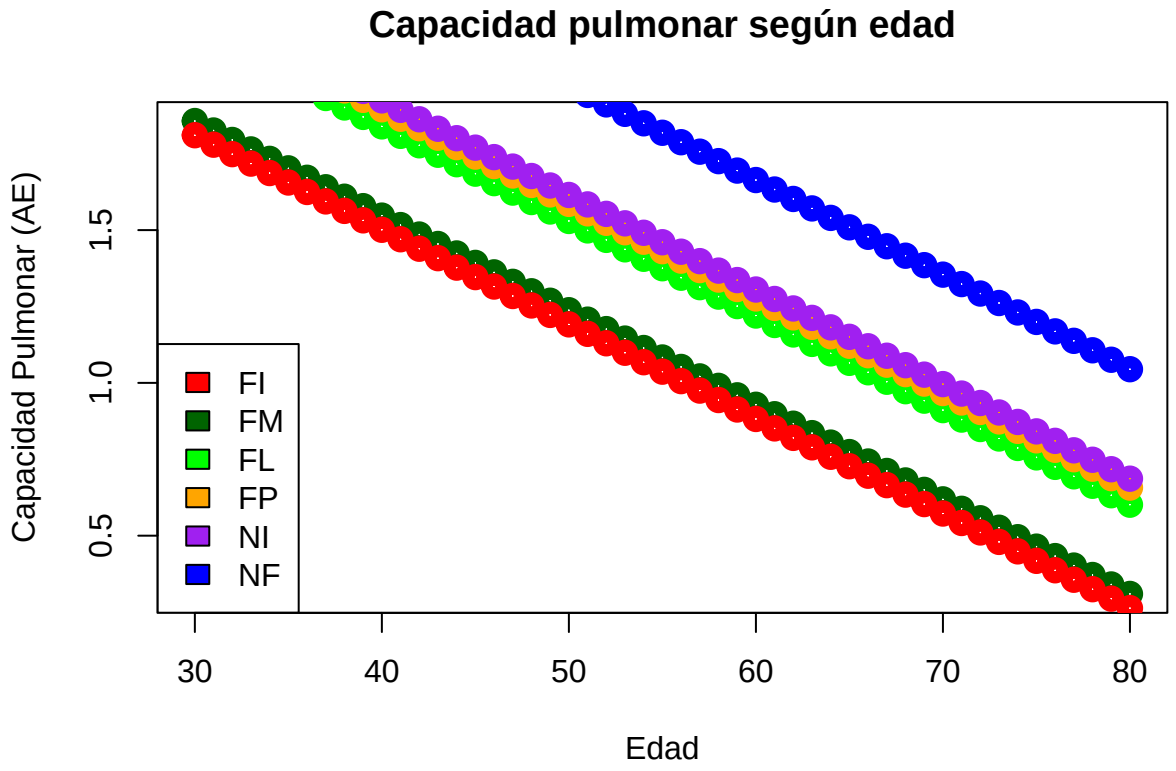
```r
registro_fm<-data.frame(Tipo="FM",edad=seq(30,80,by=1),genero="M")
registro_fi<-data.frame(Tipo="FI",edad=seq(30,80,by=1),genero="M")
registro_fl<-data.frame(Tipo="FL",edad=seq(30,80,by=1),genero="M")
registro_fp<-data.frame(Tipo="FP",edad=seq(30,80,by=1),genero="M")
registro_ni<-data.frame(Tipo="NI",edad=seq(30,80,by=1),genero="M")
registro_nf<-data.frame(Tipo="NF",edad=seq(30,80,by=1),genero="M")

predictions_fm<-predict(modelo1,registro_fm,type="response")
predictions_fi<-predict(modelo1,registro_fi,type="response")
predictions_fl<-predict(modelo1,registro_fl,type="response")
predictions_fp<-predict(modelo1,registro_fp,type="response")
predictions_ni<-predict(modelo1,registro_ni,type="response")
predictions_nf<-predict(modelo1,registro_nf,type="response")
```

Ploteamos la gráfica

```
plot(c(30:80), as.numeric(predictions_fm),main="Capacidad pulmonar según edad", xlab="Edad",
     ylab="Capacidad Pulmonar (AE)",pch=19, col="white")
points(c(30:80),as.numeric(predictions_fm), col="darkgreen",lwd=6)
points(c(30:80),as.numeric(predictions_fi), col="red",lwd=6)
points(c(30:80),as.numeric(predictions_fl), col="green",lwd=6)
points(c(30:80),as.numeric(predictions_fp), col="orange",lwd=6)
points(c(30:80),as.numeric(predictions_ni), col="purple",lwd=6)
points(c(30:80),as.numeric(predictions_nf), col="blue",lwd=6)

legend("bottomleft",legend = c("FI","FM","FL","FP","NI","NF"), fill= c("red","darkgreen","green","orang
```



Para la tabla creo un dataframe con todos los valores creados y hago un head

```
table_fm<-data.frame(edad=c(30:80),AE=as.numeric(predictions_fm),Tipo="FM",genero="M")
table_fi<-data.frame(edad=c(30:80),AE=as.numeric(predictions_fi),Tipo="FI",genero="M")
table_fl<-data.frame(edad=c(30:80),AE=as.numeric(predictions_fl),Tipo="FL",genero="M")
table_fp<-data.frame(edad=c(30:80),AE=as.numeric(predictions_fp),Tipo="FP",genero="M")
table_ni<-data.frame(edad=c(30:80),AE=as.numeric(predictions_ni),Tipo="NI",genero="M")
table_nf<-data.frame(edad=c(30:80),AE=as.numeric(predictions_nf),Tipo="NF",genero="M")
head(table_fm)
```

```
##   edad       AE Tipo genero
## 1   30 1.856905   FM      M
## 2   31 1.825953   FM      M
## 3   32 1.795002   FM      M
```

```
## 4   33 1.764050   FM      M
## 5   34 1.733099   FM      M
## 6   35 1.702148   FM      M
```

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:

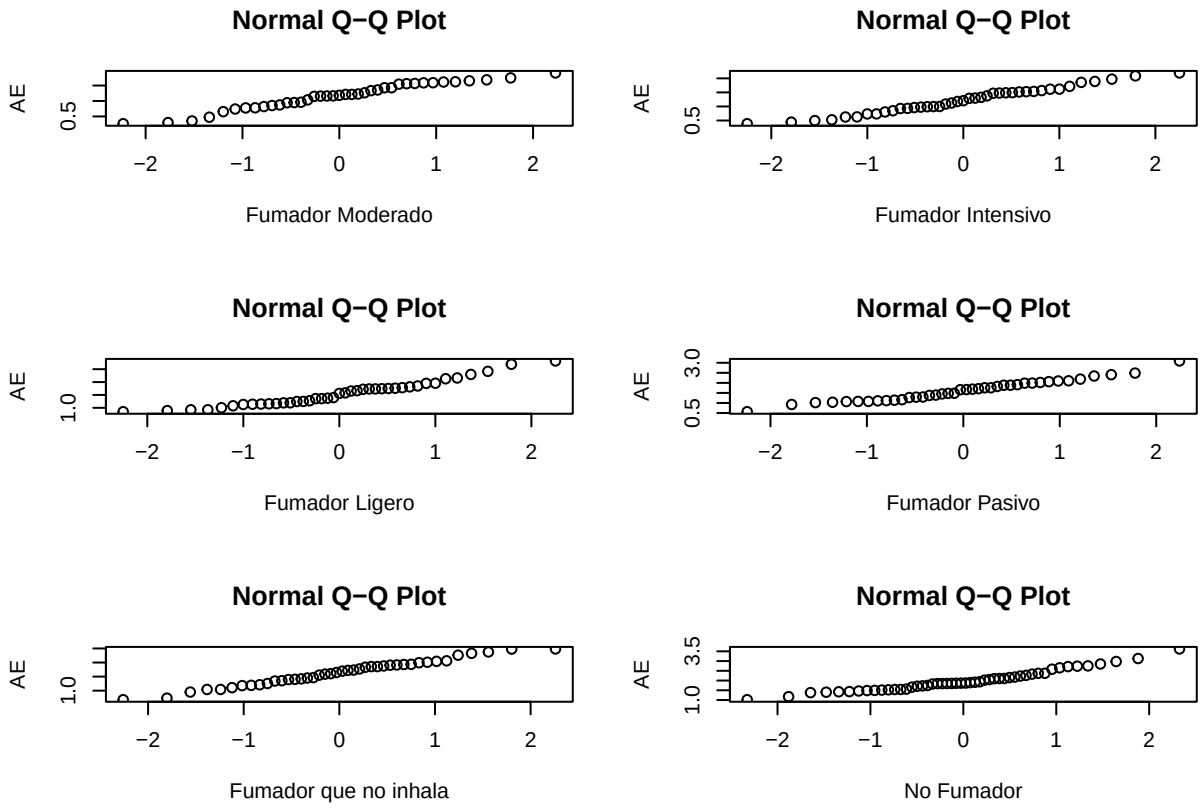
- *¿Existen diferencias entre la capacidad pulmonar (AE) entre los distintos tipos de fumadores/no fumadores?*
- *Si existen diferencias, ¿entre qué grupos están estas diferencias?*

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.

Para evaluar la normalidad plotamos los graficps QQplots que nos permiten observar similitudes con una distribución normal.

```
par(mfrow=c(3,2))
qqnorm(data[data$Tipo == "FM",]$AE,  xlab='Fumador Moderado', ylab='AE')
qqnorm(data[data$Tipo == "FI",]$AE,  xlab='Fumador Intensivo', ylab='AE')
qqnorm(data[data$Tipo == "FL",]$AE,  xlab='Fumador Ligero', ylab='AE')
qqnorm(data[data$Tipo == "FP",]$AE,  xlab='Fumador Pasivo', ylab='AE')
qqnorm(data[data$Tipo == "NI",]$AE,  xlab='Fumador que no inhala', ylab='AE')
qqnorm(data[data$Tipo == "NF",]$AE,  xlab='No Fumador', ylab='AE')
```




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

El gráfico Normal Q-Q Plot muestra que los datos se ajustan a una normal para cada Tipo de fumador y no muestra un patrón aleatorio. No obstante, realizaremos el test Shapiro para asegurarnos de que esto es cierto.

```
sapply(c("FM","FI","FL","FP","NI","NF"), function(x) shapiro.test(data[data$Tipo == x,]$AE))
```

```
##           FM                           FI
## statistic 0.9635552                    0.9781902
## p.value   0.233875                     0.6074302
## method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
## data.name "data[data$Tipo == x, ]$AE"   "data[data$Tipo == x, ]$AE"
##           FL                           FP
## statistic 0.9435042                    0.9716144
## p.value   0.04151539                   0.4042803
## method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
## data.name "data[data$Tipo == x, ]$AE"   "data[data$Tipo == x, ]$AE"
##           NI                           NF
## statistic 0.9831918                    0.9506917
## p.value   0.7833408                    0.03642425
## method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
## data.name "data[data$Tipo == x, ]$AE"   "data[data$Tipo == x, ]$AE"
```

Para los Tipos *FM,FI,FP* y *NI* podemos asumir normalidad dado que sigue una distribución normal, ya que el p-valor es mayor que 0.05 (nivel de significancia normalmente utilizado).

Para *NF* y *FL* encontramos un p-value menor de 0.05.

7.2 Homoscedasticidad: Homogeneidad de varianzas

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

Para comprobar la igualdad de varianzas aplicaremos el test de fligner:

```
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
```

Como tenemos un p-value mayor al nivel de significancia de 0.05, asumimos que las varianzas son iguales.

7.3 Hipótesis nula y alternativa

Independientemente de los resultados sobre la normalidad e homoscedasticidad 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.

El factor Tipo tiene 6 niveles: FM, FI, FL, FP, NI y NF. Las hipótesis son:

· HO: mean_FM = mean_FI = mean_FL = mean_FP = mean_NI = mean=NF

· H1: mean_i != mean_j

Donde i,j son posibles valores de los 6 niveles que tiene Tipo.

7.4 Cálculo ANOVA

Podéis usar la función aov.

```
mod1<-aov(AE~Tipo,data)
mod1_info<-anova(mod1)
mod1_info
```

```
## Analysis of Variance Table
##
## Response: AE
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Tipo        5 20.856  4.1712  17.877 4.026e-15 ***
## Residuals 247 57.630  0.2333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7.5 Interpretación

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

```
mod1_info
```

```
## Analysis of Variance Table
##
## Response: AE
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Tipo        5 20.856  4.1712  17.877 4.026e-15 ***
## Residuals 247 57.630  0.2333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los resultados de la prueba son: Sum Sq= 20.856, Mean Sq=4.1712, F value=17.877 y pvalor 4.026e-15.

Dado que el pvalor es mucho menor que 0.05 concluimos que el facto analizado es siginificativo. Rechazamos la hipótesis nula y asumimos que si hay diferencias significativas.

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.

SSW

```
ssw<-sum(sum((data$AE[data$Tipo=="FM"] - mean(data$AE[data$Tipo=="FM"]))^2),
sum((data$AE[data$Tipo=="FI"] - mean(data$AE[data$Tipo=="FI"]))^2),
sum((data$AE[data$Tipo=="FL"] - mean(data$AE[data$Tipo=="FL"]))^2),
sum((data$AE[data$Tipo=="FP"] - mean(data$AE[data$Tipo=="FP"]))^2),
sum((data$AE[data$Tipo=="NI"] - mean(data$AE[data$Tipo=="NI"]))^2),
sum((data$AE[data$Tipo=="NF"] - mean(data$AE[data$Tipo=="NF"]))^2))
ssw
```

```
## [1] 57.63008
```

SST

```
sst<-sum(sum((data$AE[data$Tipo=="FM"] - mean(data$AE))^2),
sum((data$AE[data$Tipo=="FI"] - mean(data$AE))^2),
sum((data$AE[data$Tipo=="FL"] - mean(data$AE))^2),
sum((data$AE[data$Tipo=="FP"] - mean(data$AE))^2),
sum((data$AE[data$Tipo=="NI"] - mean(data$AE))^2),
sum((data$AE[data$Tipo=="NF"] - mean(data$AE))^2))
sst
```

```
## [1] 78.48591
```

SSB

```
ssb<-sum(sum((rep(mean(data$AE[data$Tipo=="FM"]), times=length(data$AE[data$Tipo=="FM"])) - mean(data$AE))^2),
sum((rep(mean(data$AE[data$Tipo=="FI"]), times=length(data$AE[data$Tipo=="FI"])) - mean(data$AE))^2),
sum((rep(mean(data$AE[data$Tipo=="FL"]), times=length(data$AE[data$Tipo=="FL"])) - mean(data$AE))^2),
sum((rep(mean(data$AE[data$Tipo=="FP"]), times=length(data$AE[data$Tipo=="FP"])) - mean(data$AE))^2),
sum((rep(mean(data$AE[data$Tipo=="NI"]), times=length(data$AE[data$Tipo=="NI"])) - mean(data$AE))^2),
sum((rep(mean(data$AE[data$Tipo=="NF"]), times=length(data$AE[data$Tipo=="NF"])) - mean(data$AE))^2))
ssb
```

```
## [1] 20.85584
```

Comprobación SST = SSB+SSW

```
ssb+ssw
```

```
## [1] 78.48591
```

```
sst
```

```
## [1] 78.48591
```

Para la variable SSW es el cuadrado de la diferencia entre cada valor de un tipo y la media de ese mismo tipo. Para la variable SST es el cuadrado de la diferencia entre cada valor de un tipo y la media de todos los tipos. Para la variable SSB es el cuadrado de la diferencia entre la media de cada tipo por la cantidad de elementos y la media de ese mismo tipo.

Básicamente se utilizan para representar diferentes tipos de variaciones que pueden existir dentro de los diferentes grupos de datos.

Los grados de libertado son:

· SSB > 1

· SSW > n - 2 > 251

· SST > n-1 > 252

valor F

```
mse<-ssw/251
f<-ssb/mse # ssb es igual a MSR
f
```

```
## [1] 90.83477
```

El valor estadístico de F es 90.83. Usando una alpha de 0.05 tenemos un pvalue de:

· F_0.05;2,252 = 0.01095

Que podemos comprobar en la tabla de distribución de F ***link***

7.7 Fuerza de la relación

Calcular la fuerza de la relación e interpretar el resultado.

La fuerza de la relación es el cociente entre SSB/ SST.

```
ssb/sst
```

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

El resultado de 0.26 indica que el 26% explica la variabilidad en la capacidad pulmonar de cada tipo de fumador.

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.

Comparamos parejas de tipo de fumadores usando la función pairwise

```
attach(data)
```

```
## The following objects are masked from data (pos = 3):
##
##     AE, edad, genero, Tipo
```

```
pairwise.t.test(AE, Tipo)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  AE and Tipo
##
##    FI      FL      FM      FP      NF
## FL 0.00826 -       -       -       -
## FM 1.00000 0.00218 -       -       -
## FP 0.00189 1.00000 0.00031 -       -
## NF 7.5e-12 0.00031 4.0e-13 0.00244 -
## NI 0.00107 1.00000 0.00017 1.00000 0.00287
##
## P value adjustment method: holm
```

Cómo podemos observar la mayoría de los valores son de un pvalue inferior al nivel de significación alpha=0.05

Por lo que concluímos que la diferencia de la media entre diferentes parejas de tipo de fumadora es estadisticamente significante

El valor que vemos es el valor pvalue. Es decir, para el primer ejemplo 0.00826 indica la media de diferencia para la capacidad pulmonar en personas de tipo FI y tipo FL

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.

Para ello usaremos el metodo ajustado bonferroni dentro de la función pairwise

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

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

El valor p ajustado para la diferencia media en las personas entre tipo FI y FL es 0.02477.

Por norma general, encontramos un incremento en pvalue usando el metodo de bonferroni.

9 ANOVA multifactorial

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

Agrupación por género y tipo de fumador

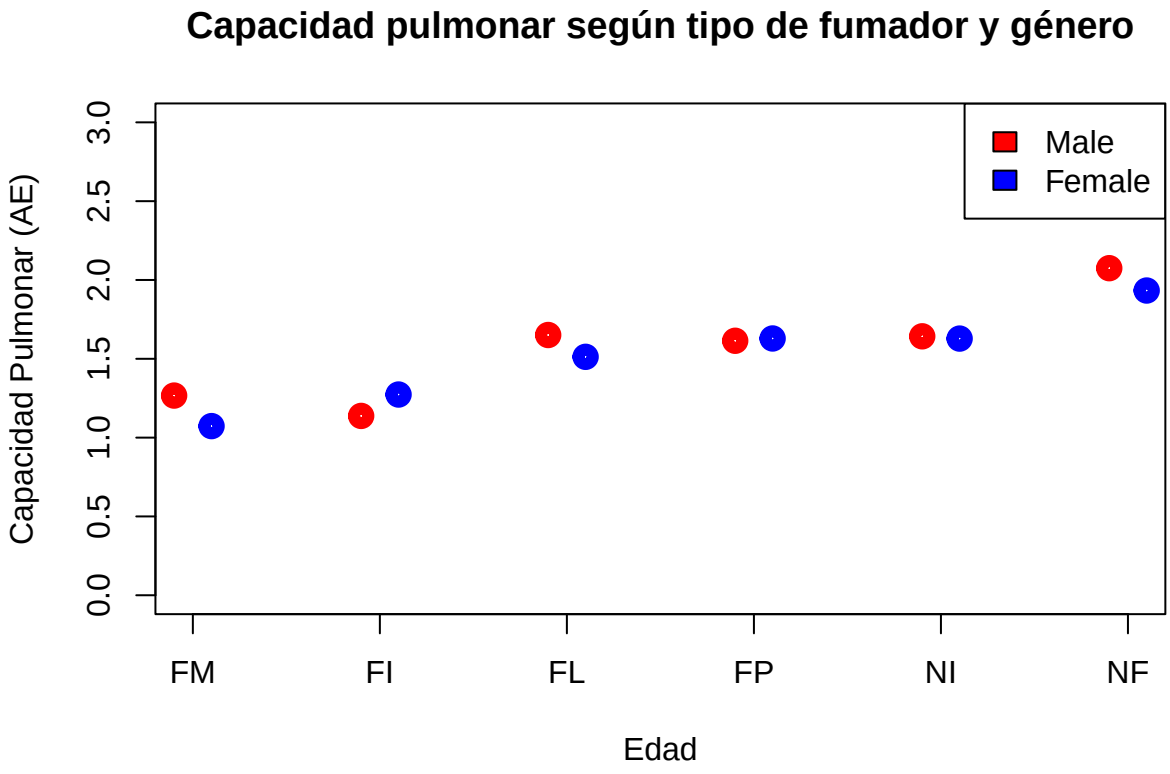
TABLE	Male	Female
FM	1.266695	1.073019
FI	1.137577	1.273318
FL	1.650977	1.512600
FP	1.614373	1.628993
NI	1.642543	1.628288
NF	2.074773	1.933138

Gráfico del valor AE para cada tipo de fumador y género

```
plot(c(1:6), c(0,1,2,3,2.5,1.5),main="Capacidad pulmonar según tipo de fumador y género", xlab="Edad", 
     ylab="Capacidad Pulmonar (AE)",pch=19, col="white")
points(c(0.9,1.1),c(FM[1],FM[2]), col=c("red","blue"),lwd=6)
points(c(1.9,2.1),c(FI[1],FI[2]), col=c("red","blue"),lwd=6)
points(c(2.9,3.1),c(FL[1],FL[2]), col=c("red","blue"),lwd=6)
points(c(3.9,4.1),c(FP[1],FP[2]), col=c("red","blue"),lwd=6)
points(c(4.9,5.1),c(NI[1],NI[2]), col=c("red","blue"),lwd=6)
points(c(5.9,6.1),c(NF[1],NF[2]), col=c("red","blue"),lwd=6)

xtick<-c("FM","FI","FL","FP","NI","NF")
axis(1, at=1:6, labels=xtick)

legend("topright",legend = c("Male","Female"), fill= c("red","blue"))
```



No encontramos una fuerte interacción en el genero-tipo. La media según el tipo de fumador aveces es superior y otras inferior para las mujeres.

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.

```
mod2<-aov(AE~Tipo*genero,data)
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: AE
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Tipo          5 20.856  4.1712 17.7392 5.809e-15 ***
## genero        1  0.197  0.1970  0.8378    0.3610
## Tipo:genero   5  0.765  0.1529  0.6504    0.6615
## Residuals   241 56.668  0.2351
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9.3 Interpretación

Interpretad el resultado.

```
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: AE
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Tipo          5 20.856  4.1712 17.7392 5.809e-15 ***
## genero        1  0.197  0.1970  0.8378    0.3610
## Tipo:genero   5  0.765  0.1529  0.6504    0.6615
## Residuals   241 56.668  0.2351
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tanto los factores principales como la interacción entre factores son significativos. Por tanto, la capacidad pulmonar en función del tipo, es ligeramente diferente según si se trata de un hombre o de una mujer.

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
p2.1	Diferencias capacidad pulmonar según género	M: 1.583762 F: 1.52328	Mujeres tienen levemente menor media
p2.2	Diferencias capacidad pulmonar según edad	Gráfico Apartado 2.2	A mayor edad, menor capacidad pulmonar
p2.3	Media de tipos de fumadores	Entre 1.16 y 1.63	Orden: FM<FI<FL<FP<NI<NF
p3	Intervalo confianza según género al 95%	M: [1.48,1.68], F:[1.42,1.61]	95% de las muestras se encuentran en este rango
p4.2	Comparación varianzas	pvalue = 0.4152	Asumimos igualdad de varianzas. Test bilateral
p4.4	Contraste de medias	T:-0.85,v-crit:1.96,pvalue:0.39	Aceptamos hipotesis nula. Mujeres tiene en promedio capacidad igual a la de hombres
p5.2	Comparación varianzas	pvalue = 0.3399	Asumimos igualdad de varianzas. Test unilateral por la izquierda
p5.5	Contraste de medias	T:-0.7,v-crit:-1.65,pvalue:8e-12	Rechazamos la hipotesis nula. La capacidad pulmonar de los fumadores es inferior a la de no fumadores
p6.2	Diagnosis modelo	Anova, test chi-cuadrado, R-squared=0.58	Modelo significativo con capacidad explicativa moderadamente buena
p6.4	Predicción	Todas las edades de 30 a 80 con todos los tipos de fumador	A mayor edad peor capacidad. Según tipo peor capacidad.
p7.1	Anova normalidad	Q-Q Plot y shapiro test	Siguen una normal para cada Tipo. pvalue>0.05
p7.2	Homoscedasticidad	Test de fligner	pvalue>0.05. Varianzas iguales
p7.5	Interpretación ANOVA	S=20.8,M=4.1,F=17.8,p=4e-15	Rechazamos hipótesis nula. Hay diferencias significativas. pvalue<0.05
p7.6	SST,SSW,SSB	T=78,W=57,B=20,p=0.01	nivel de significación <0.05
p7.7	Fuerza de la relación	0.26	26% explica la variabilidad en la capacidad pulmonar de cada tipo
p8.1	Test pairwise	pvalue<0.05	La diferencia de la media entre tipos es estadísticamente significante
p8.2	Corrección de Bonferroni	pvalue<0.05	La diferencia de la media entre tipos es estadísticamente significante
p9.3	ANOVA multifactorial	Variables Tipo y edad	Capacidad pulmonar según Tipo es ligeramente diferentes según género

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.

• Las mujeres tienen de media levemente una peor capacidad pulmonar, pero por norma general no hay diferencias significativas según género en este aspecto.

- A medida que la persona tenga más años peor será la capacidad pulmonar. Está tendencia es claramente notable para los datos.

- La media de los tipos de los fumadores nos permite ordenarlos según capacidad pulmonar. FM<FI<FL<FP<NI<NF

- En los modelos creados se observa una clara tendencia. Las personas NO fumadores tienen mayor capacidad pulmonar que las personas fumadoras.

- Las variables Tipo y edad tienen un impacto directo sobre la variable a predecir AE. La variable independiente genero se podría omitir en un modelo de predicción.

- Se han realizado diferentes predicciones según tipo de fumador y edad. La tendencia es a mayor edad y a mayor consumo al fumar, menor la capacidad pulmonar. En estas predicciones una persona No fumadora con 70 años tiene mejor capacidad pulmonar que una persona Fumadora Moderada con 50 años.