Resolución de la A4 - Análisis estadístico avanzado

Clara Matilde Roca de la Concha

2022-01-23

Índice

Antes de empezar, procedemos a instalar las librerías necesarias para esta práctica.

```
# Librerías
library(agricolae)
library(car)
```

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

```
library(dplyr)
```

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

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

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

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

```
library(ggplot2)
library(ggpubr)
library(knitr)
library(MASS)
```

```
##
## Attaching package: 'MASS'

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

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

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

```
library(reshape2)
library(stats)
```

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.

A continuación, tendremos que leer nuestros datos.

```
# Leemos el archivo
smokers <-read.csv("Fumadores.csv", sep=';')

# Examinamos los tipos de datos
str(smokers)
```

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

Revisaremos variable a variable:

AE

En primer lugar, analizamos si hay valores no numéricos y cuáles son.

```
# Vemos qué valores se transforman en NA al pasarlos a 'numeric'
AE.errores <- which(is.na(as.numeric(smokers$AE)))
```

```
## Warning in which(is.na(as.numeric(smokers$AE))): NAs introduced by coercion
```

```
smokers$AE[AE.errores]
```

```
##  [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"
```

Detectamos decimales separados por coma. Vamos a unificar todos los valores estableciendo el punto como separador decimal.

```
# Reemplazamos las comas por puntos
smokers$AE[AE.errores] <- gsub(",", ".", smokers$AE[AE.errores])

# Convertimos a formato numérico
smokers$AE <- round(as.numeric(smokers$AE), 2)

 # Revisamos las conversiones de las comas
smokers$AE[AE.errores]
```

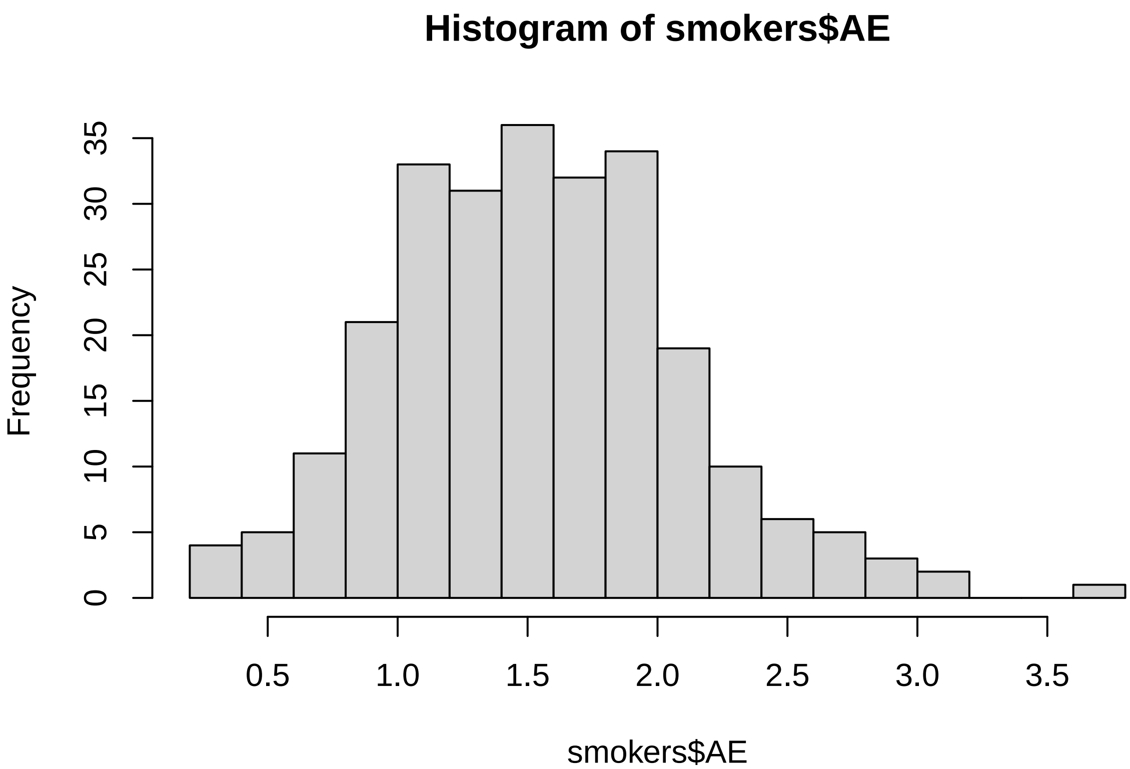
```
##  [1] 1.89 1.99 2.09 1.71 1.25 1.59 1.64 1.00 1.58 1.67 0.94 1.59 1.09 0.44 1.71
```

Hacemos una estadística descriptiva de la variable para comprobar que no hay valores inesperados y para hacernos una idea de la distribución.

```
summary(smokers$AE)
```

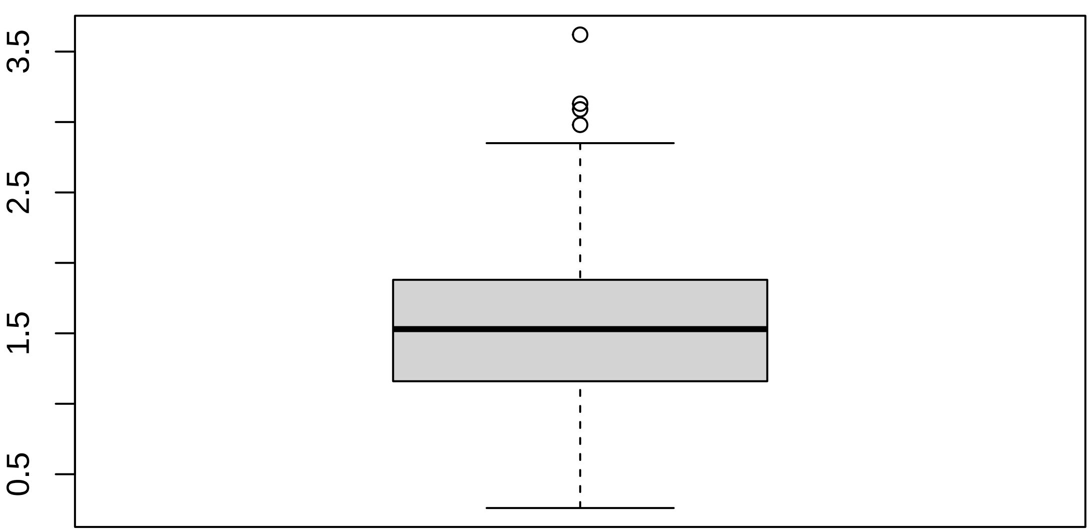
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.260   1.160   1.530   1.549   1.880   3.620
```

```
hist(smokers$AE, breaks=15)
```



Observamos posibles *outliers* en la cola derecha. Veámoslo en un diagrama de cajas.

```
boxplot(smokers$AE)
```



Llama la atención el último valor, 3.62, más aislado del resto.

Veamos en detalle esta fila.

```
max.AE.idx =  which(smokers$AE == max(smokers$AE))
smokers[max.AE.idx,]
```

```
##      AE Tipo genero edad
## 37 3.62   NF      F   17
```

Vemos que se trata de una mujer no fumadora de 17 años. Como veremos cuando analicemos la variable `age`, se trata de la persona más joven de la base de datos. Si a eso le añadimos que no fuma, podemos confirmar que el alto resultado de la prueba de de capacidad pulmonar es consistente con la investigación.

Tipo

En primer lugar, vamos a cambiar el nombre de la columna para que esté en minúsculas. No es de vital importancia, pero prefiero que trabajemos con nombres de columna con la misma lógica. Dejamos la primera variable en mayúsculas, ya que se trata de siglas, pero en este caso no se justifica.

```r
smokers$tipo <- smokers$Tipo
```

Hecho esto, analizamos los valores con los que trabajamos en una tabla.

```r
kable(table(smokers$tipo),
      col.names = c("Tipo", "#"),
      caption = "Niveles de la variable 'tipo'")
```

Cuadro 1: Niveles de la variable 'tipo'

Tipo	#
FM	1
fi	4
FI	37
FL	41
fm	9
FM	28
FM	1
FP	40
NF	50
NI	42

Observamos que existen problemas de uniformidad en los nombres de las categorías, por lo que procedemos a eliminar estos errores.

```r
# Eliminamos los espacios en blanco de los valores y cambiamos
# la variable a tipo factor una vez convertidos en mayúsculas
smokers$tipo <- as.factor(toupper(trimws(smokers$tipo)))

# Vemos los resultados en una tabla
kable(table(smokers$tipo),
      col.names = c("Tipo", "#"),
      caption = "Niveles de la variable 'tipo' tras la corrección")
```

Cuadro 2: Niveles de la variable 'tipo' tras la corrección

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

El número de categorías es la esperada. Seguimos.

genero

Procedemos de la misma forma que con la anterior variable.

```
kable(table(smokers$genero),
      col.names = c("Género", "#"),
      caption = "Niveles de la variable 'genero'")
```

Cuadro 3: Niveles de la variable 'genero'

Género	#
F	144
M	109

Vemos que está todo correcto, por lo que procedemos a convertir la variable a factor.Nótese, sin embargo, que hay un desbalance en la muestra: un mayor número de mujeres que de hombres.

```
smokers$genero <- as.factor(smokers$genero)
```

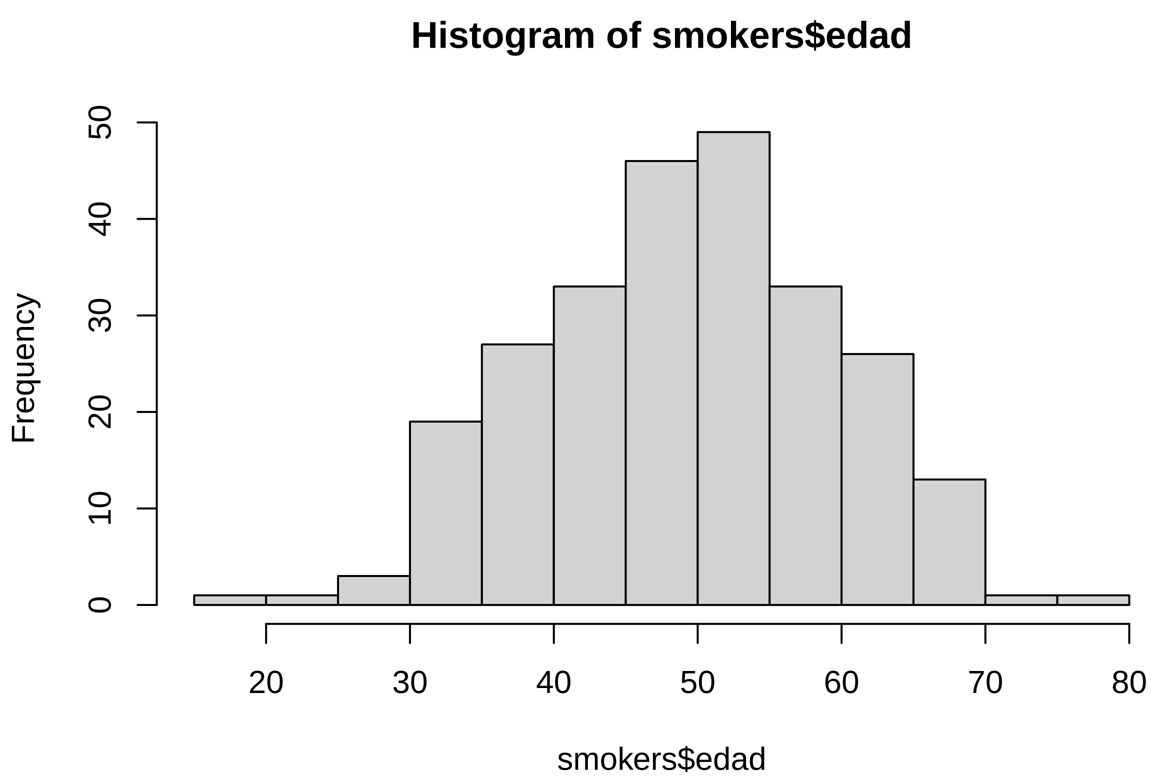
edad

Para esta última variable, realizaremos una estadística descriptiva y observaremos su distribución con un histograma.

```
summary(smokers$edad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   43.00   50.00   49.76   57.00   78.00
```

```
hist(smokers$edad, breaks=20)
```



Todos los valores se encuentran dentro de valores esperables. Como podemos observar, la distribución se acerca a una normal.

Terminado el preprocesado, pasamos a volcar los datos en el espacio de R, para poder trabajar con mayor agilidad.

```
attach(smokers)
```

2 Análisis descriptivo de la muestra

2.1 Capacidad pulmonar y género

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

En primer lugar mostramos las principales medidas de centro y de dispersión (robustas y no robustas).

```
grouped.gend <-smokers %>%
        group_by(genero) %>%
        summarise(n=n(),
                  min = min(AE),
                  Q1= round(quantile(AE, 0.25), 2),
                  media=round(mean(AE),2),
                  mediana=quantile(AE, 0.5),
                  Q3=quantile(AE, 0.75),
                  max = max(AE),
                  sd=round(sd(AE),2),
                  RIC = round(IQR(AE), 2),
                  DAM = round(mad(AE), 2))

kable(grouped.gend,
      caption ="Análisis descriptivo de 'AE' en relación al género")
```

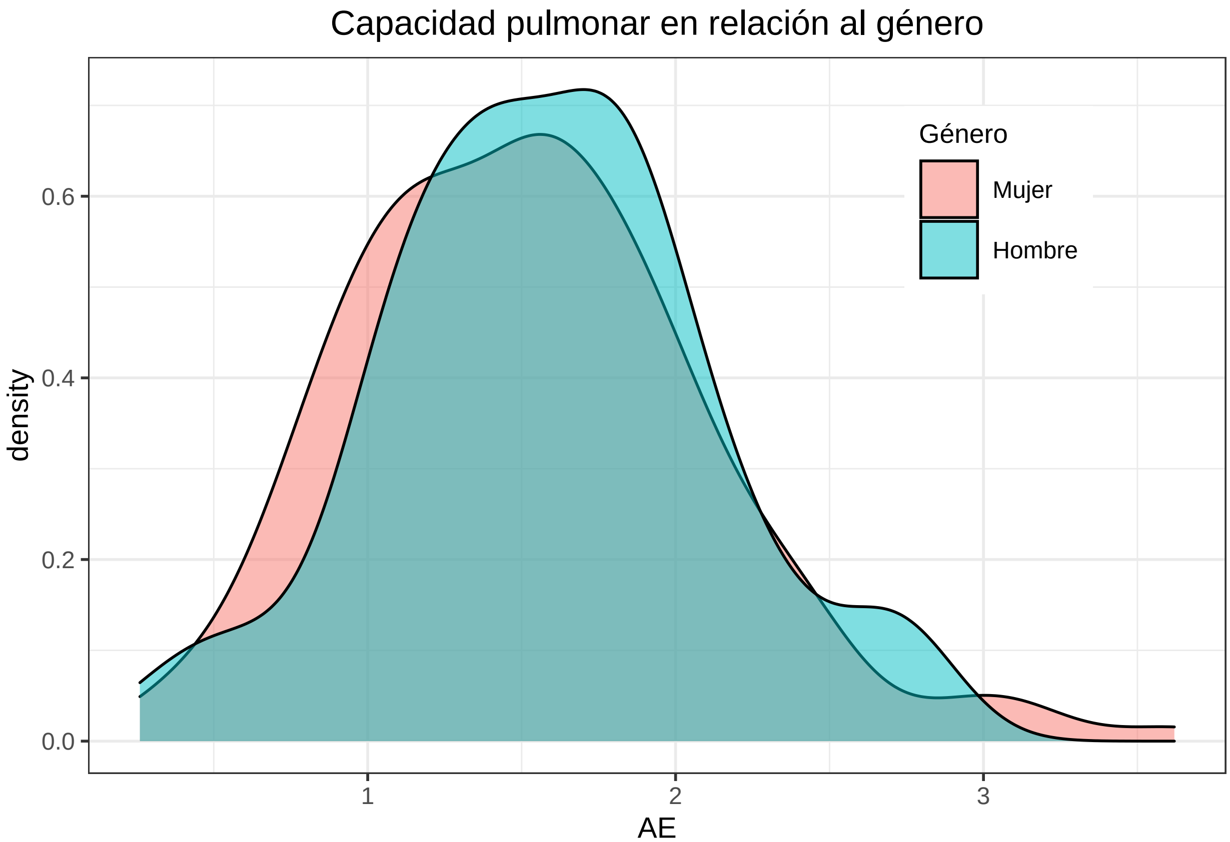
Cuadro 4: Análisis descriptivo de 'AE' en relación al género

genero	n	min	Q1	media	mediana	Q3	max	sd	RIC	DAM
F	144	0.30	1.13	1.52	1.50	1.88	3.62	0.58	0.75	0.56
M	109	0.26	1.21	1.58	1.57	1.89	2.85	0.53	0.68	0.50

Como vemos, el grupo de mujeres tiene un mínimo y un máximo mayores, pero presentan menores cifras en los estimadores de centro respecto a los hombres. La dispersión es mayor en los datos de las mujeres que en los de los hombres.

A contingunación, presentamos un grafico de densidad, ya que, al haber diferencia en el tamaño de las muestras, un histograma sería más complicado de interpretar.

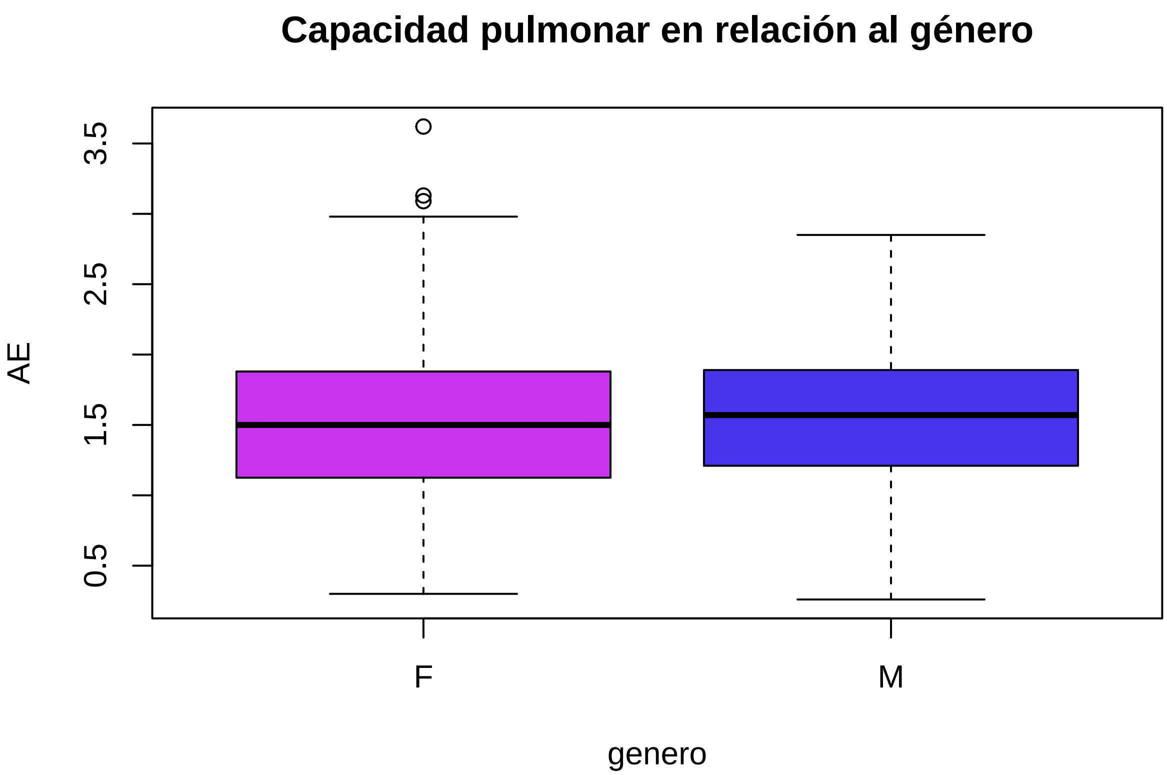
```r
# Gráfico de AE por género
ggplot(smokers, aes(x=AE,
              fill=genero)) +
              geom_density(alpha = 0.5) +
              theme_bw() +
              scale_fill_discrete(name="Género",labels=c("Mujer","Hombre"))+
              ggtitle("Capacidad pulmonar en relación al género") +
              theme(plot.title =element_text(hjust =0.5),
                    legend.title = element_text(size=10),
                    legend.key.size = unit(8, 'mm'),
                    legend.position = c(0.8, 0.8))
```



Con estos resultados, resulta difícil predecir si hay o no diferencias entre las poblaciones. Vemos cómo la muestra masculina concentra más sus valores hacia la media, mientras que la muestra femenina presenta mayor volumen de valores hacia la cola izquierda, si bien también representa los valores más altos.

Decidimos examinar los datos en un boxplot.

```r
boxplot(AE ~ genero,
        main="Capacidad pulmonar en relación al género",
        col=c("#c634eb","#4634eb"))
```



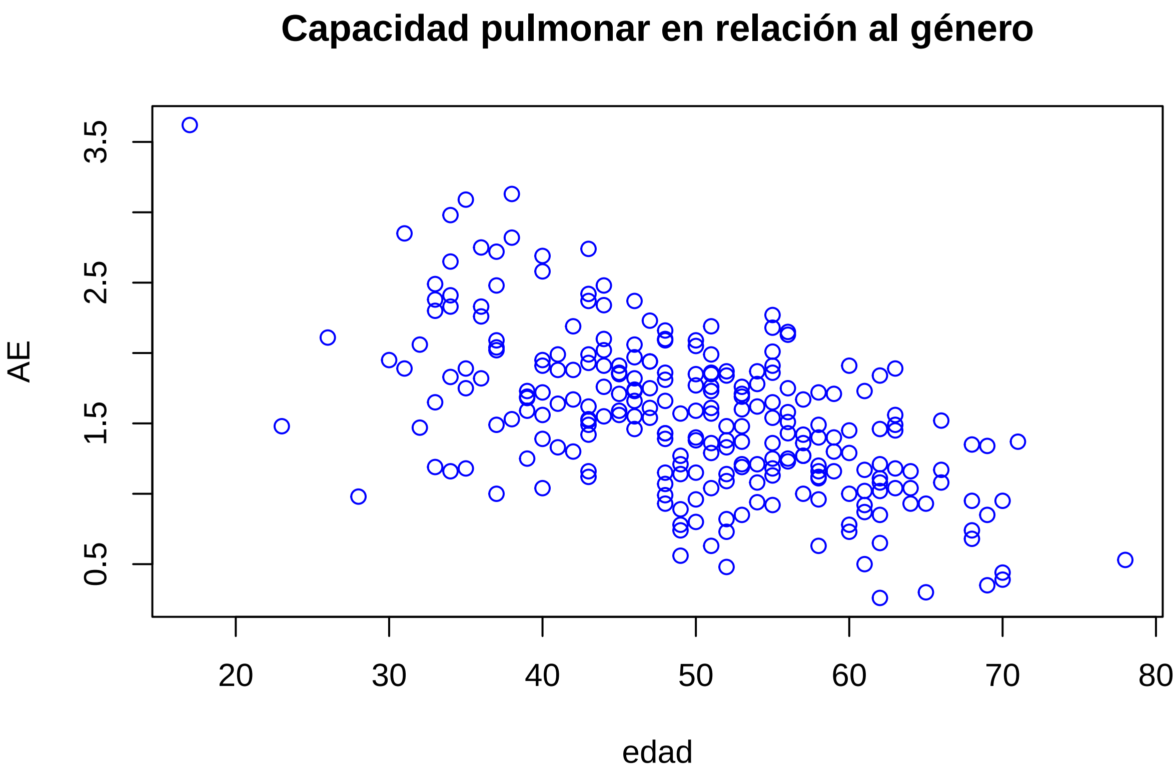
Efectivamente, las medianas y los rangos intercuartílicos son muy similares, si bien la muestra de mujeres presenta valores extremos (*outliers*) en la cola izquierda. El más extremo se corresponde con el comentado en el Ejercicio 1.

2.2 Capacidad pulmonar y edad

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

A continuación, presentamos un gráfico de dispersión para estudiar la relación entre la variable `AE` y la variable `edad`.

```r
plot(AE ~ edad,
     col='blue',
     main="Capacidad pulmonar en relación al género")
```



Al realizar el análisis visual, se observa una posible relación lineal con pendiente negativa entre la capacidad pulmonar y la edad, donde la capacidad pulmonar disminuye a medida que la edad del sujeto aumenta.

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.

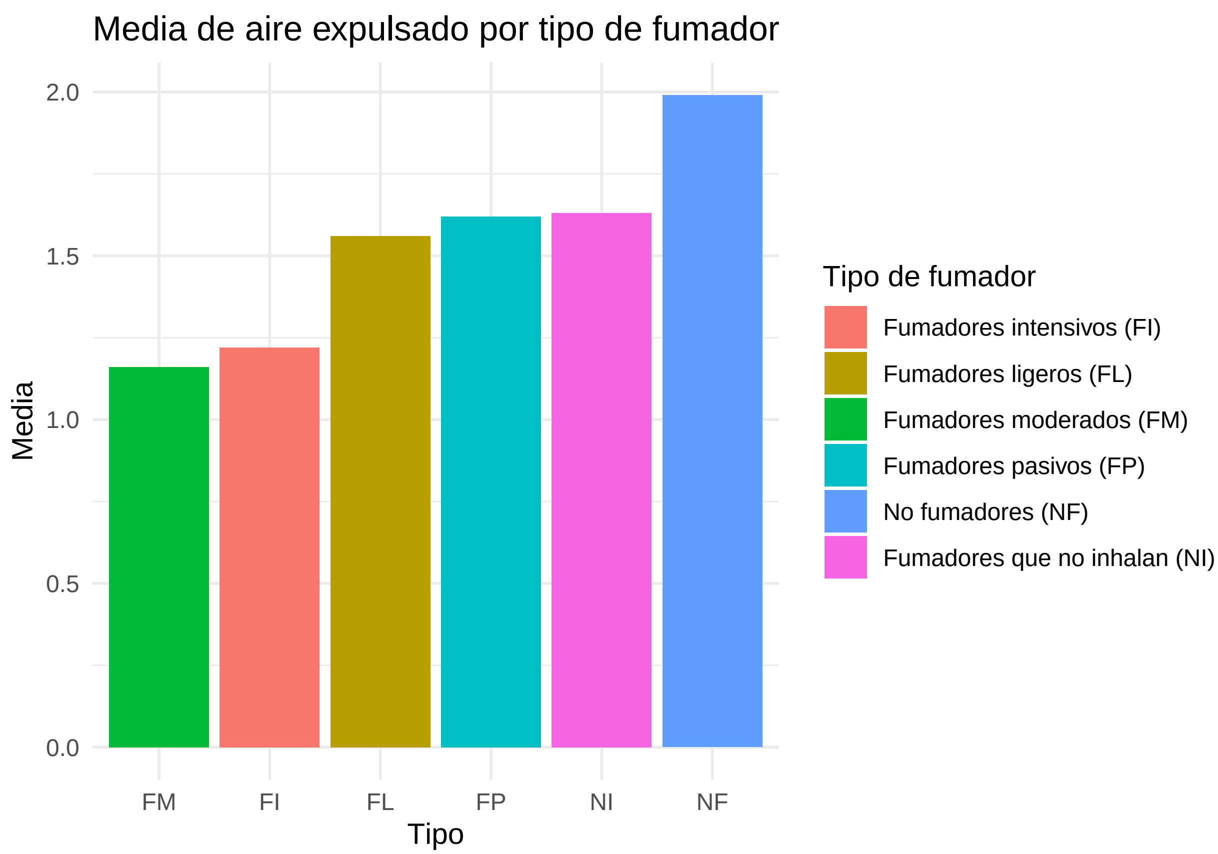
```
tipo.media <- smokers %>%
                group_by(tipo) %>%
                summarise(n=n(),
                          mean=round(mean(AE),2))
                          # Descomentar si se desea la tabla ordenada por media.
                          #%>% arrange(mean)

kable(tipo.media,
      col.names =c("tipo","#", 'AE media'),
      caption ="Número de personas y capacidad pulmonar media por tipo de fumador")
```

Cuadro 5: Número de personas y capacidad pulmonar media por tipo de fumador

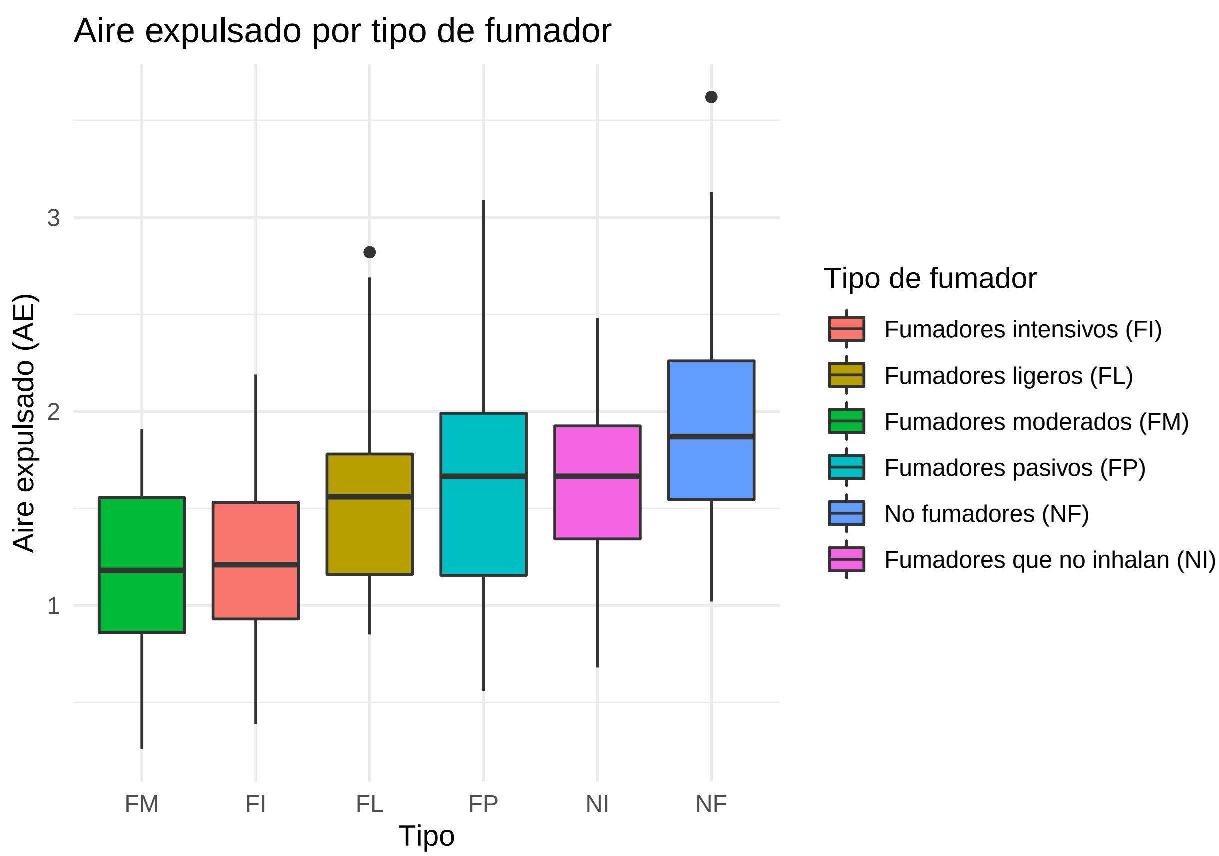
tipo	#	AE media
FI	41	1.22
FL	41	1.56
FM	39	1.16
FP	40	1.62
NF	50	1.99
NI	42	1.63

```r
# Reordenamos para que queden ordenados por media
ggplot(tipo.media, aes(x=reorder(tipo, mean),
                       y=mean, fill=tipo)) +
      geom_bar(stat="identity") +
      theme_minimal() +
      labs(title="Media de aire expulsado por tipo de fumador",
           x="Tipo", y="Media") +
      scale_fill_discrete(name="Tipo de fumador",
                          labels=c("Fumadores intensivos (FI)",
                                   "Fumadores ligeros (FL)",
                                   "Fumadores moderados (FM)",
                                   "Fumadores pasivos (FP)",
                                   "No fumadores (NF)",
                                   "Fumadores que no inhalan (NI)"))
```



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

```
ggplot(smokers, aes(x=reorder(tipo, AE),
                    y=AE, fill=tipo)) +
      geom_boxplot() +
      theme_minimal() +
      labs(title="Aire expulsado por tipo de fumador",
           x="Tipo",
           y="Aire expulsado (AE)") +
      scale_fill_discrete(name="Tipo de fumador",
                          labels=c("Fumadores intensivos (FI)",
                                   "Fumadores ligeros (FL)",
                                   "Fumadores moderados (FM)",
                                   "Fumadores pasivos (FP)",
                                   "No fumadores (NF)",
                                   "Fumadores que no inhalan (NI)"))
```



La distribución de las medias divide a los tipos de fumador en tres grupos, ordenados de menor a mayor capacidad pulmonar:

- **Fumadores del grupo 1**: Incluye los fumadores moderados (FM) y los fumadores intensivos (FI).
- **Fumadores del grupo 2**: Incluye los fumadores ligeros (FL), los fumadores pasivos (FP) y los fumadores que no inhalan (NI).
- **No fumadores**: Incluye únicamente los no fumadores (NF).

Llama la atención que los fumadores moderados presenten una media inferior a la de los fumadores intensivos. Habría que analizar otros factores, como la edad por grupo.

Sin ánimo de extendernos demasiado en el análisis, presentamos una tabla para ver las medias de edad de los grupos de fumadores en el anexo.

También llama la atención el amplio rango intercuartílico de los fumadores pasivos (FP) frente a los demás. Aunque son conjeturas, podría deberse a que hay mucha variedad en este grupo. Se entiende que no es lo mismo un fumador pasivo porque convive con un fumador intensivo o porque lo hace con un fumador ligero.

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?

Este ejercicio nos pide buscar el intervalo de confianza para una media poblacional μ con varianza σ desconocida. En primer lugar, recordemos que la variable `AE` es de naturaleza continua, por lo que permite el cálculo del intervalo de confianza.

El tamaño de las muestras es de 144 observaciones en el caso de las mujeres y de 109 en el caso de los hombres. Como ambas muestras tienen más de 30 observaciones, podemos asumir normalidad aplicando el Teorema del Límite Central [1].

Al no conocer la varianza poblacional, tendremos que aplicar el cálculo con la distribución *t-Student* con $n-1$ grados de libertad para encontrar el intervalo de confianza; siendo n el tamaño de la muestra.

Para este ejercicio y el siguiente, vamos a aprovechar la función que desarrollamos en la Actividad 2. Como en su momento ya ofrecimos la explicación del desarrollo de las fórmulas, para aligerar la lectura de esta actividad, dejaremos las fórmulas en el anexo, por si fueran de interés, pero las obviaremos aquí.

Aclarado esto, pasamos a formular nuestra función.

```r
# Función para el intervalo de confianza de una variable.
# Parámetros: x = Variable (muestra); NC = nivel de confianza (1-a)/100
IC <- function( x, NC ){
  n <- length(x)                          # Tamaño de la muestra
  if(n < 30) {
    print("Muestra menor de 30: realizar test de normalidad Shapiro-Wilk")
  } else {
    sErr <- sd(x)/sqrt(n)                 # Error estándar de la media
    quant <- (1-NC)/2                     # Cuantil en base al nivel de confianza
    z <- qt(quant, df=n-1,                # Valor crítico
          lower.tail=FALSE)
    merror <- z * sErr                    # Margen de error
    L <- round(mean(x)-merror, 2)         # Lower tail
    U <- round(mean(x)+merror, 2)         # Upper tail
    cat("El intervalo de confianza es de: [",
        L, ",", U, "]")                        # Usamos la fórmula cat+return para poder
    return(c(L, U))                       # almacenar el resultado en una variable
}}
```

Ahora vamos a aplicar la fórmula a nuestras muestras.

```r
fem <- smokers[smokers$genero == "F",]
male <- smokers[smokers$genero == "M",]

AE.f <- fem$AE
```

[1] Con esto asumimos que nuestra variable se *comportará* de forma similar a una normal, no quiere decir que *lo sea*.

```r
AE.m <- male$AE

# Aplicamos la función para AE en mujeres al 95% de confianza
IC.AEfem <- IC(AE.f, .95)
```

```
## El intervalo de confianza es de: [ 1.43 , 1.62 ]
```

```r
# Aplicamos la función para AE en hombres al 95% de confianza
IC.AEmale <- IC(AE.m, .95)
```

```
## El intervalo de confianza es de: [ 1.48 , 1.69 ]
```

Como ya habíamos observado en el análisis visual, la muestra para las mujeres muestra un mayor rango que la de los hombres, si bien comparten un intervalo muy similar. Da la impresión de que las poblaciones pueden tener una capacidad pulmonar similar.

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 de investigación que se plantea es: ¿La capacidad pulmonar reflejada en la variable `AE` de los hombres y de las mujeres es significativamente diferente? Por lo tanto, las hipótesis serían:

Hipótesis nula: La capacidad pulmonar media de las mujeres es igual a la de los hombres.

Hipótesis alternativa: La capacidad pulmonar media de las mujeres es diferente a la de los hombres.

Dicho de otro modo:

$$H_0 : \mu_{male} = \mu_{fem}$$
$$H_1 : \mu_{male} \neq \mu_{fem}$$

4.2 Contraste

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

Podemos afirmar que nos encontramos ante dos muestras independientes con varianza desconocida.

Tamaño de la muestra

En primer lugar, deberemos comprobar si los tamaños de las muestras son los adecuados, cosa que ya hemos hecho en el apartado anterior.

¿Varianzas iguales?

Hecho esto, nos preguntamos si las varianzas son iguales. Para responder a esta pregunta, tendremos que llevar a cabo un test de homocedasticidad [2].

```r
var.test(AE.f, AE.m)
```

[2] Este test ha sido incluido en la función que vamos a utilizar, pero lo vamos a ver por separado en este apartado.

```
##
##  F test to compare two variances
##
## data:  AE.f and AE.m
## F = 1.1609, num df = 143, denom df = 108, p-value = 0.4156
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8103257 1.6482865
## sample estimates:
## ratio of variances
##           1.160855
```

Como vemos, el valor 1 de varianza se encuentra en el intervalo de confianza con un 95% de confianza, por lo que podemos asumir que las varianzas son iguales.

Tipo de hipótesis

Por último, deberemos determinar si vamos a hacer un test bilateral o unilateral. Sabemos que esto viene dado por la formulación de la hipótesis alternativa. Si $H_1 : \mu_{male} \neq \mu_{fem}$, nos encontramos antes un test bilateral.

En resumen, tenemos que hacer un **contraste de hipótesis bilateral de dos muestras independientes sobre la media con varianzas desconocidas pero iguales**.

4.3 Cálculos

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

Como vamos a realizar otro contraste en el próximo ejercicio, elegimos desarrollar la función que escribimos para la Actividad 2. De nuevo, el desarrollo de las fórmulas se encuentra en el Anexo.

```r
# Función para el contraste de medias.
# Parámetros:  x.1 - Variable 1; x.2 - Variable 2;
# NC - nivel de confianza (1-a)/100 (default: 95%);
# var - varianza poblacional (default: NULL);
# norm - distribución normal (default: FALSE)
# alternative - opciones: "two.sided"(default), "less", "greater"
test.twomeans <- function(x.1, x.2, NC=.95,
                          var=NULL, norm=FALSE,
                          alternative="two.sided") {
  # Varianza conocida
  if(!is.null(var)) {
    print("Esta función no es apropiada. Se requiere un z-test")

  # Muestras pequeñas (menores de 30) no normales
  } else if(norm==FALSE) {
      n.1 <- length(x.1)     # Tamaño muestral 1
      n.2 <- length(x.2)     # Tamaño muestral 2
      if(n.1 < 30 | n.2 < 30) {
        print("Esta función no es apropiada. Se requiere un test no paramétrico")
      } else {
        mean.1 <- mean(x.1)     # Media muestral 1
        mean.2 <-  mean(x.2)    # Media muestral 2
        var.1 <- var(x.1)       # Varianza muestral 1
        var.2 <- var(x.2)       # Varianza muestral 2
        a <- 1-NC               # Nivel de significación
```

```r
        # TIPO DE VARIANZA
        # Varianzas iguales
        if(min(var.test(x.1,x.2)$conf) < 1 & max(var.test(x.1,x.2)$conf) > 1) {
          contraste <- c("independientes sobre \n", "la media con varianzas",
                         "desconocidas e iguales.\n\n ")
          gl <- n.1+n.2-2                  # Grados de libertad
          s <- sqrt(((n.1-1)*var.1+        # Desviación estándar
                       (n.2-1)*var.2)/
                      (n.1+n.2-2))
          sErr <- s*sqrt(1/n.1 + 1/n.2)    # Error estándar de la media

        # Varianzas diferentes
        } else if(min(var.test(x.1,x.2)$conf) > 1 | max(var.test(x.1,x.2)$conf) < 1) {
        contraste <- c("independientes sobre \n", "la media con varianzas",
                         "desconocidas y diferentes.\n\n")
        gl <- (var.1/n.1 + var.2/n.2)^2 /    # Grados de libertad
              ((var.1/n.1)^2 / (n.1-1) +
              (var.2/n.2)^2 / (n.2-1))
        sErr <- sqrt(var.1/n.1 + var.2/n.2) # Error estándar de la media
        }
      t <- (mean.1-mean.2)/sErr              # Estadístico de contraste

      # TIPO DE HIPÓTESIS
      # Cálculo de valor crítico + p-valor
      if(alternative=="two.sided") {
        vc <- c(qt(a/2, df=gl, lower.tail=TRUE),
              qt(a/2, df=gl, lower.tail=FALSE))
        pv <-  2*pt(abs(t), df=gl, lower.tail=FALSE)
      } else if (alternative=="less") {
        vc <- qt(a, df=gl, lower.tail=TRUE)
        pv <-  pt(t, df=gl, lower.tail=TRUE)
      } else if (alternative=="greater") {
      vc <- qt(a, df=gl, lower.tail=FALSE)
      pv <-  pt(t,df=gl, lower.tail=FALSE)
        }}
      cat("Contraste de dos muestras", contraste,
        "Estadístico de contraste:", t,
        "\n Valor crítico:", vc,
        "\n p-valor:", pv)
    return(c(t, vc, pv))
  } else {
    print("Esta función no es apropiada")
}}
```

Dicho esto, pasamos a aplicar nuestra función.

```r
test.AEg <- test.twomeans(AE.f, AE.m, .95)
```

```
## Contraste de dos muestras independientes sobre
##  la media con varianzas desconocidas e iguales.
##
##   Estadístico de contraste: -0.8616831
##  Valor crítico: -1.96946 1.96946
##  p-valor: 0.3896844
```

4.4 Interpretación

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

Rescatamos la pregunta planteada inicialmente: ¿La capacidad pulmonar reflejada en la variable `AE` de los hombres y de las mujeres es significativamente diferente?

Dado que el p-valor (0.39) > nivel de confianza α (0.05), no rechazaremos la hipótesis nula. Lo mismo ocurre con los valores críticos: dado que t (-0.86) es mayor que el valor crítico de la cola izquierda (-1.97) y menor que el valor crítico de la cola derecha (1.97), se encuentra dentro del intervalo de confianza.

Concluimos que la capacidad pulmonar de los hombres y de las mujeres no es significativamente diferente con una confianza del 95%.

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.

5.1 Hipótesis

Escribir la hipótesis nula y alternativa.

La pregunta de investigación que se plantea es: ¿La capacidad pulmonar registrada en la variable `AE` de los fumadores es inferior a la de los no fumadores? Por lo tanto, las hipótesis serían:

Hipótesis nula: La capacidad pulmonar media de los fumadores es igual a la de los no fumadores.

Hipótesis alternativa: La capacidad pulmonar media de los fumadores es inferior a la de los no fumadores.

Dicho de otro modo:

$$H_0 : \mu_{smokers} = \mu_{non\ smokers}$$
$$H_1 : \mu_{smokers} < \mu_{non\ smokers}$$

5.2 Contraste

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

Podemos afirmar que nos encontramos ante dos muestras independientes con varianza desconocida.

Tamaño de la muestra

En primer lugar, deberemos comprobar si los tamaños de las muestras son los adecuados. Para ello, pasamos a separar el dataframe en dos muestras, fumadores y no fumadores, de la forma que se indica en el enunciado.

```r
# Separamos las muestras por fumador y no fumador
nonsmoke <- droplevels(smokers[smokers$tipo == "FP" |
                        smokers$tipo == "NF",])
smoke <- droplevels(smokers[smokers$tipo == "FI" |
                    smokers$tipo == "FL" |
                    smokers$tipo == "FM" |
                    smokers$tipo == "NI",])
```

```r
AE.smoke <- smoke$AE
AE.nonsmoke <- nonsmoke$AE
```

La muestra de los fumadores es de 163, mientras que la de no fumadores es de 90, por lo que podemos asumir que tienen una distribución similar a una normal.

¿Varianzas iguales?

Pero, ¿estas varianzas son iguales? Para responder a esta pregunta, tendremos que llevar a cabo un test de homocedasticidad [3].

```r
var.test(smoke$AE, nonsmoke$AE)
```

```
##
##  F test to compare two variances
##
## data:  smoke$AE and nonsmoke$AE
## F = 0.79903, 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:
##  0.5477417 1.1426530
## sample estimates:
## ratio of variances
##          0.7990301
```

Como vemos, el valor 1 de varianza se encuentra en el intervalo de confianza con un 95% de confianza, por lo que podemos asumir que las varianzas son iguales.

Tipo de hipótesis

Por último, deberemos determinar si vamos a hacer un test bilateral o unilateral. Sabemos que esto viene dado por la formulación de la hipótesis alternativa. Si $H_1 : \mu_{smokers} < \mu_{non\ smokers}$, nos encontramos antes un test unilateral.

En resumen, tenemos que hacer un contraste de hipótesis unilateral de dos muestras independientes sobre la media con varianzas desconocidas pero iguales.

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.

Este apartado se ha resuelto en el punto anterior.

```r
kable(table(smoke$tipo),
      col.names = c("Fumadores", "#"),
      caption = "Niveles incluidos en la muestra de fumadores")
```

Cuadro 6: Niveles incluidos en la muestra de fumadores

Fumadores	#
FI	41
FL	41
FM	39
NI	42

[3] Este test ha sido incluido en la función que vamos a utilizar, pero lo vamos a ver por separado en este apartado.

```r
kable(table(nonsmoke$tipo),
      col.names = c("No fumadores", "#"),
      caption = "Niveles incluidos en la muestra de no fumadores")
```

Cuadro 7: Niveles incluidos en la muestra de no fumadores

No fumadores	#
FP	40
NF	50

5.4 Cálculos

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

Aplicamos la fórmula desarrollada en el ejercicio anterior.

```r
test.AEs <- test.twomeans(AE.smoke, AE.nonsmoke, .95, alternative='less')
```

```
## Contraste de dos muestras independientes sobre
##  la media con varianzas desconocidas e iguales.
##
##   Estadístico de contraste: -6.327631
##  Valor crítico: -1.650947
##  p-valor: 5.680665e-10
```

5.5 Interpretación

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

Rescatamos la pregunta planteada inicialmente: ¿La capacidad pulmonar registrada en la variable `AE` de los fumadores es inferior a la de los no fumadores?

Dado que el p-valor $(5.68 \times 10^{-10}) < \alpha$ (0.05), rechazaremos la hipótesis nula. Lo mismo ocurre con los valores críticos. Dado que t (-6.33) es menor que el valor crítico (-1.65), no se encuentra dentro del intervalo de confianza.

Es decir, la capacidad pulmonar de los fumadores es inferior a la de los no fumadores con una 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 `genero`). 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`.

Aplicaremos aquí un modelo de regresión múltiple, cuya ecuación de la recta sería:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_{tipo}x_{tipo} + \hat{\beta}_{edad}x_{edad} + \hat{\beta}_{genero}x_{genero} + e$$

Siendo:

$\hat{y}$, la variable explicada (`AE`);

β_0, la ordenada en el origen;

β_i, la pendiente de la recta para el valor x_i de la variable explicativa X_i; y
e, el vector de los residuos.

A continuación pasamos a realizar el cálculo en R del modelo de regresión lineal.

```r
lm_a <- lm(AE~tipo+edad+genero)
summary(lm_a)
```

```
##
## Call:
## lm(formula = AE ~ tipo + edad + genero)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05912 -0.25196  0.00082  0.23074  1.03739
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.743031   0.128802  21.296  < 2e-16 ***
## tipoFL       0.337122   0.080854   4.170 4.24e-05 ***
## tipoFM       0.045123   0.082137   0.549    0.583
## tipoFP       0.393302   0.081473   4.827 2.44e-06 ***
## tipoNF       0.780125   0.077007  10.131  < 2e-16 ***
## tipoNI       0.421557   0.080263   5.252 3.26e-07 ***
## edad        -0.030964   0.002276 -13.605  < 2e-16 ***
## generoM     -0.001756   0.047035  -0.037    0.970
## ---
## 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:  0.5829, Adjusted R-squared:  0.5709
## F-statistic:  48.9 on 7 and 245 DF,  p-value: < 2.2e-16
```

6.2 Interpretación

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

Antes de hacer una interpretación del modelo, queremos analizar si existen problemas de multicolinealidad en nuestro modelo. Para ello, proponemos examinar el factor de inflación de la varianza (VIF). Recordemos que toma valores del 1 al infinito. Un VIF a partir de 5 puede indicar que nuestro modelo puede presentar errores de estimación. A partir de 10 será problemático.

```
car::vif(lm_a)
```

```
##            GVIF Df GVIF^(1/(2*Df))
## tipo   1.034090  5        1.003358
## edad   1.023736  1        1.011798
## genero 1.027446  1        1.013630
```

Como podemos observar, no existen problemas de multicolinealidad en nuestro modelo.

Dicho esto, podemos ver que las variables `edad` y `tipo` son significativas, si bien se señala el nivel `FM` como no significativo. Por su parte, la variable `genero` no es significativa para nuestro modelo. Es un resultado que va acorde al contraste realizado entre géneros.

Eliminaremos esta última variable de nuestro modelo. Podríamos comprobar que nuestra decisión es la correcta aplicando el criterio de información de Akaike (AIC), pero con el contraste realizado anteriormente este paso sería redundante.

```
# Descomentar si se quiere aplicar el AIC
# step(object = lm_a, direction = "backward", trace = 1)
```

Nuestro modelo final se llamará `lm_smokers`.

```
lm_smokers <- lm(AE~tipo+edad)

summary(lm_smokers)
```

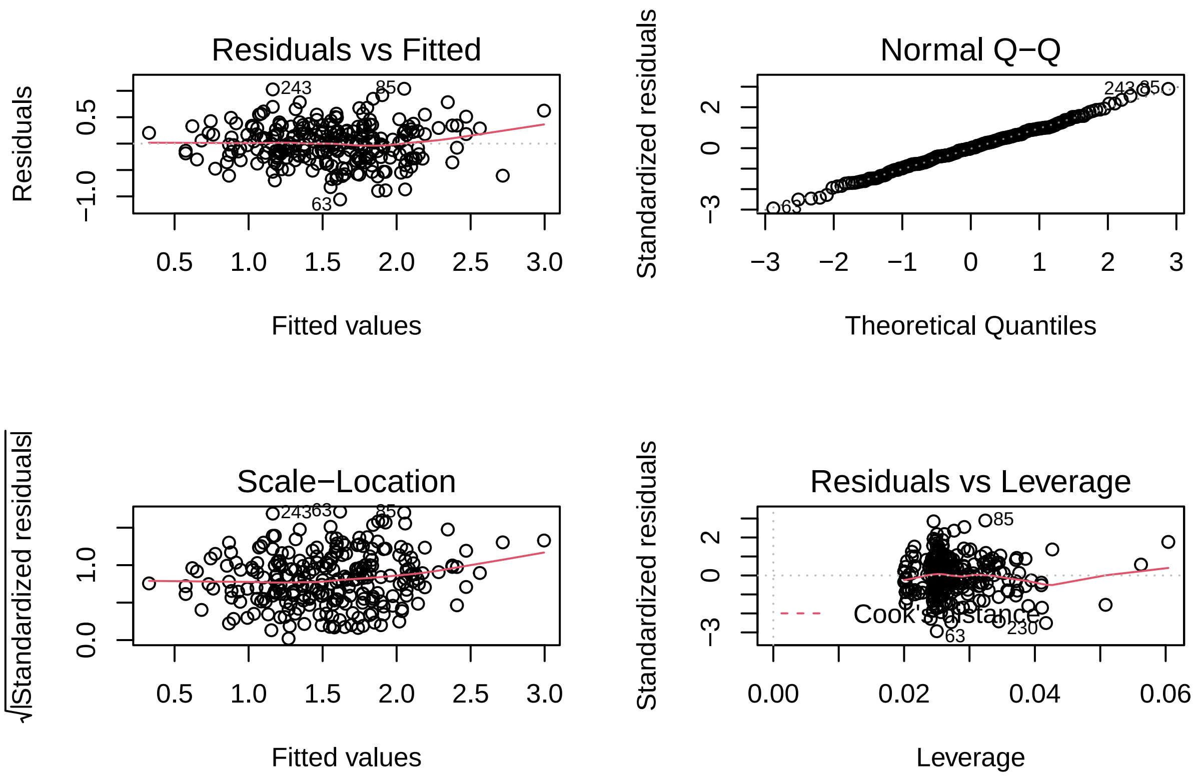
```
##
## Call:
## lm(formula = AE ~ tipo + edad)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05815 -0.25127 -0.00015  0.23153  1.03847
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.741916   0.125037  21.929  < 2e-16 ***
## tipoFL       0.337294   0.080559   4.187 3.94e-05 ***
## tipoFM       0.045058   0.081952   0.550    0.583
## tipoFP       0.393067   0.081064   4.849 2.20e-06 ***
## tipoNF       0.780114   0.076850  10.151  < 2e-16 ***
## tipoNI       0.421489   0.080079   5.263 3.08e-07 ***
## edad        -0.030956   0.002261 -13.688  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3647 on 246 degrees of freedom
## Multiple R-squared:  0.5828, Adjusted R-squared:  0.5727
## F-statistic: 57.29 on 6 and 246 DF,  p-value: < 2.2e-16
```

6.3 Bondad de ajuste

Evaluad la calidad del modelo.

Como podemos ver, el *p*-valor (<2e-16) es muy inferior a al nivel de significación, por lo que el modelo es significativo para explicar la capacidad pulmonar. Sin embargo, nos encontramos con un *R-squared* [4], de 0.58. Según este coeficiente de determinación, el modelo de regresión explicaría el 58.29% de la variabilidad total de las observaciones. Se trata de una cifra no muy alta, así pues, podemos considerar que **el modelo tiene un ajuste mejorable**.

```
par(mfrow = c(2, 2))
plot(lm_smokers)
```



Vemos que los residuos se distribuyen mayoritariamente de forma aleatoria, mientras que la mayoría de los datos siguien una distribución normal, lo que supone un buen resultado para nuestro modelo.

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.

Para hacer esto, en primer lugar crearemos una función que realice una predición por tipo de fumador y el rango de edad establecido.

Hecho esto, crearemos un dataframe con los resultados de la predicción por edades en las filas y los tipos de fumador en las columnas. Para hacer la gráfica, se han recurrido a fuentes externas.

```
# Función para la predicción por tipo
pred.tipo <- function(smoke.type) {
                    prediction <- predict(lm_smokers,
```

[4] La descripción más detallada sobre este coeficiente y su fórmula se encuentran en el Anexo

```r
                                        newdata = data.frame(tipo=smoke.type,
                                                             edad=30:80))
                    return(prediction)
}

# Creamos un dataframe vacío
df.pred <- data.frame()

# Hacemos el cálculo de predicción para cada tipo y lo almacenamos en variables
FI <- pred.tipo('FI')
FL <- pred.tipo('FL')
FM <- pred.tipo('FM')
FP <- pred.tipo('FP')
NF <- pred.tipo('NF')
NI <- pred.tipo('NI')

# Creamos el dataset con las predicciones
df.pred <- data.frame(FI, FL, FM, FP, NF, NI)

# Mostramos las primeras filas
head(df.pred)
```

```
##         FI       FL       FM       FP       NF       NI
## 1 1.813246 2.150539 1.858304 2.206312 2.593359 2.234735
## 2 1.782290 2.119584 1.827348 2.175357 2.562403 2.203779
## 3 1.751334 2.088628 1.796393 2.144401 2.531448 2.172824
## 4 1.720379 2.057672 1.765437 2.113445 2.500492 2.141868
## 5 1.689423 2.026717 1.734481 2.082490 2.469536 2.110912
## 6 1.658467 1.995761 1.703526 2.051534 2.438581 2.079957
```

```r
# Presentamos un resumen estadísitico de los resultados
kable(summary(df.pred),
      caption = c("Resumen estadístico de los resultados de la predicción",
                  "de la capacidad pulmonar por tipo de fumador"))
```

Cuadro 8: Resumen estadístico de los resultados de la predicción
Table: de la capacidad pulmonar por tipo de fumador

	FI	FL	FM	FP	NF	NI
	Min. :0.2655	Min. :0.6028	Min. :0.3105	Min. :0.6585	Min. :1.046	Min. :0.687
	1st Qu.:0.6524	1st Qu.:0.9897	1st Qu.:0.6975	1st Qu.:1.0455	1st Qu.:1.433	1st Qu.:1.074
	Median :1.0394	Median :1.3766	Median :1.0844	Median :1.4324	Median :1.819	Median :1.461
	Mean :1.0394	Mean :1.3766	Mean :1.0844	Mean :1.4324	Mean :1.819	Mean :1.461
	3rd Qu.:1.4263	3rd Qu.:1.7636	3rd Qu.:1.4714	3rd Qu.:1.8194	3rd Qu.:2.206	3rd Qu.:1.848
	Max. :1.8132	Max. :2.1505	Max. :1.8583	Max. :2.2063	Max. :2.593	Max. :2.235

```r
## Creamos una columna para la edad
df.pred$age = c(30:80)

# Unimos las columnas 'tipo' en una sola
final_data <- melt(df.pred, id='age')
```

```r
head(final_data)
```

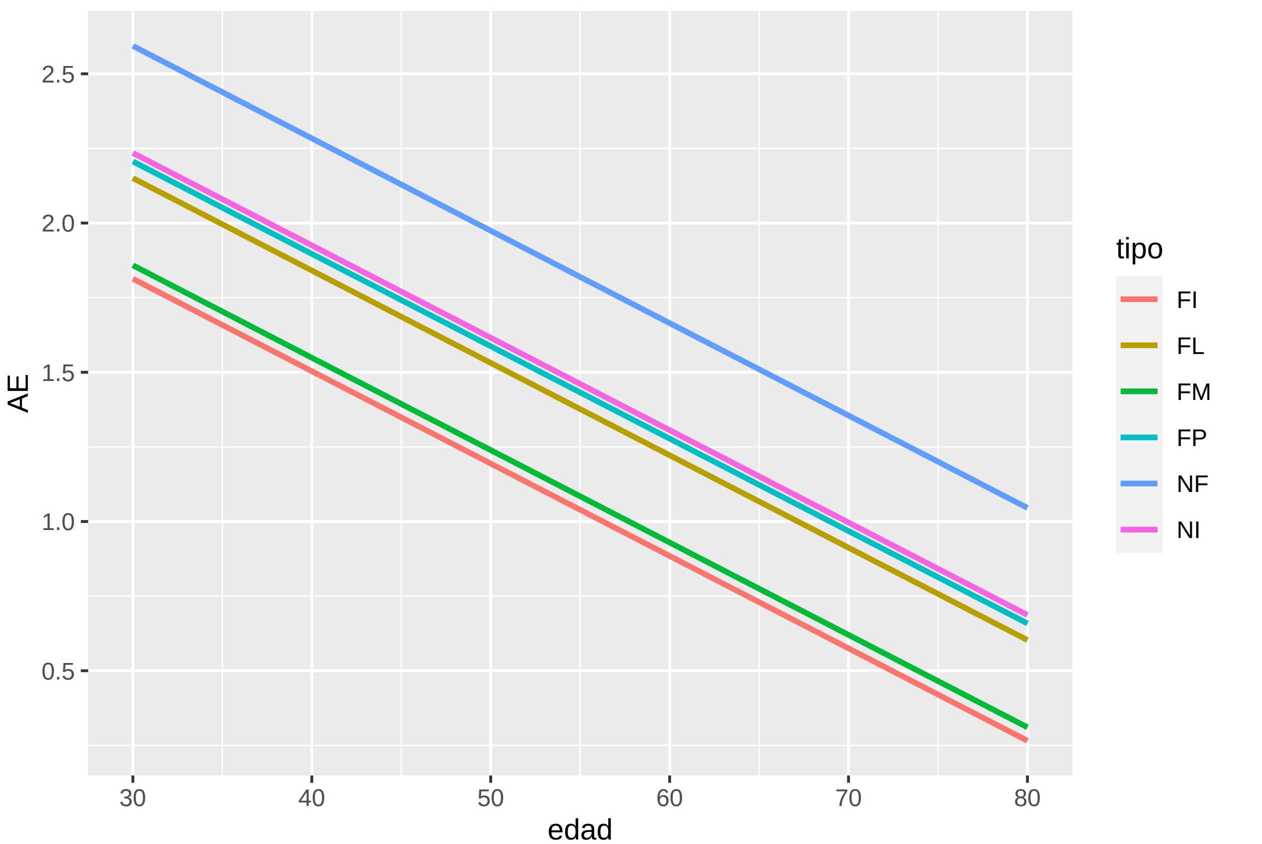
```
##   age variable    value
## 1  30       FI 1.813246
## 2  31       FI 1.782290
## 3  32       FI 1.751334
## 4  33       FI 1.720379
## 5  34       FI 1.689423
## 6  35       FI 1.658467
```

```r
tail(final_data)
```

```
##     age variable     value
## 301  75       NI 0.8417298
## 302  76       NI 0.8107742
## 303  77       NI 0.7798185
## 304  78       NI 0.7488628
## 305  79       NI 0.7179071
## 306  80       NI 0.6869515
```

```r
# Añadimos los  nombres de las correspondientes
names(final_data) <- c('edad', 'tipo', 'AE')

ggplot() + geom_line(data = final_data,
                     aes(x = edad, y = AE,
                         color = tipo, group = tipo), size = 1)
```



Vemos que en la predicción se conservan los tres grupos detectados al calcular las medias: FI/FM, FL/FM/NI, NF.

7 ANOVA unifactorial

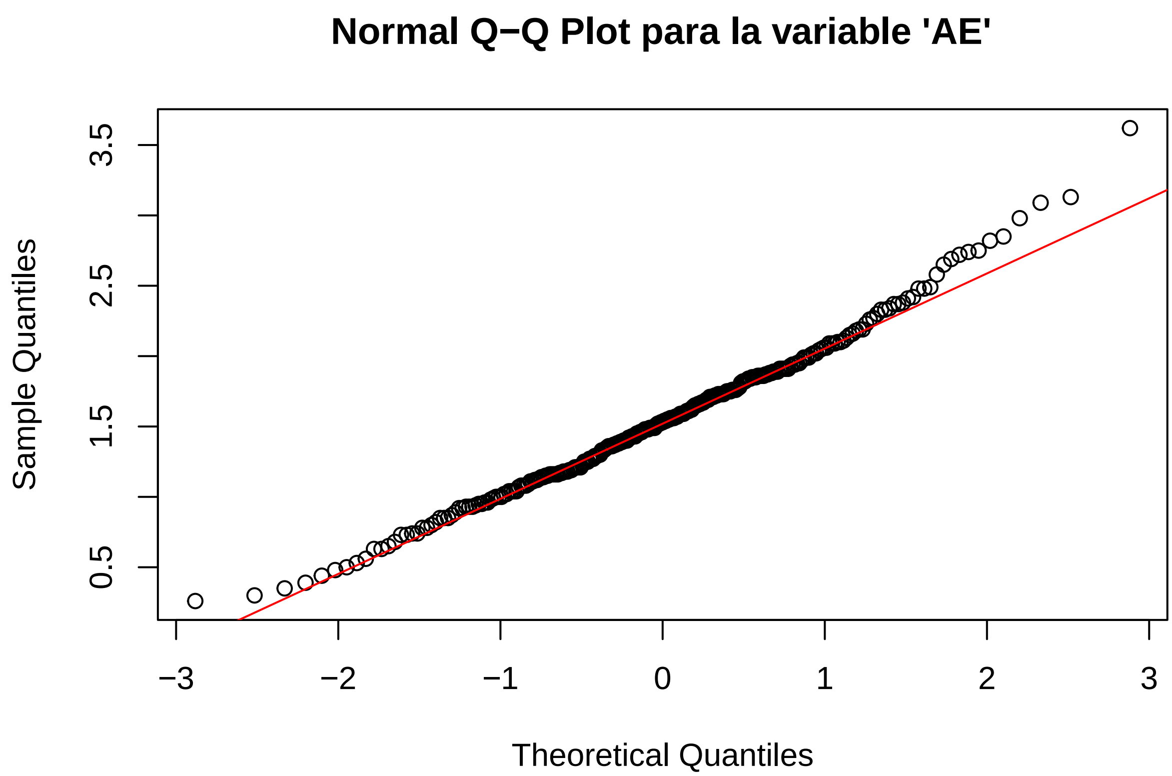
A continuación se realizará un análisis de varianza, donde se desea comparar la capacidad pulmonar entre los seis tipos de fumadores/no fumadores clasificados previamente. El análisis de varianza consiste en evaluar si la variabilidad de una variable dependiente puede explicarse a partir de una o varias 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. *Nota*: podéis usar el gráfico normal Q-Q y el test Shapiro-Wilk para evaluar la normalidad de la muestra.

```
qqnorm(AE, main="Normal Q-Q Plot para la variable 'AE'")
qqline(AE, col='red')
```




```
shapiro.test(AE)
```

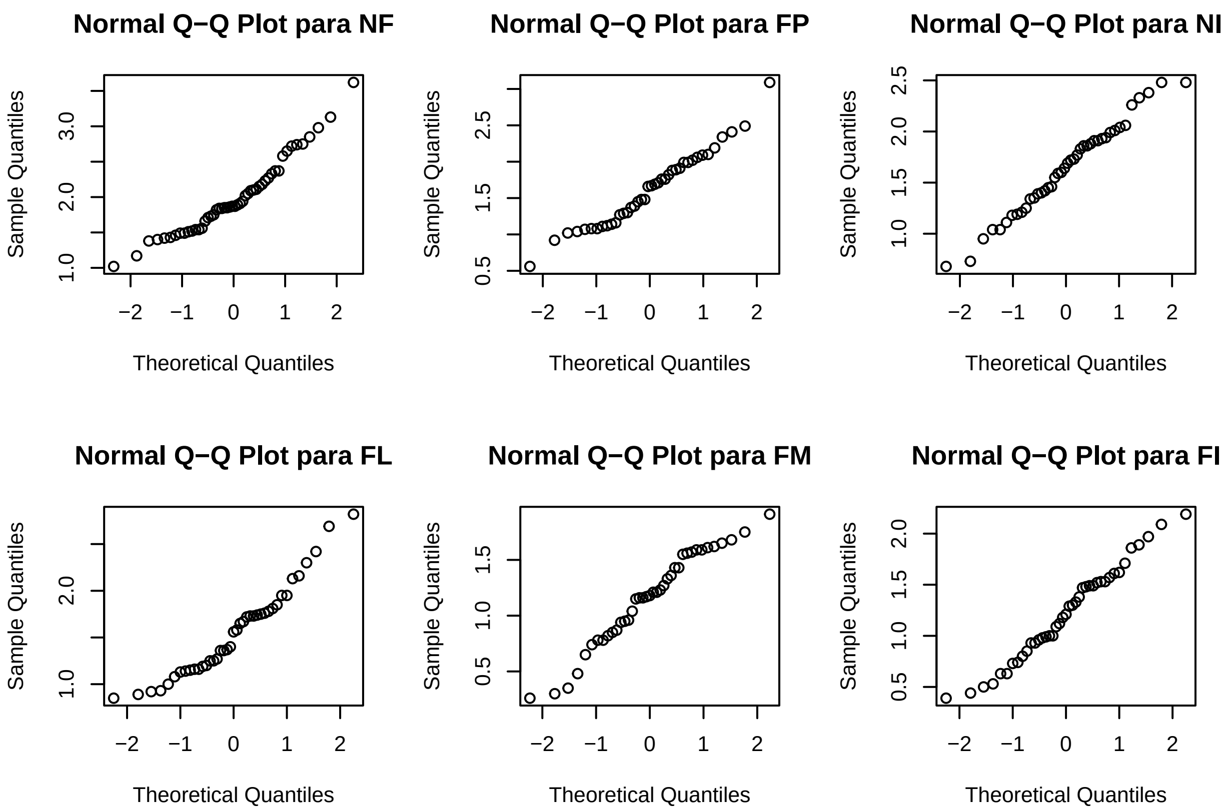
```
##
##  Shapiro-Wilk normality test
##
## data:  AE
## W = 0.98875, p-value = 0.04599
```

La muestra general del estudio presenta una distribución normal con un 90% de confianza. Como vemos en el gráfico normal Q-Q, los valores tienden a seguir una normal, con excepción de los valores extremos, especialmente en la cola derecha, lo que hace que si subimos a un nivel de confianza del 0.05 al analizar el Test de Shapiro-Wilk, afirmemos que no hay normalidad.

Recordemos, no obstante, que se recomienda hacer estos test sobre los *residuos* no sobre las observaciones, por lo que deberíamos hacerlo después de realizar el test ANOVA: "Cabe señalar que debemos aplicar la función shapiro.test a los residuos, y observemos que será equivocado aplicarlo sobre las observaciones $y_i j$", leemos en la página 22 de *Análisis de la varianza (ANOVA)*.

Analizamos a continuación las muestras por tipo de fumador.

```
par(mfrow = c(2, 3))
qqnorm(smokers[smokers$tipo == "NF",]$AE, main="Normal Q-Q Plot para NF")
qqnorm(smokers[smokers$tipo == "FP",]$AE, main="Normal Q-Q Plot para FP")
qqnorm(smokers[smokers$tipo == "NI",]$AE, main="Normal Q-Q Plot para NI")
qqnorm(smokers[smokers$tipo == "FL",]$AE, main="Normal Q-Q Plot para FL")
qqnorm(smokers[smokers$tipo == "FM",]$AE, main="Normal Q-Q Plot para FM")
qqnorm(smokers[smokers$tipo == "FI",]$AE, main="Normal Q-Q Plot para FI")
```




```
shapiro.test(smokers[smokers$tipo == "NF",]$AE)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  smokers[smokers$tipo == "NF", ]$AE
## W = 0.95062, p-value = 0.03618
```

```
shapiro.test(smokers[smokers$tipo == "FP",]$AE)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  smokers[smokers$tipo == "FP", ]$AE
## W = 0.97181, p-value = 0.4101
```

```
shapiro.test(smokers[smokers$tipo == "NI",]$AE)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  smokers[smokers$tipo == "NI", ]$AE
## W = 0.98272, p-value = 0.7656
```

```
shapiro.test(smokers[smokers$tipo == "FL",]$AE)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  smokers[smokers$tipo == "FL", ]$AE
## W = 0.94334, p-value = 0.04099
```

```
shapiro.test(smokers[smokers$tipo == "FM",]$AE)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  smokers[smokers$tipo == "FM", ]$AE
## W = 0.96332, p-value = 0.2297
```

```
shapiro.test(smokers[smokers$tipo == "FI",]$AE)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  smokers[smokers$tipo == "FI", ]$AE
## W = 0.97789, p-value = 0.5962
```

Algunas de las muestras tienen un p-valor por debajo del 0.05, por lo que no siguen una normal. Se trata de los grupos `NF`, `FL`.

Los gráficos Q-Q muestran que todos los niveles siguen una línea recta con ligeras desviaciones, lo que indica una distribución normal salvo en los valores extremos.

```
kable(table(tipo), caption ="Niveles de la variable tipo")
```

Cuadro 9: Niveles de la variable tipo

tipo	Freq
FI	41
FL	41
FM	39
FP	40
NF	50
NI	42

Si tuviéramos muestras más pequeñas, deberíamos tratar de forma diferente los grupos que no han mostrado normalidad, sin embargo, todos los niveles superan las 30 observaciones y el tamaño de las muestras es bastante similar. Con esto, aplicando el Teorema del Límite Central, podemos asumir normalidad de las distribuciones y aplicar los siguientes tests.

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.

Como vimos con el boxplot en el ejercicio 2.3, la mayoría de los niveles tienen un tamaño de caja similar, lo que indicaría homoscedasticidad; si bien es cierto que el tamaño de caja de los fumadores pasivos (FP) es mayor que las del resto.

Comprobaremos la igualdad de varianzas aplicando un test de Bartlett. Descartamos el test no paramétrico de Levene, ya que hemos considerado nuestras muestras muy similares a una normal.

Recordemos que la hipótesis nula de Bartlett es que existe igualdad de varianzas, mientras que la alternativa asume que no hay igualdad de varianzas en, por lo menos, dos de los niveles.

Para rechazar la hipótesis nula necesitaremos un p-valor menor a 0.05 para un nivel de confianza del 95%.

```
var.tipo <- bartlett.test(AE~tipo)
var.tipo
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  AE by tipo
## Bartlett's K-squared = 3.2386, df = 5, p-value = 0.6633
```

Como vemos, el p-valor es de 0.6633, significativamente mayor a 0.05, por lo que no rechazamos la hipótesis nula y asumimos homoscedasticidad.

Con estos resultados, podemos realizar el test de ANOVA. Nótese que no hemos hablado de el último supuesto para hacer el test: la incorrelación, ya que no se nos ha pedido. Se supone que las muestras son independientes, si bien existen muchas sospechas de que hay varias correlacionadas entre sí.

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.

$H_0 : \mu_{NF} = \mu_{FP} = \mu_{NI} = \mu_{FL} = \mu_{FM} = \mu_{FI}$

$H_1 : \mu_i \neq \mu_j$ para algún $i \neq j$

Dicho de otro modo:

Hipótesis nula: Las medias poblacionales de los niveles de la variable `tipo` son iguales.

Hipótesis alternativa: Las medias poblacionales de al menos dos de los niveles de la variable `tipo` son significativamente distintas.

7.4 Cálculo ANOVA

Podéis usar la función `aov`.

Como vemos, el modelo ANOVA evalúa si existen diferencias significativas entre las medias poblacionales de los grupos estudiados.

Si lo aplicamos a una variable independiente X, se formula de la siguiente manera (López-Roldán y Fachelli, 2015, p.28):

$$Y_{ij} = \mu + \alpha_j + e_{ij}$$

Donde:

Y_{ij} son los valores de la variable dependiente para el indivuo i del grupo j

μ es la media poblacional

α_j es la media del efecto de cada valor de la variable X que viene definida por:

$$\alpha_j = \mu_j - \mu$$

Siendo μ_j la media de cada grupo j

Por último, e_{ij} es la parte residual del modelo. Es decir, lo que X_j no puede explicar.

Explicado esto, las hipótesis anteriores se puede reformular:

$H_0 : \alpha_{NF} = \alpha_{FP} = \alpha_{NI} = \alpha_{FL} = \alpha_{FM} = \alpha_{FI} = 0$

$H_1 : \alpha_i \neq \alpha_j$ para algún $i \neq j$

Dicho de otro modo:

Hipótesis nula: La variable `tipo` no es significativa para el cálculo de `AE`, ya que no existe diferencia entre sus niveles.

Hipótesis alternativa: La variable `tipo` es significativa para el cálculo de `AE`, ya que existe diferencia entre sus niveles.

Ahora que ya sabemos lo que vamos a calcular, procedemos a aplicar la función `aov()` en R.

```
results<-aov(AE~tipo)
results
```

```
## Call:
##    aov(formula = AE ~ tipo)
##
## Terms:
##                      tipo Residuals
## Sum of Squares  20.79972  57.65638
## Deg. of Freedom        5       247
##
## Residual standard error: 0.4831425
## Estimated effects may be unbalanced
```

```r
# Obtenemos el detalle de los resultados
anov.res<-anova(results)
anov.res
```

```
## Analysis of Variance Table
##
## Response: AE
##            Df Sum Sq Mean Sq F value    Pr(>F)
## tipo        5 20.800  4.1599  17.821 4.449e-15 ***
## Residuals 247 57.656  0.2334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

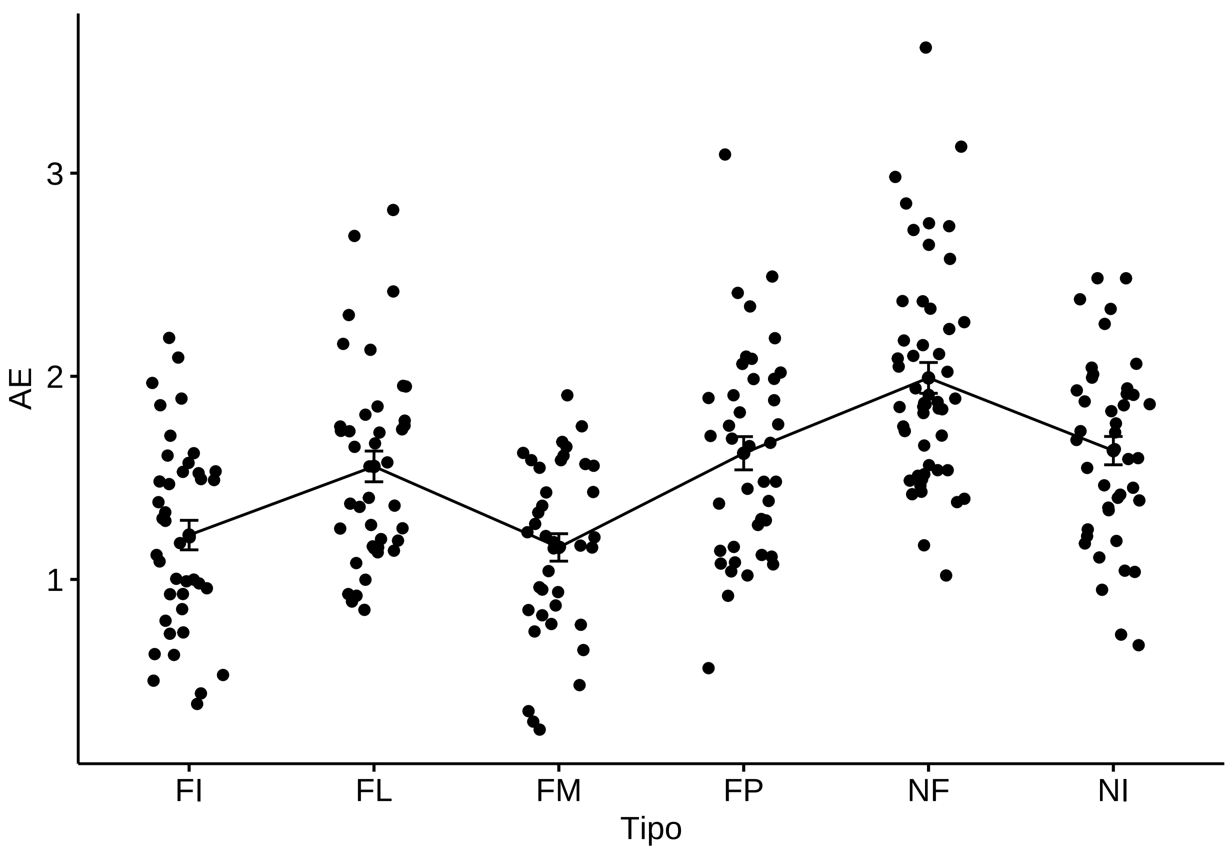
7.5 Interpretación

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

En primer lugar, vamos a fijarnos en el resultado del p-valor, de 4.449e-15. Como es notablemente menor a nuestro nivel de significación, aceptamos la hipótesis alternativa y concluimos que la variable `tipo` es significativa. Es decir, existe diferencia de medias entre al menos dos de los niveles.

Como pudimos comprobar, en el boxplot del apartado 2.3 se apreciaba diferencia de medias. Si bien, seguimos preocupados por la posibildad de incorrelación entre las variables.

```r
ggline(smokers, x = 'tipo', y = 'AE',
       add = c("mean_se", "jitter"),
       ylab = "AE", xlab = "Tipo")
```



También podemos visualizar estos datos usando `ggline`, lo que nos permite tener una mejor idea del volumen de datos y distrubición de cada nivel.

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.

Recordamos la tabla ANOVA y sus fórmulas, que se pueden encontrar en la página 12 de los apuntes.

Fuentes	**gl**	**SS**	**MS**	**F**
Tratamientos	$a-1$	SSA	$MSA = SSA/(a-1)$	$F = MSA/MSE$
Error	$N-a$	SSE	$MSE = SSE/(N-a)$	

Donde:

gl = Grados de libertad

a = Número de niveles

N = Tamaño de la muestra

SS = Suma de los cuadrados

SSA = Suma de cuadrados de los tratamientos (equivalente a SSB)

SSE = suma de cuadrados del error (equivalente a SSW)

MS = Cuadrados medios

MSA = Cuadrados medios de los tratamientos

MSE = Cuadrados medios de los errores

F = Valor F

La suma total de los cuadrados (SST) se calculará sumando SSA y SSE.

$$SST = SSA + SSE$$

Aclarado esto, pasamos a extraer los datos del test para luego aplicar las fórmulas. Hemos querido calcular manualmente los MS también, para tener un paso adicional manual y comprobar que nuestros cálculos son correctos, si bien, como veremos, se pueden obtener directamente de la tabla:

```r
# Recordamos los resultados:
anov.res
```

```
## Analysis of Variance Table
##
## Response: AE
##            Df Sum Sq Mean Sq F value    Pr(>F)
## tipo        5 20.800  4.1599  17.821 4.449e-15 ***
## Residuals 247 57.656  0.2334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Valores de la tabla
anov.res$"Df"      # gl
```

```
## [1]   5 247
```

```r
anov.res$"Sum Sq"  # SS
```

```
## [1] 20.79972 57.65638
```

```r
anov.res$"Mean Sq" # MS
```

```
## [1] 4.1599432 0.2334267
```

```r
anov.res$"F value" # F-valor
```

```
## [1] 17.8212      NA
```

```r
anov.res$"Pr(>F)" # p-valor
```

```
## [1] 4.448806e-15           NA
```

```r
# Suma de los cuadrados
SSB <- anov.res$"Sum Sq"[1]  #SSA
SSW <- anov.res$"Sum Sq"[2]  #SSE

# Suma total de los cuadrados
SST <- SSB + SSW

# Grados de libertad
df.fact <- anov.res$"Df"[1] # a-1
df.err <- anov.res$"Df"[2] #  N-a

# Nivel de significación
alpha <- 0.05

# Cuadrados medios
MSA = SSB/df.fact # Se puede extraer directamente con anov.res$"Mean Sq"[1]
MSE = SSW/df.err  # Se puede extraer directamente con anov.res$"Mean Sq"[2]

# Valor F
F.val = MSA/MSE
F.val
```

```
## [1] 17.8212
```

```r
# Valor P
p.val <- pf(F.val, df.fact, df.err, lower.tail = FALSE)
p.val
```

```
## [1] 4.448806e-15
```

```r
# Valor crítico
val.crit <- qf(1 - alpha, df.fact, df.err)
val.crit
```

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

Los resultados del p-valor, el valor crítico y el F-valor indican que:

p-valor $4.4 \times 10^{-15} < 0.05$ alpha

F-valor 17.821 > 2.251 valor crítico

En ambos casos se rechaza la hipótesis nula en favor de la alternativa, es decir, la variable `tipo` es significativa para explicar la variación de la variable `AE`. Sin embargo, hay matices que se pueden señalar analizando las sumas de los cuadrados.

La SSA (o SSB) es de 20.8, frente una SSE (o SSW), de 57.66. La suma total (SST) es de 78.46. Es decir, de una variabilidad de 78.46, 20.8 se puede explicar por la variable `tipo`, mientras que 57.66 es la variación que no se logra explicar con el modelo.

7.7 Fuerza de la relación

Calcular la fuerza de la relación e interpretar el resultado.

La fuerza de la relación se calcularía:

$$\eta^2 = SSA/SST \approx 0.265$$

Y representa el porcentaje de la variación total de la capacidad pulmonar que se explica conociendo el tipo de fumador.

Con esto podemos confirmar que la variable `tipo` es significativa, pero solo explica el 26.51% de la variación total, por lo que haría falta estudiar otros factores para terminar de comprender la tendencia en la capacidad pulmonar.

Se trata de un resultado dentro de los parámetros normales, teniendo en cuenta que se ha tratado una sola variable explicativa. En estos casos, "difícilmente se superan los valores de 0,5 o 0,6, es decir, de poco más del 50% de variabilidad explicada" (López-Roldán y Fachelli, 2015, p.37).

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.

Tal como se pide en el enunciado, aplicamos el Test Pairwise sin ningún tipo de corrección.

```
pairwise.t.test(AE, tipo, p.adj=c("none"))
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  AE and tipo
##
##    FI      FL      FM      FP      NF
## FL 0.00173 -       -       -       -
## FM 0.57385 0.00027 -       -       -
## FP 0.00022 0.54606 2.8e-05 -       -
## NF 6.0e-13 2.7e-05 2.7e-14 0.00036 -
## NI 0.00011 0.46451 1.4e-05 0.90466 0.00048
##
## P value adjustment method: none
```

Estos resultados no hacen más que corroborar lo que venimos diciendo: existen tres grupos diferenciados dentro de los seis tipos de fumadores. Recordemos que si el p-valor es mayor a 0.05 (para un 95% de confianza) no existirá diferencia entre los niveles. No existen diferencias reseñables entre los niveles de los siguientes grupos:

- FL, FP y NI;
- FM y FI;
- NF.

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.

Con esta corrección debemos ser cautos ya que, ante la probabilidad de calcular un resultado significativo por azar, tiende a ser muy conservadora y existe un riesgo de falsos negativos (*Análisis de la varianza (ANOVA)*, p.17).

```
LSD.test(results,"tipo",group=T,p.adj="bonferroni",console=T)
```

```
##
## Study: results ~ "tipo"
##
## LSD t Test for AE
## P value adjustment method: bonferroni
##
## Mean Square Error:  0.2334267
##
## tipo,  means and individual ( 95 %) CI
##
##          AE       std  r      LCL      UCL  Min  Max
## FI 1.218293 0.4648650 41 1.069677 1.366908 0.39 2.19
## FL 1.556341 0.4843850 41 1.407726 1.704957 0.85 2.82
## FM 1.157436 0.4217240 39 1.005057 1.309815 0.26 1.91
## FP 1.621250 0.5176336 40 1.470788 1.771712 0.56 3.09
## NF 1.992200 0.5356678 50 1.857623 2.126777 1.02 3.62
## NI 1.634048 0.4515289 42 1.487212 1.780883 0.68 2.48
##
## Alpha: 0.05 ; DF Error: 247
## Critical Value of t: 2.964024
##
## Groups according to probability of means differences and alpha level( 0.05 )
##
## Treatments with the same letter are not significantly different.
##
##          AE groups
## NF 1.992200      a
## NI 1.634048      b
## FP 1.621250      b
## FL 1.556341      b
## FI 1.218293      c
## FM 1.157436      c
```

Si analizamos los grupos ofrecidos por la Corrección de Bonferroni, coinciden con los que llevamos detectando desde el principio de la práctica y, en concreto, con el test de comparaciones múltiples sin corrección.

Grupo a	Grupo b	Grupo c
No fumadores (NF)	Fumadores pasivos (FP) Fumadores que no inhalan (NI) Fumadores ligeros (FL)	Fumadores moderados (FM) Fumadores intensivos (FI)

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.**

Pasamos a desarrollar el código requerido en `R`.

```
# Agrupar el conjunto de datos por tipo y género
grouped_data <- smokers %>%
  group_by(tipo, genero)

# Calcular la media de AE en cada grupo
mean_data <- grouped_data %>%
  summarise(mean_AE = round(mean(AE),3))
```

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

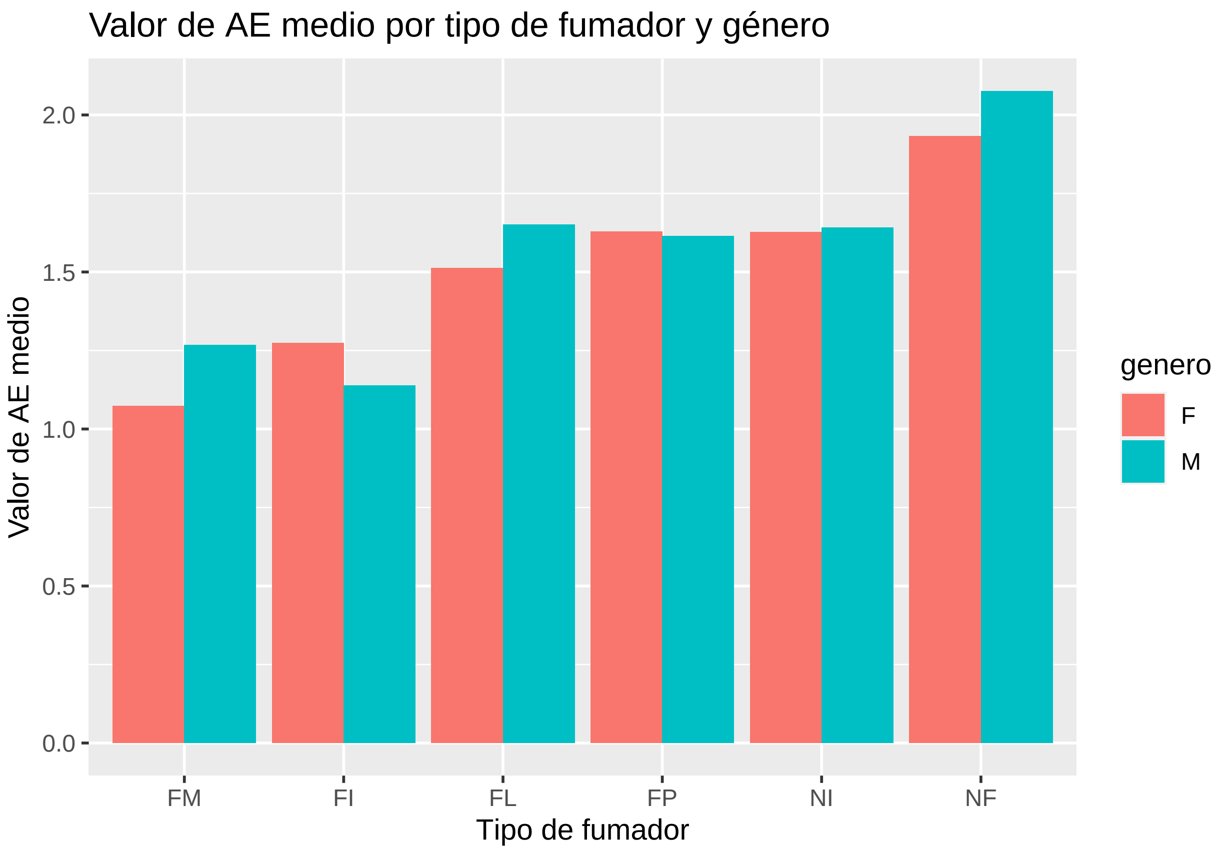
```
# Mostrar el conjunto de datos en forma de tabla
kable(mean_data, caption = 'Media de la capacidad pulmonar por tipo de fumador y género')
```

Cuadro 12: Media de la capacidad pulmonar por tipo de fumador y género

tipo	genero	mean_AE
FI	F	1.274
FI	M	1.139
FL	F	1.512
FL	M	1.651
FM	F	1.073
FM	M	1.267
FP	F	1.629
FP	M	1.615
NF	F	1.932
NF	M	2.075
NI	F	1.627
NI	M	1.642

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.**

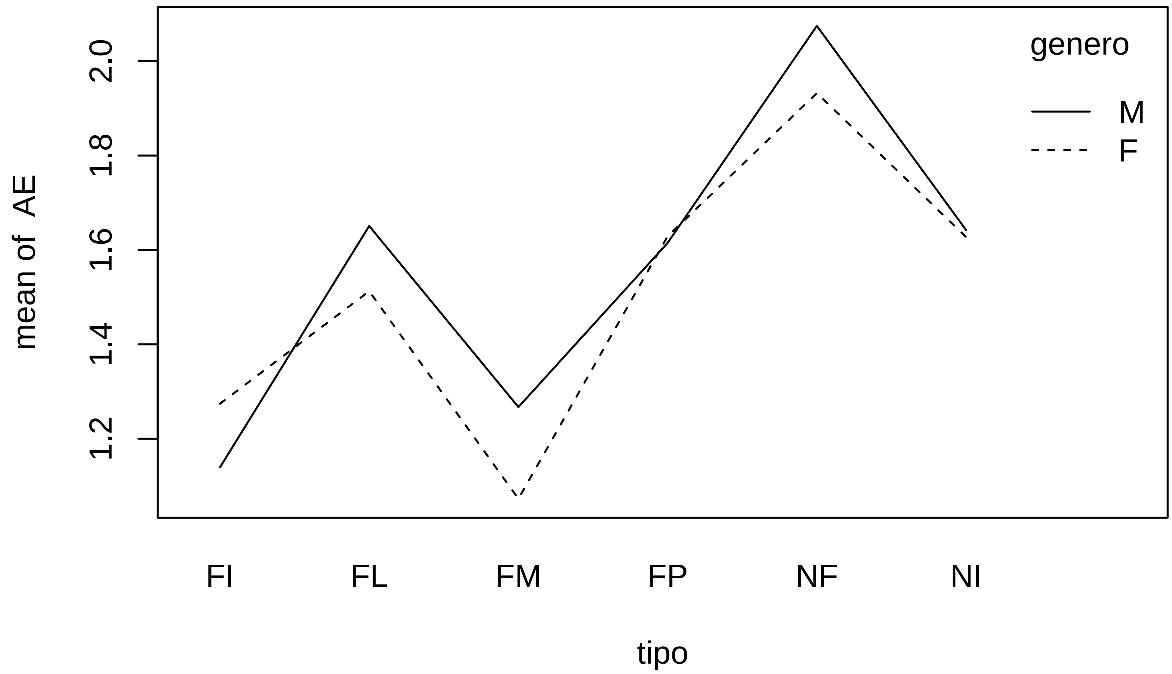
```
ggplot(mean_data, aes(x=reorder(tipo,
                                mean_AE),
                      y = mean_AE,
                      fill = genero)) +
  geom_bar(stat = "identity",
           position = "dodge") +
  ggtitle("Valor de AE medio por tipo de fumador y género") +
  xlab("Tipo de fumador") +
  ylab("Valor de AE medio")
```



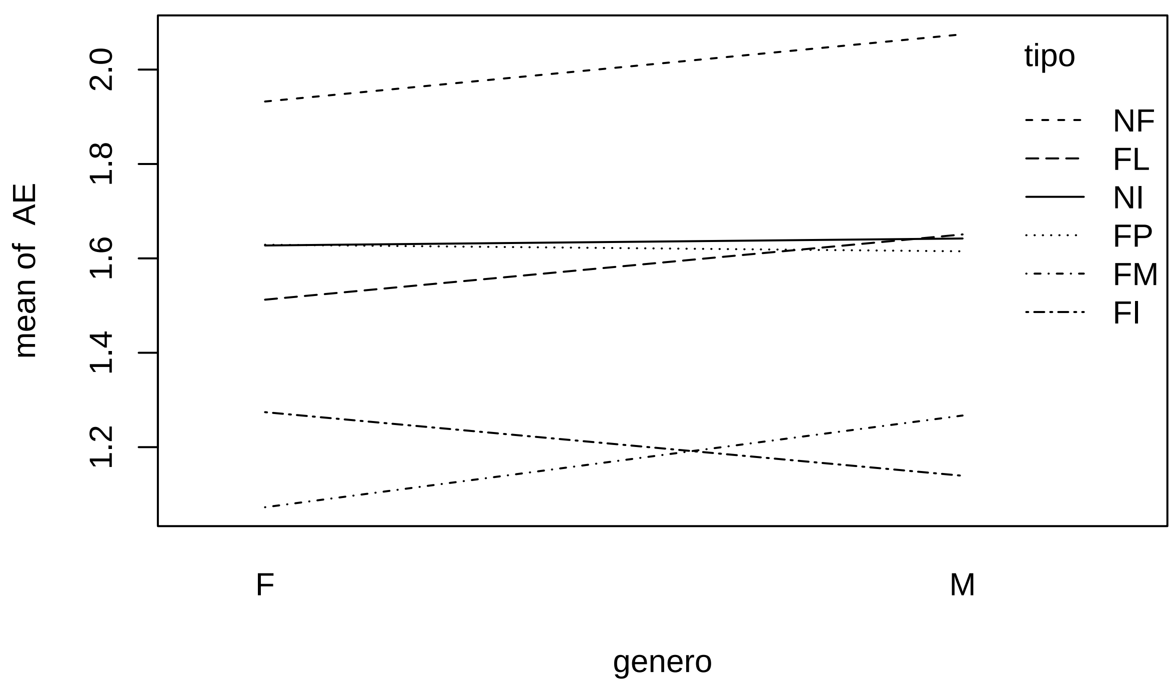
Vemos cómo, mientras en algunos grupos la media máxima es masculina, en otros es femenina.

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.**

```
interaction.plot(tipo, genero, AE)
```




```
interaction.plot(genero, tipo, AE)
```



Con estos gráficos, podemos decir que hay efectos principales en las dos variables, ya que hay distancia entre las líneas de los dos gráficos, si bien existen niveles del factor `tipo` cuyas líneas están prácticamente unidas. Véase NI y FP.

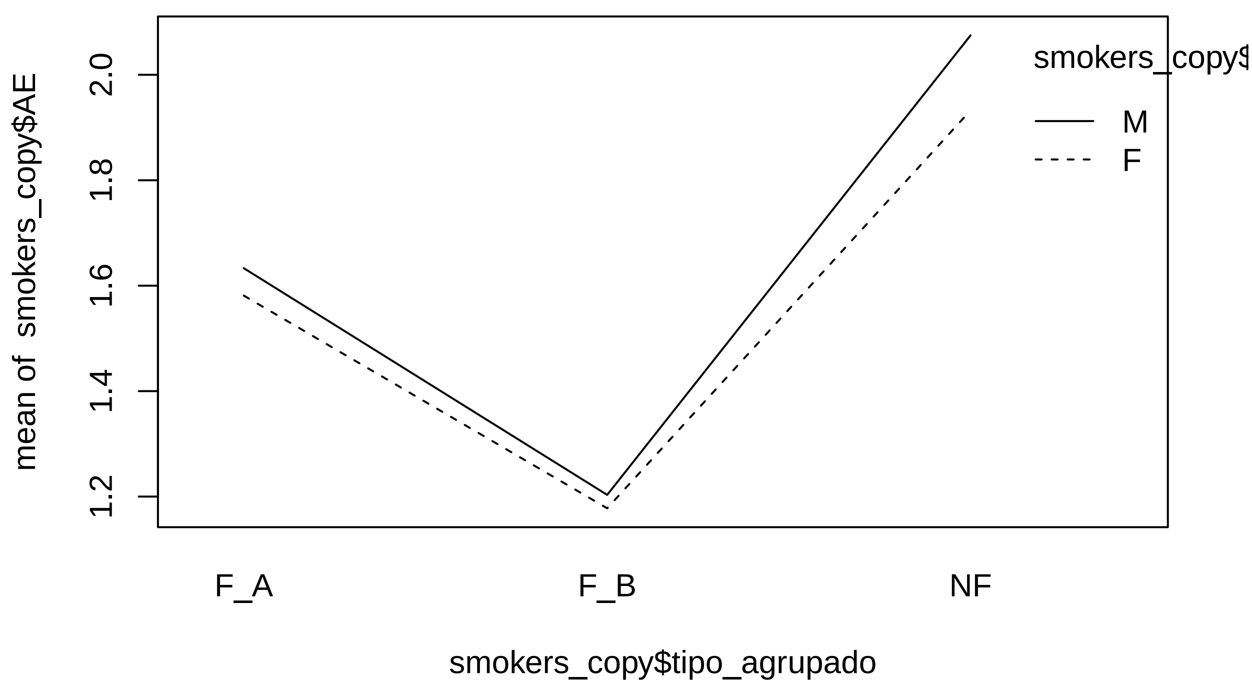
Parece que además existe interacción, dado que las líneas se cruzan. Sin embargo, es difícil interpretar los resultados con los datos que tenemos. Una interacción sugiere que los efectos de una variable dependen del valor de la otra. Viendo el segundo gráfico, podríamos interpretar que los efectos de fumar perjudican más a los hombres, ya que en los No Fumadores(NF) muestran una mayor capacidad pulmonar en hombres que en mujeres, mientras que la media en los Fumadores Intensivos (FI) se invierte y las mujeres tienen mejor media que en los hombres. Sin embargo, esto se vuelve a invertir con el grupo de Fumadores Moderados (FM), lo que resulta confuso. Algo similar ocurre con el grupo central, donde los Fumadores Pasivos (FP) tienen mejor media que las mujeres, pero donde la línea prácticamente se mantiene paralela en Fumadores Pasivos (FP) y Fumadores que no Inhalan (NI). Fundamentalmente, parece que hay inversión de medias dentro de los subgrupos detectados en la práctica.

Ante esta confusión, decidimos experimentar y crear un tipo agrupado por los tres grupos diferenciados a lo largo de la actividad.

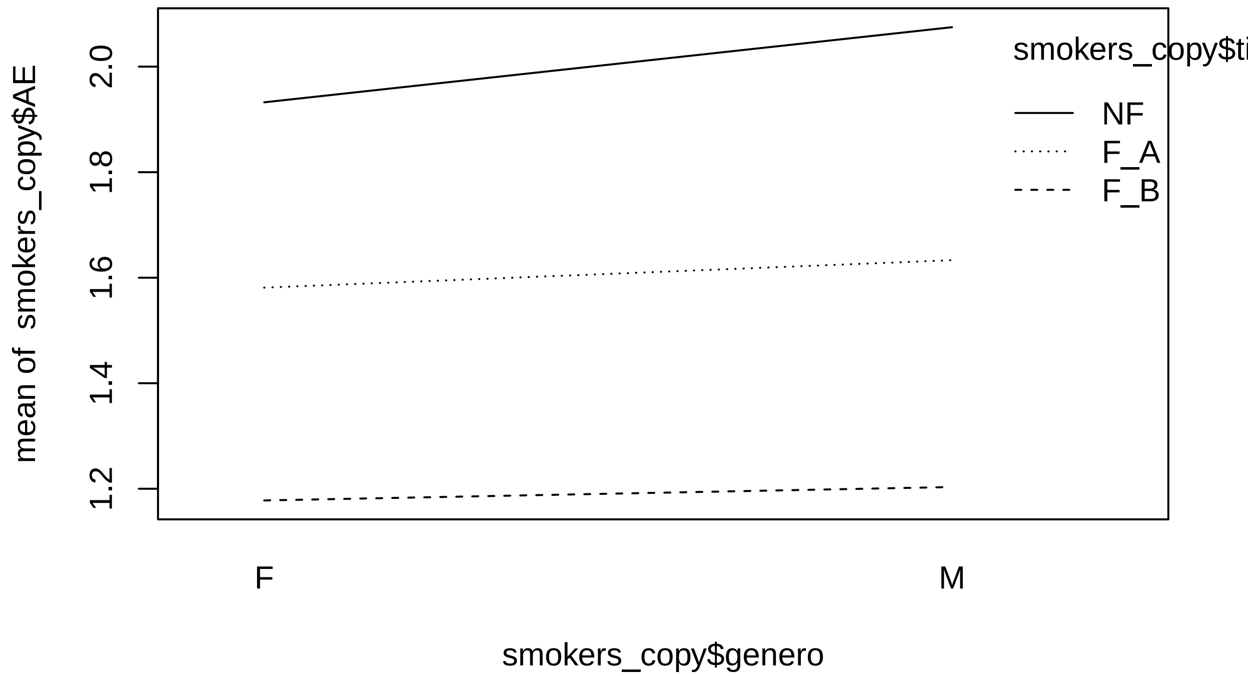
```r
# Crear copia del dataset original
smokers_copy <- smokers

# Crear nueva variable 'tipo_agrupado'
smokers_copy$tipo_agrupado <- ifelse(smokers_copy$tipo == "NF",
                                     "NF",
                              ifelse(smokers_copy$tipo %in% c("NI",
                                                              "FP",
                                                              "FL"), "F_A",
                              ifelse(smokers_copy$tipo %in% c("FM",
                                                              "FI"), "F_B", NA)))

interaction.plot(smokers_copy$tipo_agrupado,
                 smokers_copy$genero,
                 smokers_copy$AE)
```




```r
interaction.plot(smokers_copy$genero,
                 smokers_copy$tipo_agrupado,
                 smokers_copy$AE)
```



Efectivamente, vemos que esta interacción disminuye notablemente. Las líneas del segundo gráfico prácticamente se mantienen paralelas entre sí, por lo que podríamos descartar la interacción.

En cuanto a los efectos principales, están claros en la variable `tipo`. Nótese cómo eliminamos la superposición de líneas de los distintos niveles.

En conclusión, y basándonos en todo lo analizado en los ejercicios anterior, podemos ver que existen efectos principales en la variable `tipo`, cuyo análisis mejora mucho si se separa en 3 grupos; y podemos decir que la variable `genero` carece de efectos principales, por lo que es confusora en el primer análisis de interacción.

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.

Aplicamos el modelo con interacción, señalada por *.

```
modelo.92 <- aov(AE  ~ tipo*genero)
summary(modelo.92)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## tipo          5  20.80   4.160  17.684 6.41e-15 ***
## genero        1   0.20   0.201   0.855    0.356    
## tipo:genero   5   0.76   0.152   0.648    0.663    
## Residuals   241  56.69   0.235                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

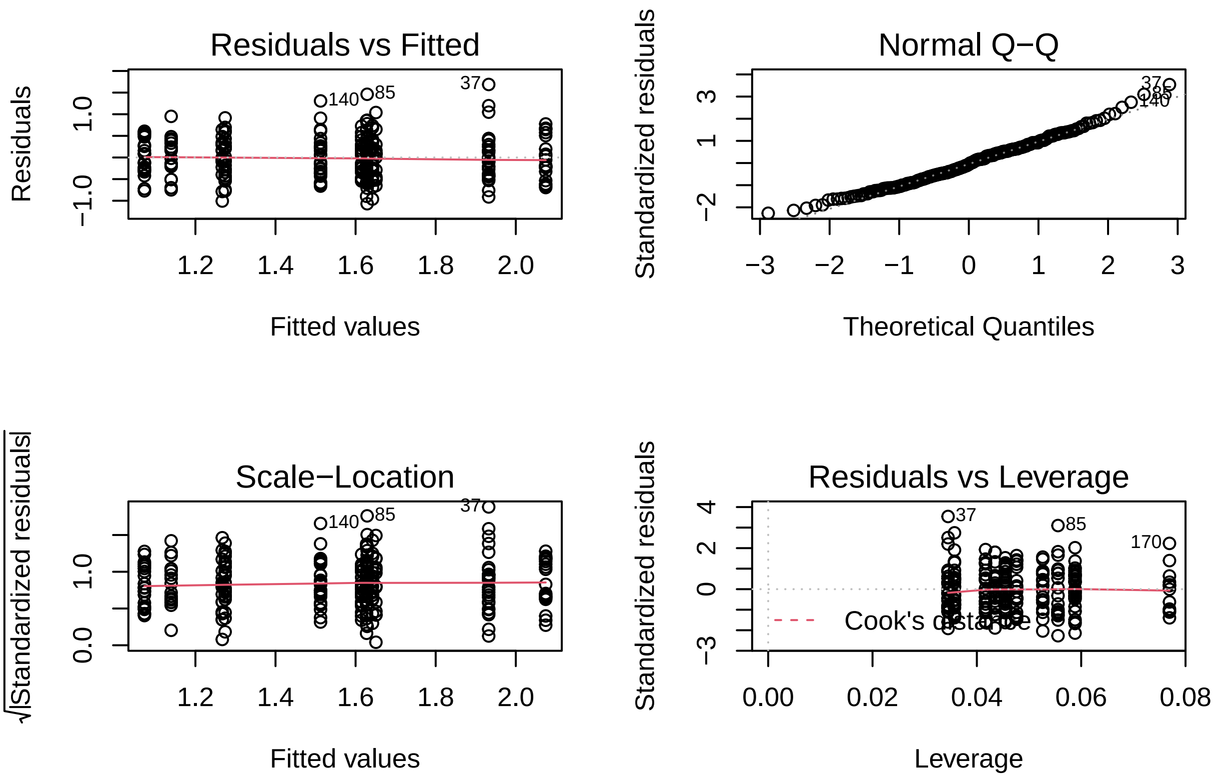
9.3 Interpretación

Interpretad el resultado.

Por mucho que hemos insistido a nuestro dataset en esta PEC, los resultados continúan siendo los mismos: la variable dependiente `AE` se puede explicar a partir de la variable independiente `tipo`, pero no se puede explicar a partir de la variable **genero**; tampoco si suponemos interacción. Así lo señalan los p-valores: la única variable con un valor p por debajo de 0.05 (incluso 0.1) es el tipo de fumador.

Analizamos los residuos para evaluar nuestro modelo.

```
par( mfrow=c(2,2), cex=.8)
plot(modelo.92)
```



Vemos que quizás se podría interpretar una cierta forma de embudo hacia los valores mayores, por lo que realizamos un test Bartlett.

```
condition <- with(smokers, interaction(tipo, genero))

bartlett.test(AE  ~ condition, data = smokers)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  AE by condition
## Bartlett's K-squared = 10.363, df = 11, p-value = 0.4981
```

Vemos que hay homocedasticidad, por lo que pasamos a evaluar la distrubución de los residuos.

```
shapiro.test(residuals(modelo.92))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modelo.92)
## W = 0.98732, p-value = 0.02481
```

De nuevo, volvemos a ver la misma gráfica y el mismo resultado del test de normalidad. La gráfica muestra una distribución que sigue la normal, salvo en los extremos. Esto hace que el test de Shapiro-Wilk rechace la hipótesis nula. Dado que estamos trabajando con muestras > 30, no creemos relevante realizar una transformación de la variable. Sin embargo, dejamos en el anexo un ejemplo de transformación de Box-Cox (que, como veremos, no altera en absoluto nuestras conclusiones). Finalmente, las últimas dos gráficas no encienden ninguna alarma.

10 Resumen técnico

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

Al ser tan grande la tabla, esta vez me ha quedado un poco apretada en las primeras columnas...

#	P	Pregunta	Resultado	Conclusión
1	1	Preprocesado (estado original)	smokers = 253 rows * 4 columns	El dataset *Fumadores.csv* se carga en la variable *smokers*, que contiene 253 observaciones de 4 variables distintas: 'AE' (chr), 'Tipo' (chr), 'genero' (chr), 'edad' (int).
2	-	Preprocesado 'AE'	Filas corregidas: 53, 58, 62, 68, 97, 107, 121, 158, 163, 165, 184, 195, 234, 250, 251	'AE': de *char* a *numeric*. Corrección de coma a punto.
3	-	Preprocesado 'Tipo'	Todas las filas relacionadas	'tipo': Cambio de nombre de variable de 'Tipo' a 'tipo' para unificar criterios en nomenclatura. Cambio de character a factor. { FM, fi, FI, FL, fm, FM, FM , FP, NF, NI} -> {FI, FL, FM, FP, NF, NI}
4	-	Preprocesado 'genero'	Todas las filas relacionadas.	'genero': De character a factor. { F, M}
5	-	Preprocesado 'edad'	-	No se detectan cambios necesarios.
6	2.1	Capacidad pulmonar en relación al género	Resumen estadístico y gráfico a partir de la p. 8	No se observan diferencias reseñables entre los dos grupos.
7	2.2	Relación entre capacidad pulmonar y edad	Gráfico disponible en la p. 11	Se observa una posible relación lineal con pendiente negativa entre la capacidad pulmonar y la edad, donde la capacidad pulmonar disminuye a medida que la edad del sujeto aumenta.

#	P	Pregunta	Resultado	Conclusión
8	2.3	Número por cada tipo de fumador y media de AE	FI 41 1.22 FL 41 1.56 FM 39 1.16 FP 40 1.62 NF 50 1.99 NI 42 1.63 Gráficos disponibles a partir de la p. 12	La distribución de las medias divide a los tipos de fumador en tres grupos: - Grupo 1: FM y FI. - Grupo 2: FL, FP y NI. - Grupo 3: NF.
9	3	IC de la capacidad pulmonar de hombres y mujeres	IC F (95%): [1.43, 1.62] IC M (95%): [1.48, 1.69]	Se cumplen lo supuestos para realizar el análisis para una t-Student con n-1 grados de libertad. El intervalo de confianza para la muestra de las mujeres es de [1.43, 1.62]; para los hombres es de [1.48, 1.69]. Las poblaciones pueden tener una capacidad pulmonar similar.
10	4	Diferencias en capacidad pulmonar entre mujeres y hombres	Estadístico: -0.86 Valores críticos: -1.97, 1.97 p-valor: 0.39	$H_0 : \mu_{male} = \mu_{fem}$ $H_1 : \mu_{male} \neq \mu_{fem}$ Tras hacer contraste de hipótesis bilateral de dos muestras independientes sobre la media con varianzas desconocidas pero iguales, concluimos que la capacidad pulmonar de los hombres y de las mujeres no es significativamente diferente con una confianza del 95%.
11	5	Diferencias en la capacidad pulmonar entre Fumadores y No Fumadores	Estadístico: -6.33 Valor crítico: -1.65 p-valor: 5.68×10^{-10}	$H_0 : \mu_{smokers} = \mu_{non\ smokers}$ $H_1 : \mu_{smokers} < \mu_{non\ smokers}$ Tras hacer un contraste de hipótesis unilateral de dos muestras independientes sobre la media con varianzas desconocidas pero iguales, concluimos que la capacidad pulmonar de los fumadores es inferior a la de los no fumadores con una confianza del 95%.

#	P	Pregunta	Resultado	Conclusión
12	6	Análisis de regresión lineal para `AE` y el resto de variables	p-values: modelo: < 2e-16 'tipo': < 4.24e-05, menos en FM 'edad': < 2e-16 'genero': 0.970 Multiple R-squared: 0.5829 El gráfico de la predicción se puede ver en la p.25	El *p*-valor del es muy inferior a al nivel de significación, por lo que el modelo es significativo para explicar la capacidad pulmonar. Pero nos encontramos con un *R-squared*, de 0.58. El modelo de regresión explicaría el 58.29% de la variabilidad total de las observaciones. Podemos considerar que el modelo tiene un ajuste mejorable. En cuanto a las variables, vemos que `edad` y `tipo` son significativas, si bien se señala el nivel `FM` como no significativo. `genero` no es significativa, por lo que la eliminaremos de nuestro modelo final: *lm_smokers*.
13	7.1	Normalidad	Test Shapiro-Wilk: W = 0.98875 p-value = 0.04599	Los datos no siguen una normal por muy poco. El p-valor de NF es de 0.036 y el de FL, de 0.4. El resto de niveles sí se ajustan a una distribución normal. Todos los niveles superan las 30 observaciones.
14	7.2	Homoscedasticidad	Test de Bartlett: p-value = 0.6633	No rechazamos la hipótesis nula y asumimos homoscedasticidad
15	7.3 a 7.5	Prueba ANOVA unifactorial	p-value = 4.449e-15	$H_0 : \alpha_{NF} = \alpha_{FP} = \alpha_{NI} = \alpha_{FL} = \alpha_{FM} = \alpha_{FI} = 0$ $H_1 : \alpha_i \neq \alpha_j$ para algún $i \neq j$ Aceptamos la hipótesis alternativa y concluimos que la variable `tipo` es significativa. Es decir, existe diferencia de medias entre al menos dos de los niveles.
16	7.6	Profundizando en ANOVA	SST = 78.46 SSW = 57.66 SSB = 20.8 Valor F = 17.821 Valor crítico (95%) = 2.251 p-value = 4.4×10^{-15}	De una variabilidad de 78.46, 20.8 se puede explicar por la variable `tipo`, mientras que 57.66 es la variación que no se logra explicar con el modelo.
17	7.7	Fuerza de la relación	$\eta^2 = SSA/SST \approx 0.265$	La variable `tipo` es significativa, pero solo explica el 26.51% de la variación total, por lo que haría falta estudiar otros factores para terminar de comprender la tendencia en la capacidad pulmonar.
18	8	Comparaciones múltiples	Test pairwise (p. 34) Corrección de Bonferroni en p. 35	Confirmamos que no existen diferencias significativas entre los niveles de los siguientes grupos: - Grupo 1: FM y FI. - Grupo 2: FL, FP y NI. - Grupo 3: NF.

#	P	Pregunta	Resultado	Conclusión
19	9	ANOVA Multi-facto-rial	p-value: 'tipo' = 6.41e-15 'genero' = 0.356 interacción 'tipo'/'genero' = 0.663	Tras el análisis visual, concluimos que no hay interacción (véase la explicación más detallada, p.38) y existen efectos principales en la variable 'tipo', pero no en la variable 'genero'. La variable dependiente `AE` se puede explicar a partir de la variable independiente `tipo`, pero no se puede explicar a partir de la variable `genero`; tampoco si suponemos 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.

Realizamos una serie de pruebas sobre una muestra de 253 sujetos donde recopilamos su capacidad pulmonar, el tipo de fumador, su edad y su género. Tras realizar múltiples pruebas estadísticas concluimos que:

- La edad juega un papel importante en la capacidad pulmonar: cuanto mayor es el paciente, menor capacidad pulmonar tiene.
- El género no parece ser un factor determinante sobre la capacidad pulmonar. Serían necesarias más pruebas para investigar posibles relaciones.
- El tipo de fumador sí es determinante en la capacidad pulmonar. Sin embargo, existen grupos de fumadores entre los que no parece haber diferencias.

Hay notables diferencias entre los los no fumadores y aquellos que fuman o inhalan humo. El grupo no fumador es el que mayor capacidad pulmonar tiene. A continuación, existe un segundo grupo, donde la capacidad respiratoria se ve disminuida, que incluye aquellos fumadores que no inhalan, los fumadores pasivos y los fumadores ligeros (aquellos que fuman o inhalan de uno a 10 cigarrillos al día durante 20 años o más). El tercer grupo tiene su capacidad pulmonar notablemente reducida respecto al primero y engloba a los fumadores moderados y los intesivos. Es decir, parece ser que a partir de un consumo de más de los 11 cigarrillos al día durante 20 años la capacidad pulmonar se ve disminuida al mismo nivel.

Para hacernos una idea, mientras la capacidad pulmonar de una persona de 30 años no fumadora es de 2.59, la de un fumador moderado/intensivo es de 1.8. Si nos vamos a una edad de 80, el no fumador tendrá una capacidad pulmonar de 1.04 y el de un fumador moderado/intensivo, de aproximadamente 0.28. Es decir, un fumador intensivo de 30 años tiene una edad pulmonar similar a la de alguien de 55 años que ni fuma ni inhala.

Dicho esto, tenemos que recordar que la edad y el tipo de fumador únicamente explican el 58% de la variabilidad de la capacidad pulmonar. Es decir, si queremos profundizar sobre este dato, deberemos hacer más estudios.

12 Anexo

Ejercicio 2.3

Presentamos una tabla para ver las medias de edad de los grupos de fumadores en el anexo.

```
tipo.edad <- smokers %>%
                group_by(tipo) %>%
                summarise(n=n(), mean=round(mean(edad), 2))

kable(tipo.edad,
      col.names =c("tipo","#", 'edad media'),
      caption ="Número de personas y media de edad por tipo de fumador")
```

Cuadro 14: Número de personas y media de edad por tipo de fumador

tipo	#	edad media
FI	41	49.22
FL	41	49.20
FM	39	52.64
FP	40	48.90
NF	50	49.42
NI	42	49.40

Efectivamente, la media de los fumadores moderados es mayor que la del resto, lo que puede explicar esta variación frente a los fumadores intensivos.

Fórmulas del Ejercicio 3

Siendo t_{n-1} una variable Student con $n-1$ grados de libertad y α, el nivel de signifación, tenemos que traducir la siguiente fórmula en una función en R:

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot s_{\bar{x}}$$

Donde:

$\bar{x}$ es la media muestral

$t_{\alpha/2,n-1}$ es el valor crítico tal que: $P(t_{n-1} \geq t_{\alpha/2,n-1}) = \alpha/2$

$s_{\bar{x}}$ es el error estándar, que se calcula con la fórmula siguiente:

$$\frac{s}{\sqrt{n}}$$

Donde s es la desviación típica muestral y n, el tamaño de la muestra.

Recordemos, por último que la desviación típica muestral se calcula:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Fórmulas del Ejercio 4

Como ya hemos comentado, necesitamos hacer un contraste de hipótesis bilateral de dos muestras independientes sobre la media con varianzas desconocidas pero iguales. A continuación presentamos los cálculos necesarios.

Deberemos calcular el **estadístico de contraste** de forma que:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}$$

Para una observación de una distribución t de Student con $n_1 + n_2 - 2$ grados de libertad. Donde:

$\bar{x}_1$ y $\bar{x}_2$ son las medias muestrales

$s_{\bar{x}_1 - \bar{x}_2}$ es el **error estándar**, que se calcula:

$$s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Donde:

n_1 y n_2 son los tamaños muestrales

s es la **desviación típica común**, que se calcula bajo la fórmula:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

El **p-valor** se calculará en base a la hipótesis alternativa planteada. Siendo $H_1 : \mu_1 - \mu_2 \neq 0$, el p-valor será:

$$p = P(|t_{n_1+n_2-2}| > |t|)$$

Los **valores críticos** son: $\pm t_{\alpha/2, n_1+n_2-2}$ tal que:

$$P(|t| > t_{\alpha/2,n_1+n_2-2}) = P(t < -t_{\alpha/2,n_1+n_2-2}) + P(t > t_{\alpha/2,n_1+n_2-2}) = \alpha$$

Aceptaremos H_0 si $|t| \leq t_{\alpha,n_1+n_2-2}$.

Fórmulas del ejercicio 6

El coeficiente de determinación R^2, según los apuntes (*Regresión lineal simple*, p. 23), "es la proporción entre la varianza explicada por la recta de regresión" y "la varianza total de datos", y nos ayudará a evaluar si el modelo de regresión lineal explica las variaciones que se generan los retrasos de los vuelos respecto a la distancia recorrida. Viene dado por la fórmula:

$$R^2 = \frac{SCR}{SCT} = \frac{\sum_{i=1}^{n}(\hat{y_1} - \bar{y})^2}{\sum_{i=1}^{n}(y_1 - \bar{y})^2}$$

Donde:

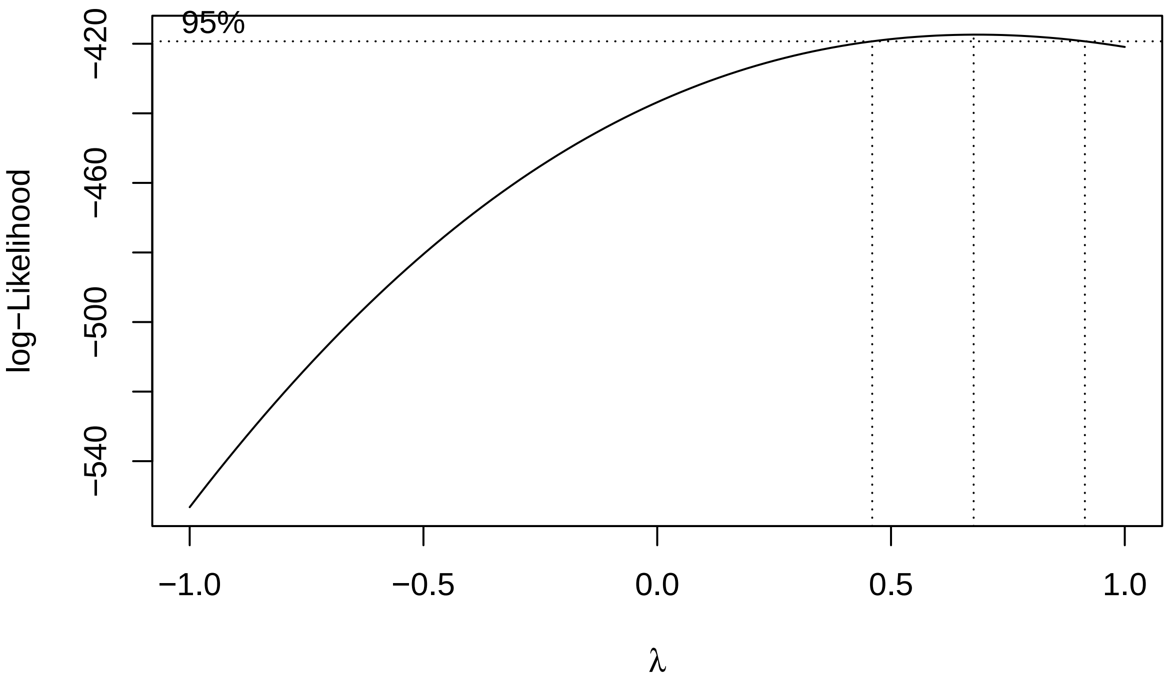
SCR = Suma de cuadrados de la regresión
SCT = Suma de cuadrados totales

El coeficiente de determinación se encuentra entre 0 y 1. Sus extremos significan:

$R^2 = 1 \rightarrow$ El ajuste es perfecto: todos los puntos se encuentran sobre la recta; no hay residuos.
$R^2 = 0 \rightarrow$ La relación entre las variables X e Y es inexistente.

Transformación de Box-Cox

```
library(MASS)
AEc<-boxcox(AE~tipo,lambda = seq(-1, 1, length = 10),plotit=T)
```




```
lambda<-AEc$x[which.max(AEc$y)]
lambda
```

```
## [1] 0.6767677
```

```
library(psych)

gm<-geometric.mean(AE)
gm
```

```
## [1] 1.438414
```

```
AEtrans<-(AE^lambda-1)/(lambda*gm^(lambda-1))

modelo.92c <- aov(AEtrans  ~ tipo*genero)
summary(modelo.92c)
```

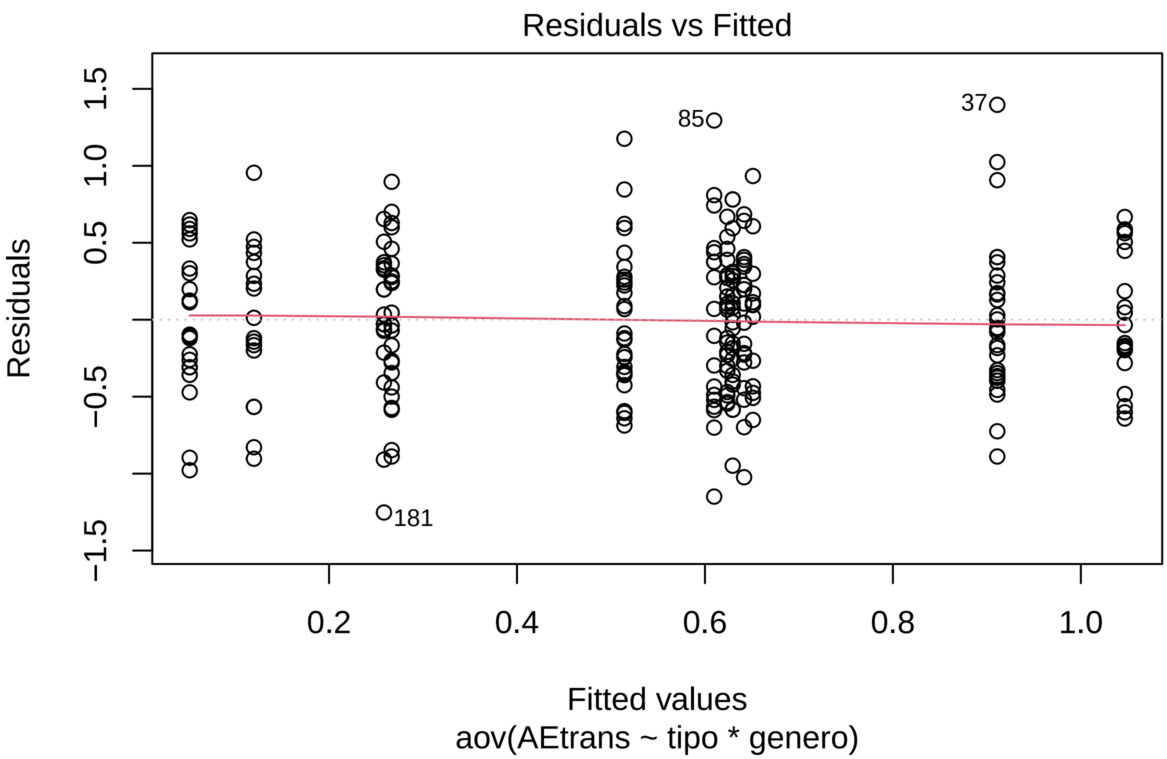
```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## tipo          5  20.55   4.111  17.997 3.68e-15 ***
## genero        1   0.22   0.219   0.960    0.328    
## tipo:genero   5   0.80   0.160   0.699    0.625    
## Residuals   241  55.05   0.228                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

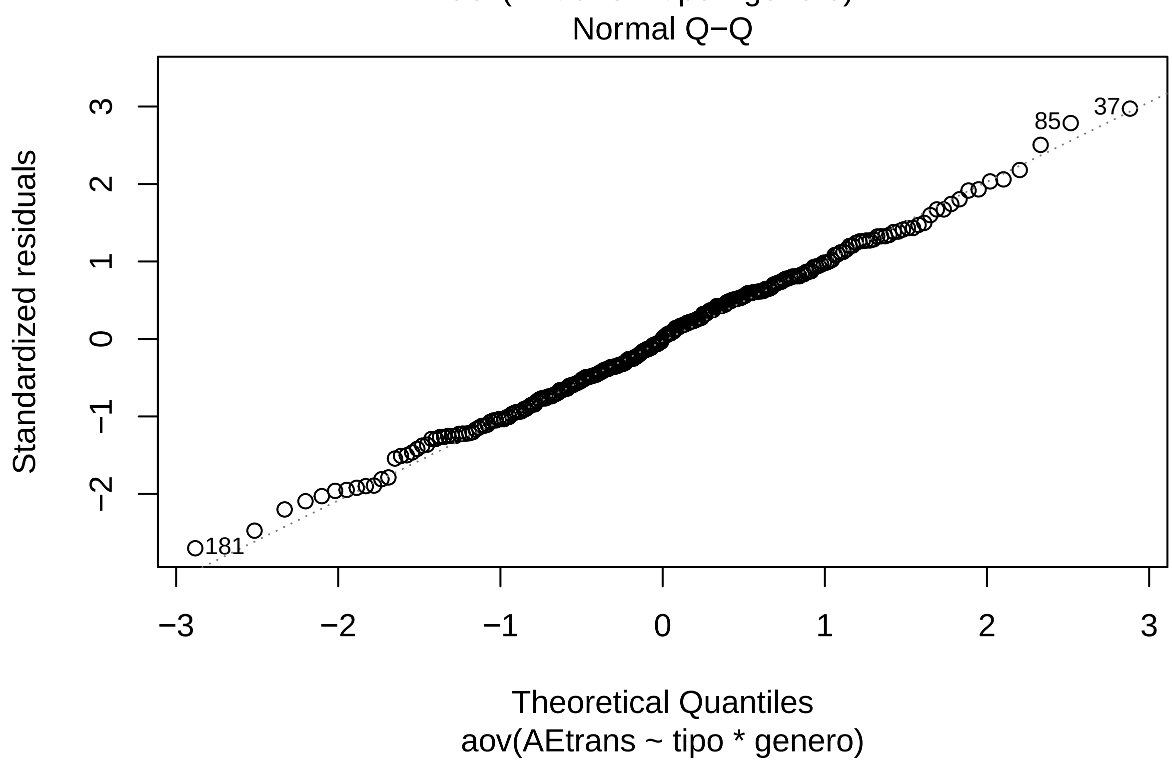
```
shapiro.test(residuals(modelo.92c))
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  residuals(modelo.92c)
## W = 0.99648, p-value = 0.8456
```

```r
plot(modelo.92c,which=c(1,2))
```





13 Referencias bibliográficas

Fox , J. (2007) [R] How to extract numbers from ANOVA tables?, [r] how to extract numbers from anova tables? Available at: https://stat.ethz.ch/pipermail/r-help/2007-December/147943.html (Accessed: January 27, 2023).

Legends (ggplot2) (no date) Cookbook for R . Available at: http://www.cookbook-r.com/Graphs/Legends_(ggplot2)/ (Accessed: January 27, 2023). R Documentation (no date) The F Distribution, R: The F distribution. Available at: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Fdist.html (Accessed: January 27, 2023).

Sinha, A. (2018) Plotting n columns of a data frame as lines with ggplot in R, Stack Overflow. Available at: https://stackoverflow.com/questions/50410524/plotting-n-columns-of-a-data-frame-as-lines-with-ggplot-in-r (Accessed: January 27, 2023).

Wickham, H., François, R. and Henry, L. (no date) Summarise each group to fewer rows - summarise, - summarise • dplyr. Available at: https://dplyr.tidyverse.org/reference/summarise.html (Accessed: January 27, 2023).

Fuentes empleadas para escribir el código en RMarkdown

TablesGenerator.com. (n.d.). Tables Generator. Retrieved November 2, 2022, from https://www.tablesgenerator.com/markdown_tables/ (Accessed: December 23, 2022).