A4 - Análisis estadístico avanzado

Estadística Avanzada

Leroy Deniz

Actualizado: 26 December, 2022

Contents

0 Contexto **3**
0.1 Importación de librerías . . . 3
0.2 Funciones auxiliares . . . 3
0.3 Configuraciones . . . 3

1 Preprocesado **4**
1.1 Lectura del fichero . . . 4
1.2 Consulta de tipos y transformaciones . . . 4

2 Análisis de la muestra **7**
2.1 Capacidad pulmonar y género . . . 7
2.2 Capacidad pulmonar y edad . . . 7
2.3 Tipos de fumadores y capacidad pulmonar . . . 8

3 Intervalo de confianza de la capacidad pulmonar **11**

4 Diferencias en capacidad pulmonar entre mujeres y hombres **12**
4.1 Hipótesis . . . 12
4.2 Contraste . . . 12
4.3 Cálculos . . . 13
4.3 Cálculos . . . 13

5 Diferencias en la capacidad pulmonar entre Fumadores y No Fumadores **14**
5.1 Hipótesis . . . 14
5.2 Contraste . . . 14
5.3 Preparación de los datos . . . 14
5.4 Cálculos . . . 15
5.5 Interpretación . . . 15

6 Análisis de regresión lineal **16**
6.1 Cálculo . . . 16
6.2 Interpretación . . . 16
6.3 Bondad del ajuste . . . 17
6.4 Predicción . . . 17

7 ANOVA unifactorial **19**
7.1 Normalidad . . . 19
7.2 Homocedasticidad: Homogeneidad de varianzas . . . 20
7.3 Hipótesis nula y alternativa . . . 20
7.4 Cálculo ANOVA . . . 20
7.5 Interpretación . . . 21
7.6 Profundización en ANOVA . . . 21
Cálculo del F value . . . 21

Cálculo del Valor crítico 21
Cálculo del P value 22
7.7 Fuerza de la relación 22

8 Comparaciones múltiples **23**
8.1 Test pairwise 23
8.2 Corrección de Bonferroni 24

9 ANOVA multifactorial **25**
9.1 Análisis visual 25
9.2 ANOVA multifactorial 26
9.3 Interpretación 27

10 Resumen técnico **28**

11 Resumen ejecutivo **29**

0 Contexto

0.1 Importación de librerías

```
library(ggplot2)
library(tidyverse)
library(reshape2)
library(stats)
library(dplyr)
```

0.2 Funciones auxiliares

```
# Función para mostrar información en vertical
vertical <- function(tbl) {
    t(t(tbl))
}
```

0.3 Configuraciones

```
options(dplyr.summarise.inform = FALSE)
```

1 Preprocesado

1.1 Lectura del fichero

```
df <- read.csv("Fumadores.csv", sep = ";", dec = ".")
```

Muestra del dataset generado a raíz de la lectura:

```
head(df)
```

```
##         AE Tipo genero edad
## 1 1.871878   NF      M   54
## 2  1.91312   NF      F   60
## 3  2.58114   NF      M   40
## 4  2.17827   NF      F   55
## 5 1.707732   NF      F   59
## 6 1.561215   NF      F   63
```

1.2 Consulta de tipos y transformaciones

```
vertical(sapply(df, class))
```

```
##        [,1]
## AE     "character"
## Tipo   "character"
## genero "character"
## edad   "integer"
```

Una vez conocidos los tipos de datos en función de su contenido, se evalúa por separado los tipos *character* para estandarizar los valores si corresponde. Los valores presentes en la variable *genero* son dos y correctos como puede verse a continuación.

```
unique(df$genero)
```

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

Sin embargo en la variable *Tipo* se encuentran los valores con espacios y con mayúsculas y minúsculas.

```
unique(df$Tipo)
```

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

A continuación se eliminan los espacios y se convierte el contenido todo a mayúsculas.

```
df$Tipo = sapply(df$Tipo, toupper)
df$Tipo <- sapply(df$Tipo, trimws, which = c("both"))
tipos <- unique(df$Tipo)
tipos
```

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

La variable *AE* se identificaba como *character* porque tenía comas en lugar de puntos en su separador de decimales. Se aplica el cambio para corregirlo.

```
df$AE <- sub(",", ".", df$AE, fixed = TRUE)
```

Una vez procesados todas las variables con las correcciones previas, se realizrá la conversión de la variable *AE* a tipo *Number*; la columna *edad* es de tipo entera ya está correctamente definida por R, así como las variables *Tipo* y *genero* que son de tipo *character* pero mantienen siempre un conjunto finito de valores, por lo que podemos pasarlas a *factor*.

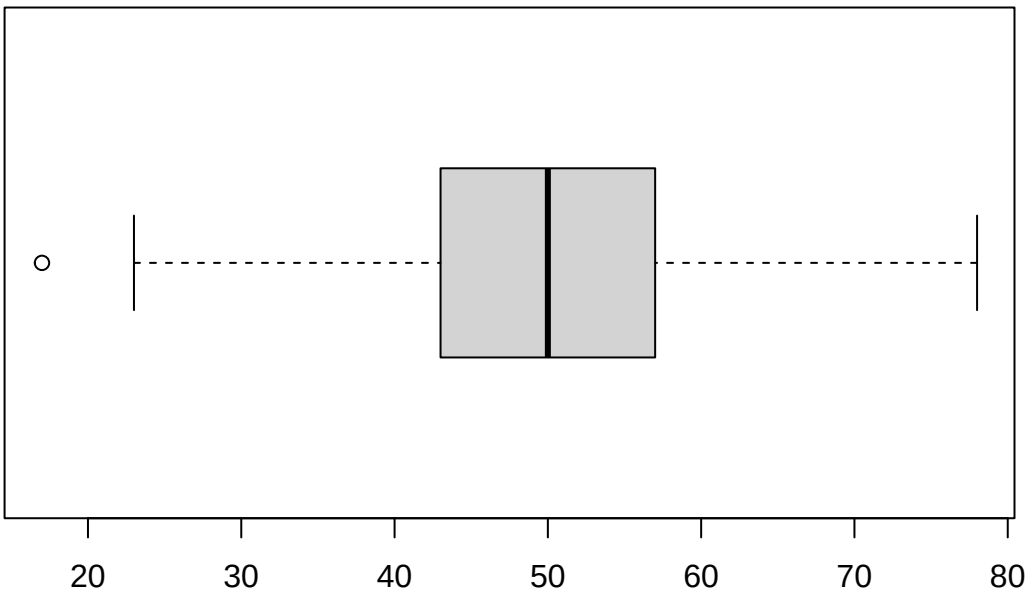
```
df$AE <- as.numeric(df$AE)
df$Tipo <- factor(df$Tipo)
df$genero <- factor(df$genero)

vertical(sapply(df, class))
```

```
##        [,1]
## AE     "numeric"
## Tipo   "factor"
## genero "factor"
## edad   "integer"
```

Para encontrar posibles valores atípicos en la variable *edad*, se utiliza un boxplot y se contebilizan.

```
boxplot(df$edad, horizontal = TRUE)
```




```
outliers = boxplot.stats(df$edad)$out
cat("Hay un total de", length(outliers), "outliers en la variable edad")
```

```
## Hay un total de 1 outliers en la variable edad
```

Al único outlier encontrado, se le asigna NaN como valor y se muestra a través de la función *summary*.

```
df$edad[which(df$edad %in% outliers)] = NaN
summary(df$edad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   23.00   43.00   50.00   49.89   57.00   78.00       1
```

Al outlier se le imputa el de la media de la serie y se verifica nuevamente cuántos outliers hay.

```
df$edad[is.na(df$edad)] <- mean(df$edad, na.rm = T)
outliers = boxplot.stats(df$edad)$out
cat("Hay un total de", length(outliers), "outliers en la variable edad")
```

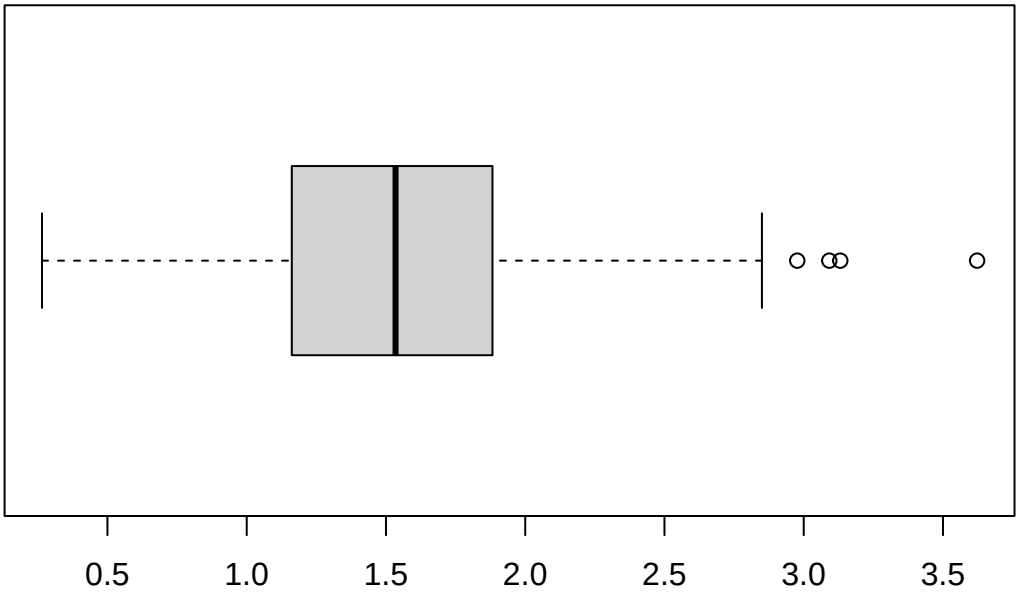
```
## Hay un total de 0 outliers en la variable edad
```

Ahora bien, para el caso de la variable AE se realiza otro boxplot para verificar la existencia de outliers y contarlos.

```
outliers = boxplot.stats(df$AE)$out
cat("Hay un total de", length(outliers), "outliers en la variable AE")
```

```
## Hay un total de 4 outliers en la variable AE
```

```
boxplot(df$AE, horizontal = TRUE)
```



Se realiza el mismo procedimiento de imputación que para la variable *edad*, imputándole el valor de la media.

```
df$AE[is.na(df$AE)] <- mean(df$AE, na.rm = T)
```

Estructura final del dataset procesado.

```
summary(df)
```

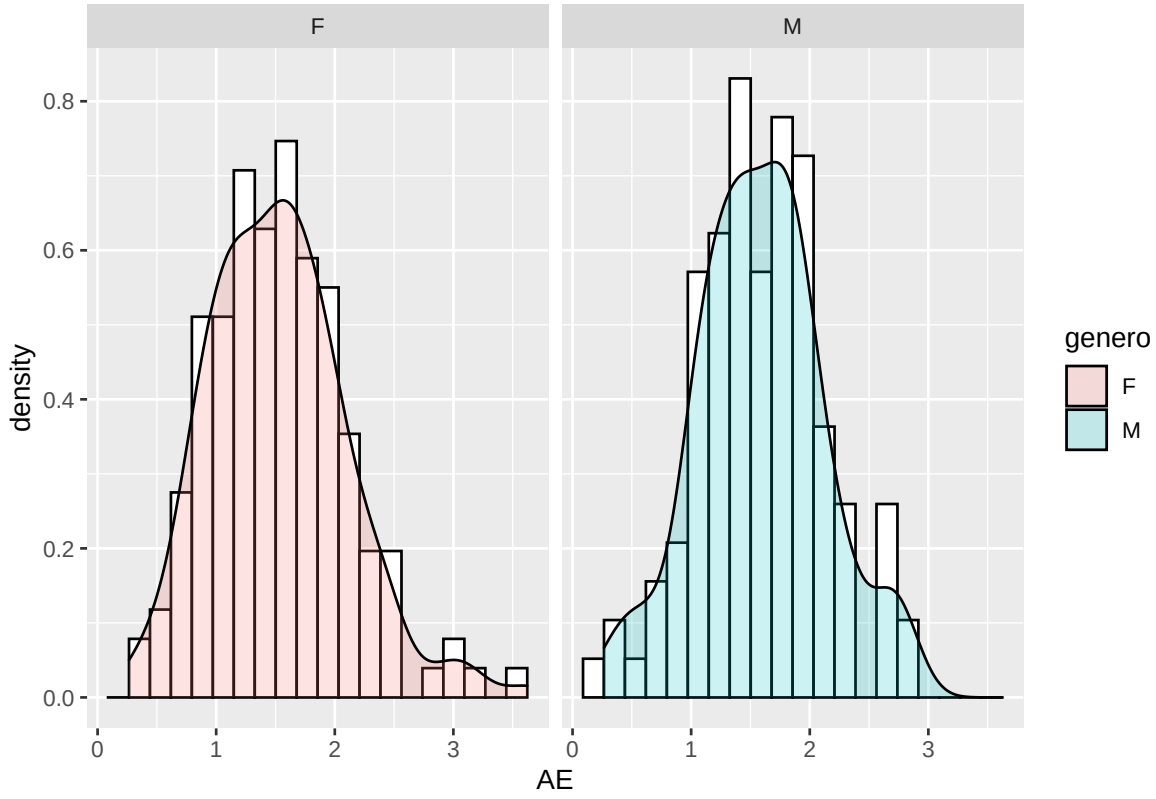
```
##        AE         Tipo    genero       edad
##  Min.   :0.2649   FI:41   F:144   Min.   :23.00
##  1st Qu.:1.1618   FL:41   M:109   1st Qu.:43.00
##  Median :1.5344   FM:39           Median :50.00
##  Mean   :1.5493   FP:40           Mean   :49.89
##  3rd Qu.:1.8824   NF:50           3rd Qu.:57.00
##  Max.   :3.6226   NI:42           Max.   :78.00
```

2 Análisis de la muestra

2.1 Capacidad pulmonar y género

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

```
ggplot(df, aes(AE)) + geom_histogram(aes(y = ..density..), bins = 20,
    color = "black", fill = "white") + geom_density(aes(fill = genero),
    alpha = 0.2) + facet_wrap(~genero)
```

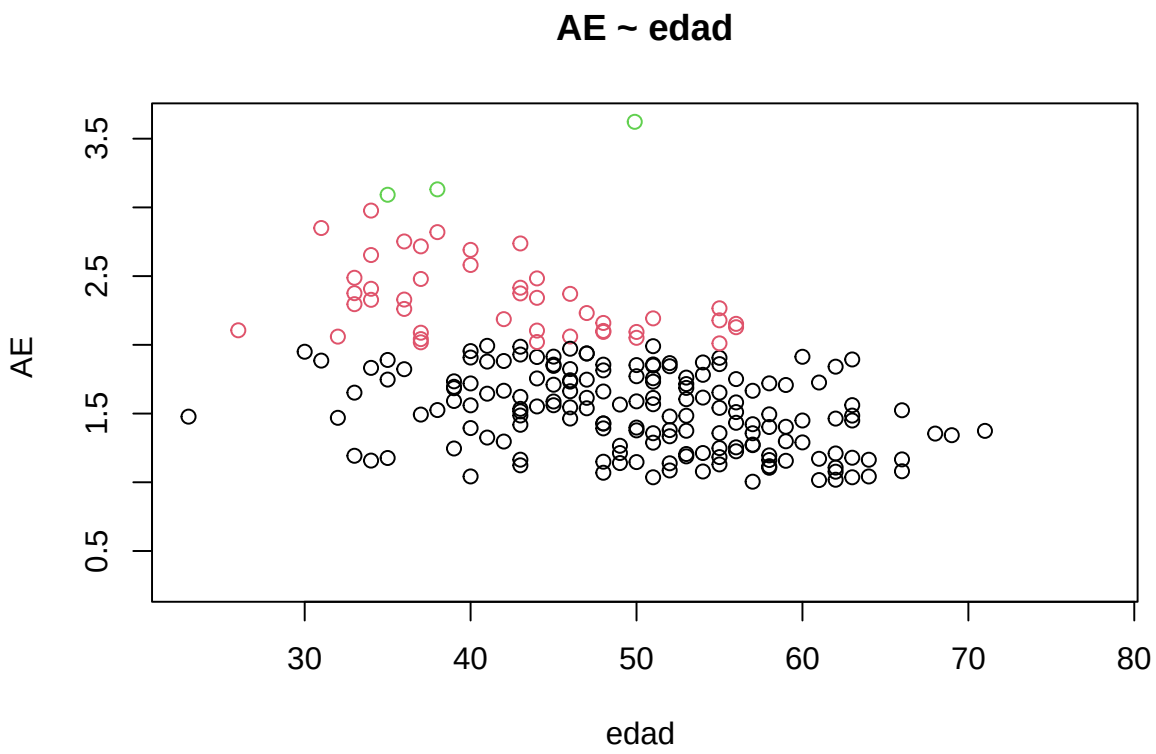


Las distribuciones parecen centradas en la media del intervalo [0;3], aunque sí existen más casos atípicos cuando el género es *F*.

2.2 Capacidad pulmonar y edad

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

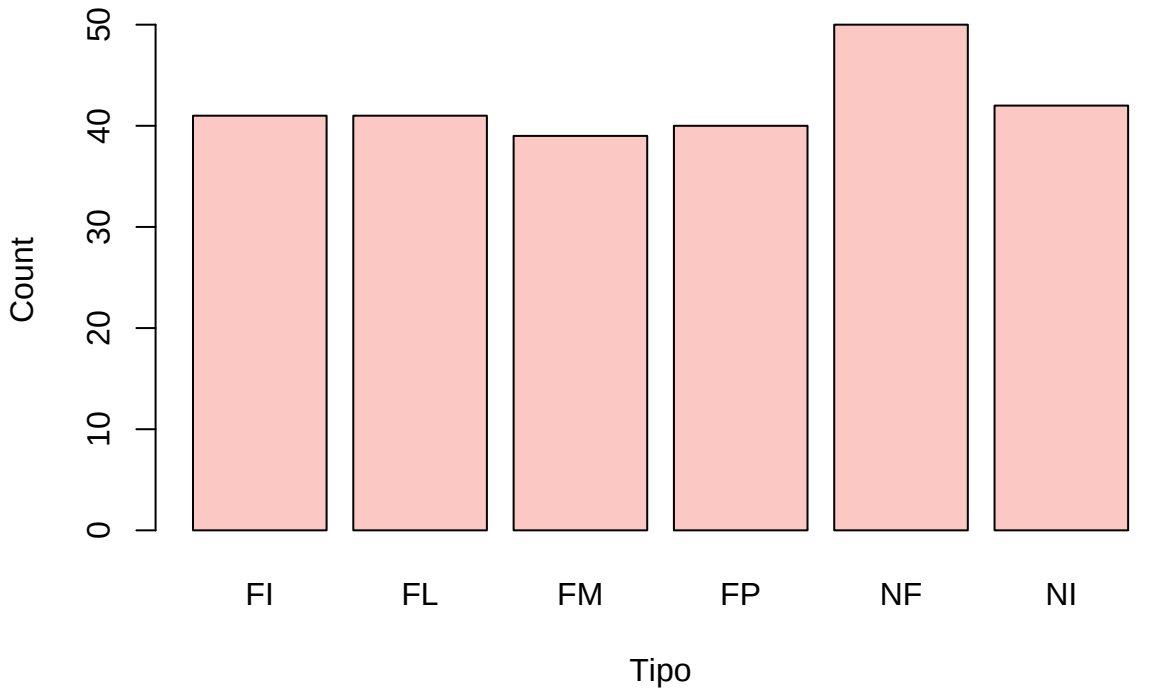
```
plot(df$edad, df$AE, main = "AE ~ edad", xlab = "edad ", ylab = "AE ",
    col = df$AE)
```



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.

```
barplot(table(df$Tipo), col = "#F7766C66", xlab = "Tipo", ylab = "Count")
```



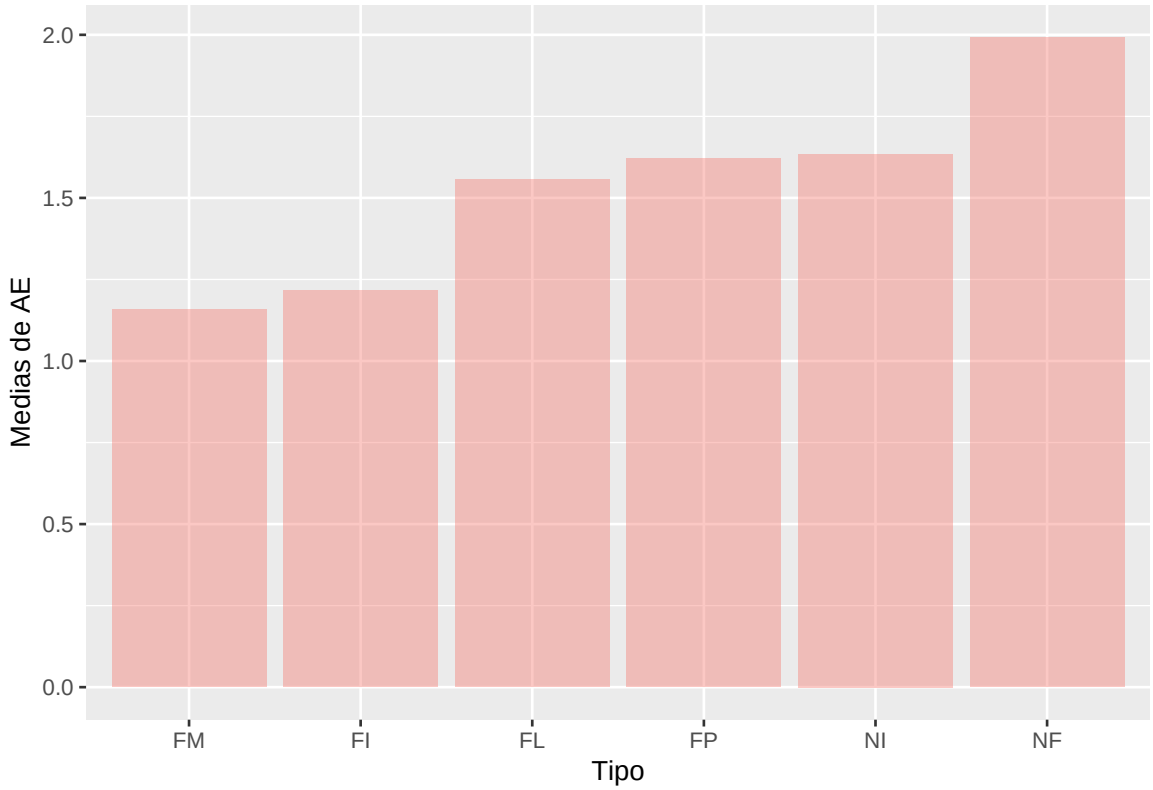
Cálculo de la media de capacidad pulmonar AE para cada tipo de fumador.

```
resumen <- df %>%
    group_by(Tipo) %>%
    summarize(means = mean(AE), counts = length(AE))
print(resumen)
```

```
## # A tibble: 6 x 3
##   Tipo  means counts
##   <fct> <dbl>  <int>
## 1 FI     1.22     41
## 2 FL     1.56     41
## 3 FM     1.16     39
## 4 FP     1.62     40
## 5 NF     1.99     50
## 6 NI     1.63     42
```

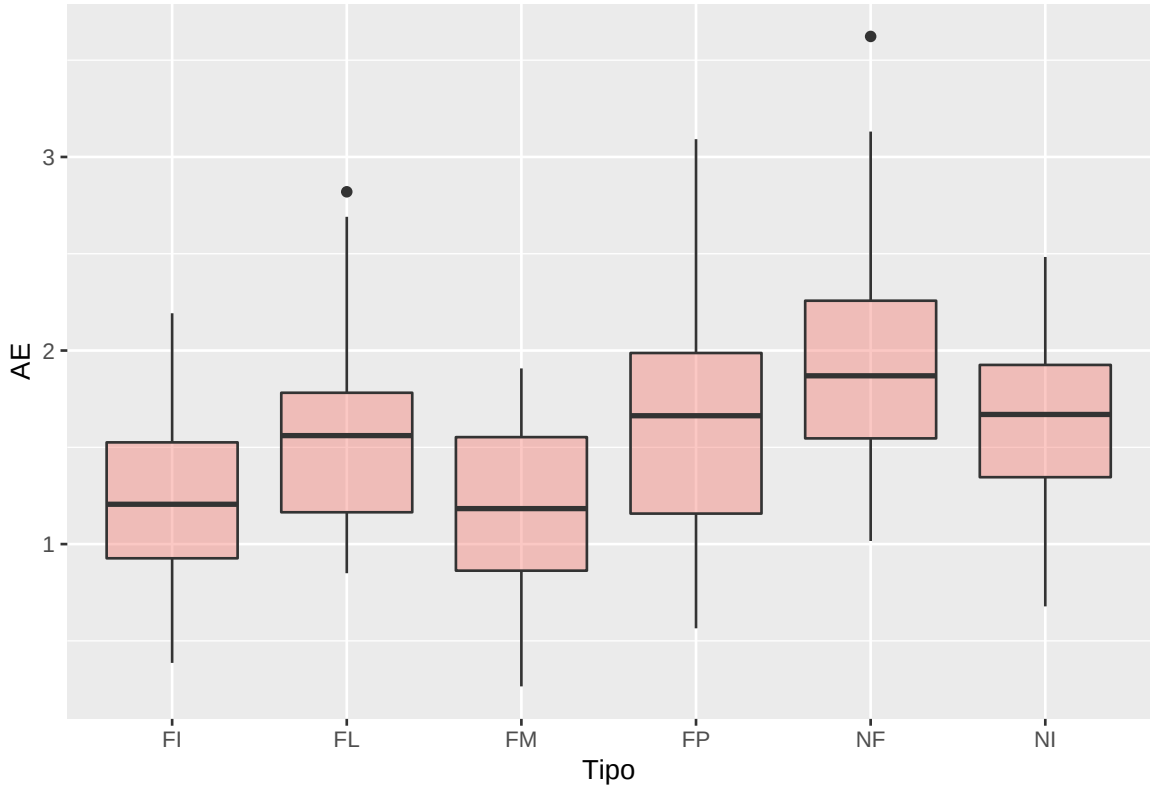
Gráfico de las medias según el tipo de fumador.

```
ggplot(resumen, aes(x = reorder(Tipo, means), y = means)) + geom_col(fill = "#F7766C66") +
    xlab("Tipo") + ylab("Medias de AE")
```



Distribución de *AE* según el tipo de fumador.

```
ggplot(df, aes(x = Tipo, y = AE)) + geom_boxplot(fill = "#F7766C66")
```



A partir del boxplot anterior, se puede ver que existen dos outliers en la variable *Tipo* cuando toma los valores FL y NF. Además, se podría ver que la media de capacidad pulmonar para los No fumadores (NF) es la más alta de todos los demás tipos y que podría haber una relación entre los tipos FI y FM.

3 Intervalo de confianza de la capacidad pulmonar

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

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

Se verifican las condiciones para la aplicación del intervalo de confianza, donde se tiene un total de 253 registros en el dataset, con 109 casos masculinos y 144 casos Femeninos. Se tiene entonces más de 30 casos para cada tipo y una varianza descnocida, por lo que se verifica así que puede construir a través de una distribución normal.

```
# Función de cálculo del intervalo de confianza para
# distribución normal
IC <- function(x, NC) {
    n <- length(x)
    alpha <- 1/(NC/100)
    SE <- sd(x)/sqrt(n)

    z <- qnorm(alpha/2, lower.tail = TRUE)
    z_SE <- z * SE
    Low <- mean(x) - z_SE
    Up <- mean(x) + z_SE

    return(c(Low, Up))
}
```

```
int_conf_f <- IC(df$AE[df$genero == "F"], 95)
int_conf_m <- IC(df$AE[df$genero == "M"], 95)
```

El intervalo de confianza al 95% para el género M es [1.5803834, 1.5871413] mientras que para el género F es [1.5201125, 1.5264477].

Ambos intervalos son relativamente similares puesto que varían recién en la segunda cifra decimal, ambos rondan el 1.55 como valor central.

4 Diferencias en capacidad pulmonar entre mujeres y hombres

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

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

4.1 Hipótesis

Escribir la hipótesis nula y alternativa.

$$H_0 : \mu_{AE\_M} = \mu_{AE\_F}$$
$$H_1 : \mu_{AE\_M} \neq \mu_{AE\_F}$$

4.2 Contraste

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

Se utiliza para este apartado un contraste de media de dos distribuciones, porque son dos separadas, independientes una de la otra aunque ambas pertenecen a la misma muestra.

```
contraste_medias <- function(s1, s2, alt, CL) {

    # Cálculo de Medias
    mean1 <- mean(s1)
    mean2 <- mean(s2)

    # Cálculo del tamaño de la muestra
    n1 <- length(s1)
    n2 <- length(s2)

    # Cálculo de la desviación estándar
    sd1 <- sd(s1)
    sd2 <- sd(s2)

    # Cálculo del nivel de significancia
    alpha <- (1 - CL/100)

    # Cálculo de los grados de libertad (Apartado 5.2.2 de
    # la teoría)
    denominador <- ((sd1^2/n1)^2/(n1 - 1) + (sd2^2/n2)^2/(n2 -
        1))
    df <- ((sd1^2/n1 + sd2^2/n2)^2)/denominador

    # Cálculo del valor t (z según la distribución normal
    # estandarizada)
    sb <- sqrt(sd1^2/n1 + sd2^2/n2)
    t <- (mean1 - mean2)/sb

    # Evaluación de la condición =
    if (alt == "bilateral") {
```

```
        t_critical <- qt(alpha/2, df, lower.tail = FALSE)
        p_value <- pt(abs(t), df, lower.tail = FALSE) * 2

        # Evaluación de la condición <
    } else if (alt == "<") {
        t_critical <- qt(alpha, df, lower.tail = TRUE)
        p_value <- pt(t, df, lower.tail = TRUE)

        # Evaluación de la condición > (alt == '>')
    } else {
        t_critical <- qt(alpha, df, lower.tail = FALSE)
        p_value <- pt(t, df, lower.tail = FALSE)
    }

    # Definición del vector resultado
    vector_data <- c(mean1, mean2, t, t_critical, p_value, alpha,
        df)
    names(vector_data) <- c("mean1", "mean2", "t", "t_critical",
        "p_value", "alpha", "df")
    return(vector_data)
}
```

4.3 Cálculos

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

```
x <- df$AE[df$genero == "F"]
y <- df$AE[df$genero == "M"]
datos <- contraste_medias(x, y, "bilateral", 95)
vertical(datos)
```

```
##                   [,1]
## mean1        1.5232801
## mean2        1.5837624
## t           -0.8620365
## t_critical   1.9698650
## p_value      0.3895251
## alpha        0.0500000
## df         240.7863233
```

4.3 Cálculos

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

Como el *p_value* es mayor que el nivel de significancia, se debe aceptar la hipótesis nula porque no hay evidencia suficiente para poder descartarla. Por lo tanto, lo único que se puede decir es que la capacidad pulmonar de ambos grupos se muestra igual.

5 Diferencias en la capacidad pulmonar entre Fumadores y No Fumadores

---

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

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

5.1 Hipótesis

Escribir la hipótesis nula y alternativa.

$$H_0 : \mu_{AE\_FUM} \geq \mu_{AE\_NOFUM}$$
$$H_1 : \mu_{AE\_FUM} < \mu_{AE\_NOFUM}$$

5.2 Contraste

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

Se aplica un contraste de hipótesis de dos muestras independientes, ya que hay suficientes casos en cada distribución para poder afirmar que siguen una distribución normal. Además, se tienen varianzas poblacionales desconocidas diferentes.

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.

```
indeces <- df$Tipo == "NF" | df$Tipo == "FP"
no_fumadores <- df$AE[indeces]
fumadores <- df$AE[!indeces]
cat("No Fumadores: ", length(no_fumadores))
```

```
## No Fumadores:  90
```

```
cat("Fumadores: ", length(fumadores))
```

```
## Fumadores:  163
```

5.4 Cálculos

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.

```
datos <- contraste_medias(fumadores, no_fumadores, "<", 95)
vertical(datos)
```

```
##                      [,1]
## mean1       1.395786e+00
## mean2       1.827437e+00
## t          -6.128091e+00
## t_critical -1.654029e+00
## p_value     3.095169e-09
## alpha       5.000000e-02
## df          1.669948e+02
```

5.5 Interpretación

Interpretar el resultado del contraste

Ya que el *p_value* tiene un valor menor al nivel de significancia, se puede rechazar la hipótesis nula y aceptar la hipótesis alternativa. Por lo tanto, se tiene evidencia estadística suficiente para inferir que la capacidad pulmonar de los fumadores es menor que 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.

```
model <- lm(formula = AE ~ Tipo + edad + genero, df)
summary(model)
```

```
##
## Call:
## lm(formula = AE ~ Tipo + edad + genero, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05722 -0.24940 -0.00566  0.22740  1.62120
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.704758   0.135750  19.924  < 2e-16 ***
## TipoFL       0.337958   0.083609   4.042 7.10e-05 ***
## TipoFM       0.043769   0.084954   0.515    0.607
## TipoFP       0.395316   0.084249   4.692 4.50e-06 ***
## TipoNF       0.801520   0.079657  10.062  < 2e-16 ***
## TipoNI       0.423578   0.082998   5.103 6.69e-07 ***
## edad        -0.030162   0.002407 -12.533  < 2e-16 ***
## generoM     -0.007653   0.048702  -0.157    0.875
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3779 on 245 degrees of freedom
## Multiple R-squared:  0.5541, Adjusted R-squared:  0.5414
## F-statistic: 43.49 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.

- el *p_value* es de 2.26e-16, es decir, que podemos considerar que se ha obtenido una muy buena regresión ya que está muy por debajo del nivel de significación.
- la *recta de regresión* que se desprende de la información del modelo es *y = 2.704758 + 0.337958 * TipoFL + 0.043769 * TipoFM + 0.395316 * TipoFP + 0.801520 * TipoNF + 0.423578 * TipoNI - 0.030162 * edad - 0.007653 * generoM*
- para un nivel de significancia de 0.05, las variables TipoFM y generoM tienen un valor *Pr(>|t|)* mayor que el nivel de significancia, por lo que no son relevantes para el modelo.

6.3 Bondad del ajuste

Evaluad la calidad del modelo.

El *R-quared* tiene un valor de 0.5541, por lo que no se podría decir que es un modelo ajustado ya que está alejado del 1, pero tampoco es poco ajustado porque está alejado del 0.

6.4 Predicción

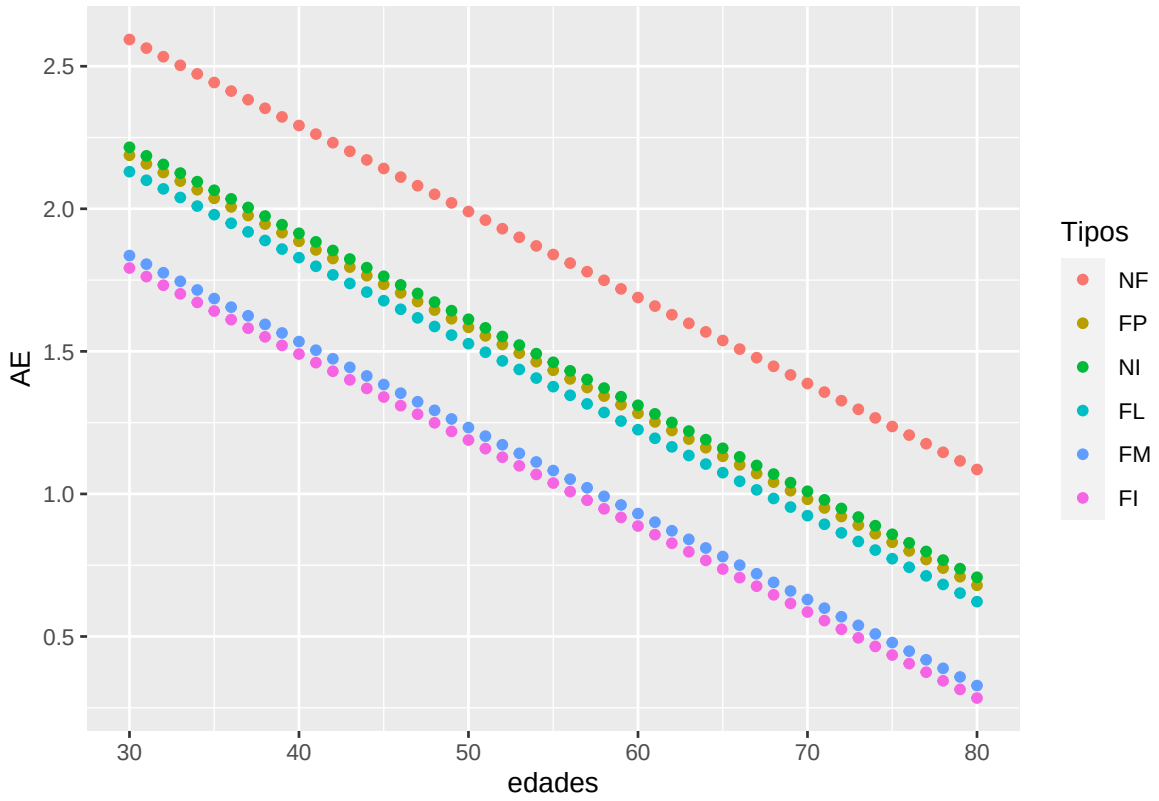
Realizad una predicción de la capacidad pulmonar para cada tipo de fumador desde los 30 años de edad hasta los 80 años de edad (podéis asumir género hombre). Mostrad una tabla con los resultados. Mostrad también visualmente la simulación.

```
names(tipos) <- tipos
edades <- 30:80
names(edades) <- edades
predictions <- data.frame(outer(edades, tipos, function(edad,
    tipo) {
    return(predict(model, data.frame(Tipo = tipo, edad = edad,
        genero = "M")))
}))
predictions["edades"] <- edades
predictions
```

```
##           NF        FP        NI        FL        FM        FI edades
## 30 2.593771 2.1875670 2.2158288 2.1302092 1.8360199 1.7922509     30
## 31 2.563609 2.1574052 2.1856670 2.1000474 1.8058581 1.7620891     31
## 32 2.533447 2.1272434 2.1555052 2.0698855 1.7756963 1.7319272     32
## 33 2.503285 2.0970816 2.1253433 2.0397237 1.7455345 1.7017654     33
## 34 2.473124 2.0669197 2.0951815 2.0095619 1.7153726 1.6716036     34
## 35 2.442962 2.0367579 2.0650197 1.9794001 1.6852108 1.6414418     35
## 36 2.412800 2.0065961 2.0348579 1.9492382 1.6550490 1.6112799     36
## 37 2.382638 1.9764343 2.0046960 1.9190764 1.6248872 1.5811181     37
## 38 2.352476 1.9462724 1.9745342 1.8889146 1.5947253 1.5509563     38
## 39 2.322314 1.9161106 1.9443724 1.8587528 1.5645635 1.5207945     39
## 40 2.292153 1.8859488 1.9142106 1.8285909 1.5344017 1.4906326     40
## 41 2.261991 1.8557870 1.8840487 1.7984291 1.5042399 1.4604708     41
## 42 2.231829 1.8256251 1.8538869 1.7682673 1.4740780 1.4303090     42
## 43 2.201667 1.7954633 1.8237251 1.7381055 1.4439162 1.4001472     43
## 44 2.171505 1.7653015 1.7935633 1.7079436 1.4137544 1.3699853     44
## 45 2.141344 1.7351397 1.7634014 1.6777818 1.3835926 1.3398235     45
## 46 2.111182 1.7049778 1.7332396 1.6476200 1.3534307 1.3096617     46
## 47 2.081020 1.6748160 1.7030778 1.6174582 1.3232689 1.2794999     47
## 48 2.050858 1.6446542 1.6729160 1.5872963 1.2931071 1.2493380     48
## 49 2.020696 1.6144924 1.6427541 1.5571345 1.2629453 1.2191762     49
## 50 1.990534 1.5843305 1.6125923 1.5269727 1.2327834 1.1890144     50
## 51 1.960373 1.5541687 1.5824305 1.4968109 1.2026216 1.1588526     51
## 52 1.930211 1.5240069 1.5522687 1.4666490 1.1724598 1.1286907     52
## 53 1.900049 1.4938451 1.5221068 1.4364872 1.1422980 1.0985289     53
## 54 1.869887 1.4636832 1.4919450 1.4063254 1.1121361 1.0683671     54
## 55 1.839725 1.4335214 1.4617832 1.3761636 1.0819743 1.0382053     55
## 56 1.809563 1.4033596 1.4316214 1.3460017 1.0518125 1.0080434     56
## 57 1.779402 1.3731978 1.4014595 1.3158399 1.0216507 0.9778816     57
## 58 1.749240 1.3430359 1.3712977 1.2856781 0.9914888 0.9477198     58
## 59 1.719078 1.3128741 1.3411359 1.2555163 0.9613270 0.9175580     59
## 60 1.688916 1.2827123 1.3109741 1.2253544 0.9311652 0.8873961     60
## 61 1.658754 1.2525505 1.2808122 1.1951926 0.9010034 0.8572343     61
```

```
## 62 1.628593 1.2223886 1.2506504 1.1650308 0.8708415 0.8270725     62
## 63 1.598431 1.1922268 1.2204886 1.1348690 0.8406797 0.7969107     63
## 64 1.568269 1.1620650 1.1903268 1.1047071 0.8105179 0.7667488     64
## 65 1.538107 1.1319032 1.1601649 1.0745453 0.7803561 0.7365870     65
## 66 1.507945 1.1017413 1.1300031 1.0443835 0.7501942 0.7064252     66
## 67 1.477783 1.0715795 1.0998413 1.0142217 0.7200324 0.6762634     67
## 68 1.447622 1.0414177 1.0696795 0.9840598 0.6898706 0.6461015     68
## 69 1.417460 1.0112559 1.0395176 0.9538980 0.6597088 0.6159397     69
## 70 1.387298 0.9810940 1.0093558 0.9237362 0.6295469 0.5857779     70
## 71 1.357136 0.9509322 0.9791940 0.8935744 0.5993851 0.5556161     71
## 72 1.326974 0.9207704 0.9490322 0.8634125 0.5692233 0.5254542     72
## 73 1.296812 0.8906086 0.9188703 0.8332507 0.5390615 0.4952924     73
## 74 1.266651 0.8604467 0.8887085 0.8030889 0.5088996 0.4651306     74
## 75 1.236489 0.8302849 0.8585467 0.7729271 0.4787378 0.4349688     75
## 76 1.206327 0.8001231 0.8283849 0.7427652 0.4485760 0.4048069     76
## 77 1.176165 0.7699613 0.7982230 0.7126034 0.4184142 0.3746451     77
## 78 1.146003 0.7397994 0.7680612 0.6824416 0.3882523 0.3444833     78
## 79 1.115841 0.7096376 0.7378994 0.6522798 0.3580905 0.3143215     79
## 80 1.085680 0.6794758 0.7077376 0.6221179 0.3279287 0.2841596     80
```

```
predictions.melt <- melt(predictions, id.vars = "edades")
ggplot(predictions.melt, aes(edades, value, colour = variable)) +
    geom_point() + ylab("AE") + labs(color = "Tipos")
```



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:

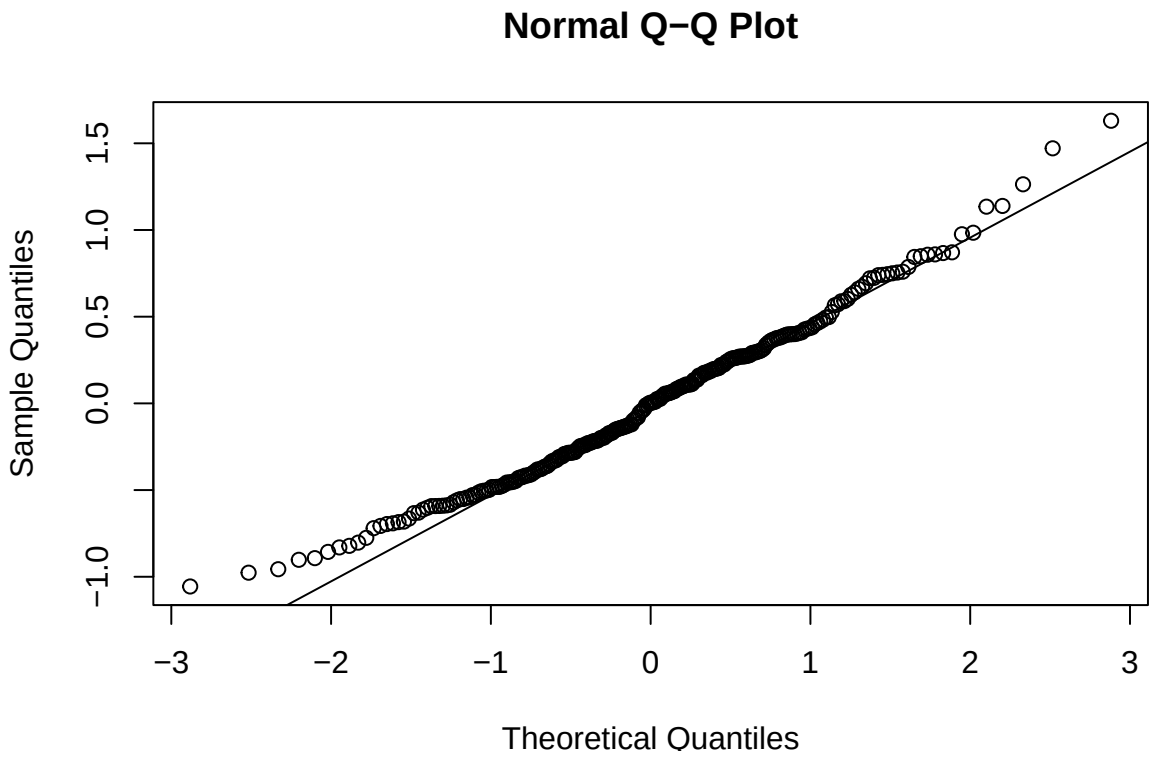
- ¿Existen diferencias entre la capacidad pulmonar (AE) entre los distintos tipos de fumadores/no fumadores?
- Si existen diferencias, ¿entre qué grupos están estas diferencias?

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 los residuos.

```
linear_model <- lm(AE ~ Tipo, df)
linear_residuals <- residuals(linear_model)
qqnorm(linear_residuals)
qqline(linear_residuals)
```



Se observa que la mayoría de los residuos se aproximan a la recta, por lo que no se ve un comportamiento que vaya en contra del supuesto de normalidad. Sin embargo, se procede al test de Shapiro-Wilk para confirmarlo.

```
shapiro.test(linear_residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  linear_residuals
## W = 0.9869, p-value = 0.02071
```

Como el *p_value* del test de Shapiro-Wilk es menor que el nivel de significancia del 5%, se rechaza la hipótesis nula y se acepta la hipótesis alternativa. En este caso, se tiene que la variable aleatoria que representa los errores del modelo no sigue una distribución normal.

7.2 Homocedasticidad: 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.

Nota: podéis usar tests como el de Levene o Bartlett test.

```
bartlett.test(AE ~ Tipo, data = df)
```

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

Como el *p_value* es mayor que el nivel de significancia del 5%, aceptamos la hipótesis nula de 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.

$$H_0 : \mu_{NF} = \mu_{FP} = \mu_{NI} = \mu_{FL} = \mu_{FM} = \mu_{FI} = \mu$$
$$H_1 : \exists \mu_j \neq \mu,\ j = \{NF, FP, NI, FL, FM, FI\}$$

7.4 Cálculo ANOVA

Podéis usar la función aov.

```
anova_model <- aov(AE ~ Tipo, data = df)
summarized_model <- summary(anova_model)
summarized_model
```

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

7.5 Interpretación

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

Dado que el *p_value* es mucho más pequeño que el nivel de significancia del 5%, se puede rechazar la hipótesis nula y aceptar la hipótesis alternativa. Esto es, que hay al menos un valor de la variable *Tipo* cuya media no es igual a las demás. Por tanto, *Tipo* es un factor significativo en el modelo para predecir *AE*.

7.6 Profundización 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.

Se obtiene la información a partir de la tabla de anova.

```
# Se obtienen los Sum Squares
SS <- summarized_model[[1]]$"Sum Sq"

# Primer elemento SSA, segundo SSE
SSB <- SS[1]
SSW <- SS[2]
SST <- SSB + SSW

# Obtener los grados de libertad
DFs <- summarized_model[[1]]$Df
df.SSB <- DFs[1]
df.SSW <- DFs[2]
```

Cálculo del F value

```
F <- (SSB/df.SSB)/(SSW/df.SSW)
cat("F value: ", F)
```

```
## F value:  17.87744
```

Cálculo del Valor crítico

```
critical <- qf(p = 0.05, df1 = df.SSB, df2 = df.SSW, lower.tail = FALSE)
cat("Critical value: ", critical)
```

```
## Critical value:  2.250576
```

Cálculo del P value

```
p_value <- pf(q = F, df1 = df.SSB, df2 = df.SSW, lower.tail = FALSE)
cat("P value: ", p_value)
```

```
## P value:  4.025786e-15
```

Los resultados obtenidos coinciden con la tabla de *aov*. El *P value* es más pequeño que el α, por lo que se confirma el rechazo a la hipótesis nula y se acepta la hipótesis alternativa (apartado 7.3).

La variable SST es la suma de los errores cuadrados totales, como se puede ver la suma de SSB y SSW. La variable SSW es lo mismo que SSE, que muestra la suma de los errores cuadrados, mientras que la variable SSB es lo mismo que SSA, que es la suma de los cuadrados de los tratamientos.

7.7 Fuerza de la relación

Calcular la fuerza de la relación e interpretar el resultado.

```
R2 <- SSB/SST
R2
```

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

El coeficiente de determinación es la proporción de variación de la variable AE frente al predictor *Tipo*, que para este caso asciende a un 26.57%.

8 Comparaciones múltiples

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

8.1 Test pairwise

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

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

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

- Como podemos observar el valor del *p_value* que hemos obtenido para el par FM-FI es mayor que un nivel de significancia del 5%, por lo que no tenemos evidencia estadística para rechazar la hipótesis de que ambos son similares.
- Como podemos observar el valor del *p_value* que hemos obtenido para el par FP-FL es mayor que un nivel de significancia del 5%, por lo que no tenemos evidencia estadística para rechazar la hipótesis de que ambos son similares.
- Como podemos observar el valor del *p_value* que hemos obtenido para el par NI-FL es mayor que un nivel de significancia del 5%, por lo que no tenemos evidencia estadística para rechazar la hipótesis de que ambos son similares.
- Como podemos observar el valor del *p_value* que hemos obtenido para el par NI-FP es mayor que un nivel de significancia del 5%, por lo que no tenemos evidencia estadística para rechazar la hipótesis de que ambos son similares.
- Para todos los demás pares tenemos valores de *p_value* menores que un nivel de significancia del 5%, por tanto se puede rechazar la hipótesis de que entre estas parejas se tiene una media similar. Esto nos indica que estos grupos son diferentes entre ellos.

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.

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

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  df$AE and df$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
```

Obtenemos los mismos resultados pero con valores más fáciles de interpretar, que tienen más fortaleza a la hora de rechazar la hipótesis.

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:

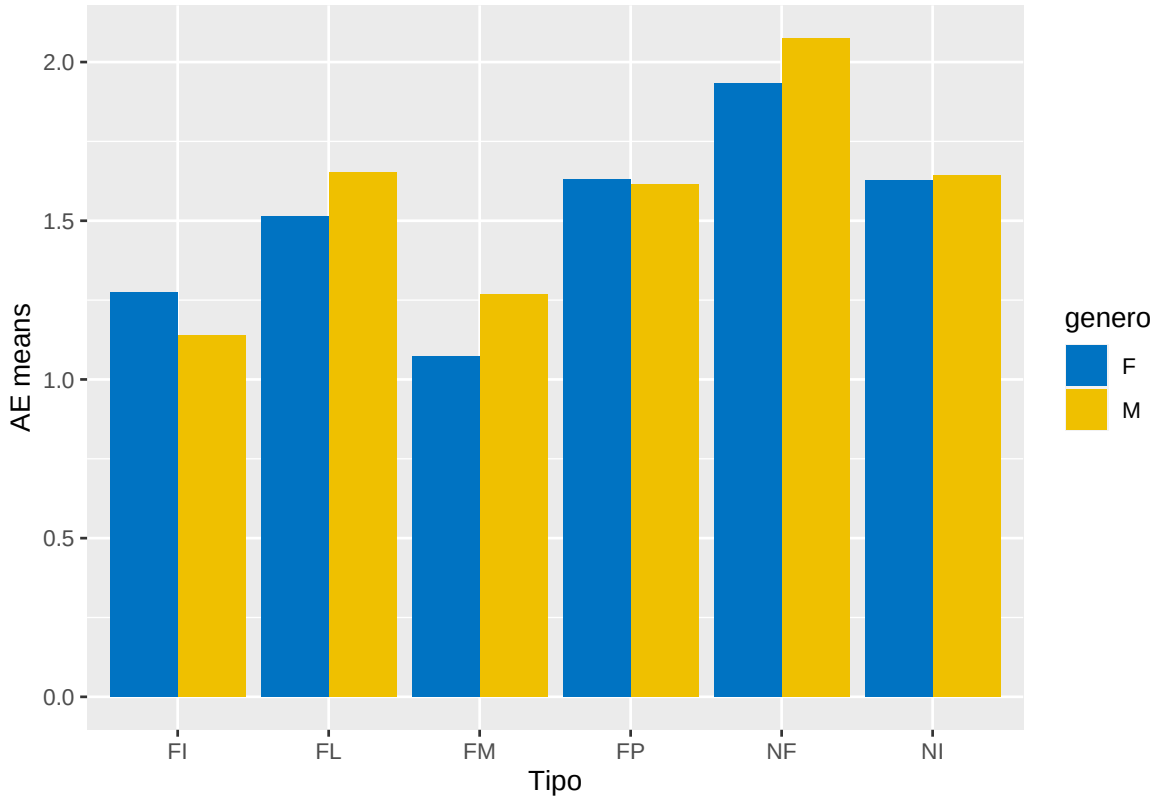
1. Agrupar el conjunto de datos por tipo de fumador y género y calcular la media de AE en cada grupo. Podéis usar las instrucciones group_by y summarise de la librería dplyr para realizar este proceso. Mostrar el conjunto de datos en forma de tabla, donde se muestre la media de cada grupo según el género y tipo de fumador.

```
resumen <- df %>%
    group_by(genero, Tipo) %>%
    summarise(means = mean(AE), counts = length(AE))
resumen
```

```
## # A tibble: 12 x 4
## # Groups:   genero [2]
##    genero Tipo  means counts
##    <fct>  <fct> <dbl>  <int>
##  1 F      FI     1.27     24
##  2 F      FL     1.51     28
##  3 F      FM     1.07     22
##  4 F      FP     1.63     18
##  5 F      NF     1.93     29
##  6 F      NI     1.63     23
##  7 M      FI     1.14     17
##  8 M      FL     1.65     13
##  9 M      FM     1.27     17
## 10 M      FP     1.61     22
## 11 M      NF     2.07     21
## 12 M      NI     1.64     19
```

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(resumen, aes(fill = genero, x = Tipo, y = means)) + geom_bar(position = "dodge",
    stat = "identity") + scale_color_manual(values = c("#0073C2FF",
    "#EFC000FF")) + scale_fill_manual(values = c("#0073C2FF",
    "#EFC000FF")) + ylab("AE means")
```



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.

Para los valores NI y FP, la diferencia de la media de AE para el género M y F no parece ser significativa. Sin embargo, sí puede verse en la gráfica anterior que para valores como FI, FL, FM y NF puede existir una diferencia considerable entre estas medias.

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.

```
multi_anova <- aov(AE ~ Tipo * genero, data = df)
summary(multi_anova)
```

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

9.3 Interpretación

Interpretad el resultado.

A priori se puede decir que *genero* no tiene evidencia estadística suficiente para afirmar que es significativa para el modelo, eso ya que el *p_value* es mayor que un nivel de significancia de 5%. Esto también le ocurre a la interacción. Por tanto, es mejor utilizar un anota de un único factor en el que solamente se toma el *Tipo*.

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 y conclusión
1	¿Se observan diferencias en la capacidad pulmonar en relación al género? Interpretación desde gráficas.	Las distribuciones parecen centradas en la media del intervalo [0;3], aunque sí existen más casos atípicos cuando el género es F.
2	¿Se observan diferencias significativas entre los intervalos de confianza de la capacidad pulmonar de mujeres y hombres?	Ambos intervalos son relativamente similares puesto que varían recién en la segunda cifra decimal,ambos rondan el 1.55 como valor central.
3	¿Se observan diferencias significativas en la capacidad pulmonar de mujeres y hombres? Interpretación a través de contraste de hipótesis.	Como el *p_value* es mayor que el nivel de significancia, se debe aceptar la hipótesis nula porque nohay evidencia suficiente para poder descartarla. Por lo tanto, lo único que se puede decir es que la capacidad pulmonar de ambos grupos se muestra igual.
4	¿Podemos afirmar que la capacidad pulmonar de los fumadores es inferior a la de no fumadores?	Ya que el *p_value* tiene un valor menor al nivel de significancia, se puede rechazar la hipótesis nula y aceptar la hipótesis alternativa. Por lo tanto, se tiene evidencia estadística suficiente para inferir que la capacidad pulmonar de los fumadores es menor que la de los no fumadores.
5	¿Existen diferencias entre la capacidad pulmonar (AE) entre los distintos tipos de fumadores/no fumadores? Si existen diferencias, ¿entre qué grupos están estas diferencias?	Dado que el *p_value* es mucho más pequeño que el nivel de significancia del 5%, se puede rechazar la hipótesis nula y aceptar la hipótesis alternativa. Esto es, que hay al menos un valor de la variable *Tipo* cuya media no es igual a las demás. Por tanto, *Tipo* es un factor significativo en el modelo para predecir *AE*. La diferencia está en los pares FL-FI, FP-FI, NF-FI, NI-FI, FM-FL, NF-FL, FP-FM, NF-FM, NI-FM, NF-FP, NI-NF.

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.

A partir de toda la información obtenida durante la realización de este informe, puede resumirse brevemente destacando la vinculación que existe entre el tipo de fumador y la capacidad pulmonar de los individuos. Incluso se puede ir un poco más allá pudiendo resaltar la estrecha relación entre los tipos específicos de individuo con la capacidad pulmonar.

La capacidad pulmonar presenta los mayores valores de la serie para aquellos individuos de la muestra que están clasificados como No Fumador (NF), mientras que los más bajos están presentes en los individuos clasificados como Fumador Moderado (FM) y Fumador Intensivo (FI), lo cual es observable en un contexto real.

En cuanto a género se refiere, no se ha determinado un tipo de vínculo que pueda dar lugar a una relación entre el género y la capacidad pulmonar.