PEC 4 - Estadística avanzada

Gloria Manresa

2023-01-27

Índice

```
# Librerías
library(ggplot2)
library(dplyr)
library(kableExtra)
```

1. Preprocesado

Cargamos el archivo de datos y lo guardamos bajo el nombre "dat". Sabemos que el delimitador de los datos es punto y coma por lo que así lo especificamos.

Mostramos las primeras y las últimas líneas para asegurarnos de que se han cargado todas las entradas y comprobamos que el número de entradas en la base da datos original coincide con el número de líneas cargadas en R.

```
dat <- read.csv('Fumadores.csv', sep=";")
knitr::kable(head(dat,3))
```

AE	Tipo	genero	edad
1.871878	NF	M	54
1.91312	NF	F	60
2.58114	NF	M	40

```
knitr::kable(tail(dat,3))
```

	AE	Tipo	genero	edad
251	1,714654	FI	F	53
252	1.205738	FI	F	53
253	1.49247	FI	F	37

Comprobamos el tipo de datos con el que R ha interpretado cada variable:

```
knitr::kable(sapply(dat, class),col.names="Tipo de datos")
```

	Tipo de datos
AE	character
Tipo	character
genero	character
edad	integer

- **AE**

Buscando inconsistencias en los datos nos damos cuentas que la variable AE presenta dos delimitadores (punto en algunas entradas y coma en otras). Se corrige para que el único delimitador sea el punto.

Cogemos un dato como referencia y para la posterior comprobación.

```
dat$AE[251]
```

```
## [1] "1,714654"
```

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

Confirmamos que se ha realizado el cambio correctamente:

```
dat$AE[251]
```

```
## [1] "1.714654"
```

A continuación modificamos el tipo de la variables AE de caracter a numeric.

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

- **Tipo**

Estudiamos las diferentes entradas que puede tomar la variable Tipo:

```
table(dat$Tipo)
```

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

Observamos la necesidad de pasar a mayúsculas todas las entradas y de borrar espacios:

```
dat$Tipo <- trimws(toupper(dat$Tipo))
table(dat$Tipo)
```

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

A continuación cambiamos el tipo de la variable de character a factor.

```
dat$Tipo <- as.factor(dat$Tipo)
```

- **Género**

Estudiamos las diferentes entradas que puede tomar la variable "genero":

```
table(dat$genero)
```

```
##
##   F   M
## 144 109
```

Cambiamos el tipo de variable de character a factor:

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

- **Edad**

Estudiamos los valores que toma la variable edad:

```
summary(dat$edad)
```

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

Comprobamos que no existen valores faltantes:

```
table(is.na(dat$edad))
```

```
##
## FALSE
##   253
```

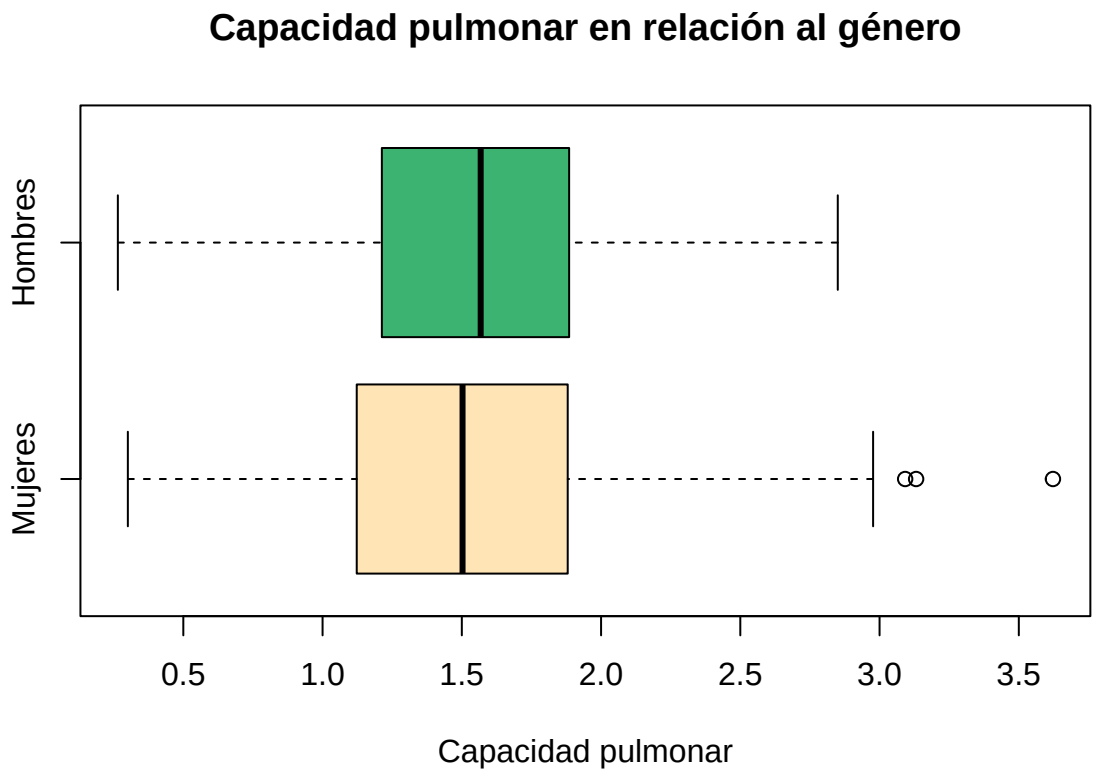
El tipo de la variable es integer, que es correcto.

2. Análisis descriptivo de la muestra

2.1 Capacidad pulmonar y género

El diagrama de cajas siguiente muestra la distribución de la variable capacidad pulmonar (AE) en el caso de los hombres y de las mujeres por separado.

```
boxplot(dat$AE~dat$genero,
main = "Capacidad pulmonar en relación al género",
xlab = "Capacidad pulmonar",
ylab = NULL,
col=c("moccasin","mediumseagreen"),
horizontal = T,
names = c("Mujeres","Hombres") )
```



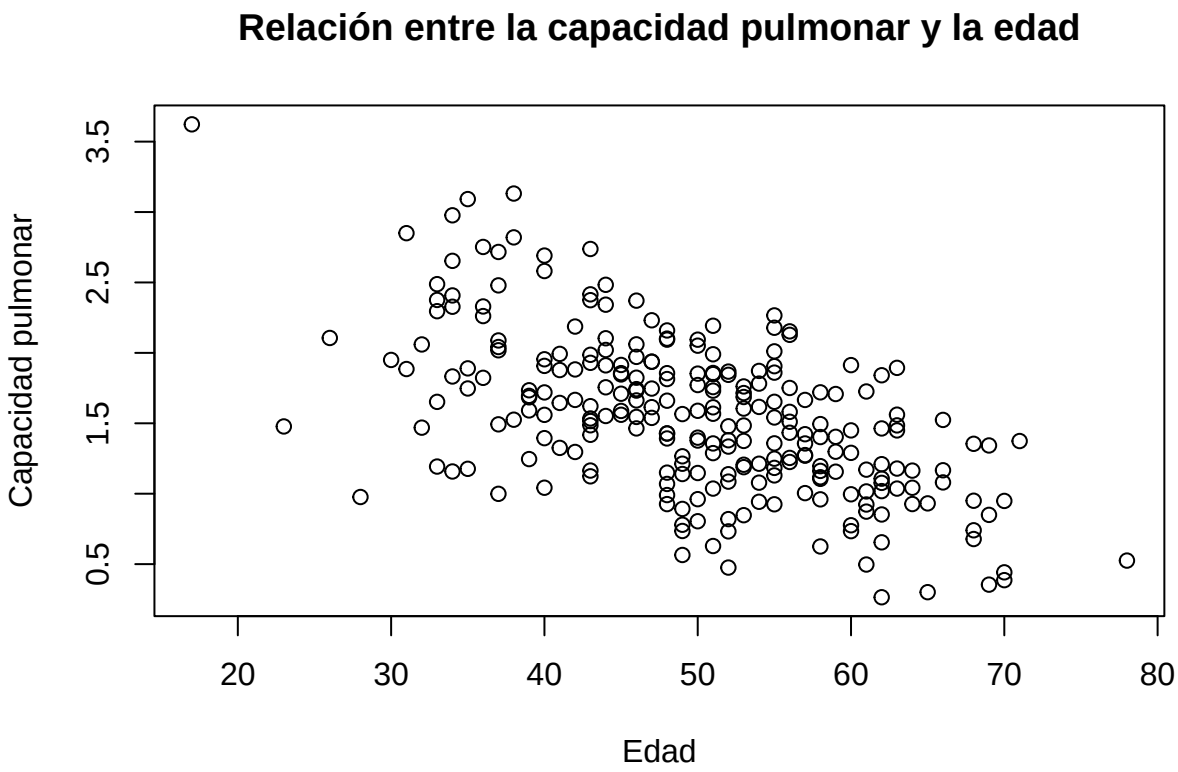
Se observa que la media de la capacidad pulmonar en las mujeres es ligeramente inferior a la de los hombres. Por el contrario el límite inferior es ligeramente inferior en los hombres.

Los valores atípicos los encontramos en el caso de las mujeres y en el límite superior.

2.2 Capacidad pulmonar y edad

El gráfico siguiente muestra la relación entre la capacidad pulmonar y la edad.

```
plot(dat$edad,dat$AE,
     main = "Relación entre la capacidad pulmonar y la edad",
     xlab = "Edad",
     ylab = "Capacidad pulmonar"
)
```



Se observa cierta tendencia donde la capacidad pulmonar se reduce con la edad.

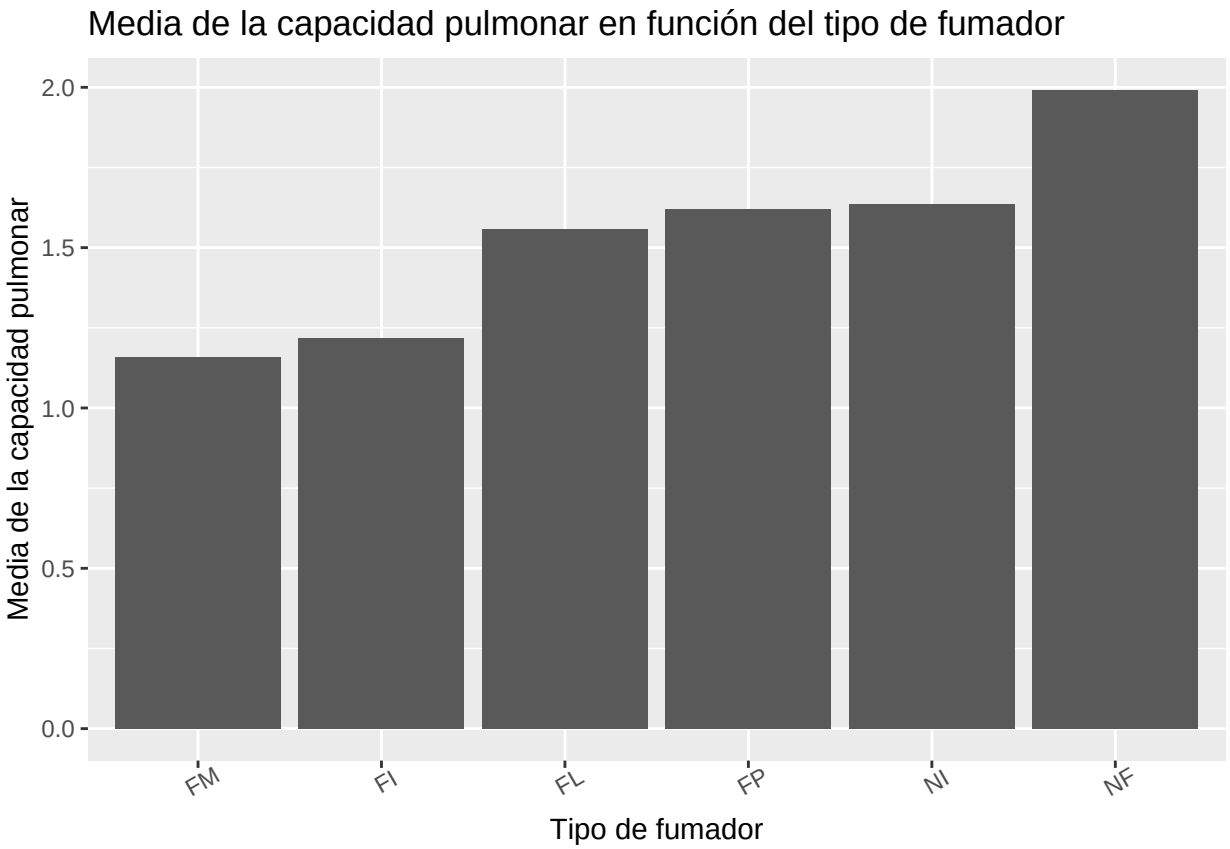
2.3 Tipos de fumadores y capacidad pulmonar.

A continuación se muestra el número de personas en cada tipo de fumador y las medias de la capacidad pulmonar según el tipo de fumador.

```
# Group by
df <- dat %>% group_by(Tipo) %>%
  summarise(mean_AE=mean(AE),n=n())

# Convert to dataframe
df1 <- df %>% as.data.frame()
#df1$num_tipo <- length(dat$Tipo[dat$Tipo %in% df1$Tipo])
knitr::kable(df1,col.names=c("Tipo de fumador","Media capacidad pulmonar","Número de personas por Tipo")
```

Tipo de fumador	Media capacidad pulmonar	Número de personas por Tipo
FI	1.217035	41
FL	1.556475	41
FM	1.157442	39
FP	1.620952	40
NF	1.992625	50
NI	1.634737	42

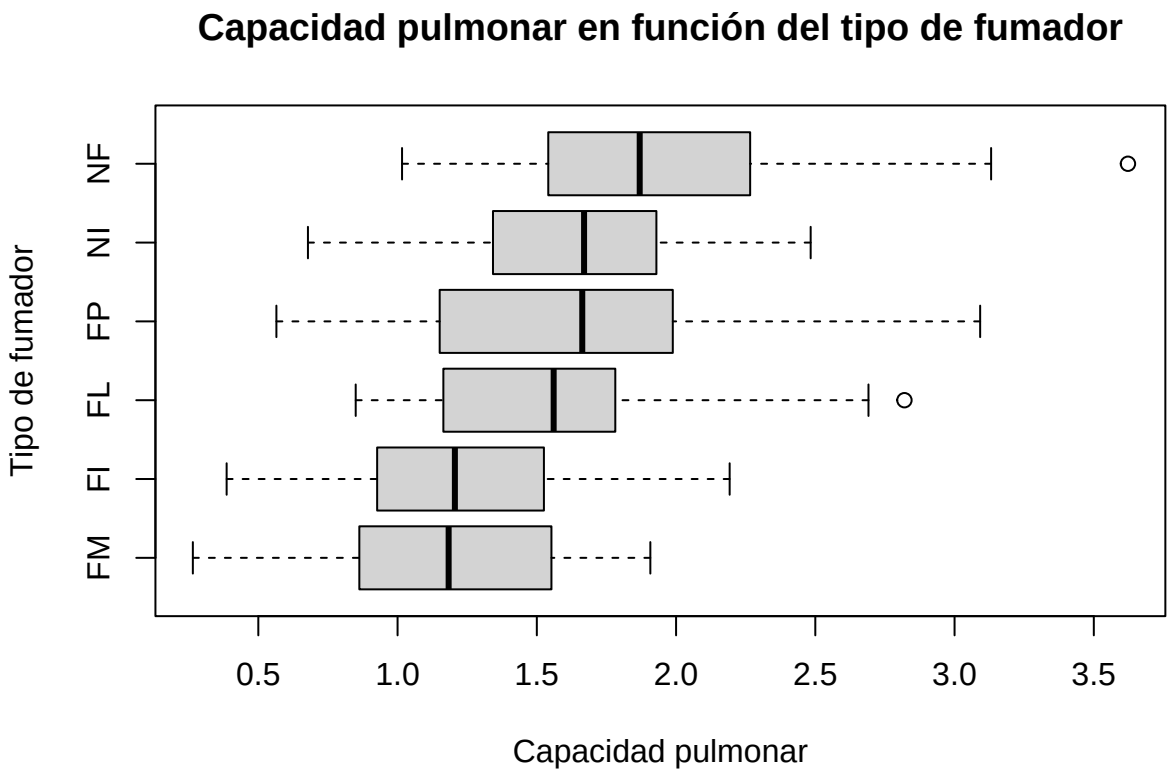


Observamos del gráfico anterior que los tipos de fumadores moderados e intensivos presentan las medias de capacidad pulmonar más bajas. En orden les siguen los fumadores ligeros, fumadores pasivos y fumadores que no inhalan. Por último y con bastante diferencia se encuentran los no fumadores.

A continuación se estudia la distribución de la capacidad pulmonar en función del tipo de fumador, ordenado según la mediana en este caso.

```
# Ordemamos de más bajo a más alto según la mediana
medianas <- reorder(dat$Tipo, dat$AE, median)

boxplot(dat$AE ~ medianas,
        main = "Capacidad pulmonar en función del tipo de fumador",
      xlab = "Capacidad pulmonar",
      ylab = "Tipo de fumador",
      horizontal = T)
```



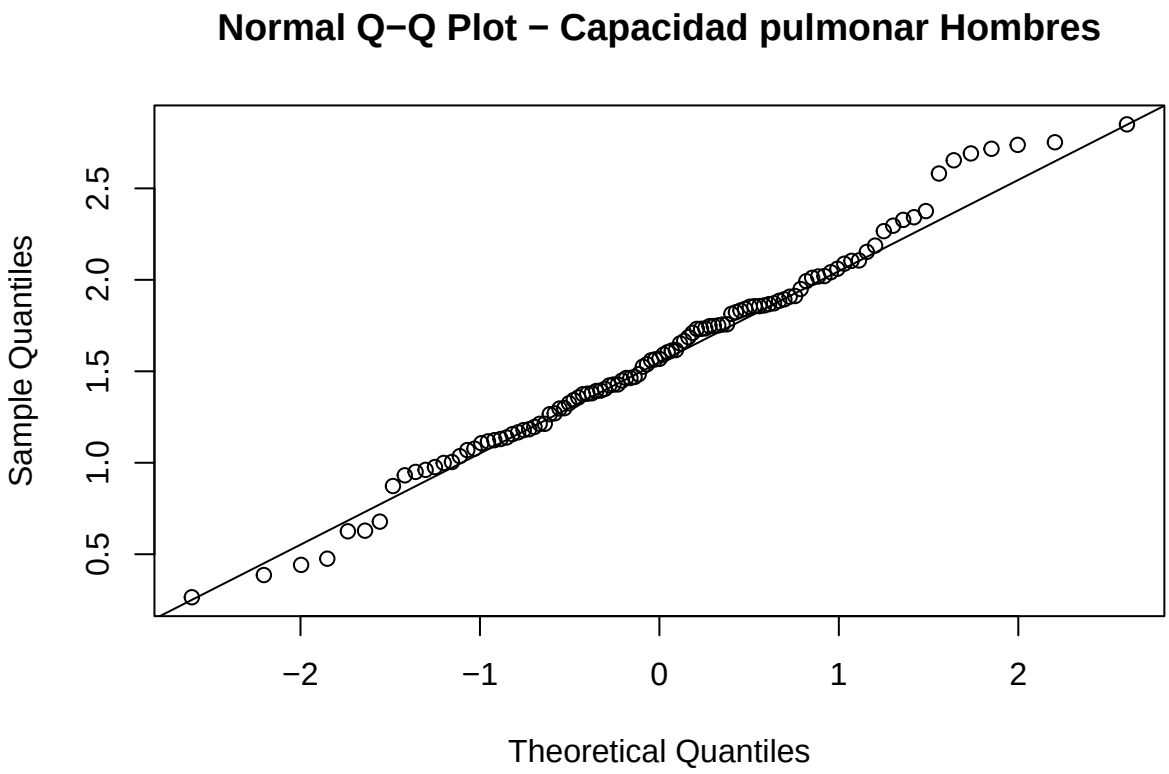
En el diagramas de caja anterior sacamos diferentes conclusiones:

- Los fumadores moderados son los que presentan los datos de capacidad pulmonar inferior, incluso menor a los fumadores intensivos.
- La capacidad pulmonar correspondiente al límite inferior de los no fumadores corresponde a una capacidad pulmonar que se encuentra en el rango intercuartílico de los fumadores moderados e intensivos.
- Los fumadores pasivos presentan unos datos muy dispersos. Su límite inferior es incluso menor a los fumadores que no inhalan y el límite superior casi alcanza el límite superior de los no fumadores.

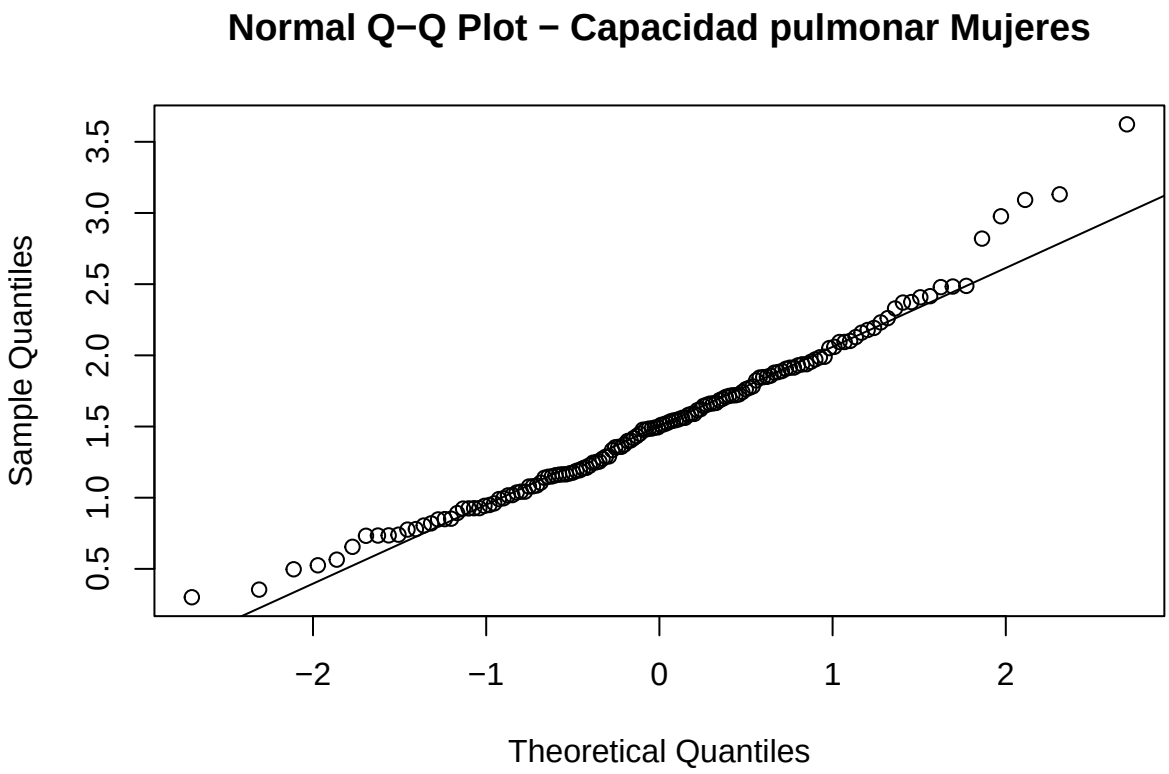
3. Intervalo de confianza de la capacidad pulmonar

Comprobamos si podemos asumir la normalidad de los datos a partir de las gráficas Q-Q de la variable AE en los hombres y en las mujeres.

```
qqnorm(dat$AE[dat$genero=="M"],main="Normal Q-Q Plot - Capacidad pulmonar Hombres")
qqline(dat$AE[dat$genero=="M"])
```




```
qqnorm(dat$AE[dat$genero=="F"],main="Normal Q-Q Plot - Capacidad pulmonar Mujeres")
qqline(dat$AE[dat$genero=="F"])
```



En ambas gráficas observamos que los datos se asemejan bastante a una distribución normal por lo que asumiremos la normalidad de los datos.

Con respecto a la varianza, la estimaremos a partir de la varianza de la muestra por lo que seguiremos una distribución t de Student con n-1 grados de libertad.

A continuación escribimos la función para el cálculo del intervalo de confianza sobre la media con varianza desconocida.

```
IC <- function(x,NC){
alfa <- 1-NC
sd <- sd(x)
n <- length(x)
SE <- sd/sqrt(n)
z <- qt(alfa/2,df=n-1,lower.tail=FALSE)
L <- mean(x)-z*SE
U <- mean(x)+z*SE
return(round(c(L,U),2))
}
```

A continuación se calcula el intervalo de confianza de la media poblacional de la variable “capacidad pulmonar” de las mujeres con un intervalo de confianza del 95%.

```
IC_AE_F <- IC(dat$AE[dat$genero=="F"], 0.95)
IC_AE_F
```

```
## [1] 1.43 1.62
```

En el caso de los hombres:

```
IC_AE_M <- IC(dat$AE[dat$genero=="M"], 0.95)
IC_AE_M
```

```
## [1] 1.48 1.69
```

De estos intervalos de confianza interpretamos que, si se realiza un muestreo elevado de la población, el 95% de los intervalos obtenidos contendrían el valor de la media poblacional de la capacidad pulmonar de hombres/mujeres.

Observamos que el intervalo de los hombres es ligeramente más amplio al de las mujeres para el mismo nivel de confianza. En el caso de los hombres el intervalo corresponde a valores mayores de capacidad pulmonar.

No se observan diferencias significativas ya que observamos que los resultados obtenidos comparten más de la mitad del intervalo.

4. Diferencias en capacidad pulmonar entre mujeres y hombres

En este apartado se aplicará un contraste de hipótesis para evaluar si existen diferencias significativas entre la capacidad pulmonar de mujeres y de hombres.

La pregunta de investigación que tratamos de responder es la siguiente:

¿Presentan los hombres una capacidad pulmonar significativamente diferente a las mujeres?

4.1. Hipótesis

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

4.2. Contraste

Estamos ante dos poblaciones independientes (hombres y mujeres) donde asumimos distribución normal ya que según el teorema del límite central podemos asumir la normalidad de los datos en el caso de que todas las muestras sean superiores a 30, como es nuestro caso.

Puesto que la pregunta es si las capacidades pulmonares son diferentes se trata de un test bilateral.

Nos queadaría realizar el test de igualdad de varianzas para determinar el test a aplicar.

```
var.test(dat$AE[dat$genero=='M'],dat$AE[dat$genero=='F'])
```

```
##
##  F test to compare two variances
##
## data:  dat$AE[dat$genero == "M"] and dat$AE[dat$genero == "F"]
## F = 0.86133, num df = 108, denom df = 143, p-value = 0.4152
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6066144 1.2339167
## sample estimates:
## ratio of variances
##           0.861326
```

Podemos asumir que las varianzas son iguales puesto que el valor observado cae dentro del intervalo de aceptación del test. El valor p es mayor que el valor de significancia por lo que confirmamos que las varianzas son iguales con un nivel de confianza del 95%.

Con esto confirmamos que el test a realizar es el test sobre la media de dos poblaciones independientes con varianza desconocida igual.

4.3. Cálculo

```
# Test sobre la media de dos poblaciones independientes con varianza desconocida igual,
# test bilateral:
test4.3 <- function(mu1,mu2,alfa){
S <- sqrt(((length(mu1)-1)*(sd(mu1))^2+(length(mu2)-1)*(sd(mu2))^2)/(length(mu1)+length(mu2)-2))
tobs <- (mean(mu1)-mean(mu2))/(S*sqrt((1/length(mu1))+(1/length(mu2))))   #Estadístico
tcritL <- qt( alfa/2, (length(mu1))+(length(mu2))-2)
tcritU <- qt( 1-alfa/2, (length(mu1))+(length(mu2))-2)
pvalue <-pt( abs(tobs), df=(length(mu1))+(length(mu2))-2, lower.tail=FALSE)*2
return(c(tobs,tcritL,tcritU,pvalue))
}
```

Cálculo con un nivel de confianza del 95%.

```
calc4.3 <- test4.3(dat$AE[dat$genero=='M'],dat$AE[dat$genero=='F'],0.05)
calc4.3
```

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

```
#tobs,tcritL,tcritU,pvalue
```

El valor observado se encuentra dentro de la zona de aceptación por lo que todo apunta a que deberíamos aceptar la hipótesis nula. El valor p es superior al valor de significancia por lo que confirmamos que podemos aceptar la hipótesis nula.

4.4. Interpretación

Según los resultados del test concluimos que no existen diferencias significativas en la capacidad pulmonar entre los hombres y las mujeres con un nivel de confianza del 95%.

5. Diferencias en la capacidad pulmonar entre fumadores y no fumadores.

En este apartado se aplicará un contraste de hipótesis para evaluar si la capacidad pulmonar de los fumadores es inferior a los no fumadores.

La pregunta de investigación que tratamos de responder es la siguiente:

¿Presentan las personas fumadoras una capacidad pulmonar menor a las personas no fumadoras?

5.1. Hipótesis

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 < \mu_2$

5.2. Contraste

Estamos ante dos poblaciones independientes (personas fumadoras y personas no fumadoras) donde asumimos distribución normal ya que según el teorema del límite central podemos asumir la normalidad de los datos en el caso de que todas las muestras sean superiores a 30, como es nuestro caso.

Se trata de un test unilateral por la izquierda.

Nos queadaría realizar el test de igualdad de varianzas para determinar el test a aplicar.

```
fum <- c('NI','FL','FM','FI')
nofum <- c('NF','FP')
var.test(dat$AE[dat$Tipo %in% fum],dat$AE[dat$Tipo %in% nofum])
```

```
##
##  F test to compare two variances
##
## data:  dat$AE[dat$Tipo %in% fum] and dat$AE[dat$Tipo %in% nofum]
## F = 0.79901, num df = 162, denom df = 89, p-value = 0.2187
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.5477312 1.1426311
## sample estimates:
## ratio of variances
##          0.7990148
```

El valor observado cae dentro del intervalo de aceptación del test y el p valor es mayor al valor de significancia por lo que podemos confirmar que las varianzas son iguales con un nivel de confianza del 95%.

Con esto confirmamos que el test a realizar es el test sobre la media de dos poblaciones independientes con varianza desconocida igual.

5.3. Preparación de los datos

Según se indica en el ejercicio tomaremos los valores de los no fumadores y los fumadores pasivos como la muestra de los no fumadores. El resto se encontraran en la muestra de fumadores.

Esto es:

```
datfum <- dat$AE[dat$Tipo %in% fum]
datnotfum <- dat$AE[dat$Tipo %in% nofum]
```

5.4. Cálculos

```
# Test sobre la media de dos poblaciones independientes con varianzas desconocidas iguales,
#test unilateral:
test5.4 <- function(mu1,mu2,alfa){
S <- sqrt(((length(mu1)-1)*(sd(mu1))^2+(length(mu2)-1)*(sd(mu2))^2)/(length(mu1)+length(mu2)-2))
tobs <- (mean(mu1)-mean(mu2))/(S*sqrt((1/length(mu1))+(1/length(mu2))))   #Estadístico
tcrit <- qt( alfa, (length(mu1))+(length(mu2))-2)
pvalue <-pt( abs(tobs), df=(length(mu1))+(length(mu2))-2, lower.tail=FALSE)
return(c(tobs,tcrit,pvalue))
}
```

En los cálculos realizados a continuación observamos que el valor de p es más pequeño que el nivel de significancia por lo que deberíamos rechazar la hipótesis nula y aceptar la hipótesis alternativa.

```
calc5.4 <- test5.4(datfum,datnotfum,0.05)
calc5.4
```

```
## [1] -6.329761e+00 -1.650947e+00  5.613478e-10
```

```
#tobs,tcrit,pvalue
```

5.5. Interpretación

El test acepta la hipótesis alternativa, esto es que los fumadores presentan una capacidad pulmonar menor a la de los no fumadores.

6. Análisis de regresión lineal

Se realizará un análisis de regresión lineal para investigar la relación entre la capacidad pulmonar y el resto de las variables.

6.1. Cálculo

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

```
##
## Call:
## lm(formula = AE ~ Tipo + genero + edad, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05421 -0.25126 -0.00321  0.23288  1.03947
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.741411   0.128797  21.285  < 2e-16 ***
## TipoFL       0.338459   0.080850   4.186 3.96e-05 ***
## TipoFM       0.046357   0.082133   0.564    0.573
## TipoFP       0.394342   0.081470   4.840 2.30e-06 ***
## TipoNF       0.781808   0.077004  10.153  < 2e-16 ***
## TipoNI       0.423523   0.080259   5.277 2.89e-07 ***
## generoM     -0.002321   0.047033  -0.049    0.961
## edad        -0.030951   0.002276 -13.601  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3655 on 245 degrees of freedom
## Multiple R-squared:  0.583,  Adjusted R-squared:  0.5711
## F-statistic: 48.94 on 7 and 245 DF,  p-value: < 2.2e-16
```

6.2. Interpretación

Observamos a partir del valor p las variables que son significativas en el modelo. Observamos que el género, no es significativa ya que tiene un valor p mayor al valor de significancia. También observamos que la edad y el hecho de no ser fumador influyen bastante en el modelo.

6.3. Bondad de ajuste

El R cuadrado nos indica el porcentaje de variabilidad de la variable predicha que es explicada por la variable predictora. En este caso se trata del 58'3%.

6.4. Predicción

A continuación realizamos las predicciones para cada tipo de fumador.

```
# Fumadores que no inhalan
dfni <- data.frame(
  'Tipo' = c(rep('NI',51)),
  'genero' = c(rep('M',51)),
  'edad' = 30:80
  )

predNI <- predict(modelo1,dfni)

# Fumadores pasivos
```

```
dffp <- data.frame(
  'Tipo' = c(rep('FP',51)),
  'genero' = c(rep('M',51)),
  'edad' = 30:80
  )

predFP <- predict(modelo1,dffp)

# Fumadores moderados
```

```
dffm <- data.frame(
  'Tipo' = c(rep('FM',51)),
  'genero' = c(rep('M',51)),
  'edad' = 30:80
  )

predFM <- predict(modelo1,dffm)

# Fumadores ligeros
```

```
dffl <- data.frame(
  'Tipo' = c(rep('FL',51)),
  'genero' = c(rep('M',51)),
  'edad' = 30:80
  )
```

```
predFL <- predict(modelo1,dffl)

# Fumadores intensivos
```

```
dffi <- data.frame(
  'Tipo' = c(rep('FI',51)),
  'genero' = c(rep('M',51)),
  'edad' = 30:80
  )

predFI <- predict(modelo1,dffi)

# No fumadores
```

```
dfnf <- data.frame(
  'Tipo' = c(rep('NF',51)),
  'genero' = c(rep('M',51)),
  'edad' = c(30:80)
  )

predNF <- predict(modelo1,dfnf)
```

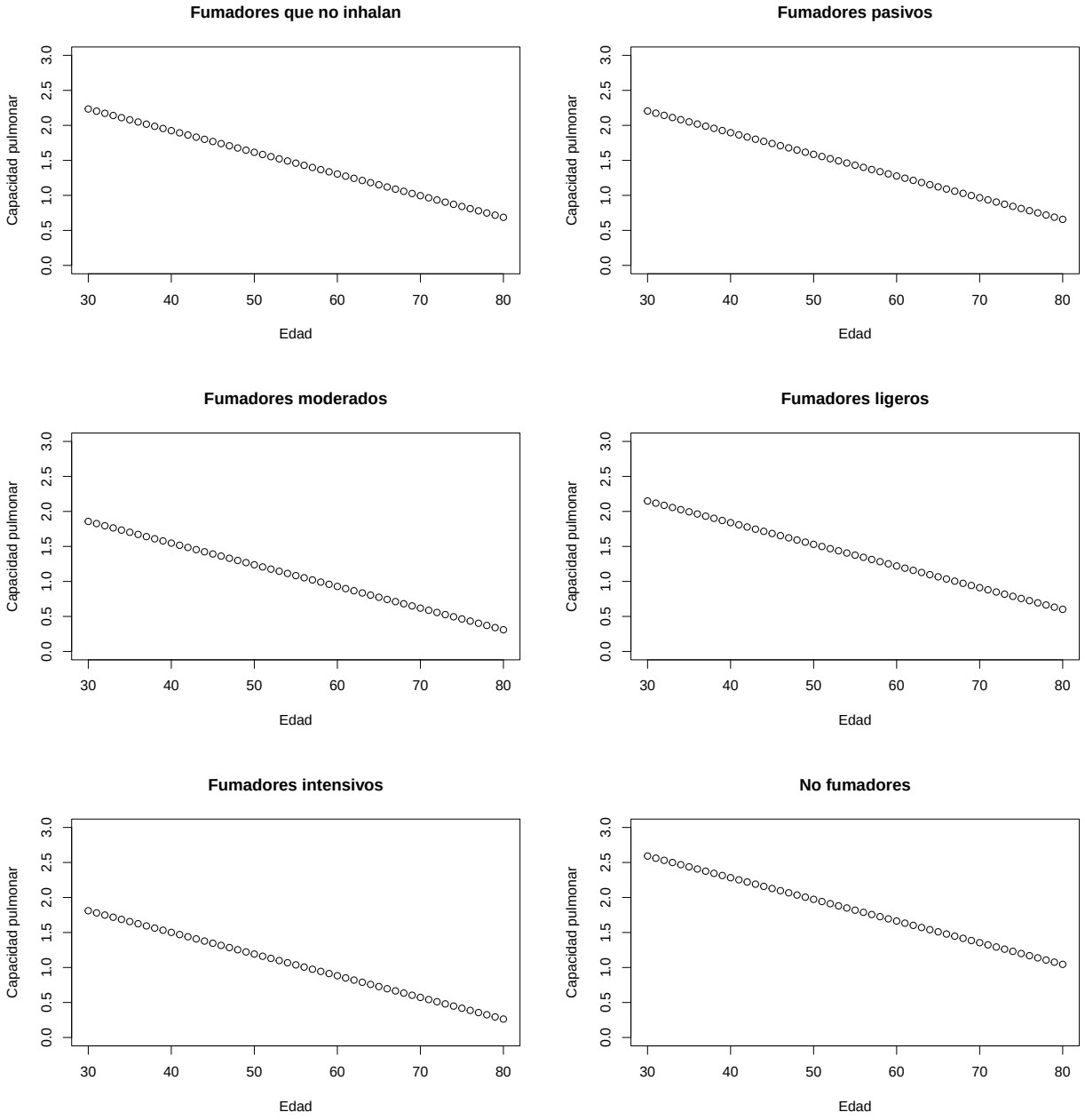
- **Tabla con los resultados**

```
prediccion <- data.frame(
  'Edad' = 30:80,
  'NI' = c(data.frame(predNI)[,1]),
  'FP' = c(data.frame(predFP)[,1]),
  'FM' = c(data.frame(predFM)[,1]),
  'FL' = c(data.frame(predFL)[,1]),
  'FI' = c(data.frame(predFI)[,1]),
  'NF' = c(data.frame(predNF)[,1])
)
kable(prediccion)
```

Edad	NI	FP	FM	FL	FI	NF
30	2.2340704	2.2048896	1.8569047	2.1490066	1.8105474	2.592355
31	2.2031189	2.1739381	1.8259533	2.1180552	1.7795960	2.561403
32	2.1721675	2.1429867	1.7950019	2.0871038	1.7486446	2.530452
33	2.1412161	2.1120353	1.7640504	2.0561524	1.7176932	2.499501
34	2.1102647	2.0810839	1.7330990	2.0252009	1.6867417	2.468549
35	2.0793133	2.0501325	1.7021476	1.9942495	1.6557903	2.437598
36	2.0483619	2.0191811	1.6711962	1.9632981	1.6248389	2.406646
37	2.0174104	1.9882296	1.6402448	1.9323467	1.5938875	2.375695
38	1.9864590	1.9572782	1.6092934	1.9013953	1.5629361	2.344744
39	1.9555076	1.9263268	1.5783419	1.8704439	1.5319847	2.313792
40	1.9245562	1.8953754	1.5473905	1.8394924	1.5010332	2.282841
41	1.8936048	1.8644240	1.5164391	1.8085410	1.4700818	2.251889
42	1.8626534	1.8334726	1.4854877	1.7775896	1.4391304	2.220938
43	1.8317019	1.8025211	1.4545363	1.7466382	1.4081790	2.189986
44	1.8007505	1.7715697	1.4235849	1.7156868	1.3772276	2.159035
45	1.7697991	1.7406183	1.3926334	1.6847354	1.3462762	2.128084
46	1.7388477	1.7096669	1.3616820	1.6537839	1.3153247	2.097132
47	1.7078963	1.6787155	1.3307306	1.6228325	1.2843733	2.066181
48	1.6769449	1.6477641	1.2997792	1.5918811	1.2534219	2.035229
49	1.6459934	1.6168126	1.2688278	1.5609297	1.2224705	2.004278
50	1.6150420	1.5858612	1.2378764	1.5299783	1.1915191	1.973327
51	1.5840906	1.5549098	1.2069249	1.4990269	1.1605677	1.942375
52	1.5531392	1.5239584	1.1759735	1.4680754	1.1296162	1.911424
53	1.5221878	1.4930070	1.1450221	1.4371240	1.0986648	1.880472
54	1.4912364	1.4620556	1.1140707	1.4061726	1.0677134	1.849521
55	1.4602849	1.4311041	1.0831193	1.3752212	1.0367620	1.818569
56	1.4293335	1.4001527	1.0521679	1.3442698	1.0058106	1.787618
57	1.3983821	1.3692013	1.0212164	1.3133184	0.9748592	1.756667
58	1.3674307	1.3382499	0.9902650	1.2823669	0.9439077	1.725715
59	1.3364793	1.3072985	0.9593136	1.2514155	0.9129563	1.694764
60	1.3055279	1.2763471	0.9283622	1.2204641	0.8820049	1.663812
61	1.2745764	1.2453956	0.8974108	1.1895127	0.8510535	1.632861
62	1.2436250	1.2144442	0.8664594	1.1585613	0.8201021	1.601910
63	1.2126736	1.1834928	0.8355079	1.1276099	0.7891507	1.570958
64	1.1817222	1.1525414	0.8045565	1.0966584	0.7581992	1.540007
65	1.1507708	1.1215900	0.7736051	1.0657070	0.7272478	1.509055
66	1.1198194	1.0906386	0.7426537	1.0347556	0.6962964	1.478104
67	1.0888679	1.0596871	0.7117023	1.0038042	0.6653450	1.447153
68	1.0579165	1.0287357	0.6807509	0.9728528	0.6343936	1.416201
69	1.0269651	0.9977843	0.6497994	0.9419014	0.6034422	1.385250
70	0.9960137	0.9668329	0.6188480	0.9109499	0.5724907	1.354298
71	0.9650623	0.9358815	0.5878966	0.8799985	0.5415393	1.323347
72	0.9341109	0.9049301	0.5569452	0.8490471	0.5105879	1.292396
73	0.9031594	0.8739786	0.5259938	0.8180957	0.4796365	1.261444
74	0.8722080	0.8430272	0.4950424	0.7871443	0.4486851	1.230493
75	0.8412566	0.8120758	0.4640909	0.7561929	0.4177337	1.199541
76	0.8103052	0.7811244	0.4331395	0.7252414	0.3867822	1.168590
77	0.7793538	0.7501730	0.4021881	0.6942900	0.3558308	1.137638
78	0.7484024	0.7192216	0.3712367	0.6633386	0.3248794	1.106687
79	0.7174509	0.6882701	0.3402853	0.6323872	0.2939280	1.075735
80	0.6864995	0.6573187	0.3093339	0.6014358	0.2629766	1.044784

- **Gráficos de los resultados**

```
plot(30:80,predNI,main = 'Fumadores que no inhalan',xlab='Edad',ylab='Capacidad pulmonar',ylim = c(0,3))
plot(30:80,predFP,main = 'Fumadores pasivos',xlab='Edad',ylab='Capacidad pulmonar',ylim = c(0,3))
plot(30:80,predFM,main = 'Fumadores moderados',xlab='Edad',ylab='Capacidad pulmonar',ylim = c(0,3))
plot(30:80,predFL,main = 'Fumadores ligeros',xlab='Edad',ylab='Capacidad pulmonar',ylim = c(0,3))
plot(30:80,predFI,main = 'Fumadores intensivos',xlab='Edad',ylab='Capacidad pulmonar',ylim = c(0,3))
plot(30:80,predNF,main = 'No fumadores',xlab='Edad',ylab='Capacidad pulmonar',ylim = c(0,3))
```



En gráficos anteriores podemos observar como el modelo ha predicho los datos en función de la edad y el Tipo de fumador (se ha asumido genero hombre según especificado en el enunciado). Observamos como las personas de menor edad, 30 años en este caso, presentan más capacidad pulmonar que el resto de personas del mismo Tipo. Vemos también como varía este dato según el tipo de fumador que sea. Observamos que

el dato menor lo encontramos en los fumadores moderados e intensivos.

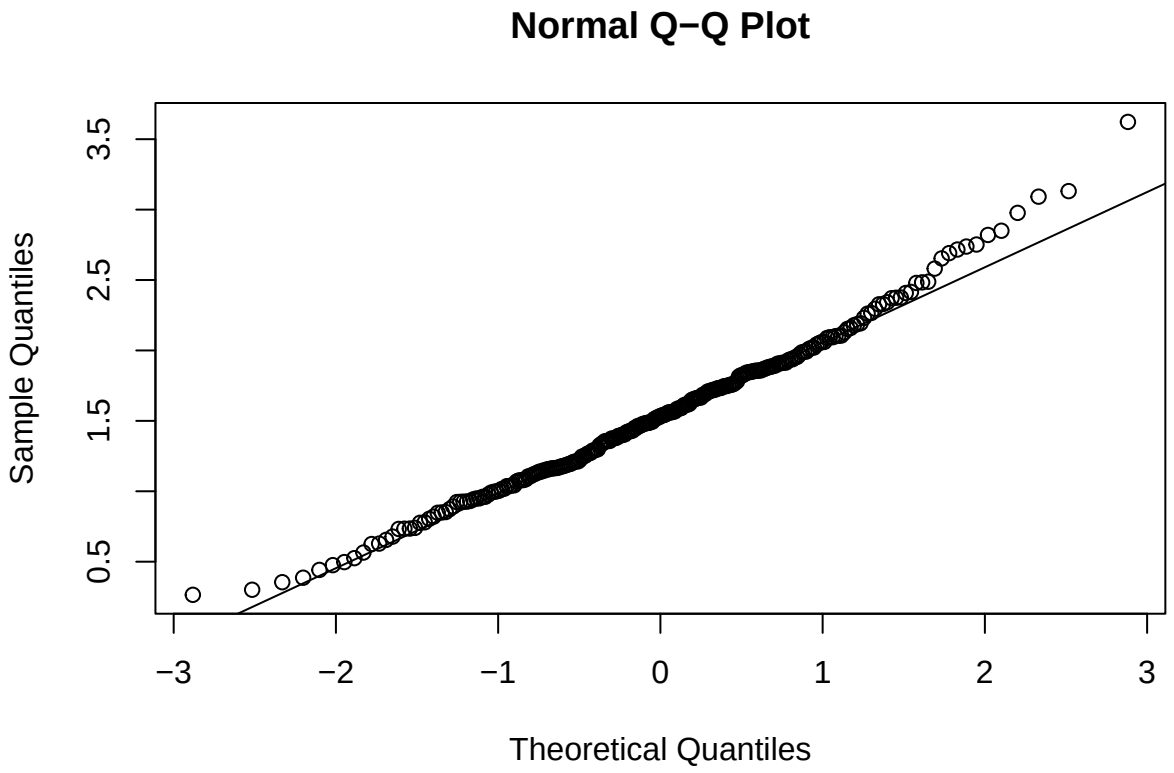
7. ANOVA unifactorial

Trataremos de responder las siguientes preguntas:

- ¿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

```
qqnorm(dat$AE)
qqline(dat$AE)
```



Podemos asumir la normalidad de la muestra ya que como vemos en el gráfico se asemeja bastante a la distribución normal.

7.2. Homoscedasticidad: Homogeneidad de varianzas

Procedemos a estuadiar la igualdad de varianzas con el test Bartlett.

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

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

Observamos que el valor p es mayor a 0.05 por lo que podemos confirmar la igualdad de varianzas.

7.3. Hipótesis nula y alternativa.

ANOVA de un factor

$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_a = 0$

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

7.4. Cálculo ANOVA

```r
set.seed(111)
library(lsr)
# Análisis de varianza
AV <- aov(AE ~ Tipo, data = dat)
summary(AV)
```

```
##              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

Los gráficos representados en el apartado 3 dejaban entrever la posible influencia de la variable Tipo en la variable de la capacidad pulmonar. Los datos de la capacidad pulmonar llegaban a ser bastante diferentes sobre todo entre los grupos (no fumadores) y los grupos (fumadores moderados y fumadores intensivos).

Con el resultado del ANOVA realizado en el apartado anterior confirmamos nuestras sospechas ya que el valor p es menor a 0.05 por lo que podemos rechazar la hipótesis nula en favor de la alternativa. Es decir, confirmamos que el tipo de fumador es significativo y que existen al menos dos tipos de fumadores cuyos efectos son significativamente distintos.

7.6. Profundizando en ANOVA

A raíz de los calculos realizados en el apartado 7.4. devuelto por el modelo aov() podemos observar los siguientes resultados:

```
SST <- 78.49
SSW <- 57.63
SSB <- 20.86
```

La suma de cuadrados total (SST) se trata de la suma de los cuadrados de los tratamientos más la suma de cuadrados del error. Se puede interpretar como la cantidad de variación en la variable dependiente.

"Within sum of squares" (SSW), es la suma de las diferencias al cuadrado entre un valor y su media muestral. Cuantifica la variabilidad entre los diferentes grupos.

En cambio, "Between sum of squares" SSB es la suma de las diferencias al cuadrado entre un valor y la media de todos los valores independientemente de la muestra.

Los grados de libertad como se puede observar se trata de el número de grupos - 1, es decir 6 grupos diferentes de fumadores/no fumadores - 1 = 5. El cálculo del grado de libertad 'residuals' se lleva a cabo con el número de observaciones (253) menos el número de grupos (6), es decir 247 grados de libertad.

```
SSB_df <- 6-1
SSW_df <- 253-6
```

El cálculo del valor f sería de la siguiente manera:

```
f_value <- (SSB/SSB_df)/(SSW/SSW_df)
f_value
```

```
## [1] 17.88103
```

A continuación calculamos el valor crítico de f

```
f_critical_value <- qf(0.05,SSB_df,SSW_df,lower.tail = FALSE)
f_critical_value
```

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

El valor f obtenido como resultado del test es mayor al valor f crítico calculado lo que significa que podemos rechazar la hipótesis nula a favor de la hipótesis alternativa. El test ANOVA realizado nos indica que al menos un grupo difiere significativamente del resto aunque, de momento, no sabemos cual.

7.7. Fuerza de la relación

La fuerza de la relación entre la variable independiente se mide mediante el estadístico eta cuadrado. Es decir, que parte de la variación de la capacidad pulmonar es atribuible al tipo de fumador.

En este caso como vemos a continuación tenemos un eta cuadrado de 0.2657. Esto es que un 26.57% de la variación de la variable AE es explicada por la variable Tipo.

```
# Capacidad explicativa
etaSquared(AV)
```

```
##         eta.sq eta.sq.part
## Tipo 0.2657271   0.2657271
```

8. Comparaciones múltiples

8.1. Test pairwise

Con la test pairwise veremos una comparación de medias por parejas, en este caso se realizará el test sin ninguna correción.

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

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

Los valores p calculados en el test anterior menores a 0.05 indican la presencia de diferencias entre dichos grupos. Observamos que las mayores diferencias se encuentran entre el grupo de no fumadores con los grupos de fumadores intensivos y moderados.

8.2. Corrección de Bonferroni

A continuación realizamos el mismo test con la corrección de Bonferroni, ésta divide el nivel de significancia entre el número de comparaciones dos a dos realizadas.

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

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

Como era de esperar el valor p ha aumentado en general. Sin embargo, seguimos encontrando las diferencias comentadas anteriormente entre el grupo de no fumadores con los grupos de fumadores intensivos y moderados.

Se podría decir que todos las comparaciones tienen diferencias significativas menos las siguientes:

- Fumador intensivo - fumador moderado.
- Fumador ligero - fumador pasivo - fumador que no inhala.

9. ANOVA multifactorial

9.1 Análisis visual

En la tabla a continuación podemos ver la media de la capacidad pulmonar según el tipo de fumador y el genero.

```
# Group by
df2 <- dat %>% group_by(Tipo,genero) %>%
  summarise(mean_AE=mean(AE))

# Convert to dataframe
df3 <- df2 %>% as.data.frame()

df3_M <- subset(df3,df3$genero=='M',c(Tipo,mean_AE))
df3_F <- subset(df3,df3$genero=='F',c(Tipo,mean_AE))
kable(df3_M,caption = "Media AE Hombres")%>%
  kable_styling(latex_options = "HOLD_position")
```

Table 1: Media AE Hombres

	Tipo	mean_AE
2	FI	1.137577
4	FL	1.650977
6	FM	1.266695
8	FP	1.614373
10	NF	2.074773
12	NI	1.642543

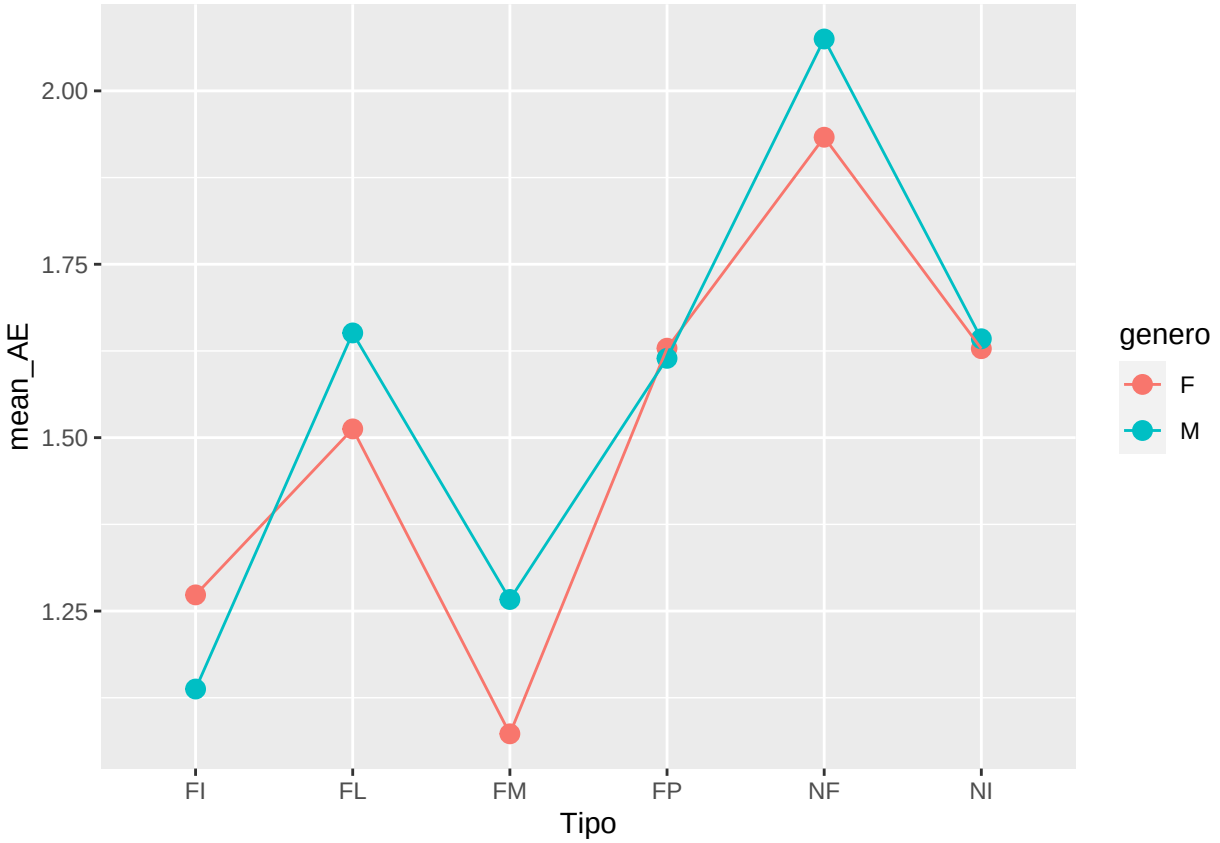
```
kable(df3_F,caption = "Media AE Mujeres")%>%
  kable_styling(latex_options = "HOLD_position")
```

Table 2: Media AE Mujeres

	Tipo	mean_AE
1	FI	1.273318
3	FL	1.512600
5	FM	1.073020
7	FP	1.628993
9	NF	1.933138
11	NI	1.628288

En la gráfica siguiente se muestra la media de la capacidad pulmonar para cada tipo de fumador y genero.

```
ggplot(df3, aes(x=Tipo, y=mean_AE, group='genero')) + geom_point(aes(colour=genero),size=3)+
  geom_line(aes(group=genero,colour=genero))
```



El gráfico parece indicar la posible existencia de interacción entre los factores genero y tipo de fumador. Pasamos a realizar un ANOVA multifactorial teniendo en cuenta la interacción para confirmar o no dicha teoria.

9.2. ANOVA multifactorial

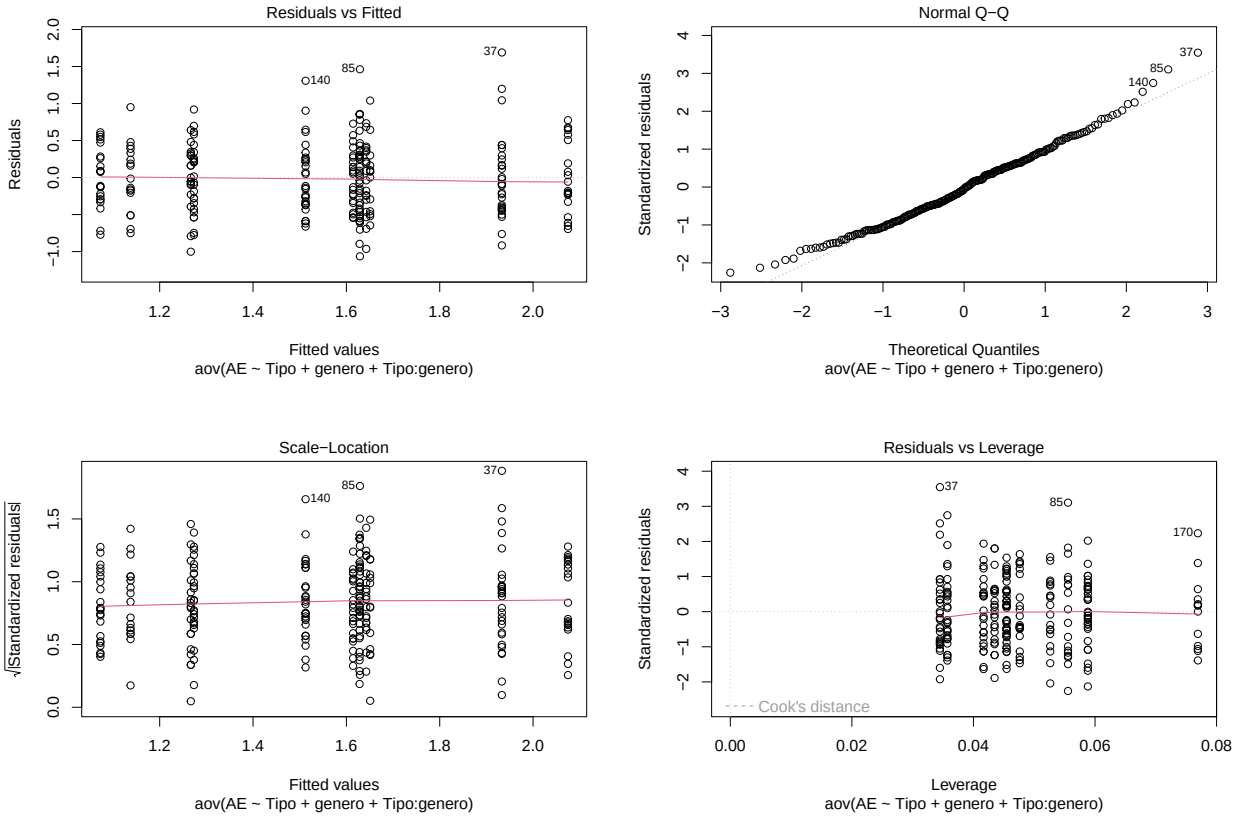
A continuación realizamos un ANOVA multifactorial teniendo en cuenta el efecto de la interacción.

```
#Ajustamos el modelo de dos factores con interacción

AV1 <- aov(AE~Tipo+genero+Tipo:genero,data = dat)
```

Comprobamos el comportamiento de los datos en relación a los supuestos del ANOVA.

```
plot(AV1) # Gráficos de ajuste del modelo
```



En el primer gráfico de residuos no se observan distribuciones demasiado diferentes entre sí por lo que asumiremos homocedasticidad. Por el resultado del gráfico Normal Q-Q asumiremos también normalidad ya que los datos se asemejan los suficiente a esta distribución.

```
summary(AV1)
```

```
##              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

Los resultados obtenidos en el ANOVA del apartado anterior rechazan la hipótesis de que existe interacción entre las variables Tipo y genero ya que el valor p es mayor al valor de significancia acordado en un 0.05.

En relación a los efectos principales, observamos que únicamente la variable Tipo es significativa por lo parece existir una asociación entre la variable "AE" (capacidad pulmonar) y este factor, el tipo de fumador. Por el contrario, la variable genero no es significativa.

10.Resumen técnico

N	Pregunta	Resultado	Conclusión
P1	IC AE mujeres 95%	1.43-1.62	El intervalo de confianza al 95% de la variable AE es entre 1.43 y 1.62
P2	IC AE hombres 95%	1.48-1.69	El intervalo de confianza al 95% de la variable AE es entre 1.48 y 1.69
P3	¿Presentan los hombres una capacidad pulmonar significativamente diferente a las mujeres?	tobs= 0.85, tcritL= -1.96, tcritU= 1.96, pvalue= 0.39	No existen diferencias significativas en la capacidad pulmonar entre los hombres y las mujeres con un nivel de confianza del 95%.
P4	¿Presentan las personas fumadoras una capacidad pulmonar menor a las personas no fumadoras?	tobs= -6.32, tcrit= -1.65, pvalue= 5e-10	Los fumadores presentan una capacidad pulmonar menor a la de los no fumadores.
P5	¿Existe relación entre la variable AE y el resto de variables (tipo, edad y genero)?	R-squared: 0.583	La edad y el hecho de no ser fumador influyen bastante en el modelo. El 58.3% de la variabilidad de la variable predicha es explicada por la variable predictora.
P6	¿Existen diferencias entre la capacidad pulmonar (AE) entre los distintos tipos de fumadores/no fumadores?	P valor= 4e-15, eta cuadrado= 0.2657	La variable factor Tipo es significativa: existen al menos dos niveles cuyos efectos son significativamente distintos. Un 26.57% de la variación de la variable AE es explicada por la variable Tipo.
P7	¿Puede la variabilidad de la variable AE explicarse a partir de la variable Tipo y genero?	P valor(Tipo) = 5e-15, P valor(genero) = 0.36, P valor(Tipo:genero) = 0.66	No existe interacción significativa entre los niveles de los factores Tipo y genero. La variable Tipo es significativa mientras que la variable genero no lo es.

11. Resumen ejecutivo

Del análisis realizado a partir de la muestra proporcionada se han sacado las siguientes conclusiones:

- La capacidad pulmonar de una persona se ve afectada por el grupo de tipo de fumador al que pertenezca.
- Las personas fumadoras (intensivas, moderadas, ligeras, que no inhalan) presentan una capacidad pulmonar menor a las personas no fumadoras(no fumadoras, fumadoras pasivas).
- Se observa cierta tendencia a la reducción de la capacidad pulmonar a medida que aumenta la edad.
- No existen diferencias significativas en la capacidad pulmonar de los hombres y de las mujeres.
- Las mayores diferencias de la capacidad pulmonar teniendo en cuenta el tipo de fumador las encontramos en los siguientes grupos: no fumadores-fumadores intensivos y no fumadores-fumadores moderados.
- Entre los siguientes tipos de fumadores no se encuentran diferencias significativas de la capacidad pulmonar: fumador intensivo-fumador moderado y fumador ligero-fumador pasivo-fumador que no inhala.