Práctica 10. ANOVA. Soluciones

Estadística (650008). Grado en Biología Sanitaria. UAH

Actualizado: 2024-12-03

PASOS PARA RESOLVER ESTOS EJERCICIOS. En general, se trata de analizar si la medida de centralización (media o mediana) de cierta variable cuantitativa se comporta de la misma manera en ciertos grupos.

- Pequeña exploración inicial para hacerte una idea de lo que pueden decir o no los datos.
- Analiza si se cumplen ciertas hipótesis (normalidad y homocedasticidad de los residuos del modelo ANOVA). Si se cumplen, hacer un contraste ANOVA para comparar las medias en cada grupo. Si no se cumplen, se comparan las medianas en cada grupo con el contraste de Kruskal-Wallis.
- Si el contraste es significativo, lo siguiente es ordenar las medias/medianas.
- En el caso de las medias puedes usar los métodos de Bonferroni/Tuckey. Para las medianas el método de Dunnet.

OJO: en las soluciones he ordenado las medias/medianas de distintas formas (diagrama, con desigualdades, explicándolo de palabra...). NO es necesario hacerlo de todas las formas posibles, aunque tienes que entenderlas todas (porque se usan indistintamente unas u otras).

Ejercicio 1

Resuelve el ejemplo ANOVA, cortesía de la unidad de Bioestadística Clínica del hospital Ramón y Cajal. Aquí están los fichero de datos.

En concreto:

1. **Descarga y lee el fichero de datos**. Puedes usar el botón `Import Datase`, o bien la orden

```r
datos = read.table(file = "Practica10-RamonCajal.csv", sep = ";", header = T)
```

para guardar los datos en una variable llamada `datos`. Nos aseguramos de que han sido bien leídos

```r
head(datos)
```

```
   X1  X2  X3  X4  X5
1 180 172 163 158 147
2 173 158 170 146 152
3 175 167 158 160 143
4 182 160 162 171 155
5 181 175 170 155 160
```

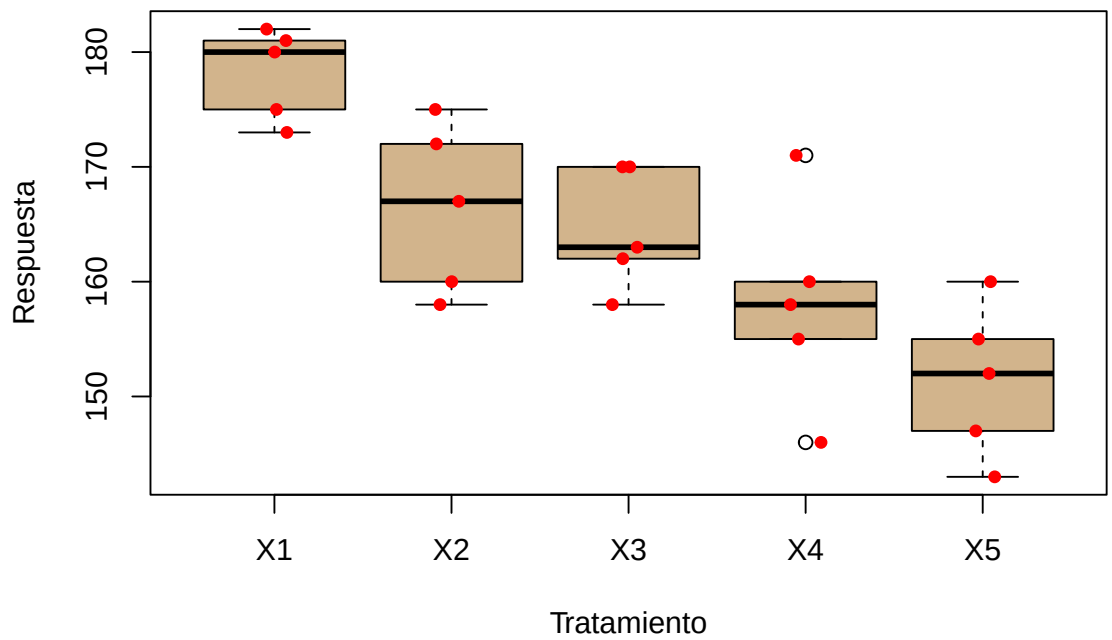
2. **Decide si hay algún tratamiento que produce una respuesta distinta de los demás**. Vamos a ir usando distintos fragmentos de la plantilla, y comentando los resultados obtenidos:

```r
####--- poner datos en formato correcto
# columnas de con el Tratamiento y la Respuesta
datos = stack(datos)
datos = datos[, c(2,1)]
colnames(datos) <- c("Tratamiento", "Respuesta")
head(datos)
```

```
  Tratamiento Respuesta
1          X1       180
2          X1       173
3          X1       175
4          X1       182
5          X1       181
6          X2       172
```

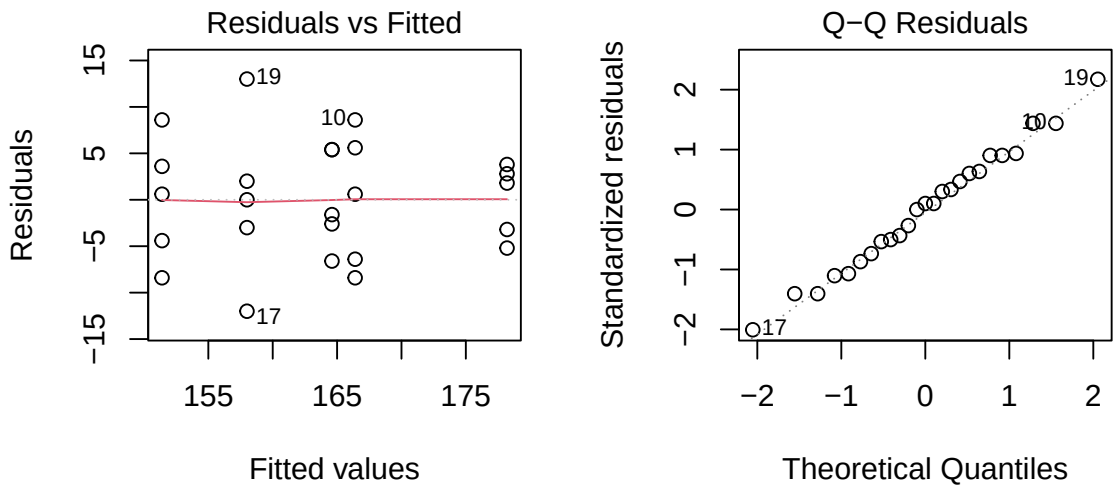
Procedemos a explorar los datos

```r
 ###--- EXPLORACION
# boxplots
bp = boxplot(Respuesta ~ Tratamiento,
         data=datos, col="tan")
stripchart(Respuesta ~ Tratamiento,
       data=datos, col="red",
       vertical = TRUE, method = "jitter",
       cex=0.8, add=TRUE, pch=19)
```



Observa que hay pocos datos en cada grupo, lo que permite que los boxplots puedan tener formas más o menos "raras". Nosotros **asumimos la hipótesis independencia** en la toma de las muestras, y tenemos que **comprobar la normalidad y la homocedasticidad de la muestra**. Podemos usar métodos gráficos

```r
###--- Graficos diagnostico condiciones ANOVA
modelo = lm(datos$Respuesta ~ datos$Tratamiento)
par(mfrow = c(1, 2), mar = c(5,5,2,1))
for(i in 1:2){
  plot(modelo, which = i)
}
```




```
par(mfrow = c(1, 1))
```

El qqplot (panel derecho) no muestra desviaciones respecto de la normalidad. En cuanto a la homocedasticidad, fíjate (panel izquierdo) en que el tratamiento que tiene una media mayor (algo más de 175) parece tener una dispersión menor que el resto. Podemos usar los contrastes de hipótesis para salir de dudas. A continuación, se incluye todo el bloque de contrastes que hay en el script: se hace uno de normalidad (aunque el método gráfico no hace sospechar nada) y otro de homocedasticidad (ahí, el método grafico puede generar alguna duda)

```
###--- contrastes normalidad
# test Shapiro-Wilks
# H0: las muestras provienen de poblaciones normales
shapiro.test(modelo$residuals)
```

```
    Shapiro-Wilk normality test

data:  modelo$residuals
W = 0.98911, p-value = 0.9927
```

```
# test Kolmogorov-Smirnov-Lilliefors
## H0: las muestras provienen de poblaciones normales
# library(nortest)
# lillie.test(modelo$residuals)
```

es coherente con el qqplot que hemos observado (en principio, basta con hacer uno de ellos). Para la homocedasticidad usamos el contraste de Levene

```
###--- contrastes homocedasticidad
## H0: varianzas iguales (caso de poblaciones normales)
 bartlett.test(Respuesta ~ Tratamiento, data = datos)
```

```
    Bartlett test of homogeneity of variances

data:  Respuesta by Tratamiento
Bartlett's K-squared = 2.6811, df = 4, p-value = 0.6125
```

```r
## H0: varianzas iguales (caso de poblaciones cualesquiera)
# library(car)
# leveneTest(Respuesta ~ Tratamiento, data = datos)
```

tampoco hay porqué dudar de la igualdad de varianzas. Ambos p-valores son grandes, por lo que no hay porqué dudar de que se cumplan las hipótesis. Procedemos a hacer el **contraste ANOVA**:

$$H_0: \text{ Las medias son iguales en las cuatro zonas}$$

$$H_1: \text{ Al menos una media es distinta en alguna de las cuatro zonas}$$

```r
###--- TABLA ANOVA
# H0: las medias son iguales
# H1: al menos una media es diferente
modelo = lm(datos$Respuesta ~ datos$Tratamiento)

# p-valor de contraste
anova(modelo)
```

```
Analysis of Variance Table

Response: datos$Respuesta
                  Df Sum Sq Mean Sq F value    Pr(>F)
datos$Tratamiento  4 2010.6  502.66   11.24 6.062e-05 ***
Residuals         20  894.4   44.72
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

que resulta ser significativo: al menos un tratamiento produce una respuesta diferente de la de los demás. Podemos calcular el coeficiente de determinación r^2

```r
# coeficiente de determinacion r2
resumen = summary(modelo)
resumen$adj.r.squared
```

```
[1] 0.6305455
```

que es Multiple R-squared: 0.6921; bastante bueno; el modelo tiene bastante poder explicativo.

3. **En caso afirmativo, ordena los tratamientos de acuerdo con su efectividad para reducir la hipertensión (puedes usar Bonferroni o Tukey); prueba con $\alpha = 0.05$ y $\alpha = 0.01$.** Ahora nos interesa **ordenar las medias poblacionales**, es decir, la respuesta que producen en la presión arterial los distintos tratamientos. Esencialmente, se hacen comparaciones de medias 2 a 2, pero teniendo la precaución de repartir el error de tipo I entre todas las comparaciones para que el error de tipo I acumulado no se dispare. **Ten en cuenta** que el método de Bonderroni es más conservador que el de Tukey: el primero necesita de evidencias muestrales más fuertes para rechazar H_0 (las 2 medias que se compara son iguales), mientras que el segundo rechaza la igualdad de medias en cada comparación con menor evidencia muestral (es más fácil dar por diferentes dos medias poblacionales). Podemos usar el ajuste de **Bonferroni diagrama de medias ordenadas**

```r
# Ajuste Bonferroni
(ptt=pairwise.t.test(datos$Respuesta, datos$Tratamiento,
                 p.adj="bonferroni", pool.sd=FALSE))
```

```
    Pairwise comparisons using t tests with non-pooled SD

data:  datos$Respuesta and datos$Tratamiento
```

```
   X1     X2     X3     X4
X2 0.1914 -      -      -
X3 0.0210 1.0000 -      -
X4 0.0472 1.0000 1.0000 -
X5 0.0016 0.0982 0.0906 1.0000

P value adjustment method: bonferroni
```

Podemos calcular las medias muestrales con

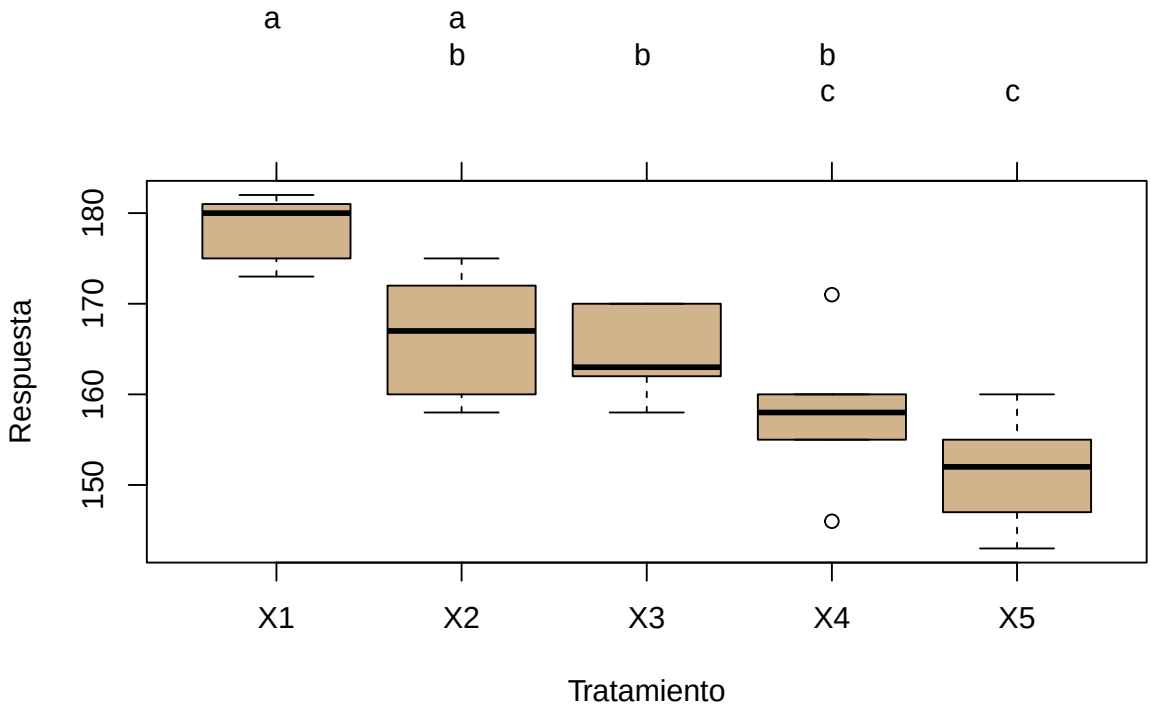
```r
# Medias muestrales
aggregate(datos$Respuesta ~ datos$Tratamiento, FUN = mean)
```

```
  datos$Tratamiento datos$Respuesta
1                X1           178.2
2                X2           166.4
3                X3           164.6
4                X4           158.0
5                X5           151.4
```

- Para $\alpha = 0.05$ se tiene:
  - La media μ_{X_1} no es diferente de μ_{X_2}.
  - Las medias μ_{X_2}, μ_{X_3}, μ_{X_4} y μ_{X_5} no son diferentes entre sí.
  - Además, μ_{X_1} es mayor que μ_{X_3}, μ_{X_4} y μ_{X_5}.
  - Podemos usar el código de letras que conoces del método de Tukey para resumir lo anterior: $\mu_{X_1} - B$, $\mu_{X_2} - B - A$, $\mu_{X_3} - A$, $\mu_{X_4} - A$ y $\mu_{X_5} - A$.
- Para $\alpha = 0.01$, si embargo, resulta que:
  - La media μ_{X_1} no es diferente de μ_{X_2}, μ_{X_3} y μ_{X_4}, pero sí es mayor que μ_{X_5}.
  - Pero μ_{X_2}, μ_{X_3}, μ_{X_4} y μ_{X_5} no son diferentes entre sí.
  - Con el mismo código de letras, el resumen es: $\mu_{X_1} - B$, $\mu_{X_2} - B - A$, $\mu_{X_3} - B - A$, $\mu_{X_4} - B - A$ y $\mu_{X_5} - A$.

  Si usamos el ajuste de **Tuckey** el resultado es algo diferente; **diagrama de medias ordenadas**. Para $\alpha = 0.05$

```r
# Ajuste Tukey, representaciones graficas
# install.packages("multcomp")  # por si no lo tienes instalado
library(multcomp)
datos.aov = aov(Respuesta ~ Tratamiento, data = datos)
glh = summary(glht(model=datos.aov, linfct = mcp(Tratamiento = "Tukey")))
par(mar=c(4,4,8,1))
plot(cld(glh, decreasing = FALSE, level = 0.05), col = "tan")
```



- Los grupos que comparten letra incluyen tratamientos entre los que no hay una respuesta significativamente diferente a nivel poblacional (ojo, que las medias muestrales sí son distintas, pero la diferencia no es relevante).
- X1 y X2 forman un grupo, X2, X3 y X4 otro grupo, y X4 y X5 el último grupo.
- Efectivamente, hay tratamientos que están en 2 grupos, y eso puede parecer raro. Por ejemplo, X2:
  - Se puede considerar que la respuesta de X2 no es significativamente diferente de la de X1 y, a la vez, no es significativamente distinta de la X3 y X4.
  - Sin embargo, X1 no comparte letra con X3 y X4, luego entre X1 y X3, X4 sí hay diferencias significativas.
  - Podemos pensar X2 como un tratamiento con respuesta intermedia; no muy diferente de X1, no muy diferente de X3 y X4, pero esa "poca" diferencia acumulada hace que entre X1 y X3, X4 sí haya diferencias significativas.
- Esto se puede simbolizar como

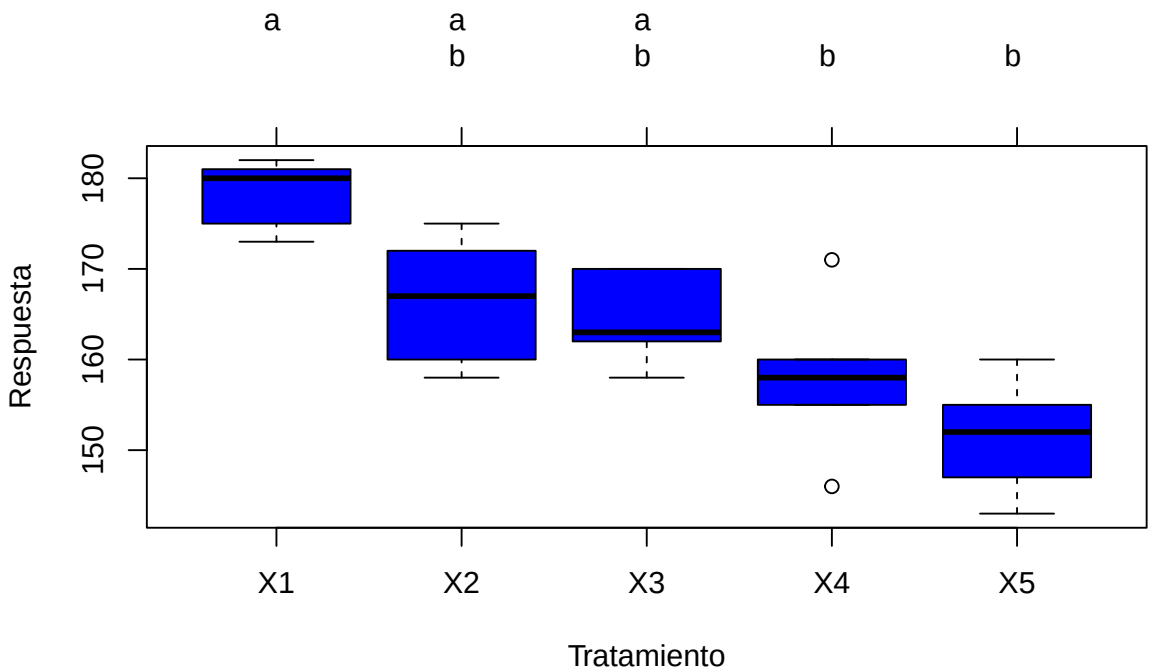
$$X4,\ X5 \leq X2,\ X3,\ X4 \leq X1,\ X2$$

- Recuerda que eran tratamientos para reducir la hipertensión, por tanto:
  - Hay 2 mejores tratamientos: X4 y X5.
  - Si un paciente no pudiera tomar X5 (por cualquier motivo), entonces tomaría X4. Pero ese tratamiento no produce respuesta significativamente distinta de X2 y X3, por lo que podría administrasele también cualquiera de estos dos.

Para $\alpha = 0.01$

```r
# Ajuste Tukey, representaciones graficas
# install.packages("multcomp")  # por si no lo tienes instalado
library(multcomp)
```

```r
datos.aov = aov(Respuesta ~ Tratamiento, data = datos)
glh = summary(glht(model=datos.aov, linfct = mcp(Tratamiento = "Tukey")))
par(mar=c(4,4,8,1))
plot(cld(glh, decreasing = FALSE, level = 0.01), col = "blue")
```



Se observan dos grupos (dentro de cada uno de ellos diferencias significativas entre las respectivas medias): un grupo está formado por las medias μ_{X_1}, μ_{X_2} y μ_{X_3}, y otro por μ_{X_2}, μ_{X_3}, μ_{X_4} y μ_{X_5}. La conclusión es que μ_{X_1} es mayor que μ_{X_4} y μ_{X_5}.

Ejercicio 2

Este fichero de datos fue extraído de aquí el 16 de diciembre de 2020 a las 9:15, y contiene los datos de concentración de NOx medidos en 5 estaciones de la Comunidad de Madrid durante las 24 horas anteriores. Se han eliminado las mediciones de las 10 de la mañana porque una de las estaciones no registró datos a esa hora. En esta página hay datos de contaminantes atmosféricos en tiempo real D.G. del Medio Ambiente Area de Calidad Atmosférica - Red de Calidad del Aire de la comunidad de Madrid.

Se pide:

1. **Descarga y lee el fichero de datos.** Puedes usar el botón `Import Datase`, o bien la orden

```r
datos = read.table(file = "Practica10-NOx-16-dic-2020_limpio.csv", sep = ";", header = T)
```

nos aseguramos de que han sido bien leídos

```r
head(datos)
```

```
           Fecha Getafe Leganes Alcala Alcobendas Fuenlabrada Mostoles Torrejon
1 15/12/20 09:00     44      38     29         50          21       32       61
2 15/12/20 11:00     32      19     24         31          13       12       72
```

```
3 15/12/20 12:00     18      19     21         25          15       13       46
4 15/12/20 13:00     18      21     13         23          17       14       34
5 15/12/20 14:00     21      21     15         15          15       10       33
6 15/12/20 15:00     28      21     16         12          15       11       31
  Alcorcon Coslada
1       32      60
2       12      55
3       14      24
4       16      24
5       12      31
6       11      24
```

Nos quedamos con las columnas 2, 3, 4, 5 y 6; y hacemos una pequeña exploración inicial para asegurar la integridad de los datos:

```r
datos = datos[ , 2:6]
summary(datos)
```

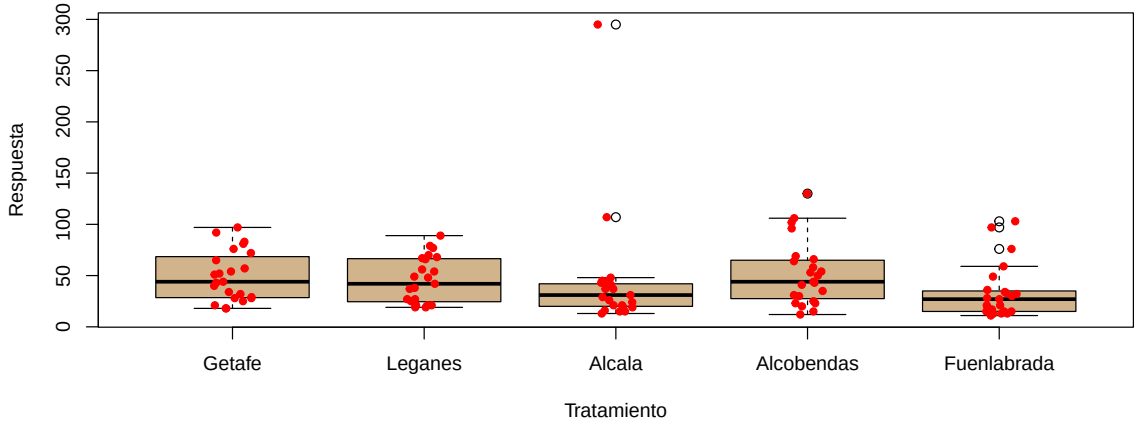
```
     Getafe         Leganes          Alcala        Alcobendas
 Min.   :18.00   Min.   :19.00   Min.   : 13.0   Min.   : 12.00
 1st Qu.:28.50   1st Qu.:24.50   1st Qu.: 20.0   1st Qu.: 27.50
 Median :44.00   Median :42.00   Median : 31.0   Median : 44.00
 Mean   :49.57   Mean   :45.39   Mean   : 44.7   Mean   : 51.74
 3rd Qu.:68.50   3rd Qu.:66.50   3rd Qu.: 42.0   3rd Qu.: 65.00
 Max.   :97.00   Max.   :89.00   Max.   :295.0   Max.   :130.00
  Fuenlabrada
 Min.   : 11.00
 1st Qu.: 15.00
 Median : 27.00
 Mean   : 33.78
 3rd Qu.: 35.00
 Max.   :103.00
```

2. **Analiza si la concentración media del contaminante es la misma en todas las localizaciones**. Hay que poner los poner datos en formato correcto, con una columna para los tratamientos (niveles del factor) y otra para la respuesta (valor numérico)

```r
datos = stack(datos)
datos = datos[, c(2,1)]
colnames(datos) <- c("Tratamiento", "Respuesta")
```

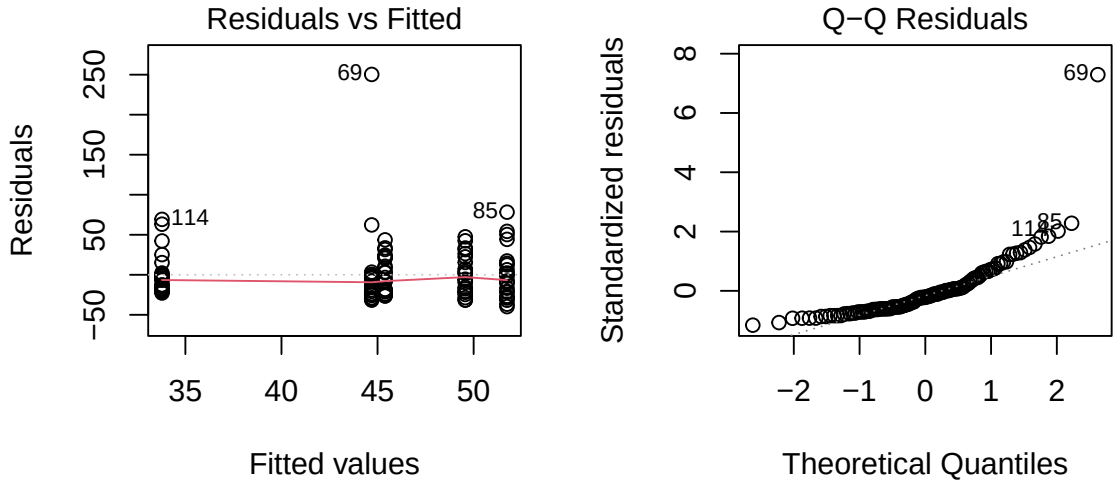
A partir de aquí, seguiremos el script

```r
##--- EXPLORACION
# boxplots
bp = boxplot(Respuesta ~ Tratamiento,
         data=datos, col="tan")
stripchart(Respuesta ~ Tratamiento,
       data=datos, col="red",
       vertical = TRUE, method = "jitter",
       cex=0.8, add=TRUE, pch=19)
```



Los boxplots sugieren que no se va a cumplir ni la normalidad (bigote inferior mucho más corto que el superior) ni la homocedasticidad (boxplots de anchuras muy diferentes). Se puede corroborar mediante las herramientas gáficas y numéricas (contrastes) que conoces.

```
###--- Graficos diagnostico condiciones ANOVA
modelo = lm(datos$Respuesta ~ datos$Tratamiento)
par(mfrow = c(1, 2), mar = c(5,5,2,1))
for(i in 1:2){
  plot(modelo, which = i)
}
```




```
par(mfrow = c(1, 1))
```

Llama la atención un valor muy atípico en Alcalá de Henares que, tal vez, cabría eliminar. De momento, lo mantenemos. En todo caso, ni las columnas el gráfico de la izquierda tienen alturas similares (no hay homocedasticidad) ni la mayor parte de los puntos se disponen sobre la recta en el qqplot (gráfico de la derecha, no hay normalidad). Podemos corroborar esto con los correspondientes contrastes de hipótesis

```
    ###--- contrastes normalidad
# test Shapiro-Wilks N < 50
# H0: las muestras provienen de poblaciones normales
```

```r
# shapiro.test(modelo$residuals)

# test Kolmogorov-Smirnov-Lilliefors, N >= 50
## H0: las muestras provienen de poblaciones normales
library(nortest)
lillie.test(modelo$residuals)
```

```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  modelo$residuals
D = 0.17136, p-value = 7.53e-09
```

La normalidad falla estrepitosamente, no así la homocedasticidad. Se suele decir que ANOVA es robusto frente a pequeñas desviaciones de la normalidad (y siempre que se alcance la homocedasticidad). En este caso desvición respecto de la normalidad es muy fuerte, por lo que hacer un ANOVA no tiene sentido. Es importante resaltar que, aunque la teoría no respalde el uso de ANOVA, matemáticamente es posible hacer los cálculos. Es decir, si se lo pides, el ordenador te hará el contraste ANOVA (tú mandas) aunque del resultado que arroje NO se puedan extraer conclusiones fidedignas (respaldadas por la teoría).

En cuanto a la homocedsaticidad

```r
###--- contrastes homocedasticidad
## H0: varianzas iguales (caso de poblaciones normales)
 # bartlett.test(Respuesta ~ Tratamiento, data = datos)

## H0: varianzas iguales (poblaciones NO normales)
library(car)
leveneTest(Respuesta ~ Tratamiento, data = datos)
```

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   4  0.2446 0.9124
      110
```

es irrelevante que se de la homocestaticidad, porque al no cumplirse la condición de normalidad se abandona la vía del ANOVA.

1. **En caso de no serlo, ordena los municipios de mayor a menor contaminación media. Puedes usar Bonferroni o Tukey**. Estas técnicas están asociadas a las medias, pero no se cumplen las condiciones ANOVA, por lo que hay que trabajar con las medianas. Además, las submuestras (datos de cada estación) son pequeñas, y habría que usar una t de Student para las comparaciones 2 a 2. Sin embargo

   ```r
   shapiro.test(datos$Respuesta)
   ```

   ```
       Shapiro-Wilk normality test

   data:  datos$Respuesta
   W = 0.71756, p-value = 1.422e-13
   ```

   por lo que no podemos asumir que las muestras provengan de una población normal, lo que impide usar Bonferroni o Tukey.

2. **Si no se cumplieran las condiciones ANOVA, utiliza el test de Kruskal-Wallis sobre la mediana y el Dunnet para ordenar las mediansa**. Optamos, pues, por hacer un contraste de

Kruskal-Wallis. Recuerda que ahora la hipótesis son

$$H_0: \text{ Las medinaas son iguales en las cuatro zonas}$$

$$H_1: \text{ Al menos una mediana es distinta en alguna de las cuatro zonas}$$

```r
###--- NO PARAMETRICOS: si falla la normalidad
# KRUSKAL-WALLIS equivale al ANOVA
# H0: medianas iguales
# H1: al menos una mediana diferente
kruskal.test(datos$Respuesta, datos$Tratamiento)
```

```
    Kruskal-Wallis rank sum test

data:  datos$Respuesta and datos$Tratamiento
Kruskal-Wallis chi-squared = 11.871, df = 4, p-value = 0.01834
```

el p-valor no es excesivamente pequeño: al 1% de significación no se rechazaría H0, pero sí al 5% y al 10% (que es el que indica el enunciado), por lo que se decide que rechazar H0. Para ordenar las medianas y determinar qué estaciones registran concentraciones de NOx significativamente diferentes (superiores/inferiores) usaremos el test de Dunnet con un **nivel de significación** $\alpha = 0.1$ (aqui tienes el **diagrama de medias ordenadas**)

```r
# Dunnett es equivalente a ordenar las medias
# install.packages("dunn.test")   # por si no lo tienes instalado
library(dunn.test)
dunn.test(datos$Respuesta, datos$Tratamiento, alpha = .1)
```

```
  Kruskal-Wallis rank sum test

data: x and group
Kruskal-Wallis chi-squared = 11.8712, df = 4, p-value = 0.02

                           Comparison of x by group
                                (No adjustment)
Col Mean-|
Row Mean |     Alcala   Alcobend   Fuenlabr     Getafe
---------+-------------------------------------------
Alcobend |  -1.864647
         |    0.0311*
         |
Fuenlabr |   0.721085   2.585733
         |     0.2354    0.0049*
         |
  Getafe |  -1.999574  -0.134927  -2.720660
         |    0.0228*     0.4463    0.0033*
         |
 Leganes |  -1.501892   0.362754  -2.222978   0.497681
         |     0.0666     0.3584    0.0131*     0.3094

alpha = 0.1
Reject Ho if p <= alpha/2
```

junto con el valor muestral de la mediana de cada problación

```r
aggregate(datos$Respuesta ~ datos$Tratamiento, FUN = median)
```

```
  datos$Tratamiento datos$Respuesta
1            Getafe              44
2           Leganes              42
3            Alcala              31
4        Alcobendas              44
5       Fuenlabrada              27
```

Observa que desde un punto de vista muestral, de mayor a menor, las medianas se ordenan como sigue: Alcalá, Leganés, Getafe, Alcobendas y Fuenlabrada. Sin embargo, a nivel poblacional, no todas las diferencias son significativas. **Ten en cuenta que** en el test de Dunnet la diferencia es signifcativa si el p-valor es menor que $\alpha/2$, y que esto se indica en la tabla poniendo un asterisco junto al p-valor correspondiente. Observa que, referido a la mediana,

- Getafe y Alcobendas tienen la mediana muestral más grande (coinciden). Entre las medianas de Getafe y Alcobendas no hay diferencias significativas (a nivel poblacional), como muestra el p-valor de la tercera fila y segunda columna.
- Getafe presenta diferencias significativas con Alcalá y Fuenlabrada, pero no con Alcobendas (fila 3 de la tabla de Dunnet).
- Lo mismo le sucede a Alcobendas (entrada 1 y 1 de la tabla y columna 2).
- Leganés presenta diferencias significativas con Fuenlabrada, pero no con Alcalá (fila 4 de la tabla).
- Y Alcalá no presenta diferencias significativas con Fuenlabrada.
- Se podría resumir esta información con el código de letras que ya conoces:
  - Getafe-C, Alcobandas-C, Leganés-C-B
  - Alcalá-B-A
  - Fuenlabrada-A

Ahora, para **el nivel de significación** $\alpha = 0.05$

```r
# Dunnett es equivalente a ordenar las medias
# install.packages("dunn.test")   # por si no lo tienes instalado
library(dunn.test)
dunn.test(datos$Respuesta, datos$Tratamiento, alpha = .05)
```

```
  Kruskal-Wallis rank sum test

data: x and group
Kruskal-Wallis chi-squared = 11.8712, df = 4, p-value = 0.02

                           Comparison of x by group
                                (No adjustment)
Col Mean-|
Row Mean |     Alcala   Alcobend   Fuenlabr     Getafe
---------+-------------------------------------------
Alcobend |  -1.864647
         |     0.0311
         |
Fuenlabr |   0.721085   2.585733
         |     0.2354    0.0049*
         |
  Getafe |  -1.999574  -0.134927  -2.720660
         |    0.0228*     0.4463    0.0033*
         |
```

```
 Leganes |  -1.501892   0.362754  -2.222978   0.497681
         |     0.0666     0.3584    0.0131*     0.3094

alpha = 0.05
Reject Ho if p <= alpha/2
```

En resumen, la diferencia con el caso $\alpha = 0.1$ es que con ese nivel de significación no había diferencias significativas ente Alcalá y Alcobendas y con $\alpha = 0.05$ sí las hay.

- Se podría resumir esta información con el código de letras que ya conoces:
  - Getafe-C, Alcobandas-C-B Leganés-C-B
  - Alcalá B-A
  - Fuenlabrada A