Actividad 9. ANOVA

José Carlos Sánchez Gómez

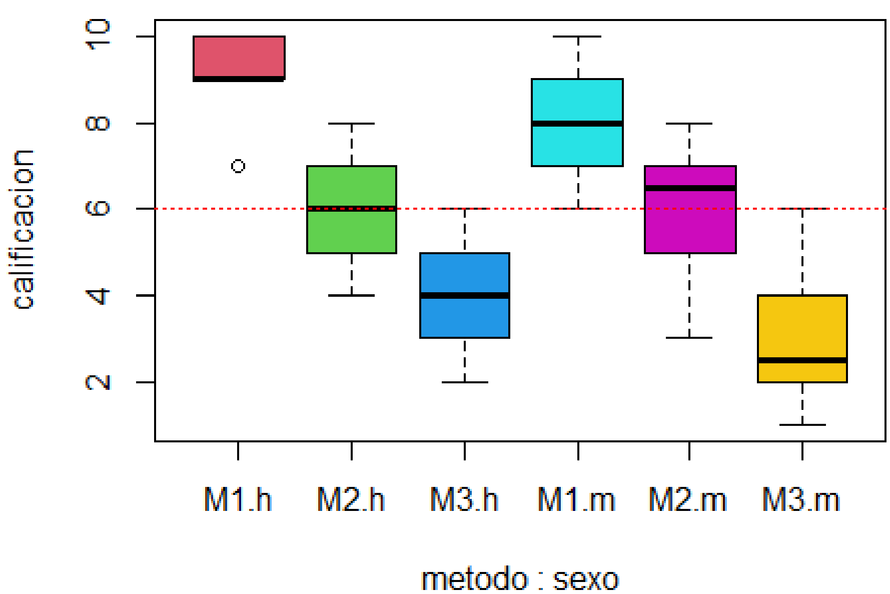
2024-08-28

Problema 1. Resuelve las dos partes del problema “El rendimiento”.

```
# Ingreso de datos
calificacion=c(10,7,9,9,9,10,5,7,6,6,8,4,2,6,3,5,5,3,9,7,8,8,10,6,8,3,5,6
,7,7,2,6,2,1,4,3)
metodo=c(rep("M1",6),rep("M2",6),rep("M3",6),rep("M1",6),rep("M2",6),rep(
"M3",6))
sexo = c(rep("h", 18), rep("m",18))
metodo = factor(metodo)
sexo = factor(sexo)
```

Analisis exploratorio.

```
datos = data.frame(calificacion, metodo, sexo)
boxplot(calificacion ~ metodo : sexo, datos, col = 2:8)
abline(h = mean(calificacion), lty= 3, col = "red")
```



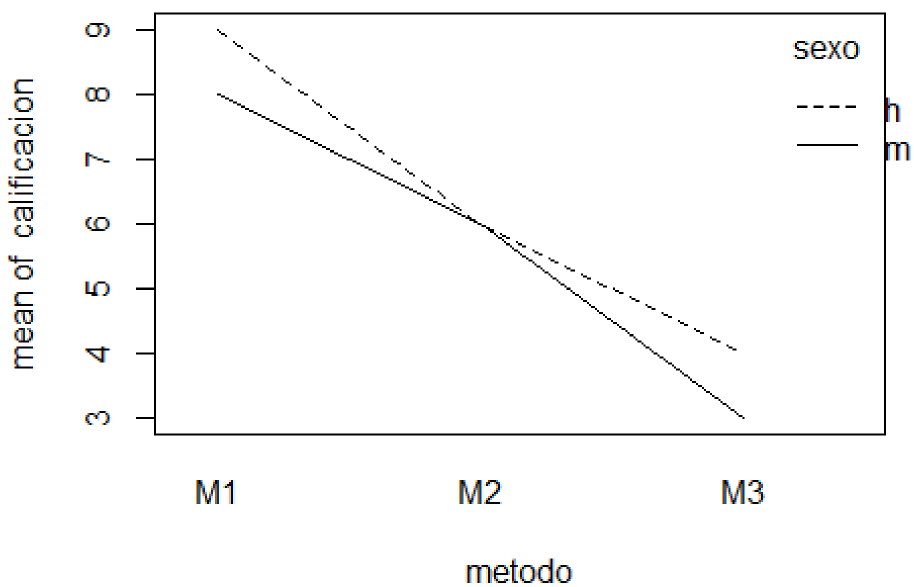
Viendo la distribución por sexo y metodo, observamos que el método que arrojo mejores calificaciones tanto para hombres como para mujeres es el primero, el segundo método da calificaciones menores, y el tercero es el que peor calificaciones tiene, especialmente para mujeres. Los primeros dos métodos ofrecen calificaciones alrededor o mejor de la media. Viendo los datos podemos concluir que el mejor método es el primero, ya sea para mujeres o para hombres.

Establecimiento de hipotesis.

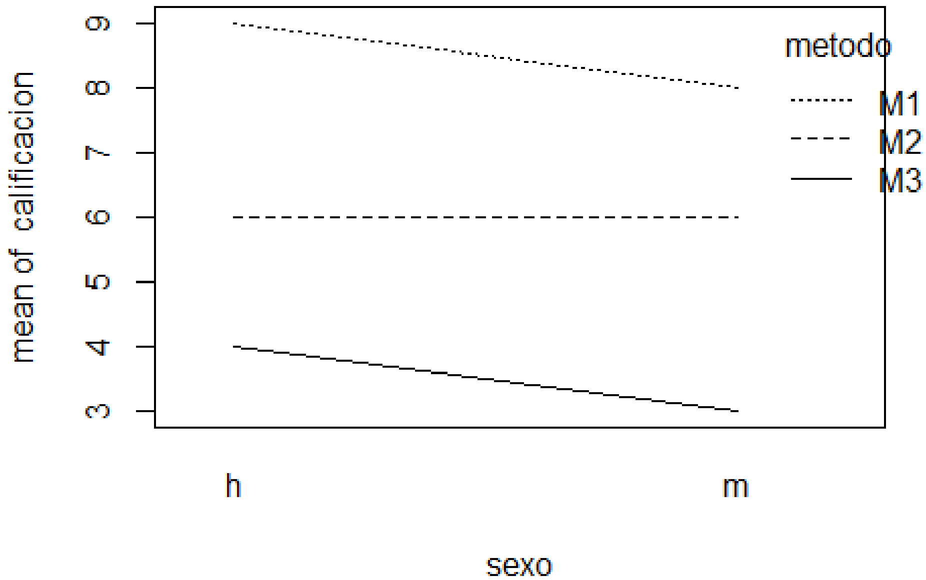
- $h_0: T_i = 0$ h_1: algún T_i es distinto de cero
- $h_0: a_j = 0$ h_1: algún a_j es distinto de cero
- $h_0: t_i a_j = 0$ h_1: aglún $t_i a_j$ es distinto de cero

ANOVA con dos niveles de interacción

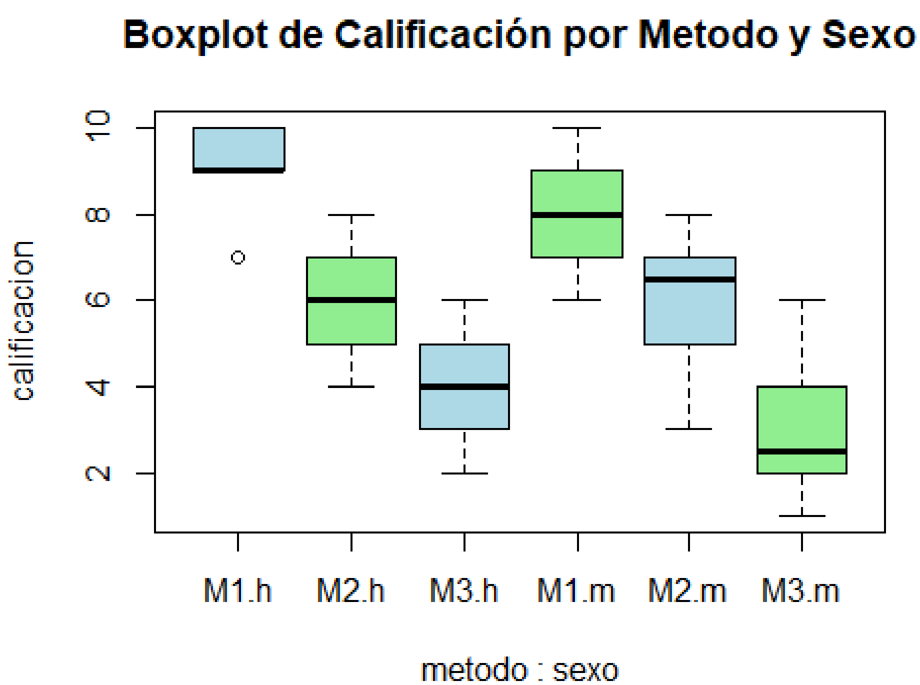
```
anova = aov(calificacion ~ metodo*sexo, datos)
summary(anova)

##             Df Sum Sq Mean Sq F value   Pr(>F)    
## metodo       2    150   75.00  32.143 3.47e-08 ***
## sexo         1      4    4.00   1.714    0.200    
## metodo:sexo  2      2    1.00   0.429    0.655    
## Residuals   30     70    2.33                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

interaction.plot(metodo, sexo, calificacion)
```




```
interaction.plot(sexo, metodo, calificacion)
```



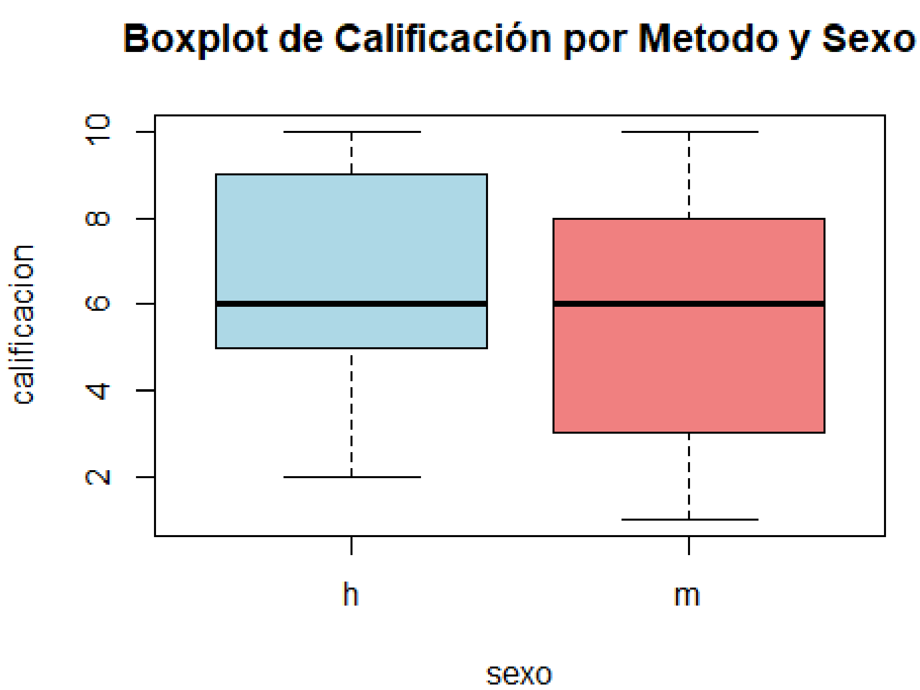

```
boxplot(calificacion ~ metodo * sexo, data = datos, col = c("lightblue",
"lightgreen"), main = "Boxplot de Calificación por Metodo y Sexo")
```



En el resumen de Anova podemos observar que lo que más afecta a los resultados (el valor F) es el método y no el sexo. Esto no se puede visualizar de manera sencilla en la gráfica de caja anterior, por lo que vamos a seguir analizando los datos pero por separado para encontar mayores relaciones entre los datos, y confirmar que el método es lo que más afecta a las calificaciones.

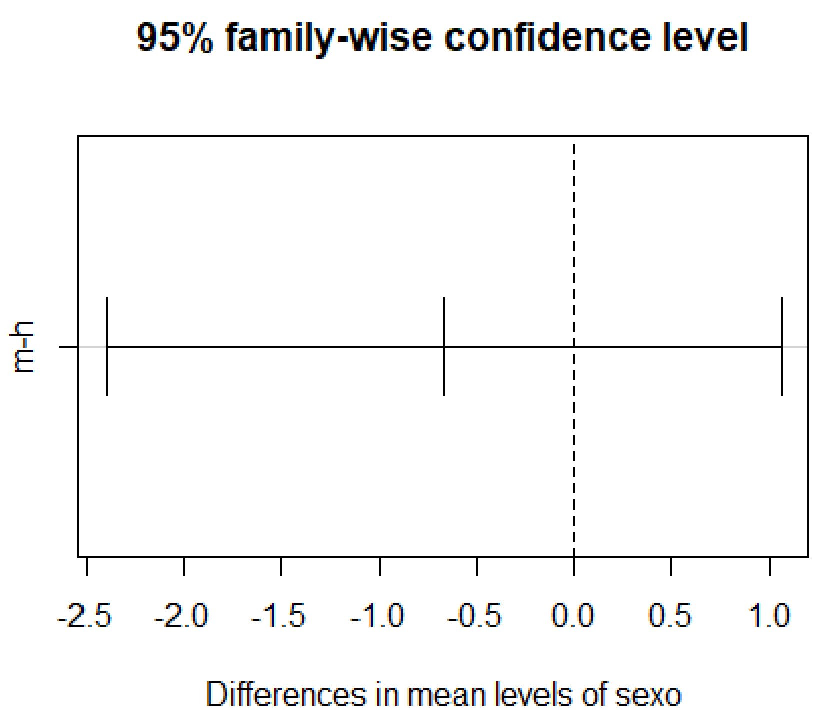
ANOVA con dos niveles sin interacción

```
anova = aov(calificacion ~ metodo + sexo, datos)
summary(anova)

##             Df Sum Sq Mean Sq F value  Pr(>F)    
## metodo       2    150   75.00  33.333 1.5e-08 ***
## sexo         1      4    4.00   1.778   0.192    
## Residuals   32     72    2.25                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Rendimiento por sexo
boxplot(calificacion ~ sexo, data = datos, col = c("lightblue",
"lightcoral"), main = "Boxplot de Calificación por Metodo y Sexo")
```




```
C<-aov(calificacion~sexo)
summary(C)

##             Df Sum Sq Mean Sq F value Pr(>F)
## sexo         1      4   4.000   0.613  0.439
## Residuals   34    222   6.529               

tapply(calificacion,sexo,mean)
```

```
##        h        m
## 6.333333 5.666667

mean(calificacion)

## [1] 6

I = TukeyHSD(aov(calificacion ~ sexo))
I

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = calificacion ~ sexo)
##
## $sexo
##           diff       lwr      upr     p adj
## m-h -0.6666667 -2.397645 1.064312 0.4392235

plot(I)
```



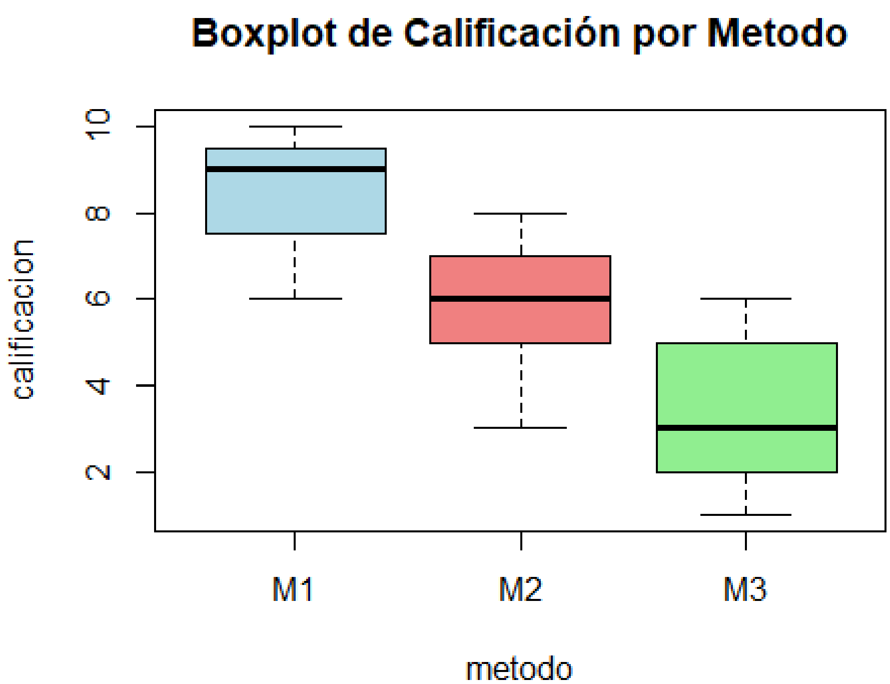
Viendo la comparacion de los datos, nos queda claro que tanto hombres como mujeres tienen calificaciones similares, y que independientemente del método, se desempañaran igual.

ANOVA con un efecto principal

```
anova = aov(calificacion ~ metodo, datos)
summary(anova)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## metodo        2    150    75.0   32.57 1.55e-08 ***
## Residuals    33     76     2.3                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Rendimiento por metodo
boxplot(calificacion ~ metodo, data = datos, col = c("lightblue",
"lightcoral", "lightgreen"), main = "Boxplot de Calificación por Metodo")
```




```r
C<-aov(calificacion~metodo)
summary(C)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## metodo        2    150    75.0   32.57 1.55e-08 ***
## Residuals    33     76     2.3                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

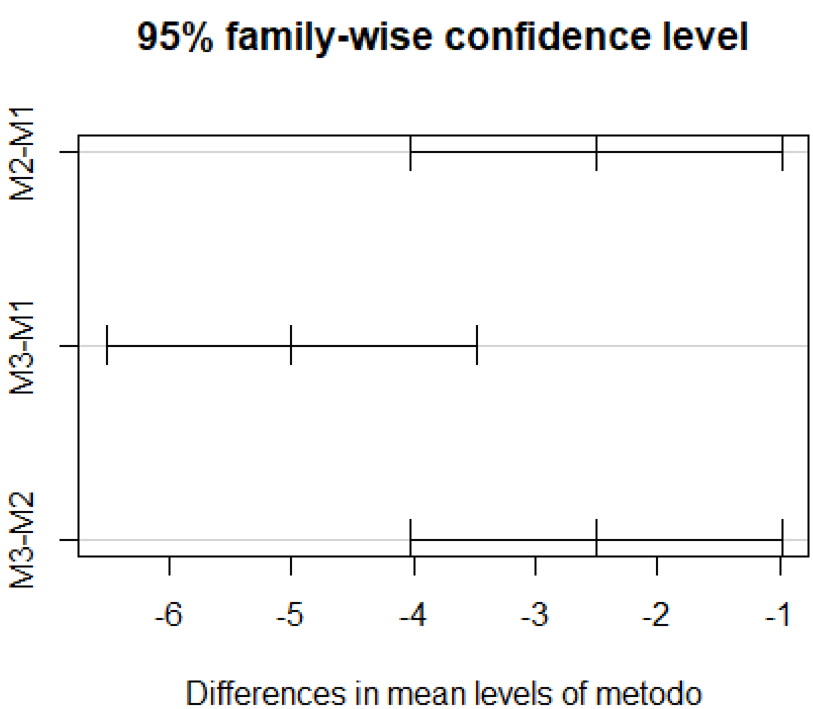
```r
tapply(calificacion,metodo,mean)
```

```
##  M1  M2  M3 
## 8.5 6.0 3.5
```

```r
mean(calificacion)
```

```
## [1] 6
```

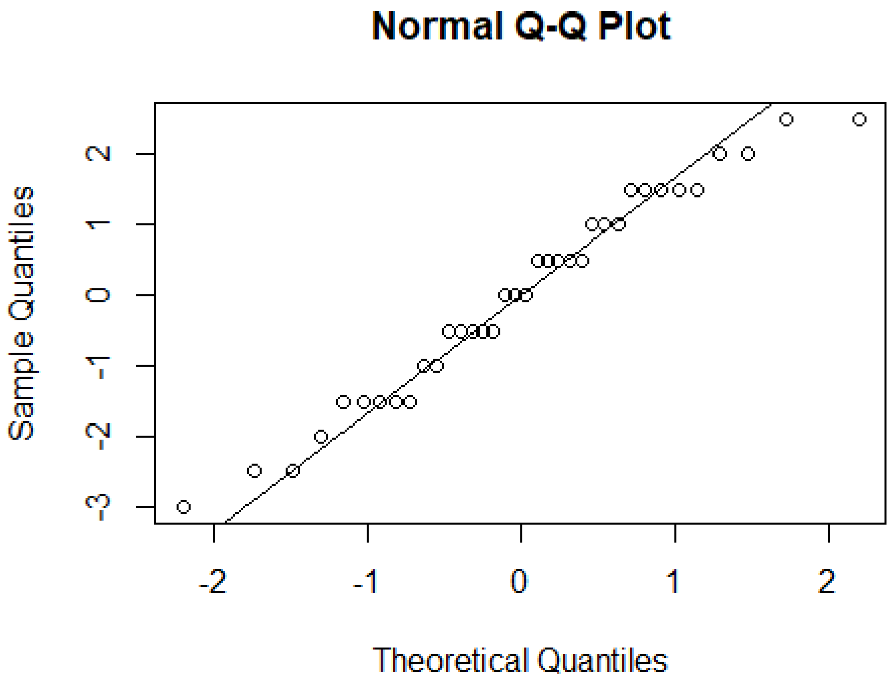
```
I = TukeyHSD(aov(calificacion ~ metodo))
I

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = calificacion ~ metodo)
##
## $metodo
##       diff       lwr        upr     p adj
## M2-M1 -2.5 -4.020241 -0.9797592 0.0008674
## M3-M1 -5.0 -6.520241 -3.4797592 0.0000000
## M3-M2 -2.5 -4.020241 -0.9797592 0.0008674

plot(I)
```



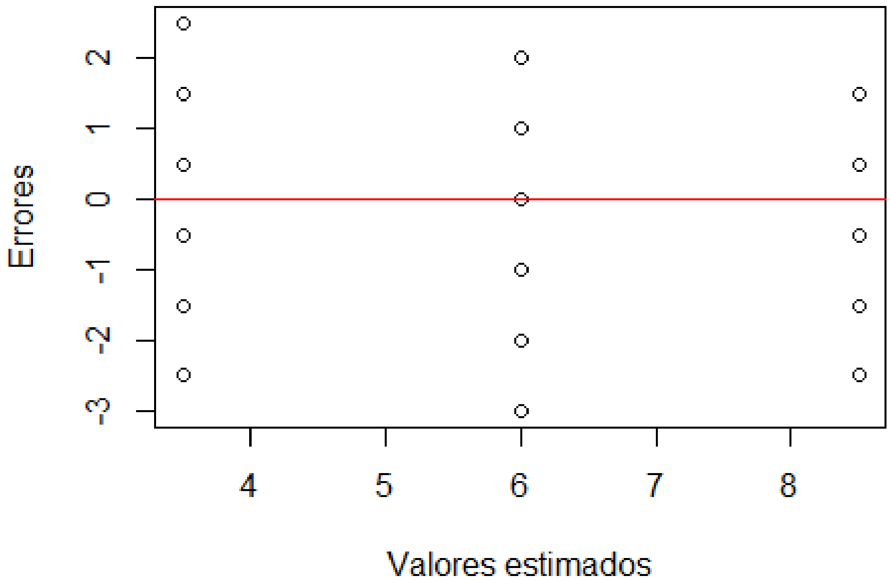
Aqui podemos ver claramente como las calificaciones varian de acuerdo con el metodo. El mejor método para obtener calificaciones es el primero, el cual otorga una media aproximada de 9, mientras que el peor es el tercero el cuál su peor valor es 1. Parece ser que el método dos no ayuda, ni empeora el rendimiento de un alumno.

Pruebas de Normalidad

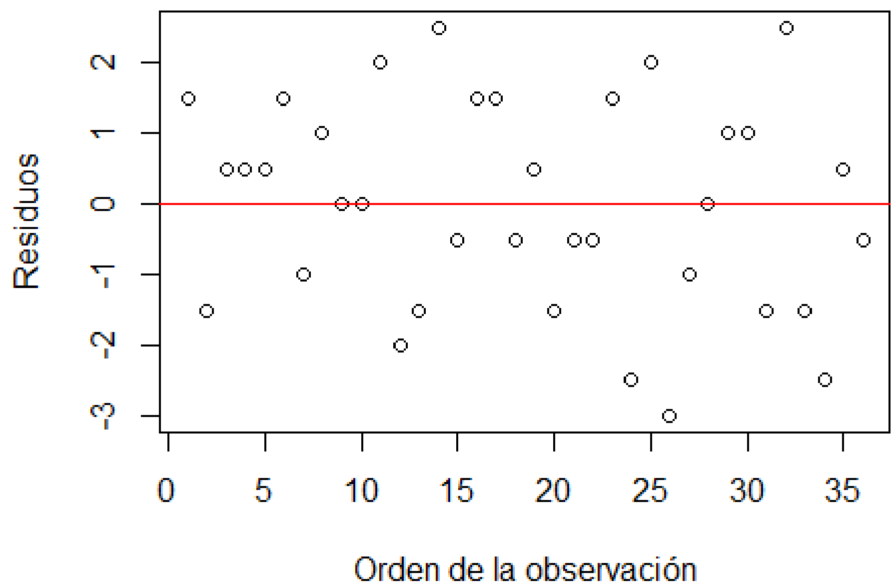
```
# Normalidad
residuos=anova$residuals
qqnorm(residuos)
qqline(residuos)
```




```
# Homocedastidad
plot(anova$fitted.values,anova$residuals,ylab="Errores",xlab="Valores
estimados")
abline(h=0,col="red")
```




```
#Independencia
n = tapply(calificacion, sexo, length)
plot(c(1:sum(n)),anova$residuals,xlab="Orden de la
observación",ylab="Residuos")
abline(h=0,col="red")
```




```
# Relacion lineal entre variables
modelo = lm(calificacion ~ metodo)
r2 = summary(modelo)$r.squared
r2

## [1] 0.6637168
```

Mientras que el problema muestra una clara relacion entre el método y las calificaciones, parece ser que el sexo no tiene ninguna relación con las calificaciones que se puedan llegar a obtener, ni con el método. El mejor modelo como previamente se dijo es el primero, el cuál mejora el rendimiento tanto en hombres como mujeres. Además este comportamiento los podemos observar en el Anova de la interaccion entre las dos efectos, pues el valor p de sexo, y la conjuncion de sexo y metodo son muy altos, mientras que el de método es muy alto, lo qe indica que este último es el que más afecta a los datos. Con esta información podemos rechazar nuestras hipotesis de que$h_0: t_i a_j = 0$ y de $h_0: T_i = 0$ diciendo que los valores p de sexo y la conjuncion de los dos efectos es 0. Mientras que la de método se aproxima mucho al valor de 0.

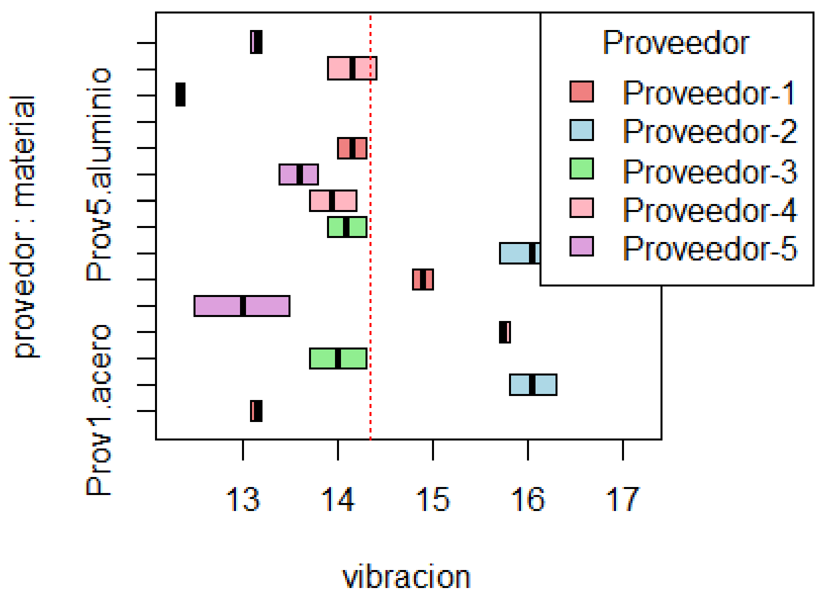
Problema 2. Resuelve las dos partes del problema “Vibración de motores”.

```
# Ingreso de datos
vibracion = c(13.1, 13.2, 15.0, 14.8, 14.0, 14.3, 16.3, 15.8, 15.7, 16.4,
17.2, 16.7, 13.7, 14.3, 13.9, 14.3, 12.4, 12.3, 15.7, 15.8, 13.7, 14.2,
14.4, 13.9, 13.5, 12.5, 13.4, 13.8, 13.2, 13.1)
provedor = c(rep("Prov1", 6), rep("Prov2", 6), rep("Prov3", 6),
rep("Prov4", 6), rep("Prov5", 6))
material = c(rep("acero", 2), rep("aluminio", 2), rep("plastico", 2))
```

```r
provedor = factor(provedor)
material = factor(material)
```

Analisis exploratorio.

```r
par(mar = c(5, 4, 4, 8))

datos = data.frame(vibracion, provedor, material)
boxplot(vibracion ~ provedor : material, datos, col = c("lightcoral",
"lightblue", "lightgreen", "lightpink", "plum"), horizontal = TRUE)
abline(v = mean(vibracion), lty= 3, col = "red")
legend("topright", legend = c("Proveedor-1", "Proveedor-2", "Proveedor-
3", "Proveedor-4", "Proveedor-5"),
       fill = c("lightcoral", "lightblue", "lightgreen", "lightpink",
"plum"),
       title = "Proveedor", inset = c(-0.3, 0), xpd = TRUE)
```




```r
# Va por proveedor y luego materiales (prov1-acero, prov2-acero, ...,
prov1-plastico, prov2-plastico)
```

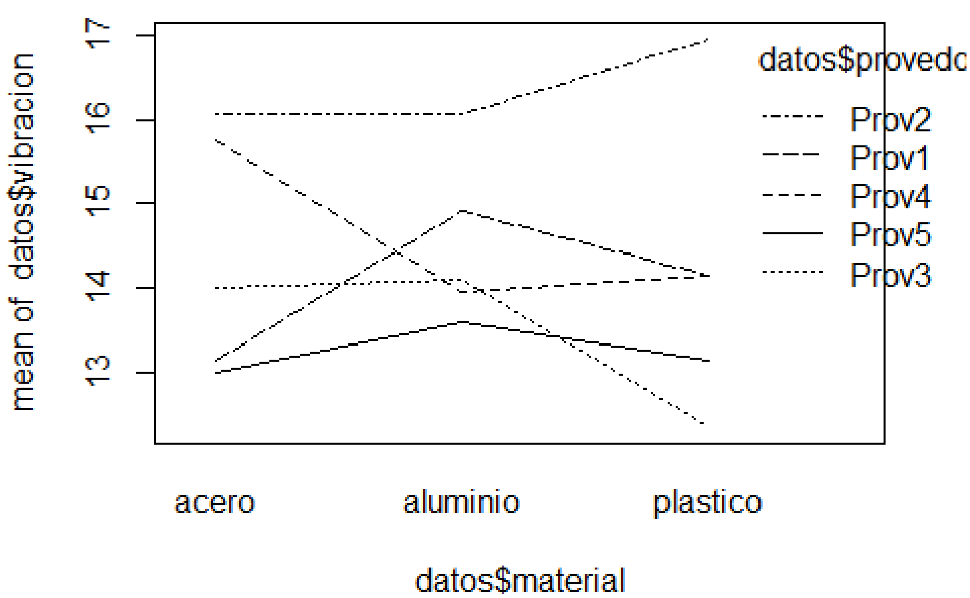
Viendo la distribucion por material y proveedor, observamos que lo que genera más vibracion son los productos creados por el proveedor 2, mientras que el proveedor 5, y el 4 son los que generan productos que tienen menores vibraciones. Los demás proveedores crean productos cuya vibracion esta más cercana a la media. De momento, podemos concluir que la causante del problema es el proveedor y no los materiales.

Establecimiento de hipotesis.

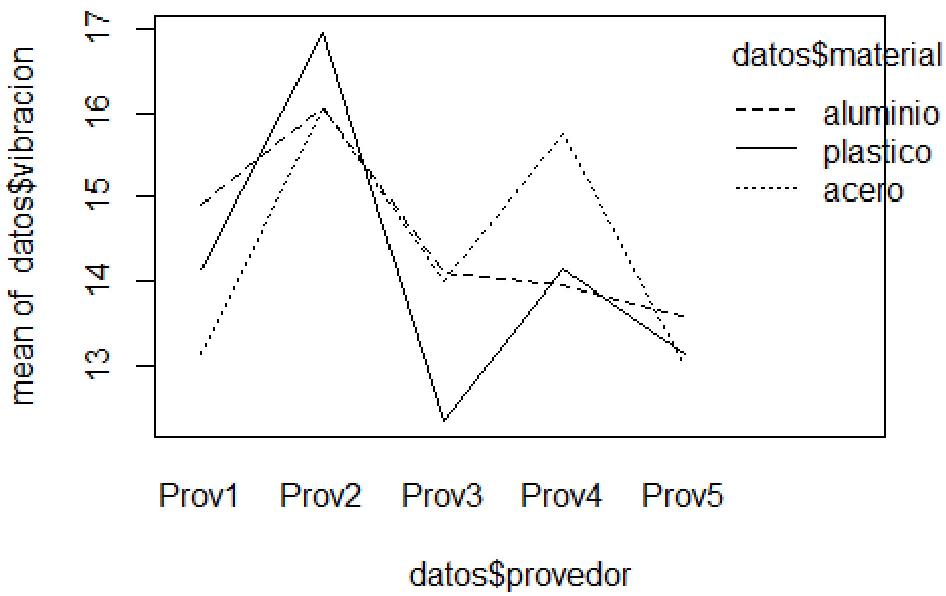
- $h_0: T_i = 0$ h_1: algún T_i es distinto de cero
- $h_0: a_j = 0$ h_1: algún a_j es distinto de cero
- $h_0: t_i a_j = 0$ h_1: aglún $t_i a_j$ es distinto de cero

ANOVA con dos niveles de interacción

```
anova = aov(vibracion ~ provedor*material, datos)
summary(anova)

##                   Df Sum Sq Mean Sq F value   Pr(>F)    
## provedor           4  36.67   9.169  82.353 5.07e-10 ***
## material           2   0.70   0.352   3.165   0.0713 .  
## provedor:material  8  11.61   1.451  13.030 1.76e-05 ***
## Residuals         15   1.67   0.111                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

interaction.plot(datos$material, datos$provedor, datos$vibracion)
```



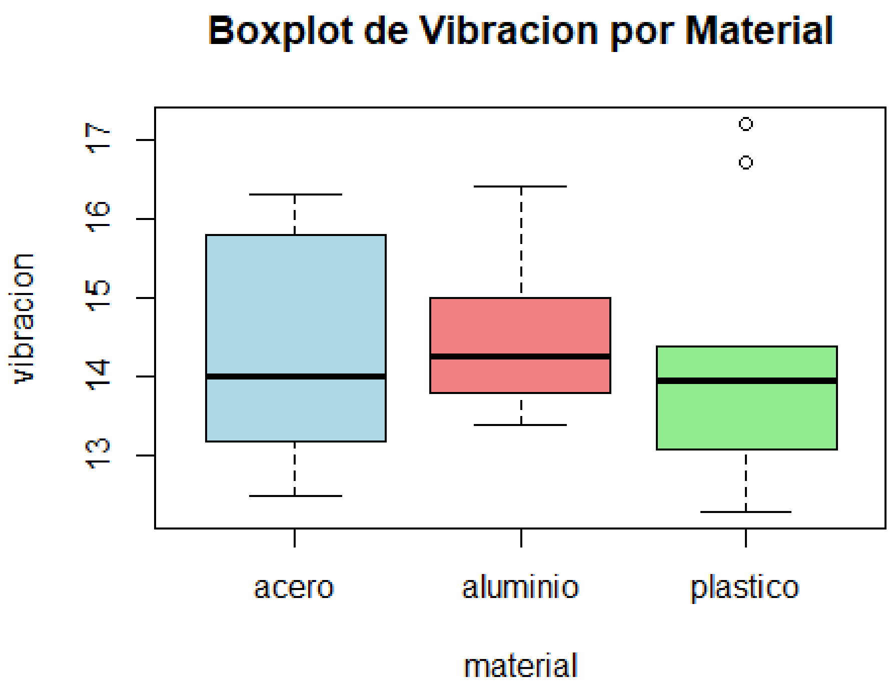

```
interaction.plot(datos$provedor, datos$material, datos$vibracion)
```



Basandonos de la información que nos otorga la gráfica de caja anterior, podemos decir que el provedor dos es el que genera productos que vibran más independientemente del material, sin embargo, no sabemos que material puede ser el causante de que se genere mayores vibraciones, o que otro provedor podría estar haciendo lo mismo; por lo que analizaremos uno por uno para encontrar cual de los dos efectos tiene un mayor o impacto, o si la conjunción de ambos genera más vibración.

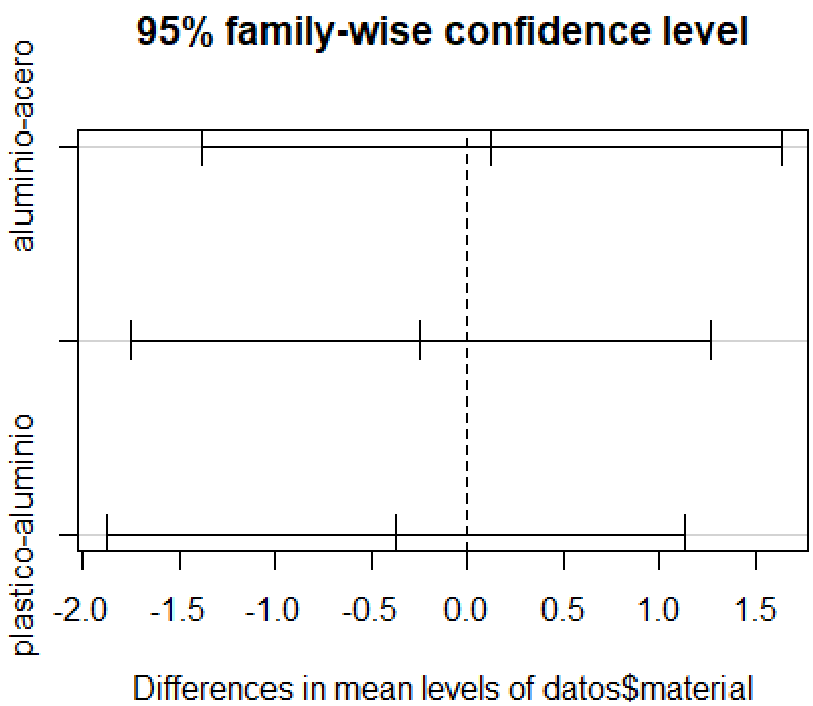
ANOVA con dos niveles sin interacción

```
anova = aov(vibracion ~ provedor + material, datos)
summary(anova)

##              Df Sum Sq Mean Sq F value   Pr(>F)    
## provedor      4  36.67   9.169   15.88 2.28e-06 ***
## material      2   0.70   0.352    0.61    0.552    
## Residuals    23  13.28   0.577                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Rendimiento por material
boxplot(vibracion ~ material, data = datos, col = c("lightblue",
"lightcoral", "lightgreen"), main = "Boxplot de Vibracion por Material")
```




```
C<-aov(vibracion ~ material, datos)
summary(C)

##             Df Sum Sq Mean Sq F value Pr(>F)
## material     2   0.70  0.3523    0.19  0.828
## Residuals   27  49.95  1.8500

tapply(datos$vibracion, datos$material, mean)

##    acero aluminio plastico
##    14.39    14.52    14.15

mean(vibracion)

## [1] 14.35333

I = TukeyHSD(aov(datos$vibracion ~ datos$material))
I

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = datos$vibracion ~ datos$material)
##
## $`datos$material`
##                     diff       lwr      upr     p adj
## aluminio-acero      0.13 -1.378171 1.638171 0.9751575
```

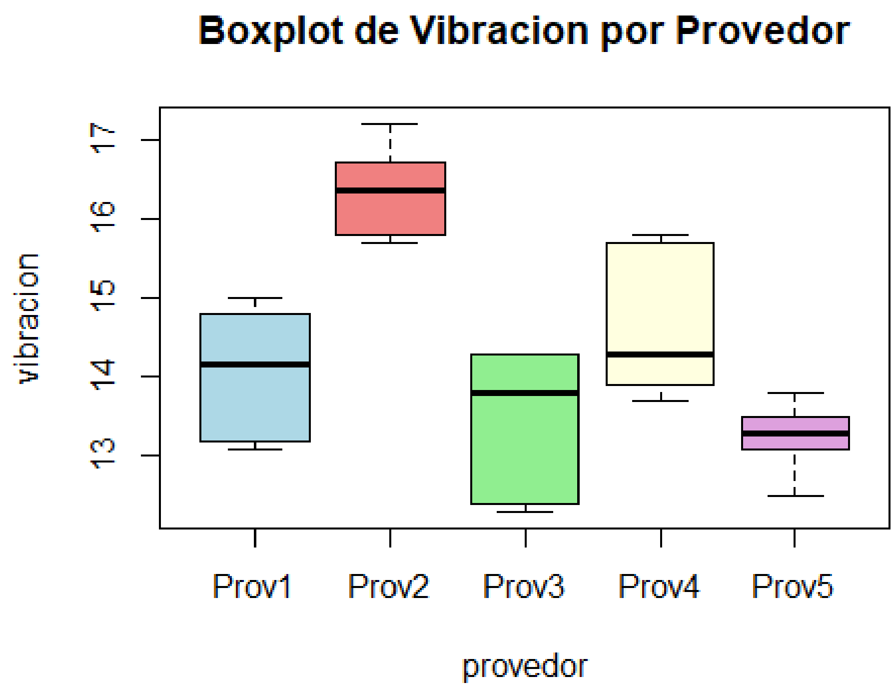
```
## plastico-acero    -0.24 -1.748171 1.268171 0.9180284
## plastico-aluminio -0.37 -1.878171 1.138171 0.8168495

plot(I)
```



Con el valor F y las gráficas dadas, podemos entender que no necesariamente los materiales son los responsables de la vibración. Si bien unos vibran más que otros, sucede por la naturaleza del mismo material, teniendo al acero como el material que más varía de los tres.

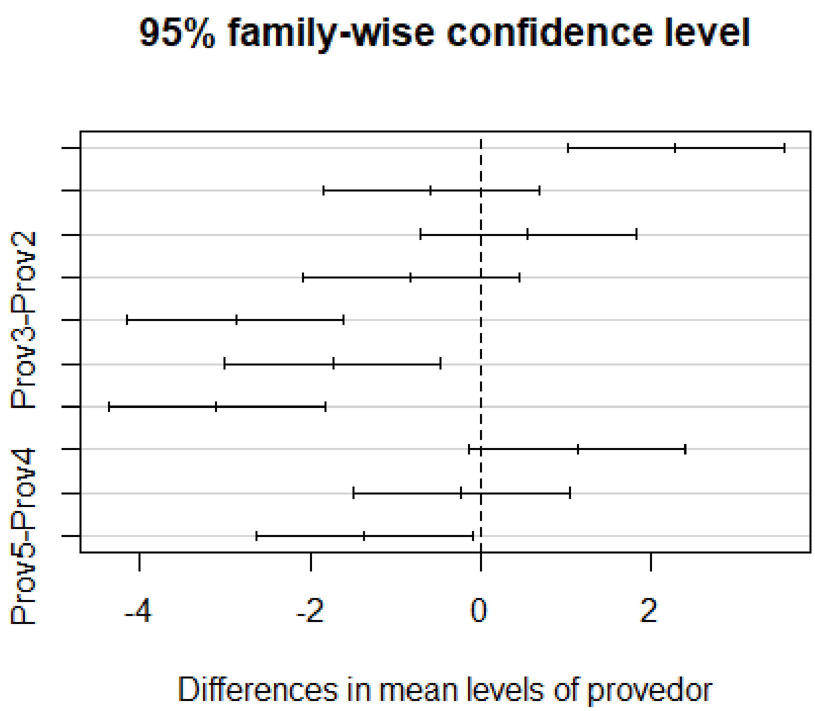
ANOVA con un efecto principal

```
anova = aov(vibracion ~ provedor, datos)
summary(anova)

##             Df Sum Sq Mean Sq F value   Pr(>F)    
## provedor     4  36.67   9.169    16.4 1.03e-06 ***
## Residuals   25  13.98   0.559                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Vibracion por provedor
boxplot(vibracion ~ provedor, data = datos, col = c("lightblue",
"lightcoral", "lightgreen", "lightyellow", "plum"), main = "Boxplot de
Vibracion por Provedor")
```




```
C<-aov(vibracion ~ provedor)
summary(C)

##             Df Sum Sq Mean Sq F value   Pr(>F)
## provedor     4  36.67   9.169    16.4 1.03e-06 ***
## Residuals   25  13.98   0.559
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

tapply(vibracion, provedor,mean)

##    Prov1    Prov2    Prov3    Prov4    Prov5
## 14.06667 16.35000 13.48333 14.61667 13.25000

mean(vibracion)

## [1] 14.35333

I = TukeyHSD(aov(vibracion ~ provedor))
I

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = vibracion ~ provedor)
##
## $provedor
##                    diff         lwr         upr     p adj
```

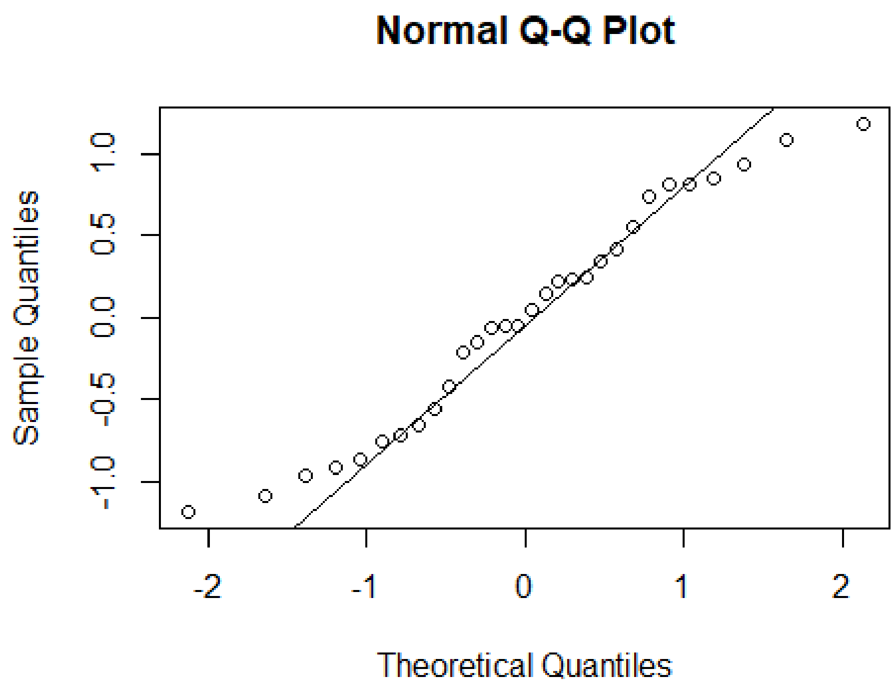
```
## Prov2-Prov1  2.2833333  1.0153666  3.55130006 0.0001595
## Prov3-Prov1 -0.5833333 -1.8513001  0.68463339 0.6630108
## Prov4-Prov1  0.5500000 -0.7179667  1.81796672 0.7089904
## Prov5-Prov1 -0.8166667 -2.0846334  0.45130006 0.3474956
## Prov3-Prov2 -2.8666667 -4.1346334 -1.59869994 0.0000055
## Prov4-Prov2 -1.7333333 -3.0013001 -0.46536661 0.0039774
## Prov5-Prov2 -3.1000000 -4.3679667 -1.83203328 0.0000015
## Prov4-Prov3  1.1333333 -0.1346334  2.40130006 0.0959316
## Prov5-Prov3 -0.2333333 -1.5013001  1.03463339 0.9821261
## Prov5-Prov4 -1.3666667 -2.6346334 -0.09869994 0.0301318

plot(I)
```



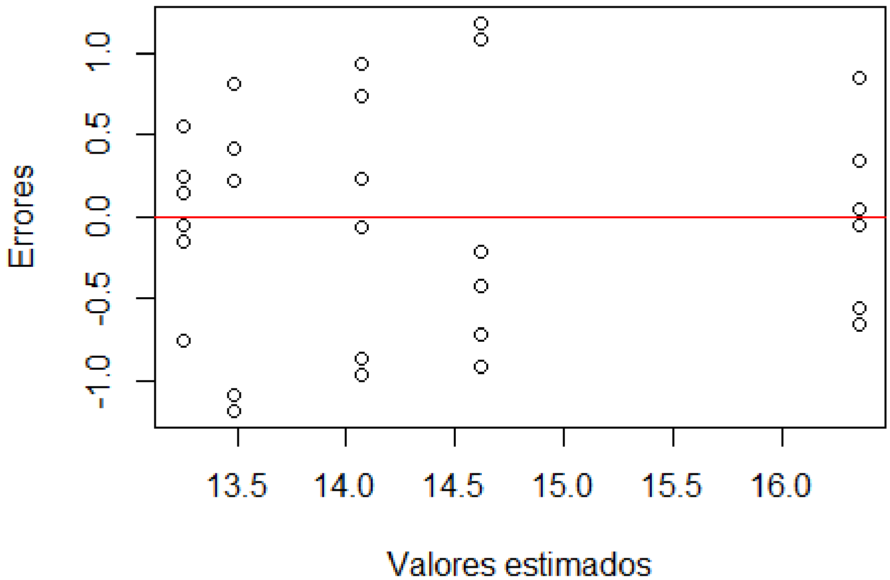
El valor p de provedor es muy bajo, lo que nos indica que tiene una gran relación con los datos. Esto se nos confirma con la gráfica de caja, la cual dice que el provedor dos es el que genera mayores vibraciones, independientemente del material que utilice, mientras que el provedor cinco es el que menos lo hace; el provedor tres llega a tener un minimo menor que el tres, sin embargo, cuenta con una variación mayor que la del cinco.

Pruebas de Normalidad

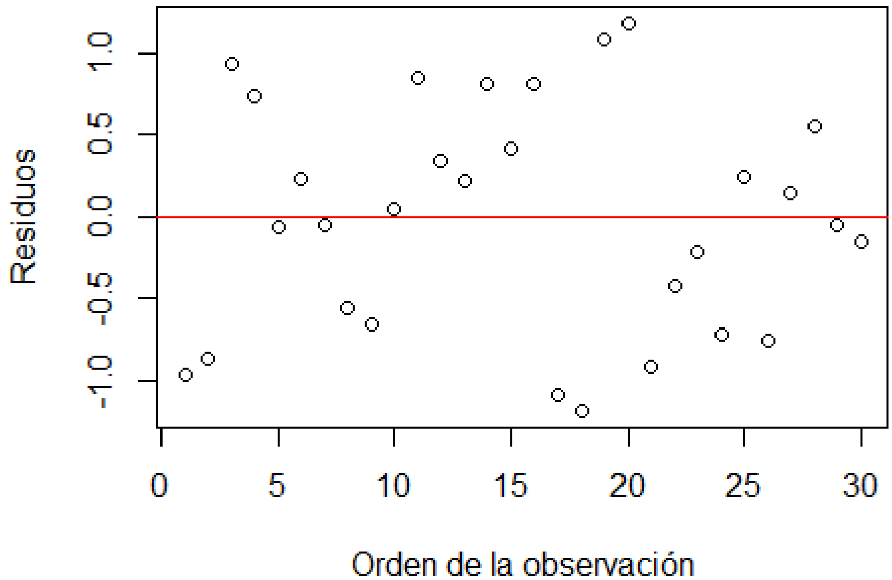
```
# Normalidad
residuos=anova$residuals
qqnorm(residuos)
qqline(residuos)
```




```r
# Homocedastidad
plot(anova$fitted.values,anova$residuals,ylab="Errores",xlab="Valores estimados")
abline(h=0,col="red")
```




```
#Independencia
n = tapply(vibracion, provedor, length)
plot(c(1:sum(n)),anova$residuals,xlab="Orden de la
observación",ylab="Residuos")
abline(h=0,col="red")
```




```r
# Relacion lineal entre variables
modelo = lm(vibracion ~ provedor)
r2 = summary(modelo)$r.squared
r2

## [1] 0.7240136
```

Dado la gráfica de QQPlot, y el valor de coeficiente de determinacion, podemos entender que dentro de los datos no existe una normalidad. Con toda la información recabada podemos decir que lo afecta principalmente a la vibracion es el provedor, que en este caso fue el segundo provedor, pues todos sus materiales eran los que más vibraban entre los demás provedores. Con esta información igual podemos rechazar una hipotesis nula; la de la interaccion entre los dos efectos, pues entre los dos efectos no exisitia una relación fuerte.