

Problemas de Análisis de la varianza

1. Doce personas son distribuidas en 4 grupos de personas 3 cada uno. A cada grupo, se le asigna aleatoriamente un tiempo distinto de entrenamiento antes de realizar una tarea. Los resultados en la mencionada tarea, con el correspondiente tiempo de entrenamiento, son los siguientes:

| 0.5 horas | 1 hora | 1.5 horas | 2 horas |
|-----------|--------|-----------|---------|
| 1 | 4 | 3 | 8 |
| 3 | 6 | 5 | 10 |
| 5 | 2 | 7 | 6 |

Ver si podemos rechazar la hipótesis nula $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$.

Solución

En primer lugar, tenemos que definir la tabla de datos para poder aplicar el test ANOVA:

```
tarea=c(1,3,5,4,6,2,3,5,7,8,10,6)
tiempo = as.factor(rep(c("0.5","1","1.5","2"),each=3))
(datos=data.frame(tarea,tiempo))
```

```
##      tarea tiempo
## 1         1    0.5
## 2         3    0.5
## 3         5    0.5
## 4         4     1
## 5         6     1
## 6         2     1
## 7         3    1.5
## 8         5    1.5
## 9         7    1.5
## 10        8     2
## 11       10     2
## 12        6     2
```

Una vez definida la tabla, realizamos el contraste ANOVA:

```
summary(aov(datos$tarea ~ datos$tiempo))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## datos$tiempo  3      42      14      3.5 0.0695 .
## Residuals    8      32       4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-valor está en la zona de penumbra, es decir, está entre 0.05 y 1. Por tanto, no podemos tomar una decisión clara. Si ponemos como umbral 0.05, podríamos concluir que no tenemos evidencias suficientes para rechazar que los resultados en el entrenamiento son distintos según el tiempo usado.

Aunque no se pide comprobaremos la igualdad de varianzas

```
bartlett.test(datos$tarea ~ datos$tiempo)
```

```
##  
## Bartlett test of homogeneity of variances  
##  
## data:  datos$tarea by datos$tiempo  
## Bartlett's K-squared = 0, df = 3, p-value = 1
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.3
```

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

```
leveneTest(datos$tarea ~ datos$tiempo)
```

```
## Levene's Test for Homogeneity of Variance (center = median)  
##      Df F value Pr(>F)  
## group 3      0      1  
##      8
```

Comprobemos las sumas de los cuadrados

```
ni=c(3,3,3,3)  
k=4  
N=sum(ni)  
SST= sum(datos$tarea^2)- sum(datos$tarea)^2/N  
SST
```

```
## [1] 74
```

```
Sumas_col=aggregate(datos$tarea,by=list(datos$tiempo),sum)  
Sumas_col$x/ni
```

```
## [1] 3 4 5 8
```

```
SSTr=sum(Sumas_col$x^2/ni)-sum(datos$tarea)^2/N  
SSTr
```

```
## [1] 42
```

```
SSE=SST-SSTr  
SSE
```

```
## [1] 32
```

2. Se registraron las frecuencias de los días que llovió a diferentes horas, durante los meses de enero, marzo, mayo y julio. Los datos obtenidos, durante un periodo de 10 años, fueron los siguientes:

| Hora | enero | febrero | marzo | julio | Total |
|-------|-------|---------|-------|-------|-------|
| 9 | 22 | 25 | 24 | 11 | 82 |
| 10 | 21 | 19 | 18 | 16 | 74 |
| 11 | 17 | 23 | 26 | 17 | 83 |
| 12 | 20 | 31 | 25 | 24 | 100 |
| 13 | 16 | 15 | 23 | 24 | 78 |
| 14 | 21 | 35 | 23 | 20 | 99 |
| Total | 117 | 148 | 139 | 112 | 536 |

Estudiar la variabilidad entre meses y entre horas.

Solución

En primer lugar, tenemos que definir la tabla de datos para poder aplicar el test ANOVA:

```
frecuencias = c(22,25,24,11,21,19,18,16,17,23,26,17,20,31,25,24,16,15,23,24,21,35,23,20)
horas = as.factor(rep(c("9","10","11","12","13","14"),each=4))
meses = as.factor(rep(c("enero","febrero","marzo","julio"),6))
(datos = data.frame(horas,meses,frecuencias))
```

```
##      horas  meses frecuencias
## 1      9  enero      22
## 2      9 febrero      25
## 3      9  marzo      24
## 4      9  julio      11
## 5     10  enero      21
## 6     10 febrero      19
## 7     10  marzo      18
## 8     10  julio      16
## 9     11  enero      17
## 10    11 febrero      23
## 11    11  marzo      26
## 12    11  julio      17
## 13    12  enero      20
## 14    12 febrero      31
## 15    12  marzo      25
## 16    12  julio      24
## 17    13  enero      16
## 18    13 febrero      15
## 19    13  marzo      23
## 20    13  julio      24
## 21    14  enero      21
## 22    14 febrero      35
## 23    14  marzo      23
## 24    14  julio      20
```

Una vez definida la tabla, realizamos el contraste ANOVA:

```
summary(aov(datos$frecuencias ~ datos$horas + datos$meses))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## datos$horas   5   149.5    29.90   1.395  0.282
## datos$meses   3   149.0    49.67   2.317  0.117
## Residuals    15   321.5    21.43
```

Como los p-valores por horas y por meses son grandes, concluimos que no tenemos evidencias para rechazar que el número de días que llueve por mes no depende ni del mes ni de la hora del día en que llueve.

3. Se realizó un estudio para determinar el nivel de agua y el tipo de planta sobre la longitud global del tronco de las plantas de guisantes. Se utilizaron 3 niveles de agua y 2 tipos de plantas. Se dispone para el estudio de 18 plantas sin hojas. Las plantas se dividen aleatoriamente en 3 subgrupos y después se los asigna los niveles de agua aleatoriamente. Se sigue un procedimiento parecido con 18 plantas convencionales. Se obtuvieron los resultados siguientes (la longitud del tronco se da en centímetros):

| FACTOR PLANTA | Sin Hojas | FACTOR AGUA | | |
|------------------|--------------|-------------|-------|-------|
| | | bajo | medio | alto |
| | | 69.0 | 96.1 | 121.0 |
| | | 71.3 | 102.3 | 122.9 |
| | | 73.2 | 107.5 | 123.1 |
| | | 75.1 | 103.6 | 125.7 |
| | | 74.4 | 100.7 | 125.2 |
| | Con Hojas | 75.0 | 101.8 | 120.1 |
| | | 71.1 | 81.0 | 101.1 |
| | | 69.2 | 85.8 | 103.2 |
| | | 70.4 | 86.0 | 106.1 |
| | | 73.2 | 87.5 | 109.7 |
| | | 71.2 | 88.1 | 109.0 |
| | | 70.9 | 87.6 | 106.9 |

Se desea saber si hay diferencias entre los niveles de agua y entre los diferentes tipos de planta. También se quiere saber si hay interacción entre los niveles de agua y el tipo de planta.

Solución

En primer lugar, tenemos que definir la tabla de datos para poder aplicar el test ANOVA:

```
longitud = c(69,96.1,121,71.3,102.3,122.9,73.2,107.5,123.1,75.1,103.6,125.7,74.4,100.7,125.2,
            75,101.8,120.1,71.1,81,101.1,69.2,85.8,103.2,70.4,86,106.1,73.2,87.5,109.7,
            71.2,88.1,109,70.9,87.6,106.9)
factor.agua = as.factor(rep(c("bajo", "medio", "alto"), 12))
factor.planta = as.factor(rep(c("sin hojas", "con hojas"), each=18))
(datos=data.frame(factor.agua, factor.planta, longitud))
```

```
##   factor.agua factor.planta longitud
## 1      bajo    sin hojas    69.0
## 2     medio    sin hojas    96.1
## 3      alto    sin hojas   121.0
## 4      bajo    sin hojas    71.3
## 5     medio    sin hojas   102.3
## 6      alto    sin hojas   122.9
## 7      bajo    sin hojas    73.2
## 8     medio    sin hojas   107.5
## 9      alto    sin hojas   123.1
## 10     bajo    sin hojas    75.1
## 11     medio    sin hojas   103.6
## 12     alto    sin hojas   125.7
## 13     bajo    sin hojas    74.4
## 14     medio    sin hojas   100.7
## 15     alto    sin hojas   125.2
## 16     bajo    sin hojas    75.0
## 17     medio    sin hojas   101.8
## 18     alto    sin hojas   120.1
## 19     bajo    con hojas    71.1
## 20     medio    con hojas    81.0
## 21     alto    con hojas   101.1
```

```
## 22      bajo      con hojas      69.2
## 23      medio     con hojas      85.8
## 24      alto      con hojas     103.2
## 25      bajo      con hojas      70.4
## 26      medio     con hojas      86.0
## 27      alto      con hojas     106.1
## 28      bajo      con hojas      73.2
## 29      medio     con hojas      87.5
## 30      alto      con hojas     109.7
## 31      bajo      con hojas      71.2
## 32      medio     con hojas      88.1
## 33      alto      con hojas     109.0
## 34      bajo      con hojas      70.9
## 35      medio     con hojas      87.6
## 36      alto      con hojas     106.9
```

Una vez definida la tabla, realizamos el contraste ANOVA:

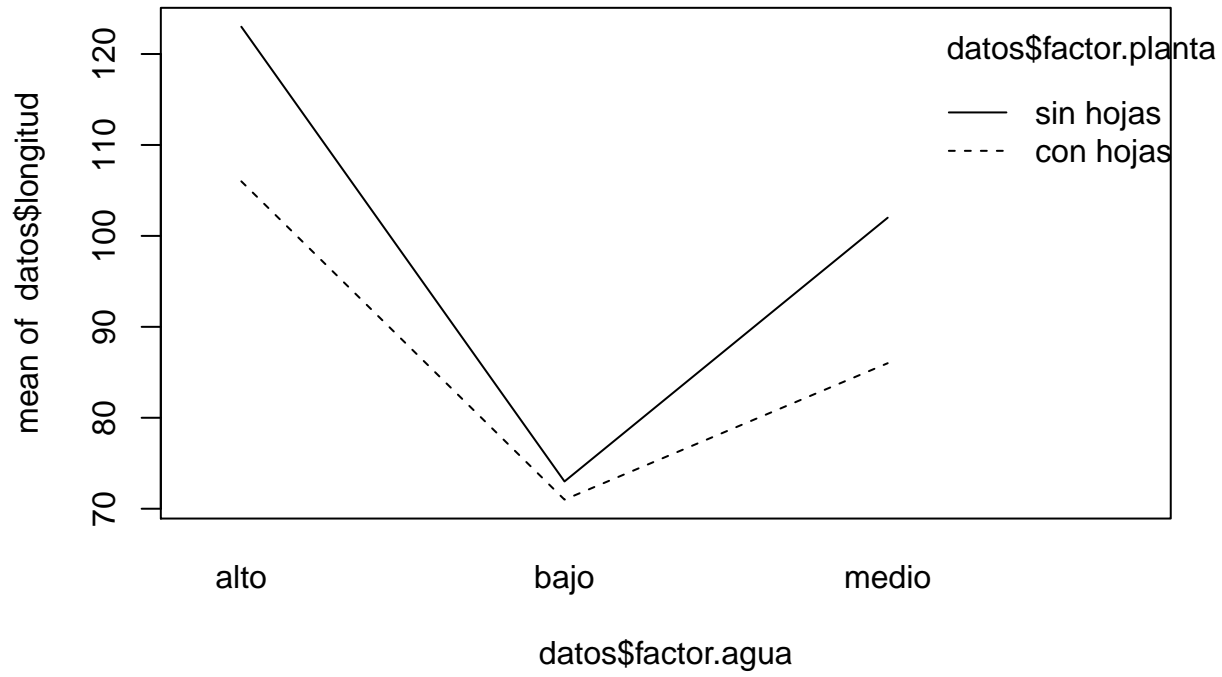
```
summary(aov(datos$longitud ~ datos$factor.agua * datos$factor.planta))
```

```
##                                Df Sum Sq Mean Sq F value    Pr(>F)
## datos$factor.agua              2  10842     5421   734.49 < 2e-16 ***
## datos$factor.planta            1   1225     1225   165.97 9.27e-14 ***
## datos$factor.agua:datos$factor.planta 2    422      211    28.59 1.12e-07 ***
## Residuals                     30    221         7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como todos los p-valores son pequeños, concluimos lo siguiente:

- tenemos evidencias suficientes para afirmar que la longitud de la planta depende del nivel de agua,
- tenemos evidencias suficientes para afirmar que la longitud de la planta depende del tipo de planta, es decir, si ésta es sin hojas o con hojas y,
- tenemos evidencias suficientes para afirmar que existe interacción entre el nivel de agua y el tipo de planta. Realizemos un gráfico de la interacción para comprobar gráficamente dicha evidencia:

```
interaction.plot(datos$factor.agua,datos$factor.planta,datos$longitud)
```



Observamos que los segmentos anteriores están lejos de ser paralelos.

4. Las variables aleatorias X_i siguen la distribución $N(m_i, \sigma^2)$, $i = 1, 2, 3, 4$. Consideramos las siguientes muestras de tamaños $n_i = 7$ de las mencionadas variables aleatorias:

| | | | | | | | |
|-------|----|----|----|----|----|----|----|
| X_1 | 20 | 26 | 26 | 24 | 23 | 26 | 21 |
| X_2 | 24 | 22 | 20 | 21 | 21 | 22 | 20 |
| X_3 | 16 | 18 | 20 | 21 | 24 | 15 | 17 |
| X_4 | 19 | 15 | 13 | 16 | 12 | 11 | 14 |

- Comprobar si las varianzas son iguales.
- Contrastar la igualdad de medias.

Solución

En primer lugar, tenemos que definir la tabla de datos para poder aplicar el test ANOVA:

```
valores=c(20,26,26,24,23,26,21,24,22,20,21,21,22,20,16,18,20,21,
          24,15,17,19,15,13,16,12,11,14)
variable.aleatoria = as.factor(rep(c("X1","X2","X3","X4"),each=7))
(datos=data.frame(valores,variable.aleatoria))
```

```
##      valores variable.aleatoria
## 1         20                X1
## 2         26                X1
## 3         26                X1
## 4         24                X1
## 5         23                X1
## 6         26                X1
## 7         21                X1
## 8         24                X2
## 9         22                X2
## 10        20                X2
## 11        21                X2
## 12        21                X2
## 13        22                X2
## 14        20                X2
## 15        16                X3
## 16        18                X3
## 17        20                X3
## 18        21                X3
## 19        24                X3
## 20        15                X3
## 21        17                X3
## 22        19                X4
## 23        15                X4
## 24        13                X4
## 25        16                X4
## 26        12                X4
## 27        11                X4
## 28        14                X4
```

Para contrastar si las varianzas son iguales, usamos el test de Bartlett:

```
bartlett.test(valores ~ variable.aleatoria)
```



```
##
## Bartlett test of homogeneity of variances
##
## data:  valores by variable.aleatoria
## Bartlett's K-squared = 3.4291, df = 3, p-value = 0.3301
```

Como el p-valor es grande, concluimos que no tenemos evidencias suficientes para rechazar que las varianzas de las muestras de las 4 variables aleatorias no sean iguales.

Contrastemos a continuación si las medias son iguales usando el test ANOVA:

```
summary(aov(valores ~ variable.aleatoria))
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## variable.aleatoria  3     345   114.99    18.16 2.29e-06 ***
## Residuals        24     152     6.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como el p-valor es muy pequeño concluimos que tenemos evidencias suficientes para afirmar que las medias de las 4 variables aleatorias no son iguales.

Comprobemos las sumas de cuadrados del ANOVA

```
summary(aov(valores ~ variable.aleatoria))->sol_aov
ni=c(7,7,7,7)
k=4
N=sum(ni)
SST= sum(valores^2)- sum(valores)^2/N
SST
```

```
## [1] 496.9643
```

```
Sumas_col=aggregate(valores,by=list(variable.aleatoria),sum)
Sumas_col$x/ni
```

```
## [1] 23.71429 21.42857 18.71429 14.28571
```

```
SSTr=sum(Sumas_col$x^2/ni)-sum(valores)^2/N
SSTr
```

```
## [1] 344.9643
```

```
SSE=SST-SSTr
SSE
```

```
## [1] 152
```

Comparamos con los resultados del summary

```
summary(aov(valores ~ variable.aleatoria))
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## variable.aleatoria  3     345   114.99    18.16 2.29e-06 ***
## Residuals        24     152     6.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pairwise.t.test(valores,variable.aleatoria,p.adjust.method = "none")
```

```
##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data:  valores and variable.aleatoria  
##  
##      X1      X2      X3  
## X2 0.1022 -      -  
## X3 0.0011 0.0549 -  
## X4 3.0e-07 1.9e-05 0.0031  
##  
## P value adjustment method: none
```

```
pairwise.t.test(valores,variable.aleatoria,p.adjust.method = "bonferroni" )
```

```
##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data:  valores and variable.aleatoria  
##  
##      X1      X2      X3  
## X2 0.61328 -      -  
## X3 0.00644 0.32957 -  
## X4 1.8e-06 0.00011 0.01842  
##  
## P value adjustment method: bonferroni
```

```
pairwise.t.test(valores,variable.aleatoria,p.adjust.method = "holm" )
```

```
##  
## Pairwise comparisons using t tests with pooled SD  
##  
## data:  valores and variable.aleatoria  
##  
##      X1      X2      X3  
## X2 0.1099 -      -  
## X3 0.0043 0.1099 -  
## X4 1.8e-06 9.5e-05 0.0092  
##  
## P value adjustment method: holm
```

```
library(agricolae)
```

```
## Warning: package 'agricolae' was built under R version 3.6.3
```

```
resultado.anova=aov(valores~variable.aleatoria)  
duncan.test(resultado.anova,"variable.aleatoria",group=TRUE,alpha = 0.05)$group
```

```
##      valores groups  
## X1 23.71429      a  
## X2 21.42857     ab  
## X3 18.71429      b  
## X4 14.28571      c
```