DISEÑO ESTADÍSTICO DE EXPERIMENTOS

Caterine Melissa Guerrero España

2022-09-27

Diseño Completamente Aleatorio con efectos fijos (Diseño unifactorial de efectos fijos)

El primer diseño que presentamos es el diseño completamente aleatorio de efectos fijos y la técnica estadística es el análisis de la varianza de una vía o un factor. La descripción del diseño así como la terminología subyacente la vamos a introducir mediante el siguiente supuesto práctico.

Supuesto práctico 1

La contaminación es uno de los problemas ambientales más importantes que afectan a nuestro mundo. En las grandes ciudades, la contaminación del aire se debe a los escapes de gases de los motores de explosión, a los aparatos domésticos de la calefacción, a las industrias,... El aire contaminado nos afecta en nuestro vivir diario, manifestándose de diferentes formas en nuestro organismo. Con objeto de comprobar la contaminación del aire en una determinada ciudad, se ha realizado un estudio en el que se han analizado las concentraciones de monóxido de carbono (CO) durante cinco días de la semana (lunes, martes, miércoles, jueves y viernes).

Días de la semana	Concentraciones de monóxido de carbono							
Lunes	420	390	480	430	440	324	450	460
Martes	450	390	430	521	320	360	342	423
Miércoles	355	462	286	238	344	423	123	196
Jueves	321	254	412	368	340	258	433	489
Viernes	238	255	366	389	198	256	248	324

En el ejemplo disponemos de una colección de 40 unidades experimentales y queremos estudiar el efecto de las concentraciones de monóxido de carbono en 5 días distintos. Es decir, estamos interesados en contrastar el efecto de un solo factor, que se presenta con cinco niveles, sobre la variable respuesta.

```
setwd("C:/Users/user/OneDrive/Paquete R/PRACTICAS-S9")
contaminacion<-read.table("supuesto1.txt",header = TRUE)
contaminacion
```

```
##    Concentracion Dia
## 1            420   1
## 2            390   1
## 3            480   1
## 4            430   1
## 5            440   1
## 6            324   1
## 7            450   1
## 8            460   1
```

```
## 9            450   2
## 10           390   2
## 11           430   2
## 12           521   2
## 13           320   2
## 14           360   2
## 15           342   2
## 16           423   2
## 17           355   3
## 18           462   3
## 19           286   3
## 20           238   3
## 21           344   3
## 22           423   3
## 23           123   3
## 24           196   3
## 25           321   4
## 26           254   4
## 27           412   4
## 28           368   4
## 29           340   4
## 30           258   4
## 31           433   4
## 32           489   4
## 33           238   5
## 34           255   5
## 35           366   5
## 36           389   5
## 37           198   5
## 38           256   5
## 39           248   5
## 40           324   5
```

1. Tranformar la variable referente a los niveles del factor fijo como factor

```
contaminacion$dia<-factor(contaminacion$Dia)
contaminacion$dia
```

```
##  [1] 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 5 5 5 5 5 5
## [39] 5 5
## Levels: 1 2 3 4 5
```

Para calcular la tabla ANOVA primero hacemos uso de la función "aov" de la siguiente forma:

```
mod<-aov(Concentracion~Dia, data = contaminacion)
```

donde:

- Concentracion = nombre de la columna de las observaciones.
- Dia = nombre de la columna en la que están representados los tratamientos.
- data= data.frame en el que están guardados los datos

```
mod
```

```
## Call:
##    aov(formula = Concentracion ~ Dia, data = contaminacion)
##
## Terms:
```

```
##                       Dia Residuals
## Sum of Squares   84565.01 253868.09
## Deg. of Freedom         1        38
##
## Residual standard error: 81.73579
## Estimated effects may be unbalanced
```

Se puede mostrar un resumen de los resultados con la función "summary" (verdadera tabla ANOVA)

```
summary(mod)#tabla anova
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Dia          1  84565   84565   12.66 0.00102 **
## Residuals   38 253868    6681
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si el valor de F es mayor que uno quiere decir que hay un efecto positivo del factor día. Se observa que el P-valor (Sig.) tiene un valor de 0.00102, que es menor que el nivel de significación 0.05. Por lo tanto, hemos comprobado estadísticamente que estos cinco grupos son distintos. Es decir, existen diferencias significativas en las concentraciones medias de monóxido de carbono entre los cinco días de la semana. Por lo tanto no se puede rechazar la hipótesis alternativa que dice que al menos dos grupos son diferentes, pero ¿Cuáles son esos grupos? ¿Los cinco grupos son distintos o sólo alguno de ellos? Pregunta que resolveremos más adelante mediante los contrastes de comparaciones múltiples.

2. En la expresión del comando "aov" indicar el factor

```
mod1<-aov(Concentracion~factor(Dia),data = contaminacion)
mod1
```

```
## Call:
##    aov(formula = Concentracion ~ factor(Dia), data = contaminacion)
##
## Terms:
##                 factor(Dia) Residuals
## Sum of Squares     119484.4  218948.8
## Deg. of Freedom           4        35
##
## Residual standard error: 79.09285
## Estimated effects may be unbalanced
```

También se puede utilizar el comando "anova" y no es necesario el comando "summary"

```
mod2<-anova(lm(Concentracion~factor(Dia),data = contaminacion))
mod2
```

```
## Analysis of Variance Table
##
## Response: Concentracion
##             Df Sum Sq Mean Sq F value   Pr(>F)
## factor(Dia)  4 119484 29871.1   4.775 0.003518 **
## Residuals   35 218949  6255.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los datos pueden venir dados en diferentes formatos:

1. Caso en el que los datos se muestran de forma que se analiza la contaminación con cada uno de los dias de la semana (de lunes a viernes). Como se muestra a continuación

```
contaminacion<-read.table("supuesto1-1.txt",header = TRUE)
contaminacion
```

```
##   Lunes Martes Miercoles Jueves Viernes
## 1   420    450       355    321     238
## 2   390    390       462    254     255
## 3   480    430       286    412     366
## 4   430    521       238    368     389
## 5   440    320       344    340     198
## 6   324    360       423    258     256
## 7   450    342       123    433     248
## 8   460    423       196    489     324
```

En primer lugar apilaremos las columnas, para ello utilizamos el comando “stack” de la siguiente forma

```
trats<-stack(contaminacion)
trats
```

```
##    values       ind
## 1     420     Lunes
## 2     390     Lunes
## 3     480     Lunes
## 4     430     Lunes
## 5     440     Lunes
## 6     324     Lunes
## 7     450     Lunes
## 8     460     Lunes
## 9     450    Martes
## 10    390    Martes
## 11    430    Martes
## 12    521    Martes
## 13    320    Martes
## 14    360    Martes
## 15    342    Martes
## 16    423    Martes
## 17    355 Miercoles
## 18    462 Miercoles
## 19    286 Miercoles
## 20    238 Miercoles
## 21    344 Miercoles
## 22    423 Miercoles
## 23    123 Miercoles
## 24    196 Miercoles
## 25    321    Jueves
## 26    254    Jueves
## 27    412    Jueves
## 28    368    Jueves
## 29    340    Jueves
## 30    258    Jueves
## 31    433    Jueves
## 32    489    Jueves
## 33    238   Viernes
## 34    255   Viernes
## 35    366   Viernes
## 36    389   Viernes
```

```
## 37    198   Viernes
## 38    256   Viernes
## 39    248   Viernes
## 40    324   Viernes
```

Nos muestra dos columnas:

- La primera columna: values nos muestra los valores de la variable respuesta. En este caso la contaminación
- La segunda columna: ind nos muestra los diferentes tratamientos Podemos realizar el Análisis de la varianza utilzando el comando anova

```
anova(lm(values~ind,data = trats))
```

```
## Analysis of Variance Table
##
## Response: values
##           Df Sum Sq Mean Sq F value   Pr(>F)
## ind        4 119484 29871.1   4.775 0.003518 **
## Residuals 35 218949  6255.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. . Los datos vienen dados de la siguiente forma:

Lunes: 420, 390, 480, 430, 440, 324, 450, 460

Martes: 450, 390, 430, 521, 320, 360, 342, 423

Miércoles: 355, 462, 286, 238, 344, 423, 123, 196

Jueves: 321, 254, 412, 368, 340, 258, 433, 489

Viernes: 238, 255, 366, 389, 198, 256, 248, 324

Se crean cinco vectores, cada uno de ellos representando la contaminación con un tratamiento.

```
Lu=c(420,390,480,430,440,324,450,460)
Ma=c(450,390,430,521,320,360,342,423)
Mi=c(355,462,286,238,344,423,123,196)
Ju=c(321,254,412,368,340,258,433,489)
Vi=c(238,255,366,389,198,256,248,324)
```

Acontinuación creamos un data.frame para poder resolver el ANOVA

```
datos=data.frame(Lu,Ma,Mi,Ju,Vi)
datos
```

```
##    Lu  Ma  Mi  Ju  Vi
## 1 420 450 355 321 238
## 2 390 390 462 254 255
## 3 480 430 286 412 366
## 4 430 521 238 368 389
## 5 440 320 344 340 198
## 6 324 360 423 258 256
## 7 450 342 123 433 248
## 8 460 423 196 489 324
```

De esta forma hemos creado una nueva base de datos que hemos llamado “datos“. Para resolver el ANOVA tenemos primero que apilar las columnas con el comando “stack”

```
datos1<-stack(datos)
datos1
```

```
##    values ind
## 1     420  Lu
## 2     390  Lu
## 3     480  Lu
## 4     430  Lu
## 5     440  Lu
## 6     324  Lu
## 7     450  Lu
## 8     460  Lu
## 9     450  Ma
## 10    390  Ma
## 11    430  Ma
## 12    521  Ma
## 13    320  Ma
## 14    360  Ma
## 15    342  Ma
## 16    423  Ma
## 17    355  Mi
## 18    462  Mi
## 19    286  Mi
## 20    238  Mi
## 21    344  Mi
## 22    423  Mi
## 23    123  Mi
## 24    196  Mi
## 25    321  Ju
## 26    254  Ju
## 27    412  Ju
## 28    368  Ju
## 29    340  Ju
## 30    258  Ju
## 31    433  Ju
## 32    489  Ju
## 33    238  Vi
## 34    255  Vi
## 35    366  Vi
## 36    389  Vi
## 37    198  Vi
## 38    256  Vi
## 39    248  Vi
## 40    324  Vi
```

Resolvemos el ANOVA como en el caso anterior

```
anova(lm(values~ind,data = datos1))
```

```
## Analysis of Variance Table
##
## Response: values
##           Df Sum Sq Mean Sq F value   Pr(>F)
## ind        4 119484 29871.1   4.775 0.003518 **
## Residuals 35 218949  6255.7
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. Los datos se muestren en un solo vector que tiene todos los datos de la contaminación tanto si se ha medido el lunes, el martes, el miércoles, el jueves o el viernes

```
contaminacion=c(Lu,Ma,Mi,Ju,Vi)
contaminacion
```

```
##  [1] 420 390 480 430 440 324 450 460 450 390 430 521 320 360 342 423 355 462 286
## [20] 238 344 423 123 196 321 254 412 368 340 258 433 489 238 255 366 389 198 256
## [39] 248 324
```

Este vector esta formado por los 40 datos que podemos comprobarlo con el comando length

```
length(contaminacion)
```

```
## [1] 40
```

Para realizar el ANOVA, ya tenemos los datos de la variable respuesta y a continuación tenemos que crear el factor tratamiento, para ello vamos a utilizar la función generador de niveles, gl, y le decimos que nos genere 5 niveles que son los cinco tratamientos, cada uno repetido 8 veves con un total de 40 datos y para identificar que nivel es cada uno, creamos las etiquetas L, M, Mi, J y V

```
trat=gl(5,8,40,labels = c("L","M","Mi","J","V"))
trat
```

```
##  [1] L  L  L  L  L  L  L  L  M  M  M  M  M  M  M  M  Mi Mi Mi Mi Mi Mi Mi Mi J
## [26] J  J  J  J  J  J  J  V  V  V  V  V  V  V  V
## Levels: L M Mi J V
```

```
anova(lm(contaminacion~trat))
```

```
## Analysis of Variance Table
##
## Response: contaminacion
##           Df Sum Sq Mean Sq F value   Pr(>F)
## trat       4 119484 29871.1   4.775 0.003518 **
## Residuals 35 218949  6255.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El modelo que hemos propuesto hay que validarlo, para ello hay que comprobar si se verifican las hipótesis básicas del modelo, es decir, si las perturbaciones son variables aleatorias independientes con distribución normal de media 0 y varianza constante (homocedasticidad).

Estudio de la Idoneidad del modelo

Como hemos dicho anteriormente, validar el modelo propuesto consiste en estudiar si las hipótesis básicas del modelo están o no en contradicción con los datos observados. Es decir si se satisfacen los supuestos del modelo: Normalidad, Independencia, Homocedasticidad. Para ello utilizamos procedimientos gráficos y analíticos.

Hipótesis de Normalidad

en primer lugar, analizamos la normalidad de las concentraciones y continumos con el análisis de la normalidad de los residuos

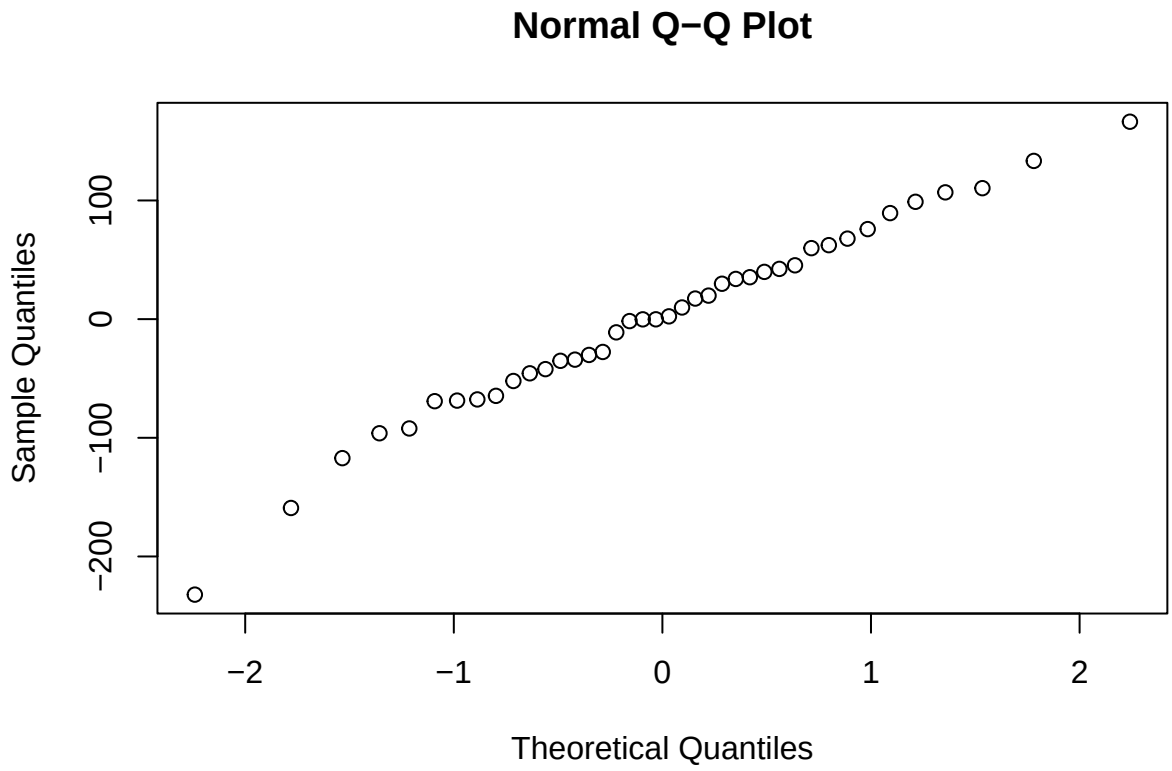
```
shapiro.test(mod$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mod$residuals
## W = 0.98479, p-value = 0.8578
```

Observamos con el test de Shapiro-Wilk que es adecuado cuando las muestras son pequeñas ($n < 50$) y es una alternativa más potente que el test de Kolmogorov-Smirnov. El p-valor es mayor que el nivel de significación del 5%, concluyendo que las muestras de las concentraciones se distribuyen de forma normal en cada día de la semana.

Podemos verlo también gráficamente con la orden "qqnorm"

```
qqnorm(mod$residuals)
```



Podemos apreciar en este gráfico que los puntos aparecen próximos a la línea diagonal. Esta gráfica no muestra una desviación marcada de la normalidad.

Hipótesis de homocedasticidad

Para comprobar la hipótesis de igualdad entre las varianzas del factor utilizamos el Test de Barlett.

```
contaminacion<-read.table("supuesto1.txt",header = TRUE)
bartlett.test(contaminacion$Concentracion, contaminacion$Dia)
```

```
##
##  Bartlett test of homogeneity of variances
```

```
##
## data:  contaminacion$Concentracion and contaminacion$Dia
## Bartlett's K-squared = 5.4942, df = 4, p-value = 0.2402
```

El p-valor es del 0.2402 que al ser mayor del nivel significación usual del 5% no podemos rechazar la hipótesis de igualdad de varianzas, es decir, se acepta la igualdad de varianzas en el factor.

Hipótesis de independencia

Para comprobar que se satisface el supuesto de independencia entre los residuos analizamos el gráfico de los residuos frente a los valores pronosticados o predichos por el modelo. El empleo de este gráfico es útil puesto que la presencia de alguna tendencia en el mismo puede ser indicio de una violación de dicha hipótesis. En R obtenemos varios gráficos a la vez que están incluidos en la estimación del modelo.

Para verlos de forma correcta hacemos uso de las siguientes órdenes:

```
contaminacion<-read.table("supuesto1.txt",header = TRUE)
mod<-aov(Concentracion~Dia, data = contaminacion)
layout(matrix(c(1,2,3,4),2,2))#para que salgan en la misma pantalla
plot(mod)
```