UNIVERSIDAD DE EL SALVADOR.

FACULTAD MULTIDISCIPLINARIA DE OCCIDENTE.

DEPARTAMENTO DE MATEMATICAS.

PRACTICA 7. SEMANA 9

CARRERA:

LICENCIATURA EN ESTADISTICA.

ASIGNATURA:

ANALISIS ESTADISTICO CON EL PAQUETE R

DOCENTE:

JAIME ISAAC PEÑA

PRESENTADO POR:

NELSON DE JESUS MAGAÑA GODINEZ

FECHA:

20 SEPTIEMBRE DE 2022

1 DISEÑO ESTADÍSTICO DE EXPERIMENTOS

1.1 Introducción al Diseño Estadístico de Experimentos

En la práctica 6 hemos descrito métodos de inferencias sobre la media y la varianza de una población y de dos poblaciones. En esta práctica 7 ampliamos dichos métodos a más de dos poblaciones e introducimos algunos aspectos elementales del Diseño Estadístico de Experimentos y del Análisis de la Varianza.

El diseño estadístico de experimentos incluye un conjunto de técnicas de análisis y un método de construcción de modelos estadísticos que, conjuntamente, permiten llevar a cabo el proceso completo de planificar un experimento para obtener datos apropiados, que puedan ser analizados con métodos estadísticos, con objeto de obtener conclusiones válidas y objetivas.

El análisis de la varianza o abreviadamente ANOVA (del inglés analysis of variance) es un procedimiento estadístico que permite dividir la variabilidad observada en componentes independientes que pueden atribuirse a diferentes causas de interés. Es una técnica estadística para comparar más de dos grupos, es decir un método para comparar más de dos tratamientos y la variable de estudio o variable respuesta es numérica.

En esta práctica presentamos el Diseño Completamente Aleatorio con efectos fijos y con efectos aleatorios, el Diseño en Bloques Completos Aleatorizados, Diseño en Bloques Incompletos Balanceados (BIB), el Diseño en Cuadrados Latinos, el Diseño en Cuadrados Greco-Latinos, el Diseño en Cuadrados de Jouden, el Diseño Bifactorial de efectos fijos y el Diseño Trifactorial de efectos fijos.

1.1.1 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).

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.

Nos interesa saber si las concentraciones medias de monóxido de carbono son iguales en los cinco días de la semana, para ello realizamos el siguiente contraste de hipótesis:

$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu$

$\mu_i \neq \mu_j$ prara alguin $i \neq j$

Es decir, contrastamos que no hay diferencia en las medias de los cinco tratamientos frente a la alternativa de que al menos una media difiere de otra.

Variable respuesta: Concentración de CO.
Factor: Día de la semana que tiene cinco niveles. Es un factor de efectos fijos ya que viene decidido qué niveles concretos se van a utilizar (5 días de la semana).
Modelo equilibrado: Los niveles de los factores tienen el mismo número de elementos (8 elementos).
Tamaño del experimento: Número total de observaciones, en este caso 40 unidades experimentales. El problema planteado se modeliza a través de un diseño unifactorial totalmente aleatorizado de efectos fijos equilibrado.

Para realizar este supuesto en R debemos introducir primero los datos de forma correcta. Podemos realizarlo directamente en R de forma manual o introducirlos previamente en un archivo de texto o Excel y leerlos en R.

En este caso lo hacemos en un archivo de texto:

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

Se puede realizar de dos formas:

Transformar la variable referente a los niveles del factor fijo como factor

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

```
##  [1] Lunes     Lunes     Lunes     Lunes     Lunes     Lunes     Lunes
##  [8] Lunes     Martes    Martes    Martes    Martes    Martes    Martes
## [15] Martes    Martes    Miercoles Miercoles Miercoles Miercoles Miercoles
## [22] Miercoles Miercoles Miercoles Jueves    Jueves    Jueves    Jueves
## [29] Jueves    Jueves    Jueves    Jueves    Viernes   Viernes   Viernes
## [36] Viernes   Viernes   Viernes   Viernes   Viernes
## Levels: Jueves Lunes Martes Miercoles Viernes
```

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

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

```
## Call:
##    aov(formula = Concentracion ~ Dia, data = contaminacion)
##
## Terms:
##                       Dia Residuals
## Sum of Squares  119416.4  219710.4
## Deg. of Freedom        4        35
##
## Residual standard error: 79.23029
## Estimated effects may be unbalanced
```

se puede mostrar un resumen de los resultados con la funcion "summary"

```
summary(mod)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Dia          4 119416   29854   4.756 0.0036 **
## Residuals   35 219710    6277
## ---
## 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 0.003524, 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)
summary(mod1)

##             Df Sum Sq Mean Sq F value Pr(>F)
## factor(Dia)  4 119416   29854   4.756 0.0036 **
## Residuals   35 219710    6277
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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 119416 29854.1  4.7558 0.003598 **
## Residuals   35 219710  6277.4
## ---
## 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

```
contaminacion1 <- read.table("Supuesto1.txt", header = TRUE)
contaminacion1

##   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

```
tats <- stack(contaminacion1)
tats
```

```
##    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 = tats))

## 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
```

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.

```
Lunes= c(420, 390, 480, 430, 440, 324, 450, 460)
Martes=c(450, 390, 430, 521, 320, 360, 342, 423)
Miercoles<-c(355, 462, 286, 238, 344, 423, 123, 196)
Jueves = c(321, 254, 412, 368, 340, 258, 433, 489)
Viernes<-c(238, 255, 366, 389, 198, 256, 248, 324)
```

Acontinuación creamos un data.frame para poder resolver el ANOVA

```
datos <- data.frame(Lunes, Martes, Miercoles, Jueves, Viernes)
datos
```

```
##   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
```

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     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
```

Recordemos el anova del 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

```
contaminacion2 <- c(Lunes, Martes, Miercoles, Jueves, Viernes)
contaminacion2

##  [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(contaminacion2)

## [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 Lunes, Martes, Miercoles, Jueves y Viernes.

```
trat <- gl(5, 8, 40, labels = c("Lunes", "Martes",
                                "Miercoles", "Jueves",
                                "Viernes"))
trat
```

```
##  [1] Lunes     Lunes     Lunes     Lunes     Lunes     Lunes     Lunes
##  [8] Lunes     Martes    Martes    Martes    Martes    Martes    Martes
## [15] Martes    Martes    Miercoles Miercoles Miercoles Miercoles Miercoles
## [22] Miercoles Miercoles Miercoles Jueves    Jueves    Jueves    Jueves
## [29] Jueves    Jueves    Jueves    Jueves    Viernes   Viernes   Viernes
## [36] Viernes   Viernes   Viernes   Viernes   Viernes
## Levels: Lunes Martes Miercoles Jueves Viernes
```

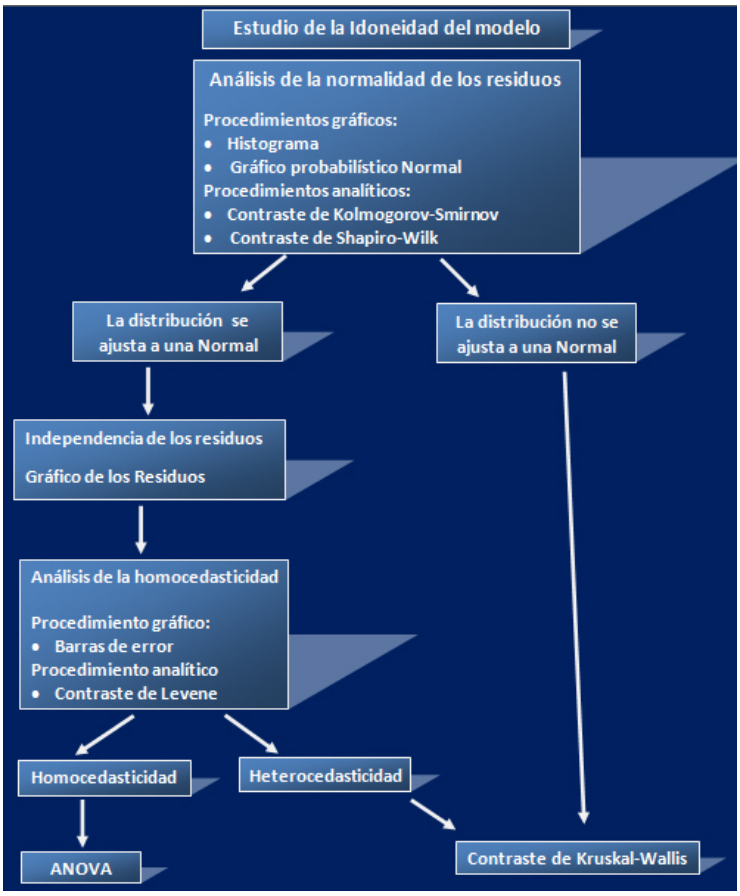
```
anova(lm(contaminacion2 ~ trat))
```

```
## Analysis of Variance Table
##
## Response: contaminacion2
##           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).

1.2 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.



1.3 Hipótesis de normalidad

En primer lugar, analizamos la normalidad de las concentraciones y continuamos con el análisis de la normalidad de los residuos.

Para analizar la normalidad de las concentraciones utilizamos el test de Shapiro-Wilks

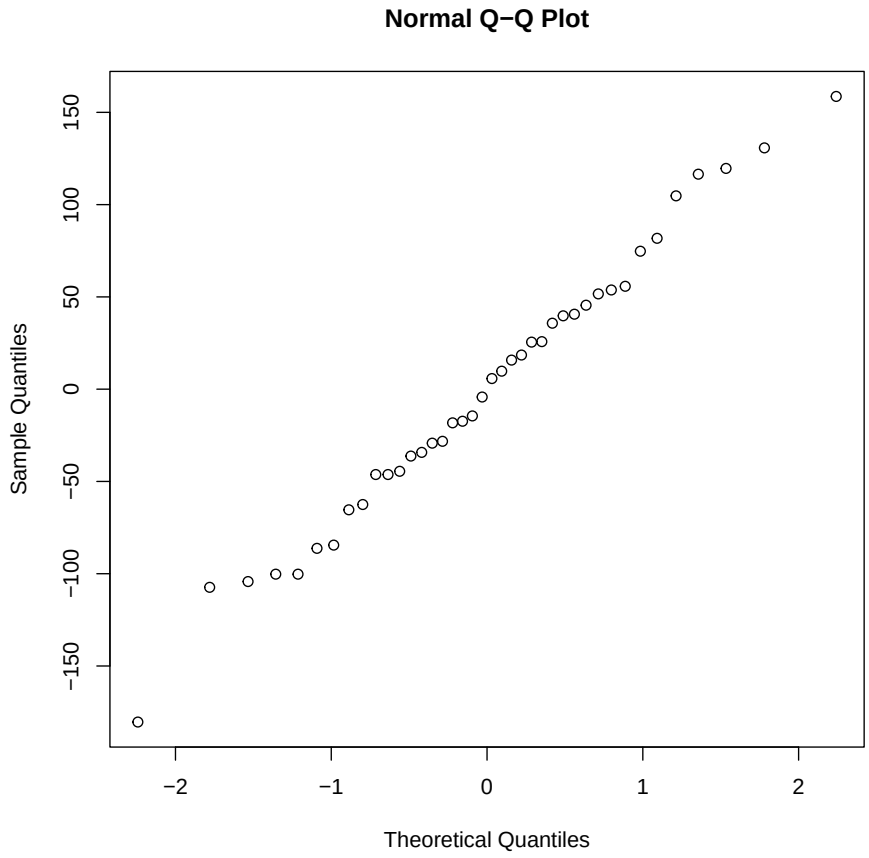
```
shapiro.test(mod$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  mod$residuals
## W = 0.98937, p-value = 0.966
```

Observamos el contraste 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.

1.4 Hipótesis de homocedasticidad

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

```
bartlett.test(contaminacion$Concentracion, contaminacion$Dia)

##
##  Bartlett test of homogeneity of variances
##
## data:  contaminacion$Concentracion and contaminacion$Dia
## Bartlett's K-squared = 5.5055, df = 4, p-value = 0.2392
```

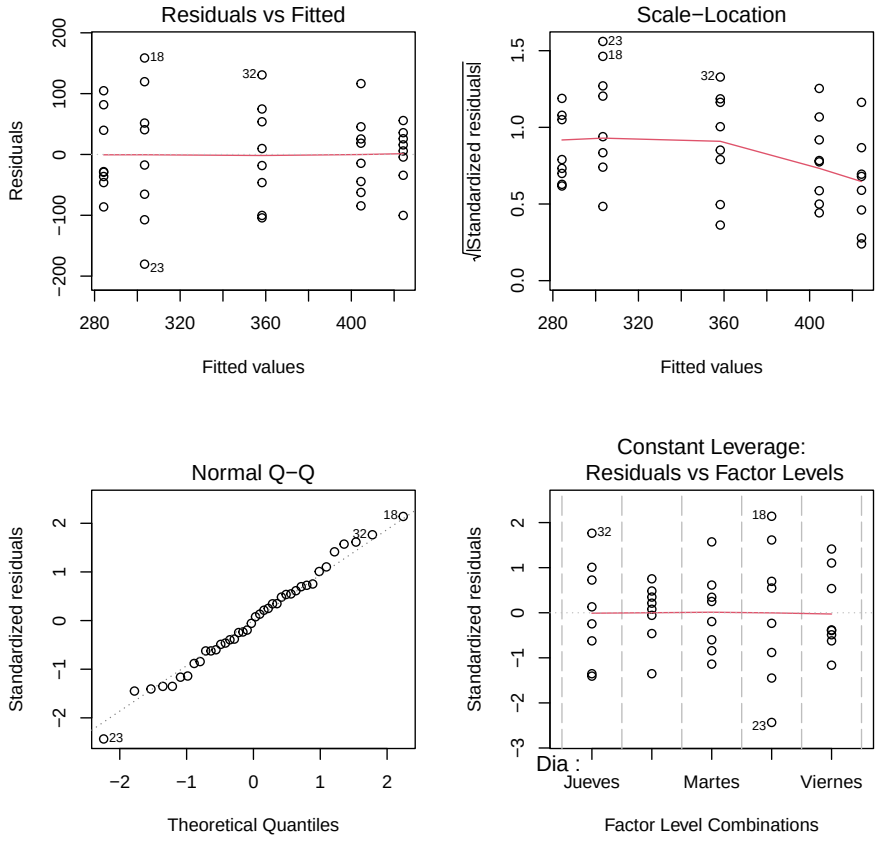
El p-valor es del 0.2402 que al ser mayor del nivel significación usual del 5% no podemos rechazar la hipótesis de igualdad de varianzas, es decir, se acepta la igualdad de varianzas en el factor.

1.5 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:

```
layout(matrix(c(1,2,3,4),2,2)) # para que salgan en la misma pantalla
plot(mod)
```



En la Figura 5 se muestran cuatro gráficos, en el primero de ellos que se representan los residuos en el eje de ordenadas y los valores pronosticados en el eje de abscisas. No observamos, en dicho gráfico, ninguna tendencia sistemática que haga sospechar del incumplimiento de la suposición de independencia.

Anteriormente, hemos comprobado estadísticamente que estos cinco grupos son distintos. Es decir 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.

1.6 Comparaciones múltiples

Para saber entre que parejas de días las diferencias entre concentraciones medias de CO son significativas aplicamos la prueba Post-hoc de Tukey

```
mod.tukey <- TukeyHSD(mod, ordered = TRUE)
mod.tukey

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##     factor levels have been ordered
##
## Fit: aov(formula = Concentracion ~ Dia, data = contaminacion)
##
## $Dia
##                       diff        lwr      upr     p adj
## Miercoles-Viernes   19.125 -94.770937 133.0209 0.9884567
## Jueves-Viernes      74.000 -39.895937 187.8959 0.3528788
## Martes-Viernes     120.250   6.354063 234.1459 0.0342270
## Lunes-Viernes      140.000  26.104063 253.8959 0.0096790
## Jueves-Miercoles    54.875 -59.020937 168.7709 0.6408983
## Martes-Miercoles   101.125 -12.770937 215.0209 0.1019544
## Lunes-Miercoles    120.875   6.979063 234.7709 0.0329483
## Martes-Jueves       46.250 -67.645937 160.1459 0.7694367
## Lunes-Jueves        66.000 -47.895937 179.8959 0.4672516
## Lunes-Martes        19.750 -94.145937 133.6459 0.9869747
```

Esta salida nos muestra los intervalos de confianza simultáneos construidos por el método de Tukey. En la tabla se muestra un resumen de las comparaciones de cada tratamiento con los restantes. Es decir, aparecen comparadas dos a dos las cinco medias de los tratamientos.

En esta tabla, las columnas:

diff: muestra las medias de cada par
p adj: muestra los p-valores de los contrastes, que permiten conocer si la diferencia entre cada pareja de medias es significativa al nivel de significación considerado (en este caso 0.05)
lwr y upr: proporcionan los intervalos de confianza al 95% para cada diferencia.

Así por ejemplo, si comparamos la concentración media de CO del Lunes con el Martes, tenemos una diferencia entre ambas medias de 19.750, un p-valor (Sig.) de 0.9868896 no significativo puesto que la concentración de CO no difiere significativamente el lunes del martes y un intervalo de confianza con un límite inferior negativo y un límite superior positivo y por lo tanto contiene al cero de lo que también deducimos que no hay diferencias significativas entre los dos

grupos que se comparan o que ambos grupos son homogéneos.

En cambio si observamos el grupo formado por el Lunes y el Miércoles, vemos que ambos extremos del intervalo son del mismo signo y el p-valor es significativo deduciendo que si hay diferencias significativas entre ambos. Las otras comparaciones se interpretan de forma análoga.

Por lo tanto la tabla se interpreta observando los valores de p adj menores que el 5%, o si el intervalo de confianza contiene al cero.

Concluimos que se detectan diferencias significativas en las concentraciones de CO entre lunes y miércoles; lunes y viernes; martes y viernes.

Supuesto práctico 2

Los medios de cultivo bacteriológico en los laboratorios de los hospitales proceden de diversos fabricantes. Se sospecha que la calidad de estos medios de cultivo varía de un fabricante a otro. Para comprobar esta teoría, se hace una lista de fabricantes de un medio de cultivo concreto, se seleccionan aleatoriamente los nombres de cinco de los que aparecen en la lista y se comparan las muestras de los instrumentos procedentes de éstos. La comprobación se realiza colocando sobre una placa dos dosis, en gotas, de una suspensión medida de un microorganismo clásico, Escherichia coli, dejando al cultivo crecer durante veinticuatro horas, y determinando después el número de colonias (en millares) del microorganismo que aparecen al final del período. Se quiere comprobar si la calidad del instrumental difiere entre fabricantes.

```
bacterias <- read.table("supuesto2.txt", header = TRUE)
bacterias
```

```
##    Calidad Fabricante
## 1      120          1
## 2      240          2
## 3      240          3
## 4      300          4
## 5      300          5
## 6      240          1
## 7      360          2
## 8      270          3
## 9      240          4
## 10     360          5
## 11     300          1
## 12     180          2
## 13     300          3
## 14     300          4
## 15     240          5
## 16     360          1
```

```
## 17     180          2
## 18     360          3
## 19     360          4
## 20     360          5
## 21     240          1
## 22     300          2
## 23     360          3
## 24     360          4
## 25     360          5
## 26     180          1
## 27     240          2
## 28     300          3
## 29     360          4
## 30     360          5
## 31     144          1
## 32     360          2
## 33     360          3
## 34     360          4
## 35     360          5
## 36     300          1
## 37     360          2
## 38     360          3
## 39     360          4
## 40     300          5
## 41     240          1
## 42     360          2
## 43     300          3
## 44     300          4
## 45     360          5
```

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

```
mod <- aov(Calidad ~ Fabricante, data = bacterias)
```

donde:

Calidad = nombre de la columna de las observaciones.
Fabricante = 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 = Calidad ~ Fabricante, data = bacterias)
```

```
##
## Terms:
##                 Fabricante Residuals
## Sum of Squares     49561.6  152073.6
## Deg. of Freedom          1        43
##
## Residual standard error: 59.46928
## Estimated effects may be unbalanced
```

y posteriormente mostramos 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)
## Fabricante   1  49562   49562   14.01 0.000534 ***
## Residuals   43 152074    3537
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Esta tabla muestra los resultados del contraste planteado. El valor del estadístico de contraste es igual a 3.976 que deja a la derecha un p-valor de 0.00827, así que la respuesta dependerá del nivel de significación que se fije. Si fijamos un nivel de significación de 0.05 se concluye que hay evidencia suficiente para afirmar la existencia de alguna variabilidad entre la calidad del material de los diferentes fabricantes. Si fijamos un nivel de significación de 0.001, no podemos hacer tal afirmación.

En el modelo de efectos aleatorios no se necesitan llevar a cabo más contrastes incluso aunque la hipótesis nula sea rechazada. Es decir, en el caso de rechazar H0 no hay que realizar comparaciones múltiples para comprobar que medias son distintas, ya que el propósito del experimento es hacer un planteamiento general relativo a las poblaciones de las que se extraen las muestras.

Diseño en Bloques Aleatorizados En los diseños estudiados anteriormente hemos supuesto que existe bastante homogeneidad entre las unidades experimentales. Pero puede suceder que dichas unidades experimentales sean heterogéneas y contribuyan a la variabilidad observada en la variable respuesta. Si en esta situación se utiliza un diseño completamente aleatorizado, no sabremos si la diferencia entre dos unidades experimentales sometidas a distintos tratamientos se debe a una diferencia real entre los efectos de los tratamientos o a la heterogeneidad de dichas unidades. Como resultado, el error experimental reflejará esta variabilidad. En esta situación se debe sustraer del error experimental la variabilidad producida por las unidades experimentales y para ello el experimentador puede formar bloques de manera que las unidades experimentales de cada bloque sean lo más homogéneas posible y los bloques entre sí sean

heterogéneos.

En el diseño en bloques Aleatorizados, primero se clasifican las unidades experimentales en grupos homogéneos, llamados bloques, y los tratamientos son entonces asignados aleatoriamente dentro de los bloques. Esta estrategia de diseño mejora efectivamente la precisión en las comparaciones al reducir la variabilidad residual.

Distinguimos dos tipos de diseños en bloques aleatorizados:

Los diseños en bloques completos aleatorizados (Todos los tratamientos se prueban en cada bloque exactamente vez). Los diseños por bloques incompletos aleatorizados (Todos los tratamientos no están representados en cada bloque, y aquellos que sí están en uno en particular se ensayan en él una sola vez).

Diseño en Bloques Completos Aleatorizados

En esta sección presentamos el diseño en Bloques Completos Aleatorizados. La palabra bloque se refiere al hecho de que se ha agrupado a las unidades experimentales en función de alguna variable extraña; aleatorizado se refiere al hecho de que los tratamientos se asignan aleatoriamente dentro de los bloques; completo implica que se utiliza cada tratamiento exactamente una vez dentro de cada bloque y el término efectos fijos se aplica a bloques y tratamientos. Es decir, se supone que ni los bloques ni los tratamientos se eligen aleatoriamente. Además una caracterización de este diseño es que los efectos bloque y tratamiento son aditivos; es decir no hay interacción entre los bloques y los tratamientos.

La descripción del diseño así como la terminología subyacente la vamos a introducir mediante el siguiente supuesto práctico.

Supuesto práctico 3

Abeto blanco, Abeto del Pirineo, es un árbol de gran belleza por la elegancia de sus formas y el exquisito perfume balsámico que destilan sus hojas y cortezas. Destilando hojas y madera se obtiene aceite de trementina muy utilizado en medicina contra torceduras y contusiones. En estos últimos años se ha observado que la producción de semillas ha descendido y con objeto de conseguir buenas producciones se proponen tres tratamientos. Se observa que árboles diferentes tienen distintas características naturales de reproducción, este efecto de las diferencias entre los árboles se debe de controlar y este control se realiza mediante bloques. En el experimento se utilizan 10 abetos, dentro de cada abeto se seleccionan tres ramas semejantes. Cada rama recibe exactamente uno de los tres tratamientos que son asignados aleatoriamente. Constituyendo cada árbol un bloque completo. Los datos obtenidos se presentan en la siguiente

tabla donde se muestra el número de semillas producidas por rama.

- Son diez Abetos en los que se aplican cuatro tratamientos distintos
- No hay ningún otro factor que pueda afectar de forma significativa a los resultados
- Los tratamientos se asignan en orden aleatorio a cada abeto
- El número de semillas observadas se muestra en la Figura 8.

1. El experimentador forma bloques de manera que las unidades experimentales de cada bloque sean lo más homogéneas posible.
2. Los bloques entre sí han de ser heterogéneos
3. Variable o factor bloque: Variable cuyo efecto sobre la variable respuesta no es directamente de interés, pero que se introduce en el experimento para obtener comparaciones homogéneas.
4. Se reduce la variabilidad residual

Distinguimos dos tipos de diseños en bloques aleatorizados:

- Los diseños en bloques completos aleatorizados (Todos los tratamientos se prueban en cada bloque exactamente vez). Los diseños por bloques incompletos aleatorizados (Todos los tratamientos no están representados en cada bloque, y aquellos que sí están en uno en particular se ensayan en él una sola vez).
- En este caso se trata de un diseño en bloques completos aleatorizados. El objetivo del estudio es comparar los tres tratamientos, por lo que se trata de un factor con tres niveles. Sin embargo, al realizar la medición sobre los distintos abetos, es posible que estos influyan sobre el número se semillas observadas. Por ello, y al no ser directamente motivo de estudio, los abetos es un factor secundario que recibe el nombre de bloque.

Nos interesa saber si los distintos tratamientos influyen en la producción de semillas, para ello realizamos el siguiente contraste de hipótesis:

$H_0 : \tau_1 = \tau_2 = \tau_3$
$H_1 : \tau_i \neq \tau_j$ para algun $i \neq j$

Es decir, contrastamos que no hay diferencia en las medias de los tres tratamientos frente a la alternativa de que al menos una media difiere de otra.

Pero, previamente hay que comprobar si la presencia del factor bloque (los abetos) está justificada. Para ello, realizamos el siguiente contraste de hipótesis:

$H_0 : \beta_1 = \beta_2 = ... = \beta_1 0$
$H_1 : \beta_i \neq \beta_j$ para algun $i \neq j$
En este caso lo hacemos en un archivo de texto:

Tenemos en cuenta que para que el ejercicio esté realizado de forma correcta los datos tienen que estar introducidos tal y como vienen en la imagen, es decir, las observaciones en una sola columna y a continuación especificado su tratamiento y su bloque correspondiente.

Para cargar los datos utilizamos la función read.table indicando el nombre del archivo (que debe de estar en el directorio de trabajo) e indicando además que tiene cabecera.

```
semillas<-read.table("supuesto3.txt", header = TRUE)
semillas
```

```
##     y Tratamiento Abeto
## 1   7           1     1
## 2   9           2     1
## 3  10           3     1
## 4   8           1     2
## 5   9           2     2
## 6  10           3     2
## 7   9           1     3
## 8   9           2     3
## 9  12           3     3
## 10 10           1     4
## 11  9           2     4
## 12 12           3     4
## 13 11           1     5
## 14 12           2     5
## 15 14           3     5
## 16  8           1     6
## 17 10           2     6
## 18  9           3     6
## 19  7           1     7
## 20  8           2     7
## 21  7           3     7
## 22  8           1     8
## 23  8           2     8
## 24  7           3     8
## 25  7           1     9
## 26  9           2     9
## 27 10           3     9
## 28  8           1    10
## 29  9           2    10
## 30 10           3    10
```

A continuación debemos transformar tanto la columna de los tratamientos como la de los bloques en un factor para podemos realizar los cálculos posteriores

adecuadamente.

```
semillas$Tratamiento = factor(semillas$Tratamiento)
semillas$Tratamiento
```

```
##  [1] 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3
## Levels: 1 2 3
```

```
semillas$Abeto = factor(semillas$Abeto)
semillas$Abeto
```

```
##  [1] 1  1  1  2  2  2  3  3  3  4  4  4  5  5  5  6  6  6  7  7  7  8  8  8  9
## [26] 9  9  10 10 10
## Levels: 1 2 3 4 5 6 7 8 9 10
```

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

```
mod = aov(y ~ Tratamiento + Abeto, data = semillas)
```

donde:

y es el nombre de la columna de las observaciones.
Tratamiento es el nombre de la columna en la que están representados los tratamientos.
Abeto es el nombre de la columna en la que están representados los bloques.
data = data.frame en el que están guardados los datos

```
mod
```

```
## Call:
##    aov(formula = y ~ Tratamiento + Abeto, data = semillas)
##
## Terms:
##                 Tratamiento Abeto Residuals
## Sum of Squares         16.2  54.8      15.8
## Deg. of Freedom           2     9        18
##
## Residual standard error: 0.936898
## Estimated effects may be unbalanced
```

y a continuación mostramos un resumen de los resultados con la función "summary" (verdadera tabla ANOVA):

```r
summary(mod) # TABLA ANOVA
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Tratamiento  2   16.2   8.100   9.228 0.00174 **
## Abeto        9   54.8   6.089   6.937 0.00026 ***
## Residuals   18   15.8   0.878
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Puesto que la construcción de bloques se ha diseñado para comprobar el efecto de una variable, nos preguntamos si ha sido eficaz su construcción. En caso afirmativo, la suma de cuadrados de bloques explicaría una parte sustancial de la suma total de cuadrados. También se reduce la suma de cuadrados del error dando lugar a un aumento del valor del estadístico de contraste experimental utilizado para contrastar la igualdad de medias de los tratamientos y posibilitando que se rechace la Hipótesis nula, mejorándose la potencia del contraste.

La construcción de bloques puede ayudar cuando se comprueba su eficacia pero debe evitarse su construcción indiscriminada. Ya que, la inclusión de bloques en un diseño da lugar a una disminución del número de grados de libertad para el error, aumenta el punto crítico para contrastar la Hipótesis nula y es más difícil rechazarla. La potencia del contraste es menor.

La Tabla *ANOVA*, muestra que:

- El valor del estadístico de contraste de igualdad de bloques, F = 6.937 deja a su derecha un p-valor menor que 0.001, menor que el nivel de significación del 5%, por lo que se rechaza la Hipótesis nula de igualdad de bloques. La eficacia de este diseño depende de los efectos de los bloques. Un valor grande de F de los bloques (6.937) implica que el factor bloque tiene un efecto grande. En este caso el diseño es más eficaz que el diseño completamente aleatorizado ya que si el cuadrado medio entre bloques es grande (6.089), el término residual será mucho menor (0.878) y el contraste principal de las medias de los tratamientos será más sensible a las diferencias entre tratamientos. Por lo tanto la inclusión del factor bloque en el modelo es acertada. Así, la producción de semillas depende del abeto.

Si los efectos de los bloques son muy pequeños, el análisis de bloque quizás no sea necesario y en caso extremo, cuando el valor de F de los bloques es próximo a 1, puede llegar a ser perjudicial, ya que el número de grados de libertad, (I-1)(J-1), del denominador de la comparación de tratamientos es menor que el número de grados de libertad correspondiente, IJ-I, en el diseño completamente aleatorizado. Pero, ¿Cómo saber cuándo se puede prescindir de los bloques? La respuesta la tenemos en el valor de la F experimental de los bloques, se ha comprobado que si dicho valor es mayor que 3, no conviene prescindir de los

bloques para efectuar los contrastes.

- El valor del estadístico de contraste de igualdad de tratamiento, F = 9.228 deja a su derecha un p-valor de 0.002, menor que el nivel de significación del 5%, por lo que se rechaza la Hipótesis nula de igualdad de tratamientos. Así, los tratamientos influyen en el número de semillas. Es decir, existen diferencias significativas en el número de semillas entre los tres tratamientos.

El modelo que hemos propuesto hay que validarlo, para ello hay que comprobar si se verifican los cuatros supuestos expresados anteriormente.

1.6.1 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.

En este modelo se ha supuesto otra hipotesis adicional: Aditividad de los efectos de tratamiento y bloque (no existe interaccion entre tratamiento y bloque). Por lo que hay que contrastar la hipotesis de aditividad de los efectos de tratamiento y bloque.

1.6.2 Hipótesis de aditividad entre los bloques y tratamientos

La interacción entre el factor bloque y los tratamientos vamos a estudiarla analíticamente mediante el Test de Interacción de un grado de Tukey

Para realizar este test en R tenemos que utilizar la library "daewr" y dentro de ella la función "Tukey1df". De la siguiente forma:

- Primero hay que instalar el paquete **daewr**

Para ello, seleccionar **Paquetes/Instalar paquetes** y de la lista escoger **daewr**. O bien utilizar la siguiente orden

```
utils:::menuInstallPkgs()

## Error in install.packages(lib = .libPaths()[1L], dependencies =
NA, type = type):  no packages were specified
```

Para realizar este contraste hay que utilizar la libray daewr, para ello realizamos la siguiente orden

```r
library(daewr)

## Registered S3 method overwritten by 'DoE.base':
##   method           from
##   factorize.factor conf.design

Tukey1df(semillas)

## Source           df     SS        MS        F     Pr>F
## A                 2   16.2    8.1
## B                 9   54.8    6.0889
## Error            18   15.8    71.1
## NonAdditivity     1   3.5573    3.5573    4.94    0.0401
## Residual         17   12.2427    0.7202
```

Puesto que el p-valor (Pr ¿ F) es 1 no rechazamos la hipótesis nula de no interacción, es decir, no hay interacción entre los tratamientos aplicados y los abetos.

1.6.3 Hipótesis de Normalidad

La normalidad las vamos a comprobar analíticamente y gráficamente.

Analíticamente mediante el contraste de Shapiro-Wilk que es adecuado cuando las muestras son pequeñas (n¡50)

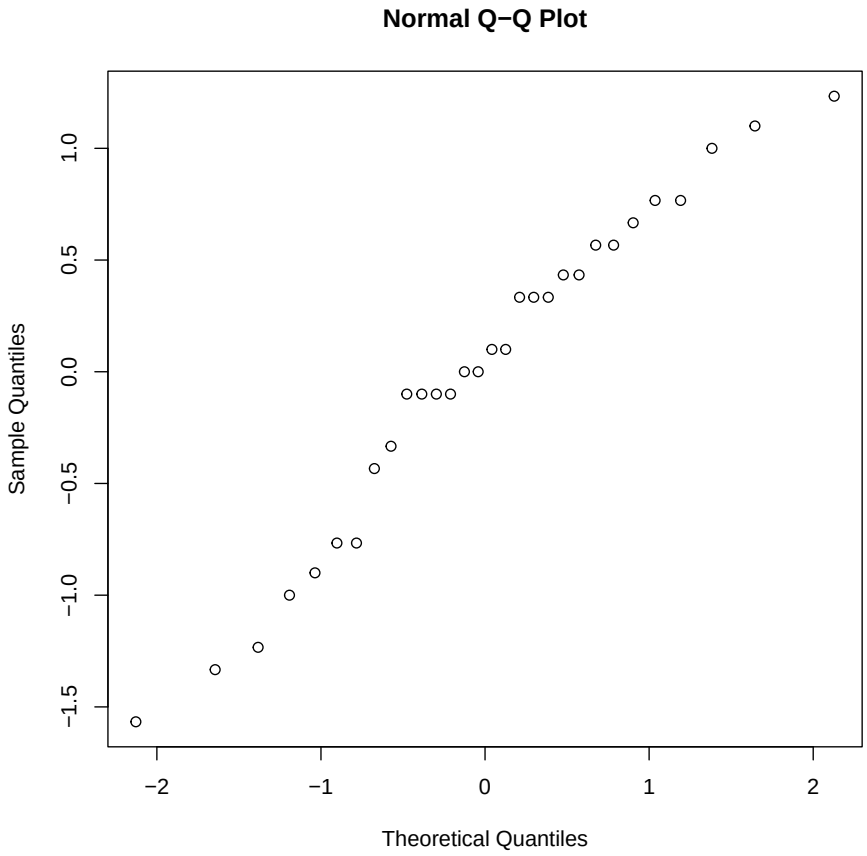
```r
shapiro.test(mod$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  mod$residuals
## W = 0.96415, p-value = 0.3935
```

Como podemos observar tenemos un p-valor de 0.3935 que aceptaría la hipótesis de normalidad por ser mayor al 5% (nivel de significación usual).

Gráficamente mediante el gráfico probabilístico normal. Para ello utilizamos la orden "qqnorm"

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



En esta gráfica observamos que prácticamente todos los puntos se encuentran sobre la diagonal por lo tanto podemos decir que no muestra una desviación marcada de la normalidad.

Hipótesis de Homogeneidad de Varianzas Para comprobar la hipótesis de homocedasticidad utilizamos el Test de Barlett distinguiendo entre la igualdad entre varianzas del factor principal y la igualdad de varianzas del factor bloque.

En nuestro ejemplo, el test para igualdad de varianzas del factor principal sería:

```
bartlett.test(semillas$y, semillas$Tratamiento)

##
##  Bartlett test of homogeneity of variances
##
## data:  semillas$y and semillas$Tratamiento
```

```
## Bartlett's K-squared = 4.1729, df = 2, p-value = 0.1241
```

El p-valor es del 0.1241 que al ser mayor del nivel significación usual del 5% no podemos rechazar la hipótesis de igualdad de varianzas en el factor principal.

De la misma manera procedemos para el factor bloque:

```
bartlett.test(semillas$y, semillas$Abeto)

##
##  Bartlett test of homogeneity of variances
##
## data:  semillas$y and semillas$Abeto
## Bartlett's K-squared = 4.0723, df = 9, p-value = 0.9066
```

El p-valor es mayor que 0.05 por lo que no podemos rechazar la hipótesis de igualdad de varianzas en el factor bloque.

1.7 Hipótesis de Independencia

Comprobaremos si se satisface el supuesto de independencia entre los residuos. Para ello tenemos que representar un gráfico de los residuos tipificados frente a los pronosticados. 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:

```
#layout(matrix(c(1,2,3,4),2,2))
#plot(mod)
```

Nos fijamos en el primer gráfico que representa los residuos frente a los valores ajustados y observamos que no hay ninguna tendencia sistemática. Concluimos que no hay sospechas para que se incumpla la hipótesis de independencia.

1.8 Comparaciones múltiples

Hemos probado anteriormente que se rechaza la Hipótesis nula de igualdad de tratamientos. Así, los tratamientos influyen en el número de semillas. Es decir, existen diferencias significativas en el número de semillas entre los tres tratamientos. Para saber entre que parejas de días estas diferencias son significativas aplicamos una prueba **Post-hoc**.

El contraste de Comparaciones múltiples que vamos a utilizar es el Test de

Duncan. Para poder hacer uso de él en R tenemos que instalar en primer lugar el paquete "agricolae" y dentro de él la función "duncan.test".

Destacar que este test hace las comparaciones especificándole si es para el factor principal o el factor bloque.

Comenzamos con el factor principal:

```
(duncan=duncan.test(mod, "Tratamiento" , group = T))

## Error in duncan.test(mod, "Tratamiento", group = T): no se pudo
encontrar la función "duncan.test"
```

En el apartado "$groups" concluimos que los tres tratamientos difieren significativamente entre sí.

Se observa que la concentración media del número de semillas es mayor con el Tratamiento3 (10.1) y menor con el Tratamiento1 (8.3).

Para el factor bloque:

```
(duncan=duncan.test(mod, "Abeto" , group = T))

## Error in duncan.test(mod, "Abeto", group = T): no se pudo encontrar
la función "duncan.test"
```

Se observa que la prueba de Duncan ha agrupado los abetos 7, 8, 1, 9, 2, 6 y 10 en un mismo grupo, 1, 9 ,2 6, 10, 3 y 4, en otro grupo y un tercer está formada únicamente por el Abeto5. Inmediatamente se ve que por ejemplo el Abeto5 difiere de todos los demás, siendo en este abeto donde se produce el mayor número de semillas (12.333) y el menor en el Abeto (7.333).

2 Diseño en bloques Incompletos Aleatorizados

En los diseños en bloques Aleatorizados, puede suceder que no sea posible realizar todos los tratamientos en cada bloque. En estos casos es posible usar diseños en bloques Aleatorizados en los que cada tratamiento no está presente en cada bloque. Estos diseños reciben el nombre de diseño en bloque incompleto aleatorizado siendo uno de los más utilizados el diseño en bloque incompleto balanceado (BIB)

El diseño de bloques incompletos balanceado (BIB) compara todos los tratamientos con igual precisión.

Este diseño experimental debe verificar:

- Cada tratamiento ocurre el mismo número de veces en el diseño.
- Cada par de tratamientos ocurren juntos el mismo número de veces que cualquier otro par.

Supongamos que se tienen I tratamientos de los cuales sólo pueden experimentar K tratamientos en cada bloque (K ¡ I). Los parámetros que caracterizan este modelo son:

- I, J y K son el número de tratamientos, el número de bloques y el número de tratamientos por bloque, respectivamente.
- R, número de veces que cada tratamiento se presenta en el diseño, es decir el número de réplicas de un tratamiento dado.
- λ, número de bloques en los que un par de tratamientos ocurren juntos.
- N, número de observaciones.

Estos parámetros deben verificar las siguientes relaciones:

$\lambda = R\frac{K-1}{I-1}$

donde $J \geq I$ y $N = IR = JK$

Si $J = I$ el diseño recibe el nombre de simétrico. Al igual que en el diseño en bloques completo, la asignación de los tratamientos a las unidades experimentales en cada bloque se debe realizar en forma aleatoria.

Este diseño lo estudiaremos a continuación mediante el supuesto práctico 4

Supuesto práctico 4 Se realiza un estudio para comprobar la efectividad en el retraso del crecimiento de bacterias utilizando cuatro soluciones diferentes para lavar los envases de la leche. El análisis se realiza en el laboratorio y sólo se pueden realizar seis pruebas en un mismo día. Como los días son una fuente de variabilidad potencial, el investigador decide utilizar un diseño aleatorizado por bloques, pero al recopilar las observaciones durante seis días no ha sido posible aplicar todos los tratamientos en cada día, sino que sólo se han podido aplicar dos de las cuatro soluciones cada día. Se decide utilizar un diseño en bloques incompletos balanceado, donde $I = 4$ y $K = 2$.

Un posible diseño para estos parámetros lo proporciona la tabla correspondiente al Diseño 5 del Fichero-Adjunto, con $R = 3$, $J = 6$ y $\lambda = 1$. La disposición del diseño y las observaciones obtenidas se muestran en la siguiente tabla.

En el ejemplo:

- $N = IR = JK$. En efecto, ya que $N = 12; I = 4, J = 6; R = 3 y K = 2$.

$\lambda = 31/3 = 1$

El objetivo principal es estudiar la efectividad en el retraso del crecimiento de bacterias utilizando cuatro soluciones, por lo que se trata de un factor con cuatro niveles. Sin embrago, como los días son una fuente de variabilidad potencial, consideramos un factor bloque con seis niveles.

- **Variable respuesta: Número de bacterias**
- **Factor**: Soluciones que tiene cuatro niveles. Es un factor de efectos fijos ya que viene decidido qué niveles concretos se van a utilizar.
- **Bloque: Día**s que tiene seis niveles. Es un factor de **efectos fijos** ya que viene decidido qué niveles concretos se van a utilizar.
- **Modelo incompleto**: Todos los tratamientos no se prueban en cada bloque.
- **Tamaño del experimento**: Número total de observaciones (12).

Podemos introducir los datos directamente en R de forma manual o introducirlos previamente en un archivo de texto o Excel y leerlos en R.

```
bacterias = read.table("supuesto4.txt", header = TRUE)
bacterias
```

```
##     y Soluciones Dias
## 1  12          1    1
## 2  24          1    2
## 3  31          1    3
## 4  21          2    1
## 5  20          2    5
## 6  21          2    6
## 7  19          3    3
## 8  18          3    4
## 9  19          3    6
## 10 15          4    2
## 11 19          4    4
## 12 47          4    5
```

A continuación debemos transformar tanto la columna de los tratamientos como la de los bloques en un factor para poder realizar los cálculos posteriores adecuadamente.

```
bacterias$Soluciones = factor(bacterias$Soluciones)
bacterias$Dias = factor(bacterias$Dias)
```

Para poder analizar los datos mediante un diseño BIB debemos instalar y cargar dos paquetes de R especializados en este tipo de diseños:

```
library(daewr)
library(AlgDesign)
```

La función "BIBsize(t , k)" de la librería daewr nos permite saber si el diseño puede realizarse. Calcula los parámetros del diseño donde

- t = número de niveles del factor tratamiento.
- k = número de tratamientos por bloque.

Ejecutamos:

```
BIBsize(t = 4 , k = 2)

## Posible BIB design with b= 6  and r= 3  lambda= 1
```

El análisis de este modelo lo podemos realizar en R de dos formas:

1. Realizaremos el análisis evaluando primero el efecto de los tratamientos y después el de los bloques utilizando dos funciones
   - Para evaluar el efecto de los tratamientos, la suma de cuadrados de tratamientos debe ajustarse por bloques, por lo tanto primero se introducen los bloques y después los tratamientos.
   - Para calcular la tabla ANOVA hacemos uso de la función "aov" ($aov(y\ A + B, data = mydataframe)$ asume suma de cuadrados tipo I) de la siguiente forma:

   ```
   mod1 <- aov(y ~ Dias + Soluciones, data = bacterias)
   ```

   donde:

   - y = nombre de la columna de las observaciones
   - Soluciones = nombre de la columna en la que están representados los tratamientos
   - Dias = nombre de la columna en la que están representados los bloques
   - data = data.frame en el que están guardados los datos

```
mod1

## Call:
##    aov(formula = y ~ Dias + Soluciones, data = bacterias)
##
## Terms:
##                      Dias Soluciones Residuals
## Sum of Squares  387.6667   123.2500  396.7500
## Deg. of Freedom        5          3         3
##
## Residual standard error: 11.5
## Estimated effects may be unbalanced
```

y posteriormente mostramos un resumen de los resultados con la función "summary" (verdadera tabla ANOVA)

```
summary(mod1)

##             Df Sum Sq Mean Sq F value Pr(>F)
## Dias         5  387.7   77.53   0.586  0.720
## Soluciones   3  123.3   41.08   0.311  0.819
## Residuals    3  396.7  132.25
```

El valor del estadístico de contraste de igualdad de Soluciones, F = 0.311, deja a su derecha un p-valor 0.819, mayor que el nivel de significación del 5%, por lo que no se rechaza la Hipótesis Nula de igualdad de tratamientos. Por lo tanto el tipo de solución para lavar los envases de la leche no influye en el retraso del crecimiento de bacterias.

- Para evaluar el efecto de los bloques, la suma de cuadrados de bloques debe ajustarse por los tratamientos, por lo tanto primero se introducen los tratamientos y después los bloques:

```
mod2 <- aov(y ~ Soluciones + Dias, data = bacterias)
mod2

## Call:
##    aov(formula = y ~ Soluciones + Dias, data = bacterias)
##
## Terms:
##                 Soluciones     Dias Residuals
## Sum of Squares    113.6667 397.2500  396.7500
## Deg. of Freedom          3        5         3
```

```
##
## Residual standard error: 11.5
## Estimated effects may be unbalanced
```

```
summary(mod2)

##             Df Sum Sq Mean Sq F value Pr(>F)
## Soluciones   3  113.7   37.89   0.286  0.834
## Dias         5  397.2   79.45   0.601  0.712
## Residuals    3  396.7  132.25
```

El valor del estadístico de contraste de igualdad de Días, F = 0.601, deja a su derecha un p-valor 0.712, mayor que el nivel de significación del 5%, por lo que no se rechaza la Hipótesis nula de igualdad de bloques. Por lo tanto los días en los que se realiza la prueba para lavar los envases de la leche no influyen en el retraso del crecimiento de bacterias.

Con este ejemplo se ilustra el hecho de decidir si se prescinde o no de los bloques. Hay situaciones en las que, aunque los bloques no resulten significativamente diferentes no es conveniente prescindir de ellos. Pero ¿cómo saber cuándo se puede prescindir de los bloques? La respuesta la tenemos en el valor de la F de los bloques, experimentalmente se ha comprobado que si dicho valor es mayor que 3, no conviene prescindir de los bloque para efectuar los contrastes.

En esta situación si se puede prescindir del efecto de los bloques y estudiar el modelo unifactorial correspondiente, cuyo único factor es: Soluciones.

2. Realizaremos el análisis evaluando tanto para los tratamientos como para los bloques ejecutando solo una función.

   Para ello necesitamos instalar y cargar el paquete "car":

   IMPORTANTE: Hemos comprobado que utilizando la versión del paquete "car" 3.0-0, encontramos un error y no permite su utilización, por lo que descargamos una versión anterior, concretamente car_2.1-6.tar.gz. Pinchamos en este enlace y guardamos el archivo en el escritorio

   Recordad que para la utilización de un paquete es necesario instalarlo y cargarlo. Para ello:

(a) Accedemos a Paquetes/Install package(s) from local file y elegimos el paquete descargado.

(b) Posteriormente lo cargamos con Paquetes/Cargar paquetes

Nota: Para instalar un paquete directamente de R, procederemos de la forma siguiente

- Accedemos a la página https://cran.r-project.org/index.html.
- Seleccionamos **Packages**
- Seleccionamos Table of available packages, sorted by name
- Seleccionamos paquete car
- Seleccionamos Old sources car archive
- Seleccionamos el paquete car_2.1-6.tar.gz y lo guardamos en el escritorio.

Una vez instalado cargado el paquete realizamos el ANOVA

```
mod3 <- lm(y ~ Soluciones + Dias, data = bacterias)
mod3

##
## Call:
## lm(formula = y ~ Soluciones + Dias, data = bacterias)
##
## Coefficients:
## (Intercept)  Soluciones2  Soluciones3  Soluciones4        Dias2        Dias3
##      20.000       -7.000       -6.750        1.750       -1.375        8.375
##       Dias4        Dias5        Dias6
##       1.000       16.125        6.875
```

```
car::Anova(mod3, type="III")

## Anova Table (Type III tests)
##
## Response: y
##             Sum Sq Df F value Pr(>F)
## (Intercept) 533.33  1  4.0328 0.1382
## Soluciones  123.25  3  0.3106 0.8187
## Dias        397.25  5  0.6008 0.7118
## Residuals   396.75  3
```

Los resultados obtenidos coinciden con los realizados primero a los tratamientos y después a los bloques

3 Diseño de cuadrado latino

Hemos estudiado en el apartado anterior que los diseños en bloques completos aleatorizados utilizan un factor de control o variable de bloque con objeto de eliminar su influencia en la variable respuesta y así reducir el error experimental. Los diseños en cuadrados latinos utilizan dos variables de bloque para reducir el error experimental.

Un inconveniente que presentan a veces los diseños es el de requerir excesivas unidades experimentales para su realización. Un diseño en bloques completos con un factor principal y dos factores de bloque, con $K_1, K_2 y K_3$ niveles en cada uno de los factores, requiere $K_1 K_2 K_3$ unidades experimentales. En un experimento puede haber diferentes causas, por ejemplo de índole económico, que no permitan emplear demasiadas unidades experimentales, ante esta situación se puede recurrir a un tipo especial de diseños en bloques incompletos aleatorizados. La idea básica de estos diseños es la de fracción es decir, seleccionar una parte del diseño completo de forma que, bajo ciertas hipótesis generales, permita estimar los efectos que interesan.

Uno de los diseños en bloques incompletos aleatorizados más importante con dos factores de control es el modelo en cuadrado latino, dicho modelo requiere el mismo número de niveles para los tres factores.

En general, para K niveles en cada uno de los factores, el diseño completo en bloques aleatorizados utiliza K^2 bloques, aplicándose en cada bloque los K niveles del factor principal, resultando un total de K^3 unidades experimentales.

Los diseños en cuadrado latino reducen el número de unidades experimentales a K^3 utilizando los K^2 bloques del experimento, pero aplicando sólo un tratamiento en cada bloque con una disposición especial. De esta forma, si K fuese 4, el diseño en bloques completos necesitaría $4^3 = 64$ observaciones, mientras que el diseño en cuadrado latino sólo necesitaría $4^2 = 16$ observaciones.

Los diseños en cuadrados latinos son apropiados cuando es necesario controlar dos fuentes de variabilidad. En dichos diseños el número de niveles del factor principal tiene que coincidir con el número de niveles de las dos variables de bloque o factores secundarios y además hay que suponer que no existe interacción entre ninguna pareja de factores.

Recibe el nombre de cuadrado latino de orden K a una disposición en filas y columnas de K letras latinas, de tal forma que cada letra aparece una sola vez en cada fila y en cada columna.

En resumen, podemos decir que un diseño en cuadrado latino tiene las siguientes características:

- Se controlan tres fuentes de variabilidad, un factor principal y dos factores de bloque.
- Cada uno de los factores tiene el mismo número de niveles, K.
- Cada nivel del factor principal aparece una vez en cada fila y una vez en cada columna.
- No hay interacción entre los factores.

En el Fichero-Adjunto se muestran algunos cuadrados latinos estándares para los órdenes 3, 4, 5, 6, 7, 8 y 9.

Este diseño lo estudiaremos a continuación mediante el supuesto práctico 5

Supuesto práctico 5

Se estudia el rendimiento de un proceso químico en seis tiempos de reposo, A, B, C, D, E y F. Para ello, se consideran seis lotes de materia prima que reaccionan con seis concentraciones de ácido distintas, de manera que cada lote de materia prima en cada concentración de ácido se somete a un tiempo de reposo. Tanto la asignación de los tiempos de reposo a los lotes de materia prima, como la concentración de ácido, se hizo de forma aleatoria. Los datos del rendimiento del proceso químico se muestran en la siguiente tabla.

	Concentraciones de ácido					
Lote	1	2	3	4	5	6
Lote 1	12 A	24 B	10 C	18 D	21 E	18 F
Lote 2	21 B	26 C	24 D	16 E	20 F	21 A
Lote 3	20 C	16 D	19 E	18 F	16 A	19 B
Lote 4	22 D	15 E	14 F	19 A	27 B	17 C
Lote 5	15 E	13 F	17 A	25 B	21 C	22 D
Lote 6	17 F	11 A	12 B	22 C	14 D	20 E

El objetivo principal es estudiar la influencia de seis tiempos de reposo en el rendimiento de un proceso químico, por lo que se trata de un factor con seis niveles. Sin embargo, como los lotes de materia prima y las concentraciones son dos fuentes de variabilidad potencial, consideramos dos factores de bloque con seis niveles cada uno.

- Variable respuesta: Rendimiento.
- Factor: Tiempo de reposo que tiene seis niveles. Es un factor de efectos fijos ya que viene decidido que niveles concretos se van a utilizar.
- Bloques: Lotes y Concentraciones, ambos con seis niveles y ambos son factores de efectos fijos.
- Tamaño del experimento: Número total de observaciones (36).

Para realizar este supuesto en R debemos introducir primero los datos de forma correcta. Podemos introducir los datos directamente en R de forma manual o introducirlos previamente en un archivo de texto o Excel y leerlos en R.

Tenemos en cuenta que para que el ejercicio esté realizado de forma correcta los datos tienen que estar introducidos tal y como vienen en la imagen, es decir, las observaciones en una sola columna y a continuación especificado su tratamiento, su bloque y después la letra latina correspondiente.

Para cargar los datos utilizamos la función read.table indicando el nombre del archivo (que debe de estar en el directorio de trabajo) e indicando además que tiene cabecera.

```
latino <- read.table("supuesto5.txt", header = TRUE, dec= ",")
latino
```

```
##    Observaciones  Lote Concentraciones Tiempo_de_reposo
## 1             12 Lote1               1                A
## 2             24 Lote1               2                B
## 3             10 Lote1               3                C
## 4             18 Lote1               4                D
## 5             21 Lote1               5                E
## 6             18 Lote1               6                F
## 7             21 Lote2               1                B
## 8             26 Lote2               2                C
## 9             24 Lote2               3                D
## 10            16 Lote2               4                E
## 11            20 Lote2               5                F
## 12            21 Lote2               6                A
## 13            20 Lote3               1                C
## 14            16 Lote3               2                D
## 15            19 Lote3               3                E
## 16            18 Lote3               4                F
## 17            16 Lote3               5                A
## 18            19 Lote3               6                B
## 19            22 Lote4               1                D
## 20            15 Lote4               2                E
```

```
## 21            14 Lote4                3                F
## 22            19 Lote4                4                A
## 23            27 Lote4                5                B
## 24            17 Lote4                6                C
## 25            15 Lote5                1                E
## 26            13 Lote5                2                F
## 27            17 Lote5                3                A
## 28            25 Lote5                4                B
## 29            21 Lote5                5                C
## 30            22 Lote5                6                D
## 31            17 Lote6                1                F
## 32            11 Lote6                2                A
## 33            12 Lote6                3                B
## 34            22 Lote6                4                C
## 35            14 Lote6                5                D
## 36            20 Lote6                6                E
```

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

```
mod1 <- aov(Observaciones~ Lote + Concentraciones + Tiempo_de_reposo, data = latino)
```

donde:

- Observaciones: Nombre de la columna de las observaciones
- Lote: Nombre de la columna en la que están representados los tratamientos
- Concentraciones : Nombre de la columna en la que está representado el primer factor bloque.
- Tiempo_de_reposo: Nombre de la columna en la que está representado el segundo factor bloque (letras latinas).
- data = data.frame en el que están guardados los datos

```
mod1

## Call:
##    aov(formula = Observaciones ~ Lote + Concentraciones + Tiempo_de_reposo,
##     data = latino)
##
## Terms:
```

```
##                      Lote Concentraciones Tiempo_de_reposo Residuals
## Sum of Squares   99.5556         30.9429         117.8889  386.1683
## Deg. of Freedom        5               1                5        24
##
## Residual standard error: 4.011277
## Estimated effects may be unbalanced
```

y posteriormente mostramos un resumen de los resultados con la función "summary" (verdadera tabla ANOVA):

```
summary(mod1)

##                   Df Sum Sq Mean Sq F value Pr(>F)
## Lote               5   99.6   19.91   1.237  0.323
## Concentraciones    1   30.9   30.94   1.923  0.178
## Tiempo_de_reposo   5  117.9   23.58   1.465  0.238
## Residuals         24  386.2   16.09
```

Observando los valores de los p-valores, 0.281, 0.368 y 0.553; mayores respectivamente que el nivel de significación del 5%, deducimos que ningún efecto es significativo.

4 Diseño de Cuadrado Greco-Latino

El modelo en cuadrado greco-latino se puede considerar como una extensión del modelo en cuadrado latino en el que se incluye una tercera variable control o variable de bloque. En este modelo como en el diseño en cuadrado latino, todos los factores deben tener el mismo número de niveles, K, y el número de observaciones necesarias sigue siendo K^2. Este diseño es, por tanto, una fracción del diseño completo en bloques aleatorizados con un factor principal y tres factores secundarios que requeriría K_4 observaciones.

Los cuadrados greco-latinos se obtienen por superposición de dos cuadrados latinos del mismo orden y ortogonales entre sí, uno de los cuadrados con letras latinas el otro con letras griegas. Dos cuadrados reciben el nombre de ortogonales si, al superponerlos, cada letra latina y griega aparecen juntas una sola vez en el cuadrado resultante.

En el Fichero-Adjunto se muestra una tabla de cuadrados latinos que dan lugar, por superposición de dos de ellos, a cuadrados greco-latinos. Notamos que no es posible formar cuadrados greco-latinos de orden 6.

La Tabla siguiente ilustra un cuadrado greco-latino para $K = 4$

Supuesto práctico 6

Para comprobar el rendimiento de un proceso químico en cinco tiempos de reposo, se consideran cinco lotes de materia prima que reaccionan con cinco concentraciones de ácido distintas a cinco temperaturas distintas, de manera que cada lote de materia prima con cada concentración de ácido y cada temperatura se someten a un tiempo de reposo. Tanto la asignación de los tiempos de reposo a los lotes de materia prima, como las concentraciones de ácido, y las temperaturas, se hizo de forma aleatoria. En este estudio el científico considera que tanto los lotes de materia prima, las concentraciones y las temperaturas pueden influir en el rendimiento del proceso, por lo que los considera como variables de bloque cada una con cinco niveles y decide plantear un diseño por cuadrados greco-latinos como el que muestra en la siguiente tabla.

La variable respuesta que vamos a estudiar es el rendimiento del proceso químico. El factor principal es tiempo de reposo que se presenta con cinco niveles.

- Variable respuesta: Rendimiento
- Factor: Tiempos de reposo que tiene cinco niveles. Es un factor de efectos fijos ya que viene decidido que niveles concretos se van a utilizar.
- Bloques: Lotes, Concentraciones y Temperaturas, cada uno con cinco niveles y de efectos fijos.
- Tamaño del experimento: Número total de observaciones (25).

Para realizar este supuesto en R debemos introducir primero los datos de forma correcta. Podemos introducir los datos directamente en R de forma manual o introducirlos previamente en un archivo de texto o Excel y leerlos en R.

En este caso lo hacemos en un archivo de texto.

Tenemos en cuenta que para que el ejercicio esté realizado de forma correcta los datos tienen que estar introducidos tal y como vienen en la imagen, es decir, las observaciones en una sola columna y a continuación especificado su tratamiento, su bloque correspondiente y después la letra latina y griega correspondiente (En este caso hemos cambiado las letras griegas como las últimas del alfabeto latino por facilidad a la hora de escribirlas).

Para cargar los datos utilizamos la función read.table indicando el nombre del archivo (que debe de estar en el directorio de trabajo) e indicando además que tiene cabecera.

```
greco <- read.table("supuesto6-2.txt", header = TRUE, dec= ",")
greco

##    Observaciones Lotes Concentraciones Tiempo_de_reposo Temperaturas
```

```
## 1             26 Lote1              1               A          Z
## 2             21 Lote1              2               B          Y
## 3             19 Lote1              3               C          X
## 4             13 Lote1              5               D          W
## 5             21 Lote1              5               E          V
## 6             22 Lote2              1               B          X
## 7             26 Lote2              2               C          W
## 8             24 Lote2              3               D          V
## 9             16 Lote2              4               E          Z
## 10            20 Lote2              5               A          Y
## 11            29 Lote3              1               C          V
## 12            26 Lote3              2               D          Z
## 13            19 Lote3              3               E          Y
## 14            18 Lote3              4               A          X
## 15            16 Lote3              5               B          W
## 16            32 Lote4              1               D          Y
## 17            15 Lote4              2               E          X
## 18            14 Lote4              3               A          W
## 19            19 Lote4              4               B          V
## 20            27 Lote4              5               C          Z
## 21            25 Lote5              1               E          W
## 22            18 Lote5              2               A          V
## 23            19 Lote5              3               B          Z
## 24            25 Lote5              4               C          Y
## 25            21 Lote5              5               D          X
```

A continuación debemos transformar tanto la columna de los tratamiento como la de los bloques en un factor para podemos realizar los cálculos posteriores adecuadamente.

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

```
greco <- read.table("supuesto6-2.txt", header = TRUE, dec = ",")
greco$Lote <- factor(greco$Lote)
greco$Temperaturas <- factor(greco$Temperaturas)
greco$Tiempo_de_reposo <- factor(greco$Tiempo_de_reposo)
greco$Concentraciones <- factor(greco$Concentraciones)
mod1 <- aov(Observaciones~ Lote + Concentraciones + Tiempo_de_reposo + Temperaturas, data =
mod1

## Call:
##    aov(formula = Observaciones ~ Lote + Concentraciones + Tiempo_de_reposo +
##     Temperaturas, data = greco)
##
## Terms:
```

```
##                      Lote Concentraciones Tiempo_de_reposo Temperaturas
## Sum of Squares    9.7600        207.7607         155.0085      97.2516
## Deg. of Freedom        4               4                4            4
##                 Residuals
## Sum of Squares   100.7792
## Deg. of Freedom         8
##
## Residual standard error: 3.549281
## Estimated effects may be unbalanced
```

donde:

- Observaciones: Nombre de la columna de las observaciones.
- Lote: Nombre de la columna en la que están representados los tratamientos.
- Concentraciones = Nombre de la columna en la que está representado el primer factor bloque.
- Tiempo_de_reposo = Nombre de la columna en la que está representado el segundo factor bloque (letras latinas).
- Temperaturas: Nombre de la columna en la que está representado el tercer factor bloque.
- Data: data.frame en el que están guardados los datos

y posteriormente mostramos un resumen de los resultados con la función "summary" (verdadera tabla ANOVA):

```
summary(mod1)
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## Lote              4   9.76    2.44   0.194 0.9349
## Concentraciones   4 207.76   51.94   4.123 0.0420 *
## Tiempo_de_reposo  4 155.01   38.75   3.076 0.0825 .
## Temperaturas      4  97.25   24.31   1.930 0.1988
## Residuals         8 100.78   12.60
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observando los valores de los p-valores, 0.150, 0.053, 0.912 y 0.020, deducimos que el único efecto significativo, al nivel de significación del 5%, es el efecto de las distintas concentraciones sobre el rendimiento del proceso químico.

5 Diseño de cuadrados de Youden

Hemos estudiado que en el diseño en cuadrado latino se tiene que verificar que los tres factores tengan el mismo número de niveles, es decir que hay el mismo número de filas, de columnas y de letras latinas. Sin embargo, puede suceder que el número de niveles disponibles de uno de los factores de control sea menor que el número de tratamientos, en este caso estaríamos ante un diseño en cuadrado latino incompleto. Estos diseños fueron estudiados por W.J. Youden y se conocen con el nombre de cuadrados de Youden.

Este diseño lo estudiaremos a continuación mediante el supuesto práctico 7.

Supuesto práctico 7

Consideremos de nuevo el experimento sobre el rendimiento de un proceso químico en el que se está interesado en estudiar seis tiempos de reposo, A, B, C, D, E y F y se desea eliminar estadísticamente el efecto de los lotes materia prima y de las concentraciones de ácido distintas. Pero supongamos que sólo se dispone de cinco tipos de concentraciones. Para analizar este experimento se decidió utilizar un cuadrado de Youden con seis filas (los lotes de materia prima), cinco columnas (las distintas concentraciones) y seis letras latinas (los tiempos de reposo). Los datos correspondientes se muestran en la siguiente tabla.

	concentracion de acido				
Lote	1	2	3	4	5
Lote 1	12 A	24 B	10 C	18 D	21 E
Lote 2	21 B	26 C	24 D	16 E	20 F
Lote 3	20 C	16 D	19 E	18 F	16 A
Lote 4	22 D	15 E	14 F	19 A	27 B
Lote 5	15 E	13 F	17 A	25 B	21 C
Lote 6	17 F	11 A	12 B	22 C	14 D

Tenemos en cuenta que para que el ejercicio esté realizado de forma correcta los datos tienen que estar introducidos tal y como vienen en la imagen, es decir, las observaciones en una sola columna y a continuación especificado su tratamiento, su bloque correspondiente y después la letra latina correspondiente.

Para cargar los datos utilizamos la función read.table indicando el nombre del archivo (que debe de estar en el directorio de trabajo) e indicando además que tiene cabecera.

```
youden <- read.table("supuesto7-1.txt", header = TRUE)
youden
```

```
##    Observaciones  Lote Concentraciones Tiempo_de_reposo
## 1             12 Lote1               1                A
## 2             24 Lote1               2                B
## 3             10 Lote1               3                C
## 4             18 Lote1               4                D
## 5             21 Lote1               5                E
## 6             21 Lote2               1                B
## 7             26 Lote2               2                C
## 8             24 Lote2               3                D
## 9             16 Lote2               4                E
## 10            20 Lote2               5                F
## 11            20 Lote3               1                C
## 12            16 Lote3               2                D
## 13            19 Lote3               3                E
## 14            18 Lote3               4                F
## 15            16 Lote3               5                A
## 16            22 Lote4               1                D
## 17            15 Lote4               2                E
## 18            14 Lote4               3                F
## 19            19 Lote4               4                A
## 20            27 Lote4               5                B
## 21            15 Lote5               1                E
## 22            13 Lote5               2                F
## 23            17 Lote5               3                A
## 24            25 Lote5               4                B
## 25            21 Lote5               5                C
## 26            17 Lote6               1                F
## 27            11 Lote6               2                A
## 28            12 Lote6               3                B
## 29            22 Lote6               4                C
## 30            14 Lote6               5                D
```

A continuación debemos transformar tanto la columna de los tratamiento como la de los bloques en un factor para podemos realizar los cálculos posteriores

adecuadamente.

```
youden$Lote <- factor(youden$Lote)
youden$Concentraciones <- factor(youden$Concentraciones)
youden$Tiempo_de_reposo <- factor(youden$Tiempo_de_reposo)
```

Para cada factor realizamos una tabla ANOVA:

- Factor principal:
  Para evaluar el efecto de los tratamientos, la suma de cuadrados de tratamientos debe ajustarse por bloques, por lo tanto primero se introducen los bloques y después los tratamientos.

  Para calcular la tabla ANOVA hacemos uso de la función "aov" (asume suma de cuadrados tipo I) de la siguiente forma:

```
mod1 <- aov(Observaciones~ Tiempo_de_reposo + Lote + Concentraciones, data = youden)
mod1

## Call:
##    aov(formula = Observaciones ~ Tiempo_de_reposo + Lote + Concentraciones,
##     data = youden)
##
## Terms:
##                 Tiempo_de_reposo      Lote Concentraciones Residuals
## Sum of Squares         151.76667 112.73333        61.66667 282.00000
## Deg. of Freedom                5         5               4        15
##
## Residual standard error: 4.335897
## Estimated effects may be unbalanced
```

```
summary(mod1)

##                  Df Sum Sq Mean Sq F value Pr(>F)
## Tiempo_de_reposo  5 151.77   30.35   1.615  0.216
## Lote              5 112.73   22.55   1.199  0.356
## Concentraciones   4  61.67   15.42   0.820  0.532
## Residuals        15 282.00   18.80
```

donde:

- Observaciones: Nombre de la columna de las observaciones.
- Lote: Nombre de la columna en la que están representados los tratamientos.

- Concentraciones: Nombre de la columna en la que está representado el primer factor bloque.
- Tiempo_de_reposo: Nombre de la columna en la que está representado el segundo factor bloque (letras latinas).
- data = data.frame en el que están guardados los datos.
  El p-valor, 0.532, es mayor que el nivel de significación del 5%, deducimos que el factor principal: Concentraciones no es significativo.

- Factor Bloque: Lotes.

  Para evaluar el efecto del primero de los bloques, la suma de cuadrados de bloques debe ajustarse por los tratamientos, por lo tanto primero se introducen los tratamientos y después los bloques:

```
mod3 <- aov(Observaciones~ Concentraciones  + Lote +Tiempo_de_reposo , data = youden)
mod3

## Call:
##    aov(formula = Observaciones ~ Concentraciones + Lote + Tiempo_de_reposo,
##     data = youden)
##
## Terms:
##                 Concentraciones      Lote Tiempo_de_reposo Residuals
## Sum of Squares         61.66667 111.36667        153.13333 282.00000
## Deg. of Freedom               4         5                5        15
##
## Residual standard error: 4.335897
## Estimated effects may be unbalanced
```

```
 summary(mod3)

##                   Df Sum Sq Mean Sq F value Pr(>F)
## Concentraciones    4  61.67   15.42   0.820  0.532
## Lote               5 111.37   22.27   1.185  0.362
## Tiempo_de_reposo   5 153.13   30.63   1.629  0.213
## Residuals         15 282.00   18.80
```

El p-valor es 0.213; mayor que el nivel de significación del 5%, deducimos que Factor Bloque:Tiempo_de_reposo no es significativo.

6 Diseños Factoriales

En muchos experimentos es frecuente considerar dos o más factores y estudiar el efecto conjunto que dichos factores producen sobre la variable respuesta. Para resolver esta situación se utiliza el Diseño Factorial.

Se entiende por diseño factorial aquel diseño en el que se investigan todas las posibles combinaciones de los niveles de los factores en cada réplica del experimento. En estos diseños, los factores que intervienen tienen la misma importancia a priori y se supone por tanto, la posible presencia de interacción. En este epígrafe vamos a considerar únicamente modelos de efectos fijos.

6.1 Diseños factoriales con dos factores

6.1.1 Modelo sin recuperacion

Supuesto práctico 8

En unos laboratorios se está investigando sobre el tiempo de supervivencia de unos animales a los que se les suministra al azar tres tipos de venenos y cuatro antídotos distintos. Se pretende estudiar si los tiempos de supervivencia de los anímales varían en función de las combinaciones veneno-antídoto. Los datos que se recogen en la tabla adjunta son los tiempos de supervivencia en horas.

El objetivo principal es estudiar la influencia de tres tipos de venenos y 4 tipos de antídotos en el tiempo de supervivencia de unos determinados animales, por lo que se trata de un modelo con dos factores: el veneno (con tres niveles) y el antídoto (con cuatro niveles). La variable que va a medir las diferencias entre los tratamientos es el tiempo que sobreviven los animales. Se combinan todos los niveles de los dos factores por lo que tenemos en total doce tratamientos.

- Variable respuesta: Tiempo de supervivencia
- Factor: Tipo de veneno que tiene tres niveles. Es un factor de efectos fijos ya que viene decidido qué niveles concretos se van a utilizar.
- Factor: Tipo de antídoto que tiene cuatro niveles. Es un factor de efectos fijos ya que viene decidido qué niveles concretos se van a utilizar.
- Tamaño del experimento: Número total de observaciones (12).

Para realizar este supuesto en R debemos introducir primero los datos de forma correcta. Podemos introducir los datos directamente en R de forma manual o introducirlos previamente en un archivo de texto o Excel y leerlos en R.

En este caso lo hacemos en un archivo de texto:

Tenemos en cuenta que para que el ejercicio esté realizado de forma correcta los datos tienen que estar introducidos tal y como vienen en la imagen, es decir, las observaciones en una sola columna y a continuación especificado sus factores correspondientes.

Para cargar los datos utilizamos la función read.table indicando el nombre del archivo (que debe de estar en el directorio de trabajo) e indicando además que tiene cabecera.

```
getwd()

## [1] "C:/Users/Usuario/Desktop/respaldo/Desktop/PAQUETE R/PRACTICAS_S9"

setwd("C:/Users/Usuario/Desktop/respaldo/Desktop/PAQUETE R/PRACTICAS_S9")
factorial <- read.table("supuesto8.txt", header = TRUE)
factorial

##    Tiempo Veneno Antidoto
## 1     4.5      1        1
## 2     2.9      2        1
## 3     2.1      3        1
## 4    11.0      1        2
## 5     6.1      2        2
## 6     3.7      3        2
## 7     4.5      1        3
## 8     3.5      2        3
## 9     2.5      3        3
## 10    7.1      1        4
## 11   10.2      2        4
## 12    3.6      3        4
```

A continuación debemos transformar todas las columnas que contienen a los factores en un factor para podemos realizar los cálculos posteriores adecuadamente.

```
factorial$Antidoto <- factor(factorial$Antidoto)
factorial$Veneno <- factor(factorial$Veneno)
```

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

```
mod <- aov(Tiempo~ Veneno + Antidoto , data = factorial)
mod
```

```
## Call:
##    aov(formula = Tiempo ~ Veneno + Antidoto, data = factorial)
##
## Terms:
##                    Veneno Antidoto Residuals
## Sum of Squares  30.58667 39.40917  23.89333
## Deg. of Freedom        2        3         6
##
## Residual standard error: 1.995551
## Estimated effects may be unbalanced
```

y posteriormente mostramos un resumen de los resultados con la función "summary" (verdadera tabla ANOVA):

```
summary(mod)

##             Df Sum Sq Mean Sq F value Pr(>F)
## Veneno       2  30.59  15.293   3.840 0.0844 .
## Antidoto     3  39.41  13.136   3.299 0.0995 .
## Residuals    6  23.89   3.982
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Esta Tabla ANOVA recoge la descomposición de la varianza considerando como fuente de variación los doce tratamientos o grupos que se forman al combinar los niveles de los dos factores. Mediante esta tabla se puede estudiar sí varían los tiempos que sobreviven los animales en función de las combinaciones veneno-antídoto. Es decir, se pueden estudiar si existen diferencias significativas entre los tiempos medios de supervivencia con los distintos tipos de venenos y antídotos, pero no se puede estudiar si la efectividad de los antídotos es la misma para todos los venenos. Observando los p-valores, 0.084 y 0.099; mayores respectivamente que el nivel de significación del 5%, deducimos que ningún efecto es significativo. Por lo tanto, no existen diferencias en los tiempos medios de supervivencia de los animales, en función de la pareja veneno-antídoto que se les suministra.

6.2 El modelo con replicación

El modelo estadístico para este diseño es:

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + u_{ijk}, i = 1, 2, ...a, j = 1, 2...b, k = 1, ...r$$

donde r es el número de replicaciones y N = abr es el número de observaciones.

El número de parámetros de este modelo es, como en el modelo de dos factores sin replicación, $ab+1$ pero en este caso el número de observaciones es abr.

La descripción del diseño así como la terminología subyacente la vamos a introducir mediante el siguiente supuesto práctico.

Supuesto práctico 9

Consideremos el supuesto práctico anterior en el que realizamos dos réplicas por cada tratamiento. Los datos que se recogen en la tabla adjunta son los tiempos de supervivencia en horas de unos animales a los que se les suministra al azar tres venenos y cuatro antídotos. El objetivo es estudiar qué antídoto es el adecuado para cada veneno.

	Antídoto			
Veneno	Antídoto 1	Antídoto 2	Antídoto 3	Antídoto 4
Veneno 1	4.5	11.0	4.5	7.1
	4.3	7.2	7.6	6.2
Veneno 2	2.9	6.1	3.5	10.2
	2.3	12.4	4.0	3.8
Veneno 3	2.1	3.7	2.5	3.6
	2.3	2.9	2.2	3.3

Para realizar este supuesto en R debemos introducir primero los datos de forma correcta. Podemos introducir los datos directamente en R de forma manual o introducirlos previamente en un archivo de texto o Excel y leerlos en R.

En este caso lo hacemos en un archivo de texto:

enemos en cuenta que para que el ejercicio esté realizado de forma correcta los datos tienen que estar introducidos tal y como vienen en la imagen, es decir, las observaciones en una sola columna y a continuación especificado sus factores correspondientes.

Para cargar los datos utilizamos la función read.table indicando el nombre del archivo (que debe de estar en el directorio de trabajo) e indicando además que tiene cabecera.

```
factorial <- read.table("supuesto8.txt", header = TRUE)
factorial
```

```
##    Tiempo Veneno Antidoto
## 1     4.5      1        1
## 2     2.9      2        1
## 3     2.1      3        1
## 4    11.0      1        2
## 5     6.1      2        2
## 6     3.7      3        2
## 7     4.5      1        3
## 8     3.5      2        3
## 9     2.5      3        3
## 10    7.1      1        4
## 11   10.2      2        4
## 12    3.6      3        4
```

A continuación debemos transformar todas las columnas que contienen a los factores en un factor para podemos realizar los cálculos posteriores adecuadamente.

```
factorial$Veneno <- factor(factorial$Veneno)
factorial$Antidoto <- factor(factorial$Antidoto)
```

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

```
mod <- aov(Tiempo~ Veneno * Antidoto , data = factorial)
mod
```

```
## Call:
##    aov(formula = Tiempo ~ Veneno * Antidoto, data = factorial)
##
## Terms:
##                   Veneno Antidoto Veneno:Antidoto
## Sum of Squares  30.58667 39.40917        23.89333
## Deg. of Freedom        2        3               6
##
## Estimated effects may be unbalanced
```

```
summary(mod)
```

```
##                 Df Sum Sq Mean Sq
## Veneno           2  30.59  15.293
## Antidoto         3  39.41  13.136
## Veneno:Antidoto  6  23.89   3.982
```

La Tabla ANOVA muestra las filas de Tipo_veneno, Tipo_antídoto y Tipo_veneno × Tipo_antídoto que corresponde a la variabilidad debida a los efectos de cada uno de los factores y de la interacción entre ambos.

Las preguntas que nos planteamos son: ¿Los venenos son igual de peligrosos? ¿Y los antídotos son igual de efectivos? La efectividad de los antídotos, ¿es la misma para todos los venenos? Para responder a estas preguntas, comenzamos comprobando si el efecto de los antídotos es el mismo para todos los venenos. Para ello observamos el valor del estadístico (Fexp= 0.761) que contrasta la hipótesis correspondiente a la interacción entre ambos factores ($H0 : (\tau\beta)_{ij} = 0$). Dicho valor deja a la derecha un Sig. = 0.614, mayor que el nivel de significación 0.05. Por lo tanto la interacción entre ambos factores no es significativa y debemos eliminarla del modelo. Construimos de nuevo la Tabla ANOVA en la que sólo figurarán los efectos principales

```
mod <- aov(Tiempo~ Veneno + Antidoto , data = factorial)
mod
```

```
## Call:
##    aov(formula = Tiempo ~ Veneno + Antidoto, data = factorial)
##
## Terms:
##                    Veneno Antidoto Residuals
## Sum of Squares  30.58667 39.40917  23.89333
## Deg. of Freedom        2        3         6
##
## Residual standard error: 1.995551
## Estimated effects may be unbalanced
```

```
summary(mod)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Veneno       2  30.59  15.293   3.840 0.0844 .
## Antidoto     3  39.41  13.136   3.299 0.0995 .
## Residuals    6  23.89   3.982
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Esta tabla muestra dos únicas fuentes de variación, lo efectos principales de los dos factores (Tipo_veneno y Tipo_antídoto), y se ha suprimido la interacción entre ambos. Se observa que el valor de la Suma de Cuadrados del error de este modelo (73.873) se ha formado con los valores de las Sumas de cuadrados del error y de la interacción del modelo anterior (20.363 + 53.510 = 73.873). Observando los valores de los p-valores, 0.0046 y 0.0117 asociados a los contrastes principales, se deduce que los dos efectos son significativos a un nivel de

significación del 5%. Deducimos que ni la gravedad de los venenos es la misma, ni la efectividad de los antídotos, pero dicha efectividad no depende del tipo de veneno con el que se administre ya que la interacción no es significativa.

6.3 Diseños factoriales con tres factores

Supongamos que hay a niveles para el factor A, b niveles del factor B y c niveles para el factor C y que cada réplica del experimento contiene todas las posibles combinaciones de tratamientos, es decir contiene los abc tratamientos posibles.

6.3.1 El modelo sin replicación

Supuesto práctico 10

En una fábrica de refrescos está haciendo unos estudios en la planta embotelladora. El objetivo es obtener más uniformidad en el llenado de las botellas. La máquina de llenado teóricamente llena cada botella a la altura correcta, pero en la práctica hay variación, y la embotelladora desea entender mejor las fuentes de esta variabilidad para eventualmente reducirla. En el proceso se pueden controlar tres factores durante el proceso de llenado: El % de carbonato (factor A), la presión del llenado (factor B) y el número de botellas llenadas por minuto que llamaremos velocidad de la línea (factor C). Se consideran tres niveles para el factor A (10%, 12%, 14%), dos niveles para el factor B (25psi, 30psi) y dos niveles para el factor C (200bpm, 250bpm). Los datos recogidos de la desviación de la altura objetivo se muestran en la tabla adjunta

	Presión (B)			
	25 psi		30 psi	
	Velocidad (C)		Velocidad (C)	
% de Carbono (A)	200	250	200	250
10	10	3	5	-1
12	11	2	5	-3
14	2	4	-3	1

La variable respuesta de este experimento es la Desviación que se produce en la altura de llenado en las botellas de refresco, siendo dichas botellas las unidades experimentales. En estas desviaciones de la altura de llenado marcada como objetivo intervienen tres factores: Porcentaje de carbono que presenta tres niveles 10%, 12% y 14%; Presión, con dos niveles 25 psi y 30 psi y Velocidad, con dos niveles 200 y 250. Los niveles de los factores han sido fijados por el experimentador, por lo que todos los factores son de efectos fijos. Se trata de un diseño trifactorial de efectos fijos, donde el número de tratamientos es $322 = 12$.

Para realizar este supuesto en R debemos introducir primero los datos de forma correcta. Podemos introducir los datos directamente en R de forma manual o introducirlos previamente en un archivo de texto o Excel y leerlos en R.

Tenemos en cuenta que para que el ejercicio esté realizado de forma correcta los datos tienen que estar introducidos tal y como vienen en la imagen, es decir, las observaciones en una sola columna y a continuación especificado sus factores correspondientes.

Para cargar los datos utilizamos la función read.table indicando el nombre del archivo (que debe de estar en el directorio de trabajo) e indicando además que tiene cabecera.

```
factorial <- read.table("supuesto10.txt", header = TRUE)
factorial

##    Altura Carbono Presion Velocidad
## 1      10      10      25       200
## 2      11      12      25       200
## 3       2      14      25       200
## 4       3      10      25       250
## 5       2      12      25       250
## 6       4      14      25       250
## 7       5      10      30       200
## 8       5      12      30       200
## 9      -3      14      30       200
## 10     -1      10      30       250
## 11     -3      12      30       250
## 12      1      14      30       250
```

A continuación debemos transformar la tres columnas en factores para poder realizar los cálculos posteriores adecuadamente.

```
factorial$Carbono <- factor(factorial$Carbono)
factorial$Velocidad <- factorial(factorial$Velocidad)
factorial$Presion <- factor(factorial$Presion)
```

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

```
mod=aov(Altura ~ Carbono + Presion + Velocidad + Carbono*Presion + Carbono*Velocidad + Presi

## Error in lm.fit(x, y, offset = offset, singular.ok = singular.ok,
...):  NA/NaN/Inf in 'x'
```

donde:

- Altura: Nombre de la columna de las observaciones.
- Carbono: Nombre de la columna en la que está representado el primer factor.
- Presion: Nombre de la columna en la que está representado el segundo factor.
- Velocidad: Nombre de la columna en la que está representado el tercer factor
- Carbono*Presion, Carbono*Velocidad y Presion*Velocidad hace referencia a las distintas interacciones.
- data= data.frame en el que están guardados los datos

```
mod

## Call:
##    aov(formula = Tiempo ~ Veneno + Antidoto, data = factorial)
##
## Terms:
##                   Veneno Antidoto Residuals
## Sum of Squares  30.58667 39.40917  23.89333
## Deg. of Freedom        2        3         6
##
## Residual standard error: 1.995551
## Estimated effects may be unbalanced
```

```
summary(mod)

##             Df Sum Sq Mean Sq F value Pr(>F)
## Veneno       2  30.59  15.293   3.840 0.0844 .
## Antidoto     3  39.41  13.136   3.299 0.0995 .
## Residuals    6  23.89   3.982
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La Tabla ANOVA muestra las filas de Carbono, Presión, Velocidad, Carbono*Presión, Carbono*Velocidad y Presión*Velocidad que corresponden a la variabilidad debida a los efectos de cada uno de los factores y a las interacciones de orden dos entre ambos. En dicha Tabla se indica que para un nivel de significación del 5% los efectos que no son significativos del modelo planteado son las interacciones entre los factores Carbono*Presión y Presión*Velocidad ya que los p-valores correspondientes a estos efectos son 0.125 y 0.057 mayores que el nivel de significación.

Como consecuencia de este resultado, replanteamos el modelo suprimiendo en primer lugar el efecto Carbono*Presión, cuya significación es mayor.
donde los efectos deben cumplir las condiciones expuestas anteriormente. Para resolverlo suprimimos la interacción Carbono*Presión. La tabla ANOVA que corresponde a este modelo es la siguiente:

```
mod <- aov(Altura~ Carbono + Presion + Velocidad + Carbono*Velocidad + Presion*Velocidad , d

## Error in lm.fit(x, y, offset = offset, singular.ok = singular.ok,
...):  NA/NaN/Inf in 'x'

summary(mod)

##             Df Sum Sq Mean Sq F value Pr(>F)
## Veneno       2  30.59  15.293   3.840 0.0844 .
## Antidoto     3  39.41  13.136   3.299 0.0995 .
## Residuals    6  23.89   3.982
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El efecto Presión*Velocidad sigue siendo no significativo por lo que lo suprimimos del modelo y replanteamos el siguiente modelo matemático
donde los efectos deben cumplir las condiciones expuestas anteriormente. Para resolverlo suprimimos la interacción Presión*Velocidad. La tabla ANOVA que corresponde a este modelo es la siguiente:

```
mod <- aov(Altura~ Carbono + Presion + Velocidad + Carbono*Velocidad, data = factorial )

## Error in lm.fit(x, y, offset = offset, singular.ok = singular.ok,
...):  NA/NaN/Inf in 'x'

summary(mod)

##             Df Sum Sq Mean Sq F value Pr(>F)
## Veneno       2  30.59  15.293   3.840 0.0844 .
## Antidoto     3  39.41  13.136   3.299 0.0995 .
## Residuals    6  23.89   3.982
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Todos los efectos de este último modelo planteado son significativos y por lo tanto es en este modelo donde vamos a realizar el estudio. Existen diferencias significativas entre los distintos porcentajes del Carbono, los dos tipos de presión, las dos velocidades de llenado y la interacción entre el porcentaje de Carbono y la Velocidad de llenado.

Supuesto pracico 11.

Consideremos el supuesto práctico anterior en el que realizamos dos réplicas por cada tratamiento. En la Tabla adjunta se muestran los datos recogidos de la desviación de la altura objetivo de las botellas de refresco. En el proceso de llenado, la embotelladora puede controlar tres factores durante el proceso: El porcentaje de carbonato (factor A) con tres niveles (10%, 12%, 14%), la presión del llenado (factor B) con dos niveles (25psi, 30psi) y el número de botellas llenadas por minuto que llamaremos velocidad de la línea (factor C) con dos niveles (200bpm, 250bpm).

	Presión (B)			
	25 psi		30 psi	
	Velocidad (C)		Velocidad (C)	
% de Carbono (A)	200	250	200	250
10	10 20	3 5	5 9	-1 -3
12	11 9	2 5	5 4	-3 2
14	2 -1	4 7	-3 -2	1 3

Para realizar este supuesto en R debemos introducir primero los datos de forma correcta. Podemos introducir los datos directamente en R de forma manual o introducirlos previamente en un archivo de texto o Excel y leerlos en R.

En este caso lo hacemos en un archivo de texto. Tenemos en cuenta que para que el ejercicio esté realizado de forma correcta los datos tienen que estar introducidos tal y como vienen en la imagen, es decir, las observaciones en una sola columna y a continuación especificado sus factores correspondientes.

Para cargar los datos utilizamos la función read.table indicando el nombre del archivo (que debe de estar en el directorio de trabajo) e indicando además que tiene cabecera.

```
factorial <- read.table("supuesto11.txt", header = TRUE)
factorial

##    Altura Carbono Presion Velocidad
## 1      10      10      25       200
## 2      20      10      25       200
## 3      11      12      25       200
## 4       9      12      25       200
## 5       2      14      25       200
```

```
## 6      -1      14      25       200
## 7       3      10      25       250
## 8       5      10      25       250
## 9       2      12      25       250
## 10      5      12      25       250
## 11      4      14      25       250
## 12      7      14      25       250
## 13      5      10      30       200
## 14      9      10      30       200
## 15      5      12      30       200
## 16      4      12      30       200
## 17     -3      14      30       200
## 18     -2      14      30       200
## 19     -1      10      30       250
## 20     -3      10      30       250
## 21     -3      12      30       250
## 22      2      12      30       250
## 23      1      14      30       250
## 24      3      14      30       250
```

A continuación debemos transformar las tres columnas en factores para poder realizar los cálculos posteriores adecuadamente.

```
factorial$Carbono <- factor(factorial$Carbono)
factorial$Velocidad <- factor(factorial$Velocidad)
factorial$Presion <- factor(factorial$Presion)
```

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

```
mod <- aov(Altura~ Carbono + Presion +
             Velocidad + Carbono*Presion +
             Carbono*Velocidad + Presion*Velocidad +
             Carbono*Velocidad*Presion, data = factorial)
```

donde:

- Altura: Nombre de la columna de las observaciones.
- Carbono: Nombre de la columna en la que está representado el primer factor.
- Presion: Nombre de la columna en la que está representado el segundo factor.
- Velocidad: Nombre de la columna en la que está representado el tercer factor

- Carbono*Presion, Carbono*Velocidad, Presion*Velocidad y Carbono*Velocidad*Presion hace referencia a las distintas interacciones.
- data= data.frame en el que están guardados los datos

y posteriormente mostramos un resumen de los resultados con la función "summary" (verdadera tabla ANOVA):

```
summary(mod)
```

```
##                            Df Sum Sq Mean Sq F value   Pr(>F)
## Carbono                     2  88.08   44.04   5.683 0.018350 *
## Presion                     1 150.00  150.00  19.355 0.000866 ***
## Velocidad                   1  80.67   80.67  10.409 0.007270 **
## Carbono:Presion             2  14.25    7.12   0.919 0.425122
## Carbono:Velocidad           2 230.58  115.29  14.876 0.000564 ***
## Presion:Velocidad           1   1.50    1.50   0.194 0.667799
## Carbono:Presion:Velocidad   2   1.75    0.88   0.113 0.894175
## Residuals                  12  93.00    7.75
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La Tabla ANOVA muestra las filas de Carbono, Presión, Velocidad, Carbono*Presión, Carbono*Velocidad, Presión*Velocidad y Carbono*Presión*Velocidad que corresponden a la variabilidad debida a los efectos de cada uno de los factores, a las interacciones de orden dos y orden tres entre los factores. En dicha Tabla se indica que para un nivel de significación del 5% los efectos que no son significativos del modelo planteado son las interacciones entre los factores, Carbono*Presión y Presión*Velocidad y Carbono*Presión*Velocidad ya que los p-valores correspondientes a estos efectos son 0.425, 0.668 y 0.894 mayores que el nivel de significación.

Para resolverlo suprimimos la interacción Carbono*Presión*Velocidad. La tabla ANOVA que corresponde a este modelo es la siguiente:

```
mod <- aov(Altura~ Carbono + Presion +
             Velocidad + Carbono*Presion +
             Carbono*Velocidad + Presion*Velocidad,
           data = factorial)
summary(mod)
```

```
##                   Df Sum Sq Mean Sq F value   Pr(>F)
## Carbono            2  88.08   44.04   6.507 0.010038 *
## Presion            1 150.00  150.00  22.164 0.000336 ***
## Velocidad          1  80.67   80.67  11.919 0.003886 **
## Carbono:Presion    2  14.25    7.12   1.053 0.375033
## Carbono:Velocidad  2 230.58  115.29  17.035 0.000178 ***
```

```
## Presion:Velocidad  1   1.50    1.50   0.222 0.645047
## Residuals         14  94.75    6.77
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los efectos Carbono*Presión y Presión*Velocidad siguen siendo no significativos.

Para resolverlo suprimimos la interacción Presión*Velocidad. La tabla ANOVA que corresponde a este modelo es la siguiente:

```
mod <- aov(Altura~ Carbono + Presion +
             Velocidad + Carbono*Presion +
             Carbono*Velocidad, data = factorial)
summary(mod)
```

```
##                   Df Sum Sq Mean Sq F value   Pr(>F)
## Carbono            2  88.08   44.04   6.864 0.007647 **
## Presion            1 150.00  150.00  23.377 0.000218 ***
## Velocidad          1  80.67   80.67  12.571 0.002935 **
## Carbono:Presion    2  14.25    7.12   1.110 0.355049
## Carbono:Velocidad  2 230.58  115.29  17.968 0.000104 ***
## Residuals         15  96.25    6.42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El efecto Carbono*Presión sigue siendo no significativo por lo tanto lo suprimimos y replanteamos

```
mod <- aov(Altura~ Carbono + Presion +
             Velocidad +  Carbono*Velocidad,
           data = factorial)
summary(mod)
```

```
##                   Df Sum Sq Mean Sq F value   Pr(>F)
## Carbono            2  88.08   44.04   6.776 0.006856 **
## Presion            1 150.00  150.00  23.077 0.000166 ***
## Velocidad          1  80.67   80.67  12.410 0.002612 **
## Carbono:Velocidad  2 230.58  115.29  17.737 6.91e-05 ***
## Residuals         17 110.50    6.50
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Todos los efectos de este último modelo planteado son significativos y por lo tanto es en este modelo donde vamos a realizar el estudio. Existen diferencias significativas entre los distintos porcentajes del Carbono, los dos tipos de

presión, las dos velocidades de llenado y la interacción entre el porcentaje de Carbono y la Velocidad de llenado.

7 Ejercicios Guiados.

Se realiza un estudio del contenido de azufre en cinco yacimientos de carbón. Se toman muestras aleatoriamente de cada uno de los yacimientos y se analizan. Los datos del porcentaje de azufre por muestra se indican en la tabla adjunta.

Yacimientos	Porcentaje de azufre
1	151 192 108 204 214 176 117
2	169 64 90 141 101 128 159 156
3	122 132 139 133 154 104 225 149 130
4	75 126 69 62 90 120 32 73
5	80 90 124 82 72 57 118 54 130

Para un nivel de significación del 5%.

1. ¿Se puede confirmar que el porcentaje de azufre es el mismo en los cinco yacimientos?
2. Si se rechaza la hipótesis nula que las medias de porcentaje de azufre en los cinco yacimientos es la misma, determinar que medias difieren entre sí utilizando el método de comparaciones múltiples de Tukey.
3. Estudiar las hipótesis de modelo: Homocedasticidad (Homogeneidad de las varianzas por grupo), Independencia y Normalidad.

Solucion: 1

El problema planteado se modeliza a través de un diseño unifactorial totalmente aleatorizado de efectos fijos no-equilibrado.

- Variable respuesta: Contenido de Azufre
- Factor: Tipo de yacimiento con cinco niveles. Es un factor de Efectos fijos ya que viene decidido qué niveles concretos se van a utilizar.
- Modelo no-equilibrado: Los niveles de los factores tienen distinto número de elementos.
- Tamaño del experimento: Número total de observaciones, en este caso 41 unidades experimentales.

Para realizar este supuesto en R debemos introducir primero los datos de forma correcta. Podemos realizarlo directamente en R de forma manual o introducirlos previamente en un archivo de texto o Excel y leerlos en R.
En este caso lo hacemos en un archivo de texto:

En primer lugar describimos los cinco grupos que tenemos que comparar, los cinco yacimientos, la variable respuesta es el porcentaje de azufre en estos cinco yacimientos. Los yacimientos no tienen todos el mismo número de observaciones, en total tenemos 41 observaciones. La hipótesis nula es que el porcentaje de azufre es el mismo en los cinco yacimientos... Es decir, no hay diferencias en los porcentajes de azufre con respecto a los distintos yacimientos y la hipótesis alternativa es que el porcentaje de azufre es diferente al menos en dos yacimientos.

Tenemos en cuenta que para que el ejercicio esté realizado de forma correcta los datos tienen que estar introducidos tal y como vienen en Figura 27, es decir, las observaciones en una sola columna y a continuación especificado su tratamiento y su bloque correspondiente.

Para cargar los datos utilizamos la función read.table indicando el nombre del archivo (que debe de estar en el directorio de trabajo) e indicando además que tiene cabecera.

Nota: La ruta hasta llegar al fichero varía en función del ordenador. Utilizar la orden setwd() para situarse en el directorio de trabajo

```
porcentaje <- read.table("guiado1-1.txt", header = TRUE)
porcentaje
```

```
##    Azufre Yacimiento
## 1     151          1
## 2     192          1
## 3     108          1
## 4     204          1
## 5     214          1
## 6     176          1
## 7     117          1
## 8     169          2
## 9      64          2
## 10     90          2
## 11    141          2
## 12    101          2
## 13    128          2
## 14    159          2
## 15    156          2
## 16    122          3
## 17    132          3
```

```
## 18    139          3
## 19    133          3
## 20    154          3
## 21    104          3
## 22    225          3
## 23    149          3
## 24    130          3
## 25     75          4
## 26    126          4
## 27     69          4
## 28     62          4
## 29     90          4
## 30    120          4
## 31     32          4
## 32     73          4
## 33     80          5
## 34     90          5
## 35    124          5
## 36     82          5
## 37     72          5
## 38     57          5
## 39    118          5
## 40     54          5
## 41    130          5
```

Debemos transformar la variable referente a los niveles del factor fijo como factor para poder hacer los cálculos de forma adecuada:

```
porcentaje$Yacimiento<-factor(porcentaje$Yacimiento)
porcentaje$Yacimiento
```

```
##  [1] 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 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 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(Azufre ~ Yacimiento, data = porcentaje)
```

donde:

- Azufre: Nombre de la columna de las observaciones.
- Yacimiento: 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 = Azufre ~ Yacimiento, data = porcentaje)
##
## Terms:
##                 Yacimiento Residuals
## Sum of Squares    40432.68  42639.76
## Deg. of Freedom          4        36
##
## Residual standard error: 34.41566
## Estimated effects may be unbalanced
```

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

```
summary(mod)

##             Df Sum Sq Mean Sq F value   Pr(>F)
## Yacimiento   4  40433   10108   8.534 5.97e-05 ***
## Residuals   36  42640    1184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En la Tabla ANOVA, el valor del estadístico de contraste de igualdad de medias, F = 8.534 deja a su derecha un p-valor menor que 0.001, menor que el nivel de significación del 5%, por lo que se rechaza la Hipótesis nula de igualdad de medias. Es decir, existen diferencias significativas en el contenido medio de azufre entre los cinco yacimientos. La pregunta que nos planteamos es si el contenido de azufre es significativamente distinto en los cinco yacimientos o sólo en alguno de ellos. Para responder a esta pregunta utilizamos algún procedimiento de comparaciones múltiples. En el apartado siguiente responderemos a esta cuestión.