Ejercicio 1: Interacción Bloque-factor con una réplica
A una compañía de contabilidad le interesaba potenciar las habilidades de sus auditores pero no sabe cuál de los entrenamientos existentes resulta más efectivo. Los auditores se someten a tres metodologías de entrenamiento:

1: Estudio individual con materiales de entrenamiento programados

2: Sesiones de entrenamiento individuales en sucursales

3: Sesiones de entrenamiento colectivos en Chicago

Treinta auditores fueron agrupados en 10 bloques dependiendo de su antiguedad y dentro de cada bloque los auditores fueron entrenados con alguno de los tres entrenamientos anteriores. Al finalizar el entrenamiento, se les realizó una prueba de habilidades adquiridas. La información se encuentra en el archivo `puntajes.txt`.
En la ayudantía anterior verificamos que nos encontramos un caso balanceado con una réplica:

```
Puntajes<-puntajes$respuesta
Metodo<-factor(puntajes$metodo)  #Factor fijo
Bloque<-factor(puntajes$bloque)  #Variable Bloque (antiguedad)
```

```
addmargins(table(Metodo,Bloque),1)
       Bloque
Metodo 1 2 3 4 5 6 7 8 9 10
   1   1 1 1 1 1 1 1 1 1  1
   2   1 1 1 1 1 1 1 1 1  1
   3   1 1 1 1 1 1 1 1 1  1
   Sum 3 3 3 3 3 3 3 3 3  3
```

a) ¿Existe evidencia para rechazar la no interacción entre la antiguedad y el método? Comente.
Respuesta
Una variable bloque no corresponde realmente al objeto de interes en un estudio, sino que esta presente de manera natural en las unidades experimentales (en este caso, las unidades experimentales son los auditores) y su presencia logra explicar diferencias que pudieran surgir antes de realizar los experimentos. Dado lo anterior, se piensa a priori, que una variable bloque no tiene interaccion con el factor en estudio. Además, notar que nos encontramos en un caso de una réplica. Dado este contexto, se realiza el test de aditividad de Tukey.

```
#Test de Tukey

(a<-length(levels(Metodo))) #Niveles del factor Metodo
(b<-length(levels(Bloque))) #Niveles de la variable bloque

library(dae)
options(contrasts = c("contr.sum", "contr.sum"))
Error.aov<-aov(Puntajes~Metodo+Bloque+Error(Metodo/Bloque))

(Tukey<-tukey.1df(aov.obj=Error.aov,data=puntajes,error.term='Metodo:Bloque'))
$Tukey.SS
[1] 0.1266074

$Tukey.F
[1] 0.01918179
```

```
$Tukey.p
[1] 0.8914739

$Devn.SS
[1] 112.2067

Tukey$Tukey.F>qf(0.95,1,a*b-a-b)
[1] FALSE
```

Por lo tanto, no existe evidencia para rechazar la hipótesis nula de la no interacción, utilizando un 95% de confianza.

b) Utilizando el modelo correspondiente, estime la eficiencia de la variable bloque, sin corregir por grados de libertad. ¿Qué interpretación entrega la eficiencia? Comente. *Respuesta:* E indica qué tan grande debe ser la muestra de un diseño completamente aleatorizado con respecto a un diseño de bloques para que las varianzas sean iguales.
Bajo el modelo de bloques, los estimadores insesgados de los errores experimentales σ_r^2 y σ_b^2 son:

$$\hat{\sigma}_b^2 = MCE$$
$$\hat{\sigma}_r^2 = \frac{(n-1)MCBloque + n(r-1)MCE}{rn-1}$$

Donde n corresponde a la cantidad de bloques y r corresponde a la cantidad de tratamientos (niveles del factor método).
Luego es fácil ver que

$$\hat{E} = \frac{\hat{\sigma}_r^2}{\hat{\sigma}_b^2}$$

Con la corrección por grados de libertad:

$$E =\approx \frac{(gl_2+1)(gl_1+3)}{(gl_1+3)(gl_1+1)}\hat{E}$$

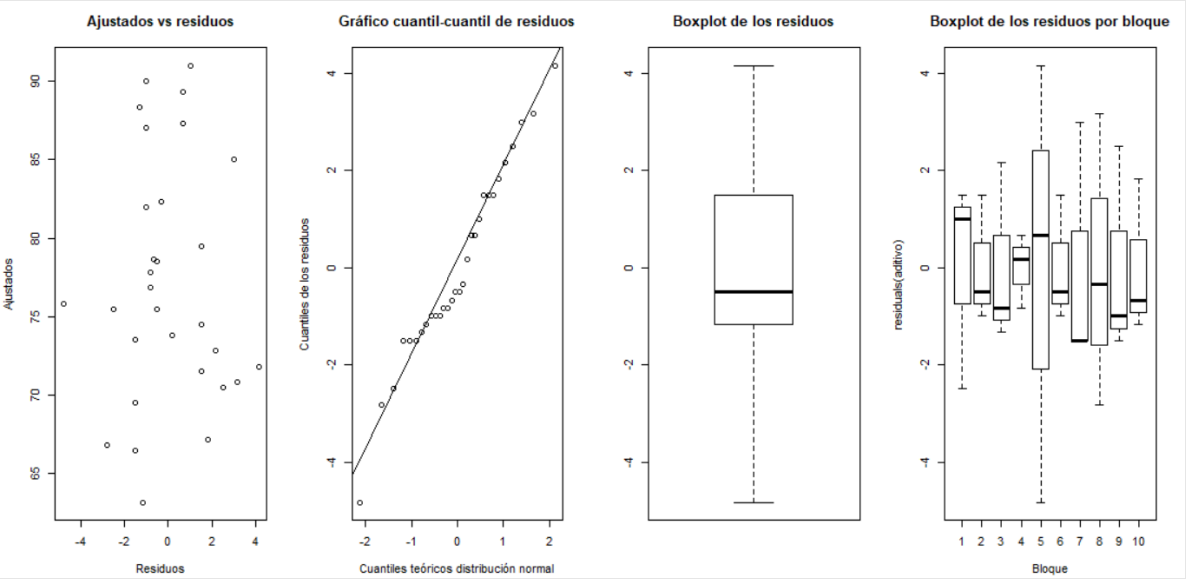
donde $gl_1 = r(n-1)$ y $gl_2 = (r-1)(n-1)$

```
#### Eficiencia de la variable bloque

r<-length(levels(Metodo))  #Niveles de tratamientos
n<-length(levels(Bloque))  #Niveles asociados a la variable bloque

sigmar<-((n-1)*anova(aditivo)[2,3]+n*(r-1)*anova(aditivo)[3,3])/(r*n-1)

sigmab<- anova(aditivo)[3,3]  #MCE

(E<-sigmar/sigmab)  #Eficiencia
[1] 3.084191

#Se requiere 3.084191 veces el tama o muestral en una muestra completamente aleatorizada
# para explicar la misma variabilidad o precision que un dise o de bloque


#Eficiencia corregida por grados de libertad

gl1<-r*(n-1)
gl2<-(r-1)*(n-1)

(Ecorregida<-(gl2+1)*(gl1+3)*E/((gl2+3)*(gl1+1)))
[1] 2.989777
```

d) ¿Cómo se comportan los residuos del modelo? ¿Cómo se comporta el ajuste en cada bloque? Comente.
Respuesta:
Supuestos sobre los residuos modelo ANOVA son:

- Normalidad de ϵ_{ij}
- Homocedasticidad de los residuos
- Independencia

```
### Analisis de residuos

par(mfrow=c(1,4))

plot(residuals(aditivo), fitted(aditivo), main="Ajustados vs residuos",xlab="Residuos",
ylab="Ajustados")

qqnorm(residuals(aditivo),main="Grafico cuantil-cuantil de residuos",
xlab="Cuantiles teoricos distribucion normal",
ylab="Cuantiles de los residuos")
qqline(residuals(aditivo))

boxplot(residuals(aditivo),main="Boxplot de los residuos")
boxplot(residuals(aditivo)~Bloque,main="Boxplot de los residuos por bloque")
```



Respecto al análisis de residuos es posible afirmar que:

- En el gráfico de Residuos y valores ajustados no se observa ninguna tendencia o relación sistemática en los residuos, pero éste análisis debería complementarse con un test de homocedasticidad, análisis de ACF, entre otros.
- En el gráfico cuantil-cuantil, es posible observar que si bien en las colas se observa un comportamiento similar a los de una distribución normal, en el centro se observan diferencias que necesitan ser estudiadas. Este análisis debe complementarse con un test de normalidad de los residuos.

- Al observar el boxplot general de los residuos, no se observan outliers, pero es posible observar que la distribución de los residuos estaría inclinada hacia un lado, pudiera ser interesante realizar un análisis de simetría de la distribución residual.
- En cuanto a los residuos por bloque, no se observan outliers, pero se observan diferencias importantes en término de rangos en los que se mueven los residuos por bloque y además, algunos boxplots se encuentran mayormente inclinados hacia la izquierda o derecha, como por ejemplo, el bloque 1 y el bloque 3. También es posible observar que los diagramas de caja y bigote presentan distintos largos, lo que podría sugerir, que existen diferencias en las variabilidades intra bloques.

Ejercicio 2 Bloques en diseños factoriales

Un científico agrícola quiere estudiar el efecto de 3 fertilizantes diferentes (1,2,3) y además el efecto de dos tipos de aguas de regadío (Agua de regadío 1 que no fue filtrada (contiene minerales) y el agua de regadío 2 que fue filtrada (no contiene minerales)) en la producción de maíz (promedio de mazorcas producidas en las plantas) en distintos campos. Estos campos se agrupan en campos de Tipo A, Tipo B, Tipo C, dependiendo de sus características (humedad de la tierra, densidad de plantas, cantidad de cosechas rendidas, etcétera). La información se encuentra en la base de datos `plantacion`.

a) Determine la naturaleza de las variables en este problema.
Respuesta:

- La variable respuesta corresponde al promedio de mazorcas producidas por terreno/-campo.
- La variable fertilizante corresponde a una variable de tipo factor con 3 niveles, de efectos fijos.
- La variable Agua de regadío corresponde a una variable de tipo factor con 2 niveles, de efectos fijos.
- La variable Campo, corresponde a una variable de tipo bloque con Tipo A, Tipo B y Tipo C.

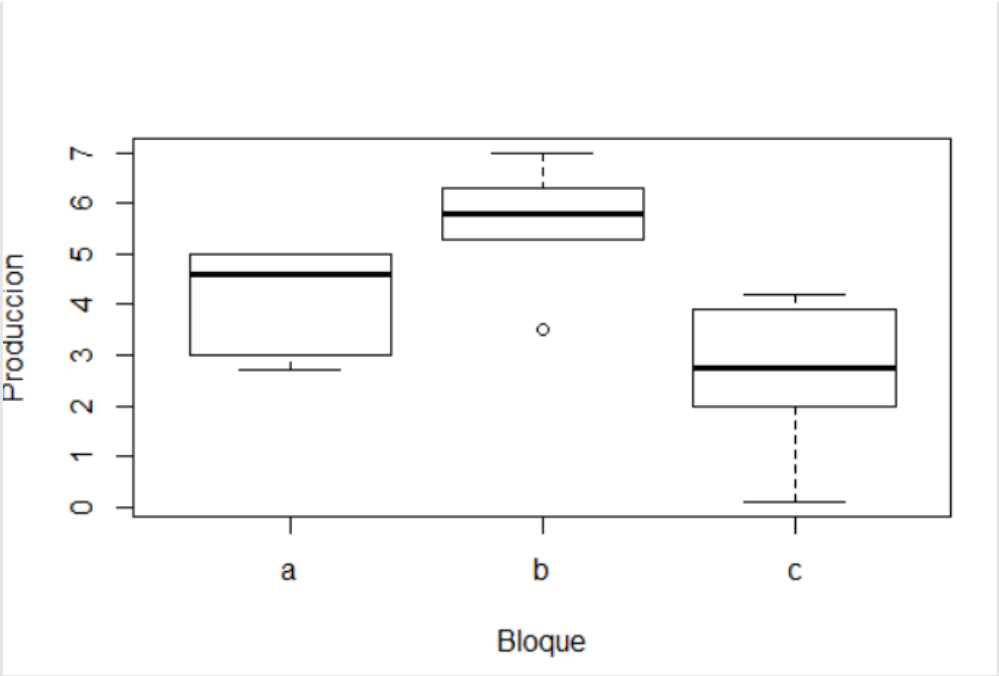
b) Realice una tabla que muestre cómo se distribuye la información en término de los factores y los bloques.
Respuesta:
La tabla, según visto en clases, se puede representar de la siguiente forma:

	Agua de regadío 1			Agua de regadío 2		
	Fertilizante 1	Fertilizante 2	Fertilizante 3	Fertilizante 1	Fertilizante 2	Fertilizante 3
Tipo A						
Tipo B						
Tipo C						

c) ¿Cómo se distribuye la producción de maíz por bloque? Comente.
Respuesta:

```
print(Plantacion)
# A tibble: 18 x 4
   Block  Fert Water     Yield
   <chr> <dbl> <chr>     <dbl>
 1 a         1 Minerals    4.7
```

```
 2 b          1 Minerals    3.5
 3 c          1 Minerals    0.1
 4 a          1 Clean       3
 5 b          1 Clean       5.6
 6 c          1 Clean       2
 7 a          2 Minerals    2.7
 8 b          2 Minerals    6.3
 9 c          2 Minerals    2.2
10 a          2 Clean       5
11 b          2 Clean       7
12 c          2 Clean       4.2
13 a          3 Minerals    4.5
14 b          3 Minerals    6
15 c          3 Minerals    3.3
16 a          3 Clean       5
17 b          3 Clean       5.3
18 c          3 Clean       3.9

Bloque<-factor(Plantacion$Block)

levels(Bloque)
[1] "a" "b" "c"

Fert<-factor(paste("Fertilizante",Plantacion$Fert))

levels(Fert)
[1] "Fertilizante 1" "Fertilizante 2" "Fertilizante 3"

Agua<-factor(Plantacion$Water)

levels(Agua)
[1] "Clean"    "Minerals"

Produccion<-Plantacion$Yield



table(table(Fert,Agua,Bloque))   #Una replica por tratamiento dentro de cada bloque
 1
18

a<-length(levels(Fert))
b<-length(levels(Agua))

#En cada bloque hay a*b observaciones

a*b
[1] 6

df%>%group_by(Bloque)%>%summarise(desviacion=sd(Produccion),
                                  Media=mean(Produccion),
                                  n=length(Produccion),
                                  Maximo=max(Produccion),
                                  Minimo=min(Produccion))
# A tibble: 3 x 6
  Bloque desviacion Media     n Maximo Minimo
  <fct>       <dbl> <dbl> <int>  <dbl>  <dbl>
1 a            1.03  4.15     6    5       2.7
2 b            1.19  5.62     6    7       3.5
3 c            1.52  2.62     6    4.2     0.1
#Es posible observar que en el Bloque c se alcanzan en general menores valores de produccion
# en comparacion a los otros bloques (considerando minimo, maximo y media), y ademas,
#la variabilidad dentro de dicho bloque es la mas alta.

boxplot(Produccion~Bloque)
```



Es posible observar diferencias importantes en términos de producción por bloques. El boxplot asociado al bloque b por ejemplo, se encuentra superior a los demás y presenta una observación anómala. El boxplot asociado al bloque c alcanza valores bastante bajos en comparación a los demás.

d) ¿Qué modelo podría plantearse en este contexto? Plantéelo y definalo en R. Analice su tabla ANOVA y los grados de libertad. Interprete los coeficientes obtenidos.
Respuesta:
El modelo que podría plantearse sería el siguiente:

$$Y_{ijk} = \mu + \rho_i + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk} \quad i = 1,2,3 \quad j = 1,2,3 \quad k = 1,2$$

Con los supuestos usuales de identificabilidad.

```
#Modelo


contrasts(Bloque)<-contr.sum
contrasts(Agua)<-contr.sum
contrasts(Fert)<-contr.sum


modelo<-aov(Produccion~Agua*Fert+Bloque)

anova(modelo)
anova(modelo)
Analysis of Variance Table

Response: Produccion
           Df  Sum Sq Mean Sq F value  Pr(>F)
Agua        1  3.2939  3.2939  3.2286 0.10258
Fert        2  8.6344  4.3172  4.2316 0.04661 *
Bloque      2 27.0044 13.5022 13.2346 0.00155 **
Agua:Fert   2  1.7811  0.8906  0.8729 0.44729
Residuals  10 10.2022  1.0202
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1
```

```
(r<-a*b)   # existen r=a*b tratamientos
[1] 6

#Existen r-1 grados asociados a los tratamientos

a - 1 # Grados asociados al factor agua
[1] 2

b - 1 # Grados asociados al factor fert
[1] 1

(a-1)*(b-1)   #Grados asociados a la interacci n de Agua y fertilizante
[1] 2

#Si se suman, es equivalente a r-1:

((a-1)+(b-1)+(a-1)*(b-1))==(r-1)
[1] TRUE



### Analisis de los efectos

coef(modelo)
(Intercept)       Agua1       Fert1       Fert2     Bloque1     Bloque2 Agua1:Fert1 Agua1:Fert2
 4.12777778  0.42777778 -0.97777778  0.43888889  0.02222222  1.48888889 -0.04444444  0.40555556


# El intercepto corresponde a la media global de la variable Produccion de mazorca

levels(Agua)
[1] "Clean"     "Minerals"

# Agua 1 corresponde al efecto asociado a un agua filtrada en la produccion dejando
# como referencia en el factor Agua, al agua no filtrada (minerales)

#Note que Agua filtrada posee un efecto estimado de 0.427777 y por las restricciones
# de identificabilidad, Agua no filtrada (o mineral) tendr  como efecto asociado
# -0.427777, por lo tanto, se observa un efecto negativo en la producci n de mazorcas
# utilizar un agua no filtrada (con muchos minerales)

#Considerar que la celda de referencia es el Agua no filtrada (con minerales)

#Respecto al Fertilizante es posible observar que el Fertilizante 1 tiene un efecto
# negativo -0.9777 en la produccion de mazorcas, mientras que, el fertilizante 2
# tiene un efecto positivo de 0.4388889, por lo que, podria decirse que el fertilizante
# 2 tendr a un mejor efecto que el fertilizante 1 (de hecho el fertilizante 1 empeora
# el rendimiento)
#Respecto al fertilizante 3 su efecto seria -(-0.9777778+0.4388889)=0.53888889
# el cual es positivo y mayor que el efecto del fertilizante 2, por lo tanto
# el fertilizante 3 tendr a mayor efectividad en la produccion de mazorcas

levels(Bloque)
[1] "a" "b" "c"


# Se tiene que el efecto asociado al bloque C es de -(0.02222+1.488889)=-1.511109
# que corresponde al  nico  efecto negativo en los bloques, por lo tanto, aquellos
# terrenos o campos que se clasifican dentro del Bloque C por sus caracter sticas
# son aquellos predispuestos a un menor rendimiento en terminos de producci n

#Y aquellos terrenos o campos caracterizados en el bloque B son aquellos con mejor
#predisposici n a la produccion de mazorcas


#Tambien se pueden observar las interacciones que corresponden al efecto en una
# combinacion de factores



#tenemos (alphabeta)_11 y (alphabeta)_12

#Las restricciones de identificabilidad:
```

```
#sum_{j}{(alphabeta)_jk}=0   con k=1,2
#sum_{k}{(alphabeta)_jk}=0   con j=1,2,3
#Se resuelve el "sistema" y se pueden obtener los demas.
```