Ejercicio 1: Modelo de efectos fijos con una réplica Complemento Complemento
La base de datos `lapaeggs.csv` contiene información de un experimento realizado por G.P. Quinn en 1988, en el cual era de particular interés conocer cuál es el efecto de la estación del año (Primavera y Verano) y de la Densidad (8, 15, 30 o 45 animales por cada recinto de 225 cm) en la producción promedio de huevos de una lapa de la especie *Siphonaria diemenensis*.

a) Determine cuál es la variable respuesta, cuáles son los factores a estudiar y sus niveles.

Respuesta:
La variable respuesta corresponde a la producción de huevos promedio por cada ambiente (combinación de factores). Los factores a estudiar son:

- Estación del año: 2 niveles (Primavera y Verano)
- Densidad: 4 niveles (8, 12, 30 y 45)

b) Cargue la base de datos. ¿Cuántas réplicas tiene cada ambiente? ¿qué implica esto al realizar un modelo ANOVA?

Respuesta:

```
library(readr)
lapaeggs <- read_csv("C:/Users/HP/Downloads/lapaeggs.csv")
View(lapaeggs)
```

```
Estacion<-factor(lapaeggs$Estacion)
Densidad<-factor(lapaeggs$Densidad)

addmargins(table(Estacion,Densidad),1)
           Densidad
Estacion     8  15  30  45
  Primavera  1   1   1   1
  Verano     1   1   1   1
  Sum        2   2   2   2
```

Se tiene una réplica por ambiente, es decir, una observación por celda. La implicancia directa que tiene esto se notará al momento de especificar el modelo, pues al no tener varias observaciones por celda, el promedio de la variable de interés (variable respuesta) corresponderá a la observación única.

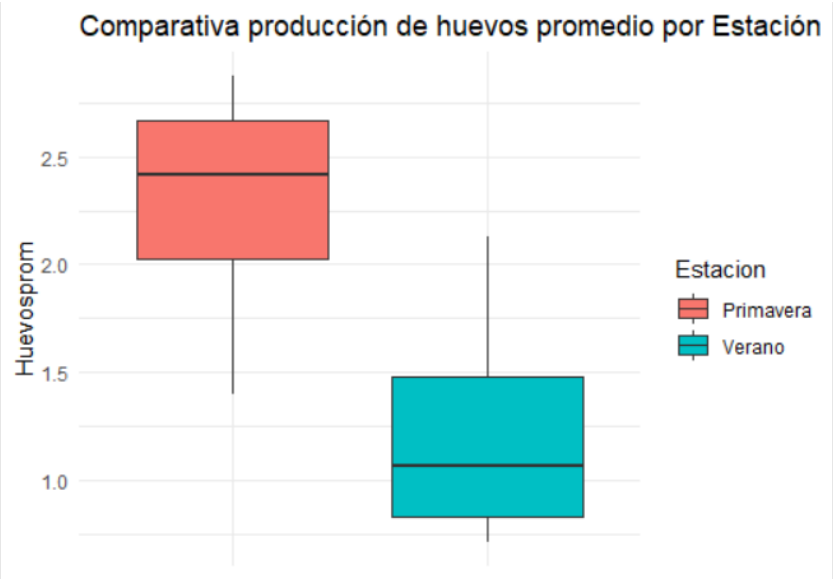
c) Realice análisis gráfico de cada factor y su posible efecto en la variable respuesta. Comente.

Respuesta:

```
df<-data.frame(Huevosprom,Estacion,Densidad)

library(ggplot2)

ggplot(aes(y = Huevosprom, x = Estacion,fill=Estacion), data = df) +
  geom_boxplot()+theme_minimal()+
  theme(axis.text.x = element_blank())+xlab("")+
  ggtitle("Comparativa produccion de huevos promedio por Estaci n")
```



Se puede observar que los boxplots graficados por estación no se encuentran alineados, en particular, se muestra que aquellos experimentos realizados en la Estación Primavera presentaron una mayor producción de huevos que aquellos realizados en Verano. Sin embargo, note que ambos boxplot se solapan (la diferencia no es radical), es pertinente realizar un análisis ANOVA más profundo para determinar el grado del efecto del factor Estación y producción de huevos.

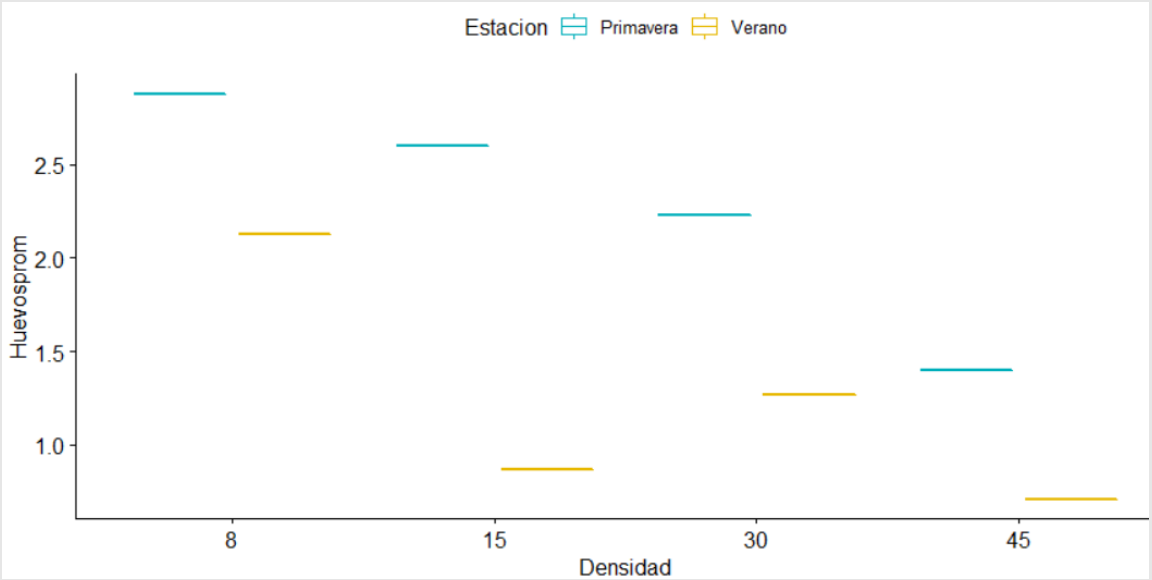
```
ggplot(aes(y = Huevosprom, x = Densidad, fill=Densidad), data = df) +
  geom_boxplot()+theme_minimal()+
  theme(axis.text.x = element_blank())+xlab("")+
  ggtitle("Comparativa produccion de huevos promedio por Densidad")
```

Se pueden observar vastas diferencias. Por ejemplo, los boxplots Densidad 8 y Densidad 45 no se solapan, por lo cual existiría un efecto considerable en aquellos experimentos con Densidad de 45 animales respecto a 8 niveles en la producción de huevos. Parecería que a mayor Densidad menor es la producción de huevos. Sin embargo, existen boxplots que se

solapan entre sí y por ende, sería necesario realizar un análisis a mayor profundidad para dilucidar cuál es el efecto del factor Densidad en la producción de huevos.

d) En el siguiente gráfico se observan boxplots divididos por Estación y Densidad. Comente por qué los boxplots muestran esta estructura y qué podría decir sobre una posible interacción entre los dos factores de interés.

Respuesta:



Note que al tener una réplica por ambiente (combinación entre factores) los boxplots corresponden sólo a una línea dado que cada boxplot representa una observación.
Respecto a la interacción, es fácil ver que en un mismo nivel del factor Densidad se observan diferencias considerables en la producción de huevos, dicha diferencia podría deberse al factor Estación. Es necesario realizar un análisis de interacción.

e) ¿Qué tipo de modelo utilizaría en este contexto? ¿por qué? Plantéelo. ¿Cómo sería el modelo si hubieran n réplicas?

Respuesta:
El factor estación se considera efecto fijo pues no corresponde a estaciones aleatorias, las estaciones son o Primavera o Verano fijadas con certeza. Respecto a las densidades, éstas se encuentran fijadas, si fueran densidades distintas y aleatorias o que incluyeran animales aleatorios (lo cual el ejercicio no lo afirma) pudiera considerarse un efecto aleatorio. Por lo tanto, el modelo a utilizar corresponde a un modelo ANOVA *Two Way* de efectos fijos con una réplica.

El modelo a utilizar sería el modelo aditivo (dado que estamos en el caso balanceado de una réplica):

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad i = 1, 2 \quad j = 1, 2, 3, 4$$

Donde:
α_i corresponde al efecto principal del nivel i del factor Estación. Note que $i = 1, 2$.
β_j corresponde al efecto principal del nivel j del factor Densidad. Note que $j = 1, 2, 3, 4$
ϵ_{ij} corresponde al error asociado en cada observación ij, y es sabido que $\epsilon_{ij} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$

Las restricciones de identificabilidad son las siguientes:

$$\sum_{i=1}^{2} a_i = 0$$

$$\sum_{j=1}^{4} b_j = 0$$

Si hubieran n réplicas, la especificación del modelo incorporaría el subíndice k donde en el caso balanceado con $n > 1$ réplicas sería $k = 1, ..., n$ y el modelo resultaría como sigue:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \quad \imath = 1, 2 \quad j = 1, 2, 3, 4 \quad k = 1, ..., n$$

f) Realice el test de Tukey de aditividad. Comente.

Respuesta:

Pasos:

1. Obtener $\hat{\alpha}_i$ y $\hat{\beta}_j$ del modelo aditivo.
2. Estimar D utilizando que:

$$\hat{D} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{4} \hat{\alpha}_i \hat{\beta}_j Y_{ij}}{\sum_{i=1}^{2} \hat{\alpha}_i^2 \sum_{j=1}^{4} \hat{\beta}_j^2}$$

3. Definir SCABp como sigue:

$$SCABp = \sum_{i=1}^{2}\sum_{j=1}^{4} \hat{D}^2 \hat{\alpha}_j \hat{\beta}_j$$

4. Desde la tabla anova del modelo aditivo extraer SCT, SCA, SCB.
5. Definir SCEp como sigue:

$$SCEp = SCT - SCA - SCB - SCABp$$

6. Definir el estadístico Fp:

$$Fp = \frac{\frac{SCABp}{1}}{\frac{SCEp}{ab-a-b}}$$

7. Regla de decisión:

$$\text{Si} \quad Fp > F_{(1,ab-a-b)} \quad \text{Rechazo la hipótesis nula}$$

En R:

```
options(contrasts = c("contr.sum", "contr.sum"))  #Restricciones de identificabilidad
av = aov(Huevosprom ~ Estacion+Densidad, df)      #Modelo aditivo

summary(av)                                       #Importante revisar los df de cada factor
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
Estacion     1 2.1373  2.1373  18.496 0.0231 *
Densidad     3 2.0895  0.6965   6.028 0.0871 .
Residuals    3 0.3467  0.1156
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

(isBalanced=!is.list(replications(av, data=df)))  #Otra forma de verificar que es balanceado
[1] TRUE

coef(av)    #El ultimo nivel de cada factor lo expresa en funcion de los demas niveles
(Intercept)   Estacion1   Densidad1   Densidad2   Densidad3
   1.759375    0.516875    0.740625   -0.025875   -0.010875

#El intercepto corresponde a la media global de la produccion de huevos
#Estacion 1 corresponde al efecto de Primavera en la media de la produccion de huevos
#Densidad 1 corresponde al efecto de la densidad=8 en la media de la produccion de huevos
#Notar que los efectos Densidad2 y Densidad3 son negativos, mayor densidad disminuye
#promedio de produccion
# tiende a provocar una disminucion en el promedio de la produccion de huevos

#Clase 20-04-2020

#Paso 1) Extraer vectores de los coeficientes estimados para cada factor

avalues<-ifelse(df$Estacion=="Primavera",unname(coef(av)[2]),-unname(coef(av)[2]))
bvalues<-ifelse(df$Densidad=="8",unname(coef(av)[3]),
ifelse(df$Densidad=="15",unname(coef(av)[4]),
ifelse(df$Densidad=="30",unname(coef(av)[5]),
-(unname(coef(av)[3])+unname(coef(av)[4])+unname(coef(av)[5])))))

#Paso 2) Estimar D

(Dhat<-sum(avalues*bvalues*Huevosprom)/(sum(unique(avalues)^{2})*sum(unique(bvalues)^{2})))
[1] 0.01405989

#Paso 3) Definir SCABp

(SCABp<-Dhat^{2}*sum(avalues^{2}*bvalues^{2}))
[1] 0.0001103516

#Paso 4) Extraer SCA, SCB y SCT del modelo aditivio

SCA<-anova(av)[1,2]
SCB<-anova(av)[2,2]
SCT<-sum(anova(av)[,2])

#Paso 5) Definir SCEp

(SCEp<-SCT-SCA-SCB-SCABp)
[1] 0.346551

#Paso 6) Definir el estadistico Fp

(a<-length(levels(Estacion)))   #Niveles del factor Estacion
[1] 2
(b<-length(levels(Densidad)))   #Niveles del factor Densidad
[1] 4

(Fp=(SCABp/1)/(SCEp/(a*b-a-b)))
[1] 0.0006368565

#Paso 7) Regla de decision

Fp>qf(0.95,1,a*b-a-b)
[1] FALSE

#No se rechaza la hipotesis nula, el modelo aditivo es correcto


#Otra alternativa (funcion del paquete dae)


install.packages("dae")
```

```
library(dae)

Error.aov<-aov(Huevosprom~Estacion+Densidad+Error(Estacion/Densidad))

tukey.1df(aov.obj=Error.aov,data=df,error.term='Estacion:Densidad')
$Tukey.SS
[1] 0.0001103516

$Tukey.F
[1] 0.0006368565

$Tukey.p
[1] 0.9821583

$Devn.SS
[1] 0.346551

#Entrega los mismos resultados recien calculados
```

Ejercicio 2: Modelo de efectos fijos caso desbalanceado Complemento

Suponga que el investigador a cargo del estudio anterior logró replicar algunos ambientes y obtuvo más información de éstos. La información completa se encuentra en el archivo `lapaegs2.csv`.

a) Verifique que éste corresponde a un caso no balanceado.

Respuesta:

```
### Caso desbalanceado


lapaeggs2 <- read_csv("C:/Users/HP/Downloads/lapaeggs2.csv")

View(lapaeggs2)

Estacion2<-factor(lapaeggs2$Estacion)
Densidad2<-factor(lapaeggs2$Densidad)
Huevosprom2<-lapaeggs2$Huevos

addmargins(table(Estacion2,Densidad2),1)

           Densidad2
Estacion2   8 15 30 45
  Primavera 2  2  1  2
  Verano    1  2  1  2
  Sum       3  4  2  4
```

b) Plantee el modelo completo.

Respuesta:
La cantidad de réplicas por combinación de tratamientos (nivel i del factor A (Estación) y nivel j del factor B (Densidad)) es n_{ij}. El modelo propuesto es:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \quad i = 1,2 \quad j = 1,2,3,4 \quad k = 1,...,n_{ij}$$

Donde $\epsilon_{ijk} \overset{\text{i.i.d}}{\sim} N(0,\sigma^2)$ y n_{ij} puede extraerse de la tabla anterior. Las restricciones de identificabilidad se aplican a cada uno de los componentes del modelo:

$$\sum_{i=1}^{2} \alpha_i = 0$$

$$\sum_{j=1}^{4} \beta_j = 0$$

$$\sum_{i=1}^{2} (\alpha\beta)_{ij} = 0 \quad j = 1, 2, 3, 4$$

$$\sum_{j=1}^{4} (\alpha\beta)_{ij} = 0 \quad i = 1, 2$$

c) Realice un test de hipótesis que permita dilucidar si en presencia de la variable Estación, el efecto de la variable Densidad es significativo (en un modelo aditivo).

Respuesta:
Debemos utilizar el test F de modelos anidados. Considere:

$$\text{Modelo 1 : Huevos} \sim \text{Estacion} \quad Y_{ik} = \mu + \alpha_i + \epsilon_{ik}$$

$$\text{Modelo 2 : Huevos} \sim \text{Estacion+Densidad} \quad Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

Queremos estudiar si en presencia de la variable Estación, la variable Densidad es significativa, es decir:

$$\text{H}_0 : \text{El modelo reducido es correcto}$$

$$\text{H}_1 : \text{El modelo completo es correcto}$$

Pasos:

1. Plantear el modelo 1 (modelo reducido) y plantear el modelo 2 (modelo completo aditivo)
2. Extraer $SCE_{\text{Modelo 1}}$ y $SCE_{\text{Modelo 2}}$ con sus respectivos grados de libertad $glE_{\text{Modelo 1}}$ y $glE_{\text{Modelo 2}}$
3. Obtener el estadístico F:

$$F = \frac{(SCE_{\text{Modelo 1}} - SCE_{\text{Modelo 2}}/(glE_{\text{Modelo 1}} - glE_{\text{Modelo 2}})}{SCE_{\text{Modelo 2}}/glE_{\text{Modelo 2}}}$$

4. Regla de decisión:

$$\text{Si} \quad F > F_(glE_{\text{Modelo 1}} - glE_{\text{Modelo 1}}, glE_{\text{Modelo 2}}) \quad \text{Rechazo la hipótesis nula}$$

En R:

```
#Test F modelos anidados

#Paso 1) Planteo un modelo reducido (sin Densidad) y otro completo

options(contrasts = c("contr.sum", "contr.sum"))
modelo1<-aov(Huevosprom ~ Estacion, df)  #Modelo reducido
modelo2<-aov(Huevosprom ~ Estacion+Densidad, df)

#Paso 2) Extraigo SCEm1, SCEm2, glEm1 y glEm2

SCEm1<-anova(modelo1)[2,2]
glEm1<-anova(modelo1)[2,1]

SCEm2<-anova(modelo2)[3,2]
glEm2<-anova(modelo2)[3,1]
```

```
#Paso 3) Plantear el Estadistico F

(F=((SCEm1-SCEm2)/(glEm1-glEm2))/(SCEm2/glEm2))
[1] 6.027514

#Paso 4) Regla de decision

F>qf(0.95,glEm1-glEm2,glEm2)
[1] FALSE
#No se rechaza la hipotesis nula

#Notar que al utilizar un 90% de confianza, se obtiene otro resultado

F>qf(0.9,glEm1-glEm2,glEm2)
[1] TRUE
```

Utilizando un 95% de confianza, el modelo simple Huevos $\sim$ Estacion es correcto.

Ejercicio 3: Diseño en Bloque Complemento

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`.

a) Realice un gráfico de interacción. Comente.
Respuesta:

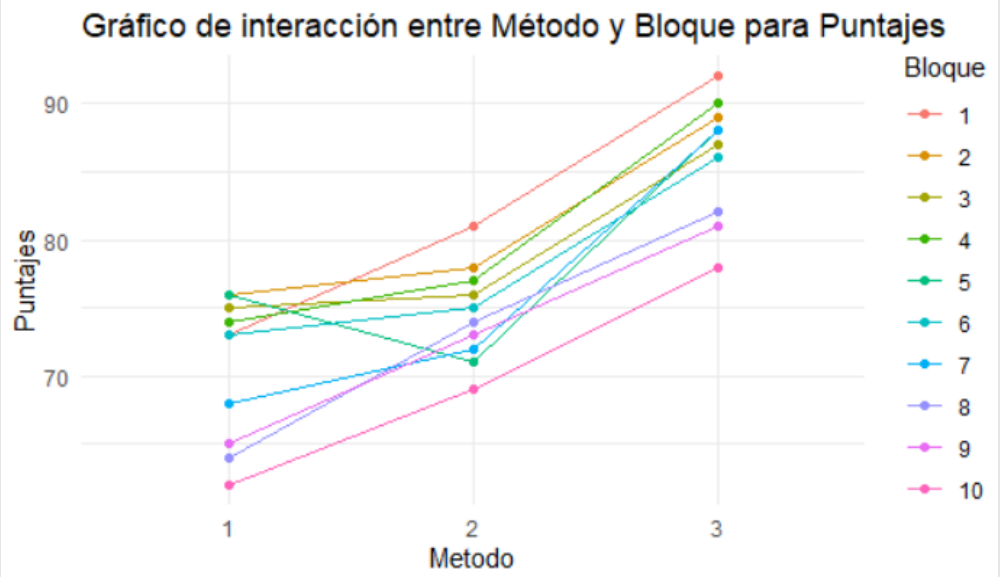
```
#Carga la data

library(readr)
puntajes <- read_delim("C:/Users/HP/Downloads/puntajes.txt",
                       "\t", escape_double = FALSE, trim_ws = TRUE)
View(puntajes)

Puntajes<-puntajes$respuesta
Metodo<-factor(puntajes$metodo)
Bloque<-factor(puntajes$bloque)


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

puntajes %>%
  ggplot() +
  aes(x = Metodo, y = Puntajes, color = Bloque) +
  geom_line(aes(group = Bloque)) +
  geom_point()+theme_minimal()+
  ggtitle("Grafico de interaccion entre Metodo y Bloque para Puntajes")
```



Es posible observar que existen ciertos cruces entre grupos. Sin embargo, no todas las líneas se cruzan, por lo tanto no es posible afirmar desde ya que existiría una interacción significativa entre las variables Método y Bloque. Es necesario realizar un test de aditividad.

b) Realice un test de Tukey de aditividad para cuantificar el grado de interacción.

Respuesta:

```
#Test de Tukey

(a<-length(levels(Metodo))) #Niveles del factor Metodo
[1] 3
(b<-length(levels(Bloque))) #Niveles de la variable bloque
[1] 10

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

#Utilizando un 95% de confianza la interaccion no es significativa
```

c) Ajuste el modelo que corresponda y realice un test para verificar si la metodología sugiere un efecto significativo en los puntajes.

Respuesta:

Dada la conclusión anterior, el modelo adecuado sería el modelo aditivo:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad i = 1, 2, 3 \quad j = 1, ..., 10$$

Con los contrastes usuales $\sum_{i=1}^{3} \alpha_i = 0$ y $\sum_{j=1}^{10} \beta_j = 0$, además $\epsilon_{ij} \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$.
Las hipótesis son:

$$\mathrm{H}_0 : \alpha_i = 0 \quad \forall i = 1, 2, 3$$

$$\mathrm{H}_1 : \exists\ \alpha_i \neq 0 \quad \text{con} \quad i = 1, 2, 3$$

Pasos:

1. Plantear el modelo aditivo
2. Extraer SC_{Metodo} y SCE y definir sus grados de libertad
3. Calcular

$$MC_{\text{Metodo}} = SC_{\text{Metodo}}/(a-1)$$

$$MCE = \frac{SCE}{ab - a - b}$$

4. Calcular estadístico F

$$F = \frac{MC_{\text{Metodo}}}{MCE}$$

5. Regla de decisión

$$\text{Si} \quad F > F_{(a-1,(b-1)(a-1))} \qquad \text{Se rechaza la hipótesis nula}$$

En R:

```
#Test F de significancia del efecto

#Paso 1) Definir el modelo aditivo
options(contrasts = c("contr.sum", "contr.sum"))
aditivo<-aov(Puntajes~Metodo+Bloque)


#Paso 2) Extraer SCE y definir grados de libertad
anova(aditivo)

SCmetodo<-anova(aditivo)[1,2]
SCE<-anova(aditivo)[3,2]

a<-length(levels(Metodo)) #Niveles del factor Metodo
r<-(b-1)*(a-1)    #grados de libertad asociados a SCE

#Paso 3) Calcular MCmetodo  y MCE

MCmetodo<-SCmetodo/(a-1)

MCE<-SCE/r

#Paso 4) Calcular estadistico F

(Fmetodo<-MCmetodo/MCE)
[1] 103.7537

#Paso 5) Regla de decisi n

Fmetodo>qf(0.95,a-1,(b-1)*(a-1))
[1] TRUE

#Con un 95% de confianza, la metodolog a presenta un efecto
# significativo en los puntajes
```

d) Evalúe si la media en todos los bloques es la misma, es decir, si la estratificación por bloques realizada sugiere un efecto significativo en los puntajes.
Respuesta:

```
#Para bloque es analogo, pero la lectura es distinta

SCbloque<-anova(aditivo)[2,2]
b<-length(levels(Bloque)) #Niveles de la variable bloque

MCbloque<-SCbloque/(b-1)

(Fbloque<-MCbloque/MCE)
[1] 7.715727

Fbloque>qf(0.95,b-1,(b-1)*(a-1))
[1] TRUE
```

Utilizando un 95% de confianza, estratificar por la variable bloque sí sugiere un efecto significativo.