

Laboratorio. ANOVA de dos vías

En este ejercicio se irán dando instrucciones que deberá ejecutar en RStudio.

El reporte es un documento con un formato uniforme y coherente, en el que se incluye texto explicativo del trabajo que se va realizando.

Deberá crear un documento reporte de entrega donde se incluye el código ejecutado y las salidas obtenidas en RStudio. Puede usar este mismo documento como base e ir agregando los resultados y gráficos obtenidos.

Además, deberá explicar los resultados cuando así se le solicite en el documento (preguntas que empiezan con "Q*").

El código R viene precedido de ">". Debe incluir en el reporte los resultados de la ejecución de código R (debajo del texto del código), así como los gráficos que se le soliciten.

El trabajo puede entregarse en grupos de dos personas.

Las entregas se realizarán en formato PDF. La letra debe ser Times New Roman, Arial o Cambria. Los tamaños permitidos son 10-12.

Debe incluir una portada de página completa en la que incluya el nombre del curso, del laboratorio, así como los nombres completos de los estudiantes. Esta página será utilizada por el profesor o el asistente para la calificación y comentarios, en caso de tener que incluir comentarios generales del informe.

El reporte se debe entregar en Mediación Virtual, en un archivo .pdf que pueda ser leído en programas comerciales de uso habitual. Debe verificar que el .pdf que subió a Mediación Virtual contiene los ejercicios resueltos y que el archivo puede abrirse correctamente. En caso de problemas con el archivo .pdf (no abre correctamente, está corrupto, etc.) se considerará que no entregó la tarea.

Las entregas tardías se penalizarán con un 10% de la nota luego de vencida la fecha y hora de entrega, más un 10% adicional por cada hora de retraso.

Ejercicio en R.

Para este ejercicio usaremos el conjunto de datos adjunto en el archivo IDE2025.csv. Contiene datos **inventados** de un estudio que evalúa el efecto del IDE de programación y la experiencia de los programadores en el tiempo en que duran en codificar un programa particular. El experimento se realizó con 60 programadores, donde cada uno se categorizó en 0.5 (menos de medio año de experiencia), 1 (alrededor de 1 año de experiencia) y 2 (2 ó más años de experiencia). Los programadores utilizaron el entorno VSCode o IntelliJ IDEA.

La variable de respuesta es la duración en horas en que resolvieron la tarea (variable Duracion).

Los factores de diseño son el entorno de programación utilizado (Herramienta) y la experiencia del programador (Experiencia).

```
# Cargar los datos del archivo IDE2025.csv
```

```
>
```

```
library(tidyverse)
```

```
>
```

```
IDE= (read.csv(file.choose(), header=T, encoding = "UTF-8"))  
attach(IDE)
```

```
# Verificar la estructura
```

```
>
```

```
str(IDE)
```

De la salida anterior, R considera la variable "Experiencia" como una variable numérica.

La convertiremos como una variable de **factor** (es decir, variable de agrupación) de la siguiente manera.

```
# Convertir Experiencia como factor y recodificar los niveles
```

```
# como "Novato", "Intermedio", "Avanzado" (Etiquetas Nov, Int, Ava)
```

```
>
```

```
IDE$Experiencia <- factor (IDE$Experiencia,  
                           levels = c(0.5, 1, 2),  
                           labels = c("Nov", " Int", " Ava"))
```

Validamos el cambio en la variable

```
>
```

```
str(IDE)
```

Revisión general de datos

Creamos tabla de frecuencia.

```
# Tabla de Frecuencias
```

```
>
```

```
table(IDE$Herramienta, IDE$Experiencia)
```

Q1 ¿Qué información obtenemos de la tabla anterior?

Calculamos algunas estadísticas generales de las variables

```
#Calcule la media, la varianza y la desviación estándar por grupos
```

```
>
```

```
group_by(IDE, Experiencia) %>%  
  summarise(  
    count = n(),  
    mean = mean(Duracion, na.rm = TRUE),  
    var = var(Duracion, na.rm = TRUE),  
    sd = sd(Duracion, na.rm = TRUE)  
  )
```

Q2. Similar al anterior, escriba el código R que presente en filas y columnas la media, varianza y desviación estándar del factor “Herramienta”. Muestre el código y el resultado obtenido.

Ahora veamos el detalle agrupando tanto “Herramienta” como “Experiencia”:

>

```
group_by(IDE, Herramienta, Experiencia) %>%
  summarise(
    count = n(),
    mean = mean(Duracion, na.rm = TRUE),
    var = var(Duracion, na.rm = TRUE),
    sd = sd(Duracion, na.rm = TRUE)
  )
```

Creamos boxplots para las variables

Tiempo en codificación por Herramienta usada

>

```
boxplot(Duracion ~ Herramienta, data=IDE, frame = FALSE,
        col = c("#00AFBB", "#E7B800"), ylab=" Duracion")
```

Q3 ¿Qué parece indicar el boxplot anterior viendo las medianas y la variabilidad?

Tiempo en codificación por Experiencia del programador

Q4 Cree un diagrama similar al anterior para la variable “Experiencia”. Presente el código y el gráfico resultante.

Q5 ¿Qué parece indicar el boxplot anterior viendo las medianas y la variabilidad?

Duración en horas por combinación de Herramienta y Experiencia

>

```
boxplot(Duracion ~ Herramienta * Experiencia, data=IDE, frame = FALSE,
        col = c("#00AFBB", "#E7B800"), ylab="Duracion")
```

Invirtiendo las variables independientes

>

```
boxplot(Duracion ~ Experiencia * Herramienta, data=IDE, frame = FALSE,  
        col = c("#00AFBB", "#E7B800"), ylab="Duracion")
```

Q6 ¿Hay información adicional que pueda extraer de los dos boxplots anteriores?

Seguidamente se construirán gráficos para visualizar posibles interacciones de las variables de diseño. Líneas paralelas indicarían que no existe interacción, mientras que líneas que se cruzan sí mostrarían una **posible** interacción.

>

```
IDE %>%  
  ggplot() +  
  aes(x = Herramienta, color = Experiencia, group = Experiencia, y = Duracion) +  
  stat_summary(fun = mean, geom = "point") +  
  stat_summary(fun = mean, geom = "line")
```

Q7 Cree un diagrama similar al anterior pero agrupando por la variable “Herramienta”. Presente el código y el gráfico resultante.

Q8 ¿Qué puede deducir de los gráficos anteriores? ¿Existe interacción entre las variables para alguno de los niveles de los factores?

Otra herramienta para crear estos gráficos es `interaction.plot()`

>

```
interaction.plot(x.factor = IDE$Experiencia, trace.factor = IDE$Herramienta,  
               response = IDE$Duracion, fun = mean,  
               type = "b", legend = TRUE,  
               xlab = "Experiencia", ylab="Duración en horas",  
               pch=c(1,19), col = c("#00AFBB", "#E7B800"))
```

Q9. Construya un gráfico equivalente para ver las interacciones donde el factor x es de la variable “Herramienta” y el factor de trace es la variable “Experiencia”. Muestre el código y el gráfico resultante.

Calculamos la prueba ANOVA de dos vías

Queremos saber si la duración en horas depende del Herramienta y la Experiencia.

La función `aov()` se puede usar para responder esta pregunta.

La función `summary.aov()` se utiliza para resumir el modelo de análisis de varianza. También puede usar la función `summary()`.

Vamos a establecer un nivel de significación (Alpha) de 0.05 para todas las pruebas Anova que se presentan seguidamente.

#ANOVA de dos vías **sin interacción** de las variables independientes (signo +)

>

```
res.aov <- aov(Duracion ~ Herramienta + Experiencia, data = IDE)  
summary(res.aov)
```

La salida incluye las columnas F value y Pr (> F) correspondientes al valor p de la prueba.

Q10 ¿Qué información nos da esta prueba ANOVA sin interacción? ¿Depende la duración en horas del Herramienta y/o de la Experiencia?

ANOVA de 2 vías **con** efecto de interacción

ANOVA de dos vías con interacción de las variables independientes (signo *)

Estas dos llamadas son equivalentes

metodo 1

>

```
res.aov_inter <- aov(Duracion ~ Herramienta * Experiencia, data = IDE)
summary(res.aov_inter)
```

metodo 2

>

```
res.aov_inter2 <- aov(Duracion ~ Herramienta + Experiencia + Herramienta:Experiencia, data = IDE)
summary(res.aov_inter2)
```

Q11 ¿Qué información nos da esta prueba ANOVA con interacción? ¿Depende la duración en horas del Herramienta y/o la Experiencia? ¿Hay interacción entre Experiencia y Herramienta?

Múltiples comparaciones por pares

Como la prueba ANOVA es significativa, podemos calcular TukeyHSD (Tukey Honest Significant Differences, R function: TukeyHSD ()) para realizar múltiples comparaciones por pares entre las medias de los grupos. La función TukeyHD () toma el ANOVA ajustado como argumento.

Dado que el ANOVA sí encontró interacción entre variables, utilizaremos el modelo que sí incluye la interacción de los factores (**res.aov_inter**).

La prueba la podemos realizar indicando que se quiere comparar solo los niveles del factor "Experiencia".

>

```
TukeyHSD(res.aov_inter , which = "Experiencia")
```

El gráfico de intervalos de confianza para la variable Experiencia sería:

>

```
par(mar = c(2, 6, 2, 2))  
plot(TukeyHSD(res.aov, conf.level=.95, which = "Experiencia"), las = 1)
```

Q12 ¿Qué se puede deducir del Tukey anterior sobre el factor Experiencia?

También se pueden realizar múltiples comparaciones para los niveles de “Herramienta” (que solo son 2).

>

```
TukeyHSD(res.aov_inter, which = "Herramienta")
```

Q13 ¿Tiene sentido analizar las comparaciones múltiples de “Herramienta”?

Finalmente, se podrían realizar comparaciones Tukey para el modelo con interacción. Estas comparaciones múltiples incluyen las pruebas recién hechas para “Experiencia” y “Herramienta” y además agregan la información de las combinaciones de Experiencia y Herramienta.

>

```
TukeyHSD(res.aov_inter)
```

Q14 Presenta la parte de los resultados con las combinaciones Herramienta:Experiencia, señalando en cuáles comparaciones de tratamientos sí hay diferencia significativa.

Verificación de supuestos

Normalidad

Para verificar la normalidad de los residuales construiremos un gráfico de histograma y un gráfico QQ.

>

```
hist(res.aov_inter$residuals)
```

>

```
qqnorm(res.aov_inter$residuals)
```

```
qqline(res.aov_inter$residuals)
```

Q15 ¿Qué puede deducir de los gráficos anteriores? ¿Parecen seguir una distribución normal?

También se ejecuta una prueba Shapiro Wilks

>

```
shapiro.test(res.aov_inter$residuals)
```

Q16 ¿Siguen los residuales una distribución normal? Justifique su respuesta.

Independencia

Puede graficar los residuales para ver si se presentan patrones. También puede usar el hecho de que las observaciones fueron correctamente aleatorizadas.

>

```
plot(res.aov_inter$residuals)
```

Q17 ¿Dado el gráfico anterior podría intuir que son independientes?

Homogeneidad de varianzas

Se grafican los residuales versus los valores ajustados (fitted). Se puede ejecutar el comando en R `plot ()`, que genera 4 gráficos, o señalar específicamente cuál gráfico desplegar (en este caso el 1).

>

```
plot(res.aov_inter, 1)
```

Q18 ¿Dado el gráfico anterior se podría intuir que las varianzas son homogéneas?

También se ejecuta la prueba de Bartlett, dado que ya se probó normalidad. Recuerde que también puede usar la prueba de Levene.

>

```
inter <- interaction(IDE$Experiencia, IDE$Herramienta)
bartlett.test(Duracion ~ inter, data = IDE)
```

Q19 ¿Se puede afirmar que las varianzas de los residuales son homogéneas? Explique.

Calcule ahora el tamaño del efecto. Se puede calcular como $SS \text{ entre grupos} / (SS \text{ total} + SS \text{ error})$ o usar la función `etaSquared` del paquete `lsr`.

>

```
library(lsr)
etaSquared(res.aov_inter, anova = TRUE)
```

Q20 ¿Qué puede decir del tamaño del efecto de cada factor? ¿Y la interacción?

Debe usar el valor `eta.sq.part`, que hace el ajuste cuando son varios factores.

También puede usar la biblioteca `effectsize`.

>

```
library(effectsize)
effectsize::eta_squared(res.aov_inter)
```

Ahora, ¿qué pasa con la **potencia** de la prueba?

Para este caso en el que tenemos un ANOVA de dos vías, usaremos la biblioteca `pwr2` y el estadístico `pwr.2way`.

Asociaremos la variable `a` al factor Herramienta, y `b` al factor Experiencia.

```
# pwr.2way(a=a, b=b, alpha=alpha, size.A=size.A, size.B=size.B, f.A=NULL, f.B=NULL, delta.A=NULL,
delta.B=NULL, sigma.A=NULL, sigma.B=NULL)
```

En este caso necesitaremos al menos los siguientes parámetros:

- En `a`, `b` se indica la cantidad de grupos (niveles) por factor.
- En `alpha` pondremos el nivel de significación, 0.05.
- En `size.A` y `size.B` se indica la cantidad de observaciones de cada grupo.
- En `f.A` y `f.B` se deben indicar el efecto de los factores A y B, **expresados con la *f* de Cohen**.
Este efecto no es igual al calculado en `eta.square`, por lo que se debe hacer un ajuste:

$$f = \sqrt{\frac{\eta^2}{1 - \eta^2}}$$

Entonces, si tenemos los efectos del **eta square parcial** para Herramienta y Experiencia iguales a 0.15 y 0.35 respectivamente (**es un solo un ejemplo, no son los valores reales de este ejercicio!**), calculamos los `f.A` y `f.B` de la siguiente manera:

>

```
library(pwr2)
efectoHerr = 0.15 # se toma del eta.squared.part que usted obtuvo
efectoExp = 0.35 # se toma del eta.squared.part que usted obtuvo
fA = sqrt(efectoHerr / (1 - efectoHerr))
fB = sqrt(efectoExp / (1 - efectoExp))
print(fA)
print(fB)
```

También podemos calcular f_A y f_B usando el paquete `effectsize`, con la función `cohens_f`:

>

```
effectsize::cohens_f(res.aov_inter)
```

Finalmente podemos calcular la potencia de la prueba:

>

```
pwr.2way(a=2, b=3, alpha=0.05, size.A=30, size.B=20, f.A=fA, f.B=fB)
```

Q22 ¿Qué potencia se obtuvo dados los datos de la muestra? ¿Es un valor apropiado para un experimento?

Fin del laboratorio.