ANOVA para muestras correlacionadas

Ejemplo de solución ejercicio prático N°6

Enunciado

Un equipo de investigación del área de interacción humano-información está estudiando si el área temática y el nivel de dificultad del problema de información influyen en el tiempo (en segundos) que le toma a una persona en formular una consulta de búsqueda para resolver dicho problema.

Para ello, han reclutado a un grupo voluntario de participantes, asignados aleatoriamente a distintos grupos. Cada participante debe resolver tres problemas de información con diferentes niveles de dificultad: baja, media y alta. A su vez, cada grupo debe resolver problemas relacionados a una temática diferente. Los datos recolectados contemplan las siguientes variables:

Descripción de los datos

Variable	Descripción
id	Identificador único de cada participante.
area	Área temática de los problemas que cada participante debe responder. Variable categórica con los niveles Arquitectura, Biología, Computación, Economía, Física, Leyes, Literatura, Matemáticas, Música, Pedagogía, Psicología, Química.
dificultad	Nivel de dificultad del problema resuelto. Variable categórica con los niveles Baja, Media y Alta.
tiempo	Tiempo, en segundos, que toma a cada participante formular la consulta

En este momento, el equipo de investigación busca determinar si existen diferencias en el tiempo que tardan las personas en formular consultas para problemas con diferentes niveles de dificultad en el área de matemáticas.

En esta pregunta se pide inferir acerca de medias de una variable numérica (tiempo) medida en condiciones distintas (niveles de dificultad) para un conjunto de las mismas personas, lo que correlaciona las mediciones. Luego se requiere usar un **procedimiento ANOVA para muestras correlacionadas**. Las hipótesis serían:

H_0: no hay diferencia en los tiempos requeridos por las mismas personas para formular consultas asociadas a un problema de información en el área de las matemáticas al considerar niveles de dificultad bajo (B), medio (M) y alto (A); es decir: $\mu_{(B-M)} = \mu_{(B-A)} = \mu_{(M-A)} = 0$.

H_A: hay diferencia en los tiempos requeridos por las mismas personas para formular consultas asociadas a problemas de información en el área de las matemáticas con diferentes niveles de dificultad, es decir $\exists\, i, j \in \{B, M, A\} : \mu_{(i-j)} \neq 0.$

Comencemos cargando los paquetes que vamos a utilizar.

```
library(dplyr)
library(emmeans)
library(ez)
library(ggpubr)
library(nlme)
```

Luego, obtengamos la muestra de datos (desde el archivo disponible para el ejercicio práctico anterior) que debemos utilizar.

```
src_dir <- "~/Downloads"
src_basename <- "EP06 Datos.csv"
src_file <- file.path(src_dir, src_basename)
datos <- read.csv(file = src_file, stringsAsFactors = TRUE)
```

Y seleccionemos datos de interés, aprovechando de especificar el orden deseado de los niveles del factor (que por defecto, R ordena alfabéticamente).

```
# Los datos ya vienen en formato largo
datos_largos <- datos |>
  filter(area == "Matemáticas") |>
  select(id, dificultad, tiempo) |>
  droplevels()
datos_largos[["id"]] <- factor(datos_largos[["id"]])
datos_largos[["dificultad"]] <- factor(datos_largos[["dificultad"]],
                                       levels = c("Baja", "Media", "Alta"))

# Mostramos las primeras filas para comprobar que todo va bien
head(datos_largos)
```

```
  id dificultad tiempo
1 17       Baja     93
2 17      Media     97
3 17       Alta     92
4 24       Baja     67
5 24      Media     95
6 24       Alta     91
```

Procedemos a verificar las condiciones para asegurar que podemos aplicar el procedimiento para muestras correlacionadas con validez.

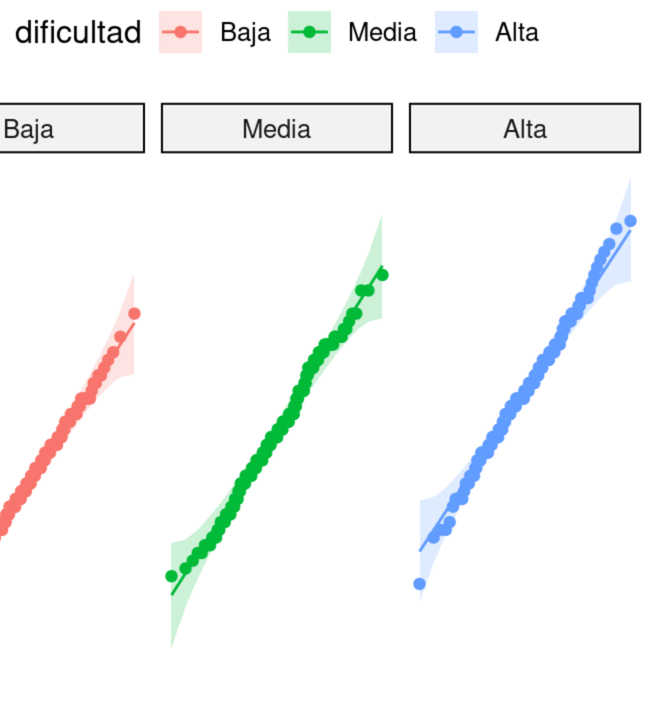
La variable dependiente corresponde a tiempo que, como vimos, se mide en una escala continua de intervalos iguales.

Por otro lado, los tríos de observaciones son independientes entre sí, pues provienen de personas diferentes que fueron elegidos de manera aleatoria.

Revisemos ahora la condición de normalidad por medio de un gráfico Q-Q.

```
g <- ggqqplot(datos_largos, x = "tiempo", y = "dificultad",
              color = "dificultad")

g <- g + facet_wrap(~ dificultad)
g <- g + rremove("x.ticks") + rremove("x.text")
g <- g + rremove("y.ticks") + rremove("y.text")
g <- g + rremove("axis.title")
print(g)
```



El gráfico sugiere que los datos siguen una distribución cercana a la normal, puesto que se encuentran dentro de la región aceptable del gráfico Q-Q y no se observan patrones no aleatorios, aunque se observa cierta desviación en el extremo superior de las preguntas con dificultad alta. Conviene entonces que usemos pruebas de normalidad para confirmar.

```
# Realizar el test de Shapiro-test para cada país
tests_normalidad <- by(datos_largos[["tiempo"]],
                       datos_largos[["dificultad"]],
                       shapiro.test)
print(tests_normalidad)
```

```
datos_largos[["dificultad"]]: Baja

    Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.99652, p-value = 0.9344

------------------------------------------------------------
datos_largos[["dificultad"]]: Media

    Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.99093, p-value = 0.2436

------------------------------------------------------------
datos_largos[["dificultad"]]: Alta

    Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.99442, p-value = 0.6629
```

Vemos que estas pruebas de Shapiro-Wilk descartan que debamos temer que alguna de estas muestras no provenga de una población con una distribución normal.

En cuanto a la condición de esfericidad, se posterga su discusión hasta ver el resultado de la prueba de Mauchly efectuada por `ezAnova()` .

Así, vamos a proceder con el procedimiento ANOVA para muestras correlacionadas considerando un nivel de significación de 0,05.

```
alfa <- 0.05

omnibus <- ezANOVA(
  data = datos_largos,
  dv = tiempo, within = dificultad, wid = id,
  return_aov = TRUE
)
```

Veamos el resultado del procedimiento por pantalla.

```
cat("Resultado de la prueba de Mauchly:\n\n")
print(omnibus[2])
cat("Resultado de la prueba ANOVA:\n\n")
print(omnibus[1])
cat("Tabla ANOVA tradicional:\n")
print(summary(omnibus[["aov"]]))
```

```
Resultado de la prueba de Mauchly:

$`Mauchly's Test for Sphericity`
      Effect         W         p p<.05
2 dificultad 0.9849307 0.2224141

Resultado de la prueba ANOVA:

$ANOVA
      Effect DFn DFd        F            p p<.05       ges
2 dificultad   2 398 114.8477 4.213847e-40     * 0.2827608

Tabla ANOVA tradicional:

Error: id
           Df Sum Sq Mean Sq F value Pr(>F)
Residuals 199  11233   56.45

Error: id:dificultad
            Df Sum Sq Mean Sq F value Pr(>F)
dificultad   2  13974    6987   114.8 <2e-16 ***
Residuals  398  24214      61
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Podemos ver que la prueba de esfericidad de Mauchly resulta no significativa con 97,5% de confianza ($W = 0{,}985$; $p = 0{,}222$), por lo que se falla en rechazar la hipótesis nula de esta prueba. Así, debemos concluir que **no hay suficiente evidencia estadística para descartar** que se cumple la condición de esfericidad en estos datos.

Interpretemos este resultado ómnibus.

> El procedimiento ANOVA correlacionado resultó significativo ($F(2, 398) = 114{,}848$; $p < 0{,}001$). En consecuencia, con 95% de confianza, rechazamos la hipótesis nula en favor de la hipótesis alternativa y concluimos que hay diferencias en el tiempo requerido por las mismas personas para formular consultas asociadas a un problema de información de en el área de las matemáticas con diferentes niveles de dificultad (baja, media y alta).

Puesto que el procedimiento ómnibus encuentra diferencias estadísticamente significativas, es necesario realizar un procedimiento post-hoc. Puesto que no requerimos hacer contrastes adicionales, usaremos la prueba HSD de Tukey (haciendo uso de un modelo mixto y de la estimación de medias marginales, implementadas en `R` en los paquetes `nlme` y `emmeans` respectivamente).

```
mixto <- lme(tiempo ~ dificultad , data = datos, random = ~1 | id)
medias <- emmeans (mixto , "dificultad")
post_hoc <- pairs(medias , adjust = "tukey")
conf_int <- confint(post_hoc, level = 1 - alfa)

print(post_hoc)
print(conf_int)
```

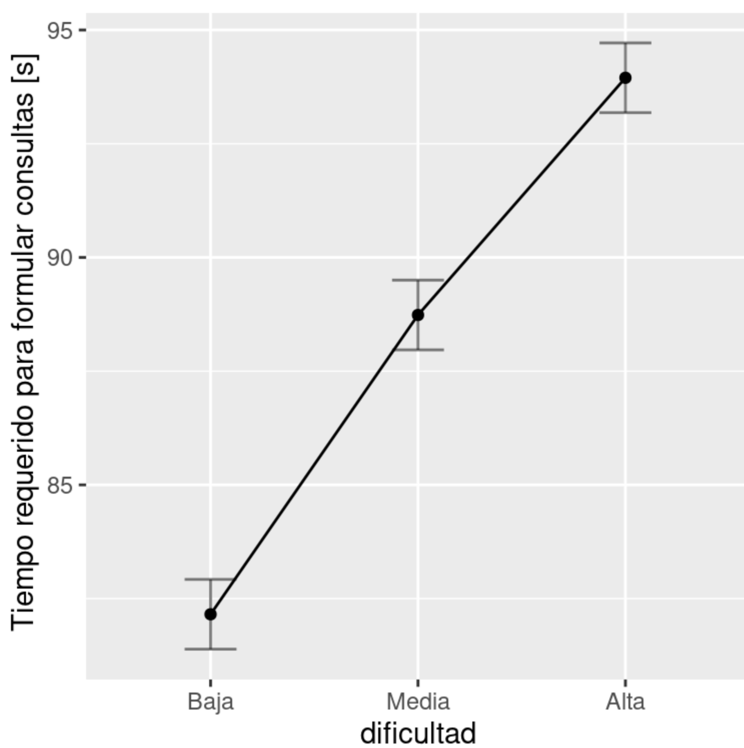
```
 contrast     estimate    SE   df t.ratio p.value
 Alta - Baja     6.244 0.238 4398  26.232  <.0001
 Alta - Media    5.338 0.238 4398  22.428  <.0001
 Baja - Media   -0.905 0.238 4398  -3.804  0.0004

Degrees-of-freedom method: containment
P value adjustment: tukey method for comparing a family of 3 estimates
 contrast     estimate    SE   df lower.CL upper.CL
 Alta - Baja     6.244 0.238 4398     5.69    6.802
 Alta - Media    5.338 0.238 4398     4.78    5.896
 Baja - Media   -0.905 0.238 4398    -1.46   -0.347

Degrees-of-freedom method: containment
Confidence level used: 0.95
Conf-level adjustment: tukey method for comparing a family of 3 estimates
```

Veamos si estos resultados coincide con el efecto (que tiene la variable independiente `dificultad` en la variable dependiente `tiempo`) encontrado en el procedimiento ANOVA para muestras correlacionadas.

```
g_efecto <- ezPlot(data = datos_largos, x = dificultad,
                   dv = tiempo, within = dificultad, wid = id,
                   y_lab = " Tiempo requerido para formular consultas [s]"
)
print(g_efecto)
```



Vemos que el gráfico del efecto coincide bien con los resultados de la prueba post-hoc. Redactemos la conclusión.

El análisis post-hoc usando el método de la diferencia honestamente significativa de Tukey indica que, en el área de las matemáticas, el tiempo requerido por una persona para formular consultas aumenta con el nivel de dificultad del problema de información (Alta-Baja: 95% CI: [5,69; 6,80], $t(4.398) = 26{,}232, p < 0{,}001$; Alta-Media: 95% CI: [4,78; 5,90], $t(4.398) = 22.428, p < 0{,}001$; Media-Baja: 95% CI: [0,35; 1,46], $t(4.398) = 3.804, p < 0{,}001$).