EP06-grupo-7

2024-10-24

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
datos <- read.csv("EP06 Datos.csv")
datosE6 <- datos %>% filter(area=="Literatura")
```

Para este ejercicio se proponen las siguientes Hipótesis:

H_0: No existen diferencias significativas entre los promedios de tiempo que tardan los usuarios en formular consultas para problemas con diferente nivel de dificultad en el área de literatura.

H_A: Existe al menos un promedio de tiempo de una dificultad distinto al resto.

Matematicamente:

$H_0 : \mu_A = \mu_M = \mu_B$

$H_A : \exists i, j \in Alta, Media, Baja, i \neq j \mid \mu_i \neq \mu_j$

Para poder aplicar el test ANOVA de variables correlacionadas debemos verificar las siguientes condiciones

1. La escala con que se mide la variabale dependiente tiene las propiedades de una escala de intervalos iguales.
2. Las mediciones son independientes al interior de cada grupo.
3. Se puede suponer razonablemente que las poblaciones de origen siguen una distribucion normal.
4. La matriz de varianza-covarianzas es esférica.

Para la condición n°1, vemos que esto sí se cumple, ya que la escala del tiempo esta en segundos, y la misma está en escala de intervalos, es más, sigue una escala de razón.

Para la condición n°2 se procede a calcular los gráficos QQ.

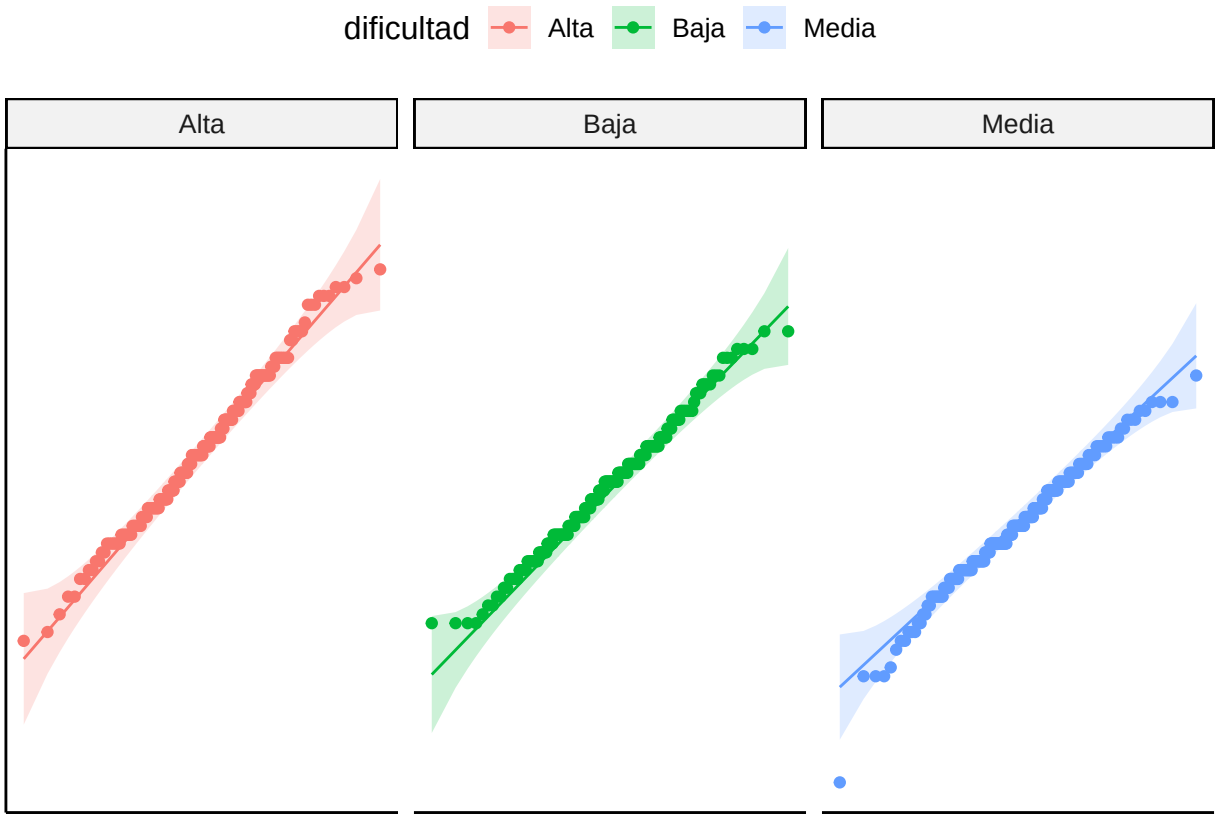
```
library(tidyverse)
library(ggpubr)
datosE61 <- datosE6
```

```r
datosE61[["dificultad"]] <- factor(datosE61[["dificultad"]])

datosE61[["id"]] <- factor(1:nrow(datosE61))

g <- ggqqplot(datosE61,
              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")
g
```



Como se puede observar en el grafico QQ, existen algunos valores atípicos en la dificultad media, por lo tanto trabajaremos con un $\alpha = 0.025$

Con respecto a la tercera condición, esta se cumple, debido a que, en el enunciado, se dice que cada voluntario fue asignado de manera aleatoria en cada grupo.

Para la cuarta condicón, se debe verificar la esfericidad con el test de Mauchly, generado por ezANOVA.

```r
library(ez)
library(nlme)
```

```
##
```

```
## Adjuntando el paquete: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
library(emmeans)
```

```
## Welcome to emmeans.
## Caution: You lose important information if you filter this package's results.
## See '? untidy'
```

```
prueba <- ezANOVA(data = datosE6, dv = tiempo, within= dificultad,
                  wid = id, return_aov = TRUE)

print(prueba[["Mauchly's Test for Sphericity"]])
```

```
##       Effect         W          p p<.05
## 2 dificultad 0.9736834 0.07134443
```

Como $p > 0.025$, se puede asegurar que los datos cumplen la condición de esfericidad.

Aplicacion de ANOVA

```
summary(prueba$aov)
```

```
##
## Error: id
##            Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 199  11081   55.68
##
## Error: id:dificultad
##             Df Sum Sq Mean Sq F value Pr(>F)
## dificultad   2   7569    3784   68.24 <2e-16 ***
## Residuals  398  22072      55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como $p < 0.025$ se rechaza la hipótesis nula en favor a la hipótesis alternativa. Por lo tanto, se puede asegurar con un 97.5% de confianza que existe al menos un promedio de tiempo distinto al resto.

Como la prueba ANOVA nos arrojo que existe una difrencia entre los promedios, procederemos a hacer una prueba post-hoc para verificar dónde se encuentra esta diferencia.

```
mixto <- lme(tiempo ~ dificultad, data = datosE6, random = ~1 | id)
medias <- emmeans(mixto, "dificultad")
tukey <- pairs(medias, adjust = "tukey")

print(tukey)
```

```
##  contrast     estimate    SE  df t.ratio p.value
##  Alta - Baja      4.09 0.745 398   5.499  <.0001
##  Alta - Media     8.70 0.745 398  11.676  <.0001
##  Baja - Media     4.60 0.745 398   6.177  <.0001
##
## Degrees-of-freedom method: containment
## P value adjustment: tukey method for comparing a family of 3 estimates
```

Luego de la realización de la prueba post-hoc HSD de Tukey, se puede afirmar con un 99% de confianza que todos los promedios de tiempos son distintos.