Taller - 1 ANOVA

Tomas Montealegre

2022-05-18

```
library(ggplot2)
library(readr)
library(nortest)
```

Taller 1 - ANOVA

- Tomas Montealegre

Punto 1

Los miembros de un equipo de fútbol se dividen al azar en tres grupos que realizan su plan de entrenamiento con métodos diferentes para mejorar su rendimiento físico. El primer grupo entrena con sesiones largas de carrera de resistencia, el segundo grupo se entrena series cortas de alta intensidad y el tercero hace trabajo de fuerza en el gimnasio. Después de dos meses de entrenamiento se realiza un test de rendimiento en carrera de 3km. Los tiempos de cada grupo fueron los siguientes:

Metodo 1	Metodo 2	Metodo 3
15	14	13
16	13	12
14	15	11
15	16	14
17	14	11

A un nivel de confianza del 95% ¿Puede considerarse que los tres métodos producen resultados equivalentes?

Suma de cuadrados

Paso 0 Creación de variables

```
metodo_1 <- c(15, 16, 14, 15, 17)
metodo_2 <- c(14, 13, 15, 16, 14)
metodo_3 <- c(13, 12, 11, 14, 11)
obs <- c(metodo_1, metodo_2, metodo_3)
num_g <- 3
```

Paso 1 Calculo de sumas por grupo y total.

```r
sum_1 <- sum(metodo_1)
sum_2 <- sum(metodo_2)
sum_3 <- sum(metodo_3)
sum_T <- sum_1 + sum_2 + sum_3 #T
```

Paso 2 Calculo de la suma total al cuadrado sobre el numero de obvservaciones.

```r
sum_Tcn <- (sum_T^2)/length(obs) #T^2/n
```

Paso 3 Calculo de la suma por grupo al cuadrado sobre el numero de obvservaciones y suma total de estos.

```r
#sum 1
sum_1cn <- (sum_1^2)/length(metodo_1)
#sum 2
sum_2cn <- (sum_2^2)/length(metodo_2)
#sum 3
sum_3cn <- (sum_3^2)/length(metodo_3)
#suma total
sum_tcn <- sum_1cn + sum_2cn + sum_3cn
```

Paso 4 Calculo de sumas cuadradas por grupo y suma total de estas.

```r
# Metodo 1
sum_c1 <- 0
for (i in metodo_1){
  sum_c1 = sum_c1 + (i^2)
}
# Metodo 2
sum_c2 <- 0
for (i in metodo_2){
  sum_c2 = sum_c2 + (i^2)
}
# Metodo 3
sum_c3 <- 0
for (i in metodo_3){
  sum_c3 = sum_c3 + (i^2)
}
# Total
sum_ct <- sum_c1 + sum_c2 + sum_c3
```

Paso 5 Calculo de las sumas cuadradas total, de tratamientos y de error muestral.

```r
SST <- sum_ct - sum_Tcn #SST: Suma cuadrada total.
SStr <- sum_tcn - sum_Tcn #SStr: Suma cuadrada tratamientos.
SSE <- SST - SStr #SSE: Suma cuadrada error muestral.
```

Paso 6 Calculo de las medias cuadradas de tratamientos y de error muestral.

```
MStr <- SStr/(num_g - 1) #MStr: Media cuadrada tratamientos.
MSE <- SSE/(length(obs)-num_g) #MSE: Media cuadrada error muestral.
```

Paso 7 Calculo de F.

```
f <- MStr/MSE
f
```

```
## [1] 9.348837
```

Librerias

```
tiempo <- c(15, 16, 14, 15, 17, 14, 13, 15, 16, 14, 13, 12, 11, 14, 11)
metodo <- c("metodo_1", "metodo_1", "metodo_1", "metodo_1", "metodo_1", "metodo_2", "metodo_2", "metodo_

datos <- data.frame(metodo = metodo, tiempo = tiempo)
```

```
anova <- aov(datos$tiempo ~ datos$metodo)
summary(anova)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## datos$metodo  2   26.8  13.400   9.349 0.00357 **
## Residuals    12   17.2   1.433
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test de hipótesis

El valor de F teórica en (2,12) es de 3.89 y el valor obtenido es de 9.348837, por lo tanto se rechaza la hipótesis nula de que los tres métodos producen resultados equivalentes y se puede afirmar a un nivel de confianza del 95% que existe al menos un método que produce resultados diferentes a los del los demás.

Mediante librerias el p-valor es igual a 0.00357 que es menor a 0.05 por lo tanto se rechaza la hipotesis nula.

Punto 2

El dataset "students" consiste en las calificaciones obtenidas por los estudiantes en varias materias. Se recogieron las siguientes variables:

- Gender: Sexo del estudiante
- Race/ethnicity: Etnia del estudiante
- Parental level of education: Nivel de educaci´on del padre y la madre
- Lunch: Tipo de almuerzo
- Test preparation course: Curso de preparaci´on completado o no completado
- Nota de matemáticas
- Nota de lectura
- Nota de escritura

Cargar dataset

```
students <- read_csv("D:/001_Maestria/Semestre_01/Estadistica/Talleres/Taller 1 - ANOVA/students.csv")
```

```
## Rows: 1000 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): gender, race/ethnicity, parental level of education, lunch, test pr...
## dbl (3): math score, reading score, writing score
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Utilizando el dataset de "students" se quiere contestar a las siguientes preguntas:

Pregunta 1

¿Hay diferencias significativas en la nota media de matemáticas entre hombres y mujeres?

Matemáticas Comenzamos con la validación de las condiciones de normalidad y homocedasticidad.

```
lillie.test(students$`math score`[students$gender=='male'])
```

Normalidad de los datos

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$`math score`[students$gender == "male"]
## D = 0.038781, p-value = 0.08004
```

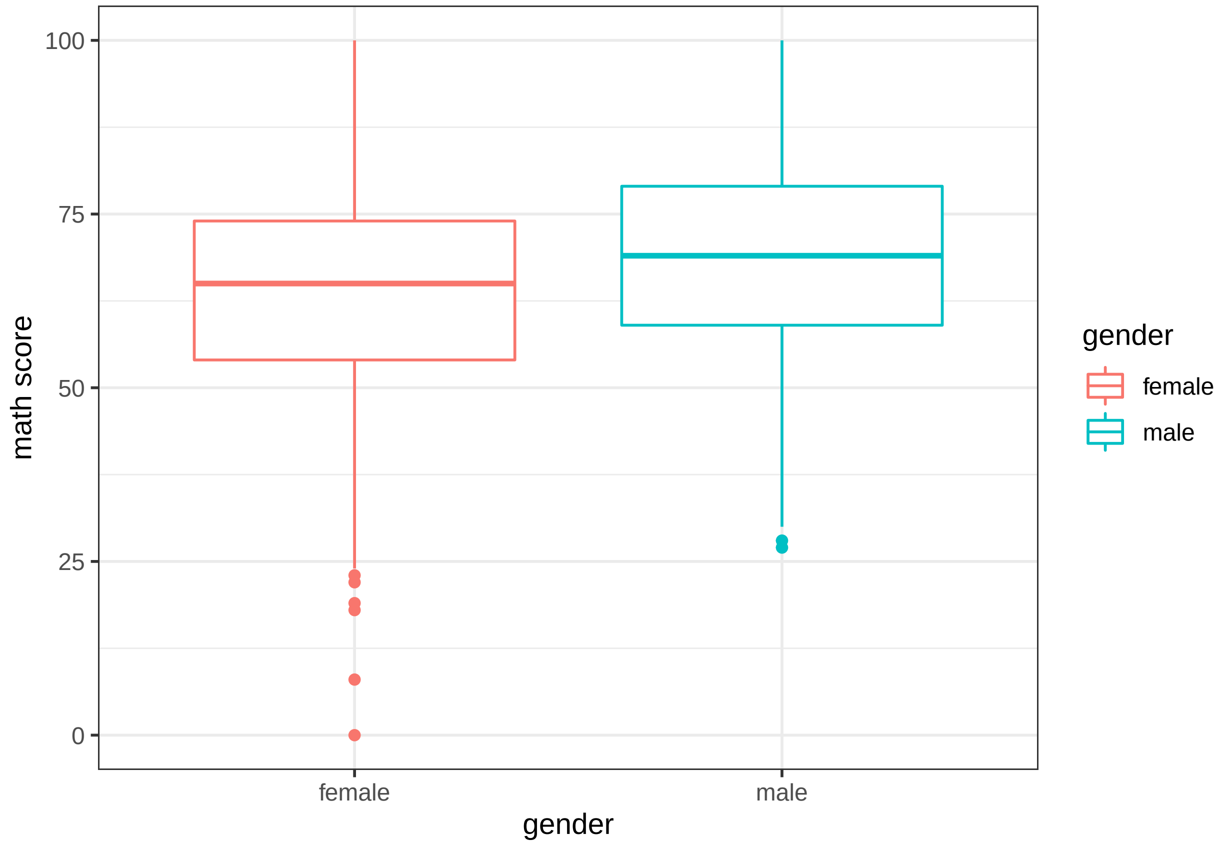
```
lillie.test(students$`math score`[students$gender=='female'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$`math score`[students$gender == "female"]
## D = 0.043394, p-value = 0.02107
```

Los datos de la categoría "female" presentan un p-valor de $0.02107 < 0.05$, no se confirma la normalidad de los datos.

Se grafican los datos en un boxplot en búsqueda de outliners.

```
ggplot(data = students, aes(x = gender, y = `math score`, color = gender)) +
    geom_boxplot() + theme_bw()
```



Se identifican como outliners los datos < 25 y se eliminan 7 datos.

```
students_hm<-students[students$`math score`>=25,]
```

Se realizan la prueba de normalidad despues de quitar los outliners

```
lillie.test(students_hm$`math score`[students_hm$gender=='male'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students_hm$`math score`[students_hm$gender == "male"]
## D = 0.038781, p-value = 0.08004
```

```
lillie.test(students_hm$`math score`[students_hm$gender=='female'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students_hm$`math score`[students_hm$gender == "female"]
## D = 0.030473, p-value = 0.2972
```

Se valida normalidad en los datos, p-valor > 0.05.

```
fligner.test(students_hm$`math score` ~ students_hm$gender,students)
```

Homocedasticidad (varianza constante entre grupos)

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  students_hm$‘math score‘ by students_hm$gender
## Fligner-Killeen:med chi-squared = 0.046273, df = 1, p-value = 0.8297
```

Se valida homocedasticidad en los datos, p-valor > 0.05.

ANOVA Se realiza el análisis ANOVA después de validar que las condiciones se cumplan.

```
anova_hm <- aov(students_hm$`math score` ~ students_hm$gender)
summary(anova_hm)
```

```
##                      Df Sum Sq Mean Sq F value   Pr(>F)
## students_hm$gender    1   4904    4904    23.5 1.45e-06 ***
## Residuals           991 206819     209
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

p-valor 1.45e-06 < 0.05, por lo tanto se rechaza la hipotesis nula, se afirma que hay diferencias significativas en la nota media de matemáticas entre hombres y mujeres.

Pregunta 2

¿Hay diferencias significativas en la media para alguna de las notas (individualmente) entre el nivel de educación parental?

Matemáticas

```
lillie.test(students$`math score`[students$`parental level of education`=="master's degree"])
```

Normalidad de los datos

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$‘math score‘[students$‘parental level of education‘ ==     "master’s degree"]
## D = 0.1246, p-value = 0.02331
```

```r
lillie.test(students$`math score`[students$`parental level of education`=="bachelor's degree"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$`math score`[students$`parental level of education` ==     "bachelor's degree"]
## D = 0.05844, p-value = 0.4149
```

```r
lillie.test(students$`math score`[students$`parental level of education`=="associate's degree"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$`math score`[students$`parental level of education` ==     "associate's degree"]
## D = 0.058941, p-value = 0.05887
```

```r
lillie.test(students$`math score`[students$`parental level of education`=='some college'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$`math score`[students$`parental level of education` ==     "some college"]
## D = 0.062697, p-value = 0.03118
```

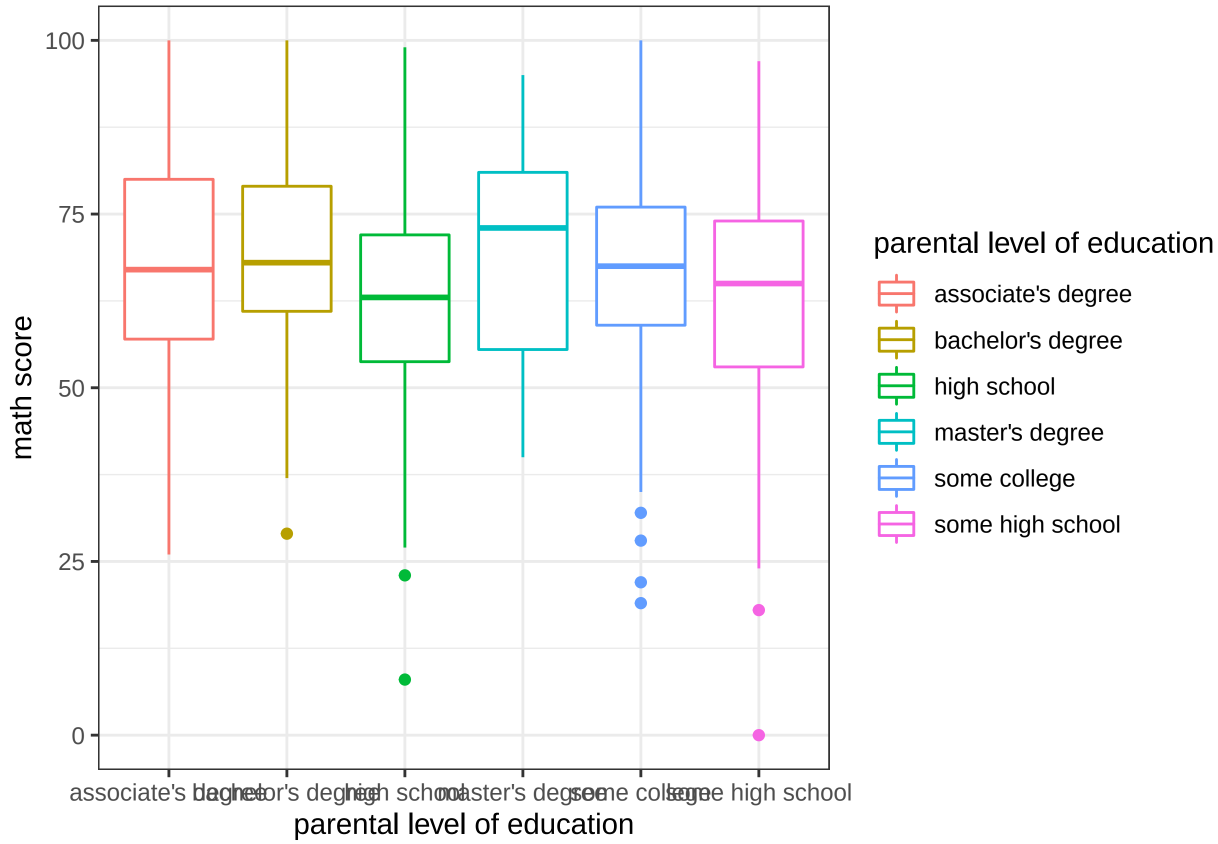
```r
lillie.test(students$`math score`[students$`parental level of education`=='high school'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$`math score`[students$`parental level of education` ==     "high school"]
## D = 0.06599, p-value = 0.03724
```

```r
lillie.test(students$`math score`[students$`parental level of education`=='some high school'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$`math score`[students$`parental level of education` ==     "some high school"]
## D = 0.081576, p-value = 0.005525
```

```r
ggplot(data = students, aes(x = `parental level of education`, y = `math score`, color = `parental level
    geom_boxplot() + theme_bw()
```




```
students_math<-students[students$`math score`>=30,]
```

```
lillie.test(students_math$`math score`[students_math$`parental level of education`=="master's degree"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students_math$`math score`[students_math$`parental level of education` ==     "master's degre
## D = 0.1246, p-value = 0.02331
```

```
lillie.test(students_math$`math score`[students_math$`parental level of education`=="bachelor's degree"]
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students_math$`math score`[students_math$`parental level of education` ==     "bachelor's deg
## D = 0.051806, p-value = 0.6172
```

```
lillie.test(students_math$`math score`[students_math$`parental level of education`=="associate's degree
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students_math$`math score`[students_math$`parental level of education` ==     "associate's deg
## D = 0.060103, p-value = 0.05085
```

```
lillie.test(students_math$`math score`[students_math$`parental level of education`=='some college'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students_math$‘math score‘[students_math$‘parental level of education‘ ==     "some college"]
## D = 0.049839, p-value = 0.194
```

```
lillie.test(students_math$`math score`[students_math$`parental level of education`=='high school'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students_math$‘math score‘[students_math$‘parental level of education‘ ==     "high school"]
## D = 0.046465, p-value = 0.3965
```

```
lillie.test(students_math$`math score`[students_math$`parental level of education`=='some high school'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students_math$‘math score‘[students_math$‘parental level of education‘ ==     "some high scho
## D = 0.054787, p-value = 0.2282
```

```
fligner.test(students$`math score` ~ students$`parental level of education`,students)
```

Homocedasticidad (varianza constante entre grupos)

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  students$‘math score‘ by students$‘parental level of education‘
## Fligner-Killeen:med chi-squared = 4.8993, df = 5, p-value = 0.4283
```

```
anova_hm <- aov(students$`math score` ~ students$`parental level of education`)
summary(anova_hm)
```

ANOVA

```
##                                         Df Sum Sq Mean Sq F value   Pr(>F)
## students$‘parental level of education‘   5   7296  1459.1   6.522 5.59e-06 ***
## Residuals                              994 222394   223.7
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Lectura

```
lillie.test(students$`reading score`[students$`parental level of education`=="master's degree"])
```

Normalidad de los datos

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$‘reading score‘[students$‘parental level of education‘ ==     "master’s degree"]
## D = 0.071291, p-value = 0.6427
```

```
lillie.test(students$`reading score`[students$`parental level of education`=="bachelor's degree"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$‘reading score‘[students$‘parental level of education‘ ==     "bachelor’s degree"]
## D = 0.062972, p-value = 0.3
```

```
lillie.test(students$`reading score`[students$`parental level of education`=="associate's degree"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$‘reading score‘[students$‘parental level of education‘ ==     "associate’s degree"]
## D = 0.059382, p-value = 0.0553
```

```
lillie.test(students$`reading score`[students$`parental level of education`=='some college'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$‘reading score‘[students$‘parental level of education‘ ==     "some college"]
## D = 0.051415, p-value = 0.154
```

```
lillie.test(students$`reading score`[students$`parental level of education`=='high school'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$‘reading score‘[students$‘parental level of education‘ ==     "high school"]
## D = 0.054308, p-value = 0.1701
```

```
lillie.test(students$`reading score`[students$`parental level of education`=='some high school'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$‘reading score‘[students$‘parental level of education‘ ==     "some high school"]
## D = 0.047623, p-value = 0.4127
```

```r
fligner.test(students$`reading score` ~ students$`parental level of education`,students)
```

Homocedasticidad (varianza constante entre grupos)

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  students$‘reading score‘ by students$‘parental level of education‘
## Fligner-Killeen:med chi-squared = 2.2205, df = 5, p-value = 0.8179
```

```r
anova_hm <- aov(students$`reading score` ~ students$`parental level of education`)
summary(anova_hm)
```

ANOVA

```
##                                         Df Sum Sq Mean Sq F value   Pr(>F)
## students$‘parental level of education‘   5   9506  1901.3   9.289 1.17e-08 ***
## Residuals                              994 203446   204.7
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Escritura

```r
lillie.test(students$`writing score`[students$`parental level of education`=="master's degree"])
```

Normalidad de los datos

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$‘writing score‘[students$‘parental level of education‘ ==     "master’s degree"]
## D = 0.079012, p-value = 0.477
```

```r
lillie.test(students$`writing score`[students$`parental level of education`=="bachelor's degree"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$‘writing score‘[students$‘parental level of education‘ ==     "bachelor’s degree"]
## D = 0.046884, p-value = 0.7574
```

```r
lillie.test(students$`writing score`[students$`parental level of education`=="associate's degree"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$`writing score`[students$`parental level of education` ==     "associate's degree"]
## D = 0.053846, p-value = 0.1207
```

```r
lillie.test(students$`writing score`[students$`parental level of education`=='some college'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$`writing score`[students$`parental level of education` ==     "some college"]
## D = 0.053277, p-value = 0.1221
```

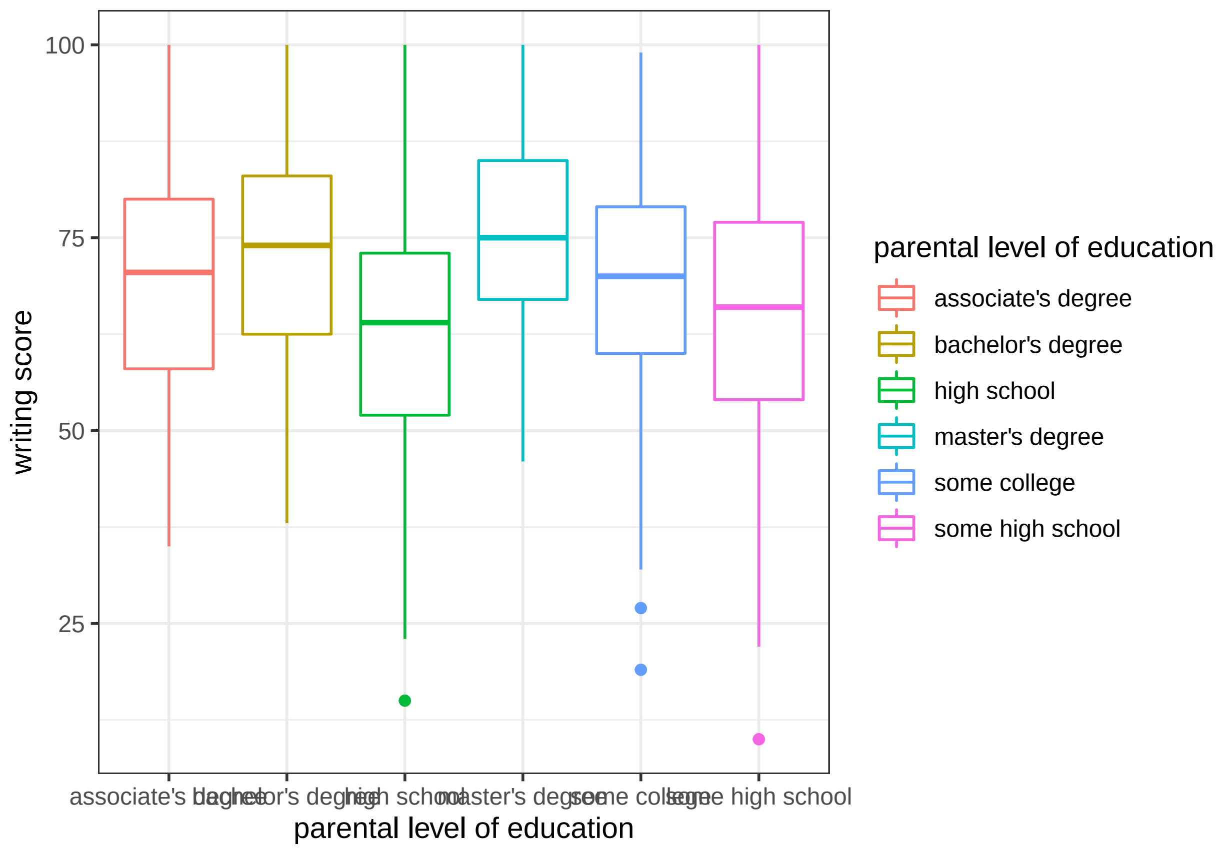
```r
lillie.test(students$`writing score`[students$`parental level of education`=='high school'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$`writing score`[students$`parental level of education` ==     "high school"]
## D = 0.064248, p-value = 0.04748
```

```r
lillie.test(students$`writing score`[students$`parental level of education`=='some high school'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students$`writing score`[students$`parental level of education` ==     "some high school"]
## D = 0.080533, p-value = 0.006573
```

```r
ggplot(data = students, aes(x = `parental level of education`, y = `writing score`, color = `parental le
    geom_boxplot() + theme_bw()
```




```
students_writing<-students[students$`writing score`>=30,]
```

```
lillie.test(students_writing$`writing score`[students_writing$`parental level of education`=="master's d
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students_writing$`writing score`[students_writing$`parental level of education` ==     "maste
## D = 0.079012, p-value = 0.477
```

```
lillie.test(students_writing$`writing score`[students_writing$`parental level of education`=="bachelor's
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students_writing$`writing score`[students_writing$`parental level of education` ==     "bache
## D = 0.046884, p-value = 0.7574
```

```
lillie.test(students_writing$`writing score`[students_writing$`parental level of education`=="associate
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students_writing$`writing score`[students_writing$`parental level of education` ==     "assoc
## D = 0.053846, p-value = 0.1207
```

```
lillie.test(students_writing$`writing score`[students_writing$`parental level of education`=='some coll
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students_writing$`writing score`[students_writing$`parental level of education` ==     "some c
## D = 0.047511, p-value = 0.2505
```

```
lillie.test(students_writing$`writing score`[students_writing$`parental level of education`=='high scho
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students_writing$`writing score`[students_writing$`parental level of education` ==     "high s
## D = 0.059555, p-value = 0.09248
```

```
lillie.test(students_writing$`writing score`[students_writing$`parental level of education`=='some high
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  students_writing$`writing score`[students_writing$`parental level of education` ==     "some h
## D = 0.084933, p-value = 0.003655
```

```
fligner.test(students$`writing score` ~ students$`parental level of education`,students)
```

Homocedasticidad (varianza constante entre grupos)

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  students$`writing score` by students$`parental level of education`
## Fligner-Killeen:med chi-squared = 3.2656, df = 5, p-value = 0.6591
```

```
anova_hm <- aov(students$`writing score` ~ students$`parental level of education`)
summary(anova_hm)
```

ANOVA

```
##                                          Df Sum Sq Mean Sq F value   Pr(>F)
## students$`parental level of education`   5  15623  3124.6   14.44 1.12e-13 ***
## Residuals                               994 215054   216.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```