

# Análisis de Varianza

Juan C. Correa

3/14/2021

El análisis de varianza es una de las técnicas más populares para evaluar la relación entre variables, a través de la comparación de tres o más grupos (e.g., de hecho, esa es la gran diferencia que tiene con técnicas como t de Student o U de Mann-Whitney que se especializan particularmente en la comparación entre dos grupos).

```
setwd("~/Documents/GitHub/Pantaleon")
library(readxl)
AOV <- read_excel("PantaleonDatos2.xls")
```

```
## New names:
## * TOTAL -> TOTAL...25
## * CONTAR -> CONTAR...26
## * edad -> edad...35
## * edad -> edad...41
## * TOTAL -> TOTAL...90
## * ...
```

En esta nueva base de datos hemos creado una variable más llamada “Entrenamiento” que puede adoptar los siguientes tres valores: Poco, Intermedio y Alto. Cuando se habla de “Poco Entrenamiento” la producción de esa finca no requiere una gran experiencia por parte de la gerencia, por tratarse de una producción bastante controlada. En cambio, cuando se habla de “Entrenamiento Alto” la producción de la finca necesita ser gerenciada por alguien con bastante experiencia para lograr dominar todas las dificultades asociadas con la producción en esa finca. Ahora, queremos ver si el entrenamiento se asocia o no con la producción en TCH, TAH. Acá vamos a concentrarnos en TCH y luego, usted, deberá proceder con el análisis de TAH.

```
resultado_TCH <- aov(AOV$TCH ~ AOV$Entrenamiento)
summary(resultado_TCH)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## AOV$Entrenamiento    2   34302    17151   41.71 <2e-16 ***
## Residuals          1527  627938     411
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los resultados muestran que, efectivamente TCH cambia sistemáticamente su comportamiento estadístico dependiendo del entrenamiento requerido por la finca ( $F = 41.71$ ;  $gl = (2, 1547)$ ;  $p = 2e-16$ ). Al generar el histograma de los residuales del modelo, se observa que el comportamiento de dichos residuales se asemeja a una curva normal, de modo que el uso de un modelo de análisis de varianza para los datos de TCH resulta ser adecuado. Sin embargo, una limitación del resultado de cálculo realizado es que la diferencia significativa observada aún no queda completamente evidenciada dada las múltiples comparaciones que realiza un análisis de varianza. En este caso concreto, por ejemplo, se hacen las siguientes comparaciones: 1) Entrenamiento Poco versus Entrenamiento Intermedio, 2) Entrenamiento Poco versus Entrenamiento Alto, 3) Entrenamiento Intermedio versus Entrenamiento Alto.

Al ejecutar un análisis de varianza, entonces, se hace fundamental conocer cuál es el origen de las diferencias, dado que el anova es sensible incluso cuando no hay diferencias entre pares de grupos, y solo se observa una

diferencia entre al menos un par de grupos. Para ello, se requiere complementar los cálculos con lo que se llama comparación pareada o de grupos.

```
Entrenamiento <- AOV$Entrenamiento
tapply(AOV$TCH, Entrenamiento, mean)

##      Alto Intermedio      Poco
## 92.00134 103.82373 109.77311

tapply(AOV$TCH, Entrenamiento, sd)

##      Alto Intermedio      Poco
## 16.95219 20.37895 23.00351

pairwise.t.test(AOV$TCH, Entrenamiento, p.adj = "none")

##
## Pairwise comparisons using t tests with pooled SD
##
## data:  AOV$TCH and Entrenamiento
##
##           Alto      Intermedio
## Intermedio 1.7e-14 -
## Poco      < 2e-16 0.00026
##
## P value adjustment method: none
```

Al ejecutar estas comparaciones pareadas, observamos que las diferencias son significativas para cualquier par comparado, pero la significancia estadística cambia según el par comparado. Por ejemplo, cuando se compara Entrenamiento Poco con Intermedio su significancia estadística muestra un valor significativo aunque ligeramente más alto ( $p = 0.00026$ ) que la significancia de comparar Entrenamiento Alto con Intermedio ( $P = 1.7e-14$ ) o Alto con Poco ( $p < 2e-16$ ). Un detalle que debe considerarse de la sintaxis anterior es que en el argumento `p.adj` se establece “none” o ninguno para introducir algún tipo de ajuste en el cálculo de las probabilidades. Esta elección, sin embargo, puede llevarnos a un problema que se conoce como error tipo 1 en estadística. El método de Bonferroni-Holm se utiliza para contrarrestar el problema de no controlar la tasa de error familiar (es decir, la probabilidad de realizar uno o más descubrimientos falsos). Para controlar ese problema, podríamos simplemente introducir una pequeña modificación a la sintaxis anterior.

```
pairwise.t.test(AOV$TCH, Entrenamiento, p.adj = "bonf")

##
## Pairwise comparisons using t tests with pooled SD
##
## data:  AOV$TCH and Entrenamiento
##
##           Alto      Intermedio
## Intermedio 5.1e-14 -
## Poco      < 2e-16 0.00079
##
## P value adjustment method: bonferroni
```

Obsérvese que ahora, las probabilidades de la significancia estadística en la comparación por pares han cambiado ligeramente. Una muy buena práctica es usar el método de Bonferroni-Holm.

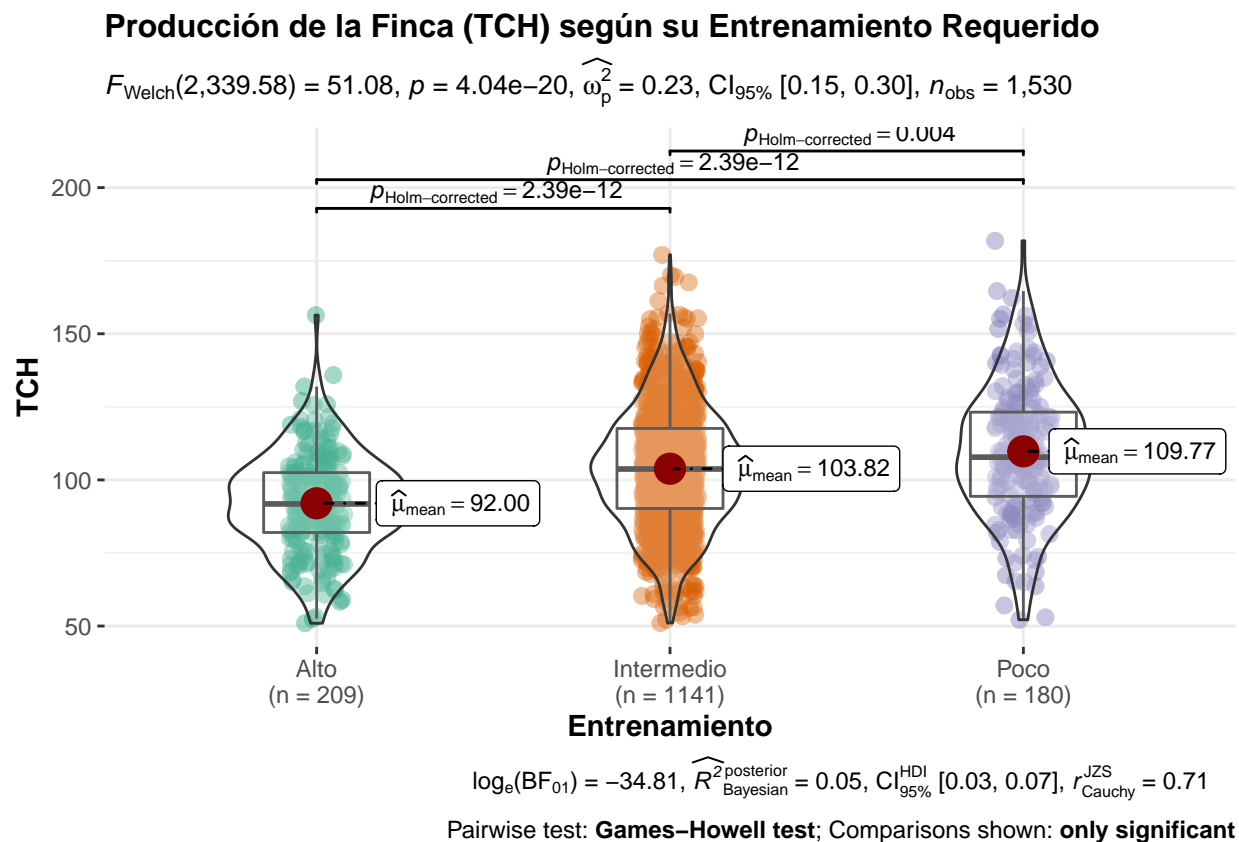
Más allá del cálculo, y tal como hemos sugerido en sesiones anteriores, siempre es recomendable acompañar los cálculos estadísticos con gráficos que ilustren o complementen la información que se obtiene con los mismos cálculos. Todo lo que hemos realizado anteriormente puede evidenciarse con el siguiente gráfico realizado con la librería `ggstatplot`.

```
library(ggstatsplot)

## Registered S3 methods overwritten by 'lme4':
##   method                                  from
##   cooks.distance.influence.merMod         car
##   influence.merMod                        car
##   dfbeta.influence.merMod                 car
##   dfbetas.influence.merMod                car

## In case you would like cite this package, cite it as:
##   Patil, I. (2018). ggstatsplot: "ggplot2" Based Plots with Statistical Details. CRAN.
##   Retrieved from https://cran.r-project.org/web/packages/ggstatsplot/index.html

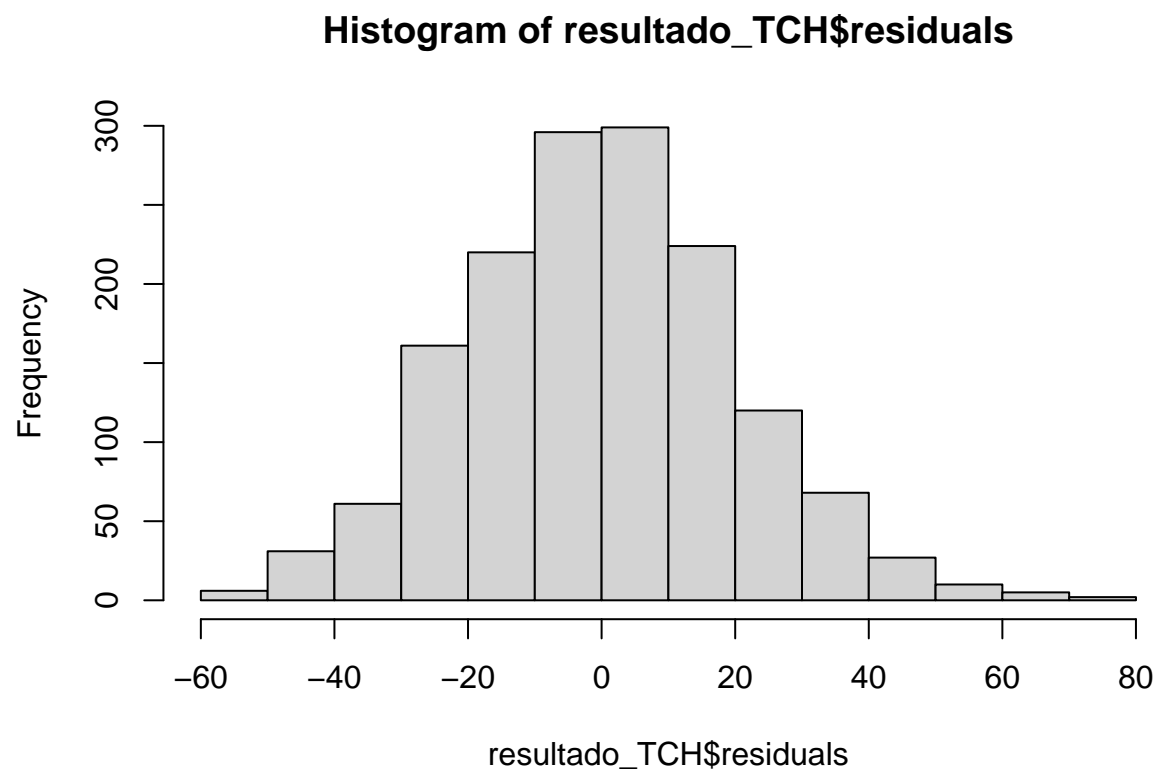
ggbetweenstats(
  data = AOV,
  x = Entrenamiento,
  y = TCH,
  title = "Producción de la Finca (TCH) según su Entrenamiento Requerido"
)
```



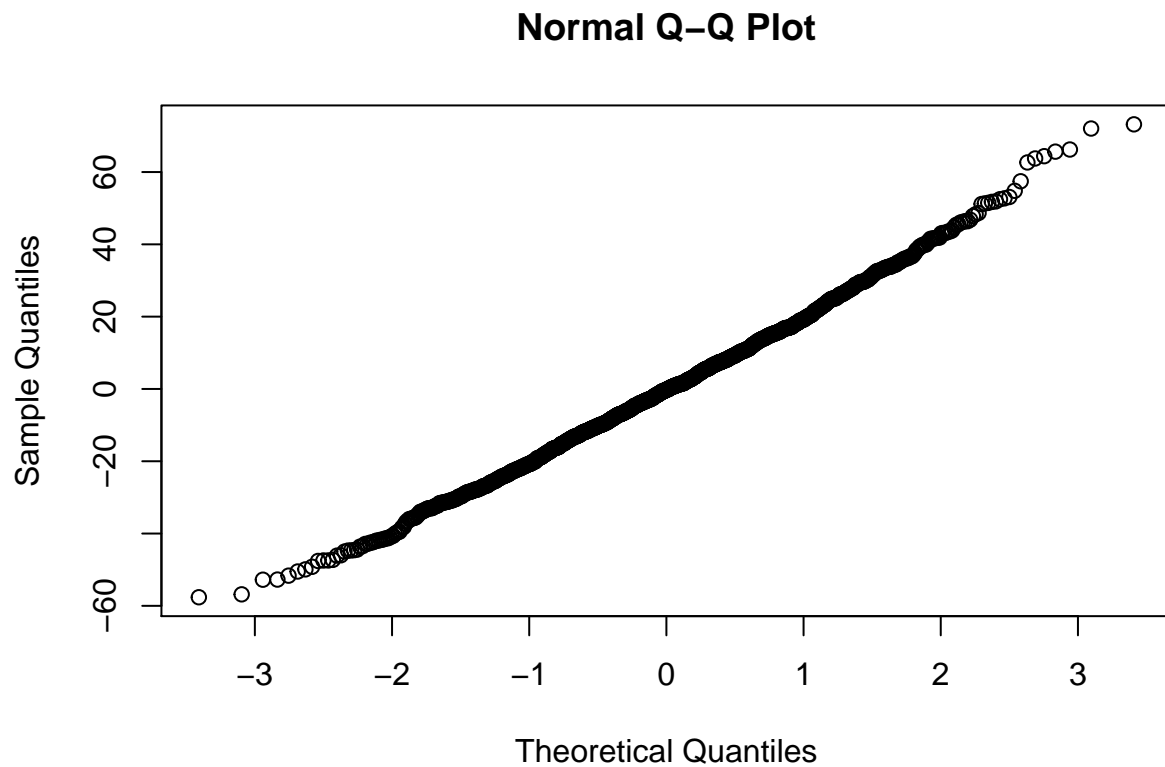
## Chequeo de supuestos en los que se apoya el ANOVA

Tal y como se ha insistido en sesiones anteriores, en estadística casi todas las técnicas se apoyan en un conjunto de supuestos o condiciones que deben chequearse para que los análisis no sean sesgados. Un supuesto sencillo de verificar es a través de la observación de los residuales del modelo (parte del modelo que encaja o se ajusta a los datos observados).

```
hist(resultado_TCH$residuals)
```



```
qqnorm(resultado_TCH$residuals)
```



Recuerde que también pueden calcularse los estadísticos de asimetría y curtosis para verificar que la distribución observada de los residuales no presenta una grotesca desviación de la curva normal (distribución subyacente al cálculo del ANOVA).

```
library(e1071)
skewness(resultado_TCH$residuals)
```

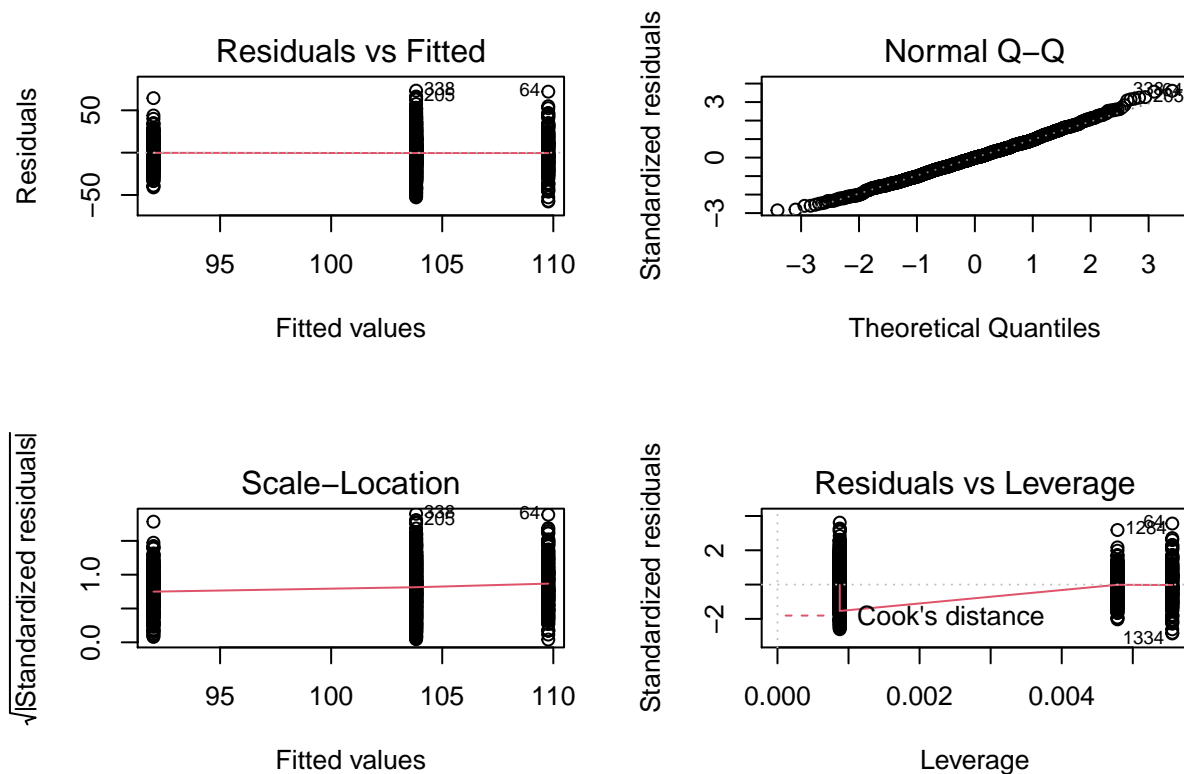
```
## [1] 0.1870317
```

```
kurtosis(resultado_TCH$residuals)
```

```
## [1] 0.1296595
```

También podría evaluar el cumplimiento de la homoscedasticidad de la siguiente manera

```
par(mfrow=c(2,2))
plot(resultado_TCH)
```



```
par(mfrow=c(1,1))
```

Cada gráfico proporciona una información específica sobre el ajuste del modelo, pero es suficiente saber que la línea roja que representa la media de los residuos debe ser horizontal y centrada en cero (o en uno, en el gráfico de “ubicación de escala” location shape), lo que significa que no hay grandes valores atípicos que causen sesgos en el modelo.

La gráfica Q-Q normal traza una regresión entre los residuos teóricos de un modelo perfectamente heterocedástico y los residuos reales de su modelo, por lo que cuanto más cerca de una pendiente de 1, mejor. Como hemos visto, los valores de asimetría y kurtosis no son adecuados y además los gráficos parecen sugerir la violación al supuesto de homoscedasticidad. Una forma de verificar definitivamente eso es a través de la prueba Breusch-Pagan con la función `bptest` de la librería `lmtest` o la prueba de varianza de error no constante con la función `ncvTest` de la librería `car`.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
```

```
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
## as.Date, as.Date.numeric
```

```
bptest(resultado_TCH)
```

```
##
```

```
## studentized Breusch-Pagan test
```

```
##
## data: resultado_TCH
## BP = 15.782, df = 2, p-value = 0.000374
```

```
library(car)
```

```
## Loading required package: carData
```

```
ncvTest(lm(AOV$TCH ~ AOV$Entrenamiento))
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 16.10725, Df = 1, p = 5.9854e-05
```

La prueba Breusch-Pagan nos arroja como resultado un estadístico significativo (BP = 15.782, df = 2, p = 0.000374) y la prueba de varianza de error no constante también nos muestra un estadístico significativo ( $\chi^2 = 16.10725$ , df = 1, p = 5.98e-05) que en conjunto sugiere que el modelo está arrojando resultados sesgados debido a que el supuesto de homoscedasticidad no se cumple para este caso, lo cual nos introduce en la necesidad de correr otros tipos de modelos que corrijan estos sesgos. Una alternativa pudiera ser la siguiente

```
library(WRS2)
tiway(AOV$TCH ~ AOV$Entrenamiento)
```

```
## Call:
## tiway(formula = AOV$TCH ~ AOV$Entrenamiento)
##
## Test statistic: F = 44.5781
## Degrees of freedom 1: 2
## Degrees of freedom 2: 204.91
## p-value: 0
##
## Explanatory measure of effect size: 0.43
## Bootstrap CI: [0.34; 0.51]
```

Desde luego, uno como analista también pudiera decidir usar una técnica alternativa no-paramétrica, como por ejemplo la prueba Kruskal-Wallis que aplica cuando se desea evaluar si las muestras se originan a partir de la misma distribución. Dicho de otra manera, la presencia de diferencia significativas sería el indicador de que las muestras a comparar no provienen de una misma distribución.

```
kruskal.test(AOV$TCH ~ AOV$Entrenamiento)
```

```
##
## Kruskal-Wallis rank sum test
##
## data: AOV$TCH by AOV$Entrenamiento
## Kruskal-Wallis chi-squared = 81.046, df = 2, p-value < 2.2e-16
pairwise.wilcox.test(AOV$TCH, AOV$Entrenamiento, p.adjust.method = "BH")
```

```
##
## Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data: AOV$TCH and AOV$Entrenamiento
##
##           Alto      Intermedio
## Intermedio 4.1e-15 -
## Poco       3.1e-15 0.00086
##
## P value adjustment method: BH
```

La limitación de aplicar pruebas no-paramétricas es que aún cuando sus cálculos se apoyen en suposiciones menos estrictas (no se asume distribución normal para la estimación de estadísticos que indiquen la asociación entre variables), dicho tratamiento no resuelve el problema que se observa con las pruebas que chequean si se violan o no el supuesto de homoscedasticidad.