Herramientas de Estadística (No)Paramétrica

(Parte 4)

Juan C. Correa

Material de uso exclusivo para
INGENIO PANTALEON, S.A.
Diagonal 6, 10-31, Zona 10

Ciudad de Guatemala

Agenda

1. Comparación entre tres o más grupos
2. Consideraciones Conceptuales
3. Contextualización
4. Chequeo de Supuestos

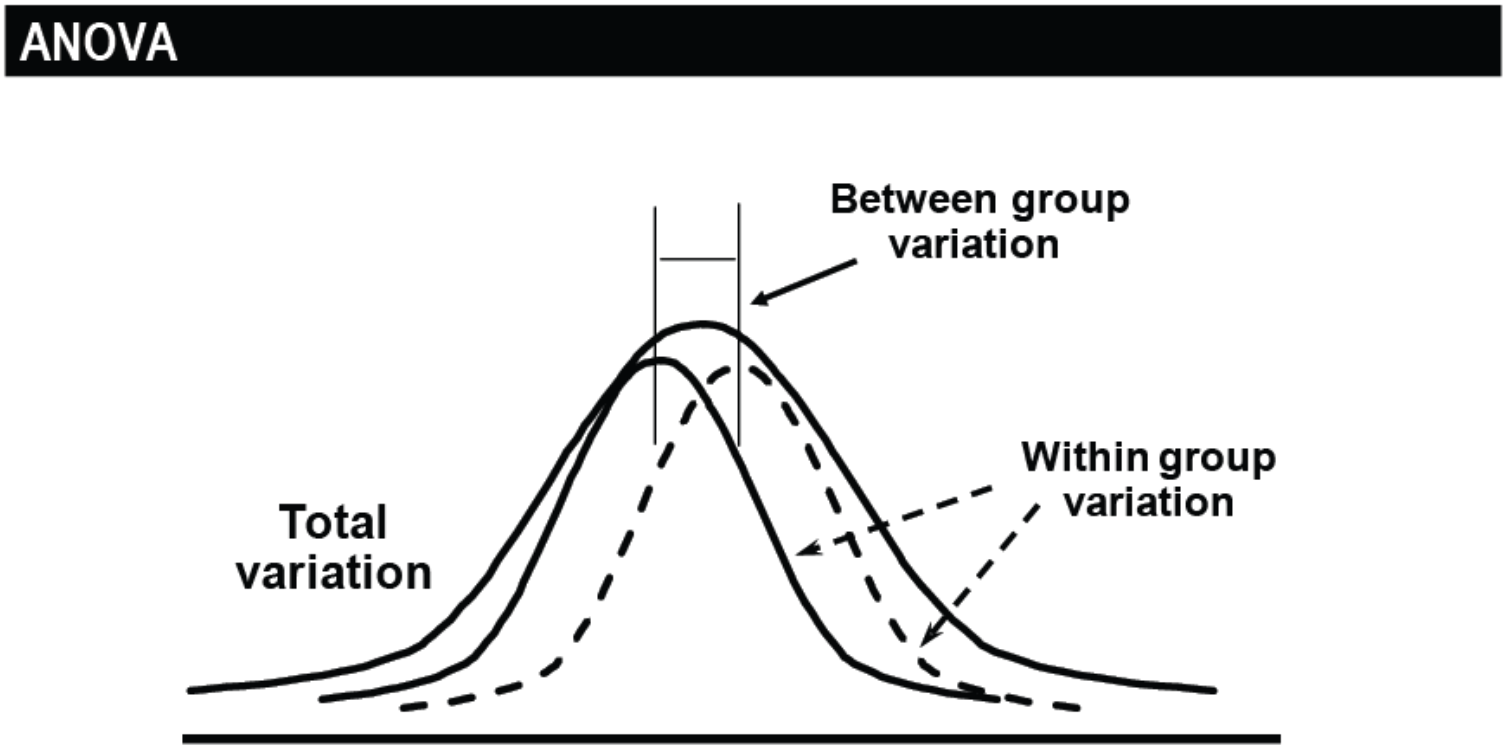
Comparación entre tres o más grupos

En Estadística (No)Paramétrica Parte 3 vimos la comparación entre dos grupos como una de las maneras más sencillas de establecer la relación entre variables, siguiendo ciertas consideraciones conceptuales que aquí seguirán manteniéndose como válidas.

Ahora, vamos a concentrar nuestra atención a la comparación entre tres o más grupos, haciendo uso de una de las técnicas más populares llamada Análisis de Varianza que en realidad abarca a un vasto conjunto de técnicas.

Consideraciones Conceptuales

La asociación entre variables, al comparar tres o más grupos, se mide a través del estadístico F, cuyo valor oscila entre $-\infty$ y ∞ y cuanto más alejado de cero, mayor será la asociación entre variables.



Consideraciones Conceptuales

La aplicación del análisis de varianza en R, es bastante sencilla como se hará evidente en nuestra documentación de GitHub.

https://github.com/jcorrean/Pantaleon/blob/main/ANOVA.Rmd

Tal sencillez, sin embargo, resulta algo difícil de maniobrar por las recientes consideraciones que deben tenerse a la hora de interpretar sus resultados.

Contextualización

Ahora hay una variable "Entrenamiento" con tres valores (Poco, Intermedio y Alto) que hacen referencia a la cantidad de experiencia acumulada por una gerencia para garantizar la producción de la 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.

Contextualización

```
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
```

Contextualización

La diferencia significativa observada aún no queda completamente evidenciada dada las múltiples comparaciones que realiza un análisis de varianza.

Se hace fundamental conocer cuál es el origen de las diferencias, dado que el anova es sensible a reflejar diferencias incluso cuando no hayan diferencias entre todos los pares de grupos. Para ello, se requiere complementar los cálculos con lo que se llama comparación pareada o de grupos.

Contextualización

```
Entrenamiento <- AOV$Entrenamiento
tapply(AOV$TCH, Entrenamiento, mean)
tapply(AOV$TCH, Entrenamiento, sd)
pairwise.t.test(AOV$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
```

Contextualización

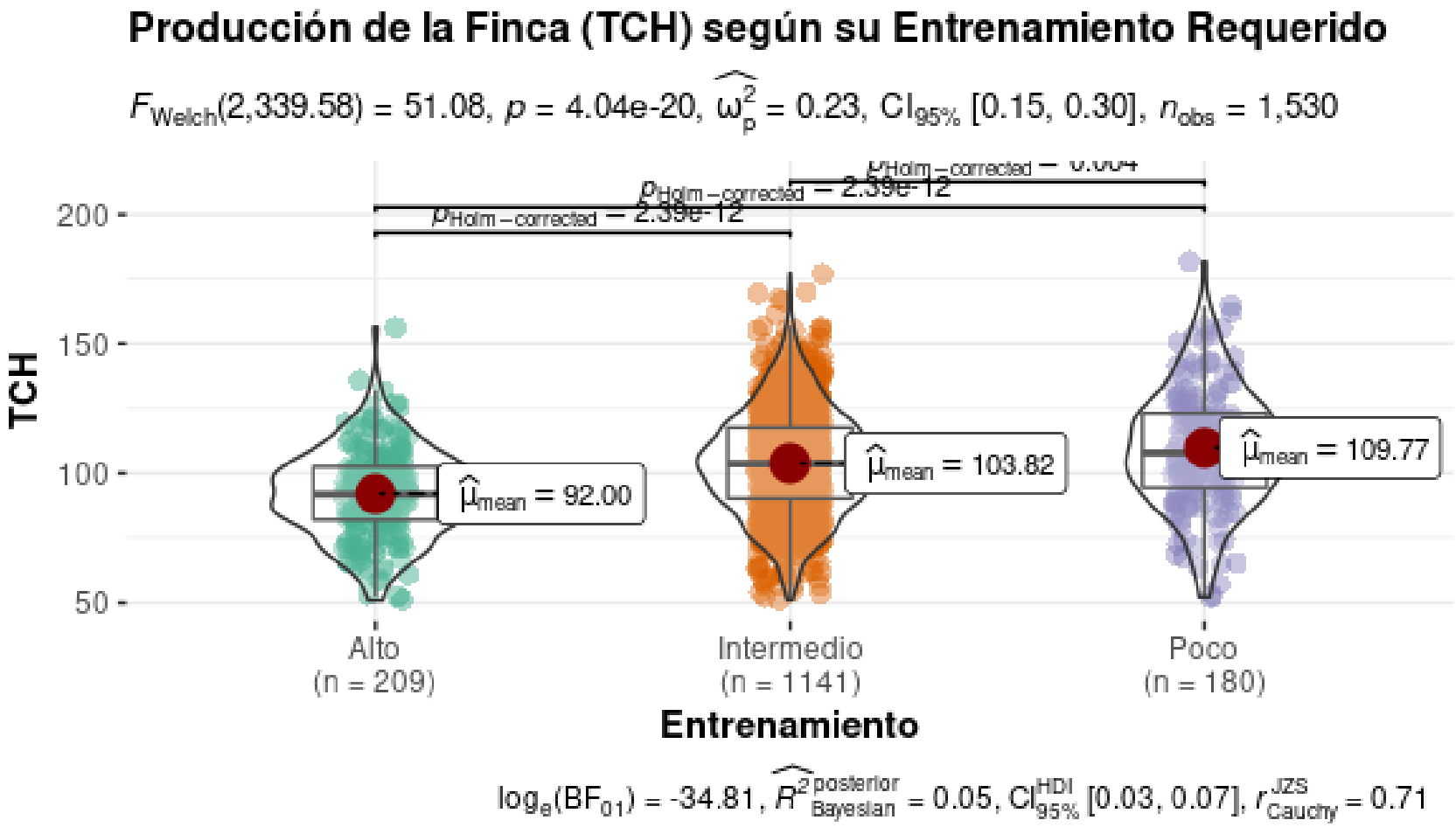
Un detalle que debe considerarse de la sintaxis anterior es que en el argumento p.adj se establece “none” o ningun ajuste en el cálculo de las probabilidades. Esta elección, sin embargo, puede llevarnos a un problema que se conoce como error tipo 1 en estadística. El método de Bonferroni-Holm se utiliza para contrarrestar el problema de no controlar la tasa de error familiar (es decir, la probabilidad de realizar uno o más descubrimientos falsos).

Contextualización

```
        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
```

Contextualización

Siempre es recomendable acompañar los cálculos estadísticos con gráficos. Por ejemplo, el siguiente gráfico realizado con la librería ggstatplot.



Chequeo de Supuestos

Aunque el análisis de varianza es una técnica suficientemente robusta a problemas como datos perdidos, datos extremos (outliers), basta que los residuales (parte del modelo que no se ajusta a los datos) muestren una distribución ligeramente fuera de la curva normal, para que genere resultados sesgados.

Behavior Research Methods (2020) 52:464–488
https://doi.org/10.3758/s13428-019-01246-w

Robust statistical methods in R using the WRS2 package

Patrick Mair[1] · Rand Wilcox[2]

Published online: 31 May 2019

https://doi.org/10.3758/s13428-019-01246-w

Chequeo de Supuestos

Un chequeo de supuestos básicos en el análisis de varianza es el cálculo de la prueba de heteroscedasticidad.

```
library(lmtest)
bptest(resultado_TCH)
library(car)
ncvTest(lm(AOV$TCH ~ AOV$Entrenamiento))
```

```
        studentized Breusch-Pagan test

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

Chequeo de Supuestos

Siempre debe evaluarse la violación al supuesto de homoscedasticidad. Si la prueba bp o la prueba ncvTest nos arroja un estadístico considerablemente alejado de cero con p-value significativo, debemos desechar el modelo anova y correr algún alternativo robusto o un alternativo no-paramétrico.

Chequeo de Supuestos

```
library(WRS2)
t1way(AOV$TCH ~ AOV$Entrenamiento)
```

```
Call:
t1way(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: 0.41
Bootstrap CI: [0.35; 0.5]
```