1 Estudiando ANOVA

(x emilopez)

¿Cuáles son los límites del universo? ¿Dios existe? ¿Cuál es el punto exacto del arroz? ¿Qué carajo es ANOVA? preguntas que la humanidad ha intentado responder desde sus orígenes, en este opúsculo pretendo responder la única sin respuesta.

1.1 ¿Qué carajo es ANOVA?

Es un método para comparar distintos tratamientos, ni más, ni menos. La idea general es separar la variación total en las partes con las que contribuye cada fuente de variación en el experimento. Se separan la variabilidad debida a los tratamientos y la debida al error.

- Cuando la primera predomina "claramente" sobre la segunda, es cuando se concluye que los tratamientos tienen efecto, o dicho de otra manera, las medias son diferentes.
- Cuando los tratamientos no dominan, contribuyen igual o menos que el error, por lo que se concluye que las medias son iguales

Esto se plantea en modo de test, con H0 suponiendo que las medias son iguales y H1 diciendo que al menos un par son distintas, pero **¿cómo hacemos el test, cuál es nuestro estadístico de prueba?**

Bueno, acá está la posta, lo que tenemos que hacer es armar ese estadístico dividiendo algo que me indica la variabilidad entre los tratamientos sobre la variabilidad dentro de cada tratamiento.

ANOVA significa análisis de la varianza, pero ¿dónde entra en juego la varianza? Bueno, debe venir del supuesto para poder realizar el test, que dice que los tratamientos poseen varianza constante.

Es bueno entender lo que significa cada notación, ya que luego lo usamos a troche y moche, si lo entendemos de entrada ganamos vida.

1.2 Notación

Tenemos k cantidad de tratamientos, y en cada tratamiento tenemos un número n_i observaciones. n_i es la cantidad de observaciones del tratamiento *i*.

- Sumamos todas las observaciones n_i de cada tratamiento (medias)

$$N = \sum_{i=1}^{t} n_i$$

- Media de cada tratamiento (muestra)

$$\overline{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

- Media global, promedio de todas las observaciones

$$\overline{Y}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_j} y_{ij}$$

1.3 Modelo

¿Cómo podemos suponer la distribución de las medias de las observaciones? Bueno, podemos armar un modelo/ecuación que nos describa este comportamiento:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

donde y_{ij} es el valor de la observación *j* del tratamiento *i*, μ_i es la media de cada tratamiento *i* y ϵ_{ij} es el error. Otra manera de expresarlo es:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

donde μ es una media global y τ_i es el efecto de cada tratamiento *i*. En base a este modelo lo que buscamos es probar la hipótesis de igualdad de los tratamientos con respecto a la media, es decir:

- H0: $\mu_1 = \mu_2 \cdots \mu_k = \mu$
- H1: $\mu_i \neq \mu_j$

Lo que es equivalente a:

- H0: $\tau_1 = \tau_2 \cdots \tau_k = 0$
- H1: $\tau_i \neq 0$

Donde τ_i es el efecto del tratamiento *i* sobre la variable respuesta, que significa el desplazamiento de una media global μ, es decir, $\tau_i = \mu_i - \mu$

1.3.1 Supuestos de ANOVA

Los supuestos son:

- Los ϵ_{ij} siguen una **distribución normal** con media cero.
- Los ϵ_{ij} son **independientes** entre sí.
- Los residuos de cada tratamiento tienen la **misma varianza** σ^2.

El test sería:

- H0: $\mu_i = \mu_j$
- H1: $\mu_i \neq \mu_j$

Los supuestos deben comprobarse, no podemos mandarnos a hacer ANOVA si no comprobamos que sean normales y de igual varianza, eso se hace así:

- **Normalidad:** con el **test de shapiro**, queremos obtener un *pvalor* grande para aceptar H0, es decir normalidad en la distribución de los datos. Se hace: `shapiro.test(modelo$residuals)`
- **Varianzas iguales:** con el **test de levene**, también queremos un *pvalor* grande. Se hace: `leveneTest(modelo)` , dentro de la lib `car` .

1.4 Método

Para probar las hipótesis dadas lo primero es decomponer la variabilidad total de los datos en sus dos componentes:

- debida a tratamientos (entre tratamientos)
- debida al error aleatorio (dentro de tratamientos)

Suma de Cuadrados de Tratamientos (SC_{ENTRE})

Mide la diferencia entre los tratamientos, si son muy diferentes entre sí, entonces el valor tenderá a ser grande. Tiene k-1 grados de libertad.

$$SC_{ENTRE} = \sum_{i=1}^{k} n_i \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot} \right)^2$$

Suma de Cuadrados del Error (SC_{DENTRO})

Mide la variación entre las observaciones de cada tratamiento, si hay mucha diferencia entonces tenderá a ser un valor grande. Tiene N-k grados de libertad.

$$SC_{DENTRO} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{i\cdot} \right)^2$$

1.5 Estadístico de prueba

Si dividimos las cada suma de cuadrados por sus respectivos grados de libertad obtenemos algo que denominamos *Suma de los cuadrados medios, MCentre y MCdentro*, veamos:

$$MC_{ENTRE} = \frac{SC_{ENTRE}}{k-1}$$

$$MC_{DENTRO} = \frac{SC_{DENTRO}}{N-k}$$

donde k es la cantidad de tratamientos y N la cantidad de observaciones totales. Che, acordate que *MCdentro* es un estimador de la varianza.

Ahora sí, armamos el estadístico F0:

$$F0 = \frac{MC_{ENTRE}}{MC_{DENTRO}} \approx F(k-1, N-k)$$

Con F0 calculamos el pvalor, y con esto sabremos si **RECHAZAMOS o ACEPTAMOS H0** para un nivel α prefijado.

- Si rechazamos H0: tamo al horno, hay que hacer **contrastes y comparaciones múltiples**
- Si aceptamos H0: listo, pero vamos a tener que saber qué potencia tenemos, tal vez nos damos cuenta que hay que aumentar el n.

1.6 Contrastes

Estamos acá porque rechazamos H0 y es lo último en guarachas, o sea, las medias no son todas iguales, entonces vamos a contrastar entre algunas en particular para ver esa diferencia de medias.

Un contraste es esto:

$$C = \sum_{i=1}^{t} k_i \mu_i$$

pero teniendo la condición que sumando los k nos dan 0: $\sum_{i=1}^{t} k_i = 0$.

Pero pensemos, ¿qué significan estos contrastes? Bueno, bien bien en claro no lo tengo, pero sí se que tiene sentido restar las medias para saber si hay diferencia o no. Entonces, supongamos que queremos contrastar el tratamiento 1 con el 3, esto sería: $C1 = \mu 1 - \mu 3$, suponiendo que tenemos 4 tratamientos, los k que tenemos entonces son:

$$k = (1, 0, -1, 0)$$

y ahora tenemos que volver a hacer un test para saber si nuestra nueva H0 se rechaza o acepta, sería:

- H0: $C1 = 0$
- H1: $C1 \neq 0$

¿Cómo resolvemos este test? Mamita querida, esto no termina nunca:

- tenemos que hacer un Intervalo de Confianza (IC) para este C1 y vemos si el 0 está dentro.
- Para hacer el IC necesitamos un estimador del C1: $\overline{C} = \sum_{i=1}^{t} k_i \overline{y}_{i\cdot}$

El IC siempre es de la forma: $\theta \pm A\sqrt{var(\theta)}$ que termina quedando de la siguiente forma:

$$\overline{C} \pm t_{\alpha/2}(N-t) * S_p * \sqrt{\sum_{i=1}^{t} \frac{k_i^2}{n_i}}$$

Notar que el Sp^2 (que es el MSE) estaba dentro de la raíz y fue sacado fuera por lo que queda Sp.

OJITO: *no confundir la primer t, de la distribución t-student con la t del extremo de la sumatoria que refiere a la cantidad de tratamientos. Tampoco confundir la k que usamos acá , que son los coeficientes de los constrastes con la k de la sección previa que hacía referencia a la cantidad de tratamientos.*

En este caso, denominamos a como **errores estandares de los contrastes** a la raiz cuadrada de las varianzas, que nos quedaría:

$$ErroresContraste_j = \sqrt{MSE \sum \frac{k_{ji}^2}{n_i}}$$

donde r es la cantidad de obsrevaciones si en cada tratamiento son iguales.

1.7 Comparaciones simultáneas

Acá lo que hacemos son IC para contrastes simultáneos. ¿WTF? Claro, como vamos a hacer los contrastes simultáneos acá entra en juego cómo tomamos el α. ¿Por qué? porque el nivel de significancia (α_0) que tenemos es para analizar cada muestra, no así para la familia de muestras como necesitamos. Lo que terminaremos haciendo es armar un IC usando este nuevo nivel de significancia. Hay diferentes métodos, cada uno se adapta mejor según el problema que tengamos.

Recordemos que los IC siempre tienen esta forma: $\theta \pm A_\alpha \sqrt{var(\theta)}$

1.7.1 Bonferroni

- Sirve para comparar todos contra todos.
- Es el mas conservador, toma un nuevo $\overline{\alpha} = \alpha_0/k$ donde k son la cantidad de tratamientos.
- Lo uso si tengo pocos tratamientos, k es chico

1.7.2 Tukey

- Lo usamos cuando tenemos un control y comparamos cada uno contra ese control

1.7.3 Dunnett

- Para comparar con el mejor o un control
- Para comparar con un control se usa: `T2 = glht(modelo, linfct = mcp(etapa = "Dunnet"), alternative="two.sided")` y vemos `summary(T2)` para luego `plot(confint(T2, level = .95))` . Se debe cargar `library(multcomp)` .
- Para ver el mejor, máximo o mínimo usamos `minHSU(Parasitos, Tratamiento, alpha=0.05, MSE, 20)` habiendo previamente cargado `source("EstadisticaAplicada/practica/mymultcomp.R")`

1.7.4 Scheffe

- Me dice directamente el valor de A_α
- Supone todos los contrastes posibles.

1.7.5 Contrastes Ortogonales

Si todos mis contrastes son simultáneamente ortogonales entonces puedo calcular $\alpha = 1 - (1 - \alpha_0)^{1/M}$ donde M es la cantidad de tratamientos.

1.8 Ejercicio de ejemplo reducido

Para estudiar si el tipo de dieta influye en el tiempo de supervivencia de ratas, se midió el tiempo de supervivencia de 349 ratas femeninas que fueron asignados aleatoriamente a uno de los 6 tipos diferentes de dietas das a continuación:

In [1]:

```
datos = read.table('/home/emiliano/EstadisticaAplicada/Estadistica.Aplicada.201
```

In [2]:

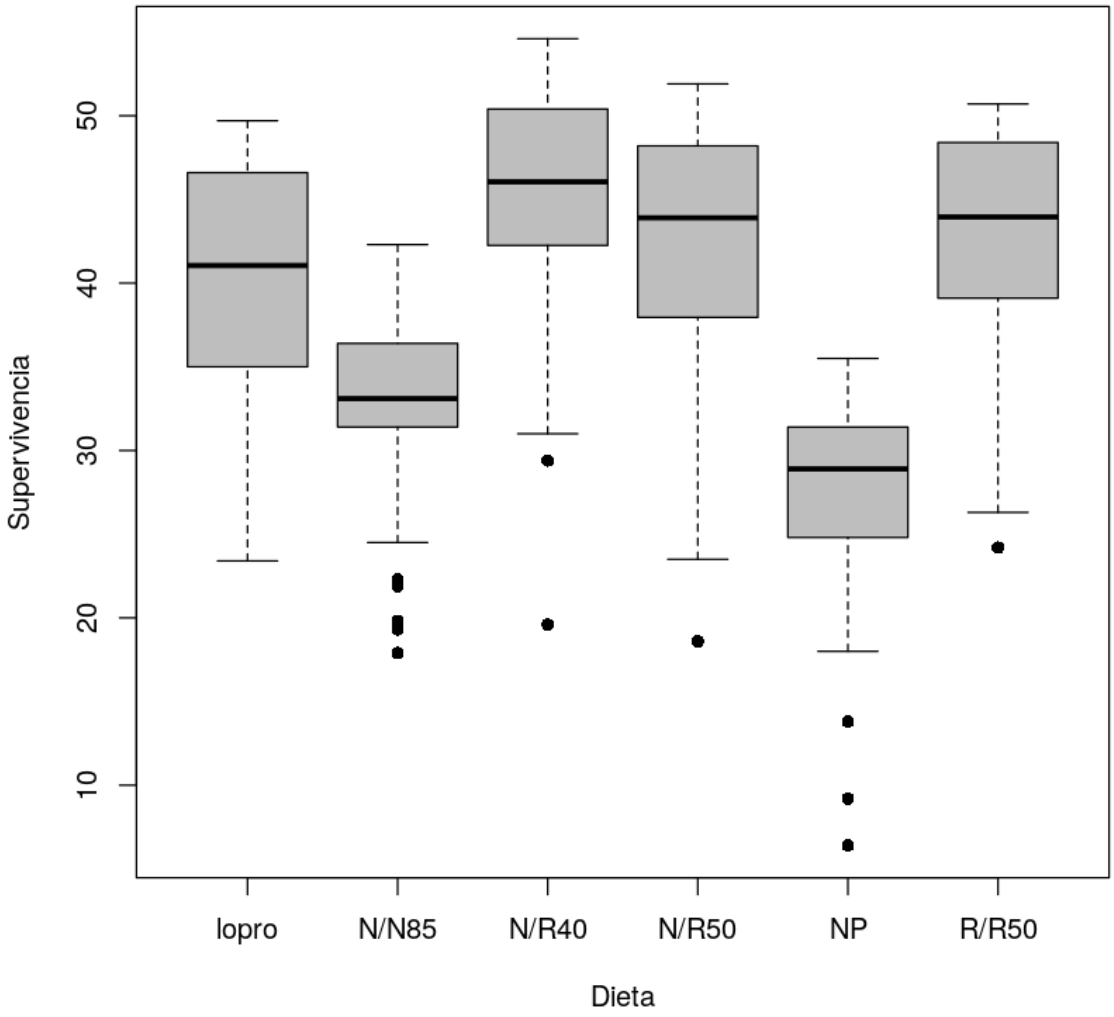
```
head(datos)
attach(datos)
```

cuantovive	dieta
35.5	NP
35.4	NP
34.9	NP
34.8	NP
33.8	NP
33.5	NP

1.9 Resumen gráfico

In [3]:

```
1 boxplot(cuantovive~dieta, col = 'gray', pch = 16, xlab = 'Dieta', ylab = 'Super
```



1.10 Resumen numérico

Veamos el n de cada tipo de dieta, su promedio y su desvío

In [4]:

```
# cantidad de elementos de cada muestra
ns = tapply(cuantovive, dieta, length)
# promedios de cada tipo de dieta
promedio = tapply(cuantovive, dieta, mean)
# desvios
desvio = tapply(cuantovive, dieta, sd)
print(cbind(ns, promedio, desvio), digits = 3)

```

```
      ns promedio desvio
lopro 56     39.7   6.99
N/N85 57     32.7   5.13
N/R40 60     45.1   6.70
N/R50 71     42.3   7.77
NP    49     27.4   6.13
R/R50 56     42.9   6.68
```

1.11 Tabla ANOVA

Mostramos la tabla ANOVA

In [5]:

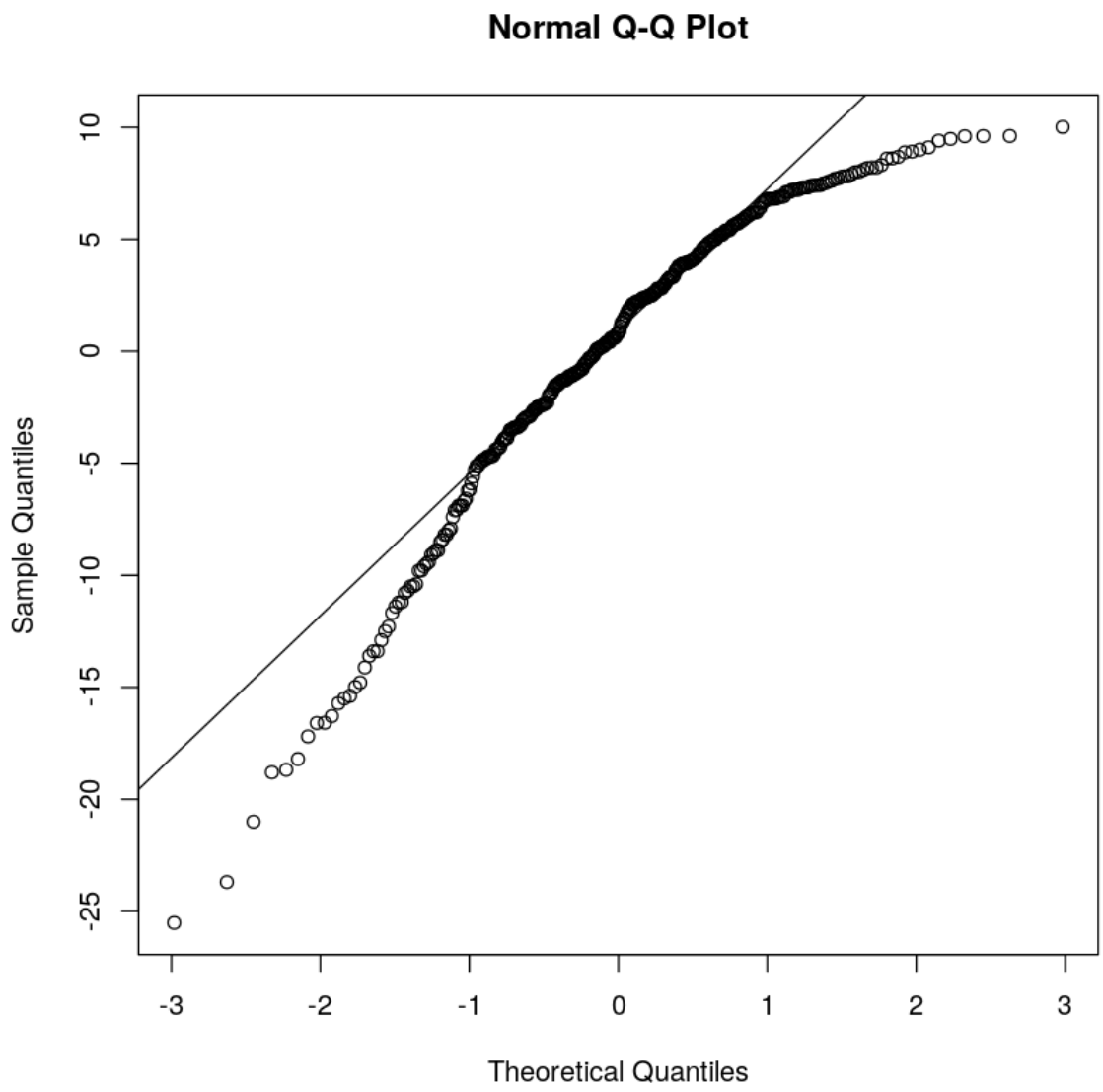
```
dieta = as.factor(dieta)
modelo = aov(cuantovive~dieta)
summary(modelo)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
dieta         5  12734  2546.8    57.1 <2e-16 ***
Residuals   343  15297    44.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Chequeo de normalidad

In [6]:

```
qqnorm(modelo$residuals)
qqline(modelo$residuals)
```



Chequeo de varianza constante

In [9]:

```
library(car)
leveneTest(modelo)
```

	Df	F value	Pr(>F)
group	5	2.721249	0.01989421
	343	NA	NA

Otra forma de llegar a lo mismo es hacer el modelo para H1 y compararlo con el modelo H0

In [8]:

```
modeloH1 = lm(cuantovive~dieta)
modeloH0 = lm(cuantovive~1)
anova(modeloH1, modeloH0)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
343	15297.42	NA	NA	NA	NA
348	28031.36	-5	-12733.94	57.10431	4.111744e-43

In [ ]:

```

```