Clase 11 - Análisis de varianza

DBT 845 - Investigación reproducible y análisis de datos biotecnológicos con R.

Dr. José Gallardo Matus

Pontificia Universidad Católica de Valparaíso

15 May 2022

PLAN DE LA CLASE

1.- Introducción

- ¿Qué es un análisis de varianza?.
- Anova como un modelo lineal.
- Hipótesis y supuestos.
- Interpretar resultados de análisis de varianza con R.

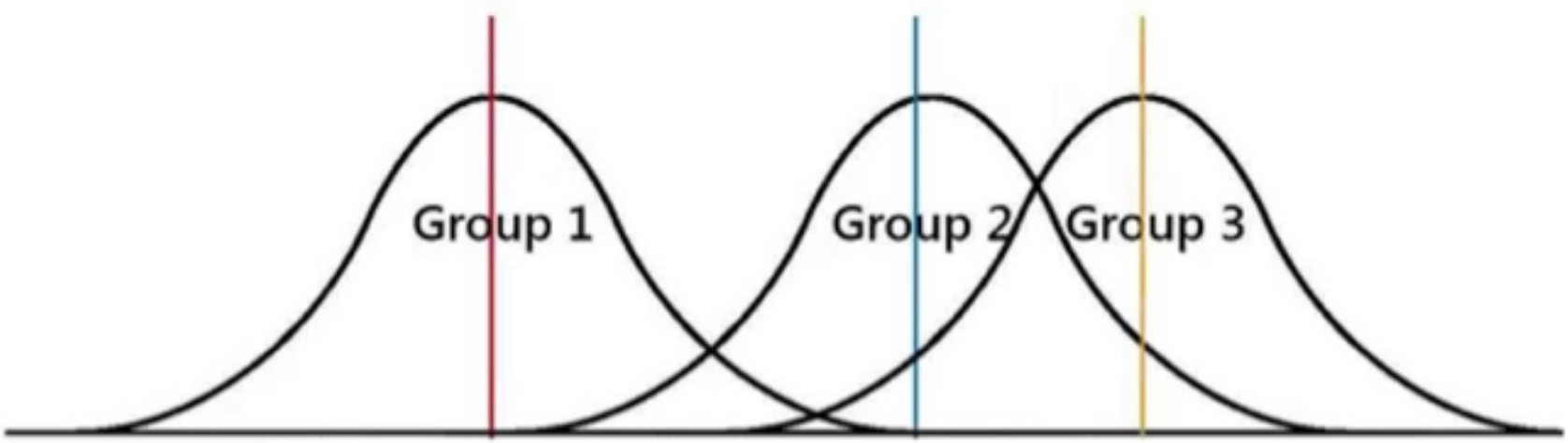
2.- Práctica con R y Rstudio cloud

- Realizar pruebas de hipótesis: Anova y posteriores.
- Realizar gráficas avanzadas con ggplot2.
- Elaborar reporte dinámico en formato html.

ANOVA

¿Qué es un análisis de varianza?

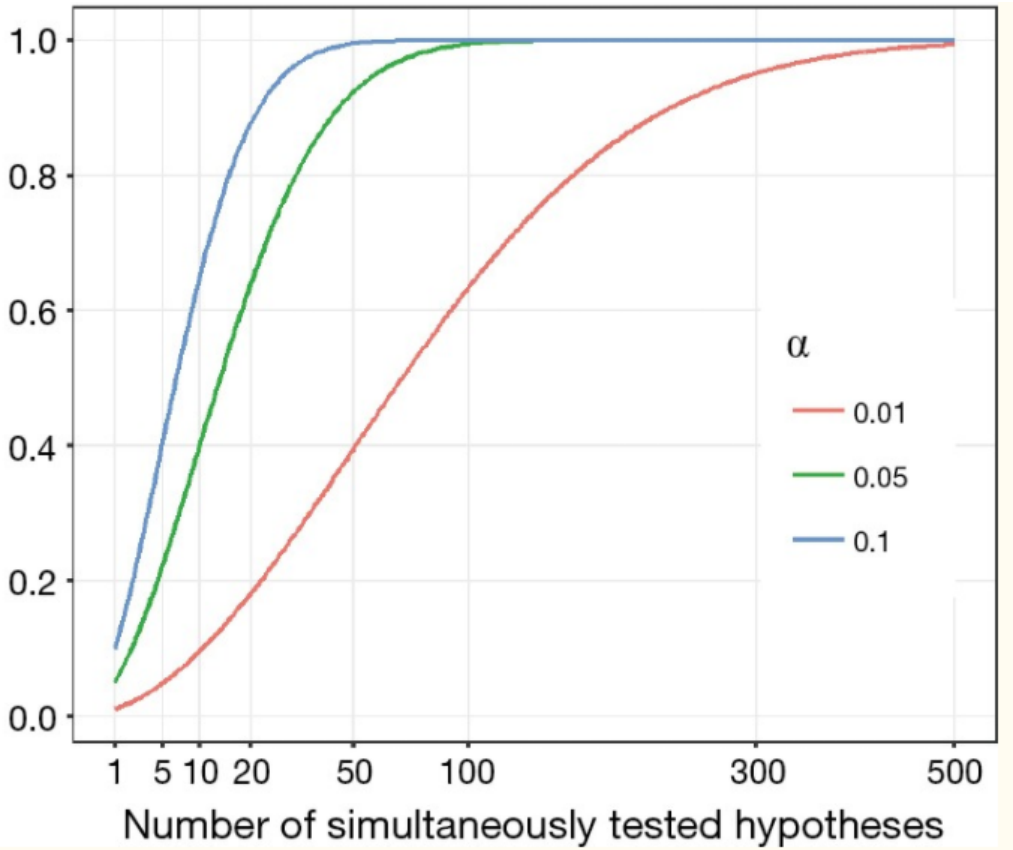
Herramienta básica para analizar el efecto de uno o más factores (cada uno con dos o más niveles) en un experimento.



PROBLEMA DE LAS COMPARACIONES MÚLTIPLES

¿Por qué preferir anova y no múltiples t-test?

Porque con una t-test normal al aumentar el número de comparaciones múltiples se incrementa la tasa de error tipo I.



ANOVA COMO UN MODELO LINEAL

¿Qué es un modelo lineal?

Modelo estadístico que define una relación matemática lineal entre variables de interes.

Modelo lineal para ANOVA de una vía

$y \sim \mu + \alpha + \epsilon$

Modelo lineal para ANOVA de dos vías

$y \sim \mu + \alpha + \beta + \epsilon$

Modelo lineal para ANOVA de dos vías con interacción

$y \sim \mu + \alpha + \beta + \alpha^{*}\beta + \epsilon$

HIPÓTESIS EN UNA ANOVA

Hipótesis factor 1

$\mathbf{H_0}$: $\alpha_{1.1} = \alpha_{1.2} = \alpha_{1.3}$

Hipótesis factor 2

$\mathbf{H_0}$: $\beta_{2.1} = \beta_{2.2} = \beta_{2.3}$

Hipótesis interacción

$\mathbf{H_0}$: $\alpha^*\beta = 0$

Hipótesis Alternativa

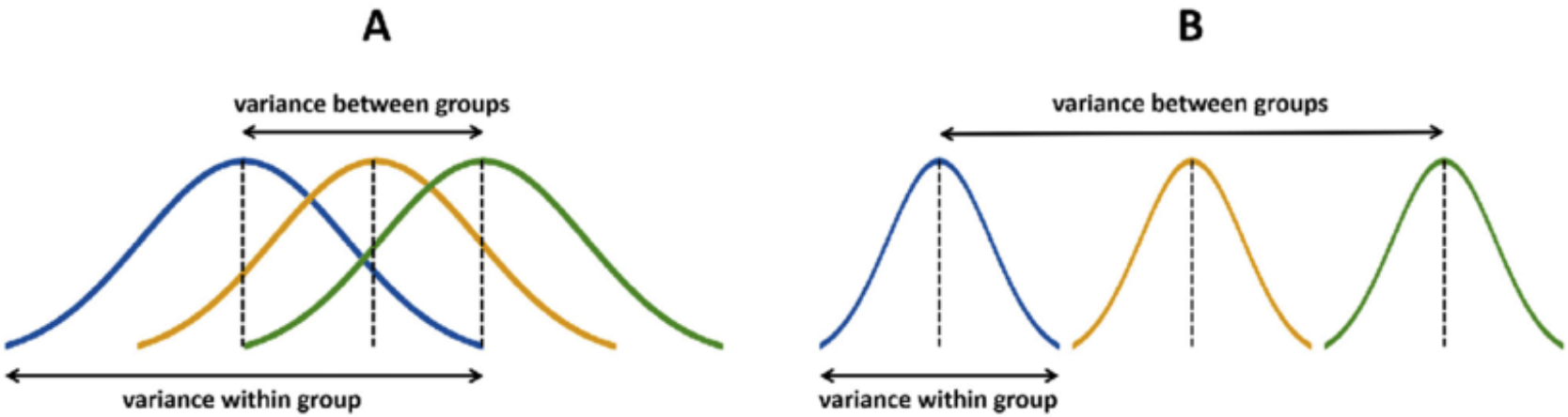
$\mathbf{H_A}$: No todas las medias son iguales

ANOVA PARA COMPARAR MEDIAS

¿Por qué se llama ANOVA si se comparan medias?
Por que el estadístico **F** es un cociente de varianzas.

$$\mathbf{F} = \frac{\sigma^2_{entregrupos}}{\sigma^2_{dentrogrupos}}$$

Mientras mayor es el estadístico **F**, más es la diferencia de medias entre grupos.



SUPUESTOS DE UNA ANOVA

1) Independencia de las observaciones.

2) Normalidad.

3) Homocedasticidad: homogeneidad de las varianzas.

TEST POSTERIORES (PRUEBAS A POSTERIORI)

¿Para qué sirven?

Para identificar que pares de niveles de uno o más factores son significativamente distintos entre sí.

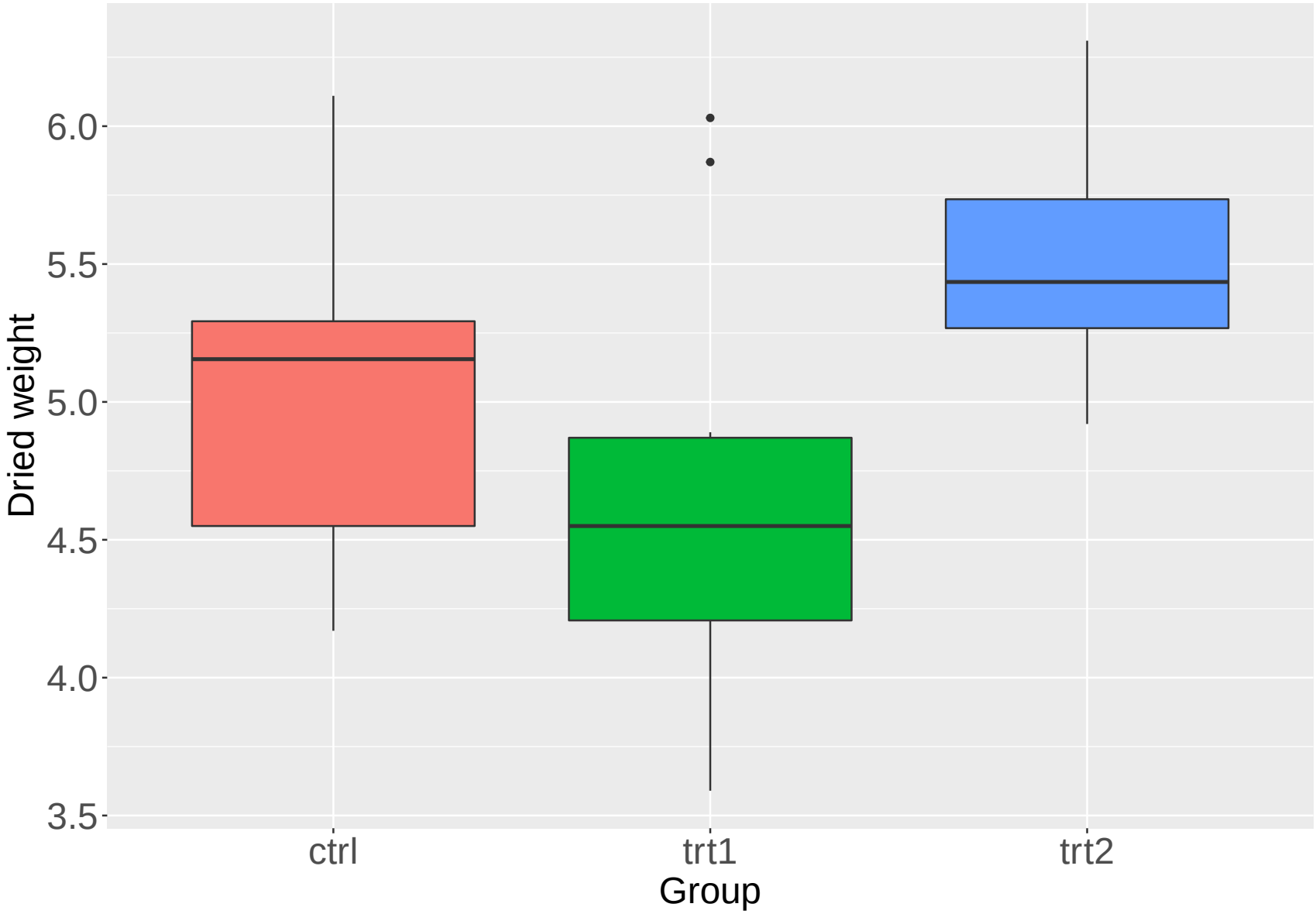
¿Cuando usarlos?

Sólo cuando se rechaza $\mathbf{H_0}$ del ANOVA.

Tukey test

Es uno de los más usados, similar al *t-test*, pero corrige la tasa de error por el número de comparaciones.

ESTUDIO DE CASO: CRECIMIENTO DE PLANTAS



ANOVA

```
res.aov <- lm(`Dried weight` ~ Group, data = my_data)
anova(res.aov)
```

```
## Analysis of Variance Table
##
## Response: Dried weight
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## Group      2  3.7663  1.8832  4.8461 0.01591 *
## Residuals 27 10.4921  0.3886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

ANOVA COMO MODELO LINEAL

summary(res.aov)

```
Call:
lm(formula = `Dried weight` ~ Group, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0710 -0.4180 -0.0060  0.2627  1.3690

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0320     0.1971  25.527   <2e-16 ***
Grouptrt1    -0.3710     0.2788  -1.331   0.1944
Grouptrt2     0.4940     0.2788   1.772   0.0877 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6234 on 27 degrees of freedom
Multiple R-squared:  0.2641,    Adjusted R-squared:  0.2096
F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591
```

MODELO LINEAL SIN INTERCEPTO

res.aov <- lm(`Dried weight` ~ -1 + Group, data = my_data)
summary(res.aov)

```
Call:
lm(formula = `Dried weight` ~ -1 + Group, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0710 -0.4180 -0.0060  0.2627  1.3690

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
Groupctrl   5.0320     0.1971   25.53   <2e-16 ***
Grouptrt1   4.6610     0.1971   23.64   <2e-16 ***
Grouptrt2   5.5260     0.1971   28.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6234 on 27 degrees of freedom
Multiple R-squared:  0.9867,    Adjusted R-squared:  0.9852
F-statistic: 665.5 on 3 and 27 DF,  p-value: < 2.2e-16
```

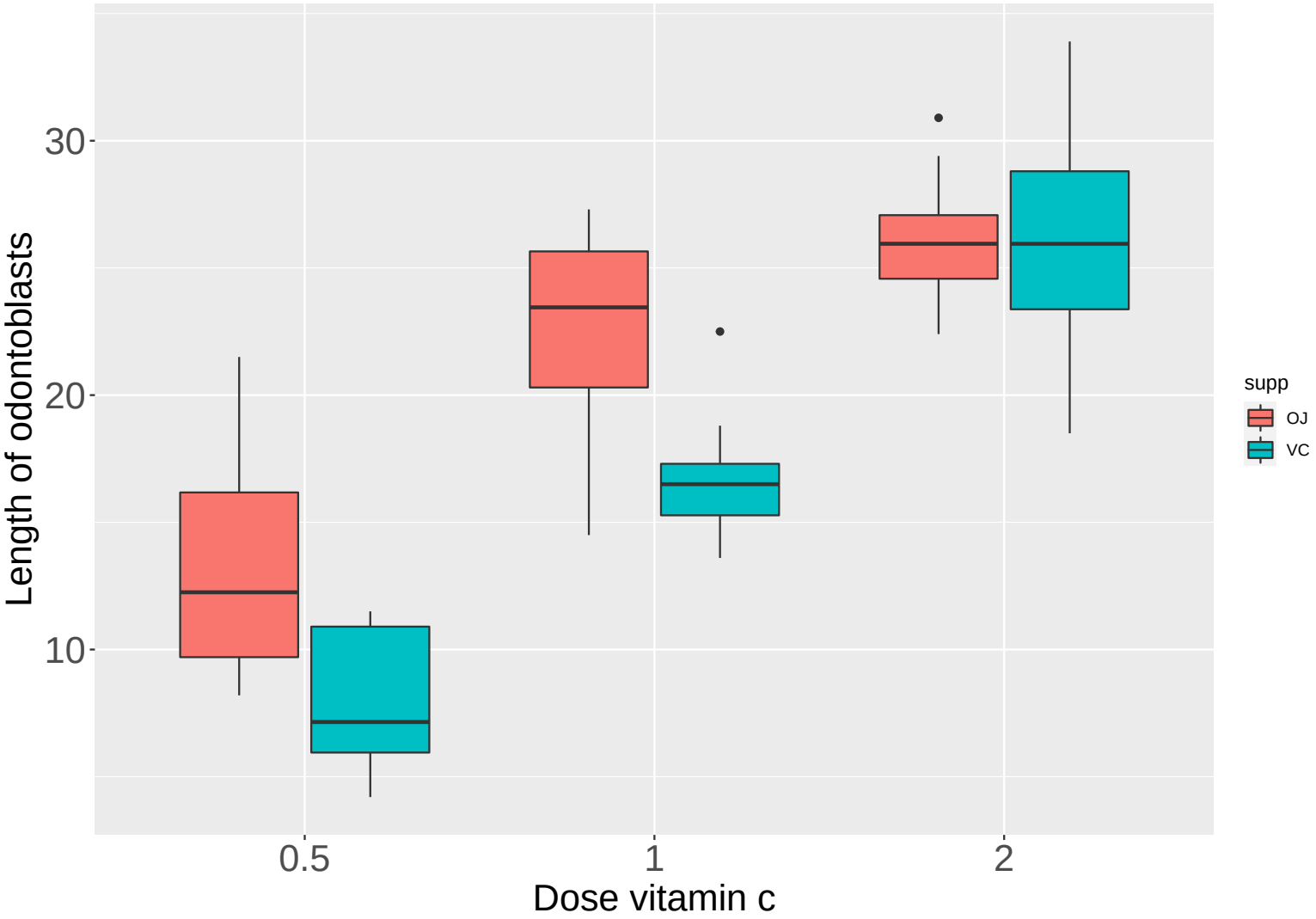
COMPARACIONES MÚLTIPLES

```
fit_anova <- aov(res.aov)
tk <- TukeyHSD(fit_anova)
```

Table 1: Prueba de Tukey.

Trat.	Contraste	H0	Diferencia	IC-bajo	IC-alto	p-ajustado
Group	trt1-ctrl	0	-0.37	-1.06	0.32	0.39
Group	trt2-ctrl	0	0.49	-0.20	1.19	0.20
Group	trt2-trt1	0	0.86	0.17	1.56	0.01

ESTUDIO DE CASO: GUINEA PIGS



OJ: Orange jouice - VC: Vitamin C

ANOVA DOS VIAS CON INTERACCIÓN

```
res.aov2 <- lm(len ~ dose * supp, data = my_data1)
anova(res.aov2)
```

```
Analysis of Variance Table

Response: len
           Df  Sum Sq Mean Sq F value    Pr(>F)
dose        2 2426.43 1213.22  92.000 < 2.2e-16 ***
supp        1  205.35  205.35  15.572 0.0002312 ***
dose:supp   2  108.32   54.16   4.107 0.0218603 *
Residuals  54  712.11   13.19
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA COMO MODELO LINEAL

summary(res.aov)

```
Call:
lm(formula = len ~ dose * supp, data = my_data1)

Residuals:
   Min     1Q Median     3Q    Max
 -8.20  -2.72  -0.27   2.65   8.27

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    13.230      1.148  11.521 3.60e-16 ***
dose1           9.470      1.624   5.831 3.18e-07 ***
dose2          12.830      1.624   7.900 1.43e-10 ***
suppVC         -5.250      1.624  -3.233  0.00209 **
dose1:suppVC   -0.680      2.297  -0.296  0.76831
dose2:suppVC    5.330      2.297   2.321  0.02411 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.631 on 54 degrees of freedom
Multiple R-squared:  0.7937,    Adjusted R-squared:  0.7746
F-statistic: 41.56 on 5 and 54 DF,  p-value: < 2.2e-16
```

PRÁCTICA ANÁLISIS DE DATOS

- El trabajo práctico se realiza en Rstudio.cloud.
  Guía 11 Anova y posteriores

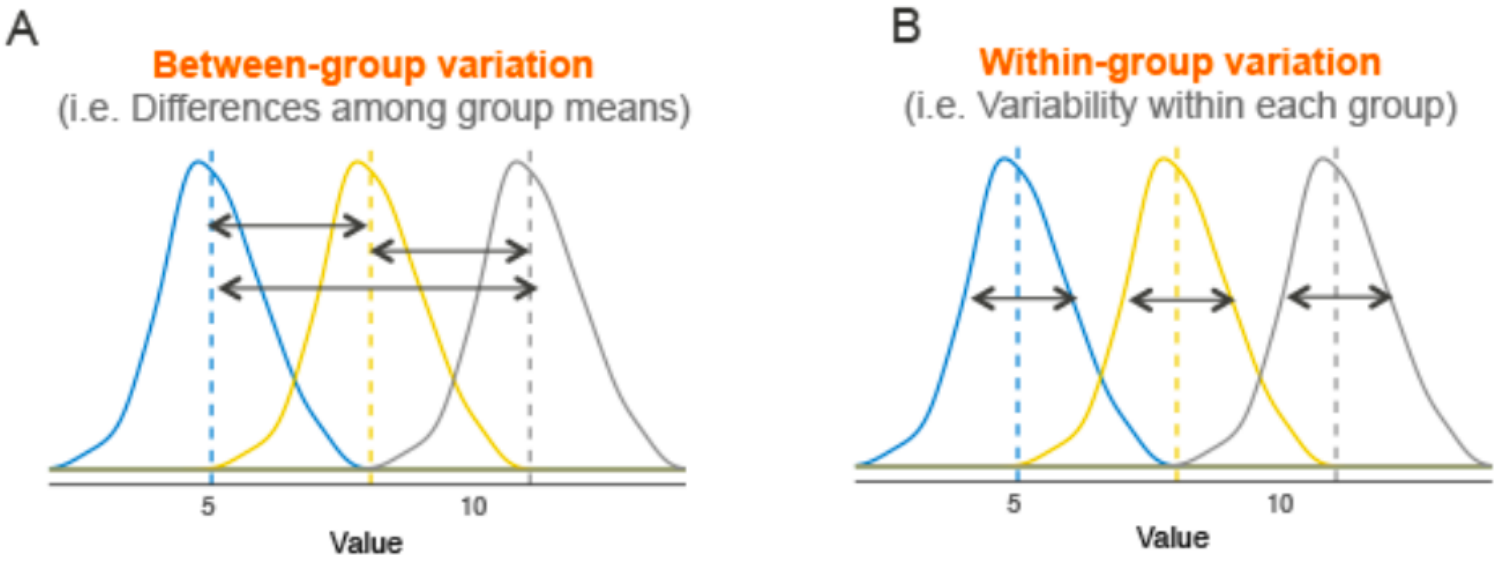


Figura: datanovia.com

RESUMEN DE LA CLASE

- **Elaborar hipótesis de anova**
- **Realizar análisis de varianza**
  - 1 factor.
  - 2 factores.
  - pruebas *a posteriori*
- **Realizar gráficas avanzadas con ggplot2**