

# ANOVA

análisis de varianza

Carlos Carleos, Norberto Corral

22 de marzo de 2022

# Planteamiento

- ▶ Sean  $q$  poblaciones  $x_i$  ( $i = 1, \dots, q$ )
  - ▶ niveles o modalidades de un “factor”
- ▶  $x_i = \mu_i + \epsilon \equiv \mathcal{N}(\mu_i, \sigma)$  con  $\epsilon \equiv \mathcal{N}(0, \sigma) = \mathcal{N}_1(0, \sigma^2)$
- ▶ De cada  $x_i$  se conoce una muestra aleatoria simple de tamaño  $n_i$ :  $x_{i1}, \dots, x_{in_i}$
- ▶ Tamaño muestral total:  $n = \sum_{i=1}^q n_i$
- ▶ Otra parametrización:  $\mu_i = \mu + \alpha_i$  con  $\mu = \frac{n_i \mu_i}{n}$
- ▶  $\sum_{i=1}^q n_i \alpha_i = 0$
- ▶ Si  $\forall i, n_i = \frac{n}{q}$  el modelo se dice *equilibrado* o *balanceado*

# Planteamiento

Hipótesis previas

$$\forall i = 1, \dots, q \quad \vec{x}_i = (x_{i1}, \dots, x_{in_i})^t \equiv \mathcal{N}_{n_i} \left( \mu_i \vec{1}_{n_i}, \sigma^2 I_{n_i} \right)$$

- ▶ independencia (muestras aleatorias simples)
- ▶ gaussianidad
- ▶ homoscedasticidad (igualdad de varianzas)

# Planteamiento

Hipótesis del análisis de varianza (ANOVA)

$$H_0 \quad \equiv \quad \forall i, j, \mu_i = \mu_j \quad \equiv \quad \forall i, \alpha_i = 0$$

$$H_1 \quad \equiv \quad \exists i, j, \mu_i \neq \mu_j \quad \equiv \quad \exists i, \alpha_i \neq 0$$

- ▶  $H_0$ : el factor no influye en la respuesta
- ▶  $H_1$ : el factor sí influye en la respuesta

# Repaso de álgebra lineal

- ▶ Matriz simétrica  $A = [a_{ij}]_{i,j}$  si  $a_{ij} = a_{ji} \equiv A = A^t$ 
  - ▶ todos sus autovalores  $(\lambda_i)_i$  son reales; autovectores:  $(\vec{u}_i)_i$
  - ▶  $A = U \Lambda U^t$  con  $\Lambda = \text{diag}(\dots, \lambda_i, \dots)$  y  $U = [\dots, \vec{u}_i \dots]$
- ▶ Matriz idempotente:  $A A = A$ 
  - ▶ (autovalor  $A$ )  $\in \{0, 1\}$  pues si  $\vec{x} \neq 0$  es autovector no nulo

$$\begin{aligned} \lambda \vec{x} &= A \vec{x} = A A \vec{x} = A \lambda \vec{x} = \lambda A \vec{x} = \lambda^2 \vec{x} \\ \implies \quad \lambda &= 0 \quad \cup \quad \left\{ \vec{x} = \lambda \vec{x} \implies \lambda = 1 \right\} \end{aligned}$$

- ▶ Traza:  $A = [a_{ij}]_{i,j} \implies \text{tr } A = \sum_i a_{ii}$ 
  - ▶  $a = \text{tr}[a]$
  - ▶  $\text{tr}(A B) = \text{tr}(B A)$
  - ▶  $\text{tr}(A B C) = \text{tr}(B C A) = \text{tr}(C A B)$
  - ▶  $E[\text{tr } A] = \text{tr } E[A]$

## Matriz de varianzas y covarianzas

$$\begin{aligned}\text{Var}(\vec{x}) &= \text{Cov}(\vec{x}) = \text{Cov}(\vec{x}, \vec{x}) = \text{E} [(\vec{x} - \text{E}[\vec{x}])(\vec{x} - \text{E}[\vec{x}])^t] \\ &= \text{E} \left[ \begin{pmatrix} (x_1 - \text{E}[x_1]) \\ \vdots \\ (x_q - \text{E}[x_q]) \end{pmatrix} [(x_1 - \text{E}[x_1]) \quad \dots \quad (x_q - \text{E}[x_q])] \right] = \\ &= \begin{bmatrix} \text{E} [(x_1 - \text{E}[x_1])^2] & \dots & \text{E} [(x_1 - \text{E}[x_1])(x_q - \text{E}[x_q])] \\ \vdots & \ddots & \vdots \\ \text{E} [(x_1 - \text{E}[x_1])(x_q - \text{E}[x_q])] & \dots & \text{E} [(x_q - \text{E}[x_q])^2] \end{bmatrix}\end{aligned}$$

- ▶ intradiagonal:  $\text{E} [(x_i - \text{E}[x_i])^2] = \text{Var}(x_i)$
- ▶ extradiagonal:  $\text{E} [(x_i - \text{E}[x_i])(x_j - \text{E}[x_j])] = \text{Cov}(x_i, x_j)$

## Matriz de varianzas y covarianzas

$$\begin{aligned}\text{Var}(\vec{x}) &= \text{Cov}(\vec{x}) = \text{Cov}(\vec{x}, \vec{x}) = \mathbf{E} [(\vec{x} - \mathbf{E}[\vec{x}])(\vec{x} - \mathbf{E}[\vec{x}])^t] = \\ &= \mathbf{E} [\vec{x} \vec{x}^t] - \mathbf{E} [\vec{x} \mathbf{E}[\vec{x}]^t] - \mathbf{E} [\mathbf{E}[\vec{x}] \vec{x}^t] + \mathbf{E} [\mathbf{E}[\vec{x}] \mathbf{E}[\vec{x}]^t] = \\ &= \mathbf{E} [\vec{x} \vec{x}^t] - \mathbf{E}[\vec{x}] \mathbf{E}[\vec{x}]^t - \mathbf{E}[\vec{x}] \mathbf{E}[\vec{x}]^t + \mathbf{E}[\vec{x}] \mathbf{E}[\vec{x}]^t = \\ &= \mathbf{E} [\vec{x} \vec{x}^t] - \mathbf{E}[\vec{x}] \mathbf{E}[\vec{x}]^t\end{aligned}$$

## Formas cuadráticas

Dados una matriz  $A$  ( $n, n$ ) simétrica e idempotente con  $\text{rango}(A) = r$  y un vector aleatorio  $\vec{x} \equiv N_n(\vec{\mu}, \sigma^2 I)$ , se verifican las siguientes propiedades:

- ▶  $A = U \Lambda U^t = (U_1 \ U_2) \begin{pmatrix} I_r & \\ & 0 \end{pmatrix} \begin{pmatrix} U_1^t \\ U_2^t \end{pmatrix} = U_1 U_1^t$   
con  $U_1$  vectores propios asociados al valor propio 1  
con  $U_2$  vectores propios asociados al valor propio 0  
y  $U^t U = I_n$
- ▶  $\vec{x}^t A \vec{x} = \vec{x}^t U_1 U_1^t \vec{x}$



# Formas cuadráticas

► esperanza

$$\begin{aligned} \mathbb{E} [\vec{x}^t A \vec{x}] &= \mathbb{E} [\vec{x}^t U_1 U_1^t \vec{x}] = \mathbb{E} [\text{tr} \{ \vec{x}^t U_1 U_1^t \vec{x} \}] \\ &= \mathbb{E} [\text{tr} \{ U_1 U_1^t \vec{x} \vec{x}^t \}] = \text{tr} \{ U_1 U_1^t \mathbb{E} [\vec{x} \vec{x}^t] \} \\ &= \text{tr} \{ U_1 U_1^t (\text{Cov}(\vec{x}) + \mathbb{E} [\vec{x}] \mathbb{E} [\vec{x}^t]) \} \\ &= \text{tr} \{ U_1 U_1^t (\sigma^2 I + \mathbb{E} [\vec{x}] \mathbb{E} [\vec{x}^t]) \} \\ &= \text{tr} \{ \sigma^2 U_1 U_1^t \} + \text{tr} \{ U_1 U_1^t \mathbb{E} [\vec{x}] \mathbb{E} [\vec{x}^t] \} \\ &= \sigma^2 \text{tr}(U_1 U_1^t) + \text{tr} \{ \mathbb{E} [\vec{x}^t] U_1 U_1^t \mathbb{E} [\vec{x}] \} \\ &= \sigma^2 \text{rango}(A) + \mathbb{E} [\vec{x}^t] A \mathbb{E} [\vec{x}] \end{aligned}$$

## Formas cuadráticas

- Si  $A \vec{\mu} = \vec{0}$  entonces  $\vec{x}^t A \vec{x} \equiv \sigma^2 \chi_r^2$ .

$$\vec{x}^t A \vec{x} = \vec{x}^t U_1 U_1^t \vec{x} = \vec{y}_1^t \vec{y}_1, \quad \text{con } \vec{y}_1 = U_1^t \vec{x}$$

$$E[\vec{y}_1] = U_1^t E[\vec{x}] = U_1^t E[\vec{x}] = \vec{\gamma}$$

$$U_1 \vec{\gamma} = U_1 U_1^t E[\vec{x}] = A E[\vec{x}] = \vec{0}$$

y, por ser  $U_1$  una matriz de rango completo, se tiene que  $\vec{\gamma} = \vec{0}$ .

$$\text{Cov}(\vec{y}_1) = U_1^t \text{Cov}(\vec{x}) U_1 = \sigma^2 U_1^t U_1 = \sigma^2 I_r$$

En consecuencia,  $\vec{y}_1 = U_1^t N(\vec{\mu}, \sigma^2 I) \equiv N_r(\vec{0}, \sigma^2 I_r)$

$$\vec{y}_1^t \vec{y}_1 = \sum_{j=1}^r y_{1j}^2 = \sum_{j=1}^r \sigma^2 \left( \frac{y_{1j}}{\sigma} \right)^2 = \sigma^2 \sum_{j=1}^r N^2(0, 1) \equiv \sigma^2 \chi_r^2$$

# Análisis de varianza

Sea el vector columna  $n \times 1$ :  $\vec{x} = \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \\ \vec{x}_q \end{pmatrix}$

yuxtaposición de los vectores  $n_i \times 1$ :  $\vec{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{in_i} \end{pmatrix}$

para  $i = 1, \dots, q$

# Análisis de varianza

$$\vec{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{n \times 1} \quad J = \frac{1}{n} \vec{1}_n \vec{1}_n^t = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix}$$

$J$  es simétrica e idempotente

al multiplicar, genera un vector columna con la media:

$$J \vec{x} = \frac{1}{n} \vec{1}_n \vec{1}_n^t \vec{x} = \vec{1}_n \frac{1}{n} \vec{1}_n^t \vec{x} = \bar{x} \vec{1}_n$$

# Análisis de varianza

$$H = I_n - J = \begin{pmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \dots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & 1 - \frac{1}{n} \end{pmatrix}$$

$H$  es simétrica e idempotente

# Análisis de varianza

$H$  es la matriz de centrado

al multiplicar, sustrae la media:

$$\begin{aligned} H \vec{x} &= \left( I_n - \frac{1}{n} \vec{1}_n \vec{1}_n^t \right) \vec{x} = \vec{x} - \vec{1}_n \frac{1}{n} \vec{1}_n^t \vec{x} \\ &= \vec{x} - \vec{1}_n \bar{x} = \begin{pmatrix} \vec{x}_1 - \bar{x} \vec{1}_{n_1} \\ \vec{x}_2 - \bar{x} \vec{1}_{n_2} \\ \vdots \\ \vec{x}_q - \bar{x} \vec{1}_{n_q} \end{pmatrix} \end{aligned}$$

# Análisis de varianza

$$\begin{aligned}\vec{x}^t H \vec{x} &= \vec{x}^t H H \vec{x} = \vec{x}^t H^t H \vec{x} \\ &= \left[ (\vec{x}_1 - \bar{x} \vec{1}_{n_1})^t, \dots, (\vec{x}_q - \bar{x} \vec{1}_{n_q})^t \right] \begin{pmatrix} \vec{x}_1 - \bar{x} \vec{1}_{n_1} \\ \vec{x}_2 - \bar{x} \vec{1}_{n_2} \\ \vdots \\ \vec{x}_q - \bar{x} \vec{1}_{n_q} \end{pmatrix} \\ &= \sum_{i=1}^q (\vec{x}_i - \bar{x} \vec{1}_{n_i})^t (\vec{x}_i - \bar{x} \vec{1}_{n_i}) \\ &= \sum_{i=1}^q \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \text{SCT}\end{aligned}$$

# Análisis de varianza

$$H = I - J = I - D + D - J$$

con

$$D = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_q \end{pmatrix}$$

con  $J_i = \frac{1}{n_i} \vec{1}_{n_i} \vec{1}_{n_i}^t$

- ▶  $I - D$  es simétrica e idempotente
- ▶  $\text{tr}(I - D) = \text{tr} I - \text{tr} D = n - q = \text{rango}(I - D)$
- ▶  $D - J$  es simétrica e idempotente
- ▶  $\text{tr}(D - J) = \text{tr} D - \text{tr} J = q - 1 = \text{rango}(D - J)$



# Análisis de varianza

$$D \vec{x} = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_q \end{pmatrix} \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \\ \vec{x}_q \end{pmatrix} = \begin{pmatrix} J_1 \vec{x}_1 \\ J_2 \vec{x}_2 \\ \vdots \\ J_q \vec{x}_q \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \vec{1}_{n_1} \\ \bar{x}_2 \vec{1}_{n_2} \\ \vdots \\ \bar{x}_q \vec{1}_{n_q} \end{pmatrix}$$

# Análisis de varianza

$$(I - D) \vec{x} = \vec{x} - D \vec{x} = \begin{pmatrix} \vec{x}_1 - \bar{x}_1 \vec{1}_{n_1} \\ \vec{x}_2 - \bar{x}_2 \vec{1}_{n_2} \\ \vdots \\ \vec{x}_q - \bar{x}_q \vec{1}_{n_q} \end{pmatrix}$$

# Análisis de varianza

$$\begin{aligned}(D - J) \vec{x} &= D \vec{x} - J \vec{x} = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_q \end{pmatrix} \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \\ \vec{x}_q \end{pmatrix} - J \vec{x} \\ &= \begin{pmatrix} J_1 \vec{x}_1 \\ J_2 \vec{x}_2 \\ \vdots \\ J_q \vec{x}_q \end{pmatrix} - \bar{x} \vec{1}_n = \begin{pmatrix} \bar{x}_1 \vec{1}_{n_1} \\ \bar{x}_2 \vec{1}_{n_2} \\ \vdots \\ \bar{x}_q \vec{1}_{n_q} \end{pmatrix} - \bar{x} \vec{1}_n = \begin{pmatrix} (\bar{x}_1 - \bar{x}) \vec{1}_{n_1} \\ (\bar{x}_2 - \bar{x}) \vec{1}_{n_2} \\ \vdots \\ (\bar{x}_q - \bar{x}) \vec{1}_{n_q} \end{pmatrix}\end{aligned}$$

# Análisis de varianza

$$\begin{aligned}\text{SCT} &= \vec{x}^t H \vec{x} = \vec{x}^t (I - J) \vec{x} = \vec{x}^t (I - D + D - J) \vec{x} \\ &= \vec{x}^t (I - D) \vec{x} + \vec{x}^t (D - J) \vec{x} \\ &= \sum_{i=1}^q \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + \sum_{i=1}^q n_i (\bar{x}_i - \bar{x})^2 \\ &= \text{SCE} + \text{SCF}\end{aligned}$$

ya que...

# Análisis de varianza

$$\begin{aligned}\vec{x}^t (I - D) \vec{x} &= \\ &= \vec{x}^t \left[ \begin{pmatrix} I_{n_1} & & & \\ & I_{n_2} & & \\ & & \ddots & \\ & & & I_{n_q} \end{pmatrix} - \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_q \end{pmatrix} \right] \vec{x} \\ &= (\vec{x}_1^t, \dots, \vec{x}_q^t) \begin{pmatrix} I_{n_1} - J_1 & & & \\ & I_{n_2} - J_2 & & \\ & & \ddots & \\ & & & I_{n_q} - J_q \end{pmatrix} \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vdots \\ \vec{x}_q \end{pmatrix} \\ &= \sum_{i=1}^q \vec{x}_i^t (I_{n_i} - J_i) \vec{x}_i = \sum_{i=1}^q \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = \text{SCE}\end{aligned}$$

# Análisis de varianza

Por ser  $D - J$  simétrica e idempotente se tiene que

$$\begin{aligned}\vec{x}^t (D - J) \vec{x} &= \vec{x}^t (D - J)^t (D - J) \vec{x} \\ &= \left[ (\bar{x}_1 - \bar{x}) \vec{1}_{n_1}^t, \dots, (\bar{x}_q - \bar{x}) \vec{1}_{n_q}^t \right] \begin{pmatrix} (\bar{x}_1 - \bar{x}) \vec{1}_{n_1} \\ \vdots \\ (\bar{x}_q - \bar{x}) \vec{1}_{n_q} \end{pmatrix} \\ &= \sum_{i=1}^q n_i (\bar{x}_i - \bar{x})^2 = \text{SCF}\end{aligned}$$

## Análisis de varianza

$$\begin{aligned} E[\text{SCE}] &= E[\vec{x}^t (I - D) \vec{x}] = E[\text{tr}(\vec{x}^t (I - D) \vec{x})] \\ &= E[\text{tr}([I - D] \vec{x} \vec{x}^t)] = \text{tr}\{E[(I - D) \vec{x} \vec{x}^t]\} \\ &= \text{tr}\{(I - D) E[\vec{x} \vec{x}^t]\} \\ &= \text{tr}\{(I - D) (\text{Cov}(\vec{x}) + E[\vec{x}] E[\vec{x}^t])\} \\ &= \text{tr}\{(I - D) (\sigma^2 I + E[\vec{x}] E[\vec{x}^t])\} \\ &= \sigma^2 \text{tr}(I - D) + \text{tr}\{(I - D) E[\vec{x}] E[\vec{x}^t]\} \\ &= \sigma^2 (n - q) + \text{tr}\{E[\vec{x}^t] (I - D) E[\vec{x}]\} \\ &= \sigma^2 (n - q) + E[\vec{x}^t] (I - D) E[\vec{x}] \end{aligned}$$

## Análisis de varianza

$$\mathbf{E}[\vec{x}] = \begin{pmatrix} \mathbf{E}[\vec{x}_1] \\ \mathbf{E}[\vec{x}_2] \\ \vdots \\ \mathbf{E}[\vec{x}_q] \end{pmatrix} = \begin{pmatrix} \mu_1 \vec{1}_{n_1} \\ \mu_2 \vec{1}_{n_2} \\ \vdots \\ \mu_q \vec{1}_{n_q} \end{pmatrix}$$

$$\begin{aligned} (I - D) \mathbf{E}[\vec{x}] &= \begin{pmatrix} I_{n_1} - J_1 & & & \\ & I_{n_2} - J_2 & & \\ & & \ddots & \\ & & & I_{n_q} - J_q \end{pmatrix} \begin{pmatrix} \mu_1 \vec{1}_{n_1} \\ \mu_2 \vec{1}_{n_2} \\ \vdots \\ \mu_q \vec{1}_{n_q} \end{pmatrix} \\ &= \begin{pmatrix} \vdots \\ (I_{n_i} - J_i) \mu_i \vec{1}_{n_i} \\ \vdots \end{pmatrix} \begin{pmatrix} \vdots \\ H_i \mu_i \vec{1}_{n_i} \\ \vdots \end{pmatrix} = \vec{0} \end{aligned}$$

$$\implies \mathbf{E}[\vec{x}^t] (I - D) \mathbf{E}[\vec{x}] = \mathbf{E}[\vec{x}^t] \vec{0} = 0 \implies \mathbf{E}[\text{SCE}] = \sigma^2 (n - q)$$



## Análisis de varianza

$$\begin{aligned} \text{E [SCF]} &= \text{E} [\vec{x}^t (D - J) \vec{x}] = \text{E} [\text{tr} \{ \vec{x}^t (D - J) \vec{x} \}] \\ &= \text{E} [\text{tr} \{ (D - J) \vec{x} \vec{x}^t \}] = \text{tr} \{ (D - J) \text{E} [\vec{x} \vec{x}^t] \} \\ &= \text{tr} \{ (D - J) (\text{Cov}(\vec{x}) + \text{E} [\vec{x}] \text{E} [\vec{x}^t]) \} \\ &= \text{tr} \{ (D - J) (\sigma^2 I + \text{E} [\vec{x}] \text{E} [\vec{x}^t]) \} \\ &= \text{tr} \{ \sigma^2 (D - J) \} + \text{tr} \{ (D - J) \text{E} [\vec{x}] \text{E} [\vec{x}^t] \} \\ &= \sigma^2 \text{tr}(D - J) + \text{tr} \{ \text{E} [\vec{x}^t] (D - J) \text{E} [\vec{x}] \} \\ &= \sigma^2 (q - 1) + \text{E} [\vec{x}^t] (D - J) \text{E} [\vec{x}] \end{aligned}$$

## Análisis de varianza

$$\mathbf{E}[\vec{x}] = \begin{pmatrix} \mathbf{E}[\vec{x}_1] \\ \vdots \\ \mathbf{E}[\vec{x}_q] \end{pmatrix} = \begin{pmatrix} \mu \vec{1}_{n_1} + \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \mu \vec{1}_{n_q} + \alpha_q \vec{1}_{n_q} \end{pmatrix} = \mu \vec{1}_n + \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix}$$

## Análisis de varianza

$$\begin{aligned}(D - J) E[\vec{x}] &= \mu (D - J) \vec{1} + (D - J) \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix} \\ &= \mu (D - J) \vec{1} + D \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix} - J \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix} \\ &= D \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix} = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_q \end{pmatrix} \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix} = \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix}\end{aligned}$$

puesto que...

## Análisis de varianza

$$\blacktriangleright (D - J) \vec{1} = \vec{0}$$

$$\blacktriangleright J \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix} = \frac{1}{n} \vec{1} \vec{1}^t \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix} = \frac{1}{n} \vec{1} \sum n_i \alpha_i = \vec{0}$$

Por tanto,

$$\begin{aligned} E[\vec{x}^t] (D - J) E[\vec{x}] &= \left( \alpha_1 \vec{1}_{n_1}^t, \dots, \alpha_q \vec{1}_{n_q}^t \right) \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix} \\ &= \sum_{i=1}^q \alpha_i^2 \vec{1}_{n_i}^t \vec{1}_{n_i} = \sum_{i=1}^q n_i \alpha_i^2 \end{aligned}$$

y se obtiene

$$E[\text{SCF}] = \sigma^2 (q - 1) + \sum n_i \alpha_i^2$$

# Análisis de varianza

Distribución de la SCE:

Suponiendo que  $\vec{x}_i \equiv N_{n_i}(\mu_i \vec{1}, \sigma^2 I_{n_i})$  para  $i = 1, \dots, q$  y que las  $\vec{x}_i$  son independientes, se verifica que

$$\text{SCE} = \vec{x}^t (I - D) \vec{x} = \sigma^2 \chi_{n-q}^2$$

Para demostrar esta propiedad, bastará ver que  $E[(I - D) \vec{x}] = \vec{0}$ .

# Análisis de varianza

$$\begin{aligned} \mathbf{E}[(I - D) \vec{x}] &= (I - D) \mathbf{E}[\vec{x}] = \mathbf{E}[\vec{x}] - D \mathbf{E}[\vec{x}] \\ &= \begin{pmatrix} \mu_1 \vec{1}_{n_1} \\ \vdots \\ \mu_q \vec{1}_{n_q} \end{pmatrix} - \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_q \end{pmatrix} \begin{pmatrix} \mu_1 \vec{1}_{n_1} \\ \vdots \\ \mu_q \vec{1}_{n_q} \end{pmatrix} \\ &= \begin{pmatrix} \mu_1 \vec{1}_{n_1} \\ \vdots \\ \mu_q \vec{1}_{n_q} \end{pmatrix} - \begin{pmatrix} \mu_1 J_1 \vec{1}_{n_1} \\ \vdots \\ \mu_q J_q \vec{1}_{n_q} \end{pmatrix} \\ &= \vec{0} \end{aligned}$$

## Análisis de varianza

Distribución de la SCF bajo  $H_0$ :

$$\text{SCF} = \vec{x}^t (D - J) \vec{x} \stackrel{H_0}{\equiv} \sigma^2 \chi_{q-1}^2$$

Dado que  $D - J$  es idempotente, bastará comprobar que  $E[(D - J) \vec{x}] = \vec{0}$ .

$$E[(D - J) \vec{x}] = (D - J) E[\vec{x}] = D E[\vec{x}] - J E[\vec{x}] = \vec{0}$$

ya que bajo  $H_0 \equiv \mu_1 = \dots = \mu_q$  y  $E[\vec{x}] = \mu \vec{1}$ , con lo cual

$$D \mu \vec{1} = \mu \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_q \end{pmatrix} \begin{pmatrix} \vec{1}_{n_1} \\ \vdots \\ \vec{1}_{n_q} \end{pmatrix} = \mu \begin{pmatrix} J_1 \vec{1}_{n_1} \\ \vdots \\ J_q \vec{1}_{n_q} \end{pmatrix} = \mu \vec{1}$$

$$J E[\vec{x}] = J \mu \vec{1} = \mu J \vec{1} = \mu \vec{1}$$

## Análisis de varianza

$$\text{CME} = \frac{\text{SCE}}{n - q}$$

Verifica que  $E[\text{CME}] = \sigma^2$  y por tanto es un estimador insesgado de la varianza de los residuos. Además,

$$\frac{\text{CME} (n - q)}{\sigma^2} \equiv \chi_{n-q}^2$$



## Análisis de varianza

$$\text{CMF} = \frac{\text{SCF}}{q - 1}$$

$$\text{E}[\text{CMF}] = \sigma^2 + \frac{\sum_{i=1}^q n_i \alpha_i^2}{q - 1}$$

Bajo  $H_0 \equiv \alpha_i = 0 \forall i$ , se cumple que  $\text{E}[\text{CMF} \mid H_0] = \sigma^2$ . Además

$$\frac{\text{CMF} (q - 1)}{\sigma^2} \stackrel{H_0}{\equiv} \chi_{q-1}^2$$

# Análisis de varianza

CMF y CME son independientes:

- ▶  $\vec{x} \equiv \mathcal{N}(\cdot, \cdot) \implies \begin{cases} (I - D)\vec{x} \equiv \mathcal{N}(\cdot, \cdot) \\ (D - J)\vec{x} \equiv \mathcal{N}(\cdot, \cdot) \end{cases}$
- ▶  $\text{Cov}[(I - D)\vec{x}, (D - J)\vec{x}] = (I - D)\text{Var}(\vec{x})(D - J) = (I - D)\sigma^2 I(D - J) = \sigma^2(I - D)(D - J) = \sigma^2(D - J - DD + DJ) = \sigma^2(D - J - D + J) = \mathbf{0}$
- ▶  $(I - D)\vec{x}$  independiente de  $(D - J)\vec{x}$
- ▶  $\vec{x}^t(I - D)\vec{x}$  independiente de  $\vec{x}^t(D - J)\vec{x}$
- ▶ SCE independiente de SCF

## Análisis de varianza

El cociente

$$\frac{\text{CMF}}{\text{CME}} = \frac{\text{CMF}/\sigma^2}{\text{CME}/\sigma^2}$$

tiende a tomar valores cercanos a uno bajo  $H_0$ , y más grandes bajo  $H_1$ .

Por otra parte, CMF y CME son independientes, luego

$$\frac{\text{CMF}}{\text{CME}} \stackrel{H_0}{\equiv} \frac{\frac{\chi_{q-1}^2}{q-1}}{\frac{\chi_{n-q}^2}{n-q}} = F_{q-1, n-q}$$

En consecuencia, la región crítica del contraste viene dada por la expresión

$$\text{R.C.} = \left\{ \frac{\text{CMF}}{\text{CME}} > k \right\} \text{ con } P[\text{R.C.} \mid H_0] = \alpha$$

# Análisis de varianza

Tabla ANOVA

Fuente de variación	S.C.	g.l.	C.M.	$F$
Entre / Factor	SCF	$q - 1$	CMF	$\frac{CMF}{CME}$
Dentro / Error	SCE	$n - q$	CME	
Total	SCT	$n - 1$		

## Ejemplo

Considérense los datos siguientes incluidos en R:

```
> aves <- data.frame (peso = chickwts$weight,  
                      come = factor(chickwts$feed,  
                                    labels =  
                                    c("caseína", "fabona", "linaza",  
                                      "har.hueso", "soja", "girasol")))  
  
> summary (aves)
```

	peso		come
Min.	:108.0	caseína	:12
1st Qu.	:204.5	fabona	:10
Median	:258.0	linaza	:12
Mean	:261.3	har.hueso	:11
3rd Qu.	:323.5	soja	:14
Max.	:423.0	girasol	:12

## Ejemplo

- ▶ Se repartió aleatoriamente en seis grupos una remesa de pollos recién nacidos.
- ▶ Cada grupo recibió un complemento alimenticio distinto.
- ▶ Se registró el peso en gramos tras seis semanas.
- ▶ ¿Influye el complemento en el peso?

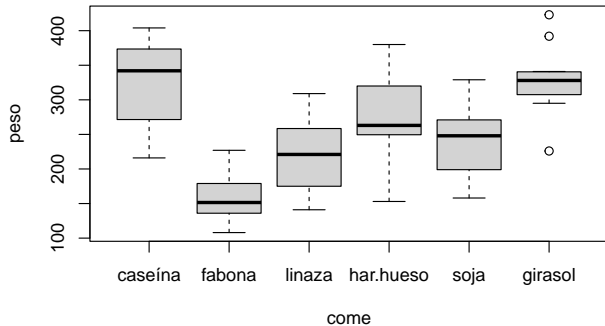
## Ejemplo

```
> options (digits = 3)
> RcmdrMisc::numSummary (aves$peso, groups=aves$come)
```

	mean	sd	IQR	0%	25%	50%	75%	100%	data:n
caseína	324	64.4	93.5	216	277	342	371	404	12
fabona	160	38.6	39.2	108	137	152	176	227	10
linaza	219	52.2	79.8	141	178	221	258	309	12
har.hueso	277	64.9	70.5	153	250	263	320	380	11
soja	246	54.1	63.2	158	207	248	270	329	14
girasol	329	48.8	27.5	226	313	328	340	423	12

# Ejemplo

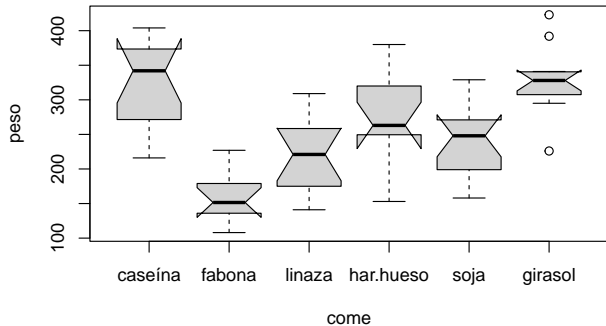
```
> boxplot (peso ~ come, aves)
```





# Ejemplo

```
> boxplot (peso ~ come, aves, notch=TRUE)
```



## Ejemplo

```
> bartlett.test (peso ~ come, aves)
```

```
    Bartlett test of homogeneity of variances
```

```
data:  peso by come
```

```
Bartlett's K-squared = 3, df = 5, p-value = 0.7
```

```
> car::leveneTest (peso ~ come, aves)
```

```
Levene's Test for Homogeneity of Variance (center = median)
```

	Df	F	value	Pr(>F)
group	5	0.75	0.59	
	65			

## Ejemplo

```
> options (width = 60)
> summary (aov (peso ~ come, aves))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
come	5	231129	46226	15.4	5.9e-10 ***
Residuals	65	195556	3009		

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Contrastes a posteriori

$$H_0^{\text{ANOVA}} : \mu_1 = \cdots = \mu_q \quad \equiv \quad \bigcap_{1 \leq i < j \leq q} H_0^{ij} : \mu_i = \mu_j$$

¿Qué  $H_0^{ij}$  se rechazan?

# Contrastes a posteriori

criterio de Bonferroni

- ▶ hay  $\frac{q(q-1)}{2}$  parejas de grupos
- ▶ contrastar  $H_0^{ij}$  a nivel  $\alpha^* = \frac{2\alpha}{q(q-1)}$

$$P \left[ \text{rechazar alguna } H_0^{ij} \mid H_0^{\text{ANOVA}} \right] \leq \sum_{1 \leq i < j \leq q} P \left[ \text{rechazar } H_0^{ij} \mid H_0^{ij} \right] = \frac{q(q-1)}{2} \alpha^* = \alpha$$

- ▶ estadígrafo de contraste

$$\frac{\bar{X}_i - \bar{X}_j}{\sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j}\right) \text{CME}}} \stackrel{H_0^{ij}}{=} t_{n-q}$$

# Contrastes a posteriori

En el ejemplo

```
> q <- length (levels (aves$come))
```

```
> q
```

```
[1] 6
```

```
> q * (q-1) / 2
```

```
[1] 15
```

## Contrastes a posteriori

```
> pairwise.t.test (aves$peso, aves$come, "none")
```

```
Pairwise comparisons using t tests with pooled SD
```

```
data:  aves$peso and aves$come
```

	caseína	fabona	linaza	har.hueso	soja
fabona	2e-09	-	-	-	-
linaza	1e-05	0.02	-	-	-
har.hueso	0.05	7e-06	0.01	-	-
soja	7e-04	3e-04	0.20	0.17	-
girasol	0.81	8e-10	6e-06	0.03	3e-04

```
P value adjustment method: none
```

## Contrastes a posteriori

```
> pairwise.t.test (aves$peso, aves$come, "bonferroni")
      Pairwise comparisons using t tests with pooled SD

data:  aves$peso and aves$come

      caseína fabona linaza har.hueso soja
fabona 3e-08  -      -      -      -
linaza 2e-04  0.228 -      -      -
har.hueso 0.684 1e-04 0.202 -      -
soja    0.010  0.005 1.000 1.000 -
girasol 1.000  1e-08 9e-05 0.397 0.004

P value adjustment method: bonferroni
```



# Contrastes a posteriori

criterio de Tukey

- ▶ basado en la distribución del *rango estudentizado*
  - ▶  $Y_1, \dots, Y_q \equiv N(0, 1)$  independientes
  - ▶  $Z \equiv \chi_r^2$  independiente de las  $Y_1, \dots, Y_q$
  - ▶ entonces  $\frac{Y_{(q)} - Y_{(1)}}{\sqrt{Z/r}} \equiv Q_{q,r}$  `ptukey(,q,r)` en R
- ▶ en ANOVA para contrastar  $H_0^{ij}$  se calcula el P-valor

$$P \left[ Q_{q,n-q} > \frac{|\bar{X}_i - \bar{X}_j|}{\sqrt{\text{CME} \left( \frac{1}{n_i} + \frac{1}{n_j} \right)}} \right]$$

## Contrastes a posteriori

```
> r <- nrow (aves) - q                # q=6 r=65
> distro <- replicate (1e5,
  {
    medias <- rnorm (q)
    numerador <- diff (range (medias))
    denominador <- sqrt (rchisq (1, r) / r)
    numerador / denominador
  })
> alfas <- c (0.01, 0.025, 0.05, 0.1, 0.5,
              0.9, 0.95, 0.975, 0.99)
> rbind (sim = quantile (distro, alfas),
         num = qtkey (alfas, q, r))
```

	1%	2.5%	5%	10%	50%	90%	95%	97.5%	99%
sim	0.861	1.06	1.25	1.48	2.49	3.76	4.16	4.54	5.00
num	0.862	1.06	1.24	1.48	2.49	3.75	4.15	4.52	4.97

# Contrastes a posteriori

```
> a <- aov (peso ~ come, aves)
> TukeyHSD (a)
```

## Contrastes a posteriori

Tukey multiple comparisons of means  
95% family-wise confidence level

Fit: aov(formula = peso ~ come, data = aves)

\$come

	diff	lwr	upr	p	adj
fabona-caseína	-163.38	-232.35	-94.4	0.000	
linaza-caseína	-104.83	-170.59	-39.1	0.000	
har.hueso-caseína	-46.67	-113.91	20.6	0.332	
soja-caseína	-77.15	-140.52	-13.8	0.008	
girasol-caseína	5.33	-60.42	71.1	1.000	
linaza-fabona	58.55	-10.41	127.5	0.141	
har.hueso-fabona	116.71	46.34	187.1	0.000	
soja-fabona	86.23	19.54	152.9	0.004	
girasol-fabona	168.72	99.75	237.7	0.000	
har.hueso-linaza	58.16	-9.07	125.4	0.128	
soja-linaza	27.68	-35.68	91.0	0.793	

## Contrastes a posteriori

```
> set.seed (122)
> mu <- c (0, 0, 0.2)
> q <- length(mu)
> m <- 100
> g <- factor (rep (1:q, each=m))
> x <- unlist (lapply (mu,
                      function (mui) rnorm (m, mui)))
> a <- aov (x ~ g)
> summary (a)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
g	2	9	4.29	3.8	0.024 *
Residuals	297	336	1.13		

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Contrastes a posteriori

```
> options (digits = 7)
> pairwise.t.test (x, g, "bonf")
```

Pairwise comparisons using t tests with pooled SD

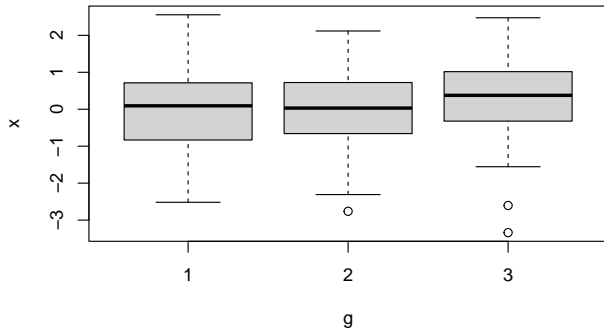
```
data:  x and g
```

```
   1      2
2 1.000 -
3 0.056 0.050
```

P value adjustment method: bonferroni

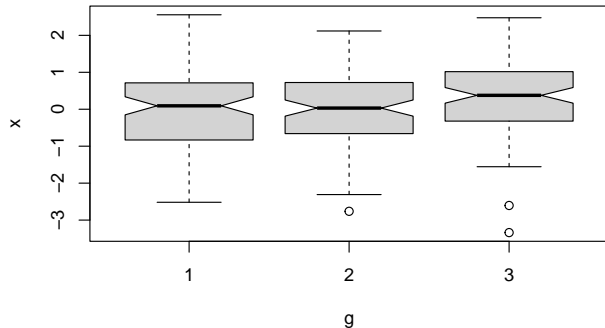
# Contrastes a posteriori

```
> boxplot (x ~ g)
```



# Contrastes a posteriori

```
> boxplot(x ~ g, notch=TRUE)
```





## Contrastes a posteriori

```
> set.seed (614)
> mu <- c (0, 0, 0.2)
> q <- length(mu)
> m <- 100
> g <- factor (rep (1:q, each=m))
> x <- unlist (lapply (mu,
                      function (mui) rnorm (m, mui)))
> a <- aov (x ~ g)
> summary (a)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
g	2	5.99	2.995	2.955	0.0536 .
Residuals	297	301.09	1.014		

---

Signif. codes:  
0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Contrastes a posteriori

```
> pairwise.t.test (x, g, "bonf")
```

```
Pairwise comparisons using t tests with pooled SD
```

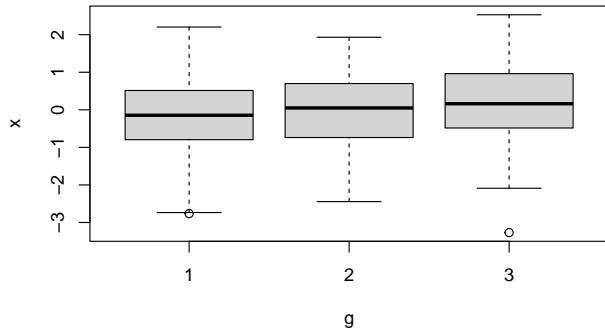
```
data: x and g
```

```
  1      2  
2 0.546 -  
3 0.048 0.831
```

```
P value adjustment method: bonferroni
```

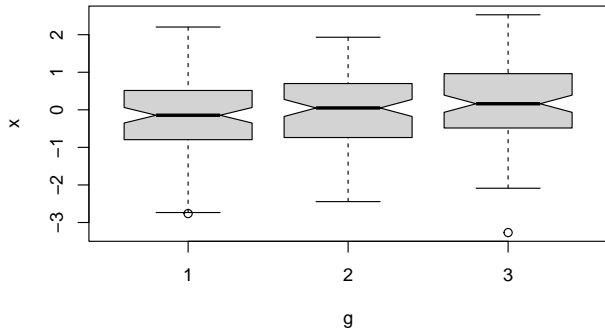
# Contrastes a posteriori

```
> boxplot (x ~ g)
```



# Contrastes a posteriori

```
> boxplot(x ~ g, notch=TRUE)
```



## Contrastes a posteriori

```
> alfa <- 0.05 ; mu <- c (0, 0, 0.2) ; q <- length(mu)
> m <- 100 ; g <- factor (rep (1:q, each=m))
> table (data.frame (t (replicate (10000, {
  x <- unlist (lapply (mu,
    function (mui) rnorm (m, mui)))
  a <- aov (x ~ g)
  H1anova <- summary(a) [[1]] ["g", "Pr(>F)"] < alfa
  H1bonfe <- any (na.omit (c (pairwise.t.test
    (x,g,"bonf")$p.value))
    < alfa)
  c (H1anova=H1anova, H1bonfe=H1bonfe)
}))))
```

```
      H1bonfe
H1anova FALSE TRUE
FALSE   7099   37
TRUE    308 2556
```

## Contrastes a posteriori

```
> rechazos <- sapply (1:1000, function(semilla){
  set.seed (semilla)
  x <- unlist (lapply (mu,
    function (mui) rnorm (m, mui)))
  a <- aov (x ~ g)
  H1anova <- summary(a) [[1]] ["g", "Pr(>F)"] < alfa
  H1bonfe <- any (na.omit (c (pairwise.t.test
    (x,g,"bonf")$p.value))
    < alfa)
  c (H1anova=H1anova, H1bonfe=H1bonfe)
})
```

## Contrastes a posteriori

```
> options (width = 55)
> (semillas <- which (xor (rechazos[1,], rechazos[2,])))
 [1] 122 129 153 218 226 246 267 268 343 369 371 377
 [13] 383 397 414 437 457 513 610 614 619 646 654 658
 [25] 670 677 706 741 744 762 777 779 793 895 906 930
 [37] 973
> rechazos [, semillas]
      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
H1anova TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
H1bonfe FALSE FALSE FALSE FALSE FALSE FALSE FALSE
      [,8] [,9] [,10] [,11] [,12] [,13] [,14]
H1anova TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
H1bonfe FALSE FALSE FALSE FALSE FALSE FALSE FALSE
      [,15] [,16] [,17] [,18] [,19] [,20] [,21]
H1anova TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
H1bonfe FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
      [,22] [,23] [,24] [,25] [,26] [,27] [,28]
H1anova TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```