

ANOVA

análisis de varianza

Carlos Carleos, Norberto Corral

20 de febrero de 2023

1. 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 $\mu = \frac{\sum n_i \mu_i}{n}$
- $\sum_{i=1}^q n_i \alpha_i = 0$
- Si $\forall i, n_i = \frac{n}{q}$ el modelo se dice *equilibrado* o *balanceado*

Hipótesis previas

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

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

Hipótesis del análisis de varianza (ANOVA)

$$\begin{aligned} H_0 &\equiv \forall i, j, \mu_i = \mu_j \equiv \forall i, \alpha_i = 0 \\ H_1 &\equiv \exists i, j, \mu_i \neq \mu_j \equiv \exists i, \alpha_i \neq 0 \end{aligned}$$

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

2. Repaso de álgebra lineal

- Matriz simétrica $A = [a_{ij}]_{i,j}$ si $a_{ij} = a_{ji} \iff A = A^t$
 - 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 $\vec{x} \neq \vec{0}$ 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: $A = [a_{ij}]_{i,j} \implies \text{tr} A = \sum_i a_{ii}$
 - $a = \text{tr}[a]$
 - $\text{tr}(A B) = \text{tr}(B A)$
 - $\text{tr}(A B C) = \text{tr}(B C A) = \text{tr}(C A B)$
 - $\text{E}[\text{tr} A] = \text{tr} \text{E}[A]$

3. Formas cuadráticas

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

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

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

$$= \mathbf{E} [\vec{x} \vec{x}^t] - \mathbf{E} [\vec{x}] \mathbf{E} [\vec{x}]^t$$

Dados una matriz A (n, n) simétrica e idempotente con $\text{rango}(A) = r$ y un vector aleatorio $\vec{x} \equiv \mathcal{N}_n(\vec{\mu}, \sigma^2 I_n)$, 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 $U^t U = I_n$
- $\vec{x}^t A \vec{x} = \vec{x}^t U_1 U_1^t \vec{x}$
- esperanza

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

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

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

$$\begin{aligned} \vec{\gamma} &= \mathbf{E} [\vec{y}_1] = \mathbf{E} [U_1^t \vec{x}] = U_1^t \mathbf{E} [\vec{x}] = U_1^t \vec{\mu} \\ U_1 \vec{\gamma} &= U_1 U_1^t \vec{\mu} = A \vec{\mu} = \vec{0} \\ \vec{\gamma} &= I_r \vec{\gamma} = U_1^t U_1 \vec{\gamma} = U_1^t \vec{0} = \vec{0} \end{aligned}$$

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

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

$$\vec{y}_1^t \vec{y}_1 = \sum_{j=1}^r y_{1j}^2 = \sum_{j=1}^r \sigma^2 \left(\frac{y_{1j}}{\sigma} \right)^2 \equiv \sigma^2 \sum_{j=1}^r \underbrace{[\mathcal{N}(0, 1)]^2}_{\text{indep.}} = \sigma^2 \chi_r^2$$

4. 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 $i = 1, \dots, q$

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

J es simétrica e idempotente

al multiplicar, genera un vector columna con la media:

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

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

H es simétrica e idempotente

H es la matriz de centrado

al multiplicar, sustrae la media:

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

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

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

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
- $\text{tr}(D - J) = \text{tr} D - \text{tr} J = q - 1 = \text{rango}(D - J)$

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

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

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

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

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

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

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

```

summary(sleep)                                # ejemplo en R

## Vamos a suponer (incorrectamente; véase la ayuda
## "?sleep") que "sleep" contiene una variable
## respuesta que representa el tiempo "extra" de
## sueño en un grupo de control (group=1) de cierto
## experimento y en otro grupo de personas a las que
## se les ha administrado un medicamento (group=2).

X <- sleep$extra                                # respuesta
N <- length(X)                                  # tamaño muestral
G <- sleep$group                                # factor

## H0: mu1=mu2 ; H1: mu1<>mu2
t.test (X ~ G, var.equal=TRUE)                 # =aov <= q=2

adeva <- aov(X~G)
I <- diag(N)
definirJ <- function (n) matrix(1/n, n, n)
J <- definirJ(N)
SCT <- X %*% (I-J) %*% X
tabla <- summary(adeva)[[1]]
sum(tabla[, "Sum Sq"])                          # SCT
stopifnot (identical (G, sort(G)))
Ni <- table(G)
D <- as.matrix(Matrix::bdiag(lapply(Ni, definirJ)))
SCE <- X %*% (I-D) %*% X
tabla["Residuals", "Sum Sq"]                   # SCE
SCF <- X %*% (D-J) %*% X
tabla[1, "Sum Sq"]                             # SCF

```

A simétrica idempotente de rango $r \implies E[\bar{x}^t A \bar{x}] = \sigma^2 r + \bar{\mu} A \bar{\mu} \implies$

$E[\text{SCE}] = E[\vec{x}^t (I - D) \vec{x}] = \sigma^2 (n - q) + \vec{\mu} (I - D) \vec{\mu}$ En detalle:

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

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

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

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

A simétrica idempotente de rango $r \implies E[\vec{x}^t A \vec{x}] = \sigma^2 r + \vec{\mu} A \vec{\mu} \implies$
 $E[\text{SCF}] = E[\vec{x}^t (D - J) \vec{x}] = \sigma^2 (q - 1) + \vec{\mu} (D - J) \vec{\mu}$ En detalle:

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

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

$$\begin{aligned}
J \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix} &= \frac{1}{n} \vec{1} \vec{1}^t \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix} = \vec{1} \frac{1}{n} \sum n_i \alpha_i = \vec{0} \\
\Rightarrow (D - J) E[\vec{x}] &= \mu (D - J) \vec{1} + (D - J) \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix} \\
&= \vec{0} + D \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix} - J \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix} \\
&= D \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix} - \vec{0} = \begin{pmatrix} J_1 & & \\ & J_2 & \\ & & \ddots \\ & & & J_q \end{pmatrix} \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix} = \begin{pmatrix} \alpha_1 \vec{1}_{n_1} \\ \vdots \\ \alpha_q \vec{1}_{n_q} \end{pmatrix}
\end{aligned}$$

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

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

4.1. Distribución de la SCE

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

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

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

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

4.2. Distribución de la SCF bajo H_0

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

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

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

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

$$\begin{aligned}
D\mu \vec{1} &= \mu \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_q \end{pmatrix} \begin{pmatrix} \vec{1}_{n_1} \\ \vdots \\ \vec{1}_{n_q} \end{pmatrix} = \mu \begin{pmatrix} J_1 \vec{1}_{n_1} \\ \vdots \\ J_q \vec{1}_{n_q} \end{pmatrix} = \mu \vec{1} \\
J\mathbb{E}[\bar{x}] &= J\mu \vec{1} = \mu J \vec{1} = \mu \vec{1}
\end{aligned}$$

5. Estadístico del contraste

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

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

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

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

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

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

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

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

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

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

Tabla ANOVA

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

6. 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)
```

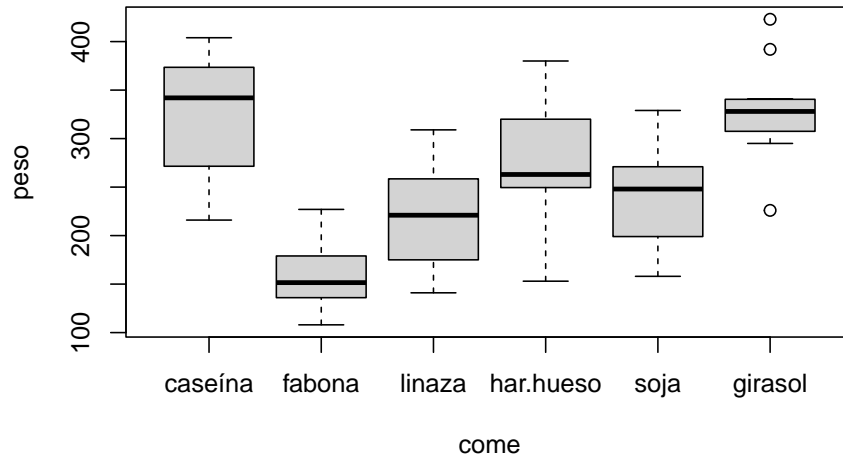
peso		come	
Min.	:108.0	caseína	:12
1st Qu.	:204.5	fabona	:10
Median	:258.0	linaza	:12
Mean	:261.3	har.hueso	:11
3rd Qu.	:323.5	soja	:14
Max.	:423.0	girasol	:12

- 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?

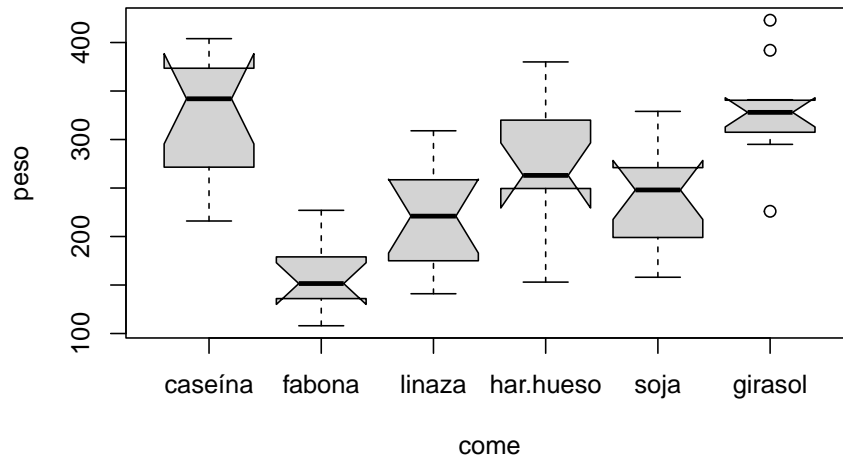
```
> options (digits = 3)
> RcmdrMisc::numSummary (aves$peso, groups=aves$come)
```

	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

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



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



```
> 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
```

```
> 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

7. Contrastes a posteriori

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

¿Qué H_0^{ij} se rechazan?

7.1. Criterio de Bonferroni

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

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

- estadístico de contraste

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

En el ejemplo

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

[1] 6

> q * (q-1) / 2

[1] 15

> 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

```
> ## 12 de 15 rechazos al 5%
```

```
> 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

```
> ## 8 de 15 rechazos al 5%
```

7.2. Criterio de Tukey

- basado en la distribución del *recorrido estudentizado*
 - $Y_1, \dots, Y_q \equiv \mathcal{N}(0, 1)$ 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 \mathbb{R}
- en ANOVA para contrastar H_0^{ij} se calcula el P-valor

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

```
> 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	0.862	1.06	1.24	1.48	2.48	3.74	4.15	4.51	4.96
num	0.862	1.06	1.24	1.48	2.49	3.75	4.15	4.52	4.97

```
> a <- aov (peso ~ come, aves)
> TukeyHSD (a)           # 8 rechazos de 15 al 5%
```

```
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
girasol-linaza   110.17  44.41 175.9 0.000
soja-har.hueso  -30.48  -95.38 34.4 0.739
girasol-har.hueso 52.01  -15.22 119.2 0.221
girasol-soja     82.49   19.13 145.9 0.004

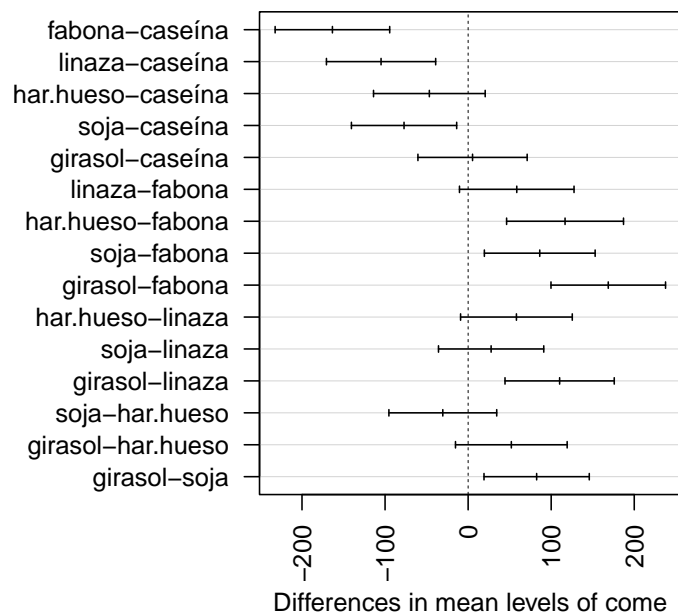
```

```

> m <- par("mar") ; par(mar=m+c(0,10,0,0)) # aumenta margen izq.
> plot(TukeyHSD(a), las=2) # etiquetas horizontales (?par)
> par(mar=m) # recupera margen izq. habitual

```

95% family-wise confidence level



7.3. Incoherencias

Es posible que se produzcan resultados contradictorios, es decir, que el ANOVA rechace su H_0 mientras que ningún contraste a posteriori por parejas dé resultado significativo, o viceversa. Veamos un par de ejemplos.

```
> set.seed(129) # anova => H1 ; bonferroni => H0
```

```

> 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	7	3.35	3.11	0.046 *
Residuals	297	320	1.08		

```

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

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

```

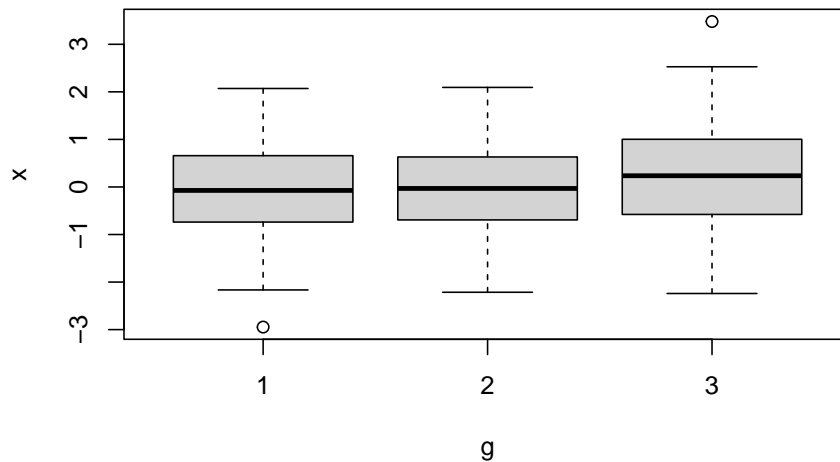
Pairwise comparisons using t tests with pooled SD

data: x and g

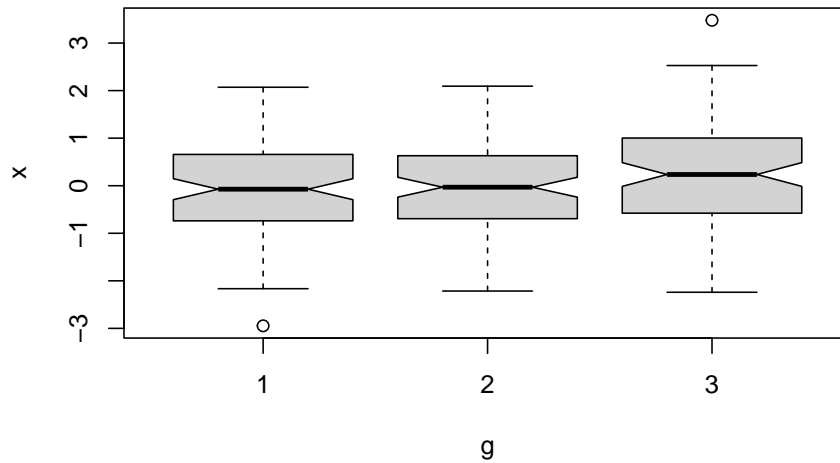
1	2
2	1.000 -
3	0.063 0.157

P value adjustment method: bonferroni

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



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



```
> set.seed (614)      # anova => H0 ; bonferroni => H1
> 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
```

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

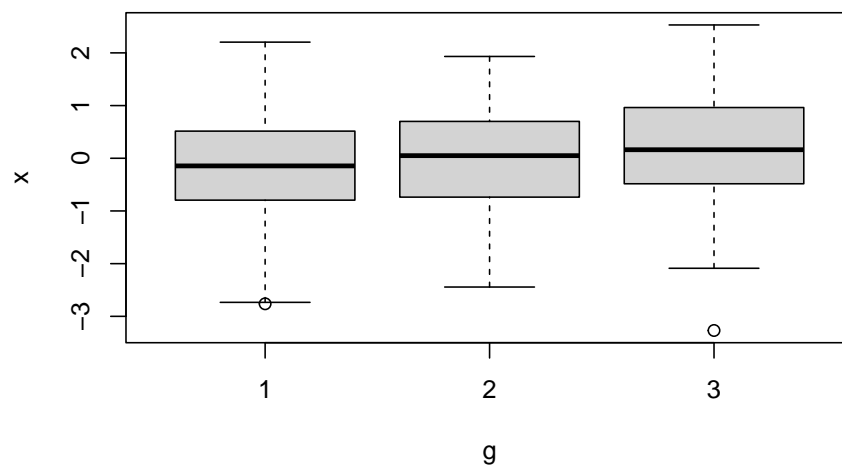
Pairwise comparisons using t tests with pooled SD

data: x and g

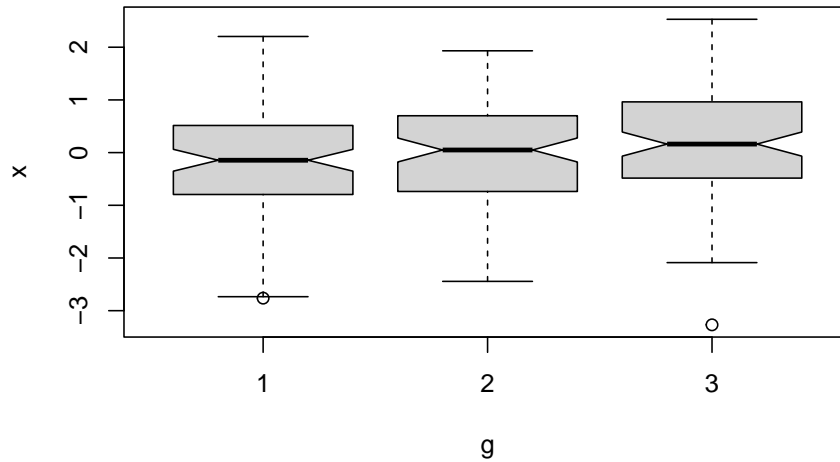
```
1 2
2 0.546 -
3 0.048 0.831
```

P value adjustment method: bonferroni

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



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



```

> 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

```

7.4. Apéndice: cómo hallar la semillas

Para encontrar muestras que produzcan los resultados anómalos, primero guardamos toda la distribución simulada:

```

> 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)
})

```

A continuación, buscamos qué columnas contienen TRUE-FALSE o FALSE-TRUE:

```

> 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
H1bonfe	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
	[,29]	[,30]	[,31]	[,32]	[,33]	[,34]	[,35]
H1anova	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
H1bonfe	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
	[,36]	[,37]					
H1anova	TRUE	TRUE					
H1bonfe	FALSE	FALSE					