15.ANOVA

Estadística
Grau en Matemàtiques

Josep A. Sanchez
Dept. Estadística i I.O.(UPC)

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Anàlisi de la variància

Volem quantificar la influència en el temps total de procès segons:

- 2 algoritmes de programació de CPUs
- 3 arquitectures de CPUs
- 3 tipus de càrregues de feina en una certa mètrica

Aquestes variables són **factors** (variables explicatives/predictores) de tipus categòric, que poden prendre un limitat nombre de valors (**nivells**)

L'objectiu de l'ANOVA (*ANalysis Of VAriance*) és modelitzar l'impacte de cada factor en la variable resposta i determinar si existeixen diferències significatives entre els nivells d'un factor donat

ANOVA d'un factor

El model més simple considera un únic factor com a variable explicativa, i s'anomena **ANOVA d'un factor**

Exemple: Volem comparar el temps d'execució basat en diferents tipus de CPUs. Suposem que el nombre de nivells del factor és a i que pel nivell i-éssim hi ha n_i observacions.

Nivell	Observacions	Mitjanes
1	$y_{11}, y_{12}, \ldots, y_{1n_1}$	$\bar{y}_{1.}$
2	$y_{21}, y_{22}, \ldots, y_{2n_2}$	$\bar{y}_{2.}$
$\vdots$		
a	$y_{a1}, y_{a2}, \ldots, y_{an_a}$	$\bar{y}_{a.}$

on $N = \sum_{i=1}^{a} n_i, \quad \bar{y}_{i.} = \sum_{j=1}^{n_i} y_{ij}/n_i, \quad \bar{y}_{..} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}/N$

ANOVA d'un factor

El valor de n_i correspon al número de **rèpliques** de cada nivell

Si $n_i = n \quad \forall i \in \{1, \ldots, a\}$ es diu que el model està **balancejat**. En aquest cas,

$$\bar{y}_{..} = \frac{1}{a} \sum_{i=1}^{a} \bar{y}_{i.}$$

En el cas no balancejat, la mitjana global és una mitjana ponderada de les mitjanes per cada nivell:

$$\bar{y}_{..} = \sum_{i=1}^{a} \frac{n_i}{N} \bar{y}_{i.}$$

L'expressió com a model lineal seria:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \qquad i = 1, \ldots, a, \quad j = 1, \ldots, n_i$$

on

- $\mu_i = \mu + \tau_i$ és el valor esperat de la resposta pel nivell i
- Els errors, ϵ_{ij} són indep. amb distribució $N(0, \sigma^2)$

ANOVA d'un factor

Objectiu: establir si hi ha diferències significatives pel valor esperat de la resposta segons els nivells del factor:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_a \quad vs. \quad H_1 : \exists(i,j) \quad \mu_i \neq \mu_j$$

fixat el nivell de significació α

Aquest test és equivalent a:

$$H_0 : \tau_i = 0 \quad \forall i \quad vs. \quad H_1 : \exists i \quad \tau_i \neq 0$$

Observació: per $a = 2$, el problema és equivalent a la comparació de les mitjanes de dues poblacions independents, amb variància desconeguda

Observació: No és correcte el procediment que no rebutja H_0 si s'han fet tots els test de comparacions 2 a 2 i no s'ha rebutjat en cap cas

Això és degut a que:

$$\alpha^* = P(\text{rebutjar } H_0 | H_0 \text{ és certa}) = 1 - (1-\alpha)^{a(a-1)/2} \geq \alpha$$

Per exemple, si a=5 i $\alpha = 0.05$, fent servir el procediment descrit tenim $P(\text{rebutjar } H_0 | H_0 \text{ és certa}) = 0.4$

ANOVA d'un factor

La variabilitat total de les dades (SS_{Total}) es pot sperara en dues parts, una deguda als diferents nivells del factor i un altra deguda a l'error:

Com $(y_{ij} - \bar{y}_{..}) = (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$ si fem el quadrat de l'expressió i verifiquem que el doble producte és zero, tindrem:

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{a} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$$

Habitualment, aquesta descomposició de sumes de quadrats és denota per:

$$SS_{Total} = SS_{Factor} + SS_{Error}$$

També es coneixen com:

- SS_{Factor}: Variació entre nivells (*between*) o deguda al factor
- SS_{Error}: Variació dintre de nivells (*within*) o residual

Els graus de llibertat d'aquestes sumes són respectivament, $N-1$,$a-1$ i $N-a$

ANOVA d'un factor

La taula de l'ANOVA d'un factor és:

	SS	d. f.	MSE	F
Factor A	SS_A	$a-1$	$\frac{SS_A}{a-1}$	$F = \frac{SS_A/(a-1)}{SS_E/(N-a)}$
Error	SS_E	$N-a$	$\frac{SS_E}{N-a}$	
Total	SS_T	$N-1$	$\frac{SS_T}{N-1}$	

i es verifica que...

- ...sota H_0, $SS_A/\sigma^2 \sim \chi^2_{a-1}$
- ...sempre $SS_E/\sigma^2 \sim \chi^2_{N-a}$
- ... SS_A i SS_E són independents

Per tant, la distribució de referència de l'estadístic F és una $F_{a-1,N-a}$. Així, el procediment per resoldre el test té per regió crítica:

$$\text{rebutjar } H_0 \quad \text{si } F \geq F_{1-\alpha,a-1,N-a}$$

Observacions:

- És un test unilateral, perquè quan H_0 no és certa, el numerador és significativament més gran que el denominador
- Este test generalitza el test de la t-Student de comparació de dues mitjanes a més de dues mitjanes, sota normalitat, homoscedasticitat i variàncies desconegudes

ANOVA d'un factor

Expressió del model lineal lligat a l'ANOVA d'un factor

Si fem servir variables indicadores ($X_i = \mathbb{I}_{\{A=A_i\}}$, on A_i és el nivell i-éssim del factor A)

$$y_{ij} = \mu + \tau_1 \mathbb{I}_{\{A=A_1\}} + \ldots \tau_a \mathbb{I}_{\{A=A_a\}} + \epsilon_{ij} \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

Matriu de disseny completa

A	$\mathbb{I}$	X_1	X_2	...	X_a
A_1	1	1	0	...	0
A_1	1	1	0	...	0
.	.	.	.		.
.	.	.	.		.
.	.	.	.		.
A_2	1	0	1	...	0
A_2	1	0	1	...	0
.	.	.	.		.
.	.	.	.		.
.	.	.	.		.
A_a	1	0	0	...	1

En forma matricial,

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{a1} \\ \vdots \\ Y_{an_a} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_a \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2n_2} \\ \vdots \\ \epsilon_{a1} \\ \vdots \\ \epsilon_{an_a} \end{pmatrix}$$

Multicol · linealitat: La matriu de dissney X no té rang màxim, ja que les seves columnes són linealment dependents (la primera columna és la suma de la resta de columnes)

Les equacions normals tenen infinites solucions, ja que per resoldre-les hem de fer servir una inversa generalitzada:

$$\hat{\beta} = (X'X)^{-}X'y$$

on $\hat{\beta} = (\hat{\mu}, \hat{\tau}_1, \ldots, \hat{\tau}_a)$ i $(X'X)^{-}$ representa una inversa generalitzada

Per obetnir una solució única de l'estimació del model hem d'introduir restriccions en forma de **contrastos**

Els més comuns:

- Restriccions de tipus **baseline** (categoria de referència): $\tau_1 = 0$ o $\tau_a = 0$
- Restricció de **suma zero**: $\sum_{i=1}^{a} \tau_i = 0$

ANOVA d'un factor

- Imposar una restricció de tipus *baseline*, suposa eliminar de la matriu de disseny la columna associada a la categoria de referència. Amb això aconseguim que sigui de rang màxim.
- Pel mateix model, l'estimació i interpretació canvia segons el contrast aplicat
- Le prediccions ($\hat{y}_{ij} = \hat{\mu}_i = \hat{\mu} + \hat{\tau}_i$) són les mateixes, independent del contrast actiu
- En l'ANOVA d'un factor,l'estimació $\hat{\mu}_i - \hat{\mu}_j$ tampoc depén del contrast

ANOVA d'un factor

Interpretació de paràmetres:

- Contrast *treatment/baseline*:
  - $\hat{\mu}$ Valor esperat de la resposta per a la categoria de referència

$$\hat{\mu} = \bar{y}_{1.}$$

  - $\hat{\tau}_i$ Canvi en el valor esperat de la resposta si enlloc de la categoria de referència considerem la i-èssima categoria

$$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{1.}$$

- Contrast de suma zero:
  - $\hat{\mu}$ Valor esperat de la resposta per tots els individus, sense indicar el grup (mitjana global)

$$\hat{\mu} = \bar{y}_{..}$$

  - $\hat{\tau}_i$ Canvi en el valor esperat de la resposta si enlloc de la mitjana global considerem la i-èssima categoria

$$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}$$

ANOVA d'un factor

Predicció (independentment del contrast):

Predicció pel nivell i: $\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i = \bar{y}_{i.}$

Predicció per la diferència de dos nivells: $\hat{\mu}_i - \hat{\mu}_j = \hat{\tau}_i - \hat{\tau}_j = \bar{y}_{i.} - \bar{y}_{j.}$

Si el test de l'ANOVA no rebutja H_0 la conclusió pràctica és que no s'han trobat diferències significatives en els valors esperats de la resposta per al diferents grups. D'altra banda, si rebutgem H_0 cal explorar on s'han trobat les diferències.

Comparancions múltiples: Es fan les comparacions dels nivells 2 a 2 ($H_0 : \mu_i = \mu_j$ vs. $H_1 : \mu_i \neq \mu_j$)

- **Mètode de Tukey**
  - Ordenem les mitjanes en ordre creixent
  - Calculem $S^2_{\bar{y}_k} = \frac{SS_{Error}}{(N-a)n_h}$ on $n_h = a/(1/n_1 + 1/n_2 + \cdots + 1/n_a)$ és la mitjana harmònica de les mides mostrals
  - Rebutgem H_0 quan:

$$|\bar{y}_{i.} - \bar{y}_{j.}| \geq q_\alpha(a, N-a) S_{\bar{y}_k}$$

on $q_\alpha(a, N-a)$ és la quantila α en la distribució dels **Rangs de Student** (tabulada)

ANOVA d'un factor

Validació dels supòsits (anàlisi dels residus):

- Plot de quantiles-quantiles Normals per verificar la normalitat dels residus
- Plot dels residus respecte a les prediccions per verificar que no hi ha tendències i que la variància és constant
- Plot del residus respecte als nivells del factor, on la distribució de residus ha de ser similar en tots els grups

ANOVA de dos factors

Exemple: volem comparar el **temps d'execució** bast en a diferents **càrregues de treball**(X_1) i b diferents **CPUs** (X_2)

Factors A i B tenen a and b nivells respect. i per a cada combinació (i, j) tenim n_{ij} observacions.

Cas Balancejat $n_{ij} = n \qquad \forall (i, j)$

A | B	1	2	$\cdots$	b	
1	$y_{111} \cdots y_{11n}$	$y_{121} \cdots y_{12n}$	$\cdots$	$y_{1b1} \cdots y_{1bn}$	$\overline{y}_{1..}$
2	$y_{211} \cdots y_{21n}$	$y_{221} \cdots y_{22n}$	$\cdots$	$y_{2b1} \cdots y_{2bn}$	$\overline{y}_{2..}$
$\vdots$					
a	$y_{a11} \cdots y_{a1n}$	$y_{a21} \cdots y_{a2n}$	$\cdots$	$y_{ab1} \cdots y_{abn}$	$\overline{y}_{a..}$
	$\overline{y}_{.1.}$	$\overline{y}_{.2.}$		$\overline{y}_{.b.}$	$\overline{y}_{...}$

ANOVA de dos factors

ANOVA de dos factors additiu:

$$y_{ijk} = \mu + \tau_i + \beta_j + e_{ijk}$$

Supòsits:

- $e_{ijk} \sim \mathrm{N}(0, \sigma^2)$,
- errors independents
- $\sum_i \tau_i = 0$ $\sum_j \beta_j = 0$ /, $\tau_1 = 0$ i $\beta_1 = 0$ / $\tau_a = 0$ i $\beta_b = 0$.

Observacions:

- Es possible tenir $n = 1$ (una rèplica per condició experimental)
- Si els efectes dels factors són multiplicatius, s'ha de considerar $log(Y)$ com a resposta

$$\mu_{ijk} = \mu \cdot \tau_i \cdot \beta_j \iff \log(\mu_{ijk}) = \log(\mu) + \log(\tau_i) + \log(\beta_j)$$

ANOVA de dos factors

Objectiu: establir si existeixen diferències signifactives entre els nivells de cada factor

Volem testar, per un valor fixat de α,

$$H_0^1 : \tau_1 = \tau_2 = \cdots = \tau_a = 0 \ \ vs \ \ H_1^1 : \exists i \qquad \tau_i \neq 0,$$

$$H_0^2 : \beta_1 = \beta_2 = \cdots = \beta_b = 0 \ \ vs \ \ H_1^2 : \exists i \qquad \beta_i \neq 0,$$

ANOVA de dos factors

La taula de l'ANOVA de dos factors additius és:

	SS	d. f.	MSE	F
Factor A	SS_A	$a-1$	$\frac{SS_A}{a-1}$	$F_0^1 = \frac{SS_A/(a-1)}{SS_E/(N-(a+b-1))}$
Factor B	SS_B	$b-1$	$\frac{SS_B}{b-1}$	$F_0^2 = \frac{SS_B/(b-1)}{SS_E/(N-(a+b-1))}$
Error	SS_E	$N-a-b+1$	$\frac{SS_E}{N-a-b+1}$	
Total	SS_T	$N-1$	$\frac{SS_T}{N-1}$	

Regles de decisió:

$$\text{rebutjar } H_0^1 \quad \text{si} \quad F_0^1 \geq F_{1-\alpha,a-1,N-(a+b-1)}$$

$$\text{rebutjar } H_0^2 \quad \text{si} \quad F_0^2 \geq F_{1-\alpha,b-1,N-(a+b-1)}$$

ANOVA de dos factors

Si es pensa que l'efecte d'un factor depén del nivell de l'altre factor amb el que es combina, s'ha de considerar el **model amb interacció** (no additiu)

$$y_{ijk} = \mu + \tau_i + \beta_j + (\gamma)_{ij} + e_{ijk}$$

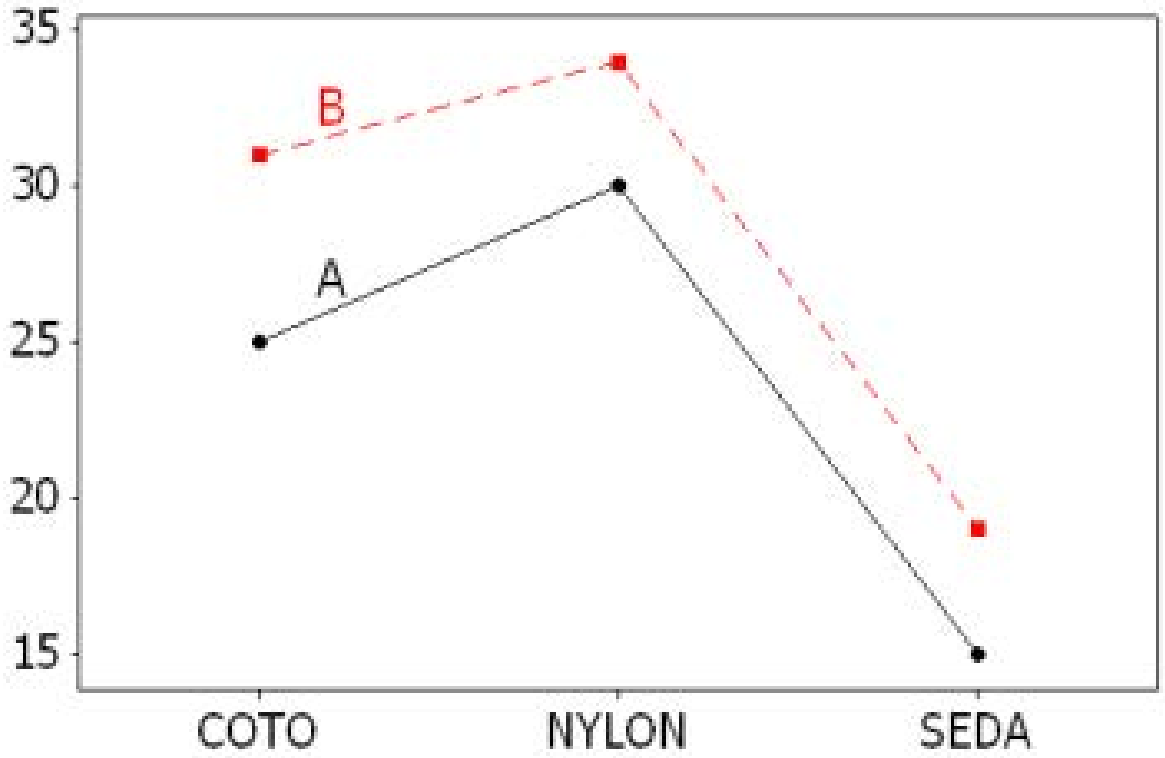
amb els supòsits:

- $e_{ijk} \sim \mathrm{N}(0, \sigma^2)$,
- errors independents
- $\sum_i \tau_i = 0,\ \sum_j \beta_j = 0, \forall\, i,\ \sum_j \gamma_{ij} = 0, \forall\, j,\ \sum_i \gamma_{ij} = 0,$

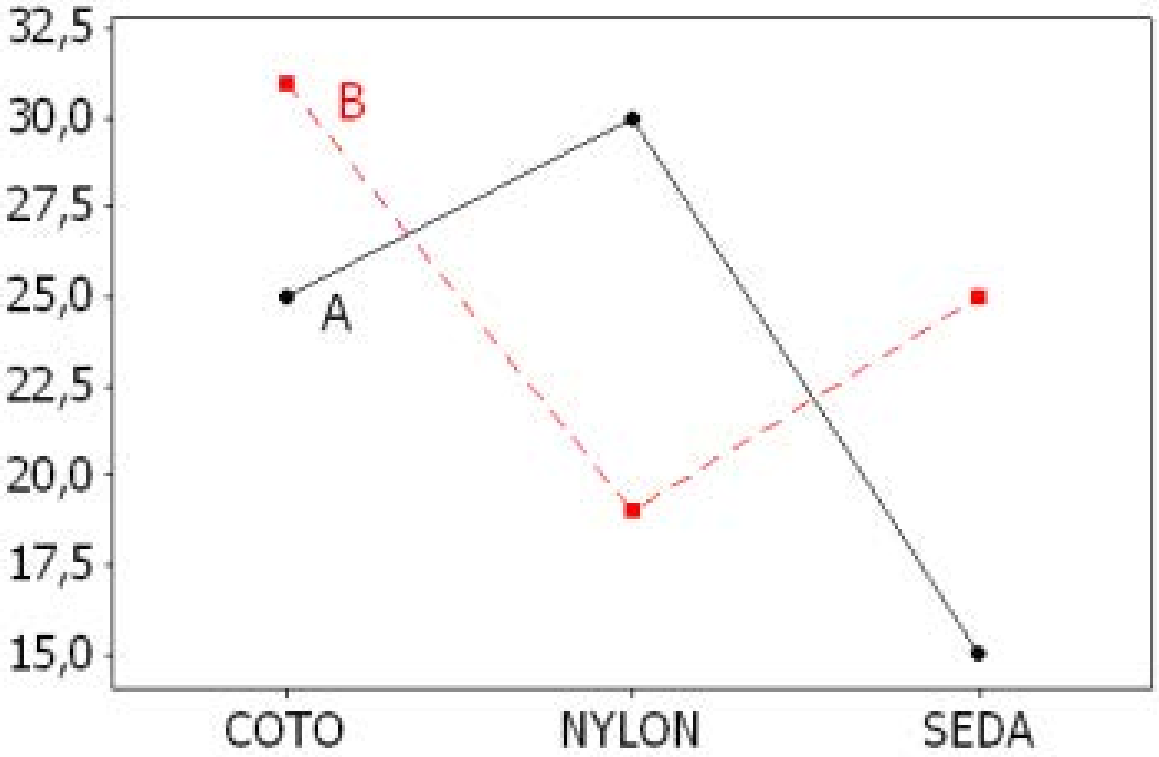
Gràficament, la interacció s'observa amb la manca de paral · lelisme dels plots dels polígons corresponents a les mitjanes de cada cel · la

ANOVA de dos factors

Model additiu:



Model amb interacció:



ANOVA de dos factors

En l'ANOVA de dos factors amb interacció, també fem el test:

$$H_0^3 : \gamma_{ij} = 0, \forall (i,j) \;\; vs \;\; H_1^3 : \exists (i,j) \qquad \gamma_{ij} \neq 0,$$

Si la interacció és significativa, significa que els factors no actuen de forma independent

Observació: El model amb interacció requereix de més d'una rèplica per condició experimental ($n \geq 2$). En cas contrari, no tindrem graus de llibertat pel term de l'error

ANOVA de dos factors

La taula de l'ANOVA de dos factors amb interacció és:

	SS	d. f.	MSE	F
Factor A	SS_A	$a-1$	$\frac{SS_A}{a-1}$	$F_0^1 = \frac{SS_A/(a-1)}{SS_E/(ab(n-1)))}$
Factor B	SS_B	$b-1$	$\frac{SS_B}{b-1}$	$F_0^2 = \frac{SS_B/(b-1)}{SS_E/(ab(n-1)))}$
AB	SS_{AB}	$(a-1)(b-1)$	$\frac{SS_{AB}}{(a-1)(b-1)}$	$F_0^3 = \frac{SS_{AB}/(a-1)(b-1)}{SS_E/(ab(n-1))}$
Error	SS_E	$ab(n-1)$	$\frac{SS_E}{ab(n-1)}$	
Total	SS_T	$N-1$	$\frac{SS_T}{N-1}$	

Les regles de decisió són similars a les anteriors.

ANOVA de dos factors

Si la interacció és estadísticament diferent de zero, podem fer servir els mateixos mètodes per comparar els nivells d'un factor, si especifiquem quin nivell de l'altre factor hem fixat

Això implica que, per exemple, les comparacions dos a dos de les mitjanes pel factor A es poden aplicar per a cada nivell del factor B, i reciprocament.

ANOVA de dos factors

Exemple: Duració de bateríes, Y, com a funció del material (3 nivells), X_1 i temperatura (3 nivells) X_2

Mat. | Temp.	15	70	125
1	$130, 74, 155, 180$	$34, 40, 80, 75$	$20, 70, 82, 58$
2	$150, 188, 159, 126$	$136, 122, 106, 115$	$25, 70, 58, 45$
3	$138, 110, 168, 160$	$174, 120, 150, 139$	$96, 104, 82, 60$

Volem saber:

- Existeixen diferències en la duració segons el material?
- Existeixen diferències en la duració segons la temperatura?
- És l'efecte de la temperatura independent del material (o viceversa)?

ANOVA de dos factors

	SS	d. f.	MSE	F	p-val.
Material	10683.7	2	5341.9	$F_0^1 = 7.91$	0.002
Temperature	39118.7	2	19559.4	$F_0^2 = 28.97$	0.000
Interaction	9613.8	4	2403.4	$F_0^3 = 3.56$	0.019
Error	18230.7	27	675.2		
Total	77647.0	35	$\frac{SS_T}{N-1}$		

Existeixen diferències ente els nivells dels dos factors i també existeix una interacció significativa amb $\alpha = 0.05$

Podem aplicar comparacions múltiples (Tukey) per un factor si hem fixat el nivell de l'altre factor

ANOVA: Matrius de disseny

ANOVA additiu de dos factors creuats amb 2 i tres nivells

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \qquad i = 1, 2 \quad j = 1, 2, 3$$

Matriu completa

A	B		A1	A2	B1	B2	B3
A1	B1	1	1	0	1	0	0
A1	B2	1	1	0	0	1	0
A1	B3	1	1	0	0	0	1
A2	B1	1	0	1	1	0	0
A2	B2	1	0	1	0	1	0
A2	B3	1	0	1	0	0	1

Contrast de tipus *treatment* (baseline) amb la primera categoria com a referència:

$$\alpha_1 = 0, \quad \beta_1 = 0$$

A	B		A2	B2	B3
A1	B1	1	0	0	0
A1	B2	1	0	1	0
A1	B3	1	0	0	1
A2	B1	1	1	0	0
A2	B2	1	1	1	0
A2	B3	1	1	0	1

ANOVA: Matrius de disseny

ANOVA de dos factors amb 2 i tres nivells amb interacció

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \qquad i = 1, 2 \quad j = 1, 2, 3$$

Matriu completa

A	B		A1	A2	B1	B2	B3	A1B1	A1B2	A1B3	A2B1	A2B2	A2B3
A1	B1	1	1	0	1	0	0	1	0	0	0	0	0
A1	B2	1	1	0	0	1	0	0	1	0	0	0	0
A1	B3	1	1	0	0	0	1	0	0	1	0	0	0
A2	B1	1	0	1	1	0	0	0	0	0	1	0	0
A2	B2	1	0	1	0	1	0	0	0	0	0	1	0
A2	B3	1	0	1	0	0	1	0	0	0	0	0	1

Contrast de tipus *treatment* (baseline) amb la primera categoria com a referència:

$$\alpha_1 = 0, \quad \beta_1 = 0, \quad \gamma_{1j} = 0 \quad \forall j, \quad \gamma_{i1} = 0 \quad \forall i$$

A	B		A2	B2	B3	A2B2	A2B3
A	B		A2	B2	B3	A2B2	A2B3
A1	B1	1	0	0	0	0	0
A1	B2	1	0	1	0	0	0
A1	B3	1	0	0	1	0	0
A2	B1	1	1	0	0	0	0
A2	B2	1	1	1	0	1	0
A2	B3	1	1	0	1	0	1

ANOVA: Tipus de Suma de quadrats

Suma de quadrats de **Tipus I**: Considera l'efecte dels factors a mesura que es van introduint. El seu valor depén de l'ordre en que s'introdueixen els factors. També coneguda com a suma de quadrats **seqüencial**. Dona una descomposició de la suma de quadrats total

Variable	Suma de Quadrats
A	$SS_{Nul} - SS_A$
B	$SS_A - SS_{A,B}$
AB	$SS_{A,B} - SS_{A,B,AB}$

ANOVA: Tipus de Suma de quadrats

Suma de quadrats de **Tipus II**: Calculats pels efectes principals en totes les seves possibles permutacions, assumint que no hi ha interacció

Variable	Suma de Quadrats
A	$SS_B - SS_{A,B}$
B	$SS_A - SS_{A,B}$
AB	$SS_{A,B} - SS_{A,B,AB}$

ANOVA: Tipus de Suma de quadrats

Suma de quadrats de **Tipus III**: Calculats pels efectes principals en totes les seves possibles permutacions però incloent les possibles interaccions. També coneguda com a suma de quadrats **marginal**

Variable	Suma de Quadrats
A	$SS_{B,AB} - SS_{A,B,AB}$
B	$SS_{A,AB} - SS_{A,B,AB}$
AB	$SS_{A,B} - SS_{A,B,AB}$