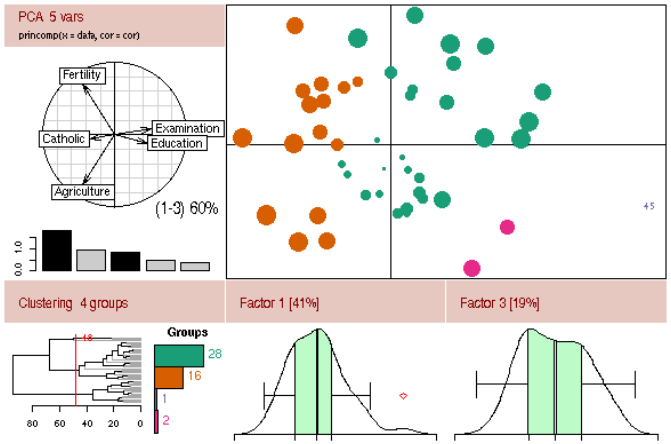
La idea de l'anàlisi de la variància
Model d'un factor

$$y_{ij} = \mu + \alpha_i + e_{ij}$$



Exemple 1: situació experimental

- Es desitja comparar l'eficàcia de tres fàrmacs = *tractaments*
- la variable resposta és una mesura d'eficàcia (major valor major eficàcia)
- una mostra de 24 pacients s'aleatoritza totalment respecte al tractament.

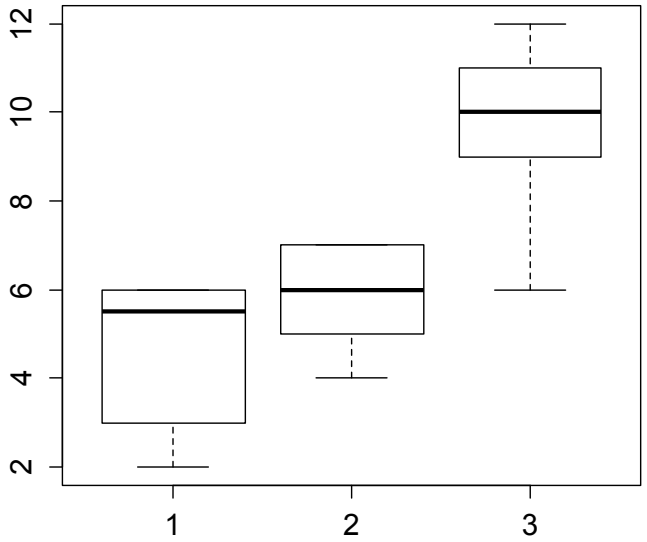
Resultats experiment

	Trat 1	Trat 2	Trat 3
	4	7	9
	2	6	12
	6	5	6
	6	7	11
	5	6	10
	6	4	11
	2	7	9
	6	5	10

- en terminologia de l'anàlisis de la variància s'anomena:
  - **nivell** a cada grup
  - **factor** al conjunt de tots els grups

Descriptiva bàsica de les dades

```
> by(dades$resp, dades$tract, summary)
dades$tract: 1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.500   5.500   4.625   6.000   6.000
------------------------------------------------------------
dades$tract: 2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.000   5.000   6.000   5.875   7.000   7.000
------------------------------------------------------------
dades$tract: 3
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00    9.00   10.00    9.75   11.00   12.00
> boxplot(resp~tract, dades)
```



Comentaris a l'exemple 1

- Es vol comparar l'eficàcia de 3 tractaments.
- Acceptant normalitat, i considerant 3 contrastos t-Student **per parelles** consecutius:
  - fàrmac 1 contra fàrmac 2 (α=5%)
  - fàrmac 1 contra fàrmac 3 (α=5%)
  - fàrmac 2 contra fàrmac 3 (α=5%)
- aparentment es resol la qüestió plantejada. Però …
- aquesta forma directa de plantejar el problema comporta un error metodològic: el nivell de significació de las 3 proves juntes és superior a α

Nivell de significació global

- si plantegem els dos testos següents
  - fàrmac 1 contra fàrmac 2 (α=5%)
  - fàrmac 1 contra fàrmac 3 (α=5%)

  les proves són independents essent l'error de tipus I global:

  α_G = nivell de significació global = $1-(1-0.05)^2 = 0.0975$

- si plantegem les 3 alhora, en no ser independents, només podem afirmar que α_G està entre 0.0975 i 0.1426.

- en general, quant més gran sigui el número de grups, més gran serà α_G acostant-se a 1!. Cal doncs una **tècnica alternativa**.

El model completament aleatoritzat

- Generalització a k mostres del test t de Student para 2 mostres normals independents.
- **Objectiu del test**: comprovar si existeixen diferencies significatives entre k grups experimentals.
- Els individus han estat assignats a l'atzar a un dels possibles tractaments
- Aquesta aleatorització dels individus als tractaments dona un nom alternatiu al disseny: ***completament aleatoritzat***

Codificació amb R

- Es requereixen dues variables
  - 1a variable: codi, per cada rèplica, del tractament aplicat
  - 2a variable: variable resposta per cada rèplica

	Trat 1	Trat 2	Trat 3
	4	7	9
	2	6	12
	6	5	6
	6	7	11
	5	6	10
	6	4	11
	2	7	9
	6	5	10

	tract	resp
1	1	4
2	1	2
3	1	6
4	1	6
5	1	5
6	1	6
7	1	2
8	1	6
9	2	7
10	2	6
11	2	5
12	2	7
13	2	6
14	2	4
15	2	7
16	2	5
17	3	9

Resultats de l'ANOVA 1F

```
> result <- aov(resp~tract, dades)
> summary(result)
            Df Sum Sq Mean Sq F value   Pr(>F)
tract        1 105.06  105.06   36.44 4.48e-06 ***
Residuals   22  63.44    2.88
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Prenent un α del 5%:
  - L'estadístic de test es **F = 22.113**
  - el p-valor subministrat per R és **0.0000068**
  - com p-valor < α , **podem rebutjar H_0**

Model ANOVA 1 Factor

- El model lineal assumit per a les dades és

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad i = 1, \cdots, a; \quad j = 1, \cdots, n_i$$

μ = mitjana general; τ_i = efecte tractament *i*; ε_{ij} = error aleatori

- Cada grup experimental s'associa a una N(μ_i ,σ).

- El nombre de tractaments *a* = *T* (total tractaments possibles). Si *a* < *T*, el model és d'efectes *aleatoris*

- L'*efecte del tractament i* és la diferència entre $\tau_i = \mu_i - \mu$

Terminologia elemental

- **Rèplica**: observacions de la variable resposta fetes sota les mateixes condicions experimentals.
- **Disseny balancejat**: situació experimental on es presenten el mateix nombre de rèpliques en cada grup/tractament .
- **Aleatorització**: un **requisit bàsic**. Cada unitat observada (cada rèplica) s'ha d'assignar aleatòriament a un tractament.
- El disseny Anova 1F es designa també com disseny **completament aleatoritzat**.

Contrast de hipòtesis

- L'Anova de 1 factor contrasta la hipòtesis que **no hi ha efecte dels tractaments**. En forma paramètrica, H_0 es:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \longrightarrow H_0: \ \tau_1 = \tau_2 = \cdots = \tau_a = 0$$

- Si rebutgem H_0, el test **només indica** que existeix una **diferència global** en els grups, però no concretament entre quins d'ells.

- Una forma equivalent d'expressar la hipòtesi nul·la és mitjançant les mitjanes poblacionals de cada grup μ_i

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_a$$

Paràmetres del model i estimacions

- Paràmetres que intervenen

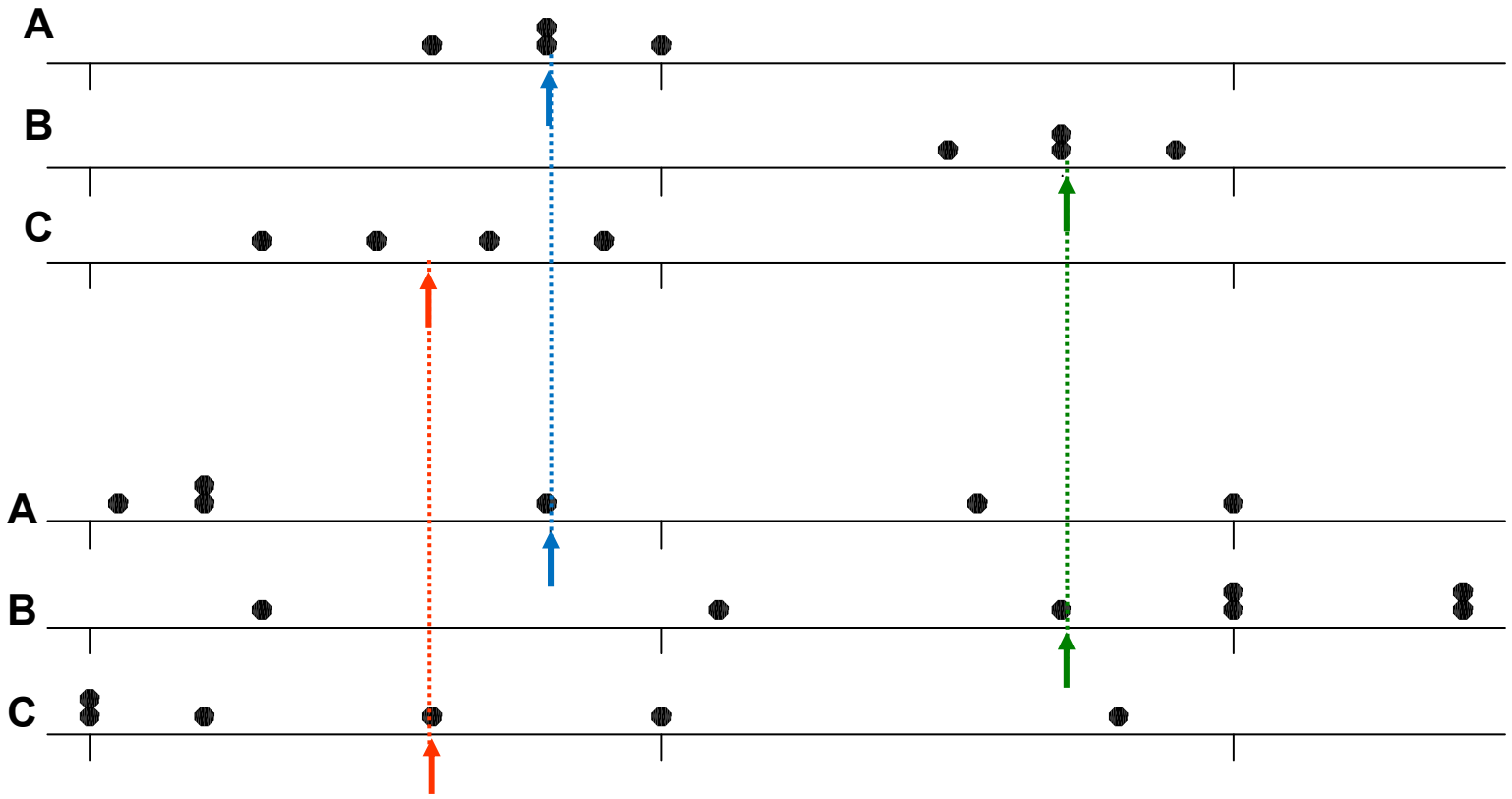
Tractament	Mitjana	Var.	Efecte tract.
1	μ_1	σ_1^2	$\tau_1=\mu_1-\mu$
2	μ_2	σ_2^2	$\tau_2=\mu_2-\mu$
....			
a=T	μ_a	σ_a^2	$\tau_a=\mu_a-\mu$
Mitjana	μ		0

- i les seves estimacions:

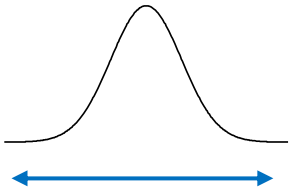
Tractament	Mitjana	Var.	Efecte tract.
1	$\bar{y}_{1\bullet}$	s_1^2	$\bar{y}_{1\bullet}-\bar{y}_{\bullet\bullet}$
2	$\bar{y}_{2\bullet}$	s_2^2	$\bar{y}_{2\bullet}-\bar{y}_{\bullet\bullet}$
....			
a=T	$\bar{y}_{a\bullet}$	s_a^2	$\bar{y}_{a\bullet}-\bar{y}_{\bullet\bullet}$
Mitjana	$\bar{y}_{\bullet\bullet}$		0

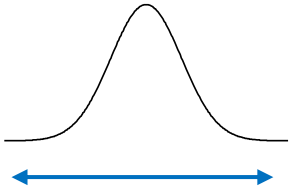
Per què el nom d'anàlisi de la variància?

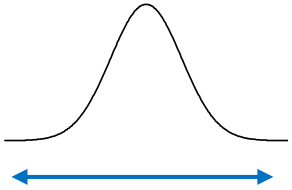
Hi ha diferències significatives entre μ_A, μ_B i μ_C?

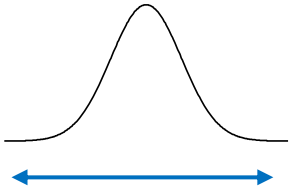


Idea intuitiva de l'anàlisi de la variància





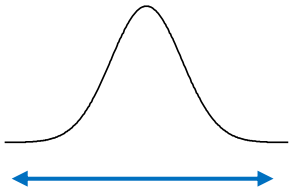


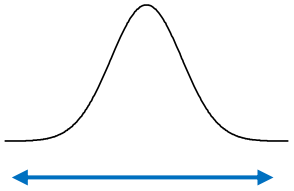


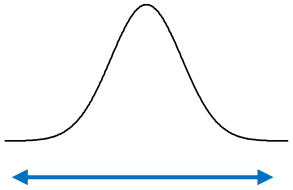
Com que

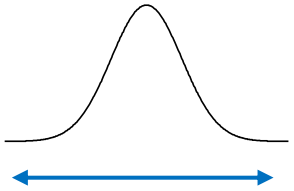
i

són iguals diem que *no hi ha diferència* entre les mitjanes tractaments









Com que

i

no són iguals diem que *hi ha diferència* entre les mitjanes dels tractaments

Idea intuitiva de l'anàlisi de la variància

Veure si les mitjanes poblacionals són iguals:

$$\mu_1 = \mu_2 = \ldots = \mu_K$$

és equivalent a comprovar si...

les diferències (variabilitat) entre les mitjanes de les mostres són superiors al que s'hauria d'esperar a partir de la variabilitat dins de cada mostra.

La tècnica d'anàlisi per comparar les mitjanes poblacionals és, per tant, una anàlisi de les variabilitats o anàlisi de la variança (ANOVA).

El mètode ANOVA permet dividir la variabilitat observada en components independents, que poden atribuir-se a diferents causes. Volem determinar si la variabilitat d'una variable és atribuible al tractament.

Passos que cal seguir per comparar més de 2 mitjanes

1. Plantejar les hipòtesis

   Ara, la hipòtesi nul·la serà que totes les mitjanes són iguals, i l'alternativa que alguna és diferent.

2. Recollir les dades

   Cal aleatoritzar o bloquejar, segons el cas en què ens trobem.

3. Anàlisi exploratori de les dades.

   Per veure quin aspecte tenen les dades, detectar possibles valors anòmals, etc.

4. Verificació dels supòsits en que es basa la metodologia.

5. Construcció de la taula ANOVA.

6. Decisió.

   Rebutgem o no la hipòtesi nul·la segons el p-valor obtingut.

Comparació de k tractaments

1. Plantejar les hipòtesis

Hipòtesis nul·la $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$

Hipòtesis alternativa $H_1 :$ **alguna és diferent**

2. Recollir les dades

Tractament 1	$y_{11}, y_{12}, \ldots, y_{1n_1}$	$\rightarrow \quad \overline{y_{1\bullet}}$
Tractament 2	$y_{21}, y_{22}, \ldots, y_{2n_2}$	$\rightarrow \quad \overline{y_{2\bullet}}$
⋮		
Tractament t	$y_{t1}, y_{t2}, \ldots, y_{tn_t}$	$\rightarrow \quad \overline{y_{t\bullet}}$
⋮		
Tractament k	$y_{k1}, y_{k2}, \ldots, y_{kn_k}$	$\rightarrow \quad \overline{y_{k\bullet}}$

Tenim k tractaments.
El tractament t té n_t observacions

Mitjana global $\overline{y}_{\bullet\bullet}$

$N = n_1 + n_2 + \ldots + n_k$

Comparació de k tractaments

3. Fer anàlisi exploratòria de dades:
   - Diagrama de punts.
   - Boxplot...

4. Verificar els supòsits en que es basa la metodologia
   - Les poblacions de les que venen les dades són normals.
   - Les poblacions són independents.
   - Les mostres són m.a.s.
   - Les variances poblacionals són iguals.

Comparació de k tractaments

5. Construir la taula ANOVA:

Font de Variacio	Suma Quadrats	g,l,	Quadrats Mitjans	F	p-valor
Entre tractaments	SS_T	k – 1	S_T^2	S_T^2 / S_R^2	
Error	SS_R	N – k	S_R^2		
Total	SS_{Tot}	N – 1			

Veurem com “omplir” aquesta taula amb números...

6. Decidir en base al p-valor obtingut.

Petit → rebutgem H_0, alguna mitjana és diferent

Gran → no rebutgem H_0, no podem dir que les mitjanes siguin diferents

Construcció de la taula ANOVA

Si els supòsits del model són vàlids,

Tractament 1	$y_{11}, y_{12}, \cdots, y_{1n_1}$	$\sim N(\mu_1;\sigma)$
Tractament 2	$y_{21}, y_{22}, \cdots, y_{2n_2}$	$\sim N(\mu_2;\sigma)$
	⋮	
Tractament t	$y_{t1}, y_{t2}, \cdots, y_{tn_t}$	$\sim N(\mu_t;\sigma)$
	⋮	
Tractament K	$y_{k1}, y_{k2}, \cdots, y_{kn_k}$	$\sim N(\mu_K;\sigma)$

S_R^2 i SS_R

Estimar σ^2 mitjançant una mesura ponderada de les s^2,

És a dir estimar la variació dintre dels tractaments

$$s_R^2 = \frac{\sum_{t=1}^{k}(n_t - 1)s_t^2}{\sum_{t=1}^{k}(n_t - 1)} = \frac{\sum_{t=1}^{k}(n_t - 1)\frac{\sum_{i=1}^{n_t}(y_{ti} - \bar{y}_{t\bullet})^2}{n_t - 1}}{N - k} = \frac{\sum_{t=1}^{k}\sum_{i=1}^{n_t}(y_{ti} - \bar{y}_{t\bullet})^2}{N - k} = \frac{SS_R}{N - k}$$

És com la s-combinada en comparació de 2 tractaments

S_R^2 se l'anomena variabilitat residual, perquè no és atribuible a cap causa en concret (és la variabilitat deguda a causes comuns),

SS_R se l'anomena Suma de Quadrats Residuals

S_T^2 i SS_T

Estimar σ^2 basant-nos en les variacions entre tractaments,

Si H_0 certa $y_{ti} \sim N(\mu;\sigma)$ (la μ és la mateixa per tots els tractaments)

Per tant, si les mostres tenen el mateix tamany:

$$\bar{y}_{t\cdot} \sim N(\mu; \sigma/\sqrt{n}) \qquad \frac{\sum_{t=1}^{k}(\bar{y}_{t\cdot} - \bar{y}_{\cdot\cdot})^2}{k-1} \text{ és un estimador de } \sigma^2/n$$

I llavors:

$$\frac{n\sum_{t=1}^{k}(\bar{y}_{t\cdot} - \bar{y}_{\cdot\cdot})^2}{k-1} \text{ és un estimador de } \sigma^2$$

En general, si el tamany de les mostres és diferent:

$$S_T^2 = \frac{\sum_{t=1}^{k} n_t(\bar{y}_{t\cdot} - \bar{y}_{\cdot\cdot})^2}{k-1} = \frac{SS_T}{k-1}$$

S_T^2 és la variabilitat entre tractaments, i

SQ_T és la Suma de Quadrats dels Tractaments

Dues formes d'estimar σ^2 quan H_0 és certa

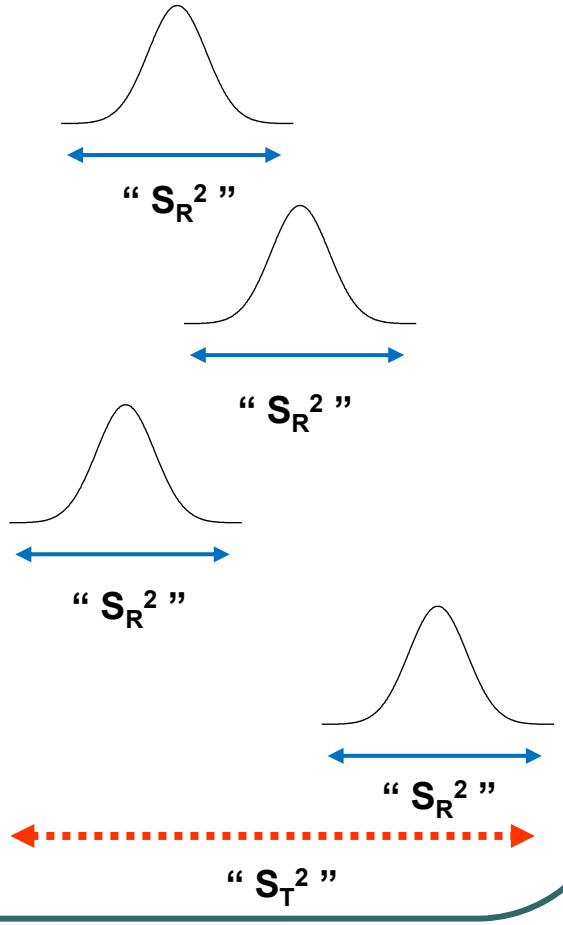
Si H_0 és falsa llavors:

$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$

H_1: **alguna és diferent**

S_T^2 serà més gran que S_R^2 ja que en aquest cas no estarà només afectada per la variabilitat dintre dels tractaments sino també per la variabilitat entre tractaments

S_R^2	sempre és bon estimador de σ^2
S_T^2	només és bon estimador de σ^2 si H_0 és certa



Càlcul del p-valor a la taula ANOVA

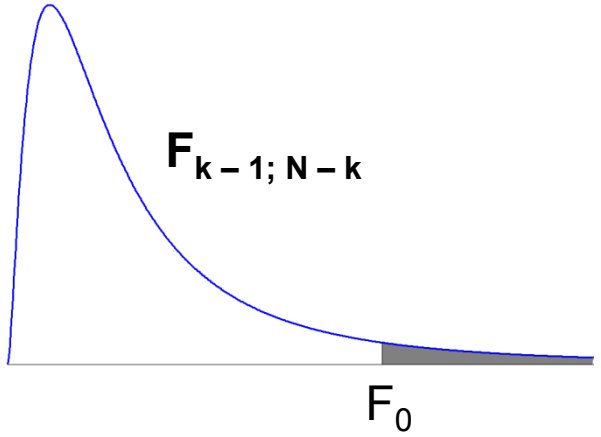
Per tant, cal comparar 2 variances.

$$H_0 : \sigma_T^2 = \sigma_R^2$$
$$H_1 : \sigma_T^2 > \sigma_R^2$$

Atenció, aquí el p-valor és només una àrea de cua!

L'estadístic de prova és:

$$F_0 = \frac{S_T^2}{S_R^2}$$



La distribució de referència és una $F_{k-1;\,N-k}$

Així doncs, $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ és equivalent a $H_0 : \sigma_T^2 = \sigma_R^2$

H_1: **alguna és diferent** $H_1 : \sigma_T^2 > \sigma_R^2$

Taula ANOVA

Font de Variació	Suma Quadrats	g.l.	Quadrats Mitjans	F	p-valor
Entre tractaments	$SS_T = \sum_{t=1}^{k}\sum_{i=1}^{n_t}(\bar{y}_{t\cdot} - \bar{y}_{\cdot\cdot})^2$	k – 1	$S_T^2 = \frac{\sum_{t=1}^{k}\sum_{i=1}^{n_t}(\bar{y}_{t\cdot} - \bar{y}_{\cdot\cdot})^2}{k-1}$	S_T^2 / S_R^2	
Error	$SS_R = \sum_{t=1}^{k}\sum_{i=1}^{n_t}(y_{ti} - \bar{y}_{t\cdot})^2$	N – k	$S_R^2 = \frac{\sum_{t=1}^{k}\sum_{i=1}^{n_t}(y_{ti} - \bar{y}_{t\cdot})^2}{N-k}$		
Total	$SS_{Tot} = \sum_{t=1}^{k}\sum_{i=1}^{n_t}(y_{ti} - \bar{y}_{\cdot\cdot})^2$	N – 1			

Error = variabilitat dins de cada tractament, la "tremolor" que sempre tenen les dades...

El p-valor surt d'enfrontar l'estadístic de prova (S_T^2 / S_R^2) a la distribució de referència, que en aquest cas és una $F_{k-1;\, N-k}$

Exemple 2

Suposem que es vol comparar la *productivitat mitjana per hora* en el muntatge d'un cert mecanisme, segons sigui el procediment de muntatge emprat:

A, B o C

Com es resoldrà el problema?

Plantejar hipòtesis i recollir dades

1. Plantejar les hipòtesis

Hipòtesis nul·la $H_0: \mu_A = \mu_B = \mu_C$

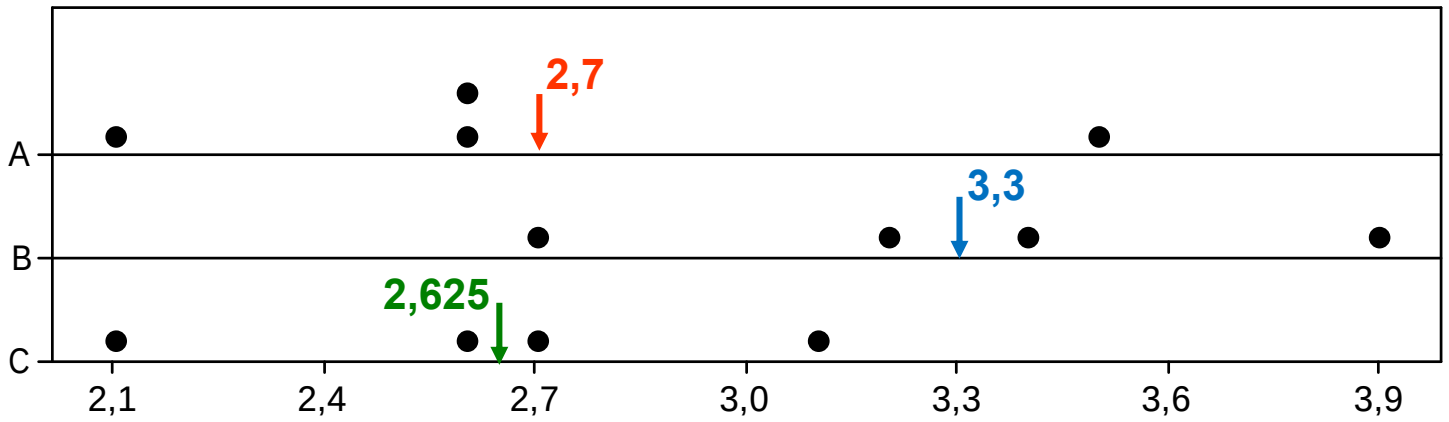
Hipòtesis alternativa $H_1:$ **alguna és diferent**

2. Recollir les dades

A	B	C
$2,6^{(6)}$	$3,2^{(12)}$	$2,6^{(4)}$
$2,1^{(10)}$	$2,7^{(1)}$	$2,1^{(8)}$
$3,5^{(2)}$	$3,9^{(3)}$	$3,1^{(11)}$
$2,6^{(9)}$	$3,4^{(7)}$	$2,7^{(5)}$
$\bar{Y}_A = 2,7$	$\bar{Y}_B = 3,3$	$\bar{Y}_C = 2,625$
$S_A = 0,58$	$S_B = 0,5$	$S_C = 0,41$

Anàlisi exploratòria i supòsits

3. Anàlisi exploratoria de dades



4. Verificació dels supòsits

- Poblacions normals
- Poblacions independents
- Aleatorietat
- Amb la mateixa variança

Queda pendent comprovar-ho!

Construcció de la taula ANOVA

5. Construcció de la taula ANOVA

$$S_R^2 = \frac{SS_R}{N-k} = \frac{(n_A-1)S_A^2 + (n_B-1)S_B^2 + (n_C-1)S_C^2}{N-k} =$$

$$= \frac{3(0{,}34+0{,}247+0{,}169)}{9} = \frac{2{,}268}{9} = 0{,}252$$

$$S_T^2 = \frac{SS_T}{K-1} = \frac{\sum_{t=1}^{k} n_t(\bar{y}_{t\bullet} - \bar{y}_{\bullet\bullet})^2}{k-1} = \frac{4\cdot(2{,}7-2{,}875)^2 + 4\cdot(3{,}3-2{,}875)^2 + 4\cdot(2{,}625-2{,}875)^2}{3-1} =$$

$$= \frac{1{,}095}{2} = 0{,}5475$$

$$SS_{Tot} = SS_R + SS_T = \sum_{t=1}^{k}\sum_{i=1}^{n_t}(y_{ti} - \bar{y}_{\bullet\bullet})^2 = 3{,}362$$

Font de Variacio	Suma Quadrats	g.l.	Quadrats mitjans	F	p-valor
Entre tractaments	SS_T=1,095	2	S_T^2=0,5475	S_T^2/S_R^2=2,17	0,17
Error	SS_R=2,268	9	S_R^2=0,252		
Total	SS_{Tot}=3,362	11			

La Taula ANOVA

Font de variació	Suma de quadrats	g.l.	Quadrats mitjans	F
Entre grups (tractament)	SS_T	$k-1$	$MS_T = \frac{SS_T}{k-1}$	$\frac{MS_T}{MS_R}$
Dins grups (Error)	SS_R	$N-k$	$MS_R = \frac{SS_R}{N-k}$	
Total	SS_{Tot}	$N-1$		

- SS_{Tot} recull la variabilitat **total** de les dades
- SS_T recull la variabilitat **entre** tractaments
- SS_R recull la variabilitat **dins** dels tractaments

- H_0 es rebutja si la variabilitat **entre** (SS_T) supera significativament la variabilitat **dins** dels grups (SS_R).
- En les taules ANOVA la variabilitat

$$MS = Quadrats\ mitjans = \frac{variació}{graus\ de\ llibertat} = \frac{SS}{g.l.}$$