ANOVA

Problema bàsic

- Tenim $k > 2$ poblacions. Usualment són subpoblacions d'una única població, definides pels nivells de factors.
- Volem decidir si el valor mitjà d'un cert paràmetre és el mateix a totes aquestes poblacions o no
- Siguin $\mu_1, \ldots, \mu_k$ les mitjanes d'aquest paràmetre en aquestes poblacions. Volem fer el contrast:

$$\left\{ \begin{array}{l} H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1 : \exists i, j \mid \mu_i \neq \mu_j \end{array} \right.$$

- Prendrem una mostra aleatòria de cada població, i a partir d'aquestes mostres ho decidirem

ANOVA

La tècnica que emprarem és l'Anàlisi de la Variància (ANOVA, de l'anglès *ANalysis Of VAriance*)

Aquesta tècnica es pot aplicar sota diferents dissenys d'experiments:

- Segons quants factors empram per separar la població en subpoblacions
- Segons com triam els nivells dels factors
- Segons com escollim les mostres

Veurem els dissenys més bàsics. En un problema concret, s'ha de decidir primer el tipus d'experiment que s'ha de realitzar.

ANOVA

Per comparar les mitjanes de dues poblacions, calculàvem les mitjanes de dues mostres i les comparàvem

Per comparar les mitjanes de $k \geqslant 3$ poblacions, podríem fer-ho per parelles, però són moltes: $\binom{k}{2}$

I les hem de comparar totes, perquè pot passar que

$$\mu_1 \approx \mu_2, \mu_2 \approx \mu_3, \mu_3 \approx \mu_4 \text{ però } \mu_1 \not\approx \mu_4$$

Volem un test que ens digui en un pas si totes són iguals, o si n'hi ha de diferents (en aquest darrer cas, ja cercarem les diferents si volem)

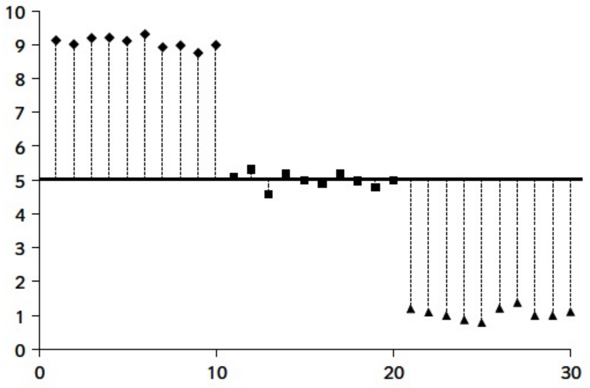
ANOVA

Per comparar les mitjanes de 3 o més poblacions, ens concentram en la variabilitat de les dades per grups:

- Variabilitat de les dades (respecte de la mitjana global)
- Variabilitat dins cada població (respecte de la mitjana dins la població)
- Variabilitat de les mitjanes per poblacions (respecte de la mitjana global)

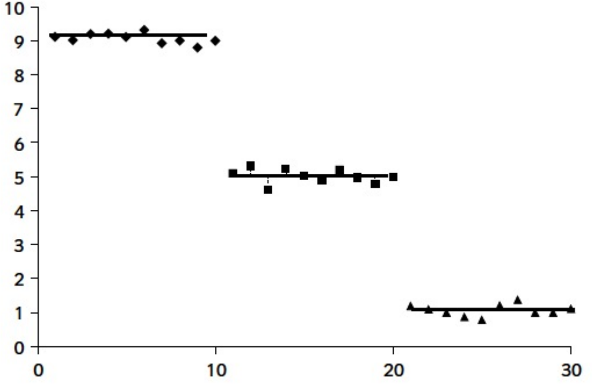
Si la variabilitat total de les dades és explicada per la variabilitat de les mitjanes de les poblacions i la "poca variabilitat" dins cada població, és indici que les mitjanes són diferents

ANOVA



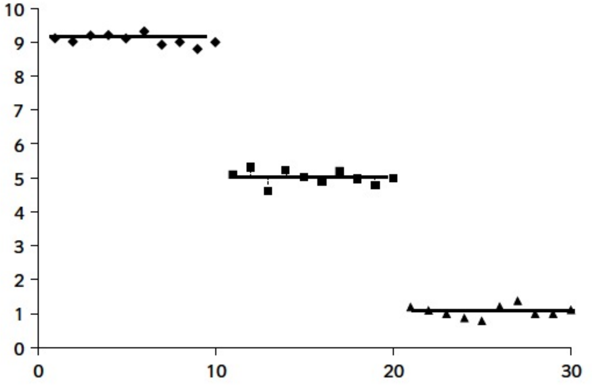
Molta variabilitat. . .

ANOVA



... concentrada entorn de les mitjanes per nivells

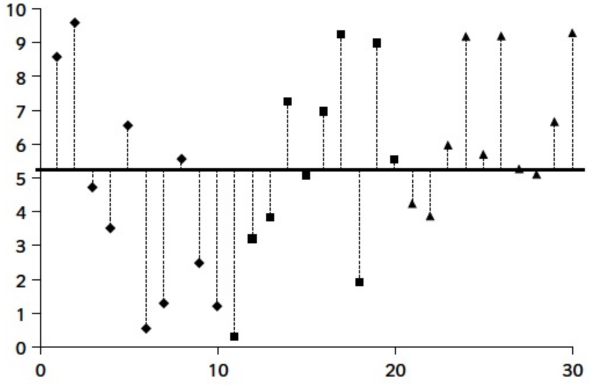
ANOVA



... concentrada entorn de les mitjanes per nivells

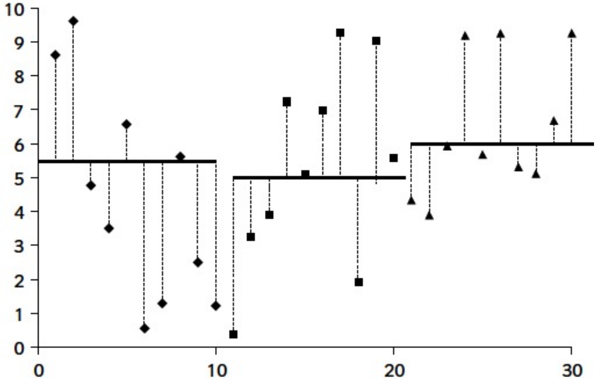
Les mitjanes són (òbviament) diferents

ANOVA



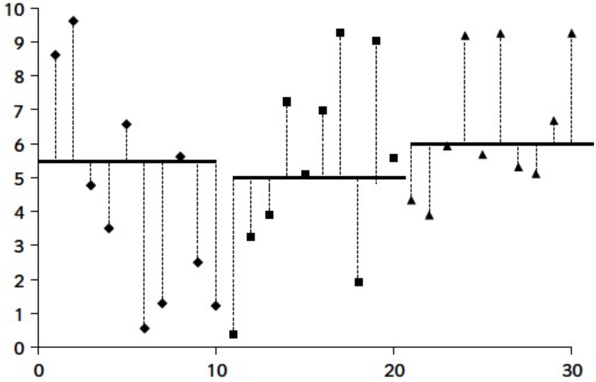
Molta variabilitat...

ANOVA



... que no és explicada per les diferències de les mitjanes

ANOVA



... que no és explicada per les diferències de les mitjanes

Les mitjanes no semblen diferents

ANOVA

Com quantificar-ho?

Classificació simple, efectes fixats, disseny completament aleatori

En els experiments amb classificació simple, efectes fixats i disseny completament aleatori:

- Empram un sol factor per classificar la població en subpoblacions (classificació simple)
- L'investigador decideix quins nivells (o tractaments) del factor emprarà (efectes fixats)
- Es pren una m.a.s. de cada subpoblació, de manera independent unes de les altres (completament aleatori)

Exemple 1

Es realitzà un estudi per investigar l'efecte del CO_2 sobre la taxa de creixement de *Pseudomonas fragi* (un corruptor d'aliments). Es creu que el creixement es veu afectat per la quantitat de CO_2 en l'aire.

Per contrastar-ho, en un experiment s'administrà CO_2 a 5 pressions atmosfèriques diferents a 10 cultius diferents per cada nivell, i s'anotà el canvi (en %) de la massa cel·lular al cap d'una hora

Exemple 1: Dades obtingudes

Pressió de CO_2 (en atmosferes)				
0.0	0.083	0.29	0.50	0.86
62.6	50.9	45.5	29.5	24.9
59.6	44.3	41.1	22.8	17.2
64.5	47.5	29.8	19.2	7.8
59.3	49.5	38.3	20.6	10.5
58.6	48.5	40.2	29.2	17.8
64.6	50.4	38.5	24.1	22.1
50.9	35.2	30.2	22.6	22.6
56.2	49.9	27.0	32.7	16.8
52.3	42.6	40.0	24.4	15.9
62.8	41.6	33.9	29.6	8.8

Exemple 2

Disposam de quatre tractaments genètics diferents per corregir un cert gen defectuós responsable d'una malaltia. Els investigadors creuen que els quatre tractaments tenen una eficàcia similar.

Per contrastar-ho, prenen 20 malalts diferents a l'atzar, els reparteixen aleatòriament en 4 grups de 5 malalts, i assignen de forma aleatòria un dels quatre tractaments a cada grup

Després d'aplicar el tractament, mesuren l'expressió del gen defectuós sota estudi

Exemple 2: Dades obtingudes

Tractament			
A	B	C	D
96	93	70	78
99	90	90	87
100	75	84	57
104	80	76	66
84	90	78	76

Tipus d'experiments

Són experiments amb classificació simple, efectes fixats i disseny completament aleatori

- Empram un sol factor (pressió, tractament)
- Decidim quins nivells empram (els que hem decidit emprar)
- Hem pres una m.a.s. de cada nivell del factor, i de manera independent

Exemple 1

Emmagatzemaré les dades en un *dataframe* amb dues variables:

- `Inc`: increment massa cel·lular (en %)
- `Pre`: Nivell de pressió, com a factor

Pressió de CO_2 (en atmosferes)				
0.0	0.083	0.29	0.50	0.86
62.6	50.9	45.5	29.5	24.9
59.6	44.3	41.1	22.8	17.2
64.5	47.5	29.8	19.2	7.8
59.3	49.5	38.3	20.6	10.5
58.6	48.5	40.2	29.2	17.8
64.6	50.4	38.5	24.1	22.1
50.9	35.2	30.2	22.6	22.6
56.2	49.9	27.0	32.7	16.8
52.3	42.6	40.0	24.4	15.9
62.8	41.6	33.9	29.6	8.8

Exemple 1

Emmagatzemaré les dades en un *dataframe* amb dues variables:

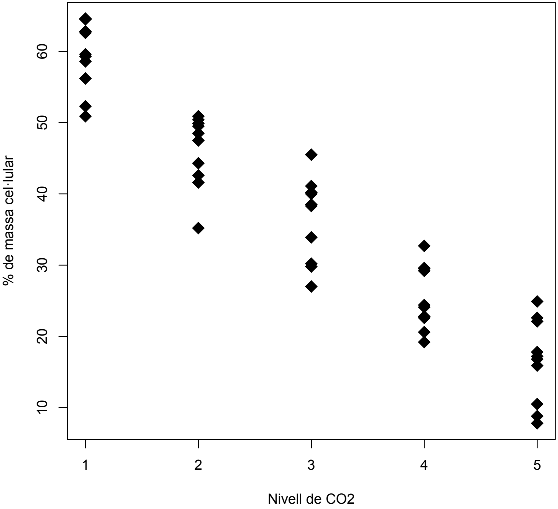
- `Inc`: increment massa cel·lular (en %)
- `Pre`: Nivell de pressió, com a factor

```
> Inc=c(62.6,50.9,45.5,
 29.5,24.9,59.6,44.3, 41.1,22.8,
 17.2,64.5,47.5,29.8,19.2,
 7.8,59.3, 49.5,38.3,20.6,10.5,58.6,48.5,40.2,
 29.2,17.8, 64.6,50.4,38.5,24.1,22.1,50.9,35.2,
 30.2,22.6, 22.6,56.2,49.9,27.0,32.7,16.8,52.3,
 42.6,40.0, 24.4,15.9,62.8,41.6,33.9,29.6,8.8)
> Pre=rep(c("0.0","0.083","0.29","0.50","0.86"),
 times=10)
```

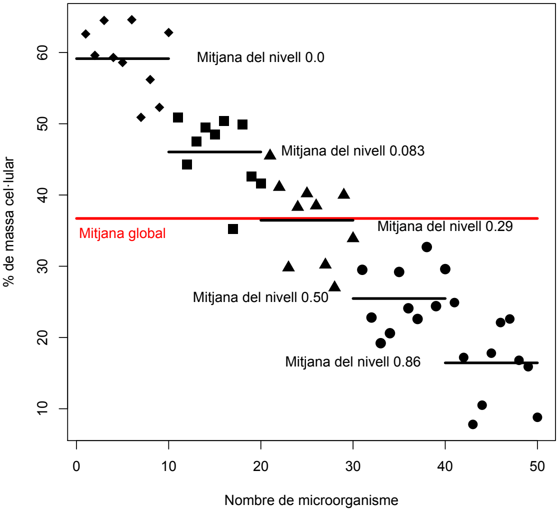
Exemple 1

```
> CO2=data.frame(Inc,Pre)
> str(CO2)
'data.frame': 50 obs. of  2 variables:
 $ Inc: num  62.6 50.9 45.5 29.5 24.9 59.6 44.3
   41.1 22.8 17.2 ...
 $ Pre: Factor w/ 5 levels "0.0","0.083",..: 1 2 3
   4 5 1 2 3 4 5 ...
> head(CO2,4)
   Inc   Pre
1 62.6   0.0
2 50.9 0.083
3 45.5  0.29
4 29.5  0.50
```

Exemple 1: Donau una ullada a les dades



Exemple 1: Donau una ullada a les dades



Variabilitat de les dades

Exemple 2

Vosaltres l'anireu fent a mà ☺

Notacions

Suposarem les dades donades amb l'estructura següent

Nivell del factor			
1	2	$\cdots$	k
X_{11}	X_{21}	$\cdots$	X_{k1}
X_{12}	X_{22}	$\cdots$	X_{k2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$
X_{1n_1}	X_{2n_2}	$\cdots$	X_{kn_k}

on

- n_i és la mida de la mostra del nivell i
- X_{ij} és el valor de la característica sota estudi a l'individu j del nivell i

ALERTA! Notacions diferents que a les matrius. A X_{ij}, i indica la columna i j la filera

Estadístics

- Suma total de les dades del nivell i-èsim:

$$T_{i\bullet} = \sum_{j=1}^{n_i} X_{ij}$$

- Mitjana mostral per al nivell i-èsim:

$$\overline{X}_{i\bullet} = \frac{T_{i\bullet}}{n_i}$$

- Suma total de les dades:

$$T_{\bullet\bullet} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} = \sum_{i=1}^{k} T_{i\bullet}$$

- Mitjana mostral de totes les dades:

$$\overline{X}_{\bullet\bullet} = \frac{T_{\bullet\bullet}}{N}, \qquad \text{on } N = n_1 + \cdots + n_k$$

Exemple 1

```
> str(CO2)
'data.frame': 50 obs. of  2 variables:
 $ Inc: num  62.6 50.9 45.5 29.5 24.9 59.6
   44.3 41.1 22.8 17.2 ...
 $ Pre: Factor w/ 5 levels "0.0","0.083",..:
   1 2 3 4 5 1 2 3 4 5 ...
```

Exemple 1

Calculam els estadístics:

- Sumes per nivells:

```
> sumes.nivells=
   aggregate(Inc~Pre,data=CO2,sum)
> sumes.nivells
    Pre   Inc
1   0.0 591.4
2 0.083 460.4
3  0.29 364.5
4  0.50 254.7
5  0.86 164.4
```

Exemple 1

Calculam els estadístics:

- Mitjanes per nivells:

```
> mitjanes.nivells=
   aggregate(Inc~Pre,data=CO2,mean)
> mitjanes.nivells
    Pre   Inc
1   0.0 59.14
2 0.083 46.04
3  0.29 36.45
4  0.50 25.47
5  0.86 16.44
```

Exemple 1

Calculam els estadístics:

- Suma total de les dades:

```
> suma.total=sum(CO2$Inc)
> suma.total
[1] 1835.4
```

- Mitjana de totes les dades:

```
> mitjana.total=mean(CO2$Inc)
> mitjana.total
[1] 36.708
```

Exemple 2

Calculau els estadístics:

- Sumes totals de les dades per nivells:

$T_{1\bullet}$	$T_{2\bullet}$	$T_{3\bullet}$	$T_{4\bullet}$

- Mitjanes mostrals per nivells:

$\overline{X}_{1\bullet}$	$\overline{X}_{2\bullet}$	$\overline{X}_{3\bullet}$	$\overline{X}_{4\bullet}$

- Suma total de les dades: $T_{\bullet\bullet} =$
- Mitjana mostral de totes les dades: $\overline{X}_{\bullet\bullet} =$

Exemple 2

Calculau els estadístics:

- Sumes totals de les dades per nivells:

$T_{1\bullet}$	$T_{2\bullet}$	$T_{3\bullet}$	$T_{4\bullet}$
483	428	398	364

- Mitjanes mostrals per nivells:

$\overline{X}_{1\bullet}$	$\overline{X}_{2\bullet}$	$\overline{X}_{3\bullet}$	$\overline{X}_{4\bullet}$
96.6	85.6	79.6	72.8

- Suma total de les dades: $T_{\bullet\bullet} = 1673$
- Mitjana mostral de totes les dades: $\overline{X}_{\bullet\bullet} = 83.65$

El model

Els paràmetres que intervindran en el contrast són:

- μ: mitjana poblacional del conjunt de la població (ignorant els nivells)
- μ_i: mitjana poblacional dins el nivell i-èsim, $i = 1, \ldots, k$

Els estimadors dels paràmetres són els següents:

- De μ: $\overline{X}_{\bullet\bullet}$
- De cada μ_i: $\overline{X}_{i\bullet}$

El model

Les suposicions del model són:

- Les k mostres són m.a.s. independents extretes de k poblacions específiques amb mitjanes $\mu_1, \ldots, \mu_k$
- Cadascuna de les k poblacions segueix una llei normal
- Totes aquestes poblacions tenen la mateixa variància σ^2

El model

Expressió matemàtica del model a estudiar en aquest cas:

$$X_{ij}-\mu=(X_{ij}-\mu_i)+(\mu_i-\mu),\ i=1,\ldots,k,\ j=1,\ldots,n_i,$$

on

- X_{ij}: valor del j-èsim individu dins del nivell i-èsim
- $X_{ij}-\mu$: desviació de l'individu respecte de la mitjana global
- $X_{ij}-\mu_i$: desviació de l'individu respecte de la mitjana del seu grup
- $\mu_i-\mu$: desviació de la mitjana del grup i-èsim respecte de la mitjana global

El model

Expressió matemàtica del model a estudiar en aquest cas:

$$X_{ij}-\mu=(X_{ij}-\mu_i)+(\mu_i-\mu),\ i=1,\ldots,k,\ j=1,\ldots,n_i,$$

i

- $X_{ij}-\overline{X}_{\bullet\bullet}$ estima $X_{ij}-\mu$
- $X_{ij}-\overline{X}_{i\bullet}$ estima $X_{ij}-\mu_i$
- $\overline{X}_{i\bullet}-\overline{X}_{\bullet\bullet}$ estima $\mu_i-\mu$

Identitat de la suma de quadrats

Teorema

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{\bullet\bullet})^2=\sum_{i=1}^{k}n_i(\overline{X}_{i\bullet}-\overline{X}_{\bullet\bullet})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\bullet})^2$$

- $SS_{Total}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{\bullet\bullet})^2$ (Suma Total de Quadrats)
- $SS_{Tr}=\sum_{i=1}^{k}n_i(\overline{X}_{i\bullet}-\overline{X}_{\bullet\bullet})^2$ (Suma de Quadrats dels Tractaments)
- $SS_{E}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{i\bullet})^2$ (Suma de Quadrats dels Residus o Errors)

Identitat de la suma de quadrats

Teorema

$SS_{Total} = SS_{Tr} + SS_E$

- $SS_{Total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{\bullet\bullet})^2$
- $SS_{Tr} = \sum_{i=1}^{k} n_i (\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2$
- $SS_E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{i\bullet})^2$

Identitat de la suma de quadrats

Teorema

$SS_{Total} = SS_{Tr} + SS_E$

A mà, s'empren les fórmules, equivalents, següents:

- $SS_{Total} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}^2 - \frac{T_{\bullet\bullet}^2}{N}$
- $SS_{Tr} = \sum_{i=1}^{k} \frac{T_{i\bullet}^2}{n_i} - \frac{T_{\bullet\bullet}^2}{N}$
- $SS_E = SS_{Total} - SS_{Tr}$

Usualment escriurem, per abreviar,

$$T_{\bullet\bullet}^{(2)} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}^2$$

Exemple 1

- $SS_{Total} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{\bullet\bullet})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{T_{\bullet\bullet}^2}{N}$

```
> SSTotal1=sum((CO2$Inc-mitjana.total)^2)
> SSTotal1
[1] 12522.36
> SSTotal=sum(CO2$Inc^2)-suma.total^2/50
> SSTotal
[1] 12522.36
```

Exemple 1

- $SS_{Tr} = \sum_{i=1}^{k} n_i(\overline{X}_{i\bullet} - \overline{X}_{\bullet\bullet})^2 = \sum_{i=1}^{k} \frac{T_{i\bullet}^2}{n_i} - \frac{T_{\bullet\bullet}^2}{N}$

```
> SSTr1=sum(table(CO2$Pre)*
 (mitjanes.nivells[,2]-mitjana.total)^2)
> SSTr1
[1] 11274.32
> SSTr=sum(sumes.nivells[,2]^2/table(CO2$Pre))
   -(suma.total^2)/50
> SSTr
[1] 11274.32
```

Exemple 1

- $SS_E = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\bullet})^2 = SS_{Total} - SS_{Tr}$

```
> SSE1=sum((CO2$Inc-mitjanes.nivells[,2])^2)
> SSE1
[1] 1248.038
> SSE=SSTotal-SSTr
> SSE
[1] 1248.038
```

Exemple 2

$n_i = 5,\ i = 1, \ldots, 4,\ N = 20$

$T_{1\bullet}$	$T_{2\bullet}$	$T_{3\bullet}$	$T_{4\bullet}$	$T_{\bullet\bullet}$	$T_{\bullet\bullet}^{(2)}$
483	428	398	364	1673	142713

- $SS_{Total} = T_{\bullet\bullet}^{(2)} - \dfrac{T_{\bullet\bullet}^2}{N} =$
- $SS_{Tr} = \displaystyle\sum_{i=1}^{k} \frac{T_{i\bullet}^2}{n_i} - \frac{T_{\bullet\bullet}^2}{N} =$
- $SS_E = SS_{Total} - SS_{Tr} =$

Exemple 2

$n_i = 5,\ i = 1, \ldots, 4,\ N = 20$

$T_{1\bullet}$	$T_{2\bullet}$	$T_{3\bullet}$	$T_{4\bullet}$	$T_{\bullet\bullet}$	$T_{\bullet\bullet}^{(2)}$
483	428	398	364	1673	142713

- $SS_{Total} = T_{\bullet\bullet}^{(2)} - \dfrac{T_{\bullet\bullet}^2}{N} = 2766.55$
- $SS_{Tr} = \displaystyle\sum_{i=1}^{k} \frac{T_{i\bullet}^2}{n_i} - \frac{T_{\bullet\bullet}^2}{N} = 1528.15$
- $SS_E = SS_{Total} - SS_{Tr} = 1238.4$

Estadístics del contrast

Emprarem els estadístics següents:

- Quadrat mitjà dels tractaments:

$$MS_{Tr} = \frac{SS_{Tr}}{k-1}$$

- Quadrat mitjà residual:

$$MS_E = \frac{SS_E}{N-k}$$

Estadístics del contrast

Aquests estadístics són variables aleatòries, i es té que

- $$E(MS_{Tr}) = \sigma^2 + \sum_{i=1}^{k} \frac{n_i(\mu_i - \mu)^2}{k-1}$$
- $$E(MS_E) = \sigma^2$$

En particular, es pot usar MS_E per estimar la variància comuna σ^2

Estadístics del contrast

Aquests estadístics són variables aleatòries, i es té que

- $E(MS_{Tr}) = \sigma^2 + \sum_{i=1}^{k} \frac{n_i(\mu_i - \mu)^2}{k-1}$
- $E(MS_E) = \sigma^2$

En particular, es pot usar MS_E per estimar la variància comuna σ^2

Si $H_0 : \mu_1 = \cdots = \mu_k = \mu$ és certa,

$$\sum_{i=1}^{k} \frac{n_i(\mu_i - \mu)^2}{k-1} = 0,$$

i si H_0 no és certa, aquesta quantitat és > 0

Estadístics del contrast

Per tant

- si H_0 és certa, $E(MS_E) = E(MS_{Tr})$ i hauríem d'esperar que aquests dos estadístics tinguessin valors propers, és a dir
$$\frac{MS_{Tr}}{MS_E} \approx 1$$
- si H_0 és falsa, $E(MS_E) < E(MS_{Tr})$ i hauríem d'esperar que
$$\frac{MS_{Tr}}{MS_E} > 1$$

Estadístics del contrast

Considerarem com a estadístic de contrast el quocient

$$F = \frac{MS_{Tr}}{MS_E}$$

Si H_0 és certa:

- la seva distribució és $F_{k-1,N-k}$ (F de Fisher amb $k-1$ i $N-k$ graus de llibertat)
- el seu valor serà proper a 1

Per tant, rebutjarem la hipòtesi nul·la si F és prou més gran que 1

Contrast ANOVA

- Calculam les sumes de quadrats

$$SS_{Total}, SS_{Tr}, SS_E$$

- Calculam

$$MS_{Tr} = \frac{SS_{Tr}}{k-1},\ MS_E = \frac{SS_E}{N-k},\ F = \frac{MS_{Tr}}{MS_E}$$

- Calculam el p-valor

$$P(F_{k-1,N-k} \geqslant F)$$

- Si el p-valor és més petit que el nivell de significació α rebutjam H_0 i concloem que no totes les mitjanes són iguals. En cas contrari, acceptam H_0.

Exemple 1

- Ja sabem que $N = 50$, $k = 5$, $SS_{Total} = 12522.36$, $SS_{Tr} = 11274.32$ i $SS_E = 1248.038$.
- Els quadrats mitjans són:

```
> N=50; k=5
> MSTr=SSTr/(k-1); MSTr
[1] 2818.58
> MSE=SSE/(N-k); MSE
[1] 27.73418
```

- L'estadístic de contrast F val:

```
> EstF=MSTr/MSE; EstF
[1] 101.6284
```

Exemple 1

- El p-valor $P(F_{k-1,N-k} \geqslant F)$ val

```
> 1-pf(101.6284,4,45)
[1] 0
```

- Per tant, rebutjam H_0 i concloem que el nivell de pressió de CO_2 pot influir en el creixement del microorganisme *Pseudomonas fragi*

  Alerta! Només concloem que no totes les mitjanes són iguals: no que totes les mitjanes són diferents. No és el mateix!

Exemple 2

- Recordem que $k = 4$, $N = 20$, $SS_{Tr} = 1528.15$, $SS_E = 1238.4$

- Quadrats mitjans:

$$MS_{Tr} = \qquad\qquad , MS_E =$$

- Estadístic de contrast

$$F = \frac{MS_{Tr}}{MS_E} =$$

- p-valor

- Per tant,

Exemple 2

- Recordem que $k = 4$, $N = 20$, $SS_{Tr} = 1528.15$, $SS_E = 1238.4$

- Quadrats mitjans:

$$MS_{Tr} = 509.38, MS_E = 77.4$$

- Estadístic de contrast

$$F = \frac{MS_{Tr}}{MS_E} = 6.58$$

- p-valor $P(F_{3,16} \geqslant F) = 0.0042$

- Per tant, rebutjam la hipòtesi nul·la: no tots els tractaments tenen la mateixa efectivitat

Taula ANOVA

El contrast ANOVA es resumeix en la taula següent:

Origen Variació	Graus llibertat	Sumes de quadrats	Quadrats mitjans	Estadístic de contrast	p-valor
Nivell	$k-1$	SS_{Tr}	$MS_{Tr} = \frac{SS_{Tr}}{k-1}$	$F = \frac{MS_{Tr}}{MS_E}$	p-valor
Residu	$N-k$	SS_E	$MS_E = \frac{SS_E}{N-k}$		

Taula ANOVA

El contrast ANOVA es resumeix en la taula següent:

Origen Variació	Graus llibertat	Sumes de quadrats	Quadrats mitjans	Estadístic de contrast	p-valor
Nivell	$k-1$	SS_{Tr}	$MS_{Tr} = \frac{SS_{Tr}}{k-1}$	$F = \frac{MS_{Tr}}{MS_E}$	p-valor
Residu	$N-k$	SS_E	$MS_E = \frac{SS_E}{N-k}$		

Exemple 1

Origen Variació	Graus llibertat	Sumes de quadrats	Quadrats mitjans	Estadístic de contrast	p-valor
Nivell	4	11274.32	2818.58	101.63	0
Residu	45	1248.04	27.73		

Taula ANOVA

El contrast ANOVA es resumeix en la taula següent:

Origen Variació	Graus llibertat	Sumes de quadrats	Quadrats mitjans	Estadístic de contrast	p-valor
Nivell	$k-1$	SS_{Tr}	$MS_{Tr}=\frac{SS_{Tr}}{k-1}$	$F=\frac{MS_{Tr}}{MS_E}$	p-valor
Residu	$N-k$	SS_E	$MS_E=\frac{SS_E}{N-k}$		

Exemple 2

Origen Variació	Graus llibertat	Sumes de quadrats	Quadrats mitjans	Estadístic de contrast	p-valor
Nivell	3	1528.15	509.38	6.58	0.004
Residu	16	1238.4	77.4		

Amb R: Exemple 1

Amb R el contrast ANOVA de l'Exemple 1 es pot realitzar amb

```
> summary(aov(CO2$Inc~CO2$Pre))
            Df Sum Sq Mean Sq F value Pr(>F)
CO2$Pre      4  11274  2818.6   101.6 <2e-16 ***
Residuals   45   1248    27.7
---
Signif. codes:0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1
```

El valor de `Pr(> F)` és el p-valor del contrast

Amb R: Exemple 2

```
> Expr=c(96,93,70,78,99,90,90,87,100,75,84,57,104,
  80,76,66,84,90,78,76)
> Tract=rep(c("A","B","C","D"),5)
> EG=data.frame(Expr,Tract)
> str(EG)
'data.frame': 20 obs. of  2 variables:
 $ Expr : num  96 93 70 78 99 90 90 87 100 75 ...
 $ Tract: Factor w/ 4 levels "A","B","C","D": 1 2 3
  4 1 2 3 4 1 2 ...
```

Amb R: Exemple 2

```
> summary(aov(Expr~Tract,data=EG))
            Df Sum Sq Mean Sq F value  Pr(>F)
EG$Tract     3   1528   509.4   6.581 0.00417 **
Residuals   16   1238    77.4
---
Signif. codes:0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1
```

Comparacions per parelles

Si hem rebutjat la hipòtesi nul·la $H_0 : \mu_1 = \cdots = \mu_k$, podem demanar-nos quins són els nivells diferents

Ho podem fer de diverses maneres, aquí en veurem tres

Comparacions per parelles

Es realitzen els $\binom{k}{2}$ contrastos

$$\left.\begin{array}{ll} H_0 & : \mu_i = \mu_j \\ H_1 & : \mu_i \neq \mu_j \end{array}\right\}$$

L'estadístic de cada contrast és:

$$T = \frac{\overline{X}_{i\bullet} - \overline{X}_{j\bullet}}{\sqrt{MS_E \cdot (\frac{1}{n_i} + \frac{1}{n_j})}}$$

que segueix una distribució t de Student amb $N - k$ graus de llibertat, t_{N-k}

El p-valor de cada contrast és $2P(t_{N-k} \geqslant |t_{i,j}|)$, on $t_{i,j}$ és el valor que hi pren l'estadístic

Comparacions per parelles

Alerta! Si es realitzen c contrastos a un nivell de significació α, la probabilitat d'Error de Tipus I a qualcun és major que α: de fet, és $1-(1-\alpha)^c$

Per exemple, a l'exemple del CO_2, si realitzam $c=\binom{5}{2}=10$ contrastos amb nivell de significació $\alpha=0.05$, la probabilitat d'Error de Tipus I a qualcun és $1-(1-0.05)^{10}\approx 0.4$.

Haurem de reduir el nivell de significació de cada contrast perquè la probabilitat final d'Error de Tipus I sigui α

Test T de Bonferroni

Emprarem l'aproximació $1-(1-x)^c \approx cx$ i aleshores, si volem efectuar c contrastos amb nivell de significació (global) α, els farem amb nivell de significació α/c

Per exemple, a l'exemple del CO_2, si realitzam els 10 contrastos, per obtenir un nivell de significació global $\alpha = 0.05$, efectuam cada contrast amb nivell de significació 0.005

Test T de Holm (més potent)

1. Siguin $C_1, \ldots, C_c$ els contrastos i $P_1, \ldots, P_c$ els p-valors corresponents
2. Ordenam aquests p-valors en ordre creixent $P_{(1)} \leqslant \cdots \leqslant P_{(c)}$ i reenumeram consistentment els contrastos $C_{(1)}, \ldots, C_{(c)}$
3. Per a cada $j = 1, \ldots, c$, calculam el p-valor ajustat $\widetilde{P}_{(j)} = (c + 1 - j)P_{(j)}$.
4. Aleshores rebutjam la hipòtesi nul·la als contrastos $C_{(j)}$ on $\widetilde{P}_{(j)} < \alpha$

Exemple 1

$\alpha = 0.05$

Fem els càlculs per passes

En primer lloc cream una matriu amb fileres les parelles de nivells les mitjanes dels quals contrastarem:

```
> pars=rbind(c(1,2),c(1,3),c(1,4),c(1,5),
   c(2,3),c(2,4),c(2,5),c(3,4),c(3,5),c(4,5))
```

Exemple 1

A continuació calculam els valors de tots els estadístics de contrast

```
> est.contrast.par =
 (mitjanes.nivells[pars[,1],2]
      -mitjanes.nivells[pars[,2],2])/
  (sqrt(MSE*(1/10+1/10)))
> est.contrast.par
 [1]  5.562226  9.634116 14.296195 18.130310
 [5] 4.071889  8.733969 12.568084 4.662080
 [9] 8.496194  3.834115
```

Exemple 1

Els afegim com a columna a la matriu de parelles de nivells

```
> pars=cbind(pars,est.contrast.par)
> pars
            est.contrast.par
 [1,] 1 2          5.562226
 [2,] 1 3          9.634116
 [3,] 1 4         14.296195
 [4,] 1 5         18.130310
 [5,] 2 3          4.071889
 [6,] 2 4          8.733969
 [7,] 2 5         12.568084
 [8,] 3 4          4.662080
 [9,] 3 5          8.496194
[10,] 4 5          3.834115
```

Exemple 1

Calculam els p-valors:

```
> p.valCO2=function(x){2*(1-pt(abs(x),N-k))}
> p.vals=sapply(est.contrast.par,p.valCO2)
> p.vals
 [1] 1.387522e-06 1.649791e-12 0.000000e+00
 [4] 0.000000e+00 1.863744e-04 3.021050e-11
 [7] 2.220446e-16 2.808032e-05 6.609646e-11
[10] 3.892218e-04
```

Exemple 1

Ho afegim com a columna a la matriu de parelles de nivells i estadístics

```
> pars=cbind(pars,p.vals)
> pars
          est.contrast.par       p.vals
 [1,] 1 2         5.562226 1.387522e-06
 [2,] 1 3         9.634116 1.649791e-12
 [3,] 1 4        14.296195 0.000000e+00
 [4,] 1 5        18.130310 0.000000e+00
 [5,] 2 3         4.071889 1.863744e-04
 [6,] 2 4         8.733969 3.021050e-11
 [7,] 2 5        12.568084 2.220446e-16
 [8,] 3 4         4.662080 2.808032e-05
 [9,] 3 5         8.496194 6.609646e-11
[10,] 4 5         3.834115 3.892218e-04
```

Exemple 1

Bonferroni: Quins p-valors són inferiors a 0.005?

```
> pars[which(p.vals<0.005),c(1,2)]
 [1,] 1 2
 [2,] 1 3
 [3,] 1 4
 [4,] 1 5
 [5,] 2 3
 [6,] 2 4
 [7,] 2 5
 [8,] 3 4
 [9,] 3 5
[10,] 4 5
```

Tots. Per tant, rebutjam totes les hipòtesis nul·les, i concloem que els nivells tenen mitjanes diferents dues a dues

Exemple 1

Holm: Ordenam les fileres de pars ordenant els p-valors de menor a major:

```
> pars.ord=pars[order(pars[,4]),]
> pars.ord
           est.contrast.par       p.vals
 [1,] 1 4         14.296195 0.000000e+00
 [2,] 1 5         18.130310 0.000000e+00
 [3,] 2 5         12.568084 2.220446e-16
 [4,] 1 3          9.634116 1.649791e-12
 [5,] 2 4          8.733969 3.021050e-11
 [6,] 3 5          8.496194 6.609646e-11
 [7,] 1 2          5.562226 1.387522e-06
 [8,] 3 4          4.662080 2.808032e-05
 [9,] 2 3          4.071889 1.863744e-04
[10,] 4 5          3.834115 3.892218e-04
```

Exemple 1

Holm: Calculam els p-valors ajustats i els afegim com a columna a `pars.ord`

```
> p.vals.adjust=pars.ord[,4]*(10+1-1:10)
> pars.ord=cbind(pars.ord,p.vals.adjust)
> pars.ord
             est.contrast.par        p.vals p.vals.adjust
 [1,] 1 4           14.296195 0.000000e+00  0.000000e+00
 [2,] 1 5           18.130310 0.000000e+00  0.000000e+00
 [3,] 2 5           12.568084 2.220446e-16  1.776357e-15
 [4,] 1 3            9.634116 1.649791e-12  1.154854e-11
 [5,] 2 4            8.733969 3.021050e-11  1.812630e-10
 [6,] 3 5            8.496194 6.609646e-11  3.304823e-10
 [7,] 1 2            5.562226 1.387522e-06  5.550090e-06
 [8,] 3 4            4.662080 2.808032e-05  8.424096e-05
 [9,] 2 3            4.071889 1.863744e-04  3.727488e-04
[10,] 4 5            3.834115 3.892218e-04  3.892218e-04
```

Exemple 1

Holm: A quins contrastos $C_{(k)}$ tenim que $\widetilde{P}_k \leqslant 0.05$?

```
> pars.ord[which(pars.ord[,5]<=0.05),c(1,2)]
 [1,] 1 4
 [2,] 1 5
 [3,] 2 5
 [4,] 1 3
 [5,] 2 4
 [6,] 3 5
 [7,] 1 2
 [8,] 3 4
 [9,] 2 3
[10,] 4 5
```

A tots, per tant rebutjam totes les hipòtesis nul·les, i concloem que els nivells tenen mitjanes diferents dues a dues

Exemple 1

Amb R, per calcular tots els p-valors de cop podem fer

```
> pairwise.t.test(CO2$Inc,CO2$Pre,
 p.adjust.method="none")

Pairwise comparisons using t tests
with pooled SD

data:  CO2$Inc and CO2$Pre

      0.0     0.083   0.29    0.50
0.083 1.4e-06 -       -       -
0.29  1.6e-12 0.00019 -       -
0.50  < 2e-16 3.0e-11 2.8e-05 -
0.86  < 2e-16 2.5e-16 6.6e-11 0.00039

P value adjustment method: none
```

Exemple 1

Si no ens volem preocupar de dividir, fem `p.adjust.method="bonferroni"` i R multiplica els p-valors obtinguts pel nombre de comparacions, i això ha de ser més petit que α

```
> pairwise.t.test(CO2$Inc,CO2$Pre,
   p.adjust.method="bonferroni")

...
      0.0      0.083   0.29    0.50
0.083 1.4e-05 -        -       -
0.29  1.6e-11 0.00186 -       -
0.50  < 2e-16 3.0e-10 0.00028 -
0.86  < 2e-16 2.5e-15 6.6e-10 0.00389

P value adjustment method: bonferroni
```

Exemple 1

Fent `p.adjust.method="holm"` (Alerta! és el per defecte) dóna els p-valors ajustats del mètode de Holm

```
> pairwise.t.test(CO2$Inc,CO2$Pre,
   p.adjust.method="holm")

...
      0.0      0.083   0.29    0.50
0.083 5.6e-06 -        -       -
0.29  1.2e-11 0.00037 -        -
0.50  < 2e-16 1.8e-10 8.4e-05 -
0.86  < 2e-16 2.0e-15 3.3e-10 0.00039

P value adjustment method: holm
```

Exemple 2

- $N = 20,\ k = 4$
- Mitjanes mostrals per nivells:

$\overline{X}_{1\bullet}$	$\overline{X}_{2\bullet}$	$\overline{X}_{3\bullet}$	$\overline{X}_{4\bullet}$
96.6	85.6	79.6	72.8

- $MS_E = 77.4$

Decidiu quines parelles de mitjanes són diferents amb $\alpha = 0.05$

Exemple 2

		Est. contr.	p-valor
1	2		
1	3		
1	4		
2	3		
2	4		
3	4		

Exemple 2

		Est. contr.	p-valor
1	2	1.977	0.0655
1	3	3.055	0.0076
1	4	4.277	0.0006
2	3	1.078	0.2968
2	4	2.3	0.0352
3	4	1.222	0.2394

Exemple 2

		Est. contr.	p-valor
1	2	1.977	0.0655
1	3	3.055	0.0076
1	4	4.277	0.0006
2	3	1.078	0.2968
2	4	2.3	0.0352
3	4	1.222	0.2394

Bonferroni: Només les comparacions (1,3) i (1,4) davallen de $0.05/6 = 0.0083$, són les úniques per a les que tenim evidència que són diferents. És a dir, concloem que $\mu_A \neq \mu_C$, $\mu_A \neq \mu_D$ i que no podem afirmar que les altres siguin diferents.

Exemple 2

Holm:

		Est. contr.	p-valor	p-valor ajustat
1	4	4.277	0.0006	
1	3	3.055	0.0076	
2	4	2.3	0.0352	
1	2	1.977	0.0655	
3	4	1.222	0.2394	
2	3	1.078	0.2968	

Exemple 2

Holm:

		Est. contr.	p-valor	p-valor ajustat
1	4	4.277	0.0006	0.0036
1	3	3.055	0.0076	0.0380
2	4	2.3	0.0352	0.1408
1	2	1.977	0.0655	0.1965
3	4	1.222	0.2394	0.4788
2	3	1.078	0.2968	0.2968

Rebutjam la hipòtesi nul·la a (1,4) i (1,3) i l'acceptam a tots els altres

Exemple 2

```
> pairwise.t.test(EG$Expr,EG$Tract,
  p.adjust.method="bonferroni")
...
  A      B      C
B 0.3933 -      -
C 0.0453 1.0000 -
D 0.0035 0.2113 1.0000

P value adjustment method: bonferroni
```

Amb l'ajustament fet, els únics que davallen de 0.05 són (A,C) i (A,D)

Exemple 2

```
> pairwise.t.test(EG$Expr,EG$Tract,
  p.adjust.method="holm")
...
  A      B      C
B 0.1966 -      -
C 0.0378 0.4787 -
D 0.0035 0.1409 0.4787

P value adjustment method: holm
```

Amb l'ajustament de Holm, els únics que davallen de 0.05 també són (A,C) i (A,D)

Contrast de Duncan

El contrast de Duncan és un altre mètode per veure en quins nivells hi ha diferències

Es realitza amb la funció `duncan.test` del paquet `agricolae`. La sintaxi és

```
duncan.test(aov,"factor",group=...)$sufix
```

on

- `aov` és el resultat de l'ANOVA de partida
- El `factor` és el factor de l'ANOVA
- `group` pot ser `TRUE` o `FALSE`, presenta el resultat de forma diferent
- El `sufix` és `group` si `group=TRUE` i `comparison` si `group=FALSE`

Exemple 2

```
> install.packages("agricolae",dep=TRUE)
> library("agricolae")
> EG.aov=aov(EG$Expr~EG$Tract)
> duncan.test(EG.aov,"EG$Tract",
  group=FALSE)$comparison
      Difference   pvalue sig.        LCL      UCL
A - B       11.0 0.065545    . -0.7955155 22.79552
A - C       17.0 0.009774   **  4.6308251 29.36917
A - D       23.8 0.000971  *** 11.0722298 36.52777
B - C        6.0 0.296877      -5.7955155 17.79552
B - D       12.8 0.042338    *  0.4308251 25.16917
C - D        6.8 0.239368      -4.9955155 18.59552
```

Dóna uns p-valors: els petits permeten rebutjar la hipòtesi nul·la corresponent

Exemple 2

```
> duncan.test(EG.aov,"EG$Tract",group=TRUE)$groups
  trt means   M
1   A  96.6   a
2   B  85.6  ab
3   C  79.6  bc
4   D  72.8   c
```

Diu que B i A, C i B, i C i D no són significativament diferents (els altres sí)

Exemple 1

```
> CO2.aov=aov(CO2$Inc~CO2$Pre)
> duncan.test(CO2.aov,"CO2$Pre",
  group=FALSE)$comparison
             Difference   pvalue sig.      LCL      UCL
0.0 - 0.083       13.10 0.000001  ***  8.35644 17.84356
0.0 - 0.29        22.69 0.000000  *** 17.70152 27.67848
0.0 - 0.50        33.67 0.000000  *** 28.52085 38.81915
0.0 - 0.86        42.70 0.000000  *** 37.43466 47.96534
0.083 - 0.29       9.59 0.000186  ***  4.84644 14.33356
0.083 - 0.50      20.57 0.000000  *** 15.58152 25.55848
0.083 - 0.86      29.60 0.000000  *** 24.45085 34.74915
0.29 - 0.50       10.98 0.000028  ***  6.23644 15.72356
0.29 - 0.86       20.01 0.000000  *** 15.02152 24.99848
0.50 - 0.86        9.03 0.000389  ***  4.28644 13.77356
```

Exemple 1

```
> duncan.test(CO2.aov,"CO2$Pre",group=TRUE)$groups
    trt  means M
1 0.0    59.14 a
2 0.083 46.04 b
3 0.29  36.45 c
4 0.50  25.47 d
5 0.86  16.44 e
```

Efectes aleatoris

En el model d'efectes fixats, l'experimentador tria els nivells a estudiar

Quan el nombre de nivells és molt gran, i es vol esbrinar si els nivells del factor tenen influència en el valor mitjà del paràmetre amb el contrast

$$\begin{cases} H_0 : \text{Les mitjanes de tots els nivells són iguals} \\ H_1 : \text{No és veritat que}\ldots \end{cases}$$

una possibilitat és triar una m.a.s. de k nivells, i aplicar ANOVA a aquests nivells

És el model d'efectes aleatoris

Efectes aleatoris

Les suposicions del model són:

- Els k nivells triats formen una m.a.s. del conjunt de nivells
- Les mitjanes μ_i de tots els nivells segueixen una distribució normal amb valor mitjà μ (el valor mitjà de tota la població) i desviació típica σ_{Tr}
- Totes les poblacions, per a tots els nivells, segueixen lleis normals
- Totes les poblacions, per a tots els nivells, tenen la mateixa variància σ^2
- Les k mostres són m.a.s. independents extretes de les k poblacions triades

Efectes aleatoris

Calculam MS_{Tr} i MS_E com abans. Amb les hipòtesis anteriors, en aquest cas

- $E(MS_{Tr}) = \sigma^2 + \dfrac{N - \sum_{i=1}^{k} \frac{n_i^2}{N}}{k-1} \cdot \sigma_{Tr}^2$
- $E(MS_E) = \sigma^2$

Si H_0 és certa, totes les mitjanes de tots els nivells són iguals, és a dir, $\sigma_{Tr}^2 = 0$, i per tant

$$F = \frac{MS_{Tr}}{MS_E} \approx 1$$

Aquest F té distribució $F_{k-1,N-k}$ si H_0 és certa

Per tant el test és el mateix que al cas d'efectes fixats, però amb els nivells seleccionats