Supòsits de l'ANOVA 1 factor

- Aleatorietat = cada mostra és una mostra aleatòria simple de la seva població
- Homocedasticitat = igualtat de variàncies en tots els tractaments
- Normalitat dels residus
- Independència de les observacions.

un disseny *balancejat* pot esmorteir en part l'impacte que té sobre el nivell de significació real les desviacions moderades de la *homocedasticitat*:

Test d'hipòtesis d'homocedasticitat

- Hi ha diferents tests per comprovar la homocedasticitat, per exemple, el **test de Bartlett** o el **test de Levene**.

- H_0: els grups presenten variàncies homogènies
- H_1: els grups no presenten variàncies homogènies

- El test de Bartlett requereix que les dades siguin normals. El test de Levene és més robust a desviacions de la normalitat.

- En R, tenim les funcions bartlett.test (package stats, es carrega per defecte) i leveneTest (package car)

Test de Bartlett

H_0: $\sigma_1{}^2 = \sigma_2{}^2 = \cdots = \sigma_a{}^2$
H_1: alguna variància és diferent

Estadístic de prova

$$\chi_0^2 = 2.3026\,\frac{q}{c}$$

$$q = (N-a)\log_{10} S_p^2 - \sum_{i=1}^{a}(n_i-1)\log_{10} S_i^2$$

$$c = 1 + \frac{1}{3(a-1)}\left(\sum_{i=1}^{a}(n_i-1)^{-1} - (N-a)^{-1}\right)$$

$$S_p^2 = \frac{\sum_{i=1}^{a}(n_i-1)S_i^2}{N-a}$$

on $S_p{}^2$ és la variància mostral del tractament *i*

Distribució de referència

Xi-quadrat amb *a-1* graus de llibertat

Test de Levene

$$H_0:\ \sigma_1{}^2 = \sigma_2{}^2 = \cdots = \sigma_a{}^2$$
$$H_1: \text{alguna variància és diferent}$$

1. Calculem la mediana de cada tractament, $\tilde{y}_i$

2. Calculem les desviacions absolutes de cada dada dins cada tractament respecte la seva mediana:

$$d_{ij} = \left|y_{ij} - \tilde{y}_i\right| \begin{cases} i = 1,2, \dots, a \\ j = 1,2, \dots, n_i \end{cases}$$

3. Trobem el p-valor amb la ANOVA d'un factor, fent servir com a tractaments les desviacions calculades.

Validació gràfica igualtat variàncies

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

Amb dades de l'exemple 2 (productivitat mitjana per hora)

Dades:

mostra	A	B	C
	2,6	3,2	2,6
	2,1	2,7	2,1
	3,5	3,9	3,1
	2,6	3,4	2,7
mitjanes	2,7	3,3	2,625

Residus:

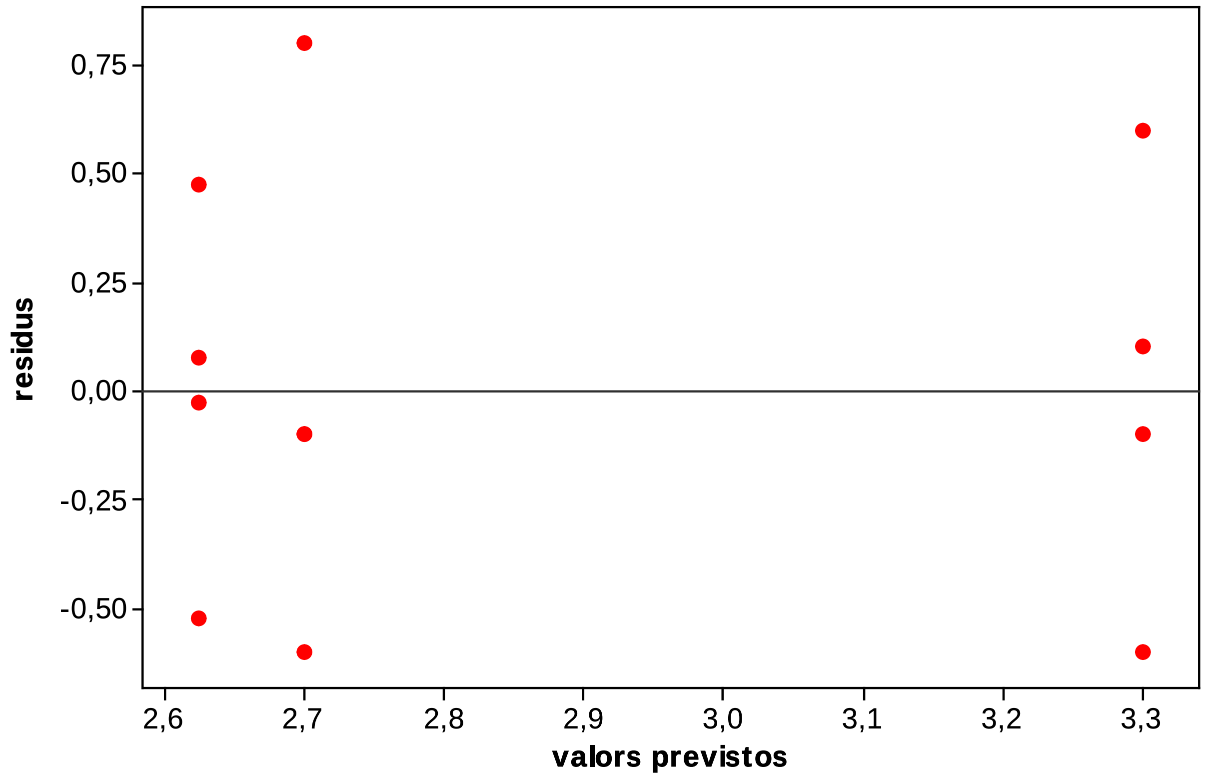
2,6 – 2,7 = – 0,1

A	B	C
-0,100	-0,100	-0,025
-0,600	-0,600	-0,525
0,800	0,600	0,475
-0,100	0,100	0,075

3,1 – 2,625 = 0,475

Residus enfront de valors previstos

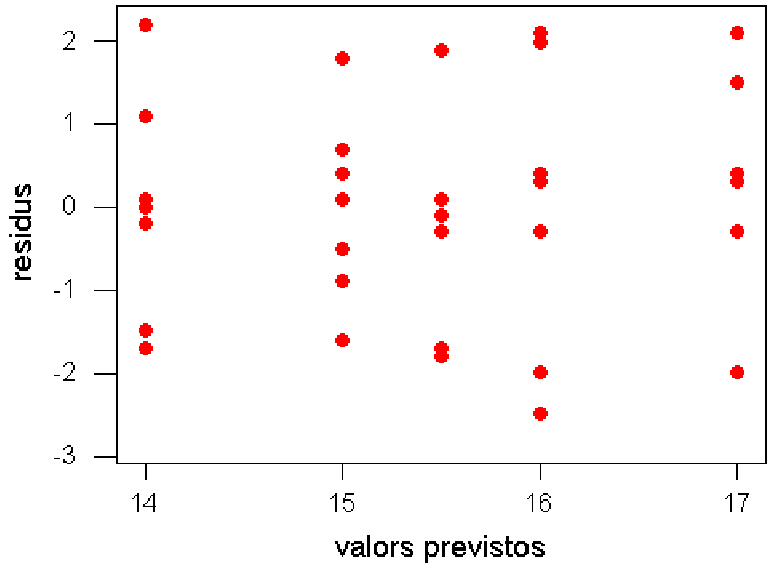
Representem els residus respecte els valors previstos per cada tractament:



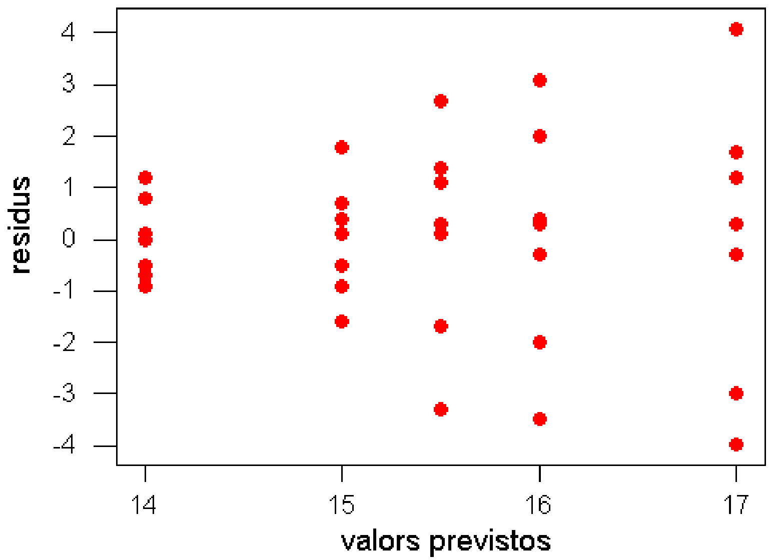
No sembla que la variabilitat augmenti o disminueixi amb el nivell de la resposta: s'acompleix el supòsit d'igualtat de variances poblacionals

Comprovació gràfica igualtat variàncies

Bé!
La variabilitat es manté constant a mesura que augmenta la resposta.



Malament! Hi ha heterocedasticitat.
La variabilitat augmenta a mesura que augmenta la resposta.



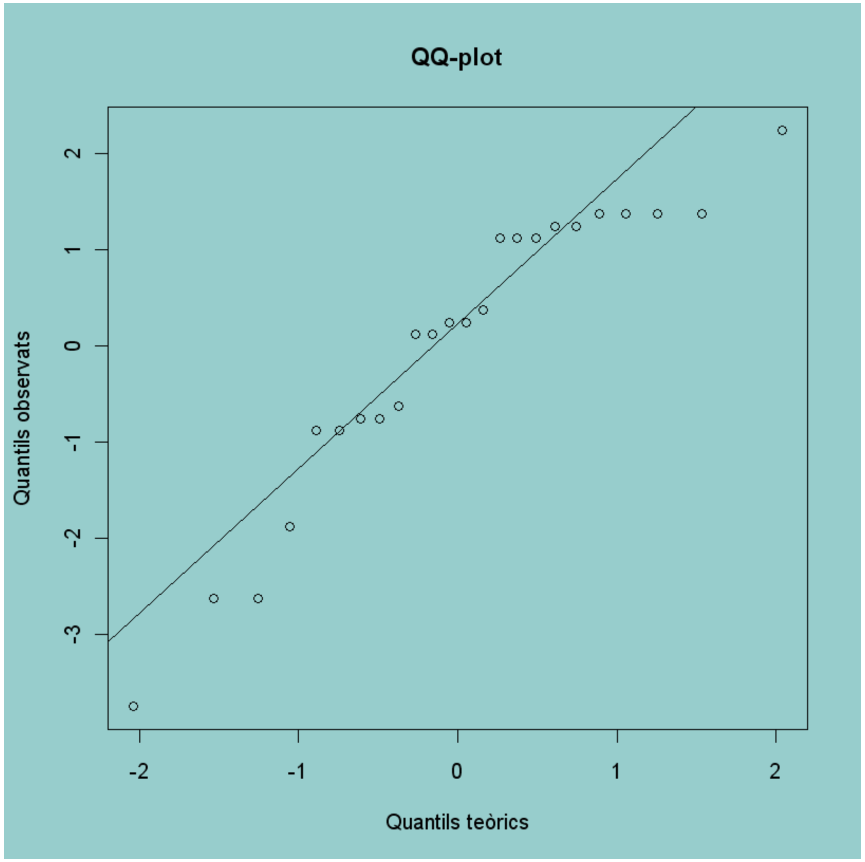
Normalitat dels residus

- La normalitat dels **residus** és un altre requisit. En el model d'un factor els residus corresponen a restar a cada observació la mitjana del seu grup

$$e_{ij} = y_{ij} - \bar{y}_{i\cdot}$$

- Malgrat existeixen en la bibliografia testos de normalitat (i.e. Shapiro-Wilk) és més informatiu en mostres petites traçar un ***qq-plot***

QQ-plot



Independència

- Existeixen en la bibliografia testos que permeten per exemple identificar la presencia de ratxes. No poden testar de forma absoluta la independència de les mostres.

- La millor forma de garantir la independència és amb un model adient de mostratge i amb l'aleatorització prèvia a l'assignació dels tractaments i també, si s'escau, en el moment de la lectura de la variable resposta.

Què fer si les variàncies no són iguals?

Hi ha diverses possibilitats:

- Transformar les dades. Sovint, fer el logaritme de les dades ajuda a estabilitzar la variància.
- Ponderar les observacions. Es pot donar més pes a les dades dels tractaments que tenen menys variància i menys pes als tractaments amb més variància. Això es fa abordant l'anàlisi de la variància des de la perspectiva dels models lineals, amb una regressió amb pesos (weighted regression).
- Fer servir mètodes no paramètrics. En particular, l'equivalent no paramètric de la taula ANOVA d'un factor és el test de Kruskal-Wallis.
- Fer servir una versió de la taula ANOVA (Welch ANOVA) que permet que les variàncies no siguin constants. Això està implementat a R en la funció oneway.test (package stats).

Transformar les dades: un exemple

Uns dies de pluges intenses fan que un riu tingui una gran crescuda. Un enginyer està interessat en determinar si 4 mètodes d'estimar la crescuda d'un riu produeixen els mateixos valors de descàrrega màxima d'aigua.

Cada mètode es fa servir 6 vegades (tenim per tant 6rèpliques de cada mètode). Les dades estan en peus cúbics per segon.

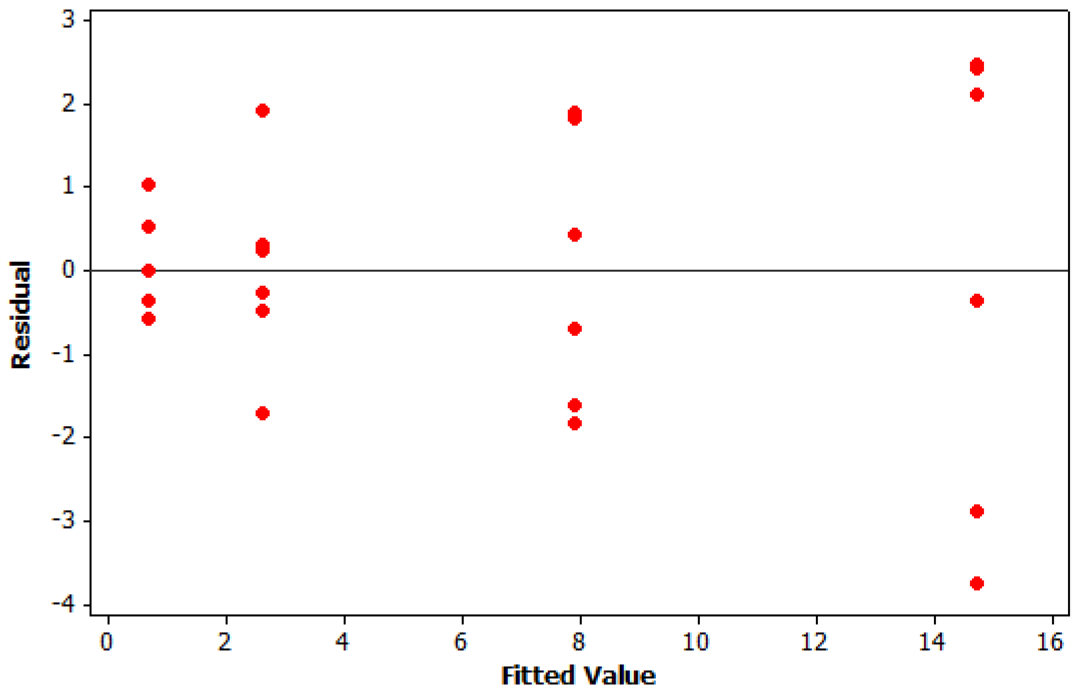
Mètode d'estimació						
1	0,34	0,12	1,23	0,70	1,75	0,12
2	0,91	2,94	2,14	2,36	2,86	4,55
3	6,31	8,37	9,75	6,09	9,82	7,24
4	17,15	11,82	10,95	17,20	14,35	16,82

Transformar les dades: un exemple

El resultat amb la taula ANOVA és:

Source	DF	SS	MS	F	P
Factor	3	708,35	236,12	76,07	0,000
Error	20	62,08	3,10		
Total	23	770,43			

La gràfica de residus respecte a valors previstos mostra que el supòsit d'homocedasticitat no s'acompleix.



Test de Bartlett i Levene

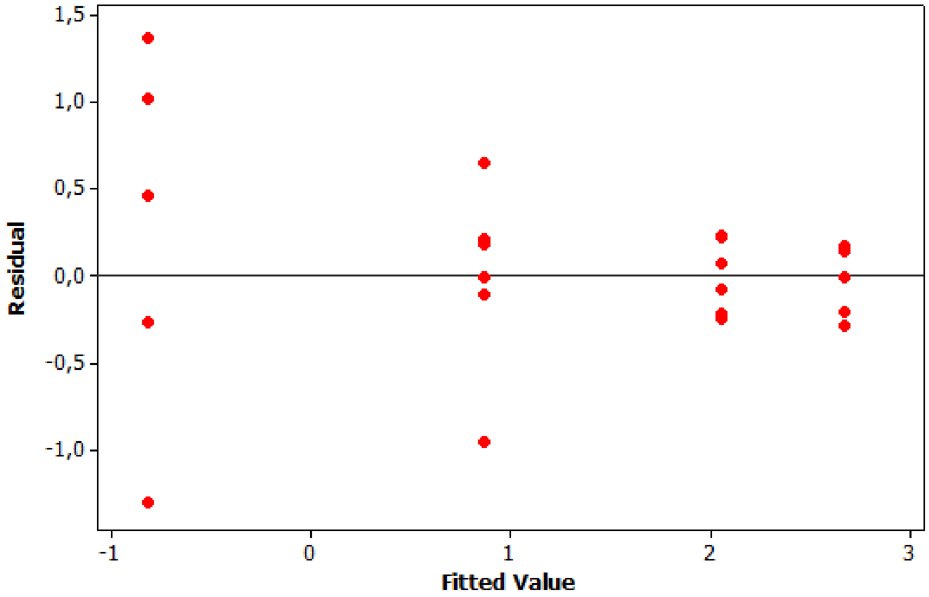
```
> bartlett.test(Data~Treatment, data=water)

 Bartlett test of homogeneity of variances

data:  Data by Treatment
Bartlett's K-squared = 8.9958, df = 3, p-value = 0.02935


> library(car)
> leveneTest(Data~Treatment, data=water)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  3  4.5684 0.01357 *
      20
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tant el test de Bartlett com el de Levene mostren que no s’acompleix el supòsit d’homocedasticitat

Transformar les dades: un exemple

Transformant amb el logaritme…

Source	DF	SS	MS	F	P
Treatment	3	42,499	14,166	33,43	0,000
Error	20	8,475	0,424		
Total	23	50,973			

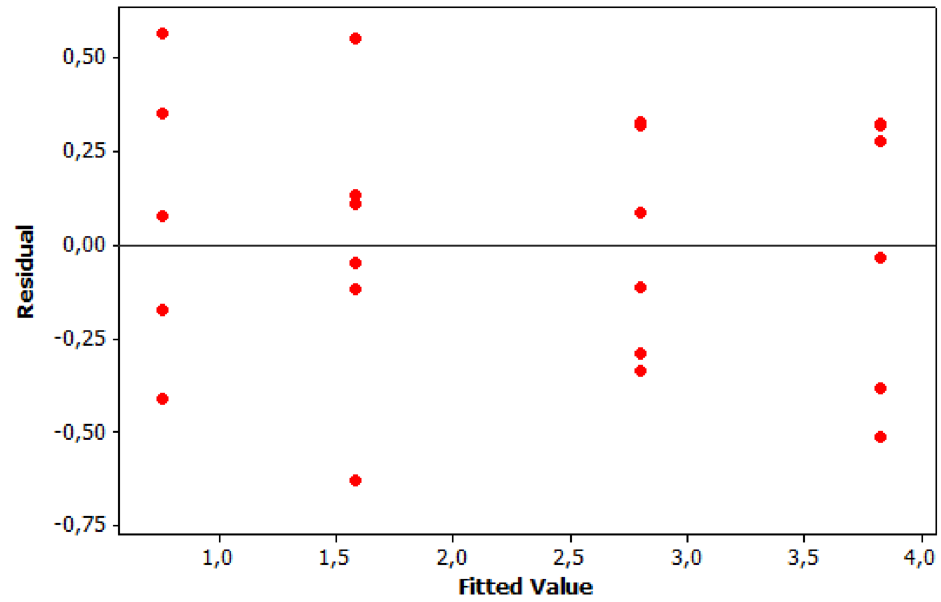


Els residus no milloren.

Transformar les dades: un exemple

Transformant amb l'arrel quadrada

Source	DF	SS	MS	F	P
Treatment	3	32,684	10,895	81,05	0,000
Error	20	2,688	0,134		
Total	23	35,373			



Els residus milloren molt. Ens quedem amb aquesta transformació.

Welch ANOVA

D'un white paper de Minitab (googlejar Welch ANOVA per més informació…)

Random samples of sizes n_1, …, n_k from k populations are observed. Let $\mu_1,\ldots,\mu_k$ denote the population means and let $\sigma_1^2,\ldots,\sigma_k^2$ denote the population variances. Let $\bar{x}_1,\ldots,\bar{x}_k$ denote the sample means and let $s_1^2,\ldots,s_k^2$ denote the sample variances. We are interested in testing the hypotheses:

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

H_1: $\mu_i \neq \mu_j$ for some i, j.

The Welch test for testing the equality of k means compares the statistic

$$W^* = \frac{\sum_{j=1}^{k} w_j(\bar{x}_j - \hat{\mu})^2/(k-1)}{1+[2(k-2)/(k^2-1)]\sum_{j=1}^{k} h_j}$$

to the F(k − 1, f) distribution, where

$$w_j = \frac{n_j}{s_j^2},$$

$$W = \sum_{j=1}^{k} w_j,$$

$$\hat{\mu} = \frac{\sum_{j=1}^{k} w_j \bar{x}_j}{W},$$

$$h_j = \frac{(1-w_j/W)^2}{n_j-1}, \text{ and}$$

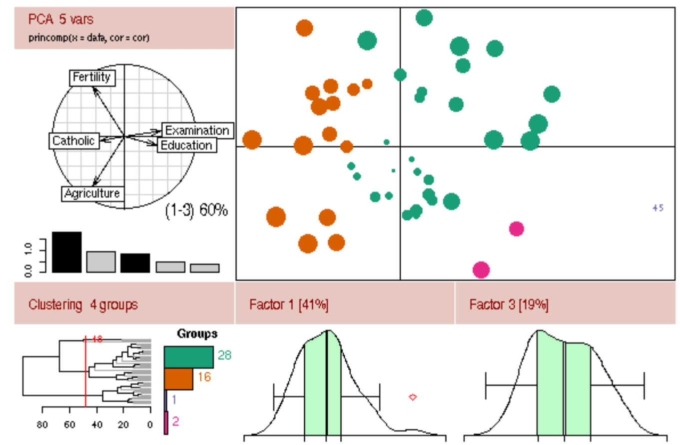
$$f = \frac{k^2-1}{3\sum_{j=1}^{k} h_j}.$$

The Welch test rejects the null hypothesis if $W^* \geq F_{k-1,f,1-\alpha}$, the percentile of the F distribution that is exceeded with probability α.

En R: comanda oneway.test (package stats), amb paràmetre var.equal=FALSE

El model amb efectes aleatoris

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



Exemple: situació experimental

- Es desitja comparar si el procediment operatori i post-operatori associat a una intervenció vertebral és igualment eficaç quan és aplicat per diferents equips de cirurgia de la xarxa hospitalària.
- Es seleccionen a l'atzar quatre hospitals (de forma genèrica correspondrien a 4 *tractaments)*
- La variable resposta mesura els dies fins aconseguir l'alta definitiva. Assumirem la seva normalitat.
- Observem que el nombre de nivells estudiats (4 hospitals) és inferior al nombre de *tractaments possibles* (tots els hospitals)

Estem en un cas d'ANOVA d'un factor amb efectes aleatoris

Resultats experiment

Observació	Hosp 1	Hosp 2	Hosp 3	Hosp 4
1	98	91	96	95
2	97	90	95	96
3	99	93	97	99
4	96	92	95	98
$\sum_{j=1}^{r} y_{ij} = y_{i\bullet}$	390	366	383	388
$\sum_{j=1}^{r} y_{ij}^2$	38030	33494	36675	37646

- La notació és la mateixa que en el model d'efectes fixos

El model d'efectes aleatoris

- 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

- L'equació és idèntica al model d'efecte fix, ara bé, els paràmetres tenen les següents especificacions:

τ_i segueix una Normal $(0, \sigma_\tau)$

ε_{ij} segueix una Normal $(0, \sigma_\varepsilon)$

τ_i i ε_{ij} són independents

- L'**objectiu primari** del disseny és estimar σ_τ i σ_ε, **no** té sentit estimar els efectes individuals τ_i

Taula Anova efectes aleatoris

- L'objectiu del disseny és el de comprovar la contribució relativa de cada font de variació a la variabilitat total:

$$E(y_{ij}) = E(\mu) + E(\tau_i) + E(\varepsilon_{ij}) = \mu$$
$$var(y_{ij}) = var(\mu) + var(\tau_i) + var(\varepsilon_{ij}) = \sigma_\tau^2 + \sigma_\varepsilon^2$$

- En general els models d'efectes fixos i d'efectes aleatoris presenten taules ANOVA diferents. En el cas d'un sol factor les dues taules son idèntiques.

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

Components de la variància

Font de variació	g.d.l.	Quadrats mitjans	$E(MS)$
Entre tractaments	$a-1$	MS_T	$\sigma_\varepsilon^2 + n\sigma_\tau^2$
Dins tractaments	$n-a$	MS_R	σ_ε^2

- Estimadors **(no esbiaixats)** de les variàncies residuals i del tractament

$$MS_T = \sigma_\varepsilon{}^2 + n\sigma_\tau{}^2$$
$$MS_R = \sigma_\varepsilon{}^2$$

Per tant:

$$\widehat{\sigma_\varepsilon{}^2} = MS_R$$
$$\widehat{\sigma_\tau{}^2} = \frac{MS_T - MS_R}{n}$$

Resultats exemple

Font de variació	Suma de quadrats	Graus de llibertat	Quadrats mitjans	F	Pvalor
Entre tractaments	89.19	3	29.73	15.68	0.0001878
Error	22.75	12	1.9		
Total	111.94	15			

- Estimadors de les variàncies entre tractaments i residual:

$$\widehat{\sigma_\varepsilon^{\ 2}} = 1.9 \qquad \widehat{\sigma_\tau^{\ 2}} = \frac{29.73 - 1.9}{4} = 6.96$$

- la variabilitat entre tractaments explica un 78,5% de la variabilitat total

Resultats exemple (amb R)

Amb la funció lmer (paquet lme4) de R:

```
> result <- lmer(Dies~1+1|Hospital, data=hospital)
> result
Linear mixed model fit by REML
Formula: Dies ~ 1 + 1 | Hospital
   Data: hospital
   AIC   BIC logLik deviance REMLdev
 69.19 71.51 -31.60    65.62   63.19
Random effects:
 Groups   Name        Variance Std.Dev.
 Hospital (Intercept) 6.9583   2.6379
 Residual             1.8958   1.3769
Number of obs: 16, groups: Hospital, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept)   95.437      1.363   70.01
> plot(fitted(result), residuals(result))
```