ANALISI DELLA VARIANZA

Per il confronto tra le medie aritmetiche di più gruppi, non è possibile ricorrere al test **t**, suddividendo l'analisi in tanti confronti a coppie quante sono le combinazioni degli **n** gruppi 2 a 2.

Se i gruppi sono numerosi, la probabilità complessiva che almeno uno di essi sia significativo per caso aumenta proporzionalmente (ad es., con α=0,05 e 20 confronti, mediamente uno risulterà significativo per caso, pur essendo vera H_0)

Nel confronto tra più medie, H_0 e H_1 assumono la formulazione :

H_0: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$

le medie delle popolazioni dalle quali sono estratti casualmente i campioni sono tra loro uguali

H_1: non tutte le medie aritmetiche sono uguali

si possono realizzare varie situazioni, e le più estreme sono:

- le medie sono tutte differenti tra loro
- una sola media è diversa dalle altre, tra loro uguali

Per verificare la significatività delle differenze tra le medie aritmetiche di vari gruppi si conduce un'**ANALISI DELLA VARIANZA** (sintetizzato in ANOVA, acronimo di **AN**alysis **O**f **VA**riance)

La distribuzione utilizzata è la distribuzione F in onore di Sir Ronald Aylmer Fisher (1890-1962), il più eminente statistico contemporaneo padre della statistica moderna

La metodologia attuale del test F è dovuta a Snedecor, un allievo di Fisher che ne perfezionò il metodo e ne semplificò la forma

Nel 1925 Fisher completò il metodo di Student per il confronto tra due medie, elaborando nel contempo il concetto di gdl: è suo il metodo attualmente utilizzato

ANOVA è la metodologia alla base della statistica moderna : gli stessi principi si applicano dalle analisi più semplici a quelle più complesse dell'analisi multivariata

IN ANOVA :

- si posono scomporre e misurare con precisione le fonti di variazioni sui valori osservati di due o più gruppi
- la fonte di variazione è detta FATTORE SPERIMENTALE (o TRATTAMENTO) e può essere a più livelli
- ogni unità od osservazione del fattore sperimentale è detta REPLICAZIONE

ANOVA AD UN CRITERIO DI CLASSIFICAZIONE (COMPLETAMENTE RANDOMIZZATA)

E’ il modello più semplice di ANOVA

E’ così chiamato in quanto si confrontano due o più livelli dello stesso fattore

E' detto anche MODELLO COMPLETAMENTE RANDOMIZZATO :

- prevede un campionamento in cui gli **n** individui omogenei (o repliche) sono assegnati casualmente ai vari livelli del fattore (o trattamenti)
- nel gruppo di soggetti da sottoporre ai diversi trattamenti per confrontarne gli effetti, l'attribuzione di ogni soggetto ad uno specifico trattamento va effettuato per estrazione casuale
- tutto il gruppo deve essere completamente randomizzato
- i vari gruppi possono non avere lo stesso n° di osservazioni o repliche ($\mathbf{n_1}$, $\mathbf{n_2}$, …, $\mathbf{n_p}$ sono in generale diversi tra loro)
- i dati sperimentali vanno riportati secondo la tabella sottostante

	LIVELLI DEL FATTORE SPERIMENTALE O TRATTAMENTI				
	T_1	T_2	T_3	...	T_p
UNITÀ' SPERIMENTALI (o REPLICAZIONI)	X_{11}	X_{12}	X_{13}	...	X_{1p}
	X_{21}	X_{22}	X_{23}	...	X_{2p}
	X_{31}	X_{32}	X_{33}	...	X_{3p}
	...	...	...	...	...
	$X_{n_1 1}$	$X_{n_2 2}$	$X_{n_3 3}$	...	$X_{n_p p}$
	n_1	n_2	n_3	...	n_p
medie dei trattamenti	$\overline{X}_{\bullet 1}$	$\overline{X}_{\bullet 2}$	$\overline{X}_{\bullet 3}$	...	$\overline{X}_{\bullet p}$
media generale	$X_{\bullet\bullet}$				

Secondo questo semplice modello di ANOVA, ogni singola osservazione X_{ij}

$$X_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

è composta da

- MEDIA GENERALE μ
- FATTORE α_j dovuto all'EFFETTO del j-esimo TRATTAMENTO misurato come

$$\alpha_j = \mu_j - \mu \qquad \text{con :}$$

μ_j media del trattamento
μ media generale

- un FATTORE CASUALE ε_{ij} detto RESIDUO o ERRORE SPERIMENTALE
  (errore non è sinonimo di sbaglio, ma di fattore sconosciuto o non valutato o non controllato nell'esperimento)

Gli errori ε_{ij} devono :

- ESSERE TRA LORO INDIPENDENTI: la variazione casuale di ogni replica non deve essere influenzata da quella di un'altra

  (è una indipendenza che può essere ottenuta solamente con una corretta distribuzione casuale delle repliche e quindi di una loro distribuzione secondo la normale)

- DARE VARIANZE OMOGENEE tra loro entro ogni trattamento

- ESSERE DISTRIBUITI NORMALMENTE

La metodologia di ANOVA prevede il calcolo di :

- devianza TOTALE scomposta in :
- devianza TRA TRATTAMENTI (o BETWEEN) con i suoi gdl e la varianza relativa
- devianza ENTRO TRATTAMENTI (o WITHIN o ERRORE) con i suoi gdl e la varianza relativa

Queste quantità abitualmente vengono presentate in uno specchietto :

devianza totale	**gdl = n-1** (n = n° dati)	
devianza tra trattamenti	**gdl = p-1** (p = n° gruppi)	**“varianza tra”**
devianza entro trattamenti	**gdl = n-p**	**“varianza entro”**

Devianza **TOTALE** (o SQ o Somma dei Quadrati degli scarti, o *Sum of Squares*) :

$$SQ_{tot} = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \bar{\bar{X}})^2 = \sum_{j=1}^{p}\sum_{i=1}^{n_j}X_{ij}^2 - \frac{(\sum_{j=1}^{p}\sum_{i=1}^{n_j}X_{ij})^2}{n}$$

- la prima formula, EURISTICA, definisce il significato di devianza totale
- la seconda formula, ABBREVIATA, è matematicamente equivalente alla prima, ma rende più semplici e rapidi i calcoli necessari

Devianza **TRA** TRATTAMENTI :

$$SQ_{tra} = \sum_{j=1}^{p} n_j\ (\bar{X}_j - \bar{\bar{X}})^2 = \sum_{i=1}^{nj}\sum_{j=1}^{p}\left((X_{ij})^2 / n_j \right) - \frac{(\sum_{i=1}^{nj}\sum_{j=1}^{p}X_{ij})^2}{n}$$

Devianza **ENTRO** TRATTAMENTI :

$$SQ_{entro} = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2$$

Dividendo “devianza tra” e “devianza entro” per i rispettivi gdl si ottengono “varianza tra” e “varianza entro” :

- la “varianza tra” misura le differenze esistenti tra un gruppo e l'altro
- la “varianza entro” misura la variabilità esistente attorno alla media aritmetica di ogni gruppo

“Varianza tra” e “varianza entro” dipendono dalla variabilità esistente nei dati ed essendo due misure della stessa variabilità, dovrebbero avere lo stesso valore

Indice dell'uguaglianza tra queste due componenti di varianza, è il

TEST F DI FISHER fondato sul rapporto $\frac{\text{“VARIANZA TRA”}}{\text{”VARIANZA ENTRO”}}$:

- se è vera H_0 dovrebbe risultare F = 1
- se è vera H_1 dovrebbe risultare F > 1
  - con un n° infinito di trattamenti e di repliche, è sufficiente F > 1 per rifiutare l'H_0
  - con un numero ridotto di repliche, può essere F > 1 per effetto delle variazioni casuali

ESEMPIO

In un'analisi della qualità dell'aria in tre diverse zone di città, è stata misurata anche la quantità di ferro (in μg/N mc a 0°C e 1013 mbar)

D.:

Esiste una differenza significativa tra le tre zone A, B e C ?

	FATTORE SPERIMENTALE				
	A	**B**	**C**		
	2,71	1,75	2,22		
	2,06	2,19	2,38		
	2,84	2,09	2,56		
	2,97	2,75	2,60		
	2,55		2,72		
	2,78				
$\sum X_j$	15,91	8,78	12,48	$\sum X$	37,17
$\mathbf{n_i}$	6	4	5	**n**	15
$\overline{X}_{\bullet j}$	2,652	2,195	2,496	$\overline{X}_{\bullet\bullet}$	2,478

La **DEVIANZA TOTALE** è data dalla somma dei quadrati degli scarti di ognuna delle 15 osservazioni rispetto alla media totale

A	B	C
$(2{,}71 - 2{,}478)^2$	$(1{,}75 - 2{,}478)^2$	$(2{,}22 - 2{,}478)^2$
$(2{,}06 - 2{,}478)^2$	$(2{,}19 - 2{,}478)^2$	$(2{,}38 - 2{,}478)^2$
$(2{,}84 - 2{,}478)^2$	$(2{,}09 - 2{,}478)^2$	$(2{,}56 - 2{,}478)^2$
$(2{,}97 - 2{,}478)^2$	$(2{,}75 - 2{,}478)^2$	$(2{,}60 - 2{,}478)^2$
$(2{,}55 - 2{,}478)^2$		$(2{,}72 - 2{,}478)^2$
$(2{,}78 - 2{,}478)^2$		

Quindi, svolgendo i calcoli e sommando i risultati

A	B	C
0,053824	0,529984	0,066564
0,174724	0,082944	0,009604
0,131044	0,150544	0,006724
0,242064	0,073984	0,014884
0,005184		0,058564
0,091204		
0,698040	**0,837456**	**0,156340**

$$\text{Devianza totale} = 0{,}698040 + 0{,}837456 + 0{,}156340 = 1{,}691836$$

Il metodo è lungo e produce stime non precise quando la media sia approssimata; per il calcolo manuale è conveniente utilizzare la formula abbreviata che comporta la somma dei quadrati di ogni replica

	A	B	C	
	7,3441	3,0625	4,9284	
	4,2436	4,7961	5,6644	
	8,0656	4,3681	6,5536	
	8,8209	7,5625	6,7600	
	6,5025		7,3984	
	7,7284			
Σx^2	**42,7051**	**19,7892**	**31,3048**	**93,7991**

$$Devianza_{tot} = Sx^2 - \frac{(Sx)^2}{n} = 93{,}7991 - \frac{(37{,}17)^2}{15} = 1{,}69184$$

“**DEVIANZA TRA**” **:**

- misura la variabilità esistente tra la media aritmetica di ogni gruppo e la media aritmetica generale, ponderata per il **n° di osservazioni** presenti in ciascun gruppo

- è la somma degli scarti di ogni media di gruppo rispetto alla media generale, ponderata per il **n° di repliche**

- ipotizza che, in assenza di variabilità d'errore, i dati sperimentali assumano i valori

A	**B**	**C**
2,652	2,195	2,496
2,652	2,195	2,496
2,652	2,195	2,496
2,652	2,195	2,496
2,652		2,496
2,652		
	media totale **2,478**	

Pertanto con la formula euristica il calcolo diventa :

$$\text{Devianza}_{tra} = \sum_{J=1}^{P} n_j(\overline{X}_j - \overline{\overline{X}})^2$$

$$\text{Devianza}_{tra} = 6\ (2,652 - 2,478)^2 + 4\ (2,195 - 2,478)^2 + 5\ (2,496 - 2,478)^2 =$$
$$= 6\quad 0,030276 + 4\quad 0,080089 + 5\quad 0,000324 =$$
$$= 0,181656 + 0,320356 + 0,00162 = 0,503632$$

La formula abbreviata è più rapida e precisa :

$$\text{Devianza}_{tra} = \sum \frac{(Sx)_j^2}{n_j} - \frac{(Sx)^2}{n}$$

$$\text{Devianza}_{tra} = \frac{(15,91)^2}{6} + \frac{(8,78)^2}{4} + \frac{(12,48)^2}{5} - \frac{(37,17)^2}{15} = 92,610196 - 92,10726 = 0,502936$$

“**DEVIANZA ENTRO**” :

- misura la variazione tra il valore di ciascuna replica e la media aritmetica del proprio gruppo

- è la somma di queste differenze elevate al quadrato per ogni gruppo

A	B	C
$(2,71-2,652)^2$	$(1,75-2,195)^2$	$(2,22-2,496)^2$
$(2,06-2,652)^2$	$(2,19-2,195)^2$	$(2,38-2,496)^2$
$(2,84-2,652)^2$	$(2,09-2,195)^2$	$(2,56-2,496)^2$
$(2,97-2,652)^2$	$(2,75-2,195)^2$	$(2,60-2,496)^2$
$(2,55-2,652)^2$		$(2,72-2,496)^2$
$(2,78-2,652)^2$		

Sviluppando i calcoli e sommando si ottiene

	A	B	C
	0,003364	0,198025	0,076176
	0,350464	0,000025	0,013456
	0,035344	0,011025	0,004096
	0,101124	0,308025	0,010816
	0,010404		0,050176
	0,015376		
Devianza $_{entro}$	0,516076	0,517100	0,154720

- con la formula euristica (somma degli scarti al quadrato) risulta

$$\text{Devianza}_{entro} = 0,516076 + 0,517100 + 0,154720 = 1,187896$$

- può essere ottenuta sottraendo la “devianza tra” dalla devianza totale

$$\text{Devianza}_{entro} = \text{Devianza}_{totale} - \text{Devianza}_{tra} = 1,69184 - 0,502936 = 1,188904$$

Per riassumere i calcoli effettuati, si imposta una tabella che riporta le tre devianze con i rispettivi gdl :

- totale : n° di repliche meno 1
- “tra” : n° di trattamenti meno 1
- “entro” : n° di repliche meno il n° di trattamenti, equivalente ai gdl della devianza totale meno quelli della “devianza tra”

	DEVIANZE	GDL	**VARIANZE**
totale	1,69184	14	
“devianza tra” (between)	0,502936	2	0,251468
“devianza entro” (within)	1,188904	12	0,0990753

Dividendo “varianza tra” per “varianza entro”, si calcola il rapporto $F_{(2,\ 12)}$

$$F_{(2,12)} = \frac{0,251468}{0,0990753} = 2,538$$

- il valore critico di **F** (2 gdl al numeratore; 12 gdl al denominatore) per α=0,05 è 3,89

- il valore calcolato di **F** è inferiore a quello tabulato: la probabilità che H_0 sia vera è **p**>5% e di conseguenza si accetta H_0 (i tre campioni sono stati estratti dalla stessa popolazione)

CONFRONTO TRA ANOVA CON DUE TRATTAMENTI E TEST t PER DUE CAMPIONI INDIPENDENTI

ANOVA può essere applicata anche a due soli trattamenti, in alternativa alla metodologia mediante test **t**

Test **t** e test **F** sono due modi solo apparentemente differenti per fare la stessa cosa: il test **t** è un caso speciale di ANOVA applicata a due gruppi

Tra **t** ed **F** esiste la precisa relazione matematica :

$$F_{(1,\, n)} = t^2_{(n)}$$

ovvero, il valore **F** (**un** gdl al numeratore e **n** gdl al denominatore) è uguale al quadrato di **t** con **n** gdl

ESEMPIO

Due gruppi di 10 uova di *Daphnia magna*, estratte casualmente dallo stesso clone, sono stati allevati in due diverse concentrazioni di cromo esavalente

Dopo un mese sono stati misurati gli individui sopravvissuti: 7 nel gruppo A e 8 nel gruppo B

A	**B**
2,7	2,2
2,8	2,1
2,9	2,2
2,5	2,3
2,6	2,1
2,7	2,2
2,8	2,3
	2,6

D.:

Verificare se le loro dimensioni sono statisticamente diverse

1- Medie:

$$\text{media del gruppo A} = 2,714$$
$$\text{media del gruppo B} = 2,250$$

2- Verifica di omogeneità delle due varianze, mediante il calcolo di devianze, gdl e rapporto **F** tra varianza maggiore e varianza minore

	A	**B**
devianze	0,10857	0,18000
gdl	6	7
varianze	0,018095	0,02571

$$F_{(7,6)} = \frac{0,02571}{0,018095} = 1,42$$

Con 7 gdl della varianza maggiore e 6 della varianza minore, per α=0,05 l'**F** critico è 4,21 > 1,42 (**F** calcolato): dunque le varianze sono omogenee

3 - Varianza "pooled" $$s_p^2 = \frac{0,10825 + 0,18000}{6 + 7} = 0,022173$$

4 - **t** con 13 gdl $$t_{13} = \frac{2,714 - 2,250}{\sqrt{0,022173 \left(\frac{1}{7} + \frac{1}{8}\right)}} = 6,02$$

5 - Controllo della probabilità sulle tabelle dei valori critici: **p** << 0,001

6 - Prospetto di ANOVA

	devianze	**gdl**	**varianze**
totale	1,093333	14	
tra	0,804762	1	0,804761
entro	0,288571	13	0,022198

7 - **F** con 1 e 13 gdl $$F_{(1,13)} = \frac{0,804761}{0,022198} = 36,25$$

8 - Verifica che a tale valore corrisponde alla stessa probabilità, inferiore a 0,001

9 - Verifica che $t^2 = F$ infatti $t^2 = 6,02^2 = 36,24$

ANOVA A DUE CRITERI DI CLASSIFICAZIONE
(BLOCCHI RANDOMIZZATI)

Nella pratica sperimentale, spesso è utile prendere in considerazione più di un fattore di variabilità quando si intende analizzare gli effetti di due o più cause contemporaneamente, oppure ridurre la varianza d'errore isolando gli effetti dovuti ad altre cause note

L'estensione più semplice è rappresentata da due criteri di classificazione, una struttura che si evidenza nel disegno sperimentale a blocchi randomizzati, dove una classificazione riguarda i trattamenti e l'altra i blocchi

	p TRATTAMENTI					
k BLOCCHI	**1**	**2**	**3**	**...**	**p**	**medie**
1	X_{11}	X_{12}	X_{13}	...	X_{1p}	$\overline{X}_{1\bullet}$
2	X_{21}	X_{22}	X_{23}	...	X_{2p}	$\overline{X}_{2\bullet}$
...	...	...	...	...	...	...
k	X_{k1}	X_{k2}	X_{k3}	...	X_{kp}	$\overline{X}_{k\bullet}$
medie	$\overline{X}_{\bullet 1}$	$\overline{X}_{\bullet 2}$	$\overline{X}_{\bullet 3}$	...	$\overline{X}_{\bullet p}$	$\overline{X}_{\bullet\bullet}$

Nel caso più semplice si ha con una sola osservazione $\mathbf{x_{ij}}$ ad ogni intersezione della i-esima riga (blocco) per la j-esima colonna (trattamento)

Il modello lineare additivo, che considera l'effetto del trattamento e del blocco su ogni osservazione, è rappresentato da

$X_{ij} = \mu + \alpha_j + \beta_i + R_{ij}$ con

- μ media generale
- α_j effetto del trattamento stimato come differenza della sua media rispetto alla media generale $\alpha_j = \overline{X}_{\bullet j} - \overline{\overline{X}}$
- β_i effetto del blocco stimato come differenza della sua media risptto alla media generale $\beta_i = \overline{X}_{i\bullet} - \overline{\overline{X}}$
- $\mathbf{R_{ij}}$ quota residua che ingloba, oltre a quelli considerati nei blocchi e nei trattamenti, altri fattori non considerati e la loro interazione insieme con gli effetti di campionamento o di errore ε_{ij}

La metodologia ANOVA a due criteri di classificazione (**p** fattori e **k** blocchi) con una sola osservazione per casella prevede il calcolo delle seguenti quantità:

- devianza totale, con p•k – 1 = n - 1 gdl
- devianza tra trattamenti, con p - 1 gdl, e rispettiva varianza
- devianza tra blocchi, con k - 1 gdl, e rispettiva varianza
- devianza d'errore, con (p-1) (k-1) = (n-1) – (p-1) – (k-1) = p•k–p-k+1 gdl, e rispettiva varianza

Devianze e gdl godono della proprietà additiva :

- Devianza $_{tot}$ = Devianza $_{tra\ tratt}$ + Devianza $_{tra\ blocchi}$ + Devianza $_{errore}$
- gdl_{tot} = $gdl_{tra\ tratt}$ + $gdl_{tra\ blocchi}$ + gdl_{errore}

devianza totale	gdl: n - 1 = p•k - 1	
devianza tra trattamenti	gdl: p - 1	varianza tra trattamenti
devianza tra blocchi	gdl: k - 1	varianza tra blocchi
devianza d'errore	gdl: (p - 1) (k - 1)	varianza d'errore

DEVIANZA TOTALE : variazione totale tra le osservazioni

$$\sum_{j=1}^{p}\sum_{i=1}^{k}(X_{ij}-\overline{\overline{X}})^2=\sum_{j=1}^{p}\sum_{i=1}^{k}X_{ij}^2-\frac{(\sum_{j=1}^{p}\sum_{i=1}^{k}X_{ij})^2}{n}$$

DEVIANZA TRA TRATTAMENTI : variazione tra le medie dei trattamenti

$$\sum_{j=1}^{p}k(\overline{X}_{.j}-\overline{\overline{X}})^2=\sum_{j=1}^{p}(\frac{\sum_{i=1}^{k}X_{.j}^2}{k})-\frac{(\sum_{i=1}^{k}\sum_{j=1}^{p}X_{ij})^2}{n}$$

DEVIANZA TRA BLOCCHI : variazione tra le medie dei blocchi

$$\sum_{i=1}^{k}p(\overline{X}_{i.}-\overline{\overline{X}})^2=\sum_{i=1}^{k}(\frac{\sum_{j=1}^{p}X_{i.}^2}{p})-\frac{(\sum_{i=1}^{k}\sum_{j=1}^{p}X_{ij})^2}{n}$$

DEVIANZA D'ERRORE (RESIDUO) : variazione di ogni osservazione dopo avere tolto l'effetto dovuto alla media generale, alla media del trattamento e alla media del blocco

$$\text{Devianza}_{err} = \text{Devianza}_{tot} - \text{Devianza}_{tra\ tratt} - \text{Devianza}_{tra\ blocchi}$$

Le varianze (tra trattamenti, tra blocchi, errore) si ottengono dividendo le rispettive devianze per i loro gdl

Il test **F** consiste nel confrontare sia la varianza tra trattamenti che quella tra blocchi separatamente con la varianza d'errore

• tra trattamenti : $F_{(p-1)\,,\,(p-1)\cdot(k-1)} = \dfrac{\text{varianza tra tratt}}{\text{varianza d' errore}}$

• tra blocchi : $F_{(k-1)\,,\,(p-1)\cdot(k-1)} = \dfrac{\text{varianza tra blocchi}}{\text{varianza d' errore}}$

ESEMPIO

Confrontare la quantità di Pb in sospensione nell'aria di 5 zone urbane, sapendo che esistono differenze durante la giornata; a distanza di 6 ore (alle 6, 12, 18 e 24) è stata fatta una rilevazione in ogni zona

D.:

C'è differenza tra ore e tra zone considerando i due fattori contemporaneamente ?

	TRATTAMENTI (ZONE)					X_{ij}	
BLOCCHI (ORE)	**1**	**2**	**3**	**4**	**5**	**totali**	**medie**
ore 6	28	25	30	22	26	131	26,2
ore 12	34	32	37	31	30	164	32,8
ore 19	22	21	24	20	19	106	21,2
ore 24	36	31	40	33	29	169	33,8
totali	120	109	131	106	104	570	
medie	30,00	27,25	32,75	26,50	26,00		28,50

DEVIANZA TOTALE con 19 gdl :

$$(28-28,5)^2 + (34-28,5)^2 + (22-28,5)^2 + \ldots + (29-28,5)^2 = 683,0$$

oppure $$(28^2 + 34^2 + 22^2 + 36^2 + 25^2 + 32^2 + \ldots + 29^2) - \frac{570^2}{20} = 683,0$$

La quantità $\frac{(SX)^2}{n} = \frac{570^2}{20}$ compare sia nel calcolo della devianza totale che nelle due "devianze tra" è detta **TERMINE DI CORREZ. GENERALE** (TCG)

DEVIANZA TRA TRATTAMENTI (zone) con 4 gdl :

$$4\ (30,00-28,5)^2 + 4\ (27,25-28,5)^2 + \ldots + 4\ (26,00-28,5)^2 = 128,5$$

oppure $$\frac{120^2}{4} + \frac{109^2}{4} + \frac{131^2}{4} + \frac{106^2}{4} + \frac{104^2}{4} - \frac{570^2}{20} = 128,5$$

DEVIANZA TRA BLOCCHI (ore) con 3 gdl :

$$5\ (26,2-28,5)^2 + 5\ (32,8-28,5)^2 + \ldots + 5\ (33,8-28,5)^2 = 525,8$$

oppure $$\frac{131^2}{5} + \frac{164^2}{5} + \frac{106^2}{5} + \frac{169^2}{5} - \frac{570^2}{20} = 525,8$$

DEVIANZA D'ERRORE e relativi gdl : ottenuti per differenza

683,0 - 128,5 - 525,8 = 28,7 con 19 - 4 - 3 = 12 gdl

	DEVIANZE	**GDL**	**VARIANZE**
totale	**683,0**	**19**	
tra trattamenti (zone)	**128,5**	**4**	**32,125**
tra blocchi (ore)	**525,8**	**3**	**175,266**
errore	**28,7**	**12**	**2,39**

La significatività della differenza tra zone è verificata con $F_{4,12} = \frac{32,125}{2,39} = 13,44$

La significatività della differenza tra ore è verificata con $F_{3,12} = \frac{175,266}{2,39} = 73,33$

Poiché i valori ottenuti superano quelli critici per α=0,05
[$F_{4,12} = 3,26$ $F_{3,12} = 3,49$]
le differenze tra le zone e le differenze tra le ore sono significative

Per comprenderne più esattamente il significato, è utile vedere quanto di ogni osservazione sia imputabile agli effetti congiunti [media generale, media di riga, media di colonna] e quanto ai rimanenti effetti espressi dal residuo

Conoscendo le medie marginali e totale, è possibile calcolare per ogni casella quale sarebbe il valore atteso se agissero solo i tre effetti noti :

media di riga + media di colonna - media generale

	TRATTAMENTI					
BLOCCHI	**1**	**2**	**3**	**4**	**5**	**medie**
I	27,70	24,95	30,45	24,20	23,70	**26,20**
II	34,30	31,55	37,05	30,80	30,30	**32,80**
III	22,70	19,95	25,45	19,20	18,70	**21,20**
IV	35,30	32,55	38,05	31,80	31,30	**33,80**
medie	**30,00**	**27,25**	**32,75**	**26,50**	**26,00**	**28,50**

Utilizzando questi dati per calcolare le devianze, si avrebbero valori identici a quelli dell'esempio per la devianza totale, per quella tra trattamenti e per quella tra blocchi, mentre la devianza d'errore risulterebbe 0, infatti …

… la devianza d'errore calcolata precedentemente è la somma dei quadrati degli scarti tra questi valori stimati e quelli precedenti osservati

In questa tabella, ogni valore è la somma degli effetti $\mu + \alpha_j + \beta_i$ mentre è privo dell'effetto R_{ij} determinato da fattori di interazione e da variazioni casuali

CONFRONTO TRA ANOVA A DUE CRITERI DI CLASSIFICAZIONE E TEST t PER DUE CAMPIONI DIPENDENTI

In una tabella 2 x N, l'ANOVA a due criteri di classificazione fornisce i medesimi risultati del test **t** per due campioni dipendenti, in riferimento al fattore che definisce i due gruppi a confronto; inoltre essa offre il vantaggio di analizzare contemporaneamente anche gli effetti dell'altro fattore

ESEMPIO

Durante una giornata lavorativa, in una stazione di rilevamento sono state misurate le quantità di inquinamento in quattro ore differenti, mentre il successivo giorno festivo, sono state ripetute le misure alle stesse ore per verificare se esiste una differenza significativa tra i due giorni

Per utilizzare il test **t** per due campioni dipendenti, vanno calcolate le differenze tra i due gruppi e si opera esclusivamente su di esse

	I^ giorno	**II^giorno**	**differenze**
ore 6	150	120	30
ore 10	172	151	21
ore 14	193	165	28
ore 18	175	150	25

$\overline{d} = 104 / 4 = 26$

Devianza diff. $(30 - 26)^2 + (21 - 26)^2 \ (28 - 26)^2 + (25 - 26)^2 = 16 + 25 + 4 + 1 = 46$

$s^2 = \frac{46}{3} = 15,33$ $\qquad s = 3,916$ $\qquad t_3 = \frac{26}{\frac{3,916}{\sqrt{4}}} = 13,28$

	I^ giorno	**II^ giorno**	**totali**	**medie**
ore 6	150	120	270	135,0
ore 10	172	151	323	161,5
ore 14	193	165	358	179,0
ore 18	175	150	325	162,5
totali	690	586	1276	
medie	172,5	146,5		159,5

Devianza TOTALE con 7 gdl :

$$(-9,5)^2 + (-39,5)^2 + (12,5)^2 + (-8,5)^2 + (33,5)^2 + (5,5)^2 + (15,5)^2 + (-9,5)^2 =$$
$$= 90,25 + 1560,25 + 156,25 + 72,25 + 1122,25 + 30,25 + 240,25 + 90,25 = 3362$$

Devianza TRA GIORNI con 1 gdl :

$$4(172,5 - 159,5)^2 + 4(146,5 - 159,5)^2 = 4x169 + 4x169 = 1352$$

Devianza TRA ORE con 3 gdl :

$$2\ (135-159,5)^2 + 2\ (161,5-159,5)^2 + 2\ (179-159,5)^2 + 2\ (162,5-159,5)^2 =$$
$$= 2\ 600,25 + 2\ 4 + 2\ 380,25 + 2\ 9 = 1200,50 + 8 + 760,5 + 18 = 1987$$

Devianza d'ERRORE con (7 - 1 - 3) = 3 gdl :

$$3362 - 1352 - 1987 = 23$$

	DEVIANZA	GDL	VARIANZA
totale	3362	7	
tra giorni	1352	1	1352
tra ore	1987	3	662,333
errore	23	3	7,666

Per la differenza tra giorni $F_{1,3} = \frac{1352}{7,666} = 176,36$

E' possibile verificare che $t_3^2 = F_{1,3}$: $(13,28)^2 = 176,36$

Non solo si elimina dalle differenze tra giorni l'effetto delle **differenze tra ore**, ma si verifica anche se queste ultime siano statisticamente significative

$$F_{3,3} = \frac{662,33}{7,666} = 86,40$$

QUADRATI LATINI

- TRE CRITERI DI CLASSIFICAZIONE
- DOPPIO DISEGNO A BLOCCHI

Analizzare contemporaneamente due fattori di variazione a **p** livelli nel disegno a blocchi randomizzati richiede $\mathbf{p^2}$ osservazioni, mentre, con le stesse modalità di programmazione, un esperimento con tre fattori di variazione a **p** livelli ne richiederebbe $\mathbf{p^3}$

I quadrati latini furono applicati per la prima volta in esperimenti di agraria, dove la suddivisione in righe e colonne di un appezzamento di terreno erano visualizzate in strisce di terreno tra loro perpendicolari; da qui il nome, per la somiglianza del frazionamento dell'area in una figura tipica dell'accampamento romano

Il disegno a quadrati latini permette di analizzare contemporaneamente tre fattori a **p** livelli con p^2 osservazioni solamente

Al vantaggio di un risparmio di materiale si contrappone lo svantaggio di una notevole rigidità, infatti tutti i tre criteri (“trattamenti”, “blocchi”, “fattore principale”) devono avere lo stesso n° di livelli

In un esperimento con 3 criteri, due sono rappresentati da righe e da colonne (i fattori secondari), mentre il terzo (il fattore principale) è distribuito entro lo schema della tabella in modo casuale ma bilanciato, e compare una volta sola sia in ogni riga e in ogni colonna

Indicando con **A, B, C, D** i 4 livelli di un fattore principale, la rappresentazione grafica bidimensionale dell'esperimento può essere :

	COLONNE			
RIGHE	**1**	**2**	**3**	**4**
1	**D**	**B**	**C**	**A**
2	**C**	**D**	**A**	**B**
3	**B**	**A**	**D**	**C**
4	**A**	**C**	**B**	**D**

Così come in un disegno a due criteri di classificazione, la randomizzazione è attuata assegnando a caso i livelli dei trattamenti entro ciascun blocco, in un quadrato latino, la randomizzazione è attuata permutando i diversi livelli del fattore principale nello schema ordinato di righe e colonne

Sono state costruite tabelle di distribuzione casuale, da utilizzare nel caso di più esperimenti a quadrati latini con schemi differenti

Il limite più pesante a questo modo di programmare l'esperimento è dato dalla sua rigidità: ad esempio, volendo analizzare un fattore a 5 livelli, occorrerà un n° uguale di livelli anche negli altri due criteri organizzati per righe e colonne

Il modello additivo lineare di ANOVA in un disegno sperimentale a quadrato latino richiede che la generica osservazione X_{ijk} appartenente al i-esimo "blocco", al j-esimo "trattamento" e al k-esimo fattore, sia data da

$X_{ijk} = \mu + \alpha_j + \beta_i + \gamma_k + \varepsilon_{ijk}$ con:

- μ media generale
- α_j effetto medio del "trattamento" i-esimo
- β_i effetto medio del "blocco" j-esimo
- γ_k effetto medio del fattore k-esimo
- ε_{ijk} variabilità residua

Il calcolo delle devianze è semplice: la devianza totale, la devianza tra righe e quella tra colonne sono calcolate con la stessa metodologia utilizzata nel disegno a blocchi randomizzati; la devianza tra trattamenti viene calcolata rispetto alla somma e alla media dei vari trattamenti

ESEMPIO

Confrontare la produttività di 5 (A, B, C, D, E) varietà di sementi in rapporto al tipo di concime (1,2,3,4,5) e ad un diverso trattamento del terreno (I, II, III, IV, V)

Si è diviso l'appezzamento in 5 strisce equivalenti e in ognuna è stata condotta un'aratura di profondità differente; perpendicolarmente a queste strisce sono state tracciate altre 5 strisce concimate in modo diverso; nei 25 quadrati sono state seminate le 5 varietà di sementi secondo lo schema

	TRATTAMENTO DEL TERRENO						
CONCIME	**I**	**II**	**III**	**IV**	**V**	**totali**	**medie**
1	**A 42**	**C 47**	**B 55**	**D 51**	**E 44**	239	47,8
2	**E 45**	**B 54**	**C 52**	**A 44**	**D 50**	245	49,0
3	**C 41**	**A 46**	**D 57**	**E 47**	**B 48**	239	47,8
4	**B 56**	**D 52**	**E 49**	**C 50**	**A 43**	250	50,0
5	**D 47**	**E 49**	**A 45**	**B 54**	**C 46**	241	48,2
totali	231	248	258	246	231	1214	
medie	46,2	49,6	51,6	49,2	46,2		48,56

sementi	**A**	**B**	**C**	**D**	**E**
totali	220	267	236	257	234
medie	44,0	53,4	47,2	51,4	46,8

I risultati di ANOVA sono

	DEVIANZE	**GDL**	**VARIANZE**
totale	480,16	24	
tra sementi	286,16	**4**	71,54
tra concimi	109,36	**4**	27,34
tra arature	17,76	**4**	4,44
errore	66,88	12	5,57

Si possono calcolare tre F, tutti con 4 e 12 gdl :

• tra sementi: $F_{4,\,12} = \dfrac{71,54}{5,57} = 12,84$

• tra concimi: $F_{4,\,12} = \dfrac{27,34}{5,57} = 4,91$

• tra arature: questa varianza è minore della varianza d'errore e pertanto è inutile calcolare il rapporto F per verificare se gli sia significativamente superiore

Per 4 e 12 gdl la tabella dei valori critici per α=0,05 fornisce il valore 3,26, per cui risultano significative :

- la differenza tra sementi (F = 12,84)
- la differenza tra concimi (F = 4,91)

ma non quella tra i diversi tipi di aratura (F < 1)

ESEMPIO

Tra le numerose applicazioni, con i quadrati latini si possono analizzare gli effetti di diversi farmaci (“fattore principale”) da somministrare ad alcune persone (“blocchi”) in giorni diversi (“trattamenti”), e accertare se l'effetto di un farmaco dipenda anche dal tempo in cui è somministrato

Si sperimentano gli effetti di 4 diversi farmaci (A, B, C, D) somministrati in 4 giorni diversi a 4 diverse persone :

	GIORNI			
PERSONE	1	2	3	4
I	**A 48**	**C 35**	**D 40**	**B 51**
II	**D 37**	**B 50**	**C 33**	**A 45**
III	**B 42**	**D 64**	**A 53**	**C 39**
IV	**C 31**	**A 40**	**B 42**	**D 37**

I risultati di ANOVA sono

	DEVIANZE	GDL	VARIANZE
totale	1098	15	
tra farmaci	389	***3***	129,7
tra giorni	125	***3***	41,7
tra persone	303	***3***	101,0
errore	281	6	46,8

Il disegno sperimentale a quadrati latini impone che le sue dimensioni non possano essere nè troppo piccole, né troppo grandi :

• il limite **minimo** è imposto dai gdl della varianza d'errore [= n^2 - (n-1) 3 - 1]:

- un quadrato latino **2x2** avrebbe in totale 3 gdl: 1 per il fattore principale, 1 per le colonne e 1 per le righe, senza più gdl per la varianza d'errore
- un quadrato latino **3x3**, avrebbe la varianza d'errore con solo 2 gdl, troppo pochi per rendere significative differenze tra medie non molto grandi

• il limite massimo è determinato dalla complessità dell'esperimento e viene abitualmente fissato per un quadrato **12x12**

La replica di un esperimento a quadrati latini determina i quadrati greco-latini, che sono la sovrapposizione di due quadrati latini; con più repliche si parla di QUADRATI CON PIÙ ALFABETI

A parte la crescente complessità dell'esperimento, un punto importante da ricordare è il n° di gdl della varianza d'errore che diminuisce proporzionalmente

DATI MANCANTI

Nel disegno a blocchi randomizzati e in quello a quadrati latini :

• la mancanza di una osservazione pone un problema di elaborazione dei dati

• si richiede un numero prefissato di osservazioni, a differenza di quanto avviene nel disegno sperimentale ad un criterio di classificazione, dove la validità di ANOVA non dipende dall'eguaglianza del n° di repliche

I dati possono mancare per :

• selezioni contro determinati valori (ad es. quelli molto grandi o molto piccoli) che uno strumento può non registrare perché troppo differenti dalla scala su cui è tarato
--> il campione raccolto è viziato in modo irrimediabile

• cause accidentali
--> è possibile rimpiazzare i dati mancanti

In una tabella a due fattori o a due entrate senza replicazioni (come nel disegno a blocchi randomizzati) il dato da stimare dipende dagli effetti di riga e di colonna calcolati dagli altri dati

Si stima un valore X'_{ij} che dipende dalla media generale μ, dall'effetto riga α_i e dall'effetto colonna β_j, che sarà privo della variazione casuale ε_{ij}

$$X'_{ij} = \mu + \alpha_j + \beta_i$$

In un disegno con **r** righe e **c** colonne, indicando con R_i il totale di riga, con C_j il totale di colonna e con T il totale generale, calcolati senza il dato mancante, X_{ij} può essere sostituito con X'_{ij}

$$X'_{ij} = \frac{r \ R_i + c \ C_j - T}{(r-1) \ (c-1)}$$

Se manca l'osservazione del trattamento ZONA 3 e blocco ORA II

TRATTAMENTI

BLOCCHI	ZONA 1	ZONA 2	ZONA 3	ZONA 4	ZONA 5	TOTALI
ORA I	28	25	30	22	26	131
ORA II	34	32	**?**	31	30	127
ORA III	22	21	24	20	19	106
ORA IV	36	31	40	33	29	169
TOTALI	120	109	94	106	104	533

il valore con cui sostituire tale osservazione è $X'_{ij} = \frac{4\ 127 + 5\ 94 - 533}{(4-1)\ (5-1)} = 37,08$

La sostituzione del dato mancante permette di eseguire i calcoli di ANOVA nel modo abituale: cambiano i gdl totale e i gdl della varianza d'errore, che saranno diminuiti di 1 (restano immutati quelli tra trattamenti e tra blocchi)

L'operazione di sostituzione ha il solo scopo di permettere di effettuare i calcoli di ANOVA in modo corretto, ma non aggiunge alcuna informazione che non fosse già contenuta nei dati osservati

Se manca più di un dato:
• si sostituiscono i dati mancanti meno uno con dati inventati, anche se logici
• il dato non sostituito viene stimato con la modalità su illustrata
• si stima un altro dato in sostituzione di un dato inventato
• si itera il procedimento per tutti i dati originariamente mancanti, finchè i valori stimati restano stabili

Nel caso di un disegno a quadrati latini **NxN**, il dato mancante nella i-esima riga, j-esima colonna e k-esimo fattore può essere sostituito da

$$X'_{ijk} = \frac{n\ (R_i + C_j + T_k) - 2G}{(n-1)\ (n-2)}$$ con :

n : dimensione del quadrato latino
R_i C_j T_k : totali riga, colonna, trattamento cui appartiene il dato mancante
G : totale generale

Anche qui la varianza tra trattamenti e l'errore standard della differenza tra due trattamenti andrebbero ridotti

EFFICIENZA RELATIVA DI UN TEST

Il disegno sperimentale a blocchi randomizzati o a quadrati latini è indubbiamente molto vantaggioso quando si intende analizzare contemporaneamente 2 o 3 fattori di variazione:

• in un quadrato latino 5x5, i gdl di ognuno dei 3 test F per verificare la significatività delle differenze tra le medie di ogni fattore sono 4 e 12

• se si facesse un esperimento ad un criterio di classificazione per esaminare un solo fattore, per avere un test con la medesima potenza servirebbero in totale **17 dati**: 16 gdl per la devianza totale, suddivisi in 4 gdl per la devianza tra trattamenti e 12 per quella d'errore

• se si analizzassero separatamente i 3 fattori con 3 esperimenti diversi, sarebbero richiesti gli stessi gdl per ogni esperimento: un totale di **51 dati**, più del doppio di quelli utilizzati nel quadrato latino

Lo scopo di allestire un disegno sperimentale più complesso è quello di rendere più significativo il test F, mediante il controllo delle maggiori fonti di variazione e la riduzione della varianza d'errore

In una ANOVA a blocchi randomizzati o a quadrati latini, rispetto all'analisi ad un criterio di classificazione si ha sempre un abbassamento sia della devianza d'errore sia dei gdl

•• non sempre la varianza d'errore, determinata dal loro rapporto, diminuisce: se la devianza d'errore diminuisce meno dei gdl, la varianza aumenta

•• al vantaggio derivante dall'eventuale abbassamento della varianza d'errore si associa sempre lo svantaggio dovuto alla perdita di gdl, per cui il valore di F richiesto per dimostrare la significatività del test aumenta

•• quando i gdl sono pochi, lo svantaggio derivante dal calo dei gdl della varianza d'errore può essere grave

ESEMPIO

Supponiamo che un'ANOVA a blocchi randomizzati abbia fornito il risultato:

	DEVIANZE	GDL	VARIANZE
totale	3362	7	
tra giorni	1352	1	1352
tra ore	1987	3	662,333
errore	23	3	7,666

Con i medesimi dati è sempre possibile condurre un'ANOVA completamente randomizzata, calcolando solo la devianza tra giorni; la devianza tra ore ed i suoi gdl sono cumulati con quelli d'errore, con il seguente risultato:

	DEVIANZE	GDL	VARIANZE
totale	3362	7	
tra giorni	1352	1	1352
errore	2010	6	335

Nel primo caso, per verificare la differenza tra giorni si ottiene un test $F_{1,3}$

$$F_{1,3} = \frac{1352}{7,666} = 176,36$$

risulta altamente significativo, essendo il valore critico, per α=0,05, uguale a 10,13

Nel secondo caso la differenza tra giorni è verificata con $F_{1,6}$

$$F_{1,6} = \frac{1352}{335} = 4,035$$

non risulta significativo, essendo il valore critico, per α=0,05 e per gli stessi gdl, uguale a 5,99

La convenienza ad utilizzare lo schema a blocchi rispetto a quello completamente randomizzato, misurata in termini di **EFFICIENZA RELATIVA**, è data da

$$E.R. = \frac{(n_1+3)\ (n_2+1)\ s_{e1}^2}{(n_1+1)\ (n_2+3)\ s_{e2}^2} \qquad \text{dove :}$$

$\mathbf{n_1}$ n° di gdl della varianza d'errore nell'analisi ad un criterio di classificazione o a disegno completamente randomizzato

$\mathbf{n_2}$ n° di gdl della varianza d'errore nell'analisi a 2 criteri di classificazione o a blocchi randomizzati

$\mathbf{s_{e1}^2}$ varianza d'errore nell'analisi ad un criterio di classificazione

$\mathbf{s_{e2}^2}$ varianza d'errore nell'analisi a due criteri di classificazione

Con i dati su riportati, l'efficienza relativa del disegno a blocchi randomizzati rispetto a quello completamente randomizzato è

$$E.R. = \frac{(6+3)\ (3+1)\ 335}{(6+1)\ (3+3)\ 7,666} = \frac{12060}{391,97} = 30,77$$

In questo caso, l'esperimento a due criteri di classificazione risulta circa 31 volte più efficace di quello ad un solo criterio (ovvero: per ottenere la stessa potenza del test a blocchi randomizzati, con il test a disegno completamente randomizzato occorrono quasi 31 volte più dati: non 8 osservazioni, bensì 248)

E' possibile calcolare l'efficienza di un quadrato latino sia rispetto al disegno a blocchi randomizzati che a quello completamente randomizzato

Supponiamo che un' ANOVA a quadrati latini abbia fornito i risultati :

	DEVIANZE	GDL	VARIANZE
totale	480,16	24	
tra sementi	286,16	4	71,54
tra concimi	109,36	4	27,34
tra arature	17,76	4	4,44
errore	66,88	12	5,57

Con gli stessi dati sarebbe stato possibile ignorare la differenza tra arature; i risultati di ANOVA sarebbero diventati :

	DEVIANZE	GDL	VARIANZE
totale	480,16	24	
tra sementi	286,16	4	71,54
tra concimi	109,36	4	27,34
errore	84,64	16	5,29

Sarebbe stato possibile stimare solamente due F:

tra sementi: $F_{4,16} = \frac{71,54}{5,29} = 13,52$

tra concimi: $F_{4,16} = \frac{27,34}{5,29} = 5,17$

che risultano ancor più significative sia per la riduzione della varianza d'errore sia per aumento dei gdl che abbassa il livello del valore critico: per α=0,05 da 4 e 12 gdl a 4 e 16 gdl il valore di F diminuisce da 3,26 a 3,01 e per α=0,01 da 5,41 a 4,77

Dopo avere verificato con i quadrati latini che non esiste differenza significativa tra arature, è più opportuno, per dimostrare la significatività delle differenze tra sementi e tra concimi, presentare i risultati con lo schema a blocchi randomizzati, anche se l'esperimento è stato condotto con lo schema più complesso dei Q.L.

Il fattore principale può essere o il tipo di semente (A) o il concime (B)

CASO A:

	DEVIANZA	GDL	VARIANZA
totale	480,16	24	
tra sementi	286,16	4	71,54
errore	194	20	9,7

$$F_{4,20} = \frac{71,54}{9,7} = 7,38$$

CASO B:

	DEVIANZA	GDL	VARIANZA
totale	480,16	24	
tra concimi	109,36	4	27,34
errore	370,80	20	18,54

$$F_{4,20} = \frac{27,34}{18,54} = 1,47$$

Le differenze tra sementi rimangono significative, ma con un valore di F assai minore; le differenze tra i concimi non risultano più significative, poiché nella varianza d'errore è stata cumulata anche la grande variabilità tra specie di sementi

L'efficienza relativa di uno specifico esperimento a quadrati latini rispetto al corrispondente disegno a blocchi randomizzati può essere stimata con una formula analoga a quella presentata precedentemente

$$E.R. = \frac{(n_2+3)\ (n_3+1)\ s_{e2}^2}{(n_2+1)\ (n_3+3)\ s_{e3}^2}$$

dove :

s_{e2}^2 e s_{e3}^2 varianze d'errore rispettive

$\mathbf{n_3}$ gdl della varianza d'errore a 3 criteri di classific.

$\mathbf{n_2}$ gdl della varianza d'errore a 2 criteri di classific.