ANOVA A DUE FATTORI

TWO-WAY ANOVA

L'analisi della varianza a una via è un metodo statistico per testare l'ipotesi nulla (H0) in base alla quale le medie di tre o più popolazioni sarebbero uguali, contro l'ipotesi alternativa (H1) secondo cui almeno una media sarebbe diversa.

Servendoci della notazione formale per le ipotesi statistiche, per k medie scriviamo:

$$H0\text{: } media_1 = media_2 = \cdots = media_k$$

H1: non tutte le medie sono uguali

Può essere necessario prendere in considerazione due fattori di variabilità. In questo caso si ricorre ad una analisi della varianza a due vie (ANOVA a due vie, two-way ANOVA).

L'ANOVA a due vie è un test di ipotesi in cui la classificazione dei dati si basa su due fattori.

Ad esempio, le due basi di classificazione per valutare l'effetto di un farmaco (variabile dipendente) sono in primo luogo sulla base della dose usata (variabile indipendente 1) e in secondo sulla base delle diverse patologie (variabile indipendente 2) per cui può essere somministrato.
È una tecnica statistica utilizzata per confrontare diversi livelli (condizioni) delle due variabili indipendenti che coinvolgono più osservazioni.

Dose Farmaco		
Dose 1	Dose 2	Dose 3
a1	b1	c1
a2	b2	c2
a3	b3	-
-	b4	-

ANOVA A UNA VIA

		Dose Farmaco		
		Dose 1	Dose 2	Dose 3
Patologia	Patologia A	X11	X12	X13
	Patologia B	X21	X22	X23
	Patologia C	X31	X32	X33

ANOVA A DUE VIE

		Dose Farmaco		
		Dose 1	Dose 2	Dose 3
Patologia	Patologia A	X11	X12	X13
	Patologia B	X21	X22	X23
	Patologia C	X31	X32	X33

Assumiamo che testiamo un numero uguale di soggetti per ogni dose e per ogni patologia. Questo viene definito balanced study

Per quanto riguarda i valori in tabella ➔ ad esempio, il valore riportato in posizione 1,1 è inteso come media dei valori acquisiti. Ossia se ho n soggetti con patologia A e sono stati testati con dose 1, il valore riportato in tabella sarà:

$$X_{11} = \frac{1}{n}\sum_{i=1}^{n} misura_{i\ (patologia\ A, dose\ 1)} = \frac{1}{n}\sum_{i=1}^{n} misura_{i\ (j,k)}$$

Nell'analisi della varianza a una via abbiamo considerato:

$\text{SSE} = \sum_{i=1}^{k}(n_i - 1)s_i^2$

$\text{SSB} = \sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2$

Consideriamo un gruppo specifico dei k a disposizione, ad esempio il gruppo h, e indichiamo con $x_{i,h}$ la misura i-esima del gruppo h.

Ricordiamo che abbiamo indichiamo con $\bar{x}$ la media totale (tutti i gruppi) e con $\bar{x}_h$ la media aritmetica del gruppo h-esimo. Allora:

$$\left(x_{i,h} - \bar{x}\right) = \left(x_{i,h} - \bar{x} + \bar{x}_h - \bar{x}_h\right) = (\bar{x}_h - \bar{x}) + \left(x_{i,h} - \bar{x}_h\right)$$

Radice della devianza

Distanza di ogni misura dalla media totale

effetto

distanza delle medie di ciascun gruppo dalla media totale

errore

distanza di ogni misura dalla media di ogni gruppo

Nell'analisi della varianza a due vie consideriamo:

$misura_{i\,(j,k)}$

$\bar{x}_{(j,k)}$	media cella (j,k)
$\bar{x}_{(j)}$	media riga j
$\bar{x}_{(k)}$	media colonna k
$\bar{x}$	media totale

effetto principale 1 — interazione

$$\left(x_{i(j,k)} - \bar{x}\right) = \left(\bar{x}_{(j)} - \bar{x}\right) + \left(\bar{x}_{(k)} - \bar{x}\right) + \left(\bar{x}_{(j,k)} - \bar{x}_{(j)} - \bar{x}_{(k)} + \bar{x}\right) + \ldots$$

effetto principale 2

Radice della devianza

Distanza di ogni misura dalla media totale

$$+\left(x_{i(j,k)} - \bar{x}_{(j,k)}\right)$$

errore

distanza di ogni misura dalla media di ogni gruppo

Passando ora alla devianza:

$$SS_{TOT} = SS_{righe} + SS_{colonne} + SS_{interazione} + SSE = SS_A + SS_B + SS_{AB} + SSE$$

e ipotizzando **a righe** e **b colonne** **e n valori totali per ogni cella/posizione:**

Sorgente della variabilità	Devianza	Gradi di libertà	Varianza
Tra righe	$SS_A = b \cdot n \cdot \sum(\bar{x}_{(j)} - \bar{x})^2$	a-1	S_A^2
Tra colonne	$SS_B = a \cdot n \cdot \sum(\bar{x}_{(k)} - \bar{x})^2$	b-1	S_B^2
Interazione	$SS_{AB} = n \cdot \sum\sum(\bar{x}_{(j,k)} - \bar{x}_{(j)} - \bar{x}_{(k)} + \bar{x})^2$	(a-1)(b-1)	S_{AB}^2
Errore	$SSE = \sum\sum\sum(x_{i(j,k)} - \bar{x}_{(j,k)})^2$	ab(n-1)	S_E^2
Totale	$SS_{TOT} = \sum\sum\sum(x_{i(j,k)} - \bar{x})^2$	abn-1	

Le ipotesi nulle sono:

H_{0A}: nessuna differenza tra le medie per il fattore A, ossia effetti fattore A nulli (effetti riga)
H_{0B}: nessuna differenza tra le medie per il fattore B, ossia effetti fattore B nulli (effetti colonna)
H_{0AB}: nessuna differenza tra le medie per l'interazione tra i fattori A e B

Sorgente della variabilità	Devianza	Gradi di libertà	Varianza	Statistica test
Tra righe	$SS_A = b \cdot n \cdot \sum(\bar{x}_{(j)} - \bar{x})^2$	a-1	S_A^2	$F_A = \frac{S_A^2}{S_E^2}$
Tra colonne	$SS_B = a \cdot n \cdot \sum(\bar{x}_{(k)} - \bar{x})^2$	b-1	S_B^2	$F_B = \frac{S_B^2}{S_E^2}$
Interazione	$SS_{AB} = n \cdot \sum\sum(\bar{x}_{(j,k)} - \bar{x}_{(j)} - \bar{x}_{(k)} + \bar{x})^2$	(a-1)(b-1)	S_{AB}^2	$F_{AB} = \frac{S_{AB}^2}{S_E^2}$
Errore	$SSE = \sum\sum\sum(x_{i(j,k)} - \bar{x}_{(j,k)})^2$	ab(n-1)	S_E^2	
Totale	$SS_{TOT} = \sum\sum\sum(x_{i(j,k)} - \bar{x})^2$	abn-1		

F_A è una statistica di Fisher con gradi di libertà (a-1, ab(n-1))
F_B è una statistica di Fisher con gradi di libertà (b-1, ab(n-1))

ESEMPIO:

Si vuole verificare se insetti adulti della stessa specie che vivono in 4 località differenti (A, B, C, D) hanno differenze significative nelle loro dimensioni, considerando pure che dalla primavera all'autunno continuano ad aumentare.

	LOCALITA'			
	A	**B**	**C**	**D**
Primavera	**45**	**63**	**70**	**48**
	50	**57**	**65**	**52**
Estate	**57**	**77**	**74**	**60**
	65	**69**	**80**	**56**
Autunno	**70**	**82**	**88**	**70**
	79	**75**	**82**	**77**

In altri termini, ci si chiede se la crescita è diversa nelle 4 località in rapporto alle stagioni (per semplificare al massimo i calcoli, sono state riportate solamente 2 misure per località e stagione).

LOCALITA'

	A	B	C	D	Totali	Medie
Primavera	**45**	**63**	**70**	**48**		
	50	**57**	**65**	**52**	450	56,250
Totali	(95)	(120)	(135)	(100)		
Medie	(47,5)	(60,0)	(67,5)	(50,0)		
Estate	**57**	**77**	**74**	**60**		
	65	**69**	**80**	**56**	538	67,250
Totali	(122)	(146)	(154)	(116)		
Medie	(61,0)	(73,0)	(77,0)	(58,0)		
Autunno	**70**	**82**	**88**	**70**		
	79	**75**	**82**	**77**	623	77,875
Totali	(149)	(157)	(170)	(147)		
Medie	(74,5)	(78,5)	(85,0)	(73,5)		
Totali gen.	366	423	459	363	1611	
Medie gen.	61,00	70,50	76,50	60,50		67,125

- La devianza totale , con 23 gdl (per tutti e 24 i dati) è ottenuta con

$$(45-67{,}125)^2 + (50-67{,}125)^2 + \ldots + (70-67{,}125)^2 + (77-67{,}125)^2$$

e corrisponde a :

$$45^2 + 50^2 + \ldots + 70^2 + 77^2 - \frac{1611^2}{24} = \mathbf{3.300}$$

- La devianza tra le medie delle caselle o dei vari livelli dei due fattori, con 11 gdl (le medie delle caselle all'incrocio località per stagione sono 12) è data da

$$2(47{,}5 - 67{,}125)^2 + 2(60{,}0 - 67{,}125)^2 + \ldots + 2(73{,}5 - 67{,}125)^2$$

e corrisponde a (con i totali delle 12 caselle)

$$\frac{95^2}{2} + \frac{120^2}{2} \ldots + \frac{147^2}{2} - \frac{1611^2}{24} = \mathbf{3.052}$$

- La devianza tra trattamenti o del fattore A, con 3 gdl è calcolata (per le 4 medie di colonna) da

$$\text{Devianza}_A = 6\cdot(61{,}00 - 67{,}125)^2 + 6\cdot(70{,}50 - 67{,}125)^2 + \ldots + 6\cdot(60{,}50 - 67{,}125)^2$$

ed è uguale a (con i totali dei 4 trattamenti):

$$\text{Devianza}_A = \frac{366^2}{6} + \frac{423^2}{6} + \frac{459^2}{6} + \frac{363^2}{6} - \frac{1611^2}{24} = \mathbf{1.084}$$

- La devianza tra blocchi o del fattore B, con 2 gdl è ottenuta mediante

$$\text{Devianza}_B = 8\cdot(56{,}250 - 67{,}125)^2 + 8\cdot(67{,}250 - 67{,}125)^2 + 8\cdot(77{,}875 - 67{,}125)^2$$

e corrisponde a (con i totali dei 3 blocchi)

$$\text{Devianza}_B = \frac{450^2}{8} + \frac{538^2}{8} + \frac{623^2}{8} - \frac{1611^2}{24} = \mathbf{1.871}$$

- La devianza d'interazione AB viene stimata per differenza:

$$\text{Devianza}_{AB} = \text{Devianza}_{\text{tra medie}} - \text{Devianza}_A - \text{Devianza}_B = 3.052 - 1.084 - 1.871 = \mathbf{97}$$

e nello stesso modo vengono calcolati i rispettivi gdl

$$\text{gdl}_{AB} = \text{gdl}_{\text{tra medie}} - \text{gdl}_A - \text{gdl}_B = 11 - 3 - 2 = \mathbf{6}$$

- La devianza d'errore o residuo nei calcoli manuali viene quasi sempre stimata per differenza

$$\text{Devianza}_{\text{errore}} = \text{Devianza}_{\text{totale}} - \text{Devianza}_{\text{tra medie}} = 3.300 - 3.052 = \mathbf{248}$$

e nello stesso modo sono calcolati i suoi gdl

$$\text{gdl}_{\text{errore}} = \text{gdl}_{\text{totale}} - \text{gdl}_{\text{tra medie}} = 23 - 11 = \mathbf{12}$$

	DEVIANZA	GDL	VARIANZA
Totale	3.300	23	---
Tra medie	3.052	11	---
Tra tratt. (A)	1.084	3	361,33
Tra blocchi (B)	1.870	2	935,50
Interazione (AB)	97	6	16,16
Errore	248	12	20,66

Per le differenze tra trattamenti o effetto del fattore A si calcola un test F con gdl 3 e 12

$$F_{3,12} = \frac{361,33}{20,66} = 17,49$$

il cui valore critico alla probabilità $\alpha = 0.05$ è uguale a 3,49.

Per le differenze tra blocchi o effetto del fattore B si calcola un test F con gdl 2 e 12

$$F_{2,12} = \frac{935,5}{20,66} = 45,28$$

il valore critico del quale alla probabilità $\alpha = 0.05$ è 3,89.

Per l'effetto dell'interazione AB si calcola un test F con gdl 6 e 12

$$F_{6,12} = \frac{16,16}{20,66} = 0,78$$

che fornisce un rapporto inferiore ad 1 e

quindi non è significativo.

Con i risultati dell'esempio, si possono trarre le conclusioni relative ai tre test F:

1 - nelle 4 località, le dimensioni medie degli insetti sono significativamente differenti;

2 - tra primavera, estate ed autunno le dimensioni medie degli insetti variano significativamente;

3 - non esiste interazione: nelle 4 località le dimensioni medie degli insetti variano con intensità simile durante le stagioni.

Una chiave per i test parametrici sulle medie