Anova ad una via

L'analisi della varianza viene utilizzata per verificare l'uguaglianza delle medie fra diverse gruppi. Questo viene fatto scomponendo la variabilità dei dati osservati tenendo conto della suddivisione in gruppi. In particolare, si può pensare di stimare la variabilità delle osservazioni, sia a partire da un'analisi dei dati all'interno dei diversi gruppi e sia a partire da un'analisi della variabilità tra i gruppi. Nell'ipotesi nulla di appartenenza dei diversi gruppi alla stessa popolazione, questa stima deve essere equivalente.

In un modello completamente randomizzato, quale quello che analizzeremo, le varie unità (ad esempio gli individui) si differenziano semplicemente per un fattore sperimentale, ad esempio l'assegnazione ad un trattamento, in modo del tutto causale. I diversi gruppi, ad esempio associati ai diversi trattamenti, possono avere anche numerosità differenti.

La singola misura può essere schematizzata in questo modo

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

dove α_i è un fattore costante, legato ad esempio al trattamento e ε_{ij} è la variabile che rappresenta la variabilità casuale del dato, non controllabile e non legata al trattamento. L'ipotesi è che la variabilità sia della stessa entità nei diversi gruppi, che le ε_{ij} siano tra loro incorrelate. Faremo inoltre l'ipotesi che ε_{ij} sia una variabile gaussiana a valore medio nullo e varianza σ^2.

La media di ogni gruppo è pari a

$$\mu_i = \mu + \alpha_i$$

con i=1,2, .., k, dove k è il numero di gruppi e μ è la media generale. In questo modo $\sum_i \alpha_i = 0$.

Quindi l'ipotesi nulla H_0 si riferisce allo scenario nel quale la media dei singoli gruppi sia pari alla media generale e quindi che α_i=0 per ogni i.

Nel modello che prenderemo in considerazione i fattori legati a i gruppi sono delle costanti caratteristiche di ogni gruppo (effetti fissi); sono possibili modelli nei quali i fattori sono descritti come variabili aleatorie (effetti causali o random). Nella seguente tabella si mostrano le misure, classificate in funzione del trattamento. In questo caso il fattore di classificazione è uno, e per questo si parla di anova ad una via. Le diverse misure in ogni gruppo prendono il nome di repliche.

È importante inoltre ricordare che dovranno essere verificate le ipotesi del test, tra le quali la gaussianità delle osservazioni e l'uguaglianza delle varianze nei vari gruppi.

		Osservazioni o repliche			
Gruppi o Trattamenti	T1	X_{11}	X_{12}	…	X_{1n1}
	T2	X_{21}	X_{22}	…	X_{2n2}
	T3				X_{3n3}
	…			..	…
	Tk	X_{k1}	X_{k2}	…	X_{knk}

La devianza totale (o somma dei quadrati SS totale) si può stimare come

$$SS_{Tot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{\bar{x}}\right)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}^2 - \frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}\right)^2}{n}$$

dove $n = \sum_i n_i$ è il numero totale di misure, e n_i sono il numero di repliche nel gruppo *i*-esimo.

La media di ciascun gruppo è $m_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$

La media generale di tutte le osservazioni è $\bar{\bar{x}} = \frac{1}{n} \sum_{i=1}^{k} n_i m_i = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$.

La varianza in ogni gruppo è $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} \left(x_{ij} - m_i\right)^2$.

La devianza entro i gruppi (o somma dei quadrati SS entro i gruppi)

$$SS_{entro} = \sum_{i=1}^{k} (n_i - 1) s_i^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2$$

dove $\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$.

Mentre la devianza tra gruppi è scrivibile come

$$SS_{tra} = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{\bar{x}})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\left(x_{ij}\right)^2 / n_i\right) - \frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}\right)^2}{n}$$

Le varianze possono essere ora ottenute dividendo le devianze per i rispettivi gradi di libertà.
In particolare, la devianza entro i gruppi possiede *n-k* gradi di libertà, perché è necessario calcolare i *k* valori medi dei vari gruppi. Quindi la varianza entro i gruppi può essere scritta come

$$S_{entro}^2 = \frac{SS_{entro}}{n - k}$$

La devianza tra gruppi possiede *k*-1 gradi di libertà, visto che la media totale può essere vista come la media pesata, in funzione della numerosità, delle medie dei singoli gruppi. Quindi la varianza tra gruppi può essere calcolata come

$$S_{tra}^2 = \frac{SS_{tra}}{k - 1}$$

Nell'ipotesi nulla di medie dei gruppi uguale a quella generale il rapporto

$$F = \frac{S_{tra}^2}{S_{entro}^2}$$

segue una distribuzione di Fisher con gradi di libertà pari a $(k-1, n-k)$. Sotto l'ipotesi nulla il valore campionario dovrebbe assumere un valore vicino ad 1. Nel caso in cui l'ipotesi nulla non sia vera, la statistica tenderà ad avere valori maggiori di 1.
Il tipo di test è quindi unilatero. Quindi è possibile trovare il valore critico di F sotto l'ipotesi nulla, corrispondente ad un valore di significatività pari ad α. Se il valore campionario di *F* è superiore è possibile rifiutare l'ipotesi nulla con una significatività pari ad α. Avendo a disposizione un calcolatore è inoltre possibile calcolare l'area sotto la curva *F* per valori superiori al valore *F* campionario, determinando quindi il *p* value. Avendo a disposizione le tabelle, tale valore può essere solo approssimato.

La tabella ANOVA, viene costruita in questo modo

Sorgente della variabilità	*Somma dei Quadrati (SS) Devianza*	*df o gdl (degrees of freedom/ gradi di libertà)*	*MS (media dei quadrati) Stima della Varianza*	*F*	*p*
Tra gruppi	SS_{entro}	*n-k*	S^2_{entro}	*Valore di F campionario*	*P value corrispondente*
Entro Gruppi	SS_{tra}	*k-1*	S^2_{tra}		
Totale	SS_{tot}	*n-1*			

Si fa notare che tale tabella vista ha la stessa forma della tabella fornita dal comando matlab *anova1*.

Anova a due criteri di classificazione (due vie)

Nel caso in cui le misure siano classificate secondo due fattori, ad esempio si classificano i soggetti di uno studio rispetto a due trattamenti, allora si parla di ANOVA a due vie. Se per ciascuna combinazione dei due fattori sono presenti più misure, o repliche, allora sarà possibile studiare l'interazione tra i due fattori.
Vedremo per prima l'anova senza repliche, in questo caso la tabella delle misure può essere vista in questo modo, dove il fattore *A* presenta *a* livelli e il fattore *B* presenta *b* livelli.

		Fattore B			
		1	2	…	b
Fattore A	1	X_{11}	X_{12}	…	X_{1b}
	2	X_{21}	X_{22}	…	X_{2b}
	3				X_{3b}
	…			..	…
	a	X_{a1}	X_{a2}	…	X_{ab}

Il modello che non prevede interazioni può essere scritto in questo modo

$$x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

dove α_i è l'effetto imputabile al fattore *A*, e β_j è l'effetto imputabile al trattamento *B*. Anche in questo caso la variabilità a fattori non controllabili e non imputabile ai due fattori è descritta dal termine ε_{ij}. Come nel caso dell'anova ad una via, studieremo l'effetto dei fattori *A* e *B* analizzando la variabilità nei dati, considerandoli nel loro insieme o raggruppando in funzione dei due fattori. Analizzeremo un modello ad *effetti fissi*, quindi considereremo α_i e β_j come costanti e quindi non descritti da variabili aleatorie. Tali modelli si distinguono da quelli casuali o random, nei quali entrambi i fattori sono descritti da variabili aleatorie, e da quelli misti, nei quali alcuni fattori sono random e altri sono fissi. È importante sottolineare che nel caso di effetti fissi, siamo interessati a verificare l'effetto dei

trattamenti nei gruppi sotto studio e non vogliamo estendere gli eventuali risultati alla popolazione dalla quale il nostro campione è estratto.
Le ipotesi nulle sono

H_a: effetti fattore *A* nulli (effetti di riga α_i pari a 0, per ogni *i*)
H_b: effetti fattore *B* nulli (effetti di colonna β_j pari a 0, per ogni *j*)
Definita la media totale $\bar{\bar{x}} = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b} x_{ij}$ dove *n* è il numero totale delle misure, si ottiene la devianza totale

$$SS_{Tot} = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(x_{ij} - \bar{\bar{x}}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} x_{ij}^2 - \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{b} x_{ij}\right)^2}{n}$$

La stima dei fattore di riga α_i è pari a $\hat{\alpha}_i = \frac{1}{b}\sum_{j=1}^{b} x_{ij} - \bar{\bar{x}}$

In maniera analoga si definiscono le stime dei fattori di colonna $\hat{\beta}_j = \frac{1}{a}\sum_{i=1}^{a} x_{ij} - \bar{\bar{x}}$

la devianza spiegata dalle differenze tra le righe, fattore *A*, è

$$SS_A = b\sum_{i=1}^{a}\hat{\alpha}_i^2 = b\sum_{i=1}^{a}(\bar{x}_i - \bar{\bar{x}})^2$$

dove $\bar{x}_i$ è la media della generica riga *i*.

Analogamente per il fattore *B*

$$SS_B = a\sum_{j=1}^{b}\hat{\beta}_i^2 = a\sum_{j=1}^{b}\left(\bar{x}_j - \bar{\bar{x}}\right)^2$$

dove $\bar{x}_j$ è la media della generica colonna *j*.
La devianza dell'errore o dei residui è

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\varepsilon_{ij}^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}}\right)^2 = SS_{Tot} - SS_A - SS_B$$

si fa notare che una opportuna scelta di μ permette di avere $\sum_i \alpha_i = 0$ e $\sum_j \beta_j = 0$.

Da queste si possono ricavare le rispettive varianze, tenendo conto dei gradi di libertà,

quindi

$$S_{Tot}^2 = \frac{SS_{Tot}}{n-1} = \frac{SS_{Tot}}{ab-1}$$

$$S_A^2 = \frac{SS_A}{a-1}\,, \qquad S_B^2 = \frac{SS_B}{b-1},\;\; S_E^2 = \frac{SSE}{(a-1)(b-1)}$$

Sotto le ipotesi nulle prima viste si ha che il rapporto tra le varianze viste segue una distribuzione di Fisher e possiamo applicare lo stesso approccio visto per l'anova ad una via. Tale approccio viene esemplificato nella tabella seguente.

Sorgente della variabilità	*Somma dei Quadrati (SS) Devianza*	*df o gdl (degrees of freedom/ gradi di libertà)*	*MS (media dei quadrati) Stima della Varianza*	*Valore di F campionario F*
Tra righe	SS_A	*a-1*	S_A^2	$F_A = \frac{S_A^2}{S_E^2}$
Tra colonne	SS_B	*b-1*	S_B^2	$F_B = \frac{S_B^2}{S_E^2}$
Errore	*SSE*	*(a-1)(b-1)*	S_E^2	
Totale	SS_{tot}	*ab-1*		

F_A è una statistica di Fisher con gradi di libertà *F(a-1;(a-1)(b-1))*
F_B è una statistica di Fisher con gradi di libertà *F(b-1;(a-1)(b-1))*

Nel caso di ipotesi nulla valida entrambe le statistiche valgono circa 1 e assumono valori più elevati nel caso ci si allontani dallo scenario di ipotesi nulla valida.

Anova a due vie con interazione

Nel caso in cui si consideri un modello non semplicemente additivo, ma con presente un contributo legato all'interazione tra i fattori, è necessario che per ogni cella della tabella precedente siano presenti più misure. In pratica, noi faremo l'ipotesi di avere lo stesso numero di misure per ogni cella in modo da avere un modello bilanciato.
Il modello sarà così modificato

$$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

dove è stato aggiunto un fattore $(\alpha\beta)_{ij}$ che descrive l'effetto imputabile all'interazione dei due fattori. Si suppone di avere per ogni cella (*i*,*j*), *r* repliche.

Analogamente a quanto visto per il modello precedente si possono calcolare le devianze calcolate tra le medie di riga, fattore *A*, di colonna o fattore *B*, e considerando le medie per ogni cella fattore di interazione *AB*.

Le ipotesi nulle sono:
H_a: effetti fattore *A* nulli (effetti di riga α_i pari a 0, per ogni *i*)
H_b: effetti fattore *B* nulli (effetti di colonna β_j pari a 0, per ogni *j*)
$H_{(ab)}$: effetti fattore *AB* nulli (effetti $(\alpha\beta)_{ij}$ pari a 0, per cella *ij*)

Se definiamo le seguenti quantità: $\bar{\bar{x}}$ è la media generale, $\bar{x}_i$ è la media sulle righe, $\bar{x}_j$ è la media sulle colonne, $\bar{x}_{ij}$ è la media di ogni cella.

Si può costruire la tabella per lo studio delle variabilità in maniera analoga a quanto fatto precedentemente per il caso di anova a due vie.

Sorgente della variabilità	*Somma dei Quadrati (SS) Devianza*	*gdl*	*MS (media dei quadrati) Stima della Varianza*	*Valore di F campionario F*
Tra righe	$SS_A = br\sum_i (\bar{x}_i - \bar{\bar{x}})^2$	*a-1*	S_A^2	$F_A = \frac{S_A^2}{S_E^2}$
Tra colonne	$SS_B = ar\sum_j \left(\bar{x}_j - \bar{\bar{x}}\right)^2$	*b-1*	S_B^2	$F_B = \frac{S_B^2}{S_E^2}$
Interazione	$SS_{AB} = r\sum_j \sum_i \left(\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}}\right)^2$	*(a-1)(b-1)*	S_{AB}^2	$F_{AB} = \frac{S_{AB}^2}{S_E^2}$
Errore	$SSE = \sum_j \sum_i \sum_k \left(\bar{x}_{ijk} - \bar{x}_{ij}\right)^2$	*ab(r-1)*	S_E^2	
Totale	$SS_{Tot} = \sum_j \sum_i \sum_k \left(\bar{x}_{ijk} - \bar{\bar{x}}\right)^2$	*abr-1*		