
Analisi della varianza

Mediante un singolo esperimento vengono confrontate fra loro più popolazioni (gruppi, tesi).

Consente di valutare quantitativamente l'importanza delle diverse fonti di variazione nella variabilità osservata nel corso di un esperimento. Le **fonti di variazione** possono essere:

- **sistematiche** (sotto controllo dello sperimentatore);
- **casuali** (variabilità biologica, condizioni ambientali, errore di misura, ecc..)

Fattore sperimentale e disegni sperimentali

Fattore sperimentale: fonte di variabilità il cui effetto si vuole determinare sulla base dei risultati dell'esperimento

Il fattore assume più valori, detti **livelli** o **modalità** (per es. dosi)

In generale si considerano più fattori sperimentali ed i **trattamenti** sono determinati dalle **combinazioni dei livelli** dei fattori sperimentali

Ogni trattamento deve essere applicato a **più unità sperimentali (replicazioni)**

Il disegno completamente randomizzato

Il disegno sperimentale più semplice è detto **disegno completamente randomizzato**

Si utilizza quando si considera **un solo fattore** sperimentale a più livelli, che in questo caso coincidono coi trattamenti

I trattamenti sono assegnati alle unità sperimentali in modo casuale (**randomizzazione**)

Se il **numero di repliche** è **uguale** per tutti i **trattamenti** il disegno è detto **bilanciato** (preferibile), altrimenti è detto sbilanciato

I dati

	Trattamenti						
	1	2	...	i	...	p	
	Y_{11}	Y_{21}	...	Y_{i1}	...	Y_{p1}	
	Y_{12}	Y_{22}	...	Y_{21}	...	Y_{p1}	
	...	...	...	...	...	...	
	Y_{1j}	Y_{2j}	...	Y_{ij}	...	Y_{pj}	
	...	...	...	...	...	...	
	Y_{1n}	Y_{2n}	...	Y_{in}	...	Y_{pn}	
Medie	$\bar{Y}_{1.}$	$\bar{Y}_{2.}$	...	$\bar{Y}_{i.}$	...	$\bar{Y}_{p.}$	$\bar{Y}$

Esempio

Si sono messi a confronto 4 diversi tipi di atmosfera modificata (aria normale: A, 5%O₂+3%CO₂: B, 3%O₂+3%CO₂: C e 1%O₂+3%CO₂: D) per identificare le migliori condizioni per la conservazione dei fagioli. I risultati, relativi alla concentrazione di proteine totali dopo 11 giorni di conservazione, sono espressi in g/100 g

Per ogni tesi sono state effettuate 6 repliche. I trattamenti sono stati assegnati a caso alle unità sperimentali.

Il disegno dell'esperimento è detto completamente casualizzato (randomizzato). Il disegno è bilanciato perché tutti i trattamenti presentano lo stesso numero di repliche

Trattamenti				
A	B	C	D	
1.54	1.57	1.55	1.61	
1.54	1.56	1.66	1.65	
1.62	1.66	1.64	1.60	
1.56	1.56	1.53	1.89	
1.55	1.56	1.60	1.61	
1.54	1.57	1.67	1.65	
1.56	1.58	1.60	1.67	1.60

Il modello lineare - 1

I dati derivanti dall'esperimento possono essere rappresentati mediante un **modello lineare**

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

in cui

- Y_{ij} è la generica osservazione dell' i -esimo trattamento sulla j -esima unità sperimentale
- μ_i è la **media del trattamento**
- ϵ_{ij} è l'**errore sperimentale**

Il modello lineare - 2

Il modello può essere rappresentato anche in un'altra forma

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

con μ media di tutte le popolazioni rappresentate nell'esperimento
e $\alpha_i = \mu - \mu_i$ **effetto** dell' i -esimo trattamento

Generalmente si assume $i = 1, \dots, p$ (p numero di trattamenti) e
 $j = 1, \dots, n_i$ (n_i numero di repliche per l' i -esimo trattamento). Se
il disegno è **bilanciato**, $n_1 = n_2 = \dots = n_p = n$

Le assunzioni

Gli errori sperimentali devono soddisfare **tre assunzioni**:

- devono essere **mutualmente indipendenti**
- devono essere a **varianza costante** (σ^2) entro trattamento e tra trattamenti
- devono avere **distribuzione normale**

Inoltre, il modello stesso impone l'**additività** tra componente sistematica e componente casuale

Come fare inferenza

Il modello lineare di analisi della varianza è un **modello teorico** che descrive le caratteristiche del fenomeno che stiamo studiando. Possiamo essere interessati a:

- **stimare** i parametri del modello, ossia **gli effetti dei trattamenti**
- **sottoporre a verifica** ipotesi sulle caratteristiche del fenomeno studiato, tradotte in opportune **ipotesi sui parametri** del modello stesso

Test delle ipotesi - 1

Le ipotesi che vengono sottoposte a verifica sono:

H_0 : i trattamenti sono equivalenti

H_1 : i trattamenti non sono equivalenti

che, in termini di parametri del modello, si possono formulare nel modo seguente:

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

H_1 : almeno un μ_i diverso dagli altri

Test delle ipotesi - 2

Notiamo che:

- l'ipotesi alternativa comprende molteplici situazioni, per cui viene specificata semplicemente come negazione dell'ipotesi nulla
- entrambe le ipotesi si possono esprimere in termini di α , per esempio

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$$

Come costruire il test?

Il test è basato sulla seguente considerazione:

Se è **vera** l'ipotesi **nulla**, i dati differiscono tra loro per il **solo** effetto della variabilità **casuale**

Se invece è **vera** l'ipotesi **alternativa**, **entrambe** le fonti di variabilità contribuiscono a determinare la variabilità **complessiva**

Il test è quindi basato sull'analisi della variabilità complessiva in funzione delle diverse cause (da cui il termine **Analisi della Varianza**)

Misura della variabilità

La variabilità dei dati osservati può essere misurata mediante gli scostamenti dei dati dalla media. La **devianza totale** è definita nel modo seguente:

$$\sum_i \sum_j (Y_{ij} - \bar{Y})^2$$

La devianza totale può essere scomposta nel modo seguente:

$$\begin{aligned} \sum_i \sum_j (Y_{ij} - \bar{Y})^2 &= n \sum_i (\bar{Y}_i - \bar{Y})^2 + \sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2 \\ SS(\mathbf{y}) &= SS(\mathbf{a}) + SS(\mathbf{e}) \end{aligned}$$

Le devianze

Le due quantità sono dette rispettivamente:

- **Devianza tra gruppi (trattamenti)**, $SS(a)$: misura la quota di variabilità attribuibile alle differenze tra i trattamenti
- **Devianza entro gruppi (d'errore)**, $SS(e)$: misura la quota di variabilità imputabile a tutte le cause non controllate nell'esperimento e all'errore di campionamento

Cosa ci aspettiamo

Se l'ipotesi nulla è **vera**, ci possiamo attendere uno **scarso contributo della devianza tra gruppi** alla devianza totale

Sel'ipotesi nulla è **falsa**, ci possiamo attendere che **entrambe le devianze** contribuiscano a determinare la devianza totale

A questo livello non è però possibile fare confronti, perché le devianze hanno un **numero di addendi diverso**

I gradi di libertà

Ad ognuna delle devianze sono associati i gradi di libertà:

- la devianza **totale** ha $np - 1$ gradi di libertà
- la devianza **tra gruppi** ha $p - 1$ gradi di libertà
- la devianza **d'errore** ha $p(n - 1)$ gradi di libertà

I gradi di libertà si **scompongono additivamente** come le devianze

Le varianze

Le **varianze** (**Mean Squares** o **Quadrati Medi**) si ottengono dividendo le devianze per i gradi di libertà. Avremo quindi:

- $MS(\mathbf{a}) = \frac{SS(\mathbf{a})}{p-1}$, **varianza tra trattamenti**

- $MS(\mathbf{e}) = \frac{SS(\mathbf{e})}{p(n-1)}$, **varianza d'errore**

Il test - 1

L'ipotesi nulla di **equivalenza dei trattamenti** è formulata nel modo seguente:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p = \mu$$

o, alternativamente:

$$H_0 : \alpha_i = 0, \quad i = 1, 2, \cdots, p$$

Sotto l'ipotesi nulla i dati provengono quindi da un'**unica popolazione** di media μ e varianza σ^2

Il test - 2

Il test è basato sul confronto tra la varianza tra trattamenti e la varianza d'errore, sulla base delle considerazioni seguenti:

Se H_0 è **vera**, allora ci aspettiamo che nel campione estratto
 $MS(\mathbf{a}) \simeq MS(\mathbf{e})$

Se H_0 è **falsa**, allora ci aspettiamo che nel campione estratto
 $MS(\mathbf{a}) \gg MS(\mathbf{e})$

Si può infatti dimostrare che:

$$E[MS(\mathbf{a})] = \sigma^2 + \frac{n \sum \alpha_i^2}{p-1} \quad \text{e} \quad E[MS(\mathbf{e})] = \sigma^2$$

Il confronto

Per confrontare le due varianze si utilizza il rapporto

$$\frac{MS(\mathbf{a})}{MS(\mathbf{e})}$$

rifiutando H_0 quando esso è molto **elevato**

Il valore critico

Sotto l'ipotesi nulla vale che (rapporto tra varianze):

$$\frac{MS(\mathbf{a})}{MS(\mathbf{e})} \sim F_{p-1, p(n-1)}$$

dove $F_{p-1, p(n-1)}$ indica la distribuzione F di Fisher con $p - 1$ e $p(n - 1)$ gradi di libertà

Si **rifiuta** l'ipotesi nulla quando

$$\frac{MS(\mathbf{a})}{MS(\mathbf{e})} > F_{p-1, p(n-1); \alpha}$$

Esempio

Tabella di analisi della varianza

Fonti di variazione	GdL	SS	MS	F _{oss}
Tra trattamenti	3	0.041	0.014	2.99
Errore	20	0.091	0.005	
Totale	23	0.132		

Per un test con un livello di protezione pari al 5%, bisogna confrontare il valore osservato di F con il valore critico che si ricava dalle tavole. In questo caso, $F_{p-1, p(n-1); 0.05} = 3.10$ e quindi l'ipotesi nulla viene accettata