ANALISI DELLA VARIANZA

Introduzione

- Ora abbiamo abbastanza strumenti per cominciare a porre domande più complesse ai nostri dati. Al momento siamo in grado di confrontare coppie di oggetti (t test, F test), vogliamo confrontare gli effetti di più cause sui nostri dati
- Ad esempio quando mediante un singolo esperimento vengono confrontate fra loro più popolazioni (gruppi, tesi).
- Voglio valutare quantitativamente l'importanza delle diverse fonti di variazione nella variabilità osservata nel corso di un esperimento. Le **fonti di variazione** possono essere:
    1. **sistematiche** (sotto controllo dello sperimentatore);
    2. **casuali** (variabilità biologica, condizioni ambientali, errore di misura, ecc..)

La tecnica per ottenere questo è l'**Analisi della Varianza**

Alcune definizioni

- **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**).

Alcune definizioni

- **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**.

Organizzazione dei dati

Formalmente diremo che abbiamo p trattamenti e n repliche

1	2	...	i	...	p
Y_{11}	Y_{21}	...	Y_{i1}	...	Y_{p1}
Y_{12}	Y_{22}	...	Y_{i2}	...	Y_{p2}
...	...	...	...	...	...
Y_{1j}	Y_{2j}	...	Y_{ij}	...	Y_{pj}
...	...	...	...	...	...
Y_{1n}	Y_{2n}	...	Y_{in}	...	Y_{pn}

Indicheremo con $\bar{Y}_{i.}$ le medie calcolate sui trattamenti ad esempio la media del trattamento 1 la indicheremo con $\bar{Y}_{1.}$

Esempio I

- Si sono messi a confronto 4 diversi tipi di atmosfera modificata (aria normale: A; $5\% O_2 + 3\% CO_2 2$ B; $3\% O_2 + 3\% CO_2$ C; $1\% O_2 + 3\% CO_2$; 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 $g/100g$.
- Per ogni tesi sono state effettuate 6 replicazioni. 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 replicazioni.

Esempio I

	A	B	C	D
Risultati	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

Valori medi: $\bar{Y}_{A.} = 1.56$, $\bar{Y}_{B.} = 1.58$, $\bar{Y}_{C.} = 1.61$ e $\bar{Y}_{D.} = 1.67$
La domanda è: **I trattamenti applicati sono alla base delle differenze osservate tra le medie?**

Modelli Statistici

Per poter rispondere a questa domanda in modo rigoroso dobbiamo riformulare il problema in termini di rappresentazione matematica dell'intera questione. Come possiamo rappresentare il problema formalmente? Come possiamo rappresentare le osservazioni in funzione delle loro medie?

Modello Lineare

$$Y_{ij} = \mu_{i.} + \varepsilon_{ij}$$

oppure

$$Y_{ij} = \mu + \nu_{i.} + \varepsilon_{ij}$$

dove

- nella prima espressione $\mu_{i.}$ è l'effetto trattamento i e ε_{ij} l'errore sperimentale
- nella seconda espressione μ è un effetto medio generale, $\nu_{i.}$ è l'effetto trattamento i, ottenibile come $\nu_{i.} = \mu - \mu_{i.}$ e ε_{ij} l'errore sperimentale

Modelli Statistici

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.

Modelli Statistici

- Il modello lineare di analisi della varianza ‘e un modello teorico che descrive le caratteristiche del fenomeno che stiamo studiando. Possiamo essere interessati a:
  1. stimare i parametri (elementi) del modello, ossia gli effetti dei trattamenti;
  2. sottoporre a verifica ipotesi sulle caratteristiche del fenomeno studiato, tradotte in opportune ipotesi sui parametri del modello stesso.

Test ANOVA

- 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:
  $H_0 : \mu_1 = \mu_2 = \ldots = \mu_p$
  H_1 : almeno un μ_i è diverso dagli altri
- l'ipotesi alternativa comprende molteplici situazioni, per cui viene specificata semplicemente come negazione dell'ipotesi nulla
- le ipotesi possono essere riformulate anche in termini di ν_i, ad esempio:
  $H_0 : \nu_1 = \nu_2 = \ldots = \nu_p = 0$

Test ANOVA

Come costruiamo 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 s**ull'analisi della variabilità complessiva in funzione delle diverse cause** (da cui il termine Analisi della Varianza).

Test ANOVA

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

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

Questa misura è scomponibile in:

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

Test ANOVA

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 (dell'errore)**, $SS(e)$: misura la quota di variabilità imputabile a tutte le cause non controllate nell'esperimento e all'errore di campionamento

Ci aspettiamo che:

- Se è vera H_0, ci possiamo attendere uno scarso contributo della devianza tra gruppi ($SS(a)$) alla devianza totale.
- Se è falsa H_0, entrambe le devianze dovrebbero contribuire a determinare la devianza totale.

Attenzione però le devianze hanno un numero di addendi diverso, non posso confrontarle.

Test ANOVA

Ad ognuna delle devianze sono associati i suoi 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.

Test ANOVA

Posso ora definire le varianze dividendo le devianze per i gradi di libertà:

$$
\begin{aligned}
MS(a) &= \frac{SS(a)}{p-1} \text{ varianza tra gruppi} \\
MS(e) &= \frac{SS(e)}{p(n-1)} \text{ varianza dell'errore}
\end{aligned}
$$

Test ANOVA

L'ipotesi di eguaglianza tra i trattamenti è formulata come:

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

oppure

$$H_0 : \nu_i = 0; \ i = 1, \ldots, p$$

Sotto l'ipotesi nulla i dati provengono quindi da un'unica popolazione di media μ e varianza σ^2. Il test è basato sul confronto tra la varianza tra trattamenti e la varianza dell'errore, sulla base delle considerazioni seguenti:

- <u>Se è vera</u> H_0 mi aspetto che $MS(a)$ sia quasi uguale a $MS(e)$
- <u>Se è falsa</u> H_0 mi aspetto ch $MS(a) >> MS(e)$

Test ANOVA

In termini tecnici ... Si può infatti dimostrare che:

$$E[MS(a)] = \sigma^2 + \frac{n \sum_i \nu_i^2}{p-1}$$

e che

$$E[MS(e)] = \sigma^2$$

quindi se gli effetti (ν_i) sono trascurabili $MS(a) \simeq MS(e)$.
Per confrontare queste due varianze useremo di nuovo il loro rapporto, quindi la statistica test è:

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

che si distribuisce come una F con $p-1$ e $p(n-1)$ gradi di libertà.

Fissato l'errore ammissibile α rifiuteremo l'ipotesi nulla quando $p-value < \alpha$.

Test ANOVA: Esempio I

Torniamo al nostro esempio delle atmosfere per la conservazione dei fagioli. E vediamo che risultato da l'analisi della varianza sui 4 trattamenti applicati:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Trattamento	3	0.040912	0.0136375	2.9956	0.05509
Residuals	20	0.091050	0.0045525		

Il p-value (colonna Pr(>F)) è pari a 0.05509, quindi ad un livello di $\alpha = 0.05$ accetto H_0. Se invece fisso una soglia d'errore più alta ad esempio $\alpha = 0.1$ rifiuto H_0

Disegno a blocchi randomizzati

Introducendo l'analisi delle varianza abbiamo parlato di sorgenti variabilità controllabili da chi conduce l'esperimento.
Questo viene fatto tramite il **disegno sperimentale** ovvero tramite il modo in cui somministriamo i trattamenti alle unità sperimentali
Nell'esempio uno abbiamo preso un **fattore sperimentale** (l'atmosfera in cui conservare i fagioli) e abbiamo costruito un **disegno completamente randomizzato**.
Come facciamo a controllare la situazione quando abbiamo più di un fattore sperimentale?

Disegno a blocchi randomizzati

Supponiamo di avere due fattori sperimentali, come nell'esempio dei topi. Voglio vedere se la variazione di peso nei ratti è determinata dalla sola quantità di cibo o anche dal tipo di cibo. Quindi ho $DA = 1, 2$ (1=alto, 2=basso) e $DT = 1, 2, 3$ (1=carne, 2=maiale, 3=cereali), ho 60 ratti e per tener conto dei due fattori in modo appropriato devo *assegnare lo stesso numero di individui ad ogni combinazione dei livelli dei due fattori*.
Ho $2 \times 3 = 6$ combinazioni. In pratica creo 6 gruppi di 10 ratti (blocchi) e a ciascun gruppo somministro una delle possibili combinazioni in modo casuale.
L'esempio dei ratti può essere descritto nel modo seguente:

- Due fattori: DA e DT. Considero l'ammontare come *fattore di raggruppamento (bloccaggio)* e il tipo di dieta come fattore sperimentale
- Costituisco due gruppi di 30 ratti ciascuno a cui assegno un ammontare di cibo,
- all'interno di ciascuno dei due gruppi distribuisco il tipo di dieta in modo casuale.

Disegno a blocchi randomizzati

Più in generale cercheremo di bloccare (raggruppare) le nostre unità sperimentali rispetto a quelle caratteristiche che non possiamo controllare direttamente e che potrebbero influire sulla variabilità.

Consideriamo un altro esempio: voglio studiare l'influenza di 3 fertilizzanti (A,B,C) sul tempo di fioritura di un tipo di piante di soia. I fagioli di soia vengono studiati in una serra che non ha condizioni climatiche del tutto uniformi (illuminazione, temperatura etc.). Queste quindi possono influire sulla variabilità del mio esperimento. Supponiamo di poter individuare 4 zone omogenee per condizioni ambientali allora disegno l'esperimento come segue:

Blocco 1	Blocco 2	Blocco 3	Blocco 4
B	A	B	C
A	C	C	A
C	B	A	B

Inciso: ANOVA a due vie

Abbiamo due fattori con p e q livelli rispettivamente ed n repliche. Il modello di analisi della varianza con due fattori è:

$$Y_{ij} = \mu + \nu_{i1} + \nu_{i2} + \varepsilon_{ij}$$

Dove ν_{i1} = effetto del fattore sperimentale 1 e ν_{i2} = effetto del fattore sperimentale 2. Con due fattori potrei chiedermi se esiste anche un effetto combinato dei due ovvero *se esiste un'interazione tra i due fattori* e quindi se il modello è del tipo:

$$Y_{ij} = \mu + \nu_{i1} + \nu_{i2} + \eta_{i1*2} + \varepsilon_{ij}$$

la devianza totale si scompone di conseguenza, con gradi di libertà:

$$\begin{array}{ccccccccc} SS(y) & = & SS(a_1) & + & SS(a_2) & + & SS(a_1 \times a_2) & + & SS(e) \\ npq-1 & = & (p-1) & + & (q-1) & + & (p-1)(q-1) & + & pq(n-1) \end{array}$$

Per poter stimare tutti i termini del modello è fondamentale come si organizza il disegno sperimentale e il numero di repliche.

Disegno a blocchi randomizzati

Risultati in giorni

Blocco	A	B	C	Totali
1	30	35	23	88
2	35	33	22	90
3	43	42	32	117
4	42	48	41	131
Totali	150	158	118	426

Disegno a blocchi randomizzati

Risultati ANOVA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Fertilizzante	2	228.17	114.083	14.512	0.005027 **
Blocco	3	429.58	143.194	18.215	0.002038 **
Residuals	6	47.17	7.861		

E' possibile rispondere alla domanda *Blocchi e trattamenti interagiscono?* . Con questi dati no, sono troppo pochi (calcolare i gdl se inseriamo anche l'interazione).

Quadrati Latini

Il Quadrato Latino è un disegno sperimentale utile quando si vuole controllare la variabilità lungo due direzioni. Abbiamo un uguale numero di righe, colonne e trattamenti (r)

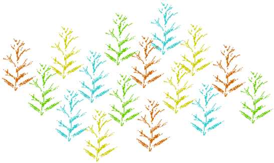

Figura: colori diversi corrispondono a diversi trattamenti

A	B	C	D
B	A	D	C
C	D	A	B
D	C	B	A

Quadrati Latini

Source of variation	Degrees of freedom[a]	Sums of squares (SSQ)	Mean square (MS)	F
Rows (*R*)	r-1	SSQ_R	$SSQ_R/(r-1)$	MS_R/MS_E
Columns (*C*)	r-1	SSQ_C	$SSQ_C/(r-1)$	MS_C/MS_E
Treatments (*Tr*)	r-1	SSQ_{Tr}	$SSQ_{Tr}/(r-1)$	MS_{Tr}/MS_E
Error (*E*)	(r-1)(r-2)	SSQ_E	$SSQ_E/((r-1)(r-2))$	
Total (*Tot*)	r^2-1	SSQ_{Tot}		

[a]where r=number of treatments, rows, and columns.

Figura: Tavola di scomposizione della varianza