ANOVA a blocchi randomizzati

Andrea Onofri

9 gennaio 2012

Indice

1 Disegno sperimentale

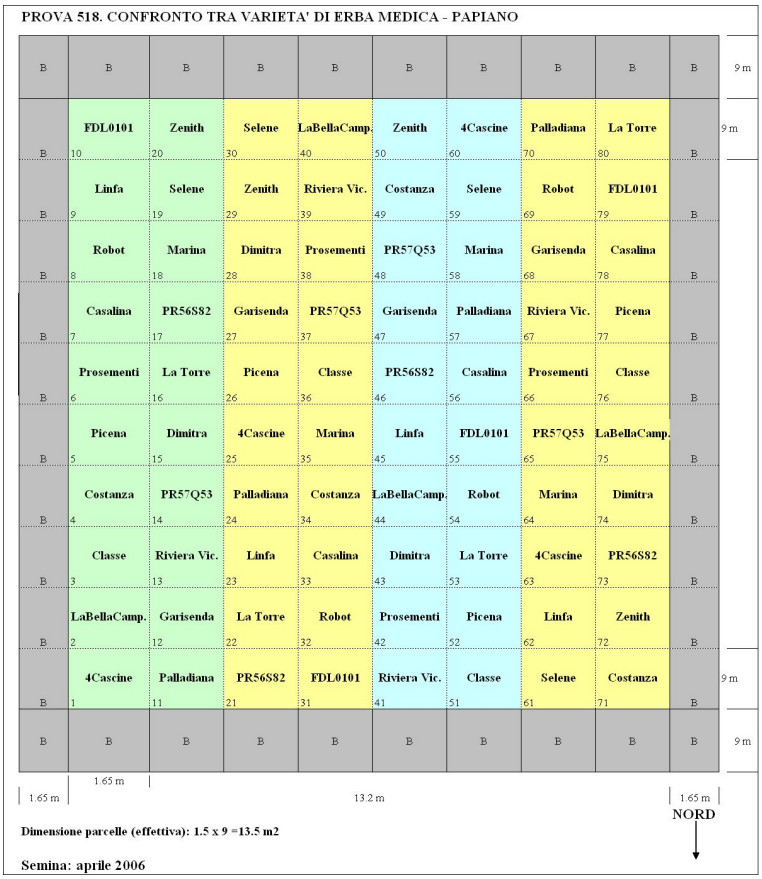
Protocollo sperimentale

- TIPO DI PROGETTO: Rete Varietale Nazionale
- IPOTESI SCIENTIFICA: la produttività dell'erba medica è fortemente dipendente dalla scelta varietale
- OBIETTIVO: confrontare tra loro le 20 varietà più promettenti in Italia
- CONTROLLO: non è necessario
- UNITA' SPERIMENTALI: Parcelle di 9 m^2 minimo
- DISEGNO SPERIMENTALE: Blocchi randomizzati con 4 repliche
- CONDUZIONE PROVA: ordinarietà per semina, concimazione, diserbo e altre operazioni colturali
- RILIEVI: produzione di fieno su una subparcella di almeno 6 m^2

Date le dimensioni della prova, è lecito ipotizzare che, pur scegliendo un appezzamento il più omogeneo possibile, saranno comunque evidenti dei gradienti di fertilità, soprattutto procedendo dai lati del campo (vicino alle fosse) verso il centro. In questa situazione, se l'esperimento fosse disegnato a randomizzazione completa, l'anzidetto gradiente di fertilità non verrebbe tenuto in considerazione e finirebbe nell'errore sperimentale, diminuendo l'efficienza dell'esperimento.

Di conseguenza, si impiega un disegno a blocco randomizzato con quattro repliche. Il campo viene suddiviso in tante sezioni (dette blocchi) quante sono le repliche, trasversalmente al gradiente di fertilità. In questo modo, ci si attende che all'interno di ciascuna sezione l'ambiente sia molto omogeneo, mentre potrebbero comparire delle eterogeneità tra una sezione e l'altra. In ciascun blocco si sistema una replica, randomizzando i trattamenti tra loro.

CRBD



E' opportuno precisare che esperimenti con questo layout sono molto diffusi in campo, ma potrebbero essere altrettanto utili considerando i ripiani di una cella climatica, i bancali di una serra, oppure le sezioni di una stalla, o gli appezzamenti di una coltura arborea. Rispetto alla randomizzazione completa, esiste un vincolo in più nel disegno, in quanto mentre nella randomizzazione completa i trattamenti sono assegnati casualmente alle 80 unità sperimentali, nel blocco randomizzato bisogna rispettare la suddivisione in blocchi e mettere in ogni blocco una sola replica di ciascun trattamento.

La suddivisione in blocchi evidenzia che (oltre al trattamento sperimentale, cioè la concimazione) le unità sperimentali differiscono anche per l'appartenenza ad uno dei blocchi, cioè per il fatto che si trovano in una sezione che, per la sua posizione o per il tipo di terreno o per altri aspetti, differisce in qualche modo dalle altre.

2 Elaborazione dei dati

ANOVA a blocchi randomizzati

- dataset 'medica.xls'
- Su ogni unità sperimentale, oltre all'errore sperimentale, agiscono due effetti: il trattamento ed il blocco.
- L'effetto varietale è altamente significativo ($P \leq 0.05$)

EFFECT	SS	DF	MS	F	ProbF	
Blocks	0.100	3.000	0.033	0.911		
Tesi	2.776	19.000	0.146	3.993	0.00003	**
Residual	2.086	57.000	0.0366			
Total	4.962	79.000	0.063			

Verifica assunzioni di base

- Prima di procedere, verifichiamo gli assunti di base per l'ANOVA
- Non ci sono outliers
- Il grafico dei residui non evidenzia 'patologie' di rilievo.
- Il test di Bartlett e il test di Levene non sono significativi ($P > 0.05$)
- I valori consigliati di lambda oscillano da -0.25 a più di 2.5 (infatti ln(RSS) è sempre minore del valore massimo pari a 0.795). Quindi la non trasformazione (lambda = 1) è una opzione accettabile

Non-additività dei blocchi

- L'ANOVA assume l'effetto 'additivo dei blocchi'
  - l'appartenenza ad un blocco implica un incremento/decremento costante per tutte itrattamenti nclusi in quel blocco;
  - La somma degli effetti dei diversi blocchi è zero
- può capitare che l'effetto del blocco non sia lo stesso per tutti i trattamenti (ad esempio: in una prova di concimazione, nei blocchi molto fertili, il testimone non concimato è favorito)

- l'assunto di addittività dei blocchi deve essere verificato, con il test di Tukey
- In questo caso, anche questo test è non significativo ($P > 0.05$)
- CONCLUSIONE: questo dataset sembra rispettare tutti gli assunti di base dell'ANOVA

Conclusioni

- Verificare i gradi di libertà. E' il modo migliore di capire se abbiamo fatto errori
- La deviazione standard del residuo è $\sqrt{(0.0366)} = 0.191$ q/ha
- La SEM = 0.096 q/ha
- Il coefficiente di variabilità è pari al 6.5% (che è ottimo, per questo tipo di prove varietali)

Efficienza del blocco randomizzato

- *Analyse them as you have randomised them (Fisher)*
- Un disegno a blocco randomizzato NON DOVREBBE essere elaborato a randomizzazione completa, perchè manca l'indipendenza (le unità nello stesso blocco sono più simili di quelle in blocchi diversi)
- Tuttavia è lecito chiedersi che cosa abbiamo guadagnato introducendo un vincolo nella randomizzazione
- Il criterio di giudizio è: abbiamo guadagnato se siamo riusciti ad ottenere un errore sperimentale più piccolo (controllo locale degli errori)

Se rifacciamo l'ANOVA a randomizzazione completa vediamo che la devianza del trattamento (e la varianza) non cambia, mentre la devianza del residuo somma le devianze del blocco e quella del residuo nell'ANOVA a blocchi randomizzati, ed è pari a $0.1 + 2.086 = 2.186$, alla quale corrisponde una SEM pari a 0.095. Insomma, non abbiamo guadagnato molto!!!!

Calcolo efficienza - 1

- Confrontiamo la varianza residua del CRD e del CRBD
- Per il confronto non possiamo utilizzare la varianza residua dell'ANOVA a randomizzazione completa (il disegno NON E' a randomizzazione completa, ma è vincolato)

- Stimiamo il residuo del disegno a randomizzazione completa, riunendo la varianza del blocco e quella del residuo nell'ANOVA a blocchi randomizzati:

$$s^2_{cr} = \frac{SS_{blocks} + r(t-1)MSE}{rt-1} = \frac{0.1 + 4 \cdot (20-1) \cdot 0.037}{79} = 0.0369$$

dove r è il numero delle repliche e t è il numero dei trattamenti.

Calcolo efficienza (RE) - 2

$$RE = \frac{0.0369}{0.0366} = 1.007$$

- Sostanzialmente, i due disegni hanno la stessa efficienza
- Bisognerebbe anche tener conto della semplicità (Occam's razor) e considerare che il blocco randomizzato è più complesso (usa più gradi di libertà) e quindi dovrebbe essere penalizzato per questo:

$$RE = \frac{(57+1)(60+3)}{(57+3)(60+1)} \frac{0.0369}{0.0366} = 0.998 \cdot 1.007 = 1.005$$

57 e 60 sono i gradi di libertà del residuo nei due disegni.

Insomma, il blocco randomizzato è poco più efficiente della randomizzazione completa, in questo esempio specifico.

Test di confronto multiplo

- Con 20 medie, abbiamo 190 confronti possibili
- In assenza di correlazioni, se i confronti sono fatti per P=0.05, la probabilità di sbagliarne almeno 1 è del 99.9 %
- Il tasso d'errore per esperimento è altissimo!
- Se ci interessano proprio tutti i 190 confronti, non possiamo utilizzare la MDS, ma preferiamo utilizzare la HSD
- In alternativa, se fossimo interessati solo al confronto con la varietà migliore (Pick the best: 19 confronti in totale) potremmo utilizzare il test di Dunnet (MCB) a una via.

MCP

Varietà	Produzione (t/ha)	MDS	HSD	MCB
PR57Q53	3.16	a	a	a
Casalina	3.13	a	a	a
FDL 0101	3.12	ab	a	a
Prosementi	3.10	ab	ab	a
PR56S82	3.09	ab	ab	a
Dimitra	3.07	abc	ab	a
Picena GR	3.04	abcd	ab	a
Zenith	3.02	abcde	ab	a
La Bella Campagnola	3.00	abcde	ab	a
Selene	2.99	abcdef	ab	a
La Torre	2.98	abcdef	ab	a
Classe	2.91	abcdef	abc	a
Marina	2.86	bcdefg	abc	a
Garisenda	2.81	cdefg	abc	a
Palladiana	2.80	cdefg	abc	a
Robot	2.79	defg	abc	
Linfa	2.77	efg	abc	
4 Cascine	2.73	fgh	abc	
Riviera Vicentina	2.60	gh	bc	
Costanza	2.46	h	c	
Differenza critica		0.271	0.503	0.366

3 Particolarità

Dati mancanti o aberranti

- Nei disegni a randomizzazione completa, il dato aberrante/mancante può essere semplicemente trascurato, senza nocumento per l'analisi
- Nel blocco randomizzato la mancanza di un dato crea problemi:
  - le medie aritmetiche sono distorte (*biased*; si pensi alla possibilità che ad una tesi venga a mancare il dato del blocco più produttivo (o meno produttivo);
  - le devianze (del blocco e del trattamento) nell'ANOVA divengono sensibili all'ordine con cui gli effetti vengono immessi nel modello (se mettiamo prima il blocco e poi la tesi otteniamo devianze diverse che non mettendo prima la tesi e poi il blocco)

Ricostruzione del dato mancante

La metodica più semplice per gestire lo sbilanciamento è quella 'ricostruire' il dato mancante con la seguente formula:

$$Y = \frac{tT + rR - G}{(t-1)(r-1)}$$

dove t è il numero delle tesi, r è il numero delle repliche, T è la somma dei dati relativi alla tesi che contiene il dato mancante (ovviamente escluso quest'ultimo), R è la somma dei dati relativi al blocco che contiene il dato mancante (sempre escluso quest'ultimo), G è il totale generale (escluso il dato mancante).

NB: Quando si corregge un dato bisognerebbe parallelamente ridurre di una unità il numero dei gradi di libertà della varianza residua e ricalcolare SEM e SED di conseguenza.

Se si vuole evitare la ricostruzione del dato, si possono utilizzare metodiche più avanzate, basate sul modello lineare e sulle medie '*Least Squares*', che tuttavia non possono essere trattate in questo corso, per motivi di spazio.