Sottocampionamento e misure ripetute

Andrea Onofri

February 8, 2012

Contents

1 Sub-sampling

Il sottocampionamento è una situazione che si verifica molto di frequente nella sperimentazione di pieno campo e richiede attenzioni particolari in termini di analisi dei dati, in relazione all'assunto fondamentale di indipendenza dei residui.

Sottocampionamento

- Molto di frequente lo sperimentatore esegue più di una singola misura per parcella.
- ESEMPIO: confronto varietale su frumento, blocco randomizzato con tre repliche. Raccolta di tre subsampioni di granella per parcella, per misurare il peso di 1000 semi (**File TKW.xls**).
- NON DOBBIAMO MAI ANALIZZARE IL DATASET COME SE AVESSE NOVE REPLICHE INDIPENDENTI PER OGNI VARIETA'!
- I dati che provengono dalla stessa parcella (e quindi i loro residui) non sono indipendenti, ma condividono la stessa parcella e sono più 'simili' tra loro che non quelli di parcelle diverse.

Sottocampionamento: soluzioni - 1

- Fare la media dei sottocampioni ed analizzarla.
- Va bene SE E SOLO SE il numero di sottocampioni è costante per ogni parcella. Se no, assegneremmo lo stesso peso a dati caratterizzati da diversa precisione.

- Perdiamo l'informazione relativa alla variabilità entro parcella (che talvolta è interessante!).

```
> model <- lm(Mean ~ Block + Genotype, data=TKW)
> anova(model)
Analysis of Variance Table

Response: Mean
              Df  Sum Sq Mean Sq F value    Pr(>F)
factor(Block)  2   36.78  18.390  2.0046    0.1439
Genotype      29 2408.24  83.043  9.0522 9.943e-13 ***
Residuals     58  532.08   9.174
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Sottocampionamento: soluzioni - 2

- Analizziamo l'intero dataset, ma considerando due 'strati' di errore sperimentale: tra parcelle ed entro parcella.
- Ricordate il concetto di 'repliche vere'? In questo caso le repliche vere sono le parcelle, perchè ricevono il trattamento in modo indipendente.
- L'effetto del trattamento va quindi testato sull'errore tra parcelle, non su quello entro parcella.
- Se il numero di sottocampioni è costante, i risultati sono gli stessi che nel caso precedente, ma l'informazione sull'errore entro-parcella non è persa.

Sottocampionamento: soluzioni - 2

```
> model <- aov(TKW ~ factor(Block) + Genotype + Error(factor(Plot)), data=TKW)
> summary(model)

Error: factor(Plot)
              Df Sum Sq Mean Sq F value    Pr(>F)
factor(Block)  2  110.3  55.169  2.0046    0.1439
Genotype      29 7224.7 249.129  9.0522 9.943e-13 ***
Residuals     58 1596.2  27.521
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: Within
           Df Sum Sq Mean Sq F value Pr(>F)
Residuals 180 152.15 0.84526
>
```

Sottocampionamento: soluzioni - 3

- Utilizzare un modello misto, con le parcelle inserite come effetto random.
- E' esattamente identica alla soluzione precedente, cambia solo la tipologia di stima (REML invece che minimi quadrati).
- La piattaforma di lavoro è più flessibile e consente una più efficiente gestione dei dati sbilanciati.
- E' una soluzione avanzata, ma non viene trattata in questa sede.

Sottocampionamento: test F, SEM e SED

- Il sottocampionamento non pone particolari problemi in termini di inferenza, se abbiamo analizzato le medie o qualche altra statistica riassuntiva per parcella
- Se si è analizzato l'intero dataset, ricordare che la voce d'errore esatta (per test F, SEM e SED) è quella relativa alle parcelle
- RICORDA: l'errore sperimentale è rappresentato dalla variazione tra unità sperimentali che hanno ricevuto lo stesso trattamento in modo indipendente!

2 Misure ripetute

Misure ripetute

- Anche qui le misure vengono ripetute sulle stesse unità sperimentali, ma non sono randomizzate e seguono una metrica spazio-temporale (**DATI LONGITUDINALI**).
- ESEMPI: raccolta di colture poliennali; prelievi a profondità diverse.
- Come nel caso del sottocampionamento, la misure ripetute non sono indipendenti;
- a differenza che nel caso precedente, però, le misure sono ordinate e non casuali.
- IMPORTANTE. Nel sottocampionamento le diverse misure sono spesso solo repliche, mentre in questo caso le misure ripetute hanno interesse individuale (es. i raccolti annuali di una coltura poliennale)

Cosa fare?

- MAI ANALIZZARE UN DISEGNO A MISURE RIPETUTE COME FOSSE UN FATTORIALE! Questo è un errore perchè le misure ripetute non sono indipendenti.
- DOBBIAMO DISTINGUERE DUE CASI:
- PRIMO CASO: le misure ripetute sono relativamente indipendenti tra loro (non c'è indipendenza nell'intera prova, ma c'è indipendenza entro parcella)
- SECONDO CASO: le misure prese sulla stessa unità sono tanto più correlate (cioè simili) quanto più sono 'vicine' (spazialmente o temporalmente) tra loro. Si può quindi parlare di 'autocorrelazione' seriale.

DATI LONGITUDINALI SENZA CORRELAZIONE SERIALE

- Il disegno è uno split-plot con il fattore sperimentale sulle main-plots a il tempo/spazio nelle sub-plots (*split-plot in time*).
- Utilizzando uno schema a split-plot, le osservazioni sperimentali sono naturalmente raggruppate per parcella, il che da conto della loro mancata indipendenza.
- Questa soluzione si può adottare solo in assenza di correlazione seriale (la condizione di Huynh-Feldt è valida, cioè stesso SED per tutte le possibili coppie di misure in qualunque tempo/spazio)
- Questa soluzione è generalmente adottata nel caso dei raccolti multipli delle colture poliennali.

Split-plot in time (alfalfa.xls)

```
> summary(aov(Yield~Block+Var*Year + Error(Block/Var/Year),
 data=Alfalfa))

Error: Block
      Df Sum Sq Mean Sq
Block  3 7.6511  2.5504

Error: Block:Var
          Df  Sum Sq Mean Sq F value   Pr(>F)
Var       19 104.269  5.4878  4.6024 3.75e-06 ***
Residuals 57  67.966  1.1924
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Error: Block:Var:Year
           Df  Sum Sq Mean Sq   F value Pr(>F)
Year        2 2602.53 1301.27 1847.1002 <2e-16 ***
Var:Year   38   31.83    0.84    1.1888 0.2389
Residuals 120   84.54    0.70
---
```

DATI LONGITUDINALI CON CORRELAZIONE SERIALE

- La condizione di Huynh-Feldt non è valida
- E' necessario utilizzare un modello misto, introducendo l'unità sperimentale (parcella) come fattore random.
- Introdurre nell'analisi la necessaria struttura di correlazione seriale (autoregressiva o altro), facendo attenzione a non utilizzare strutture troppo complesse (a questo scopo si utilizzano indicatori come AIC o BIC: più basso è il valore, migliore è il modello).
- Questi aspetti non vengono trattati in questa sede!

SPLIt-plot in time: SEM e SED

- Lo split-plot in time pone gli stessi problemi dello split-plot, in termini di SED multiple e approssimazione dei gradi di libertà (vedere la lezione apposita)
- Tener presente che, per le colture poliennali, l'anno può essere considerato un fattore random
- DSAASTAT consente di elaborare uno split-plot in time, sia con gli 'anni' fissi che random

Further readings

- LITTELL RC, MILLIKEN GA, STROUP WW, WOLFINGER RD & SCHABANBERGER O (2006) SAS for mixed models. SAS Publishing, 2nd Edition, Cary, NC, USA.
- PINHEIRO JC & BATES DM (2000) Mixed-effects models in S and S-Plus. Springer-Verlag New York Inc., New York.