Un'analisi del modello di cattura-ricattura

Patrick Zecchin

Università di Trento

26 settembre 2014

relatore: prof. Pier Luigi Novi Inverardi

Piano della presentazione

La stima della numerosità di popolazioni

Una questione di viva importanza, nella statistica e non solo

Varie le tecniche messe a punto, tra cui

- il campionamento *per centri*
- lo *snowballing*
- il metodo di *cattura ricattura*

Il modello iniziale

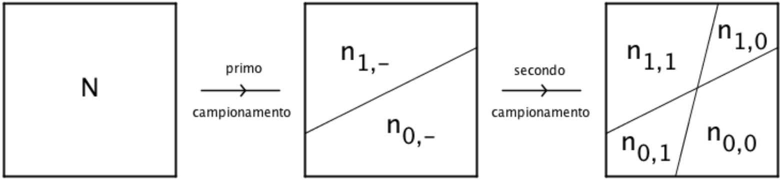
Uno sguardo storico

- Si tratta di un modello già proposto da Pierre Laplace nel 1802
- ma "ufficializzato" nel 1896 da Carl Petersen
  e nel 1930 da Frederick Lincoln
- utilizzato inizialmente in ambito ecologico: platesse e anatre

Il modello iniziale

Funzionamento del metodo di cattura-ricattura

Il metodo originariamente ideato è piuttosto semplice:



Il modello iniziale

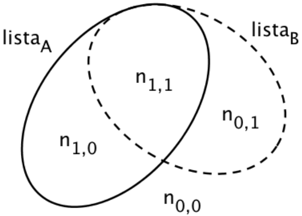
Le ipotesi da soddisfare

Per formulare il modello occorrono prima di alcune ipotesi:

1. popolazione chiusa
2. marcatura efficiente e non invasiva
3. omogeneità/equicatturabilità tra i soggetti,
   risolvibile con la stratigrafia
4. indipendenza delle liste,
   stimabile tramite l'*odds ratio* della tabella di contingenza

Il modello iniziale

Lo stimatore di Lincoln-Petersen



$$\frac{n_{1,1}}{n_{1,0}+n_{1,1}} = \mathbb{P}[x \in B \mid x \in A] \overset{indip.}{=} \mathbb{P}[x \in B \mid x \notin A] = \frac{n_{0,1}}{n_{0,0}+n_{0,1}}$$

$$\rightarrow \hat{N} = n_{1,1} + n_{1,0} + n_{0,1} + \hat{n}_{0,0} = \frac{n_A n_B}{n_{1,1}}$$

Il modello iniziale

Lo stimatore di Lincoln-Petersen: comportamento

Lo stimatore di Lincoln-Petersen $\hat{N} = \frac{n_A n_B}{n_{1,1}}$

- è asintoticamente non distorto:

$$\mathbb{E}[\hat{N}] \approx \frac{\mathbb{E}[n_A]\mathbb{E}[n_B]}{\mathbb{E}[n_{1,1}]} = \frac{N p_A p_B}{p_{A,B}} \overset{indip.}{=} N$$

- ma purtroppo è distorto per piccoli valori del campione

Il modello iniziale

Altri stimatori proposti

- stimatore di *Chapman-Seber* (1951)

$$\hat{N}_{Chapman} = \frac{(n_A + 1)(n_B + 1)}{n_{1,1} + 1} - 1$$

- stimatore di *Chao* (1987)

$$\hat{N}_{Chao} = n_{1,0} + n_{0,1} + n_{1,1} + \frac{(n_{1,0} + n_{0,1})^2}{4n_{1,1}}$$

L'evoluzione del modello

Importanti sviluppi nel XX secolo

Dopo Petersen (1896) e Lincoln (1930)

- Schnabel propone la versione generalizzata a k-liste (1938)
- Sekar e Deming stimano il numero di nascite e morti vicino a Calcutta (1949)
- il cattura-ricattura viene ampiamente utilizzato fuori dall'ecologia

L'evoluzione del modello

... e conseguenti difficoltà

Gli importanti sviluppi del XX secolo pongono nuove problematiche, ma si forniscono possibili nuove soluzioni, quali

- il *two-samples method*, proposto da Wittes
- il *sample coverage approach*, come ricorda Chao
- i modelli *log-lineari*, suggeriti da Fienberg

I modelli log-lineari

variabile B	variabile A cat_{A1}	cat_{A2}	totale
cat_{B1}	$n_{1,1}$	$n_{0,1}$	$n_{\_,1}$
cat_{B2}	$n_{1,0}$	$n_{0,0}$	$n_{\_,0}$
totale	$n_{1,\_}$	$n_{0,\_}$	n

$$\mu_{i,j} = \mathbb{E}[n_{ij}] = np_{i,j} \overset{indip.}{=} np_{i,\_}p_{\_,j}$$

$$\log[\mu_{i,j}] = \lambda + \lambda_i^A + \lambda_j^B \quad + \underbrace{\lambda_{ij}^{AB}}_{interazione}$$

con la necessità di testare tramite i test di *goodness of fit* χ^2 oppure L^2

L'evoluzione del modello

Utilizzo dei modelli log-lineari

Ci si trova di fronte alle relazioni che identificano il modello

$$\log \mathbb{E}[n_{1,1}] = \lambda + \lambda^A + \lambda^B + \lambda^{AB} \qquad \log \mathbb{E}[n_{0,0}] = \lambda - \lambda^A - \lambda^B + \lambda^{AB}$$

$$\log \mathbb{E}[n_{1,0}] = \lambda + \lambda^A - \lambda^B - \lambda^{AB} \qquad \log \mathbb{E}[n_{0,1}] = \lambda - \lambda^A + \lambda^B - \lambda^{AB}$$

Si tratta di trovare i termini λ per cui il modello meglio si adatta ai dati conosciuti

L'evoluzione del modello

Ulteriori questioni teoriche finali:

- la selezione del modello
  - criterio di *Akaike* $AIC = 2k - 2\log[L]$
  - criterio di *Bayes* $BIC \approx k\log[n] - 2\log[L]$

- la costruzione di intervalli di confidenza
  - con la soluzione classica
  - tramite la verosimiglianza profilo $\log[L_1(N_0)] > \log[L(\hat{N}, \hat{\delta})] - \frac{1}{2}\chi^2_{1-\alpha}(1)$
  - utilizzando il metodo *bootstrap*

Un caso concreto: una stima dell'incidenza del diabete

Inquadramento del problema

L'OMS stima in 380 milioni il numero di persone con diabete nel mondo, ponendolo come 8ª causa di morte

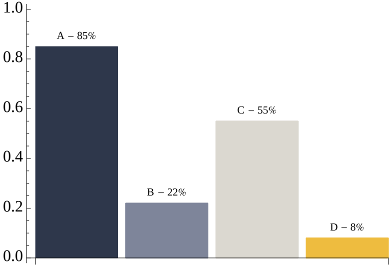
Una stima dell'incidenza del diabete

Dati di partenza

Nello studio del caso in oggetto sono state considerate 4 liste

a. elenco pazienti fornito da medici di famiglia (1754 casi)
b. registri di diagnosi di diabete fornito da ospedali piemontesi (452 casi)
c. database con prescrizioni di insulina e ipoglicemizzanti (1135 casi)
d. lista con richieste di rimborso per medicinali (173 casi)

per un totale di 2069 casi unici



		A	Yes	Yes	No	No
C	D	B	Yes	No	Yes	No
Yes	Yes		58	46	14	8
Yes	No		157	650	20	182
No	Yes		18	12	7	10
No	No		104	709	74	-

Una stima dell'incidenza del diabete

Primo approccio: il two-samples method

B	A: Yes	A: No
Yes	337	115
No	1417	-

Lincoln-Petersen: $\hat{N} = 2353$
Chao: $\hat{N} = 3610$
Chao C.I.: $3342 \div 3878$
Chapman: $\hat{N} = 2351$
Chapman C.I.: $2238 \div 2464$
odds ratio $= 1{,}6$

(a) *Liste A - B.*

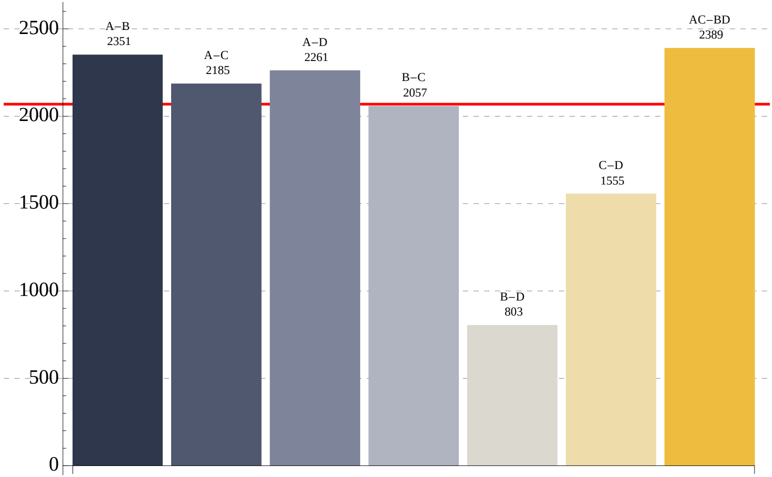
C	A: Yes	A: No
Yes	911	224
No	843	-

Lincoln-Petersen: $\hat{N} = 2185$
Chao: $\hat{N} = 2290$
Chao C.I.: $2235 \div 2345$
Chapman: $\hat{N} = 2185$
Chapman C.I.: $2141 \div 2229$
odds ratio $= 3{,}7$

(b) *Liste A - C.*

Una stima dell'incidenza del diabete

Primo approccio: il two-samples method



Una stima dell'incidenza del diabete

Un secondo approccio: i modelli log-lineari

Si possono utilizzare i modelli log-lineari, con una procedura *backward*, per determinare un modello dall'espressione

$$\begin{aligned}\log \mathbb{E}[n_{i,j,k,l}] =& \lambda + \lambda^{A} + \lambda^{B} + \lambda^{C} + \lambda^{D}\\ &+\lambda^{AB} + \lambda^{AC} + \lambda^{AD} + \lambda^{BC} + \lambda^{BD} + \lambda^{CD}\\ &+\lambda^{ABC} + \lambda^{ABD} + \lambda^{ACD} + \lambda^{BCD} + \lambda^{ABCD}\end{aligned}$$

Una stima dell'incidenza del diabete

Un secondo approccio: i modelli log-lineari

Dall'analisi deriva che

- il modello migliore è della forma:

$$\begin{aligned}\log \mathbb{E}[n_{i,j,k,l}] =& \lambda + \lambda^{A} + \lambda^{B} + \lambda^{C} + \lambda^{D} \\ &+ \lambda^{AB} + \lambda^{AC} + \cancel{\lambda^{AD}} + \lambda^{BC} + \lambda^{BD} + \lambda^{CD} \\ &+ \cancel{\lambda^{ABC}} + \cancel{\lambda^{ABD}} + \cancel{\lambda^{ACD}} + \cancel{\lambda^{BCD}} + \cancel{\lambda^{ABCD}}\end{aligned}$$

  con rapporto di verosimiglianza (la "precisione") $L^2 = 7,6$
- questo ci fornisce una stima di 2 771 casi (C.I. 2 492 - 3 051)
- il modello log-lineare di indipendenza ha $L^2 = 217,5$, da cui le incongruenze precedenti

Una stima dell'incidenza del diabete

I modelli log-lineari con stratigrafia

Si possono ulteriormente dividere i pazienti in base al tipo di trattamento e ripetere l'analisi

- dieta: 360 casi (C.I. 303 - 442)
- ipoglicemizzanti: 1 890 casi (C.I. 1 785 - 2 014)
- insulina: 333 casi (C.I. 328 - 341)
- totale: 2 583 casi (C.I. 2 416 - 2 798)

Questa è la migliore stima ottenibile con questo metodo.

Conclusioni

Con una tabella riassuntiva si vogliono schematizzare i risultati ottenuti tramite le diverse analisi

provenienza	casi accertati	precisione
elenco da medici di famiglia (lista A)	1 754	68%
registro ospedaliero (lista B)	452	17%
database prescrizioni (lista C)	1 135	44%
elenco rimborsi (lista D)	173	7%
two-samples method corretto (AC-BD)	2 389	92%
modello loglineare	2 771	107%
modello loglineare con stratigrafia	2 583	100%

Si evidenzia il deciso miglioramento nella stima dell'entità del problema.

Conclusioni

Anche dall'analisi del caso concreto risulta che

- innumerevoli sono le applicazioni della statistica e molti sono i modelli applicabili
- vi è una continua evoluzione e un continuo miglioramento della tecnica e delle metodologie
- tali metodi forniscono un'idea più corretta dell'entità dei problemi in esame.