UNIVERSITA' DEGLI STUDI DI TRENTO

DIPARTIMENTO DI MATEMATICA

CORSO DI LAUREA IN

MATEMATICA

TESI FINALE

TITOLO

UN'ANALISI DEL MODELLO DI CATTURA-RICATTURA

RELATORE

NOME

PROF. PIER LUIGI NOVI INVERARDI

LAUREANDO

NOME

PATRICK ZECCHIN

ANNO ACCADEMICO 2013/2014

Indice

Introduzione

L'argomento oggetto di questa tesi è il metodo statistico denominato *cattura-ricattura*, finalizzato alla stima della numerosità di una data popolazione: una metodologia nata verso la fine dell'Ottocento in campo ecologico, ma che nel corso del XX secolo si è sviluppata sia in termini di tecniche sia in termini di campo di applicazione.

Un primo capitolo è dedicato, a premessa, ai metodi di stima della numerosità delle popolazioni, tematica di vivo interesse sia teorico ma soprattutto applicativo. Segue un'analisi sotto l'aspetto dapprima storico e successivamente operativo del modello di cattura-ricattura, dove particolare rilievo è dato alle ipotesi sottostanti.
Attenzione viene poi rivolta alla versione generalizzata del modello, con accenni alle diverse soluzioni presentate dai vari ricercatori, ma soprattutto rileva l'utilizzo dei *modeli log-lineari*.
La trattazione termina con l'analisi di un caso contreto di utilizzo di questo metodo, fornendo così una visione più completa e non solo teorica dell'argomento.

Capitolo 1

La necessità di metodi per la stima per la numerosità delle popolazioni

Attualità e centralità della Statistica

Nel mondo complesso e variegato che l'umanità si trova a dover affrontare quotidianamente, una disciplina di cui certamente si sentirebbe la mancanza qualora non esistesse è la **Statistica**. La Statistica è infatti quella scienza che studia il modo in cui raccogliere, interpretare e rappresentare i dati che riguardano i fenomeni che ci circondano, provando a sintetizzarli e comprenderli, per poterne successivamente estrarre delle previsioni.

Proprio perché scienza interpretativa e in un certo senso predittiva, la Statistica è da sempre impiegata in diversi campi dell'agire umano, dove fornisce un supporto qualitativo e soprattutto quantitativo allo studio dei diversi fenomeni. La troviamo ampiamente utilizzata ad esempio

- in ambito economico, dove fornisce strumenti per l'analisi di fenomeni per l'appunto economici, per la verifica di modelli di previsione e di comportamento, per l'analisi del mercato o del territorio, per lo studio del modello aziendale;

- in campo fisico e chimico, dove numerosi processi, riguardanti reazioni nucleari, hanno natura stocastica e sono studiati all'interno della teoria della Fisica Statistica;

- in psicologia, in particolare in psicometria, per la misura delle abilità di una persona, delle caratteristiche dei suoi comportamenti e della sua personalità;

- in ambito sociologico, per la comprensione delle dinamiche socioeconomiche che intercorrono nella popolazione, come pure per sondaggi e ricerche di mercato;

- in ambiente medico, dove la biostatistica permette tra l'altro di studiare l'incidenza di una determinata malattia o l'efficacia di una nuova cura.

Tra i vari campi di studio in cui si è sviluppata la Statistica troviamo anche quello dedicato alla **stima della grandezza di una popolazione** (umana e non) che presenta specifiche caratteristiche oggetto di studio. Questa è un'applicazione di non scarsa importanza, basti pensare alla necessità di organizzare e strutturare una determinata azione od optare per una politica dedicata per l'appunto ad una data popolazione, che dipenderà necessariamente dalla sua grandezza. È inoltre un ambito che anche storicamente ha sempre interessato l'uomo, il quale si è cimentato periodicamente in censimenti.
Un piano di rilievo è occupato poi dallo studio di quelle popolazioni che vengono definite *rare* o *elusive*.

Le popolazioni rare e la stima della loro numerosità

Le popolazioni rare e/o elusive sono quelle popolazioni per le quali si riscontrano difficoltà di identificazione e di conteggio: non se ne conosce la dislocazione, né la composizione, né la dimensione. Classici esempi di tali popolazioni sono l'insieme degli immigrati senza regolare permesso, non volendo questi ovviamente comparire in liste che potrebbero identificarli, o dei senzatetto; lo sono pure le persone affette da malattie rare e particolari, o ancora le specie animali in via di estinzione.
Si tratta quindi di riuscire a dare una stima della grandezza di questo insieme, sfruttando gli elementi in possesso: classicamente, occorre riuscire a dedurre da un numero scarso di liste, spesso incomplete e sovrapposte, oppure da elenchi ampi e sovrabbondanti, il numero complessivo di membri della popolazione in esame.

Dal momento che il conteggio di una popolazione, tanto più se del tipo appena enunciato, è un problema sentito e di vivo interesse, sono state ideate numerose tecniche per cercare di risolvere tale questione. Tra le principali, possiamo ricordare il *campionamento per centri*, il cosiddetto *snowballing* e il metodo di *cattura-ricattura*.

Il **campionamento per centri** è stato introdotto principalemente per analizzare dati quali quelli riferiti alla popolazione di immigrati presenti in una comunità. Ne esistono tuttavia anche altre applicazioni, come la stima delle presenze turistiche.

Si procede in questo modo: volendo stimare il numero N di unità con una data caratteristica, all'interno di una popolazione P, si considerano L gruppi, detti *centri*, che siano punto di aggregazione delle N unità. I centri potranno quindi essere ad esempio luoghi di svago, di culto o di assistenza sociale. Da ogni centro viene poi estratto casualmente un numero prefissato di elementi, su cui ci si accerta se è soddisfatta o meno la caratteristica che stiamo cercando e da cui si parte per fare inferenza sull'intera popolazione.

La tecnica dello **snowballing** è utilizzata ogni qual volta si possa assumere che i membri della popolazione in esame si conoscano l'un l'altro, come nel caso in cui oggetto di studio sono alcune minoranze etniche o religiose, i consumatori di sostanze stupefacenti o le persone con particolari malattie. Lo snowballing prevede la creazione di una lista di elementi della popolazione con la data caratteristica a partire da un campione iniziale, preidentificato: dalle unità di questo campione, ad esempio tramite un questionario, si viene a conoscenza di altri individui che presentano la caratteristica da studiare, i quali a loro volta potranno fornire i nomi di altre persone da contattare e inserire nell'elenco. Eventuali problemi ovviamente sorgono qualora ci siano individui emarginati dal gruppo, che quindi non vengono segnalati e contattati e che portano allora ad una distorsione nella stima.

Il metodo di **cattura-ricattura** fu utilizzato inizialmente per le popolazioni animali e solo successivamente applicato allo studio di caratteristiche e comportamenti umani. È questo il modello che andremo ad enunciare e analizzare in questa trattazione.

Capitolo 2

Il metodo di cattura-ricattura

Breve introduzione e sviluppo storico

Il modello di cattura-ricattura, nella letteratura indicato come *capture recapture method* o con altri nomi simili quali *mark-recapture method* e *dual-system method*, gode di una lunga storia.

Il primo esempio giunto fino a noi di applicazione di tale metodo risale al 1802, quando Pierre S. Laplace (matematico e fisico della Francia napoleonica) lo utilizzò per stimare la grandezza della popolazione umana vivente in Francia, utilizzando i dati delle nascite raccolti a livello nazionale e i dati sul censimento e sulle nascite catalogati da diverse comunità locali.

La data che tuttavia segna davvero la nascita di questo modello è il 1896, quando il biologo danese Carl G. J. Petersen (1860-1928) pubblica *The Yearly Immigration of Young Plaice Into the Limfjord From the German Sea* [27], un articolo in cui suggeriva l'applicazione di tale approccio per la stima della numerosità di una popolazione di pesci (platesse, in particolare). Al suo lavoro fece seguito dopo qualche decennio un altro articolo [23], questa volta di Frederick C. Lincoln (1892-1960) sulla stima della popolazione di uccelli acquatici.
Da allora, il modello è stato notevolmente sviluppato, diventando sempre più usato nell'ecologia, ogni qualvolta si voleva stimare la grandezza di una determinata specie animale in un ambiente. Successivamente, a partire soprattutto dagli anni cinquanta del secolo scorso, la strategia del cattura-ricattura ha avuto nuovo vigore nell'ambito delle scienze mediche e sociali, in particolar modo nell'epidemiologia, dove è spesso importante avere un'idea quanto più accurata della grandezza della popolazione affetta da determinate malattie.

Il modello base e le ipotesi da soddisfare

La versione più semplice del modello di cattura-ricattura è il cosiddetto *Lincoln-Petersen model*, sviluppato per l'appunto da Petersen e da Lincoln (indipendentemente) e che prevede l'uso dei dati presenti in due sole liste, parzialmente sovrapposte, per dare una stima della grandezza complessiva della popolazione.

Il **metodo** ideato dai due ricercatori è in realtà piuttosto semplice. Una volta individuata la popolazione che vogliamo stimare, si installano delle trappole o delle reti in modo da campionare una prima volta la popolazione, marcando ogni esemplare trovato e compilando una prima lista. Successivamente, una volta che l'ambiente in esame è tornato allo stato naturale, si effettua un secondo campionamento, andando a registrare in modo particolare quanti esemplari appartenevano anche alla prima lista e quanti no.
Abbiamo dunque effettuato un procedimento di *cattura*, *marcatura* e *ricattura*, da cui il nome del modello, che ci permette di valutare il numero complessivo di individui nella popolazione: infatti il numero di individui trovati anche nel secondo campionamento è proporzionale al numero di individui marcati sull'intera popolazione, fatto da cui possiamo facilmente ricavare uno stimatore.

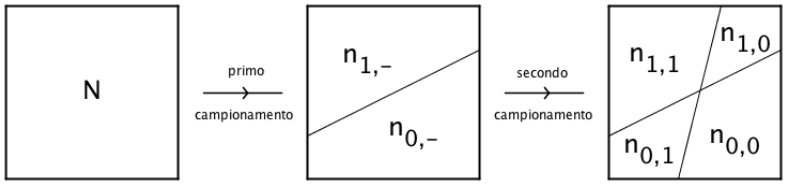

Figura 2.1: funzionamento del modello di cattura-ricattura

Prima di procedere con la ricerca dello stimatore è conviente soffermarsi però sulle **ipotesi** che devono essere verificate per il buon funzionamento del modello.

1. Occorre innanzitutto che la popolazione in esame sia una **popolazione chiusa** (*assumption of closure*): assumiamo dunque che non ci siano nascite, morti o migrazioni durante l'analisi, per cui il numero di individui rimane costante in grandezza e composizione durante lo studio.
   Si tratta comunque di una ipotesi solitamente facilmente verificata nei casi concreti, soprattutto se non si considerano tra i due campionamenti intervalli temporali troppo ampi.

2. In secondo luogo abbiamo bisogno di una **marcatura efficiente** e, soprattutto nelle applicazioni in ambito ecologico, non invadente. In

altre parole occorre che, una volta segnato l'individuo, la marcatura permanga fino al secondo campionamento della popolazione, per poter così effettuare il *matching* tra le liste, e che non impedisca al soggetto il normale svolgersi delle sue attività.
Nel caso questo non sia verificato avremo una riduzione del numero di individui ricatturati, che porterà ad una perdita di accoppiamenti tra le due liste e quindi a una sovrastima della grandezza della popolazione. Tuttavia questa ipotesi, come la precedente, non presenta solitamente grandi problemi né per le applicazioni ecologiche, dove esistono diversi metodi che vanno dalla colorazione all'installazione di appositi bracciali, né per quelle in campo umano, dove la terna nome, cognome e data di nascita identifica quasi univocamente gran parte della popolazione.

3. Necessitiamo poi che tutti i soggetti abbiano, per ogni singolo campionamento, la **medesima probabilità di essere inclusi nella lista** (*equal-catchability*). Occorre notare però che questo non corrisponde al fatto che la probabilità di trovarsi nella prima lista e la probabilità di trovarsi nella seconda sia la medesima, bensì che i componenti della popolazione siano *omogenei* e quindi ugualmente rilevabili da ogni singolo metodo di campionamento.
A differenza delle prime due ipotesi, questa richiesta non è facilmente soddisfatta, soprattutto quando si affronta l'ambito medico-sociale. Si intuisce subito che non tutti i malati hanno ad esempio la stessa probabilità di essere rilevati da una determinata indagine, così come non tutti i tossicodipendenti godono di omogeneità rispetto a un dato metodo campionario. Per ovviare a questo problema si fa spesso uso della *stratigrafia*, introducendo una partizione sullo spazio campionario Ω in base ad alcune qualità comuni: si stima quindi N separatamente per ciascuno strato creato, sommando infine i risultati ottenuti. Questo metodo, seppur riesca spesso ad abbattere l'eterogeneità originaria, riduce il numero di individui su cui applicare di volta in volta il modello di cattura e matching, facendogli perdere sensibilità. La stratigrafia, che viene ad esempio discussa da Sekar e Deming in un articolo per il *Journal of the American Statistical Association* [31], non è ad ogni modo l'unica soluzione: varie alternative sono state proposte, come ad esempio l'introduzione di coefficienti correttivi oppure l'uso del modello di regressione logistica.

4. L'ultima ipotesi di cui abbiamo bisogno è l'**indipendenza** tra le liste: il fatto che un individuo x della popolazione compaia nella lista i non deve influenzare il fatto che compaia o meno nella lista j, ossia

$$\mathbb{P}[x \in j \,|\, x \in i] = \mathbb{P}[x \in j] \quad \forall x \, \forall i \, \forall j$$

Questa ipotesi è in realtà una diretta conseguenza di quella al paragrafo precedente, dal momento che l'equicatturabilità implica che i soggetti marcati e i non marcati hanno la stessa probabilità di essere ricatturati nel secondo campionamento, per cui la prima cattura (cioè l'essere nella prima lista) non influenza la seconda. È tuttavia conveniente tenerla separata, sia per darle maggior risalto sia per l'importanza e la criticità che vedremo avere nel modello. Inoltre, proprio come la richiesta precedente da cui deriva, anche l'indipendenza presenta alcune difficoltà al momento dell'applicazione pratica, soprattutto nel campo epidemiologico: basti pensare al medico che tiene un suo elenco di pazienti e che li indirizza spesso verso la medesima struttura, creando così una correlazione tra la sua lista e quella dell'ospedale in questione.

lista B	lista A		totale
	sì	no	
sì	$n_{1,1}$	$n_{0,1}$	$n_{\_,1}$
no	$n_{1,0}$	$n_{0,0}$	$n_{\_,0}$
totale	$n_{1,\_}$	$n_{0,\_}$	

Tabella 2.1: esempio di tabella di contingenza nel caso di due liste

Un modo per evidenziare eventuali relazioni di dipendenza tra le liste viene suggerito nella letteratura da Hook e Regal in [20] e in [2] e sfrutta il cosidetto *odds ratio*. Una volta rappresentati i dati dei due campionamenti in una tabella di contingenza (*contingency table*), come è ad esempio la Tabella 2.1, è possibile calcolare la probabilità di successo p e conseguentemente l'*odd* di ogni riga della tabella, sfruttando

$$p_I = n_{1,1}/n_{\_,1} \quad odd_I = p_I/(1-p_I)$$

$$p_{II} = n_{1,0}/n_{\_,0} \quad odd_{II} = p_{II}/(1-p_{II})$$

L'odd che ne risulta è un numero non negativo, con valore maggiore di 1 qualora il successo sia più probabile del fallimento e minore di 1 in caso contrario.
A questo punto è possibile calcolare l'*odds ratio*, ossia il rapporto tra i due odds, dato appunto da

$$\theta = \frac{odds_I}{odds_{II}} = \frac{p_I/(1-p_I)}{p_{II}/(1-_{II})} = \frac{n_{1,1}n_{0,0}}{n_{1,0}n_{0,1}}$$

Maggiori informazioni sull'odds ratio si possono trovare nel Glossario finale o in [2]. Quello che qui ci interessa sapere è che qualora l'odds ratio sia

uguale a 1 si ha l'indipendenza tra le due liste, mentre più ci si discosta da tale valore più forte è la relazione tra di esse. Tuttavia, una volta accertata la dipendenza tra due liste – cosa che come abbiamo già sottolineato risulta tutt'altro che rara, rimane la necessità di ovviare a tale problema: una possibile soluzione fortunatamente c'è e fa uso dei *modelli log-lineari*, di cui discuteremo in seguito, nel caso più generale del modello di cattura-ricattura con un numero arbitrario di liste.

Gli stimatori di Lincoln-Petersen, di Chao e di Chapman-Seber

Ci apprestiamo ora ad introdurre lo stimatore più semplice, anche noto come *stimatore di Lincoln-Petersen*, per la grandezza della popolazione a seguito di un campionamento di cattura-ricattura con due liste. Riprendiamo a tal proposito la notazione usata nella Tabella 2.1, denotando quindi con $n_{1,0}$ il numero di individui presenti nella lista A ma non nella lista B; per semplicità, rinominiamo $n_{1,\_} =: n_A$ e $n_{\_,1} =: n_B$ il numero di elementi rispettivamente della prima e della seconda lista. Indicheremo inoltre con $\mathbb{E}[\,\cdot\,]$ il valore di aspettazione (o speranza matematica) di una data variabile casuale.

Assumiamo innanzitutto che tutte e quattro le sopracitate ipotesi siano soddisfatte: siamo quindi in presenza di una popolazione chiusa e omogenea al campionamento, in cui abbiamo usato una marcatura efficiente e da cui abbiamo ricavato due liste A e B indipendenti tra loro. I valori $n_{i,j}$ della tabella di contingenza altro non sono che delle variabili aleatorie che seguono una distribuzione multinomiale di parametri $p_{i,j}$, con questi parametri tali che $\sum_{i,j} p_{i,j} = 1$.
A questo punto se supponiamo che il rapporto tra il numero di elementi della lista A e il numero di elementi totali sia sostanzialmente uguale al rapporto ristretto alla lista B, fatto che discende dall'omogeneità richiesta dalla terza assunzione fatta, abbiamo immediatamente che

$$\frac{n_A}{N} = \frac{n_{1,1}}{n_B} \quad \rightarrow \quad \hat{N} = \frac{n_A n_B}{n_{1,1}}$$

cioè che uno stimatore per la grandezza della popolazione è dato dalla grandezza della prima lista moltiplicata per un fattore di correzione.

Lo stesso risultato è per di più ottenuto concentrandosi sull'ipotesi di indipendenza: sotto questa proprietà per un dato soggetto x abbiamo che

$$\mathbb{P}[x \in B \,|\, x \in A] = \mathbb{P}[x \in B \,|\, x \notin A]$$

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

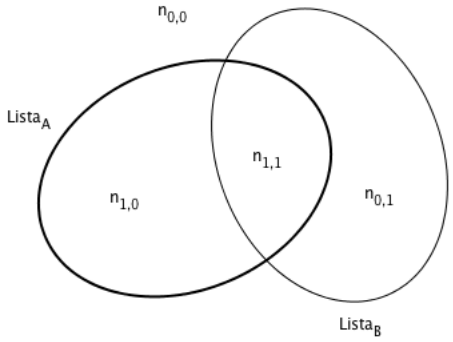

Figura 2.2: rappresentazione grafica della tabella di contingenza 2.1

da cui abbiamo uno stimatore del valore ignoto del numero di elementi non rilevati da nessun campionamento $\hat{n}_{0,0} = \frac{n_{0,1}n_{1,0}}{n_{1,1}}$ e quindi nuovamente lo **stimatore (o indice) di Lincoln-Petersen**

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

Valutiamo a questo punto se lo stimatore trovato è soddisfacente: ci aspettiamo infatti che per grandi valori del campione l'indice di Petersen ci fornisca una risposta pressoché corretta. Ricordando quanto definito all'inizio della sezione abbiamo immediatamente che

$$\mathbb{E}[n_A] = Np_A = N(p_{1,1} + p_{1,0}) \qquad \mathbb{E}[n_B] = Np_B = N(p_{1,1} + p_{0,1})$$

$$\mathbb{E}[n_{1,1}] = Np_{A,B} = Np_{1,1}$$

relazioni che possiamo usare ora per studiare il comportamento asintotico di $\mathbb{E}[\hat{N}]$. Per grandi valori di N abbiamo d'altra parte che il valore di aspettazione di una funzione di variabili casuali è asintoticamente uguale alla funzione dei valori di aspettazione, per cui

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

A questo punto, sotto l'ipotesi di indipendenza delle due liste, abbiamo banalmente che $p_{A,B} = p_A p_B$, per cui $\mathbb{E}[\hat{N}] = N$ e dunque lo stimatore è asintoticamente non distorto, come speravamo.

L'espressione dello stimatore di Petersen ci fornisce anche un'idea sugli effetti che la non-indipendenza tra le liste crea sul risultato fornito: in questo caso non ci è purtroppo possibile scrivere $p_{A,B} = p_A p_B$ e quindi procedere come visto sopra. Possiamo tuttavia usare la relazione sulla probabilità

condizionata $p_{A,B} = p_{B|A}p_A$, che ci permette di riscrivere $\mathbb{E}[\hat{N}] = \frac{Np_B}{p_{B|A}}$. Risulta a questo punto chiaro che qualora essere nella lista A aumenti la probabilità di essere anche nella lista B, avremo che $p_{B|A} > p_B$ e quindi $\mathbb{E}[\hat{N}] < N$, cioè $\hat{N}$ sottostimerà il reale valore di N. Considerazioni analoghe si possono fare nel caso in cui $p_{B|A} < p_B$ o in cui sia la lista B a condizionare la presenza o meno di elementi nella lista A.

Si può dimostrare, tramite le proprietà della distribuzione multinomiale o dell'ipergeometrica come proposto ad esempio da Chapman in [10], che l'indice di Petersen, per quanto asintoticamente non distorto, è distorto per piccoli valori di N. Questo, insieme ad altri problemi sorti nel corso degli anni durante le varie applicazioni del modello di cattura-ricattura, hanno portato all'introduzione di altri stimatori che ne correggessero le storture. Due che vale sicuramente la pena citare, anche se non saranno trattati nello specifico, sono lo **stimatore di Chapman-Seber** e lo **stimatore di Chao**. Il primo, introdotto per ovviare al rischio del denominatore nullo e al problema della distorsione del sopracitato stimatore di Petersen, stima il valore di N tramite la formula

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

e presenta una varianza pari a

$$\sigma^2_{Chapman} = \frac{(n_A + 1)(n_B + 1)(n_A - n_{1,1})(n_B - n_{1,1})}{(n_{1,1} + 1)^2(n_{1,1} + 2)^2}$$

Infine, lo stimatore di Chao [8] è stato introdotto espressamente per superare l'inconveniente della possibile non validità dell'ipotesi sull'omogeneità della popolazione al campionamento e segue l'espressione

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

con varianza

$$\sigma^2_{Chao} = \frac{(n_A + n_B)^2(n_A + n_B - 2n_{1,1})^2}{16n^3_{1,1}}$$

Capitolo 3

L'evoluzione del modello

Un ampliamento dei propri orizzonti...

Nel corso del XX secolo il modello di cattura-ricattura ha avuto importanti sviluppi. Abbiamo visto nel capitolo precedente come fosse originariamente un metodo con applicazioni in ambito ecologico di conta del numero di esemplari di una determinata specie, un modo per poter stimare la grandezza della popolazione in base alla sovrapposizione di due singoli campionamenti parziali. Ben presto il modello si rivelò però più generale e di ben più ampio respiro.

Già nel 1938, ovvero solo dopo otto anni dalla pubblicazione dell'articolo di Petersen, la studiosa Zoe E. Schnabel pubblica un articolo [29] in cui introduce il **modello di cattura-ricattura a k-liste**, una generalizzazione di quello finora usato. Partendo sempre dalle quattro ipotesi enunciate nelle pagine precedenti, Schnabel ebbe l'idea di considerare più di due campionamenti sulla popolazione, così da avere una cronistoria delle catture di ogni singolo soggetto della popolazione e poter utilizzare queste maggiori informazioni per meglio stimare il numero di individui totale.

La vera estensione dell'area di ricerca in cui poter applicare il modello di cattura-ricattura si ha però nel 1949 quando C. Chandra Sekar e W. Edwards Deming [31] utilizzano questa metodologia per studiare i tassi di natalità e di mortalità e la percentuale di copertura del corrispettivo registro in una zona vicino a Calcutta, in India. Questo lavoro segna infatti l'inizio dell'applicazione del modello al campo delle **scienze umane, mediche e sociali**: il cattura-ricattura può essere infatti utilizzato, in linea di principio, in ogni situazione dove vi siano delle liste incomplete. Basta quindi osservare come «essere nel campione A» sia totalmente analogo a «essere nell'elenco A» per poter comprendere come il metodo finora descritto sia integrabile in realtà quali il censimento della popolazione di uno Stato oppure la stima del numero di soggetti affetti da una data malattia.
È stato in particolare proprio nell'epidemiologia che si sono avuti maggiori

esempi di utilizzo, come nel calcolo del numero di infetti dalla febbre dengue [28], dell'incidenza del cancro tra le forze armate [26] o dell'ampiezza di un'epidemia di epatite A [9]. Lavori interessati hanno riguardato anche la stima del numero di persone rimaste uccise durante la guerra in Kosovo [4], la quantità di tossicodipendenti in alcune città francesi [33] o il numero di bambini di strada in brasile [18]. Maggiori casi si possono ad ogni modo ritrovare nella Bibliografia allegata, in particolare nella rassegna presentata dall'*International Working Group for Disease Monitoring and Forecasting* in [16], mentre nell'ultimo capitolo verrà affrontato come esempio specifico la stima del numero effettivo di diabetici nell'area di Casale Monferrato.

... che presenta però qualche difficoltà

L'introduzione di tutti questi cambiamenti ha portato anche alla necessità di nuovi sistemi e tecniche per affrontare il problema, metodi ora spesso nominati con *multiple-record system*, soprattutto per la gestione di un numero di liste superiore alle due proposte da Lincoln e da Petersen e per il controllo dell'ipotesi sottostanti. Nel passaggio dal campo ecologico a quello medico-sociale sorgono infatti alcuni **problemi**.

Innanzitutto perdiamo nella maggioranza dei casi quell'ordine temporale "naturale" che caratterizzava i campionamenti degli animali e che invece è difficilmente presente nelle liste epidemiologiche.
Ma ciò che malauguratamente non vale più, come già accennato in precedenza, è la validità delle ipotesi di omogeneità e soprattutto di indipendenza, assunzioni che abbiamo visto essere alla base del metodo con cui abbiamo ricavato gli stimatori precedenti. D'altra parte i pazienti tendono a essere eterogenei circa la possibilità di essere segnalati in un dato registro, senza contare che data la struttura della nostra società e del sistema assistenzialistico, è quasi certo che la presenza di un soggetto in una lista ne influenzi la presenza in un'altra.
Occorre inoltre un modo per tenere conto del crescere del numero di liste, che solitamente in questo campo di indagine provengono da due a quattro fonti diverse (un numero superiore infatti comporta nella maggioranza dei casi più costi rispetto ai benefici prodotti).

Da questo punto la letteratura sull'argomento si divide ed inizia a prendere strade diverse.

Da un lato J. T. Wittes propone nel 1974 sul *Journal of Chronic Diseases* di confrontare **due liste per volta**, riducendosi in sostanza al caso noto di Lincoln-Petersen: nel momento in cui le ipotesi sono soddisfatte dovremmo infatti ottenere dei risultati molto simili nell'analisi di ciascuna delle $\binom{n}{2}$ coppie di liste. Questa strategia permette inoltre di evidenziare l'eventuale dipendenza fra le liste, caso in cui avremo che la stima della grandezza N della popolazione ottenuta dalla coppia di liste (i, j) sarà sensibilmente

diversa alla stima ottenuta dalle altre coppie. Tuttavia questo *two-samples method* è difficilmente applicabile nell'epidemiologia causa la presenza dei sopracitati problemi, motivo per cui rimane un modello utilizzabile più in singole specifiche situazioni che a livello generale.

Altri ricercatori come K. H. Pollock, D. L. Otis e G. C. White suggeriscono invece di adottare, in particolar modo per il settore ecologico, una famiglia di modelli basati sulla regressione logistica e sulla funzione **logit**$[\cdot]$, che corrisponde al logaritmo naturale dell'odd.
Altri ancora, come ricorda Anne Chao in [9], propongono di utilizzare il **Sample Coverage Approach**, un metodo che nasce espressamente per avere una misura della sovrapposizione reciproca delle varie liste e per quantificare la possibile dipendenza tra di esse. L'idea sottostante, che risale ancora a Turing e Good (1953), è che la stima della grandezza della popolazione possa essere ricavata dalle relazioni che ci sono tra il numero di elementi della popolazione e il *sample coverage*, la copertura del campione. Questa strategia presenta alcuni punti di forza come il poter misurare tramite il coefficiente di covarianza (CCV) la dipendenza tra i campioni, la facile generalizzazione al caso con k-liste e il non dover scegliere tra una famiglia di modelli.

Infine Stephen E. Fienberg propone su *Biometrika* nel 1972 di utilizzare la famiglia dei **modelli log-lineari** per risolvere il problema della dipendenza tra le liste, sfruttando la teoria dell'analisi delle tabelle di contingenza multidimensionali. Una soluzione che risulta essere generale e facilmente applicabile a tutti i campi d'indagine in cui è possibile sfruttare il metodo di cattura-ricattura.

I modelli log-lineari

I modelli log-lineari fanno parte dei *modelli lineari generalizzati* (GLM), una famiglia di modelli che estendono quello classico della regressione lineare. A differenza del modello lineare, in cui le variabili casuali si suppongono distribuite secondo una distribuzione normale, la versione generalizzata si disfa di questa necessità, per essere così applicabile ad ogni variabile casuale appartenente ad una famiglia esponenziale: oltre alla gaussiana la troviamo dunque applicata anche in contesti di distribuzione bi- o multinomiale, poissonia, gamma, ecc.

Il principale utilizzo dei modelli log-lineari si ha nello **studio delle tabelle di contingenza**, in particolare nella modellizzazione dei valori delle sue celle. La sua forza sta nel poter essere applicabile a ogni generica tabella, qualsiasi sia la sua dimensione e le interdipendenze dei valori al suo interno. Daremo qui di seguito una breve introduzione alla teoria di questi

modelli, studiando in particolare il caso semplice delle tabelle 2×2 da cui è tuttavia facilmente desumibile il caso generale n-dimensionale.

Consideriamo dunque una generica tabella di contingenza 2×2 che classifica n soggetti secondo le categorie $cat_{\_\,\_}$, relative a due variabili categoriche A e B. Per il momento, a differenza del caso della Tabella 2.1, supponiamo che tutti i valori n_{ij} siano a noi noti.

variabile B	variabile A		totale
	cat_{A1}	cat_{A2}	
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

Tabella 3.1: esempio di tabella di contingenza 2×2

Osserviamo innanzitutto come, sotto l'ipotesi di indipendenza, per ogni scelta di $i, j = 1, 2$ la probabilità $p_{i,j}$ di ciascuna cella sarà data dal prodotto delle probabilità marginali $p_{\_,j}$ e $p_{i,\_}$, ossia

$$p_{i,j} = p_{\_,j} \cdot p_{i,\_}$$

Come già osservato in precedenza, le probabilità $p_{i,j}$ costituiscono i parametri di una distribuzione multinomiale, motivo per cui $\sum_{i,j} p_{i,j} = 1$. Occorre notare come si potrebbe usare, come alcuni autori fanno, la distribuzione poissoniana al posto di quella multinomiale; preferiamo qui tuttavia usare quest'ultima, essendo la poissoniana più adatta al caso di eventi rari.
Considerate ora le frequenze attese

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

possiamo considerarne il logaritmo naturale per avere dunque l'espressione, ora additiva e non più moltiplicativa,

$$\log[\mu_{i,j}] = \log[n] + \log[p_{i,\_}] + \log[p_{\_,j}] = \lambda + \lambda_i^A + \lambda_j^B \qquad i, j = 1, 2$$

che identifica il **modello log-lineare di indipendenza** per una tabella 2×2. In particolare λ_i^A rappresenta l'effetto della variabile A classificata come i e analogo discorso si può fare per λ_j^B; il termine λ corrisponde invece ad un "fattore di scala" per la tabella.

Nel caso tuttavia in cui le variabili a cui si riferisce la tabella non siano indipendenti è necessario ricorrere ad una versione più generale del modello, chiamato **modello log-lineare saturo**, che tenga conto di un fattore di correzione λ_{ij}^{AB} che rappresenti l'interazione tra le due variabili A e B:

$$\log[\mu_{i,j}] = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \qquad i, j = 1, 2$$

Il caso precedente di indipendenza si può far semplicemente derivare da quest'ultimo ponendo $\lambda_{ij}^{AB} = 0$.
Possiamo inoltre osservare come esista una diretta relazione tra il parametro λ_{ij}^{AB} e l'odds ratio, che avevamo già visto a pag. 15 essere una misura dell'indipendenza delle variabili indicate nella tabella. Calcolando direttamente il logaritmo dell'odds ratio si ottiene infatti che

$$\begin{aligned}\log[odds\,ratio] &= \log\Big[\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}\Big] = \log[\mu_{11}] + \log[\mu_{22}] - \log[\mu_{12}] - \log[\mu_{21}] \\ &= (\lambda + \lambda_1^A + \lambda_1^B + \lambda_{11}^{AB}) + (\lambda + \lambda_2^A + \lambda_2^B + \lambda_{22}^{AB}) \\ &\quad - (\lambda + \lambda_1^A + \lambda_2^B + \lambda_{12}^{AB}) - (\lambda + \lambda_2^A + \lambda_1^B + \lambda_{21}^{AB}) \\ &= \lambda_{11}^{AB} + \lambda_{22}^{AB} - \lambda_{12}^{AB} - \lambda_{21}^{AB}\end{aligned}$$

da cui qualora $\lambda_{ij}^{AB} = 0\ \forall i, j$ avremo che il logaritmo dell'odds ratio sarà nullo, ossia $odds\,ratio = 1$ e le variabili categoriche A e B sono indipendenti.

Quanto visto finora può essere facilmente generalizzato a **tabelle n-dimensionali**, semplicemente inserendo termini per ogni variabile aggiunta e per ogni possibile relazione con le altre variabili. Ad esempio, per una tabella $2 \times 2 \times 2$ (una *three-way table*) il modello saturo sarà della forma

$$\log[\mu_{ijk}] = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC} \quad i, j, k = 1, 2$$

dove possiamo osservare esserci le medesime componenti λ e $\lambda_{\cdots}^{\cdots}$ del caso 2×2 più la componente λ_{ijk}^{ABC} che rappresenta la interdipendenza congiunta tra le tre variabili categoriche in esame.

Si tratta quindi a questo punto di valutare i parametri $\lambda_{\cdots}^{\cdots}$ che determinano **il modello migliore**. Un modello che però sia contemporaneamente anche il più parsimonioso possibile in termini di numero di parametri $\lambda_{\cdots}^{\cdots}$ utilizzati: se da una parte il modello saturo si adatta perfettamente alle frequenze delle celle della tabella, d'altro canto può contenere un numero spropositato di termini costitutivi, spesso ridondanti. Obiettivo del ricercatore è dunque quello di trovare modelli più semplici di quello saturo, rinunciando ad alcuni parametri ma non perdendo di vista il risultato finale, che deve comunque essere non significativamente diverso dai dati raccolti.

Come spiegano ad esempio G. Chiari e P. Peri [11], si tratta di un lavoro che per tabelle di contingenza 2×2 è eseguibile anche "manualmente", utilizzando gli odds della tabella, distribuzioni marginali e medie, ma che già per le tabelle a tre entrate richiede l'uso di sistemi di calcolo automatici e di algoritmi iterativi, come quello di *Stephan-Deming* o quello di *Newton Raphson*.
Che si possa agire manualmente o che si faccia uso di software specifici, ciò che viene fatto è analizzare i possibili valori di $\lambda_{\cdots}^{\cdots}$, ponendoli eventualmente uguale a zero se ininfluenti, e testare quanto ogni modello si adatti alle

frequenze osservate nella tabella. Quest'ultimo passaggio in particolare è svolto solitamente attraverso l'utilizzo di due **test** per la verifica delle ipotesi H_0 contro H_1: il *test chi-quadrato* o *test di Pearson* χ^2 e il *rapporto di verosimiglianza* L^2 o G^2, che seguono l'espressione:

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \qquad L^2 = 2 \sum_{i,j} n_{ij} \log\Big[\frac{n_{ij}}{\hat{\mu}_{ij}}\Big]$$

Questi due indicatori, comuni nella pratica statistica anche al di fuori del problema dei modelli loglineari e delle tabelle di contingenza, stimano il cosiddetto *goodness of fit* del modello rispetto alla realtà, che è tanto maggiore quanto più χ^2 o L^2 si avvicina a 0. Al contrario, quanto più l'indicatore si discosta dallo 0 si ha una contraddizione dell'ipotesi H_0 che si voleva testare e dunque il modello proposto va rifiutato in favore di altri.

L'utilizzo dei modelli log-lineari nel metodo di cattura ricattura

Quanto appena introdotto si è rivelato fondamentale per l'estensione del modello di cattura-ricattura al caso più generale di k-liste con eventuale dipendenza. Abbiamo infatti una struttura – i modelli log-lineari – che si adatta a tabelle di contingenza di qualsiasi dimensione e che contiene parametri che giustificano ogni iterazione tra le variabili categoriche in esame.

Possiamo dunque partire dal modello di indipendenza per il caso a due liste A e B, che prevede $\mathbb{E}[n_{1,1}] = N p_A p_B$, e prendere il logaritmo di ambo i membri

$$\log\big[\mathbb{E}[n_{1,1}]\big] = \log[N] + \log[p_A] + \log[p_B]$$

a cui possiamo aggiungere un termine $\log[i_{A,B}]$ corrispondente all'interazione tra le due. In modo del tutto analogo possiamo ad esempio scrivere

$$\log\big[\mathbb{E}[n_{0,1}]\big] = \log[N] + \log[1 - p_A] + \log[p_B] + \mathit{interazione}$$

per arrivare così all'usuale formulazione del modello log-lineare saturo:

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

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

dove per comodità consideriamo $\log[1 - p_A] = -\lambda^A$, in modo che si elimino gran parte dei termini in caso di somma complessiva.

Il caso a più di due liste si generalizza con facilità, proprio come visto nel paragrafo precedente. Per tre liste A, B e C si hanno ad esempio otto diversi parametri: il termine comune λ, i termini di effetto principale λ^A, λ^B, λ^C, i parametri di interazione "a due" λ^{AB}, λ^{AC}, λ^{BC} e un parametro

di interazione congiunta λ^{ABC}.
Volendo stimare n_{000}, ci troviamo tuttavia con solo sette termini conosciuti all'interno della tabella, motivo per cui occorre fare delle ipotesi a priori sul valore di qualche parametro. Queste ipotesi riguardano solitamente il termine di interazione di grado più alto, che viene posto uguale a 0, e vengono comunque poi testati sui dati disponibili tramite l'applicazione del test chi-quadrato o del rapporto di verosimiglianza implementati in software statistici.
Una volta trovati i parametri $\lambda^{\cdots}$ che meglio si adattano ai dati conosciuti della tabella di contingenza in esame e che siano in numero minore possibile, abbiamo identificato il modello che meglio descrive la situazione e possiamo sfruttarlo per avere una stima del valore ignoto di n_{000} e dunque di N, come speravamo.

Aspetti teorici conclusivi

Prima di concludere l'esposizione del modello di cattura-ricattura generalizzato rimangono ancora due aspetti da trattare e definire: la scelta del modello da utilizzare e la costruzione di intervalli di confidenza.

Le tecniche di **selezione del modello** sono entrate nella teoria del metodo di cattura-ricattura soltanto di recente: se in origine infatti solo uno o due modelli erano disponibili ai ricercatori, il crescere dell'interesse sull'argomento ha portato ad un miglioramento della tecnica e allo sviluppo di nuove metodologie, quali quella del Sample Coverage Approach o quella log-lineare. Avendo dunque a disposizione almeno otto modelli diversi, come cita ad esempio l'*International Working Group for Disease Monitoring and Forecasting* in [15], occorre ai ricercatori un metodo per identificare il modello che meglio si adatta al caso in esame.
Al giorno d'oggi si utilizzano principalmente due indicatori, realizzati tramite il test di massima verosimiglianza. Esiste innanzitutto il **criterio di Akaike**, meglio noto come *Akaike Information Criterion* (AIC), sviluppato dal matematico Hirotsugu Akaike nel 1974 e che segue l'espressione

$$AIC = 2k - 2\log[L]$$

dove k è il numero di parametri del modello e L la funzione di verosimiglianza. Il criterio di Akaike non solo valuta quanto il modello si adatta alla situazione reale ma include nell'analisi anche il numero di parametri presenti, apprezzando maggiormente un modello parsimonioso e penalizzando quello con un numero eccessivo. In particolare, il modello con indice AIC più basso sarà un modello preferibile rispetto agli altri.

Il secondo metodo, ideato da Gideon E. Schwarz nel 1978, prende il nome di **criterio di Bayes**, o *Bayes Information Criterion* (BIC). Questa tecnica

che, come la precedente, prevede di scegliere il modello con indice più basso e che studia il numero di parametri k e la funzione di verosimiglianza L, incorpora in più la dimensione del campione n:

$$BIC = k(\log[n] - \log[2\pi]) - 2\log[L] \approx k\log[n] - 2\log[L]$$

Si può anche osservare come generalmente il criterio BIC penalizza maggiormente i modelli più grandi rispetto ad AIC.

Infine esiste il problema della costruzione di un **intervallo di confidenza** per il valore di $\hat{N}$, o equivalentemente di $\hat{n}_{00}$, trovato: senza tale intervallo non avremmo d'altra parte idea della precisione con cui stimiamo il numero totale di individui.
Storicamente la costruzione degli intervalli di confidenza avveniva assumendo che $\hat{N}$ fosse una variabile casuale con distribuzione asintotica normale, per cui si poteva applicare la formula standard

$$stima \pm (percentile \times S.E.) \quad \overset{95\%}{\rightarrow} \quad \hat{N} \pm 1,96\, S.E.$$

Questa procedura tuttavia funziona malamente in tutti quei modelli che presentano una distribuzione asimmetrica, proprio come lo sono ad esempio praticamente tutti quelli proposti finora nell'ambito del cattura-ricattura. Occorre pertanto pensare ad una qualche tecnica alternativa per costruire tali intervalli, come quella del *verosimiglianza profilo* o del *bootstrap* [13].

L'idea alla base della costruzione di intervalli di confidenza attraverso il **profile likelihood** è quella di invertire il test del rapporto di verosimiglianza per costruire l'intervallo in questione. Un intervallo di confidenza a $1-\alpha$ per il parametro N è infatti l'insieme di tutti i valori N_0 per cui il test bilaterale per l'ipotesi nulla $H_0\colon N = N_0$ non sarà rifiutato per un livello di significatività pari ad α. Considerata la distribuzione asintoticamente chi-quadrato del test del rapporto di verosimiglianza e posto $L_1(N) = \max_\delta L(N, \delta)$ la funzione di verosimiglianza profilo (in cui sostanzialmente alcuni parametri vengono espressi in funzione di altri, riducendone così il numero complessivo), abbiamo che H_0 non sarà rifiutata per tutti i valori N_0 che soddisfano

$$\log[L_1(N_0)] > \log[L(\hat{N}, \hat{\delta})] - \frac{1}{2}\chi^2_{1-\alpha}(1)$$

dove $\chi^2_{1-\alpha}(1)$ indica il quantile di una distribuzione χ^2 con 1 grado di libertà e $\hat{N}$ il valore stimato dal modello. L'insieme di tali valori N_0 costituirà dunque il nostro intervallo di confidenza cercato.

Il metodo **bootstrap** fu invece sviluppato dallo statistico Bradley Efron nel 1979 e sfrutta il ricampionamento (con reimissione) per generare un insieme di stime di N, da cui poi creare gli intervali di confidenza. Partendo

infatti da un campione $x = (x_1, \ldots, x_m)$ effettivamente osservato, da questo si possono estrarre con reimissione m valori x_i^* e considerare questi nuovi valori $x^* = (x_1^*, \ldots, x_m^*)$ come campione osservato su cui fare l'analisi statistica desiderata. Ripetendo questo procedimento più volte, avremo più di una stima e potremo calcolare da esse una media e una varianza *bootstrap*, valori che useremo poi per costruire l'intervallo di confidenza.
Tale procedimento, applicato al metodi di cattura-ricattura, prevede dunque di effettuare un ricampionamento tra i dati osservati, spesso nell'ordine di $400 - 1000$ volte, e per ognuna di queste situazioni generare la stima bootstrap $\hat{N}^*$. Una volta ottenuti questi valori, si può banalmente calcolarne la media e la varianza campionaria e, dopo aver riordinato in modo crescente i $\hat{N}^*$, costruire l'intervallo di confidenza prendendo il percentile desiderato.
L'unico vero problema che può sorgere riguarda la quantità di lavoro computazionale da svolgere: ciascuna situazione generata dal bootstrap è un nuovo campionamento su cui occorre effettuare da capo un'analisi secondo le procedure del cattura-ricattura, compresa la scelta del modello migliore e la stima dei parametri che lo identificano.

Capitolo 4

Una stima dell'incidenza del diabete

Affrontiamo in questo capitolo finale l'analisi di un caso pratico di applicazione delle metodologie viste finora; ci occuperemo in particolare della stima della quantità di popolazione affetta da diabete nel comune di Casale Monferrato, in provincia di Alessandria, Piemonte, sfruttando l'analisi effettuata Graziella Bruno e altri ricercatori in due articoli pubblicati nel 1992 e nel 1994 [6, 7].

Il **diabete mellito** corrisponde ad una malattia cronica che comporta un aumento della concentrazione di glucosio nel sangue, frequentemente causato da problemi nella produzione di un ormone indispensabile per il metabolismo degli zuccheri, l'insulina. Secondo un articolo recentemente pubblicato sulla nota rivista *The Lancet*, nel 2013 questo disturbo accomunava circa 382 milioni di persone nel mondo; uno studio del 2012 dell'*Organizzazione Mondiale per la Sanità* lo classificava come ottava causa di morte, con un milione e mezzo di decessi annuali a livello globale.

Date le dimensioni e le implicazioni del fenomeno, le autorità sovrannazionali hanno dato il via da diversi anni a studi sull'effettiva estensione del problema, in modo da poter adottare politiche adeguate. È proprio in questo contesto che si inserisce lo studio sopra citato *"Application of capture-recapture to count diabetes?"*. Da una parte la necessità di stimare l'incidenza del diabete sulla popolazione, dall'altra parte il problema dell'inaccuratezza o del costo eccessivo degli usuali metodi campionari: uno scenario che lascia ben sperare per una vantaggiosa applicazione del metodo di cattura-ricattura appena presentato.

Occorre innanzitutto **identificare due o più liste**, da cui partire per fare inferenza sul numero complessivo dei malati. Nei due articoli in esame vengono proposte quattro liste, parzialmente sovrapposte, contenenti i soggetti affetti da diabete mellito (tipo 1 o 2, indifferentemente) in data 1 ottobre 1987:

- l'elenco di tutti i pazienti con una diagnosi di diabete mellito fornita dai pediatri, dai medici di famiglia e dalle cliniche per il diabete dell'area in questione, raccolti in modo organizzato e standardizzato dal gruppo di ricerca (lista A)
- il registro di tutte le persone con una diagnosi di diabete mellito dimesse da qualsiasi ospedale pubblico o privato situato in Piemonte (lista B)
- il database informatico contenente le prescrizioni di insulina e ipoglicemizzanti orali, principali metodi di trattamento in caso di diabete, fornito dal sistema sanitario nazionale (lista C)
- la lista di tutti i residenti di Casale Monferrato che hanno fatto richiesta di rimborso per siringhe o strisce reagenti *reagent strips* (lista D).

Questi quattro elenchi sono stati vagliati dai ricercatori, che hanno eliminato i casi che non erano validi nella data di riferimento scelta, accertato che si trattasse del disturbo in esame e rintracciato le cartelle cliniche degli eventuali deceduti o trasferiti. Sono poi stati riconosciuti i soggetti presenti in più liste, sono stati collegati tra loro e si è proceduto infine a costruire la tabella di contingenza 4.1.

		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	-

Tabella 4.1: tabella di contingenza sul campionamento effettuato [7]

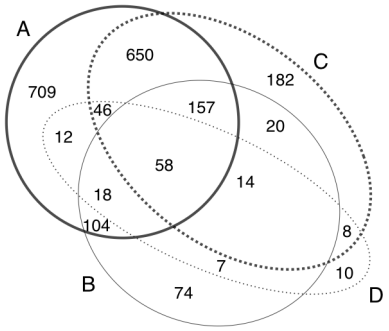

Figura 4.1: rappresentazione grafica dei dati del campionamento [7]

È emerso così che nell'area di Casale Monferrato, che secondo la stima del 1987 era abitata da 93 477 persone, risultavano sicuramente affette dal diabete mellito almeno 2 069 soggetti distinti, corrispondenti agli individui presenti nelle quattro liste private dei duplicati.
La figura 4.2 ci mostra in particolare il livello di copertura di ciascuna lista, dove possiamo osservare che nessuna di esse copre completamente il numero di diabetici che sappiamo per certo vivere nell'area sotto studio. Anche la lista A, che risulta essere la più numerosa, lascia comunque fuori almeno il 15% dei diabetici e questo ci evidenza ancora una volta come possa essere oltremodo sbagliato affidarsi ad un singolo registro per avere un'idea complessiva dell'incidenza del diabete. Questo errore sarebbe ancora più grave - e potenzialmente assai pericoloso - nel caso in cui si stia stimando la numerosità degli affetti da patologie particolarmente gravi e/o trasmissibili.

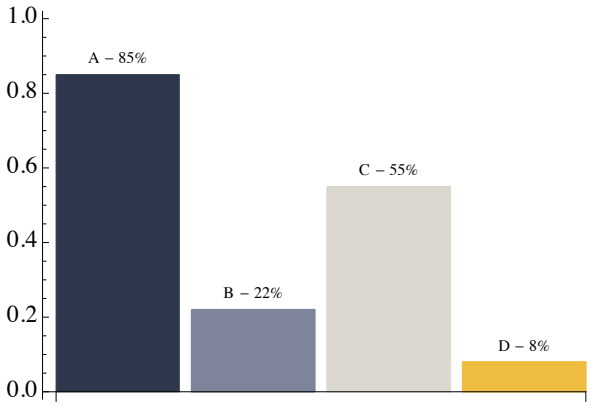

Figura 4.2: copertura dei quattro registri rispetto al numero totale di casi unici accertati dal campionamento (2069)

Una prima analisi tramite il two-samples method

Analizziamo ora i dati secondo le metodologie descritte nel capitolo precedente, iniziando in particolare dal two-samples method proposto da Wittes. Sfruttando i tre stimatori di Lincoln-Petersen, di Chao e di Chapman di pag. 16 ricaviamo facilmente i risultati per ciascuna possibile coppia di liste, come riassunto nella Tabella 4.2 e nella Figura 4.3.
Possiamo notare subito come vi siano delle criticità nei casi 4.2e e 4.2f, in cui la stima $\hat{N}$ proposta dagli stimatori di Lincoln-Petersen e di Chapman è rispettivamente di $803 - 806$ e di $1\,555 - 1\,558$, valori entrambi sotto il valore minimo accertato di 2 069 soggetti. Da questo fatto deduciamo quindi la presenza di dipendenza tra le liste B e D e tra le liste C e D: i pazienti ricoverati in ospedale sono con più probabilità anche quelli che fanno richiesta

di rimborso per siringhe e altro materiale, così come pure vi è una relazione tra le prescrizioni mediche e le richieste di rimborso stesse.

	A	
B	Yes	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.*

	A	
C	Yes	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.*

	A	
D	Yes	No
Yes	134	39
No	1620	-
Lincoln-Petersen: $\hat{N} = 2264$		
Chao: $\hat{N} = 6928$		
Chao C.I.: $5918 \div 7938$		
Chapman: $\hat{N} = 2261$		
Chapman C.I.: $2088 \div 2434$		
odds ratio = 1,2		

(c) *Liste A - D.*

	B	
C	Yes	No
Yes	249	886
No	293	-
Lincoln-Petersen: $\hat{N} = 2060$		
Chao: $\hat{N} = 2529$		
Chao C.I.: $2313 \div 1744$		
Chapman: $\hat{N} = 2057$		
Chapman C.I.: $1907 \div 2208$		
odds ratio = 1,5		

(d) *Liste B - C.*

	B	
D	Yes	No
Yes	97	76
No	335	-
Lincoln-Petersen: $\hat{N} = 806$		
Chao: $\hat{N} = 1007$		
Chao C.I.: $869 \div 1145$		
Chapman: $\hat{N} = 803$		
Chapman C.I.: $711 \div 896$		
odds ratio = 6,9		

(e) *Liste B - D.*

	C	
D	Yes	No
Yes	126	1009
No	47	-
Lincoln-Petersen: $\hat{N} = 1558$		
Chao: $\hat{N} = 3395$		
Chao C.I.: $2916 \div 3873$		
Chapman: $\hat{N} = 1555$		
Chapman C.I.: $1423 \div 1687$		
odds ratio = 3,4		

(f) *Liste C - D.*

Tabella 4.2: analisi dei dati secondo il *two-sample method*, utilizzando gli stimatori di pag. 16 con intervalli di confidenza (C.I.) al 95%

Tutto ciò è anche evidenziato dai valori dell'odds ratio, che sono particolarmente elevati nei casi 4.2e (OR=6,9), 4.2b (OR=3,7) e 4.2f (OR=3,4), dove supponiamo pertanto esserci una dipendenza positiva. Questo ci spinge a provare a eseguire un'analisi della tabella ottenuta accorpando le due coppie di liste che presentano una maggiore dipendenza reciproca (ossia $A - C$ e $B - D$) generando così, sempre tramite il two-samples method, una stima corretta del valore $\hat{N}$ cercato.

	AC	
BD	Yes	No
Yes	437	91
No	1541	-
Lincoln-Petersen: $\hat{N} = 2390$		
Chao: $\hat{N} = 3593$		
Chao C.I.: $3373 \div 3812$		
Chapman: $\hat{N} = 2389$		
Chapman C.I.: $2307 \div 2471$		
odds ratio = 3,4		

Tabella 4.3: analisi dei dati aggregati secondo il two-samples method.

Notiamo anche come gli stimatori di Lincoln-Petersen e di Chapman siano in sostanza identici nei risultati forniti, essendo per di più molto simili sia per formulazione che per derivazione, mentre lo stimatore di Chao restituisce spesso valori che si discostano di molto da quest'ultimi (uno su tutti il caso 4.2c, in cui lo la numerosità proposta da Chao è tre volte quella degli altri indici). Analizzando maggiormente i dati, appare però evidente come tale discrepanza sia da attribuire alla struttura dei dati che compongono le singole tabelle. Poiché dove i valori sono tutti dello stesso ordine di grandezza Chao restituisce dei risultati apprezzabili (come ad esempio in 4.2b), possiamo affermare che questo stimatore risente in modo particolare della grandezza reciproca dei valori della tabella, fornendo stime inaccurate nel caso questi non siano simili.

Quanto visto finora ci permette di fare alcune considerazioni. Innanzitutto osserviamo che nessun metodo di accertamento - e conseguentemente nessuna lista - incontrato finora ci permette di avere una buona stima del numero di diabetici nell'area in esame. Il fatto poi che le liste di persone affette da tale malattia non sono tra loro indipendenti ma che anzi esista un complesso sistema di relazioni tra i vari campioni suggerisce che l'uso di solo due liste fornisca una stima piuttosto carente della numerosità totale: pertanto il metodo proposto da Wittes produce in questo caso una sottostima del valore cercato.

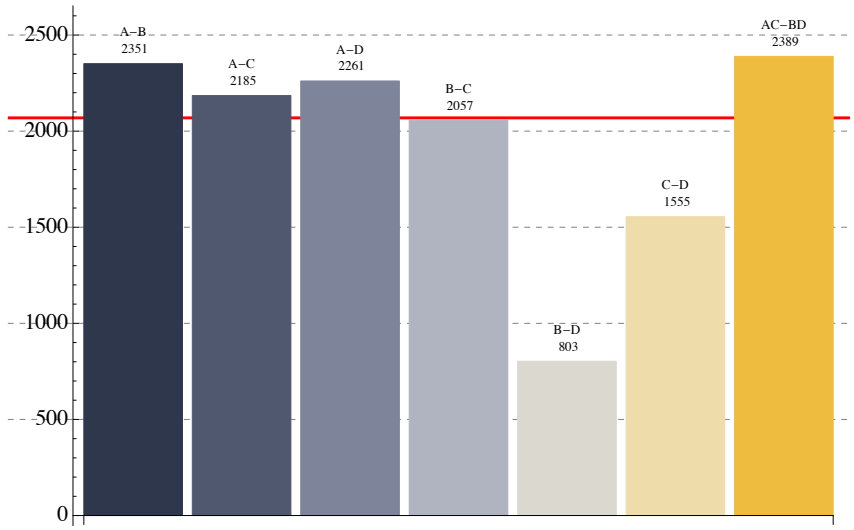

Figura 4.3: numero di diabetici ricavato tramite lo stimatore di Chapman su ciascuna coppia di liste, in rosso il livello 2069

Un'analisi più completa tramite i modelli log-lineari

A questo punto possiamo quindi tentare un secondo approccio, quello dei modelli loglineari. Seguendo sempre il lavoro svolto da Bruno e altri nell'articolo citato, procediamo alla ricerca del modello del tipo

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

che meglio si adatti alla nostra tabella incompleta 4.1: una volta ottenuti i valori $\lambda^{\cdots}$ cercati, sarà immediato prevedere il valore di $n_{0,0,0,0}$.
Consideriamo in particolar modo come assunzione base l'assenza di iterazione contemporanea tra tutte le quattro liste, per cui $\lambda^{ABCD} = 0$, per poi ricercare il modello che meglio si adatta partendo da quello saturo e procedendo eliminando man mano i vari termini, in modo gerarchico, secondo una procedura di tipo *backward*.

Tramite uno dei diversi software di statistica disponibili, arriviamo ad avere che il modello più parsimonioso in termini di componenti e che meglio prevede i dati noti all'interno della tabella è quello costituito da

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

il quale fornisce una stima di 2 771 casi, con un intervallo di confidenza al 95% tra 2 492 e 3 051 e un rapporto di verosimiglianza $L^2 = 7,6$. Possiamo facilmente osservare come questo valore sia maggiore di circa il 30% il numero di casi unici identificati dalle quattro liste in esame, che quindi fornivano una sottostima del valore dei diabetici nell'area di Casale Monferrato. Notiamo anche come il modello di indipendenza tra le liste, ossia quello della forma

$$\log \mathbb{E}[n_{i,j,k,l}] = \lambda + \lambda^A + \lambda^B + \lambda^C + \lambda^D$$

sia alquanto fuorviante, fornendo una stima di 2 250 casi (CI 2 215 - 2 288) ma con un eccessivo valore del rapporto di verosimiglianza $L^2 = 217,5$.

Si può procedere tuttavia con la **stratigrafia** del nostro campione in esame, come spiegato a pag. 13, in modo da provare a ridurre probabili casi di eterogeneità nel campione e ottenere una stima più accurata. Si parte pertanto costruendo la Tabella 4.4, ottenuta suddividendo i casi della Tabella 4.1 in base al tipo di trattamento: dieta, ipoglicemizzanti, insulina. Sarebbe ugualmente possibile una stratigrafia sulla base dell'età del paziente o una ottenuta dalla combinazione età-trattamento, ma si è visto non portare ad alcun miglioramento apprezzabile.

Ottenute queste tre nuove tabelle di contingenza possiamo rieffettuare l'analisi seguendo la medesima procedura *backward* vista prima, per ottenere così i risultati in Tabella 4.5.
Qui troviamo che la migliore stima ottenuta tramite l'utilizzo del modello di cattura-ricattura e dei modelli log-lineari è di 2 583 casi di diabete nell'area di Casale Monferrato, fatto che porta l'incidenza della malattia dal $2,2\%$ (in base ai casi accertati) al $2,8\%$ (C.I. $2,4 - 3,1\%$) della popolazione presente nell'area in esame nel 1987.

La Tabella 4.6 mostra infine un confronto tra le diverse stime ottenute in quest'ultimo capitolo, evidenziando come il metodo esposto in queste pagine possa essere di estremo aiuto per identificare i casi sfuggiti al campionamento e quindi fornire una stima più accurata circa la numerosità di una popolazione.

strato	modello	$\hat{N}$	C.I.	L^2
dieta	$\lambda^{AC} + \lambda^{3D}$	360	$303 - 442$	5,7
ipoglicemizzanti	$\lambda^{AC} + \lambda^{BD}$	1890	$1785 - 2014$	13,5
insulina	$\lambda^{AC} + \lambda^{BC} + \lambda^{BD} + \lambda^{CD}$	333	$328 - 341$	7,5
totale	-	2583	$2416 - 2798$	-

Tabella 4.5: analisi loglineare con stratigrafia per tipo di trattamento (in modello sono indicati solo i termini di grado massimo).

C	D	A / B	Yes / Yes	Yes / No	No / Yes	No / No
Yes	Yes		0	1	1	1
Yes	No		1	11	4	11
No	Yes		1	0	0	0
No	No		11	150	13	-

(a) *Sottoposti a dieta.*

C	D	A / B	Yes / Yes	Yes / No	No / Yes	No / No
Yes	Yes		6	9	3	3
Yes	No		98	751	13	163
No	Yes		3	5	1	4
No	No		68	509	58	-

(b) *Sottoposti a ipoglicemizzanti.*

C	D	A / B	Yes / Yes	Yes / No	No / Yes	No / No
Yes	Yes		51	36	10	4
Yes	No		57	62	3	6
No	Yes		14	7	6	6
No	No		22	42	2	-

(c) *Sottoposti a insulina.*

Tabella 4.4: tabelle di contingenza del campione, suddivisa per tipo di trattamento.

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	2 389	92%
modello loglineare	2 771	107%
modello loglineare con stratigrafia	2 583	100%

Tabella 4.6: confronto tra le varie stime del numero di diabetici in Casale Monferrato

Conclusioni

Si è voluto in queste pagine esporre una delle innumerevoli possibili metodologie delle scienze statistiche, il modello di cattura-ricattutra, per la **descrizione** e la **comprensione** della realtà a noi circostante.

Obiettivo preminente del ricercatore è quello di ricavare, dalla complessità osservata e rilevata, una visione chiara e semplificata dello specifico fenomeno che si vuole analizzare, un modello che sia al tempo stesso il più veritiero possibile.
Questo è lo scopo che hanno perseguito gli studiosi susseguitisi nel corso del tempo, alcuni dei quali sono nominati in questa tesi. Di fronte a un problema di stima della numerosità di una popolazione, Petersen e Lincoln hanno dato un loro primo contributo, fornendo una loro innovativa visione del problema e formulando un loro modello. Con il procedere degli anni numerosi altri ricercatori hanno dato ulteriori apporti, portando ad oggi la metodologia del cattura-ricattura ad un livello, una complessità e un respiro decisamente più grande di quanto non fosse nei primi anni del '900.

Seguendo le linee tracciate da G. Bruno e altri studiosi, nell'ultimo capitolo si è voluto riportare uno studio concreto, quello relativo all'incidenza del diabete nell'area di Casale Monferrato, ripercorrendo le singole fasi. Di fronte ad un problema di carattere generale già affrontato con altre metodologie più "classiche", si è provato ad applicare la tecnica del cattura-ricattura nelle sue varie forme, nel tentativo ben riuscito di migliorare la stima circa la dimensione del fenomeno epidemiologico in esame. Lo specifico studio citato è stato sviluppato costantemente provando a coniugare **semplicità** del modello - dapprima utilizzando il sistema a due liste, poi scegliendo il modello più parsimonioso quanto a termini λ costitutivi - e sua **aderenza alla realtà**, testando di volta in volta i risultati con diverse metodologie.

Questo singolo caso specifico ci mostra quindi le possibilità offerteci da questa scienza, che ci può fornire una visione più dettagliata e completa della circostanza analizzata, in modo da poter meglio comprendere e all'occorrenza dirigere il fenomeno: ecco l'importanza della Statistica.

Glossario

campione date X_i variabili casuali, il vettore casuale $(X_1, X_2, \ldots, X_n)$ costituisce un campione casuale relativo ad una variabile casuale X se i suoi elementi sono indipendenti e hanno la stessa distribuzione. 9

distorsione dato uno stimatore T del parametro θ, si dice distorsione di T il valore

$$B_\theta(T) = \mathbb{E}(T) - \theta$$

Uno stimatore per cui $B_\theta(T) = 0$ risulta essere "non distorto", proprietà rilevante per la corretta stima di θ. 17

distribuzione multinomiale si tratta di una distribuzione di probabilità discreta che generalizza la distribuzione binomiale

$$\mathbb{P}(k) = \binom{n}{k} p^k (1-p) n - k$$

In particolare, dati c parametri, la distribuzione multinomiale di parametri $(p_1, \ldots, p_m)$ tali che $\sum_i p_i = 1$ descrive la probabilità di estrarre con ripetizione un campione di n unità e di osservarvi n_1 unità nella categoria c_1, $\ldots$, n_m unità nella categoria c_m. Questa probabilità è data da

$$\mathbb{P}(n_1, n_2, \ldots, n_m) = \frac{n!}{n_1! n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} \cdots p_c^{n_m}$$

Si osserva facilmente come per $m = 2$ ci si riconduce alla distribuzione binomiale. 15

famiglia esponenziale è una famiglia di funzioni di densità di probabilità per cui, fissati k parametri $(\theta_1, \ldots, \theta_k)$, la corrispondente funzione di massa può essere scritta come

$$f(x, \theta) = C^*(x) D^*(\theta) \exp\Big\{ \sum_{n=1}^{k} A_n(\theta) B_n(x) \Big\}$$

con $A_n(\theta)$ per $n = 1, \ldots, k$ e $D^*(\theta)$ funzioni dei soli parametri $\theta_1, \ldots, \theta_k$ e $B_n(x)$ per $n = 1, \ldots, k$ e $C^*(x)$ funzioni della sola x. 21

indipendenza due eventi A e B si dicono indipendenti qualora

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

o, in forma del tutto equivalente, quando vale $\mathbb{P}(A \,|\, B) = \mathbb{P}(A)$ e $\mathbb{P}(B \,|\, A) = \mathbb{P}(B)$. 13

intervallo di confidenza corrisponde ad un insieme di valori in cui la probabilità di trovarvi un qualche parametro θ sia maggiore di un valore α fissato. 26

odd è il rapporto fra la frequenza di coloro che sono in una categoria e la frequenza di coloro che non sono in quella categoria.
La sua interpretazione è la possibilità che un individuo scelto a caso si trovi in una data categoria prestabilita piuttosto che in una qualunque delle altre. (*Log-Linear Models*, David Knoke, Peter J. Burke, SAGE Publications, 1980). 14

odds ratio è una misura di associazione per le tabelle di contingenza a due entrate ed è dato dal rapporto tra due odds

$$\vartheta = \frac{odd_1}{odd_2} = \frac{p_1/(1-p_1)}{p_2(1-p_2)}$$

L'odds ratio è un numero non negativo che vale 1 nel caso di indipendenza, in cui $odd_1 = odd_2$; per $\vartheta > 1$ la probabilità di successo nella prima riga è più alta che nella seconda e più tale valore è distante da 1 più è forte l'associazione tra le due variabili. L'odds ratio non cambia valore in base all'orientazione della tabella . 14

popolazione indica l'insieme degli elementi oggetti dello studio. 8

regressione logistica spesso indicata anche come "regressione logit", è un particolare modello lineare generalizzato che mette in relazione la probabilità p di un evento X con le variabili $X_1, \ldots, X_n$ tramite la funzione $logit[\,\cdot\,]$, funzione che corrisponde al logaritmo naturale dell'*odd*

$$logit\big(p(x)\big) = \beta_0 + \sum_{j=1}^{n} \beta_j x_j$$

Viene usata quando la relazione tra i valori $p(x)$ e le variabili X_i non è lineare . 13, 21

stimatore è una funzione che ad ogni campione statistico associa un valore del parametro θ da stimare. 15

tabella di contingenza è una tabella a due o più entrate che riporta le frequenze congiunte delle variabili, spesso utilizzata in statistica per rappresentare e analizzare le relazioni tra variabili categoriche. 14

variabile casuale è una funzione sullo spazio di probabilità che associa ad ogni valore del dominio la probabilità del suo manifestarsi; in particolare $X\colon \Omega \to \mathbb{R}$ variabile casuale soddisfa $\forall B \in \mathscr{B}$, $X^{-1}(B) \in \mathcal{E}$, dove $X^{-1}(B) = \{\omega \in \Omega\colon X(\omega) \in B\} \subseteq \Omega$. 15

verosimiglianza è una funzione, più nota col termine inglese "likelihood function" e indicata con $L(\theta)$, che rappresenta la probabilità di osservare il campione al variare del parametro θ. Nel caso di osservazioni campionarie, che sono per definizione indipendenti ed equidistribuite, possiamo scrivere la funzione come

$$\begin{aligned} L(\theta) &= \mathbb{P}(\textit{dati osservati}, \theta) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n, \theta) \\ &= \mathbb{P}(X_1 = x_1) \cdot \mathbb{P}(X_n = x_n) = \prod_k f(x_k, \theta) \end{aligned}$$

con $f(x_k, \theta)$ funzione di densità di X_k. 24, 25

Bibliografia

[1] Damiano D. Abeni, Giovanna Brancato e Carlo A. Perucci. ≪Capture-Recapture to Estimate the Size of the Population with Human Immunodeficiency Virus Type 1 Infection≫. In: *Epidemiology Resources Inc* 5.4 (1994).

[2] Alan Agresti. *An introduction to categorical data analysis*. Wiley Interscience, 2007 (cit. a p. 14).

[3] Alessandra Andreotti. *Campionamento per popolazioni rare o elusive*. 2004.

[4] Patrick Ball, Wendy Betts et al. *Killings and Refugee Flow in Kosovo: March - June 1999*. A Report to the International Criminal Tribunal for the Former Yugoslavia. American Association for the Advancement of Science, 2002 (cit. a p. 20).

[5] Hermann Brenner, Christa Stegmaier e Hartwig Ziegler. ≪Estimating completeness of cancer registration: an empirical evaluation of the two source capture-recapture approach in Germany≫. In: *Journal of Epidemiology And Community Health* 49 (1995).

[6] G. Bruno, G Bargero et al. ≪A population-based prevalence survey of known diabetes mellitus in northern Italy based upon multiple independent sources of ascertainment.≫ In: *Diabetologia* 35.9 (1992) (cit. a p. 29).

[7] Graziella Bruno, Ronald E. LaPorte, Franco Merletti et al. ≪Application of capture-recapture to count diabetes?≫ In: *Diabetes Care* 17.6 (1994), p. 548 (cit. alle pp. 29, 30).

[8] Anne Chao. ≪Estimating the Population Size for Capture-Recapture Data with Unequal Catchability≫. In: *Biometrics* 43.4 (1987) (cit. a p. 17).

[9] Anne Chao, P. K. Tsay, Sheng-Hsiang Ling et al. ≪The applications of capture-recapture models to epidemiological data (Tutorial in biostatistics)≫. In: *Statistics in medicine* (2001) (cit. alle pp. 20, 21).

[10] Douglas George Chapman. ≪Some Properties of the Hypergeometric Distribution With Applications to Zoological Censuses≫. In: *University of California Publications in Statistics* 1 (1951) (cit. a p. 17).

[11] Giorgio Chiari e Pierangelo Peri. *I modelli loglineari nella ricerca sociologica*. Università di Trento, 1987 (cit. a p. 23).

[12] Ronald Christensen. *Log-Linear Models and Logistic Regression*. Springer, 1997.

[13] Bradley Efron. ≪Second thoughts on the bootstrap≫. In: *Statistial Science* 18.2 (2003) (cit. a p. 26).

[14] N. Fisher et al. ≪Estimating numbers of homeless and homeless mentally ill people in north east Westminster by using capture-recapture analysis≫. In: *British Medical Journal* (1994).

[15] IWGDMF International Working Group for Disease Monitoring an Forecasting. ≪Capture-Recapture and Multiple-Record Systems Estimation I: History and Theoretical Development≫. In: *American Journal of Epidemiology* 142 (1995) (cit. a p. 25).

[16] IWGDMF International Working Group for Disease Monitoring an Forecasting. ≪Capture-Recapture and Multiple-Record Systems Estimation II: Applications in Humand Diseases≫. In: *American Journal of Epidemiology* 142.10 (1995) (cit. a p. 20).

[17] Maria Letizia Giarrizzo et al. ≪Stima di prevalenza di diabete mellito in una provincia del Lazio attraverso i modelli di cattura e ricattura≫. In: *Epidemiol Prev* (2007).

[18] R. Q. Gurgel, J. D. C. da Fonseca et al. ≪Capture-recapture to estimate the number of street children in a city in Brazil≫. In: *Arch Dis Child* (2004) (cit. a p. 20).

[19] Thomas N. Herzog. *Applications of Capture-Recapture Methods*. 2013.

[20] Ernest B. Hook e Ronald R. Regal. ≪The value of capture-recapture methods even for apparent exhaustive surveys. The need for adjustment for source of ascertainment intersection in attempted complete prevalence studies.≫ In: *American Journal of Epidemiology* (1992) (cit. a p. 14).

[21] Angela Jeansonne. *Loglinear Models*. 2007.

[22] Marieke Lettink e Doug P. Armstrong. ≪An introduction to using mark-recapture analysis for monitoring threatened species≫. In: *Department of Conservation Technical Series 28A* (2003).

[23] Frederick C. Lincoln. *Calculating Waterfowl Abundance on the Basis of Banding Returns*. Rapp. tecn. United States Department of Agriculture, 1930 (cit. a p. 11).

[24] *Notes on using capture-recapture techniques to assess the sensitivity of rapid case-finding methods*. Valid International's Workshop on Coverage Assessment Methods. 2006.

[25] Antony Overstall e Ruth king. *Capture-Recapture Models for Population Estimates*. Report. University of St Andrews, 2011.

[26] Mario Stefano Peragallo, Francesco Urbano, Florigio Lista et al. ≪Evaluation of cancer surveillance completeness among the Italian army personnel, by capture-recapture methodology≫. In: *Cancer Epidemiology* (2010) (cit. a p. 20).

[27] Carl Georg Johannes Petersen. ≪The Yearly Immigration of Young Plaice Into the Limfjord From the German Sea≫. In: *Report of the Danish Biological Station* 5 (1896) (cit. a p. 11).

[28] Guy La Ruche, Dominique Dejour-Salamanca, Pascale Bernillon et al. *Capture-Recapture Method for Estimating Annual Incidence of Imported Dengue, France, 2007-2010*. Rapp. tecn. Centers for disease control e prevention, 2013 (cit. a p. 20).

[29] Zoe Emily Schnabel. ≪The Estimation of Total Fish Population of a Lake≫. In: *The American Mathematical Monthly* 45.6 (1938) (cit. a p. 19).

[30] Leo J. Schouten, Huub Straatman, Lambertus A. L. M. Kiemeney et al. ≪The Capture-Recapture Method for Estimation of Cancer Registry Completeness: A Useful Tool?≫ In: *International Journal of Epidemiology* 23.6 (1994).

[31] C. Chandra Sekar e W. Edwards Deming. ≪On a Method of Estimating Birth and Death Rates and the Extent of Registration≫. In: *Journal of the American Statistical Association* 44.245 (1949) (cit. alle pp. 13, 19).

[32] Laura Terzera. ≪Le indagini campionarie≫.

[33] Laure Vaissade e Stéphane Legleye. ≪Capture-recapture estimates of the local prevalence of problem drug use in six French cities≫. In: *European Journal of Public Health* 19.1 (2008) (cit. a p. 20).