UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

LAUREA TRIENNALE IN STATISTICA E TECNOLOGIE INFORMATICHE

Campionamento per popolazioni rare o elusive

Relatore: Ch.mo Prof. Giancarlo Diana

Laureanda: Alessandra Andreotti

ANNO ACCADEMICO 2003 – 2004

È con immensa gioia e felicità che porgo un ringraziamento speciale ai miei genitori, a Marco, agli amici, ai compagni di Università e a tutti i professori che hanno contribuito al raggiungimento di questo importante traguardo.

INDICE

1. INTRODUZIONE

La moderna teoria del campionamento da popolazioni finite considera non solo insiemi di individui di ampiezza e dislocazione note, per i quali siano presenti liste complete ed esaustive, ma anche le cosiddette *popolazioni rare o elusive*. Per queste ultime sorgono problemi sia in termini di identificabilità dei soggetti, sia in termini di conteggio della numerosità.

Dopo aver introdotto le caratteristiche basilari di queste popolazioni ed accennato ad alcuni casi reali che sono stati oggetto di studio in letteratura, l'obiettivo che ci si propone è quello di fornire un elenco ed una breve descrizione dei metodi di campionamento impiegati nel contesto delle popolazioni rare o elusive, intendendo che una popolazione può essere definita rara, elusiva o, al contempo, rara ed elusiva.

In primo luogo viene presentato il *campionamento per centri*, largamente utilizzato in indagini relative alla popolazione degli immigrati, al fine di poter controllare il fenomeno dell'irregolarità, o alla popolazione degli *homeless.*

In seguito viene descritta la tecnica nota in letteratura come *multiple frames*, che è stata utilizzata, ad esempio, per stimare il numero di individui affetti dal morbo di Alzheimer.

Si fornisce, poi, una descrizione del *campionamento per network*, che ha visto, in particolare, il suo utilizzo in indagini relative al numero di scomparsi in una prefissata località.

Successivamente viene descritto il *campionamento adattivo*, utilizzato per la conduzione di indagini con lo scopo di stimare il numero di animali appartenenti ad una razza in via di estinzione.

Viene poi discusso il metodo dello *snowballing*, che è stato utilizzato per la stima del numero di consumatori di eroina in una data località.

Infine viene descritto il *metodo di cattura e ricattura*, prevalentemente utilizzato per stimare il numero totale di unità appartenenti ad una certa popolazione animale in una particolare area geografica.

Per tutte le metodologie sopra elencate verranno descritte le principali caratteristiche e le tecniche di stima di grandezze di interesse quali la media di una caratteristica Y (campionamento per centri) o il suo totale (multiple frames, campionamento per network e adattivo) o il numero totale di individui appartenenti alla popolazione (campionamento per centri, multiple frames, snowballing, metodo di cattura e ricattura). Inoltre, per ognuna di queste metodologie, si farà accenno alle applicazioni a casi reali suggerite in letteratura, ragionando eventualmente sulla possibilità di poterle intercambiare per utilizzarle indifferentemente in un ambito piuttosto che in un altro.

2. LE POPOLAZIONI RARE ED ELUSIVE

Come accennato nell'introduzione, vi sono numerose popolazioni che, per loro natura e per caratteristiche proprie, rientrano tra le popolazioni rare o elusive. Sono esempi di popolazione rara la collettività di coloro che sono affetti dal morbo di Alzheimer, oppure l'insieme delle famiglie con un componente scomparso; sono, invece, esempi di popolazioni elusive la collettività degli immigrati senza permesso di soggiorno, oppure l'insieme di coloro che non hanno fissa dimora (i cosiddetti homeless). Infine, un esempio reale di popolazione contemporaneamente rara ed elusiva è l'insieme degli animali di una razza in via d'estinzione.

Nel trattare popolazioni di questo tipo che, per loro natura, raccolgono individui che sfuggono al controllo diretto, ci si trova di fronte ad un problema fondamentale: non si può in alcun modo disporre di una lista esaustiva di tutte le unità di popolazione. Al più potranno essere presenti una o più liste incomplete, a volte parzialmente sovrapposte che, generalmente, non forniscono una copertura totale della popolazione. Questo è quello che accade prevalentemente per le popolazioni elusive.

Per le popolazioni rare, invece, potrebbero anche essere disponibili liste complete, ma la difficoltà sta proprio nella loro costruzione: le unità possono, infatti, essere contenute in liste più ampie, a cui appartengono molti altri soggetti che non sono d'interesse per l'indagine.

3. METODI DI CAMPIONAMENTO

L'attenzione verrà posta su sei tecniche principali, e precisamente:

i. Il campionamento per centri, impiegato in assenza di liste;

ii. Il multiple frames, utilizzato in presenza di liste incomplete;

iii. Il campionamento per network, usato quando la numerosità della lista risulta eccessiva;

iv. Il campionamento adattivo, impiegato in assenza di liste, o quando si ha un'eccessiva ampiezza delle liste a disposizione;

v. Lo snowballing, utilizzato quando gli individui della popolazione oggetto di studio si conoscono l'un l'altro e quindi possono fornire informazioni aggiuntive;

vi. Il metodo di cattura e ricattura, applicato, generalmente, alle popolazioni animali per cui non si hanno liste a disposizione.

3.1 Campionamento per centri

Il campionamento per centri, viene impiegato in assenza di liste, ed è stato introdotto, in particolare, riferendosi alle popolazioni di immigrati (popolazioni elusive). In Italia, negli ultimi vent'anni, il fenomeno migratorio è diventato sempre più complicato e vario, essendo aumentato in modo considerevole il numero di ingressi di stranieri all'interno del territorio. Inoltre, è estremamente difficile ottenere delle informazioni affidabili, a causa della continua crescita della componente illegale.

Il campionamento per centri si basa sull'ipotesi fondamentale che gli immigrati tendano ad instaurare una serie di relazioni fra loro, frequentando almeno un ambiente di aggregazione (centro) per motivi legati alla vita qu tidiana: contatti sociali, salute, religione, svago. Il campionamento per centri prevede, quindi, l'esistenza di L centri (individuati, ad esempio, attraverso un'indagine preliminare), che sono punti di aggregazione delle N unità comprese nella prefissata popolazione P oggetto di studio. È necessario precisare che la numerosità totale N e la dimensione di alcuni centri possono essere ignote. Pensando al problema dell'immigrazione, N denota il numero di immigrati, che si distinguono in regolari (con permesso di soggiorno), ed irregolari (senza permesso).

Da ognuno dei centri individuati viene estratto, attraverso uno schema di campionamento casuale semplice senza reinserimento, un numero prefissato di unità: dal centro G_i si estraggono n_i unità, (i = 1, …, L). Su ciascuna unità si rileva poi la caratteristica d'interesse Y.

Viene quindi costruito un campione di dimensione n ottenuto nel modo seguente

$$n = \sum_{i=1}^{L} n_i .$$

Si deve notare, inoltre, che

$$N \leq \sum_{i=1}^{L} N_i ,$$

cioè, il numero totale di unità della popolazione è minore, o al più uguale, della somma delle unità che frequentano i singoli centri, dal momento che vi possono essere individui che frequentano più di un centro.

Ad ogni individuo risulta associato un *profilo* di afferenza ai centri. Il generico profilo viene indicato con $u_r = [u_{r1}, \ldots, u_{ri}, \ldots, u_{rL}]$ ed è un vettore di lunghezza L il cui elemento i-esimo vale 1 se il soggetto frequenta il centro G_i e 0 altrimenti. Poiché l'ipotesi di base è che ogni straniero frequenti almeno un centro, il profilo con tutti elementi nulli non viene preso in considerazione, ed il numero totale dei profili risulta essere $p = 2^L - 1$. Di conseguenza l'indice r varia da 1 fino a p.

Si indichi, inoltre, con N_{ur} l'incognito numero di unità della popolazione con profilo u_r, e con $\{Y_{rq}; q = 1, \ldots, N_{ur}\}$ il sottoinsieme dei valori delle caratteristiche di interesse della popolazione rilevati sulle unità con profilo u_r. Ad ogni unità campionata viene chiesto di compilare un questionario che consenta, fra le altre cose, di evidenziarne il profilo. In questo modo si può disporre della quantità $f_{ur,\,i}$, (r = 1, .., p; i = 1, …, L), che denota la frequenza del profilo u_r nel centro G_i, e della quantità $f_{ur} = \sum_{i=1}^{L} f_{ur,\,i}$ che rappresenta la frequenza del profilo u_r nella popolazione.

Inoltre si può disporre anche della serie di valori campionari della caratteristica Y osservati per le unità con profilo u_r indicata con $\{y_{rs};\ s = 1, \ldots, f_{ur}\}$.

Una stima per la media $\overline{Y}$ di popolazione della caratteristica Y è la seguente

$$\overline{y} = \frac{1}{n} \sum_{r=1}^{p} \sum_{s=1}^{f_{ur}} y_{rs}\, \hat{c}_{ur} \ , \qquad (1)$$

dove con $\hat{c}_{ur} = \left(\sum_{i=1}^{L} \frac{n_i}{n\alpha_i} u_{ri} \right)^{-1}$ si indicano i coefficienti pesati in funzione del profilo u_r e con $\alpha_i = \frac{N_i}{N}$, i = 1, … L, si indicano le percentuali di individui che frequentano i centri $G_1, G_2, \ldots, G_L$. I coefficienti α_i si ipotizzano noti da precedenti indagini.

Nonostante le unità siano estratte da ogni centro attraverso un campionamento casuale semplice senza reinserimento, è possibile che il campione complessivo non possieda i requisiti di un campione casuale semplice. Questo perché un soggetto che frequenta più centri può essere estratto più di una volta. Per questo motivo vengono inseriti i pesi $\hat{c}_{ur}$ che riaggiustano la stima sulla base del numero di centri frequentati da ogni individuo: il campione pesato riacquista quindi le proprietà volute.

Per la stima della numerosità N, è possibile procedere singolarmente per ogni nazionalità, essendo questa una caratteristica sempre osservata in ogni indagine riguardante gli stranieri. In particolare, il metodo ricorre ai documenti di registrazione dove la dimensione N_A è nota e rappresenta la generica nazionalità l-esima.

Si indichino con $_{nat(l)}Y$ e $_{A}Y$ le variabili dicotomiche che indicano, rispettivamente, se l'unità appartiene alla l-esima nazionalità, e se l'unità della nazionalità l-esima è registrata. Risultano quindi determinate le seguenti quantità campionarie

$$n'_{l} = \sum_{r=1}^{p} \sum_{s=1}^{f_{ur}} {}_{nat(l)}y_{rs}\hat{c}_{ur} \quad , \qquad n'_{Al} = \sum_{r=1}^{p} \sum_{s=1}^{f_{ur}} {}_{A}y_{rs}\, {}_{nat(l)}y_{rs}\hat{c}_{ur}$$

dove:

- n'_{l} rappresenta il numero di unità campionarie che appartengono alla l-esima nazionalità;
- n'_{Al} rappresenta il numero di unità della nazionalità l-esima ufficialmente registrate;

Inoltre, partendo dalla l-esima nazionalità è possibile definire un elemento di una ipotetica macro-area (per esempio, se la nazionalità in questione è Albanese, la relativa macro-area è Est Europa). Quindi, indicando con $_{j}n'$ la dimensione campionaria della j-esima macro-area e con $_{j}n'_{A}$ il numero di individui campionati appartenenti alla macro-area j e ufficialmente registrati, si identifica la quantità $_{j}\tilde{z}_{Al}$ data da

$${}_{j}\tilde{z}_{Al} = \frac{n'_{Al}}{n'_{l}} - \frac{\sqrt{{}_{j}n'}}{\left(\sqrt{{}_{j}n'} + \sqrt{n'_{l}}\right)}\left(\frac{n'_{Al}}{n'_{l}} - \frac{{}_{j}n'_{A}}{{}_{j}n'}\right),$$

dove n'_{Al}/n'_{l} è la proporzione di unità appartenenti alla l-esima nazionalità e presenti nelle liste di registrazione.

Infine, una stima della dimensione totale della nazionalità l-esima all'interno della macro-area j può essere ottenuta nel seguente modo

$${}_{j}\hat{N}_{l} = N_{A}/{}_{j}\tilde{z}_{Al} \,,$$

e, aggregando, si ottiene una stima dell'intera dimensione

$$\hat{N} = \sum_{j} \sum_{l} {}_{j}\hat{N}_{l} . \qquad (2)$$

Applicazione a dati reali

Dal 1990 la tecnica del campionamento per centri è stata applicata a varie rilevazioni condotte sulla popolazione degli immigrati.

Il più importante contributo, è stato dato dall'indagine condotta da I.S.M.U (Fondazione per l'Integrazione e Multi-etnicità), come parte delle attività di "osservatore provinciale" lanciate nel 1996 in alcune aree Lombarde, e unite all'esperienza dell'Osservatorio Regionale nel 2001.

Questo ha reso possibile la conoscenza e la stima delle caratteristiche della presenza straniera nelle sue componenti, legale ed illegale.

Attraverso la metodologia illustrata nella sezione precedente, è stato stimato il tasso di illegalità in alcune aree geografiche e le indagini hanno fornito i risultati presentati in Tabella 1. Questi dati consentono di capire quali siano stati gli effetti di contenimento della componente illegale prodotti dalle sanatorie del 1998 e 2002. Infatti, osservando la tabella sottostante, si può notare come, dopo la sanatoria del 1998, le percentuali tendano a diminuire, per poi aumentare nuovamente. Per esempio, la percentuale di illegalità passa dal 21% in tutta la Lombardia al 1 Gennaio 2001, al 31% dell'anno dopo. Solo dopo la sanatoria del 2002 le stime tendono a diminuire, infatti, la percentuale di illegalità passa dal 31% in tutta la Lombardia al 1 Gennaio 2002, al 11% al 1 Luglio 2003.

Tab. 1 – Stima del numero di immigrati illegali per ogni 100 stranieri presenti nell'area

Area	31-12-1998	30-06-1999	30-06-2000	01-01-2001	01-01-2002	01-07-2003
Milano città	20	17	16	22	36	14
Provincia di Milano	30	24	20	22	35	13
Varese			22	17	26	7
Como				20	29	8
Sondrio				24	26	13
Bergamo				22	23	8
Brescia				18	27	9
Pavia				27	28	11
Cremona			14	19	29	8
Mantova			16	15	19	8
Lecco				16	26	9
Lodi		29		23	25	7
Lombradia				21	31	11

Fonte: Lombardia – I.S.M.U. Fondazione e Osservatorio Regionale per l'Integrazione e la Multi-etnicità.

3.2 Il Multiple frames

La metodologia denominata multiple frames viene tipicamente utilizzata per indagini condotte su una popolazione i cui membri risultano affetti da una particolare malattia (popolazione rara).

Come si è già notato in precedenza, può accadere che non sia disponibile alcuna lista completa della popolazione oggetto di studio, ma che, ad esempio, esistano una o più liste parziali.

In questo tipo di tecnica si dispone di L liste (incomplete) della medesima popolazione P, e da ciascuna di esse si estrae un campione attraverso campionamento casuale semplice.

Una stima proposta per il totale della caratteristica Y è la seguente:

$$\hat{y} = \sum_{s=1}^{f_{ur}} \sum_{r=1}^{p} \frac{y_{rs}}{\delta_r} \ , \qquad (3)$$

dove $\delta_r = \sum_{i=1}^{L} \frac{n_i u_{ri}}{N_i}$.

Il fatto che con la (1) si stimi $\bar{y}$ e con la (3) $\hat{y}$ non deve stupire; nel contesto del campionamento per centri è possibile al più avere informazioni a proposito di $\alpha_i = N_i/N$, nel caso del multiple frames sono note, invece, le numerosità N_i ma non N. Infatti qualche membro della popolazione può essere inserito in più di una lista. Si pensi, ad esempio, ad una persona con una particolare malattia e alla possibilità che essa sia stata ricoverata in più di un ospedale. Per questo motivo si ha, nuovamente,

$$N \leq \sum_{i=1}^{L} N_i \, .$$

È proprio a causa di questa parziale sovrapposizione che il multiple frames può essere inteso, assimilando la lista al centro, come un caso particolare del campionamento per centri; la particolarità consiste nel fatto che sono note le dimensioni N_i delle liste.

Qui di seguito vengono brevemente illustrati due metodi per affrontare il problema delle sovrapposizioni: il primo propone di ridefinire le liste in modo da eliminare la sovrapposizione, il secondo suggerisce di effettuare una compensazione delle liste.

Eliminazione delle sovrapposizioni

Un primo modo per eliminare le sovrapposizioni è di inserire le varie liste disponibili in una singola lista senza duplicati. Un metodo per contrastare l'aumento degli errori è quello di definire le liste in modo da semplificarne le unioni, anche se questo si realizza a costo di una perdita in termini di efficienza delle stime.

A Washington, nel 1960, è stata applicata questa metodologia per condurre uno studio sulla popolazione sorda. Per tale studio è stata raccolta una lista completa delle persone colpite da sordità avvalendosi di informazioni fornite da organizzazioni e scuole per sordi, da informatori e da agenzie sociali.

Una procedura alternativa per eliminare le sovrapposizioni è di usare un'unica identificazione per specificare coloro che sono realmente inclusi nella lista. Si considerino, per esempio, le liste fornite dalle aziende ospedaliere di pazienti ricoverati in un certo reparto. Ipotizzando che vi sia una lista prioritaria (relativa ad esempio all'azienda ospedaliera di dimensioni maggiori), si considererà questa come *lista di base*. Tutti gli individui appartenenti a questa lista entrano nella *lista unica* che si vuole costruire. Con lo stesso criterio verranno ordinate progressivamente tutte le altre liste a disposizione. Si considera allora la lista che, nell'ordine stabilito, viene dopo la lista di base e la si confronta con essa. Se l'individuo preso in considerazione è già presente nella lista di base, l'elemento viene respinto dalla lista unica, se invece questo non si verifica, l'elemento viene accettato. Questa procedura evita il bisogno di inserire le diverse liste; l'indagine viene svolta solo per campionare i listati presenti nella lista prioritaria.

La creazione di una lista combinata senza duplicati può essere particolarmente utile quando lo scopo dell'indagine è stimare la dimensione della popolazione, N, e il totale della popolazione, $Y = N\overline{Y}$. Infatti, quando la lista combinata è ampia e non contiene spazi vuoti (unità cancellate), N è nota e Y può essere stimata da $N\overline{y}$.

Compensazione delle sovrapposizioni

Quando un campione viene estratto da due o più liste sovrapposte, la probabilità che un elemento della popolazione venga selezionato dipende dal numero di liste in cui esso appare.

Un primo metodo utilizzato per tenere conto di questo ulteriore aspetto consiste nell'assegnare dei *pesi* agli elementi del campione. Tali pesi vengono costruiti in modo tale che siano inversamente proporzionali o alle probabilità di inclusione, o al presunto numero di selezioni: sono inversamente proporzionali alle probabilità di inclusione quando nessun elemento della popolazione può comparire nel campione più di una volta, mentre, sono inversamente proporzionali al numero atteso di selezioni quando un'unità campionaria, selezionata da più di una lista, è già inclusa nel campione.

L'uso di questi pesi richiede la conoscenza delle liste, dalle quali viene estratto ogni elemento del campione.

Un approccio generale utilizzato in presenza di liste sovrapposte, verrà analizzato qui di seguito con riferimento al caso in cui siano presenti due liste di campionamento, A e B.

Gli elementi della popolazione oggetto di studio apparterranno ad uno dei seguenti tre sottoinsiemi: i membri della sola lista A (in numero pari a N_a), i membri della sola lista B (in numero pari a N_b), ed i membri di entrambe le liste (in numero pari a N_{ab}). Una o tutte le liste possono contenere, inoltre, membri che non appartengono alla popolazione di interesse.

L'obiettivo principale di questa procedura è dividere i membri della popolazione rispetto alle due liste disponibili. La caratteristica di interesse Y viene, quindi, divisa in due parti: *a*Y associata agli elementi della lista A e *b*Y associata agli elementi della lista B.

Quando il numero degli elementi della popolazione è noto per i tre sottoinsiemi (sono cioè note le quantità N_a, N_b, e N_{ab}), allora il totale della caratteristica Y nella popolazione viene stimato da

$$\hat{Y}_1 = N_a \bar{y}_a + N_b \bar{y}_b + N_{ab}\left(a\bar{y}'_{ab} + b\bar{y}''_{ab}\right), \qquad (4)$$

dove:

- $\bar{y}_a$ è la media campionaria di Y degli elementi che appartengono solo alla lista A;
- $\bar{y}_b$ è la media campionaria di Y degli elementi che appartengono solo alla lista B;
- $\bar{y}'_{ab}$ è la media campionaria di Y degli elementi appartenenti ad entrambe le liste, ma campionati dalla lista A;
- $\bar{y}''_{ab}$ è la media campionaria di Y degli elementi appartenenti ad entrambe le liste, ma campionati dalla lista B.

Nel caso in cui N_a, N_b e N_{ab} siano ignoti, il totale si può stimare nel seguente modo

$$\hat{Y}_2 = F_A\left(y_a + a y'_{ab}\right) + F_B\left(y_b + b y''_{ab}\right), \qquad (5)$$

dove:

- F_A e F_B sono le inverse delle frazioni di campionamento per le due liste;
- y_a, y_b, y'_{ab}, y''_{ab} sono i totali campionari.

Caso reale

Facendo riferimento all'esempio citato nell'introduzione, l'indagine sul numero di individui affetti dal morbo di Alzheimer viene condotta applicando la tecnica del multiple frames, ma potrebbe anche essere realizzata utilizzando il campionamento per centri, essendo il multiple frames un caso particolare di esso. Quindi, nell'esempio, gli ospedali possono essere visti come centri e i malati come unità.

Nella situazione appena considerata, si può disporre di liste realizzate dagli ospedali, le quali però possono presentare intersezioni dal momento che uno stesso malato può essere stato ricoverato in più di un ospedale. Si può quindi utilizzare il metodo dell'eliminazione delle sovrapposizioni per poter disporre di una lista unica.

3.3 Il campionamento per network

Il campionamento per network è un metodo che si applica quando la numerosità delle liste è troppo elevata e viene utilizzato, ad esempio, in indagini sul numero di scomparsi in una prefissata località (popolazione rara). Con tale campionamento vengono estratti, con tecnica casuale semplice, n gruppi fra gli L di cui si compone la popolazione P oggetto di studio, costituita da N unità.

In riferimento alla caratteristica d'interesse Y di cui si cercano informazioni, vengono esaminate direttamente tutte le unità contenute in ciascuno degli n gruppi estratti e poi, nel campione, vengono inserite altre unità che sono connesse alle prime da un prefissato legame. Le unità coinvolte da questo legame formano un *network* di molteplicità pari al numero dei gruppi che entrano in gioco nella formazione dei legami. Ad esempio, se il gruppo è la famiglia ed il legame che si vuole studiare è "essere fratelli", il soggetto compreso nella famiglia estratta e tutti i suoi fratelli, appartenenti o meno ad essa, formano un network. Al contrario, una coppia di coniugi osservata nella stessa famiglia, poiché non soddisfa il legame prefissato, non può appartenere allo stesso network.

Il fatto che un'unità possa essere compresa nel campione, o direttamente perché è stato scelto il gruppo nella quale è inserita, o indirettamente perché è compresa in un network, fa spostare l'attenzione dalle unità ai network.

Per la stima del totale si ricorre allo stimatore Horvitz-Thompson modificato. La probabilità di inclusione dei network, che per il k-esimo network osservato di numerosità m_k, risulta essere

$$\pi_k = 1 - \frac{\binom{L - m_k}{n}}{\binom{L}{n}} \qquad (\text{con } 1 \leq k \leq n), \qquad (6)$$

conduce alla seguente stima del totale della caratteristica Y

$$\hat{y} = \sum_{k=1}^{n} \frac{y_k}{\pi_k} \qquad (7)$$

dove y_k è il totale della caratteristica Y rilevato sul k-esimo network esaminato.

Caso reale

Facendo riferimento all'esempio citato nell'introduzione, riguardante l'indagine sul numero di scomparsi in una determinata località, le famiglie potrebbero essere definite come i gruppi di campionamento, mentre il legame potrebbe essere la specificazione di un certo grado di parentela.

3.4 Il campionamento adattivo

Il campionamento adattivo è un metodo che si applica quando si presenta il problema della mancanza di liste oppure quando esse hanno ampiezza eccessiva. Questa metodologia è stata utilizzata, ad esempio, per stimare il numero di animali di una razza in via di estinzione (popolazione rara ed elusiva).

Con tale campionamento vengono estratti, con tecnica casuale semplice, n unità fra le N che costituiscono la popolazione P oggetto di studio.

Sia y'_k il valore della caratteristica Y osservata sulla k-esima unità, (k = 1, .., n). Se è soddisfatta la condizione $y'_k > h$, essendo *h* una costante fissata a priori, si esaminano le unità ad essa "*adiacenti*". Ad esempio, se le unità sono zone geografiche di forma rettangolare, il criterio di "adiacenza" prevede di esaminare le quattro zone confinanti. Quindi, se in una o più unità adiacenti, il corrispondente valore di Y risulta maggiore di *h*, si procede nell'analizzare le unità adiacenti a queste ultime; non appena si trova un'unità il cui valore di Y è minore di *h*, l'indagine si arresta. In questo modo, accade che dalla k-esima unità osservata inizialmente se ne osservino altre ad essa adiacenti.

L'insieme di tutte le unità esaminate conseguentemente alla k-esima, per le quali l'osservazione del fenomeno soddisfa la condizione definita precedentemente, viene detto *network* di molteplicità pari al numero delle unità che lo formano. Le unità fra le n iniziali, dove l'indagine si interrompe, poiché il corrispondente valore di Y non soddisfa la condizione, forma un network di molteplicità pari a 1.

La stima del totale di Y coincide con la (7), dove y_k rappresenta il totale della caratteristica Y rilevata nel k-esimo network osservato, il quale ha probabilità di inclusione pari alla (6). Se il network k-esimo ha molteplicità pari a 1, poiché $y'_k < h$, risulta che $y'_k = y_k$.

Ovviamente, se ciascuno degli L gruppi è formato da una sola unità, così che L = N, il campionamento per network coincide formalmente con quello adattivo; in tal caso si ha $1 \leq k \leq n$. Si può, quindi, considerare il campionamento adattivo un caso particolare del campionamento per network.

Caso reale

Facendo riferimento all'esempio citato nell'introduzione, relativo all'indagine sul numero di animali di una razza in via di estinzione, le unità sono gli appezzamenti di terra sui quali si trovano gli animali e la condizione per continuare l'indagine è che un territorio contenga almeno *h* animali; mentre la condizione di "adiacenza" riguarda l'esame delle zone limitrofe. Nella situazione appena considerata, le unità su cui si lavora sono gli appezzamenti di terreno, ma ciò che interessa stimare è il numero di animali che si trovano sugli appezzamenti. In pratica, gli stimatori utilizzati sono gli stessi citati nel campionamento per network.

3.5 Lo snowballing

La tecnica dello snowballing viene impiegata in studi su popolazioni rare ed, in particolare, è stata utilizzata in indagini condotte per stimare il numero di consumatori di eroina in una data località.

Una condizione necessaria affinché questa tecnica abbia successo presuppone che gli individui della popolazione si conoscano l'un l'altro.

Questa condizione non è verificata per tutte le popolazioni rare, ma solo per qualche minoranza etnica, gruppi religiosi, persone invalide (ad esempio persone sorde), ecc.

Lo scopo dello snowballing è quello di creare una lista di elementi della popolazione che possiedano la caratteristica Y d'interesse. Questo viene realizzato identificando alcune unità della popolazione oggetto di studio, alle quali viene chiesto, attraverso un questionario, di individuare degli altri membri da contattare, a cui verrà applicata la stessa procedura.

L'unico problema nell'usare questo metodo, è la *completezza* della lista. Le unità mancanti dalla lista sono, probabilmente, coloro che vengono socialmente isolati dagli altri membri della popolazione.

Si assume, quindi, che sia disponibile un campione iniziale, di n elementi, dalla popolazione rara e che ciascuna di queste unità identifichi il nome di altri soggetti. Alcuni di questi individui menzionati sono inclusi nel campione iniziale, mentre altri no. Coloro che non appartengono al campione iniziale e che vengono nominati da almeno un individuo, si dicono "appartenenti alla prima ondata dello snowballing". Coloro che non appartengono al campione iniziale o alla prima ondata dello snowballing, ma vengono nominati da almeno un individuo appartenente alla prima ondata, si dicono "appartenenti alla seconda ondata dello snowballing". E così via.

Un'ondata si dice "finale" quando i suoi membri non nominano nessun individuo che non sia già stato menzionato precedentemente. Lo snowballing risulta incompleto quando il campionamento si ferma prima dell'ultima ondata.

Per attuare il procedimento appena descritto, gli individui della popolazione vengono rappresentati attraverso un *grafico diretto*, in cui le unità sono i vertici di questo grafico. Il numero dei vertici, *v*, è stimato dalle informazioni campionarie.

Il vertice *i* fa un "arco" verso il vertice *j* se l'individuo *i*-esimo, quando viene sottoposto al questionario, identifica l'elemento *j* come un membro della popolazione. I vertici vengono identificati da numeri interi e l'insieme dei vertici è $V = \{1, \ldots, v\}$.

Gli archi sono ordinati a coppie (i, j) di vertici V; se $i = j$, l'arco viene chiamato "laccio". L'insieme degli archi, W, è un sottoinsieme di V^2 che contiene tutti i lacci $\{ (i, i): i \in V \}$ e gli archi $\{ (i, j): i, j \in V^2 \text{ e } i \neq j \}$. Il campione iniziale, S_0, è un sottoinsieme di V.

Il campione iniziale e l'insieme degli archi sono rappresentati da variabili indicatrici dicotomiche $x = (x_i: i \in V)$ e $y = (y_{ij}: (i, j) \in V^2)$, rispettivamente. Quindi: x_i è 1 se il vertice *i* è nel campione iniziale, oppure 0 se non lo è, e y_{ij} è 1 se il grafico contiene un arco da *i* a *j*, oppure 0 altrimenti. La matrice *y* è definita come la "*matrice adiacente al grafico*", dove gli elementi della diagonale sono tutti uguali a 1.

Si definiscono A_j e B_j i sottoinsiemi dei vertici rispettivamente dopo e prima il vertice *j*. Più precisamente

$$A_j = \{i \in V: y_{ji} = 1\}, \qquad B_j = \{i \in V: y_{ij} = 1\}$$

Così che A_j è indicato dalla riga *j*-esima della matrice *y* e B_j è indicato dalla colonna *j*-esima della matrice *y*. Le dimensioni di A_j e B_j sono indicate da a_j e b_j, rispettivamente.

Queste ultime sono ottenute dalla somma per riga e per colonna della *matrice adiacente y*

$$a_j = |A_j| = \sum_{i=1}^{v} y_{ji} \quad , \quad b_j = |B_j| = \sum_{i=1}^{v} y_{ij} \tag{8}$$

Per ogni sottoinsieme di vertici S di V, si denotano con A(S) e B(S) i rispettivi sottoinsiemi dei vertici che stanno dopo e prima gli elementi di S; infatti

$$A(S) = \bigcup_{j \in S} A_j \quad , \quad B(S) = \bigcup_{j \in S} B_j \tag{9}$$

La prima ondata di snowballing iniziata da S_0 è data da $S_1 = A(S_0) \bigcap \overline{S}_0$, dove $\overline{S}_0$ indica il complementare di S_0.

La seconda ondata, invece, è data da $S_2 = A(S_1) \bigcap \overline{S}_0 \bigcap \overline{S}_1$. E così via.

Lo snowballing, iniziato da S_0, è dato da $S_0 \bigcup S_1 \bigcup \ldots \bigcup S_K$, dove K è il numero di ondate e S_{K+1} è il primo insieme vuoto nella sequenza S_1, S_2, … .

						prima ondata								seconda ondata						terza ondata					
	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
campione	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
iniziale	0	1	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0
	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0
	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0
	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0
	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1
	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0
	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1

Figura 1 – *matrice adiacente al grafico* su 25 vertici. I vertici sono stati ordinati per semplificare l'illustrazione dello snowballing a tre ondate, di dimensione totale 21, che genera un nuovo campione dal campione iniziale di dimensione 5

La figura 1 mostra un esempio di *matrice adiacente al grafico y* su $v = 25$ vertici. I vertici sono stati ordinati nel modo seguente: i primi 5 sono i vertici del campione iniziale S_0, gli 8 che seguono sono i vertici della prima ondata S_1, ed i successivi 6 sono i vertici della seconda ondata S_2.

I restanti 2 vertici appartengono alla terza ondata S_3; essendo arrivati a 21 vertici, che è la dimensione totale del campione definita precedentemente, ne risulta che S_4 è vuoto.

Per il calcolo della stima del numero di vertici v è necessario definire il campione iniziale S_0 come un sottoinsieme Bernoulliano di V con probabilità di selezione (inclusione) α. Questo significa che gli indicatori $x_1, \ldots, x_v$ sono indipendenti ed identicamente distribuiti (i.i.d.) da una Bernoulli (α). Ne risulta che il campione iniziale ha dimensione $n = |S_0|$ da una Binomiale (v, α).

Sia, inoltre, W, un sottoinsieme Bernoulliano di V^2 con probabilità di selezione pari a 1 per i lacci e pari a β per tutti gli altri archi. Questo significa che gli indicatori (y_{ij}) sono 1 nella diagonale, mentre sono i.i.d. da una Bernoulli (β) fuori dalla diagonale. I parametri del modello statistico sono v, α, e β.

Ponendo pari a r il numero degli archi che non formano un laccio nel campione iniziale, accade che, dato S_0, r è una Binomiale (n (n–1), β), ed il suo valore atteso condizionato rispetto ad n risulta essere

$$E(r \mid n) = n\,(n-1)\beta.$$

Il valore atteso di r è invece dato da

$$E(r) = v\,(v-1)\,\alpha^2\beta.$$

Si definisca, inoltre, s il numero di archi formati dal campione iniziale alla prima ondata di snowballing, per il quale, dato S_0, s è una Binomiale (n (v – n), β). Il suo valore atteso condizionato ad n è

$$E(s \mid n) = n\,(v-n)\,\beta,$$

mentre incondizionatamente ad n si ha

$$E(s) = v\,(v-1)\,\alpha\,(1-\alpha)\beta.$$

Quindi, dato n, le stime di β e *v* sono date dalle seguenti equazioni

$r = n(n-1)\beta \rightarrow$ che corrisponde a E (r | n)

$s = n(v-n)\beta \rightarrow$ che corrisponde a E (s | n) ,

da cui:

$$\hat{\beta}_1 = r / n(n-1)$$

$$\hat{v}_1 = [nr + (n-1)s] / r \,. \qquad (10)$$

Incondizionatamente a n, le stime di α, β, *v* sono date dalle seguenti equazioni:

$n = v\alpha$

$r = v(v-1)\alpha^2\beta \rightarrow$ che corrisponde a E(r)

$s = v(v-1)\alpha(1-\alpha)\beta \rightarrow$ che corrisponde a E(s) ,

da cui:

$$\hat{\alpha}_2 = r / (r+s)$$

$$\hat{\beta}_2 = r(r+s) / n[(n-1)r + ns] \qquad (11)$$

$$\hat{v}_2 = n(r+s) / r \,.$$

Applicazione a dati reali

In questa sezione si fa riferimento ad un'indagine realizzata per stimare il numero di consumatori di eroina nella città di Groningen. Il campione, formato da coloro che fanno abitualmente uso di questa droga, è stato realizzato utilizzando un campione iniziale di n = 34 persone.

Gli individui del campione sono stati rintracciati attraverso contatti con agenzie di assistenza sociale, medici, e visitando gruppi di incontro di consumatori di eroina.

Dopo un'estesa intervista, nella quale la fedeltà dell'intervistatore era chiara agli utenti, agli intervistati è stato chiesto di indicare altri consumatori di eroina appartenenti alla stessa città. Questi "candidati" sono stati identificati attraverso nome, nickname, professione ed età.

Il numero totale di unità entrate a far parte del campione è $t = 311$ (dove $t = r + s$), di cui $r = 15$ appartenenti al campione iniziale.

Le stime di v risultanti dall'indagine sono le seguenti

$$\hat{v}_1 = 685,$$

$$\hat{v}_2 = 705.$$

3.6 Metodo di Cattura e Ricattura

Il metodo di cattura e ricattura, prima di essere stato adattato per altri scopi più generali, fu applicato per la prima volta in uno studio sulla popolazione dei pesci e sulla popolazione di animali e piante selvatici.

Le assunzioni richieste per il calcolo della stima della dimensione totale della popolazione oggetto di studio, N, sono le seguenti:

a. Durante l'indagine non avvengono cambiamenti nella popolazione (la popolazione è *chiusa*);

b. La marchiatura permane almeno fino alla fine del periodo di studio;

c. Sia gli animali marchiati che i non marchiati hanno la stessa probabilità di essere ricatturati: ciò implica che la cattura, la manipolazione e la marchiatura non hanno alcun effetto sulla successiva probabilità che un individuo sia catturato di nuovo;
d. gli animali marchiati si mescolano omogeneamente con il resto della popolazione;
e. Il campionamento viene effettuato in modo casuale.

In letteratura si trovano molti modelli utilizzabili in questo metodo di campionamento, ma i principali, citati in questo elaborato sono: il *modello a due liste* ed il *modello log-lineare*.

Il modello più semplice, usato per stimare la dimensione totale della popolazione oggetto di studio, è quello a due liste: la prima relativa agli animali catturati e alla loro successiva marchiatura, la seconda relativa alle unità ricatturate.

Un secondo modello usato nella cattura e ricattura è il modello log-lineare, che utilizza i logaritmi dei parametri presenti nel modello a due liste e viene impiegato solitamente nell'analisi di tabelle di contingenza.

Modello a due liste

Per dimostrare alcune delle difficoltà che si incontrano quando si utilizzano le liste di campionamento, consideriamo il modello con due liste A e B.

Siano n_{11}, n_{10}, n_{01} e n_{00} i numeri di individui estratti da entrambe le liste, solo dalla prima lista, solo dalla seconda lista, e da nessuna delle due liste, rispettivamente.

Se si assume che non esista eterogeneità ma che esista dipendenza tra le liste di campionamento, allora queste numerosità hanno distribuzione multinomiale con probabilità p_{ij} (i = 1, 0 e j = 1, 0) di appartenere alle rispettive categorie.

Sia $r = n_{11} + n_{10} + n_{01}$ il numero dei diversi individui listati, assumendo $n_{00} = N - r$.

Ora, si indichi con $n_A = n_{11} + n_{10}$ il numero di individui della prima lista e con $n_B = n_{11} + n_{01}$ il numero di individui della seconda lista.

Dato che la cattura è un procedimento casuale, allora si può facilmente dedurre che anche n_A e n_B sono variabili casuali.

Il valore atteso di queste due variabili risulta essere

$$E[n_A] = Np_A$$

$$E[n_B] = Np_B$$

e, quindi

$$E[n_{11}] = Np_{AB} .$$

Dove $p_A = p_{11} + p_{10}$ è la probabilità di appartenere alla prima lista, $p_B = p_{11} + p_{01}$ è la probabilità di appartenere alla seconda lista, e $p_{AB} = p_{11}$ è la probabilità di appartenere ad entrambe le liste di campionamento.

Assumendo che i membri della lista A coprano la stessa proporzione di popolazione della lista B, allora

$$\frac{n_A}{N} = \frac{n_{11}}{n_B},$$

la quale ci porta alla seguente stima

$$\hat{N} = \frac{n_A n_B}{n_{11}} \quad . \qquad (12)$$

Sapendo che, approssimativamente, il valore atteso di $\hat{N}$ è

$$E[\hat{N}] = \frac{N\, p_A p_B}{p_{AB}} = NR ,$$

allora, se le liste sono indipendenti, R = 1 e $\hat{N}$ è approssimativamente non distorto. Se, invece, le liste non sono indipendenti, allora possiamo scrivere

$$R = \frac{p_A\, p_B}{p_{AB}} = \frac{p_B}{p_{B\,|\,A}} ,$$

dove $p_{B|A}$ è la probabilità che un individuo appartenga alla lista B dato che appartiene già alla lista A.

Un ulteriore modo per scrivere la stima di N è il seguente

$$\hat{N} = r + \hat{n}_{00} \quad , \tag{13}$$

dove $\hat{n}_{00} = \frac{n_{01} n_{10}}{n_{11}}$.

Un problema che si può manifestare nell'uso di questo modello, è la "*dipendenza apparente*". Questa avviene quando è presente l'eterogeneità della popolazione, così che la probabilità di appartenere ad una lista varia da individuo ad individuo. Si assume, quindi, che ad ognuno degli N membri della popolazione siano associate le seguenti variabili casuali: p_A, p_B e p_{AB}. Quindi, sotto opportune condizioni (ad esempio che p_A, p_B e p_{AB} possano assumere i valori 0 o 1), si trova che

$$R = \frac{E[p_A]E[p_B]}{E[p_{AB}]} .$$

Per esaminare cosa accade ad R, è utile assumere l'indipendenza delle liste, così che $E[p_{AB}] = E[p_A p_B]$.

Allora accade che

$$R = 1 - \frac{\text{cov}[p_A, p_B]}{E[p_A p_B]},$$

dove cov $[p_A,p_B] = E[p_A p_B] - E[p_A]\, E[p_B]$. Se la covarianza è positiva, allora $R<1$.

L'effetto dell'eterogeneità, che produce la dipendenza apparente, può, a volte, essere ridotta dalla stratificazione. In questo caso, per ottenere la stima della dimensione totale della popolazione, basta trovare la stima della numerosità di ogni strato, e poi N viene calcolato come somma delle stime delle numerosità di ogni strato.

Concludendo, il modello a due liste è tipicamente utilizzato in presenza di eterogeneità e dipendenza tra le liste. Quando sono presenti più di due liste, il problema dell'eterogeneità e della dipendenza diventa più difficile da trattare. La chiave di questo quesito è la relazione tra le probabilità p_{ij} (i = 1, 0 e j = 1, 0) di appartenere alle varie categorie.

Modello log-lineare

Mentre il modello a due liste si utilizza in presenza di eterogeneità e di dipendenza delle liste, il modello log-lineare si utilizza in presenza di omogeneità e di indipendenza delle liste di campionamento, e viene impiegato in analisi di tabelle di contingenza.

Se esiste interazione tra le liste, allora si può affermare che

$$\log[n_{11}] = \log(N) + \log(p_A) + \log(p_B) + \log(i_{AB}),$$

dove i_{AB} è il parametro che denota la presenza o meno di interazione.

Un primo tipo di modello proposto per l'analisi di una tabella di contingenza completa è il seguente

$$\begin{aligned} \log E(n_{11}) &= u + u_A + u_B + u_{AB} \\ \log E(n_{01}) &= u - u_A + u_B - u_{AB} \\ \log E(n_{10}) &= u + u_A - u_B - u_{AB} \\ \log E(n_{00}) &= u - u_A - u_B + u_{AB}, \end{aligned} \qquad (14)$$

dove il termine u rappresenta l'effetto principale, e i termini u_A, u_B e u_{AB} rappresentano le variabili marginali delle probabilità di apparire nella lista A, nella lista B e in entrambe le liste, rispettivamente.

Una parametrizzazione alternativa alla precedente utilizza il parametro di interazione u_{AB} solo per formare la categoria mancante, cioè

$$\begin{aligned} \log E(n_{11}) &= u \\ \log E(n_{01}) &= u + u_A \\ \log E(n_{10}) &= u + u_B \\ \log E(n_{00}) &= u + u_A + u_B + u_{AB} . \end{aligned} \qquad (15)$$

Quando n_{00} non è osservato, questa parametrizzazione contiene solo i tre parametri u, u_A, u_B per descrivere le tre osservazioni n_{11}, n_{01}, n_{10}. Dalle prime tre equazioni è possibile stimare u, u_A, u_B. In ogni caso, l'unica restrizione che si può applicare è $u_{AB} = 0$; in questo modo si può stimare n_{00} dalla quarta equazione.

Nel caso in cui n_{00} sia osservato, il termine di interazione u_{AB} è il seguente

$$u_{AB} = \log \frac{E(n_{00})E(n_{11})}{E(n_{01})E(n_{10})} = \frac{p_{00}p_{11}}{p_{01}p_{10}} .$$

E' possibile notare, inoltre, che quando u_{AB} è zero, u, u_A e u_B sono rispettivamente:

$$u = \log (N\, p_A\, p_B)$$

$$u_A = \log \frac{1 - p_A}{p_A}$$

$$u_B = \log \frac{1 - p_B}{p_B}.$$

L'estensione di questo modello per più di due liste è di facile determinazione. Ad esempio, con tre liste di campionamento esistono 8 possibili combinazioni di liste nelle quali appaiono gli individui della popolazione oggetto di studio. In generale, ci sono 8 parametri nel modello: u, che rappresenta il logaritmo del numero atteso di individui appartenenti a tutte le liste, u_A, u_B, u_C, che rappresentano i tre effetti principali, u_{AB}, u_{AC}, u_{BC}, che rappresentano i tre fattori di interazione tra due liste, e u_{ABC}, che rappresenta il fattore di interazione tra le tre liste.

Metodi di selezione del modello

Inizialmente i metodi di selezione erano scarsamente utilizzati a causa dell'esistenza di pochi modelli, e quindi quello più appropriato era di facile identificazione. Oggi, invece, esiste una vasta scelta di modelli utilizzati per trattare la dipendenza tra le liste e l'eterogeneità spesso osservata tra gli individui. La selezione del modello è, quindi, parte della procedura di stima, e viene completamente integrata con essa.

I metodi di selezione del modello citati di seguito, vengono realizzati attraverso il test di massima verosimiglianza.

Un primo metodo che utilizza questo test viene chiamato *Criterio di Akaike* (AIC), il quale è calcolato nel modo seguente

$$AIC = -2 \times [\log(L) - q], \tag{16}$$

dove log(L) rappresenta la log-verosimiglianza che permette di calcolare le stime di massima verosimiglianza dei parametri del modello, e q definisce il numero di parametri del modello preso in considerazione.

Il valore che ne risulta rappresenta la misura di quanto bene si adatta il modello ai dati, il quale viene penalizzato dalla somma dei parametri utilizzati (chiamata *complessità del modello*). Il modello migliore è quello relativo al più piccolo valore risultante dall'indice AIC.

Esiste, inoltre, un aggiustamento di questo metodo, proposto nel caso in cui il modello sia composto da un numero esiguo di parametri.

Questo indice viene calcolato nel modo seguente

$$CAIC = -2 \log(L) + q [\log(n) + 1],$$

dove log(L) e q sono gli stessi termini utilizzati nell'indice AIC, mentre n rappresenta la dimensione del campione.

Un secondo metodo di selezione, simile al precedente, viene denominato *Criterio di Bayes* (BIC), il quale risulta essere

$$BIC = -2 \log(L) + q \times \log(n), \tag{17}$$

dove log(L), q e n sono gli stessi termini utilizzati in precedenza.

Limiti del metodo di cattura e ricattura

La validità delle stime calcolate con il metodo di cattura e ricattura, come in tante altre procedure statistiche, dipende strettamente dalle assunzioni del modello. Se queste assunzioni non vengono prese in considerazione, le stime risultanti potrebbero non essere del tutto corrette.

- *Popolazione aperta*: quando è possibile che gli individui di una determinata popolazione possano spostarsi, morire, riprodursi all'interno o al di fuori di essa, la popolazione in questione viene definita "*popolazione aperta*". L'assunzione principale della cattura e ricattura definisce che la popolazione oggetto di studio sia chiusa, nella quale, quindi, non ci siano nascite, decessi o spostamenti. Questa assunzione non è un problema in uno studio di incidenza di una certa malattia, dato che in un'indagine di questo tipo si considera il numero di nuovi casi in un determinato periodo di tempo.

  Al contrario, questa assunzione può essere un problema in uno studio di prevalenza, nel quale si considera il numero di malati in un determinato periodo di tempo, e quindi le persone in gravi condizioni di salute tendono a spostarsi per cercare cure migliori.

- *Fonte di dipendenza*: una delle assunzioni più importanti del metodo di cattura e ricattura è l'indipendenza delle liste di campionamento. Liste positivamente dipendenti e liste negativamente dipendenti possono provocare sottostime o sovrastime, rispettivamente. Se sono disponibili solo liste positivamente dipendenti, allora il metodo a due liste serve come limite inferiore del vero numero di casi. Invece, se sono disponibili solo liste negativamente dipendenti, non è raccomandato utilizzare il metodo di cattura e ricattura.

- *Liste "riportate volontariamente"*: una delle assunzioni del metodo di cattura e ricattura è l'omogeneità della popolazione, la quale afferma che, per una determinata lista, tutti i membri della popolazione abbiano la stessa possibilità di essere estratti. Ma non è sempre è così. Può capitare, ad esempio, che a causa della gravità di una malattia, i membri della popolazione malata differiscano notevolmente circa la possibilità di essere estratti dalla lista.

  Inoltre, se la lista viene riportata volontariamente, possono sorgere alcuni problemi. Ad esempio, una lista ottenuta dai centri che curano l'abuso di droga, sorge principalmente dallo sforzo volontario dei tossicodipendenti di cercare aiuti sanitari. Quando sorgono questi problemi, è suggerito utilizzare un campionamento stratificato. In ogni caso, le liste riportate volontariamente devono essere usate molto prudentemente nel metodo di cattura e ricattura.

Applicazione a dati reali

In questa sezione viene presentato uno studio realizzato da alcuni studenti americani dell'Università di Keele (Nottingham), il quale è stato svolto per stimare, attraverso il metodo di cattura e ricattura, il numero di caramelle presenti in una scatola di Smarties.

La formula utilizzata per il calcolo della dimensione totale della popolazione è la seguente

$$\hat{N} = \frac{a\,b}{c},$$

dove a rappresenta il numero di marchiati e rilasciati nella popolazione, b rappresenta il numero della seconda cattura (ricattura), e c rappresenta il numero di ricatturati nella seconda cattura.

La procedura che è stata seguita per determinare $\hat{N}$ è la seguente:

1. Si è versato il contenuto della scatola di caramelle in un piatto e si sono poi scelte, contate e marchiate tutte le caramelle di un determinato colore, in questo caso caramelle rosse, (a);
2. Sono state poi versate tutte le caramelle, comprese quelle rosse, in una borsa di carta, la quale è stata accuratamente agitata per disperdere le caramelle marchiate nell'intera popolazione;
3. Successivamente è stato estratto il secondo campione, indicando, poi, la dimensione campionaria (b) e il numero di caramelle rosse ricatturate in esso (c);
4. A questo punto è stata calcolata la stima del numero di Smarties;
5. Dopo aver reinserito nella borsa il secondo campione estratto, si è ripetuto questo procedimento fino al raggiungimento di dieci estrazioni;
6. Successivamente, è stata calcolata la media dei dieci campioni;
7. In fine, è stato contato il numero di Smarties realmente presenti nella popolazione iniziale ed è stato confrontato con le stime derivate dai campionamenti.

La conta delle Smarties è servita come verifica pratica della stima della dimensione totale della popolazione e quindi dell'affidabilità del metodo di campionamento.

Questo divertente esperimento è stato citato principalmente per far capire il procedimento basilare del metodo di cattura e ricattura. Ovviamente, in letteratura si trovano molte altre applicazioni utilizzate nel contesto principale di questo metodo di campionamento (popolazioni animali).

4. CONCLUSIONI

In questa tesi sono stati descritti i metodi principali studiati in letteratura, impiegati nel contesto delle popolazioni rare o elusive. In particolare, per quanto riguarda le varie applicazioni a casi reali, l'obiettivo che ci si propone è quello di capire se esiste la possibilità di poter utilizzare, in modi diversi, un metodo di campionamento piuttosto che un altro.

È stata, quindi, costruita la Tabella 2 in modo da poter avere subito un'idea di come i vari metodi si possono intercambiare tra loro.

In particolare, nella tabella sottostante, il simbolo **X** indica che quel metodo di campionamento è stato utilizzato per quella specifica indagine, mentre il simbolo ● indica la metodologia da utilizzare in alternativa.

Tab. 2

	Per centri	Multiple Frames	Network	Adattivo	Snowballing	Cattura-Ricattura
Numerosità immigrati	**X**					
Morbo di Alzheimer	●	**X**				
Scomparsi in una località			**X**			
Razza in via di estinzione			●	**X**		●
Consumo eroina					**X**	
Smarties						**X**

Il campionamento per centri, utilizzato in indagini relative alla popolazione degli immigrati, può essere impiegato anche in altri studi. Ad esempio, l'analisi condotta per stimare il numero di individui affetti dal morbo di Alzheimer (multiple frames) potrebbe essere realizzata utilizzando le stime del campionamento per centri.

Anche il campionamento per network, utilizzato in indagini sul numero di scomparsi in una prefissata località, può essere impiegato per altri scopi. Ad esempio, l'indagine condotta per stimare il numero di animali di una razza in via di estinzione (campionamento adattivo) potrebbe essere realizzata utilizzando le stime del campionamento per network.

Inoltre, si può applicare lo stesso ragionamento tra campionamento adattivo e metodo di cattura e ricattura. Dato che quest'ultimo si applica, generalmente, alle popolazioni animali, e che il campionamento adattivo è stato utilizzato in un'indagine sul numero di animali appartenenti ad una razza in via di estinzione, si può concludere che questo studio può anche essere realizzato utilizzando il metodo di cattura e ricattura.

5. NOTE BIBLIOGRAFICHE

La breve introduzione riguardante le popolazioni rare ed elusive presentata nel paragrafo 2 è stata tratta essenzialmente da Nicolini e Marasini (2003).

Il materiale riguardante il campionamento per centri è stato ricavato da Nicolini e Marasini (2003) e da Blangiardo et al. (2004). Per ulteriori approfondimenti circa gli stimatori utilizzati, si rimanda a Mecatti (2002) e a Marasini e Migliorati (2003).

Per quanto riguarda il multiple frames, il paragrafo è stato ispirato da Kalton e Anderson (1986) e da Nicolini e Marasini (2003). Ulteriori approfondimenti si possono trovare in Kish (1965). Con riferimento al metodo di compensazione delle sovrapposizioni, sono stati citati due casi. Nel secondo di questi, è stata affrontata solamente l'ipotesi con due liste di campionamento, A e B; per ulteriori approfondimenti riguardo al metodo con più di due liste, si rimanda a Hartley (1962).

In relazione al campionamento per network, la sessione è stata suggerita da Nicolini e Marasini (2003). Per ulteriori chiarimenti in riferimento allo stimatore di Horvitz-Thompson modificato, si rimanda a Thompson (1992).

Per il materiale riguardante il campionamento adattivo, è stato utilizzato Nicolini e Marasini (2003). Per ulteriori chiarimenti con riferimento alla probabilità di inclusione, si rimanda a Thompson (1990).

Per quanto riguarda lo snowballing, la sezione è stata suggerita da Kalton e Anderson (1986) e da Ove e Snijders (1994). Per ulteriori informazioni si rimanda a Snow et al. (1981) e Goodman (1961); mentre per approfondimenti riguardanti l'applicazione a dati reali, si veda Bieleman et al. (1993).

Per la sessione riguardante il metodo di cattura e ricattura, sono stati utilizzati, i siti internet http://www.pitt.edu/~yuc2/cr/history.htm e http://science.ntu.ac.uk/rsscse/ts/bts/dudley2/text.htm, Cormack (1989) e Burnham et al. (1995). Per ulteriori approfondimenti riguardo alle parametrizzazioni utilizzate nel modello log-lineare si rimanda a Fienberg (1972) e Cormack (1981), rispettivamente.

RIFERIMENTI BIBLIOGRAFICI

BIELEMAN B. et al. (1993). “Lines across Europe: nature and extent of cocaine use in Barcelona, Rotterdam, and Turin”, *Amsterdam: Swets and Zeitlinger*.

BLANGIARDO G.C., MIGLIORATI S, TERZERA L. (2004). “Center sampling: from applicative issues to methodological aspects”.

BURNHAM P. et al. (1995). “Model selection strategy in the analysis of capture-recapture data”, *Biometrics*, 51, 888-898.

CORMACK R.M. (1981). “Loglinear models for capture-recapture experiments on open populations”, *The Mathematical Theory of the Dynamics of Biological Populations II*, R.W. Hiorns and D. Cooke, eds. London, Academic Press.

CORMACK R.M. (1989). “Log-linear models for capture-recapture”, *Biometrics*, 45, 395-413.

FIENBERG S.E. (1972). “The multiple recapture census for closed populations and incomplete 2^k contingency tables”, *Biometrika*, 59, 591-603.

GOODMAN L.A. (1961). “Snowball sampling”, *Ann. Math. Statist.*, 32, 148-170.

HARTLEY H.O. (1962). “Multiple frame surveys”, *Proc. Of the Social Statistics Section*, Amer. Statist. Ass., 203-206.

KALTON G., ANDERSON W. (1986). “ Sampling rare populations” *J. R. Statist. Soc.*, 149, Part 1, pp. 65-82.

KISH L. (1965). “Survey Sampling”, *New York: Wiley*.

MARASINI D., MIGLIORATI S. (2003). “ Small area estimation for immigrant people”.

MECATTI F. (2002). “La stima della media nel campionamento per centri”, *Statistica*, 2, 285-297.

NICOLINI G., MARASINI D. (2003). “Campionamento per popolazioni rare ed elusive: la matrice dei profili”, *Working Paper n. 02.2003 – gennaio*.

OVE F., SNIJDERS T. (1994). “Estimating the size of hidden populations using snowball sampling”, *Journal of Official Statistics*, Vol. 10, No. 1, pp. 53-67.

SNOW R.E. et al. (1981). “Using reputational sampling to identify residential clusters of minorities dispersed in a large urban region: Hispanics in Atlanta, Georgia”, *Proc. Of the Section on Survey Research Methods, Amer. Statist. Ass.*, 101-106.

THOMPSON S.K. (1990). “Adaptive cluster sampling”, *Journal of the American Statistical Association 85*, 1050-1059.

THOMPSON S.K. (1992). “Sampling”, *Wiley, New York*, 148-158.

ABBREVIAZIONI E SIMBOLI UTILIZZATI

Campionamento per centri:

L	numero di centri o liste o gruppi
P	popolazione oggetto di studio
N	dimensione della popolazione P
n	dimensione del campione o dei gruppi estratti
n_i	numero di unità estratte dal generico centro o gruppo G_i (i = 1, …, L)
N_i	numerosità del centro o gruppo i-esimo
Y	caratteristica oggetto di interesse
u_r	*profilo* di afferenza ai centri (r = 1, …, p)
	$u_r = \begin{cases} 1 \text{ se l'unità frequenta il centro } G_i \\ 0 \text{ altrimenti} \end{cases}$
p	numero totale dei profili, $p = 2^L - 1$
u_{ri}	generico profilo riferito al centro i-esimo
N_{ur}	numero di unità della popolazione con profilo u_r
Y_{rq}	valori della caratteristica di interesse per le unità con profilo u_r (q = 1, …, N_{ur})
$f_{ur,\,i}$	frequenza del profilo u_r nel centro G_i
f_{ur}	frequenza del profilo u_r nella popolazione
y_{rs}	valori campionari della caratteristica Y osservati per le unità con profilo u_r (s = 1, …, f_{ur})
$\hat{c}_{ur}$	coefficienti pesati in funzione del profilo u_r

α_i opportuno peso che rappresenta le percentuali di individui che frequentano i centri G_i

N_A generica nazionalità l-esima

$_{nat(l)}Y$ variabile dicotomica che denota la presenza della l-esima nazionalità

$$_{nat(l)}Y = \begin{cases} 1 \text{ se l'unità è dell'l-esima nazionalità} \\ 0 \text{ altrimenti} \end{cases}$$

$_AY$ variabile dicotomica che indica se l'unità l-esima è registrata

$$_AY = \begin{cases} 1 \text{ se l'unità dell'l-esima nazione è registrata} \\ 0 \text{ altrimenti} \end{cases}$$

n'_l numero di unità campionarie che appartengono alla l-esima nazionalità

n'_{Al} numero di unità della nazionalità l-esima ufficialmente registrato

$_jn'$ dimensione campionaria della j-esima macro-area

$_jn'_A$ numero di individui campionati appartenenti alla macro-area j e ufficialmente registrati

Multiple frames:

A, B liste di campionamento

N_a membri della lista A

N_b membri della lista B

N_{ab} membri di entrambe le liste

aY caratteristica di interesse associata agli elementi della lista A

bY	caratteristica di interesse associata agli elementi della lista B
$\bar{y}_a$	media campionaria di Y degli elementi che appartengono solo alla lista A
$\bar{y}_b$	media campionaria di Y degli elementi che appartengono solo alla lista B
$\bar{y}'_{ab}$	media campionaria di Y degli elementi appartenenti ad entrambe le liste, ma campionati dalla lista A
$\bar{y}''_{ab}$	media campionaria di Y degli elementi appartenenti ad entrambe le liste, ma campionati dalla lista B
F_A	inversa della frazione di campionamento per la lista A, cioè $F_A = \frac{1}{f_A}$, dove $f_A = \frac{n_a + n'_{ab}}{N_a + N_{ab}}$;
F_B	inversa della frazione di campionamento per la lista B, cioè $F_B = \frac{1}{f_B}$, dove $f_B = \frac{n_b + n''_{ab}}{N_b + N_{ab}}$;
y_a, y_b, y'_{ab}, y''_{ab}	totali campionari

Campionamento per network:

m_k	molteplicità (numerosità) del k-esimo network osservato ($1 \leq k \leq n$)
π_k	probabilità di inclusione dei network
y_k	totale della caratteristica Y rilevato sul k-esimo network esaminato

Campionamento adattivo:

y'_k	valore della caratteristica Y osservata sulla k-esima unità ($k = 1, .., n$)
h	costante fissata a priori

Snowballing:

v	numero di vertici (unità)
V	rappresenta l'insieme dei vertici, V = {1, …, v}
W	rappresenta l'insieme degli archi
V^2	rappresenta l'insieme dei lacci e degli archi
S_0	campione iniziale
x_i	variabile indicatrice dicotomica del campione iniziale
	$x_i = \begin{cases} 1 \text{ se il vertice } i \text{ è nel campione iniziale} \\ 0 \text{ altrimenti} \end{cases}$
y_{ij}	variabile indicatrice dicotomica dell'insieme degli archi
	$y_{ij} = \begin{cases} 1 \text{ se il grafico contiene un arco da } i \text{ a } j \\ 0 \text{ altrimenti} \end{cases}$
y	matrice adiacente al grafico
A_j	sottoinsieme dei vertici presenti dopo il vertice j
B_j	sottoinsieme dei vertici presenti prima del vertice j
a_j	dimensione di A_j
b_j	dimensione di B_j
S	sottoinsieme di vertici di V
A(S)	sottoinsieme dei vertici dopo gli elementi di S
B(S)	sottoinsieme dei vertici prima gli elementi di S
S_1	prima ondata di snowballing
S_2	seconda ondata di snowballing
K	numero di ondate
i.i.d.	indipendente ed identicamente distribuito
α	probabilità di inclusione di S_0
β	probabilità di inclusione degli archi (i lacci hanno probabilità di inclusione pari a 1)

r	numero degli archi che non formano un laccio
s	numero di archi formati dal campione iniziale alla prima ondata di snowballing

Metodo di cattura e ricattura:

n_{11}	numero di individui estratti da entrambe le liste di campionamento
n_{10}	numero di individui estratti solo dalla lista A
n_{01}	numero di individui estratti solo dalla lista B
n_{00}	numero di individui estratti da nessuna delle due liste di campionamento
p_{ij}	probabilità di appartenere alle rispettive categorie (i = 1, 0 e j = 1, 0)
r	numero dei diversi individui listati
n_A	numero di individui della prima lista
n_B	numero di individui della seconda lista
p_A	probabilità di appartenere alla lista A
p_B	probabilità di appartenere alla lista B
p_{AB}	probabilità di appartenere ad entrambe le liste di campionamento
$p_{B\|A}$	probabilità che un individuo appartenga alla lista B dato che appartiene già alla lista A
i_{AB}	parametro che denota la presenza o meno di interazione
u	effetto principale
u_A	variabile marginale della probabilità di apparire nella lista A
u_B	variabile marginale della probabilità di apparire nella lista B

u_{AB}	variabile marginale della probabilità di apparire in entrambe le liste di campionamento
log(L)	funzione di log-verosimiglianza
q	numero di parametri del modello preso in considerazione
a	numero di marchiati e rilasciati nella popolazione
b	numero della seconda cattura (ricattura)
c	numero di ricatturati nella seconda cattura