STATISTICA, anno LXII, n. 2, 2002

LA STIMA DELLA MEDIA NEL CAMPIONAMENTO PER CENTRI (*)

Fulvia Mecatti

1. INTRODUZIONE

La presenza straniera in Italia è attualmente argomento di crescente interesse per demografi e sociologi e, come si intuisce, la popolazione di riferimento nelle indagini in tale campo presenta caratteristiche che richiedono, in generale, tecniche di campionamento *ad hoc* più o meno discoste da quelle tradizionalmente offerte dalla teoria dei campioni da popolazioni finite.

Infatti se da un lato, nell'istante di rilevazione, la popolazione straniera si presenta composta da un numero finito N di unità statistiche, dall'altro un tale numero è solitamente ignoto; inoltre ciascuna unità non risulta identificabile mediante un'etichetta poiché mancano liste di tipo "tradizionale" che possano assumersi aggiornate ed esaustive; infine, sotto il profilo operativo, va tenuto presente che il successo dell'osservazione in tale campo è strettamente legato a garanzie di anonimato (si pensi ad esempio alla popolazione clandestina).

In tale situazione sembra particolarmente utile la tecnica, proposta da Blangiardo (1996), nota come "campionamento per centri" e nel seguito indicata con il simbolo CxC. Secondo tale tecnica il contatto con l'unità statistica, supporto delle informazioni di interesse, può avvenire mediante i così detti "centri o ambienti di aggregazione", diffusi sul territorio, che la popolazione straniera è spinta a frequentare con una certa regolarità per motivi religiosi, sociali, sanitari, per rispetto di consuetudini, o più semplicemente per le necessità della vita quotidiana.

Tali ambienti di aggregazione, o più semplicemente "centri", che si suppongono pari a K, sono identificati mediante opportuna pre-indagine; il k-esimo centro risulta composto da un ignoto numero (finito) H_k di unità statistiche che non sono identificabili e che possono, ed in generale è ciò che accade, frequentare più di un centro, $\left(K > 1;\ k = 1, \cdots, K;\ H_k \geq 1;\ \sum_{k=1}^{K} H_k \geq N \right)$.

(*) Ricerca finanziata tramite MURST COFIN99 Metodi di inferenza statistica in problemi complessi

Sotto il profilo metodologico, pertanto, i centri sono entità differenti sia dagli strati sia dai clusters ed in generale non esiste alcuna applicazione algebrica fra soggetti e centri.

Col proposito di stimare la media μ di un fenomeno quantitativo e non negativo y (eventualmente dicotòmico) presente su una popolazione avente le caratteristiche suddette, Blangiardo (1996) ha proposto una procedura che è estremamente complessa da trattare sotto il profilo metodologico ma che, ampiamente applicata in indagini sul campo, ha fornito risultati attendibili e coerenti con quanto noto (AA.VV., 1999, 2000).

Scopo del presente lavoro è un'analisi metodologica della procedura suddetta così da garantirne una giustificazione formale (paragrafo 3); oltre a ciò, viene fornita un'opportuna stima della varianza dello stimatore per μ che consente di effettuare ulteriori inferenze quali stime intervallari e verifiche di ipotesi (paragrafo 4).

Il lavoro ha come premessa una breve formalizzazione della teoria generale del CxC (paragrafo 2) e si chiude con i risultati di una simulazione condotta al calcolatore allo scopo di valutare la *performance* degli stimatori proposti (paragrafo 5).

2. TEORIA GENERALE DEL CXC

Strumento fondamentale del CxC è la matrice $\mathbf{U}$ di dimensioni $N \times K$ detta "matrice di afferenza ai centri" (Blangiardo, 1996) le cui righe, formate dalle cifre 1 e 0, prendono il nome di "profili di afferenza", o semplicemente profili, ed informano circa la struttura di frequentazione dei K centri da parte di ciascuno degli N soggetti; in particolare 1 indica la frequentazione e 0 la non frequentazione di un centro. Ad esempio, se K=4, un possibile profilo è $[1,0,0,1]$ e indica che il soggetto che lo possiede frequenta i centri contrassegnati con 1 e 4 ma non i centri contrassegnati con 2 e 3, secondo un qualche prefissato ordinamento.

Si definisce "spazio dei profili" l'insieme $\mathcal{U}$ delle $2^K - 1$ disposizioni con ripetizioni di classe K delle cifre 1 e 0 che configurano tutti i possibili profili di afferenza.

Lo spazio $\mathcal{U}$ è dunque perfettamente identificabile e poiché ogni individuo presenta uno ed un solo profilo esiste una applicazione dalle righe della matrice $\mathbf{U}$ allo spazio $\mathcal{U}$ che consente di perdere di vista i soggetti (non identificabili) e di porre l'attenzione sui profili (perfettamente identificabili).

Così, indicata con $H_{\mathbf{u}_r}$ la frequenza (assoluta) nella popolazione del generico profilo $\mathbf{u}_r \in \mathcal{U}$ si ha: $\sum_{r=1}^{2^K-1} H_{\mathbf{u}_r} = N$ e $\sum_{r=1}^{2^K-1} H_{\mathbf{u}_r} u_{rk} = H_k$, dove u_{rk} è il k-esimo elemento del profilo $\mathbf{u}_r$ e vale 1 o 0 secondo la struttura di frequentazione dei centri indicata da $\mathbf{u}_r$ medesimo. Nell'esempio precedente si ha: $u_{r1} = 1,\ u_{r2} = 0,\ u_{r3} = 0,\ u_{r4} = 1.$

Con riguardo alla tecnica di estrazione secondo lo schema CxC, in letteratura sono reperibili due procedure; la prima proposta da Blangiardo (1996) e la seconda da Migliorati (1997). La seconda, alla quale si farà riferimento nel seguito, è applicabile qualora si disponga di informazioni ausiliarie interpretabili come "misure di ampiezza" dei centri che consentano la determinazione di K interi n_k proporzionali a tali misure di ampiezza e tali che $\sum_{k=1}^{K} n_k = n$, dove n è la prefissata ampiezza campionaria. Da ciascun centro k si estraggono poi casualmente e indipendentemente n_k soggetti. supponendo che, nell'istante di rilevazione, sia "quasi certo" che nel k-esimo centro siano presenti tutti gli H_k soggetti il che può, ad esempio, realizzarsi usando l'accorgimento di effettuare la rilevazione nel momento di massima affluenza al centro, $(k = 1, \cdots, K)$. Viceversa, la presenza di tutti i soggetti è senz'altro verificata quando il centro si identifica con una lista parziale come accade, ad esempio, nel caso in cui il centro sia l'anagrafe della popolazione straniera residente.

Il CxC prevede, come parte integrante dell'indagine, che i soggetti campionati in ciascuno dei K centri forniscano, accanto alle informazioni oggetto dell'indagine stessa, anche il proprio profilo di afferenza ai rimanenti $K-1$ centri, il che sembra realizzabile e non viola le garanzie di anonimato di cui si è detto nell'introduzione. Ne segue che, una volta formato il campione, è possibile la valutazione delle seguenti frequenze campionarie.

La frequenza assoluta ${}_k f_{\mathbf{u}_r}$ del profilo $\mathbf{u}_r \in \boldsymbol{\mathcal{U}}$ fra gli n_k soggetti campionati nel centro k; la "frequenza campionaria di profilo" $f_{\mathbf{u}_r} = \sum_{k=1}^{K} {}_k f_{\mathbf{u}_r}$ e la "frequenza condizionata campionaria" $f_{m|k} = \sum_{r=1}^{2^K-1} {}_k f_{\mathbf{u}_r} u_{rm}$ dei soggetti che, essendo stati campionati in k, dichiarano di frequentare anche il generico centro $m, (m \neq k)$. Conviene osservare che, con riferimento alla generica coppia di centri k e m, può anche accadere che ${}_k f_{\mathbf{u}_r}$ sia costituita dagli stessi soggetti che danno luogo a ${}_m f_{\mathbf{u}_r}$

In Migliorati (1997) si dimostra che la frequenza (relativa) campionaria di profilo $f_{\mathbf{u}_r}/n$ è stima distorta per la corrispondente frequenza $H_{\mathbf{u}_r}/N$ nella popolazione. E ciò come conseguenza del fatto che le probabilità di inclusione dei singoli individui variano direttamente col numero di centri frequentati e inversamente in ragione dell'affollamento di ciascun centro. Così, ad esempio, potrebbe paradossalmente accadere che nella popolazione vi sia un unico soggetto con profilo $\mathbf{u}_r$ costituito da tutti "1", cioè vi sia un solo soggetto che frequenta tutti i centri. Si ha allora $H_{\mathbf{u}_r} = 1$ e $f_{\mathbf{u}_r} = K$ se detto soggetto entra a far parte di ognuno dei K campioni osservati.

Col proposito di ottenere una stima corretta del rapporto $H_{\mathbf{u}_r}/N$, in Blangiardo (1996) è fornita la seguente quantità:

$$h'_{\mathbf{u}_r} = \frac{h_{\mathbf{u}_r}}{\sum_{r=1}^{2^K-1} h_{\mathbf{u}_r}} \tag{1}$$

dove:

$$h_{\mathbf{u}_r} = f_{\mathbf{u}_r} \left[\sum_{k=1}^{K} \frac{u_{rk}\ n_g f_{g|k}}{f_{k|g}} \right]^{-1} \tag{2}$$

con g detto "centro base" che viene scelto fra i K sotto l'ipotesi: $f_{k|g} \neq 0, \forall k$ e che rimane costante per ognuno dei $2^K - 1$ profili.

Si dimostra (Migliorati, 1997) che, al divergere della numerosità campionaria del centro base $\left(n_g \to +\infty\right)$, la (1) è stima (asintoticamente) corretta e consistente del rapporto $H_{\mathbf{u}_r} / N$ mentre la (2) è stima (asintoticamente) corretta e consistente del rapporto $H_{\mathbf{u}_r} / H_g$.

3. STIMA DELLA MEDIA DELLA POPOLAZIONE

Sia $\mathbf{Y}_r = \left[Y_{r1}, \ldots, Y_{rq}, \ldots Y_{rH_{\mathbf{u}_r}} \right]$ il vettore di manifestazioni del fenomeno di interesse $\boldsymbol{\mathcal{U}}$ reperibili sugli $H_{\mathbf{u}_r}$ soggetti che nella popolazione presentano profilo $\mathbf{u}_r$; poiché ciascun soggetto presenta uno ed un solo profilo, l'insieme dei vettori $\mathbf{Y}_r$, rappresenta una partizione della popolazione e la media μ oggetto di stima risulta dall'associazione, sullo spazio dei profili $\boldsymbol{\mathcal{U}}$, delle medie $\mu_{\mathbf{u}_r}$ di profilo: $\frac{1}{N} \sum_{r=1}^{2^K-1} H_{\mathbf{u}_r} \mu_{\mathbf{u}_r} = \frac{1}{N} \sum_{r=1}^{2^K-1} \sum_{q=1}^{H_{\mathbf{u}_r}} Y_{rq} = \mu$.

Analogamente, ma con riferimento al campione, sia $\mathbf{y}_r = \left[y_{r1}, \ldots, y_{rs}, \ldots, y_{rf_{\mathbf{u}_r}} \right]$ il vettore delle osservazioni di $\boldsymbol{\mathcal{U}}$ rilevate sugli $f_{\mathbf{u}_r}$ soggetti che, fra gli n campionati, presentano profilo $\mathbf{u}_r$.

Nella versione originaria (Blangiardo, 1996) il CxC prevede di assegnare il "peso" $c_{\mathbf{u}_r} = \frac{h'_{\mathbf{u}_r}}{f_{\mathbf{u}_r} / n}$ a ciascun soggetto campionato che presenta profilo $\mathbf{u}_r$, pervenendo alla seguente stima per μ:

$$\bar{y}_{CxC} = \frac{1}{n} \sum_{r=1}^{2^K-1} c_{\mathbf{u}_r} \sum_{s=1}^{f_{\mathbf{u}_r}} y_{rs} = \sum_{r=1}^{2^K-1} h'_{\mathbf{u}_r} \frac{1}{f_{\mathbf{u}_r}} \sum_{s=1}^{f_{\mathbf{u}_r}} y_{rs} = \sum_{r=1}^{2^K-1} h'_{\mathbf{u}_r} \frac{y_{\mathbf{u}_r}}{f_{\mathbf{u}_r}} \tag{3}$$

ottenuta associando, sullo spazio dei profili $\mathcal{U}$, le medie campionarie di profilo ponderate con le stime $h'_{\mathbf{u}_r}$ dei rapporti $\frac{H_{\mathbf{u}_r}}{N}$ richiamate con la (1).

Col proposito di studiare le proprietà dello stimatore descritto dalla (3) e di stimarne la varianza, verrà ora mostrato che la (3) medesima può vedersi come "approssimazione" della stima corretta per μ costruita con metodi tradizionali sfruttando la partizione della popolazione indotta dai profili e che verrà proposta con la (5).

Ricordato che il CxC prevede l'indipendenza delle estrazioni all'interno dei centri e che lo stesso soggetto può essere estratto da più centri, sia ${}_kI_{rq}$ la variabile casuale (v.c.) indicatore della frequenza con cui il valore Y_{rq} è presente nel campione di ampiezza n_k scelto dal centro k; tale v.c., nota in letteratura come "Sample membership indicator" di Y_{rq}, (Särndal *et al.*, 1992, p.36), stanti le premesse, è Binomiale di parametri n_k e $\frac{u_{rk}}{H_k}$. Il valore atteso: $\mathrm{E}\left(\sum_{k=1}^{K} {}_kI_{rq}\right) = \sum_{k=1}^{K} \frac{n_k\, u_{rk}}{H_k} = \phi_{\mathbf{u}_r}$, noto in letteratura come "frequenza attesa di inclusione" del valore Y_{rq} nell'intero campione di ampiezza n, (Frosini et al., 1994, p.21), consente di verificare che lo stimatore:

$$\bar{Y} = \sum_{r=1}^{2^K-1} \frac{1}{N\phi_{\mathbf{u}_r}} \sum_{q=1}^{H_{\mathbf{u}_r}} Y_{rq} \sum_{k=1}^{K} {}_kI_{rq} \tag{4}$$

descritto dalla quantità:

$$\bar{y} = \sum_{r=1}^{2^K-1} \frac{1}{N\phi_{\mathbf{u}_r}} \sum_{s=1}^{f_{\mathbf{u}_r}} y_{rs}, \tag{5}$$

dove le somme risultano nulle se $f_{\mathbf{u}_r} = 0$, è corretto per μ.

Ma la (5) non è utilizzabile quale "stima" poiché funzione di quantità ignote; tuttavia, in prima approssimazione, all'ignoto rapporto $\frac{H_{\mathbf{u}_r}}{N}$ può sostituirsi $f_{\mathbf{u}_r}\left(\sum_{k=1}^{K} \frac{n_k\; u_{rk}\; N}{H_k}\right)^{-1} = f_{\mathbf{u}_r} \big/ N\phi_{\mathbf{u}_r}$, (Migliorati, 1997, eq. (3)), che a sua volta può essere stimato tramite la (1), così che un'approssimazione per l'ignoto coefficiente $\frac{1}{N\phi_{\mathbf{u}_r}}$ che figura nella (5) è il rapporto $\frac{h'_{\mathbf{u}_r}}{f_{\mathbf{u}_r}}$. Sostituendo tale rapporto nella

stessa (5), si riproduce la (3) potendosi concludere che la stima $\bar{y}_{CxC}$ è una "approssimazione" della stima corretta $\bar{y}$.

Sulla base di una tale osservazione, nel seguito l'attenzione è fissata sulla v.c. $\overline{Y}$ definita con la (4) che, avendo struttura meno complessa dello stimatore $\overline{Y}_{CxC}$ descritto dalla (3), consente la determinazione in via analitica della sua varianza e, di conseguenza, di pervenire ad una stima di questa. Ovviamente la sostituzione di $\overline{Y}$ con $\overline{Y}_{CxC}$ è tanto più "adeguata" quanto "migliore" risulta l'approssimazione di $\bar{y}$ mediante $\bar{y}_{CxC}$.

L'obiettivo è ora la determinazione della varianza $\mathrm{V}(\overline{Y})$.

Con riguardo alla v.c. ${}_kI_{rq}$, Binomiale di parametri n_k e $\frac{u_{rk}}{H_k}$, si osserva che $\sum_{r=1}^{2^K-1}\sum_{q=1}^{H_{\mathbf{u}_r}} {}_kI_{rq} = n_k$ e $\sum_{r=1}^{2^K-1}\sum_{q=1}^{H_{\mathbf{u}_r}} \frac{u_{rk}}{H_k} = \frac{1}{H_k}\sum_{r=1}^{2^K-1} H_{\mathbf{u}_r} u_{rk} = 1$; essa è pertanto una marginale di una v.c. Multinomiale a $2^K - 1 + N$ componenti. Per l'indipendenza delle estrazioni sia all'interno dei centri sia fra un centro e l'altro (dunque sia rispetto all'indice *q* che rispetto all'indice *k*), dopo alcuni laboriosi passaggi algebrici si prova che:

$$\mathrm{Cov}\left(\sum_{q=1}^{H_{\mathbf{u}_r}} Y_{rq}\ {}_kI_{rq},\ \sum_{v=1}^{H_{\mathbf{u}_t}} Y_{tv}\ {}_kI_{tv}\right) = Y_{\mathbf{u}_r} Y_{\mathbf{u}_t} \mathrm{Cov}\left({}_kI_{rq},\ {}_kI_{tv}\right) = -Y_{\mathbf{u}_r} Y_{\mathbf{u}_t} \frac{n_k\ u_{rk}\ u_{tk}}{H_k^2} \quad (6)$$

dove $Y_{\mathbf{u}_r} = \sum_{q=1}^{H_{\mathbf{u}_r}} Y_{rq}$. Si ha, allora:

$$\mathrm{V}\left(\overline{Y}\right) = \frac{1}{N^2}\sum_{k=1}^{K} \mathrm{V}\left(\sum_{r=1}^{2^K-1} \frac{1}{\phi_{\mathbf{u}_r}} \sum_{q=1}^{H_{\mathbf{u}_r}} Y_{rq}\ {}_kI_{rq}\right) =$$

$$= \frac{1}{N^2}\sum_{k=1}^{K}\left[\sum_{r=1}^{2^K-1} \frac{1}{\phi_{\mathbf{u}_r}^2} \sum_{q=1}^{H_{\mathbf{u}_r}} Y_{rq}^2\ \mathrm{V}\left({}_kI_{rq}\right) + \sum_{r=1}^{2^K-1}\sum_{\substack{t=1\\ t\neq r}}^{2^K-1} \frac{1}{\phi_{\mathbf{u}_r}\phi_{\mathbf{u}_t}} \mathrm{Cov}\left(\sum_{q=1}^{H_{\mathbf{u}_r}} Y_{rq}\ {}_kI_{rq}, \sum_{v=1}^{H_{\mathbf{u}_t}} Y_{tv}\ {}_kI_{tv}\right)\right] =$$

$$= \frac{1}{N^2}\left[\sum_{r=1}^{2^K-1} \frac{1}{\phi_{\mathbf{u}_r}^2} \sum_{q=1}^{H_{\mathbf{u}_r}} Y_{rq}^2\ A_r - \sum_{r=1}^{2^K-1}\sum_{\substack{t=1\\ t\neq r}}^{2^K-1} \frac{Y_{\mathbf{u}_r} Y_{\mathbf{u}_t}}{\phi_{\mathbf{u}_r}\phi_{\mathbf{u}_t}}\ B_{rt}\right]. \quad (7)$$

dove $A_r = \sum_{k=1}^{K} \frac{n_k\ u_{rk}}{H_k}\left(1 - \frac{u_{rk}}{H_k}\right)$ e $B_{rt} = \sum_{k=1}^{K} \frac{n_k\ u_{rk}\ u_{tk}}{H_k^2}$

4. STIMA DELLA VARIANZA DELLO STIMATORE

Dopo alcuni passaggi algebrici e tenuto sempre conto dell'indipendenza delle estrazioni fra i centri, si prova che:

$$\mathrm{E}\left[\left(\sum_{q=1}^{H_{\mathbf{u}_r}} Y_{rq}\ {}_kI_{rq}\right)\left(\sum_{v=1}^{H_{\mathbf{u}_t}} Y_{tv}\ {}_kI_{tv}\right)\right] = Y_{\mathbf{u}_r} Y_{\mathbf{u}_t}\left(\phi_{\mathbf{u}_r}\phi_{\mathbf{u}_t} - B_{rt}\right)$$

il che consente di verificare che la seguente:

$$v\left(\overline{Y}\right) = \frac{1}{N^2}\left[\sum_{r=1}^{2^K-1}\sum_{s=1}^{f_{\mathbf{u}_r}} \frac{y_{rs}^2}{\phi_{\mathbf{u}_r}^3}\ A_r - \sum_{r=1}^{2^K-1}\sum_{\substack{t=1\\ t\neq r}}^{2^K-1} \frac{y_{\mathbf{u}_r}\, y_{\mathbf{u}_t}}{\phi_{\mathbf{u}_r}\phi_{\mathbf{u}_t}}\left(\frac{B_{rt}}{\phi_{\mathbf{u}_r}\phi_{\mathbf{u}_t} - B_{rt}}\right)\right] \tag{8}$$

è "stima corretta" per $\mathrm{V}(\overline{Y})$.

Tuttavia la (8), dipendendo da quantità ignote, non è stima operativa così che, procedendo in analogia con il precedente paragrafo, è possibile approssimarla con la seguente quantità:

$$v_{CxC}\left(\overline{Y}\right) = \sum_{r=1}^{2^K-1}\sum_{s=1}^{f_{\mathbf{u}_r}} y_{rs}^2\left(\frac{h'_{\mathbf{u}_r}}{f_{\mathbf{u}_r}}\right)^2 + - \sum_{r=1}^{2^K-1}\sum_{\substack{t=1\\ t\neq r}}^{2^K-1} y_{\mathbf{u}_r}\, y_{\mathbf{u}_t}\,\frac{h'_{\mathbf{u}_r}\, h'_{\mathbf{u}_t}}{f_{\mathbf{u}_r} f_{\mathbf{u}_t}} \times$$

$$\left(\frac{\sum_{k=1}^{K} n_k\ u_{rk}\ u_{tk}\ h_{\mathbf{u}_r,k}\ h_{\mathbf{u}_t,k}}{\sum_{k=1}^{K} n_k\, u_{rk}\, h_{\mathbf{u}_r,k}\sum_{k=1}^{K} n_k\, u_{tk}\, h_{\mathbf{u}_t,k} - \sum_{k=1}^{K} n_k\ u_{rk}\ u_{tk}\ h_{\mathbf{u}_r,k}\ h_{\mathbf{u}_t,k}}\right) \tag{9}$$

dove $h_{\mathbf{u}_r,k}$ ha la stessa struttura del secondo membro della (2) che nel presente contesto assume la forma: $h_{\mathbf{u}_r,k} = f_{\mathbf{u}_r}\left[\sum_{m=1}^{K} \frac{u_{rm}\ n_k\, f_{k|m}}{f_{m|k}}\right]^{-1}$ cioè k gioca, di volta in volta, il ruolo di centro base, $(k=1,\cdots,K)$.

La (9), proposta come stima dell'ignota varianza dello stimatore $\overline{Y}_{CxC}$ descritto dalla (3), si ottiene effettuando nella (8) le seguenti approssimazioni:

$$\frac{1}{N^2\phi_{\mathbf{u}_r}^2} \approx \left(\frac{h'_{\mathbf{u}_r}}{f_{\mathbf{u}_r}}\right)^2 \text{ e } \frac{A_r}{\phi_{\mathbf{u}_r}} \approx 1 \tag{10}$$

ed essendo:

$$\frac{B_{rt}}{\phi_{\mathbf{u}_r}\phi_{\mathbf{u}_t}-B_{rt}}=\frac{\sum_{k=1}^{K}\frac{n_k\ u_{rk}\ u_{tk}}{H_k^2}}{\sum_{k=1}^{K}\frac{n_k\ u_{rk}}{H_k}\sum_{k=1}^{K}\frac{n_k\ u_{tk}}{H_k}-\sum_{k=1}^{K}\frac{n_k\ u_{rk}\ u_{tk}}{H_k^2}}=$$

$$=\frac{\sum_{k=1}^{K}n_k\ u_{rk}\ u_{tk}\frac{H_{\mathbf{u}_r}}{H_k}\frac{H_{\mathbf{u}_t}}{H_k}}{\sum_{k=1}^{K}n_k\ u_{rk}\frac{H_{\mathbf{u}_r}}{H_k}\sum_{k=1}^{K}n_k\ u_{tk}\frac{H_{\mathbf{u}_t}}{H_k}-\sum_{k=1}^{K}n_k\ u_{rk}\ u_{tk}\frac{H_{\mathbf{u}_r}}{H_k}\frac{H_{\mathbf{u}_t}}{H_k}}$$

dove ciascuno dei rapporti $\frac{H_{\mathbf{u}_r}}{H_k}$ può essere sostituito, in accordo con quanto precisato nel paragrafo 2, con la stima $h_{\mathbf{u}_r,k}$ avendosi in definitiva:

$$\frac{B_{rt}}{\phi_{\mathbf{u}_r}\phi_{\mathbf{u}_t}-B_{rt}}\approx\frac{\sum_{k=1}^{K}n_k\ u_{rk}\ u_{tk}\ h_{\mathbf{u}_r,k}\ h_{\mathbf{u}_t,k}}{\sum_{k=1}^{K}n_k\ u_{rk}\ h_{\mathbf{u}_r,k}\sum_{k=1}^{K}n_k\ u_{tk}\ h_{\mathbf{u}_t,k}-\sum_{k=1}^{K}n_k\ u_{rk}\ u_{tk}\ h_{\mathbf{u}_r,k}\ h_{\mathbf{u}_t,k}} \tag{11}$$

5. RISULTATI DI UNA SIMULAZIONE

Con lo scopo di indagare circa la *performance* degli stimatori e delle approssimazioni proposte nei precedenti paragrafi 3 e 4, e tenuto conto della complessità strutturale dei medesimi, si è proceduto con simulazioni al calcolatore che, accanto ai noti vantaggi, consentono fra l'altro di dare una valutazione circa le dimensioni campionarie necessarie affinché si realizzino gli eventuali risultati asintotici.

La situazione di partenza è stata la seguente.

La matrice $\mathbf{U}$ delle afferenze ai centri è stata prodotta casualmente come il risultato di N scelte da ciascuna di K v.c. indipendenti di Bernoulli di parametro π_k, $(k=1,\cdots,K)$. Variando i valori π_k, sotto il vincolo $\sum_{k=1}^{K}\pi_k\geq 1$, si ottengono centri più o meno affollati; un semplice test di controllo provvede all'aggiunta di un valore "1" in posizione casuale qualora una riga di $\mathbf{U}$ si presenti completamente nulla.

Il fenomeno $\boldsymbol{y}$ oggetto di attenzione è stato prodotto con N scelte casuali da una v.c. Uniforme che assume gli interi compresi fra 16 e 60, assimilabile, ad esempio, al fenomeno "Età ".

Le ampiezze campionarie n_k risultano "approssimativamente proporzionali" alle numerosità H_k dei centri poiché sono fissate proporzionalmente al risultato

della somma fra dette numerosità e una componente casuale ε determinazione della Normale Standardizzata, $(k=1,\ldots K)$.

La simulazione consiste nell'estrazione di $p=500$ campioni secondo la tecnica del secondo tipo descritta nel paragrafo 2, per diverse combinazioni di valori della numerosità N, della frazione di sondaggio n/N, del numero dei centri K e dei parametri π_k; su ciascuno dei p campioni estratti, si è proceduto al calcolo dei valori $\bar{y}$, $\bar{y}_{CxC}$, $v(\bar{Y})$ e $v_{CxC}(\bar{Y})$ impiegando, rispettivamente, le (5), (3), (8) e (9).

La media aritmetica dei p valori ottenuti per ciascuna delle suddette quantità ha fornito una stima Monte Carlo dei valori attesi dei corrispondenti stimatori.

Il programma per la simulazione è stato implementato con il codice *Mathematica 3.0* ed il valore $p=500$ è stato scelto per mantenere entro limiti ragionevoli i tempi di elaborazione; in ogni caso tale scelta si è rivelata sufficiente per gli obiettivi della simulazione.

Nel complesso sono state effettuate circa 30 simulazioni esplorando, in tal modo, una gamma sufficientemente varia di scenari. Alcuni fra i risultati più interessanti sono riportati nella Tavola 1.

TAVOLA 1

Alcuni risultati simulativi

Parametri della Simulazione	μ	Stime Monte Carlo		$V(\bar{Y})$	Stime Monte Carlo	
		$E(\bar{Y})$	$E(\bar{Y}_{CxC})$		$E[v(\bar{Y})]$	$E[v_{CxC}(\bar{Y})]$
$N=300$ $K=3$ $n/N=0.3$ $\pi_k=\{0.8,\ 0.95,\ 0.9\}$	38.723	38.770	38.798	11.553	11.605	11.607
$N=300$ $K=3$ $n/N=0.3$ $\pi_k=\{0.3,\ 0.3,\ 0.4\}$	38.637	38.700	38.681	12.697	12.737	15.086
$N=300$ $K=5$ $n/N=0.3$ $\pi_k=\{0.5, k=1\ldots K\}$	38.613	38.631	38.576	7.656	7.7372	8.754
$N=300$ $K=8$ $n/N=0.3$ $\pi_k=\{0.5, k=1\ldots K\}$	38.177	38.265	38.312	5.000	5.039	6.143
$N=800$ $K=5$ $n/N=0.3$ $\pi_k=\{0.6, 0.7, 0.8, 0.9, 0.95\}$	38.501	38.507	38.500	2.208	2.215	2.193
$N=1500$ $K=5$ $n/N=0.3$ $\pi_k=\{0.6, 0.7, 0.8, 0.9, 0.95\}$	37.753	37.735	37.738	1.709	1.707	1.710
$N=800$ $K=5$ $n/N=0.8$ $\pi_k=\{0.6, 0.7, 0.8, 0.9, 0.95\}$	38.267	38.250	38.2437	0.796	0.797	0.799

I risultati delle simulazioni mostrano che le stime y_{CxC} e $v_{CxC}(\bar{Y})$ risultano, in generale, "buone approssimazioni" dei valori di $\bar{y}$ e $v(\bar{Y})$. I valori riportati nella tabella consentono di osservare che, in media, i risultati migliori si ottengono, a parità di *K*, per valori elevati dell'insieme dei pesi π_k (si vedano le prime due righe della tabella) mentre si evidenzia una tendenza alla sovrastima all'aumentare del numero dei centri K (terza e quarta riga della tavola). Ciò pare coerente con il fatto che le stime (1) e (2) risultano tanto più utili quanto più "distorta" è la frequenza (relativa) campionaria di profilo il che avviene quanto più affollati sono i centri e più sensibili le loro sovrapposizioni; d'altro canto, l'aumento del numero dei centri considerati comporta un maggior numero di profili e dunque un maggior numero di rapporti $H_{\mathbf{u}_r}/N$ e $H_{\mathbf{u}_r}/H_g$ da stimare.

Poiché non si evidenziano variazioni significative al variare dei rimanenti parametri della simulazione, si può ritenere che la capacità approssimativa si mantiene stabile rispetto all'aumento della numerosità N (quinta e sesta riga della tabella) e della frazione di sondaggio n/N (settima riga). Ne segue, quale utile indicazione operativa, che gli stimatori proposti trovano opportuno impiego nel caso di un numero contenuto di centri ma molto affollati e con notevoli sovrapposizioni.

Sotto il profilo inferenziale, le suddette considerazioni consentono di concludere positivamente circa la correttezza delle stime $\bar{y}_{CxC}$ e $v_{CxC}(\bar{Y})$ nei riguardi dei corrispondenti parametri μ e $\mathrm{V}(\bar{Y})$ e anche circa la proprietà della *c*-consistenza (Cochran, 1977, p. 21), osservando l'effetto dell'aumento della frazione di sondaggio n/N sui valori di $v(\bar{Y})$ e $v_{CxC}(\bar{Y})$

Con riguardo alla consistenza, nei grafici che seguono sono posti a confronto i valori dei parametri μ e $\mathrm{V}(\bar{Y})$ con l'insieme dei p valori delle stime $\bar{y}_{CxC}$ e $v_{CxC}(\bar{Y})$ ottenute ponendo $K=5$, $n/N=0.3$ e $\pi_k=0.5$, $(k=1,...,K)$ e nell'ordine $N=300$, 800 e 1500 mantenendo, viceversa, fissi i valori μ e H_k/N, $(k=1,\cdots,K)$;

I grafici evidenziano una tendenza alla concentrazione dei p valori, intorno al corrispondente parametro, via via più marcata all'aumentare di N (e conseguentemente di *n*) il che giustifica una valutazione positiva circa la consistenza degli stimatori proposti.

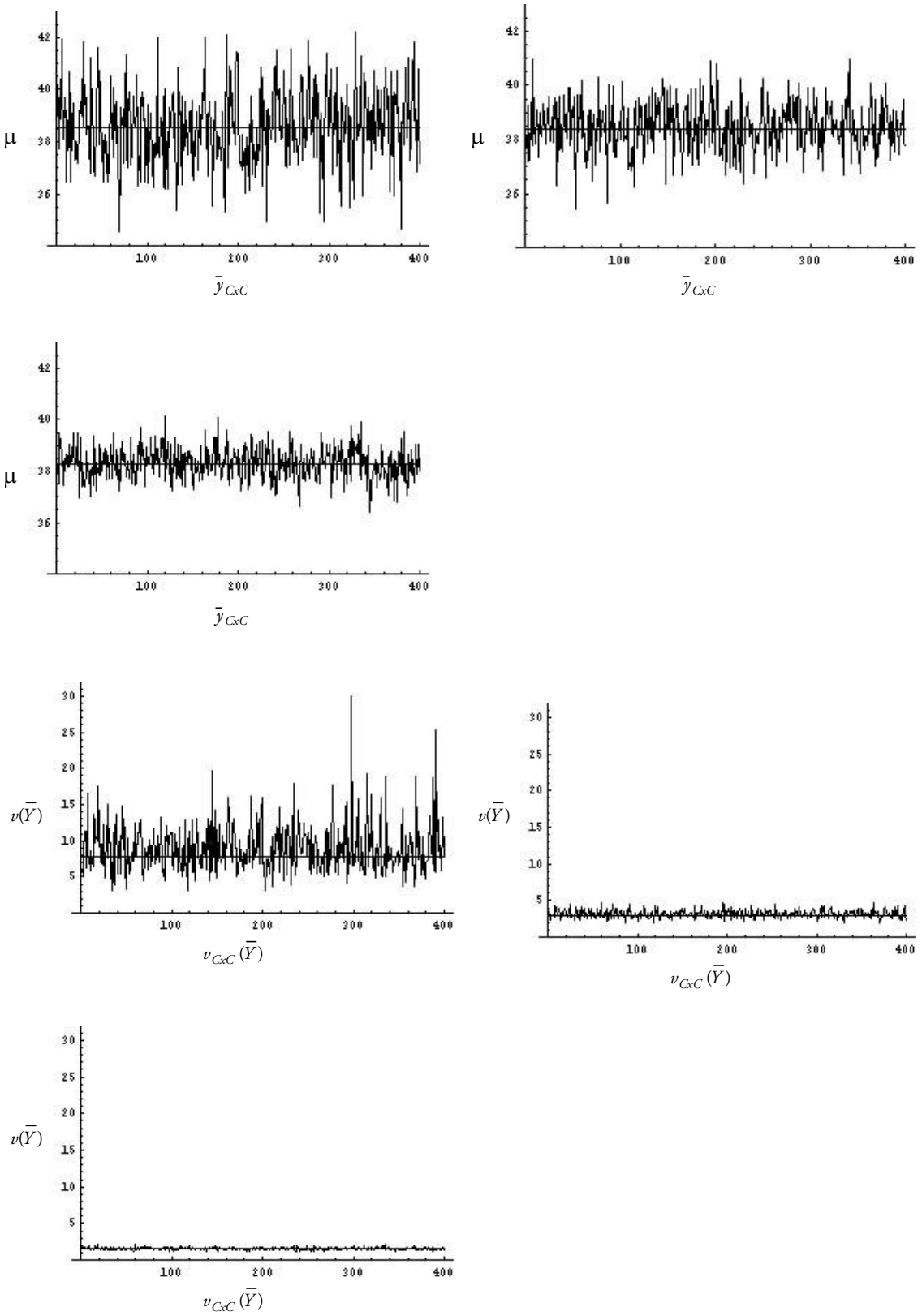

Figura 1 – Distribuzioni Monte Carlo degli stimatori $\bar{y}_{CxC}$ e $v_{CxC}(\bar{Y})$ per valori crescenti a N ed n.

Infine, dette approssimazioni sono state impiegate per costruire intervalli di confidenza per μ assumendo (previo opportuno test di adattamento) la distribuzione asintoticamente Normale dello stimatore descritto da $\bar{y}_{CxC}$. I p valori simulati hanno consentito di determinare la "copertura" di tali intervalli di confidenza come frequenza relativa degli intervalli che, fra i p costruiti, contengono μ. In tutti i casi simulati la copertura è risultata non inferiore al prefissato livello di confidenza, $(\alpha = 0.1,\ 0.05,\ 0.01)$, confermando l'attesa secondo cui l'approssimazione $v_{CxC}(\bar{Y})$, proposta nel paragrafo 4, consente la costruzione di intervalli di confidenza conservativi per l'ignoto valore di μ.

Dipartimento di Statistica
Università di Milano-Bicocca

FULVIA MECATTI

RIFERIMENTI BIBLIOGRAFICI

AA. VV. (1999), *Push and pull factors of international migration. Country report, Italy,* "European Commission, Statistical Office of the European Communities", Eurostat

AA. VV. (2000), *L'immigrazione straniera nell'area milanese 1999,* "Rapporto statistico dell'Osservatorio Fondazione Cariplo-I.S.Mu.Provincia di Milano"

G.C. BLANGIARDO (1996), *Il campionamento per centri o ambienti di aggregazione nelle indagini sulla presenza straniera*, in "Studi in onore di Giampiero Landenna", Giuffré, Milano, pp. 15-30.

W.G. COCHRAN (1977), *Sampling techniques* , 3rd ed., J. Wiley, New York.

B.V. FROSINI, M. MONTINARO E G. NICOLINI (1994), *Il campionamento da popolazioni finite*, UTET, Torino.

S. MIGLIORATI (1997), *Alcune considerazioni sul campionamento per centri*, "Statistica Applicata", 9, pp. 369-386.

C.E. SÄRNDAL, B. SWENSSON, J. WRETMAN (1992), *Model assisted survey sampling*, Springer, New York.

RIASSUNTO

La stima della media nel campionamento per centri

La tecnica nota come "campionamento per centri" è utile per indagini su popolazioni, come ad esempio la popolazione straniera irregolarmente presente in Italia, che risultano composte da unità statistiche in numero finito ma ignoto, non identificabili tramite un'etichetta e reperibili soltanto presso "centri" o "ambienti di aggregazione" diffusi sul territorio. Inoltre, ogni unità statistica può frequentare più di un centro.

Nel presente lavoro, è proposta una stima per la media di un fenomeno quantitativo presente su una popolazione avente le caratteristiche suddette, nonché una stima per la varianza del corrispondente stimatore. Sono forniti alcuni risultati di una simulazione che mettono in luce "buone" proprietà degli stimatori proposti soprattutto nel caso di un numero contenuto di centri, molto affollati e con notevoli sovrapposizioni.

SUMMARY

Estimation of the mean in the aggregation points sampling

The "aggregation points" sampling design applies, for instance, in survey of irregular immigrants i.e. of populations composed by a finite but unknown number of units which do not consent labelling and that can be reached only through a set of known but overlapping frames called "aggregation points".

Dealing with the "aggregation points" sampling design, the problem of estimating the mean of a quantitative character is concerned; an estimate of the estimator's variance is also proposed.

Some results from a simulation study are presented. Simulations indicate that estimators proposed perform better in case of not too large number of aggregation points but extensively overlapping.