CONFRONTI FRA STIMATORI PER LA MEDIA NEL CAMPIONAMENTO PER CENTRI (*)

F. Mecatti, S. Migliorati

1. INTRODUZIONE

Il Campionamento per Centri (CxC) è stato sviluppato in Italia (Blangiardo, 1996; Eurostat, 2000) nell'ambito delle indagini sulla popolazione straniera, composta da regolari e irregolari, le cui caratteristiche, in particolare la mancanza di una lista completa ed esaustiva delle unità statistiche, la loro non identificabilità nonché le evidenti esigenze di anonimato, non consentono il ricorso alle tradizionali tecniche di campionamento né alle più moderne tecniche di tipo cattura-ricattura.

I centri si identificano con luoghi naturali di aggregazione delle unità statistiche (ad esempio luoghi di culto, la stazione ferroviaria ecc…) o anche con liste parziali (ad esempio l'anagrafe della popolazione straniera residente); essi sono preidentificati in numero finito (L), si suppone coprano l'intera popolazione di numerosità N ignota e, di norma, si presentano sovrapposti nel senso che la medesima unità può frequentare più di un centro. In generale sono altresì ignote sia le numerosità N_l dei centri sia le suddette sovrapposizioni, $(l = 1, \cdots, L)$.

Col proposito di stimare la media μ di una qualche caratteristica presente su tale popolazione, si estraggono L campioni casuali semplici, di prefissata ampiezza n_l, indipendentemente da ciascun centro.

Sia $\alpha_l = N_l / N$ il "peso" dell'l-esimo centro. Tale peso può realisticamente supporsi noto, ad esempio sulla base di esperienze passate, nonostante siano ignoti i valori assoluti N ed N_l.

Sotto tale ipotesi, una stima corretta per μ è stata recentemente proposta (Mecatti e Migliorati, 2001) insieme alla varianza esatta del corrispondente stimatore e ad una stima per questa.

Una seconda stima corretta è derivabile, sotto tale ipotesi, da una precedente proposta (Mecatti, 2002).

Il presente lavoro si propone di fornire la varianza esatta dello stimatore relati-

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

vo a quest'ultima, nonché una stima della medesima di modo che sia possibile realizzare un confronto fra le due stime corrette per μ sia sotto il profilo teorico sia a livello applicativo.

In particolare nel paragrafo 2, dopo aver brevemente richiamato i fondamenti teorici del CxC ed introdotta la necessaria notazione, sono presentate le due stime suddette e riassunti alcuni risultati già apparsi in letteratura. Nel paragrafo 3 è fornita la forma esatta della varianza del secondo stimatore nonché una sua stima. I due stimatori sono discussi e confrontati nel paragrafo 4. Infine, il paragrafo 5 è dedicato all'implementazione ed alla descrizione dei risultati di una simulazione avente lo scopo di confrontare i due stimatori sia con riferimento alle proprietà inferenziali sia sotto il profilo operativo. Le dimostrazioni sono raccolte in appendice.

2. NOTAZIONE E STIME DELLA MEDIA NEL CxC

Sia Y la caratteristica quantitativa di interesse e sia $\{Y_1,\cdots,Y_N\}$ il "parametro" della popolazione se le unità statistiche fossero identificabili tramite un'etichetta (Cassel *et al.*, 1977, p. 6).

Nell'ambito del CxC le unità non sono identificabili e in generale non esiste alcuna applicazione fra soggetti e centri. Un tale problema può essere superato affiancando ai due insiemi "teorici" etichette e parametro della popolazione, la matrice "reale" **U** di dimensioni $N \times L$, detta "matrice di afferenza ai centri" la cui generica cella *il*-esima $(i=1,\cdots,N; l=1,\cdots,L)$ è 1 se l'*i*-esimo soggetto frequenta l'*l*-esimo centro ed è 0 in caso contrario di modo che la somma delle colonne di **U** riproduce la numerosità N_l dei centri. Si noti che in generale risulta $\sum_{l=1}^{L} N_l \geq N$ a causa delle sovrapposizioni fra centri. Le righe di **U** sono dette "profili" ed informano circa la struttura di afferenza agli L centri da parte di ciascun soggetto.

Indicato con $\mathbf{u}_r$ $(r=1,\cdots,2^L-1)$ il generico profilo ovvero una delle 2^L-1 disposizioni con ripetizione delle cifre 0 e 1, ad esclusione dell'L-upla nulla, e con u_{rl} il generico elemento di tale profilo, $(u_{rl}=0,1; l=1,\cdots,L)$, siano $N_{\mathbf{u}_r}$ il numero di soggetti che nella popolazione di riferimento presentano profilo $\mathbf{u}_r$ e $\{Y_{rq}; q=1,\cdots,N_{\mathbf{u}_r}\}$ il sottoinsieme del parametro della popolazione con riferimento ai soli soggetti che nella popolazione presentano profilo $\mathbf{u}_r$.

Essendo $\sum_{r=1}^{2^L-1} N_{\mathbf{u}_r} = N$ e poiché gli insiemi $\{Y_{rq}; q=1,\cdots,N_{\mathbf{u}_r}\}$ formano una partizione del parametro della popolazione, il ricorso ai profili consente l'eliminazione delle sovrapposizioni così che alla quantità μ oggetto di stima può darsi la seguente rappresentazione:

$$\mu = \frac{1}{N} \sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} Y_{rq} .$$

Un'ulteriore quantità d'interesse è la "molteplicità" dell'*r*-esimo profilo $m_r = \sum_{l=1}^{L} u_{rl}$ vale a dire il numero di centri frequentati dai soggetti con profilo $\mathbf{u}_r$ (Thompson, 1992, p. 149).

Il CxC prevede di estrarre un campione di prefissata ampiezza da ciascun centro. Ne segue che un disegno di campionamento realistico consiste nell'estrazione di un campione casuale semplice (senza reinserimento) di ampiezza n_l scelto nell'*l*-esimo centro in cui si suppone che tutti gli N_l soggetti siano presenti "quasi certamente". E ciò può realizzarsi effettuando l'estrazione indipendentemente in ciascuno degli *L* centri nel momento di massimo affollamento, mentre una tale ipotesi è certamente verificata nel caso di liste parziali.

Su ciascun soggetto estratto sono rilevati contemporaneamente il valore di *Y* e il profilo.

Con riferimento al campione dal centro *l*, siano $f_{\mathbf{u}_r,l}$ la frequenza campionaria del profilo $\mathbf{u}_r$ e $\{y_{rs}; s = 1, \cdots, f_{\mathbf{u}_r,l}\}$ l'insieme dei valori osservati sui soli soggetti con profilo $\mathbf{u}_r$ che, quando *r* varia da 1 a $2^L - 1$, costituiscono una partizione del campione dall'*l*-esimo centro.

Ciò premesso, la seguente è stima corretta per μ (Mecatti e Migliorati, 2001):

$$\overline{y}_m = \sum_{l=1}^{L} \frac{\alpha_l}{n_l} \sum_{r=1}^{2^L-1} \frac{y_{r,l}}{m_r} = \sum_{l=1}^{L} \sum_{r=1}^{2^L-1} \frac{y_{r,l}}{\beta_{rl}} \tag{1}$$

dove: $y_{r,l} = \sum_{s=1}^{f_{\mathbf{u}_r,l}} y_{rs}$ è il totale campionario relativo all'*r*-esimo profilo e al campione dal centro *l* e $\beta_{rl} = \frac{n_l m_r}{\alpha_l} = \frac{n_l}{\alpha_l} \sum_{l=1}^{L} u_{rl}$.

Lo stimatore $\overline{Y}_m$ descritto dalla (1), ha la seguente varianza esatta:

$$\mathrm{V}(\overline{Y}_m) = \sum_{l=1}^{L} \frac{n_l (N_l - n_l)}{N_l (N_l - 1)} \left[\sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y_{rq}^2 u_{rl}}{\beta_{rl}^2} - \frac{1}{N_l} \left(\sum_{r=1}^{2^L-1} \frac{Y_r u_{rl}}{\beta_{rl}} \right)^2 \right] \tag{2}$$

dove $Y_r = \sum_{q=1}^{N_{\mathbf{u}_r}} Y_{rq}$ rappresenta il totale dell'*r*-esimo profilo.

Una semplice stima per la (2) è la seguente:

$$\hat{v}(\overline{Y}_m) = \sum_{l=1}^{L} \frac{1}{(n_l - 1)} \left[n_l \sum_{r=1}^{2^L-1} \sum_{s=1}^{f_{\mathbf{u}_r,l}} \frac{y_{rs}^2}{\beta_{rl}^2} - \left(\sum_{r=1}^{2^L-1} \frac{y_{r,l}}{\beta_{rl}} \right)^2 \right] \tag{3}$$

che risulta corretta a meno dei fattori di correzione per popolazioni finite $\left(1 - \frac{n_l}{N_l}\right)$, $(l = 1, \cdots, L)$, e risulta essere, pertanto, di tipo conservativo.

Sotto l'ipotesi α_l noto, $(l = 1, \cdots, L)$, con riferimento ad una precedente proposta (Mecatti, 2002), è altresì corretta per μ la seguente stima:

$$\overline{y}_c = \sum_{l=1}^{L} \sum_{r=1}^{2^L-1} \frac{1}{\lambda_r} \sum_{s=1}^{f_{\mathbf{u}_r,l}} y_{rs} = \sum_{l=1}^{L} \sum_{r=1}^{2^L-1} \frac{y_{r,l}}{\lambda_r} \tag{4}$$

dove $\lambda_r = \sum_{l=1}^{L} \frac{n_l u_{rl}}{\alpha_l}$.

Nel successivo paragrafo è fornita la forma esatta della varianza dello stimatore $\overline{Y}_c$ descritto dalla (4) ed è altresì proposta una opportuna stima per detta varianza.

3. VARIANZA DI $\overline{Y}_c$ E STIMA DI $V(\overline{Y}_c)$

Sia $\delta_{rq,l}$ la variabile casuale indicatore del valore Y_{rq} nel campione dall'l-esimo centro $(l = 1, \cdots, L)$; sotto il disegno di campionamento assunto si ha che il vettore $[\delta_{rq,l};\ r = 1, \cdots, 2^L - 1;\ q = 1, \cdots, N_{u_r}]$ ha distribuzione Multi-Ipergeometrica con parametri $\left(n_l, \frac{u_{rl}}{N_l}\right)$. Qui di seguito vengono fornite le espressioni di $\mathrm{V}(\overline{Y}_c)$ e della corrispondente stima rinviando all'appendice le relative dimostrazioni. In particolare risulta:

$$\mathrm{V}(\overline{Y}_c) = \sum_{l=1}^{L} V\left(\sum_{r=1}^{2^L-1} \frac{1}{\lambda_r} \sum_{q=1}^{N_{u_r}} Y_{rq}\, \delta_{rq,l} \right) = \sum_{l=1}^{L} \frac{n_l (N_l - n_l)}{N_l (N_l - 1)} \left[\sum_{r=1}^{2^L-1} \sum_{q=1}^{N_{\mathbf{u}_r}} \frac{Y_{rq}^2 u_{rl}}{\lambda_r^2} - \frac{1}{N_l} \left(\sum_{r=1}^{2^L-1} \frac{Y_r u_{rl}}{\lambda_r} \right)^2 \right] \tag{5}$$

Col proposito di ottenere una stima per $\mathrm{V}(\overline{Y}_c)$, si osservi che la seguente quantità:

$$\sum_{l=1}^{L} \frac{1}{n_l - 1} \left(1 - \frac{n_l}{N_l}\right) \left[n_l \sum_{r=1}^{2^L-1} \sum_{s=1}^{f_{\mathbf{u}_r,l}} \frac{y_{rs}^2}{\lambda_r^2} - \left(\sum_{r=1}^{2^L-1} \frac{y_{r,l}}{\lambda_r} \right)^2 \right] \tag{6}$$

è corretta per $\mathrm{V}(\overline{Y}_c)$ come emergerà nell'appendice; dalla (6) segue che la seguente stima per $\mathrm{V}(\overline{Y}_c)$:

$$\hat{v}(\overline{Y}_c) = \sum_{l=1}^{L} \frac{1}{n_l - 1} \left[n_l \sum_{r=1}^{2^L-1} \sum_{s=1}^{f_{\mathbf{u}_r,l}} \frac{y_{rs}^2}{\lambda_r^2} - \left(\sum_{r=1}^{2^L-1} \frac{y_{r,l}}{\lambda_r} \right)^2 \right] \tag{7}$$

è corretta a meno dei fattori di correzione per popolazioni finite $\left(1 - \frac{n_l}{N_l}\right)$, $(l = 1, \cdots, L)$, presentandosi, pertanto, come stima conservativa.

4. CONFRONTI FRA I DUE STIMATORI $\overline{Y}_m$ E $\overline{Y}_c$

Dalle definizioni (1) e (4) si evince che entrambi gli stimatori $\overline{Y}_m$ e $\overline{Y}_c$ si presentano come combinazioni lineari dei totali campionari di profilo degli L campioni (indipendenti) di centro ma utilizzano pesi di profilo differenti. In particolare, in $\overline{Y}_m$ ogni profilo ha un peso specifico $\beta_{rl} = \frac{n_l m_r}{\alpha_l}$ secondo il centro in cui è stato osservato, mentre in $\overline{Y}_c$ a ciascun profilo è assegnato un peso "medio" $\lambda_r = \sum_{l=1}^{L} \frac{n_l u_{rl}}{\alpha_l}$ costante per ogni centro. I due stimatori differiscono, pertanto, per la logica di sintesi delle informazioni campionarie di centro.

E' immediato osservare che $\overline{Y}_m \equiv \overline{Y}_c$ se e solo se $\frac{n_l}{\alpha_l} \propto \frac{n_l}{N_l} = k, (l = 1, \ldots L)$, ovvero se la frazione di campionamento è costante in ognuno degli L centri, i due stimatori coincidono mentre forniscono stime differenti in ogni altro caso.

Essendo, per ipotesi, α_l variabili esogene, indicata con $n = \sum_{l=1}^{L} n_l$ l'ampiezza campionaria complessiva, un confronto fra i due stimatori può basarsi su un'allocazione di n fra gli L centri diversa dall'allocazione proporzionale.

In Mecatti e Migliorati (2001) è proposta una scomposizione della varianza di Y che consente di esprimere $V(\overline{Y}_m)$ in funzione della sola variabilità interna ai centri opportunamente "corretta con le molteplicità" di modo che risulta individuabile l'allocazione ottima sia sotto la semplice funzione di costo $n = \sum_{l=1}^{L} n_l$ sia sotto funzioni di costo più complesse, ad esempio lineare $c = c_0 + \sum_{l=1}^{L} c_l n_l$ dove c_0 rappresenta i costi fissi di indagine e c_l indica il costo per unità osservata nell'l-esimo centro.

Viceversa, la struttura di $\overline{Y}_c$ non consente una analoga espressione di $\mathrm{V}(\overline{Y}_c)$ in funzione della variabilità interna ai centri, di modo che il problema: $\min_{n_1,\cdots,n_L} \mathrm{V}(\overline{Y}_c)$, tanto sotto il vincolo di costo semplice quanto sotto il vincolo di costo lineare, non ha soluzione analitica.

Una prima conclusione generale è, pertanto, la seguente: qualora si disponga di informazioni a priori circa la variabilità interna ai centri, provenienti ad esempio da indagini precedenti, $\overline{Y}_m$ risulta preferibile a $\overline{Y}_c$ sotto il profilo dell'allocazione ottima dell'ampiezza campionaria complessiva fra gli L centri.

E' importante notare che sono rilevabili svariate analogie fra il CxC ed il problema noto come *multiple frame* (Mecatti e Migliorati, 2001) nel caso particolare in cui i centri si identificano con liste incomplete e sovrapposte. In tale ambito e ponendo $L = 2$ (*dual frame*), $\overline{Y}_m$ ha struttura che concorda con la stima proposta da Hartley (1962, 1974) mentre $\overline{Y}_c$ coincide con la stima proposta da Lund (1968). Con riferimento al caso generale $(L \geq 2)$, l'attenzione è ora fissata sul confronto fra i due stimatori, entrambi corretti, sotto il profilo dell'efficienza.

E' immediato verificare che non esiste una relazione d'ordine fra i pesi β_{rl} e λ_r uniforme per ogni centro e per ogni profilo, inoltre un confronto analitico fra le corrispondenti varianze $\mathrm{V}(\overline{Y}_m)$ e $\mathrm{V}(\overline{Y}_c)$ non consente di individuare fra i due lo stimatore uniformemente più efficiente né esiste la condizione analitica esatta sotto la quale l'uno risulta relativamente più efficiente dell'altro.

Interessanti considerazioni generali possono, tuttavia, essere tratte esplorando il problema per via simulativa.

5. IMPLEMENTAZIONE DELLA SIMULAZIONE E PRINCIPALI RISULTATI

Col proposito di confrontare i due stimatori $\overline{Y}_m$ e $\overline{Y}_c$ sia sotto il profilo delle proprietà inferenziali sia sotto quello operativo si è proceduto con simulazioni che consentono, fra l'altro, di valutare le dimensioni campionarie necessarie affinché si realizzino i risultati asintotici.

Con riguardo all'implementazione, realizzata mediante il codice *Mathematica 3.0*, la situazione di partenza è la seguente:

- La matrice **U** di afferenza ai centri è prodotta casualmente come il risultato di N scelte da ciascuna di L variabili casuali indipendenti di Bernoulli di parametro π_l, $(l = 1,\cdots,L)$, eliminando le eventuali righe nulle; i valori assegnati ai parametri π_l sotto il vincolo $\sum_{l=1}^{L} \pi_l \geq 1$, consentono di controllare il grado di affollamento degli L centri.
- La caratteristica Y di interesse è prodotta mediante N scelte casuali da una variabile casuale con opportuna distribuzione, ad esempio Uniforme discreta sugli interi da 16 a 60 assimilabile al fenomeno "età".

– Fissata la numerosità campionaria totale *n*, le ampiezze campionarie di centro n_l sono prodotte secondo tre differenti criteri di allocazione non proporzionale: 1) allocazione costante $n_l = N/L$ con opportuna approssimazione nel caso non intero, 2) allocazione ottima semplice per lo stimatore $\overline{Y}_m$, 3) allocazione ottima con funzione di costo lineare per lo stimatore $\overline{Y}_m$. Con tali numerosità sono calcolate le due varianze esatte $\mathrm{V}(\overline{Y}_m)$ e $\mathrm{V}(\overline{Y}_c)$ impiegando rispettivamente la (2) e la (5).
– La simulazione consiste nell'estrazione di p campioni casuali semplici senza reinserimento secondo la tecnica descritta nel paragrafo 2, vale a dire indipendentemente da ciascuno degli L centri e secondo le 3 allocazioni suddette. Su ciascuno di tali campioni si sono calcolate le stime $\overline{y}_m$, $\overline{y}_c$, $\hat{v}(\overline{Y}_m)$ e $\hat{v}(\overline{Y}_c)$ impiegando, rispettivamente, le (1), (4), (3) e (7). La media dei *p* valori per ciascuna di dette quantità fornisce una stima Monte Carlo dei valori attesi dei corrispondenti stimatori.
– Nel complesso sono state realizzate circa 50 simulazioni per diverse combinazioni di valori della numerosità *N*, della frazione di campionamento n/N, del numero dei centri *L*, dei parametri π_l e della funzione di costo.

L'esercizio simulativo si è articolato in due *step* successivi:

1. Analisi dell'efficienza relativa dei due stimatori corretti $\overline{Y}_m$ e $\overline{Y}_c$ mediante il confronto delle due varianze esatte nei diversi scenari considerati.
2. Analisi delle proprietà delle stime per le suddette varianze $\hat{v}(\overline{Y}_m)$ e $\hat{v}(\overline{Y}_c)$, proposte rispettivamente con la (3) e la (7), mediante l'impiego dei *p* valori simulati per ciascuno scenario.

La tabella 1 riporta alcuni dei risultati più interessanti riguardo allo *step* 1.

La riga (I) della tabella 1 mostra le varianze dei due stimatori nel caso di centri ugualmente affollati (π_l costanti): qualora l'allocazione della numerosità campionaria fra i centri conservi una situazione prossima alla costanza i due stimatori tendono a coincidere (prima e seconda colonna) mentre, in caso contrario (terza colonna) $\overline{Y}_m$ risulta più efficiente $\overline{Y}_c$.

Nella riga (II) in cui i parametri π_l sono ben differenziati dando luogo a centri diversamente affollati, è confermata la maggiore efficienza di $\overline{Y}_m$ rispetto a $\overline{Y}_c$ qualunque sia l'allocazione scelta sebbene l'allocazione ottima per il primo riduca sensibilmente anche la varianza del secondo (seconda e terza colonna).

La riga (III) evidenzia l'effetto dell'aumento del numero dei centri: da un lato un aumento nella variabilità di entrambi gli stimatori rilevabile solo nel caso di allocazione costante (prima colonna) poiché negli altri casi la variabilità degli stimatori viene a dipendere dalla variabilità interna ai centri; dall'altro una inversione di tendenza nell'efficienza relativa fra i due stimatori poiché $\overline{Y}_c$ risulta più efficiente di $\overline{Y}_m$ nel caso di allocazione ottima per quest'ultimo (seconda e terza colonna)

TABELLA 1

Efficienza relativa dei due stimatori corretti $\overline{Y}_m$ *e* $\overline{Y}_c$

Parametri della simulazione	Allocazione costante		Allocazione ottima semplice		Allocazione ottima con costi lineari	
$N=800$ $L=3$ $n/N=0.15$ $\pi_l=\{0.5, 0.5, 0.5\}$	$\frac{n_l}{N_l}=\{0.09, 0.1, 0.1\}$		$\frac{n_l}{N_l}=\{0.09, 0.1, 0.09\}$		$\frac{n_l}{N_l}=\{0.13, 0.1, 0.06\}$ $c_l=\{4, 2, 1\}$	
(I)	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$
	3.37	3.37	3.38	3.39	4.40	5.00
$N=800$ $L=3$ $n/N=0.15$ $\pi_l=\{0.05, 0.5, 0.9\}$	$\frac{n_l}{N_l}=\{0.69, 0.1, 0.06\}$		$\frac{n_l}{N_l}=\{0.1, 0.06, 0.12\}$		$\frac{n_l}{N_l}=\{0.06, 0.05, 0.13\}$ $c_l=\{4, 2, 1\}$	
(II)	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$
	3.94	5.21	2.26	2.33	4.06	4.55
$N=800$ $L=5$ $n/N=0.15$ $\pi_l=\{0.05, 0.1, 0.15, 0.85, 0.95\}$	$\frac{n_l}{N_l}=\{0.62\ 0.28, 0.19, 0.04, 0.03\}$		$\frac{n_l}{N_l}=\{0.026, 0.03, 0.05, 0.06, 0.09\}$		$\frac{n_l}{N_l}=\{0.03, 0.08, 0.06, 0.09, 0.06\}$ $c_l=\{2, 1, 2, 2, 10\}$	
(III)	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$
	4.78	7.55	2.64	1.98	2.02	1.92
	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$
$N=1500$ $L=3$ $n/N=0.15$ $\pi_l=\{0.05, 0.5, 0.9\}$	$\frac{n_l}{N_l}=\{0.84, 0.1, 0.06\}$		$\frac{n_l}{N_l}=\{0.11, 0.07, 0.12\}$		$\frac{n_l}{N_l}=\{0.07, 0.05, 0.13\}$ $c_l=\{4, 2, 1\}$	
(IV)	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$
	2.25	2.94	1.17	1.21	1.73	2.01
$N=1500$ $L=3$ $n/N=0.4$ $\pi_l=\{0.05, 0.5, 0.9\}$	$\frac{n_l}{N_l}=\{1, 0.32, 0.18\}$		$\frac{n_l}{N_l}=\{0.32, 0.16, 0.33\}$		$\frac{n_l}{N_l}=\{0.17, 0.12, 0.36\}$ $c_l=\{4, 2, 1\}$	
(V)	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$	$V(\overline{Y}_m)$	$V(\overline{Y}_c)$
	0.57	0.72	0.44	0.44	0.60	0.66

mantenendosi, viceversa, meno efficiente nel caso di allocazione costante (prima colonna).

Nella riga (IV) è stata aumentata la numerosità N della popolazione a parità di frazione di campionamento n/N con il conseguente aumento anche della numerosità campionaria n; la sensibile riduzione delle varianze di entrambi gli stimatori

corretti conferma la consistenza dei medesimi così come, d'altro canto, si evince dalle (2) e (5).

Infine la riga (V) mostra l'effetto dell'aumento della frazione di campionamento n/N a parità di numerosità N della popolazione: la notevole contrazione di entrambe le varianze consente di concludere positivamente anche circa la *c*-consistenza (Cochran, 1977, p. 21) dei due stimatori.

I risultati in tabella 1 confermano la non esistenza dello stimatore uniformemente più efficiente fra $\overline{Y}_m$ e $\overline{Y}_c$ ma suggeriscono, altresì, la seguente considerazione: come già osservato nel paragrafo 4 ciò che discrimina i due stimatori sono i pesi $\frac{n_l}{\alpha_l} \propto \frac{n_l}{N_l}$ avendosi, in particolare, che se le frazioni di campionamento di centro risultano fra loro "molto" differenziate $\overline{Y}_m$ risulta più efficiente di $\overline{Y}_c$ mentre accade il viceversa se le frazioni di campionamento di centro risultano fra loro "poco" differenziate. Il grafico in figura 1 conferma una tale osservazione riportando sulle ascisse una misura di variabilità delle frazioni di campionamento di centro n_l/N_l e sulle ordinate la corrispondente efficienza relativa $\mathrm{V}(\overline{Y}_c)/\mathrm{V}(\overline{Y}_m)$ per 25 differenti casi (a parità di *N*, di *n* e di μ):

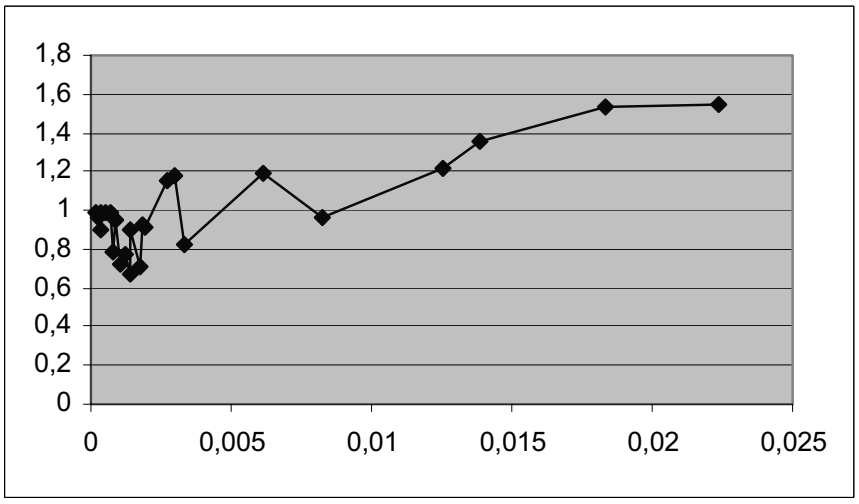

Figura 1 – Efficienza relativa dei due stimatori corretti $\overline{Y}_m$ e $\overline{Y}_c$ all'aumentare della variabilità delle frazioni di campionamento di centro n_l/N_l .

Si osserva che l'efficienza relativa $\mathrm{V}(\overline{Y}_c)/\mathrm{V}(\overline{Y}_m)$ partendo da 1 in corrispondenza di variabilità nulla fra le frazioni di campionamento di centro $n_l/N_l, (l=1,\cdots L)$, corrispondente all'allocazione proporzionale, risulta inferiore all'unità fintanto che la variabilità fra le frazioni di campionamento di centro è "piccola" mentre tende a crescere, dopo un periodo di assestamento, mantenendosi costantemente sopra l'unità per valori via via crescenti di detta variabilità. Il medesimo andamento si osserva altresì su insiemi maggiori di casi e per livelli di

variabilità per le frazioni di campionamento di centro superiori a quelle rappresentate nella figura 1. Ciò risulta coerente col fatto che, come già osservato nel paragrafo 4, in $\overline{Y}_c$ le informazioni campionarie di centro sono sintetizzate mediante un "peso medio" funzione delle frazioni di campionamento n_l/N_l mentre $\overline{Y}_m$ utilizza "pesi" diversi e specifici per ciascun centro.

Con riguardo allo *step* 2, in tabella 2 sono riportati i risultati di alcune fra le numerose simulazioni eseguite:

TABELLA 2

Proprietà delle stime $\hat{v}(\overline{Y}_m)$ *e* $\hat{v}(\overline{Y}_c)$

Parametri della Simulazione	Medie			$\overline{Y}_m$		$\overline{Y}_c$	
(I)							
$N=800;\ L=3$	μ	$\mathrm{E}(\overline{Y}_m)$	$\mathrm{E}(\overline{Y}_c)$	$\mathrm{V}(\overline{Y}_m)$	$\mathrm{E}[\hat{v}(\overline{Y}_m)]$	$\mathrm{V}(\overline{Y}_c)$	$\mathrm{E}[\hat{v}(\overline{Y}_c)]$
$n/N=0.15$							
$\pi_l=\{0.5,\ 0.5,\ 0.5\}$							
$\frac{n_l}{N_l}=\{0.14,\ 0.12,\ 0.24\}$	38.64	38.73	38.72	1.15	1.56	1.16	1.46
(II)							
$N=800;\ L=5$	μ	$\mathrm{E}(\overline{Y}_m)$	$\mathrm{E}(\overline{Y}_c)$	$\mathrm{V}(\overline{Y}_m)$	$\mathrm{E}[\hat{v}(\overline{Y}_m)]$	$\mathrm{V}(\overline{Y}_c)$	$\mathrm{E}[\hat{v}(\overline{Y}_c)]$
$n/N=0.15$							
$\pi_l=\{0.5,\ 0.5,\ 0.5,\ 0.5,\ 0.5\}$							
$\frac{n_l}{N_l}=\{0.058,\ 0.057,\ 0.059,\ 0.062,\ 0.063\}$	38.04	37.94	37.95	4.43	4.59	4.45	4.62
(III)							
$N=200;\ L=3$	μ	$\mathrm{E}(\overline{Y}_m)$	$\mathrm{E}(\overline{Y}_c)$	$\mathrm{V}(\overline{Y}_m)$	$\mathrm{E}[\hat{v}(\overline{Y}_m)]$	$\mathrm{V}(\overline{Y}_c)$	$\mathrm{E}[\hat{v}(\overline{Y}_c)]$
$n/N=0.15$							
$\pi_l=\{0.5,\ 0.5,\ 0.5\}$							
$\frac{n_l}{N_l}=\{0.089,\ 0.103,\ 0.096\}$	38.01	37.66	37.65	11.94	13.62	12.08	13.75
(IV)							
$N=2000;\ L=3$	μ	$\mathrm{E}(\overline{Y}_m)$	$\mathrm{E}(\overline{Y}_c)$	$\mathrm{V}(\overline{Y}_m)$	$\mathrm{E}[\hat{v}(\overline{Y}_m)]$	$\mathrm{V}(\overline{Y}_c)$	$\mathrm{E}[\hat{v}(\overline{Y}_c)]$
$n/N=0.15$							
$\pi_l=\{0.5,\ 0.5,\ 0.5\}$							
$\frac{n_l}{N_l}=\{0.093,\ 0.093,\ 0.088\}$	38.14	38.11	38.11	1.39	1.52	1.39	1.52
(V)							
$N=800;\ L=3$	μ	$\mathrm{E}(\overline{Y}_m)$	$\mathrm{E}(\overline{Y}_c)$	$\mathrm{V}(\overline{Y}_m)$	$\mathrm{E}[\hat{v}(\overline{Y}_m)]$	$\mathrm{V}(\overline{Y}_c)$	$\mathrm{E}[\hat{v}(\overline{Y}_c)]$
$n/N=0.6$							
$\pi_l=\{0.5,\ 0.5,\ 0.5\}$							
$\frac{n_l}{N_l}=\{0.375,\ 0.383,\ 0.376\}$	38.1	38.14	38.14	0.603	0.969	0.604	0.968

I risultati riportati in tabella 2 si riferiscono ad un'unica allocazione della numerosità campionaria *n*, in particolare l'allocazione ottima per $\overline{Y}_m$ con funzione di costo semplice, di modo che le ampiezze campionarie n_l di centro, ovvero la variabilità fra le frazioni di campionamento di centro $n_l/N_l, (l=1,\cdots L)$, vengono a dipendere dalla variabilità interna ai centri medesimi. In tal modo si è prodotta una sufficiente varietà di scenari anche mantenendo costante il valore assegnato ai

parametri π_l che risulta, comunque, scarsamente influente sulle proprietà delle stime $\hat{v}(\overline{Y}_m)$ e $\hat{v}(\overline{Y}_c)$.

Per ciascuno scenario prospettato, ovvero ciascuna riga della tabella 2, i valori della media μ confrontati con le stime Monte Carlo dei valori attesi dei due stimatori corretti $\mathrm{E}(\overline{Y}_m)$ e $\mathrm{E}(\overline{Y}_c)$ (seconda colonna) consentono una valutazione positiva circa la bontà della simulazione condotta, in merito al generatore di numeri pseudo-casuali impiegato nonché con riguardo alla scelta del numero p=500 delle iterazioni di cui si compone la simulazione.

Nella seconda e terza colonna sono posti a confronto le stime Monte Carlo dei valori attesi degli stimatori descritti da $\hat{v}(\overline{Y}_m)$ e $\hat{v}(\overline{Y}_c)$ con il corrispondente oggetto di stima $\mathrm{V}(\overline{Y}_m)$ e $\mathrm{V}(\overline{Y}_c)$ di modo che è possibile trarre conclusioni in merito alla loro correttezza.

In particolare, dalla riga (I) si evince che già con un'ampiezza campionaria pari n=120 si ottengono stime pressoché corrette mentre l'aumento del numero dei centri (riga (II)) non sembra produrre effetti sulla correttezza delle stime medesime.

Nelle righe (III) e (IV) si mettono in evidenza gli effetti di una drastica variazione della numerosità N della popolazione, a parità di frazione di campionamento n/N, come conseguenza del fatto che nelle (3) e (7) sono trascurati i fattori di correzione per popolazioni finite $(1-n_l/N_l),(l=1,\cdots L)$: la distorsione può essere sensibile per popolazioni poco numerose (riga III) mentre tende a scomparire per popolazioni più numerose (riga IV).

Infine, l'aumento della frazione di campionamento a parità di numerosità della popolazione (riga (V)) ha l'effetto di contrarre sensibilmente tanto le varianze $\mathrm{V}(\overline{Y}_m)$ e $\mathrm{V}(\overline{Y}_c)$ quanto le rispettive stime $\hat{v}(\overline{Y}_m)$ e $\hat{v}(\overline{Y}_c)$.

Per concludere, con riguardo alla consistenza delle stime $\hat{v}(\overline{Y}_m)$ e $\hat{v}(\overline{Y}_c)$, nei grafici che seguono (figure 2 e 3) sono posti a confronto i valori reali delle varianze $\mathrm{V}(\overline{Y}_m)$ e $\mathrm{V}(\overline{Y}_c)$ con insiemi di p=250 valori simulati delle suddette stime ottenute ponendo L=3, n/N=0.15 e $\pi_l = 0.5, (l=1,\cdots,L)$ e nell'ordine N=400, 800, 1600 restando, viceversa, fisso il valore del parametro μ:

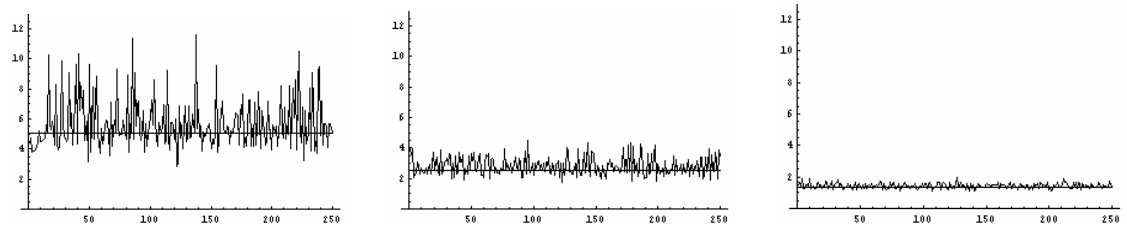

Figura 2 – p=*250* valori simulati della stima $\hat{v}(\overline{Y}_m)$ rispetto a $\mathrm{V}(\overline{Y}_m)$ per N=400, 800 e 1600.

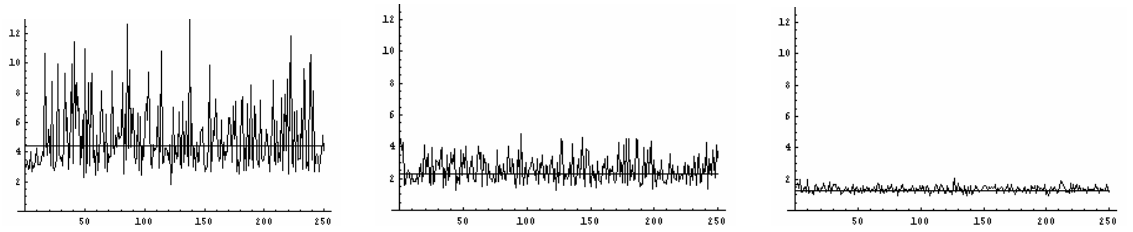

Figura 3 – Rappresentazione dei 250 valori simulati della stima $\hat{v}(\overline{Y}_c)$ rispetto a $\mathrm{V}(\overline{Y}_c)$ per N=400, 800 e 1600.

I grafici mostrano, per entrambe le stime $\hat{v}(\overline{Y}_m)$ (figura 2) e $\hat{v}(\overline{Y}_c)$ (figura 3), una tendenza alla concentrazione dei p valori simulati intorno ai corrispondenti oggetti di stima $\mathrm{V}(\overline{Y}_m)$ e $\mathrm{V}(\overline{Y}_c)$ via via più evidente all'aumentare di N, consentendo di dare valutazione positiva anche circa la consistenza delle medesime.

Può essere infine sottolineato che le simulazioni confermano la c-consistenza anche per le stime $\hat{v}(\overline{Y}_m)$ e $\hat{v}(\overline{Y}_c)$.

Dipartimento di Statistica FULVIA MECATTI
Università degli Studi di Milano-Bicocca

Dipartimento di Statistica SONIA MIGLIORATI
Università degli Studi di Milano-Bicocca

APPENDICE

La dimostrazione della (5) può avvenire attraverso i seguenti passaggi:

$$\mathrm{V}(\overline{Y}_c)=\sum_{l=1}^{L} V\left(\sum_{r=1}^{2^L-1}\frac{1}{\lambda_r}\sum_{q=1}^{N_{u_r}} Y_{rq}\,\delta_{rq,l}\right)=\sum_{l=1}^{L}\left\{\sum_{r=1}^{2^L-1}\frac{1}{\lambda_r^2}\left[\sum_{q=1}^{N_{\mathbf{u}_r}} Y_{rq}^2\; V(\delta_{rq,l})\;+\right.\right.$$

$$\left.\left.+\sum_{q=1}^{N_{\mathbf{u}_r}}\sum_{\substack{v=1\\ v\neq q}}^{N_{\mathbf{u}_r}} Y_{rq}Y_{rv}\,\mathrm{Cov}(\delta_{rq,l}\,\delta_{rv,l})\right]+\sum_{r=1}^{2^L-1}\sum_{\substack{t=1\\ t\neq r}}^{2^L-1}\frac{1}{\lambda_r\,\lambda_t}\,\mathrm{Cov}\left(\sum_{q=1}^{N_{\mathbf{u}_r}} Y_{rq}\,\delta_{rq,l},\sum_{v=1}^{N_{\mathbf{u}_t}} Y_{tv}\,\delta_{tv,l}\right)\right\}=$$

$$=\sum_{l=1}^{L}\left[\sum_{r=1}^{2^L-1}\frac{1}{\lambda_r^2}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2\,n_l\,u_{rl}}{N_l-1}-\sum_{r=1}^{2^L-1}\frac{Y_r^2\,n_l\,u_{rl}}{\lambda_r^2\,N_l(N_l-1)}-\sum_{r=1}^{2^L-1}\frac{1}{\lambda_r^2}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2\,n_l^2\,u_{rl}}{N_l(N_l-1)}+\right.$$

$$\left.+\sum_{r=1}^{2^L-1}\frac{Y_r^2\,n_l^2\,u_{rl}}{\lambda_r^2\,N_l^2(N_l-1)}-\sum_{r=1}^{2^L-1}\sum_{\substack{t=1\\ t\neq r}}^{2^L-1}\frac{Y_r\,Y_t\,n_l\,u_{rl}\,u_{tl}}{\lambda_r\lambda_t\,N_l(N_l-1)}+\sum_{r=1}^{2^L-1}\sum_{\substack{t=1\\ t\neq r}}^{2^L-1}\frac{Y_r\,Y_t\,n_l^2\,u_{rl}\,u_{tl}}{\lambda_r\lambda_t\,N_l^2(N_l-1)}\right]=$$

$$=\sum_{l=1}^{L}\frac{n_l(N_l-n_l)}{N_l(N_l-1)}\left[\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2\,u_{rl}}{\lambda_r^2}-\frac{1}{N_l}\left(\sum_{r=1}^{2^L-1}\frac{Y_r\,u_{rl}}{\lambda_r}\right)^2\right].$$

Col proposito di ottenere una stima per $\mathrm{V}(\overline{Y}_c)$, si noti che la variabile casuale descritta dalla (6) ha la forma seguente:

$$\sum_{l=1}^{L}\frac{1}{n_l-1}\left(1-\frac{n_l}{N_l}\right)\left[n_l\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2}{\lambda_r^2}\delta_{rq,l}-\left(\sum_{r=1}^{2^L-1}\frac{1}{\lambda_r}\sum_{1=1}^{N_{\mathbf{u}_r}}Y_{rq}\,\delta_{rq,l}\right)^2\right]$$

ed il suo valore atteso risulta:

$$\sum_{l=1}^{L}\frac{1}{n_l-1}\left(1-\frac{n_l}{N_l}\right)\left[n_l\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2\,n_l\,u_{rl}}{\lambda_r^2\,N_l}-\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\frac{Y_{rq}^2\,n_l\,u_{rl}}{\lambda_r^2\,N_l}+\right.$$

$$\left.-\sum_{r=1}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\sum_{\substack{v=1\\v\neq q}}^{N_{\mathbf{u}_r}}\frac{Y_{rq}\,Y_{rv}\,n_l(n_l-1)\,u_{rl}}{N_l(N_l-1)}-\sum_{r=1}^{2^L-1}\sum_{\substack{t=1\\t\neq r}}^{2^L-1}\sum_{q=1}^{N_{\mathbf{u}_r}}\sum_{v=1}^{N_{\mathbf{u}t}}\frac{Y_{rq}\,Y_{tv}\,n_l(n_l-1)\,u_{rl}\,u_{tl}}{N_l(N_l-1)}\right]$$

che, dopo semplici riduzioni algebriche, viene a coincidere con la (5), ovvero la (6) risulta corretta per $\mathrm{V}(\overline{Y}_c)$. Dalla (6) medesima segue quindi la seguente stima per $\mathrm{V}(\overline{Y}_c)$:

$$\hat{v}(\overline{Y}_c)=\sum_{l=1}^{L}\frac{1}{n_l-1}\left[n_l\sum_{r=1}^{2^L-1}\sum_{s=1}^{f_{\mathbf{u}_r,l}}\frac{y_{rs}^2}{\lambda_r^2}-\left(\sum_{r=1}^{2^L-1}\frac{y_{r,l}}{\lambda_r}\right)^2\right]$$

che coincide con la (7).

RIFERIMENTI BIBLIOGRAFICI

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.

C.M. CASSEL, C.E. SÄRNDAL AND J.H. WRETMAN (1977), *Foundations of Inference in Survey Sampling,* John Wiley & Sons, New York.

W.G. COCHRAN, (1977), *Sampling Techniques*, 3rd ed., John Wiley & Sons, New York.

EUROSTAT (2000), *Push and Pull Factors of International Migration. Country Report* – Italy*,* 3/2000/E/n. 5 Bruxelles: European Communities Printing Office.

H.O. HARTLEY (1962), *Multiple Frame Surveys*, "Proceedings of the Social Statist. Sec. ASA", pp. 203-206.

H.O. HARTLEY (1974), *Multiple Frame Methodology and Selected Applications*, "Sankhya", series C., 36, pp. 99-118.

R.E. LUND (1968), *Estimators in Multiple Frame Surveys*, "Proceedings of the Social Statist. Sec. ASA", pp. 282-288.

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

F. MECATTI e S. MIGLIORATI (2001), *Center Sampling: Theory and Estimation*, "Technical Report 01-06", Department of Statistics, Pennsylvania State University.

S.K. THOMPSON (1992), *Sampling*, Wiley, New York.

RIASSUNTO

Confronto fra stimatori per la media nel campionamento per centri

Nel campionamento per centri la popolazione di riferimento si presenta naturalmente aggregata in insiemi sovrapposti di unità statistiche non identificabili ed in genere di numerosità ignota. Sotto l'ipotesi che sia noto almeno il peso di tali centri rispetto alla numerosità totale della popolazione, è noto uno stimatore corretto $\bar{Y}_m$ per la media di una qualche caratteristica quantitativa presente su tale popolazione. Un secondo stimatore corretto $\bar{Y}_c$ è derivabile, sotto tale ipotesi, da una precedente proposta.

Nel presente lavoro è fornita la varianza esatta di tale secondo stimatore ed una stima per questa. Inoltre, i due stimatori, che differiscono per la logica di sintesi delle informazioni campionarie di centro e coincidono solo nel caso di allocazione proporzionale della numerosità campionaria fra i centri, sono posti a confronto tanto sotto il profilo delle proprietà inferenziali quanto sotto quello operativo.

Nel caso generale di L centri ($L \geq 2$) i risultati di una simulazione mostrano che non esiste, fra i due, lo stimatore uniformemente più efficiente ma $\bar{Y}_c$ risulta più efficiente di $\bar{Y}_m$ se la variabilità delle frazioni di campionamento di centro è prossima a 0 mentre $\bar{Y}_m$ è più efficiente di $\bar{Y}_c$ all'aumentare di tale variabilità.

La simulazione consente di valutare anche le stime proposte per le varianze di tali stimatori che risultano asintoticamente corrette e consistenti, sia in senso classico sia secondo Cochran.

SUMMARY

Center sampling: a comparison between mean estimators

The center sampling technique is well suited when a population is naturally gathered in overlapping groups of units for which the units can not be labeled and the group size as well as the population size are unknown.

An unbiased estimator $\bar{Y}_m$ for the mean of a quantitative characteristic of interest has been proposed under the simple hypothesis that the relative weight of each center is known. A second estimator $\bar{Y}_c$ can be deduced from a previous proposal under the same hypothesis.

In the present paper the exact variance of $\bar{Y}_c$ together with an estimate of it are given. The two estimators are based on different ways of summarizing data and they coincide in the case of proportional allocation of the overall sample size only.

A comparison between the two estimators is accomplished both from the inferential and from the practical point of view. Through a simulation study it is shown that no estimator is uniformly more efficient than the other in the general case of $L \geq 2$ centers. Besides it comes out that $\bar{Y}_c$ is more efficient when there is a "small" variability of the center sampling fractions, while $\bar{Y}_m$ is more efficient as this variability increases. Simulation also shows that the proposed estimators for the variance are asymptotically unbiased, consistent and *c*-consistent.