BIBLIOTECA DOCUMENTAZIONE RELAZIONI INTERNAZIONALI

quaderni di discussione

N. 85.05

Su un metodo di stima degli effetti stratificazione e clustering e dell'effetto complessivo del disegno di campionamento nei campioni a due stadi con stratificazione delle unità di primo stadio

Aldo Russo (*)

istat

I quaderni di discussione sono a circolazione ristretta e non impegnano la responsabilità dell'ISTAT ma riflettono solo il punto di vista degli autori. Non possono, quindi, essere citati e fatti circolare senza il permesso degli autori.

Le richieste vanno indirizzate a :
«ISTAT - Centro Documentazione - Dr.ssa Borgnino-Valenzano
Via Balbo, 16 - 00100 - ROMA

N. 85.05

Su un metodo di stima degli effetti stratificazione e clustering e dell'effetto complessivo del disegno di campionamento nei campioni a due stadi con stratificazione delle unità di primo stadio

Aldo Russo (*)

(*) Reparto Studi, Progetto 1: Studio dei campioni

Settembre 1985

Indice

Sommario

Dopo una breve premessa in cui si analizza criticamente la metodologia proposta da Verma, Scott e O'Muircheartaigh per la valutazione dei piani di campionamento di indagini basate su campioni "complessi", si propone una metodologia alternativa che consente di pervenire alla determinazione di stime più soddisfacenti, dal punto di vista statistico, dell'effetto stratificazione, dell'effetto clustering e dell'effetto complessivo del disegno di campionamento.

Per ciascuno di questi effetti, che sono alla base della problematica della valutazione dei piani di campionamento, viene esposta in dettaglio la derivazione algebrica nell'ipotesi di un'indagine campionaria basata su uno schema a due stadi con stratificazione delle unità di primo stadio.

Infine, allo scopo di dare una visione unificante, vengono ricavate alcune relazioni tra gli effetti menzionati.

I. Introduzione

Il presente studio trae la sua origine dal tipo di analisi condotta a suo tempo sui risultati delle indagini ISTAT "sulle condizioni di salute e sul ricorso ai servizi sanitari, I980" [4], e "sulle vacanze e gli sports, I983" [5]; tali analisi avevano lo scopo di offrire una valutazione della "bontà" dei piani di campionamento adottati, mediante uno studio approfondito degli effetti stratificazione, clustering, ponderazione e dell'effetto complessivo del disegno di campionamento.

Dette analisi ricalcavano integralmente l'impostazione metodologica proposta da Verma, Scott e O'Muircheartaigh [6], in un ampio studio finalizzato fondamentalmente alla valutazione dei piani di campionamento utilizzati da vari Paesi per effettuare l'indagine mondiale sulla fecondità.

Ciò che desideriamo esporre in questa sede é il frutto di una ulteriore riflessione su tale tipo di metodologia, la quale non ha ancora trovato - a nostro avviso - una sistemazione del tutto soddisfacente.

Ricavando alcuni risultati originali riguardanti la determinazione degli errori standard propri di disegni campionari ipotetici-costruiti espressamente-il presente lavoro illustra una metodologia alternativa che consente di pervenire ad una valutazione meno approssimata delle entità degli effetti studiati, rendendo i risultati dell'analisi più accurati rispetto a quelli conseguibili con la metodologia di Verma, Scott e O'Muircheartaigh.

Tra questi effetti non viene preso in considerazione l'effetto ponderazione, che sarà trattato in un prossimo articolo.

Affinché risultino più chiaramente le modifiche proposte, premettiamo alla loro esposizione una breve descrizione dell'impostazione dell'analisi condotta dagli autori citati per l'indagine mondiale sulla fecondità.

2. Impostazione di Verma, Scott e O'Muircheartaigh

2.I- Premessa

L'impianto metodologico muove dalla considerazione che gli errori di campionamento delle stime ottenute da un campione complesso di una prefissata numerosità differiscono da quelli di un ipotetico campione casuale semplice di pari numerosità in conseguenza di tre fattori che si possono cosi sintetizzare:

i)- la stratificazione della popolazione porta in genere, almeno nei casi in cui é stata effettuata utilizzando criteri appropriati, ad una riduzione della varianza delle stime (effetto stratificazione);

ii)- se si esclude l'effetto della stratificazione, l'introduzione

di più stadi di campionamento produce un aumento della varianza delle stime(effetto clustering);questo effetto é tanto più grande quanto più grande é la correlazione interna alle unità appartenenti ai diversi stadi di campionamento;

iii)- infine é da considerare l'effetto ponderazione,che si ha nei casi in cui é necessario introdurre dei pesi per la determinazione delle stime;tale effetto misura la variazione della varianza di un campione autoponderante rispetto alla varianza delle stime di un campione di uguale numerosità ma con pesi variabili.

I tre effetti sopra indicati interagiscono in tutti i disegni campionari complessi dando origine a quello che é comunemente chiamato "effetto complessivo del disegno di campionamento" (o deft)[3], tipico di ciascuna variabile oggetto d'indagine.

2.2- Procedimenti di calcolo

Illustreremo ora in dettaglio i procedimenti attraverso i quali vengono stimati i valori degli effetti in questione riferendoci alla seguente situazione:immaginiamo di aver effettuato una indagine(che indicheremo per comodità,d'ora innanzi,come indagine A) per stimare il totale del carattere Y in una determinata popolazione P mediante un campione a due stadi con stratificazione delle unità di primo stadio.

Allo scopo di rendere più chiara l'esposizione supponiamo che le unità di primo stadio siano i comuni e quelle di secondo stadio gli individui.

Indichiamo inoltre con $\hat{Y}$ la stima del totale Y e con $\hat{G}(\hat{Y})$ la stima del corrispondente errore di campionamento.

Consideriamo poi le seguenti strategie di campionamento fitti

zie desumibili dalla struttura campionaria dell'indagine A:

a)- due stadi semplici con numerosità campionaria di primo e secondo stadio uguali a quelle del campione usato per l'indagine A;

b)- uno stadio stratificato con estrazione di soli individui, pur restando ferma la condizione che il loro numero in ogni strato sia uguale al numero complessivo di individui-campione dell'indagine A;

c)- campione casuale semplice con estrazione di soli individui e di dimensione uguale al numero complessivo di individui-campione dell'indagine A.

A queste diverse strategie campionarie corrispondono, generalmente, diversi errori di campionamento che indicheremo rispettivamente con: $\sigma(\hat{Y}_{\bar{s}})$, $\sigma(\hat{Y}_{\bar{c}})$ e $\sigma(\hat{Y}_{\bar{s},\bar{c}})$

Una stima di tali errori é ottenuta in base ai dati forniti dall'indagine A e facendo le seguenti ipotesi di lavoro:

-per la strategia a) : supponendo che i comuni-campione dell'indagine A siano stati estratti da un universo unitario di comuni, ossia da un universo non stratificato;

-per la strategia b) : assumendo che gli individui-campione dell'indagine A, siano stati estratti - in ogni strato - da un universo di individui non suddiviso in comuni;

-per la strategia c) : considerando il campione totale di individui dell'indagine A come un campione casuale semplice estratto dalla popolazione P.

Ottenute le stime di $\sigma(\hat{Y}_{\bar{s}})$, $\sigma(\hat{Y}_{\bar{c}})$ e $\sigma(\hat{Y}_{\bar{s},\bar{c}})$ che indicheremo rispettivamente con $\hat{\sigma}(\hat{Y}_{\bar{s}})$, $\hat{\sigma}(\hat{Y}_{\bar{c}})$ e $\hat{\sigma}(\hat{Y}_{\bar{s},\bar{c}})$ é possibile determinare una stima degli effetti in esame confrontando tali stime con la stima dell'errore di campionamento dell'indagine A.

Si perviene pertanto alle espressioni seguenti:

(I) $$\hat{E}_s = \frac{\hat{\sigma}(\hat{Y})}{\hat{\sigma}(\hat{Y}_{\bar{s}})}$$

per l'effetto esercitato dalla stratificazione;

(2) $$\hat{E}_c = \frac{\hat{\sigma}(\hat{Y})}{\hat{\sigma}(\hat{Y}_{\bar{c}})}$$

per l'effetto clustering imputabile all'introduzione dei comuni come primo stadio di campionamento;

(3) $$\widehat{deft} = \frac{\hat{\sigma}(\hat{Y})}{\hat{\sigma}(\hat{Y}_{\bar{s},\bar{c}})}$$

per l'effetto complessivo del disegno di campionamento, dovuto alla interazione di tutti gli effetti.

3. Impostazione alternativa

3.I- Premessa

Le considerazioni in base alle quali sono state ottenute le stime $\hat{E}_s$, $\hat{E}_c$ e $\widehat{deft}$ non sono del tutto rigorose dal punto di vista campionario in rapporto alla determinazione delle stime $\hat{\sigma}(\hat{Y}_{\bar{s}})$, $\hat{\sigma}(\hat{Y}_{\bar{c}})$ e $\hat{\sigma}(\hat{Y}_{\bar{s},\bar{c}})$. Infatti i procedimenti di calcolo di queste ultime, richiedendo il ricorso a particolari ipotesi, conducono a stime di $\sigma(\hat{Y}_{\bar{s}})$, $\sigma(\hat{Y}_{\bar{c}})$ e $\sigma(\hat{Y}_{\bar{s},\bar{c}})$ approssimate e di dubbio significato statistico.

Da questa constatazione prende lo spunto il presente lavoro che ha lo scopo di suggerire un metodo più soddisfacente per la valutazione degli effetti $\hat{E}_s$, $\hat{E}_c$ e $\widehat{deft}$ e che fornisce stime di $\sigma(\hat{Y}_{\bar{s}})$, $\sigma(\hat{Y}_{\bar{c}})$ e $\sigma(\hat{Y}_{\bar{s},\bar{c}})$ basate su procedimenti di calcolo statisticamente fondati.

Nelle pagine seguenti descriveremo i risultati di tale studio nel caso - significativamente generale almeno nel quadro delle indagini ISTAT -di un disegno di campionamento a due stadi, strati

ficato al primo stadio, con estrazione casuale semplice e senza reimmissione delle unità in ciascuno stadio.

Dopo aver richiamato-senza dilungarci in sviluppi teorici e dimostrativi che si possono agevolmente trovare in [2],[7]-l'espressione dello stimatore del valore complessivo di un dato carattere e le formule concernenti l'errore di campionamento e la corrispondente stima del suddetto stimatore (par.3.2), si determinano-sfruttando le informazioni desumibili dal campione suddetto ed eliminando le ipotesi di cui al paragrafo 2.2.- le espressioni esatte delle stime $\hat{\sigma}(\hat{Y}_{\bar{s}})$, $\hat{\sigma}(\hat{Y}_{\bar{c}})$ e $\hat{\sigma}(\hat{Y}_{\bar{s},\bar{c}})$ e si offre la dimostrazione che si tratta di stime corrette (par.3.3,3.4 e 3.5).

3.2- Campionamento a due stadi con stratificazione delle unità di primo stadio: stima del totale e sua varianza

3.2.I- Simbologia

Indichiamo con:

i	indice di unità di primo stadio (PSU)
j	indice di unità di secondo stadio (SSU)
h	indice di strato (h=I,.....,H)
N_h	numero di PSU-universo nello strato h
$N = \sum_{h}^{H} N_h$	numero totale di PSU-universo
M_{hi}	numero di SSU-universo appartenenti alla PSU i dello strato h
$M_h = \sum_{i}^{N_h} M_{hi}$	numero di SSU-universo nello strato h
$M = \sum_{h}^{H} M_h$	numero totale di SSU-universo
n_h	numero di PSU-campione nello strato h ($n = \sum_{h}^{H} n_h$)

m_{hi} numero di SSU-campione nella PSU i dello strato h

$m_h = \sum_i^{n_h} m_{hi}$ numero di SSU-campione nello strato h

$m = \sum_h^H m_h$ numero totale di SSU-campione

Y_{hij} valore del carattere Y osservato sulla SSU j della PSU i dello strato h

$Y_{hi} = \sum_j^{M_{hi}} Y_{hij}$ totale del carattere Y nella PSU i dello strato h

$Y_h = \sum_i^{N_h} Y_{hi}$ totale del carattere Y nello strato h

$Y = \sum_h^H Y_h$ totale generale del carattere Y nella popolazione

$\bar{Y}_{hi} = Y_{hi} / M_{hi}$ media nella PSU i dello strato h

$\bar{Y}_h = Y_h / M_h$ media dello strato h .

$\bar{Y} = Y / M$ media generale della popolazione

3.2.2 - <u>Stima corretta del totale Y</u>

Nel campionamento a due stadi, stratificato al primo stadio, si dimostra che una stima non distorta del totale Y, nella classe degli stimatori lineari, é data dal seguente stimatore:

$$(4) \qquad \hat{Y} = \sum_h^H \sum_i^{n_h} \sum_j^{m_{hi}} \frac{Y_{hij}}{W_{hij}}$$

in cui W_{hij} indica la probabilità che il generico valore Y_{hij} si presenti comunque in una m - pla campionaria.

Per rendere esplicita la (4) occorre ora determinare i valori W_{hij} che conseguono dalla struttura probabilistica del campio-

namento a due stadi.

Tenendo presenti le caratteristiche del meccanismo di estrazione qui considerato (PSU e SSU estratte senza reimmissione e con probabilità uguali) W_{hij} è data dal prodotto della probabilità di estrarre la PSU i dallo strato h per la probabilità di estrarre la SSU j dentro a quella PSU, e cioè:

$$(5) \qquad W_{hij} = \left\{ \frac{n_h}{N_h} \right\} \cdot \left\{ \frac{m_{hi}}{M_{hi}} \right\}$$

Sostituendo tale espressione nella (4) si ottiene:

$$(6) \qquad \hat{Y} = \sum_{h}^{H} \frac{N_h}{n_h} \sum_{i}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j}^{m_{hi}} Y_{hij} =$$

$$= \sum_{h}^{H} \frac{N_h}{n_h} \sum_{i}^{n_h} \hat{Y}_{hi} = \sum_{h}^{H} \hat{Y}_h$$

in cui si é posto:

$$(7) \qquad \hat{Y}_{hi} = \frac{M_{hi}}{m_{hi}} \sum_{j}^{m_{hi}} Y_{hij}$$

$$(8) \qquad \hat{Y}_h = \frac{N_h}{n_h} \sum_{i}^{n_h} \hat{Y}_{hi}$$

espressioni che rappresentano, rispettivamente, le stime corrette dei totali Y_{hi} e Y_h .

3.2.3 - Varianza della distribuzione dello stimatore $\hat{Y}$

La varianza dello stimatore $\hat{Y}_h$ è definita dalla relazione:

$$(9) \quad V(\hat{Y}_h) = N_h^2 \frac{N_h - n_h}{N_h} \frac{{}_a S_h^2}{n_h} + \frac{N_h}{n_h} \sum_{i}^{N_h} M_{hi}^2 \frac{M_{hi} - m_{hi}}{M_{hi}} \frac{{}_b S_{hi}^2}{m_{hi}}$$

in cui:

$$(10) \qquad {}_aS_h^2 = \frac{1}{N_h - I} \sum_i^{N_h} \left(Y_{hi} - \frac{Y_h}{N_h} \right)^2$$

$$(11) \qquad {}_bS_{hi}^2 = \frac{I}{M_{hi} - I} \sum_j^{M_{hi}} \left(Y_{hij} - \bar{Y}_{hi} \right)^2$$

Conseguentemente la varianza dello stimatore $\hat{Y}$ è da da:

$$(12) \qquad V(\hat{Y}) = V\left(\sum_h^H \hat{Y}_h\right) = \sum_h^H V(\hat{Y}_h) =$$

$$= \sum_h^H N_h^2 \frac{N_h - n_h}{N_h} \frac{{}_aS_h^2}{n_h} + \sum_h^H \frac{N_h}{n_h} \sum_i^{N_h} M_{hi}^2 \frac{M_{hi} - m_{hi}}{M_{hi}} \frac{{}_bS_{hi}^2}{m_{hi}}$$

La (12) consta di due parti ciascuna delle quali esprime il contributo di una diversa fonte di variabilità alla variabilità totale. Il primo addendo rappresenta il contributo dovuto alla diversità fra le popolazioni globalmente considerate, che si esprime attraverso la"varianza dei totali ${}_aS_h^2$ ". Il secondo addendo, che dipende dai parametri ${}_bS_{hi}^2$, cioè dalle varianze delle singole sub-popolazioni, rappresenta, invece, una sintesi (per qualche aspetto simile ad una media ponderata) della variabilità all'interno di queste ultime.

3.2.4 - Stima corretta della varianza $V(\hat{Y})$

La stima di $V(\hat{Y})$ si ottiene "traducendo in termini campionari" la (9), ossia:

$$(13) \qquad \hat{V}(\hat{Y}_h) = N_h^2 \frac{N_h - n_h}{N_h} \frac{{}_aS_h^2}{n_h} + \frac{N_h}{n_h} \sum_i^{n_h} M_{hi}^2 \frac{M_{hi} - m_{hi}}{M_{hi}} \frac{{}_bS_{hi}^2}{m_{hi}}$$

in cui:

$$(14)\quad {}_a s_h^2 = \frac{1}{n_h - I} \sum_i^{n_h} \left(\hat{Y}_{hi} - \frac{\sum_i^{n_h} \hat{Y}_{hi}}{n_h} \right)^2$$

$$(15)\quad {}_b s_{hi}^2 = \frac{1}{m_{hi} - I} \sum_j^{m_{hi}} (Y_{hij} - \hat{\bar{Y}}_{hi})^2$$

$$(16)\quad \hat{\bar{Y}}_{hi} = \frac{\hat{Y}_{hi}}{M_{hi}} = \frac{1}{m_{hi}} \sum_j^{m_{hi}} Y_{hij}$$

Dalla (13) consegue subito che la stima di $V (\hat{Y})$ è data dalla seguente espressione:

$$(17)\quad \hat{V} (\hat{Y}) = \hat{V} \left(\sum_h^H \hat{Y}_h \right) = \sum_h^H \hat{V} (\hat{Y}_h) =$$

$$= \sum_h^H N_h^2 \frac{N_h - n_h}{N_h} \frac{{}_a s_h^2}{n_h} + \sum_h^H \frac{N_h}{n_h} \sum_i^{n_h} M_{hi}^2 \frac{M_{hi} - m_{hi}}{M_{hi}} \frac{{}_b s_{hi}^2}{m_{hi}}$$

E' possibile dimostrare che lo stimatore così definito rappresenta uno stimatore non distorto di $V (\hat{Y})$, per quanto le sue componenti non siano gli stimatori corretti dei corrispondenti termini di $V (\hat{Y})$

3.3- Effetto stratificazione

Supponiamo di aver effettuato un'indagine basata su un campione a due stadi con stratificazione delle unità di primo stadio.

Per valutare l'efficacia della stratificazione in termini di guadagno nella precisione delle stime fornite dall'indagine, con-

sideriamo un ipotetico campione a due stadi semplici con modalità di estrazione delle unità e numerosità campionarie di primo e secondo stadio uguali a quelle del campione usato per effettuare l'indagine in esame.

Per i fini delineati potremo individuare struttura e caratteristiche di tale strategia campionaria mediante una particolarizzazione pura e semplice delle formule ottenute per il campione a due stadi stratificato in primo stadio. Precisamente col simbolismo di cui al paragrafo 3.2.I, eliminando però l'indice h di strato, si ha che una stima corretta del totale generale Y della popolazione - che indicheremo nel presente contesto con $\hat{Y}_{\bar{S}}$ - é espressa da:

$$(I8) \qquad \hat{Y}_{\bar{S}} = \frac{N}{n} \sum_{i}^{n} \frac{M_i}{m_i} \sum_{j}^{m_i} Y_{ij}$$

Dalla formula (I2) si ottiene poi la corrispondente espressione della varianza di campionamento:

$$(I9) \qquad V(\hat{Y}_{\bar{S}}) = N^2 \frac{N - n}{N} \frac{{}_aS^2}{n} + \frac{N}{n} \sum_{i}^{N} M_i^2 \frac{M_i - m_i}{M_i} \frac{{}_bS^2}{m_i}$$

in cui:

$$(20) \qquad {}_aS^2 = \frac{I}{N - I} \sum_{i}^{N} \left(Y_i - \frac{Y}{N} \right)^2$$

$$(2I) \qquad {}_bS_i^2 = \frac{I}{M_i - I} \sum_{j}^{M_i} \left(Y_{ij} - \bar{Y}_i \right)^2$$

$$(22) \qquad \bar{Y}_i = \sum_{j}^{M_i} Y_{ij} \Big/ M_i = Y_i \Big/ M_i$$

A questo punto ci poniamo il problema di determinare una stima corretta di $V(\hat{Y}_{\bar{S}})$ sulla base delle sole informazioni desumibili dal campione a due stadi con stratificazione delle unità primarie.

Per risolverlo conviene intanto porre in una forma più conveniente l'espressione di ${}_aS^2$. Si ha:

$$(23)\quad (N-I)\,{}_aS^2 = \sum_i^N \left(Y_i - \frac{Y}{N}\right)^2 = \sum_h^H \sum_i^{N_h} \left(Y_{hi} - \frac{Y}{N}\right)^2 =$$

$$= \sum_h^H \sum_i^{N_h} \left(Y_{hi} - \frac{Y_h}{N_h} + \frac{Y_h}{N_h} - \frac{Y}{N}\right)^2 =$$

$$= \sum_h^H \sum_i^{N_h} \left(Y_{hi} - \frac{Y_h}{N_h}\right)^2 + \sum_h^H N_h \left(\frac{Y_h}{N_h} - \frac{Y}{N}\right)^2 + 2\sum_h^H \left(\frac{Y_h}{N_h} - \frac{Y}{N}\right) \sum_i^{N_h} \left(Y_{hi} - \frac{Y_h}{N_h}\right)$$

avendo aggiunto e tolto la media dei valori totali relativi agli N_h comuni universo del generico strato h e sviluppato il quadrato.

Osserviamo ora che l'ultimo termine della (23) é nullo in quanto la somma degli scarti di Y_{hi} dalla media aritmetica Y_h / N_h é evidentemente nulla.

Tenendo presente poi che il secondo termine della (23) può scriversi:

$$(24)\quad \sum_h^H N_h \left(\frac{Y_h}{N_h} - \frac{Y}{N}\right)^2 = \sum_h^H N_h \left(\frac{Y_h^2}{N_h^2} + \frac{Y^2}{N^2} - 2\,\frac{Y}{N}\,\frac{Y_h}{N_h}\right) =$$

$$= \sum_h^H \frac{Y_h^2}{N_h} + \frac{Y^2}{N} - 2\,\frac{Y^2}{N} = \sum_h^H \frac{Y_h^2}{N_h} - \frac{Y^2}{N}$$

la (23) diviene in conclusione:

$$(25)\quad (N-I)\,{}_aS^2 = \sum_h^H (N_h - I)\,{}_aS_h^2 + \sum_h^H \frac{Y_h^2}{N_h} - \frac{Y^2}{N}$$

Pertanto alla (I9) si può dare la forma seguente:

$$(26)\quad V(\hat{Y}_{\bar{s}}) = N\,\frac{N-n}{(N-I)n}\left[\sum_h^H (N_h - I)\,{}_aS_h^2 + \sum_h^H \frac{Y_h^2}{N_h} - \frac{Y^2}{N}\right] +$$

$$+ \frac{N}{n}\sum_i^N M_i^2\,\frac{M_i - m_i}{M_i}\,\frac{{}_bS_i^2}{m_i} =$$

$$= \frac{N(N-n)}{n(N-I)}\left[\sum_h^H (N_h - I)\,{}_aS_h^2 + \sum_h^H \frac{Y_h^2}{N_h} - \frac{Y^2}{N}\right] + \frac{N}{n}\sum_h^H\sum_i^{N_h} M_{hi}^2\,\frac{M_{hi} - m_{hi}}{M_{hi}}\,\frac{{}_bS_{hi}^2}{m_{hi}}$$

avendo espresso l'ultimo termine della (26) col simbolismo del campione a due stadi stratificato.

La formula (26) é strutturata in modo tale da consentire agevolmente l'ottenimento di una stima corretta di $V(\hat{Y}_{\bar{s}})$. A tal fine é sufficiente determinare una stima corretta dei parametri ${}_aS_h^2$, Y_h^2, Y^2 e della combinazione lineare

$$\frac{N}{n}\sum_h^H\sum_i^{N_h} M_{hi}^2\,\frac{M_{hi} - m_{hi}}{M_{hi}}\,\frac{{}_bS_{hi}^2}{m_{hi}}$$

Per il paramento ${}_aS_h^2$ una stima corretta, come é noto dalla teoria dei campioni [I], é espressa da:

$$(27)\quad {}_aS_h^2 - \frac{I}{n_h}\sum_i^{n_h} \frac{M_{hi}(M_{hi} - m_{hi})}{m_{hi}}\,{}_bS_{hi}^2$$

in cui ${}_aS_h^2$ e ${}_bS_{hi}^2$ sono rispettivamente definite dalle formule (I5) e (I6).

Riguardo ai parametri Y_h^2 e Y^2 una stima corretta può ottenersi sulla base delle relazioni seguenti:

$$(28)\quad V(\hat{Y}_h) = E\left[\hat{Y}_h - E(\hat{Y}_h)\right]^2 = E\left[\hat{Y}_h - Y_h\right]^2 = E\,(\hat{Y}_h^2) - Y_h^2$$

$$(29)\quad V(\hat{Y}) = E\left[\hat{Y} - E(\hat{Y})\right]^2 = E\left[\hat{Y} - Y\right]^2 = E\,(\hat{Y}^2) - Y^2$$

dalle quali si trae:

$$(30)\qquad Y_h^2 = E(\hat{Y}_h^2) - V(\hat{Y}_h)$$

$$(3I)\qquad Y^2 = E(\hat{Y}^2) - V(\hat{Y})$$

Consegue immediatamente che:

$$(32)\quad \hat{Y}_h^2 - \hat{V}(\hat{Y}_h) \qquad e \qquad \hat{Y}^2 - \hat{V}(\hat{Y})$$

sono rispettivamente stime corrette di Y_h^2 e Y^2.

Occupiamoci infine della stima della combinazione lineare; lo stimatore avrà la forma:

$$(33)\qquad \frac{N}{n}\sum_h^H \sum_i^{n_h} \lambda_{hi}\, {}_bs_{hi}^2$$

Si ha;

$$E\left[\frac{N}{n}\sum_h^H \sum_i^{n_h} \lambda_{hi}\, {}_bs_{hi}^2\right] = E\left[\frac{N}{n}\sum_h^H \sum_i^{n_h} \lambda_{hi}\, E({}_bs_{hi}^2)/i\right] =$$

$$= E\left[\frac{N}{n}\sum_h^H \sum_i^{n_h} \lambda_{hi}\, {}_bS_{hi}^2\right] = \frac{N}{n}\sum_h^H \sum_i^{n_h} \frac{I}{N_h}\sum_i^{N_h} \lambda_{hi}\, {}_bS_{hi}^2 = \frac{N}{n}\sum_h^H \frac{n_h}{N_h}\sum_i^{N_h} \lambda_{hi}\, {}_bS_{hi}^2$$

uguagliando quest'ultima espressione alla combinazione lineare segue che:

$$(34)\qquad \lambda_{hi} = \frac{M_{hi}(M_{hi} - M_{hi})}{m_{hi}} \frac{N_h}{n_h}$$

e quindi una stima corretta della combinazione suddetta é for-

nita dall'espressione:

$$(35)\qquad \frac{N}{n}\sum_{h}^{H}\frac{N_h}{n_h}\sum_{i}^{n_h}\frac{M_{hi}(M_{hi}-m_{hi})}{m_{hi}}\,{}_bs^2_{hi}$$

Introducendo la (27), le (32) e la (34) nella (26) si ottiene pertanto una stima corretta di $V(\hat{Y}_{\bar{s}})$:

$$(36)\quad \hat{V}(\hat{Y}_{\bar{s}}) = \frac{N(N-n)}{n(N-I)}\left[\sum_{h}^{H}(N_h-I)\left({}_as^2_h-\frac{I}{n_h}\sum_{i}^{n_h}\frac{M_{hi}(M_{hi}-m_{hi})}{m_{hi}}\,{}_bs^2_{hi}\right)+\right.$$

$$\left.+\sum_{h}^{H}\frac{I}{N_h}\left(\hat{Y}_h^2-\hat{V}(\hat{Y}_h)\right)-\frac{I}{N}\left(\hat{Y}^2-\hat{V}(\hat{Y})\right)\right]+\frac{N}{n}\sum_{h}^{H}\frac{N_h}{n_h}\sum_{i}^{n_h}\frac{M_{hi}(M_{hi}-m_{hi})}{m_{hi}}\,{}_bs^2_{hi}$$

Sviluppando ed aggregando i termini simili si ha in definitiva:

$$(37)\quad \hat{V}(\hat{Y}_{\bar{s}}) = \frac{N(N-n)}{n(N-I)}\left[\sum_{h}^{H}(N_h-I)\,{}_as^2_h+\sum_{h}^{H}\frac{\hat{Y}_h^2}{N_h}-\frac{\hat{Y}^2}{N}+\frac{\hat{V}(\hat{Y})}{N}-\sum_{h}^{H}\frac{\hat{V}(\hat{Y}_h)}{N_h}\right]+$$

$$+\sum_{h}^{H}\frac{N\left[(N-n)+N_h(n-I)\right]}{n\,n_h(N-I)}\sum_{i}^{n_h}\frac{M_{hi}(M_{hi}-m_{hi})}{m_{hi}}\,{}_bs^2_{hi}$$

Sostituendo nella (37) le espressioni di $\hat{V}(\hat{Y}_h)$ e $\hat{V}(\hat{Y})$, date rispettivamente da (I3) e (I7), si ricava la formula alternativa:

$$(38)\quad \hat{V}(\hat{Y}_{\bar{s}}) = \frac{(N-n)}{n(N-I)}\left[\sum_{h}^{H}\frac{N_h\left[N(n_h-I)+(N_h-n_h)\right]}{n_h}\,{}_as^2_h+\sum_{h}^{H}\frac{\hat{Y}_h^2}{N_h}-\frac{\hat{Y}^2}{N}\right]+$$

$$+\sum_{h}^{H}\frac{N_h}{n_h}\sum_{i}^{n_h}\frac{M_{hi}(M_{hi}-m_{hi})}{m_{hi}}\,{}_bs^2_{hi}$$

A questo punto non resta che determinare, sulla base dei risultati ottenuti, una stima dell'effetto stratificazione E_s. A tale scopo basterà dividere l'errore di campionamento della stima $\hat{Y}$ per l'errore di campionamento della stima $\hat{Y}_{\bar{s}}$, ossia:

$$(39) \qquad \hat{E}_s = \frac{\sqrt{\hat{V}.(\hat{Y})}}{\sqrt{\hat{V}(\hat{Y}_{\bar{s}})}} = \frac{\hat{\sigma}(\hat{Y})}{\hat{\sigma}(\hat{Y}_{\bar{s}})}$$

3.4- Effetto clustering

Per valutare l'effetto clustering-dovuto all'introduzione delle PSU come primo stadio di campionamento-sulla precisione delle stime ottenute mediante il disegno d'indagine descritto al punto 3.2 abbiamo bisogno di definire una strategia campionaria fittizia basata su uno schema ad uno stadio stratificato, le cui unità di rilevazione siano costituite da sole SSU.

Lo studio di tale effetto-basato quindi sul confronto tra campionamento a due stadi e campionamento ad uno stadio-ha pieno significato solo a parità di numerosità campionarie. Per adeguarci a questa esigenza basterà supporre che in ogni strato il numero di SSU estratte per il campione ad uno stadio stratificato sia uguale al numero di SSU estratte dalle n_h PSU prescelte per il campione a due stadi con stratificazione delle unità primarie.

Con questa premessa ed utilizzando il simbolismo descritto al paragrafo 3.2.I salvo l'eliminazione dell'indice i di PSU, richiamiamo le espressioni della stima del totale Y della popolazione e la corrispondente varianza di campionamento per la suddetta strategia fittizia.

Per il totale Y della popolazione é possibile dimostrare che una stima corretta é data da:

$$(40) \qquad \hat{Y}_{\bar{c}} = \sum_{h}^{H} \frac{M_h}{m_h} \sum_{j}^{m_h} Y_{hj}$$

La varianza di detta stima é fornita dall'espressione:

$$(4I) \qquad V(\hat{Y}_{\bar{c}}) = \sum_{h}^{H} M_h^2 \frac{M_h - m_h}{M_h} \frac{S_h^2}{m_h}$$

in cui:

$$(42) \qquad S_h^2 = \frac{I}{M_h - I} \sum_{j}^{M_h} \left(Y_{hj} - \bar{Y}_h\right)^2$$

Nel seguito mostreremo come sia possibile, sulla base delle sole informazioni ricavabili dal campione a due stadi stratificato al primo stadio, stimare in modo non distorto la varianza $V(\hat{Y}_{\bar{c}})$.

A tal fine conviene porre la (42) nella forma equivalente:

$$(43) \quad S_h^2 = \frac{I}{M_h - 1} \sum_{j}^{M_h} (Y_{hj} - \bar{Y}_h)^2 = \frac{I}{M_h - 1} \sum_{i}^{N_h} \sum_{j}^{M_{hi}} (Y_{hij} - \bar{Y}_h)^2 =$$

$$= \frac{I}{M_h - I} \sum_{i}^{N_h} \sum_{j}^{M_{hi}} \left[(Y_{hij} - \bar{Y}_{hi}) + (\bar{Y}_{hi} - \bar{Y}_h) \right]^2 =$$

$$= \frac{I}{M_n - I} \sum_{i}^{N_h} \sum_{j}^{M_{hi}} (Y_{hij} - \bar{Y}_{hi})^2 + \frac{I}{M_h - I} \sum_{i}^{N_h} \sum_{j}^{M_{hi}} (\bar{Y}_{hi} - \bar{Y}_h)^2 +$$

$$+ \frac{2}{M_h - I} \sum_{i}^{N_h} \sum_{j}^{M_{hi}} (Y_{hij} - \bar{Y}_{hi})(\bar{Y}_{hi} - \bar{Y}_h)$$

avendo utilizzato il simbolismo descritto al punto 3.2.I ed aggiunto e tolto la media relativa alla generica PSU i dello strato h.

Notiamo ora che l'ultimo termine della (43) risulta nullo in quanto le somme degli scarti di Y_{hij} dalle rispettive medie $\bar{Y}_{hi}$ sono evidentemente nulle.

Se teniamo poi conto della definizione di ${}_bS^2_{hi}$,formula (II), la (43) diviene in conclusione:

$$(44) \qquad S_h^2 = \frac{I}{M_h - 1}\left[\sum_i^{N_h}(M_{hi} - I)\,{}_bS^2_{hi} + \sum_i^{N_h} M_{hi}\,(\bar{Y}_{ni} - \bar{Y}_h)^2\right]$$

Pertanto si può dare alla (4I) la forma seguente:

$$(45) \qquad V(\hat{\bar{Y}}_c) = \sum_h^H \frac{M_h(M_h - m_h)}{m_h(M_h - I)}\left[\sum_i^{N_h}(M_{hi}-I)\,{}_bS^2_{hi} + \sum_i^{N_h} M_{hi}(\bar{Y}_{hi} - \bar{Y}_h)^2\right]$$

Tenendo presente inoltre che $\bar{Y}_{hi} = Y_{hi}/M_{hi}$ e $\bar{Y}_h = Y_h / M_h$ e sviluppando il quadrato,il secondo termine della (45) può porsi nella forma:

$$(46) \qquad \sum_i^{N_h} M_{hi}\,(\bar{Y}_{ni} - \bar{Y}_n)^2 = \sum_i^{N_h} M_{hi}\left(\frac{Y_{hi}}{M_{hi}} - \frac{Y_n}{M_h}\right)^2 =$$

$$= \sum_i^{N_h} \frac{Y_{hi}^2}{M_{hi}} - \frac{Y_h^2}{M_h}$$

Pertanto la (45) può riscriversi nel modo seguente :

$$(47)\quad V(\hat{\bar{Y}}_c) = \sum_h^H \frac{M_h(M_h - m_h)}{m_h(M_h - I)} \left[\sum_i^{N_h} (M_{hi} - I)\, {}_bS^2_{hi} + \sum_i^{N_h} \frac{Y^2_{hi}}{M_{hi}} - \frac{Y^2_h}{M_h} \right]$$

A questo punto non resta che determinare una stima corretta della (47);a tale scopo sarà sufficiente stimare in modo non distorto le combinazioni lineari:

$$(48)\qquad \sum_i^{N_h} (M_{hi} - I)\, {}_bS^2_{hi}$$

$$(49)\qquad \sum_i^{N_h} \frac{Y^2_{hi}}{M_{hi}}$$

e il parametro Y^2_h .

Consideriamo,in primo luogo,la combinazione lineare (48); lo stimatore avrà la forma:

$$(5I)\qquad \sum_i^{n_h} \lambda_{hi}\ {}_bs^2_{hi}$$

Si ha:

$$E\left[\sum_i^{n_h} \lambda_{hi}\ {}_bs^2_{hi}\right] = E\left[\sum_i^{n_h} \lambda_{hi}\ E({}_bs^2_{hi}/i)\right] = E\left[\sum_i^{n_h} \lambda_{hi}\ {}_bS^2_{hi}\right] =$$

$$= \sum_i^{n_h} \frac{I}{N_h} \sum_i^{N_h} \lambda_{hi}\ {}_bS^2_{hi} = \frac{n_h}{N_h} \sum_i^{N_h} \lambda_{hi}\ {}_bS^2_{hi}$$

ed uguagliando alla (48) segue che:

$$(52)\qquad \lambda_{hi} = \frac{(M_{hi} - I)\, N_h}{n_h}$$

e quindi una stima corretta della (48) é espressa da:

$$(53) \qquad \frac{N_h}{n_h} \sum_{i}^{n_h} (M_{hi} - I)_b s^2_{hi}$$

Per quanto riguarda la combinazione lineare (49) osserviamo intanto che l'espressione:

$$(54) \qquad \hat{Y}^2_{hi} - \hat{V}(\hat{Y}_{hi})$$

rappresenta una stima corretta del parametro Y^2_{hi} .

Infatti dalla nota identità:

$$(55) \qquad V(\hat{Y}_{hi}) = E\left[\hat{Y}_{hi} - E(\hat{Y}_{hi})\right]^2 = E(\hat{Y}^2_{hi}) - Y^2_{hi}$$

segue che:

$$(56) \qquad Y^2_{hi} = E(\hat{Y}^2_{hi}) - V(\hat{Y}_{hi})$$

e quindi una stima corretta di Y^2_{hi} é data da:

$$(57) \qquad \hat{Y}^2_{hi} - \hat{V}(\hat{Y}_{hi})$$

Per calcolare poi una stima corretta della (49) basterà determinare l'espressione di λ_{hi} del seguente stimatore:

$$(58) \qquad \sum_{i}^{n_h} \lambda_{hi} \left[\hat{Y}^2_{hi} - \hat{V}(\hat{Y}_{hi})\right]$$

Si ha:

$$E\left\{\sum_{i}^{n_h} \lambda_{hi}\left[\hat{Y}^2_{hi} - \hat{V}(\hat{Y}_{hi})\right]\right\} = E\left\{\sum_{i}^{n_h} \lambda_{hi}\; E\left[\hat{Y}^2_{hi} - \hat{V}(\hat{Y}_{hi})\right]/i\right\} =$$

$$= E\left\{\sum_{i}^{n_h} \lambda_{hi}\; Y^2_{hi}\right\} = \sum_{i}^{n_h} \frac{I}{N_h} \sum_{i}^{N_h} \lambda_{hi}\; Y^2_{hi} = \frac{n_h}{N_h} \sum_{i}^{N_h} \lambda_{hi}\; Y^2_{hi}$$

ed uguagliando l'ultimo membro di tale espressione alla (49) si

ottiene:

$$\lambda_{hi} = \frac{N_h}{n_h} \frac{1}{M_{hi}} \tag{59}$$

In definitiva una stima corretta é fornita da:

$$\frac{N_h}{n_h} \sum_{i}^{n_h} \frac{1}{M_{hi}} \left[\hat{Y}_{hi}^2 - \hat{V}(\hat{Y}_{hi}) \right] \tag{60}$$

Infine per il parametro Y_h^2 vale quanto già dimostrato al paragrafo 3.3, formula (32).

Sulla base dei risultati ottenuti consegue che una stima corretta della (47) é definita dall'espressione:

$$\hat{V}(\hat{Y}_{\bar{c}}) = \sum_{h}^{H} \frac{M_h(M_h-m_h)}{m_h(M_h-I)} \left[\frac{N_h}{n_h} \sum_{i}^{n_h} (M_{hi}-I)\ {}_b s_{hi}^2 + \frac{N_h}{n_h} \sum_{i}^{n_h} \frac{\hat{Y}_{hi}^2}{M_{hi}} + \right. \tag{6I}$$

$$\left. - \frac{N_h}{n_h} \sum_{i}^{n_h} \frac{\hat{V}(\hat{Y}_{hi})}{M_{hi}} - \frac{\hat{Y}_h^2}{M_h} + \frac{\hat{V}(\hat{Y}_h)}{M_h} \right]$$

Essendo peraltro:

$$\hat{V}(\hat{Y}_{hi}) = M_{hi}^2 \frac{M_{hi} - m_{hi}}{M_{hi}} \frac{{}_b s_{hi}^2}{m_{hi}} \tag{62}$$

é possibile dare alla (62) la formula alternativa:

$$\hat{V}(\hat{Y}_{\bar{c}}) = \sum_{h}^{H} \frac{M_h(M_h-m_h)}{m_h(M_h-I)} \left[\frac{N_h}{n_h} \sum_{i}^{n_h} \frac{M_{hi}(m_{hi}-I)}{m_{hi}}\ {}_b s_{hi}^2 + \frac{N_h}{n_h} \sum_{i}^{n_h} \frac{\hat{Y}_{hi}^2}{M_{hi}} - \frac{\hat{Y}_h^2}{M_h} + \frac{\hat{V}(\hat{Y}_h)}{M_h} \right] \tag{63}$$

Se teniamo poi conto della definizione:

$$(64)\quad {}_b s^2_{hi} = \frac{I}{m_{hi} - I} \sum_{j}^{m_{hi}} (Yhij - \hat{\bar{Y}}_{hi})^2 = \frac{I}{m_{hi} - I} \sum_{j}^{m_{hi}} Y^2_{hij} - \frac{m_{hi}}{m_{hi} - I} \hat{\bar{Y}}^2_{hi}$$

la (63), dopo alcune riduzioni, può scriversi pure nella forma:

$$(65)\quad \hat{V}(\hat{\bar{Y}}_{\bar{c}}) = \sum_{h}^{H} \frac{M_h(M_h - m_h)}{m_h(M_h - I)} \left[\frac{N_h}{n_h} \sum_{i}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j}^{m_{hi}} Y^2_{hij} - \frac{\hat{Y}^2_h}{M_h} + \frac{\hat{V}(\hat{Y}_h)}{M_h} \right]$$

Si può a questo punto definire una stima dell'effetto clustering attraverso il rapporto:

$$(66)\quad \hat{E}_c = \frac{\sqrt{\hat{V}(\hat{Y})}}{\sqrt{\hat{V}(\hat{\bar{Y}}_{\bar{c}})}} = \frac{\hat{\sigma}(\hat{Y})}{\hat{\sigma}(\hat{\bar{Y}}_{\bar{c}})}$$

in cui si é posto :

$$\sqrt{\hat{V}(\hat{Y})} = \hat{\sigma}(\hat{Y}) \qquad \text{e} \qquad \sqrt{\hat{V}(\hat{\bar{Y}}_{\bar{c}})} = \hat{\sigma}(\hat{\bar{Y}}_{\bar{c}})$$

3.5- Effetto complessivo del disegno di campionamento

Supponiamo che dalla popolazione definita al paragrafo 3.2 venga estratto un campione casuale semplice di SSU di dimensione uguale a quella del campione a due stadi stratificato al primo stadio, e cioé a $m = \sum_h \sum_i m_{hi}$.

Per tale schema campionario é possibile dimostrare-riferendo ad esso il simbolismo di cui al paragrafo 3.2.I salvo l'eliminazione dell'indice i di PSU e l'indice h di strato- che una stima corretta del totale della popolazione Y, che nel presente contesto indicheremo con $\hat{\bar{Y}}_{\bar{s},\bar{c}}$, é espressa da:

$$(67)\quad \hat{\bar{Y}}_{\bar{s},\bar{c}} = \frac{M}{m} \sum_{j}^{m} Y_j$$

La corrispondente varianza di campionamento é espressa da:

$$(68) \qquad V(\hat{\bar{Y}}_{\bar{s},\bar{c}}) = M^2 \frac{M-m}{M} \frac{S^2}{m}$$

in cui:

$$(69) \qquad S^2 = \frac{I}{M-I} \sum_{j}^{M} (Y_j - \bar{\bar{Y}})^2$$

Ciò premesso ci proponiamo di stimare in modo non distorto la suddetta varianza, sfruttando la conoscenza delle sole informazioni ottenibili dal campione a due stadi stratificato al primo stadio.

A tale scopo conviene, in primo luogo, porre in una forma diversa la (69). Si ha:

$$(70) \qquad S^2 = \frac{I}{M-I} \sum_{j}^{M} (Y_j - \bar{\bar{Y}})^2 = \frac{I}{M-I} \sum_{h}^{H} \sum_{i}^{N_h} \sum_{j}^{M_{hi}} (Y_{hij} - \bar{\bar{Y}})^2 =$$

$$= \frac{I}{M-I} \sum_{h}^{H} \sum_{i}^{N_h} \sum_{j}^{M_{hi}} (Y_{hij} - \bar{Y}_{hi} + \bar{Y}_{hi} + \bar{Y}_h - \bar{Y}_h - \bar{\bar{Y}})^2 =$$

$$= \frac{I}{M-I} \left[\sum_{h}^{H} \sum_{i}^{N_h} \sum_{j}^{M_{hi}} (Y_{hij} - \bar{Y}_{hi})^2 + \sum_{h}^{H} \sum_{i}^{N_h} \sum_{j}^{M_{hi}} (\bar{Y}_{hi} - \bar{Y}_h)^2 + \sum_{h}^{H} \sum_{i}^{N_h} \sum_{j}^{M_{hi}} (\bar{Y}_h - \bar{\bar{Y}})^2 + \right.$$

$$+ \sum_{h}^{H} \sum_{i}^{N_h} \sum_{j}^{M_{hi}} (Y_{hij} - \bar{Y}_{hi})(\bar{Y}_{hi} - \bar{Y}_h) + \sum_{h}^{H} \sum_{i}^{N_h} \sum_{j}^{M_{hi}} (Y_{hij} - \bar{Y}_{hi})(\bar{Y}_h - \bar{\bar{Y}}) +$$

$$\left. + \sum_{h}^{H} \sum_{i}^{N_h} \sum_{j}^{M_{hi}} (\bar{Y}_{hi} - \bar{Y}_h)(\bar{Y}_h - \bar{\bar{Y}}) \right]$$

avendo introdotto il simbolismo utilizzato per il campione a due stadi ed aggiunto e tolto le medie $\bar{Y}_{hi}$ e $\bar{Y}_h$.

Poiché con facili passaggi é facile accertare che gli ultimi tre termini di tale espressione sono nulli, la(70) si riduce alla:

$$(7I)\quad S^2 = \frac{I}{M - I}\left[\sum_h^H \sum_i^{N_h} \sum_j^{M_{hi}} (Y_{hij} - \bar{Y}_{hi})^2 + \sum_h^H \sum_i^{N_h} M_{hi}(\bar{Y}_{hi} - \bar{Y}_h)^2 + \right.$$

$$\left. + \sum_h^H M_h(\bar{Y}_h - \bar{Y})^2\right]$$

In forza poi della definizione di ${}_bS^2_{hi}$, la (7I) può scriversi:

$$(72)\quad S^2 = \frac{I}{M - I}\left[\sum_h^H \sum_i^{N_h} (M_{hi} - I)\ {}_bS^2_{hi} + \sum_h^H \sum_i^{N_h} M_{hi}(\bar{Y}_{hi} - \bar{Y}_h)^2 + \sum_h^H M_h(\bar{Y}_h - \bar{Y})^2\right]$$

Tenendo presente inoltre che $\bar{Y}_{hi} = Y_{hi}/M_{hi}$, $\bar{Y}_h = Y_h/M_h$ e $\bar{Y} = Y/M$ il secondo e il terzo termine della (72) possono essere posti nella forma:

$$(73)\quad \sum_h^H \sum_i^{N_h} M_{hi}(\bar{Y}_{hi} - \bar{Y}_h)^2 = \sum_h^H \sum_i^{N_h} M_{hi}\left(\frac{Y_{hi}}{M_{hi}} - \frac{Y_h}{M_h}\right)^2 =$$

$$= \sum_h^H \sum_i^{N_h} \frac{Y_{hi}^2}{M_{hi}} - \sum_h^H \frac{Y_h^2}{M_h}$$

$$(74)\quad \sum_h^H M_h(\bar{Y}_h - \bar{Y})^2 = \sum_h^H M_h\left(\frac{Y_h}{M_h} - \frac{Y}{M}\right)^2 = \sum_h^H \frac{Y_h^2}{M_h} - \frac{Y^2}{M}$$

In definitiva la (72) può riscriversi nel modo seguente:

$$(75)\quad S^2 = \frac{I}{M-I}\left[\sum_h^H \sum_i^{N_h} (M_{hi} - I)\, {}_bS^2_{hi} + \sum_h^H \sum_i^{N_h} \frac{Y^2_{hi}}{M_{hi}} - \frac{Y^2}{M}\right]$$

e quindi la (68) assume la forma equivalente:

$$(76)\quad V(\hat{Y}_{\bar{s},\bar{c}}) = \frac{M(M-m)}{m(M-I)}\left[\sum_h^H \sum_i^{N_h} (M_{hi}-I)\, {}_bS^2_{hi} + \sum_h^H \sum_i^{N_h} \frac{Y^2_{hi}}{M_{hi}} - \frac{Y^2}{M}\right]$$

Una stima corretta della (76), che indicheremo con $\hat{V}(\hat{Y}_{\bar{s},\bar{c}})$, può ottenersi immediatamente stimando in modo non distorto le combinazioni lineari:

$$(77)\quad \sum_i^{N_h} (M_{hi} - I)\, {}_bS^2_{hi}$$

$$(78)\quad \sum_i^{N_h} \frac{Y^2_{hi}}{M_{hi}}$$

e il parametro Y^2.

Tali stime, come abbiamo già dimostrato nelle pagine precedenti, sono date rispettivamente da:

$$(79)\quad \frac{N_n}{n_n} \sum_i^{n_h} (M_{hi} - I)\, {}_bs^2_{hi} \qquad \text{(formula (53))}$$

$$(80)\quad \frac{N_n}{n_n} \sum_i^{n_h} \frac{I}{M_{hi}} \left(\hat{Y}^2_{hi} - \hat{V}(\hat{Y}_{hi})\right) \qquad \text{(formula (60))}$$

$$(8I)\quad \hat{Y}^2 - \hat{V}(\hat{Y}) \qquad \text{(formula (32))}$$

Introducendo la (79),(80) e (8I) nella (76) si ottiene pertanto una stima corretta di $V(\hat{\bar{Y}}_{\bar{s},\bar{c}})$ espressa da:

$$(82)\quad \hat{V}(\hat{\bar{Y}}_{\bar{s},\bar{c}}) = \frac{M(M-m)}{m(M-I)}\left[\sum_{h}^{H}\frac{N_h}{n_h}\sum_{i}^{n_h}(M_{hi}-I)\,{}_b s^2_{hi} - \frac{I}{M}\left(\hat{Y}^2 - \hat{V}(\hat{Y})\right) + \right.$$

$$\left. + \sum_{h}^{H}\frac{N_h}{n_h}\sum_{i}^{n_h}\frac{I}{M_{hi}}\left(\hat{Y}^2_{hi} - \hat{V}(\hat{Y}_{hi})\right)\right]$$

Essendo peraltro:

$$(83)\quad {}_b s^2_{hi} = \frac{I}{m_{hi}-I}\sum_{j}^{m_{hi}}(Y_{hij} - \bar{\bar{Y}}_{hi})^2$$

e

$$(84)\quad \hat{V}(\hat{Y}_{hi}) = M^2_{hi}\,\frac{M_{hi}-m_{hi}}{M_{hi}}\,\frac{{}_b s^2_{hi}}{m_{hi}}$$

é possibile dare alla (82) la formula alternativa:

$$(85)\quad \hat{V}(\hat{\bar{Y}}_{\bar{s},\bar{c}}) = \frac{M(M-m)}{m(M-I)}\left[\sum_{h}^{H}\frac{N_h}{n_h}\sum_{i}^{n_h}\frac{M_{hi}}{m_{hi}}\sum_{j}^{m_{hi}}Y^2_{hij} - \frac{\hat{Y}^2}{M} + \frac{\hat{V}(\hat{Y})}{M}\right]$$

Possimamo ora definire una stima dell'effetto complessivo di campionamento mediante il rapporto:

$$(86)\quad \hat{deft} = \frac{\sqrt{\hat{V}(\hat{Y})}}{\sqrt{\hat{V}(\hat{\bar{Y}}_{\bar{s},\bar{c}})}} = \frac{\hat{\sigma}(\hat{Y})}{\hat{\sigma}(\hat{\bar{Y}}_{\bar{s},\bar{c}})}$$

avendo posto:

$$\sqrt{\hat{V}(\hat{Y})} = \hat{\sigma}(\hat{Y}) \qquad \text{e} \qquad \sqrt{\hat{V}(\hat{\bar{Y}}_{\bar{s},\bar{c}})} = \hat{\sigma}(\hat{\bar{Y}}_{\bar{s},\bar{c}})\ .$$

4. Relazioni formali tra gli effetti stratificazione, clustering e l'effetto complessivo del disegno di campionamento

Nei paragrafi precedenti abbiamo determinato le espressioni (38), (6I) e (82) mediante le quali-sfruttando le informazioni desumibili da un'indagine basata su un disegno campionario a due stadi con stratificazione delle unità di primo stadio-é possibile ottenere una stima corretta degli errori di campionamento $\sigma(\hat{Y}_{\bar{s}})$, $\sigma(\hat{Y}_{\bar{c}})$, e $\sigma(\hat{Y}_{\bar{s},\bar{c}})$, corrispondenti alle tre diverse strategie campionarie fittizie espressamente costruite per la valutazione degli effetti $\hat{E}_s$, $\hat{E}_c$ e $\hat{deft}$.

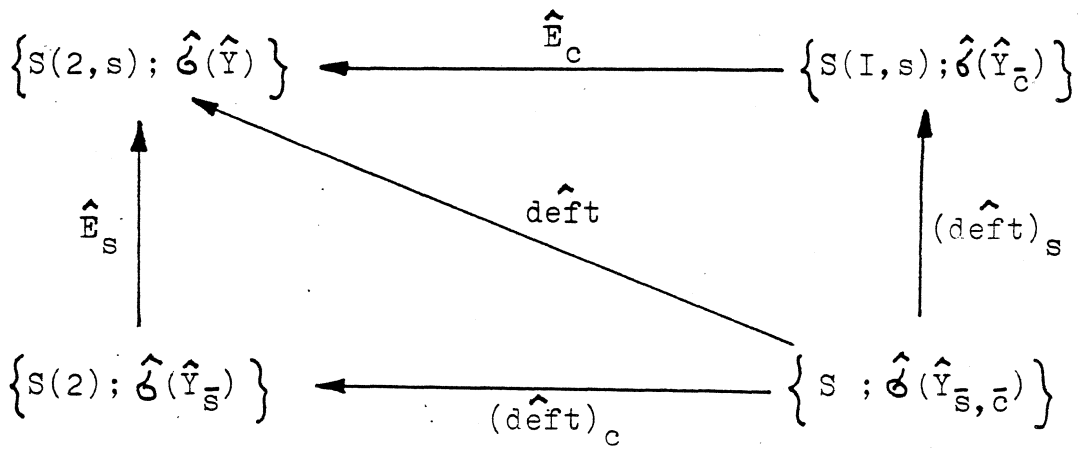
Questo paragrafo, invece, si ispira all'esigenza di rendere compatta la trattazione svolta introducendo in essa i legami esistenti tra gli effetti in questione.

A tal fine si ricavano le relazioni:

a) tra l'effetto complessivo del disegno di campionamento e l'effetto stratificazione o l'effetto clustering;

b) tra l'effetto stratificazione e l'effetto clustering.

Per facilitare la lettura degli sviluppi metodologici successivi introduciamo il diagramma seguente:



composto di un insieme di vertici-che rappresentano le strutture campionarie introdotte nei paragrafi precedenti- e di segmenti orientati congiungenti i vertici-che rappresentano gli effetti già esaminati e gli effetti $(\widehat{deft})_s$ e $(\widehat{deft})_c$ il cui significato sarà chiarito in seguito.

Più precisamente, il significato dei simboli posti nei vertici del diagramma é il seguente:

$\{S(2,s); \hat{\sigma}(\hat{Y})\}$ = disegno a due stadi con stratificazione delle unità di primo stadio in cui $\hat{\sigma}(\hat{Y})$ indica la corrispondente stima dell'errore di campionamento ottenibile dalla formula (I7);

$\{S(2); \hat{\sigma}(\hat{Y}_{\bar{s}})\}$ = disegno a due stadi semplici in cui $\hat{\sigma}(\hat{Y}_{\bar{s}})$ indica la corrispondente stima dell'errore di campionamento;

$\{S(I,s); \hat{\sigma}(\hat{Y}_{\bar{c}})\}$ = disegno ad uno stadio stratificato in cui $\hat{\sigma}(\hat{Y}_{\bar{c}})$ indica la corrispondente stima dell'errore di campionamento;

$\{S ; \hat{\sigma}(\hat{Y}_{\bar{s},\bar{c}})\}$ = campione casuale semplice in cui $\hat{\sigma}(\hat{Y}_{\bar{s},\bar{c}})$ indica la corrispondente stima dell'errore di campionamento.

Prima di iniziare la nostra indagine appare pure utile trascrivere le espressioni di $\hat{E}_s$, $\hat{E}_c$ e $\widehat{deft}$:

(39) $$\hat{E}_s = \frac{\hat{\sigma}(\hat{Y})}{\hat{\sigma}(\hat{Y}_{\bar{s}})}$$

(66) $$\hat{E}_c = \frac{\hat{\sigma}(\hat{Y})}{\hat{\sigma}(\hat{Y}_{\bar{c}})}$$

$$(86) \qquad \widehat{deft} = \frac{\hat{\sigma}(\hat{Y})}{\hat{\sigma}(\hat{Y}_{\bar{s},\bar{c}})}$$

Nella trattazione che segue considereremo dapprima le relazioni di cui al punto a) e successivamente quelle di cui al punto b).

Per ottenere quanto ci proponiamo é sufficiente sostituire nella (86) la (39) o la (66); si ottiene:

$$(87) \qquad \widehat{deft} = \frac{\hat{\sigma}(\hat{Y})}{\hat{\sigma}(\hat{Y}_{\bar{s},\bar{c}})} = \frac{\hat{E}_s \cdot \hat{\sigma}(\hat{Y}_{\bar{s}})}{\hat{\sigma}(\hat{Y}_{\bar{s},\bar{c}})} = \hat{E}_s \cdot (\widehat{deft})_c$$

in cui si é posto

$$(88) \qquad (\widehat{deft})_c = \frac{\hat{\sigma}(\hat{Y}_{\bar{s}})}{\hat{\sigma}(\hat{Y}_{\bar{s},\bar{c}})}$$

e

$$(89) \qquad \widehat{deft} = \frac{\hat{\sigma}(\hat{Y})}{\hat{\sigma}(\hat{Y}_{\bar{s},\bar{c}})} = \frac{\hat{E}_c \cdot \hat{\sigma}(\hat{Y}_{\bar{c}})}{\hat{\sigma}(\hat{Y}_{\bar{s},\bar{c}})} = \hat{E}_c \cdot (\widehat{deft})_s$$

in cui si é posto

$$(90) \qquad (\widehat{deft})_s = \frac{\hat{\sigma}(\hat{Y}_{\bar{c}})}{\hat{\sigma}(\hat{Y}_{\bar{s},\bar{c}})}$$

Ai parametri (88) e (90) é possibile dare una interpretazione statistica che ne chiarisce l'importanza e il significato.

Il parametro $(\widehat{deft})_c$ rappresenta la perdita di precisio

ne-determinata dall'introduzione dell'estrazione a più stadi ma, più ancora, dall'effetto di clustering dentro le unità di primo stadio-che si ha rispetto ad un campione casuale semplice di uguale numerosità. In pratica, $(\hat{deft})_c$ rappresenta il coefficiente per cui moltiplicare l'errore di campionamento di un campione casuale semplice al fine di ottenere quello di un campione a due stadi semplici di pari numerosità.

Il parametro $(\hat{deft})_s$ rappresenta il guadagno, in termini di riduzione degli errori campionari, che si ottiene passando da un campione casuale semplice a quello ad uno stadio stratificato di uguale dimensione.

Con lo studio fin qui svolto abbiamo mostrato che $\hat{deft}$ può scindersi nel prodotto di due componenti secondo le relazioni (87) o (89). Tali relazioni possono essere anche illustrate mediante il diagramma precedente, nel quale é possibile individuare i seguenti due "cammini":

$$\left\{S\ ;\ \hat{\sigma}(\hat{Y}_{\bar{s},\bar{c}})\right\} \xrightarrow{(\hat{deft})_c} \left\{S(2);\ \hat{\sigma}(\hat{Y}_{\bar{s}})\right\} \xrightarrow{\hat{E}_s} \left\{S(2,s);\hat{\sigma}(\hat{Y})\right\}$$

$$\left\{S\ ;\ \hat{\sigma}(\hat{Y}_{\bar{s},\bar{c}})\right\} \xrightarrow{(\hat{deft})_s} \left\{S(I,s);\ \hat{\sigma}(\hat{Y}_{\bar{c}})\right\} \xrightarrow{\hat{E}_c} \left\{S(2,s);\hat{\sigma}(\hat{Y})\right\}$$

In ciascuno di tali cammini, l'effetto complessivo del disegno di campionamento é il prodotto dei due effetti parziali mediante i quali si passa dalla struttura campionaria più "semplice" $\left\{S;\ \sigma(\hat{Y}_{\bar{s},\bar{c}})\right\}$ a quella più "complessa" $\left\{S(2,s);\ \hat{\sigma}(\hat{Y})\right\}$ attraverso la struttura "intermedia" $\left\{S(2);\ \hat{\sigma}(\hat{Y}_{\bar{s}})\right\}$ o $\left\{S(I,s);\ \hat{\sigma}(\hat{Y}_{\bar{c}})\right\}$.

A completamento del paragrafo siamo ora in grado di esprimere il legame tra $\hat{E}_s$ e $\hat{E}_c$; a tal fine ponendo a confronto la (87) con la (89) si trae la relazione:

$$\hat{E}_s \ (\widehat{deft})_c = \hat{E}_c \ (\widehat{deft})_s \tag{9I}$$

da cui seguono le due relazioni:

$$\hat{E}_s = \hat{E}_c \frac{(\widehat{deft})_s}{(\widehat{deft})_c} \tag{92}$$

$$\hat{E}_c = \hat{E}_s \frac{(\widehat{deft})_c}{(\widehat{deft})_s} \tag{93}$$

Tenendo poi presente le formule (39) e (66) si ottengono le relazioni alternative:

$$\hat{E}_s = \hat{E}_c \frac{\hat{\sigma}(\hat{Y}_{\bar{c}})}{\hat{\sigma}(\hat{Y}_{\bar{s}})} \tag{94}$$

$$\hat{E}_c = \hat{E}_s \frac{\hat{\sigma}(\hat{Y}_{\bar{s}})}{\hat{\sigma}(\hat{Y}_{\bar{c}})} \tag{95}$$

5. Futuri itinerari di studio

Nel presente lavoro abbiamo proposto-nell'ambito della problematica della valutazione dei piani di campionamento-una metodologia che consente di inquadrare in una visione unitaria lo studio dell'effetto stratificazione, dell'effetto clustering e dell'effetto complessivo del disegno di campionamento, metodologia quindi che conduce a stime più soddisfacenti, dal punto di vista statistico, di tali effetti.

L'impostazione metodologica suggerita si riferisce al caso di indagini campionarie basate su disegni a due stadi con stratificazione delle unità primarie, con estrazione casuale semplice e senza reimmissione in ciascuno stadio, nonché al caso di stimatori diretti.

Tale impostazione verrà successivamente estesa alle indagini che utilizzano stimatori indiretti (del rapporto, di regressione) e campioni in cui le unità primarie vengono estratte con probabilità variabili.

Un altro punto da approfondire, che in un certo senso costituisce una naturale estensione della nostra ricerca, é quello della messa a punto di un idoneo metodo statistico di stima dell'effetto ponderazione e della conseguente determinazione dei legami esistenti tra quest'effetto e gli effetti già trattati in questa nota; questi problemi saranno argomento di un prossimo articolo.

Allo scopo di valutare il grado di approssimazione dei criteri adottati da Verma, Scott e O'Muircheartaigh -già da noi utilizzati nelle indagini cui si é fatto cenno nell'Introduzione- sarà tentata un'applicazione dei procedimenti di stima qui esposti alle suddette indagini.

BIBLIOGRAFIA

[I] Cochran W.G.(I977),Sampling techniques,Wiley,New York.

[2] Hansen M.H.,Hurwitz W.N.,Madow W.G.(I953),Sample Survey Methods and Theory,Vol.II,Wiley,New York.

[3] Kish L.(I965),Survey Sampling,Wiley,New York.

[4] Napolitano P.,Russo A.,Zannella F.(I983),Calcolo,presentazione ed analisi degli errori di campionamento nell'indagine ISTAT sulle condizioni di salute della popolazione e sul ricorso ai servizi sanitari,Atti del Convegno S.I.S. ,Trieste.

[5] Russo A.(I984),Calcolo ed analisi degli errori di campionamento nell'indagine ISTAT sulle vacanze e gli sports degli italiani, gennaio I983,Atti della XXXII Riunione scientifica della S.I.S., Sorrento.

[6] Verma V.,Scott C.,O'Muircheartaigh C.(I980),Sample designs and sampling errors for the World Fertility Survey,J.R.Statist.Soc. A.I43,Part.4.

[7] Yamane T. (1967), Elementary Sampling Theory, Prentice-Hall, Inc. Englewood Cliffs, N.J.

"QUADERNI DI DISCUSSIONE PUBBLICATI"

84.01 REY G. M.
Le statistiche ufficiali e l'attività della Pubblica Amministrazione
Giugno 1984

85.01 CRESCENZI F.
Nota su alcune metodologie per la classificazione di unità territoriali
Febbraio 1985

85.02 CORTESE A.
Alcune considerazioni sulle prospettive del censimento della popolazione
Marzo 1985

85.03 MATURANI G.
Stima delle ore di lavoro effettivamente prestate dai lavoratori occupati negli anni 1960-1983
Aprile 1985

85.04 NAPOLITANO P.
Esposizione di alcune tecniche per la investigazione dei dati
Maggio 1985

85.05 RUSSO A.
Su un metodo di stima degli effetti stratificazione e clustering e dell'effetto complessivo del disegno di campionamento nei campioni a due stadi con stratificazione delle unità di primo stadio
Settembre 1985