manuale di tecniche di indagine

4 - tecniche di campionamento: teoria e pratica

istituto nazionale di statistica

note e relazioni
anno 1989 n. 1

La preparazione del Fascicolo e il coordinamento redazionale dei testi sono stati curati da Aldo Russo
Autore dei Capitoli: 1, 2, 3, 13 Mario Di Traglia
4, 8, 12, 14 Aldo Russo
5, 10 Stefano Falorsi
6, 7 Piero Demetrio Falorsi
9, 11 Giuliana Coccia

Editing di
Mario Nanni e Claudio Antonio Pajer

Supplemento all'Annuario Statistico Italiano

ISSN: 0535-9856

abete grafica s.p.a. - Roma - Contratto n. 14762 del 6-8-1988 - copie 3.000

INDICE

PRESENTAZIONE

Il *Manuale di tecniche di indagine* la cui preparazione è stata curata dal Reparto Studi dell'Istituto, si configura come guida per la razionalizzazione delle operazioni di rilevazione ed è stato pure concepito quale strumento didattico da utilizzare ai fini della formazione dei funzionari dell'Istat. Poiché nell'effettuazione di indagini statistiche sono impegnati molti altri organismi pubblici e privati, si ritiene che esso possa costituire uno strumento utile anche per l'attività di questi organismi, in particolare di quelli che hanno un qualche ruolo nel sistema informativo socio-economico del Paese.

Il *Manuale* prende in esame i vari segmenti del *ciclo produttivo* nei quali si sviluppa normalmente ogni indagine statistica cogliendo aspetti che vanno dalla costruzione del disegno campionario al controllo della qualità dei dati, dall'analisi delle caratteristiche delle varie tecniche di indagine alla definizione di criteri standardizzati per la presentazione dei risultati. Pensato inizialmente per le indagini condotte con il metodo del campione, in particolare per quelle sulle famiglie, nella sua definitiva articolazione esso detta norme valide per fasi di lavoro riscontrabili nelle rilevazioni totali ed allarga pertanto il suo campo di applicazione che finisce per comprendere le generalità delle indagini.

La sua impostazione riflette il desiderio di colmare il divario fra il *libro di testo* ed il *manuale operativo*. Se da un lato infatti non si rinuncia al rigore della formalizzazione e si introducono spunti di innovazione sul piano metodologico, dall'altro si tengono ben presenti le esigenze del lavoro sul campo e risulta quindi ampio lo spazio riservato alle esemplificazioni.

Il *Manuale* consta dei seguenti fascicoli:

1. **Pianificazione della produzione di dati**
2. **Il questionario: progettazione, redazione, verifica**
3. **Tecniche di somministrazione del questionario**
4. **Tecniche di campionamento: teoria e pratica**
5. **Tecniche di stima della varianza campionaria**
6. **Il sistema di controllo della qualità dei dati**
7. **Le rappresentazioni grafiche di dati statistici**

In ogni caso va precisato che il *Manuale* non è da considerarsi completato in quanto è previsto che ai fascicoli programmati se ne aggiungano altri mano a mano che l'attività di ricerca avrà portato a termine l'esplorazione di aspetti per ora solo individuati.

PARTE 1

INTRODUZIONE ALLA TEORIA DEL CAMPIONAMENTO

CAPITOLO 1 - CONSIDERAZIONI GENERALI

Introduzione

La fase iniziale nello studio dei fenomeni statistici è costituita dai processi di raccolta delle informazioni sulle modalità attraverso le quali i fenomeni stessi si manifestano; da tali processi scaturiscono, alla fine, i dati statistici elementari.

Una prima suddivisione nello studio dei fenomeni statistici (siano essi naturali, ambientali, socio-demografici, economici, ecc.), sulla base dei processi di raccolta dei dati, è in: i) Indagine statistica; ii) Esperimento statistico.

L'indagine statistica è caratterizzata dalle seguenti peculiarità:

— esistenza di unità individuabili sulle quali effettuare la misurazione (rilevazione) della caratteristica (o delle caratteristiche) di interesse;
— l'insieme di tali unità, denominato popolazione, è costituito da un numero finito di elementi.

L'esperimento statistico è invece caratterizzato dai seguenti aspetti:

— le osservazioni (dati) riguardano risultati di prove (ad esempio, testa e croce nel lancio di una moneta) o di esperimenti (ad esempio, dosi-risposte nella sperimentazione dei farmaci);
— non è possibile parlare di popolazione finita, in quanto è teoricamente possibile prolungare all'infinito il numero delle prove o degli esperimenti stessi.

Talvolta i due processi vengono indicati con i termini *situazioni osservazionali*, nel caso di indagine statistica e di *situazioni sperimentali* nel caso di esperimenti statistici (Coppi, 1979). In questo volume ci occuperemo esclusivamente dell'indagine statistica. In una indagine statistica, si hanno a disposizione N unità, che costituiscono la popolazione oggetto di indagine; la popolazione è indicata con la lettera *U*, mentre una sua generica unità sarà denotata con la lettera *u*.

Ad ogni unità *u* di *U* viene associato un numero identificativo *i* ($i = 1, \ldots, N$); pertanto, ad ogni i corrisponde l'unità u_i di *U* e viceversa.

Introduciamo inoltre i simboli y e Y_i per indicare rispettivamente la caratteristica da rilevare (variabile oggetto di indagine) ed il valore osservato sulla generica unità u_i.

Se si è interessati allo studio di un solo carattere, e quindi l'indagine viene condotta su una sola caratteristica di *u*, allora y

è una variabile unidimensionale; se invece si intende misurare su ciascuna unità caratteristiche diverse, allora y sarà multidimensionale. Se, ad esempio, interessa misurare k caratteristiche, la variabile oggetto di indagine può essere formalmente espressa da:

$$\mathbf{y} = (y_1, \ldots, y_j, \ldots\ldots, y_k) \qquad (1)$$

ed il valore osservato sull'unità i da:

$$\mathbf{Y}_i = (Y_{i1}, \ldots\ldots, Y_{ij}, \ldots\ldots, Y_{ik}) \qquad (2)$$

Ad esempio, con riferimento ad una ipotetica indagine sui consumi delle famiglie immaginiamo che i caratteri da rilevare siano:

y_1 = reddito;
y_2 = spesa per generi alimentari;
y_3 = numero di componenti il nucleo familiare.

In tal caso, la variabile oggetto di indagine sarà:

$$\mathbf{y} = (y_1, y_2, y_3) \qquad (3)$$

Sulla generica famiglia i, dell'insieme delle N famiglie, una particolare modalità della variabile **y** assumerà la forma:

$$\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}) \qquad (4)$$

Altri due aspetti molto importanti nelle indagini statistiche sono: il periodo di riferimento delle variabili oggetto di studio ed il periodo di rilevazione.

Per periodo di riferimento si intende l'istante o l'intervallo di tempo cui viene riferita l'osservazione della variabile (o delle variabili) oggetto di studio.

Nel caso dell'ultimo censimento della popolazione, ad esempio, le variabili relative al foglio di famiglia sono state rilevate con riferimento alla data del 25 ottobre 1981; nei casi in cui era necessario ricorrere ad un momento preciso ci si è riferiti alla mezzanotte tra il 24 ed il 25 ottobre 1981.

Ad esempio, i bambini nati prima della mezzanotte del 24 dovevano essere censiti, quelli nati dopo la mezzanotte del 24 ottobre dovevano essere esclusi dal censimento.

Nel caso dell'indagine Istat sui consumi delle famiglie (ISTAT, 1988) il periodo di riferimento, relativamente alla spesa per generi alimentari, è la decade del mese. Più precisamente, su un campione di famiglie viene rilevata la spesa sostenuta per pane, pasta, ecc., nei primi dieci giorni del mese di gennaio; su un secondo campione di famiglie, la spesa sostenuta nella seconda decade di gennaio; e così via per le restanti decadi dei mesi successivi.

Il periodo di rilevazione viene definito come l'intervallo di tempo in cui viene materialmente effettuata la rilevazione.

Riprendendo l'esempio relativo all'ultimo censimento della popolazione, il periodo di rilevazione era definito dall'intervallo 25 ottobre - 11 novembre, 1981.

La popolazione

Nel momento in cui si decide di effettuare un'indagine statistica occorre individuare gli obiettivi che si vogliono perseguire.

Per obiettivi si deve intendere il complesso delle informazioni che l'indagine dovrà fornire.

Questi possono dipendere dall'insieme delle informazioni disponibili a priori per ogni unità di *U,* dalla possibilità di individuare le unità da rilevare, dall'esistenza di liste, ecc..

Dalla individuazione degli obiettivi discende la definizione di *unità di analisi* e quindi della *popolazione oggetto d'indagine* (o di studio).

Se ad esempio, con riferimento alla popolazione residente di una data area geografica, l'obiettivo della indagine è quello di ottenere il reddito medio per sesso e classi di età ($15 \vdash 45$; $45 \vdash 70$; $70 \vdash$), l'unità di analisi è l'individuo, e la popolazione oggetto d'indagine è l'insieme degli individui di età maggiore di 14 anni.

Dopo aver definito la popolazione oggetto di indagine e prima di effettuare la rilevazione sulle unità definite, è necessario disporre di una lista attraverso la quale identificare e raggiungere le unità stesse. Bisogna tuttavia sottolineare che non sempre è disponibile una lista della popolazione oggetto di indagine, per cui si rende necessario l'utilizzazione di liste costituite da unità diverse da quelle di analisi, che consentano però di raggiungere queste ultime.

Le unità costituenti la lista sono chiamate *unità di campionamento.*

Con riferimento all'esempio precedente, si supponga di non disporre di una lista di individui ma di una lista delle famiglie residenti in quell'area geografica.

Tale lista verrà utilizzata per raggiungere ed intervistare gli individui appartenenti alla popolazione oggetto di indagine.

Nelle situazioni concrete, tuttavia, non sempre è possibile rilevare, qualunque sia la lista, tutte e soltanto le unità costituenti la popolazione suddetta. L'insieme delle unità individuabili attraverso la lista costituisce la *popolazione base.*

Riprendendo ancora l'esempio precedente, può accadere sia che la lista delle famiglie non contenga tutte quelle effettivamen-

te residenti nell'area geografica, sia che contenga delle famiglie che risiedono di fatto in altre aree; conseguentemente si ha che la popolazione base può non coincidere con la popolazione oggetto di indagine.

È da sottolineare, a questo punto, che le due popolazioni sinora definite sono del tutto teoriche, nel senso che esse non rappresentano (per il fenomeno della mancata intervista) ancora l'insieme delle unità sulle quali vengono osservate le modalità delle variabili oggetto di studio. Tale insieme, nella letteratura statistica sull'argomento, è noto sotto il nome di *popolazione osservata*.

È possibile, infine, che non tutte e soltanto le unità della popolazione osservata vengano considerate ai fini dell'ottenimento degli obiettivi dell'indagine, in quanto il processo di controllo della qualità dei dati può determinare l'aggiunta o l'eliminazione di una o più unità; quest'ultima popolazione viene denominata *popolazione indotta*.

Dalle considerazioni svolte discende, pertanto, che la definizione di popolazione oggetto di indagine è una operazione tanto importante dal punto di vista operativo quanto complessa dal punto di vista teorico. Nella generalità dei casi concreti, quindi, la popolazione utilizzata per la determinazione dei risultati dell'indagine può differire da quella oggetto di studio (Kish, 1965).

A margine è utile accennare ad alcuni particolari casi di indagini statistiche in cui, definiti gli obiettivi dell'indagine, risulta impossibile o estremamente complesso determinare la popolazione oggetto d'indagine e/o quella base.

Se, ad esempio, interessa misurare il grado di inquinamento atmosferico su un determinato territorio o di presenza di sostanze chimiche nel terreno, le unità di analisi vengono a determinarsi sulla base dei volumi di aria o di terreno analizzabili, e le unità di rilevazione sono le aree individuate da apposite suddivisioni (reticoli) del territorio di riferimento; questo tipo di indagine viene chiamata *indagine areale*. La popolazione è formata dall'insieme degli elementi del reticolo, mentre la lista diventa la successione generata dalla numerazione degli elementi stessi.

Esistono poi le indagini biologiche, in cui la popolazione è formata da batteri o virus, e quelle fisiche, in cui la popolazione può essere formata da molecole, atomi, ecc; anche per queste indagini sussistono le difficoltà accennate.

Parametri oggetto di indagine

Come accennato precedentemente, gli obiettivi di una indagine statistica sono la conoscenza di alcuni parametri di interesse della popolazione (ad esempio, reddito medio, numero di occupati, ecc.). Il raggiungimento di tali obiettivi avviene attraverso

la raccolta di misurazioni delle modalità assunte dalla variabile oggetto d'indagine sulle unità di *U*. Successivamente, l'insieme delle informazioni raccolte deve essere elaborato, sintetizzato e presentato in una forma (grafica, tabellare, ecc.) atta a fornire gli obiettivi dell'indagine stessa.

Più in generale, definiamo parametro oggetto di indagine una funzione $F[\cdot]$ delle modalità osservabili sull'insieme degli elementi di *U*, che indichiamo con:

$$\theta = F[Y_1, \ldots, Y_i, \ldots, Y_N] \tag{5}$$

I parametri il cui calcolo è più frequentemente richiesto sono: totale, media e rapporto tra due totali (Fabbris, 1989).

Per il totale, la funzione $F[\cdot]$ diventa:

$$Y = \sum_{i=1}^{N} Y_i \tag{6}$$

Per la media si ha:

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i \tag{7}$$

Per il rapporto tra due totali (o tra due medie)

$$R = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} \tag{8}$$

Esistono inoltre diverse tipologie di variabili oggetto di indagine; una generale suddivisione può comunque effettuarsi in *variabili quantitative* e *variabili qualitative*. Le variabili quantitative possono suddividersi in *discrete* e *continue*, mentre le qualitative in *ordinabili* e *non ordinabili*. Per quanto riguarda le variabili continue è appena il caso di sottolineare che è una distinzione del tutto teorica, in quanto le variabili che interessano le inda-

gini concrete vengono sempre rese discrete attraverso opportuni arrotondamenti.

Il parametro Y diventa l'ammontare totale di un carattere in *U* se tale carattere è quantitativo; una frequenza assoluta delle unità con una data modalità del carattere y, se è qualitativo.

Il parametro $\overline{Y}$ diventa invece: la media, se y è quantitativo; una frequenza relativa, delle unità che possiedono le modalità di interesse, se y è qualitativo.

Allo scopo di chiarire quanto appena detto consideriamo il seguente esempio.

Con riferimento all'indagine Istat sui consumi delle famiglie (ISTAT, 1988) tra le variabili oggetto di studio è compresa la variabile y: numero di componenti della famiglia.

Essa può essere decomposta in:

y_1: numero di componenti = 1
y_2: numero di componenti = 2
y_3: numero di componenti = 3
e così via.

Ognuna di queste ultime variabili diventa in tal modo un carattere qualitativo, denominato *variabile indicatrice dell'insieme di componenti della famiglia*, che assume valore 1 oppure 0 a seconda che la famiglia osservata abbia (oppure no) la modalità 1, 2, 3, ..., ecc..

In altri termini, con riferimento ad esempio a y_1, si ha:

$$y_1 = \begin{cases} 1 & \text{se la famiglia ha un solo componente} \\ 0 & \text{se la famiglia ha un numero di componenti diverso da 1} \end{cases}$$

Con riferimento alla variabile y, la relazione:

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i \qquad (9)$$

assume il significato di *numero medio di individui per famiglia*, mentre l'espressione:

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_{1i} \qquad (10)$$

riferita alla variabile y_1, rappresenta la *frequenza relativa delle famiglie con un numero di componenti pari ad 1*.

Per quanto riguarda il totale del carattere y si ha

$$Y = \sum_{i=1}^{N} Y_i \qquad (11)$$

indica l'*ammontare di individui* nella popolazione. L'espressione:

$$Y = \sum_{i=1}^{N} Y_{1i} \qquad (12)$$

fornisce invece la *frequenza assoluta delle famiglie con un componente*.

Indagini totali ed indagini campionarie

Una valutazione esatta dei parametri descritti può aversi (se viene ipotizzata l'assenza di *errori di misura*) soltanto osservando tutte le unità di U; in altri termini, conducendo una indagine denominata *Totale* o *Censuaria*. Le indagini totali consistono nell'osservare tutte le unità di U, attraverso uno specifico piano di rilevazione, dopo averle individuate utilizzando gli strumenti esposti precedentemente (liste, aree, strade, ecc.).

Se, invece, si è disposti ad avere una valutazione del parametro di interesse in termini di incertezza è allora possibile condurre una indagine compiendo le osservazioni della variabile su una parte delle unità di *U*; questo tipo di indagine viene chiamata *parziale* o *campionaria*. Solitamente si pensa che l'indagine censuaria consenta di arrivare ad una valutazione certa del parametro di interesse; in realtà non è esattamente così.

Infatti, numerosi sono i fattori di errore che intervengono nelle varie fasi di espletamento dell'indagine stessa ed il loro effetto può essere di maggiore influenza in una indagine totale che non in una indagine campionaria.

Indichiamo, come già fatto in precedenza, con:

$$\theta^* = \theta + e(N) = F\,[Y_1, \ldots, Y_i, \ldots, Y_N] + e(N) \qquad (13)$$

il parametro oggetto di indagine osservato; il termine e(N) indica l'errore che può essere commesso durante tutte le fasi di preparazione ed espletamento dell'indagine, ossia l'errore di misura.

Nel caso di indagine campionaria, oltre all'errore di misura, è presente anche l'errore campionario dovuto alla natura parziale della rilevazione. Per tali indagini, quindi, si ha un *errore totale*, simbolicamente espresso da:

$$e_T = \varepsilon(N, n, S^2) + e(N, n) \qquad (14)$$

dove $\varepsilon(N, n, S^2)$ indica l'errore campionario visto come funzione della dimensione del campione n e della variabilità del carattere S^2 (Cochran, 1977; Signore, 1988).

In questa sede verranno trattati soltanto quegli errori che derivano dalla mancata osservazione di una parte delle unità di U, ossia gli errori campionari.

La forma della (14) è di solito additiva poiché le due componenti, in genere, sono indipendenti; non si possono comunque escludere a priori forme diverse da quella additiva.

Per quanto riguarda le indagini censuarie le loro caratteristiche positive possono così riassumersi:

— consentono di ottenere una valutazione del parametro θ con un elevato grado di accuratezza, purché venga prestata massima attenzione agli aspetti organizzativi ed ai processi di misurazione e registrazione dei dati;
— consentono di costruire e controllare le liste.

Le caratteristiche negative invece sono:

— costo elevato;
— tempi di preparazione dell'indagine e di esecuzione ed elaborazione dei dati elevati;
— minor dettaglio nelle informazioni rilevate;
— complessità nella gestione dei records.

È evidente che quelle che sono caratteristiche negative per una indagine censuaria diventano positive per una indagine campionaria.

Bisogna comunque aggiungere che l'indagine censuaria risulta necessaria per rilevare, ad intervalli di tempo, quelle variabili strutturali utili alla conoscenza dei principali fenomeni riguardanti la popolazione; i risultati possono essere poi utilizzati per condurre indagini campionarie più efficienti su particolari fenomeni di interesse. L'efficienza che deriva dall'utilizzo delle informazioni censuarie dipende (come del resto l'intervallo intercensuario) dalla dinamica temporale delle variabili strutturali.

L'indagine per campione

Nella letteratura statistica con il termine *campione* si intende un sottoinsieme di U.

Come detto in precedenza, per procedere ad una indagine statistica (totale o campionaria) bisogna disporre di unità identificabili attraverso una lista o base dell'indagine. La lista viene

creata attribuendo a ciascuna unità di U un numero identificativo i, per cui scrivere ad esempio:

$$U = \{ 1, 2, 3, 4 \} \qquad (15)$$

equivale ad indicare, qualunque cosa esse rappresentino (famiglie, individui, aree, ecc.), le unità di U con i numeri da 1 a 4. Viene da chiedersi se esistono criteri per assegnare un numero ad una unità invece che ad un'altra, ma questo porta a considerazioni che esulano dagli scopi del manuale. Estrarre un campione di n unità da U significa estrarre n numeri da:

$$L = \{ 1, 2, 3, \ldots\ldots, N \} \qquad (16)$$

($n < N$) ed osservare (intervistare) le unità di U che corrispondono all'insieme selezionato:

$$s = \{ i_1, \ldots\ldots, i_n \}; \qquad i = 1, \ldots, N \qquad (17)$$

Si ottiene quindi l'insieme dei valori osservati (non tutti distinti), che rappresenta il risultato campionario:

$$Y(s) = \{ Y_1, \cdots, Y_i, \cdots, Y_n \} \qquad (18)$$

L'insieme di tutti gli insiemi di unità che è possibile formare, avendo a disposizione una popolazione di N unità, viene chiamato universo campionario ed indicato con U_c. Ad esempio, se:

$$U = \{ 1, 2, 3, 4 \} \qquad (19)$$

$$U_c = [\{1\}, \{2\}, \{3\}, \{4\}, \{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\}, \{1,2,3\}, \{1,2,4\}, \{1,3,4\}, \{2,3,4\}, \{1,2,3,4\}] \qquad (20)$$

Una prima suddivisione nella tipologia di campioni, per lo studio del carattere y, può essere effettuata sulla base del tipo di meccanismo utilizzato nella scelta delle unità di *U* da osservare. A seconda che le unità appartenenti ad s siano state scelte per certe loro caratteristiche specifiche od a caso, si hanno le due seguenti procedure di campionamento (e quindi due diverse tipologie di campioni):

— campionamento *ragionato*;
— campionamento *casuale*.

Nel primo caso vengono scelte, come campione, delle particolari unità di *U* sulla base di informazioni a priori sulle unità stesse, in modo da soddisfare alcune predeterminate esigenze;

nel secondo caso, invece, le unità da rilevare sono scelte a caso.

Il campionamento ragionato

I motivi che possono indurre lo statistico a ricorrere ad un campione ragionato sono di varia natura, ma tutti comunque rispondono all'esigenza di controllare il risultato ottenuto Y(s) sulla base di informazioni a priori sulle unità di *U*.

Per meglio chiarire tale concetto viene, di seguito, presentato un esempio in cui si ha una popolazione di 8 unità (N = 8) ed un campione di 2 unità (n = 2). Il carattere oggetto di indagine viene indicato con y, mentre il carattere ausiliario (informazione a priori) viene indicato con x.

Sia

$$\{1,2,3,4,5,6,7,8\} \tag{21}$$

la popolazione oggetto d'indagine e sia:

$$\{0,1,1,0,1,1,0,1\} \tag{22}$$

l'insieme dei valori che il carattere y (supposto dicotomico) assume in *U*.

Supponiamo di conoscere le modalità assunte dal carattere ausiliario x su tutte le unità di *U* e di essere a conoscenza inoltre che tra i caratteri y ed x esiste un «legame» sufficientemente elevato.

Sia poi:

$$\{0,0,1,0,1,0,1,1\} \tag{23}$$

l'insieme dei valori assunti dal carattere *x* in *U*.

Nel nostro caso, il significato del termine legame è che le modalità dei caratteri y ed x sulle stesse unità di *U* risultano, con frequenza elevata, identiche. In altri termini

$$P\{y(i) = 1 | x(i) = 1\} > P\{y(i) = 1\}$$

dove il simbolo P indica la probabilità.

Poiché è conosciuto a priori che esiste un legame abbastanza stretto tra x ed y (infatti le unità che presentano lo stesso valore sia della y che della x sono 7 su 8) si può scegliere un campione in modo che risulti:

$$\hat{X} = \frac{N}{n} \sum_{i=1}^{n} X_i = \sum_{i=1}^{N} X_i = X = 4 \tag{24}$$

Infatti, scegliendo:

$$s = \{3, 4\}$$

si ha:

$$X(s) = \{0,1\}$$

e quindi:

$$\hat{X} = \frac{8}{2}(1 + 0) = 4 = X$$

$$\hat{Y} = \frac{8}{2}(1 + 0) = 4 \neq Y = 5$$

Come si vede, la stima campionaria $\hat{Y}$ risulta diversa dal valore vero del parametro Y, ma abbastanza vicina ad esso. Inoltre, nel nostro caso, sono pochi i campioni in U_c in cui $\hat{X} = X$ e, contemporaneamente $\hat{Y}$ molto diverso da Y.

Come si vedrà meglio in seguito, anche non adottando un campione ragionato, esiste la possibilità di utilizzare l'informazione disponibile su x.

Campionamento casuale

Nella popolazione *U*, su ciascuna unità *u*, la caratteristica oggetto di indagine y, assume una determinata modalità Y.

L'insieme delle modalità assunte dall'insieme delle unità {u} viene simbolicamente indicato con:

$$Y(L) = (Y_1, \cdots, Y_i, \cdots, Y_N) \tag{25}$$

che indica un ben determinato punto nell'insieme di tutti i possibili punti derivanti dall'associazione, a ciascuna delle unità di *U*, delle possibili modalità del carattere y.

L'insieme Y(L) è, in altri termini, *deterministico*, nel senso che $Y_1, \cdots\cdots, Y_N$ sono valori certi (se pur sconosciuti) del carattere y. Il parametro oggetto di indagine è, come già visto, una funzione di Y(L), mentre una stima campionaria di θ risulta una funzione dei dati campionari Y(s):

$$\hat{\theta} = F[Y(s)] \tag{26}$$

Supponiamo che {s} non sia stato scelto aleatoriamente; poiché:

$$Y(s) = (Y_1, \ldots, Y_n) \qquad (27)$$

è un insieme di valori certi di y, allora la (27) è una funzione deterministica; ossia $\hat{\theta}$ non può essere considerata una variabile aleatoria. Questo fatto risulta abbastanza grave in quanto, non essendo $\hat{\theta}$ una variabile aleatoria, non è possibile applicare la teoria dell'inferenza statistica; teoria che è alla base dello studio dei fenomeni statistici tramite le tecniche campionarie.

Per fare in modo che $\hat{\theta}$ sia una variabile aleatoria bisogna rendere casuale l'insieme Y(s); ciò viene realizzato semplicemente rendendo aleatorio l'insieme s delle etichette che individuano le unità di *U* che faranno parte del campione.

È allora evidente che, scegliere a caso le unità di *U* che faranno parte del campione, significa scegliere a caso l'insieme s in modo che Y (s) rappresenti una realizzazione della variabile aleatoria y.

Poiché $\hat{\theta}$ è una funzione di Y(s) di conseguenza risulta essa stessa una variabile aleatoria, la quale (dato il suo ruolo di *stimatore* del parametro θ) gode di particolari proprietà di cui verrà trattato nel prosieguo. Il motivo quindi di *casualizzare* la scelta delle unità di *U* risiede nel rendere applicabile la teoria dell'inferenza statistica al campionamento da popolazioni finite.

L'inferenza è, come ben noto, quel processo attraverso il quale, disponendo di un numero limitato di osservazioni (dati campionari) di un certo fenomeno (carattere oggetto di indagine), è possibile fare delle valutazioni più generali (sulla popolazione), la cui validità è misurabile in termini probabilistici.

La validità di tali valutazioni dipende, in larga misura, dal modo in cui vengono scelte le unità da osservare. Questa fase viene definita *disegno di campionamento*.

Una definizione più precisa di disegno di campionamento verrà fornita in seguito; per il momento esso viene inteso come una distribuzione di probabilità su U_c.

Ad esempio, la distribuzione di probabilità, indotta dalla scelta ragionata delle unità di *U* (campione ragionato), su U_c è:

$$d(s) = \begin{cases} 1 & s = s_o \\ 0 & s \neq s_o \end{cases} \qquad (28)$$

dove s_o rappresenta un insieme di unità di *U* caratterizzate da una qualche proprietà di interesse per l'indagine.

È da aggiungere che nel campionamento ragionato il disegno non ha una particolare rilevanza; quello che invece è rilevante sono le motivazioni che inducono a scegliere delle unità invece di altre.

CAPITOLO 2 - CONCETTI BASE DELLA TEORIA DEL DISEGNO DI CAMPIONAMENTO

Introduzione

Come accennato precedentemente, ciò che distingue l'indagine statistica dall'esperimento statistico è l'esistenza di una popolazione U, di unità ben definite, sulle quali rilevare le modalità Y_i assunte dal carattere oggetto di indagine y.

Questo fatto fa si che, nel caso dell'indagine campionaria, possa essere definito un insieme U_c di sottoinsiemi di {U} chiamato, come già detto, universo campionario.

Tale insieme, poiché le unità da osservare vengono scelte aleatoriamente, risulta dotato di una distribuzione di probabilità d (s), che associa ad ogni elemento «s_o» di U_c un numero d (s_o) che esprime la probabilità di osservare il campione s_o; d (s) viene chiamato disegno campionario.

Più precisamente, per disegno campionario si intende una distribuzione di probabilità su U_c tale che:

$$\begin{aligned} &d(s \mid N, n, Y) > 0 \quad \text{per} \quad (Y_1, \ldots, Y_n) \in Y(L),\ n(s) = n \\ &d(s \mid N, n, Y) = 0 \qquad \text{altrove} \end{aligned} \qquad (1)$$

dove n (s) indica la numerosità campionaria.

La coppia {d(s), t[Y(s)}, dove t [Y(s)] è una funzione dei dati campionari denominata *stimatore,* viene definita strategia di campionamento.

È appena il caso di aggiungere che in letteratura, a volte, per quest'ultima definizione, vengono utilizzati indifferentemente i termini *strategia campionaria* e piano di campionamento.

In questa sede l'insieme {d(s); t[Y(s)]} verrà sempre indicato con il termine *strategia di campionamento*.

Una ulteriore distinzione che è possibile fare riguarda la numerosità campionaria a seconda che sia fissata a priori oppure sia essa stessa una variabile aleatoria.

Sulla base di ciò vengono distinti i disegni a numerosità fissata ed i disegni a numerosità variabile.

Nella prima tipologia rientrano la maggior parte dei disegni campionari di uso più frequente presso l'Istat. Nella seconda tipologia rientrano, invece, i disegni campionari per il controllo industriale (disegni di campioni sequenziali) e per la rilevazione di

caratteristiche rare (campionamento inverso o poissoniano). In questa sede, sia pure mantenendo una trattazione adattabile ad entrambe le tipologie, verranno esposti i disegni a numerosità fissata.

Il campionamento casuale semplice

Per quanto detto precedentemente, la scelta del disegno campionario è effettuato dal ricercatore prima di condurre materialmente l'indagine. È evidente che, come ogni scelta, questa debba essere supportata da informazioni a priori: tali informazioni riguarderanno, ovviamente, le unità di U. La quantità e la qualità delle informazioni disponibili rende possibile l'applicazione di disegni semplici oppure complessi.

È quindi necessario definire il legame tra informazione a priori e disegno.

In una prima fase viene definito l'insieme minimo di informazione richiesto per l'effettuazione di una indagine statistica; tale insieme verrà genericamente indicato con la lettera maiuscola A.

Per effettuare una indagine statistica, l'insieme minimo risulta essere:

$$A = \{U, y, n\} \qquad (2)$$

In altri termini è possibile condurre una indagine statistica solo se sono univocamente definibili le unità su cui rilevare la caratteristica oggetto di indagine y, una popolazione di riferimento U ed il numero di unità da campionare n. Bisogna aggiungere che per i campioni a numerosità variabile non è necessario aver definito a priori la numerosità campionaria n.

La situazione (2) è abbastanza anomala nel campionamento da popolazioni finite, ma riteniamo comunque necessario trattarla per motivi di omogeneità.

Tale situazione è più vicina all'esperimento statistico che all'indagine statistica vera e propria. Poiché però esistono delle unità ben definite ed una popolazione U, non è possibile parlare di esperimento statistico. Il campionamento verrà quindi condotto individuando a caso delle unità di U e, su queste, rilevando la caratteristica di interesse y.

Secondo questo meccanismo agisce di fatto un disegno campionario semplice, nel senso che la probabilità di un campione risulta dipendere dalla numerosità (sconosciuta) di U e dal numero di unità osservate, dove però d (s) non può essere esplicitata.

Inoltre, l'unico parametro che è possibile stimare è la media della caratteristica y nella popolazione U. Verrà dimostrato in

seguito che, in tale situazione, lo stimatore ottimale nella classe degli stimatori lineari risulta essere la media aritmetica semplice delle osservazioni campionarie:

$$t\,[Y\,(s)] = \frac{1}{n} \sum_{i=1}^{n} Y_i \qquad (3)$$

In altri termini è possibile ottenere soltanto stime di medie (o proporzioni). Questo è reso evidente dal fatto che, per la stima di totali o frequenze assolute, è necessario conoscere l'ampiezza N della popolazione di riferimento. Dalla teoria statistica, inoltre, si ha che t gode delle seguenti proprietà:

— Per n abbastanza piccolo ($n < 30$) la variabile aleatoria:

$$Z = \frac{t - \theta}{\sigma} \qquad (4)$$

risulta distribuita secondo una *t* di Student con n-1 gradi di libertà;

— Per $n > 30$ si ha:

$$Z \sim N\,(0,1)$$

Ossia «Z» si distribuisce secondo una normale standardizzata (Cochran, 1977).

Nelle precedenti formule θ rappresenta il generico parametro oggetto di studio, t rappresenta la sua stima campionaria ottenuta utilizzando, nella funzione t [Y(s)], i dati del campione $\{s\}$ e σ^{-2} rappresenta una stima della varianza di t.

Sulla base di quanto detto è possibile fare inferenza sul parametro oggetto di indagine utilizzando la teoria dell'inferenza statistica per popolazioni finite; bisogna aggiungere che è sempre possibile ottenere delle stime di N, attraverso opportune tecniche.

Supponiamo ora di conoscere la numerosità di U e di avere a disposizione una lista L di etichette per U; quindi:

$$A = \{U, N, n, y, L\} \qquad (5)$$

In questo caso è possibile adottare un disegno campionario

semplice in cui d (s) è conosciuta (dipende da N ed n) ed utilizzare il seguente stimatore (Yamane, 1967):

$$t\,[Y\,(s)] = \begin{cases} \dfrac{N}{n}\displaystyle\sum_{i=1}^{N} Y_i & \text{per totali o frequenze assolute} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} Y_i & \text{per medie o frequenze relative} \end{cases} \qquad (6)$$

Se, inoltre, il disegno adottato prevede la non ripetizione delle unità osservate, a parità di stimatore utilizzato, si ha una diminuzione della variabilità della stima espressa da:

$$f_c = \frac{N - n}{N} \qquad (7)$$

denominato *fattore di correzione,* della varianza degli stimatori, per popolazioni finite.

In realtà le informazioni disponibili sono solitamente più numerose di quelle indicate nella (5) per cui i disegni che vengono adottati sono generalmente più complessi: disegni stratificati, a più stadi ed a probabilità variabili. Tali disegni, in letteratura, vengono denominati appunto «disegni complessi».

Riteniamo comunque necessario descrivere il campionamento casuale semplice in quanto esso è la base della teoria del campionamento e punto di partenza per la costruzione di disegni complessi.

Per disegno casuale semplice viene intesa una distribuzione di probabilità su U_c che dipende soltanto dalla numerosità della popolazione N e del campione n.

Nell'ambito di tale disegno è possibile distinguere due tipologie di campionamento:

— disegno casuale semplice con ripetizione delle unità;
— disegno casuale semplice senza ripetizione delle unità.

Nel primo caso si ha:

$$d\,(s) = \begin{cases} \dfrac{1}{N^n} & \text{per } n\,(s) = n;\ Y\,(s) \in Y\,(L) \\ 0 & \text{altrove} \end{cases} \qquad (8)$$

ricordando che Y(L) indica l'insieme delle modalità di y assunta nella popolazione U:

$$Y(L) = \{ Y_1, \ldots\ldots\ldots\ldots, Y_N \} \qquad (9)$$

con n (s) una funzione delle etichette campionate che fornisce la numerosità campionaria n, e con Y (s) i dati campionari:

$$Y\,(s) = (Y_1\ , \ldots\ldots\ldots,\ Y_n) \qquad (10)$$

La probabilità di estrarre una generica unità u_i etichettata dal numero i, (i = 1, 2, ..., N) – ricordando che da un punto di vista *frequentista,* la probabilità di un evento è data dal numero dei casi favorevoli al verificarsi dell'evento diviso il numero dei casi possibili (Hansen, Hurwitz e Madow, 1953) – è data da:

$$P\,\{i\} = \frac{1}{N} \qquad (11)$$

La probabilità di estrarre una unità u_j (etichettata con il numero j, $j \neq i$) condizionatamente al fatto che è stata estratta, nella precedente selezione, l'unità u_i, è data da:

$$P\,\{j \mid i\} = \frac{1}{N} \qquad (12)$$

questo perché, all'unità estratta precedentemente, viene data la possibilità di essere estratta nuovamente, e quindi il numero di unità di U non varia al variare dell'estrazione campionaria.

La probabilità dell'evento *uscita dell'unità contrassegnata dall'etichetta j alla seconda estrazione* rimane quindi immutata.

Da quanto detto si ha:

$$P\,\{i, j\} = P\,\{i\}\,P\,\{j\} \qquad (13)$$

poiché i due eventi sono indipendenti; generalizzando al caso di n unità, si ha:

$$P\,\{i_1, \ldots\ldots i_n\} = \frac{1}{N^n} \qquad (14)$$

che è appunto la (8).

Nel campionamento senza reimmissione il disegno è dato da (Cassel, Särndal e Wretman, 1977):

$$d(s) = \begin{cases} \dfrac{n!\,(N-n)!}{N!} & \text{per } n(s) = n,\ Y(s) \in Y(L) \\ 0 & \text{altrove} \end{cases} \qquad (15)$$

In questo caso la probabilità di estrarre l'unità u_i alla prima estrazione è ancora:

$$P\{i\} = \frac{1}{N} \qquad (16)$$

mentre la probabilità di estrarre l'unità j condizionatamente al fatto di avere estratto l'unità i nella precedente estrazione è data da:

$$P\{j \mid i\} = P\{j, i\}/P\{i\} = \frac{1}{N-1} \qquad (17)$$

La probabilità di estrarre l'unità u_j condizionatamente al fatto di avere estratta l'unità u_i non è costante, poiché i due eventi non sono indipendenti. In altri termini, se si vuole effettuare un campionamento aleatorio semplice mediante n estrazioni (prove) indipendenti, allora bisogna utilizzare il disegno di campionamento con ripetizione delle unità di U (nel senso che una stessa unità può essere rilevata più di una volta); altrimenti, si utilizzerà il campionamento senza ripetizione delle unità, dove però le estrazioni risulteranno dipendenti. Il grado di dipendenza è inversamente proporzionale ad N ed è pari a:

$$-\frac{1}{N-1} \qquad \text{per ogni coppia } (i, j) \in \{L\} \qquad (18)$$

La (18) è significativa per N piccolo, mentre tende a zero per N tendente ad infinito. Calcoliamo ora la probabilità di estrarre un campione formato dalle unità u_i, u_j, ed u_h:

$s = \{i, j, h\}$; $(i, j, h) \in \{L\}$,

$n(s) = 3$

Indipendentemente dall'ordine di estrazione si ottiene:

$$P'\{i, j, h\} = P\{i, j, h\} + P\{j, i, h\} + P\{i, h, j\} + P\{h, i, j\} + P\{j, h, i\} + P\{h, j, i\} \qquad (19)$$

Poiché i 3! campioni sono indipendenti ed incompatibili, la (19) deriva dall'applicazione del *teorema delle probabilità totali*.

Sostituendo nella (19) le espressioni della probabilità congiunta delle unità u_i, u_j, u_h che per ciascuno degli addendi è pari a:

$$\frac{1}{N(N-1)(N-2)} \qquad (20)$$

e considerando una popolazione di numerosità N ed n(s) = n, segue immediatamente che:

$$d(s) = \frac{n!}{N(N-1)(N-2)\ldots\ldots(N-n+1)} \qquad (21)$$

che è possibile riscrivere come:

$$d(s) = \left(\frac{N(N-1)(N-2)\ldots\ldots(N-n+1)}{n!}\right)^{-1} \qquad (22)$$

La frazione della (22) con semplici passaggi algebrici diventa:

$$\frac{N(N-1)(N-2)\ldots\ldots(N-n+1)}{n!} = \frac{N!}{n!\,(N-n)!} \qquad (23)$$

Il secondo membro della (23), che esprime il numero di combinazioni di N elementi di classe n, deriva dal fatto che ai fini del disegno viene trascurato l'ordine di estrazione delle unità.

Invertendo si ha quindi:

$$d(s) = \frac{n!\,(N-n)!}{N!}$$

che è esattamente la (15).

Avendo illustrato gli aspetti fondamentali del disegno di campionamento casuale semplice, passiamo ora a descrivere

alcuni disegni relativi a campioni complessi di uso più frequente nelle indagini reali.

Il campionamento stratificato

Il fatto di essere in possesso di informazioni a priori, in una indagine statistica, consente l'utilizzo delle stesse al fine di ottenere disegni più efficienti. È evidente che avere informazioni su U e non utilizzarle (o peggio utilizzarle male) è equivalente alla situazione di assenza di informazione. È quindi molto importante, durante la progettazione di una indagine campionaria, ricercare e controllare le informazioni disponibili e finalizzarle alla costruzione, sia del disegno di campionamento che degli stimatori dei parametri di interesse. Le informazioni a priori infatti possono essere utilizzate nella costruzione di d(s), nella costruzione di t[Y(s)] o in entrambe le fasi, come meglio vedremo nel prosieguo. Supponiamo che l'insieme di informazioni sia:

$$A = \{ N, n, Y(L), X(L), L \} \qquad (24)$$

in cui:

$$X(L) = \{ X_1, \ldots\ldots, X_N \} \qquad (25)$$

In altri termini, sono conosciuti, per tutti gli elementi di U, i valori assunti da una variabile che pur non essendo una caratteristica oggetto di indagine, risulta, in qualche modo, ad essa correlata.

In questo caso un disegno che consente di migliorare l'efficienza di uno stimatore è il disegno campionario stratificato. L'argomento della stratificazione verrà ripreso in seguito con maggiori dettagli; in questo capitolo verranno esplicitati soltanto alcuni concetti riguardanti il disegno di campionamento.

Viene definito con il termine *stratificazione* un processo di suddivisione delle unità di U in gruppi omogenei (strati).

Per variabile di stratificazione viene invece intesa quella caratteristica (o insieme di caratteristiche), non oggetto di indagine, in base alle quali classificare le unità di U in gruppi omogenei rispetto alle variabili oggetto d'indagine (Castellano ed Herzel, 1981).

È da aggiungere che il processo di stratificazione non incide sulla correttezza degli stimatori, bensì sulla loro efficienza. Questo significa che lo stimatore rimane corretto anche con una stratificazione *sbagliata,* risultando però meno efficiente. Per quanto riguarda tali concetti, si rimanda alla illustrazione delle proprietà degli stimatori, inserita nel Capitolo 3.

Il primo problema da affrontare, in un procedimento di stratificazione, è la scelta della (o delle) variabile di stratificazione.

Poiché l'effetto è tanto maggiore quanto più è forte il legame tra variabile oggetto di indagine e variabile di stratificazione, il criterio guida, nella scelta di quest'ultima deve essere basato su tale legame. La variabile di stratificazione ottimale sarebbe allora, quella per cui la correlazione, con la variabile oggetto di indagine, risulta massima (prossima all'unità). Questa è tuttavia una situazione ipotetica in quanto essa si verificherebbe soltanto in due casi poco realistici:

— quando esiste un legame deterministico tra la variabile oggetto di indagine e quella di stratificazione;
— quando la variabile di stratificazione coincide con quella oggetto di indagine.

Nelle indagini reali, la variabile di stratificazione è scelta sulla base di risultati, sulla variabile oggetto di studio, ottenuti in indagini effettuate in tempi precedenti, oppure sulla base di opinioni personali del ricercatore sul loro grado di dipendenza.

Supponiamo ora di avere una popolazione di N unità e di stratificarla, sulla base della variabile definita dalla (25), in H strati di numerosità $N_1, N_2, \ldots, N_H$.

Si hanno in questo modo H sub-popolazioni, ciascuna più omogenea rispetto alla popolazione originaria ed in ognuna delle quali può essere selezionato un campione di n_h unità mediante un disegno casuale semplice.

Ricordando la formula della probabilità congiunta delle n unità del campione nel disegno campionario semplice si ha, per ogni strato h:

$$P(s_h) = d(s_h) = \frac{n_h!\,(N_h - n_h)!}{N_h!} \qquad (26)$$

Poiché gli H campioni sono indipendenti, ricordando che la probabilità congiunta di H eventi indipendenti è data dal prodotto delle probabilità dei singoli eventi, si ha:

$$_{g}d(s) = \prod_{h=1}^{H} \frac{n_h!\,(N_n - n_h)!}{N_h!} \qquad (27)$$

È da aggiungere che il numero di campioni (nell'universo dei campioni) con probabilità maggiore di zero in un disegno campionario semplice è più grande rispetto a quello di un disegno stratificato.

È molto importante sottolineare che l'efficienza della stratificazione è direttamente funzione del legame esistente tra variabile di stratificazione e variabile oggetto di indagine. Se la variabile oggetto d'indagine è multidimensionale bisognerebbe disporre di una variabile di stratificazione significativamente legata a ciascuna delle variabili oggetto di studio. Questa è una situazione abbastanza rara nella realtà in quanto nelle situazioni concrete accade spesso che solo alcune delle variabili oggetto d'indagine risultano correlate con quella di stratificazione e quindi solo per queste il disegno risulterà efficiente. È opportuno, inoltre, sottolineare che l'eventuale introduzione di ulteriori variabili di stratificazione, significativamente legate con le rimanenti variabili oggetto di indagine, pur aumentando l'efficienza delle stime relative a queste ultime può produrre una diminuzione di efficienza delle stime riferite al gruppo di variabili correlate con la prima variabile di stratificazione.

Il campionamento a due stadi

Supponiamo ora che le unità che compongono la popolazione oggetto di studio siano raggruppate secondo un criterio che non dipende da chi programma l'indagine. I gruppi siano preesistenti all'indagine stessa, completamente individuati (o individuabili) e che, per ciascun gruppo, esista la lista delle unità che ne fanno parte.

Supponiamo ancora che il disegno di campionamento preveda la selezione di un campione di unità soltanto in alcuni dei gruppi in cui la popolazione oggetto d'indagine risulta suddivisa. Questo tipo di disegno viene denominato *disegno a due stadi* (Fabbris, 1989).

Ciascun raggruppamento di unità costituisce un'unità primaria o di primo stadio; le unità *finali* di campionamento costituiscono le unità secondarie o di secondo stadio.

Facciamo notare che nel disegno a due stadi i gruppi di unità sono preesistenti all'indagine, mentre nel disegno stratificato sono costruiti ad hoc; nel primo caso, inoltre, il campionamento viene effettuato solo in alcuni gruppi, mentre nel secondo riguarda tutti i gruppi.

In un campione a tre stadi, invece, le unità di secondo stadio non coincidono con le unità finali, ma bisogna effettuare, per ogni unità di secondo stadio, un ulteriore campionamento per avere le unità finali; queste vengono chiamate anche unità di terzo stadio. In questa sede verrano trattati unicamente i disegni a due stadi, poiché sono quelli più utilizzati nelle indagini concrete (Istat, 1978).

Un esempio di popolazione clusterizzata è l'insieme delle famiglie italiane. Infatti, si possono avere come unità di primo stadio i comuni (o addirittura le province o le regioni) e come unità di secondo stadio le famiglie o gli individui.

È da osservare che le unità di primo stadio sono delle unità fittizie, create raggruppando le unità finali attraverso criteri di vicinanza territoriale, amministrativi, ecc. L'esigenza di adottare disegni a più stadi non è di natura statistica, mentre lo sono i riflessi del disegno sugli stimatori.

Infatti i motivi che giustificano il campionamento a due stadi possono essere: economici, logistici oppure legati alla disponibilità o meno di liste.

I motivi di natura economico-logistici sorgono quando la popolazione è distribuita su un territorio molto vasto ed adottare un campione non stadificato implicherebbe il coinvolgimento di molte unità amministrative (provincie, comuni, ecc.) ed un gran numero di rilevatori i quali dovrebbero raggiungere posti molto lontani tra loro per intervistare in alcuni casi pochissime unità finali. Come risulta evidente, tutto ciò implicherebbe costi elevati ed un enorme lavoro di organizzazione per lo svolgimento dell'indagine.

L'altro caso riguarda situazioni in cui non esistono liste per tutte le unità di secondo stadio, oppure siano difficilmente gestibili, ma sia conosciuto invece l'ammontare delle unità di secondo stadio per tutte le unità di primo stadio.

Il caso più semplice di disegno a due stadi è quello in cui le unità, sia di primo che di secondo stadio, vengono estratte attraverso un disegno campionario semplice. Indichiamo con:

N	il numero di unità di primo stadio
M_i	il numero di unità di secondo stadio nell'unità di primo stadio i (i=1,2,.....,N)
n	il numero di unità di primo stadio campionate
m_i	il numero di unità di secondo stadio campionate nell'unità di primo stadio i

Essendo le unità di primo stadio un raggruppamento di unità di secondo stadio, la probabilità di selezionare un insieme $\{s_i\}$ di unità di secondo stadio dipende dalla probabilità dell'evento «selezione di un insieme $\{s\}$ di unità di primo stadio che contenga l'unità primaria i». In altri termini, l'evento «selezione dell'unità primaria i» condiziona l'evento «selezione dell'insieme $\{s_i\}$».

Supponiamo di voler condurre un'indagine sulle famiglie mediante un disegno a due stadi, in cui le unità primarie siano costituite dai comuni e quelle secondarie dalle famiglie.

La probabilità di selezionare l'insieme delle $\{m_i\}$ famiglie, avendo già scelto l'unità primaria i, (i = 1, ..., n) è:

$$P\{m_i|\, i\} = \frac{m_i!\,(M_i - m_i)!}{M_i!} \qquad (28)$$

Quindi, la probabilità di aver l'insieme $\{m_1, ..., m_i, ..., m_n\}$, condizionatamente all'insieme costituito dagli n comuni estratti, è:

$$P\{m_1, ..., m_i, ..., m_n|\, 1..., i..., n\} = \prod_{i=1}^{n} \frac{m_i!\,(M_i - m_i)!}{M_i!} \qquad (29)$$

in quanto le probabilità $P\{m_i|\, i\}$ sono condizionatamente indipendenti.

Poiché la probabilità di selezionare l'insieme delle n unità primarie è data da:

$$P\{1,...., i,..., n\} = \frac{n!\,(N - n)!}{N!} \qquad (30)$$

In conclusione, la probabilità di selezionare l'insieme $\{m_1, ..., m_i, ..., m_n\}$ al variare dell'insieme delle n unità primarie è espressa da:

$$P\{m_1, ..., m_n\} = P\{1, ..., n\}\, P\{m_1, ..., m_n|\, 1, ..., n\} =$$

$$= \frac{n!\,(N - n)!}{N!} \prod_{i=1}^{n} \frac{m_i!\,(M_i - m_i)!}{M_i!} \qquad (31)$$

La (31), pertanto, definisce il disegno a due stadi (che denotiamo con il simbolo d" (s)) in cui sia le unità primarie che quelle secondarie sono selezionate senza reimmissione e probabilità uguali.

È possibile infine dimostrare che:

$${}_gd''(s) < d(s) < {}_gd(s) \qquad (32)$$

Il campionamento a due stadi con stratificazione

Il campionamento a due stadi come abbiamo precedentemente illustrato si può intendere come un insieme di N subpopolazioni, delle quali ne vengono campionate n e, su ciascuna di queste viene effettuato un campione di m_i unità (i=1,..,n).

È evidente che, per ciascun campione (di unità primarie o secondarie) può essere adottato un disegno campionario semplice o stratificato. Il caso appena trattato riguarda l'adozione, sia per le unità di primo stadio che di secondo stadio, di un disegno campionario semplice; qui di seguito tratteremo del disegno di un campionamento a due stadi con stratificazione delle unità primarie.

La stratificazione è utilizzabile, nel disegno a due stadi, nelle seguenti modalità:

- — stratificazione solo delle unità primarie;
- — stratificazione solo delle unità secondarie nell'ambito delle unità primarie campionate;
- — stratificazione sia delle unità primarie che di quelle secondarie.

Nel primo caso le informazioni a priori, necessarie per attuare il processo di stratificazione, possono essere simbolicamente espresse da:

$$X = [X_1,...,X_i,...,X_N] \qquad (33)$$

dove X_i indica l'ammontare del carattere di stratificazione X relativo all'unità primaria i (i = 1,..., N).

In questa situazione il disegno diventa:

$${}_gd''(s) = \prod_{h=1}^{H} \left[\frac{(N_h - n_h)!\, n_h!}{N_h!} \prod_{i=1}^{n_h} \frac{(M_{hi} - m_{hi})!\, m_{hi}!}{M_{hi}!} \right] \qquad (34)$$

Nel secondo caso devono essere disponibili informazioni a livello disaggregato; ossia, per ogni unità elementare, si dovrà disporre di un insieme di informazioni del tipo:

$$X = \left[X_{11},...., X_{1M_1},...., X_{N1}.., X_{NM_N}\right] \qquad (35)$$

Di tali informazioni, verranno utilizzate soltanto quelle relative alle unità primarie campionate. L'espressione della distribuzione di probabilità nell'universo dei campioni, in questo caso, risulta abbastanza complicata; pertanto non si ritiene utile fornirla. È da aggiungere, comunque, che i motivi che possono indurre lo statistico ad adottare un disegno del genere sono essenzialmente due: l'elevata variabilità del carattere oggetto d'indagine all'in-

terno delle unità primarie; facilità di reperire e gestire le informazioni di cui alla (35).

Il terzo caso, infine, consiste nel suddividere le N unità primarie in H gruppi (strati) e, nell'ambito di ogni unità primaria campione in ciascun gruppo, nella suddivisione delle unità di secondo stadio in K_i strati. In un tale disegno le variabili di stratificazione utilizzate per le unità di primo stadio, possono essere diverse da quelle utilizzate per le unità di secondo stadio.

In conclusione, è da sottolineare tuttavia che laddove il disegno a due stadi non sia necessario, esso è da evitare in quanto le stime che fornisce hanno una attendibilità inferiore a quella di qualsiasi altro disegno (a parità di numerosità campionaria in termini di unità finali).

Questo fatto è dovuto alla omogeneità delle unità di secondo stadio appartenenti alla stessa unità primaria. In altri termini, le unità che appartengono alla stessa unità primaria tendono a presentare simili modalità della variabile oggetto di indagine. Una misura di questo fenomeno è fornita dal coefficiente di correlazione intraclasse, ed indicato con il simbolo rho. La relazione che lega la varianza di uno stimatore derivante da un campione a due stadi $V_c[t(s)]$ e quella di un campione casuale semplice V[t(s)] è data dalla seguente espressione approssimata (Kish, 1965):

$$V_c[t(s)] = V\,[t(s)]\,\{\,1 + (\overline{m}-1)\ rho\,\} \qquad (36)$$

dove $\overline{m}$ rappresenta il numero medio di unità di secondo stadio campionate nelle unità di primo stadio.

Come si evince dalla formula precedente, il campionamento a due stadi è «migliore» di quello semplice soltanto se *rho* è negativo. Se questo si verifica, significa, che le unità appartenenti ad unità primarie diverse tendono ad avere le stesse modalità del carattere oggetto di indagine, mentre quelle appartenenti alla stessa unità primaria tendono ad essere «diverse» (disomogenee). Le ricerche condotte su questo terreno hanno mostrato che questa situazione è molto rara.

In generale, il disegno a due stadi è conveniente quando la rilevazione dei dati viene effettuata tramite intervista *diretta* (con intervistatore sul posto). Quando invece l'intervista viene effettuata con il metodo telefonico o postale, in genere, è *conveniente* il disegno ad uno stadio.

CAPITOLO 3 - CONCETTI BASE DELLA TEORIA DELLA STIMA

Introduzione

Le quantità che, solitamente, interessano in una indagine statistica sono particolari funzioni delle osservazioni (misurazioni) sulle unità di U, relative alle variabili oggetto di indagine. Per comodità del lettore ne riportiamo brevemente le formule:

— totale o frequenza assoluta

$$Y = \sum_{i=1}^{N} Y_i \qquad (1)$$

— media o frequenza relativa

$$\overline{Y} = \sum_{i=1}^{N} \frac{Y_i}{N} \qquad (2)$$

— rapporto tra medie o totali

$$R = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} \qquad (3)$$

Nel prosieguo, ogni volta che verrà trattato genericamente del parametro oggetto di stima, esso verrà indicato con θ; quando invece ci si riferirà specificatamente ai parametri sopra descritti, essi verranno indicati rispettivamente con Y, $\overline{Y}$ ed R.

L'obiettivo dell'indagine campionaria è quello di fornire delle informazioni sui parametri (1), (2) e (3) attraverso i dati campio-

nari Y(s). Questo viene fatto utilizzando delle funzioni t[Y(s)] denominate *stimatori*. La funzione t[(·)] è quasi sempre del tipo:

$$\sum_{i=1}^{n} K_i Y_i \qquad (4)$$

Il valore fornito dalla (4), sostituendo in essa i dati campionari, viene chiamato *stima* del parametro oggetto di indagine. La quantità K_i rappresenta il *peso* associato all'unità i. Il significato di tale quantità è quello di indicare quante unità della popolazione U sono rappresentate da ciascuna unità campione.

Metodi di stima che non utilizzano informazioni ausiliarie: gli stimatori diretti

Il numero di unità non incluse nel campione, «rappresentate» da ciascuna unità campionata attraverso il peso associato K_i, dipende dalla probabilità che, dato il disegno d(s), l'unità i ha di essere inclusa nel campione s. Tale peso infatti risulta essere l'inverso delle probabilità di inclusione dell'unità i di U in s.

Riprendiamo la formula del disegno casuale semplice senza ripetizione, relativa ad un campione di ampiezza n, selezionato da una popolazione di N unità, trattata nel capitolo 2:

$$d(s) = \frac{n!\,(N-n)!}{N!} \qquad (5)$$

Supponiamo di voler ottenere una stima di:

$$Y = \sum_{i=1}^{N} Y_i \qquad (6)$$

Inoltre, le informazioni disponibili siano:

$$A = \{N, n, y, L\} \qquad (7)$$

ossia non si dispone di alcuna informazione ausiliaria sulle unità di U.

È possibile dimostrare che in questo caso, l'unico stimatore lineare omogeneo, che ha la proprietà di essere ottimale nel senso dei minimi quadrati, della (6) è dato da:

$$\hat{Y} = \sum_{i=1}^{n} K_i Y_i \qquad (8)$$

dove:

$$K_i = \frac{N}{n} \qquad (9)$$

È quindi possibile affermare che ogni unità campionata rappresenta K_i unità di U; vediamo ora come K_i discende da d (s).

La probabilità di inclusione relativa all'unità i è data dalla somma di d(s) nell'insieme dei campioni che, nell'universo U_c, contengono l'unità i; indichiamo tale insieme con il simbolo $\{\tilde{s}\}$. Il numero di campioni contenuti in $\{\tilde{s}\}$ viene indicato con N_s e la probabilità di inclusione relativa alla generica unità i, con Π_i; in formula si ha:

$$\Pi_i = \sum_{j=1}^{N_s} d\,(s_{ij}) \qquad (10)$$

in cui s_{ij} rappresenta il generico campione j di U_c contenente l'unità i, e dove:

$$\tilde{s} = \{s : i \in s\}$$

Il numero di campioni dell'universo U_c che appartengono all'insieme $\{\tilde{s}\}$ viene calcolato associando all'unità i tutti i campioni di ampiezza n - 1 che è possibile formare dopo aver tolto, dalla popolazione U, l'unità i. In formule:

$$N_s = \frac{(N-1)!}{(N-n)!\,(n-1)!} \qquad (11)$$

$$\Pi_i = \sum_{j=1}^{N_s} d\,(s_{ij}) = \sum_{j=1}^{N_s} d\,(s) \qquad (12)$$

Poiché d(s) è costante in U_c la (12) diventa:

$$\Pi_i = N_s\, d\,(s) = N_s \frac{(N - n)\,!\; n!}{N!} \qquad (13)$$

Sostituendo nella (13) ad N_s la sua espressione data dalla (11) si ha:

$$\Pi_i = \frac{(N - 1)\,!}{(N - n)!\,(n - 1)!} \cdot \frac{(N - n)!\; n!}{N!} = \frac{n}{N} \qquad (14)$$

Come la (14) mostra, la probabilità di inclusione dell'unità contrassegnata dall'etichetta i non dipende dal valore dell'etichetta stessa.

Questo significa che le unità della popolazione, se compariranno nel campione, rappresenteranno ciascuna lo stesso numero di unità di U. In altri termini, il peso K_i è costante per tutte le unità della popolazione (si può anche dire che le unità sono *scambiabili*) ossia per quanto riguarda la stima di θ, ciascuna unità di U vale l'altra.

D'altra parte, la conclusione era già nelle premesse, poiché quando nel disegno non viene utilizzata (non si dispone di) alcuna informazione relativa alle unità di U, il disegno è definito «non informativo». Dalla teoria dell'inferenza statistica, le situazioni non informative vengono descritte (formalizzate) attraverso distribuzioni di probabilità uniformi.

Vediamo ora che, seguendo un ragionamento leggermente differente da quello utilizzato finora, l'unico stimatore di θ che è possibile utilizzare se lo statistico si trova a predisporre un piano di campionamento nella situazione descritta dalla (7), è dato dalla (8): tale stimatore infatti risulta «ottimale» nel senso dei minimi quadrati (Fuller, 1975).

Come già accennato nel capitolo 2, l'insieme dei valori esistenti nella popolazione viene indicato con:

$$Y\,(L) = (Y_1, \ldots\ldots, Y_N)$$

mentre quelli del campione con:

$$Y\,(s) = (Y_1, \ldots\ldots, Y_n)$$

Supponiamo che il parametro oggetto di stima sia il totale Y, definito dalla (1); avendo inoltre indicato con $\overline{Y}$ la media del

carattere y espressa dalla (2), per ciascuna unità i si può scrivere:

$$Y_i - \overline{Y} = \varepsilon_i \qquad (i = 1, \ldots, N) \qquad (15)$$

risultando, come è noto

$$\sum_{i=1}^{N} \varepsilon_i = 0 \qquad (16)$$

Ogni elemento di Y (L) può allora essere scritto come:

$$Y_i = \overline{Y} + \varepsilon_i \qquad (i = 1, \ldots, N) \qquad (17)$$

Utilizzando la notazione matriciale

$$\mathbf{Y} = \overline{\mathbf{Y}}\,\mathbf{1} + \boldsymbol{\varepsilon} \qquad (18)$$

dove $\mathbf{1}' = (1, \ldots\ldots, 1)$ è un vettore di N elementi.

Una stima di $\overline{Y}$, attraverso il metodo dei minimi quadrati in presenza di osservazioni dipendenti, viene fornita da (Kendall e Stuart, 1977):

$$\hat{\overline{Y}} = (\mathbf{1}'\,\mathbf{D}^{-1}\,\mathbf{1})^{-1}\,(\mathbf{1}'\,\mathbf{D}^{-1}\,\mathbf{Y}\,(\mathbf{s})) \qquad (19)$$

dove $\mathbf{1}' = (1, \ldots, 1)$ è un vettore di n elementi.

La dipendenza tra le oservazioni deriva dal fatto che le unità di U, una volta selezionate, non vengono più considerate per selezioni successive (selezione senza ripetizione). Tale metodo di selezione, nella pratica del campionamento, è quello maggiormente adottato. Come risulta dalla relazione (18) del capitolo 2, la dipendenza tra coppie di osservazioni sulle unità di U è costante, per cui si ha:

$$V\,(\varepsilon) = V\,[y] \begin{bmatrix} 1\ \rho \ldots\ldots\ldots \rho \\ \rho\ 1 \ldots\ldots\ldots \rho \\ \ldots\ldots\ldots \\ \ldots\ldots\ldots \\ \rho\ \rho \ldots\ldots\ldots 1 \end{bmatrix}^{-1} = V\,[y]\,\mathbf{D}^{-1} \qquad (20)$$

in cui:

$$V[y] = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \overline{Y})^2 \qquad (21)$$

Si può dimostrare, con semplici passaggi ed invertendo la matrice **D**, che:

$$\hat{\overline{Y}} = \frac{\{1 + (n - 1)\, V[y]\}}{n} \frac{1}{\{1 + (n - 1)\, V[y]\}} \mathbf{1}' \mathbf{Y}(s) =$$

$$= \frac{1}{n} \sum_{i=1}^{n} Y_i \qquad (22)$$

Tenendo presente che:

$$s = \{1,2 \ldots, n\}$$

rappresenta l'insieme delle unità campionate, indichiamo ora con:

$$\bar{s} = \{n + 1, n + 2, \ldots, N\} \qquad (23)$$

l'insieme delle unità di U non campionate.

Per una stima del totale del carattere nella popolazione, associamo a ciascuna unità di $\bar{s}$ la media campionaria ottenuta tramite la (22); si ha:

$$\hat{Y} = \sum_{i=1}^{n} Y_i + (N - n)\, \hat{\overline{Y}} =$$

$$= \sum_{i=1}^{n} Y_i + (N - n) \sum_{i=1}^{n} \frac{Y_i}{n} =$$

$$= \frac{N}{n} \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} K_i\, Y_i \qquad (24)$$

Poichè gli n valori osservati non hanno errore campionario ed $\hat{\overline{Y}}$ è ottimale nel senso dei minimi quadrati, lo stimatore del totale fornito dalla (24) è ottimale in tal senso.

La costante K_i rappresenta (come già osservato) il peso associato ad ogni unità di U.

È interessante ora verificare come il coefficiente di correlazione tra le osservazioni di y sia esattamente quello introdotto nel precedente capitolo 2.

Dalla teoria dei minimi quadrati si ha che:

$$V[\hat{\overline{Y}}] = V[y]\, (\mathbf{1}'\mathbf{D}^{-1}\mathbf{1})^{-1} \qquad (25)$$

e poiché:

$$(\mathbf{1}'\mathbf{D}^{-1}\mathbf{1})^{-1} = \frac{1 + (n - 1)\, \rho}{n} \qquad (26)$$

si ha:

$$V[\hat{\overline{Y}}] = \frac{V[y]}{n} \{1 + (n - 1)\, \rho\} \qquad (27)$$

Poiché nelle indagini campionarie, V[y] non è generalmente noto, una sua stima si ottiene, utilizzando i dati campionari, attraverso:

$$\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \hat{\overline{Y}})^2 \qquad (28)$$

Nel campionamento senza ripetizione delle unità si ha, quando n = N:

$$V[\hat{\overline{Y}}] = 0 \qquad (29)$$

Quindi la (27), poiché ρ non dipende da n, diventa:

$$\frac{S^2}{N} \{1 + (N - 1)\, \rho\} = 0 \qquad (30)$$

dalla quale si ricava:

$$\rho = - \frac{1}{N - 1}$$

Ovviamente nel campionamento con ripetizione si avrà $\rho = 0$ e la (19) diventerà:

$$\hat{\bar{Y}} = (\mathbf{1}'\mathbf{1})^{-1}\,[\mathbf{1}'\,\mathbf{Y}\,(s)] = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad (31)$$

e quindi, per la stima del totale del carattere si avrà:

$$\hat{Y} = N\,\hat{\bar{Y}} = \frac{N}{n}\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n}\frac{Y_i}{\Pi_i} \qquad (32)$$

in cui:

$$\Pi_i = \frac{n}{N} \qquad (33)$$

Se il parametro oggetto di stima è la media del carattere nella popolazione, ossia:

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i \qquad (34)$$

uno stimatore della (34), come è facile derivare dalla (19), è la media campionaria. Riteniamo comunque più utile far discendere lo stimatore della media del carattere y dalla (32). Questo per avere un quadro generale degli stimatori lineari dei parametri della popolazione; ossia stimatori del tipo:

$$\hat{\theta} = \sum_{i=1}^{n} K_i\,Y_i \qquad (35)$$

Da quanto detto, la (31) può essere riscritta, in termini di probabilità di inclusione Π_i; dalla (32) si ha infatti:

$$\hat{\bar{Y}} = \frac{\hat{Y}}{N} = \sum_{i=1}^{n}\frac{Y_i}{N\,\Pi_i} \qquad (36)$$

Uno stimatore della forma (36) viene chiamato di Horvitz-Thompson, i quali, per primi (1952), lo introdussero per disegni a probabilità variabili.

Nella (36) il peso associato a ciascuna unità è:

$$K_i = \frac{1}{N\,\Pi_i}$$

Poiché, nella situazione descritta dalla (7);

$$\Pi_i = \frac{n}{N} \qquad (i = 1, 2 \ldots n)$$

si ha:

$$\hat{\bar{Y}} = \sum_{i=1}^{n}\frac{Y_i}{n} \qquad (38)$$

Quindi nella situazione (7) gli stimatori della media e del totale di un carattere nella popolazione oggetto di indagine sono dati rispettivamente dalla (38) e dalla (32).

È importante sottolineare il fatto che la (38) e la (32) sono un caso particolare dello stimatore di Horvitz-Thompson, e precisamente, sono riferite ad un campione aleatorio semplice con selezione, senza ripetizione e con probabilità costante, delle unità (Horvitz e Thompson, 1952).

Lo stimatore di Horvitz-Thompson è stato introdotto, come appena detto, per disegni a probabilità variabili, ossia per disegni in cui ciascuna unità *rappresenta* un numero diverso di unità di U. Un disegno di tal genere non può essere adottato nella situazione (7) poiché necessita della conoscenza di informazioni ausiliarie per ciascuna unità di U; in altri termini bisogna trovarsi nella condizione espressa da:

$$A = \{N, n, L, Y, X\} \qquad (39)$$

In tale circostanza è comunque possibile operare sia in termini di disegno che di stimatore, e le indicazioni che è possibile dare per adottare stretegie campionarie ottimali sono legate al modo in cui viene «sfruttata» l'informazione disponibile.

Per quanto riguarda il disegno, l'informazione X può essere sfruttata essenzialmente in due modi:

— suddividendo la popolazione U in gruppi omogenei (stratificazione);
— tenendo conto della diversa rappresentatività che hanno le

unità di U relativamente alla variabile oggetto d'indagine (selezione con probabilità variabili).

Per quanto riguarda lo stimatore, l'informazione a priori può essere utilizzata ai fini della costruzione di:

— stimatori del rapporto o di regressione;
— stimatori post-stratificati.

In questa sede verrà trattato maggiormente dell'utilizzazione dell'informazione X nella fase di costruzione degli stimatori.

L'ipotesi sottesa alla metodologia della stratificazione, come è facile verificare, è quella di scambiabilità parziale delle unità di U.

Supponiamo ora che l'informazione X della (39) non consenta di raggruppare efficacemente le unità di U in strati «omogenei». Questa situazione si verifica quando, rispetto alla y, non esistono in U dei gruppi veri e propri pur esistendo un legame tra la variabile di stratificazione x e la variabile oggetto di indagine y.

È comunque sempre possibile creare una stratificazione delle unità di U sulla base di classi di valori della x, ma essa risulta in questo caso meno efficace. È quindi consigliabile utilizzare l'informazione x, se quanto detto si verifica, per selezionare le unità di U con probabilità variabili. In tal caso ogni unità ha un peso K_i diverso; in altri termini ciascuna unità campione rappresenta un numero diverso di unità di U e le formule degli stimatori diventano:

$$\hat{Y} = \sum_{i=1}^{n} \frac{Y_i}{\Pi_i} = \sum_{i=1}^{n} K_i Y_i \qquad (40)$$

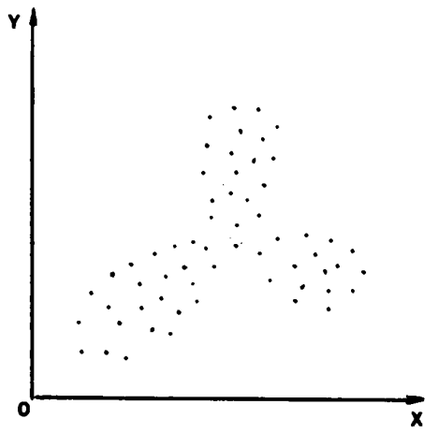
per la stima del totale del carattere y in U, e:

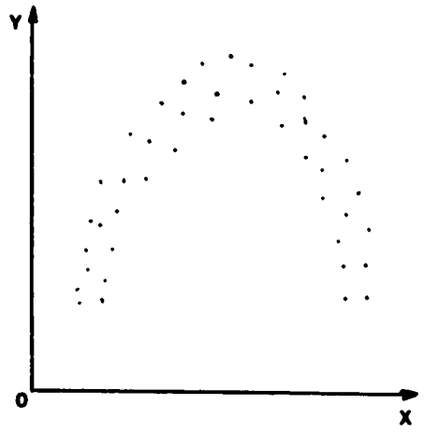
$$\hat{\bar{Y}} = \frac{1}{N} \sum_{i=1}^{n} \frac{Y_i}{\Pi_i} = \frac{1}{N} \sum_{i=1}^{n} K_i Y_i \qquad (41)$$

per la stima della media.

Per meglio chiarire quanto detto, vengono graficamente esposte le due situazioni, rappresentando le unità di U tramite punti nel piano cartesiano P (X, Y).

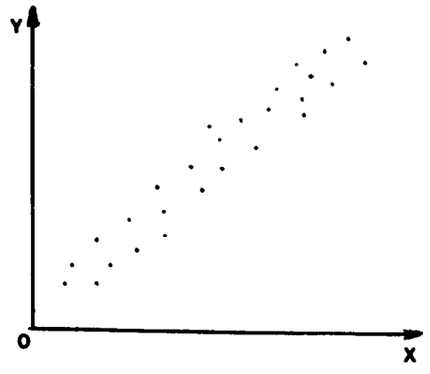
Nella prima, e cioè quando l'insieme delle unità di U assume una delle forme di seguito schematizzate:





conviene utilizzare la x per stratificare.

Se invece le unità di U assumono una configurazione del tipo



conviene adottare un disegno a probabilità variabile. La scelta di un disegno a probabilità variabile viene giustificata dal fatto che la varianza dello stimatore sia del totale Y che della media $\overline{Y}$ è una funzione omogenea della differenza tra Y_i/Π_i e Y_j/Π_j per ogni (i, j) $\in$ U con i≠j.

In altri termini:

$$V(\hat{\theta}) = \Phi\left(\frac{Y_i}{\Pi_i} - \frac{Y_j}{\Pi_j}\right) \qquad (42)$$

Riteniamo utile sottolineare che l'omogeneità della (42) comporta $\Phi(O) = O$.

Dimostriamo ora che la varianza dello stimatore $\hat{\theta}$ è pari a zero se i valori assunti dalla variabile ausiliaria x sono proporzionali ai valori assunti dalla variabile y. Siano quindi le probabilità di inclusione per le generiche unità i e j date rispettivamente da:

$$\Pi_i = \frac{n X_i}{\sum_{i=1}^{N} X_i} \qquad (43)$$

$$\Pi_j = \frac{n X_j}{\sum_{i=1}^{N} X_i} \qquad (44)$$

Per la relazione di proporzionalità tra y ed x si ha:

$$\frac{Y_i}{X_i} = \frac{Y_j}{X_j} \qquad (i, j = 1, 2, \ldots, N \qquad \text{con } i \neq j) \qquad (45)$$

Moltiplicando primo e secondo membro della (45) per:

$$\frac{\sum_{i=1}^{N} X_i}{n} \qquad (46)$$

si ottiene:

$$\frac{Y_i}{\Pi_i} = \frac{Y_j}{\Pi_j} \qquad (47)$$

da cui segue immediatamente che $V(\hat{\theta}) = O$.

Inoltre la (45) implica una relazione lineare tra x e y, in base alla quale sarebbe sufficiente campionare due sole unità di U per conoscere esattamente le modalità assunte dalla variabile oggetto di indagine y sulle rimanenti (N-2) unità di U.

Nelle indagini reali l'equazione (45) non è mai soddisfatta; può però verificarsi che al variare di i in U la variabilità di Y_i/Π_i risulti sufficientemente contenuta; in conseguenza di ciò risulta contenuta anche la varianza di $\hat{\theta}$ espressa dalla (42). Lo stimatore

$$\hat{\theta} = \sum_{i=1}^{n} K_i Y_i \quad \text{con} \quad K_i = n X_i / \sum_{i=1}^{N} X_i$$

risulta più efficiente dell'analogo stimatore, definito dalla (35), in cui K_i è costante.

Da quanto detto risulta abbastanza evidente che lo statistico, dovendo progettare un'indagine campionaria, debba studiare il tipo di relazione che lega la variabile oggetto di studio y a quella ausiliaria x.

È altrettanto evidente che per fare ciò necessita, oltre che della conoscenza teorica del fenomeno da indagare, anche di dati oggettivi sia sulla y che sulla x. Per quanto riguarda la conoscenza del fenomeno il problema viene risolto attraverso la consultazione di esperti; per quanto concerne la disponibilità di dati su y, il problema è più complesso in quanto se non sono state effettuate precedenti indagini sullo stesso fenomeno tali dati non esistono. Inoltre, pur esistendo dati di indagini effettuate su y in tempi precedenti, questi possono essere inutilizzabili se il fenomeno è molto variabile nel tempo e la data di effettuazione dell'indagine precedente risulta remota. In questo caso lo statistico deve *fidarsi* delle affermazioni dell'esperto, senza la possibilità di verificarle sperimentalmente nella progettazione del piano di campionamento.

Riepilogando, l'informazione X può essere utilizzata efficacemente nella fase della costruzione del disegno campionario se essa influenza, in qualche modo, le modalità della caratteristica oggetto di indagine y. Accertato che tale informazione ha qual-

che influenza su y, conviene utilizzarla per stratificare le unità di U se la relazione è di tipo non lineare o se l'interpolazione lineare dell'insieme di punti (X_i, Y_i); i = 1, 2, ... N non è *soddisfacente;* mentre conviene utilizzare un disegno a probabilità variabili quando esiste un legame lineare tra x ed y con una ridotta *dispersione* dei punti.

Metodi di stima basati sull'uso di informazioni ausiliarie: gli stimatori indiretti

Vediamo ora come utilizzare l'informazione X nella fase di costruzione degli stimatori dei parametri di interesse. Supponiamo che tale parametro sia:

$$Y = \sum_{i=1}^{N} Y_i$$

Sia inoltre conosciuta, per ciascuna unità di U, la modalità assunta dalla caratteristica ausiliaria. Dopo aver selezionato il campione di n unità si disporrà delle seguenti quantità:

$$\hat{\bar{X}} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$\hat{\bar{Y}} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

$$X = \sum_{i=1}^{N} X_i$$

Ricordando che uno stimatore lineare di Y ha la forma (Cassel, Sarndal e Wretman, 1977):

$$\hat{Y} = \sum_{i=1}^{n} K_i Y_i \qquad (48)$$

con:

$$K_i = \frac{1}{\Pi_i}$$

allo stesso modo una stima di X data da:

$$\hat{X} = \sum_{i=1}^{n} K_i X_i \qquad (49)$$

Come si nota il coefficiente della combinazione lineare (49) è identico al coefficiente della (48). Questo perché il disegno campionario è identico in quanto gli insieme di valori X (s) ed Y (s) sono rilevazioni effettuate sulle medesime unità di U, ed inoltre K_i dipende esclusivamente da d (s) e non dalle caratteristiche che vengono rilevate su $\{s\}$.

Poiché, però, è conosciuto il parametro X (in generale diverso da $\hat{X}$) bisognerà cercare un coefficiente di correzione della (49) affinché si verifichi $X = \hat{X}$; indichiamo con λ tale coefficiente. Risolvendo in λ l'equazione:

$$\sum_{i=1}^{N} X_i = \lambda \sum_{i=1}^{n} K_i X_i \qquad (50)$$

si avrà:

$$\lambda = \frac{\sum_{i=1}^{N} X_i}{\sum_{i=1}^{n} K_i X_i} \qquad (51)$$

Lo stimatore di Y che soddisfa il vincolo $X = \hat{X}$ è quindi dato da:

$$\hat{\bar{Y}} = \lambda \sum_{i=1}^{n} K_i Y_i =$$

$$= \frac{\sum_{i=1}^{N} X_i}{\sum_{i=1}^{n} K_i X_i} \sum_{i=1}^{n} K_i Y_i = \frac{\hat{Y}}{\hat{X}} X \qquad (52)$$

da cui discende il rapporto di proporzionalità:

$$\frac{\hat{\bar{Y}}}{\hat{Y}} = \frac{X}{\hat{X}} \qquad (53)$$

Ossia la stima vincolata $\hat{\bar{Y}}$ è nello stesso rapporto con la stima $\hat{Y}$, come il valore vero del parametro X lo è con $\hat{X}$.

È facile notare che se $K_i = N/n$ (per i = 1, 2,...., N) allora la (52) diventa:

$$\hat{\bar{Y}} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} \sum_{i=1}^{N} X_i \qquad (54)$$

Le (52) e (54) vengono chiamate *stimatore rapporto del totale del carattere* y *in* U.

Nel caso quindi di disegno a probabilità costante lo *stimatore rapporto* non dipende dalle probabilità di inclusione delle unità di U nel campione.

Se il parametro oggetto di indagine è la media del carattere y in U, sapendo che:

$$\bar{X} = \frac{X}{N} \quad \text{e} \quad \hat{\bar{Y}} = \frac{\hat{\bar{Y}}}{N} \qquad (55)$$

Dalla (52) si ha:

$$\hat{\bar{Y}} = \frac{\hat{Y}}{\hat{X}} \bar{X} \qquad (56)$$

Nella pratica del campionamento, lo stimatore rapporto ha largo uso in quanto più efficiente dello stimatore diretto anche se non gode della proprietà di correttezza, della quale verrà trattato nel prosieguo. Vi sono però delle ipotesi che, se soddisfatte, rendono lo stimatore rapporto ottimale (Royall e Cumberland, 1971). Esse sono:

$$y = \beta x + \varepsilon; \qquad (57)$$

$$\varepsilon \sim N(0, \sigma^2) \qquad (58)$$

$$V[y] = \alpha x \qquad (59)$$

In altri termini se esiste una relazione stocastica, tra il carattere y ed il carattere x, di tipo lineare ed omogenea ed inoltre se la varianza di y è proporzionale ad x, allora il modo migliore di utilizzare, in fase di stima, l'informazione X è attraverso lo stimatore rapporto.

Se infatti riprendiamo la (54) e scomponiamo la sommatoria del totale del carattere x, nella parte osservata nel campione s e la parte rimanente (non osservata) della popolazione $\bar{s}$, si ottiene:

$$\sum_{i=1}^{N} X_i = \sum_{i=1}^{n} X_i + \sum_{i=n+1}^{N} X_i \qquad (60)$$

in base alla quale la (54) si può riscrivere:

$$\hat{\bar{Y}} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} \left[\sum_{i=1}^{n} X_i + \sum_{i=n+1}^{N} X_i \right] = \qquad (61)$$

$$= \sum_{i=1}^{n} Y_i + \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} \sum_{i=n+1}^{N} X_i$$

Ponendo:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} \qquad (62)$$

si ha:

$$\hat{\bar{Y}} = \sum_{i=1}^{n} Y_i + \hat{\beta} \sum_{i=n+1}^{n} X_i \qquad (63)$$

La (63) significa che il totale del carattere y nella popolazione U, viene stimato utilizzando i dati osservati nel campione e la relazione (57) per le unità non campionate; l'attenzione quindi si sposta sulla stima del parametro β. Utilizzando i minimi quadrati ponderati si ha:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i (\alpha X_i)^{-1} Y_i}{\sum_{i=1}^{n} X_i^2 (\alpha X_i)^{-1}} \qquad (64)$$

dalla quale si ricava la (62), che introdotta nella (63) consente il calcolo di $\hat{\bar{Y}}$.

Poiché $\hat{\beta}$ è ottenuta mediante il metodo dei minimi quadrati, la $\hat{\bar{Y}}$ è ottimale nel senso che ha varianza minima nella classe degli stimatori lineari ed omogenei.

Lo statistico che deve progettare il piano di campionamento, avendo a disposizione risultati di indagini precedenti od altre informazioni che lo inducono a ritenere valide le ipotesi (57) e (59) può applicare lo stimatore rapporto con la quasi certezza di aver utilizzato uno stimatore ottimale nel senso appena detto.

Supponiamo ora che esista (nelle indagini Istat di solito esiste) una variabile z della quale non è stato possibile tenere conto in fase di disegno. Supponiamo ancora che z assuma un numero limitato di modalità $Z_1, \ldots, Z_a \ldots, Z_A$, denominate post-strati. È evidente che lo statistico debba ritenere che la variabile oggetto di indagine y e la variabile z siano tra loro, in qualche modo, legate (altrimenti non avrebbe senso l'utilizzazione di quest'ultima nel piano di campionamento). Dopo aver selezionato il campione si avranno n_a unità che prendono le modalità Z_a ($a = 1, \ldots A$), con:

$$\sum_{a=1}^{A} n_a = n$$

Indichiamo inoltre con N_a il numero di unità (noto) che assumono la generica modalità Z_a nella popolazione, con:

$$\sum_{a=1}^{A} N_a = N$$

Il risultato campionario:

$$Y(s) = (Y_1, \ldots, Y_n)$$

può essere suddiviso in:

$$\{ Y(s_1), Y(s_2), \ldots\ldots, Y(s_A) \}$$

dove s_1 è l'insieme delle n_1 unità campionate che presentano il carattere $z = Z_1$, s_2 l'insieme delle n_2 unità con $z = Z_2$, ecc. Il campione s risulta così suddiviso in modo aleatorio in quanto non è stato programmato a priori quante unità con la caratteristica $z = Z_a$ ($a = 1, \ldots, A$) dovevano far parte del sub-campione s_a tra le A sub-popolazioni di numerosità $N_1, N_2, \ldots, N_A$. Il parametro oggetto di stima sarà, nel caso del totale:

$$Y = \sum_{i=1}^{N} Y_i = \sum_{a=1}^{A} \sum_{i=1}^{N_a} {}_aY_i = \sum_{a=1}^{A} {}_aY \qquad (65)$$

in cui ${}_aY_i$ indica il valore osservato sull'unità i del post-strato a.

Nel caso della media del carattere y si ha:

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i = \frac{1}{N} \sum_{a=1}^{A} \sum_{i=1}^{N_a} {}_aY_i = \frac{1}{N} \sum_{a=1}^{A} {}_aY \qquad (66)$$

Come risulta dalle (65) e (66), i parametri oggetto di studio possono essere ottenuti come somma dei parametri dei post-strati.

È possibile quindi procedere separatamente alla stima dei parametri dei post-strati stessi ed ottenere la stima del parametro di interesse come somma delle stime (Holt e Smith, 1979).

In questo caso la (40) diventa:

$${}_p\hat{Y} = \sum_{a=1}^{A} \left[\frac{N_a}{n_a} \sum_{i=1}^{n_a} {}_aY_i \right] \qquad (67)$$

Poiché, dopo aver fissato i post-strati, n_a è una variabile casuale le quantità nelle parentesi della (57) si possono scrivere:

$$\frac{N_a}{n_a} \sum_{i=1}^{n_a} {}_aY_i \qquad (a = 1, \ldots, A) \qquad (68)$$

Come è possibile osservare la (68) è uno stimatore rapporto del totale del carattere y in U_a.

La (67) diventa:

$${}_p\hat{Y} = \sum_{a=1}^{A} \frac{N_a}{n_a} \sum_{i=1}^{n_a} {}_aY_i \qquad (69)$$

Lo stimatore post-stratificato, condiziona la stima del parametro Y alla distribuzione delle unità di U rispetto alla variabile z.

Supponiamo ora di disporre anche dell'informazione x e di voler condizionare la stima di Y sia rispetto ad x che a z. Abbiamo già visto che il condizionamento rispetto alla variabile x conduce allo stimatore rapporto e quello rispetto alla variabile z allo stimatore post-stratificato.

In questo caso la (52) può porsi nella forma:

$${}_p\hat{\bar{Y}} = \sum_{a=1}^{A} \frac{\sum_{i=1}^{n} K\ {}_aY_i}{\sum_{i=1}^{n} K\ {}_aX_i} \sum_{i=1}^{N_a} {}_aX_i = \sum_{a=1}^{A} \frac{{}_a\hat{Y}}{{}_a\hat{X}}\ {}_aX \qquad (70)$$

in cui:

$${}_a\hat{Y} = \sum_{i=1}^{n} K\ {}_aY_i \qquad \text{e} \qquad {}_a\hat{X} = \sum_{i=1}^{n} K\ {}_aX_i \qquad (71)$$

L'espressione (70) è nota in letteratura con il nome di *stimatore del rapporto post-stratificato.*

Riteniamo utile osservare che in presenza di post-stratificazione alcuni post-strati possono risultare vuoti in questo caso l'inconveniente può essere superato aggregando opportunamente i post-strati stessi (Cochran, 1977).

È da aggiungere, infine, che nella situazione in cui si disponga di un insieme di informazioni a priori è preferibile utilizzarne alcune per il disegno ed altre per la costruzione dello stimatore.

Se, ad esempio, ci troviamo nella situazione:

$$\{N,n,y,X,z\}$$

è stato dimostrato che se l'informazione x viene utilizzata per stratificare le N unità di u (a meno di una forte relazione stocastica di tipo lineare tra y ed x), non conviene utilizzarla nuovamente nella costruzione dei procedimenti di stima.

Precisione delle stime dei parametri oggetto d'indagine

Come affermato nei capitoli precedenti, la stima dei parametri di interesse nella popolazione viene effettuata utilizzando la media campionaria della variabile oggetto di indagine, nel senso che alle unità di u non campione, viene attribuita la media del carattere y calcolata sulla base delle unità campione.

Se ad esempio, il parametro oggetto di studio è la media del carattere y in U, si avrà (Royall, 1971):

$$\hat{\bar{Y}} = \frac{1}{N} \left[\sum_{i=1}^{n} Y_i + (N - n) \sum_{i=1}^{n} \frac{Y_i}{n} \right] = \frac{1}{n} \sum_{i=1}^{n} Y_i \qquad (72)$$

mentre se siamo interessati alla stima di Y

$$\hat{Y} = \sum_{i=1}^{n} Y_i + (N - n)\,\hat{\bar{Y}} = \frac{N}{n} \sum_{i=1}^{n} Y_i \qquad (73)$$

In presenza di un disegno stratificato si avrà (per brevità ipotizziamo che si voglia stimare il totale del carattere y):

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} Y_{hi} + \sum_{h=1}^{H} (N_h - n_h)\;\hat{\bar{Y}}_h \qquad (74)$$

dove:

$$\hat{\bar{Y}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}$$

Se invece si tratta di un disegno a due stadi, per la stima di Y si ha:

$$\hat{Y} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} Y_{ij} + \sum_{i=1}^{n} (M_i - m_i)\,\hat{\bar{Y}}_i + \frac{N - n}{n} \sum_{i=1}^{n} M_i \hat{\bar{Y}}_i \qquad (75)$$

Come è facile vedere, la stima del totale Y si compone di tre parti; la prima rappresenta i dati osservati dal campione, la seconda è l'estensione alle unità di secondo stadio «non osservate», per ciascuna unità di primo stadio campionata, della media:

$$\hat{\bar{Y}}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} Y_{ij} \qquad (76)$$

delle rilevazioni sulle unità di secondo stadio *osservate* ed infine la terza parte è l'estensione alle unità di primo stadio *non osservate* della media delle *stime* dei totali delle unità di primo stadio *osservate*.

È noto che sintetizzando un fenomeno attraverso la sua media si ha una perdita di informazioni sul fenomeno stesso; tale perdita è direttamente legata alla sua variabilità.

In altre parole, volendo prevedere le modalità che il carattere y assumerà in una qualsiasi unità di U non osservata nel campione, la probabilità di sbagliare sarà tanto maggiore quanto maggiore è la variabilità del fenomeno. In assenza di informazioni ausiliarie, se viene attribuita la media campionaria ad una qualsiasi unità non osservata si è sicuri di avere minor probabilità di allontanarsi molto dal valore vero, che non attribuendogli qualsiasi altra modalità di y.

Una misura che quantifica la perdita di informazione sul parametro Y è la varianza dello stimatore. Come è stato già precisato nel capitolo 1, il campionamento casuale ha lo scopo di far sì che il valore assunto dal carattere y nella generica estrazione campionaria i (i=1,..., n) sia una variabile casuale: indichiamola con y_i. Le variabili casuali (v.c.) le cui realizzazioni danno luogo agli n risultati campionari possono indicarsi con $(y_1,\ldots,y_n)$.

La v.c. media campionaria $\bar{y}$ tenderà (per il teorema del limite centrale) ad avere una distribuzione normale (Vitali, 1987) con media:

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

e varianza σ^2/n, dove (Diana e Salvan, 1987):

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \overline{Y})^2 \qquad (77)$$

Infatti, quando le n selezioni sono indipendenti, si ha:

$$V(\bar{y}) = V\left[\sum_{l=1}^{n} \frac{y_l}{n}\right] = \frac{1}{n^2} \sum_{l=1}^{n} V[y_l] = \frac{\sigma^2}{n} \qquad (78)$$

Invece, nel caso in cui le n selezioni sono dipendenti occorre considerare la covarianza tra le coppie di v.c. $(y_l\ y_{l'})$ con $l \neq l'$ $(= 1,2,\ldots, N)$. Si ha:

$$C(y_l\ y_{l'}) = \sum_{i=1}^{N} \sum_{j=1}^{N} P\{y_i, y_j\} [(y_i - \bar{y})\ (y_j - \bar{y})] =$$

$$= \frac{1}{N\ (N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} [(y_i - \bar{y})\ (y_j - \bar{y})] = \qquad (79)$$

$$= \frac{1}{N\ (N-1)} \sum_{i=1}^{N} (y_i - \bar{y}) \quad \sum_{j=1}^{N} (y_j - \bar{y}) - \sum_{i=1}^{N} (y_i - \bar{y})^2$$

Poiché:

$$\sum_{i=1}^{N} (y_i - \bar{y}) = \sum_{j=1}^{N} (y_j - \bar{y}) = 0 \quad \text{e} \quad \sum_{i=1}^{N} (y_i - \bar{Y})^2 = N\sigma^2 \qquad (80)$$

la (79) diventa:

$$C(y_l\ y_{l'}) = -\ \frac{\sigma^2}{N-1} \qquad (81)$$

La varianza della v.c. $\bar{y}$ è pertanto definita dalla espressione:

$$V(\bar{y}) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 + \sum_{l=1}^{n} \sum_{l'=1}^{n} C(y_l\ y_{l'}) =$$

$$= \frac{1}{n^2}\ n\ \sigma^2 - n\ (n\text{-}1)\ \frac{\sigma^2}{N\text{-}1} = \qquad (82)$$

$$= \frac{1}{n}\ \sigma^2 - \frac{n\text{-}1}{N\text{-}1}\ \sigma^2 = \frac{N-n}{N-1}\ \frac{\sigma^2}{n}$$

I risultati finora ottenuti si basano sull'ipotesi che la N-upla $Y_1,\ldots, Y_N$ sia fissata e che la casualità dei valori campionari sia dovuta al processo di selezione.

Mostriamo ora che gli stessi risultati si possono ottenere anche nell'ambito di un approccio che considera la N-upla $Y_1,\ldots, Y_N$ come determinazione di N v.c. $y_1,\ldots,y_N$ indipendenti ed identicamente distribuite.

Seguendo tale approccio si può porre

$$y_l = \mu + \varepsilon_l \qquad (i = 1,\ldots,N) \qquad (83)$$

dove μ è la media (costante) delle N v.c. ed ε_i ha media zero e varianza costante.

La selezione del campione implica quindi la selezione di n delle N v.c. che hanno generato la popolazione; ciò induce un legame tra la n v.c. selezionate, espresso da:

$$\rho(y_l\ y_{l'}) = -\ \frac{C(y_l\ y_{l'})}{\sigma_l\ \sigma_{l'}} \qquad (84)$$

Poiché $\sigma_l = \sigma_{l'}$ e $C(y_l\ y_{l'})$ è data dalla (81) si ha:

$$\rho(y_l\ y_{l'}) = -\ \frac{1}{N-1} \qquad (85)$$

che costituiscono gli elementi della matrice **D** inserita nella (20).

Dalla (27) segue poi:

$$V(\bar{y}) = \frac{\sigma^2}{n} [1 + (n - 1)\ \rho]$$

da cui si ricava (Kendall, 1977):

$$V(\bar{y}) = \frac{\sigma^2}{n}\ \frac{N-n}{N-1}$$

che è esattamente la (82).

A questo punto è opportuno sottolineare che in entrambe le formule viene assunta la conoscenza della variabilità del carattere y in U, ossia viene posto σ^2 noto. Nella quasi totalità delle indagini campionarie ciò non si verifica per cui si rende necessario pervenire ad una stima della precisione dello stimatore invece che ad un calcolo esatto della stessa. Ponendo:

$$S^2 = \frac{N}{N-1}\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2 \qquad (86)$$

Sostituendo S^2 con σ^2 nella (82) si ottiene

$$V(\bar{y}) = \frac{N-n}{N}\frac{S^2}{n} \qquad (87)$$

Sapendo che la stima campionaria corretta di S^2 è data da s^2, dove:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \hat{\overline{Y}})^2 \qquad (88)$$

si ha che una stima della (82) è ottenuta semplicemente sostituendo s^2 ad S^2 nella (87) ottenendo:

$$\hat{V}(\hat{\overline{Y}}) = \frac{N-n}{N}\frac{s^2}{n} \qquad (89)$$

Per la varianza della stima del totale del carattere y in U, basta tener presente che:

$$\hat{Y} = N\,\hat{\overline{Y}} \qquad (90)$$

e poiché N è una costante conosciuta si ha:

$$V(\hat{Y}) = N^2\,V(\hat{\overline{Y}}) \qquad (91)$$

Per cui una stima di $V(\hat{Y})$ si ha sostituendo al secondo termine della (91) la (89), ottenendo (Cochran, 1977):

$$\hat{V}(\hat{Y}) = N\frac{N-n}{n}s^2 \qquad (92)$$

Come è facile osservare una misura della precisione della stima è ottenuta attraverso una stima della variabilità del carattere oggetto di indagini y, espressa dal fattore s^2 e dal fattore $(N-n)/N$ che risulta dipendere dal numero di unità appartenenti a $\bar{s}$, ossia la parte non osservata della popolazione. È immediato osservare che per $n = N$ il campione copre l'intera poplazione e la (92) è pari a zero, ossia la varianza campionaria è nulla.

Nei successivi capitoli verranno descritte le espressioni della varianza campionaria con riferimento a strategie campionarie più complesse.

Infine, riteniamo opportuno aggiungere che una volta determinata la stima della varianza, definita dalla (92), è possibile calcolare altre statistiche di grande importanza quali:

- l'errore di campionamento assoluto, convenzionalmente indicato con il simbolo $\hat{\delta}(\hat{Y})$, espresso come radice quadrata della $\hat{V}(\hat{Y})$;
- l'errore di campionamento relativo, convenzionalmente indicato con $\hat{\varepsilon}(\hat{Y})$, espresso come rapporto tra l'errore assoluto $\hat{\delta}(\hat{Y})$ e la stima $\hat{Y}$;
- l'intervallo di confidenza, generalmente riferito al livello di fiducia P=0,95, i cui estremi sono così definiti: $\hat{Y}-2\hat{\delta}(\hat{Y})$; $\hat{Y}+2\hat{\delta}(\hat{Y})$.

Proprietà degli stimatori

Lo stimatore del parametro di interesse nella popolazione è stato definito come *funzione dei dati campionari* ed indicato con t[Y(s)]. È evidente che tale definizione è molto generale e con scarsi risvolti applicativi, in quanto lo stimatore t non può essere una arbitraria funzione di Y(s) ma deve necessariamente rispondere a precisi requisiti. In generale questi sono (Fabbris, 1989):

- appartenenza ad una classe ben definita di funzioni (ad esempio, la classe delle funzioni lineari ed omogenee),
- dovere avere almeno una delle seguenti proprietà: correttezza, consistenza, efficienza e robustezza.

Si può notare che tali proprietà, riferite agli stimatori di una certa classe C ne individuano un sottoclasse C'. Ad esempio se C è la classe delle funzioni lineari omogenee C' potrebbe rappresentare la classe delle funzioni t' tali che (Giusti, 1983):

$$E[t'] = \theta \qquad (93)$$

Ossia la classe degli stimatori corretti di θ, nell'ambito di C.

È evidente quindi che la ricerca di quelle funzioni che godono delle proprietà elencate riduce notevolmente la classe degli stimatori utilizzabili nella pratica. È comunque da osservare che, nella teoria classica del campionamento da popolazione finite,

non è possibile ridurre tale classe ad una unica funzione t* (Godambe, 1965).

Dal punto di vista operativo, le proprietà più rilevanti sono: la correttezza e l'efficienza (Vitali, 1987).

La prima è stata definita dalla (93) mentre la seconda è una proprietà relativa che mette a confronto la variabilità di due stimatori t_1 e t_2. Si dirà che t_1 è più efficente di t_2 se, a parità di disegno campionario, si ha:

$$\frac{V[t_1]}{V[t_2]} < 1 \qquad (94)$$

Per quanto riguarda la consistenza e la robustezza, essendo la prima una proprietà asintotica e la seconda resa quasi superflua dalla apparente assenza di ipotesi di base nell'approccio classico all'inferenza da popolazione finite, queste non vengono solitamente prese in considerazione nel campionamento. Riteniamo comunque opportuno, per ragioni di completezza, trattarle in questa sede.

Della consistenza esistono almeno due definizioni. La prima, dovuta Cochran (1977), considera consistenti quegli stimatori la cui varianza si annulla per n = N; questa viene chiamata, in letteratura, C-consistenza. La seconda, dovuta a Murthy (1967) ed Hansen, Hurwitz e Madow (1953), ipotizza una successione di campioni (dalla stessa popolazione) e la corrispondente successione di stime $\theta_1, \ldots, \theta_m, \ldots$ ottenute dallo stesso stimatore t, in cui:

$$\theta_m = \frac{1}{m} \sum_{i=1}^{m} \theta_i \qquad (95)$$

Lo stimatore viene definito consistente se:

$$\lim_{m \to \infty} \theta_m = \theta \qquad (96)$$

$$\lim_{m \to \infty} V[\theta_m] = 0 \qquad (97)$$

Ritornando ora al concetto di correttezza dimostriamo che lo

stimatore diretto del totale del carattere y, nella popolazione oggetto di indagine, definito da:

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} Y_i \qquad (98)$$

é corretto. Riscriviamo la (98) in termini dell'insieme delle etichette {s}:

$$\hat{Y}(s) = \frac{N}{n} \sum_{i \varepsilon s} Y_i \qquad (99)$$

Ricordando che U_c indica l'universo campionario e considerando il caso in cui:

$$d(s) = \frac{n!\,(N-n)!}{N!}$$

la correttezza si può esprimere con:

$$\sum_{s \in U_c} d(s)\, \hat{Y}(s) = \sum_{i=1}^{N} Y_i \qquad (100)$$

La (100) si può scrivere:

$$\sum_{s \in U_c} \frac{n!\,(N-n)!}{N!} \frac{N}{n} \sum_{i \in s} Y_i = \sum_{i=1}^{N} Y_i \qquad (101)$$

Il primo membro della (101) può riscriversi nella forma:

$$\frac{n!\,(N-n)!}{N!} \frac{N}{n} \sum_{s \in U_c} \sum_{i \in s} Y_i \qquad (102)$$

dalla quale segue che:

$$\sum_{s \in U_c} \sum_{i \in s} Y_i = \sum_{i=1}^{N} \binom{N-1}{n-1} Y_i \qquad (103)$$

poiché i campioni in U_c, che contengono la generica unità i, sono pari a:

$$\binom{N-1}{n-1}$$

che rappresenta tutti i modi di associare all'unità i tutti i possibili campioni di numerosità (n-1) selezionabili dalla popolazione, di N-1 unità. Sostituendo la (103) nella (102) e tenendo presente che:

$$\binom{N-1}{n-1}\binom{N}{n}^{-1} = \frac{n!\,(N-n)!}{N!}\,\frac{(N-1)!}{(n-1)!\,[(N-1)-(n-1)]!} =$$

$$= \frac{n(n-1)!(N-n)!}{N(N-1)!}\,\frac{(N-1)!}{(n-1)!(N-n)!} = \frac{n}{N} \qquad (104)$$

il primo membro della (101) diventa:

$$\binom{N}{n}^{-1}\frac{N}{n}\binom{N-1}{n-1}\sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} Y_i \qquad (105)$$

che dimostra il risultato cercato.

L'estensione a disegni più complessi è relativamente immediata.

Vediamo ora che lo stimatore *rapporto* non è corretto. Intanto uno stimatore non corretto viene definito *distorto* ed una misura della distorsione viene fornita da:

$$B(t) = E(t) - \theta \qquad (106)$$

Consideriamo il caso della stima del totale del carattere y ed esprimiamo la (106) relativamente allo stimatore *rapporto*:

$$B(\hat{\bar{Y}}) = E\left(\frac{\hat{Y}}{\hat{X}}X\right) - Y \qquad (107)$$

Questa si può scrivere come:

$$B(\hat{\bar{Y}}) = -\left[E(\hat{Y}) - E(\hat{X})\,E\left(\frac{\hat{Y}}{\hat{X}}\right)\right] = -C\left[\frac{\hat{Y}}{\hat{X}}, \hat{X}\right] \qquad (108)$$

Infatti si ha:

$$C\left[\frac{\hat{Y}}{\hat{X}}, \hat{X}\right] = E\left\{\left[\frac{\hat{Y}}{\hat{X}} - E\left(\frac{\hat{Y}}{\hat{X}}\right)\right]\left[\hat{X} - E(\hat{X})\right]\right\} =$$

$$= E\left\{\left(\frac{\hat{Y}}{\hat{X}}\hat{X}\right) - E\left(\frac{\hat{Y}}{\hat{X}}\right)E(X)\right\} = E(\hat{Y}) - E(\hat{X})\,E\left(\frac{\hat{Y}}{\hat{X}}\right) \qquad (109)$$

Inoltre, date due generiche variabili u e v, il valore massimo della covarianza C (u, v) è pari a $[V(u)V(v)]^{1/2}$, ossia quando $\rho = 1$ (Leti, 1983).

In virtù di questo risultato il valore massimo di $B(\hat{\bar{Y}})$ è:

$$\max B(\hat{\bar{Y}}) = -\left[V\left(\frac{\hat{Y}}{\hat{X}}\right)V(\hat{X})\right]^{1/2} \qquad (110)$$

Dalle considerazioni appena svolte segue che la distorsione è massima quando tra le due variabili y ed x esiste la relazione:

$$\frac{y}{x} = k\,x \qquad (111)$$

e quindi y=k, ossia quando non esiste alcun legame lineare tra i due caratteri.

Per contro, come accennato precedentemente, tale distorsione si annulla se tra il carattere y ed il carattere x esiste una relazione lineare stocastica del tipo:

$$y = \alpha x + \varepsilon \qquad (112)$$

$$E[\varepsilon] = O$$

Inoltre, lo stimatore rapporto ha varianza minima quando:

$$V[y] = \beta x^g \qquad \text{con } 1 < g < 2 \qquad (113)$$

I motivi che rendono lo stimatore rapporto più vantaggioso di altri stimatori (diretto, regressione, ecc.) poggiano sulle seguenti circostanze: risulta facilmente trattabile dal punto di vista informatico; generalmente, è più efficiente degli altri stimatori citati. Infatti, affinché lo stimatore del rapporto risulti più efficiente dello stimatore diretto è sufficiente che (Cochran, 1977):

$$C[\hat{Y}, \hat{X}] > \frac{1}{2} R V(\hat{X}) \qquad (114)$$

in cui:

$$R = \frac{Y}{X}$$

PARTE 2

CRITERI DI SELEZIONE DEL CAMPIONE, METODI DI STIMA E VARIANZE DI CAMPIONAMENTO

CAPITOLO 4 - CONSIDERAZIONI INTRODUTTIVE E NOTAZIONI SIMBOLICHE

Introduzione

Nella prima parte di questo volume, didatticamente preliminare, abbiamo offerto alcuni concetti generali e introduttivi relativi alle indagini campionarie e proposto i fondamenti della teoria statistica dell'inferenza per quanto attiene al disegno di campionamento ed ai metodi di stima dei parametri della popolazione.

Questa seconda parte racchiude una trattazione sufficientemente completa dei criteri di selezione delle unità campionarie e di alcuni fondamentali metodi di stima (stimatori diretti, del rapporto e del rapporto post-stratificati), di uso più frequente nelle indagini su larga scala condotte dai maggiori centri di informazione statistica a livello internazionale. Relativamente a ciascuno degli stimatori suddetti, si studiano inoltre le espressioni della varianza e della corrispondente stima, le quali svolgono un ruolo di grande importanza sia ai fini della valutazione del livello di precisione dei risultati prodotti da un'indagine campionaria, sia nell'ambito della problematica della determinazione della dimensione del campione.

Su tali contenuti qualcosa va aggiunta circa il criterio metodologico con il quale essi vengono proposti. L'esperienza ci ha suggerito di usare il principio generale di esporre ogni metodo partendo dal campionamento casuale semplice estendendo poi l'illustrazione al campionamento ad uno stadio stratificato e a quello a due stadi con stratificazione al livello delle unità primarie, che costituiscono i contesti campionari più comuni nella realtà di ricerca; d'altra parte, il lettore non dovrebbe faticare nell'individuare le considerazioni che si applicano a campioni su qualsivoglia numero di stadi.

L'idea di iniziare l'esposizione con riferimento al campionamento semplice ci pare didatticamente efficace e storicamente valida per mostrare l'utilità dei metodi e delle tecniche statistiche proposti (e in genere adottati in contesti campionari più complessi) e per far nascere l'opportunità di ulteriori approfondimenti.

Anche se i capitoli appaiono densi sul piano degli sviluppi formali, l'esposizione si concentra sull'essenziale di quanto occorre affinché i metodi descritti risultino comprensibili a chi abbia interessi statistici di natura operativa. Per non appesantire matematicamente la trattazione non ci siamo dilungati in sviluppi teorici dimostrativi, rinviando alla bibliografia per gli approfondimenti o per chi volesse acquisire metodi più sofisticati.

Questa seconda parte inoltre condiziona la successiva ove si affrontano i problemi riguardanti la stratificazione e il calcolo

della numerosità campionaria, che costituiscono le fasi operative fondamentali nella progettazione del disegno di campionamento per indagini statistiche.

Infine, nel presente capitolo, abbiamo ritenuto importante premettere, con riferimento ai tipi di campionamento sopra menzionati, le caratteristiche strutturali della popolazione e del campione, nonché le notazioni simboliche alle quali si farà continuo riferimento per la descrizione dei metodi proposti.

Campionamento casuale semplice

Data una popolazione costituita da un numero finito di elementi, l'operazione di campionamento semplice consiste nel selezionare a caso dalla popolazione un prefissato numero di elementi.

Tra i vari criteri di selezione delle unità da includere nel campione, maggiormente si conviene alle indagini campionarie effettive quello che va sotto il nome di *scelta a caso senza reimmissione* che consiste nel separare dalla popolazione un elemento per volta, in modo che l'elemento stesso non possa cadere sotto osservazione più di una sola volta.

Tuttavia, in teoria, è possibile un altro criterio di selezione noto con il nome di *scelta a caso con reimmissione*, in base al quale le osservazioni su un prefissato numero di unità campionarie avvengono come se fossero fatte una per volta, dopo aver estratto a caso un solo elemento e averlo subito ricollocato nella popolazione in modo che uno stesso elemento possa cadere più volte sotto osservazione.

Nel seguito del presente volume, per rimanere in condizioni di maggiore generalità, ci atterremo al primo criterio di selezione; d'altra parte, tutte le espressioni che scriveremo, mediante opportuni passaggi al limite, si possono trasformare nelle corrispondenti espressioni relative al criterio di scelta con reimmissione.

Passiamo ora a descrivere le caratteristiche strutturali della popolazione e del campione e i simboli di cui si farà uso nel prosieguo:

i	indice di unità
N	dimensione della popolazione
$f = \frac{n}{N}$	frazione di campionamento
y	generica variabile oggetto di rilevazione
Y_i	valore della variabile y relativo all'unità i

$Y = \sum_{i=1}^{N} Y_i$	totale della variabile y nella popolazione
$\overline{Y} = \frac{Y}{N}$	media della variabile y nella popolazione
$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \overline{Y})^2$	varianza della variabile y nella popolazione
$\hat{\overline{Y}} = \frac{1}{n} \sum_{i=1}^{n} Y_i$	media della variabile y nel campione
$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \hat{\overline{Y}})^2$	varianza della variabile y nel campione.

Campionamento ad uno stadio stratificato

Ogni stima, ottenuta mediante l'esame di un campione estratto a caso da una popolazione assegnata, comporta necessariamente il rischio di commettere un errore (denominato *errore di campionamento*), dovuto al fatto di selezionare dalla popolazione una parte soltanto di elementi.

Il mezzo più semplice per ridurre l'errore di campionamento consiste nell'aumentare la numerosità campionaria, in quanto a mano a mano che la rilevazione parziale tende a diventare totale la struttura del campione diventa più simile a quella della popolazione, ovvero il campione diventa sempre più rappresentativo della popolazione da cui proviene; conseguentemente il livello di precisione della stima aumenta.

A parità di dimensione del campione, però, si può migliorare la precisione della stima ricorrendo alla stratificazione della popolazione, cioè alla sua scomposizione in parti omogenee (o strati) rispetto al carattere oggetto di indagine, e selezionando da ciascuno strato un numero prefissato di unità con cui formare in complesso un campione di determinata dimensione, denominato *campione ad uno stadio stratificato*.

Per questo tipo di campionamento possiamo individuare caratteristiche strutturali e notazioni simboliche partendo da quelle definite per il campionamento semplice, salvo l'aggiunta di un indice distintivo di strato.

Pertanto indichiamo con:

h	indice di strato ($h = 1,\ldots,H$)
i	indice di unità
N_h	dimensione della sub-popolazione dello strato h
$N = \sum_{h=1}^{H} N_h$	dimensione della popolazione
n_h	dimensione del campione dello strato h
$n = \sum_{h=1}^{H} n_h$	dimensione complessiva del campione
$f_h = \frac{n_h}{N_h}$	frazione di campionamento relativa allo strato h
$f = \frac{n}{N}$	frazione di campionamento totale
y	generica variabile oggetto d'indagine
Y_{hi}	valore della variabile y relativo all'unità i appartenente allo strato h
$Y_h = \sum_{i=1}^{N_h} Y_{hi}$	totale della variabile y relativo alla sub-popolazione dello strato h

$\overline{Y}_h = \frac{Y_h}{N_h}$	media della variabile y relativa alla sub-popolazione dello strato h
$Y = \sum_{h=1}^{H} Y_h = \sum_{h=1}^{H} \sum_{i=1}^{N_h} Y_{hi}$	totale della variabile y nella popolazione
$\overline{Y} = \frac{Y}{N}$	media della variabile y nella popolazione
$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \overline{Y}_h)^2$	varianza della variabile y relativa alla sub-popolazione dello strato h
$\hat{\overline{Y}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}$	media della variabile y relativa al campione dello strato h
$s^2_h = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (Y_{hi} - \hat{\overline{Y}}_h)^2$	varianza della variabile y relativa al campione dello strato h

Le notazioni sopra illustrate possono essere riassunte mediante i quadri simbolici seguenti:

Rappresentazione simbolica relativa alla popolazione

Strato	Distribuzione della variabile y	Dimensione	Totale	Media	Varianza
1	$Y_{11} \ldots Y_{1i} \ldots Y_{1N_1}$	N_1	Y_1	$\overline{Y}_1$	S_1^2
h	$Y_{h1} \ldots Y_{hi} \ldots Y_{hN_h}$	N_h	Y_h	$\overline{Y}_h$	S_h^2
H	$Y_{H1} \ldots Y_{Hi} \ldots Y_{HN_H}$	N_H	Y_H	$\overline{Y}_H$	S_H^2
Totale	— . . — . . —	N	Y	$\overline{Y}$	—

Rappresentazione simbolica relativa al campione

Strato	Distribuzione della variabile y	Dimensione	Frazione di campionamento	Media	Varianza
1	Y_{11} . . Y_{1i} . . Y_{1n_1}	n_1	f_1	$\hat{\bar{Y}}_1$	s_1^2
.	. . . . .	.	.	.	.
.	. . . . .	.	.	.	.
h	Y_{h1} . . Y_{hi} . . Y_{hn_h}	n_h	f_h	$\hat{\bar{Y}}_h$	s_h^2
.	. . . . .	.	.	.	.
.	. . . . .	.	.	.	.
H	Y_{H1} . . Y_{Hi} . . Y_{Hn_H}	n_H	f_H	$\hat{\bar{Y}}_H$	s_H^2
Totale	— . . — . . —	n	f	—	—

Campionamento a due stadi con stratificazione delle unità primarie

Come la stratificazione ha generalmente l'effetto di aumentare la precisione delle stime fornite dall'indagine, così il ricorso al campionamento a due stadi è determinato non dal desiderio di conseguire il medesimo effetto, in senso assoluto, rispetto al campionamento ad uno stadio, bensì dalla necessità di ridurre il costo dell'operazione o, meglio, di conseguire la maggiore precisione possibile, pur contenendo il costo nei limiti delle disponibilità finanziarie.

Allo scopo di rendere più efficienti le stime, un campione a due stadi è generalmente stratificato al livello delle unità primarie.

Per trattare questo tipo di campionamento nel seguito si farà uso dei seguenti simboli:

h indice di strato ($h = 1,\ldots, H$)

i indice di unità primaria

j indice di unità secondaria

N_h numero di unità primarie dello strato h

$N = \sum_{h=1}^{H} N_h$ numero complessivo di unità primarie

n_h numero di unità primarie campione dello strato h

$n = \sum_{h=1}^{H} n_h$ numero complessivo di unità primarie campione

M_{hi} numero di unità secondarie dell'unità primaria i dello strato h

$M_h = \sum_{i=1}^{N_h} M_{hi}$ numero di unità secondarie dello strato h

$M = \sum_{h=1}^{H} M_h$ numero complessivo di unità secondarie

m_{hi} numero di unità secondarie campione dell'unità primaria i dello strato h

$m_h = \sum_{i=1}^{n_h} m_{hi}$ numero di unità secondarie campione dello strato h

$m = \sum_{h=1}^{H} m_h$ numero complessivo di unità secondarie campione

$f_{1h} = \frac{n_h}{N_h}$ frazione di campionamento primaria dello strato h

$f_{2hi} = \frac{m_{hi}}{M_{hi}}$ frazione di campionamento secondaria nell'unità primaria i dello strato h

$f_h = \frac{m_h}{M_h}$ frazione di campionamento totale dello strato h

$f = \frac{m}{M}$ frazione di campionamento totale

y generica variabile oggetto d'indagine

Y_{hij} valore della variabile y dell'unità secondaria j dell'unità primaria i dello strato h

$$Y_{hi} = \sum_{j=1}^{M_{hi}} Y_{hij}$$ totale della variabile y dell'unità primaria i dello strato h

$$Y_h = \sum_{i=1}^{N_h} Y_{hi}$$ totale della variabile y dello strato h

$$Y = \sum_{h=1}^{H} Y_h$$ totale della variabile y nella popolazione

$$\overline{Y}_{hi} = \frac{Y_{hi}}{M_{hi}}$$ media della variabile y nell'unità primaria i dello strato h

$$\overline{Y}_h = \frac{Y_h}{M_h}$$ media della variabile y dello strato h

$$\overline{Y} = \frac{Y}{M}$$ medie della variabile y nella popolazione

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{hi})^2$$ varianza della variabile y fra i totali delle unità primarie dello strato h

$$S_{hi}^2 = \frac{1}{M_{hi} - 1} \sum_{j=1}^{M_{hi}} (Y_{hij} - \overline{Y}_{hi})^2$$ varianza della variabile y dentro l'unità primaria i dello strato h

$$\hat{\overline{Y}}_{hi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} Y_{hij}$$ media della variabile y relativa alle m_{hi} unità campionarie

$$s_{hi}^2 = \frac{1}{m_{hi} - 1} \sum_{j=1}^{m_{hi}} (Y_{hij} - \hat{\overline{Y}}_{hi})^2$$ varianza della variabile y relativa alle m_{hi} unità campionarie

Quadro simbolico relativo alla popolazione del generico strato h

Unità primarie	Distribuzione della variabile y	Numero di unità secondarie	Totale	Media	Varianza fra unità primarie	Varianza interna alle unità primarie
1	$Y_{h11} \cdot\cdot Y_{h1j} \cdot\cdot Y_{h1M_{h1}}$	M_{h1}	Y_{h1}	$\overline{Y}_{h1}$	—	S_{h1}^2
i	$Y_{hi1} \cdot\cdot Y_{hij} \cdot\cdot Y_{hiM_{hi}}$	M_{hi}	Y_{hi}	$\overline{Y}_{hi}$	—	S_{hi}^2
N_h	$Y_{hN_{h1}} \cdot\cdot Y_{hN_{hj}} \cdot\cdot Y_{hN_hM_{hN_h}}$	M_{hN_h}	Y_{hN_h}	$\overline{Y}_{hN_h}$	—	$S_{hN_h}^2$
Totale	— .. — .. —	M_h	Y_h	$\overline{Y}_h$	S_h^2	—

Quadro simbolico relativo al campione del generico strato h

Unità primarie	Distribuzione della variabile y	Numero di unità secondarie	Frazione di campionamento	Media	Varianza interna alle unità primarie
1	$Y_{h1i} \cdot\cdot Y_{h1j} \cdot\cdot Y_{h1m_{h1}}$	m_{h1}	f_{2h1}	$\hat{\overline{Y}}_{h1}$	s_{h1}^2
i	$Y_{hi1} \cdot\cdot Y_{hij} \cdot\cdot Y_{him_{hi}}$	m_{hi}	f_{2hi}	$\hat{\overline{Y}}_{hi}$	s_{hi}^2
n_h	$Y_{hn_h1} \cdot\cdot Y_{hn_hj} \cdot\cdot Y_{hn_hm_{hn_h}}$	m_{hn_h}	f_{2hn_h}	$\hat{\overline{Y}}_{hn_h}$	$s_{hn_h}^2$
Totale	— .. — .. —	m_h	f_{2h}	—	—

CAPITOLO 5 - CRITERI DI SELEZIONE

Introduzione

La precisione dei risultati ottenuti da una indagine campionaria dipende non solo dalla dimensione del campione ma anche da altri elementi, quali il criterio di selezione del campione, il procedimento di stima, la stratificazione, ecc..

Il criterio di selezione è la procedura mediante la quale si effettua il sorteggio delle unità della popolazione.

Tale operazione può essere effettuata solo se le unità statistiche costituenti la popolazione oggetto di indagine sono identificabili, vale a dire se esistono elenchi, schedari, ecc. utilizzabili come *basi di campionamento*, mediante i quali ciascuna unità è identificabile univocamente (Kish, 1965).

Il criterio di selezione dipende sia dal tipo di estrazione che dal sistema probabilistico di selezione delle unità costituenti la popolazione oggetto d'indagine (Castellano ed Herzel, 1981).

Il tipo di estrazione può essere con reimmissione o senza reimmissione delle unità; il sistema probabilistico di selezione può essere a probabilità uguali o a probabilità variabili.

Nelle pagine seguenti illustreremo il significato dei criteri sopra introdotti. Esaminiamo in primo luogo il tipo di estrazione.

L'estrazione con reimmissione consiste nell'ottenere un campione di n unità mediante n estrazioni successive, reinserendo nella popolazione l'unità estratta prima di procedere alla successiva estrazione. Operando in tal modo la composizione della popolazione resta invariata ad ogni estrazione; ovviamente, con questo tipo di estrazione, la stessa unità può essere sorteggiata anche più di una volta.

Nella tabella seguente è rappresentato l'insieme dei possibili campioni U_c (noto con il nome di universo dei campioni) di ampiezza n = 2 estraibili da una popolazione costituita da N = 3 unità u_1, u_2 e u_3:

C_1	C_2	C_3	C_4	C_5	C_6
u_1	u_1	u_1	u_2	u_2	u_3
u_1	u_2	u_3	u_2	u_3	u_3

Nell'estrazione senza reimmissione le n unità del campione si ottengono mediante n estrazioni successive non reinserendo nella popolazione l'unità estratta prima di procedere alla successiva estrazione.

Pertanto con questo tipo di estrazione la popolazione si riduce di un'unità ad ogni estrazione e una stessa unità non può essere sorteggiata che una volta sola.

Nella seguente tabella sono rappresentati i diversi campioni C_i di ampiezza n = 2, che si possono ottenere estraendo le unità senza reimmissione da una popolazione di N = 3 unità u_1, u_2, u_3:

C_1	C_2	C_3
u_1	u_1	u_2
u_2	u_3	u_3

L'estrazione con probabilità uguali consiste nell'attribuire a ciascuna unità della popolazione la stessa probabilità di essere selezionata.

Per una maggiore comprensione di tale concetto consideriamo una popolazione costituita da N = 3 unità u_1, u_2 e u_3. Se si attribuisce a ciascuna delle tre unità la stessa probabilità di selezione, ovviamente pari a 0,33, si ottiene che il relativo sistema probabilistico di selezione è definito dall'insieme delle tre probabilità:

0,33; 0,33; 0,33

Nell'estrazione con probabilità variabili, invece, viene attribuita una probabilità di selezione generalmente diversa da un'unità all'altra della popolazione.

Con riferimento all'esempio sopra riportato uno dei possibili sistemi a probabilità variabili potrebbe essere costituito dall'insieme:

0,10; 0,40; 0,50

in cui i tre valori rappresentano rispettivamente le probabilità di selezione di u_1, u_2, u_3.

Campionamento casuale semplice

Le considerazioni appena svolte avevano la finalità di illustrare i concetti relativi alle due componenti, tipo e sistema probabilistico di estrazione, che definiscono un criterio di selezione. Tali considerazioni, sviluppate a fini di chiarezza separatamente per le due componenti, non consentono, tuttavia, di comprendere in modo pieno il concetto di criterio di selezione che, come sottolineato all'inizio, si basa invece sulla definizione congiunta di tali componenti.

Nel seguente quadro simbolico sono indicati i quattro possibili criteri di selezione casuale:

Tipo di estrazione	Sistema probabilistico	
	Probabilità uguali (pu)	Probabilità variabili (pv)
Con reimmissione (cr)	(cr, pu)	(cr, pv)
Senza reimmissione (sr)	(sr, pu)	(sr, pv)

in cui:

(cr, pu): con reimmissione e probabilità uguali
(cr, pv): con reimmissione e probabilità variabili
(sr, pu): senza reimmissione e probabilità uguali
(sr, pv): senza reimmissione e probabilità variabili.

È da osservare tuttavia che nelle indagini effettive condotte su larga scala i criteri di selezione più frequentemente utilizzati sono quelli senza reimmissione, cioè (sr, pu) e (sr, pv).

L'opportunità di ricorrere a questi ultimi criteri risiede nel fatto che essi generalmente conducono a risultati più precisi di quelli che si otterrebbero selezionando le unità con reimmissione.

Nei successivi paragrafi, pertanto, la trattazione sarà limitata allo studio dei criteri (sr, pu) e (sr, pv).

Introduciamo le seguenti due definizioni:

— criterio di selezione (sr, pu): l'unità estratta non viene reintrodotta nella popolazione e ciascuna unità della popolazione ha la stessa probabilità di selezione alla prima estrazione;
— criterio di selezione (sr, pv): l'unità estratta non viene reintrodotta nella popolazione e le N unità della popolazione hanno probabilità di selezione diverse alla prima estrazione.

Esaminiamo in primo luogo il criterio (sr, pu).

A tal fine indichiamo con u_1, u_2,...., u_i,...., u_N le unità di un'ipotetica popolazione di dimensione N.

Supponiamo inoltre di voler estrarre un campione di ampiezza n.

In tale circostanza e sotto l'ipotesi che il tipo di estrazione sia senza reimmissione si ha la seguente situazione: alla prima estrazione è possibile selezionare una qualsiasi delle N unità della popolazione; alla seconda estrazione una delle rimanenti N-1 unità, e così via fino alla n-esima estrazione in cui è possibile selezionare una delle N-(n-1) unità rimaste.

Sotto l'ulteriore ipotesi che le N unità della popolazione abbiano la stessa probabilità di selezione alla prima estrazione, che indichiamo con il simbolo P_1, si ha:

$$P(u_1) = P(u_2) = \cdot\cdot = P(u_i) = \cdot\cdot = P(u_N) = P_1 = \frac{1}{N} \qquad (1)$$

Conseguentemente, la probabilità di selezione alla seconda estrazione, P_2, di ciascuna delle rimanenti N-1 unità è uguale a:

$$P_2 = \frac{1}{N-1} \qquad (2)$$

Procedendo in modo analogo per le successive estrazioni, si ottiene:

$$P_3 = \frac{1}{N-2}, \cdots\cdots, \quad P_n = \frac{1}{N-(n-1)} \qquad (3)$$

Le precedenti considerazioni avevano lo scopo di definire le probabilità di selezione relative alle n estrazioni; tali probabilità sono di fondamentale importanza anche per la definizione della *probabilità di inclusione*, la cui determinazione sta alla base dei procedimenti di stima dei parametri oggetto d'indagine (Horwitz e Thompson, 1952).

Definiamo probabilità d'inclusione di una generica unità u_i della popolazione la probabilità che tale unità sia inclusa nel campione.

Con riferimento, ad esempio, ad un campione di dimensione n tale probabilità risulta costituita dalla somma delle probabilità seguenti:

P_1^*, che la generica unità u_i sia selezionata alla prima estrazione;
P_2^*, che la generica unità u_i sia selezionata alla seconda estrazione, non essendo stata selezionata alla prima;
P_3^*, che la generica unità u_i sia selezionata alla terza estrazione, non essendo stata selezionata alle prime due estrazioni;

. . . . . .
. . . . . .

P_n^*, che la generica unità u_i sia selezionata alla n-esima estrazione, non essendo stata selezionata alle prime n-1 estrazioni.

La probabilità di inclusione della generica unità u_i della popolazione è pertanto espressa da:

$$\Pi(u_i) = P_1^* + P_2^* + \cdots\cdots + P_n^* \qquad (4)$$

Ovviamente P_1^* coincide con P_1; P_2^* è definito dal prodotto della probabilità che alla prima estrazione sia selezionata una unità diversa da u_i per la probabilità che u_i sia selezionata alla seconda estrazione; in formula cioè:

$$P_2^* = \frac{N-1}{N} P_2 = \frac{N-1}{N}\frac{1}{N-1} = \frac{1}{N} \qquad (5)$$

Procedendo in modo analogo si ha dunque:

$$P_3^* = \frac{N-2}{N} P_3 = \frac{N-2}{N}\frac{1}{N-2} = \frac{1}{N} \qquad (6)$$

. . . . . .
. . . . . .

$$P_n^* = \frac{N-(n-1)}{N} P_n = \frac{N-(n-1)}{N}\frac{1}{N-(n-1)} = \frac{1}{N}$$

In conclusione $\Pi(u_i)$ è uguale a:

$$\Pi(u_i) = \frac{1}{N} + \frac{1}{N} + \cdots\cdots + \frac{1}{N} = \frac{n}{N} \qquad (7)$$

Per chiarire il significato del criterio di selezione testè esaminato è utile sviluppare un esempio.

Consideriamo una popolazione costituita da N = 3 unità u_1, u_2, u_3, e supponiamo di voler estrarre da essa senza reimmissione e con probabilità uguali un campione di ampiezza n = 2; in questo caso si ha:

$$P_1 = \frac{1}{3} \quad \text{e} \quad P_2 = \frac{1}{2}$$

e quindi

$$P_1^* = \frac{1}{3} \quad \text{e} \quad P_2^* = \frac{1}{2}\frac{2}{3} = \frac{1}{3}$$

Conseguentemente la probabilità di inclusione della generica unità u_i della popolazione in un campione di ampiezza n = 2 è data da:

$$\Pi(u_i) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3} \qquad (i = 1,2,3)$$

Esaminiamo ora il criterio di selezione (sr, pv).

L'illustrazione sarà limitata al caso di due sole estrazioni, ossia al caso che si voglia estrarre un campione di dimensione n = 2 da una popolazione costituita dalle N unità $u_1, u_2, \ldots, u_i, \ldots, u_N$.

Tale scelta è dovuta sia alla complessità delle espressioni delle probabilità di selezione e di inclusione nel caso $n > 2$, sia perché nella maggior parte delle indagini concrete, generalmente basate su disegni campionari complessi stratificati, si ricorre all'estrazione di una o di due sole unità da ogni strato (Cochran, 1977).

Sotto l'ipotesi che la selezione venga effettuata senza reimmissione, la popolazione si riduce di un'unità dopo la prima estrazione.

Supponiamo inoltre che le probabilità di selezione alla prima estrazione relative alle N unità della popolazione siano diverse, ossia:

$$P(u_1) \neq P(u_2) \neq \cdots \neq P(u_i) \neq \cdots \neq P(u_N) \qquad (8)$$

Tenendo presente che l'esposizione è limitata al caso n = 2, si ha che la probabilità di selezione alla prima estrazione dell'unità u_i è uguale a:

$$P_1(u_i) = P(u_i) \qquad (i=1,\ldots,N) \qquad (9)$$

Indichiamo con u_j (j = 1,2, ...,i,..., N) l'unità estratta alla prima selezione; la probabilità di selezione relativa alla seconda estrazione di una qualsiasi delle N - 1 unità u_i è data da:

$$P_2(u_i) = \frac{P_1(u_i)}{1 - P_1(u_j)} \quad (i = 1, \ldots, N) \quad \text{e} \quad (i \neq j) \qquad (10)$$

La probabilità d'inclusione della generica unità u_i (i = 1, ..., N) della popolazione in un campione di ampiezza n = 2 è costituita dalla somma dei seguenti due termini:

— probabilità che l'unità u_i sia selezionata alla prima estrazione: $P_1^*(u_i)$

— probabilità che l'unità u_i sia selezionata alla seconda estrazione, non essendo stata selezionata alla prima: $P_2^*(u_i)$.

La probabilità d'inclusione della generica unità u_i della popolazione è pertanto espressa da:

$$\Pi(u_i) = P_1^*(u_i) + P_2^*(u_i) \qquad (i = 1,2,\ldots,N) \qquad (11)$$

Ovviamente $P_1^*(u_i)$ coincide con $P_1(u_i)$; $P_2^*(u_i)$ è definita dal prodotto della probabilità che alla prima estrazione sia selezionata un'unità diversa da u_i per la probabilità che u_i sia selezionata alla seconda estrazione; in formula cioè:

$$P_2^*(u_i) = \sum_{j \neq i}^{N} P_1(u_j) \quad \frac{P_1(u_i)}{1 - P_1(u_j)} \qquad (12)$$

In conclusione $\Pi(u_i)$ è uguale a:

$$\Pi(u_i) = P_1(u_i) + \sum_{j \neq i}^{N} P_1(u_j) \quad \frac{P_1(u_i)}{1 - P_1(u_j)} \qquad (13)$$

Nel caso in cui si estragga un campione di ampiezza n = 1, la probabilità di inclusione della generica unità u_i della popolazione è data ovviamente dell'espressione:

$$\Pi(u_i) = P_1^*(u_i) = P(u_i) \qquad (i = 1,..,N) \qquad (14)$$

Per chiarire il significato del criterio di selezione esaminato consideriamo l'esempio seguente.

Data una popolazione costituita da N = 3 unità u_1, u_2, u_3 supponiamo di voler estrarre, senza reimmissione, da essa un campione di ampiezza n = 2, assegnando alle unità le seguenti probabilità di selezione alla prima estrazione:

$P(u_1) = 0,6; \quad P(u_2) = 0,3; \quad P(u_3) = 0,1$

In tal caso si ha:

$P_1^*(u_1) = 0,6; \quad P_1^*(u_2) = 0,3; \quad P_1^*(u_3) = 0,1$

e dalla (12):

$$P_2^*(u_1) = \frac{0,6}{1-0,3}\,0,3 + \frac{0,6}{1-0,1}\,0,1 = 0,324$$

$$P_2^*(u_2) = \frac{0,3}{1-0,6}\,0,6 + \frac{0,3}{1-0,1}\,0,1 = 0,483$$

$$P_2^*(u_3) = \frac{0,1}{1-0,6}\,0,6 + \frac{0,1}{1-0,3}\,0,3 = 0,193$$

Le probabilità d'inclusione sono quindi:

$\Pi(u_1) = 0,6 + 0,324 = 0,924$

$\Pi(u_2) = 0,3 + 0,483 = 0,783$

$\Pi(u_3) = 0,1 + 0,193 = 0,293$

A conclusione di queste prime considerazioni sui criteri di selezione riteniamo utile illustrare alcuni metodi di selezione di rilevante importanza nell'ambito delle indagini condotte su larga scala.

Prima di affrontare tale argomento è opportuno sottolineare che il criterio che ha condotto alla relazione (13) si basa sulle seguenti probabilità:

— $P_1(u_i) \qquad (i = 1, \ldots, N)$

— $P_2(u_i) = \dfrac{P_1(u_i)}{1 - P_1(u_j)}$

dove $P_1(u_i)$ rappresenta la probabilità di selezione, alla prima estrazione, dell'unità u_i e $P_2(u_i)$ la probabilità di selezione, alla seconda estrazione, dell'unità u_i, non essendo stata selezionata u_i alla prima estrazione.

Ciò premesso, nel caso di indagini concrete, le probabilità $P_1(u_i)$ e $P_2(u_i)$ sono spesso definite come grandezze proporzionali alle dimensioni delle unità stesse (Brewer e Hanif, 1983).

Questo modo di definire le probabilità di selezione presenta una certa utilità, in quanto generalmente consente di aumentare il livello di precisione delle stime, a parità di condizioni.

A tale scopo si cerca una variabile, i cui valori siano noti prima di effettuare la selezione, correlata con la variabile oggetto di studio.

Così, ad esempio, se le unità sono i comuni, come nel caso delle indagini Istat sulle famiglie, la dimensione è definita dalla popolazione dei comuni stessi; conseguentemente un comune con popolazione maggiore di un altro ha una probabilità di selezione più elevata (Russo e Falorsi, 1985).

Indichiamo con 'a' la variabile in questione e con A_i il valore che essa assume sull'unità u_i. Pertanto la probabilità $P_1(u_i)$ è espressa da:

$$P_1(u_i) = \frac{A_i}{\sum_{i=1}^{N} A_i} \tag{15}$$

La probabilità $P_2(u_i)$ è data da:

$$P_2(u_i) = \frac{A_i}{\left(\sum_{i=1}^{N} A_i\right) - A_j} \tag{16}$$

Le probabilità appena definite possono essere, poi, utilizzate per la determinazione della probabilità di inclusione $\Pi(u_i)$ secondo la relazione (13).

Nelle situazioni concrete, tuttavia, data la complessità dell'espressione (13), le probabilità di inclusione vengono determinate attraverso criteri che, pur rendendo più agevole il calcolo di tali probabilità, sono basati su probabilità di selezione approssimativamente proporzionali alle dimensioni delle unità.

Secondo quest'ottica si muove, ad esempio, il *metodo di Brewer*, che descriveremo sempre con riferimento al caso di un campione costituito da due sole unità.

In base a tale metodo, alla generica unità u_i ($i = 1, \ldots, n$) viene attribuita:

— alla prima estrazione, una probabilità di selezione proporzionale al rapporto:

$$\frac{P_1(u_i)\,(1 - P_1(u_i))}{1 - 2\,P_1(u_i)} \qquad (17)$$

che indichiamo col nome di *probabilità corretta* (*revised probability* in lingua inglese);

— alla seconda estrazione, non essendo stata selezionata u_i alla prima estrazione, una probabilità di selezione data da:

$$\frac{P_1(u_i)}{1 - P_1(u_j)} \qquad (18)$$

dove u_j ($j = 1, \ldots, N$ con $j \neq i$) indica la generica unità diversa da u_i (selezionata alla prima estrazione).

Inoltre, si assume che ogni $P_1(u_i)$ ($i = 1,\ldots,N$) sia minore di 1/2, al fine di evitare che la corrispondente probabilità corretta definita dalla (17) possa assumere valori negativi. Osserviamo che le probabilità definite dalla (17) pur risultando inferiori ad uno non godono della proprietà che la loro somma risulti pari ad uno.

Allo scopo di avere un sistema di probabilità che rispetti quest'ultima condizione basta normalizzare le probabilità in oggetto dividendo ciascuna probabilità corretta per la quantità:

$$C = \sum_{i=1}^{N} \frac{P_1(u_i)\,(1 - P_1(u_i))}{1 - 2\,P_1(u_i)} \qquad (19)$$

Moltiplicando e dividendo per 2 la precedente espressione si ha:

$$C = \frac{1}{2} \sum_{i=1}^{N} \frac{P_1(u_i)\,(1 + 1 - 2\,P_1(u_i))}{1 - 2\,P_1(u_i)} \qquad (20)$$

ed essendo $\sum_{i=1}^{N} P_1(u_i) = 1$ si ottiene che:

$$C = \frac{1}{2}\left[\,1 + \sum_{i=1}^{N} \frac{P_1(u_i)}{1 - 2\,P_1(u_i)}\right] \qquad (21)$$

alla quale si fa generalmente riferimento per la normalizzazione delle probabilità in questione.

A questo punto possiamo considerare la probabilità di inclusione della generica unità u_i, che in base alle relazioni (17), (18) e (21), è espressa da:

$$\Pi(u_i) = \frac{P_2(u_i)\,(1 - P_1(u_i))}{C\,(1 - 2\,P_1(u_i))} + \frac{1}{C} \sum_{j \neq i}^{N} \frac{P_1(u_j)\,(1 - P_1(u_j))}{1 - 2\,P_1(u_j)} \frac{P_1(u_i)}{1 - P_1(u_j)} \qquad (22)$$

Mettendo in evidenza il termine $P_1(u_i)/C$ e sommando e sottraendo $P_1(u_i)$ nel numeratore del primo addendo, si ha:

$$\Pi(u_i) = \frac{P_1(u_i)}{C}\left[\,1 + \frac{P_1(u_i)}{1 - 2\,P_1(u_i)} + \sum_{j \neq i}^{N} \frac{P_1(u_i)}{1 - 2\,P_1(u_j)}\right] \qquad (23)$$

Infine, introducendo il secondo addendo, tra parentesi quadre, nella sommatoria e considerando la relazione (21) abbiamo che:

$$\Pi(u_i) = \frac{P_1(u_i)}{C}\left[\,1 + \sum_{j=1}^{N} \frac{P_1(u_j)}{1 - 2\,P_1(u_j)}\right] = \frac{P_1(u_i)}{C} 2C = 2P_1(u_i) \qquad (24)$$

dalla quale si deduce immediatamente l'estrema semplicità formale della probabilità $\Pi(u_i)$, la qualcosa rende considerevolmente più agevole il calcolo delle $\Pi(u_i)$.

L'illustrazione del metodo di Brewer è stata limitata al caso in cui vengono estratte due sole unità campione; tuttavia, la sua applicazione può essere facilmente estesa al caso $n > 2$, ponendo $\Pi(u_i) = n\,P_1(u_i)$.

Consideriamo, adesso, un altro metodo molto utile nelle situazioni concrete.

In base a tale metodo, che descriveremo nel caso $n > 2$, a ciascuna unità u_i viene assegnato un'intervallo proporzionale alla sua ampiezza A_i mediante le seguenti operazioni:

— si calcolano le quantità ($n\,A_i$) per $i = 1, \ldots, N$
— si considerano i totali cumulati

$$B_i = \sum_{j=1}^{i} n\, A_j \qquad \text{per } i = 1, \ldots, N \qquad (25)$$

— alle unità 1, 2, ..., i, ... si assegnano rispettivamente i seguenti intervalli di scelta:

$$[1 - B_1],\ [(B_1 + 1) - B_2],\ \ldots,\ [(B_{i-1} + 1) - B_i] \ldots \qquad (26)$$

Estratto un numero casuale C compreso tra 1 e $\sum_{i=1}^{N} A_i$, vengono selezionate le n unità i cui intervalli comprendono i numeri

$$C,\ C + \sum_{i=1}^{N} A_i\ ,\ \ldots,\ C + (n - 1) \sum_{i=1}^{N} A_i \qquad (27)$$

Se $n\, P_1 (u_i) < 1$, ovvero $n\, A_i < \sum_i A_i$ (per i = 1, ..., N), ciascuna unità ha una probabilità di inclusione pari a $n\, P_1 (u_i)$ e non può essere selezionata più di una volta.

Campionamento ad uno stadio stratificato

La precedente trattazione sui criteri di selezione è stata svolta nell'ipotesi di campione casuale semplice, ossia nel caso che da una popolazione di N unità venga estratto un campione di n unità.

Nelle indagini concrete si ricorre a disegni di campionamento più complessi di quello casuale semplice e la determinazione delle probabilità d'inclusione per tali indagini risulta conseguentemente più complessa (Russo e Falorsi, 1989).

Nel presente paragrafo illustreremo il calcolo di tali probabilità nel caso di indagini effettuate mediante disegni di campionamento a uno stadio stratificato (Vajani, 1969).

Per tali disegni il criterio di selezione più frequentemente adottato nella pratica delle indagini campionarie è il criterio (sr, pu).

Più precisamente da ciascuno strato si estrae senza reimmissione e probabilità uguali un campione di n_h unità.

Tenendo presente le considerazioni appena svolte e le notazioni simboliche introdotte nel Capitolo 4 relativamente al disegno a uno stadio stratificato, la probabilità di inclusione relativa alla generica unità della popolazione appartenente allo strato h è data dal rapporto:

$$\Pi (u_i) = \frac{n_h}{N_h} \qquad \begin{matrix} (i = 1, \ldots, N_h) \\ (h = 1, \ldots, H) \end{matrix} \qquad (28)$$

Riteniamo utile aggiungere che nella predisposizione di disegni campionari per l'effettuazione di indagini concrete risulta, in alcune situazioni, vantaggioso fissare la dimensione campionaria dei singoli strati in modo che sia rispettata la condizione seguente:

$$\frac{n_h}{N_h} = \frac{n}{N} \qquad (h = 1, \ldots, H) \qquad (29)$$

Tale condizione esprime il fatto che ciascuna unità della popolazione presenta la stessa probabilità di inclusione a prescindere dallo strato di appartenenza (Castellano ed Herzel, 1981).

Campionamento a due stadi con stratificazione delle unità primarie

Illustreremo, ora, il calcolo delle probabilità di inclusione nel caso di indagini basate su disegni di campionamento a due stadi con stratificazione delle unità primarie.

Per tali disegni, nell'ambito di ciascuno strato di unità primarie, si ha un criterio di selezione composito basato su i due seguenti stadi di selezione:

- nel primo, si estrae un campione di unità primarie;
- nel secondo, da ciascuna delle unità primarie campione si procede all'estrazione di un campione di unità secondarie.

I criteri di selezione di uso più frequente nella pratica campionaria sono i seguenti due:

- sia le unità primarie che le unità secondarie vengono estratte senza reimmissione e con probabilità uguali;
- le unità primarie vengono estratte senza reimmissione e probabilità variabili e quelle secondarie senza reimmissione e probabilità uguali.

Relativamente al secondo criterio la situazione più ricorrente è quella in cui si estrae, da ciascuno strato, una o al più due unità primarie.

Ciò premesso, passiamo a descrivere la determinazione delle probabilità di inclusione con riferimento ai due criteri di selezione appena definiti.

Tenendo presente le precedenti considerazioni relative al criterio di selezione (sr, pu) e la simbologia introdotta nel Capitolo 4 per descrivere la struttura del disegno campionario in esame, segue che il rapporto (Russo, 1984a):

$$\Pi_1 (u_{hi}) = \frac{n_h}{N_h} \qquad \begin{matrix}(i = 1, \ldots, N_h)\\(h = 1, \ldots, H)\end{matrix} \qquad (30)$$

esprime la probabilità di inclusione della generica unità primaria i appartenente allo strato h.

In modo analogo, la probabilità di inclusione della generica unità secondaria j dentro l'unità primaria i, è data da:

$$\Pi_2 (u_{hij}) = \frac{m_{hi}}{M_{hi}} \qquad (j = 1, \ldots, M_{hi}) \qquad (31)$$

Pertanto la probabilità di inclusione *finale* della generica unità secondaria j appartenente all'unità primaria i è data dal prodotto della probabilità di inclusione dell'unità primaria i per la probabilità di inclusione della generica unità secondaria j appartenente all'unità primaria i, ossia da:

$$\Pi (u_{hij}) = \frac{n_h}{N_h} \frac{m_{hi}}{M_{hi}} = \Pi_1 (u_{hi})\, \Pi_2 (u_{hij}) \qquad (32)$$

Le precedenti considerazioni possono facilmente essere estese al caso in cui le unità primarie sono selezionate con probabilità variabile senza reimmissione e quelle secondarie con probabilità uguali e senza reimmissione.

Infatti, in questo caso, seguendo Brewer, la probabilità $\Pi_1(u_{hi})$ è espressa da:

$$\Pi_1(u_{hi}) = n_h\, P(u_{hi}) \qquad (33)$$

Conseguentemente la probabilità di inclusione finale dell'unità u_{hij} è data da:

$$\Pi (u_{hij}) = n_h\, P (u_{hi}) \frac{m_{hi}}{M_{hi}} \qquad (34)$$

La (34) può essere particolarizzata al caso in cui le probabilità $P(u_{hi})$ sono proporzionali alle dimensioni delle unità primarie.

A tale scopo, indichiamo con:

— A_{hi}, la dimensione dell'unità primaria i dello strato h;

— $A_h = \sum_{i=1}^{N_h} A_{hi}$

Pertanto la probabilità $P(u_{hi})$ è fornita dal rapporto:

$$P (u_{hi}) = \frac{A_{hi}}{A_h} \qquad (35)$$

A conclusione riteniamo utile scrivere la relazione (34) nei casi particolari, di fondamentale importanza nelle situazioni concrete, di una o due unità primarie campione in ciascuno strato (Istat, 1978).

Dalla (34) seguono immediatamente le due relazioni:

$$\Pi (u_{hij}) = 2\, P (u_{hi}) \frac{m_{hi}}{M_{hi}} \qquad (36)$$

$$\Pi (u_{hij}) = P (u_{hi}) \frac{m_{hi}}{M_{hi}} \qquad (37)$$

CAPITOLO 6 - METODI DI STIMA DIRETTI

Introduzione

Ogni indagine campionaria condotta su larga scala ha, generalmente, la finalità di fornire un elevato numero di stime di parametri della popolazione, che possono essere di tipo differente (frequenze assolute, totali, proporzioni, medie, ecc.).

Poiché, indipendentemente dal metodo di stima adottato, le stime di frequenze assolute, di proporzioni o di medie si possono ricavare da quella di un totale, limiteremo la descrizione soltanto a quest'ultimo tipo di stima.

Il principio su cui è fondato qualsiasi metodo di stima campionaria è quello che il sottoinsieme delle unità della popolazione incluse nel campione deve *rappresentare* anche il sottoinsieme complementare costituito dalle rimanenti unità della popolazione stessa (Statistics Canada, 1976).

Tale principio viene realizzato attribuendo a ciascuna unità inclusa nel campione un peso, che può essere visto come il numero di elementi della popolazione rappresentati da detta unità.

Se, ad esempio, ad una unità campionaria viene attribuito un peso pari a 50, ciò indica che tale unità *rappresenta* se stessa ed altri 49 elementi della popolazione che non sono stati sottoposti ad indagine.

In generale, per ottenere la stima di un totale (ad esempio il reddito totale) si devono eseguire le seguenti tre operazioni:

i) determinare il peso da attribuire a ciascuna unità inclusa nel campione;

ii) moltiplicare il valore relativo ad una data variabile oggetto di indagine, rilevato sulla generica unità inclusa nel campione, per il peso attribuito alla medesima unità (nell'esempio in questione, il reddito di ciascuno individuo campionato moltiplicato per il corrispondente peso);

iii) effettuare la somma dei prodotti di cui al punto ii).

Nelle indagini effettive, generalmente basate su disegni di campionamento complessi, il peso da attribuire a ciascuna unità è ottenuto in base ad una procedura articolata in più passi (Bureau of the Census, 1978; Falorsi, 1989; Russo, 1988a; Russo e Falorsi, 1985):

a) in primo luogo, viene calcolato un peso iniziale, definito *base,* determinato in funzione del disegno di campionamento;

b) successivamente, vengono calcolati alcuni fattori correttivi del peso base, che possono essere distinti in fattori:

— per mancata risposta totale;

— che consentono di rispettare la condizione di uguaglianza

tra alcuni parametri noti della popolazione e le corrispondenti stime campionarie;

c) infine, viene determinato un peso, noto sotto il nome di *peso finale,* espresso come prodotto del peso base per i fattori correttivi.

Nel presente capitolo illustriamo i metodi di stima diretti, così denominati in quanto essi si basano sull'utilizzazione dei parametri strutturali della popolazione e del campione e dei soli valori (delle variabili oggetto di studio) rilevati sulle unità incluse nel campione; tali metodi, che consentono la determinazione di stime corrette, si fondano sull'uso dei pesi base.

Campionamento casuale semplice

La stima diretta di un parametro della popolazione è determinata mediante l'utilizzo dei pesi base.

In generale, per ottenere la stima diretta di un totale riferito ad una data popolazione (ad esempio il reddito totale) si devono eseguire le seguenti tre operazioni:

i) calcolare il peso base per ogni unità inclusa nel campione;

ii) moltiplicare il dato relativo a ciascuna unità facente parte del campione (nell'esempio in questione il reddito di ciascun individuo campionato) per il peso base corrispondente;

iii) effettuare la somma dei prodotti ottenuti al punto ii).

Relativamente al punto i), il peso base di una unità facente parte del campione viene calcolato come reciproco della probabilità di inclusione di tale unità. Ricordiamo che il termine *probabilità di inclusione* indica la probabilità, calcolata sulla base del disegno di campionamento, che un elemento della popolazione sia incluso fra gli elementi del campione (Cochran, 1977; Horvitz e Thompson, 1952).

Le stime campionarie dirette, ottenute mediante il metodo precedentemente descritto, sono stime corrette nel senso che la media delle stime campionarie nell'universo dei campioni è uguale al totale della popolazione.

Per illustrare quanto detto consideriamo una popolazione finita di N elementi a ciascuno dei quali è associato il valore del carattere y:

$$Y_1, Y_2, \ldots, Y_i, \ldots, Y_N$$

in cui Y_i (i = 1, ..., N) esprime il valore assunto dal carattere y sull'elemento i della popolazione.

Il totale del carattere y è dato da

$$Y = \sum_{i=1}^{N} Y_i \qquad (1)$$

Consideriamo ora un campione formato da n elementi, estratti senza reimmissione e con probabilità uguali dalle N unità costituenti la popolazione. Esso presenta come risultato gli n valori:

$$Y_1, Y_2, \ldots, Y_i, \ldots, Y_n$$

Tenendo presenti le caratteristiche del meccanismo di selezione, la probabilità di inclusione di ciascuna delle n unità campione è espressa da:

$$\Pi = \frac{n}{N} \qquad (2)$$

di conseguenza il peso base è definito da:

$$K = \frac{N}{n} = \frac{1}{\Pi} \qquad (3)$$

Pertanto la stima campionaria diretta del parametro Y, definito dalla (1), è data da:

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} Y_i = \frac{1}{\Pi} \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} K\, Y_i \qquad (4)$$

Consideriamo, come esempio, una popolazione costituita da N = 6 unità, per le quali il carattere y oggetto di indagine assume i valori:

$$Y_1 = 6;\ Y_2 = 5;\ Y_3 = 10;\ Y_4 = 7;\ Y_5 = 8;\ Y_6 = 2$$

Il totale Y è pari a:

$$6 + 5 + 10 + 7 + 8 + 2 = 38$$

Supponiamo di aver estratto, da detta popolazione, un campione di n = 2 elementi, selezionati senza reimmissione e con probabilità uguali. Supponiamo, inoltre, di avere selezionato il campione identificato dalla coppia:

$$(Y_1 = 6;\ Y_3 = 10)$$

La probabilità di inclusione Π è:

$$\Pi = \frac{2}{6} = 0{,}33$$

Conseguentemente, il peso base K risulta uguale a:

$$K = \frac{1}{\Pi} = 3$$

La stima campionaria $\hat{Y}$ è data da:

$$\hat{Y} = (3 \times 6) + (3 \times 10) = 48$$

Campionamento ad uno stadio stratificato

Indichiamo con y il generico carattere oggetto di indagine e supponiamo di voler determinare, mediante un campionamento ad uno stadio stratificato, la stima diretta del totale del carattere y della popolazione.

Tenendo presenti le notazioni simboliche introdotte nel Capitolo 4, ricordiamo che tale parametro è definito da:

$$Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} Y_{hi} = \sum_{h=1}^{H} Y_h \qquad (5)$$

in cui si è posto:

$$Y_h = \sum_{i=1}^{N_h} Y_{hi} \qquad (6)$$

Ciascuna delle n_h unità campione dello strato h viene inclusa nel campione con probabilità uguale a (Cochran, 1977; Yamane, 1967):

$$\Pi_h = \frac{n_h}{N_h} \qquad (7)$$

In base a quanto esposto nel caso del campionamento semplice, il peso base da attribuire a ciascuna delle n_h unità è pertanto uguale a:

$$K_h = \frac{1}{\Pi_h} = \frac{N_h}{n_h} \qquad (8)$$

La stima campionaria diretta del totale Y_h è fornita dall'espressione:

$$\hat{Y}_h = \sum_{i=1}^{n_h} \frac{N_h}{n_h} Y_{hi} = \sum_{i=1}^{n_h} K_h Y_{hi} \qquad (9)$$

Conseguentemente la stima diretta del totale Y, definito dalla (5) è data da (Russo, 1982):

$$\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h = \sum_{h=1}^{H} \sum_{i=1}^{n_h} K_h Y_{hi} \qquad (10)$$

Campionamento a due stadi con stratificazione delle unità primarie

Supponiamo di voler determinare, mediante un campione a due stadi con stratificazione delle unità di primo stadio, la stima diretta del totale del carattere y nella popolazione.

Tenendo presenti le notazioni simboliche introdotte nel Capitolo 4, ricordiamo che tale parametro è definito da:

$$Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{hij} = \sum_{h=1}^{H} \sum_{i=1}^{N_h} Y_{hi} = \sum_{h=1}^{H} Y_h \qquad (11)$$

in cui si è posto:

$$Y_h = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{hij} = \sum_{i=1}^{N_h} Y_{hi} \qquad (12)$$

$$Y_{hi} = \sum_{j=1}^{M_{hi}} Y_{hij} \qquad (13)$$

Ciascuna delle n_h unità primarie dello strato h viene inclusa nel campione con probabilità Π_{1hi} ($i = 1, \ldots, n_h$).

Nel caso in cui le n_h unità primarie siano estratte con probabilità uguale, la probabilità Π_{1hi} è data da (Hansen, Hurwitz e Madow, 1953):

$$\Pi_{1hi} = \frac{n_h}{N_h} \qquad (14)$$

Se invece le n_h unità primarie sono estratte con probabilità proporzionale alla *dimensione* delle unità stesse, la probabilità Π_{1hi} è espressa da (Kish, 1965):

$$\Pi_{1hi} = n_h \frac{A_{hi}}{\sum_{i=1}^{N_h} A_{hi}} \qquad (15)$$

in cui A_{hi} rappresenta il totale della variabile *a*, utilizzata per definire la dimensione, dell'unità primaria i dello strato h.

In ognuna delle n_h unità primarie campione dello strato h vengono estratte m_{hi} unità secondarie con probabilità pari a:

$$\Pi_{2hi} = \frac{m_{hi}}{M_{hi}} \qquad (16)$$

La probabilità di inclusione della generica unità secondaria j, selezionata nella unità primaria campione i dello strato h, è pertanto definita da:

$$\Pi_{hi} = \Pi_{1hi}\,\Pi_{2hi} \qquad (17)$$

Conseguentemente, il peso base attribuito alla medesima unità è dato da:

$$K_{hi} = \frac{1}{\Pi_{hi}} \qquad (18)$$

La stima campionaria diretta del parametro Y_h è ottenuta mediante l'espressione seguente:

$$\hat{Y}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \frac{1}{\Pi_{hi}} Y_{hij} = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi}\, Y_{hij} \qquad (19)$$

La (19) può porsi nella forma equivalente:

$$\hat{Y}_h = \sum_{i=1}^{n_h} \frac{\hat{Y}_{hi}}{\Pi_{1hi}} \qquad (20)$$

in cui:

$$\hat{Y}_{hi} = \frac{1}{\Pi_{2hi}} \sum_{j=1}^{m_{hi}} Y_{hij} = \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} Y_{hij} \qquad (21)$$

rappresenta la stima diretta del totale Y_{hi} del carattere y nell'unità primaria i dello strato h.

La stima diretta del parametro Y è definita dalla seguente espressione (Russo, 1988b):

$$\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{\hat{Y}_{hi}}{\Pi_{1hi}} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi}\, Y_{hij} \qquad (22)$$

Campioni auto-ponderanti

Le indagini campionarie condotte su larga scala sono, in genere, basate su *campioni autoponderanti* in cui le unità della popolazione hanno la medesima probabilità di essere incluse nel campione (Cochran, 1977).

Un campione casuale semplice è naturalmente autoponderante, in quanto, come risulta dalle formule (2) e (3), tutte le unità della popolazione hanno una probabilità di inclusione uguale a $\Pi = n/N$. Conseguentemente tutte le unità incluse nel campione hanno un peso costante pari a $K = 1/\Pi = N/n$.

Nel caso di un disegno ad uno stadio stratificato, la realizzazione di un campione autoponderante implica che tutte le unità della popolazione abbiano una probabilità di inclusione costante, a prescindere dallo strato di appartenenza, ossia:

$$\Pi_1 = \cdots = \Pi_h = \cdots = \Pi_H = \overline{\Pi}_h \qquad (23)$$

Esprimendo le probabilità di inclusione in base alla relazione (7) abbiamo che:

$$\frac{n_1}{N_1} = \cdots = \frac{n_h}{N_h} = \cdots = \frac{n_H}{N_H} = \frac{n}{N} \qquad (24)$$

in cui: $N = \sum_{h=1}^{H} N_h$ e $n = \sum_{h=1}^{H} n_h$

La relazione (24) esprime la condizione che il numero di unità campionarie in ciascuno strato risulti proporzionale alla numerosità N_h.

Il peso base relativo alle n_h unità campionarie dello strato h (h = 1,...,H) è dato da:

$$K_h = \frac{1}{\Pi_h} = \frac{N}{n} = \overline{K}_h \qquad (25)$$

Pertanto la stima diretta del totale Y, definita dalla (10), può essere riformulata nel seguente modo:

$$\hat{Y} = \overline{K}_h \sum_{h=1}^{H} \sum_{i=1}^{n_h} Y_{hi} = \frac{N}{n} \sum_{h=1}^{H} \sum_{i=1}^{n_h} Y_{hi} \qquad (26)$$

Dalla precedente relazione si desume che la stima diretta di un campione ad uno stadio stratificato autoponderante viene ottenuta come se le n unità campionarie fossero state estratte mediante un campionamento casuale semplice.

Infine, nel caso di un disegno a due stadi con stratificazione delle unità primarie, l'autoponderazione del campione comporta che tutte le unità della popolazione abbiano la medesima probabilità di inclusione, a prescindere dallo strato e dalla unità primaria di appartenenza.

Esaminiamo dapprima il caso in cui le unità primaria siano estratte con probabilità uguali. Tenendo presenti le relazioni (14), (16) e (17) si deduce che:

$$\Pi_{hi} = \frac{n_h}{N_h} \frac{m_{hi}}{M_{hi}} = \overline{\Pi}_{hi} = \frac{m}{M} \qquad (27)$$

in cui:

$$M = \sum_{h=1}^{H} \sum_{i=1}^{N_h} M_{hi} \quad e \quad m = \sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi} \qquad (28)$$

Esaminiamo ora il caso in cui le unità primarie siano estratte con probabilità proporzionali alle dimensioni delle unità stesse. Tenendo presenti le relazioni (15) (16) e (17) abbiamo che:

$$\Pi_{hi} = n_h \frac{A_{hi}}{A_h} \frac{m_{hi}}{M_{hi}} = \overline{\Pi}_{hi} = \frac{m}{M} \qquad (29)$$

Pertanto, sia nel caso in cui le unità primarie siano estratte con probabilità uguali che nel caso in cui le unità primarie siano estratte con probabilità proporzionale alla dimensione, il peso base relativo alle m_{hi} unità campionarie selezionate nell'unità primaria i appartenente allo strato h è espresso da:

$$K_{hi} = \frac{1}{\Pi_{hi}} = \frac{M}{m} = \overline{K}_{hi} \qquad (30)$$

Quindi la stima diretta del totale Y, definita dalla (22), può essere riformulata nel seguente modo:

$$\hat{Y} = \overline{K}_{hi} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} Y_{hij} = \frac{M}{m} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} Y_{hij} \qquad (31)$$

Riassumendo le precedenti considerazioni si osserva che se l'indagine ha la finalità di fornire soltanto stime corrette, così come sono quelle descritte nel presente capitolo, l'utilizzazione di campioni autoponderanti, che attribuiscono alle unità campionarie il medesimo peso, comporta una notevole riduzione dei tempi necessari per il calcolo delle stime e dei corrispondenti errori di campionamento.

Inoltre, l'introduzione di pesi costanti conduce a stime che generalmente, a parità di condizioni, risultano più efficienti di quelle ottenibili attraverso disegni basati su pesi variabili.

CAPITOLO 7 - FATTORI CORRETTIVI PER MANCATA RISPOSTA TOTALE

Introduzione

La situazione in cui non viene effettuata la rilevazione su un'unità inclusa nel campione è nota con la denominazione di *mancata risposta totale*. Sotto il nome di *mancata risposta parziale* viene invece indicata la situazione in cui, relativamente ad una o più unità finali, non vengono rilevati i valori di alcune delle caratteristiche oggetto di studio.

Le mancate risposte parziali o totali, introducono nelle stime effetti distorsivi che possono essere attenuati mediante l'utilizzazione di teniche speciali. Inoltre, la mancata risposta, riducendo la dimensione del campione, comporta un aumento della varianza di campionamento e di conseguenza una minore precisione delle stime oggetto di indagine.

Nel presente capitolo illustreremo alcuni criteri da adottare nel caso in cui si verifichi il fenomeno della mancata risposta totale (Cochran, 1977; Kish, 1965; Little, 1983; Platek e Gray, 1983; Politz e Simmons, 1949; Statistics Canada, 1976).

Campionamento casuale semplice

Un metodo frequentemente adottato consiste nel sostituire le unità non rispondenti con unità estratte da una lista, comunemente nota sotto il nome di *Elenco suppletivo*.

Tale criterio, conosciuto come *metodo di sostituzione*, presenta le seguenti caratteristiche:

- il numero di unità rispondenti coincide, in genere, con l'ampiezza teorica del campione; di conseguenza la precisione delle stime oggetto di indagine risulta uguale a quella prevista dal disegno di campionamento;
- può introdurre distorsione nel caso in cui le caratteristiche oggetto di indagine dei non rispondenti siano diverse da quelle delle unità sostitutive.

Per illustrare il metodo delle sostituzioni, consideriamo una popolazione finita composta di N elementi, a ciascuno dei quali è associato il valore del carattere y:

$$Y_1, Y_2, \ldots, Y_i, \ldots, Y_N$$

Da detta popolazione viene selezionato un campione di n elementi estratti senza reimmissione e con probabilità uguali.

Indichiamo con n_{1r} il numero di unità rispondenti alla prima intervista; di conseguenza il numero di unità non rispondenti è uguale a $\bar{n}_{1r} = n - n_{1r}$.

Il metodo delle sostituzioni consiste nell'estrarre $\bar{n}_{1r}$ unità aggiuntive dall'elenco suppletivo.

Supponiamo inoltre che il numero di unità rispondenti, fra le $\bar{n}_{1r}$, sia pari a n_{2r}; conseguentemente il numero di unità non rispondenti alla seconda intervista è uguale a $\bar{n}_{2r} = \bar{n}_{1r} - n_{2r}$.

In tale circostanza può essere effettuata una ulteriore selezione di $\bar{n}_{2r}$ unità dall'elenco suppletivo, da intervistare al posto delle unità non-rispondenti alla seconda intervista.

Il procedimento di sostituzione in esame può essere iterato fino ad ottenere un numero di unità rispondenti pari ad n.

Tuttavia, poiché in alcune situazioni concrete non si raggiunge un numero di unità rispondenti uguale ad n, la trattazione successiva sarà sviluppata sotto l'ipotesi di avere un numero di unità rispondenti pari ad ${}_rn$, con ${}_rn \leq n$.

Si possono pertanto presentare due situazioni:

${}_rn = n$

${}_rn < n$

Nel primo caso, in cui il numero di unità rispondenti è pari all'ampiezza teorica del campione, il peso base delle unità rispondenti rimane uguale a:

$$K = \frac{N}{n}$$

e le stime dirette possono ottenersi adottando la procedura già descritta nel Capitolo 6.

Nel secondo caso il peso base viene in genere corretto mediante un fattore moltiplicativo.

I fattori correttivi possono essere ottenuti adottando metodologie diverse. Nel presente capitolo ci limiteremo, tuttavia, alla descrizione di un criterio che, pur non essendo molto raffinato, viene comunemente adottato nelle indagini su larga scala per la sua semplicità computazionale.

Riprendendo quanto prima illustrato, il fattore correttivo del peso base, ottenuto dall'applicazione del criterio in oggetto, è dato da:

$$W = \frac{n}{{}_rn} \qquad (1)$$

Pertanto il peso base corretto viene ottenuto come:

$${}_cK = K\ W = \frac{N\,n}{n\ {}_rn} = \frac{N}{{}_rn} \qquad (2)$$

Di conseguenza la stima campionaria diretta nel parametro $Y = \sum_{i=1}^{N} Y_i$, corretta per mancata risposta totale, è espressa da:

$$\hat{Y} = \sum_{i=1}^{{}_rn} {}_cK\ Y_i = \frac{N}{{}_rn} \sum_{i=1}^{{}_rn} Y_i \qquad (3)$$

Campionamento ad uno stadio stratificato

Riprendiamo quanto illustrato nel Capitolo 6 a proposito del campionamento ad uno stadio stratificato ed indichiamo con ${}_rn_h$ il numero di unità rispondenti nel generico strato h (eventualmente comprensivo delle sostituzioni).

Si possono verificare i due seguenti casi:

— ${}_rn_h = n_h$

— ${}_rn_h < n_h$

Nel prosieguo limiteremo la trattazione al secondo caso, in quanto nel primo non è necessaria l'introduzione di un fattore correttivo.

Con riferimento al caso in cui ${}_rn_h < n_h$ si possono presentare le due seguenti situazioni:

${}_rn_h > o$
${}_rn_h = o$

Esaminiamo dapprima il caso in cui ${}_rn_h > o$. In tale circostanza, il fattore correttivo per mancata risposta totale relativo allo strato h è dato da:

$$W_h = \frac{n_h}{{}_rn_h} \qquad (4)$$

Pertanto il peso base corretto per mancata risposta totale è definito da:

$${}_cK_h = K_h\ W_h = \frac{N_h}{{}_rn_h} \qquad (5)$$

Di conseguenza la stima diretta del totale Y è fornita dall'espressione:

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{{}_rn_h} {}_cK_h \quad Y_{hi}$$

Passiamo ora ad esaminare il caso in cui ${}_rn_h = o$, ossia la situazione in cui non sia stata effettuata la rilevazione sulle unità campionarie appartenenti allo strato h.

Il criterio che viene comunemente adottato è quello di far *rappresentare* la popolazione dello strato h dalle unità rispondenti di un altro strato. Tale criterio consiste, sostanzialmente, nell'aggregare i due strati in questione in un solo strato e nel definire un fattore correttivo del peso base dello strato in cui è stata effettuata la rilevazione.

Per illustrare il criterio in questione indichiamo u lo strato in cui non è stata effettuata la rilevazione (${}_rn_u = o$)e con v lo strato al quale si è deciso di aggregare lo strato u (ovviamente sarà ${}_rn_v > o$).

Indichiamo inoltre con g lo strato formato dall'unione degli strati u e v, e con G il numero di strati risultante dopo l'operazione di aggregazione degli strati.

Per determinare il peso base corretto per mancata risposta totale occorre calcolare i due seguenti fattori correttivi:

$$- \quad W_v = \frac{n_v}{{}_rn_v} \tag{6}$$

$$- \quad W'_g = \frac{N_u + N_v}{N_v} \tag{7}$$

che consentono di definire un fattore complessivo espresso da:

$$W_g = W_v \, W'_g = \frac{N_u + N_v}{N_v \, {}_rn_v} n_v \tag{8}$$

Conseguentemente il peso base corretto per mancata risposta con riferimento allo strato g è dato da:

$${}_cK_g = K_v \, W_g = \frac{N_u + N_v}{{}_rn_v} \tag{9}$$

Pertanto la stima diretta del totale Y è data da:

$$\hat{Y} = \sum_{g=1}^{G} \sum_{i=1}^{{}_rn_g} {}_cK_g \quad Y_{gi} \tag{10}$$

Al fine di chiarire il significato della relazione (10) ricordiamo che l'indice g può indicare alternativamente uno strato elementare oppure uno strato risultante dall'unione di due strati elementari.

Campionamento a due stadi con stratificazione delle unità primarie

In un disegno a due stadi si presentano le due seguenti situazioni di mancata risposta:

- in una o più unità primarie non viene effettuata la rilevazione. In tale circostanza si perdono, ovviamente, tutte le informazioni sulle caratteristiche oggetto di studio relative alle unità finali che dovevano essere rilevate nell'ambito di tali unità primarie. Indichiamo tale situazione con il termine di *mancata risposta di unità primaria.*
- in una o più unità primarie non viene effettuata la rilevazione per alcune delle unità secondarie incluse nel campione. Denominiamo tale situazione con *mancata risposta di unità secondaria.*

Esaminiamo dapprima la mancata risposta di unità primaria ed indichiamo con ${}_rn_h$ il numero di unità primarie rispondenti (comprensivo anche delle sostituzioni) nel generico strato h (h = 1,2,...,H).

Si possono presentare le due situazioni:

${}_rn_h > 0$

${}_rn_h = 0$

Nel primo caso il fattore correttivo, per mancata risposta, per il generico strato h è dato da:

$$W_h = \frac{n_h}{{}_rn_h} \tag{11}$$

Nel secondo caso in cui ${}_rn_h = 0$, non è stata effettuata la rilevazione sulle unità campionarie dello strato h. Come già illustrato precedentemente il criterio che viene comunemente adottato

è quello di far rappresentare la popolazione dello strato h dalle unità rispondenti di un altro strato. Tale criterio consiste nell'aggregare i due strati in questione in un solo strato e nel definire un fattore correttivo del peso base dello strato in cui è stata effettuata la rilevazione.

Per descrivere il criterio in esame, indichiamo con u lo strato in cui non è stata eseguita la rilevazione e con v lo strato al quale si è deciso di aggregare lo strato u (ovviamente sarà ${}_rn_v > 0$). Indichiamo, inoltre, con g lo strato formato dall'unione degli strati u e v, e con G il numero di strati che si ottiene dopo l'operazione di aggregazione degli strati mancanti.

Per determinare il peso base corretto per mancata risposta totale occorre calcolare due fattori correttivi.

Il primo fattore è dato da:

$$W_v = \frac{n_v}{{}_rn_v} \qquad (12)$$

Il secondo fattore correttivo è espresso da uno dei due seguenti rapporti:

$$- \quad W'_g = \frac{N_u + N_v}{N_v} \qquad (13)$$

$$- \quad W'_g = \frac{A_u + A_v}{A_v} \qquad (14)$$

dove il primo rapporto si riferisce al caso in cui le unità primarie sono estratte con probabilità uguale, mentre il secondo si riferisce al caso in cui le unità primarie sono estratte con probabilità proporzionale alla dimensione. Riteniamo opportuno ricordare che N_u ed A_u indicano, con riferimento allo strato u, rispettivamente il numero di unità primarie ed il totale della variabile «a» che rappresenta la dimensione del medesimo strato; analogo significato hanno N_v e A_v.

Esaminiamo ora il problema della mancata risposta di unità secondaria ed indichiamo con ${}_rm_{hi}$ il numero di unità secondarie rispondenti nell'unità primaria i dello strato h. In tale circostanza si calcola un fattore correttivo definito da (Russo, 1984b; Russo, 1988a):

$$W_{hi} = \frac{m_{hi}}{{}_rm_{hi}} \qquad (15)$$

In base a quanto sopra illustrato segue che il peso base corretto per mancata risposta è uguale a:

$$- \quad {}_cK_{hi} = K_{hi}\, W_h\, W_{hi} \qquad (16)$$

oppure a:

$$- \quad {}_cK_{gi} = K_{vi}\, W_v\, W'_g\, W_{vi} \qquad (17)$$

che si riferiscono rispettivamente al primo e al secondo caso di mancata risposta di unità primaria.

Nella (16) i fattori W_h e W_{hi} sono espressi rispettivamente dalla (11) e (15), mentre il peso K_{hi} è definito dall'espressione (18) del Capitolo 6. Nella (17) i fattori W_v e W'_g sono definiti dalle formule (12), (13) o (14); W_{vi} e K_{vi} sono espressi da relazioni analoghe alla (15) ed alla (18) dal Capitolo precedente, salvo la sostituzione dell'indice h con indice v.

In definitiva, tenendo presente la (16) e la (17), la stima del totale Y è data da:

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{{}_rn_h} \sum_{j=1}^{{}_rm_{hi}} {}_cK_{hi}\; Y_{hij} \qquad (18)$$

oppure da:

$$\hat{Y} = \sum_{g=1}^{G} \sum_{i=1}^{{}_rn_g} \sum_{j=1}^{{}_rm_{gi}} {}_cK_{gi}\; Y_{gij} \qquad (19)$$

Allo scopo di chiarire le precedenti relazioni sottolineamo che la (18) è la formula da usare nel caso in cui in ciascuno strato h ($h = 1, \ldots, H$) si sia verificato ${}_rn_h > o$, mentre la (19) è la formula da utilizzare nel caso in cui in qualcuno degli H strati si sia verificato ${}_rn_h = o$. Pertanto nella (19) l'indice g di strato può indicare alternativamente uno strato elementare oppure uno strato risultante dall'unione di due strati elementari.

CAPITOLO 8 - METODI DI STIMA INDIRETTI: GLI STIMATORI DEL RAPPORTO

Introduzione

Generalmente, nella fase preparatoria di un'indagine campionaria si dispone di informazioni riguardanti uno o più caratteri.

Tali informazioni, note sotto il nome di *informazioni ausiliarie,* possono essere usate nel processo di stratificazione, nell'attribuzione delle probabilità di selezione e nei metodi di stima delle caratteristiche della popolazione oggetto d'indagine.

Dell'utilizzazione delle informazioni ausiliarie ai fini della statificazione e delle probabilità di selezione ci siamo già occupati nei precedenti capitoli 2 e 3; in questo e nel successivo Capitolo 9, invece, parleremo di alcuni metodi di stima fondati appunto sull'uso delle suddette informazioni.

L'opportunità di ricorrere all'impiego di tali metodi - denominati metodi di stima indiretti — sta nella maggiore precisione che in generale si consegue, a parità di ogni altra condizione, rispetto a quella ottenibile mediante i metodi di stima diretti descritti nel Capitolo 6 (De Cristofaro, 1979).

Attualmente i procedimenti di stima indiretti in uso sono due: quello fondato sul metodo della regressione e quella basato sul metodo del rapporto (con e senza post-stratificazione).

Nel presente volume ci occuperemo del secondo metodo del quale illustreremo la struttura fondamentale e le caratteristiche essenziali limitatamente — come abbiamo più volte detto — al campionamento casuale semplice, ad uno stadio stratificato e a due stadi con stratificazione delle unità primarie.

Questo metodo, da tempo noto e frequentemente applicato sia all'Istat sia presso i più grossi centri di informazione statistica a livello internazionale, è da preferire al metodo fondato sulla regressione, in quanto quest'ultimo è di più difficile realizzazione pratica soprattutto nelle indagini condotte su larga scala.

Campionamento casuale semplice

In parecchie circostanze si osserva che il rapporto di due grandezze presenta una variabilità minore di quella delle singole grandezze: su questa semplice osservazione è fondato il metodo del rapporto (Cochran, 1977).

Il principio alla base del metodo è il seguente: consideriamo una popolazione finita di N elementi, cui siano associate le N coppie di valori dei due caratteri y (oggetto d'indagine) ed x (ausiliario):

(Y_1, X_1), ..., (Y_i, X_i), ..., (Y_N, X_N)

in cui i simboli Y_i ed X_i (i = 1, ..., N) esprimono i valori assunti rispettivamente da y ed x sull'unità i, che nel seguito si riterranno tutti positivi.

Siano poi:

$$Y = \sum_{i=1}^{N} Y_i \qquad (1)$$

e

$$X = \sum_{i=1}^{N} X_i \qquad (2)$$

i corrispondenti valori complessivi.

Consideriamo ora un campione costituito da n elementi, estratti senza reimmissione e con probabilità uguali dalle N unità costituenti la popolazione; esso presenta, in questo caso, come risultati n coppie:

(Y_1, X_1), ..., (Y_i, X_i), ..., (Y_n, X_n)

Ciò posto, supponiamo di voler stimare il totale del carattere y, definito dalla (1).

Si può sempre ignorare il carattere x e prendere come stima del totale Y la quantità:

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} Y_i \qquad (3)$$

che definisce la stima diretta di Y.

In alternativa alla (3) consideriamo invece la stima $\hat{\hat{Y}}$ definita dall'espressione:

$$\hat{\hat{Y}} = \frac{\hat{Y}}{\hat{X}} X \qquad (4)$$

in cui:

$$\hat{X} = \frac{N}{n} \sum_{i=1}^{n} X_i \qquad (5)$$

rappresenta la stima diretta del totale del carattere x, espresso dalla (2).

La (4) definisce la stima del totale Y ottenuta con il metodo del rapporto, così denominato in quanto si basa sul rapporto di due stime. È evidente, inoltre, che per utilizzare la (4) occorre che sia noto il totale del carattere x nella popolazione, che costituisce l'*informazione ausiliaria esterna*.

La (4) può anche riscriversi in una forma molto interessante ai fini pratici del calcolo delle stime.

In primo luogo, la (4) può porsi nella forma:

$$\hat{\hat{Y}} = \frac{\frac{N}{n} \sum_{i=1}^{n} Y_i}{\hat{X}} X = \frac{\sum_{i=1}^{n} K\, Y_i}{\hat{X}} X \qquad (6)$$

in cui:

$$K = \frac{N}{n} \qquad (7)$$

indica il peso base.

Posto inoltre:

$$B = \frac{X}{\hat{X}} \qquad (8)$$

la (6) può scriversi:

$$\hat{\hat{Y}} = \sum_{i=1}^{n} K\, B\, Y_i = \sum_{i=1}^{n} {}_fK\, Y_i \qquad (9)$$

in cui:

$${}_fK = K\, B \qquad (10)$$

rappresenta il *peso finale* da attribuire a ciascuna delle n unità incluse nel campione.

Introducendo nella (10) la (5) e la (7), il peso ${}_fK$ assume la forma abbastanza semplice;

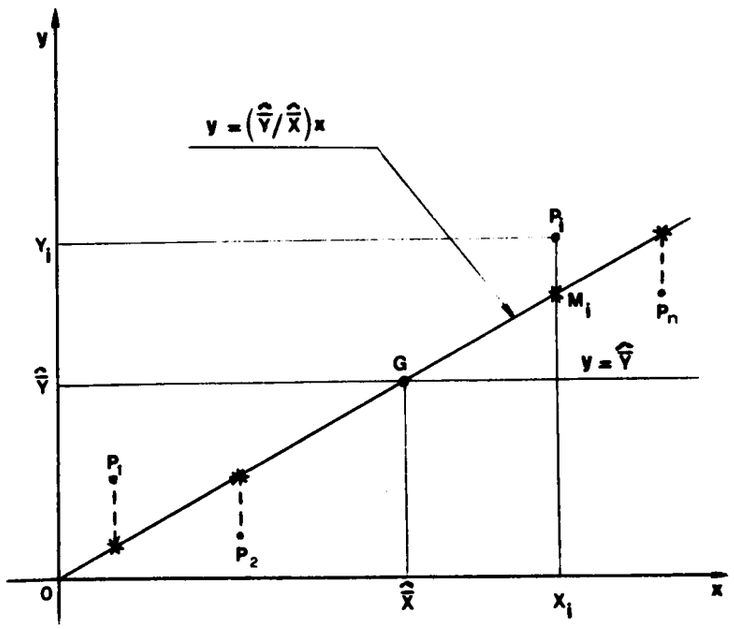
$$ {}_fK = \frac{X}{\sum_{i=1}^{n} X_i} \qquad (11) $$

La teoria del metodo del rapporto si basa su dimostrazioni algebricamente abbastanza complesse; non si può, nè si intende, fare qui dettagliata ed esauriente esposizione teorica del metodo in esame, per la qual cosa il lettore desideroso di più ampi ragguagli potrà consultare la letteratura in materia (Yamane, 1967).

Riteniamo tuttavia utile illustrarne le principali proprietà.

In primo luogo alcune semplici considerazioni di ordine geometrico che consentiranno di chiarire meglio il discorso.

In un sistema di riferimento ortogonale dei punti del piano si individuino i punti di coordinate (Y_i, X_i); il campione si rappresenta allora come una nuvola di punti (Y_i, X_i), i = 1,...,n. La figura sotto riportata, nella quale abbiamo posto $P_i = (X_i, Y_i)$, illustra tale situazione.



Consideriamo ora la retta passante per l'origine O ed il baricentro della nuvola G $(\hat{\bar{X}}, \hat{\bar{Y}})$; l'equazione di tale retta è:

$$ y = \frac{\hat{\bar{Y}}}{\hat{\bar{X}}}\, x \qquad (12) $$

in cui:

$$ \hat{\bar{Y}} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \text{e} \qquad \hat{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (13) $$

che rappresentano rispettivamente le stime corrette di $\bar{Y} = Y/N$ e $\bar{X} = X/N$.

Osserviamo intanto che quando x assume il valore X, totale del carattere ausiliario x, l'ordinata corrispondente, fornita dalla (12), coincide con $\hat{Y}$, definito attraverso la (4).

Infatti, la (4) può scriversi nella forma equivalente:

$$ \hat{Y} = \frac{\hat{\bar{Y}}}{\hat{\bar{X}}}\, X \qquad (14) $$

che coincide con la (12), quando in quest'ultima si pone x = X.

Ora, come vedremo meglio in seguito (Capitoli 10 e 11), la stima della varianza della stima diretta $\hat{Y}$ è fornita dalla espressione:

$$ \hat{V}(\hat{Y}) = N\,\frac{N-n}{n(n-1)}\sum_{i=1}^{n}(Y_i - \hat{\bar{Y}})^2 \qquad (15) $$

e la stima della varianza della stima rapporto $\hat{\bar{Y}}$ da:

$$ \hat{V}(\hat{\bar{Y}}) = N\,\frac{N-n}{n(n-1)}\sum_{i=1}^{n}\left(Y_i - \frac{\hat{\bar{Y}}}{\hat{\bar{X}}}\,X_i\right)^2 \qquad (16) $$

Dalla figura si osserva che la differenza

$$ (Y_i - \hat{\bar{Y}}) \qquad (17) $$

è rappresentata dal segmento $P_i H_i$, e la differenza:

$$(Y_i - \frac{\hat{\bar{Y}}}{\hat{\bar{X}}} X_i) \qquad (18)$$

dal segmento $P_i M_i$. Ciò mostra che la varianza $\hat{V}(\hat{\bar{Y}})$ è più piccola della varianza $\hat{V}(\hat{Y})$ quando la nuvola dei punti (X_i, Y_i), $i = 1, \ldots, n$, si dispone più attorno la retta passante per i punti O e G che alla retta passante per G e parallela all'asse delle ascisse: cioè, quando le variabili x ed y sono fortemente correlate.

È possibile mostrare (Zanella, 1974) che la maggiore efficienza della stima $\hat{\bar{Y}}$ rispetto alla stima $\hat{Y}$ è assicurata quando, sotto la solita condizione $n > 30 \div 40$, risulta:

$$\rho > \frac{1}{2} \frac{\frac{S_x}{\bar{X}}}{\frac{S_y}{\bar{Y}}} = \frac{1}{2} \frac{\text{coefficiente di variazione di x}}{\text{coefficiente di variazione di y}} \qquad (19)$$

dove ρ indica il coefficiente di correlazione fra le componenti della variabile doppia (x, y), essendo inoltre:

$$S_x^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2; \quad S_y^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \qquad (20)$$

La condizione (19) mostra come la maggiore efficienza della stima $\hat{\bar{Y}}$ rispetto alla stima $\hat{Y}$ possa essere assicurata da una *notevole* correlazione fra i due caratteri x ed y e/o dall'essere il coefficiente di variazione di x *piccolo* rispetto a quello di y: ciò che si verifica, in altre parole, quando la variabilità dei valori di x, riferiti al proprio ordine di grandezza, è poco rilevante rispetto alla variabilità dei valori di y.

Una seconda proprietà della stima $\hat{\bar{Y}}$ è legata al fatto che è possibile dimostrare (Sukhatme e Sukhatme, 1970) le due seguenti relazioni:

$$B(\hat{\bar{Y}}) = E(\hat{\bar{Y}}) - Y \doteq \frac{N-n}{N} \frac{1}{n} \left[\frac{S_x^2}{\bar{X}^2} - \frac{S_{xy}}{\bar{X}\bar{Y}} \right] \qquad (21)$$

$$B(\hat{Y}) = E(\hat{Y}) - Y = 0 \qquad (22)$$

in cui nella (21) S_x^2 è definita dalla prima espressione delle (20) e S_{xy} è espressa da:

$$S_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$$

Si dice, allora, che la $\hat{Y}$ fornisce una stima corretta, nel senso che, per la (22), il valore medio di tutte le stime che si possono ottenere con un campione di determinata ampiezza coincide con il valore Y oggetto di stima, mentre la stima $\hat{\bar{Y}}$ introduce una distorsione pari alla (21). Inoltre, mentre la (22) è una espressione esatta, la (21) è approssimata a meno di termini di grado non inferiore a $1/n^2$.

Osservazioni non prive di qualche interesse si possono fare sulla distorsione che il metodo del rapporto comporta e sulla bontà dell'approssimazione raggiunta con la (21).

In primo luogo, dalla (21) si deduce che la distorsione della stima $\hat{\bar{Y}}$ decresce al crescere della dimensione del campione, annullandosi per $n = N$.

Dalla (21) si deduce anche che la distorsione di $\hat{\bar{Y}}$ decresce in valore assoluto col ridursi della differenza assoluta tra i termini compresi tra le parentesi quadre e si annulla quando:

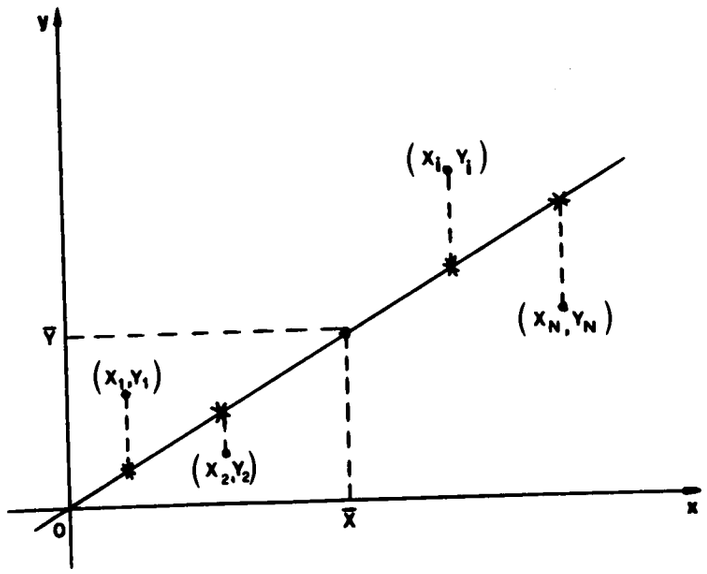
$$\frac{S_x^2}{n} \frac{1}{\bar{X}^2} - \frac{S_{xy}}{n} \frac{1}{\bar{X}\bar{Y}} = 0 \qquad (23)$$

che può riscriversi nella forma:

$$\frac{S_{xy}}{S_x^2} = \frac{\bar{Y}}{\bar{X}} \qquad (24)$$

La relazione (24) consente pertanto di concludere che la distorsione di $\hat{\bar{Y}}$ è nulla quando la retta di regressione di y rispetto ad x passa per l'origine del sistema di riferimento; la figura seguente illustra tale situazione.

In definitiva, possiamo concludere che la distorsione di $\hat{\bar{Y}}$ tende a diminuire sia all'aumentare della dimensione n del campione, sia quando la funzione di regressione di y rispetto ad x è



presso a poco lineare con coefficiente di regressione pari al rapporto $\bar{Y}/\bar{X}$, o, in altre parole, quando tra y e x esiste un rapporto diretto di proporzionalità quasi costante.

Sulla base delle precedenti considerazioni teoriche, sviluppate con riferimento alla varianza e alla distorsione della stima $\hat{\bar{Y}}$, si può giustificare la regola pratica secondo la quale il metodo del rapporto si presenta raccomandabile per valori della dimensione del campione sufficientemente grandi (in letteratura è indicata la solita condizione $n > 30 \div 40$) in presenza di un coefficiente di variazione della x inferiore a quello della y e di valori (X_i, Y_i), i = 1, ..., n, che tendono ad essere allineati lungo una retta passante per l'origine.

Tali condizioni possono, infatti, far ritenere la stima $\hat{\bar{Y}}$ approssimativamente corretta e più efficiente della stima $\hat{Y}$.

Le precedenti considerazioni si possono facilmente illustrare con l'esempio seguente.

Consideriamo una popolazione costituita da N = 3 contribuenti; siano:

$$Y_1 = 1,\ Y_2 = 3,\ Y_3 = 15$$

le imposte sul reddito pagate dai tre contribuenti e:

$$X_1 = 10,\ X_2 = 25,\ X_3 = 100$$

i rispettivi redditi individuali.

Immaginiamo di conoscere a priori il reddito totale della popolazione (X = 135) e supponiamo di voler stimare, per mezzo di un campione di ampiezza n = 2, l'ammontare totale di imposta sul reddito corrisposto dai tre contribuenti.

Estratto il campione, siano:

$$(Y_2, X_2) \text{ e } (Y_3, X_3)$$

le coppie di valori dei due caratteri (imposta e reddito) associate ai due contribuenti inclusi nel campione.

La stima dell'imposta totale, mediante la procedura basata sul metodo del rapporto, è data da:

$$\hat{Y} = \frac{\frac{3}{2}(3 + 15)}{\frac{3}{2}(25 + 100)} 135 = 19{,}44$$

mentre l'imposta totale (incognita) è pari a $Y = Y_1 + Y_2 + Y_3 = 19$.

La stima diretta di Y sarebbe risultata uguale a:

$$\hat{Y} = \frac{3}{2}(3 + 15) = 27$$

Campionamento ad uno stadio stratificato

Passando dal campionamento casuale semplice a quello ad uno stadio stratificato si presentano, come possibili, due diversi modi di applicare il metodo del rapporto.

Consideriamo la prima alternativa.

Innanzi tutto, immaginiamo che la popolazione di N unità, considerata nel caso del campionamento semplice, sia ripartita in H strati.

Fermo restando tutto quanto è stato precisato sulla struttura e sulle caratteristiche della popolazione potremo, in maniera affatto analoga, individuare strutture e caratteristiche della parte di popolazione racchiusa in ciascuno strato con gli stessi simboli, salvo l'aggiunta di un indice distintivo di strato.

Perciò, con riferimento allo strato generico h (h = 1, ..., H) indichiamo con N_h il numero di elementi contenuti in detto strato e con:

$$(Y_{h1}, X_{h1}), \ldots., (Y_{hi}, X_{hi}), \ldots., (Y_{hN_h}, X_{hN_h})$$

le coppie dei due caratteri y ed x.

Siano inoltre:

$$Y_h = \sum_{i=1}^{N_h} Y_{hi} \qquad (25)$$

e

$$Y = \sum_{h=1}^{H} Y_h \qquad (26)$$

rispettivamente i totali del carattere y relativi al generico strato h e alla popolazione.

Analoghe quantità possono definirsi per il carattere x, avendosi:

$$X_h = \sum_{i=1}^{N_h} X_{hi} \qquad (27)$$

$$X = \sum_{h=1}^{H} X_h \qquad (28)$$

Supponiamo ora che da ogni strato vengano estratti n_h elementi senza reimmissione e probabilità uguali; siano:

$$(Y_{h1}, X_{h1}), \ldots, (Y_{hi}, X_{hi}), \ldots (Y_{hn_h}, X_{hn_h})$$

le n_h coppie di valori dei due caratteri y ed x osservati sulle unità estratte dallo strato h.

Assumendo noti (a priori) i totali $X_1, \ldots, X_h, \ldots, X_H$ si vuole stimare, al solito, il totale del carattere y, espresso dalla (26).

La stima di Y, mediante il metodo del rapporto, è data da:

$${}_s\hat{Y} = \sum_{h=1}^{N} \frac{\hat{Y}_h}{\hat{X}_h} X_h \qquad (29)$$

in cui:

$$\hat{Y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} Y_{hi} = \sum_{i=1}^{n_h} K_h Y_{hi} \qquad (30)$$

$$\hat{X}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} X_{hi} = \sum_{i=1}^{n_h} K_h X_{hi} \qquad (31)$$

$$K_h = \frac{N_h}{n_h} \qquad (32)$$

essendo K_h il peso base relativo allo strato h.

Posto:

$$B_h = \frac{X_h}{\hat{X}_h} \qquad (33)$$

la (29), tenendo presente la (30), può scriversi nella forma equivalente:

$${}_s\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} K_h B_h Y_{hi} \qquad (34)$$

che può riscriversi:

$${}_s\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} {}_fK_h Y_{hi} \qquad (35)$$

in cui:

$${}_fK_h = K_h B_h = \frac{X_h}{\sum_{i=1}^{n_h} X_{hi}} \qquad (36)$$

è il peso finale da attribuire a ciascuna delle n_h unità campione dello strato h.

Esaminiamo ora il secondo modo. Esso consiste nel considerare le due stime dirette di Y e X, ossia:

$$\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h \tag{37}$$

$$\hat{X} = \sum_{h=1}^{H} \hat{X}_h \tag{38}$$

e nell'ottenere la stima di Y mediante l'espressione:

$${}_c\hat{Y} = \frac{\hat{Y}}{\hat{X}} X \tag{39}$$

Posto:

$$B' = \frac{X}{\hat{X}} \tag{40}$$

ed introducendo nella (39) la (37) si ha:

$${}_c\hat{Y} = \sum_{h=1}^{H} B' \hat{Y}_h = \sum_{h=1}^{H} \sum_{i=1}^{n_h} B' K_h Y_{hi} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} {}_fK'_h Y_{hi} \tag{41}$$

in cui:

$${}_fK'_h = B' K_h \tag{42}$$

è il peso finale da attribuire a ciascuna delle n_h unità campione dello strato h.

Dopo aver descritto la struttura fondamentale dei due metodi, noti rispettivamente con il nome di *metodo del rapporto separato* e *metodo del rapporto combinato,* riteniamo opportuno analizzarne alcune caratteristiche differenziali, sì da fornire al lettore utili criteri di scelta.

La più importante di esse è rappresentata dalla precisione di un metodo rispetto all'altro.

A tale scopo dovremmo fare approfondita analisi delle rispettive varianze campionarie, le cui espressioni saranno illustrate in modo dettagliato nel Capitolo 11.

Se il numero n_h delle unità che si prelevano da ciascuno strato è sufficientemente grande (diciamo $n > 30 + 40$), dal confronto tra le due varianze campionarie si deduce che se i rapporti $\hat{Y}_h/\hat{X}_h$ sono costanti nei singoli strati i due metodi forniscono uguale precisione, mentre se i rapporti $\hat{Y}_h/\hat{X}_h$ sono variabili da strato a strato, il metodo del rapporto separato risulta più preciso; in altre parole, la varianza campionaria della stima ${}_s\hat{Y}$ risulta più piccola di quella della stima ${}_c\hat{Y}$.

Per analizzare più compiutamente le caratteristiche differenziali dei due metodi di stima è opportuno fare anche qualche considerazione sulla distorsione insita nelle stime ${}_s\hat{Y}$ e ${}_c\hat{Y}$.

Si dimostra (Sukhatme e Sukhatme, 1970), utilizzando lo sviluppo in serie di Taylor limitato ai termini di ordine quadratico, che la distorsione delle stime ${}_s\hat{Y}$ e ${}_c\hat{Y}$ è rispettivamente fornita dalle espressioni approssimate:

$$B\,({}_s\hat{Y}) = \sum_{h=1}^{H} Y_h \frac{N_h - n_h}{N_h} \frac{1}{n_h} \left(\frac{S_{xh}^2}{\overline{X}_h^2} - \rho_h \frac{S_{xh}}{\overline{X}_h} \frac{S_{yh}}{\overline{Y}_h} \right) \tag{43}$$

$$B\,({}_c\hat{Y}) = Y \sum_{h=1}^{H} \frac{N_h - n_h}{N_h} \left(\frac{N_h}{N} \right)^2 \frac{1}{n_h} \left(\frac{S_{xh}^2}{\overline{X}^2} - \rho_h \frac{S_{xh}}{\overline{X}} \frac{S_{yh}}{\overline{Y}} \right) \tag{44}$$

in cui:

$$S_{xh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(X_{hi} - \overline{X}_h \right)^2 \tag{45}$$

$$S_{yh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(Y_{hi} - \overline{Y}_h \right)^2 \tag{46}$$

essendo ρ_h il coefficiente di correlazione (Leti, 1983) tra le componenti della variabile doppia (x_h, y_h).

Per studiare il comportamento delle distorsioni B $(_s\hat{\bar{Y}})$ e B $(_c\hat{\bar{Y}})$ al variare della dimensione del campione è opportuno scrivere in una forma più semplice le espressioni (43) e (44).

A tal fine la (43), trascurando il fattore di correzione finito e ponendo:

$$n_h = \frac{n}{H} = \bar{n} = \text{cost.} \tag{47}$$

$$\frac{S_{xh}^2}{\overline{X}_h^2} = C_x^2 = \text{cost.}; \frac{S_{yh}^2}{\overline{Y}_h^2} = C_y^2 = \text{cost.}; \rho_h = \rho = \text{cost.} \tag{48}$$

può riscriversi nella forma:

$$B\,(_s\hat{\bar{Y}}) = \frac{Y}{\bar{n}}\,(C_x^2 - \rho\, C_x\, C_y) \tag{49}$$

dalla quale si deduce che $_s\hat{\bar{Y}}$ fornisce una stima soddisfacente di Y se la dimensione del campione di ciascuno strato è sufficientemente grande.

In modo analogo, la (44), trascurando il fattore di correzione finito e ponendo:

$$\frac{n_h}{N_h} = \frac{n}{N} = \text{cost.} \tag{50}$$

$$\frac{S_{xh}^2}{\overline{X}^2} = \overline{C}_x^2 = \text{cost.}; \frac{S_{yh}^2}{\overline{Y}^2} = \overline{C}_y^2 = \text{cost.}; \rho_h = \rho = \text{cost.} \tag{51}$$

può porsi nella forma:

$$B\,(_c\hat{\bar{Y}}) = \frac{Y}{n}\,(\overline{C}_x^2 - \rho\, \overline{C}_x\, \overline{C}_y) \tag{52}$$

dalla quale segue che, anche quando la dimensione n_h dei singoli strati è piccola, $_c\hat{\bar{Y}}$ fornisce una stima soddisfacente di Y se la dimensione campionaria complessiva n è sufficientemente grande.

Dalla (43), inoltre, si deduce che la distorsione di $_s\hat{\bar{Y}}$ si annulla quando, per ciascuno degli H strati, risulta soddisfatta la relazione:

$$\frac{S_{xh}^2}{\overline{X}_h^2} - \rho_h\, \frac{S_{xh}}{\overline{X}_h}\, \frac{S_{yh}}{\overline{Y}_h} = o \tag{53}$$

ossia quando:

$$\rho_h\, \frac{S_{xh}\, S_{yh}}{S_{xh}^2} = \frac{S_{xyh}}{S_{xh}^2} = \frac{\overline{Y}_h}{\overline{X}_h} \tag{54}$$

in cui si è posto:

$$\rho_h\, S_{xh}\, S_{yh} = S_{xyh} \tag{55}$$

con:

$$S_{xyh} = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (X_{hi} - \overline{X}_h)\,(Y_{hi} - \overline{Y}_h) \tag{56}$$

Il risultato espresso dalla (54) consente pertanto di concludere che la distorsione di $_s\hat{\bar{Y}}$ è nulla se ciascuna delle H rette di regressione passa per l'origine con coefficiente di regressione S_{xyh} / S_{xh}^2 pari al rapporto $\overline{Y}_h / \overline{X}_h$.

Analogamente, dalla (44) segue che la distorsione di $_c\hat{\bar{Y}}$ si annulla quando, per ciascuno degli H strati, risulta soddisfatta la relazione:

$$\frac{S_{xh}}{\overline{X}^2} - \rho_h\, \frac{S_{xh}}{\overline{X}}\, \frac{S_{yh}}{\overline{Y}} = o \tag{57}$$

ossia se:

$$\rho_h\, \frac{S_{xh}\, S_{yh}}{S_{xh}^2} = \frac{S_{xyh}}{S_{xh}^2} = \frac{\overline{Y}}{\overline{X}} \tag{58}$$

dalla quale si deduce che la distorsione di $_c\hat{\bar{Y}}$ si annulla se le H rette di regressione passano tutte per l'origine con coefficiente di regressione pari al rapporto $\overline{Y} / \overline{X}$.

Le precedenti considerazioni teoriche sul livello di precisione e sulla distorsione delle stime ${}_s\hat{\bar{Y}}$ e ${}_c\hat{\bar{Y}}$ possono essere così sintetizzate:

- se i rapporti $\bar{Y}_h / \bar{X}_h$ sono molto variabili da strato a strato il metodo del rapporto separato conduce a stime più precise di quelle ottenibili con il metodo del rapporto combinato;
- se il numero di strati H è elevato e le dimensioni campionarie n_h di alcuni strati sono piccole (diciamo, $n_h < 30$) la distorsione complessiva B (${}_s\hat{\bar{Y}}$) potrebbe essere influenzata da una forte distorsione in detti strati e risultare quindi più corretto il metodo del rapporto combinato.

Concludendo, tali risultati conducono a formulare la seguente regola di scelta fra i due metodi: se i rapporti $\bar{Y}_h / \bar{X}_h$ sono poco variabili da strato a strato, conviene usare il metodo del rapporto combinato in quanto fornisce stime caratterizzate da una distorsione trascurabile e da una precisione presso a poco uguale a quella del metodo del rapporto separato.

È stata, altresì, suggerita (Desabie, 1959) una regola pratica in base alla quale il metodo del rapporto separato è da preferire al metodo del rapporto combinato qualora per ogni strato risulta:

$$\sqrt{H}\,\frac{\sigma(\hat{X}_h)}{X_h} < 0{,}2 \qquad (59)$$

in cui $\sigma(\hat{X}_h)$ indica l'errore di campionamento della stima diretta $\hat{X}_h$.

In caso contrario, la distorsione complessiva della stima ${}_s\hat{\bar{Y}}$ rischia di non essere trascurabile; in tale circostanza conviene usare il metodo combinato.

Riteniamo, infine, utile suggerire una regola di decisione, di carattere generale, per scegliere fra diversi stimatori (corretti e/o distorti) dello stesso parametro; essa può essere stabilita attraverso lo studio dell'errore quadratico medio (EQM), definito dall'espressione (Kish, 1965):

$$EQM(\hat{\Theta}) = E(\hat{\Theta} - \Theta)^2 = V(\hat{\Theta}) + B^2(\hat{\Theta}) \qquad (60)$$

in cui $\hat{\Theta}$ indica lo stimatore del parametro Θ.

Questa espressione è composta da due quantità: la varianza dello stimatore e il quadrato della sua distorsione. Naturalmente se $\hat{\Theta}$ è non distorto, allora l'errore quadratico medio coinciderà con la varianza.

La (60), oltre a mettere in evidenza il ruolo della distorsione nella misura della rappresentatività di uno stimatore, fornisce una regola per la scelta fra stimatori diversi. Infatti, dati gli stimatori $\hat{\Theta}_1$ e $\hat{\Theta}_2$ del parametro Θ, si sceglierà quello per cui la (60) è più piccola. Quanto detto ci porta a definire l'*efficienza relativa* fra due stimatori. Più precisamente: dati due stimatori $\hat{\Theta}_1$ e $\hat{\Theta}_2$ di Θ, si chiama *efficienza relativa* la quantità:

$$E(\hat{\Theta}_1 / \hat{\Theta}_2) = \frac{EQM(\hat{\Theta}_2)}{EQM(\hat{\Theta}_1)} \qquad (61)$$

In tal modo si sceglierà $\hat{\Theta}_1$ se la (61) è maggiore di uno, si sceglierà $\hat{\Theta}_2$ se è minore di uno; altrimenti la scelta è indeterminata, cioè nessuno dei due stimatori è preferibile.

La (61) dunque ci fornisce un altro strumento per poter scegliere fra i metodi del rapporto separato e combinato; naturalmente, l'analisi differenziale, volutamente limitata ai due suddetti metodi, poteva essere estesa, così come abbiamo fatto nel caso del disegno casuale semplice, anche allo stimatore diretto:

$$\hat{Y} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} Y_{hi} \qquad (62)$$

descritto nel precedente Capitolo 6 e che presenta la proprietà di fornire stime corrette.

A questo riguardo osserviamo che se nel definire formalmente la stima di un parametro si cerca sempre di fare in modo che essa risulti corretta, però, l'eventuale distorsione che talvolta non si riesce ad eliminare non costituisce necessariamente un attributo peggiorativo della stima.

In altri termini, una stima corretta non è sicuramente migliore di una stima distorta, in quanto è l'errore quadratico medio che concorre a misurare la bontà di una stima, nel senso che dovrà ritenersi migliore quella stima che comporti più piccolo EQM.

Campionamento a due stadi con stratificazione delle unità primarie

Nelle pagine che seguono ci proponiamo di illustrare il metodo del rapporto con riferimento al caso di indagini basate su disegni a due stadi con stratificazione delle unità di primo stadio.

Innanzi tutto, è opportuno richiamare alcune caratteristiche della popolazione sulla quale si effettua l'operazione di campionamento.

Si dispone di H strati, ciascuno dei quali contenente N_h unità primarie; sia M_{hi} il numero di unità secondarie appartenenti alla generica unità primaria i (i= 1,...., N_h) dello strato h (h= 1,..., H).

Indichiamo con y il generico carattere oggetto d'indagine e con x un secondo carattere, in relazione con y, e del quale si posseggono, a priori, informazioni ausiliarie necessarie per la stima del totale Y del carattere y.

Alle M_{hi} unità sono associate le coppie:

$$(Y_{hi1}, X_{hi1}),\ldots, (Y_{hij}, X_{hij}),\ldots, (Y_{hiM_{hi}}, X_{hiM_{hi}})$$

in cui Y_{hij} e X_{hij} esprimono, rispettivamente, i valori assunti dai due caratteri y e x sulla generica unità secondaria j dell'unità primaria i dello strato h.

Introduciamo, poi, le quantità:

$$Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{hij} = \sum_{h=1}^{H} Y_h \qquad (63)$$

$$X = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} X_{hij} = \sum_{h=1}^{H} X_h \qquad (64)$$

che definiscono rispettivamente il totale del carattere y e il totale del carattere x, avendo indicato con:

$$Y_h = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{hij} \qquad (65)$$

$$X_h = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} X_{hij} \qquad (66)$$

i corrispondenti totali relativi allo strato h.

Passando all'estrazione del campione, la scelta di primo stadio consiste nell'estrarre, senza reimmissione, un numero prefissato di n_h unità primarie dallo strato h (h=1, ..., H); la scelta di secondo stadio consiste nell'estrarre, senza reimmissione, da ciascuna delle n_h unità prescelte, un numero di unità prefissate secondo un determinato criterio: cioè alla generica unità primaria i dello strato h è associato a priori il numero m_{hi} ($m_{hi} < M_{hi}$)

delle unità da estrarre nel secondo stadio, qualora l'unità primaria medesima sia estratta nel primo stadio.

Per l'impiego del metodo del rapporto si richiede una contemporanea rilevazione sia del carattere y che del carattere x, sulle m_{hi} unità secondarie, avendosi le coppie:

$$(Y_{hi1}, X_{hi1}),.., (Y_{hij}, X_{hij}),....,(Y_{him_{hi}}, X_{him_{hi}})$$

per i= 1,....,n_h e j= 1,....,m_{hi}.

Ciò posto, osserviamo che anche questa volta vi sono due modi di procedere.

La prima alternativa, nota con il nome di procedimento di stima secondo il *metodo del rapporto separato*, conduce all'espressione (Cochran, 1977):

$${}_s\hat{Y} = \sum_{h=1}^{H} \frac{\hat{Y}_h}{\hat{X}_h} X_h \qquad (67)$$

dove ${}_s\hat{Y}$ indica appunto la stima del totale Y.

Nella (67) $\hat{Y}_h$ e $\hat{X}_h$ rappresentano rispettivamente le stime dirette di Y_h e X_h; tenendo presente quanto già illustrato nel precedente Capitolo 6, le suddette stime sono espresse da:

$$\hat{Y}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi}\, Y_{hij} \qquad (68)$$

$$\hat{X}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi}\, X_{hij} \qquad (69)$$

in cui:

$$K_{hi} = \frac{1}{\Pi_{hi}} = \frac{1}{\Pi_{1hi}\,\Pi_{2hi}} \qquad (70)$$

è il peso base relativo a ciascuna delle m_{hi} unità.

Posto:

$$B_h = \frac{X_h}{\hat{X}_h} \qquad (71)$$

la (67) può scriversi nella forma:

$$ {}_s\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi} B_h Y_{hij} \qquad (72) $$

che può porsi nella forma:

$$ {}_s\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} {}_fK_{hi} Y_{hij} \qquad (73) $$

in cui:

$$ {}_fK_{hi} = K_{hi} B_h \qquad (74) $$

è il peso finale da attribuire a ciascuna delle m_{hi} unità.

Come si vede dalla (67), per utilizzare il procedimento di stima testè descritto, occorre che siano noti, a priori, i totali $X_1, \ldots, X_h, \ldots, X_H$, che costituiscono l'*informazione ausiliaria esterna.*

Il secondo metodo di stima si basa sull'utilizzazione della seguente formula (Hansen, Hurwitz e Madow, 1953):

$$ {}_c\hat{Y} = \frac{\hat{Y}}{\hat{X}} X \qquad (75) $$

che fornisce la stima ${}_c\hat{Y}$ del totale Y, nota in letteratura con il nome di *stima del rapporto combinato.*

Nella (75) i simboli $\hat{Y}$ e $\hat{X}$ indicano le stime corrette rispettivamente di Y e X; esse sono definite da:

$$ \hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi} Y_{hij} \qquad (76) $$

$$ \hat{X} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi} X_{hij} \qquad (77) $$

Posto:

$$ B' = \frac{X}{\hat{X}} \qquad (78) $$

la (75) si può scrivere nella forma:

$$ {}_c\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi} B' Y_{hij} \qquad (79) $$

che può riscriversi:

$$ {}_c\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} {}_fK'_{hi} Y_{hij} \qquad (80) $$

in cui:

$$ {}_fK'_{hi} = K_{hi} B' \qquad (81) $$

è il peso finale relativo a ciascuna delle m_{hi} unità appartenenti all'unità primaria i dello strato h.

A conclusione delle precedenti considerazioni sul metodo del rapporto, nel contesto del campionamento a due stadi, sarebbe opportuno svolgere anche un confronto sistematico fra gli stimatori ${}_s\hat{Y}$ e ${}_c\hat{Y}$, definiti dalla (67) e (75), e lo stimatore diretto, definito dalla (22) del capitolo 6.

Tale confronto, da condurre in termini di varianza, distorsione ed errore quadratico medio, avrebbe lo scopo di far meglio risaltare le proprietà di ciascuno dei suddetti stimatori e di analizzarne compiutamente le caratteristiche differenziali, in modo da fornire criteri validi di scelta.

Tuttavia, poichè le espressioni dell'errore quadratico medio di ${}_s\hat{Y}$ e ${}_c\hat{Y}$ sono algebricamente abbastanza complesse, non riteniamo opportuno fare qui un'esauriente esposizione teorica.

Vogliamo però sottolineare il fatto che gli stimatori ${}_s\hat{Y}$ e ${}_c\hat{Y}$ sono rispettivamente affetti da una distorsione di ordine $1/n_h$ e $1/n$, in cui n_h e n indicano il numero di unità primarie nello strato h e il numero complessivo delle stesse. Da ciò segue che: ${}_s\hat{Y}$ fornisce una stima soddisfacente del totale Y se n_h è sufficientemente grande in ciascuno degli H strati; ${}_c\hat{Y}$ fornisce una stima soddisfacente del totale Y se n è sufficientemente grande.

In base a tali osservazioni si può giustificare la regola pratica secondo la quale, se il numero di strati H è elevato e le dimensioni campionarie n_h di alcuni strati sono piccole, si presenta raccomandabile lo stimatore ${}_c\hat{Y}$. In tale situazione, infatti, la distorsione complessiva di ${}_s\hat{Y}$ potrebbe essere influenzata da una forte distorsione derivante dagli strati in cui n_h è piccolo.

CAPITOLO 9 - METODI DI STIMA INDIRETTI: GLI STIMATORI DEL RAPPORTO POST-STRATIFICATI

Introduzione

Nel precedente capitolo sono stati illustrati gli stimatori basati sul metodo del rapporto, che consentono di ottenere stime più precise rispetto a quelle ottenibili mediante l'uso di stimatori diretti, a parità di condizioni (numerosità campionaria, stratificazione, probabilità di selezione, ecc.).

Un ulteriore miglioramento del livello di precisione delle stime può ottenersi attraverso l'adozione di stimatori noti in letteratura con il nome di *stimatori del rapporto post-stratificati*.

A questo riguardo, già nel Capitolo 3, abbiamo introdotto il concetto di post-stratificazione, nel contesto del campionamento casuale semplice, con riferimento sia allo stimatore diretto che a quello del rapporto.

I problemi fondamentali della post-stratificazione sono la scelta delle variabili e la formazione dei sottogruppi di popolazione da utilizzare come post-strati.

Le variabili scelte, fra quelle disponibili, devono presentare un livello elevato di correlazione con le variabili oggetto di indagine; tale condizione consente la formazione di post-strati omogenei. Due variabili che frequentemente vengono utilizzate a tal fine sono il sesso e l'età; in una indagine sulle forze di lavoro una post-stratificazione basata su tali caratteri consente di creare gruppi omogenei rispetto alle variabili di interesse (ad esempio le sotto-popolazioni individuate dalla combinazione congiunta delle modalità del sesso e dell'età - espresse generalmente in classi — risultano più omogenee rispetto alla variabile oggetto di indagine *condizione professionale*)

In questo capitolo approfondiremo i concetti già dati, estendendo la trattazione agli altri due disegni di campionamento presi in considerazione in questo volume.

Campionamento casuale semplice

Consideriamo un campione costituito da n elementi, estratti senza reimmissione e probabilità uguali, da una popolazione di N elementi; supponiamo inoltre di voler stimare il totale del carattere y.

Nel Capitolo 8 abbiamo esaminato come stimare detto totale mediante lo stimatore del rapporto, basato sull'utilizzazione di una variabile x correlata con la variabile oggetto di studio y; ora illustreremo l'utilizzazione dello stimatore del rapporto post-stratificato.

Consideriamo una variabile ausiliaria z, correlata con il carat-

tere y, e suddividiamola in modo da formare *a* sotto-gruppi (a = 1, ..., A) nei quali classificare le N unità della popolazione e le n unità campione (dopo aver effettuato l'indagine). Indicando con ${}_aN$ ed ${}_an$ le dimensioni della sub-popolazione e del campione relative al post-strato a, valgono le relazioni:

$$N = {}_1N + \ldots + {}_aN + \ldots {}_AN$$

$$n = {}_1n + \ldots + {}_an + \ldots {}_An$$

Su ciascuna unità campione i, appartenente al post-strato a, vengono osservate le coppie di valori $({}_aY_i, {}_aX_i)$ dei due caratteri y — oggetto di indagine — ed x — ausiliario.

Definiamo per ogni post-strato i totali dei due caratteri mediante le espressioni seguenti:

$${}_aY = \sum_{i=1}^{N} {}_aY_i \qquad {}_aX = \sum_{i=1}^{N} {}_aX_i \tag{1}$$

le cui stime dirette, e corrette, sono fornite rispettivamente da:

$${}_a\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} {}_aY_i \qquad {}_a\hat{X} = \frac{N}{n} \sum_{i=1}^{n} {}_aX_i \tag{2}$$

È da notare che le sommatorie figuranti nelle espressioni (1) e (2) potrebbero essere limitate ad ${}_aN$ e ${}_an$, che esprimono rispettivamente l'ammontare della popolazione e del campione nel sub-strato a; abbiamo invece esteso tali sommatorie al complesso delle unità (N ed n), considerando che le variabili ${}_aY_i$ ed ${}_aX_i$ assumono valore zero quando l'unità selezionata non appartiene al post-strato a, per il quale vogliamo ottenere l'informazione.

Utilizzando le espressioni (1) e (2) costruiamo lo stimatore del rapporto post-stratificato:

$${}_p\hat{Y} = \sum_{a=1}^{A} \frac{{}_a\hat{Y}}{{}_a\hat{X}} {}_aX \tag{3}$$

Dalla (3) si deduce la necessità della conoscenza a priori del totale ${}_aX$, che rappresenta *l'informazione ausiliaria*, desumibile generalmente da registri, liste anagrafiche, ecc..

L'espressione (3) può essere riscritta in una forma più conveniente ai fini pratici del calcolo delle stime; a tale scopo introduciamo il *peso base*:

$$K = \frac{N}{n} \tag{4}$$

inoltre esprimiamo, per ciascun post-strato, il rapporto fra il totale ${}_aX$ e la corrispondente stima ${}_a\hat{X}$ nel modo seguente:

$${}_aB = \frac{{}_aX}{{}_a\hat{X}} \tag{5}$$

Possiamo così costruire ${}_aK$, il *peso finale* da attribuire a ciascuna delle ${}_an$ unità del campione:

$${}_aK = K \cdot {}_aB \tag{6}$$

Utilizzando quest'ultima relazione, possiamo riscrivere la (3) nel modo seguente:

$${}_p\hat{Y} = \sum_{a=1}^{A} \sum_{i=1}^{n} {}_aK \, {}_aY_i \tag{7}$$

Campionamento ad uno stadio stratificato

Supponiamo che la popolazione di N unità, utilizzata nel campionamento casuale semplice, sia suddivisa in H strati, da ciascuno dei quali vengono estratti n_h elementi, senza reimmissione e probabilità uguali.

L'obiettivo è sempre stimare il totale del carattere y mediante lo stimatore del rapporto post-stratificato. Come è stato già descritto nel Capitolo 8, nel campionamento ad uno stadio stratificato si presentano due modi diversi di utilizzare lo stimatore del rapporto; prendiamo in esame la prima alternativa, che consiste nel costruire una stima del rapporto per il totale di ogni strato ed addizionare tali totali *(stimatore del rapporto separato)*.

Consideriamo la variabile ausiliaria z, correlata con il carattere y, in base alla quale definiamo a sottogruppi (a = 1, ..., A) nei quali classificare, dopo aver effettuato l'indagine, sia le N_h unità della popolazione che le n_h unità campione.

A ciascuna unità i dello strato h, appartenente al post-strato a, vengono associate le coppie ai valori $({}_aY_{hi}, {}_aX_{hi})$. Siano inoltre:

$$ {}_aX_h = \sum_{i=1}^{N_h} {}_aX_{hi} \qquad {}_aY_h = \sum_{i=1}^{N_h} {}_aY_{hi} \tag{8} $$

i totali dei caratteri x ed y relativi alla popolazione dello stato h, appartenti al post-strato a. Assumendo noti (a priori) i totali ${}_aX_h$, possiamo stimare il totale del carattere Y mediante lo stimatore del rapporto post-stratificato separato (${}_{ps}\hat{Y}$) espresso da:

$$ {}_{ps}\hat{Y} = \sum_{a=1}^{A} \sum_{h=1}^{H} \frac{{}_a\hat{Y}_h}{{}_a\hat{X}_h} {}_aX_h \tag{9} $$

in cui:

$$ {}_a\hat{Y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} {}_aY_{hi} \tag{10} $$

$$ {}_a\hat{X}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} {}_aX_{hi} \tag{11} $$

sono le stime dirette rispettivamente del totale del carattere y e del totale del carattere x.

Per riscrivere la (9) in una forma più conveniente per il calcolo delle stime, poniamo:

$$ K_h = \frac{N_h}{n_h} \tag{12} $$

che definiamo *peso base*; e

$$ {}_aB_h = \frac{{}_aX_h}{{}_a\hat{X}_h} \tag{13} $$

il rapporto fra il totale del carattere x, noto a priori, e la stima diretta di questo totale ottenuta dalle osservazioni campionarie, per ciascuno strato h e sottogruppo a; in tal caso il corrispondente *peso finale* è definito da:

$$ {}_aK_h = K_h \cdot {}_aB_h \tag{14} $$

Tenendo presente quest'ultima espressione, possiamo riscrivere la (9) nella forma equivalente:

$$ {}_{ps}\hat{Y} = \sum_{a=1}^{A} \sum_{h=1}^{H} \sum_{i=1}^{n_h} {}_ak_h \, {}_aY_{hi} \tag{15} $$

Esaminiamo ora la seconda alternativa nota come *stimatore del rapporto combinato*, tale metodo non richiede la conoscenza (a priori) dei totali ${}_aX_h$, ma solamente di ${}_aX$, cioè dell'informazione per l'intero sottogruppo a; la stima di Y si ottiene mediante l'espressione:

$$ {}_{pc}\hat{Y} = \sum_{a=1}^{A} \frac{{}_a\hat{Y}}{{}_a\hat{X}} {}_aX \tag{16} $$

nella quale compaiono le stime dirette dei totali Y ed X fornite rispettivamente da:

$$ {}_a\hat{Y} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} {}_aY_{hi} \tag{17} $$

$$ {}_a\hat{X} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} {}_aX_{hi} \tag{18} $$

posto

$$ {}_aB' = \frac{{}_aX}{{}_a\hat{X}} \tag{19} $$

il rapporto fra il totale del carattere x e la sua stima per il sottogruppo a, definiamo il peso finale ${}_aK'_h$ da attribuire a ciascuna delle ${}_an_h$ unità campione dello strato h, post-stratificate nel gruppo a:

$$ {}_aK'_h = {}_aB' \cdot K_h \tag{20} $$

Tenendo conto della (17) e della (20) si ha:

$$_{pc}\hat{Y} = \sum_{a=1}^{A} \sum_{h=1}^{H} \sum_{i=1}^{n_h} {}_aK'_h \, {}_aY_{hi} \qquad (21)$$

Campionamento a due stadi con stratificazione delle unità primarie

In tale disegno campionario, la popolazione su cui si conduce l'indagine è divisa in H strati, ognuno dei quali contiene N_h unità primarie; sia inoltre M_{hi} il numero di unità secondarie appartenenti alla generica unità primaria i (i=1,..., N_h) dello strato h (h = 1,... H).

Supponiamo di estrarre, senza reimmissione, un numero prefissato di n_h unità primarie da ciascuno strato h e che da ognuna delle n_{hi} unità selezionate vengano estratte m_{hi} unità di secondo stadio.

Consideriamo ora la variabile z, correlata con il carattere y, del quale vogliamo stimare il totale, e definiamo *a* post-strati (a = 1, ..., A), nei quali classificare le M_h unità della popolazione ($M_h = \Sigma M_{hi}$ per i = 1,..., N_h); indicando con $_aM_h$ la sub-popolazione dello strato h, appartenente al post-strato a, possiamo definire le due espressioni:

$$_aY_h = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} {}_aY_{hij} \qquad {}_aX_h = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} {}_aX_{hij} \qquad (22)$$

che rappresentano rispettivamente i totali del carattere y ed x relativi alla sub-popolazione costituita dalle $_aM_h$ unità, a ciascuna delle quali vengono associate le coppie di valori ($_aY_{hij}$, $_aX_{hij}$).

In modo analogo possiamo classificare le m_h unità campione dello strato h nei post-strati *a*.

Assumendo noti (a priori) i totali $_aX_h$, possiamo utilizzare lo stimatore del rapporto post-stratificato separato ($_{ps}\hat{Y}$) espresso da:

$$_{ps}\hat{Y} = \sum_{a=1}^{A} \sum_{h=1}^{H} \frac{{}_a\hat{Y}_h}{{}_a\hat{X}_h} \, {}_aX_h \qquad (23)$$

L'espressione (23) ha la stessa struttura formale della (9), ma nel campionamento a due stadi, con stratificazione delle unità primarie, le stime dirette dei totali dei due caratteri ($_a\hat{Y}_h$, $_a\hat{X}_h$) tengono conto della complessità del disegno e sono espresse da:

$$_a\hat{Y}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi} \, {}_aY_{hij} \qquad (24)$$

$$_a\hat{X}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi} \, {}_aX_{hij} \qquad (25)$$

in cui il *peso base* K_{hi} è espresso da (vedi capitolo 6):

$$K_{hi} = \frac{1}{\Pi_{hi}} = \frac{1}{\Pi_{1hi} \, \Pi_{2hi}} \qquad (26)$$

Se inoltre poniamo:

$$_aB_h = \frac{{}_aX_h}{{}_a\hat{X}_h} \qquad (27)$$

è possibile definire il *peso finale* ($_aK_{hi}$) da attribuire a ciascuna unità secondaria m_{hij}, selezionata nell'unità primaria i dello strato h, appartenente al post-strato a:

$$_aK_{hi} = K_{hi} \, {}_aB_h \qquad (28)$$

Tenendo presente la (28), riscriviamo l'espressione dello stimatore $_{ps}\hat{Y}$ in forma più conveniente ai fini pratici del calcolo delle stime:

$$_{ps}\hat{Y} = \sum_{a=1}^{A} \sum_{h=1}^{H} \sum_{i=1}^{m_h} \sum_{j=1}^{m_{hi}} {}_aK_{hi} \, {}_aY_{hij} \qquad (29)$$

Passiamo ora ad illustrare la seconda alternativa nota come metodo del rapporto combinato. Come è stato già detto in tal caso non è necessario conoscere il totale ($_aX_h$) del carattere

ausiliario separatamente per ciascuno strato, ma utilizziamo l'informazione ausiliaria ($_aX$) per l'intero sottogruppo *a*, definiamo quindi l'espressione:

$$_{pc}\hat{Y} = \sum_{a=1}^{A} \frac{_a\hat{Y}}{_a\hat{X}} \, _aX \qquad (30)$$

in cui:

$$_a\hat{Y} = \sum_{h=1}^{H} {}_a\hat{Y}_h = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi} \, {}_aY_{hij}$$

$$_a\hat{X} = \sum_{h=1}^{H} {}_a\hat{Y}_h = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi} \, {}_aX_{hij} \qquad (31)$$

rappresentano le stime dirette dei totali dei due caratteri ottenuti utilizzando il peso base K_{hi} espresso dalla (28).

Ponendo il rapporto fra il totale del carattere x e la sua stima nel post-strato *a*, pari a:

$$_aB' = \frac{_aX}{_a\hat{X}} \qquad (32)$$

possiamo costruire il peso finale

$$_aK'_{hi} = K_{hi} \, {}_aB'_h \qquad (33)$$

necessario per riscrivere l'espressione (30) nella seguente forma equivalente:

$$_{pc}\hat{Y} = \sum_{a=1}^{A} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} {}_aK'_{hi} \, {}_aY_{hij} \qquad (34)$$

CAPITOLO 10 - VARIANZA DEGLI STIMATORI DIRETTI

Introduzione

Una volta ottenute le stime oggetto di indagine è opportuno determinare, per ciascuna di esse, una statistica mediante la quale è possibile valutarne l'affidabilità.

Tale statistica, nota con il nome di varianza totale, è costituita dalla somma di due componenti: la prima, denominata varianza di campionamento, è dovuta alla natura parziale della rilevazione; la seconda, derivante da numerosi e spesso incontrollabili fattori di disturbo, è nota con il nome di varianza di risposta.

In particolare la varianza di campionamento, che consente di valutare il livello di precisione di una stima campionaria, viene utilizzata anche per il calcolo dell'errore di campionamento, assoluto e relativo, e dell'intervallo di confidenza, che costituiscono ulteriori elementi di giudizio del livello di precisione dei risultati forniti da un'indagine campionaria.

Si ritiene utile sottolineare, inoltre, che la varianza di campionamento riveste un ruolo di grande importanza anche nell'ambito dei complessi problemi della stratificazione e della determinazione della numerosità campionaria.

Nel seguito illustreremo il calcolo della varianza della distribuzione campionaria della stima del totale e della stima di tale varianza, con riferimento ai criteri di selezione delle unità più utilizzati nelle indagini concrete su larga scala.

Campionamento casuale semplice

Affrontiamo, in primo luogo, il problema della determinazione della varianza nel contesto del campionamento casuale semplice in cui le unità si suppongono estratte senza reimmissione e probabilità uguali. A tale scopo introduciamo le seguenti espressioni:

$$Y = \sum_{i=1}^{N} Y_i \qquad (1)$$

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i \qquad (2)$$

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \overline{Y})^2 \qquad (3)$$

che definiscono rispettivamente il totale, la media e la varianza del carattere y nella popolazione.

Una stima corretta di Y è fornita dall'espressione:

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} Y_i \qquad (4)$$

È possibile dimostrare (Cochran, 1977; Yamane, 1967) che la varianza di $\hat{Y}$ è espressa dalla relazione:

$$V(\hat{Y}) = N^2 \frac{N-n}{N} \frac{S^2}{n} \qquad (5)$$

Se la varianza della popolazione non è nota occorre stimarla attraverso il campione; una stima corretta di S^2 è data da:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \hat{\overline{Y}})^2 \qquad (6)$$

dove:

$$\hat{\overline{Y}} = \frac{1}{n} \sum_{i=1}^{n} Y_i \qquad (7)$$

è una stima corretta di $\overline{Y}$.

Pertanto una stima corretta di $V(\hat{Y})$ è data dall'espressione:

$$\hat{V}(\hat{Y}) = N^2 \frac{N-n}{N} \frac{s^2}{n} \qquad (8)$$

Allo scopo di chiarire quanto sopra illustrato consideriamo, ad esempio, una popolazione costituita da N = 3 unità sulle quali

il carattere y oggetto di indagine assume i valori $Y_1 = 6$, $Y_2 = 5$ e $Y_3 = 10$.

Supponiamo, inoltre, di voler estrarre senza reimmissione e probabilità uguali un campione di numerosità n = 2 per stimare il totale Y; siano $Y_1 = 6$ e $Y_3 = 10$ i valori osservati sulle due unità estratte.

Il valore della varianza della popolazione è uguale a:

$$S^2 = \frac{1}{2}\left[(6-7)^2 + (5-7)^2 + (10-7)^2 \right] = \frac{14}{2} = 7$$

La stima del totale risulta pari a:

$$\hat{Y} = \frac{3}{2}(6+10) = 24$$

e la corrispondente varianza di campionamento:

$$V(\hat{Y}) = 3^2 \frac{3-2}{2} \frac{7}{2} = 15{,}7$$

Nelle situazioni concrete, in cui non si conoscono i valori di y nella popolazione, è possibile determinare soltanto una stima di $V(\hat{Y})$. Con riferimento al caso esemplificato si ha:

$$s^2 = (6-8)^2 + (10-8)^2 = 8$$

e

$$\hat{V}(\hat{Y}) = 3^2 \frac{3-2}{2} \frac{8}{2} = 18$$

Campionamento ad uno stadio stratificato

Illustreremo ora il calcolo della varianza della distribuzione campionaria della stima del totale, e della stima di questa varianza, nel caso di un campione a uno stadio stratificato, in cui le unità si suppongono estratte senza reimmissione e probabilità uguali da ogni strato.

Tenendo presente, le notazioni simboliche introdotte nel Capitolo 4, ricordiamo che Y_{hi} rappresenta il valore del carattere y oggetto di studio relativo alla generica unità i appartenente allo strato h.

Siano inoltre:

$$Y_h = \sum_{i=1}^{N_h} Y_{hi} \qquad (9)$$

$$\overline{Y}_h = \frac{1}{N_h} Y_h \qquad (10)$$

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \overline{Y}_h)^2 \qquad (11)$$

il totale, la media e la varianza relativi al generico strato h.

Dato il meccanismo probabilistico di selezione delle unità appena descritto e tenuto conto anche di quanto illustrato a proposito del campionamento casuale semplice, segue immediatamente (Cochran, 1977; Kish, 1965; Castellano ed Herzel, 1981) che la varianza di campionamento della stima $\hat{Y}_h$ è fornita dalla relazione:

$$V(\hat{Y}_h) = N_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h} \qquad (12)$$

essendo $\hat{Y}_h$ definita dall'espressione:

$$\hat{Y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} Y_{hi} \qquad (13)$$

Tenuto conto, poi, che la stima del totale

$$Y = \sum_{h=1}^{H} Y_h \qquad (14)$$

è data da:

$$\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h \qquad (15)$$

e che le estrazioni dei campioni dagli strati 1, 2,..., H sono indipendenti, si ricava che la varianza della (15) è espressa dalla relazione:

$$V(\hat{Y}) = V\left(\sum_{h=1}^{H} \hat{Y}_h\right) = \sum_{h=1}^{H} V(\hat{Y}_h) \qquad (16)$$

In base alla (12) la (16) può riscriversi nella forma:

$$V(\hat{Y}) = \sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h} \qquad (17)$$

Se la varianza della popolazione nello strato h (h= 1,...,H) non è nota, occorre stimarla attraverso il campione; poiché per il generico strato h valgono le considerazioni svolte per il campionamento casuale semplice segue che una stima corretta di S_h^2 è data da (Russo, 1982):

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (Y_{hi} - \hat{\overline{Y}}_h)^2 \qquad (18)$$

in cui:

$$\hat{\overline{Y}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi} \qquad (19)$$

Pertanto una stima corretta di V (Ŷ) è data dall'espressione:

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h} \frac{s_h^2}{n_h} \qquad (20)$$

Campionamento a due stadi con stratificazione delle unità primarie

Riteniamo utile ricordare, intanto, che il campionamento in oggetto si basa su un processo che consiste nella selezione da ciascuno strato di un prefissato numero di unità primarie, dalle quali vengono successivamente estratti i campioni di unità secondarie.

Tale processo comporta che la varianza di una data stima campionaria risulta costituita, come vedremo meglio nel seguito, da due componenti: la prima, dovuta alla variabilità tra le unità primarie; la seconda, alla variabilità tra le unità secondarie.

Nelle pagine che seguono verranno illustrate le espressioni della varianza di campionamento della stima di un totale, con riferimento ai due seguenti casi fondamentali:

- le unità primarie e quelle secondarie sono estratte senza reimmissione e probabilità uguali;
- le unità primarie sono estratte senza reimmissione con probabilità variabile e le unità secondarie senza reimmissione con probabilità uguali.

Esaminiamo, ora, il primo caso.

A tale scopo riteniamo utile riscrivere le seguenti quantità già definite nei Capitoli 4 e 6:

$$Y_{hi} = \sum_{j=1}^{M_{hi}} Y_{hij} \qquad (21)$$

che indica il totale del carattere y relativo all'unità primaria i dello strato h

$$Y_h = \sum_{i=1}^{N_h} Y_{hi} \qquad (22)$$

$$\hat{Y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} Y_{hij} = \frac{N_h}{n_h} \sum_{i=1}^{n_h} \hat{Y}_{hi} \qquad (23)$$

che rappresentano rispettivamente il totale del carattere y nello strato h e la corrispondente stima.

La varianza di primo stadio è definita dall'espressione (Hansen, Hurwitz e Madow, 1953):

$$V_I(\hat{Y}_h) = N_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h} \qquad (24)$$

in cui:

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(Y_{hi} - \frac{Y_h}{N_h} \right)^2 \qquad (25)$$

Al secondo stadio, viene selezionato un campione casuale semplice di m_{hi} unità secondarie, nell'ambito di ciascuna delle n_h unità primarie estratte al primo stadio. Si ha allora che la varianza della stima del totale nella generica unità primaria i, è data da:

$$M_{hi}^2 \frac{M_{hi} - m_{hi}}{M_{hi}} \frac{S_{hi}^2}{m_{hi}} \qquad (26)$$

in cui:

$$S_{hi}^2 = \frac{1}{M_{hi} - 1} \sum_{j=1}^{M_{hi}} \left(Y_{hij} - \frac{Y_{hi}}{M_{hi}} \right)^2 \qquad (27)$$

È possibile dimostrare (Hansen, Hurwitz e Madow, 1953) che la varianza di secondo stadio nel generico strato h è data dall'espressione:

$$V_{II}(\hat{Y}_h) = \frac{N_h}{n_h} \sum_{i=1}^{N_h} M_{hi}^2 \frac{M_{hi} - m_{hi}}{M_{hi}} \frac{S_{hi}^2}{m_{hi}} \qquad (28)$$

Pertanto la varianza della stima del totale si ottiene sommando la (24) e la (28), ossia in formula:

$$V(\hat{Y}_h) = V_I(\hat{Y}_h) + V_{II}(\hat{Y}_h) \tag{29}$$

Nel caso in cui le varianze S_h^2 e S_{hi}^2 non sono note occorre stimarle in base al campione; le stime corrette di queste varianze sono date rispettivamente da:

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \hat{\overline{Y}}_h \right)^2 \tag{30}$$

in cui:

$$\hat{\overline{Y}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{Y}_{hi} \tag{31}$$

è una stima corretta di:

$$\overline{Y}_h = \frac{Y_h}{N_h} \tag{32}$$

e da:

$$s_{hi}^2 = \frac{1}{m_{hi} - 1} \sum_{j=1}^{m_{hi}} \left(Y_{hij} - \hat{\overline{Y}}_{hi} \right)^2 \tag{33}$$

in cui:

$$\hat{\overline{Y}}_{hi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} Y_{hij} \tag{34}$$

è una stima corretta di $\overline{Y}_{hi}$.

Sostituendo la (30) e (33), rispettivamente nella (24) e nella (28) si ottiene una stima corretta di $V(\hat{Y}_h)$, data da:

$$\hat{V}(\hat{Y}_h) = \hat{V}_I(\hat{Y}_h) + \hat{V}_{II}(\hat{Y}_h) \tag{35}$$

dove:

$$\hat{V}_I(\hat{Y}_h) = N_h^2 \frac{N_h - n_h}{N_h} \frac{s_h^2}{n_h} \tag{36}$$

è una stima corretta di $V_I(\hat{Y}_h)$, e:

$$\hat{V}_{II}(\hat{Y}_h) = \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi}^2 \frac{M_{hi} - m_{hi}}{M_{hi}} \frac{s_{hi}^2}{m_{hi}} \tag{37}$$

è una stima corretta di $V_{II}(\hat{Y}_h)$.

Poiché l'estrazione in ciascuno strato è indipendente dalle estrazioni negli altri strati, la varianza della stima $\hat{Y}$ è data da:

$$V(\hat{Y}) = V\left(\sum_{h=1}^{H} \hat{Y}_h \right) = \sum_{h=1}^{H} V(\hat{Y}_h) \tag{38}$$

Una stima corretta di $V(\hat{Y})$ è invece fornita dall'espressione:

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H} \left(\hat{V}_I(\hat{Y}_h) + \hat{V}_{II}(\hat{Y}_h) \right) \tag{39}$$

Esaminiamo adesso il caso in cui le unità primarie vengono estratte senza reimmissione e probabilità variabile e quelle secondarie senza reimmissione e probabilità uguale.

È possibile dimostrare (Cochran, 1977) che la varianza di primo stadio nel generico strato h, è data dall'espressione:

$$V_I(\hat{Y}_h) = \sum_{i=1}^{N_h} \sum_{i>i'}^{N_h} (\Pi_{1hi}\,\Pi_{1hi'} - \Pi_{1hii'}) \left(\frac{Y_{hi}}{\Pi_{1hi}} - \frac{Y_{hi'}}{\Pi_{1hi'}} \right) \qquad (40)$$

in cui ricordiamo che Π_{1hi} e $\Pi_{1hi'}$ rappresentano le probabilità di inclusione delle unità primarie i ed i' appartenenti allo strato h e $\Pi_{1hii'}$ indica la probabilità di inclusione congiunta delle unità i e i'.

L'estrazione al secondo stadio avviene esattamente come nel caso precedente; perciò, la varianza della stima del totale nella generica unità primaria i è uguale alla (28). Inoltre è possibile dimostrare che quando le unità primarie vengono estratte senza reimmissione e probabilità variabili, la varianza di secondo stadio ha la seguente espressione:

$$V_{II}(\hat{Y}) = \sum_{i=1}^{N_h} \frac{M_{hi}(M_{hi} - m_{hi})}{m_{hi}\,\Pi_{1hi}} S_{hi}^2 \qquad (41)$$

Pertanto la varianza della stima del totale nello strato h è data dalla somma della (45) con la (46), e cioé da:

$$V(\hat{Y}_h) = V_I(\hat{Y}_h) + V_{II}(\hat{Y}_h) \qquad (42)$$

Una stima corretta della $V_{II}(\hat{Y}_h)$ è data da:

$$\hat{V}_{II}(\hat{Y}_h) = \sum_{i=1}^{n_h} \frac{M_{hi}(M_{hi} - m_{hi})}{m_{hi}\,\Pi_{1hi}} s_{hi}^2 \qquad (43)$$

dove s_{hi}^2 è data dalla (33), e una stima corretta di $V_I(\hat{Y}_h)$ è data da:

$$\hat{V}_I(\hat{Y}_h) = \sum_{i=1}^{n_h} \sum_{i>i'}^{n_h} \frac{(\Pi_{1hi}\,\Pi_{1hi'} - \Pi_{1hii'})}{\Pi_{1hii'}} \left(\frac{\hat{Y}_{hi}}{\Pi_{1hi}} - \frac{\hat{Y}_{hi'}}{\Pi_{1hi'}} \right)^2 \qquad (44)$$

Sostituendo la (43) insieme alla (44) nella (42) si ottiene una stima corretta di $V(\hat{Y}_h)$, data da:

$$\hat{V}(\hat{Y}_h) = \hat{V}_I(\hat{Y}_h) + \hat{V}_{II}(\hat{Y}_h) \qquad (45)$$

L'espressione della varianza della stima del totale è pertanto:

$$V(\hat{Y}) = \sum_{h=1}^{H} \left(V_I(\hat{Y}_h) + V_{II}(\hat{Y}_h) \right) \qquad (46)$$

e una sua stima è:

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H} \left(\hat{V}_I(\hat{Y}_h) + \hat{V}_{II}(\hat{Y}_h) \right) \qquad (47)$$

La tecnica del collassamento degli strati

A conclusione di questo capitolo riteniamo utile soffermarci su un problema che ricorre spesso con riferimento ad indagini basate su disegni campionari stratificati che prevedono la selezione di una sola unità da ogni strato.

In tal caso, infatti, la varianza di campionamento nello strato non può essere stimata e le formule precedentemente considerate non sono più valide.

È possibile, tuttavia, applicare una tecnica nota in letteratura come *metodo degli strati collassati* (*collapsed strata* in lingua inglese) che consiste nell'accoppiamento degli H strati originari, creando un'insieme di H/2 nuovi strati detti *superstrati*.

Nel seguito ci riferiremo ad un disegno di campionamento ad uno stadio stratificato, tuttavia le formule ottenute possono essere adattate facilmente sia a disegni più complessi, quali il disegno a due stadi con stratificazione delle unità primarie nel caso in cui venga estratta una sola unità primaria da ciascuno strato, che ad altri tipi di stimatori.

Supponiamo di aver effettuato l'accoppiamento degli H strati originari e indichiamo con t, (t = 1, ..., H/2) il generico *superstrato* costituito dall'unione di due strati originari, che denotiamo con u e w; indichiamo inoltre con $Y_{tu.}$ e $Y_{tw.}$ i valori della variabile y relativi all'unità selezionata nello strato u e a quella selezionata dallo strato w.

Una stima approssimata della (17) è fornita dall'espressione:

$$\hat{V}(\hat{Y}) = \sum_{t=1}^{H/2} \left(\hat{Y}_{tu} - \hat{Y}_{tw} \right)^2 \qquad (48)$$

dove

$$\hat{Y}_{tu} = N_u Y_{tu.} \quad e \quad \hat{Y}_{tw} = N_w Y_{tw.} \qquad (49)$$

Infatti, dall'identità:

$$\hat{Y}_{tu} - \hat{Y}_{tw} = (Y_{tu} - Y_{tw}) + (\hat{Y}_{tu} - Y_{tu}) - (\hat{Y}_{tw} - Y_{tw}) \qquad (50)$$

segue che:

$$E(\hat{Y}_{tu} - \hat{Y}_{tw})^2 = (Y_{tu} - Y_{tw})^2 + N_w(N_w - 1)S^2_w + N_u(N_u - 1)S^2_u \quad (51)$$

dove Y_{tu} e Y_{tw} indicano i totali del carattere y relativi agli strati u e w appartenenti al generico superstrato t (t = 1, ..., H/2).

Tenendo presente la (51), si ha poi:

$$E \left[\sum_{t=1}^{H/2} (\hat{Y}_{tu} - \hat{Y}_{tw})^2 \right] = \sum_{t=1}^{H/2} E (\hat{Y}_{tu} - \hat{Y}_{tw})^2 =$$

$$= \sum_{h=1}^{H} N_h (N_h - 1) S^2_h + \sum_{t=1}^{H/2} (Y_{tu} - Y_{tw})^2 =$$

$$= V(\hat{Y}) + \sum_{t=1}^{H/2} (Y_{tu} - Y_{tw})^2 \qquad (52)$$

Tale relazione dimostra che l'espressione compresa fra le parentesi quadre rappresenta in media la varianza più una componente distorsiva espressa dal secondo addendo della (52). Inoltre, si deduce che l'entità della distorsione dipende, sostanzialmente, da come vengono accoppiati gli strati originari; l'insieme degli H/2 superstrati dovrebbe essere formato in modo tale che la somma delle differenze al quadrato tra i totali Y_{tu} e Y_{tw} per (t=1,...,H/2) sia minima.

CAPITOLO 11 — VARIANZA DEGLI STIMATORI RAPPORTO

Introduzione

In questo capitolo illustreremo le espressioni delle varianze delle stime indirette, basate sul metodo del rapporto, esaminate nel Capitolo 8.

La trattazione sarà sviluppata nell'ottica della teoria cosiddetta *del primo ordine,* in base alla quale la varianza di una stima rapporto viene derivata come varianza di una sua forma linearizzata, ottenuta mediante lo sviluppo in serie di Taylor limitato ai termini di ordine lineare. Espressioni meno approssimate delle varianze sono state determinate nell'ambito della teoria *del secondo ordine,* che tuttavia non riteniamo opportuno inserire in questa sede; per ulteriori approfondimenti si rinvia alla letteratura specializzata (Sukhatme e Sukhatme, 1970).

Il contenuto di questo capitolo è strettamente legato a quello del Capitolo 10; infatti, come vedremo nel seguito, le espressioni delle stime rapporto coinvolgono nella loro struttura quelle già descritte per le stime dirette.

Prima di affrontare il problema con riferimento al campionamento casuale semplice é utile sottolineare che, nel corso di questi ultimi anni, allo scopo di snellire i procedimenti di calcolo, sono stati studiati alcuni metodi approssimati per il calcolo della varianza che hanno la caratteristica di non tenere conto del disegno di campionamento. Tali metodi sono noti in letteratura (Cochran 1977; Kish e Frankel, 1974) come:

- metodo dello sviluppo in serie di Taylor;
- metodo Jack-Knife;
- metodo delle replicazioni ripetute bilanciate (B.R.R.).

Campionamento casuale semplice

Consideriamo una popolazione di N elementi sulla quale osserviamo due caratteri y ed x ed indichiamo con (Y_i, X_i) le coppie di valori dei due caratteri rilevati sulla generica unità i della popolazione; siano poi:

$$Y = \sum_{i=1}^{N} Y_i \qquad (1)$$

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i \qquad (2)$$

il totale e la media del carattere y nella popolazione;

$$X = \sum_{i=1}^{N} X_i \qquad (3)$$

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \qquad (4)$$

il totale e la media del carattere x nella popolazione.

Consideriamo ora un campione casuale semplice costituito da n elementi estratti senza reimmissione e con probabilità uguali dalle N unità che costituiscono la popolazione.

La stima del totale del carattere y mediante il metodo del rapporto è data da:

$$\hat{\hat{Y}} = \frac{\hat{Y}}{\hat{X}} X \qquad (5)$$

in cui:

$$\hat{Y} = \sum_{i=1}^{n} \frac{N}{n} Y_i \qquad (6)$$

$$\hat{X} = \sum_{i=1}^{n} \frac{N}{n} X_i \qquad (7)$$

rappresentano rispettivamente le stime dirette dei totali Y ed X.

Come è noto per applicare la (5) dobbiamo conoscere il totale del carattere x nella popolazione che rappresenta *l'informazione ausiliaria* esterna.

È possibile dimostrare (Cochran, 1977; Yamane, 1967) che la varianza di $\hat{\hat{Y}}$ è espressa dalla relazione:

$$V(\hat{\hat{Y}}) = N^2 \frac{(N - n)}{N n} S_r^2 \qquad (8)$$

in cui si è posto:

$$S_r^2 = S_y^2 + R^2 S_x^2 - 2RS_{yx} \qquad (9)$$

Nella (9) abbiamo posto:

$$R = \frac{Y}{X} \qquad (10)$$

$$S_y^2 = \frac{1}{N - 1} \sum_{i=1}^{N} (Y_i - \overline{Y})^2 \qquad (11)$$

$$S_x^2 = \frac{1}{N - 1} \sum_{i=1}^{N} (X_i - \overline{X})^2 \qquad (12)$$

$$S_{yx}^2 = \frac{1}{N - 1} \sum_{i=1}^{N} (Y_i - \overline{Y}) \quad (X_i - \overline{X}) \qquad (13)$$

Si fa, tuttavia, presente che, nella generalità dei casi concreti, non disponendo dei valori della popolazione si ha la sola possibilità di determinare una stima campionaria di detta varianza. A tale scopo traduciamo in termini campionari gli elementi che figurano nell'espressione (9).

Una stima corretta di S_y^2 è data da:

$$s_y^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (Y_i - \hat{\overline{Y}})^2 \qquad (14)$$

in cui:

$$\hat{\bar{Y}} = \frac{1}{n} \sum_{i=1}^{n} Y_i \qquad (15)$$

è una stima corretta di $\bar{Y}$.

Analogamente per il carattere x si ottiene:

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{\bar{X}})^2 \qquad (16)$$

$$\hat{\bar{X}} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad (17)$$

Conseguentemente una stima corretta della covarianza S_{yx} è definita dall'espressione

$$s_{yx} = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \hat{\bar{Y}})(X_i - \hat{\bar{X}}) \qquad (18)$$

Infine una stima di R è data dal rapporto

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} \qquad (19)$$

in cui $\hat{Y}$ ed $\hat{X}$ sono le stime dirette dei totali delle due variabili, espresse rispettivamente dalla (6) e dalla (7).

Possiamo ora riscrivere la (9) in termini campionari

$$s_r^2 = s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R} s_{yx} \qquad (20)$$

pertanto una stima della V $(\hat{\hat{Y}})$ è data dall'espressione:

$$\hat{V}(\hat{\hat{Y}}) = N^2 \frac{(N-n)}{Nn} s_r^2 \qquad (21)$$

L'uso dello stimatore $\hat{\hat{Y}}$ conduce in generale a stime più precise di quelle ottenibili mediante lo stimatore diretto $\hat{Y}$. Infatti, per campioni di grandi dimensioni, confrontando le varianze di $\hat{\hat{Y}}$ e di $\hat{Y}$ si ottiene la relazione (Capitolo 8):

$$\rho > \frac{1}{2} \frac{\frac{S_x}{\bar{X}}}{\frac{S_y}{\bar{Y}}}$$

dove ρ è il coefficiente di correlazione fra le due variabili y ed x.

Nel caso in cui $S_x/\bar{X} \simeq S_y/\bar{Y}$ la precedente relazione diventa $\rho > ½$; da ciò si deduce che ogni qual volta la correlazione fra i due caratteri è superiore a 0,5, lo stimatore $\hat{\hat{Y}}$ conduce a stime con varianza minore rispetto a quella dello stimatore $\hat{Y}$.

Infine riteniamo opportuno sottolineare che nell'espressione (21) è presente una distorsione di ordine 1/n, che ovviamente per campioni sufficientemente numerosi (n> 30 unità) diventa trascurabile (Rao, 1968).

Allo scopo di chiarire quanto sopra illustrato, consideriamo, ad esempio, una popolazione costituita da N = 3 contribuenti; siano:

$$Y_1 = 1, \qquad Y_2 = 3, \qquad Y_3 = 15$$

le imposte sul reddito pagate da ogni contribuente e:

$$X_1 = 10, \qquad X_2 = 25, \qquad X_3 = 100$$

i rispettivi redditi individuali.

Estratto un campione di n = 2 contribuenti, siano

$$(Y_2, X_2) \text{ e } (Y_3, X_3)$$

le coppie osservate di valori.

La stima dell'imposta totale, mediante la procedura basata sul metodo del rapporto, è data da:

$$\hat{\hat{Y}} = 19{,}44$$

La stima della varianza di campionamento di $\hat{Y}$ si ottiene attraverso il calcolo delle seguenti quantità:

$$\hat{R} = \frac{27}{187,5} = 0,144$$

$$s_y^2 = 6^2 + 6^2 = 72$$

$$s_x^2 = 37,5^2 + 37,5^2 = 2812,5$$

$$s_{yx} = 6 \cdot 37,5 + 6 \cdot 37,5 = 450$$

$$s_r^2 = 72 + (0,144)^2\ 2812,5 - 2\ (0,144)\ 450 = 0,72$$

Pertanto, dalla relazione (21), si ottiene

$$\hat{V}(\hat{Y}) = 3^2 \cdot \frac{1}{6}\ 0,72 = 1,08$$

Campionamento ad uno stadio stratificato

Nel precedente Capitolo 8 abbiamo illustrato, con riferimento al campionamento ad uno stadio stratificato, i procedimenti di stima indiretti basati sul metodo del rapporto separato e combinato. Per agevolare la comprensione degli sviluppi successivi riscriviamo alcune relazioni utili ai fini della determinazione delle varianze relative ai suddetti metodi di stima.

Indichiamo con Y_{hi} ed X_{hi} la coppia di valori dei due caratteri rilevati sulla unità i appartenente al generico strato h (h= 1,....,H); siano inoltre:

$$Y_h = \sum_{i=1}^{N_h} Y_{hi} \qquad (22)$$

e

$$Y = \sum_{h=1}^{H} Y_h \qquad (23)$$

rispettivamente i totali del carattere y relativi al generico strato h ed alla popolazione.

Analogamente per il carattere x abbiamo:

$$X_h = \sum_{i=1}^{N_h} X_{hi} \qquad (24)$$

$$X = \sum_{h=1}^{H} X_h \qquad (25)$$

Supponiamo inoltre che da ogni strato vengano estratti n_h elementi senza reimmissione e con probabilità uguali.

Come già descritto nel Capitolo 8, con riferimento al disegno ad uno stadio stratificato, possiamo ottenere la stima del totale del carattere y applicando il metodo del rapporto in due modi diversi. Esaminiamo in primo luogo lo stimatore del rapporto separato.

Assumendo noti a priori i totali $X_1..., X_h,...,X_H$ si stima il totale del carattere y mediante l'espressione:

$${}_s\hat{Y} = \sum_{h=1}^{H} \frac{\hat{Y}_h}{\hat{X}_h} X_h \qquad (26)$$

in cui

$$\hat{Y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} Y_{hi} \qquad (27)$$

e

$$\hat{X}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} X_{hi} \qquad (28)$$

Per il calcolo della varianza della stima ${}_s\hat{\bar{Y}}$ poiché l'estrazione delle n_h unità in ciascuno strato è indipendente dall'estrazione negli altri strati, vale la relazione (Cochran, 1977).

$$V({}_s\hat{\bar{Y}}) = \sum_{h=1}^{H} V(\hat{\bar{Y}}_h) = \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} {}_sS_h^2 \qquad (29)$$

in cui si è posto

$${}_sS_h^2 = S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h S_{yxh} \qquad (30)$$

dove:

$$R_h = \frac{\bar{Y}_h}{\bar{X}_h} \qquad (31)$$

$$S_{yh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 \qquad (32)$$

$$S_{xh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2 \qquad (33)$$

$$S_{yxh} = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)(X_{hi} - \bar{X}_h) \qquad (34)$$

Una stima di $V({}_s\hat{\bar{Y}})$ si ottiene traducendo in termini campionari ${}_sS_{xh}^2$ definita dalla (30)

$$\hat{V}({}_s\hat{\bar{Y}}) = \sum_{H=1}^{H} N_h \frac{N_h - n_h}{n_h} {}_ss_h^2 \qquad (35)$$

nella quale la varianza ${}_ss_h^2$, è espressa da:

$${}_ss_h^2 = s_{yh}^2 + \hat{R}_h^2 s_{xh}^2 - 2\hat{R}_h s_{yxh} \qquad (36)$$

in cui:

$$s_{yh}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (Y_{hi} - \hat{\bar{Y}}_h)^2 \qquad (37)$$

$$s_{xh}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (X_{hi} - \hat{\bar{X}}_h)^2 \qquad (38)$$

$$s_{yxh}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (Y_{hi} - \hat{\bar{Y}}_h)(X_{hi} - \hat{\bar{X}}_h) \qquad (39)$$

$$\hat{\bar{Y}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi} \qquad (40)$$

$$\hat{\bar{X}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} X_{hi} \qquad (41)$$

$$\hat{R} = \frac{\hat{Y}_h}{\hat{X}_h} \qquad (42)$$

La formula della varianza $V({}_s\hat{\bar{Y}})$, definita dalla (29), è valida solo se la numerosità del campione in ciascuno strato è abbastanza grande. Infatti, quando gli n_h sono piccoli ed il numero

degli stati H è grande, la distorsione della stima ${}_s\hat{\bar{Y}}$ non è più trascurabile rispetto al suo errore di campionamento.

Passiamo ora a descrivere l'espressione della varianza dello stimatore del rapporto combinato. A tale scopo richiamiamo alcune relazione fondamentali già illustrate nel Capitolo 8. Consideriamo le stime dirette di Y e di X, cioé:

$$\hat{Y} = \sum_{h=1}^{H} \hat{Y}_h \qquad (43)$$

$$\hat{X} = \sum_{h=1}^{H} \hat{X}_h \qquad (44)$$

La stima del rapporto combinato del totale Y é definita da:

$${}_c\hat{\bar{Y}} = \frac{\hat{Y}}{\hat{X}} X \qquad (45)$$

la (45) non richiede la conoscenza di X_h (ammontare del carattere x in ogni strato h), ma solo del totale X nella popolazione; inoltre, essa è molto meno soggetta al rischio di distorsione rispetto a quella ottenuta con il metodo del rapporto separato.

Se l'ampiezza totale del campione n è grande, poiché l'estrazione delle n_h unità in ciascuno strato è indipendente dalle estrazioni negli altri strati, si può dimostrare che la varianza della stima ${}_c\hat{\bar{Y}}$ è data da (Cochran, 1977):

$$V({}_c\hat{\bar{Y}}) = \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} {}_cS^2 \qquad (46)$$

in cui:

$${}_cS^2 = S^2_{yh} + R^2 S^2_{xh} - 2RS_{yxh} \qquad (47)$$

$$R = \frac{Y}{X} \qquad (48)$$

Le espressioni (47) e (30) hanno la stessa struttura, con la differenza che nella (30) compare R_h e nella (47) R.

Una stima della varianza $V({}_c\hat{\bar{Y}})$ è data da:

$$\hat{V}({}_c\hat{\bar{Y}}) = \sum_{h=1}^{H} N_h \frac{N_h - n_h}{n_h} {}_cs^2 \qquad (49)$$

in cui:

$${}_cs^2 = s^2_{yh} + \hat{R}^2 s^2_{xh} - 2\hat{R} s_{yxh} \qquad (50)$$

essendo inoltre:

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} \qquad (51)$$

una stima di R, in cui $\hat{X}$ ed $\hat{Y}$ sono le stime dirette dei totali X e Y dei due caratteri espresse rispettivamente dalla (43) e dalla (44); s^2_{yh}, s^2_{xh} e s_{yxh} sono le stime corrette di S^2_{yh}, S^2_{xh} e S_{yxh} fornite dalle formule (37), (39) e (41).

Si può facilmente dimostrare (Yamane, 1967) che la varianza delle stime ottenute con il metodo del rapporto separato è inferiore a quella delle stime ottenute con il metodo del rapporto combinato se i rapporti R_h sono molto variabili da strato a strato, e se le numerosità n_h sono sufficientemente elevate.

In tutti gli altri casi, in base alle considerazioni esposte precedentemente relative alla distorsione della stima ${}_s\hat{\bar{Y}}$, ottenuta con il metodo del rapporto separato, si preferisce la stima del rapporto combinato (Hansen, Hurwitz, Madow, 1953).

Campionamento a due stadi con stratificazione delle unità primarie

In questo tipo di campionamento la popolazione oggetto di indagine è divisa in H strati, ognuno dei quali contenente N_h unità primarie, ciascuna delle quali comprendente M_{hi} unità secondarie.

Sia Y_{hij} il valore del generico carattere oggetto di studio osservato sulla unità secondaria j appartenente all'unità primaria i dello strato h, e sia X_{hij} il corrispondente valore relativo alla variabile ausiliaria x.

Introduciamo, poi le quantità:

$$Y = \sum_{h=1}^{H} Y_h = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{hij} \qquad (52)$$

$$X = \sum_{h=1}^{H} X_h = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} X_{hij} \qquad (53)$$

che definiscono rispettivamente il totale del carattere y ed il totale del carattere x nella popolazione.

Supponiamo di estrarre, da ciascuno strato, n_h unità primarie senza reimmissione e probabilità uguali; da ciascuna delle n_h unità selezionate, vengono estratte m_{hi} unità di secondo stadio senza reimmissione e con probabilità uguali.

Sappiamo che possiamo ottenere le stime del totale del carattere y applicando il metodo del rapporto in due modi diversi (Vedi Capitolo 8).

Assumendo noti a priori i totali $X_1,\ldots,X_h,\ldots,X_H$, utilizzando lo stimatore del rapporto separato si stima il totale del carattere y mediante l'espressione:

$$_s\hat{Y} = \sum_{h=1}^{H} \frac{\hat{Y}_h}{\hat{X}_h} X_h \qquad (54)$$

in cui $\hat{Y}_h$ ed $\hat{X}_h$ rappresentano le stime corrette di Y_h ed X_h espresse rispettivamente da:

$$\hat{Y}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi}\, Y_{hij} \qquad (55)$$

$$\hat{X}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi}\, X_{hij} \qquad (56)$$

dove

$$K_{hi} = \frac{N_h}{n_h} \frac{M_{hi}}{m_{hi}} \qquad (57)$$

rappresenta il peso base.

Poiché l'estrazione delle n_h unità in ciascuno strato è indipendente da quella negli altri strati, la varianza del totale $_s\hat{Y}$ è data da (Cochran, 1977):

$$V\,(_s\hat{Y}) = \sum_{h=1}^{H} \left[N_h \frac{N_h - n_h}{n_h} {}_sS_h^2 + \frac{N_h}{n_h} \sum_{i=1}^{N_h} M_{hi} \frac{M_{hi} - m_{hi}}{m_{hi}} {}_sS_{hi}^2 \right] \qquad (58)$$

La precedente espressione risulta composta da due parti, ciascuna delle quali esprime una diversa variabilità: il primo addendo esprime la varianza di primo stadio e, poiché dipende da $_sS_h^2$, che rappresenta la variabilità fra i totali delle unità primarie nello strato h, può essere molto elevato se la variabilità tra le dimensioni delle unità primarie è forte.

Il secondo addendo rappresenta la varianza di secondo stadio; esso dipende dai parametri $_sS_{hi}^2$ che indicano la variabilità interna in ciascuna unità primaria.

In particolare la varianza di primo stadio è fornita da:

$$_sS_h^2 = S_{yh}^2 + R_h^2\, S_{xh}^2 - 2R_h\, S_{yxh} \qquad (59)$$

in cui:

$$S_{yh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(Y_{hi} - \frac{Y_h}{N_h} \right)^2 \qquad (60)$$

è la varianza del carattere y nella popolazione appartenente allo strato h; analogamente, per il carattere x si ha:

$$S_{xh}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(X_{hi} - \frac{X_h}{N_h} \right)^2 \qquad (61)$$

L'espressione:

$$S_{yxh} = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(Y_{hi} - \frac{Y_h}{N_h} \right) \left(X_{hi} - \frac{X_h}{N_h} \right) \qquad (62)$$

è la covarianza fra i due caratteri nello strato h; nelle formule appena descritte i simboli Y_{hi}, X_{hi}, Y_h e X_h sono definiti dalle relazioni seguenti:

$$Y_{hi} = \sum_{j=1}^{M_{hi}} Y_{hij} \quad ; \quad Y_h = \sum_{i=1}^{N_h} Y_{hi} \tag{63}$$

$$X_{hi} = \sum_{j=1}^{M_{hi}} X_{hij} \quad ; \quad X_h = \sum_{i=1}^{N_h} X_{hi} \tag{64}$$

ed R_h rappresenta il rapporto fra i due totali Y_h e X_h.
Per quanto riguarda la varianza di secondo stadio si ha:

$${}_sS^2_{hi} = S^2_{yhi} + R^2_h S^2_{xhi} - 2R_h S_{yxhi} \tag{65}$$

Ricordando che $\overline{Y}_{hi}$ ed $\overline{X}_{hi}$ sono le medie aritmetiche dei due caratteri in ciascuna unità primaria i, le varianze e la covarianza che figurano nell'espressione (65) sono date da:

$$S^2_{yhi} = \frac{1}{M_{hi}-1} \sum_{j=1}^{M_{hi}} (Y_{hij} - \overline{Y}_{hi})^2 \tag{66}$$

$$S^2_{xhi} = \frac{1}{M_{hi}-1} \sum_{j=1}^{M_{hi}} (X_{hij} - \overline{X}_{hi})^2 \tag{67}$$

$$S_{yxhi} = \frac{1}{M_{hi}-1} \sum_{j=1}^{M_{hi}} (Y_{hij} - \overline{Y}_{hi})\,(X_{hij} - \overline{X}_{hi}) \tag{68}$$

Una stima della varianza di ${}_s\hat{\hat{Y}}$ si ottiene traducendo in termini campionari ${}_sS^2_h$ e ${}_sS^2_{hi}$, cioè:

$$\hat{V}({}_s\hat{\hat{Y}}) = \sum_{h=1}^{H} \left[N_h \frac{N_h - n_h}{n_h} {}_ss^2_h + \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi} \frac{M_{hi} - m_{hi}}{m_{hi}} {}_ss^2_{hi} \right] \tag{69}$$

Nella (69) ${}_ss^2_h$ rappresenta una stima della varianza di primo stadio ed è espressa da:

$${}_ss^2_h = s^2_{yh} + \hat{R}_h s^2_{xh} - 2\hat{R}_h S_{yxh} \tag{70}$$

i cui addendi sono forniti dalle seguenti espressioni:

$$s^2_{yh} = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (\hat{Y}_{hi} - \hat{\overline{Y}}_h)^2 \tag{71}$$

dove:

$$\hat{\overline{Y}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{Y}_{hi} \tag{72}$$

è una stima corretta di $\overline{Y}_h = Y_h/N_h$;

$$s^2_{xh} = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (\hat{X}_{hi} - \hat{\overline{X}}_h)^2 \tag{73}$$

in cui:

$$\hat{\overline{X}}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{X}_{hi} \tag{74}$$

è una stima corretta di $\overline{X}_h = X_h/N_h$;

$$s_{yxh} = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (\hat{Y}_{hi} - \hat{\overline{Y}}_h)\,(\hat{X}_{hi} - \hat{\overline{X}}_h) \tag{75}$$

Infine:

$$\hat{R}_h = \frac{\hat{Y}_h}{\hat{X}_h} \tag{76}$$

in cui $\hat{Y}_h$ ed $\hat{X}_h$ sono espresse rispettivamente dalla (55) e dalla (56).

Una stima della varianza di secondo stadio ${}_sS^2_{hi}$ è data dall'ultimo termine della (69) ${}_ss^2_{hi}$ che può esplicitarsi come segue:

$${}_ss^2_{hi} = s^2_{yhi} + \hat{R}^2_h s^2_{xhi} - 2\hat{R}_h s_{yxhi} \tag{77}$$

i cui addendi sono descritti dalle espressioni seguenti:

$$s^2_{yhi} = \frac{1}{m_{hi} - 1} \sum_{j=1}^{m_{hi}} (Y_{hij} - \hat{\bar{Y}}_{hi})^2 \tag{78}$$

$$s^2_{yhi} = \frac{1}{m_{hi} - 1} \sum_{j=1}^{m_{hi}} (X_{hij} - \hat{\bar{X}}_{hi})^2 \tag{79}$$

$$s_{yxhi} = \frac{1}{m_{hi} - 1} \sum_{j=1}^{m_{hi}} (Y_{hij} - \hat{\bar{Y}}_{hi})(X_{hij} - \hat{\bar{X}}_{hi}) \tag{80}$$

dove:

$$\hat{\bar{Y}}_{hi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} Y_{hij} \tag{81}$$

$$\hat{\bar{X}}_{hi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} X_{hij} \tag{82}$$

sono rispettivamente le stime corrette di $\bar{Y}_{hi}$ ed $\bar{X}_{hi}$.

Consideriamo ora il caso in cui la stima del totale Y sia ottenuta mediante lo stimatore combinato definito da (Vedi Capitolo 8):

$${}_c\hat{Y} = \frac{\hat{Y}}{\hat{X}} X \tag{83}$$

Nella (83) $\hat{Y}$ ed $\hat{X}$ rappresentano le stime corrette di Y ed X, che sono definite rispettivamente da:

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi} Y_{hij} \tag{84}$$

$$\hat{X} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi} X_{hij} \tag{85}$$

La varianza della stima ${}_c\hat{Y}$ è data da:

$$V({}_c\hat{Y}) = \sum_{h=1}^{H} \left[N_h \frac{N_h - n_h}{n_h} {}_cS^2_h + \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi} \frac{M_{hi} - m_{hi}}{m_{hi}} {}_cS^2_{hi} \right] \tag{86}$$

in cui la varianza di primo stadio ${}_cS^2_h$ è definita da:

$${}_cS^2_h = S^2_{yh} + R^2 S^2_{xh} - 2R S_{yxh} \tag{87}$$

e la varianza di secondo stadio ${}_cS^2_{hi}$ è espressa da:

$${}_cS^2_{hi} = S^2_{yhi} + R^2 S^2_{xhi} - 2R S_{yxhi} \tag{88}$$

Le due espressioni (87) ed (88), hanno la stessa struttura rispettivamente della (59) e della (65) con la differenza che nelle espressioni relative allo stimatore del rapporto combinato compare il termine $R = Y/X$ anziché $R_h = Y_h/X_h$.

Una stima della varianza è fornita da:

$$\hat{V}({}_c\hat{Y}) = \sum_{h=1}^{H} \left[N_h \frac{N_h - n_h}{n_h} {}_cs^2_h + \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi} \frac{M_{hi} - m_{hi}}{m_{hi}} {}_cs^2_{hi} \right] \tag{89}$$

dove:

$$_cs_h^2 = s_{yh}^2 + \hat{R}^2 s_{xh}^2 - 2\hat{R}\, s_{yxh} \qquad (90)$$

in cui $\hat{R} = \hat{Y}/\hat{X}$ è una stima di R e:

s_{yh}^2, s_{xh}^2 e s_{yxh} sono fornite rispettivamente dalle relazioni (71), (73) e (75); infine

$$_cs_{hi}^2 = s_{yhi}^2 + \hat{R}^2 s_{xhi}^2 - 2\hat{R}\, s_{yxhi} \qquad (91)$$

i cui addendi sono definiti dalle espressioni (78), (79) ed (80).

Esaminiamo ora il caso in cui le unità di primo stadio vengono estratte senza reimmissione e probabilità variabili, generalmente proporzionali alle dimensioni delle unità stesse.

Indichiamo con Π_{1hi} la probabilità di inclusione dell'unità primaria i appartenente allo strato h e con $\Pi_{1hii'}$ la probabilità che n unità primarie i ed i', entrambe appartenenti allo strato h, entrino nel campione.

Supponiamo che da ciascuna delle n_h unità primarie venga selezionato un numero di unità di secondo stadio m_{hi} senza reimmissione e con probabilità uguali.

In tale situazione, la stima del totale Y utilizzando lo stimatore del rapporto separato è fornita dalle espressioni (54) (55) e (56) nelle quali tuttavia il peso K_{hi} assume la forma:

$$K_{hi} = \frac{M_{hi}}{\Pi_{1hi}\, m_{hi}} \qquad (92)$$

La varianza di detta stima è (Cochran, 1977):

$$V\,(_s\hat{Y}) = \sum_{h=1}^{H} \Bigg[\sum_{\substack{i=1 \\ i \neq i'}}^{N_h} \sum_{i'=1}^{N_h} (\Pi_{1hi}\,\Pi_{1hi'} - \Pi_{1hii'})\; {}_sS_h^2 + \qquad (93)$$

$$+ \sum_{i=1}^{N_h} \frac{M_{hi}\,(M_{hi} - m_{hi})}{\Pi_{1hi}\, m_{hi}}\; {}_sS_{hi}^2 \Bigg]$$

Si osservi che il primo addendo della (93) è la variabilità dovuta al campionamento di primo stadio, in cui $_sS_h^2$ è data da da:

$$_sS_h^2 = S_{yh}^2 + R_h^2 S_{xh}^2 - 2R_h S_{yxh} \qquad (94)$$

dove si è indicato con:

$$S_{yh}^2 = \sum_{i=1}^{N_h} \sum_{i'=1}^{N_h} \left(\frac{Y_{hi}}{\Pi_{1hi}} - \frac{Y_{hi'}}{\Pi_{1hi'}} \right)^2 \qquad (i \neq i') \qquad (95)$$

la varianza dei totali del carattere y fra le unità primarie dello strato h; analogamente, per il carattere x:

$$S_{xh}^2 = \sum_{i=1}^{N_h} \sum_{i'=1}^{N_h} \left(\frac{X_{hi}}{\Pi_{1hi}} - \frac{X_{hi'}}{\Pi_{1hi'}} \right)^2 \qquad (i \neq i') \qquad (96)$$

Infine:

$$S_{yxh} = \sum_{i=1}^{N_h} \sum_{i'=1}^{N_h} \left(\frac{Y_{hi}}{\Pi_{1hi}} - \frac{Y_{hi'}}{\Pi_{1hi'}} \right) \left(\frac{X_{hi}}{\Pi_{1hi}} - \frac{X_{hi'}}{\Pi_{1hi'}} \right) (i \neq i') \qquad (97)$$

rappresenta la covarianza fra i totali dei due caratteri nello strato h.

Il secondo addendo della (93) è la variabilità dovuta al campionamento di secondo stadio; poiché le unità di secondo stadio vengono selezionate senza reimmissione e probabilità uguali, $_sS_{hi}^2$ è espressa dalla formula (65).

Se le varianze $_sS_h^2$ ed $_sS_{hi}^2$ non sono note, occorre stimarle in base al campione; per quanto riguarda la varianza $_sS_h^2$, una stima consistente è data da:

$$_ss_h^2 = s_{yh}^2 + \hat{R}_h^2\, s_{xh}^2 - 2\,\hat{R}_{syxh} \qquad (98)$$

in cui:

$$s_{yh}^2 = \sum_{i=1}^{n_h} \sum_{i'=1}^{n_h} \left(\frac{\hat{Y}_{hi}}{\Pi_{1hi}} - \frac{\hat{Y}_{hi'}}{\Pi_{1hi'}} \right)^2 \qquad (i \neq i') \qquad (99)$$

dove $\hat{Y}_{hi}$ ed $\hat{Y}_{hi'}$ sono le stime dirette del totale del carattere y relativo alle unità primarie i ed i' dello strato h, fornite da un'epressione del tipo:

$$\hat{Y}_{hd} = \sum_{j=1}^{m_{hd}} K_{hd}\, Y_{hdj} \qquad (d = i, i') \qquad (100)$$

Analogamente, per il carattere x si ha:

$$s^2_{xh} = \sum_{i=1}^{n_h} \sum_{i'=1}^{n_h} \left(\frac{\hat{X}_{hi}}{\Pi_{1hi}} - \frac{\hat{X}_{hi'}}{\Pi_{1hi'}} \right)^2 \qquad (i \neq i') \qquad (101)$$

dove $\hat{X}_{hi}$ ed $\hat{X}_{hi}$, sono stime fornite da espressioni analoghe alla (100), salvo ovviamente la sostituzione di Y_{hdj} con X_{hdj}.

Infine, nell'ultimo termine della (98) è presente una stima della covarianza, S_{yxh} espressa da:

$$s_{yxh} = \sum_{\substack{i=1 \\ (i \neq i')}}^{n_h} \sum_{i'=1}^{n_h} \left(\frac{\hat{Y}_{hi}}{\Pi_{1hi}} - \frac{\hat{Y}_{hi}}{\Pi_{1hi'}} \right) \left(\frac{\hat{X}_{hi}}{\Pi_{1hi}} - \frac{\hat{X}_{hi}}{\Pi_{1hi'}} \right) \qquad (102)$$

Un stima consistente di ${}_sS^2_{hi}$ è fornita dalla (77), in quanto le unità di secondo stadio sono selezionate sempre senza reimmissione e con probabilità uguali.

Conseguentemente una stima della V $({}_s\hat{\bar{Y}})$, definita dalla (93), è data da:

$$\hat{V}\,({}_s\hat{\bar{Y}}) = \sum_{h=1}^{H} \Bigg[\sum_{\substack{i=1 \\ i' \neq i}}^{n_h} \sum_{i'=1}^{n_h} \left(\frac{\Pi_{1hi}\,\Pi_{1hi'} - \Pi_{1hii'}}{\Pi_{1hii'}} \right) {}_ss^2_h + \qquad (103)$$

$$+ \sum_{i=1}^{n_h} \frac{M_{hi}\,(M_{hi} - m_{hi})}{m_{hi}\,\Pi_{1hi}}\, {}_ss^2_{hi} \Bigg]$$

Consideriamo infine il caso in cui le unità di primo stadio sono estratte senza reimmissione e probabilità variabili e viene utilizzata la procedura dello stimatore del rapporto combinato.

Tenendo presente che in tale situazione la stima ${}_c\hat{\bar{Y}}$ è fornita dalle espressioni (83), (84) e (85), nelle quali è presente il peso K_{hi} costruito in base alle probabilità di selezione variabili definito dalla (92).

La varianza della stima ${}_c\hat{\bar{Y}}$ è data da:

$$V\,({}_c\hat{\bar{Y}}) = \sum_{h=1}^{H} \Bigg[\sum_{i=1}^{N_h} \sum_{i'=1}^{N_h} (\Pi_{1hi}\,\Pi_{1hi'} - \Pi_{1hii'})\, {}_sS^2_h + \qquad (104)$$

$$+ \sum_{i=1}^{N_h} \frac{M_{hi}\,(M_{hi} - m_{hi})}{\Pi_{1hi}\,m_{hi}}\, {}_cS^2_{hi} \Bigg] \qquad (i \neq i')$$

in cui:

$${}_cS^2_h = S^2_{yh} + R^2\, S^2_{xh} - 2R\, S_{yxh} \qquad (105)$$

esprime la varianza di primo stadio, che ha la stessa struttura della (94) con la sola eccezione che quest'ultima comprende R_h mentre nella (105) appare R.

Inoltre, poiché le unità di secondo stadio sono estratte senza reimmissione e probabilità uguali, la varianza di secondo stadio è sempre definita dalla (88).

Quindi una stima della varianza V $({}_c\hat{\bar{Y}})$ è fornita da:

$$\hat{V}\,({}_c\hat{\bar{Y}}) = \sum_{h=1}^{H} \Bigg[\sum_{i=1}^{n_h} \sum_{i'=1}^{n_h} \left(\frac{\Pi_{1hi}\,\Pi_{1hi'} - \Pi_{1hii'}}{\Pi_{1hii'}} \right) {}_cs^2_h + \qquad (106)$$

$$+ \sum_{i=1}^{n_h} \frac{M_{hi}\,(M_{hi} - m_{hi})}{m_{hi}\,\Pi_{1hi}}\, {}_cs^2_{hi} \Bigg] \qquad (i \neq i')$$

CAPITOLO 12 – VARIANZA DEGLI STIMATORI RAPPORTO POST-STRATIFICATI

Introduzione

Nel Capitolo 9 abbiamo studiato gli stimatori del rapporto (separato e combinato) post-stratificati che rivestono attualmente un ruolo di grande importanza, essendo alla base delle procedure di stima della maggior parte delle indagini effettuate dai più importanti centri di informazione statistica a livello nazionale ed internazionale (Bureau of Census, INSEE, Istat, ecc.).

In questo capitolo illustreremo le espressioni della varianza e della corrispondente stima degli stimatori in oggetto, che, in un certo senso, costituiscono una naturale estensione degli stimatori rapporto descritti nel Capitolo 8.

Tali stimatori, come abbiamo visto, sono caratterizzati da una struttura formale molto complessa che non consente la determinazione di una espressione corretta della loro varianza; tuttavia, nell'ottica di una filosofia del compromesso fra rigore formale e praticità empirica, sono stati studiati alcuni metodi mediante i quali è possibile pervenire all'ottenimento di espressioni che, nel caso di indagini effettive condotte su larga scala, consentono di determinare il valore della varianza con un elevato grado di approssimazione.

Nel seguito, ai fini della derivazione delle suddette espressioni, faremo uso di un criterio fondato sull'utilizzazione dello sviluppo in serie di Taylor di una funzione, limitato ai termini di ordine lineare; in questo modo, pertanto, trascureremo tutti i termini non lineari dello sviluppo in serie e ciò equivale a dire che approssimeremo gli stimatori del rapporto post-stratificati con le loro forme linearizzate.

Campionamento casuale semplice

Lo studio della varianza dello stimatore ${}_p\hat{Y}$, definito dalla (3) del Capitolo 9, verrà svolto generalizzando un criterio suggerito da alcuni studiosi (Hansen, Hurwtiz e Madow, 1953) per la derivazione della varianza dello stimatore rapporto, nel contesto del campionamento casuale semplice.

Nelle pagine che seguono daremo una traccia di tale studio.

A tale scopo riscriviamo l'espressione esplicita di ${}_p\hat{Y}$:

$$ {}_p\hat{Y} = \sum_{a=1}^{A} \frac{{}_a\hat{Y}}{{}_a\hat{X}} \; {}_aX \qquad (1) $$

dalla quale segue immediatamente che:

$$V({}_p\hat{\bar{Y}}) = \sum_{a=1}^{A} {}_aX^2 \; V\left(\frac{{}_a\hat{Y}}{{}_a\hat{X}}\right) + \tag{2}$$

$$+ \sum_{a=1}^{A} \sum_{b=1}^{A} {}_aX \, {}_bX \; C\left(\frac{{}_a\hat{Y}}{{}_a\hat{X}} \; , \; \frac{{}_b\hat{Y}}{{}_b\hat{X}}\right)$$

con $a \neq b$.

Esaminiamo, in primo luogo, l'espressione del primo addendo della (2). Posto:

$$\frac{{}_a\hat{Y} - {}_aY}{{}_aY} = \Delta y_a \qquad e \qquad \frac{{}_a\hat{X} - {}_aX}{{}_aX} = \Delta x_a \tag{3}$$

si ricavano le relazioni:

$${}_a\hat{Y} = {}_aY\,(1 + \Delta y_a) \qquad e \qquad {}_a\hat{X} = {}_aX\,(1 + \Delta x_a) \tag{4}$$

Indicando poi, per brevità, con W_1 il primo addendo della (2), per definizione di varianza possiamo scrivere:

$$W_1 = \sum_{a=1}^{A} {}_aX^2 \, E\left[\frac{{}_a\hat{Y}}{{}_a\hat{X}} - E\left(\frac{{}_a\hat{Y}}{{}_a\hat{X}}\right)\right]^2 = \tag{5}$$

$$= \sum_{a=1}^{A} {}_aY^2 \; E\left[\frac{1 + \Delta y_a}{1 + \Delta x_a} - E\left(\frac{1 + \Delta y_a}{1 + \Delta x_a}\right)\right]^2$$

dove l'ultimo passaggio consegue dall'impiego delle relazioni (3).

Per calcolare il valore medio della (10) conviene porre:

$$\frac{1}{1 + \Delta x_a} = 1 - \Delta x_a \tag{6}$$

dove $(1 - \Delta x_a)$ rappresenta lo sviluppo in serie di Taylor limitato ai termini di ordine lineare (Ghizzetti, 1965).

Introducendo la (6) nella (5) si ottiene:

$$W_1 = \sum_{a=1}^{A} {}_aY^2 \, E\left[\Delta y_a - \Delta x_a - \Delta y_a \, \Delta x_a + E\,(\Delta y_a \, \Delta x_a)\right]^2 \tag{7}$$

e trascurando i termini di ordine superiore al secondo si ha:

$$W_1 = \sum_{a=1}^{A} \left[V({}_a\hat{Y}) + \left(\frac{{}_aY}{{}_aX}\right)^2 V({}_a\hat{X}) - 2 \left(\frac{{}_aY}{{}_aX}\right) C({}_a\hat{Y}, {}_a\hat{X})\right] \tag{8}$$

in cui:

$$V({}_a\hat{Y}) = \frac{N\,(N - n)}{n} \, {}_aS_y^2 \tag{9}$$

$$V({}_a\hat{X}) = \frac{N\,(N - n)}{n} \, {}_aS_x^2 \tag{10}$$

$$C({}_a\hat{Y}, {}_a\hat{X}) = \frac{N\,(N - n)}{n} \, {}_aS_{xy} \tag{11}$$

$${}_aS_y^2 = \frac{1}{N - 1} \sum_{i=1}^{N} \left({}_aY_i - \frac{1}{N} \sum_{i=1}^{N} {}_aY_i\right)^2 \tag{12}$$

$$ {}_aS_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left({}_aX_i - \frac{1}{N} \sum_{i=1}^{N} {}_aX_i \right)^2 \qquad (13) $$

$$ {}_aS_{xy}^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left({}_aY_i - \frac{1}{N} \sum_{i=1}^{N} {}_aY_i \right) \left({}_aX_i - \frac{1}{N} \sum_{i=1}^{N} {}_aX_i \right) \qquad (14) $$

Introducendo poi la (9), (10) e (11) nella (8), W_1 può riscriversi nella forma equivalente:

$$ W_1 = \frac{N(N-n)}{n(N-1)} \sum_{a=1}^{A} \sum_{i=1}^{N} \left[\left({}_aY_i - \frac{1}{N} \sum_{i=1}^{N} {}_aY_i \right)^2 + \right. $$

$$ + {}_aR^2 \left({}_aX_i - \frac{1}{N} \sum_{i=1}^{N} {}_aX_i \right)^2 + $$

$$ \left. - 2\,{}_aR \left({}_aY_i - \frac{1}{N} \sum_{i=1}^{N} {}_aY_i \right) \left({}_aX_i - \frac{1}{N} \sum_{i=1}^{N} {}_aX_i \right) \right] \qquad (15) $$

nella quale si è posto:

$$ {}_aR = \frac{{}_aY}{{}_aX} \qquad (16) $$

Ponendo infine:

$$ {}_aD_i = {}_aY_i - {}_aR\,{}_aX_i \qquad (17) $$

è possibile scrivere la (15) nella forma più compatta espressa da:

$$ W_1 = \frac{N}{n} \frac{N-n}{N-1} \sum_{a=1}^{A} \sum_{i=1}^{N} \left[{}_aD_i - \frac{1}{N} \sum_{i=1}^{N} {}_aD_i \right]^2 = \qquad (18) $$

$$ = \frac{N}{n} \frac{N-n}{N-1} \sum_{a=1}^{A} \sum_{i=1}^{N} {}_aD_i^2 $$

poiché:

$$ \sum_{i=1}^{N} {}_aD_i = 0 \qquad (19) $$

A questo punto possiamo passare allo studio del secondo addendo della (2), che per brevità indicheremo con W_2.

Sulla base delle relazioni (3) si può scrivere:

$$ W_2 = \sum_{a=1}^{A} \sum_{b=1}^{A} {}_aX\,{}_bX \; E\left[\frac{{}_a\hat{Y}}{{}_a\hat{X}} - E\left(\frac{{}_a\hat{Y}}{{}_a\hat{X}} \right) \right] \left[\frac{{}_b\hat{Y}}{{}_b\hat{X}} - E\left(\frac{{}_b\hat{Y}}{{}_b\hat{X}} \right) \right] = \qquad (20) $$

$$ = \sum_{a=1}^{A} \sum_{b=1}^{A} {}_aY\,{}_bY \quad E\left[\frac{1+\Delta y_a}{1+\Delta x_a} - E\left(\frac{1+\Delta y_a}{1+\Delta x_a} \right) \right] \cdot $$

$$ \left[\frac{1+\Delta y_b}{1+\Delta x_b} - \quad E\left(\frac{1+\Delta y_b}{1+\Delta x_b} \right) \right] $$

Introducendo poi la (6) e l'analoga relazione riferita al post-strato b, segue che:

$$W_2 = \sum_{a=1}^{A} \sum_{b=1}^{A} {}_aY\ {}_bY\ E\ [\Delta y_a - \Delta x_a - \Delta y_a\ \Delta x_a + E\ (\Delta y_a\ \Delta x_a)\ ] \cdot$$

$$\cdot\ [\Delta y_b - \Delta x_b - \Delta y_b\ \Delta x_b + E\ (\Delta y_b\ \Delta x_b)] = \tag{21}$$

$$= \sum_{a=1}^{A} \sum_{b=1}^{A} \left[C\ ({}_a\hat{Y},\ {}_b\hat{Y}) - \left(\frac{{}_bY}{{}_bX} \right) C\ ({}_a\hat{Y},\ {}_b\hat{X}) + \right.$$

$$\left. - \left(\frac{{}_aY}{{}_aX} \right) C\ ({}_a\hat{X},\ {}_b\hat{Y}) + \frac{{}_aY\ {}_bY}{{}_aX\ {}_bX} C\ ({}_a\hat{X},\ {}_b\hat{X}) \right]$$

Sostituendo nella (21) le forme esplicite delle covarianze, che possono ottenersi agevolmente mediante semplice adattamento formale della (11), segue che:

$$W_2 = \frac{N}{n} \frac{N-n}{N-1} \sum_{a=1}^{A} \sum_{b=1}^{A} \sum_{i=1}^{N} {}_aD_i\ {}_bD_i \quad (a \neq b) \tag{22}$$

Riunendo, infine, la (19) e la (22) segue che un'utile approssimazione della varianza $V\ ({}_p\hat{Y})$ è definita dall'espressione:

$$V\ ({}_p\hat{Y}) = \frac{N}{n} \frac{N-n}{N-1} \sum_{i=1}^{N} \left(\sum_{a=1}^{A} {}_aD_i \right)^2 \tag{23}$$

La (23) è estremamente interessante dal punto di vista calcolatorio in quanto, evitando il calcolo delle covarianze fra i diversi post-strati, consente in modo abbastanza agevole la determinazione della varianza $V\ ({}_p\hat{Y})$.

Infine, è possibile dimostrare (Cochran, 1977) che una stima distorta benché consistente della (23) è fornita dall'espressione:

$$\hat{V}\ ({}_p\hat{Y}) = \frac{N}{n} \frac{N-n}{n-1} \sum_{i=1}^{n} \left(\sum_{a=1}^{A} {}_a\hat{D}_i \right)^2 \tag{24}$$

in cui:

$${}_a\hat{D}_i = {}_aY_i - {}_a\hat{R}\ {}_aX_i \tag{25}$$

$${}_a\hat{R} = \frac{{}_a\hat{Y}}{{}_a\hat{X}} \tag{26}$$

risultando, in analogia alla (19), anche:

$$\sum_{i=1}^{n} {}_a\hat{D}_i = 0 \tag{27}$$

Campionamento ad uno stadio stratificato

Affrontiamo ora, nel contesto del disegno di campionamento ad uno stadio stratificato, la descrizione delle espressioni della varianza degli stimatori del rapporto (separato e combinato) post-stratificati, illustrati nel Capitolo 9.

Si aprono al riguardo due possibilità. Da un lato, per il generico strato elementare h (h = 1,...:, H) si può utilizzare un procedimento simile a quello finora svolto, sommando poi opportunamente le varianze degli strati semplici per ottenere la varianza del campione stratificato.

Dall'altro si può seguire una via più rapida fondata sul semplice adattamento formale delle espressioni appena illustrate.

Riteniamo pertanto utile seguire quest'ultimo procedimento.

Esaminiamo in primo luogo lo stimatore del rapporto separato post-stratificato, definito dalla (9) del Capitolo 9 e che riteniamo tuttavia opportuno riscrivere:

$${}_{ps}\hat{\bar{Y}} = \sum_{a=1}^{A} \sum_{h=1}^{H} \frac{{}_a\hat{Y}_h}{{}_a\hat{X}_h} {}_aX_h \tag{28}$$

Ai fini della determinazione della varianza di ${}_{ps}\hat{\bar{Y}}$ conviene anzitutto porre la (28) nella forma:

$${}_{ps}\hat{\bar{Y}} = \sum_{h=1}^{H} \sum_{a=1}^{A} \frac{{}_a\hat{Y}_h}{{}_a\hat{X}_h} {}_aX_h \tag{29}$$

dalla quale si ottiene:

$$V\,({}_{ps}\hat{\bar{Y}}) = \sum_{h=1}^{H} V\left(\sum_{a=1}^{A} \frac{{}_a\hat{Y}_h}{{}_a\hat{X}_h}\,{}_aX_h\right) \qquad (30)$$

L'espressione fra parentesi tonde a secondo membro della (30), relativa al generico strato h, è formalmente e statisticamente uguale alla (1) scritta a proposito del campionamento semplice; quindi per la varianza di tale espressione potremo riferirci alla (23) salvo l'aggiunta dell'indice h.

Si ottiene pertanto:

$$V\,({}_{ps}\hat{\bar{Y}}) = \sum_{h=1}^{H} \frac{N_h}{n_h}\frac{N_h - n_h}{N_h - 1} \sum_{i=1}^{N_h} \left(\sum_{a=1}^{A} {}_aD_{hi} - \frac{1}{N_h}\sum_{a=1}^{A}\sum_{i=1}^{N_h} {}_aD_{hi}\right)^2 =$$

$$= \sum_{h=1}^{H} \frac{N_h}{n_h}\frac{N_h - n_h}{N_h - 1} \sum_{i=1}^{N_h} \left(\sum_{a=1}^{A} {}_aD_{hi}\right)^2 \qquad (31)$$

in cui:

$${}_aD_{hi} = {}_aY_{hi} - \frac{{}_aY_h}{{}_aX_h}\,{}_aX_{hi} \qquad (32)$$

$$\sum_{i=1}^{N_h} {}_aD_{hi} = 0 \qquad (33)$$

$${}_aY_h = \sum_{i=1}^{N_h} {}_aY_{hi} \quad ; \quad {}_aX_h = \sum_{i=1}^{N_h} {}_aX_{hi} \qquad (34)$$

In definitiva, una stima consistente anche se distorta della (31) è data, in analogia alla (24), dall'espressione:

$$\hat{V}\,({}_{ps}\hat{\bar{Y}}) = \sum_{h=1}^{H} \frac{N_h}{n_h}\frac{N_h - n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\sum_{a=1}^{A} {}_a\hat{D}_{hi} - \frac{1}{n_h}\sum_{a=1}^{A}\sum_{i=1}^{n_h} {}_a\hat{D}_{hi}\right)^2 =$$

$$= \sum_{h=1}^{H} \frac{N_h}{n_h}\frac{N_h - n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\sum_{a=1}^{A} {}_a\hat{D}_{hi}\right)^2 \qquad (35)$$

dove:

$${}_a\hat{D}_{hi} = {}_aY_{hi} - \frac{{}_a\hat{Y}_h}{{}_a\hat{X}_h}\,{}_aX_{hi} \quad ; \quad \sum_{i=1}^{n_h} {}_a\hat{D}_{hi} = 0 \qquad (36)$$

$${}_a\hat{Y}_h = \sum_{i=1}^{n_h} \frac{N_h}{n_h}\,{}_aY_{hi} \quad ; \quad {}_a\hat{X}_h = \sum_{i=1}^{n_h} \frac{N_h}{n_h}\,{}_aX_{hi} \qquad (37)$$

Le formule istituite per la varianza dello stimatore del rapporto separato post-stratificato ${}_{ps}\hat{\bar{Y}}$ possono ora adattarsi agevolmente per lo stimatore combinato espresso da:

$${}_{pc}\hat{\bar{Y}} = \sum_{a=1}^{A} \frac{{}_a\hat{Y}}{{}_a\hat{X}}\,{}_aX \qquad (38)$$

in cui:

$${}_a\hat{Y} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} {}_aY_{hi} \quad ; \quad {}_a\hat{X} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} {}_aX_{hi} \qquad (39)$$

rappresentano rispettivamente le stime dirette di:

$${}_aY = \sum_{h=1}^{H}\sum_{i=1}^{N_h} {}_aY_{hi} \quad ; \quad {}_aX = \sum_{h=1}^{H}\sum_{i=1}^{N_h} {}_aX_{hi} \qquad (40)$$

Dalla (31) si ottiene immediatamente:

$$V\left({}_{pc}\hat{\bar{Y}}\right) = \sum_{h=1}^{H} \frac{N_h}{n_h}\frac{N_h - n_h}{N_h - 1} \sum_{i=1}^{N_h} \left(\sum_{a=1}^{A} {}_aD'_{hi} - \frac{1}{N_h} \sum_{a=1}^{A} \sum_{i=1}^{N_h} {}_aD'_{hi} \right)^2 \quad (41)$$

in cui:

$${}_aD'_{hi} = {}_aY_{hi} - \frac{{}_a\bar{Y}}{{}_a\bar{X}} {}_aX_{hi} \quad (42)$$

Osserviamo che, a differenza dei casi precedentemente esaminati, risulta ora:

$$\sum_{i=1}^{N_h} {}_aD'_{hi} \neq o$$

In conclusione, una stima consistente della (40) è fornita da:

$$\hat{V}\left({}_{pc}\hat{\bar{Y}}\right) = \sum_{h=1}^{H} \frac{N_h}{n_h}\frac{N_h - n_h}{N_h - 1} \sum_{i=1}^{n_h} \left(\sum_{a=1}^{A} {}_a\hat{D}'_{hi} - \frac{1}{n_h} \sum_{a=1}^{A} \sum_{i=1}^{n_h} {}_a\hat{D}'_{hi} \right)^2 \quad (43)$$

dove:

$${}_a\hat{D}'_{hi} = {}_aY_{hi} - \frac{{}_a\hat{\bar{Y}}}{{}_a\hat{\bar{X}}} {}_aX_{hi} \quad (44)$$

Campionamento a due stadi con stratificazione delle unità primarie

Affrontiamo, infine, lo studio della varianza dello stimatore del rapporto (separato e combinato) post-stratificato, nel contesto del campionamento a due stadi stratificato al livello delle unità primarie.

Un modo di procedere immediato, che verrà adottato nel seguito, è quello di eseguire l'adattamento formale delle espressioni ottenute precedentemente.

Esaminiamo, in primo luogo, lo stimatore del rapporto separato post-stratificato definito da:

$${}_{ps}\hat{\bar{Y}} = \sum_{a=1}^{A} \sum_{h=1}^{H} \frac{{}_a\hat{Y}_h}{{}_a\hat{X}_h} {}_aX_h \quad (45)$$

in cui:

$${}_a\hat{Y}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi} \; {}_aY_{hij} \quad (46)$$

$${}_a\hat{X}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi} \; {}_aX_{hij} \quad (47)$$

indicano rispettivamente le stime corrette dei totali:

$${}_aY_h = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} {}_aY_{hij} \quad ; \quad {}_aX_h = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} {}_aX_{hij} \quad (48)$$

Ai fini degli sviluppi algebrici successivi è opportuno richiamare anche le espressioni esplicite del peso K_{hi} relative ai due schemi probabilistici di selezione delle unità di primo stadio esaminati nei precedenti capitoli.

In generale, possiamo scrivere:

$$K_{hi} = K_{1hi} \; K_{2hi} \quad (49)$$

in cui il primo fattore è il reciproco della probabilità di inclusione dell'unità primaria i appartenente al generico strato h ed il secondo fattore è il reciproco della probabilità di inclusione dell'unità secondaria j dentro l'unità primaria i.

Il primo fattore assume la forma (Capitoli 5 e 6):

$$K_{1hi} = \frac{1}{\Pi_{1hi}} = \frac{N_h}{n_h} \quad o \quad K_{1hi} = \frac{1}{\Pi_{1hi}} = \frac{1}{n_h A_{hi}} \sum_{i=1}^{N_h} A_{hi} \quad (50)$$

a seconda che le n_h siano estratte senza reimmissione e probabilità uguali oppure senza reimmissione e probabilità proporzionale alla dimensione.

Per il secondo fattore si ha invece la sola forma:

$$K_{2hi} = \frac{1}{\Pi_{2hi}} = \frac{M_{hi}}{m_{hi}} \quad (51)$$

Poiché dalle (50) si deduce che la prima delle due espressioni si può ottenere come caso particolare della seconda, nel seguito ci limiteremo allo studio della varianza limitatamente al caso in

cui le unità primarie vengono selezionate con probabilità proporzionale alla dimensione e senza reimmissione e le unità secondarie con probabilità uguali e senza reimmissione.

Particolarizzando questa varianza si può facilmente ottenere l'espressione della varianza relativa al caso in cui le unità primarie vengono estratte senza reimmissione e probabilità uguali.

Ciò premesso, dalla (45) si ottiene:

$$V\left({}_{ps}\hat{\bar{Y}}\right) = V\left(\sum_{a=1}^{A}\sum_{h=1}^{H}\frac{{}_a\hat{Y}_h}{{}_a\hat{X}_h}\,{}_aX_h\right) = \sum_{h=1}^{H} V\left(\sum_{a=1}^{A}\frac{{}_a\hat{Y}_h}{{}_a\hat{X}_h}\,{}_aX_h\right) =$$

$$= \sum_{h=1}^{H}\left[\sum_{a=1}^{A}{}_aX_h^2\; V\left(\frac{{}_a\hat{Y}_h}{{}_a\hat{X}_h}\right) +\right.$$

$$\left. + \sum_{a=1}^{A}\sum_{b=1}^{A}{}_aX_h\;{}_bX_h\; C\left(\frac{{}_a\hat{Y}_h}{{}_a\hat{X}_h}\,,\,\frac{{}_b\hat{Y}_h}{{}_b\hat{X}_h}\right)\right] \qquad (52)$$

con $a \neq b$.

Introduciamo le relazioni:

$$\frac{{}_a\hat{Y}_h - {}_aY_h}{{}_aY_h} = \Delta\, y_{ha} \qquad \text{e} \qquad \frac{{}_a\hat{X}_h - {}_aX_h}{{}_aX_h} = \Delta\, x_{ha} \qquad (53)$$

Mediante un procedimento simile a quello che ha condotto alla (8) e alla (21) si ottiene che la (52) si può porre nella forma (Russo, 1988 b):

$$V\left({}_{ps}\hat{\bar{Y}}\right) = \sum_{h=1}^{H}\left\{\sum_{a=1}^{A}\left[V\left({}_a\hat{Y}_h\right) + \left(\frac{{}_aY_h}{{}_aX_h}\right)^2 V\left({}_a\hat{X}_h\right) + \right.\right. \qquad (54)$$

$$\left. - 2\left(\frac{{}_aY_h}{{}_aX_h}\right) C\left({}_a\hat{Y}_h\,,\,{}_a\hat{X}_h\right)\right] + \sum_{a=1}^{A}\sum_{b=1}^{A}\left[C\left({}_a\hat{Y}_h\,,\,{}_b\hat{Y}_h\right) +\right.$$

$$\left. - \left(\frac{{}_bY_h}{{}_bX_h}\right) C\left({}_a\hat{Y}_h\,,\,{}_b\hat{X}_h\right) - \left(\frac{{}_aY_h}{{}_aX_h}\right) C\left({}_a\hat{X}_h\,,\,{}_b\hat{Y}_h\right) +\right.$$

$$\left.\left. + \frac{{}_aY_h\;{}_bY_h}{{}_aX_h\;{}_bX_h} C\left({}_a\hat{X}\,,\,{}_b\hat{X}_h\right)\right]\right\}$$

in cui ad esempio, con $i' > i$, si ha:

$$V\left({}_a\hat{Y}_h\right) = \sum_{i=1}^{N_h}\sum_{i'>i}^{N_h}\left(\Pi_{1hi}\,\Pi_{1hi'} - \Pi_{1hii'}\right)\left(\frac{{}_aY_{hi}}{\Pi_{1hi}} - \frac{{}_aY_{hi'}}{\Pi_{1hi'}}\right)^2 +$$

$$+ \sum_{i=1}^{N_h}\frac{M_{hi}^2}{\Pi_{1hi}\,m_{hi}}\left(1 - \frac{m_{hi}}{M_{hi}}\right){}_aS_{yhi}^2 \qquad (55)$$

$$C\left({}_a\hat{Y}_h\,,\,{}_a\hat{X}_h\right) = \sum_{i=1}^{N_h}\sum_{i'>i}^{N_h}\left(\Pi_{1hi}\,\Pi_{1hi'} - \Pi_{1hii'}\right)\left(\frac{{}_aY_{hi}}{\Pi_{1hi}} - \frac{{}_aY_{hi'}}{\Pi_{1hi'}}\right)\cdot$$

$$\left(\frac{{}_aX_{hi}}{\Pi_{1hi}} - \frac{{}_aX_{hi'}}{\Pi_{1hi'}}\right) + \sum_{i=1}^{N_h}\frac{M_{hi}^2}{\Pi_{1hi}\,m_{hi}}\left(1 - \frac{m_{hi}}{M_{hi}}\right){}_aS_{yxhi} \qquad (56)$$

$${}_aS_{yhi}^2 = \frac{1}{M_{hi} - 1}\sum_{j=1}^{M_{hi}}\left({}_aY_{hij} - \frac{1}{M_{hi}}\sum_{j=1}^{M_{hi}}{}_aY_{hij}\right)^2 \qquad (57)$$

$${}_aS_{yxhi} = \frac{1}{M_{hi}-1}\sum_{j=1}^{M_{hi}}\left({}_aY_{hij} - \frac{1}{M_{hi}}\sum_{j=1}^{M_{hi}}{}_aY_{hij}\right)\left({}_aX_{hij} - \frac{1}{M_{hi}}\sum_{j=1}^{M_{hi}}{}_aX_{hij}\right) \qquad (58)$$

Le espressioni di $V\left({}_a\hat{X}_h\right)$ e delle rimanenti covarianze comprese nella (54) possono ottenersi attraverso un semplice adattamento formale della (55) e della (56).

Introducendo tali espressioni nella (54) e ponendo:

$${}_aD_{hi} = {}_aY_{hi} - \frac{{}_aY_h}{{}_aX_h}\,{}_aX_{hi} \quad ; \quad {}_aD_{hij} = {}_aY_{hij} - \frac{{}_aY_h}{{}_aX_h}\,{}_aX_{hij} \qquad (59)$$

la (54) può riscriversi nella forma (Russo, 1988):

$$V\left({}_{ps}\hat{Y}\right) = \sum_{h=1}^{H}\left[\sum_{i=1}^{N_h}\sum_{i'>i}^{N_h}\left(\Pi_{1hi}\,\Pi_{1hi'} - \Pi_{1hii'}\right)\left(\sum_{a=1}^{A}\frac{{}_aD_{hi}}{\Pi_{1hi}} - \sum_{a=1}^{A}\frac{{}_aD_{hi'}}{\Pi_{1hi'}}\right)^2 + \right. \tag{60}$$

$$\left. + \sum_{i=1}^{N_h}\frac{M_{hi}}{\Pi_{1hi}}\,\frac{M_{hi}-m_{hi}}{m_{hi}\,(M_{hi}-1)}\sum_{j=1}^{M_{hi}}\left(\sum_{a=1}^{A}{}_aD_{hij} - \frac{1}{M_{hi}}\sum_{a=1}^{A}\sum_{j=1}^{M_{hi}}{}_aD_{hij}\right)^2\right]$$

Infine, una stima consistente della (60) è fornita dall'espressione:

$$\hat{V}\left({}_{ps}\hat{Y}\right) = \sum_{h=1}^{H}\left[\sum_{i=1}^{n_h}\sum_{i'>1}^{n_h}\left(\frac{\Pi_{1hi}\quad\Pi_{1hi'}}{\Pi_{1hii'}} - 1\right)\cdot\right.$$

$$\cdot\left(\sum_{a=1}^{A}\frac{{}_a\hat{D}_{hi}}{\Pi_{1hi}} - \sum_{a=1}^{A}\frac{{}_a\hat{D}_{hi'}}{\Pi_{1hi'}}\right)^2 + \sum_{i=1}^{n_h}\frac{M_{hi}}{\Pi_{1hi}}\,\frac{M_{hi}-m_{hi}}{m_{hi}\,(m_{hi}-1)}\cdot$$

$$\left.\cdot\sum_{j=1}^{m_{hi}}\left(\sum_{a=1}^{A}{}_a\hat{D}_{hij} - \frac{1}{m_{hi}}\sum_{a=1}^{A}\sum_{j=1}^{m_{hi}}{}_a\hat{D}_{hij}\right)^2\right] \tag{61}$$

in cui si è fatto uso dello stimatore di Yates e Grundy (1953), avendo inoltre posto:

$${}_a\hat{D}_{hi} = {}_a\hat{Y}_{hi} - \frac{{}_a\hat{Y}_h}{{}_a\hat{X}_h}\,{}_a\hat{X}_{hi} \quad\text{e}\quad {}_a\hat{D}_{hij} = {}_aY_{hij} - \frac{{}_a\hat{Y}_h}{{}_a\hat{X}_h}\,{}_a\hat{X}_{hij} \tag{62}$$

essendo:

$${}_a\hat{Y}_{hi} = \frac{M_{hi}}{m_{hi}}\sum_{j=1}^{m_{hi}}{}_aY_{hij} \qquad\text{e}\qquad {}_a\hat{X}_{hi} = \frac{M_{hi}}{m_{hi}}\sum_{j=1}^{m_{hi}}{}_aX_{hij} \tag{63}$$

le stime dirette rispettivamente di:

$${}_aY_{hi} = \sum_{j=1}^{M_{hi}}{}_aY_{hij} \qquad\text{e}\qquad {}_aX_{hi} = \sum_{j=1}^{M_{hi}}{}_aX_{hij} \tag{64}$$

Per quanto riguarda poi la varianza dello stimatore del rapporto combinato post-stratificato definito dall'espressione (Capitolo 9):

$${}_{pc}\hat{Y} = \sum_{a=1}^{A}\frac{{}_a\hat{Y}}{{}_a\hat{X}}\,{}_aX \tag{65}$$

si può seguire un procedimento analogo a quello adottato per la derivazione della varianza dello stimatore ${}_{ps}\hat{Y}$.

Si ottiene un'espressione con struttura uguale alla (60), salvo il fatto che in luogo delle (59) figurano le quantità:

$${}_aD'_{hi} = {}_aY_{hi} - \frac{{}_aY}{{}_aX}\,{}_aX_{hi} \quad\text{e}\quad {}_aD'_{hij} = {}_aY_{hij} - \frac{{}_aY}{{}_aX}\,{}_aX_{hij} \tag{66}$$

Anche per la stima della varianza di ${}_{pc}\hat{Y}$ possiamo riferirci alla (61) introducendo al posto delle (62) le espressioni:

$${}_a\hat{D}'_{hi} = {}_a\hat{Y}_{hi} - \frac{{}_a\hat{Y}}{{}_a\hat{X}}\,{}_a\hat{X}_{hi} \quad\text{e}\quad {}_a\hat{D}'_{hij} = {}_aY_{hij} - \frac{{}_a\hat{Y}}{{}_a\hat{X}}\,{}_aX_{hij} \tag{67}$$

Fino ad ora abbiamo descritto le espressioni della varianza degli stimatori ${}_{ps}\hat{Y}$ e ${}_{pc}\hat{Y}$ nel caso in cui le unità di primo stadio vengono selezionate con probabilità proporzionale alla dimensione.

Sebbene tale meccanismo probabilistico di selezione sia quello di uso più frequente nell'ambito delle indagini a due o più stadi di campionamento, ci sembra tuttavia utile descrivere anche le espressioni della varianza nel caso in cui le unità primarie siano selezionate con probabilità uguali.

In tal caso sappiamo che le probabilità di inclusione sono espresse da:

$$\Pi_{1hi} = \frac{n_h}{N_h} \qquad\text{e}\qquad \Pi_{1hii'} = \frac{n_h}{N_h}\,\frac{n_h-1}{N_h-1} \tag{68}$$

in base alle quali la (60) può riscriversi nella forma:

$$V\,({}_{ps}\hat{\bar{Y}}) = \sum_{h=1}^{H} \left[\left(1 - \frac{N_h}{n_h} \frac{n_h - 1}{N_h - 1} \right) \sum_{i=1}^{N_h} \sum_{i' > i}^{N_h} \left(\sum_{a=1}^{A} {}_aD_{hi} - \sum_{a=1}^{A} {}_aD_{hi'} \right)^2 + \right.$$

$$\left. + \frac{N_h}{n_h} \sum_{i=1}^{N_h} \frac{M_{hi}\,(M_{hi} - m_{hi})}{m_{hi}\,(M_{hi} - 1)} \sum_{j=1}^{M_{hi}} \left(\sum_{a=1}^{A} {}_aD_{hij} - \frac{1}{M_{hi}} \sum_{a=1}^{A} \sum_{j=1}^{M_{hi}} {}_aD_{hij} \right)^2 \right] \quad (69)$$

Ricordiamo ora la ben nota relazione:

$$^2\Delta^2_R = 2\ \sigma^2 \quad (70)$$

in cui $^2\Delta^2_R$ indica il quadrato della differenza quadratica media con ripetizione e σ^2 la varianza.

Tenedo presente la (70) e la doppia sommatoria del primo membro della (69) possiamo scrivere:

$$\frac{1}{N_h} \sum_{i=1}^{N_h} \sum_{i' > i}^{N_h} \left(\sum_{a=1}^{A} {}_aD_{hi} - \sum_{a=1}^{A} {}_aD_{hi'} \right)^2 =$$

$$= \sum_{i=1}^{N_h} \left(\sum_{a=1}^{A} {}_aD_{hi} - \frac{1}{N_h} \sum_{a=1}^{A} \sum_{i=1}^{N_h} {}_aD_{hi} \right)^2 \quad (71)$$

In virtù della (71) la (60) diviene:

$$V\,({}_{ps}\hat{\bar{Y}}) = \sum_{h=1}^{H} \left[\frac{N_h}{n_h} \frac{N_h - n_h}{N_h - 1} \sum_{i=1}^{N_h} \left(\sum_{a=1}^{A} {}_aD_{hi} - \sum_{a=1}^{A} \frac{1}{N_h} \sum_{i=1}^{N_h} {}_aD_{hi} \right)^2 + \right.$$

$$\left. + \frac{N_h}{n_h} \sum_{i=1}^{N_h} \frac{M_{hi}\,(M_{hi} - m_{hi})}{m_{hi}\,(M_{hi} - 1)} \sum_{j=1}^{M_{hi}} \left(\sum_{a=1}^{A} {}_aD_{hij} - \frac{1}{M_{hi}} \sum_{a=1}^{A} \sum_{j=1}^{M_{hi}} {}_aD_{hij} \right)^2 \right] \quad (72)$$

Seguendo un procedimento simile a quello che ha condotto alla (72) si possono poi ottenere le espressioni della varianza di ${}_{pc}\hat{\bar{Y}}$ e delle stime di $V\,({}_{ps}\hat{\bar{Y}})$ e $V\,({}_{pc}\hat{\bar{Y}})$.

Forme alternative e casi particolari

A conclusione di questo capitolo illustreremo alcune espressioni alternative della stima della varianza, che sono di un certo interesse in quanto da un lato si presentano in una foma più adatta ai fini del calcolo mediante elaboratori elettronici e dall'altra si riferiscono a disegni campionari molto diffusi nell'ambito delle indagini condotte su larga scala.

La trattazione sarà limitata allo stimatore ${}_{pc}\hat{\bar{Y}}$ nel contesto del campionamento ad uno stadio stratificato, ma le considerazioni che svolgeremo possono facilmente essere estese agli altri tipi di campionamento considerati nel presente volume e allo stimatore ${}_{ps}\hat{\bar{Y}}$.

Consideriamo in primo luogo l'espressione (43). Sviluppando il quadrato si ottiene:

$$\hat{V}\,({}_{pc}\hat{\bar{Y}}) = \sum_{h=1}^{H} \frac{N_h}{n_h} \frac{N_h - n_h}{n_h - 1} \sum_{i=1}^{n_h} \left[\left(\sum_{a=1}^{A} {}_a\hat{D}'_{hi} \right)^2 + \right.$$

$$\left. \frac{1}{n_h^2} \left(\sum_{i=1}^{n_h} \sum_{a=1}^{A} {}_a\hat{D}'_{hi} \right)^2 - 2 \frac{1}{n_h} \left(\sum_{a=1}^{A} {}_a\hat{D}'_{hi} \right) \left(\sum_{i=1}^{n_h} \sum_{a=1}^{A} {}_a\hat{D}'_{hi} \right) \right] =$$

$$= \sum_{h=1}^{H} \frac{N_h}{n_h} \frac{N_h - n_h}{n_h - 1} \left[\sum_{i=1}^{n_h} \left(\sum_{a=1}^{A} {}_a\hat{D}'_{hi} \right)^2 + \frac{1}{n_h} \left(\sum_{i=1}^{n_h} \sum_{a=1}^{A} {}_a\hat{D}'_{hi} \right)^2 + \right.$$

$$\left. - \frac{2}{n_h} \left(\sum_{i=1}^{n_h} \sum_{a=1}^{A} {}_a\hat{D}'_{hi} \right)^2 \right] = \sum_{h=1}^{H} \frac{N_h}{n_h} \frac{N_h - n_h}{n_h - 1} \left[\sum_{i=1}^{n_h} \left(\sum_{a=1}^{A} {}_a\hat{D}'_{hi} \right)^2 + \right.$$

$$\left. - \frac{1}{n_h} \left(\sum_{i=1}^{n_h} \sum_{a=1}^{A} {}_a\hat{D}'_{hi} \right)^2 \right] = \sum_{h=1}^{H} \frac{N_h}{n_h} \frac{N_h - n_h}{n_h - 1} \left[\sum_{i=1}^{n_h} \left(\hat{D}'_{hi} \right)^2 + \right.$$

$$\left. - \frac{1}{n_h} \left(\sum_{i=1}^{n_h} \hat{D}'_{hi} \right)^2 \right] \quad (73)$$

nella quale si è posto:

$$\sum_{a=1}^{A} {}_a\hat{D}'_{hi} = \hat{D}'_{hi} \tag{74}$$

Un'ulteriore espressione alternativa della (43), oltre la (73), può ricavarsi mediante alcuni semplici passaggi. Infatti dalla (73) segue che:

$$\hat{V}\left({}_{pc}\hat{\bar{Y}}\right) = \sum_{h=1}^{H} \frac{N_h}{n_h} \frac{N_h - n_h}{n_h - 1} \left[\sum_{i=1}^{n_h} \left(\hat{D}'_{hi}\right)^2 - \frac{1}{n_h} \left(\sum_{i=1}^{n_h} \left(\hat{D}'_{hi}\right)^2 \right] = \tag{75}$$

$$= \sum_{h=1}^{H} \frac{n_h}{N_h} \frac{N_h - n_h}{n_h - 1} \frac{N_h^2}{n_h^2} \left[\sum_{i=1}^{n_h} \left(\hat{D}'_{hi}\right)^2 - \frac{1}{n_h} \left(\sum_{i=1}^{n_h} \left(\hat{D}'_{hi}\right)^2 \right]$$

avendo moltiplicato e diviso per N_h/n_h.

La (75) può inoltre riscriversi come:

$$\hat{V}\left({}_{pc}\hat{\bar{Y}}\right) = \sum_{h=1}^{H} \frac{N_h - n_h}{N_h} \frac{n_h}{n_h - 1} \left[\sum_{i=1}^{n_h} \left(\frac{N_h}{n_h} \hat{D}'_{hi} \right)^2 + \right.$$

$$\left. - \frac{1}{n_h} \left(\sum_{i=1}^{n_h} \frac{N_h}{n_h} \hat{D}'_{hi} \right)^2 \right] = \sum_{h=1}^{H} \frac{N_h - n_h}{N_h} \frac{n_h}{n_h - 1} \left[\sum_{i=1}^{n_h} (K_h \hat{D}'_{hi})^2 + \right.$$

$$\left. - \frac{1}{n_h} \left(\sum_{i=1}^{n_h} K_h \hat{D}'_{hi} \right)^2 \right] \tag{76}$$

in cui $K_h = N_h/n_h$.

Se le frazioni di campionamento n_h/N_h (h = 1,...,H) sono molto piccole la (76) può porsi nella forma:

$$\hat{V}\left({}_{pc}\hat{\bar{Y}}\right) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \left[\sum_{h=1}^{n_h} (K_h \hat{D}'_{hi})^2 - \frac{1}{n_h} \left(\sum_{i=1}^{n_h} K_h \hat{D}'_{hi} \right)^2 \right] \tag{77}$$

Inoltre, nel caso in cui le dimensioni campionarie n_h sono elevate, dalla (77) si ricava l'espressione dalla forma abbastanza semplice:

$$\hat{V}\left({}_{pc}\hat{\bar{Y}}\right) = \sum_{h=1}^{H} \left[\sum_{i=1}^{n_h} (K_h \hat{D}'_{hi})^2 - \frac{1}{n_h} \left(\sum_{i=1}^{n_h} K_h \hat{D}'_{hi} \right)^2 \right] \tag{78}$$

Riteniamo, ancora, utile descrivere una forma piacevolmente simmetrica e semplificata che si ottiene quando il campione è formato soltanto da due unità.

In tal caso, dalla (74) si ha:

$$\hat{V}\left({}_{pc}\hat{\bar{Y}}\right) = \sum_{h=1}^{H} \frac{N_h}{2} (N_h - 2) \left[(\hat{D}'_{h1})^2 + (\hat{D}'_{h2})^2 + \right.$$

$$\left. - \frac{1}{2} (\hat{D}'_{h1})^2 - \frac{1}{2} (\hat{D}'_{h2})^2 + 2\hat{D}'_{h1} \hat{D}'_{h2} \right] = \tag{79}$$

$$= \sum_{h=1}^{H} \frac{N_h}{4} (N_h - 2) \left[\hat{D}'_{h1} - \hat{D}'_{h2} \right]^2$$

La (79) può anche riscriversi come segue:

$$\hat{V}\left({}_{pc}\hat{\bar{Y}}\right) = \sum_{h=1}^{H} \frac{N_h - 2}{2} \frac{N_h}{2} [\hat{D}'_{h1} - \hat{D}'_{h2}]^2 = \tag{80}$$

$$= \sum_{h=1}^{H} \frac{N_h - 2}{N_h} [K_h \hat{D}'_{h1} - K_h \hat{D}'_{h2}]^2$$

che nel caso $N_h \gg 1$ diviene:

$$\hat{V}\,({}_{pc}\hat{Y}) = \sum_{h=1}^{H} [K_h\,\hat{D}'_{h1} - K_h\,\hat{D}'_{h2}]^2 \qquad (81)$$

Prima di concludere osserviamo che le espressioni delle varianze qui descritte possono anche derivarsi con altro ragionamento basato su una variante del procedimento seguito da Hansen, Hurwitz e Madow e da noi adottato. Detta variante, che illustreremo con riferimento allo stimatore ${}_p\hat{Y}$ e al campionamento semplice, può esprimersi come segue: sviluppiamo, in primo luogo, in serie di Taylor di punto iniziale:

$({}_1Y, {}_1X, \ldots, {}_aY, {}_aX, \ldots, {}_AY, {}_AX)$

la stima ${}_p\hat{Y}$, interpretata come funzione delle stime:

$({}_1\hat{Y}, {}_1\hat{X}, \ldots, {}_a\hat{Y}, {}_a\hat{X}, \ldots, {}_A\hat{Y}, {}_A\hat{X})$

Posto:

$${}_p\hat{Y} = f\,({}_1\hat{Y}, {}_1\hat{X}, \ldots, {}_a\hat{Y}, {}_a\hat{X}, \ldots, {}_A\hat{Y}, {}_A\hat{X}) = \sum_{a=1}^{A} \frac{{}_a\hat{Y}}{{}_a\hat{X}}\,{}_aX \qquad (82)$$

si ottiene:

$$f\,({}_1\hat{Y}, {}_1\hat{X}, \ldots, {}_a\hat{Y}, {}_a\hat{X}, \ldots, {}_A\hat{Y}, {}_A\hat{X}) \doteq$$
$$= f\,({}_1Y, {}_1X, \ldots, {}_aY, {}_aX, \ldots, {}_AY, {}_AX) +$$
$$+ \sum_{a=1}^{A} \left[\frac{\partial f}{\partial\,{}_a\hat{Y}} ({}_a\hat{Y} - {}_aY) + \frac{\partial f}{\partial\,{}_a\hat{X}} ({}_a\hat{X} - {}_aX) \right] \qquad (83)$$

in cui le derivate parziali si inntendono calcolate nel punto iniziale.

Dalla (83) si ricava immediatamente che:

$${}_p\hat{Y} = \sum_{a=1}^{A} {}_aY + \sum_{a=1}^{A} ({}_a\hat{Y} - \frac{{}_aY}{{}_aX}\,{}_a\hat{X}) \qquad (84)$$

dalla quale segue:

$$V\,({}_p\hat{Y}) = V\left[\sum_{a=1}^{A} \left({}_a\hat{Y} - \frac{{}_aY}{{}_aX}\,{}_a\hat{X} \right) \right] \qquad (85)$$

in quanto il primo addendo a secondo membro è costante.
Tenendo poi presente che:

$${}_a\hat{Y} = \sum_{i=1}^{n} K\,{}_aY_i \quad \text{e} \quad {}_a\hat{X} = \sum_{i=1}^{n} K\,{}_aX_i \qquad (86)$$

la (85) può riscriversi nella forma:

$$V\,({}_p\hat{Y}) = V\left[\sum_{a=1}^{A} \left(\sum_{i=1}^{n} K\,{}_aY_i - \frac{{}_aY}{{}_aX} \sum_{i=1}^{n} K\,{}_aX_i \right) \right] =$$

$$= V\left[\sum_{i=1}^{n} K \sum_{a=1}^{A} \left({}_aY_i - \frac{{}_aY}{{}_aX}\,{}_aX_i \right) \right] =$$

$$= V\left[\sum_{i=1}^{n} K \sum_{a=1}^{A} {}_aD_i \right] = V\left[\sum_{i=1}^{n} K\,D_i \right] \qquad (87)$$

nella quale si è tenuto conto della (17) e si è posto

$$D_i = \sum_{a=1}^{A} {}_aD_i \qquad (88)$$

In definitiva si ha:

$$V\,({}_p\hat{Y}) = \frac{N\,(N-n)}{n}\,S_D^2 \qquad (89)$$

in cui:

$$S_D^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(D_i - \frac{1}{N} \sum_{i=1}^{N} D_i \right)^2 = \frac{1}{N-1} \sum_{i=1}^{N} D_i^2 \qquad (90)$$

in quanto risulta:

$$\frac{1}{N-1} \sum_{i=1}^{N} D_i = \frac{1}{N-1} \sum_{i=1}^{N} \sum_{a=1}^{A} \left({}_aY_i - \frac{{}_aY}{{}_aX} {}_aX_i \right) = 0 \qquad (91)$$

Pertanto la formula (89) si può scrivere:

$$V\left({}_p\hat{\bar{Y}}\right) = \frac{N\,(N-n)}{n\,(N-1)} \sum_{i=1}^{N} D_i^2 = \frac{N\,(N-n)}{n\,(N-1)} \sum_{i=1}^{N} \left(\sum_{a=1}^{A} {}_aD_i \right)^2 \qquad (92)$$

che coincide con la (23).

PARTE 3

TECNICHE DI FORMAZIONE DEL CAMPIONE

CAPITOLO 13 - LA STRATIFICAZIONE

Introduzione

La precisione di uno stimatore espressa dal suo errore quadratico medio è misurata da:

$$EQM\ (\hat{\theta}) = \sum_{s \in U_c} d\ (s)\ \{\hat{\theta}\ (s) - E\ [\hat{\theta}\ (s)]\}^2 \qquad (1)$$

Nel caso di uno stimatore corretto, essendo $E\ [\hat{\theta}\ (s)] = \theta$, la (1) si può scrivere:

$$V\ (\hat{\theta}) = \sum_{s \in U_c} d\ (s)\ [\hat{\theta}\ (s) - \theta]^2 \qquad (2)$$

La (2), come si vede, dipende da due elementi (al variare di s): $d\ (s)$ e $\hat{\theta}(s)$. Il primo indica il disegno campionario, il secondo la stima di θ dato il campione s.

Vediamo ora come modificando l'universo campionario U_c ed il disegno $d\ (s)$ è possibile aumentare la precisione di $\hat{\theta}$.

Un modo per realizzare ciò consiste nell'applicare delle tecniche note in letteratura con il nome *tecniche di stratificazione*.

Tali tecniche si basano sulla suddivisione della popolazione oggetto di indagine in H gruppi (o strati) di numerosità $N_1, N_2, \ldots N_H$, tali che

$$\sum_{h=1}^{H} N_h = N \qquad (3)$$

Da ciascuno strato viene estratto un campione di n_h unità sotto il vincolo

$$\sum_{h=1}^{H} n_h = n \qquad (4)$$

Affinché si massimizzi l'azione esercitata dalla stratificazione, gli strati dovranno contenere unità il più possibile simili rispetto al carattere oggetto di indagine y.

Vediamo ora come, in termini più espliciti, la varianza dello stimatore si riduce in presenza di un disegno stratificato. Una formulazione generale della varianza di una stima $\hat{Y}$ (Yates e Grundy, 1953), con riferimento ad un disegno casuale semplice, è la seguente:

$$V(\hat{Y}) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\Pi_i \Pi_j - \Pi_{ij}) \left(\frac{Y_i}{\Pi_i} - \frac{Y_j}{\Pi_j} \right)^2 \quad (i \neq j) \qquad (5)$$

Utilizziamo la (5) per un disegno stratificato ed indichiamo le probabilità di inclusione con Π'_i, Π'_j e Π'_{ij}. Tutte quelle coppie di unità (i, j) che appartengono a strati diversi hanno probabilità di inclusione $\Pi_{ij} = \Pi_i \Pi_j$; questo perché esse vengono selezionate indipendentemente l'una dall'altra. In altre parole la selezione dell'unità i nello strato h non influisce, come è facilmente verificabile, sulla probabilità di selezione dell'unità j appartenente allo strato h' con $h \neq h'$. Per il teorema delle probabilità composte si ha che la probabilità Π'_{ij} è pari a $\Pi'_i \Pi'_j$ che, sostituita nella (5) rende nullo il contributo alla varianza della stima $\hat{Y}$ relativo a tutte le coppie (i, j) appartenenti a strati diversi.

Questo fatto, insieme alla proprietà di un buon processo di stratificazione che a coppie (i, j) di unità appartenenti a strati diversi fa corrispondere valori massimi della differenza

$$\left(\frac{Y_i}{\Pi_i} - \frac{Y_j}{\Pi_j} \right) \qquad (6)$$

i quali, a loro volta, si associano a valori nulli del fattore:

$$(\Pi'_i \Pi'_j - \Pi'_{ij})$$

conduce ad una riduzione di $V(\hat{Y})$.

Il problema che ora si presenta allo statistico è quello di raggruppare in strati omogenei le unità di U. Se fossero conosciute le modalità che il carattere oggetto di indagine y assume per ciascuna unità il problema sarebbe risolto in quanto basterebbe applicare un qualsiasi algoritmo di classificazione per ottenere strati il più possibile omogenei.

Nelle situazioni reali in cui le modalità di y sono sconosciute, bisogna ricercare qualche altro carattere (o insieme di caratteri) le cui modalità siano conosciute per ciascuna unità di U. Questo (o questi) viene chiamato carattere di stratificazione.

È ovvio che tra il carattere di stratificazione e quello oggetto di indagine debba esistere un qualche legame sulla base del quale raggruppando le unità in strati omogenei rispetto al carattere di stratificazione (indichiamolo con x) risulti allo stesso modo omogenea la classificazione rispetto al carattere y. Allora, quanto più stretto è il legame tra y ed x tanto maggiore risulterà l'effetto della stratificazione in termini di riduzione della varianza di $\hat{Y}$.

Nelle pagine che seguono vengono affrontati, in primo luogo, due fondamentali aspetti di un processo di stratificazione: determinazione del numero di strati e delimitazione degli stessi. Seguiranno poi alcune considerazioni relative ai disegni a due stadi e alle indagini *multiscopo*; alcuni cenni riguardanti la stima dell'effetto della stratificazione concluderanno il capitolo.

Determinazione del numero di strati

Come già accennato, la procedura di stratificazione consiste nel raggruppare le N unità della popolazione oggetto di indagine in H gruppi (strati) il più possibile omogenei rispetto alla variabile (od alle variabili) di interesse. A tal fine si ricorre in genere all'uso di un carattere non oggetto di indagine, le cui modalità siano conosciute per ciascuna unità della popolazione, che sia legato (correlato) con il carattere di interesse (Grosbras, 1987).

Nella pratica la prima operazione consiste nel reperire tali informazioni, solitamente disponibili da fonti amministrative (ad esempio, le anagrafi dei comuni o delle camere di commercio) o da precedenti indagini statistiche (Censimenti).

La seconda operazione consiste nella scelta, tra quelle disponibili, delle variabili da utilizzare nella formazione degli strati; questo argomento verrà affrontato nel seguito.

Il problema successivo che si presenta allo statistico che progetta un'indagine campionaria è quello di determinare il numero di strati in cui suddividere la popolazione. Tale problema è di facile soluzione se la variabile di stratificazione è di tipo qualitativo in quanto, se le modalità che questa può assumere non sono in numero eccessivo (come avviene ad esempio, se i caratteri di stratificazione sono il sesso o la ripartizione geografica), gli strati risultano automaticamente determinati dalle diverse modalità del carattere stesso.

È molto frequente però il caso in cui la variabile di stratificazione è di tipo quantitativo, per cui la determinazione del numero di strati non è così immediata. È stato ampiamente dimostrato (Cochran, 1977; Fabbris, 1989) che all'aumentare del numero degli strati il contributo aggiuntivo alla diminuzione della varianza campionaria risulta decrescente; oltre un certo limite, pertanto, il guadagno in efficienza non giustifica la mole di calcoli necessari per una stratificazione più fine.

Vediamo ora come determinare il numero di strati da adottare in una indagine campionaria. Tenendo presente che ciascuno strato deve contenere almeno un'unità campione e che la determinazione della numerosità campionaria dipende, tramite la varianza, dalla stessa stratificazione, si vede che il problema è di difficile soluzione se non si dispone di informazioni relative alla variabilità del carattere oggetto di indagine.

In letteratura, infatti, non esiste una metodologia diretta alla determinazione del numero ottimale di strati da adottare in un disegno stratificato, ma vengono date indicazioni derivanti in larga parte da esperienze empiriche (Kish, 1965; Hansen, Hurwitz e Madow, 1953).

In particolare può essere utilizzata una formula che consente di valutare l'incremento di efficienza all'aumentare del numero di strati. Indicando con ρ la correlazione lineare tra la variabile di stratificazione e quella oggetto d'indagine, con H il numero di strati e con $\hat{Y}$ e $\hat{Y}_{cs}$, rispettivamente, le stime dirette relative al campionamento stratificato e a quello casuale semplice, si dimostra (Cochran, 1977) che:

$$\frac{V(\hat{Y})}{V(\hat{Y}_{cs})} = \left[\frac{\rho^2}{H^2} + (1 - \rho^2)\right] \qquad (8)$$

Poiché il rapporto $V(\hat{Y})/V(\hat{Y}_{cs})$, per un dato valore di ρ, è funzione del reciproco di H^2 si ha che l'incremento di efficienza dovuto alla stratificazione diminuisce velocemente all'aumentare di H e tende ad $(1-\rho^2)$ per H che tende ad infinito.

Se invece $\rho = 1$ (che si verifica quando esiste un legame lineare tra variabile di stratificazione e variabile oggetto d'indagine) per H che assume valori molto elevati, il rapporto in questione diventa prossimo allo zero; per contro, se $\rho = 0$ il rapporto assume il valore 1 qualunque sia H.

In quest'ultimo caso, pertanto, la stratificazione non influisce sull'efficienza della stima $\hat{Y}$.

Ai fini della scelta del numero di strati, alle considerazioni appena svolte, occorre aggiungere che all'aumentare di H aumentano generalmente sia i costi di elaborazione che quelli relativi alla raccolta delle informazioni.

In definitiva, combinando i diversi aspetti illustrati, la soluzione deve essere cercata determinando quel valore H^* tale che per $H > H^*$ l'aumento di efficienza, in un'analisi costi-benefici, risulta improduttivo.

Determinazione dei limiti degli strati

Per determinare, sulla base della relazione (8), il numero di strati in cui suddividere la popolazione, occorre aver definito i limiti degli strati, ossia i relativi estremi inferiori e superiori.

Esistono diversi metodi per risolvere tale problema.

Una prima soluzione è stata data da Dalenius (1957), attraverso ipotesi sulla distribuzione teorica del carattere oggetto di indagine nella popolazione.

Indicando con H il numero di strati (per il momento, fissato in modo arbitrario), i limiti degli stessi vengono trovati minimizzando la quantità:

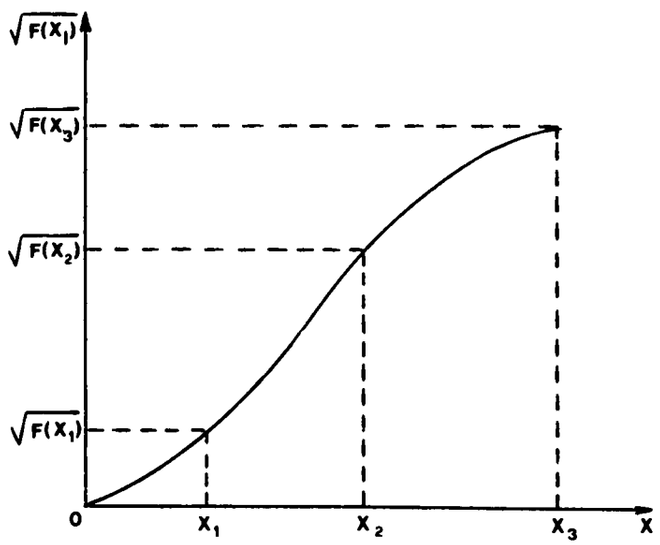
$$\sum_{h=1}^{H} W_h S_h \qquad (9)$$

in cui $W_h = N_h/N$ e S_h indica la radice quadrata della varianza della variabile di stratificazione definita da:

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2 \qquad (10)$$

A tal fine, utilizzando una metodologia ampiamente descritta in letteratura, gli strati vengono definiti in modo che la radice quadrata della cumulata della distribuzione della variabile di stratificazione risulti suddivisa in intervalli di ampiezza costante.

Indicando con $F(X_i)$ la cumulata della distribuzione delle frequenze relative della variabile x, il criterio può essere descritto attraverso il seguente grafico, in cui $H = 3$.



I tre punti, individuati sull'asse delle ordinate, che soddisfano la condizione:

$$\sqrt{F(X_3)} - \sqrt{F(X_2)} = \sqrt{F(X_2)} - \sqrt{F(X_1)} =$$

$$= \sqrt{F(X_1)} - 0 = \text{costante}$$

consentono la determinazione dei punti X_1, X_2 e X_3 sulla base dei quali risultano definiti i seguenti limiti:

$$(0 \dashv X_1),\ (X_1 \dashv X_2) \quad \text{e} \quad (X_2 \dashv X_3)$$

È possibile dimostrare che tale criterio ha come conseguenza la formazione di strati caratterizzati dall'avere un numero di unità decrescente al crescere del valore medio del carattere x nello strato.

Un criterio alternativo, che consente di minimizzare la (9), è quello di determinare il limite degli strati, avendo ordinato le unità di U secondo valori crescenti della variabile x, in modo da avere:

$$\sum_{i=1}^{N_1} X_{1i} = \sum_{i=1}^{N_2} X_{2i} = \cdots = \sum_{i=1}^{N_H} X_{Hi} \qquad (11)$$

ossia, in altri termini, imponendo che il totale del carattere di stratificazione risulti approssimativamente costante negli H strati.

L'ampiezza degli strati può essere determinata attraverso il rapporto:

$$\overline{H} = \frac{\sum_{h=1}^{H} \sum_{i=1}^{N_h} X_{hi}}{H} \qquad (12)$$

Può accadere che, per alcune unità di U, si verifichi che $X_i \geqslant \overline{H}$; tali unità costituiranno allora strati a sé e saranno pertanto incluse nel campione con certezza. In letteratura queste sono definite ''unità auto-rappresentative''.

Il criterio di stratificazione appena illustrato porta generalmente ad attribuire una probabilità di selezione, al variare di H, crescente al crescere del carattere di stratificazione; è appunto tale proprietà che consente di minimizzare la (9).

Infatti, esistendo nella maggior parte dei casi un legame diretto tra la variabile di stratificazione e quella oggetto d'indagine e risultando in genere negli strati la variabilità di quest'ultima crescente al crescere delle modalità della variabile di stratificazione, dalla (5) si deduce come la suddetta proprietà consente la minimizzazione della (9).

Infine, riteniamo utile sottolineare che se tra il carattere di stratificazione e quello oggetto di indagine dovesse esistere una relazione inversa, questo criterio di stratificazione dovrebbe essere modificato facendo in modo da selezionare, con probabilità maggiore, quelle unità che assumono valori più piccoli del carattere di stratificazione.

Le considerazioni sopra svolte, sotto l'ipotesi di H arbitrario, ci forniscono indicazioni sul modo ottimale di formare i limiti degli strati.

Tali indicazioni sono di estrema importanza in quanto consentono di definire l'utilizzazione della relazione (8) ai fini della effettiva determinazione del numero e dei limiti degli strati stessi.

Infatti, stabilito un criterio di formazione dei limiti, con riferimento ad un numero iniziale di strati H_o, e suddividendo la popolazione in base a tale criterio si ha la possibilità di calcolare la varianza $V(\hat{Y})$ che compare nelle (8).

Successivamente, il processo viene iterato per valori di $H > H_o$ fino alla determinazione del valore H^* già precedentemente trattato.

Alcune considerazioni sulla stratificazione nei disegni a due stadi

Come abbiamo già sottolineato, nei disegni a due stadi esistono due differenti popolazioni: quella costituita dall'insieme delle unità di primo stadio e quella delle unità di secondo stadio. È evidente che in tal caso la procedura di stratificazione può riguardare una delle due oppure entrambe le popolazioni.

Si possono avere così i seguenti casi:

a) stratificazione delle sole unità di primo stadio;

b) stratificazione delle sole unità di secondo stadio;

c) stratificazione sia delle unità di primo stadio che di quelle di secondo stadio;

d) stratificazione delle unità di secondo stadio solo in alcune unità di primo stadio.

Ovviamente la scelta di una delle tipologie elencate deve rispondere a delle particolari situazioni accertate dallo statistico che programma l'indagine.

Nel primo caso, ossia quando la stratificazione riguarda solo le unità primarie, i motivi che possono giustificare tale scelta sono essenzialmente due: non si dispone di una variabile di stratificazione le cui modalità siano note a livello di unità secondarie;

scarsa variabilità del carattere oggetto di indagine all'interno di ciascuna unità primaria.

Il primo motivo è abbastanza frequente nelle situazioni concrete; infatti, pur esistendo generalmente informazioni aggregate a livello di unità primarie (U.S.L., Comuni, ecc.), spesso risulta difficile se non impossibile disporre di informazioni utilizzabili a livello di unità secondarie. Per quanto riguarda la seconda motivazione è da osservare che talvolta la natura della variabile oggetto di indagine è tale che essa tende ad assumere, con riferimento alle unità appartenenti alla stessa unità primaria, modalità pressoché simili; questo fenomeno comporta una scarsa variabilità dentro ciascuna unità primaria. In tal caso, la stratificazione delle unità secondarie darebbe luogo ad un contributo irrilevante alla riduzione della varianza complessiva dello stimatore, risultando, pertanto, improduttiva.

Il fenomeno appena descritto, spesso riscontrabile nelle indagini di natura socio-economica, è valutabile mediante il coefficiente di correlazione intra-classi, rho, già introdotto nel capitolo 2.

Esaminiamo ora il caso, più raro ma pur sempre possibile, in cui conviene adottare la stratificazione delle sole unità secondarie. Questo disegno può essere adottato quando la variabilità del carattere oggetto di indagine è elevata all'interno delle unità primarie e conseguentemente bassa fra le stesse; in tal caso, è evidente allora che una stratificazione delle unità primarie non risulterebbe efficace ai fini della riduzione della varianza dello stimatore.

Tale contributo può scaturire invece dalla stratificazione delle unità secondarie all'interno di ciascuna unità primaria.

Il fenomeno appena descritto può essere misurato mediante il coefficiente di correlazione intra-classi che, nella situazione descritta, assume valori negativi.

Indicando con V_{cc} e V_{cs} la varianza dello stimatore, rispettivamente con riferimento ad un disegno casuale complesso e ad un disegno casuale semplice, sulla base della nota relazione:

$$V_{cc} = V_{cs} [1 + (\bar{n} - 1) \text{ rho}] \qquad (13)$$

si ha $V_{cc} < V_{cs}$. Questo implica che dall'adozione di un disegno a due stadi con stratificazione delle unità secondarie si ha il duplice effetto, in termini di riduzione della varianza dello stimatore, dovuto sia al processo di stratificazione dentro le unità primarie e sia alla stadificazione.

I casi descritti finora sono delle situazioni estreme; di solito, invece, si hanno delle situazioni in cui esiste sia una certa variabilità del carattere a livello di unità primarie, sia, condizionalmente a ciascuna di esse, a livello di unità secondarie. In questo

caso, se si dispone delle informazioni necessarie, è consigliabile utilizzare la procedura di stratificazione in entrambi gli stadi. È da sottolineare comunque che non esistono metodi analitici rigorosi per decidere quale disegno adottare.

Sulla base della scomposizione della varianza complessiva dello stimatore V_{cc} in termini di varianza di primo stadio V_I e di secondo stadio V_{II} essendo:

$$V_{cc} = V_I + V_{II} \qquad (14)$$

si possono tuttavia suggerire alcune soluzioni di larga massima:

— se V_I è approssimativamente uguale a V_{II} allora è opportuno stratificare sia le unità primarie che quelle secondarie;

— se, invece, V_I è relativamente più elevata di V_{II} conviene adottare un disegno che preveda la stratificazione solo per le unità primarie e, viceversa, nel caso contrario.

Per quanto riguarda il caso d), esso può interessare disegni che prevedano o meno la stratificazione delle unità primarie. Questa tipologia di disegno è consigliabile quando solo dentro ad alcune unità primarie esiste una forte variabilità del carattere oggetto d'indagine. È ovvio, quindi, che soltanto in queste unità la stratificazione può risultare vantaggiosa, mentre scarsi guadagni sono da attendersi per le altre unità primarie. In conclusione, si può affermare che il disegno a due (o più) stadi, data la sua elevata flessibilità, risulta uno dei disegni maggiormente utilizzabili nella pratica del campionamento.

Vediamo ora come determinare il numero ed i limiti degli strati in un disegno a due stadi con stratificazione delle unità di primo stadio nel caso in cui viene imposto il vincolo di un numero minimo di rilevazioni da effettuare in ciascuna unità primaria campione. Lo scopo di tale vincolo è quello di assicurare, ai rilevatori che si recano presso le unità di primo stadio campionate, un congruo numero di rilevazioni e, di conseguenza, un consistente compenso che possa fungere da incentivo ad effettuare le rilevazioni stesse. Un ulteriore motivo, che giustifica l'imposizione di un numero minimo di interviste, è quello di ridurre i costi di spostamento dei rilevatori.

Infatti, se il numero complessivo di rilevazioni è fissato, l'imposizione del vincolo comporta, negli strati in cui le unità primarie sono di piccole dimensioni, una diminuzione del numero di unità di primo stadio campione. Questo vincolo, inoltre, rende facilmente risolvibile il problema della individuazione del numero e dei limiti degli strati in cui suddividere la popolazione delle unità di primo stadio.

Indicando con s l'insieme delle unità di primo stadio campione e con m_i il numero di unità di secondo stadio campione da rilevare nella unità di primo stadio i, l'imposizione del vincolo del

numero minimo di unità può essere formalizzata nel seguente modo:

$$m^* = \min_{i \varepsilon s} \{m_i\} \qquad (15)$$

È da aggiungere che considerare il disegno che prevede la stratificazione delle sole unità di primo stadio non comporta alcuna perdita in generalità per il metodo che verrà esposto in seguito; tale metodo infatti rimane valido anche se la stratificazione è prevista per le unità di entrambi gli stadi.

Supponiamo che, come variabile di stratificazione venga considerato il numero di unità di secondo stadio universo all'interno delle unità di primo stadio.

Indichiamo inoltre con:

$$M = \sum_{i=1}^{N} M_i \qquad (16)$$

$$m = \sum_{i=1}^{n} m_i$$

rispettivamente l'insieme delle unità di secondo stadio universo e campione. Utilizzando la (15) e la (16) viene determinata una soglia (che indichiamo con S) che delimita l'ampiezza degli strati, rispetto alla variabile di stratificazione. Tale soglia risulta data da:

$$S = \frac{m^* M}{m} \qquad (17)$$

Rimane ora da individuare quali e quante unità primarie vanno a formare i vari strati. A tale fine le N unità, che costituiscono la popolazine di primo stadio, vengono poste in ordine decrescente delle modalità M_i ($i = 1,..,N$) della variabile di stratificazione. Il primo degli H strati risulta formato da quelle unità per le quali si ha:

$$\sum_{i=1}^{N_1} M_i \geqslant S \qquad (18)$$

con $\qquad M_1 \geqslant M_2 \geqslant .... \geqslant M_{N_1}.$

Per il secondo strato ($h = 2$) si avra:

$$\sum_{i=N_1+1}^{N_2} M_i \geqslant S \qquad (19)$$

e così via per gli altri strati.

In generale, tenendo presente la (17), si può scrivere:

$$M_h \geqslant \frac{m^* M}{m} \qquad (20)$$

in cui M_h indica il totale del carattere di stratificazione nel generico strato h.

Dalla (20) si ottiene:

$$\frac{m}{M} M_h \geqslant m^* \qquad (21)$$

da cui segue che:

$$\frac{m}{M} > \frac{m^*}{M_h} \qquad (22)$$

Ossia l'ampiezza degli strati risulta approssimativamente costante con oscillazioni ΔM_h che verificano la seguente disuguaglianza:

$$\Delta M_h < \min_{i \varepsilon h} M_i \qquad (23)$$

che, ovviamente, diminuisce man mano che gli strati comprendono unità primarie di ampiezza minore (ossia all'aumentare di H).

Tale approssimazione risulta inferiore a quella ottenibile senza l'introduzione del vincolo (15) in quanto in tal caso, per il generico strato h, questa sarebbe pari a:

$$\Delta M_h < \frac{1}{2} \left[\max_{i \varepsilon (h+1)} M_i \right] \qquad (24)$$

È opportuno inoltre aggiungere che nelle situazioni concrete accade in genere che alcune unità primarie hanno un'ampiezza M_i superiore alla soglia S; ciascuna di tali unità (denominate «unità auto-rappresentative») forma strato a sé e quindi viene inclusa nel campione con probabilità 1.

Si dimostra ora che il principio generale di costruire strati di uguale ampiezza, in termini di ammontare del carattere di stratificazione, nei quali è costante il numero di unità primarie campione rappresenta, sotto opportune ipotesi, una proprietà ottimale del disegno stratificato (Hansen, Hurwitz e Madow, 1953).

Assumiamo, a tal fine, di avere H strati suddivisi in G gruppi all'interno dei quali la varianza relativa di una stima $\hat{Y}$ è costante.

Indichiamo con z_{ghi} la probabilità di inclusione alla prima estrazione associata alla unità primaria i dello strato h appartenente al gruppo g (g = 1,...,G); supponiamo poi che all'interno di ogni unità primaria campione siano rilevate tutte le unità secondarie.

Sia inoltre Y_{ghi} il totale dal carattere oggetto d'indagine nell'unità primaria i dello strato h incluso nel gruppo g. Una stima del totale del carattere y nella popolazione è data da:

$$\hat{Y} = \sum_{g=1}^{G} \sum_{h=1}^{H_g} \frac{1}{n_g} \sum_{i=1}^{\bar{n}_g} \frac{Y_{ghi}}{Z_{ghi}} = \sum_{g=1}^{G} \sum_{h=1}^{H_g} \hat{Y}_{gh} \qquad (25)$$

in cui $\bar{n}_g$ indica il numero di unità primarie campione in ciascuno degli strati del gruppo g.

La varianza della stima $\hat{Y}_{gh}$ è data da:

$$V(\hat{Y}_{gh}) = \sum_{g=1}^{G} \frac{1}{\bar{n}_g} \sum_{h=1}^{H_g} Y_{gh}^2 \; B_{gh}^2 \qquad (26)$$

in cui:

$$B_{gh}^2 = \frac{\sum_{i=1}^{N_{gh}} Z_{ghi} \left(\frac{Y_{ghi}}{Z_{ghi}} - Y_{gh} \right)^2}{N_{gh} \; Y_{gh}^2} \qquad (27)$$

Poiché, per ogni gruppo g, la varianza relativa B_{gh}^2 risulta costante, il valore di Y_{gh} che minimizza la (26) sarà quel valore che minimizza la quantità:

$$\sum_{h=1}^{H_g} Y_{gh}^2 \; B_{gh}^2 \qquad (28)$$

sotto il vincolo:

$$\sum_{h=1}^{H_g} Y_{gh} = Y_g \qquad (29)$$

Applicando il metodo dei moltiplicatori di Lagrange si ha:

$$F = \sum_{h=1}^{H_g} Y_{gh}^2 \; B_{gh}^2 + \lambda \left(\sum_{h=1}^{H_g} Y_{gh} - Y_g \right) \qquad (30)$$

il minimo della (28), sotto il vincolo (29), si ottiene come soluzione del sistema:

$$\begin{cases} \dfrac{\partial F}{\partial Y_{gh}} = 0 \\ \dfrac{\partial F}{\partial \lambda} = 0 \end{cases} \qquad (31)$$

che fornisce Y_{gh} proporzionale a $1/B_{gh}^2$; essendo B_{gh}^2 costante per ipotesi si deduce che Y_{gh} è costante.

Poiché la variabile di stratificazione x e quella oggetto di indagine y risultano generalmente correlate, possiamo affermare che il criterio di stratificazione che consente di minimizzare la (26), sotto le ipotesi sopra elencate, è quello di costruire strati di ampiezza approssimativamente costante in termini di ammontare del carattere di stratificazione.

La stratificazione in presenza di unità di grandi dimensioni

Supponiamo ora di essere in grado di suddividere la popolazione oggetto di indagine, sulla base della variabile y (oggetto di indagine), in due gruppi: il primo contenente tutte quelle unità che presentano valori elevati del carattere (grandi unità); il secondo comprendente le rimanenti (piccole unità). Nel primo viene effettuata una indagine totale, nel secondo una indagine campionaria.

Poiché la variabile y esprime la dimensione delle unità, l'inclusione nel campione delle unità grandi (con probabilità uno) assicura la rilevazione di una larga parte dell'ammontare totale del carattere oggetto di indagine con un numero relativamente limitato di unità campione.

Il primo problema da affrontare è quello di scegliere una modalità Y_o del carattere y in base alla quale tutte le unità che presentano modalità di y minore di Y_o vengono considerate «piccole unità». Indichiamo a tal fine, con N il numero di unità della popolazione, con n il numero complessivo delle unità campione e con n_1 il numero delle unità del primo gruppo (grandi unità). Il parametro Y, totale del carattere y nella popolazione, può essere scritto come:

$$Y = \sum_{i=1}^{n_1} Y_i + \sum_{i=n_1+1}^{N} Y_i \qquad (32)$$

La stima diretta della (32) risulta quindi data da:

$$\hat{Y} = \sum_{i=1}^{n_1} Y_i + \frac{N - n_1}{n - n_1} \sum_{i=1}^{n-n_1} Y_i = Y_1 + \hat{Y}_2 \qquad (33)$$

La varianza della (33) risulta:

$$V(\hat{Y}) = \frac{(N - n_1)(N - n)}{n - n_1} S_2^2 \qquad (34)$$

in cui:

$$S_2^2 = \frac{1}{N - n_1 - 1} \sum_{i=n_1+1}^{N} (Y_i - \overline{Y}_2)^2 \qquad (35)$$

$$\overline{Y}_2 = \frac{1}{N - n_1} \sum_{i=n_1+1}^{N} Y_i \qquad (36)$$

Dalla (34), ricordando che $\varepsilon^2 = V(\hat{Y})/\hat{Y}^2$, segue la relazione:

$$\varepsilon^2 \hat{Y} = \frac{(N-n_1)(N-n)}{n - n_1} S_2^2 \qquad (37)$$

Dalla (37), inoltre, si ricava:

$$n(n_1) = n_1 + \frac{(N - n_1)^2 S_2^2}{\varepsilon^2 \hat{Y}^2 + (N - n_1) S_2^2} \qquad (38)$$

Premesso quanto sopra determiniamo ora il valore Y_o.

A tale scopo viene utilizzato un criterio finalizzato alla individuazione di un insieme $s = \{ i : Y_i < Y_o ; i = 1,\ldots,N \}$ in modo che, fissato un certo livello di precisione della stima, l'ampiezza globale del campione (che qui indichiamo come funzione di n_1) $n(n_1) = n_2 + n_1$ sia minima.

Come è noto il minimo si ha quando:

$$n(n_1 - 1) \geq n(n_1) \qquad (39)$$

$$n(n_1 + 1) \geq n(n_1) \qquad (40)$$

Nella (39), $(n_1 - 1)$ rappresenta l'ampiezza del gruppo delle unità di grandi dimensioni dal quale manca una unità che è stata invece classificata come «piccola»; viceversa per la quantità $(n_1 + 1)$ della (40). Per consentire una maggiore flessibilità all'ampiezza campionaria $n(n_1)$ viene introdotta, nella (39) e nella (40), la quantità (b - 1) con b reale arbitrario; avendosi così:

$$n(n_1 - 1) - b + 1 > n(n_1) \qquad (41)$$

$$n(n_1 + 1) - b - 1 > n(n_1) \qquad (42)$$

Per procedere al calcolo del valore Y_o ottimale, bisogna ordinare le unità della popolazione in senso decrescente di y. Determinare allora il valore di y che delimita i due gruppi (in modo tale

che la (38) sia minima), significa individuare una regola di arresto ottimale nella costruzione dell'insieme *s*.

A questo scopo definiamo le seguenti relazioni (Hidiroglou, 1979):

$$\begin{cases} S^2_{n_1+1} = \alpha_{n_1} S^2_{n_1} + \beta_{n_1} (\hat{Y}_{n_1} - \mu_{n_1})^2 & (43) \\ S^2_{n_1-1} = \alpha'_{n_1} S^2_{n_1} - \beta'_{n_1} (\hat{Y}_{n_1+1} - \mu_{n_1})^2 & (44) \end{cases}$$

dove:

$$S^2_{n_1} = \frac{1}{N - n_1 - 1} \sum_{i=n_1+1}^{N} (Y_i - \mu_{n_1})^2 \qquad (45)$$

$$\mu_{n_1} = \frac{1}{N - n_1} \sum_{i=n_1+1}^{N} Y_i \qquad (46)$$

$$\hat{Y}_{n_1} = \frac{N - n_1}{n - n_1} \sum_{i=1}^{n-n_1} Y_i \qquad (47)$$

$$\alpha = \frac{N - n_1 - 1}{N - n_1} \quad ; \quad \beta = \frac{N - n_1}{N - n_1 + 1}$$

$$\alpha' = \frac{N - n_1 - 1}{N - n_1 - 2} \quad ; \quad \beta' = \frac{N - n_1}{N - n_1 - 1} \qquad (48)$$

Sostituendo le (43) e (44) nella (39) ed utilizzando le (41) e (42) si perviene al seguente sistema di disequazioni:

$$(Y_{n_1} - \mu_{n_1})^2 \geqslant \left[\frac{(bN - n - n_1 b - n_1)(N - n_1)}{(n - n_1)(N - n - b + 1)} + \frac{1}{N - n_1} \right] S_2^2$$

$$(Y_{n_1+1} - M_{n_1})^2 \leqslant \left[\frac{(bN - n - n_1 b - n_1)}{(n - n_1)(N - n + b - 1)} + \frac{1}{N - n_1} \right] S_2^2 \qquad (49)$$

Una soluzione della (49) risulta data da:

$$Y_o = \mu_{n_1} + \left\{ \frac{b(N-n_1-1)}{(n-n_1)} + \frac{(b-1)(N-n_1)}{2[N-n_1-b+1]} + \frac{(b-1)(N-n_1-2)}{2[N-n+b-1]} + \right.$$

$$\left. + \frac{1}{2} \frac{b(b-1)}{(n-n_1)} \left[\frac{N-n_1}{N-n_1-b+1} - \frac{N-n_1-2}{N-n+b-1} \right] \right\}^{1/2} \qquad (50)$$

Questa può però non essere unica, per cui possono presentarsi problemi di scelta tra soluzioni diverse (Glasser, 1962).

La formula (50), come si vede facilmente, non risulta molto agevole nei casi concreti; una approssimazione può essere ottenuta utilizzando i parametri S^2, μ, ε ed Y riferiti all'intera popolazione invece che alla sub-popolazione delle N — n, unità del gruppo *s*. In questo caso la regola di arresto ottimale, nella formazione del gruppo *s*, prevede che l'inclusione delle unità della popolazione oggetto di indagine in *s* avrà termine quando verrà esaminata quella unità con modalità Y_o tale che:

$$Y_o > \mu + \left[\frac{b \varepsilon^2 \hat{Y}^2}{N} + S^2 \left\{ (2b - 1) + \frac{N(b-1) S^2}{\varepsilon^2 \hat{Y}^2} \right\} \right]^{1/2} \qquad (51)$$

È stato fin qui ipotizzato che i due gruppi siano stati individuati tramite la stessa variabile oggetto di indagine y ; è evidente che in una indagine reale questo non è possibile, per cui il limite dei due gruppi dovrà essere determinato sulla base di una variabile ausiliaria x, strettamente correlata con quella oggetto di indagine, le cui modalità sono conosciute per tutte le N unità della popolazione. Ad esempio, in una indagine sulle aziende agricole in cui l'obiettivo dell'indagine è quello di rilevare dati sulla produzione in complesso, come variabile ausiliaria (se disponibile) può essere considerata la superficie agricola utilizzata di ciascuna azienda. È evidente che includendo con probabilità uno le aziende di grandi dimensioni (in termini di superficie agricola utilizzata) si ha l'eliminazione di una grossa componente di variabilità nello stimatore: attraverso la metodologia esposta è possibile massimizzare tale riduzione di variabilità. Un ulteriore campo di applicazione della metodologia esposta è quello dei disegni e due stadi per le indagine sulla popolazione. In tal caso la variabile ausiliaria dovrebbe essere la dimensione demografica dei comuni e la metodologia dovrà essere applicata alle unità di primo stadio (comuni).

Alcune considerazioni relative al caso di variabili multidimensionali

Può accadere che le «informazioni disponibili» per stratificare una popolazione siano in numero elevato: ossia, la variabile di stratificazione **x** sia una variabile multidimensionale:

$$\mathbf{x} = (x_1, x_2, \ldots, x_k) \qquad (52)$$

Si pone, così, il problema di suddividere la popolazione in gruppi omogenei utilizzando la (52) in modo ottimale (Cochran, 1977; Zannella, 1987).

A tale fine bisogna sottolineare che i problemi che sorgono sono di duplice natura: problemi di calcolo e problemi di numerosità campionaria. I due problemi sono legati poiché il numero di strati cresce, a parità del numero di classi (o di modalità), al crescere del numero di variabili di stratificazione.

Il numero di strati H è ricavabile dalla seguente relazione:

$$H = \prod_{i=1}^{K} F_i \qquad (53)$$

in cui F_i indica il numero di classi (o modalità) della i-esima variabile di stratificazione.

La numerosità campionaria minima, che si ottiene nella situazione di una sola unità campione per strato, è ovviamente pari ad H; nel caso in cui si voglia ottenere una stima corretta della varianza, che comporta la selezione di almeno due unità per strato, la dimensione camionaria minima è pari a 2H.

Utilizzando un numero consistente di variabili di stratificazione può accadere che la numerosità teorica, determinata sulla base di un prefissato livello atteso di precisione delle stime, debba essere successivamente aumentata per il verificarsi delle due seguenti situazioni:

- l'allocazione del campione complessivo negli strati può comportare che, in alcuni di questi, il numero di unità da campionare sia inferiore ad uno, mentre per la determinazione delle stime è necessario avere almeno un'unità campione per strato;
- l'allocazione del campione complessivo negli strati può comportare che, in alcuni di questi, il numero di unità da campionare sia inferiore a due, mentre l'esigenza di avere una stima corretta della varianza richiede che, in ogni strato, le unità selezionate siano almeno due.

Il presentarsi dell'una o dell'altra circostanza comporta che la numerosità effettiva supera quella teorica. L'incremento è tanto maggiore quanto più elevato è il numero di strati. Questo è, in genere, tanto più elevato quanto maggiore è il numero delle variabili di stratificazione e quanto più elevato è il numero di modalità delle stesse.

D'altra parte un numero di strati elevato, come si evince dalla relazione (8), non porta alcun contributo aggiuntivo apprezzabile alla diminuzione della varianza campionaria.

Un metodo che è possibile adottare in tal caso è quello di scegliere come variabili di stratificazione solo quelle maggiormente correlate con le variabili oggetto d'indagine.

Alcuni studiosi (Kish e Anderson, 1978), in determinate situazioni, suggeriscono un metodo basato sulla determinazione di nuove variabili (da utilizzare poi per la stratificazione) espresse come combinazioni lineari delle variabili di stratificazione originarie.

Ai fini della costruzione degli strati, tra queste nuove variabili, denominate «componenti principali», si scelgono quelle che rappresentano una quota elevata della variabilità dei caratteri di stratificazione stessi.

Un altro caso, molto frequente nell'ambito delle indagini campionarie su larga scala, è quello dell'indagine «multiscopo», che ha la finalità di fornire informazioni su un numero elevato di variabili.

Per tali indagini, le variabili di stratificazione possono risultare efficaci per alcuni caratteri e non esserlo (od essere addirittura dannose) per altri. La scelta delle variabili di stratificazione, anche in questo caso, deve essere effettuata in modo che esse risultino correlate con la maggior parte dei caratteri oggetto di indagine od almeno con quelli più importanti rispetto agli obiettivi dell'indagine stessa.

Effetto della stratificazione

L'obiettivo fondamentale del disegno di campionamento stratificato come abbiamo più volte sottolineato, è quello di aumentare, a parità di tutte le altre condizioni, l'efficienza dello stimatore. Non sempre, comunque, per motivi di diversa natura, tale disegno raggiunge pienamente il suo scopo; da qui la necessità di valutare, volta per volta, rispetto ad una determianta stratificazione, l'effetto che essa ha avuto sulla varianza.

Una misura di tale effetto è data dal rapporto tra la varianza del campione stratificato (utilizzato per l'indagine effettiva) e quella di un (ipotetico) simile campione senza stratificazione.

Indicando con V_s e $V_{\bar{s}}$ rispettivamente la varianza della stima del disegno stratificato e quella del disegno non stratificato, una misura dell'effetto della stratificazione è data da:

$$E_s = \frac{V_s}{V_{\bar{s}}} \qquad (54)$$

I problemi sorgono quando si procede al calcolo di una stima della (54) in conseguenza del fatto che $V_{c\bar{s}}$ viene stimata utilizzando i dati campionari provenienti dal disegno stratificato e non, come sarebbe invece corretto, i dati derivanti da un campione casuale semplice di pari numerosità.

La strada più semplice per ottenere una stima di $V_{c\bar{s}}$ è quella basata sull'utilizzazione delle informazioni disponibili trascurando che le stesse derivano da un disegno stratificato.

In termini analitici nel caso particolare di un disegno ad uno stadio stratificato ciò comporta l'uso della formula (8) del capitolo 10, mediante la quale si perviene però ad una stima distorta di $V_{\bar{s}}$.

Una soluzione, statisticamente valida, che consente di ottenere una stima corretta di $V_{\bar{s}}$ è fornita dall'espressione (Cochran, 1977):

$$\hat{V}_{\bar{s}} = \frac{N\,(N-n)}{n\,(N-1)}\left[\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{i=1}^{H}Y^2_{hi} - \hat{Y}_s + \hat{V}_s\right] \qquad (55)$$

in cui $\hat{Y}_s$ indica la stima diretta ottenuta con il campionamento ad uno stadio stratificato e $\hat{V}_s$ la corrispondente stima della varianza campionaria, espressa dalla (20) del capitolo 10.

Svolgiamo ora qualche considerazione sul calcolo dell'effetto della stratificazione nei disegni a due stadi.

Come si è avuto modo di accennare in precedenza, nei disegni a due stadi, la procedura di stratificazione può riguardare entrambe oppure una sola delle due popolazioni (di primo e secondo stadio) che sono oggetto di campionamento. È evidente che le considerazioni sulle metodologie di costruzione degli strati rimangono valide e possono essere indipendentemente applicate sia sulle sole unità di primo o di secondo stadio che su entrambe.

Quello che risulta invece modificato è il procedimento di stima dell'effetto della stratificazione sulla varianza degli stimatori.

Quando la procedura di stratificazione riguarda solo il secondo stadio, l'effetto della stratificazione può essere calcolato, per ciascuna unità di primo stadio campione, allo stesso modo di un disegno ad uno stadio stratificato utilizzando le (55). Per quanto concerne il caso del disegno a due stadi, in cui la costruzione degli strati sia prevista per le sole unità di primo stadio, il calcolo dell'effetto della stratificazione viene effettuato utilizzando formule diverse rispetto a quelle di un disegno ad uno stadio stratificato. Ricordiamo che il problema del calcolo (e quindi della stima) dell'effetto della stratificazione risiede nel calcolo della varianza dello stimatore relativo all'analogo disegno senza stratificazione utilizzando i dati campionari relativi al disegno stratificato.

Con E''_s viene indicato l'effetto della stratificazione, con V''_s la varianza complessiva, in presenza di un disegno a due stadi con stratificazione delle unità di primo stadio, e con $V''_{\bar{s}}$ la varianza in un analogo disegno senza stratificazione.

Si ha:

$$E''_s = \frac{V''_s}{V''_{\bar{s}}} \qquad (56)$$

ovviamente una stima della (56) si ottiene sostituendovi la stima di V''_s e $V''_{\bar{s}}$. Per quanto riguarda V''_s viene utilizzata la formula già indicata nel capitolo 10 del presente volume; per $V''_{\bar{s}}$ questa può essere calcolata attraverso la seguente formula (Russo, 1985; Russo, 1986 b):

$$V''_{\bar{s}} = \frac{1}{n}\sum_{h=1}^{H}\frac{n_h}{P_h}\,V_h + \frac{1}{n}\sum_{h=1}^{H}\frac{Y_h^2}{P_h} - \frac{Y_2}{n} \qquad (57)$$

Si può dimostrare che una stima corretta della (57) risulta data da (Russo, 1986 b):

$$\hat{V}''_{\bar{s}} = \frac{1}{n}\left[\sum_{h=1}^{H}\frac{\hat{V}_h}{P_h}(n_h-1) + \sum_{h=1}^{H}\frac{\hat{Y}_h^2}{P_h} - \hat{Y}^2 + \hat{V}\right] \qquad (58)$$

Infine, una stima dell'effetto stratificazione risulta espressa dal rapporto:

$$\hat{E}''_s = \frac{\hat{V}''_s}{\hat{V}''_{\bar{s}}}$$

CAPITOLO 14 - DETERMINAZIONE DELLA DIMENSIONE DEL CAMPIONE

Introduzione

Nei precedenti Capitoli 10, 11 e 12 abbiamo definito le espressioni della varianza campionaria necessarie sia per valutare probabilisticamente il livello di precisione delle varie stime considerate nei Capitoli 6, 8 e 9, sia per formulare in maniera razionale i piani di campionamento considerati nel presente volume.

L'illustrazione della struttura formale di tali espressioni ha messo in luce che la varianza di una stima dipende anche dalla dimensione campionaria complessiva (nel caso di campionamento casuale semplice) e dalle dimensioni dei campioni dei singoli strati (nel caso di campioni stratificati).

In tutte le considerazioni svolte era tacita l'ipotesi che le suddette dimensioni campionarie fossero quantità arbitrarie, anche se di entità inferiori a quelle delle corrispondenti popolazioni.

Nel Capitolo 3 abbiamo, altresì, sottolineato che, una volta calcolata la varianza, è possibile determinare l'errore di campionamento (assoluto e relativo) e l'intervallo di confidenza, che costituiscono ulteriori elementi obiettivi di giudizio circa l'attendibilità delle stime fornite da un'indagine campionaria.

Ora, nel caso di indagini concrete, non è infrequente che tale giudizio sia negativo, nel senso che gli intervalli di confidenza sono eccessivamente grandi e che le stime ottenute risultano molto imprecise; in tal caso, si sarebbero spesi tempo e danaro per ottenere risultati di scarsa attendibilità e quindi inutilizzabili ai fini pratici.

Per cautelarsi contro il rischio di ottenere, ad indagine effettuata, un'ampiezza troppo grande degli intervalli di confidenza è necessario fissare in anticipo l'errore massimo di campionamento che si è disposti a tollerare e determinare conseguentemente una dimensione del campione sufficientemente elevata perché detto errore non venga superato. A tale scopo, come vedremo in dettaglio nel prosieguo, è necessario ricorrere all'impiego della varianza campionaria, corrispondente al disegno di campionamento (ad uno stadio stratificato, a due stadi, ecc.) e al procedimento di stima (metodo diretto, metodo del rapporto, ecc.) che si vogliono adottare per la realizzazione dell'indagine e per la determinazione delle stime delle caratteristiche della popolazione oggetto di studio.

Nei paragrafi che seguono, limitatamente ai disegni di campionamento considerati, illustreremo i criteri comunemente seguiti per la determinazione della dimensione complessiva del campione e per la ripartizione di tale dimensione nei singoli strati.

L'illustrazione, per maggiore semplicità, sarà svolta con riferimento al caso fondamentale della stima diretta del totale Y del carattere y oggetto d'indagine; tuttavia, i risultati e le dimostrazioni relativi a tale tipo di stima sono più che sufficienti per comprendere lo spirito e le tecniche di determinazione della dimensione del campione sia negli altri casi particolari di stima precedentemente trattati, sia per affrontare e risolvere gli altri casi di cui non ci siamo occupati.

Campionamento casuale semplice

Immaginiamo di aver estratto, senza reimmissione e con probabilità uguali, un campione di ampiezza n da una popolazione costituita da N unità.

Supponiamo ancora di voler stimare il totale Y del generico carattere y oggetto d'indagine, definito da:

$$Y = \sum_{i=1}^{N} Y_i \qquad (1)$$

in cui Y_i indica il valore del carattere y relativo all'unità i.

La stima diretta del totale Y è fornita dall'espressione:

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} Y_i \qquad (2)$$

che verifica la relazione:

$$E(\hat{Y}) = Y \qquad (3)$$

cioè: il valore medio, nell'universo dei campioni, di tutte le possibili stime del totale Y, ottenibili per mezzo di un campione di determinata ampiezza n, coincide con il totale Y.

Inoltre, la varianza della stima $\hat{Y}$ è espressa da:

$$V(\hat{Y}) = N^2 \frac{N - n}{N} \frac{S^2}{n} \qquad (4)$$

in cui:

$$S^2 = \frac{1}{N - 1} \sum_{i=1}^{N} (Y_i - \overline{Y})^2 \qquad (5)$$

$$\overline{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i \qquad (6)$$

Dalla (4) si ricavano poi l'errore campionario assoluto e l'errore campionario relativo definiti, rispettivamente, dalle relazioni:

$$\sigma(\hat{Y}) = \sqrt{V(\hat{Y})} \qquad (7)$$

e

$$\varepsilon(\hat{Y}) = \frac{\sigma(\hat{Y})}{Y} \qquad (8)$$

L'errore $\sigma(\hat{Y})$, come è stato già illustrato, misura la variabilità, rispetto alla propria media $Y = E(\hat{Y})$, della distribuzione delle stime $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_k, \ldots$ ottenibili da tutti i possibili campioni di ampiezza n estraibili dalla popolazione di dimensione N (Grosbras, 1987).

L'importanza dell'errore $\sigma(\hat{Y})$ sta soprattutto nella constatazione che, sotto ipotesi assai generali e abbastanza verosimili nella maggior parte delle indagini concrete, si può affermare che su 1.000 stime, relative pertanto a 1.000 campioni fatti tutti nelle stesse condizioni, circa 683 danno luogo a una stima il cui errore, in valore assoluto, non supera $\sigma(\hat{Y})$, circa 954 danno luogo ad una stima il cui errore, in valore assoluto, non supera $2\,\sigma(\hat{Y})$, ed, infine, 997 danno luogo ad una stima il cui errore, in valore assoluto, non supera $3\,\sigma(\hat{Y})$.

In particolare, oltre il 95% delle stime campionarie $\hat{Y}_k$ cade in un intervallo, intorno al parametro oggetto di stima Y, definito da:

$$[Y - 2\,\sigma(\hat{Y}),\ Y + 2\,\sigma(\hat{Y})] \qquad (9)$$

In altri termini, $2\,\sigma(\hat{Y})$ delimita un intervallo, noto con il nome di ''intervallo di fiducia'', nel quale, nella maggior parte dei casi concreti, cade con probabilità non inferiore a 0,95 la stima campionaria $\hat{Y}$; ossia, $2\,\sigma(\hat{Y})$ segna un limite massimo, in senso assoluto, all'errore di stima, che è da ritenere non superato nel 95% dei casi (Singh e Chaudhary, 1986).

Ma l'espressione (4) può vedersi anche sotto altro aspetto molto interessante in pratica. Dalla (4), infatti, si evince una relazione inversa tra la varianza $V(\hat{Y})$ e la dimensione del campione: in altri termini, per una data popolazione, la varianza $V(\hat{Y})$ è tanto più piccola quanto più grande è la dimensione del campione,

fino ad annullarsi per n = N. Quest'ultimo fatto è molto importante perché, quando si può scegliere a piacere la dimensione campionaria, conviene senz'altro scegliere n in modo che il margine di errore e quindi la varianza campionaria risultino il più possibile piccoli.

In questo secondo caso, ossia quando si è liberi di far variare n, si usa parlare di "errore ammesso" e lo si indica con il simbolo 2Θ: esso rappresenta, pertanto, l'errore massimo in senso assoluto che si è disposti a tollerare nel 95% dei casi.

Dall'essere:

$$2\,\sigma\,(\hat{Y}) \leq 2\,\Theta \tag{10}$$

segue:

$$\sigma^2\,(\hat{Y}) \leq \Theta^2 \tag{11}$$

e quindi, tenendo presente la (4), la (11) diviene:

$$N^2\ \frac{N-n}{N}\ \frac{S^2}{n} \leq \Theta^2 \tag{12}$$

Risolvendo quest'ultima relazione rispetto ad n si ottiene la diseguaglianza:

$$n \geq \frac{N^2\ S^2}{\Theta^2 + N\ S^2} \tag{13}$$

a cui deve soddisfare la numerosità n del campione in base all'errore ammesso $2\,\Theta$.

Circa poi il modo di fissare l'errore massimo 2Θ si usa commisurare tale errore ad una percentuale in relazione al livello di precisione che si assegna alla stima; in altre parole, se commisuriamo l'errore 2Θ ad una percentuale $2\,\varepsilon\,(\hat{Y})$ del parametro da stimare Y, ossia (De Lucia, 1958):

$$2\,\Theta = 2\,\varepsilon\,(\hat{Y})\,Y \tag{14}$$

si ha:

$$\Theta^2 = \varepsilon^2\,(\hat{Y})\,Y^2 \tag{15}$$

Introducendo la (15) nella (13) segue quindi:

$$n \geq \frac{N^2\ S^2}{\varepsilon^2\,(\hat{Y})\,Y^2 + N\ S^2} \tag{16}$$

Come può osservarsi dalla (16), la richiesta dimensione campionaria, che garantisce un errore non superiore ad $2\,\varepsilon\,(\hat{Y})$ al livello di probabilità P = 95%, può essere determinata purché si conoscano la varianza S^2 ed il totale Y.

Il problema viene generalmente risolto utilizzando una stima di S^2 e di Y ottenuta da una precedente indagine.

In mancanza di tali informazioni è necessario ricorrere ad un'indagine pilota su un campione preliminare relativamente piccolo, al solo scopo di ricavare una stima di S^2 e Y.

Supponiamo ora, tanto per svolgere un esempio, che si voglia effettuare un'indagine campionaria per stimare il reddito totale Y di una popolazione di 10.000 unità, in modo che l'errore ammesso, al livello di probabilità P = 95%, sia pari al 5% del reddito totale della popolazione in oggetto, cioè:

$$2\,\Theta = 0{,}05\,Y \tag{17}$$

Supponiamo inoltre di conoscere, da una indagine precedente, una stima di Y, uguale a 10.000.000 di lire e una stima di S^2, uguale a 900.000 lire.

Dalla (17), tenendo presente la (15), segue che $\varepsilon\,(\hat{Y}) = 0{,}05/2 = 0{,}025$; applicando poi la (16) si ricava:

$$n \geq \frac{10.000^2 \cdot 900.000}{0{,}025^2 \cdot 10.000.000^2 + 10.000 \cdot 900.000} = 1.440 \tag{18}$$

Campionamento ad uno stadio stratificato

Nel precedente Capitolo 10 abbiamo illustrato l'espressione della varianza campionaria della stima $\hat{Y}$ del totale Y, che riteniamo tuttavia utile riscrivere:

$$V\,(\hat{Y}) = \sum_{h=1}^{H} N_h^2\ \frac{N_h - n_h}{N_h}\ \frac{S_h^2}{n_h} \tag{19}$$

in cui:

$$\hat{Y} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} Y_{hi} \tag{20}$$

rappresenta una stima corretta di:

$$Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} Y_{hi} \qquad (21)$$

Ricordiamo ancora che la (19) si riferisce al caso in cui le n_h unità vengono estratte con probabilità uguale e senza reimmissione.

Ciò premesso, dalla relazione (19) discende che per una data popolazione la varianza V (Ŷ) varia sia in funzione della dimensione complessiva del campione n e sia, per un assegnato valore di n, in funzione delle numerosità n_1, ..., n_h, ..., n_H, vincolate dalla condizione che la loro somma risulti uguale ad n. Sotto tale vincolo, le quantità n_1, ..., n_h, ..., n_H, possono prendersi in molti modi differenti a cui corrispondono valori diversi di V (Ŷ).

Tra questi vari modi di fissare le dimensioni campionarie degli H strati, due sono i più comunemente usati (Pompilj, 1952).

Il primo è basato sul criterio di prelevare da ciascuno strato la stessa percentuale di elementi, sicché il numero di elementi scelti risulta proporzionale alla numerosità N_h di ciascuno strato (criterio proporzionale); l'altro è basato sul criterio di prelevare da ciascuno strato una percentuale variabile di elementi in modo da ottenere il minimo valore di V(Ŷ) (criterio di Neyman).

Illustriamo, in primo luogo, il criterio proporzionale (De Lucia, 1958).

Supponiamo per un momento che la dimensione campionaria complessiva n sia assegnata.

Secondo questo criterio il numero complessivo n, con cui formare il campione stratificato, viene ripartito tra i vari strati in proporzione alla numerosità di ciascuno strato; deve risultare cioè:

$$\frac{n_1}{N_1} = \ldots\ldots\ldots = \frac{n_h}{N_h} = \ldots\ldots\ldots = \frac{n_H}{N_H} \qquad (22)$$

dalla quale segue che:

$$n_h = \frac{N_h}{N} n \qquad (h = 1,\ldots\ldots,H) \qquad (23)$$

con:

$$n = \sum_{h=1}^{H} n_h \qquad \text{ed} \qquad N = \sum_{h=1}^{H} N_h \qquad (24)$$

In tal modo si viene a dare maggiore importanza agli strati aventi dimensione più elevata, prescindendo pertanto dalla variabilità di ciascuno strato.

Inserendo la relazione (23) nella (19) e nella (20) si ottiene:

$$V(\hat{Y}) = \frac{N}{n}\left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} N_h S_h^2 \qquad (25)$$

$$\hat{Y} = \frac{N}{n} \sum_{h=1}^{H} \sum_{i=1}^{n_h} Y_{hi} \qquad (26)$$

da cui si vede che, al crescere di n, la varianza V(Ŷ) descresce annullandosi per n = N.

Tutto ciò, per l'ipotesi sopra introdotta, presuppone la conoscenza della numerosità totale del campione n.

Supponiamo ora di voler ottenere una stima del totale Y, definito dalla (21), mediante una rilevazione campionaria basata su un disegno ad uno stadio stratificato e sul metodo diretto di stima, espresso dalla (20).

Supponiamo, inoltre, di voler determinare la dimensione complessiva del campione, avendo stabilito di procedere alla formazione del campione secondo il criterio proporzionale.

Utilizzando la relazione (25) e fissato l'errore massimo 2 Θ (al livello di probabilità P = 95%) che si è disposti a tollerare nella stima di Y, si ha:

$$2\,\sigma(\hat{Y}) \leq 2\,\Theta \qquad (27)$$

da cui segue:

$$\sigma^2(\hat{Y}) \leq \Theta^2 \qquad (28)$$

che per la (25) può scriversi nella forma equivalente:

$$\frac{N}{n}\left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} N_h S_h^2 \leq \Theta^2 \qquad (29)$$

da cui si ricava che la dimensione complessiva del campione stratificato deve soddisfare alla disuguaglianza:

$$n \geqslant \frac{N \sum_{h=1}^{H} N_h S_h^2}{\Theta^2 + \sum_{h=1}^{H} N_h S_h^2} \qquad (30)$$

La dimensione del campione relativa al generico strato h si ottiene poi mediante la relazione (23).

Passiamo ora a descrivere il criterio di Neyman (Grigoletto, 1976).

Secondo questo criterio si tende a definire le numerosità $n_1, \ldots, n_h, \ldots, n_H$ in modo da minimizzare la varianza campionaria espressa dalla (19); anche in questo caso supporremo, per un momento, che la dimensione complessiva del campione n sia assegnata.

Si deve cioè minimizzare la funzione:

$$V(\hat{Y}) = \sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h}$$

intesa come funzione delle variabili $n_1, \ldots, n_h, \ldots, n_H$, sotto il vincolo:

$$\sum_{h=1}^{H} n_h = n \qquad (31)$$

Ora è possibile dimostrare (Castellano ed Herzel, 1981) che il minimo valore di $V(\hat{Y})$ rispetto alle variabili $n_1, \ldots, n_h, \ldots, n_H$, vincolate dalla condizione (31) e per un assegnato valore di n, si ottiene per:

$$n_h = n \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h} \qquad (32)$$

È questa la formula che dà la ripartizione di Neyman della numerosità campionaria n tra gli strati; secondo tale criterio, quindi, si tiene conto non soltanto del numero delle unità N_h, ma anche della diversa variabilità espressa dagli scarti S_h: ciò equivale, in pratica, a prendere più unità negli strati in cui la variabilità è elevata che non in quelli in cui la variabilità è contenuta.

Se l'espressione fornita dalla (32) per n_h viene utilizzata nella (19), si ha che la varianza $V(\hat{Y})$ può riscriversi nella forma equivalente:

$$V(\hat{Y}) = \frac{1}{n}\left(\sum_{h=1}^{H} N_h S_h\right)^2 - \sum_{h=1}^{H} N_h S_h^2 \qquad (33)$$

Utilizzando tale relazione è poi possibile determinare la dimensione n in funzione dell'errore massimo ammesso nella stima di Y (Russo e Falorsi, 1989).

Supponiamo, infatti, di voler determinare la numerosità n di un campione, avendo stabilito di procedere alla formazione del campione secondo il criterio di Neyman. Indicando con 2Θ l'errore massimo ammesso, al livello di probabilità P = 95%, nella stima di Y, si ha:

$$\frac{1}{n}\left(\sum_{h=1}^{H} N_h S_h\right)^2 - \sum_{h=1}^{H} N_h S_h^2 \leq \Theta^2 \qquad (34)$$

da cui si ricava che la dimensione n del campione deve soddisfare alla disuguaglianza:

$$n \geq \frac{\left(\sum_{h=1}^{H} N_h S_h\right)^2}{\Theta^2 + \sum_{h=1}^{H} N_h S_h^2} \qquad (35)$$

Una volta determinata la dimensione n, si ottiene la dimensione del campione per ciascuno degli strati 1,...,h,...,H, per mezzo della (32).

Riteniamo utile aggiungere che, talvolta, nell'ambito di questo secondo criterio, si definiscono le dimensioni $n_1,...,n_h,...,n_H$ ancora in modo da minimizzare la varianza della stima $\hat{Y}$, espressa dalla (19), ma non più considerando il vincolo dato da una fissata numerosità totale del campione bensì avendo fisso un costo totale C per la rilevazione; tale criterio, (noto con il nome di *criterio ottimale*), è opportuno quando il costo unitario di rilevazione varia notevolmente tra i vari strati (Droesbeke, Fichet e Tassi, 1987).

Indicando con c_h il costo dovuto alla rilevazione di una unità nello strato h, il vincolo è dunque espresso da:

$$C = \sum_{h=1}^{H} c_h n_h \qquad (36)$$

È possibile dimostrare che, sotto il vincolo (36), il minimo della varianza $V(\hat{Y})$, definita dalla (19), si ottiene (Castellano ed Herzel, 1981) per:

$$n_h = n \frac{\dfrac{N_h S_h}{\sqrt{c_h}}}{\displaystyle\sum_{i=1}^{H} \frac{N_h S_h}{\sqrt{c_h}}} \qquad (h = 1,...,H) \qquad (37)$$

La (37) suggerisce la regola di assegnare valori di n_h più grandi a quegli strati con più alta variabilità, a costi unitari più bassi e di dimensione più elevata.

Il confronto della (37) con la (32) mette in rilievo che, a parità di n, la composizione del campione ottimo non è quella del campione a varianza minima (o di Neyman), se le c_h non sono tutte uguali; se invece i costi c_h sono costanti segue immediatamente che la (37) si semplifica divenendo uguale alla (32).

Inoltre, sfruttando la (37), poiché il costo totale C è fissato, potrà porsi:

$$\sum_{h=1}^{H} c_h n_h = \frac{n}{\displaystyle\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}}} \sum_{h=1}^{H} N_h S_h \sqrt{c_h} = C \qquad (38)$$

da cui si ricava la relazione:

$$n = C \frac{\displaystyle\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}}}{\displaystyle\sum_{h=1}^{H} N_h S_h \sqrt{c_h}} \qquad (39)$$

mediante la quale è possibile determinare la dimensione n del campione complessivo, che introdotta nella (37) consente il calcolo delle dimensioni n_h (h = 1,...., H).

Se si introduce poi la (37) nell'espressione (19), si ricava che $V(\hat{Y})$ può riscriversi nella forma:

$$V(\hat{Y}) = \frac{1}{n} \left(\sum_{h=1}^{H} N_h S_h \sqrt{c_h} \right) \left(\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}} \right) - \sum_{h=1}^{H} N_h S_h^2 \qquad (40)$$

È utile osservare che del criterio appena descritto esiste anche una seconda formulazione finalizzata alla determinazione delle numerosità $n_1, ..., n_h, ..., n_H$ che rendono minimo il costo totale C per un prefissato valore Θ dell'errore di campionamento $\sigma(\hat{Y})$.

In concreto può forse ritenersi più realistica la prima formulazione in cui si ritiene prefissato il costo totale. Le trattazioni analitiche dei due problemi sono, in ogni caso, strettamente collegate ed è possibile svilupparle in gran parte in modo unitario, nel senso che è agevole trovare con pochi passaggi aggiuntivi la soluzione sia per l'una che per l'altra delle formulazioni (Castellano ed Herzel, 1981).

È possibile dimostrare che la dimensione n del campione complessivo che risolve il problema della minimizzazione del costo C, sotto la condizione che l'errore di campionamento abbia il valore prefissato Θ, è fornita dall'espressione:

$$n \geq \frac{\left(\displaystyle\sum_{h=1}^{H} N_h S_h \sqrt{c_h} \right) \left(\displaystyle\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}} \right)}{\Theta^2 + \displaystyle\sum_{h=1}^{H} N_h S_h^2} \qquad (41)$$

Partendo da questa relazione si dimostra anche che la formula che dà la ripartizione ottimale della numerosità n tra gli strati è la (37) definita nell'ambito della prima formulazione.

Prima di concludere l'illustrazione della problematica in esame riteniamo opportuno aggiungere le tre seguenti considerazioni.

La prima riguarda il fatto che, per attuare i criteri di ripartizione della numerosità campionaria n o di determinazione di quest'ultima, occorre conoscere, oltre naturalmente ai costi c_h, gli scarti quadratici medi degli strati. Questi, almeno in forma approssimata, possono essere conosciuti mediante un preliminare campione pilota o attraverso l'utilizzazione delle informazioni desumibili da precedenti indagini.

La seconda concerne la circostanza che, nella generalità dei casi concreti, l'utilizzazione delle (32), (35), (37), (39) e (41) viene attuata fissando l'errore $2\,\Theta$ uguale ad una percentuale $2\,\varepsilon\,(\hat{Y})$ del parametro da stimare Y, ossia ponendo $2\,\Theta = 2\,\varepsilon\,(\hat{Y})\,Y$; in queste situazioni, pertanto, occorre conoscere una stima del parametro Y.

La terza, infine, concerne il caso in cui le formule (32) e (37) possono indicare di assegnare ad uno strato una dimensione n_h superiore alla dimensione N_h dello strato stesso. In questa circostanza si ricalcola la numerosità campionaria degli altri strati escludendo lo strato da selezionare esaustivamente.

Confronto tra campionamento ad uno stadio stratificato e campionamento semplice

La finalità di un processo di stratificazione, come abbiamo già avuto occasione di dire (Capitolo 13), è quella di aumentare la rappresentatività del campione, vale a dire di ridurre la varianza campionaria delle stime oggetto d'indagine, a parità di dimensione del campione.

Abbiamo poi appena visto che detta varianza è definita da espressioni con struttura formale diversa a seconda del criterio usato per la ripartizione della numerosità complessiva n, negli strati in cui la popolazione è suddivisa.

Ci proponiamo, ora, di esaminare l'azione esercitata dalla stratificazione sulla varianza campionaria, derivante dai due criteri di ripartizione: proporzionale e di Neyman.

Peraltro, potendo far rientrare tra i diversi criteri di formazione del campione anche quello relativo al caso di campione senza stratificazione, verremo a confrontare tra loro non solo i due criteri di ripartizione appena citati, ma anche ciascuno di essi con il campionamento casuale semplice.

Confrontiamo, in primo luogo, il campionamento stratificato proporzionale con quello casuale semplice (Yamane, 1967).

Le espressioni da porre a confronto sono la (19) e la (25), che riteniamo tuttavia utile riscrivere:

$$V(\hat{Y}_{cs}) = N^2 \quad \frac{N-n}{N}\,\frac{S^2}{n} \tag{42}$$

$$V(\hat{Y}_p) = \frac{N}{n}\left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} N_h\, S_h^2 \tag{43}$$

nelle quali abbiamo introdotto i simboli $\hat{Y}_{cs}$ e $\hat{Y}_p$ per distinguere con maggiore chiarezza la stima diretta relativa al campionamento casuale semplice e la stima diretta ottenuta con il campionamento stratificato proporzionale.

Consideriamo ora la nota identità:

$$S^2 = \frac{1}{N-1}\sum_{h=1}^{H}(N_h - 1)\,S_h^2 + \frac{1}{N-1}\sum_{h=1}^{H} N_h\,(\overline{Y}_h - \overline{Y})^2 \tag{44}$$

Quando $N \gg 1$ e $N_h \gg 1$ la (44) può porsi nella forma:

$$S^2 = \frac{1}{N}\sum_{h=1}^{H} N_h\, S_h^2 + \frac{1}{N}\sum_{h=1}^{H} N_h\,(\overline{Y}_h - \overline{Y})^2 \tag{45}$$

Introducendo poi la (45) nella (42) si deduce:

$$V(\hat{Y}_{cs}) = \left(\frac{N-n}{N}\right)\left[\frac{N}{n}\sum_{h=1}^{H} N_h\, S_h^2 + \frac{N}{n}\sum_{h=1}^{H} N_h\,(\overline{Y}_h - \overline{Y})^2\right] \tag{46}$$

e tenendo presente la (43) si ricava immediatamente che:

$$V(\hat{Y}_{cs}) = V(\hat{Y}_p) + \frac{N-n}{n}\sum_{h=1}^{H} N_h\,(\overline{Y} - \overline{Y})^2 \tag{47}$$

Da tale risultato si deduce pertanto che se la popolazione è ripartita in più strati ed il campione è formato secondo il criterio

proporzionale, si ottiene una varianza generalmente minore di quella che si otterrebbe con il campionamento casuale semplice e tanto più piccola quanto maggiore è la variabilità tra le medie $\hat{Y}_h$ dei singoli strati. Conseguentemente, una riduzione di $V(\hat{Y}_p)$ si ottiene esaltando il secondo addendo figurante a secondo membro della (47), ossia ripartendo la popolazione in strati il più possibile omogenei rispetto alla variabile oggetto di studio y.

Per quanto riguarda il confronto tra i due criteri di ripartizione proporzionale e di Neyman è opportuno in primo luogo riscrivere la (43) nella forma:

$$V(\hat{Y}_p) = \frac{N}{n} \sum_{h=1}^{H} N_h S_h^2 - \sum_{h=1}^{H} N_h S_h^2 \qquad (48)$$

Aggiungendo e sottraendo nella (48) la quantità:

$$\frac{1}{n} \left(\sum_{h=1}^{H} N_h S_h \right)^2 \qquad (49)$$

segue immediatamente che la (48) può riscriversi come:

$$V(\hat{Y}_p) = V(\hat{Y}_N) + \frac{N}{n} \left[\sum_{h=1}^{H} N_h S_h^2 - \frac{1}{N} \left(\sum_{h=1}^{H} N_h S_h \right)^2 \right] \qquad (50)$$

nella quale abbiamo introdotto il simbolo $\hat{Y}_N$ per indicare la stima diretta relativa al campionamento stratificato secondo Neyman. Ponendo ancora:

$$\overline{S} = \frac{1}{N} \sum_{h=1}^{H} N_h S_h \qquad (51)$$

segue che la (50) può porsi nella forma:

$$V(\hat{Y}_p) = V(\hat{Y}_N) + \frac{N}{n} \sum_{h=1}^{H} N_h (S_h - \overline{S})^2 \qquad (52)$$

dalla quale si deduce che, se il campione è formato secondo il criterio di Neyman, si ottiene una varianza generalmente minore di quella che si otterrebbe con il campione proporzionale, e tanto più piccola quanto maggiore risulta la varianza tra gli scarti quadratici medi.

Infine, il confronto tra il campione formato secondo il criterio di Neyman ed il campione semplice è immediato.

Infatti sostituendo la (52) nella (47) si ha:

$$V(\hat{Y}_{cs}) = V(\hat{Y}_N) + \left[\frac{N-n}{n} \sum_{h=1}^{H} N_h (\overline{Y}_h - \overline{Y})^2 + \right. \qquad (53)$$

$$\left. + \frac{N}{n} \sum_{h=1}^{H} N_h (S_h - \overline{S})^2 \right]$$

La (53) suggerisce che se il campione è formato secondo il criterio di Neyman si ottiene una varianza minore di quella che si otterrebbe con il campione semplice, e tanto più piccola quanto maggiore risultano sia la varianza tra le medie dei singoli strati sia la varianza tra gli scarti quadratici medi.

Campionamento a due stadi con stratificazione delle unità primarie

In un disegno di campionamento a due stadi, con stratificazione al livello delle unità primarie, la logica di base per il calcolo della dimensione campionaria è analoga a quella che governa l'allocazione ottimale in un disegno di campionamento ad uno stadio stratificato.

Il problema, cioè, si può porre nei seguenti modi alternativi (Sukhatme e Sukhatme, 1970):

a) determinare la dimensione del campione che, per un prestabilito valore del costo totale dell'indagine, rende minima la varianza di campionamento della stima $\hat{Y}$;
b) determinare la dimensione del campione che, per un prestabilito valore della varianza di campionamento della stima $\hat{Y}$, rende minimo il costo totale dell'indagine.

Per quanto riguarda il costo totale, nelle diverse fasi di un'indagine campionaria, si sostengono spese di vario genere classificabili nel modo seguente:

— spese generali, indipendenti dalle modalità di estrazione e dalla dimensione del campione;

— spese dipendenti dal numero di unità primarie incluse nel campione;
— spese dipendenti dal numero di unità secondarie incluse nel campione.

Le spese generali riducono semplicemente la disponibilità finanziaria; le altre spese sono generalmente proporzionali rispettivamente ai numeri di unità di primo e di secondo stadio incluse nel campione.

Pertanto, indicando con c_0 l'insieme delle spese generali, con c_{1h} e c_{2h} rispettivamente i costi di inclusione nel campione di un'unità primaria e di un'unità secondaria nello strato h, il costo totale C dell'indagine può definirsi nel modo seguente:

$$C = c_0 + \sum_{h=1}^{H} c_{1h}\, n_h + \sum_{h=1}^{H} c_{2h} \sum_{i=1}^{n_h} m_{hi} \qquad (54)$$

Osserviamo che in alcune indagini può esistere un costo di trasferimento dall'una all'altra unità primaria che non è sempre agevole inglobare nel costo c_{1h}; in tali casi si richiede lo studio di funzioni di costo più generali della (54) (Hansen, Hurwitz e Madow, 1953).

Per quanto concerne la definizione dell'espressione della varianza di campionamento da utilizzare ai fini del nostro studio bisogna evidentemente definire gli obiettivi dell'indagine e precisare le caratteristiche generali della strategia campionaria mediante la quale ottenere gli obiettivi stessi.

A tal fine, supponiamo che l'obiettivo della indagine sia l'ottenimento di una stima del totale del carattere y, definito da:

$$Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{hij} \qquad (55)$$

Supponiamo inoltre che per l'effettuazione dell'indagine si sia deciso di ricorrere all'impiego di un campione a due stadi, stratificato al primo stadio, con selezione sia delle unità primarie che di quelle secondarie secondo un meccanismo probabilistico del tipo: senza reimmissione con probabilità uguali.

Immaginiamo ancora che per la determinazione di una stima del totale Y si voglia adottare uno stimatore diretto, espresso da:

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{M_{hi}} K_{hi}\, Y_{hij} \qquad (56)$$

in cui:

$$K_{hi} = \frac{N_h}{n_h} \frac{M_{hi}}{m_{hi}} \qquad (57)$$

Per quanto premesso sopra, alla luce delle considerazioni svolte nei precedenti capitoli, la varianza della stima $\hat{Y}$ è definita dall'espressione:

$$V(\hat{Y}) = \sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h} + \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{N_h} M_{hi}^2 \frac{M_{hi} - m_{hi}}{M_{hi}} \frac{S_{hi}^2}{m_{hi}} \qquad (58)$$

Conviene ora attribuire a tale espressione una forma più appropriata per il seguito, che consente di rendere più agevole la nostra analisi.

A tale scopo supponiamo che — in ciascuno strato h (h = 1, ..., H) — per la ripartizione della dimensione campionaria complessiva delle unità secondarie tra le unità primarie campione si voglia utilizzare il *criterio proporzionale,* secondo il quale da ogni unità primaria si seleziona la medesima frazione di unità secondarie. Più precisamente, assumiamo che:

$$\frac{m_{h1}}{M_{h1}} = \cdots = \frac{m_{hi}}{M_{hi}} = \cdots = \frac{m_{hN_h}}{M_{hN_h}} = f_{2h} \qquad (59)$$

con h = 1, ..., H.

La (59) può essere posta nella forma:

$$\frac{\frac{1}{N_h}\left(m_{h1} + \cdots + m_{hi} + \cdots + m_{hN_h} \right)}{\frac{1}{N_h}\left(M_{h1} + \cdots + M_{hi} + \cdots + M_{hN_h} \right)} = f_{2h} \qquad (60)$$

che può riscriversi nella forma semplificata:

$$\frac{m_{hi}}{M_{hi}} = \frac{\overline{m}_h}{\overline{M}_h} = f_{2h} \qquad (i = 1, \ldots, N_h \; ; h = 1, \ldots, H) \qquad (61)$$

in cui $\overline{M}_h$ è la media di unità secondarie per unità primaria nello strato h, che può essere considerata anche come il valore atteso delle M_{hi}. Essa può essere scritta come:

$$\overline{M}_h = \frac{M_h}{N_h} = \frac{1}{N_h} \sum_{i=1}^{N_h} M_{hi} \qquad (62)$$

In modo analogo, il simbolo $\overline{m}_h$ può essere considerato come il valore atteso delle m_{hi}. È opportuno sottolineare che il significato di $\overline{m}_h$ è diverso da quello espresso da:

$$\overline{m}'_h = \frac{1}{n_h} \sum_{i=1}^{n_h} m_{hi} \qquad (63)$$

che è semplicemente una media campionaria delle n_h unità secondarie $m_{h1}, \ldots, m_{hi}, \ldots, m_{hnh}$.

Sotto l'ipotesi (59), introducendo la (61) nella (58), segue immediatamente che $V(\hat{Y})$ può riscriversi nella forma:

$$V(\hat{Y}) = \sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h} +$$

$$+ \sum_{h=1}^{H} \frac{N_h}{n_h} \frac{\overline{M}_h}{\overline{m}_h} \frac{\overline{M}_h - \overline{m}_h}{\overline{M}_h} \sum_{i=1}^{N_h} M_{hi} S_{hi}^2 \qquad (64)$$

Attraverso semplici passaggi il secondo addendo della (64) può riscriversi in una forma più conveniente.

Si ha:

$$\sum_{h=1}^{H} \frac{N_h}{n_h} \frac{\overline{M}_h}{\overline{m}_h} \frac{\overline{M}_h - \overline{m}_h}{\overline{M}_h} \sum_{i=1}^{N_h} M_{hi} S_{hi}^2 =$$

$$= \sum_{h=1}^{H} \frac{M_h}{n_h \overline{m}_h} \frac{\overline{M} - \overline{m}_h}{\overline{M}_h} \sum_{i=1}^{N_h} M_{hi} S_{hi}^2 = \qquad (65)$$

$$= \sum_{h=1}^{H} \frac{M_h^2}{n_h \overline{m}_h} \frac{\overline{M} - \overline{m}_h}{\overline{M}_h} \frac{1}{N_h \overline{M}_h} \sum_{i=1}^{N_h} M_{hi} S_{hi}^2 =$$

$$= \sum_{h=1}^{H} \frac{M_h^2}{n_h \overline{m}_h} \frac{\overline{M} - \overline{m}_h}{\overline{M}_h} S_{2h}^2$$

in cui si è posto:

$$S_{2h}^2 = \frac{1}{N_h \overline{M}_h} \sum_{i=1}^{N_h} M_{hi} S_{hi}^2 \qquad (66)$$

In definitiva la (64) assume l'espressione:

$$V(\hat{Y}) = \sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h} + \sum_{h=1}^{H} \frac{M_h^2}{n_h \overline{m}_h} \frac{\overline{M}_h - \overline{m}_h}{\overline{M}_h} S_{2h}^2 \qquad (67)$$

Dalle precedenti considerazioni discende che il costo totale C, espresso dalla (54), varia al variare delle unità selezionate al primo stadio; si dovrà pertanto fare nel seguito riferimento al costo totale medio, che può definirsi nel modo seguente:

$$\overline{C} = M(C) = c_0 + \sum_{h=1}^{H} c_{1h}\, n_h + \sum_{h=1}^{H} M\left(c_{2h} \sum_{i=1}^{n_h} m_{hi}\right) =$$

$$= c_0 + \sum_{h=1}^{H} c_{1h}\, n_h + \sum_{h=1}^{H} c_{2h}\, n_h \quad M\left(\frac{1}{n_h} \sum_{i=1}^{n_h} m_{hi}\right) = \qquad (68)$$

$$= c_0 + \sum_{h=1}^{H} c_{1h}\, n_h + \sum_{h=1}^{H} c_{2h}\, n_h\, \overline{m}_h$$

Premesso quanto sopra, ci proponiamo ora — seguendo il criterio a) — di determinare i valori di n_h e $\overline{m}_h$ che rendono minima la varianza $V(\hat{Y})$, definita dalla (67), sotto il vincolo di un prefissato costo totale medio $\overline{C}^*$, espresso dalla (68).

Formalmente deve, in conseguenza, risolversi il seguente problema di minimo condizionato:

$$\begin{aligned} &\min V(\hat{Y}) = \min\,[V(n_h, \overline{m}_h\,;\, h = i, \ldots., H)] \\ &\overline{C}\,(n_h, \overline{m}_h\,;\, h = 1, \ldots., H) = \overline{C}^* \end{aligned} \qquad (69)$$

Per risolvere tale problema di minimo condizionato, si può ricorrere alla regola dei moltiplicatori di Lagrange (Pizzetti, 1962) secondo cui un punto di minimo condizionato per la funzione (67) è un punto di minimo libero per la funzione:

$$G\,(n_h, \overline{m}_h\,;\, h = 1, \ldots., H) = V(\hat{Y}) + \lambda\,(\overline{C} - \overline{C}^*) \qquad (70)$$

in cui $V(\hat{Y})$ e $\overline{C}$ sono fornite dalle espressioni (67) e (68), essendo λ un parametro costante.

Derivando quindi la (70) rispetto ad n_h ed $\overline{m}_h$ e ponendo le derivate uguale a zero, si ottiene:

$$\frac{\partial G}{\partial n_h} = -\frac{N_h^2\, S_h^2}{n_h^2} - \frac{M_h^2\, S_{2h}^2}{n_h^2\, \overline{m}_h} + \frac{M_h^2\, S_{2h}^2}{n_h^2\, \overline{M}_h} + \lambda\,(c_{1h} + c_{2h}\, \overline{m}_h) = 0 \qquad (71)$$

$$\frac{\partial G}{\partial \overline{m}_h} = -\frac{M_h^2\, S_{2h}^2}{n_h\, \overline{m}_h^2} + \lambda\, c_{2h}\, n_h = 0 \qquad (72)$$

Dalla seconda equazione del sistema sopra descritto si ricava:

$$n_h^2 = \frac{M_h^2\, S_{2h}^2}{\lambda\, c_{2h}\, \overline{m}_h^2} \qquad (73)$$

Sottraendo poi la (72), moltiplicata per $\overline{m}_h$, dalla (71), moltiplicata per n_h, si ottiene:

$$n_h^2 = \frac{1}{\lambda\, c_{1h}} \left(N_h^2\, S_h^2 - \frac{M_h^2\, S_{2h}^2}{\overline{M}_h}\right) \qquad (74)$$

Dalla relazione che si ottiene imponendo l'uguaglianza tra la (73) e la (74) consegue subito:

$$\overline{m}_h = \overline{M}_h \sqrt{\frac{c_{1h}}{c_{2h}} \frac{S_{2h}^2}{S_h^2 - \overline{M}_h\, S_{2h}^2}} \qquad (75)$$

Infine, per determinare il valore di n_h dalla (73) si ha:

$$n_h = \frac{M_h\, S_{2h}}{\sqrt{\lambda}\, \sqrt{c_{2h}}\, \overline{m}_h} \qquad (76)$$

Sostituendo la (76) nella funzione dei costi si ottiene poi:

$$\sqrt{\lambda} = \frac{1}{\overline{C}^* - c_0} \sum_{h=1}^{H} \frac{M_h\, S_{2h}}{\overline{m}_h\, \sqrt{c_{2h}}}\, (c_{1h} + c_{2h}\, \overline{m}_h) \qquad (77)$$

che introdotta nella (76) conduce alla relazione:

$$n_h = \frac{(\overline{C}^* - c_0)\, M_h\, S_{2h}/\overline{m}_h\, \sqrt{c_{2h}}}{\sum_{h=1}^{H} (c_{1h} + c_{2h}\, \overline{m}_h)\, (M_h\, S_{2h}/\overline{m}_h\, \sqrt{c_{2h}})} \qquad (78)$$

Il valore di n_h fornito da quest'ultima relazione ed il valore di $\overline{m}_h$ dato dalla (75) risolvono quindi il problema della minimizzazione della varianza $V(\hat{Y})$ sotto la condizione che il costo totale medio abbia valore prefissato $\overline{C}^*$.

Le soluzioni analitiche, definite dalle (75) e (78), hanno il pregio di mettere in evidenza in qual modo la dimensione del campione dipende dai vari fattori che su di essa influiscono, come meglio vedremo in seguito. È utile osservare che la dimensione del campione, però, si può anche determinare in maniera diversa mediante semplice calcolo numerico, conveniente (in quanto più rapido ed efficace) soprattutto quando si considerano funzioni di costo più generali che comportano complesse soluzioni iterative.

Più precisamente, si parte dall'espressione (78) che esprime n_h in funzione di $\overline{m}_h$, essendo note e costanti le altre quantità coinvolte nella formula medesima. Attraverso la (78) si costruisce, in funzione di opportuni valori assegnati ad $\overline{m}_h$, una tabella dei corrispondenti valori di n_h e della varianza della stima $\hat{Y}_h$, definita ancora dall'espressione (67) salvo l'eliminazione del simbolo di sommatoria. In tal modo, per ciascuno strato h (h = 1, ..., H), si individuano rapidamente i valori di n_h ed $\overline{m}_h$ ai quali corrisponde la varianza minima.

Ritorniamo ora ai risultati espressi dalle relazioni (75) e (78), in quanto esse si prestano ad alcuni commenti.

Supponiamo, in primo luogo, che siano prefissati i rapporti c_{1h}/c_{2h} per ogni h (h = 1, ..., H).

Secondo la (75) sussiste pertanto la relazione di proporzionalità:

$$\overline{m}_h = \frac{S_{2h}^2}{S_h^2 - \overline{M}_h S_{2h}^2} \qquad (79)$$

dalla quale si vede che, per determinati valori di S_{2h}^2 (cioè, della variabilità media dentro le N_h unità primarie), $\overline{m}_h$ decresce al crescere di S_h^2 (cioè, della variabilità tra le stesse) e cresce in corrispondenza, secondo la (78), il valore n_h. Ne segue, pertanto, la regola secondo la quale al crescere di S_h^2, rispetto ad S_{2h}^2, diviene più efficiente ed economico il campionamento stratificato, in quanto il valore ottimale di n_h tende ad N_h; nel caso limite, $n_h = N_h$, ogni unità primaria costituisce strato a sé stante.

Se invece consentiamo anche ai rapporti c_{1h}/c_{2h} di variare, dalla (75) si evince che se al decrescere del fattore $S_{2h}^2/S_h^2 - \overline{M}_h S_{2h}^2$, il rapporto c_{1h}/c_{2h} cresce convenientemente, $\overline{m}_h$ può rimanere costante o addirittura crescere. In concreto, ciò significa che se anche la variabilità tra le unità primarie risulta elevata rispetto a quella dentro le unità stesse, ma il costo c_{2h} è molto piccolo rispetto al costo c_{1h}, può non essere conveniente il campionamento stratificato in cui ogni unità primaria costituisce strato a sé.

A conclusione delle precedenti considerazioni si ritiene utile aggiungere che i campioni formati secondo il criterio sopra descritto sono noti in letteratura con il nome di *campioni autoponderati a livello di singolo strato*.

Infatti, ricordando che in generale un campione è definito autoponderante se la probabilità di inclusione è uguale per tutte le unità, segue immediatamente che il tipo di campionamento appena esaminato è autoponderante in ciascuno degli H strati, poiché ciascuna unità secondaria ha la stessa probabilità di inclusione a prescindere dall'unità primaria a cui appartiene.

In termini analitici si ha:

$$f_{2hi} = f_{2h} = \frac{m_{hi}}{M_{hi}} = \frac{\overline{m}_h}{\overline{M}_h} = \text{cost} \qquad (80)$$

e quindi:

$$\Pi_{hi} = \frac{n_h}{N_h}\frac{m_{hi}}{M_{hi}} = \text{cost} \qquad (81)$$

per i = 1, ..., N_h e per ogni h (h = 1, ..., H).

Un altro criterio (Desabie, 1959) comunemente seguito nella formazione di un campione a due stadi, stratificato al livello di unità primarie, è quello secondo cui dalle n_h unità primarie campione viene selezionato un numero costante di unità secondarie, ossia:

$$m_{h1} = \cdots = m_{hi} = \cdots = m_{hn_h} = \overline{m}_h \qquad (82)$$

In tal caso, la probabilità di inclusione di una unità secondaria dell'unità primaria i dello strato h risulta uguale a:

$$f_{2hi} = \frac{m_{hi}}{M_{hi}} = \frac{\overline{m}_h}{M_{hi}} \qquad (83)$$

Da ciò, volendo formare un campione autoponderante in ogni strato, la convenienza di selezionare le n_h unità primarie con probabilità proporzionale alla dimensione, assumendo come misura l'ammontare di unità secondarie M_{hi} (i = 1, ..., N_h). Posto $Z_{hi} = M_{hi}/M_h$, si ha infatti:

$$\Pi_{hi} = n_h Z_{hi} \frac{m_{hi}}{M_{hi}} = n_h \frac{M_{hi}}{M_h}\frac{\overline{m}_h}{M_{hi}} = \frac{n_h \overline{m}_h}{M_h} \qquad (84)$$

dalla quale appare chiaramente che la probabilità di inclusione nel campione è uguale per tutte le unità secondarie appartenenti ad un dato strato h (h = 1, ..., H).

In questa situazione lo stimatore diretto, espresso dalla (56), assume la forma:

$$\hat{Y} = \sum_{h=1}^{H} K_h \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} Y_{hij} = \sum_{h=1}^{H} \hat{Y}_h \qquad (85)$$

in cui:

$$K_h = K_{hi} = \frac{1}{n_h Z_{hi}} \frac{M_{hi}}{m_{hi}} = \frac{M_h}{n_h \overline{m}_h} \qquad (86)$$

Ai fini degli sviluppi successivi, per rendere più agevole la trattazione, faremo l'ipotesi che le unità primarie siano estratte con reimmissione. Sotto tale ipotesi la varianza dello stimatore $\hat{Y}$ è fornita dall'espressione:

$$V(\hat{Y}) = \sum_{h=1}^{H} V(\hat{Y}_h) = \sum_{h=1}^{H} \left[\frac{1}{n_h} \sum_{i=1}^{N_h} Z_{hi} \left(\frac{Y_{hi}}{Z_{hi}} - Y_h \right)^2 + \right.$$

$$\left. + \sum_{i=1}^{N_h} \frac{1}{n_h Z_{hi}} M_{hi}^2 \frac{M_{hi} - m_{hi}}{M_{hi}} \frac{S_{hi}^2}{m_{hi}} \right] \qquad (87)$$

Trascurando poi il fattore di correzione finito relativo al secondo stadio di campionamento (nelle situazioni concrete in genere è $m_{hi} \ll M_{hi}$), la (87) può essere posta nella forma:

$$V(\hat{Y}) = \sum_{h=1}^{H} \left[\frac{S_h^2}{n_h} + \sum_{i=1}^{N_h} \frac{M_{hi}}{n_h Z_{hi} \overline{m}_h} M_{hi} S_{hi}^2 \right] =$$

$$= \sum_{h=1}^{H} \left[\frac{S_h^2}{n_h} + \frac{M_h^2}{n_h \overline{m}_h} \frac{1}{M_h} \sum_{i=1}^{N_h} M_{hi} S_{hi}^2 \right] = \qquad (88)$$

$$= \sum_{h=1}^{H} \left[\frac{S_h^2}{n_h} + \frac{M_h^2}{n \overline{m}_h} S_{2h}^2 \right]$$

in cui abbiamo posto:

$$S_h^2 = \sum_{i=1}^{N_h} Z_{hi} \left(\frac{Y_{hi}}{Z_{hi}} - Y_h \right)^2 \; ; \; S_{2h}^2 = \frac{1}{M_h} \sum_{i=1}^{N_h} M_{hi} S_{hi}^2 \qquad (89)$$

Per quanto riguarda la funzione dei costi, tenendo presente che in ogni strato è $m_{hi} = \overline{m}_h$ = costante, osserviamo che il costo d'inclusione complessivo (relativo alle $n_h \overline{m}_h$ unità secondarie) non varia al variare delle unità primarie scelte, così come accadeva nel criterio proporzionale. Pertanto, dalla (54) discende subito che il costo totale è dato da:

$$C = c_0 + \sum_{h=1}^{H} c_{1h} n_h + \sum_{h=1}^{H} c_{2h} n_h \overline{m}_h \qquad (90)$$

Premesso quanto sopra, supponiamo di voler determinare i valori di n_h e $\overline{m}_h$ che rendono minima la varianza, definita dalla (88), sotto il vincolo che il costo totale C, espresso dalla (90), sia uguale ad un prestabilito valore C^*.

Ricorrendo al procedimento applicato in precedenza si trova, mediante semplici passaggi, che i valori ottimali di n_h e $\overline{m}_h$ sono rispettivamente forniti dalle espressioni:

$$n_h = \frac{(C^* - c_0) M_h S_{2h} / \overline{m}_h \sqrt{c_{2h}}}{\sum_{h=1}^{H} (c_{1h} + c_{2h} \overline{m}_h) (M_h S_{2h} / \overline{m}_h \sqrt{c_{2h}})} \qquad (91)$$

$$\overline{m}_h = \sqrt{\frac{c_{1h}}{c_{2h}} \frac{M_h^2 S_{2h}^2}{S_h^2}} \qquad (92)$$

Passiamo ora a considerare un caso particolare molto importante, caratterizzato dalle seguenti condizioni:

— le unità, in ciascuno stadio, sono estratte senza reimmissione e probabilità uguali;
— la frazione di campionamento per le unità secondarie, in ogni strato, è costante; ossia, in formula:

$$f_{2hi} = f_{2h} = \frac{m_{h1}}{M_{h1}} = \cdots = \frac{m_{hi}}{M_{hi}} = \cdots = \frac{m_{hN_h}}{M_{hN_h}} \qquad (93)$$

— da ogni strato viene selezionato la stesso numero $\hat{n}$ di unità primarie, ossia:

$$\hat{n} = n_h = \frac{n}{H} \tag{94}$$

in cui n indica il numero complessivo di unità primarie campione;
— gli strati hanno la medesima ampiezza $\hat{M}$, cioè:

$$\hat{M} = M_h = \sum_{h=1}^{H} M_{hi} = \frac{M}{H} \tag{95}$$

— il campione è completamente autoponderante, nel senso che ciascuna unità secondaria ha la stessa probabilità di inclusione indipendentemente dall'unità primaria e dallo strato a cui appartiene; ossia:

$$\Pi = \Pi_{hi} = \frac{n_h}{N_h} \frac{m_{hi}}{M_{hi}} = \text{cost} \tag{96}$$

La (96), tenendo presente la (61) e le precedenti ipotesi, può scriversi nella forma equivalente:

$$\Pi = \frac{\hat{n}}{N_h} \frac{\overline{m}_h}{\overline{M}_h} = \frac{\hat{n}\,\overline{m}_h}{M_h} = \frac{\hat{n}\,\overline{m}_h}{\hat{M}} = \frac{n\,\overline{m}_h}{M} \tag{97}$$

dalla quale segue:

$$\overline{m}_h = \frac{\Pi\, M}{n} = \text{cost} \tag{98}$$

ossia, il valore atteso $\overline{m}_h$ è costante su tutti gli strati; indichiamo tale valore con il simbolo $\overline{m}$. Pertanto la (97) può riscriversi come:

$$\Pi = \frac{n\,\overline{m}}{M} \tag{99}$$

Alla luce delle posizioni fatte la varianza di campionamento dello stimatore:

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi}\, Y_{hij} = \frac{M}{n\,\overline{m}} \sum_{h=1}^{H} \sum_{i=1}^{\hat{n}} \sum_{j=1}^{m_{hi}} Y_{hij} \tag{100}$$

con:

$$K_{hi} = \frac{1}{\Pi} = \frac{M}{n\,\overline{m}} \tag{101}$$

può essere posta nella forma:

$$V(\hat{Y}) = \sum_{h=1}^{H} V(\hat{Y}_h) = \sum_{h=1}^{H} \left[N_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h} + \right.$$

$$\left. + \frac{N_h}{n_h} \sum_{i=1}^{N_h} M_{hi}^2 \frac{M_{hi} - m_{hi}}{M_{hi}} \frac{S_{hi}^2}{m_{hi}} \right] = \sum_{h=1}^{H} \left[\frac{N_h\,(H\,N_h - n)}{n} S_h^2 + \right.$$

$$\left. + \frac{M}{n\,\overline{m}} \frac{M - \overline{m}\,H\,N_h}{H} S_{2h}^2 \right] \tag{102}$$

Indichiamo poi con:

$$\overline{C} = c_0 + c_1\, n + c_2\, n\, \overline{m} \tag{103}$$

la funzione dei costi in cui c_1 e c_2 rappresentano rispettivamente i costi di inclusione nel campione di un'unità primaria e di un'unità secondaria.

Nella situazione cosi delineata il sistema (69) diviene:

$$\begin{cases} \min V(\hat{Y}) = \min [V(n, \overline{m})] \\ \overline{C}\,(n, \overline{m}) = \overline{C}^* \end{cases} \tag{104}$$

Seguendo il procedimento già utilizzato si ricava che i valori ottimali di n ed $\overline{m}$ sono rispettivamente definiti dalle espressioni:

$$n = \frac{\overline{C}^* - c_0}{c_1 + c_2\, \overline{m}} \tag{105}$$

$$\bar{m} = \sqrt{\frac{\hat{M}^2 \sum_{h=1}^{H} S_{2h}^2}{\frac{c_1}{c_2} \sum_{h=1}^{H} N_h^2 \, (S_h^2 - \bar{M}_h \, S_{2h}^2)}} \qquad (106)$$

Un altro caso particolare di considerevole interesse è quello fondato sulle condizioni seguenti (Hansen, Hurwitz e Madow, 1953):

— le unità primarie sono selezionate con probabilità proporzionale alla dimensione e senza reimmissione; le secondarie, con probabilità uguali e senza reimmissione;
— da ogni strato viene selezionato lo stesso numero $\hat{n}$ di unità primarie:

$$\hat{n} = n_h = \frac{n}{H} \qquad (107)$$

— gli strati hanno la medesima ampiezza $\hat{M}$:

$$\hat{M} = M_h = \frac{M}{H} \qquad (108)$$

— il campione è completamente auto-ponderante, ossia:

$$\Pi = \Pi_{hi} = n_h \, Z_{hi} \, \frac{m_{hi}}{M_{hi}} = \text{cost} \qquad (109)$$

Tenendo presente le suddette condizioni, la (109) può riscriversi nella forma:

$$\Pi = \frac{n}{H} \frac{M_{hi}}{M_h} \frac{m_{hi}}{M_{hi}} = \frac{n \, m_{hi}}{M} \qquad (110)$$

dalla quale si ricava:

$$m_{hi} = \frac{\Pi \, M}{n} \qquad (111)$$

da cui segue immediatamente che m_{hi} = costante per i = 1, ..., N_h ed h = 1, ..., H ; d'ora innanzi indicheremo tale valore costante con il simbolo $\bar{m}$.

Lo stimatore di Y nel presente contesto si scrive:

$$\hat{Y} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hi} \, Y_{hij} = \frac{M}{n \, \bar{m}} \sum_{h=1}^{H} \sum_{i=1}^{\hat{n}} \sum_{j=1}^{\bar{m}} Y_{hij} \qquad (112)$$

in cui:

$$K_{hi} = \frac{M}{n \, \bar{m}} \qquad (113)$$

La varianza di $\hat{Y}$ assumendo, per semplicità, che le unità primarie siano estratte con reimmissione è data da:

$$V(\hat{Y}) = \sum_{h=1}^{H} \left[\frac{1}{n_h} \sum_{i=1}^{N_h} Z_{hi} \left(\frac{Y_{hi}}{Z_{hi}} - Y_h \right)^2 + \right. \qquad (114)$$

$$\left. + \sum_{i=1}^{N_h} \frac{1}{n_h \, Z_{hi}} \, M_{hi}^2 \, \frac{M_{hi} - m_{hi}}{M_{hi}} \, \frac{S_{hi}^2}{m_{hi}} \right]$$

La (114), tenendo le posizioni fatte e trascurando il fattore di correzione finito, può pertanto scriversi nella forma:

$$V(\hat{Y}) = \sum_{h=1}^{H} \left[\frac{H}{n} \sum_{i=1}^{N_h} Z_{hi} \left(\frac{Y_{hi}}{Z_{hi}} - Y_h \right)^2 + \right.$$

$$\left. + \sum_{i=1}^{N_h} \frac{M_{hi}}{n_h \, Z_{hi} \, m_{hi}} \, M_{hi} \, S_{hi}^2 \right] = \sum_{h=1}^{H} \left[\frac{H}{n} \, S_h^2 + \right. \qquad (115)$$

$$\left. + \frac{M}{n \, \bar{m}} \sum_{i=1}^{N_h} M_{hi} \, S_{hi}^2 \right] = \frac{S_1^2}{n} + \frac{S_2^2}{n \, \bar{m}}$$

in cui abbiamo posto:

$$S_1^2 = H \sum_{h=1}^{H} S_h^2 \quad e \quad S_2^2 = M \sum_{i=1}^{N_h} M_{hi} S_{hi}^2 \qquad (116)$$

Quindi, volendo determinare la dimensione del campione che per un costo C* prestabilito fornisca la massima precisione per la stima $\hat{Y}$ basta determinare n ed $\overline{m}$ in modo da rendere minima la varianza $V(\hat{Y})$ con il vincolo espresso dalla funzione di costo definita da:

$$C = c_0 + c_1 n + c_2 n \overline{m} \qquad (117)$$

Da quest'ultima si ricava:

$$n = \frac{C^* - c_0}{c_1 + c_2 \overline{m}} \qquad (118)$$

mentre per la determinazione di $\overline{m}$, dopo aver introdotto la funzione di Lagrange

$$G(n, \overline{m}) = V(\hat{Y}) + \lambda (c_0 + c_1 n + c_2 n \overline{m} - C^*) \qquad (119)$$

e risolto il sistema che si ottiene uguagliando a zero le derivate parziali, si ricava:

$$\overline{m} = \frac{S_2}{S_1} \sqrt{\frac{c_1}{c_2}} \qquad (120)$$

A conclusione di tutte le considerazioni svolte, con riferimento alla determinazione della dimensione del campione nel contesto dei disegni a due stadi, riteniamo opportuno aggiungere quanto segue.

Tutti i risultati raggiunti sono stati ottenuti mediante una formulazione che conduce, per un prefissato costo totale C*, alla determinazione delle dimensioni campionarie che rendono minima la varianza di campionamento $V(\hat{Y})$.

D'altra parte, come abbiamo precedentemente osservato, è possibile adottare una seconda formulazione in base alla quale si determinano le dimensioni campionarie che rendono minimo il costo totale C per un valore prefissato della varianza di campionamento $V(\hat{Y})$.

Le trattazioni analitiche dei due problemi sono in ogni caso strettamente collegate ed è possibile svilupparle in gran parte in modo unitario, nel senso che è agevole trovare, con semplici passaggi, la soluzione sia per l'una che per l'altra formulazione.

Una seconda considerazione può svilupparsi con riferimento alla circostanza in cui il numero di unità primarie N_h (h = 1, ..., H) è sufficientemente grande. In tal caso, infatti, le espressioni precedenti si possono trasformare, in maniera interessante e significativa, introducendo un apportuno coefficiente di omogeneità, che costituisce un'estensione del coefficiente di correlazione intra-classi (Hansen, Hurwitz e Madow, 1953).

A tal fine, tenendo presente l'espressione (67), consideriamo la varianza di $\hat{Y}_h$ definita da:

$$V(\hat{Y}_h) = N_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h} + \frac{M_h^2}{n_h \overline{m}_h} \frac{\overline{M}_h - \overline{m}_h}{\overline{M}_h} S_{2h}^2 \qquad (121)$$

Se n_h è piccolo rispetto ad N_h la (121) può riscriversi come:

$$V(\hat{Y}_h) = N_h^2 \frac{N_h - 1}{N_h} \frac{S_h^2}{n_h} + \frac{M_h^2}{n_h \overline{m}_h} \frac{\overline{M}_h - \overline{m}_h}{\overline{M}_h} S_{2h}^2 \qquad (122)$$

Posto:

$$V_h' = N_h^2 \frac{N_h - 1}{N_h} S_h^2 + M_h^2 \frac{\overline{M}_h - 1}{\overline{M}_h} S_{2h}^2 \qquad (123)$$

si definisce la quantità:

$$(rho)_h = \frac{N_h^2 \frac{N_h - 1}{N_h} S_h^2 - \frac{V_h^1}{\overline{M}_h}}{(\overline{M}_h - 1) \frac{V_h^1}{\overline{M}_h}} \qquad (124)$$

nota con il nome di *coefficiente di omogeneità*.

Utilizzando la (123) e la (124), mediante semplici passaggi, si ricava che il primo addendo della (122) è dato da:

$$N_h^2 \frac{N_h - 1}{N_h} \frac{S_h^2}{n_h} = \frac{V_h'}{\overline{M}_h} [1 + (\overline{M}_h - 1) (rho)_h] \qquad (125)$$

Introducendo poi la (125) nella (122) segue subito che:

$$S_{2h}^2 = \frac{V_h'}{M_h^2} [1 - (rho)_h] \qquad (126)$$

In definitiva, introducendo la (125) e la (126) nella (122) si ottiene che:

$$V(\hat{Y}_h) = \frac{V_h'}{n_h \bar{m}_h} [1 + (\bar{M}_h - 1) (rho)_h] \qquad (127)$$

dalla quale segue poi che $V(\hat{Y})$ può essere posta nella forma alternativa:

$$V(\hat{Y}) = \sum_{h=1}^{H} V(\hat{Y}_h) = \sum_{h=1}^{H} \frac{V_h'}{n \bar{m}_h} [1 + (\bar{M}_h - 1) (rho)_h] \qquad (128)$$

In conclusione, sfruttando la (128) in luogo della (67) e seguendo il procedimento che ha condotto alle (75) e (78), si ricavano le espressioni:

$$n_h = \frac{(C^* - c_0) M_h S_{2h} / \bar{m}_h \sqrt{c_{2h}}}{\sum_{h=1}^{H} (c_{1h} + c_{2h} \bar{m}_h) (M_h S_h / \bar{m}_h \sqrt{c_{2h}})} \qquad (129)$$

$$\bar{m}_h = \sqrt{\frac{c_{1h}}{c_{2h}} \frac{1 - (rho)_h}{(rho)_h}} \qquad (130)$$

che forniscono i valori ottimali di n_h e $\bar{m}_h$.

Dalla (130) si deduce che la dimensione media $\bar{m}_h$ varia nello stesso senso del rapporto tra i costi unitari c_{1h} e c_{2h}, proporzionalmente alla radice quadrata del rapporto stesso, ed in senso inverso a quello del coefficiente di omogeneità $(rho)_h$, proporzionalmente al fattore:

$$\left[\frac{1 - (rho)_h}{(rho)_h}\right]^{1/2} \qquad (131)$$

Vale a dire che $\bar{m}_h$ varia in misura relativamente minore del rapporto c_{1h}/c_{2h} e maggiore del coefficiente $(rho)_h$.

Prima di concludere svolgiamo qualche ulteriore considerazione sul coefficiente rho per meglio chiarirne il significato.

Supponiamo che le N_h unità primarie siano tutte della stessa ampiezza, ossia:

$$M_{h1} = \cdots = M_{hi} = \cdots = M_{hN_h} = \bar{M}_h$$

In tal caso, il coefficiente di omogeneità si semplifica assumento la forma (Hansen, Hurwitz e Madow, 1953):

$$(rho)_h = \frac{\sigma_h^2 - \frac{{}_t\sigma_h^2}{\bar{M}_h}}{\frac{\bar{M}_h - 1}{\bar{M}_h} {}_t\sigma_h^2} \qquad (132)$$

in cui:

$${}_t\sigma_h^2 = \sigma_h^2 + \sigma_{2h}^2 \qquad (133)$$

$$\sigma_h^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} \left(\frac{Y_{hi}}{\bar{M}_h} - \frac{Y_h}{M_h} \right)^2 \qquad (134)$$

$$\sigma_{2h}^2 = \frac{1}{N_h \bar{M}_h} \sum_{i=1}^{N_h} \sum_{j=1}^{\bar{M}_h} \left(Y_{hij} - \frac{Y_{hi}}{\bar{M}_h} \right)^2 \qquad (135)$$

La (132) è nota in letteratura con il nome di coefficiente di correlazione intra-classi, intendendo per classi le unità primarie; alcuni studiosi (Kish, 1965), per evitare confusione, hanno suggerito l'uso del simbolo rho per il coefficiente di correlazione intra-classi e del simbolo roh per il coefficiente di omogeneità.

Per quanto riguarda il coefficiente di correlazione intra-classi è interessante rilevare che la (132) può derivarsi anche con altro ragionamento; detta variante può esprimersi definendo il coefficiente in esame come segue:

$$(rho)_h = \frac{E\,(Y_{hij} - Y_h/M_h)\,(Y_{hil} - Y_h/M_h)}{E\,(Y_{hij} - Y_h/M_h)^2} = \tag{136}$$

$$= \frac{2\sum_{i=1}^{N_h}\sum_{j<l}^{\bar{M}_h}(Y_{hij} - Y_h/M_h)\,(Y_{hil} - Y_h/M_h)}{(\bar{M}_h - 1)\sum_{i=1}^{N_h}\sum_{j=1}^{\bar{M}_h}(Y_{hij} - Y_h/M_h)^2}$$

il cui numeratore si ottiene confrontando tutte le possibili coppie di unità secondarie appartenenti alla stessa unità primaria.

Utilizzando poi le (133), (134) e (135) si trova mediante semplici passaggi la (132).

Quest'ultima, che misura pertanto il grado medio di omogeneità tra le osservazioni interne alle unità primarie, varia nell'intervallo:

$$-\frac{1}{\bar{M}_h - 1} \leqslant (rho)_h \leqslant 1 \tag{137}$$

Il valore massimo si raggiunge per $\sigma^2_{2h} = o$, ossia quando tutte le unità secondarie di ciascuna delle N_h unità primarie hanno lo stesso valore; raggiunge il valore minimo per $\sigma^2_h = 0$.

Per siffatte considerazioni, il coefficiente definito dalla (124) costituisce in generale una misura dell'omogeneità internamente alle unità primarie.

Osserviamo infine che, attraverso un procedimento analogo a quello sopra descritto, è possibile trasformare in funzione di rho anche le espressioni della varianza definite con riferimento agli altri criteri di formazione del campione illustrati in precedenza; l'utilizzazione delle suddette espressioni consente di esprimere il valore ottimale della dimensione campionaria di secondo stadio in funzione del coefficiente di omogeneità.

Campionamento a due stadi con unità auto-rappresentative

Nel formare un campione a due stadi viene generalmente adottata la decisione di includere con certezza nel campione le unità primarie la cui dimensione supera una data soglia S*; tali unità, come abbiamo già sottolineato nel precedente capitolo

13, sono definite auto-rappresentative in quanto ciascuna di esse costituisce strato a sé. Ad esempio, nel campione per l'indagine Istat sulle forze di lavoro le unità auto-rappresentative sono costituite dai comuni con popolazione uguale o superiore a 20.000 abitanti.

L'opportunità di selezionare con certezza le unità primarie di più grandi dimensioni si può far dipendere sia dall'efficienza attesa per le stime che dall'organizzazione del lavoro sul campo.

Nelle pagine che seguono descriveremo le linee metodologiche essenziali per il calcolo della dimensione del campione nel contesto di un campionamento fondato sull'unione dei seguenti due disegni:

- a due stadi, stratificato al livello delle unità primarie, in cui le unità in ciascuno stadio vengono selezionate senza reimmissione e probabilità uguali;
- ad uno stadio stratificato, costituito dalle unità primarie con dimensione superiore ad una data soglia S*. Anche in questo caso supponiamo che le unità in ciascuno strato vengano estratte senza reimmissione con probabilità uguali.

Ai fini degli sviluppi successivi, indichiamo con h (h = 1,..., H) il generico strato per il disegno a due stadi e con l (l = 1,..., L) il generico strato per il disegno ad uno stadio.

Sia poi:

$$Y = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} Y_{hij} + \sum_{l=1}^{L}\sum_{j=1}^{M_l} Y_{lj} \tag{138}$$

il parametro oggetto di stima.

Lo stimatore diretto di Y è definito dall'espressione:

$$\hat{Y} = \sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}}{m_{hi}}\sum_{j=1}^{m_{hi}} Y_{hij} + \sum_{l=1}^{L}\frac{M_l}{m_l}\sum_{j=1}^{m_l} Y_{lj} \tag{139}$$

Sotto l'ipotesi che la frazione di campionamento secondaria sia costante, tenendo presente la (67), segue immediatamente che la varianza di $\hat{Y}$ è data da:

$$V(\hat{Y}) = \sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h} + \sum_{h=1}^{H} \frac{M_h^2}{n_h \bar{m}_h} \frac{\bar{M} - \bar{m}_h}{\bar{M}_h} S_{2h}^2 + \tag{140}$$

$$+ \sum_{l=1}^{H} M_l^2 \frac{M_l - m_l}{M_l} \frac{S_l^2}{m_l}$$

in cui:

$$S_l^2 = \frac{1}{M_l - 1} \sum_{j=1}^{M_l} \left(Y_{lj} - \frac{Y_l}{M_l} \right)^2 \tag{141}$$

Indicando con c_l il costo di inclusione nel campione di un'unità appartente al generico strato l, tenendo presente la (68), il costo totale medio può definirsi attraverso l'espressione:

$$\bar{C} = c_o + \sum_{h=1}^{H} c_{1h} n_h + \sum_{h=1}^{H} c_{2h} n_h \bar{m}_h + \sum_{l=1}^{L} c_l m_l \tag{142}$$

Quindi, volendo determinare la dimensione media del campione che per un costo $\bar{C}^*$ prestabilito fornisca la massima precisione per la stima $\hat{Y}$, occorre determinare n_h, $\bar{m}_h$ ed m_l in modo da rendere minima la varianza $V(\hat{Y})$, definita dalla (140), con il vincolo espresso dalla funzione di costo (142).

Dopo aver introdotto la funzione di Lagrange

$$G(n_h, \bar{m}, m_l) = V(\hat{Y}) + \lambda (\bar{C} - \bar{C}^*) \tag{143}$$

uguagliando a zero le derivate parziali rispetto ad n_h, $\bar{m}_h$ ed m_l si ottiene:

$$\frac{\partial G}{\partial n_h} = - \frac{N_h^2 S_h^2}{n_h^2} - \frac{M_h^2 S_{2h}^2}{n_h^2 \bar{m}_h} + \frac{M_h^2 S_{2h}^2}{n_h^2 \bar{M}_h} + \lambda (c_{1h} + c_{2h} \bar{m}_h) = 0$$

$$\frac{\partial G}{\partial \bar{m}_h} = - \frac{M_h^2 S_{2h}^2}{n_h \bar{m}_h^2} + \lambda n_h c_{2h} = 0 \tag{144}$$

$$\frac{\partial G}{\partial m_l} = - \frac{M_l^2 S_l^2}{\bar{m}_l^2} + \lambda c_l = 0$$

Risolvendo il sistema (144) si ottengono i valori ottimali definiti dalle espressioni:

$$n_h = \frac{N_h \sqrt{S_h^2 - \bar{M}_h S_{2h}^2}}{\sqrt{c_{1h}}} a \tag{145}$$

$$\bar{m}_h = \bar{M}_h \sqrt{\frac{c_{1h}}{c_{2h}} \frac{S_{2h}^2}{S_h^2 - \bar{M}_h S_{2h}^2}} \tag{146}$$

$$m_l = \frac{M_l S_l}{\sqrt{c_l}} a \tag{147}$$

$$a = \frac{1}{\sqrt{\lambda}} = \tag{148}$$

$$= \frac{\bar{C}^* - c_o}{\sum_{h=1}^{H} N_h \sqrt{c_{1h}} \sqrt{S_h^2 - \bar{M}_h S_{2h}^2} + \sum_{h=1}^{H} M_h S_{2h} \sqrt{c_{2h}} + \sum_{l=1}^{L} M_l S_l \sqrt{c_l}}$$

A conclusione di quest'ultimo caso osserviamo che le argomentazioni esposte possono facilmente essere estese agli altri criteri di formazione del campione precedentemente esaminati.

Il caso delle indagini multiscopo

Tutte le precedenti considerazioni si riferiscono al caso in cui l'indagine campionaria si pone l'obiettivo di fornire la stima di un solo parametro della popolazione oggetto d'indagine.

La situazione più ricorrente, tuttavia, è quella di indagini che si pongono molteplici obiettivi, nel senso che sono finalizzate all'ottenimento di stime di diversa natura (totali, frequenze assolute, rapporti, ecc., riferiti sia all'intera popolazione sia a un numero elevato di sub-popolazioni) su una vasta gamma di variabili oggetto di studio.

Per tali indagini, note in letteratura con il nome di *indagini multiscopo*, la determinazione della dimensione campionaria presenta problemi di gran lunga più complessi e delicati rispetto al caso in cui si debba stimare un solo parametro.

La soluzione, come mostrano anche le esperienze condotte in altri Paesi, viene cercata nell'ottica di una filosofia del buon senso e del compromesso fra rigore metodologico e praticità empirica.

Un primo elemento di complessità è dovuto al fatto che un criterio di determinazione della dimensione campionaria può essere efficace per la stima di un dato parametro, ma inefficace per la stima di altri parametri.

Un secondo elemento di complessità, tanto per citarne un altro, nasce quando l'indagine deve produrre stime sia per la popolazione oggetto di rilevazione, sia per particolari sub-insiemi della popolazione medesima. È il caso, ad esempio, della maggior parte delle indagini Istat, che debbono fornire simultaneamente stime nazionali e regionali dello stesso parametro. Per tali indagini, i due obiettivi risultano generalmente in conflitto, nel senso che il campione ottimale per l'intero territorio nazionale (in cui le regioni, pertanto, intervengono come strati) non presenta in genere anche la capacità di soddisfare le attese nel livello di precisione delle stime regionali.

Il problema della determinazione della dimensione del campione e della scelta del criterio di allocazione tra gli strati per indagini multiscopo fu studiato dapprima da Neyman (1934). Egli osservò che se le variabili oggetto di studio sono positivamente correlate, le varianze delle variabili stesse risultano positivamente correlate; in questo caso, l'allocazione secondo Neyman per una variabile risulta efficiente per le altre variabili. Se, invece, le variabili non sono correlate, Neyman suggerisce l'allocazione proporzionale.

In ogni caso, quale che sia il criterio scelto per allocare il campione tra gli strati, resta sempre aperto il problema della scelta della dimensione campionaria fra quelle calcolate.

Infatti, ad ogni parametro oggetto di stima corrisponde una determinata dimensione campionaria che garantisce il livello atteso di precisione prefissato per la stima del parametro stesso. Alla suddetta dimensione campionaria, in base al criterio di allocazione adottato, corrisponde quindi una certa dimensione campionaria per ciascuno degli strati in cui la popolazione è suddivisa.

Per risolvere il problema in esame, Cochran (1977) propone il procedimento seguente:

a) calcolare, in funzione del livello atteso di precisione, l'allocazione ottima per ogni parametro oggetto di stima;
b) trovare per ogni strato il compromesso più ragionevole tra le numerosità calcolate (ad esempio, la media aritmetica).

Nel caso di indagini che devono fornire un numero elevato di stime, alcuni studiosi suggeriscono una procedura simile a quella di Cochran ma comprendente una fase iniziale consistente nel selezionare un sottoinsieme di parametri (quelli più importanti per l'analisi) fra quelli oggetto di stima.

Yates (1953) suggerisce un criterio basato sulla minimizzazione del costo totale $\Sigma\, c_h\, n_h$ sotto il vincolo che le varianze delle stime dei parametri da stimare siano uguali a certe prestabilite quantità. Tale criterio, però, presuppone che il numero di strati sia più grande del numero di parametri oggetto di stima.

Un criterio più ragionevole è proposto da Dalenius (1957): esso si basa sulla minimizzazione del costo totale sotto il vincolo che le varianze delle stime dei parametri da stimare non eccedano certe prefissate quantità.

Riteniamo utile osservare che il problema dell'allocazione ottimale di un campione stratificato può essere anche affrontato mediante l'uso di tecniche di programmazione matematica; nessuna di esse, però, ha mostrato di essere ottimale in ogni caso (Fabbris, 1989).

RIFERIMENTI BIBLIOGRAFICI

Brewer K. W. e Hanif M. (1983), *Sampling with Unequal Probabilities,* Springer-Verlag, New York.

Bureau of the Census (1978), The Current Population Survey: Design and Methodology, *Technical Paper 40,* U.S. Department of Commerce, Washington.

Cassel C. M. Sarndal C. E. e Wretman J. H. (1977), *Foundations of Inference in Survey Sampling,* Wiley, New York.

Castellano V. e Herzel A. (1981), *Elementi di teoria dei campioni,* Edizioni Sistema, Roma.

Cochran W. G. (1977), *Sampling Techniques,* Wiley, New York.

Coppi R. (1979), Alla base dei metodi statistici: *la formalizzazione dei dati,* Quaderni di Statistica Sanitaria, anno II, N. 1, 81-98.

Dalenius T. (1957), *Sampling in Sweden,* Almqvist and Wicksell, Stockholm.

De Cristofaro R. (1979), *Rilevazioni campionarie,* CLUEB, Bologna.

De Lucia L. (1958), *Problemi di tecnica campionaria nelle analisi di mercato. La stima delle proporzioni,* Giuffrè Editore, Milano.

Desabie M. J. (1959), *Theorie et pratique des sondages,* INSEE, Parigi.

Diana G. e Salvan A. (1987), *Campionamento da popolazioni finite,* CLEUP, Padova.

Droesbeke J. J., Fichet B. e Tassi P. (1987), *Les sondages,* Ed. Économica, Paris.

Fabbris L. (1989), *L'indagine campionaria. Metodi, disegni e tecniche di campionamento,* La Nuova Italia Scientifica, Roma.

Falorsi S. (1989), *Stimatori utilizzati nelle indagini Istat condotte sulle famiglie: contributi metodologici e principali risultati empirici, Giornata di studio sul campionamento statistico,* Rapporto tecnico N. 4, Istat, Roma.

Fuller W. A. (1975), Regression analysis for sample survey, *Sankhyā,* C 37, 117-132.

Ghizzetti A. (1965), *Lezioni di Analisi Matematica,* Vol. I, Libreria Eredi Veschi, Roma.

Giusti F. (1983), *Introduzione alla statistica,* Loescher Editore, Torino.

Glasser G. J. (1962), On the Complete Coverage of Large Units in a Statistical Study, in *Review of International Statistical Institute,* 28-38.

Godambe V. B. (1965), *A unified theory of sampling from finite populations,* Journal Royal Statistical Association Society, B. 17, 269-278.
Grigoletto F. (1976), *Appunti di statistica. Parte prima: La stima,* Serie di statistica n. 6, Cleup, Padova.
Grosbras J. M. (1987), *Methodes statistiques des sondages,* Ed. Economica, Paris.
Hansen M. H., Hurwitz W. N. e Madow W. G. (1953), *Sample Survey Methods and Theory*, Wiley, New York.
Hidiroglou M. A. (1979), *On the inclusion of Large Units in Simple Random Samply,* American Statistical Association, Proceedings of the Sections on Survey Research Methods, 305-8.
Holt D. e Smith T. F. M. (1979), *Post-stratification,* Journal of the Royal Statistical Association, A 142, 33-46.
Horvitz D. G. e Thompson D. J. (1952), *A generalization of sampling without replacement from a finite universe*, Journal of the American Statistical Association, 47, 663-685.
Istat (1978), *Rilevazioni campionarie delle forze di lavoro,* Metodi e norme, serie A, N. 15, Istat, Roma.
Istat (1988), *I consumi delle famiglie, anno 1986,* in Collana di informazione, N. 16, Istat, Roma.
Kendall M. G. e Stuart A. (1977), *The Advanced Theory of Statistics*, 3 voll, Ch. Griffin e Co., London.
Kish L. (1965), *Survey Sampling*, Wiley, New York.
Kish L. e Frankel M. R. (1974), *Inference from Complex Samples,* Journal of the Royal Statistical Society, 1-37.
Kish L. e Anderson D. W. (1978), *Multivariate and multipourpose stratification,* Journal of the American Statistical Association, 73, 24-34.
Leti G. (1983), *Statistica descrittiva*, Il Mulino, Bologna.
Little R. J. A. (1983), *Superpopulation Models for Non response,* in Incomplete Data in Sample Surveys, vol. 2, Academic Press, New York.
Murthy M. N. (1967), *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta.
Platek R. e Gray G. B. (1983), *Imputation Methodology: Total survey Error,* in Incomplete Data in sample surveys, vol. 2, Academic Press, New York.
Pizzetti E. (1962), *Lezioni di Analisi Matematica*, Parte II, Libreria Editrice de Santis, Roma.
Politz A. N. e Simmons W. R. (1949), *An attempt to «get not at homes» into the sample without callbacks,* Journal of the American Statistical Association, 44, 9-31.
Pompilj G. (1967), *Teoria dei campioni*, Veschi, Roma.
Rao J. N. K. (1968), *Some small sample results in ratio and regression estimation,* Journal Indian Statistics Association, 6, 160-168.
Royall R. M. (1971), *Linear regression models in finite population sampling theory,* in Foundations of statistical inference, Holt, Rinehart e Winston, Toronto.
Royall R. M. e Cumberland D. (1981), *An empirical study of the ratio estimator and estimators of its variance,* Journal of the American Statistical Association, 76, 66-88.
Russo A. (1982), *Metodologia utilizzata per il riporto dei dati all'universo e per il calcolo degli errori di campionamento, in Indagine statistica sulla ricerca scientifica, anni 1980 e 1981,* Supplemento al Bollettino mensile di statistica, Istat, Roma.
Russo A. (1984a), *Indagine sulle vacanze, i viaggi e gli sport degli italiani nel 1982: Piano della rilevazione campionaria ed errori di campionamento,* in Supplemento al Bollettino Mensile di Statistica, n. 15, Istat, Roma.
Russo A. (1984b), *Considerazioni metodologiche sul problema del riporto dei dati all'universo,* in *Confronti internazionali a nuove indagini in Italia*, Dipartimento di Scienze Statistiche, Università di Padova.
Russo A. (1985), *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,* Quaderni di Discussione, N. 5, Istat, Roma.
Russo A. e Falorsi P. D. (1985), *Rilevazione campionaria delle forze di lavoro: Metodologia del campionamento, calcolo e presentazione degli errori campionari,* Quaderni di Discussione, n. 6, Istat, Roma.
Russo A. (1986a), *Su un metodo di stima dell'effetto ponderazione nei campioni a due stadi con stratificazione delle unità di primo stadio,* Quaderni di Discussione, N. 1, Istat, Roma.
Russo A. (1986b), *Una metodologia per la stima degli effetti stratificazione, clustering, ponderazione e dell'effetto complessivo dal disegno di campionamento nei campioni a due stadi con selezione delle unità primarie con reimmissione e probabilità variabile,* Quaderni di Discussione, N. 2, Istat, Roma.
Russo A. (1986c), *Un metodo di stima dell'effetto della stratificazione nei campioni complessi,* in Atti della XXXIII Riunione della Società Italiana di Statistica, Bari.
Russo A. (1988a), *Descrizione algebrica della procedura di stima, in Indagine sugli sport e sulle vacanze degli italiani nel 1985,* Note e Relazioni, N. 2, Istat, Roma.

Russo A. (1988b), *Un metodo di stima dell'effetto della post-stratificazione nei campioni a due stadi,* in Atti della XXXIV Riunione Scientifica della Società Italiana di Statistica, Siena.

Russo A. e Falorsi P. D. (1989), *Principali aspetti metodologici delle rilevazioni campionarie nel settore agricolo e zootecnico,* in Atti del III Corso Nazionale di Aggiornamento e Formazione Statistica, per il personale regionale e provinciale responsabile delle rilevazioni statistiche nel settore agricolo, Istat, Roma.

Russo A. e Falorsi S. (1989), *Indagine su alcune specie di alberi da frutto ed agrumi, anno 1987: Metodologia del campionamento, procedimento di stima, calcolo e presentazione degli errori di campionamento,* Rapporto interno, Istat, Roma.

Signore M. (1988), *Stima dell'errore di misura: alcune riflessioni sui problemi teorici e pratici per l'applicazione ad indagini su larga scala,* in Atti della XXXIV Riunione Scientifica della Società Italiana di Statistica, Siena.

Singh D. e Chaudhary F. S. (1986), *Theory and Analysis of Sample Survey Designs*, Wiley, New York.

Statistics Canada (1976), *Methodology of the Canadian Labour Force Survey,* Ottawa.

Sukhatme P. V. e Sukhatme B. V. (1970), *Sampling Theory of Survey with Applications*, The Iowa State University Press, Ames, Iowa, U.S.A.

Vajani L. (1969), *Metodi statistici nelle ricerche di mercato*, Etas Kompass, Milano.

Vitali O. (1987), *Elementi di statistica per le scienze sociali,* Cacucci Editori, Bari.

Yamane T. (1967), *Elementary Sampling Theory*, Prentice — Hall, Inc. Englewood Cliffs, New York.

Yates F. e Grundy P. M. (1953), *Selection without replacement from within strata with probability proportional to size,* Journal of the Royal Statistical Society, B 15, 253-261.

Zanella A. (1974), *Elementi di teoria del campionamento da popolazioni finite,* Cleup, Padova.

Zannella F. (1987), *Metodologia, programmi e sperimentazioni relativi alla progettazione di una procedura generalizzata per la stratificazione dei comuni, Commissione di studio avente il compito di formulare proposte in merito alla progettazione e applicazioni di campioni,* Documento n. 1, Istat, Roma.