manuale di tecniche di indagine

5 - tecniche di stima della varianza campionaria

istat
istituto nazionale di statistica

note e relazioni
anno 1989 n. 1

Autore del fascicolo:
Francesco Zannella

Elaborazione computerizzata
delle formule statistiche:
Patrizia De Lellis

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.

INTRODUZIONE

Errori campionari ed altri tipi di errore

Lo scopo principale delle indagini campionarie condotte dall'Istat è quello di fornire le stime di numerosi parametri descrittivi per l'intera popolazione e per subpopolazioni costituite da particolari raggruppamenti territoriali o da sottoclassi individuate dalle caratteristiche strutturali delle unità rilevate.

I risultati dell'indagine vengono, generalmente, presentati sotto forma tabellare con le caselle delle tavole contenenti stime di medie, totali, frequenze assolute o relative, e in alcuni casi stime di rapporti fra totali.

Le stime, che si ottengono come funzioni dei valori osservati sulle unità del campione, possono essere soggette a diversi tipi di errore, classificabili in tre ampie categorie:

a) errori di misura
b) errori di campionamento
c) errori d'implementazione del campione

Gli *errori di misura* derivano dal fatto che si è osservato un valore diverso da quello che s'intendeva misurare e sono presenti anche nelle indagini in cui vengono rilevate tutte le unità della popolazione. Gli errori di questo tipo si possono verificare in qualsiasi fase del processo di produzione del dato statistico: dalla predisposizione del questionario alla rilevazione, della codifica alla registrazione delle risposte.

Gli *errori di campionamento* sono dovuti esclusivamente al fatto che l'indagine è stata condotta su un campione e non sull'intera popolazione. Essi derivano dal processo di estrapolazione dei dati dalle unità campionate alla popolazione e la loro grandezza dipende dal disegno campionario, dalle dimensioni del campione e dal procedimento di stima adottati.

Gli *errori d'implementazione del campione* possono verificarsi per la presenza di errori nelle liste dalle quali devono essere estratte le unità da rilevare (non completezza, duplicazioni, errori nei nominativi o negli indirizzi, ecc.), o quando l'estrazione del campione non viene effettuata seguendo le regole previste dal disegno campionario, o perché alcune unità selezionate non possono essere rilevate. Questi errori producono una modifica nelle probabilità di selezione e possono comportare una distorsione nelle stime.

Da quanto detto deriva che l'errore di campionamento costituisce soltanto una componente dell'errore da cui possono essere affette le stime derivate da un'indagine campionaria, e pertanto il suo valore può essere riguardato come l'estremo inferiore dell'errore complessivo.

Utilizzazione degli errori campionari

Pur con i limiti sopra accennati, gli errori di campionamento hanno una notevole importanza nell'interpretazione e nell'analisi dei risultati di un'indagine campionaria.

La conoscenza degli errori campionari è indispensabile, in primo luogo, nella fase di predisposizione delle tavole che devono essere pubblicate, in quanto dall'esame di tali errori è possibile stabilire fino a quale dettaglio si può scendere nella pubblicazione dei risultati.

Occorre, infatti, tener presente che le stime di parametri riferite a sottoclassi della popolazione vengono derivate da subcampioni, la cui numerosità non sempre è stata programmata a priori in modo da contenere l'errore delle stime entro limiti stabiliti. Può così accadere che per certe suddivisioni della popolazione il campione risulti poco numeroso e gli errori campionari così elevati da rendere le stime non attendibili e quindi non pubblicabili o pubblicabili con particolari avvertenze.

In secondo luogo la conoscenza degli errori di campionamento è essenziale per una corretta utilizzazione dei risultati pubblicati. A tale proposito occorre tener presente che esistono diverse categorie di utilizzatori interessate ad informazioni diversificate sugli errori campionari.

Un primo gruppo è costituito da coloro che non hanno un particolare interesse per la metodologia dell'indagine ed utilizzano i risultati esclusivamente da un punto di vista pratico. Per questo tipo di utilizzatori le informazioni sugli errori campionari possono riguardare soltanto il grado di attendibilità delle principali stime, che può essere riportato in apposite tabelle o essere segnalato direttamente all'interno delle tavole pubblicate, evidenziando con un qualche carattere le stime con errori relativi superiori a certi limiti stabiliti (ad esempio 5%, 10%, 20%).

Un secondo gruppo è rappresentato da esperti che devono analizzare in modo più approfondito i risultati e che pertanto devono essere messi in grado di valutare l'errore campionario per una qualsiasi stima che può essere derivata dall'indagine. Per questa categoria di utilizzatori è necessario, inoltre, fornire gli elementi che consentono di valutare la significatività delle differenze che si riscontrano tra le stime relative a due sottogruppi della popolazione e, più in generale, gli elementi per poter applicare in modo corretto i più comuni test statistici per la verifica delle ipotesi.

Un terzo gruppo va individuato negli statistici esperti di campionamento, che sono interessati alle informazioni necessarie per la programmazione di nuovi campioni. Per questo tipo di utilizzatori occorre fornire elementi per la valutazione dell'efficienza del disegno campionario adottato e degli effetti imputabili alla struttura del campione (stadificazione, stratificazione e ponderazione).

Stima degli errori campionari

Le espressioni per la stima degli errori campionari dipendono sia dal piano di campionamento che dal tipo di stimatore utilizzati.

I piani di campionamento che vengono impiegati nell'Istat sono i più diversi: da quelli ad un stadio stratificato che trovano più frequente applicazione nelle indagini condotte su popolazioni di dimensioni limitate e nelle indagini postali, a quelli più complessi come i disegni campionari adottati per le indagini sulle famiglie che sono a due stadi con stratificazione delle unità di primo stadio.

Gli stimatori che vengono utilizzati sono a volte piuttosto semplici, mentre altre volte l'uso di aggiustamenti, quali la stratificazione a posteriori e la riponderazione per le mancate risposte o il ricorso a particolari procedimenti di stima come il metodo del rapporto, possono portare ad espressioni più complesse.

Per la stima degli errori campionari possono essere seguiti due diversi approcci: il primo basato sulla metodologia standard, il secondo sulle replicazioni del campione.

L'utilizzazione della metodologia standard richiede la determinazione per via analitica delle espressioni delle stime degli errori campionari per ogni specifico stimatore e piano di campionamento adottati.

Nel caso di stimatori lineari, ossia di stimatori esprimibili come combinazioni lineari dei valori osservati sulle unità del campione, ciò non comporta eccessive difficoltà; infatti la teoria mette a disposizione le formule esatte per i più comuni piani di campionamento utilizzati nella pratica.

Per gli stimatori non lineari non si dispone di formule esatte, ed espressioni approssimate possono essere ricavate mediante il metodo della linearizzazione. Questo metodo, basato sullo sviluppo in serie di Taylor dello stimatore, è applicabile a tutti gli stimatori che possono essere espressi come funzioni regolari di totali (rapporti, differenze tra rapporti, coefficienti di regressione e di correlazione). Lo stimatore viene ad essere approssimato mediante una combinazione lineare dei totali a cui vengono applicate le formule specifiche del piano di campionamento adottato.

La tecnica basata sulle replicazioni del campione comprende metodi diversi (gruppi casuali, replicazioni bilanciate ripetute, jackknife e bootstrap) e consiste nel generare dei subcampioni mediante le unità del campione totale. Su ogni subcampione viene stimato il parametro d'interesse adottando la stessa forma funzionale impiegata per il campione totale. La variabilità tra le stime dei subcampioni viene utilizzata per ottenere la stima della variabilità campionaria dello stimatore desunto dal campione totale.

Questa metodologia, che può essere utilizzata per stimare l'errore campionario di un qualsiasi stimatore, sia esso lineare che non lineare, richiede per ogni specifico piano di campionamento la predisposizione di un'apposita tecnica per la generazione dei subcampioni.

Procedure informatiche per il calcolo degli errori campionari

A partire dalla fine degli anni 70 si è cominciato da parte dell'Istat ad affrontare il problema del calcolo degli errori di campionamento in modo sistematico ed organico.

Una prima serie di esperienze è stata effettuata con la messa a punto di metodologie e programmi informatici ad hoc per alcuni disegni campionari utilizzati nelle indagini sulle famiglie.

Allo scopo di superare le difficoltà derivanti dalla predisposizione di metodologie e programmi specifici per ciascun tipo d'indagine, si è cominciato a sperimentare il programma generalizzato «CLUSTERS», basato su un metodo approssimato di stima degli errori, ma applicabile ai più diversi disegni campionari.

Le applicazioni effettuate sui risultati di diverse indagini, hanno confermato le buone prestazioni di questo pacchetto, peraltro già evidenziate in letteratura (cfr. Francis e Sedransk, 1979). Pertanto il CLUSTERS è oggi impiegato dall'Istituto come procedura standard per il calcolo degli errori di campionamento per disegni campionari complessi.

Il software in questo settore della statistica ha avuto negli ultimi anni un notevole sviluppo, per cui oggi sono disponibili programmi informatici generalizzati basati su metodi diversi quali: i gruppi casuali, le replicazioni bilanciate ripetute e il jackknife.

Attualmente sono in corso di sperimentazione alcuni di questi prodotti per valutare l'opportunità, sia in termini di efficienza che di flessibilità e semplicità di utilizzazione, di un loro impiego nell'Istituto.

Presentazione degli errori campionari

Se è una buona regola nel pubblicare i risultati di un'indagine campionaria mettere l'utilizzatore in condizione di valutare l'attendibilità di tutte le stime pubblicate, ciò non significa che devono essere forniti gli errori di campionamento per tutti i parametri stimati e per tutte le possibili classificazioni previste dal piano di pubblicazione dei risultati. Infatti, ciò richiederebbe costi e tempi di elaborazione eccessivi e la stampa di un numero di tavole uguale se non superiore a quello in cui sono riportati i risultati dell'indagine.

Utilizzando la *portabilità* degli errori campionari, ossia la possibilità di poter trasferire da una sottoclasse ad un'altra certe conclusioni circa la variabilità campionaria delle stime, si può limitare il calcolo degli errori campionari a un numero ristretto di sottoclassi, e stimare mediante opportuni modelli gli errori per le altre sottoclassi.

Scopi ed articolazione del fascicolo

In questo fascicolo, dopo aver introdotto i concetti che sono alla base del campionamento probabilistico, vengono descritti:

a) la metodologia standard per la stima degli errori di campionamento e la sua applicazione ai più comuni disegni campionari;

b) alcuni metodi basati sulle replicazioni del campione (gruppi casuali, replicazioni bilanciate ripetute e il loro adattamento ai principali disegni campionari;

c) la metodologia e il programma informatico attualmente impiegati nell'Istat;

d) i metodi che possono essere adottati per la presentazione sintetica degli errori di campionamento;

e) gli standards per la pubblicazione degli errori campionari nei rapporti contenenti i risultati finali di un'indagine.

Il lavoro non rappresenta una trattazione esaustiva delle metodologie che possono essere impiegate per la stima degli errori di campionamento, ma piuttosto una guida ragionata per i reparti responsabili delle indagini, finalizzata alla loro autonomia in questa importante fase di elaborazione dei risultati.

CAPITOLO 1 - IL CAMPIONAMENTO PROBABILISTICO: DEFINIZIONI E SIMBOLOGIA

In questo capitolo vengono riportati sinteticamente i concetti che sono alla base del campionamento probabilistico, inoltre vengono date le definizioni ed introdotta la simbologia che verranno utilizzate in seguito.

Per una trattazione più esaustiva si rimanda al quarto fascicolo del Manuale di tecniche di indagine (Tecniche di campionamento: teoria e pratica).

1. Popolazione e variabili di rilevazione

Etichettazione delle unità

Un'indagine campionaria ha come oggetto un insieme finito di unità che prende il nome di popolazione e viene indicato con $U = \{U_1, \ldots, U_i, \ldots, U_N\}$ dove U_i rappresenta la generica unità e N la numerosità o ampiezza della popolazione. Le unità della popolazione, la cui natura è irrilevante ai fini della trattazione, possono essere persone, famiglie, aziende agricole, imprese industriali o commerciali, ecc.

Anche se nel testo verrà fatto quasi sempre specifico riferimento alle famiglie, è bene sottolineare che ogni procedura riportata può essere applicata ai risultati di una qualsiasi indagine campionaria che utilizzi lo stesso disegno di campionamento cui la procedura si riferisce.

Ciò premesso, per poter effettuare un campionamento da una popolazione finita occorre che le unità della popolazione siano identificabili mediante un'opportuna etichetta. Il modo in cui è possibile effettuare l'etichettazione dipende dal tipo di lista di cui si dispone.

Ad esempio nel caso di un'indagine campionaria sulle famiglie si può considerare la lista come costituita da tutti gli elenchi anagrafici comunali ed assegnare a ciascuna famiglia un'etichetta formata dal codice della provincia, da quello del comune e dal numero progressivo della famiglia all'interno dell'elenco comunale. Così alla 113ma famiglia dell'elenco anagrafico del comune di Chieri (codice 078) della provincia di Torino (codice 01) corrisponderà l'etichetta 01078113.

Il modo più semplice di etichettare le unità della popolazione è, tuttavia, quello di far corrispondere ad ognuna di esse un numero progressivo da 1 a N, cosicché ciascuna unità verrà identificata dal corrispondente numero ordinale. È questo il metodo a cui si farà riferimento; pertanto nel seguito una popolazione di N unità verrà indicata con $U = (1, 2, \ldots, N)$.

La matrice dei dati

Le indagini campionarie condotte dall'Istituto hanno come obiettivo quello di fornire informazioni su numerose caratteristiche della popolazione oggetto di studio, per cui su ogni unità vengono rilevate un gran numero di variabili di natura diversa.

Indicando con p il numero delle variabili oggetto di rilevazione e con $Y_{(r)i}$ il valore che l'r.ma variabile assume nell'unità i.ma, l'insieme delle informazioni relative all'intera popolazione può essere riguardato come una matrice di dati (o file):

Unità	valori delle variabili					
1	$Y_{(1)1}$	$Y_{(2)1}$	...	$Y_{(r)1}$	...	$Y_{(p)1}$
2	$Y_{(1)2}$	$Y_{(2)2}$	...	$Y_{(r)2}$	...	$Y_{(p)2}$
.	.	.		.		.
.	.	.	...	.	...	.
.	.	.		.		.
i	$Y_{(1)i}$	$Y_{(2)i}$	...	$Y_{(r)i}$	...	$Y_{(p)i}$
.	.	.		.		.
.	.	.	...	.	...	.
.	.	.		.		.
N	$Y_{(1)N}$	$Y_{(2)N}$	...	$Y_{(r)N}$	...	$Y_{(p)N}$

dove i vettori riga (o record) rappresentano i valori associati alle singole unità e i vettori colonna le distribuzioni semplici delle singole variabili.

Variabili categoriche

È bene tener presente che spesso alcuni dei caratteri oggetto d'indagine possono essere qualitativi o quantitativi divisi in classi per cui il valore $Y_{(r)i}$, anche se espresso in termini numerici, sta a rappresentare il codice della modalità o della classe di appartenenza dell'unità i.ma. Questi valori non possono essere trattati allo stesso modo di quelli relativi alle variabili quantitative, in quanto su di essi non si possono eseguire le usuali operazioni aritmetiche il cui risultato non avrebbe alcun significato logico.

Per poter procedere ad una trattazione uniforme, qualunque sia la natura dei caratteri oggetto di studio, è necessario effettuare una quantificazione dei caratteri qualitativi mediante l'introduzione delle *variabili indicatrici delle modalità* (o classi).

Una variabile indicatrice costituisce una particolare ricodifica delle modalità di un carattere che assume il valore 1 nelle unità che presentano la modalità considerata e 0 nelle altre.

Così un carattere A che classifica le unità secondo T modalità esaustive e che si escludono mutualmente viene ad essere espresso mediante T variabili indicatrici, una per ciascuna modalità. La variabile indicatrice $I_{(t)}$ relativa alla modalità A_t (t = 1, 2, . . ., T) risulta così definita:

$$I_{(t)i} = \begin{cases} 1 & \text{se l'unità i.ma presenta la modalità } A_t \\ 0 & \text{altrimenti} \end{cases}$$

Esempio 1.1. Si consideri una popolazione costituita da N = 4 unità su ciascuna delle quali vengono rilevati tre caratteri, due quantitativi ($Y_{(1)}$ = spesa mensile per l'alimentazione e $Y_{(2)}$ = spesa mensile totale, entrambi espressi in migliaia di lire) e uno qualitativo ($Y_{(3)}$ = ripartizione geografica con tre modalità così codificate: 1 = nord, 2 = centro 3 = sud):

Unità	$Y_{(1)i}$	$Y_{(2)i}$	$Y_{(3)i}$
1	700	2000	1
2	740	1800	2
3	640	1600	2
4	720	1600	3

Al posto del carattere $Y_{(3)}$ possono essere considerate le variabili indicatrici delle tre modalità. La variabile $I_{(1)}$ indicatrice della modalità «nord» assume il valore 1 in corrispondenza della prima unità e il valore 0 nelle altre due. In modo analogo si ricavano le distribuzioni delle altre due variabili indicatrici. La popolazione viene così ad essere rappresentata dalla seguente matrice di dati:

Unità	$Y_{(1)i}$	$Y_{(2)i}$	$I_{(1)i}$	$I_{(2)i}$	$I_{(3)i}$
1	700	2000	1	0	0
2	740	1800	0	1	0
3	640	1600	0	1	0
4	720	1600	0	0	1

I parametri della popolazione

L'obiettivo di un'indagine campionaria su una popolazione finita non è tanto quello di ricavare informazioni sulle distribuzioni semplici o congiunte delle variabili, quanto quello di fornire le stime di alcuni parametri della popolazione; dove con il termine parametro s'intende indicare una qualsiasi funzione reale dei valori che le variabili di rilevazione assumono nelle unità della popolazione.

Sebbene il campionamento possa venire effettuato per molteplici scopi l'interesse è, generalmente, centrato sui seguenti parametri:

- ammontare totale o medio di un carattere quantitativo;
- frequenza assoluta o relativa delle unità che appartengono ad una determinata classe di un carattere quantitativo o presentano una determinata modalità di un carattere qualitativo;
- rapporto tra due totali o tra due medie;
- rapporto tra due frequenze assolute o relative;
- rapporto tra un ammontare totale (o medio) e una frequenza assoluta (o relativa);
- differenza tra due totali, tra due medie o tra due rapporti.

È bene evidenziare che il caso della stima di una frequenza assoluta (o relativa) può sempre essere ricondotto a quello di un totale (o di una media) introducendo la variabile indicatrice della classe (o della caratteristica) cui la frequenza si riferisce.

La frequenza assoluta delle unità che appartengono alla classe considerata viene a coincidere con il totale della variabile indicatrice e la frequenza relativa con la sua media. Ovviamente il rapporto fra due frequenze assolute (o relative) risulterà uguale al rapporto tra i totali (o le medie) delle variabili indicatrici corrispondenti.

Pertanto nel prosieguo si farà riferimento soltanto ai seguenti parametri: media, totale e rapporto tra due medie o totali, che verranno indicati rispettivamente con $\overline{Y}$, Y e R, mentre verrà utilizzata la lettera θ per indicare un generico parametro.

Le sottoclassi

In genere tra gli obiettivi di un'indagine campionaria non c'è solo quello di fornire le stime dei parametri per l'intera popolazione ma anche per particolari sottoclassi della popolazione stessa; dove per sottoclasse s'intende un insieme di unità che presentano la stessa modalità di un carattere o le stesse modalità di due o più caratteri oggetto di rilevazione.

Possono essere sottoclassi sia le ripartizioni geografiche, le regioni, le province o altri particolari raggruppamenti territoriali che le persone aventi la stessa caratteristica socio-demografica. Esempi di quest'ultimo tipo di sottoclassi sono: la popolazione distinta per sesso, per particolari gruppi di età, secondo le modalità del titolo di studio, della condizione professionale e così via.

Ai fini del campionamento non ha alcuna rilevanza il fatto che le sottoclassi siano definite dalle modalità di un solo carattere o dall'incrocio delle modalità di due o più caratteri, mentre, come si vedrà meglio in seguito, si hanno sviluppi diversi a seconda che sia conosciuto o meno il numero delle unità della popolazione che appartengono alla sottoclasse.

2. Campione e universo dei campioni

Esistono diverse definizioni formali dei termini campione e universo dei campioni, quelle che riteniamo più semplici e più rispondenti alla pratica dell'Istituto sono le seguenti:

a) un campione **c** è un sottoinsieme della popolazione **U**

b) l'universo dei campioni **C** è l'insieme di tutti i possibili campioni che possono essere derivati dalla popolazione.

Se si conteggia l'insieme vuoto e l'insieme coincidente con l'intera popolazione il numero dei possibili campioni è 2^N.

Esempio 1.2. Per una popolazione di N = 4 unità, si hanno $2^4 = 16$ possibili campioni:

$c_1 = \{\cdot\}$	$c_2 = \{1\}$	$c_3 = \{2\}$	$c_4 = \{3\}$
$c_5 = \{4\}$	$c_6 = \{1,2\}$	$c_7 = \{1,3\}$	$c_8 = \{1,4\}$
$c_9 = \{2,3\}$	$c_{10} = \{2,4\}$	$c_{11} = \{3,4\}$	$c_{12} = \{1,2,3\}$
$c_{13} = \{1,2,4\}$	$c_{14} = \{1,3,4\}$	$c_{15} = \{2,3,4\}$	$c_{16} = \{1,2,3,4\}$

dove i numeri entro parentesi indicano le unità della popolazione comprese nel campione e $\{\cdot\}$ sta a rappresentare l'insieme vuoto.

Il numero delle unità che costituiscono il campione prende il nome di ampiezza o numerosità del campione.

Nell'esempio riportato si hanno, oltre all'insieme vuoto, quattro campioni di numerosità 1, sei campioni di numerosità 2, quattro campioni di numerosità 3 e un campione di numerosità 4.

Il disegno campionario

Una procedura campionaria è il processo casuale attraverso il quale si perviene alla selezione del campione. La procedura campionaria assegna ad ogni campione **c** una probabilità P(**c**) di essere estratto.

La distribuzione di probabilità $\{P(\mathbf{c}), \mathbf{c} \in \mathbf{C}\}$ è chiamata disegno campionario (od anche piano di campionamento). Il campionamento conforme alle preassegnate probabilità P(**c**) è detto campionamento probabilistico, dove le probabilità sono numeri non negativi che soddisfano la condizione:

$$P(c) \geq 0 \quad e \quad \sum_{c \in C} P(c) = 1 \qquad [1.1]$$

Esempio 1.3. Sempre con riferimento ad una popolazione di quattro unità si supponga di voler estrarre un campione di numerosità n = 2 mediante un procedimento di selezione senza reimmissione che assegni ad ogni campione la stessa probabilità di essere estratto. I campioni con una numerosità diversa da 2 avranno ovviamente una probabilità uguale a zero, mentre i campioni costituiti da due unità avranno una probabilità di selezione uguale a $1/\binom{4}{2} = 1/6$, dove $\binom{4}{2} = 6$ è il numero dei possibili campioni di numerosità 2.

Pertanto il disegno campionario risulta così definito:

$P(c_1) = 0$	$P(c_2) = 0$	$P(c_3) = 0$	$P(c_4) = 0$
$P(c_5) = 0$	$P(c_6) = 1/6$	$P(c_7) = 1/6$	$P(c_8) = 1/6$
$P(c_9) = 1/6$	$P(c_{10}) = 1/6$	$P(c_{11}) = 1/6$	$P(c_{12}) = 0$
$P(c_{13}) = 0$	$P(c_{14}) = 0$	$P(c_{15}) = 0$	$P(c_{16}) = 0$

Come si può immediatamente verificare la somma della probabilità di selezione è uguale a 1.

Nel seguito, quando verrà preso in considerazione un piano di campionamento si farà riferimento ai soli campioni con probabilità diversa da zero, che costituiscono il supporto del disegno campionario. Nell'esempio fatto il supporto è rappresentato dai 6 campioni di numerosità 2.

Probabilità d'inclusione

Per un determinato disegno campionario si definisce probabilità d'inclusione del primo ordine π_i, la probabilità che l'unità i.ma sia compresa nel campione. La probabilità d'inclusione per l'unità i.ma è data dalla somma della probabilità di tutti i campioni che contengono tale unità.

Così, per il disegno campionario considerato in precedenza, la probabilità d'inclusione della prima unità della popolazione è data da:

$$\pi_1 = P(c_6) + P(c_7) + P(c_8) = 1/2$$

Nello stesso modo si possono ricavare le probabilità d'inclusione delle altre tre unità, ottenendo i seguenti valori:

$$\pi_2 = 1/2 \qquad \pi_3 = 1/2 \qquad \pi_4 = 1/2$$

Si può verificare che la somma delle probabilità d'inclusione è uguale a 2, e coincide con la numerosità del campione.

Si definisce probabilità d'inclusione del secondo ordine π_{ij}, la probabilità che l'unità i.ma e j.ma siano entrambe comprese nel campione. La probabilità d'inclusione delle unità i e j è data dalla somma delle probabilità di tutti i campioni che le contengono.

Per l'esempio 1.3 si hanno i seguenti valori:

$$\pi_{12} = P(c) = 1/6 \qquad \pi_{23} = P(c) = 1/6$$
$$\pi_{13} = P(c) = 1/6 \qquad \pi_{24} = P(c) = 1/6$$
$$\pi_{14} = P(c) = 1/6 \qquad \pi_{34} = P(c) = 1/6$$

3. Stimatore e stima di un parametro

Matrice dei dati campionari

Si supponga di dover effettuare un'indagine campionaria su una popolazione di N unità utilizzando un campione di numerosità n generato da un determinato disegno campionario.

Si indichi con $\mathbf{c} = \{1, 2, \ldots, n\}$ il generico campione, le cui unità non coincidono necessariamente con le prime N unità della popolazione, con P(c) la sua probabilità di selezione e con p il numero delle variabili oggetto d'indagine.

All'i.ma unità del campione risulta associato il seguente vettore riga (o record):

$$i, y_{(1)i}, y_{(2)i}, \ldots, y_{(p)i}$$

dove sono state utilizzate le lettere minuscole per indicare che i valori delle variabili sono riferiti alle unità del campione.

L'insieme degli n vettori riga costituisce la matrice (o file) dei dati campionari.

Esempio 1.4. Si consideri la popolazione dell'esempio 1.1 ed un disegno campionario che genera campioni di numerosità 2 con uguale probabilità. Per ognuno dei 6 campioni che formano l'universo dei campioni si avrà una matrice di dati formata da due righe (una per ciascuna unità campione) e 6 colonne (una per l'etichetta, due per le variabili quantitative e tre per le variabili indicatrici):

Campione	Unità	valori delle variabili				
1	1	700	2000	1	0	0
	2	740	1800	0	1	0
2	1	700	2000	1	0	0
	3	640	1600	0	1	0
3	1	700	2000	1	0	0
	4	720	1600	0	0	1
4	2	740	1800	0	1	0
	3	640	1600	0	1	0
5	2	740	1800	0	1	0
	4	720	1600	0	0	1
6	3	640	1600	0	1	0
	4	720	1600	0	0	1

Distribuzione campionaria di uno stimatore

Si definisce stimatore $\hat{\theta}$ di un parametro una funzione reale dei valori osservati sulle unità del campione.

Il valore $\hat{\theta}(\mathbf{c})$ che lo stimatore assume in corrispondenza di uno specifico campione prende il nome di stima.

Per un dato disegno campionario si avrà quindi un valore della stima per ognuno dei possibili campioni e a ciascuno di questi valori, non tutti necessariamente distinti, sarà associata una probabilità uguale a quella di selezione del campione cui la stima si riferisce.

L'insieme delle coppie $\{\hat{\theta}(\mathbf{c}), P(\mathbf{c})\}$ costituisce la distribuzione campionaria dello stimatore per il piano di campionamento considerato.

Esempio 1.5. Con riferimento all'esempio 1.4 si supponga di voler stimare la spesa media mensile per l'alimentazione, la spesa media mensile totale e la frequenza relativa delle unità appartenenti a ciascuna ripartizione territoriale. Utilizzando come stimatore di una media la media calcolata sulle unità osservate nel campione e ricordando che la stima di una frequenza si ottiene stimando la media della variabile indicatrice corrispondente, si hanno le seguenti distribuzioni campionarie degli stimatori:

Campione	P(c)	$\hat{\bar{Y}}_{(1)}$	$\hat{\bar{Y}}_{(2)}$	$\hat{\bar{I}}_{(2)}$	$\hat{\bar{I}}_{(2)}$	$\hat{\bar{I}}_{(3)}$
1	1/6	720	1900	0.50	0.50	0.00
2	1/6	670	1800	0.50	0.50	0.00
3	1/6	710	1800	0.50	0.00	0.50
4	1/6	690	1700	0.00	1.00	0.00
5	1/6	730	1700	0.00	0.50	0.50
6	1/6	680	1600	0.00	0.50	0.50

È bene sottolineare che non sempre nello stimare un parametro $\theta = f(Y_1, Y_2, \ldots, Y_N)$ si utilizza come stimatore la stessa funzione f applicata ai dati campionari; quando ciò avviene, lo stimatore è detto naturale.

4. Valore medio e varianza campionaria di uno stimatore

Valore medio di uno stimatore

La media delle stime relative ai singoli campioni ponderata con le rispettive probabilità prende il nome di valore medio dello stimatore rispetto al disegno campionario:

$$E(\hat{\theta}) = \sum_{c \in C} \hat{\theta}(c) P(c) \qquad [1.2]$$

dove con la lettera E si è indicato il simbolo di valore medio.

Uno stimatore è detto centrato o corretto se il suo valore medio è uguale al valore che il parametro assume nella popolazione:

$$E(\hat{\theta}) = \theta \qquad [1.3]$$

altrimenti è detto distorto, e la differenza tra il valore medio e il valore vero è chiamata distorsione:

$$B = E(\hat{\theta}) - \theta \qquad [1.4]$$

È facile verificare che per l'esempio 1.5 gli stimatori considerati sono tutti corretti.

Varianza campionaria ed errore standard

Si definisce varianza di uno stimatore rispetto al disegno campionario la media ponderata dei quadrati delle differenze fra la stima e il loro valore medio:

$$Var(\hat{\theta}) = \sum_{c \in C} \left[\hat{\theta}(c) - E(\hat{\theta})\right]^2 P(c) \qquad [1.5]$$

La varianza dello stimatore viene anche chiamata varianza campionaria e la sua radice quadrata prende il nome di errore standard o errore di campionamento:

$$SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})} \qquad [1.6]$$

Distribuzione campionario degli errori

L'errore che si commette nello stimare un parametro della popolazione mediante i valori osservati in un determinato campione è misurato dalla differenza tra la stima e il valore vero del parametro:

$$\xi(c) = \hat{\theta}(c) - \theta \qquad [1.7]$$

Sia l'intensità che il segno dell'errore possono variare al variare del campione, l'insieme delle coppie $\{\xi(c), P(c)\}$ rappresenta la distribuzione campionaria degli errori.

Errore quadratico medio

Si definisce errore quadratico medio (MSE) la media aritmetica ponderata dei quadrati degli errori delle stime derivate dai singoli campioni:

$$MSE(\hat{\theta}) = \sum_{c \in C} \xi^2(c) P(c) \qquad [1.8]$$

Se i valori rilevati sulle unità del campione non sono affetti da errori di misura, si può facilmente verificare che:

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + B^2 \qquad [1.9]$$

La [1.9] sta ad indicare che in assenza di errori di misura, l'errore quadratico medio è somma di due componenti: la varianza campionaria e il quadrato della distorsione dello stimatore.

Nel caso in cui lo stimatore è corretto l'errore quadratico medio coincide con la varianza campionaria e la sua radice quadrata è uguale all'errore standard.

Poiché nelle indagini campionarie usualmente viene calcolata la sola varianza campionaria, questa fornisce una misura corretta dell'errore quadratico medio solo nel caso in cui lo stimatore è centrato, negli altri casi ne rappresenta soltanto l'estremo inferiore.

Esempio 1.6. Riprendendo l'esempio 1.5, per la stima della spesa media mensile per l'alimentazione si ha la seguente distribuzione campionaria degli errori:

Campione	P(c)	ζ(c)	ζ^2(c)
1	1/6	+20	400
2	1/6	-30	900
3	1/6	+10	100
4	1/6	-10	100
5	1/6	+30	900
6	1/6	-20	400

La varianza campionaria è data:

$$Var(\hat{\bar{Y}}_{(1)}) = (400+900+100+100+900+400) / 6 = 467$$

e l'errore standard:

$$SE(\hat{\bar{Y}}_{(1)}) = \sqrt{467} = 21{,}6$$

Poiché lo stimatore considerato è centrato, la varianza campionaria può essere presa come misura dell'errore quadratico medio delle stime.

Errore relativo

Un altro indicatore che viene, generalmente, calcolato per valutare l'attendibilità di una stima, è costituito dall'errore relativo, Esso si ottiene come rapporto tra l'errore standard e il valore vero del parametro che comunemente viene espresso in forma percentuale:

$$RE(\hat{\theta}) = \frac{SE(\hat{\theta})}{\theta} \cdot 100 \qquad [1.10]$$

Per l'esempio in esame l'errore relativo è dato da:

$$RE = 100 \cdot \frac{246}{700} = 3{,}1\%$$

Occorre osservare che l'errore standard è espresso nella stessa unità di misura del parametro, nel caso in esame in migliaia di lire, mentre l'errore relativo è un numero puro e quindi risulta indipendente dall'unità di misura.

È bene sottolineare che l'errore standard e l'errore relativo non costituiscono una misura dell'errore per uno specifico campione, ma soltanto l'errore medio che ci si aspetta stimando il parametro d'interesse con quel disegno campionario e quello stimatore.

Così l'errore standard calcolato nell'esempio precedente, sta ad indicare che il piano di campionamento adottato (campione di ampiezza 2, estrazione con uguale probabilità senza reimmissione) fornisce una stima per la spesa media mensile per l'alimentazione con un errore medio di 21.600 lire, mentre il valore di RE indica che l'errore è pari al 3,1% del valore che deve essere stimato.

L'errore di campionamento non dipende soltanto dal disegno campionario ma anche dalla natura della variabile cui si riferisce la stima e dalla dimensione del campione.

Esempio 1.7. Nel prospetto che segue sono riportati gli errori di campionamento per le stime dei cinque parametri presi in esame nell'esempio 1.5:

Parametro da stimare	valore del parametro	errore standard	errore relativo
Spesa media per l'alimentazione	700	21,60	3,1%
spesa media totale	1750	95,70	5,5%
freq. rel. 'nord'	0,25	0,25	100,0%
freq. rel. 'centro'	0,50	0,29	58,0%
freq. rel. 'sud'	0,25	0,25	100,0%

Come si osserva, gli errori relativi sono abbastanza contenuti per le stime delle medie mentre risultano elevatissimi per le stime delle frequenze, tanto da poter affermare che un campione di 2 unità (tasso di campionamento del 50%) non è idoneo a fornire stime attendibili della distribuzione territoriale delle famiglie.

Ci si può chiedere come variano gli errori di campionamento al variare dell'ammontare della dimensione del campione.

Esempio 1.8. Utilizzando lo stesso disegno campionario e la stessa procedura di stima, ma con una numerosità campionaria n = 3, sono stati calcolati gli errori standard assoluti e relativi delle stime dei cinque parametri:

Parametro da stimare	valore del parametro	errore standard	errore relativo
Spesa media per l'alimentazione	700	12,50	1,8%
spesa media totale	1750	55,30	1,4%
freq. rel. 'nord'	0,25	0,14	57,7%
freq. rel. 'centro'	0,50	0,17	33,4%
freq. rel. 'sud'	0,25	0,14	57,7%

Si può verificare la sensibile riduzione degli errori di campionamento che si ha aumentando l'ampiezza campionaria da 2 a 3 unità. La riduzione relativa, misurata come complemento all'unità del rapporto tra gli errori standard corrispondenti, risulta costante per tutte le caratteristiche considerate *e uguale* a circa 0,42.

Come si vedrà in seguito, in popolazioni di piccole dimensioni la riduzione dipende sia dalla numerosità del campione che da quella della popolazione, mentre per popolazioni di ampiezza elevata, come quelle che vengono generalmente prese in considerazione nelle indagini Istat, la riduzione dipende soltanto dalla dimensione del campione.

5. Intervalli di confidenza

Nella pratica del campionamento probabilistico accanto alle stime puntuali finora considerate, rivestono una particolare im-

portanza le stime per intervallo, comunemente denominate intervalli di confidenza.

Gli intervalli di confidenza vengono calcolati sulla base della distribuzione campionaria dello stimatore. In molte situazioni concrete questa distribuzione è approssimativamente normale, e se lo stimatore è centrato la media e lo scostamento quadratico medio sono uguali rispettivamente al valore del parametro nella popolazione e all'errore standard.

Se sono valide queste assunzioni, è possibile calcolare in corrispondenza di ciascun campione un intervallo in modo che il valore vero del parametro sia interno all'intervallo in una frazione P dei campioni ed esterno nella frazione complementare.

Per determinare gli estremi dell'intervallo si procede nel seguente modo:

1) si fissa il livello di confidenza desiderato, ossia la probabilità P che, estraendo un campione in base al disegno campionario adottato, l'intervallo contenga il valore vero del parametro;

2) utilizzando i valori tabulati della distribuzione normale standardizzata si determina il valore t (semi-ampiezza dell'intervallo) in modo che l'area della normale compresa tra – t e + t sia uguale a P.

3) si calcolano gli estremi inferiori e superiore dell'intervallo

$$estremo \quad inferiore = \hat{\theta}(c) - t\,SE(\theta)$$

$$estremo \quad superiore = \hat{\theta}(c) + t\,SE(\theta) \qquad [1.11]$$

L'ampiezza dell'intervallo dipende, oltre che dall'errore standard, anche dal livello di confidenza fissato, nel senso che più è elevato il valore di P più ampio è l'intervallo.

Di seguito si riportano i valori di t in corrispondenza ad alcuni valori di P più frequentemente utilizzati nella pratica:

P (%)	t
68	1,00
90	1,64
95	1,96
99	2,58

Gli intervalli di confidenza vengono calcolati sotto l'assunzione della normalità della distribuzione campionaria del parametro d'interesse. Il grado di approssimazione dipende dalla forma della distribuzione della variabile nella popolazione, dal disegno campionario, dal tipo di stimatore e dalla dimensione del campione. L'approssimazione migliora con l'aumentare dell'ampiezza campionaria e nelle indagini su larga scala, l'assunzione di normalità può essere considerata praticamente soddisfatta.

Esempio 1.9. Con riferimento ai dati dell'esempio 1.4 sono stati calcolati gli intervalli di confidenza relativi alla spesa media mensile per l'alimentazione, prendendo un valore di t=1 corrispondente ad un livello di confidenza P=68% nel caso di distribuzione normale.

Ricordando che SE = 21,6 gli estremi dell'intervallo di confidenza per il generico campione sono dati da:

estremo inferiore $= \hat{\theta}$ (c) – 21,6

estremo superiore $= \hat{\theta}$ (c) + 21,6

Di seguito sono riportate per ciascun campione la stima degli estremi dell'intervallo di confidenza, e sono stati contrassegnati con (*) gli intervalli esatti, ossia quelli che contengono il valore vero

Campione	P (c)	stima	estremo inferiore	estremo superiore	intervalli esatti
1	1/6	720	698,4	741,6	(*)
2	1/6	670	648,4	691,6	
3	1/6	710	688,4	731,6	(*)
4	1/6	690	668,4	711,6	(*)
5	1/6	730	708,4	751,6	
6	1/6	680	658,4	701,6	(*)

Nell'esempio riportato la probabilità di estrarre un campione a cui è associato un intervallo di confidenza esatto è data da 4/6 pari al 66,7% ed è circa uguale a quella attesa (68%).

CAPITOLO 2 - METODOLOGIA STANDARD PER LA STIMA DELLA VARIANZA CAMPIONARIA

1. Premessa

Una volta effettuata l'indagine campionaria, si dispone soltanto dei risultati relativi al campione selezionato, che costituisce uno dei possibili campioni che possono essere generati dal disegno campionario adottato. Tuttavia, in un campionamento probabilistico è possibile stimare la varianza campionaria attraverso i dati rilevati nel campione. Ciò è dovuto al fatto che la variabilità che si riscontra tra le unità del campione può essere utilizzata per ricavare una stima della variabilità campionaria.

La teoria del campionamento da popolazioni finite fornisce, per i principali disegni campionari, le espressioni per il calcolo della varianza degli stimatori comunemente utilizzati nella pratica. Nel caso di stimatori di un totale e di una media, o di loro combinazioni lineari, vengono date le espressioni esatte, mentre per altri stimatori come il rapporto o la differenza di rapporti si dispone di espressioni approssimate ma ugualmente valide nel caso di campioni di numerosità elevata.

Le formule per la stima della varianza campionaria, e di conseguenza quella delle altre statistiche che da questa possono essere derivate (errore standard, errore relativo ed intervalli di confidenza) dipendono, oltre che dal tipo di stimatore, anche dal disegno campionario utilizzato. Pertanto per poter procedere concretamente al calcolo degli errori di campionamento per una qualsiasi indagine condotta dall'Istituto, è necessario predisporre, per ciascun disegno campionario, un'opportuna procedura informatica.

Per superare le difficoltà derivanti dalla diversa tipologia degli stimatori e quindi dalla molteplicità delle formule che devono essere impiegate, possono essere introdotte delle opportune trasformazioni delle variabili così da ricondurre i vari stimatori ad una stessa forma funzionale. In tal modo è possibile utilizzare, per uno specifico piano di campionamento, un'unica espressione per la stima della varianza campionaria.

Questa metodologia, che viene definita standard in quanto utilizza le formule specifiche (standard) di ciascun disegno campionario, verrà illustrata, nel presente capitolo, con riferimento ad un generico piano di campionamento e agli stimatori del totale, della media, della differenza tra due medie, del rapporto e della differenza tra due rapporti. Nei prossimi tre capitoli verranno, invece, descritte le modalità che devono essere seguite per la sua applicazione ai più comuni disegni campionari.

2. Varianza campionaria dello stimatore di un totale

Stimatori lineari

Come è stato detto, lo stimatore di un parametro della popolazione è una funzione dei valori osservati nel campione e in quanto tale può assumere le forme più diverse.

Una classe di stimatori che riveste una particolare importanza nel campionamento da popolazioni finite è quella degli stimatori lineari, ossia degli stimatori che possono essere espressi come funzioni lineari dei valori osservati nel campione. Più precisamente, secondo Cassel, Särndal e Wretman (1977, pp. 21-22) uno stimatore è lineare se è del tipo:

$$\hat{\theta}_{(r)} = a_{co} + \sum_{i=1}^{n} a_{ci}\, y_{(r)i} \qquad [2.1]$$

mentre è lineare omogeneo se assume la forma:

$$\hat{\theta}_{(r)} = \sum_{i=1}^{n} a_{ci}\, y_{(r)i} \qquad [2.2]$$

dove la sommatoria è estesa a tutte le unità del campione, le a_{ci} sono dei coefficienti definiti precedentemente all'indagine per tutti i campioni $c \in C$ e per tutte le unità campionate $i \in c$, e $y_{(r)i}$ è il valore della variabile r.ma osservato nell'i.ma unità del campione.

Stimatori di un totale

Uno stimatore lineare che gioca un ruolo fondamentale nel campionamento da popolazioni finite è quello dell'ammontare totale di un carattere, che è dato da:

$$\hat{Y}_{(r)} = \sum_{i=1}^{n} w_i\, y_{(r)i} \qquad [2.3]$$

dove w_i è il coefficiente di espansione associato all'unità i.ma, e il cui valore dipende dal disegno campionario e dal metodo di stima utilizzati.

Di seguito verranno descritti i due metodi di stima di un totale che trovano il più largo impiego nella pratica: il primo dovuto ad Hansen e Hurwitz (1943) è applicabile ai disegni campionari in cui le unità vengono estratte con reimmissione, il secondo introdotto da Horvitz e Thompson (1952) trova, invece, impiego nei casi in cui le unità sono estratte senza reimmissione.

Lo stimatore di Hansen e Hurwitz

Da una popolazione di N unità si estrae un campione di numerosità n con reimmissione (o mediante n prove indipendenti). Indicando con p_i la probabilità di selezione dell'i.ma unità, lo stimatore del totale proposto da Hansen e Hurwitz è dato da:

$$\hat{Y}_{HH} = \sum_{i=1}^{n} \frac{y_i}{p_i} = \sum_{i=1}^{n} w_i\, y_i \qquad [2.4]$$

dove $w_i = 1/p_i$ e y_i è il valore della generica variabile osservato nell'unità i.ma e si è omesso l'indice r di variabile per non appesantire troppo la simbologia.

Si dimostra (cfr. ad es. Cochran 1977, pp. 252-254) che lo stimatore è non distorto ed ha varianza campionaria:

$$Var(\hat{Y}_{HH}) = \frac{1}{n} \sum_{i=1}^{N} \sum_{j>i} \left(\frac{Y_i}{p_i} - \frac{Y_j}{p_j} \right)^2 p_i p_j \qquad [2.5]$$

Si dimostra, inoltre, che uno stimatore non distorto della varianza campionaria è dato da:

$$\hat{Var}(\hat{Y}_{HH}) = \frac{1}{n(n-1)} \sum_{i=i}^{n} \left(\frac{y_i}{p_i} - \hat{Y}_{HH} \right)^2 \qquad [2.6]$$

che può anche essere scritto:

$$\hat{Var}(\hat{Y}_{HH}) = \frac{1}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j>i} \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2 \qquad [2.7]$$

Come si verifica facilmente lo stimatore di Hansen e Hurwitz è lineare omogeneo e la sua varianza campionaria si annulla quando per tutte le unità della popolazione le probabilità di selezione p_i sono proporzionali ai corrispondenti valori Y_i.

Lo stimatore di Horvitz e Thompson

Da una popolazione di N unità si estrae senza reimmissione un campione di numerosità n. Indicando con π_i la probabilità di inclusione del primo ordine, lo stimatore del totale proposto da Horvitz e Thompson è dato da:

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} w_i y_i \qquad [2.8]$$

dove $w_i = 1/\pi_i$.

Sen (1953) e Yates-Grunding (1953) hanno ricavato, indipendentemente l'uno dagli altri, la seguente espressione della varianza campionaria:

$$Var(\hat{Y}_{HT}) = \sum_{i=1}^{N} \sum_{j>i} \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \qquad [2.9]$$

Gli stessi autori hanno altresì proposto di utilizzare come stimatore della varianza campionaria:

$$\hat{Var}(\hat{Y}_{HT}) = \sum_{i=1}^{n} \sum_{j>i} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \qquad [2.10]$$

dove con π_{ij} sono state indicate le probabilità d'inclusione del secondo ordine.

Questo stimatore gode delle seguenti proprietà:

- è lineare omogeneo;
- è il solo stimatore non distorto nella classe degli stimatori in cui il coefficiente di espansione associato a ciascuna unità della popolazione è fissato prima della selezione del campione;
- non esiste uno stimatore nella classe degli stimatori lineari omogenei non distorti che ha una varianza più piccola di $\hat{Y}_{HT}$;
- se tutti i valori Y_i (i = 1, . . ., N) sono esattamente proporzionali ai corrispondenti valori π_i la varianza di $\hat{Y}_{HT}$ è uguale a zero.

Le formule [2.7] e [2.10] che danno la stima della varianza campionaria dello stimatore di un totale, rispettivamente nel caso di selezione con e senza reimmissione, possono essere facilmente estese ai disegni campionari più complessi che prevedono la stratificazione e più stadi di campionamento.

Lo stimatore con il metodo del rapporto

Accanto agli stimatori di Hansen-Hurwitz e Horvitz-Thompson, definiti stimatori «diretti» in quanto utilizzano le sole informazioni desunte dal campione, vengono spesso utilizzati degli stimatori «indiretti» che tengono conto anche di informazioni supplementari. Il ricorso a questo tipo di stimatori, che risultano molto più complessi di quelli diretti, trova giustificazione nella loro maggiore efficienza.

Tra gli stimatori indiretti, quello che trova più frequente utilizzazione nella pratica, è lo stimatore con il metodo del rapporto. Questo stimatore si applica quando è noto l'ammontare totale nella popolazione di una o più variabili ausiliarie correlate con le variabili di rilevazione.

Limitando la trattazione al caso di una sola variabile ausiliaria, si indichi con x_i il valore che essa assume nell'i.ma unità del campione e con X il suo ammontare totale nella popolazione. Lo stimatore del totale con il metodo del rapporto è dato da:

$$\hat{Y}_R = \frac{\hat{Y}}{\hat{X}} X = \hat{R} X \qquad [2.11]$$

dove $\hat{Y}$ ed $\hat{X}$ sono le stime dei totali ottenute applicando uno dei due metodi diretti descritti in precedenza.

Lo stimatore $\hat{Y}_R$ è non lineare, infatti esso è dato dal prodotto tra il rapporto $\hat{R} = \hat{Y} / \hat{X}$ e la costante X. La procedura che deve essere seguita per la stima della sua varianza campionaria, e più in generale della varianza campionaria di stimatori non lineari, verrà descritta nel paragrafo 4.

3. Varianza campionaria di una combinazione lineare di totali

Stimatori combinazioni lineari di totali

L'importanza dello stimatore di un totale deriva dal fatto che gli stimatori di numerosi parametri, e tra questi tutti quelli considerati nel presente fascicolo, possono essere espressi come combinazioni lineari di totali:

$$\hat{\theta} = b_1 \hat{Y}_{(1)} + b_2 \hat{Y}_2 + \cdots + b_m \hat{Y}_{(m)} \qquad [2.12]$$

Varianza campionaria

e la loro varianza campionaria può essere calcolata in funzione delle stime delle varianze e covarianze campionarie degli stimatori dei totali:

$$\hat{Var}(\hat{\theta}) = \sum_{r=1}^{m} b_r^2 \, \hat{Var}\left(\hat{Y}_{(r)}\right) + \sum_{r=1}^{m} \sum_{s \neq r} b_r \, b_s \, \hat{Cov}\left(\hat{Y}_{(r)}, \hat{Y}_{(s)}\right) \qquad [2.13]$$

Sostituendo nella [2.13] a $\mathrm{Var}(\hat{Y}_{(r)})$ e $\mathrm{Cov}(\hat{Y}_{(r)}, \hat{Y}_{(s)})$ le corrispondenti stime derivate dai dati osservati nel campione, si ottiene la stima della varianza campionaria:

$$\hat{\hat{Var}}(\hat{\theta}) = \sum_{r=1}^{m} b_r^2 \, \hat{\hat{Var}}\left(\hat{Y}_{(r)}\right) + \sum_{r=1}^{m} \sum_{s \neq r} b_r \, b_s \, \hat{\hat{Cov}}\left(\hat{Y}_{(r)}, \hat{Y}_{(s)}\right) \qquad [2.14]$$

Alcuni esempi

Esempi piuttosto semplici di stimatori combinazioni lineari di totali sono dati dalla media, dalla somma e dalla differenza di due medie, per i quali si ha:

Stimatore di una media $\overline{Y}_{(k)}$

coefficienti della combinazione lineare

$$b_r = \begin{cases} 1/N & \text{se } r = k \\ 0 & \text{altrimenti} \end{cases}$$

stimatore

$$\hat{\overline{Y}}_{(k)} = \frac{1}{N} \hat{Y}_{(k)} \qquad [2.15]$$

varianza campionaria

$$Var\left(\hat{\overline{Y}}_{(k)}\right) = \frac{1}{N^2} Var\left(\hat{Y}_{(k)}\right) \qquad [2.16]$$

stimatore della varianza campionaria

$$\hat{Var}\left(\hat{\overline{Y}}_{(k)}\right) = \frac{1}{N^2} \hat{Var}\left(\hat{Y}_{(k)}\right) \qquad [2.17]$$

Stimatore della somma o della differenza di due medie $\overline{Y}_{(k)} \pm \overline{Y}_{(h)}$

coefficienti della combinazione lineare

$$b_r = \begin{cases} 1/N & \text{se } r = k \\ \pm 1/N & \text{se } r = h \\ 0 & \text{altrimenti} \end{cases}$$

stimatore

$$\hat{\overline{Y}}_{(k)} \pm \hat{\overline{Y}}_{(h)} = \frac{1}{N} \hat{Y}_{(k)} \pm \frac{1}{N} \hat{Y}_{(h)} \qquad [2.18]$$

varianza campionaria

$$Var\left(\hat{\overline{Y}}_{(k)} \pm \hat{\overline{Y}}_{(h)}\right) = \frac{Var\left(\hat{Y}_{(k)}\right) + Var\left(\hat{Y}_{(h)}\right) \pm 2Cov\left(\hat{Y}_{(k)}, \hat{Y}_{(h)}\right)}{N^2} \qquad [2.19]$$

stimatore della varianza campionaria

$$\hat{Var}\left(\hat{\overline{Y}}_{(k)} \pm \hat{\overline{Y}}_{(h)}\right) = \frac{\hat{Var}\left(\hat{Y}_{(k)}\right) + \hat{Var}\left(\hat{Y}_{(h)}\right) \pm 2\hat{Cov}\left(\hat{Y}_{(k)}, \hat{Y}_{(h)}\right)}{N^2} \qquad [2.20]$$

4. Varianza campionaria degli stimatori non lineari

Linearizzazione degli stimatori non lineari

Non tutti gli stimatori che vengono utilizzati nella pratica possono essere espressi come combinazioni di stimatori lineari, si pensi infatti allo stimatore di un rapporto o a quello di un totale mediante il metodo del rapporto.

Per gli stimatori non lineari non si dispone di espressioni esatte per il calcolo della varianza campionaria per cui è necessario ricorrere a procedimenti approssimati. Nel caso di stimatori che sono funzioni regolari di totali, il metodo che viene comunemente utilizzato è quello basato sulla linearizzazione dello stimatore mediante il suo sviluppo in serie di Taylor arrestato al primo termine.

È bene sottolineare che il metodo dello sviluppo in serie di Taylor non è una tecnica per la stima della varianza campionaria, ma solo il procedimento per approssimare una funzione non linearè mediante una combinazione di termini lineari, alla quale vanno poi applicate le formule per il calcolo della varianza proprie del piano di campionamento in esame.

Indicando con $\theta = f(Y_{(1)}, Y_{(2)}, \ldots, Y_{(m)})$ una funzione non lineare dei totali di m caratteri e con $\hat{\theta}$ il suo stimatore naturale, si dimostra che sotto condizioni non molto restrittive l'approssimazione di $\hat{\theta} - \theta$ arrestata ai termini lineari è data da:

$$\hat{\theta} - \theta \doteq \sum_{r=1}^{m} b_r \left(\hat{Y}_{(r)} - Y_{(r)}\right) \qquad [2.21]$$

dove $b_r = \dfrac{\delta f}{\delta Y_{(r)}}$ è la derivata prima della funzione rispetto al paramento $Y_{(r)}$.

Stima di MSE

Poiché $\hat{\theta}$ non è uno stimatore corretto, applicando la (2.14) si ottiene una stima approssimata dell'errore quadratico medio:

$$\hat{MSE}\left(\hat{\theta}\right) \doteq \sum_{r=1}^{m} \hat{b}_r^2 \, \hat{Var}\left(\hat{Y}_{(r)}\right) + \sum_{r=1}^{m} \sum_{s \neq r} \hat{b}_r \, \hat{b}_s \, \hat{Cov}\left(\hat{Y}_{(r)}, \hat{Y}_{(s)}\right) \qquad [2.22]$$

dove $\hat{b}_r$ sono i valori delle derivate prime nelle quali i totali $Y_{(r)}$ sono stati sostituiti con le rispettive stime $\hat{Y}_{(r)}$.

Alcuni esempi

Così per gli stimatori di un rapporto e di un totale con il metodo del rapporto, si ha:

Stimatore di un rapporto

$$\hat{R} = \frac{\hat{Y}_{(1)}}{\hat{Y}_{(2)}} \qquad [2.23]$$

approssimazione lineare:

$$\hat{R} - R \doteq \frac{1}{Y_{(2)}} \left(\hat{Y}_{(1)} - Y_{(1)}\right) - \frac{Y_{(1)}}{Y_{(2)}^2} \left(\hat{Y}_{(2)} - Y_{(2)}\right) \qquad [2.24]$$

stimatore dell'errore quadratico medio

$$\hat{MSE}(\hat{R}) \doteq \hat{R}^2 \left[\frac{\hat{Var}\left(\hat{Y}_{(1)}\right)}{\hat{Y}_{(1)}^2} + \frac{\hat{Var}\left(\hat{Y}_{(2)}\right)}{\hat{Y}_{(2)}^2} - 2 \frac{\hat{Cov}\left(\hat{Y}_{(1)}, \hat{Y}_{(2)}\right)}{\hat{Y}_{(1)} \, \hat{Y}_{(2)}} \right] \qquad [2.25]$$

Stimatore di un totale con il metodo del rapporto

$$\hat{Y}_R = \frac{\hat{Y}}{\hat{X}} X = \hat{R} X \qquad [2.26]$$

approssimazione lineare:

$$\hat{Y}_R - Y \doteq \left(\hat{Y} - Y\right) - \hat{R}\left(\hat{X} - X\right) \qquad [2.27]$$

stimatore dell'errore quadratico medio

$$\hat{MSE}\left(\hat{Y}_R\right) = \hat{Var}\left(\hat{Y}\right) + \hat{R}^2 \hat{Var}\left(\hat{X}\right) - 2\hat{R}\hat{Cov}\left(\hat{Y}, \hat{X}\right) \qquad [2.28]$$

5. Il metodo di Woodruff

Da *quanto* fin qui esposto si evince che tutti gli stimatori che vengono usualmente presi in considerazione nelle indagini campionarie condotte dall'Istat o sono stimatori di totali o possono essere espressi come combinazioni lineari di stimatori di totali.

Risulta, altresì, evidente che una procedura informatica che utilizzi la metodologia fin qui descritta, deve prevedere la stima della matrice delle varianze e covarianze campionarie degli stimatori che intervengono nella combinazione lineare.

Trasformata di Woodruff

Per semplificare ulteriormente la procedura è conveniente utilizzare il metodo proposto da Woodruff (1971), mediante il quale è possibile ricondurre il calcolo della varianza campionaria di un qualsiasi stimatore combinazione lineare di totali a quello della varianza campionaria di un totale.

Sia $\hat{\theta} = b_1 \hat{Y}_{(1)} + b_2 \hat{Y}_{(2)} + \ldots + b_m \hat{Y}_{(m)}$ lo stimatore in esame, la stima della sua varianza campionaria può anche essere scritta:

$$\hat{Var}\left(\hat{\theta}\right) = \hat{Var}\left(\sum_{r=1}^{m} b_r \hat{Y}_{(r)}\right) \qquad [2.29]$$

Ricordando che:

$$Y_{(r)} = \sum_{i \in c}^{+} w_i y_{(r)i} \qquad [2.30]$$

si ha:

$$\hat{Var}\left(\hat{\theta}\right) = \hat{Var}\left(\sum_{r=1}^{m} b_r \sum_{i \in c} w_i y_{(r)i}\right) =$$

$$= \hat{Var}\left(\sum_{i \in c} w_i \sum_{r=1}^{m} b_{(r)} y_{(r)i}\right) = \qquad [2.31]$$

$$= \hat{Var}\left(\sum_{i \in c} w_i z_i\right) = \hat{Var}\left(\hat{Z}\right)$$

dove:

$$z_i = b_1 y_{(1)i} + b_2 y_{(2)i} + \cdots + b_m y_{(m)i} \qquad [2.32]$$

In pratica si tratta di determinare per ciascuna unità del campione i valori di una nuova variabile combinazione lineare delle m variabili e calcolare quindi la varianza campionaria dello stimatore del totale della variabile così ottenuta.

È ovvio che questo metodo consente di semplificare notevolmente la procedura di calcolo riconducendo un problema di stima multidimensionale a un problema unidimensionale.

Alcuni esempi

Di seguito si riportano le trasformazioni lineari che devono essere utilizzate per calcolare la varianza campionaria con il metodo di Woodruff degli stimatori comunemente considerati nelle indagini Istat.

Stimatore di una media

$$z_i = \frac{1}{N} y_i \qquad [2.33]$$

Stimatore della somma o della differenze di due medie

$$z_i = \frac{1}{N} y_{(1)i} \pm \frac{1}{N} y_{(2)i} \qquad [2.34]$$

Stimatore di un rapporto

$$z_i = \frac{1}{\hat{Y}_{(2)}} \left(y_{(1)i} - \hat{R} y_{(2)i}\right) \qquad [2.35]$$

Stimatore di un totale con il metodo del rapporto

$$z_i = y_i - \hat{R} x_i \quad con \quad \hat{R} = \frac{\hat{Y}}{\hat{X}} \qquad [2.36]$$

Stimatore della differenza tra due rapporti

$$z_i = \frac{1}{\hat{Y}_{(2)}} \left(y_{(1)i} - \hat{R}_1 y_{(2)i}\right) - \frac{1}{\hat{Y}_{(4)}} \left(y_{(3)i} - \hat{R}_2 y_{(4)i}\right) \qquad [2.37]$$

6. Varianza campionaria degli stimatori relativi a sottoclassi

Variabili indicatrici delle sottoclassi

La stima del totale di un carattere e la relativa varianza campionaria possono essere calcolati sia con riferimento all'intera popolazione che a delle sottoclassi.

Le espressioni che sono state finora riportate consentono il calcolo della varianza campionaria per gli stimatori relativi alla popolazione totale, e la loro utilizzazione nel caso di stimatori per sottoclassi richiede il ricorso ad alcuni adattamenti.

L'appartenenza di un'unità ad una determinata sottoclasse può essere individuata mediante l'introduzione di un'opportuna variabile indicatrice, che assume il valore 1 se l'unità appartiene alla sottoclasse e il valore 0 nel caso contrario. Indicando con t la sottoclasse e con $I_{(t)}$ la variabile indicatrice si ha:

$$I_{(t)i} = \begin{cases} 1 & \text{se } i \in t \\ 0 & \text{altrimenti} \end{cases}$$

Il numero totale delle unità della popolazione che appartengono alla sottoclasse è allora dato da:

$$N_t = \sum_{i=1}^{N} I_{(t)i} \qquad [2.38]$$

e quello del campione

$$n_t = \sum_{i=1}^{n} I_{(t)i} \qquad [2.39]$$

Il procedimento per il calcolo della varianza è diverso a seconda che N_t sia conosciuto o meno, e poiché entrambi i casi sono molto frequenti nella pratica verrà effettuata una specifica trattazione per ciascuno di essi.

N_t è conosciuto

In pratica è come se si fosse effettuato un campionamento nella sottoclasse t, estraendo dalle N_t unità della popolazione n_t unità campione utilizzando lo stesso disegno campionario impiegato per selezionare il campione totale.

Si possono, quindi, applicare le stesse formule che vengono adottate per gli stimatori riferiti all'intera popolazione, con l'avvertenza di prendere in considerazione le sole unità appartenenti alla sottoclasse (quelle unità, cioè, per le quali la variabile indicatrice assume il valore 1).

N_t non è conosciuto

Nel trattare questo caso verrà seguito il procedimento riportato dal Cochran (1977, pp. 35-37), che risulta il più idoneo per una successiva trattazione informatica.

Per ciascuna variabile oggetto d'indagine viene definita una nuova variabile che per le unità appartenenti alla sottoclasse assume lo stesso valore della variabile originaria, mentre per le altre unità assume il valore 0:

$$y'_{(r)i} = \begin{cases} y_{(r)i} & \text{se } i \in t \\ 0 & \text{altrimenti} \end{cases}$$

Nel caso in cui la natura dello stimatore comporta il ricorso ad una trasformazione lineare, al posto delle variabili originarie verranno utilizzate le corrispondenti variabili con apice:

$$z_i = b_1\, y'_{(1)i} + b_2\, y'_{(2)i} + \cdots + b_m\, y'_{(m)i} \qquad [2.40]$$

Le stime dei totali e delle relative varianze campionarie per la sottoclasse si ottengono applicando alle variabili con apice le formule valide per l'intera popolazione.

Nei prossimi tre capitoli verrà decritta l'applicazione della metodologia standard ai seguenti piani di campionamento:

- casuale semplice
- ad uno stadio stratificato

- a due stadi con stratificazione delle unità di primo stadio (PSU), con le seguenti modalità di selezione della PSU in ciascuno stadio:
  - senza reimmissione e con uguale probabilità d'inclusione
  - con reimmissione e probabilità di selezione proporzionali all'ampiezza
  - senza reimmissione e con probabilità di inclusione proporzionali all'ampiezza

Particolare rilievo verrà dato all'organizzazione del file dei dati di base e alla soluzione dei problemi computazionali. Al fine di illustrare con esempi numerici la metodologia riportata, per ciascuno dei disegni campionari presi in considerazione è stato predisposto un apposito programma SAS per il calcolo degli errori di campionamento.

CAPITOLO 3 - APPLICAZIONE DELLA METODOLOGIA STANDARD AL CAMPIONAMENTO CASUALE SEMPLICE

1. Premessa

Il campionamento casuale semplice riveste una particolare importanza non tanto per le sue applicazioni pratiche, quanto perché costituisce la base dei disegni campionari più complessi usualmente adottati nelle indagini su larga scala.

La selezione del campione viene effettuata mediante n estrazioni, in ognuna delle quali le unità della popolazione hanno la stessa probabilità di essere scelte.

Se ad ogni prova possono essere estratte solo le unità della popolazione non selezionate in precedenza, si ha un campionamento casuale semplice senza reimmissione (o senza ripetizione). Alla prima estrazione ciascuna delle N unità della popolazione avrà una probabilità di selezione uguale a 1/N, alla seconda estrazione ognuna delle rimanenti N-1 unità avrà una probabilità uguale a 1/(N-1) e così via.

Se, invece, ad ogni prova possono essere estratte anche le unità selezionate nelle prove precedenti si ha un campionamento casuale semplice con reimmissione (o con ripetizione). Le N unità della popolazione avranno sempre la stessa probabilità di selezione, uguale a 1/N, in ciascuna delle n estrazioni.

In pratica, per entrambi i disegni campionari, le unità che formano il campione vengono selezionate generando una serie di numeri casuali compresi tra 1 ed N, mediante una tavola di numeri aleatori od utilizzando un'apposita procedura informatica. Il campione senza reimmissione sarà costituito dalle unità le cui etichette corrispondono ai primi n numeri diversi che sono stati estratti, mentre quello con reimmissione sarà formato dalle unità con etichette corrispondenti ai primi n numeri anche se alcuni sono ripetuti.

Le formule della varianza campionaria e del relativo stimatore sono in genere molto più semplici nel caso di un campionamento con reimmissione. Però questo disegno campionario risulta meno efficiente di quello senza reimmissione ed, inoltre, crea delle complicazioni nella raccolta e nel trattamento delle informazioni quando la stessa unità della popolazione è compresa più volte nel campione. Per questo motivo i disegni campionari adottati nella pratica sono generalmente senza reimmissione, mentre, spesso, nello stimare la varianza campionaria si uti-

lizzano delle formule approssimate derivate dal campionamento con ripetizione.

Comunque, indipendentemente dalle modalità seguite per la selezione del campione, una volta effettuata la rilevazione e svolte tutte le successive operazioni di trattamento dei dati raccolti (codifica, revisione, registrazione e correzione degli errori), si disporrà di un file dei dati di base costituito da tanti records quante sono le unità rilevate. Ciascun record contiene, oltre al codice identificativo delle unità campione, i valori delle variabili di rilevazione.

Indicando con p il numero delle variabili, all'unità i.ma corrisponderà il seguente vettore di valori:

$$i,\ y_{(1)1},\ \ldots\ y_{(n)i},\ \ldots,\ y_{(p)i} \qquad (i = 1,2,\ \ldots,\ n)$$

Prima di procedere oltre è utile osservare che non sempre la numerosità del campione rilevato coincide con quella del campione programmato, può accadere infatti che per un certo numero di unità non sia possibile effettuare la rilevazione. In questo caso n sta ad indicare il numero delle unità effettivamente osservate.

Come è stato più volte detto, il piano di elaborazione dei risultati di un'indagine prevede la programmazione di numerose tavole, più o meno complesse, le cui caselle corrispondono alle modalità di un carattere qualitativo (o alle classi di un carattere quantitativo) o all'incrocio di modalità di due o più caratteri. All'interno di ciascuna casella sono riportate le stime di frequenza assolute o relative, di totali o medie di variabili quantitative, o rapporti. Le stime all'interno delle caselle sono riferite a sottoclassi della popolazione, mentre quelle relative al totale delle caselle fanno riferimento all'intera popolazione.

Nell'illustrare la procedura che deve essere seguita per la stima della varianza campionaria i due casi verranno trattati separatamente, prendendo in esame dapprima il caso delle stime relative alla popolazione e successivamente quello delle stime per sottoclassi. Per non appesantire troppo il testo, le formule per la stima della varianza campionaria verranno, in genere, date senza dimostrazione, rimandando per una trattazione più completa al testo del Cochran (1977) e al fascicolo 4 del Manuale di Tecniche di indagine.

Nei prossimi paragrafi, dopo aver riportato le espressioni per la stima della varianza campionaria per i due schemi di campionamento, verrà descritta, mediante un'applicazione su dati fittizi, la procedura informatica che è stata predisposta per il calcolo degli errori di campionamento.

2. Il campionamento senza reimmissione

Si può facilmente verificare (cfr. Hajek, 1981 pag. 53) che l'universo dei campioni è costituito da $\binom{N}{n}$ campioni distinti ciascuno dei quali ha la stessa probabilità $1/\binom{N}{n}$ di essere estratto e che le probabilità di inclusione del primo e del secondo ordine sono date rispettivamente da:

$$\pi_i = \frac{n}{N} \qquad (i = 1.2....,N) \qquad [3.1]$$

$$\pi_{ij} = \frac{n}{N}\,\frac{n-1}{N-1} \qquad (i = 1,2,...,N;\ j \neq i) \qquad [3.2]$$

Il rapporto f = n/N prende il nome di tasso di campionamento o frazione sondata e il suo reciproco w = N/n costituisce il coefficiente di espansione che viene assegnato a ciascuna unità campionata.

Di seguito vengono fornite le espressioni per il calcolo della varianza campionaria della stima di un totale per l'intera popolazione e per sottoclassi. La varianza campionaria di una stima esprimibile come combinazione lineare di stime di totali potrà essere determinata utilizzando il metodo di Woodruff riportato nel secondo capitolo.

Stime riferite all'intera popolazione

Trattandosi di un campionamento senza reimmissione per la stima dell'ammontare totale di un carattere si utilizza lo stimatore di Horvitz-Thompson, pertanto applicando la [2.7] si ha:

$$\hat{Y}_{(r)} = \sum_{i=1}^{n} w_i\ y_{(r)i} \qquad [3.3]$$

dove: $w_i = \frac{N}{n}$

La stima della varianza campionaria si ottiene dalla [2.10] sostituendo alle probabilità d'inclusione del primo e del secondo i corrispondenti valori dati dalle [3.1] e [3.2].

$$\hat{Var}(\hat{Y}_{(r)}) = \frac{N-n}{n^2\ (N-1)} \sum_{i=1}^{n} \sum_{j>i} (y_i - y_j)^2 = \frac{1-f}{n}\, N^2\ \hat{S}^2_{(r)} \qquad [3.4]$$

dove $\hat{S}^2_{(r)}$ è la stima della varianza della variabile r.ma nella popolazione:

$$\hat{S}^2_{(r)} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad [3.5]$$

e $\bar{y}$ è la media aritmetica semplice dei valori osservati sulle n unità del campione.

Le stime dell'errore standard e dell'errore relativo sono date da:

$$\hat{SE}(\hat{Y}_{(r)}) = \sqrt{\frac{1-f}{n}}\ N\ \hat{S}_{(r)} \qquad [3.6]$$

e

$$\hat{RE}(\hat{Y}_{(r)}) = \sqrt{\frac{1-f}{n}}\ \frac{N\ \hat{S}_{(r)}}{\hat{Y}_{(r)}} = \sqrt{\frac{1-f}{n}}\ \hat{C}_{(r)} \qquad [3.7]$$

dove $\hat{C}_{(r)}$ è la stima del coefficiente di variazione.

Le formule riportate mettono in evidenza che, per una determinata popolazione, le diverse stime ricavate da un campione di una prefissata numerosità non hanno tutte la stessa precisione, infatti gli errori campionari risultano tanto più elevati quanto più grande è la variabilità nella popolazione dei caratteri cui le stime si riferiscono.

Da questo deriva che la precisione di un campione non è un concetto assoluto esprimibile mediante un unico indicatore, ma che va esplicitato con riferimento ad ogni singola stima oggetto d'indagine.

Stime riferite a sottoclassi

Sia A una partizione delle unità della popolazione in T sottoclassi disgiunte: $A_1, A_2, \ldots, A_T$, si indichi con N_t e n_t il numero delle unità della popolazione e del campione che cadono nella sottoclasse A_t ($t = 1,2,\ldots,T$). Si vuole stimare l'ammontare totale di una variabile in corrispondenza di ciascuna sottoclasse.

Come indicato nel secondo capitolo, occorre distinguere il caso in cui N_t è conosciuto da quello in cui tale numero è incognito.

Se N_t è conosciuto si considera il subcampione costituito dalle sole unità che appartengono alla sottoclasse in esame e a ciascuna di esse si attribuisce un nuovo peso $w_i = N_t/n_t$. La stima del totale per la sottoclasse t è allora data da:

$$\hat{Y}_{(r)t} = \sum_{i \in A_t} w_i\ y_{(r)i} \qquad [3.8]$$

dove la somma è estesa alle sole unità appartenenti alla sottoclasse.

La stima della varianza campionaria è data da:

$$\hat{Var}(\hat{Y}_{(r)t}) = \frac{1-f_t}{n_t}\ N_t^2\ \hat{S}^2_{(r)t} \qquad [3.9]$$

dove:

$$\hat{S}^2_{(r)t} = \frac{1}{n_t - 1} \sum_{i \in A_t} (y_{(r)i} - \bar{y}_{(r)t})^2 \qquad [3.10]$$

e $\bar{y}_{(r)t}$ è la media artitmetica semplice dei valori osservati nelle n_t unità campione della sottoclasse.

Se N_t non è conosciuto per ogni unità del campione vengono definite le nuove variabili:

$$y'_{(r)i} = \begin{cases} y_{(r)i} & \text{se } i \in A_t \\ 0 & \text{altrimenti} \end{cases}$$

a cui vanno applicate le stesse formule valide per le stime relative all'intera popolazione.

3. Il campionamento con reimmissione

Poiché ad ogni estrazione può essere scelta una qualsiasi delle N unità della popolazione, indipendentemente dal risultato che si è verificato nelle prove precedenti, il numero dei campioni che possono essere generati da questo disegno campionario è dato da N^n e ciascuno di essi ha la stessa probabilità $1/N^n$ di essere estratto.

La probabilità di selezione dell'unità i.ma è sempre la stessa in ciascuna delle n estrazioni ed è uguale a 1/N.

Trattandosi di un campionamento con reimmissione per la stima di un totale si applica lo stimatore di Hansen-Hurwitz, per cui ponendo nella (2.4) $p_i = 1/N$ si ha:

$$\hat{Y}_{(r)} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_{(r)i}}{p_i} = \sum_{i=1}^{n} w_i \, y_{(r)i} \qquad [3.11]$$

dove $w_i = N/n$

Come si vede confrontando la [3.11] con la [3.3] si tratta dello stesso stimatore che viene utilizzato nel campionamento senza reimmissione.

Applicando la [2.7] si ricava la seguente espressione dello stimatore della varianza campionaria:

$$\hat{Var}(\hat{Y}_{(r)}) = \frac{N^2}{n^2(N-1)} \sum_{i=1}^{n} \sum_{j>i} (y_i - y_j)^2 \qquad [3.12]$$

che può anche essere scritta:

$$\hat{Var}(\hat{Y}_{(r)}) = N^2 \frac{\hat{S}^2_{(r)}}{n} \qquad [3.13]$$

Gli stimatori dell'errore standard assoluto e relativo sono dati rispettivamente da:

$$\hat{SE}(\hat{Y}_{(r)}) = \frac{N \, \hat{S}_{(r)}}{\sqrt{n}} \qquad [3.14]$$

e

$$\hat{RE}(\hat{Y}_{(r)}) = \frac{\hat{C}_{(r)}}{\sqrt{n}} \qquad [3.15]$$

La (3.13) può essere utilizzata per calcolare le varianze campionarie degli stimatori combinazioni lineari di totali e stimatori relativi a sottoclassi, per i quali devono essere seguiti gli stessi procedimenti riportati per il campionamento senza reimmissione.

4. Il fattore di correzione per popolazioni finite

Ritornando al campionamento senza reimmissione, dalla (3.7) si evince che l'errore relativo delle stime che si ottengono mediante un campione causale semplice di dimensione n è crescente con l'aumentare della numerosità della popolazione; infatti per valori di N più elevati diminuisce il tasso di campionamento e quindi risulta più grande il valore del fattore (1-f).

La quantità (1-f) prende il nome di fattore di correzione per popolazioni finite e può essere trascurata quando la numerosità della popolazione è molto grande rispetto a quella del campione, in quanto il suo valore risulta prossimo all'unità.

In genere, se si trascura il fattore di correzione per popolazioni finite, si sovrastimano gli errori di campionamento di una quantità che diventa trascurabile quando il tasso di campionamento è poco elevato. Nel prospetto che segue sono riportati i valori della sovrastima dell'errore relativo nel caso di omissione del fattore di correzione, in corrispondenza a diversi valori di f.

f	sovrastima	f	sovrastima
0,001	0,0005	0,030	0,0153
0,002	0,0010	0,040	0,0206
0,003	0,0015	0,050	0,0260
0,004	0,0020	0,100	0,0541
0,005	0,0025	0,200	0,1180
0,010	0,0050	0,300	0,1952
0,020	0,0102	0,400	0,2910

Dall'esame di questi valori si evince che il fattore di correzione può essere praticamente ignorato per valori di f inferiori al 5%, in quanto per essi si ha una sovrastima dell'errore relativo minore del 2,6%.

In questi casi la stima della varianza campionaria coincide con quella del campionamento con reimmissione e, come si evince dalla [3.15], l'errore relativo delle stime non dipende dal tasso di campionamento ma soltanto dalla numerosità del campione.

A titolo esemplificativo si considerino due popolazioni costituite rispettivamente 100.000 e 10.000 unità e aventi lo stesso coefficiente di variazione CV = 1,42 per una determinata variabile. Se si effettua un campionamento casuale semplice di

numerosità n = 100 in ciascuna delle due popolazioni, gli errori relativi attesi delle stime dei totali saranno uguali rispettivamente a 14,19% e 14,13%. Gli errori relativi risultano approssimativamente uguali anche se sono stati adottati due tassi di campionamento molto diversi ($f_1 = 1/1000$ e $f_2 = 1/100$).

Questo risultato contraddice un'idea assai diffusa tra coloro che non hanno molta dimestichezza con la teoria dei campioni, secondo cui è il tasso di campionamento che condiziona la precisione delle stime.

La [3.15] è, inoltre, molto utile da un punto di vista pratico, in quanto essa consente di valutare molto rapidamente la variazione che si ha nell'errore delle stime quando si passa da un campione di numerosità n_1 ad uno di numerosità n_2. La variazione, misurata come rapporto tra i due errori relativi, è infatti data dalla radice quadrata del rapporto tra n_1 e n_2. Così se si quadrupla la dimensione del campione si dimezza l'errore relativo delle stime.

5. Stabilità dello stimatore della varianza campionaria

Nelle indagini campionarie l'obiettivo principale è quello di fornire una stima dei parametri d'interesse, mentre la stima della varianza campionaria costituisce un obiettivo secondario, che assume rilevanza soltanto nell'analisi dei risultati e nella programmazione del campione per indagini successive. Tuttavia anche nel dare informazioni sugli errori di campionamento ci si deve interrogare sulla loro attendibilità, in quanto essi stessi sono stime desunte dai dati campionari e quindi affetti da errore. In altri termini si pone il problema della stabilità o della precisione dello stimatore della varianza campionaria, che nel campionamento casuale semplice coincide con quella di $\hat{S}^2_{(n)}$.

Si dimostra (cfr. Kish 1965, pag. 289) che la precisione dello stimatore, misurata dal suo coefficiente di variazione è data da:

$$CV^2 = \frac{1}{n}\left(\beta - \frac{n-3}{n-1}\right) \qquad [3.16]$$

dove β è l'indice di curtosi della variabile in esame:

Nella tavola 3.1 sono riportati i valori di β per alcune delle distribuzioni statistiche che più frequentemente si incontrano nella pratica.

Tavola 3.1 - Valori di β per alcune distribuzioni statistiche

Tipo di distribuzione			β
Uniforme			1,80
Triangolare			2,40
Normale			3,00
Binomiale: $\beta = (1\text{-}3\,PQ)/PQ$			
	valori di P		
	0,01	0,99	98,01
	0,05	0,95	18,05
	0,10	0,90	8,11
	0,15	0,85	4,84
	0,20	0,80	3,5
	0,25	0,75	2,33
	0,30	0,70	1,76
	0,35	0,65	1,0
	0,40	0,40	1,17
	0,45	0,55	1,04
	0,05	0,50	1,00

Le prime tre distribuzioni, alle quali si può fare riferimento nel caso della stima dell'ammontare totale di un carattere quantitativo, presentano valori di β compresi in un intervallo piuttosto ristretto ($1{,}80 < \beta < 3{,}00$). Pertanto, anche considerando il caso più sfavorevole $\beta = 3$, è sufficiente un campione di 200 unità per avere una stima della varianza campionaria con un errore relativo inferiore al 10%.

Per la distribuzione binomiale, che viene utilizzata nel caso di stime di frequenze, i valori di β risultano variabili con P; pertanto la precisione della stima della varianza campionaria dipende, oltre che dalla numerosità del campione, anche dall'ordine di grandezza della frequenza relativa che deve essere stimata.

Così per valori di P compresi tra 0,10 e 0,90 è sufficiente un campione di 700 unità per assicurare un errore relativo inferiore al 10%, mentre per valori di P inferiori a 0,01 o superiori a 0,99 occorre un campione di oltre 9750 unità per avere la stessa precisione.

6. Procedura informatica per il calcolo degli errori campionari

Utilizzando la metodologia descritta nei precedenti paragrafi, l'autore ha predisposto una procedura informatica che consente il calcolo degli errori campionari di stimatori esprimibili come combinazioni lineari di totali, con riferimento sia all'intera popolazione che a sottoclassi.

Per illustrare la procedura e le modalità che devono essere seguite per il calcolo degli errori di campionamento si è fatto

riferimento ad un esempio basato su dati fittizi, ma che non si discosta, se non nelle dimensioni, dai casi che si possono presentare nel concreto.

I dati di base

Si supponga di avere effettuato un'indagine campionaria su una popolazione di N = 800 medici di famiglia che operano sul territorio di una USL, utilizzando un campione casuale semplice senza reimmissione di n = 40 medici e di aver rilevato per ciascuno di essi: il sesso, l'età in anni compiuti, il numero di pazienti assistiti, il numero medio giornaliero sia delle visite ambulatoriali che di quelle domiciliari.

Una volta registrati i dati rilevati si disporrà di un file con il seguente tracciato record:

colonne	variabili
1-2	numero d'ordine
3	sesso
4-5	età
6-9	numero di assistiti
10-11	numero di visite ambulatoriali
12	numero di visite domiciliari

Le variabili rilevate sono tutte quantitative ad esclusione di quella relativa al sesso, per la quale sono stati utilizzati i codici 1 e 2 per indicare rispettivamente le modalità maschio e femmina. I valori sono riportati nella tavola 3.2.

Parametri e sottoclassi

Si ipotizzi, inoltre, che obiettivo dell'indagine sia quello di fornire le stime per l'intera popolazione e per classi di età dei medici, dei seguenti parametri:

θ_1 = % di medici di sesso maschile
θ_2 = numero medio di assistiti per medico
θ_3 = numero medio di visite per medico
θ_4 = numero medio di visite per assistito
θ_5 = % di visite ambulatoriali sulle visite complessive

dove le sottoclassi sono così definite:

A_1 = fino a 40 anni
A_2 = 41-50 anni
A_3 = oltre 50 anni

Tavola 3.2 - Valori osservati nel campione

unità campione	Sesso	Età	Assistiti	visite ambulat.	visite domicil.
1	1	61	817	60	18
2	1	52	1144	84	24
3	2	37	397	48	12
4	1	40	512	48	12
5	1	46	1237	72	18
6	2	38	615	60	18
7	2	41	797	54	12
8	1	38	649	60	18
9	1	62	581	42	6
10	2	58	747	66	18
11	2	46	832	60	12
12	1	57	678	42	18
13	1	63	954	66	24
14	1	60	1132	66	18
15	2	64	974	54	12
16	2	54	549	48	6
17	1	53	512	42	6
18	1	45	817	60	18
19	1	42	438	48	12
20	1	48	986	60	12
21	1	60	815	66	18
22	1	51	1142	72	24
23	2	38	395	54	12
24	1	39	510	54	12
25	1	45	1235	78	18
26	2	37	614	66	24
27	2	42	796	60	12
28	1	39	647	66	18
29	1	63	580	48	6
30	2	59	746	72	18
31	2	45	831	66	12
32	1	56	677	48	18
33	1	61	952	72	12
34	1	62	1130	72	18
35	2	60	975	60	12
36	2	53	660	54	6
37	1	51	514	48	6
38	1	47	815	66	18
39	1	43	440	54	12
40	1	44	985	66	12

In primo luogo è necessario introdurre nel file dei dati di base la variabile indicatrice della modalità «maschio», la variabile che identifica le tre sottoclassi per la quali devono essere prodotte le stime e la variabile «numero di visite complessive» ottenuta come somma del numero delle visite ambulatoriali e di quelle domiciliari.

Variabile indicatrice della modalità «maschio»

$$y_{(1)i} = \begin{cases} 1 & \text{se sesso} = 1 \\ 0 & \text{altrimenti} \end{cases}$$

Variabile che identifica le sottoclassi

$$\text{CLASSE} = \begin{cases} 1 & \text{se ETA} < 41 \\ 2 & \text{se } 40 < \text{ETA} < 51 \\ 3 & \text{se } 50 < \text{ETA} \end{cases}$$

Occorre quindi stimare i totali delle variabili che intervengono nella stima dei parametri e, successivamente, determinare le opportune trasformazioni lineari.

Indicando con $y_{(2)i}$ $y_{(3)i}$ e $y_{(4)i}$ i valori che le variabili relative al numero di assistiti, di visite mediche complessive e di visite ambulatoriali, assumono in corrispondenza dell'i.mo medico intervistato, la stima dei totali è data da:

$$\hat{Y}_{(r)} = \sum_{i=1}^{n} w_i \, y_{(r)i} \qquad (r = 1, 2, 3, 4, 5)$$

I valori delle variabili combinazioni lineari delle variabili originali sono dati da:

$$z_{(1)i} = \frac{y_{(1)i}}{N} 100 \qquad \text{per la stima di } \theta_1$$

$$z_{(2)i} = \frac{y_{(2)i}}{N} 100 \qquad \text{per la stima di } \theta_2$$

$$z_{(3)i} = \frac{y_{(3)i}}{N} 100 \qquad \text{per la stima di } \theta_3$$

$$z_{(4)i} = \frac{\hat{Y}_{(3)}}{\hat{Y}_{(2)} N} + \frac{y_{(3)i}}{\hat{Y}_{(2)} N} - \frac{\hat{Y}_{(3)}}{\hat{Y}^2_{(2)}} y_{(2)i} \qquad \text{per la stima di } \theta_4$$

$$z_{(5)i} = 100 \left(\frac{\hat{Y}_{(4)}}{\hat{Y}_{(3)} N} + \frac{y_{(4)i}}{\hat{Y}_{(3)}} - \frac{\hat{Y}_{(4)}}{\hat{Y}^2_{(3)}} y_{(3)i} \right) \qquad \text{per la stima di } \theta_5$$

Per le stime relative alle sottoclassi vengono prima definite le variabili con apice introdotte nel paragrafo precedente, e successivamente vengono operate le trasformazioni lineari sulle quali vengono calcolate le stime dei parametri e le rispettive varianze campionarie mediante le formule [3.3] e [3.10].

Input del programma

Il programma, che utilizza procedure SAS, prevede come input:

- il file dei dati di base
- il numero delle sottoclassi (NC)
- il numero di variabili di tipo y (NY)
- il numero di variabili di tipo z (NZ)
- le istruzioni per la ricodifica delle variabili indicatrici e delle sottoclassi
- le istruzioni per le trasformazioni lineari

Output del programma

L'output è costituito da una tavola per la popolazione e una per ciascuna sottoclasse, contenenti per ogni parametro considerato:

- la stima
- l'errore standard
- l'errore relativo
- gli estremi dell'intervallo di confidenza per P = 95%.

Le tavole con i risultati finali delle elaborazioni effettuate sono riportate alla fine del capitolo.

Trattandosi di un'applicazione su dati fittizi non è necessario effettuare un'analisi approdondita dei risultati, anche se è opportuno svolgere alcune considerazioni che hanno carattere di generalità. A tale scopo nella tavola che segue sono stati riportati la numerosità del campione e gli errori relativi delle stime per la popolazione totale e ciascuna delle tre sottoclassi considerate.

Tavola 3.3 - Numerosità del campione ed errori relativi delle stime per l'intera popolazione e per sottoclassi

	Popolazione	Classi di età		
		fino a 40	41-50	oltre 50
n. medici campione	40	8	12	20
Parametri		Errori Relativi		
n. medici nella popolaz.	0.00	31.21	23.84	15.61
% medici maschi	11.45	46.82	31.21	21.27
n. assistiti per medico	4.72	31.85	25.14	16.61
n.visite per medico	2.98	31.62	24.13	16.64
n. visite per 100 assis.	3.34	3.65	5.22	3.24
% visite ambulatoriali	0.83	1.29	0.82	1.45

Dall'esame dei dati si evince quanto segue:

a) sia per l'intera popolazione che per le sottoclassi gli errori relativi variano notevolmente con il parametro oggetto di stima, e ciò è dovuto alla diversa variabilità dei caratteri oggetto d'indagine
b) gli errori relativi della popolazione sono sempre minori dei corrispondenti errori delle sottoclassi, e questo perché le stime della popolazione sono state derivate da un campione di numerosità maggiore
c) per lo stesso motivo gli errori relativi delle sottoclassi risultano, generalmente, decrescenti con l'aumentare del numero dei medici campionati.

Quanto riscontrato nell'applicazione effettuata si ritrova nelle indagini reali, anche quando si utilizzano disegni campionari più complessi.

Come si vedrà in seguito queste regolarità consentono di costruire opportuni modelli, mediante i quali si possono trasferire gli errori campionari, calcolati su un numero ristretto di variabili e sottoclassi, ad altre variabili e sottoclassi. In questo modo è possibile ottenere le informazioni sugli errori di campionamento per un gran numero di stime senza dover procedere al calcolo diretto, con un notevole risparmio dei tempi di elaborazione e del numero delle tavole che devono essere pubblicate.

Tavola 3.4 - Errori campionari per l'intera popolazione

	STIMA	SE	RE	INF	SUP
Parametri					
n. medici	800.00	0.00	0.00	800.00	800.00
% Maschi	65.00	7.44	11.45	57.56	72.44
n. Assistiti per medico	773.42	36.53	4.72	736.89	809.96
n. Visite per medico	74.10	2.21	2.98	71.89	76.31
n. Visite per 100 assistiti	9.58	0.32	3.34	9.26	9.90
% Visite ambulatoriali	80.36	0.67	0.83	79.69	81.03

Tavola 3.5 - Errori campionari per la sottoclasse: fino a 40 anni

	STIMA	SE	RE	INF	SUP
Parametri					
n. medici	160.00	49.94	31.21	110.06	209.94
% Maschi	50.00	23.41	46.82	26.59	73.41
n. Assistiti per medico	542.37	172.75	31.85	369.62	715.13
n. Visite per medico	72.75	23.01	31.62	49.74	95.76
n. Visite per 100 assistiti	13.41	0.49	3.65	12.92	13.90
% Visite ambulatoriali	78.35	1.01	1.29	77.34	79.36

Tavola 3.6 - Errori campionari per la sottoclasse: 41-50 anni

	STIMA	SE	RE	INF	SUP
Parametri					
n. medici	240.00	57.22	23.84	182.78	297.22
% Maschi	66.67	20.81	31.21	45.86	87.48
n. Assistiti per medico	850.75	213.89	25.14	636.86	1064.64
n. Visite per medico	76.00	18.34	24.13	57.66	94.34
n. Visite per 100 assistiti	8.93	0.47	5.22	8.47	9.40
% Visite ambulatoriali	81.58	0.67	0.82	80.91	82.25

Tavola 3.7 - Errori campionari per la sottoclasse: oltre 50 anni

	STIMA	SE	RE	INF	SUP
Parametri					
n. medici	400.00	62.43	15.61	337.57	462.43
% Maschi	70.00	14.89	21.27	55.11	84.89
n. Assistiti per medico	819.45	136.12	16.61	683.33	955.57
n. Visite per medico	73.50	12.08	16.44	61.42	85.58
n. Visite per 100 assistiti	8.97	0.29	3.24	8.68	9.26
% Visite ambulatoriali	80.41	1.16	1.45	79.25	81.57

CAPITOLO 4 – APPLICAZIONE DELLA METODOLOGIA STANDARD AL CAMPIONAMENTO STRATIFICATO

1. Il campionamento stratificato

Scopi della stratificazione

La stratificazione è una tecnica frequentemente utilizzata nella progettazione dei piani di campionamento e consiste nel suddividere le N unità della popolazione in un certo numero di gruppi o strati non sovrapposti. Il campione stratificato si ottiene estraendo un campione causale semplice da ciascuno strato.

La stratificazione viene introdotta, in primo luogo, per programmare il campione in corrispondenza di determinate subpopolazioni (domini di studio), in modo da contenere l'errore atteso delle stime entro certi limiti stabiliti.

Così nell'esempio riportato nel capitolo precedente si è visto che, adottando un campionamento casuale semplice, le numerosità dei subcampioni relativi alle tre classi di età dei medici sono risultate pari a 8, 12 e 20. Tali numerosità non risultano determinate a priori ma sono affidate al caso, per cui gli errori campionari delle stime relative alle tre sottoclassi, non possono essere tenuti sotto controllo, a meno che non si scelga una numerosità del campione totale molto elevata e quindi non economica.

Se si dispone di una lista contenente, oltre ai nominativi degli 800 medici che costituiscono la popolazione, anche il loro anno di nascita, è possibile suddividere la popolazione totale in tre subpopolazioni in base alla classe di età. Si può allora estrarre da ciascuna di esse un campione casuale semplice della numerosità desiderata.

In secondo luogo la stratificazione viene utilizzata per ridurre gli errori campionari delle stime.

Come si è visto nel campionamento casuale semplice la varianza campionaria della stima è direttamente proporzionale alla varianza del carattere nella popolazione; pertanto è possibile ridurre la varianza campionaria, senza modificare la numerosità complessiva del campione, diminuendo la variabilità del carattere. Ciò può essere ottenuto suddividendo le unità della popolazione in strati che al loro interno siano il più possibile omogenei rispetto alle variabili oggetto di indagine.

Ovviamente tale obiettivo può essere perseguito soltanto se per ogni unità della popolazione si dispone di informazioni supplementari, costituite dai valori di una o più variabili ausiliarie correlate con le variabili di rilevazione (cfr. Zannella, 1983).

Un'ultima ragione, non meno importante delle precedenti, che motiva il ricorso alla stratificazione va ricercata nelle esigenze amministrative ed organizzative che possono richiedere una suddivisione del territorio, ad esempio in regioni o province, in modo da decentrare le operazioni connesse con l'implementazione del campione e l'esecuzione della rilevazione stessa.

Selezione del campione

I problemi che devono essere affrontati per la programmazione e la selezione di un campione stratificato riguardano da un lato la formazione degli strati (individuazione della variabili di stratificazione, determinazione del numero degli strati, scelta dei criteri per il raggruppamento delle unità), dall'altro il calcolo della numerosità del campione totale e la sua ripartizione tra gli strati (cfr. Zannella, 1985).

Senza entrare nel merito delle diverse soluzioni che possono essere adottate per le quali si rimanda al fascicolo 4 del Manuale di tecniche di indagine, qui si sofferma l'attenzione sulle modalità che devono essere seguite per la selezione del campione.

L'estrazione del campione può essere effettuata in due modi diversi a seconda delle informazioni di base di cui si dispone:

a) stratificazione prima della selezione

b) stratificazione dopo selezione

Stratificazione prima della selezione

Il primo metodo presuppone la disponibilità di un file di dati di base contenente per ogni unità della popolazione, oltre ai codici identificativi, i valori delle variabili ausiliarie che vengono utilizzate per la stratificazione. L'applicazione del criterio di stratificazione prescelto comporta l'attribuzione ad ogni unità di un ulteriore codice, che identifica lo strato di appartenenza. In genere il file viene ordinato per codice di strato e ad ogni unità viene assegnato un numero d'ordine progressivo all'interno dello strato.

Elaborando le informazioni contenute nel file dei dati di base si ricava il file degli strati, con tanti record quanti sono gli strati che sono stati formati. Ciascun record contiene il codice identificativo dello strato e tutte le informazioni necessarie per la determinazione delle dimensioni del campione (numerosità della popolazione in ogni strato, varianza delle variabili «guida», etc.).

Sulla base dei criteri scelti viene determinata la numerosità complessiva del campione e la sua ripartizione nei singoli strati, ottenendo un file così costituito:

Strato	Unità nella popolazione	Unità nel campione	Tasso di campionamento
1	N_1	n_1	$f_1 = n_1/N_1$
2	N_2	n_2	$f_2 = n_2/N_2$
.	.	.	.
.	.	.	.
.	.	.	.
L	N_L	n_L	$f_L = n_L/N_L$

Si ritorna sul file dei dati di base e mediante un programma che genera numeri casuali vengono selezionate le unità da campionarie in ciascuno strato.

Si predispone, quindi, la lista delle unità campione, contenente per ciascuna unità il codice di strato, il codice identificativo e tutte le informazioni necessarie per la sua corretta individuazione sul territorio.

Stratificazione dopo selezione

La stratificazione dopo selezione viene utilizzata quando il file dei dati di base non contiene le informazioni necessarie per attribuire a ciascuna unità il codice di strato.

Si supponga, ad esempio, che per effettuare un'indagine campionaria sulle imprese industriali si voglia utilizzare un campione ad uno stadio stratificato, con stratificazione delle imprese in L classi di fatturato. Si conosce il numero complessivo N delle unità della popolazione e la numerosità N_h (h = 1, 2, . . ., L) di ogni singolo strato. Si dispone, inoltre, di una lista contenente la denominazione e l'indirizzo di ciascuna impresa, ma non il fatturato, per cui non è possibile procedere ad una stratificazione delle unità della popolazione.

Tale inconveniente può essere superato se la raccolta delle informazioni per l'attribuzione del codice di strato ad ogni unità campione, non comporta un costo eccessivo.

In questo caso, dopo aver predeterminato la numerosità campionaria nei singoli strati, si procede all'estrazione del campione, con l'avvertenza di selezionare un'unità campione alla volta e di assegnare a ciascuna di esse un numero progressivo, in base all'ordine di estrazione.

In ciascuna unità campione vengono rilevate le informazioni relative alle variabili di stratificazione, e in base ai valori osservati si procede alla loro classificazione negli strati, seguendo l'ordine di estrazione.

Una volta che in uno strato sia stata raggiunta la numerosità campionaria programmata, tutte le successive unità del campione che cadono nello strato non vengono più prese in considerazione.

Questo procedimento risulta dispendioso, in quanto richiede la selezione di un numero di unità molto più elevato di quello su cui poi viene condotta la rilevazione, ma, se i costi per rilevare le variabili di stratificazione sono contenuti, può trovare ugualmente utile applicazione.

Il file dei dati campionari

Indipendentemente dal metodo seguito per la selezione del campione stratificato, una volta effettuata la rilevazione e svolte tutte le successive operazioni di trattamento dei dati raccolti, si disporrà di un file costituito da n record, uno per ciascuna unità del campione totale. Ogni record conterrà il codice di strato, il numero identificativo all'interno dello strato dell'unità campionata e i valori osservati delle variabili oggetto di indagine.

Indicando con p il numero delle variabili rilevate, all'unità i.ma dello strato h corrisponderà il seguente vettore di valori:

$$h,\ i,\ y_{(1)hi},\ y_{(2)hi},\ \ldots,\ y_{(p)hi}$$

Le probabilità d'inclusione

Poiché il campione stratificato può essere riguardato come l'unione di L campioni casuali semplici selezionati indipendentemente l'uno dall'altro, le probabilità d'inclusione del primo ordine delle unità dello strato h sono date da:

$$\pi_{hi} = \frac{n_h}{N_h} \qquad [4.1]$$

e coincidono con il tasso di campionamento f_h.

A ciascuna unità del campione viene quindi attribuito un coefficiente di espansione dato da:

$$w_{hi} = \frac{1}{\pi_{hi}} = \frac{N_h}{n_h} \qquad [4.2]$$

Le espressioni che danno le probabilità d'inclusione del secondo ordine sono diverse a seconda che le due unità appartengono allo stesso strato o a due strati diversi. Nel primo caso si ha:

$$\pi_{hij} = \frac{n_h}{N_h}\ \frac{n_h - 1}{N_h - 1} \qquad [4.3]$$

Nel secondo caso indicando con h e k ≠ h i due strati, si ha:

$$\pi_{hi,\,kj} = \frac{n_h}{N_h}\ \frac{n_k}{N_k} \qquad [4.4]$$

2. Stima della varianza campionaria

Nel fornire le formule per la stima della varianza campionaria verrà, dapprima, trattato il caso della stima del totale in un singolo strato e, successivamente i casi della stima di un totale riferito all'intera popolazione, a un dominio di studio e ad una sottoclasse.

Nel caso di stime che possono essere espresse, o approssimate, mediante combinazioni lineari di totali, si dovrà, in primo luogo, procedere alla costruzione di nuove variabili utilizzando le formule [2.33] ÷ [2.37] del secondo capitolo; quindi, alle variabili così ottenute, dovranno essere applicate le formule relative alla stima di un totale, che sono riportate in questo paragrafo.

Stimatore del totale in uno strato

Poiché in ciascuno strato viene effettuato un campionamento casuale semplice senza reimmissione per derivare la stima di un totale e la corrispondente stima della varianza campionaria si possono utilizzare le formule [3.3] e [3.4] riportate nel capitolo precedente:

Stimatore del totale

$$\hat{Y}_{(r)h} = \sum_{i=1}^{n_h} w_{hi}\ y_{(r)hi} \qquad [4.5]$$

stimatore della varianza campionaria

$$\hat{Var}(\hat{Y}_{(r)h}) = \frac{1 - f_h}{n_h}\ N_h^2\ \hat{S}^2_{(r)h} \qquad [4.6]$$

dove:

$$\hat{S}^2_{(r)h} = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{(r)hi} - \bar{y}_{(r)h})^2 \qquad [4.7]$$

è la stima corretta della varianza della variabile r.ma nello strato h.

Stimatore del totale nell'intera popolazione

Lo stimatore dell'ammontare totale di un carattere nella popolazione è dato dalla somma degli stimatori dei totali nei singoli strati:

$$\hat{Y}_{(r)} = \sum_{h=1}^{L} \hat{Y}_{(r)h} \qquad [4.8]$$

Poiché gli stimatori $\hat{Y}_{(r)h}$ (h = 1, . . ., L) sono indipendenti, la stima della varianza della loro somma è data dalla somma delle stime delle varianze:

$$\hat{Var}(\hat{Y}_{(r)}) = \sum_{h=1}^{L} \hat{Var}(\hat{Y}_{(r)h}) \qquad [4.9]$$

Stimatore del totale in un dominio di studio

Per dominio di studio si intende una subpopolazione costituita da uno o più strati. Il numero delle unità della popolazione appartenenti ad un determinato dominio di studio è sempre conosciuto, infatti esso è dato dalla somma del numero delle unità contenute negli strati che formano il dominio.

Lo stimatore di un totale è dato dalla somma degli stimatori dei totali degli strati compresi nel dominio:

$$\hat{Y}_{(r)d} = \sum_{h \in d} \hat{Y}_{(r)h} \qquad [4.10]$$

In modo analogo si ottiene lo stimatore della varianza campionaria:

$$\hat{Var}(\hat{Y}_{(r)d}) = \sum_{h \in d} \hat{Var}(\hat{Y}_{(r)h}) \qquad [4.11]$$

Stimatore del totale in una sottoclasse

Per sottoclasse della popolazione, e quindi del campione, si intende un insieme di unità appartenenti a strati differenti ma caratterizzate dal presentare la stessa modalità di uno o più caratteri rilevati. Il numero delle unità della popolazione che appartengono ad una determinata sottoclasse è, generalmente sconosciuto, e le unità campionate nella sottoclasse sono distribuite, più o meno uniformemente, su tutti gli strati.

Per derivare lo stimatore del totale e la relativa varianza campionaria viene seguito lo stesso procedimento adottato per il campionamento casuale semplice nel caso in cui il numero delle unità della popolazione che cadono nella sottoclasse non è conosciuto.

Per ogni variabile oggetto di stima si definisce una nuova variabile che assume lo stesso valore di quella originaria nelle unità appartenenti alla sottoclasse e il valore zero nelle altre unità:

$$y'_{(r)hi} = \begin{cases} y_{(r)hi} & \text{se } i \in A_t \\ 0 & \text{altrimenti} \end{cases}$$

dove con A_t si è indicata la sottoclasse in esame.

È bene evidenziare che le stime per una determinata sottoclasse possono essere calcolate sia con riferimento alla popolazione totale che a un particolare dominio di studio. Così ad esempio, in una indagine campionaria sulla fecondità potrebbe essere richiesta la stima del numero medio di figli per la sottoclasse costituita dalle donne coniugate che lavorano, con riferimento sia all'intero territorio nazionale che ai domini di studio costituiti dalle singole regioni.

Per stimare la varianza campionaria nel caso di sottoclassi vanno, quindi, applicate alle variabili con apice definite precedentemente, le formule [4.9] o [4.11] a seconda che lo stimatore sia riferito all'intera popolazione o ad un dominio di studio.

3. Strati con una sola unità campione

A volte può accadere che in uno o più strati venga rilevata una sola unità campione, e questo per i seguenti motivi:

a) una o più unità della popolazione presentano valori particolari delle variabili di stratificazione, per cui ciascuna di esse costituisce uno strato a se stante ($N_h = 1$, $n_h = 1$)

b) si è effettuata una stratificazione molto fine che ha comportato la ripartizione delle unità della popolazione in un gran numero di strati, per cui in alcuni di essi il campione programmato prevede la selezione di una sola unità campione ($N_h > 1$, $n_h = 1$)

c) il campione programmato prevede due o più unità campione per strato, ma a causa delle cadute campionarie in alcuni strati è stato possibile rilevare una sola unità ($N_h > 1$, $n_h = 1$)

Il primo caso non comporta complicazioni nella stima della varianza campionaria, che ovviamente risulta uguale a zero, in quanto la rilevazione è stata condotta su tutta la popolazione

dello strato, che d'altra parte presenta variabilità nulla essendo costituita da una sola unità.

Negli altri due casi non è possibile stimare la varianza $S^2_{(r)h}$, in quanto occorre disporre di almeno due unità campione per strato.

Per superare questo inconveniente si applica la tecnica denominata «collapsed strata», introdotta da Hansen, Hurwitz e Madow (1953), che consiste nel raggruppare gli strati in modo da assicurare almeno due unità per strato.

Siano h e k i due strati che devono essere raggruppati, si indichino con N_h e N_k il numero delle unità della popolazione in essi comprese e con n_h e n_k le numerosità campionarie, con almeno uno dei due numeri uguale a 1.

Sia g il nuovo strato che deve essere formato, esso sarà costituito da $N_g = N_h + N_k$ unità della popolazione in cui sono state campionate $n_g = n_h + n_k$ unità.

Lo stimatore del totale nello strato g si ottiene applicando la [4.5]:

$$\hat{Y}_{(r)g} = \sum_{i=1}^{n_g} w_{gi}\, y_{(r)gi} \qquad [4.12]$$

dove i pesi w_{gi} sono così definiti:

$$w_{gi} = \begin{cases} w_{hi} & i \in h \\ w_{gi} & i \in g \end{cases} \qquad [4.13]$$

Si verifica facilmente che $\hat{Y}_{(r)g} = \hat{Y}_{(r)h} + \hat{Y}_{(r)k}$ e che quest'ultimo è uno stimatore corretto del totale dello strato ottenuto dal collassamento.

La varianza campionaria di $\hat{Y}_{(r)g}$ si ottiene applicando [4.6]

$$\hat{Var}\left(\hat{Y}_{(r)g}\right) = \frac{1-f_g}{n_g} N_g^2\, \hat{S}^2_{(r)g} \qquad [4.14]$$

dove:

$$\hat{S}^2_{(r)g} = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (y_{(r)gi} - \bar{y}_{(r)g})^2 \qquad [4.15]$$

Si dimostra (cfr. Cochran 1977, pp. 138-139) che la [4.15] fornisce una stima distorta per eccesso della varianza nello strato g, per cui sostituendo tale valore nella [4.9] si ha una sovrastima della varianza campionaria dello stimatore del totale.

Le formule date per n_h e n_k generici, rimangono valide quando uno o entrambi gli strati presentano una numerosità del campione uguale ad 1.

4. Valutazione dell'effetto della stratificazione

Come è stato detto nella premessa, uno degli scopi della stratificazione è quello di ridurre la varianza campionaria degli stimatori. Ci si può, pertanto, chiedere quel'è il guadagno che si realizza adottando un campionamento stratificato al posto di un campionamento casuale semplice.

Per valutare l'effetto della stratificazione occorre in primo luogo stimare la varianza campionaria per il campionamento casuale semplice, utilizzando i valori osservati su un campione stratificato.

Si dimostra (cfr. Cochran 1977, pp. 136-137) che un'espressione approssimata di tale stima è data da:

$$\hat{V0}\left(\hat{\bar{Y}}_{(r)}\right) = \frac{1-f}{n}\left[N \sum_{h=1}^{L} \frac{N_h}{n_h}\, y^2_{(r)hi} - \hat{Y}^2_{(r)}\right] \qquad [4.16]$$

dove f = n/N è il tasso di campionamento totale.

Nel caso di un campionamento stratificato proporzionale (o autoponderante), per il quale vale la relazione:

$$n_h = n \frac{N_h}{N} \qquad [4.17]$$

la [4.16] si semplifica nel seguente modo:

$$\hat{V0}\left(\hat{\bar{Y}}_{(r)}\right) = \frac{1-f}{n} N^2\, \hat{S}^2_{(r)} \qquad [4.18]$$

dove:

$$\hat{S}^2_r = \frac{1}{n-1} \sum_{h=1}^{L} \sum_{i=1}^{n_h} (y_{(r)hi} - \bar{y}_{(r)})^2 \qquad [4.19]$$

e

$$\bar{y}_r = \frac{1}{n} \sum_{h=1}^{L} \sum_{i=1}^{n_h} y_{(r)hi} \qquad [4.20]$$

Effetto del disegno

Un indicatore, molto utilizzato nella pratica, per valutare l'influenza del piano di campionamento sugli errori campionari è l'effetto del disegno (deff), introdotto da Kish (1965, pp.257-258), dato da:

$$deff\left(\hat{\bar{Y}}_{(r)}\right) = \frac{\hat{Var}\left(\hat{\bar{Y}}_{(r)}\right)}{\frac{1-f}{n} N^2 \hat{S}^2_{(r)}} \qquad [4.21]$$

dove con $\hat{V}ar(\hat{Y}_r)$ si è indicata la varianza campionaria dello stimatore del totale per il piano di campionamento in esame.

Nel caso di un campionamento stratificato la [4.21] misura l'effetto complessivo della stratificazione e dell'allontanamento dell'autoponderazione. Se si vuole una misura separata delle due componenti, occorre utilizzare anche la [4.16], ottenendo:

$$deff\left(\hat{\bar{Y}}_{(r)}\right) = \frac{\hat{Var}\left(\hat{\bar{Y}}_{(r)}\right)}{\hat{V}0\left(\hat{\bar{Y}}_{(r)}\right)} \cdot \frac{\hat{V}0\left(\hat{\bar{Y}}_{(r)}\right)}{\frac{1-f}{n} N^2 S^2_{(r)}} \qquad [4.22]$$

in cui i due fattori e secondo membro misurano rispettivamente l'effetto dovuti alla stratificazione e quello della non autoponderazione.

5. Procedura informatica per il calcolo degli errori campionari

Anche per il campionamento stratificato, così come per quello casuale semplice, è stata messa a punto dall'autore una procedura informatica per il calcolo degli errori di campionamento.

Il programma, che utilizza istruzioni SAS, richiede come input il file dei dati rilevati nel campione, contenente per ogni unità: i valori delle variabili rilevate, il codice di strato e il coefficiente di espansione.

Il programma calcola la numerosità del campione e della popolazione in ciascuno strato e stampa la lista degli eventuali strati nei quali $n_h = 1$ e $N_h > 1$. Per questi strati si dovrà procedere al raggruppamento seguento il metodo descritto nel paragrafo 3.

Per non generare confusione fra gli strati originari e quelli ottenuti dopo collassamento, il programma provvede ad inserire nel file una nuova variabile contenente il codice identificativo degli strati che vengono utilizzati per il calcolo della varianza campionaria.

Devono, quindi essere inputati:

a) il numero di domini territoriali (ND)
b) il numero di sottoclassi (NC)
c) il numero delle variabili che devono essere lette (NY)
d) il numero delle stime per le quali è richiesto il calcolo della varianza campionaria (NZ).

Mediante apposite istruzioni MACRO vengoni introdotte le seguenti nuove variabili:

«DOMINIO» che assume i valori 0, 1, 2, ..., ND (dove 0 indica il dominio totale);

«CLASSE» che assume i valori 0, 1, 2, ... NC (dove 0 indica la sottoclasse totale);

le variabili indicatrici delle modalità nel caso di stime di frequenza.

Successivamente, il programma provvede al calcolo dei valori delle variabili trasformazioni lineari, che verranno utilizzati per ottenere le stime dei parametri e le corrispondenti varianze campionarie.

L'output è costituito da (ND + 1) (NC + 1) tavole, una per ciascun dominio e sottoclasse, compresi il dominio totale e la sottoclasse totale. In ciascuna tavola viene riportato per ogni parametro:

- la denominazione del parametro che è stato stimato
- la stima del parametro
- la stima dell'errore standard (SE)
- la stima dell'errore relativo (RE)
- gli estremi dell'intervallo di confidenza al 95% (INF e SUP)

Applicazione della procedura

A chiarimento di quanto fino qui esposto, si riporta un'applicazione della procedura ai risultati dell'indagine campionaria sull'inserimento professionale dei laureati (Istat, 1989).

La popolazione, da indagare è costituita dai 72.124 laureati nell'anno 1986, stratificati secondo la sede dell'università e il

corso di laurea, per un totale di 717 strati. La rilevazione è stata condotta, mediante questionario postale, su un campione di 13.514 laureati, ripartiti in modo proporzionale tra i singoli strati.

Ad indagine ultimata si sono avuti di ritorno 9712 questionari con un tasso di risposta di circa il 72%. Poiché le cadute campionarie non si sono distribuite proporzionalmente tra gli strati, il campione effettivo non è più autoponderante. Pertanto i coefficienti di espansione hanno assunto valori variabili da strato a strato.

Inoltre, a causa della non risposte, in alcuni strati si è avuta una solo unità campione, per cui prima di procedere al calcolo degli errori campionari si è reso necessario il raggruppamento degli strati, che ne ha ridotto il numero a 699.

Parametri, domini e sottoclassi

L'applicazione prevede la stima dei seguenti parametri:

θ_1 = Età media alla laurea
θ_2 = Voto medio di laurea
θ_3 = % di laureati fuori corso
θ_4 = % di laureati che hanno cambiato corso di laurea
θ_5 = % di laureati che attualmente lavorano
θ_6 = % di laureati che lavorano stabilmente calcolata sui laureati che attualmente lavorano
θ_7 = % di laureati con contratto di formazione calcolata sui laureati che attualmente lavorano
θ_8 = % di laureati che lavorano saltuariamente calcolata sui laureati che attualmente lavorano
θ_9 = % di laureati che lavoravano prima della laurea calcolata sui laureati che attualmente lavorano

Le stime sono state calcolate per due domini di studio, costituiti dai seguenti raggruppamenti dei corsi di laurea:

Dominio 1 = corsi di laurea scientifici

Dominio 2 = altri corsi di laurea

e per una sola sottoclasse costituita dai laureati provenienti dai licei classici o scientifici.

File dei dati di base

Per l'esecuzione del programma non è necessario leggere tutti i campi del file CMS dei dati rilevati, ma è sufficiente limitarsi a quelli contenenti le variabili che devono essere utilizzate per ricavare le stime.

Di seguito si riporta il tracciato record limitato ai campi che devono essere letti:

Colonne	Variabili	
5-7	Gruppi di laurea	
	Scientifico	= 101-112
	medico	= 201-202
	ingegneria	= 301-316
	agrario	= 401-406
	economico	= 501-512
	politico-sociale	= 601-603
	giuridico	= 701
	letterario	= 801-817
21	Attualmente lavora	
	sì = 1	
	no = 2	
22	Tipo di lavoro	
	stabile	= 1
	contratto di formazione lavoro	= 2
	precario	= 3
	occasionale	= 4
23	Lavorava prima della laurea	
	no = 1	
	sì = 2	
24	Ha cambiato lavoro dopo la laurea	
	sì = 1	
	no = 2	
111-113	Voto di laurea	
117	Ha cambiato corso di laurea	
	sì = 1	
	no = 2	
118	Si è laureato in corso	
	sì = 1	
	no = 2	
128-129	Diploma di scuola secondaria superiore	
	maturità tecnico-professionale	= 1-5
	maturità magistrale	= 6
	maturità scientifica	= 7
	maturità classica	= 8
	maturità linguistica	= 9
	maturità artistica	= 10
	altro	= 11
173-174	Anno di nascita	
206-214	Coefficiente di espansione	
220-224	Codice di strato	

Input del programma

Un primo blocco d'istruzioni provvede a leggere i dati dal file CMS e a costruire il data set SAS contenente le variabili necessarie per le successive elaborazioni.

Vengono quindi forniti in input i seguenti valori:

numero di domini	ND = 2
numero di sottoclassi	NC = 1
numero di variabili utilizzate	NY = 11
numero di parametri da stimare	NZ = 10

Successivamente, nelle apposite MACRO, vengono inserite le istruzioni per definire:

i domini di studio (MACRO DOMINI)

la sottoclasse (MACRO CLASSI)

le variabili da utilizzare nelle trasformazioni lineari (MACRO VARY)

le trasformazioni lineari (MACRO VARZ)

Output del programma

Le MACRO ERRORI provvede al calcolo degli errori di campionamento, mentre la MACRO STAMPA predispone l'output finale, che è costituito dalle 6 tavole allegate alla fine del paragrafo.

La procedura informatica, come si è avuto modo di verificare, è semplice da utilizzare e presenta una notevole flessibilità sia rispetto alla natura dei parametri che devono essere stimati che ai domini di studio e alle sottoclassi.

Tavola 4.1 Errori di campionamento *Dominio = 0* *Classe = 0*

	STIMA	SE	RE	INF	SUP
Parametri					
N. Laureati	72124	0,00	0,00	72124	72124
Età media alla laurea	27,08	0,05	0,17	27,03	27,12
Voto medio di laurea	103,82	0,08	0,07	103,75	103,90
% Laureati fuori corso	79,56	0,46	0,58	79,10	80,03
% Hanno cambiato corso	9,61	0,32	3,37	9,28	9,93
% Laureati che lavorano	78,29	0,46	0,59	77,73	78,65
% Stabilmente	64,08	0,59	0,92	63,49	64,68
% Contratto formazione	7,17	0,33	4,54	6,85	7,50
% Saltuariamente	28,74	0,54	1,89	28,20	29,29
% Lavorano in precedenza	25,36	0,48	1,89	24,88	25,84
% Hanno cambiato lavoro	37,52	1,15	3,06	36,37	38,67

Tavola 4.2 Errori di campionamento *Dominio = 0* *Classe = 1*

	STIMA	SE	RE	INF	SUP
Parametri					
N. Laureati	52087,7	339,65	0,65	51748,1	52427,4
Età media alla laurea	26,51	0,04	0,16	26,47	26,55
Voto medio di laurea	104,30	0,09	0,08	104,21	104,39
% Laureati fuori corso	78,91	0,55	0,70	78,36	79,46
% Hanno cambiato corso	9,04	0,37	4,09	8,67	9,41
% Laureati che lavorano	76,24	0,56	0,73	75,69	76,80
% Stabilmente	60,72	0,73	1,20	59,99	61,45
% Contratto formazione	7,60	0,39	5,19	,21	8,00
% Saltuariamente	31,68	0,68	2,14	31,00	32,35
% Lavorano in precedenza	17,65	0,52	2,95	17,13	18,18
% Hanno cambiato lavoro	48,39	1,69	3,49	46,71	50,08

Tavola 4.3 Errori di campionamento *Dominio = 1* *Classe = 0*

	STIMA	SE	RE	INF	SUP
Parametri					
N. Laureati	36043	0,00	0,00	36043	36043
Età media alla laurea	27,07	0,05	0,17	27,02	27,11
Voto medio di laurea	103,43	0,11	0,10	103,32	103,54
% Laureati fuori corso	77,61	0,66	0,85	76,95	78,26
% Hanno cambiato corso	8,27	0,43	5,16	7,84	8,69
% Laureati che lavorano	77,14	0,65	0,85	76,49	77,80
% Stabilmente	60,80	0,83	1,36	59,97	61,563
% Contratto formazione	6,83	0,42	6,09	6,42	7,25
% Saltuariamente	32,37	0,78	2,42	31,58	33,15
% Lavorano in precedenza	17,61	0,59	3,36	17,02	18,20
% Hanno cambiato lavoro	50,41	1,96	3,88	48,45	52,36

Tavola 4.4 Errori di campionamento *Dominio = 1* *Classe = 1*

	STIMA	SE	RE	INF	SUP
Parametri					
N. Laureati	28104,8	226,94	0,81	27877,9	8331,8
Età media alla laurea	26,82	0,05	0,17	26,78	26,87
Voto medio di laurea	104,03	0,12	0,11	103,91	104,14
% Laureati fuori corso	76,69	0,76	0,99	75,94	77,45
% Hanno cambiato corso	7,85	0,46	5,87	7,93	8,31
% Laureati che lavorano	75,38	0,76	1,01	74,62	76,14
% Stabilmente	58,20	0,97	1,66	57,24	59,17
% Contratto formazione	7,01	0,47	6,71	6,54	7,48
% Saltuariamente	34,79	0,93	2,66	33,86	35,71
% Lavorano in precedenza	12,14	0,59	4,86	11,55	12,73
% Hanno cambiato lavoro	58,24	2,62	4,50	55,62	60,86

Tavola 4.5 Errori di campionamento *Dominio = 2* *Classe = 0*

	STIMA	SE	RE	INF	SUP
Parametri					
N. Laureati	36081	0,00	0,00	36081	36081
Età media alla laurea	27,09	0,08	0,30	27,01	27,17
Voto medio di laurea	104,22	0,11	0,11	104,11	104,33
% Laureati fuori corso	81,52	0,65	0,80	80,86	82,17
% Hanno cambiato corso	10,95	0,49	4,46	10,46	11,44
% Laureati che lavorano	79,24	0,64	0,81	78,60	79,88
% Stabilmente	67,28	0,84	1,25	66,44	68,12
% Contratto formazione	7,50	0,50	6,65	7,00	8,00
% Saltuariamente	25,22	0,75	2,98	24,47	25,97
% Lavorano in precedenza	33,10	0,75	2,27	32,35	33,86
% Hanno cambiato lavoro	30,67	1,40	4,58	29,27	32,08

Tavola 4.6 Errori di campionamento *Dominio = 2* *Classe = 1*

	STIMA	SE	RE	INF	SUP
Parametri					
N. Laureati	23982,9	252,71	1,05	23730,2	24235,6
Età media alla laurea	26,14	0,08	0,29	26,07	26,22
Voto medio di laurea	104,62	0,13	0,13	104,48	104,75
% Laureati fuori corso	81,51	0,81	0,99	80,70	82,31
% Hanno cambiato corso	10,44	0,59	5,69	9,85	11,03
% Laureati che lavorano	77,26	0,82	1,06	76,44	78,08
% Stabilmente	63,60	1,09	1,72	62,51	64,69
% Contratto formazione	8,28	0,65	7,87	7,63	8,93
% Saltuariamente	28,12	0,98	3,50	27,14	29,11
% Lavorano in precedenza	24,12	0,89	3,69	23,23	25,01
% Hanno cambiato lavoro	42,59	2,18	5,12	40,40	44,77

CAPITOLO 5 - APPLICAZIONE DELLA METODOLOGIA STANDARD AL CAMPIONAMENTO A DUE STADI

1. Premessa

Quando la popolazione oggetto d'indagine è molto ampia e le unità elementari che la costituiscono presentano una notevole dispersione territoriale, il campionamento ad uno stadio, semplice o stratificato, risulta scarsamente applicabile per una serie di ragioni di natura economica ed organizzativa.

In primo luogo non sempre si dispone di una lista attendibile delle unità elementari e la sua costruzione può comportare costi eccessivi rispetto all'economia generale dell'indagine.

Ad esempio se si vuole effettuare un'indagine compionaria sulla fecondità delle donne coniugate appartenenti ad una determinata classe di età, per costruire la lista occorre elaborare i dati delle anagrafi di tutti i comuni italiani, o di tutte le sezioni elettorali se si ricorre a quest'ultima base informativa.

In secondo luogo con la selezione diretta delle unità elementari si ha una disseminazione del campione su tutto il territorio e di conseguenza un elevato numero di aree interessate alla rilevazione, con un numero medio d'interviste per area molto ridotto. Ciò se da un lato comporta una maggiore efficienza delle stime, dall'altro fa aumentare sia le difficoltà organizzative che i costi per unità campionata.

Per superare, o quantomeno limitare, tali inconvenienti si può utilizzare il campionamento a grappolo ad uno o più stadi. Il territorio viene suddiviso in unità areali, ciascuna delle quali forma un grappolo di unità elementari. Una volta formata la lista delle unità areali si procede ad un loro campionamento semplice o stratificato, mediante una delle tecniche di selezione che verranno illustrate nei prossimi paragrafi.

Il modo in cui può essere suddiviso il territorio dipende dalla natura delle unità elementari, dal livello territoriale di disponibilità delle informazioni per la costruzione delle liste, dai vincoli organizzativi, etc. Così le unità areali possono essere costituite da sezioni di censimento, sezioni elettorali, circoscrizioni comunali, comuni, unità sanitarie locali, distretti scolastici, ecc.

Non sempre i grappoli sono formati da unità territoriali, anche se questo costituisce il caso più frequente nella pratica. Ad esempio i grappoli possono essere costituiti dalle unità in cui è organizzata l'attività lavorativa (le unità locali sono grappoli di lavoratori dipendenti) o di studio (le scuole o le facoltà universitarie sono grappoli di studenti).

Ai fini del campionamento più che la natura dei grappoli è rilevante il loro numero, la loro ampiezza e il fatto che questa risulti costante oppure variabile. Quasi sempre i grappoli coincidono con unità amministrative o con suddivisioni del territorio effettuate a fini statistici, comunque preesistenti all'indagine e generalmente di ampiezza variabile.

In ciascun grappolo selezionato la rilevazione può essere condotta sulla totalità o su un campione delle unità elementari. Nel primo caso si ha un disegno campionario a grappolo ad uno stadio o semplicemente a grappolo, nel secondo caso un campionamento a due stadi.

È ovvio, che a parità di numero di unità elementari che devono essere rilevate, il campionamento a due stadi consente di estendere la rilevazione su un numero più elevato di grappoli. La scelta tra l'uno e l'altro disegno campionario va effettuata tenendo conto di come la variabilità dei caratteri oggetto d'indagine si scompone tra ed entro i grappoli e dei costi di rilevazione sia delle unità elementari che dei grappoli.

Ai fini degli sviluppi teorici si farà riferimento al solo campionamento a due stadi, che comprende come caso particolare anche il campionamento a grappolo.

2. Il campionamento a due stadi: generalità e simbologia

Si indichi con M il numero delle unità elementari della popolazione e con N il numero dei grappoli in cui sono state raggruppate. I grappoli costituiscono le unità di primo stadio (PSU), quelle elementari le unità di secondo stadio o finali.

Stratificazione delle PSU

Come per il campionamento ad uno stadio anche per quello a due stadi è frequente il ricorso alla stratificazione, sia per ridurre la varianza campionaria degli stimatori che per programmare il campione in corrispondenza di particolari domini di studio.

Così per le indagini campionarie sulle famiglie usualmente viene utilizzato un disegno campionario a due stadi con stratificazione delle unità di primo stadio.

Le PSU, che sono costituite dai comuni, vengono stratificate per provincia o per regione a seconda del livello territoriale cui devono essere riferite le stime, e sulla base di una o più variabili correlate con i caratteri oggetto di rilevazione.

In genere come variabile di stratificazione viene utilizzata l'ampiezza demografica che, oltre ad essere un buon indicatore sintetico delle caratteristiche socio-demografiche dei comuni, consente di tenere sotto controllo in ciascun strato sia l'ammontare totale della popolazione che la variabilità delle dimensioni dei comuni (cfr. Zannella, 1989).

Sia L il numero degli strati in cui sono state raggruppate le

PSU e con riferimento al generico strato h (h = 1, 2, . . . , L) si indichi con:

- N_h = numero di PSU
- i = indice di PSU (i = 1, 2, . . . , N_h)
- M_{hi} = numero di unità elementari nell'ima PSU
- M_h = numero di unità elementari nello strato
- j = indice di unità elementare (j = 1, 2, . . . , M_{hi})
- $Y_{(r)hij}$ = valore che l'i.ma variabile di rilevazione assume nella j-ma unità elementare dell'i.ma PSU
- $Y_{(r)hi}$ = ammontare totale dell'r.ma variabile nell'i-ma PSU
- $Y_{(r)h}$ = ammontare totale dell'r.ma variabile nello strato

Valgono le seguenti relazioni:

$$M_h = \sum_{i=1}^{n_h} M_{hi} \qquad [5.1]$$

$$Y_{(r)hi} = \sum_{j=1}^{M_{hi}} Y_{(r)hij} \qquad [5.2]$$

$$Y_{(r)h} = \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{(r)hij} = \sum_{i=1}^{N_h} Y_{(r)hi} \qquad [5.3]$$

L'ammontare totale della variabile r.ma nella popolazione è dato da:

$$Y_{(r)} = \sum_{h=1}^{L} Y_{(r)h} \qquad [5.4]$$

Tecniche di selezione delle PSU

Nello strato h vengono selezionate n_h PSU campione e in ciascuna di esse si effettua un campionamento casuale semplice senza reimmissione e con uguale probabilità delle unità elementari. A seconda delle modalità di selezione delle PSU si possono avere i seguenti disegni campionari:

- con reimmissione e uguale probabilità di selezione
- con reimmissione e probabilità di selezione variabile
- senza reimmissione e uguale probabilità d'inclusione
- senza reimmissione e probabilità d'inclusione variabile

I disegni campionari con probabilità variabile, con o senza reimmissione, trovano applicazione quando all'interno di ciascu-

no strato i grappoli presentano dimensioni diverse. Essi risultano più efficienti dei corrispondenti disegni campionari con probabilità uguali e l'efficienza è tanto maggiore quanto più elevata è la correlazione tra le variabili oggetto d'indagine e le probabilità di selezione o d'inclusione.

Per quanto riguarda la scelta tra il campionamento con o senza reimmissione valgono le stesse considerazioni svolte per il campionamento ad uno stadio: la selezione con reimmissione è meno efficiente ma conduce ad espressioni più semplici dello stimatore della varianza campionaria.

Nei prossimi paragrafi verranno sviluppati in modo dettagliato i due disegni campionari con probabilità variabile.

Le espressioni degli stimatori della varianza campionaria nel caso di probabilità uguali verranno *derivate* da quelle del corrispondente campionamento con probabilità variabile, del quale rappresentano un caso particolare.

3. Campionamento delle PSU con reimmissione

Nello strato h (h = 1, 2, . . . , L) vengono estratte n_h PSU campione con reimmissione e con probabilità di selezione variabile. Dall'i.ma PSU campione vengono estratte m_{hi} unità elementari mediante un campionamento casuale semplice senza reimmissione.

Stimatore di un totale

Indicando con p_{hi} la probabilità di selezione dell'i.ma PSU campionata e con $y_{(r)hij}$ il valore dell'r.ma variabile osservato sulla j.ma unità elementare rilevata, si ha:

Stimatore del totale nell'i.ma PSU

$$\hat{Y}_{(r)hi} = \sum_{j=1}^{m_h} \frac{M_{hi}}{m_{hi}} y_{(r)hij} \qquad [5.5]$$

Stimatore del totale nello strato h

$$\hat{Y}_{(r)h} = \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\hat{Y}_{(r)hi}}{p_{hi}} = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}\, y_{(r)hij} \qquad [5.6]$$

dove il coefficiente di espansione w_{hij} è dato da:

$$w_{hij} = \frac{M_{hi}}{n_h\, p_{hi}\, m_{hi}} \qquad [5.7]$$

Stimatore del totale nella popolazione

$$\hat{Y}_{(r)} = \sum_{h=1}^{L} \hat{Y}_{(r)h} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}\, y_{(r)hij} \qquad [5.8]$$

Stimatore del totale in un dominio di studio

$$\hat{Y}_{(r)d} = \sum_{h \in d} \hat{Y}_{(r)} = \sum_{h \in d} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}\, y_{(r)hij} \qquad [5.9]$$

Stimatore della varianza campionaria

Uno stimatore non distorto della varianza campionaria dello stimatore del totale nello strato h è dato da (cfr. Cochran, 1977 pag. 307):

$$\hat{Var}\left(\hat{Y}_{(r)h}\right) = \frac{1}{n_h(n_h - 1)} \sum_{i=1}^{n_h} \left(\frac{\hat{Y}_{(r)hi}}{p_{hi}} - \hat{Y}_h \right)^2 \qquad [5.10]$$

da cui si ricavano le espressioni degli stimatori della varianza campionaria dello stimatore di un totale nell'intera popolazione e in un dominio di studio:

$$\hat{Var}\left(\hat{Y}_{(r)}\right) = \sum_{h=1}^{L} \hat{Var}\left(\hat{Y}_{(r)h}\right) \qquad [5.11]$$

e

$$\hat{Var}\left(\hat{Y}_{(r)d}\right) = \sum_{h \in d} \hat{Var}\left(\hat{Y}_{(r)h}\right) \qquad [5.12]$$

Per calcolare la varianza campionaria dello stimatore di un totale riferito ad una sottoclasse si applicano, le formule precedenti dopo aver introdotto le nuove variabili $y'_{(r)hij}$, dove $y'_{(r)hij} = y_{(r)hij}$ se l'unità elementare appartiene alla sottoclasse e $y'_{(r)hij} = 0$ nel caso contrario.

Per gli stimatori non lineari, prima di applicare le formule relative alla stima di un totale, si dovrà procedere alla loro linearizzazione utilizzando le espressioni riportate nel secondo capitolo.

Campionamento autoponderante

Le formule [5.6], [5.8] e [5.9] mettono in evidenza che gli stimatori del totale negli strati, nella popolazione e nei domini di studio, si ottengono come somma ponderata dei valori osservati sulle unità elementari, dove i pesi sono costituiti dai coefficienti di espansione.

L'introduzione di coefficienti di espansione variabile, oltre ad appesantire le elaborazioni, fa aumentare la varianza campionaria degli stimatori (Kish, 1965). È pertanto preferibile, quando è possibile ricorrere ad un campionamento autoponderante, in cui ogni unità elementare ha lo stesso coefficiente di espansione.

Nel caso, molto frequente nella pratica, in cui le probabilità di selezione sono proporzionali all'ampiezza dei grappoli: $p_{hi} = M_{hi}/M_h$, ciò può essere ottenuto adottando i seguenti criteri per la ripartizione della numerosità del campione di secondo stadio:

a) le m unità elementari del campione totale vengono ripartite tra gli strati proporzionalmente all'ampiezza di ciascuno strato: $m_h = m\ M_h/M$
b) a ciascuna PSU campione dello strato h viene assegnato lo stesso numero di unità elementari da intervistare: $m_{hi} = m_h/n_h$

In questo caso a ciascuna unità elementarie viene assegnato un coefficiente di espansione costante ed uguale a M/m.

Campionamento con probabilità uguali

Le formule fin quì riportate rimangono valide anche nel caso in cui le PSU appartenenti allo stesso strato vengono estratte con uguale probabilità. È sufficiente effettuare nelle precedenti espressioni le seguenti sostituzioni:

$$p_{hi} = \frac{1}{N_h} \qquad [5.13]$$

e

$$w_{hi} = \frac{N_h}{n_h} \cdot \frac{M_{hi}}{m_{hi}} = \frac{1}{f_{1h}} \cdot \frac{1}{f_{2hi}} \qquad [5.14]$$

dove le quantità $f_{1h} = n_h/N_h$ e $f_{2hi} = m_{hi}/M_{hi}$ rappresentano i tassi di campionamento rispettivamente di primo e di secondo stadio.

4. Procedura informatica per il calcolo degli errori campionari nel compionamento con reimmissione

Input del programma

Anche per questo disegno campionario è stato predisposta un'apposita procedura informatica per la stima degli errori campionari. La precedura che utilizza istruzioni SAS prevede come input:

1) Il file degli strati formato da tanti records quanti sono gli strati in cui sono state raggruppate le PSU. Ogni record contiene il codice di strato, il numero di PSU universo e campione, il numero di unità elementari universo e campione;

2) Il file delle PSU formato da tanti records quante sono le unità di primo stadio campionate. Per ogni PSU campione sono riportati i codici di strato e di PSU, il numero di unità elementari universo e campione, la probabilità di selezione;

3) Il file delle unità elementari costituiscono da tanti records quante sono le unità elementari rilevate, Per ogni unità elementare sono riportati i codici di strato, di PSU e di unità elementare e i valori delle variabili oggetto dell'indagine.

Deve inoltre essere imputato il numero di parametri per i quali è richiesta la stima degli errori campionari (NZ), il numero delle variabili che vengono utilizzate per i calcoli (NY), il numero delle sottoclassi (NC) e dei domini di studio (ND).

Costruzione del file dei dati di base

Mediante un'operazione di merge fra i tre files il programma procede alla formazione di un unico file, in cui per ogni unità elementare sono riportate anche le informazioni contenute nel file degli strati e in quello delle PSU. Quindi attribuisce ad ogni unità elementare il coefficiente di espansione calcolato mediante la [5.7].

Il programma provvede all'introduzione delle variabili «DOMINIO» e «CLASSE» e delle variabili indicatrici delle modalità nel caso di stime di frequenze.

Se sono previsti stimatori non lineari il programma procede alla costruzione di nuove variabili mediante le trasformazioni lineari riportate nel secondo capitolo e nel caso di sottoclassi costruisce le variabili con apice.

Il file così ottenuto, denominato file dei dati di base, contiene per ogni unità elementare oltre ai codici identificativi (strato, PSU, unità elementare) le seguenti altre informazioni: numero delle PSU universo (N_h) e campione (n_h); numero delle unità elementari universo (M_h) e campione (m_h) dello strato; numero delle unità elementari universo (M_{hi}) e campione (m_{hi}) e probabilità di selezione (P_{hi}) della PSU; coefficiente di espansione (w_{hij}) e valori delle variabili dopo eventuale trasformazione ($z_{(r)hij}$) dell'unità elementare; codice di dominio e di sottoclasse.

Stima dei parametri e delle varianze campionarie

Mediante una procedura SUMMARY viene costruito un file contenente per ogni PSU i codici identificativi (strato e PSU), le stime $\hat{Z}_{(r)hi}$ dei totali, date dalla [5.5] e le quantità $\hat{Z}_{(r)hi}/p_{hi}$, che costituiscono stime indipendenti dell'ammontare totale nello strato.

È facile verificare che la stima $\hat{Z}_{(r)h}$ del totale nello strato h è uguale alla media aritmetica semplice degli n_h valori $\hat{Z}_{(r)hi}/p_{hi}$ e che la sua varianza campionaria è data dalla varianza corretta di questi valori divisa per n_h. Pertanto con una successiva procedura SUMMARY viene formato un file contenente per ogni strato le stime $\hat{Z}_{(r)h}$ e le stime delle corrispondenti varianze campionarie.

Le stime per l'intera popolazione e per i domini di studio si ottengono come somma delle stime degli strati.

Output del programma

L'output è costituito da (MD+1) (NC+1) tavole, una per ciascun dominio di studio e sottoclasse, contenenti per ogni parametro: la denominazione, la stima, la stima dell'errore standard assoluto e percentuale, gli estremi dell'intervallo di confidenza al 95%.

Esempio 5.1 Si vuole stimare l'ammontare totale di un carattere in una popolazione costituita da M = 24000 unità elementari raggruppate in 19 PSU di ampiezza variabile da 300 a 3200. A tale scopo si utilizza un disegno campionario autoponderante a due stadi con stratificazione delle unità di primo stadio.

Le PSU sono state stratificate in tre classi di ampiezza: fino a 1000, 1001-2000 e oltre 2000, ottenendo:

Strato 1		Strato 2		Strato 3	
PSU	M_{1i}	PSU	M_{2i}	PSU	M_{31}
1	300	1	1300	1	2400
2	360	2	1500	2	2400
3	480	3	1500	3	3200
4	510	4	1800		
5	570	5	1900		
6	600	6	2000		
7	720				
8	780				
9	810				
10	870				

Da ciascuno strato vengono selezionate 2 PSU con reimmissione e con probabilità di selezione proporzionale all'ampiezza. In totale vengono campionate 24 unità elementari, con un tasso di campionamento costante negli strati ed uguale a 1/1000, ottenendo:

Tavola 5.1 File degli strati

Strati	N_h	M_h	n_h	m_h
1	10	6000	2	6
2	6	10000	2	10
3	3	8000	2	8
Totale	19	24000	6	24

Effettuata l'estrazione si hanno le seguenti PSU campione: nel primo strato la 3ª e la 6ª, nel secondo strato la 2ª e la 4ª, nel terzo strato la 1ª e la 3ª.

Per ottenere un campione autoponderante alle due PSU estratte in ciascuno strato viene assegnato lo stesso numero di unità elementari campione.

Nel prospetto che segue per ciascuna PSU selezionata è riportato il numero delle unità elementari universo e campione e la probabilità di selezione.

Tavola 5.2 File delle unità di primo stadio

Strati	PSU	M_{hi}	p_{hi}	m_{hi}
1	3	480	0,08	3
1	6	600	0,10	3
2	2	1500	0,15	5
2	4	1800	0,18	5
3	1	2400	0,30	4
3	3	3200	0,40	4

dove $p_{hi} = M_{hi}/M_h$.

Si può immediatamente verificare che ogni unità elementare ha la stessa probabilità di essere estratta, uguale a 1/1000 e lo stesso coefficiente di ponderazione $w_{hij} = 1000$.

Una volta estratte le unità elementari ed effettuata la rilevazione, si dovrà predisporre un file contenente per ciascuna unità elementare i codici identificativi (strato, PSU, unità elementare), il coefficiente di ponderazione e i valori delle variabili.

Tavola 5.3 File delle unità elementari

Strato	PSU campione	unità elementare	y_{hij}	w_{hij}
1	3	37	3,0	1000
1	3	152	3,4	1000
1	3	318	2,6	1000
1	6	214	3,6	1000
1	6	397	3,5	1000
1	6	412	2,9	1000
2	2	97	3,7	1000
2	2	112	2,5	1000
2	2	418	2,3	1000
2	2	1024	3,0	1000
2	2	1397	2,5	1000
2	4	615	2,6	1000
2	4	837	3,8	1000
2	4	1114	3,5	1000
2	4	1536	2,3	1000
2	4	1722	3,8	1000
3	1	228	2,8	1000
3	1	754	3,8	1000
3	1	1618	3,6	1000
3	1	2312	3,0	1000
3	3	874	4,4	1000
3	3	1132	3,2	1000
3	3	2018	4,3	1000
3	3	2937	3,1	1000

Unendo i tre files si avrà un nuovo file formato da quello delle unità elementari a cui sono state aggiunte le seguenti altre variabili: N_h, M_h, n_h, m_h, M_{hi}, p_{hi}, m_{hi}.

Poiché è richiesta la stima di un totale e non sono previste sottoclassi non occorre effettuare trasformazioni di variabili, pertanto il file dei dati di base coincide con quello generato precedentemente ($z_{hij} = y_{hij}$).

All'interno di ciascuna PSU vengono sommati i valori z_{hij} e z_{hij}/p_{hij} ponderati con pesi uguali a M_{hi}/m_{hi}, ottenendo il seguente file:

Tavola 5.4 File delle stime per PSU

Strato	PSU	$\hat{Z}_{hi}$	$\hat{Z}_{h1}/p_{hi}$
1	3	1440	18000
1	6	2000	20000
2	2	4200	28000
2	4	5760	32000
3	1	7800	26000
3	3	12000	30000

All'interno di ciascuno strato vengono calcolate le medie aritmetiche e le varianze corrette dei valori $\hat{Z}_{hi}/p_{hi}$. Questi ultimi valori vengono divisi per n_h ottenendo così la stima della varianza campionaria:

Tavola 5.5 File delle stime per strato

Strato	$\hat{Z}_h$	Var ($\hat{Z}_h$)
1	19000	1000000
2	30000	4000000
3	28000	4000000

Poiché ND = O e NC = O, l'output è costituito dalla sola tavola relativa alla popolazione totale:

Tavola 5.6 Stime ed errori standard delle stime

	dominio = O		sottoclasse = O		
Parametri	Stima	SE	RE	INF	SUP
θ_1	77.000	3.000	3,9	71.000	83.000

5. Campionamento delle PSU senza reimmissione

Nello strato h (h = 1, 2, . . . , L) vengono estratte senza reimmissione e con probabilità d'inclusione variabile n_h PSU e dall'i.ma PSU campione (i = 1, 2, . . . , n_h) vengono selezionate

m_{hi} unità elementari mediante un campioramento casuale semplice senza reimmissione e con uguale probabilità.

La stima del totale dell'i.ma PSU campione è dato, come per il compionamento con reimmissione, da:

$$\hat{Y}_{(r)hi} = \sum_{j=1}^{m_{hi}} \frac{M_{hi}}{m_{hi}} \, y_{(r)hij} \qquad [5.15]$$

Stimatore del totale in uno strato

Per stimare il totale nello strato h si applica lo stimatore di Horvitz-Thompson:

$$\hat{Y}_{(r)h} = \sum_{i=1}^{n_h} \frac{\hat{Y}_{(r)hi}}{\pi_{hi}} \qquad [5.16]$$

dove π_{hi} è la probabilità d'inclusione del primo ordine dell'i.ma PSU dello strato h. Indicando con p_{hi} la probabilità di selezione si ha:

$$\pi_{hi} = n_h \, p_{hi} \qquad [5.17]$$

La [5.16] può anche essere scritta:

$$\hat{Y}_{(r)h} = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} \, y_{(r)hij} \qquad [5.18]$$

dove i coefficienti di espansione sono dati da:

$$w_{hij} = \frac{M_{hi}}{m_{hi} \, \pi_{hi}} \qquad [5.19]$$

Campione autoponderante

Se le probabilità d'inclusione sono proporzionali all'ampiezza volgono le seguenti relazioni:

$$\pi_{hi} = n_h \, \frac{M_{hi}}{M_h} \qquad [5.20]$$

e

$$w_{hij} = \frac{M_h}{n_h \, m_{hi}} \qquad [5.21]$$

In questo caso per ottenere un campione autoponderante occorre che il numero delle unità elementari campione sia riparti-

to tra gli strati proporzionalmente alla loro ampiezza e che all'interno di ciascuno strato ad ogni PSU campione sia assegnato lo stesso numero di unità elementari da rilevare. È facile verificare che il coefficiente di espansione risulta costante ed uguale al reciproco del tasso di campionamento totale (M/m).

Lo stimatore della varianza campionaria di $\hat{Y}_{(r)h}$ è dato dalla somma di due componenti, attribuibili rispettivamente al primo e al secondo stadio di campionamento:

$$\hat{Var}\left(\hat{Y}_{(r)h}\right) = \hat{V}_1\left(\hat{Y}_{(r)h}\right) + \hat{V}_2\left(\hat{y}_{(r)h}\right) \qquad [5.22]$$

Indicando con π_{hik} la probabilità d'inclusione del secondo ordine dell'i.ma e della k.ma PSU campione dello strato h, valgono le seguenti relazioni (cfr. Wolter, 1985, pag. 15):

$$\hat{V}_1\left(\hat{Y}_{(r)h}\right) = \sum_{i=1}^{n_h} \sum_{k>i} \left(\frac{\hat{Y}_{(r)hi}}{\pi_{hi}} - \frac{\hat{Y}_{(r)hk}}{\pi_{hk}} \right)^2 \frac{\pi_{hi}\,\pi_{hk} - \pi_{hik}}{\pi_{hik}} \qquad [5.23]$$

e

$$\hat{V}_2\left(\hat{Y}_{(r)h}\right) = \sum_{i=1}^{n_h} \frac{M_{hi}(M_{hi} - m_{hi})}{m_{hi}\;\pi_{hi}}\;\hat{S}^2_{2(r)hi} \qquad [5.24]$$

dove $\hat{S}^2_{2(r)hi}$ è lo stimatore corretto della varianza delle unità elementari nell'i.ma PSU campione dello strato h:

$$\hat{S}^2_{2(r)hi} = \frac{1}{m_{hi} - 1} \sum_{j=1}^{m_{hi}} \left(y_{(r)hij} - \hat{\bar{Y}}_{(r)hi} \right)^2 \qquad [5.25]$$

e

$$\hat{\bar{Y}}_{(r)hi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{(r)hij} \qquad [5.26]$$

Stimatore del totale nella popolazione

Lo stimatore $\hat{Y}_{(r)}$ del totale per l'intera popolazione è dato dalla somma degli stimatori dei totali dei singoli strati, e lo stimatore della sua varianza campionaria è dato da:

$$\hat{Var}\left(\hat{Y}_{(r)}\right) = \hat{V}_1\left(\hat{Y}_{(r)}\right) + \hat{V}_2\left(\hat{Y}_{(r)}\right) \qquad [5.27]$$

dove:

$$\hat{V}_1\left(\hat{Y}_{(r)}\right) = \sum_{h=1}^{L} \hat{V}_1\left(\hat{Y}_{(r)h}\right) \qquad [5.28]$$

e

$$\hat{V}_2\left(\hat{Y}_{(r)}\right) = \sum_{h=1}^{L} \hat{V}_2\left(\hat{Y}_{(r)h}\right) \qquad [5.29]$$

Espressioni analoghe si hanno per lo stimatore del totale in un dominio di studio. Per gli stimatori non lineari e per stimatori relativi a sottoclassi valgono le stesse considerazioni svolte per il campionamento con reimmissione.

Estrazione di due PSU per strato

Dalla [5.23] si evince che la stima della varianza campionaria di primo stadio richiede il calcolo delle probabilità d'inclusione del secondo ordine per tutte le coppie di PSU campionate in ciascuno strato.

Esistono numerosi metodi per la selezione delle PSU senza reimmissione e con probabilità d'inclusione variabile (cfr. ad es. Brewer e Hanif, 1987), per alcuni dei quali si dispone delle espressioni esplicite per il calcolo delle π_{hik} mentre per altri è necessario utilizzare procedimenti iterativi.

In genere questo tipo di campionamento viene applicato per $n_h = 2$ e il calcolo delle probabilità d'inclusione viene così ad essere limitato ad una sola coppia di PSU per strato. Lo stimatore della varianza di primo stadio è dato da:

$$\hat{V}_1\left(\hat{Y}_{(r)h}\right) = \left(\frac{\hat{Y}_{(r)h1}}{\pi_{h1}} - \frac{\hat{Y}_{(r)h2}}{\pi_{h2}} \right)^2 \frac{\pi_{h1}\,\pi_{h2} - \pi_{h12}}{\pi_{h12}} \qquad [5.30]$$

dove sono stati utilizzati gli indici 1 e 2 per indicare rispettivamente la prima e la seconda PSU estratte nello strato.

Un metodo di selezione, per il quale si hanno le soluzioni esplicite per calcolare le probabilità d'inclusione del secondo ordine è quello dovuto a Brewer (1963) e consiste nell'estrarre la prima PSU con probabilità:

$$p_{hi} = \frac{M_{hi}\;(M_h - M_{hi})}{M_h\;(M_h - 2M_{hi})} \qquad [5.31]$$

e la seconda con probabilità:

$$p_{hk} = \frac{M_{hk}}{M_h - M_{hi}} \qquad [5.32]$$

dove con i è stato indicato il numero d'ordine della PSU estratta per prima.

Si dimostra che le probabilità d'inclusione del primo ordine sono proporzionali all'ampiezza:

$$\pi_{hik} = \frac{2M_{hi}}{M_h} \qquad [5.33]$$

e che le probabilità d'inclusione del secondo ordine sono date da:

$$\pi_{hik} = \frac{2\ M_{hi}\ M_{hk}}{M_h^2\ D_h} \cdot \frac{M_h - M_{hi} - M_{hk}}{(M_h - M_{hi})(M_h - M_{hk})} \qquad [5.34]$$

dove:

$$D_h = \frac{1}{2}\left(1 + \sum_{i=1}^{N_h} \frac{M_{hi}}{M_h - 2M_{hi}}\right) \qquad [5.35]$$

Le espressioni che danno le probabilità d'inclusione del secondo ordine risultano molto più semplici quando le PSU appartenenti allo stesso strato vengono estratte con probabilità uguali. In questo caso la prima PSU campione viene estratta con probabilità $1/N_h$, la seconda con probabilità $1/(N_h-1)$ e così via. Le probabilità d'inclusione del primo e del secondo ordine sono date rispettivamente da:

$$\pi_{hi} = \frac{n_h}{N_h} \qquad [5.36]$$

e

$$\pi_{hij} = \frac{n_h(n_h - 1)}{N_h(N_h - 1)} \qquad [5.37]$$

Come si verifica immediatamente sostituendo la [5.36] nella [5.21] il coefficiente di espansione risulta uguale a:

$$w_{hij} = \frac{N_h}{n_h} \cdot \frac{M_{hi}}{m_{mhi}} = \frac{1}{f_{1h}} \cdot \frac{1}{f_{2hi}} \qquad [5.38]$$

dove f_{1h} e f_{2hi} sono i tassi di campionamento di primo e di secondo stadio.

Le stime delle varianze campionarie di primo e di secondo stadio si semplificano nel sequente modo:

$$\hat{V}_1\left(\hat{Y}_{(r)h}\right) = \frac{(1 - f_{1h})\ N_h^2}{n_h(n_h - 1)} \sum_{i=1}^{n_h} \left(\hat{Y}_{(r)hi} - \frac{\hat{Y}_{(r)h}}{N_h}\right)^2 \qquad [5.39]$$

e

$$\hat{V}_2\left(\hat{Y}_{(r)h}\right) = \frac{N_h}{n_h} \sum_{i=1}^{n_h} M_{hi}^2\ \frac{1 - f_{2hi}}{m_{hi}}\ \hat{S}_{2(r)hi} \qquad [5.40]$$

6. Procedura informatica per il calcolo degli errori campionari nel campionamento senza reimmissione

La procedura informatica che è stata predisposta fa riferimento al caso più generale in cui vengono estratte 2 o più PSU per strato con probabilità d'inclusione variabile, e può essere applicata anche per il campionamento con probabilità uguali.

Input del programma

L'input è costituito dal file degli strati, dal file delle PSU e da quello delle unità elementari.

Il file degli strati è formato da tanti records quanti sono gli strati in cui sono state raggruppare le PSU e ciascun record contiene: il codice dello strato, il numero di PSU universo (N_h) e campione (n_h), il numero delle unità elementari universo (M_h) e campione (m_h).

Il file delle PSU è formato da tanti records quante sono le PSU campione, in ciascuno dei quali sono riportati: i codici identificativi (strato e PSU), la probabilità d'inclusione del primo ordine (π_{hi}), le probabilità d'inclusione del 2° ordine (π_{hik}) tra la PSU cui il record si riferisce e tutte le altre PSU campione dello strato, il numero delle unità elementari universo (M_{hi}) e campione (m_{hi}).

Il file delle unità elementari è costituito da tanti records quante sono le unità elementari campionate. Ciascun record

contiene: i codici identificativi (strato, PSU, unità elementare) e i valori delle variabili rilevate.

Come negli altri programmi che sono stati predisposti deve essere inputato il numero dei parametri per i quali è richiesta la stima (NZ), il numero delle variabili che vengono utilizzate per i calcoli (NY), il numero di domini di studio (ND) e delle sottoclassi (NC).

Stima dei parametri e delle varianze campionarie

Il programma, dopo aver effettuato il merge fra i tre files input, provvede al calcolo del coefficiente di espansione, alla costruzione delle variabili «DOMINIO» e «CLASSE» e delle variabili indicatrici per la stima di frequenze, ad effettuare le trasformazioni lineari nel caso di stimatori non lineari e alla costruzione delle variabili con apice per le stime relative a sottoclassi.

Mediante una procedura SUMMARY viene generato il file delle stime a livello di PSU, nel quale per ogni PSU campione sono riportate: le stime delle variabili trasformate ($\hat{Z}_{(r)hi}$), le stime delle varianze tra le unità elementari ($\hat{S}^2_{2(r)hi}$) e tutte le informazioni contenute nel file input delle PSU.

Quindi per ogni PSU campione vengono calcolate le seguenti quantità:

$$A_{hi} = \sum_{k>i} \left(\frac{\hat{Y}_{(r)hi}}{\pi_{hi}} - \frac{\hat{Y}_{(r)hk}}{\pi_{hk}} \right)^2 \frac{\pi_{hi}\,\pi_{hk} - \pi_{hik}}{\pi_{hik}} \qquad [5.41]$$

e

$$B_{hi} = \frac{M_{hi}(M_{hi} - m_{hi})}{m_{hi}\,\pi_{hi}} \; \hat{S}^2_{2(r)hi} \qquad [5.42]$$

Una successiva procedura SUMMARY forma il file delle stime a livello di strato, contenente per ciascuna sottoclasse le stime dei parametri e le stime delle varianze campionarie di primo stadio, di secondo stadio e totali.

Output del programma

L'output è costituito da una serie di (ND+1) (NC+1) tavole, una per dominio di studio e sottoclasse, contenenti per ciascun parametro: la denominazione del parametro, la stima, l'errore standard assoluto e percentuale e gli estremi dell'intervallo di confidenza al 95%. Una seconda serie, sempre di (MD+1) (MC+1) tavole, riporta per ciascun parametro la stima della varianza campionaria totale e la sua scomposizione nei due stadi di campionamento.

Esempio 5.2 Con riferimento ai dati dell'esempio 5.1 si supponga che le PSU siano state estratte senza reimmissione e con probabilità d'inclusione proporzionale all'ampiezza. In ciascun strato le due PSU campione sono state selezionate con il metodo di Brewer illustrato nel precedente paragrafo.

Il numero delle unità elementari da campionare è stato ripartito fra gli strati proporzionalmente all'ampiezza e in ciascuno strato alle due PSU selezionate è stato assegnato lo stesso numero di unità elementari campione.

I files input degli strati e delle unità elementari sono gli stessi riportati rispettivamente nelle tavole 5.1 e 5.3. Il file input delle PSU campione è dato da:

Tavola 5.7 File delle PSU

Strati	PSU	M_{hi}	m_{hi}	π_{hi}	π_{hi1}	π_{hi2}
1	3	480	3	0,16	0,0000	0,0172
1	6	600	3	0,20	0,0172	0,0000
2	2	1.500	5	0,30	0,0000	0,0069
2	4	1.800	5	0,36	0,0699	0,0000
3	1	2.400	4	0,60	0,0000	0,4000
3	3	3.200	4	0,80	0,4000	0,0000

Dopo il merge dei tre files input e le successive elaborazioni si ottiene il seguente file:

Tavola 5.8 File delle stime a livello di PSU

Strati	PSU	Z_{hi}	$\hat{S}^2_{2hi}$	$\hat{Z}_{hi}/\pi_{hi}$	A_{hi}	B_{hi}
1	3	1.440	0,16	9.000	435.650	76.320
1	6	2.000	0,14	10.000	435.650	85.580
2	2	4.200	0,32	14.000	1.090.200	478.400
2	4	5.760	0,50	16.000	1.090.200	897.500
3	1	7.800	0,23	13.000	400.000	551.084
3	3	12.000	0,48	15.000	400.000	1.534.084

Mediante la procedura SUMMARY si ricava il file delle stime a livello di strato:

Tavola 5.9 File delle stime a livello di strato

Strati	$\hat{Z}$	$\hat{V}_1(\hat{Z}_h)$	$\hat{V}_2(\hat{Z}_h)$	$\hat{V}ar(\hat{Z}_h)$
1	19.000	871.300	159.000	1.031.200
2	30.000	2.180.400	1.375.900	3.556.300
3	28.000	800.000	2.085.160	2.885.160

Poiché non sono previsti né domini di studio (ND = 0) né sottoclassi (NC = 0), l'output è costituito da due sole tavole:

Tavola 5.10 Stime ed errori standard

	dominio = 0		sottoclasse = 0		
Parametri	Stima	SE	RE	INF	SUP
θ1	77.000	2.734	3,55	71.532	82.468

Tavola 5.11 Scomposizione della varianza campionaria

	dominio = 0		sottoclasse = 0	
Parametri	1° stadio	2° stadio	totale	% 1° stadio
θ1	3.851.700	3.620.960	7.472.660	51,5

CAPITOLO 6 – LA METODOLOGIA BASATA SULLE REPLICAZIONI DEL CAMPIONE

1. Premessa

Come si è visto nei precedenti tre capitoli, l'utilizzazione della metodologia standard per la stima della varianza campionaria comporta l'adozione di formule specifiche per ogni piano di campionamento. Inoltre, per gli stimatori non lineari tali formule risultano approssimate essendo basate sullo sviluppo in serie di Taylor dello stimatore arrestato al termine di primo grado.

Una tecnica alternativa che consente la stima della varianza campionaria sia di stimatori lineari che non lineari, senza dover ricorrere per questi ultimi ad approssimazioni, è quella basata sulle replicazioni del campione.

Essa comprende diversi metodi (i gruppi casuali, le replicazioni bilanciate ripetute, il jackknife e il bootstrap) che consistono nel formare, seguendo opportune regole, un certo numero di subcampioni con le unità del campione rilevato e nel calcolare per ognuno di essi la stima del parametro d'interesse.

La variabilità delle stime dei subcampioni viene utilizzata per calcolare la varianza campionaria dello stimatore derivato dal campione totale.

In questo capitolo verranno descritti i due metodi che hanno trovato finora più larga applicazione nel campionamento da popolazioni finite, e precisamente:

a) il metodo dei gruppi casuali
b) il metodo delle replicazioni bilanciate ripetute.

2. Il metodo dei gruppi casuali

Il metodo dei gruppi casuali costituisce il fondamento di tutti i procedimenti per il calcolo degli errori di campionamento che utilizzano le replicazioni del campione. Esso consiste nel dividere il campione totale in un certo numero di subcampioni indipendenti, ciascuno ricavato mediante lo stesso disegno che ha generato il campione complessivo, e calcolare la stima di interesse (media, totale, ecc.) per ciascun subcampione.

Ogni subcampione fornisce, quindi, una stima indipendente del parametro della popolazione e la varianza fra queste stime dà una misura della varianza campionaria della stima totale.

Questo metodo è stato introdotto da Mahalanobis (1944) che ha chiamato i subcampioni «campioni interpenetranti» e

successivamente ripreso dalla Sottocommissione dell'ONU per il Campionamento Statistico (1949) con il nome di «campioni replicati».

Il termine «gruppi casuali», con il quale il metodo viene attualmente indicato, è stato utilizzato per la prima volta da Hansen, Hurwitz e Madow (1953) nel loro manuale sul campionamento statistico.

Il metodo dei gruppi casuali verrà illustrato con riferimento ad un campionamento casuale semplice con reimmissione, e successivamente verranno descritte le modalità da seguire per adattare il procedimento ai principali disegni campionari utilizzati nella pratica.

La tecnica di base

Da una popolazione di N unità viene scelto, mediante estrazione casuale semplice con reimmissione, un campione di numerosità n. Si vuole calcolare la varianza campionaria mediante il metodo dei gruppi casuali, utilizzando k subcampioni indipendenti.

Se n è un multiplo di k, ossia n = mk, con m e k interi, si avranno k gruppi di m unità ciascuno e il primo gruppo casuale si otterrà estraendo senza reimmissione m unità dalle n che costituiscono il campione totale; il secondo gruppo estraendo m unità dalle rimanenti n – m, e così via. Se n/k non è un intero, ossia n = km + r (O<r<k), si formeranno (k-r) gruppi di ampiezza m e r di ampiezza (m + 1).

Così per n = 1320 e k = 25 il rapporto m = 1320/25 = 52,4 non è un intero e n = 52·25 + 20 per cui si avranno 5 gruppi di 52 unità e 20 gruppi di 53.

Si indichi con θ il valore del parametro che deve essere stimato, con $\hat{\theta}$ lo stimatore relativo al campione totale e con $\hat{\theta}_s$ (s = 1, 2, . . ., k) gli stimatori ottenuti applicando ai k subcampioni la stessa forma funzionale di $\hat{\theta}$.

Sulla forma degli stimatori non viene posta alcuna restrizione potendo essere sia lineari che non lineari.

Per la stima di θ possono essere utilizzati due diversi stimatori: il primo costituito da $\hat{\theta}$, il secondo dalla media aritmetica semplice degli stimatori $\hat{\theta}_s$.

Si dimostra (cfr. Wolter, 1985, pp. 85-87) che i due stimatori così ottenuti coincidono nel caso in cui sono funzioni lineari dei valori osservati, mentre portano a due risultati differenti nel caso di funzioni non lineari.

Poiché nella pratica è preferibile adottare uno stesso tipo di stimatore per tutti i parametri in esame, viene comunemente utilizzato quello basato sul campione totale.

Lo stimatore della varianza campionaria con il metodo dei gruppi casuali è dato da:

$$Var\left(\hat{\theta}\right) = \frac{1}{k(k-1)} \sum_{s=1}^{k} \left(\hat{\theta}_s - \hat{\theta}\right)^2 \qquad [6.1]$$

Prima di descrivere l'adattamento del metodo a disegni campionari più complessi, è bene illustrare con un esempio numerico la sua applicazione al campionamento casuale semplice.

Esempio 6.1. Da una popolazione di N = 800 unità è stato estratto un campione casuale semplice di numerosità 20. Su ciascuna unità selezionata sono stati osservati i valori di due caratteri quantitativi ottenendo il seguente insieme di dati:

Unità	$y_{(1)i}$	$y_{(2)i}$	Unità	$y_{(1)i}$	$y_{(2)i}$
1	3	23	11	3	29
2	5	14	12	4	43
3	2	12	13	3	56
4	4	24	14	2	58
5	6	17	15	5	24
6	8	27	16	6	18
7	6	28	17	6	29
8	7	42	18	8	56
9	6	44	19	10	36
10	5	48	20	1	12

Si vogliono calcolare gli errori standard delle stime dei totali dei due caratteri e del rapporto $R = Y_{(2)}/Y_{(1)}$.

Applicando la metodologia standard per il campionamento casuale semplice, illustrata nel terzo capitolo, e ricordando che ad ogni unità è associato il coefficiente di espansione:

$$w_i = 800/20 = 40$$

si ricavano i seguenti valori:

Parametro	Stima	Stima dell'errore standard
$Y_{(1)}$	400	517,16
$Y_{(2)}$	25600	2718,42
R	6,40	0,87

Volendo calcolare l'errore standard mediante il metodo dei gruppi casuali occorre in primo luogo fissare il numero k dei subcampioni e determinare il numero m delle unità che compongono ciascuno di essi. Poiché l'applicazione del metodo richiede la disponibilità di almeno due gruppi casuali, limitandoci al caso in cui m è un intero, si hanno le seguenti possibilità:

k	m
2	10
4	5
5	4
10	2
20	1

Il numero dei gruppi casuali che viene utilizzato per la stima della varianza influisce sulla stabilità dello stimatore e, come si vedrà in seguito, è possibile, sotto certe ipotesi, determinare k in funzione della precisione desiderata nella stima della varianza campionaria.

Sia k = 5 il numero dei subcampioni in cui si è deciso di suddividere il campione totale e m = 20/5 = 4, il numero delle unità che compongono ciascuno di essi. Il primo gruppo casuale viene determinato estraendo a caso senza reimmissione 4 unità

tra le 20 campionate; il secondo gruppo estraendo con la stessa tecnica 4 unità tra le rimanenti 16 e così via. Si supponga che siano stati formati i seguenti gruppi casuali:

gruppi casuali	unità			
1	20	11	17	16
2	18	1	6	2
3	7	13	15	9
4	8	14	10	3
5	5	12	4	19

ottenendo:

Tavola 6.1 - Costruzione dei gruppi casuali

gruppi casuali	unità	$Y_{(1)i}$	$Y_{(2)i}$	w_i
1	20	1	12	200
	11	3	29	200
	17	6	29	200
	16	6	18	200
2	18	8	56	200
	1	3	23	200
	6	8	27	200
	2	5	14	200
3	7	6	28	200
	13	3	56	200
	15	5	24	200
	9	6	44	200
4	8	7	42	200
	14	2	58	200
	10	5	48	200
	3	2	12	200
5	5	6	17	200
	12	4	43	200
	4	4	24	200
	19	10	36	200

Per ciascun gruppo casuale vengono calcolate le stime dei parametri d'interesse utilizzando lo stesso procedimento di stima che viene impiegato per il campione totale, dopo aver attribuito ad ogni unità il nuovo coefficiente di espansione $w_i = 800/4 = 200$.

Per stimare i tre parametri d'interesse si possono utilizzare o le medie aritmetiche delle stime ottenute dai cinque gruppi casuali o le stime calcolate sull'intero campione.

Come si verifica immediatamente, i due metodi danno lo stesso risultato nel caso della stima di totali mentre portano a due risultati diversi per la stima del rapporto:

se si utilizza la media dei cinque gruppi

$\hat{R} = 33,10 : 5 = 6,62$

se si calcola il rapporto tra le stime dei due totali

$\hat{R} = 25600/4000 = 6,40$

Utilizzando come stime dei parametri quelle calcolate sull'intero campione, le stime delle varianze campionarie si ottengono mediante la [6.1].

Nella tav. 6.2 sono riassunti i risultati dell'applicazione del metodo dei gruppi casuali:

Tavola 6.2 - Distribuzione delle stime e calcolo della varianza campionaria

gruppi casuali	$\hat{Y}_{(1)}$	$\hat{Y}_{(2)}$	$\hat{R}$
1	3200	17600	5,50
2	4800	24000	5,00
3	2000	30400	7,60
4	3200	32000	10,00
5	4800	24000	5,00
totale	20000	128000	33,10
Stima dei parametri	4000	25600	6,40
Stima dell'errore standard	357,78	2579,92	0,98

Le diversità che si riscontrano fra le stime dell'errore standard ottenute con il metodo dei gruppi casuali e le corrispondenti stime ricavate con la procedura standard sono di tipo accidentale, imputabili all'operazione di scelta casuale dei gruppi utilizzati per i calcoli.

La stabilità dello stimatore

Il numero k dei gruppi casuali in cui viene suddiviso il campione totale, che finora è stato considerato come una quantità nota, influisce sulla precisione dello stimatore della varianza campionaria.

Più in generale, se i k subcampioni sono indipendenti, il coefficiente di variazione della stima della varianza campionaria è dato da (Wolter 1985, pag. 55):

$$CV = \sqrt{\frac{\beta\left(\hat{\theta}_s\right) - (k-3)(k-1)}{k}} \qquad [6.2]$$

dove $\beta(\hat{\theta}_s)$ è l'indice di curtosi dello stimatore $\hat{\theta}_s$.

Dalla [6.2] si evidenzia che il coefficiente di variazione dello stimatore della varianza, a parità di β, è decrescente con il crescere del numero dei gruppi casuali. Pertanto più k è elevato maggiore sarà la precisione dello stimatore.

Nel caso di campionamento casuale semplice con reimmissione, l'indice di curtosi dello stimatore è dato da:

$$\beta\left(\hat{\theta}_r\right) = \frac{1}{m}\left[\beta + 3(m-1)\right] \qquad [6.3]$$

dove β è l'indice di curtosi del carattere in esame nella popolazione.

Sostituendo la [6.2] nella [6.4] e ricordando che m=n/k, si ricava:

$$CV = \sqrt{\frac{\beta}{n} + \frac{3(n-k)}{nk} - \frac{k-3}{k(k-1)}} \qquad [6.4]$$

Determinazione del numero dei gruppi casuali

La [6.4] può essere utilizzata per determinare il numero dei gruppi casuali che devono essere formati per ottenere una stima della varianza campionaria con un errore relativo non superiore a CV.

$$k \geq 1 + \frac{2n}{nCV^2 + (3-\beta)} \qquad [6.5]$$

Se si fissa CV e si ha una qualche informazione su β è possibile calcolare k mediante la [6.5].

Nel caso in cui il carattere in esame ha una distribuzione normale ($\beta = 3$), il numero k dei gruppi casuali è indipendente da n, ed è dato da:

$$k \geq 1 + \frac{2}{CV^2} \qquad [6.6]$$

da cui si ricava la seguente distribuzione di valori di k in funzione di CV:

CV	k
0.02	5000
0.04	1.251
0.05	801
0.10	201
0.15	90
0.20	51

Se si vuole stimare la varianza campionaria con un errore non superiore al 10% è necessario utilizzare almeno 201 gruppi casuali.

Nel caso di stime di frequenza si deve fare riferimento alla distribuzione binomiale, per la quale β risulta variabile con i valori del parametro P (frequenza relativa nella popolazione).

Nella tav. 6.3 sono stati tabulati i valori di k, per un errore della stima della varianza non superiore al 10%, in corrispondenza a diverse numerosità campionarie e per valori di β da 1 a 8 che, come si è visto nel terzo capitolo, comprendono la maggior parte dei casi che si possono presentare nella pratica.

Dall'esame dei valori si evince che per grandi campioni, o per campioni di ridotte dimensioni ma con valori di β non elevati, sono necessari all'incirca 200 gruppi casuali per ottenere stime della varianza sufficientemente stabili.

Questi risultati ottenuti nell'ipotesi di gruppi casuali indipendenti possono essere estesi, con approssimazioni trascurabili, ai più comuni disegni campionari adottati dall'Istituto.

Tavola 6.3 - Valori di k per CV = 0,10 in funzione di n e di β

n \ β	1	2	3	4	5	6	7	8
1000	168	183	201	223	251	287	334	401
2000	183	191	201	212	223	236	251	268
3000	189	195	201	208	215	223	232	241
4000	191	196	201	206	212	217	223	230
5000	193	197	201	205	209	214	218	223
10000	197	199	201	203	205	207	209	212
20000	199	200	201	202	203	204	205	206
30000	200	200	201	202	202	203	204	204
40000	200	201	201	202	202	203	203	204
50000	200	201	201	201	202	202	203	203
100000	201	201	201	201	201	202	202	202
200000	201	201	201	201	201	201	201	202

3. Estensione del metodo dei gruppi casuali ai disegni campionari complessi

Affinché lo stimatore della varianza mediante il metodo dei gruppi casuali abbia proprietà statistiche accettabili, i subcampioni casuali non devono essere formati arbitrariamente, ma è necessario che siano estratti in modo che ciascuno di essi abbia lo stesso disegno campionario del campione totale.

Questa condizione può essere soddisfatta per la quasi totalità dei piani di campionamento utilizzati dall'Istat seguendo le regole appresso riportate.

a) Piano di campionamento ad uno stadio stratificato

Si indichi con L il numero degli strati, con n_h il numero delle unità estratte dallo strato h senza reimmissione con uguale probabilità o con probabilità proporzionale all'ampiezza, e con k_h il numero di subcampioni casuali che deve essere formato nello strato.

Nello strato h il primo subcampione viene ricavato estraendo un campione casuale semplice senza reimmissione di ampiezza $m_h = n_h/k_h$, assegnando uguale probabilità a ciascuna delle n_h unità campionate nello strato.

Il secondo subcampione casuale è ottenuto estraendo un campione d'ampiezza m_h dalle rimanenti $n_h - m_h$ unità e così via.

Se n_h/k_h non è intero, ossia $n_h = k_h + q_h$ allora le rimanenti q_h unità vengono distribuite tra i primi q_h gruppi casuali.

Il metodo dei gruppi casuali viene quindi applicato a ciascuno degli L strati, ricavando così la stima del parametro e della varianza campionaria per lo strato h.

Ripetendo il procedimento per tutti gli strati, le stime relative al campione totale sono date da:

$$\hat{\theta} = \sum_{h=1}^{L} \hat{\theta}_h \qquad [6.7]$$

$$Var\left(\hat{\theta}\right) = \sum_{h=1}^{L} Var\left(\hat{\theta}_h\right) \qquad [6.8]$$

Il procedimento descritto fornisce le stime a livello di strato e quindi consente di calcolare la varianza, oltre che per lo stimatore riferito all'intera popolazione, anche per quelli relativi a subpopolazioni definite dall'aggregazione di due o più strati, come ad esempio i domini territoriali.

b) Piano di campionamento a due stadi con stratificazione delle unità di primo stadio.

Nel campionamento a due stadi i subcampioni casuali devono essere formati suddividendo, in ciascuno strato, soltanto le unità di primo stadio e lasciando aggregate le unità elementari che formano la stessa PSU.

Così se in primo stadio di campionamento vengono estratti i comuni e in secondo stadio le famiglie, un subcampione casuale è costituito da una parte dei comuni campionati nello strato e da tutte le famiglie estratte da quei comuni.

La formazione dei gruppi casuali viene, quindi, effettuata operando sulle sole PSU, e non tiene conto dei successivi stadi di campionamento. Pertanto la procedura che viene descritta per un disegno campionario a due stadi può essere applicata direttamente anche a campioni basati su tre o più stadi.

In ciascuno strato i subcampioni vengono generati seguendo la stessa regola riportata nel punto a) sia se le PSU sono state estratte con probabilità uguale che con probabilità proporzionale all'ampiezza.

c) Unità autorappresentative

A volte vengono utilizzati piani di campionamento in cui le PSU di dimensioni maggiori, dette autorappresentative, sono tutte comprese nel campione, ossia vengono selezionate con probabilità uguale ad 1.

Quando il disegno campionario prevede la presenza di unità autorappresentative è necessaria una particolare attenzione nella formazione dei subcampioni. Infatti ogni unità autorappresentativa va considerata come uno strato a se stante, per cui al loro interno i subcampioni devono essere formati raggruppando le unità di secondo stadio, che in questo caso rappresentano le PSU.

Così nei disegni campionari comunemente utilizzati nelle indagini sulle famiglie, i gruppi casuali all'interno di ciascun comune autorappresentativo devono essere costituiti mediante raggruppamenti casuali delle famiglie campione.

Poiché i tempi di elaborazione sono proporzionali al numero di PSU che vengono utilizzate per il calcolo degli errori di campionamento, risulta evidente che la presenza di unità autorappresentative aumentando il numero di PSU comporta un notevole appesantimento dei tempi di esecuzione dei programmi.

È possibile superare questo inconveniente e rendere l'elaborazione più economica raggruppando, in modo opportuno, le unità di secondo stadio all'interno delle unità autorappresentative, così da formare delle «pseudo» PSU di dimensioni maggiori ma meno numerose.

L'adozione di questo accorgimento comporta una sovrastima della varianza campionaria, che può essere ridotta a valori trascurabili se per la formazione delle pseudo PSU si utilizza un procedimento di raggruppamento casuale (cfr. ad es. Verma, 1982, pag. 22).

d) Estrazione di una sola unità per strato

Un'altra tecnica frequentemente utilizzata nei piani di campionamento a due stadi con stratificazione delle unità di primo stadio, è quella di estrarre una sola PSU da ciascuno strato.

Tale procedimento, efficiente e semplice da utilizzare, crea però dei problemi nalla stima della varianza campionaria, in quanto disponendo di una sola PSU campione per strato non è possibile stimare la variabilità negli strati.

Per superare questo inconveniente e ricavare comunque una stima della varianza campionaria, si procede alla formazione dei nuovi strati mediante il raggruppamento di due o più strati. Le PSU che cadono nel nuovo pseudo strato vengono riguardate come unità selezionate in modo indipendente.

4. Il metodo delle replicazioni bilanciate ripetute

Uno dei maggiori limiti del metodo dei gruppi causali va ricercato nella difficoltà di costruire un numero sufficientemente elevato di *sub*campioni per quei *disegni* campionari che prevedono l'estrazione di poche unità di primo stadio per strato.

Così per i disegni campionari in cui vengono estratte due sole PSU per strato, è possibile formare soltanto due subcampioni indipendenti, per cui la stima della varianza che si ottiene con questo metodo risulta poco stabile.

Per superare questo inconveniente il Bureau of Census degli Stati Uniti ha messo a punto, attorno agli inizi degli anni 60, un metodo basato sulle replicazioni bilanciate ripetute di metà campione.

Il metodo, ripreso e perfezionato da numerosi autori, viene ormai utilizzato correntemente nelle indagini campionarie su larga scala, anche perché supportato da diversi pacchetti informatici.

Nel riportare i principi base del metodo si farà riferimento all'impostazione data da P. J. McCarthy (1966, 1969), cui è dovuta la prima trattazione organica di questa tecnica.

Per descrivere il metodo verrà preso in esame un disegno campionario ad uno stadio stratificato in cui vengono estratte

con reimmissione due unità per strato; successivamente verranno riportate le regole che dovranno essere seguite per adattare il metodo a disegni campionari più complessi.

Le replicazioni di metà campione

Si indichi con il L il numero degli strati, con u_{h1} e u_{h2} le due unità campione dello strato h; il campione è, quindi formato da 2 L osservazioni indipendenti. Una replicazione di metà campione si ottiene mediante scelta casuale di una unità da ciascuno degli L strati.

Il numero delle possibili replicazioni di metà campione è dato dalle disposizioni con ripetizione di 2 elementi di classe L.

Si indichi con la a_{hr} la variabile che assume il valore + 1 se la prima unità dello stato h è compresa nell'r.ma replicazione e il valore – 1 se invece vi è compresa la seconda unità. L'insieme delle possibili replicazioni può essere descritto da una matrice di dimensioni 2 × L di valori + 1 e – 1.

Si può verificare che i coefficienti a_{hr} soddisfano le seguenti uguaglianze:

$$\sum_{r=1}^{2^L} a_{hr} = 0 \qquad (h = 1,...,L) \qquad [6.9]$$

$$\sum_{r=1}^{2^L} a_{hr}\ a_{kr} = 0 \qquad (h = 1,...,L;\ k \neq h) \qquad [6.10]$$

La prima uguaglianza sta da indicare che le due unità campione di ciascuno strato sono comprese lo stesso numero di volte nell'insieme delle possibili replicazioni;

la seconda uguaglianza esprime, invece, la condizione di ortogonalità delle colonne della matrice.

Esempio 6.2. Per L = 3 si ha il seguente campione costituito complessivamente da 6 unità:

Strato	Unità del campione		Peso
1	u_{11}	u_{12}	w_1
2	u_{12}	u_{22}	w_2
3	u_{31}	u_{33}	w_3

Ciascuna replicazione di metà campione è costituita da 3 unità, una per ogni strato, e il numero delle possibili replicazioni è dato da 2 = 8:

Replicazione	Strati 1	2	3
1	u_{11}	u_{21}	u_{31}
2	u_{11}	u_{21}	u_{32}
3	u_{11}	u_{22}	u_{31}
4	u_{11}	u_{22}	u_{32}
5	u_{12}	u_{21}	u_{31}
6	u_{12}	u_{21}	u_{32}
7	u_{12}	u_{22}	u_{31}
8	u_{12}	u_{22}	u_{32}

che possono essere rappresentate mediante la seguente matrice:

Replicazione	Strati 1	2	3
1	+1	+1	+1
2	+1	+1	−1
3	+1	−1	+1
4	+1	−1	−1
5	−1	+1	+1
6	−1	+1	−1
7	−1	−1	+1
8	−1	−1	−1

si può verificare che i coefficienti a_{hr} soddisfano le condizioni [6.9] e [6.10].

Poiché ogni replicazione riproduce in scala ridotta il disegno complessivo dell'intero campione, è possibile ricavare da ciascuna di esse uno stimatore del parametro della popolazione, applicando lo stesso procedimento usato per l'intero campione.

Indicando con $\hat{\theta}$ lo stimatore relativo al campione totale e con $\hat{\theta}_r$ quello che si desume dall'r.ma replicazione si dimostra che vale l'uguaglianza:

$$\hat{\theta} = \frac{1}{2^L} \sum_{r=1}^{2^L} \hat{\theta}_r \qquad [6.11]$$

Lo stimatore della varianza di $\hat{\theta}$ è dato da:

$$Var(\hat{\theta}) = \frac{1}{2^L} \sum_{r=1}^{2^L} (\hat{\theta}_r - \hat{\theta})^2 \qquad [6.12]$$

Quando gli stimatori non sono lineari la [6.11] non è valida e la [6.12] costituisce uno stimatore dell'errore quadratico medio.

Appare evidente che un tale modo di procedere comporta notevoli costi computazionali non appena L è un poco elevato, basti pensare che per L = 20 si ha un insieme di oltre un milione di replicazioni.

Una via naturale per superare questo inconveniente è quella di utilizzare soltanto un numero R ridotto di replicazioni, semplificando così le difficoltà di calcolo.

Pertanto se le R replicazioni vengono scelte a caso lo stimatore della varianza campionaria risulta più elevato di quello che si avrebbe operando sulle 2^L replicazioni.

Le replicazioni bilanciate ripetute

La soluzione per ottenere stimatori corretti utilizzando un numero ridotto di replicazioni è stato fornita da McCarthy con l'introduzione delle replicazioni bilanciate ripetute.

Un sottoinsieme di R replicazioni si dice bilanciato se soddisfa la condizione di ortogonalità:

$$\sum_{r=1}^{R} a_{hr}\, a_{kr} = 0 \qquad (h = 1, ..., L;\ k \neq h) \qquad [6.13]$$

e completamente bilanciato se soddisfa anche la condizione:

$$\sum_{r=1}^{R} a_{hr} = 0 \qquad (h = 1, ..., L) \qquad [6.14]$$

Si dimostra che lo stimatore della varianza campionaria basato su R replicazioni bilanciate ripetute riproduce esattamente lo stimatore che si ottiene utilizzando l'insieme completo delle 2^L replicazioni.

Il problema è quindi ricondotto alla costruzione di un insieme di replicazioni bilanciate per un prefissato numero di strati. Si può infatti, verificare che non tutte le R replicazioni che possono essere scelte dall'insieme totale soddisfano le due condizioni precedenti.

Esempio 6.3. Così per L = 3 e per R = 4 ci sono 70 possibili sottoinsiemi di 4 replicazioni e soltanto alcuni di essi hanno la caratteristica di essere bilanciati.

Ad esempio risulta bilanciato il sottoinsieme costituito dalle prime 4 replicazioni:

Replicazione	Strati 1	2	3
1	+1	+1	+1
2	+1	+1	−1
3	+1	−1	+1
4	+1	−1	−1

per il quale è possibile verificare che risulta soddisfatta la condizione (7.6), ma non completamente bilanciato in quanto non tutte le somme per colonna sono nulle.

Mentre il sottoinsieme costituito dalle replicazioni 1, 2, 4, 5 non è bilanciato:

Replicazione	Strati		
	1	2	3
1	+1	+1	+1
2	+1	+1	−1
4	+1	−1	+1
5	+1	+1	+1

La matrice di Hadamard

La costruzione di sottoinsiemi di replicazioni bilanciate ripetute può essere effettuata utilizzando le matrici di Hadamard. Queste sono matrici quadrate che hanno la proprietà di avere le colonne ortogonali a due a due, e possono essere generate per dimensioni multiple di 4.

Così utilizzando il metodo proposto di Plackett e Burman (1946) per ottenere le matrici di Hadamard, la matrice di ordine 4 è data da:

+1	+1	+1	+1
−1	+1	−1	+1
−1	−1	+1	+1
+1	−1	−1	+1

Per L = 3 si può ottenere un sottoinsieme di 4 replicazioni bilanciate scegliendo 3 delle quattro colonne della matrice, ad esempio le prime 3:

Replicazione	Strati		
	1	2	3
1	+1	+1	+1
2	+1	+1	−1
4	−1	−1	+1
5	+1	−1	−1

Si verifica immediatamente che i coefficienti a_{hr} soddisfano le due condizioni di bilanciamento completo.

La matrice di Hadamard di ordine 4 può essere impiegata per costruire insiemi di 4 replicazioni completamente bilanciate nel caso il cui il numero degli strati è minore di 4, e semplicemente bilanciato per L = 4 in quanto la somma dei coefficienti dell'ultima colonna è diversa da zero.

Più in generale se si hanno L strati si possono ricavare R replicazioni bilanciate utilizzando la matrice di Hadamard di ordine K, dove K è un intero multiplo di 4 tale che: $L < K < 2^L$.

Se ad esempio si vuole ricavare un insieme di 8 replicazioni bilanciate per un disegno campionario con 5 strati, si utilizzano 5 colonne della matrice di Hadamard di ordine 8:

Matrice di Hadamard di ordine 8

+1	+1	+1	+1	+1	+1	+1	+1
−1	−1	−1	+1	−1	+1	−1	+1
−1	−1	+1	+1	−1	−1	+1	+1
+1	−1	−1	+1	+1	−1	−1	+1
+1	+1	+1	−1	−1	−1	−1	+1
−1	+1	−1	−1	+1	−1	+1	+1
−1	−1	+1	−1	+1	+1	−1	+1
+1	−1	−1	−1	−1	+1	+1	+1

Quando L < K si ottiene un bilanciamento completo, mentre se L = K si ricava un insieme semplicemente bilanciato.

Nell'appendice A del volume di Wolter (op. cit.) sono riportate le matrici di Hadamard fino a quella di ordine 100. È possibile costruire matrici di ordine superiore utilizzando la procedura passo-passo sviluppata da Gurney e Jewett del Bureau of Census (1975), che è facilmente implementabile.

Stimatori della varianza campionaria

Si indichino con $\hat{\theta}$ e $\hat{\theta}_r$ gli stimatori relativi al campione totale e all'r.ma replicazione bilanciata (r = 1, 2, . . ., R), ottenuti utilizzando la stessa forma funzionale non necessariamente lineare.

Lo stimatore della varianza basato sulle R replicazioni bilanciate è dato da:

$$\hat{V}^{(1)}_{BRR}\left(\hat{\theta}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \hat{\theta}\right)^2 \qquad [6.15]$$

Se accanto ad ogni replicazione r si considera la replicazione complementare formata dalle unità del campione non comprese in r, è possibile definire altri tre stimatori della varianza:

$$V^{(2)}_{BRR}\left(\hat{\theta}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}^{(c)}_r - \hat{\theta}\right)^2 \qquad [6.16]$$

$$V^{(3)}_{BRR}\left(\hat{\theta}\right) = \frac{V^{(1)}_{BRR}\left(\hat{\theta}\right) + V^{(2)}_{BRR}\left(\hat{\theta}\right)}{2} \qquad [6.17]$$

$$V^{(4)}_{BRR}\left(\hat{\theta}\right) = \frac{1}{4R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \hat{\theta}^{(c)}_r\right)^2 \qquad [6.18]$$

dove $\hat{\theta}^{(c)}_r$ è la stima che si ottiene dalla replicazione complementare.

I primi due stimatori della varianza sono dell'identica forma e differiscono soltanto per le replicazioni che vengono utilizzate. Il terzo stimatore è una media aritmetica semplice dei primi due ed è più preciso degli altri anche se richiede un maggior numero di calcoli.

Quando la funzione di stima è lineare i quattro stimatori coincidono, mentre nel caso non lineare portano i valori differenti.

Per ricapitolare quanto fin qui esposto è bene fare riferimento ad un esempio numerico utilizzando dati artificiali.

Esempio 6.4. Sia U una popolazione di $N = 100$ unità suddivisa in $L = 5$ strati di numerosità $N_1 = 10$, $N_2 = 20$, $N_3 = 40$, $N_4 = 20$ e $N_5 = 10$. Da ogni strato vengono selezionate due unità mediante estrazione casuale semplice con reimmissione e su ogni unità del campione vengono rilevati i valori di due caratteri quantitativi. I risultati della rilevazione sono riportati nel prospetto che segue:

Strato	unità	$y_{(1)hi}$	$y_{(2)hi}$	w_{hi}
1	1	3	5	5
1	2	4	5	5
2	1	7	8	10
2	2	5	7	10
3	1	5	6	20
3	2	2	4	20
4	1	4	6	10
4	2	5	9	10
5	1	3	4	5
5	2	4	6	5

Si vogliono stimare i totali dei due caratteri e calcolare la varianza campionaria degli stimatori mediante il metodo delle BRR.

Utilizzando lo stimatore del totale per il campionamento stratificato (cfr. cap. 4) si ha: $\hat{Y}_{(1)} = 420$ e $\hat{Y}_{(2)} = 600$.

Per stimare la varianza campionaria con il metodo delle replicazioni bilanciate ripetute occorre:

1. Stabilire il numero R di replicazioni, dove R deve essere un multiplo di 4 compreso tra 5 e 32. Possono quindi essere scelti i seguenti valori di R: 8, 12, 16, 20, 24, 28 e 32, si supponga di prendere $R = 8$;

2. Costruire la matrice di Hadamard di ordine 8;

3. Scegliere 5 delle 8 colonne della matrice di Hadamard, ad esempio, le colonne de 2 a 6, ottenendo la matrice 8×5:

	Strati				
Replicazione	1	2	3	4	5
1	+1	+1	+1	+1	+1
2	+1	−1	+1	−1	+1
3	−1	+1	+1	−1	−1
4	−1	−1	+1	+1	−1
5	+1	+1	−1	−1	−1
6	+1	−1	−1	+1	−1
7	−1	+1	−1	+1	+1
8	−1	−1	−1	−1	+1

4. Ricordando che per ogni strato i valori + 1 e − 1 stanno ad indicare la scelta rispettivamente della prima e della 2 unità, si ricava il seguente insieme di 8 replicazioni completamente bilanciate:

Tavola 6.4 – Replicazioni bilanciate ripetute di metà campione

Replicazione	strato	unità	$y_{(1)hi}$	$y_{(2)hi}$	w_{hi}
1	1	1	3	5	10
	2	1	7	8	20
	3	1	5	6	40
	4	1	4	6	20
	5	1	3	4	10
2	1	1	3	5	10
	2	2	5	7	20
	3	1	5	6	40
	4	2	5	9	20
	5	1	3	4	10
3	1	2	4	5	10
	2	1	7	8	20
	3	1	5	6	40
	4	2	5	9	20
	5	2	4	6	10
4	1	2	4	5	10
	2	2	5	7	20
	3	1	5	6	40
	4	1	4	6	20
	5	2	4	6	10
5	1	1	3	5	10
	2	1	7	8	20
	3	2	2	4	40
	4	2	5	9	20
	5	2	4	6	10
6	1	1	3	5	10
	2	2	5	7	20
	3	2	2	4	40
	4	1	4	6	20
	5	2	4	6	10
7	1	2	4	5	10
	2	1	7	8	20
	3	2	2	4	40
	4	1	4	6	20
	5	1	3	4	10
8	1	2	4	5	10
	2	2	5	7	20
	3	2	2	4	40
	4	2	5	9	20
	5	1	3	4	10

5. Per ogni replicazione si calcolano le stime $\hat{Y}_{(1)}$ e $\hat{Y}_{(2)}$ seguento lo stesso procedimento utilizzato per il campione totale, ossia una somma ponderata con pesi uguali a w_{hi}. Così per la prima replicazione si ha:

$$\hat{Y}_{(1)1} = 3 \cdot 10 + 7 \cdot 20 + 5 \cdot 40 + 4 \cdot 20 + 2 \cdot 10 = 480$$

$$\hat{Y}_{(2)1} = 5 \cdot 10 + 8 \cdot 20 + 6 \cdot 40 + 6 \cdot 20 + 4 \cdot 10 = 610$$

In modo analogo si ricavano le stime relativo alle altre replicazioni, ottenendo:

Replicazioni	$\hat{Y}_{(1)r}$	$\hat{Y}_{(2)r}$
1	480	610
2	460	650
3	520	690
4	460	610
5	390	610
6	330	530
7	370	530
8	350	570

6. Infine si calcolano le medie e le varianze delle due distribuzioni, ottenendo le stime dei due totali e le corrispondenti varianze campionarie:

$$\hat{Y}_{(1)} = 420 \quad \hat{V}_{BRR}(\hat{Y}_{(1)}) = 4150$$
$$\hat{Y}_{(2)} = 600 \quad \hat{V}_{BRR}(\hat{Y}_{(2)}) = 2700$$

5. Estensione del metodo delle BRR ai disegni campionari complessi

Finora il metodo delle BRR è stato introdotto con riferimento ad un piano di campionamento ad uno stadio stratificato con estrazione di due unità per strato con reimmissione. Da quanto esposto risulta evidente che in una situazione così semplice l'impiego delle BRR non è molto vantaggioso, in quanto questa tecnica riproduce i risultati che possono essere ottenuti in modo molto semplice con i metodi standard di stima della varianza campionaria.

I maggiori vantaggi della tecnica BRR si hanno nella sua applicazione a piani di campionamento complessi, nei quali i metodi diretti per la stima della varianza presentano notevoli difficoltà di calcolo. In questi casi, comunque, la metodologia fin qui descritta deve essere opportunamente adattata.

Senza perdere di generalità, per semplificare l'esposizione delle regole che devono essere seguite per applicare il metodo delle BRR ai più comuni disegni campionari utilizzati nella pratica, si farà riferimento allo stimatore di un totale.

a) Piani di campionamento a due o più stadi

Si consideri un piano di campionamento a più stadi in cui le unità di primo stadio sono suddivise in L strati. In ciascuno strato vengono selezionate due PSU senza reimmissione e con probabilità d'inclusione proporzionali all'ampiezza. L'usuale stimatore del totale della popolazione è dato da:

$$\hat{Y} = \sum_{h=1}^{L} \left(\frac{\hat{Y}_{h1}}{\pi_{h1}} + \frac{\hat{Y}_{h2}}{\pi_{h2}} \right) \qquad [6.19]$$

dove $\hat{Y}_{h1}$ e $\hat{Y}_{h2}$ sono le stime dei totali delle due PSU campione dello strato h.

Si formano R replicazioni bilanciate ripetute mediante le 2^L PSU campione e per ogni replicazione si calcola la stima $\hat{Y}_r$:

$$\hat{Y}_r = \sum_{h=i}^{m_{hi}} \sum_{i=1}^{2} a_{hri} \frac{2\hat{Y}_{hi}}{\pi_{hi}} \qquad [6.20]$$

dove a_{hri} assume il valore 1 se l'i.ma PSU è compresa nella replicazione r.ma e O altrimenti.

La stima della varianza campionaria di $\hat{Y}$ con il metodo delle BRR è data da:

$$\hat{V}_{BRR}(\hat{Y}) = \frac{1}{R} \sum (\hat{Y}_r - \hat{Y})^2 \qquad [6.21]$$

La [6.21] è una stima corretta della varianza campionaria dello stimatore di un totale solo nel caso in cui le PSU vengono estratte da ciascuno strato con reimmissione. Quando le PSU sono selezionate senza reimmissione si ha una distorsione che può essere considerata come trascurabile nelle applicazioni pratiche.

b) Estrazione di una sola PSU per strato

Quando viene estratta una sola PSU per strato lo stimatore del totale della popolazione è dato da:

$$\hat{Y} = \sum_{h=1}^{L} \hat{Y}_h \qquad [6.22]$$

dove $\hat{Y}_h$ è lo stimatore del totale dello strato h, desunto dai valori osservati nell'unica PSU campionata nello strato.

Come è noto per ottenere uno stimatore della varianza campionaria di $\hat{Y}$ occorre raggruppare gli strati, ed indicando con G il numero dei nuovi gruppi ottenuti collapsando coppie di strati originali si ha:

$$\hat{Var}\left(\hat{Y}\right) = \sum_{g=1}^{G} \left(\hat{Y}_{g1} - \hat{Y}_{g2}\right)^2 \qquad [6.23]$$

dove $\hat{Y}_{g1}$ e $\hat{Y}_{g2}$ sono gli stimatori dei totali dei due strati appartenenti al g.mo gruppo.

Lo stimatore della varianza ottenuto raggruppando gli strati può essere riprodotto esattamente utilizzando il metodo delle BRR.

Si costruiscono R replicazioni bilanciate ripetute, ciascuna delle quali sarà formata da G valori ottenuti prendendo uno strato originale da ciascun gruppo.

La stima della varianza è data da:

$$\hat{V}_{BRR}\left(\hat{Y}\right) = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{Y}_r - \hat{Y}\right)^2 \qquad [6.24]$$

dove $\hat{Y}_r$ è la stima del totale calcolata per l'r.ma replicazione.

Come è noto per questo disegno campionario non esistono stimatori centrati della varianza campionaria, ed anche lo stimatore con il metodo delle BRR presenta una disorsione positiva.

c) Unità autorappresentative

Quando nel disegno campionario sono presenti unità autorappresentative, queste vanno trattare come strati e le PSU sono costituite dalla unità di secondo stadio che vengono estratte da ciascuna unità autorappresentativa.

In questo caso si hanno più di due PSU per strato per cui il metodo deve essere opportunatamente adattato.

Una prima via che può essere seguita è quella di suddividere le unità di secondo stadio in un certo numero di strati artificiali ciascuno costituito da due sole unità, e costruire le R replicazioni bilanciate prendendo un'unità da ciascun pseudostrato.

Il procedimento molto semplice da utilizzare diventa assai laborioso quando il numero di unità di secondo stadio campionate nelle PSU autorappresentative è elevato.

In questo caso si può adottare un'altra procedura, altrettanto semplice ma che comporta un minor numero di calcoli. Le unità di secondo stadio di ciascuna PSU comune autorappresentativa vengono suddivise in due gruppi causali, possibilmente della stessa ampiezza, e la coppia di pseudo PSU così ottenuta viene utilizzata per formare l'insieme delle R replicazioni bilanciate.

CAPITOLO 7 - LA PROCEDURA GENERALIZZATA UTILIZZATA DALL'ISTAT

1 Premessa

Nella pratica del campionamento spesso può accadere che il campione realizzato si discosti da quello programmato, per cui non sono più soddisfatte le condizioni richieste per l'applicazione degli usuali stimatori della varianza.

In questi casi non è conveniente ricercare le formule esatte per la stima della varianza campionaria, le quali, d'altro canto richiederebbero poi la predisposizione del relativo programma informatico.

Bisogna, infatti, sempre tener presente che l'obiettivo principale di un'indagine è *quello* di fornire le stime, il più possibile precise, dei parametri d'interesse e che gli errori di campionamento rappresentano soltanto gli indicatori di tale precisione.

La stima degli errori campionari non deve quindi comportare un costo eccessivo rispetto all'economia generale dell'indagine, trattandosi in definitiva di un sottoprodotto dell'indagine stessa.

Per questo motivo all'Istat dall'inizio degli anni '80 si è deciso di ricorrere a procedure generalizzate applicabili a tutte le situazioni concrete e che nello stesso tempo presentino caratteristiche di semplicità, flessibilità ed economicità di tempo e di costo.

Dopo un esame comparato dei diversi pacchetti allora disponibili (cfr. Francis e Sedransk, 1979), si è deciso di adottare il programma CLUSTERS (Computation and Listing of USeful STatistics for Error of Sampling), sviluppato da Verma e Pearce (1978) nell'ambito della World Fertility Survey.

Il programma, introdotto all'Istat nel 1981 e utilizzato in via sperimentale in occasione della prima indagine sulle condizioni di salute della popolazione (Zannella, 1982), visti i buoni risultati, viene ormai utilizzato correntemente per il calcolo degli errori di campionamento.

In questo capitolo vengono descritte la metodologia per la stima della varianza campionaria che sta alla base del programma CLUSTERS e le regole che devono essere seguite per il suo adattamento ai principali disegni campionari impiegati dall'Istat.

Nell'appendice 1 è riportata una descrizione particolareggiata del programma CLUSTERS e delle modalità per la sua utilizzazione sotto CMS; mentre nell'appendice 2 viene illustrata un'applicazione della procedura ai risultati della seconda indagine sulle condizioni di salute della popolazione italiana.

2. Metodologia per la stima della varianza campionaria

Nel descrivere la metodologia si farà riferimento ai lavori di Kalton (1977) e di Verma (1982) e le formule per il calcolo della varianza campionaria verranno sviluppate per un disegno campionario a due o più stadi con stratificazione delle unità di primo stadio. Ovviamente nel caso di un campionamento ad un solo stadio le PSU verranno ad identificarsi con le unità elementari di rilevazione.

Le ipotesi di base

Per la stima della varianza vengono fatte due ipotesi, spesso non valide nella pratica ma che l'esperienza empirica porta a giustificare:

1) in ciascuno strato vengono selezionate due o più unità di primo stadio;
2) le unità di primo stadio sono scelte mediante estrazioni indipendenti e con probabilità variabile.

La prima assunzione è sempre soddisfatta tranne nel caso dei piani di campionamento attualmente utilizzati per alcune indagini sulle famiglie, in cui viene estratto un solo comune per strato. Si è visto nei capitoli precedenti, che in questi casi è possibile superare l'inconveniente mediante un opportuno raggruppamento degli strati (collapsed strata).

La seconda ipotesi implica che la selezione delle PSU venga effettuata con reimmissione. Questa assunzione generalmente non è soddisfatta, perché il ricorso ad un campionamento con reimmissione comporta una perdita di efficienza ed anche delle notevoli complicazioni pratiche nel caso in cui la stessa unità è compresa più volte nel campione. L'uso del metodo quando questa condizione non è valida porta ad una sovrastima della varianza, che è trascurabile nel caso in cui la frazione di campionamento in primo stadio è piccola.

Stimatore del totale in uno strato

Si supponga comunque che tali condizioni siano soddisfatte e che le PSU siano state raggruppate in L strati. Facendo riferimento all'h.mo strato si indichi con:

N_h = numero di PSU nella popolazione
n_h = numero di PSU nel campione
i = indice di PSU
M_{hi} = numero di unità elementari della popolazione nell'i.ma PSU
m_{hi} = numero di unità elementari del campione nell'i.ma PSU

j = indice di unità elementare
y_{hij} = valore della variabile r.ma osservato nella j.ma unità elementare

In ogni strato le PSU campione vengono estratte con probabilità variabile e con reimmissione, mentre le unità elementari vengono selezionate da ciascuna PSU campione con uguale probabilità e senza reimmissione.

Le unità elementari all'interno di ciascuna PSU campione hanno tutte la stessa probabilità di selezione, che è data da:

$$p_{hij} = \frac{n_h\, p_{hi}\, m_{hi}}{M_{hi}} \qquad [7.1]$$

dove con p_{hi} è stata indicata la probabilità di selezione dell'i.ma PSU campione in ciascuna delle n_h estrazioni indipendenti.

Nel caso, molto frequente nella pratica, in cui le PSU vengono selezionate con probabilità proporzionale all'ampiezza, la [7.1] diventa:

$$p_{hij} = n_h \frac{m_{hi}}{M_h} \qquad [7.2]$$

dove m_h è il numero di unità elementari della popolazione che sono comprese nello strato h.

Lo stimatore del totale nello strato h è dato da:

$$\hat{Y}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \frac{y_{hij}}{p_{hij}} = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij} \qquad [7.3]$$

dove $w_{hij} = 1/p_{hij}$ è il coefficiente di espansione associato alla j.ma unità elementare dell'i.ma PSU.

La [7.3] può anche essere scritta:

$$\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi} \qquad [7.4]$$

dove:

$$\hat{Y}_h = \sum_{j=1}^{m_{hi}} w_{hij}\, y_{hij} \qquad [7.5]$$

Poiché gli stimatori $\hat{Y}_{hi}$ sono indipendenti ed hanno uguale varianza $Var(\hat{Y}_{hi})$, si ha:

$$Var\left(\hat{Y}_h\right) = n_h\, Var\left(\hat{Y}_{hi}\right) \qquad [7.6]$$

Uno stimatore corretto di $\hat{Var}(\hat{Y}_h)$ è dato da:

$$\hat{Var}\left(\hat{Y}_h\right) = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2 \qquad [7.7]$$

che, sostituita nella [7.6], permette di ricavare uno stimatore corretto della varianza campionaria:

$$\hat{Var}\left(\hat{Y}_h\right) = \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2 \qquad [7.8]$$

Le stime a livello di strato possono essere utilizzate per ricavare quelle relative all'intera popolazione e ai domini di studio ottenuti come raggruppamento di due o più strati.

Stime riferite all'intera popolazione

Stimatore del totale:

$$\hat{Y} = \sum_{i=1}^{L} \hat{Y}_h \qquad [7.9]$$

Stimatore della varianza campionaria

$$\hat{Var}\left(\hat{Y}\right) = \sum_{h=1}^{L} \hat{Var}\left(\hat{Y}_h\right) \qquad [7.10]$$

Stime riferite ad un dominio di studio

Stimatore del totale:

$$\hat{Y}_d = \sum_{h \in d} \hat{Y}_h \qquad [7.11]$$

Stimatore della varianza campionaria

$$\hat{Var}\left(\hat{Y}_d\right) = \sum_{h \in d} \hat{Var}\left(\hat{Y}_h\right) \qquad [7.12]$$

Stime riferite a sottoclassi

Nel caso di sottoclassi, la varianza campionaria dello stimatore di un totale si ottiene applicando le formule precedenti con l'avvertenza di trascurare in ciascuna PSU le unità elementari che non appartengono alla sottoclasse in esame. Se una PSU non ha alcuna unità elementare compresa nella sottoclasse, è l'intera PSU a non essere considerata nei calcoli.

Stimatori non lineari

Per la stima della varianza campionaria di stimatori non lineari, la procedura utilizza il metodo dello sviluppo in serie di Taylor. In effetti, la metodologia su cui è basato il programma CLUSTERS prevede che ogni stimatore venga definito come un rapporto $R = \hat{Y}/\hat{X}$, in corrispondenza del quale viene determinata la variabile combinazione lineare:

$$z_{hij} = y_{hij} - \hat{R} x_{hij} \qquad [7.13]$$

a cui vengono applicate le formule [7.8], [7.10] e [7.12] per le stime della varianza canpionaria a livello di strato, per l'intera popolazione e per i domini di studio.

Così per le stime riferite all'intera popolazione si ha:

$$\hat{Z}_{hi} = \sum_{j=1}^{m_{hi}} w_{hij}\, z_{hij} = \hat{Y}_{hi} - \hat{R}\hat{X}_{hi}$$

$$\hat{Z}_h = \sum_{i=1}^{n_h} \hat{Z}_{hi} = \hat{Y}_h - \hat{R}\hat{X}_h$$

$$\hat{Z} = \sum_{h=1}^{L} \hat{Z}_h = \hat{Y} - \hat{R}\hat{X} = 0 \qquad [7.14]$$

e la stima della varianza campionaria è data da:

$$\hat{Var}\left(\hat{\bar{R}}\right) = \frac{n_h}{n_{h-1}} \sum_{i=1}^{n_h} \left(\hat{Z}_{hi} - \frac{\hat{Z}_h}{n_h} \right)^2 \qquad [7.15]$$

La [7.15] viene utilizzata per calcolare la varianza campionaria delle stime di medie, frequenze relative e percentuali e di rapporti, mediante opportuna definizione delle variabili che sono a numeratore e a denominatore del rapporto.

Così nel caso di una media la variabile a denominatore viene definita come una variabile di conto, una variabile, cioè, che assume sempre il valore 1 per ogni unità elementare del campione. In tal modo i valori $\hat{X}_{hi}$, $\hat{X}_h$ e $\hat{X}$ vengono a concidere con la somma di coefficienti di espansione nella PSU, nello strato e nell'intero campione.

Ovviamente nel caso di frequenza relative o percentuali, si dovrà procedere all'introduzione della variabile indicatrice della modalità a cui si riferisce la stima.

3. Il fattore del disegno e il rapporto di omogeneità

La procedura generalizzata calcola sia i valori della stima della varianza campionaria e delle altre statistiche che da questa possono essere derivate (errore standard, errore relativo e intervalli di confidenza), che quello del fattore del disegno campionario (deft).

Il fattore del disegno

Come è noto, il deft costituisce una misura sintetica della «bontà» del piano di campionamento adottato, e si ottiene rapportando l'errore standard calcolato per il disegno campionario effettivo (SE) a quello di un campionamento casuale semplice di uguale numerosità (SEO):

$$deft = \frac{SE}{SEO} \qquad [7.16]$$

dove SEO è dato da:

$$SEO = \sqrt{\frac{1-f}{m\sum w_{hij}} \sum z_{hij}^2 \, w_{hij}} \qquad [7.17]$$

in cui la somma è rispetto agli indici h, i, j e s'intende estesa a tutte le unità del campione appartenenti alla sottoclasse e al dominio in esame.

Risulta evidente che il deft non è altro che la radice quadrata dell'effetto del disegno (deff) introdotto nel quarto capitolo.

Il rapporto di omogeneità

Accanto al fattore del disegno, la procedura prevede il calcolo di un altro indicatore, chiamato «rapporto di omogeneità» (roh), che è legato al deft dalla seguente relazione (cfr. Kish, 1965):

$$deft^2 = 1 + (b-1)roh \qquad [7.18]$$

dove b è l'ampiezza media delle PSU, ed è data dal rapporto tra il numero delle unità elementari e il numero di PSU campionate.

Dalla [7.18] si evince che, per una determinata variabile, il fattore del disegno dipende dall'ampiezza media delle PSU e dal rapporto di omogeneità, il quale è una misura sintetica della correlazione all'interno delle PSU tra i valori osservati nelle unità elementari (correlazione intraclasse).

Il rapporto di omogeneità può essere espresso in funzione del deft:

$$roh = \frac{deft^2 - 1}{b-1} \qquad [7.19]$$

Quest'ultima equazione diventa indefinita per b = 1, e comunque per avere stime attendibili di roh è necessario che i valori di b siano sufficientemente elevati. Il programma CLUSTERS calcola i valori di roh solo nel caso in cui b > 5.

Ritornando all'ampiezza media delle PSU, per il campione totale essa è data da: b = m/n, dove n e m sono rispettivamente il numero complessivo delle PSU e delle unità elementari campionate. Nel caso di domini e sottoclassi il rapporto va calcolato tra il numero di PSU e il numero di unità elementari del campione comprese nel dominio o nella sottoclasse.

Con riferimento al campione totale, il valore di deft varia con la variabile oggetto di stima, in quanto si modifica il valore di roh. Lo stesso disegno campionario può quindi risultare molto efficiente (deft prossimo all'unità) per certe variabili e meno efficiente (deft elevato) per altre.

Una misura complessiva dell'influenza del piano di campionamento sugli errori standard delle stime è data dalla media aritmetica semplice dei deft relativi alle singole variabili (Verma, Scott e O'Muircheartaigh, 1980).

Il fattore del disegno per sottoclassi

Il disegno campionario ha un diverso impatto sull'errore standard delle stime a seconda che queste siano riferite al campione totale o a sottoclassi, e questo per due ragioni:

a) le unità elementari di una sottoclasse presentano, in genere, una maggiore omogeneità rispetto al totale delle unità campionate.

b) l'ampiezza media delle PSU è più bassa nelle sottoclassi rispetto al campione totale.

Questi motivi fanno sì che il deft delle sottoclassi assuma valori sempre inferiori di quelli del corrispondente deft calcolato per il campione totale.

Le sperimentazioni condotte sui risultati di numerose indagini campionarie (Verma, 1982), hanno, inoltre, evidenziato che, nel caso di sottoclassi distribuite uniformemente sulle PSU i valori di deft non si discostano molto dall'unità. Ciò sta ad indicare che, per le stime relative a sottoclassi, la varianza campionaria può essere ben approssimata da quella relativa ad un campionamento *casuale semplice*.

4. Stima degli effetti clustering e stratificazione

La formula per la stima della varianza campionaria utilizzata dalla procedura generalizzata, se da un lato ha l'indubbio vantaggio di essere molto semplice, in quanto richiede il calcolo della varianza tra le sole PSU, dall'altro, non consente di avere informazioni separate sugli effetti delle diverse componenti la struttura del campione.

Nel caso di disegni campionari a due o più stadi stratificati, la varianza campionaria delle stime differisce da quella di un campionamento casuale semplice per due fattori: l'effetto stratificazione e l'effetto clustering (o stadificazione).

L'effetto stratificazione comporta, almeno nei casi in cui questa è stata effettuata utilizzando criteri appropriati, una riduzione della varianza delle stime.

L'effetto clustering, in genere, produce un aumento della varianza campionaria a causa delle correlazioni esistenti all'interno delle unità relative ai diversi stadi di campionamento.

Per valutare gli effetti dei singoli fattori e delle loro interazioni occorrerebbe procedere ad una scomposizione della varianza campionaria nelle diverse componenti, operazione questa piuttosto complessa e per la quale non sempre sono disponibili le soluzioni esplicite.

Un metodo approssimato, ma molto semplice da utilizzare, è quello proposto da Verma ed altri per l'analisi dei piani di campionamento adottati dai diversi Paesi nel quadro dell'indagine mondiale sulla fecondità (Verma, Scott e O'Muircheartaigh, 1980).

Si supponga di aver effettuato un'indagine campionaria utilizzando un piano di campionamento a due stadi con stratificazione delle unità di primo stadio. L'errore standard di una stima viene calcolato ipotizzando di volta in volta che il campione sia stato generato:

- dal disegno campionario effettivamente adottato (SE)
- da un campionamento casuale semplice (SEO)
- da un campionamento a due stadi non stratificati (SE1)
- da un campionamento ad uno stadio stratificato (SE2)

I valori dei deft per i diversi disegni campionari sono dati da:

deft = SE/SEO fattore del disegno per il piano di campionamento effettivamente utilizzato

deft1 = SE1/SEO fattore del disegno per un piano di campionamento a due stadi senza stratificazione

deft2 = SE2/SEO fattore del disegno per un piano di campionamento ad uno stadio stratificato

Gli effetti clustering e stratificazione nel disegno campionario effettivamente adottato si ottengono dai seguenti rapporti:

effstr = deft/deft1 effetto della stratificazione nel campione a due stadi

effclu = deft/deft2 effetto clustering nel campione stratificato

Esempio 7.1 Si riporta un'applicazione del metodo ai risultati dell'indagine sulla salute del 1980 (Napolitano, Russo, Zannella, 1983). Il piano di campionamento utilizzato è a due stadi con stratificazione delle unità di primo stadio (i comuni) e campionamento casuale semplice delle unità di secondo stadio (le famiglie) che costituiscono grappoli di unità elementari di rilevazione (i componenti).

Si supponga di voler valutare gli effetti medi della stratificazione e del clustering su 5 gruppi di stime relative a 41 parametri riguardanti i diversi argomenti oggetto d'indagine e più precisamente:

1) otto parametri relativi a variabili socio-demografiche
2) undici relativi alle malattie acute e croniche in atto
3) sei relativi alle invalidità permanenti
4) sette relativi al ricorso ai servizi sanitari
5) nove relativi all'abitudine al fumo

Mediante il programma CLUSTERS, sono stati determinati per ciascuna stima i valori di deft, deft1 e deft2 e successivamente sono stati calcolati i valori medi per gruppo di variabili.

Tavola 7.1 – Valori medi del fattore del disegno per gruppi di variabili

gruppi di variabili	deft	deft1	deft2
1. socio-demografiche	1,460	1,473	1,338
2. malattie in atto	1,510	1,512	1,304
3. invalidità permanenti	1,392	1,393	1,380
4. ricorso ai servizi sanitari	1,531	1,534	1,281
5. abitudine al fumo	1,410	1,412	1,240
totale	1,469	1,856	1,309

Mediante i valori dei deft calcolati sono stati stimati gli effetti stratificazione e clustering:

Tavola 7.2 – Effetti stratificazione e clustering per gruppi di variabili

Gruppi di variabili	effstr	effclu
1. socio-demografiche	0,991	1,091
2. malattie in atto	0,999	1,158
3. invalidità permanenti	0,999	1,009
4. ricorso ai servizi sanitari	1,998	1,195
5. abitudine al fumo	1,999	1,137
totale	1,997	1,122

Dall'esame di questi valori si evince che:

a) la stratificazione dei comuni non sembra comportare una riduzione significativa degli errori standard, infatti sul complesso delle stime considerate si ha una riduzione media di appena lo 0,3% con un valore massimo dello 0,9% per le stime relative alle variabili socio-economiche;

b) l'adozione di un campionamento a due stadi al posto di un campionamento casuale semplice comporta un aumento medio dell'errore standard del 12,2%, con un massimo del 19,5% per le stime relative al gruppo «ricorso ai servizi sanitari» e un minimo del 9,1% per quelle relative al gruppo «socio-demografico».

Il procedimento descritto per un campionamento a due stadi con stratificazione delle unità di primo stadio, può essere facilmente esteso a disegni campionari più complessi.

5. Scomposizione della varianza campionaria

Il procedimento descritto nel precedente paragrafo consente di determinare gli effetti imputabili ai diversi fattori caratterizzanti il piano di campionamento che è stato utilizzato. Tuttavia, per chi deve programmare i disegni campionari può risultare più

utile la scomposizione della varianza tra le sue componenti, che può essere derivata dai valori del fattore del disegno.

Seguendo l'impostazione suggerita da Butcher, riportata nella discussione che accompagna il lavoro di Verma, Scott e O'Muirchertaigh (pp. 466-467), per un campionamento a due stadi si ha:

$$100\, deft^2 = m\left(\frac{p_1}{n} + \frac{p_2}{m}\right) = bp_1 + p_2 \qquad [7.20]$$

dove p_1 e p_2 sono le percentuali di varianza per elemento associate rispettivamente al primo e al secondo stadio di campionamento, e soddisfano la condizione:

$$p_1 + p_2 = 100 \qquad [7.21]$$

Una volta calcolato il deft è possibile mediante le equazioni [7.20] e [7.21] determinare i valori di p_1 e p_2:

$$p_1 = 100\,\frac{deft^2 - 1}{b - 1} \qquad [7.22]$$

dove b = m/n è il numero medio di unità elementari per PSU.

Per un campionamento a tre stadi valgono le seguenti relazioni:

$$100\, deft = n_3\left(\frac{p_1}{n_1} + \frac{p_2}{n_2} + \frac{p_3}{n_3}\right)$$

$$100\, deft1^2 = n_3\left(\frac{p_1 + p_2}{n_2} + \frac{p_3}{n_3}\right) \qquad [7.23]$$

$$p_1 + p_2 + p_3 = 100$$

dove con n_i e p_i sono state indicate rispettivamente la numerosità campionaria e la percentuale di varianza per elemento dell'i.mo stadio di campionamento (i = 1, 2, 3), e con deft1 il fattore del disegno per il piano di campionamento a due stadi ottenuto ignorando le unità di primo stadio.

Esempio 7.2 Per meglio chiarire quanto esposto in questo paragrafo si riporta un'applicazione ai dati dell'indagine campionaria sulle fecondità del 1979 (Zannella, 1982). Il piano di campionamento utilizzato è a tre stadi con stratificazione delle unità di pri-

mo stadio (i comuni), scelta sistematica delle unità di secondo stadio (le sezioni elettorali) ed estrazione casuale semplice senza reimmissione delle unità elementari (donne non nubili in età 18-44 anni). Di seguito si riportano le numerosità campionarie relative a ciascuno

comuni campione	$n_1 = 236$
sezioni elettorali campione	$n_2 = 1473$
donne intervistate	$n_3 = 5499$

Si supponga di voler stimare le percentuali di varianza campionaria per elemento imputabili a ciascuno stadio di campionamento, per le stime di 36 parametri riguardanti i diversi argomenti oggetto d'indagine e precisamente:

1) sette relativi alla nuzialità e all'esposizione al concepimento
2) dieci relativi alla fecondità
3) cinque relativi alle preferenze sul numero di figli
4) cinque relativi alla conoscenza di metodi contraccettivi
5) nove relativi all'uso di metodi contraccettivi

Mediante il programma CLUSTERS, per ciascuna stima sono stati determinati i valori del fattore del disegno per i seguenti piani di campionamento:

a) campionamento effettivamente utilizzato (deft)

b) campionamento a due stadi con stratificazione delle unità di primo stadio, costituite dalle sezioni elettorali (deft1)

Per non appesantire il testo vengono riportati i soli valori medi dei deft per gruppo di variabili:

Tavola 7.3 - Valori medi del fattore del disegno per gruppi di variabili

gruppi di variabili	deft	deft1
1. nuzialità	1,378	1,263
2. fecondità	1,799	1,479
3. preferenze	1,725	1,474
4. conoscenza	1,919	1,611
5. uso	1,939	1,566
totale	1,756	1,476

Sostituendo nella [7.23] a deft e deft1 i valori stimati, per ognuno dei gruppi considerati si ha un sistema di tre equazioni in tre incognite:

$$100\ \text{deft}^2 = 5499\left(\frac{p_1}{236}+\frac{p_2}{1473}+\frac{p_3}{5499}\right)$$

$$100\ \text{deft1}^2 = 5499\left(\frac{p_1+p_2}{1473}+\frac{p_3}{5499}\right)$$

$$p_1 + p_2 + p_3 = 100$$

che risolto permette di ottenere le percentuali di varianza cercate:

Tavola 7.4 - Percentuale di varianza per elemento associata a ciascuno stadio di campionamento

gruppi di variabili	p_1	p_2	p_3	totale
1. nuzialità	1,55	20,22	78,23	100,00
2. fecondità	5,36	38,08	56,56	100,00
3. preferenze	4,10	38,80	57,10	100,00
4. conoscenza	5,56	52,81	41,63	100,00
5. uso	6,68	46,46	46,86	100,00
totale	4,626	38,50	56,88	100,00

6. Adattamento della metodologia a particolari disegni campionari

La metodologia descritta è basata, come si è detto, sull'assunzione che vengano scelte due o più PSU per strato mediante estrazione con reimmissione. Spesso i disegni campionari che vengono adottati non soddisfano queste condizioni e le stime della varianza così ottenute possono risultare distorte.

Quando ciò si verifica è necessario far ricorso a particolari accorgimenti per eliminare, o almeno ridurre, questo inconveniente.

a) Campionamento senza reimmissione

Quasi sempre i piani di campionamento utilizzati nella pratica prevedono l'estrazione delle PSU senza reimmissione, con uguale probabilità o probabilità proporzionale all'ampiezza. La metodologia proposta fornisce in questi casi una sovrastima della varianza campionaria.

Per ridurre la distorsione è conveniente introdurre il coefficiente di correzione per popolazioni finite: $1-f_h$, dove f_h è il tasso di campionamento finale nello strato h. La [7.8] risulta così modificata:

$$\hat{Var}\left(\hat{Y}_h\right) = (1-f_h)\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\left(\hat{Y}_{hi}-\frac{\hat{Y}_h}{n_h}\right)^2 \qquad [7.24]$$

Nel caso di un campione proporzionale (f_h = costante), la correzione può essere introdotta alla fine delle eleborazioni, moltiplicando la varianza campionaria per la quantità (1-f) dove f è il tasso di campionamento complessivo.

Per campioni che utilizzano tassi di campionamento molto bassi, e ciò si verifica spesso nelle indagini su larga scala, il fattore di correzione può essere omesso in quanto assume valori molto vicini all'unità.

b) Selezione di una PSU per strato

Come più volte detto, spesso per aumentare l'efficienza del disegno campionario la stratificazione viene spinta ad un punto tale da rendere necessaria l'estrazione di una sola PSU per strato per contenere la numerosità del campione in primo stadio.

In questi casi non è possibile ottenere una stima corretta della varianza e una stima distorta per eccesso può essere ricavata mediante la tecnica del «collapsed strata».

Gli strati contenenti una sola PSU campione vengono raggruppati a due a due in modo da formare un certo numero di pseudo strati contenenti ciascuno due PSU campione, che vengono riguardate come una coppia di osservazioni indipendenti. Ovviamente se il numero complessivo delle PSU è dispari ci sarà uno pseudo strato formato da tre PSU e tutti gli altri da due.

Le modalità in base alle quali raggruppare gli strati devono essere stabilite prima della rilevazione e non devono tener conto dei valori osservati nel campione. Infatti se gli pseudo strati vengono formati mettendo assieme le PSU che presentano valori molto prossimi delle variabili rilevate, si ha una sottostima della varianza campionaria.

Sia g un generico pseudo strato formato dal collassamento degli strati h e k, e $M_g = M_h + M_k$ la sua popolazione complessiva. Si indichino con r ed s le PSU campionate nei due strati, con m_{hr} e m_{ks} il numero delle unità elementari rilevate nelle due PSU e con y_{hri} ed y_{ksj} i valori osservati.

Alle unità elementari delle due PSU sono associati i coefficienti di espansione originari w_{hri} e w_{ksj}.

Una corretta applicazione della metodologia proposta richiede che dopo il raggruppamento vengano calcolati i nuovi coefficienti di ponderazione mediante opportuno riproporzionamento dei precedenti, in modo che le quantità $2\,\hat{Y}^*_{hr}$ e $2\,\hat{Y}^*_{hs}$ forniscano ciascuna una stima del totale dello pseudo strato, dove:

$$\hat{Y}^*_{hr} = \sum_{i=1}^{m_{hr}} y_{hri}\; w^*_{hri}$$

$$\hat{Y}^*_{ks} = \sum_{j=1}^{m_{ks}} y_{ksj}\; w^*_{ksj} \qquad [7.25]$$

Così nel caso di selezione di una sola PSU per strato con probabilità proporzionale all'ampiezza, i coefficienti di ponderazione originali sono dati da:

$$w_{hri} = \frac{M_h}{m_{hr}}$$

$$w_{kri} = \frac{M_k}{m_{ks}} \qquad [7.26]$$

Mentre i coefficienti corretti per tener conto del raggruppamento sono dati da:

$$w^*_{hri} = w_{hri}\;\; \frac{M_h + M_k}{2\;M_h} = \frac{M_g}{2\;m_{hr}}$$

$$w^*_{ksj} = w_{ksj}\;\; \frac{M_h + M_k}{2M_k} = \frac{M_g}{2\;m_{mks}} \qquad [7.27]$$

Più in generale se lo pseudo strato g si ottiene dal raggruppamento di $n_g > 2$ strati originali, si ha:

$$w^*_{hri} = \frac{M_g}{n_g\;m_{hr}} \qquad [7.28]$$

c) Unità autorappresentative

Quando il piano di campionamento prevede la presenza di unità autorappresentative ciascuna di esse va considerata come uno strato a se stante e le PSU sono costituite dalle unità che vengono selezionate al loro interno nel primo stadio di campionamento.

Così nel caso delle indagini campionarie sulle famiglie i comuni di grandi dimensioni spesso costituiscono delle unità autorappresentative e le PSU vengono, quindi, ad essere formate dalle famiglie. In questo caso l'indice h sta ad indicare il comune autorappresentativo, l'indice i la generica famiglia campione, mentre j è l'indice che individua il componente all'interno della famiglia.

La presenza di unità autorappresentative non comporta inconvenienti teorici, ma può comportare notevoli difficoltà di cal-

colo, quando, come spesso accade, il numero delle PSU campione risulta piuttosto elevato.

È possibile *ridurre i* tempi di elaborazione raggruppando le unità di secondo stadio all'interno di ogni unità autorappresentativa, formando così delle pseudo PSU.

Come è stato detto alla fine del sesto capitolo, questo modo di procedere introduce una distorsione positiva nella stima della varianza campionaria. La distorsione è tanto più grande quanto maggiore è il numero delle unità di secondo stadio che formano ciascuna PSU.

Nel formare i raggruppamenti è pertanto necessario contemperare l'esigenza di ridurre i tempi di calcolo con la necessità di ottenere stime attendibili. Le sperimentazioni effettuate sui risultati di alcune indagini campionarie sulle famiglie, hanno mostrato che una dimensione accettabile dello pseudo PSU è di circa 10 famiglie.

CAPITOLO 8 - MODELLI PER LA PRESENTAZIONE DEGLI ERRORI CAMPIONARI

1. Premessa

Nel pubblicare i risultati di un'indagine campionaria occorre fornire agli utilizzatori tutte le informazioni necessarie per poter valutare l'attendibilità delle stime prodotte. Queste informazioni devono riguardare sia gli errori dovuti alla variabilità campionaria delle stime che quelli imputabili ad altre cause, quali: le risposte errate, le *mancate risposte*, ecc.

Limitatamente agli aspetti campionari, per consentire un uso corretto dei risultati sarebbe necessario pubblicare per ogni stima il corrispondente errore di campionamento.

È evidente che il calcolo e la presentazione degli errori campionari di tutte le stime pubblicate comporterebbe tempi e costi eccessivi e d'altra parte non risulterebbe di facile consultazione per il lettore. Ciò, inoltre, non esaurirebbe il problema della completezza dell'informazione, in quanto non sarebbero disponibili gli errori campionari delle stime ottenibili mediante successive elaborazioni.

Un modo per risolvere il problema di una presentazione esaustiva e concisa degli errori campionari è quello di ricorrere ad opportuni modelli che mettano in relazione l'errore relativo con il valore medio di uno stimatore:

$$RE\left(\hat{Y}\right) = g\left[E\left(\hat{Y}\right);\ a_1, a_2, \ldots, a_k\right] \qquad [8.1]$$

I parametri che compaiono nel modello vengono stimati utilizzando un certo numero di coppie di valori $\hat{Y}$ e $\widehat{RE}(\hat{Y})$ calcolati mediante uno dei metodi descritti nei precedenti capitoli.

La funzione:

$$\hat{RE}\left(\hat{Y}\right) = g\left(\hat{Y};\ \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_k\right) \qquad [8.2]$$

può essere utilizzata per calcolare l'errore relativo di una qualsiasi stima e, quindi, l'errore standard e tutti gli altri indicatori che da questo possono essere derivati.

Gli errori campionari possono venire tabulati per valori crescenti delle stime, così da fornire un'informazione di semplice e rapida consultazione.

Nei paragrafi che seguono verranno riportati i diversi modelli adottati nella pratica e gli standard che devono essere seguiti per la pubblicazione dei risultati di un'indagine campionaria.

2. Scelta del modello

Verranno dapprima descritti i modelli per la presentazione sintetica degli errori campionari delle stime di frequenze, che costituiscono i casi di gran lunga più importanti nelle indagini Istat, e successivamente quelli delle stime di totali e di medie.

a) Stima di una frequenza assoluta

Si indichino con N e P rispettivamente il numero delle unità e la frequenza relativa di una certa caratteristica nella popolazione, e con n la numerosità del campione.

Ricordando la relazione che intercorre tra l'effetto del disegno, la varianza campionaria effettiva e quella di un campionamento casuale semplice di uguale numerosità, si ha:

$$deff = \frac{n\ Var\left(N\hat{P}\right)}{(1-f)N^2\ P(1-P)} \qquad [8.3]$$

da cui si ricava:

$$Var\left(N\hat{P}\right) = \frac{(1-f)\ deff}{n}\ \left(N^2P - N^2P^2\right) \qquad [8.4]$$

Dividendo entrambi i membri per (NP) si ottiene:

$$RE^2 = -\frac{(1-f)\ deff}{n} + \frac{N(1-f)\ deff}{n\ NP} \qquad [8.5]$$

Ponendo:

$$a = -\frac{(1-f)deff}{n} \qquad [8.6]$$

$$b = \frac{N(1-f)deff}{n}$$

si ottiene il seguente modello:

$$RE^2 = a + b/NP \qquad [8.7]$$

Il modello [8.7] mette in evidenza che l'errore relativo può essere espresso come una funzione decrescente dell'ammontare totale della stima e per la sua utilizzazione bisogna procedere nel seguente modo:

1. calcolare la stima e l'errore relativo per numerosi parametri in corrispondenza di differenti domini territoriali e sottoclassi della popolazione;
2. calcolare le stime a e b dei parametri del modello mediante il metodo dei minimi quadrati ponderati, utilizzando come pesi i reciproci dei quadrati degli errori relativi (Russo, 1987);
3. tabulare gli errori relativi per valori crescenti delle stime delle frequenze assolute mediante la funzione:

$$\hat{RE} = \sqrt{\hat{a} + \frac{\hat{b}}{N\hat{P}}} \qquad [8.8]$$

Nel presentare gli errori campionari è preferibile riportare l'errore standard al posto di quello relativo, in quanto questo consente di determinare rapidamente gli estremi dell'intervallo di confidenza. L'errore standard si ottiene moltiplicando l'errore relativo per il valore della stima.

b) Stima di una frequenza relativa o di una percentuale

La [8.8] può essere utilizzata per determinare gli errori campionari delle stime di frequenze relative o di percentuali. Infatti, l'errore relativo di queste stime è uguale all'errore relativo delle corrispondenti frequenze assolute, mentre l'errore standard si ottiene moltiplicando l'errore relativo per il valore della percentuale o della frequenza relativa.

c) Stima del rapporto tra due frequenze

Nel caso in cui anche il denominatore è una frequenza stimata mediante i dati del campione, per il calcolo dell'errore standard deve essere utilizzato un altro procedimento. Siano Y ed X le stime di due frequenze assolute e R = Y/X il loro rapporto, se si ipotizza che R ed X sono incorrelati, ossia che Cov(R,X) = 0, approssimativamente vale la seguente relazione (Woodruff 1985, pp. 204-205):

$$\hat{RE}^2\left(\hat{R}\right) = \hat{RE}^2\left(\hat{Y}\right) - \hat{RE}^2\left(\hat{X}\right) \qquad [8.9]$$

Sostituendo agli errori relativi delle frequenze assolute i corrispondenti valori calcolati mediante la [8.8], si ha:

$$\hat{RE}\left(\hat{R}\right) = \sqrt{\frac{\hat{b}\left(1-\hat{R}\right)}{\hat{R}\ \hat{X}}} \qquad [8.10]$$

da cui si ricava:

$$\hat{SE}\left(\hat{R}\right) = \sqrt{\frac{\hat{b}\left(1-\hat{R}\right)\ \hat{R}}{\hat{X}}} \qquad [8.11]$$

Dalla [8.11] si evince che l'errore standard dipende sia dal valore del rapporto (R) che da quello della base (X) sulla quale il rapporto è stato calcolato e che a parità di R l'errore è tanto più piccolo quanto più grande è il valore della base.

La [8.11] gode, inoltre, dell'importante proprietà di essere simmetrica rispetto ad R, per cui l'errore standard del rapporto percentuale R è uguale all'errore standard del rapporto complementare R = 100 — R.

Per consentire un'utilizzo immediato la [8.11] può essere tabulata in corrispondenza di diversi valori di X e di R.

d) Modelli per gli errori relativi di medie e di totali

Nel caso della stima di un totale, studi empirici hanno evidenziato l'esistenza di una relazione tra l'errore relativo e l'ampiezza della stima, nel senso che l'errore relativo decresce all'aumentare della stima.

I modelli che trovano più frequente applicazione sono (Wolter, 1985, pag. 203):

$$RE^2 = a + \frac{b}{Y} \qquad [8.12]$$

$$RE^2 = a + \frac{b}{Y} + \frac{c}{Y^2} \qquad [8.13]$$

$$\log\left(RE^2\right) = a - b\ \log(Y) \qquad [8.14]$$

La stima dei parametri dei modelli e la successiva tabulazione degli errori viene effettuata utilizzando la stessa procedura descritta per le stime delle frequenze assolute.

L'errore relativo di una media è uguale a quello del totale corrispondente, mentre l'errore standard si ottiene moltiplicando l'errore relativo per il numero delle unità della popolazione cui la media si riferisce. È bene evidenziare che questi modelli non hanno alcuna giustificazione teorica, anche se la loro validità è stata verificata mediante le numerose applicazioni condotte sui risultati delle indagini campionarie dell'Istat.

Esempio 8.1 Per illustrare quanto fin qui esposto, viene riportata un'applicazione ai risultati dell'indagine campionaria sugli sbocchi professionali dei laureati.

Applicando la procedura descritta nel quarto capitolo sono stati calcolati gli errori di campionamento di un considerevole numero di stime di frequenze assolute riferite sia al totale dei laureati che a diversi domini di studio e sottoclassi.

I valori ottenuti con il calcolo diretto sono stati utilizzati per stimare i parametri del modello [8.7] con il metodo dei minimi quadrati ponderati, ottenendo i seguenti valori:

a = -0,000044759 e b = 6,45645799

L'indice di determinazione è risultato uguale a 0.8259, e ciò sta ad indicare un buon adattamento del modello alle stime ottenute con il calcolo diretto.

È stata, quindi, effettuata la tabulazione degli errori standard in corrispondenza di valori crescenti delle stime di frequenze assolute (tav. 8.1) e mediante la [8.11] quella degli errori standard delle stime di rapporti tra frequenze, per diversi valori del rapporto e della base (tav. 8.2).

Frequenza assoluta	Errore standard	Frequenza assoluta	Errore standard
10	8	900	76
20	11	1000	80
30	14	2000	113
40	16	3000	138
50	18	4000	158
60	20	5000	177
70	21	6000	193
80	23	7000	207
90	24	8000	221
100	25	9000	233
200	36	10000	245
300	44	20000	334
400	51	30000	392
500	57	40000	432
600	62	50000	459
700	67	60000	476
800	72	70000	482

Tavola 8.2 Errori standard di stime di rapporti tra frequenze assolute

Rapporto		Base del rapporto						
R	1-R	1000	2000	5000	10000	20000	50000	70000
0,1	99,9	0,254	0,180	0,114	0,080	0,057	0,036	0,030
0,2	99,8	0,359	0,254	0,161	0,114	0,080	0,051	0,043
0,3	99,7	0,439	0,311	0,197	0,139	0,098	0,062	0,053
0,4	99,6	0,507	0,359	0,227	0,160	0,113	0,072	0,061
0,5	99,5	0,567	0,401	0,253	0,179	0,127	0,080	0,068
0,6	99,4	0,621	0,439	0,278	0,296	0,139	0,088	0,074
0,7	99,3	0,670	0,474	0,300	0,212	0,150	0,095	0,080
0,8	99,2	0,716	0,506	0,320	0,226	0,160	0,101	0,086
0,9	99,1	0,759	0,537	0,339	0,240	0,170	0,107	0,091
1,0	99,0	0,799	0,565	0,358	0,253	0,179	0,113	0,096
2,0	98,0	1,125	0,795	0,503	0,356	0,252	0,159	0,134
3,0	97,0	1,371	0,969	0,613	0,433	0,306	0,194	0,164
4,0	96,0	1,575	1,113	0,704	0,498	0,352	0,223	0,188
5,0	95,0	1,751	1,238	0,783	0,554	0,392	0,248	0,209
6,0	94,0	1,908	1,349	0,853	9,603	0,427	0,270	0,228
7,0	93,0	2,050	1,450	0,917	0,648	0,458	0,290	0,245
8,0	92,0	2,180	1,541	0,975	0,689	0,487	0,308	0,261
9,0	91,0	2,300	1,626	1,028	0,727	0,514	0,325	0,275
10,0	90,0	2,411	1,705	1,078	0,762	0,539	0,341	0,288
20,0	80,0	3,214	2,273	1,607	1,016	0,719	0,455	0,384
30,0	70,0	3,682	2,604	1,841	1,164	0,823	0,521	0,440
40,0	60,0	3,936	2,783	1,968	1,245	0,880	0,557	0,470
50,0	50,0	4,018	2,841	2,009	1,270	0,898	0,568	0,480

3. Standard per la presentazione dei risultati

Ogni pubblicazione dell'Istat, in cui sono riportati i risultati di un'indagine campionaria, deve comprendere una nota metodologica in cui vengono descritti: il disegno campionario adottato, la metodologia utilizzata per la determinazione delle stime e la procedura seguita per il calcolo degli errori di campionamento.

Nella nota, inoltre, devono essere riportate le definizioni dei termini utilizzati, come ad esempio: *errore di campionamento, errore standard, intervallo di confidenza*, ecc.. Questi concetti devono essere illustrati mediante esempi numerici in modo da consentire al lettore una loro corretta utilizzazione.

Gli errori di campionamento devono essere presentati nel contesto dell'errore totale e occorre sempre evidenziare che essi costituiscono soltanto una componente di quest'ultimo, anche se per stime basate su campioni di dimensioni ridotte possono rappresentarne l'aspetto più importante.

Le informazioni sugli errori campionari non devono generare confusione nell'interpretazione dei risultati dell'indagine. L'obiettivo che si deve raggiungere con questo tipo d'informazione è quello di chiarire i limiti di affidabilità dei risultati e non di renderli meno intellegibili. Inoltre, la presentazione degli errori campionari deve essere effettuata in modo da facilitarne ed incoraggiarne l'utilizzazione. È, pertanto, preferibile fornire un'informazione approssimata ma facilmente utilizzabile che un'informazione esatta ma di difficile uso.

I modi di presentazione e il grado di dettaglio devono tener conto delle diverse categorie di utilizzatori.

Ricordando quanto detto nel capitolo introduttivo circa i potenziali utilizzatori, nel calcolo e nella presentazione degli errori di campionamento è opportuno seguire le seguenti regole:

1. Il calcolo degli errori campionari deve essere effettuato per un gran numero di stime di natura diversa (medie, totali, frequenze relative e assolute, rapporti), riferite sia all'intera popolazione che a domini territoriali e sottoclassi di ampiezza variabile.
2. In una tavola devono essere riportate le principali stime con l'errore standard, l'errore relativo e l'intervallo di confidenza. Per la costruzione di quest'ultimo va utilizzato lo stesso livello di fiducia per tutte le stime e, tenuto conto di quanto raccomandato a livello internazionale (Gonzales et al., 1975), è opportuno che tale livello sia uguale al 95%.
3. Deve essere consentita la valutazione dell'errore campionario per una qualunque stima, mediante la pubblicazione di tabelle in cui sono riportate le stime e gli errori aspettati. Deve essere riportata la metodologia seguita per la costruzione delle tabelle sintetiche e ne deve essere illustrata l'utilizzazione mediante esempi.
4. Per permettere al lettore una immediata valutazione dell'attendibilità delle stime pubblicate, è opportuno inserire nelle tavole, in cui compaiono stime con errori elevati, dei caratteri speciali o delle note a fondo pagina. Ad esempio, come suggerisce Verma (1982, pag. 49), si potrebbero segnalare con un asterisco, o riportare tra parentesi, le stime con un errore relativo superiore ad un limite prestabilito. Ovviamente tale limite può variare da indagine ad indagine, tenuto conto del livello di precisione delle stime che è stato scelto per programmare il campione.
5. Infine, per consentire agli esperti di tecnica dei campioni di acquisire le informazioni necessarie per la valutazione del disegno campionario utilizzato e per la programmazione di successivi campioni, è opportuno riportare una tavola con i valori dell'effetto del disegno per le principali stime oggetto d'indagine in corrispondenza di domini territoriali e sottoclassi di diversa dimensione.

APPENDICE 1 – IL PROGRAMMA CLUSTERS

1. Caratteristiche generali

Il CLUSTERS è un insieme di programmi scritti in linguaggio FORTRAN IV per il calcolo degli errori di campionamento per campioni ad uno o più stadi semplici e stratificati.

La versione che viene descritta è quella attualmente operante nell'Istituto ed è stata messa a punto da Verma e Pearce nel 1977 per conto della World Fertility Survey (WFS, TECH. 770 1978). Gli stessi autori hanno predisposto nel 1986 una nuova versione che è stata recentemente acquistata dall'Istituto, ma non ancora implementata. Quest'ultima versione è più flessibile della precedente ma ne mantiene sostanzialmente le caratteristiche, sia per quanto riguarda la metodologia utilizzata che l'input e l'output previsti.

Il programma è compatibile con i sistemi IBM 360/370 e con i sistemi operativi CMS/OS/DOS; attualmente lavora sotto CMS.

La procedura prevede come input una serie di schede parametro (KINPUT) predisposte dall'utilizzatore, con le quali vengono descritti il formato dei dati di base, la struttura del campione e le stime per le quali si richiede il calcolo degli errori standard.

Il programma CLUST1 legge le schede parametro e controlla la correttezza della loro compilazione. Se non ci sono errori nelle schede e lo spazio di memoria è sufficiente per l'elaborazione dei dati, il programma legge il file dei dati di base (KDATA) e accumula i totali che verranno utilizzati dal programma CLUST2.

Il programma CLUST1 prevede come output tre files:

a) KPRINT è formato da records lunghi 133 (compreso il carattere di controllo), contiene la lista delle schede parametro e le informazioni sul campione;

b) KCNTRL contiene le informazioni di controllo necessarie per il programma CLUST2;

c) KSUMS contiene le somme utilizzate da CLUST2 per calcolare gli errori di campionamento ed ha le seguenti caratteristiche:

RECFM = VB
LRECL = 4 * (4 + 8 * (1 + NVARI)) + 4
BLKSIZE = 4 + K * LRECL

dove NVARI è il numero di schede parametro di tipo «VARI» utilizzate per il calcolo degli errori e K è il fattore di bloccaggio scelto (K è un intero positivo).

Il file KSUMS, generato da CLUST1, deve essere ordinato in modo particolare prima di essere letto dal programma. Il CLUSTERS non prevede un programma specifico per l'ordinamento ma l'utilizzazione di un programma di utility, pertanto il sort dovrà essere effettuato sotto OS/USI fornendo i seguenti parametri:

SORT FIELDS = (5, 16, BI, A)

L'output è costituito dal file KAREA, che è identico a KSUMS tranne che nell'ordinamento.

Infine il programma CLUST2 legge il file KCNTRL prodotto da CLUST1 e il file KAREA con le somme ordinate e calcola gli errori di campionamento richiesti.

I risultati vengono forniti in due forme:

a) mediante un file di stampa (KPRINT) formato da records lunghi 133 (compreso il carattere di controllo);
b) un file (KTAPE) su disco o su nastro da utilizzare per successive elaborazioni e con le seguenti caratteristiche:

RECFM = VB
LRECL = 72
BLKSIZE = 4 + 72 * K

dove K è il fattore di bloccaggio scelto.

I file KINPUT, KPRINT, KCNTRL e KTAPE sono sempre di modeste dimensioni per cui se ne può prevedere l'ingresso o l'uscita su disco.

Il file KDATA dei dati di base è in genere di grosse dimensioni in quanto è costituito da tanti records quante sono le unità elementari di osservazione (dalle 80.000 alle 300.000 nelle indagini sulle famiglie attualmente effettuate), con una lunghezza di centinaia di caratteri, per cui la sua lettura deve essere prevista da nastro.

Le dimensioni del file KSUMS dipendono dal numero delle unità di primo stadio (PSU) e dal numero delle variabili e sottoclassi per le quali è richiesto l'errore di campionamento. Il numero dei records è dato da:

$$NRK = (1 + NCLAS) * NPSU + 1$$

dove NCLAS è il numero di schede parametro di tipo 'CLAS' predisposto e NPSU è il numero di unità di primo stadio, mentre la lunghezza di ciascun record è data, come scritto in precedenza, da:

$$LRECL = 4 * (4 + 8 * (1 + NVARI)) + 4$$

Ad esempio per l'indagine sulla salute del 1983 in cui NPSU = 15863 volendo calcolare gli errori di campionamento utilizzando 40 schede 'VARI' e 10 schede 'CLAS', si ha:

$NRK = (1 + 10) * 15863 + 1 = 174.494$

$LRECL = 4 * (4 + 8 * (1 + 40)) + 4 = 1332$

Pertanto, anche per KSUMS deve essere prevista l'uscita su nastro.

Da quanto esposto risulta evidente che l'impiego del CLUSTERS sotto CMS comporta l'utilizzazione simultanea di due unità nastro, almeno per le indagini di medie e grandi dimensioni come quelle attualmente condotte sulle famiglie. Soltanto nel caso di campioni di ridotte dimensioni con un numero limitato di PSU e il calcolo degli errori è eichiesto per poche variabili e sottoclassi, è possibile ricorrere ad una sola unità nastro prevedendo l'uscita del file KSUMS su disco.

Di seguito sono riportati i numeri simbolici assegnati ai files di input-output dopo l'implementazione del CLUSTERS e che consentano la loro identificazione da parte dei programmi CLUST1, CLUST2 e sort:

file	n. ident.	I/O
	programma CLUST1	
KINPUT	1	input disco
KDATA	2	input disco o nastro
KPRINT	4	output disco
KCONTRL	5	output disco
KSUMS	6	output disco o nastro
	programma SORT	
KSUMS	1	input disco o nastro
Parametri	2	input disco
Karea	7	output nastro
	programma CLUST2	
KCNTRL	1	input disco
KAREA	2	input nastro
KPRINT	7	output disco
KTAPE	6	output nastro

Un'ultima considerazione va svolta sul 'work space' necessario per far girare i programmi. L'ammontare dello spazio di lavoro dipende soltanto dal numero delle variabili delle sotto-

classi e dai domini territoriali per i quali si richiede il calcolo degli errori di campionamento. Nella versione implementata lo spazio di lavoro disponibile è di 10.000 parole sia in CLUST1 che in CLUST2.

I programmi CLUST1 e CLUST2 prevedono la stampa dello spazio di lavoro usato e nel caso in cui quello disponibile non è sufficiente viene prodotto un messaggio di errore e precisamente il codice 111 nel programma CLUST1 e CLUST2.

Per l'esecuzione del programma occorre fornire al CLUSTERS una serie d'informazioni necessarie per la lettura del file dei dati di base, per la specificazione della struttura del campione e dei calcoli che devono essere effettuati.

Queste informazioni vengono date sotto forma di schede parametro, ciascuna lunga 80 caratteri, che devono essere predisposte nel seguente ordine:

TITL(e)	*contiene il titolo che verrà* usato per l'output
FORM(at)	descrive il formato che deve essere usato per la lettura del file dei dati di base
PROB(lem)	dà le informazioni sulla struttura del campione e i calcoli da effettuare
FACT(or)	riporta il fattore di scala da applicare ai pesi e le indicazioni relative al record indicante la fine dei dati
CLAS(s)	contiene il nome di una sottoclasse o i nomi di una coppia di sottoclass
VARI(able)	definisce il nome della variabile ed indica se viene letta direttamente dal record e in quale campo o se deve essere ricodificata, contiene gli eventuali valori eccezionali che devono essere esclusi dai calcoli
RECO(de)	specifica la ricodifica che deve essere usata per definire una sottoclasse o una variabile
DOMA(in)	indica il nome da utilizzare per ciascun dominio che è stato specificato
AREA	definisce aree di riferimento, PSU, strati, domini e i pesi, quando questi non vengono letti direttamente sui dati di base.

Ciascuna scheda parametro riporta nelle prime quattro posizioni il nome che caratterizza il tipo di scheda, mentre il tracciato relativo alle restanti 76 posizioni è variabile da scheda a scheda.

Nei prossimi paragrafi, dopo aver descritto la struttura del file dei dati di base, verranno illustrati i contenuti e le modalità di compilazione delle schede parametro.

2. Struttura del file dei dati di base

Il file dei dati di base deve essere formato da records di lunghezza fissa leggibili in FORTRAN, un record per ogni unità elementare di rilevazione.

Non possono essere utilizzati files gerarchici.

Il file deve terminare con un record fittizio, utilizzato per indicare la fine dei dati. Questo record è formato da tutti blanks tranne che in uno specifico campo, dove viene riportato il valore indicante la fine dei dati. Questo valore deve apparire nel campo specificato solo nell'ultimo record del file, altrimenti si ha una fine anticipata delle elaborazioni. Ad esempio se nelle prime due colonne è riportato il codice di provincia che assume i valori da 1 a 95, si può utilizzare il valore 99 da leggere in questo campo (valore che non è presente in nessun altro record) per indicare la fine dei dati. Il record fittizio con cui termina il file presenterà quindi il valore 99 nel primo campo e blank in tutti gli altri.

Ad ogni record corrisponde l'insieme completo dei dati rilevati per una delle unità del campione. Un record risulta quindi costituito da un certo numero di campi nei quali sono riportati i valori delle variabili rilevate (tanti campi quante sono le variabili) e deve contenere soltanto valori numerici.

Nel record devono essere riportate oltre alle variabili di rilevazione anche quelle relative alla struttura del campione, ed in particolare:

a) il numero d'ordine delle unità di primo stadio cui l'unità elementare appartiene;
b) il codice di strato, dopo l'eventuale raggruppamento degli strati originali nel caso sia stata selezionata una sola unità per strato;
c) il coefficiente di ponderazione da assegnare a ciascuna unità elementare per il riporto dei dati all'universo.

Come si vedrà meglio nel prossimo paragrafo in alcuni casi è possibile che queste informazioni non siano contenute nei records individuali ma che vengano acquisite dalla lettura delle schede 'AREA'.

Nel file dei dati che viene predisposto per l'elaborazione delle tavole di pubblicazione è sempre previsto il coefficiente di ponderazione, mentre, generalmente, non sono previsti il nume-

ro d'ordine dell'area di riferimento (o della PSU) e il codice di strato, almeno secondo le modalità richieste dal CLUSTERS. È necessario, quindi, procedere all'inserimento di queste informazioni mediante un opportuno programma.

Poiché il file deve risultare ordinato per area di riferimento (o per PSU quando questa coincide con l'area di riferimento), si deve procedere all'ordinamento utilizzando una procedura SORT.

Ai fini dell'elaborazione non è necessario che vengano lette tutte le informazioni contenute nel file dei dati di base, ma solo quelle che intervengono nel calcolo degli errori richiesti.

3. Schede parametro

a) Il titolo

La prima scheda che deve essere predisposta è quella contenente il titolo del lavoro, e ha il seguente tracciato:

Scheda 1 'TITL' (e)

Campo	Colonne	Descrizione
a	1-4	TITL
b	11-78	Titolo dell'elaborazione

b) il formato dei dati

La lettura dei dati di base viene effettuata dal CLUSTERS attraverso la scheda 'Form' in cui sono riportati i campi che devono essere letti e il loro formato. A volte per la lettura del formato sono necessarie due o più schede 'FORM', in questo caso le schede successive alla prima non riportano la dicitura 'FORM' nelle colonne 1-4:

Scheda 2 'FORM' (at)

Campo	Colonne	Descrizione
	prima scheda	
a	1-4	FORM
b	8	Numero sk necessarie
c	11-78	formato
	seconda scheda	
a	1-4	blank
b	8	blank
c	11-78	formato
	n.ma scheda	
a	1-4	blank
b	8	blank
c	11-78	formato

c) la struttura del campione

Per effettuare le elaborazioni occorre fornire al programma, mediante un apposita scheda parametro (PROB), le informazioni relative alla struttura del campione (PSU, strati e coefficiente di ponderazione) e ai calcoli da effettuare (numero di sottoclassi, variabili e domini territoriali).

La PSU cui l'unità elementare appartiene, può essere identificata dal CLUSTERS in due modi:

a) direttamente sul record individuale mediante la lettura di un apposito campo;
b) dall'esterno del file mediante la lettura delle schede AREA, una per ciascuna area di riferimento.

Le aree di riferimento sono delle unità intermedie di campionamento, in cui vengono suddivise le singole PSU e sono caratterizzate dal fatto che in ciascuna di esse le unità elementari devono presentare lo stesso coefficiente di ponderazione. In particolare quando si ha una sola area per PSU questa viene a coincidere con l'area di riferimento, e questo è un caso molto frequente nella pratica.

Il numero d'identificazione dell area di riferimento può essere riportato in un certo campo del record individuale, oppure può essere determinato attraverso i valori minimo e massimo del numero d'ordine delle unità elementari che sono comprese nell'area. In questo caso le unità elementari devono essere ordinate per area di appartenenza.

La scheda parametro PROB(lem) ha la seguente struttura:

Scheda 3 'PROB' (lem)

Campo	Colonne	Descrizione
a	1-4	PROB
b	6-10	numero di campi che devono essere letti

segue: **Scheda 3 'PROB' (lem)**

Campo	Colonne	Descrizione
c	11-15	numero di sk CLAS
d	16-20	numero di sk VARI
e	21-25	numero di AREE nel campione
f	26-30	identificazione dell'AREA da un campo input = 0 dalle sk AREA = 2
g	31-35	numero del campo con il nº ident. dell'area soltanto se il campo f = 0 altrimenti blank
h	36-40	identificazione della PSU da un campo input = 0 PSU = area di rifer. = 1 dalle sk AREA = 2
i	41-45	numero del campo con il nº ident. della PSU soltanto se il campo h = 0 altrimenti blank
j	46-50	identificazione dello strato da un campo input = 0 coppie di PSU adiacenti = 1 dalle sk AREA = 2
k	51-55	numero del campo con il nº ident. dello strato soltanto se il campo j = 0 altrimenti blank
l	56-60	identificazione del coeff. di ponderazione da un campo input = 0 dati non ponderati = 1 dalle sk AREA = 2
m	61-65	numero del campo contenente il peso soltanto se il campo k = 0 altrimenti blank
n	66-70	lettura del numero del dominio da un campo input = 0 non ci sono domini = 1 dalle sk AREA = 2
o	71-75	numero del campo input contenente il dominio soltanto se il campo n = 0 altrimenti blank
p	76-80	numero di sk DOMA soltanto se ci sono domini altrimenti blank

d) la fine dei dati e i valori anomali

Per definire il record della fine dei dati si utilizza la scheda FACT, nella quale va anche riportato l'eventuale fattore di scala per il quale vanno moltiplicati i pesi e il numero di records errati che possono essere accettati senza interrompere l'elaborazione.

La scheda FACT ha il seguente tracciato:

Scheda 4 'FACT' (or)

Campo	Colonne	Descrizione
a	1-4	FACT
b	11-20	fattore di scala da applicare ai pesi
c	21-25	numero del campo contenente il valore indicante la fine dei dati
d	26-30	valore indicante la fine dei dati
e	31-35	numero di rekords invalidati

e) le sottoclassi

Per sottoclasse del campione o della popolazione di riferimento s'intende un insieme di unità elementari, appartenenti anche a PSU e strati differenti, caratterizzato dal presentare la stessa modalità di un carattere o le stesse modalità di più caratteri rilevati. Nell'esempio sono state considerate due sottoclassi: i maschi e le femmine; altri esempi di sottoclassi sono dati da particolari gruppi di età del titolo di studio od anche i disoccupati in età 14-29 anni, etc.

È bene evidenziare che per il calcolo degli errori di campionamento non è necessario che le sottoclassi siano esaustive e disgiunte, per cui è possibile definire delle sottoclassi utilizzando soltanto alcune modalità di un carattere e due sottoclassi possono avere anche unità elementari in comune (sottoclassi sovrapposte).

Il programma CLUSTERS legge i valori di uno o più campi, specificati nelle schede 'CLAS' in cui vengono definite le sottoclassi, e determina se una unità elementare appartiene o no alla sottoclasse definendo una nuova variabile che assume il valore 1 se l'unità vi appartiene e 0 se non vi appartiene.

La scheda CLAS ha il seguente tracciato:

Scheda 5 'CLAS' (s)

Campo	Colonne	Descrizione
a	1-4	CLAS
b	9-12	nome di una sottoclasse
c	17-18	numero di sk RECO usate per la sottoclasse
d	19-22	nome di una sottoclasse
e	27-28	numero di sk RECO usate per la sottoclasse

La scheda CLAS è predisposta per due sottoclassi, se viene riportata una sola sottoclasse i campi d ed e vanno lasciati vuoti. Nel caso di due sottoclassi l'errore di campionamento viene fornito anche per la differenza tra le due sottoclassi.

Ogni scheda CLAS è seguita dalle schede RECO che descrivono la sottoclasse o le sottoclassi elencate nella scheda CLAS.

Ciascuna sottoclasse viene identificata mediante il nome su quattro posizioni riportato nella scheda CLAS. Poiché nel presentare i risultati quattro caratteri non sono in genere sufficiente per identificare le sottoclassi, è quindi consigliabile utilizzare delle sigle e successivamente rielaborare l'output del CLUSTERS in modo da poter scrivere i nomi in chiaro.

Ad ogni scheda CLAS seguono una o più schede RECO per la ricodifica delle sottoclassi.

f) Le variabili

Le variabili possono essere lette direttamente da un campo del record input che deve essere specificato nelle schede VARI, o possono essere definite attraverso la ricodifica dei valori letti in uno o più campi. Nella scheda VARI è previsto inoltre un dispositivo per indicare un numero di valori eccezionali, che se presenti in un record portano all'esclusione del caso dai calcoli che devono essere effettuati. Così se si vogliono escludere le mancate risposte fra i valori eccezionali va inserito il codice corrispondente alla modalità 'nessuna risposta'. Se si tratta di una stima riferita ad una particolare popolazione si possono escludere i casi che non appartengono a quella popolazione, ad esempio se si vuole calcolare il numero di sigarette per fumatore si possono escludere dai calcoli i non fumatori di sigarette.

Le frequenze relative sono trattate come medie definendo una nuova variabile che assume il valore 1 se il caso presenta la caratteristica per la quale si vuole calcolare la frequenza e il valore 0 altrimenti.

Le percentuali possono essere ottenute codificando la nuova variabile con i valori 100 e 0.

Il rapporto tra due variabili viene specificato mediante due schede VARI, una per la variabile a numeratore e l'altra per quella a denominatore.

In genere il calcolo degli errori riguarda medie o frequenze, ciascuna delle quali richiede una sola scheda VARI per la variabile relativa al numeratore, essendo il denominatore semplicemente la somma dei coefficienti di ponderazione delle unità elementari. La scheda VARI ha il seguente tracciato:

Scheda 6 'VARI' (able)

Campo	Colonne	Descrizione
a	1-4	VARI
b	7-10	Va riportato il nome della variabile, se si tratta del denominatore di una variabile rapporto va lasciato in bianco
c	11-15	numero del campo input in cui viene presa la variabile, altrimenti blank o zero
d	16-20	numero di sk RECO utilizzate se la variabile è creata mediante ricodifica, altrimenti blank o zero
e	21-25	numero di valori della variabile che devono essere esclusi dal calcolo
f	26-30	valori che devono essere esclusi dal calcolo
g	31-35	"
h	36-40	"
i	41-45	"
j	46-50	"
k	51-55	"
l	56-60	"
m	61-65	"
n	66-70	"
o	71-75	"
p	76-80	"

g) la ricodifica di variabili e sottoclassi

Le schede RECO vengono utilizzate per la costruzione delle sottoclassi e delle variabili, quando queste non possono essere ricavate direttamente dalla lettura di un campo input.

Il programma CLUSTERS prevede quattro diversi procedimenti di ricodifica con i quali possono venir risolti i più comuni problemi che si riscontrano nella pratica, per la cui illustrazioni si rimanda al manuale sull'uso del CLUSTERS (Verma and Pearce 1970, op. cit.).

Ciascuna scheda RECO dovrà seguire la rispettiva scheda CLAS o VARI.

h) i domini di studio

Spesso il calcolo degli errori di campionamento viene effettuato oltre che per l'intero territorio anche per particolari suddivisioni (province, regioni, ripartizioni). Nella terminologia del campionamento, queste sottoclassi di popolazione assumono la denominazione di domini di studio. I domini differiscono dalle sottoclassi vere e proprie per il fatto che i primi devono essere non sovrapposti *ed esaustivi*, nel senso che ciascuna unità del campione può appartenere ad un solo dominio, e i domini considerati devono comprendere tutte le unità del campione.

La presenza dei domini viene specificata negli ultimi tre campi della scheda PROB in cui viene indicato se il dominio viene letto da un campo input e in caso affermativo il numero del campo e il numero di schede DOMA utilizzate per descrivere i domini.

Ciascun dominio viene descritto mediante una scheda DOMA, che ha il seguente tracciato:

Scheda 8 'DOMA' (in)

Campo	Colonne	Descrizione
a	1-4	DOMA
b	10-21	nome del dominio
c	29-30	Numero di modalità (max 10) assegnate al dominio
d	31-35	
e	36-40	codici assegnati al dominio (il numero
.	.	deve essere uguale a quello definito
.	.	in c)
.	.	
p	76-80	

i) le aree di riferimento

Quando il disegno campionario è a più stadi, la struttura del campione può essere definita mediante unità territoriali successive al primo stadio di campionamento (aree).

Ciascuna «area» può essere letta direttamente da un campo input, o quando i records del file input sono ordinati, mediante il numero di ordine minimo e massimo dei records che cadono nell'area. Per ciascuna area viene predisposta una scheda «AREA» che ha il seguente tracciato:

Scheda 9 'AREA'

Campo	Colonne	Descrizione
a	1-4	AREA
b	6-10	Numero d'identificazione dell'area, se il campo f della sk PROB è uguale a zero; Estremo superiore del n. d'ordine dei casi che cadono nell'AREA, se il campo f della sk PROB è uguale a 2
c	11-15	Numero della PSU di appartenenza dell'AREA se il campo h della sk PROB è uguale a 2 altrimenti blank o zero
d	16-20	Numero dello strato cui appartiene l'AREA se il campo j della sk PROB è uguale a 2 altrimenti blank o zero
e	21-25	Numero del dominio cui appartiene l'AREA se il campo n della sk PROB è uguale a 2 altrimenti blank o zero
f	26-30	Valore del peso che deve essere attribuito ai casi di quest'AREA se il campo 1 della sk PROB è uguale a 2, altrimenti blank o zero

4. Descrizione dell'output

L'output è costituito da 3 liste:

- La prima riproduce le schede parametro; se ci sono errori le schede errate vengono stampate per ultime e sono seguite da un messaggio relativo all'errore.
- La seconda riporta la struttura del campione, ossia il numero identificativo dell'AREA, delle PSU dello strato e del dominio territoriale, il valore del peso e il numero di osservazioni che cadono in ciascuna AREA.
- L'ultima è quella contenente i risultati relativi al calcolo degli errori di campionamento.

I risultati stampati a blocchi, uno per ciascuna classe (o coppie di sottoclassi) indicate nelle schede CLAS, dove il primo blocco è relativo all'intero campione.

Ciascun blocco è suddiviso in sezioni, una per ciascun dominio territoriale.

In definitiva per ciascun dominio territoriale (compreso il dominio totale) e per ciascuna sottoclasse (compreso l'intero campione) viene stampata una riga relativa alle caratteristiche della sottoclasse e una per ogni variabile di cui è stato richiesto il calcolo dell'errore campionario.

Ciascuna riga contiene le seguenti quantità:

R	= Valore della stima (media, percentuale, etc)
SE	= Errore standard di R per il disegno campionario effettivamente utilizzato
N	= Numero delle osservazioni non ponderate usate per il calcolo
WN	= Numero delle osservazioni ponderate usate per il calcolo
SER	= Errore standard di R nel caso di un piano di campionamento semplice della stessa numerosità
SD = SER· N	Deviazione standard
DEFT = SE/SER	= Effetto del piano di campionamento
ROH	= Rapporto di omogeneità, viene stampato soltanto se il numero medio di osservazioni per PSU non è inferiore a 6
SE/R	= Errore relativo
R – 2SE	= Estremo inferiore dell'intervallo di coefficienza
R + 2SE	= Estremo superiore dell'intervallo di coefficienza (P = 95%)
B	= Numero medio di osservazioni non ponderate per PSU
CV	= Coefficiente di variazione del numero di osservazioni per PSU, viene riportato soltanto nella riga relativa alle caratteristiche della sottoclasse.

APPENDICE 2 – APPLICAZIONE DELLA PROCEDURA GENERALIZZATA AI RISULTATI DELLA SECONDA INDAGINE SULLA SALUTE

1. Obiettivi dell'indagine e disegno campionario utilizzato

Per rendere più chiare le modalità d'impiego della procedura generalizzata attualmente utilizzata dall'Istat per il calcolo degli errori di campionamento, viene riportata una sua applicazione ai dati della seconda indagine sulle condizioni di salute della popolazione.

Per l'esecuzione dell'indagine è stato impiegato un disegno campionario a due stadi con stratificazione delle unità di primo stadio (i comuni) e selezione di una sola PSU per strato con probabilità proporzionale all'ampiezza demografica. All'interno di ciascuna PSU campione, le unità di secondo stadio (le famiglie) sono state estratte mediante un campionamento casuale semplice senza reimmissione.

La stratificazione dei comuni, effettuata in base all'ampiezza demografica, ha comportato la formazione di strati costituiti da un solo comune. In questi strati (autorappresentativi) il campione è ad un solo stadio e le PSU sono costituite dalle famiglie.

La rilevazione, condotta mediante intervista diretta, ha interessato tutti i componenti delle famiglie campione, i quali, pertanto, costituiscono le unità elementari. La struttura del campione su cui è stata condotta l'indagine è sintetizzata dal seguente prospetto:

Unità campione	Tipo di comune		Totale
	autorappr.	non autorappr.	
Comuni	429	735	1.164
Famiglie	16.307	14.714	31.021
Individui	47.329	42.436	89.765

Le informazioni raccolte hanno riguardato:

- le caratteristiche socio-demografiche (sesso, età, stato civile, titolo di studio, attività lavorativa, ecc.);
- le condizioni di salute (stato di salute al momento dell'intervista, malattie acute, malattie cronico degenerative, invalidità permanenti);

- il ricorso ai servizi sanitari (cure mediche, accertamenti diagnostici, ricoveri ospedalieri);
- il consumo di farmaci;
- alcune abitudini connesse con lo stato di salute (fumo, consumo di bevande alcoliche, caffè, ecc.).

I risultati dell'indagine sono disponibili su nastro sotto forma di un file CMS costituito da 89.765 record contenenti:

— i codici identificativi delle unità di rilevazione;
— il codice di strato;
— i valori delle variabili oggetto d'indagine;
— il coefficiente di espansione.

Per limitare il numero delle tavole prodotte dall'output del CLUSTERS, il calcolo degli errori campionari è stato effettuato per un numero ridotto di stime, sottoclassi e domini.

Sono state considerate 2 sottoclassi (maschi e femmine), 3 domini territoriali (nord, centro e sud) e le stime dei seguenti parametri:

1. SALU = numero di persone in non buono stato di salute nelle 4 settimane precedenti l'intervista per 100 abitanti
2. GIOM = numero medio di giornate di malattia per persona in non buono stato di salute
3. DIAB = numero di casi di diabete per 1000 abitanti
4. IPER = numero di casi di ipertensione per 1000 abitanti
5. BRON = numero di casi di bronchite per 1000 abitanti
6. ARTR = numero di casi di artrosi per 1000 abitanti
7. RICO = numero di persone che nel corso dell'anno sono state ricoverate in ospedale per 100 abitanti
8. DEGM = numero medio di giorni di degenza per persona ricoverata in ospedale
9. FUMA = numero di fumatori per 100 abitanti
10. SIGM = numero medio di sigarette fumate giornalmente

2. Predisposizione del file dei dati di base

Come riportato nell'appendice 1, il calcolo degli errori di campionamento mediante il programma CLUSTERS richiede come input:

a) il file dei dati di base
b) il file delle schede parametro

Per poter applicare la metodologia su cui si basa la procedura CLUSTERS è stato necessario raggruppare gli strati dei comuni non autorappresentativi, in modo da avere almeno due

comuni campione per strato. Ciò ha comportato l'inserimento, nel file dei dati di base, di un nuovo campo contenente il codice di strato dopo raggruppamento.

Per contenere il numero delle PSU, che risulta molto elevato per la presenza di comuni autorappresentativi (NPSU = 16307 + 735 = 17042), sono state formate delle pseudo PSU all'interno di ciascun comune autorappresentativo mediante raggruppamento casuale delle famiglie campione.

La costruzione delle pseudo PSU è stata effettuata in modo di avere almeno 2 unità primarie per comune autorappresentativo e in ciascuna di esse non più di 10 famiglie campione. Sono state così formate 1978 pseudo PSU, con una composizione media di circa 8 famiglie per PSU. Il numero complessivo di PSU, dato dalla somma del numero di comuni non autorappresentativi (735) e del numero di pseudo PSU, si è così ridotto a 2713. Si è provveduto, quindi, all'inserimento di un ulteriore campo contenente il numero d'ordine (da 1 a 2713) delle nuove PSU.

Poiché il CLUSTERS richiede che i records contengano soltanto valori numerici, è stato necessario ricodificare alcune variabili per le quali erano stati utilizzati codici alfabetici per indicare valori errati o mancanti. Infine è stato aggiunto un record fittizio indicante la fine dei dati, contenente il valore 99 nelle prime due colonne e tutti blanks nelle altre.

Le modifiche non sono state apportate direttamente sul nastro, ma si è provveduto alla costruzione su disco del file SALUTE DATI (LRECL 256).

Di seguito viene descritto il tracciato record, limitatamente alle variabili necessarie per le elaborazioni, in cui sono riportati: le colonne occupate dalla variabile, il nome della variabile, gli eventuali codici e le modalità corrispondenti a ciascuno di essi, il numero progressivo del campo e il formato.

Colonne	Variabile	Campo	Formato
1-2	PROVINCIA	1	I2
3-21	campi che non devono essere letti		19x
22	SESSO	2	I1
	maschi = 8		
	femmine = 9		
23-46	campi che non devono essere letti		24x
47	STATO DI SALUTE	3	I1
	buono = 1		
	non buono = 2		
	buono dopo correzione = 3		
	non buono senza gg. malat. = 4		
	non buono dopo correzione = 6		

48-51	campi che non devono essere letti		4x
52-53	GIORNI DI MALATTIA	4	I2
54-60	campi che non devono essere letti		7x
61	DIABETE	5	I1
	insorta prima del 1983 = 1		
	insorta durante il 1983 = 2		
	malattia assente = blank		
62	IPERTENSIONE	6	I1
	insorta prima del 1983 = 1		
	insorta durante il 1983 = 2		
	malattia assente = blank		
63-65	campi che non devono essere letti		3x
66	BRONCHITE	7	I1
	insorta prima del 1983 = 1		
	insorta durante il 1983 = 2		
	malattia assente = blank		
67-77	campi che non devono essere letti		11x
78	ARTROSI	8	I1
	insorta prima del 1983 = 1		
	insorta durante il 1983 = 2		
	malattia assente = blank		
79-130	campi che non devono essere letti		52x
131	È STATO RICOVERATO	9	I1
	si = 1		
	no = 2		
	non indicato = blank		
132	campo che non deve essere letto		1x
133-135	GIORNI DI DEGENZA	10	I3
	(999 = non indicati)		
136-154	campi che non devono essere letti		19x
155	ABITUDINE AL FUMO	11	I1
	non fumatore = 1		
	fumatore = 2		
	non indicato = blank		
156-160	campi che non devono essere letti		5x
161-162	SIGARETTE FUMATE GIORNALMENTE	12	I2
	(99 = non indicate)		
163-183	campi che non devono essere letti		21x
184	RIPARTIZIONI TERRITORIALI	13	I1
	Italia Nord Occidentale = 1		
	Italia Nord Orientale = 2		
	Italia Centrale = 3		
	Italia Meridionale = 4		
	Italia Insulare = 5		

185-223	campi che non devono essere letti		39x
224-228	COEFFICIENTE DI PONDERAZIONE	14	I5
229-238	campi che non devono essere letti		10x
239-243	N. D'ORDINE DELLE PSU	15	I5
244-247	campi che non devono essere letti		4x
248-251	CODICE DI STRATO	16	I4
252-256	campi che non devono essere letti		5x

Per la descrizione del tracciato record deve essere utilizzato il formato FORTRAN, dove con kIn vengono indicati k campi ciascuno di n cifre intere e con *nx un campo* di n caratteri che non va letto. Così I2 indica che deve essere letto un campo di 2 cifre intere e 19x che non devono essere lette le successive 19 colonne.

La variabile provincia è stata letta perché in questo campo è riportato il codice identificativo della fine dei dati.

3. Compilazione delle schede parametro

Si è passati, quindi, alla compilazione del file SALUTE PARAM, formato da tanti record di 80 caratteri quante sono le schede parametro che è necessario compilare per l'esecuzione del programma.

Le schede parametro vengono riportate con un numero progressivo indicante la posizione che esse occupano nel file. Per la loro descrizione non vengono indicate le colonne del record ma i singoli campi, rimandando all'appendice 1 per la corrispondenza tra campi e colonne. Per indicare un campo che deve essere lasciato in bianco è stata utilizzata la lettera 'b'.

Schede per la lettura dei dati e per la descrizione del problema

La prima scheda che deve essere compilata è quella contenente il titolo del lavoro, che ha il seguente tracciato:

Campi sk 1

a TITL
b b
c errori di campionamento per l'indagine sulla salute del 1983

La descrizione del formato per la lettura dei dati richiede la compilazione di due schede:

Campi	sk 2
a	FORM
b	2
c	(l2,19x,l1,24x,l1,4x,l2,7x,2l1,3x,l1,11x,l1,52x,i1
	sk 3
a	b
b	b
c	8x,l3,12x,l1,5x,l2,21x,l1,39x,l5,10x,l5,4x,i4)

Nella quarta scheda vengono descritte la struttura del campione e le elaborazioni richieste:

Campi	sk 4	
a	PROB	
b	16	n. di campi letti
c	2	n. di sk CLAS
d	10	n. di sk VARI
e	2713	n. di PSU
f	0	lettura dell'AREA da un campo input
g	15	campo identificativo dell'AREA
h	1	PSU = AREA di riferimento
i	b	
j	0	lettura dello strato da un campo input
k	16	campo identificativo dello strato
l	0	lettura del coeff. di ponderazione da un campo input
m	14	campo identificativo del coeff. di ponderazione
n	0	lettura dei domini da un campo input
o	13	campo identificativo dei domini
p	3	n. di sk DOMA

Segue la scheda 'FACT' in cui sono riportati l'eventuale fattore di scala, il numero del campo e il valore che indica la fine dei dati, il numero degli eventuali record invalidati:

Campi	sk 4	
a	FACT	
b	b	non è previsto un fattore di scala
c	1	campo identificativo della fine dei dati
d	99	valore indicante la fine dei dati
e	0	numero di record invalidati

Schede per la descrizione delle sottoclassi

Per ogni sottoclasse viene predisposta una scheda CLAS in cui è riportato il nome della sottoclasse, formato da quattro caratteri, e il numero di schede RECO utilizzate per la ricodifica. Ciascuna scheda CLAS è immediatamente seguita dalla corrispondente scheda RECO utilizzata per ricodificare la sottoclasse.

Così per definire la sottoclasse delle unità del campione di sesso maschile, occorre costruire una nuova variabile 'MASC' che assume il valore 1 quando nel secondo campo input (SESSO) viene letto il codice 8 (maschio) e il valore 0 negli altri casi. In modo analogo bisogna procedere per definire la sottoclasse 'FEMM'.

Per la definizione delle due sottoclassi sono state, pertanto, predisposte due schede VARI e due schede CLAS:

Campi	sk 5	sk 6	sk 7	sk 8
a	CLAS	RECO	CLAS	RECO
b	MASC	1	FEMM	1
c	1	b	1	b
d	b	b	b	b
e	b	1	b	1
f		0		0
g		2		2
h		b		b
i		b		b
j		b		b
k		b		b
l		8		9
m-p		b		b

Schede per la descrizione delle stime

Per ciascuna delle 10 stime, per le quali è richiesto il calcolo degli errori campionari, deve essere predisposta una scheda VARI contenente: il nome del parametro che deve essere stimato, il campo identificativo se questa viene letta da un campo input, il numero delle schede RECO nei casi in cui è necessaria la ricodifica, il numero degli eventuali valori che devono essere esclusi dai calcoli e i valori da escludere.

Per ognuna delle tre stime di percentuali (SALU, RICO e FUMA), occorre definire una nuova variabile che assume il valore 100 se l'unità rilevata presenta la caratteristica in esame e il valore 0 altrimenti.

Analogamente, per DIAB, IPER, BRON e ARTR, che sono stime di frequenze relative moltiplicate per 1000, occorre definire quattro nuove variabili che assumono i valori 1000 e 0 a seconda che l'unità rilevata presenti o meno la malattia considerata.

Le restanti tre stime (GIOM, DEGM e SIGM), essendo delle medie, vengono lette direttamente da un campo input, nell'ordine il 4°, il 10° e il 12°. Alcuni valori devono essere esclusi dai calcoli, in modo che le tre medie siano riferite rispettivamente alle sole persone: in non buona salute, ricoverate in ospedale e fumatrici di sigarette. Pertanto, i valori esclusi dai calcoli sono: 0 (nessun giorno di malattia) per GIOM, 0 (non ricoverato) e 999 (non indicato) per DEGM, 0 (nessuna sigaretta) e 99 (non indicato) per SIGM.

In totale si hanno 10 schede VARI e 7 schede RECO, ciascuna delle quali dovrà essere posta immediatamente dopo la corrispondente scheda VARI:

Campi	sk 9	sk 10	sk 11	sk 12	sk 13	sk 14	sk 15	sk 16	sk 17
a	VARI	RECO	VARI	VARI	RECO	VARI	RECO	VARI	RECO
b	SALU	1	GIOM	DIAB	1	IPER	1	BRON	1
c	0	b	4	0	b	0	b	0	b
d	1	b	0	1	b	1	b	1	b
e	0	3	1	0	2	0	2	0	2
f		0	0		2		2		2
g		3			5		6		7
h		b			b		b		b
i		b			b		b		b
j		b			b		b		b
k		100			1000		1000		1000
l		2			1		1		1
m		4			2		2		2
n		6			b		b		b
o-p					b		b		b

Campi	sk 18	sk 19	sk 20	sk 21	sk 22	sk 23	sk 24	sk 25
a	VARI	RECO	VARI	RECO	VARI	VARI	RECO	VARI
b	ARTR	1	RICO	1	DEGM	FUMA	1	SIGM
c	0	b	0	b	10	0	b	12
d	1	b	1	b	0	1	b	0
e	0	2	0	1	2	0	1	2
f		2		1	0		1	0
g		8		9	999		11	99
h		b		b			b	
i		b		b			b	
j		b		b			b	
k		1000		100			100	
l		1		1			2	
m		2		b			b	
n-p		b		b			b	

Schede per la descrizione dei domini

Ciascun dominio viene descritto mediante una scheda DOMA in cui sono riportati il nome del dominio, il numero di modalità che identificano il dominio e i codici corrispondenti:

Campo	sk 26	sk 27	sk 28
a	DOMA	DOMA	DOMA
b	NORD	CENTRO	SUD
c	2	1	2
d	1	3	4
e	2	b	5
f-p	b	b	b

Una volta ultimata la compilazione delle schede parametro è bene controllare che non siano stati commessi errori. In particolare occorre verificare che il numero di schede CLAS, VARI e DOMA compilate coincida con i numeri riportati rispettivamente nei campi 'c', 'd' e 'p' della scheda PROB.

4. Esecuzione del programma ed elaborazione delle tavole

I due programmi CLUST1 e CLUST2 e il SORT vengono fatti eseguire mediante apposite exec, che devono essere di volta in volta modificate per lo specifico problema trattato.

Le modifiche riguardano i nomi dei file input ed output e le unità (disco o nastro) su cui si trovano, e le specifiche del file KCSUMS generato da CLUST1 che variano con il numero delle schede VARI compilate.

Il programma CLUST1 viene lanciato mediante l'EXEC CLUST1:

```
CP LINK PRODSOFT 204 204 RR FORT
ACC 204 B
SET DOS OFF
GLOBAL LOADLIB VSF2LOAD TXTLIB VSF2FORT VSF2LINK
GLOBAL MOD1EEH VSF2MATH TFORTLIB FORTMOD1
FI 1 DISK SALUTE PARAM A (LRECL 80
* file input delle schede parametro
FI 2 DISK SALUTE DATI A  (LRECL 256 RECFM FB BLOCK
2560
* file input dei dati di base
FI 4 DISK KPRINT DATI A  (LRECL 133 RECFM FB BLOCK 133
* file output con le sk parametro e la struttura del campione
```

```
FI 5 DISK KCNTRL DATI  A (LRECL 88
• file output con le informazioni di controllo per CLUST2
FI 6 DISK KCSUMS DATI A  (LRECL 372  RECFM VB  BLOCK
3724
• file output che deve essere sortato
FI 7 DUMMY (LRECL 72
LOAD CLUST1
START
```

La lunghezza dei record e il bloccaggio del file KCSUMS sono dati da:

LRECL = 4 (4 + 8 (1 + NVARI)) + 4 = 372
BLKSIZE = 4 + 10 LRECL = 3724

dove NVARI=10 è il numero delle stime richieste.

Come è stato detto nell'appendice 1, prima di mandare in esecuzione il programma CLUST2 occorre predisporre su nastro il file KSUMS ordinato per sottoclasse, dominio, strato e PSU. Ciò viene fatto mediante il programma SORT che viene lanciato da CLUSORT EXEC:

```
TAPE REW
EXEC DTRIPF NOPAN
FI SORTIN  DISK KCSUMS DATI A (LRECL 372 RECFM VB
BLOCK 3724
• file input generato da CLUST1
FI SORTOUT TAP1 (LRECL 372 RECFM V BLOCK 3724
• file output ordinato
FI SYSIN DISK CASORT CONTROL A (RECFM FB LRECL 80
FI SORTLIST TERM
CASORT
```

fornendo i seguenti parametri: (5, 16, BI, A).

Infine, mediante CLUST2 EXEC viene mandato in esecuzione il programma CLUST2:

```
TAPE REW
CP LINK PRODSOFT 204 204 RR FORT
ACC 204 B
SET DOS OFF
GLOBAL LOADLIB VSF2LOAD TXTLIB VSF2FORT VSF2LINK
GLOBAL MOD1EEH VSF2MATH TFORTLIB FORTMOD1
FI 1 DISK KCNTRL DATI A
• file input generato da CLUST1
FI 2 TAP1 (LRECL 372 RECFM VB BLOCK 3724
• file input generato da SALSORT
```



```
FI 7 DISK ERRORI DATI  A (LRECL 133 RECFM FB BLOCK 133
• file output con gli errori di campionamento
FI 6  DISK A A A (LRECL 72 RECFM VB
LOAD CLUST2
START
```

La struttura del file CMS 'ERRORI DATI' contenente gli errori campionari è descritta nell'appendice 1.

Poiché il file ERRORI DATI non può essere utilizzato direttamente per la pubblicazione, è stato predisposto un programma SAS che trasforma il file CMS in un data set SAS e successivamente, mediante la PROC TABULATE, elabora le tavole standard da pubblicare. Il data set SAS costituisce, inoltre, l'input per la stima degli errori mediante modelli.

Il numero delle tavole che vengono elaborate è dato dal prodotto (NCLAS + 1)•(NDOM + 1), dove con NCLAS e NDOM si è indicato rispettivamente il numero delle classi e dei domini.

Ciascuna tavola riporta per ognuna delle stime richieste: la denominazione del parametro, il valore della stima, l'zerrore standard assoluto e percentuale, l'intervallo di confidenza al 95%, il fattore del disegno, la numerosità del campione e il numero medio di unità campione per PSU.

Nell'applicazione effettuata si hanno 12 tavole divise in 4 gruppi: il primo relativo all'Italia, il secondo al Nord, il terzo al Centro e il quarto al Sud. Ciascun gruppo è costituito da 3 tavole: la prima con i risultati per la sottoclasse totale, la seconda per i maschi e la terza per le femmine.

Tavola A1.1 – Errori di campionamento: indagine sulla salute del 1983

Dominio = Italia *Classe = Totale*

PARAMETRI	STIMA	SE	RE(%)	INF	SUP	DEFT	N	B
SALU	19,406	0,552	2,8	18,301	20,511	2,331	89765	33,1
GIOM	10,489	0,279	2,7	9,932	11,046	1,730	16785	6,8
DIAB	33,074	1,478	4,5	30,117	36,031	1,794	89765	33,1
IPER	65,385	2,833	4,3	59,719	71,051	2,113	89765	33,1
BRON	45,425	2,264	5,0	40,898	49,952	2,058	89765	33,1
ARTR	157,807	5,244	3,3	147,320	168,294	2,367	89765	33,1
RICO	7,703	0,240	3,1	7,223	8,183	1,875	89765	33,1
DEGM	18,771	0,943	5,0	16,885	20,657	1,838	7021	2,6
FUMA	25,578	0,306	1,2	24,966	26,190	1,654	89765	33,1
SIGM	15,722	0,144	0,9	15,434	16,010	1,778	27590	8,1

Tavola A1.2 – Errori di campionamento: indagine sulla salute del 1983

Dominio = Italia *Classe = Maschi*

PARAMETRI	STIMA	SE	RE(%)	INF	SUP	DEFT	N	B
SALU	17,415	0,499	2,9	16,416	18,414	1,892	43567	16,1
GIOM	10,657	0,362	3,4	9,933	11,391	1,581	7364	3,0
DIAB	29,002	1,512	5,2	25,977	32,027	1,564	43567	16,1
IPER	53,223	3,001	5,6	47,220	59,226	1,905	43567	16,1
BRON	62,370	3,054	4,9	56,261	68,479	1,851	43567	16,1
ARTR	131,676	4,524	3,4	122,627	140,725	1,906	43567	16,1
RICO	7,302	0,267	3,7	6,767	7,837	1,672	43567	16,1
DEGM	19,983	1,313	6,6	17,357	22,609	1,794	3249	1,2
FUMA	37,174	0,459	1,2	36,256	38,092	1,606	43567	16,1
SIGM	17,440	0,167	1,0	17,107	17,773	1,754	20042	5,6

Tavola A1.3 – Errori di campionamento: indagine sulla salute del 1983

Dominio = Italia *Classe = Femmine*

PARAMETRI	STIMA	SE	RE(%)	INF	SUP	DEFT	N	B
SALU	21,259	0,595	2,8	20,069	22,449	2,016	46198	17,0
GIOM	10,364	0,333	3,2	9,698	11,030	1,656	9421	3,8
DIAB	36,851	2,031	5,5	32,790	40,912	1,736	46198	17,0
IPER	76,710	3,225	4,2	70,261	83,159	1,840	46198	17,0
BRON	29,436	1,975	6,7	25,486	33,386	1,807	46198	17,0
ARTR	182,102	5,642	3,1	170,818	193,386	2,022	46198	17,0
RICO	8,068	0,296	3,7	7,476	8,660	1,741	46198	17,0
DEGM	17,744	1,082	6,1	15,581	19,907	1,682	3772	1,4
FUMA	14,650	0,332	2,3	13,985	15,315	1,621	46198	17,0
SIGM	11,050	0,188	1,7	10,673	11,427	1,682	7548	2,4

Tavola A2.1 – Errori di campionamento: indagine sulla salute del 1983

Dominio = Nord *Classe = Totale*

PARAMETRI	STIMA	SE	RE(%)	INF	SUP	DEFT	N	B
SALU	19,804	0,829	4,2	18,145	21,463	2,310	38920	33,4
GIOM	10,142	0,385	3,8	9,373	10,911	1,772	7319	6,9
DIAB	31,563	2,084	6,6	27,395	35,731	1,749	38920	33,4
IPER	73,837	4,433	6,0	64,971	82,703	2,085	38920	33,4
BRON	38,398	2,688	7,0	33,023	43,773	1,894	38920	33,4
ARTR	154,506	8,070	5,2	138,366	170,646	2,393	38920	33,4
RICO	9,039	0,395	4,4	8,249	9,829	1,880	38920	33,4
DEGM	18,751	1,483	7,9	15,785	21,717	1,911	3577	3,1
FUMA	26,443	0,496	1,9	25,451	27,435	1,698	38920	33,4
SIGM	15,233	0,236	1,5	14,761	15,705	1,859	12505	8,3

Tavola A2.2 – Errori di campionamento: indagine sulla salute del 1983

Dominio = Nord *Classe = Maschi*

PARAMETRI	STIMA	SE	RE(%)	INF	SUP	DEFT	N	B
SALU	17,696	0,808	4,6	16,080	19,312	1,942	18760	16,1
GIOM	10,510	0,552	5,3	9,406	11,614	1,643	3195	3,0
DIAB	28,293	2,239	7,9	23,814	32,772	1,551	18760	16,1
IPER	59,632	4,738	7,9	50,156	69,108	1,888	18760	16,1
BRON	51,656	3,560	6,9	44,537	58,775	1,693	18760	16,1
ARTR	121,352	7,046	5,8	107,261	135,443	1,960	18760	16,1
RICO	8,812	0,450	5,1	7,911	9,713	1,681	18760	16,1
DEGM	19,134	2,251	11,8	14,633	23,635	1,976	1691	1,5
FUMA	36,558	0,699	1,9	35,157	37,953	1,607	18760	16,1
SIGM	17,024	0,280	1,6	16,463	17,585	1,820	8690	5,5

Tavola A2.3 – Errori di campionamento: indagine sulla salute del 1983

Dominio = Nord *Classe = Femmine*

PARAMETRI	STIMA	SE	RE(%)	INF	SUP	DEFT	N	B
SALU	21,726	0,878	4,0	19,971	23,481	1,981	20160	17,3
GIOM	9,867	0,449	4,6	8,968	10,766	1,736	4124	3,5
DIAB	34,511	2,799	8,1	28,913	40,109	1,882	20160	17,3
IPER	86,818	4,939	5,7	76,940	96,696	1,799	20160	17,3
BRON	25,974	2,642	10,2	20,689	31,259	1,751	20160	17,3
ARTR	184,849	8,772	4,7	167,306	202,392	2,042	20160	17,3
RICO	9,223	0,472	5,1	8,280	10,166	1,734	20160	17,3
DEGM	18,414	1,496	8,1	15,417	21,411	1,820	1886	1,6
FUMA	16,982	0,545	3,2	15,891	18,073	1,838	20160	17,3
SIGM	11,066	0,271	2,5	10,523	11,609	1,886	3815	2,8

Tavola A3.1 – Errori di campionamento: indagine sulla salute del 1983

Dominio = Centro *Classe = Totale*

PARAMETRI	STIMA	SE	RE(%)	INF	SUP	DEFT	N	B
SALU	20,801	0,886	4,3	19,028	22,574	2,131	25546	34,0
GIOM	10,635	0,410	3,9	9,815	11,455	1,518	5173	7,4
DIAB	34,882	2,549	7,3	29,785	39,979	1,699	25546	34,0
IPER	75,757	4,639	6,1	66,478	85,036	1,908	25546	34,0
BRON	51,705	3,550	6,9	44,605	58,805	1,825	25546	34,0
ARTR	201,309	9,669	4,8	181,972	220,646	2,239	25546	34,0
RICO	7,664	0,395	5,2	6,874	8,454	1,756	25546	34,0
DEGM	18,211	1,414	7,8	15,383	21,039	1,723	1908	2,5
FUMA	26,854	0,504	1,9	25,845	27,863	1,538	25546	34,0
SIGM	15,723	0,268	1,7	15,188	16,258	1,785	8363	8,8

Tavola A3.2 – Errori di campionamento: indagine sulla salute del 1983

Dominio = Centro *Classe = Maschi*

PARAMETRI	STIMA	SE	RE(%)	INF	SUP	DEFT	N	B
SALU	19,065	0,781	4,1	17,503	20,627	1,698	12432	16,5
GIOM	10,765	0,578	5,4	9,609	11,921	1,442	2304	3,3
DIAB	34,417	2,990	8,7	28,438	40,396	1,542	12432	16,5
IPER	63,773	4,403	6,9	54,966	72,580	1,616	12432	16,5
BRON	72,668	5,292	7,3	62,085	83,251	1,719	12432	16,5
ARTR	171,785	8,342	4,9	155,101	188,469	1,790	12432	16,5
RICO	7,022	0,531	7,6	5,960	8,084	1,736	12432	16,5
DEGM	20,258	1,926	9,5	16,405	24,111	1,613	865	1,2
FUMA	37,367	0,687	1,8	35,992	38,742	1,435	12432	16,5
SIGM	17,429	0,302	1,7	16,826	18,032	1,713	5983	6,0

Tavola A3.3 – Errori di campionamento: indagine sulla salute del 1983

Dominio = Centro *Classe = Femmine*

PARAMETRI	STIMA	SE	RE(%)	INF	SUP	DEFT	N	B
SALU	22,445	1,020	4,5	20,404	24,486	1,907	13114	17,4
GIOM	10,537	0,494	4,7	9,550	11,524	1,465	2869	4,1
DIAB	35,330	3,923	11,1	27,4846	43,1764	1,778	13114	17,4
IPER	87,123	5,808	6,7	75,506	98,740	1,751	13114	17,4
BRON	31,845	2,918	9,2	26,009	37,681	1,573	13114	17,4
ARTR	229,280	10,74	4,7	207,799	250,761	1,950	13114	17,4
RICO	8,273	1	6,0	7,280	9,266	1,639	13114	17,4
DEGM	16,561	0,497	9,4	13,447	19,675	1,587	1043	1,4
FUMA	16,895	1,557	4,0	15,539	18,251	1,642	13110	17,4
SIGM	11,526	0,678	2,9	10,852	12,200	1,677	2380	2,8
		0,337						

Tavola A4.1 – Errori di campionamento: indagine sulla salute del 1983

Dominio = Sud *Classe = Totale*

PARAMETRI	STIMA	SE	RE(%)	INF	SUP	DEFT	N	B
SALU	18,163	0,969	5,3	16,226	20,100	2,279	25299	31,8
GIOM	10,871	0,552	5,1	9,767	11,975	1,633	4293	6,0
DIAB	34,009	2,724	8,0	28,562	39,456	1,763	25299	31,8
IPER	49,245	5,063	10,3	39,118	59,372	2,200	25299	31,8
BRON	50,903	4,838	9,5	41,226	60,580	2,133	25299	31,8
ARTR	138,772	8,654	6,2	121,464	156,080	2,275	25299	31,8
RICO	6,047	0,361	6,0	5,325	6,769	1,771	25299	31,8
DEGM	19,189	1,509	7,9	16,170	22,208	1,594	1536	1,9
FUMA	23,811	0,512	2,2	22,787	24,835	1,576	25299	31,8
SIGM	16,468	0,202	1,2	16,065	16,871	1,505	6722	7,1

Tavola A4.2 – Errori di campionamento: indagine sulla salute del 1983

Dominio = Sud *Classe = Maschi*

PARAMETRI	STIMA	SE	RE(%)	INF	SUP	DEFT	N	B
SALU	16,188	0,814	5,0	14,560	17,816	1,789	12375	15,5
GIOM	10,786	0,649	6,0	9,487	12,085	1,469	1865	2,7
DIAB	27,001	2,602	9,6	21,796	32,206	1,524	12375	15,5
IPER	39,636	5,477	13,8	28,682	50,590	2,015	12375	15,5
BRON	70,211	6,354	9,0	57,503	82,919	1,897	12375	15,5
ARTR	123,161	7,220	5,9	108,721	137,601	1,782	12375	15,5
RICO	5,574	0,368	6,6	4,837	6,311	1,522	12375	15,5
DEGM	21,478	1,629	7,6	18,221	24,735	1,357	693	0,9
FUMA	37,841	0,803	2,1	36,235	39,447	1,547	12375	15,5
SIGM	18,008	0,244	1,4	17,520	18,496	1,573	5369	5,5

Tavola A4.3 – Errori di campionamento: indagine sulla salute del 1983

Dominio = Sud *Classe = Femmine*

PARAMETRI	STIMA	SE	RE(%)	INF	SUP	DEFT	N	B
SALU	20,032	1,052	5,3	17,929	22,135	1,970	12924	16,2
GIOM	10,939	0,666	6,1	9,608	12,270	1,530	2428	3,3
DIAB	40,633	3,665	9,0	33,303	47,963	1,656	12924	16,2
IPER	58,324	5,657	9,7	47,009	69,639	1,889	12924	16,2
BRON	32,550	3,972	12,2	24,606	40,494	1,819	12924	16,2
ARTR	153,471	9,300	6,1	134,871	172,071	1,953	12924	16,2
RICO	6,491	0,467	7,2	5,557	7,425	1,674	12924	16,2
DEGM	17,341	2,252	13,0	12,836	21,846	1,678	843	1,1
FUMA	10,492	0,494	4,7	9,504	11,480	1,543	12924	16,2
SIGM	10,582	0,363	3,4	9,8573	11,307	1,563	1353	1,6

RIFERIMENTI BIBLIOGRAFICI

Bean, J.A. (1975), *Distribution and Properties of Variances Estimators for Complex Multistate Probability Sample.* «Vital and Health statistics», Series 2, No. 65, National Center for Health Statistic, Public Health Service, Washington. D.C.

Brewer, K.R.W., and Hanif, M., (1983), *Sampling with Unequal Probabilities,* Springer-Verlag, New York.

Cassel, C.M., Sarndal, C.E., and Wretman, J.H., (1977), *Foundations of Inference in Survey Sampling,* John Wiley & Sons, New York.

Castellano V. e Herzel A. (1981), *Elementi di teoria dei campioni,* Edizioni Sistema, Roma.

Cochran, W.G. (1977), *Sampling Tecniques,* John Wiley & Sons, New York.

Dippo, C.S., Fay R.E. and Morganstein, D.H. (1984), *Computing Variances from Complex Samples with Replicate Weights.* «Proceedings of the Section on Survey Research Methods», American Statistical Association.

Francis, I. and Sedransk, J. (1976), *Software Requirements for the Analysis of Surveys.* «Proceedings 9th International Biometric Conference», 228-253.

Francis, I. and Sedransk, J. (1979), *Comparing Software for Processing and Analyzing Survey Data.* «Bulletin of the International statistical Institute».

Gonzales, M., Ogus J., Shapiro, G. and Tepping, B. (1975), *Standards for Discussion and Presentation of Errors in Survey and Census Data.* «Journal of the American Statistical Association, Supplement» 70, 5-23.

Gourney, M. (1970a), *McCarthy's Orthogonal Replications for Variances with Grounded Strata.* «Technical Notes» no. 3, 13-16, U.S. Bureau of the Census, Washington, D.C. 20233.

Gourney, M. (1970b), *The Variance of the Replications Method for Estimating Variances for CPS Sample Design.* «Technical Notes» No. 3, 7-12, U.S. Bureau of the Census, Washington D.C. 20233.

Gourney M., and Jewett, R.S. (1975), *Constructing Orthogonal Replications for Variance Estimation.* «Journal of the American Statistical Association» 70, 819-821.

Hajek, J. (1981), *Sampling From a Finite Population,* Marcel Dekker, Inc. New York and Basel.

Hansen, M.H., Hurwitz W.N., and Bershad, M.A. (1961), *Measurement Errors in Censuses and Surveys,* «Bulletin of the International Statistical Institute» 38, Part. II, 359-374.

Hartley, H.O., Rao, J.N.K., and Kiefer, G. (1969), *Variance Estimation with One Unit per Stratum,* «Journal of the American Statistical Association» 64, 841-851.

Kaplan B., Francis, I. and Sedransk, J. (1979a), *A Comparison of Methods and Programs for Computing Variances of Estimators from Complex Sample Surveys.* «Proceeding of the Section on Survey Research Methods», American Statistical Association, 97-100.

Kaplan B., Francis I. and Sedransk, J. (1979b), *Criteria for Comparing Programs for Computing Variances of Estimators from complex Sample Surveys.* «Proceeding of the 12th Annual Symposium on the Interface of Computer Science and Statistics» 390-395.

Kish, L., (1965), *Survey Sampling,* John Wiley & Sons, New York.

Kish, L. and Frankel, M.R. (1970), *Balanced Repetead Replication for Standard Errors.* «Journal of the American Statistical Association» 65, 1071-1094.

Krewski, D, (1978a), *On the Stability of Some Replication Variance Estimators in the linear case.* «Journal of Statistical Planning and Inference» 2, 45-51.

Krewski, D. and Rao, J.N.K. (1981), *Inference from Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods.* «Annals of Statistics» 9, 1010-1019.

Lee, K.H. (1972), *The use of partially Balanced Designs for Half-Sample Replication Method of Variance Estimation.* «Journal of the American Statistical Association» 67, 324-334.

Lee, K.H. (1973a), *Using Partially Balanced Designs for the Half-Sample Method of Variance Estimation.* «Journal of the American Statistical Association» 68, 612-614.

Lee, K.H. (1973b), *Variance Estimation in Stratified Sampling* «Journal of the American Statistical Association» 66, 336-342.

Lemenshow, S. (1976), *The use of Unique statistical Weights for Estimating Variances with the Balanced Half-Sample Tecnique.* «Proceeding of the Social Statistics Section», American Statistical Association.

Leti, G. (1983), *Statistica Descrittiva,* il Mulino, Bologna.

Little, R.J.A. (1982), *Sampling Errors of Fertility Rates from the WFS.* «WFS Tecnical Bulletins» no. 10.

McCarthy, P.J. (1966), *Replication: An Approach to the Analysis of Data from Complex Surveys.* «Vital and Health Statistics» Series 2. No. 14, National Center for Health Statistics, Public Health Service, Washington, D.C.

McCarthy, P.J. (1969a), *Pseudoreplication: Further Evaluation and Application of the Balanced Half-Sample Tecnique.* «Vital and Health Statistics» Series 2. No. 31, National Center for Health Statistics, Public Health Service, Washington, D.C.

McCarthy, P.J. (1969b), *Pseudoreplication: Half-Samples.* «Review of the International Statistical Institute» 37, 239-264.

O'Muircheartaigh, C.A. (1982), *Methodology of the Response Errors Project.* «WFS Scientific Reports» n. 28.

O'Muircheartaigh, C.A. and A.M. Markwardt (1981), *An Assessment of the Reliability of WFS Data.* «Word Fertility Survey Conference 1980: Record of Proceeding», vol. 3:313. Voorburg, Netherlands: International Statistical Institute.

Royall, R.M. and Cumberland, W.G. (1978). *Variance Estimation in Finite Population Sampling.* «Journal of the American Statistical Association» 73, 351-358.

Russo, A. (1987). Sulla presentazione degli errori di campionamento *mediante modelli.* «Quaderni di Discussione» n. 4, Istat, Roma.

Seth, G.R. (1966), *On collapsing Strata.* «Journal of the Indian Society of Agricultural Statistics» 18, 1-3.

Shah, B.V. (1978), *Variance Estimates for Complex Statistics from Multistage Sample Surveys.* «Survey Sampling and Measurement», N. Krishnan Namboodiri (ed.) Academic Press: New York.

Shapiro, G.M. and Bateman, D.V. (1978), *A Better Alternative to the Collapsed Stratum Variance Estimate.* «Proceedings of the Section on Survey Research Methods» American Statistical Association.

Tepping, B.J. (1968), *Variance Estimation in Complex Surveys.* «Proceeding of the Social Statistics Section», American Statistical Association.

United Nations (1949), *The Preparation of Sampling Survey Reports.* Statistical Papers, Series C, No. 1, Statistical Office of the United Nations, New York.

Verma, V. (1978), *CLUSTERS: a Package Program for the Computation of Sampling Errors.* United Nations Economic Commission for Europe Conference of European Statisticians, meeting on problems relating to household surveys, Geneva.

Verma, V. (1981a), *Assessment of Errors in Household Surveys.* «Bull. International Statistical Institute» 49.

Verma, V. (1981b), *Sampling for National Fertility Surveys.* «World Fertility Survey Conference 1980: Record of Proceeding». vol. 3: 389.

Verma V. (1982), *The Estimation and Presentation of Sampling Errors.* «Technical Bullettins, World Fertility Survey, december 1982 n. 11».

Verma V. and M.C. Pearce (1978), *Users' Manual for CLUSTERS.* WFS Technical Paper no. 770.

Verma, V., C. Scott and C. O' Muircheartaigh (1980), *Sample Designs and Sampling Errors for the World Fertility Survey.* «J. Royal Statistical Society A», 143, pt. 4:431-73.

Wolter, K.M., (1985), *Introduction to Variance Estimation,* Springer-Verlag, New York.

Woodruff, R. (1971), *A Simple Method for Approximating the Variance of Complicated Estimate.* «J. American Statistical Association» 67:411-14.

Woodruff, R. and B. Casey (1976), *Computerized Method dor Approximating the Variance of Complicated Estimate.* «J. American Statistical Association» 71:315-21.

Zannella, f. (1982), *Calcolo degli errori di campionamento* in «Indagine statistica sulle condizioni di salute della popolazione, e sul ricorso ai servizi sanitari», Istat, Supplemento al Bollettino mensile di statistica, anno 1982, n. 12.

Zannella, F. (1984), *La misura dell'errore delle stime nelle indagine campionarie «multipurpose» e l'utilizzazione di variabili ausiliarie nei procedimenti di stratificazione,* Atti della XXXII Riunione Scientifica della SIS, Sorrento, 11-13 aprile 1984.

Zannella, F. (1985), *Problemi relativi alla stratificazione dei comuni italiani nelle indagini sulla popolazione* (Istat, Commissione «campioni» doc. n. 6, 1985).

PUBBLICAZIONI ISTAT

BOLLETTINO MENSILE DI STATISTICA

La più completa ed autorevole raccolta di dati congiunturali concernenti l'evoluzione dei fenomeni demografici, sociali, economici e finanziari.

Abbonamento annuo L. 115.000 (Estero L. 139.000) Ogni fascicolo L. 15.000

INDICATORI MENSILI

Forniscono dati riassuntivi e tempestivi sull'andamento mensile dei principali fenomeni interessanti la vita nazionale.

Abbonamento annuo L. 29.000 (Estero L. 35.000) Ogni fascicolo L. 3.700

NOTIZIARI ISTAT

Forniscono i primi risultati delle rilevazioni ed elaborazioni statistiche riguardanti l'attività produttiva, i prezzi, il commercio interno, gli scambi internazionali come pure lo stato ed il movimento della popolazione e le sue caratteristiche sociali e sanitarie.

I dati, esposti in grafici e tabelle, sono accompagnati da commenti, illustrazioni e note interpretative.

Serie 1 - Statistiche demografiche e sociali

Abbonamento annuo L. 22.000 (Estero L. 29.000) una copia L. 1.600

Serie 2 - Statistiche dell'attività produttiva

Abbonamento annuo L. 64.000 (Estero L. 85.000) una copia L. 1.600

Serie 3 - Statistiche del lavoro, delle retribuzioni e dei prezzi

Abbonamento annuo L. 22.000 (Estero L. 29.000) una copia L. 1.600

Serie 4 - Argomenti vari

Abbonamento annuo L. 13.000 (Estero L. 17.000) una copia L. 1.600

Abbonamento annuo a tutte le serie L. 106.000 (Estero L. 144.000).

INDICATORI TRIMESTRALI

Conti economici trimestrali

Abbonamento annuo L. 11.000 (Estero L. 13.000) Ogni fascicolo L. 3.700

STATISTICA DEL COMMERCIO CON L'ESTERO

Documentazione statistica ufficiale, a periodicità trimestrale, sul commercio dell'Italia con l'estero; fornisce, per tutte le merci comprese nella classificazione merceologica della tariffa dei dazi doganali, l'andamento delle importazioni e delle esportazioni da e per i principali Paesi.

Abbonamento annuo L. 99.000 (Estero L. 112.000) Ogni fascicolo L. 31.000

Abbonamento annuo cumulativo a tutti i periodici, compresa la «Statistica del commercio con l'estero»: L. 300.000 (Estero L. 390.000); esclusa la «Statistica del commercio con l'estero»: L. 209.000 (Estero L. 286.000)

Gli abbonamenti decorrono dal 1° gennaio anche se sottoscritti nel corso dell'anno. In tal caso l'abbonato riceverà i numeri dell'annata già pubblicati. L'abbonato ai periodici ISTAT ha diritto a ricevere gratuitamente i fascicoli non pervenutigli soltanto se ne segnalerà il mancato arrivo entro 10 giorni dal ricevimento del fascicolo successivo. Decorso tale termine, si spediscono solo contro rimessa dell'importo. Le variazioni di indirizzo devono essere segnalate dall'abbonato per iscritto. Nel sottoscrivere l'abbonamento cumulativo, gli interessati possono chiedere che l'ISTAT provveda, senza ulteriori richieste, all'invio di tutte le pubblicazioni non periodiche non appena liberate dalle stampe, contro assegno o con emissione di fattura, con lo sconto del 30%. Le singole pubblicazioni possono essere richieste direttamente all'Istituto nazionale di statistica (Via Cesare Balbo, 16 - 00100 Roma) versando il relativo importo, maggiorato del 10% per spese di spedizione, sul c/c postale n. 619007.

Tutti i prezzi sono riferiti all'anno 1991.

ANNUARIO STATISTICO ITALIANO - Edizione 1990 - L. 46.000

Sintetizza in semplici tabelle numeriche di facile lettura ed attraverso appropriate note illustrative e rappresentazioni grafiche, i dati fondamentali della vita economica, demografica e sociale e fornisce un quadro panoramico della corrispondente situazione degli altri principali Paesi del mondo.

COMPENDIO STATISTICO ITALIANO - Edizione 1990 - L. 22.000

Sintetizza i risultati delle rilevazioni ed elaborazioni statistiche di maggior interesse nazionale.

ITALIAN STATISTICAL ABSTRACT - Edition 1990 - L. 22.000

Fornisce i principali risultati delle rilevazioni ed elaborazioni statistiche concernenti la situazione sociale ed economica italiana - Edizione in lingua inglese.

I CONTI DEGLI ITALIANI - Vol. 24, edizione 1990 - L. 16.000

Illustra in forma divulgativa i principali aspetti quantitativi dell'economia italiana.

LE REGIONI IN CIFRE - Edizione 1990 - Distribuzione gratuita

Fornisce i dati delle singole regioni e delle due grandi ripartizioni geografiche: Nord-Centro e Mezzogiorno.

ANNUARI

STATISTICHE DEMOGRAFICHE
n. 33 - Anno 1984
Tomo I, parte prima - Movimento e calcolo della popolazione secondo gli atti anagrafici - L. 11.000
Tomo I, parte seconda - Iscrizioni e cancellazioni anagrafiche per trasferimento di residenza, 1983 - Espatriati e rimpatriati, 1984 - L. 9.000
n. 34 - Anno 1985
Tomo I, parte prima - Movimento e calcolo della popolazione secondo gli atti anagrafici - L. 11.000
Tomo I, parte seconda - Iscrizioni e cancellazioni anagrafiche per trasferimento di residenza, 1984 - Espatriati e rimpatriati, 1985 - L. 9.500
n. 33/34 - Anni 1984 e 1985
Tomo 2, parte prima - Nascite e decessi - L. 38.000
Tomo 2, parte seconda - Matrimoni, separazioni e divorzi - L. 15.000
n. 35 - Anno 1986
Tomo I, parte prima - Popolazione residente e movimento anagrafico dei Comuni - L. 11.500
Tomo I, parte seconda - Iscrizioni e cancellazioni anagrafiche, 1985 e 1986 - Espatriati e rimpatriati, 1986 - L. 15.800
n. 36 - Anno 1987
Tomo I, parte prima - Popolazione residente e movimento anagrafico dei Comuni - L. 18.900
Tomo I, parte seconda - Iscrizioni e cancellazioni anagrafiche - Espatri e rimpatri, 1987 - L. 15.000
Raccoglie i dati sulla dinamica demografica italiana, sia naturale che migratoria, nonché dei dati sintetici sul movimento annuale della popolazione residente anagrafica comunale e sul suo ammontare.

POPOLAZIONE E MOVIMENTO ANAGRAFICO DEI COMUNI - n. 2 - Anno 1989 - L. 20.000
Riporta i dati dell'ammontare della popolazione residente, desunti dall'analisi del movimento naturale e di quello migratorio, nonché la stima della popolazione residente per sesso ed età a livello regionale.

STATISTICHE DELLA SANITÀ - n. 3 - Anno 1987 - L. 23.000
Riunisce le statistiche sulle strutture e sull'attività degli Istituti di cura, sulle malattie infettive e diffusive soggette a denuncia obbligatoria, sulle interruzioni volontarie della gravidanza e sugli aborti spontanei.

CAUSE DI MORTE - n. 3 - Anno 1987 - L. 25.000
Raccoglie i dati relativi alle statistiche sulle cause di morte e di nati-mortalità.

STATISTICHE DELLA PREVIDENZA, DELLA SANITÀ E DELL'ASSISTENZA SOCIALE
n. 28 - Anni 1987, 1988 - L. 20.000
Vengono illustrate alcune forme di attività svolte dai vari Istituti nel settore della previdenza sociale, i conti economici delle Unità Sanitarie Locali e degli Istituti ospedalieri pubblici, nonché i principali aspetti dell'assistenza sociale.

STATISTICHE DELL'ISTRUZIONE - n. 40 - Anno scolastico 1986-87
Tomo I - Dati analitici: nazionali, regionali e provinciali - L. 23.000
Tomo 2 - Dati riassuntivi comunali - L. 18.000
Quadro statistico completo ed aggiornato della situazione scolastica del Paese, attraverso dati sui vari rami d'insegnamento esaminati sotto i più interessanti aspetti dell'ordinamento degli studi e dei risultati conseguiti dagli iscritti.

STATISTICHE CULTURALI - n. 29 - Anno 1987 - L. 14.000
Documentazione ufficiale completa sulle principali attività culturali concernenti, tra l'altro, la produzione libraria, la pubblicazione di riviste scientifiche, la stampa periodica e le biblioteche.

STATISTICHE GIUDIZIARIE - n. 36 - Anno 1988 - L. 41.000
Ampia documentazione statistica dell'attività giudiziaria nonché dei principali fenomeni in materia civile e penale nel campo della criminalità e degli Istituti di prevenzione e pena.

STATISTICHE DELL'AGRICOLTURA, ZOOTECNIA E MEZZI DI PRODUZIONE - n. 36 - Anno 1988 - L. 41.000
Contiene i dati relativi ai vari aspetti dell'agricoltura nazionale, nonché i dati sulla consistenza e produttività degli allevamenti.

STATISTICHE FORESTALI - n. 40 - Anno 1987 - L. 14.000.
Fornisce un quadro completo sulla struttura delle foreste italiane e delle relative utilizzazioni legnose, unitamente ad alcuni aspetti economici.

STATISTICHE METEOROLOGICHE - n. 24 - Anno 1983 - L. 15.800
Raccoglie i dati relativi alle temperature, piovosità e altri fattori climatici rilevati da una rete di stazioni ed osservatori distribuiti nel territorio nazionale.

STATISTICHE DELLA CACCIA E DELLA PESCA - n. 3 - Anno 1987 - L. 10.000
Raccoglie i dati sull'attività della pesca e sulla consistenza del relativo naviglio, nonché su alcuni aspetti del settore venatorio.

STATISTICHE INDUSTRIALI - n. 28 - Anni 1986 e 1987 - L. 41.000
Nel suo genere, unica e veramente preziosa pubblicazione in cui sono organicamente raccolte tutte le informazioni statistiche fondamentali concernenti il complesso ed importante settore dell'industria.

STATISTICHE DELL'ATTIVITA' EDILIZIA - n. 2 - Anno 1987 - L. 14.000
Fornisce i risultati del settore dell'attività edilizia relativamente ai fabbricati residenziali e non residenziali.

STATISTICHE DELLE OPERE PUBBLICHE - n. 2 - Anno 1987 - L. 10.000
Statistica ufficiale delle opere pubbliche effettuate dallo Stato e da Enti pubblici, nonché da privati con finanziamento parziale dello *Stato*.

STATISTICHE DEL COMMERCIO INTERNO - n. 30 - Anni 1987, 1988 - L. 15.000
Fornisce i risultati delle rilevazioni correnti relativi al fenomeno della distribuzione. Vi figurano gli indici mensili delle vendite al minuto, nonché la più recente distribuzione per Comune delle licenze di esercizio.

STATISTICHE DEL TURISMO - n. 3 - Anno 1988 - L. 11.000
Descrive il sistema delle informazioni statistiche sul turismo ed espone, in un quadro organico, statistiche, dati ed indicatori aventi per oggetto i principali aspetti di questo fenomeno.

STATISTICHE DELLA NAVIGAZIONE MARITTIMA - n. 42 - Anno 1987 - L. 20.000
Contiene i dati statistici sul movimento dei natanti e del relativo carico avvenuto nei porti marittimi e negli altri approdi autorizzati del territorio nazionale.

STATISTICA DEGLI INCIDENTI STRADALI - n. 37 - Anno 1989 - L. 20.000
La più completa ed aggiornata raccolta di dati su una materia di viva attualità.

STATISTICA ANNUALE DEL COMMERCIO CON L'ESTERO - n. 44 - Anno 1987
Tomo I - Dati generali e riassuntivi - L. 41.000
Tomo 2 - Merci per Capitoli merceologici e Paesi
- Parte prima: da Cap. 1 a Cap. 24 - L. 14.000
- Parte seconda: da Cap. 25 a Cap. 40 - L. 18.000
- Parte terza: da Cap. 41 a Cap. 67 - L. 21.000
- Parte quarta: da Cap. 68 a Cap. 83 - L. 18.000
- Parte quinta: da Cap. 84 a Cap. 85 - L. 25.000
- Parte sesta: da Cap. 86 a Cap. 99 - L. 18.000
- Appendice: L. 10.000
Riporta i dati definitivi sull'andamento delle importazioni e delle esportazioni con l'analisi completa del movimento per merci e per Paesi. Nel tomo primo è riportata, tra l'altro, un'ampia documentazione sul movimento delle merci nei depositi doganali e sul commercio di transito.

STATISTICHE DEI BILANCI DELLE AMMINISTRAZIONI REGIONALI, PROVINCIALI E COMUNALI - n. XXVII - Anno 1982 - L. 14.000
Espone i dati relativi ai bilanci delle Amministrazioni, tenendo conto dell'aspetto contabile, funzionale ed amministrativo dei documenti contabili. Per le Amministrazioni provinciali e comunali è stata dedicata particolare attenzione ai dati riguardanti i servizi sociali, i settori d'intervento nel campo economico ed il personale.

STATISTICHE DEL LAVORO - n. 26 - Anno 1984 - L. 12.000
Organica ed aggiornata documentazione statistica su tutti i principali aspetti del *mondo del lavoro*.

CONTABILITÀ NAZIONALE - n. 15 - Anni 1960-85 - L. 17.000
Contiene i dati sulla struttura e sulla evoluzione delle principali grandezze del sistema economico italiano.

COLLANA D'INFORMAZIONE

Anno 1990

n. 4 - STRUTTURA E POTENZIALE PRODUTTIVO DELLE PRINCIPALI COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 - L. 15.000
n. 5 - STRUTTURA DELLE AZIENDE AGRICOLE - Anno 1986 - L. 27.000
n. 6 - STATISTICHE DEL MOVIMENTO DELLA NAVIGAZIONE NEI PORTI ITALIANI - Anno 1987 - L. 11.000
n. 7 - LA DISTRIBUZIONE QUANTITATIVA DEL REDDITO IN ITALIA NELLE INDAGINI SUI BILANCI DI FAMIGLIA - Anno 1988 - L. 11.000
n. 8 - BILANCI CONSUNTIVI DELLE AMMINISTRAZIONI PROVINCIALI E COMUNALI - Anno 1986 - L. 20.000
n. 9 - STRUTTURA DELLE AZIENDE AGRICOLE - Anno 1987 - L. 27.000
n. 10 - CONTI ECONOMICI NAZIONALI - Anni 1970-89 - L. 11.000
n. 11 - STATISTICHE DELLA ZOOTECNIA E DEI MEZZI DI PRODUZIONE IN AGRICOLTURA - Anno 1988 - L. 11.000
n. 12 - CONTI DELLE AMMINISTRAZIONI PUBBLICHE E DELLA PROTEZIONE SOCIALE - Anni 1983-88 - L. 15.000
n. 13 - STATISTICHE DEI SUICIDI E DEI TENTATIVI DI SUICIDIO - Anni 1984-88 - L. 11.000
n. 14 - INDAGINE STATISTICA SULL'INNOVAZIONE TECNOLOGICA NELL'INDUSTRIA ITALIANA - Anni 1981-85 - L. 20.000
n. 15 - CONTI ECONOMICI DELLE IMPRESE CON ADDETTI DA 10 A 19 - Anno 1987 - L. 11.000
n. 16 - LE AZIENDE AGRICOLE SECONDO LA CLASSIFICAZIONE TIPOLOGICA - Anno 1986 - L. 20.000
n. 17 - INDAGINE 1989 SUGLI SBOCCHI PROFESSIONALI DEI LAUREATI - L. 20.000
n. 18 - RILEVAZIONE DELLE FORZE DI LAVORO - Ottobre 1989 - L. 11.000
n. 19 - STATISTICHE SUI TRATTAMENTI PENSIONISTICI AL 31 DICEMBRE 1988 - L. 11.000
n. 20 - RILEVAZIONE DELLE FORZE DI LAVORO - Media 1989 - L. 20.000
n. 21 - CONTI ECONOMICI REGIONALI - Anni 1980-87 - L. 20.000
n. 22 - OCCUPAZIONE E REDDITI DA LAVORO DIPENDENTE - Anni 1980-89 - L. 20.000
n. 23 - STATISTICHE DEL MOVIMENTO DELLA NAVIGAZIONE NEI PORTI ITALIANI - Anno 1988 - L. 11.000
n. 24 - LE AZIENDE AGRICOLE SECONDO LA CLASSIFICAZIONE TIPOLOGICA - Anno 1987 - L. 20.000
n. 25 - VALORE AGGIUNTO DELL'AGRICOLTURA PER REGIONE - Anni 1980-89 - L. 11.000
n. 26 - STATISTICHE SUL TRASPORTO AEREO - Anni 1987, 1988 - L. 11.000
n. 27 - STATISTICHE DELL'AGRICOLTURA E DEI MEZZI DI PRODUZIONE - Anno 1989 - L. 11.000
n. 28 - CONTI NAZIONALI ECONOMICI E FINANZIARI DEI SETTORI ISTITUZIONALI - Anni 1983-87 - L. 11.000

NOTE E RELAZIONI

Anno 1989

n. 1 - MANUALE DI TECNICHE DI INDAGINE (n. 7 fascicoli)
1. Pianificazione della produzione dei dati - L. 10.000 - 2. Il questionario: progettazione, redazione e verifica - L. 11.000 - 3. Tecniche di somministrazione del questionario - L. 11.000 - 4. Tecniche di campionamento: teoria e pratica - L. 20.000 - 5. Tecniche di stima della varianza campionaria *(in corso di stampa)* - 6. Il sistema di controllo della qualità dei dati *(in corso di stampa)* - 7. Le rappresentazioni grafiche di dati statistici - L. 15.000
n. 2 - DISTRIBUZIONE PER ETÀ DELLA POPOLAZIONE SCOLASTICA - Anno scolastico 1984-85 - L. 10.000
n. 3 - LA CRIMINALITÀ ATTRAVERSO LE STATISTICHE - Anni 1971-87 - L. 14.000
n. 4 - PREVISIONI DELLA POPOLAZIONE RESIDENTE PER SESSO, ETÀ E REGIONE - Base 1-1-1988
Tomo 1 - L. 18.000
Tomo 2 - L. 38.000
n. 5 - STATISTICHE SUI MINORENNI - Anni 1984-86 - L. 18.000
n. 6 - ANALISI DELLE FONTI STATISTICHE PER LA MISURA DELL'IMMIGRAZIONE STRANIERA IN ITALIA: ESAME E PROPOSTE - L. 10.000
n. 7 - NUMERI INDICI DEI PREZZI ALLA PRODUZIONE DEI PRODOTTI INDUSTRIALI - Base 1980 = 100 - L. 10.000

Anno 1990

n. 1 - METODOLOGIA E ANALISI DEI RISULTATI DELL'INDAGINE SULLE COLTIVAZIONI LEGNOSE AGRARIE - Anno 1987 - L. 11.000
n. 2 - LA MORTALITÀ DIFFERENZIALE SECONDO ALCUNI FATTORI SOCIO-ECONOMICI - Anni 1981-82 - L. 11.000

METODI E NORME

Serie A

n. 18 - NUMERI INDICI DEL COSTO DI COSTRUZIONE DI UN FABBRICATO RESIDENZIALE: Base 1976 = 100 - L. 1.500
n. 20 - NUMERI INDICI DEI PREZZI: Base 1980 = 100 - L. 4.500
n. 21 - NUMERI INDICI DEI PREZZI DEI PRODOTTI VENDUTI E DEI BENI ACQUISTATI DAGLI AGRICOLTORI: Base 1980 = 100 - L. 5.000
n. 23 - NUMERI INDICI DEI PREZZI AL CONSUMO: Base 1985 = 100 - L. 6.300
n. 25 - NUMERI INDICI DELLA PRODUZIONE INDUSTRIALE: Base 1985 = 100 - L. 11.000
n. 26 - NUMERI INDICI DEI PREZZI ALLA PRODUZIONE DEI PRODOTTI INDUSTRIALI: Base 1980 = 100 - L. 11.000
n. 27 - NUMERI INDICI DEL FATTURATO, DEGLI ORDINATIVI E DELLA CONSISTENZA DEGLI ORDINATIVI: Base 1985 = 100 - L. 11.000

Serie B

n. 21 - ISTRUZIONI PER LA RILEVAZIONE STATISTICA DEL MOVIMENTO DELLA POPOLAZIONE L. 4.000
n. 22 - ISTRUZIONI PER LA RILEVAZIONE DEI DATI DELLE STATISTICHE FORESTALI - L. 6.000
n. 23 - ISTRUZIONI PER LA RILEVAZIONE DELL'ATTIVITÀ EDILIZIA - L. 8.400
n. 24 - ISTRUZIONI PER LE RILEVAZIONI DELLE STATISTICHE GIUDIZIARIE
Tomo 1 - Procedura di rilevazione - L. 15.800
Tomo 2 - Modelli di rilevazione - L. 15.800
n. 25 - MANUALE PER LA PROGETTAZIONE DEI DATI STATISTICI - L. 10.000
n. 26 - ISTRUZIONI PER LE COMMISSIONI COMUNALI DI CONTROLLO DELLE RILEVAZIONI DEI PREZZI AL CONSUMO - L. 10.000
n. 27 - ISTRUZIONI PER LA RILEVAZIONE DELLE OPERE PUBBLICHE - L. 11.000
n. 28 - ISTRUZIONI PER LA RILEVAZIONE STATISTICA DEGLI INCIDENTI STRADALI - L. 11.000

Serie C

n. 8 - CLASSIFICAZIONE DELLE ATTIVITÀ ECONOMICHE - L. 6.500
n. 9 - CLASSIFICAZIONE DELLE PROFESSIONI - L. 6.500
n. 10 - CLASSIFICAZIONI DELLE MALATTIE, TRAUMATISMI E CAUSE DI MORTE - Ristampa 1986
Vol. 1: Introduzione e parte sistematica - L. 16.000
Vol. 2: Indici alfabetici - L. 25.000

ANNALI DI STATISTICA

Serie IX

Vol. 3 - STUDI STATISTICI SUI CONSUMI - Edizione 1983 - Dati dal 1959 al 1974 - L. 9.500
Vol. 4 - CONTABILITÀ NAZIONALE - FONTI E METODI - L. 9.000 *(esaurito)*
Vol. 5 - ATTI DEL SEMINARIO SULLA VALUTAZIONE DEI RISULTATI E DELLA METODOLOGIA DEI CENSIMENTI (Roma, 7-11 maggio 1984) - L. 25.000
Vol. 6 - ATTI DEL CONVEGNO «LA FAMIGLIA IN ITALIA» (Roma, 29-30 ottobre 1985) - L. 14.000
Vol. 7 - ATTI DEL CONVEGNO SULL'INFORMAZIONE STATISTICA E I PROCESSI DECISIONALI (Roma, 11-12 dicembre 1986) - L. 15.000
Vol. 8 - ATTI DEL SEMINARIO SULLE STATISTICHE ECOLOGICHE (Roma, 28 marzo-1 aprile 1988) - *(in corso di stampa)*
Vol. 9 - NUOVA CONTABILITÀ NAZIONALE Roma, 1990 - L. 23.000

CENSIMENTI

12° CENSIMENTO GENERALE DELLA POPOLAZIONE - 25 ottobre 1981
DATI SULLE CARATTERISTICHE STRUTTURALI DELLA POPOLAZIONE E DELLE ABITAZIONI - Campione al 2% dei fogli di famiglia - Dati provvisori - L. 5.000
Vol. I - Primi risultati provinciali e comunali sulla popolazione e sulle abitazioni *(dati provvisori)* - L. 6.500
Vol. II - Dati sulle caratteristiche strutturali della popolazione e delle abitazioni:
Tomo 1 - Fascicoli provinciali - Prezzi vari
Tomo 2 - Fascicoli regionali - Prezzi vari
Tomo 3 - Fascicolo nazionale - Italia - L. 25.000

Vol. III - Popolazione delle frazioni geografiche e delle località abitate dei comuni - Fascicoli regionali e nazionale - Prezzi vari
Vol. IV - Atti del censimento - L. 26.500
Vol. V - Relazione generale sul censimento - L. 25.000

POPOLAZIONE LEGALE DEI COMUNI - L. 8.000

6° CENSIMENTO GENERALE DELL'INDUSTRIA, DEL COMMERCIO, DEI SERVIZI E DELL'ARTIGIANATO - 26 ottobre 1981

Vol. I - Primi risultati sulle imprese e sulle unità locali - Dati provvisori
Tomo 1 - Dati nazionali, regionali e provinciali *(esaurito)*
Tomo 2 - Dati comunali *(esaurito)*
Vol. II - Dati sulle caratteristiche strutturali delle imprese e delle unità locali
Tomo 1 - Fascicoli provinciali - Prezzi vari
Tomo 2 - Fascicoli regionali - Prezzi vari
Tomo 3 - Fascicolo nazionale - Italia - L. 14.000
Vol. III - Atti del censimento - L. 11.000
Vol. IV - Relazione generale sul censimento - L. 26.500

3° CENSIMENTO GENERALE DELL'AGRICOLTURA - 24 ottobre 1982

CARATTERISTICHE STRUTTURALI DELLE AZIENDE AGRICOLE - L. 14.000
Vol. I - Primi risultati provinciali e comunali - Dati provvisori - L. 8.000
Vol. II - Caratteristiche strutturali delle aziende agricole:
Tomo 1: Fascicoli provinciali - Prezzi vari
Tomo 2: Fascicoli regionali - Prezzi vari
Tomo 3: Fascicolo nazionale - Italia - L. 11.000
Vol. III - Atti del censimento - L. 33.500

TIPOLOGIA DELLE AZIENDE AGRICOLE - Campione al 10% dei questionari d'azienda - L. 6.000

INDAGINE SULLE SUPERFICI A VITE
Vol. I - Caratteristiche delle aziende con vite
Tomo 1: Dati provinciali, regionali e nazionali - L. 33.500
Tomo 2: Dati comunali - L. 15.000
Vol. II - Caratteristiche dei vitigni - L. 33.500

L'ITALIA DEI CENSIMENTI - L. 10.000

ALTRE

INFORMAZIONE STATISTICA - Parliamone con l'ISTAT - Edizione 1988 - L. 12.000
CONOSCERE L'ITALIA - INTRODUCING ITALY - Edizione 1990 - Distribuzione gratuita
SOMMARIO DI STATISTICHE STORICHE - 1926-1985 - L. 35.000
ATLANTE STATISTICO ITALIANO 1988 - L. 50.000
COMUNI, COMUNITÀ MONTANE, REGIONI AGRARIE AL 31 DICEMBRE 1988 - Edizione 1990 - L. 20.000
STATISTICHE AMBIENTALI - Vol. I, 1984 - L. 9.000 *(esaurito)*
POPOLAZIONE RESIDENTE E PRESENTE DEI COMUNI - Censimenti dal 1861 al 1981 - L. 14.000
SOMMARIO STORICO DI STATISTICHE SULLA POPOLAZIONE - Anni 1951-1987 - L. 41.000
IMMAGINI DELLA SOCIETÀ ITALIANA - Edizione 1988 - L. 30.000
SINTESI DELLA VITA SOCIALE ITALIANA - Edizione 1990 - L. 15.000
MORTALITÀ PER CAUSA E UNITÀ SANITARIA LOCALE - Anni 1980-82 - L. 35.000
ELEZIONI DELLA CAMERA DEI DEPUTATI E DEL SENATO DELLA REPUBBLICA, 14 giugno 1987 - L. 10.000
45 ANNI DI ELEZIONI IN ITALIA 1946-90 - Edizione 1990 - L. 20.000
IL VALORE DELLA LIRA DAL 1861 al 1982 - L. 5.000
STATISTICHE SULLA AMMINISTRAZIONE PUBBLICA - Anni 1985-87 - L. 21.000

Stampato a Roma nel mese di gennaio 1991
abete grafica s.p.a. - via Prenestina, 683