I lavori raccolti in questa collana hanno avuto origine nell'ambito del progetto di ricerca dell'ISTAT «Le statistiche dei consumi privati nel sistema statistico nazionale» e del progetto di ricerca MURST 40% «La misura dei consumi privati: uno studio sull'accuratezza, coerenza e qualità dei dati».
Al progetto di ricerca hanno partecipato i ricercatori dell'ISTAT e dei seguenti Dipartimenti e Istituti universitari:

- Dipartimento di Scienze Statistiche, Bologna
- Dipartimento Statistico, Firenze
- Dipartimento di Scienze Statistiche, Perugia
- Dipartimento di Contabilità Nazionale, Roma
- Istituto di Statistica e Matematica, Istituto Universitario Navale, Napoli
- Istituto di Statistica, Messina.

LA MANCATA RISPOSTA TOTALE NEI CAMPIONI COMPLESSI: UN'APPLICAZIONE ALL'INDAGINE CAMPIONARIA SUI CONSUMI DELLE FAMIGLIE

Piero Demetrio Falorsi* Aldo Russo*

Rapporto di ricerca n.23

CON PRI - La misura dei consumi privati

* Servizio Studi Metodologici - Istituto Nazionale di Statistica

Dipartimento di Scienze Statistiche "Paolo Fortunati"
dell' Università degli Studi di Bologna
Settembre 1992

Finito di stampare nel mese di Settembre 1992
presso le Officine Grafiche TECNOPRINT S.N.C.
Via del Legatore 3, Bologna.

INDICE

1. Introduzione

Nel corso di un'indagine campionaria, qualunque sia il metodo di raccolta delle informazioni (intervista diretta, telefonica, postale, ecc.) ed il tipo di indagine (occasionale, trasversale ripetuta, parziale, ecc.), si deve far fronte agli errori di rilevazione.

Un errore di rilevazione frequente e' la mancanza di espressione del dato (*dato mancante*), che si manifesta come:

(i) assenza di partecipazione alla rilevazione da parte dell'unita' statistica (*mancata risposta totale*) e per questo mancheranno tutte le informazioni inerenti all'unita';
(ii) assenza di un valore validamente espresso per una singola variabile (*mancata risposta parziale*).

Le mancate risposte totali, oggetto del presente lavoro, hanno un duplice effetto sull'affidabilita' delle stime fornite da un'indagine campionaria ; esse, infatti, producono un aumento della varianza campionaria ed introducono una distorsione. Il primo effetto e' dovuto alla ridotta numerosita' del campione; il secondo effetto nasce dalla circostanza che, in generale, la popolazione dei mancati rispondenti ha - sulla variabile oggetto di studio - una distribuzione diversa da quella della popolazione dei rispondenti.

Seguendo la copiosa bibliografia sull'argomento (Kalsbeek 1980; Little 1982; Little 1986; Oh e Scheuren 1983; Platek e Gray 1985), per il trattamento della mancata risposta totale emergono due diversi approcci, noti in letteratura con i nomi di *modello di risposta aleatorio* e *modello di risposta fisso*. Il primo approccio assume che ogni unita' della popolazione ha una certa probabilita' di risposta se l'unita' viene selezionata; nel secondo approccio, invece, la popolazione viene suddivisa in due strati identificati rispettivamente con le locuzioni *strato dei rispondenti* e *strato dei non rispondenti*. Le unita' della popolazione appartenenti al primo strato partecipano con certezza all'indagine se incluse nel campione; quelle del secondo strato non partecipano con certezza all'indagine se incluse nel campione. Il modello di risposta fisso, pertanto, puo' essere riguardato come un caso particolare del modello di risposta aleatorio in cui le probabilita' di risposta sono 0 ed 1.

Nel nostro studio, fondato sul modello di risposta fisso, dopo aver descritto gli aspetti caratterizzanti il disegno campionario di riferimento (par. 2) e la struttura formale del parametro oggetto di stima $\overline{Y}$ (par. 3), si espone una decomposizione di $\overline{Y}$ espressa in funzione dei rispondenti e dei non rispondenti (par. 4). Nei successivi paragrafi, vengono derivate le espressioni della distorsione e della varianza dello stimatore diretto (par. 5), dello stimatore post-stratificato (par. 6) e dello stimatore del rapporto post-stratificato (par. 7).

La trattazione si conclude (par. 8) con uno studio empirico basato sull'utilizzazione di dati desunti dall'indagine Istat sui consumi delle famiglie.

2. Disegno campionario

Nella maggior parte degli studi condotti sul tema delle mancate risposte totali si assume che il disegno campionario sia casuale semplice.

In questo lavoro, allo scopo di offrire una metodologia statistica piu' aderente alla realta' di ricerca, svilupperemo i nostri asserti con riferimento ad un disegno campionario complesso a due stadi di selezione, con stratificazione delle unita' primarie.

Le unita' primarie (UP) sono costituite da grappoli di unita' secondarie (US); queste ultime sono costituite da grappoli di unita' finali.

Il meccanismo probabilistico di selezione del campione prevede la selezione delle UP con probabilita' variabile e senza reimmissione e quello delle US con probabilita' uguale e senza reimmissione; tutte le unita' finali, appartenenti alle US selezionate, sono incluse nel campione.

Riteniamo utile sottolineare che il disegno campionario sopra illustrato e' di rilevante interesse essendo alla base di importanti indagini nazionali, quali sono - ad esempio - quelle condotte dall'Istat sulle forze di lavoro e sui consumi delle famiglie.

Il campione delle suddette indagini e' infatti formato su due stadi (i comuni italiani, le famiglie residenti), con stratificazione delle unita' di primo stadio, selezione PPS delle unita' primarie e sistematica di quelle di secondo stadio; tutti i membri delle famiglie selezionate sono inclusi nel campione.

3. Parametro oggetto di stima

Supponiamo che mediante un'indagine, basata sul disegno campionario descritto nel precedente paragrafo, si voglia ottenere una stima della media $\overline{Y}$, per unita' secondaria, della variabile oggetto di studio y.

Ai fini dei successivi sviluppi algebrici, esprimiamo $\overline{Y}$ utilizzando il simbolismo del disegno di campionamento considerato.

Introduciamo, quindi, le seguenti notazioni simboliche:

h = indice di strato ($h=1,\dots,H$)
i = indice di UP
j = indice di US
N_h = numero di UP nello strato h
n_h = numero di UP campione nello strato h
M_{hi} = numero di US nella UP (hi)
m_{hi} = numero di US campione nella UP (hi)
M = numero di US nell'intera popolazione
Y_{hij} = valore della variabile di studio y nella US (hij)

Il valore medio $\overline{Y}$ - per unita' secondaria - puo' essere pertanto espresso nella forma:

$$\overline{Y} = \frac{1}{M} \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{hij} \tag{1}$$

4. Decomposizione di $\overline{Y}$

Suddividiamo le M_{hi} US della UP (hi) in L sottoclassi definite sulla base dei valori di una determinata variabile (ad esempio, le famiglie - considerate come US - possono essere raggruppate secondo l'ampiezza). Assumiamo, inoltre, che ciascuna sottoclasse sia divisa in due strati:

- il primo, chiamato *strato dei rispondenti* (SR), consiste di tutte le unita' sulle quali, se incluse nel campione, puo' essere rilevata la variabile y;

- il secondo, chiamato *strato dei non rispondenti* (SNR), si compone di tutte le unita' sulle quali non e' possibile rilevare la variabile y.

Relativamente alla sottoclasse l (l=1,...,L), indichiamo con ${}_rM_{hil}$ il numero di US in SR e con ${}_{\bar{r}}M_{hil}$ il numero di US in SNR; in tal modo e' possibile riscrivere la (1) nella forma:

$$\bar{Y} = \frac{1}{M}\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{l=1}^{L}\left(\sum_{j=1}^{{}_rM_{hil}} Y_{hilj} + \sum_{j=1}^{{}_{\bar{r}}M_{hil}} Y_{hilj}\right) \qquad (2)$$

dove:

$$Y_{hilj} = Y_{hij}\ {}_l\delta_{hij}$$

essendo ${}_l\delta_{hij}$ una variabile dicotomica che assume valore 1 se l'US (hij) appartiene alla sottoclasse l, e valore 0 altrimenti.

Indichiamo con:

$${}_r\bar{Y}_{hil} = \frac{1}{{}_rM_{hil}}\sum_{j=1}^{M_{hil}} Y_{hilj}\ {}_r\delta_{hij}$$

$${}_{\bar{r}}\bar{Y}_{hil} = \frac{1}{{}_{\bar{r}}M_{hil}}\sum_{j=1}^{M_{hil}} Y_{hilj}\ {}_{\bar{r}}\delta_{hij}$$

$$t_{hil} = \frac{{}_rM_{hil}}{M_{hil}} \quad , \quad W_{hi} = \frac{M_{hi}}{M} \quad , \quad W_{hil} = \frac{M_{hil}}{M_{hi}}$$

in cui: ${}_r\delta_{hij}$ denota una variabile indicatrice che e' uguale ad 1 se l'US (hij) appartiene allo strato SR, e valore 0 altrimenti; ${}_{\bar{r}}\delta_{hij} = 1 - {}_r\delta_{hij}$. Utilizzando le precedenti relazioni e' possibile esprimere la (2) nella seguente forma alternativa:

$$\bar{Y} = \sum_{h=1}^{H}\sum_{i=1}^{N_h} W_{hi}\sum_{l=1}^{L} W_{hil}\left(t_{hil}\ {}_r\bar{Y}_{hil} + (1 - t_{hil})\,{}_{\bar{r}}\bar{Y}_{hil}\right)$$

5. Stimatore diretto

Nelle indagini concrete, basate sul disegno campionario qui considerato, per stimare il valor medio $\bar{Y}$ - in presenza di mancata risposta totale - viene spesso usato lo stimatore diretto espresso da:

$${}_d\hat{\bar{Y}} = \frac{1}{M}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{M_{hi}}{\pi_{hi}\ {}_rm_{hi}}\sum_{j=1}^{m_{hi}} Y_{hij}\ {}_r\delta_{hij} \qquad (3)$$

in cui: ${}_rm_{hi}$ indica il numero di US campione rispondenti nella UP (hi); π_{hi} denota la probabilita' di inclusione della UP (hi).

5.1 Distorsione

Denotiamo con ${}_rm_{hil}$ il numero di US campione rispondenti in (hil) ed introduciamo le seguenti ipotesi:

$$\Pr(\ {}_rm_{hi} > 0) = 1 \ (h=1,\dots,H\ ;\ i=1,\dots,N_h)$$

$$\Pr(\ {}_rm_{hil} > 0) = 1 \ (h=1,\dots,H\ ;\ i=1,\dots,N_h\ ; l=1,\dots,L)$$

dove $\Pr({}_rm_{hi} > 0)$ e $\Pr({}_rm_{hil} > 0)$ indicano le probabilita' che le quantita' ${}_rm_{hi}$ e ${}_rm_{hil}$ siano maggiori di 0; esse risultano valide nel caso in cui si adotti un campione sufficientemente numeroso in ciascuna UP. Tuttavia, nelle appendici vengono fornite le espressioni esatte relative ai valori attesi ed alle varianze degli stimatori oggetto di studio, derivate rimuovendo le suddette ipotesi.

Cio' detto, sfruttando la condizione $\Pr(\ {}_rm_{hi} > 0) = 1$, il valore atteso dello stimatore diretto e' definito da (si veda Appendice A1):

$$E({}_d\hat{\bar{Y}}) = \sum_{h=1}^{H}\sum_{i=1}^{N_h} W_{hi}\ {}_r\bar{Y}_{hi} \qquad (4)$$

essendo:

$$ {}_r\overline{Y}_{hi} = \frac{1}{{}_rM_{hi}} \sum_{j=1}^{M_{hi}} Y_{hij} \; {}_r\delta_{hij} \quad ; \quad {}_rM_{hi} = \sum_{l=1}^{L} {}_rM_{hil} \quad . $$

Il valore medio ${}_r\overline{Y}_{hi}$ puo' essere espresso come:

$$ {}_r\overline{Y}_{hi} = \frac{1}{t_{hi}} \sum_{l=1}^{L} t_{hil} \, W_{hil} \; {}_r\overline{Y}_{hil} $$

dove:

$$ t_{hi} = \frac{{}_rM_{hi}}{M_{hi}} $$

Da cio' segue che il valore atteso di ${}_d\hat{\overline{Y}}$ si puo' scrivere:

$$ E({}_d\hat{\overline{Y}}) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \frac{W_{hi}}{t_{hi}} \sum_{l=1}^{L} t_{hil} \, W_{hil} \; {}_r\overline{Y}_{hil} \tag{5} $$

Conseguentemente la distorsione $B\,({}_d\hat{\overline{Y}})$ di ${}_d\hat{\overline{Y}}$ e' data da:

$$ B({}_d\hat{\overline{Y}}) = E({}_d\hat{\overline{Y}}) - \overline{Y} = B_1 + B_2 \tag{6} $$

in cui:

$$ B_1 = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \frac{W_{hi}}{t_{hi}} \sum_{l=1}^{L} W_{hil} \, (t_{hil} - t_{hi}) \; {}_r\overline{Y}_{hil} \tag{7} $$

$$ B_2 = \sum_{h=1}^{H} \sum_{i=1}^{N_h} W_{hi} \sum_{l=1}^{L} W_{hil} \, (1 - t_{hil}) \, ({}_r\overline{Y}_{hil} - {}_{\bar{r}}\overline{Y}_{hil}) \tag{8} $$

La formula (6) mostra che la distorsione dello stimatore diretto puo' essere decomposta nella somma di due componenti: la prima, B_1, deriva dal fatto che le sottoclassi di ciascuna UP presentano differenti tassi di risposta; la seconda, B_2, e' dovuta alla circostanza che, in ciascuna UP e per ogni sottoclasse l, i rispondenti ed i non rispondenti assumono valori medi diversi.

E' possibile dimostrare che uno stimatore consistente della componente B_1 e' dato da:

$$ \hat{B}_1 = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{W_{hi}}{\pi_{hi} \; \hat{t}_{hi}} \sum_{l=1}^{L} W_{hil} \, (\hat{t}_{hil} - \hat{t}_{hi}) \; {}_r\hat{\overline{Y}}_{hil} \tag{9} $$

dove:

$$ \hat{t}_{hi} = \frac{{}_rm_{hi}}{m_{hi}} \quad ; \quad \hat{t}_{hil} \; \frac{{}_rm_{hil}}{m_{hil}} \quad ; \quad {}_r\hat{\overline{Y}}_{hil} = \frac{{}_r\hat{Y}_{hil}}{{}_r\hat{M}_{hil}} $$

$$ {}_r\hat{Y}_{hil} = \frac{M_{hi}}{m_{hi}} \sum_{j=1}^{m_{hi}} Y_{hilj} \; {}_r\delta_{hij} \; ; \quad {}_r\hat{M}_{hil} = \frac{M_{hi}}{m_{hi}} \; {}_rm_{hil} \, . $$

5.2 Varianza

Sotto l'ipotesi $Pr({}_rm_{hi} > 0) = 1$, la varianza di ${}_d\hat{\overline{Y}}$ e' data da (si veda Appendice A1):

$$ V({}_d\hat{\overline{Y}}) \doteq \sum_{h=1}^{H} \sum_{i=1}^{N_h} \left(\sum_{i'>i}^{N_h} (\pi_{hi} \, \pi_{hi'} - \pi_{hii'}) \left(\frac{W_{hi}}{\pi_{hi}} \, {}_r\overline{Y}_{hi} - \frac{W_{hi'}}{\pi_{hi'}} \, {}_r\overline{Y}_{hi'} \right)^2 + \right. $$

$$ \left. + \frac{W^2_{hi}}{\pi_{hi}} \, \frac{(M_{hi} - m_{hi}) \, M_{hi}}{(M_{hi} - 1) \; {}_rM^2_{hi} \, m_{hi}} \sum_{j=1}^{M_{hi}} ((Y_{hij} - {}_r\overline{Y}_{hi}) \; {}_r\delta_{hij})^2 \right) \tag{10} $$

in cui $\Pi_{hii'}$ denota la probabilita' di inclusione congiunta delle UP (hi) ed (hi').

Utilizzando il teorema 11.2, riportato in Cochran (1977), si deriva il seguente stimatore consistente della varianza $V({}_d\hat{Y})$:

$$\hat{V}({}_d\hat{Y})=\sum_{h=1}^{H}\sum_{i=1}^{n_h}\Big(\sum_{i'>i}^{n_h}\big(\frac{\Pi_{hi}\,\Pi_{hi'}}{\Pi_{hii'}}-1\big)\big(\frac{W_{hi}}{\Pi_{hi}}\;{}_r\hat{Y}_{hi}-\frac{W_{hi'}}{\Pi_{hi'}}\;{}_r\hat{Y}_{hi'}\big)^2+$$

$$+\quad\frac{W^2_{hi}}{\Pi_{hi}}\;\frac{(M_{hi}-m_{hi})\,M_{hi}}{m_{hi}\,(m_{hi}-1)\;{}_r\hat{M}^2_{hi}}\;\sum_{j=1}^{m_{hi}}\big((Y_{hij}-{}_r\hat{Y}_{hi})\;{}_r\delta_{hij})^2\Big)\qquad(11)$$

dove:

$${}_r\hat{Y}_{hi}=\frac{{}_r\hat{Y}_{hi}}{{}_r\hat{M}_{hi}}\;;\quad {}_r\hat{Y}_{hi}=\frac{M_{hi}}{m_{hi}}\sum_{j=1}^{m_{hi}}Y_{hij}\;\;{}_r\delta_{hij}$$

$${}_r\hat{M}_{hi}=\frac{M_{hi}}{m_{hi}}\;\;{}_rm_{hi}\;.$$

6. Stimatore post-stratificato

Qualora i valori M_{hil} siano conosciuti in ciascuna UP campione, essi possono essere utilizzati per tentare di ridurre la distorsione dovuta alla mancata risposta totale. Definiamo a tal fine lo stimatore post-stratificato (Holt e Smith 1979):

$${}_p\hat{Y}=\frac{1}{M}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{l=1}^{L}\frac{M_{hil}}{\Pi_{hi}\;{}_rm_{hil}}\sum_{j=1}^{m_{hil}}Y_{hilj}\;\;{}_r\delta_{hij}\qquad(12)$$

6.1 Distorsione

Tenendo presente l'ipotesi $Pr({}_rm_{hil}>0)=1$, si dimostra (si veda Appendice A2) che il valore atteso di ${}_p\hat{Y}$ e' dato da:

$$E({}_p\hat{Y})=\sum_{h=1}^{H}\sum_{i=1}^{N_h}W_{hi}\sum_{l=1}^{L}W_{hil}\;\;{}_r\bar{Y}_{hil}\qquad(13)$$

Pertanto la distorsione di ${}_p\hat{Y}$ puo' essere espressa nella forma:

$$B({}_p\hat{Y})=E({}_p\hat{Y}-\bar{Y})=$$

$$=\sum_{h=1}^{H}\sum_{i=1}^{N_h}W_{hi}\sum_{l=1}^{L}W_{hil}\,(1-t_{hil})\,(\,{}_r\bar{Y}_{hil}-{}_{\bar{r}}\bar{Y}_{hil})=B_2\qquad(14)$$

ossia, l'uso dello stimatore post-stratificato consente di eliminare la componente B_1 della distorsione totale dello stimatore diretto.

Sulla base di tale risultato, inoltre, e' possibile trarre le seguenti conclusioni (Thomsen, 1973):

(a) se B_1 ed B_2 presentano il medesimo segno, l'uso dello stimatore post-stratificato non accresce la distorsione;

(b) se le due componenti B_1 ed B_2 hanno segni discordi, lo stimatore post-stratificato riduce la distorsione se e solo se:

$$2\,|\,B_2\,|<|B_1|\;;$$

(c) valendo la condizione (a) o (b) il limite superiore della riduzione della distorsione e' dato da:

$$|\,|B_1+B_2|-|B_1|\,|<|B_1|\quad.$$

6.2 Varianza

Sempre sotto l'ipotesi $Pr(\ {}_r m_{hil} > 0) = 1$, la varianza campionaria dello stimatore ${}_p\hat{\bar{Y}}$ e' data da (si veda Appendice A2):

$$V({}_p\hat{\bar{Y}}) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \left(\sum_{i'>i}^{N_h} (\Pi_{hi}\, \Pi_{hi'} - \Pi_{hii'}) \left(\frac{\bar{X}_{hi}}{\Pi_{hi}} - \frac{\bar{X}_{hi'}}{\Pi_{hi'}} \right)^2 + \right.$$

$$\left. + \quad \frac{M_{hi}(M_{hi}-m_{hi})}{\Pi_{hi}\,(M_{hi}-1)m_{hi}} \sum_{j=1}^{M_{hi}} Z^2{}_{hij} \right) \qquad (15)$$

dove:

$$\bar{X}_{hi} = \frac{1}{M} \sum_{l=1}^{L} M_{hil}\ {}_r\bar{Y}_{hil}$$

$$Z_{hij} = \frac{1}{M} \sum_{l=1}^{L} ((Y_{hilj} - \ {}_r\bar{Y}_{hil})\ {}_l\delta_{hij})\ \ {}_r\delta_{hij} \frac{M_{hil}}{{}_rM_{hil}}$$

Uno stimatore consistente di $V({}_p\hat{\bar{Y}})$ puo' essere ottenuto mediante la seguente espressione:

$$\hat{V}({}_p\hat{\bar{Y}}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \left(\sum_{i'>i}^{n_h} \left(\frac{\Pi_{hi}\ \Pi_{hi'}}{\Pi_{hii'}} - 1 \right) \left(\frac{\hat{\bar{X}}_{hi}}{\Pi_{hi}} - \frac{\hat{\bar{X}}_{hi'}}{\Pi_{hi'}} \right)^2 + \right.$$

$$\left. + \quad \frac{M_{hi}\,(M_{hi}-m_{hi})}{\Pi_{hi}(m_{hi}-1)m_{hi}} \sum_{j=1}^{m_{hi}} \hat{Z}^2{}_{hij} \right) \qquad (16)$$

in cui:

$$\hat{\bar{X}}_{hi} = \frac{1}{M} \sum_{l=1}^{L} \frac{M_{hil}}{{}_rm_{hil}} \sum_{j=1}^{m_{hi}} Y_{hilj}\ \ {}_r\delta_{hij} = \frac{1}{M} \sum_{l=1}^{L} M_{hil}\ \ {}_r\hat{\bar{Y}}_{hil}$$

$$\hat{Z}_{hij} = \frac{1}{M} \sum_{l=1}^{L} ((Y_{hilj} - \ {}_r\hat{\bar{Y}}_{hil})\ {}_l\delta_{hij})\ \ {}_r\delta_{hij} \frac{m_{hil}}{{}_rm_{hil}}$$

7. Stimatore del rapporto post-stratificato

Uno stimatore di rilevante interesse, in quanto alla base della maggior parte delle indagini su larga scala, e' quello noto sotto il nome di stimatore del rapporto post-stratificato.

Nel caso in cui sia noto il totale M_l della popolazione in ciascuna sottoclasse l, lo stimatore in questione - in presenza di mancata risposta totale - e' definito da:

$${}_{rp}\hat{\bar{Y}} = \frac{1}{M} \sum_{l=1}^{L} \frac{\hat{Y}_l}{\hat{M}_l} M_l \qquad (17)$$

in cui:

$$\hat{Y}_l = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{M_{hi}}{\Pi_{hi}\ \ {}_rm_{hi}} \sum_{j=1}^{{}_rm_{hi}} Y_{hilj}$$

$$\hat{M}_l = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{M_{hi}}{\Pi_{hi}\ \ {}_rm_{hi}}\ \ {}_rm_{hil}$$

$$M_l = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} M_{hilj}$$

Riteniamo utile sottolineare che uno stimatore simile ad ${}_{rp}\hat{\bar{Y}}$ viene adottato nell'indagine Istat sui consumi delle famiglie, dove le famiglie vengono post-stratificate secondo l'ampiezza.

7.1 Distorsione

Sotto le approssimazioni

$$E({}_{rp}\hat{\bar{Y}}) \doteq \frac{1}{M} \sum_{l=1}^{L} \frac{E(\hat{Y}_l)}{E(\hat{M}_l)} M_l \quad ; \quad Pr({}_r m_{hi} > 0) = 1 \qquad (18)$$

si ha che il valore atteso di ${}_{rp}\hat{\bar{Y}}$ e' dato da (si veda Appendice A3):

$$E({}_{rp}\hat{\bar{Y}}) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} W_{hi} \sum_{l=1}^{L} {}_r W_{hil} \; {}_r\bar{Y}_{hil} \frac{M_l}{M^*_l} \qquad (19)$$

dove:

$${}_r W_{hil} = \frac{{}_r M_{hil}}{{}_r M_{hi}} \quad ; \quad M^*_l = \sum_{h=1}^{H} \sum_{i=1}^{N_h} M_{hi} \; {}_r W_{hil} \qquad (20)$$

Pertanto, la distorsione di ${}_{rp}\hat{\bar{Y}}$ e' espressa da:

$$B({}_{rp}\hat{\bar{Y}}) = B^*_1 + B_2 \qquad (21)$$

in cui:

B_2 e' definito dalla formula (8) e B^*_1 e' dato da:

$$B^*_1 = \sum_{h=1}^{H} \sum_{i=1}^{N_h} W_{hi} \sum_{l=1}^{L} \left({}_r W_{hil} \frac{M_l}{M^*_l} - W_{hil} \right) {}_r\bar{Y}_{hil} \qquad (22)$$

La componente B^*_1 deriva dal fatto che, in ciascuna UP e per ogni sottoclasse, il rapporto ${}_r W_{hil} \, M_l/M^*_l$ e' in genere diverso da W_{hil}.

Uno stimatore consistente di B^*_1 e' definito da:

$$\hat{B}^*_1 = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{W_{hi}}{\pi_{hi}} \sum_{l=1}^{L} \left({}_r\hat{W}_{hil} \frac{M_l}{\hat{M}_l} - W_{hil} \right) {}_r\hat{\bar{Y}}_{hil} \qquad (23)$$

essendo:

$${}_r\hat{W}_{hil} = \frac{{}_r m_{hil}}{{}_r m_{hi}}$$

7.2 Varianza

Supponiamo ancora valide le (18) e sfruttando, inoltre, le approssimazioni lineari di $\hat{Y}_l$ e $\hat{M}_l$ definite rispettivamente dalle espressioni:

$$\hat{Y}_l \doteq \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{1}{\pi_{hi}} \frac{M_{hi}}{{}_r M_{hi}} \left({}_r\hat{Y}_{hil} - \frac{{}_r Y_{hil}}{{}_r M_{hi}} \, {}_r\hat{M}_{hi} \right)$$

$$\hat{M}_l \doteq \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{1}{\pi_{hi}} \frac{M_{hi}}{{}_r M_{hi}} \left({}_r\hat{M}_{hil} - {}_r W_{hil} \; {}_r\hat{M}_{hi} \right)$$

si ha che la varianza di ${}_{rp}\hat{\bar{Y}}$ e' data da (si veda Appendice A3):

$$V({}_{rp}\hat{\bar{Y}}) =$$

$$= \frac{1}{M^2}\Bigg(\sum_{l=1}^{L} \left(\frac{M_l}{M^*_l}\right)^2 \Big(V(\hat{Y}_l) + R^2_l\, V(\hat{M}_l) - 2\, R_l\, C(\hat{Y}_l\ ,\ \hat{M}_l)\Big) +$$

$$+ \sum_{l'\neq l}^{L} \frac{M_l\, M_{l'}}{M^*_l\, M^*_{l'}} \Big(C(\hat{Y}_l\ , \hat{Y}_{l'}) - R_{l'}\, C(\hat{Y}_l\ , \hat{M}_{l'}\ ) +$$

$$- \ R_l\ \ C\,(\hat{Y}_{l'}\ \ , \hat{M}_l) + \ R_l\, R_{l'}\, C(\hat{M}_l\ , \hat{M}_{l'})\Big)\Bigg) \qquad (24)$$

dove le formule delle covarianze $C(\hat{Y}_l\ , \hat{M}_l)$, $C(\hat{Y}_l\ , \hat{Y}_{l'})$, $C(\hat{Y}_l\ , \hat{M}_{l'})$, $C(\hat{M}_l\ , \hat{M}_{l'})$ sono riportate in appendice (espressioni (A3.9), (A3.10), (A3.11) ed (A3.12)) essendo inoltre;

$$V(\hat{Y}_l) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \Bigg(\sum_{i'>i}^{N_h} (\pi_{hi}\ \pi_{hi'} - \pi_{hii'}) \left(\frac{M_{hi}\ \ {}_rW_{hil}\ \ {}_r\bar{Y}_{hil}}{\pi_{hi}} + \right.$$

$$\left. - \ \frac{M_{hi'}\ \ {}_rW_{hi'l}\ \ {}_r\bar{Y}_{hi'l}}{\pi_{hi'}} \right)^2 +$$

$$+ \ \frac{M_{hi}\ (M_{hi} - m_{hi})}{\pi_{hi}\ (M_{hi} - 1)\ m_{hi}\ \ {}_rM^2_{hi}} \ \sum_{j=1}^{M_{hi}} \left(\left(Y_{hilj} - \frac{{}_rY_{hil}}{{}_rM_{hi}} \right)\ {}_r\delta_{hij} \right)^2 \Bigg) \qquad (25)$$

Uno stimatore consistente della varianza $V({}_{rp}\hat{\bar{Y}})$ e' dato da:

$$\hat{V}({}_{rp}\hat{\bar{Y}}) =$$

$$= \frac{1}{M^2}\Bigg(\sum_{l=1}^{L} \left(\frac{M_l}{\hat{M}_l}\right)^2 \Big(\hat{V}\ (\hat{Y}_l) + \hat{R}^2_l\, \hat{V}(\hat{M}_l\ ) - 2\ \hat{R}_l\ \ \ C(\hat{Y}_l\ , \hat{M}_l\ )\Big) +$$

$$+ \ \sum_{l'\neq l}^{L} \ \frac{M_l\, M_{l'}}{\hat{M}_l\, \hat{M}_{l'}} \ \Big(\hat{C}\ (\hat{Y}_l\ , \hat{Y}_{l'}) - \hat{R}_l\ \ \hat{C}\ (\hat{Y}_l\ , \hat{M}_{l'}\ ) +$$

$$- \ \ \hat{R}_l\ \ \hat{C}\ (\hat{Y}_{l'}\ \ , \hat{M}_l) + \ \hat{R}_l\ \hat{R}_{l'}\ \hat{C}\ (\hat{M}_l\ , \hat{M}_{l'})\Big)\Bigg) \qquad (26)$$

dove:

$$\hat{R}_l = \frac{\hat{Y}_l}{\hat{M}_l} \qquad (27)$$

Esaminiamo ora, nella (26), i termini relativi alle varianze; la varianza $\hat{V}(\hat{Y}_l)$ e' espressa da:

$$\hat{V}(\hat{Y}_l) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \Bigg(\sum_{i'>i}^{n_h} \left(\frac{\pi_{hi}\ \pi_{hi'} - \pi_{hii'}}{\pi_{hii'}}\right) \left(\frac{M_{hi}\ \ {}_r\hat{W}_{hil}\ \ {}_r\hat{\bar{Y}}_{hil}}{\pi_{hi}} + \right.$$

$$\left. - \ \frac{M_{hi'}\ \ {}_r\hat{W}_{hi'l}\ \ {}_r\hat{\bar{Y}}_{hi'l}}{\pi_{hi'}} \right)^2 + \ \frac{M_{hi}\ (M_{hi} - m_{hi})}{\pi_{hi}\ (m_{hi} - 1)\ m_{hi}\ \ {}_r\hat{M}^2_{hi}} \ \cdot$$

$$\cdot \ \sum_{j=1}^{m_{hi}} \left(\left(Y_{hilj} - \frac{{}_r\hat{Y}_{hil}}{{}_r\hat{M}_{hi}}\ \ {}_r\delta_{hij} \right)\ {}_r\delta_{hij} \right)^2 \Bigg) \qquad (28)$$

La varianza $\hat{V}(\hat{M}_l)$ puo' essere ottenuta con un'espressione simile alla (28) sostituendo la variabile M_{hilj} alla variabile Y_{hilj}.

Per i termini relativi alle covarianze le corrispondenti espressioni esplicite sono:

$$\hat{C}(\hat{Y}_l\ ,\ \hat{M}_l) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{m^2_{hi}}{(\pi_{hi}\ {}_r m_{hi})^2} \left(\frac{M_{hi}}{m_{hi}}\right)^2 \left(\sum_{j=1}^{m_{hi}} \hat{Z}_{hilj}\, \hat{Z}'_{hilj} + \right.$$

$$\left. + \sum_{j' \neq j}^{m_{hi}} \hat{Z}_{hilj}\, \hat{Z}'_{hilj'} \right) \tag{29}$$

$$\hat{C}(\hat{Y}_l\ ,\ \hat{Y}_{l'}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{m^2_{hi}}{(\pi_{hi}\ {}_r m_{hi})^2} \left(\frac{M_{hi}}{m_{hi}}\right)^2 \left(\sum_{j=1}^{m_{hi}} \hat{Z}_{hilj}\, \hat{Z}_{hil'j} + \right.$$

$$\left. + \sum_{j' \neq j}^{m_{hi}} \hat{Z}_{hilj}\, \hat{Z}_{hil'j'} \right) \tag{30}$$

$$\hat{C}(\hat{Y}_l\ ,\ \hat{M}_{l'}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{m^2_{hi}}{(\pi_{hi}\ {}_r m_{hi})^2} \left(\frac{M_{hi}}{m_{hi}}\right)^2 \left(\sum_{j=1}^{m_{hi}} \hat{Z}_{hilj}\, \hat{Z}'_{hil'j} + \right.$$

$$\left. + \sum_{j' \neq j}^{m_{hi}} \hat{Z}_{hilj}\, \hat{Z}'_{hil'j'} \right) \tag{31}$$

$$\hat{C}(\hat{Y}_{l'}\ ,\ \hat{M}_l) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{m^2_{hi}}{(\pi_{hi}\ {}_r m_{hi})^2} \left(\frac{M_{hi}}{m_{hi}}\right)^2 \left(\sum_{j=1}^{m_{hi}} \hat{Z}_{hil'j}\, \hat{Z}'_{hilj} + \right.$$

$$\left. + \sum_{j' \neq j}^{m_{hi}} \hat{Z}_{hil'j}\ \hat{Z}'_{hilj'} \right) \tag{32}$$

dove:

$$\hat{Z}_{hilj} = \left(Y_{hilj} - \frac{{}_r\hat{Y}_{hil}}{{}_r\hat{M}_{hi}} \right) {}_r\delta_{hij}$$

$$\hat{Z}'_{hilj} = (M_{hilj} - {}_r\hat{W}_{hil})\ {}_r\delta_{hij}$$

8. Studio empirico

8.1 Premessa

Lo studio empirico degli stimatori proposti nel presente lavoro e' stato condotto sui dati rilevati nel mese di aprile 1989 mediante l'indagine Istat sui consumi delle famiglie (Istat 1992).

L'indagine in oggetto si fonda su di un campione a due stadi analogo a quello descritto nel precedente paragrafo 2.

Le UP sono i comuni che, nell'ambito di ciascuna regione geografica, sono divisi nei seguenti due gruppi:

(1) comuni con popolazione uguale o superiore a 50.000 abitanti;

(2) rimanenti comuni.

I comuni del primo gruppo sono inclusi con certezza nel campione e partecipano all'indagine in tutti i mesi dell'anno.

I comuni del secondo gruppo vengono stratificati secondo le modalita' congiunte della zona altimetrica (montagna, collina e pianura) e dell'attivita' economica prevalente (agricoltura, industria ed altre attivita').

Da ciascuno degli strati elementari cosi' definiti si procede alla selezione di 3 comuni:

- il primo, effettua l'indagine nei mesi di gennaio, aprile, luglio ed ottobre;
- il secondo, nei mesi di febbraio, maggio, agosto e novembre;
- il terzo, nei mesi di marzo, giugno, settembre e dicembre.

Il numero complessivo di comuni-campione per mese e' pari a 274, di cui 144 appartengono al primo gruppo e 130 al secondo gruppo.

Le US sono costituite dalle famiglie, selezionate mediante campionamento sistematico dalle anagrafi dei comuni campione. Le caratteristiche d'interesse sono rilevate su tutti i componenti delle famiglie estratte.

In ciascun mese vengono intervistate 3200 famiglie circa.

8.2 Variabili sotto studio

Lo studio e' stato condotto su 50 variabili suddivise nei seguenti quattro gruppi cosi' costituiti:

A: 7 variabili relative al possesso di beni durevoli;
B: 22 variabili concernenti le spese per beni non alimentari;
C: 20 variabili inerenti le spese per beni alimentari;
D: reddito mensile della famiglia.

8.3 Post-stratificazioni

Ai fini della valutazione delle proprieta' empiriche degli stimatori proposti, sono state considerate le seguenti post-stratificazioni alternative:

P1: due post-strati (il primo contenente le famiglie con un componente ed il secondo le rimanenti famiglie);
P2: due post-strati (il primo contenente le famiglie con uno e due componenti ed il secondo le rimanenti famiglie);
P3: tre post-strati (il primo, contenente le famiglie con un componente; il secondo, le famiglie con due componenti; il terzo, le rimanenti famiglie);
P4: quattro post-strati (il primo contenente le famiglie con un componente, il secondo le famiglie con due componenti, il terzo le famiglie con tre componenti ed il quarto le famiglie rimanenti).

8.4 Criteri di valutazione

8.4.1 Distorsione

Le distorsioni e le varianze di campionamento degli stimatori considerati sono state analizzate mediante alcune statistiche calcolate su ciascun gruppo di variabili e per ogni post-stratificazione.

Ai fini di chiarire i criteri di costruzione delle suddette statistiche, consideriamo una determinata post-stratificazione, per esempio P1, ed indichiamo con $\hat{B}_1({}_d\hat{\bar{Y}}_\alpha)$ la stima della componente B_1 della distorsione dello stimatore diretto ${}_d\hat{\bar{Y}}_\alpha$, della media $\bar{Y}_\alpha$ della generica variabile y_α del gruppo A ($\alpha = 1,...,7$).

Le statistiche in oggetto sono cosi' definite:

(a) valore medio, nelle variabili del gruppo A e sotto la post-stratificazione P1, dei valori assoluti delle stime $\hat{B}_1$:

$$E(\hat{B}_1) = \frac{1}{7}\sum_{\alpha=1}^{7} | \hat{B}_1 ({}_d\hat{\bar{Y}}_\alpha) |$$

(b) valore minimo, nelle variabili del gruppo A e sotto la post-stratificazione P1, delle stime $\hat{B}_1$:

$$MIN(\hat{B}_1) = MIN\ (\hat{B}_1({}_d\hat{\bar{Y}}_\alpha))$$

(c) valore massimo, nelle variabili del gruppo A e sotto la post-stratificazione P1, delle stime $\hat{B}_1$:

$$MAX(\hat{B}_1) = MAX\ (\hat{B}_1 ({}_d\hat{\bar{Y}}_\alpha))$$

(d) scarto quadratico medio, nelle variabili del gruppo A e sotto la post-stratificazione P1, delle stime $\hat{B}_1$:

$$S(\hat{B}_1) = \left(\frac{1}{7} \sum_{\alpha=1}^{7} \left(\hat{B}_1 ({}_d\hat{\bar{Y}}_\alpha) - \frac{1}{7}\sum_{\alpha=1}^{7} \hat{B}_1 ({}_d\hat{\bar{Y}}_\alpha)\right)^2 \right)^{0.5}$$

Per lo stimatore del rapporto post-stratificato sono state calcolate le medesime statistiche, le cui espressioni sono formalmente analoghe rispettivamente alle (a), (b), (c) e (d), salvo la sostituzione di $\hat{B}_1({}_d\hat{\bar{Y}}_\alpha)$ con $\hat{B}^*_1 ({}_{rp}\hat{\bar{Y}}_\alpha)$.

8.4.2 Errore di campionamento

Per la valutazione degli errori campionari, indichiamo con $\hat{V}({}_d\hat{\bar{Y}}_\alpha)$ la stima della varianza campionaria dello stimatore diretto della media $\bar{Y}_\alpha$ della generica variabile y_α del gruppo A.

Indicando con :

$$\hat{\sigma} ({}_d\hat{\bar{Y}}_\alpha) = (\hat{V}({}_d\hat{\bar{Y}}_\alpha))^{0.5}$$

la stima dell'errore di campionamento di ${}_d\hat{\bar{Y}}_\alpha$, le statistiche adottate a tal fine sono:

(a_1) il valore medio, nelle variabili del gruppo A, delle stime degli errori campionari:

- $$E(\hat{\sigma}) = \frac{1}{7}\sum_{\alpha=1}^{7} \hat{\sigma} ({}_d\hat{\bar{Y}}_\alpha)$$

(b_1) il valore minimo, nelle variabili del gruppo A, delle stime degli errori campionari:

- $$MIN(\hat{\sigma}) = MIN \left(\hat{\sigma} ({}_d\hat{\bar{Y}}_\alpha)\right)$$

(c_1) il valore massimo, nelle variabili del gruppo A, delle stime degli errori campionari:

- $$MAX(\hat{\sigma}) = MAX \left(\hat{\sigma} ({}_d\hat{\bar{Y}}_\alpha)\right)$$

(d_1) lo scarto quadratico medio, nelle variabili del gruppo A, delle stime degli errori campionari:

- $$S(\hat{\sigma}) = \left(\frac{1}{7}\sum_{\alpha=1}^{7} \left(\hat{\sigma} ({}_d\hat{\bar{Y}}_\alpha) - \frac{1}{7}\sum_{\alpha=1}^{7} \hat{\sigma} ({}_d\hat{\bar{Y}}_\alpha)\right)^2\right)^{0.5}$$

Per lo stimatore post-statificato e quello del rapporto post-stratificato si e' proceduto al calcolo di analoghe statistiche, le cui espressioni sono ottenibili dalle precedenti mediante semplice adattamento formale.

8.5 Presentazione dei risultati

8.5.1 Tassi di risposta e valori medi delle variabili

Dalla tavola 1 si desume una certa correlazione fra i tassi di risposta e l'ampiezza della famiglia; il valore piu' basso (84.5%) registra per le famiglie con un solo componente e quello piu' elevato

(90%) per le famiglie con un numero di componenti superiore a quattro.

Nella tavola 2, per ciascun gruppo di variabili e distintamente per ampiezza della famiglia, sono indicati i valori medi per famiglia.

Piu' precisamente, con riferimento ad esempio al gruppo A, i suddetti valori sono stati ottenuti attraverso la formula:

$$\hat{Y}_l = \frac{1}{7}\sum_{\alpha=1}^{7} \hat{Y}_{l\alpha} = \frac{1}{7}\sum_{\alpha=1}^{7} \frac{\hat{Y}_{l\alpha}}{\hat{M}_l}$$

in cui le stime $\hat{Y}_{l\alpha}$ e $\hat{M}_l$ sono state calcolate utilizzando le espressioni di $\hat{Y}_l$ e $\hat{M}_l$ riportate nel paragrafo 7.

La tavola 2 mostra che i valori medi aumentano al crescere dell'ampiezza della famiglia.

8.5.2 Distorsione

Nella tavola 3 sono riportati i valori della componente distorsiva B_1 dello stimatore diretto.

Dall'esame della tavola si evince che:

- la scelta del tipo di post-stratificazione ha una sensibile influenza sulla distorsione; infatti, i valori piu' elevati di $E(\hat{B}_1)$ si ottengono con le post-stratificazioni P1 e P2, mentre i valori piu' piccoli con P3 e P4;

- dalla precedente considerazione segue che per ridurre il piu' possibile la distorsione totale dello stimatore diretto, sostituendolo con lo stimatore post-stratificato, sembra che le post-stratificazioni piu' efficaci siano quelle basate su un numero piccolo di post-strati;

- $\hat{B}_1$ assume sia valori positivi che negativi.

Nella tavola 4 sono indicati i valori della componente distorsiva $\hat{B}^*_1$ dello stimatore del rapporto post-stratificato.

Dall'esame della tavola si evince che:

- $\hat{B}^*_1$ assume valori piu' grandi di $\hat{B}_1$;

- l'effetto di ciascuna post-stratificazione varia a seconda del gruppo di variabili considerato. Per il gruppo A, i valori piu'alti di $E(\hat{B}^*_1)$ si riferiscono alle post-stratificazioni P1 e P2; mentre, per il gruppo B , i valori piu' alti si registrano con le post-stratificazioni P2 e P4.

8.5.3 Errore di campionamento

Nelle tavole 5, 6 e 7 sono mostrati i valori delle statistiche usate per l'analisi degli errori campionari, rispettivamente per gli stimatori diretto, post-stratificato e del rapporto post-stratificato.

Dall'esame dei valori ivi contenuti osserviamo che:

- lo stimatore del rapporto post-stratificato e' quello piu' efficiente;

- per i gruppi di variabili A e C, risulta piu' efficiente lo stimatore post-stratificato rispetto al diretto, mentre per i gruppi di variabili B e D e' piu' efficiente lo stimatore diretto rispetto a quello post-stratificato;

- lo stimatore post-stratificato raggiunge l'efficienza massima con la post-stratificazione P1; mentre la post-stratificazione P4 e' quella che da' i risultati migliori con l'applicazione dello stimatore del rapporto post-stratificato.

8.5.4 Considerazioni conclusive

I risultati ottenuti sono parziali e vanno riferiti unicamente allo specifico contesto sperimentale.

Emerge, comunque, il risultato che nelle indagini concrete, in presenza di mancata risposta totale, la scelta del tipo di stimatore con cui ottenere le stime campionarie risulta piuttosto difficile. Ad esempio, nel caso studiato, lo stimatore del rapporto post-stratificato che risulta il piu' efficiente, e' anche quello maggiormente affetto da distorsione. Inoltre, la situazione e' ulteriormente complicata dal fatto che i risultati variano al modificarsi delle post-stratificazioni e dei gruppi di variabili di volta in volta presi in considerazione.

Tav. 1 - *Tassi di risposta percentuali secondo il numero di componenti della famiglia*

Numero di componenti delle famiglie	Tassi di risposta (%)
1	84,50
2	88,18
3	86,96
4 e piu'	90,00

Tav. 2 - *Valori medi per gruppi di variabili, secondo il numero di componenti della famiglia*

Numero di componenti della famiglia	Gruppi di variabili A	B	C	D
1	0,22	8.823	10.512	1.447.750
2	0,29	14.571	18.546	2.332.212
3	0,41	22.804	24.320	3.096.182
4 e piu'	0,42	26.476	30.100	352.000

Tav.3 - *Distorsione dello stimatore diretto*

Statistiche	Post-stratificazioni P1	P2	P3	P4
	Gruppo A			
$E(\hat{B}_1)$	0,0048	0,0043	0,0013	0,0019
$MIN(\hat{B}_1)$	0,0009	0,0014	0,0006	0,0009
$MAX(\hat{B}_1)$	0,0086	0,0098	0,0022	0,0046
$S(\hat{B}_1)$	0,0027	0,0032	0,0009	0,0020
	Gruppo B			
$E(\hat{B}_1)$	460	423	290	260
$MIN(\hat{B}_1)$	-249	-289	- 161	-457
$MAX(\hat{B}_1)$	5.509	4.629	3.302	1.452
$S(\hat{B}_1)$	1.176	992	717	433
	Gruppo C			
$E(\hat{B}_1)$	318	251	66	105
$MIN(\hat{B}_1)$	52	34	-70	- 265
$MAX(\hat{B}_1)$	740	662	367	275
$S(\hat{B}_1)$	205	184	93	115
	Gruppo D			
$E(\hat{B}_1)$	41.093	39.460	11.851	28.874

Tav. 4 - *Distorsione dello stimatore post-stratificato*

Statistiche	Post-stratificazioni P1	P2	P3	P4
	Gruppo A			
$E(\hat{B}^*_1)$	0,0082	0,0110	0,0112	0,0106
$MIN(\hat{B}^*_1)$	-0,0225	-0,0203	-0,0266	-0,0299
$MAX(\hat{B}^*_1)$	0,0011	0,0010	0,0018	0,0020
$S(\hat{B}^*_1)$	0,0081	0,0079	0,0099	0,0096
	Gruppo B			
$E(\hat{B}^*_1)$	522	666	625	675
$MIN(\hat{B}^*_1)$	- 2.268	- 3.919	-3.207	- 4.473
$MAX(\hat{B}^*_1)$	611	434	523	496
$S(\hat{B}^*_1)$	709	1.080	797	1.047
	Gruppo C			
$E(\hat{B}^*_1)$	741	718	919	891
$MIN(\hat{B}^*_1)$	- 3.143	-2.953	- 3.906	- 4.041
$MAX(\hat{B}^*_1)$	66	69	100	77
$S(\hat{B}^*_1)$	699	667	879	885
	Gruppo D			
$E(\hat{B}^*_1)$	77.245	73.223	97.679	79.467

Tav. 5 - *Errore dello stimatore diretto*

Statistiche	Gruppi di variabili			
	A	B	C	D
$E(\hat{\sigma})$	0,012	3.305	974	58232
$MIN(\hat{\sigma})$	0,005	141	125	-
$MAX(\hat{\sigma})$	0,017	24.636	2.572	-
$S(\hat{\sigma})$	0,004	5.776	711	-

Tav. 6 - *Errore dello stimatore post-stratificato*

Statistiche	Post-stratificazioni			
	P1	P2	P3	P4
		Gruppo A		
$E(\hat{\sigma})$	0,014	0,015	0,015	0,016
$MIN(\hat{\sigma})$	0,007	0,005	0,006	0,006
$MAX(\hat{\sigma})$	0,018	0,019	0,002	2,370
$S(\hat{\sigma})$	0,003	0,005	0,005	0,160
		Gruppo B		
$E(\hat{\sigma})$	3.231	3.450	3.222	3.577
$MIN(\hat{\sigma})$	139	148	184	185
$MAX(\hat{\sigma})$	23.002	20.932	20.978	24.438
$S(\hat{\sigma})$	5.441	5.330	4.987	5.994
		Gruppo C		
$E(\hat{\sigma})$	977	1.033	1.076	1.084
$MIN(\hat{\sigma})$	126	129	137	143
$MAX(\hat{\sigma})$	2.577	2.608	3.110	2.998
$S(\hat{\sigma})$	710	726	790	804
		Gruppo D		
$E(\hat{\sigma})$	58.444	70.722	71.973	78.638

Tav. 7 - *Errore dello stimatore del rapporto post-stratificato*

Statistiche	Post-stratificazioni P1	P2	P3	P4
		Gruppo A		
$E(\hat{\sigma})$	0,013	0,012	0,014	0,013
$MIN(\hat{\sigma})$	0,006	0,006	0,006	0,006
$MAX(\hat{\sigma})$	0,017	0,017	0,018	0,017
$S(\hat{\sigma})$	0,003	0,004	0,004	0,004
		Gruppo B		
$E(\hat{\sigma})$	3.148	3.125	3.583	3.095
$MIN(\hat{\sigma})$	133	132	153	132
$MAX(\hat{\sigma})$	23.315	22.327	27.845	22.377
$S(\hat{\sigma})$	5.454	5.247	6.509	5.262
		Gruppo C		
$E(\hat{\sigma})$	923	931	1.058	913
$MIN(\hat{\sigma})$	119	118	137	117
$MAX(\hat{\sigma})$	2.447	2.518	2.818	2.434
$S(\hat{\sigma})$	657	691	758	670
		Gruppo D		
$E(\hat{\sigma})$	53.925	54.133	67.539	52.999

Note

Il lavoro e' frutto della collaborazione dei due Autori. Per quanto riguarda la sua stesura, A. Russo ha redatto i paragrafi 1, 2, 3, 4, 5, 8.5 e l'Appendice A1; P. D. Falorsi i rimanenti paragrafi e le Appendici A2 e A3.

Riferimenti bibliografici

COCHRAN W.G. (1977), *Sampling Techiniques*, New York, John Wiley.

HOLT D., SMITH T.M.F. (1979), *Post-stratification*, in "Journal of the Royal Statistical Society", A, 142, pagg. 33-46.

ISTAT (1992), *I consumi delle famiglie* - anno 1989 - in "Collana di informazione", n. 1, Roma, Istat.

KALSBEEK W.D. (1980), *A conceptual review of survey error due to nonresponse*, in "Proceedings of the Section on Survey Research Methods of the American Statistical Association", pagg. 131-136.

LITTLE R.J.A. (1982), *Models for nonresponse in sample survey*, in "Journal of the American Statistical Association", 77, pagg. 237-250.

LITTLE R.J.A. (1986), *Survey nonresponse adjustements for estimates of means*, in "International Statistical Review", 54, pagg. 139-157.

OH H.L., SCHEUREN F. (1983), *Weighting adjustment for unit nonresponse*, in "Incomplete data in sample surveys", Volume 2, pagg. 143-184, New York, Academic Press.

PLATEK R., GRAY G.B. (1985), *Some aspects of nonresponse adjustment*, in "Survey Methodology", 11, pagg. 1-14.

THOMSEN I. (1973), *A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data*, in "Statistik Tidskrift", 4, pagg. 278-283.

WOODRUFF R.S. (1971), *Simple method for approximating the variance of a complicated estimate*, in "Journal of the American Statistical Association, 66, pagg. 411-414.

Appendici

A1. Valore atteso e varianza dello stimatore diretto

Il valore atteso dello stimatore diretto puo' essere dedotto dalla seguente espressione:

$$E({}_d\hat{Y}) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \frac{M_{hi}}{M\,\Pi_{hi}} E_1(\lambda_{hi})\, E_2(E_2({}_r\hat{Y}_{hi} \mid {}_r m_{hi})) \quad (A1.1)$$

dove: λ_{hi} e' una variabile dicotomica che assume valore 1 se l'UP (hi) e' inclusa nel campione, e valore 0 altrimenti; E_1 indica l'operatore di valore atteso fra tutte le selezioni relative al primo stadio; E_2 denota l'operatore di valore atteso fra tutte le selezioni relative al secondo stadio di campionamento, condizionate ad un insieme fisso di UP selezionate al primo stadio; il simbolo

$${}_r\hat{Y}_{hi}$$

denota una statistica campionaria che assume valore

$${}_r\hat{Y}_{hi} = \frac{1}{{}_r m_{hi}} \sum_{j=1}^{m_{hi}} Y_{hij}\ {}_r\delta_{hij}$$

se ${}_r m_{hi} > 0$, e valore 0 altrimenti;

$E_2({}_r\hat{Y}_{hi} \mid {}_r m_{hi})$ indica il valore atteso di ${}_r\hat{Y}_{hi}$ condizionato al fatto che il numero di US campione rispondenti nell'UP (hi) sia uguale ad ${}_r m_{hi}$.

Si ha che:

(i) se ${}_r m_{hi} > 0$, allora $E_2({}_r\hat{\bar{Y}}_{hi}| {}_r m_{hi}) = {}_r\bar{Y}_{hi}$

(ii) se ${}_r m_{hi} = 0$, allora $E_2({}_r\hat{\bar{Y}}_{hi}| {}_r m_{hi}) = 0$.

Da cio' deriva che:

$$E_2 (E_2 ({}_r\hat{\bar{Y}}_{hi}| {}_r m_{hi})) = {}_r\bar{Y}_{hi} \; Pr({}_r m_{hi} > 0) \qquad (A1.2)$$

dove la probabilita' $Pr({}_r m_{hi} > 0)$ e' espressa da:

$$Pr({}_r m_{hi} > 0) = 1 - \prod_{j=0}^{m_{hi}-1} \left(\frac{{}_{\bar{r}}M_{hi} - j}{M_{hi} - j} \right) \qquad (A1.3)$$

Conseguentemente, usando il risultato $E_1(\lambda_{hi})= \pi_{hi}$, si ottiene:

$$E({}_d\hat{Y}) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} W_{hi} \, {}_r\bar{Y}_{hi} \, P_r({}_r m_{hi} > 0) \qquad (A1.4)$$

Pertanto, sotto l'ipotesi $Pr({}_r m_{hi} > 0) = 1$ e' possibile derivare l'espressione (4).

Al fine di derivare la varianza dello stimatore diretto, facciamo riferimento al teorema 10.2 riportato in Cochran (1977). Abbiamo che:

$$V({}_d\hat{Y}) = V_1 (E_2 ({}_d\hat{Y})) + E_1 (V_2 ({}_d\hat{Y})) \qquad (A1.5)$$

dove:

$$V_1(E_2 ({}_d\hat{Y}))$$

indica la varianza di primo stadio tra i valori medi relativi al secondo stadio:

$$V_2({}_d\hat{Y})$$

denota la varianza tra tutte le possibili selezioni di US per un dato insieme di UP.

Usando il risultato (A1.2), si ottiene:

$$V_1 (E_2 ({}_d\hat{Y})) = V \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{M_{hi}}{M \, \pi_{hi}} \, {}_r\bar{Y}_{hi} \, Pr({}_r m_{hi} >0) \right) \qquad (A1.6)$$

Mediante il teorema 9A.5 riportato in Cochran (1977), si ottiene:

$$V_1 (E_2({}_d\hat{Y}))=$$

$$= \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{i'>i}^{N_h} (\pi_{hi} \, \pi_{hi'} - \pi_{hii'}) \left(\left(\frac{W_{hi}}{\pi_{hi}} \, {}_r\bar{Y}_{hi} \, Pr ({}_r m_{hi} > 0) \right) + \right.$$

$$\left. - \left(\frac{W_{hi'}}{\pi_{hi'}} \; {}_r\bar{Y}_{hi'} \, Pr (m_{hi'} >0) \right) \right)^2 \qquad (A1.7)$$

Il secondo addendo a secondo membro della (A1.5) si puo' scrivere:

$$\sum_{h=1}^{H} \sum_{i=1}^{N_h} E_1 (\lambda_{hi}) \quad \frac{W^2_{hi}}{\pi^2_{hi}} \quad V_2 ({}_r\hat{\bar{Y}}_{hi}) \qquad (A1.8)$$

L'approssimazione lineare di $V_2({}_r\hat{\bar{Y}}_{hi})$ (Woodruff 1971) e' data da:

$$E_2\left(\frac{d\ {}_r\hat{\bar{Y}}_{hi}}{d\ {}_r\hat{Y}_{hi}}\ ({}_r\hat{Y}_{hi} - {}_rY_{hi}) + \frac{d\ {}_r\hat{\bar{Y}}_{hi}}{d\ {}_r\hat{M}_{hi}}\ ({}_r\hat{M}_{hi} - {}_rM_{hi})\right)^2 \qquad \text{(A1.9)}$$

dove le derivate parziali (denotate con il simbolo d):

$$d\ {}_r\hat{\bar{Y}}_{hi}\ /d\ \hat{Y}_{hi} \qquad \text{e} \qquad d\ {}_r\hat{\bar{Y}}_{hi}/d\ {}_r\hat{M}_{hi}$$

sono calcolate in corrispondenza dei valori attesi ${}_rY_{hi}$ e ${}_rM_{hi}$.

Abbiamo che:

$$\frac{d\ {}_r\hat{\bar{Y}}_{hi}}{d\ {}_r\hat{Y}_{hi}} = \frac{1}{{}_rM_{hi}}\ ; \qquad \frac{d\ {}_r\hat{Y}_{hi}}{d\ {}_r\hat{M}_{hi}} = \frac{{}_r\bar{Y}_{hi}}{{}_rM_{hi}}$$

Usando il precedente risultato, segue che:

$$V_2({}_r\bar{Y}_{hi}) \doteq E_2\left(\frac{M_{hi}}{m_{hi}}\sum_{j=1}^{m_{hi}} (Y_{hij} - {}_r\bar{Y}_{hi})\ \frac{{}_r\delta_{hij}}{{}_rM_{hi}}\right)^2 =$$

$$\doteq \frac{M_{hi}\ (M_{hi} - m_{hi})}{(M_{hi} - 1)\ \ {}_rM^2_{hi}\ m_{hi}} \sum_{j=1}^{M_{hi}} ((Y_{hij} - {}_r\bar{Y}_{hi})\ \ {}_r\delta_{hij})^2 \qquad \text{(A1.10)}$$

Sostituendo le espressioni (A1.6), (A1.7), (A1.8) ed (A1.10) nella (A1.5), e introducendo l'approssimazione $Pr({}_rm_{hi} > 0)$, si perviene alla relazione (10).

A2. Valore atteso e varianza dello stimatore post-stratificato

Dalla (A1.1) si desume:

$$E({}_p\hat{Y}) = \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{l=1}^{L} \frac{M_{hil}}{M}\ E_2\ (E_2\ (\ {}_r\hat{\bar{Y}}_{hil}|{}_rm_{hil})) \qquad \text{(A2.1)}$$

dove: ${}_r\hat{\bar{Y}}_{hil}$ e' una statistica che e' uguale a:

$${}_r\hat{\bar{Y}}_{hil} = \frac{1}{{}_rm_{hil}} \sum_{j=1}^{m_{hi}} Y_{hilj}\ \ {}_r\delta_{hij}$$

se ${}_rm_{hil} > 0$, ed a 0 altrimenti.

Abbiamo che:

$$E_2\ (E_2\ (\ {}_r\hat{\bar{Y}}_{hil}|\ \ {}_rm_{hil})) = {}_r\bar{Y}_{hil}\ Pr\ (\ {}_rm_{hil} > 0) \qquad \text{(A2.2)}$$

dove:

$$Pr(\ {}_rm_{hil} > 0) = 1 - \prod_{j=0}^{m_{hi}} \left(\frac{\bar{{}_r}M_{hil} - j}{M_{hi} - j}\right)$$

Pertanto, il valor medio dello stimatore post-stratificato e' dato da:

$$E({}_p\hat{Y}) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{l=1}^{L} W_{hi} \quad W_{hil} \quad {}_r\overline{Y}_{hil} \, Pr(\, {}_r m_{hil} > 0) \qquad (A2.3)$$

Sulla base di tale relazione, sotto l'ipotesi $P_r(\, {}_r m_{hil} > 0) = 1$, si perviene facilmente all'espressione (13).

Per quanto riguarda la varianza di ${}_p\hat{Y}$, dalla (A1.5) si ottiene:

$$V({}_p\hat{Y}) = V_1 \, (E_2({}_p\hat{Y})) + E_1 \, (V_2({}_p\hat{Y})) \qquad (A2.4)$$

Mediante la (A2.2) si ha:

$$V_1 \, (E_2({}_p\hat{Y})) = V \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{Z_{hi}}{\Pi_{hi}} \right) \qquad (A2.5)$$

dove:

$$Z_{hi} = \frac{1}{M} \sum_{l=1}^{L} M_{hil} \quad {}_r\overline{Y}_{hil} \, Pr({}_r m_{hil} > 0)$$

Conseguentemente si ottiene:

$$V_1(E_2({}_p\hat{Y})) =$$

$$= \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{i'>i}^{N_h} (\Pi_{hi} \, \Pi_{hi'} - \Pi_{hii'}) \left(\frac{Z_{hi}}{\Pi_{hi}} - \frac{Z_{hi'}}{\Pi_{hi'}} \right)^2 \qquad (A2.6)$$

Il secondo addendo della (A2.4) si puo' scrivere:

$$\sum_{h=1}^{H} \sum_{i=1}^{N_h} \frac{1}{\Pi_{hi}} V_2 \, (\hat{Z}_{hi}) \qquad (A2.7)$$

essendo:

$$\hat{Z}_{hi} = \frac{1}{M} \sum_{l=1}^{L} M_{hil} \quad {}_r\hat{\overline{Y}}_{hil}$$

Consideriamo poi l'approssimazione lineare della varianza $V_2(\hat{Z}_{hi})$ definita da:

$$V_2(\hat{Z}_{hi}) \doteq E_2 \left(\frac{1}{M} \sum_{l=1}^{L} M_{hil} \left(\frac{d \; {}_r\hat{\overline{Y}}_{hil}}{d \; {}_r\hat{Y}_{hil}} (\, {}_r\hat{Y}_{hil} - \, {}_r Y_{hil}) + \right. \right.$$

$$\left. \left. + \quad \frac{d \; {}_r\hat{\overline{Y}}_{hil}}{d \; {}_r\hat{M}_{hil}} (\, {}_r\hat{M}_{hil} - \, {}_r M_{hil}) \right) \right)^2$$

Abbiamo che:

$$\frac{d \; {}_r\hat{\overline{Y}}_{hil}}{d \; {}_r\hat{Y}_{hil}} = \frac{1}{{}_r M_{hil}} \quad ; \quad \frac{d \; {}_r\hat{\overline{Y}}_{hil}}{d \; {}_r\hat{M}_{hil}} = \frac{{}_r\overline{Y}_{hil}}{{}_r M_{hil}}$$

Pertanto segue che:

$$V_2(\hat{Z}_{hi}) \doteq \frac{M_{hi}(M_{hi} - m_{hi})}{(M_{hi} - 1)\, m_{hi}} \sum_{j=1}^{M_{hi}} Z^2_{hij} \qquad (A2.8)$$

in cui:

$$Z_{hij} = \frac{1}{M} \sum_{l=1}^{L} \left((Y_{hij} - {}_r\overline{Y}_{hil})\, {}_l\delta_{hij} \right) \; {}_r\delta_{hij} \frac{M_{hil}}{{}_rM_{hil}} \qquad (A2.9)$$

Sotto l'ipotesi $Pr({}_rm_{hil}>0) = 1$ e sostituendo le espressioni (A2.6), (A2.7), (A2.8) e (A2.9) nella (A2.4), e' possibile derivare la relazione (15).

A3. Valore atteso e varianza dello stimatore del rapporto post-stratificato

Al fine di ottenere il valore atteso dello stimatore del rapporto post-stratificato, introduciamo l'approssimazione lineare:

$$E({}_{rp}\hat{\hat{Y}}) \doteq \sum_{l=1}^{L} \frac{E(\hat{Y}_l)}{E(\hat{M}_l)} M_l \qquad (A3.1)$$

Si ha inoltre:

$$E(\hat{Y}_l) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} M_{hi}\, E_2\left(E_2\left({}_{rp}\hat{\hat{Y}}_{hil} \mid {}_rm_{hi}\right)\right) \qquad (A3.2)$$

dove: ${}_{rp}\hat{\hat{Y}}_{hil}$ e' una statistica che assume valore:

$${}_{rp}\hat{\hat{Y}}_{hil} = \frac{{}_r\hat{Y}_{hil}}{{}_r\hat{M}_{hi}} = \frac{\sum_{j=1}^{m_{hi}} Y_{hilj}\; {}_r\delta_{hij}}{{}_rm_{hi}} \qquad (A3.3)$$

se ${}_rm_{hi} > 0$, e valore 0 altrimenti.

Conseguentemente si ottiene:

$$E_2\left(E_2\left({}_{rp}\hat{\bar{Y}}_{hil} \mid {}_r m_{hi}\right)\right) = \frac{{}_r Y_{hil}}{{}_r M_{hi}} \Pr({}_r m_{hi} > 0) =$$

$$= \quad {}_r W_{hil}\ {}_r \bar{Y}_{hil} \Pr({}_r m_{hi} > 0)$$

Pertanto segue che:

$$E(\hat{Y}_l) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} M_{hi}\ {}_r W_{hil}\ {}_r \bar{Y}_{hil} \Pr({}_r m_{hi} > 0)$$

Con analogo procedimento si ricava che:

$$E(\hat{M}_l) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} M_{hi}\ {}_r W_{hil}\ \Pr({}_r m_{hi} > 0)$$

Il valore atteso dello stimatore del rapporto post-stratificato assume quindi la forma:

$$E({}_{rp}\hat{\bar{Y}}) =$$

$$= \sum_{h=1}^{H} \sum_{i=1}^{N_h} W_{hi} \sum_{l=1}^{L} {}_r W_{hil}\ {}_r \bar{Y}_{hil} \Pr({}_r m_{hi} > 0) \frac{M_l}{E(\hat{M}_l)} \qquad (A3.4)$$

Sfruttando tali risultati e tenendo inoltre presente l'ipotesi $\Pr({}_r m_{hi}>0)$, e' possibile ottenere l'espressione (19).

Sotto l'usuale approssimazione lineare, la varianza dello stimatore del rapporto post-stratificato e' definita dall'espressione:

$$V({}_{rp}\hat{\bar{Y}}) =$$

$$= \frac{1}{M^2}\left(\sum_{l=1}^{L} \frac{M_l}{E^2(\hat{M}_l)}\left(V(\hat{Y}_l) + R_l^2\, V(\hat{M}_l) - 2\ R_l\, C(\hat{Y}_l, \hat{M}_l)\right) +\right.$$

$$+ \sum_{l' \neq l}^{L} \frac{M_l\, M_{l'}}{E(\hat{M}_l)\ E(\hat{M}_{l'})}\left(C(\hat{Y}_l, \hat{Y}_{l'}) - R_{l'}\, C(\hat{Y}_l, \hat{M}_{l'}) +\right.$$

$$\left.\left. - R_l\ C(\hat{Y}_{l'}, \hat{M}_l) + R_l\ R_{l'}\ C(\hat{M}_l, \hat{M}_{l'})\right)\right) \qquad (A3.5)$$

dove:

$$R_l = \frac{E(\hat{Y}_l)}{E(\hat{M}_l)}$$

Esaminiamo ora i termini coinvolti nella formula (A3.5). Abbiamo che:

$$V(\hat{Y}_l) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \left(\sum_{j>i}^{N_h} (\Pi_{hi}\, \Pi_{hi'} - \Pi_{hii'}) \left(\frac{M_{hi}\ {}_rW_{hi'l}\ {}_r\overline{Y}_{hil}}{\Pi_{hi}} \right. \right. .$$

$$. \ Pr({}_rm_{hi'} > 0) - \frac{M_{hi'}\ {}_rW_{hi'l}\ {}_r\overline{Y}_{hi'l}}{\Pi_{hi'}} Pr({}_rm_{hi'} > 0) \Bigg)^2 +$$

$$\frac{M_{hi}\,(M_{hi} - m_{hi})}{(M_{hi} - 1)\, m_{hi}\ {}_rM^2_{hi}} \sum_{j=1}^{M_{hi}} \left(\left(Y_{hilj} - \frac{{}_rY_{hil}}{{}_rM_{hi}} \right) {}_r\delta_{hij} \right)^2 \Bigg) \qquad (A3.6)$$

La varianza $V(\hat{M}_l)$ puo' essere ottenuta con un'espressione simile alla (A3.6), sostituendo Y_{hilj} con la variabile M_{hilj} .

Al fine di derivare le espressioni esplicite relative alle covarianze della relazione (A3.5), introduciamo le seguenti approssimazioni lineari:

$$\hat{Y}_l \doteq \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{1}{\Pi_{hi}} \frac{M_{hi}}{{}_rM_{hi}} \left({}_r\hat{Y}_{hil} - \frac{{}_rY_{hil}}{{}_rM_{hi}}\, {}_r\hat{M}_{hi} \right) \qquad (A3.7)$$

$$\hat{M}_l \doteq \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{1}{\Pi_{hi}} \frac{M_{hi}}{{}_rM_{hi}} \left({}_r\hat{M}_{hil} - {}_rW_{hil}\ {}_r\hat{M}_{hi} \right) \qquad (A3.8)$$

Conseguentemente, poiche' $E(\hat{Y}_l) = E(\hat{M}_l) = 0$, si ottiene:

$$C(\hat{Y}_l\, , \hat{M}_l) = E((\hat{Y}_l\ \hat{M}_l) =$$

$$= \sum_{h=1}^{H} \sum_{i=1}^{N_h} \left(\frac{1}{\Pi_{hi}} \frac{M_{hi}}{{}_rM_{hi}} \left(\sum_{j=1}^{M_{hi}} Z_{hilj}\, Z'_{hilj}\ \frac{M_{hi}}{m_{hi}} + \right. \right.$$

$$+ \sum_{j \neq j'}^{M_{hi}} Z_{hilj}\, Z'_{hilj'} \frac{M_{hi}\,(m_{hi} - 1)}{m_{hi}\,(M_{hi} - 1)} \Bigg) \Bigg) \qquad (A3.9)$$

$$C(\hat{Y}_l\, , \hat{Y}_{l'}) \doteq E(\hat{Y}_l\ \hat{Y}_{l'}) =$$

$$= \sum_{h=1}^{H} \sum_{i=1}^{N_h} \left(\frac{1}{\Pi_{hi}} \frac{M_{hi}}{{}_rM_{hi}} \left(\sum_{j=1}^{M_{hi}} Z_{hilj}\, Z_{hil'j}\quad \frac{M_{hi}}{m_{hi}} + \right. \right.$$

$$+ \sum_{j' \neq j}^{M_{hi}} Z_{hilj}\, Z_{hil'j'} \frac{M_{hi}\,(m_{hi} - 1)}{m_{hi}\,(M_{hi} - 1)} \Bigg) \Bigg) \qquad (A3.10)$$

$$C(\hat{Y}_l\, , \hat{M}_{l'}) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \left(\frac{1}{\Pi_{hi}} \frac{M_{hi}}{{}_rM_{hi}} \left(\sum_{j=1}^{M_{hi}} Z_{hilj}\, Z'_{hil'j}\ \frac{M_{hi}}{m_{hi}} + \right. \right.$$

$$\sum_{j'\neq j}^{M_{hi}} Z_{hilj}\, Z'_{hil'j'} \frac{M_{hi}\,(m_{hi} - 1)}{m_{hi}\,(M_{hi} - 1)}\Bigg)\Bigg) \qquad (A3.11)$$

$$C(\hat{M}_l\,,\, \hat{M}_{l'}) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \Bigg(\frac{1}{\pi_{hi}} \frac{M_{hi}}{{}_rM_{hi}} \Bigg(\sum_{j=1}^{M_{hi}} Z'_{hilj}\, Z'_{hil'j} \frac{M_{hi}}{m_{hi}} +$$

$$+ \sum_{j'\neq j}^{M_{hi}} Z'_{hilj}\, Z'_{hil'j'} \frac{M_{hi}\,(m_{hi} - 1)}{m_{hi}\,(M_{hi} - 1)}\Bigg)\Bigg) \qquad (A3.12)$$

in cui:

$$Z_{hilj} = \left(Y_{hilj} - \frac{{}_rY_{hil}}{{}_rM_{hi}}\right) {}_r\delta_{hij}$$

$$Z'_{hilj} = (M_{hilj} - {}_rW_{hil})\, {}_r\delta_{hij} \qquad (A3.13)$$