Statistica & Applicazioni
Vol. II, n. 2, 2004
pp. 59-75

Alcune considerazioni sull'efficienza degli stimatori rapporto-cum-prodotto

Pier Francesco Perri[§]

Summary: *In order to estimate the unknown mean of a study variable using two auxiliary variables, the efficiency of a modified version of the ratio-cum-product estimators is analysed when the sampling fraction is less then 0.5. Approximate expressions for the bias and the mean square error are obtained and specific conditions under which the modified estimators are more efficient then the original ones are analytically defined. Also a comparison with the multivariate regression estimator is included.*

Keywords: *auxiliary variables, efficiency, sampling fraction, ratio-cum-product estimators, regression estimator.*

1. Introduzione

Nel campionamento da popolazioni finite le informazioni ausiliarie ricoprono un ruolo di primaria importanza in quanto consentono, se opportunamente impiegate, di definire strategie campionarie altamente efficienti. Tali informazioni possono essere utilizzate nella fase di progettazione dell'indagine campionaria, direttamente nella fase di costruzione degli stimatori (*fase di stima*), oppure in entrambe. Limitando l'attenzione alla seconda possibilità, i metodi del *rapporto*, del *prodotto* e della *regressione* (Cochran, 1977; Murthy, 1977) rappresentano i procedimenti comunemente impiegati nelle applicazioni.

In questo lavoro, supponendo di voler stimare la media incognita di una variabile oggetto di interesse, si analizzerà l'utilizzo delle informazioni

[§] Dipartimento di Economia e Statistica – Università degli Studi della Calabria – Via P. Bucci, 87036 Arcavacata di Rende (CS) (e-mail: pierfrancesco.perri@unical.it).

ausiliarie nella fase di stima tramite l'impiego congiunto degli stimatori rapporto e prodotto. Tale tecnica è stata inizialmente proposta da Singh (1965) nel caso in cui siano disponibili informazioni supplementari su due variabili ausiliarie correlate a quella di indagine.

Il contributo che viene presentato consiste, principalmente, nell'analizzare l'efficienza di una versione modificata degli *stimatori rapporto-cum-prodotto* introdotti da Singh (1965, 1967). L'idea di una ridefinizione di questi stimatori è motivata dal tentativo di utilizzare, per piani di campionamento che prevedono una *frazione di campionamento* inferiore a 0.5, il contenuto informativo presente nelle unità della popolazione non inserite nel campione.

Il lavoro è organizzato come segue. Nel paragrafo 2 introdurremo gli stimatori rapporto-cum-prodotto originari, mentre nel paragrafo 3 presenteremo gli stimatori modificati ricavandone, per un generico piano di campionamento, l'espressione approssimata della distorsione e dell'errore quadratico medio. Nel paragrafo 4 confronteremo gli stimatori modificati con quelli originari e con lo stimatore per regressione multivariato. Alcune considerazioni finali concluderanno il lavoro.

2. Gli stimatori rapporto-cum-prodotto

Sia $P = \{1,2,...,N\}$ una popolazione finita di N unità e siano Y, X e Z, rispettivamente, la variabile di indagine e due variabili ausiliarie che assumono valori Y_i, X_i e Z_i sulla *i-esima* unità della popolazione. Supponiamo che le variabili ausiliarie siano correlate a Y e che le loro medie $\bar{X} = N^{-1}\sum_{i=1}^{N} X_i$ e $\bar{Z} = N^{-1}\sum_{i=1}^{N} Z_i$ siano note. Al fine di stimare la media incognita di Y, $\bar{Y} = N^{-1}\sum_{i=1}^{N} Y_i$, Singh (1965, 1967), combinando il metodo di stima del rapporto e del prodotto, ha proposto, per un generico piano di campionamento, gli *stimatori rapporto-cum-prodotto*:

$$\hat{\bar{Y}}_{R1} = \left(\frac{\hat{\bar{Y}}}{\hat{\bar{X}}} \bar{X} \right) \frac{\hat{\bar{Z}}}{\bar{Z}} \quad e \quad \hat{\bar{Y}}_{R2} = \left(\frac{\hat{\bar{Y}}}{\hat{\bar{X}}} \bar{X} \right) \frac{\bar{Z}}{\hat{\bar{Z}}} \tag{1}$$

$$\hat{\bar{Y}}_{P1} = \left(\frac{\hat{\bar{Y}}}{\bar{X}} \hat{\bar{X}} \right) \frac{\hat{\bar{Z}}}{\bar{Z}} \quad e \quad \hat{\bar{Y}}_{P2} = \left(\frac{\hat{\bar{Y}}}{\bar{X}} \hat{\bar{X}} \right) \frac{\bar{Z}}{\hat{\bar{Z}}} \tag{2}$$

in cui $\hat{\bar{Y}}$, $\hat{\bar{X}}$ e $\hat{\bar{Z}}$ sono stimatori corretti, rispettivamente, di $\bar{Y}$, $\bar{X}$ e $\bar{Z}$. Questi stimatori, analogamente agli stimatori *rapporto*, $\hat{\bar{Y}}_R = \left(\hat{\bar{Y}} / \hat{\bar{X}}\right)\bar{X}$, e *prodotto*, $\hat{\bar{Y}}_P = \left(\hat{\bar{Y}}\hat{\bar{X}}\right)/\bar{X}$, sono distorti anche se, per grandi campioni, l'entità della distorsione può essere ritenuta trascurabile.

Impiegando opportunamente lo sviluppo in serie di Taylor è possibile determinare le espressioni approssimate fino all'ordine desiderato della distorsione, *B*, e dell'errore quadratico medio, *MSE*, degli stimatori (1) e (2). Se la numerosità campionaria è sufficientemente elevata, al primo grado di approssimazione, si ha (Singh, 1965, 1967):

$$B\left(\hat{\bar{Y}}_{R1}\right) \doteq B\left(\hat{\bar{Y}}_{R}\right) + \bar{Y}\left(C_{011} - C_{101}\right)$$

$$MSE\left(\hat{\bar{Y}}_{R1}\right) \doteq MSE\left(\hat{\bar{Y}}_{R}\right) + \bar{Y}^2\left(C_{002} + 2C_{011} - 2C_{101}\right); \qquad (3)$$

$$B\left(\hat{\bar{Y}}_{R2}\right) \doteq B\left(\hat{\bar{Y}}_{R}\right) + \bar{Y}\left(C_{002} - C_{011} + C_{101}\right)$$

$$MSE\left(\hat{\bar{Y}}_{R2}\right) \doteq MSE\left(\hat{\bar{Y}}_{R}\right) + \bar{Y}^2\left(C_{002} - 2C_{011} + 2C_{101}\right); \qquad (4)$$

$$B\left(\hat{\bar{Y}}_{P1}\right) \doteq B\left(\hat{\bar{Y}}_{P}\right) + \bar{Y}\left(C_{011} + C_{101}\right)$$

$$MSE\left(\hat{\bar{Y}}_{P1}\right) \doteq MSE\left(\hat{\bar{Y}}_{P}\right) + \bar{Y}^2\left(C_{002} + 2C_{011} + 2C_{101}\right); \qquad (5)$$

$$B\left(\hat{\bar{Y}}_{P2}\right) \doteq B\left(\hat{\bar{Y}}_{P}\right) + \bar{Y}\left(C_{002} - C_{011} - C_{101}\right)$$

$$MSE\left(\hat{\bar{Y}}_{P2}\right) \doteq MSE\left(\hat{\bar{Y}}_{P}\right) + \bar{Y}^2\left(C_{002} - 2C_{011} - 2C_{101}\right), \qquad (6)$$

dove

$$C_{rst} = \frac{E\left[\left(\hat{\bar{X}} - \bar{X}\right)^r \left(\hat{\bar{Y}} - \bar{Y}\right)^s \left(\hat{\bar{Z}} - \bar{Z}\right)^t\right]}{\bar{X}^r \bar{Y}^s \bar{Z}^t} \qquad (7)$$

mentre $B\left(\hat{\bar{Y}}_R\right)$, $B\left(\hat{\bar{Y}}_P\right)$, $MSE\left(\hat{\bar{Y}}_R\right)$ e $MSE\left(\hat{\bar{Y}}_P\right)$ rappresentano le espressioni della distorsione e dell'errore quadratico medio degli stimatori rapporto e prodotto che impiegano solo la variabile ausiliaria X (si veda, ad esempio, Murthy, 1977).

La possibilità di utilizzare due variabili ausiliarie tramite l'impiego congiunto dei metodi del rapporto e del prodotto conduce, come è facile constatare, a stimatori caratterizzati da una maggiore complessità teorica e computazionale rispetto agli stimatori rapporto e prodotto che utilizzano una sola variabile ausiliaria. Tuttavia, questa maggiore complessità non determina necessariamente stime più precise. Infatti, confrontando l'errore quadratico medio degli stimatori (1) e (2) con quello di $\hat{\bar{Y}}_R$ e $\hat{\bar{Y}}_P$ si prova agevolmente (Singh, 1965) che:

(1) $\hat{\bar{Y}}_{R1}$ è più efficiente di $\hat{\bar{Y}}_R$ se $2\dfrac{C_{011}-C_{101}}{C_{002}} < -1$;

(2) $\hat{\bar{Y}}_{R2}$ è più efficiente di $\hat{\bar{Y}}_R$ se $2\dfrac{C_{011}-C_{101}}{C_{002}} > 1$;

(3) $\hat{\bar{Y}}_{P1}$ è più efficiente di $\hat{\bar{Y}}_P$ se $2\dfrac{C_{011}+C_{101}}{C_{002}} < -1$;

(4) $\hat{\bar{Y}}_{P2}$ è più efficiente di $\hat{\bar{Y}}_P$ se $2\dfrac{C_{011}+C_{101}}{C_{002}} > 1$.

2. Gli stimatori rapporto-cum-prodotto modificati

Gli stimatori ottenuti tramite il metodo del rapporto e del prodotto richiedono, almeno nella loro formulazione più semplice, la conoscenza della media (o del totale) dalla variabile ausiliaria e l'impiego di stimatori corretti per la media sia della variabile di indagine che di quella ausiliaria. Tali stimatori impiegano, pertanto, una duplice fonte di informazione: quella derivante dalla popolazione, tramite la conoscenza della media della variabile ausiliaria, e quella campionaria derivante dagli stimatori corretti delle medie. Sembra abbastanza intuitivo, allora, attendersi stime sempre più precise quando maggiore è il contenuto informativo sulla variabile di indagine impiegato nella struttura degli stimatori rapporto e prodotto. Da qui il tentativo di introdurre negli stimatori un'altra possibile fonte di informazione: quella derivante dalla variabile ausiliaria e relativa alle unità della popolazione non inserite nel campione. Pertanto, detta W una generica variabile ausiliaria, $\overline{W}$ la sua media nota e $\hat{\overline{W}}$ uno stimatore corretto di $\overline{W}$, supponiamo di selezionare, secondo un prefissato piano di campionamento, un campione di ampiezza *n*. Sia, dunque, $\hat{\overline{W}}^* = \left(N\overline{W} - n\hat{\overline{W}}\right)/(N-n)$ una trasformazione lineare di $\hat{\overline{W}}$ in cui *N* è la numerosità della popolazione oggetto di studio. Tale trasformazione si configura come uno stimatore corretto di $\overline{W}$ basato sulle unità della popolazione non inserite nel campione. Se la frazione di campionamento, $f = n/N$, è inferiore a 0.5 le unità non campionate risultano essere più numerose di quelle inserite nel campione e, di conseguenza, lo stimatore $\hat{\overline{W}}^*$ risulta più preciso di $\hat{\overline{W}}$ [si veda la (8) e la (10)]. Pertanto, nel costruire gli stimatori rapporto e prodotto si potrebbe pensare di impiegare nella loro struttura stimatori corretti tipo $\hat{\overline{W}}^*$ piuttosto che $\hat{\overline{W}}$. Una proposta in merito è stata avanzata da Bandyopadhyay (1980) e Srivenkataramana (1980) i quali, dopo aver modificato la struttura degli stimatori rapporto e prodotto nel senso sopra specificato, hanno analizzato l'efficienza degli stimatori modificati definendo le condizioni sotto le quali questi ultimi risultano più precisi di quelli originari.

Alla luce di questi lavori, nel presente si vuole estendere la medesima modifica agli stimatori rapporto-cum-prodotto con l'intento di stabilire se il loro impiego possa, in qualche modo, aumentare la precisione delle stime. Pertanto, a partire dagli stimatori originari (1) e (2) definiremo quelli modificati e, dopo averne ricavato l'errore quadratico medio, effettueremo alcuni confronti di efficienza con gli stimatori originari e con lo stimatore per regressione multivariato.

Siano, dunque, *X* e *Z* due variabili ausiliarie correlate a quella di indagine, *Y*, per le quali siano note le rispettive medie, $\overline{X}$ e $\overline{Z}$, sull'intera popolazione. Supponiamo che le variabili *Y*, *X* e *Z* assumano valori positivi e che la frazione di campionamento, come usualmente accade nelle indagini reali, sia inferiore a 0.5. A partire dagli stimatori corretti $\hat{\overline{X}}$ e $\hat{\overline{Z}}$, rispettivamente di $\overline{X}$ e $\overline{Z}$, definiamo i due stimatori corretti:

$$\hat{\overline{X}}^* = \frac{N\overline{X} - n\hat{\overline{X}}}{N-n} \quad e \quad \hat{\overline{Z}}^* = \frac{N\overline{Z} - n\hat{\overline{Z}}}{N-n}$$

per i quali, come è facile verificare, valgono le seguenti relazioni:

$$
\begin{aligned}
&Var\left(\hat{\overline{W}}^*\right) = m^2 Var\left(\hat{\overline{W}}\right) \\
&Cov\left(\hat{\overline{W}}^*, \hat{\overline{Y}}\right) = -mCov\left(\hat{\overline{W}}, \hat{\overline{Y}}\right), \quad \hat{\overline{W}}^* = \hat{\overline{X}}^*, \hat{\overline{Z}}^*
\end{aligned} \tag{8}
$$

$$
Cov\left(\hat{\overline{X}}^*, \hat{\overline{Z}}^*\right) = m^2 Cov\left(\hat{\overline{X}}, \hat{\overline{Z}}\right) \tag{9}
$$

in cui

$$
m = \frac{n}{N-n} = \frac{f}{1-f}. \tag{10}
$$

Se in (1) e (2) sostituiamo $\hat{\overline{X}}$ e $\hat{\overline{Z}}$, rispettivamente, con $\hat{\overline{X}}^*$ e $\hat{\overline{Z}}^*$ otteniamo i seguenti stimatori:

$$
\hat{\overline{Y}}^*_{R1} = \left(\frac{\hat{\overline{Y}}}{\hat{\overline{X}}^*}\overline{X}\right)\frac{\hat{\overline{Z}}^*}{\overline{Z}} \quad e \quad \hat{\overline{Y}}^*_{R2} = \left(\frac{\hat{\overline{Y}}}{\hat{\overline{X}}^*}\overline{X}\right)\frac{\overline{Z}}{\hat{\overline{Z}}^*} \tag{11}
$$

$$
\hat{\overline{Y}}^*_{P1} = \left(\frac{\hat{\overline{Y}}}{\overline{X}}\hat{\overline{X}}^*\right)\frac{\hat{\overline{Z}}^*}{\overline{Z}} \quad e \quad \hat{\overline{Y}}^*_{P2} = \left(\frac{\hat{\overline{Y}}}{\overline{X}}\hat{\overline{X}}^*\right)\frac{\overline{Z}}{\hat{\overline{Z}}^*} \tag{12}
$$

che indicheremo con il termine di stimatori *rapporto-cum-prodotto modificati*.

In virtù della *ratio* sottostante la costruzione degli stimatori rapporto e prodotto (Murthy, 1977), nonché degli stimatori rapporto-cum-prodotto originari (Singh, 1965, 1967), possiamo osservare che, in base alle relazioni definite dalla (8), lo stimatore $\hat{\overline{Y}}^*_{R1}$ può essere considerato come il complementare di $\hat{\overline{Y}}_{P2}$ sulle unità non campionate. Lo stesso dicasi per gli stimatori $\hat{\overline{Y}}^*_{R2}$ e $\hat{\overline{Y}}_{P1}$, $\hat{\overline{Y}}^*_{P1}$ e $\hat{\overline{Y}}_{R2}$, $\hat{\overline{Y}}^*_{P2}$ e $\hat{\overline{Y}}_{R1}$.

Al fine di valutare l'effetto sulla precisione delle stime derivante dalla modifica apportata, determiniamo l'errore quadratico medio degli stimatori proposti.

Siano:

$$
\hat{\overline{Y}}^*_1 = \frac{\hat{\overline{Y}}}{\overline{X}}\hat{\overline{X}}^* \quad e \quad \hat{\overline{Y}}^*_2 = \frac{\hat{\overline{Y}}}{\hat{\overline{X}}^*}\overline{X} \tag{13}
$$

gli stimatori *tipo-prodotto* e *tipo-rapporto* analizzati da Bandyopadhyay (1980) e Srivenkataramana (1980) per i quali, al primo ordine di approssimazione, si ha:

$$MSE\left(\hat{\bar{Y}}_1^*\right) \doteq \bar{Y}^2\left(m^2 C_{200} - 2mC_{110} + C_{020}\right) \tag{14}$$

$$MSE\left(\hat{\bar{Y}}_2^*\right) \doteq \bar{Y}^2\left(m^2 C_{200} + 2mC_{110} + C_{020}\right). \tag{15}$$

Consideriamo lo stimatore $\hat{\bar{Y}}_{R1}^*$ ed esprimiamolo nella forma:

$$\hat{\bar{Y}}_{R1}^* = \bar{Y}\left(1+\delta_y\right)\left(1+\delta_{z^*}\right)\left(1+\delta_{x^*}\right)^{-1} \tag{16}$$

dove $\delta_u = \left(\hat{\bar{U}} - \bar{U}\right)/\bar{U}$, $u = x^*, y, z^*$. Assumendo $|\delta_{x^*}| < 1$, la funzione $\left(1+\delta_{x^*}\right)^{-1}$ può essere espressa come una serie convergente di infiniti termini, $\left(1+\delta_{x^*}\right)^{-1} = \sum_{k=1}^{\infty}(-1)^k \delta_{x^*}^k$. Allora, sostituendo nella (16) lo sviluppo in serie di Taylor di $\left(1+\delta_{x^*}\right)^{-1}$ arrestato al primo ordine, e ammettendo che la dimensione campionaria sia sufficiente a garantire che $\hat{\bar{U}}$ approssimi "abbastanza bene" $\bar{U}$, è possibile trascurare nelle successive operazioni i termini di grado superiore al secondo. Pertanto, dopo alcuni passaggi si perviene alle seguenti espressioni, approssimate al primo ordine, della distorsione e dell'errore quadratico medio:

$$\begin{aligned} B\left(\hat{\bar{Y}}_{R1}^*\right) &\doteq \bar{Y}E\left(\delta_y\delta_{z^*} - \delta_y\delta_{x^*} - \delta_{x^*}\delta_{z^*} + \delta_{x^*}^2\right) \\ &= m\bar{Y}\left[m\left(C_{200} - C_{101}\right) + C_{110} - C_{011}\right]; \end{aligned}$$

$$\begin{aligned} MSE\left(\hat{\bar{Y}}_{R1}^*\right) &\doteq \bar{Y}^2 E\left(\delta_y - \delta_{x^*} + \delta_{z^*}\right)^2 \\ &= \bar{Y}^2\left[m^2\left(C_{200} + C_{002} - 2C_{101}\right) - 2m\left(C_{011} - C_{110}\right) + C_{020}\right]. \end{aligned} \tag{17}$$

Riordinando in maniera opportuna i termini di quest'ultima, si ottiene l'espressione equivalente:

$$MSE\left(\hat{\bar{Y}}_{R1}^{*}\right) \doteq MSE\left(\hat{\bar{Y}}_{2}^{*}\right) + m\bar{Y}^{2}\left[m\left(C_{002} - 2C_{101}\right) - 2C_{011}\right]. \qquad (18)$$

Fissati i valori C_{rst}, dalla (17) si osserva che l'errore quadratico medio di $\hat{\bar{Y}}_{R1}^{*}$ è funzione, tramite *m*, della frazione di campionamento. Pertanto, può essere interessante determinare il valore ottimale di m che minimizza $MSE\left(\hat{\bar{Y}}_{R1}^{*}\right)$. E' chiaro che, una volta ottenuto tale valore, diciamo m_{opt}, rimane implicitamente definito anche il valore ottimale della frazione di campionamento

$$f_{opt} = \frac{m_{opt}}{1 + m_{opt}} \qquad (19)$$

che massimizza la precisione dello stimatore.

L'errore quadratico medio dello stimatore è minimizzato per:

$$m_{opt} = \frac{C_{011} - C_{110}}{C_{200} + C_{002} - 2C_{101}} \qquad (20)$$

a patto ovviamente che tale valore sia *ammissibile*, ovvero ricada nell'intervallo (0,1). In questo caso, sostituendo l'espressione di m_{opt} nella (17) si ottiene:

$$\min MSE\left(\hat{\bar{Y}}_{R1}^{*}\right) \doteq \bar{Y}^{2}\left[C_{020} - \frac{\left(C_{011} - C_{110}\right)^{2}}{C_{200} + C_{002} - 2C_{101}}\right]. \qquad (21)$$

Si osservi che il denominatore nella (20) altro non è che la derivata seconda (rispetto a *m*) di $MSE\left(\hat{\bar{Y}}_{R1}^{*}\right)$ e che tale derivata risulta essere sempre positiva. Infatti, indicato con $C(.)$ il *coefficiente di variazione* dello stimatore specificato in argomento e con $\rho(.,.)$ il *coefficiente di correlazione* tra due stimatori, è agevole stabilire in base alla (7) che:

$$C_{200} + C_{002} - 2C_{101} = C^{2}\left(\hat{\bar{X}}\right) + C^{2}\left(\hat{\bar{Z}}\right) - 2\rho\left(\hat{\bar{X}}, \hat{\bar{Z}}\right)C\left(\hat{\bar{X}}\right)C\left(\hat{\bar{Z}}\right).$$

Allora, per $\rho\left(\hat{\bar{X}}, \hat{\bar{Z}}\right) < 0$ l'espressione in esame è certamente positiva mentre, per $\rho\left(\hat{\bar{X}}, \hat{\bar{Z}}\right) > 0$ si ha:

$$C^2\left(\hat{\bar{X}}\right)+C^2\left(\hat{\bar{Z}}\right)-2\rho\left(\hat{\bar{X}},\hat{\bar{Z}}\right)C\left(\hat{\bar{X}}\right)C\left(\hat{\bar{Z}}\right)>\left[C\left(\hat{\bar{X}}\right)-C\left(\hat{\bar{Z}}\right)\right]^2>0\,.$$

Per gli altri tre stimatori, procedendo in maniera analoga, si perviene ai seguenti risultati:

- Stimatore $\hat{\bar{Y}}^*_{R2}$

$$B\left(\hat{\bar{Y}}^*_{R2}\right)\doteq m\bar{Y}\left[m\left(C_{200}+C_{002}+C_{101}\right)+C_{110}+C_{011}\right]$$

$$MSE\left(\hat{\bar{Y}}^*_{R2}\right)\doteq MSE\left(\hat{\bar{Y}}^*_2\right)+m\bar{Y}^2\left[m\left(C_{002}+2C_{101}\right)+2C_{011}\right] \tag{23}$$

$$m_{opt}=-\frac{C_{011}+C_{110}}{C_{200}+C_{002}+2C_{101}}$$

$$min\,MSE\left(\hat{\bar{Y}}^*_{R2}\right)\doteq\bar{Y}^2\left[C_{020}-\frac{\left(C_{011}+C_{110}\right)^2}{C_{200}+C_{002}+2C_{101}}\right];$$

- Stimatore $\hat{\bar{Y}}^*_{P1}$

$$B\left(\hat{\bar{Y}}^*_{P1}\right)\doteq m\bar{Y}\left(mC_{101}-C_{110}-C_{011}\right)$$

$$MSE\left(\hat{\bar{Y}}^*_{P1}\right)\doteq MSE\left(\hat{\bar{Y}}^*_1\right)+m\bar{Y}^2\left[m\left(C_{002}+2C_{101}\right)-2C_{011}\right] \tag{25}$$

$$m_{opt}=\frac{C_{011}+C_{110}}{C_{200}+C_{002}+2C_{101}}$$

$$min\,MSE\left(\hat{\bar{Y}}^*_{P1}\right)\doteq\bar{Y}^2\left[C_{020}-\frac{\left(C_{011}+C_{110}\right)^2}{C_{200}+C_{002}+2C_{101}}\right];$$

- Stimatore $\hat{\bar{Y}}^*_{P2}$

$$B\left(\hat{\bar{Y}}^*_{P2}\right)\doteq m\bar{Y}\left[m\left(C_{002}-C_{101}\right)+C_{011}-C_{110}\right]$$

$$MSE\left(\hat{\bar{Y}}_{P2}^{*}\right) \doteq MSE\left(\hat{\bar{Y}}_{1}^{*}\right) + m\bar{Y}^{2}\left[m\left(C_{002} - 2C_{101}\right) + 2C_{011}\right] \qquad (26)$$

$$m_{opt} = -\frac{C_{011} + C_{110}}{C_{200} + C_{002} - 2C_{101}}$$

$$\min MSE\left(\hat{\bar{Y}}_{P2}^{*}\right) \doteq \bar{Y}^{2}\left[C_{020} - \frac{\left(C_{011} - C_{110}\right)^{2}}{C_{200} + C_{002} - 2C_{101}}\right].$$

4. Confronti di efficienza

Le espressioni dell'errore quadratico medio ottenute per gli stimatori rapporto-cum-prodotto modificati consentono di realizzare utili confronti di efficienza al fine di individuare quale stimatore sia più conveniente utilizzare in relazione ai dati di cui il ricercatore dispone. Con tale intento, verranno di seguito definite le condizioni che permettono di stabilire in quali situazioni gli stimatori modificati risultano essere più efficienti degli stimatori $\hat{\bar{Y}}_{1}^{*}$ e $\hat{\bar{Y}}_{2}^{*}$, che impiegano la sola variabile ausiliaria *X*, e degli stimatori rapporto-cum-prodotto originari . Per quest'ultimo confronto, si analizzeranno solo gli stimatori che, in virtù di quanto già detto nel precedente paragrafo, sono da ritenersi complementari.

Un ulteriore confronto verrà realizzato, nell'ambito del campionamento casuale semplice, con lo stimatore per regressione multivariato provando come quest'ultimo sia più preciso degli stimatori rapporto-cum-prodotto sia originari che modificati.

Il confronto tra gli stimatori rapporto-cum-prodotto modificati e gli stimatori $\hat{\bar{Y}}_{1}^{*}$ e $\hat{\bar{Y}}_{2}^{*}$ consente di mettere in luce quanto segue:

TEOREMA 1

- $\hat{\bar{Y}}_{R1}^{*}$ *è più efficiente di* $\hat{\bar{Y}}_{2}^{*}$ *se:*

$$\begin{aligned} \frac{m}{2} &< \frac{C_{011}}{C_{002} - 2C_{101}} \quad \textit{quando} \quad C_{002} - 2C_{101} > 0 \\ \frac{m}{2} &> \frac{C_{011}}{C_{002} - 2C_{101}} \quad \textit{quando} \quad C_{002} - 2C_{101} < 0; \end{aligned} \qquad (27)$$

- $\hat{\bar{Y}}_{R2}^{*}$ *è più efficiente di* $\hat{\bar{Y}}_{2}^{*}$ *se:*

$$\begin{aligned} &\frac{m}{2} < -\frac{C_{011}}{C_{002}+2C_{101}} \quad \textit{quando} \quad C_{002}+2C_{101}>0 \\ &\frac{m}{2} > -\frac{C_{011}}{C_{002}+2C_{101}} \quad \textit{quando} \quad C_{002}+2C_{101}<0; \end{aligned} \tag{28}$$

- $\hat{\bar{Y}}_{P1}^{*}$ *è più efficiente di* $\hat{\bar{Y}}_{1}^{*}$ *se:*

$$\begin{aligned} &\frac{m}{2} < \frac{C_{011}}{C_{002}+2C_{101}} \quad \textit{quando} \quad C_{002}+2C_{101}>0 \\ &\frac{m}{2} > \frac{C_{011}}{C_{002}+2C_{101}} \quad \textit{quando} \quad C_{002}+2C_{101}<0; \end{aligned} \tag{29}$$

- $\hat{\bar{Y}}_{P2}^{*}$ *è più efficiente di* $\hat{\bar{Y}}_{1}^{*}$ *se:*

$$\begin{aligned} &\frac{m}{2} < -\frac{C_{011}}{C_{002}-2C_{101}} \quad \textit{quando} \quad C_{002}-2C_{101}>0 \\ &\frac{m}{2} > -\frac{C_{011}}{C_{002}-2C_{101}} \quad \textit{quando} \quad C_{002}-2C_{101}<0; \end{aligned} \tag{30}$$

DIMOSTRAZIONE
Si proverà solo la prima condizione essendo le altre ottenibili allo stesso modo. Dalla (18) si ha che $MSE\left(\hat{\bar{Y}}_{R1}^{*}\right) < MSE\left(\hat{\bar{Y}}_{2}^{*}\right)$ se $m^2\left(C_{002}-2C_{101}\right)-2mC_{011}<0$, ovvero se $m\left(C_{002}-2C_{101}\right)<2C_{011}$. Da quest'ultima disuguaglianza, considerando il segno di $C_{002}-2C_{101}$ e risolvendo rispetto a *m*, si perviene immediatamente alla (27). ◆

I risultati ottenuti da questi primi confronti sottolineano, come peraltro già evidenziato per gli stimatori originari (1) e (2), che l'impiego di due variabili ausiliarie non conduce necessariamente a stime più precise di quelle ottenibili considerandone una sola.

Estendendo l'analisi al confronto tra gli stimatori originari e quelli modificati si perviene alle seguenti conclusioni:

TEOREMA 2

- $\hat{\bar{Y}}^{*}_{R1}$ *è più efficiente di* $\hat{\bar{Y}}_{P2}$ *se:*

$$\frac{1+m}{2} > \frac{C_{011}-C_{110}}{C_{200}+C_{002}-2C_{101}}\,; \tag{31}$$

- $\hat{\bar{Y}}^{*}_{R2}$ *è più efficiente di* $\hat{\bar{Y}}_{P1}$ *se:*

$$\frac{1+m}{2} > -\frac{C_{011}+C_{110}}{C_{200}+C_{002}+2C_{101}}\,; \tag{32}$$

- $\hat{\bar{Y}}^{*}_{P1}$ *è più efficiente di* $\hat{\bar{Y}}_{R2}$ *se:*

$$\frac{1+m}{2} > \frac{C_{011}+C_{110}}{C_{200}+C_{002}+2C_{101}}\,; \tag{33}$$

- $\hat{\bar{Y}}^{*}_{P2}$ *è più efficiente di* $\hat{\bar{Y}}_{R1}$ *se:*

$$\frac{1+m}{2} > \frac{C_{110}-C_{011}}{C_{200}+C_{002}-2C_{101}}\,. \tag{34}$$

DIMOSTRAZIONE

Relativamente alla (31), posto $\Delta\left(\hat{\bar{Y}}_{P2},\hat{\bar{Y}}^{*}_{R1}\right) = MSE\left(\hat{\bar{Y}}_{P2}\right) - MSE\left(\hat{\bar{Y}}^{*}_{R1}\right)$, per la (6) e la (17) si ha:

$$\Delta\left(\hat{\bar{Y}}_{P2},\hat{\bar{Y}}^{*}_{R1}\right) = \bar{Y}^{2}\left[\left(1-m^{2}\right)\left(C_{200}+C_{002}-2C_{101}\right)-2(1-m)\left(C_{011}-C_{110}\right)\right].$$

Tenuto conto che, per ipotesi, $f < 0.5$ segue che $(1-m) > 0$. Pertanto $\Delta\left(\hat{\bar{Y}}_{P2},\hat{\bar{Y}}^{*}_{R1}\right) > 0$ se $(1+m)\left(C_{200}+C_{002}-2C_{101}\right) > 2\left(C_{011}-C_{110}\right)$. Ma, per precedenti considerazioni, $C_{200}+C_{002}-2C_{101} > 0$ e, quindi, la disuguaglianza resta provata.
Le condizioni (32), (33) e (34) si provano procedendo in maniera analoga. ◆

Il confronto pone in evidenza come la modifica apportata agli stimatori rapporto-cum-prodotto può risultare vantaggiosa se la frazione di campionamento è scelta in maniera appropriata. In particolare, se si considera la situazione di massima precisione degli stimatori modificati, valutandone l'errore quadratico medio in $m_{opt} \in (0,1)$, è agevole verificare che le condizioni (31), (32), (33) e (34) sono sempre vere. Pertanto, in tale situazione, gli stimatori modificati possono essere considerati sempre più efficienti dei corrispondenti stimatori originari.

Le espressioni della distorsione e dell'errore quadratico medio degli stimatori modificati, nonché le condizioni di superiorità di uno stimatore rispetto ad un altro, pur essendo state ricavate in termini generali, possono essere adattate ad un particolare piano di campionamento sostituendo a C_{rst} le corrispondenti espressioni indotte dallo schema di selezione prescelto. Nel caso in cui si decida di ricorrere al *campionamento casuale semplice senza reimmissione* (*srswor*), risulta abbastanza semplice particolarizzare C_{rst}. Infatti, detta *U* una generica variabile che assume valori positivi e indicato con C_u il corrispondente coefficiente di variazione:

$$C_u = \frac{S_u}{\overline{U}} \quad \text{con} \quad S_u = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(U_i - \overline{U}\right)^2}$$

è agevole provare (Cochran, 1977) che:

$$C_{200} = C_x^2 \text{ , } C_{020} = C_y^2 \text{ , } C_{002} = C_z^2 \text{ ,}$$

$$C_{110} = \rho_{xy} C_x C_y \text{ , } C_{101} = \rho_{xz} C_x C_z \text{ , } C_{011} = \rho_{yz} C_y C_z$$

essendo ρ_{uv} il coefficiente di correlazione tra le variabili *U* e *V*.

Alla luce di queste nuove posizioni possiamo stabilire, relativamente al confronto tra stimatori rapporto-cum-prodotto originari e modificati, che:

(1) $\hat{\overline{Y}}^*_{R1}$ è più efficiente di $\hat{\overline{Y}}_{P2}$ se $\dfrac{1+m}{2} > C_y \dfrac{C_z \rho_{yz} - C_x \rho_{xy}}{C_x^2 + C_z^2 - 2C_x C_z \rho_{xz}}$;

(2) $\hat{\overline{Y}}^*_{R2}$ è più efficiente di $\hat{\overline{Y}}_{P1}$ se $\dfrac{1+m}{2} > -C_y \dfrac{C_z \rho_{yz} + C_x \rho_{xy}}{C_x^2 + C_z^2 + 2C_x C_z \rho_{xz}}$;

(3) $\hat{\overline{Y}}^*_{P1}$ è più efficiente di $\hat{\overline{Y}}_{R2}$ se $\dfrac{1+m}{2} > C_y \dfrac{C_z \rho_{yz} + C_x \rho_{xy}}{C_x^2 + C_z^2 + 2C_x C_z \rho_{xz}}$;

(4) $\hat{\overline{Y}}^*_{P2}$ è più efficiente di $\hat{\overline{Y}}_{R1}$ se $\dfrac{1+m}{2} > -C_y \dfrac{C_z \rho_{yz} - C_x \rho_{xy}}{C_x^2 + C_z^2 - 2C_x C_z \rho_{xz}}$.

4.1 *Confronto di efficienza con lo stimatore per regressione multivariato*

Confrontiamo ora nel *srswor* gli stimatori rapporto-cum-prodotto modificati con lo *stimatore per regressione multivariato*, $\hat{\bar{Y}}_{reg}$, che impiega le variabili ausiliarie *X* e *Z* e ricorre alle sole osservazioni campionarie.

Nella sua tradizionale formulazione lo stimatore $\hat{\bar{Y}}_{reg}$ si presenta come (Sukhatme *et al*., 1984):

$$\hat{\bar{Y}}_{reg} = \bar{y} + \hat{\beta}_{yx.z}\left(\bar{X} - \bar{x}\right) + \hat{\beta}_{yz.x}\left(\bar{Z} - \bar{z}\right) \tag{35}$$

dove $\bar{u} = n^{-1}\sum_{i=1}^{n} u_i$ è la media campionaria della variabile *U* (*U*=*X*,*Y*,*Z*) mentre $\hat{\beta}_{yx.z}$ e $\hat{\beta}_{yz.x}$ rappresentano le stime dei minimi quadrati, rispettivamente, del *coefficiente di regressione parziale* di *Y* su *X*, $\beta_{yx.z}$, e di *Y* su *Z*, $\beta_{yz.x}$. Lo stimatore $\hat{\bar{Y}}_{reg}$ è distorto per $\bar{Y}$ anche se, per grandi campioni, la distorsione risulta trascurabile. L'errore quadratico medio è esprimibile nella forma:

$$MSE\left(\hat{\bar{Y}}_{reg}\right) \doteq \frac{1-f}{n} S_y^2\left(1 - R_{y.xz}^2\right) \tag{36}$$

dove

$$R_{y.xz} = \sqrt{\frac{\rho_{xy}^2 + \rho_{yz}^2 - 2\rho_{xy}\rho_{yz}\rho_{xz}}{1-\rho_{xz}^2}}$$

rappresenta il *coefficiente di correlazione multipla* tra *Y* e *X*, *Z*.

Per confrontare lo stimatore $\hat{\bar{Y}}_{reg}$ con gli stimatori rapporto-cum-prodotto modificati analizziamo la situazione più favorevole a questi ultimi, ovvero supponiamo che la frazione di campionamento possa essere determinata in modo da massimizzarne la precisione. Consideriamo, quindi, l'errore quadratico medio degli stimatori (11) e (12) valutati in $m = m_{opt}$.

Il confronto analitico tra $MSE\left(\hat{\bar{Y}}_{reg}\right)$ e $min\, MSE\left(\hat{\bar{Y}}^*_{R-P}\right)$, con $\hat{\bar{Y}}^*_{R-P} = \hat{\bar{Y}}^*_{R1}, \hat{\bar{Y}}^*_{R2}, \hat{\bar{Y}}^*_{P1}, \hat{\bar{Y}}^*_{P2}$, consente di stabilire quanto segue:

TEOREMA 3

Nel srswor lo stimatore $\hat{\bar{Y}}_{reg}$ è sempre più efficiente degli stimatori rapporto-cum-prodotto modificati.

DIMOSTRAZIONE

Posto $\Delta\left(\hat{\bar{Y}}_{reg}, \hat{\bar{Y}}^*_{R-P}\right) = MSE\left(\hat{\bar{Y}}_{reg}\right) - min\, MSE\left(\hat{\bar{Y}}^*_{R-P}\right)$, lo stimatore $\hat{\bar{Y}}^*_{R-P}$ risulta essere più efficiente di $\hat{\bar{Y}}_{reg}$ se $\Delta\left(\hat{\bar{Y}}_{reg}, \hat{\bar{Y}}^*_{R-P}\right) > 0$. Dopo alcuni passaggi algebrici è possibile provare che tale disuguaglianza è soddisfatta se:

(1) $\left[\rho_{xz}\left(C_z\rho_{yz} + C_x\rho_{xy}\right) - \left(C_x\rho_{yz} + C_z\rho_{xy}\right)\right]^2 < 0$ per $\hat{\bar{Y}}^*_{R-P} = \hat{\bar{Y}}^*_{R1}$;

(2) $\left[\rho_{xz}\left(C_z\rho_{yz} - C_x\rho_{xy}\right) + \left(C_x\rho_{yz} - C_z\rho_{xy}\right)\right]^2 < 0$ per $\hat{\bar{Y}}^*_{R-P} = \hat{\bar{Y}}^*_{R2}$;

(3) $\left[\rho_{xz}\left(C_z\rho_{yz} - C_x\rho_{xy}\right) - \left(C_x\rho_{yz} - C_z\rho_{xy}\right)\right]^2 < 0$ per $\hat{\bar{Y}}^*_{R-P} = \hat{\bar{Y}}^*_{P1}$;

(4) $\left[\rho_{xz}\left(C_z\rho_{yz} + C_x\rho_{xy}\right) + \left(C_x\rho_{yz} + C_z\rho_{xy}\right)\right]^2 < 0$ per $\hat{\bar{Y}}^*_{R-P} = \hat{\bar{Y}}^*_{P2}$.

Ovviamente, nessuna della precedenti condizioni è realizzabile, per cui $\hat{\bar{Y}}_{reg}$ è sempre più efficiente degli stimatori modificati con errore quadratico minimo. Ma allora, lo stimatore per regressione sarà certamente più efficiente degli stimatori modificati anche nelle situazioni in cui $m \neq m_{opt}$. ◆

Ricordiamo che nella situazione ottimale presa in esame, gli stimatori rapporto-cum-prodotto modificati sono più efficienti di quelli originari. Pertanto, in virtù di quanto stabilito dall'ultimo teorema, possiamo affermare per transitività che nel *srswor* lo stimatore per regressione multivariato è anche più efficiente degli stimatori rapporto-cum-prodotto originari.

Quest'ultima considerazione può essere chiaramente estesa a tutte le altre situazioni, diverse da quelle ottimali, in cui gli stimatori modificati risultano, in base alle condizioni riportate nel teorema 2, più efficienti di quelli originari.

Per concludere, facciamo osservare che la modifica apportata agli stimatori rapporto-cum-prodotto non ha alcun effetto sullo stimatore per regressione, nel senso che se si considera lo *stimatore per differenza multivariato:*

$$\hat{\bar{Y}}_D = \bar{y} + B_1\left(\bar{X} - \bar{x}^*\right) + B_2\left(\bar{Z} - \bar{z}^*\right)$$

con

$$\bar{x}^* = \frac{N\bar{X} - n\bar{x}}{N-n} \quad \text{e} \quad \bar{z}^* = \frac{N\bar{Z} - n\bar{z}}{N-n}$$

e si determinano i valori delle costanti B_1 e B_2 che ne minimizzano la varianza, si ottiene lo stimatore per regressione:

$$\hat{\bar{Y}}'_{reg} = \bar{y} + \beta_{yx.z}\left(\bar{X} - \bar{x}\right) + \beta_{yz.x}\left(\bar{Z} - \bar{z}\right)$$

che coincide esattamente con l'espressione definita nella (35) se ai coefficienti $\beta_{yx.z}$ e $\beta_{yz.x}$ si sostituiscono le rispettive stime ottenibili unicamente dal campione selezionato.

5. Conclusioni

Con il presente lavoro si è inteso condurre un'analisi esplorativa sull'efficienza di una versione modificata degli stimatori rapporto-cum-prodotto. Gli stimatori analizzati presentano la stessa struttura di quelli originari ma, sotto la realistica ipotesi che la frazione di campionamento sia inferiore a 0.5, utilizzano un maggiore contenuto informativo derivante, non solo dalle unità campionate, ma anche dalle unità della popolazione non inserite nel campione. Se la frazione di campionamento è opportunamente scelta, gli stimatori modificati possono risultare più efficienti di quelli originari. In particolare, l'impiego degli stimatori modificati risulta sempre vantaggioso se si adotta una frazione di campionamento ottimale che consente di minimizzarne l'errore quadratico medio. Tuttavia, in questa situazione ottimale si è provato nell'ambito del *srswor* che gli stimatori rapporto-cum-prodotto originari e modificati non sono mai più efficienti dello stimatore ottenuto con il metodo della regressione. Questo risultato può essere esteso, ovviamente, a tutte le altre situazioni (non ottimali) in cui gli stimatori modificati sono da ritenersi più efficienti degli originari.

Riferimenti Bibliografici

Bandyopadhyay S. (1980). Improved ratio and product estimators. *Sankhyā, C*, **42**, 45-49.

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

Murthy M.N. (1977). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

Singh M.P. (1965). On the estimation of ratio and product of the population parameters. *Sankhyā, B*, **27**, 321-328.

Singh M.P. (1967). Ratio cum product method of estimation. *Metrika*, **12**, 34-42.

Srivenkataramana T. (1980). A dual to ratio estimator in sample surveys. *Biometrika*, **67**, 199-204.

Sukhatme P.V., Sukhatme B.V., Sukhatme S., Asok C. (1984). *Sampling Theory of Surveys with Applications*. 3rd edt., Iowa State University Press, Ames.