Il Campionamento a due stadi

Campionamento a due stadi

- Il campionamento a 2 stadi nasce per superare un problema che si verifica con il campionamento a grappoli:
  - nel caso in cui i grappoli (unità primarie) non presentano dimensione uguale (caso frequente), si ottengono stime distorte.
- Per risolvere tale problema si possono percorrere due strade:
  - utilizzare il campionamento a probabilità variabili;
  - utilizzare il campionamento a 2 stadi;

Campionamento a due stadi

- il campionamento a 2 stadi si può considerare un'estensione del campionamento a grappoli:
  - nel caso in cui dalle Unità Primarie estraiamo tutte le Unità Secondarie, otteniamo un campione a grappoli di dimensioni pari al numero delle unità secondarie estratte.
  - Se si effettua un ulteriore campionamento delle unità secondarie otteniamo un campionamento a due stadi

Campionamento a due stadi: caratteristiche

- 1° stadio: estrazione, senza ripetizione, dalla popolazione iniziale, di un campione casuale di unità primarie.

- 2° stadio: selezione, senza ripetizione, dalle unità primarie, di un certo numero di unità elementari (unità secondarie).

- Esempio: si estraggono alcuni Comuni (unità primarie) e da ciascuno di essi si selezionano alcune famiglie (unità secondarie). Si parla, in generale, di sottocampionamento.

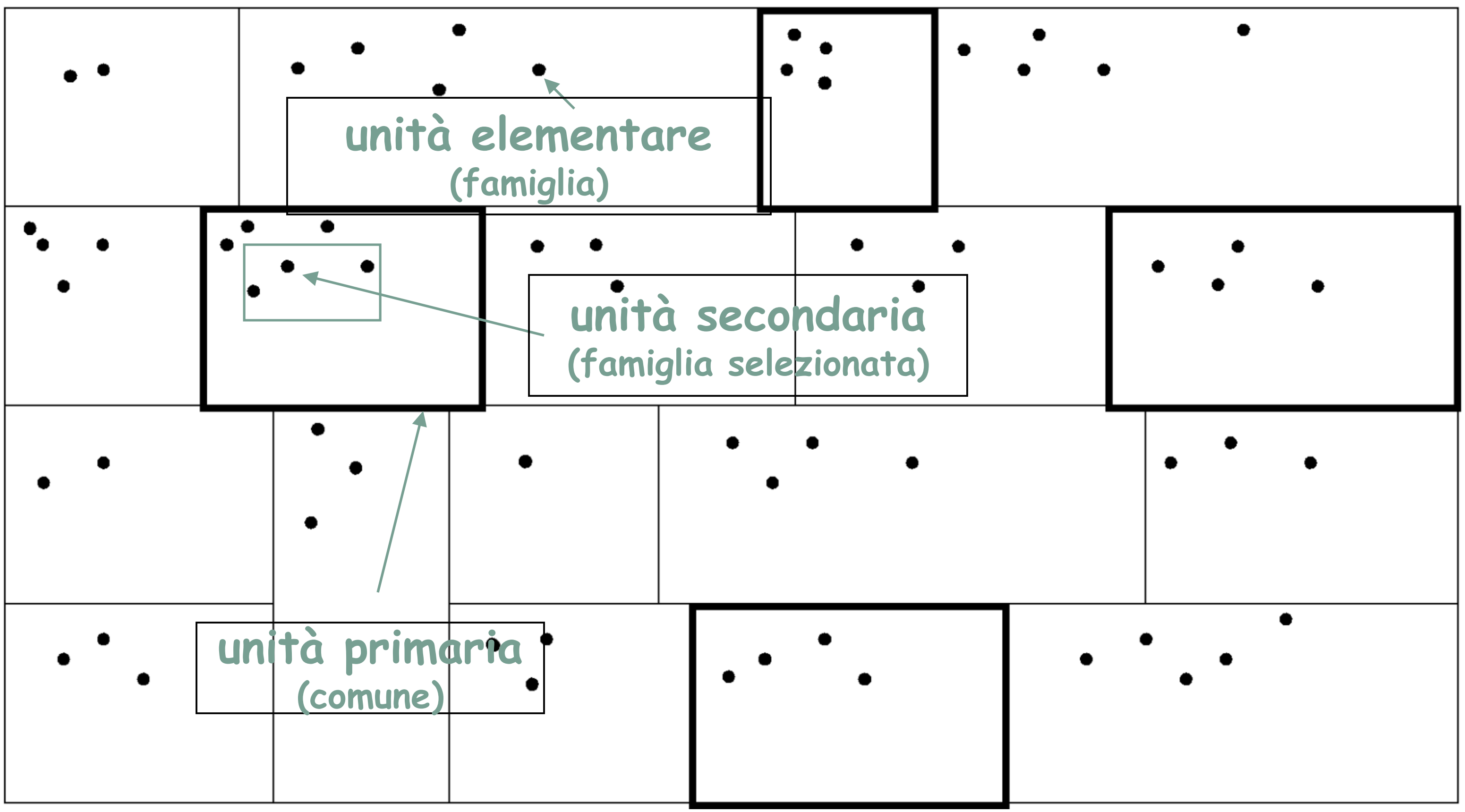
Campionamento a due stadi: definizione

Si consideri una popolazione P:

Si chiama Campionamento a due stadi senza ripetizione il piano che consiste nell'estrarre, senza ripetizione, un campione casuale di unità primarie e nel selezionare, senza ripetizione, da ogni unità primaria estratta un certo numero di unità elementari.

molto spesso, in un campionamento a 2 stadi, il campionamento di I stadio è un campionamento a grappoli, ma potrebbe essere qualsiasi altro tipo di campionamento casuale.

Campionamento a due stadi: esempio



Campionamento a due stadi: caratteristiche

- In generale, la definizione di campionamento a due stadi (CDS) può facilmente essere estesa a tre o più stadi, in ognuno dei quali si può scegliere la tecnica più idonea (elasticità):

  - Casuale semplice
  - Casuale stratificato
  - Casuale sistematico

- Campionamento a grappoli (CGR): le unità primarie sono rappresentate dai grappoli e le unità secondarie sono tutte quelle appartenenti ai grappoli estratti (e non solo una parte).

  Il CDS può essere definito come un CGR con sottocampionamento.

Campionamento a due stadi: pro e contro

- Vantaggi dal punto di vista economico e organizzativo:

  - Riduzione dei costi e dei tempi necessari per la scelta del campione.

- Svantaggi dal punto di vista statistico:

  - Diminuzione della precisione delle stime all'aumentare del numero degli stadi.

  Possibile soluzione: aumentare la dimensione campionaria in ogni stadio.

Notazioni

- U.P. = Unità Primarie U.S. /u.e. = unità Secondarie o elementari
- N = numero di unità primarie nella popolazione
- n = numero di unità primarie nel campione
- M_i = numero totale di unità elementari dell'unità primaria i-esima
- m_i = numero totale di unità elementari estratte dall'unità primaria i-esima
- $M.$ = numero di unità elementari nella popolazione
- $\sum_{i=1}^{n} m_i$ ampiezza campionaria (in termini di unità elementari)
- $f_1 = \frac{n}{N}$ tasso di sondaggio di primo stadio
- $f_{2i} = \frac{m_i}{M_i}$ tasso di sondaggio di secondo stadio

Notazioni

Y_{ij} Valore della variabile indagata per la j-esima unità secondaria nell'unità primaria i (dato di popolazione)

y_{ij} Valore della variabile indagata per la j-esima unità secondaria nell'unità primaria i (dato campionario)

$\overline{Y}_i$ Media della variabile indagata nell'unità primaria j (dato di popolazione)

$\overline{y}_i$ Media della variabile indagata nell'unità primaria j (dato campionario)

$\overline{\overline{Y}}$ Media della variabile indagata nella popolazione

$\overline{\overline{y}}$ Media della variabile indagata nel campione

Notazioni

$$\hat{Y}_{ij} = \sum_{i=1}^{N} \sum_{j=1}^{M} Y_{ij}$$

Totale della variabile indagata per la j-esima unità secondaria nell'unità primaria i (dato di popolazione)

$$\hat{y}_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij}$$

Valore della variabile indagata per la j-esima unità secondaria nell'unità primaria i (dato campionario)

Notazioni

$$S_1^2 = \frac{\sum_{i=1}^{N}\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2}{N-1}$$

Varianza fra le unità primarie (dato di popolazione)

$$S_2^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(Y_{ij} - \bar{Y}_i\right)^2}{N(M-1)}$$

Varianza fra le unità secondarie all'interno delle unità primarie (dato di popolazione)

$$s_1^2 = \frac{\sum_{i=1}^{n}\left(\bar{y}_i - \bar{\bar{y}}\right)^2}{n-1}$$

Stima della varianza fra le unità primarie (dato campionario)

$$s_2^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij} - \bar{y}_i\right)^2}{n(m-1)}$$

Stima della varianza fra le unità secondarie all'interno delle unità primarie (dato campionario)

Probabilità di inclusione del 1°ordine

La probabilità di inclusione del primo ordine dell'unità *j* è la probabilità congiunta di estrarre l'unità primaria *i* nel primo stadio di campionamento e l'unità secondaria *j* (tra le M_i appartenenti all'unità primaria considerata) nel secondo stadio.

$$\pi_{ij} = P_i P_{j/i} = \frac{n}{N} \frac{m_i}{M_i} = f_1 f_{2i}$$

dove: *Pi* è *la* probabilità di estrarre l'unità primaria *i*
Pj/i è *la* probabilità di estrarre l'unità secondaria *j* una volta estratta l'unità primaria i

Nota: il CDS sarà autoponderante quando

$$f_{21} = f_{22} = \dots\dots\dots = f_{2N} = \frac{m_1}{M_1} = \frac{m_2}{M_2} = \dots\dots\dots = \frac{m_N}{M_N}$$

Probabilità di inclusione del 2°ordine

Per calcolare la probabilità di inclusione del secondo ordine, considerata la generica coppia di unità ij ed i'k, distinguiamo 2 casi:

- Le due unità appartengono alla stessa unità primaria (allo stesso grappolo) (*i* = *i'*)
- Le unità appartengono a unità primarie diverse (grappoli diversi) (*i* ≠ *i'*).

$$\pi_{(ij)(i'k)} = \begin{cases} \dfrac{n}{N}\dfrac{m_i}{M_i}\dfrac{m_i-1}{M_i-1} & se \quad i=i' \\[2ex] \dfrac{n}{N}\dfrac{n-1}{N-1}\dfrac{m_i}{M_i}\dfrac{m_{i'}}{M_{i'}} & se \quad i\neq i' \end{cases}$$

Stimatori nel caso di unità di primo stadio aventi la stessa dimensione

Consideriamo il caso più semplice in cui una popolazione è costituita da N unità primarie aventi tutte la medesima dimensione **M**, per cui risulta:

$$M_1=M_2= \ldots\ldots=M_i=\ldots\ldots M_N = \mathbf{M}$$

Per effettuare il campionamento a due stadi estrarremo, con uno dei metodi noti, *n* fra queste *unità primarie* e successivamente in ciascuna di ciascuna di esse estrarremo *m* unità secondarie.

Se θ è il parametro della popolazione

$$\hat{\theta}$$

è uno stimatore di θ di cui vogliamo trovare valore atteso e la varianza. Un modo per ricavarli consiste nel considerare tutte le unità di secondo stadio che provengono da un fissato insieme di *n* unità di primo stadio

Stimatori nel caso di unità di primo stadio aventi la stessa dimensione

$\hat{\theta}$ uno stimatore di θ

$E_2(.)=$ Valor medio, calcolato sulle unità di secondo stadio estratte dalle n unità primarie

quindi

$E_1(.)=$ Valor medio calcolato su tutte le n unità di primo stadio estratte

Stimatori nel caso di unità di primo stadio aventi la stessa dimensione

Il **valore atteso dello stimatore** $\hat{\theta}$

ovvero il valor medio su tutti i possibili campioni sarà espresso da:

$$E\left(\hat{\theta}\right)=E_1\left[E_2\left(\hat{\theta}\right)\right]$$

Per quanto riguarda la **varianza dello stimatore** $\hat{\theta}$

si può dimostrare che essa è data dalla:

$$V\left(\hat{\theta}\right)=V_1\left[E_2\left(\hat{\theta}\right)\right]+E_1\left[V_2\left(\hat{\theta}\right)\right]$$

dove:
E è il valore atteso su tutti i possibili campioni
E_2 e V_2 sono il valore atteso e la varianza sui dati del campione di secondo stadio
E_1 e V_1 sono il valore atteso e la varianza sui dati del campione di primo stadio

Stimatori nel caso di unità di primo stadio aventi la stessa dimensione

Supponiamo $\theta = E\left(\widehat{\theta}\right)$

Per definizione $V\left(\widehat{\theta}\right) = E\left(\widehat{\theta} - \theta\right)^2 = E_1\left[E_2\left(\widehat{\theta} - \theta\right)^2\right]$

Ma

$$E_2\left(\widehat{\theta} - \theta\right)^2 = E_2\left(\widehat{\theta}^2\right) - 2\theta E_2\left(\hat{\theta}\right) + \theta^2$$

Ed essendo

$$V_2(\theta) = E_2\left(\widehat{\theta}^2\right) - [E_2(\widehat{\theta})]^2$$

se ne deduce che

$$E_2\left(\widehat{\theta}^2\right) = V_2(\theta) + [E_2(\widehat{\theta})]^2$$

Stimatori nel caso di unità di primo stadio aventi la stessa dimensione

Risulterà

$$E_2\left(\widehat{\theta}-\theta\right)^2 = V_2(\theta)+[E_2(\widehat{\theta})]^2 - 2\theta E_2\left(\widehat{\theta}\right)+\theta^2$$

Sostituendo a E_2 il suo valore e facendo la media delle unità del campione di primo stadio si ottiene:

$$V(\theta)=E_1\left[E_2\left(\widehat{\theta}\right)\right]^2-\theta^2+E_1\left[V_2\left(\widehat{\theta}\right)\right]$$

Stimatori nel caso di unità di primo stadio aventi la stessa dimensione

Ed essendo $E_1[E_2(\hat{\theta})] = \theta$

$$V\left(\hat{\theta}\right) = E_1\left[E_2\left(\hat{\theta}\right)\right]^2 + E_1\left[V_2\left(\hat{\theta}\right)\right] - \theta^2 = E_1\left[E_2\left(\hat{\theta}\right)\right]^2 - \left\{E_1\left[E_2\left(\hat{\theta}\right)\right]\right\}^2 + E_1\left[V_2\left(\hat{\theta}\right)\right]$$

$$= V_1\left[E_2\left(\hat{\theta}\right)\right] + E_1\left[V_2\left(\hat{\theta}\right)\right]$$

Nel caso di tre stadi si avrà:

$$V(\theta) = V_1\left\{E_2\left[E_3\left(\hat{\theta}\right)\right]\right\} + E_1\left\{V_2\left[E_3\left(\hat{\theta}\right)\right]\right\} + E_1\left\{E_2\left[V_3\left(\hat{\theta}\right)\right]\right\}$$

Stima dei parametri

- STIMA DELLA MEDIA
- STIMA DEL TOTALE
- STIMA DELLA PROPORZIONE

Primo caso:

Campionamento casuale semplice in entrambi gli stadi

Stima dei parametri (ccs)

CASO DI ESTRAZIONE CASUALE DELLE UNITA' PRIMARIE E SECONDARIE

Stima della media

y_{ij} → valore di Y osservato nell'unità secondaria j estratta dall'unità primaria i-esima

$$\bar{y}_i = \frac{\sum_{j=1}^{m} y_{ij}}{m}$$

→ media delle unità secondarie (elementi del campione estratto dall'i-esima unità primaria)

$$\bar{\bar{y}} = \frac{\sum_{i=1}^{n} \bar{y}_i}{n} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm}$$

→ media di Y in tutte le unità del campione, vista come media di Y calcolata nelle unità primarie o come media delle medie di Y calcolata in tutte le unità secondarie

Stima dei parametri (ccs)

CASO DI ESTRAZIONE CASUALE DELLE PRIME E SECONDE UNITA'

Stima della media

$$\bar{\bar{y}} = \frac{\sum_{i=1}^{n} \bar{y}_i}{n} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}}{nm} \qquad \bar{\bar{Y}} = \frac{\sum_{i=1}^{N} \bar{Y}_i}{N} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} Y_{ij}}{NM}$$

<u>Proprietà</u>

$$E\left(\bar{\bar{y}}\right) = \bar{\bar{Y}}$$

Stimatore non distorto

$$V\left(\bar{\bar{y}}\right) = \left(\frac{N-n}{N}\right)\frac{S_1^2}{n} + \left(\frac{M-m}{M}\right)\frac{S_2^2}{mn}$$

Varianza dello stimatore media

Stima dei parametri (ccs)

Stima della media

PROPRIETA' DELLO STIMATORE:

I campioni di I e II stadio sono entrambi casuali semplici.

Uno stimatore non distorto della media nei due stadi è la media campionaria di I e II stadio.

Applicando la scomposizione della media si ottiene:

$$E\left(\bar{\bar{y}}\right) = E_1\left[E_2\left(\bar{\bar{y}}\right)\right] = E_1\left(\frac{\sum_{i=1}^{n} \bar{Y}_i}{n}\right) = \frac{\sum_{i=1}^{N} \bar{Y}_i}{N} \bar{\bar{Y}}$$

$$E_2\left(\bar{\bar{y}}\right) = \frac{\sum_{i=1}^{n} \bar{Y}_i}{n}$$

Stima dei parametri (ccs)

Varianza dello stimatore media

$$V(\theta) = V_1\left[E_2\left(\hat{\theta}\right)\right] + E_1\left[V_2\left(\hat{\theta}\right)\right]$$

$$V\left(\bar{\bar{y}}\right) = V_1\left[E_2\left(\bar{\bar{y}}\right)\right] + E_1\left[V_2\left(\bar{\bar{y}}\right)\right]$$

Poiché:

$$E_2\left(\bar{\bar{y}}\right) = \frac{\sum_{i=1}^{n} \bar{Y}_i}{n} \qquad V_2\left(\bar{\bar{y}}\right) = \frac{M-m}{Mn^2} \frac{\sum_{i=1}^{n} S_{2i}^2}{m}$$

allora:

$$V_1\left[E_2\left(\bar{\bar{y}}\right)\right] = \left(\frac{N-n}{n}\right) \frac{S_1^2}{n}$$

Stima dei parametri (ccs)

Varianza dello stimatore media

Dove:

$$S_1^2 = \frac{\sum_{i=1}^{N}\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2}{N-1} \qquad S_{2i}^2 = \frac{\sum_{j=1}^{M}\left(y_{ij} - \bar{Y}_i\right)^2}{M-1}$$

S_{2i}^2 è la varianza tra le unità secondarie per l'unità primaria i-esima

Quando ne facciamo la media (fra tutte le unità primarie), si ottiene:

$$\frac{\sum_{i=1}^{N} S_{2i}^2}{N} = S_2^2$$

per cui:

$$E_1\left[V_2\left(\bar{\bar{y}}\right)\right] = \left(\frac{M-m}{M}\right)\frac{S_2^2}{mn}$$

Stima dei parametri (ccs)

Varianza dello stimatore media

se $$f_1 = \frac{n}{N} \qquad f_2 = \frac{m}{M}$$

allora:

$$V\left(\bar{\bar{y}}\right) = \frac{1-f_1}{n} S_1^2 + \frac{1-f_2}{mn} S_2^2$$

La cui stima sui dati del campione è:

$$\hat{v}\left(\bar{\bar{y}}\right) = \frac{1-f_1}{n} s_1^2 + \frac{f_1\left(1-f_2\right)}{mn} s_2^2$$

Se f_1 è trascurabile (popolazione infinita), allora:

$$v\left(\bar{\bar{y}}\right) = \frac{s_1^2}{n} = \frac{\sum_{i=1}^{n}\left(\bar{y}_i - \bar{\bar{y}}\right)^2}{n\left(n-1\right)}$$

Stima dei parametri (ccs)

Stima del totale

$$\hat{Y}_{ts} = \frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} = N\bar{y}_{ts} = N\bar{\bar{y}}$$

$$E\left(\hat{Y}_{ts}\right) = \hat{Y}$$

Stimatore non distorto

$$V\left(\hat{Y}_{ts}\right) = N^2 V\left(\bar{Y}_{ys}\right) = N^2\left(\frac{1-f}{n}S_1^2 + \frac{1-f_2}{nm}S_2^2\right)$$

Varianza dello stimatore

Stima dei parametri (ccs)

Stima della proporzione

$P_i = \frac{A_i}{M}$ $\rightarrow$ Proporzione nell'unità i-esima della popolazione

$p_i = \frac{a_i}{m}$ $\rightarrow$ Proporzione nell'unità i-esima del campione

Il parametro da stimare è:

$$P = \frac{\sum_{i=1}^{N} A_i}{\sum_{i=1}^{N} M} = \frac{1}{N}\sum_{i=1}^{N}\frac{A_i}{M} = \frac{1}{N}\sum_{i=1}^{N} P_i$$

Lo stimatore della proporzione è:

$$p_{TS} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} M} = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i}{M} = \frac{1}{n}\sum_{i=1}^{n} p_i$$

La varianza si ottiene dalla formula dello stimatore della media sostituendo

P_i a Y_i

Stima dei parametri (ccs)

Stima della proporzione

Lo stimatore è non distorto, per cui:

$$E(p) = E_1\left[E_2\left(\frac{1}{n}\sum_{i=1}^{n} p_i\right)\right] = E_1\left[\frac{1}{n}\sum_{i=1}^{n} P_i\right] = \frac{1}{N}\sum_{i=1}^{N} P_i = P$$

ricordando che:

$$V(\bar{\bar{y}}) = \frac{(1-f_1)}{n} S_1^2 + \frac{(1-f_2)}{nm} S_2^2$$

e $\bar{\bar{y}} = p_{TS}$ si avrà:

$$S_1^2 = \frac{\sum_{i=1}^{N} (P_i - P)^2}{N-1} \qquad e \qquad S_2^2 = \frac{M}{N(M-1)} \sum_{i=1}^{N} P_i Q_i$$

Stima dei parametri (ccs)

Varianza dello stimatore della proporzione

$$V(p_{ts}) = \frac{1-f_1}{n} \frac{\sum_{i=1}^{N} \left(P_i - \overline{P}\right)^2}{N-1} + \frac{1}{nN} \sum_{i=1}^{N} \frac{1-f_{2i}}{M_i} P_i Q_i$$

La stima è ottenuta sostituendo

$$P_i Q_i = m_i p_i q_i / (m_i - 1)$$

$$\hat{v}(p_{ts}) = \frac{1-f_1}{n} \frac{\sum_{i=1}^{n} \left(p_i - \overline{p}\right)^2}{n-1} + \frac{1}{nN} \frac{\sum_{i=1}^{n} \left(1-f_{2i}\right)}{m_i - 1} p_i q_i$$

se $f_{2i} = f_2$

$$\hat{v}(p_{ts}) = \frac{1-f_1}{n} \frac{\sum_{i=1}^{n} \left(p_i - \overline{p}\right)^2}{n-1} + \frac{1}{nN} \frac{\sum_{i=1}^{n} \left(1-f_2\right)}{m-1} p_i q_i$$

Stima dei parametri (ccs)

Nota bene:

Se le unità di primo stadio hanno numerosità variabile

$$S_{2i}^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} \left(y_{ij} - \bar{Y}_i \right)^2$$

Stima dei parametri (ccstr)

- STIMA DELLA MEDIA

Secondo caso:

Campionamento casuale stratificato per le unità primarie e casuale semplice per le unità secondarie

Stima dei parametri (ccstr)

Notazioni

Y_{hij} valore della variabile oggetto di studio nella *j*-esima unità elementare della popolazione appartenente all'*i*-esima unità primaria dello strato *h*-esimo

y_{hij} valore della variabile oggetto di studio nella *j*-esima unità elementare del campione appartenente all'*i*-esima unità primaria dello strato *h*-esimo

Stima dei parametri (ccstr)

Notazioni

N_h le unità primarie nell'h-esimo strato della popolazione

M_h le unità di secondo stadio contenute nello strato h della popolazione

$$f_{1h} = \frac{n_h}{N_h}$$ le frazioni di campionamento per ogni strato nel primo stadio

$$f_{2h} = \frac{m_h}{M_h}$$ le frazioni di campionamento per ogni strato nel secondo stadio

Stima dei parametri (ccstr)

Stima della media

si ha che $\bar{\bar{y}}_{ts} = \dfrac{\sum_{h=1}^{L} N_h M_h \bar{\bar{y}}_h}{\sum_{h=1}^{L} N_h M_h} = \sum_{h=1}^{L} W_h \bar{\bar{y}}_h$ è lo stimatore non distorto di $\bar{\bar{Y}}_{ts}$

dove

$\bar{\bar{y}}_h$ media campionaria dello strato *h*

N_h ampiezza relativa dello strato *h* nella popolazione

M_h numerosità delle unità di secondo stadio nella popolazione

$W_h = \dfrac{N_h M_h}{\sum_{h=1}^{L} N_h M_h}$ è l'ampiezza relativa dello strato *h* in termini di unità di secondo stadio ed *L* è il numero di strati

lo strato *h* contiene Nh unità di primo stadio a loro volta comprendenti Mh unità di secondo stadio

Stima dei parametri (ccstr)

Stima della media

In un campionamento a due stadi con estrazione di un campionamento casuale stratificato al primo stadio e al secondo stadio, ponendo

$$\bar{y}_{hi} = \frac{\sum_{h=1}^{n_h} \sum_{j=1}^{m_h} y_{ijh}}{m_h}$$ media campionaria nell'unità *i*-esima dello strato *h*

e

$$\bar{\bar{y}}_h = \frac{\sum_{i=1}^{n_h} \sum_{j=1}^{m_h} y_{hij}}{n_h m_h} = \frac{1}{n_h} \sum_{i=1}^{n_h} \bar{y}_{hi}$$ media campionaria nello strato *h*

Stima dei parametri (ccstr)

Stima della media

dato che possiamo considerare le $\bar{\bar{y}}_h$

come stime indipendenti della media dello strato *h*-esimo della popolazione mediante un campionamento casuale a due stadi, si avrà:

$$E_1[E_2(\bar{\bar{y}}_{ST})] = \sum_{h=1}^{L} W_h \bar{\bar{Y}}_h = \sum_{h=1}^{L} W_h \left(\frac{\sum_{i=1}^{N_h} \sum_{j=1}^{M_h} Y_{hij}}{N_h M_h} \right) = \sum_{h=1}^{L} \frac{N_h M_h}{\sum_{h=1}^{L} N_h M_h} \left(\frac{\sum_{i=1}^{N_h} \sum_{j=1}^{M_h} Y_{hij}}{N_h M_h} \right) =$$

$$= \frac{1}{\sum_{h=1}^{L} N_h M_h} \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{M_h} Y_{hij} = \bar{\bar{Y}}$$

dove L è il numero di strati

Stima dei parametri (ccstr)

Varianza dello stimatore media

se le estrazioni avvengono in modo indipendente dai diversi strati, la varianza dello stimatore della media sarà data dalla somma (pesata ovviamente per W_h^2)
delle varianze delle medie campionarie di strato; quindi:

$$V\left(\bar{\bar{y}}_{TS}\right)=\sum_{h=1}^{L}W_h^2\left[\frac{\left(1-f_{1h}\right)}{n_h}S_{1h}^2+\frac{\left(1-f_{2h}\right)}{n_h m_h}S_{2h}^2\right]$$

la varianza delle unità di primo stadio

$$S_{1h}^2=\frac{\sum_{i=1}^{N_h}\left(\bar{Y}_{hi}-\bar{\bar{Y}}_h\right)^2}{N_h-1}$$

viene stimata con

$$s_{1h}^2=\frac{\sum_{h=1}^{N_h}\left(\bar{y}_{hi}-\bar{y}_h\right)^2}{n_h-1}$$

la varianza delle unità di secondo stadio

$$S_{2h}^2=\frac{\sum_{i=1}^{N_h}\sum_{j=1}^{M_h}\left(y_{ijh}-\bar{Y}_{hi}\right)^2}{N_h\left(M_h-1\right)}$$

viene stimata con

$$s_{2h}^2=\frac{\sum\sum\left(y_{ijh}-\bar{y}_{hi}\right)^2}{n_h\left(m_h-1\right)}$$

Stima dei parametri (ccstr)

Varianza dello stimatore media

$$s_{1h}^2 = \frac{\sum_{h=1}^{N_h} \left(\bar{y}_{hi} - \bar{y}_h\right)^2}{n_h - 1}$$

è la stima della varianza delle medie delle unità di primo stadio nello strato *h*

$$s_{2h}^2 = \frac{\sum \sum \left(y_{ijh} - \bar{y}_{hi}\right)^2}{n_h \left(m_h - 1\right)}$$

è la stima della varianza fra le unità secondarie dentro le unità primarie dello strato *h*

Stima dei parametri (ccstr)

Varianza dello stimatore media

dove $$f_{1h} = \frac{n_h}{N_h}$$

$$f_{2h} = \frac{m_h}{M_h}$$

Una stima non distorta della varianza dello stimatore è pertanto:

$$\hat{v}\left(\bar{\bar{y}}_{ts}\right) = \sum_{h=1}^{L} W_h^2 \left[\frac{(1 - f_{1h})}{n_h} s_{1h}^2 + \frac{(1 - f_{2h})}{n_h m_h} s_{2h}^2 \right]$$

Stima dei parametri (ccstr)

Stima del totale

si ha che $\hat{y}_{ts} = \sum_{h=1}^{L} N_h M_h \bar{\bar{y}}_h$ è uno stimatore non distorto di $\hat{Y}_{ts}$

La varianza dello stimatore totale è:

$$V(\hat{y}_{ts}) = \left(\sum_{h=1}^{L} N_h M_h \right)^2 \left\{ \sum_{h=1}^{L} W_h^2 \left[\frac{(1 - f_{1h})}{n_h} S_{1h}^2 + \frac{(1 - f_{2h})}{n_h m_h} S_{2h}^2 \right] \right\}$$

Una stima della varianza dello stimatore totale è:

$$\hat{v}(\hat{y}_{ts}) = \left(\sum_{h=1}^{L} N_h M_h \right)^2 \left\{ \sum_{h=1}^{L} W_h^2 \left[\frac{(1 - f_{1h})}{n_h} s_{1h}^2 + \frac{(1 - f_{2h})}{n_h m_h} s_{2h}^2 \right] \right\}$$

Campionamento a due stadi e campione casuale semplice: confronti

Ipotesi sul CDS:

Piano di campionamento autoponderante quindi:

$$f_1 f_{2i} = \pi \qquad \forall i$$

$\overline{m} = \frac{\sum_{i=1}^{N} m_i}{N}$ Media di mi nelle N unità primarie e $f_{2i} = f_2 = \frac{\overline{m}}{M}$

$$V(\hat{Y}_{CDS}) = \frac{M^2}{n\overline{m}} S^2 [1 + (\overline{m} - 1)\delta] \qquad V(\overline{\overline{y}}_{CDS}) = \frac{S^2}{n\overline{m}} [1 + (\overline{m} - 1)\delta]$$

$\delta \cong \rho = 1 - \frac{S^2_W}{S^2}$ Coefficiente di omogeneità nelle unità primarie

Campionamento a due stadi e campione casuale semplice: confronti

Sotto le ipotesi iniziali si avrà:

$$deff \cong 1 + (\overline{m} - 1)\rho$$

$$\overline{m} = m_i$$

- Anche il CDS, così come il CGR, da luogo ad una perdita di efficienza rispetto al CCS qualora vi sia omogeneità nelle unità primarie

- La riduzione di efficienza è però inferiore rispetto al CGR e dipende anche da $\overline{m}$, quantità che chi costruisce il campione può determinare nel modo più conveniente.

Campionamento a tre stadi: generalizzazione

Esempio:

- Regioni (unità di primo stadio)
- Comuni (delle regioni estratte, unità di secondo stadio)
- Famiglie (dei comuni estratti, unità di terzo stadio)

Si può generalizzare la procedura del campionamento a due stadi

- La popolazione è costituita da:
  - N elementi di primo stadio (regioni)
  - M elementi di secondo stadio (ogni regione è costituita da M comuni)
  - K elementi di secondo stadio (ogni comune è formato da K famiglie)

- Il campione è costituito da:
  - n elementi di primo stadio (regioni)
  - m elementi di secondo stadio (ogni regione è costituita da m comuni)
  - k elementi di secondo stadio (ogni comune è formato da k famiglie)

Campionamento a tre stadi: generalizzazione

Y_{iju} Valore della variabile oggetto di indagine nella popolazione, misurata sull'unità di terzo stadio

y_{iju} Valore della variabile oggetto di indagine nel campione, misurata sull'unità di terzo stadio estratta dalle unità di secondo stadio estratte dalle unità di primo stadio

Le medie di interesse per le unità di terzo stadio, nella popolazione, sono:

$$\overline{Y}_{ij} = \frac{\sum_{u=1}^{K} Y_{ijk}}{K}$$

Media delle unità di terzo stadio nell'unità di secondo stadio j e di primo stadio i

$$\overline{\overline{Y}}_{i} = \frac{\sum_{j=1}^{M}\sum_{u=1}^{K} Y_{ijk}}{MK}$$

Media delle unità di terzo stadio nelle unità di secondo stadio e di primo stadio i

$$\overline{\overline{\overline{Y}}} = \frac{\sum_{i=}^{N}\sum_{j=1}^{M}\sum_{u=1}^{K} Y_{ijk}}{NMK}$$

Media delle unità di terzo stadio nelle unità di secondo stadio e di primo stadio

Il campionamento a tre stadi generalizzazione

Le varianze di interesse per le unità di terzo stadio sono:

$$S_1^2 = \frac{\sum_{i=1}^{N}\left(\overline{Y}_i - \overline{\overline{\overline{Y}}}\right)^2}{N-1}$$ Varianza nelle unità di primo stadio della popolazione

$$S_2^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\overline{Y}_{ij} - \overline{Y}_i\right)^2}{N(M-1)}$$ Varianza nelle unità di secondo stadio della popolazione

$$S_3^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{u=1}^{K}\left(\overline{Y}_{iju} - \overline{Y}_{ij}\right)^2}{NM(K-1)}$$ Varianza nelle unità di terzo stadio della popolazione

il campionamento a tre stadi generalizzazione

Estrazione di un campione casuale semplice nei tre stadi

$$\bar{y}_{ij} = \frac{\sum_{u=1}^{k} y_{ijk}}{k}$$

Media delle unità di terzo stadio nell'unità di secondo stadio j e di primo stadio i nel campione è uno stimatore non distorto di

$$\bar{Y}_{ij} = \frac{\sum_{u=1}^{K} Y_{ijk}}{K}$$

$$E\left(\bar{y}_{ij}\right) = \bar{Y}_{ij}$$

$$\bar{\bar{y}}_{i} = \frac{\sum_{j=1}^{m}\sum_{u=1}^{k} y_{ijk}}{mk}$$

Media delle unità di terzo stadio nell'unità di secondo stadio e di primo stadio i nel campione è uno stimatore non distorto di

$$\bar{\bar{Y}}_{i} = \frac{\sum_{j=1}^{M}\sum_{u=1}^{K} Y_{ijk}}{MK}$$

$$E\left(\bar{\bar{y}}_{i}\right) = \bar{\bar{Y}}_{i}$$

$$\bar{\bar{\bar{y}}} = \frac{\sum_{i=}^{n}\sum_{j=1}^{m}\sum_{u=1}^{k} y_{ijk}}{nmk}$$

Media delle unità di terzo stadio nell'unità di secondo stadio e di primo stadio nel campione è uno stimatore non distorto di

$$\bar{\bar{\bar{Y}}} = \frac{\sum_{i=}^{N}\sum_{j=1}^{M}\sum_{u=1}^{K} Y_{ijk}}{NMK}$$

$$E\left(\bar{\bar{\bar{y}}}\right) = \bar{\bar{\bar{Y}}}$$

Il campionamento a tre stadi generalizzazione

Estrazione di un campione casuale semplice nei tre stadi

La varianza dello stimatore media $\bar{\bar{\bar{y}}}$ è:

$$V\left(\bar{\bar{\bar{y}}}\right) = \frac{1-f_1}{n} S_1^2 + \frac{1-f_2}{nm} S_2^2 + \frac{1-f_3}{nmk} S_3^2$$

dove:

$$f_1 = \frac{n}{N} \qquad f_2 = \frac{m}{M} \qquad f_3 = \frac{k}{K}$$

Una stima della varianza dello stimatore è:

$$\hat{v}\left(\bar{\bar{\bar{y}}}\right) = \frac{1-f_1}{n} s_1^2 + \frac{1-f_2}{nm} s_2^2 + \frac{1-f_3}{nmk} s_3^2$$

il campionamento a tre stadi generalizzazione

con:

$$s_1^2 = \frac{\sum_{i=1}^{n}\left(\bar{y}_i - \bar{\bar{\bar{y}}}\right)^2}{n-1}$$ Varianza nelle unità di primo stadio del campione

$$s_2^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\bar{y}_{ij} - \bar{y}_i\right)^2}{n(m-1)}$$ Varianza nelle unità di secondo stadio del campione

$$s_3^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{u=1}^{k}\left(\bar{y}_{iju} - \bar{y}_{ij}\right)^2}{nm(k-1)}$$ Varianza nelle unità di terzo stadio del campione

Alcune considerazioni sul CDS

- In teoria, in ogni stadio le unità di campionamento potrebbero essere raggruppate in strati da cui campionare separatamente;
- la stratificazione delle unità primarie
- il grappolo nelle unità primarie

sono le pratiche più ricorrente nelle indagini reali.

Motivazioni:

1. Abbatte la varianza delle stime in misura maggiore rispetto alla stratificazione delle u.e. o delle unità successive al 1° stadio.
2. È meno costoso stratificare grappoli di unità elementari anziché le unità elementari stesse, anche in considerazione delle maggiori informazioni a priori disponibili.

N.B. Se le unità primarie non sono di grandi dimensioni, è raro che si ricorra alla stratificazione delle unità di campionamento al loro interno; anche perché è possibile sfruttare la stratificazione implicita con un opportuno campionamento sistematico.

Campionamento a due stadi con probabilità variabili

Viene utilizzato se:

- Gli M_i sono diversi per ogni *i*
- Vi è proporzionalità fra Y (= variabile di dimensione o proxy di quella da investigare) e M

Si assegna ad ogni campione di primo stadio (grappolo) una probabilità di estrazione $P_i = M_i / M_0$

Si devono conoscere a priori le numerosità di tutti i grappoli

Campionamento a due stadi con probabilità variabili

Ipotesi:

estrazione senza ripetizione

- probabilità variabili sia delle unità primarie sia delle unità secondarie

Lo stimatore del totale $\hat{Y}_{TS}$ sarà:

$$\hat{y}_{TS} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{y_{ij}}{\pi_i \pi_{j/i}}$$

dove

π_i è la probabilità di inclusione della i-esima unità primaria e

$\pi_{j/i}$ è la probabilità di inclusione della j-esima unità secondaria dopo aver estratto l'unità primaria i

Campionamento a due stadi a probabilità variabili

Indichiamo con

$$\hat{Y}_i = \sum_{j=1}^{m_i} \frac{Y_{ij}}{\pi_{j/i}}$$

lo stimatore di secondo stadio del totale dell'i-esima unità. Allora lo stimatore del totale sarà:

$$\hat{Y}_{TS} = \sum_{i=1}^{n} \frac{\hat{Y}_i}{\pi_i}$$

stimatore di Horwitz e Thompson

Nel caso di probabilità costante si avrà: $\pi_i = \frac{n}{N} = f_1 \quad \pi_{j/i} = \frac{m_i}{M_i} = \frac{m}{M} = f_{2i}$

Campionamento a due stadi a probabilità variabili

La varianza dello stimatore del totale sarà:

$$V(\hat{Y}_{TS}) = V_1\left(\sum_{i=1}^{n} \frac{\sum_{j=1}^{m_i} y_{ij}}{\pi_i}\right) + \sum_{i=1}^{N} \frac{V_2\left(\hat{Y}_i\right)}{\pi_i}$$

Dove:

il primo termine a secondo membro è la varianza dello stimatore HT del totale della popolazione nel campione delle unità primarie

il secondo membro è la varianza di secondo stadio dello stimatore del totale dell'unità primaria i

La dimostrazione deriva dalla:

$$V\left(\hat{Y}_{TS}\right) = V_1\left[E_2\left(\hat{Y}_{TS}\right)\right] + E_1\left[V_2\left(\hat{Y}_{TS}\right)\right]$$

Campionamento a due stadi a probabilità variabili

$$E_2\left(\hat{Y}_{TS}\right)=E_2\left(\sum_{i=1}^{n}\frac{\hat{Y}_i}{\pi_i}\right)=\sum_{i=1}^{n}\frac{1}{\pi_i}\left[E\left(\hat{Y}_{TS}\right)\right]=\sum_{i=1}^{n}\frac{\sum_{j=1}^{m_i}y_{ij}}{\pi_i}$$

e

$$V_2\left(\hat{Y}_{TS}\right)=V_2\left(\sum_{i=1}^{n}\frac{\sum_{j=1}^{m_i}y_{ij}}{\pi_i}\right)=\sum_{i=1}^{n}\frac{1}{\pi_i^2}\left[V_2\left(\hat{Y}_{TS}\right)\right]$$

E_2 e V_2 sono gli operatori di valore atteso e varianza dei dati campionari nelle unità elementari

Campionamento a due stadi a probabilità variabili

sostituendo:

$$V\left(\hat{Y}_{TS}\right) = V_1\left[\sum_{i=1}^{n} \frac{\hat{Y}_i}{\pi_i}\right] + E_1\left[\sum_{i=1}^{n} \frac{1}{\pi_i^2} V_2\left(\hat{Y}_i\right)\right]$$

con

$$E_1\left[\sum_{i=1}^{n} \frac{1}{\pi_i^2} V_2\left(\hat{Y}_i\right)\right] = \sum_{i=1}^{N} \frac{1}{\pi_i^2} V_2\left(\hat{Y}_i\right)\pi_i$$

Campionamento a due stadi a probabilità variabili

si dimostra che una stima non distorta della varianza del totale è:

$$\hat{v}\left(\hat{Y}_{TS}\right) = \hat{v}_1\left[\sum_{i=1}^{n}\frac{\hat{Y}_i}{\pi_i}\right] + \sum_{i=1}^{n}\frac{\hat{v}_2\left(\hat{Y}_i\right)}{\pi_i}$$

con

$$\hat{v}_1\left[\sum_{i=1}^{n}\frac{\hat{Y}_i}{\pi_i}\right] = \sum_{i=1}^{n}\frac{1-\pi_i}{\pi_i^2}\hat{Y}_i^2 + \sum_{i=1}^{n}\sum_{i'\neq i}^{n}\left(\frac{1}{\pi_i\pi_{ii'}} - \frac{1}{\pi_{ii'}}\right)\hat{Y}_i\hat{Y}_{i'}$$

$$\hat{v}_2\left(\hat{Y}_i\right) = \sum_{i=1}^{n}\frac{1}{\pi_i}\hat{v}_2\left(\hat{Y}_i\right)$$

Esempio

- Si vuole stimare il numero medio di figli in Italia. La popolazione è suddivisa in aree geografiche. Si effettua un campionamento a due stadi:
  - Al primo stadio si effettua una stratificazione per aree geografiche dalle quali si estraggono n_h comuni con $n=750$.
  - Al secondo stadio si procede all’estrazione stratificata di 1540 famiglie sulla base del reddito.

Esempio

- Dati:

Aree geografiche	Nh	Nh/N	Comuni	Famiglia Mh	Famiglia mh
NO 1	257	0,132	99	1258	185
NO 2	268	0,138	103	2310	254
NE 1	159	0,082	61	541	100
NE 2	251	0,129	97	1000	142
C 1	120	0,062	46	236	58
C 2	154	0,079	59	562	120
C 3	180	0,093	69	758	145
S 1	200	0,103	77	954	210
S 2	210	0,108	81	1240	147
I 1	90	0,046	35	124	90
I 2	56	0,029	22	100	89
	1945	**1,000**	**750**	**9083**	**1540**

Esempio

- formule:

$$\bar{\bar{y}}_{ST} = \frac{\sum_h N_h M_h \bar{y}_h}{\sum_h N_h M_h} = \sum_h W_h \bar{\bar{y}}_h$$

$$v(\bar{\bar{y}}_{ST}) = \sum_h W_h^2 \left[\frac{1 - f_{1h}}{n_h} s_{1h}^2 + \frac{f_{1h}(1 - f_{2h})}{n_h m_h} s_{2h}^2 \right]$$

Esempio

- Dati aggiuntivi calcolati sui dati campionari:

s1h	s2h	n°medio di figli
1,26	1,12	2,51
1,13	1,28	1,84
1,23	1,25	1,8
1,08	1,98	2
1,34	1,78	1,89
1,69	1,56	2,7
1,87	1,54	1,8
1,23	1,54	3
1,21	1,34	3,5
1,02	1,39	2,9
1,32	1,57	2,7

Esempio

- Dati aggiuntivi calcolati sui dati campionari per il calcolo dello stimatore media:

Wh	Nh×Mh	wh×yh medio
0,162	323306	0,406
0,310	619080	0,570
0,043	86019	0,077
0,126	251000	0,251
0,014	28320	0,027
0,043	86548	0,117
0,068	136440	0,123
0,095	190800	0,286
0,130	260400	0,456
0,006	11160	0,016
0,003	5600	0,008
SOMMA	1998673	**2,337**

Esempio

- Dati aggiuntivi calcolati sui dati campionari per il calcolo della varianza della media:

f1h	f2h	wh^2	1-fih/nh*s1h	f1h(1-f2h)	nh*mh	(f1h(1-f2h)/nh*mh)*s2h	Wh^2(1-f1h/nh)s1h)+(f1h(1-f2h)/nh*mh)*s2h
0,386	0,147	0,026	0,008	0,329	18333,548	0,000020092	0,000204929
0,386	0,110	0,096	0,007	0,343	26248,843	0,000016736	0,000646161
0,386	0,185	0,002	0,012	0,314	6131,105	0,000064085	2,29494E-05
0,386	0,142	0,016	0,007	0,331	13743,702	0,000047664	0,000108875
0,386	0,246	0,000	0,018	0,291	2683,805	0,000192894	3,61091E-06
0,386	0,214	0,002	0,017	0,303	7125,964	0,000066391	3,29116E-05
0,386	0,191	0,005	0,017	0,312	10064,267	0,000047717	7,73617E-05
0,386	0,220	0,009	0,010	0,301	16195,373	0,000028595	8,95613E-05
0,386	0,119	0,017	0,009	0,340	11903,599	0,000038262	0,000156487
0,386	0,726	0,000	0,018	0,106	3123,393	0,000047053	5,64468E-07
0,386	0,890	0,000	0,038	0,042	1921,851	0,000034651	2,95111E-07
		0,173					**0,001343707**