Il campionamento sistematico

syst

Il campionamento sistematico

Generalità

- Consiste nella scelta casuale di una unità tra le N che formano la popolazione e nella selezione successiva delle altre n-1 unità, secondo un criterio predefinito.
  - **vantaggi**:
    - rapido e semplice procedimento di selezione delle unità, anche se non si dispone di un elenco dettagliato delle unità che costituiscono la popolazione
    - nelle indagini ripetute nel tempo (panel)
    - quando le unità campionarie possono essere selezionate sul luogo in cui vengono effettuate le rilevazioni dei dati
    - flessibile alle esigenze di rilevazione
    - può condurre alla formazione di un campionamento 'proporzionato[1] se traiamo vantaggio dall'ordinamento della popolazione

[1] Il campione proporzionato è quel campione che riproduce la stessa proporzionalità nelle unità presenti nella popolazione

Il campionamento sistematico

Generalità

- Esempi tipici di campionamento sistematico si riferiscono ad unità statistiche come:
  - le abitazioni (una abitazione ogni k all'interno di un blocco di abitazioni)
  - individui da estrarre da un elenco nominativo
- È un campionamento ideale su dati di tipo amministrativo-burocratico perché le liste sono il risultato dell'arrivo della 'pratica' che si presume rispetti la casualità.
- Ambiti di applicazione:
  - indagini di tipo assicurativo, quando le unità devono essere selezionate in base all'arrivo presso supermercati (negozi in genere, musei, porti, banchi di accettazione all'aeroporto, alle stazioni, ecc..), o ancora gli ingressi in ospedale, ecc.

Il campionamento sistematico

Generalità

- Per poter effettuare un campionamento sistematico occorre disporre delle N unità della popolazione numerate da 1 a N secondo un ordine specifico.

- Il campionamento casuale sistematico potrebbe essere definito un **campionamento pseudo casuale** perché la selezione casuale delle unità è rispettata solo per la prima, mentre il campione è determinato a partire da essa.

La casualità è garantita per la prima unità estratta e non per tutto il campione.

Il campionamento sistematico

Il processo di selezione nel campionamento casuale sistematico

- si supponga che le N unità della popolazione P siano numerate da 1 a N e presenti nella lista con un certo ordine.
- Si suppnga di volere estrarre un campione di dimensioni n dalla P
  - dobbiamo determinare **il passo di campionamento**

$$k = \frac{\mathrm{N}}{\mathrm{n}}$$

  - si procede estraendo un numero casuale fra 1 e k secondo uno dei criteri citati nell'ambito del piano di campionamento casuale semplice.

Il campionamento sistematico

Il processo di selezione nel campionamento casuale sistematico

- Se il numero casuale è
$$r \leq k$$
- si definisce il **campionamento casuale sistematico** quel *piano che permette di assegnare una probabilità costante ai campioni determinati a partire da un numero casuale iniziale, indicato con r.*
- Se r è il numero casuale scelto e k è il passo di campionamento, il campione, di dimensione n, sarà il seguente sottinsieme:

$$s = \{r, r+k, r+2k, ..., r+(n-1)k\}$$

1 2 3 n

Il campionamento sistematico

Esempio

- Si supponga una popolazione costituita da 200 unità etichettate e ordinate secondo un criterio stabilito a priori:

  N=200 (1, 2, 3, 4……,55, ……100, …… 200)

- Si supponga, inoltre, di volere estrarre, in modo sistematico, un campione di dimensione n=50.
- Il passo di campionamento è quindi:

$$k = \frac{\mathrm{N}}{\mathrm{n}} = \frac{200}{50} = 4$$

- Si procede estraendo un numero casuale compreso fra 1 e 4.
  - Si supponga che r=2.
  - Il campione sarà costituito dalle unità identificate dalle etichette:

$$\{2; 2+4; 2+2\times 4; .....; 2+(50-1)\times 4\} = \{2; 6; 10; .....; 198\}$$

Il campionamento sistematico

Esempio

- il passo di campionamento k permette una suddivisione della popolazione in n zone costituite, ciascuna da k unità:

$$\{|1,2,\ldots,k|\ |k+1,k+2,\ldots,k+k|.|.\ldots.|.|.\ldots\ldots,N|\}$$

- il passo di campionamento k=4, permette di suddividere la popolazione in 50 parti, ciascuna costituita da 4 elementi.

$$|5,6,7,8|.|.\ldots.|.|197,198,199,200|$$

L'ordine di selezione riflette l'ordine con cui sono numerati gli elementi nella popolazione. La **frazione di selezione** indica la distanza fra gli elementi inclusi nel campione ed è pari a:

$$F = k = \frac{\mathrm{N}}{\mathrm{n}}$$ **coincide con il passo di campionamento**

$$f = \frac{1}{\mathrm{F}}$$

Il campionamento sistematico

Spazio campionario e probabilità di inclusione

Poiché ogni r, primo e unico numero casuale, è estratto a caso da 1 a k, ogni unità ha la stessa probabilità di essere estratta pari a:

$$\frac{1}{k}$$

Date le premesse, il campionamento sistematico può essere considerato:

– **un campionamento casuale stratificato** con n strati in cui viene scelta una unità per strato;

– **un campionamento casuale con un unico grappolo** (k grappoli ciascuno con n unità);

–più semplicemente **un campionamento casuale semplice** con estrazione casuale della prima unità.

Il campionamento sistematico

Spazio campionario e probabilità di inclusione

Se estraiamo tutti i possibili campioni sistematici con passo k variando r (1,2,….,k), si può considerare la seguente tabella:

	Numero casuale	Strati							Grappoli
		1	2	3	..	..	..	n	
primo campione sistematico	r=1	1	1+k	2k	..	..	..	n	1
secondo campione "	r=2	2	2+k	2+2k	..	..	..	2n	2
.....	...	..	..	..	..	..	..	..	..
.....	...	..	..	..	..	..	..	..	..
k-esimo campione "	r=k	k	2k	3k	..	..	..	kn	k

Ogni campione (grappolo) ha la stessa probabilità di essere estratto pari a: $p(s)=\frac{1}{k}$

Lo **spazio campionario**, costituito da tutti i possibili campioni che si possono estrarre, è costituito da $k=\frac{N}{n}$ campioni diversi.

Il campionamento sistematico

Spazio campionario e probabilità di inclusione

Le probabilità di inclusione del primo e del secondo ordine vengono dedotte dalle probabilità corrispondenti del campionamento casuale semplice o a grappoli, con un unico grappolo.

Probabilità di inclusione del primo ordine

La probabilità di inclusione del primo ordine dell'unità i nel campione k è:

$$\pi_i = \frac{n}{N} = \frac{1}{k}$$

Probabilità di inclusione del secondo ordine

La probabilità di inclusione del secondo ordine, coincide col probabilità di una coppia (i e j) nello stesso grappolo (k) ovvero:

$$\pi_{(ij)(k)} = \frac{n}{N} = \frac{1}{k}$$

Se le due unità appartengono a due campioni diversi, la probabilità di inclusione del secondo ordine è pari a 0.

Il campionamento sistematico

Problematiche relative alla scelta del passo di campionamento

Fino a questo momento, abbiamo ipotizzato che k sia multiplo di N, ovvero

$$N = n \times k$$

Nel caso in cui, invece, **k** non è un intero, non tutti gli elementi della popolazione hanno una probabilità di essere estratti.

Se $N \neq n \times k$???

Primo problema

Si supponga che la popolazione sia composta da 15 unità, etichettate e ordinate da 1 a 15 e che la dimensione campionaria sia pari a 4.

Il passo di campionamento sarà pari a:

$$k = \frac{N}{n} = \frac{15}{4} = 3{,}75 \cong 4$$

Se il numero casuale estratto è tale che:

$$r = k = 4$$

Il campionamento sistematico

Problematiche relative alla scelta del passo di campionamento

Se consideriamo la popolazione suddivisa in n=4 porzioni, ognuna costituita da k elementi, si avrà

$$P = \{|1,2,3,4| \ |5,6,7,8| \ |9,10,11,12| \ |13,14,15|\}$$

e il numero casuale estratto è 4 ….. il campione selezionato sarà: $s = \{|4| \ |8| \ |12| \ |?|\}$

Ci sarebbe il problema dell'unità mancante nell'ultima porzione di popolazione

Secondo problema
il valore del passo di campionamento k è fissato a priori indipendentemente dalla numerosità della popolazione o del campione (caso frequente quando non si conosce la numerosità della popolazione).

Si ipotizza che k sia pari a 4 e la dimensione campionaria pari a 3. Se la popolazione rimane la stessa dell'esempio precedente:

$$P = \{|1,2,3,4| \ |5,6,7,8| \ |9,10,11,12| \ |13,14,15|\}$$

Il campionamento sistematico

Problematiche relative alla scelta del passo di campionamento

e il numero casuale estratto è

$$r = 2$$

il campione selezionato sarà:

$$s = \{2 \mid 6 \mid 10\}$$

dal quale, come si può facilmente vedere, vengono escluse, a priori, le unità

$$P = \{1,2,3,4 \mid 5,6,7,8 \mid 9,10,11,12 \mid \sout{13,14,15}\}$$

Tale problema determina la mancanza del rispetto della condizione fondamentale per estrarre un campione probabilistico, **ovvero che sia nota e diversa da 0 la probabilità che ogni elemento della popolazione abbia di far parte del campione**.

Gli elementi 13, 14 e 15 della popolazione hanno probabilità pari a 0 di essere estratti.

Il campionamento sistematico

Possibili soluzioni – 1 –

Una *possibile soluzione* consiste nel consentire la variabilità della dimensione campionaria fra n e n+1, ovvero si può scegliere k in modo tale che il numero casuale estratto è

$$nk < N < (n+1)k$$ cioè scegli k in modo che si verifica questa condizione

In questo modo, la scelta di k e la determinazione casuale di r determinano una dimensione campionaria di n o n+1.

Esempio: si supponga che la popolazione sia costituita da 7 elementi mentre la dimensione campionaria n sia, per il momento, fissata a 2.
In questo caso k sarà:

$$k = \frac{N}{n} = \frac{7}{2} = 3{,}5$$

Se si sceglie k'=3 allora è verificata la condizione per cui

$$nk < N < (n+1)k$$

$$(2\times 3) < 7 < (3\times 3) \text{ quindi } 6 < 7 < 9$$

Il campionamento sistematico

Possibili soluzioni – 1 –

Se r=2 allora la dimensione campionaria sarà pari a 2. Infatti la popolazione sarà costituita da:

$$P = \{|1,2,3| \ |4,5,6,| \ |7|\}$$

e il campione costituito dagli elementi:

$$s = \{|2| \ |5|\}$$

Se r=1 allora la dimensione campionaria sarà pari a 3. Infatti dalla popolazione, si estraggono gli elementi tali che il campione sia costituito da:

$$s = \{|1| \ |4| \ |7|\}$$

l'esclusione dell'elemento 7 nella selezione del campione nel primo esempio è dovuta al caso, ovvero alla selezione casuale di r.

Il campionamento sistematico

Possibili soluzioni – 2 –

Un'altra possibile soluzione consiste nell'eliminare, secondo un procedimento casuale, tante unità fino a ridurre la lista della popolazione in modo tale che

$$N = n \times k$$

elimina unità dalla popolazione così che N/n = k sia un intero

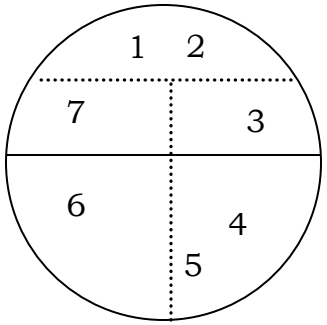
Ciò significa che prima di estrarre le unità da includere nel campione ***si estraggono casualmente*** le unità da eliminare nella popolazione.

in modo casuale

Possibili soluzioni – 3 –

Un'ulteriore soluzione potrebbe essere quella di considerare la lista circolare, in modo che dopo l'ultima unità si ricominci dalla prima.

Nella popolazione costituita da 7 elementi la lista potrebbe essere considerata nel seguente modo:



cioè se sfori ricominci dalla prima unità

Il campionamento sistematico

Possibili soluzioni – 4 –

Si può risolvere il problema usando la tecnica degli intervalli frazionali.
Tecnica che permette di generare numeri casuali in modo da garantire a tutti gli elementi della popolazione una probabilità diversa da 0 di essere inclusa nel campione.

Si supponga di dover estrarre un campione di 100 elementi da una popolazione di 920 unità. In questo caso:

$$k = \frac{920}{100} = 9{,}2$$

Si procede ipotizzando $k = 92$ e si sceglie a caso un numero r compreso fra 1 e 92.
Se, il numero casuale estratto, ad esempio, è pari a 4, il campione sarà identificato dalle seguenti unità:

$$s' = \{|4|\ |4+92|\ |4+2\times 92|.....\}$$

cioè

$$s' = \{|4|\ |96|\ |188|.....\}$$

non ho capito XD

Si procede eliminando l'ultima cifra dell'etichetta individuata, ottenendo così il campione:

$$s' = \{|4|\ |9|\ |18|.....\}$$

Il campionamento sistematico

Problematiche relative alla struttura della popolazione

La selezione sistematica delle unità può essere "pericolosa" in termini di assenza di casualità, anche in funzione della struttura della popolazione.

a) Una difficoltà che si può presentare nella selezione sistematica delle unità si ha quando nella popolazione vi sono ***informazioni mancanti***
Dal punto di vista statistico è un problema di incompletezza della lista, ovvero di una grande differenza fra popolazione su cui fare campionamento e popolazione effettivamente indagata.
Può succedere, infatti che

$$n \times k = N$$

cioè la lista è incompleta, N = M + B dove B indica la parte mancante e M quella conosciuta, se i B sono distribuiti casualmente allora non è un problema

ma la popolazione è costituita da alcuni dati mancanti: $N = M + B$

M rappresenta la parte della popolazione i cui elementi presentano caratteristiche misurabili
B è la parte della popolazione con dati mancanti (*blank*).
Se i B sono distribuiti casualmente nella popolazione dicotoma (dati mancanti e non), si tratta, per il campionamento, di scegliere un valore m tale che

$$m \times k = M$$

m indica il campione "misurabile"

ovvero circoscrivere ai dati 'misurabili' l'operazione di campionamento.
Se si può procedere in questo modo, si hanno m unità con probabilità di selezione pari a $\frac{1}{k}$

Il campionamento sistematico

Problematiche relative alla struttura della popolazione

m è una variabile casuale con media e errore standard dati rispettivamente da:

$$E(m) = \frac{M}{K} \qquad SE(m) = \sqrt{n\overline{M}\left(1-\overline{M}\right)}$$

dove $\overline{M} = \frac{M}{N}$

La dimensione campionaria m* sarà in funzione di M ovvero $k = \frac{M}{m^*}$

M deve essere nota ed i dati mancanti B devono essere distribuiti casualmente, altrimenti non si applica il sistematico

Condizione necessaria per seguire tale procedimento è che ***sia nota M e che B sia distribuita casualmente***. ***Se non è possibile fare assunzioni di questo tipo, il campionamento casuale sistematico, non è un "buon" piano di campionamento***.

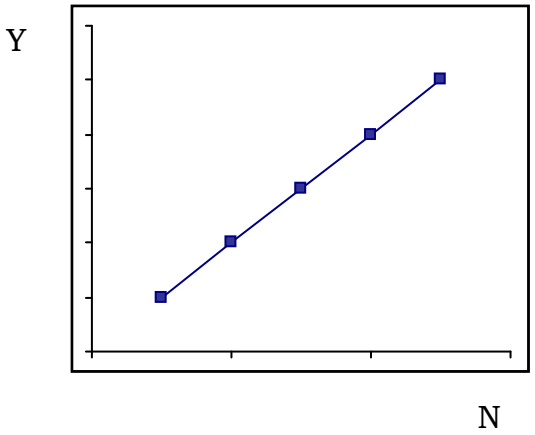
Il campionamento sistematico

Problematiche relative alla struttura della popolazione

b) popolazione con trend lineare

Se l'ordinamento delle unità della popolazione è casuale, il campionamento sistematico si riconduce al campionamento casuale semplice e di solito viene preferito per la semplicità di estrazione delle unità. Se le unità della popolazione vengono elencate in modo da rispettare un ordine (crescente o decrescente) della variabile di interesse Y, ovvero in presenza di un trend lineare, in genere il campionamento sistematico è da preferire ad altri tipi di campionamento.

La popolazione risulta ordinata secondo il grafico seguente (trend monotono):



se la popolazione è ordinata in modo casuale allora il sistematico equivale al ccs ma più semplice. Se invece, le liste seguono un qualche ordine, per esempio in ordine crescente di qualche variabile Y (presenza di trend lineari), in genere si preferisce il campionamento sistematico (per popolazione ordinata).

Il campionamento sistematico

Problematiche relative alla struttura della popolazione

- Il problema sussiste quando, nel caso più comune, non è nota la variabile di interesse Y

    - si sceglie l'ordinamento in funzione di una variabile ausiliaria che si ipotizza fortemente correlata con quella di interesse (incognita).

spesso non conosciamo Y (è la variabile di interesse dopotutto xdd), quindi usiamo una variabile fortemente correlata con essa (una proxy)

Esempio di popolazione ordinata:

-***i dirigenti d'azienda in funzione della loro anzianità di servizio***
(se l'obiettivo dell'indagine campionaria è il reddito medio dei dirigenti. In questo caso, infatti, vengono inclusi nel campione i dirigenti con bassa, media e alta anzianità che si presume abbiano livelli medi di reddito diversi fra loro).

obiettivo: reddito medio dirigenti, quindi ordiniamo la lista per età (proxy del reddito medio) e si applica il campionamento sistematico.

Una scelta errata della variabile ausiliaria, però, potrebbe causare un allentamento dalla casualità. Se per lo stesso obiettivo, ***si ordinano i dirigenti in base all'età***, non è necessariamente detto che dirigenti giovani guadagnino meno dei dirigenti anziani.

attenzione però alla scelta della variabile ausiliaria (proxy)

Il campionamento sistematico

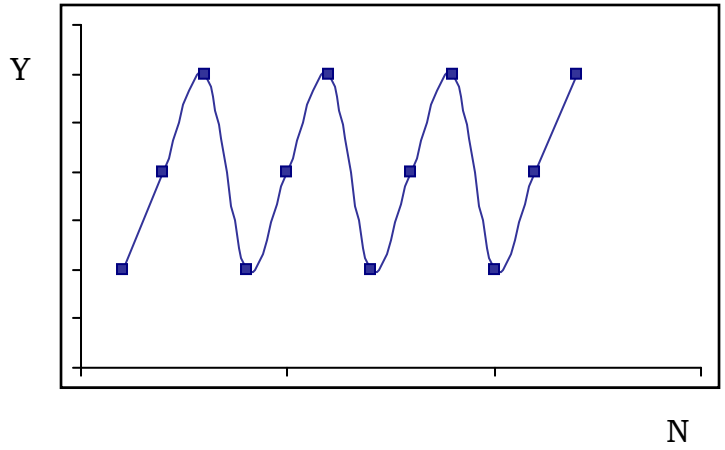
Problematiche relative alla struttura della popolazione

b) popolazione con trend periodico

Se nella popolazione è possibile ipotizzare che la variabile di interesse Y segua una fluttuazione periodica (tutti i casi in cui il fenomeno presenta una certa stagionalità), il campionamento sistematico può essere pericoloso, perché anche in questo caso ci si potrebbe allontanare dalla casualità.

trend lineare = good, trend periodico = bad, perché? si rischia di allontanarsi dalla casualità, selezionando per esempio elementi che si trovano sempre in cima all'onda, od in fondo od al centro.

La popolazione può essere rappresentata secondo il grafico seguente (fluttuazioni periodiche):



Il campionamento sistematico

Problematiche relative alla struttura della popolazione

•Il trend periodico rischia di far selezionare, nel campione, elementi che si trovano sempre al vertice o sempre al punto medio dell'onda

•È conveniente, invece, ricorrere al campionamento sistematico, in questo caso, se è noto il periodo del trend. Infatti ***scegliendo k come un multiplo dispari del semiperiodo***, si è sicuri che le unità sono scelte in tutti i punti dell'onda.

per rimediare a questo, se conosciamo il semiperiodo, possiamo rimediare selezionando k come multiplo dispari del semi periodo

possibili soluzioni per evitare che la popolazione presenti un trend periodico di cui non si conosce il periodo:

soluzioni per rimediare ad una popolazione con trend periodico

- ***casualizzare la lista della popolazione con una procedura di rimescolamento***, anche se si corre il rischio di non tenere conto di qualche tipo di stratificazione naturale della popolazione che invece viene considerata se il trend è lineare rimescola la lista
- **cambiare il valore di r parecchie volte**. Si procede infatti estraendo le prime x unità in funzione del primo numero casuale r, si estraggono le successive x unità a partire da un r' diverso da r e così via fino ad arrivare ad un campione di dimensione n. Nel caso di trend periodico esso viene "distrutto" con diverse spezzate con diverse pendenze scegli un r come numero casuale, estrai un certo numero di unità, e poi cambia r
- ***replicare le selezioni*** e determinare C differenti campioni (soluzione dispendiosa)

Il campionamento sistematico

Stima dei parametri

Anche per il campionamento sistematico si affrontano le problematiche della stima dei parametri *media, totale e proporzione* e delle loro varianze. Non si affronterà, invece, l'argomento della determinazione della dimensione campionaria perché sostanzialmente coincidente con quanto già descritto sul campionamento casuale semplice.

ampiezza campionaria per campionamento sistematico = come il ccs

Media

La media della variabile di interesse Y, nel campionamento sistematico è pari a:

$$\bar{y}_{SI} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Essa è una stima della media della popolazione pari a:

$$\bar{Y}_{SI} = \frac{\sum_{i=1}^{N} Y_i}{N}$$

Si dimostra che quando $N = n \times k$

$E(\bar{y}_{SI}) = \bar{Y}_{SI}$ ***la media campionaria è una stima non distorta della media della popolazione***

quindi solo quando k è intero la media campionaria è una stima non distorta

Il campionamento sistematico

Varianza della media

Per quanto riguarda la varianza dello stimatore vale il seguente teorema:

Teorema 1:

$$V(\bar{y}_{SI}) = \frac{N-1}{N} S^2 - \frac{k(n\text{-}1)}{N} S_w^2$$

dove

$$S_w^2 = \frac{1}{k(n\text{-}1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2$$

varianza all'interno (within) di ogni campione sistematico

esprime la varianza fra le unità all'interno dello stesso campione sistematico.

y_{ij} rappresenta il j-esimo elemento dell'i-esimo campione sistematico

$\bar{y}_{i.}$ indica la media dell'i-esimo campione sistematico

Il denominatore della varianza interna fra le unità in ogni campione sistematico, indica che ogni campione dei k possibili, contribuisce, con n-1 gradi di libertà, alla somma dei quadrati.

Il campionamento sistematico

Varianza della media

Dimostrazione:

Dall'abituale scomposizione della varianza possiamo scrivere:

$$(N-1)S^2 = \sum_{j=1}^{n}\sum_{i=1}^{k}(y_{ij}-\bar{Y})^2 = \sum_{j=1}^{n}\sum_{i=1}^{k}(y_{ij}-\bar{y}_{i.}+\bar{y}_{i.}-\bar{Y})^2 = \sum_{j=1}^{n}\sum_{i=1}^{k}\left[(y_{ij}-\bar{y}_{i.})+(\bar{y}_{i.}-\bar{Y})\right]^2 =$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{k}\left[(y_{ij}-\bar{y}_{i.})^2+(\bar{y}_{i.}-\bar{Y})^2+2(y_{ij}-\bar{y}_{i.})(\bar{y}_{i.}-\bar{Y})\right] = n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{Y})^2 + \sum_{j=1}^{n}\sum_{i=1}^{k}(y_{ij}-\bar{y}_{i.})^2$$

varianza nei campioni (within)

varianza fra campioni (between)

= 0 per la prima proprietà della media campionaria

il primo membro dopo l'uguaglianza è l'usuale varianza fra i campioni
Il secondo è la varianza entro i campioni

la varianza dello stimatore media, per definizione, si può scrivere anche nella seguente maniera:

$$V(\bar{y}_{SI}) = \frac{\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{Y})^2}{k}$$

var_between = var tot - var_within

$$(N-1)S^2 = nkV(\bar{y}_{SI}) + k(n-1)S_w^2$$

var_tot var_within

Quindi ricordando che $N = n \times k$

$$V(\bar{y}_{SI}) = \frac{(N-1)}{N}S^2 - \frac{k(n-1)}{N}S_w^2$$

Il campionamento sistematico

Varianza della media

Un altro modo per definire la varianza dello stimatore viene dal seguente teorema:

Teorema 2:

$$V(\bar{y}_{SI}) = \frac{S^2}{n}\frac{N-1}{N}\left[1+(n-1)\rho_w\right]$$

dove ρ_w è il ***coefficiente di correlazione lineare fra le coppie di unità che appartengono allo stesso campione sistematico*** ed è definito come:

$$\rho_w = \frac{E\left(y_{ij}-\bar{Y}\right)\left(y_{iu}-\bar{Y}\right)}{E\left(y_{ij}-\bar{Y}\right)^2}$$

il numeratore è calcolato su tutte le coppie distinte di elementi ed il denominatore su tutti gli N elementi di y_{ij} ed è pari a $\frac{N-1}{N}S^2$

il numero di tutte le coppie distinte di elementi per i k campioni di n unità è dato da: $\binom{n}{2}k = \frac{kn(n-1)}{2}$

per cui ρ_w è pari a: $\rho_w = \frac{2\sum_{i=1}^{k}\sum_{i<u}\left(y_{ij}-\bar{Y}\right)\left(y_{iu}-\bar{Y}\right)N}{kn(n-1)S^2(N-1)} = \frac{2}{(n-1)(N-1)S^2}\sum_{i=1}^{k}\sum_{i<u}\left(y_{ij}-\bar{Y}\right)\left(y_{iu}-\bar{Y}\right)$

Il campionamento sistematico

Varianza della media

Dimostrazione:

$$V(\bar{y}_{SI}) = \frac{\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{Y})^2}{k} \Rightarrow kV(\bar{y}_{SI}) = \sum_{i=1}^{k}(\bar{y}_{i.} - \bar{Y})^2$$

moltiplico tutto per n^2 e ottengo:

$$n^2 kV(\bar{y}_{SI}) = n^2 \sum_{i=1}^{k}(\bar{y}_{i.} - \bar{Y})^2 = n^2 \sum_{i=1}^{k}\left[(\bar{y}_{i1} - \bar{Y}) + (\bar{y}_{i2} - \bar{Y}) + ... + (\bar{y}_{in} - \bar{Y})\right]^2 =$$

ricordando che $E(y_{ij} - \bar{Y})^2 = \frac{N-1}{N}S^2$

i prodotti al quadrato e poi sommati diventano $(N-1)S^2$

ricordando che $\rho_w = \frac{2}{(n-1)(N-1)S^2}\sum_{i=1}^{k}\sum_{i<u}(y_{ij} - \bar{Y})(y_{iu} - \bar{Y})$

i doppi prodotti sommati sono pari a: $(n-1)(N-1)S^2\rho_w$

Il campionamento sistematico

Varianza della media

Dimostrazione:

per cui possiamo scrivere

$$n^2 \mathrm{k} V\left(\bar{y}_{SI}\right) = (N-1)S^2 - (n-1)(N-1)S^2\rho_w$$

vale che: $n^2 k = nnk = nN$

quindi:

$$n\mathrm{N} V\left(\bar{y}_{SI}\right) = (N-1)S^2 - (n-1)(N-1)S^2\rho_w$$

$$V\left(\bar{y}_{SI}\right) = \frac{N-1}{N}\frac{S^2}{n} + \frac{(n-1)(N-1)}{N}\frac{S^2}{n}\rho_w == \frac{N-1}{N}\frac{S^2}{n}\left[1+(n-1)\rho_w\right]$$

Il campionamento sistematico

Varianza della media

La presenza di una correlazione positiva fra le unità appartenenti allo stesso campione fa aumentare il valore della varianza dello stimatore

quindi se la correlazione fra unita appartenenti allo stesso campione è positiva, allora aumenta la varianza dello stimatore

Anche un piccolo valore di ρ_w può provocare un forte incremento della varianza a causa del fattore moltiplicativo n-1. anche se rho è piccolo si può avere un grande aumento di varianza (fattore moltiplicativo (n-1))

Sia il primo che il secondo teorema esprimono la varianza della media ottenuta con il campionamento sistematico in termini della varianza S^2

e quindi è fortemente legata alla varianza della media nel caso di campionamento casuale semplice.

Teorema 3:

si può ipotizzare di esprimere la varianza in *termini di varianza ottenuta con il campionamento stratificato in cui gli strati sono composti, ciascuno da k unità* (il primo strato è composto dalle prime k unità, il secondo dalle seconde k, e così via).

in questo caso si guarda l'angolo dove abbiamo n strati ognuno di k unità

$$V(\bar{y}_{SI}) = \frac{S^2_{wstr}}{n}\frac{N-n}{N}\left[1+(n-1)\rho_{wstr}\right]$$

Il campionamento sistematico

Varianza della media

dove

$$S^2_{wstr} = \frac{1}{n(k\text{-}1)} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(y_{ij} - \bar{y}_{.j} \right)^2$$

esprime la varianza tra le unità all'interno dello stesso j-esimo strato e

$$\rho_{wstr} = \frac{E\left(y_{ij} - \bar{y}_{.j}\right)\left(y_{iu} - \bar{y}_{.u}\right)}{E\left(y_{ij} - \bar{y}_{.j}\right)^2} = \frac{2}{n(n-1)(k-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} \frac{\left(y_{ij} - \bar{y}_{.j}\right)\left(y_{iu} - \bar{y}_{.u}\right)}{S^2_{wstr}}$$

è il ***coefficiente di correlazione tra gli scarti ottenuti per ogni elemento nel campione sistematico rispetto alla medie per strato*** *cui appartengono.*

$\bar{y}_{.j}$ indica la media del j-esimo strato

Il campionamento sistematico

Confronti fra piani di campionamento

Stima della media

Si procederà effettuando i confronti fra il campionamento sistematico e gli altri considerati (casuale semplice e stratificato).

Primo confronto: **campionamento casuale semplice e campionamento sistematico**

Corollario del teorema 1: la media ottenuta con il campione sistematico è più precisa della media ottenuta con il campione casuale semplice solo se vale la seguente relazione

$$S_w^2 > S^2$$

cioè ***se la varianza entro i campioni sistematici è più grande della varianza della popolazione*** considerata nella sua globalità.

Dimostrazione:

La varianza dello stimatore nel campionamento casuale semplice è: $V(\bar{y}_{CCS}) = \frac{S^2}{n}\frac{N-n}{N}$

La varianza dello stimatore nel campionamento sistematico è: $V(\bar{y}_{SI}) = \frac{(N-1)}{N}S^2 - \frac{k(n-1)}{N}S_w^2$

Il campionamento sistematico

Confronti fra piani di campionamento

Stima della media

Si ponga

$$V(\bar{y}_{SI}) < V(\bar{y}_{CCS})$$

sostituendo si ottiene:

$$\frac{N-1}{N}S^2 - \frac{k(n\text{-}1)}{N}S_w^2 < \frac{S^2}{n}\frac{N-n}{N}$$

$$\frac{N-1}{N}S^2 - \frac{k(n\text{-}1)}{N}S_w^2 < \frac{S^2}{n}\frac{N-n}{N} = \frac{k(n\text{-}1)}{N}S_w^2 > \frac{N-1}{N}S^2 - \frac{S^2}{n}\frac{N-n}{N}$$

$$\frac{Nk(n\text{-}1)}{N}S_w^2 > (N-1)S^2 - \frac{S^2}{n}(N-n) = k(n\text{-}1)S_w^2 > S^2\left[(N-1) - \frac{N-n}{n}\right]$$

$$k(n\text{-}1)S_w^2 > k(n\text{-}1)S^2$$

se S_w > S allora il sistematico è più efficiente

cioè se la varianza all'interno dei campioni sistematic
è più grande di quella generale

dato che $N = n \times k$

per cui:

$$\left[(N-1) - \frac{N-n}{n}\right] = nk - 1 - \frac{nk-n}{n} = \frac{nnk - n - nk + n}{n} = k(n-1)$$

Il campionamento sistematico

Confronti fra piani di campionamento

Stima della media

-il campionamento casuale sistematico è più preciso rispetto al casuale semplice se

- la varianza all'interno dei campioni sistematici è più elevata della varianza della popolazione

-se le unità all'interno dello stesso campione sono eterogenee

se le unità della popolazione sono omogenee allora avremmo potuto diminuire la dimensione campionaria, oppure, l'ordinamento della lista non ha determinato lo stratificamento della popolazione.

- se non accade, ovvero ***se le unità sono omogenee***, vuol dire che le unità della popolazione, ordinate secondo il criterio scelto, non sono molto diverse tra loro.

-Le cause:

- una sostanziale omogeneità all'interno dei dati della popolazione
  - si può ottenere lo stesso risultato riducendo la dimensione campionaria

-l'ordinamento non ha determinato una stratificazione della popolazione
  - occorrerebbe scegliere un altro criterio di ordinamento o casualizzare la lista

Il ***confronto*** fra campionamento sistematico e casuale semplice può essere effettuato anche ***in termini di coefficiente di correlazione***. Infatti se poniamo l'uguaglianza fra le due relazioni:

$$V(\bar{y}_{SI}) = \frac{S^2}{n}\frac{N-1}{N}\left[1+(n-1)\rho_w\right] = \frac{S^2}{n}\frac{N-n}{N}$$

Il campionamento sistematico

Confronti fra piani di campionamento

Stima della media

possiamo scrivere:

$$\left[1+(n-1)\rho_w\right]=\frac{N-n}{N-1}\rightarrow(n-1)\rho_w=\frac{N-n}{N-1}-1=\frac{N-n-N+1}{N-1}=-\frac{n-1}{N-1}$$

quindi i due piani di campionamento coincidono se:

$$\rho_w=-\frac{n-1}{N-1}\frac{1}{n-1}=-\frac{1}{N-1}$$

in pratica le varianze della media per ccs e sistematico sono uguali quando rho = - 1 (N-1), in questo caso conviene il sistematico. Se rho è maggiore di questa quantità conviene ccs

Una guida per la scelta fra le due tecniche è riportata, in sintesi, nella tabella seguente:

Valori eterogenei	Valori omogenei (correlazione elevata)
$V(\bar{y}_{SI})<V(\bar{y}_{CCS})$	$V(\bar{y}_{SI})>V(\bar{y}_{CCS})$
$\rho_w=-\frac{1}{N-1}$	$\rho_w>-\frac{1}{N-1}$
$S_w^2>S^2$	$S_w^2<S^2$

se c'è eterogeneità conviene il sistematico, se i valori sono omogenei allora si ha una forte correlazione, e quindi conviene il ccs
per lo stesso ragionamento se le varianze negli strati sono alte (maggiori di quella) generale, abbiamo strati eterogenei, e quindi conviene il sistematico

Il campionamento sistematico

Confronti fra piani di campionamento

Stima della media

Secondo confronto: **campionamento casuale stratificato e campionamento sistematico**

Corollario del teorema 3: il campionamento sistematico ha la stessa precisione del campionamento stratificato con una unità per strato, per la stima della media, se

$$\rho_{wstr} = 0$$

Ciò si verifica perché la varianza della media nel campionamento stratificato con una unità per strato è pari a:

$$V(\bar{y}_{SI}) = \frac{(N-n)}{N}\frac{S^2_{2wstr}}{n}$$

Se

$$\rho_{wstr} > 0$$

il campionamento causale sistematico non migliora la precisione della stima della media

Il campionamento sistematico

Confronti fra piani di campionamento

Stima della media

Secondo confronto: **campionamento casuale stratificato e campionamento sistematico**

Individuiamo il valore che deve assumere

$$\rho_{wstr}$$

a parità di n, per scegliere il campionamento casuale sistematico.
La varianza della media nel campione causale sistematico

$$V(\bar{y}_{SI}) = \frac{S^2_{wstr}}{n}\frac{N-n}{N}\left[1+(n-1)\rho_{wstr}\right] \geq 0$$

da qua si vede meglio, se Rho = 0 la varianza della media è come lo stratificato, altrimenti è maggiore.

è positiva se

$$\left[1+(n-1)\rho_{wstr}\right] \geq 0 \rightarrow (n-1)\rho_{wstr} \geq -1 \rightarrow \rho_{wstr} \geq -\frac{1}{n-1}$$

quindi se

$$-\frac{1}{n-1} \leq \rho_{wstr} \leq 0$$

questa relazione ci dice se conviene il sistematico

il valore minimo dipende da n. Se $n\rightarrow\infty$ il minimo$\rightarrow$0

Il campionamento sistematico

Stima dei parametri

Totale

Una stima non distorta, per $N = n \times k$ del totale della variabile di interesse nella popolazione indagata:

$$\hat{Y} = \sum_{j=1}^{N} Y_j$$

è data dalla seguente

$$\hat{Y}_{SI} = \frac{N}{n} \sum_{j=1}^{n} y_j = N\bar{y}$$

che esprime ***il totale della variabile indagata nel campione casuale sistematico***

Varianza del totale

se $N = n \times k$

$$V(\hat{y}_{SI}) = E\left(\hat{y}_{SI} - \hat{Y}\right)^2 = \frac{N^2}{k} \sum \left(\bar{y}_i - \bar{Y}\right)^2$$

Il campionamento sistematico

Varianza del totale

Un altro modo per indicare la varianza del totale nel campionamento sistematico è:

$$V\left(\hat{y}_{SI}\right) = N\left(N-1\right)S^2 - Nk\left(n-1\right)S_w^2$$

dove

$$S_w^2 = \frac{1}{k\left(n-1\right)}\sum_{i=1}^{k}\sum_{j}\left(y_{ij} - \bar{Y}_j\right)^2$$

è la varianza tra le unità che costituiscono uno stesso campione sistematico.

Il campionamento sistematico

Confronti fra piani di campionamento

Stima del totale

Primo confronto: **campionamento casuale semplice e campionamento sistematico**

La varianza dello stimatore nel campionamento sistematico è:

$$V(\hat{y}_{SI}) = N(N-1)S^2 - Nk(n-1)S_w^2$$

La varianza dello stimatore nel campionamento casuale semplice è:

$$V(\hat{y}_{CCS}) = N^2 \frac{1-f}{n} S^2$$

Ne consegue che

$$V(\hat{y}_{SI}) < V(\hat{y}_{CCS})$$

se e solo se

$$S_w^2 > S^2$$

se la variabilità all'interno dei campioni sistematici è maggiore della variabilità sull'intera popolazione

Per il confronto fra campionamento stratificato e sistematico valgono le stesse considerazioni effettuate per la varianza della media praticamente le stesso considerazioni fatte con la media

Il campionamento sistematico

Stima dei parametri

Proporzione

Nella popolazione, la proporzione è pari a: $P = \frac{A}{N}$

Viene stimata con la proporzione campionaria che è pari a: $p = \frac{a}{n}$

che esprime ***la proporzione della variabile indagata nel campione casuale sistematico***

se $N = n \times k$ $E(p) = P$

Quindi la proporzione campionaria è uno stimatore non distorto della proporzione nella popolazione

Varianza della proporzione

$$V(p_{SI}) = \frac{\sum_{i=1}^{k} (\bar{p}_i - P)^2}{k}$$

Il campionamento sistematico

Varianza della proporzione

che si può esprimere anche in termini di S^2 e S_w^2

$$V(\bar{y}_{SI}) = \frac{N-1}{N} S^2 - \frac{k(n\text{-}1)}{N} S_w^2$$

dove

$$S^2 = \frac{P(1-P)}{N}$$

$$S_w^2 = \frac{1}{k(n\text{-}1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (p_{ij} - \bar{p}_{i.})^2$$

I confronti fra piani di campionamento seguono i criteri e la logica di quelli già seguiti per la stima della media come per la media anche qui (eterogeneità = good sistematico, omogeneità = bad sistematico)

Il campionamento sistematico

Considerazioni sul campionamento sistematico

Il campionamento casuale sistematico è un buon metodo quando

- la popolazione indagata segue un ordine casuale
  (ad esempio in ordine alfabetico per cognome).

In questo caso, infatti, sebbene non si sfrutti l'eventuale stratificazione naturale presente nella popolazione, **non si riscontra alcun trend lineare e non esiste correlazione fra elementi contigui**

Nel caso di popolazioni non ordinate

- che seguono un ordine casuale

il campionamento sistematico, in media, ***fornisce gli stessi risultati del campione casuale semplice***

Il campionamento sistematico

Considerazioni sul campionamento sistematico

Infatti:

consideriamo tutte le N! finite popolazioni che si possono formare con le N! permutazioni delle osservazioni

$$Y_1 Y_2 Y_3 \ldots\ldots Y_N$$

in media

$$E\left(V_{SI}\right) = V_{CCS}$$

la varianza nel campionamento casuale semplice è uguale in tutte le permutazioni

Nel caso di popolazioni con trend lineare

- la varianza del campione casuale sistematico è più efficiente della stessa nel campione casuale semplice (ma meno dello stratificato)

se è presente un trend lineare nella lista della popolazione allora : str > sistematico > ccs (cioè str migliore)

Il campionamento sistematico

Considerazioni sul campionamento sistematico

Si ipotizzi il caso più semplice di popolazione con trend lineare

i primi N numeri naturali

1,2,3,4,5,.....,N

la cui media aritmetica è $$M = \frac{N(N+1)}{2N}$$

mentre la varianza è pari a $$S^2 = \frac{N(N+1)}{12}$$

infatti $$S^2 = \frac{1}{N-1}\left[\sum i^2 - \frac{\left(\sum i\right)^2}{N}\right] = \frac{1}{N-1}\left[\frac{N(N+1)(2N+1)}{6} - \frac{N(N+1)^2}{4}\right] = \frac{N(N+1)}{12}$$

Il campionamento sistematico

Considerazioni sul campionamento sistematico

Ipotizzando un confronto fra

il campionamento casuale semplice e il campionamento stratificato con $N = n \times k$

la varianza dello stimatore media nel campione casuale semplice è pari a:

$$V_{CCS} = \frac{S^2}{n}\frac{N-n}{N} = \frac{N-n}{N}\frac{N(N+1)}{12n} = nk - n\frac{N+1}{12n} = n(k-1)\frac{N+1}{12n} = \frac{(k-1)(N+1)}{12}$$

Nel caso di campione casuale stratificato, invece, dove $N = k$

$$V_{CCStrat} = \frac{S_w^2}{n}\frac{N-n}{N} = \frac{N-n}{N}\frac{k(k+1)}{12n} = \frac{n(k-1)}{nk}\frac{k(k+1)}{12n} = \frac{(k^2-1)}{12n}$$

Nel caso di campione casuale sistematico la media del secondo campione è pari a:

media secondo campione = media primo campione +1

analogamente la media del terzo campione è

media terzo campione = media secondo campione +1

Il campionamento sistematico

Considerazioni sul campionamento sistematico

Per cui:

$$\sum\left(\bar{y}_i - \bar{Y}\right)^2 = \frac{k\left(k^2-1\right)}{12}$$

le medie infatti possono essere sostituite dai numeri 1,2,…,k. La varianza della media è dunque pari a:

$$V\bar{y}_{SI} = \frac{1}{k}\sum\left(\bar{y}_i - \bar{Y}\right)^2 = \frac{\left(k^2-1\right)}{12}$$

possiamo dedurre che:

$$\frac{\left(k^2-1\right)}{12n} \le \frac{\left(k^2-1\right)}{12} \le \frac{(k-1)(N+1)}{12}$$

cioè:

$$V\bar{y}_{Str} \le V\bar{y}_{SI} \le V\bar{y}_{CSS}$$

L'uguaglianza si verifica per n=1.

Quando c'è un trend lineare:

il campione casuale sistematico

è più efficiente del campione casuale semplice

meno efficiente del campione casuale stratificato