TEORIA E TECNICA DELL'INDAGINE STATISTICA E DEL CAMPIONAMENTO (MATR.DISPARI)

CAMPIONAMENTO A STADI
CAMPIONAMENTO A GRAPPOLI (CONTINUO)

MANUELA SCIONI
manuela.scioni@unipd.it
Dipertimento di Scienze Statistiche

Università degli Studi di Padova

Dipartimento di Scienze Statistiche

CAMPIONAMENTO A GRAPPOLI:
IL COEFFICIENTE DI CORRELAZIONE INTRACLASSE

- Il coefficiente di correlazione intraclasse misura il grado di concordanza fra unità organizzate entro gruppi:

$$\rho = \frac{Cov(y_{ij}, y_{ij'})}{\sqrt{\left(Var(y_{ij})Var(y_{ij'})\right)}} \quad i = 1, \dots K, j \neq j'$$

K è il numero di UPS nella popolazione

M è il numero di USS in ciascuna UPS (semplificazione: dimensione costante)

La stima campionaria di ρ è data da:

$$r = \frac{\sum_{i=1}^{k} \sum_{j=1}^{M} \sum_{\substack{l=1 \\ l \neq j}}^{M} (y_{ij} - \bar{y})(y_{il} - \bar{y})}{\sum_{i=1}^{k} (M-1) \sum_{j=1}^{M} (y_{ij} - \bar{y})^2}$$

IL COEFFICIENTE DI CORRELAZIONE INTRACLASSE

Formulazione ANOVA:

$$\rho = 1 - \frac{M}{M-1}\frac{SSW}{SSTO}; \qquad -\frac{1}{M-1} \leq \rho \leq 1$$

Cluster totalmente omogenei → SSW=0 → ρ = 1
Cluster totalmente eterogenei → SSB=0 → SSW=SSTO → ρ =-1/(M-1)
USS distribuite casualmente → ρ = 0

$$\mathrm{MSB} = \frac{NM-1}{M(N-1)} S^2 \left[1 + (M-1)\rho\right]$$

DEFF NEL CAMPIONAMENTO A STADI

$$DEFF(stadi) = \frac{V(\hat{t}_{stadi})}{V(\hat{t}_{ccs})} = \frac{MSB}{S^2} = \frac{KM-1}{M(K-1)}[1+(M-1)\rho]$$

Se K è grande, MK>>M, per cui (MK-1)≅(MK-M), e:

DEFF ≅ 1 + (M-1) ρ

Se m<M, al posto di M si utilizza la dimensione $\bar{m}$ media delle unità di I stadio

Già piccoli valori positivi di ρ implicano un aumento considerevole della varianza di stima rispetto al CCS. Ad esempio, per ρ=0,1 e M=31, la varianza di stima è già quattro volte quella del CCS; per ρ=0,2 è sette volte.

Come ridurre l'effetto della somiglianza delle unità entro i gruppi?

- Distribuire i sub-campioni tra un numero maggiore di unità di primo stadio
- Evitare cluster troppo numerosi
- Creare sub-campioni eterogenei

CASO 2: CAMPIONAMENTO AD UNO STADIO CON NUMEROSITÀ DI II STADIO DIVERSE

Due possibilità per la stima:

1. Stimatori non distorti, del tutto simili a quelli visti per il caso 1, con la differenza che il denominatore della stima della media è $K=\Sigma M_i$, quantità non sempre nota. Inoltre, la varianza di stima tende ad essere molto più grande nei cluster più numerosi.

1. Stimatori rapporto: $\hat{\bar{y}}_r = \frac{\sum_{i \in S} t_i}{\sum_{i \in S} M_i}$ $\qquad \hat{t}_r = K \cdot \hat{\bar{y}}_r$

   la cui varianza è molto inferiore a quella dello stimatore non distorto

TEORIA E TECNICA DELL'INDAGINE STATISTICA E DEL CAMPIONAMENTO (MATR.DISPARI)

CAMPIONAMENTO A STADI

CAMPIONAMENTO A DUE STADI

MANUELA SCIONI
Dipartimento di Scienze Statistiche
manuela.scioni@unipd.it

Università degli Studi di Padova

Dipartimento di Scienze Statistiche

CAMPIONAMENTO A DUE STADI

- Nel campionamento a due stadi si procede a selezionare un campione di USS da ogni UPS selezionata, per cui i totali all'interno di ogni UPS, che nel campionamento a grappoli erano noti, sono ora delle variabili casuali che contribuiranno alla varianza di stima.
- La varianza di stima è data da due addendi: il primo è lo stesso del campionamento ad uno stadio e rappresenta la variabilità fra le UPS, il secondo è la variabilità dovuta al campionamento effettuato al secondo stadio

PROBABILITÀ DI SELEZIONE

- Campione su 2 stadi:

$$p_{ij} = p_i \cdot p_{j|i}$$

dove

- p_{ij} è la probabilità congiunta di selezionare l'unità di secondo stadio j e quella di primo stadio i
- p_i è la probabilità di selezionare l'unità di primo stadio i
- $p_{j|i}$ è la probabilità di selezionare l'unità di secondo stadio j condizionatamente al fatto che è stata selezionata quella di primo stadio i

- Per un campione su 3 stadi:

$$p_{ijk} = p_i \cdot p_{j|i} \cdot p_{k|ij}$$

CAMPIONE A STADI AUTOPONDERANTE

Caso 1: campione a grappoli con CCS al I stadio

$p_i = \frac{k}{K}; \quad p_{j|i} = 1; \quad p_{ij} = \frac{k}{K}$

Caso II: prob. di selezione costanti ad ogni stadio

$\Rightarrow p_i = \frac{k}{K}$, dove K = numerosità popolazione di I stadio e k = numerosità campionaria di I stadio

$\Rightarrow p_{j|i} = p_j = \frac{m}{M}$, dove $M =$ numerosità popolazione di II stadio contenuta in un'unità di I stadio e $m =$ numerosità campionaria di secondo stadio

$$\Rightarrow p_{ij} = \frac{k}{K} \cdot \frac{m}{M} = \frac{km}{KM}$$

PROBABILITÀ COSTANTE AD OGNI STADIO: NOTE

- La condizione $p_{j|i} = p_j = \frac{m}{M}$ implica che ogni unità di primo stadio contenga un numero fisso M di unità.
- Tale condizione, se non è realizzata nella realtà, può essere creata attraverso opportuni piani di campionamento che controllino la dimensione delle unità primarie, ad esempio:
  - ordinare le unità primarie sulla base della dimensione (es: ordinare le città sulla base della popolazione residente)
  - formare *K* strati di dimensione pressoché uguale e pari a *M*, unendo le unità primarie di piccole dimensioni e suddividendo quelle più grandi.
  - selezionare *k* strati
  - selezionare un numero fisso di m unità per strato

CAMPIONE A DUE STADI AUTOPONDERANTE (2)

Caso III: prob. di selezione proporzionali alla dimensione delle unità (PPS)

$\Rightarrow p_i = k \cdot \frac{M_i}{N}$, dove $n =$ numerosità campionaria UPS, M_i numero USS nell'i-esima UPS e

$N = \sum_{i=1}^{K} M_i$

$\Rightarrow p_{j|i} = \frac{m}{M_i}$, dove $m =$ numerosità campionaria USS

$\Rightarrow p_{ij} = p_i \cdot p_{j|i} = k \cdot \frac{M_i}{N} \cdot \frac{m}{M_i} = \frac{k \cdot m}{N}$

Se non si conosce la dimensione delle UPS, ma una misura B_i proporzionale a M_i (esempio il bilancio dell'azienda)

$\Rightarrow p_{ij} = p_i \cdot p_{j|i} = k \cdot \frac{B_i}{B} \cdot \frac{m}{B_i} = \frac{k \cdot m}{B}$, dove $B = \sum_{i=1}^{K} B_i$

TEORIA E TECNICA DELL'INDAGINE STATISTICA E DEL CAMPIONAMENTO (MATR.DISPARI)

CAMPIONAMENTO A STADI:
CAMPIONAMENTO SISTEMATICO

MANUELA SCIONI
Dipartimento di Scienze Statistiche
manuela.scioni@unipd.it

Università degli Studi di Padova
Dipartimento di Scienze Statistiche

IL CAMPIONAMENTO SISTEMATICO COME CASO SPECIALE DEL CAMPIONAMENTO A STADI

- Determino il passo di campionamento t=N/n, che è il numero di possibili campioni che posso estrarre
- Una volta estratto il n° casuale r ($1 \leq r \leq t$), tutti gli elementi del campione sistematico sono determinati. Ciò equivale ad avere un campione ad uno stadio, dove il numero di UPS è dato dai possibili campioni che possono essere costruiti (=*t*) e M, l'ampiezza del cluster, è la numerosità campionaria. Il numero di UPS selezionate è pari a k=1, perché dei *t* possibili campioni ne viene estratto solo uno.
- Si osserva allora una sola media di cluster, della quale non si può stimare la varianza dato che k=1. È necessario stimarla da altre fonti. In alternativa si possono compenetrare i campioni sistematici, ovvero selezionare più campioni partendo da punti casuali differenti.

CAMPIONAMENTO SISTEMATICO VS. CCS

1. La lista è in ordine casuale: $\rho \cong 0$ e il camp. sistematico dà risultati simili al CCS. Si possono usare le formule del CCS per stimare la varianza.
2. La lista è in ordine crescente o decrescente: presumibilmente $\rho < 0$, perché le unità simili sono vicine e col camp. Sistematico selezioniamo unità distanti fra loro. Il camp. sistematico è allora più efficiente del CCS. Le formule del CCS forniranno int. di confidenza più ampi del vero.
3. La lista è ciclica e il ciclo è un multiplo di t. Gli elementi campionati sono circa uguali e la stima della varianza (formula CCS) risulterebbe quasi nulla (erroneamente). Sarebbe lo stesso selezionare una sola unità dalla popolazione.