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

CAMPIONAMENTO SISTEMATICO E A PROBABILITÀ VARIABILI

camp SISTEMATICO._ molto simile ccs spesso tratatto come ccs

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

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

CAMPIONAMENTO SISTEMATICO

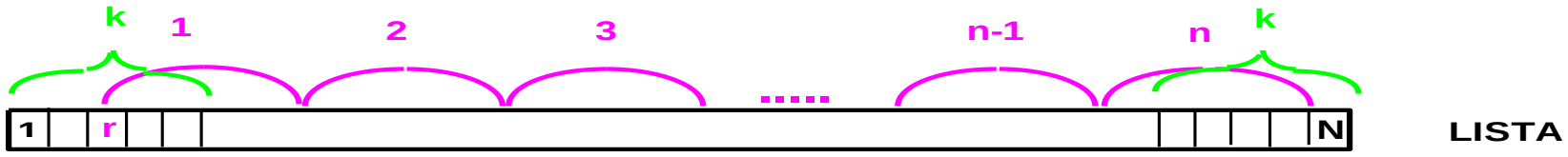
basta conoscere la N e la mia numerosità ottimale e posso anche senza avere la lista completa es: supermarket voglio indagare su tutti i clienti allora determino N in una settimana ottengo n e mi calc passo di camionamento dopo me metto fuori da l supermarket e a partire dal cliente r_esimo scelgo uno ogno top es: se ogni 10 viene 7 ogni volta

Selezione casuale in blocco prendendo un'unità ogni tante a partire da una scelta a caso. Approssima il CCS.

Procedimento:

1. messa in sequenza delle unità della lista
2. determinazione del "passo di campionamento" $k = N/n \;\; (= 1/\pi_i)$
3. identificazione delle *n* unità:
   a) selezione del numero casuale *r*: $1 \leq r \leq k$
   b) identificazione delle unità campionarie $r;\; r + k;\; r + 2\,k;\; \ldots, r + (n\text{-}1)\,k$

SE decimale arrotondo al piu piccolo non al piu grande



camp casuale hanno prob non nulla di entrare di entrare a far parte del camp MA
camp sistematico è pseudocasuale xke casualita solo in r poi tutti quelli multipli di k hanno prob 1 e gli altri prob zero quindi non nulla fino al primo numero casuale dopodiche puo essere nulla cioe pseudocasuale

CAMPIONAMENTO SISTEMATICO VS. CCS

VANTAGGI

- Semplicità della procedura (si estraggono uno o due numeri casuali)
- Se la lista è disposta casualmente, lo si tratta come un CCS
- Se la lista è ordinata, è generalmente più efficiente di un CCS, perché si attua implicitamente un controllo sulle caratteristiche della popolazione

se la lista puo essere ordinata da giovane a anzianno mi permette di applocare un controllo, la v. rispetto a cui faccio ordinamento la chiamo v.ausiliaria che è correllata

SVANTAGGI

- È in realtà un campione pseudo-casuale, dato che viene estratto un solo numero casuale e tutti gli altri sono automaticamente determinati
- Rischio di distorsione del campione dato da ciclicità della lista
- Manca uno stimatore corretto della varianza di stima. Si adottano stimatori asintotici

xke ce ciclicità
es lista neo diplomati padova al interno sono ordinati per voto alla maturita, se io ho un passso di camp = numerosita di una classe ce riskio che vado a prender sempre piu alti o bassi cioe contrario alla stratificazione implicita, invece con ciclica questa vado a prendere sempre unita con caratteristiche simili.

STRATIFICAZIONE IMPLICITA: LA STRATIFICAZIONE A SERPENTINA

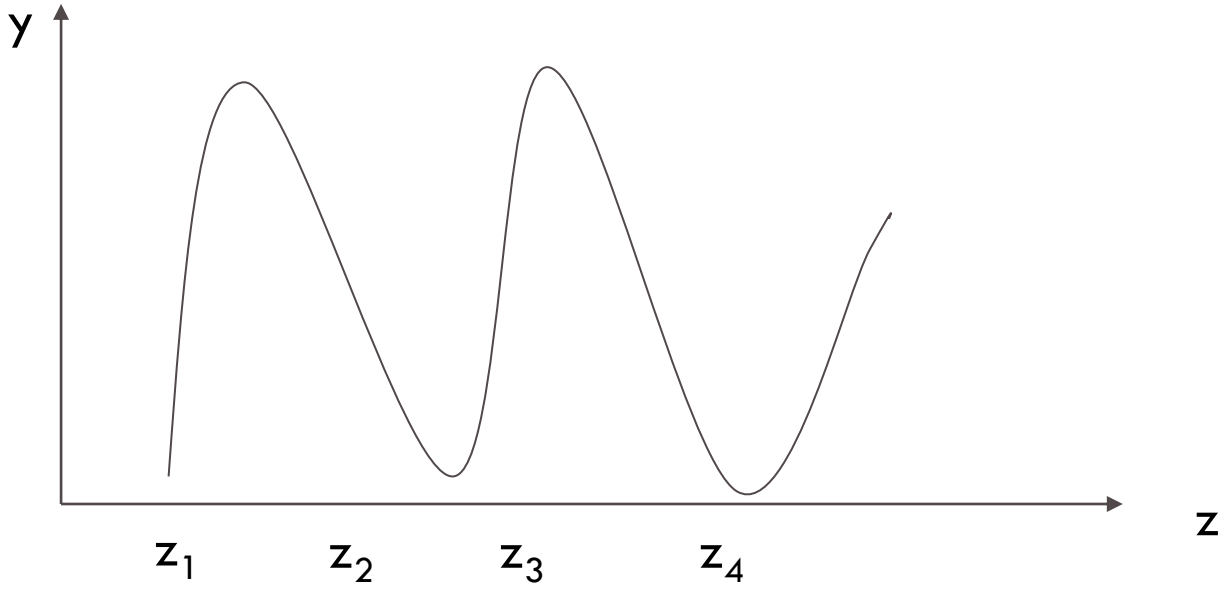
Con una sola variabile di stratificazione (quantitativa):

1. Ordinamento delle unità da quella con il valore più elevato a quella con il valore minore
2. Selezione sistematica delle unità

caso in cui piu v.ausiliare in cui una puo esssere qualitativa
es: eta,sesso ordino a sepentina ordino da giov a vecchio per maschi e pio vecchio a giov per femmine dopodiche scelgo un unita ogni top, cosi questa procedura controlla entrambe le 2 variabili si vedra piu avanti
es: grafico: 4 modalita per la qualitattiva e y=quantitativa come v.ausilitaria se ordino le mie unita picolo grande prima categoria poi grand pic seconda categoria e cosi si crea la serpentina

Con una variabile di stratificazione quantitativa e una o più variabili qualitative:

Ordinamento "a serpentina"



SELEZIONE CON PROBABILITÀ VARIABILI

Selezione casuale di un campione realizzata assegnando alle unità della popolazione probabilità variabili

Assegnazione all'unità *i* (*i*=1, ..., *N*) della probabilità p_i di selezione in un passo di selezione mediante assegnazione di M_i numeri casuali su un totale di *M*.

massa es: tot di residenti in italia

$$M = \sum_i^N M_i$$

$$p_i = \frac{M_i}{M}$$

i=comune i_esimo ogni comune ha una prob proporzionale al numero di abitante cioe abitanti/abitatntiTotItalia

lle unita che fanno parte della pop non hanno la stessa prob di far parte del campione quindi non siamo nella situcaszione ccs che tutte hanno porob costante, ma è il ricercatore che assegna questa prob per tenere conto della dim cioe della loro importanza rispetto ad altre:
es:devo fare un camp di turni , P comuni italiani , tutti hanno la stessa prob di essere estratta ma morbello molto piu picoolo di Roma allora non posso ignirare la dim xke roma piu informativo,

SELEZIONE CON O SENZA REINSERIMENTO

poco senso per pop finite

pi prob di inclusione

Selezione casuale con reinserimento $\pi_i = np_i$

- Selezione di *n* numeri casuali
- identificazione delle unità a cui sono stati assegnati i numeri casuali (è possibile che un'unità sia estratta più volte)

Selezione casuale senza reinserimento (π_i= ?)

- Selezione di un numero casuale e identificazione dell'unità corrispondente
- i numeri casuali corrispondenti alle unità estratte si considerano "bianchi" (se estratti vengono ignorati)

quello che facciamo noi sempre senza reinserimento, non sono piu in gradi di calc inclusione di prob dal passo 2 in poi, al primo va bene xke v.ausiliare ma dal 2 in poi cambiano xke dip da cosa è estratto al primo passo es:: padova200000 al secondo passso se estratto roma al denominatore 70 milione - 1 millione quindi dal 2 passo mi cambiano le prob

SELEZIONE P.P.S.

- **Probability Proportional to Size** = campionamento con probabilità proporzionali alla dimensione
  - Rappresenta il disegno più efficiente fra quelli a probabilità variabile
  - È anche il disegno più logico: le unità più "importanti" hanno maggiore probabilità di essere selezionate
- Come determinare le probabilità da attribuire ad ogni unità? Mediante una variabile ausiliaria *X*, correlata al fenomeno *Y* in analisi (o che comunque determina una > o < rappresentatività delle unità rispetto a Y)
- Esempi:
- Y=fatturato impresa; X=n° addetti
- Y=soddisfazione studenti; X=n° esami sostenuti

CAMPIONAMENTO CON PROBABILITA' VARIABILI: STIMA DEL TOTALE

P PIU FACILE LAVORARE SUL TOTALE CHE SULLA MEDIA nel camp con prob variabile

Nel campionamento con probabilità variabili (senza ripetizione) si fa riferimento allo **stimatore (del totale) di Horvitz-Thompson** (1952)

somma del mio campione ciascuna fratto la sua prob di inclusione

$$\hat{Y} = \sum_{i \in c} \frac{y_i}{\pi_i} = \sum_{i=1}^{N} \frac{Y_i}{\pi_i} I_i$$

inverso della prob di seleizione cioe maggior è la parte di entrare minore sara il peso cioe chi ha meno prob avra piu peso invece nel ccs N/n ed è uguale su tutte le unita

$1/\pi$ all'universo

La stima ottenibile con l'applicazione di questo stimatore è molto efficiente se si trova una grandezza (circa) proporzionale alla grandezza da stimare.

Se $\pi_i = n / N$, lo stimatore del totale diventa il più comune stimatore basato sulla media campionaria:

$$\hat{Y} = \sum_{i}^{n} \frac{y_i}{n} N = N\bar{y}$$

CAMPIONAMENTO CON PROB VARIABILI VS. PROB. COSTANTI

es imprese ccs se fiat avrebbe la stessa prob di un azienda piccola cioe ha poca informazione rispettp alla fiat

VANTAGGI

- Controllo nella selezione rispetto alla dimensione delle unità
- Guadagno in efficienza commisurato alla relazione esistente fra la dimensione delle unità e la variabile oggetto di studio (se esiste proporzionalità, il campione può essere più piccolo, a parità di efficienza attesa)

non sempre è possibile consocere la pop ad es dei comuni o num di dipendenti è difficele da raggiungere

SVANTAGGI

- È necessario conoscere la dimensione delle unità della popolazione
- Il campione risultante non è **auto-ponderante** (quindi le unità NON hanno tutte la stessa probabilità di selezione)
- Gli stimatori da adottare sono complessi, soprattutto se la selezione è senza re-immissione

ESERCIZIO CAMPIONAMENTO SISTEMATICO

Tutti i risarcimenti rilasciati da un'assicurazione in un giorno dell'anno 2018 (estratto a caso) sono, in euro:
400, 600, 570, 960, 780, 800, 460, 650, 440, 530, 470, 810, 625, 510, 700.

- Si elenchino tutti i possibili campioni sistematici di numerosità campionaria pari a 3 che possono essere estratti dalla popolazione di risarcimenti elencata.
- Calcolare le corrispondenti medie campionarie.
- Come si può calcolare la varianza di stima? Motivare la risposta

p sempre casuale

ESERCIZIO CAMPIONAMENTO A PROBABILITÀ VARIABILI

Formare un campione di 3 professori da una lista di 7, con probabilità proporzionale al numero di studenti che seguono il rispettivo corso:

voglio formare un camp di 3 prof,
devo definire le varie prob di selezione
quindi prima devo calc num di studetnti totali quinidi sommatoria

Professore	**N studenti**
Pippo	120
Topolino	45
Pluto	89
Paperino	54
Paperoga	134
Paperina	67
Gastone	23