

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

PRINCIPI DI CAMPIONAMENTO

siamo arrivati fino al piano di pubblicazione
ci manca il piano di campionamento:
cioè insieme di procedure che si adottano quando facciamo
un'indagine campionaria piuttosto che esaustiva,
esaustiva se è coinvolta tutta la pop.

es: esami sono domande a campione = indagine
campionaria le domande debbono essere diversificate così
rapp la realtà

NON voglio trarre conclusioni dal campione MA dalla pop
tramite il campione

MANUELA SCIONI

Dipartimento di Scienze Statistiche

manuela.scioni@unipd.it



UNIVERSITÀ
DEGLI STUDI
DI PADOVA



da adesso in poi popolazione =
pop obiettivo e pop statistica
non distinguo piu, pop in campionamento parlo di pop statistica
quindi suppongo che le 2 pop coincidono da adesso

PROGRAMMA

quando tutte le unit facenti parte della pop. o
meglio della mia lista hanno prob non nulla
cioe tutti hanno la possibilita di entrare nella
campione

CAMPIONAMENTI PROBABILISTICI

unita selez casualmente dalla lista

■ **Campionamento casuale semplice**

selez casuale al interno di ogni stratto,
sono indipendenti es: laureati padova gli
suddivido rispetto alla variabile faccio una
selezione casuale al interno di questo gruppo

■ **Campionamento stratificato**

■ **Campionamento a stadi**

TUTTI 3
CASUALI

le unit stat sono aggregate
in un unita meno ampie
es: laureati appartengono universita corsi di
lauree che app a univ che app a atenee, seleziono
casualmente le unita aggregate es: corsi di
laurea, utile quando non abb l intera lista della pop,
es: indagine sulle scuole superiori non sapp ogii i
nomi cognomi di tutti pero sappiamo tutte le
scuole allora selz 5 scuole poi mi approccio alla
singola scuola dalla lista degli studenti quindi
senza avere la lista di partenza sono riuscito a
fare un campionamento

ALTRI TIPI DI CAMPIONAMENTO

- **Campionamento sistematico (pseudocasuale)**
- **Campionamento per quote (non probabilistico)**

guardare esempio lezione
26 aprile

CAMPIONAMENTO PROBABILISTICO

- Punto di partenza è P , una popolazione di dimensione N
- c = campione di dimensione n tratto dalla popolazione P

in maiuscolo = popolazione
minuscolo = campione

- **Campionamento probabilistico:**

1. Definire l'insieme dei possibili campioni $C = \{c_1, c_2, \dots, c_M\}$
2. Assegnare ad ogni possibile campione una probabilità $p(c)$
3. Ad ogni unità è associata una probabilità $\pi_i > 0$ di far parte del campione

coef binomiale

t.c la somma = 1
e ogni unità ha prob non nulla di essere estratta

NON FAREMO LA PROCEDURA cmq LA VEDIAMO
PRIMA LEZIONE
questi 4 step formano il DISEGNO DI CAMPIONAMENTO

$$\pi_i = \sum_j P(c_j | i \in c_j)$$

cmq può avere zero ma deve avere la possibilità di entrare

4. Definire una procedura per ottenere un campione c con probabilità $p(c)$

La distribuzione di probabilità $p(c)$ definisce il **disegno di campionamento**

esempio slide a mano controllare:
fattoriale; 6 numeri di possibili campioni che posso estrarre da 4 senza rimpiazzamento

CAMPIONAMENTO DA POPOLAZIONI FINITE

- P è una popolazione di numerosità **finita** N
- Di conseguenza, il numero di campioni che possiamo costruire è finito, e ad ognuno di essi è associata una probabilità $p(c)$
- Lo stimatore è funzione dei valori Y_i di un campione c :

$$\hat{\theta}_c = f(Y_i, i \in c)$$

parleremo spesso di stimatore della media e totale
il camp deve rapp lo pop così lo stimatore ha senso, posso associare alcuni parametri allo stimatore che descrivono la distribuzione

- La distribuzione dello stimatore è composta da un insieme finito di probabilità, ed è perciò assimilabile ad una funzione di probabilità di v.c. discrete

Il obiettivo dell'indagine è andare a stimare questi parametri
per un'indagine avremo più variabili di interesse nella realtà con più parametri di interesse, cioè la var. di interesse può essere descritta da uno o più parametri

POPOLAZIONE E PARAMETRI (IGNOTI)

P = Popolazione di N elementi

Y = caratteristica di P che vogliamo studiare

Lo studio di Y avviene mediante la stima dei suoi parametri:

- Media
- Totale
- Mediana
- Percentili

θ è il parametro relativo a Y, nella popolazione P, che intendiamo stimare con l'indagine campionaria

θ Parametro ignoto

$$\text{Media: } \bar{Y} = \sum_{i=1}^N \frac{Y_i}{N}$$

$$\text{Totale: } Y = \sum_{i=1}^N Y_i$$

$$\text{Varianza: } S^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2; \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2$$

$$\text{Covarianza: } \sigma_{XY} = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$$

$$\text{Coefficiente di correlazione: } \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Se Y è dicotomica (Y = 0 assenza, Y = 1 presenza):

$$\bar{Y} = \text{proporzione} = P = \frac{N_1}{N} \quad \sigma^2 = P \cdot (1 - P)$$

in realta non faremo mai cosi,
ma prima estraiamo un
campione e poi vediamo errore

PARAMETRI DELLO STIMATORE

esempio
altezza media
lezione 26
aprile

| | |
|----------------------------|--|
| | |
| Distribuzione campionaria: | $P(\hat{\theta} = k) = \sum_{c: \hat{\theta}_c = k} p(c)$ |
| Valore atteso: | $E(\hat{\theta}) = \sum_c \hat{\theta}_c p(c)$ |
| Distorsione: | $B(\hat{\theta}) = E(\hat{\theta}) - \theta$ |
| Varianza: | $Var(\hat{\theta}) = \sum_c (\hat{\theta}_c - E(\hat{\theta}))^2 p(c)$ |
| Errore Quadratico Medio: | $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \sum_c (\hat{\theta}_c - \theta)^2 p(c) = Var(\hat{\theta}) + (B(\hat{\theta}))^2$ |

ci serve per capire se distorto

entrambi gli errori danno varianza di stima più varianza di rilevazione, quello calc nel esempio era varianza di stima, in nostro errore camp deve essere il più piccolo possibile è questo l'obiettivo di tutte le procedure possibili, errore camp err non camp vanno controllati sempre, di solito quello di rilevazione è più alto di solito, es: xke le persone non dichiarano il vero es: quando berlusconi vinse

vanno
contrallati tutti
e due gli errori

COMPONENTI DELL'ERRORE

xke indagine
camp

ERRORE CAMPIONARIO

- Al cambiare del campione cambia la stima di θ , quindi varia $\hat{\theta}$:

- Varianza di stima:

$$Var_S(\hat{\theta}) = \sum_c^{N^*} \{\theta_c - E(\hat{\theta})\}^2 p_c$$

- Errore campionario di stima

- $\sqrt{Var_S(\hat{\theta})}$

- Errore quadratico medio di stima:

$$Mse_S(\hat{\theta}) = \sum_c (\hat{\theta}_c - \theta)^2 p(c) = Var(\hat{\theta}) + (B(\hat{\theta}))^2$$

ERRORE DI RILEVAZIONE

- È dovuto a errate dichiarazioni (volontarie, dimenticanze, interpretazioni non corrette da parte dell'intervistatore, ...)

- Anch'esso contribuisce alla varianza di $\hat{\theta}$

errori per errore del
rispondente o nostro,,
puo anche essere
sistematico

varianza di rilevazione, in un infagine camp non posso calcolarla

$$Var(\hat{\theta}) = Var_S(\hat{\theta}) + Var_R(\hat{\theta})$$

varianza stima calcolata
prima nel esempio 40 min