

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

MANUELA SCIONI

Dipartimento di Scienze Statistiche

Manuela.scioni@unipd.it



UNIVERSITÀ
DEGLI STUDI
DI PADOVA



POST-STRATIFICAZIONE

- Talvolta, si possono ipotizzare ragionevoli variabili di stratificazione, ma il valore di queste non è noto a priori per tutti gli elementi della popolazione e quindi non si possono utilizzare nella fase di disegno del campione.
- Ad esempio, in un'indagine sulle famiglie, si vuole stimare la spesa mensile per l'acquisto di beni alimentari. Sarebbe interessante stratificare per dimensione delle famiglie, che è però ignota per le unità della lista. Da una fonte amministrativa è disponibile la distribuzione della dimensione familiare nella popolazione. Tale distribuzione fornisce i W_h .

POST-STRATIFICAZIONE (2)

- Se si prende un CCS piuttosto grande, il campione assomiglierà a uno stratificato proporzionale con riferimento alla variabile “numero di membri della famiglia” (ciò varrebbe per qualsiasi variabile). Possiamo stimare y_h in ogni strato $h = 1, 2, \dots, H$ e poi combinare y_h come in un campione stratificato

$$\bar{y}_{post} = \sum_{h=1}^H W_h \bar{y}_h$$

utilizzando i valori di W_h dati nella distribuzione della popolazione.

POST-STRATIFICAZIONE (3)

- In generale allora, se: W_h è noto per strati definiti a valle della rilevazione; n_h non è piccolo (>30); n è grande, possiamo considerare la seguente approssimazione per la stima della varianza:

$$\text{Var}(\bar{y}_{post}) \cong (1 - f) \sum_{h=1}^H W_h \frac{S_h^2}{n}$$

che è essenzialmente ottenuta utilizzando la varianza dello stratificato proporzionale.

- Questa tecnica è anche utilizzata per correggere gli effetti di mancate risposte che dipendano esclusivamente dalle variabili di post-stratificazione.

ES. DATI MANCANTI

	M	F	TOTALE
Popolazione	40%	60%	100%
Campione	160 (40%)	240 (60%)	400 (100%)
Rispondenti	90 (30%)	210 (70%)	300 (100%)
Pesi di post-stratificazione	$\frac{N_1}{N} = 0.4$	$\frac{N_2}{N} = 0.6$	1
Effetto della post-stratificazione	$0.4/0.3=1,33$	$0.6/0.7=0,857$	

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

CAMPIONAMENTO NON PROBABILISTICO PER QUOTE

MANUELA SCIONI

Dipartimento di Scienze Statistiche

manuela.scioni@unipd.it



UNIVERSITÀ
DEGLI STUDI
DI PADOVA



CAMPIONAMENTO NON PROBABILISTICO

- Alcune unità della popolazione oggetto d'indagine hanno probabilità nulla di entrare a far parte del campione
- Esempi:
 - non si dispone di una lista della popolazione, per cui si reclutano unità campionarie senza partire da una lista
 - Ma anche: si parte da una lista della popolazione che non coincide con la popolazione d'interesse
 -

CAMPIONAMENTO PER QUOTE

Il riferimento è il campionamento stratificato proporzionale:

- Si suddivide la popolazione secondo alcune variabili delle quali si conosce la distribuzione nella popolazione (es: titolo studio dal Censimento)
- Il campione viene costruito rispettando le proporzioni delle variabili di stratificazione nella popolazione
- Non è disponibile la lista dettagliata secondo le var. di stratificazione, per cui si prosegue a reclutare fino a che tutte le quote sono riempite

CAMPIONAMENTO PER QUOTE (2)

Il campionamento non è probabilistico perché:

- Non c'è una selezione casuale da una lista
- L'intervistatore è libero di scegliere chi intervistare, presumibilmente chi è più comodo da raggiungere o dà sicurezza di risposta (quindi alcuni soggetti avranno probabilità nulla di essere interpellati)
- Infine, nel camp. per quote è impossibile analizzare la qualità del campione mediante un confronto con la distribuzione nella popolazione, perché tale distribuzione è comunque rispettata in virtù del meccanismo delle quote. Non esiste una distribuzione dei dati mancanti.

COME SELEZIONARE LE UNITÀ

- Indagini telefoniche (tipicamente sondaggi di opinione): le unità sono selezionate a partire dagli elenchi telefonici.
 - Si chiede della presenza di un soggetto con determinate caratteristiche (es. donna fra i 20 e i 40 anni); se assente, si salta alla famiglia successiva.
- Indagini telefoniche: a partire da numeri generati casualmente
- Indagini via web
- Interviste faccia a faccia: all'uscita del supermercato, del seggio elettorale, ...
 - Si scelgono più luoghi, più giorni e più orari, persone con caratteristiche diverse

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

I SONDAGGI

MANUELA SCIONI

Dipartimento di Scienze Statistiche

manuela.scioni@unipd.it



UNIVERSITÀ
DEGLI STUDI
DI PADOVA



COME SI SVOLGONO I SONDAGGI

dal Mattino di Padova del 29/5/2016

XXX ha realizzato questa indagine che si è svolta a livello nazionale dal 22 marzo al 4 aprile 2016 su un **campione rappresentativo** della popolazione residente in Italia, con età superiore ai 18 anni. I rispondenti totali sono stati **1997 (su 13.287 contatti)**, l'analisi dei dati è stata **riproporzionata** sulla base del genere, del territorio, delle classi d'età, della condizione professionale e del titolo di studio. Il **marginale di errore** è pari a $\pm 2.2\%$. La rilevazione è avvenuta con un'indagine attraverso i principali social network e con un **campione casuale** raggiungibile con i metodi **CAWI e CATI**.



Menu

Home

HELP

Sondaggi

Elenco Sondaggi

Log In

Registrazione

Vecchio sito

Privacy

SONDAGGIO

Dati Sondaggio

Domande

Conclusioni

Titolo del sondaggio

La situazione politica - 2/9/2016

Soggetto che ha realizzato il sondaggio

Soggetto committente

Agorà-RAI 3

Soggetto acquirente

Agorà-RAI 3

Data o periodo in cui è stato realizzato il sondaggio - Da

31/08/2016

Data o periodo in cui è stato realizzato il sondaggio - A

31/08/2016

Mezzo(i) di comunicazione di massa sul quale(i) è stato pubblicato o diffuso il sondaggio

Agorà-RAI 3

Data di pubblicazione o diffusione

02/09/2016

Popolazione di riferimento

Popolazione residente in Italia, di 18 anni e oltre

Estensione territoriale del sondaggio

Nazionale (Italia)

Metodo di campionamento, inclusa l'indicazione se trattasi di campionamento probabilistico o non probabilistico, del panel e l'eventuale ponderazione

Campione casuale probabilistico stratificato di 1.000 soggetti maggiorenni rappresentativo rispetto ai parametri di sesso, età e macro area di residenza

Consistenza numerica del campione di intervistati, numero dei non rispondenti e delle sostituzioni effettuate

1.000 soggetti maggiorenni (su 8.914 contatti complessivi)

Rappresentatività del campione, inclusa l'indicazione del margine d'errore

Margine di errore (livello di rappresentatività del campione al livello di confidenza del 95%): $\pm 3,1 \%$

Metodo raccolta delle informazioni

Interviste telefoniche su utenze fisse e cellulari (CATI/CAMI)

Toolbar

Scarica Sondaggio

Gestione Sondaggi

ANALIZZIAMO PUNTO PER PUNTO

- **Campione casuale (o probabilistico)**: Ogni unità della popolazione ha probabilità non nulla di essere selezionata: $0 < p_i \leq 1, \sum p_i = 1$.
- Per estrarre un campione casuale serve una **LISTA ESAUSTIVA*** della popolazione

* a meno di campionamento a stadi

LA LISTA

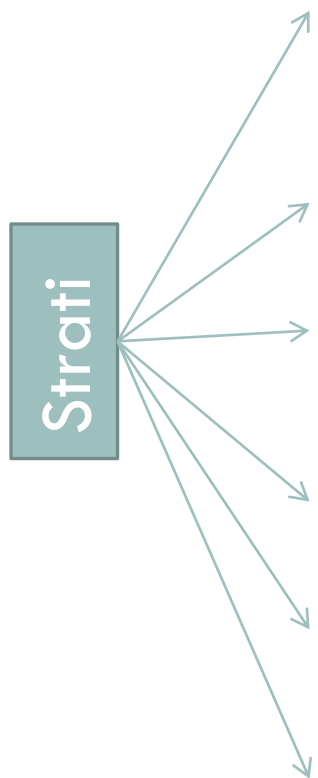
Popolazione residente in Italia, con età superiore ai 18 anni.

- Possibili liste da cui trarre il campione:
 - Anagrafi della popolazione
 - Liste elettorali
 - Censimento
- Non disponibili per sondaggi elettorali, d'opinione, ricerche di marketing ...
 - sono acquistate e utilizzate altre liste (di numeri telefonici e indirizzi mail), oppure sono generati numeri telefonici
 - le liste sono non esaustive e presumibilmente non rappresentative

Distorsione da selezione

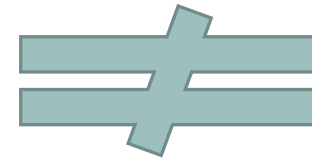
DISEGNO CAMPIONARIO STRATIFICATO VS. PER QUOTE

Suddivisione della popolazione per sesso età e titolo di studio



M <20 Alto	M 20-45 Alto	M >45 Alto
M <20 Medio	M 20-45 Medio	M >45 Medio
M <20 Basso	M 20-45 Basso	M >45 Basso
F <20 Alto	F 20-45 Alto	F >45 Alto
F <20 Medio	F 20-45 Medio	F >45 Medio
F <20 Basso	F 20-45 Basso	F >45 Basso

Stratificato: la popolazione è suddivisa in strati, da ogni strato è estratto un campione



Quote: è nota la quota della popolazione negli strati, e si cercano rispondenti finché non si riempiono tutti gli strati

DISTORSIONE DA NON RISPOSTA

*I rispondenti totali sono stati **1997 (su 13.287 contatti)***

Il tasso di non risposta è altissimo (85%), chi ha risposto (e riempito le quote) molto probabilmente non è rappresentativo di chi non ha risposto (lavoro diverso, orari diversi, rifiuto, ...)

TECNICA DI RILEVAZIONE

*La rilevazione è avvenuta con un'indagine attraverso i principali social network e con un **campione casuale** raggiungibile con i metodo **CAWI e CATI***

- CATI: Computer Assisted Telephone Interview;
- CAWI: Computer Assisted Web Interview

Rilevazione attraverso social network: pochissimi volontari, nulla di casuale