Indice

Campionamento non probabilistico

- In alcuni casi la reperibilità di liste delle unità da campionare è troppo dispendiosa in termini di denaro e/o di tempo (senza le liste non si può in genere usare un campione probabilistico) e non sempre è possibile ricorrere a quei tipi di campionamento probabilistico che possono essere adottati anche in assenza delle suddette liste, come il campionamento sistematico e quello a stadi.
- In tali casi si ricorre a tecniche non probabilistiche di formazione del campione.
- Si tratta di disegni di campionamento che pur usando alcune regole circa la scelta delle unità statistiche, consentono di scegliere tali unità 'a caso' (**non casualmente** come nel campionamento probabilistico)

Campionamento non probabilistico: motivazioni

- Nelle indagini campionarie accade spesso che una quota rilevante del campione risulta **irreperibile o rifiuta** di partecipare all'indagine (unit nonresponse).
  - Nel caso di questionari autosomministrati tale quota arriva anche al 40% del campione
  - Nelle indagini telefoniche è usuale una percentuale di unit nonresponse pari al 20%
  - Questo fenomeno dipende anche dalla composizione del campione e dalla natura dell'indagine (es. popolazioni a rischio o indagini su argomenti sensibili).
- Di solito, si ovvia a questo problema predisponendo un campione di riserva dal quale si attingono unità sostitute

Campionamento non probabilistico: motivazioni (2)

- Quando neanche il campione di riserva risolve il problema delle unit nonresponse allora si può ricorrere a disegni di campionamento non probabilistici

Naturalmente questa non è una ragione per abbandonare il campionamento probabilistico sempre e comunque. La selezione dei casi attraverso procedure oggettive garantisce che i ricercatori non scelgano gli intervistati nella cerchia dei loro conoscenti, o, comunque, tra le persone più facilmente reperibili, con gravi ed evidenti effetti distorsivi.
Questo accade invece spesso nei campioni non probabilistici, nonostante le regole più o meno rigide introdotte dai ricercatori circa la scelta delle unità statistiche.

Campionamento non probabilistico: caratteristiche e limiti

- Non vi è alcun riferimento agli aspetti probabilistici (non si può parlare di probabilità di selezione nè di probabilità di inclusione)
- Non vale la condizione di casualità
  - Ricordatevi che la scelta casuale del campione è la sola garanzia per conoscere le proprietà statistiche e l'affidabilità delle stime
  - I campioni non probabilistici riflettono in varia misura l'orientamento di chi li forma.
- Non consente una stima dell'errore di campionamento e può introdurre un errore di selezione dovuto alla non casualità del disegno di campionamento
- Non consente inferenza
  - I risultati valgono solo per il campione!

Campionamento per quote

- Il campionamento per quote è **analogo al campionamento stratificato**, proporzionale o non proporzionale
- Vi sono alcune differenze/analogie...

1. **Non occorre** la lista di campionamento
2. una volta divisa idealmente la target population in strati, le persone da intervistare in ogni strato vengono selezionate '**a caso**' (le prime che si incontrano)

Campionamento per quote: Come funziona

- È necessario conoscere la distribuzione della popolazione negli strati individuati da una variabile correlata con quella oggetto di studio (come nel campionamento stratificato)
- È necessario cioè conoscere le "quote" di popolazione che appartengono ai diversi strati
- Si fissa l'ampiezza del campione, quindi si calcolano le quote di soggetti da raggiungere in ogni strato
- Ad ogni intervistatore verranno assegnate quote di soggetti da intervistare, lasciandoli liberi di contattare chi credono, nel rispetto di tali quote
- Gli strati possono risultare da molteplici variabili.
- Le variabili di stratificazione più usate sono: il sesso, l'età (in classi), il titolo di studio, la dimensione del comune di residenza, e possono essere considerate contemporaneamente.

Campionamento per quote: Come funziona (2)

- Ad esempio: a ciascun intervistatore si assegnano 5 soggetti di sesso femminile, tra i 35 e 50 anni, laureati e residenti in comuni di 10000-50000 abitanti; 7 soggetti maschili con le stesse carattersitiche, ecc.
  Il campione complessivo naturalmente riprodurrà la distribuzione della popolazione nei differenti strati a seconda dell'allocazione desiderata (di solito in questi casi proporzionale)

Indice

Errore non campionario ed errore campionario

Ricordiamo che in un'indagine statistica, le stime oggetto possono essere soggette a due diverse fonti di errore: l'errore campionario e l'errore non campionario.
Ricordiamo gli errori non campionari più frequenti (`vedi slide Progettazione d'indagine: approccio in termini di qualità`)

- Errore di misurazione (Lato misurazione)
- Errore di copertura (Lato rappresentazione)
- Errore dovuto a mancate risposte (Lato rappresentazione)
- Errori in fase di aggiustamenti post-rilevazione (Lato rappresentazione e misurazione)

Parametri e...

In riferimento alla conoscenza della target population siamo interessati a stimare parametri ossia caratteristiche riassuntive della distribuzione della variabile di interesse nella popolazione
Ad esempio:

- se l'unità statistica è l'individuo e la variabile X è il reddito e vogliamo conoscere il reddito medio il parametro della popolazione che ci interessa è la media, μ
- se la variabile Y è la percentuale (proporzione) di voti per il PD, il parametro della popolazione che ci interessa è la proporzione, π
- Se vogliamo studiare la relazione tra etè reddito, il parametro della popolazione che ci interessa conoscere è il coefficiente di correlazione, ρ

... stimatori

Per ottenere un valore del parametro della popolazione a partire dal campione utilizziamo particolari funzioni dette stimatori

- più precisamente uno stimatore è una funzione che associa **ad ogni possibile campione un valore del parametro da stimare**
  - Ad es. per stimare la media μ (parametro) utilizziamo lo stimatore (funzione) media campionaria $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_1$ che applicata su un determinato campione ci fornisce una stima (valore) di μ che chiamiamo $\bar{x}$
- esistono numerosi stimatori per lo stesso parametro... ma solo alcuni godono di particolari proprietà
  - ad esempio μ (media della popolazione) può essere stimata anche calcolando la mediana campionaria ma le proprietà ottimali le ha solo la media campionaria!

... stimatori (2)

- anche se utilizziamo il migliore stimatore possibile per ciascun parametro però le stime che questo ci fornirà saranno **sempre** affette da errore campionario (dovuto al disegno di campionamento)
- Infatti in caso di campionamento probabilistico si può assumere che le stime ottenute sul campione sono uguali a quelle che si otterrebbero sulla popolazione a meno di un errore dovuto alla casualità, **che però sappiamo controllare**
- l'errore campionario esprime gli effetti determinati dal disegno di campionamento sulla distorsione e sulla varianza delle stime.

Componenti dell'errore campionario

L'errore campionario è costituito da:

1. distorsione (errore fisso)
   - dipende dal fatto che non tutte le unità statistiche nella lista di campionamento (frame population) sono osservate. È un problema solo se vi è una esclusione sistematica di alcune u.s. (es. probabilità di selezione prossima a zero di alcune u.s.)
2. varianza campionaria (errore variabile)
   - anche se forniamo a tutte le u.s. nella frame population la stessa probabilità di selezione, lo stesso disegno di campionamento produrrà differenti campioni e pertanto vi sarà una variabilità delle stime. Tale variabilità influenza la precisione delle stime

La precisione delle stime

- Se indichiamo con θ il valore incognito del parametro della popolazione, con $\hat{\theta}$ la stima ottenuta dal campione, con e l'errore campionario (errore variabile) abbiamo:

$$\theta = \hat{\theta} \pm \kappa_{\alpha/2} e$$

- cioè: Parametro = stima sul campione $\pm$ costante che dipende dal livello di confidenza (es. al 95% è 1,96) $\times$*errore*

Errore variabile

è funzione dell'errore standard dello stimatore (variabilità dello stimatore). l'errore standard (SE) è a sua volta funzione della varianza nella popolazione e della numerosità campionaria e dipende dal disegno di campionamento
$\Rightarrow$ alta varianza alto SE; alta numerosità campionaria basso SE

La precisione delle stime: media e proporzione

- Nel caso in cui il parametro di interesse sia μ la precisione della stima dipenderà dalla varianza $VAR(\bar{X}) = \frac{\sigma^2}{n}$ ossia dall'errore standard $SE(\mu) = \frac{\sigma}{\sqrt{n}}$:
- Nel caso in cui il parametro di interesse sia π la precisione della stima dipenderà dalla varianza $VAR(\hat{\pi}) = \frac{\hat{\pi}(1-\hat{\pi})}{n}$ ossia dall'errore standard $SE(\pi) = \sqrt{\frac{\bar{\pi}(1-\pi)}{n}}$

La precisione delle stime: media e proporzione (2)

Ma la varianza dipende anche dal **disegno di campionamento**.
Nel caso di campionamento casuale da popolazioni finite la varianza campionaria deve essere molitiplicata per un fattore di correzione che dipende dalla frazione di campionamento $f = n/N$:

$$VAR(\bar{X}) = (1 - f)\frac{\sigma^2}{n} \text{ e } VAR(\hat{\pi}) = (1 - f)\frac{\hat{\pi}(1-\hat{\pi})}{n}$$

il fattore di correzione può essere ignorato in caso di popolazioni grandi (o frazione di campionamento piccola), perché $(1 - f) \to 1$