Statistica

Introduzione al campionamento da popolazioni finite

Andrea Giommi

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università degli Studi di Firenze

Indagine

L'indagine è lo strumento mediante il quale si acquisiscono informazioni su uno o più fenomeni attinenti a una popolazione

- Indagine completa (censimento)

Semplice sul piano teorico ma complessa in pratica. Impossibile se:

  - Popolazione non finita
  - Osservazione distruttiva

- Indagine campionaria

Complessa sul piano teorico ma più facile da mettere in pratica

  - Costi limitati
  - Tempi ridotti
  - Numero elevato informazioni raccolte
  - Accuratezza nella rilevazione

Indagini sperimentali e osservazioneli

Altra importante distinzione è quella tra indagine sperimentale e osservazionale.

- Indagine sperimentale
  Studio dell'effetto sui soggetti di indagine dei diversi valori di una variabile, tenendo per quanto possibile sotto controllo altre variabili rilevanti

- Indagine osservazionale
  Osservazione di una o più variabili di studio sui soggetti di indagine senza possibilità di controllo sperimentale sugli stessi

Popolazione e Campione

Popolazione

Insieme finito o non finito di unità che non interessano prese singolarmente ma per il contributo che danno all'insieme di appartenenza

$\Downarrow$

Popolazione obiettivo

Elementi componenti

Estensione spaziale

Estensione temporale

$\Downarrow$

Lista

Popolazione e Campione (II)

Campione

Puó essere definito campione un qualsiasi sottoinsieme della popolazione

$$N \Rightarrow \text{Dimensione popolazione}$$
$$n \Rightarrow \text{Dimensione campione}$$

Importante distinzione:

- Campioni probabilistici
- Campioni non probabilistici

- Possibile enumerare i campioni estraibili
- Possibile assegnare ad ogni campione una probabilità
- Possibile assegnare ad ogni unità una prob. strettamente positiva
- Possibile disporre di un meccanismo per selezionare con le probabilità assegnate

Piano di indagine e Piano di campionamento

L'indagine è un insieme di fasi interrelate, complessivamente definite: *piano di indagine* (survey plan, survey design)

Piano di indagine:

- Definizione della popolazione obiettivo
- Scelta dei caratteri da studiare, modo di definirli e osservarli
- Scelta livelli spaziali e temporali di indagine
- Definizione del metodo di raccolta, codifica, elaborazione dei dati
- Definizione dei costi, livelli di precisione e di accuratezza desiderati
- Stime e altre analisi statistiche
- Metodologie di calcolo degli errori campionari
- Metodi di controllo e rilevazione e correzione degli errori non campionari
- Presentazione e diffusione dei risultati

Piano di campionamento

Il *piano di campionamento* riguarda la selezione del campione. Attualmente nella teoria del campionamento da popolazioni finite si definisce:

- *Schema di campionamento* (*schema di selezione*): L'insieme delle regole da seguire nella formazione del campione
- *Piano di campionamento*: la distribuzione di probabilità sull'universo dei possibili campioni S

Tale distribuzione si indica con l'espressione: $p(s)$ nella quale s è un generico campione appartenente all'insieme S

$$\{s\} = S,$$

e, naturalmente,

$$0 \leq p(s) \leq 1$$
$$\sum_{s \in S} p(s) = 1$$

Campioni non probabilistici

Tutti quelli che non rispettano le quattro condizioni viste in precedenza.
La maggior parte dei campioni nelle indagini osservazionali (spesso chiamate sondaggi) sono di natura non probabilistica

- Campioni di volontari o formati in modo spontaneo
- Campioni formati in modo fortuito
- Campioni a scelta ragionata
  - Campioni per quota
  - Campioni di unità tipo
  - Aree barometro
  - Campioni "a valanga"

Campioni probabilistici

- Sono fondamentali nella produzione delle statistiche ufficiali
- Sono i campioni prevalenti nell'ambito del SISTAN

Esempi notevoli di campioni probabilistici sono:

- La Rilevazione Continua delle Forze di Lavoro (ISTAT)
- L'indagine su consumi delle famiglie (ISTAT)
- L'indagine Multiscopo (ISTAT)
  - ✓ salute
  - ✓ aspetti della vita quotidiana
  - ✓ sicurezza
  - ✓ uso del tempo libero

Campioni probabilistici (II)

E come si forma un campione probabilistico (casuale)?

- Schema dell'urna (teorico)
- Tavole dei numeri casuali (la storia)
- Algoritmi matematici tradotti in routine informatiche

Altre classificazioni dei campioni probabilistici

- Campioni con replicazione (reimmissione, ripetizione)
- Campioni senza replicazione

E infine, quando possiamo dire che un campione è *rappresentativo*?
$\Rightarrow$ Rappresentativo è sinonimo di probabilistico (casuale)

Stima

La stima è il procedimento mediante il quale un valore (stima) ricavato come funzione (stimatore) delle osservazioni campionarie, viene assunto a rappresentare il valore incognito di una grandezza caratteristica (parametro) della popolazione obiettivo

Parametri di maggior interese nelle indagini osservazionali:

- Totali (occupati, forza lavoro)
- Medie (reddito pro-capite)
- Proporzioni (tasso di occupazione, attività)
- Rapporti (tasso di disoccupazione)
- Indici di variabilità (strumentali per la precisione degli stimatori)

Proprietà degli stimatori

Sul piano intuitivo, vorremmo che la stima (valore numerico dello stimatore) fosse *prossima* al valore numerico del parametro (incognito) o *coincidesse* con il parametro.

- Definiamo:

$$\begin{array}{ccl} \mathrm{T} & \Rightarrow & \text{Stimatore} \\ t & \Rightarrow & \text{Stima} \\ \theta & \Rightarrow & \text{Parametro} \end{array}$$

- Definiamo inoltre:

$$\mathrm{D} = t - \theta \quad \Rightarrow \quad \text{Errore di stima}$$

- Situazione ideale: D = 0

Proprietà degli stimatori (II)

- D non può essere azzerato nell'indagine campionaria, ma è = 0 nei censimenti;
- Come è possibile ridurlo nell'indagine campionaria?
  - Dimensione campionaria
  - Piano di campionamento

Dimensione Campionaria

$$\Uparrow \qquad \qquad$$
$$\boldsymbol{n} \quad \mathbf{D}$$
$$\qquad \qquad \Downarrow$$

Attenzione!

n non può essere aumentato liberamente (costi vs risorse)

Campioni probabilistici

(i) Campionamento casuale semplice
(ii) Campionamento casuale stratificato
(iii) Campionamento sistematico
(iv) Campionamento a grappoli e a più stadi

Campionamento casuale semplice (CCS)

Popolazione:

$$U_1 \quad U_2 \quad U_3 \quad U_4$$

$N = 4$; $n = 2$

Possibili campioni $\{s\}$:

$$\begin{array}{ccc} (U_1 \ \ U_2) & (U_1 \ \ U_3) & (U_1 \ \ U_4) \\ (\ U_2 \ \ U_3) & (U_2 \ \ U_4) & (U_3 \ \ U_4) \end{array}$$

Numero campioni (cardinalità S): $C_{4,2} = 6$ CCS $\Rightarrow$ p(s)=$\frac{1}{6}$

Stima della media nel CCS

Y Carattere di studio nella popolazione

$\bar{Y}$ Media del carattere Y nella popolazione

$\bar{y}$ Stima campionaria di $\bar{Y}$

Valori nella popolazione:

$$Y_1 = 20, \quad Y_2 = 40, \quad Y_3 = 36, \quad Y_4 = 48, \quad (\bar{Y} = 36)$$

Possibili campioni

$$\begin{array}{ccc} (Y_1\,Y_2) & (Y_1\,Y_3) & (Y_1\,Y_4) \\ (Y_2\,Y_3) & (Y_2\,Y_4) & (Y_3\,Y_4) \end{array}$$

Stima della media nel CCS II

Media campionaria per ciascuno dei 6 possibili campioni:

$$\begin{array}{ccc} (20,\ 40) & (20,\ 36) & (20,\ 48) \\ (40,\ 36) & (40,\ 48) & (36,\ 48) \end{array}$$

Medie campionarie

30	28	34
38	44	42

Media popolazione =36

- Nessuna media campionaria è uguale alla media della popolazione
- La media di tutte le medie è 36 (media: stimatore corretto)
- Errori di stima: come valutarli nel loro complesso?

Precisione dello stimatore

- La qualità dello stimatore è data dalla sua: precisione
- La precisione è il **reciproco** della varianza dello stimatore o della radice quadrata della varianza che prende il nome di errore standard dello stimatore
- **Maggiore** la varianza (errore standard) **minore** la precisione e viceversa

$$\begin{array}{rcl} \text{Varianza} & \longmapsto & V(\bar{y}) \\ \text{Errore standard} & \longmapsto & \sqrt{V(\bar{y})} = ES(\bar{y}) \end{array}$$

Precisione dello stimatore: applicazione

$$V(\bar{y}) = \frac{\sum_i (\bar{y}_i - \bar{Y})^2}{C_{N,n}} =$$

$$= \frac{(30-36)^2 + (28-36)^2 + ... + (42-36)^2}{6} = 34,67$$

$$V(\bar{y}) = 34,67 \quad \text{e} \quad ES(\bar{y}) = \sqrt{V(\bar{y})} = 5,89$$

$ES(\bar{y}) = 5,89 \quad \Rightarrow \quad ?$

Precisione e dimensione campionaria

$$n = 3$$

- Possibili campioni

$$(U_1\, U_2\, U_3) \quad (U_1\, U_2\, U_4)$$
$$(U_1\, U_3\, U_4) \quad (U_2\, U_3\, U_4)$$

- Valori corrispondenti

$$(20, 40, 36) \quad (20, 40, 48)$$
$$(20, 36, 48) \quad (40, 36, 48)$$

- Medie campionarie

$$32 \quad 36 \quad 34{,}67 \quad 41{,}33$$

- Media delle medie ancora uguale a 36; ma varianza ed errore standard minori:

$$V(\bar{y}) = 11,55 \quad ES(\bar{y}) = 3,4$$

Formula per la varianza dello stimatore

- Varianza elementare:

$$S^2 = \frac{\sum_i (Y_i - \bar{Y})^2}{N-1}$$

- Varianza dello stimatore:

$$V(\bar{y}) = \frac{S^2}{n}(1 - \frac{n}{N})$$

- Nell'esempio con $n = 3$

$$S^2 = ((20-36)^2 + (40-36)^2 + (36-36)^2 + (48-36)^2)/3 = 138,67$$

$$V(\bar{y}) = \frac{138,67}{3}(1 - \frac{3}{4}) = 11,55$$

Stima della varianza dello stimatore

- Nella pratica la vera varianza dello stimatore non può essere calcolata, dato che non è nota la varianza elementare S^2. Tuttavia è possibile stimarla dal campione inserendo nella sua formula la stima da campione della varianza elementare

- Stima della varianza elementare:
$$s^2 = \frac{\sum_i (y_i - \bar{y})^2}{n-1}$$

- Stima della varianza dello stimatore:
$$v(\bar{y}) = \frac{s^2}{n}(1 - \frac{n}{N})$$

Stima di una proporzione

- La proporzione, π, di unità che, nella popolazione, hanno un particolare attributo, può essere vista come la media di carattere (dicotomico) che assume valore 1 in presenza dell'attributo e valore 0 in sua assenza.

- Di conseguenza la proporzione campionaria, che possiamo indicare con p o $\hat{\pi}$, ha le stesse proprietà dalla media campionaria.
  In particolare, è uno stimatore corretto della proporzione nella popolazione

Proprietà formali dello stimatore media nel CCS

- Correttezza

$$E(\bar{y}) = \mu$$

con μ = media del carattere Y nella popolazione

- Efficienza

$$V(\bar{y}) \leq V(T)$$

con T costante diversa dalla media aritmetica (es: mediana del campione)

- Consistenza

$$V(\bar{y}) \to 0$$

al crescere della dimensione campionaria

Campionamento casuale stratificato

1. Suddivisione della popolazione in sottopopolazioni (STRATI)
2. Estrazione di campioni (casuali) indipendenti da ogni strato

Obiettivi

(a) Stimatori con elevata precisione

(b) Dominii di studio

$$
\begin{array}{ccll}
N_h & \Rightarrow & \sum_h N_h = N; & h = (1, 2..., H) \\
n_h & \Rightarrow & \sum_h n_h = n; & \\
W_h = N_h/N & \Rightarrow & \sum_h W_h = 1; &
\end{array}
$$

Stratificazione

- Parametro da stimare

$$\bar{Y} = \sum_{h=1}^{H} W_h \bar{Y}_h$$

- Stimatore

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

- Varianza

$$V(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 V(\bar{y}_h) = \sum_{h=1}^{H} W_h^2 \frac{S_h^2}{n_h}(1 - \frac{n_h}{N_h})$$

Stratificazione proporzionale

- Frazione di campionamento costante

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f; (h = 1, ..., H)$$

- Vantaggi
  - (i) Per la popolazione generale, stime più precise rispetto al CCS
  - (ii) Facilità di applicazione (con un numero limitato di strati)
- Svantaggi
  - (i) Precisione diversa per strati di diversa dimensione
  - (ii) Difficoltà di applicazione con molti strati

Stratificazione proporzionale: precisione dello stimatore

- Stimatore della media

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

- Varianza dello stimatore

$$V(\bar{y}_{stp}) = \frac{(1-f)}{n} \sum_{h=1}^{H} W_h S_h^2 = \frac{(1-f)}{n} S_w^2$$

- Stimatore campionario della varianza

$$v(\bar{y}_{stp}) = \frac{(1-f)}{n} \sum_{h=1}^{H} W_h s_h^2$$

Stratificazione non proporzionale

- Stratificazione ottimale

$$n_h \propto \frac{N_h S_h}{\sqrt{c_h}} \Rightarrow f_h \propto \frac{S_h}{\sqrt{c_h}}$$

(c_h = costo di rilevazione unitario nello strato h)

- Uguale precisione in ogni strato

Se possiamo ipotizzare che:

$$S_1 \cong S_2 \cong ... \cong S_H$$

Allora:

$$n_1 = n_2 = ... = n_H$$

Campionamento Sistematico

- Intervallo di selezione

$$k = \frac{N}{n}$$

Esempio: $N = 1500 \quad n = 100$

$$k = \frac{1500}{100} = 15$$

- Numero casuale compreso tra 1 e 15; supponiamo 6
- Campione sistematico:

$$6 \quad 6 + 15 \quad 6 + 2x15 \quad \ldots \quad 6 + 99x15$$

Cioè:

$$6 \quad 21 \quad 36 \quad \ldots \quad 1491$$

Campionamento a grappoli e a più stadi

✓Grappoli Campionamento di aggregati (*unità di rilevazione*): tutte le unità componenti l'aggregato (*unità di studio*) entrano a far parte del campione

✓Stadi Campionamento di grappoli (o unità di studio) da grappoli di livello gerarchico superiore

Due ordini di motivazioni:

- Indisponibilità della lista delle unità di studio
- Costi

Attenzione!

✓Strati $\Rightarrow$ omogenei al loro interno
✓Stadi $\Rightarrow$ eterogenei al loro interno

Confronto: Grappoli vs CCS

- Nella pratica operativa: grappoli omogenei al loro interno (es: famiglie, scuole, classi scolastiche, ecc.)
- Grappoli: la dimensione del campione risulta variabile in termini di unità di studio
- Grappoli vs CCS: a parità di dimensione (CCS, fissa; grappoli, attesa) Il CCS fornisce stime più precise, ma ad un costo ben superiore
- A parità di costi (risorse) il campionamento a grappoli può avere dimensione mediamente maggiore e conseguentemente produrre stimatori più precisi rispetto al CCS