Statistica per l'impresa

2.6 Le indagini campionarie

Le indagini campionarie

Quando non è possibile o pratico indagare l'intera popolazione: indagine *campionaria*

- Popolazione oggetto di indagine: *universo*
- Campione osservato

- Unità di rilevazione
- Unità di analisi

Le indagini campionarie - fasi

Fasi dell'indagine campionaria:

1. Definizione degli obiettivi
2. Identificazione della popolazione di riferimento
3. Scelta dei criteri di selezione (*piano di campionamento*)
4. Scelta dei metodi di stima
5. Scelta delle modalità di raccolta dei dati
6. Messa a punto del questionario
7. Organizzazione della fase di rilevazione dei dati
8. Valutazione dei costi di realizzazione dell'intera analisi

Popolazione e campione

Popolazioni:

- finite
- infinite

- Ci occupiamo di popolazioni *finite* di numerosità N
- La *lista di campionamento* definisce una *popolazione di selezione*. Esempi di liste di campionamento:
  - anagrafe
  - liste elettorali
  - elenchi telefonici
  - registro delle imprese
  - ...

  Problema: la *popolazione di selezione* può non corrispondere alla *popolazione obiettivo*!

Problemi di selezione

Se la popolazione di selezione non *rappresenta* quella obiettivo, le informazioni del campione non possono essere generalizzate a quest'ultima.

- Non corrispondenza tra pop. di selezione e pop. obiettivo
- Non-risposta (dropout, rifiuto)

La popolazione effettivamente rappresentata dalle unità sottoposte a indagine è detta *popolazione di indagine*.

Formazione del campione

Campioni probabilistici:

- sono selezionati con meccanismo casuale
- è nota (e non nulla) la probabilità di estrarre ciascuna unità

Sia Y una variabile aleatoria che rappresenta il carattere di interesse *nella popolazione*, che supponiamo di $N < \infty$ unità. Si estrae il campione casuale:

$$\{Y_1, Y_2, \ldots, Y_n\}$$

Si definiscono:

- dimensione campionaria n: il numero di unità che compongono il campione
- frazione di campionamento: $f = \frac{n}{N}$

Stima di parametri

Supponiamo di voler stimare un parametro ignoto Θ (una media, un totale, una proporzione...) di Y utilizzando il campione estratto: la stima sarà data dal valore della statistica

$$t_n = f(y_1, y_2, \ldots, y_n)$$

calcolata sui valori y_i osservati nel campione

In generale, per ottenere una stima da un particolare campione utilizzeremo una *funzione dei dati campionari* detta *stimatore*. Il corrispondente di t_n sarà:

$$T_n = f(Y_1, Y_2, \ldots, Y_n)$$

Stimatori e stime

Stimatore vs. stima:

- *stimatore*: una variabile aleatoria, funzione del campione (una "formula")
- *stima*: il valore assunto dallo stimatore in corrispondenza del campione osservato (un "numero")

Proprietà desiderabili di uno stimatore:

- correttezza
- efficienza
- (*consistenza*)

Le proprietà dello stimatore dipenderanno (anche) dal piano di campionamento!

Schemi di campionamento probabilistico

Campionamento casuale semplice:

- ogni unità i ha probabilità $\pi_i = \frac{n}{N}$ di essere inclusa

Campionamento sistematico:

- se le unità ammettono un ordine, è selezionato casualmente un intero j in $1, k = int(N/n)$ poi altre $(n-1)$ unità sono scelte ogni k unità
- La probabilità di inclusione dipende dalla variabile ordinante. (*Box: MUS*)

Campionamento stratificato:

- presuppone che nella popolazione la caratteristica di interesse sia correlata a una serie di altri attributi
- le unità vengono classificate in base a questi, ottenendo delle sottopopolazioni "omogenee" (*strati*) da cui estrarre campioni casuali

Campionamento a grappoli e a stadi

Campionamento stratificato

Denotiamo $h = 1, \ldots, H$ il singolo strato, dove H è il numero degli strati

- N_h sarà la dimensione del singolo strato h *nella popolazione*
- $W_h = \frac{N_h}{N}$ la proporzione della popolazione dello strato h nella popolazione totale
- n_h la dimensione del campione estratto dallo strato h

Risulta pertanto $\sum_{h=1}^{H} W_h = 1$ e $\sum_{h=1}^{H} n_h = n$. Si può adottare una

- stratificazione proporzionale: $f_h = \frac{n_h}{N_h} = \frac{n}{N}$
- stratificazione non proporzionale: gli strati della popolazione meno numerosi, o quelli dalla variabilità più elevata, vengono sovrarappresentati

Campioni non probabilistici

I campioni sono selezionati in base a considerazioni di ordine pratico

- Campionamento di comodo: scelta arbitraria
- Campionamento ragionato: selezione non casuale basata su informazioni a priori (es. paniere di prodotti, o interviste a opinion leaders)
- Campionamento per quote: la popolazione viene suddivisa in base a caratteristiche note, come nel campionaento stratificato, ma poi si campiona in modo non casuale da ogni quota

Soluzioni subottimali che non offrono alcuna garanzia di rappresentatività della popolazione. E' molto probabile che qualche tipo di involontaria selezione renda la popolazione di indagine diversa dalla popolazione obiettivo.

Stima dei parametri della popolazione - la media

Stima della media della popolazione sulla base di un campione casuale semplice:

$$T_{\bar{Y}} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

detto *media campionaria*

- è uno stimatore corretto ($E(\bar{Y}) = \mu$)
- qual è la sua efficienza?

Siamo interessati anche alla *precisione* dello stimatore (che, ricordiamo, è una variabile aleatoria). La precisione sarà inversamente proporzionale alla *variabilità*, misurata dall'*errore standard* $ES(T_{\bar{Y}})$. A sua volta $ES(T_{\bar{Y}})$

dipende dalla varianza di Y nella popolazione, σ^2_Y, che però è ignota. Pertanto per calcolare l'errore standard della media campionaria abbiamo bisogno di stimare la varianza della caratteristica nella popolazione.

Stima dei parametri della popolazione - la varianza

La *varianza campionaria*

$$S^2_{\bar{Y}} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

è uno stimatore corretto di σ^2_Y
L'errore standard della media campionaria (nel caso di CCS senza ripetizione) è pari a

$$ES(T_{\bar{Y}}) = \sqrt{(1-f)\frac{S_{\bar{Y}}}{n}}$$

La precisione dello stimatore $T_{\bar{Y}}$

- aumenta al diminuire di $(1-f)$
- aumenta al diminuire di σ^2_Y
- aumenta al crescere di n

Cosa succede se $f \to 1$? E perché?

Stima dei parametri della popolazione - la proporzione

La proporzione $\pi = N_K/N$ di individui che nella popolazione presentano la caratteristica k può essere stimata attraverso l'equivalente proporzione nel campione (es. *exit poll*).
Posta Z la variabile dicotomica che vale 1 se k, 0 altrimenti, lo stimatore della proporzione

$$T_P = \frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{n_k}{n}$$

non è che la media campionaria di Z, ovvero la *proporzione campionaria* che presenta la caratteristica in esame.

Altra applicazione del principio di analogia

Stima della varianza della proporzione campionaria

L'errore standard della proporzione campionaria (nel caso di CCS senza ripetizione) è pari a

$$ES(T_P) = \sqrt{(1-f)\frac{\pi(1-\pi)}{n-1}}$$

ma π non è noto. Si può alternativamente:

- porre $\pi = p$, usando la proporzione campionaria come stimatore (la cosa più naturale)
- porre $\pi = 0.5$, che è il punto di massimo di $Var(T_P)$ (la scelta più conservativa)

Osservazione

Se in una popolazione normalmente distribuita stimo "bene" media e varianza, ho descritto l'intera distribuzione.

Stima della media nel campionamento stratificato

La media campionaria:

$$T_{\bar{Y}_{ST}} = \sum_{h=1}^{H} W_h \bar{Y}_h = \sum_{h=1}^{H} W_h \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}$$

è la media *ponderata con i pesi degli strati* W_h delle medie di ciascuno strato. L'errore standard: $ES(T_{\bar{Y}_{ST}}) = \sqrt{\sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{S_h}{n_h}}$ dove

- N_h dimensione dello strato *nella popolazione*
- $W_h = \frac{N_h}{N}$ proporzione dello strato
- n_h proporzione del campione estratto dallo strato h
- $f_h = \frac{n_h}{N_h}$ frazione di campionamento dello strato h
- $S_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (Y_{ih} - \bar{Y}_h)^2$ stimatore di $Var(\bar{Y})$ nello strato h

Stima della proporzione nel campionamento stratificato

La proporzione, analogamente:

$$T_{P_{ST}} = \sum_{h=1}^{H} W_h P_h = \sum_{h=1}^{H} W_h \frac{r_h}{n_h}$$

con r_h il numero degli elementi dello strato che presentano la caratteristica in esame. L'errore standard della proporzione, nel caso di campionamento stratificato, viene derivato dalla varianza della popolazione del singolo strato h:

$$S^2_{P_h} = \frac{n_h \pi(1-\pi)}{n_h - 1}$$

da cui, sostituendo la stima P_{ST} per π,

$$ES(T_{P_{ST}}) = \sqrt{\sum_{h=1}^{H} W_h^2 (1-f_h) \frac{P_h(1-P_h)}{n_h(n_h-1)}}$$

Proporzionalità e campionamento stratificato

Se la proporzione campionaria dei singoli strati replica la proporzione degli strati nella popolazione, ovvero

$$f_h = \frac{n_h}{n} = \frac{n}{N}$$

si può dimostrare che le statistiche del campionamento stratificato coincidono con quelle del campionamento casuale semplice.

Altrimenti (ovvero: se alcuni strati sono sovra/sottorappresentati) non è così.

Stime intervallari

Ci siamo occupati finora di stime *puntuali*.

Conoscere (stimare!) la *varianza* dello stimatore ci permette di valutare la *precisione* della stima.

- Costruiamo attorno alla stima puntuale un *intervallo di confidenza* $I_{1-\alpha}$, entro il quale il "vero" valore del parametro nella popolazione cadrà con probabilità $1 - \alpha$.
- A questo scopo serve fare ipotesi sulla *distribuzione* dello stimatore

In generale, per una statistica T con distribuzione (simmetrica) Υ, l'intervallo di confidenza al livello $1 - \alpha$ è dato da

$$T \pm \tau_{\frac{\alpha}{2}} ES(T)$$

dove $\tau_{\frac{\alpha}{2}}$ è il percentile della distribuzione Υ tale per cui il 95% della distribuzione risulta compreso tra $-\tau_{\frac{\alpha}{2}}$ e $\tau_{\frac{\alpha}{2}}$.

Stime intervallari - esempio

Come sono distribuiti i nostri stimatori? Per "grandi" campioni (sul libro $N > 30$, ma è un po' ottimistico) il CLT dice che è possibile approssimare la distribuzione della media campionaria alla Normale *a prescindere dalla distribuzione della popolazione*.

In questo caso, $z_{\frac{0.05}{2}} = 1.96$ e l'intervallo di confidenza per la media campionaria al livello $1 - \alpha = 95\%$ è dato da

$$I_{\bar{Y},95\%} = T \pm 1.96 ES(T_{\bar{Y}})$$

Nel caso di piccoli campioni estratti da popolazioni normali, qualora (come in genere accade) si debba stimare la varianza, lo stimatore avrà distribuzione *t di Student* con $n - 1$ gradi di libertà, dove n è la grandezza del campione. I percentili della t saranno sostituiti a quelli della Normale. Essi sono in generale leggermente più grandi ma $t \to N$ se $n \to \infty$.

Errore campionario

La varianza dello stimatore è legata alla dimensione del campione (tende a 0 quando $f \to 1$). Definiamo *errore campionario* la semiampiezza dell'intervallo di confidenza:

$$e = \tau_{\frac{\alpha}{2}} ES(T)$$

Ribaltando quanto sopra, possiamo fissare e e determinate di quante unità campionarie abbiamo bisogno.

Nel caso di $\bar{Y}$, ipotizzando un campione "grande" e $(1 - f) \sim 1$ (popolazione "grandissima"), si ottiene

$$e = z_{\frac{\alpha}{2}} \frac{\sigma_Y}{\sqrt{n}}$$

da cui – risolvendo per n – si ha la dimensione campionaria necessaria per un prefissato livello di e, fissato α e data la varianza σ^2_Y dello stimatore.

Tecniche di rilevazione

- Intervista diretta (rapporto umano permette spiegazioni ma crea anche barriere emotive e permette arbitrarietà)
- Autocompilazione dei questionari (se l'intervistato è affidabile e collaborativo)
- Intervista telefonica (economica, automatizzabile, ma bassi tassi di copertura e possibile selezione)
- Indagine web (molto economica e facile da somministrare; seri effetti di selezione)

Il questionario - Formulazione delle domande

- Domande chiuse (costringono l'intervistato a scegliere tra modalità predefinite; facili da processare ma limitanti per l'espressione del pensiero dell'intervistato)
- Domande aperte (permettono all'intervistato di esprimere compiutamente il proprio pensiero; richiedono tempi lunghi di elaborazione e digitalizzazione)
- Domande filtro (domande chiuse che hanno la funzione di indirizzare l'intervistato verso una o l'altra sezione successiva del questionario)

Il questionario - Misurazione delle risposte

La misurazione delle modalità di risposta è un aspetto cruciale dell'indagine. In alcuni casi le risposte sono già espresse in unità di misura (Kg, anni...); altrimenti - percezioni, atteggiamenti, opinioni e caratteri qualitativi in genere - la misurazione può avvenire con l'ausilio di *scale ad hoc*.

- La *scala nominale* si limita a codificare caratteri qualitativi non ordinabili (M/F, colori...)
- Dalle *scale ordinali* si può ricavare un ordinamento naturale (es. *buono* $>$ *cattivo*, *Alfa Romeo* $>$ *Fiat*) ma non si possono calcolare differenze, distanze, valori medi; si possono calcolare, peraltro, mediane e quantili.
- La *scala a intervallo* permette di definire sia l'ordine che una *distanza* tra modalità diverse, come ad esempio in *1=molto negativo, 2=negativo, 3=neutro, 4=positivo, 5=molto positivo*.

Valutazione dei risultati

L'indagine perfetta sarebbe affetta esclusivamente da *errore campionario*, dipendente dalla natura non esaustiva dell'indagine campionaria.

Imperfezioni nel processo danno luogo a errori di diverso tipo, detti *errori non campionari*. Questi vengono distinti in:

- *errori di copertura* causati da difetti nella lista di campionamento
- *errori da mancata risposta* dovuti all'impossibilità di osservare parte delle unità campionarie
- *errori di misurazione*