La stima della frequenza di risposte casualizzate

mediante domanda incorrelata

Un argomento che potrebbe essere trattato come tematica di approfondimento in una Unità Didattica dedicata al Calcolo delle Probabilità è una particolare applicazione dei teoremi delle probabilità totali e composte, nota come "Metodo delle risposte casualizzate a domanda incorrelata" o "Schema di Simmons". Tale tecnica viene sovente utilizzata in indagini di statistica sociale e mira a ricavare informazioni attendibili su argomenti delicati (ad es., uso di droga, attitudini sessuali, redditi effettivi, ecc.).

Schema operativo

Si supponga che $\mathbf{P_X}$ sia la frequenza con cui una caratteristica delicata **X** sia presente in una data popolazione **U**; mentre $\mathbf{P_K}$ è la frequenza con cui si presenta nella stessa popolazione un attributo **K**, incorrelato con **X**, che può essere dichiarato senza alcuna riserva (ad es. **K** è un numero telefonico che finisce con una certa cifra, il colore dei capelli o degli occhi, …).

Lo schema di rilevazione prende le mosse consegnando all'intervistato un'urna contenente un certo numero di palline bianche e nere e gli si chiede di rispondere (segnandolo su una scheda) con un "sì" oppure con un "no" alla domanda delicata, se dall'urna pesca pallina bianca; mentre risponderà con un "sì" oppure con un "no" al possesso dell'attributo **K,** se dall'urna pesca pallina nera. Poiché solo l'intervistato conosce il risultato dell'estrazione dall'urna, solo lui conosce se il "sì" oppure il "no" è attribuito alla domanda delicata **X** oppure all'attributo **K**: ciò dovrebbe indurre l'intervistato a non dichiarare il falso, in quanto, essendo noto solo a lui il significato del "sì" o del "no" dato, ha la garanzia di non palesare la sua situazione personale.

Metodologia di applicazione

Ipotizziamo di utilizzare per la nostra rilevazione un'urna contenente 12 palline bianche e 6 nere e che venga richiesto ai nostri intervistati di rispondere al quesito delicato *"fai o hai mai fatto uso di stupefacenti?",* se pesca dall'urna pallina bianca; e alla domanda *"il tuo numero di telefono finisce con la cifra 5?"*, se pesca pallina nera.

Indicando con:

$\boldsymbol{\lambda}$ = 2/3 (P_B) la probabilità di estrarre dall'urna di pallina bianca;

1- $\boldsymbol{\lambda}$ = 1/3 (1-P_B) la probabilità di estrarre pallina nera;

$\mathbf{P_X}$ (incognita) la probabilità di rispondere "si" alla domanda delicata (di possedere la caratteristica X)

$\mathbf{1-P_X}$ (incognita) la probabilità di non possedere la caratteristica X;

$\mathbf{P_K}$ = 1/10 = 0,1 la probabilità che il proprio numero telefonico termini con la cifra 5;

$\mathbf{1-P_K}$ = 0,9 la probabilità che il proprio numero telefonico **non** termini con la cifra 5;

si può impostare il cosiddetto **Schema di Simmons**, che fornisce le possibili risposte, in relazione al gruppo di appartenenza e al risultato dell'esperimento aleatorio (nel caso di campionamento casuale):

Schema di Simmons

Eventi Composti	Probabilità	Risposta alla domanda "delicata" Gruppo che ha estratto pallina bianca $\lambda = 2/3$	Risposta alla domanda sul n° tel. Gruppo che ha estratto pallina nera $1 - \lambda = 1/3$ [(1)]
$X \cap K$	$P_X \bullet P_K$	SI	SI
$X \cap \overline{K}$	$P_X \bullet (1 - P_K)$	SI	NO
$\overline{X} \cap K$	$(1 - P_X) \bullet P_K$	NO	SI
$\overline{X} \cap \overline{K}$	$(1 - P_X) \bullet (1 - P_K)$	NO	NO

(1) Attenzione a non confondere questa probabilità (prob. di estrarre pallina nera - risposta alla domanda **non** "delicata"), con quella di avere il n° di telefono che termina con 5 (= 1/10 = 0,1).

In cui:

X $\Rightarrow$ Possiede la caratteristica delicata X (uso di droga);

$\overline{X}$ $\Rightarrow$ Non possiede la caratteristica X (non uso di droga);

K $\Rightarrow$ Possiede l'attributo K (n° tel. termina con 5);

$\overline{K}$ $\Rightarrow$ Non possiede l'attributo K (n° tel. non termina con 5).

Applicando congiuntamente i principi delle probabilità totali (per eventi incompatibili) e composte (per eventi indipendenti), dalla tabella (nell'incrocio righe/colonne in cui compare la risposta "si") si ricava la probabilità, indicata con **ρ**, che in una prova la risposta alla domanda delicata X o al possesso dell'attributo K, sia positiva:

$$\rho = \lambda \bullet P_X \bullet P_K + \lambda \bullet P_X \bullet (1 - P_K) + (1-\lambda) \bullet P_X \bullet P_K + (1-\lambda) \bullet (1 - P_X) \bullet P_K;$$

dalla quale, svolgendo i prodotti e semplificando, si ottiene:

$$\rho = \lambda \bullet P_X + (1-\lambda) \bullet P_K \Rightarrow \lambda \bullet P_X = \rho - (1-\lambda) \bullet P_K \Rightarrow$$

$$\Rightarrow \quad P_X = \frac{\rho - P_K(1-\lambda)}{\lambda};$$

Se si indica ora con **x** il numero delle risposte affermative ottenute su **n** unità campionarie, la frequenza **f = x/n** delle risposte positive sul totale degli intervistati, può essere considerata una stima del valore teorico **ρ**. Sostituendo, quindi, il valore di **f** nella relazione precedente si ottiene la corrispondente stima (corretta e consistente) della probabilità $\mathbf{P_X}$ di risposta affermativa alla domanda delicata (nel nostro es. uso di droga):

$$\hat{P}_X = \frac{f - P_K(1-\lambda)}{\lambda}; \quad \textbf{(*)}$$

in cui $\mathbf{P_K}$ e **λ** sono costanti note.

Osservazione

La distribuzione campionaria di $\mathbf{P_X}$ **,** al crescere di **n**, tende alla normalità con media **ρ** e varianza $\sigma^2(\hat{P}_X) = \frac{\rho(1-\rho)}{n\lambda^2}$; una cui stima, tenuto conto che **f** è una stima di **ρ**, è data da:

$$\hat{S}^2(\hat{P}_X) = \frac{f(1-f)}{(n-1)\lambda^2}$$

Se si utilizza lo stesso schema di estrazione, la quantità **(1- λ) • $\mathbf{P_K}$** diventa una costante e la stima della probabilità di risposta "si" a un qualsivoglia quesito delicato si può ottenere in modo immediato, sostituendo nella (*) il valore della frequenza delle risposte positive rilevate.

Esempio

Utilizzando lo schema di estrazione precedente (un'urna contenente 6 palline nere e 12 bianche), nell'ipotesi che siano state rilevate x = 50 risposte affermative, su un totale di n = 120 intervistati, si ha:

λ = 12/18 = 0.67 ; **1-λ** = 6/18 = 0.33;

f = **x/n** = 50/120 = 0.42; $\mathbf{P_K}$ = 1/10 = 0.1.

La stima di P_X sarà, quindi*:*

$$\hat{P}_X = \frac{0.42 - 0.1 * 0.33}{0.67} = \frac{0.42 - 0.033}{0.67} = 0.575;$$

con varianza stimata:

$$\hat{S}^2(\hat{P}_X) = \frac{0.42(1-0.42)}{119 * 0.67^2} = \frac{0.2436}{53.42} = 0.005;$$

ovvero, la probabilità di presenza della caratteristica delicata X (nel nostro caso uso di droga) nel nostro campione è del 57.5%, con una variabilità dello 0.5%.

Da notare che la stessa frequenza di risposte "si" (ad es. f = 50/120 = 0.42), conduce ad una stima differente di P_X (prob. di possedere la caratteristica delicata X), se viene cambiato lo schema di estrazione. Infatti, se cambia la composizione dell'urna, cambiano, di conseguenza, anche le probabilità di rispondere all'uno o all'altro quesito (caratteristica delicata o attributo incorrelato); se, invece, cambia la probabilità di risposta al quesito sul possesso dell'attributo K, varia il peso di tale probabilità nel calcolo della sima di P_X.

Esempio

- Se cambia la composizione dell'urna:

  4 palline nere – 6 bianche

  λ = 6/10 = 0.6; **1-λ** = 4/10 = 0.4;

  f = **x/n** = 50/120 = 0.42; $\mathbf{P_K}$ = 1/10 = 0.1;

$$\hat{P}_X = \frac{0.42 - 0.1 * 0.4}{0.6} = \frac{0.42 - 0.04}{0.6} = 0.63\ ;$$

  ovvero, il 63% possiede la caratteristica X.

- Se cambia la probabilità dell'attributo K; ad esempio, in caso di estrazione di pallina nera, rispondere alla domanda: "hai gli occhi azzurri?"; si ha:

  λ = 12/18 = 0.67; **1-λ** = 6/18 = 0.33;

  f = **x/n** = 50/120 = 0.42; $\mathbf{P_K}$ = 1/4 = 0.25;

$$\hat{P}_X = \frac{0.42 - 0.25 * 0.33}{0.67} = \frac{0.42 - 0.0825}{0.67} = 0.504\ ;$$

  cioè il 50,4% possiede la caratteristica X.

Osservazione

Si può verificare (anche empiricamente) che, utilizzando lo stesso schema di estrazione, , da un certo "livello soglia" di frequenze di risposte "si", la probabilità che la popolazione di intervistati possegga la caratteristica "delicata" X è sicuramente pari ad Uno (100%). Nell'esempio dello schema di estrazione qui utilizzato, se il n° di risposte "si" è prossimo al 70%, possiamo essere certi che la caratteristica X è posseduta da tutti gli intervistati:

$$\hat{P}_X = \frac{0.7 - 0.1 * 0.33}{0.67} = \frac{0.7 - 0.033}{0.67} = 1\ .$$