Randomized Response Model Technique.

Assegno di Ricerca area scientifico/disciplinare SO1A-Statistica R 200/01.

Massimo Guerriero
massimo.guerriero@univr.it

30 aprile 2002

Indice

Elenco delle tabelle

1 Introduzione

Il campionamento Randomized Response (Risposta Casualizzata) è una tecnica introdotta nel 1965 da Warner. L'obiettivo principale è quello di poter investigare variabili, qualitative o quantitative, ritenute delicate come ad esempio la droga, l'aborto, le abitudini sessuali, il reddito e altro.

Lo studio condotto si propone di effettuare un excursus delle metodologie proposte dalla data di introduzione della tecnica fino ai giorni nostri. In effetti l'ultimo contributo citato è riferito all'anno 2002. Va inoltre sottolineato che la produzione scientifica, per evidenti motivi, non è stata presa tutta in considerazione e ci si è quindi concetrati su quei metodi ritenuti particolarmente interessanti per l'autore del presente scritto.

Per semplificare la lettura dello scritto, in apertura di ogni metodologia proposta, viene indicato se questa si riferisce ad un piano di campionamento per variabili o per attributi.

2 I principali modelli proposti dal 1965 ad oggi.

2.1 Il modello di Warner (del 1965).

Attributi

Si supponga che ogni persona di una popolazione si possa far appartenere a due gruppi A e B (complementare di A). Il gruppo A identifica gli appartenenti ad una categoria *delicata* e si vuole stimare la proporzione di persone appartenenti a tale categoria. Per far ciò si estraggono con reinserimento n persone da sottoporre ad intervista; prima però ogni intervistato esegue un esperimento casuale E i cui eventi elementari sono A (appartenenza alla categoria delicata) e B (non appartenenza alla categoria delicata) ai quali è associata la probabilità rispettivamente di p e $1-p$. Il risultato dell'esperimento casuale risulta essere ignoto al ricercatore (intervistatore) che otterrà dall'intervistato una risposta del tipo "SI" oppure "NO", in accordo al risultato ottenuto nell'esperimento e alla sua reale situazione circa il possedere o meno l'attributo delicato A.

L'esperimento può essere di varia natura, come il lancio di una moneta opportunamente calibrata, l'estrazione di una pallina da un'urna di opportuna composizione, ecc.

Siano quindi:

π =proporzione di soggetti appartenenti ad A (ignota)

p =probabilità che si verifichi A nell'esperimento casuale E

$$X_i = \begin{cases} 1 & SI \\ 0 & NO \end{cases}$$

Così:

$$P(X_i = 1) = \pi p + (1-\pi)(1-p)$$

$$P(X_i = 0) = (1-\pi)p + \pi(1-p)$$

La stima di π_A sarà quindi (per $p \neq 1/2$)

$$\hat{\pi}_A = \frac{p-1}{2p-1} + \frac{n_1}{(2p-1)n}$$

ove

$$\frac{n_1}{n} = \pi p + (1-\pi)(1-p)$$

Le proprietà dello stimatore indicato sono:

$$E(\hat{\pi}_A) = \pi_A$$

$$Var(\hat{\pi}_A) = \frac{1}{n}\left[\frac{1}{16(p-\frac{1}{2})^2} - (\pi - \frac{1}{2})^2\right]$$

2.2 Il modello a proporzioni multiple (del 1967).

Attributi

Si supponga che ogni persona di una popolazione appartenga ad uno ed uno solo di tre gruppi (A, B, C), connessi e mutuamente esclusivi.

Si vuole stimare la proporzione degli individui appartenenti ad ognuno dei tre gruppi. Per fare ciò si estraggono dalla popolazione di origine due campioni casuali con reinserimento di dimensione rispettivamente n_1 e n_2 (non necessariamente deve essere $n_1 = n_2$) indipendenti e non *sovrapposti*.

Per proteggere la privacy del rispondente viene utilizzato un meccanismo casuale attraverso il quale vengono fornite indicazioni dai rispondenti circa l'appartenenza ad uno dei gruppi evidenziati. Si consideri ad esempio l'uso di due mazzi carte, il primo dei quali viene assegnato al primo campione e il secondo dei quali viene assegnato al secondo. Ogni mazzo di carte è formato da tre diversi tipi di carte, riportanti, a seconda del tipo, le seguenti tre affermazioni:

"Appartengo al gruppo A"

"Appartengo al gruppo B"

"Appartengo al gruppo C"

Nel caso in cui la numerosità campionaria o la dislocazione geografica delle unità statistiche siano tali da dover far intervenire più intervistatori sul medesimo campione, è bene che la composizione dei mazzi di carte sia la stessa. In ogni mazzo del medesimo campione la proporzione dei tre diversi tipi di carte deve esere uguale, mentre la proporzione tra i due campioni deve essere diversa. In sostanza la probabilità di estrarre una carta di un certo tipo deve essere uguale nel medesimo campione ma diversa tra i due campioni; infine la proporzione all'interno di ogni mazzo non deve essere pari ad $\frac{1}{3}$.

Ciò premesso l'intervistatore fa scegliere all'intervistato un mazzo di carte relativo al campione di appartenenza, gli dice di mescolarlo a piacere, di estrarre una carta a sorte e di leggere quanto riportato sulla carta estratta, senza però svelargielo. Se la frase contenuta risultasse essere in accordo con lo *status* posseduto, il rispondente dovrà rispondere con un "SI", altrimenti con un "NO". La risposta viene registrata dall'intervistatore che chiederà poi di ripristinare il mazzo di carte e di mescolarlo nuovamente.
Siano quindi:

π_1 =vera e ignota proporzione di soggetti appartenenti al gruppo A

π_2 =vera e ignota proporzione di soggetti appartenenti al gruppo B

π_3 =vera e ignota proporzione di soggetti appartenenti al gruppo C
Dove

$$\sum_{j=1}^{3} \pi_{ij} = 1$$

Inoltre sia P_{ij} la proporzione di carte riportante di tipo j nell'$i-mo$ campione, essendo $j = 1, 2, 3$ e $i = 1, 2$; così anche

$$\sum P_{ij} = 1$$

e

$$(P_{11} - P_{13})(P_{22} - P_{23}) \neq (P_{12} - P_{13})(P_{21} - P_{23})$$

Se si assume che i rispondenti dicano nella totalità dei casi la verità, la stima di massima verosimiglianza di π_1, π_2 e π_3 risulta essere la seguente:

$$X_{ir} = \begin{cases} 1 & \text{se ha risposto SI l'r-mo rispondente dell'i-mo campione} \\ 0 & \text{se ha risposto NO l'r-mo rispondente dell'i-mo campione} \end{cases}$$

$i = 1, 2$
$r = 1, 2, ..., n_1$ per $i = 1$ (primo campione)
$r = 1, 2, ..., n_2$ per $i = 2$ (secondo campione)
Inoltre la probabilità che l'r-mo rispondente nel primo campione risponda "SI" è:

$$Pr(X_{1r} = 1) = \lambda_1$$

$$Pr(X_{1r} = 0) = 1 - \lambda_1$$

e così nel secondo campione si ha:

$$Pr(X_{2r} = 1) = \lambda_2$$

$$Pr(X_{2r} = 0) = 1 - \lambda_2$$

Sia ancora:

1. n_{11} il numero dei "SI" ottenuti nel primo campione
2. $(n_1 - n_{11})$ il numero dei "NO" ottenuti nel primo campione
3. n_{21} il numero dei "SI" ottenuti nel secondo campione
4. $(n_2 - n_{21})$ il numero dei "NO" ottenuti nel secondo campione

Si ha quindi che:

$$\hat{\pi}_1 = \frac{(\frac{n_{11}}{n_1} - P_{13})(P_{22} - P_{23}) - (\frac{n_{21}}{n_2} - P_{23})(P_{12} - P_{13})}{(P_{11} - P_{13})(P_{22} - P_{23}) - (P_{12} - P_{13})(P_{21} - P_{23})}$$

$$\hat{\pi}_2 = \frac{(\frac{n_{11}}{n_1} - P_{13})(P_{21} - P_{23}) - (\frac{n_{21}}{n_2} - P_{23})(P_{11} - P_{13})}{(P_{11} - P_{13})(P_{22} - P_{23}) - (P_{12} - P_{13})(P_{21} - P_{23})}$$

$$\hat{\pi}_3 = 1 - \hat{\pi}_2 - \hat{\pi}_1$$

Ora se

$$\frac{n_{11}}{n_1} \sim Bi(\lambda_1, n_1)$$

e

$$\frac{n_{21}}{n_2} \sim Bi(\lambda_2, n_2)$$

segue che

$$E(\hat{\pi}_1) = \pi_1$$

$$E(\hat{\pi}_2) = \pi_2$$

$$E(\hat{\pi}_3) = \pi_3$$

Ovvero che tutti e tre sono stimatori corretti delle ignote proporzioni. Per quando riguarda la variabilità degli stimatori si ha:

$$Var(\hat{\pi}_1) = \frac{1}{k^2}\left[(P_{22} - P_{23})^2 \frac{\lambda_1(1-\lambda_1)}{n_1} + (P_{12} - P_{13})\frac{\lambda_2(1-\lambda_2)}{n_2}\right]$$

$$Var(\hat{\pi}_2) = \frac{1}{k^2}\left[(P_{21} - P_{23})^2 \frac{\lambda_1(1-\lambda_1)}{n_1} + (P_{12} - P_{13})\frac{\lambda_2(1-\lambda_2)}{n_2}\right]$$

ove

$$k = (P_{11} - P_{13})(P_{22} - P_{23}) - (P_{12} - P_{13})(P_{21} - P_{23})$$

La varianza di $\hat{\pi}_1$ e $\hat{\pi}_2$ raggiunge il limite inferiore di Rao-Cramer.

La varianza di $\hat{\pi}_3$ è conseguenza della relazione che lega i tre stimatori, quindi:

$$Var(\hat{\pi}_3) = \frac{1}{k^2}\left[(P_{22} - P_{21})^2 \frac{\lambda_1(1-\lambda_1)}{n_1} + (P_{12} - P_{11})^2\frac{\lambda_2(1-\lambda_2)}{n_2}\right]$$

$$Var(\hat{\pi}_1 + \hat{\pi}_2) = Var(\hat{\pi}_3) = Var(\hat{\pi}_1) + Var(\hat{\pi}_2) + 2Cov(\hat{\pi}_1, \hat{\pi}_2)$$

ove

$$\begin{aligned}[Cov(\hat{\pi}_1, \hat{\pi}_2) &= \frac{1}{k^2}\left[(P_{22} - P_{23})(P_{23} - P_{21})\frac{\lambda_1(1-\lambda_1)}{n_1}\right] + \\ &+ \frac{1}{k^2}\left[(P_{12} - P_{13})(P_{13} - P_{11})\frac{\lambda_2(1-\lambda_2)}{n_2}\right]\end{aligned}$$

Il confronto con l'indagine diretta mostra una maggiore efficienza di quest'ultima. Va detto però che il meccanismo casuale può essere utilizzato *adhoc* per migliorare la collaborazione dei rispondenti: probabilmente nel caso del metodo delle risposte casualizzate gli stessi rispondenti sono portati maggiormente a dire la verità.

Se però l'attributo di uno dei tre gruppi, diciamo A, porta con sè un carattere sociale delicato, i rispondenti che appartengono a questo gruppo non avrebbero motivo di mentire circa l'appartenenza ai gruppi B e C. Se poi si assume che la caratteristica di uno degli altri due gruppi, ad esempio C, è più imbarazzante della caratteristica del gruppo B, allora è più probabile che i rispondenti di B, che decidono di non dire la verità , non riporteranno il fatto di appartenere al gruppo C. Ancora, i rispondenti del gruppo C che decidono di non riportare la verità probabilmente si identificheranno maggiormente con il gruppo A che non con il B.

Le conclusioni che seguono sono molto importanti per utilizzare con successo la tecnica RR testè esposta.

1. Il casualizzatore utilizzato pare essere molto importante. Dopo varie sperimentazioni infatti i ricercatori hanno visto che il mazzo di carte funziona meglio con rispondenti che posseggono una certa cultura; per ovviare a questo inconveniente sono stati creati vari casualizzatori e tra essi quello formato da una scatola particolare pare abbia riscontrato un particolare successo. Si tratta di una scatola sigillata, contenente un certo numero di palline di tre diversi colori e di opportuna composizione. I rispondenti

sono chiamati a scuotere la scatola, osservare il colore che appare su di una finestra e di rispondere in accordo all'affermazione collegata a quel particolare colore e allo status posseduto. Come al solito l'intervistatore registra la risposta, ignorando il risultato dell'esperimento alla fine del quale il rispondente è chiamato a rimestare la scatola in modo di non lasciare traccia della bilia comparsa alla finestra.

2. Il modello RR presenta stime più variabili rispetto a quelle del modello ad intervista diretta. Si introduce qui una questione ancora aperta, ossia accettare il compromesso di una più alta variabilità delle stime in luogo di una maggiore protezione del rispondente e di un minor rifiuto di rispondere a questioni delicate da parte dello stesso.

3. Un'altra fonte di distorsione delle stime sono le risposte errate date, deliberatamente o per incomprensione del quesito, dai rispondenti. Anche qui introduciamo subito un argomento che spesso nel presente trattato verrà richiamato. Si tratteranno infatti metodi la cui variabilità degli stimatori sarà molto contenuta o la più contenuta, nell'ambito di apllicazione, ma la tecnica sarà talmente complessa da creare grossi problemi ai rispondenti in termini di utilizzo dei casualizzatori ma anche tale da creare forti sospetti. Sembrerebbe che tecniche complesse, per i non addetti ai lavori, ovviamente, fornirebbero ai ricercatori indicazioni esatte circa il risultato dell'esperimento effettuato dai vari rispondenti.

4. L'incapacità dell'intervistatore di comprendere la tecnica RR e il modo di proporlo aumenterebbero la probabilità di avere risposte false o evase con la conseguenza di ottenere stime maggiormente distorte.

2.3 Il modello della domanda incorrelata (di Simmons del 1969).

Attributi

Come abbiamo appena detto, rifiutarsi di rispondere o rispondere volutamente in modo errato sono due delle maggiori fonti di errore extra campionario nelle indagini sulle popolazioni umane. La tecnica proposta da Simmons, che andremo ad analizzare, fa sì che il grado di verità espresso dai rispondenti sia superiore rispetto a quello proposto da Warner. Si tratta della felice intuizione di introdurre una domanda relativa ad un carattere incorrelato con il carattere delicato, oggetto dell'investigazione. Così ad esempio in parallelo alla domanda delicata potrebbe essere introdotta una domanda circa l'aver mai visitato una certa città, circa il mese di compleanno di un parente, ecc. Gli autori suggeriscono anche una domanda relativa al luogo di nascita, ma si ritiene che questo possa far creare sospetti nei rispondenti che potrebbero evadere o eludere le domande proposte. L'obiettivo è anche in questo caso quello di stimare l'ignota proporzione degli appartenenti ad una certa categoria, ritenuta delicata. Sia quindi A questa categoria e Y sia quella incorrelata con A. L'intervistato esegue un esperimento casuale e dovrà fornire all'intervistatore un "SI" o un "NO" in accordo al suo status e alla affermazione che casualmente gli viene posta. Si noti sin d'ora che anche la proporzione di coloro che posseggono l'attributo Y è ignota e la somma di queste due proporzioni potrebbe essere inferiore, uguale o superiore all'unità. I rispondenti infatti potrebbero appartenere ad uno solo o anche ad entrambi i gruppi. Immaginiamo ad esempio il caso di aver abortito

e di essere stati a Parigi! Si estraggono quindi due campioni indipendenti di numerosità rispettivamente n_1 e n_2 (non necessariamente uguali); ognuna delle unità statistiche dei due campioni esegue un esperimento casuale il cui risultato è ignoto all'intervistatore al quale viene data la risposta "SI" o "NO". I casualizzatori utilizzati nei due campioni saranno diversi se non nella forma, almeno nella probabilità che si verifichino gli eventi o meglio le affermazioni:

Appartengo al gruppo A

Appartengo al gruppo Y

Nel caso particolare poi in cui insistano più intervistatori su di un campione, il casualizzatore deve assolutamente essere uguale in tutto e per tutto, per evidenti motivi di non far insospettire i rispondenti. Sia ora p_1 la probabilità che venga estratta dall'intervistato l'affermazione di appartenenza al gruppo A nel primo campione e p_2 l'analoga probabilità nel secondo campione. In generale sarà $p_1 \neq p_2$. Per analogia sia $1-p_1$ la probabiltà che nel primo campione venga estratta l'affermazione di appartenenza al gruppo Y (non delicato) e $1-p_2$ sia la stessa relativa al secondo campione.

Supposto che vi sia la totale veridicità da parte dei rispondenti, la probabilità di ottenere la risposta "SI" nei due campioni è rispettivemente pari a :

$$\lambda_1 = p_1\pi_A + (1-p_1)\pi_Y$$

$$\lambda_2 = p_2\pi_A + (1-p_2)\pi_Y$$

Quindi gli stimatori di massima verosimiglianza sono:

$$\pi_A = \frac{\lambda_1(1-p_2) - \lambda_2(1-p_1)}{p_1 - p_2}$$

$$\pi_Y = \frac{\lambda_1 p_2 - \lambda_2 p_1}{p_2 - p_1}$$

le cui stime si ottengono sostituendo nelle espressioni la proporzione dei "SI" ottenute nei due campioni:λ_1 e λ_2.

Se

$$\lambda_1 \sim Bi(n_1, \lambda_1)$$

e se

$$\lambda_2 \sim Bi(n_2, \lambda_2)$$

allora lo stimatore della varianza di π_A sarà

$$Var(\hat{\pi}_A) = \frac{1}{(p_1-p_2)^2}\left[\frac{\lambda_1(1-\lambda_1)(1-p_2)^2}{n_1} + \frac{\lambda_2(1-\lambda_2)(1-p_1)^2}{n_2}\right]$$

la cui stima si ottiene sostituendo λ_1 e λ_2 con la stessa proporzione di "SI" ottenuta rispettivamente nel primo e nel secondo campione.

Va detto che in generale si sceglie

$$p_1 + p_2 = 1 \tag{1}$$

e non si prendono in considerazione i casi in cui $p_2 = 0$ e $p_1 = 1$ poichè verrebbero meno le considerazioni fatte sui casualizzatori. Nei prossimi paragrafi questi casi particolari verranno trattati separatamente per la scelta di non dover stimare π_Y o per la scelta di non voler lavorare con due campioni.

La scelta di p_1 deve essere tale che esso in valore sia lontano da 0.5 ma contemporaneamente non deve creare sospetto nei rispondenti: gli studi empirici suggeriscono $p_1 = 0.2$ o $p_1 = 0.2$ con uno scostamento, sia in senso positivo che negativo di 0.1. Fissato p_1 in generale si ottiene anche p_2, ma se volessimo ignorare la (1), la scelta di p_2 deve essere tale da rendere minima la variabilità dello stimatore π_A, per un fissato livello di π_Y, n_1, n_2. La teoria ci suggerisce di prendere p_1 lontano da p_2: se l'uno stà sotto il valore 0.5, l'altro deve essere il più vicino possibile all'unità e viceversa, cercando sempre di non insospettire il rispondente.

Per quanto concerne invece la scelta di π_Y, deve essere:

1. per $\pi_A < 0.5$ si necessita di un π_Y più piccolo possibile
2. per $\pi_A > 0.5$ si necessita di un π_Y più grande possibile
3. per $\pi_A = 0.5$ si necessita di un π_Y vicino ai valori 0 o 1

Tutto questo però in coerenza con il fatto che se π_Y è prossimo a zero viene meno l'effetto della domanda incorrelata. E' bene quindi che esso non sia sotto il valore 0.1 e non superi il valore 0.9, tenendo presente che risultano più accettabili, se possibili, valori compresi tra 0.25 e 0.75.

La dimensione campionaria ottima è la seguente:

$$\frac{n_1}{n_2} = \sqrt{\frac{\lambda_1(1-\lambda_1)(1-p_2)^2}{\lambda_2(1-\lambda_2)(1-p_1)^2}}$$

Inoltre si dimostra[1] che il confronto con il modello di Warner, in termini di varianza può essere così formulato: se $p \in (p_0, 1)$ ove $p_0 = 0.339333$ è l'unica soluzione in $\left[0, \frac{1}{2}\right]$ di

$$\frac{1}{1+p^2} = 4p(1-p)$$

allora

$$Var(\hat{\pi}_S) < Var(\hat{\pi}_W).$$

2.3.1 Il caso particolare in cui π_Y è noto a priori.

Si prenda ora in considerazione il caso in cui la proporzione di coloro che posseggono l'attributo non delicato sia noto a priori. Vi possono essere infatti studi precedenti che forniscono informazioni su detto attributo oppure le informazioni sono disponibili presso organismi pubblici o privati come ad esempio l'anagrafe

[1] Dowling & Shachtman, 1975

o i centri di ricerca. In tale contesto va da sè che serve soltanto un campione e lo scenario che si presenta è il seguente. La probabilità di ottenere una risposta affermativa è data da:

$$\lambda_1 = p_1\pi_A + (1-\pi_1)\pi_Y$$

La stima di MV è data da:

$$\pi_A|\pi_Y = \frac{\lambda_1 - \pi_Y(1-p_1)}{p_1}$$

la cui varianza risulta essere pari a:

$$Var(\pi_A|\pi_Y) = \frac{\lambda_1(1-\lambda_1)}{np_1^2}$$

ove n è la dimensione dell'unico campione necessario e p_1 è la probabilità di estrarre l'affermazione legata al carattere delicato.

Anche in questo caso si dimostra che se $p \in (p_{00}, 1)$ ove $p_{00} = 0.381966$ allora

$$Var(\hat{\pi}_S) < Var(\hat{\pi}_W) \qquad \forall(\pi_Y, \pi_A) \in [0,1]^2$$

Inoltre in generale la varianza dello schema di Simmons, con proporzione nota del carattere non delicato, risulta essere sempre inferiore a quella con proporzione ignota dello stesso carattere non delicato, per ogni valore di p, π_Y e π_A.

2.4 Applicazione della tecnica RR a variabili quantitative (di Greenberg, Abernathy e Horvitz del 1969).

Variabili

Si tratta probabilmente del primo contributo in ambito di modelli a risposte casualizzate per quanto concerne le variabili. Tale metodo viene applicato ad un caso reale avente per obbiettivo la stima della quantità di aborti avuti da un certo numero di donne entrate a far parte del campione. L'impostazione di base ricalca quella classica e il casualizzatore utilizzato è una scatola, di opportuna composizione, avente una finestra attraverso la quale poter vedere il colore della pallina selezionata. A seconda del colore che compare vengono proposte le due seguenti affermazioni:

1. "Quanti aborti ha avuto fino ad oggi?"
2. "Se una donna dovesse lavorare full-time, quanti figli potrebbe avere?"

Nella fattispecie, l'urna era composta di 35 palle rosse e 15 blu alle quali sono state associate rispettivamente la prima e la seconda affermazione. La vera differenza nel caso in parola è che la risposta data dal rispondente ora non è un "SI" o un "NO" ma un numero. Tale numero è per il ricercatore di ignoto riferimento per quanto concerne le affermazioni per evidenti motivi di protezione del rispondente. Come nello schema di Simmons si necessita di due campioni indipendenti di numerosità rispettivamente n_1 e n_2 (nello studio effettuato era $n_1 = 623$ e $n_2 = 287$), stratificati e non sovrapposti.

Siano quindi:

- n_1 e n_2 le numerosità campionarie
- p_i la probabilità di estrarre la palla rossa nel campione i; $i = 1, 2$
- $1 - p_i$ la probabilità di estrarre la palla blu nel campione i; $i = 1, 2$
- $z_{i,j}$ la risposta dell'individuo j nel campione i
- $f(A)$ la funzione di densità associata al carattere delicato (numero di aborti)
- $f(Y)$ la funzione di densità associata al carattere non delicato (numero di figli) che deve essere molto simile, nella forma, ad $f(A)$ eccezion fatta per il parametro di localizzazione
- $\bar{A}$ la stima della media di $f(A)$
- $\bar{Y}$ la stima della media di $f(Y)$
- $\bar{Z}_1$ la media delle risposte ottenute nel primo campione
- $\bar{Z}_2$ la media delle risposte ottenute nel secondo campione

Ciò premesso dalla metodologia si ottengono i seguenti stimatori delle medie di A e di Y

$$m_A = \frac{(1-p_2)\bar{Z}_1 - (1-p_1)\bar{Z}_2}{p_1 - p_2}$$

$$m_Y = \frac{p_2\bar{Z}_1 - p_1\bar{Z}_2}{p_2 - p_1}$$

La varianza di detti stimatori è:

$$Var(m_A) = \frac{1}{(p_1-p_2)^2}\left[(1-p_2)^2 Var(\bar{Z}_1) + (1-p_1)^2 Var(\bar{Z}_2)\right]$$

$$Var(m_Y) = \frac{1}{(p_2-p_1)^2}\left[p_2^2 Var(\bar{Z}_1) + p_1^2 Var(\bar{Z}_2)\right]$$

Si precisa che Z_i è la distribuzione delle risposte date nei due campioni!

2.4.1 Come scegliere la variabile non delicata.

Risulta fondamentale che l'affermazione o la domanda legata alla caratteristica non delicata sia tale per cui la risposta da registrare dall'intervistatore sia espressa nella stessa unità di misura delle risposte date alla domanda delicata (dollari, sterline, numero di volte che è accaduto un certo avvenimento). Fissato quindi il criterio di scelta su p_1 e p_2, i parametri su cui si può intervenire sono n_1, n_2, μ_Y e σ_Y^2. Per alcuni valori fissati di n_1, n_2, la varianza degli stimatori diminuisce con il diminuire di σ_Y^2 e $|\mu_A - \mu_Y|$. Tutto ciò risulta essere una importante guida nella scelta della domanda o affermazione non delicata: non importa quanto il carattere Y sia diverso o assomigli al carattere delicato A *in media* ma piuttosto quanto più uniformi o più vicine possibili siano le risposte.

Operativamente un'accurata scelta potrebbe essere quella che vede convergere le medie dei due caratteri A e Y, con varianza σ_Y^2 minima. Si noti però a tal riguardo che se σ_Y^2 fosse considerevolmente più piccola di σ_A^2, si potrebbe assistere ad una caduta nella cooperazione da parte dei rispondenti. In generale quindi dovrà essere:

$$\mu_Y \longrightarrow \mu_A$$

$$\sigma_Y^2 \longrightarrow \sigma_A^2$$

2.4.2 Come scegliere le numerosità campionarie.

Le numerosità campionarie che minimizzano la varianza dello stimatore $\hat{\mu}_A$ devono essere tali da verificare la seguente relazione:

$$\frac{n_1}{n_2} = \sqrt{\frac{(1-p_2)^2 Var(Z_1)}{(1-p_1)^2 Var(Z_2)}} = \frac{(1-p_2)^2[(1+p_1(\phi_1^2-1)+p_1(1-p_1)\phi_2^2]}{(1-p_1)^2[(1+p_2(\phi_2^2-1)+p_2(1-p_2)\phi_2^2]}$$

dove:

$$\phi_1 = \frac{\sigma_A}{\sigma_Y} \quad e \quad \phi_2 = \frac{\mu_A - \mu_Y}{\sigma_Y}.$$

In molte situazioni comunque, un'accettabile approssimazione è data da:

$$\frac{n_1}{n_2} \doteq \frac{p_1}{p_2}$$

fissato $p_1 + p_2 = 1$. Nel caso specifico in cui i parametri della variabile Y siano noti a priori evidentemente non servono più due campioni ma ne basta uno solo e lo stimatore e la sua varianza risulatano essere pari a:

$$\hat{\mu}_A|\mu_Y = \frac{\bar{Z} - (1-p)\mu_Y}{p}$$

$$Var(\hat{\mu}_A|\mu_Y) = \frac{Var(\bar{Z})}{p^2} = \frac{Var(Z)}{np^2}$$

2.5 Il modello per la ottimizzazione del modello di Simmons (di Moor del 1971).

Attributi

Avendo come riferimento il modello di Simmons, si estraggono due campioni casuali semplici con reinserimento, indipendenti e di numerosità rispettivamente n_1 e n_2 in cui la probabilità di estrarre l'affermazione legata al carattere delicato è pari rispettivamente a p_1 e p_2. Se si suppone che tutti i rispondenti dicano la verità è noto che la varianza dello stimatore di π_A (proporzione di coloro che posseggono la caratteristica delicata) è:

$$Var(\hat{\pi}_A) = \frac{1}{(p_1-p_2)^2}\left[\frac{\lambda_1(1-\lambda_1)(1-p_2)^2}{n_1} + \frac{\lambda_2(1-\lambda_2)(1-p_1)^2}{n_2}\right] \qquad (2)$$

ove $\lambda_i = p_i\pi_A + (1-p_i)\pi_Y$ per i=1,2

Come al solito π_Y è la proporzione di persone che posseggono la caratteristica non delicata. La minimizzazione della (2) viene qui proposta ponendo $p_2 = 0$; questo implica che nel secondo campione si interroghi solo circa l'attributo non delicato e che nello stesso campione non venga fatto uso di casualizzatore. Un tale utilizzo del secondo campione può servire in fase preliminare per ottenere il valore di π_Y che al momento dell'indagine vera e propria su π_A potrebbe essere considerato noto a priori. Ovviamente nei casi in cui π_Y fosse già noto, tale fase preliminare risulterebbe inutile. Se volessimo fare un confronto tra quanto suggerito da Simmons, ossia $p_2 = 1 - p_1$ e la scelta di $p_2 = 0$ si possono evidenziare almeno tre vantaggi di quest'ultima impostazione:

- $\hat{\pi}_A$ ha varianza di stima minima
- i costi di intervista sono più contenuti vista l'applicazione del casualizzatore solo nel primo campione
- se la caratteristica Y può essere scelta dal ricercatore, è possibile stimare π_Y con metodi molto meno dispendiosi dell'intervista diretta.

Il confronto invece con il metodo di Warner è reso possibile ponendo $p_1 = p$ ed è evidente la maggior efficienza che si ottiene dal passaggio da $p_2 = 1 - p_1$ a $p_2 = 0$, minimizzando la (2) rispetto a n_1 e n_2 e per n costante. La minimizzazione risulta essere pari a:

$$Var(\hat{\pi}_A)^\star = \left[\frac{(1-p_1)\sqrt{\lambda_2(1-\lambda_2)} + (1-p_2)\sqrt{\lambda_1(1-\lambda_1)}}{(p_1-p_2)\sqrt{n}}\right]^2$$

ove il simbolo $(\star)$ denota la scelta ottimale delle dimensioni campionarie poste pari a:

$$\frac{n_1}{n_2} = \sqrt{\frac{\lambda_1(1-\lambda_1)(1-p_2)^2}{\lambda_2(1-\lambda_2)(1-p_1)^2}} \tag{3}$$

Per i valori più interessanti di π_Y, $(0.1; 0.5)$, fissato $\pi_A = 0.2$ e fissata la miglior allocazione per n_1 e n_2, il guadagno in termine di precisione delle stime (o se si vuole in termini di errore di stima) è circa l'80% per $P_1 = 0.7$, il 25% per $P_1 = 0.8$ e il 5% per $P_1 = 0.9$.

In sintesi si forniscono alcune utili indicazioni:

- SCELTA DI $\pi_Y \Rightarrow Y$ tale per cui $\left|\pi_Y - \frac{1}{2}\right| = max$
- SCELTA DI $p_1 \Rightarrow$ fissato $p_1 > p_2$, il primo deve essere il più possibile vicino all'unità, rispettando sempre il concetto di protezione del rispondente. Questa condizione è tale da rendere minima $Var(\hat{\pi}_A)^\star$
- SCELTA DI $p_2 \Rightarrow$ deve essere più basso possibile e come detto anche pari a zero.

In quest'ultimo caso,

$$Var(\hat{\pi}_A)^\star = \left[\frac{(1-p_1)\sqrt{\pi_Y(1-\pi_Y)} + \sqrt{\lambda_1(1-\lambda_1)}}{p_1\sqrt{n}}\right]^2$$

e la peggior scelta di π_Y è quando esso vale 0.5; in questo caso si dimostra comunque che il modello qui presentato resta preferibile a quello di Warner.

2.6 Il modello di Eriksson (del 1973).

Attributi e Variabili

Vengono qui proposti due modelli introdotti per la stima di parametri riferiti l'uno ad un piano di campoionamento per variabili e l'altro ad un piano di campionamento per attributi.

2.6.1 Il caso delle variabili.

Si consideri una popolazione di N persone alle quali sono associati i valori $X_1, X_2, ..., X_N$ di media μ_X e varianza σ_X^2. Si assume che sia possibile elencare L valori $Y_1, Y_2, ..., Y_L$ talchè tutti i valori X_i siano inclusi. La distribuzione dei valori Y deve essere scelta in modo tale che le persone siano convinte che l'intervistatore abbia scarse possibilità di indovinare se la risposta rivela la verità o meno circa un aspetto considerato delicato. Al momento dell'intervista vengono utilizzati due diversi tipi di carte, riportanti ciascuno le seguenti affermazioni:

- Dai la risposta correttamente
- Dicci se il tuo valore è Y_j

ove la proporzione di carte del primo tipo è p mentre la proporzione di carte del secondo tipo e riportante il valore Y_j è p_j, in coerenza con la relazione:

$$\sum_{j=1}^{L} p_j = 1 - p$$

La media e la varianza dei valori delle carte del secondo tipo sono:

$$\mu_Y = \sum_{j=1}^{L} Y_j \frac{p_j}{1-p}$$

$$\sigma_Y^2 = \sum_{j=1}^{L} (Y_j - \mu_Y)^2 \frac{p_j}{1-p}$$

L'intervista con la *i-ma* persona, che possiede il vero valore X_i, consiste nel chiedergli questo valore, di fargli estrarre una carta a sorte e di fargli dare la risposta in coerenza con quanto uscito dall'esperimento casuale. Per ogni intervistato l'esperimento viene replicato k volte in modo indipendente. Il valore fornito dall'*i-mo* intervistato alla *v-ma* replicazione viene considerata come una osservazione di una variabile casuale Z_{iv} avente spazio campionario $\Omega_{Z_{iv}} = (X_i, Y_1, Y_2, ..., Y_L)$ con associata probabilità rispettivamente pari a $p, p_1, p_2, ..., p_L$.

Dato quindi un campione di n persone sottoposte a questo tipo di intervista, le medie condizionate di :

$$\bar{Z}_i = \frac{\sum Z_{iv}}{k}$$

e

$$\bar{Z} = \frac{\sum\sum Z_{iv}}{nk}$$

sono:

$$E_Z(Z_{iv}) = pX_i + \sum_{j=1}^{L} Y_j p_j = pX_i + (1-p)\mu_Y$$

$$E_Z(\bar{Z}_i) = pX_i + (1-p)\mu_Y$$

$$E_Z(Z) = p\bar{X} + (1-p)\mu_Y$$

Supponendo che $\xi = \sum a_i X_i$ sia uno stimatore di μ_X, utilizzando un arbitrario campione di n persone sottoposte ad intervista diretta si ha che nel modello RR, una stima di μ_X è:

$$\hat{\mu}_X = \frac{\sum a_i(\bar{Z}_i - (1-p)\mu_Y)}{p}$$

il calcolo della varianza risulta molto complicato ma se si considera il caso in cui $a_i = \frac{1}{n}$ e $\xi = \frac{\sum X_i}{n} = \bar{X}$ si ha la seguente formula semplificata (per campioni autoponderati)

$$Var(\hat{\mu}_X) = Var(\bar{X}) + \frac{1-p}{nkp}\left(\frac{\sigma_Y^2}{p} + \sigma_X^2 + (\mu_X - \mu_Y)^2\right)$$

ove il primo termine indica la varianza che si ottiene con la intervista diretta e stimatore ξ, mentre il secondo termine fornisce l'aumento di varianza dovuto al modello RR.

2.6.2 Il caso delle mutabili.

Si consideri una popolazione suddivisibile in t gruppi disgiunti $A_1, A_2, ..., A_t$. Sia da stimare la proporzione di individui della popolazione appartenti a ciascuno di questi gruppi e per far ciò viene utilizzato lo stesso meccanismo prima descritto; l'esecuzione di un esperimento casuale porta all'estrazione di una carta contenente, con probabilità rispettivamente di p e $1-p$, le seguenti due affermazioni:

1. Dai una risposta vera
2. Appartieni al gruppo A_i? $\qquad i = 1, 2, ..., t$

Consideriamo ad esempio la stima di π_1, proporzione di coloro che appartengono al gruppo A_1. Se le carte di tipo 2 hanno una frequenza relativa pari a π_Y e quelle di tipo 1 pari a $1-\pi_Y$ e l'intervistatore registra il valore 1 alla risposta del tipo "appartengo al gruppo A_1" e il valore 0 altrimenti, allora valgono i risultati precedentemente descritti.

Sia quindi

$Z_{ij} = (1, 0)$ le possibili risposte dell'i-ma persona del campione alla j-ma estrazione della carta.

Per campioni casuali semplici si ha:

$$\hat{\pi}_1 = \frac{\bar{Z} - (1-p)\pi_Y}{p} = \frac{\bar{Z} - p_1}{p}$$

e

$$\begin{aligned} Var(\hat{\pi}_1) &= \frac{\pi_1(1-\pi_1)}{n}\frac{(N-n)}{N-1} + \\ &+ \frac{(1-p)}{nkp}\left(\frac{\pi_Y(1-\pi_Y}{p} + \pi_1(1-\pi_1) + (\pi_1 - \pi_Y)^2\right). \end{aligned}$$

L'esperienza sul campo suggerisce che per $\pi_1 \leq 0.5$ si deve prendere π_Y più piccolo possibile, in coerenza con il rischio di aumentare gli errori extra campionari. Analogamente, se $\pi_1 \geq 0.5$, allora π_Y deve essere più grande possibile.

Si dimostra inoltre che se le sottopopolazioni fossero solo due e si estraesse un campione casuale semplice con parametro $k = 1$, la varianza del presente metodo risulta essere inferiore rispetto a quella del metodo di Warner.

Un vantaggio del presente metodo è che a prescindere dal numero di proporzioni da stimare, si abbisogna solo di un unico campione; con i metodi introdotti invece si abbisognerebbe di t campioni indipendenti ($t - 1$ nel caso in cui π_Y fosse nota).

2.7 Il modello con due domande alternate (di Folson del 1973).

Attributi

Il modello che viene qui presentato sviluppa una nuova tecnica RR che utilizza in modo più efficace il secondo dei due campioni estratti. L'idea di base è quella di introdurre una seconda domanda incorrelata (Y_2) con l'attributo delicato A e di far uso sia di intervista sottoposta alle leggi probabilistiche del casualizzatore, sia di intervista diretta. Quindi nei due campioni S_1 e S_2, vengono somministrate tramite selezione casuale le affermazioni relative ai caratteri rispettivamente A, Y_1 e A, Y_2, mentre vengono fatte interviste dirette di tipo faccia a faccia circa i caratteri non delicati Y_2 e Y_1. Va da sè che ogni rispondente di entrambi i campioni potrebbe rispondere o ad una domanda per tipo, tra delicata e non, oppure ad entrambe le domande non delicate.

La tabella che segue sintetizza quanto appena esposto:

	S_1	S_2
Randomize device	Q_A	Q_A
Intervista diretta	Q_{Y_2}	Q_{Y_1}

Tabella 1: Esemplificazione del metodo di Folson

In termini probabilistici le affermazioni collegate ai caratteri delicati vengono estratte con probabilità pari a p e ciò fissato sia:

$$\lambda_i^r$$

la probabilità di ottenere una risposta del tipo "SI" alla domanda con casualizzatore nel campione $i = 1, 2$;

$$\lambda_i^d$$

la probabilità di ottenere una risposta del tipo "SI" alla domanda diretta nel campione $i = 1, 2$;

$$\lambda_i^{rd}$$

la probabilità di ottenere una risposta del tipo "SI" ad entrambe le domande nel campione $i = 1, 2$; in termini probabilistici questo significa che nel campione S_1:

$$\lambda_1^r = p\pi_A + (1-p)\pi_{Y_1}$$

$$\lambda_1^d = \pi_{Y_2}$$

$$\lambda_1^{rd} = p\pi_{AY_2} + (1-p)\pi_{Y_1Y_2}$$

e che nel campione S_2:

$$\lambda_2^r = p\pi_A + (1-p)\pi_{Y_2}$$

$$\lambda_2^d = \pi_{Y_1}$$

$$\lambda_2^{rd} = p\pi_{AY_1} + (1-p)\pi_{Y_1Y_2}$$

dove

- π_A =proporzione della popolazione con attributo delicato;
- π_{Y_i} =proporzione della popolazione con l'attributo non delicato Y_i e $i = 1, 2$;
- π_{AY_i} =proporzione della popolazione con l'attributo delicato e quello innocuo Y_i $(i = 1, 2)$;
- $\pi_{Y_1Y_2}$ =proporzione della popolazione con entrambi gli attributi innocui Y_1 e Y_2.

Due stimatori corretti di π_A, possono essere ottenuti dalle frequenze osservate dei "SI" nei due campioni:

$$\hat{\pi}_A(1) = \frac{\hat{\lambda}_1^r - (1-p)\hat{\lambda}_2^d}{p}$$

$$\hat{\pi}_A(2) = \frac{\hat{\lambda}_2^r - (1-p)\hat{\lambda}_1^d}{p}$$

Se $\hat{\pi}_A(1)$ e $\hat{\pi}_A(2)$ fossero stime statisticamente indipendenti di π_A, allora lo stimatore ottimo potrebbe essere una media pesata dei due, con pesi w_i inversamente proporzionali alla varianza di $\hat{\pi}_A(i)$. Nel caso di dipendenza tra i due stimatori dalla metodologia deriva la seguente buona approssimazione della varianza:

$$Var(\hat{\pi}_A)_{U2} = \frac{\lambda_1^r(1-\lambda_1^r) + (1-p)^2\pi_Y(1-\pi_Y)}{Np^2}$$

ove il simbolo $U2$ sta ad indicare che si tratta del modello a due domande incorrelate. L'autore propone come domande incorrelate ad A quelle del tipo:

1. Sei nato/a nel mese di aprile?
2. Tua madre è nata nel mese di aprile?

Con l'accorgimento che quando una delle due domande incorrelate sposta l'attenzione su un altro soggetto o su un altro oggetto personale rispetto al quale la risposta pottrebbe essere ignota al rispondente, l'autore suggerisce di provvedere alla messa a punto di una batteria di domande sostitutive come la data di nascita del marito e del figlio primogenito. In generale questi suggerimenti non vengono pienamente condivisi per motivi già esposti in precedenza: si pensi soltanto alla diffidenza che possano creare domande così personali; il rispondente medio potrebbe ritenere di essere individuato dando risposte legate alle date di nascita o di matrimonio che non dovrebbero pertanto essere prese in considerazione. Molte altre domande potrebbero essere utilizzate, come ad esempio, quelle relative ai gusti culinari o alla visita di talune città.

Un'altra forma della varianza di π_A può essere:

$$Var(\hat{\pi}_A)_{U2} = \frac{\Sigma_1^2\Sigma_2^2 - \Sigma_{12}^2}{\Sigma_1^2 + \Sigma_2^2 - 2\Sigma_{12}}$$

ove

$$\Sigma_1^2 = Var(\hat{\pi}_A(1)) = \frac{1}{p^2}\left[\frac{\lambda_1^r(1-\lambda_1^r)}{n_1} + \frac{(1-p)^2\pi_{Y1}(1-\pi_{Y1})}{n_2}\right]$$

$$\Sigma_2^2 = Var(\hat{\pi}_A(2)) = \frac{1}{p^2}\left[\frac{\lambda_1^r(1-\lambda_1^d)}{n_1} + \frac{(1-p)^2\pi_{Y2}(1-\pi_{Y2})}{n_2}\right]$$

$$\Sigma_{12} = Cov[\hat{\pi}_A(1), \hat{\pi}_A(2)]$$

Vediamo ora quale allocazione ϑ, con $n_1 = \vartheta N$ e $n_2 = (1-\vartheta)N$, minimizza la varianza dello stimatore. Nel caso specifico in cui $\pi_{Y1} = \pi_{Y2} = \pi_Y$, $\pi_{Y1Y2} = \pi_Y^2$ e $\pi_{AY1} = \pi_{AY2} = \pi_A\pi_Y$, tale ϑ è pari a 0.5. In generale si assume che $\vartheta = 0.5$ è tale per cui:

$$n_1 = \vartheta N$$

$$n_2 = (1 - \vartheta)N$$

$$(\hat{\pi}_A)_{U2} = \vartheta\hat{\pi}_A(1) + (1 - \vartheta)\hat{\pi}_A(2)$$

che fornisce inoltre una stima corretta di $Var(\hat{\pi}_A)$. Il valore ottimo di ϑ risulta essere pari a:

$$\vartheta_{opt} = \frac{\Sigma_2^2 - \Sigma_{12}}{\Sigma_1^2 + \Sigma_2^2 - 2\Sigma_{12}}$$

Va detto infine che una scelta accurata delle domande relative ai caratteri non delicati, positivamente correlati con il carattere delicato, permette di aumentare la precisione degli stimatori. In termini di efficienza il modello proposto è sempre migliore del modello di Moors, fatta eccezione del caso in cui π_Y sia noto a priori. Per il modello appena esposto l'autore propende per fissare $p = 0.5$ a cui corrisponde un randomizzatore molto semplice, quale il lancio di una moneta, in grado di destare meno sospetti di altri casualizzatori. Inoltre se $\pi_A = 0.2$ e $pi_Y = 0.1$ si ha un aumento in efficienza del 27% rispetto all'uso della intervista diretta; tale aumento di efficienza passerebbe al 53% se a parità di parametri p passasse da 0.5 a 0.7. Anche se molto efficiente questo modello non ha trovato molte applicazioni e probabilmente questo è dovuto al fatto che l'impianto dell'indagine risulta abbastanza pesante a fronte di altri modelli molto più semplici nella loro costruzione, anche se meno efficaci: credo che il modello di Simmons in tal senso sia molto significativo.

2.8 Tecnica RR con un nuovo casualizzatore (di Liu, Chaow e Mosley nel 1975).

Attributi

La novità che viene qui proposta riguarda il casualizzatore, formato da una grande boccia avente collegato un lungo tubo. La boccia risulta essere il contenitore delle palle e il tubo il risultato dell'esperimento casuale. Si supponga quindi che una popolazione possa essere suddivisa in k gruppi mutuamente esclusivi e per semplicità sia il colore 1 rappresentante del gruppo 1, il colore 2 rappresentante del gruppo 2 e così via. Il numero di palle per colore può essere diverso ossia $m_1 \neq m_2 \neq m_3 \neq ... \neq m_k$ con m_i=numero di palle di colore i e $\sum_{i=1}^{k} m_i = m =$ numero totale di palle. Al rispondente è richiesto di portare il casualizzatore con il tubo verso l'alto, di scuoterlo e di girarlo con il tubo verso destra in modo da far cadere tutte le palle nel tubo. Lo stesso tubo è contrassegnato in m punti, ognuno per posizione occupata dalle m palle. I rispondenti devono fornire la posizione occupata dalla prima palla il cui colore lo identifica con il gruppo a cui appartiene. Il tubo è trasparente solo su un lato in modo tale che l'intervistatore, posto lontano dall'esperimento, non possa vederne il

risultato. Per quanto concerne la stima dei parametri sia X_{ij}, la probabbilità che la prima palla di colore i sia nella prima posizione, fissata pari a:

$$X_{i,1} = \frac{m_i}{m}$$

La probabilità $X_{i,3}$ risulta essere pari a:

$$X_{i,3} = \left(\frac{m - m_i}{m}\right)\left(\frac{m - m_i - 1}{m - 1}\right)\left(\frac{m_i}{m - 2}\right)$$

In generale si ha che la probabilità che la prima palla occupi la posizione $j - ma$ è:

$$X_{i,j} = \begin{cases} \frac{(m-j)!(m-m_i)!m_i}{m!(m-m_i+1-j)!} & j = 1, 2, 3, ..., (m - m_i + 1) \\ 0 & altrove \end{cases}$$

Ora sia π_i la vera proporzione del gruppo i nella popolazione in modo tale che

$$\sum_{i=1}^{k} \pi_i = 1$$

Ora la probabilità che un intervistato risponda j, ossia P_j, è:

$$P_j = \sum_{i=1}^{k} \pi_i X_{i,j} \qquad j = 1, 2, ..., (m - m_i + 1)$$

Se un campione di dimensione n viene tratto da una popolazione e n_j è il numero di intervistati che hanno risposto j, allora $Y_j = \frac{n_j}{n}$ per $j = 1, 2, ..., m$, e inoltre

$$E(Y_j) = P_j = \sum_{i=1}^{k} \pi_i X_{i,j}$$

Si può anche scrivere il seguente modello lineare:

$$Y_j = \sum_{i=1}^{k} \pi_i X_{i,j} + e_j = \sum_{i=1}^{k-1} \pi_i X_{ij} + \pi_k X_{kj} + e_j =$$

$$= \sum_{i=1}^{k-1} \pi_i X_{ij} + (1 - \sum_{i=1}^{k-1} \pi_i X_{ij}) Xkj + e_j =$$

$$= \sum_{i=1}^{k-1} \pi_i X_{ij} + Xkj + Xkj \sum_{i=1}^{k-1} \pi_i + e_j$$

da cui

$$Y_j - X_{kj} = \sum_{i=1}^{k-1} \pi_i (X_{ij} - Xkj) + e_j$$

e in forma matriciale

$$\mathbf{Y} = \mathbf{X}\pi + \mathbf{e}$$

ove

$\mathbf{Y} = [Y_j - X_{kj}]$ è vettore m x 1
$\mathbf{X} = [X_{ij} - X_{kj}]$ è matrice m x $(k-1)$
$\pi = [\pi_i]$ è vettore $(k-1)$ x 1
$\mathbf{e} = [e_j]$ è vettore m x 1

Ovviamente $E(\mathbf{e}) = \mathbf{0}$ e la matrice di varianza-covarianza di $\mathbf{e}$, ossia $\mathbf{V} = v_{ij}$ del tipo m x m ha componenti pari a:

$$v_{ii} = \frac{P_i(1-P_i)}{n} \quad \text{e} \quad v_{ij} = \frac{-P_iP_j}{n}$$

La stima di π con il metodo dei minimi quadrati pesati è data da:

$$\hat{\pi} = (\mathbf{X'VX})^{-1}\mathbf{X'V^{-1}Y}$$

e

$$\hat{\pi}_k = 1 - \sum_{i=1}^{k-1} \hat{\pi}_i.$$

Se $\mathbf{C} = (\mathbf{X'VX})^{-1}$ e c_{ij} è il valore dell'i-ma riga e della j-ma colonna di C allora il vettore varianza-covarianza di π è:

$$Var(\hat{\pi}) = \mathbf{C}$$

e

$$Var(\hat{\pi}_k) = \sum_{i=1}^{k-1} c_{ii} + 2\sum_{i \neq j}^{k-1} c_{ij}$$

2.8.1 Alcune notazioni sull'efficienza degli stimatori.

Il numero minimo di palle necessarie per stimare i k gruppi mutuamente esclusivi è:

$$Min(k) = 1 + 2 + 3 + ... + k = \frac{k(k+1)}{2}.$$

Ovviamente questo non è il numero ottimo che rende minima la varianza degli stimatori, visto che la stessa varianza è una funzione di π e di X, e X è determinato dal numero di palle contenute nel casualizzatore. Prendendo come esempio un caso tricotomico, la varianza che ne risulta ci permette di fare alcune considerazioni di massima:

1. se $m_1 < m_2 < m_3$ e m_1, m_2 sono fissati, la varianza delle stime può essere diminuita se si aumenta m_3. Ciò nonostante, non c'è un valore ottimo di m_3 ma maggiore è quest'ultimo, migliore è l'efficienza dello stimatore. Così la varianza dello stimatore utilizzando una combinazione di palle di $1:2:3$ è maggiore di quella di $1:2:4$ e quest'ultima è più grande di quella di $1:2:5$;

2. se $m_1 < m_2 < m_3$ e m_2, m_3 sono fissati, la varianza gradatamenta aumenta all'avvicinarsi di m_1 a m_2. In ogni caso all'avvicinarsi di m_1 a 1, la varianza gradualmente diminuisce. L'ottimo si raggiunge quando $m_1 = 1$ e la varianza degli stimatori è minima quando la combinazione delle palle è $1 : 4 : 5$, aumenta per $2 : 4 : 5$ e massima per $3 : 4 : 5$;

3. se $m_1 < m_2 < m_3$ e m_1, m_3 sono fissati, la varianza di π_3 aumenterà ma quella di π_1 diminuirà al convergere di m_2 verso m_3. Contrariamente, se m_2 converge su m_1, la varianza di π_1 aumenterà ma quella di π_3 diminuirà. La varianza di π_1 è maggiore con la seguente combinazione di palle $1 : 2 : 5$ rispetto a $1 : 3 : 5$ o $1 : 4 : 5$. In generale sembrerebbe che l'aumento dell'efficienza di uno stimatore comporti il diminuire di quella di un altro;

4. se il rapporto di palle è fissato ma il numero di palle cambia, l'efficienza degli stimatori raggiunge l'ottimo con il minimo numero di palle;

5. se il numero totale di palle è fissato, è noto che l'ottimo per m_1 è $m_1 = 1$. Se abbiamo invece informazioni a priori su π, possiamo selezionare una delle possibili combinazioni che hanno $m_1 = 1$ e nelle quali il coefficente di variazione per ogni parametro è circa lo stesso; questo fornirebbe la giusta combinazione di palle;

6. per ridurre la varianza degli stimatori, i valori di m_i dovrebbero essere il più dissimili possibile. Così il colore corrispondente al più piccolo numero di palle potrebbe essere associato con la caratteristica non delicata, in modo da aumentare la collaborazione dei rispondenti.

2.9 Il modello a due stadi per variabili (di Reinmuth e Geurts del 1975).

Variabili

La necessità di introdurre alcune modifiche nei metodi fin qui proposti nasce da un sottoprodotto di uno studio condotto dagli autori per l'Associazione di Vendita al Dettaglio di Honolulu. La scopo dello studio fu di esaminare l'intensità del taccheggio nei centri commerciali di Honolulu. Più specificatamente, l'obiettivo era di determinare la frequenza del taccheggio su un settore di acquirenti che poteva essere considerato di taccheggiatori, ossia la frequenza di azioni ripetute di taccheggio tra i taccheggiatori. La proporzione di taccheggiatori nella popolazione di tutti gli acquirenti dei centri commerciali di Honolulu può essere stimata da un modello RR dicotomico applicato a dati qualitativi. Il risultato è una proporzione π_S, che stima la proporzione di taccheggiatori. La frequenza di taccheggio è invece una risposta quantitativa. Va detto che una classica indagine di tipo RR per variabili, registrerebbe un grande numero di risposte negative, indicativo del fatto che la maggioranza dei frequentatori non è solita alla pratica del taccheggio! Una stima in tale contesto fornirebbe la frequenza media di taccheggio per acquirente e risulterebbe scarsamente rappresentativa.

Per stimare invece la frequenza di taccheggiamento su coloro che praticano tali azioni si necessita di combinare i due modelli RR per variabili e per attributi. Viene qui proposto uno stimatore rapporto a due stadi del tipo:

$$\hat{\theta} = \frac{\hat{\mu}_S}{\hat{\pi}_S}$$

rapporto tra frequenza stimata di taccheggio sulla popolazione stimata dei taccheggiatori.

Utilizzando un simile approccio è indispensabile che le due stime $\hat{\mu}_S$ e $\hat{\pi}_S$ derivino da campioni indipendenti e non sovrapposti per almeno due ragioni. La prima è che un rispondente potrebbe essere più riluttante a dover dar risposta a due domande delicate; inoltre potrebbe capitare anche il caso assurdo in cui alla prima domanda si risponda di non essere taccheggiatore e alla seconda di dover rispondere al quesito circa il numero di volte in cui si ha taccheggiato. La seconda ragione è di tipo puramente teorico: se i due stimatori provengono da campioni indipendenti, allora risulterà più semplice ricavare la distribuzione del loro rapporto.

Nell'impostare il disegno dell'indagine è bene tenere presente i seguenti due aspetti:

1. la numerosità campionaria totale deve essere divisa in due sottocampioni indipendenti e non sovrapposti tali che
$$\frac{n_1}{n_2} = \frac{p_1}{p_2}$$
2. la distribuzione della domanda innoqua deve essere il più possibile simile a quella della domanda delicata per evitare di insospettire i rispondenti.

Date le premesse metodologiche si ha che:

$$E[\hat{\theta}] = E\left[\frac{\hat{\mu}_S}{\hat{\pi}_S}\right] = E[\hat{\mu}_S]E\left[\frac{1}{\hat{\pi}_S}\right]$$

Si dimostra che $\hat{mu}_S$ e $\hat{pi}_S$ sono stimatori corretti di μ_S e π_S rispettivamente. Ma

$$E\left(\frac{1}{x}\right) \geq \frac{1}{E(x)} \longrightarrow E(\hat{\theta}) = E(\hat{\mu}_S)E\left(\frac{1}{\hat{\pi}_S}\right) \geq \frac{\mu_S}{\pi_S}$$

Cochran fa due notazioni circa l'errore degli stimatori del rapporto campionario. Anzittutto ha rilevato che l'errore è dell'ordine di $\frac{1}{n}$ e lo rende trascurabile nei grandi campioni; inoltre se le quantità stimate sono linearmente dipendenti (con retta passante per l'origine), lo stimatore risulta essere corretto. Nel caso in esame questo è vero; infatti:

$$E\left(\frac{\mu_S}{\pi_S}\right) = \beta\pi_S$$

per tutti i possibili valori di π_S. Nell'applicazione di cui stiamo trattando ovviamente π_s e $E(\hat{\mu}_S|\pi_S)$ sono linearmente dipendenti con retta passante per l'origine. Se considerassimo infatti il caso in cui non vi siano taccheggiatori ($\pi_S = 0$), la frequenza attesa di taccheggiamento per cliente sarebbe zero. Inoltre se la proporzione di taccheggiatori (π_S) aumentasse ci si attenderebbe che aumentasse anche la frequenza attesa. In generale se un taccheggiatore sa che ci sono molti altri suoi *colleghi* è ragionevole pensare che il fenomeno dilaghi. Tale ragionamento vale anche per il viceversa.

Si dimostra che:

$$E(\hat{\theta} - \theta) = \frac{\theta}{\pi_S^2} Var(\hat{\pi}_S)$$

la cui stima risulta essere pari a

$$\frac{\hat{\theta}}{\hat{\pi}_S^2} \, Var(\hat{\pi}_S)$$

e fissato θ' pari alla differenza tra la stima e la distorsione di stima, esso risulterà pari a :

$$\hat{\theta} \left(1 - \frac{Var(\hat{\pi}_S)}{\hat{\pi}_S^2} \right)$$

Si noti che la distorsione di stima e π_S sono inversamente proporzionali. Se volessimo interpretare questa informazione dovremmo dire che più piccola è la proporzione di taccheggiatori sui clienti, maggiormente nasconderanno il loro comportamento, quando si osservano tutti i clienti, e più basse saranno le frequenze di taccheggio date in risposta dai taccheggiatori attuali.

Un'approssimazione della varianza di stima può essere:

$$\hat{Var}(\theta) = \frac{1}{\hat{\pi}_S^2} \left(Var(\hat{\mu}_S) + \theta^{12} Var(\hat{\pi}_S) \right)$$

2.9.1 I tratti salienti dello studio.

Sono stati selezionati 342 clienti del più grande centro commerciale di Honolulu. Lo studio è stato condotto su un periodo di 5 giorni consecutivi da due intervistatori. Il casualizzatore era costituito da un sacchetto nero contenente 75 tavolette nere e 25 bianche. Nella prima fase dello studio si è stimato π_S con una numerosità campionaria pari a 184 suddivisa in due campioni indipendenti e non sovrapposti di numerosità rispettivamente di 138 e 46. Le due domande proposte erano:

1. Negli ultimi 12 mesi ha taccheggiato qualcuno?
2. Oltre a oggi è venuto qui altre volte?

Nel primo campione venne associata la prima domanda con il colore nero e nel secondo campione con il colore bianco.

Per determinare invece l'intensità di taccheggio nella popolazione oggetto di studio, fu estratto un altro campione di dimensione pari a 158 unità e suddiviso poi in due sottocampioni indipendenti e non sovrapposti di numerosità rispettivamente pari a 126 e 42 unità. In questo caso le domande proposte furono:

1. Quante volte hai taccheggiato qualcuno in questo centro commerciale?
2. Oltre a oggi, quante volte è venuto qui?

La prima domanda fu associata nei due campioni rispettivamente al colore nero e al colore bianco. Si noti che in entrambi i campioni fu posto $p_1 = 0.75$ e $p_2 = 0.25$.

Questa tecnica potrebbe essere applicata anche ad altri ambiti come ad esempio:

1. Stimare la frequenza dell'uso di droga nella popolazione dei drogati.
2. Stimare il numero medio di acquisti di un prodotto sensibile sugli utilizzatori abituali.
3. stimare l'incidenza dei crimini commessi dai colletti bianchi tra tutti quelli commessi dai dipendenti.

2.10 Il modello di Liu e Chow (del 1976).

Variabili

Questo modello viene introdotto per la stima di parametri quantitativi e prende le mosse da altri contributi, con la modifica però dell'uso del casualizzatore. Si tratta infatti di una boccia con un tubo sul quale insiste una finestra grazie alla quale è possibile vedere il risultato dell'esperimento. Le palle contenute sono di due tipi diversi: quelle di colore bianco sulle quali è riportato un numero intero compreso tra 0 e k e quelle di colore rosso. La composizione dell'urna viene scelta a priori e pertanto le quote delle varie palle sono fissate prima di eseguire l'esperimento. Di fatto siamo in presenza di un'urna che contiene non due tipologie di palle bensì $k + 1$. Gli intervistati devono eseguire il seguente esperimento: mescolare la boccia, ruotarla verso destra in modo tale da far precipitare una palla verso la finestra, osservare l'esito dell'esperimento e nel caso in cui la palla fosse rossa debbono rispondere al quesito delicato, contrariamente debbono comunicare il numero riportato sulla palla bianca comparsa alla finestra. Come al solito l'intervistatore non conosce l'esito dell'esperimento poichè ad esso viene fornito un numero ed egli non sa se è stato letto sulla palla bianca o se è la risposta al quesito delicato.

Per quanto riguarda la stima dei parametri sia w_i il numero di palle bianche riportanti il numero i, $i = 1, 2, 3, ..., k$ e r sia il numero di palle rosse. Sia ancora $\sum_{i=1}^{k} w_i = w$ e quindi $r + w$ il numero totale di palle. Se π_i rappresenta la vera proporzione di coloro che nella popolazione oggetto di studio posseggono i come misura del carattere quantitativo delicato, in modo tale tale che $\sum_{i=1}^{k} \pi_i = 1$, allora la probabilità che un rispondente risponda i è data da:

$$p_i = \pi_i \frac{r}{r+w} + \frac{w_i}{r+w} \tag{4}$$

Ora se il campione consta di n unità, sia n_i il numero di rispondenti che rispondono i e Y_i la proporzione di coloro che rispondono i, ossia $Y_i = \frac{n_i}{n}$. Sostituendo Y_i come stima di p_i in (4), la stima di π_i sarà data da:

$$\hat{\pi}_i = Y_i \frac{r+w}{r} - \frac{w_i}{r} \tag{5}$$

le cui stime di varianza e covarianza sono:

$$\hat{Var}(\hat{\pi}_i) = \left(\frac{r+w}{r}\right)^2 \frac{\hat{Y}_i(1-\hat{Y}_i)}{n} \tag{6}$$

$$\hat{Cov}(\hat{\pi}_i, \hat{\pi}_j) = -\left(\frac{r+w}{r}\right)^2 \frac{\hat{Y}_i \hat{Y}_j}{n} \tag{7}$$

Dalla (7) si vede che il rapporto di palle rosse sul numero totale di palle è la componente principale che ha effetti sull'efficienza di stima. Se il numero totale di palle è fissato, e il numero di palle rosse aumenta, l'efficienza dello stimatore cresce; mentre se il numero totale di palle aumenta ma resta fissa la quota di palle rosse, l'efficienza dello stimatore non muta. Va detto però che fissato il rapporto $\frac{r+w}{r}$ è possibile aumentare la collaborazione dei rispondenti aumentando il numero totale di palle o cambiando la proporzione di palle bianche che riportano quel preciso numero. Inoltre aumentando il numero di replicazioni dell'esperimento si potrebbe aumentare l'efficienza di stima ma con il costo di sottoporre gli intervistati ad un meccanismo troppo complesso che potrebbe far insorgere inutili sospetti. Si noti infine che se la quota di palle rosse converge verso l'unità la tecnica RR diviene quella classica di intervista diretta!

2.11 Tre importanti modelli a contaminazione per lo studio di variabili (di Pollock e Bek del 1976).

Variabili

2.11.1 Il modello di Greenberg (G).

Si tratta di un'estensione del modello di Simmons ove vengono trattati dati binari. In questo caso invece le risposte a quesiti delicati hanno una funzione di probabilità $f(\bullet)$ ignota e le risposte a quesiti innocui hanno una funzione di probabilità $g(\bullet)$, nota. I rispondenti forniscono indicazioni circa il quesito delicato e quello innocuo con probabilità rispettivamente di p e $1-p$. Così la distribuzione del vettore delle risposte, Z è:

$$\Psi_G(z) = pf(z) + (1-p)g(z)$$

ove

$$\mu_Z = p\mu_X + (1-p)\mu_Y$$

e

$$\sigma_Z^2 = \sigma_Y^2 + p(\sigma_X^2 - \sigma_Y^2) + pq(\mu_X - \mu_Y)^2$$

Lo stimatore di μ_X è del tipo:

$$\hat{\mu}_X = \frac{\hat{\mu}_Z - q\mu_Y}{p} \qquad \text{con } \mu_Y \text{ noto}$$

In molti casi lo stimatore momento campionario *distribution free* ($\bar{Z}$) di μ_Z è dato da:

$$\hat{\mu}_X^G = \frac{\bar{Z} - q\mu_Y}{p} \tag{8}$$

che presenta una varianza pari a:

$$Var(\hat{\mu}_X^G) = \frac{\sigma_Z^2}{np^2}$$

Si noti che la (8) è stimatore corretto, consistente e asintoticamente normale di μ_X a prescindere dalla forma di $\Psi_G(z)$.

2.11.2 Il modello additivo (A).

In questo modello, come è facile intuire dal nome, il rispondente deve sommare al risultato della risposta data alla domanda delicata (X), un valore casuale (Y) preso da una distribuzione nota a priori. Le risposte così ottenute (Z) saranno del tipo:

$$Z = X + Y$$

da cui, fissata la indipendenza tra X e Y, segue che:

$$\mu_Z = \mu_X + \mu_Y$$

e

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$$

Con queste premesse uno stimatore per μ_X sarà del tipo:

$$\hat{\mu}_X = \hat{\mu}_Z - \mu_Y$$

In situazioni ove risulta difficile reperire la distribuzione di Z o non si vuol fissare una forma alla distribuzione di X, $f(x)$, $\bar{Z}$, lo stimatore dei momenti di μ_Z può essere:

$$\hat{\mu}_X^A = \bar{Z} - \mu_Y$$

e

$$Var(\hat{\mu}_X^A) = \frac{\sigma_Z^2}{n}$$

2.11.3 Il modello moltiplicativo (M).

Questo modello è una variante del modello principale ove l'operatore somma lascia il posto all'operatore moltiplicazione. Infatti le risposte Z sono frutto di una moltiplicazione i cui fattori sono il valore di X, attributo sensibile e Y, valore tratto da una distribuzione nota di numeri casuali. In definitiva si ha che:

$$Z = XY$$

e

$$\mu_Z = \mu_X \mu_Y$$

con varianza pari a

$$\sigma_Z^2 = \mu_Y^2 \sigma_X^2 + \mu_X^2 \sigma_Y^2 + \sigma_X^2 \sigma_Y^2$$

Lo stimatore per μ_X sarà:

$$\hat{\mu}_X = \frac{\hat{\mu}_Z}{\mu_Y}.$$

Lo stimatore dei momenti è:

$$\hat{\mu}_X^M = \frac{\bar{Z}}{\mu_Y}$$

la cui varianza risulta essere pari a

$$Var(\hat{\mu}_X^M) = \frac{\sigma_Z^2}{n\mu_Y^2}$$

2.11.4 Il confronto tra questi tre modelli G, A e M.

Introduciamo prima una tabella riassuntiva del metodo, dello stimatore e della varianza dello stimatore:

Metodo	Stimatore	Varianza di stima
G	$\frac{\bar{Z}-(1-p)\mu_Y}{p}$	$\frac{\sigma_Y^2+p(\sigma_X^2-\sigma_Y^2)+(p(1-p)(\mu_X-\mu_Y))^2}{np^2}$
A	$\bar{Z}-\mu_Y$	$\frac{\sigma_X^2+\sigma_Y^2}{n}$
M	$\frac{\bar{Z}}{\mu_Y}$	$\frac{\mu_Y^2\sigma_X^2-\sigma_Y^2\frac{\mu_x^2+\sigma_X^2}{\mu_Y^2}}{n}$

Tabella 2: I tre metodi proposti

Da notare che nel modello A la varianza dello stimatore dipende solo da σ_X^2 e σ_Y^2 mentre negli altri due modelli intervengono anche le medie di X e di Y. Il confronto diretto tra il modello G e il modello A necessita di una premessa. Nel modello G si necessita che le distribuzioni di X e Y siano simili ossia:

$$\mu_X \simeq \mu_Y \quad \text{e} \quad \sigma_X^2 \simeq \sigma_Y^2$$

Così la varianza dello stimatore risulta essere ragionevolmente piccola e non si introducono errori dovuti alla elusione delle risposte. Una approssimazione della varianza in un siffatto contesto potrebbe essere:

$$Var(\hat{\mu}_X^G) \simeq \frac{\sigma_Y^2}{np^2} \tag{9}$$

vista la convergenza di medie e varianze di X e Y Nel modello A, con le stesse premesse, si ha che

$$Var(\hat{\mu}_X^A) \simeq \frac{2\sigma_Y^2}{n} \tag{10}$$

Il rapporto tra la (9) e la (10) per valori di p inferiori a 0.7 mostra una maggior efficienza del modello A nei confronti del modello G.

Confrontando invece il modello A con il modello M, quest'ultimo risulta essere più efficiente, nel caso di variabili casuali positive, se:

$$\mu_Y > \sqrt{\mu_X^2+\sigma_X^2}.$$

2.12 La soluzione del conflitto tra "p" alto ed elusione delle risposte (di Morrarty e Wiseman del 1976).

Attributi

Nei modelli di Warner e Simmons nasce subitaneamente il conflitto tra avere un'alta probabilità di estrazione del quesito delicato e una bassa elusione delle domande da parte dei rispondenti. Generalmente ci si attende, a ragione, che a valori molto elevati di p coincida una altrettanto alta evasione alla domanda delicata. Gli autori propongono una soluzione, che ha attirato l'attenzione degli studiosi, per l'introduzione di un concetto nuovo. Secondo gli stessi, infatti, vi

sono due valori di p che possono essere studiati: p, ossia la già citata probabilità di estrarre il quesito legato al carattere delicato e $p^\star$, ossia la probabilità percepita dal rispondente di estrarre il quesito legato alla caratteristica delicata. Così per il rispondente la regola di decisione sarà:

$$\text{se} \qquad Pr^\star(A|SI) = \frac{p^\star \pi_A}{p^\star \pi_A + (1-p^\star)\pi_Y} = \theta$$

allora coopera

$$\text{se} \qquad Pr^\star(A|SI) > \theta$$

allora o non coopera o dice il falso se deve rispondere con un "SI".

Ora se fosse possibile reperire un casualizzatore tale per cui $p^\star < p$ si può trovare soluzione al dilemma prima esposto. In tal senso un buon casualizzatore sarà tale se il valore di p sarà molto più grande del valore di $p^\star$. Per cercare il casualizzatore in parola gli autori fecero un esperimento su 100 studenti universitari in possesso di nozioni probabilistiche, delle due maggiori realtà universitarie americane. I casualizzatori testati furono i seguenti:

1. lanciare due dadi e fare la somma dei risultati;
2. prendere un *chip* da una scatola contenentene uno bianco, uno rosso e dieci blu;
3. lanciare uno *spinner*;
4. pescare tre carte con reinserimento da un mazzo di 52 carte.

Per ogni casualizzatore fu richiesto di valutare verosimilmente che:

1. la somma dei dadi fosse compresa tra 4 e 10;
2. venga estratto un *chip* blu;
3. lo *spinner* cadesse nell'area ombrata;
4. le carte estratte fossero tutte dello stesso colore.

Dai dati sperimentali sembra che il primo casualizzatore sia quello che meglio interpreti il fatto che la probabilità percepita sia molto inferiore di quella reale. Infatti qui si ha che $p = 0.83$ mentre il valore mediano fornito dagli intervistati si collocò su un valore pari a 0.7. L'ultimo dei casualizzatori proposti risultò essere il peggiore.

2.13 Alcuni cenni al modello a risposte casualizzate multiple (MSQ) (di Kim e Flueck del 1976).

Attributi

Si consideri un campione di numerosità $N^\star$ allocato casualmente in Q sottocampioni mutuamente esclusivi. I rispondenti dei vari sottocampioni selezioneranno, attraverso l'uso di un casualizzatore, un'affermazione delicata e riferita ovviamente alla mutabile delicata oggetto di studio. A questa affermazione, non nota nei singoli casi all'intervistatore, i rispondenti risponderanno affermativamente o negativamente in coerenza con quanto risultato dall'esperimento casuale e con la loro specifica condizione nei riguardi dell'attributo delicato.

La struttura del modello è la seguente:

sottocampione i:

$(A_{i1}, \bar{A}_{i2}, A_{i3}, ..., \bar{A}_{iQ}, B_{i1}, \bar{B}_{i2}, B_{i3}, ..., \bar{B}_{iQ'})$ con $i = 1, 2, ..., (Q + Q')$ e dove $\bar{A}_{ij}$ è il complemento della affermazione A_{ij}.

Sia inoltre p_{ij} la probabilità associata alla j-ma affermazione nell'i-mo campione tale che $\sum_{j=1}^{Q} pij = 1$.

Sia ancora di interesse, per semplicità di esposizione e di notazione, solo l'attributo collegato alla affermazione A_{ij}.

Si definisce π_j la probabilità che una persona della popolazione possegga la caratteristica A_j con $j = 1, 2, ..., Q$ e λ_i la probabilità che un rispondente fornisca una risposta affermativa nel sottocampione i.

Nel caso specifico in cui si supponga che tutte le risposte date corrispondano a verità si ha che:

$$\lambda_i = p_{i1}\pi_1 + p_{i2}(1 - \pi_2) + p_{i3}\pi_3 + ... + \left(1 - \sum_{j=1}^{Q-1} p_{ij}\right)(1 - \pi_Q)$$

e nel caso di un numero dispari di addendi, l'ultimo termine va sostituito con $\pi_Q(1 - \pi_Q)$.

Con notazione matriciale:

$$\mathbf{\Lambda} = \mathbf{P\Pi}$$

ove

$$\mathbf{\Lambda}' = [\pi_1(1 - \pi_2)\pi_3(1 - \pi_4)...\pi_{Q-1}(1 - \pi_Q)]$$

$$\mathbf{\Pi}' = [\lambda_1\lambda_2...\lambda_Q]$$

e

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & 1 - \sum_{j=1}^{Q-1} p_{1j} \\ p_{21} & p_{22} & \cdots & 1 - \sum_{j=1}^{Q-1} p_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ p_{Q1} & p_{Q2} & \cdots & 1 - \sum_{j=1}^{Q-1} p_{Qj} \end{bmatrix}$$

Così i valori osservati di λ, $\hat{\lambda}$, sono combinazioni lineari di $\mathbf{\Pi}$, con un certo termine di errore $\mathbf{e}$. Dato $|\mathbf{P}| \neq \mathbf{0}$ si ha:

$$\hat{\mathbf{\Pi}} = \mathbf{P}^{-1}\hat{\mathbf{\Lambda}}$$

e utilizzando Q sottocampioni, possiamo stimare le Q caratteristiche delicate.

Lo stimatore così ricavato risulta essere corretto e con varianza pari a:

$$Var(\hat{\pi}) = \mathbf{P}^{-1}\hat{\mathbf{\Lambda}}\mathbf{P}^{-1}$$

ove $Var(\hat{\lambda})$ è la matrice diagonale con elementi

$$\frac{\lambda_i(1-\lambda_i)}{n_i} \qquad peri = 1, 2, 3, ..., Q$$

Inoltre la matrice di varianza e covarianza per π_i e π_j, indipendenti, per ogni valore di i e j diversi tra loro, si ottiene sostituendo l'elemento di posizione ij di $Var(\hat{\pi})$ con lo 0.

Si hanno così i due seguenti teoremi:

Teorema 1.
Se $\hat{\pi}_1$ e $\hat{\pi}_2$ sono indipendenti, $n_1 = n_2 = \frac{N^\star}{2}$, $p_1 + p_2 = 1$ e $p_1 \neq p_2$, allora $Var(\hat{\pi}_1 + \hat{\pi}_2)_W \geq Var(\hat{\pi}_1 + \hat{\pi}_2)_{MSQ} \quad \forall \pi_1, \pi_2$.

Infine si segnala la possibilità di utilizzare il modello presentato al caso di replicazione dell'esperiemnto casuale, al caso di analisi di un carattere quantitativo discreto e continuo e a quello di randomizzazione doppia (sia per la domanda che per la risposta).

2.14 La tecnica di Warner applicata al caso di due domande delicate (di Clickner e Iglewicz del 1976).

Attributi

Ci si propone qui di estendere il modello di Warner considerando il problema di stimare simultaneamente due proporzioni ignote riferite a due attributi delicati. Si vuole indagare cioè sul fatto che la popolazione di origine possegga almeno uno dei due attributi delicati presi in considerazione. Se A e B sono le caratteristiche delicate allora:

- π_A= probabilità che una persona possegga la caratteristica A;
- π_B= probabilità che una persona possegga la caratteristica B;
- π_{AB}= probabilità che una persona possegga le caratteristiche A e B;

Trovata la stima di queste quantità si potrà anche stimare poi tutte le altre probabilità congiunte, marginali e condizionate relativamente ad A e B.

Ai rispondenti viene dato un casualizzatore attraverso l'uso del quale debbono dare risposta affermativa o negativa ad una delle due affermazioni:

1. Appartengo al gruppo A;
2. Non appartengo al gruppo A.

Il casualizzatore seleziona la prima affermazione con probabilità p_1 e la seconda affermazione con probabilità $\bar{p}_1 = 1 - p_1$. Come al solito il ricercatore riceve una risposta che non sa a quale quesito si riferisca.

Analogamente viene applicato il meccanismo alla caratteristica B, con un diverso casualizzatore che presenta una probabilità di estrarre l'affermazione del primo tipo ma, ovviamente riferita a B, pari a p_2, con il vincolo che $p_1 \neq p_2$.

Pertanto le coppie di risposte che si ottengono sono del tipo:

1. SI-SI

2. SI-NO

3. NO-SI

4. NO-NO

Nel proseguio si assumerà per semplicità $SI = 1$ e $NO = 0$. Sia ora:
λ_{ij} = probabilità di ottenere come risposta i e j alla prima e alla seconda affermazione con $i = 0, 1$ e $j = 0, 1$.
Da cui:

$$\lambda_{11} = +\pi_A\bar{p}_2(p_1 - \bar{p}_1) + \pi_B\bar{p}_1(p_2 - \bar{p}_2) + \pi_{AB}(p_2 - \bar{p}_2)(p_1 - \bar{p}_1) + \bar{p}_2\bar{p}_1$$

$$\lambda_{10} = +\pi_A p_2(p_1 - \bar{p}_1) - \pi_B\bar{p}_1(p_2 - \bar{p}_2) - \pi_{AB}(p_2 - \bar{p}_2)(p_1 - \bar{p}_1) + p_2\bar{p}_1$$

$$\lambda_{01} = -\pi_A\bar{p}_2(p_1 - \bar{p}_1) + \pi_B p_1(p_2 - \bar{p}_2) - \pi_{AB}(p_2 - \bar{p}_2)(p_1 - \bar{p}_1) + \bar{p}_2 p_1$$

$$\lambda_{00} = -\pi_A p_2(p_1 - \bar{p}_1) - \pi_B p_1(p_2 - \bar{p}_2) + \pi_{AB}(p_2 - \bar{p}_2)(p_1 - \bar{p}_1) + p_2 p_1$$

Ora per stimare π_A, π_B e π_{AB}, si consideri un campione di n rispondenti e sia X_{ij} il numero di rispondenti che hanno risposto i e j rispettivamente alla prima e alla seconda affermazione. La distribuzione congiunta di X_{11},X_{10},X_{01} e X_{00} è di tipo multinomiale con parametri n, λ_{11}, λ_{10}, λ_{10}, λ_{00}. La stima di massima verosimiglianza di λ_{ij} è:

$$\hat{\lambda}_{ij} = \frac{X_{ij}}{n}$$

da cui e con riferimento alle equazioni prima introdotte si ha che:

$$\hat{\pi}_A = \frac{\hat{\lambda}_{11} + \hat{\lambda}_{10} - \bar{p}_1}{p_1 - \bar{p}_1}$$

$$\hat{\pi}_B = \frac{\hat{\lambda}_{11} + \hat{\lambda}_{01} - \bar{p}_2}{p_2 - \bar{p}_2}$$

$$\hat{\pi}_{AB} = \frac{\hat{\lambda}_{11}(p_1p_2 - \bar{p}_1\bar{p}_2) - \hat{\lambda}_{10}\bar{p}_2 - \hat{\lambda}_{01}\bar{p}_1 + \bar{p}_1\bar{p}_2}{(p_1 - \bar{p}_1)(p_2 - \bar{p}_2)}$$

Si noti che π_A e π_B sono gli stessi stimatori che derivano dal modello di Warner, sono corretti e hanno varianza e covarianza pari a:

$$Var(\hat{\pi}_A) = \frac{\pi_A\bar{\pi}_A + f(p_1)}{n}$$

$$Var(\hat{\pi}_B) = \frac{\pi_B\bar{\pi}_B + f(p_2)}{n}$$

$$Var(\hat{\pi}_{AB}) = \frac{\pi_{AB}\bar{\pi}_{AB} + \pi_A f(p_2) + \pi_B f(p_1) + f(p_1)f(p_2)}{n}$$

$$Cov(\hat{\pi}_A, \hat{\pi}_B) = \frac{\pi_{AB} - \pi_A\pi_B}{n}$$

$$Cov(\hat{\pi}_A, \hat{\pi}_{AB}) = \frac{\pi_{AB}\bar{\pi}_A + \pi_B f(p_1)}{n}$$

$$Cov(\hat{\pi}_B, \hat{\pi}_{AB}) = \frac{\pi_{AB}\bar{\pi}_B + \pi_A f(p_2)}{n}$$

ove

$$f(p_i) = \frac{p_i \bar{p}_i}{(p_i - \bar{p}_i)^2}$$

Sia ora $\lambda i. = \lambda_{i1} + \lambda_{i0}$ con $i = 0, 1$ e $\lambda_{.j} = \lambda_{1j} + \lambda_{0j}$ con $j = 0, 1$ allora si ha che:

$$\lambda_{11} = \lambda_{1.}\lambda_{.1} \qquad \Longleftrightarrow \qquad \pi_{AB} = \pi_A \pi_B$$

Ossia le risposte alle affermazioni si possono considerare indipendenti se e solo se sono indipendenti le mutabili in analisi. Tale indipendenza si può provare applicando il test χ^2 al modello RR qui proposto:

$$\chi^2 = \sum_{ij} \frac{(X_{ij} - n\hat{\lambda}_{i.}\hat{\lambda}_{.j})^2}{n\hat{\lambda}_{i.}\hat{\lambda}_{.j}}$$

Se l'interesse fosse rivolto solo alla stima della frazione di coloro che posseggono entrambi gli attributi, la procedura da seguire è quella appena esposta o in alternativa quella di Warner applicata al caso specifico $A \cap B$. Ovviamente ai rispondenti, dopo l'utilizzo di un opportuno casualizzatore, viene richiesto di rispondere affermativamente o negativamente alle due seguenti affermazioni:

1. Posseggo entrambe le caratteristiche A e B
2. Non posseggo nè A nè B

alle quali è associata rispettivamente una probabilità pari a p e $\bar{p} = 1 - p$. Lo stimatore che si ottiene è uguale a quello proposto da Warner e con varianza pari a:

$$Var(\pi^{\star}_{AB}) = \frac{\pi_{AB}\bar{\pi}_{AB} + f(p)}{n}$$

E' interessante notare che nessuna procedura è uniformemente migliore dell'altra. Infatti ponendo $p_1 = p_2 = p$ si ha:

$$\frac{Var(\hat{\pi}_{AB})}{Var(\pi^{\star}_{AB})} = \frac{1 + f(p)[\pi_A \pi_B + f(p) - 1]}{Var(\pi^{\star}_{AB})}$$

che è minore di 1 quando $\pi_A + \pi_B + f(p) < 1$ e questo si raggiunge quando tutti i termini π_A, π_B e $\frac{1}{2} - \left|\frac{1}{2} - p\right|$ sono prossimi a zero. Si noti però che più piccola è la varianza di $\hat{\pi}_{AB}$ maggiore sarà l'effetto di non risposte e/o di risposte non veritiere. Le due varianze poste a confronto possono essere reciprocamente l'una molto più grande dell'altra ma $Var(\pi^{\star}_{AB})$ è maggiore solo quando si esegue lo studio su caratteristiche la cui presenza sulle popolazioni è considerata rara

e con p piccolo. Più tipicamente la varianza a numeratore è considerevolmente più grande di quella a denominatore e questo significa che quando si passa a studiare due mutabili delicate congiuntamente ci si attende un incremento della variabilità degli stimatori che resta tale anche se le caratteristiche delicate investigate fossero più di due.

Circa la scelta di p_1 e p_2 si sottolinea il fatto che se questi vengono presi uguali, si garantisce maggiormente la protezione del rispondente, per ogni domanda e fissata una certa efficienza attesa dal modello. La pratica suggerisce di prendere p compreso tra i valori 0.2 e 0.3.

Dai confronti con il modello di Warner emerge chiaramente che quest'ultimo proposto è preferibile qualora provocasse un incremento di collaborazione nei rispondenti, in specie qualora inducesse a rispondere in modo veritiero.

2.15 Il modello senza casualizzatore (di Swensson del 1974).

Attributi

L'autore propone una tecnica diversa da tutte quelle elaborate in passato in quanto assicura la protezione del rispondente senza far uso del casualizzatore. Questo obiettivo viene raggiunto articolando in modo opportuno le affermazioni alle quali gli intervistati devono dar risposta affermativa o negativa. Come si vedrà, il punto debole del modello sta pobabilmente nella interpretazione corretta di dette affermazioni, spesso non così immediate.

Si vuole quindi studiare la proporzione di individui che posseggono un certo attributo delicato A, utilizzando, come nello schema di Simmons, un attributo Y non delicato, di ignota distribuzione e non correlato con A. Si considerino quindi due campioni indipendenti di numerosità rispettivamente n_1 e n_2 ai quali vengono collegate rispettivamente le seguenti due affermazioni:

1. Possiedi l'attributo delicato A o possiedi Y e non A ?
2. Possiedi l'attributo delicato A o non possiedi nè A nè Y?

La stima di massima verosimiglianza per π_A risulta essere:

$$\hat{\pi}_A = \hat{\lambda}_1 + \hat{\lambda}_2 - 1$$

ove $\hat{\lambda}_i$ è la proporzione di risposte affermative osservata nel campione $i = 1, 2$. La varianza dello stimatore sarà pari a:

$$Var(\hat{\pi}_A) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

ove

$$\pi_1 = \pi_A + \pi_{Y\bar{A}}$$

$$\pi_2 = \pi_A + \pi_{\bar{Y}\bar{A}}$$

Nel caso particolare in cui $\pi_A = 0.1$, $\pi_Y = 0.1$ e $\pi_{AY} = \pi_A\pi_Y$, il disegno proposto risulta essere 3.5 volte più efficientete di quello di Warner con $p = 0.7$.

2.16 Analisi della protezione del rispondente nei modelli RR (di Lanke 1976).

Attributi

Si tratta di un contributo che analizza nel dettaglio la protezione del rispondente. Infatti l'autore afferma che maggiore è la probabilità condizionata di appartenere al gruppo di coloro che posseggono l'attributo delicato, maggiore è l'imbarazzo prodotto dal dover dare quella risposta. Così la probabilità che un rispondente che appartiene ad A risponda affermativamente (ossia SI) è $Pr(I \in A|I = SI)$ e sarà indicata con $P(A|SI)$ e analogamente $P(A|NO)$.

In generale quindi il metodo 1 sarà considerato più protettivo del metodo 2 se:

$$max[P(A|SI), P(A|NO)]$$

è più piccolo nel metodo 1.

E' facile verificare che nel modello di Simmons, la risposta affermativa è più imbarazzante della risposta negativa; ossia

$$P(A|SI) > P(A|NO)$$

Per il metodo di Warner valgono le stesse considerazioni se $p > 0.5$. Se indichiamo con p_w la probabilità di estrarre il quesito delicato nel modello di Warner e con p_s l'analoga probabilità riferita al modello di Simmons allora si dimostra che:

$$\forall p_s,\ \pi_Y,\ \exists!\ p_w = \frac{1}{2} + \frac{p_S}{2p_s + 4(1 - p_s)\pi_Y} \tag{11}$$

tale per cui i due metodi risultano identici in termine di protezione del rispondente. Si noti però che in pratica i valori di p_w che soddisfano la (11) sono quelli compresi nel range $[0.5, 1]$ ed inoltre sempre con riferimento alla (11), dato un valore per p_w, $(p_w > 0.5)$, si possono ricavare i valori per p_s e per π_Y. Viceversa dati p_w e π_Y, si può ricavare p_s. In generale però non è detto che dati p_w, $(p_w > 0.5)$ e p_s, si possa ricavare π_Y.

Vediamo ora alcuni confronti pratici:

1. Se $p_w = p_s > 0.339$ allora $V_W > V_S$, $\forall(\pi_A, \pi_Y)$
2. Se $p_w = p_s > 0.382$ allora $V_W > V_M$, $\forall(\pi_A, \pi_Y)$
3. Se $p_w = p_s > 0.354$ allora $V_W > V_F$, $\forall(\pi_A, \pi_Y)$

Dati inoltre p_s e π_Y e preso p_w in accordo con la (11):

1. Se $\pi_Y > 0.5$ allora $V_W > V_S$, $\forall(\pi_A, p_s)$
2. Se $\pi_Y < 0.5$ allora $V_W < V_S$, $\forall(\pi_A, p_s)$
3. Se $\pi_Y > \frac{2}{3}$ allora $V_W < V_F$, $\forall(\pi_A, p_s)$
4. Se $\pi_Y < 0.5$ allora $V_W < V_F$, $\forall(\pi_A, p_s)$
5. Se $0.5 < \pi_Y < \frac{2}{3}$ allora la relazione dipende da π_A e da p_s e $\nexists \pi_S : V_W > V_M$, $\forall(\pi_A, p_s)$
6. Se $\pi_Y < 0.5$ allora $V_W < V_M$, $\forall(\pi_A, p_s)$
7. Se $\pi_Y > \frac{4}{5}$ allora $V_W < V_M$, quanto più $\pi_A < \frac{16}{25}$, $\forall p_s$

2.17 Il modello a probabilità condizionate (di Anderson del 1976).

Attributi

La popolazione oggetto di studio sia formata da individui che possano afferire solo a uno di due gruppi mutuamente esclusivi, A e $\bar{A}$, in proporzione π e $1-\pi$. Si consideri inoltre di estrarre un campione di persone cui sottoporre una tecnica RR per ottenere risposte del tipo "SI" - "NO". Una data tecnica di randomizzazione fornisce:

$$P(SI|A) = 1 - P(NO|A) \quad \text{e} \quad P(SI|\bar{A}) = 1 - P(NO|\bar{A}).$$

Il sospetto di essere classificato come un soggetto di tipo A, dopo aver risposto con un "SI" o con un "NO", viene misurato dalle seguenti probabilità:

$$P(A|SI;\pi) = \frac{P(SI|A)\pi}{P(SI|A)\pi + P(SI|\bar{A})(1-\pi)}$$

$$P(A|NO;\pi) = \frac{P(NO|A)\pi}{P(NO|A)\pi + P(NO|\bar{A})(1-\pi)}$$

Nel seguito si assume che $P(SI|A) > P(SI|\bar{A})$ ossia il fatto di avere dato una risposta affermativa aumenti il sospetto di essere classificato come un soggetto di tipo A. Vale il viceversa per la risposta negativa.

Si supponga ora di estrarre con reinserimento un campione di n unità, m delle quali abbiano risposto affermativamente alla affermazione sottoposta. La probabilità di ottenere una risposta affermativa è:

$$P(SI) = P(SI|A)\pi + P(SI|\bar{A})(1-\pi).$$

La stima di massima verosimiglianza per π si ottiene dalla seguente relazione:

$$P(SI|A)\hat{\pi} + P(SI|\bar{A})(1-\hat{\pi}) = \frac{n_1}{n},$$

ossia,

$$\hat{\pi} = \frac{\frac{m}{n} - P(SI|\bar{A})}{P(SI|A) - P(SI|\bar{A})}.$$

Tale stimatore risulta essere corretto e con varianza pari a:

$$Var(\hat{\pi}) = \frac{P(SI)[1-P(SI)}{n[P(SI|A) - P(SI|\bar{A}]^2}$$

che dopo alcune calcolazioni può essere scritto come:

$$Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n} \underbrace{\frac{1}{\frac{P(A|SI;\pi)}{\pi} - 1} \frac{1}{\frac{P(\bar{A}|NO;\pi)}{1-\pi} - 1}}_{\Psi(\pi)}$$

Il confronto con il modello binomiale indica quindi che quest'ultimo risulta essere peggiore solo se $\Psi(\pi) > 1$. Inoltre si noti che per un fissato valore di π tutti i modelli a risposta binaria con gli stessi rischi $P(A|SI)$ e $P(\bar{A}|NO)$ sono ugualmente efficaci per stimare π a prescindere dal casualizzatore. Dal punto di vista squisitamente probabilistico i due modelli differiscono per il fatto di generare diverse $P(A|SI)$ e $P(\bar{A}|NO)$.

2.18 Il modello di Warner con replicazioni multiple (di Liu e Chow del 1976).

Attributi

Il modello elaborato dagli autori prende le mosse dal modello di Warner e ne propongono un'estensione ottenuta replicando l'esperimento casuale m volte con l'obiettivo di dimuire la varianza dello stimatore senza intervenire sulla numerosità campionaria. Si supponga quindi di dover stimare π, vera proporzione di un attributo sensibile. La probabilità di selezionare la domanda o l'affermazione delicata sia p e m, come detto, il numero di replicazioni indipendenti dell'esperimento. Sia $X_i = i$ il numero di volte in cui il $j-mo$ intervistato abbia fornito una risposta affermativa. Così si ha che:

$$P(X_i = i|m) = \left(\begin{array}{c} m \\ n \end{array} \right) \left[\pi p^i (1-p)^{m-i} + (1-\pi) p^{m-i} (1-p)^i\right] = w_i$$

per $j = 1, 2, ..., n$,

$i = 1, 2, ..., m$ e

$\sum_{i=0}^{m} w_i = 1$

Se il campione consta di n unità e n_i sono quelle che hanno fornito i volte la risposta affermativa, in modo tale che

$$\sum_{i=o}^{m} n_i = n,$$

la funzione di verosimiglianza è:

$$L = \prod_{i=0}^{m} w_i^{n_i}$$

e la $log - verosomiglianza$ sarà:

$$log(L) = \sum_{i=0}^{m} n_i log(w_i)$$

Si dimostra che la varianza asintotica è tale per cui l'efficienza del modello aumenta all'aumentare delle replicazioni dell'esperimento. La varianza infatti anche in questo caso è composta da due addendi, uno dei quali è dovuto alla particolare tecnica adottata (quella RR) e raggiunge il minimo quando lo schema RR coincide all'indagine diretta. Va da sè che se $p > 0.5$ all'aumentare delle replicazioni dell'esperimento, aumentano i sospetti del rispondente nei riguardi della tecnica.

Dal punto di vista pratico questa tecnica potrebbe risultare efficace se il numero di replicazioni non supera le 3 prove; in caso contrario il metodo potrebbe appunto creare sospetti nel rispondente e anche molto pesante nella metodologia da applicare.

2.19 Una tecnica particolare senza uso di casualizzatore (di Takahasi e Sakasegawa del 1977).

Attributi

Qui viene proposta una tecnica RR che non fa uso di casualizzatore e quindi trova applicazione non solo nelle interviste di tipo $faccia-a-faccia$ ma anche con semplicità nelle interviste di tipo $telefonico$, *postale* e $autosomministrate$. Nel particolare il casualizzatore viene sostituito da una serie di domande ausiliarie, come ad esempio la preferenza circa una stagione. Dopo aver pensato a detta preferenza che non deve essere resa pubblica, il rispondente deve fornire al ricercatore uno zero (0) o un uno (1) in accordo con la lista che segue:

1. Se preferisci la primavera e
   (a) possiedi il carattere delicato A, allora rispondi 0
   (b) non possiedi A, allora rispondi 1
2. Se preferisci l'autunno e
   (a) possiedi il carattere delicato A, allora rispondi 1
   (b) non possiedi A, allora rispondi 0

Si noti che se la preferenza di una stagione è indipendente dal possedere o meno l'attributo A e la proporzione p di prefernza delle stagioni nella popolazione è nota, la tecnica proposta ricalca lo stesso modello matematico di Warner. Se la indipendenza è sostenibile, allora non vi sono problemi sul fatto che p sia ignoto, poichè sarebbe stimabile attraverso l'estrazione di un secondo campione indipendente. Ma se la indipendenza non è sostenibile, allora la stima di π è impossibile. Come si vedrà nel dettaglio è molto difficile, se non impossibile, trovare alcune domande ausiliari indipendenti da alcuni attributi e d'altro canto, la tecnica che verrà esaminita non assume l'indipendenza tra domanda ausiliaria e attributo delicato oggetto di studio.

Sia quindi π la vera proporzione dell'attributo sensibile A nella popolazione e lo si voglia stimare. Si estraggono a tal uopo tre campioni indipendenti con reinserimento dalla popolazione e ad ogni intervistato viene richiesto di scegliere una delle tre affermazioni proposte senza rilevare la scelta all'intervistatore. Fatta detta scelta il rispondente dovrà fornire in risposta un valore tra i due proposti, 0 e 1, in accordo con lo schema seguente.

Lista per il primo campione.

1. Se hai scelto il colore viola e
   (a) possiedi il carattere delicato A, allora rispondi 0
   (b) non possiedi A, allora rispondi 1
2. Se hai scelto il colore blu e
   (a) possiedi il carattere delicato A, allora rispondi 1
   (b) non possiedi A, allora rispondi 0
3. Se hai scelto il colore verde e

(a) possiedi il carattere delicato A, allora rispondi 1

(b) non possiedi A, allora rispondi 0

Lista per il secondo campione.

1. Se hai scelto il colore viola e

   (a) possiedi il carattere delicato A, allora rispondi 1

   (b) non possiedi A, allora rispondi 0

2. Se hai scelto il colore blu e

   (a) possiedi il carattere delicato A, allora rispondi 0

   (b) non possiedi A, allora rispondi 1

3. Se hai scelto il colore verde e

   (a) possiedi il carattere delicato A, allora rispondi 1

   (b) non possiedi A, allora rispondi 0

Lista per il terzo campione.

1. Se hai scelto il colore viola e

   (a) possiedi il carattere delicato A, allora rispondi 1

   (b) non possiedi A, allora rispondi 0

2. Se hai scelto il colore blu e

   (a) possiedi il carattere delicato A, allora rispondi 1

   (b) non possiedi A, allora rispondi 0

3. Se hai scelto il colore verde e

   (a) possiedi il carattere delicato A, allora rispondi 0

   (b) non possiedi A, allora rispondi 1

Le affermazioni circa le preferenze possono essere formulate anche prendendo in considerazione gli animali, i paesi, i numeri e altro.

	Campione 1		**Campione 2**		**Campione 3**	
Colori	A	$\bar{A}$	A	$\bar{A}$	A	$\bar{A}$
Viola	0	1	1	0	1	0
Blu	1	0	0	1	1	0
Verde	1	0	1	0	0	1

Tabella 3: Sintesi delle tipologie di risposte

Vediamo ora come possiamo pervenire alla stima di π. Siano:

- $P(A,i)$ = la proporzione nella popolazione di persone che posseggono l'attributo A e che scelgono B_i con $i = 1,2,3$;
- $P(\bar{A},i)$ = la proporzione nella popolazione di persone che non hanno l'attributo A e che scelgono B_i con $i = 1,2,3$;
- $n(i)$ = l'ampiezza dell'$i-mo$ campione $(i = 1,2,3)$;
- $y(i)$ = il numero di rispondenti che hanno risposto 1 nell'$i-mo$ campione.
- $q(i) = \pi + p(\bar{A},i) - p(A,i)$ per $i = 1,2,3$ è la probabilità che un rispondente nell'$i-mo$ campione risponda 1.

Ciò premesso si ha che:

$$\pi = \sum_{i=1}^{3} q(i) - 1$$

e

$$E[y(i)] = n(i)q(i) \quad \text{per} \quad i = 1,2,3$$

Lo stimatore di massima verosimiglianza per π sarà:

$$\hat{\pi} = \sum_{i=1}^{3} \frac{y(i)}{n(i)} - 1$$

Si tratta di uno stimatore corretto e con varianza pari a:

$$Var(\hat{\pi}) = \frac{3}{n}\sum_{i=1}^{3} q(i)[1-q(i)] = \frac{3}{n}\left[1 - \pi\bar{\pi} - \sum_{i=1}^{3} d(i)^2\right]$$

ove $\bar{\pi} = 1 - \pi$ e $d(i) = p(A,i) - p(\bar{A},i)$ per $i = 1,2,3$

Si noti inoltre che non si assume la indipendenza tra A e la scelta di un qualche B_i; contrariamente ci si assume un rischio molto forte se si volesse optare per la indipendenza senza l'ausilio di accurate considerazioni.

Si consideri ora che

$$n(i) = \frac{n}{3} \to \sum_{i=1}^{3} d(i)^2 \geq \frac{1}{3}(\pi - \bar{\pi})^2$$

e l'uguaglianza si ha quando

$$d(1) = d(2) = d(3) = \frac{\pi - \bar{\pi}}{3}$$

D'altro canto

$$\sum_{i=1}^{3} d(i)^2 \leq (\pi)^2 + (\bar{pi})^2$$

e l'uguaglianza si ha quando $P(A,i) = \pi$ per qualche i e, $P(\bar{A},j) = \bar{\pi}$ per qualche $j \neq i$.

Possiamo altresì scrivere che

$$\frac{3}{n}\pi\bar{\pi} \leq Var(\pi) \leq \frac{1}{n}(2+\pi\bar{\pi})$$

qui si ha l'uguaglianza a sinistra se

$$d(1) = d(2) = d(3) = \frac{\pi - \bar{\pi}}{3}$$

mentre si ha l'uguaglianza a destra se

$$P(\bar{A}, j) = \bar{\pi} \quad \text{per} j \neq i \tag{12}$$

Se ponessimo inoltre $p = \frac{1}{3}$ allora la varianza dello stimatore di Warner è data da $\frac{2+\pi\bar{\pi}}{n}$ e dalla (12) si desume che la varianza del modello qui presentato non risulta essere maggiore di quella di Warner, ferma restando la condizione di partenza.

Nel caso in cui la caratteristica delicata fosse indipendente dalla scelta di B_i allora la varianza diverrebbe:

$$Var(\pi) = \frac{3}{n}\left[1 - \pi\bar{\pi} - (\pi - \bar{\pi})^2 \sum_{i=1}^{3} a_i\right]$$

2.19.1 Due varianti al modello.

Vengono qui proposte due varianti al modello appena presentato. Nella prima di queste si estraggono due campioni indipendenti e ad ogni intervistato si chiede di scegliere silentemente una fra due affermazioni, come ad esempio la preferenza tra l'autunno e la primavera e di fornire in seguito all'intervistatore una risposta in accordo con il modello che segue:

Lista per il primo campione.

1. Se hai scelto l'autunno e
   (a) possiedi il carattere delicato A, allora rispondi 1
   (b) non possiedi A, allora rispondi 0
2. Se hai scelto la primavera rispondi 1 in ogni caso.

Lista per il secondo campione.

1. Se hai scelto la primavera e
   (a) possiedi il carattere delicato A, allora rispondi 1
   (b) non possiedi A, allora rispondi 0
2. Se hai scelto l'autunno rispondi 1 in ogni caso.

Fissati $n(i)$ e $y(i)$ come sopra si ha che:

$$\pi_1 = \hat{\pi} = \frac{y(1)}{n_1} + \frac{y(2)}{n_2} - 1$$

e risulta essere uno stimatore corretto e con varianza pari a:

$$Var(\hat{\pi}_1) = \sum_{i=1}^{2} \frac{q(i)[1-q(i)]}{n(i)}$$

ove

$$q(i) = 1 - p(\bar{A}, i)$$

e $p(\bar{A}, 1)$ è la proporzione nella popolazione delle persone che non hanno l'attributo delicato A e hanno scelto silentemente come stagione l'autunno; per analogia $p(\bar{A}, 2)$ sono quelli che hanno scelto silentemente la primavera. Nel caso specifico in cui $n(1) = n(2) = \frac{n}{2}$ si ha:

$$Var(\hat{\pi}_1) = \frac{2}{n}\left[\pi\hat{\pi} + 2P(\bar{A}, 1)P(\bar{A}, 2)\right]$$

e

$$\frac{2\pi\bar{\pi}}{n} \leq Var(\hat{\pi}_1) \leq \frac{1}{n}(1 + 2\pi\bar{\pi})$$

Si noti che nella presente modifica al modello di partenza la risposta 0 include solo la possibilità di $\bar{A}$ mentre la risposta 1 include sia A che $\bar{A}$. Quindi se il possedere o non il possedere il carattere A rappresentasse un aspetto delicato dell'intervistato allora questa tecnica non potrebbe essere applicata.

La seconda variante è simile alla prima e lo schema è quello che segue. Si estraggono tre campioni indipendenti e ad ogni intervistato viene chiesto di scegliere silentemente tra tre possibilità, B_1, B_2 e B_3, per poi rispondere in accordo con la lista seguente.

Lista per il primo campione.

1. Se hai scelto B_1 e
   (a) possiedi il carattere delicato A, allora rispondi 1
   (b) non possiedi A, allora rispondi 0
2. Se hai scelto B_2 e
   (a) possiedi il carattere delicato A, allora rispondi 1
   (b) non possiedi A, allora rispondi 1
3. Se hai scelto B_3 e
   (a) possiedi il carattere delicato A, allora rispondi 0
   (b) non possiedi A, allora rispondi 0

Lista per il secondo campione.

1. Se hai scelto B_1 e

(a) possiedi il carattere delicato A, allora rispondi 0

(b) non possiedi A, allora rispondi 0

2. Se hai scelto B_2 e

(a) possiedi il carattere delicato A, allora rispondi 1

(b) non possiedi A, allora rispondi 0

3. Se hai scelto B_3 e

(a) possiedi il carattere delicato A, allora rispondi 1

(b) non possiedi A, allora rispondi 1

Lista per il terzo campione.

1. Se hai scelto B_1 e

(a) possiedi il carattere delicato A, allora rispondi 1

(b) non possiedi A, allora rispondi 1

2. Se hai scelto B_2 e

(a) possiedi il carattere delicato A, allora rispondi 0

(b) non possiedi A, allora rispondi 0

3. Se hai scelto B_3 e

(a) possiedi il carattere delicato A, allora rispondi 1

(b) non possiedi A, allora rispondi 0

Per una opportuna sintesi si consulti la tabella che segue:

	Campione 1		**Campione 2**		**Campione 3**	
ITEM	A	$\bar{A}$	A	$\bar{A}$	A	$\bar{A}$
B_1	1	0	0	0	1	1
B_2	1	1	1	0	0	0
B_3	0	0	1	1	1	0

Tabella 4: Sintesi delle tipologie di risposte

Lo stimatore che se ne ottiene è:

$$\pi_2 = \hat{\pi} = \sum_{i=1}^{3} \frac{y(i)}{n(i)} - 1$$

che è corretto e con varianza pari a:

$$Var(\pi_2) = \sum_{i=1}^{3} \frac{q(i)[1-q(i)]}{n(i)}$$

ove $q(i)$ è la somma delle proporzioni nella popolazione cui corrispondono le risposte di tipo 1 nei vari campioni.

Alla base del metodo proposto e delle sue varianti vi sono due assunzioni particolari che devono essere sempre valide per non inficiare la validità del metodo stesso:

1. i rispondenti non devono cambiare scelta di B_i dopo aver visto la lista di appartenenza;

2. i rispondenti devono rispondere onestamente e correttamente circa la lista di appartenenza.

Si tenga infine sempre presente che la privacy dell'intervistato viene meno se uno dei B_i è molto diffuso o risulta essere molto correlato con il carattere delicato A. La condizione ottimale è che i vari B_i siano equamente distribuiti nella popolazione oggetto di studio e che inoltre siano scarsamente correlati con A.

2.20 Un modello a contaminazione di tipo additivo (di Kim e Flueck del 1978).

Attributi

Si suppone che ogni rispondente appartenga ad uno solo di k gruppi disgiunti e, considerando il caso particolare di $k = 3$ (una generalizzazione non risulta per niente difficile), sia C_j il vero gruppo di appartenenza del $j - mo$ rispondente, con $C_j = 1, 2, 3$ e $j = 1, 2, ..., n$ e a_j il valore additivo selezionato casualmente, con $a_j = 1, 2, 3$. Così la $j - ma$ risposta codificata y_j, derivante dal vero gruppo C_j sarà:

$$y_j = C_j + a_j$$

e per ottenere la riservatezza del rispondente, y_j è trasformata dal rispondente stesso nel valore R_j, ossia:

$$R_j = \begin{cases} y_j & \text{se} \quad y_j \leq 3 \\ y_j - 3 & \text{se} \quad y_j > 3 \end{cases}$$

La tabella che segue meglio spiega il meccanismo:

R	**Sorgente (c+a)**		
1	$1+3$	$2+2$	$3+1$
2	$1+1$	$2+3$	$3+2$
3	$1+2$	$2+1$	$3+3$

Tabella 5: Sintesi delle tipologie di risposte

Si definisce:

$$\pi_C = P(X \in C) \qquad \text{per } C = 1, 2, 3$$

$$P_a = P(X \text{selezioni } a) \qquad \text{per } a = 1, 2, 3$$

ove X è un rispondente. La probabilità λ_a che una persona riporti il valore R $(= 1, 2, 3)$ è:

$$\lambda_1 = p_3\pi_1 + p_2\pi_2 + p_1\pi_3$$

$$\lambda_2 = p_1\pi_1 + p_3\pi_2 + p_2\pi_3$$

$$\lambda_3 = p_2\pi_1 + p_1\pi_2 + p_3\pi_3$$

Notando poi che

$$\lambda_3 = 1 - \lambda_1 - \lambda_2$$

e che

$$\pi_3 = 1 - \pi_1 - \pi_2$$

quindi le equazioni di interesse si riducono a:

$$\lambda_1 = p_1 + (p_3 - p_1)\pi_1 + (p_2 - p_1)\pi_2$$

$$\lambda_2 = p_2 + (p_1 - p_2)\pi_1 + (p_3 - p_2)\pi_2$$

e in notazione matriciale

$$\Lambda = \mathbf{P}\Pi$$

ove

$$\Lambda' = [\lambda_1 - p_1 \quad \lambda_2 - p_2]$$

$$\Pi' = [\pi_1 \quad \pi_2]$$

e

$$P = \begin{bmatrix} p_3 - p_1 & p_2 - p_1 \\ p_1 - p_2 & p_3 - p_2 \end{bmatrix}$$

Dato che $|P| \neq 0$ (si noti che $|P| = 0$ solo se $p_1 = p_2 = p_3 = \frac{1}{3}$) si ha che:

$$\hat{\Pi} = \mathbf{P}^{-1}\hat{\Lambda}$$

e

$$Var(\hat{\Pi}) = \mathbf{P}^{-1}Var(\Lambda)(\mathbf{P}^{-1})'$$

ove

$$Var(\hat{\Lambda}) = \frac{1}{n}\begin{bmatrix} \lambda_1(1-\lambda_1) & -\lambda_1\lambda_2 \\ -\lambda_1\lambda_2 & \lambda_2(1-\lambda_2) \end{bmatrix}$$

e

$$\hat{\Lambda}' = [\hat{\lambda}_1 - p_1 \quad \hat{\lambda}_2 - p_2],$$

$$\hat{\lambda}_i = \frac{Z_i}{n}$$

e Z_i è una variabile casuale che descrive il numero di persone che si assegnano come membri dell'$i-mo$ gruppo in un campione di n rispondenti ($Z_1+Z_2+Z_3 = n$). Specificatamente lo stimatore puntuale di π_i con $i = 1, 2, 3$ è:

$$\hat{\pi}_1 = \frac{1}{|P|}\Big[(p_3 - p_2)(\hat{\lambda}_1 - p_1) + (p_1 - p_2)(\hat{\lambda}_2 - p_2)\Big]$$

$$\hat{\pi}_2 = \frac{1}{|P|}\Big[(p_2 - p_1)(\hat{\lambda}_1 - p_1) + (p_3 - p_1)(\hat{\lambda}_2 - p_2)\Big]$$

$$\hat{\pi}_3 = 1 - \hat{\pi}_1 - \hat{\pi}_2$$

Gli stimatori qui presentati sono corretti e con varianza pari a:

$$Var(\hat{\pi}_1) = \frac{(p_3 - p_2)^2\lambda_1(1-\lambda_1) - 2(p_3 - p_2)(p_1 - p_2)\lambda_1\lambda_2 + (p_1 - p_2)^2\lambda_2(1-\lambda_2)}{n|P|^2}$$

$$Var(\hat{\pi}_2) = \frac{(p_2 - p_1)^2\lambda_1(1-\lambda_1) - 2(p_2 - p_1)(p_3 - p_1)\lambda_1\lambda_2 + (p_3 - p_1)^2\lambda_2(1-\lambda_2)}{n|P|^2}$$

$$\begin{aligned} Cov(\hat{\pi}_1, \hat{\pi}_2) &= \frac{(p_3 - p_2)(p_2 - p_1)\lambda_1(1-\lambda_1) - (p_2 - p_1)^2\lambda_1\lambda_2}{n|P|^2} + \\ &+ \frac{(p_3 - p_2)(p_3 - p_1)\lambda_1\lambda_2 + (p_1 - p_2)(p_3 - p_1)\lambda_2(1-\lambda_2)}{n|P|^2}. \end{aligned}$$

Si noti che quando $k = 2$ si ha il modello di Warner.

Se dovessimo inoltre spendere qualche parola in termini di efficienza, il presente modello è molto buono quando $p_i \neq \frac{1}{3}$ ma in generale risulta di difficoltosa applicazione. Applicazioni del modello infatti non ve ne sono state.

2.21 La tecnica RR per campioni in blocco (di Kim e Flueck del 1978).

Attributi

Nel presente modello si esamineranno i quattro casi in cui domanda e rispondenti vengano estratti in blocco o con reinserimento. Si supponga quindi che N sia il numero di rispondenti, M il numero di domande e p la proporzione delle

M domande delicate. Per semplicità sia il gruppo di N persone composto da $A \in G_1$ e $(N-A) \in \bar{G}_1$ ove $\frac{A}{N} = \pi$. Analogamente il gruppo di M domande sia composto da $B \in Q_1$ e $(M-B) \in Q_2$ e $\frac{B}{M} = p$.

Siano inoltre:

$$x_i = \begin{cases} 1 & \text{se i-mo rispondente appartiene a } G_1 \\ 0 & \text{altrimenti} \end{cases}$$

$$y_i = \begin{cases} 1 & \text{se i-mo rispondente seleziona } Q_1 \\ 0 & \text{altrimenti} \end{cases}$$

$$z_i = \begin{cases} 1 & \text{se i-mo rispondente risponde } SI \\ 0 & \text{altrimenti} \end{cases}$$

allora:

$$Z_i = x_i y_i + (1-x_i)(1-y_i) = 2x_i y_i - x_i - y_i + 1.$$

Si definiscano inoltre:

$$X = \sum_{i=1}^{n} x_i$$

$$Y = \sum_{i=1}^{n} y_i$$

$$Z = \sum_{i=1}^{n} z_i$$

ove X è la variabile casuale che rappresenta il numero totale dei rispondenti che hanno la caratteristica 1, Y è la variabile casuale che rappresenta il numero totale di volte in cui Q_1 è stato selezionato e Z è la variabile casuale che rappresenta il totale dei SI. A seguire viene data la formula generale della varianza della somma di variabili casuali che potrà poi essere utilizzata in tutti e quattro i casi di studio.

$$Var(Z) = E\left(\sum_{i=1}^{n} z_i^2\right) + E\left(\sum_{i \neq j}^{n} z_i z_j\right) - \left[E\left(\sum_{i=1}^{n} z_i\right)\right]^2$$

Inoltre forniamo altre notazioni che resteranno valide per tutti i casi di specie:

$$Z \sim Bi(n, \lambda) \tag{13}$$

$$\hat{\pi} = \frac{\frac{Z}{N} - (1-p)}{2p-1} \quad \text{con} \quad p \neq \frac{1}{2} \tag{14}$$

2.21.1 Caso 1. Estrazione con reinserimento sia per i rispondenti che per la domanda.

Si tratta del modello originale di Warner che per la (13) e la (14) presenta una varianza pari a:

$$Var_1(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}$$

2.21.2 Caso 2. Estrazione della domanda con reinserimento e dei rispondenti in blocco.

In questo caso si ha che:

$$X \sim Ipj(n, N, A)$$

e

$$Y \sim Bi(n, p)$$

allora

$$E(x_i x_j) = P(X_i = 1, X_j = 1) = \frac{\pi(N\pi - 1)}{N - 1}$$

da cui

$$Var_2(\hat{\pi}) = \frac{\pi(1-\pi)}{n}\frac{N-n}{N-1} + \frac{p(1-p)}{n(2p-1)^2}$$

Da notare che Var_2 è come Var_1 ma include in più la correzione per popolazioni finite.

2.21.3 Caso 3. Estrazione dei rispondenti con reinserimento ed estrazione in blocco della domanda.

Qui si ha che:

$$X \sim Bi(n, \lambda)$$

$$Y \sim Ipj(n, N, A).$$

Si noti inoltre che y_i e y_j non sono indipendenti per $i \neq j$, pertanto:

$$E(y_i, y_j) = \frac{p(Mp - 1)}{M - 1}$$

da cui

$$Var_3(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{4\pi(1-\pi)p(1-p)}{n(2p-1)^2}\left(1 - \frac{M-n}{M-1}\right) + \frac{p(1-p)}{n(2p-1)^2}\left(\frac{M-n}{M-1}\right)$$

2.21.4 Caso 4. Estrazione in blocco sia dei rispondenti che della domanda.

Qui si ha che:

$$X \sim Ipj$$

$$Y \sim Ipj$$

Si noti inoltre che sia (x_i, x_j) che (y_i, y_j) sono dipendenti, pertanto:

$$E(x_i, x_j) = \frac{\pi(N\pi - 1)}{N - 1}$$

e

$$E(y_i, y_j) = \frac{p(Mp - 1)}{M - 1}$$

da cui

$$\begin{aligned} Var_4(\hat{\pi}) &= \frac{\pi(1-\pi)}{n(2p-1)^2}\left[(4p^2 - 4p)\frac{N(M-n) - n(M-1)}{(N-1)(M-1)} + \frac{N-n}{N-1}\right] + \\ &+ \frac{p(1-p)}{n(2p-1)^2}\frac{M-n}{M-1}. \end{aligned}$$

Si notino qui i due fattori di correzione e che all'aumentare di N, Var_4 converge verso Var_3, mentre all'aumentare di M, Var_4 converge verso Var_2

2.21.5 Il confronto tra i casi esaminati.

Se volessimo fare un confronto tra i quattro casi si ha che:

- $Var_1(\hat{\pi}) > Var_i(\hat{\pi})$ per $i = 2, 3, 4$ se $p \neq \frac{1}{2}$
- $Var_2(\hat{\pi}) \geq Var_4(\hat{\pi})$ se $p \neq \frac{1}{2}$
- $Var_3(\hat{\pi}) \geq Var_4(\hat{\pi})$ se $p \neq \frac{1}{2}$

La varianza di $\hat{\pi}$ per valori piccoli di π è spesso più piccola nei casi 3 e 4; i casi 1 e 2 sono virtualmente uguali e similmente per i casi 3 e 4; questo perchè i fattori di correzione sono prossimi a 1. Si può vedere inoltre che se π tende a 0.5 la riduzione della varianza diviene più piccola. Inoltre all'aumentare di M per n costante, la riduzione di varianza diviene minore poichè aumenta il fattore di correzione relativo alla domanda.

2.21.6 Il modello di Simmons negli stessi quattro casi.

Il metodo di Simmons è uno dei modelli più importanti e per questo motivo verrà preso in considerazione nella fattispecie delle casistiche poc'anzi analizzate. Sulla base del metodo in parola siano:

$$X_{ij} = \begin{cases} 1 & \text{se j-mo rispondente nel campione } i = 1, 2 \text{ appartiene a } G_1 \\ 0 & \text{altrimenti} \end{cases}$$

$$V_{ij} = \begin{cases} 1 & \text{se j-mo rispondente nel campione } i = 1, 2 \text{ appartiene } G_2 \\ 0 & \text{altrimenti} \end{cases}$$

$$y_{ij} = \begin{cases} 1 & \text{se j-mo rispondente nel campione } i = 1, 2 \text{ seleziona } Q_1 \\ 0 & \text{altrimenti} \end{cases}$$

$$z_{ij} = \begin{cases} 1 & \text{se j-mo rispondente nel campione } i = 1, 2 \text{ risponde } SI \\ 0 & \text{altrimenti} \end{cases}$$

Allora

$$Z_{ij} = X_{ij}Y_{ij} + V_{ij}(1 - Y_{ij}) \qquad j = 1, 2, ..., n_i \qquad i = 1, 2$$

Inoltre siano:

$$Z_{ij} = \sum_{j=1}^{n_i} z_{ij}$$

$$\pi = P(X_{ij} \in G_1)$$

e

$$\pi_2 = P(X_{ij} \in G_2)$$

Lo stimatore, proposto da *Greenberg et al.*, risulta essere pari a:

$$\pi = \frac{\frac{Z_1(1-p_2)}{n_1} + \frac{Z_2(1-p_1)}{n_2}}{p_1 - p_2} \qquad \text{con} \qquad p_1 \neq p_2$$

Definendo inoltre λ_i come la probabilità di ottenere una risposta affermativa nel campione $i = 1, 2$, allora $E(Z_i) = n_i\lambda_i$ e quindi lo stimatore di π risulta essere corretto in tutti e quattro i casi in esame. La varianza dello stimatore risulta essere la somma delle singole varianze poichè insiste la indipendenza dei due campioni. Ossia:

$$Var(\hat{\pi}) = \frac{(1-p_2)^2 Var\left(\frac{Z_1}{n_1}\right) + (1-p_1)^2 Var\left(\frac{Z_2}{n_2}\right)}{(p_1 - p_2)^2}$$

ferma restando la condizione di diseguaglianza tra p_1 e p_2. Vediamo ora, caso per caso, come si comporta la varianza dello stimatore per poter fare poi in seconda battuta alcune valutazioni.

Nel **primo caso** la varianza assume la forma seguente:

$$Var_1(\hat{\pi}) = \frac{1}{(p_1 - p_2)^2}\left[(1-p_2)^2\frac{\lambda_1(1-\lambda_1)}{n_1} + (1-p_1)^2\frac{\lambda_1(1-\lambda_2)}{n_2}\right]$$

Nel **secondo caso** invece:

$$\begin{aligned} Var_2(\hat{\pi}) &= Var_1(\hat{\pi}) - \frac{1}{(p_1 - p_2)^2} * \\ &* \frac{(1-p_2)^2(1-\frac{1}{n_1})p_1^2\pi(1-\pi) + (1-p_1)^2\pi_2(1-\pi_2)}{N-1} + \\ &+ \frac{(1-p_1)^2(1-\frac{1}{n_2})p_2^2\pi(1-\pi) + (1-p_2)^2\pi_2(1-\pi_2)}{N-1} \end{aligned}$$

Nel **terzo caso** la varianza diviene:

$$\begin{aligned} Var_3(\hat{\pi}) &= Var_1(\hat{\pi}) - \frac{1}{(p_1-p_2)^2} * \\ &* \left[(1-p_2)^2(1-\frac{1}{n_1})\frac{p_1(1-p_1)\pi^2}{M_1-1} + (1-p_1)^2(1-\frac{1}{n_2})\frac{p_2(1-p_2)\pi^2}{M_2-1}\right] \end{aligned}$$

Infine nel **quarto caso** la varianza diviene:

$$\begin{aligned} Var_4(\hat{\pi}) &= Var_1(\hat{\pi}) - \frac{1}{(p_1-p_2)^2}(1-p_2)^2\left(1-\frac{1}{n_1}\right) * \\ &* \frac{N\pi^2 p_1(1-p_1) + M_1 p_1^2 \pi(1-\pi)\pi p_1(1-\pi p_1)}{(N-1)(M_1-1)} + \\ &- \frac{2\pi\pi_2 p_1(1-p_1)}{M_1-1} + \frac{(1-2p_1)\pi_2(1-\pi_2)}{N-1} + \\ &+ \frac{N\pi_2^2 p_1(1-p_1) + M_1 p_1^2 \pi_2(1-\pi_2) - \pi_2 p_1(1-\pi_2 p_1)}{(N-1)(M_1-1)} + \\ &+ (1-p_1)^2(1-\frac{1}{n_2})\frac{N\pi^2 p_2(1-p_2) + M_2 p_2^2\pi(1-\pi) - \pi p_2(1-\pi p_2)}{(N-1)(M_2-1)} + \\ &- \frac{2\pi\pi_2 p_2(1-p_2)}{M_2-1} + \frac{(1-2p_2)\pi_2(1-\pi_2)}{N-1} + \\ &+ \frac{N\pi_2^2 p_2(1-p_2) + M_2 p_2^2 \pi_2\pi_2(1-\pi_2) - \pi_2 p_2(1-\pi_2 p_2)}{(N-1)(M_2-1)} \end{aligned}$$

I confronti tra le quattro variabilità hanno dimostrato che quella del primo caso è in generale sempre maggiore; altri confronti numerici sono stati fatti con le dimensioni campionarie, n_1 e n_2, uguali, e con la distribuzione dei caratteri A e Y identiche. In generale la variabilità del quarto caso è risultata la più bassa. In particolare si è notato che all'aumentare di π_2, l'efficienza relativa aumenta del 67% se $\pi = 0,1$, $\pi_2 = 1,0$, $p_1 = 0,7$, $p_2 = 0,3$, $N = 100.000$, $M_1 = M_2 = 50$ e $n_1 = n_2 = 50$.

2.22 Il modello RR nel campionamento a due stadi (di Marasini del 1981).

Attributi

Si suppone che la popolazione oggetto di studio sia stratificabile e il classico modello RR viene adattato al caso di due stadi semplici con scelta senza ripetizione in entrambi gli stadi. Inoltre si suppone siano note a priori le numerosità dei singoli strati. Verrà fornita una stima corretta e consistente della media del fenomeno quantitativo oggetto di studio sia a livello di strato che di popolazione complessiva.

Sia quindi P la popolazione oggetto di studio di N unità, ripartibili in k strati S_i di numerosità N_i; sia X il fenomeno quantitativo in studio e x_{iv} l'intensità con cui il fenomeno si presenta sulla $v-ma$ unità dello strato S_i, per $v = 1, 2, ..., N_i$.

Supponendo ignota la media μ_1 di X si vuole ottenere una stima $\hat{\mu}$ attraverso la seguente procedura:

1. a ciascuno strato S_i viene associato un numero intero positivo n_i di esperimenti casuali E_{ij}, tra loro indipendenti ognuno dei quali genera m eventi elementari necessari ed incompatibili di probabilità rispettivamente, di $\lambda_{ij1}, \lambda_{ij2}, ..., \lambda_{ijr}, ..., \lambda_{ijm}$; il generico evento elementare viene indicato con $E_{ijr} = (\alpha_{ijr}, \beta_{ijr})$ essendo queste coppie di numeri reali con $\alpha_{ijr} \neq 0 \quad \forall r$ e

$$\sum_{r=1}^{m} \lambda_{ijr} = 1;$$

2. considerati i k strati si effettua una scelta, quella di primo stadio, senza ripetizione di h strati; evidentemente $h \leq k$;

3. in ognuno degli h strati viene operata una scelta, quella di secondo stadio, senza ripetizione, per un numero di unità pari a quello ad essi associato;

4. per tutelare il rispondente così estratto, esso esegue un esperimento casuale del tipo sopra descritto;

Il rispondente poi trasforma il vero valore x_{ij} posseduto su X nel seguente modo:

$$y_{ij} = \alpha_{ij} x_{ij} + \beta_{ij}$$

e lo comunica all'operatore che ignora l'esito dell'esperimento casuale.
La stima della media dello strato S_i è data da:

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{y_{ij} - \beta_{ij}}{\alpha_{ij}},$$

tale stima è corretta e consistente.

2.23 Modelli RR multivariati per dati categoriali (di Bourke del 1982).

Attributi

In termini generali si vuole stimare la distribuzione congiunta di g variabili categoriali ove la variabile i consti di $m(i)$ modalità (con $i = 1, 2, ..., m(g)$). Così si vuole stimare $\pi_{c_1,c_2,...,c_g}$, la proporzione della popolazione che è nella categoria c_1 per la variabile 1, c_2 per la variabile 2, ..., c_g per la variabile g, fissato $c_1 = 1, 2, ..., m(1), c_2 = 1, 2, ..., m(2), ..., c_g = 1, 2, ..., m(g)$.

Per illustrare la metodologia nel dettaglio si considera il caso bivariato; siano quindi S e T le due variabili delicate e si voglia stimare π_{ij}, la proporzione della popolazione che è nella categoria i per S e in quella j per T; in totale quindi si debbono stimare $m_1 \text{x} m_2$ proporzioni. Si assume inoltre che $m(1) = 2$ e $m(2) = 2$. Facendo uso del casualizzatore, esso deve prevedere quattro diversi risultati dell'esperimento casuale come di seguito illustrato, essendo p_i la probabilità che esca $i = 1, 2, 3, 4$.

1. Se esce **1** allora rispondi qui:
   (a) ST_{11}
   (b) ST_{12}
   (c) ST_{21}
   (d) ST_{22}
2. Se esce **2** allora rispondi qui:
   (a) ST_{12}
   (b) ST_{21}
   (c) ST_{22}
   (d) ST_{11}
3. Se esce **3** allora rispondi qui:
   (a) ST_{21}
   (b) ST_{22}
   (c) ST_{11}
   (d) ST_{12}
4. Se esce **4** allora rispondi qui:
   (a) ST_{22}
   (b) ST_{11}
   (c) ST_{12}
   (d) ST_{21}

ove ST_{ij} denota una affermazione che porta con sè l'appartenenza in i per S e in j per T. Di fatto ogni rispondente esegue l'esperimento casuale, il cui risultato è ignoto all'intervistatore, e sceglie la lettera che lo identifica nella colonna selezionata casualmente. Se λ_i indica la probabilità che un qualsiasi rispondente fornisca la risposta i e n_i è il numero osservato di tali risposte $(i = 1, 2, 3, 4)$, allora:

$$\begin{aligned}
\lambda_1 &= p_1\pi_{11} + p_2\pi_{12} + p_3\pi_{21} + p_4\pi_{22} \\
\lambda_2 &= p_1\pi_{12} + p_2\pi_{21} + p_3\pi_{22} + p_4\pi_{11} \\
\lambda_3 &= p_1\pi_{21} + p_2\pi_{22} + p_3\pi_{11} + p_4\pi_{12} \\
\lambda_4 &= p_1\pi_{22} + p_2\pi_{11} + p_3\pi_{12} + p_4\pi_{21}
\end{aligned}$$

In notazione matriciale, si ha che:

$$\Lambda = \mathbf{D}\Pi$$

ove Λ e Π sono i vettori colonna di λ_i e π_{ij} rispettivamente, mentre $\mathbf{D}$ è così composta:

$$D = \begin{bmatrix} p_1 & p_2 & p_3 & p_4 \\ p_4 & p_1 & p_2 & p_3 \\ p_3 & p_4 & p_1 & p_2 \\ p_2 & p_3 & p_4 & p_1 \end{bmatrix}$$

La stima di massima verosimiglianza per il vettore π è data da:

$$\hat{\pi} = \mathbf{D}^{-1}\hat{\lambda}$$

ove $\hat{\lambda}$ è il vettore di elementi $\lambda_i = \frac{n_i}{n}$.
La stima è tale per cui

$$\hat{\pi}_{ij} \geq 0 \quad \forall(i,j) \quad \text{e} \quad \sum_i \sum_j \hat{\pi}_{ij} = 1$$

Tale stimatore risulta essere asintoticamente corretto. Va notato che per un numero elevato di variabili o di categorie (modalità), se si considera il caso bivariato, il numero di proporzioni da stimare cresce rapidamente e nella pratica questo si riflette nel fatto che il rispondente debba scegliere il suo status in una lista molto lunga. In talune situazioni il presente modello perde di praticità e di efficacia.

2.23.1 Il caso particolare della randomizzaizone separata per ogni variabile.

Si presenta qui il caso di tre variabili; ma una generalizzazione non risulterebbe difficile ma soltanto di pesante notazione. Sia quindi R costituito di due modalità, S di tre e T di due. Ora, vista la specificità del caso, ovvero l'utilizzo di uno schema diverso per ogni variabile, si assume lo schema di *Simmons* per R, quello di *Bourke* e *Delenius* per S e il disegno simmetrico della domanda incorrelata proposto da *Bourke* per T. Vengono inoltre utilizzati tipi diversi di carte per ogni variabile secondo il seguente schema.

Carte per la variabile R.

1. Se esce 1 allora
   (a) R_1
   (b) R_2
2. Se esce 2 allora
   (a) U_1
   (b) U_2

Carte per la variabile S.

1. Se esce 1 allora
   (a) S_1

(b) S_2

(c) S_3

2. Se esce 2 allora

   (a) S_2

   (b) S_3

   (c) S_1

3. Se esce 3 allora

   (a) S_3

   (b) S_1

   (c) S_2

Carte per la variabile T.

1. Se esce 1 allora

   (a) T_1

   (b) T_2

2. Se esce 2 allora

   (a) T_2

   (b) T_1

3. Se esce 3 allora

   (a) V_1

   (b) V_2

Vediamo un pò di notazione. S_i indica l'appartenenza alla categoria i della variabile S, $i = 1, 2, 3$. Le variabili U e V sono le variabili incorrelate, necessarie al disegno che fa uso di domanda non delicata e incorrelata. Le affermazioni U_1 e U_2 sono mutuamente esclusive e collettivamente esaustive, come anche per V_1 e V_2. Si assuma anche che vengano utilizzati casualizzatori diversi per ogni variabile oggetto di studio e sia a_i la probabilità che il casualizzatore per la variabile R produca come uscita i, $i = 1, 2$ e similmente siano b_j e f_k le corrispondenti probabilità per i casualizzatori utilizzati per le variabili S e T. Sia ora λ_{ijk} la probabilità che un rispondente scelto a caso fornisca come risposta il vettore $[i, j, k]$ e sia n_{ijk} il numero di identiche risposte. Si ha

$$\lambda_{ijk} = \sum_{C_1} \sum_{C_2} \sum_{C_3} P[risposta(i,j,k)|rispondente(C_1, C_2, C_3)]\pi_{C_1,C_2,C_3}$$

con $i = 1, 2, 3$, $j = 1, 2, 3$ e $k = 1, 2$.
Dallo sviluppo di λ si ottiene che:

$$\lambda = \mathbf{D}\pi$$

ove λ e π sono vettori colonna e $\mathbf{D}$ è definita dal prodotto di tre matrici [2]. Da qui segue che:

$$\hat{\pi} = (\mathbf{D})^{-1} Var(\hat{\lambda})(\mathbf{D}^{-1})'$$

e la forma di $Var(\hat{\lambda})$ segue facilmente dalla distribuzione asintotica multinomiale di $\hat{\lambda}$.

2.23.2 Il caso particolare della randomizzaizone unica per tutte le variabili.

L'idea si muove dal presupposto di avere due variabili, la prima delle quali di due modalità e la seconda delle quali con quattro modalità. Risulta necessario quindi un casualizzatore con almeno quattro possibili esiti e le affermazioni possono essere così riassunte:

Carte per la variabile 1.

1. Se esce **1 o 2** allora rispondi qui:
   (a) S_1
   (b) S_2
2. Se esce **3 o 4** allora rispondi qui:
   (a) S_2
   (b) S_1

Carte per la variabile 2.

1. Se esce **1** allora rispondi qui:
   (a) T_1
   (b) T_2
   (c) T_3
   (d) T_4
2. Se esce **2** allora rispondi qui:
   (a) T_2
   (b) T_3
   (c) T_4
   (d) T_1
3. Se esce **3** allora rispondi qui:
   (a) T_3
   (b) T_4

[2] Per i dettagli si rinvia a Patrick D. Bourke, 1982.

(c) T_1

(d) T_2

4. Se esce **4** allora rispondi qui:

   (a) T_4

   (b) T_1

   (c) T_2

   (d) T_3

Al rispondente è richiesto di utilizzare il casualizzatore e di memorizzare l'uscita che è ignota all'intervistatore, per poi dare la risposta definitiva in accordo al risultato del casualizzatore e allo status posseduto. La stima di π è del tutto analoga a quella del caso precedente.

2.24 Un modello RR scrambled (di Eichhorn e Hayre del 1983).

Variabili

Gli autori introducono questo modello per stimare la media e la varianza di un carattere definito delicato e indicato con X. Ogni rispondente selezionato dal campionamento viene istruito ad utilizzare un casualizzatore e a generare un numero casuale, detto S, da alcune predeterminate distribuzioni come la Chi-quadrato, la Uniforme, la Poisson, la Binomiale o la Weibul. La distribuzione della variabile casuale S è assunta nota come anche la media θ e la varianza γ^2 della variabile di camuffamento. All'$i-mo$ rispondente del campione di dimensione n, estratto con tecnica casuale semplice con reinserimento, viene chiesto di riportare il valore

$$Z_i = \frac{S_i X_i}{\theta}$$

come risposta camuffata del vero valore posseduto relativamente alla variabile sensibile X.

Uno stimatore corretto della media di X, μ_X, è dato da:

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

la cui varianza è:

$$Var(\hat{\mu}_1) = \frac{1}{n}[\sigma_X^2 + C_y^2(\sigma_X^2 + \mu_X^2)]$$

ove

$$C_y = \frac{\gamma}{\theta}$$

indica il noto coefficiente di variazione della variabile di camuffamento S.

2.25 Lo schema di Simmons in una versione modificata (di Olivieri del 1983).

Attributi

La modifica dello schema di Simmons qui proposta consiste nell'avere la domanda o l'affermazione alternativa, a quella delicata A, fissa. Ogni unità statistica infatti esegue un esperimento casuale E che consta di due eventi elementari necessari ed incompatibili di probabilità costante in ogni prova e pari rispettivamente a λ e $1-\lambda$. Al primo evento viene associata la domanda delicata relativa al possedere A, mentre al secondo evento viene associata una risposta fissa e pari a SI [3].

Passando ora agli aspetti analitici, si supponga che nella popolazione oggetto di studio una certa frazione di persone, pari a $\pi = \frac{C}{N}$, possegga la caratteristica delicata A. Per la stima del parametro ignoto viene estratto un campione di n unità e la natura della stima stessa dipenderà dal fatto che lo stesso campione sia estratto con reiserimento o in blocco. Vediamo quindi questi due casi distintamente.

Nel caso di campionamento bernulliano sia

$$f = \frac{x}{n}$$

quindi uno stimatore corretto, consistente e di massima verosimiglianza di π, con distribuzione standardizzata e asintoticamente normale è:

$$p = \frac{f - (1-\lambda)}{\lambda}$$

la cui varianza è:

$$Var(p) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)(1-\lambda)}{n\lambda}$$

ove il secondo addendo è la variabilità aggiuntiva dello stimatore dovuta alla particolare tecnica e quindi molto utile per la valutazione in una sua applicazione.

Nel caso di campionamento in blocco è da evidenziare subito che la variabile casuale che sottende a questo tipo di campionamento è la ipergeometrica che qui può essere evidenziata con parametri pari a N, $\pi = \frac{C}{N}$ e n. Va da sè che cambiano le probabilità di ottenere risposte affermative o negative in quanto uno stesso intervistato non potrà più far nuovamente parte del campione e a ragion di logica pare che questa seconda alternativa risulti anche più realistica dal punto di vista applicativo. Per quanto riguarda lo stimatore esso non muta nella forma, risulta avere le stesse proprietà prima esaltate e con varianza pari a:

$$Var(p) = \frac{\pi(1-\pi)}{n}\frac{N-n}{N-1} + \frac{(1-\pi)(1-\lambda)}{n\lambda}$$

ove si evidenziano il fattore di correzione dovuto al tipo di campionamento e la aggiunta quota di variabilità dovuta alla tecnica RR.

Il confronto del modello presentato con il modello di Warner fa emergere una maggiore efficienza del primo nel caso in cui $\lambda > 0.33$ per ogni livello di

[3] La risposta fissa sarà NO nel caso in cui la negazione dichiari l'appartenenza al gruppo A.

π. Il confronto con il modello di Simmons invece dipende dalla probabilità associata al quesito delicato: se essa è alta risulta essere più efficiente il modello di Olivieri; viceversa si presentano situazioni alterne. In generale comunque non si registrano grosse differenze, anche se qui il confronto è stato reso possibile solo in un caso specifico del modello di Simmons, ossia con π_Y noto.

2.26 Lo schema di Poole per la stima dei parametri (di Olivieri del 1984).

Con schema di Poole si arriva a determinare la distribuzione di una variabile definita delicata. L'autore invece propone la stima della media μ_X e della varianza σ_X^2 della variabile quantitativa che indicheremo con X e avente inoltre funzione di densità $f(x)$. Sia ora Y una variabile indipendente da X con funzione di probabilità nota $f(y)$, di media μ_Y e varianza σ_Y^2, entrambe note e con $(\mu_Y \neq 0)$. Come è noto da Poole, l'intervistato deve restituire all'intervistatore il prodotto di X, vero valore della variabile delicata posseduto, per Y, valore estratto a caso; ossia:

$$Z = XY$$

Da qui Poole ricava la funzione di probabilità di X, mentre se si vuol stimare la media e la varianza di X, estraendo un campione bernoulliano di numerosità n, la quantità

$$\frac{\sum_{i=1}^{n} z_i}{n\mu_Y} = \frac{m}{\mu_Y}$$

con

$$m = \frac{\sum_{i=1}^{n} z_i}{n}$$

è uno stimatore corretto di μ_X.

Lo stimatore invece della varianza di X, σ_X^2 risulta essere pari a:

$$s^{\star 2} = \frac{s^2(\sigma_Y^2 + n\mu_Y^2) - nm^2\sigma_Y^2}{n\mu_Y^2(\sigma_Y^2 + \mu_Y^2)}$$

che risulta essere corretto.

In termini di efficienza, il confronto con il campionamento bernulliano metta in evidenza che:

1. la variabilità delle stime è sempre maggiore nello schema di Poole;
2. la protezione del rispondente comporta dei costi che sono direttamente proporzionali al quadrato del *c.v.* dei numeri casuali associati alla variabile delicata;
3. l'efficienza aumenta con l'aumentare della variabilità della variabile delicata X;
4. la protezione del rispondente è tanto maggiore quanto maggiore è il numero k dei numeri casuali, mentre l'efficienza risulta decrescente all'aumentare di k.

Quindi la distribuzione dei numeri casuali da una parte risulta essere protettiva per il rispondente, dall'altra condiziona la variabilità delle stime. Pertanto l'adozione di una distribuzione piuttosto che di un'altra merita una accorta valutazione.

2.27 Il campionamento RR con risposta alternativa fissa dotato di memoria e la stratificazione della popolazione (di Olivieri del 1984).

Attributi

Ogni intervistato deve estrarre una palla, senza reinserimento, da un'urna che ne contiene N, di cui X di colore rosso. Se la palla estratta è di colore rosso allora risponde al quesito delicato, altrimenti risponde comunque "SI". Se R è il numero dei "SI" ottenuti nelle N prove, di questi $(N - X)$ derivano da soggetti che non hanno estratto la palla rossa. Ne segue che delle X persone che hanno estratto la pallina rossa si sono avute $S = R - (N - X)$, con $(R \geq N - X)$, risposte affermative $(S = 0, 1, 2, ..., X)$. Si dimostra che:

$$P = \frac{S}{X} = \frac{f - (1 - \lambda)}{\lambda} = \frac{S}{N\lambda}$$

con

$$f = \frac{R}{N} \qquad \text{e} \qquad \lambda = \frac{X}{N}$$

è stima corretta, consistente e di massima verosimiglianza di π, con distribuzione binomiale e di parametri X e π e con varianza pari a:

$$Var(P) = \frac{\pi(1 - \pi)}{\lambda N}$$

Si supponga ora che la popolazione di H unità oggetto di studio sia suddivisibile in k strati disgiunti e di numerosità H_i ciascuno, in modo tale che:

$$\sum_{i=1}^{k} H_i = H$$

e che π_i sia la ignota proporzione del carattere delicato D nello strato i.
Come è noto dalla metodologia:

$$\pi = \sum_{i=1}^{k} \frac{H_i}{H} \pi_i.$$

Da ciascuno strato ora si estragga un camione bernoulliano di N_i unità e si adotti la stessa tecnica sopra descritta. In base a quanto esposto prima la stima corretta di π_i diviene:

$$P_i = \frac{S^{(i)}}{N_i \lambda_i} = \frac{R_i - (N_i - X_i)}{X_i} = \frac{f_i - (1 - \lambda_i)}{\lambda_i}$$

per $i = 1, 2, ..., k$. Ora una stima corretta per π è la seguente:

$$P = \sum_{i=1}^{n} \frac{H_i}{H} P_i = \sum_{i=1}^{n} \frac{H_i}{H} \frac{S^{(i)}}{N_i \lambda_i}.$$

Una stima corretta della varianza negli strati è:

$$s_n^2 = \sum_{i=1}^{n} \frac{H_i}{H} s_i^2$$

ove

$$s_i^2 = \frac{1}{X_i - 1} X_i p_i (1 - p_i)$$

Mentre una stima corretta della varianza fra gli strati è:

$$s_f^2 = \sum_{i=1}^{n} \frac{H_i}{H} (p_i - p)^2 - \sum_{i=1}^{n} \frac{H_i}{H} \left(1 - \frac{H_i}{H}\right) \frac{s_i^2}{N_i \lambda_i}.$$

Come è noto dalla metodologia, l'efficienza della stratificazione risulta maggiore se si adotta una numerosità campionaria per strato proporzionale alla dimensione dello strato medesimo, ossia se:

$$N_i = \frac{H_i}{H} N;$$

in questo caso la varianza dello stimatore diviene:

$$Var(P)_{prop} = \sum_{i=1}^{n} \frac{H_i}{H} \frac{\sigma_i^2}{N \lambda_i} = \frac{1}{N\lambda} \sum_{i=1}^{n} \frac{H_i}{H} \sigma_i^2.$$

Da qui si evince che:

$$Var(P)_{prop} \leq Var(P).$$

Infine un ulteriore miglioramento potrtebbe ottenersi adottando il dimensionamento di Neymann per strato pari a:

$$N_i = N \frac{H_i \sigma_i}{\sum_{i=1}^{n} H_i \sigma_i}.$$

2.28 Un modello a contaminazione (scrambled) per ottenere dati quantitativi (di Eichhorn e Hoyre del 1993).

Variabili

Questo modello fa parte di quelli proposti in letteratura detti a contaminazione della risposta. Qui si vogliono ottenere dati quantitativi e per ciò il rispondente fornisce all'intervistatore un valore risultante da una opportuna elaborazione del vero valore posseduto, con riguardo ad una variabile ritenuta delicata. In particolare si tratta di una contaminazione moltiplicativa con nota distribuzione dei numeri casuali utilizzati come moltiplicatori. Come al solito il ricercatore conosce solo la distribuzione di detti numeri casuali ma non conosce nella fattispecie il singolo moltiplicatore estratto casualmente dall'intervistato.

Anche questa tecnica comunque è un caso speciale del modello lineare introdotto da Warner nel 1971, considerato il pioniere anche dei modelli a contaminazione.

Sia X quindi la risposta alla domanda delicata e sia S la variabile casuale indipendente da X e avente media e varianza finita. Si assuma inoltre che X sia non negativa e S sia positiva. L'intervistatore, con queste premesse, genererà, con un qualche meccanismo, un valore casuale di S che utilizzerà come fattore per fornire la risposta Y, che sarà del tipo:

$$Y = XS.$$

Siano ancora:

$$\theta = E(S) \qquad \gamma^2 = Var(S) \qquad \mu = E(X) \qquad \sigma^2 = Var(X)$$

ove (θ, γ^2) sono parametri noti, mentre (μ, σ^2) sono parametri ignoti. Allora

$$E(Y) = E(XS) = E(X)E(S) = \mu\theta \qquad \text{e}$$

$$V(Y) = E(X^2)E(S^2) - [E(X)]^2[E(S)]^2 = \gamma^2 E(X^2) + \sigma^2\theta^2 = \gamma^2(\sigma^2 + \mu^2) + \sigma^2\theta^2.$$

Considerando ora le osservazioni $y_1, y_2, ..., y_n$, uno stimatore corretto per μ è:

$$\hat{\mu} = \frac{\bar{y}}{\theta}$$

e

$$Var(\hat{\mu}) = \frac{1}{n\theta^2} Var(Y) = \frac{1}{n}\left(\sigma^2 + \frac{\gamma^2 E(X^2)}{\gamma^2}\right) = \frac{1}{n}(\sigma^2 + \rho^2 E(X^2))$$

ove

$$\rho^2 = \left(\frac{\gamma}{\theta}\right).$$

Un intervallo di confidenza per μ può essere ottenuto stimando $Var(Y)$ con:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

In molte applicazioni però n supera le 30 unità potendo così assumere la normalità asintotica di $\bar{\mu}$, quindi

$$\hat{\mu} \pm z_{\frac{\alpha}{2}} \frac{s}{\theta\sqrt{n}}.$$

In tema di privacy del rispondente, la esperienza ha insegnato che l'intervistato si sente maggiormente protetto da questi tipi di modelli a contaminazione che non da quelli come Warner e Simmons. Tuttavia resta oggettivo il fatto che la tecnica non risulta essere così semplice come appare: bisogna adottare un meccanismo di casualizzazzione che generi numeri casuali e che sia di semplice uso, bisogna adottare un calcolatore che permetta di fare di conto, e ciò potrebbe far insospettire i rispondenti (potrebbe essere un meccanismo di memorizzazione dell'operazione di moltiplicazione!).

Per quanto concerne la scelta della distribuzione di S ci si trova di fronte a conflitti oggettivi. Per contenere la frequenza di risposte evasive, S potrebbe essere assunta in un ampio range di valori con elevata probabilità. Per stimare μ nel migliore dei modi però si nota da $V(\hat{\mu})$ che dovremmo prendere $\rho = \frac{\gamma}{\theta}$ il più piccolo possibile. In generale infine sembra essere ragionevole assumere che la media di S sia pari a 1.

Consideriamo ora due distribuzioni particolari per S. La prima assume che S sia del tipo:

$$P(S = n) = P(S = n^{-1}) = 2^{-n} \quad \text{per} \quad n = 2, 3, ... \tag{15}$$

che potrebbe essere così generata: il rispondente lancia due monete fintanto che esse non presentino la stessa uscita; la probabilità che n lanci presentino il medesimo risultato è dato da $2^{-(n-1)}$ per $n \geq 2$ e, quindi se fissiamo $S = n$ allora S si distribuisce come nella (15).

Un modo alternativo per generare S è quello di avere una sacca contenente palle di due diversi colori in eguale numero e di chiedere al rispondente di estrarre con reinserimento due palle finchè esse non siano dello steso colore. Qui

$$\theta = \sum_{n=2}^{\infty} (n + n^{-1}) 2^{-n} = 1 + \log_e 2.$$

Per calcolare $\gamma^2 = Var(S)$ si noti che

$$E(S^2) = \sum_{n=2}^{\infty} (n^2 + n^{-2}) 2^{-n}$$

e facendo uso di alcune regole

$$\sum n^2 2^{-n} = \frac{11}{12} \quad \text{e} \quad \sum n^2 2^{-n} \simeq 0.08.$$

Così,

$$\gamma^2 = E(S^2) - \theta^2 = 5.58 - (1.6931)^2 = 2.7134 \qquad \text{da cui}$$

$$\rho^2 = \left(\frac{\gamma}{\theta}\right)^2 = 0.9465.$$

La distribuzione di S quindi è semplice da generare, ha mediana prossima all'unità ed ha un ragionevole valore osservato di ρ.

Si propongono ora due metodi che permettono la riduzione della varianza dello stimatore $\hat{\mu}$. Il primo consiste nel far sottrarre il valore A da X prima della moltiplicazione con S. La risposta sarà del tipo:

$$Y = S(X - A)$$

e la stima della ignota media diviene così

$$\hat{\mu} = A + \frac{\overline{y}}{\theta}$$

con

$$Var(\hat{\mu}) = n^{-1}\{\sigma^2 + \rho^2 E[(X - A)^2]\}$$

Una scelta coscienziosa di A può far ridurre notevolmente la varianza di stima e infatti se ponessimo $A = \mu$, allora

$$Var(\hat{\mu}) = n^{-1}\sigma^2(1 + \rho^2)$$

con una riduzione della varianza pari a

$$\frac{\rho^2\mu^2}{n};$$

nella pratica μ è ignoto ma possiamo avere informazioni a priori sul fenomeno oggetto di studio.

Il secondo metodo è quello di ottenere numerose informazioni per ogni rispondente; così l'intervistato i genera numeri casuali del tipo $S_{i1}, ..., S_{im}$ dalla distibuzione S e fornisce m risposte del tipo $X_iS_{i1}, X_iS_{i2}, ..., X_iS_{im}$, ove X_i è la risposta vera alla domanda delicata e m è un numero intero positivo. Ci sono due ragioni che giustificano questa scelta: la prima è che si ottiene una forte riduzione della varianza di stima e la seconda è che se m è piccolo, diciamo tra 2 e 4, il rispondente si sente ancor più protetto della medesima tecnica con risposta singola.

Sia quindi:

$$Y_{ij} = X_iS_{ij} \quad \text{per} \quad i = 1, 2, ..., n \quad \text{e} \quad j = 1, 2, ..., m.$$

Sia anche

$$\bar{Y_i} = \frac{(Y_{i1} + ... + Y_{i2})}{m} = X_i\bar{S}_i$$

In questo specifico caso $\bar{Y_i}$ è la $i - ma$ osservazione, la variabile scrambling è $\bar{S}_i$ che presenta media pari a θ e varianza pari a $\frac{\gamma^2}{m}$. Quindi:

$$Var(\hat{\mu}) = n^{-1}[\sigma^2 + m^{-1}\rho^2 E(X^2)];$$

La pratica consiglia di prendere in genere $m = 3$ e un valore di ρ^2 grande in modo tale che il rapporto tra variabilità dello stimatore e protezione del rispondente sia massimo.

2.29 Un modello per la stima di parametri di variabili che utilizza il metodo del rapporto (di Abel, Abel, Sultan e Abdel del 1985).

Variabili

Con la metodologia proposta si vuole ottenere una stima della media e della varianza di una caratteristica delicata, detta A e riferita ad una variabile quantitativa, basandosi sulla procedura di stima detta del rapporto. Si vogliono utilizzare inoltre informazioni relative ad alcune caratteristiche (variabili) ausiliare, dette B, che risultino correlate con la variabile delicata A. Come si intuisce a differenza del metodo di Simmons, B può riferirsi anche ad una variabile correlata o non correlata ad A ed anch'essa può risultare più o meno delicata.

Se consideriamo il caso in cui B è correlato con A, una stima di μ_A può essere:

$$\hat{\mu}_{a|r} = \hat{R}\hat{\mu}_{b1} = \frac{\hat{\mu}_a}{\hat{\mu}_b}\hat{\mu}_{b1}$$

ove $\hat{\mu}_a$ e $\hat{\mu}_b$ sono la media di A e di B desunte da due campioni indipendenti estratti dalla popolazione oggetto di studio e di dimensione rispettivamente pari a n_1 e n_2.

Vi sono alcuni casi in cui la caratteristica B è tale per cui possa essere nota a priori la sua media, come ad esempio, se si richiede la data di matrimonio o la data di compleanno. In questo caso specifico

$$\hat{\mu}_a = \frac{(1-p_2)\bar{Z}_1 - (1-p_1)\bar{Z}_2}{p_1 - p_2}$$

$$\hat{\mu}_b = \frac{p_2\bar{Z}_1 - p_1\bar{Z}_2}{p_2 - p_1}$$

da cui

$$\hat{\mu}_{a|r} = \mu_b \frac{(1-p_2)\bar{Z}_1 - (1-p_1)\bar{Z}_2}{p_1\bar{Z}_2 - p_2\bar{Z}_1}$$

ove:

- p_i è la probabilità di selezionare la domanda sensibile nel campione $i = 1, 2$ con $p_1 \neq p_2$;
- $\bar{Z}_i$ la media campionaria delle due variabili A e B rispettivamente.

Per quanto riguarda la varianza dello stimatore si noti che $\hat{R}$ è distorto ma si può ottenere una stima della sua varianza e quindi anche dello stimatore proposto $\hat{\mu}_{a|r}$. Ossia:

$$Var(\hat{\mu}_{a|r}) = Var(\hat{\mu}_a) + R^2 Var(\hat{\mu}_b) - 2R\rho\sqrt{Var(\hat{\mu}_a)Var(\hat{\mu}_b)}$$

ove

$$\rho = Cor(A, B) \qquad \text{e} \qquad R = \frac{\mu_a}{\mu_b}$$

e

$$Var(\hat{\mu}_a) = \frac{(1-p_2)^2 Var(\bar{Z}_1) + (1-p_1)^2 Var(\bar{Z}_2)}{(p_1 - p_2)^2}$$

$$Var(\hat{\mu}_b) = \frac{p_2^2 Var(\bar{Z}_1) + p_1^2 Var(\bar{Z}_2)}{(p_2 - p_1)^2}$$

$$Var(\bar{Z}_i) = \frac{1}{n_i}[\sigma_b^2 + p_i(\sigma_a^2 - \sigma_b^2) + p_i(1-p_i)(\mu_a - \mu_b)^2] \qquad i = 1, 2$$

Si può notare che:

1. $Var(\hat{\mu}_{a|r})$ diminuisce all'aumentare di $(n_1 + n_2)$;
2. la varianza di questo stimatore è minore di quella del modello di Simmons quando $\rho = \frac{R}{2}\frac{\sigma_{\hat{\mu}_b}}{\sigma_{\hat{\mu}_a}}$;
3. la varianza diminuisce all'avvicinarsi di ρ all'unità;
4. la varianza diminuisce se: $\mu_a = \mu_b$, $\sigma_a^2 = \sigma_b^2$ e se $p_1 \neq p_2 \neq \frac{1}{2}$.

La distorsione dello stimatore è pari a:

$$\frac{1}{\mu_b}[RVar(\hat{\mu}_b) - \rho\sqrt{Var(\mu_a)Var(\mu_b)}].$$

In generale questo stimatore è migliore di quello di Simmons quando $\rho \to 1$ ossia quando i caratteri A e B sono fortemente correlati. Inoltre dalla pratica si evince che è bene che le medie e le varianze dei due caratteri siano il più simili possibile, che presentino stessa unità di misura e che la somma delle probabilità di estrazione dell'affermazione delicata non sia pari a 1 e lontana dal valore 0.5, il più possibile in coerenza con il sospetto che ciò potrebbe creare nei rispondenti. Nel caso particolare in cui non vi sia coincidenza nelle medie di A e B è bene che la media di A sia inferiore rispetto a quella di B.

2.30 Due stimatori per campionamento RR con distribuzione continua da popolazione dicotomica (di Franklin del 1989).

Variabili

La proposta è di presentare un modello RR per una popolazione dicotomica facendo riferimento al modello di Franklin del 1977, ma facendo uso di una distribuzione dei numeri casuali di tipo normale.

Sia θ la proporzione degli appartenenti alla categoria delicata A. Un campione casuale semplice di $n \geq 1$ unità viene estratto da una popolazione potenzialmente infinita, in modo da far coincidere l'estrazione in blocco con quella con reinserimento. Ogni rispondente condurrà $n \geq 1$ prove e il rispondente i nella prova j estrae casualmnete due valori dalle densità g_{ij} e h_{ij} rispettivamente senza che l'intervistatore sappia quali siano detti valori e da dove essi provengano. Si assume inoltre l'indipendenza tra le due densità g_{ij} e h_{ij}. Il rispondente infine dovrà fornire il valore di g_{ij}, se egli possiede la cratteristica delicata A, il valore di h_{ij} altrimenti.

L'intervistatore (il ricercatore) conosce soltanto le distribuzioni esatte di g_{ij} e h_{ij} e riceve in risposta il valore z_{ij} caratterizzante il rispondente i alla prova j. Supponendo noti i parametri media e varianza delle due densità, si suppone che la loro forma sia di tipo normale, ossia:

$$g_j \sim N(\mu_{1j}, \sigma_{1j}) \qquad h_j \sim N(\mu_{2j}, \sigma_{2j}).$$

Si noti che nel caso particolare in cui

$$g_j \sim Be(\pi) \qquad \text{e} \qquad h_j \sim Be(\pi)$$

con

$$h_j(z) = 1 - g_j(z) \qquad \text{e} \qquad k = 1$$

si torna allo schema di Warner.

In generale si possono produrre due stimatori, uno basato sulle medie delle righe di Z_{ij} e l'altro basato sulle medie delle colonne sempre di Z_{ij}. Fissato che:

$$\bar{Z}_{i.} = \frac{1}{k}\sum_{j=1}^{k} E(Z_{ij}) = \frac{1}{k}\sum_{j=1}^{k} E(Z_j)$$

visto che le $k - ple$ sono indipendenti e identicamente distribuite.

Ora

$$E(Z_j) = \theta\mu_{1j} + (1-\theta)\mu_{2j} \qquad \text{per} \quad j = 1, 2, ..., k$$

$$E(\bar{Z}_{i.}) = \frac{1}{k}\left[\theta\sum_{j=1}^{k}\mu_{1j} + (1-\theta)\sum_{j=1}^{k}\mu_{2j}\right] = \frac{1}{k}[\theta m_1 + (1-\theta)m_2]$$

ove

$$m_i = \sum_{j=1}^{k}\mu_{ij} \qquad \text{per} \quad i = 1, 2 \qquad \text{da cui}$$

$$\hat{\theta}_1 = \frac{k\bar{Z} - m_2}{m_1 - m_2} = \frac{\sum_{j=1}^{k}\bar{Z}_{.j} - m_2}{m_1 - m_2}$$

che risulta essere corretto e con varianza pari a:

$$\begin{aligned} Var(\hat{\theta}_1) &= \frac{k^2}{(m_1 - m_2)^2} Var(\bar{Z}) = \\ &= \frac{1}{n(m_1 - m_2)^2} Var(Z_1 + Z_2 + ... + Z_k), \end{aligned}$$

vista la postulata indipendenza e identicità nella distribuzione delle $k - ple$.

Se crolla l'assunto di indipendenza, la distribuzione congiunta di (Z_i, Z_j) di due prove, per rispondente, è data da

$$f_{ij}(Z_i, Z_j) = \theta g_j(z_j) + (1-\theta)h_i(z_i)h_j(z_j)$$

da cui

$$Var(Z_i) = \theta(1-\theta)[\mu_{1i} - \mu_{2i}]^2 + \theta\sigma_{1i}^2 + (1-\theta)\sigma_{2i}^2$$

e

$$Cov(Z_i, Z_j) = \theta(1-\theta)(\mu_{1i} - \mu_{2i})(\mu_{1j} - \mu_{2j})$$

e

$$\begin{aligned} Corr(Z_i, Z_j) &= \frac{\theta(1-\theta)(\mu_{1i}-\mu_{2i})(\mu_{1j}-\mu_{2j})}{\sqrt{\theta(1-\theta)[\mu_{1i}-\mu_{2i}]^2+\theta\sigma_{1i}^2+(1-\theta)\sigma_{2i}^2}} * \\ * &\quad \frac{1}{\sqrt{\theta(1-\theta)[\mu_{1j}-\mu_{2j}]^2+\theta\sigma_{1j}^2+(1-\theta)\sigma_{2j}^2}} \end{aligned}$$

In generale

$$Corr(Z_i, Z_j) \to 1$$

se tutti gli scarti quadratici medi tendono a zero e se le differenze tra le medie per i e per j tendono all'infinito, ossia:

$$(\sigma_{1i}, \sigma_{1j}, \sigma_{2i}, \sigma_{2j}) \to 0$$

$$[(\mu_{1i}-\mu_{2i}), (\mu_{1j}-\mu_{2j})] \to \infty.$$

Così

$$\begin{aligned} Var(\hat{\theta}_1) &= \frac{1}{n(m_1-m_2^2)}\left[\sum_{j=1}^{k} Var(Z_j) + 2\sum_{i<j} Cov(Z_i, Z_j)\right] \\ &= \frac{\theta(1-\theta)}{n} + \frac{\theta\sum_{j=1}^{k}\sigma_{1j}^2 + (1-\theta)\sum_{j=1}^{k}\sigma_{2j}^2}{n(m_1-m_2)^2}. \end{aligned}$$

Lavorando sulle colonne si ha che:

$$\hat{\theta}_2 = w_1\hat{\theta}_{.1} + ... + w_k\hat{\theta}_{.k}$$

ove

$$w_j = \frac{|\mu_{1j}-\mu_{2j}|}{\sum_{j=1}^{k}|\mu_{1j}-\mu_{2j}|} = \frac{D_j}{D} \quad \text{per} \quad j = 1, 2, ..., k$$

e

$$\hat{\theta}_{.j} = \frac{\bar{Z}_{.j}-\mu_{2j}}{\mu_{1j}-\mu_{2j}}$$

e se le osservazioni sono di tipo i.i.d.:

$$\begin{aligned} Var(\hat{\theta}_{.j}) &= \frac{1}{n(\mu_{1j}-\mu_{2j})^2} Var(Z_j) = \\ &= \frac{\theta(1-\theta)}{n} + \frac{1}{n(\mu_{1j}-\mu_{2j})^2}\left[\theta\sigma_{1j}^2 + (1-\theta)\sigma_{2j}^2\right]. \end{aligned}$$

La varianza dello stimatore risulta essere pari a:

$$\begin{aligned} Var(\hat{\theta}_2) &= \sum_{j=1}^{k} w_j^2 \, Var(\hat{\theta}_{.j}) + 2 \sum_{i<j} w_i \, w_j \, Cov(\hat{\theta}_{.j}, \hat{\theta}_{i.}) = \\ &= \frac{\theta(1-\theta)}{n} + \frac{\theta \sum_{j=1}^{k} \sigma_{1j}^2 + (1-\theta) \sum_{j=1}^{k} \sigma_{2j}^2}{nD^2}. \end{aligned}$$

Le varianze dei due stimatori sono uguali se e solo se tutti i termini $(\mu_{1j} - \mu_{2j})$ hanno lo stesso segno per $j = 1, 2, ..., k$. In questo caso si ha:

$$\hat{\theta}_1 = \frac{\sum_{j=1}^{k} D_j \hat{\theta}_{.j}}{D} = \hat{\theta}_2.$$

Negli altri casi la varianza del secondo stimatore è sempre minore di quella del primo. Inoltre se $(m_1 - m_2)$ fosse pari a zero, o tendesse a zero la varianza del primo stimatore risulterebbe assai grande; d'altra parte però la varianza del secondo stimatore diminuisce all'aumentare di $|\mu_{1j} - \mu_{2j}|$. La scelta ottimale dei due stimatori quindi dipende da queste ultime considerazioni matematiche e dalla percezione psicologica dei rispondenti nei confronti della procedura.

2.31 Il modello di Kuk (del 1990).

Attributi

Il modello qui proposto consiste nel generare due output di tipo binario, ignoti all'intervistatore ma in accordo con due distribuzioni di bernoulli di parametri noti, θ_1 e θ_2 rispettivamente. Il rispondente riporta la prima uscita se egli appartiene al gruppo delicato A, altrimenti riporta la seconda uscita. Un modo molto semplice per generare questi tipi di output è quello di far uso di due mazzi di carte, nel primo dei quali la proporzione di carte rosse è pari a θ_1 mentre nel secondo dei quali è pari a θ_2. Si assume che, in generale, $\theta_1 \neq \theta_2$. Dopo aver mescolato i due mazzi di carte, l'intervistato deve selezionare casualmente una carta da entrambi i mazzi.

Se r è la proporzione di carte rosse ottenute da un campione di n rispondenti allora la stima (di massima verosimiglianza o dei momenti) per π_1 è data da:

$$\hat{\pi}_1 = \frac{r - \theta_2}{\theta_1 - \theta_2}$$

che risulta essere corretta e con varianza pari a:

$$Var(\hat{\pi}_1) = \frac{\phi(1-\phi)}{n(\theta_1 - \theta_2)^2}$$

ove

$$\phi = \pi\theta_1 + (1-\pi)\theta_2.$$

Ponendo $\theta_1 = p$ e $\theta_2 = 1-p$, ove p vuole ricordare la probabilità del carattere delicato del modello di Warner, si ha la coincidenza della varianza degli stimatori dei due modelli, ossia $V_W = V_1$; inoltre se $\theta_1 = p + (1-p)\pi'$ e $\theta_2 = \pi'(1-p)$ si

ha una coincidenza tra la varianza dello stimatore del modello di Kuk e quella dello stimatore del modello di Simmons, ossia $V_1 = V_S$.

L'esperimento proposto dall'autore può essere reiterato, chiedendo all'intervistato di estrarre k carte con reiserimento dai due mazzi. Sia ora r_k la proporzione di carte rosse ottenute da un campione di n elementi; uno stimatore corretto di π è dato da:

$$\hat{\pi}_k = \frac{r_k - \theta_2}{\theta_1 - \theta_2}$$

la cui varianza è:

$$Var(\hat{\pi}_k) = \frac{\phi(1-phi)}{kn(\theta_1 - \theta_2)^2} + \frac{\pi(1-\pi)}{n}\left(1 - \frac{1}{k}\right) =$$

$$= \frac{1}{k}V_1 + \frac{\pi(1-\pi)}{n}\left(1 - \frac{1}{k}\right)$$

Ora se

$$\pi(1-\pi)(\theta_1 - \theta_2)^2 < \phi(1-\phi)$$

allora $V_k < V_{k'}$ ove $k' < k$; in altre parole incrementando il numero di replicazioni si ottiene uno stimatore più preciso senza intervenire sul costo di campionamento. Per quanto riguarda invece la protezione del rispondente bisogna fare alcune considerazioni su θ_1 e θ_2: questi ultimi devono essere il più simili possibile, ottimizzando contemporaneamente valori di k e di n. In generale comunque è buona prassi non far ripetere troppe volte l'esperimento anche perchè potrebbe capitare al rispondente di dover rispondere almeno una volta al quesito delicato, insospettandolo e inducendolo così a non rispondere o a rispondere in modo evasivo.

2.32 Il metodo con due casualizzatori per intervistato (di Mangat e Singh del 1990).

Attributi

Si estrae un campione con reiserimento di n unità ad ognuna delle quali vengono dati due casualizzatori. Il primo, detto R_1 consente di selezionare le seguenti due affermazioni:

1. appartiene alla classe delicata A;
2. vai al secondo casualizzatore R_2;

con probabilità rispettivamente di T e $1 - T$. Il secondo casualizzatore, R_2, è identico a quello di Warner ed è, come noto, composto da due affermazioni (appartenenza o non apparteneza al gruppo delicato), estratte casualmente e con una associata probabilità di p e $1 - p$ rispettivamente.

L'intervistato viene istruito sull'uso di R_1 e se necessario anche sull'uso di R_2 e la risposta che esso deve dare può essere solamente un "SI" o un "NO" in accordo con il risultato dell'esperimento e allo status posseduto circa l'affermazione estratta.

Ciò premesso la probabilità di ottenere un "SI" sarà pari a:

$$\theta_1 = T\pi + (1-T)p\pi + (1-p)(1-\pi)$$

e uno stimatore per π, ignota proporzione nella popolazione oggetto di studio del carattere delicato, sarà:

$$\hat{\pi}_1 = \frac{\frac{n'}{n} - (1-T)(1-p)}{2p - 1 + 2T(1-p)}$$

ove n' è il numero di "SI" ottenuti dal campione investigato. Si noti che $\hat{\pi}_1$ può assumere valori negativi ma con una probabilità molto bassa. Ora se

$$\frac{n'}{n} \sim Bi(n, \theta_1)$$

allora $\hat{\pi}_1$ è uno stimatore corretto la cui varianza è:

$$Var(\hat{\pi}_1) = \frac{\pi(1-\pi)}{n} + \frac{(1-T)(1-p)1 - (1-T)(1-p)}{n2p - 1 + 2T(1-p)^2}$$

il cui stimatore è dato da:

$$\hat{Var}(\hat{\pi}_1) = \frac{\frac{n'}{n}\left(1 - \frac{n'}{n}\right)}{(n-1)2p - 1 + 2T(1-p)^2}.$$

In generale si ha che

$$Var(\hat{\pi}_1) < Var(\hat{\pi}_W) \quad \text{se} \quad T > \frac{1-2p}{1-p};$$

ossia per un valore accettabile di p, è sempre possibile scegliere T in modo tale che $V_1 > V_W$.

2.33 L'uso del casualizzatore a discrezione dell'intervistato (di Mangat del 1991).

Attributi

La procedura proposta prende le mosse dal modello di Simmons, introducendo però la variante che l'intervistato, all'insaputa dell'intervistatore, può decidere se far uso, oppure no, del casualizzatore. La risposta che l'intervistatore si attende, a prescindere dall'uso o meno del casualizzatore è un "SI" o un "NO".

Supponendo la veridicità delle risposte e che la probabilità di utilizzo del casualizzatore sia pari a T, la probabilità di ottenere una risposta affermativa sarà:

$$\theta_1 = T\pi + (1-T)[p\pi + (1-p)\pi_Y]$$

da cui lo stimatore di di π sarà:

$$\hat{\pi}_m = \frac{n'}{n}$$

ove

$$\frac{n'}{n} \sim Bi(n, \theta_1)$$

con il medesimo significato per n', ossia il numero di risposte affermative ottenute dal campione. Lo stimatore risulta distorto, con distorsione pari a:

$$E(\pi_m - \hat{\pi}_m) = (1 - T)(1 - p)(\pi_Y - \pi).$$

Inoltre,

$$MSE(\hat{\pi}_m) = \frac{\theta_1(1 - \theta_1)}{n} + \left[(1 - T)(1 - p)(\pi_Y - \pi)\right]^2.$$

Si noti che in questo modello T è ignoto e per quanto riguarda la scelta di p e π_Y, questa deve mirare alla riduzione della distorsione dello stimatore. Il suggerimento è di scegliere p più vicino possibile all'unità, senza peraltro far insospettire i rispondenti. Il valore di π_Y può essere scelto il più possibile vicino a π, sfruttando eventuali informazioni sul fenomeno oggetto di studio.

2.34 Una variante al modello di Simmons (di Singh e Singh del 1992).

Attributi

Il modello proposto dagli autori differisce da quello di Simmons in quanto al campione di n unità viene richiesto di restituire quale risposta il "SI", nel caso in cui essi posseggano l'attributo delicato "A" oggetto di studio, e di utilizzare un casualizzatore S e di restituire un "SI" o un "NO" in accordo con quanto risultato dall'esperimento aleatorio e con lo status posseduto nei confronto dello stesso risultato dell'esperimento aleatorio, nel caso in cui essi non posseggano A. Il casualizzatore è lo stesso proposto da Simmons. Il rispondente non deve dire come è arrivato a fornire la risposta data. La probabilità di ottenere una risposta affermativa è:

$$\alpha = \pi + (1 - \pi)(1 - p)\pi_Y$$

Da notare che la risposta affermativa prescinde dal gruppo di appartenenza e ciò garantisce maggiormente la privacy del rispondente e migliora, come impostazione, il modello di Simmons. D'altro canto però la risposta negativa può provenire soltanto dal gruppo di coloro che non posseggono l'attributo delicato A, potendo essi essere identificati ma, probabilmente, senza problema di specie. Lo stimatore della proporzione di diffusione del carattere A sarà quindi dato da:

$$\hat{\pi}_1 = \frac{\hat{\alpha} - (1 - p)\pi_Y}{1 - \pi_Y(1 - p)}$$

ove $\hat{\alpha}$ è la proporzione osservata delle risposte affermative. Inoltre se

$$\hat{\alpha} \sim Bi(n, \alpha)$$

allora lo stimatore risulta essere corretto e con varianza pari a:

$$Var(\hat{\pi}_1) = \frac{\pi(1 - \pi)}{n} + \frac{\pi_Y(1 - \pi)(1 - p)}{n[1 - \pi_Y(1 - p)]}$$

la cui stima è:

$$\hat{Var}(\hat{\pi}_1) = \frac{\hat{\alpha}(1-\hat{\alpha})}{(n-1)[1-\pi_Y(1-p)]^2}.$$

Il confronto con il modello di Simmons verrà fatto secondo due classificazioni ossia $\pi \leq 0.5$ e $\pi > 0.5$

Caso 1. $\pi \leq 0.5$**.** Il nuovo modello è più efficiente se

$$Var(\hat{\pi}_1) - Var(\hat{\pi}) < 0$$

ossia se

$$\pi > \frac{\pi_Y(p^2 - k^2)}{p(1-\pi_Y)(2k-1)} \tag{16}$$

ove

$$k = 1 - \pi_Y(1-p).$$

E per i valori consigliati in entrambi i modelli per p e π_Y, $(2k-1)$ è sempre positivo e $p^2 < k^2$ e quindi la (15) è praticamente sempre verificata.

Caso 2. $\pi > 0.5$**.** Con k definito come sopra e per π_Y e p prossimi all'unità si ha:

$$\frac{\pi_Y k k_1 - p^2(1-\pi_Y)}{[-p^2(1-\pi_Y) + pk_1(1-2\pi_Y)]} > 1$$

ove

$$k_1 = \pi_Y + \pi_Y(1-p).$$

In questo caso il valore ottimo di π_Y per lo schema di Simmons è prossimo a 1 mentre per questo modello deve essere prossimo a 0. Per poter fare un confronto prendiamo valori complementari di π_Y nelle due procedure e sostituiamo π_Y con $(1-\pi_Y)$ nella espressione della varianza. In questo caso si ha che $\hat{\pi}_1$ è più efficiente di $\hat{\pi}$ se

$$Var(\hat{\pi}_1) - Var(\hat{\pi}) < 0$$

ossia se

$$\pi - [p^2(1-\pi_Y) + (1-2\pi_Y)pk_1] < \pi_Y k k_1 - p^2(1-\pi_Y)$$

ove

$$k_1 = \pi_Y + \pi_Y(1-p).$$

Per π_Y e p entrambi convergenti a 1, il coefficiente π è positivo, ossia:

$$\pi \leq \frac{\pi_Y k k_1 - p^2(1-\pi_Y)}{[-p^2(1-\pi_Y) + pk_1(1-2\pi_Y)]}.$$

In definitiva la strategia proposta è spesso più efficiente del modello di Simmons.

2.35 Una variante al modello di Warner di (Singh, Singh e Singh del 1993).

Attributi

Facendo riferimento al modello di Warner del 1965, la proporzione di carte abbinate all'affermazione di non appartenenza alla categoria delicata viene suddivisa in ulteriori due categorie. La prima, pari a p_1, resta uguale a quella proposta da Warner, mentre la seconda, pari a p_2, suggerisce di pescare un'altra carta. Evidentemente con riferimento al modello di Warner si ha che:

$$p_1 + p_2 = 1 - p.$$

Il meccanismo è gestito in modo tale che se si estrae una carta di tipo 1 o 2 si risponde in accordo con la tecnica di Warner mentre, se si estrae una carta di tipo 3, si sceglie un'altra carta, lasciando fuori quella appena selezionata e si risponde nuovamente come appena illustrato. Se nella seconda estrazione venisse nuovamente estratta una carta di tipo 3, si fornisce comunque una risposta negativa (NO). Assumendo la sincerità dei rispondenti, la probabilità di ottenere un "SI" è data da:

$$\theta_1 = 1 + \frac{mp_2}{m-1}[p\pi + p_1(1-\pi)]$$

ove m è il numero totale di carte del casualizzatore.
Con queste premesse uno stimatore di π sarà:

$$\hat{\pi}_1 = \frac{\hat{\theta}_1 - p_1\left(1 + \frac{mp_2}{m-1}\right)}{\left(1 + \frac{mp_2}{m-1}\right)(p - p_1)}$$

ove $\hat{\theta}_1$ è la proporzione osservata dei "SI".
Se

$$\hat{\theta}_1 \sim Bi(n, \theta_1)$$

allora lo stimatore è corretto e presenta una varianza pari a

$$Var(\hat{\pi}_1) = \frac{\pi(1-\pi)}{n} + \frac{p_1\left(1 + \frac{mp_2}{m-1}\right)\left[1 - p_1\left(1 + \frac{mp_2}{m-1}\right)\right]}{n\left[(p - p_1)\left(1 + \frac{mp_2}{m-1}\right)\right]^2}$$

e una stima della quale è fornita da:

$$\hat{Var}(\hat{\pi}_1) = \frac{\hat{\theta}_1(1-\hat{\theta}_1)}{(n-1)\left[(p-p_1)(1+\frac{mp_2}{m-1})\right]^2}$$

Per confrontare questa strategia con il metodo di Warner consideriamo p prossimo all'unità, valore quest'ultimo ottimo per entrambi i modelli. Si definisce efficienza relativa il rapporto che segue:

$$RE = \frac{Var(\hat{\pi})}{Var(\hat{\pi}_1)}$$

e se $RE > 1$ il metodo proposto risulta più efficiente di quello di Warner. Questo si ottiene se:

$$p_2 > \frac{(m-1)[(2p-1)^2 p_1(1-p_1) - p(1-p)(p-p_1)^2]}{m[p(1-p)(p-p_1)^2 + p_1^2(2p-1)^2]}$$

che è soddisfatta per tutti i valori utilizzabili di p, p_1 e p_2. E' da notare che se m è sufficentemente grande esso non influenza RE che aumenta considerevolmente se la scelta di p e p_2 è tale per cui $p_1 = 1 - p - p_2$. Infine se m è grande, l'estrazione con reinserimento è perfettamente uguale a quella senza reinserimento.

2.36 Una tecnica RR ove la numerosità campionaria non è fissata a priori (di Singh, Singh del 1993).

Attributi

In questa procedura la dimensione campionaria non è fissata a priori e il campionamento continua finchè un predeterminato numero m di soggetti non abbia risposto "SI". Lo stimatore di π in questo caso è:

$$\pi_1 = \frac{\hat{\theta} - 1 + p}{2p - 1}$$

con $p \neq 0.5$ e

$$\hat{\theta} = \frac{m-1}{n-1}.$$

In questo caso la frequenza relativa dei "SI" non tende mai a zero e $\hat{\pi}_1$ resta più frequentemente nel range $[0, 1]$; lo stimatore inoltre non dipende solo da p, è corretto e con varianza pari a:

$$Var(\hat{\pi}_1) = \frac{1}{(2p-1)^2}\left[\alpha(m-1)\left(\sum_{r=2}^{m-1}\frac{(-\frac{\theta}{\alpha})^r}{m-r} - \left(-\frac{\theta}{\alpha}\right)^m \log_e \theta\right) - \theta^2\right]$$

ove θ è la probabilità di ottenere il "SI" e

$$\alpha = 1 - \theta \quad \text{con} \quad \theta = p\pi + (1-p)(1-\pi).$$

Si noti che deve essere $m \geq 3$ per l'esistenza della varianza. Inoltre all'aumentare di m, il calcolo della varianza dello stimatore diviene sempre più difficoltosa. Uno stimatore corretto di detta varianza è:

$$\hat{Var}(\hat{\pi}_1) = \frac{\hat{\theta}(1-\hat{\theta})}{(n-2)(2p-1)^2}$$

In generale la scelta di p deve essere tale per cui

$$\left|p - \frac{1}{2}\right| = max$$

e p deve cadere nella parte opposta di π se il punto 0.5 è considerato l'asse di simmetria dell'intervallo $[0, 1]$. Resta valido comunque che p non si deve avvicinare troppo nè a 0 nè a 1 per non creare sospetti nei rispondenti, favorendo

così la veridicità nelle risposte: valori buoni sono 0.2 ± 0.1 e 0.8 ± 0.1. In quanto a m, la varianza di stima è sua funzione decrescente, pertanto m deve essere abbastanza grande. Inoltre il valore di m richiesto per una data precisione dipende anche dalla scelta di p: se p è vicino a zero o all'unità, è adeguato ad assicurare la cooperazione dei rispondenti e una più piccola dimensione di m sarà necessaria nel caso in cui p si avvicini al valore 0.5.

2.37 Due modelli a contaminazione della risposta (di Singh del 1993).

Attributi

Il primo dei due modelli, denominato RRT1, è costruito in modo tale che ogni rispondente del campione estratto con reinserimento e di dimensione n, sia sottoposto all'esperimento casuale di seguito descritto. L'esito del casualizzatore può essere di due tipi: una affermazione che asserisce circa l'appartenenza al gruppo delicato A, con probabilità nota e pari a p, e un'altra affermazione che impone di rispondere comunque con un "SI", con probabilità $1 - p$. Il rispondente fornisce una risposta che l'intervistatore però non può interpretare. Ciò premesso la probabilità di ottenere una risposta affermativa risulta essere pari a:

$$\theta_1 = p\pi + (1 - p)$$

e lo stimatore di massima verosimiglianza di π è:

$$\hat{\pi}_1 = \frac{\hat{\theta}}{p} - \frac{1-p}{p}$$

e se

$$\hat{\theta} \sim Bi(n, \theta_1)$$

lo stimatore risulta essere corretto e con varianza pari a:

$$Var(\hat{\pi}_1) = \frac{\pi(1-\pi)}{n} + \frac{(1-p)(1-\pi)}{np}$$

e un suo stimatore corretto è:

$$\hat{Var}(\hat{\pi}_1) = \frac{\theta(1-\theta)}{(n-1)p^2}.$$

Il secondo dei due modelli presenta lo stesso impianto e si differisce dal primo per le probabilità associate ai due eventi elementari dell'esperimento casuale: esse infatti vengono scambiate. Così la probabilità di ottenere una risposta affermativa risulta essere pari a:

$$\theta_2 = (1 - p)\pi + p.$$

Lo stimatore di π_2 sarà del tipo:

$$\hat{\pi}_2 = \frac{\hat{\theta}}{1-p} - \frac{p}{1-p}$$

la cui varianza risulta essere pari a:

$$Var(\hat{\pi}_2) = \frac{\pi(1-\pi)}{n} + \frac{p(1-\pi)}{n(1-p)}$$

una cui stima corretta è data da:

$$\hat{Var}(\hat{\pi}_2) = \frac{\theta(1-\theta)}{(n-1)(1-p)^2}.$$

In termini di efficienza dello stimatore si ha che $\hat{\pi}_1$ è più efficiente di $\hat{\pi}$ se $(2p-1)^2 < p^2$ o se $\pm(2p-1) < p$. Se invece $\hat{\pi}_2$ è migliore di $\hat{\pi}$ se $p < \frac{2}{3}$ ed è migliore anche di $\hat{\pi}_1$ se $p < \frac{1}{2}$ e viceversa. In generale, in alternativa al metodo di Warner, si preferisce il modello deniminato RRT1 se $p < \frac{1}{2}$ altrimenti si preferisce il modello RRT2.

2.38 Il modello di Franklin in una sua generalizzazione (di Singh e Singh del 1993).

Attributi

In accordo con la tecnica di Franklin introdotto nel 1989 viene estratto un campione casuale semplice con reinserimento di n unità da una data popolazione. Ogni intervistato viene dotato di tre casualizzatori detti R_0, R_1 e R_2, ognuno dei quali fornisce due output mutuamente esclusivi. Tutti gli intervistati utilizzano il primo casualizzatore costituito dalle seguenti due affermazioni:

1. usa il casualizzatore R_1
2. usa il casualizzatore R_2

La prima affermazione viene estratta con probabilità T e la seconda con probabilità $1-T$. Se il rispondente estrae la carta riportante la prima affermazione allora viene istruito sull'uso del casualizzatore R_1: utilizzare la densità $g^{\star}_{ij}(L_{ij})$ se egli appartiene alla categoria delicata A, utilizzare la densità $h^{\star}_{ij}(L_{ij})$ altrimenti. Se l'intervistato invece seleziona la carta riportante la seconda affermazione allora viene istruito sull'uso del casualizzatore R_2: egli utilizzerà la densità $g_{ij}(L_{ij})$ se appartiene ad A, altrimenti utilizzerà la densità $h_{ij}(L_{ij})$. Inoltre vengono condotte $k \geq 1$ di prove indipendenti per ogni rispondente. Ovviamente questa seconda fase dell'uso di R_1 o R_2 viene fatta tenendo all'oscuro l'intervistatore che conosce solamente la forma esatta delle densità:

$$g_{ij} \qquad g^{\star}_{ij} \qquad h_{ij} \qquad h^{\star}_{ij}.$$

L'intervistato infatti svela solo il numero casuale ottenuto e non la distribuzione da cui è stato ottenuto. Si denoti ora con l_{ij} il valore fornito dall'intervistato. Da un totale di kxn osservazioni del tipo l_{ij} ($i = 1, 2, ..., n$ e $j = 1, 2, ..., k$) si possono ottenere alcune inferenze su π.

Le densità in parola sono così caratterizzate:

$$g^{\star}_j \sim N(u^{\star}_{1j}, \sigma^{\star 2}_{1j}) \qquad g_j \sim N(u_{1j}, \sigma^2_{1j})$$

$$h^{\star}_j \sim N(u^{\star}_{2j}, \sigma^{\star 2}_{2j}) \qquad h_j \sim N(u_{2j}, \sigma^2_{2j})$$

e per semplicità

$$u^{\star}_{1j} = u_{1j} \qquad u^{\star}_{2j} = u_{2j} \qquad j = 1, 2, ..., k.$$

A seguire si indicheranno due stimatori per π basati rispettivamente sulle medie delle righe e delle colonne di L_{ij}:

$$\hat{\pi}_{S1} = \frac{\sum_{j=1}^{k} \tau_{\cdot j} - m_2}{m_1 - m_2} \qquad \text{ove} \qquad m_i = \sum_{j=1}^{k} u_{ij} \quad i = 1, 2$$

$$\hat{\pi}_{S2} = \sum_{j=1}^{k} w_j \hat{\pi}^{\star}_{sj} \qquad \text{ove} \qquad \hat{\pi}^{\star}_{sj} = \frac{\bar{L}_{ij} - u_{2j}}{u_{1j} - u_{2j}} \qquad \text{e} \qquad \sum_{j=1}^{k} w_j = 1$$

le cui varianze sono rispettivamente:

$$\begin{aligned} Var(\hat{\pi}_{S1}) &= \frac{\pi(1-\pi)}{n} + \frac{\pi \sum_{j=1}^{k} \sigma^2_{1j} + (1-\pi) \sum_{j=1}^{k} \sigma^2_{2j}}{n(m_1 - m_2)^2} + \\ &+ \frac{T\left[\pi \sum_{j=1}^{k} (\sigma^{\star 2}_{1j} - \sigma^2_{1j}) + (1-\pi) \sum_{j=1}^{k} (\sigma^{\star 2}_{2j} - \sigma^2_{2j})\right]}{n(m_1 - m_2)^2} \end{aligned}$$

$$Var(\hat{\pi}_{S2}) = \sum_{j} w_j^2 \, Var(\hat{\pi}^{\star}_{sj}) + 2 \sum_{j<j'} w_j \, w_{j'} \, Cov(\hat{\pi}^{\star}_{sj}, \hat{\pi}^{\star}_{sj'})$$

ove

- $j \neq j'$ per $j = 1, 2, ..., k$
- $Cov(\hat{\pi}^{\star}_{sj}, \hat{\pi}^{\star}_{sj'}) = \frac{\pi(1-\pi)}{n}$
- $$\begin{aligned} Var(\hat{\pi}^{\star}_{sj}) &= \frac{\pi(1-\pi)}{n} + \frac{\pi\sigma^2_{1j} + (1-\pi)\sigma^2_{2j}}{n(u_{1j} - u_{2j})^2} + \\ &+ \frac{T[\pi(\sigma^{*2}_{1j} - \sigma^2_{1j}) + (1-\pi)(\sigma^{*2}_{2j} - \sigma^2_{2j})]}{n(u_{1j} - u_{2j})^2} \end{aligned}$$

I pesi w_j vengono scelti in accordo al metodo di Franklin, ossia:

$$w_j = \frac{D_j}{D} \quad \text{ove} \quad D_j = |u_{1j} - u_{2j}| \quad \text{e} \quad D = \sum_{j=1}^{k} D_j$$

Questa particolare scelta ci permetterà il confronto, in termini di efficienza, con il metodo di Franklin. Non si esclude, in generale, una scelta diversa dei pesi w_j. Fissati invece i pesi come detto si ha che:

$$
\begin{aligned}
Var(\hat{\pi}_{S2}) &= \frac{\pi(1-\pi)}{n} + \\
&+ \frac{\pi \sum \sigma_{1j}^2 + (1-\pi)\sum \sigma_{2j}^2 + T[[\pi(\sigma_{1j}^{*2} - \sigma_{1j}^2) + (1-\pi)(\sigma_{2j}^{*2} - \sigma_{2j}^2)]}{nD^2}
\end{aligned}
$$

A seguire alcune considerazioni:

1. i due stimatori sono uguali quando tutte le differenze tra u_{1j} e u_{2j} hanno lo stesso segno ($\forall j$); negli altri casi lo stimatore $\hat{\pi}_{S2}$ presenta una varianza minore dello stimatore $\hat{\pi}_{S2}$. Inoltre se la differenza tra m_1 e m_2 fosse pari a 0 o tendente a 0, la varianza di $\hat{\pi}_{S1}$ può divenire molto grande. D'altra parte, la varianza dello stimatore $\hat{\pi}_{S2}$ può diminuire prendendo una differenza assoluta di D_j grande;

2. lo stimatore proposto $\hat{\pi}_{S1}$ è stato trovato affinchè esso sia più efficiente di quello di Franklin se le densità $g_j^\star, h_j^\star, g_j, h_j$ sono tali per cui $\sigma_{1j}^\star < \sigma_{1j}$ e $\sigma_{2j}^\star < \sigma_{2j}$ $\forall j = 1, 2, ..., k$. Similmente $\hat{\pi}_{S2}$ è più efficiente di quello di Franklin se $\sigma_{1j}^\star < \sigma_{1j}$ e $\sigma_{2j}^\star < \sigma_{2j}$ $\forall j = 1, 2, ..., k$. Fintanto che si assumono noti detti parametri, la scelta non risulta di certo difficoltosa. Inoltre si segnala in questo metodo fa uso di quattro distribuzioni, anzichè due, con una conseguente aumentata protezione del rispondente derivante da una maggiore difficoltà nel capire da quale distribuzione possa derivare la risposta fornita.

3. nel metodo qui proposto se R_1 possedesse una unica affermazione del tipo "usa la densità $g_j^\star(L)$", e ponendo $h_j(L) = 1 - g_j(L)$ in R_2, ci si rifarebbe al caso suggerito da Mangat e Singh del 1990;

4. se $\sigma_{1j}^\star = \sigma_{1j}$ e $\sigma_{2j}^\star = \sigma_{2j}$ $\forall j$, il modello si riduce a quello di Franklin. Lo stesso risultato si ottiene con $T = 0$.

2.39 Una strategia che permette di ammettere l'appartenenza al gruppo delicato (di Mangat del 1994).

Attributi

Si assume di estrarre da una popolazione un campione casuale semplice con reinserimento di n unità, ognuna delle quali è istruita a fornire "SI" come risposta, nel caso in cui possegga l'attributo delicato A e a utilizzare un casualizzatore nel caso contrario. Questo è composto di due affermazioni, possedere e non possedere l'attributo delicato A, alle quali è associata una nota probabilità pari a p e $1-p$. In questo secondo caso il rispondente deve rispondere in accordo con l'esito dell'esperimento e con lo status posseduto. Tutto questo come al solito avviene senza che l'intervistatore sappia il risultato dell'esperimento. Con queste premesse, la probabilità di ottenere una risposta effermativa sarà:

$$\alpha = \pi + (1-\pi)(1-p)$$

e la stima di massima verosimiglianza di π sarà:

$$\hat{\pi}_m = \frac{\hat{\alpha} - 1 + p}{p}$$

ove $\hat{\alpha}$ è la proporzione osservata dei "SI" nelle n interviste. Inoltre se

$$\hat{\alpha} \sim Bi(n, \alpha)$$

allora $\hat{\pi}_m$ è corretto e con varianza pari a:

$$Var(\hat{\pi}_m) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)(1-p)}{np}$$

e una stima corretta è

$$\hat{Var}(\hat{\pi}_m) = \frac{\hat{\alpha}(1-\hat{\alpha})}{(n-1)p^2}$$

Il presente stimatore è più efficiente di quello proposto da *Mangat* e *Singh* nel 1990 se:

$$\pi > 1 - \frac{p(1-T)[1-(1-T)(1-p)]}{[2p-1+2T(1-p)]^2} \quad \text{per} \quad p \neq \frac{1}{2}$$

ove T è la probabilità di estrarre l'attributo delicato nel primo casualizzatore dei due autori e p è la stessa probabilità del presente modello. Inoltre il presente modello è più efficiente di quello di Warner se:

$$\pi > 1 - \left(\frac{p}{2p-1}\right)^2$$

che è vera per $p > \frac{1}{3}$.

2.40 Uno stimatore RR nel caso di distribuzione continua da popolazione dicotomica (di Chua Chiang del 1995).

Attributi

In accordo con lo schema di Franklin del 1989 si ponga:

$$g_{ij} = g_j \qquad \text{e} \qquad h_{ij} = h_j$$

e si definisca

$$w_{ij} = \begin{cases} 1 & \text{se} \quad z_{ij} \leq a_j \\ 0 & \text{se} \quad z_{ij} > a_j \end{cases}$$

per $i = 1, 2, ..., n,\ j = 1, 2, ..., k$ e a_j costante e scelta in modo tale che lo stimatore di Franklin abbia varianza minima.

Sia poi:

$$p_j = \int_{-\infty}^{a_j} g_j(t)dt \quad \text{e} \quad q_j = \int_{a_j}^{\infty} h_j(t)dt.$$

Per il metodo dei momenti si consideri il seguente stimatore:

$$\hat{\theta} = \frac{k\bar{w} - \sum_{j=1}^{k}(1-q_j)}{\sum_{j=1}^{k}(p_j + q_j - 1)} \quad \text{ove} \quad \bar{w} = \frac{1}{nk}\sum_{ij} w_{ij}$$

che è stimatore corretto di θ e con varianza pari a:

$$Var(\hat{\theta}) = \frac{\theta(1-\theta)}{n} + \frac{\theta \sum_{j=1}^{k} p_j(1-p_j) + (1-\theta)\sum_{j=1}^{k} q_j(1-q_j)}{n\left[\sum_{j=1}^{k}(p_j+q_j-1)\right]^2}$$

Per semplicità consideriamo ora il caso di una prova ($k = 1$) e assumiamo che g_1 e h_1 siano di tipo normale, ossia:

$$g_1 \sim N(\mu_1, \sigma^2) \qquad , \qquad g_2 \sim N(\mu_2, \sigma^2).$$

Sia inoltre:

$$b = \frac{\mu_2 - \mu_1}{\sigma} \quad \text{e} \quad a = \frac{\mu_1 + \mu_2}{2} + c\sigma.$$

Quindi:

$$p_1 = \Phi\left(\frac{b}{2} + c\right) \quad \text{e} \quad q_1 = 1 - \Phi\left(\frac{b}{2} + c\right)$$

ove $\Phi(\bullet)$ è la funzione di ripartizione di una $N(0,1)$.
Così la varianza dello stimatore proposto è:

$$\begin{aligned} Var(\hat{\theta}) &= \frac{\theta(1-\theta)}{n} + \\ &+ \frac{\theta\Phi\left(\frac{b}{2}+c\right)\left[1-\Phi\left(\frac{b}{2}+c\right)\right] + (1-\theta)\Phi\left(-\frac{b}{2}+c\right)\left[1-\Phi\left(-\frac{b}{2}+c\right)\right]}{n\left[\Phi\left(\frac{b}{2}+c\right) - \Phi\left(-\frac{b}{2}+c\right)\right]^2} \end{aligned}$$

ove c è un valore scelto in modo tale che la varianza di $\hat{\theta}$ sia minima per un dato b. In questo caso la varianza di Franklin è data da:

$$Var(\hat{\theta}_1) = \frac{\theta(1-\theta)}{n} + \frac{1}{nb^2} = Var_F.$$

Da un confronto con le due varianze si nota che:

1. sia Var_F che Var_A, diminuiscono al crescere di b;
2. al crescere di b, Var_A decresce di più di Var_F;
3. sembra che lo stimatore proposto sia migliore quando la proporzione di individui in A è piccola: $b \geq \frac{1}{4}$, quando $\theta = 0.05$.

   Fermo restando $k = 1$, si assume ora g_1 e h_1 di tipo esponenziale con media rispettivamente μ_1 e μ_2. Sia anche $b = \frac{\mu_1}{\mu_2}$ e $a = \mu_1 + c\mu_2$, quindi

   $$p_1 = 1 - e^{-\left(1+\frac{c}{b}\right)} \quad \text{e} \quad q_1 = e^{-(b+c)}.$$

   La varianza dello stimatore risulta allora

$$Var(\hat{\theta}) = \frac{\theta(1-\theta)}{n} + \frac{\theta\left[1-e^{-\left(1+\frac{c}{b}\right)}\right]e^{-\left(1+\frac{c}{b}\right)} + (1-\theta)e^{-(b+c)}\left[1-e^{-(b+c)}\right]}{n\left[e^{-(b+c)} - e^{-\left(1+\frac{c}{b}\right)}\right]^2}$$

e come sopra c viene scelto in modo tale che la varianza di $\hat{\theta}$ sia la più piccola possibile per un dato valore di b. In questo caso la varianza nel modello di Franklin è data da:

$$Var(\hat{\theta}) = \frac{\theta(1-\theta)}{n} + \frac{1+\theta(b^2-1)}{n(b-1)^2}.$$

Per $b < 1$ sia Var_A che Var_F sono funzioni decrescenti di b. Così in generale all'aumentare di θ, le due distribuzioni random debbono differire maggiormente in termini della loro media per ottenere uno stimatore migliore di quello di Franklin.

2.41 Una procedura RR a due stadi (di Chang e Liang del 1996).

Attributi

Nel metodo proposto dagli autori ogni rispondente deve fornire due risposte binarie in accordo con due distribuzioni indipendenti e di noti parametri τ e q. Ad ogni intervistato vengono consegnati due casualizzatori detti rispettivamente R_1 e R_2. Il primo consta di due affermazioni quali:

1. appartengo al gruppo delicato A;
2. utilizza il secondo casualizzatore;

di probabilità rispettivamente τ e $1-\tau$. Anche il secondo casualizzatore consta di due affermazioni:

1. appartengo al gruppo delicato A;
2. appartengo al gruppo non delicato e incorrelatao con A, U;

di probabilità rispettivamente p e $q = 1-p$.
Una stima di π sarà:

$$\pi_T = \frac{\frac{n'}{n} - (1-\tau)q\pi'}{p+q\tau}$$

ove n' è il numero totale dei "SI" ottenuti dal campione di n rispondenti. Tale stimatore è corretto e con varianza pari a:

$$\begin{aligned} Var(\pi)_T &= \frac{\pi(1-\pi)}{n} + \frac{(1-\tau)^2q^2\pi'(1-\pi') + (1-\tau)q(p+q\tau)(\pi+\pi'-2\pi\pi')}{n(p+q\tau)^2} \\ &= \frac{\phi_T(1-\phi_T)}{n[\tau+(1-\tau)p]^2} \end{aligned}$$

ove

$$\phi_T = [\tau + (1-\tau)p]\pi + (1-\tau)q\pi'$$

è la probabilità di ottenere una risposta affermativa nel presente metodo. Si noti che qui il simbolo π' coincide con π_Y del modello di Simmons. Uno stimatore corretto per la varianza è:

$$\hat{Var}(\hat{\pi}_T) = \frac{\frac{n'}{n}\left(1 - \frac{n'}{n}\right)}{(n-1)(p+q\tau)^2}.$$

Lo stimatore proposto risulta essere più efficiente di quello di Simmons se:

$$q[2(1-\tau)p + \tau]\pi'(1-\pi') + p(p+q\tau)[\pi(1-\pi') + \pi'(1-\pi)] > 0.$$

In quanto alla scelta dei parametri si nota che maggiori sono τ e p, minore è la varianza dello stimatore. In sintesi:

1. La tecnica a due stadi è uniformemente più efficiente del metodo di Simmons.
2. Per vari valori di π e π', l'efficienza relativa aumenta quando τ converge a 1 e p converge a 0.
3. In termini generali si ha la maggior efficienza relativa per tutte le combinazioni di τ e p, scegliendo π' in convergenza verso $\frac{1}{2}$.

2.42 La tecnica RR applicata alle variabili (di Singh e Joarder del 1997).

Variabili

Nella presente proposta ogni rispondente ha a disposizione la seguente scelta:

1. riportare il valore posseduto di x_i
2. riportare il valore camuffato $t_i = \frac{x_{is}}{\theta}$

evidentemente senza rilevare all'intervistatore quale opzione è stata scelta per fornire la risposta. Va detto comunque che sia x_i che t_i debbono avere lo stesso dominio altrimenti il rispondente sarebbe istantaneamente identificato. Sia dunque W la probabilità che un rispondente selezioni la prima opzione e $(1-W)$ la complementare probabilità. Così il vettore delle delle risposte Z_i ha la seguente distribuzione:

$$Z_i = \begin{cases} x_i & \text{con probabilità } W \\ t_i & \text{con probabilità } 1-W \end{cases}$$

Uno stimatore corretto di π è dato da:

$$\bar{y_1} = \frac{1}{n}\sum_{i=1}^{n} z_i$$

e

$$E(z_i) = WE(x_i) + (1-W)E(z_i) = \mu_x$$

Inoltre la distribuzione di Z_i^2 è:

$$Z_i = \begin{cases} x_i^2 & \text{con probabilità } W \\ t_i^2 & \text{con probabilità } 1-W \end{cases}$$

e

$$\begin{aligned} E(Z_i^2) &= WE(x_i^2) + (1-W)E(z_i^2) = \\ &= W(\sigma_X^2 + \mu_X^2) + (1-W)(1 + \frac{\gamma^2}{\theta^2})(\sigma_X^2 + \mu_X^2) \end{aligned}$$

La varianza dello stimatore prposto è data da:

$$Var(\bar{y}_1) = \frac{1}{n}\left[\sigma_X^2 + (\frac{\gamma}{\theta})^2(\sigma_X^2 + \mu_X^2)(1-W)\right]$$

e un suo stimatore corretto è calcolato ponendo

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{y}_1)^2 \qquad \text{e} \qquad E(s_x^2) = Var(\bar{y}_1).$$

La scelta di W è arbitraria ma l'esperienza di altre indagini svolte può essere di aiuto come anche un'indagine pilota che, intervistando un sottogruppo di rispondenti, può fornire una stima del valore in parola. Inoltre nel modello presentato il parametro W è stato fissato uguale per tutti i rispondenti; in generale questo è restrittivo e si potrebbe considerare il caso in cui W è diverso per ogni rispondente: in tal senso dovremmo impostare il modello secondo un approccio gerarchico bayesiano.

2.43 Il modello di Moors e violazione della privacy dei rispondenti. Una rettifica attraverso la strategia del gruppo casuale (di Mangat, Singh e Singh del 1997).

Attributi

Prendendo le mosse dal modello di Moors, questo metodo introduce la variante di suddividere la popolazione di N unità in due gruppi casuali di dimensione rispettivamente N_1 e N_2, utilizzando il campionamento casuale semplice senza reinserimento in modo tale che $N_1 + N_2 = N$. Vengono poi tratti due campioni di dimensione n_1 e n_2 dalle due sottopopolazioni con campionamento casuale semplice con reinserimento. Ora come nel modello di Moor, i rispondenti del primo campione vengono provvisti di un casualizzatore detto S_1 che porta ad dare informazioni circa le caratteristiche A e Y, rispettivamente delicata e non. I rispondenti del secondo campione invecevengono intervistati direttamente circa il carattere non delicato Y. La probabilità di ottenere una risposta affermativa nel primo campione è data da:

$$\theta_1^{'} = p_1\pi_1 + (1-p_1)\pi_{Y1}$$

ove π_1 e π_{Y1} è la proporzione di individui che in N_1 posseggono rispettivamente A e Y. Ovviamente la probabilità di ottenere una risposta affermativa nel secondo campione sarà:

$$\theta_2^{'} = \pi_{Y2}$$

ove π_{Y2} è la proporzione di rispondenti che posseggono la caratteristica Y. SE $\hat{\pi}_{Y2}$ è la stima di π_{Y2} ottenuta dal campione di n_2 unità, lo stimatore di π è dato da:

$$\hat{\pi}_S = \frac{\hat{\theta}_1^{'} - (1-p_1)\hat{\pi}_{Y2}}{p_1}$$

ove $\hat{\theta}_1^{'}$ è la proporzione del "SI" nel primo campione. Tale stimatore è corretto e con varianza pari a

$$\begin{aligned}
Var(\hat{\pi}_S) &= \frac{1}{p_1^2(N-1)}\left[\frac{(N-1)\theta_1(1-\theta_1)}{n_1} + \frac{N_2(n_1-1)}{N_1 n_1}p_1^2\pi(1-\pi)\right] + \\
&+ \frac{1}{p_1^2(N-1)}\left[p_1^2\pi(1-\pi)\right] + \\
&+ \frac{1}{p_1^2(N-1)}\left[\pi_Y(1-\pi_Y)(1-p_1)^2\left(-\frac{N_2}{n_1 N_1} + \frac{N(N_2-1)}{n_2 N_2} + \frac{N^2}{N_1 N_2}\right)\right] + \\
&+ \frac{1}{p_1^2(N-1)}\left[2\left(1 + \frac{N_2(n_1-1)}{N_1 n_1}\right)p_1(1-p_1)\rho_{AY}\sqrt{\pi(1-\pi)\pi_Y(1-\pi_Y)}\right]
\end{aligned}$$

ove evidentemente l'ultimo addendo risulta essere pari a zero nel caso in cui la correlazione fra A e Y fosse nulla. Si noti che la varianza dipende da p_1, π_Y, N_2, n_1 e n_2 e se scelti in modo ottimale si può renderla minima. In ogni caso, posto che i caratteri A e Y siano incorrelati, la questione più rilevante riguarda l'allocazione della popolazione nelle due sottopopolazioni e le rispettive dimensioni campionarie, visto che π_Y e p_1 possono essere scelti come indicato da *Greenberg et all. (1969)*. In generale quindi fissati gli altri parametri si ha:

$$\frac{N_1}{N_2} \approx \sqrt{\frac{p_1^2\pi(1-\pi) + \pi_Y(1-\pi_Y)(1-p_1)^2}{\pi_Y(1-\pi_Y)(1-p_1)^2}}$$

inoltre l'allocazione ottimale per i due campioni risulta essere:

$$\frac{n_1}{n_2} \approx \sqrt{\frac{\theta_1(1-\theta_1)}{\pi_Y(1-\pi_Y)(1-p_1)^2} - \frac{N_1}{N_2 N}}$$

Ancora uno stimatore corretto della varianza di stima è:

$$\begin{aligned}
\hat{Var}(\hat{\pi}_S) &= \frac{n_1 N_1}{N(N_1-1)(n_1-1)p_1^2}\left[\frac{N-1}{n_1}(\hat{\pi}_S p_1(1-\pi_{\hat{S}}p_1)\right] + \\
&+ \frac{n_1 N_1}{N(N_1-1)(n_1-1)p_1^2}\left[\frac{N-1}{n_1}\left[p_1(1-p_1)\hat{\pi}_{Y2} - 2\pi_S\pi_{Y2}p_1(1-p_1)\right]\right] + \\
&+ \frac{n_1 N_1}{N(N_1-1)(n_1-1)p_1^2}\left[\frac{N_2(n_1-1)}{N_1 n_1}p_1^2\hat{\pi}_S(1-\hat{pi}_S)\right] +
\end{aligned}$$

$$
\begin{aligned}
+ \quad & \frac{n_1 N_1}{N(N_1-1)(n_1-1)p_1^2}\frac{N_2 n_2(N-1)}{N(N_2-1)(n_2-1)}\hat{\pi}_{Y2}(1-p_1)^2\left(\frac{N-1}{n_1}\right)+ \\
- \quad & \frac{n_1 N_1}{N(N_1-1)(n_1-1)p_1^2}\frac{N_2 n_2(N-1)}{N(N_2-1)(n_2-1)}\hat{\pi}_{Y2}(1-p_1)^2\left(\frac{N_2}{n_1 N_1}\right)+ \\
+ \quad & \frac{n_1 N_1}{N(N_1-1)(n_1-1)p_1^2}\frac{N_2 n_2(N-1)}{N(N_2-1)(n_2-1)}\hat{\pi}_{Y2}(1-p_1)^2\left[\frac{N(N_2-1)}{n_2 N_2}\right]+ \\
+ \quad & \frac{n_1 N_1}{N(N_1-1)(n_1-1)p_1^2}\frac{N_2 n_2(N-1)}{N(N_2-1)(n_2-1)}\hat{\pi}_{Y2}(1-p_1)^2\left[\frac{N^2}{N_1 N_2}\right]+ \\
- \quad & \frac{n_1 N_1}{N(N_1-1)(n_1-1)p_1^2}\frac{N_2 n_2(N-1)}{N(N_2-1)(n_2-1)}\hat{\pi}_{Y2}(1-p_1)^2\left[\frac{2N(N_2+n_2-1)}{N_2 n_1 n_2}\right].
\end{aligned}
$$

Per trattare il tema dell'efficienza possiamo scrivere la varianza dello stimatore proposto in funzione della varianza dello stimatore proposto da Moors, $Var_m(\hat{\pi}_{hg})$, ossia:

$$
\begin{aligned}
\hat{Var}(\hat{\pi}_S) &= Var_m(\hat{\pi}_{hg} + \frac{N_2}{N_1(N-1)p_1^2}\left[\frac{n_1-1}{n_1}p_1^2\pi(1-\pi)\right]+ \\
&+ \frac{N_2}{N_1(N-1)p_1^2}\left[\pi_Y(1-\pi_Y)(1-p_1^2)\left(\frac{N^2}{N_2^2}-\frac{N_1^2}{N_2^2}-\frac{N_1^2}{N_2^2 n_2}\frac{1}{n_1}\right)\right] \\
Var(\hat{\pi}_{hg}) &= \frac{1}{p_1^2}\left[\frac{\theta_1(1-theta_1)}{n_1}+\frac{(1-p_1^2)\theta_2(1-\theta_2)}{n_2}\right].
\end{aligned}
$$

Ovviamente la varianza $Var(\hat{\pi}_S)$ è leggermente maggiore di Var_m ma garantisce la protezione dei rispondenti. Inoltre il confronto con il modello di Simmons mostra che lo stimatore $\hat{\pi_S}$ è più efficiente di quello di $\hat{\pi}_{hg}$.

2.44 Due modelli alternativi al modello di Moors (di Singh, Singh e Mangat del 1997).

Attributi

2.44.1 Il primo modello

Vengono estratti senza reinserimento due campioni di numerosità rispettivamente n_1 e n_2, indipendenti. Ogni rispondente incluso nel primo campione è dotato del casualizzatore S_1 che fornisce con probabilità rispettivamente di p_1 e $1-p_1$ le seguenti due affermazioni:

1. appartengo ad A
2. Appartengo ad Y.

I rispondenti del secondo campione, che non sono stati estratti nel primo campione, vengono sottoposti ad intervista diretta circa il possedere o meno l'attributo Y. Per i rispondenti del secondo campione, che sono stati inclusi anche nel primo, sono provvisti del casualizzatore S_2 e vengono interrogati circa il loro stato nei confronti di A e Y, con lo stesso metodo adottato da Simmons.

Sia quindi n_{21} il numero di unità comuni ai due campioni [4] e n_{22} il numero di unità non comuni del secondo campione, in modo tale che

[4] Nel caso particolare di $n_{21}=0$ si rientra nel modello di Moors.

$$n_{21} + n_{22} = n_2.$$

Si ha quindi:

$$\hat{\pi}_1 = \frac{\hat{\theta}_1 - (1-p_1)\hat{\pi}_{2Y}}{p_1} \qquad \text{e} \qquad \hat{\pi}_2 = \frac{\hat{\theta}_2 - (1-p_2)\hat{\pi}_{2Y}}{p_2}$$

ove $\hat{\theta}_1$ è la proporzione delle risposte affermative nel primo campione, $\hat{\theta}_2$ è la medesima proporzione riferita al secondo campione e che sono comuni al primo e $\hat{pi}_{2Y}$ è la proporzione di color che posseggono Y su gli n_{22} rispondenti che non sono comuni al primo campione.

Ciò premesso uno stimatore corretto di π è:

$$\hat{\pi}_p = W\hat{\pi}_1 + (1-W)\hat{\pi}_2$$

ove W è una costante reale. La varianza dello stimatore proposto è data da

$$Var(\hat{\pi}_p) = W^2\, Var(\hat{\pi}_1) + (1-W)^2\, Var(\hat{\pi}_2) + 2W(1-W)\, Cov(\hat{\pi}_1, \hat{\pi}_2)$$

ove

$$\begin{aligned}
Var(\hat{\pi}_1) &= \frac{1}{n_1 p_1^2}\left[\theta_1(1-\theta_1) - \frac{(n_1-1)\pi(1-\pi)}{N-1}\right] + \\
&+ \frac{1}{n_1 p_1^2}\frac{\pi_Y(1-\pi_Y)(1-p_1)^2}{N-1}\left[N n_1 E_1\left(\frac{1}{n_{22}}\right) + n_1 - 2N\right] \\
Var(\hat{\pi}_2) &= \frac{1}{p_2^2}\left[E_1\left(\frac{1}{n_{21}}\right)\theta_2(1-\theta_2) + \frac{\pi(1-\pi)}{N-1} - \frac{\pi(1-\pi)}{N-1}\right] + \\
&+ \frac{1}{p_2^2}\frac{(1-p_2)^2\pi_Y(1-\pi_Y)}{n_1(N-1)}\left[N n_1 E_1\left(\frac{1}{n_{12}}\right) + n_1 - 2N\right] \\
Cov(\hat{\pi}_1, \hat{\pi}_2) &= \frac{(N-n_1)p_1 p_2 \pi(1-\pi)}{n_1(N-1)p_1 p_2} + \\
&+ \frac{(1-p_1)(1-p_2)\pi_Y(1-\pi_Y)N\left[n_1 E_1\left(\frac{1}{n_{22}}\right) - 1\right]}{n_1(N-1)p_1 p_2}.
\end{aligned}$$

Il valore di $E_1(\bullet)$ si trova in *Gavindarajulu (1962).*

Si può facilmente vedere che il valore ottimo della costante W per minimizzare la varianza è dato da:

$$W_{opt} = \frac{Var(\hat{\pi}_2) - Cov(\hat{\pi}_1, \hat{\pi}_2)}{Var(\hat{\pi}_1) + Var(\hat{\pi}_2) - 2\, Cov(\hat{\pi}_1, \hat{\pi}_2)}$$

e

$$Var_{min}(\hat{\pi}_p) = \frac{Var(\hat{\pi}_1)\, Var(\hat{\pi}_2) - [Cov(\hat{\pi}_1, \hat{\pi}_2)]^2}{Var(\hat{\pi}_1) + Var(\hat{\pi}_2) - 2\, Cov(\hat{\pi}_1, \hat{\pi}_2)}.$$

2.44.2 Il secondo modello.

In questo secondo modello proposto viene tratto soltanto un campione, senza reiserimento e di dimensione n. Il campione così ottenuto viene a sua volta suddiviso in due sottocampioni di dimensione rispettivamente n_1 e n_2 tali che $n_1+n_2 = n$. Ogni intervistato del primo campione viene sottoposto ad intervista con il metodo di Simmons mentre, ogni intervistato del secondo campione viene sottoposto ad intervista diretta circa il possesso del carattere Y.

Uno stimatore corretto per π è dato da:

$$\hat{\pi}_R = \frac{\hat{\theta}_1 - (1-p_1)\hat{\pi}_{2Y}}{p_1}$$

ove

$\hat{\theta}_1$ e $\hat{\pi}_{2Y}$ è la proporzione di risposte affermative nel primo e nel secondo campione rispettivamente.

Qualora i due caratteri A e Y fossero incorrelati, come qui supposto, la varianza dello stimatore risulta essere pari a:

$$Var(\hat{\pi}_R) = \frac{1}{p_1^2}\left[\frac{\theta_1(1-\theta_1)}{n_1} - \frac{\pi(1-\pi)(n_1-1)}{n_1(N-1)} + \frac{(1-p_1)^2(N+n_2)\pi_Y(1-\pi_Y)}{n_2(N-1)}\right].$$

2.44.3 Alcuni confronti tra i modelli e imputazione dei valori ottimi per le dimensioni campionarie.

Si deve subito precisare che se i due sottocampioni sono tratti indipendentemente, in accordo con il metodo di Moors, possiamo fare uso solo della prima strategia proposta, ove vi sono unità comuni nei due campioni. Nel secondo modello invece, non vi sono le condizioni per essere in accordo con la strategia di Moors. In effetti l'espressione della varianza ottenuta nel secondo modello è esatta mentre nel primo modello vengono coinvolti i termini $E\left(\frac{1}{n_{21}}\right)$ e $E\left(\frac{1}{n_{22}}\right)$ che possono essere quantificati solo tramite approssimazione. A causa di ciò verrà confrontato il secondo metodo con il modedllo di Kim del 1978, il cui stimatore è uguale a quello di Simmons e la cui varianza è pari a:

$$Var_K(\hat{\pi}_G) = \frac{1}{(p_1-p_2)^2}\left[\frac{A}{n_1} + \frac{B}{n_2} - C\right]$$

ove

$$\begin{aligned}
A &= (1-p_2)^2\theta_1(1-\theta_1) + \frac{(1-p_2)^2[p_1^2\pi(1-\pi) + (1-p_1)^2\pi_Y(1-\pi_Y)]}{N-1} \\
B &= (1-p_1)^2\theta_2(1-\theta_2) + \frac{(1-p_2)^2[p_2^2\pi(1-\pi) + (1-p_2)^2\pi_Y(1-\pi_Y)]}{N-1} \\
C &= \frac{(1-p_2)^2[p_1^2\pi(1-\pi) + (1-p_1)^2\pi_Y(1-\pi_Y)]}{N-1} + \\
&+ \frac{(1-p_2)^2[p_2^2\pi(1-\pi) + (1-p_2)^2\pi_Y(1-\pi_Y)]}{N-1}.
\end{aligned}$$

I valori di n_1 e n_2 tale che $n_1 + n_2 = n$ e tali che la varianza di stima sia minima sono:

$$n_1 = \frac{n\sqrt{A}}{\sqrt{A}+\sqrt{B}} \qquad \text{e} \qquad n_2 = \frac{n\sqrt{B}}{\sqrt{A}+\sqrt{B}}$$

e il valore di minimo della varianza risulta essere pari a:

$$Var_K(\hat{\pi}_G) = \frac{1}{(p_1-p_2)^2}\left[\frac{(\sqrt{A}+\sqrt{B})^2}{n} - C\right] = min$$

Similmente la varianza del secondo metodo proposto può essere scritta come:

$$Var(\hat{\pi}_R) = \frac{1}{p_1^2}\left[\frac{D}{n_1} + \frac{E}{n_2} - F\right]$$

ove

$$D = \theta_1(1-\theta_1) + \frac{\pi(1-\pi)}{N-1}$$

$$E = \frac{N(1-p_1)^2\pi_Y(1-\pi_Y)}{N-1}$$

$$F = \frac{\pi(1-\pi) - (1-p_1)^2\pi_Y(1-\pi_Y)}{N-1}.$$

Da qui, le dimensioni campionarie che minimizzano la varianza sono:

$$\frac{n\sqrt{D}}{\sqrt{D}+\sqrt{E}} \qquad \text{e} \qquad n_2 = \frac{n\sqrt{E}}{\sqrt{D}+\sqrt{E}}$$

e il valore di minimo della varianza risulta essere pari a:

$$Var(\hat{\pi}_R) = \frac{1}{p_1^2}\left[\frac{(\sqrt{D}+\sqrt{E})^2}{n} - F\right] = min.$$

L'efficienza relativa percentuale delle strategie proposte nel confronto con il metodo di Kim del 1978 diviene:

$$\begin{aligned} RE &= \frac{Min[Var_K(\hat{\pi}_G)]}{Min[Var_K(\hat{\pi}_G)]} \times 100 = \\ &= \left(\frac{p_1}{p_1 - P_2}\right)^2 \left[\frac{(\sqrt{A}+\sqrt{B})^2 - nC}{\sqrt{D}+\sqrt{E})^2 - nF}\right] \times 100. \end{aligned}$$

Da uno studio empirico si è potuto notare che l'efficienza relativa della strategia proposta è considerevolmente più alta di quella di Simmons, utilizzando la selezione campionaria proposta da Kim. Inoltre l'efficienza relativa aumenta con la frazione campionaria. Comunque, nel caso in cui la dimensione della popolazione aumenti, ma la frazione campionaria resti invariata, l'efficienza relativa non aumenta.

2.45 Ottenere risposte veritiere indirettamente (di Chua e Tsui del 2000).

Attributi

La procedura che viene qui proposta è basata su due distinti stadi. Nel primo stadio ognuno degli n rispondenti seleziona casualmente un numero senza rinserimento da un sottoinsieme di numeri naturali $\Omega = \{1, 2, 3, ..., m\}$, ove $m \geq n$. Gli interi estratti da tutti i rispondenti non vengono svelati all'intervistatore. Nel secondo stadio, sempre all'oscuro dall'intervistatore, ad ogni rispondente, che ricorda il numero estratto, diciamo k, è richiesto di generare due insiemi di numeri, ognuno dei quali composto di m unità e messe in ordine crescente. Tali insiemi sono generati da due distribuzioni dette rispettivamente F e G. Il rispondente deve comunicare all'intervistatore il *k-mo* valore più grande degli m estratti dalla distribuzione F, se egli possiede la cararteristica oggetto di studio; altrimenti dovrà comunicare il medesimo valore ma con riferimento alla distribuzione G.

Questa tecnica anche se abbastanza complessa pare che ottenga una protezione del rispondente molto sentita dallo stesso, proprio per come viene sviluppata: il doppio stadio di cui il secondo in dipendenza del risultato del primo, la generazione di numeri casuali e la selezione ulteriore di uno solo di questi.

Siano ora

$$X_1, X_2, ..., X_n$$

i valori riportati dal campione di n intervistati. Si dimostra che

$$E(X_i) = \theta\mu_F + (1-\theta)\mu_G,$$

ove μ_F e μ_G sono le medie delle distribuzioni di F e G, rispettivamente, e θ è la proporzione di diffusione del cararttere in studio. Con il metodo dei momenti si può ottenere uno stimatore corretto di θ:

$$\hat{\theta} = \frac{\bar{X} - \mu_G}{\mu_F - \mu_G}$$

la cui varianza risulta essere:

$$\begin{aligned}
Var(\hat{\theta}) &= \frac{1}{n(\mu_F - \mu_G)^2}\theta\left(\sigma_F^2 + \mu_F^2 - \frac{1}{m}\sum_{i=1}^{m}\mu_{F[i:m]}^2\right) + \\
&+ \frac{(1-\theta)}{n(\mu_F - \mu_G)^2}\left(\sigma_G^2 + \mu_G^2 - \frac{1}{m}\sum_{i=1}^{m}\mu_{G[i:m]}^2\right) + \\
&+ \frac{\theta(1-\theta)}{n(\mu_F - \mu_G)^2}\frac{1}{m}\sum_{i=1}^{m}(\mu_{F[i:m]} - \mu_{G[i:m]})^2 + \\
&+ \frac{1}{n(\mu_F - \mu_G)^2}\frac{m-n}{(m-1)m} * \\
&* \sum_{i=1}^{m}(\theta\mu_{F[i:m]} + (1-\theta)\mu_{G[i:m]} - (\theta\mu_F + (1-\theta)\mu_G))^2.
\end{aligned}$$

ove σ_F^2 e σ_G^2 sono le varianze delle distribuzioni F e G; $\mu_{J[i:m]}$ e $\sigma^2_{J[i:m]}$ è la media e la varianza della statistica di ordine i delle m variabili casuali tratte dalla distribuzione F quando $J = F$ e dalla distribuzione G, quando $J = G$, rispettivamente. Conseguentemente $Var(\hat{\theta})$ dipende dalla distanza tra μ_F e μ_G e dalla differenza tra m e n. Inoltre quando il numero di rispondenti eguaglia il numero di valori casuali generati da F o G (si legga $m = n$), la varianza si attesta sul suo valore minimo. Ancora, quando m tende a ∞, $(m-n)(m-1) = 1$, la varianza qui presentata viene a coincidere con quella dello stimatore del metodo di Franklin nel caso particolare in cui si faccia un'unica replicazione per rispondente. Ossia:

$$Var(\hat{\theta}_f) = \lim_{x\to\infty} Var(\hat{\theta}) = \frac{\theta(1-\theta)}{n} + \frac{\theta\sigma_F^2 + (1-\theta)\sigma_G^2}{n(\mu_F - \mu_G)^2}.$$

Inoltre se

$$\frac{m-n}{m-1} < 1$$

allora $Var(\hat{\theta})$ è ovviamente minore di $Var(\hat{\theta}_f)$. Infine $Var(\hat{\theta}_f)$ può essere ridotta alla varianza dello stimatore di Warner ponendo rispettivamente:

$$\mu_F = P \quad \mu_G = 1 - P \quad \sigma_F^2 = P(1-P) \quad \sigma_G^2 = (1-P)P.$$

Concentrandosi sul caso di minima varianza, ossia $m = n$, si può valutare l'efficienza dello stimatore del metodo di Franklin relativamente alla varianza dello stimatore $\hat{\theta}$ come segue:

$$eff(\hat{\theta}_F, \hat{\theta}) = 1 - \frac{Var(\hat{\theta}_F) - Var(\hat{\theta})}{Var(\hat{\theta}_F)}$$

ove

$$Var(\hat{\theta}_F) - Var(\hat{\theta}) = \frac{\sum_{i=1}^{n}(\theta\mu_{F[i:n]} + (1-\theta)\mu_{G[i:n]} - (\theta\mu_F + \mu_G(1-\theta)))^2}{n^2(\mu_F - \mu_G)^2}$$

2.46 Stima della media e della varianza di una variabile quantitativa delicata utilizzando unità distinte nel campionamento RR (di Singh, Mahmood e Tracy del 2001).

Variabili

Nella strategia che viene qui proposta, ogni rispondente facente parte del campione, è provvisto di due scatole A e B contenenti ciascuna carte riportanti numeri del tipo $a_1, a_2, \ldots, a_l$ (tali che $a_j > 0$) e $b_1, b_2, \ldots, b_r$ le cui medie e varianze, note, sono rispettivamente $\bar{A}$, $\bar{B}$ e σ_A^2, σ_B^2. Ad ogni rispondente viene chiesto di pescare casualmente due carte, una per scatola, poi dovrà moltiplicare il valore vero posseduto, circa la variabile delicata oggetto di studio e detta (Y), per il valore ottenuto dalla scatola A ed aggiungervi il valore ottenuto dalla scatola B. In altre parole il rispondente *i-mo* fornirà come risposta il valore

$$Z_i = a_f Y_i + b_q.$$

Se un rispondente venisse scelto più di una volta, egli dovrà fornire il medesimo valore Y_i visto che anche in questa seconda volta la sua privacy, come pure nella prima, è garantita dal fatto che i valori delle carte che camufferanno la risposta, saranno quasi certamente diversi.

Indicando con E_R, V_R e C_R la media, la varianza e la covarianza nell'ambito della tecnica RR e indicando con

$$R_i = \bar{A}^{-1}(Z_i - \bar{B})$$

la risposta RR trasformata, si ha che

$$E_R(R_i) = Y_i$$

$$Var_R(R_i) = \bar{A}^{-2}(\sigma_A^2 Y_i^2 + \sigma_B^2) = \sigma_i^2$$

e

$$C_R(R_i, R_j) = 0 \quad \text{per} \quad i \neq j.$$

E' importante notare che i numeri contenuti nelle scatole potrebbero avere una variabilità simile a quella del carattere delicato: in questo caso il rispondente potrebbe essere maggiormente portato a dire la verità. Supponiamo ora che R_{di} e R_{ri} siano rispettivamente la risposta RR trasformata del rispondente distinto, ossia se egli non viene più estratto, e del rispondente che invece viene invitato nuovamente a rispondere.

Uno stimatore corretto della media di una popolazione finita e costituita da u unità distinte e campionata con tecnica di campionamento casuale semplice con reinserimento, è dato da:

$$\bar{y}_u^\star = \frac{1}{u}\sum_{i=1}^{n} R_{di}$$

la cui varianza è data da:

$$Var(\bar{y}_u^\star) = \frac{1}{N}E\left(\frac{1}{u}\right)\sum_{i=1}^{N}\sigma_i^2 + \left[E\left(\frac{1}{u}\right) - \frac{1}{N}\right]s_y^2$$

ove

$$s_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

è l'usuale varianza campionaria corretta.

Uno stimatore corretto della medesima media ma basato su tutte le selezioni campionarie, anche quelle ripetute, è dato da:

$$\bar{y}_n^\star = \frac{u\bar{y}_u^\star + (n-u)\bar{y}_2}{n}$$

ove

$$\bar{y}_2 = \frac{1}{n-u}\sum_{i=1}^{n-u} R_{ri}$$

è la media basata sulle sole prove ripetute dei rispondenti selezionati più volte, escludendo però la loro prima risposta.

Inoltre

$$R_{ri} = \frac{Z_i^\star - \bar{B}}{A}$$

per

$$Z_i^\star = a'_f Y_i + b'_q$$

ove a'_f e b'_q sono i numeri casuali selezionati dai rispondenti nelle differenti prove e gli $Z_i^\star$ sono le diverse risposte ottenute nelle medesime prove ripetute dai rispondenti.

La varianza dello stimatore è data da:

$$Var(\bar{y}_n^\star) = \frac{1}{nN}\left[\sum_{i=1}^{N}\sigma_i^2 + (N-1)s_y^2\right].$$

Si noti che lo stimatore della media basato sulle unità distinte in generale non è superiore a quello che si ottiene dalla procedura RR come ne è il caso invece dell'intervista diretta.

Per quanto concerne invece la stima della varianza della medesima popolazione,

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

utilizzando le unità distinte, si ha :

$$\hat{Var}(\hat{\sigma_y^2}) = \left[\frac{C_u(n) - C_u(n-1)}{C_u(n)}\right] s_u^2$$

ove

$$C_u(n) = u^n - \begin{pmatrix} u \\ 1 \end{pmatrix}(u-1)^n + ... + (-1)^{u-1}\begin{pmatrix} u \\ u-1 \end{pmatrix}$$

e

$$s_u^2 = (\sigma_A^2 + \bar{A}^2)^{-1}\left[\frac{\bar{A}^2}{u-1}\sum_{i=1}^{u}(R_{di} - \bar{y}_u^\star)^2\right].$$

Il problema qui esposto potrebbe essere esteso a quello della stima del coefficiente di correlazione.

2.47 Stima della proporzione di un carattere qualitativo (di Chang e Huang del 2001).

Attributi

La tecnica proposta da Chang e Huang nel 2001 nasce dall'associazione tra un'indagine a risposta diretta *(DR)* e il modello di Warner *(RR)* e si propone di stimare simultaneamente la proporzione dei soggetti appartenenti al gruppo sensibile **A** e la probabilità che i membri di **A** siano sinceri.

Nel metodo proposto da Chang e Huang un campione di n elementi viene estratto senza reinserimento dalla popolazione e scisso in due sottocampioni di dimensione n_j, $n_j = 1, 2$, tali che $n_1 + n_2 = n$.

Ciascun membro del j-esimo campione è invitato direttamente a una domanda relativa alla sua appartenenza ad **A** e, se risponde negativamente a tale quesito, a sperimentare il modello di Warner.

Ipotizzando che i rispondenti appartenenti ad A dicano sempre il vero se confrontati con la tecnica RR e lo dicano con probabilità T alla domanda posta loro direttamente, la proporzione degli intervistati che nel $j - mo$ campione rispondono affermativamente è pari a:

$$\theta_j = \pi_A T + \pi_A(1-T)P_j + (1-\pi_A)(1-P_j) \tag{17}$$

dove
P_j= probabilità di selezione della domanda "Appartieni ad A?";
$(1 - P_j)$=probabilità di selezione della domanda "Appartieni ad $\bar{A}$?".
Dalla (17) sono ottenibili uno stimatore di π_A e T:

$$\begin{aligned} \hat{\pi}_{ACH} &= \frac{(1-P_2)\bar{Z}_1 - (1-P_1)\bar{Z}_2}{(P_1 - P_2)} \\ \hat{T}_{CH} &= \frac{(1-2P_2)\bar{Z}_1 - (1-2P_1)\bar{Z}_2 - (P_1 - P_2)}{(1-P_2)\bar{Z}_1 - (1-P_1)\bar{Z}_2} \end{aligned}$$

dove
$\bar{Z}_j = \frac{\sum_{i=1}^{n_j} Z_{ij}}{n_j}$ è una variabile bernoulliana di parametri θ_j e n_j, per $i = 1, 2, ..., n_j$ e $j = 1, 2$.
La varianza dello stimatore $\hat{\pi}_{ACH}$ è pari a:

$$Var(\hat{\pi}_{ACH}) = \frac{1}{(P_1 - P_2)^2}\left[\frac{(1-P_2)^2\theta_1(1-\theta_1)}{n_1} + \frac{(1-P_1)^2\theta_2(1-\theta_2)}{n_2}\right]$$

Prendiamo ora in considerazione lo stimatore $\hat{T}_{CH}$ e determiniamo l'espressione della distorsione

$$\begin{aligned} D(\hat{T}_{CH}) &= \frac{[T(1-P_2) + (2P_2 - 1)]^2\theta_1(1-\theta_1)}{(P_1 - P_2)^2\pi^2 n_1} + \\ &+ \frac{[T(1-P_1) + (2P_1 - 1)]^2\theta_2(1-\theta_2)}{(P_1 - P_2)^2\pi^2 n_2} \end{aligned}$$

e dello scarto quadratico medio

$$\begin{aligned} \sigma(\hat{T}_{CH}) &= \frac{(1-P_2)[T(1-P_2) + (2P_2 - 1)]\theta_1(1-\theta_1)}{(P_1 - P_2)^2\pi^2 n_1} + \\ &+ \frac{[(1-P_1)T(1-P_1) + (2P_1 - 1)]\theta_2(1-\theta_2)}{(P_1 - P_2)^2\pi^2 n_2} \end{aligned}$$

Cocludiamo l'analisi di questa tecnica con lo studio della sua efficienza relativa (RE) rispetto al modello di Warner e al caso di risposta diretta.

Il primo confronto produce un valore di RE pari a:

$$RE = \frac{\theta_W(1-\theta_W)(P_1-P_2)^2}{(2P_W-1)^2[(1-P2)\sqrt{\theta_1(1-\theta_1)}+(1-P_1)\sqrt{\theta_1(1-\theta_1)}]}$$

dove $P_W = P_1$ e $P_2 = 1 - P_1$.

La tecnica di Chang e Huang risulta pertanto sempre preferibile al modello di Warner e in particolare per valori elevati di π_A e T.

Confrontando ora la tecnica descritta e la tecnica di risposta diretta, si ottiene un valore di RE pari a:

$$RE = \frac{[\theta_D(1-\theta_D)+n\pi_A(1-T)^2](P_1-P_2)^2}{[(1-P2)\sqrt{\theta_1(1-\theta_1)}+(1-P_1)\sqrt{\theta_1(1-\theta_1)}]}$$

In questo caso si osserva che RE dipende anche da n è che è notevolmente più elevata rispetto al caso in cui si ricorre a una domanda diretta, soprattutto quando aumenta P_1 o n e diminuisce T.

2.48 Un nuovo modello RR (di Gupta, Gupta e Singh del 2002).

Variabili

In questo modello ogni rispondente selezionato con campinamento casuale semplice con reinserimento può scegliere una delle seguenti due opzioni:

1. riportare il vero valore posseduto di X
2. riportare il valore camuffato SX

ove S è la variabile di camuffamento e X è la variabile delicata oggetto di studio. Qui viene supposto che X e S siano entrambe positive con $\mu_X = 1$ e $\mu_S^2 = \gamma^2$. Il modello potrebbe essere quindi così scritto:

$$Z = S^Y X$$

ove Y è una variabile casuale così definita:

$$Y = \begin{cases} 1 & \text{se la risposta è camuffata} \\ 0 & \text{altrimenti} \end{cases}$$

Se W è la probabilità che un rispondente fornirà la risposta camuffata, allora Y è una variabile casuale di tipo bernoulli con $E(Y) = W$. In generale W può essere visto come il livello di delicatezza del carattere A e il suo valore sarà prossimo a 1 nel caso di carattere molto delicato e prossimo a 0 nel caso di carattere scarsamente delicato.

Con queste premesse, uno stimatore corretto della media della popolazione, μ_X è dato da:

$$\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

la cui varianza è pari a:

$$Var(\hat{\mu}_X) = \frac{1}{n}\left[\sigma_X^2 + W\gamma^2(\sigma_X^2 + \mu_X^2)\right].$$

Uno stimatore della varianza è dato da:

$$\hat{Var}(\hat{\mu}_X) = \frac{1}{n}[s_X^2 + \hat{W}\gamma^2(s_X^2 + \hat{\mu}_X^2)]$$

ove

$$s_X^2 = \frac{s_Z^2 - \hat{W}\gamma^2\hat{\mu}_X^2}{1 + \hat{W}\gamma^2}$$

con

$$s_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$$

è uno stimatore di σ_X^2. Si noti che la varianza è funzione crescente di W
Per la stima di W riprendiamo la relazione

$$Z = S^Y X$$

e applicando la funzione logaritmica ad ambo i membri si ha:

$$\log(Z) = Y\log(S) + \log(X)$$

e dopo alcuni passaggi matematici si ottiene

$$\hat{W} = \frac{\frac{1}{n}\sum_{i=1}^{n}\log(Z_i) - \log(\frac{1}{n}\sum_{i=1}^{n}Z_i)}{\delta}$$

ove

$\delta = E[\log(S)]$la media nota della variabile logaritmo delle risposte camuffate.

Uno stimatore della varianza di $\hat{W}$ è dato da:

$$\hat{Var}(\hat{W}) = \frac{\hat{W}(1-\hat{W})}{n-1}.$$

Un test per W può essere così impostato. Siano le ipotesi nulla ed alternativa del tipo:

$$H_0 : W = W_0 \qquad \text{e} \qquad H_1 : W \neq W_0$$

il test statistico sarà:

$$Z_c = \frac{\hat{W} - W_0}{\sqrt{\frac{W_0(1-W_0)}{n}}}$$

che si distribuisce normalmente e per l'accettazione o il rigetto di H_0 si segue la metodologia classica dei test circa la regione di accettazione e quella di rifiuto.

2.49 Implementazione di un piano di campionamento RR.

Per implementare correttamente un piano di campionamento RR è bene considerare le seguenti linee guida.

1. La tecnica RR deve essere preferita ad una classica tecnica di intervista qualora:
   - si debba studiare un attributo ritenuto particolarmente imbarazzante;
   - si preveda una evasione nelle risposte superiore al 50%;
   - la numerosità campionaria sia elevata: almeno 300 rispondenti;
   - al massimo il 10% della popolazione possegga l'attributo oggetto di studio.
2. In una fase di colloquio di tipo *faccia a faccia*, il ricercatore deve spiegare ai rispondenti in modo diretto e sintetico la natura della procedura RR ed il perchè del suo utilizzo. Segue poi una illustrazione del carattere oggetto di studio, delle tipologie di risposte che debbono essere fornite e del casualizzatore utilizzato. Dopo la distribuzione di quest'ultimo si devono prevedere alcune prove di utilizzo del casualizzatore stesso affinchè l'intervistato possa familiarizzarvi; questa fase ha il duplice vantaggio di ridurre gli errori nell'uso del meccanismo di casualizzazione e di convincere maggiormente i rispondenti circa la protezione del loro anonimato.
3. A questa fase sperimentale deve seguire un piccolo incontro informativo circa l'appena avvenuta familiarizzazione; il ricercatore deve valutare la percezione del rispondente nei riguardi dell'anonimato garantito dal meccanismo e nei riguardi dell'utilizzo del casualizzatore. Ovviamente se si abbina la tecnica RR con l'intervista telefonica o postale questa fase non può perfezionarsi.
4. La implementazione di un modello RR postale richiede ovviamente l'utilizzo di un casualizzatore che possa essere veramente spedibile anche nel rispetto della privacy del rispondente. Non essendo presente il ricercatore all'atto dell'uso del casualizzatore, varie precauzioni sono d'obbligo. Il casualizzatore deve essere di facile utilizzo, immediato e deve assicurare al rispondente, più che mai, una reale casualizzazione della domanda. E' consigliato inoltre di investigare soltanto un attributo delicato per evidenti motivi di semplicità e di sicurezza. Al rispondente inoltre deve essere chiesto di spedire la risposta, non identificabile, ad una mail box all'indirizzo indicato; inoltre deve essere fornito al rispondente il nome e il numero di telefono del ricercatore per far sì che questi possa tempestivamente rispondere ad eventuali domande dell'intervistato.
5. In presenza di rispondenti di basso livello di istruzione o di rispondenti molto giovani o facilmente distraibili, è bene utilizzare modelli molto semplici, come quelli a proporzione singola o a domanda incorrelata.
6. Per ridurre il rischio di insospettire il rispondente nell'ambito del modello a domanda incorrelata, il ricercatore deve selezionare la domanda incorrelata in modo opportuno: essa deve riferirsi veramente a una caratteristica

ritenuta, in senso generale, innoqua. Per esempio chiedere se si è mancini non è di certo una buona scelta se il ricercatore prevede una fase di scrittura da parte dell'intervistato. Così anche se il rispondente ritiene che il ricercatore abbia accesso a dati riservati, non risulterebbe essere una buona scelta prevedere domande incorrelate legate ad alcune date di nascita, ai numeri telefonici o ai numeri di documenti sanitari.

7. Il modello a domanda incorrelata risulta essere superiore al modello ad intervista diretta nel ridurre il rischio di identificazione del rispondente, qualora l'attributo oggetto di studio risultasse essere particolarmente delicato. Poichè però l'efficienza relativa del primo è maggiore, il ricercatore dovrà opportunamente aumentare la dimensione campionaria se optasse per il modello a domanda incorrelata.

8. Qualora si suddividesse il campione oggetto di studio in due sottocampioni per implementare il modello a domanda incorrelata di Simmons, il maggior numero di rispondenti deve essere allocato nel sottocampione in cui si ha il più alto valore di p. Greenberg *et. al.* (1971) consigliano di rispettare la seguente relazione nell'allocazione delle unità campionarie: $\frac{n_1}{n_2} \approx \frac{p_1}{p_2}$.

2.50 Conclusioni e scenari futuri.

Non vi è dubbio che l'idea sviluppata dal suo pioniere sia molto originale in specie analizzando i contributi presenti in letteratura. Si può infatti affermare che questi ultimi prendano le mosse da alcuni modelli considerati ancora oggi di riferimento: il metodo di Warner, quello di Simmons, quello di Moors ed infine quello di Kuk. Ad oggi quindi non vi sono stati interventi di rilievo che mettessero al bando la tecnica esaminata ma piuttosto tentativi di ottimizzazione delle tecniche già presentate. Ciò sottolinea, come si diceva, la genialità della intuizione di Warner e dei suoi seguaci.

Alcuni autori sono riusciti ad escogitare qualche procedura in grado di rendere veramente minima la variabilità delle stime; purtroppo però si tratta di tecniche apprezzabili solamente dal punto di vista teorico poichè costituite da un casualizzatore o da un meccanismo talmente complicati da scoraggiare o insospettire l'intervistato.

La tecnica Randomized Response presenta in generale alcuni svantaggi ma anche alcuni vantaggi che possono essere così sommariamente descritti.

Gli svantaggi.

1. Il campione che si estrae con la tecnica RR deve essere di numerosità maggiore rispetto al campione che si estrae utilizzando l'intervista diretta (se si fissa un certo errore di stima), poichè la prima tecnica presenta una perdita di efficienza nelle stime la cui variabilità risente di un fattore addittivo dovuto proprio alla casualizzazione.

2. La tecnica RR necessità di una tempistica più estesa per selezionare il modello, per scegliere i parametri di interesse, per escogitare le domande o affermazioni delicate e non delicate, per scegliere un opportuno casualizzatore, per spiegare agli intervistati come funziona il casualizzatore e come utilizzarlo e per predisporre le procedure per la elaborazione dei dati.

3. I costi di gestione sono elavati anche facendo solamente riferimento al casualizzatore che deve essere creato, riprodotto in quantitativi elevati (bisogna prevedere alcuni casualizzatori di scorta nel caso qualcuno si rompesse o si usurasse); ma anche nel caso di intervista postale vi sono costi elevati di spedizione del questionario ma soprattutto del casualizzatore.

4. I ricercatori che vogliono fare uso di questa tecnica devono avvalersi di intervistatori esperti e di esperti della raccolta, classificazione ed elaborazione delle informazioni.

5. Utilizzando la tecnica che prende informazioni a priori di un carattere π_Y, il ricercatore necessita comunque di un *data base* affidabile.

6. La tecnica RR risulta difficoltosa da applicare alle interviste telefoniche o a quelle postali.

7. Il successo della tecnica dipende da molti fattori che potremmo dire controllabili dal ricercatore ma anche da un fattore non controllabile dal ricercatore: lo skill dell'intervistato; questo deve essere tenuto in particolare considerazione poichè alcuni casualizzatori hanno destato forti sospetti nell'intervistato per la probabilità percepita da esso di estrarre una certa domanda e non per quella reale derivante dalla metodologia.

I vantaggi

1. Il vantaggio principale è sicuramente quello legato alla riservatezza del rispondente le cui risposte non possono in nessun modo essere ricondotte ad esso. In questo modo il ricercatore può esplicitare l'oggetto dell'indagine senza dover ricorrere ad una intervista strutturata con domande indirette e di controllo.

2. La tecnica RR aumenta la fiducia nel rispondente con una relativa minimizzazione delle risposte false, delle risposte evasive, della non cooperazione in temi così delicati come la sterilizzazione, l'aborto, l'eutanasia, le abitudini sessuali, le attività illegali quali l'uso di droga, la criminalità, la prostituzione, l'immigrazione clandestina, l'evasione fiscale.

3. La generalizzazione ad altre popolazioni dei risultati ottenuti è spesso utilizzata poichè la tecnica RR fornisce stime della proporzione di popolazioni che posseggono una certa attitudine, o un certo attributo, più accurati.

4. La tecnica RR può essere utilizzata in modo più efficace se viene svolta una indagine su di una popolazione per capirne le attitudini e per indirizzare investimenti di una certa portata per tentare di risolvere o di modificare alcuni comportamenti sociali ritenuti devianti.

5. Nella particolarità della intervista postale è stato dimostrato che il tasso di risposta con tecnica RR è migliore rispetto a quello di intervista postale classica (senza considerare evidentemente le differenze di costo).

6. I costi di intervista possono essere abbattuti se considerati nel lungo periodo: una efficace impostazione iniziale dell'indagine potrebbe essere utilizzata per altre indagini; perlomeno i casualizzatori potrebbero essere riutilizzati.

Se si volessero evidenziare alcuni aspetti particolari meritevoli di approfondimento si potrebbe sicuramente tentare, alla luce delle riflessioni poc'anzi esposte, una generalizzazione delle varie procedure, con riferiemento ad un numero molto limitato di famiglie; dal punto di vista meramente metodologio alcuni metodi potrebbero essere ripresi nell'ottica della mistura di distribuzioni che in taluni casi parrebbe essere particolarmente adatta; l'ambito dello studio delle variabili, come quello degli attributi multipli, risultano essere stati studiati in modo marginale; potrebbe essere valido anche uno studio della tecnica in relazione alla legge sulla privacy per meglio chiarire i limiti che quest'ultima impone al ricercatore; collegato a questa vi è anche lo studio della misura della protezione che il metodo offre all'intervistato; infine sarebbe particolarmente interessante una applicazione della metodologia alle nuove tecnologie che di forza stanno entrando nell'uso comune. Il riferimento è rivolto ad internet, che potrebbe essere un utile mezzo per abbattere i costi di indagine e per raggiungere velocemente una grossa fetta della popolazione, certi del fatto, comunque, che questo strumento non presenta oggi una diffusione omogenea su scala mondiale.

Riferimenti bibliografici

[1] Randomized response in more than one question. *Social Statistic Section of A.S.A.*, 1975.

[2] James R. Abernathy, Bernard G. Greenberg, and Daniel G. Horvitz. Estimates of induced abortion in urban north carolina. *Demography*, 1970.

[3] Arun Kumar Adhikari, Arijit Chaudhuri, and K. Vijayan. Optimum sampling strategies for randomized response trials. *International Statistical Institute*, 1984.

[4] S.E. Ahmed. To pool or not to pool ithe proportion in randomized response surveys. *Commun. Stat. Theory and Math.*, 1997.

[5] S.E. Ahmed and S.M. Khan. Combining proportions from several randomized response model. *Interstat*, 1997.

[6] S.E. Ahmed and V.K. Rohatgi. Shrinkage estimation of the proportion in randomized response. *Metrika*, 1996.

[7] Ronald L. Akers, James Massey, William Clarke, and Ronald M. Lauer. Are self-reports of adolescent deviance valid? biochemical measures, randomized response, and the bogus pipeline in smoking behavior. *The University of North Carolina Press*, 1983.

[8] Harald Anderson. Estimation of a proportion trough randomized response. *Int. Stat. Rev.*, 1976.

[9] Richard F. Antonak and Hanoch Livneh. Randomized response technique: A review and proposed extension to disability attitude research. *Genetic, Social and General Psychology Monographs*, 1995.

[10] Raghunath arnab. On comutativity of design and model expectations in randomized response surveys. *Comm. Satistist. Theory Meth.*, 1990.

[11] Raghunath Arnab. Nonnegative variance estimation in randomized response surveys. *Commun. Statist. Theory Meth.*, 1994.

[12] Raghunath Arnab. On admissibility and optimality of sampling strategies in randomized response surveys. *Sankhya*, 1995.

[13] Raghunath Arnab. Randomized response surveys: Optimun estimation of a finite population total. *Statistical Paper*, 1998.

[14] Guy Begin and Michel Boivin. Comparison of data gathered on sensitive questions via direct questionnarie, randomized response technique, and a project method. *Psychological Reports*, 1980.

[15] David R. Bellhouse. Linear model for randomized response designs. *Journal of the American Statistical Association*, 1980.

[16] Sandra F. Belt, Wayne W. Daniel, and Bikramjit S. Garcha. The takahasi-sakasegawa randomized response technique. *Sociological Methods & Research*, 1982.

[17] P.D. Bourke. On the analysis of some multivariate randomized response designs for categorial data. *Journal of Statistical Planning and Inference*, 1981.

[18] P.D. Bourke. Randomized response multivariate designs for categorial data. *Commun. Statist. Theor. Meth.*, 1982.

[19] P.D. Bourke and T. Delenius. Some new ideas in the real of randomized inquiries. *Int. Stat. Rev.*, 1976.

[20] P.D. Bourke and Michael A. Moran. Estimating proportions from randomized response data using the em algorithm. *Journal of the American Statistical Association*, 1988.

[21] Franco Bressan. Lo schema di warner dotato di memoria. *Rivista di Statistica Applicata*, 1984.

[22] Franco Bressan. Possibilità di ampliamento dello schema di simmons nell'applicazione a gruppi di persone. *Rivista di Statistica Applicata*, 1984.

[23] Cathy Campbell and Brian L. Joiner. How to get the answer without being sure you've asked the question. *The American Statistician*, 1973.

[24] Arijit Chadhury, Arun K. Adhikary, and Tapabrata Maiti. A note on non-negative mean square error estimation of regression estimators in randomize response surveys. *Statistical Paper*, 1998.

[25] Horng-Jinh Chang and Kuo-Chung Huang. Estimation of proportion and sensivity of a qualitative character. *Metrika*, 2001.

[26] Horng-Jinh Chang and Der-Hsin Liang. A two-stage unrelated randomized response procedure. *Australian journal of Statistics*, 1996.

[27] Orng-Jinh Chang and Der-Hsin Liang. Randomized response trials: A unified approach for qualitative data. *Australian journal of Statistics*, 1996.

[28] A. Chaudhuri. Randomized response: Estimating mean square errors of linear estimators and finding optimal unbiased strategies. *Metrika*, 1992.

[29] A. Chaudhuri and R. Mukherjee. Randomized response techniques: A riview. *Indian Statistical Institute*, 1987.

[30] A. Chaudhuri and H. Stenger. *Randomize Response.* Survey Sampling: Theory and Methods, 1991.

[31] Arijit Chaudhuri. Randomized response surveys of finite populations: a unified approach with quantitative data. *Journal of Statistical Planning and Inference*, 1986.

[32] Arijit Chaudhuri. Variance estimation with randomize response. *Comm. Satistist. Theory Meth.*, 1990.

[33] Arijit Chaudhuri. Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. *Journal of Statistical Planning and Inference*, 2001.

[34] Arijit Chaudhuri and Debesh Roy. Model assisted survey sampling startegies with randomized response. *Journal of Statistical Planning and Inference*, 1997.

[35] T. Timothy Chen. Log-linear models for the categorial data obteined from randomized response techniques. *A.S.A. Proceedings of the Section on Survey Research Methods*, 1978.

[36] T.C. Chua and Albert K. Tsui. Procuring honest responses indirectly. *Journal of Statistical Planning and Inference*, 2000.

[37] Tin Chiu Chua and Kok Leong Chiang. An alternative estimator for randomized response sampling with continuous distributions from a dichotomous population. *Commun. Statist. Theory Meth.*, 1995.

[38] Robert P. Clickner and Boris Iglewicz. Warner's randomized response technique: the two sensitive question case. *Proceedings of the Social Statistic Section of A.S.A.*, 1976.

[39] R. Colombi. La casualizzazione delle risposte. *Quaderni di Statistica e Matematica Applicati alle Scienze Sociali - Università Cattolica di Milano*, 1978.

[40] T.A. Dowling and Richard H. Shachtman. On the relative efficiency of randomized response models. *Journal of the American Statistical Association*, 1975.

[41] John C. Duffy and Jennifer J. Waterton. Randomized response models for estimating yhe distribution function of a quantitative character. *International Statistical Institute*, 1984.

[42] Benjamin H. Eichhorn. Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 1983.

[43] Abdel Latif A. Abul Ela, Bernard G. Greenberg, and Daniel G. Horvitz. A multiproportions randomized response model. *J.A.S.A.*, 1967.

[44] Abel Latif A. Abul Ela and Sultan M. Abdel Hamied. A randomized response ratio estimate from quantitative data. *Proceedings of the Survey research section of A.S.A.*, 1985.

[45] Sven A. Eriksson. A new model for randomized response. *Int. Stat. Rev.*, 1973.

[46] Pieralda Ferrari. Lo schema di campionamento a due stadi con risposte casualizzate e con numerosità di primo stadio ignote. *Quaderni di Statistica e Matematica Applicata alle Scienze Economico-Sociali*, 1982.

[47] Pieralda Ferrari. Lo schema di campionamento a due stadi con risposte casualizzate e con numerosità di primo stadio ignote. *Quaderni di Statistica e Matematica Applicata alle Scienze Economico-Sociali*, 1983.

[48] Pieralda Ferrari and Donata Marasini. Il campionamento con risposte casualizzate: stima e verifica di ipotesi. *Quaderni di Statistica e Matematica Applicata alle Scienze Economico-Sociali*, 1983.

[49] Dorothy S. Fidler and Richard E.Kleinknecht. Randomized response versus direct questioning: Two data-collection methods for sensitive information. *Psychological Bulletin*, 1977.

[50] Ralph E. Folson, Bernard G. Greenberg, Daniel G. Horvitz, and James R. Abernathy. The linear randomized response model. *Journal of the American Statistical Association*, 1973.

[51] James Alan Fox and Paule E. Tracy. The randomized response approach. applicability to criminal justice research and evaluation. *Evaluation Review*, 1980.

[52] LeRoy A. Franklin. A comparison of estimators for randomized response sampling with continuous distribuitions from a dichtotomous population. *Commun. Statist. Theory Meth.*, 1988.

[53] V.P. Godambe. Estimation in randomised response trials. *International Statistical Review*, 1980.

[54] Michael S. Goodstadt and Valerie Gruson. The randomized response technique: A test on drug use. *Journal of the American Statistical Association*, 1975.

[55] A.L. Gould, B.V. Shah, and J.R. Abernathy. Unrelated question randomized response technique with two trials per respondent. *Proceedings of Social Statistic Section, American Statistical Association.*, 1969.

[56] Bernard G. Greenberg, Abdel Latif A. Abul Ela, Walt R. Simmons, and Daniel G. Horvitz. The unrelated question randomized response model: Theoretical framework. *J.A.S.A.*, 1969.

[57] Bernard G. Greenberg, Roy R. Kuebler, James R. Abernathy, and Daniel G. Horvitz. Application of the randomized response technique in obtaining quantitative data. *Journal of the American Statistical Association*, 1971.

[58] B.G. Greenberg, R.R. Kuebler, J.R. Abernathy, and D.G. Horvitz. A note on the randomized response technique. *Journal of Statisrical Planning and Inference*, 1977.

[59] Bernard G. Greengberg, James R. Abernathy, and Daniel G. Horvitz. Application of the randomized response tecnique in obtaining quantitative data. *Proocedings of the social statistic section of A.S.A.*, 1969.

[60] Sat Gupta, Bhisham Gupta, and Sarjinder Singh. Estimation of sensitivity levl of personal interview survey questions. *Journal of Statistical Planning and Inference*, 2002.

[61] Hedayat and Sinha. *Randomized Response: A Data-Gathering Tool for Sensitive Characteristics.* Design and Inference in Finite population Sampling, 1991.

[62] Peter G.M. Van Der Heijden, Ger Van Gils, Jan Bouts, and Joop J. Hox. A comparison of randomized response, computer assisted self-interview, and face-to-face direct questioning. *Sociological Methods & Research*, 2000.

[63] D.G. Horvitz, B.G. Greenberg, and J.R. Abernathy. Randomized response: a data-gathering device for sensitive questions. *Int. Stat. Rev.*, 1976.

[64] Jong IK Kim and John A. Flueck. A review of randomized response models and some new results. *Proceedings of the Social Statistic Section of A.S.A.*, 1976.

[65] Jong-Ik Kim and John A. Flueck. Modification of the randomized response technique for sampling without replacement. *Proceedings of the Section on Survey Research Method of A.S.A.*, 1978.

[66] K. Krishnamoorthy and D. Raghavarao. Untruthful answering in repeated randomized response procedures. *The Canadian Journal of Statistics*, 1993.

[67] Anthony Y.C. Kuk. Asking sensitive questions indirectly. *Biometrica*, 1990.

[68] Charles W. Lamb and Donald E. Stem. An empirical validation of the randomized response technique. *Journal of Marketing Research*, 1978.

[69] Gianpiero Landenna, Donata Marasini, and Pieralda Ferrari. Schemi di campionamento a due stadi con risposte casualizzate. *Atti della S.I.S.*, 1982.

[70] Johannes A. Landsheer, Peter Van Der Hejden, and Ger Van Gils. Trust and understanding, two psychological aspect of randomized response. *Quality and Quantity*, 1999.

[71] Jan Lanke. On the choice of the unrelated question in simmons' version of randomized response models. *Journal of the American Statistical Association*, 1975.

[72] Jan Lanke. On the degree of protection in randomized interviews. *Int. Stat. Rev.*, 1976.

[73] Ernest R. Larkins, Evelyn C. Hume, and Bikramjit S. Garcha. The validity of the randomized response method in tax ethics research. *Journal of Applied Business Research*, 1997.

[74] Frederick W. Leysieffer and Stanley L. Warner. Respondent jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association*, 1976.

[75] P.T. Liu, C.N. Chen, and L.P. Chow. A study of feasibility of hopkins randomized response models. *Proceedings of the Social Statistical section of A.S.A.*, 1976.

[76] P.T. Liu and L.P. Chow. The efficiency of the multiple trial randomized response technique. *Biometrics*, 1976.

[77] P.T. Liu and L.P. Chow. A new discrete quantitative randomized response model. *Journal of the American Statistical Association*, 1976.

[78] P.T. Liu, L.P. Chow, and W.H. Mosley. Use of randomized response technique with a new randomizing device. *Journal of the American Statistical Association*, 1975.

[79] Lars Ljungqvist. A unified approach to measures of privacy in randomized response models: A utilitarian perspective. *Journal of the American Statistical Association*, 1993.

[80] William Locander, Seymour Sudman, and Norman Bradburn. An investigation of interview method, threat and response distortion. *Journal of the American Statistical Association*, 1976.

[81] R.M. Loynes. Asymptotically optimal randomized response procedures. *Journal of the American Statistical Association*, 1976.

[82] Naurag Singh Mangat, Ravindra Singh, and Sarjinder Singh. Violation of respondent's privacy in moors' model - its ratification through a random group strategy. *Commun. Stat. Theory and Math.*, 1997.

[83] Naurang Singh Mangat. An optimal randomized response sampling technique using non-stigmatized attribute. *Statistica*, 1991.

[84] Naurang Singh Mangat and Ravindra Singh. An alternative approach to randomized response survey. *Statistica*, 1993.

[85] Naurang Singh Mangat, Ravindra Singh, and Sarjinder Singh. An improved unrelated question randomized response startegy. *Calcutta Statistical Association Bulletin*, 1992.

[86] Naurang Singh Mangat, Ravindra Singh, and Sarjinder Singh. On the use of a modified randomization device in randomized response inquireies. *Metrom*, 1993.

[87] N.S. Mangat. An improved randomized response strategy. *Royal Statistical Society*, 1994.

[88] N.S. Mangat and Ravindra Singh. An alternative randomized response procedure. *Biometrica*, 1990.

[89] Donata Marasini. Le risposte casualizzate in uno schema di campionamento a due stadi. *Quaderni di Statistica e Matematica applicata alle Scienze Economico-Sociali*, 1981.

[90] Giorgio Marbach. Sull'uso di quesiti che tutelano completamente la riservatezza dell'informazione. *Metron*, 1975.

[91] Norman S. Matloff. Use of covariates in randomized response settings. *Statistic & Probability Letters*, 1984.

[92] Aride Mazzali. Uno schema a risposte casualizzate per il campionamento di un carattere con k madalità. *Rivista Italiana di Statistica*, 1984.

[93] Helio S. Migon and Wilma M. Tachibana. Bayesian approximation in randomized response model. *Computational Statistics & Data Analysis*, 1997.

[94] J. J. A. Moors. Optimization of the unrelated question randomized response model. *Journal of the American Statistical Association*, 1971.

[95] Mark Moriarty and Frederick Wiseman. On choiceof a randomization response technique with the randomized response model. *Proceedings of the Social Statistic Section of A.S.A.*, 1976.

[96] Tapan K. Nayak. On randomized response surveys for estimating a proportion. *Commun. Statist. Theory Meth.*, 1994.

[97] Dario Olivieri. Una versione modificata dello schema di simmons. *Rivista di Statistica Applicata*, 1983.

[98] Dario Olivieri. La stratificazione nel campionamento r.r. con risposta alternativa fissa, dotato di memoria. *Rivista di Statistica Applicata*, 1984.

[99] Dario Olivieri. Stima dei parametri ed efficenza nello schema a risposte casualizzate di poole. *Riunioni della Società Italiana di Statistica*, 1984.

[100] Robert G. Orwin and Robert F. Boruch. Rrt meets rdd statistical startegies for assuring response privacy in telephone surveys. *Public Opinion Quarterly*, 1982.

[101] V.R. Padmawar and K. Vijayan. Randomized response revisited. *Journal of Statistical Planning and Inference*, 2000.

[102] Shyamal Das Peddada and K.M. Lal Saxena. Minimum norm constrained estimators in randomized response surveys. *The Indian Journal of Statistics*, 1991.

[103] K.H. Pollock and Yuksel Bek. A comparison of three randomized response models for quantitative data. *Journal of the American Statistical Association*, 1976.

[104] W. Kenneth Poole. Estimation of the distribution function of a continuos type random variable through randomized response. *Journal of the American Statistical Association*, 1974.

[105] W. Kenneth Poole. Generalizations of a contamination model for continuous type random variables. *Commun. Statist. Theor. Meth.*, 1982.

[106] D. Raghavarao. On an estimation problem in warner's randomized response technique. *Biometrics*, 1978.

[107] D. Raghavarao and W.T. Federer. Block total response as an alternative to the randomized response method in survey. *Journal of the American Statistical Association*, 1979.

[108] James E. Reinmuth and Michael D. Geurts. The collection of sensitive information using a two-stage, randomized response model. *Journal of Marketing Research*, 1975.

[109] Singh S., R. Singh, N.S. Mangat, and D.S. Tracy. An alternative device for randomized responses. *Statistica*, 1994.

[110] N.J. Scheers and C. Mitchell Dayton. Covariate randomized response models. *Journal of the American Statistical Association*, 1988.

[111] Pranab Kumar Sen. On unbiased estimation for randomized response models. *Journal of the American Statistical Association*, 1974.

[112] Pranab Kumar Sen. On unbiased estimation for randomized response models. *Journal of the American Statistical Association*, 1974.

[113] Pranab Kumar Sen. On unbiased estimation for randomized response models. *Journal of the American Statistical Association*, 1976.

[114] Iris M. Shimizu and Gordon Scott Bonham. Randomized response technique in a national survey. *Journal of the American Statistical Association*, 1978.

[115] Jeffrey S. Simonoff. Randomized response and loss of efficiency. *Internet at www.sern.nyu.edu*, 1997.

[116] Jabir Singh. A note on the randomized response technique. *Journal of the American Statistical Association*, 1976.

[117] Jabir Singh. An additive randomized response model. *Proceedings of the Section on Survey Research Methods of A.S.A.*, 1978.

[118] Jagbir Singh. A note on maxim likelihood estimation from randomized response models. *Proceedings of the Social Statistical Section of A.S.A.*, 1978.

[119] Ravindra Singh, N.S. Mangat, and Sarjinder Singh. A mail survey design for sensitive character without using randomization device. *Commun. Statist. Theory Meth.*, 1993.

[120] Ravindra Singh and Sarjinder Singh. Generalized franklin's model for randomized response sampling. *Commun. Statist. Theory Meth.*, 1993.

[121] Sarjinder Singh, Munir Mahmood, and D.S. Tracy. Estimation of mean and variance of stigmatized quantitative variable using distinct units in randomized response sampling. *Statistical Papers*, 2001.

[122] Sarjinder Singh, Ravindra Singh, and Naurag Singh Mangat. Some alternative strategie to moors' model in randomized response sampling. *Journal of Statistical Planning and Inference*, 2000.

[123] Sarjinder Singh, Ravindra Singh, and Naurag Singh Mangat. Some alternative strategie to moors' model in randomized response sampling. *Journal of Statistical Planning and Inference*, 2001.

[124] Sukhminder Singh. An alternative to warner's randomized response technique. *Statistica*, 1993.

[125] Donald E. Staim and R. Kirk Steinhorst. Telephone interview and mail questionnaire applications of the randomized response model. *Journal of the Statistical Association*, 1984.

[126] Ajit C. Tamhane. Randomized response techniques for multiple sensitive attributes. *Journal of the American Statistical Association*, 1981.

[127] Sabahat Tezcan and Abdel R. Omran. Prevalence and reporting of induced abortion in turkey: Two survey techniques. *Study in Family Planning*, 1981.

[128] N.K. Unnikrishnan. Bayesian analysis for randomized response models. *The Indian Journal of Statistics*, 1999.

[129] Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *J.A.S.A.*, 1965.

[130] Stanley L. Warner. The linear randomized response model. *Journal of the American Statistical Association*, 1971.

[131] Stanley L. Warner. Optimal randomized response models. *Int. Stat. Rev.*, 1976.

[132] Robert L. Winkler and Leroy A. Franklin. Warner's randomized response model: A bayesian approach. *Journal of the American Statistical Association*, 1979.