Rivista di Statistica Applicata, Vol. 17, N. 3 (Settembre) 1984

Un modello R. R. con risposta alternativa fissa dotato di memoria

D. Olivieri e **F. Bressan**, Istituto di Statistica e R.O., Università di Verona

Gli autori propongono un nuovo modello a risposta casualizzata che migliora l'efficienza e semplifica le modalità di rilevazione. La tecnica, che si inquadra nello schema di Simmons, pone una domanda incorrelata a risposta fissa e prevede la determinazione del numero di interpellati sul quesito delicato.

1. Introduzione

La presente nota presenta una dei risultati emersi nel primo convegno di studio sulle risposte casualizzate (R.R.), tenutosi a Verona nel giugno 1982 (*Rivista di Statistica Applicata*, V. 16, N. 1).

Per un eccellente resoconto dello sviluppo delle R.R., vedasi Horvitz *et al.* (1976).

I due aspetti peculiari della nuova proposta, che si innesta nello schema di Simmons, Greenberg *et al.*, e Horvitz *et al.*, per la stima della diffusione di un *carattere delicato* e che supporremo presente in una data popolazione in proporzione π, riguardano:

— l'opportunità di proporre una domanda incorrelata a risposta predeterminata;

— la determinazione del numero di coloro che rispondono al quesito sul carattere delicato.

I vantaggi che comporta l'adozione congiunta di questi due accorgimenti per la stima di π si riflettono sostanzialmente:

— nell'utilizzo di un solo campione rispetto ai due previsti dal modello di Simmons, quando la diffusione dell'evento incorrelato π_y è incognita;

— nella semplificazione delle modalità di rivelazione;

— in una maggior efficienza rispetto al modello di Simmons, pure nell'ipotesi di π_y nota (Greenberg *et al.* 1969, pag. 532) ed anche rispetto allo schema di Warner (1965) almeno nella maggior parte delle situazioni operative.

La procedura pare proponibile quando la manifestazione del carattere delicato può risultare imbarazzante e contemporaneamente la situazione complementare non suscita remora alcuna.

2. La nuova proposta

Lo schema di Simmons risente di alcune difficoltà operative che, sostanzialmente, si riconducono alla necessità di usare due campioni, all'opportunità di scegliere oculatamente i parametri di casualizzazione, ai problemi legati alla ricerca della domanda incorrelata ed ai riflessi indotti sulla efficienza delle stime dalla diversa diffusione che il carattre innocuo ed incorrelato y, alternativo alla domanda delicata, può presentare nella popolazione ($0 < \pi_y < 1$).

La maggior parte di questi problemi discendono direttamente dall'esigenza di operare sull'attributo incorrelato y; d'altra parte è proprio siffatto accorgimento che garantisce ai partecipanti una protezione che deriva dall'impossibilità per chiunque altro di interpretare il significato della risposta. È peraltro evidente che la risposta alla sola domanda delicata porrebbe in imbarazzo esclusivamente coloro che si trovano nella situazione delicata che verrebbe documentata mediante la manifestazione di assenso o, più raramente, di diniego (usualmente è la risposta affermativa che denuncia la situazione imbarazzante).

Partendo da queste premesse supponiamo che la situazione delicata determini, quando la risposta è sincera, il « Sì » (« No » qualora sia la negazione a denunciare la situazione imbarazzante); immaginiamo inoltre di operare su di un campione bernoulliano di n unità proveniente da una popolazione in cui il carattere delicato è presente in proporzione π. La nuova tecnica di rilevazione prevede che ogni soggetto estragga *senza reinserimento* una pallina da un'urna che contiene m palline ($m \geq n$) di cui z colorate.

Se la pallina estratta è colorata l'interpellato risponderà al quesito: « Possiedi il carattere delicato? ». In caso contrario egli risponderà sempre « Sì ». Alla fine dell'estrazione si potrà determinare il numero x di rispondenti alla domanda delicata come differenza tra z ed il numero di palline colorate residue z_1:

$$x = z - z_1.$$

Questa procedura, in estrema sintesi, è riconducibile allo schema di Simmons in cui:

— la diffusione del carattere incorrelato y sia estesa a tutta la popolazione ($\pi_y = 1$);

— si conosca il numero di coloro che hanno risposto alla domanda delicata; ciò discende dall'estrazione esaustiva, essendo nota a priori la composizione dell'urna.

3. Aspetti analitici

Si supponga che il numero di « Sì » ottenuti nelle n prove attuate

secondo le indicazioni esposte in precedenza sia r. Di questi r assensi $(n - x)$ derivano dai soggetti che, avendo trovato la pallina « non colorata », hanno risposto « Sì ». Ne consegue che sulle x persone che hanno estratto la pallina colorata, e quindi risposto alla domanda delicata, si sono avute

$$s = r - (n - x); \quad (r \geq n - x)$$

risposte affermative. In ciascuna di tali unità la probabilità di ottenere la risposta affermativa risulta π, poiché il sottoinsieme degli x soggetti è un campione bernoulliano proveniente da un universo in cui il carattere delicato è presente in proporzione π. Ne consegue che s $(s = 0, 1, 2, \ldots, x)$ definisce una variabile binomiale di parametri x e π. Allora

$$M(s) = x\pi \tag{1}$$

$$Var(s) = x\pi(1 - \pi). \tag{2}$$

Con queste premesse definiamo quale stimatore di π

$$p = \frac{s}{x} = \frac{r - (n - x)}{x} = \frac{f - (1 - \lambda)}{\lambda}$$

con $f = \dfrac{r}{n}$ e $\lambda = \dfrac{x}{n}$. (*)

Lo stimatore appare:

1) corretto. Infatti

$$M(p) = \frac{1}{x} \quad M(s) = \pi;$$

2) di varianza

$$Var(p) = \frac{1}{x^2} \quad Var(s) = \frac{\pi(1 - \pi)}{x} = \frac{\pi(1 - \pi)}{\lambda n};$$

3) consistente risultando, dai punti precedenti

$$\lim_{x \to \infty} P(|p - \pi| < \varepsilon) = 1;$$

4) asintoticamente normale per x crescente, in quanto variabile binomiale in cui il numero di prove x tende all'infinito;

(*) Si noti che λ, di valore medio z/m, non è predeterminato. È possibile peraltro fissarlo a priori qualora si ponga $m = n$ ed allora $x = z$, essendo $z_1 = 0$.

5) di massima verosimiglianza. Infatti la funzione di verosimiglianza delle s risposte affermative, sulle x prove indipendenti, risulta

$$L = \binom{x}{s} \pi'^{s} (1 - \pi')^{s-x}$$

da cui, passando ai logaritmi e derivando

$$\frac{d \log L}{d\pi'} = \frac{s}{\pi'} - \frac{(x - s)}{(1 - \pi')} = 0$$

$$\pi' = \frac{s}{x}.$$

4. Efficienza

4.1 *Simmons e nuovo modello*

Confrontiamo innanzitutto la variabilità delle stime ottenute con il nuovo metodo rispetto alla variabilità del modello di Simmons. Nell'ipotesi più favorevole per quest'ultimo, che prevede allora nota la frazione π_y, la varianza risulta (Greenberg *et al.*, 1969, p. 533)

$$Var\,(p/\pi_y) = \frac{\rho\,(1 - \rho)}{n\lambda^2} \quad \text{con} \quad \rho = \pi\lambda + \pi_y\,(1 - \lambda)$$

Pur operando in questa particolare situazione il nuovo schema presenta sempre una varianza minore delle stime di Simmons. Infatti, indicando con $Var\,(N)$ la varianza dell'estimatore in oggetto, si ha:

$$Var\,(N) < Var\,(p/\pi_y)$$

$$\frac{\pi\,(1 - \pi)}{n\lambda} < \frac{[\lambda\pi + \pi_y\,(1 - \lambda)]\,[1 - \pi\lambda - \pi_y\,(1 - \lambda)]}{n\lambda^2} \qquad (3.1)$$

$$0 < \pi^2\lambda - 2\pi\,\pi_y\lambda + \pi_y\,[1 - \pi_y\,(1 - \lambda)].$$

Questa condizione è soddisfatta se:

$$4\pi_y^2\lambda^2 - 4\lambda\pi_y\,[1 - \pi_y\,(1 - \lambda)] < 0$$

ovvero, per ogni λ, se:

$$\pi_y\,(1 - \pi_y) > 0.$$

Il vincolo risulta sempre verificato per qualsiasi valore del campo di definizione dei parametri che compaiono nella (3.1), fatto salvo il caso di $\pi = \pi_y = 1$. In questa ipotesi, che evidentemente individua una situa-

Tab. I - Rapporto fra varianza schema di Simmons (π_Y noto) e nuovo schema.

λ	π_Y	π 1%	5%	10%	20%	30%	40%	50%	70%	90%
	.1	22.78	5.46	30.33	2.36	2.13	2.14	2.29	3.20	8.31
	.3	56.44	12.24	6.76	4.11	3.33	3.07	3.07	3.87	9.24
.3	.5	76.90	16.27	8.73	5.04	3.91	3.46	3.33	3.91	8.73
	.7	84.16	17.54	9.24	5.16	3.87	3.30	3.07	3.33	6.76
	.9	78.22	16.07	8.31	4.46	3.20	2.60	2.29	2.13	3.33
	.1	10.50	2.92	2.00	1.59	1.52	1.56	1.68	2.29	5.56
	.3	26.46	6.08	3.56	2.34	2.00	1.90	1.92	2.38	5.33
.5	.5	38.38	8.40	4.67	2.84	2.29	2.06	2.00	2.29	4.67
	.7	46.26	9.87	5.33	3.09	2.38	2.06	1.92	2.00	3.56
	.9	50.10	10.50	5.56	3.09	2.29	1.90	1.68	1.52	2.00
	.1	5.14	1.83	1.43	1.26	1.24	1.27	1.35	1.70	3.56
	.3	12.64	3.29	2.13	1.58	1.43	1.39	1.41	1.66	3.20
.7	.5	19.10	4.54	2.72	1.84	1.57	1.46	1.43	1.57	2.72
	.7	24.52	5.56	3.20	2.03	1.66	1.49	1.41	1.43	2.13
	.9	28.90	6.38	3.56	2.16	1.70	1.47	1.35	1.24	1.43
	.1	2.09	1.22	1.11	1.07	1.07	1.08	1.10	1.22	1.82
	.3	4.21	1.62	1.30	1.15	1.11	1.10	1.11	1.19	1.66
.9	.5	6.23	2.01	1.49	1.23	1.15	1.13	1.12	1.15	1.49
	.7	8.17	2.38	1.66	1.30	1.19	1.14	1.11	1.11	1.30
	.9	10.01	2.73	1.82	1.37	1.22	1.15	1.10	1.07	1.11

zione degenere, le stime presentano variabilità nulla quale che sia lo schema adottato.

La Tab. I descrive il rapporto fra le varianze dei due stimatori.

4.2 *Warner e nuovo modello*

La varianza delle stime di π ricavabili con il modello R.R. di Warner, risulta (Greenberg *et al.* 1969, p. 521):

$$Var\,(W) = \frac{\pi\,(1-\pi)}{n} + \frac{\lambda\,(1-\lambda)}{n\,(2\lambda-1)^2}. \qquad \lambda \neq 0{,}5$$

Il nuovo modello appare più efficiente se

$$Var\,(N) < Var\,(W)$$

$$\frac{\pi\,(1-\pi)}{n\lambda} < \frac{\pi\,(1-\pi)}{n} + \frac{\lambda\,(1-\lambda)}{n\,(2\lambda-1)^2}$$

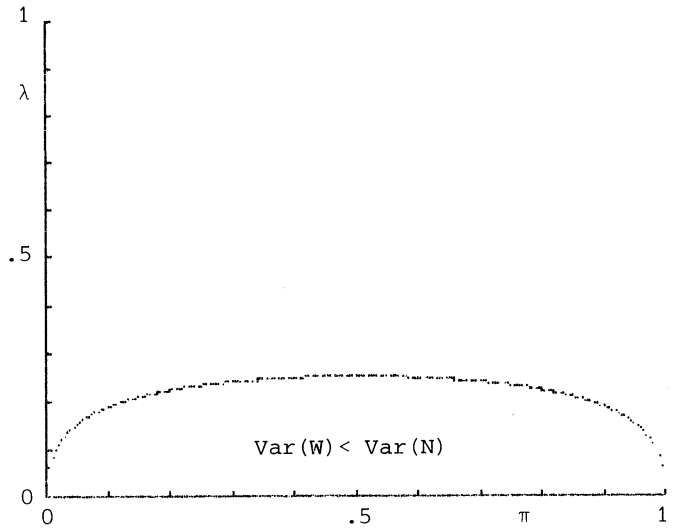

Fig. 1 - Confronto tra varianza dello schema di Warner (Var (W)) e varianza del nuovo schema (Var (N)).

Tab. II - Rapporto tra varianze: schema di Warner e nuovo schema.

λ%	π%								
	1	5	10	20	30	40	50	70	90
10	1.52	.40	.26	.19	.17	.16	0.16	.17	.26
25	19.19	4.20	2.33	1.42	1.14	1.03	1.00	1.14	2.33
40	242.82	50.93	27.07	15.40	11.83	10.40	10	11.83	27.07
55*	1375.55	287.13	151.8	85.63	65.37	57.27	55	65.37	151.8
60	364.24	76.39	40.6	23.10	17.74	15.60	15	17.74	40.6
70	93.50	20.04	10.91	6.44	5.08	4.53	4.38	5.08	10.91
80	36.72	8.28	4.75	3.02	2.49	2.28	2.22	2.49	4.75
90	13.68	3.56	2.31	1.69	1.50	1.43	1.41	1.50	2.31

Per $\lambda = 0{,}5$ le stime di Warner sono indeterminate.

ovvero per

$$\pi(1-\pi) < \left(2 - \frac{1}{\lambda}\right)^{-2}.$$

La disuguaglianza risulta sempre soddisfatta per valori di $\lambda > 0{,}25$ ($\lambda \neq 0{,}5$). Come apparirà dal confronto con lo schema bernoulliano, è evidente che le situazioni operative consigliano di porre il parametro λ

su valori quanto più elevati possibile, compatibilmente con l'opportunità di non insospettire gli intervistati.

La maggior efficienza dello schema di Warner appare dunque una ipotesi teorica relegata a situazioni concrete improponibili ($\lambda \leq 0{,}25$) di fronte all'esigenza di mantenere ridotta la variabilità delle stime. La Fig. 1 e la Tab. II illustrano il rapporto esistente fra le varianze dei due modelli.

4.3 *Confronto con lo schema bernoulliano*

I due confronti precedenti hanno largamente dimostrato il miglioramento nella precisione delle stime conseguibile con la nuova tecnica. Tuttavia per valutare meglio il rapporto tra costi che implica l'adozione del modello R.R. proposto, sembra opportuno paragonarlo con lo schema bernoulliano. In concreto la maggiore variabilità delle stime rappresnta il costo necessario per ottenere informazioni attendibili su questioni delicate, rispetto al consueto onere dei campioni tradizionali.

Detta

$$Var\,(B) = \frac{\pi\,(1 - \pi)}{n}$$

la varianza del piano bernoulliano, il rapporto

$$E = \sqrt{\frac{Var\,(N)}{Var\,(B)}} = \frac{1}{\sqrt{\lambda}}$$

esprime il costo, in termini di errore medio, della casualizzazione. Dalla Tab. III si osserva che la variabilità è tanto minore quanto maggiore è λ; la specificazione di λ sul valore 0,5 — particolarmente opportuno dal punto di vista operativo — comporta una maggiorazione dell'errore medio delle stime R.R. di poco più del 40% rispetto al campionamento non casualizzato. Porre $\lambda = 2/3$, situazione che, secondo la nostra esperienza (Olivieri, 1983), è accettata senza sospetto dagli intervistati, peggiora le stime solamente del 22,5%.

Tab. III - Rapporto fra errore medio del nuovo schema ed errore del piano bernoulliano.

λ	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
E	3,16	2,24	1,83	1,58	1,41	1,29	1,20	1,12	1,05

5. Considerazioni conclusive

Come appare dai paragrafi precedenti i vantaggi che comporta l'ap-

plicazione della nuova tecnica R.R. rispetto allo schema di Simmons riguardano la maggior semplicità di rilevazione associata ad una più elevata efficienza. Per evitare le difficoltà collegate con la scelta del carattere incorrelato, l'attuale proposta prevede che, qualora l'evento aleatorio (il colore della pallina estratta dall'urna) escluda la domanda delicata, la risposta sia in ogni caso « Sì ». Ne consegue che alla risposta « No » sono associati i soggetti che sicuramente non posseggono il carattere imbarazzante. Ciò potrebbe configurare una riduzione della protezione sulla riservatezza delle posizioni di questi ultimi. Tuttavia la critica appare in larga misura inconsistente se consideriamo che sono proprio i soggetti che si trovano nella posizione imbarazzante ad esigere di mantenere la garanzia sull'impossibilità di interpretare il significato della risposta e non gli altri. E tale prerogativa permane anche nel nuovo modello.

L'elemento essenziale della procedura consiste nell'estrazione senza reinserimento; questa tecnica appare perfettamente attuabile quando la rilevazione venga svolta su gruppi di persone che contemporaneamente assistono alla composizione preliminare dell'urna. In alternativa potrà attuarsi qualche accorgimento (una sigillata ecc.) che garantisca anche chi è interpellato individualmente, contro la possibilità di manomissione dell'urna, la cui composizione iniziale sia attestata da persone degne di fede.

Riferimenti bibliografici

Bressan, F. « Possibilità di ampliamento dello schema di Simmons nell'applicazione a gruppi di persone ». *Rivista di Statistica Applicata*, **16**, N. 1, 1983.

Greenberg, B.G., Abul-Ela, A.A., Simmons, W.R. and Horvitz, D.G. « The unrelated question randomized response model: theoretical framework ». *Journal of the American Statistical Association*, **64**, 1969, p. 520-539.

Horvitz, D.G., Shah, B.V., Simmons, W.R. « The unrelated question randomized response model », *1967 Social Statistics Section Proceedings of the American Statistical Association* 65-72.

Horvitz, D.G., Greenberg, B.G. and Albernathy, J.R. « Randomized Response: a Data-Gathering Device for Sensitive Question », *International Statistical Review* **44**, 1976, p. 181-196.

Olivieri, D. « *La diffusione della droga nelle Scuole Secondarie Superiori di Verona* », Cassa di Risparmio di Verona, Vicenza e Belluno, Verona, 1982.

Olivieri, D. « Una modifica allo schema di Simmons », Rivista di Statistica Applicata, **16**, N. 1, 1983.

Warner, S.L. « Randomized response: a survey technique for eliminating evasive answer bias », Journal of the american Statistical Association, **60**, 1965, p. 63-69.

Summary

This paper proposes a new randomized response model which offers improved efficiency and simplifies the randomizing device. Related to Simmons's method, the new technique involves an unrelatd question with a fixed response and allows for the determination of the number of responses to the sensitive question.