Relazione Assegno di Ricerca
Le tecniche di campionamento con risposta casualizzata:
metodi ed applicazioni

Anno Secondo
01/05/2002 – 30/04/2003

Efficienza e protezione della privacy nei Randomized Response Model.

Massimo Guerriero[§]

Summary.
Randomized response is a method used in statistical surveys whose goal is to reduce or eliminate respons errors which occour when respondents are queired about sensitive or highly personal matters. In this work the goals are to analyze and compare the respondent privacy protection in some randomized response models according to efficiency and privacy protection. We prove that some single and double randomization models give no advantage in terms of efficiency under the condition of equal respondent protection compared to Warner's and Simmons' models.

Keywords: *randomized, privacy protection, response models, efficiency, jeopardy.*

1. Introduzione.

Le indagini basate su tecniche a risposta casualizzata (*randomized response*, RR d'ora in poi) sono nate nel secolo scorso, intorno alla prima metà degli anni '60, come risposta all'esigenza di poter raccogliere in modo affidabile informazioni su temi di natura molto delicata e personale (ad es. comportamenti sessuali, evasione fiscale, uso di droghe, etc.), temi sui quali gli intervistati in generale mostrano forte imbarazzo o diffidenza, rifiutando di fornire una risposta o preferendo darne una falsa.

[§] Dipartimento di Economie Società ed Istituzioni – Sezione Statistica – Università degli Studi di Verona – via Dell'Artigliere, 19, 37129 VERONA (e-mail: massimo.guerriero@univr.it).

I modelli RR si basano sostanzialmente sull'idea che, garantendo agli intervistati un'adeguata protezione della privacy, questi saranno più propensi a fornire risposte e a fornirle in modo veritiero. Tale protezione nei modelli RR viene cercata mediante l'uso di un meccanismo di casualizzazione (*random device*) attraverso il quale l'intervistato seleziona in privato la domanda a cui rispondere e comunica poi all'intervistatore solo la propria risposta, impedendo di fatto di essere riconosciuto come appartenente o meno alla classe di coloro che possiedono la caratteristica delicata. Evidentemente l'obiettivo dei piani di campionamento RR è quello di stimare, tipicamente, la frazione[1] di coloro che posseggono una certa caratteristica delicata o, ritenuta tale, in una data popolazione di interesse.

Accanto all'analisi dell'efficienza, la quantificazione del livello di privacy garantita dal modello RR è quindi una caratteristica assolutamente cruciale e necessaria. Uno schema altamente efficiente ma protettivo in misura inadeguata potrebbe infatti non essere in grado di spingere gli intervistati o una parte rilevante di essi a collaborare, vanificando in tal modo il maggior costo e i maggiori sforzi richiesti per predisporre la procedura a risposta casualizzata.

Numerose sono le proposte apparse in letteratura per misurare adeguatamente il livello di privacy. Una breve ma esauriente rassegna è contenuta in Ljungqvist (1993). Tutte queste misure, più o meno direttamente, si fondano sulle probabilità condizionali $P(A|Sì)$ e $P(A|No)$, cioè sulle probabilità di essere percepiti come appartenenti al gruppo A dei soggetti aventi la caratteristica sensibile, dato che è stata fornita, rispettivamente, una risposta affermativa o negativa.

L'obiettivo del presente lavoro è di porre a confronto, mediante qualcuna di queste misure, la protezione della privacy di alcuni dei più interessanti modelli RR a singola e doppia casualizzazione, fra cui rispettivamente quelli di Warner (1965), di Greenberg *et al.* (1969), noto come modello di Simmons, di Mangat (1994) e quelli di Mangat e Singh (1990), di Chang e Liang (1996), congiuntamente al problema della loro efficienza considerata in termini di variabilità degli stimatori.

In particolare si dimostrerà come gli ultimi modelli introdotti altro non siano che casi particolari dei primi due e, ancora di più, come non migliorino l'efficienza degli stimatori e non garantiscano una maggiore protezione come invece gli autori affermano.

E' fondamentale sottolineare che l'analisi che qui condurremo poggia sull'assunto che gli intervistati siano perfettamente razionali e quindi in grado di calcolare in modo corretto le diverse probabilità in gioco e le relative misure di protezione della privacy. Come sottolineato da Ljungqvist

[1] Non si deve trascurare, tuttavia, che molti modelli RR presenti in letteratura sono riferiti ai piani di campionamento per variabili, oltre che per attributi.

(1993), tali misure sono da considerare però solo come un punto di partenza per una futura analisi della protezione della privacy che tenga conto anche dei diversi aspetti appartenenti alla sfera culturale e psicologica degli intervistati. Ad esempio, la difficoltà a comprendere gli aspetti probabilistici del casualizzatore, in particolar modo se quest'ultimo ha una struttura piuttosto complessa, può talora dar vita a reazioni di diffidenza [Landsheer *et al*. (1999)], ad una crescita del timore di essere riconosciuto o di essere erroneamente scambiato per un portatore della modalità delicata e quindi ad un aumento di risposte mancanti o false (*lying*) o date non seguendo le istruzioni del modello RR [il cosiddetto *cheating*, vedi Clark e Desharnais (1998)]. Se è vero quindi che da un punto di vista statistico è importante ottenere stimatori con variabilità minima, non è lecito però al contempo sottovalutare quegli elementi che possono scatenare una reazione potenzialmente avversa dell'intervistato di fronte alla casualizzazione della domanda.

Il lavoro qui presentato sarà così articolato. Nella sezione 2 verranno introdotte le misure di protezione della privacy di Lanke (1976) e quella di Leysieffer e Warner (1976); nella sezione 3 verranno presentati i modelli di riferimento ovvero quello di Warner (1965) e quello di Greenberg *et al.* (1969) anche in termini di misure di protezione della privacy; nella sezione 4 verrà presentato il confronto con un modello a randomizzazione singola [Mangat (1994)] mentre nella sezione 5 verranno proposti i confronti con due modelli a randomizzazione doppia [Mangat e Singh (1990), di Chang e Liang (1996)]; la sezione 6 sarà dedicata alla discussione dei risultati ottenuti.

2. Misure di protezione della privacy

Supponiamo che l'indagine abbia lo scopo di stimare mediante un campione di n unità, la proporzione π_A di soggetti che hanno un dato carattere sensibile (indichiamo con A l'insieme di coloro che nella popolazione hanno il carattere sensibile e con A^c i restanti).
Si possono configurare due casi.
Nel primo, i soggetti appartenenti ad A e quelli appartenenti ad A^c hanno la medesima riluttanza a manifestare il loro carattere e quindi necessitano di una pari protezione della privacy. Questo, ad esempio, accade quando si sta investigando il comportamento nel voto: gli appartenenti ad A sono coloro che intendono votare il partito attualmente al governo e quelli di A^c coloro che non intendono votarlo. In entrambi i gruppi i soggetti vogliono preservare l'anonimato.

Nel secondo caso invece è solo l'appartenenza ad A che vede la riluttanza del soggetto a dare una risposta veritiera, ad esempio perché implica l'aver compiuto un'azione o il possedere una caratteristica socialmente inaccettabile o punibile dalla legge. In tal caso, nessuno di coloro che appartengono ad A^c necessiterebbe di essere protetto. Il meccanismo di randomizzazione può però costringere parte di essi a rispondere "Sì", la qual cosa può negli stessi ingenerare il timore di essere erroneamente scambiati per soggetti appartenenti ad A. Se la protezione offerta dal modello RR non è quindi adeguata, una certa percentuale di soggetti sia di A che di A^c potrebbe essere indotta a contravvenire le regole imposte dal modello e a dare comunque risposta negativa (*risposte evasive*) o addirittura rifiutarsi di rispondere (*mancate risposte*).

2.1 La misura di Lanke

Partendo dall'ipotesi che ci si trovi nel secondo dei casi qui sopra delineati, Lanke (1976) sostiene che "l'imbarazzo" di un intervistato nel fornire una data risposta si possa ragionevolmente ritenere tanto più marcato quanto più elevata è la probabilità condizionata $P(A|R)$ di appartenere ad A, data la risposta R[2], e quindi propone come misura di protezione della privacy, che verrà indicata con il simbolo $MP_L(A)$, la quantità:

$$MP_L(A) = \max\{P(A|Sì), P(A|No)\} \qquad (1)$$

In effetti nella tipica situazione in cui un campione sia suddiviso in due gruppi disgiunti A e A^c, a significare il possesso o meno della modalità delicata, gli intervistati del primo gruppo temeranno di essere individuati, mentre quelli del secondo saranno preoccupati di essere erroneamente scambiati come appartenenti al primo gruppo. Si può quindi ragionevolmente ritenere che l'intervistato, in modo più o meno inconsapevole, utilizzi la regola della probabilità condizionata per decidere se cooperare o meno.

2.2. La misura di Leysieffer e Warner.

Leysieffer e Warner (1976) partono dal presupposto che ogni osservazione possa appartenere solamente o al gruppo A o al suo complementare A^c e che la procedura di intervista produca una certa risposta, R, ai fini di stimare la

[2] La risposta R è di tipo dicotomico nel senso che l'intervistato, dopo aver effettuato l'esperimento casuale, il cui risultato è non noto all'intervistatore, dovrà fornire allo stesso una risposta affermativa ovvero negativa.

ignota proporzione, π_A, degli appartenenti al gruppo *A*. Si dirà quindi che *R* è percepita come rischiosa nei confronti di *A* o A^c rispettivamente se $P(A|R) > \pi_A$ o se $P(A^c|R) > 1 - \pi_A$.

Con queste premesse gli autori propongono le seguenti misure dei livelli di pericolo basate su $P(R|A)$ e $P(R|A^c)$:

$$MP_{LW}(R, A) = \frac{P(R|A)}{P(R|A^c)} \tag{2}$$

per l'esposizione al rischio di essere identificato come appartenente al gruppo *A* se viene fornita la risposta *R* e,

$$MP_{LW}(R, A^c) = \frac{P(R|A^c)}{P(R|A)}, \tag{3}$$

analogamente, per l'esposizione al rischio di essere identificato come appartenente al gruppo A^c se viene fornita la risposta *R*.

I valori di MP_{LW} maggiori dell'unità indicano quelle risposte che aumentano il rischio del rispondente di essere individuato come appartenente al gruppo riportato nell'argomento della funzione MP_{LW}. Questo discende direttamente da come la funzione MP_{LW} è stata ideata dagli autori. In particolare nel primo caso se la funzione assume valori maggiori dell'unità ciò indica una maggiore esposizione al rischio di essere individuato come appartenente al gruppo *A* qualora venga fornita la risposta *R*. Analogamente, nel secondo caso, l'esposizione al rischio viene riferita al gruppo complementare A^c.

Ora, ipotizzando che $P(A \mid Sì) > P(A \mid No)$, $MP_L(A) = P(A \mid Sì)$ e quindi:

$$MP_L(A) = \frac{\pi_A MP_{LW}(Sì, A)}{1 - \pi_A(1 - MP_{LW}(Sì, A))}.$$

Pertanto la relazione che lega le due misure è monotona crescente: per $MP_{LW}=1$ si ha $MP_L=\pi_A$ e per valori via via crescenti di MP_{LW}, MP_L tende a 1.

3. I modelli di Warner e Simmons.

I modelli proposti da Warner (1967) e da Greenberg *et al.* (1969) sono i modelli RR più conosciuti e più diffusamente utilizzati nelle indagini riguardanti i temi delicati o ritenuti tali dagli intervistati. La loro fama è

senza dubbio dovuta al fatto che risultano essere molto semplici e, come vedremo, i non meno efficienti; è per tali ragioni che qui verranno presi come modelli di riferimento.
Il primo, quello di Warner, è così articolato. Ad ogni unità campionaria viene fatto eseguire un esperimento casuale che genera due eventi elementari incompatibili, associati alle seguenti due affermazioni: "Appartengo ad *A*", "Non appartengo ad *A*", con probabilità p_W e $1-p_W$ rispettivamente; in base al risultato ottenuto, ignoto al ricercatore, l'intervistato dovrà fornire a quest'ultimo una risposta di tipo affermativa ovvero negativa. La stima della proporzione degli appartenenti al gruppo delicato *A* sarà quindi, per $p_W \neq \frac{1}{2}$,

$$\hat{\pi}_A^W = \frac{\frac{n'}{n} - (1-p_W)}{2p_W - 1},$$

ove *n'* è il numero di risposte affermative ottenute. La media e la varianza dello stimatore sono rispettivamente pari a:

$$E\left(\hat{\pi}_A^W\right) = \pi_A \quad \text{e,}$$

$$V_W = V\left(\hat{\pi}_A^W\right) = \left[\frac{\lambda_W(1-\lambda_W)}{n(2p_W-1)^2}\right].$$

$\lambda_W = p_W\pi_A + (1-\pi_A)(1-p_W)$ è la probabilità di ottenere una risposta affermativa nella singola prova.
Per il modello di Warner la misura di protezione di Lanke (MP_L) e quella di Leysieffer e Warner (MP_{LW}) possono essere così scritte:

$$MP_L^W(A) = \begin{cases} \dfrac{\pi_A(1-p_W)}{1-\lambda_W} & se \quad p_W < \dfrac{1}{2} \\ \dfrac{\pi_A p_W}{\lambda_W} & se \quad p_W > \dfrac{1}{2} \end{cases} \qquad (4)$$

$$MP_{LW}^{W}\left(Sì, A\right) = MP_{LW}^{W}\left(No, A^{c}\right) = \frac{p_W}{1-p_W} \tag{5}$$

Il modello di Simmons introduce il concetto di carattere in correlato, detto *Y*, al carattere delicato *A*, nel senso che alla domanda circa il possesso della caratteristica delicata viene contrapposta, in luogo della sua negazione, la domanda circa il possesso di un attributo considerato non delicato. Prenderemo qui in esame il caso in cui la diffusione $\left(\pi_Y^S\right)$ del carattere non delicato sia nota a priori e quindi selezionabile opportunamente dal ricercatore[3]. Così, ad ogni unità campionaria viene fatto eseguire un esperimento casuale che genera due eventi elementari incompatibili, associati alle due affermazioni: "Appartengo ad *A*" e "Appartengo ad *Y*", con probabilità p_S e $1-p_S$ rispettivamente; in base al risultato ottenuto, ignoto al ricercatore, l'intervistato dovrà fornire a quest'ultimo una risposta di tipo affermativa ovvero negativa. La stima della proporzione degli appartenenti al gruppo delicato *A* sarà quindi:

$$\hat{\pi}_A^S = \frac{\frac{n'}{n} - \pi_Y^S\left(1-p_S\right)}{p_S}$$

ove *n'* rappresenta il numero di risposte affermative ottenute. La varianza dello stimatore sarà:

$$V_S = V\left(\hat{\pi}_A^S\right) = \frac{\lambda_S\left(1-\lambda_S\right)}{np_S^2}$$

ove $\lambda_s = p_s\pi_A + \left(1-p_s\right)\pi_Y^S$ è la probabilità di ottenere una risposta affermativa.

Per il modello di Simmons le misure di protezione di Lanke (MP_L) e quella di Leysefer e Warner (MP_{LW}) possono essere così scritte:

$$MP_L^S\left(A\right) = \frac{\pi_A\left[\pi_Y^S\left(1-p_S\right)+p_S\right]}{\lambda_S} \tag{6}$$

[3] Il parametro $\left(\pi_Y^S\right)$ è quindi un parametro completamente controllabile dallo sperimentatore.

$$MP_{LW}^{S}\left(Sì, A\right)=\frac{\left[\pi_{Y}^{S}\left(1-p_{S}\right)+p_{S}\right]}{\pi_{Y}^{S}\left(1-p_{S}\right)} \tag{7}$$

$$MP_{LW}^{S}\left(No, A^{c}\right)=\frac{\left[1-\pi_{Y}^{S}\left(1-p_{S}\right)\right]}{\left(1-\pi_{Y}^{S}\right)\left(1-p_{S}\right)} \tag{8}$$

Lanke (1976) e Leysieffer e Warner (1976) confrontano questi modelli sotto la condizione di eguale protezione secondo le due misure introdotte e arrivano alla conclusione che il modello di Simmons è più efficiente di quello di Warner quando $\pi_{Y}^{S}>\frac{1}{2}$, per ogni valore di π_A.

4. Confronto con un modello a randomizzazione singola.

Nel modello di Mangat (1994) si assume di estrarre da una popolazione un campione casuale semplice con reinserimento di *n* unità, ognuna delle quali viene istruita a fornire all'intervistatore risposta affermativa qualora appartenga al gruppo delicato *A*, ovvero di utilizzare un casualizzatore identico a quello proposto da Warner (1965), nel caso contrario. Ciò detto, la probabilità di ottenere una risposta affermativa nella singola prova è, $\lambda_{M}=\pi_{A}+\left(1-\pi_{A}\right)\left(1-p_{M}\right)$ e la stima di massima verosimiglianza di π_A sarà,

$$\hat{\pi}_{A}^{M}=\frac{\frac{n'}{n}-1+p_{M}}{p_{M}}$$

con varianza pari a:

$$V_{M}=V\left(\hat{\pi}_{A}^{M}\right)=\frac{\lambda_{M}\left(1-\lambda_{M}\right)}{np_{M}^{2}}.$$

Le misure di protezione della privacy di Lanke, MP_L, e di Leysieffer e Warner, MP_{LW}, risultano essere pari rispettivamente a:

$$MP_{L}\left(A\right)=\frac{\pi_{A}}{\lambda_{M}} \tag{9}$$

$$MP_{LW}(Sì, A) = \frac{1}{1-p_M} \quad , \quad MP_{LW}(No, A^c) = \frac{p_M}{0} \tag{10}$$

Mangat (1994) dimostra che il presente modello è più efficiente di quello di Warner (1965) se:

$$\pi_A > 1 - \left(\frac{p}{2p-1}\right)^2$$

che risulta essere vera per $p > \frac{1}{3}$.

Questo risultato è vero ma solo limitatamente al caso particolare

$$p_W = p_M = p \tag{11}$$

L'autore non ha infatti effettuato il confronto delle efficienze dei modelli sulla base del criterio di pari protezione arrivando, quindi, a interpretazioni non corrette dei risultati. Confrontando le protezioni dei due modelli trattati nel caso $p_W = p_M = p$ si ricava che il modello di Warner (1965) offre una più elevata protezione rispetto al modello di Mangat (1994), cioè

$$MP_L^W(A) < MP_L^M(A) \tag{12}$$

$$MP_{LW}^W(Sì, A) < MP_{LW}^M(Sì, A) \quad , \quad MP_{LW}^W(No, A^c) < MP_{LW}^M(No, A^c). \tag{13}$$

Utilizzando il criterio di efficienza e protezione, con alcuni passaggi algebrici, si dimostra che il modello di Mangat (1994) è invece equivalente (cioè parimenti efficiente e protettivo) al modello di Simmons quando:

$$p_S = p_M \quad e \quad \pi_Y^S = 1. \tag{14}$$

5. L'illusione della doppia randomizzazione.

5.1. Il modello di Mangat e Singh (1990).

Questo modello introduce l'utilizzo di due casualizzatori forniti ad ognuna delle n unità di un campione casuale estratto con reinserimento dalla popolazione oggetto di studio. L'esperimento casuale associato al primo casualizzatore genera, con probabilità rispettivamente pari a T_{MS} e $1-T_{MS}$, due eventi incompatibili associati alle seguenti affermazioni[4]: "Appartengo al gruppo delicato *A*", "Si passi all'uso del secondo casualizzatore". L'esperimento casuale associato al secondo casualizzatore ed identico a quello proposto da Warner (1965), genera, con probabilità rispettivamente pari a p_{MS} e $1-p_{MS}$, due eventi incompatibili associati alle usuali affermazioni a cui il rispondente dovrà fornire risposta sì ovvero no: "Appartengo al gruppo delicato *A*", "Non appartengo al gruppo delicato *A*".
Con tali premesse la probabilità di ottenere una risposta affermativa nella singola prova sarà quindi $\lambda_{MS}=\pi_A\phi+(1-\pi_A)(1-\phi)$ con $\phi=T_{MS}+(1-T_{MS})p_{MS}$ e uno stimatore corretto per l'ignota proporzione dei soggetti della popolazione appartenenti al gruppo delicato *A*, sarà:

$$\hat{\pi}_A^{MS}=\frac{\frac{n'}{n}-(1-\phi)}{2\phi-1}$$

con varianza pari a:

$$V_{MS}=V\left(\hat{\pi}_A^{MS}\right)=\frac{\lambda_{MS}(1-\lambda_{MS})}{n(2\phi-1)^2}$$

ove *n'* rappresenta il numero di risposte affermative ottenute.
Si introducono ora le misure di protezione proposte da Lanke e da Leysieffer e Warner per il modello in parola.

$$MP_L(A)=\begin{cases}\dfrac{\pi_A(1-\phi)}{1-\lambda_{MS}} & se \quad \phi<\dfrac{1}{2}\\[2ex] \dfrac{\pi_A\phi}{\lambda_{MS}} & se \quad \phi>\dfrac{1}{2}\end{cases} \qquad (15)$$

[4] Al solito, l'esito dell'esperimento casuale è non noto all'intervistatore.

$$MP_{LW}(Sì, A) = MP_{LW}(No, A^c) = \frac{\phi}{1-\phi} \tag{16}$$

Gli autori dimostrano che la varianza dello stimatore di π_A è sempre inferiore a quella del modello di Warner (1965) se:

$$T_{MS} > \frac{1-2p}{1-p}.$$

Tuttavia anche in questo caso il risultato è vero solo sotto la condizione,

$$p_{MS} = p_W = p \tag{17}$$

che è vera per ogni p, scelto opportunamente il parametro T_{MS}. Rimangono pertanto validi i commenti esposti per il modello di Mangat (1994).
Si dimostra invece che, in accordo con il criterio di efficienza e protezione della privacy del rispondente e, per ogni π_A, il modello di Mangat e Singh (1990) è equivalente a quello di Warner posto

$$p_W = \phi. \tag{18}$$

Infatti, sostituendo p_W con $T_{MS} + (1-T_{MS})p_{MS}$, si ottengono le seguenti uguaglianze:

$$P_{MS}(Sì|A) = P_W(Sì|A) \tag{19}$$

$$P_{MS}(Sì|A^c) = P_W(Sì|A^c), \tag{20}$$

ossia nei due modelli la probabilità di ottenere una risposta affermativa condizionatamente al fatto di appartenere al gruppo A o A^c risultano rispettivamente uguali. Da ciò derivano le seguenti uguaglianze:

$$
\begin{aligned}
&\lambda_{MS} = \lambda_{W,} \\
&MP_L^{MS}(A) = MP_L^{W}(A), \\
&MP_{LW}^{MS}(Sì, A) = MP_{LW}^{W}(Sì, A), \\
&MP_{LW}^{MS}(No, A^c) = MP_{LW}^{W}(No, A^c), \\
&\hat{\pi}_A^{MS} = \hat{\pi}_A^{W}, \\
&V_{MS} = V_W .
\end{aligned}
\tag{21}
$$

Pertanto per ogni modello di Mangat e Singh (1990) esiste sempre un modello di Warner con eguale protezione e pari efficienza. Resta dimostrata, quindi, la non superiorità del modello di Mangat-Singh (1990) nei confronti di quello di Warner (1965).

5.2. Il modello di Chang e Liang (1996)

Il presente metodo è analogo a quello proposto da Mangat e Singh (1990), con la differenza che il secondo casualizzatore è identico a quello proposto da Simmons. Ad ognuno degli *n* intervistati quindi vengono consegnati due casualizzatori; l'esperimento casuale associato al primo casualizzatore genera, con probabilità rispettivamente pari a τ_{CL} e $1-\tau_{CL}$, due eventi incompatibili legati, rispettivamente, alle seguenti affermazioni: "Appartengo al gruppo delicato *A*", "Si passi all'uso del secondo casualizzatore". L'esperimento casuale relativo invece al secondo casualizzatore e identico a quello proposto da Simmons, nel caso particolare di diffusione $\left(\pi_Y^{CL}\right)$ nota a priori del carattere incorrelato ad *A*, *Y*, genera, con probabilità rispettivamente pari a p_{CL} e $q_{CL} = 1 - p_{CL}$, due eventi incompatibili legati alle usuali affermazioni a cui il rispondente dovrà fornire risposta sì ovvero no: "Appartengo al gruppo delicato *A*", "Appartengo al gruppo *Y*".
Con tali premesse la stima di π_A sarà data da:

$$
\hat{\pi}_A^{CL} = \frac{\frac{n'}{n} - (1-\tau_{CL}) q_{CL} \pi_Y^{CL}}{p_{CL} + q_{CL}\tau_{CL}}
$$

ove, al solito, *n'* è il numero di risposte affermative ottenute e π_Y^{CL} è la proporzione, nota a priori, di quanti nella popolazione oggetto di studio posseggano l'attributo *Y*.
La varianza dello stimatore è:

$$V_{CL} = V\left(\hat{\pi}_A^{CL}\right) = \frac{\lambda_{CL}\left(1-\lambda_{CL}\right)}{n\left[\tau_{CL} + \left(1-\tau_{CL}\right)p_{CL}\right]^2}$$

ove $\lambda_{CL} = \left[\tau_{CL} + \left(1-\tau_{CL}\right)p_{CL}\right]\pi_A + \left(1-\tau_{CL}\right)q_{CL}\pi_Y^{CL}$ è la probabilità di ottenere una risposta affermativa nella singola prova.
Ora, le misure di protezione di Lanke e Warner e Leysieffer risultano, rispettivamente, pari a:

$$MP_L\left(A\right) = \frac{\pi_A\left[\tau_{CL} + \left(1-\tau_{CL}\right)p_{CL} + \left(1-\tau_{CL}\right)q_{CL}\pi_Y^{CL}\right]}{\lambda_{CL}} \tag{22}$$

$$MP_{LW}\left(S\grave{\imath}, A\right) = \frac{\tau_{CL} + \left(1-\tau_{CL}\right)p_{CL} + \left(1-\tau_{CL}\right)q_{CL}\pi_Y^{CL}}{\left(1-\tau_{CL}\right)q_{CL}\pi_Y^{CL}} \tag{23}$$

$$MP_{LW}\left(S\grave{\imath}, A\right) = \frac{\tau_{CL} + \left(1-\tau_{CL}\right)p_{CL} + \left(1-\tau_{CL}\right)q_{CL}\left(1-\pi_Y^{CL}\right)}{\left(1-\tau_{CL}\right)q_{CL}\left(1-\pi_Y^{CL}\right)} \tag{24}$$

Gli autori dimostrano che il modello proposto è uniformemente più efficiente di quello di Simmons se:

$$\begin{aligned} &q\left[2\left(1-\tau_{CL}\right)p + \tau_{CL}\right]\pi_Y\left(1-\pi_Y\right) + \\ &+ p\left(p + q\tau_{CL}\right)\left[\pi_A\left(1-\pi_Y\right) + \pi_Y\left(1-\pi_A\right)\right] > 0 \end{aligned}$$

Questo confronto è verificato, ancora una volta per:

$$p_{CL} = p_S = p\,. \tag{25}$$

In accordo invece con il criterio di efficienza e protezione e sulla base delle misure di protezione proposte, per ogni π_A, il modello di Chang e Liang (1996) è equivalente a quello di Simmons quando

$$p_S = \tau_{CL} + (1 - \tau_{CL}) p_{CL} \quad e \quad \pi_Y^S = \pi_Y^{CL}. \tag{26}$$

Questo si prova mediante una sostituzione algebrica, analogamente a quanto esposto per il modello di Mangat e Singh (1990).

6. Discussione.

Quanto esposto permettere di tracciare alcune considerazioni.
In circa 30 anni di produzione scientifica sono stati proposti molti modelli alternativi a quelli di Warner (1965) e Simmons (1969) ma nessun risultato è stato raggiunto se non a fronte di complicazioni nell'impianto di indagine, con riguardo al casualizzatore, tali da rendere i nuovi modelli proposti validi solo da un punto di vista metodologico ma con l'effettiva impossibilità di applicarli in indagini reali; infatti si sarebbe esposti sia ad un forte aumento dei costi d'indagine sia ad una elevata aggressività del modello; quest'ultima si rifletterebbe in un alto tasso di non risposta.
I modelli a doppia randomizzazione di Mangat e Singh (1990) e di Chang e Liang (1996), qui presi in considerazione, più complessi e più costosi di quelli di riferimento, non introducono addirittura alcun beneficio, né in relazione alla protezione della privacy, né in relazione alla miglior efficienza. Al più possono risultare parimenti efficienti ma con un costo di impianto più elevato.
Il modello di Mangat (1994), pur essendo a randomizzazione singola, risulta equivalente a quello di Simmons ed inoltre non permette di fissare un limite superiore al rischio legato alla risposta negativa nel gruppo A^c. In alcuni casi ciò potrebbe rappresentare un problema.
Questo lavoro suggerisce quindi che un'adeguata analisi dell'efficienza e della protezione della privacy è sempre necessaria qualora un nuovo modello RR venga proposto.

Bibliografia.

Chang, H-J and Liang, D-H (1996). *A two-stage unrelated randomized response procedure*. Austral. J. Statist., **38**, 43-51.

Clark S.J., Desharnais R.A. (1998). Honest answers to embarassing questions: detecting cheating in the randomized response model. *Psychological Methods*, **3(2)**, 160-168.

Greenberg, B. G., Abul-Ela, A. A., Simmons, W.R. and Horvitz, D.G. (1969). *The unrelated question randomized response model: theorical frameworks*. J. Am. Statist. Ass., **64**, 520-539.

Landsheer J.A., van der Heijden P., van Gils G. (1999), Trust and understanding, two psychological aspects of randomized response. A study of a method for improving the estimate of social security fraud. *Quality & Quantity,* **33**, 1-12.

Lanke, J. (1976). *On the degree of protection in randomized interviews*. Int. Stat. Rev., **44**, 197-203.

Leysieffer, F.W. and Warner, S.L. (1976). *Respondent jeopardy and optimal designs in randomized response modes*. J. Am. Statist. Ass., **71**, 649-656.

Ljungqvist, L. (1993). *A unified approach to measures of privacy in randomized response models: a utilitarian perspective*. J. Am. Statist. Ass., **88**, 97-103.

Mangat, N. S. (1994). *An improved randomized response strategy*. J.R. Statist. Soc. B, **56**, 93-95.

Mangat, N. S. and Ravindra Singh. (1990). *An alternative randomized response procedure.* Biometrika, **77**, 439-442.

Warner, S. L. (1965). *Randomized response: a survey technique for eliminating evasive answer bias*. J. Am. Statist. Ass., **60**, 63-69.