Università degli Studi di Padova

Facoltà di Scienze Statistiche

Corso di Laurea in Scienze Statistiche
Demografiche e Sociali

Tesi di Laurea

Metodi a risposte casualizzate per dati quantitativi

Relatore: Ch. mo Prof. Giancarlo Diana

Laureanda: Marlies Ranieri

Anno Accademico 2007-2008

a mia nonna Gretel e a Mario

Indice

Introduzione

Le tecniche a risposte casualizzate (Randomized Response Technique, RRT) nascono negli anni Sessanta per essere applicate in indagini volte a rilevare informazioni personali, imbarazzanti e delicate come, per esempio, l'assunzione di droghe, l'evasione fiscale, la frequenza di aborti o comportamenti sessuali. In tali circostanze gli intervistati possono rifiutarsi di rispondere, manifestando diffidenza, o fornire deliberatamente una risposta non veritiera: risulterebbe pertanto svantaggioso, se non inutile, svolgere un'intervista diretta, che porterebbe ad esiti altamente distorti.

Per eliminare, o almeno ridurre, la frequenza di risposte non vere, sono state sviluppate delle tecniche per acquisire l'informazione sensibile preservando la riservatezza degli intervistati. Studi empirici hanno dimostrato che le tecniche a risposte casualizzate permettono di ottenere una sostanziale riduzione della distorsione dello stimatore; in particolare si è visto che i risultati conseguiti sono tanto più affidabili, quanto più è sensibile la variabile oggetto d'indagine. Sfortunatamente, il punto critico consiste nella minore resa in termini di efficienza rispetto all'osservazione diretta, che costringe ad utilizzare campioni di dimensione maggiore per ottenere stime precise.

Le prime applicazioni si propongono di stimare la proporzione di individui appartenenti ad un gruppo associato a comportamenti socialmente non tollerati, senza violare la privacy dei rispondenti. Grazie ad un meccanismo di casualizzazione, come per esempio un dado o un mazzo di carte, l'intervistato seleziona in privato la domanda a cui rispondere, e comunica all'intervistatore solo la propria risposta, impedendo di essere riconosciuto come appartenente, o meno, al gruppo di coloro che possiedono la caratteristica sensibile.

Fino ad oggi, i modelli casualizzati sono stati principalmente impiegati in indagini che richiedono una risposta dicotomica (“si” o “no”) alla domanda sensibile, oppure una scelta tra un insieme di categorie nominali. Tuttavia, in letteratura le procedure sono state estese anche al caso in cui le risposte siano di tipo quantitativo, con l’obiettivo di stimare media e varianza della variabile sensibile. In tal modo il ricercatore non si limita a rilevare, ad esempio, se una donna ha subito o meno un aborto, ma quanti aborti questa ha avuto; oppure non solo se il soggetto assume o meno droghe, ma il numero di occasioni in cui ne fa generalmente uso.

In questa tesi tratteremo esclusivamente le strategie per rilevare variabili sensibili di tipo quantitativo, dal momento che lo studio di variabili dicotomiche, o multinomiali, è già stato abbondantemente approfondito da vari autori.

I modelli apparsi in letteratura per rilevare dati sensibili quantitativi sono numerosi, e soggetti a continue generalizzazioni, col fine di migliorare da un lato l’efficienza degli stimatori, e dall’altro, la protezione della privacy dei rispondenti. Nel seguito si descrivono dettagliatamente le principali strategie proposte, comprese alcune di recente formulazione, tutte basate sulla medesima logica di ottenere come risposta, non la variabile sensibile, ma una sua opportuna trasformazione, ricavata introducendo una componente di codifica.

Nuovi sviluppi di ricerca nell’ambito delle tecniche casualizzate sono indirizzati verso l’utilizzo di una variabile ausiliaria correlata al carattere sensibile, al fine di migliorare l’efficienza degli stimatori. Nei piani di campionamento tradizionali si ricorre spesso a stimatori costruiti in base al metodo del rapporto, o a quello della regressione, al fine di aumentare la precisione della stima del parametro ignoto. Nonostante le numerose tecniche sviluppate per stimare parametri non sensibili basandosi su una o più variabili ausiliarie, pochi tentativi sono stati attuati nel campo delle tecniche casualizzate, e questi ultimi sono quasi esclusivamente

limitati allo studio di dati qualitatitivi. Recentemente, è stata proposta in letteratura una classe di stimatori della proporzione di persone con il carattere sensibile, utilizzando una variabile ausiliaria di media nota. Nello studio di dati sensibili quantitativi, la possibilità di introdurre una variabile ausiliaria non è stata ancora valutata in modo approfondito, e ha ricevuto meno attenzione rispetto a temi quali, la stratificazione e il campionamento a probabilità variabili.
Un altro aspetto cruciale delle procedure casualizzate è la protezione della privacy garantita ai rispondenti. È prevedibile infatti che, tanto più gli intervistati si sentiranno tutelati, tanto più coopereranno fornendo una risposta veritiera. Generalmente in letteratura, i modelli casualizzati sono confrontati rapportando la varianza degli stimatori al fine di valutarne il rendimento in termini di efficienza relativa. Questo tipo di raffronto però non risulta affatto completo, in quanto trascura il livello di protezione della privacy assicurato ai rispondenti e, conseguentemente, la possibilità di applicare i modelli a casi reali.
Come sostenuto da alcuni autori, per una valido confronto dovrebbero essere presi in considerazione sia l'efficienza che il livello di privacy.
D'altra parte, la protezione della privacy e l'efficienza assumono andamenti opposti: più lo stimatore è efficiente, tanto meno riservatezza viene garantita ai rispondenti, e viceversa. Il modello casualizzato ottimale dovrebbe armonizzare questi aspetti, tentando di raggiungere un giusto compromesso che consenta di ottenere uno stimatore adeguatamente preciso tutelando, al tempo stesso, la riservatezza.

Partendo da questi concetti iniziali abbiamo articolato la tesi nel seguente modo.
Il Capitolo 1 presenta una rassegna dei principali modelli casualizzati per dati quantitativi proposti in letteratura. Viene descritta la struttura degli stimatori più comuni, evidenziando i risultati salienti che torneranno utili nel corso del lavoro.

Il Capitolo 2 ha lo scopo di introdurre una classe generale di stimatori che prevede la possibilità di utilizzare una variabile ausiliaria. La classe permette di comprendere una molteplicità di strategie includendo, come casi particolari, i modelli casualizzati senza l'informazione ausiliaria. Inoltre, si indaga analiticamente il guadagno di efficienza aquisito grazie all'uso di uno stimatore per regressione.
Il Capitolo 3 è dedicato ai confronti in termini di efficienza e di rispetto della privacy tra alcuni casi speciali della classe. Si introduce qui, una misura di protezione della privacy generalizzata, volta a valutare il grado di riservatezza quando si utilizza una variabile ausiliaria.
Nel Capitolo 4, alla luce dei risultati relativi ai confronti, si è tentato di rispondere all'esigenza di trovare un ragionevole compromesso tra efficienza e tutela della privacy, proponendo tre modelli misti. Se ne esamina il grado di flessibilità, illustrando infine il criterio di applicazione di uno dei modelli introdotti.

In appendice si completa la selezione di grafici per confrontare l'efficienza e la protezione della privacy dei casi speciali nel Capitolo 3, e delle tecniche miste nel Capitolo 4. Inoltre, si riportano alcuni grafici ottenuti con ulteriori distribuzioni delle variabili di codifica.

Notazioni ed assunzioni

Consideriamo una popolazione finita formata da N individui. Si assume di estrarre un campione casuale semplice con reinserimento *CCSCR*, di dimensione n, al fine di rilevare uno o più caratteri quantitativi. Si considera un *CCSCR*, in quanto questa, è la procedura di selezione privilegiata per le tecniche a risposte casualizzate.

- Sia $Y \geq 0$ la variabile sensibile oggetto di studio, di media e varianza ignote.

Nel presente lavoro, si è interessati alla stima della media di Y. A tale scopo, si dispone del campione di n risposte, così indicato

$$(z_1, z_2, ..., z_n)$$

- Z rappresenta la variabile risposta, e contiene tutta l'informazione ottenibile su Y dal campione. Questa è esprimibile come una combinazione tra Y ed un'altra variabile, U oppure W, o anche entrambe

$$Z = g(Y, U, W)$$

  dove $g(\ )$ è una funzione opportuna.
- $U > 0$ e $W > 0$ sono possibili variabili di codifica non sensibili, di distribuzione nota, ed indipendenti da Y e tra loro.

Su ciascuno degli N individui che compongono la popolazione, si manifestano altri caratteri quantitativi, uno dei quali può essere utilizzato come variabile ausiliaria, eventualmente codificata.

- Sia $X \geq 0$ la variabile ausiliaria di media nota, correlata ad Y ed indipendente da U e da W.

In questo caso il campione è costituito dalle coppie

$$((z_1, v_1), (z_2, v_2), ..., (z_n, v_n))$$

- V contiene tutta l'informazione ausiliaria su X. Questa potrà coincidere con X, qualora non se ne preveda la codifica; altrimenti sarà una combinazione tra X ed un'altra variabile di codifica, H oppure T, o anche entrambe. In quest'ultimo caso

$$V = g(X, H, T)$$

dove $g(\ )$ è definita come in precedenza.

- Premettiamo inoltre, che le procedure discusse nella tesi poggiano sull'assunto fondamentale, comune a tutta la letteratura specifica, che gli intervistati forniscano risposte veritiere.

Nel corso della tesi faremo ricorrente uso dei termini “modello casualizzato” e “strategia casualizzata”.

- L’accezione “modello casualizzato” è strettamente legata alla particolare funzione $g(\ )$, che lega le variabili tra loro.
- L’espressione “strategia casualizzata”, intende invece indicare l’abbinamento del modello con il rispettivo stimatore della media della variabile sensibile.
- Talvolta definiremo un modello più efficiente rispetto ad un altro, intendendo così esprimere la maggiore efficienza dello stimatore ad esso associato.

Denoteremo in genere, con le lettere maiuscole, le caratteristiche della popolazione, e con le lettere minuscole, le corrispondenti quantità nel campione. Le assunzioni descritte di seguito per la variabile Z sono valide per ciascuna delle variabili che abbiamo prima descritto.
Si indica rispettivamente valore atteso, varianza e coefficiente di variazione di Z con

$$\mu_Z = E(Z) = \frac{1}{N}\sum_{i=1}^{N} z_i \qquad \sigma_Z^2 = V(Z) = \frac{1}{N}\sum_{i=1}^{N}(z_i - \mu_Z)^2 \qquad C_Z = \frac{\sigma_Z}{\mu_Z}$$

La covarianza ed il coefficiente di correlazione tra le variabili casuali Z e X sono denotati rispettivamente con

$$\sigma_{XZ} = Cov(X, Z) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_X)(z_i - \mu_Z) \qquad \rho_{XZ} = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z}$$

Le corrispondenti quantità del campione sono

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i \qquad s_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 \qquad c_Z = \frac{s_Z}{\bar{z}}$$

$$s_{ZY} = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y}) \qquad r_{XZ} = \frac{s_{XZ}}{s_X s_Z}$$

Capitolo 1

Modelli casualizzati per dati quantitativi

1.1 Introduzione

In questo capitolo presentiamo i modelli più rilevanti proposti fino ad oggi in letteratura. Le procedure casualizzate per dati quantitativi sono soggette a continui sviluppi, al fine di conseguire sempre maggiori guadagni in termini di efficienza degli stimatori, possibilmente a parità di protezione della privacy. L'obiettivo perseguito è la stima della media della variabile sensibile, preservando, al meglio, l'anonimato dei rispondenti. Le tecniche che descriveremo si basano tutte sulla stessa logica: si intende evitare l'intervista diretta, proteggendo il carattere sensibile mediante l'introduzione di una variabile di codifica, la cui distribuzione è nota al ricercatore. Chiaramente, per ogni modello, la scelta di tale distribuzione non è affatto banale, in quanto da essa dipende l'efficienza dello stimatore.

In letteratura si trovano diversi suggerimenti per massimizzare l'efficienza di ciascuna procedura rispetto alle altre. È facile intuire però, come confronti analitici in termini di efficienza risultino alquanto difficoltosi a causa, non solo dei numerosi parametri coinvolti, ma anche della molteplicità di situazioni nelle

quali si potrebbe trovare il ricercatore. Quest'ultima considerazione costringe spesso a ricorrere a confronti numerici, sicuramente più agevoli, nel tentativo di rispecchiare il rendimento delle varie strategie nei casi reali.
Di seguito illustriamo i principali modelli, ideati da vari autori, ponendo l'attenzione ai risultati che torneranno utili nel seguito del lavoro.

1.2 Il modello a domande incorrelate

La prima tecnica formulata per rilevare una variabile sensibile quantitativa è basata sul modello a domande incorrelate. Questa semplice idea rappresenta un'estensione diretta di un modello proposto per stimare la frequenza di un attributo nella popolazione ("Unrelated Questions").
Ciascun intervistato seleziona, grazie ad un meccanismo di casualizzazione, una tra due domande, di cui una sensibile e l'altra innocua. Con probabilità p l'intervistato risponde fornendo il suo valore del carattere delicato, Y, mentre con probabilità $1-p$, risponde alla domanda neutra, fornendo un valore della variabile W. La risposta osservata è dunque

$$Z = \begin{cases} Y \quad \text{con probabilità} \quad p \\ \\ W \quad \text{con probabilità } 1-p \end{cases} \tag{1.1}$$

L'intervistatore, ignaro dell'esito della selezione, dispone del campione casuale ($z_1, \ldots, z_n$). È quindi possibile stimare la media ignota della variabile sensibile, μ_Y, con

$$\hat{\mu}_G = \frac{\bar{z} - (1-p)\mu_W}{p} \tag{1.2}$$

Si nota immediatamente che lo stimatore $\hat{\mu}_G$ è corretto e la sua varianza risulta

$$V(\hat{\mu}_G) = \frac{1}{n}\left[\sigma_Y^2 + \frac{(1-p)}{p}\left(\sigma_Y^2 + (\mu_Y - \mu_W)^2 + \frac{\sigma_W^2}{p}\right)\right] \tag{1.3}$$

La distribuzione della variabile riposta Z è un miscuglio di due distribuzioni pure, non codificate, ed indipendenti tra loro. Il trucco consiste semplicemente nell'ottenere la stima di μ_Y, separando statisticamente i valori di Y dai valori di W.

Affinchè gli intervistati si sentano adeguatamente protetti, è necessario che la distribuzione di W non sia troppo dissimile da quella di Y, altrimenti si potrebbe identificare la domanda a cui si risponde. È di fondamentale importanza inoltre, che i valori delle risposte alla domanda innocua e alla domanda sensibile siano della stessa unità di misura, pertanto il ricercatore dovrebbe scegliere una distribuzione di W che abbia lo stesso supporto della variabile sensibile.

Vantaggiosa per l'efficienza dello stimatore $\hat{\mu}_G$, è la prossimità tra la media di W e la media di Y, si vede infatti dalla (1.3) che la precisione aumenta se il termine $|\mu_Y - \mu_W|$ tende a zero.

Se la variabilità di W fosse considerevolmente più piccola di quella di Y, ci potrebbe essere un'avversione nel cooperare da parte dei rispondenti, che non si sentirebbero sufficientemente protetti. Perciò è auspicabile che σ_W^2 sia almeno tanto grande quanto σ_Y^2. Chiaramente, sia media che varianza della caratteristica sensibile sono ignote, e diviene quindi problematico scegliere i parametri di W in maniera del tutto appropriata.

Il modello presenta due principali svantaggi dal punto di vista della protezione della privacy: il primo è che gli intervistati potrebbero assumere valori di W molto simili o regolari, rendendo possibile associare la risposta alla domanda. Il secondo è che gli intervistati potrebbero manifestare reticenza nel rivelare il valore della variabile sensibile, non essendo magari pienamente convinti che l'intervistatore sia incapace di identificare le singole risposte.

1.3 I modelli a risposte casualizzate codificate

Un primo miglioramento rispetto al modello a domande incorrelate è costituito dai metodi a risposte casualizzate codificate ("Scrambled Randomized Response Method"). Tali tecniche si propongono di mascherare i dati sensibili, grazie ad una codifica indotta dalla somma o dalla moltiplicazione per un numero casuale, generato da una distribuzione nota. Tra le due strategie, è stato il modello moltiplicativo a destare maggiore interesse in letteratura, venendo successivamente generalizzato e riproposto con alcune varianti.

Al contrario del modello a domande incorrelate, in cui si osserva Y con probabilità p, con questi metodi la variabile sensibile non è mai osservata direttamente, perché sempre codificata, favorendo la percezione di tutela della privacy. Un aspetto pratico da tenere presente consiste nella generazione dei

numeri casuali senza suscitare sospetti negli intervistati. A questo scopo, può essere programmato un calcolatore che simuli una sequenza indefinita di numeri casuali da una distribuzione prescelta.

L'efficienza dei metodi a risposte codificate rispetto al modello a domande incorrelate è stata studiata in letteratura a livello analitico; tuttavia, i confronti attuati, presentando notevoli semplificazioni, non assumono carattere generale.

1.3.1 Il modello additivo

Un primo semplice modo di combinare la variabile sensibile Y con la codifica U, consiste nel sommarne i valori. L'i-esimo rispondente seleziona un valore u_i, lo somma al suo valore y_i, e comunica successivamente all'intervistatore il risultato dell'operazione. Pertanto si ottiene la risposta codificata

$$Z = Y + U \tag{1.4}$$

Uno stimatore non distorto della media della variabile sensibile è

$$\hat{\mu}_{AD} = \bar{z} - \mu_U \tag{1.5}$$

la cui varianza è data da

$$V(\hat{\mu}_{AD}) = \frac{\sigma_Y^2 + \sigma_U^2}{n} \tag{1.6}$$

Il principale vantaggio del modello additivo è che la (1.6) dipende solamente da σ_Y^2 e σ_U^2 , mentre, sia nel modello a domande incorrelate, che nel modello moltiplicativo, che presenteremo di seguito, compaiono nell'espressione della varianza di stima μ_Y e μ_U **.** In letteratura non si trovano suggerimenti in merito alla distribuzione che dovrebbe assumere U , al fine di massimizzare l'efficienza del modello. Rispetto ad altre tecniche, la scelta della variabile di codifica additiva è meno vincolante dato che la precisione di $\hat{\mu}_{AD}$ dipende unicamente da parametri di scala.

Il principale svantaggio del metodo consiste nel fatto che valori anomali di Y potrebbero non essere adeguatamente protetti dalla somma con un numero casuale. Ciò lascia intuire che la tecnica additiva potrebbe non tutelare a sufficienza la privacy, dal momento che i valori estremi necessitano di più riservatezza.

1.3.2 Il modello moltiplicativo

Il modello moltiplicativo assume particolare rilevanza, ed è stato riformulato da vari autori al fine di migliorarne l'efficienza. La procedura prevede che ciascun rispondente selezioni un numero casuale e lo moltiplichi per il suo valore sensibile, comunicando all'intervistatore il risultato dell'operazione. Pertanto, la risposta codificata è

$$Z = WY \tag{1.7}$$

Uno stimatore corretto di μ_Y è

$$\hat{\mu}_{EH} = \frac{\bar{z}}{\mu_W} \tag{1.8}$$

con varianza

$$V(\hat{\mu}_{EH}) = \frac{1}{n}\left[\sigma_Y^2 + \mu_Y^2\left(1 + C_Y^2\right)C_W^2\right] \qquad (1.9)$$

Gli autori suggeriscono di scegliere la distribuzione di W considerando due obbiettivi in conflitto tra loro, uno atto ad incrementare la percezione della privacy, l'altro l'efficienza dello stimatore. Per far sì che la frequenza di risposte non veritiere resti esigua, W dovrebbe coprire un vasto range di valori con probabilità alta. D'altro canto, per stimare μ_Y in modo accurato, si dovrebbe - considerando la (1.9) - mantenere il coefficiente di variazione C_W , il più piccolo possibile. Infine, una caratteristica ragionevole per la distribuzione di W è assumere la mediana eguale ad 1: una distribuzione continua che rispetta tali condizioni è, ad esempio, una F di Fisher con (5,5) gradi di libertà.

È stato attuato, in letteratura, un confronto numerico tra il modello moltiplicativo ed il modello a domande incorrelate, al fine di valutare il rendimento in termini di efficienza: è risultato che, per opportune scelte dei parametri, la strategia moltiplicativa fornisce uno stimatore più preciso di quella a domande incorrelate.

Si descrivono di seguito delle possibili generalizzazioni dei metodi a risposte codificate presenti in letteratura:

- il modello misto di Saha
- il modello di Bar-Lev
- il modello a risposte casualizzate opzionali
- il modello a risposte casualizzate forzate

1.4 Il modello misto di Saha

Il modello misto costituisce una versione più recente dei metodi a risposte codificate. L'idea di creare un modello misto è nata in contemporanea con tali metodi, ma ha trovato compiuta formulazione solo attualmente, grazie a Saha. L'idea consiste nel combinare il modello additivo ed il moltiplicativo, per sfruttare i punti di forza di entrambi relativamente all'efficienza e alla tutela della privacy. Il procedimento si effettua nel seguente modo: l'i-esimo rispondente seleziona un numero casuale u_i e somma il numero estratto con il valore della variabile sensibile, y_i. Di seguito, estrae un altro numero casuale, w_i, che moltiplica per la somma, precedentemente calcolata, $(y_i + u_i)$. L'intervistatore, ovviamente ignaro dei numeri selezionati, u_i e w_i, conosce tuttavia le distribuzioni delle variabili che li hanno generati, U e W. La risposta osservata è

$$Z = W(Y + U) \tag{1.10}$$

è possibile quindi stimare correttamente μ_Y con

$$\hat{\mu}_{SA} = \frac{\bar{z}}{\mu_W} - \mu_U \tag{1.11}$$

la cui varianza è

$$V(\hat{\mu}_{SA}) = \frac{1}{n\mu_W^2}\left\{\sigma_Y^2\mu_W^2\left(C_W^2 + 1\right) + \sigma_W^2\left[\mu_Y^2 + \mu_U^2\left(1 + C_U^2\right) + 2\mu_Y\mu_U\right] + \sigma_U^2\mu_W^2\right\} \tag{1.12}$$

L'utilizzo del modello misto è consigliabile quando il carattere oggetto di indagine è particolarmente sensibile, in quanto la percezione della privacy dei rispondenti è agevolata dalla presenza di una doppia codifica.

Questa considerazione trova giustificazione a livello analitico, giacché la strategia di Saha fornisce uno stimatore meno preciso, rispetto ad entrambi i metodi ad una codifica. Si verifica infatti, che lo stimatore associato al modello additivo (1.5) è più efficiente di $\hat{\mu}_{SA}$, essendo

$$V(\hat{\mu}_{AD}) = V(\hat{\mu}_{SA}) - \frac{1}{n}\left\{C_W^2\left[\mu_Y^2\left(1+C_Y^2\right)+\mu_U^2\left(1+C_U^2\right)+2\mu_U\mu_Y\right]\right\} \qquad (1.13)$$

Si osservi poi, che il modello moltiplicativo (1.7) fornisce uno stimatore almeno tanto efficiente di $\hat{\mu}_{SA}$, dal momento che

$$V(\hat{\mu}_{EH}) = V(\hat{\mu}_{SA}) - \frac{1}{n}\left\{C_W^2\left[\mu_U^2\left(1+C_U^2\right)+2\mu_U\mu_Y\right]+\sigma_U^2\right\} \qquad (1.14)$$

Tuttavia, anche se la variabile sensibile è molto delicata, è sempre necessario valutare l'utilità di una procedura così elaborata: è indispensabile soppesare con cautela, se l'uso di due variabili di codifica comporti un effettivo vantaggio, o se costituisca solo un'inutile complicazione del disegno.

1.5 Il modello di Bar-Lev

Il modello di Bar-Lev generalizza il modello moltiplicativo mediante l'introduzione di un parametro di disegno, $p \in (0,1)$, controllato dal ricercatore per casualizzare le risposte. Grazie ad un meccanismo di casualizzazione, l'i-esimo rispondente fornisce il valore y_i, con probabilità p, mentre comunica il valore codificato $w_i y_i$, con probabilità $1-p$. Lo schema pertanto è

$$Z = \begin{cases} Y & \text{con probabilità } p \\ \\ WY & \text{con probabilità } 1-p \end{cases} \qquad (1.15)$$

Si noti che, per p = 0, la procedura appena descritta si riduce al modello moltiplicativo mentre, per p =1, si ottiene l'intervista diretta.

Uno stimatore non distorto di μ_Y è

$$\hat{\mu}_{BL} = \frac{\bar{z}}{p + (1-p)\mu_W} \qquad (1.16)$$

la cui varianza può essere espressa come

$$V(\hat{\mu}_{BL}) = \frac{1}{n}\left[\sigma_Y^2 + \mu_Y^2 (1 + C_Y^2) C_W^*(p)\right] \qquad (1.17)$$

dove

$$C_W^*(p) = \frac{p + E(W^2)(1-p)}{[p + \mu_W (1-p)]^2} - 1 \qquad (1.18)$$

È possibile dimostrare che, se la distribuzione della variabile di codifica W soddisfa la condizione

$$0 < \mu_W < \frac{2E(W^2)}{[1 + E(W^2)]} \qquad (1.19)$$

allora

$$\frac{\sigma_Y^2}{n} < V(\hat{\mu}_{BL}) < V(\hat{\mu}_{EH}), \qquad \forall \quad p \in (0,1)$$

Il principale vantaggio del modello di Bar-Lev si riscontra nella maggiore efficienza rispetto al moltiplicativo, se viene soddisfatta la (1.19). Una scelta appropriata della distribuzione di W che soddisfi la condizione suddetta, è un'esponenziale di media $\mu_W = 1/\lambda_W$, dove $2-\sqrt{2} < \lambda_W < 2+\sqrt{2}$.

Come il modello a domande incorrelate, la procedura presenta il rischio che l'intervistato possa indugiare nel rispondere, dovendo talvolta rivelare il suo valore della variabile sensibile, cosa che, ovviamente, non accade adottando il moltiplicativo (1.7).

1.6 Il modello a risposte casualizzate opzionali

Una procedura sempre più efficiente del modello moltiplicativo è la tecnica a risposte casualizzate opzionali, proposta due anni prima del modello di Bar-Lev. La tecnica nasce dalla constatazione che alcuni intervistati potrebbero non manifestare alcuna reticenza nel rivelare il loro valore della variabile sensibile e dunque, non avrebbero bisogno di alcuna codifica.

Si prevede che ogni intervistato, utilizzando ad esempio un calcolatore, generi un valore dalla distribuzione di W. Di seguito, ciascuno decide autonomamente se riportare il valore di Y, oppure la risposta codificata WY. La risposta casualizzata opzionale è perciò

$$Z = W^T Y \tag{1.20}$$

dove T è una variabile dicotomica che assume valore 1 se la risposta è codificata, valore 0 altrimenti.

Se denotiamo con π la probabilità che l'intervistato scelga di codificare il suo valore sensibile, allora T è una variabile casuale di Bernoulli con $E(T)=\pi$. La probabilità π è detta "livello di sensibilità della domanda". Se una domanda è particolarmente imbarazzante, allora più persone decideranno di riportare la risposta codificata, ed il valore di π sarà alto. Al contrario, se la domanda non è notevolmente sensibile, allora il valore di π sarà basso.

Osserviamo che la tecnica a risposte opzionali può essere interpretata come un caso particolare del modello di Bar-Lev, in cui si assume $\mu_W = 1$ e $\pi = 1 - p$.

La sostanziale differenza tra le due procedure consiste nel fatto che, mentre nel modello (1.15), p è un parametro definito dal ricercatore, nella tecnica a risposte opzionali, π è ignoto, e va conseguentemente stimato, dipendendo unicamente dalla libera decisione degli intervistati di usare o meno la codifica.

Uno stimatore distorto di π è dato da

$$\hat{\pi} = \frac{\frac{1}{n}\sum_{i=1}^{n} \log(z_i) - \log(\bar{z})}{E(\log(W))} \qquad (1.21)$$

1.7 Il modello a risposte casualizzate forzate

Una tecnica di formulazione recente è il modello a risposte casualizzate forzate che rappresenta una generalizzazione del modello di Bar-Lev.

All'i-esimo rispondente si richiede, mediante un meccanismo di casualizzazione, di riportare il valore y_i con probabilità p_1, il valore $w_i y_i$ con probabilità p_2, mentre, con probabilità p_3, il valore F scelto dal ricercatore. La risposta osservata è

$$Z = \begin{cases} Y & \text{con probabilità } p_1 \\ WY & \text{con probabilità } p_2 \\ F & \text{con probabilità } p_3 \end{cases} \qquad (1.22)$$

Dove p_1 , p_2 , p_3 sono parametri di disegno, fissati dal ricercatore, che sottostanno al vincolo $p_1 + p_2 + p_3 = 1$. Il modello moltiplicativo si ottiene come caso particolare ponendo $p_1 = p_3 = 0$ e $p_2 = 1$, e il modello di Bar-Lev con $p_1 = p$, $p_2 = 1 - p$ e $p_3 = 0$.

Uno stimatore non distorto di μ_Y è

$$\hat{\mu}_{GS} = \frac{\bar{z} - p_3 F}{(p_1 + p_2 \mu_W)} \qquad (1.23)$$

con varianza

$$V(\hat{\mu}_{GS}) = \frac{1}{n(p_1 + p_2\mu_Y)^2}\left[\left(p_1 + p_2\left(\sigma_Y^2 + \mu_Y^2\right) - (p_1 + p_2\mu_Y)^2\right)\left(\sigma_W^2 + \mu_W^2\right)\right. \qquad (1.24)$$

$$\left. + p_3(1 - p_3)F^2 - 2p_3F(p_1 + p_2\mu_Y)\mu_W\right] + \frac{\sigma_W^2}{n}$$

Si noti che lo stimatore $\hat{\mu}_{GS}$ dipende da F. È possibile definire il valore ottimo di F tale da minimizzare la (1.24): il valore dipende però, a sua volta, dalla media ignota del carattere sensibile, in particolare

$$F_{opt} = \frac{(p_1 + p_2\mu_W)\mu_Y}{(1 - p_3)}$$

pertanto F_{opt} non può essere utilizzato nelle situazioni ricorrenti in pratica.

La varianza minima di $\hat{\mu}_{GS}$, raggiunta in corrispondenza di F_{opt}, è

$$V(\hat{\mu}_{GS})_{\min} = \frac{\mu_Y^2}{n}\left\{\frac{\left(1+C_Y^2\right)\left[p_1 + p_2\mu_W^2\left(1+C_W^2\right)\right]}{\left(p_1 + p_2\mu_W\right)^2} - 1 - \frac{p_3}{1-p_3}\right\} \qquad (1.25)$$

È stato inoltre dimostrato per via numerica che, per opportune scelte dei parametri, il modello (1.22) è più efficiente del modello di Bar-Lev.

1.8 Nota bibliografica

In questo capitolo abbiamo presentato i modelli più rilevanti per la rilevazione di una variabile sensibile quantitativa.

In letteratura possiamo fare riferimento alle pubblicazioni di Greenberg et al. (1971) ed Eriksson (1973), in cui viene proposta l'estensione del modello a domande incorrelate per risposte categoriali, al caso in cui la risposte siano quantitative. In particolare, nella pubblicazione di Greenberg si trova anche una formulazione del modello a domande incorrelate che prevede l'estrazione di due campioni casuali di dimensione n_1 ed n_2, indipendenti e non sovrapposti: l'impiego di due campioni, invece che di uno, si rende necessario, qualora la media della variabile di codifica non sia nota al ricercatore e deve essere stimata. In questo lavoro non trattiamo tale situazione, e supponiamo di disporre sempre di un solo campione.

Per quanto riguarda i metodi a risposte casualizzate codificate, si rimanda alla pubblicazione di Warner (1971), in cui viene trattato un modello lineare generale

che include tali metodi come casi speciali; Warner menziona qui per la prima volta l'eventualità di un modello a codifica moltiplicativa, senza però approfondirne lo studio. Una breve trattazione relativa ai modelli a risposte casualizzate codificate, si trova in una pubblicazione di Pollock e Bek (1976), in cui si tentano i primi confronti analitici di efficienza dei modelli additivo e moltiplicativo, rispetto al modello a domande incorrelate. Lo studio del modello a codifica moltiplicativa viene approfondito per la prima volta in un lavoro di Eichhorn e Hayre (1983), i quali trattano in maniera estesa il problema della scelta della variabile di codifica, fornendo i primi spunti per i modelli misti a combinazione additivo-moltiplicativa. In questo contesto viene proposta una prima semplice formulazione di modello misto, $Z = WY + U$. Inoltre, si trova qui un confronto numerico sull'efficienza tra modello moltiplicativo e modello a domande incorrelate, il quale indica la superiorità del moltiplicativo.

Il modello misto descritto nel testo è stato proposto in un recente articolo da Saha (2007), con l'obiettivo principale di individuare un metodo applicabile, sia per dati quantitativi, che qualitativi.

Il modello moltiplicativo di Eichhorn e Hayre è stato generalizzato da Bar-Lev et al. (2004). Il modello a risposte opzionali è invece pubblicato in un lavoro di Gupta, Gupta e Singh (2002), successivamente ripreso da Huang (2007), il quale ha formulato l'estensione della procedura, al caso in cui μ_W sia diverso da 1.

Di sviluppo recente sono, infine, i modelli a risposte casualizzate forzate, introdotti per la prima volta da Gjestvang e Singh (2005), e generalizzati ulteriormente in una pubblicazione di Odumade e Singh (2007). Nel primo lavoro del 2005 è possibile trovare i valori dei parametri, tali da garantire la maggiore efficienza del modello a risposte forzate rispetto al modello di Bar-Lev; inoltre, in questa stessa pubblicazione, si applica la procedura a risposte forzate anche a due campioni indipendenti.

Capitolo 2

Una classe generale di stimatori

2.1 Introduzione

I numerosi modelli proposti in letteratura si basano sostanzialmente sull'idea di rilevare una variabile risposta data dalla combinazione tra la variabile sensibile e la componente di codifica. Può esser utile individuare uno schema generale che sia in grado di rappresentare le principali soluzioni. A questo scopo, si introduce un modello generale di codifica che prevede l'eventuale utilizzo di una variabile ausiliaria, la quale può essere a sua volta codificata.

In letteratura, l'opportunità di usare l'informazione ausiliaria nei modelli casualizzati è stata finora valutata solo marginalmente in merito a disegni di campionamento diversi dal *CCSCR*, in particolare nel campionamento stratificato e a probabilità variabili. Per ovviare a tale carenza, in questo capitolo si utilizza uno stimatore basato su una variabile supplementare, applicando un *CCSCR*, e se ne evidenziano i vantaggi.

Privilegiamo l'informazione ausiliaria che non sia delicata, tale da apparire il più possibile neutrale agli occhi dei rispondenti rispetto al carattere sensibile. Inoltre, è desiderabile che questa sia di agevole reperimento, in modo da evitare costi

aggiuntivi all'indagine. Ci orientiamo pertanto verso l'uso di variabili ausiliarie poco, o moderatamente, correlate con il carattere sensibile. Le variabili che si possono considerare a questo fine sono frequenti in molte aree sociali, cliniche e mediche. Persino dati provenienti da fonti amministrative, se collegati all'oggetto di indagine, risultano validi, e presentano un interessante vantaggio: essendo di dominio pubblico, si possono osservare direttamente senza compromettere la privacy degli intervistati. Per di più, i rispondenti potrebbero non essere informati della rilevazione congiunta dei dati pubblici, evitando così l'insorgere di sospetti.

Nel presente capitolo, la definizione di una classe generale di stimatori consente di mostrare analiticamente il vantaggio di utilizzare uno stimatore per regressione, che combini linearmente l'informazione derivante dalle risposte osservate e l'informazione ausiliaria. Vedremo come gli stimatori che prevedono una variabile ausiliaria sono almeno tanto efficienti quanto gli stimatori che non la considerano, ed il guadagno in efficienza aumenta, al crescere della correlazione con la variabile sensibile.

2.2 Un modello generale di codifica

Si presenta un modello generale che esprime in maniera compatta le principali strategie casualizzate proposte in letteratura, ed altre ancora che potrebbero venire formulate in futuro. A questo scopo, si introduce una classe generale di stimatori che prevede la possibilità di usare dell'informazione ausiliaria correlata alla variabile sensibile, eventualmente codificata. Al variare dei parametri della classe si ottengono, come casi particolari, la maggior parte degli stimatori descritti nel Capitolo 1.

Si assuma di usare un qualsiasi meccanismo di casualizzazione, e sia $X > 0$, la variabile ausiliaria non sensibile di media nota, la quale è correlata positivamente, o negativamente, con la variabile sensibile Y.

Con l'obiettivo di esprimere genericamente i vari modelli, introduciamo due variabili dummy, S ed R, al variare delle quali si ottiene qualsiasi combinazione per la risposta osservata, Z, e per l'informazione ausiliaria osservata, V. Per pervenire alla stima di μ_Y, si richiede ad ogni intervistato di realizzare un processo bernuolliano con probabilità di successo p. Se si ottiene un successo, l'intervistato fornisce i suoi valori sia di Y che di X. Alternativamente, questi comunica i valori di S e di R. Naturalmente il ricercatore non è in grado di associare i valori forniti alle variabili selezionate, rimanendo all'oscuro dell'esito del processo bernoulliano.

$$Z = \begin{cases} Y & \text{con probabilità } p \\ S & \text{con probabilità } 1-p \end{cases} \qquad V = \begin{cases} X & \text{con probabilità } p \\ R & \text{con probabilità } 1-p \end{cases} \tag{2.1}$$

Disponendo del campione casuale ((z_1, v_1), (z_2, v_2), ..., (z_n, v_n)) di n risposte, costruiamo la classe di stimatori

$$\hat{\mu}_S = \frac{\bar{z}_d - c}{h} \qquad (h \neq 0) \tag{2.2}$$

dove $\bar{z}_d$ è lo stimatore per regressione di μ_Z basato sulla variabile ausiliaria V

$$\bar{z}_d = \bar{z} - b(\bar{v} - \mu_V) \tag{2.3}$$

Le costanti reali c, h e b vengono opportunamente scelte: c ed h dipendono esclusivamente dal tipo di casualizzazione effettuata, mentre b è legato principalmente all'uso della variabile ausiliaria. Per $b=0$ si rappresentano le strategie che non prevedono la variabile V, e conseguentemente la stima di μ_Z è la semplice media campionaria, ovvero

$$\hat{\mu}_M = \frac{\bar{z} - c}{h} \tag{2.4}$$

Pertanto, la classe (2.2) ha il pregio di includere, per differenti modelli casualizzati, sia gli stimatori usuali senza l'informazione supplementare, sia gli stimatori costruiti grazie ad essa.

Nella seguente tabella si riportano le espressioni per S, c ed h che specificano alcune tecniche casualizzate illustrate nel capitolo precedente, per le quali $b=0$.

Tabella 2.1 Modelli casualizzati

$\hat{\mu}_s$	p	S	c	h
Intervista diretta $\hat{\mu}_0$	1		0	1
A domande incorrelate $\hat{\mu}_G$	(0,1)	W	$(1-p)\mu_W$	p
Mod. Additivo $\hat{\mu}_{AD}$	0	$Y+U$	μ_U	1
Mod. Moltiplicativo $\hat{\mu}_{EH}$	0	WY	0	μ_W
Modello Misto di Saha $\hat{\mu}_{SA}$	0	$W(Y+U)$	$\mu_W\mu_U$	μ_W
Modello di Bar – Lev $\hat{\mu}_{BL}$	(0,1)	WY	0	$p+(1-p)\mu_W$

2.3 La stima con variabile ausiliaria

Esistono varie tecniche per stimare μ_Z, sfruttando la conoscenza dei valori assunti dalla variabile V. Si può pensare di utilizzare i classici stimatori per rapporto, prodotto o per regressione. Mediante un'appropriata specificazione di b nella (2.3), siamo in grado di ottenere ciascuno di questi stimatori, ed in particolare per $b = 0$, come già evidenziato, vengono inclusi gli stimatori senza l'uso di informazione ausiliaria.

La scelta di adottare uno stimatore per regressione poggia sull'intento di stimare μ_Z mediante uno stimatore, $\bar{z}_d$, che combini linearmente le medie campionarie $\bar{z}$ e $\bar{v}$. Se b è noto, lo stimatore per regressione di μ_Z risulta corretto ed è possibile stimarne la varianza senza distorsione. Come è risaputo, in letteratura, lo stimatore per regressione è il migliore per la stima di una media ignota di una popolazione, quando ci si avvale di una o più variabili ausiliarie.

Se b è noto, e c ed h soddisfano il vincolo $\mu_Z = c + h\mu_Y$, allora $\hat{\mu}_S$ è corretto con varianza data da

$$V(\hat{\mu}_S) = \frac{V(\bar{z}_d)}{h^2} = \frac{1}{nh^2}\left(\sigma_Z^2 - 2b\sigma_{VZ} + b^2\sigma_V^2\right) \tag{2.5}$$

che può essere stimata senza distorsione mediante

$$\hat{v}(\hat{\mu}_S) = \frac{1}{nh^2}\left(s_Z^2 - 2bs_{VZ} + b^2 s_V^2\right) \tag{2.6}$$

La varianza espressa in (2.5) dipende dalla costante b. Tra le possibili opzioni, scegliamo b come soluzione ottima del problema di minimizzazione della varianza, ovvero

$$b_0 = \arg\min_b V(\hat{\mu}_S)$$

Si verifica facilmente che il $V(\hat{\mu}_S)$ raggiunge il minimo in corrispondenza di

$$b_0 = \frac{\sigma_{VZ}}{\sigma_V^2} \tag{2.7}$$

il quale rappresenta il coefficiente di regressione lineare di Z su V. Si osservi, in particolare, che se $V = X$ abbiamo che $b_0 = h\beta_{XY}$.

Sostituendo b_0 nella (2.5), si ottiene il migliore stimatore della classe, $\hat{\mu}_{S,opt}$, il quale diventa uno stimatore per regressione. Essendo $\hat{\mu}_S$ una trasformazione lineare di $\bar{z}_d$, non sono conseguibili ulteriori miglioramenti, almeno fino al primo ordine di approssimazione, se a parità di informazione ausiliaria, si utilizza invece di $\hat{\mu}_{S,opt}$, uno stimatore non lineare di μ_Z.

In corrispondenza dello stimatore ottimo, si raggiunge il limite inferiore per la varianza della classe, dato da

$$V(\hat{\mu}_S)_{\min} = \frac{\sigma_Z^2}{nh^2}\left(1-\rho_{VZ}^2\right) \tag{2.8}$$

Uno stimatore consistente di $V(\hat{\mu}_S)_{\min}$ è

$$\hat{v}(\hat{\mu}_S)_{\min} = \frac{s_Z^2}{nh^2}\left(1 - r_{VZ}^2\right) \tag{2.9}$$

Considerando la struttura della varianza (2.8) vediamo che questa è il prodotto di due componenti: la prima, $\sigma_Z^2/\left(nh^2\right)$ è la varianza dello stimatore che non contempla l'utilizzo di X (i.e. $b=0$), mentre la seconda, $1-\rho_{VZ}^2$ rappresenta il guadagno in efficienza acquisito grazie all'uso della variabile ausiliaria. Pertanto l'efficienza dello stimatore $\hat{\mu}_S$ è funzione del grado di correlazione tra le variabili osservate, Z e V. Riscrivendo i termini della (2.8) abbiamo che

$$V(\hat{\mu}_S)_{\min} - V(\hat{\mu}_M) = -\frac{\sigma_Z^2 \rho_{ZV}^2}{nh^2} \tag{2.10}$$

Indipendentemente dal modello casualizzato adottato (per ogni h e c), l'utilizzo della variabile ausiliaria X fornisce uno stimatore per regressione che, nel caso ottimo, è almeno tanto efficiente del corrispondente stimatore costruito senza l'informazione supplementare, e si ha che il guadagno in efficienza aumenta, al crescere della correlazione tra la variabile ausiliaria e la variabile sensibile.

$$V(\hat{\mu}_M) \geq V(\hat{\mu}_S)_{\min} \tag{2.11}$$

Gli stimatori sono equivalenti solo nel caso in cui le variabili Z ed V siano incorrelate.

2.3.1 Il caso b_0 ignoto

Quando b_0 non è noto a priori, solitamente il suo valore può essere scelto considerando indagini pilota o esperienze precedenti. Comunque sia, anche utilizzando una stima affidabile per b_0, lo stimatore di μ_Z risulta distorto e la varianza dello stimatore non raggiungerà il limite inferiore nella (2.8). Se ci sono buone ragioni per credere che la scelta non sia attendibile, allora una stima può essere calcolata dai dati campionari. Una stima consistente per b_0 è la stima ai minimi quadrati

$$\tilde{b}_0 = \frac{s_{ZV}}{s_V^2} \qquad (2.12)$$

che può essere sempre quantificata, dal momento che (z_i, v_i) è la risposta dell'i-esimo rispondente del campione. Si noti che tale formulazione evita l'uso della variabile sensibile, i cui valori non possono essere raccolti direttamente.
Al primo ordine di approssimazione, lo stimatore ottenuto sostituendo $\tilde{b}_0$ al posto di b_0 ignoto, coincide con lo stimatore $\hat{\mu}_{S,opt}$ in cui b_0 è noto, e la sua varianza, a meno di termini di ordine superiore a $1/n$, è data dalla (2.8).

2.4 Nota bibliografica

In questo capitolo si è vista una classe generale di stimatori che prevede l'utilizzo di una variabile ausiliaria. In letteratura, l'attenzione dedicata all'impiego di una variabile supplementare è stata sinora limitata, e circoscritta quasi

esclusivamente allo studio di dati sensibili qualitativi. Di pubblicazione recente, è il lavoro di Diana e Perri (2007) in cui si presenta una classe generale di stimatori per la proporzione di persone con la caratteristica sensibile.

Relativamente allo studio di dati quantitativi, possiamo considerare la pubblicazione relativa alla stratificazione ottimale di Mahajan e Singh (2005). In questo articolo si utilizzano stimatori per rapporto e per regressione, combinati e separati, basando la rilevazione su un modello casualizzato moltiplicativo.

Un ulteriore contesto in cui è stata considerata l'informazione supplementare è il campionamento in due fasi: al riguardo si può reperire l'articolo di Grewal, Bansal e Singh (2003).

Capitolo 3

Efficienza e protezione della privacy

3.1 Introduzione

L’obiettivo principale dei modelli casualizzati consiste nel garantire un’adeguata protezione della privacy agli intervistati preservando, per quanto possibile, l’efficienza degli stimatori del parametro di interesse. Si può ragionevolmente supporre che, assicurando ai rispondenti un’appropriata riservatezza, questi saranno propensi a fornire risposte veritiere.

Per poter confrontare adeguatamente diverse procedure casualizzate, bisogna pertanto valutare, congiuntamente all’efficienza degli stimatori, il livello di privacy garantito; ovvero, è necessario confrontare le varianze di stima delle differenti strategie, a parità del grado di riservatezza.

Privacy ed efficienza sono due aspetti cruciali per la valutazione di un modello casualizzato: uno schema altamente efficiente, ma non sufficientemente protettivo, potrebbe non essere in grado di indurre gli intervistati a collaborare; al contrario, uno schema troppo protettivo, non fornirebbe una stima adeguatamente precisa. Intuitivamente, appare prevedibile che ogni tentativo di migliorare l’efficienza, ridurrà inevitabilmente la protezione della privacy degli intervistati, e viceversa.

È necessario che il ricercatore valuti bene il grado di precisione dello stimatore che vuole conseguire, in relazione alla delicatezza della variabile oggetto di

studio. Se l'argomento è particolarmente imbarazzante, una contenuta perdita in efficienza può essere tollerata, al fine di utilizzare degli stimatori che altrimenti risulterebbero distorti, e quindi inefficaci.

Questo capitolo ha lo scopo di valutare il rendimento di alcune procedure casualizzate, in termini di efficienza e di rispetto della privacy. Le strategie confrontate prevedono la possibilità di utilizzare una variabile ausiliaria, con o senza codifica. Al fine di valutare il livello di riservatezza si propone una misura di protezione della privacy generalizzata. Tale misura permette di confrontare sia i modelli basati su una variabile ausiliaria, eventualmente codificata, sia le usuali procedure senza l'informazione supplementare.

3.2 Misure di privacy

Il grado di riservatezza garantibile deve essere stabilito considerando aspetti di carattere pratico ed etico dell'indagine, ed effettuando valutazioni in merito alla precisione degli stimatori. È necessario, pertanto, quantificare adeguatamente la protezione della privacy mediante degli indici, che sintetizzino il livello di tutela del rispondente.

In letteratura troviamo varie proposte, tutte basate sull'idea generale che tanto più precisamente può essere stimato μ_Y, tanto meno protetto sarà l'intervistato. Numerosi autori hanno tentato di formulare delle misure del grado di protezione considerando varianze condizionate, intervalli di confidenza o indici di entropia. Alcune di queste sono basate sul modello (*model-based*), dipendendo interamente dal tipo di distribuzione della variabile sensibile, altre invece sono basate sul disegno (*design-based*).

Una prima misura basata su indici di variabilità, è stata proposta per un modello additivo, ma può essere applicata a qualsiasi procedura. L'indice è il seguente

$$R = \sqrt{\frac{\sigma_W}{\sigma_Y}} \tag{3.1}$$

dove R è la radice del rapporto tra la deviazione standard della variabile di codifica e la deviazione standard della variabile sensibile. A valori piccoli di R, corrispondono livelli relativamente bassi di protezione. Al crescere di R, viene garantita più riservatezza dalla codifica, ma peggiora la qualità dei dati raccolti. Per R =1, la varianza della variabile di codifica è uguale alla varianza di popolazione della variabile sensibile mentre, per R =0, si ottiene l'intervista diretta.

Un'altra misura di privacy costruita con delle varianze è

$$\widehat{R} = \frac{\sigma_Z^2}{\sigma_Y^2} \tag{3.2}$$

che rappresenta il rapporto tra la variabilità della risposta, rispetto a quella della caratteristica sensibile: tanto più $\widehat{R}$ è maggiore di 1, tanto minore sarà il rischio di violare la privacy. Naturalmente, al crescere del rumore causato dalla codifica, aumenterà conseguentemente $\widehat{R}$.

Per quantificare la riservatezza assicurata, si può utilizzare anche la varianza condizionata $V(Y \mid Z)$. Tanto più questa è grande, tanto più alto è il grado di protezione garantito. Si osservi che tale misura (*model-based*) richiede, non solo la specificazione della distribuzione di Y, ma anche di $Y \mid Z$.

Altri tentativi di valutare il livello di privacy si avvalgono di intervalli di confidenza. In particolare, per il modello moltiplicativo è stata proposta la seguente misura di protezione: sia [a,b] un intervallo di confidenza, per un'assegnata variabile di codifica W a livello $(1-\alpha)\%$. Allora, se $Z = WY$ è la variabile risposta, si ha che un intervallo di confidenza per Y a livello $(1-\alpha)\%$ è dato da $[\,Z/a, Z/b\,]$. Il rapporto b/a, insieme alla probabilità α, fornisce una misura della protezione garantita ai rispondenti; per un dato α, tanto più è grande il rapporto, tanto più alto è il grado di protezione. Ad esempio, assumendo che $W \sim F(5,5)$, si considerano adeguate le protezioni relative ai casi

$$\alpha = 0{,}1 \qquad b/a \approx 25\,; \qquad \alpha = 0{,}2 \qquad b/a \approx 11{,}9\,; \qquad \alpha = 0{,}5 \qquad b/a \approx 3{,}6$$

Sussistono anche misure basate sull'ampiezza dell'intervallo di confidenza della media del carattere sensibile. Sia $[l_1, l_2]$ l'intervallo di confidenza di μ_Y a livello $(1-\alpha)\%$, allora l'ampiezza $l_1 - l_2$ definisce l'ammontare di privacy a livello $(1-\alpha)\%$. Tuttavia tale metodo risulta semplicistico, e spesso non adeguato.

Richiamando la teoria dell'informazione di Shannon**,** sono state introdotte alcune misure (*model-based*), per quantificare il livello di protezione grazie all'entropia. Sostanzialmente, per misurare la tutela della privacy si utilizza il concetto di mutua informazione tra dato originale y_i e dato codificato z_i. L'ammontare medio di informazione proveniente dalla variabile sensibile non casualizzata Y, dipende dalla sua distribuzione, e può essere misurata dall'entropia differenziale

$$h(Y) = E(-\log_2 f_Y(y)) = -\int_{S_Y} f_Y(y)\log_2 f_Y(y)dy$$

L’ammontare medio di informazione che rimane in Y, dopo che è stata osservata la variabile risposta Z, può essere misurato dall’entropia differenziale condizionata

$$h(Y \mid Z) = E\left(-\log_2 f_{Y|Z=z}(y)\right) = -\int_{S_{YZ}} f_{YZ}(y,z)\log_2 f_{Y|Z=z}(y)dydz$$

Conseguentemente, la perdita media di informazione di Y che avviene rilevando Z può essere misurata mediante la differenza tra le due entropie

$$I(Y \mid Z) = h(Y) - h(Y \mid Z) = E\left(\log_2 \frac{f_{Y|Z=z}(y)}{f_Y(y)}\right)$$

Tale quantità è anche detta “mutua informazione” tra le variabili Y e Z. Sono state formulate, inoltre, le seguenti misure per valutare, rispettivamente, il livello di protezione ed il livello di perdita della privacy

$$\Pi(Y) = 2^{h(Y)} \qquad\qquad P(Y \mid Z) = 1 - 2^{-I(Y;Z)}.$$

3.3 Una misura di protezione della privacy generalizzata

Si presenta ora una misura di protezione della privacy che prevede la possibilità di utilizzare una variabile ausiliaria. La formulazione di una nuova misura è indispensabile in questo lavoro, in quanto, come più volte espresso, lo scopo precipuo è di valutare il vantaggio ottenibile utilizzando l'informazione supplementare. Abbiamo quindi generalizzato la misura (3.2), al fine di costruire uno strumento idoneo all'obiettivo.

Si vuole quantificare il livello di protezione garantito, in funzione della correlazione tra la variabile sensibile Y e le variabili osservate, V e Z.

In riferimento al modello (2.1) con variabile ausiliaria, introduciamo la seguente misura del livello di protezione

$$\tau = 1 - \rho_{Y.VZ}^2 = 1 - \frac{\rho_{YV}^2 + \rho_{YZ}^2 - 2\rho_{YV}\rho_{YZ}\rho_{VZ}}{1 - \rho_{VZ}^2} \tag{3.3}$$

dove $\rho_{Y.VZ}$ è il coefficiente di correlazione multiplo di Y con V e Z.

La misura proposta è ovviamente normalizzata, in quanto τ può variare nell'intervallo $[0,1]$. Ci si attende che all'aumentare di τ, si alzi il livello di protezione della privacy, e pertanto i rispondenti siano più propensi a cooperare.

Per $\tau = 1$, si assicura al rispondente la massima riservatezza mentre, all'estremo opposto, per $\tau = 0$, si ottiene l'intervista diretta.

Se non si utilizza alcuna informazione ausiliaria, la misura (3.3) si riduce a $\tau = 1 - \rho_{YZ}^2$, e tale indice può essere espresso come

$$\tau = 1 - \frac{h^2 \sigma_Y^2}{\sigma_Z^2} \tag{3.4}$$

In questa notazione, si evidenzia l'analogia con la (3.2).

Mentre la (3.2) rapporta la variabilità della risposta Z a quella di Y, la (3.4) è il complemento a 1 della varianza dell'intervista diretta, rispetto alla varianza dello stimatore della classe (2.2).

Inoltre, la (3.4) traduce la discordanza tra privacy ed efficienza, dove quest'ultima, è intesa in termini di variabilità dello stimatore. In un'intervista diretta, caratterizzata da efficienza massima, si ha $\sigma_Y^2 = \sigma_Z^2$, $h = 1$ e risulta $\tau = 0$; all'opposto, un'intervista che assicura un alto grado di riservatezza presenta $\sigma_Z^2 >> \sigma_Y^2$, e di conseguenza l'indice tende ad 1.

La misura esprime l'idea che qualsiasi accortezza atta ad incrementare la precisione dello stimatore, implica una perdita in termini di protezione della privacy; vale a dire che un alto grado di riservatezza comporta la rinuncia ad un'efficienza elevata, e viceversa.

Consideriamo, a titolo di esempio, il modello a domande incorrelate, il modello moltiplicativo ed il modello di Bar-Lev, discussi nel Capitolo 1. Se valgono le condizioni

$$0 < \mu_W < \frac{2E\left(W^2\right)}{\left[1 + E\left(W^2\right)\right]} \qquad C_W^* > \frac{1-p}{pE\left(Y^2\right)}\left[E\left(Y^2\right) + E\left(W^2\right) - 2\mu_Y \mu_W + \frac{1-p}{p}\sigma_W^2\right]$$

allora, l'ordinamento degli stimatori in termini di efficienza è

$$\hat{\mu}_G \succ \hat{\mu}_{BL} \succ \hat{\mu}_{EH}$$

mentre in termini di protezione della privacy, misurata con τ, l'ordine è completamente invertito.

3.4 Casi Speciali

Riportiamo di seguito una tabella che presenta alcuni esempi illustrativi relativi a differenti specificazioni del modello (2.1). I casi speciali possono essere facilmente ottenuti precisando il significato delle variabili dummy, R ed S, ed i valori del parametro di disegno p. Una volta attuate le scelte di R, S e p, le costanti c ed h sono sistematicamente definite.

Tabella 3.1 *Modelli casualizzati*

	R	S	p	c	h
Intervista diretta			1	0	1
Modello 1	T	W	$(0,1]$	$(1-p)\mu_W$	p
Modello 2	X	W	$(0,1]$	$(1-p)\mu_W$	p
Modello 3	X	$Y+U$	$[0,1]$	$(1-p)\mu_U$	1
Modello 4	X	WY	$[0,1]$	0	$p+(1-p)\mu_W$
Modello 5	$X+H$	$Y+U$	$[0,1]$	$(1-p)\mu_U$	1
Modello 6	TX	WY	$[0,1]$	0	$p+(1-p)\mu_W$

Al fine di confrontare le 6 procedure casualizzate, sia in termini di efficienza, che in termini di protezione della privacy, conviene definire preliminarmente le quantità necessarie per calcolare la varianza dello stimatore di μ_Y e la misura di protezione della privacy τ. Tali quantità possono essere espresse in forma generale, riferendosi alla classe di stimatori (2.2).

$$\mu_Z = p\mu_Y + (1-p)\mu_S \qquad \mu_V = p\mu_X + (1-p)\mu_R \tag{3.5}$$

$$\sigma_Z^2 = pE(Y^2) + (1-p)E(S^2) - \mu_Z^2 \qquad \sigma_V^2 = pE(X^2) + (1-p)E(R^2) - \mu_V^2 \tag{3.6}$$

$$\begin{aligned} \sigma_{VZ} &= pE(XY) + (1-p)E(RS) - \mu_Z\mu_V \\ &= p\sigma_{XY} + (1-p)\sigma_{RS} + p(1-p)(\mu_Y - \mu_S)(\mu_X - \mu_R) \end{aligned} \tag{3.7}$$

$$\sigma_{YV} = p\sigma_{XY} + (1-p)\sigma_{YR} \qquad \sigma_{YZ} = p\sigma_Y^2 + (1-p)\sigma_{YS} \tag{3.8}$$

Naturalmente queste devono poi essere particolarizzate a seconda del modello, come fatto di seguito.

Modello 1

Sia la variabile sensibile, che la variabile ausiliaria, si riconducono alla procedura a domande incorrelate. In tale condizione, le espressioni per μ_Z e μ_V sono facilmente ottenibili dalla (3.5), sostituendo μ_S con μ_W e μ_R con μ_T, ossia $\mu_Z = h\mu_Y + c$ e $\mu_V = h'\mu_X + c'$, dove $h = h' = p$ mentre $c = (1-p)\mu_W$ e $c' = (1-p)\mu_T$.

La varianza di Z è $\sigma_Z^2 = p\mu_Y^2\left(1+C_Y^2\right)+(1-p)\mu_W^2\left(1+C_W^2\right)-\mu_Z^2$ mentre, quella di V risulta $\sigma_V^2 = p\mu_X^2\left(1+C_X^2\right)+(1-p)\mu_T^2\left(1+C_T^2\right)-\mu_V^2$. Le covarianze in (3.7) e (3.8) divengono $\sigma_{VZ} = p\sigma_{XY} + p(1-p)(\mu_Y - \mu_W)(\mu_X - \mu_T)$, $\sigma_{YV} = p\sigma_{XY}$ e $\sigma_{YZ} = p\sigma_Y^2$, poiché $\sigma_{YT} = \sigma_{YW} = \sigma_{WT} = 0$.

Inoltre, ponendo $p = 1$, non si utilizza alcuna procedura casualizzata, ed il metodo si riduce all'intervista diretta, sia per la variabile sensibile, che per la variabile ausiliaria. In questo modo, μ_Y viene stimato dal classico stimatore per regressione di Y su X, oppure, per $b = 0$, dalla media campionaria $\bar{y}$.

Modello 2

La variabile sensibile è rilevata mediante la tecnica a domande incorrelate, mentre la variabile ausiliaria è osservata senza codifica. Ovviamente μ_Z, σ_Z^2 e σ_{YZ}, c ed h sono definite come nel Modello 1, mentre $\mu_V = \mu_X$ e $\sigma_V^2 = \sigma_X^2$. Inoltre, $\sigma_{VZ} = p\sigma_{XY}$ e $\sigma_{YV} = \sigma_{XY}$, poiché $\sigma_{XW} = 0$.

Modello 3

La variabile sensibile prevede una codifica additiva, mentre la variabile ausiliaria è osservata direttamente. L'espressione di μ_Z sarà $\mu_Z = h\mu_Y + c$, dove $h = 1$ mentre $c = (1-p)\mu_U$; per quanto riguarda la varianza di Z abbiamo $\sigma_Z^2 = \sigma_Y^2 + (1-p)\mu_U^2\left(C_U^2 + p\right)$. Per media, varianza di V e σ_{YV}, ci si riferisca al Modello 2. In più, per le covarianze si ha $\sigma_{VZ} = \sigma_{XY}$ e $\sigma_{YZ} = \sigma_Y^2$.

Notiamo inoltre che, per $b = 0$ e $p = 0$, viene riprodotto il modello di additivo.

Modello 4

La variabile sensibile presenta una codifica di tipo moltiplicativo, mentre la variabile ausiliaria ne è priva.

L' espressione della media di Z è $\mu_Z = h\mu_Y + c$, dove $h = [p + (1-p)\mu_W]$ e $c = 0$, mentre la varianza di Z è $\sigma_Z^2 = \mu_Y^2(1+C_Y^2)[p + (1-p)\mu_W^2(1+C_W^2)] - \mu_Z^2$.

μ_V, σ_V^2 e σ_{YV} sono le medesime del Modello 2. Infine, in merito alle covarianze, si ha $\sigma_{VZ} = \sigma_{XY}[p + (1-p)\mu_W]$ e $\sigma_{YZ} = \sigma_Y^2[p + (1-p)\mu_W]$.

Notiamo che, per $b = 0$, si ottiene il modello di Bar-Lev e, come caso particolare di quest'ultimo, per $p = 0$, il modello moltiplicativo.

Modello 5

La variabile sensibile, come anche la variabile ausiliaria, sono rilevate con una codifica di tipo additivo. Si veda dunque il Modello 3, per le espressioni di μ_Z, σ_Z^2 e σ_{YZ}, mentre, media e varianza di V sono, rispettivamente, $\mu_V = h'\mu_X + c'$ dove $h' = 1$, $c' = (1-p)\mu_H$, e $\sigma_V^2 = \sigma_X^2 + (1-p)\mu_H^2(C_H^2 + p)$. La covarianza tra V e Z è $\sigma_{VZ} = \sigma_{XY} + p(1-p)\mu_U\mu_H$, mentre quella tra Y e V risulta $\sigma_{YV} = \sigma_{XY}$.

Modello 6

Sia la variabile sensibile, che la variabile ausiliaria prevedono una codifica di tipo moltiplicativo. Si consideri il Modello 4, per le espressioni di μ_Z, σ_Z^2 e σ_{YZ}. In merito a V si ha $\mu_V = h'\mu_X + c'$, dove $h' = [p + (1-p)\mu_T]$ e $c' = 0$, inoltre $\sigma_V^2 = \mu_X^2(1+C_X^2)[p + (1-p)\mu_T^2(1+C_T^2)] - \mu_V^2$. Le covarianze tra V e Z e tra Y e V sono date da $\sigma_{VZ} = \sigma_{XY}[p + (1-p)\mu_T\mu_W] + p(1-p)\mu_Y\mu_X(1-\mu_W)(1-\mu_T)$ e $\sigma_{YV} = \sigma_{XY}[p + (1-p)\mu_T]$.

3.5 Confronti di efficienza e di protezione della privacy

Vari autori hanno tentato di confrontare l'efficienza tra differenti strategie casualizzate, rapportando le varianze degli stimatori del parametro di interesse. D'altra parte, è noto in letteratura che generalmente, l'efficienza di una strategia aumenta, al diminuire del livello di protezione della privacy. Dal momento che anche il grado di riservatezza garantito è un criterio essenziale di valutazione, pare ragionevole confrontare l'efficienza dei modelli, tenendo possibilmente costante il livello di protezione.

In questo lavoro si intende studiare il guadagno di efficienza che si può acquisire utilizzando una variabile ausiliaria (con o senza codifica) rispetto all'intervista diretta, nei 6 modelli descritti nella *Tabella* 3.1, per fissarne l'ordinamento in termini di efficienza, e successivamente, in termini di protezione della privacy.

Dal punto di vista dell'efficienza, lo stimatore ottimale della classe (2.2) è più preciso dello stimatore $\bar{y}$ se $\sigma_Y^2 > \sigma_Z^2\left(1-\rho_{VZ}^2\right)/h^2$.

Purtroppo, le condizioni di superiorità dello stimatore ottimale per i singoli modelli rispetto all'intervista diretta si possono rendere in forma semplice solo in situazioni banali. Data la complessità analitica dei confronti, si è deciso di valutare graficamente le procedure in particolari situazioni, le quali, seppur significative, hanno un valore puramente esemplificativo. L'impostazione seguita prevede uno studio congiunto dell'efficienza e della tutela della privacy: questo ci è parso l'approccio più ragionevole, considerata la stretta connessione tra i due aspetti.

Valuteremo l'efficienza relativa

$$\frac{\sigma_Z^2\left(1-\rho_{ZV}^2\right)}{h^2\sigma_Y^2} \tag{3.9}$$

al variare del parametro di disegno $p \in [0,1]$.

Per quantificare il grado di riservatezza, adotteremo la misura di privacy generalizzata

$$\tau = 1-\rho_{Y.VZ}^2 = 1-\frac{\rho_{YV}^2+\rho_{YZ}^2-2\rho_{YV}\rho_{YZ}\rho_{VZ}}{1-\rho_{VZ}^2}$$

al variare, ancora, del parametro di disegno $p \in [0,1]$.

I nostri obiettivi per quanto riguarda i confronti di <u>efficienza</u> sono:

- stabilire se i modelli casualizzati con una variabile ausiliaria (codificata o meno) possono essere più efficienti dell'intervista diretta.
- valutare la precisione dello stimatore di ciascun modello casualizzato rispetto agli altri.

Gli intendimenti in merito ai confronti in termini di protezione della <u>privacy</u> sono:

- esaminare il livello di protezione della privacy garantito dai modelli casualizzati con una variabile ausiliaria (codificata o meno) rispetto all'intervista diretta.
- valutare il livello di riservatezza garantito da ciascun modello casualizzato rispetto agli altri.

Per attuare i confronti tra le procedure, la scelta da effettuare con maggiore cautela riguarda la distribuzione della variabile di codifica.

Come riportato nel Capitolo 1, in letteratura non viene suggerita una particolare distribuzione per la codifica additiva U mentre varie, sono le proposte per quella moltiplicativa W. In merito a questa, è stata consigliata la distribuzione F di Fisher con (5,5) gradi di libertà per il modello moltiplicativo (1.7); invece, al fine di garantire la maggiore efficienza del modello di Bar-Lev (1.15), rispetto al moltiplicativo (1.7), è stata indicata un'esponenziale di media $\mu_W = 1/\lambda_W$, con $2-\sqrt{2} < \lambda_W < 2+\sqrt{2}$.

Dopo avere valutato attentamente queste proposte, si è giunti alla conclusione che uno studio adeguatamente rappresentativo sia possibile solo privilegiando delle distribuzioni che enfatizzino bene la molteplicità dei casi reali, al variare dei parametri. In quest'ottica, la F di Fisher è sicuramente una distribuzione idonea, in quanto offre una vasta gamma di opzioni al variare dei gradi di libertà.

Tra le 6 procedure confrontate, l'enfasi verrà posta sui modelli di tipo additivo (Modelli 3 e 5) e di tipo moltiplicativo (Modelli 4 e 6). Le tecniche a domande incorrelate (Modelli 1 e 2) sono presenti per completezza di esposizione, ma verranno considerate solo marginalmente, in quanto si ritiene siano di utilità ridotta nei casi ricorrenti nella pratica. Il nostro intento è di rappresentare, in maniera più completa possibile, i diversi rendimenti ottenibili dalle strategie additive e moltiplicative, del tutto consapevoli che le scelte effettuate sui valori dei parametri non favoriscono i modelli a domande incorrelate.

È opportuno iniziare da un semplice risultato inerente al modello additivo (1.4) e al moltiplicativo (1.7) proposti in letteratura. Lo stimatore (1.5), associato al primo modello, è più efficiente di quello definito per il secondo (1.8), se

$$\sigma_U^2 < \frac{\sigma_W^2 \mu_Y^2 \left(1 + C_Y^2\right)}{\mu_W^2} \qquad (3.10)$$

Si noti che la condizione non dipende da μ_U in quanto tale parametro non compare nella varianza di stima (1.6).

Nota 3.1 Se U e W hanno la stessa varianza, la (3.10) diviene

$$\mu_W^2 < \mu_Y^2 \left(1 + C_Y^2\right) \qquad (3.11)$$

Osserviamo che la condizione è sempre vera se $\mu_Y = \mu_W$.

Aumentando opportunamente il valore di μ_W, il modello (1.7) può divenire più efficiente del modello (1.4). Infatti, a parità degli altri parametri, la varianza (1.9) si riduce al diminuire di C_W. □

Nota 3.2 Se si valuta il livello di riservatezza dei modelli (1.4) e (1.7) mediante $\tau = 1 - h^2 \sigma_Y^2 / \sigma_Z^2$, si riscontra che il modello additivo è più protettivo del moltiplicativo quando

$$\sigma_U^2 > \frac{\sigma_W^2 \mu_Y^2 \left(1 + C_Y^2\right)}{\mu_W^2}$$

Tale condizione è identica alla (3.10) con segno inverso. Di conseguenza, a parità di efficienza relativa, il livello di protezione misurato mediante τ è equivalente. □

Svilupperemo lo studio in due casi, a seconda che le scelte effettuate sui valori dei parametri soddisfino o meno la (3.10). Il primo caso illustra delle situazioni, in cui le strategie di tipo additivo possono essere più efficienti delle moltiplicative; il secondo, invece, presenta dei contesti in cui i modelli di tipo moltiplicativo possono fornire uno stimatore più preciso.

Si realizzano i confronti in due casi al fine di rendere la trattazione il più possibile rappresentativa, data la mancanza di un modello preferibile in assoluto. La scelta della distribuzione della variabile di codifica, e dei parametri relativi, permette infatti di modellare il comportamento delle strategie nel modo desiderato.

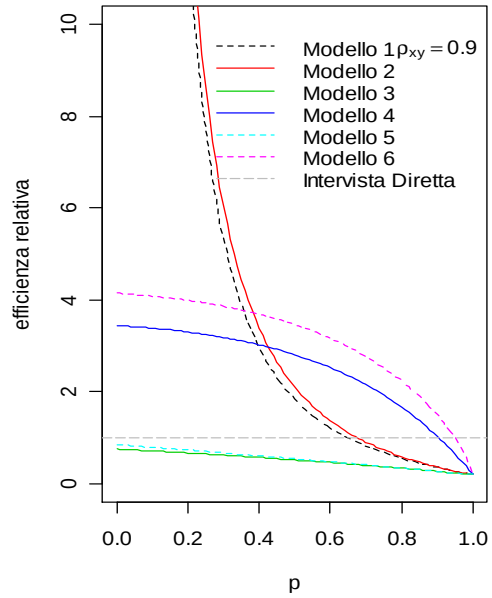
In entrambi i casi è stata adottata la distribuzione F di Fisher, fissando i gradi di libertà in modo da soddisfare (*Caso 1*) o violare (*Caso 2*) la condizione (3.10). In questo modo è possibile illustrare, quanto siano mutevoli i rendimenti delle procedure, anche se si stima la medesima quantità ignota, disponendo della stessa informazione supplementare. Nel dettaglio, i valori prescelti sono $\mu_X = 10$ e $\mu_Y = 2$, $C_X = 0.8$ e $C_Y = 2$.

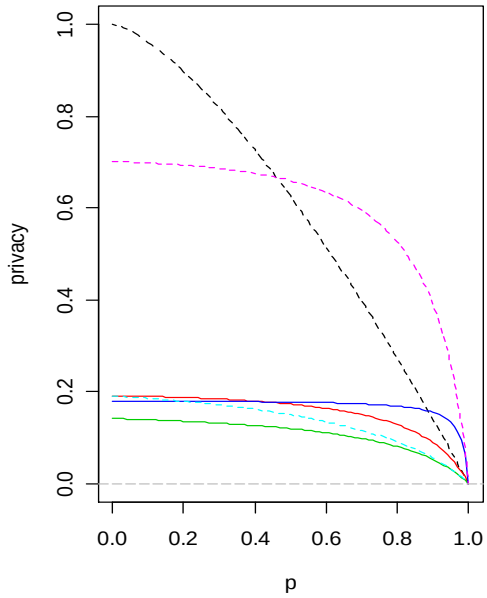
Nel *Caso 1*, si assume che la distribuzione di H ed U abbia (5,5) g.d.l., mentre quella di T e W (10,5) g.d.l.; i rispettivi coefficienti di variazione sono $C_H = C_U = 1.789$ e $C_T = C_W = 1.612$. Il *Caso 2* prevede invece che la distribuzione di H ed U sia caratterizzata da (1,5) g.d.l. mentre, quella di T e W da (10,50) g.d.l., con $C_H = C_U = 2.828$ e $C_T = C_W = 0.502$.

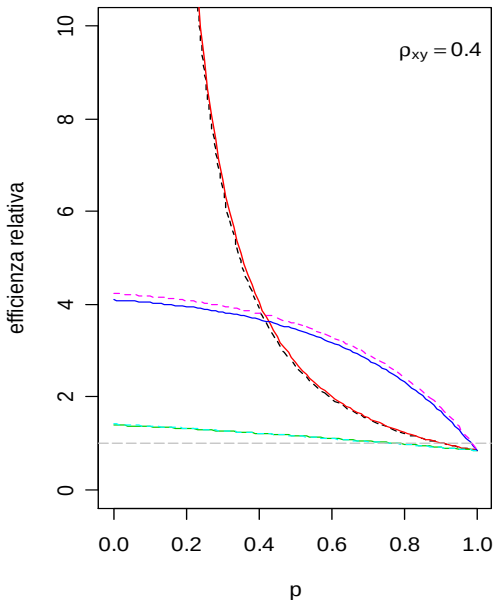
Di seguito si riportano, per brevità di esposizione, solo i grafici relativi a $\rho_{XY} = 0.0$, 0.4 e 0.9.

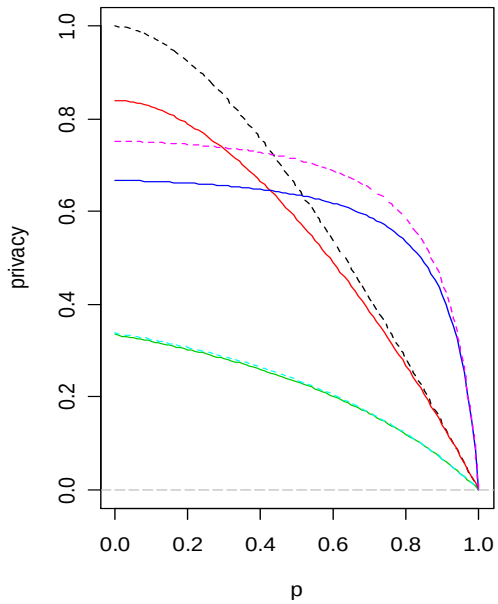
Per l'analisi completa delle situazioni esaminate con $(\rho_{XY} = 0.0, \mathbf{0.2}, 0.4, \mathbf{0.7}, 0.9)$, si rimanda all'Appendice A.

Caso 1. $H,U \sim F(5,5)$ $C_H = C_U = 1.789$ $T,W \sim F(10,5)$ $C_T = C_W = 1.612$

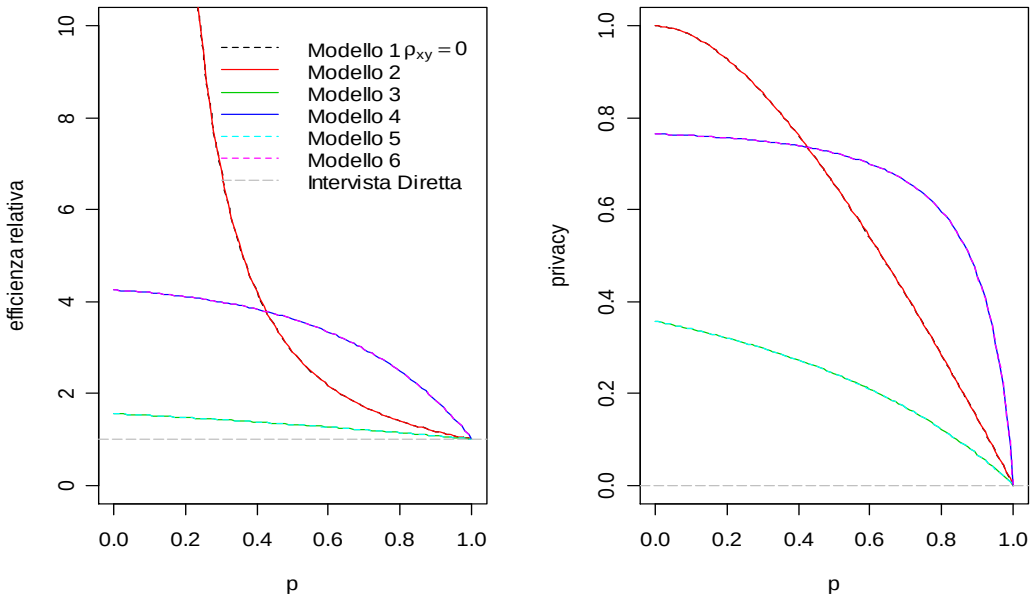




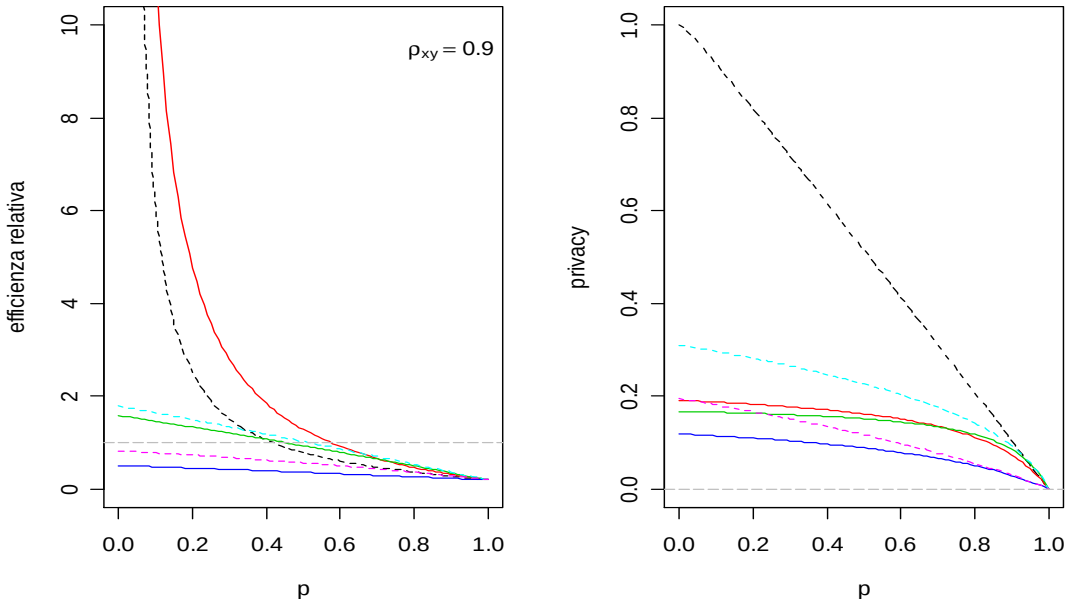




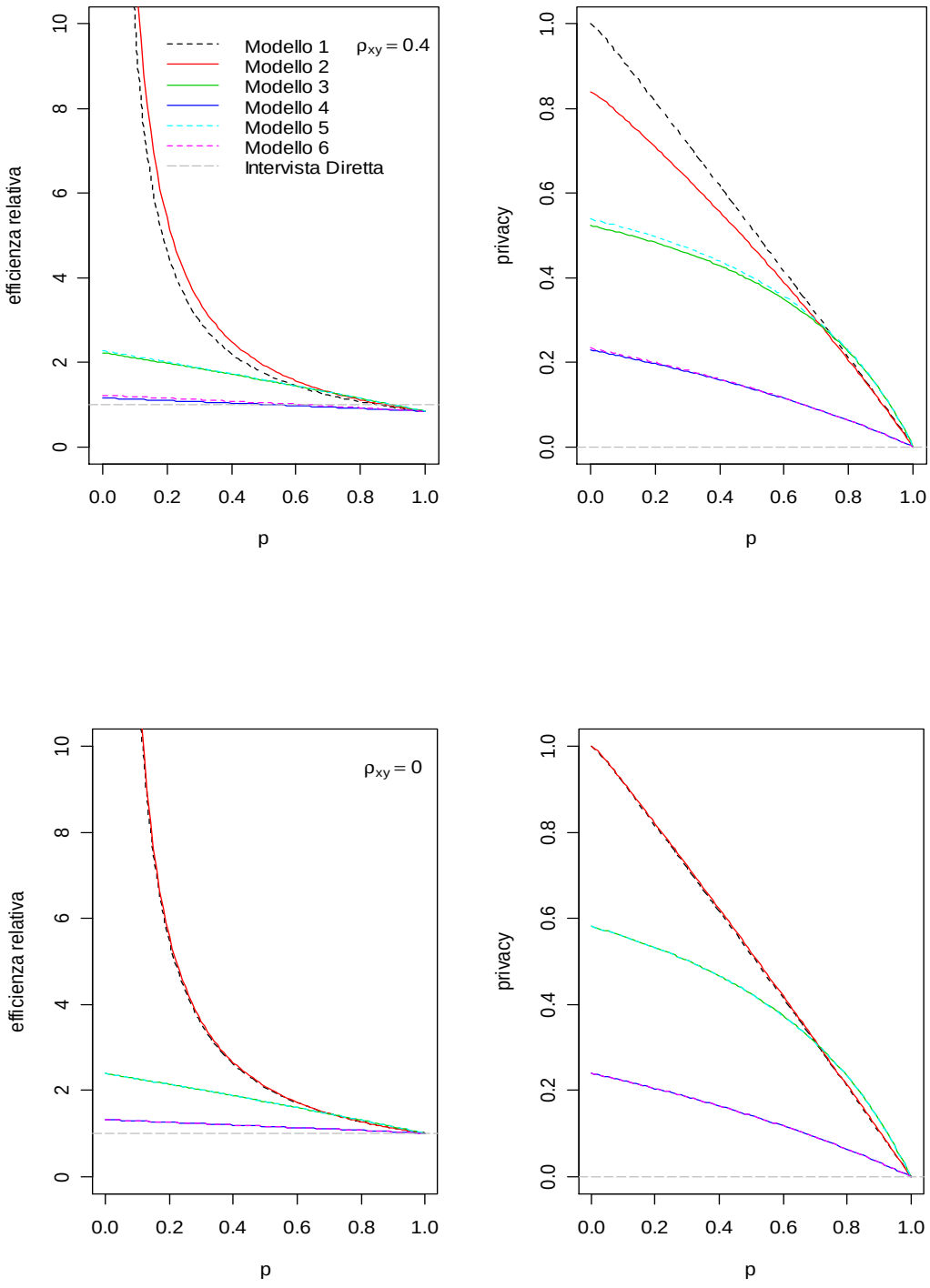
Caso 1. $H,U \sim F(5,5) \quad C_H = C_U = 1.789 \qquad T,W \sim F(10,5) \quad C_T = C_W = 1.612$



Caso 2. $H,U \sim F(1,5) \quad C_H = C_U = 2.828 \qquad T,W \sim F(10,50) \quad C_T = C_W = 0.502$



Caso 2. $H, U \sim F(1,5)$ $C_H = C_U = 2.828$ $T, W \sim F(10,50)$ $C_T = C_W = 0.502$



Come previsto, nel *Caso 1* si riscontra la scarsa efficienza delle procedure moltiplicative e a domande incorrelate, rispetto ai modelli di tipo additivo. Al contrario, nel *Caso 2*, le strategie additive possono rivelarsi poco efficienti, se confrontate con le moltiplicative, anche se spesso sono in grado di ottenere uno stimatore più preciso delle tecniche a domande incorrelate.

Osserviamo anzitutto, che i modelli casualizzati che usano una variabile ausiliaria (codificata o meno) possono essere più efficienti dell'intervista diretta. Nelle situazioni illustrate, l'efficienza relativa è una funzione decrescente di p. Quando $p=1$, tutti i modelli riproducono l'intervista diretta con variabile ausiliaria e quindi la (3.9) diviene sempre inferiore a 1, assumendo valore pari a $1-\rho_{XY}^2$. Questo comporta che, se $|\rho_{XY}|>0$, sussistono sempre dei valori di p, per i quali ciascun modello con variabile ausiliaria fornisce uno stimatore più preciso di $\bar{y}$. In particolare, per $p=0$, è possibile dimostrare che il Modello 3 è più efficiente dell'intervista diretta solo se

$$\sigma_U^2 < \rho_{XY}^2 \sigma_Y^2 \tag{3.12}$$

ciò si verifica per il Modello 4 se

$$C_W^2 \mu_Y^2 \left(1+C_Y^2\right) < \rho_{XY}^2 \sigma_Y^2 \tag{3.13}$$

Nei grafici presentati, le condizioni suddette si riscontrano in corrispondenza di $\rho_{XY}=0.9$: la (3.12) è soddisfatta nel *Caso 1*, mentre la (3.13) nel *Caso 2*.

Poiché, nelle situazioni descritte, l'efficienza relativa è decrescente in p, le condizioni (3.12) e (3.13) assicurano la maggiore efficienza dei modelli rispetto all'intervista diretta per ogni $p \in [0,1]$.

L'efficienza delle strategie si riduce al diminuire di ρ_{XY}, anche se non si osservano cambiamenti sostanziali in termini di ordinamento. In particolare, per $p = 0$, è possibile verificare che l'efficienza relativa dei Modelli 3 e 4, decresce di un fattore costante pari a ρ_{XY}^2, rispetto all'efficienza relativa raggiungibile senza variabile supplementare. Sussiste infatti la seguente uguaglianza

$$\frac{\sigma_Z^2\left(1-\rho_{ZV}^2\right)}{h^2\sigma_Y^2} = \frac{\sigma_Z^2}{h^2\sigma_Y^2} - \rho_{XY}^2 \tag{3.14}$$

Dunque, come era da attendersi, la precisione dello stimatore migliora, al crescere della correlazione tra il carattere sensibile e la variabile ausiliaria.
Tale risultato è di notevole rilevanza, in quanto anche l'introduzione di informazione ausiliaria senza costi aggiuntivi e di facile reperibilità, come dati amministrativi, può implicare comunque un guadagno di precisione. Inoltre, utilizzando informazione di dominio pubblico, non vi è neppure la necessità di una codifica.

L'introduzione delle variabili di codifica T ed H comporta, in generale, una diminuzione dell'efficienza dello stimatore. Nelle situazioni presentate, ciò diviene evidente nel passaggio dal Modello 4 (senza codifica) al Modello 6 (con codifica), ma si osserva meno vistosamente anche per il Modello 3 (senza codifica) rispetto al Modello 5 (con codifica).
A livello analitico, l'efficienza relativa dei Modelli 5 e 6, per $p = 0$, vale rispettivamente

$$\frac{\sigma_Z^2\left(1-\rho_{ZV}^2\right)}{h^2\sigma_Y^2} = \frac{\sigma_Z^2}{h^2\sigma_Y^2} - \rho_{XY}^2\left(\frac{\sigma_X^2}{\sigma_X^2+\sigma_H^2}\right) \tag{3.15}$$

$$\frac{\sigma_Z^2\left(1-\rho_{ZV}^2\right)}{h^2\sigma_Y^2}=\frac{\sigma_Z^2}{h^2\sigma_Y^2}-\rho_{XY}^2\left(\frac{\sigma_X^2}{\sigma_X^2+C_T^2\mu_X^2\left(1+C_X^2\right)}\right) \qquad (3.16)$$

Poiché, l'efficienza relativa è una funzione decrescente di p , si vede immediatamente che il Modello 3 è più efficiente del Modello 5 per ogni $p\in[0,1]$, essendo

$$\frac{\sigma_X^2}{\sigma_X^2+\sigma_H^2}<1$$

A parità degli altri parametri, tanto più grande è σ_H^2 rispetto a σ_X^2, tanto più risulta vantaggioso il Modello 3 del Modello 5.
Analogamente, il Modello 4 è più efficiente del Modello 6 per ogni $p\in[0,1]$, valendo

$$\frac{\sigma_X^2}{\sigma_X^2+C_T^2\mu_X^2\left(1+C_X^2\right)}<1$$

A parità degli altri parametri, si nota che tanto maggiore è C_T^2 rispetto a σ_X^2**,** tanto più preciso è il Modello 4 del Modello 6.

Le tecniche a domande incorrelate assumono invece un comportamento anomalo. È più efficiente il modello con variabile ausiliaria codificata (Modello 1), piuttosto che il modello con variabile ausiliaria osservata direttamente (Modello 2). Ciò risulta notevolmente interessante per il ricercatore che predilige queste strategie, dal momento che il Modello 1 fornisce uno stimatore più preciso del Modello 2, assicurando, al tempo stesso, più riservatezza.

Sul fronte della tutela della privacy, come intuibile, gli ordinamenti sono speculari: considerando le strategie additive e moltiplicative, la maggiore efficienza è garantita dal modello meno protettivo, e viceversa.
Pertanto, nel *Caso 1*, i modelli di tipo additivo si presentano svantaggiosi per tutelare la riservatezza rispetto alle tecniche moltiplicative, e a domande incorrelate. D'altronde, le situazioni descritte nel *Caso 2* evidenziano la carenza di protezione delle strategie moltiplicative, rispetto sia a modelli di tipo additivo, che a domande incorrelate.
L'effetto delle variabili di codifica T ed H è pure prevedibile: se da un lato queste implicano una diminuzione dell'efficienza, dall'altro comportano un aumento del livello di tutela della privacy. Questo aspetto si apprezza all'aumentare di ρ_{XY} nel passaggio dal Modello 3 al 5, e dal Modello 4 al 6.

3.6 Conclusioni

I confronti in termini di <u>efficienza</u> hanno messo in luce:

- l'assoluta convenienza ad utilizzare l'informazione ausiliaria, con o senza codifica. È stato accertato che le strategie basate su di una variabile supplementare, possono essere addirittura più efficienti dell'intervista diretta.
- nel *Caso 1*, la preminenza delle tecniche additive (Modelli 3 e 5), rispetto alle moltiplicative (Modelli 4 e 6).
- nel *Caso 2*, la superiorità delle procedure moltiplicative (Modelli 4 e 6), rispetto alle additive (Modello 3 e 5).

In merito ai confronti di protezione della privacy si sono riscontrati:

- i sostanziali miglioramenti in termini di protezione ottenibili anche con piccole riduzioni di efficienza.
- nel *Caso 1*, l'inadeguatezza dei modelli di tipo additivo (Modelli 3 e 5) per indagini in cui la variabile oggetto di studio è notevolmente delicata. In queste situazioni, le strategie moltiplicative (Modelli 4 e 6) sono sicuramente preferibili.
- nel *Caso 2*, la scarsa protezione delle procedure moltiplicative (Modelli 4 e 6) rispetto alle additive (Modelli 3 e 5).

Di fondamentale importanza, per la nostra trattazione, è il beneficio riscontrato nell'utilizzo di una variabile ausiliaria, con eventuale codifica: l'introduzione di informazione supplementare è sempre un vantaggio nella rilevazione di un carattere sensibile, dal momento che accresce l'efficienza dello stimatore, senza compromettere pesantemente la privacy dei rispondenti. Tale contributo diviene poi di immenso valore, se pensiamo che può permettere di ottenere stime ancora più precise di quelle fornite dall'intervista diretta.
I confronti attuati non hanno la pretesa di essere esaustivi, e pertanto non conducono ad alcun tipo di ordinamento definitivo delle strategie discusse. D'altro canto, è inverosimile pensare di poter stabilire un ordine in termini di efficienza e di protezione della privacy tra i modelli: i grafici che abbiamo riportato devono essere intesi come un utile strumento per orientarsi. Ovviamente, la scelta dipende dal contesto pratico in cui si opera. Troppi sono infatti i fattori che entrano in gioco: la sensibilità della variabile, considerazioni di carattere etico e morale, nonché aspetti di natura psicologica, che potrebbero influenzare i rispondenti. In quest'ottica, lo studio che abbiamo condotto, mira ad

evidenziare la molteplicità delle opzioni di scelta, al fine di ovviare a qualsiasi esigenza riscontrabile nella pratica. Dal momento che non sussiste un modello privilegiato in assoluto, diviene inevitabile che traspaia in modo netto la soggettività del ricercatore.
È infatti possibile stabilire la strategia a priori, basandosi su preferenze puramente personali. Riteniamo che il punto critico, in grado di compromettere la bontà dell'indagine, non sia la selezione della procedura, ma bensì, la scelta della distribuzione della variabile di codifica, e dei valori dei parametri relativi. Ciascuna strategia può di fatto conseguire gli esiti sperati, sia in termini di efficienza che di tutela della privacy: l'abilità consiste solo nell'individuare la situazione più congeniale che ottimizzi i rendimenti.

Avendone verificato l'assoluta convenienza, proponiamo l'adozione di una procedura basata su una variabile ausiliaria, e descriviamo ora un possibile percorso per il ricercatore che intende applicare questa metodica.
In fase orientativa, il ricercatore deve documentarsi adeguatamente sul contesto di indagine in cui opera. Studi compiuti in precedenza, indagini pilota, e testimoni privilegiati possono dare utili indicazioni in merito all'ordine di grandezza di media e varianza della variabile sensibile.
Successivamente, in base a considerazioni personali, il ricercatore sceglie una tecnica di rilevazione: la nostra opinione è che le strategie di tipo additivo e moltiplicativo rendano meglio nella pratica, costituendo un superamento delle procedure a domande incorrelate. In via preliminare, ci sembra preferibile non considerare la variabile ausiliaria, riferendosi ai modelli (1.4) e (1.7).
Il ricercatore è tenuto ad ipotizzare per eccesso il livello di riservatezza desiderato, che denotiamo con $\tilde{\tau}$. Sono essenzialmente due i motivi per delineare una protezione garantista, prima di effettuare una qualsiasi considerazione sulla

precisione dello stimatore. Il primo è che l'introduzione della variabile ausiliaria migliorerà comunque l'efficienza, abbassando, per quanto poco, la protezione; il secondo è che, in merito alle situazioni illustrate, l'efficienza relativa è una funzione decrescente di p. I confronti indicano quindi, che il parametro p costituisce un ulteriore elemento capace di aumentare la precisione, e lo considereremo in seguito.

Una volta scelto $\tilde{\tau}$, si ricavano automaticamente i parametri per la variabile di codifica, giacché media e varianza di Y sono già state precedentemente ipotizzate. Qualora il ricercatore scelga una tecnica additiva, si ottiene

$$\sigma_U^2 = \frac{\tilde{\tau}}{1-\tilde{\tau}}\sigma_Y^2 \qquad (3.17)$$

Nota 3.3 Adottando una strategia additiva, il metodo descritto non permette di calcolare la media μ_U; tale quantità è tuttavia ininfluente per l'efficienza dello stimatore (1.5), e pertanto può essere fissata in modo arbitrario. □

Se invece si preferisce una procedura moltiplicativa, si possono calcolare i parametri μ_W e σ_W^2, imponendo dei vincoli sugli stessi. Ad esempio, se restringiamo la scelta alle variabili W, tali che $k\mu_W = \sigma_W^2$, con k costante fissata

$$\sigma_W^2 = k^2\left(\frac{1-\tilde{\tau}}{\tilde{\tau}}\right)\left(\frac{\mu_Y^2}{\sigma_Y^2}+1\right) \qquad (3.18)$$

o, equivalentemente

$$\mu_W = k\left(\frac{1-\tilde{\tau}}{\tilde{\tau}}\right)\left(\frac{\mu_Y^2}{\sigma_Y^2}+1\right) \qquad (3.19)$$

Il ricercatore deve poi scegliere la distribuzione della variabile di codifica. In base alla nostra analisi, reputiamo che la F di Fisher sia particolarmente vantaggiosa, data la tipica flessibilità al variare dei suoi parametri. Si consiglia di calcolare una coppia di gradi di libertà approssimativamente corrispondenti a media e varianza ipotizzate. In tal modo, il ricercatore controlla graficamente il comportamento della strategia prescelta al variare di $p \in [0,1]$: ribadiamo che, in questa fase, si utilizzano i soli grafici per $\rho_{XY} = 0$.

Se la situazione trovata non è del tutto soddisfacente, suggeriamo di far variare i gradi di libertà, sino ad identificarne una migliore. Si ritiene, a proposito, che un buon criterio consista nell'armonizzare l'esigenza di precisione dello stimatore, con la riservatezza assicurata ai rispondenti.

Si valuta poi se introdurre una variabile ausiliaria osservata direttamente, oppure a sua volta codificata: naturalmente ciò dipenderà dal grado di correlazione tra l'informazione ausiliaria ed il carattere sensibile. In particolare, può essere conveniente anche una variabile supplementare con correlazione bassa. Inoltre, recuperando dati pubblici, si ottiene dell'informazione ausiliaria a costo nullo, o quasi, senza rischiare di ledere la privacy dei rispondenti.

Infine, l'ultima scelta riguarda il parametro di casualizzazione p. Stabilire un valore adeguato di p è di importanza cruciale, dipendendo da ciò sia il rendimento di efficienza, che di protezione della privacy. In letteratura, si sconsigliano valori superiori a 0.7, considerati impraticabili, mentre si ritengono applicabili 0.2 e 0.3. Poiché l'utilizzo di una variabile ausiliaria aumenta la precisione dello stimatore, si può scegliere un valore di p più basso rispetto alle situazioni in cui $\rho_{XY} = 0$. Questo aspetto rappresenta un ulteriore vantaggio ottenuto grazie all'uso di informazione supplementare, dal momento che al diminuire di p, i rispondenti si sentono più tutelati.

3.7 Nota bibliografica

Osserviamo che la tutela della privacy è un tema discusso e sentito, non solo per indagini volte a rilevare variabili sensibili, ma più in generale, per agenzie statistiche che trattano grandi basi di dati contenenti informazioni personali. Di conseguenza, le tecniche di data mining sono considerate una sfida per la protezione della privacy, data la loro naturale tendenza ad utilizzare informazioni sensibili circa gli individui. Questo è il motivo per cui i metodi di codifica sono abitualmente utilizzati da organizzazioni che forniscono dati a fini statistici, e si giustifica il fatto che molte misure di protezione della privacy sono state sviluppate e discusse in ambito di data mining.

Alcune tra le misure di privacy che abbiamo introdotto nel par. 3.2, sono nate in tale contesto: per approfondimenti, è possibile riferirsi all'articolo di Mukharjee e Duncan (1997), in cui viene trattata un'analisi di variabili sensibili a distribuzione asimmetrica. Gli autori si propongono di valutare l'effetto di una codifica additiva, suggerendo a tale scopo delle misure per quantificare il livello di riservatezza conseguito.

Le misure basate sul concetto di entropia sono sviluppate nell'articolo di Evfimievski (2002), il quale si ispira ai lavori di Shannon sui sistemi di sicurezza, e nell'articolo di Agrawal e Aggarwal (2001).

Ulteriori aspetti relativi alla privacy nel data mining sono rintracciabili nelle pubblicazioni di Fienberg e Willenborg (1998), Mukherjee e Duncan (1997), Agrawal e Srikant (2000), Mukherjee e Duncan (2000).

In merito alle misure costruite con gli intervalli di confidenza, si trova l'articolo già citato di Eichhorn e Hayre (1983), in cui gli autori del modello moltiplicativo, descrivono un metodo per quantificarne la tutela della privacy.

In questa sede, vale la pena di ricordare anche i vari tentativi attuati per misurare adeguatamente la riservatezza quando la variabile sensibile è qualitativa. In letteratura, vari autori si sono proposti di definire il livello di rischio del rispondente di essere identificato come appartenente al gruppo associato al carattere sensibile. A riguardo, si rimanda alle pubblicazioni di Leysieffer e Warner (1976) e di Ljungqvist (1993) in cui si propongono delle misure di protezione basate sulle probabilità condizionate di essere identificato come appartenente al gruppo sensibile, avendo dato una risposta affermativa o negativa.

Più recenti sono le pubblicazioni di Guerriero e Sandri (2005 e 2007) in cui si enfatizza l'importanza di considerare, non solo l'efficienza, ma anche la protezione della privacy al fine di ottenere un confronto più adeguato tra diverse procedure. In quest'ottica, gli autori paragonano l'efficienza di alcune strategie, a parità di protezione della privacy, valutata mediante le misure ideate da Lanke (1976) e da Leysieffer e Warner (1976).

La misura di protezione della privacy generalizzata (3.3) ed i casi speciali presentati nel par. 3.4 sono stati sviluppati nel corso di questo lavoro.

Capitolo 4

I modelli di Saha modificati

4.1 Introduzione

Nel Capitolo 3, si è osservato che la precisione dello stimatore e la riservatezza garantita si possono in parte modellare grazie, non solo alla scelta della distribuzione della variabile di codifica, e dei suoi parametri, ma anche imponendo dei vincoli sugli stessi. Si può dunque pensare di sfruttare tale elasticità, orientandosi verso la formulazione di alcuni modelli misti allo scopo di equilibrare al meglio efficienza e protezione della privacy.

La strada del modello misto è già stata parzialmente percorsa in letteratura, grazie ad una prima formulazione, nata in contemporanea con i metodi a risposte codificate, ed è stata ripresa recentemente da Saha. L'introduzione di una seconda variabile di codifica assicura un più alto grado di riservatezza gravando però, a sua volta, sulla precisione dello stimatore. A nostro avviso, l'impostazione di Saha appare eccessivamente rigida: la sua tendenza a favorire il rispetto della privacy, a scapito dell'efficienza, difficilmente si presta al raggiungimento di un buon compromesso.

Pertanto, in questo capitolo proponiamo alcune tecniche miste al fine di migliorare il modello di Saha, incrementandone la precisione di stima, e preservandone, possibilmente, la naturale attitudine a proteggere la riservatezza.

4.1.1 Simbologia

In questo capitolo useremo le seguenti sigle per designare i modelli:

- MDI: Modello a domande incorrelate (1.1)
- MA: Modello additivo (1.4)
- MM: Modello moltiplicativo (1.7)
- MSO: Modello di Saha originale (1.10)
- MSM: Modelli di Saha modificati. Verrà aggiunto il numero 1, 2, o 3 a seconda del modello da denotare.

- Per indicare l'introduzione di una variabile ausiliaria con o senza codifica, aggiungeremo alla sigla di un modello, rispettivamente, "Vcod" o "V".
- Per denotare un modello in presenza o in assenza di una caratteristica, indicheremo la stessa tra parentesi (.).

4.2 I modelli di Saha modificati

Ricordiamo, anzitutto, che la variabile risposta del MSO è $Z = W(Y + U)$, data da una combinazione di una codifica di tipo additivo con una di tipo moltiplicativo. Nel par. 1.4 si è dimostrato analiticamente che il MSO assicura più riservatezza sia del MA che del MM, costituendo un utile strumento per la rilevazione di un carattere molto delicato. D'altra parte, il MSO può fornire uno stimatore poco preciso, limitando notevolmente le possibilità di scelta del ricercatore. Il tentativo di migliorare l'idea di Saha ci ha indotto a formulare tre possibili sviluppi, volti a raggiungere un equilibrio più soddisfacente tra efficienza dello stimatore e rispetto della privacy.

Il primo sviluppo, il MSM1, nasce da una combinazione tra il MM ed il MDI. L'i-esimo intervistato estrae, ad esempio con un calcolatore, due numeri casuali, w_i e u_i, e in seguito comunica la risposta $z_i = w_i[ay_i + (1-a)u_i]$**,** dove a è una costante fissata dal ricercatore che varia nell'intervallo $(0,1]$. Quindi, la risposta osservata è

$$Z = W[aY + (1-a)U] \tag{4.1}$$

Si noti che per $a = 1$, la procedura diviene il MM.
Uno stimatore non distorto di μ_Y è

$$\hat{\mu}_{SA1} = \frac{\bar{z} - (1-a)\mu_W \mu_U}{a\mu_W} \tag{4.2}$$

Richiamando l'espressione della varianza della classe generale di stimatori per $b = 0$, data da

$$V(\hat{\mu}_M) = \frac{\sigma_Z^2}{nh^2} \tag{4.3}$$

è possibile ricavare $V(\hat{\mu}_{SA1})$ ponendo

$$\sigma_Z^2 = a^2\mu_W^2\left[\sigma_Y^2 + \mu_Y^2(1+C_Y^2)C_W^2\right] + (1-a)^2\mu_W^2\left[C_W^2\mu_U^2(1+C_U^2) + \sigma_U^2\right] + \\ 2a(1-a)\sigma_W^2\mu_U\mu_Y \tag{4.4}$$

ed $h = a\mu_W$.

Si noti che per $a = 1/2$ si ottiene $V(\hat{\mu}_{SA1}) = V(\hat{\mu}_{SA})$.

È possibile verificare inoltre, che

$$V(\hat{\mu}_{SA1}) = V(\hat{\mu}_{EH}) + \frac{1}{na^2}\left\{(1-a)^2\left[C_W^2\mu_U^2(1+C_U^2) + \sigma_U^2\right] + 2a(1-a)C_W^2\mu_U\mu_Y\right\} \tag{4.5}$$

ovvero

$$V(\hat{\mu}_{EH}) \le V(\hat{\mu}_{SA1})$$

Quest'ultima scrittura esplicita che, a parità dei parametri, il MSM1 può essere, al meglio, tanto efficiente quanto il MM, in $a \in (0,1]$. In maniera equivalente, potremmo anche dire che il MSM1 garantisce almeno il grado di riservatezza assicurato dal MM.

Nota 4.1 Il valore di a ottimale che minimizza $V(\hat{\mu}_{SA1})$ non è interno all'intervallo $(0,1]$, ma bensì, maggiore di 1. Poiché per $a \to 0$ si ha che $V(\hat{\mu}_{SA1}) \to +\infty$, allora $V(\hat{\mu}_{SA1})$ è una funzione monotona non crescente di $a \in (0,1]$. Di conseguenza per $a < 1/2$, il MSM2 è meno efficiente del MSO, altrimenti, per $a > 1/2$, fornisce uno stimatore più preciso. □

Il MSM2 è la combinazione lineare del MM e il MA. L'i-esimo intervistato estrae due numeri casuali, w_i e u_i, ed in seguito comunica la risposta $z_i = (1-a)w_i y_i + a(y_i + u_i)$, dove $a \in [0,1]$ è una costante fissata dal ricercatore.

$$Z = (1-a)WY + a(Y+U) \tag{4.6}$$

Ponendo $a = 0$, il ricercatore opta per il MM mentre, per $a = 1$, si ottiene il MA. Uno stimatore non distorto di μ_Y è

$$\hat{\mu}_{SA2} = \frac{\bar{z} - a\mu_U}{a + (1-a)\mu_W} \tag{4.7}$$

La varianza $V(\hat{\mu}_{SA2})$ si ottiene particolarizzando la (4.3) per

$$\begin{aligned}\sigma_Z^2 = {} & (1-a)^2 \mu_W^2 \sigma_Y^2 + 2a(1-a)\mu_W \sigma_Y^2 + (1-a)^2 \mu_Y^2 \sigma_W^2 + \\ & (1-a)^2 \left[\sigma_U^2 + \sigma_Y^2 (1 + \sigma_W^2)\right] - 2(1-a)(\sigma_U^2 + \sigma_Y^2) + \sigma_U^2 + \sigma_Y^2\end{aligned} \tag{4.8}$$

ed $h = a + (1-a)\mu_W$.

Nel Capitolo 1, si è verificato analiticamente che il MSO è meno efficiente, sia del MA che del MM. Essendo una combinazione lineare convessa di questi, il MSM2 non è mai meno efficiente della strategia meno precisa, tra il MA ed il MM. Ne consegue allora, che il MSM2 è più efficiente del MSO.

Nota 4.2 La varianza di $\hat{\mu}_{SA2}$ è una funzione convessa di a. Il valore ottimo di a, che minimizza $V(\hat{\mu}_{SA2})$ è

$$a_{opt} = \frac{\sigma_W^2 C_Y^2 (1+\mu_Y^2)}{\mu_W \sigma_U^2 + \sigma_W^2 C_Y^2 (1+\mu_Y^2)} \tag{4.9}$$

Si verifica immediatamente che $a_{opt} \in (0,1)$.

Il limite inferiore della varianza, raggiunto in corrispondenza di a_{opt}, è

$$V(\hat{\mu}_{SA2})_{\min} = \frac{1}{n}\left[\frac{\sigma_U^2\sigma_Y^2 + C_W^2\mu_Y^2(1+C_Y^2)(\sigma_U^2+\sigma_Y^2)}{\sigma_U^2 + C_W^2\mu_Y^2(1+C_Y^2)}\right] \tag{4.10}$$

Valgono inoltre,

$$V(\hat{\mu}_{SA2})_{\min} < V(\hat{\mu}_{AD}) \quad \text{e} \quad V(\hat{\mu}_{SA2})_{\min} < V(\hat{\mu}_{EH})$$

Si osservi tuttavia, che la ricerca del limite inferiore (4.10) è di scarsa utilità nella pratica, dal momento che lo scopo precipuo è di conciliare la protezione della riservatezza con l'efficienza, e non di ottimizzare quest'ultima. □

Nota 4.3 Come abbiamo già sottolineato (cfr. 3.10), se vale

$$\sigma_U^2 < \frac{\sigma_W^2 C_Y^2 \left(1+\mu_Y^2\right)}{\mu_W^2}$$

il MA diviene più efficiente del MM. Allora per $a \in (0,1]$, il MSM2 fornisce uno stimatore più preciso del MM, ma per valori di a inferiori ad a_{opt}, può essere meno efficiente del MA. Viceversa, se la condizione (3.10) non è vera, per $a \in [0,1)$ il MSM2 si presenta più efficiente del MA, ma per valori di a superiori ad a_{opt}, può essere meno preciso del MM. Ovviamente, le configurazioni descritte discendono dalla convessità di $V(\hat{\mu}_{SA2})$ al variare di a.

Inoltre, ricordando l'espressione (4.5) è possibile affermare che

$$V(\hat{\mu}_{SA2}) \leq V(\hat{\mu}_{SA1})$$

se la (3.10) è soddisfatta. □

Mentre il MSM1 non può fornire stime più precise del MM, abbiamo visto che l'efficienza del MSM2 è limitata sia inferiormente che superiormente. Queste considerazioni lasciano intuire che i due sviluppi siano ancora piuttosto rigidi, nonostante l'introduzione del parametro a, finalizzata ad incrementare l'elasticità del MSO. Il passo successivo consiste allora nel considerare un parametro di disegno, p, introdotto per casualizzare le risposte.

Proponiamo quindi un ulteriore modello, il MSM3, il quale raggiunge la flessibilità sperata, ma prevede la possibilità di osservare direttamente Y con probabilità p. È chiaro quindi che il MSM3 può compromettere molto di più la percezione di riservatezza degli intervistati rispetto ai MSM1 e MSM2.

Si generalizza il MSO, mediante l'introduzione di un parametro di disegno $p \in (0,1)$ per casualizzare le risposte. La risposta osservata è

$$Z = \begin{cases} Y & \text{con probabilità } p \\ W(Y+U) & \text{con probabilità } 1-p \end{cases} \tag{4.11}$$

Osserviamo che per p = 0, si ottiene il MSO, mentre per p =1, si applica l'intervista diretta.

Uno stimatore non distorto di μ_Y è

$$\hat{\mu}_{SA3} = \frac{\bar{z} - (1-p)\mu_W \mu_U}{p + (1-p)\mu_W} \tag{4.12}$$

è possibile ricavare $V(\hat{\mu}_{SA3})$ sostituendo nella (4.3)

$$\sigma_Z^2 = p\mu_Y^2(1+C_Y^2) + (1-p)\mu_W^2(1+C_W^2)(\sigma_Y^2 + \mu_Y^2 + \sigma_U^2 + \mu_U^2 + 2\mu_Y\mu_U) - \tag{4.13}$$
$$(p\mu_Y + (1-p)\mu_W(\mu_Y + \mu_U))^2$$

ed $h = p + (1-p)\mu_W$.

4.3 I MSM con variabile ausiliaria

Mantenendo la linea di sviluppo seguita sinora, si introduce nei MSM una variabile ausiliaria, eventualmente codificata.

A prima vista, la scelta di codificare doppiamente X può sembrare inutile dal momento che si dispone già di due variabili di codifica per Y. Tuttavia, reputiamo opportuno considerare anche questa possibilità, al fine di compensare un'eventuale perdita di protezione della privacy rispetto al MSO.

Riportiamo nella *Tabella 4.1*, i MSMV, i MSMVcod, il MSO e, naturalmente l'intervista diretta.

Tabella 4.1 Modelli casualizzati

	R	S	p	c	h
Intervista diretta			1	0	1
MSO		$W(Y+U)$	0	$\mu_W \mu_U$	μ_W
MSM1V	X	$W[aY+(1-a)U]$	0	$(1-a)\mu_W \mu_U$	$a\mu_W$
MSM2V	X	$(1-a)WY+a(Y+U)$	0	$a\mu_U$	$a+(1-a)\mu_W$
MSM3V	X	$W(Y+U)$	$[0,1]$	$(1-p)\mu_W \mu_U$	$p+(1-p)\mu_W$
MSM1Vcod	$T[aX+(1-a)H]$	$W[aY+(1-a)U]$	0	$(1-a)\mu_W \mu_U$	$a\mu_W$
MSM2Vcod	$(1-a)TX+a(X+H)$	$(1-a)WY+a(Y+U)$	0	$a\mu_U$	$a+(1-a)\mu_W$
MSM3Vcod	$T(X+H)$	$W(Y+U)$	$[0,1]$	$(1-p)\mu_W \mu_U$	$p+(1-p)\mu_W$

Allo scopo di confrontare le procedure, definiamo di seguito le quantità utili per calcolare la varianza degli stimatori, e la rispettiva misura di privacy τ (cfr. par. 3.4).

MSM1V.

La variabile sensibile è rilevata con la procedura MSM1, mentre la variabile ausiliaria è priva di codifica. Le espressioni di media e varianza di Z sono $\mu_Z = h\mu_Y + c$, dove $h = a\mu_W$ e $c = (1-a)\mu_W\mu_U$, mentre σ_Z^2 è data dalla (4.4). Per la variabile ausiliaria si ha $\mu_V = \mu_X$ e $\sigma_V^2 = \sigma_X^2$ e le covarianze sono $\sigma_{VZ} = a\mu_W\sigma_{XY}$, $\sigma_{YV} = \sigma_{XY}$ e $\sigma_{YZ} = a\mu_W\sigma_Y^2$.

MSM2V.

La variabile sensibile è codificata come previsto secondo il MSM2, mentre la variabile ausiliaria è osservata direttamente. Media e varianza di Z sono $\mu_Z = h\mu_Y + c$, dove $h = a + (1-a)\mu_W$ e $c = a\mu_U$, e σ_Z^2 è data dalla (4.8). Per media e varianza di V e σ_{YV} ci si riferisca al MSM1V, mentre, per quanto riguarda le covarianze si ha $\sigma_{VZ} = [a + (1-a)\mu_W]\sigma_{XY}$, e $\sigma_{YZ} = [a + (1-a)\mu_W]\sigma_Y^2$.

MSM3V.

La variabile sensibile prevede la codifica MSM3, mentre la variabile ausiliaria non è codificata. In merito a Z si trova che $\mu_Z = h\mu_Y + c$, dove $h = p + (1-p)\mu_W$, mentre $c = (1-p)\mu_W\mu_U$, inoltre σ_Z^2 si può ricavare dalla (4.13). Per media e varianza di V e σ_{YV}, si veda il MSM1V; le covarianze sono $\sigma_{VZ} = [p + (1-p)\mu_W]\sigma_{XY}$ e $\sigma_{YZ} = [p + (1-p)\mu_W]\sigma_Y^2$.

MSM1Vcod.

Sia la variabile sensibile che la variabile ausiliaria sono rilevate con la codifica prevista dal MSM1. Per media e varianza di Z e la covarianza tra Y e Z, si veda il MSM1V. In merito a V si ha $\mu_V = h'\mu_X + c'$, dove $h' = a\mu_T$ e $c' = (1-a)\mu_T\mu_H$, mentre la varianza è

$$\sigma_V^2 = a^2\mu_T^2\left[\sigma_X^2 + \mu_X^2(1+C_X^2)C_T^2\right] + (1-a)^2\mu_T^2\left[C_T^2\mu_H^2(1+C_H^2) + \sigma_H^2\right] + 2a(1-a)\sigma_T^2\mu_H\mu_X$$

Le rimanenti covarianze sono $\sigma_{VZ} = a^2\mu_W\mu_T$, $\sigma_{YV} = a\mu_T\sigma_{XY}$.

MSM2Vcod.

La variabile sensibile, come la variabile ausiliaria, è rilevata mediante la codifica prevista dal MSM2. Per quanto riguarda μ_Z, σ_Z^2 e σ_{YZ} si veda il MSM2, mentre per V si trova che $\mu_V = h'\mu_X + c'$, dove $h' = a + (1-a)\mu_T$ e $c' = a\mu_H$. L'espressione di σ_V^2 è

$$\sigma_V^2 = (1-a)^2\mu_T^2\sigma_X^2 + 2a(1-a)\mu_T\sigma_X^2 + (1-a)^2\mu_X^2\sigma_T^2 + (1-a)^2\left[\sigma_H^2 + \sigma_X^2(1+\sigma_T^2)\right] - 2(1-a)(\sigma_H^2 + \sigma_X^2) + \sigma_H^2 + \sigma_X^2$$

Inoltre, le covarianze risultano $\sigma_{VZ} = \sigma_{XY}\left[(1-a)^2\mu_T\mu_W + a(1-a)(\mu_W + \mu_T) + a^2\right]$, $\sigma_{YV} = h'\sigma_{XY}$.

MSM3Vcod.

Sia la variabile sensibile che la variabile ausiliaria, prevedono la codifica della procedura MSM3. In merito a media e varianza di Z si rimanda al MSM3V,

mentre per quanto riguarda *V* abbiamo $\mu_V = h'\mu_X + c'$, dove $h' = p + (1-p)\mu_T$ e $c' = (1-p)\mu_T\mu_H$; la varianza è

$$\sigma_V^2 = p\mu_X^2(1+C_X^2) + (1-p)\mu_T^2(1+C_T^2)(\sigma_X^2 + \mu_X^2 + \sigma_H^2 + \mu_H^2 + 2\mu_X\mu_H) - [p\mu_X + (1-p)\mu_T(\mu_X + \mu_H)]^2$$

Le rimanenti covarianze sono $\sigma_{YV} = [p + (1-p)\mu_T]\sigma_{XY}$, e

$$\sigma_{VZ} = (\mu_X\mu_Y + \sigma_{XY})[p + (1-p)\mu_W\mu_T] + (1-p)(\mu_W\mu_Y\mu_T\mu_H + \mu_W\mu_U\mu_T\mu_X + \mu_W\mu_U\mu_T\mu_H) - [p\mu_X + (1-p)\mu_T(\mu_X + \mu_H)][p\mu_Y + (1-p)\mu_W(\mu_Y + \mu_U)]$$

4.4 Efficienza e protezione della privacy

L'enfasi posta in letteratura sullo studio dell'efficienza, non ha portato ad approfondire le potenzialità dei modelli misti, dal momento che l'alto grado di riservatezza che garantiscono è pagato in termini di precisione dello stimatore. Ci si può allora ragionevolmente chiedere, se sia possibile raggiungere un opportuno equilibrio modificando il MSO. Pertanto, si presentano ora dei confronti che descrivono il comportamento dei MSM(V(cod)), utilizzando, come misure, l'efficienza relativa (3.9) e τ (3.3), entrambe al variare di $a \in [0,1]$.

I nostri obiettivi, in merito all'efficienza, sono:

- stabilire se i MSMV(cod) possono essere più efficienti dell'intervista diretta.
- valutare la precisione dello stimatore di ciascun MSM(V(cod)) rispetto al MSO, e confrontarli tra loro.

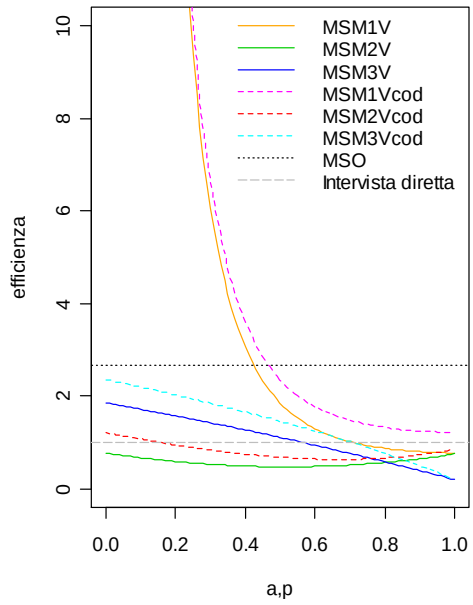
I confronti in termini di protezione della privacy sono volti a:

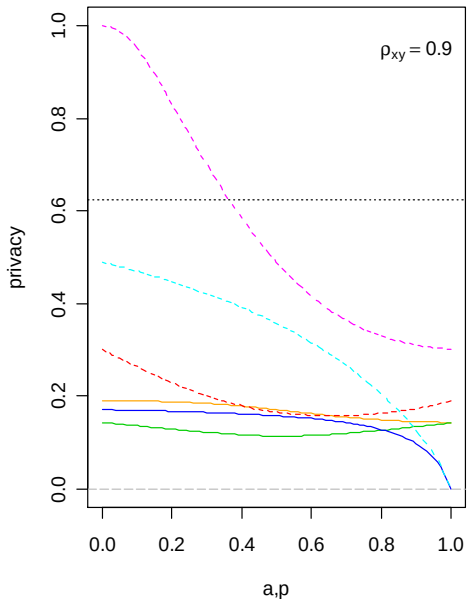
- esaminare il livello di protezione garantito dai MSMV(cod) rispetto all'intervista diretta.
- valutare il livello di riservatezza garantito da ciascun MSM(V(cod)) rispetto al MSO, e confrontarli tra loro.

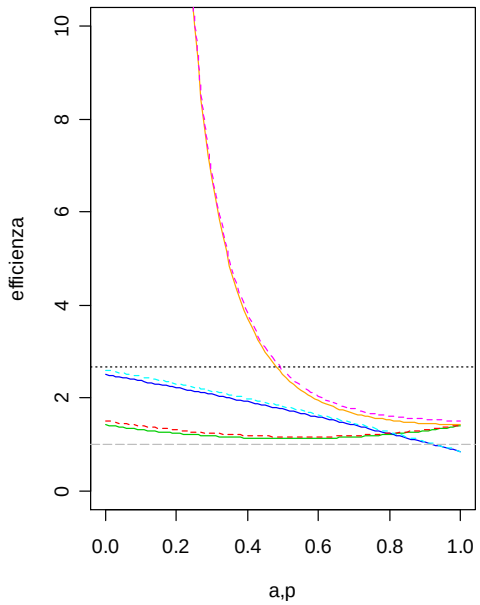
A causa della complessità dei confronti analitici, lo studio si sviluppa a livello grafico. I valori dei parametri sia della variabile sensibile, che della variabile ausiliaria, sono i medesimi del Capitolo 3, come anche la distribuzione di H,U e di T,W è la F di Fisher, assunta nel *Caso 1* con (5,5) e (5,50) g.d.l. a cui corrispondono, rispettivamente, $C_H = C_U = 1.789$ e $C_T = C_W = 0.679$. Nel *Caso 2*, sono stati posti (1,5) g.d.l. per H,U e (5,50) g.d.l. per T,W ed i rispettivi coefficienti di variazione sono $C_H = C_U = 2.828$ e $C_T = C_W = 0.679$. La scelta effettuata sui gradi di libertà vuole creare delle situazioni il più possibile rappresentative, e allo stesso tempo, vantaggiose per la ricerca di un equilibrio tra efficienza e protezione della privacy.

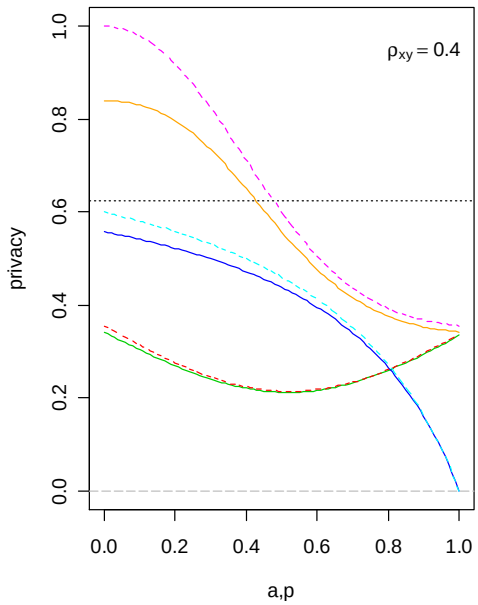
Per brevità di esposizione, riportiamo ora solo i grafici relativi a $\rho_{XY} = 0, 0.4, 0.9$. L'analisi completa con $(\rho_{XY} = 0.0, \mathbf{0.2}, 0.4, \mathbf{0.7}, 0.9)$ è illustrata nell'Appendice B.

Caso 1. $H, U \sim F(5,5) \quad C_H = C_U = 1.789 \quad T, W \sim F(5,50) \quad C_T = C_W = 0.679$

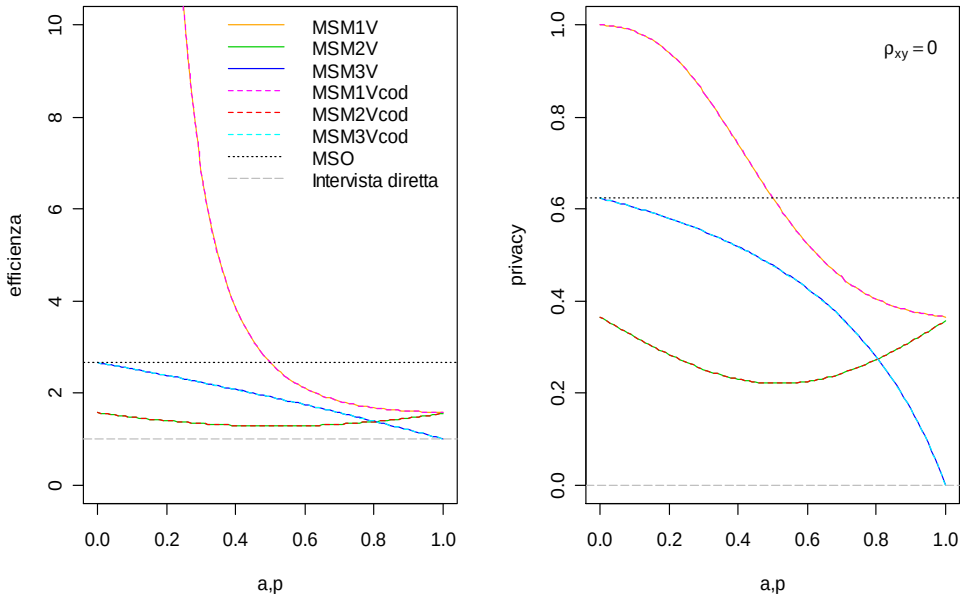




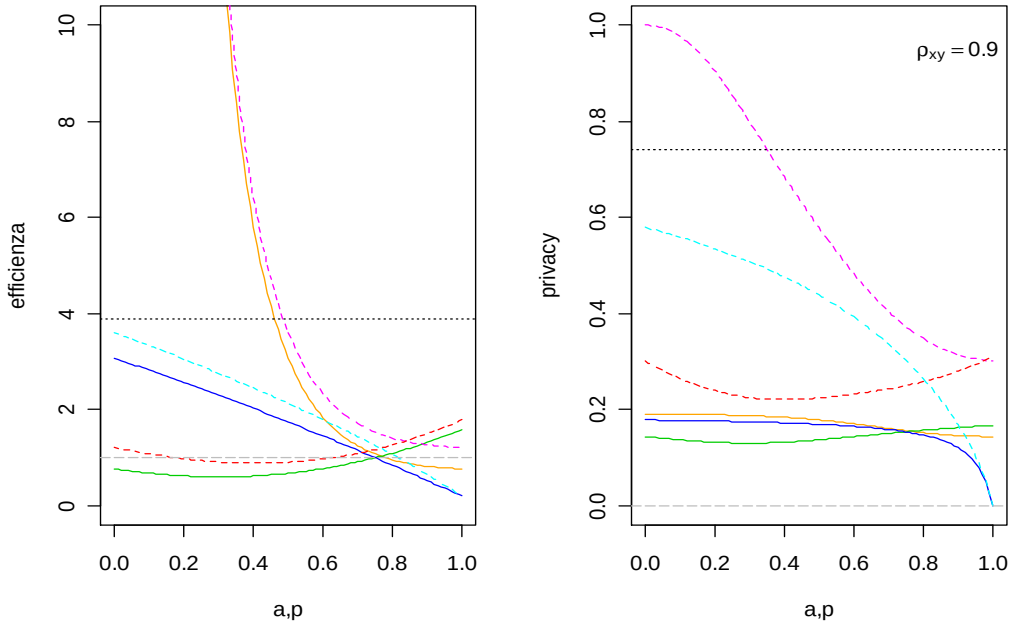




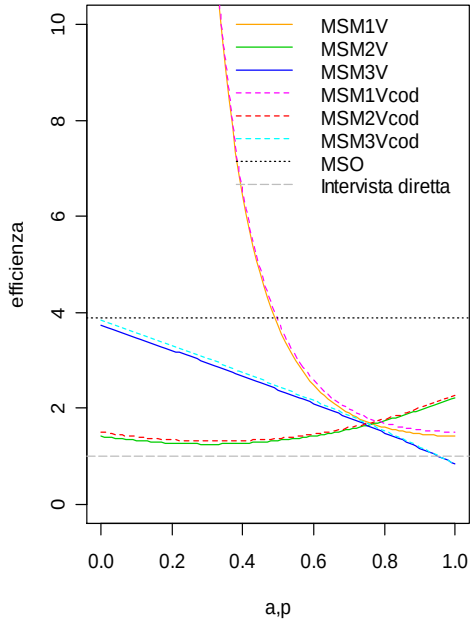
Caso 1. $H,U \sim F(5,5)$ $C_H = C_U = 1.789$ $T,W \sim F(5,50)$ $C_T = C_W = 0.679$

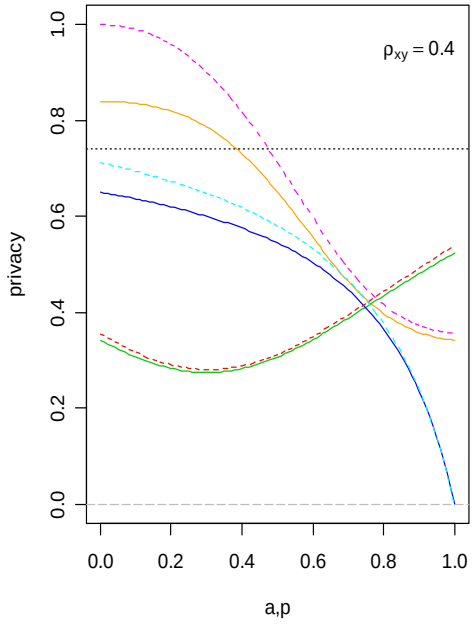


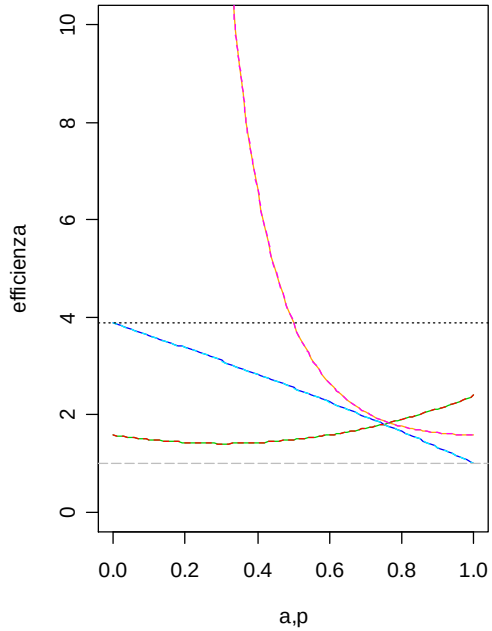
Caso 2. $H,U \sim F(1,5)$ $C_H = C_U = 2.828$ $T,W \sim F(5,50)$ $C_T = C_W = 0.679$

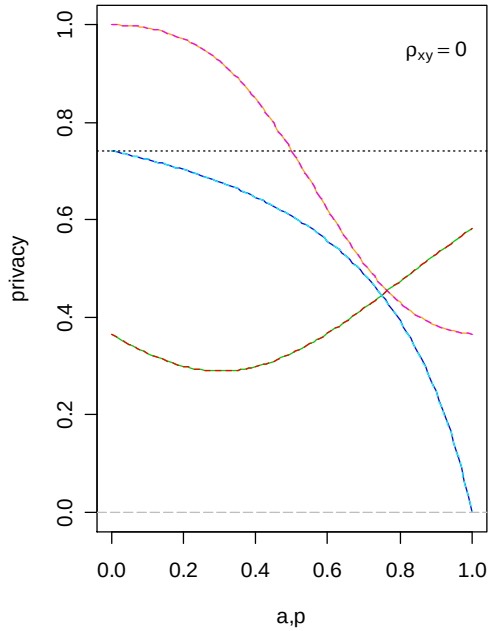


Caso 2. $H, U \sim F(1,5)$ $C_H = C_U = 2.828$ $T, W \sim F(5,50)$ $C_T = C_W = 0.679$









Precisiamo, prima di tutto, che a e p svolgono dei ruoli nettamente distinti nei modelli in cui sono introdotti, incidendo diversamente nel computo della varianza di stima. In ciascun grafico, il confronto tra i modelli può dunque avvenire per valori diversi di a e di p.

Come da attendersi, i MSMV(cod) possono fornire uno stimatore più preciso di $\bar{y}$ e, naturalmente, l'efficienza aumenta, tanto più è forte la correlazione tra X ed Y. Poiché, nei grafici illustrati, l'efficienza relativa del MSM3V(cod) è una funzione decrescente di p, è ragionevole pensare che il suo eventuale vantaggio rispetto all'intervista diretta, sia il frutto di due elementi compresenti, la variabile ausiliaria ed il parametro di disegno p. Al contrario, la possibile superiorità del MSM1V(cod) rispetto all'intervista diretta, è attribuibile interamente all'introduzione della variabile ausiliaria.

I grafici illustrano che i MSM1Vcod non sono più efficienti dei MSM3(V(cod)) ma, possono fornire stimatori più precisi dei MSM2(V(cod)) quando la (3.10) non è soddisfatta (*Caso 2*). I MSM3(V(cod)) possono essere più efficienti dei MSM2(V(cod)), ma solo per valori di p elevati, sconsigliati nei casi pratici. I MSM2(V(cod)), come anche i MSM3(V(cod)), forniscono sempre uno stimatore più preciso del MSO, mentre i MSM1(V(cod)) sono più efficienti della versione originale quando $a < 1/2$.

Consideriamo ora la protezione della privacy. In corrispondenza di p elevati, il MSM3(V(cod)) assicura meno riservatezza dei MSM1(V(cod)) e MSM2(V(cod)). Si nota poi, che per valori di p proponibili nella pratica, il MSM3Vcod tutela maggiormente del MSM2Vcod, ma meno del MSM1Vcod. Mentre, per valori di a grandi, il MSM2(V(cod)) si presenta più protettivo del MSM1(V(cod)) quando la (3.10) non è vera.

L’unico modello capace di garantire maggiore riservatezza del MSO è il MSM1(V(cod)) quando $a < 1/2$, ma per tali valori del parametro la precisione dello stimatore è gravemente carente.
Si osservi che piccole diminuzioni in termini di efficienza possono migliorare considerevolmente il livello di protezione della privacy. Si potrebbe pensare allora di raggiungere un buon equilibrio grazie ai MSMVcod, dal momento che la codifica di X implica esigue diminuzioni di precisione, acquistando un aumento, anche notevole, del grado di riservatezza.

4.5 Conclusioni

In merito ai confronti di efficienza si è riscontrato che:

- la versione di Saha originale è sempre svantaggiosa dei modelli modificati, fuorchè rispetto al MSM1(V(cod)) per $a < 1/2$.
- introducendo una variabile ausiliaria, anche codificata, le tecniche proposte possono essere più efficienti dell’intervista diretta.
- nel *Caso 1*, il MSM1(V(cod)) è più sconveniente rispetto alle altre due procedure modificate.
- nel *Caso 2*, il MSM1(V(cod)) può essere più vantaggioso del MSM2(V(cod)).
- il MSM2(V(cod)) è maggiormente efficiente del MSM3(V(cod)), tranne che per valori di p elevati, sconsigliati nella pratica.

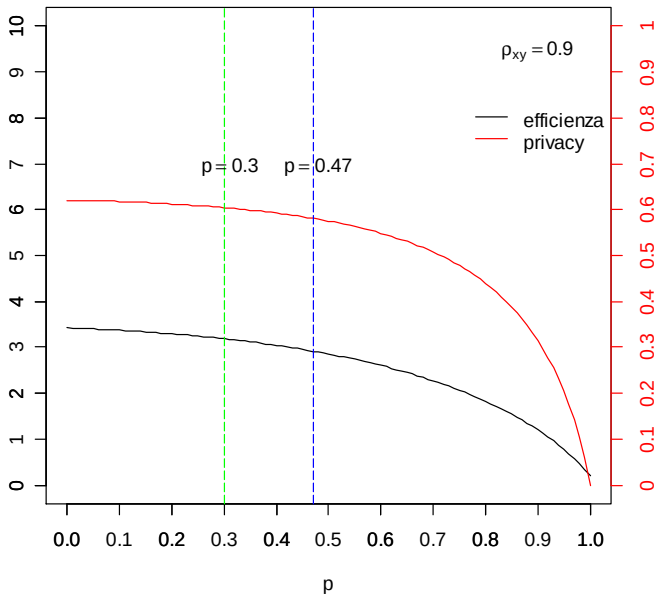
Per quanto riguarda la tutela della privacy, si è osservato che:

- il modello di Saha originale assicura meno riservatezza solo se paragonato al MSM1(V(cod)) per $a < 1/2$.
- nel *Caso 1*, il MSM1Vcod è il più protettivo tra le tecniche modificate.
- nel *Caso 2*, il MSM1Vcod può garantire meno protezione del MSM2(V(cod)).
- il MSM3(V(cod)) è svantaggioso rispetto alle altre procedure modificate quando i valori di p sono alti.

Ribadiamo che i casi illustrati non individuano degli ordinamenti definitivi, ma sono piuttosto finalizzati a delineare una possibile gerarchia tra i modelli. Il contributo dei grafici ci sembra di primaria utilità, qualora si vogliano valutare dei nuovi sviluppi, come i MSM(V(cod)). Si è visto che le procedure sono dotate di una buona flessibilità capace conciliare riservatezza e precisione dello stimatore. Per aumentare l'efficienza, mantenendo un equilibrio soddisfacente, è possibile optare per i MSMVcod, più precisi dei MSM, e più indicati a tutelare la privacy dei MSMV. Si ritiene infatti che i MSMV siano adeguati solo se la variabile supplementare introdotta è poco correlata al carattere sensibile.
In presenza di correlazione superiore a 0.4, i MSMVcod appaiono estremamente convenienti, garantendo una buona protezione della privacy, senza pregiudicare l'efficienza dello stimatore.
In particolare, il MSM3Vcod permette di raggiungere la flessibilità desiderata, anche se questa è parzialmente dovuta alla possibilità di rilevare direttamente Y.
Al contrario, i MSM1(V(cod)) sono sicuramente i più rigidi, perché ancora troppo legati all'idea originale di Saha.

Nei confronti discussi si sono utilizzate due variabili di codifica diverse, tuttavia è possibile semplificare la rilevazione adottandone una sola: riteniamo che una buona soluzione consista nell'estrarre, indipendentemente, due numeri casuali dalla medesima distribuzione. In questo modo, la percezione della privacy dei rispondenti non ne risente troppo. Inoltre, preferendo la distribuzione F di Fisher è possibile conseguire una notevole flessibilità, anche senza differenziare i parametri di H,U e T,W.

Si descrive ora una traccia per il ricercatore che vuole adottare il MSM3Vcod allo scopo di sfruttarne l'elasticità riscontrata. Assumiamo $\mu_X = 25$, $\mu_Y = 500$, $C_X = 2$, $C_Y = 15$ e $\rho_{XY} = 0.9$: tali valori sono relativi ad una popolazione già considerata in letteratura, ad esempio, in Gupta e Shabbir (2008).
Come distribuzione delle variabili di codifica H,U e T,W scegliamo la F di Fisher, rispettivamente, con (5,5) e (10,5) g.d.l.. Il grafico riportato illustra la situazione suddetta.



La principale utilità del grafico consiste nel descrivere in maniera semplice e diretta il rendimento del modello al variare del parametro. A ciascun valore di p, corrisponde un livello di efficienza e, conseguentemente, uno di protezione della privacy. Poiché non tutte le combinazioni sono proponibili nella pratica, il ricercatore è tenuto a selezionarne una, tra le possibili, ritenuta soddisfacente su entrambi i fronti.

È ragionevole scartare da subito valori di p superiori a 0.5, in modo da non compromettere troppo la percezione di riservatezza dei rispondenti. Si osservi inoltre che per $p \in (0,0.5)$ il livello di protezione della privacy non subisce un notevole abbassamento, e può dunque esser lecito escludere anche valori inferiori a 0.3, orientandosi così verso un'efficienza relativa non superiore a 3.5. Quando p è uguale a 0.3, si assicura una protezione del 60%, ed un'efficienza relativa intorno a 3.2. Per migliorare la precisione, è sufficiente posizionarsi verso dei valori di p più grandi come, ad esempio, 0.47. In tal modo, si ottiene uno stimatore del 2.7% più preciso abbassando, di conseguenza, la protezione del 2.3%. Quest'ultima opzione appare la più soddisfacente, raggiungendo un'efficienza accettabile e mantenendo, al tempo stesso, un livello di protezione relativamente alto.

4.6 Nota bibliografica

Per approfondimenti si rimanda agli articoli, già citati, di Eichhorn e Hayre (1983) e di Saha (2007), dove è trattato estesamente il modello originale di Saha.
I modelli di Saha modificati sono stati sviluppati nell'ambito di questo lavoro.

Conclusioni

I risultati ottenuti in questa tesi mettono in luce l'importante ruolo che può rivestire l'informazione ausiliaria nella rilevazione di un carattere sensibile.

L'introduzione di un modello generale di codifica ha permesso di esprimere le principali strategie casualizzate proposte in letteratura, basate su una codifica di tipo additivo, moltiplicativo, o a domande incorrelate. In riferimento a tale modello, è stata poi costruita una classe generale di stimatori per la media della variabile sensibile, la quale include accanto ad alcuni stimatori presenti in letteratura, gli usuali stimatori basati su una variabile ausiliaria a media nota.

È stato ottenuto lo stimatore ottimale a varianza minima per la classe, il quale è uno stimatore per regressione, almeno tanto efficiente quanto lo stimatore corrispondente senza variabile ausiliaria.

In linea con le tendenze più recenti, si è ritenuto opportuno valutare congiuntamente i due aspetti cruciali delle procedure casualizzate, l'efficienza e la protezione della privacy. L'approccio seguito ha quindi delineato, sin dall'inizio, il nostro obiettivo di conciliarli, nella consapevolezza che massimizzare privacy ed efficienza allo stesso tempo è un problema irrisolubile.

In coerenza con altre formulazioni in letteratura, è stata introdotta una nuova misura di privacy generalizzata valida anche quando si utilizza l'informazione supplementare.

I confronti grafici hanno confermato l'assoluta convenienza ad utilizzare una variabile ausiliaria, al fine di migliorare l'efficienza dello stimatore, senza compromettere considerevolmente la riservatezza. Ad ogni modo, il risultato più rilevante consiste nel fatto che le strategie basate su una variabile ausiliaria possono essere più efficienti dell'intervista diretta.

Focalizzando l’attenzione sulle strategie di tipo additivo e moltiplicativo, si è constatata la mancanza di un modello preferibile a priori, sia per l’efficienza, che per la tutela della privacy. Infatti, è la scelta della distribuzione, e dei relativi parametri della variabile di codifica, che muta i rendimenti dei modelli. Grazie alla condizione analitica che assicura la maggiore efficienza del modello additivo, rispetto al moltiplicativo, abbiamo sviluppato un’analisi numerica allo scopo di enfatizzare la molteplicità delle situazioni che ricorrono nella pratica.

La ricerca di un ragionevole compromesso tra protezione della privacy ed efficienza, ci ha indotto poi a formulare tre tecniche miste ispirate al modello di Saha, migliorandone l’efficienza e l’adattabilità ai casi reali.

Come da attendersi, si è appurato graficamente che i nuovi sviluppi sono più elastici, e possono essere più efficienti dell’intervista diretta quando si usa una variabile ausiliaria.

Infine, abbiamo utilizzato un dataset presente in letteratura, allo scopo di esemplificare l’applicazione di un modello di Saha modificato, basato su una variabile ausiliaria codificata. Consentendo di sfruttare al massimo l’informazione disponibile e di coniugare al meglio esigenze di efficienza e di protezione della privacy, la tecnica proposta può costituire un valido strumento di rilevazione.

Appendice A

Confronti di efficienza e di protezione della privacy

In questa appendice si presentano dettagliatamente i confronti effettuati nel corso del lavoro riguardanti le 6 strategie discusse nel par. 3.4.

Con l'intento di configurare un insieme di situazioni, il più possibile esaustivo, sono stati considerati vari set di valori dei parametri per differenti distribuzioni delle variabili di codifica.

I valori dei parametri sono stati scelti, ogni volta, al fine di soddisfare (*Caso 1*), o violare (*Caso 2*), la condizione (3.10). Completiamo anzitutto, la selezione di grafici già riportata nel Capitolo 3, considerando $(\rho_{XY} = 0.0,\ \mathbf{0.2},\ 0.4,\ \mathbf{0.7}, 0.9)$. Ricordiamo che in entrambi i casi si assume $\mu_X = 10$ e $\mu_Y = 2$, $C_X = 0.8$ e $C_Y = 2$. Per il primo, si ha che H ed U si distribuiscono come una F di *Fisher* con (5,5) g.d.l., mentre T e W con (10,5) g.d.l., ed i coefficienti di variazione sono $C_H = C_U = 1.789$ e $C_T = C_W = 1.612$. Nel secondo, la distribuzione di H ed U è caratterizzata da (1,5) g.d.l., mentre, quella di T e W da (10,50) g.d.l., con $C_H = C_U = 2.828$ e $C_T = C_W = 0.502$.

È stata poi considerata la distribuzione *Lognormale*, assumendo in entrambi i casi $\mu_X = 10$, $\mu_Y = 2$, $C_X = 2$ e $C_Y = 2$. Nel *Caso 1* si è posto $H, U \sim Ln(0.8,1)$ e $T, W \sim Ln(0.5,1)$, con $C_H = C_U = 1.311$ e $C_T = C_W = 1.311$. Invece, nel *Caso 2*, si sono scelte $H, U \sim Ln(1.8,0.8)$ e $T, W \sim Ln(0.5,0.4)$, per le quali $C_H = C_U = 1.107$ e $C_T = C_W = 0.701$. Di seguito, è stata esaminata la distribuzione *Gamma*, supponendo nel *Caso 1*, come nel *2*, $\mu_X = 100$, $\mu_Y = 2$, $C_X = 2$ e $C_Y = 2$. Nel

primo si è assunto $H,U \sim Gamma(0.5,0.2)$ e $T,W \sim Gamma(0.4,0.2)$, con $C_H = C_U = 1.414$ e $C_T = C_W = 1.581$. Nel secondo, $H,U \sim Gamma(3,0.2)$ e $T,W \sim Gamma(0.4,2)$, con $C_H = C_U = 0.577$ e $C_T = C_W = 1.581$. È stata infine valutata la distribuzione di *Poisson*, adatta alla codifica di una variabile sensibile quantitativa discreta. Nei due casi, abbiamo posto $\mu_X = 50$, $\mu_Y = 2$, $C_X = 0.8$, $C_Y = 2$; nel primo, sono state scelte $H,U \sim Pois(10)$ e $T,W \sim Pois(0.8)$ per cui, rispettivamente, si ha $C_H = C_U = 0.316$ e $C_T = C_W = 1.118$; mentre nel secondo, $H,U \sim Pois(8)$ e $T,W \sim Pois(4)$ con $C_H = C_U = 0.354$ e $C_T = C_W = 0.5$.

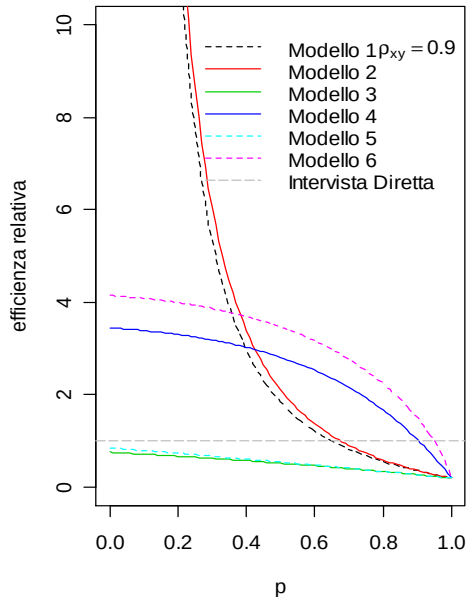
I risultati ottenuti assumendo le distribuzioni *Gamma* e *Lognormale* confermano gli ordinamenti ottenuti con la F di *Fisher*. Scegliendo invece, la distribuzione di *Poisson* si sono riscontrate alcune variazioni che illustriamo brevemente.

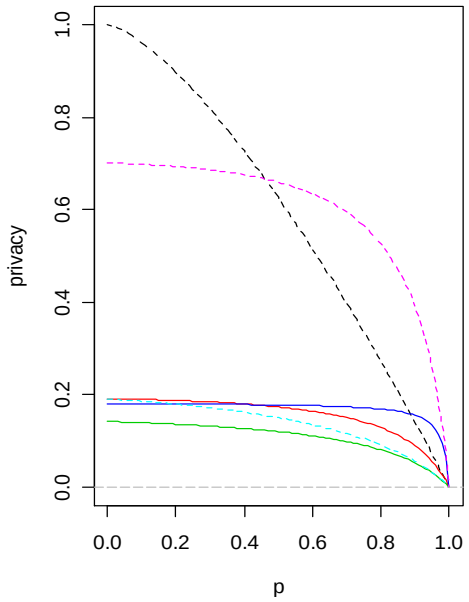
Le differenze trovate dal punto di vista dell'<u>efficienza</u> sono:

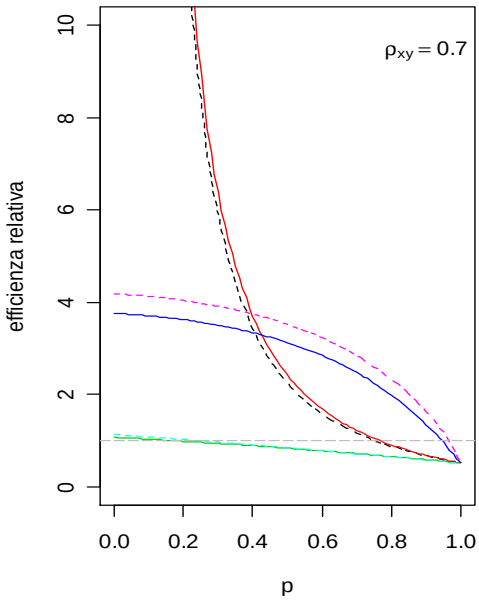
- l'efficienza relativa delle strategie di tipo additivo (Modelli 3 e 5) presenta un andamento concavo (*descrescente in* p). Di conseguenza, nel *Caso 1*, le procedure moltiplicative (Modelli 4 e 6) possono essere più vantaggiose delle additive.
- nel *Caso 2*, l'efficienza relativa dei modelli moltiplicativi (Modelli 4 e 6) mostra una concavità (*descrescente in* p). Di conseguenza, le strategie additive (Modelli 3 e 5) possono fornire uno stimatore più preciso delle moltiplicative.
- il Modello 5 (con variabile ausiliaria codificata) è più efficiente (*meno efficiente*) del Modello 3 (con variabile ausiliaria osservata direttamente).

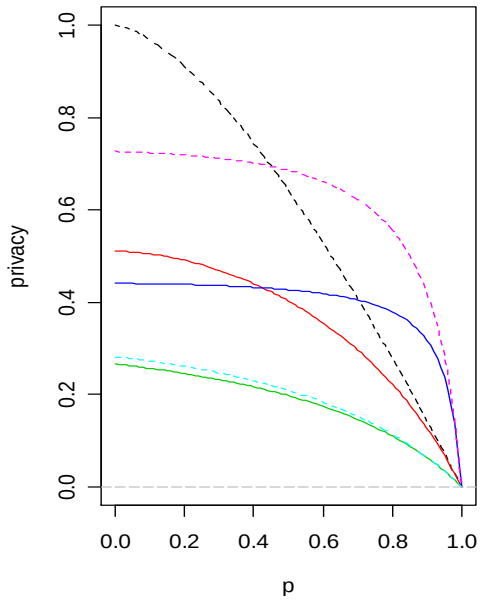
Ovviamente, i cambiamenti relativi alla protezione della <u>privacy</u> sono speculari a quelli appena descritti.

Caso 1. $H, U \sim F(5,5)$ $T, W \sim F(10,5)$

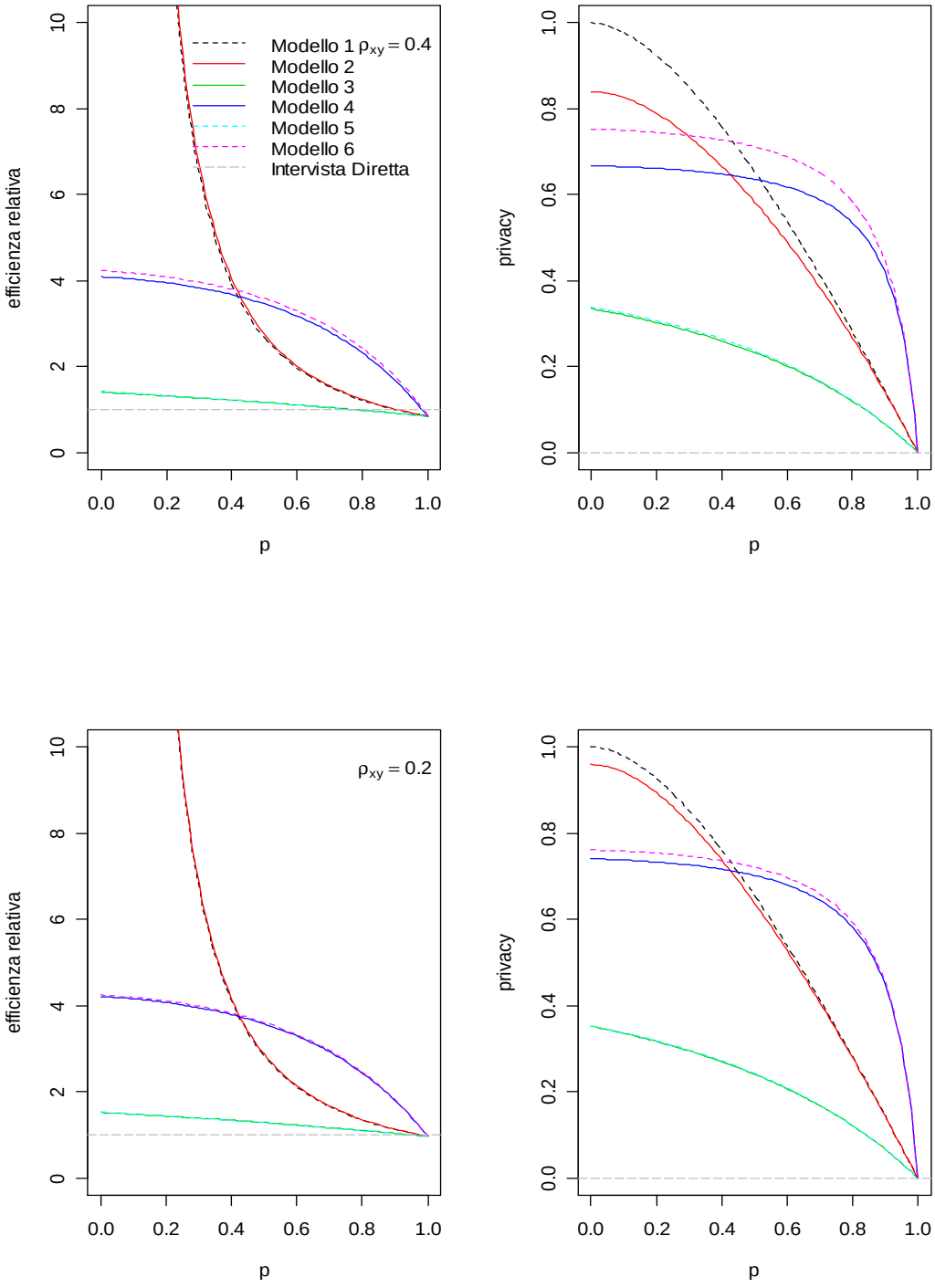




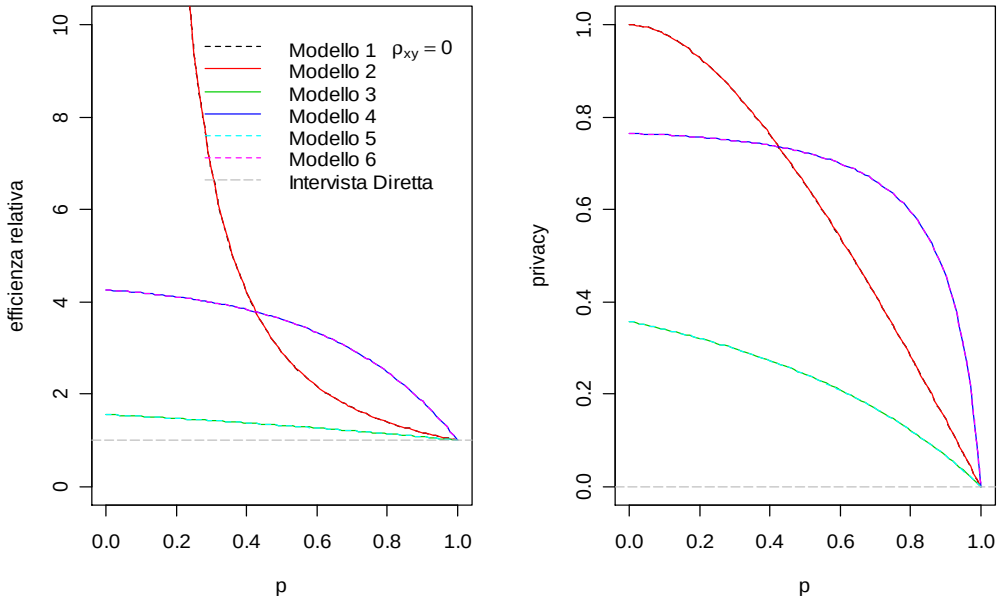




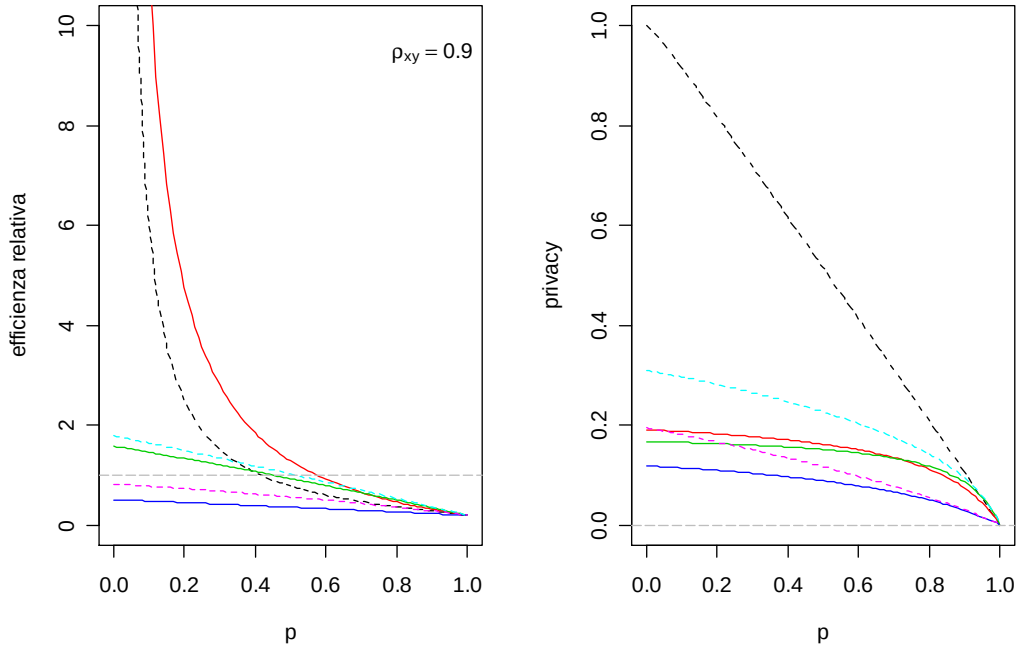
Caso 1. $H, U \sim F(5,5) \qquad T, W \sim F(10,5)$



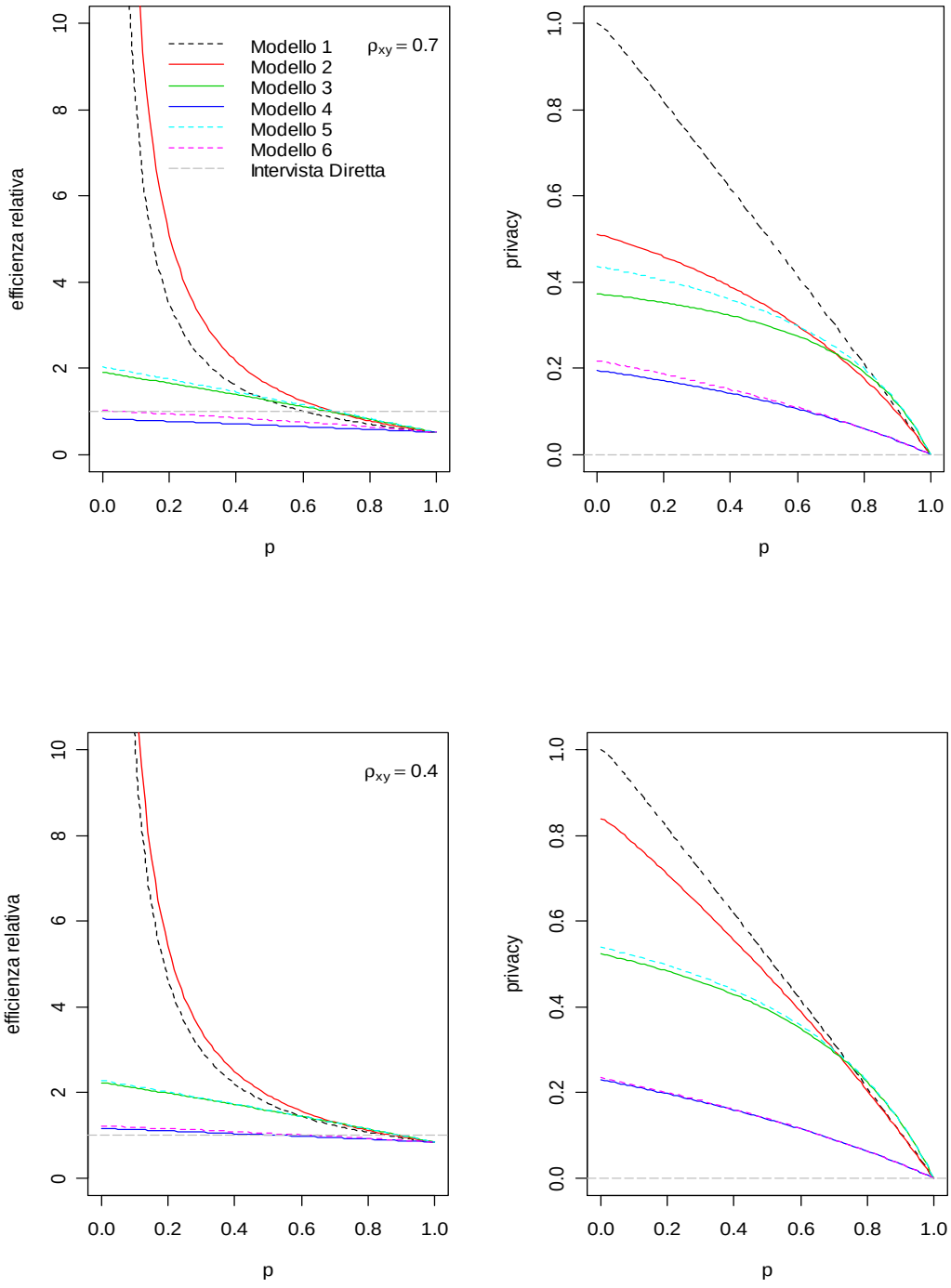
Caso 1. $H,U \sim F(5,5)$ $T,W \sim F(10,5)$



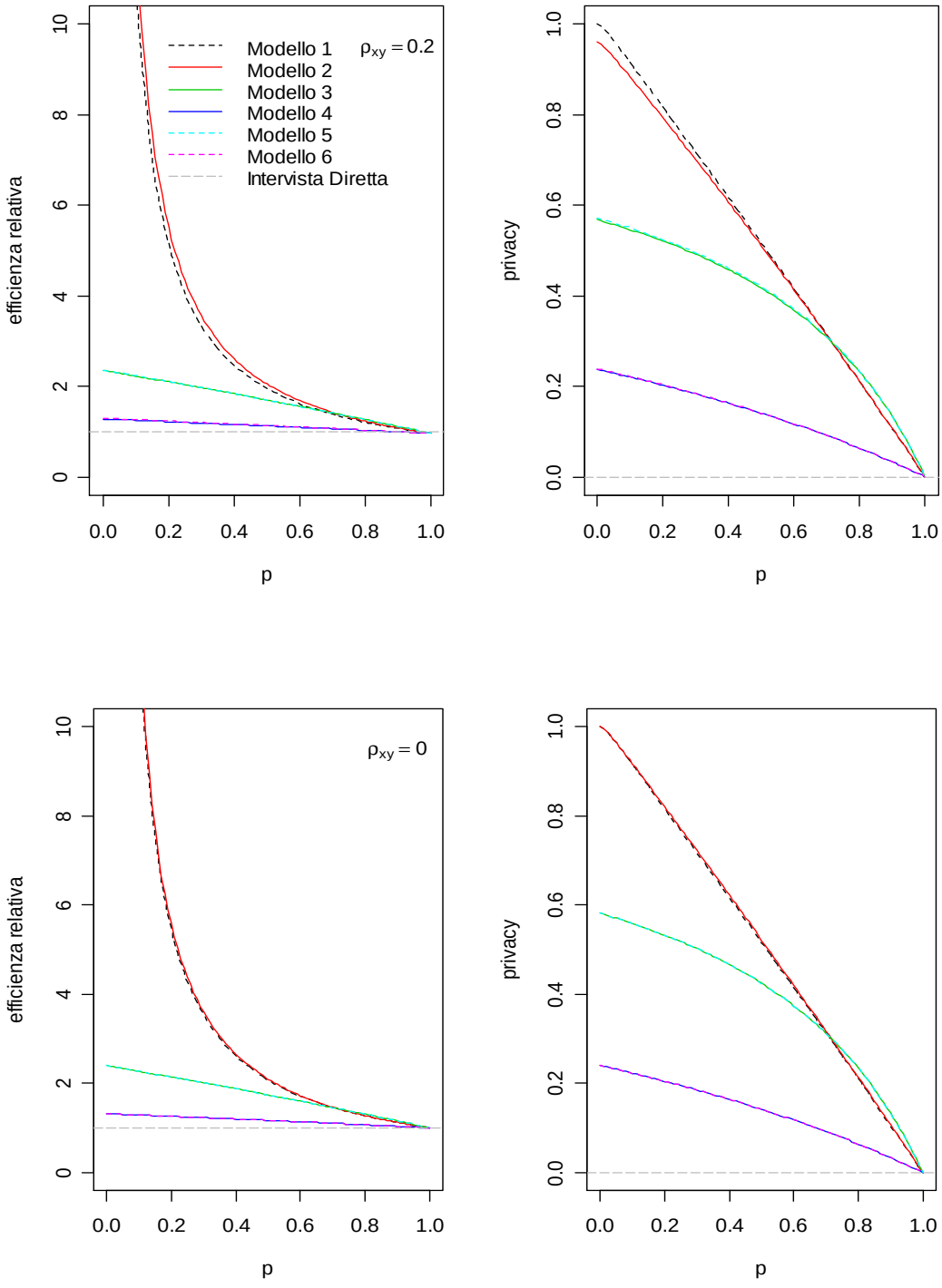
Caso 2. $H,U \sim F(1,5)$ $T,W \sim F(10,50)$



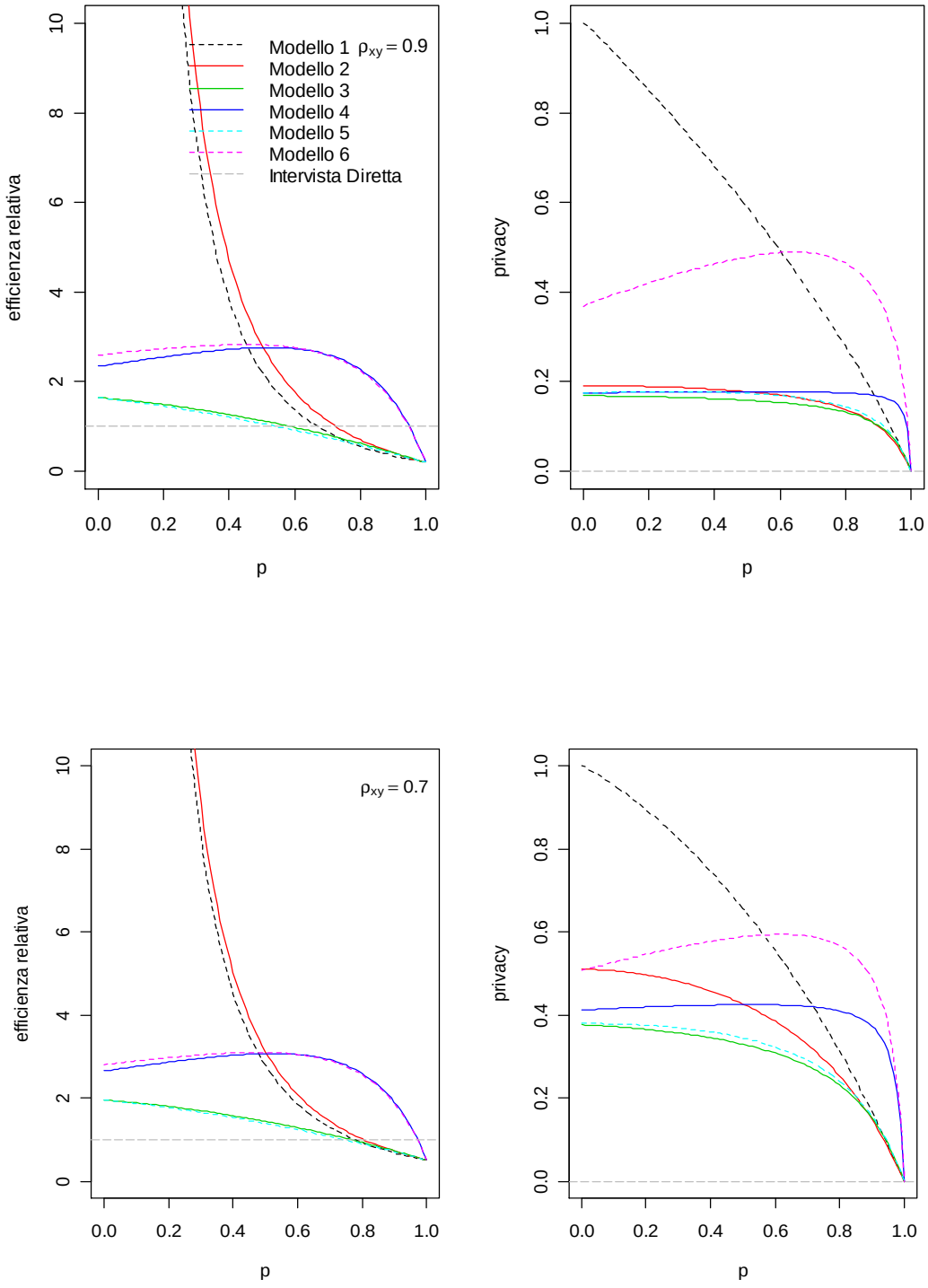
Caso 2. $H, U \sim F(1,5) \qquad T, W \sim F(10,50)$



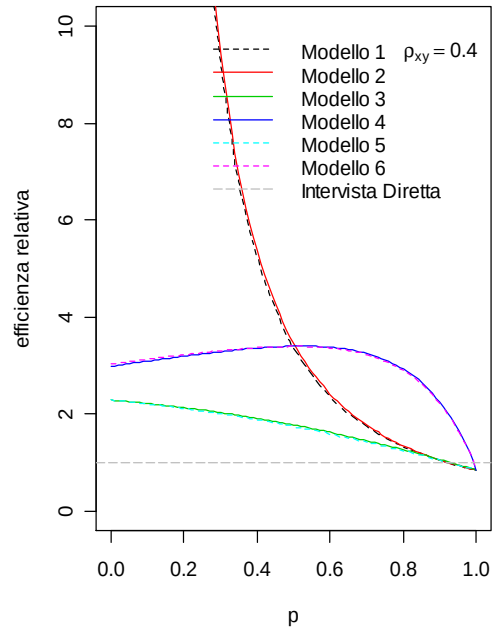
Caso 2. $H, U \sim F(1,5) \quad T, W \sim F(10,50)$

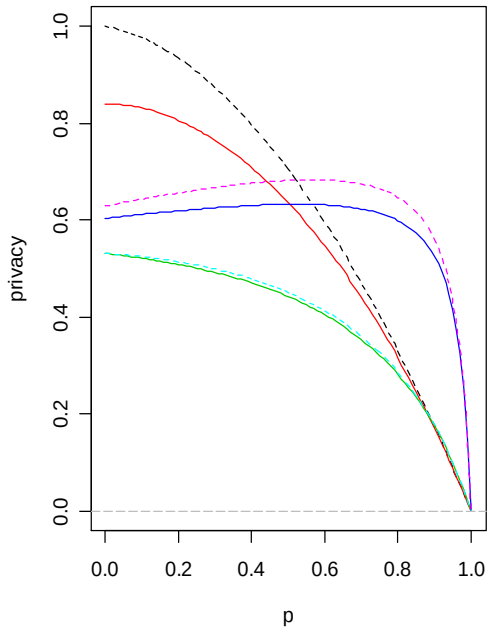


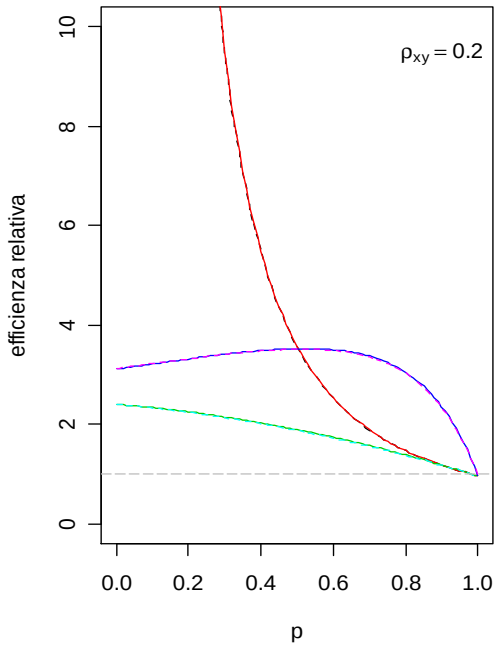
Caso 1. $H, U \sim Ln(0.8,1)$ $T, W \sim Ln(0.5,1)$

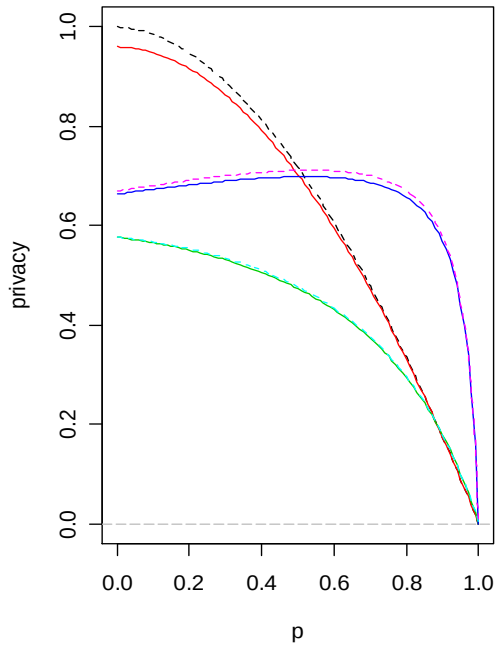


Caso 1. $H, U \sim Ln(0.8,1)$ $T, W \sim Ln(0.5,1)$

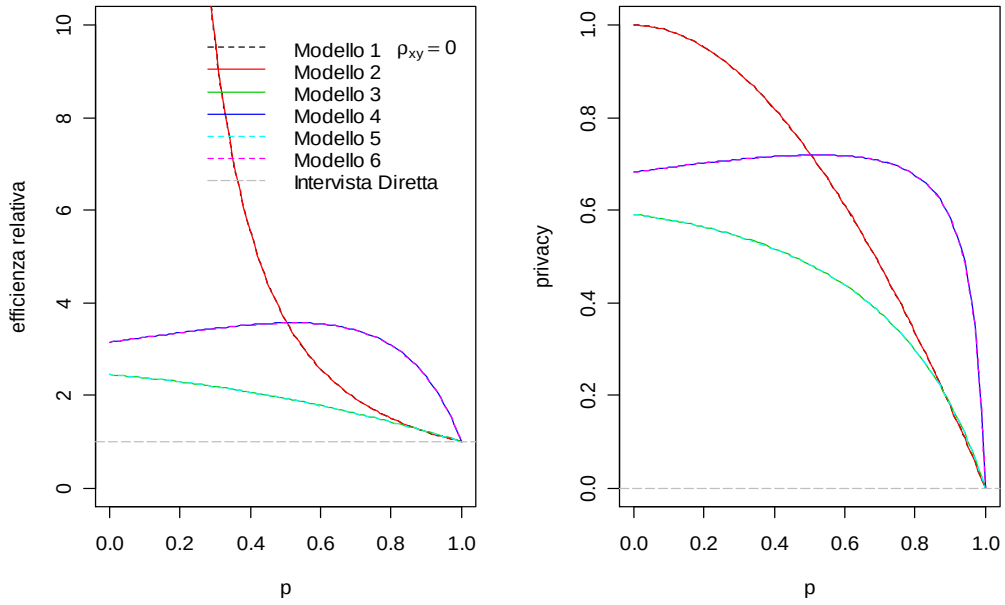




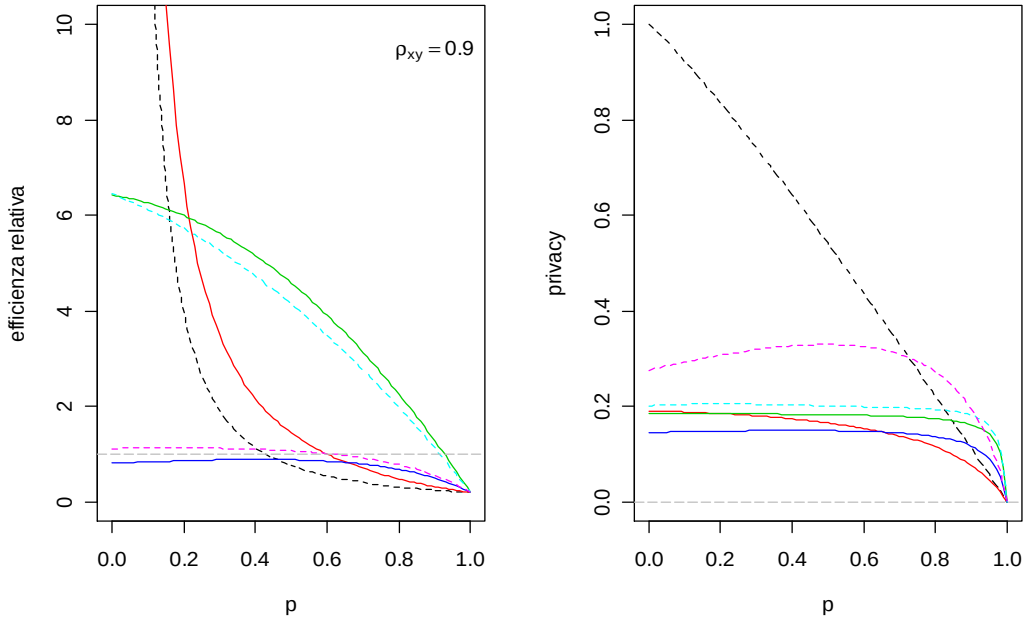




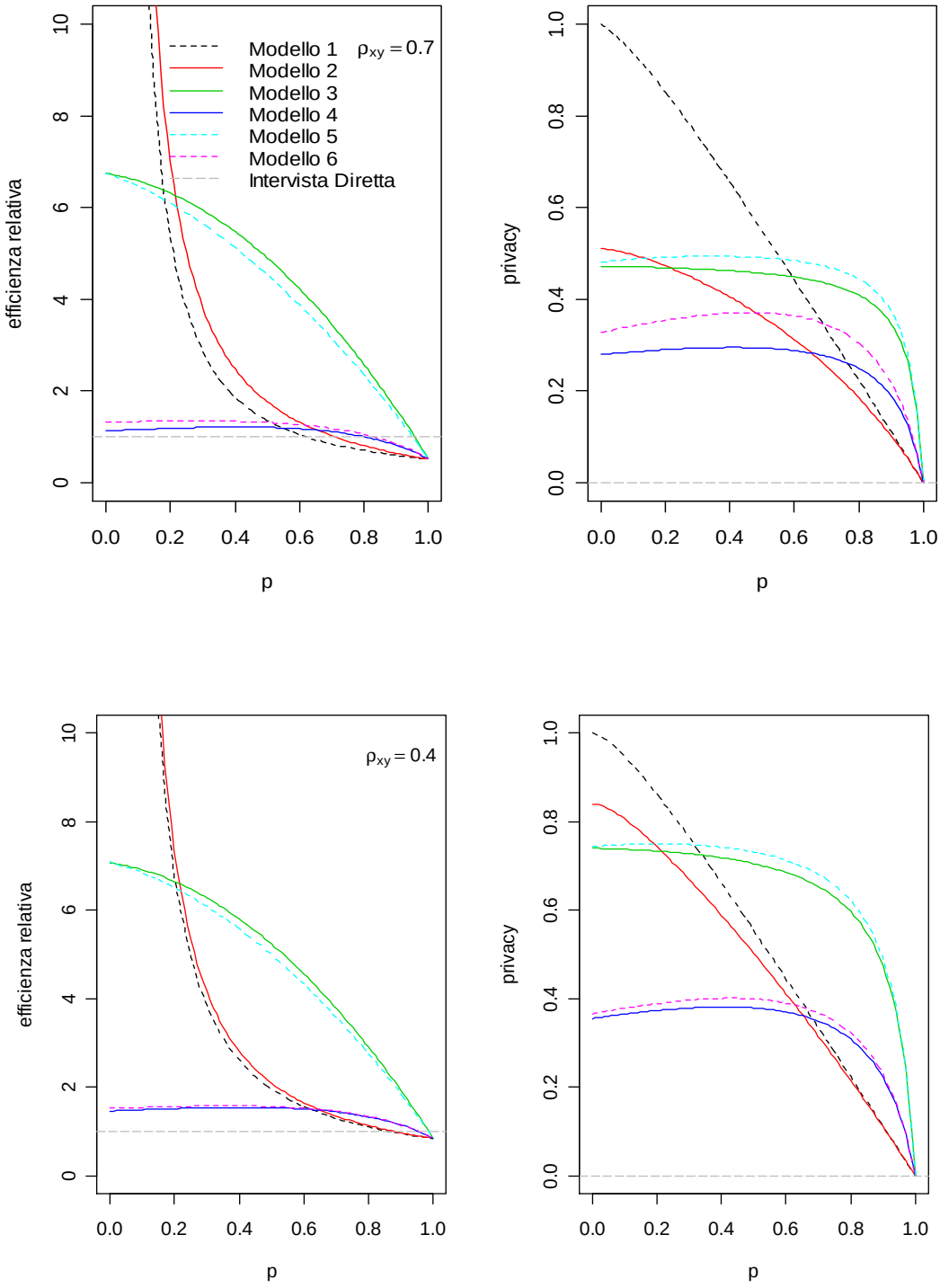
Caso 1. $H,U \sim Ln(0.8,1)$ $T,W \sim Ln(0.5,1)$



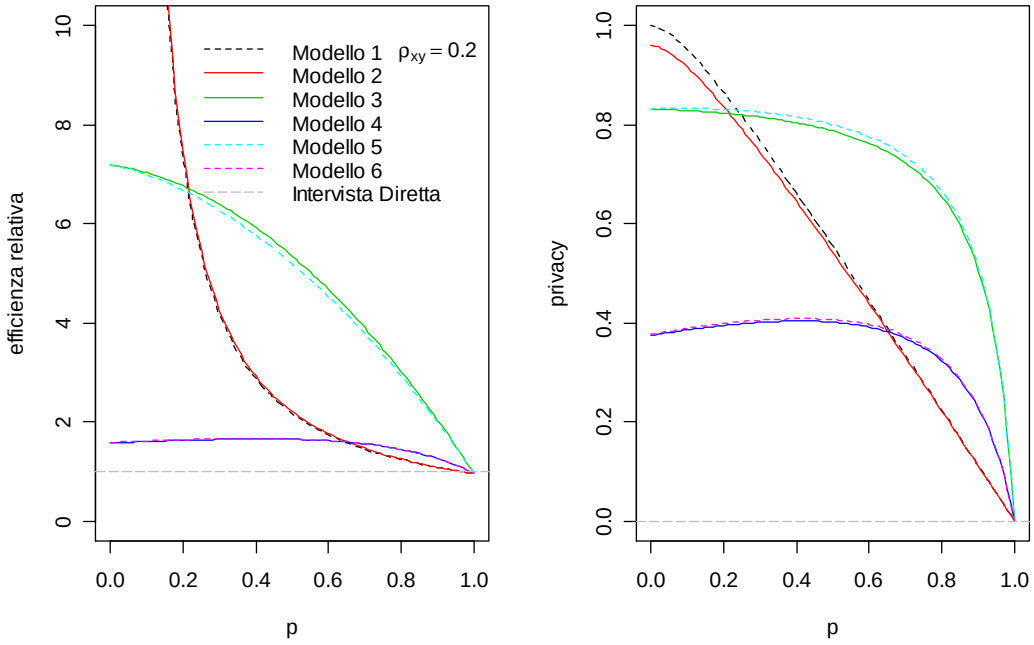
Caso 2. $H,U \sim Ln(1.8,0.8)$ $T,W \sim Ln(0.5,0.4)$

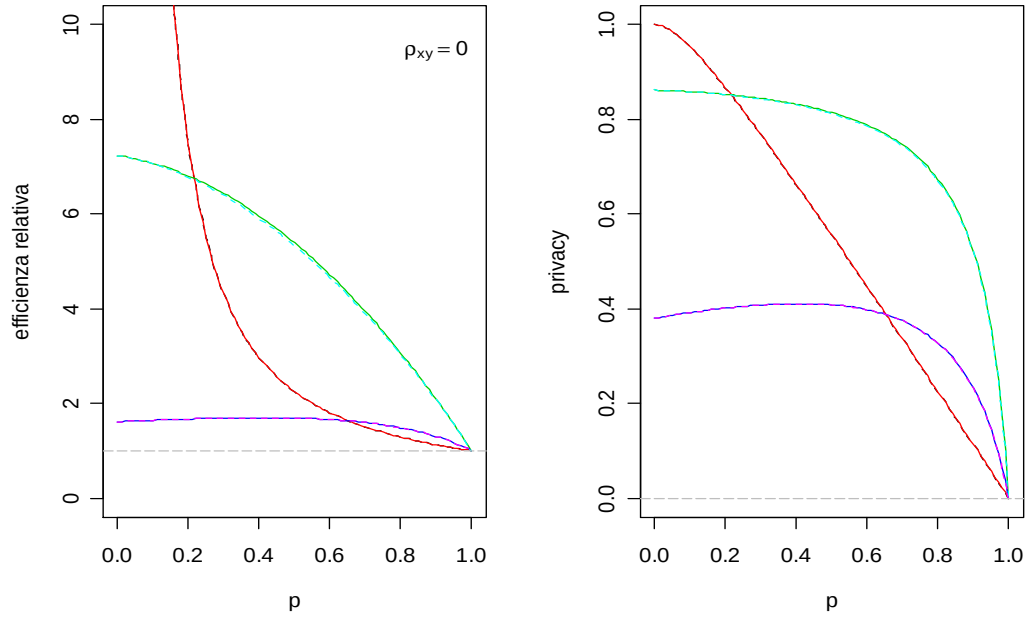


Caso 2. $H,U \sim Ln(1.8,0.8) \quad T,W \sim Ln(0.5,0.4)$

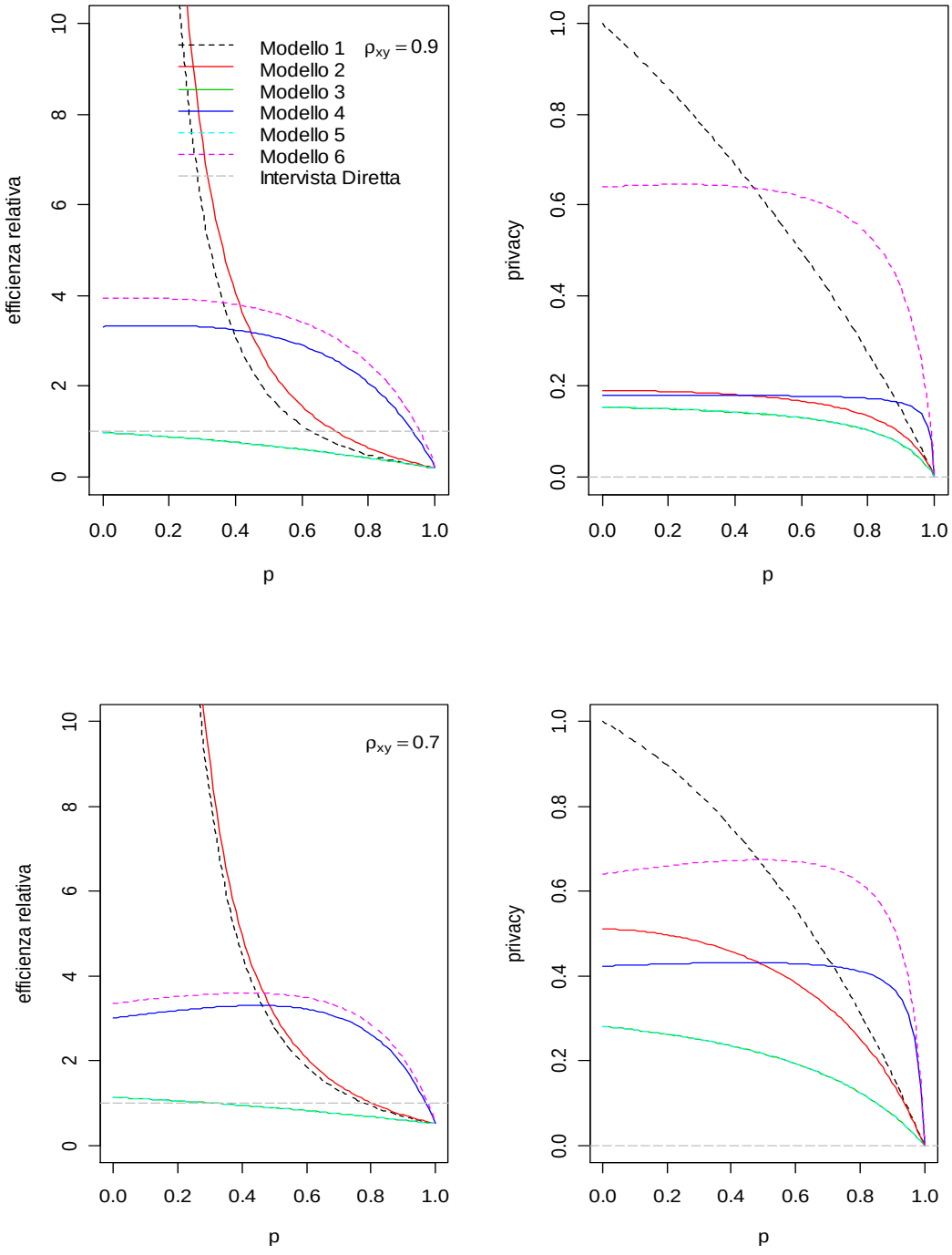


Caso 2. $H,U \sim Ln(1.8,0.8) \quad T,W \sim Ln(0.5,0.4)$

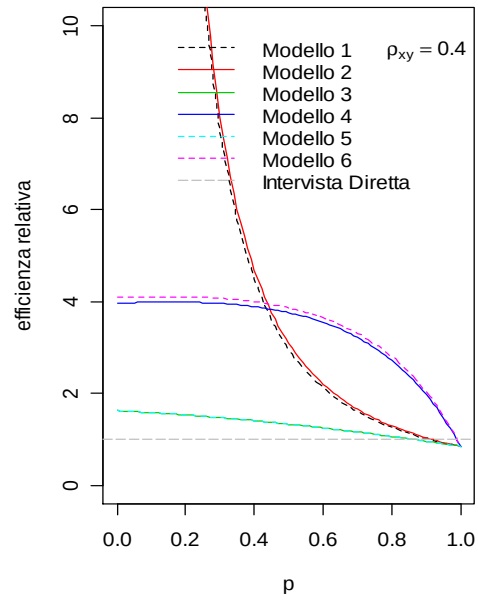


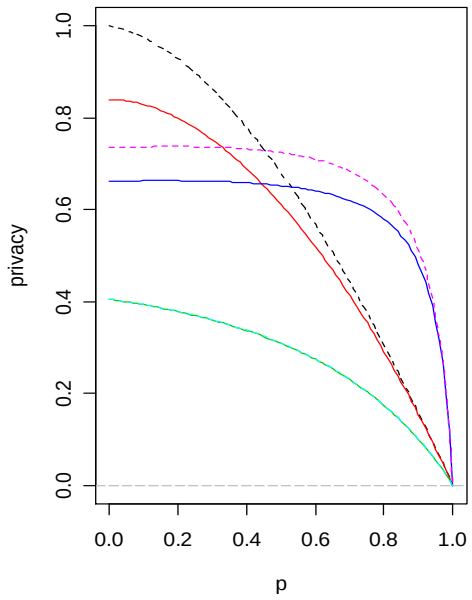


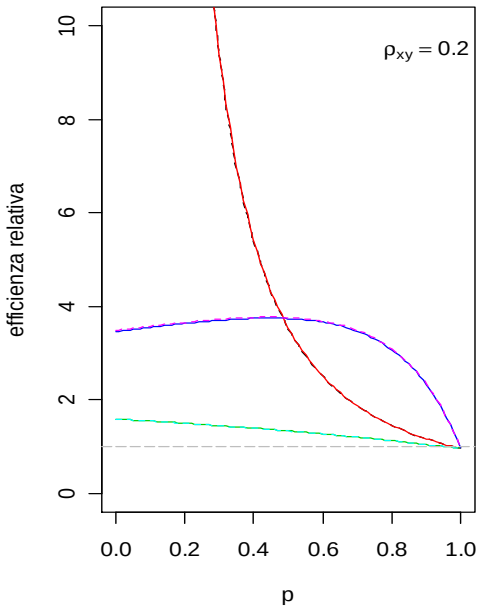
Caso 1. $H,U \sim Gamma(0.5,0.2)$ $T,W \sim Gamma(0.4,0.2)$

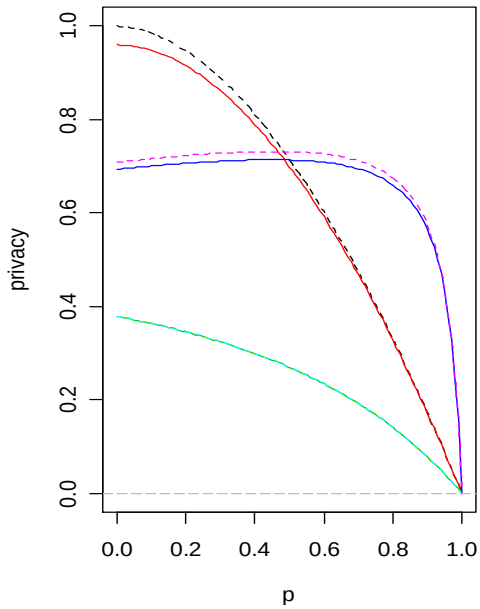


Caso 1. $H, U \sim Gamma(0.5, 0.2)$ $T, W \sim Gamma(0.4, 0.2)$

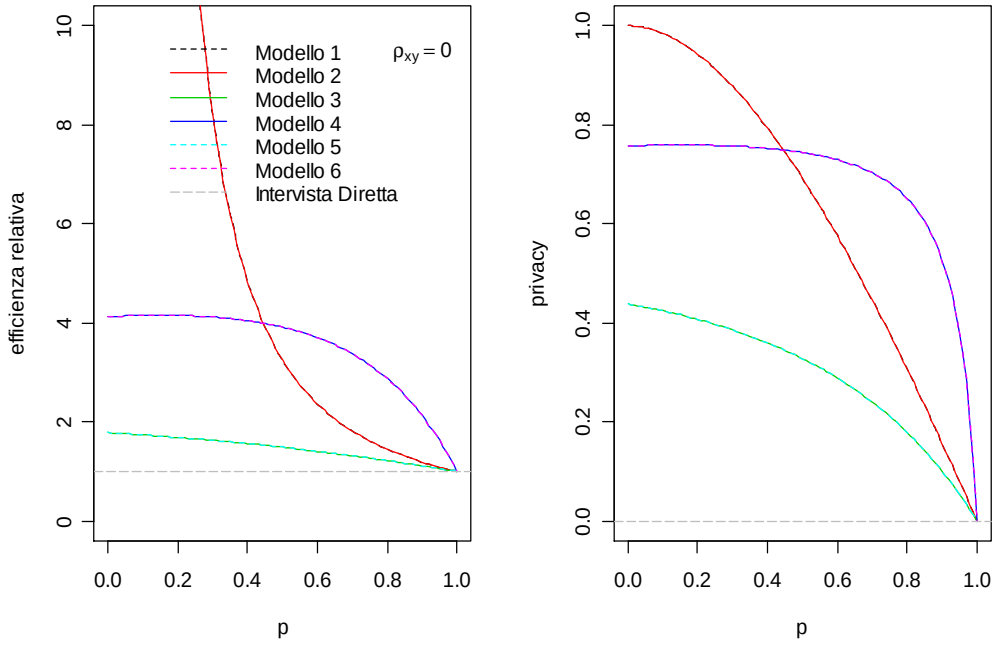




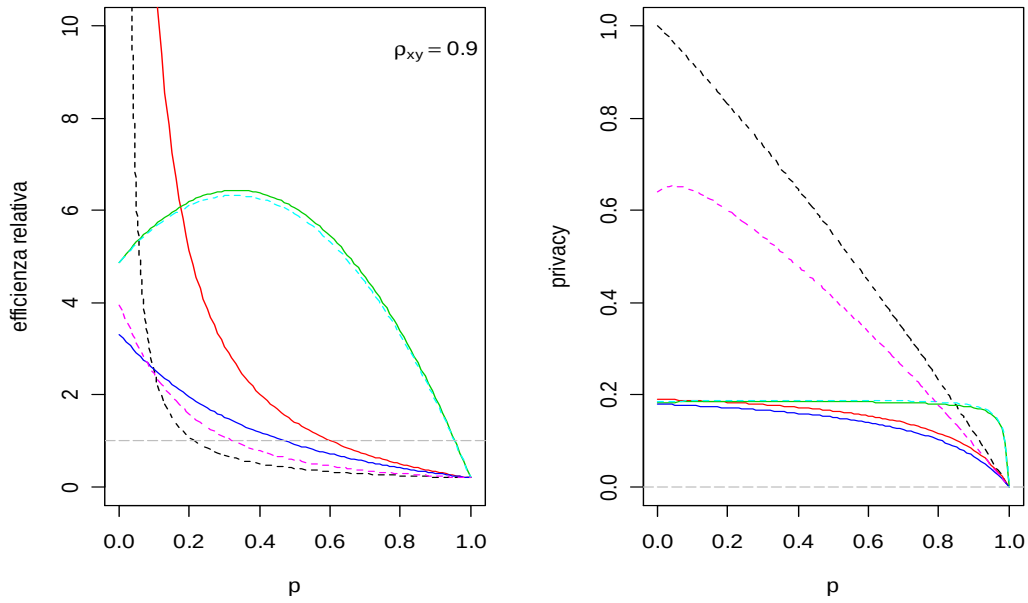




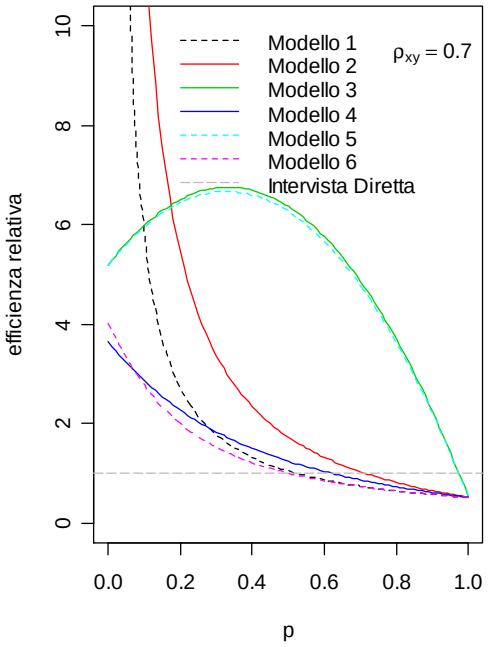
Caso 1. $H,U \sim Gamma(0.5,0.2)$ $T,W \sim Gamma(0.4,0.2)$

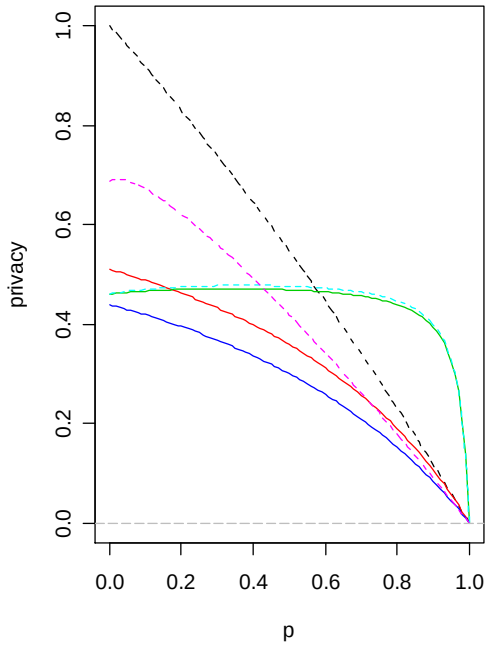


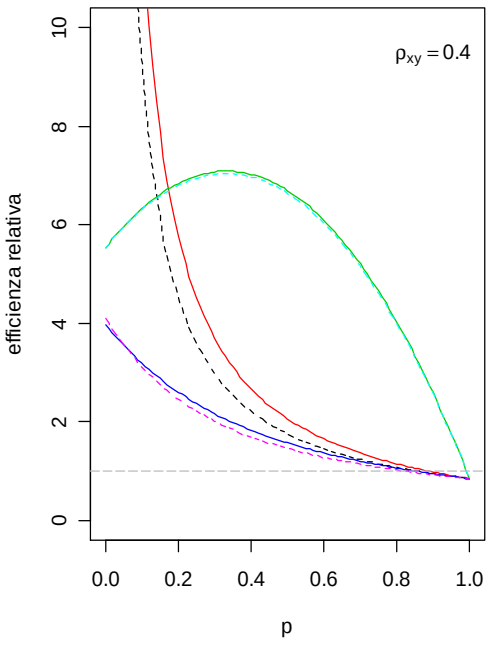
Caso 2. $H,U \sim Gamma(3,0.2)$ $T,W \sim Gamma(0.4,2)$

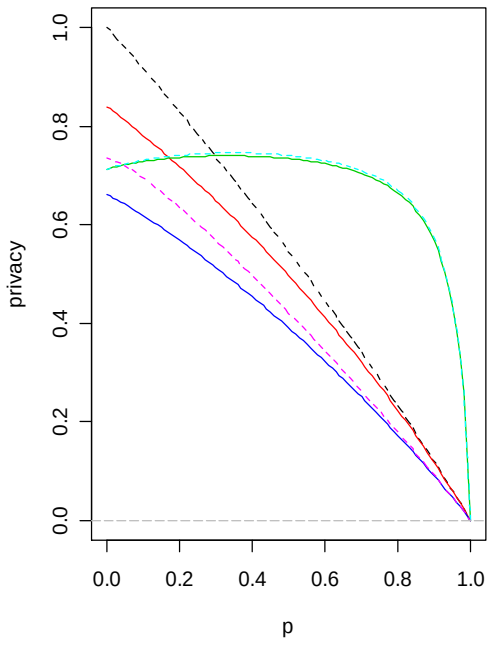


Caso 2. $H,U \sim Gamma(3,0.2)$ $T,W \sim Gamma(0.4,2)$

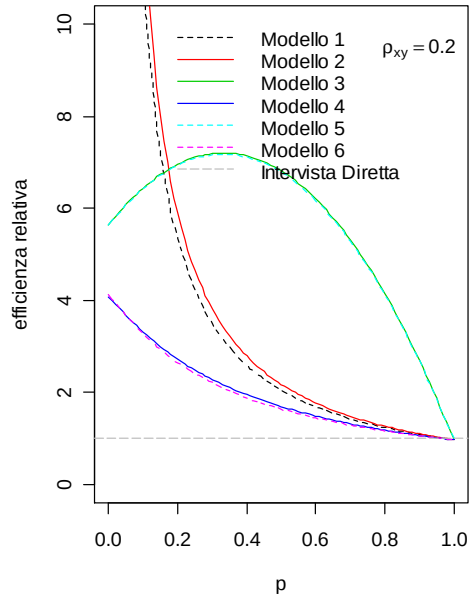


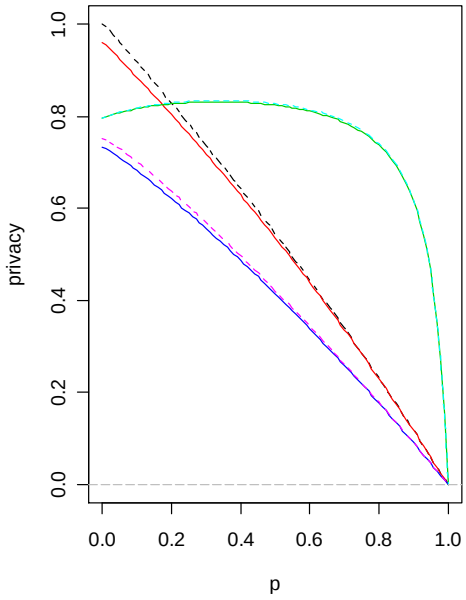


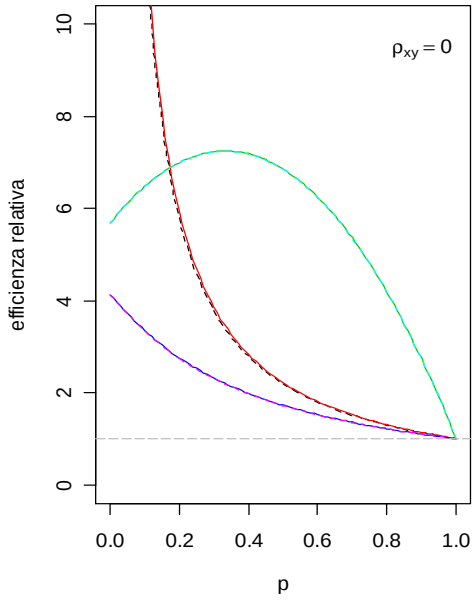


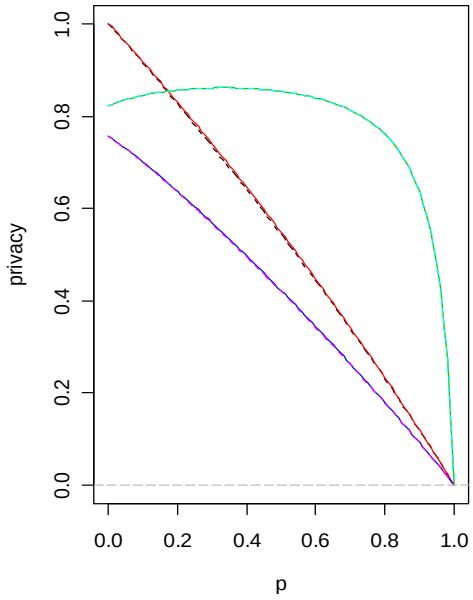


Caso 2. $H, U \sim Gamma(3,0.2)$ $T, W \sim Gamma(0.4,2)$

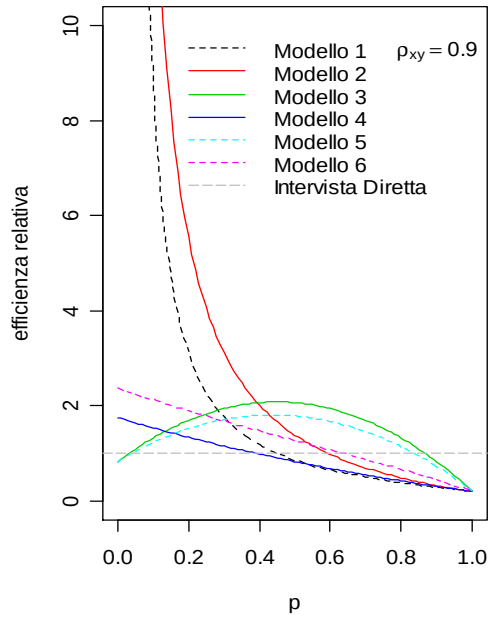


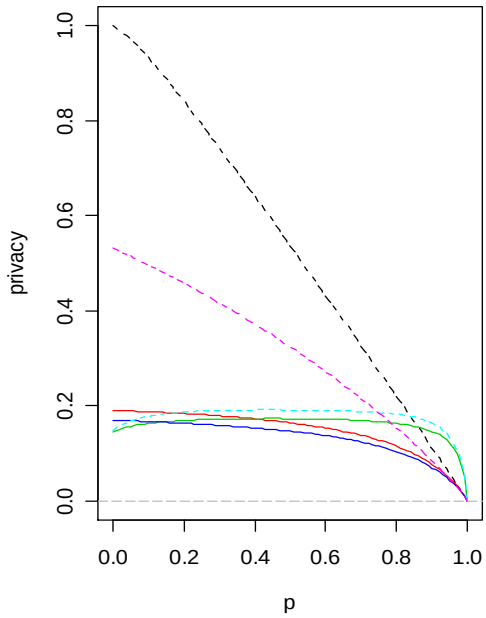


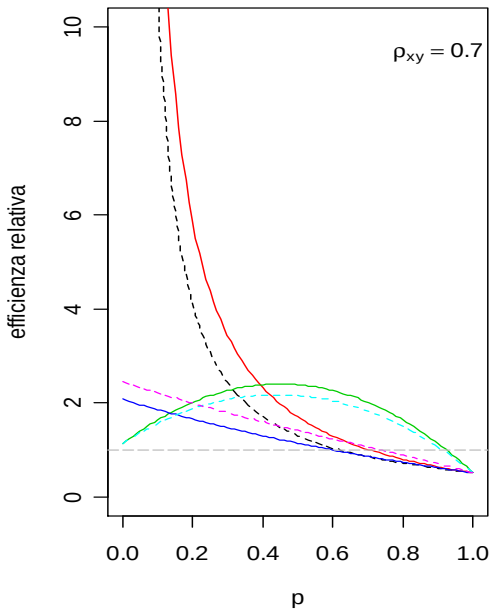


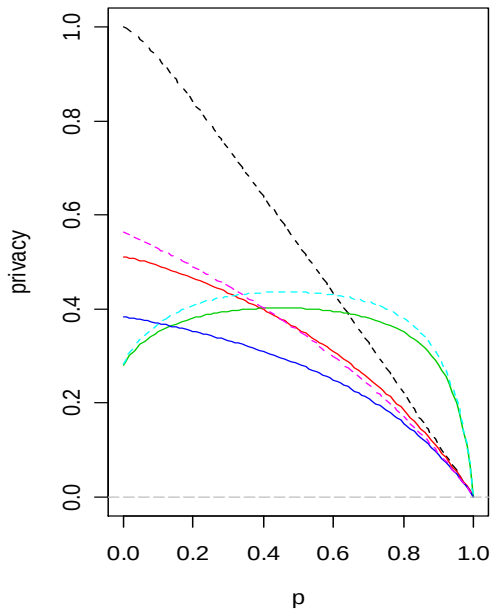


Caso 1. $H, U \sim Pois(10) \quad T, W \sim Pois(0.8)$

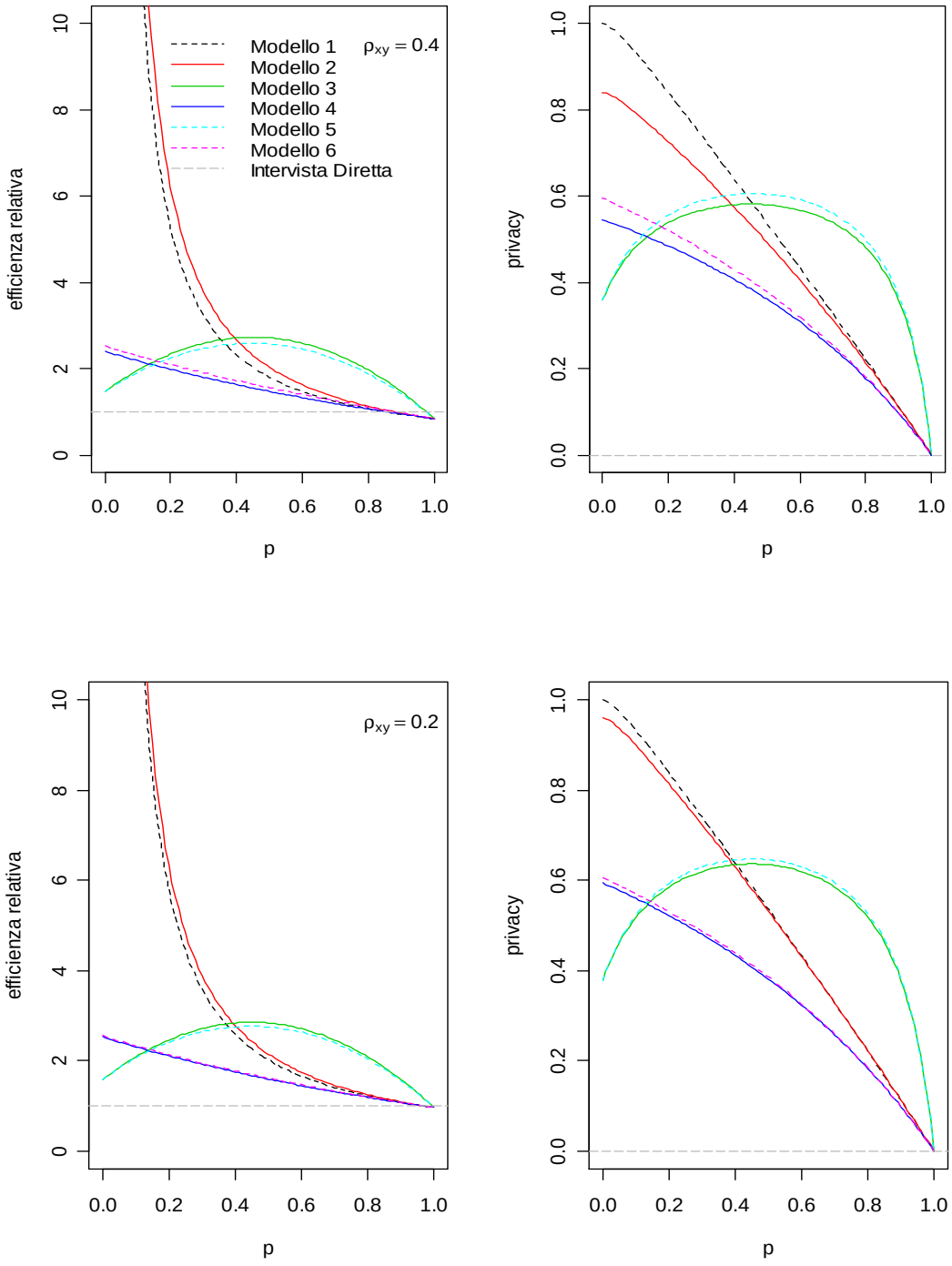




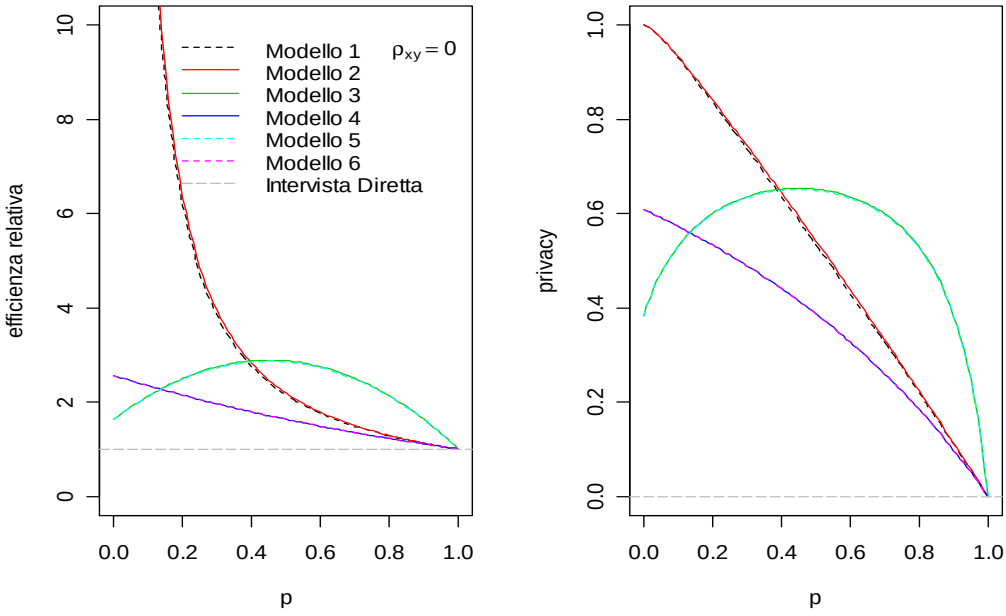




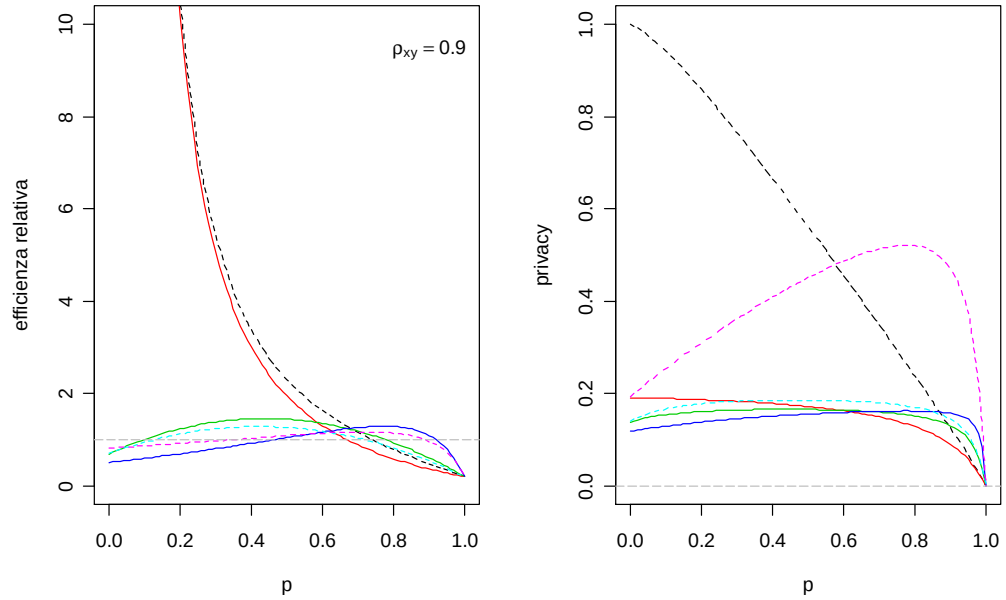
Caso 1. $H, U \sim Pois(10) \quad T, W \sim Pois(0.8)$



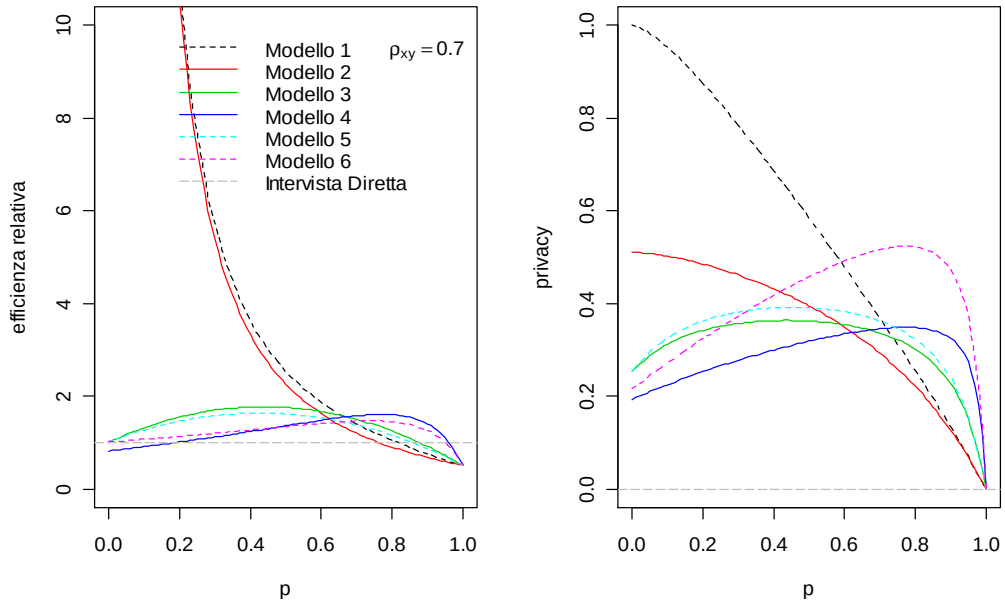
Caso 1. $H,U \sim Pois(10)$ $T,W \sim Pois(0.8)$

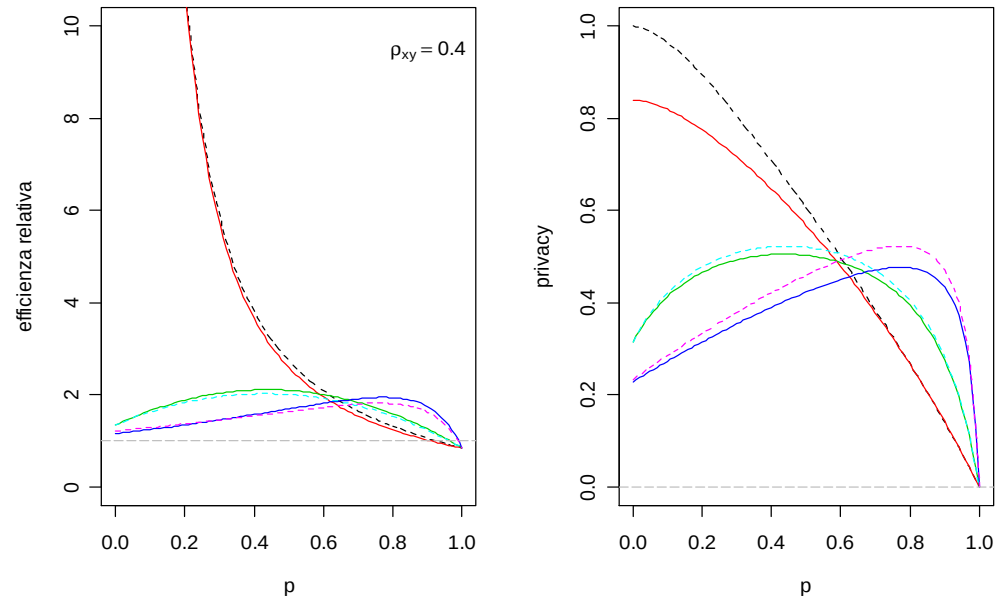


Caso 2. $H,U \sim Pois(8)$ $T,W \sim Pois(4)$

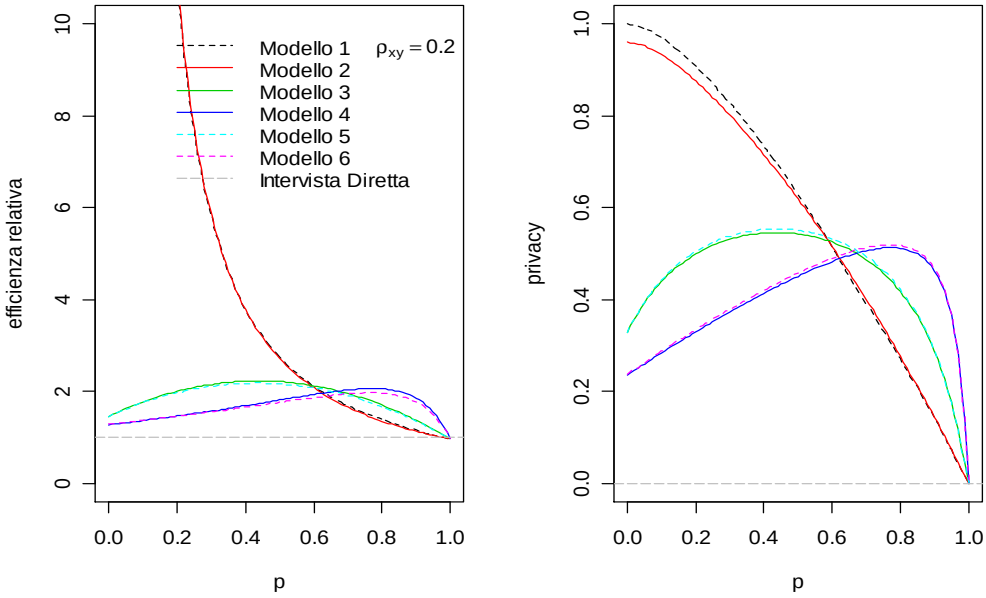


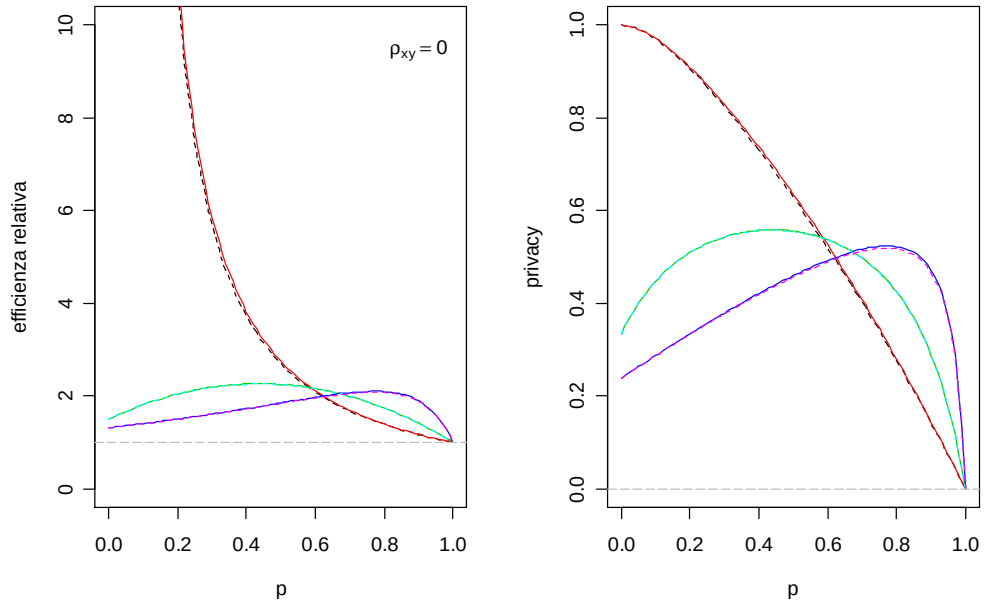
Caso 2. $H,U \sim Pois(8) \quad T,W \sim Pois(4)$





Caso 2. $H,U \sim Pois(8) \quad T,W \sim Pois(4)$





Appendice B

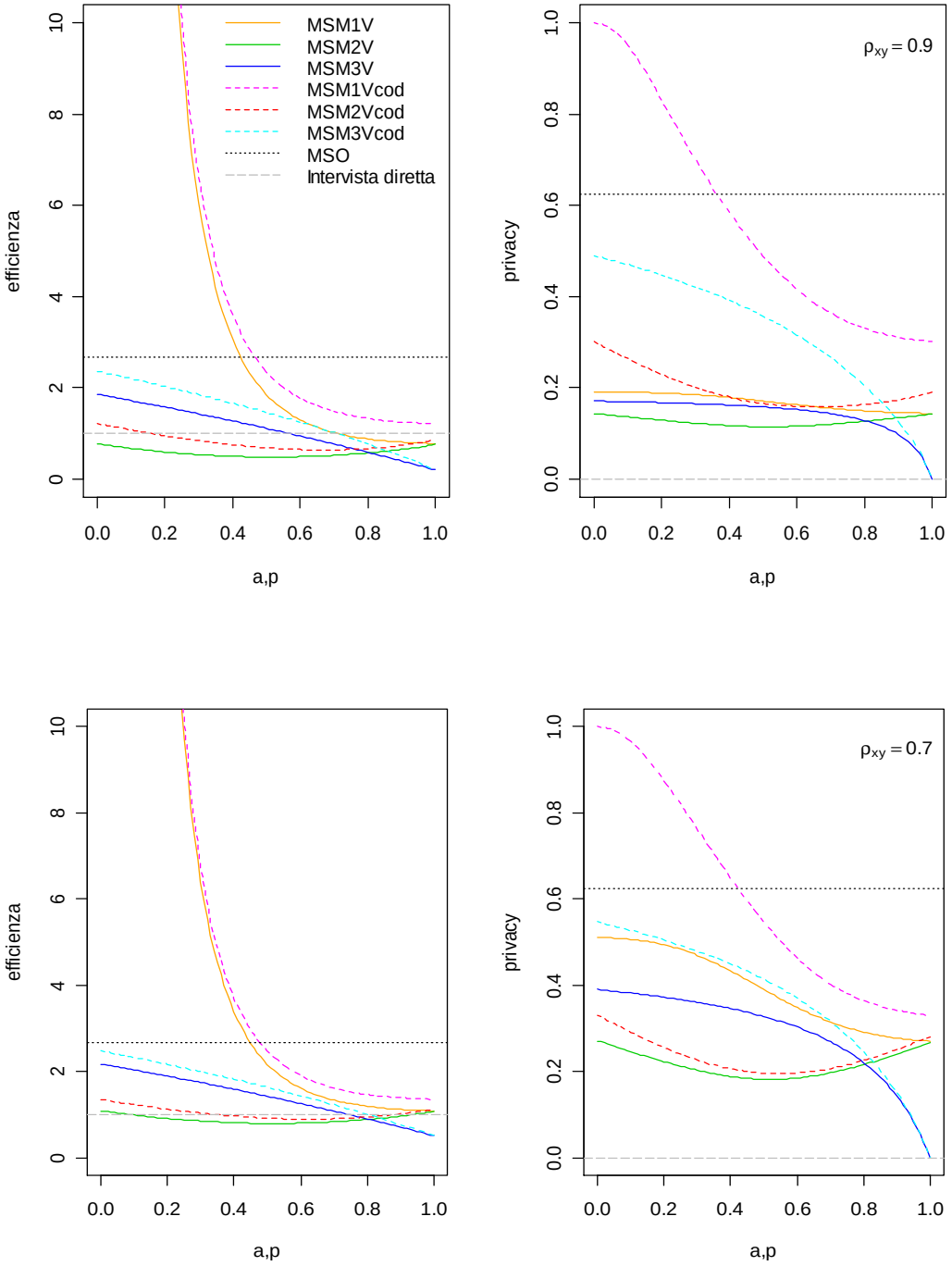
Efficienza e protezione della privacy dei MSM

Si esaurisce ora la selezione di grafici presentata nel par. 4.4, relativa ai tre modelli di Saha modificati, considerando $(\rho_{XY} = 0.0,\ \mathbf{0.2},\ 0.4,\ \mathbf{0.7},\ 0.9)$.

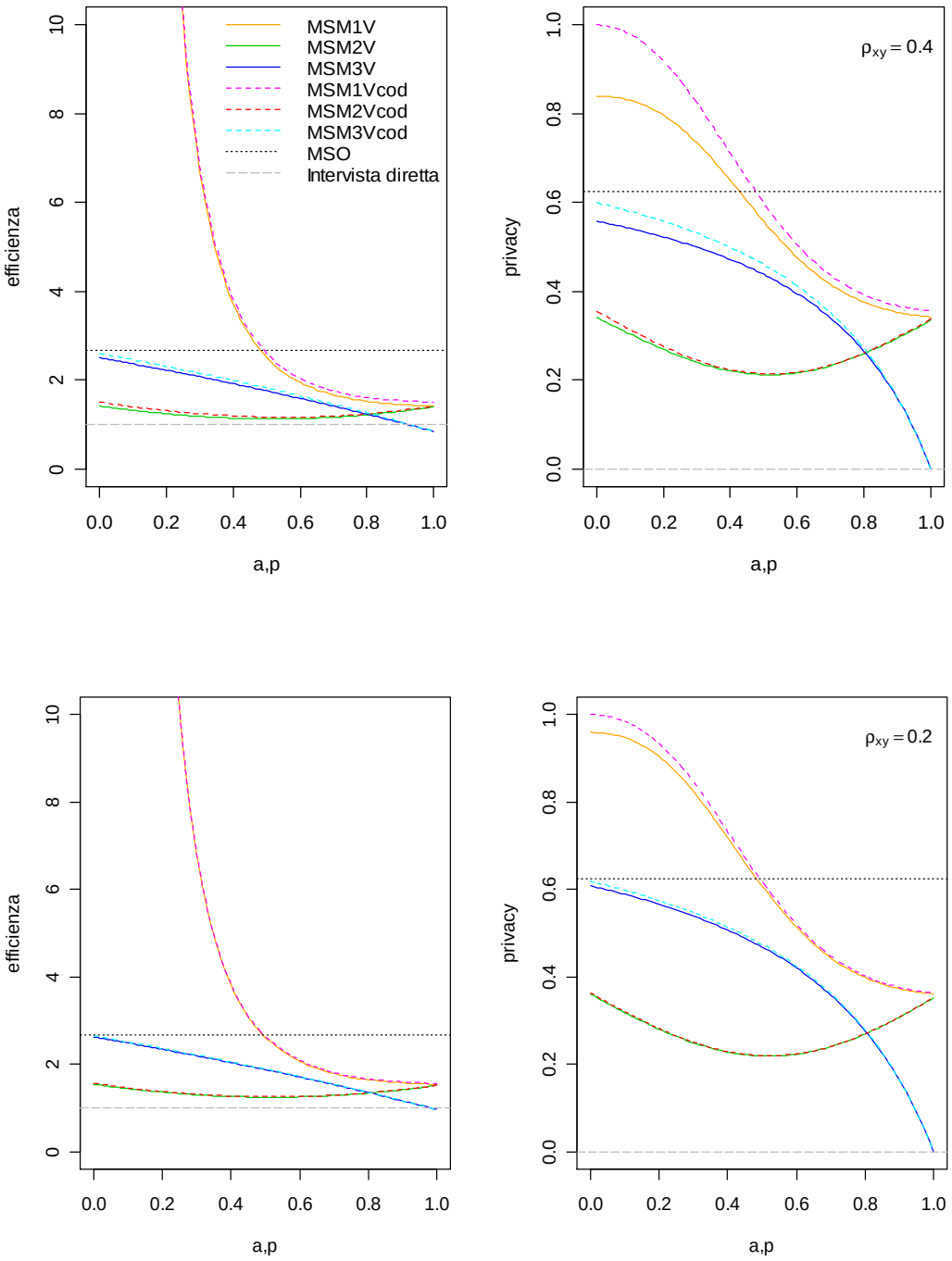
Sia per il *Caso 1* che per il *Caso 2*, abbiamo assunto $\mu_X = 10$ e $\mu_Y = 2$, $C_X = 0.8$ e $C_Y = 2$. Scegliendo come distribuzione la F di *Fisher* , nel primo caso H ed U presentano (5,5) g.d.l., mentre T e W (5,50) g.d.l.; i rispettivi coefficienti di variazione sono $C_H = C_U = 1.789$ e $C_T = C_W = 0.679$. Nel secondo caso, invece, abbiamo posto (1,5) g.d.l. per H ed U e (5,50) g.d.l. per T e W , a cui corrispondono $C_H = C_U = 2.828$ e $C_T = C_W = 0.679$.

I risultati ottenuti per $\rho_{XY} = 0.2$ e $\rho_{XY} = 0.7$ sono in accordo con quelli riportati nel par. 4.4.

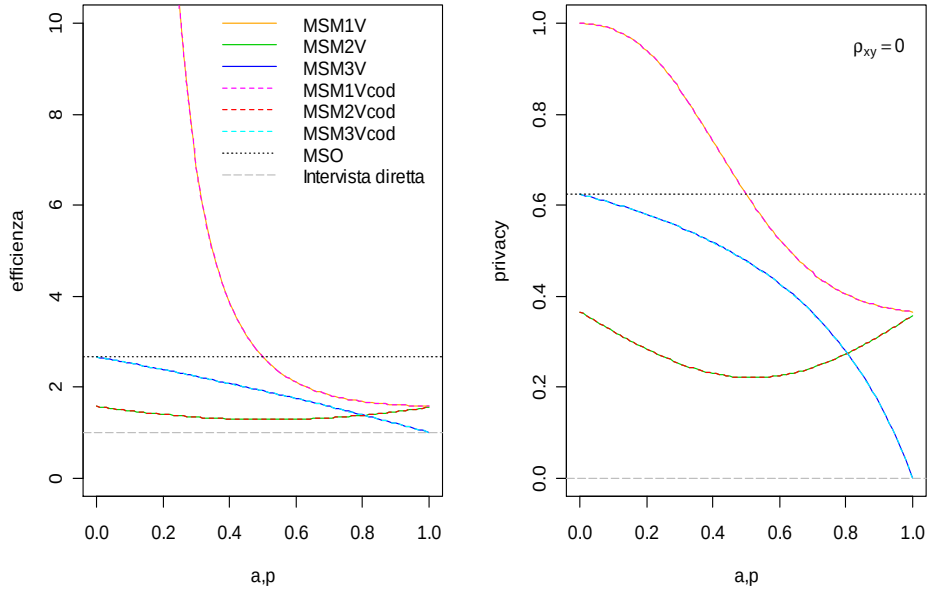
Caso 1. $H,U \sim F(5,5) \quad T,W \sim F(5,50)$



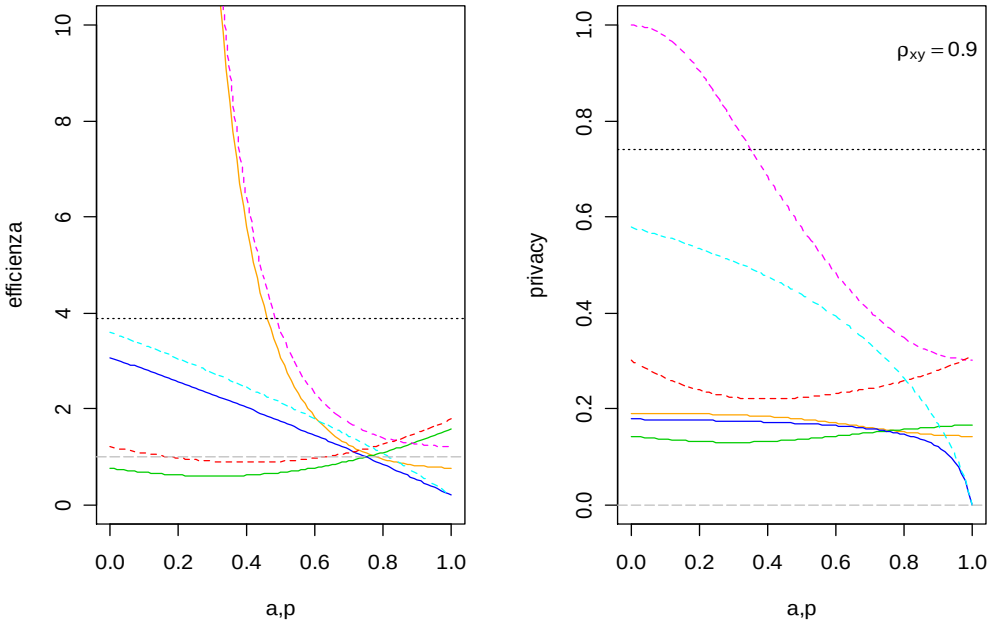
Caso 1. $H,U \sim F(5,5) \quad T,W \sim F(5,50)$



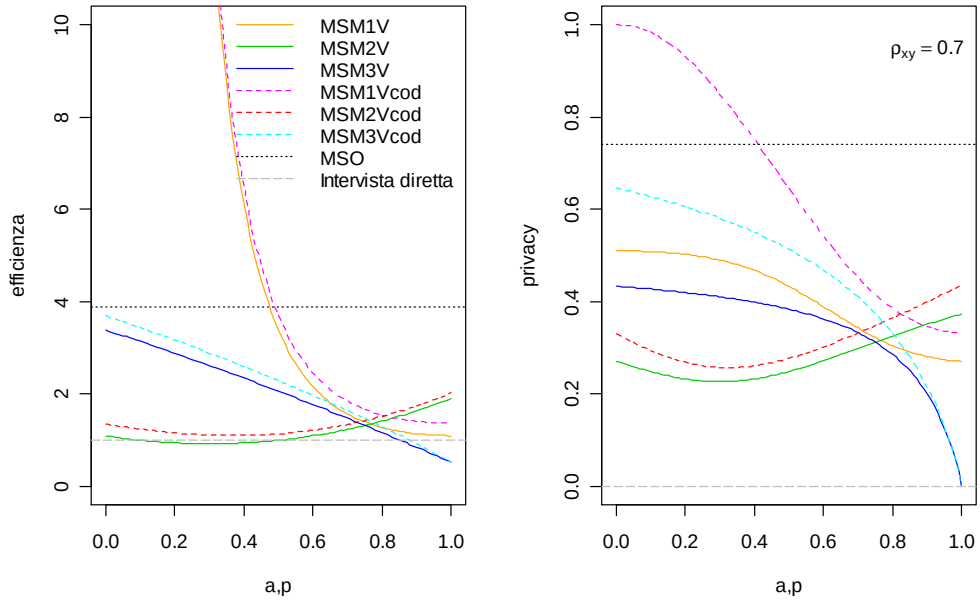
Caso 1. $H,U \sim F(5,5) \quad T,W \sim F(5,50)$

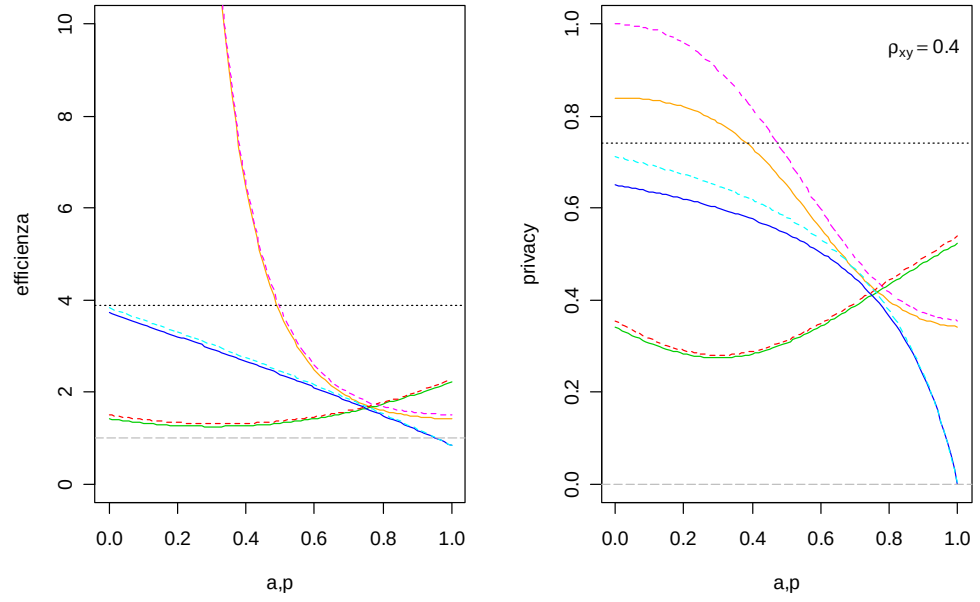


Caso 2. $H,U \sim F(1,5) \quad T,W \sim F(5,50)$

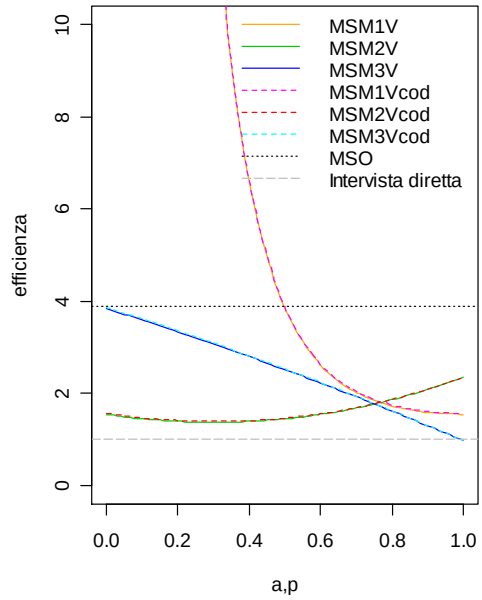


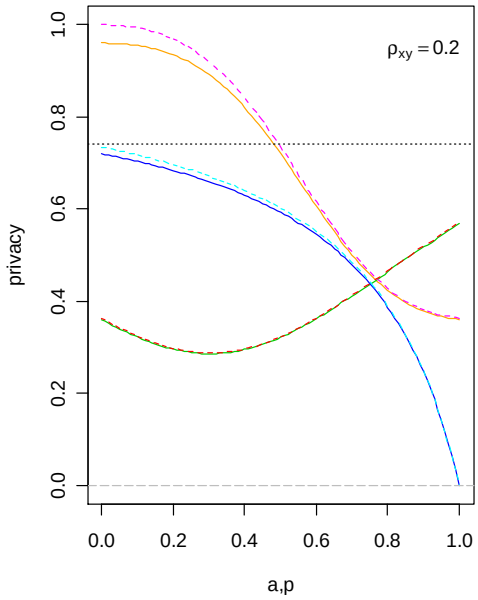
Caso 2. $H,U \sim F(1,5) \qquad T,W \sim F(5,50)$

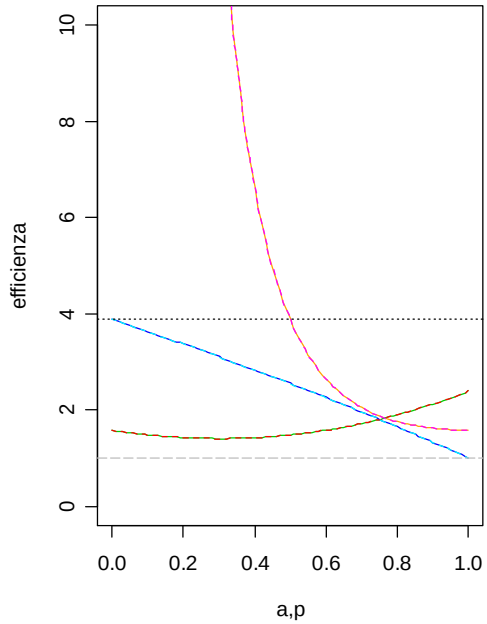


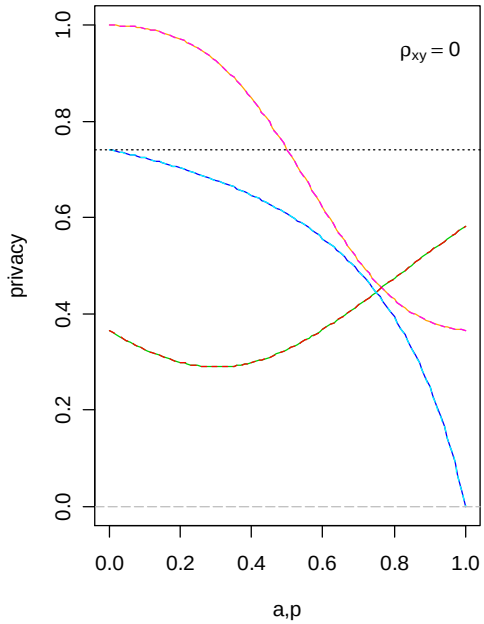


Caso 2. $H,U \sim F(1,5) \qquad T,W \sim F(5,50)$









Riferimenti bibliografici

Agrawal, R. e Srikant, R.(2000). "Privacy preserving data mining", *Proceedings of the 19th ACM SIGMOD Conference on Managment of Data*, Dallas, Texas, USA.

Agrawal, D. e Aggarwal, C.C. (2001). "On the design and quantification of privacy preserving data mining algorithms", *Proceedings of the 20th Symposium on Principles of Database System*, Santa Barbara, California, USA.

Bar-Lev, S.K., Bobovitch, E. e Boukai, B. (2004). "A note on randomized response models for quantitative data", *Metrika*, 60, 255-260.

Bhargava, M. e Singh, R. (2002). "On the efficiency comparison of certain randomized response strategies", *Metrika*, 55, 191-197.

Diana, G. e Perri, P.F. (2007a). "Estimating a sensitive proportion through randomized procedures based on auxiliary information", *Statistical Papers*, DOI 10.1007/s00362-007-0107-y, in corso di pubblicazione.

Diana, G. e Perri, P.F. (2007b). "Estimation of finite population mean using multi-auxiliary information", *Metron*, LXV, 99-112.

Duncan, G. T. e Mukerjee, S. (2000). “Optimal disclosure limitation strategy in statistical databases: deterring tracker attacks through additive noise”, *Journal of the American Statistical Association*, 95, 720-729.

Eichhorn, B. e Hayre, L.S. (1983). “Scrambled randomized response methods for obtaining sensitive quantitative data”, *Journal of Statistical Planning and Inference*, 7, 307-316.

Eriksson, S. A. (1973). “A new model for randomized response”, *International Statistical Review*, 41, 101-113.

Evfimievski, A. (2002). “Randomization in Privacy Preserving Data Mining”, *SIGKDD Explorations*, 4, 43-48.

Fox, J. A. e Tracy, P. E. (1986). “Randomized Response: A Method for Sensitive Survey”, *Sage Publication*, Inc., Newbury Park.

Gjestvang, C. e Singh, S. (2005). “Forced quantitative randomized response model: a new device”, *Metrika*, 66, 243-257.

Greenberg, B. G., Abul-Ela, A.L.A., Simmons, W. R., Horvitz, D. G. (1969). “The unrelated question randomized response model: theoretical framework”, *Journal of the American Statistical Association*, 64, 520-539.

Greenberg, B.G., Kuebler R., Abernathy J. e Horvitz D. (1971). “Application of the randomized response technique in obtaining quantitative data”, *Journal of the American Statistical Association*, 66, 243-250.

Grewal, I. S., Bansal, M. L. e Sidhu, S. S. (2006). “Population mean corresponding to Horwitz-Thompson’ s estimator for multi-characteristics using randomised response technique”, *Model Assisted Statistics and Applications*, 1, 215-220.

Grewal, I. S., Bansal, M. L. e Singh, S. (2003). “Estimation of a population mean of a stigmatized quantitative variable using double sampling”, *Statistica*, LXIII, 1, 79-88.

Guerriero, M. (2005). “Efficienza e protezione della privacy nei randomized response model”, *Statistica e applicazioni*, III, 1.

Guerriero, M. e Sandri, M. F. (2007). “A note on the comparison of some randomized response procedures”, *Journal of statistical planning and inference* , 137, 2184-2190.

Gupta, S., Gupta, B. e Singh, S. (2002). “Estimation of sensitive level of personal interview survey questions”, *Journal of Statistical Planning and Inference*, 100, 239-247.

Gupta, S. e Shabbir, J. (2004). “Sensitivity estimation for personal interview survey questions”, *STATISTICA*, LXIV, 644-653.

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

Huang, K.C. (2005). “Estimation of sensitive data from a dichotomous population”, *Statistical Papers*, 47, 149-156.

Leysieffer, F. W. e Warner, S. (1976). “Respondent Jeopardy and Optimal Designs in Randomized Response Models”, *Journal of the American Statistical Association*, 71, 649-656.

Ljungqvist, L. (1993). “A Unified Approach to Measures of Privacy in Randomized Response Models: A Utilitarian Perspective”, *Journal of the American Statistical Association*, 88, 97-103.

Mahajan, P.K. (2006). “Optimum Stratification for scrambled response with ratio and regression estimators”, *Model Assisted Statistics and Application*, 1, 17-22.

Mahajan, P.K. e Singh, R. (2005). “Optimum stratification for scrambled response in pps sampling”, *Metron*, LXIII, 103-114.

Mahajan, P. K., Gupta, J.P. e Singh, R. (1994). “Determination of optimum strata boundaries for scrambled response”, *Statistica*, 54, 375-381.

Mukharjee, S. e Duncan, G. T. (1997). “Disclosure limitation through additive noise data masking: analysis of skewed sensitive data”. *Proceedings of the 30th Annual Hawaii International Conference on System Sciences*.

Odumade, O. e Singh, S. (2007). “Generalized forced quantitative randomized response model”, *ASA Section on Survey Research Methods*, 3469-3475.

Pollock, K. H. e Bek, Y. (1976) “A comparison of three randomized response models for quantitative data”, *Journal of the American Statistical Association*, 71, 884-886.

Poole, W.K. (1974). “Estimation of the the distribution function of a continuous type random variable through randomized response”, *Journal of the American Statistical Association*, 69, 1002-1005.

Ryu, J.-B., Kim, J. M., Heo T.Y. e Park, C. G. (2006). “On stratified randomized response sampling”, *Model Assisted Statistics and Applications*, 1, 31-36.

Saha, A. (2007). “A simple randomized response technique in complex survey”, *Metron* , LXV, 59-66.

Singh, H. P. e Mathur, N. (2005) “Estimation of population mean when coefficient of variation is known using scrambled response technique”, *Journal of Statistical Planning and Inference*, 131,135-144.

Sukhatme, P. V., Sukhatme, B. V. e Sukhatme S., Asok, C. (1984). “Sampling Theory of Survey with Applications”, *Iowa State University Press*, Ames, Iowa, USA.

Warner, S. (1965). “Randomized response: a survey technique for eliminating evasive evasive answer bias”, *Journal of the American Statistical Association*, 60, 63-69.

Warner, S. (1971). “The linear randomized response model”, *Journal of the American Statistical Association*, 66, 336, 884-888.

Zaizai, Y. (2005-2006). “Ratio method of estimation of population proportion using randomized response technique”, *Model Assisted Statistics and Applications*, 1, 125-130.