STATISTICA, anno LXIII, n. 4, 2003

ALCUNE OSSERVAZIONI SU PROTEZIONE DEI DATI PERSONALI E ANALISI STATISTICA

Franco Peracchi (*)

1. INTRODUZIONE

Voglio anzitutto ringraziare la rivista *Statistica* per l'opportunità che mi offre di riflettere brevemente sul tema del rapporto tra protezione dei dati personali e analisi statistica. Ho l'impressione che, nonostante i molti interventi su un tema così importante, il dibattito italiano si sia per lo più svolto a un livello eccessivamente astratto, e quindi non abbia adeguatamente colto quali rischi concreti di violazione della *privacy* rendano necessaria una regolamentazione formale dell'uso dei dati personali nell'analisi statistica. Si tratta, a mio avviso, di un punto importante, perché una regolamentazione troppo generica può portare a un livello eccessivo di protezione, con un inutile aggravio dei costi diretti e indiretti della protezione.

2. OBIETTIVI DELL'ANALISI STATISTICA

Gli obiettivi di un'analisi statistica dipendono della natura dei dati a disposizione. Se i dati sono di tipo censuario, ottenuti cioè tramite la completa enumerazione delle unità che compongono una popolazione di interesse (individui, famiglie, imprese, ecc.), l'obiettivo dell'analisi statistica è quello di descrivere certe caratteristiche della popolazione. Se i dati a disposizione sono di tipo campionario, l'obiettivo è invece quello di fare inferenza circa certe caratteristiche ignote della popolazione. Solitamente, le caratteristiche della popolazione che costituiscono l'oggetto di un'analisi statistica possono essere ricondotte a totali, momenti (medie, varianze, ecc.), percentili, o funzioni (più o meno complicate) di momenti e percentili.

In generale, l'obiettivo di un'analisi statistica non è quello di descrivere o fare inferenza circa le caratteristiche delle singole unità che compongono la popolazione o il campione. Semplificando le cose al massimo, le caratteristiche di una singola unità sono viste come il risultato di una parte sistematica (il "segnale") che fornisce l'informazione desiderata circa le caratteristiche della popolazione, e di

(*) Ringrazio Ugo Trivellato per i commenti puntuali su una versione preliminare di questo lavoro.

una parte non sistematica (il "disturbo" o "errore") che sporca il segnale fornito dalla parte sistematica e la cui influenza deve quindi essere rimossa attraverso tecniche opportune.

Il fatto che le caratteristiche individuali non costituiscano l'oggetto di un'analisi statistica non significa però che esse siano prive di interesse o irrilevanti per l'analista. Anzitutto, come già detto, esse contengono l'informazione rilevante per descrivere o fare inferenza circa le caratteristiche della popolazione. Va inoltre tenuta presente una seconda considerazione. In un'analisi statistica descrittiva, il valore medio di una variabile di interesse può dipendere in misura determinante dal valore assunto da tale variabile per un numero ridotto di unità della popolazione. Analogamente, in un contesto inferenziale, la qualità dell'inferenza circa il valore medio di una variabile di interesse può dipendere in misura determinante dal valore assunto da tale variabile per un numero ridotto di unità presenti nel campione. In entrambi i casi, è importante poter determinare quali unità esercitino un'influenza eccessiva sui risultati di un'analisi statistica e perché.

Presentata in questi termini, l'analisi statistica non sembra dunque entrare in conflitto con il diritto degli individui alla *privacy*, e non sembra quindi richiedere una particolare regolamentazione formale che si aggiunga a quella implicita nei codici internazionali di comportamento della professione statistica. In base a questi ultimi, infatti, qualunque analisi che abbia come obiettivo lo studio delle caratteristiche di una particolare unità della popolazione verrebbe declassata al rango di "analisi non statistica", con ovvie conseguenze negative per gli autori di tale analisi.

È certamente possibile che, in un particolare paese, i codici di comportamento della professione statistica non corrispondano agli standard internazionali, oppure che le sanzioni per gli autori di "analisi non statistiche" siano deboli o addirittura inesistenti. Fortunatamente, non credo che questo sia il caso dell'Italia.

Se si escludono dunque queste due spiegazioni (e si conviene sul fatto che la giurisprudenza fornisce comunque strumenti adeguati per punire gli abusi compiuti sotto lo schermo di finalità statistiche), da quali altri rischi connessi all'analisi statistica nasce allora l'esigenza di protezione formale dei dati personali? Una possibile risposta è legata all'uso crescente di dati amministrativi nell'analisi statistica.

3. DATI CAMPIONARI E DATI AMMINISTRATIVI

Tradizionalmente, l'analisi statistica si è basata su dati raccolti tramite indagini di tipo campionario o censuario, in cui le unità della popolazione rispondono a una serie di domande relative a fenomeni oggetto di interesse statistico. Tali indagini forniscono agli individui due forme estremamente efficaci di autoprotezione della *privacy*. La prima è la possibilità di non rispondere. La mancata risposta si può manifestare nel rifiuto di partecipare all'indagine (*unit nonresponse*), oppure nel rifiuto di rispondere a specifiche domande (*item nonresponse*). L'altra è la possibilità di rispondere in modo evasivo o impreciso, o addirittura di non riportare il vero nel rispondere (*misreporting*). Sebbene esistano, in alcuni casi, forme di sanzione del rifiuto di partecipare a un indagine, la loro efficacia è assai dubbia. Non mi risulta poi che esi-

stano forme di sanzione della mancata risposta a specifiche domande o della " bugia statistica". È noto invece che violazioni, reali o percepite, del diritto alla *privacy* tendono a tradursi in un deterioramento della qualità dei dati per l'aumento dei fenomeni di mancata risposta o di *misreporting* (si vedano, per esempio Cochran, 1977, e Groves e Couper, 1998). Nel caso delle indagini statistiche, la possibilità di non rispondere o di mentire rappresenta l'equivalente del comportamento che, nel campo della teoria politica, viene a volte indicato come "votare con i piedi".

Recentemente, l'analisi statistica ha iniziato a fare ricorso in misura crescente a dati di tipo amministrativo, basati cioè sull'informazione raccolta per scopi amministrativi (e quindi non statistici) da un gran numero di enti pubblici e privati. Dal punto di vista dell'argomento di questa nota, la novità principale è che le due forme di autoprotezione tipiche delle indagini statistiche tendono a mancare per questo tipo di dati. Infatti, l'inserimento in un archivio amministrativo è molto spesso sottratto alla libera scelta individuale. Inoltre, per gli scopi propri di tali archivi, l'informazione che essi contengono tende ad avere un elevato grado di accuratezza. Queste due caratteristiche contribuiscono a spiegare il crescente interesse per i dati amministrativi nell'analisi statistica. In effetti, il loro utilizzo consente spesso di ridurre o addirittura eliminare completamente i problemi connessi alla mancata risposta e agli errori di misura, che affliggono invece in misura più o meno accentuata qualunque analisi basata su indagini statistiche.

A mio avviso, l'impossibilità di forme di autoprotezione nel caso dei dati amministrativi costituisce la principale giustificazione per l'introduzione di norme che assicurino una protezione formale dei dati personali. Come già detto, una protezione formale risulta invece assai meno necessaria per i dati ottenuti tramite indagini statistiche. In questo caso, sarebbe anzi desiderabile rimuovere l'obbligatorietà della risposta per le indagini statistiche condotte dall'Istituto Nazionale di Statistica. Sia pure scarsamente utilizzata, questa norma peculiare (essa è infatti presente in pochi altri paesi oltre all'Italia) mi pare contrasti in modo stridente con il diritto fondamentale al rispetto della *privacy* che si intende altrimenti difendere.

In ogni caso, credo che sia importante distinguere con chiarezza tra indagini statistiche e uso statistico di dati amministrativi, evitando inutili confusioni. Le due fonti di dati hanno natura e caratteristiche diverse e, come ho cercato di argomentare, il problema del rispetto della *privacy* si pone assai diversamente nei due casi. Purtroppo, la legislazione recentemente introdotta in Italia sembra avere scelto la strada di assicurare un livello generale di protezione che astrae completamente da questa distinzione. Sia pure legittima, questa strategia potrebbe portare a un livello eccessivo di protezione, con un inutile aggravio dei costi diretti (maggiore burocrazia) e indiretti (minore ricerca scientifica, e quindi minore conoscenza) della protezione.

4. CONCLUSIONI

La possibilità per gli individui di non partecipare a un'indagine statistica, oppure di rispondere in modo evasivo o impreciso, rappresentano una forma semplice

ma efficace di protezione della *privacy*, che rende non necessaria una regolamentazione formale dell'uso dei dati individuali nell'analisi statistica.

La recente tendenza a utilizzare in modo crescente i dati amministrativi nell'analisi statistica, spesso giustificata proprio con l'aumento della frequenza delle mancate risposte nelle indagini campionarie e censuarie, ha modificato il quadro ponendo in primo piano l'urgenza di nuove forme di tutela della *privacy*.

La legislazione recentemente introdotta in Italia rappresenta un passo importante in questa direzione. Ignorando tittavia la distinzione tra indagini statistiche e dati amministrativi, essa potrebbe però portare a un livello eccessivo di protezione, con un inutile aggravio dei costi diretti e indiretti della protezione.

Dipartimento SEFEMEQ FRANCO PERACCHI
Università di Roma "Tor Vergata"

RIFERIMENTI BIBLIOGRAFICI

W.G. COCHRAN, (1977), *Sampling Techniques*, Wiley, New York.
R.M. GROVES, M.P. COUPER, (1998), *Nonresponse in Household Interview Surveys*, Wiley, New York.

RIASSUNTO

Alcune osservazioni su protezione dei dati personali e analisi statistica

La libertà che gli individui hanno di non partecipare a un'indagine statistica, oppure di rispondere in modo evasivo o impreciso, rappresentano una forma semplice ma efficace di protezione della *privacy*, che rende non necessaria una regolamentazione formale dell'uso dei dati individuali nell'analisi statistica. La recente tendenza a utilizzare in modo crescente i dati amministrativi nell'analisi statistica ha modificato il quadro e ha posto in primo piano la necessità di nuove forme di tutela della *privacy*. La legislazione recentemente introdotta in Italia rappresenta un passo importante in questa direzione. Tuttavia, ignorando la distinzione tra indagini statistiche e dati amministrativi, essa potrebbe portare a un livello eccessivo di protezione, con un inutile aggravio dei costi della protezione.

SUMMARY

Some remarks on protection of personal data and statistcal analysis

The fact that individuals are free not to participate to a survey, or to answer in an imprecise or even misleading way, represents a simple but very effective form of protection of individual privacy, which makes it largely unnecessary to formally regulate the use of individual data in statistical analysis. The increasing use of administrative data in statistical analysis has modified the picture and brought to the forefront the need of new forms of protection of individual privacy. The regulation recently introduced in Italy represents an important step forward in this direction. However, because it ignores the distinction between statistical surveys and administrative data, this new regulation could lead to an excessive level of protection, with an unncessary increase of the direct and indirect costs of protection.