ires
ISTITUTO RICERCHE ECONOMICO-SOCIALI DEL PIEMONTE
VIA BOGINO 21 - 10123 TORINO - TEL.(011) 88051 - FAX (011) 8123723

Documenti Ires n. 5/96

Utilizzo e rilevanza degli strumenti statistici nella ricerca economico-sociale

Renato Miceli

Ottobre 1996

Indice

Testo della relazione presentata nell'ambito del convegno "Statistica e Pubblica Amministrazione: esperienze a confronto". Torino, 25 settembre 1996

1. La ricerca economico-sociale

Tra i compiti istituzionali dell'Istituto Ricerche Economiche e Sociali del Piemonte (IRES) vi sono:

a) l'osservazione, la documentazione e l'analisi delle principali grandezze socio-economiche e territoriali del sistema regionale;
b) la produzione di ricerche di base e finalizzate;
c) l'acquisizione di conoscenze teorico-concettuali e metodologiche e l'adattamento di queste alle esigenze della Pubblica Amministrazione locale.

Per lo svolgimento di tali compiti il supporto della statistica è indispensabile tutte le volte che occorrono strumenti di analisi quantitativa.

Non tutta l'attività di ricerca si riduce a quella che spesso viene definita (impropriamente) "ricerca quantitativa", ma è a questo segmento dell'attività dell'Istituto (per altro di gran lunga predominante) che io farò riferimento in questo intervento.

Con una definizione che si deve in sostanza a Boudon (1987) possiamo definire la ricerca empirica nelle scienze sociali come una successione di operazioni per produrre risposte a domande sulla realtà. Successione di operazioni generalmente articolate in quattro fasi: il progetto, la raccolta dei dati, l'analisi e l'esposizione dei risultati.

Questo modo di definire la ricerca nelle scienze sociali può apparire ovvio, ma così non è. Una tale definizione infatti implica almeno due considerazioni non marginali.

1) Le capacità conoscitive delle scienze sociali sono limitate. La realtà sociale in sé stessa è inconoscibile. Tutto ciò che si può fare è di formulare delle domande specifiche sulla realtà, e cercare di produrre delle risposte plausibili.
2) La ricerca empirica nelle scienze sociali non si sviluppa a partire da "problemi" teorici (un risultato empirico che mette in crisi una teoria fino

a quel punto accettata come vera - come avviene nelle scienze naturali), ma nasce da "domande" ossia da bisogni di conoscenza più o meno esplicitamente tradotti in insiemi di interrogativi sulla realtà. << E' molto comune in simili circostanze che gli interrogativi che guidano la ricerca siano di natura essenzialmente descrittiva, e non ambiscano in alcun modo a produrre spiegazioni e interpretazioni di portata generale. Si vuol conoscere meglio, più in dettaglio, un determinato fenomeno (il "disagio giovanile", la "cultura dell'alcol", gli atteggiamenti verso gli immigrati, ecc.), sovente a carattere settoriale o insediato in una realtà locale, e non si ha alcun interesse o possibilità di produrre generalizzazioni teoriche>> (L. Ricolfi, 1995).

Nell'ambito di questo tipo di ricerca empirica è opportuno distinguere, inoltre, quel sottoinsieme in cui un ruolo assolutamente cruciale viene svolto dalla "matrice dei dati". Quando la base empirica della ricerca viene, per così dire, oggettivata nella matrice dati il ricorso alla statistica e alla data analysis diventa obbligato e spesso massiccio, finendo talvolta per oscurare l'importanza delle altre fasi della ricerca (disegno, raccolta ed esposizione).

Per assolvere ai suoi compiti istituzionali un Istituto di ricerca come l'Ires si trova così a dover fare i conti con l'innumerevole serie di operazioni teoriche e pratiche che un tale tipo di ricerca impone. Ma soprattutto il fare ricorso (come accade nella maggior parte dei casi) alla ricerca basata sulla matrice dei dati comporta l'utilizzo di tre strumenti insostituibili: i dati, la statistica e l'informatica.

2. I dati

Per quanto riguarda i dati l'Ires si trova collocato in un interessante crocevia: da un lato grande utilizzatore di basi di dati prodotte altrove (spesso prodotte per scopi diversi dall'utilizzo statistico), dall'altro produttore e organizzatore di basi dati proprie ottenute tramite specifici progetti (osservatori) o tramite rilevazioni effettuate ad hoc.

Da un punto di vista tematico l'Istituto si occupa di una vasta gamma di argomenti che, a vario titolo, fanno da sfondo ai processi decisionali della Pubblica Amministrazione. Con riferimento specifico al loro ambito semantico, si spazia dai dati relativi agli aspetti economici e del mercato del lavoro, alla finanza pubblica, ai consumi culturali e all'istruzione, fino alle opinioni, agli atteggiamenti e ai comportamenti collettivi.

Anche solo questo sintetico quadro è sufficiente per comprendere come, da un punto di vista più tecnico, le basi dati su cui si deve operare siano assai differenziate. Per quanto riguarda il tipo di organizzazione delle matrici dati e le relative procedure statistiche si opera sia su serie storiche sia su matrici cross-section (o ecologiche). A livello di aggregazione dell'unità di analisi si lavora con matrici cosiddette "macro" (dati aggregati, generalmente per articolazione territoriale), "micro" (quando l'unità d'analisi è costituita da singoli protagonisti dell'attività socio-economica: individui, imprese) e, con particolare riferimento alla sfera degli studi sui comportamenti o le opinioni, con matrici "pico" (quando le unità della matrice si riferiscono tutte al medesimo individuo). Relativamente, infine, alle fonti è comune il ricorso a fonti statistiche ufficiali come l'utilizzo di dati provenienti da fonti amministrative, gestionali o da campagne di rilevazione appositamente condotte: le cosiddette "survey" o indagini campionarie.

Con riferimento all'ultimo aspetto di differenziazione tra i vari tipi di dati, quello delle fonti, mi sembra opportuno sottolineare il rapido cambiamento cui si è assistito negli ultimi anni. Cambiamento che ha riguardato sia

i fornitori ufficiali di dati statistici (l'ISTAT), sia i comportamenti e le scelte attuate dai ricercatori.

Non molti anni addietro nel campo delle fonti statistiche si assisteva, in Italia, ad una sorta di monopolio. I dati statistici provenivano quasi esclusivamente dall'ISTAT e comunque, l'accesso ad altre fonti, era molto spesso mediato da quest'ultimo organismo.

Così soltanto dieci anni fa, nel 1986, non si poteva non rilevare, come feci in un W.P. dell'Ires (scritto in collaborazione con L. Ricolfi), che << il quadro delle statistiche ufficiali del dopoguerra si presenta come un colabrodo. Indicatori disponibili per un certo periodo cessano di esserlo in quello successivo per riapparire, magari in forma leggermente modificata, in un periodo ancora successivo. Indicatori disponibili su base nazionale non lo sono su base regionale, o lo sono solo per certi sottoperiodi. Altri indicatori, magari disponibili fin dagli anni del dopoguerra, cessano improvvisamente e misteriosamente di esistere in un dato anno per riemergere, come fiumi carsici, in un periodo talora anche lontano senza alcuna plausibile spiegazione>> (Ires, W.P. n. 84, 1986). La strada percorsa da allora, soprattutto con le operazioni di ristrutturazione e decentramento avviate dall'ISTAT mi sembra molta e ha portato certamente innumerevoli vantaggi.

Accanto a ciò ha anche preso corpo un altro processo che, anche se per ragioni differenti, è altrettanto rilevante. Parallelamente allo sviluppo delle tecnologie informatiche e al rapido diffondersi di strumenti di elaborazione delle informazioni sempre più potenti si è assistito alla proliferazione di fonti di dati statistici molto differenziate. Sono diventati facilmente accessibili - anche per un trattamento statistico - dati non direttamente prodotti a tale scopo, ma provenienti da archivi amministrativi e gestionali, così come è divenuta più facilmente realizzabile la produzione di matrici dati ottenute tramite rilevazioni dirette e appositamente condotte.

Valga per tutti l'esempio di uno studio longitudinale della popolazione torinese portato avanti dall'area di epidemiologia della USL 5 di Torino, attraverso "l'aggancio" dei dati individuali del Censimento della Popolazione (per gli anni 1971, 1981 e 1991) ad altri archivi, quali quelli sulle cause di morte e sui soggetti assistiti dal Comune. Le potenzialità di una tale base di dati sono notevoli, non solo per gli studi di tipo epidemiologico, ma anche per quelli di tipo socio-anagrafico. E' per questo che l'Ires ha sviluppato una collaborazione con l'area di epidemiologia che ha già prodotto interessanti

risultati sui fenomeni di mobilità sociale nell'area torinese (Cfr. Relazione sulla situazione economica sociale e territoriale del Piemonte, 1995).

In genere si può dire che l'informazione di tipo amministrativo risulta sempre più importante nei tentativi di valutare gli effetti delle politiche pubbliche, un'esigenza sempre più sentita e spesso richiesta dagli stessi provvedimenti legislativi settoriali.

Anche grazie a questi eventi si è diffusa la consapevolezza che l'espressione "raccolta dati", comunemente usata nei manuali di metodologia, è un'espressione che non rappresenta correttamente il processo di oggettivazione della base empirica di una ricerca nella matrice dati. Con un'espressione colorita si può dire che << ... i cosiddetti "dati" non crescono nei prati e i ricercatori non li raccolgono, essi sono piuttosto "costruiti" dal ricercatore stesso attraverso procedure di interpretazione e di attribuzione di significato >> (A.P. Ercolani, A Areni, L Mannetti, 1990).

L'immagine che i dati esistessero indipendentemente e autonomamente dalle operazioni e dalle scelte che il ricercatore poteva compiere non era tuttavia soltanto il frutto di una visione "ingenua" della ricerca scientifica o di una totale dipendenza da una (o poche) fonti statistiche accreditate come "ufficiali". Essa traeva origine anche da una concezione positivista che ha pervaso (e talora ancora pervade) l'orientamento delle scienze, anche sociali. Concezione questa criticata da autori come Lakatos (1968) e Popper (1969) che hanno sostenuto la comune natura congetturale sia dei dati sia della teoria, in opposizione alla visione ottocentesca che attribuiva al dato la natura di "fatto oggettivo" (in base al quale la teoria può essere verificata o falsificata).

L'importanza del superamento di un approccio positivista consiste anche in una maggiore attenzione al dato e alle sue qualità. Nessun trattamento statistico, anche il più sofisticato e anche se condotto con il massimo rigore, può migliorare il contenuto informativo di una base di dati lacunosa o ancor peggio prodotta e gestita senza prestare la dovuta attenzione ai delicati processi di rilevazione (o oggettivazione nella matrice dati delle rilevanze empiriche).

Una scarsa attenzione alle fasi di "costruzione" del dato, alla delicatezza della catena di operazioni che lega i concetti alle variabili della matrice dati, così come alle caratteristiche delle variabili per quanto riguarda il livello di scala a cui sono espresse (nominale, ordinale, intervalli, ecc.), conduce a

rapporti di ricerca assolutamente inutili (quando non dannosi) sul piano dell'accrescimento delle conoscenze.

Come è stato ampiamente argomentato (A. Marradi, 1984; L. Ricolfi, 1985), la fase di raccolta dei dati e quella di analisi devono essere strettamente interconnesse. In questo senso è una convinzione assolutamente erronea (purtroppo talvolta presente anche tra addetti ai lavori) quella che attribuisce il ruolo della statistica solo alla fase di analisi dei dati. Uno dei tristi risultati di un tale convincimento è l'apparire di rapporti di ricerca che fanno ampio uso di tecniche statistiche di analisi dei dati anche molto sofisticate, applicate spesso su dati che non rispettano le più elementari assunzioni delle tecniche medesime. Gli esempi purtroppo sono numerosi e questa non è certamente la sede per entrare in dettagli tecnici. Tuttavia, e solo per fissare le idee, penso che a nessuno sfugga l'assurdità di una misura di centralità come la media aritmetica calcolata su una variabile espressa a livello di scala nominale, o come la stima di un modello di regressione lineare applicato a variabili categoriali di tipo ordinale.

3. L'analisi statistica

Entrando maggiormente nell'ambito dell'analisi dei dati va detto che, anche in questo caso, farò riferimento ad uno degli ambiti di attività dell'Istituto in cui l'apporto della statistica è fondamentale. Tralascerò pertanto l'utilizzo delle tecniche e dei modelli di simulazione per concentrare l'attenzione sul ruolo della statistica in tutte le ricerche che contemplano l'uso di una matrice dati come definito in precedenza. Verrà anche tralasciato quello che spesso viene considerato il più naturale (o addirittura l'unico) contributo della statistica alla ricerca socio-economica: la definizione delle caratteristiche dei campioni e le operazioni di campionamento.

Generalmente (quando si parla di matrice dati) si tratta di una matrice di profilo spesso detta "CxV" (casi per variabili), in teoria dei dati conosciuta come "two way-two mode" (Carroll e Arabie, 1980). Altri tipi di matrici (per esempio le matrici oggetto per oggetto come le matrici di distanze) vengono anch'essi trattati ma con frequenza decisamente inferiore alla classica matrice di profilo.

In un tale contesto è possibile riconoscere almeno tre tipi di operazioni di ricerca che si distinguono sia dal punto di vista epistemologico, sia per quanto riguarda l'adeguatezza e la liceità delle tecniche e dei modelli statistici che, di volta in volta, possono essere utilizzati.

Ai tre tipi di operazioni di ricerca corrispondono altrettante "domande" cui si cerca di rispondere (L. Ricolfi, 1995). Quando ci chiediamo "Com'é Y?" - dove Y è una qualche entità empirica (comportamento, fenomeno, insieme di casi) oggetto del nostro interesse - l'operazione di ricerca che stiamo compiendo è quella della descrizione. Risultati di ricerca che conducono ad affermazioni del tipo: "la maggior parte degli italiani possiede almeno un'automobile" oppure "il reddito percepito aumenta con il livello di istruzione" sono descrizioni, cioè affermazioni che non aggiungono nulla a quel che chiunque può desumere da un'ispezione della matrice dati. In questi casi sia che si utilizzino semplici strumenti statistici mono o bivariati, sia che si

faccia ricorso a tecniche di analisi più complesse come la correlazione, l'analisi in componenti principali o i modelli Log-lineari, l'analisi resta confinata, per così dire, alla superficie della matrice dati. In altri termini, cioè, la descrizione fornisce risposte che non pretendono di andare oltre, dietro o sotto ciò che i dati stessi sono in grado di indicare.

Quando invece cerchiamo di rispondere a domande del tipo "Perché Y?" o "Cos'é Y?" le operazioni di ricerca che stiamo compiendo vengono dette rispettivamente: spiegazione e interpretazione. In questi casi cerchiamo di andare in profondità, al di là, dietro o sotto il contenuto della matrice dati. Quando con una operazione di spiegazione giungiamo ad affermazioni del tipo: "l'istruzione è la principale determinante del reddito" significa che è stato possibile trovare un X (l'istruzione) che è sistematicamente collegato a Y (il reddito), ma soprattutto che ci sentiamo autorizzati ad imputare a tale relazione i caratteri della "necessarietà" (non-accidentalità) e dell'asimmetria (è X che influenza Y e non viceversa). Soprattutto quest'ultimo aspetto (la direzione della relazione causale) non appartiene al dominio delle rilevanze empiriche (dei dati), ma si tratta di "un di più" che ai dati viene aggiunto. Un nesso tra proposizioni verificabili (asserti nella terminologia di Marradi) che non può però essere verificato. <<I nessi fra asserti possono essere esposti al rischio della falsificazione, ma non possono in alcun modo essere provati. La loro verità resta ipotetica, e per così dire, negativa>> (Ricolfi, 1995). Di una spiegazione che ha superato i più stringenti e sofisticati controlli statistici tutto quel che possiamo dire è che è verosimile, o che non è stata contraddetta dai fatti.

Nel caso di un'operazione di interpretazione, inoltre, a questo - per così dire - elemento di arbitrarietà si aggiunge il fatto che una parte delle entità coinvolte sono puramente ipotetiche, inosservabili o, come si dice in gergo, latenti. Quest'ultima situazione la si incontra - per esempio - quando, nell'ambito delle operazioni di ricerca, si formulano affermazioni del tipo: "le risposte fornite dagli intervistati alle domande di atteggiamento z1, z2, ..., zk dipendono da una dimensione latente di xenofobia". Anche in questo caso andiamo ben oltre il contenuto della matrice dati. Quando si cerca di ricondurre una serie di risposte di un questionario a una o più dimensioni latenti (come avviene nei test di abilità o di intelligenza o nell'analisi degli atteggiamenti), si assume sempre l'esistenza di un livello noumenico, profondo o latente soggiacente al livello fenomenico, superficiale o manifesto

espresso dalla matrice dati. Tramite l'uso di tecniche come l'analisi fattoriale o l'analisi della struttura latente le relazioni osservate tra le variabili manifeste vengono imputate (o attribuite) a dei costrutti ipotetici non direttamente osservabili (come l'intelligenza, la xenofobia, l'abilità, ecc.).

Stabilito il quadro delle operazioni di ricerca che, tramite un appropriato uso della statistica, possono essere condotte su una base dati, vorrei qui "spezzare una lancia" a favore della descrizione. Spesso negletta e, ingiustamente considerata un livello "poco nobile" della ricerca scientifica, la descrizione viene troppo spesso relegata al ruolo di "ancella" delle più "blasonate" operazioni di spiegazione e di interpretazione. Questo stato di cose, a ben vedere, non ha alcuna giustificazione soprattutto se, come spesso si sente dire, viene motivato dalla scarsità di strumenti statistici adeguati e/o sufficientemente sofisticati.

Nelle pratiche di ricerca orientate alla descrizione esistono almeno due approcci sostanzialmente diversi e che è opportuno mantenere distinti anche con riferimento alle strategie, alle tecniche o ai modelli di analisi statistica che ne conseguono.

Il primo approccio alla descrizione può essere fatto risalire a quello che Amartya Sen chiama la descrizione come scelta: << la descrizione non è soltanto una questione di osservazione e riporto; essa comporta l'esercizio – forse difficile – della selezione>> (A. Sen, 1986). Da questo punto di vista, come ricorda lo stesso Sen, il criterio guida di una buona descrizione è la rilevanza.

Il secondo modo di intendere la descrizione è quello che Ricolfi definisce della rappresentazione. In questo secondo caso << ... alcuni compiti di selezione possono essere, per così dire, bypassati o "rimandati" ad una fase successiva. Prima di porsi domande precise su porzioni molto particolari e ristrette della matrice dati, il ricercatore può voler condurre una sorta di "esplorazione" della matrice dati stessa>> (L. Ricolfi, 1994). Da questo secondo punto di vista, il criterio guida della descrizione è l'accuratezza. Una buona descrizione, in questo caso, si misura sulla sua capacità di riprodurre i dati della matrice di partenza.

Non è difficile scorgere tra questi due "modi" della descrizione i relativi suggerimenti sulle procedure statistiche e sui modelli da adottare. Nel primo caso il ricercatore deve decidere quali sono le variabili rilevanti e soprattutto quali sono le specifiche domande cui vuole trovare una risposta. Immagi-

nando che si operi con variabili di tipo categoriale, sembra naturale prefigurare un percorso che proceda dal calcolo di distribuzioni, misure di centralità e di eterogeneità, passando per l'analisi degli indici di associazione e delle relative tabelle di contingenza, fino ad arrivare alla stima di modelli Log-lineari. Nel secondo caso, invece, quando la descrizione è intesa come rappresentazione, è più facile immaginare che il ricercatore procederà ad un'analisi delle corrispondenze (se le variabili sono categoriali) o ad una analisi in componenti principali (se le variabili sono cardinali).

Anche se con enfasi differente, entrambi gli approcci alla descrizione e le relative procedure statistiche, implicano un criterio di parsimonia. In altri termini, una descrizione, affinché sia tale, deve anche essere sintetica. Quest'ovvia constatazione mi sembra opportuna per prendere le distanze da un'altra cattiva abitudine talvolta presente nella ricerca socio-economica, così come in alcune pubblicazioni di Enti Pubblici che pensano in questo modo di sfruttare al meglio l'informazione amministrativa di base. Si tratta della, per così dire, abitudine di descrivere una matrice dati con decine e decine di tabelle a doppia entrata. Questo modo di procedere presenta almeno due gravi inconvenienti dai quali, tuttavia, è possibile sfuggire proprio grazie ad un più appropriato uso della strumentazione statistica.

Il primo inconveniente può essere battezzato come "noia". A voler presentare tutte le tabelle a doppia entrata che si possono ottenere con 4 variabili categoriali si devono riportare 6 tabelle. Ma attenzione con 7 variabili il numero di tabelle sale già a 21, e con 10 variabili il lettore verrebbe sommerso da ben 45 tabelle. La crescita esponenziale del numero delle tabelle, oltre alla noia nell'eventuale lettore, comporta però una conseguenza, per certi versi, più grave. L'elevato numero di tabelle che si dovrebbero produrre finisce per consigliare una drastica riduzione, quantomeno a livello redazionale, e il rischio che i criteri che hanno guidato tale ridimensionamento restino oscuri al lettore è spesso elevato.

Vi è tuttavia un secondo inconveniente di portata ancora più dirompente del precedente.

In statistica sono ben conosciute alcune "trappole" in cui è facile incorrere quando, come nel caso delle tabelle a doppia entrata, l'analisi è limitata al livello bivariato. Spesso vengono citate come i paradossi dell'analisi multivariata e portano i nomi degli statistici o dei metodologi che per primi

li hanno segnalati. Si parla così, solo per citarne alcuni, del paradosso di Tschuprov (1939), di Simpson (1951) e di quello di Eysenck (1970).

Senza entrare nei dettagli, si può brevemente illustrare il paradosso di Tschuprov, che ha a che fare con il concetto statistico di correlazione spuria, con questo famoso esempio. Incrociando il numero di bambini nati in alcune zone con il numero di cicogne presenti nelle stesse zone si scopre che esiste una forte e significativa relazione positiva. Dobbiamo forse mettere in dubbio le nostre convinzioni sui processi che conducono alla nascita dei bambini? Il paradosso è ben presto evidente se si accetta di passare da un'analisi bivariata ad una multivariata. Lo scopo è quello di "controllare" se la relazione tiene introducendo nell'analisi altre variabili. Così, considerando anche una variabile in grado di esprimere la ruralità o meno di ciascuna zona, è possibile constatare che le relazioni "parziali" che si ottengono tra numero di bambini nati e numero di cicogne sono decisamente più confortanti. I coefficienti di correlazione ottenuti controllando l'originaria relazione con quella terza variabile sono prossimi a zero e statisticamente non significativi.

Non si deve d'altra parte pensare che situazioni come quella appena illustrata siano rare o riguardino solo particolari contesti o specifiche variabili. Per fare un esempio concreto, in una recente ricerca dell'Ires sugli "Atteggiamenti e comportamenti verso gli immigrati in alcuni ambienti istituzionali" (Ires, 1995) è stato elaborato uno specifico indice di "apertura/chiusura" dei torinesi nei confronti degli immigrati extracomunitari. L'interesse era concentrato sulla relazione tra apertura verso gli immigrati e ambiente di lavoro (in particolare: vigili urbani e ospedalieri). Una serie di semplici analisi evidenziavano una stretta relazione tra le due variabili. In particolare emergeva come tra i vigili il grado di chiusura fosse di gran lunga più elevato di quanto lo fosse tra gli ospedalieri. Volendo attribuire all'ambiente di lavoro un effetto causale sull'atteggiamento si sarebbe (erroneamente) potuto concludere che lavorare a contatto con gli immigrati tra i vigili urbani comporta la chiusura, mentre lavorare nelle stesse condizioni in un ospedale comporta l'apertura. Utilizzando però sugli stessi dati un modello causale multivariato (l'analisi della varianza) e controllando l'originaria relazione con un insieme di altre variabili come sesso, età, istruzione, ecc., è stato possibile accertare che l'originaria relazione era spuria, non genuina. La conclusione cui si è potuti pervenire, in questa parte dello studio, è

stata così molto diversa dalla precedente suggestione. Controllando l'originaria relazione con la scolarità è stato possibile accertare che solo quest'ultima produce un effetto verso l'apertura e che, invece, lo svolgimento di un'attività lavorativa a diretto contatto con gli immigrati (sia essa tra i vigili o tra gli ospedalieri) produce una tendenza alla chiusura.

4. L'informatica

Naturalmente l'utilizzo di tecniche di analisi statistica dei dati adeguate e talvolta anche sofisticate necessita due requisiti minimi:

a) una sufficiente conoscenza della statistica e dei modelli di analisi dei dati;
b) la disponibilità di strumenti di elaborazione capaci di sopportare l'enorme quantità di calcoli che molte tecniche statistiche richiedono.

Per quanto riguarda il primo requisito vi è, tutto sommato, poco da dire e, forse, molto da fare. Si tratta ovviamente di organizzare corsi e seminari che permettano a tutti coloro che operano sui dati di agire con le necessarie competenze statistiche e metodologiche.

Il secondo requisito, quello degli strumenti informatici è certamente più un problema di dotazione che di competenza. L'enorme e rapido sviluppo ricevuto dagli strumenti di elaborazione negli ultimi anni è lì a testimoniare la sostanziale inesistenza di un problema formativo in ambito strettamente informatico. Se si esclude il campo della realizzazione di ampie basi di dati direttamente interrogabili dall'utente finale (le così dette banche dati) la competenza informatica richiesta a chi voglia trattare statisticamente la propria matrice dati è ormai quasi del tutto inesistente. Ben lontani siamo dai problemi che si dovevano risolvere quando l'uso di un main frame necessitava istruzioni di allocazione delle aree di memoria da destinare ai dati o al programma, e gli algoritmi di calcolo dovevano spesso essere tradotti in un linguaggio comprensibile sì dalla macchina, ma generalmente ostico e incomprensibile per gli esseri umani.

L'avvento dei primi packages matematico-statistici (per le scienze sociali in particolare l'SPSS e poi il SAS), nella prima metà degli anni '80, ha già drasticamente ridimensionato le difficoltà di approccio all'elaboratore. La disponibilità poi di potenzialità di calcolo enormi (ormai del tutto confrontabili con quelle di un main frame dei primi anni '80), a costi decisamente bassi e a disposizione di ciascuno sul proprio personal computer, ha

infine eliminato ogni residua esigenza di competenza informatica. L'ampia disponibilità di memoria sia per il calcolo sia per l'immagazzinamento delle informazioni, ha permesso un parallelo e impressionante sviluppo del software matematico-statistico. La stragrande maggioranza dei personal computer con i quali ci troviamo ad interagire, a casa come sul posto di lavoro, possiede risorse di gran lunga superiori a quelle necessarie per far "girare" i più diffusi programmi statistici. Questi ultimi sono stati dotati di interfacce utente molto familiari, riducendo generalmente il tempo di apprendimento all'uso del software a meno di una giornata di lavoro. Ogni package di questo tipo (mi riferisco in particolare ai due maggiormente diffusi: SAS e SPSS) dispone di una libreria di procedure statistiche così varia e ricca, che generalmente travalica di gran lunga, le esigenze di un utente medio.

In definitiva mi sembra si possa dire senza timore di smentita che, dal punto di vista, per così dire, tecnologico ci troviamo in una situazione in cui gli strumenti disponibili sono abbondantemente superiori alle necessità medie, ma anche medio-alte degli utilizzatori.

Un'analisi fattoriale, una regressione (lineare o non-lineare) che preveda qualche centinaio di variabili indipendenti con selezione stepwise dei predittori, una cluster analysis, ecc. sono eseguibili in pochi secondi e senza alcuna specifica competenza informatica.

Lo sviluppo tecnologico non ha però reso altrettanto accessibile la comprensione dei risultati delle procedure statistiche, né ha favorito una qualche sorta di controllo automatico sulla sensatezza delle operazioni che in modo così semplice e naturale vengono richieste ed eseguite dalla macchina. Per fare un esempio banale, immaginiamo di disporre di una variabile X che riporta le frequenze di adesione di una certa popolazione per ciascuna fede religiosa considerata (cattolici, protestanti, ebrei, musulmani; X=40;20;10;30). Se chiediamo a qualsiasi package in commercio di fornirci la media, esso diligentemente risponderà dicendoci che la media è pari a 25. Nessun turbamento, nessun avviso o segnalazione di errore verrà prodotta per avvisarci dell'assoluta insensatezza di quel risultato!

5. Conclusioni

Finora ho concentrato l'attenzione su due delle quattro fasi in cui, come dicevo all'inizio, si articola la ricerca socioeconomica. Vorrei concludere questo intervento con una considerazione su una delle due fasi finora non considerate: l'esposizione dei risultati. In particolare, uno dei problemi spesso sollevati a proposito di ricerche che fanno ampio ricorso all'uso della statistica, si riferisce alla cripticità, alle difficoltà di comprensione da parte dei non addetti, soprattutto di quei passaggi in cui vengono esposti o problematizzati i risultati di qualche procedura statistica. Senza dubbio il compito di farsi capire e di commentare in modo chiaro e accessibile i risultati di uno studio spetta a chi ha condotto la ricerca. Evidentemente l'esposizione deve considerare il fatto che il pubblico a cui è destinata non è tenuto ad essere un esperto delle specifiche procedure statistiche o dei modelli utilizzati, e ogni sforzo deve essere fatto per rendere comprensibile il risultato senza eccessivi appesantimenti di ordine tecnico.

Ciò detto, tuttavia mi sembra anche opportuno richiamare l'attenzione sulla "snumeratezza" che talvolta sembra affliggere i destinatari dell'informazione.

Il termine "snumeratezza" è stato introdotto in Italia, qualche anno addietro, da un simpatico volumetto del matematico americano John Allen Paulos (1992). << La snumeratezza - egli dice - cioè la mancanza di confidenza con i concetti fondamentali della matematica e della statistica, affligge un numero spropositato di cittadini per altro colti. Gli stessi che rabbrividiscono nel sentire confondere termini come "indurre" e "dedurre" non mostrano alcun disagio neppure di fronte agli svarioni numerici più macroscopici >>.

Un interessante esempio di snumeratezza ci è fornito dallo stesso Allen con questo breve aneddoto.

<< Un tizio viene aggredito in pieno centro da quello che secondo lui è un negro. Quando però gli inquirenti ripetono la scena più volte in analoghe

condizioni di luce, la vittima identifica correttamente la razza dell'assalitore soltanto nell'80% dei casi circa. Che probabilità ci sono che il suo aggressore sia veramente negro? Molti naturalmente risponderanno che le probabilità sono dell'80%, ma la risposta corretta, partendo da alcuni logici presupposti, è di gran lunga inferiore. Partiamo infatti dal presupposto che più o meno il 90% della popolazione sia bianca e che soltanto il 10% sia negra, che il centro in questione rappresenti tipicamente questa composizione, che non ci sia una razza più incline ad aggredire i passanti e che la vittima abbia le stesse probabilità di confondere la razza in entrambe le direzioni, negro al posto di bianco e bianco al posto di negro. Alla luce di tali premesse, in 100 aggressioni simili la vittima in media identificherà come negri 26 aggressori: l'80% dei 10 negri, ossia 8, più il 20% dei 90 bianchi, ossia 18, per un totale di 26. Di conseguenza, dal momento che soltanto 8 dei 26 identificati come negri sono effettivamente negri, la probabilità che la vittima sia stata aggredita da un negro è soltanto di 8/26, cioè approssimativamente del 31%!>>.

Una suggestione, infine, sul grado di snumeratezza che affligge la nostra società può essere ricavata dal numero di risposte esatte ottenute a questa domanda. << Un oggetto costa £ 1000. Un cartello indica che verrà scontato del 50%, decido di acquistarlo. Quando giungo alla cassa vengo informato di aver vinto un premio per cui l'oggetto in questione viene ulteriormente scontato del 50%. Quanto pagherò alla cassa per acquistare l'oggetto? >>.

La domanda è stata inserita in un questionario somministrato a 740 ragazzi in età compresa fra i 16-17 anni nell'ambito di uno studio dell'Ires (di prossima pubblicazione) sulle scelte scolastiche post obbligo degli adolescenti piemontesi. Tra costoro soltanto 94 ragazzi (meno del 13%) risultavano non iscritti ad alcun corso di istruzione secondaria. Ebbene il 53% (394 ragazzi) ha fornito una risposta errata!

Riferimenti bibliografici

Allen Paulos J., (1992), *Gli snumerati*, Milano, Leonardo Paperback.

Eysenck H.J., (1970), *Explanation and the Concept of Personality* In Borger R., Cioffi F. (Eds.), Explanation in the Behavioural Sciences, Cambridge, University Press, (trad. it. Borger R., Cioffi F., *La spiegazione nelle scienze del comportamento*, Milano, Franco Angeli, 1981).

Boudon R., (1970), *Metodologia della ricerca sociologica*, Bologna, Il Mulino.

Carroll J.D., Arabie P., (1980), *Multidimensional scaling*, in Annual Review of Psychology, 31, pp. 607-649.

Ercolani A.P., Areni A., Mannetti L., (1990), *La ricerca in psicologia. Modelli di indagine e di analisi dei dati*, Roma, La Nuova Italia Scientifica.

Ires, (1995), *Atteggiamenti e comportamenti verso gli immigrati in alcuni ambienti istituzionali*, Torino, Rosenberg & Sellier.

Lakatos I., (1968), *Criticism and The Methodology of Scientific Research Programmes* In Procedings of Aristotelian Society, N° 69, pp. 149-86.

Marradi A., (1984), *Concetti e metodi per la ricerca sociale* In Cardano M., Miceli R. (a cura di), Il linguaggio delle variabili, Torino, Rosenberg & Sellier, 1991.

Miceli R., Ricolfi L., (1986), *Archivio degli indicatori sociali. Un approccio costruttivista all'organizzazione dei dati*, Torino, W.P. N° 84, IRES.

Popper K.R., (1969), *Conjectures and Refutations*, London, Routledge & Kegan Paul, (trad. it. *Congetture e confutazioni*, Bologna, Il Mulino, 1972).

Ricolfi L., (1985), *Operazioni di ricerca e scale* In Cardano M., Miceli R. (a cura di), Il linguaggio delle variabili, Torino, Rosenberg & Sellier, 1991.

Ricolfi L., (1995), *L'arte della descrizione. Un'introduzione alla ricerca standard*, (dattiloscritto), Dispense del corso di Metodologia delle Scienze Sociali (a.a.93-94).

Sen A., (1986), *Scelta, benessere, equità*, Bologna, il Mulino.

Simpson E.H., (1951), *The Interpretation of Interaction in Contingency Tables* In Journal of the Royal Statistical Society, Series B, 13, 238-249.

Tschuprov A.A., (1939), *Priciples of the Mathematical Theory of Correlation*, London,.