ANDREA CICCARELLI

METODO STATISTICO, INNOVAZIONE TECNOLOGICA E LA REGOLA NON SCRITTA DEL BUON SENSO

Non è semplice condensare in poche pagine i quindici anni di vita, di esperienze e di "scappellotti" che ho passato al fianco del mio Maestro. Rimarranno nel mio cuore la passione, le (animate) discussioni, i confronti sulla Magica e gli innumerevoli consigli. La sintonia è tale che ciò che segue è uscito dalla mia penna, ma ho la presunzione di ritenere che potrebbe essere allo stesso modo il frutto del Suo pensiero.

1. Premessa. La Statistica e, in particolare, la Statistica economica traggono la loro esistenza dalla necessità di osservare, in termini quantitativi e qualitativi, tutti quei fenomeni demografici, sociali, economici, fisici, biologici, etc. che quotidianamente possiamo incontrare.

La necessità di numerare e quantificare i fatti che ci circondano non è certo una novità della società moderna, potendosi riscontrare esempi (o tentativi) di misure quantitative già in civiltà molto lontane: nell'antico Egitto, infatti, esisteva un sistema sufficientemente progredito di organizzazione statistica, che consentiva, fin dalla prima dinastia, di rilevare l'ammontare della popolazione mediante censimenti, e, successivamente, di effettuare rilevazioni a fini fiscali relative a vari beni; in Cina già in epoca pre-cristiana veniva attribuita fondamentale importanza alla conoscenza dell'ammontare della popolazione, e fin dalla dinastia dei Ming (XIV-XVII secolo) i censimenti venivano effettuati con cadenza decennale e si tenevano registri della popolazione che riportavano sesso, età e professione degli individui; ben nota è anche l'attenzione riservata dalla Roma antica alle rilevazioni censuarie[1], che fin dal VI secolo a. C. si ripetevano ogni cinque anni, e nelle quali i cittadini dovevano dichiarare il pro-

[1] Non a caso, il termine deriva proprio dal latino *censere*, e sulla base del *census* veniva ripartita la popolazione romana in cinque classi, in virtù delle quali ogni cittadino si vedeva assegnato il suo posto nell'organizzazione politica e amministrativa della società, e in base alle quali veniva stabilito il livello di tassazione da sopportare.

prio nome, quello del padre, l'età, il nome della moglie e dei figli e l'ammontare dei propri beni[2].

Taluni, tuttavia, fanno risalire la vera e propria nascita di tali discipline in epoca più recente, e più precisamente al XVII secolo, quando William Petty pose al centro dei suoi studi l'Aritmetica politica, presentandola come quella scienza che opera attraverso «numeri, pesi e misure» e definendola come «l'arte di ragionare mediante le cifre sulle cose che riguardano il governo». Fu in tal modo che si propose di effettuare delle "stime" della popolazione e delle quantità e tipologie di beni di cui questa aveva bisogno, oltre che di analizzare le caratteristiche dell'apparato produttivo preposto a fornire i suddetti beni[3].

Non è certo intenzione di chi scrive effettuare una rassegna storiografica di quanti, con il loro contributo, hanno consentito di portare la disciplina statistica al livello cui è oggi[4]; ci limiteremo a considerare il fatto che almeno nel nostro Paese la Statistica e i metodi quantitativi in genere sono visti con un certo sospetto, forse perché la nostra cultura si fonda principalmente su basi umanistiche, il che contribuirebbe a munire di un'aurea di verità (e non solo di leggenda…) una storia da sempre raccontata dal mio Maestro: in una scuola elementare di qualche anno fa, dal colloquio genitori-docenti i primi apprendono che il proprio figlio di 10 anni aveva scritto acqua con due "q". In seguito a tale "malefatta", i genitori non avrebbero lesinato al malcapitato le più "feroci" e "dure" punizioni, arrivando, pare, a proibire la televisione e la nutella[5] per numerosi giorni. Successivamente, durante il colloquio con il docente di matematica, gli stessi genitori avrebbero saputo che il proprio figliolo aveva scritto che due più due fa 27, e con sguardo perso ed aria sconsolata, aprendo le braccia in segno di resa, avrebbero replicato all'interlocutore: "professore, che vuole fare…la matematica è difficile...".

[2] Una più accurata ed esaustiva trattazione di tali argomenti può essere trovata in G. LETI, *Statistica descrittiva*, Il Mulino, Bologna, 1983, dal quale è stato ripreso quanto sopra riportato.

[3] W. PETTY, *Aritmetica politica*, Liguori ed., Napoli, 1986.

[4] Lasciamo questo arduo compito all'abile penna di Alighiero Erba, straordinario maestro e amico sincero, che ha ripercorso con grande precisione tutti quei passaggi storici che qui, forzatamente, dobbiamo evitare; si veda A. ERBA, *Scritti di statistica economica*, Giappichelli editore, Torino, 2012.

[5] Deve essere una scena di molti anni orsono, dal momento che oggi, causa olio di palma, la nutella verrebbe tolta forse con altri obiettivi.

2. La statistica tra metodologie di stima e interpretazione delle stime. Obiettivo della statistica dovrebbe essere quello di determinare in termini quantitativi la realtà osservata, sulla base, naturalmente, delle ipotesi poste in essere dal ricercatore. Modificando le ipotesi iniziali, evidentemente, può variare (in termini anche consistenti) il risultato finale. Proprio per tale motivo i principali aggregati economici e sociali vengono stimati a partire da procedure condivise a livello internazionale, non solo per quanto attiene alle metodologie statistico-matematiche da impiegare, ma soprattutto per quanto concerne la definizione degli aggregati di base.

Proprio per questo motivo, ad esempio, nel contare gli occupati, il nostro ufficio centrale di statistica non si limita semplicemente ad individuare "*quelli che lavorano*" (come l'uomo della strada sarebbe portato a fare), ma tutti coloro con almeno 15 anni di età che, nella settimana a cui l'intervista fa riferimento, hanno svolto almeno un'ora di lavoro in una qualsiasi attività che prevede un corrispettivo monetario o in natura [6], secondo una metodologia che non solo è stata condivisa in ambito Eurostat, ed è ormai consolidata da anni di rilevazioni, ma che, soprattutto, garantisce la confrontabilità degli aggregati tra tutti i paesi dell'Unione Europea.

Appare del tutto evidente che un tasso di occupazione (o di disoccupazione) eventualmente calcolato a partire da questi dati risente delle definizioni adottate e, pertanto, può fornire informazioni "corrette" nell'ambito, come già precedentemente evidenziato, delle ipotesi utilizzate in partenza.

Piuttosto, siamo talvolta costretti ad assistere ad un "disinvolto" utilizzo delle informazioni statistiche prodotte, sia a causa della scarsa attenzione che viene data al modo in cui il dato statistico viene prodotto, sia a causa della limitata disponibilità di informazioni quantitative, che a volte spinge l'utilizzatore finale a *trascinare* il significato dei numeri ben oltre il limite consentito (dal buon senso).

In tale direzione, è esemplificativo il caso del Pil (Prodotto interno lordo) che, costruito con l'intenzione di fornire una stima del valore dei beni e dei servizi finali prodotti in una certa area geografica in un determinato arco temporale, viene spesso fatto passare come indicatore di ric-

[6] Per brevità di esposizione abbiamo evidenziato solamente tale fattispecie, ma in realtà rientrano nella definizione di occupato anche coloro che hanno svolto almeno un'ora di lavoro non retribuito nella ditta di un familiare nella quale collaborano abitualmente o coloro che sono momentaneamente assenti dal lavoro (ad esempio, per ferie, malattia o Cassa integrazione). Per una dettagliata descrizione degli aggregati e delle metodologie di rilevazione si può consultare la documentazione al seguente link: http://www.istat.it/it/archivio/8263.

chezza o addirittura di benessere, spingendo ben oltre il lecito le capacità informative dell'aggregato in questione. Da questo errato utilizzo derivano a volte interpretazioni non congruenti con la realtà percepita[7], che spingono qualche commentatore a dubitare addirittura della correttezza del metodo di calcolo, ma mai a mettere in discussione l'eventualità che all'aggregato in questione sia stato chiesto di sintetizzare una realtà che, per sua stessa costruzione, non era in grado di riassumere.

Il problema della corretta comunicazione dell'informazione statistica appare oggi ancora più sentito, alla luce della grande facilità di accesso che anche i comuni cittadini hanno ai dati: manca, purtroppo, una cultura statistica che faccia comprendere, da un lato, la necessità di approfondire il contenuto degli aggregati che si vanno a studiare[8] (indispensabile al fine di una corretta interpretazione dei risultati) e che, dall'altro, consenta di prendere atto che i dati disponibili rappresentano pur sempre una *media* di situazioni che, potenzialmente, potrebbero essere anche molto lontane tra loro[9].

3. Il contributo delle nuove tecnologie alla produzione di dati statistici. L'innovazione tecnologica ha portato innumerevoli vantaggi, primo fra tutti la possibilità di processare moli di dati (anche di grandissime dimensioni) utilizzando un semplice *personal computer* casalingo. L'impatto che questa ha avuto sull'analisi statistica è stato enorme: si pensi alla possibilità di elaborare complessi calcoli matriciali che ha consentito l'utilizzo e lo sviluppo di metodologie statistiche prima confinate soprattutto nel limbo della teoria, e che ora possono essere comunemente utilizzate senza eccessivo impiego di risorse.

[7] Sempre ammesso, poi, che le percezioni dei singoli possano aver priorità sulle oggettive misurazioni…

[8] Bisogna, in altre parole, familiarizzare con i cosiddetti *metadati*, ossia quelle informazioni che descrivono i dati statistici che ci apprestiamo ad utilizzare; si tratta, in sostanza, di leggere il "libretto delle istruzioni" del dato statistico, operazione che rimane ai più solitamente indigesta, ma che faremmo perfino al momento dell'acquisto di una nuova lavatrice.

[9] La variabilità dei fenomeni…questa sconosciuta…: è una caratteristica che praticamente ogni fenomeno (sia esso di natura sociale, demografica, economica, biologica) possiede, e che, possiamo dire, giustifica l'esistenza stessa delle metodologie statistiche, che hanno come prioritario obiettivo quello di sintetizzare le diverse situazioni, per quanto lontane queste possano essere tra loro. I detrattori del metodo statistico, a questo punto, tirerebbero fuori la celeberrima storiella dei due polli rievocata da Trilussa (TRILUSSA, *Tutte le poesie*, a cura di Pietro Pancrazi, Arnoldo Mondadori Editore, Milano), il quale, fine poeta, avrebbe forse tratto giovamento dall'utilizzo di una misura di variabilità per riuscire a comprendere correttamente l'effettivo utilizzo dei due animali.

A dire il vero, l'influenza non si è avvertita solamente nella fase di trattamento dei dati, ma anche per quanto attiene alla loro produzione: oggi, infatti, la quantità di dati generata è abnorme, e deriva dall'utilizzo dei telefoni cellulari, delle carte di credito usate per gli acquisti, dei gps, della televisione, delle infrastrutture intelligenti delle città, dei sensori montati sugli edifici, sui mezzi di trasporto pubblici e privati e via discorrendo.

Si usa riferirsi a tutti i dati così potenzialmente disponibili con la locuzione *Big Data*. Con tale termine si è soliti identificare l'enorme volume di dati (strutturati o non strutturati) noti per essere (spesso) facilmente reperibili sul web e difficili da elaborare utilizzando le tradizionali tecniche per i database comuni o utilizzando metodi statistici di base. In realtà, il problema non risiede tanto (o solo) nel volume, quanto, piuttosto, nella loro frammentazione e variabilità, elementi che portano alla necessità di combinare tecniche di analisi differenti (strutturate e non strutturate) al fine di estrarre risultati significativi[10].

L'enorme mole di informazioni prodotte quotidianamente e la possibilità di trattarle ed immagazzinarle fornita dai recenti sviluppi della tecnologia, hanno sollecitato gli "appetiti" degli utenti, stimolando i ricercatori ad implementare tecniche e metodologie per consentire un utilizzo di tali informazioni a fini statistici.

Ma siamo pronti, allo stato attuale, a produrre informazioni statistiche attraverso l'utilizzazione ed elaborazione dei Big Data?[11]. Avere a disposizione questa enorme mole di dati non equivale automaticamente ad avere informazioni (statistiche) accurate o opportune, per questo è necessario che il loro trattamento avvenga con metodi rigorosi e dal valore accertato, sia sotto il profilo tecnico che etico. Solo attraverso analisi statistiche robuste è possibile scoprire modelli nascosti, correlazioni sconosciute e altre

[10] L'innovazione nelle metodologie e nella produzione di informazioni statistiche non si limitano e (lo auspichiamo!) non si fermano all'avvento dei Big Data; per limiti di spazio abbiamo scelto questo come elemento esemplificativo di una tendenza che, ci sembra, stia spostando l'asse del ragionamento non sugli obiettivi dell'analisi quanto, piuttosto, sui metodi, in una continua sovrapposizione dei due piani che, spesso, tende a confondere gli strumenti con le finalità.

[11] La produzione di informazioni statistiche è un vero e proprio processo produttivo, che prevede diverse fasi e che può essere valutato alla luce di alcune dimensioni di qualità che il dato statistico deve possedere. In accordo con quanto stabilito a livello europeo, le dimensioni della qualità dei dati statistici sono le seguenti: pertinenza; accuratezza ed affidabilità; tempestività e puntualità; coerenza e comparabilità; accessibilità e trasparenza; costi e oneri dei rispondenti. Si veda a tal proposito EUROSTAT, *ESS Handbook for Quality Reports. 2014 Edition*, Luxembourg, 2015.

informazioni potenzialmente strategiche non immediatamente ricavabili dall'insieme dei dati.

Nessuna indagine statistica, del resto, appare esente da "criticità": un'indagine campionaria, ad esempio, è affetta dal cosiddetto *errore di campionamento*, che rappresenta, in un certo senso, il "costo" che dobbiamo sopportare per non aver potuto (o voluto) effettuare l'indagine sull'intera popolazione di riferimento. Ciò che consente, tuttavia, di utilizzare tali indagini senza timori sta proprio nel fatto che il suddetto errore di campionamento è misurabile, e la sua quantificazione rappresenta elemento indispensabile per interpretare correttamente le stime prodotte.

Per quanto attiene ai Big Data, invece, ci sembra di poter evidenziare una certa tendenza ad esaltarne le potenzialità, sottovalutandone i difetti: ad esempio, metter insieme tutte le informazioni desumibili dai tracciati gps utilizzati in un certo arco di tempo non significa avere a disposizione dati sugli spostamenti di tutta la popolazione, ma solo di una parte di essa che difficilmente potrà essere considerata un campione rappresentativo della prima; si potrebbe trattare, in sostanza, di un esemplare caso di *selection bias* (ossia distorsione da selezione), che rischia di minare anche in modo consistente l'accuratezza[12] delle stime.

Tale disinvolto utilizzo dei dati di base potrebbe portare, sull'onda delle mode del momento, a far passare alcune idee a nostro avviso totalmente fuorvianti[13], ovvero: che qualsiasi analisi dei dati può produrre risultati accurati; che ogni singolo dato può essere catturato rendendo obsolete le tradizionali tecniche di indagine; che non sia necessario riconoscere il nesso di causa ed effetto tra le variabili, perché le correlazioni tra queste ci raccontano tutto ciò di cui abbiamo bisogno; che non necessitiamo di sofisticati modelli statistici in quanto, con dati sufficientemente abbondanti, i numeri parlano da soli.

Non vorremmo, insomma, che passasse l'idea che basta trovare da qualche parte un insieme più o meno grande di informazioni e che queste, dopo essere state messe in un enorme frullatore, possano in modo spontaneo produrre informazioni statistiche.

4. Le recenti disquisizioni intorno al *p-value*. Ha suscitato molto

[12] Per accuratezza si intende il grado di vicinanza delle stime al valore esatto (o "vero") che la statistica proposta si prefigge di misurare (si veda EUROSTAT, 2015, *op. cit.*).

[13] Si veda a tal proposito l'articolo di Tim Harford: *Big data: are we making a big mistake?,* apparso sul Financial Times il 28 Marzo 2014 e disponibile sul sito www.ft.com, che fornisce una panoramica particolareggiata (ancorché, forse, non ancora esaustiva) dei potenziali rischi collegati all'utilizzo dei Big Data a fini statistici.

clamore, alcuni mesi orsono, la presa di posizione dell'*American Statistical Association* (ASA) sul corretto significato da attribuire al *p-value* e, quindi, sul suo idoneo utilizzo[14]. Il tutto era stato originato da alcune recenti considerazioni sulla "capacità" del *p-value* di sostenere le evidenze empiriche e sull'arbitrarietà della scelta di soglie (come quella $p \leq 0{,}05$) in base alle quali accettare/rigettare la cosiddetta ipotesi nulla.

Nel documento, l'ASA evidenzia, tra le altre cose, come il *p-value* rappresenti una misura di quanto i dati analizzati siano incompatibili con lo specifico modello statistico testato e, soprattutto, come non sia corretto basare le proprie conclusioni scientifiche o le politiche da intraprendere solamente sul superamento (o meno) di una determinata soglia di valore, ma prendendo in considerazione anche ulteriori elementi, quali il disegno dello studio, la validità delle assunzioni sottostanti l'analisi, la qualità delle misure effettuate. In sostanza, il suggerimento sarebbe quello di non prendere decisioni in modo a*utomatico* (magari perché è stato trovato un *p-value* pari a 0.049), ma, con un po' di buon senso, mettendo a sistema tutte le informazioni ottenute dall'analisi effettuata.

Nel leggere tutti gli interventi che ne sono seguiti, ho pensato: ma ce n'era veramente bisogno? Possibile che ci siano utilizzatori dei metodi statistici che si accontentano di ragionare su una "linea di demarcazione" e non si insospettiscano quando il risultato delle analisi si trova proprio in prossimità della suddetta linea? Del resto, soprattutto recentemente, era diffusa la pratica di accompagnare sempre all'analisi il valore di *p*, piuttosto che limitarsi a decretare la significatività (o meno) dell'assunto di base[15] al superamento di una soglia in qualche modo (pre)determinata.

All'inizio di ogni mio corso mi affanno a spiegare agli studenti che la *verità* la conosce solamente il Padreterno; noi, utilizzando le metodologie quantitative, possiamo ambire a misurare la realtà circostante, approssimandola per quanto possibile attraverso le stime dei parametri che ci interessano. Il processo decisorio, poi, non può essere demandato automaticamente all'*output* della nostra analisi: appare essenziale un intervento del ricercatore, che con la sua esperienza è in grado di orientarsi attraverso il dato quantitativo, proponendone una valutazione, un'analisi ed una sintesi che, opportunamente integrate con gli assunti di base, consentano di pren-

[14] R. L. WASSERSTEIN, N. A. LAZAR, *The ASA's Statement on p-values: Context, Process, and Purpose*, in The American Statistician, 70 (2), 2016, pp. 129-133, DOI: 10.1080/00031305.2016.1154108.

[15] In verità, tale abitudine sembrerebbe essere più consueta nell'ambito della letteratura economica e sociale, mentre in campo medico e scientifico l'utilizzo della formula dicotomica (significativo/non significativo) parrebbe maggiormente riscontrabile.

dere le più corrette decisioni in condizioni di incertezza. Soprattutto, mi sforzo di far comprendere come risultati prossimi ad una soglia (in qualsiasi modo essa sia stata scelta) sono da ritenersi estremamente pericolosi, e vanno giudicati *cum grano salis*: innanzitutto, proprio perché la soglia non ci è stata fornita da un'entità divina[16], poi in quanto piccoli "spostamenti" del *p-value* intorno alla suddetta soglia potrebbero essere la conseguenza non solo dell'evidenza empirica, ma anche, ad esempio, di ingannevoli assunti di partenza o della (parziale) inadeguatezza del modello statistico utilizzato.

5. Alcune considerazioni conclusive. L'innovazione tecnologica ha consentito enormi passi in avanti nella raccolta, nella conservazione e nella distribuzione di dati statistici, rendendo (ahimè) accessibile a molti (forse a tutti!) l'utilizzo di metodologie quantitative per leggere e sintetizzare i suddetti dati.

Il problema non sta nel far girare i modelli quantitativi: oggi basta cliccare col mouse sul pulsantino giusto, e un *software* statistico è in grado di offrirci meravigliose stime dei parametri dei nostri modelli; le difficoltà sorgono quando persone poco informate (sui metodi e sulle loro potenzialità) tentano di interpretare quei risultati. Ci sentiamo, in questo, molto vicini al biostatistico Jeff Leek, che si è inserito in questo modo nel dibattito riportato nel paragrafo precedente: «Il problema non è che le persone utilizzano malamente il *p-value*; ma che la maggior parte delle analisi dei dati che possiamo leggere non viene effettuata da persone che sono state formate per realizzare analisi dei dati»[17].

Paradossalmente, sembrerebbe come se lo sviluppo nelle possibilità di applicazione delle tecniche abbia non solo reso più semplice la costruzio-

[16] Proviene spesso dalla prassi, ma in fondo altro non è che una libera scelta del ricercatore. Nessun assunto e nessun modello potranno mai essere "esatti"; ma le eventuali distorsioni generate da questa mancata "esattezza" potranno incidere, evidentemente, in modo limitato. In altre parole, se troviamo un *p-value* inferiore a 0,0001 sarà difficile immaginare che tale scostamento dal modello di riferimento possa essere stato causato dalla distorsione dovuta ad eventuali non corrette assunzioni iniziali (anche se non va mai dimenticato che il ben noto fenomeno passato alla storia come "Maledizione del Faraone" potrebbe essere sempre in agguato…). Un valore *border line*, invece, dovrebbe far suonare un campanello di allarme, e suggerire quantomeno ulteriori verifiche empiriche (utilizzando, ad esempio, modelli differenti oppure prendendo in considerazione numerosità campionarie maggiori, che, come noto, consentono di diminuire la variabilità delle stime).

[17] Leek, J. (2014), "On the Scalability of Statistical Procedures: Why the p-Value Bashers Just Don't Get It," Simply Statistics Blog, Available at http://simplystatistics.org/2014/02/14/on-the-scalability-of-statisticalprocedures- why-the-p-value-bashers-just-dont-get-it/.

ne di automatismi interpretativi, ma abbia reso automatica (mi spingo a dire dicotomica) anche la scelta dei ricercatori impegnati nel processo decisorio, rendendo possibile a chiunque fregiarsi del titolo di *data scientist* (per usare un termine tanto in voga ultimamente). Niente di più sbagliato! Come direbbe a questo punto il mio Maestro: "la vita difficilmente porta a scegliere tra il bianco e il nero: è nelle *gradazioni del grigio* che il ricercatore deve mostrare le sue capacità di intercettare correttamente la realtà"; ed in tale crepuscolo la migliore luce può essere fornita solo da uno statistico preparato. L'enorme mole di dati e la grande velocità con la quale diventano (apparentemente) obsoleti, sembrano spingere, invece, verso una richiesta di risposte sempre più immediate, elemento che contrae i momenti di riflessione e, conseguentemente, può condurre a sintesi solo parziali (e, per questo, talvolta fuorvianti).

Del resto non sembriamo esser i soli ad aver maturato tale pensiero: già nel 2009 Hal Varian (noto economista e attuale *Chief Economist* di Google), dichiarava che le aziende «più ancora che avere la capacità di raccogliere le informazioni, devono avere quella di utilizzarle per conoscere più a fondo la clientela. L'abilità di maneggiare i dati – di capirli, processarli, estrarne valore, visualizzarli e comunicarli – diventerà di gran lunga la conoscenza più importante nei prossimi dieci anni… probabilmente il lavoro più richiesto e appagante dei prossimi dieci anni sarà quello dello statistico». Colpa degli statistici è stata proprio quella di non comunicare nel modo corretto queste abilità; speriamo di non avere solamente tre anni per poter invertire la tendenza.