I DIRITTI NELLA "RETE" DELLA RETE

Collana diretta da Franco Pizzetti

GIUSEPPE D'ACQUISTO

DECISIONI ALGORITMICHE

Equità, causalità, trasparenza

G. Giappichelli Editore

Capitolo Primo
Decisioni algoritmiche: opportunità e rischi

SOMMARIO: 1. Dalle decisioni umane alle decisioni algoritmiche. – 2. Le fasi di una decisione algoritmica. – 3. Decisioni algoritmiche e fallacie logiche: il paradosso di Simpson e il paradosso di Berkson. – 4. Il problema della discriminazione algoritmica. – 5. Verso una "narrazione" dei dati.

1. Dalle decisioni umane alle decisioni algoritmiche

In questo libro affronteremo il tema delle decisioni algoritmiche, ossia analizzeremo quei casi in cui il giudizio ultimo su un fenomeno è rimesso a una macchina, che prende decisioni sulla base dell'analisi dei dati disponibili e attraverso l'impiego di modelli matematici, o algoritmi, senza l'intervento dell'uomo. Sono situazioni destinate a diventare molto frequenti, sia perché in molti ambiti il giudizio della macchina può essere più accurato di quello umano (si pensi alla diagnostica per immagini nel settore medico), e dunque avremo un progressivo effetto di sostituzione uomo-macchina nell'analisi di specifici fenomeni, sia perché il nuovo contesto tecnologico caratterizzato da un'ampia disponibilità di dati, dalla pervasività di dispositivi per la loro raccolta, e dall'incremento della capacità computazionale delle macchine (contesto tecnologico che raggruppiamo sotto il nome di intelligenza artificiale) creerà molte nuove occasioni in cui sarà necessario esprimere un giudizio con prontezza e direttamente in una forma automatizzata. Le ragioni di questa delega decisionale alla macchina da parte dell'uomo sono molteplici e forse ineluttabili: l'uomo deve infatti arrestarsi di fronte al volume di dati che viene generato costantemente, ben oltre le proprie capacità di memorizzazione e di computazione, e per estrarre una conoscenza da questi dati serve l'ausilio della macchina. Un ausilio che, proprio a causa dei volumi dei dati e della varietà e velocità delle decisioni, dovrà essere sempre più autonomo dall'intervento dell'uomo perché produca risultati efficaci. L'aspetto quantitativo è essenziale per comprendere la scala e l'ineluttabilità di questa autonomia decisionale. A livello di informazione disponibile, è sotto gli occhi di tutti la capacità di generare e scambiare informazioni resa oggi possibile dai dispositivi che impie-

ghiamo in ogni ambito, da quello lavorativo a quello ricreativo, e non inferiore è la quantità di informazione tecnica (dati di localizzazione o di temporizzazione, immagini) che le macchine si scambiano tra loro per renderci disponibili i servizi che comunemente usiamo e che sempre di più si spostano verso una fruizione in modalità digitale (i servizi bancari, le relazioni con le pubbliche amministrazioni, ogni genere di acquisto di beni, ecc.), e per il loro stesso corretto funzionamento. Si tratta di un giacimento di informazioni inestinguibile e in continua crescita all'interno del quale è possibile trovare la risoluzione di problemi specifici (possiamo migliorare l'efficienza dei trasporti e ridurre il numero di incidenti? oppure, impiegare meglio le risorse e inquinare di meno? possiamo prevedere la diffusione di un virus o addirittura prevenirlo? e così via). Già oggi la quantità di informazione prodotta ha superato la soglia degli zettabyte (10^{21} byte) su base annua ovvero, a livello pro-capite e immaginando che questa informazione sia uniformemente distribuita su ogni abitante del pianeta, qualche decina di gigabyte al giorno. L'equivalente in termini di bit di un paio di film ad alta definizione prodotti ogni giorno da una molteplicità di fonti per ciascuno di noi. Su un piano infrastrutturale, con la versione 6 del protocollo IP (IPv6), che consente a tutti i computer del mondo di scambiarsi informazioni su internet, sarà teoricamente possibile indirizzare e mettere in collegamento ogni cosa con ogni altra cosa esistente sulla faccia della Terra. Per intenderci (ma è un'immagine molto al di sotto delle potenzialità della tecnologia), è come se ogni oggetto che ci circonda, e ogni parte minuta di cui è composto, potesse essere dotato di un sensore per comunicare il proprio stato di funzionamento, o se ogni singola foglia di ogni albero piantato sulla terra potesse fare lo stesso per dirci se è stata irrorata meglio o peggio della foglia a fianco.

Siamo ben oltre la capacità dell'uomo di governare "manualmente" questi flussi di dati, di analizzarli per estrarre conoscenza e di formulare con prontezza decisioni, che dovranno essere in larga parte automatizzate, e si apre la prospettiva molto concreta di un allargamento dello spettro delle decisioni che si potranno assumere. Oggi la decisione e il giudizio dell'uomo si applicano a un numero molto limitato di fenomeni. Decidere è un'operazione lenta e costosa per l'uomo. Lo è la raccolta dei dati e la loro interpretazione e, per la presenza di questa lentezza e di questo "filtro economico", l'uomo limita le occasioni di decisione e di giudizio a circostanze ben codificate. Ad esempio, la violazione di una legge o di un principio etico può dare luogo ad un giudizio, oppure la necessità di raggiungere un obiettivo può innescare una successione di decisioni strategiche. Si tratta di un numero certamente rilevante di situazioni, ma comunque finito. A misura d'uomo.

L'intervento della macchina e l'ampia disponibilità di misure sullo "stato del mondo" aumenterà enormemente il numero di situazioni in cui si potrà prendere una decisione o esprimere un giudizio. Per stare agli stessi semplici esempi degli oggetti e delle piante prima richiamati, si potrà giudicare se quella singola vite di quel determinato componente sia stata ben serrata o se non

sia stata invece la causa di un certo malfunzionamento, oppure se la singola pianta di un campo abbia ricevuto la giusta quantità di acqua o se non sia stata l'assenza o l'abbondanza di acqua ad avere compromesso il raccolto. E da questo allargamento dello spettro discenderanno, come è facilmente intuibile, nuove forme di responsabilità. È opportuno allora riflettere sulle caratteristiche delle decisioni che una macchina potrà assumere attraverso l'impiego di algoritmi senza l'intervento dell'uomo confrontandole con le caratteristiche tipiche delle decisioni dell'uomo. Questo passaggio dalle poche decisioni umane alle molte decisioni algoritmiche richiederà, come vedremo, un sostanziale ripensamento di alcune nostre categorie concettuali. Occorre affrontare questo percorso senza preconcetti: l'automazione delle decisioni porterà grandi benefici all'uomo consentendogli la risoluzione rapida di molti problemi in diversi ambiti, che senza l'ausilio di algoritmi non sarebbero neppure affrontabili, ma è un'opzione non priva di rischi che vanno ben individuati per poter essere affrontati, sia su un piano tecnico, sia su un piano di natura normativa.

Innanzitutto, per via dei limiti naturali della capacità di memoria e di calcolo umani, le nostre decisioni sono spesso il frutto di una informazione incompleta sui fenomeni osservati e di una razionalità soltanto parziale, che non di rado possono dare luogo a contraddizioni. Il contesto in cui la decisione umana è assunta conta poi moltissimo: esistono infatti consuetudini, specificità culturali, interessi individuali di varia natura che possono condizionare, anche fortemente, le decisioni assunte dall'uomo. Ne consegue che le decisioni umane sono largamente soggettive e talora espongono coloro ai quali le decisioni o i giudizi si riferiscono al rischio della mancanza di obiettività e persino di discriminazione. In questo quadro, proprio per bilanciare queste limitazioni e per non trasformare l'atto della decisione in un arbitrio, l'uomo giudicante ha messo a tutela dell'uomo giudicato (o dell'uomo a cui i risultati della decisione su un certo fenomeno si applicano) un insieme di valori ispiratori, ovvero di principi etici e giuridici ai quali ci si richiama nel momento in cui pur disponendo di una informazione incompleta e di una razionalità parziale si esprime comunque il giudizio o si assume una decisione. Tra questi, il principio di equità, ossia l'intento di trattare ognuno secondo i meriti o le colpe con assoluta imparzialità, la causalità, ossia la capacità di individuare un nesso inequivoco tra un'azione e la sua conseguenza, la trasparenza, ossia l'impegno a far comprendere tutti i presupposti di una decisione senza volontà di occultamento o di segretezza da parte del decisore.

D'altro canto, la macchina è un decisore pienamente razionale, che non ha interessi da difendere e il cui risultato sarà sempre lo stesso se non cambiano i dati impiegati per assumere la decisone. Le decisioni automatizzate sono dunque certamente caratterizzate dall'assenza di contraddizioni. Ma questa maggiore obiettività rispetto alle decisioni umane non è priva di rischi. C'è ad esempio un problema legato alla finalità della decisione, alla scelta dei dati in ingresso e della loro qualità, che è rilevantissimo. Sono aspetti collaterali ri-

spetto all'impiego di un algoritmo e riguardano atteggiamenti dell'uomo. Se una decisione automatizzata produce effetti discriminatori a causa del modo in cui l'uomo ha impiegato un algoritmo, o ha selezionato i dati con cui farlo funzionare non possiamo dire che l'algoritmo sia discriminatorio. Saremo, in tutti questi casi, in presenza di una decisione automatizzata razionale, quindi oggettiva, ma costruita su dati sbagliati e impiegata dall'uomo per scopi deliberatamente discriminatori. Qui la macchina è strumento non artefice della decisione. Potremmo dire che siamo ancora pienamente nella sfera delle decisioni umane, ma assistite da uno strumento tecnologico. In sostanza, non siamo di fronte a un problema concettualmente nuovo: da sempre l'uomo impiega la tecnologia anche per discriminare. Si pensi alla stampa. Essa è stata lo strumento per la diffusione della conoscenza, ma può essere impiegata per veicolare informazioni scorrette e persino calunniose nei confronti di una singola persona. Non possiamo certo dire che la stampa sia una tecnologia discriminatoria in sé, ma l'uso che se ne fa certamente può esserlo. Allo stesso modo può succedere per gli algoritmi e per le decisioni automatizzate.

Vi sono, e vi saranno, però molte situazioni in cui un algoritmo sarà interamente artefice della decisione, assunta in autonomia e senza la supervisione dell'uomo. Qui si manifesta il rischio di una decisione automatizzata discriminatoria in sé, senza che vi sia intento discriminatorio da parte dell'uomo nell'impiego dell'algoritmo. La ragione di questo rischio è determinata dal fatto che non necessariamente la razionalità della decisione o l'accuratezza delle informazioni impiegate per assumerla implicano ciò che noi chiamiamo equità, causalità e trasparenza. In altri termini, il piano dei valori ispiratori per l'uomo non ha un equivalente nella sfera delle decisioni automatizzate, e potranno ben esistere (e vedremo molti esempi in questo libro) decisioni algoritmiche non contraddittorie che però agli occhi dell'uomo appaiono non eque, non causali e non trasparenti. Equità, causalità e trasparenza sono concetti dell'uomo e sono molto difficili da formalizzare in termini "comprensibili" da una macchina. Questo dunque, in definitiva, lo scenario che l'automazione delle decisioni ci offre: passare da decisioni assunte dall'uomo su un numero tutto sommato limitato di fenomeni, caratterizzate da un livello di razionalità parziale ma ispirate a valori condivisi di tutela, a decisioni assunte da una macchina su un numero pressoché illimitato di fenomeni, caratterizzate da piena razionalità ma non sostenute valori etico-giuridici molto difficili da formalizzare in termini logico-matematici.

È un salto nel vuoto? I numeri prima richiamati ci portano a dire che questo passaggio avverrà. Quando e in che modo dipende in larga misura dalla qualità del dibattito tecnico e giuridico attualmente in corso su questi temi, nel quale si fronteggiano due punti di vista nettamente contrapposti: da una parte c'è chi sostiene che l'autonomia decisionale delle macchine sia un'invasione di campo nelle prerogative umane di formulare giudizi con assunzione di responsabilità, e che tale prospettiva vada ostacolata e fortemente regimentata, dall'altra chi invece vede nella capacità decisionale della macchina una

grande duplice opportunità, ovvero quella di poter limitare l'arbitrio delle decisioni umane, fonte di grandi discriminazioni, e di poter ottenere attraverso l'impiego di algoritmi decisioni su questioni nuove per le quali il solo ingegno dell'uomo non è sufficiente.

Qualsiasi sia il punto di vista, occorre sottrarre questo dibattito alla tentazione di uno scontro ideologico di scarsa utilità concreta, e provare a formularlo in termini oggettivi partendo proprio dal senso che attribuiamo ai valori ispiratori delle decisioni e dei giudizi umani in modo da esprimerlo secondo categorie logico-matematiche interpretabili da una macchina. Questo passaggio è preliminare per qualsiasi tipo di intervento, sia di tipo tecnico sugli algoritmi che saranno impiegati per l'assunzione di decisioni automatizzate, sia di natura normativa sulle leggi che disciplineranno questo nuovo tipo di relazione uomo-macchina.

2. Le fasi di una decisione algoritmica

In una decisione automatizzata la parte algoritmica, ossia il momento in cui i dati in ingresso a un modello matematico sono impiegati per calcolare i risultati, oggetto della decisione, è sicuramente una fase molto importante e ad essa dedicheremo la maggior parte delle riflessioni che faremo in questo libro. Tuttavia molto contano le fasi precedenti di raccolta dei dati e successive di impiego del risultato, perché spesso in queste due fasi si concentra la maggior parte dei rischi di effetti indesiderati e potenzialmente discriminatori. Sono fasi in cui l'intervento dell'uomo è decisivo. La scelta dei dati da impiegare, ad esempio, è un fattore soggettivo e può influenzare significativamente l'oggettività della decisione. La macchina nell'eseguire un algoritmo eredita, potremmo dire, scelte preliminari fatte dall'uomo che possono condizionare il risultato del modello impiegato. Il risultato di un algoritmo in sé non è discriminatorio. La macchina non disponendo di valori morali non è in grado di esprimere un giudizio su un risultato, o su una certa etichetta con cui i dati o i risultati sono classificati. Essa opera su un piano puramente formale e valuta la correttezza logico-matematica di un'inferenza senza attribuire alcuno stigma al segno o al valore di quella inferenza. Questa "passività" della macchina ha però dei risvolti negativi indesiderati. Se, infatti, l'uomo, prima di impiegare un modello matematico, considera un certo parametro come rappresentativo di un fenomeno, o di una persona, allora la macchina nell'applicare il modello cristallizzerà quell'attribuzione generando uno stereotipo dal quale potrebbe essere molto difficile tornare indietro.

Questo passaggio è cruciale: vi sono parametri che non *sono* propri di un fenomeno o di una persona, ma che sono invece loro *attribuiti* dall'uomo in ragione del contesto (ad esempio, il "buon impiegato", oppure il "bravo studente" o il "buon padre di famiglia"). La macchina non può naturalmente sovvertire questo genere di attribuzioni effettuate dall'uomo, e questa passività

può tramutare le scelte iniziali in uno stigma. La decisione della macchina riflette e replica indefinitamente le disuguaglianze e i pregiudizi introdotti dall'uomo, finendo per amplificarli. Il beneficio della rapidità nella decisione offerto dall'automazione verrebbe controbilanciato negativamente da una sorta di "effetto stereotipo" di giudizi costruiti sulla base di modelli generali formalmente corretti ma viziati da *bias* iniziali, con il risultato che ciascuna persona e ciascun comportamento verrebbero incasellati in dei veri e propri cliché. Si pensi alle conseguenze di decisioni seriali di questo tipo nell'amministrazione della giustizia, senza magari neppure un contraddittorio, o nell'erogazione di crediti, ma anche nelle assunzioni o, più in generale, nel mondo del lavoro, nella scuola e persino nella pubblicità mirata che riceviamo navigando in Internet.

Inoltre vi sono parametri che risulta estremamente complesso misurare direttamente e che si preferisce stimare molto più facilmente per via indiretta, ma esponendosi a errori. Si pensi alla situazione molto frequente in epoca di pandemia Covid della misurazione della temperatura corporea prima di ammettere una persona all'interno di un luogo chiuso. Si tratta di una misura indiretta dello stato di positività al virus di una persona, che ovviamente è approssimata e dà luogo a errori. Possono infatti ben esistere persone che pur avendo la febbre non sono positive al Covid (falsi positivi), così come persone che pur non avendo la febbre risultano invece positive e dunque contagiose (falsi negativi). La ragione per la quale ci si "accontenta" di questa stima indiretta è di natura organizzativa: si tratta infatti di un processo semplice da realizzare. La classificazione in termini di veri positivi e veri negativi al virus effettuata con la temperatura corporea è tutto sommato abbastanza ben approssimata ed è veloce e molto semplice da realizzare. La strumentazione necessaria, poi, è molto economica: basta infatti un semplice termo-scanner che fornisce istantaneamente un responso, e non servono indagini più accurate come l'applicazione di un tampone o un prelievo del sangue che richiederebbero la disponibilità di un laboratorio di analisi, specifiche competenze e tempo. Eppure, il risultato approssimato di quell'algoritmo (la misurazione della temperatura) "diventa", nel momento in cui si dispone del risultato, il parametro di decisione con cui la persona è classificata e da esso dipende la sua ammissione o meno all'interno di quel locale chiuso. È un esempio relativo a un caso specifico, ma è ben rappresentativo di cosa può succedere in una decisione affidata a un algoritmo.

Muovendoci alla fase successiva alla produzione di un risultato, ossia all'uso che ne viene fatto, anche l'accettazione passiva di una decisione algoritmica può generare effetti secondari imprevisti e indesiderati. È importante non considerare la decisione come un processo input-output puramente sequenziale, e non concentrarsi unicamente sulla sua correttezza formale. Vi sono comportamenti umani che discendono dalla decisione e che possono innescare delle controreazioni, ovvero delle azioni che influenzano le premesse stesse della decisione modificandone gli effetti. Ad esempio, se un al-

goritmo prevede che la soluzione migliore per evitare congestioni in una strada sia di dirottare il traffico verso un'altra strada, l'effetto che ne potrebbe sortire è di sovraccaricare quest'ultima senza realmente risolvere il problema. In presenza di controreazioni la decisione algoritmica potrebbe semplicemente trasformare un problema in un altro problema. Oppure, se un algoritmo prevede dalla cronologia di azioni di una certa persona che questa appartiene a un determinato insieme di altre persone assimilabili per comportamenti suggerendole di instaurare un contatto con loro, o di intraprendere le loro stesse azioni (ad esempio, acquistare lo stesso libro, visitare gli stessi posti), allora c'è il rischio che, superata la fase della novità di azioni mai prima sperimentate, quella persona possa restare ingabbiata nel cliché dei comportamenti ripetuti in quel gruppo perdendo progressivamente slancio verso nuove azioni. Questa iper-profilazione sortirebbe l'effetto opposto rispetto allo scopo dell'algoritmo di suscitare azioni e, magari, stimolare il consumo di beni o servizi.

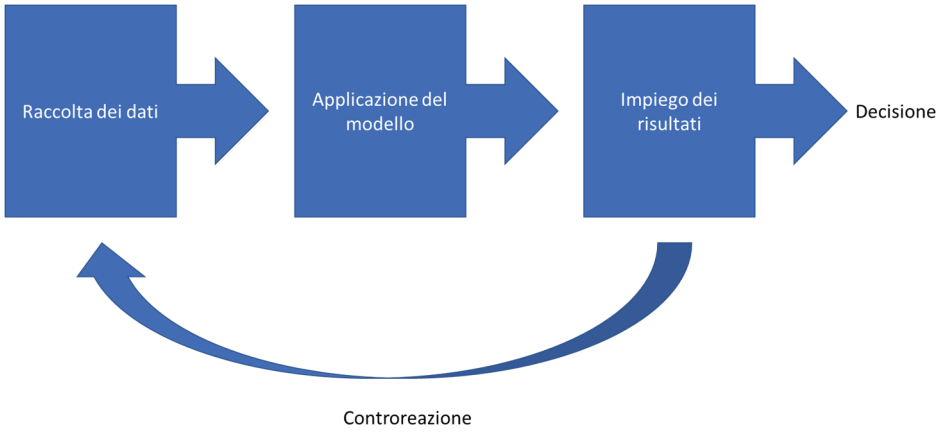

Figura 1.1: Le fasi di una decisione algoritmica

E poi c'è la fase centrale algoritmica vera e propria, di natura computazionale, che determina il rischio di risultati perfettamente razionali ma lontane dai principi decisionali dell'uomo. Vedremo nei prossimi paragrafi alcune situazioni di decisioni perfettamente razionali che tuttavia ci appaiono come paradossali. Sono esempi paradigmatici, ben noti nella letteratura, che ci mostrano il nervo scoperto di ogni forma di decisione algoritmica basata unicamente sull'analisi dei dati, ovvero la possibilità di giustificare una conclusione, ovvero un giudizio, ma anche il suo opposto esponendoci di fatto a ogni tipo di conseguenza. Per evitare queste situazioni è necessario sviluppare un certo formalismo logico-matematico.

Prima di sviluppare questo formalismo logico-matematico è importante comprendere bene le ragioni per cui esso è necessario, quale tipo di fenomeno questo formalismo dovrà rappresentare e cosa ci si aspetta dalla sua applicazione. Come detto, le decisioni algoritmiche sono il risultato dell'applicazione di modelli logico-matematici a misure relative a uno stato del mondo, o alla particolare condizione di una o più persone. Il termine misura suggerisce che i dati impiegati in questo processo siano una rappresentazione oggettiva dei fatti. Ciò di cui ci occuperemo d'ora in avanti e per tutto il resto del libro è proprio il caso della decisione automatizzata "ideale", ossia quella in cui siano state affrontate e risolte le questioni legate ai *bias* sui dati impiegati in ingresso al processo, eventualmente generati da pregiudizi dell'uomo, e le questioni legate alle controreazioni che si possono innescare dopo che la decisione sia stata assunta, e ciò che conta è la razionalità del modello impiegato per ottenere un risultato dai dati in ingresso. Entrambe le fasi, quella dei pregiudizi, preliminare alla decisone, e quella delle conseguenze e dei feedback, successiva alla decisione, non sollevano questioni concettualmente nuove, ma riguardano piuttosto aspetti di natura comportamentale dell'uomo. Essi possono essere affrontati con interventi normativi tradizionali indirizzati all'uomo, utilizzatore dei risultati della macchina, quali l'applicazione di incentivi a una condotta etica e di sanzioni nei confronti di condotte malevoli o illegali.

Nella fase centrale logico-matematica di una decisione algoritmica siamo interessati al processo che trasforma le osservazioni in misure, ossia in numeri, e al calcolo delle grandezze di uscita che costituiscono il risultato della decisione. Ciò che ci interessa comprendere è se il modello matematico in sé sia in grado di creare disparità o effetti discriminatori, e se il risultato di quel modello sia spiegabile secondo i valori umani di equità e causalità. In effetti, purtroppo, la sola fase algoritmica può dare luogo a questo tipo di inconvenienti generando situazioni di disparità. La buona notizia è che grazie all'applicazione di un preciso formalismo logico-matematico queste situazioni possono essere indentificate e i fattori che li determinano possono essere misurati in modo da comprendere quale sia il loro impatto. Ciò consente di effettuare un'interpretazione "valoriale" dei risultati di un algoritmo ed eventualmente di intervenire sui parametri dei modelli matematici impiegati in modo da mitigare potenziali effetti discriminatori. Vediamo adesso alcuni dei più comuni paradossi associati a decisioni algoritmiche, soffermandoci in particolare sui casi più frequenti e più studiati, ovvero il paradosso di Simpson, detto anche paradosso dell'inversione o del mescolamento, e il paradosso di Berkson, o paradosso della selezione. È importante partire da questi due casi, che sono dei veri e propri paradigmi, in quanto essi fanno emergere tutte le fallacie interpretative tipiche in cui si può incorrere se si mantiene un atteggiamento acritico e di accettazione passiva dei risultati prodotti da decisioni algoritmiche.

3. Decisioni algoritmiche e fallacie logiche: il paradosso di Simpson e il paradosso di Berkson

La sola analisi dei dati disponibili, effettuata senza l'ausilio di uno strumento di indagine critica dei risultati, può facilmente condurre a conclusioni contraddittorie, nonostante i ragionamenti su cui quei risultati sono stati ottenuti siano formalmente corretti e basati su presupposti logici validi. Il primo esempio che consideriamo è il paradosso di Simpson, dal nome del matematico britannico che lo formalizzò negli anni '50 del secolo scorso. Esso indica una situazione in cui una relazione tra due fenomeni appare modificata, o perfino invertita, guardando a diversi gruppi di dati a causa di altri fenomeni non presi in considerazione nell'analisi. Per comprendere la *ratio* di questo paradosso, immaginiamo il seguente caso numerico. Ipotizziamo che in una certa regione geografica siano presenti due categorie di potenziali acquirenti di un bene, tra loro ben distinte. Ad esempio, assumiamo che queste due categorie siano distinguibili per genere. Da un'indagine di mercato emerge che delle 100 donne intervistate in quella regione, 90 acquisterebbero il bene e 10 non lo acquisterebbero, mentre dei 900 uomini intervistati nella stessa regione, 720 acquisterebbero il bene e 180 non lo acquisterebbero. Quindi, in quella regione il 90% delle donne e l'80% degli uomini sarebbero interessati a quel bene. Lo stesso tipo di indagine è svolta in una regione adiacente, nella quale sono intervistate 800 donne, delle quali 160 acquisterebbero quel bene, e 200 uomini, dei quali 20 lo acquisterebbero. In pratica, il 20% delle donne e il 10% degli uomini. In entrambe le regioni è dunque rilevata una prevalenza di donne che acquisterebbero quel bene. Se però raggruppiamo le due regioni adiacenti in una macro-regione e guardiamo i dati aggregati per genere le cose cambiano. In tutto l'indagine ha riguardato 900 donne (100 della prima regione e 800 della seconda), delle quali 250 acquisterebbero il bene (90 delle prima regione e 160 della seconda), e 1100 uomini (900 della prima regione e 200 della seconda), dei quali 740 lo acquisterebbero (720 della prima regione e 20 della seconda), e se guardiamo le percentuali complessive di gradimento scopriamo che soltanto il 27,8% delle donne acquisterebbe quel bene (data dal rapporto 250/900), mentre la percentuale degli uomini acquirenti sale a 67,3% (data dal rapporto 740/1100). Come si osserva, le due indagini di mercato riferite a due sottogruppi o all'intera macro-regione danno risultati di gradimento su quel bene del tutto opposti: se consideriamo il dato disaggregato, allora siamo portati a ritenere che le donne preferiscano più degli uomini quel bene, se invece lo consideriamo in forma aggregata, allora sono gli uomini a manifestare una maggiore propensione all'acquisto. Come dunque comportarsi se uno è un venditore di quel bene per massimizzare la probabilità di successo della propria offerta? Meglio rivolgersi alle donne o meglio rivolgersi agli uomini? Entrambe le strategie commerciali sarebbero sostenute da argomenti razionali e in assenza di altri elementi non ci sarebbe modo di preferire l'una all'altra.

Questo è un esempio del paradosso di Simpson. Se il processo di scelta della strategia commerciale fosse affidato alla decisione automatizzata assunta da un algoritmo, questo non saprebbe fornire una risposta "assoluta", ma ne fornirebbe due ugualmente razionali di segno opposto, quella per regione e quella aggregata. Ci esponiamo (il nostro venditore è) dunque esposto a un arbitrio, che paradossalmente è basato su presupposti logici corretti. Come venire a capo di questa *impasse*?

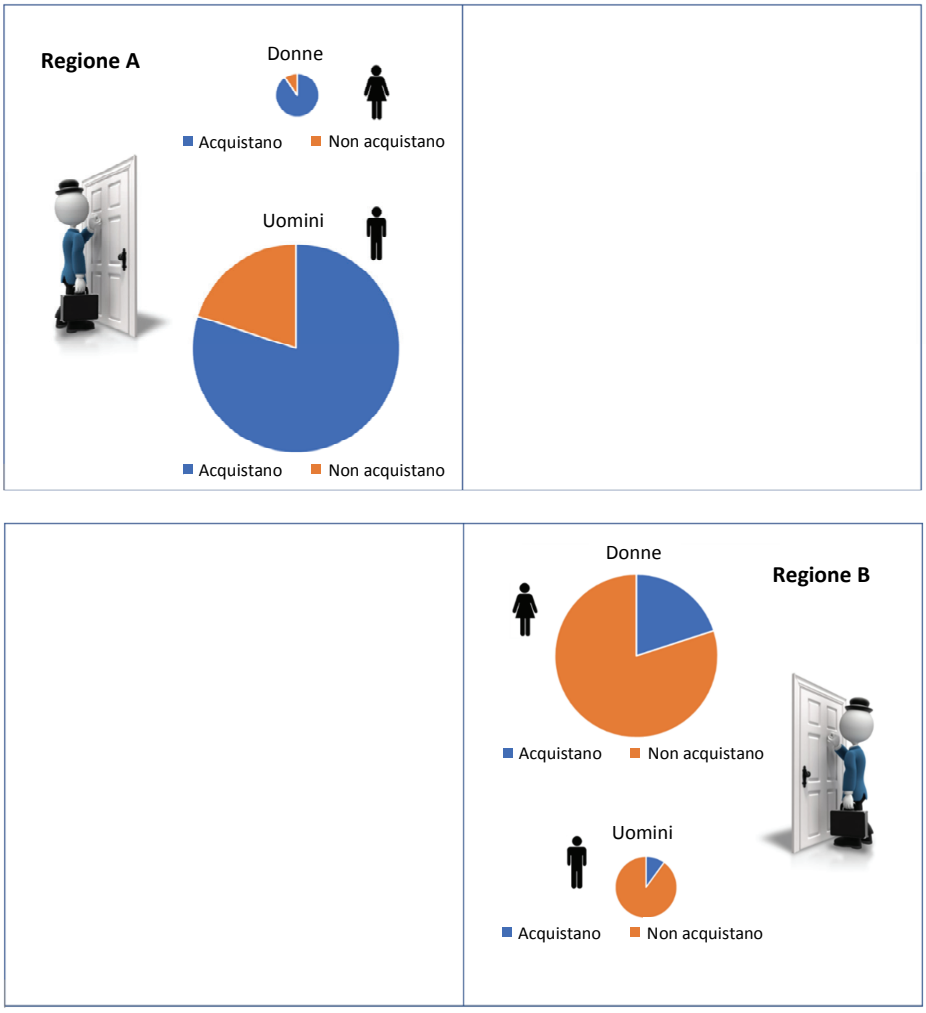

Figura 1.2: L'impiego del dato disaggregato da parte del venditore porta a porta nelle due regioni

Esiste naturalmente una via di uscita da questo paradosso. Per comprenderla muoviamoci per il momento su un piano puramente intuitivo (senza la formalizzazione logico-matematica a cui arriveremo). La prima considerazione da fare è che i due risultati, la disaggregazione per regioni e l'aggregazione per singola macro-area, si riferiscono a situazioni diverse, ovvero a due tipi di venditori di quel bene. Ad esempio, un venditore porta a porta nel momento in cui vorrà vendere il suo bene potrà trovarsi soltanto nella prima o nella seconda regione, mai in entrambe, e dunque per lui è corretto il risultato proveniente dal dato disaggregato e la strategia di vendita più efficace sarà quella di rivolgersi alle donne. La situazione è ben rappresentata nella figura 1.2. Al contrario, per un venditore stanziale che ha un negozio in un'area di confine tra le due regioni sarà la clientela a muoversi verso il negozio, e questa potrà arrivare da entrambe le regioni. Dunque ha più senso considerare il dato aggregato e sviluppare un'offerta rivolta prevalentemente agli uomini, come evidenziato nella figura 1.3. Anzi, possiamo ben dire che tra i due venditori non c'è concorrenza, in quanto essi si rivolgono primariamente a due bacini di clienti potenziali disgiunti: l'uno, ovvero il venditore porta a porta, alle donne, l'altro, il proprietario del negozio di confine, agli uomini. Si tratta di considerazioni a margine dei risultati del modello, che rendono molto più chiaro il risultato portandolo fuori dal paradosso. Esse necessitano però di conoscenze ulteriori sul "modello di business" dei venditori, che non emergono dalla semplice analisi dei dati.

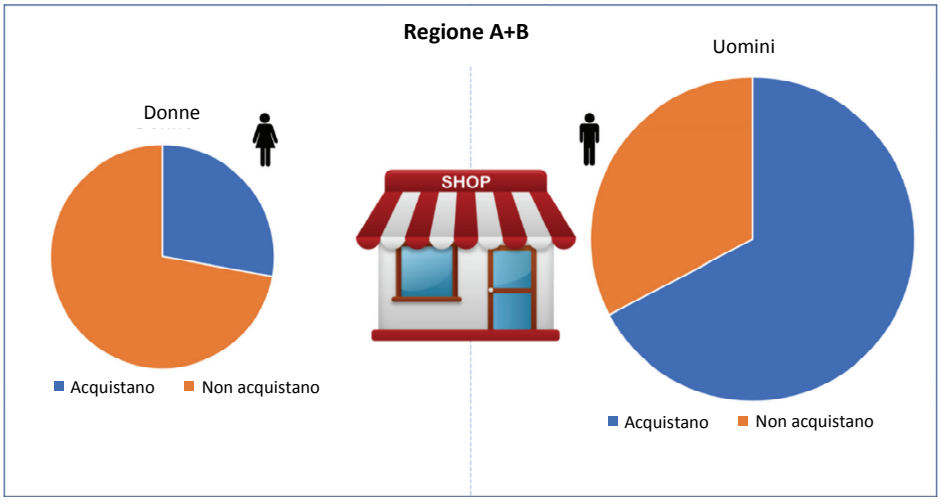

Figura 1.3: L'impiego del dato aggregato da parte negozio al confine delle due regioni

I dati, dunque, non spiegano i dati. Non ne spiegano l'origine, né l'uso che se ne farà. Ci sono "parti nascoste" che vanno al di là delle misure e di quanto osservato che devono essere "raccontate" con una vera e propria storia perché le decisioni assunte da un algoritmo non appaiano paradossali. Questa narrazione deve essere di natura tecnica, ovvero deve essere effettuata impiegando strumenti logico-matematici perché sia incorporabile negli algoritmi. In altri termini, occorre una rappresentazione formale del "non detto" dai dati in modo che l'algoritmo possa accorgersi dell'innesco di un eventuale paradosso e uscirne da solo, senza che sia l'uomo a fare questo ulteriore passo interpretativo. Una buona decisione algoritmica deve prevedere questa via di fuga dal paradosso.

Nel caso appena considerato entrambi i risultati sono plausibili, e possono essere applicati a due distinti modelli di business di un ipotetico venditore. Esistono invece situazioni in cui uno dei due risultati è privo di senso e bisogna essere più netti nella selezione. In questo caso a scegliere l'opzione sbagliata si corre il rischio di impiegare un algoritmo razionale per ottenere risultati totalmente irrazionali e inefficaci. Vediamo due esempi di questo fenomeno, uno in cui in modo inequivoco è corretto preferire l'uso di dati disaggregati, l'altro in cui invece l'unica opzione con un senso concreto è quella di fare riferimento a una statistica aggregata. Impieghiamo gli stessi valori numerici per entrambi gli esempi, in modo da mostrare come non siano soltanto i dati osservati a determinare la spiegazione di un fenomeno.

Ipotizziamo di voler determinare l'efficacia di un farmaco per la cura di una patologia. La decisione sarà assunta sulla base delle osservazioni riportate nella tabella 1.4. Nella parte di sinistra della tabella sono riportati i risultati ottenuti con un campione di 60 pazienti che non assumono il farmaco, 20 donne e 40 uomini, con le relative percentuali di successo e insuccesso (naturalmente complementari tra loro). La parte di destra riporta invece gli stessi risultati per altri 60 pazienti, 40 donne e 20 uomini, che hanno assunto il farmaco. Anche in questo caso si osserva un fenomeno di inversione tra gruppi e popolazione. Infatti, tanto nel gruppo delle donne, quanto in quello degli uomini la patologia scompare più frequentemente se non si assume il farmaco (nel 95% dei casi contro il 92,5% per le donne, e nel 70% dei casi contro il 60% per gli uomini). Se invece guardiamo l'intera popolazione senza suddivisione di genere l'assunzione del farmaco determina una maggiore efficacia nella cura di una patologia (nell'82% dei casi contro il 78%). Siamo in presenza di un paradosso e bisogna fare un passo in più di natura interpretativa rispetto all'osservazione dei dati per uscirne. In questo caso, la prima banale considerazione da fare è che nel momento in cui un paziente va dal medico per la prescrizione di una cura il suo genere, maschile o femminile, è certamente noto al medico e dunque non ci troveremo mai nella situazione della statistica aggregata in cui questo dato non è disponibile. Già per questa semplice constatazione è sicuramente preferibile la statistica disaggregata, che ci porta alla decisione di astenerci dall'uso di quel farmaco. Vi sono però altre sottigliezze in

questo campione che è bene iniziare a osservare in vista di una più rigorosa formalizzazione del problema. La prima, più evidente, è una diversa propensione all'uso di farmaci: nel campione osservato risalta il fatto che le donne assumono più farmaci degli uomini (40 donne contro 20 uomini assumono farmaci, mentre 40 uomini contro 20 donne non li assumono) e dunque non abbiamo una distribuzione uniforme delle osservazioni. In altri termini, ci sono due sottogruppi, quello delle donne che assumono farmaci e quello degli uomini che non ne assumono, che sono numericamente più rilevanti degli altri due. Possiamo dire che in questa tabella i dati non hanno tutti la stessa "forza" e che ci sono dati "forti", ossia più stabili in quanto ottenuti con un numero maggiore di osservazioni, e dati "deboli", basati su un numero inferiore di rilevazioni. Inoltre osserviamo che, indipendentemente dall'assunzione del farmaco, le donne superano più facilmente la patologia. Non abbiamo dati sulla ragione di questo fenomeno, ma potremmo ipotizzare una causa genetica che rende la guarigione delle donne più rapida, oppure potremmo immaginare che l'assunzione del farmaco determini anche effetti collaterali nocivi nei pazienti di sesso maschile che per questa ragione fanno più fatica a guarire dalla patologia. Meglio dunque disaggregare i dati in prima battuta per assumere la nostra decisione ed eventualmente formulare ipotesi più approfondite sulla presenza di fattori genetici, o sugli effetti collaterali del farmaco da incorporare nel modello decisionale.

	Non assumono il farmaco		**Assumono il farmaco**	
	Non guariscono	**Guariscono**	**Non guariscono**	**Guariscono**
Donne	*1/20 (5%)*	*19/20* ***(95%)***	*3/40 (7,5%)*	*37/40 (92,5%)*
Uomini	*12/40 (30%)*	*28/40* ***(70%)***	*8/20 (40%)*	*12/20 (60%)*
Totale	*13/60 (22%)*	*47/60 (78%)*	*11/60 (18%)*	*49/60* ***(82%)***

Tabella 1.4: Paradosso di Simpson sull'effetto di un farmaco per genere

Utilizziamo adesso gli stessi dati in un contesto diverso. Questa volta siamo sempre interessati a comprendere l'efficacia di un trattamento, ma disponiamo di osservazioni ripartite in due diversi gruppi, i pazienti a bassa e quelli ad alta pressione sanguigna alla fine del trattamento, come mostrato nella tabella 1.5. I dati numerici sono gli stessi e dunque siamo nella medesima situazione paradossale, ma le considerazioni da fare sono diverse. Per esempio, la pressione sanguigna potrebbe essere uno stato dei pazienti che si manifesta come effetto del trattamento (prima del trattamento alcuni pazienti avevano la pressione alta) e questa nuova condizione favorisce la guarigione. Ma può pure verificarsi il caso in cui il trattamento non ha effetti su un livello di pressione già basso in partenza. Inoltre, può darsi il caso di effetti tossici secondari del farmaco che ne riducono l'efficacia indipendentemente dal livello di pressione rilevato alla fine del trattamento.

Si tratta di situazioni compatibili con le osservazioni che però non emergono dalla semplice analisi dei dati e che rimangono nascoste. Non siamo in presenza di uno stato "intrinseco" per i due gruppi che si manifesta sin dall'inizio del trattamento, ma di condizioni che sono indotte dal trattamento, difficilmente quantificabili individualmente. Per tutte queste ragioni il dato disaggregato è poco rappresentativo di questa varietà di situazioni non immediatamente rilevabili ed è meglio impiegare il dato aggregato per prendere una decisione.

Su un aspetto è bene soffermarsi: nell'interpretazione che abbiamo usato per spiegare i dati c'è una sorta di "inversione logica" rispetto al precedente esempio. Nel primo esempio abbiamo osservato come lo stato dei due gruppi, ossia il loro genere, fosse intrinseco, inalterabile e noto in partenza e come questo stato condizionasse il trattamento (le donne si mostravano più propense all'uso di farmaci). Nel secondo, invece, il trattamento determina lo stato dal momento che il farmaco può condizionare il livello finale di pressione sanguigna, che può ben essere uno dei fattori della guarigione dalla patologia. Questa "inversione logica" è molto importante per la formalizzazione logico-matematica di un processo decisionale algoritmico, come vedremo nel capitolo IV dedicato alla causalità delle decisioni.

	Non assumono il farmaco		**Assumono il farmaco**	
	Non guariscono	**Guariscono**	**Non guariscono**	**Guariscono**
Alta pressione sanguigna	*1/20 (5%)*	*19/20* ***(95%)***	*3/40 (7,5%)*	*37/40 (92,5%)*
Bassa pressione sanguigna	*12/40 (30%)*	*28/40* ***(70%)***	*8/20 (40%)*	*12/20 (60%)*
Totale	*13/60 (22%)*	*47/60 (78%)*	*11/60 (18%)*	*49/60* ***(82%)***

Tabella 1.5: Paradosso di Simpson sull'effetto di un farmaco per soggetti ipertesi e a bassa pressione

A questo punto, siamo in grado di introdurre una rappresentazione formale minima di un processo decisionale algoritmico. In tutti gli esempi considerati abbiamo sempre osservato delle invarianti: una coppia di elementi mutuamente esclusivi, (vendere l'oggetto a un uomo o a una donna, assumere un farmaco o non assumerlo), che chiamiamo fattori della decisione, un certo effetto binario (il successo nella vendita o meno, la guarigione da una patologia o meno) e un terzo elemento che rappresenta uno stato in cui i fattori possono trovarsi (la localizzazione dei clienti in una certa regione nel primo esempio, oppure il genere o la pressione sanguigna negli altri due). Possiamo indicare il fattore della decisone, l'effetto e lo stato rispettivamente con le tre variabili logiche $X=\{0, 1\}$, $Y=\{0, 1\}$, $Z=\{0, 1\}$ e impiegare una tabella come quella indicata in didascalia 1.6 per rappresentare una decisione algoritmica binaria.

	X=0		***X=1***	
	Y=0	***Y=1***	***Y=0***	***Y=1***
Z=0	N_{000}	N_{010}	N_{100}	N_{110}
Z=1	N_{001}	N_{011}	N_{101}	N_{111}
Totale	$N_{000} + N_{001}$	$N_{010} + N_{011}$	$N_{100} + N_{101}$	$N_{110} + N_{111}$

Tabella 1.6: Formalizzazione di una decisione algoritmica binaria

Le quantità N_{XYZ} indicano il numero di osservazioni raccolte per le diverse combinazioni dei valori logici delle variabili *X, Y* e *Z*, considerate in questo ordine da sinistra a destra nel pedice. Si lascia al lettore l'esercizio di rappresentare gli esempi visti sinora in questa notazione.

Un altro caso di inversione riguarda la correlazione tra variabili [1]. In particolare, può verificarsi che una tendenza di un certo segno osservata in un gruppo si trasformi un una tendenza di segno opposto per l'intera popolazione, con effetti ancora una volta paradossali. La causa di questo fenomeno, come nelle manifestazioni del paradosso di Simpson che abbiamo osservato nei precedenti esempi, è la mancanza di una spiegazione sull'origine dei dati che ci consenta di comprendere se sia più corretta la rappresentazione dei dati per sottoinsiemi oppure se sia più realistico considerare i dati come misure di un fenomeno unico che riguarda tutta la popolazione. Alcuni esempi ci possono fare capire come il rischio di cattive interpretazioni dei dati, e dunque di giudizi sbagliati su un fenomeno, siano concreti.

Immaginiamo di voler sviluppare un algoritmo che sia capace di prevedere il prezzo di un appartamento noto il numero delle stanze di cui esso si compone. Ci aspetteremmo che il prezzo della casa aumenti all'aumentare del numero di stanze, eppure se osserviamo i dati raccolti attraverso un'indagine di mercato può emergere una tendenza di segno opposto. In altre parole, può ben verificarsi che in un campione concretamente rilevato la tendenza registrata sia che maggiore è il numero di stanze più basso risulta il prezzo della casa, come nel caso della figura 1.7 a sinistra in cui ogni punto rappresenta il numero di stanze (variabile in ascissa) e il prezzo di un appartamento (variabile in ordinata) normalizzati. Si tratta come è ovvio di un effetto paradossale determinato, ancora una volta, da un mescolamento acritico dei dati. Nel determinare il prezzo di un appartamento, infatti, il numero di stanze non è l'unico fattore che ne influenza il prezzo, ma altri fattori intervengono e, ad esempio, la sua posizione è importante almeno quanto il numero di stanze. Ce

[1] Sul concetto di correlazione si veda il capitolo III in G. D'ACQUISTO, *Intelligenza artificiale. Elementi*, Giappichelli, 2021, pp. 129-138.

ne accorgiamo disaggregando il dato per zona, come nella figura 1.7 a destra, in cui ogni punto è colorato diversamente se l'appartamento è in zona centrale, residenziale, periferica o extra-urbana. Se analizziamo i dati per gruppi, vediamo che le case sono più costose quanto più sono vicine al centro città. In questa disaggregazione il numero di stanze è effettivamente correlato positivamente al prezzo della casa e il risultato ci appare più intuitivo e giustificabile, e non più paradossale.

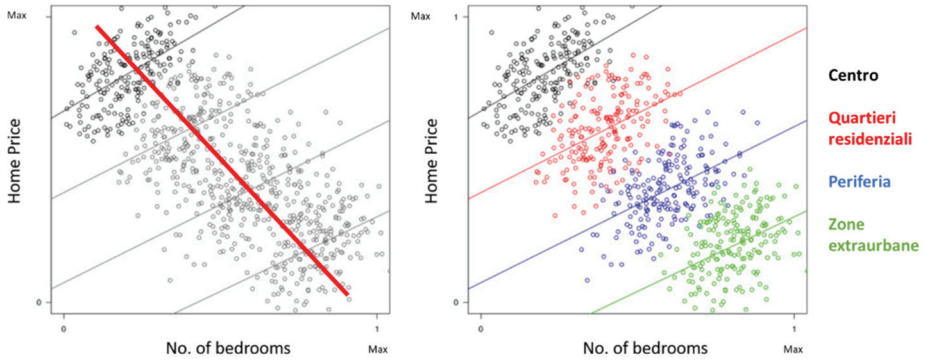

Figura 1.7: Effetto di inversione gruppi-popolazione sulla correlazione

Un'altra situazione paradossale di diversa natura, nota questa volta come paradosso di Berkson dal nome del medico e statistico americano che ne formalizzò gli effetti nei suoi studi degli anni '50, si verifica quando due grandezze che risultano indipendenti, o persino positivamente correlate in una popolazione appaiono negativamente correlate in un campione estratto da quella popolazione. Per questa ragione, il paradosso di Berkson è anche noto come paradosso della selezione.

L'errore in cui si può incorrere è quello di ipotizzare un legame tra grandezze e attribuire a un certo effetto una determinata causa, che invece è soltanto apparente. Un semplice esempio numerico ci può aiutare a comprendere l'origine di questo abbaglio interpretativo. Immaginiamo di essere interessati all'ipotetica presenza di una relazione tra il reddito e la pressione sanguigna dei residenti in una certa area geografica. In prima battuta saremmo indotti a pensare che tra queste due grandezze non esiste alcun legame statistico. Tuttavia, può succedere che l'essere ipertesi e l'avere un reddito elevato siano due caratteristiche che inducono un comportamento comune. Può infatti essere osservata una comune maggiore propensione da parte degli appartenenti ai due gruppi a visitare un medico: gli uni, ossia gli ipertesi, per ragioni legate all'esistenza di una patologia in corso, gli altri, ossia le persone con un redito elevato e che magari dispongono di una assicurazione sanitaria, per ragioni legate alla prevenzione di patologie. Questo fe-

nomeno può generare un "effetto selezione" per cui se scegliamo come campione per le nostre osservazioni le persone nella sala di attesa di un medico, può ben verificarsi che se qualcuno è al centro medico e ha un reddito relativamente basso è più probabile che abbia la pressione sanguigna relativamente alta. Al contrario, se qualcuno è lì e ha la pressione sanguigna bassa, è più probabile che abbia un reddito relativamente alto. Sulla base di questa logica diventerebbe plausibile ipotizzare una correlazione negativa tra reddito e pressione sanguigna. Viceversa, se selezioniamo in modo diverso il nostro campione, ad esempio misurando la pressione sanguigna di un campione di persone fuori da un luogo di culto o all'interno di un centro commerciale è possibile che questo "effetto selezione" non si manifesti e che le due grandezze appaiano indipendenti.

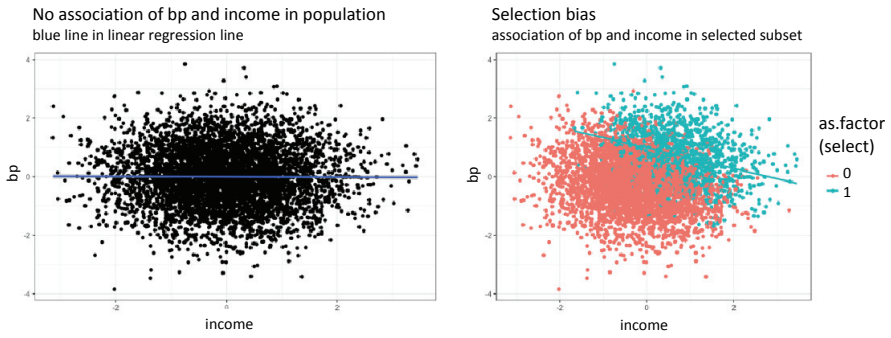

Figura 1.8: Un esempio di effetto selezione nel paradosso di Berkson

Nella figura 1.8 viene rappresentata una manifestazione di questo "effetto selezione". I punti neri a sinistra (relativi alla popolazione) mostrano che non esiste una relazione evidente tra il reddito e la pressione sanguigna. Il gruppo che è stato scelto presso la struttura medica (un sottoinsieme della popolazione originale) è mostrato in verde nel grafico a destra e mostra la presenza di una correlazione negativa tra le due caratteristiche. È un tipo di fallacia molto frequente e che può verificarsi in molti ambiti, soprattutto quelli nei quali la ricerca delle potenziali cause di un certo fenomeno osservato è l'oggetto dell'indagine o di un giudizio, come appunto il settore medico o quello economico. Le conseguenze epistemologiche di questo paradosso sono estremamente rilevanti. Vale la pena di evidenziarne almeno due. La prima è legata al criterio di scelta delle grandezze osservate. Viene da chiedersi su quali presupposti scientifici sia plausibile lo studio di una relazione tra reddito e pressione sanguigna, che intuitivamente appare piuttosto incongrua. D'altro canto, escludere a priori l'esistenza di una relazione tra grandezze misurate potrebbe essere un'ipotesi non meno azzardata. La conoscenza scientifica, in effetti, avanza

sulla base della congettura dell'esistenza di legami anche non apparenti o intuitivi, e scartarne taluni su base puramente aprioristica potrebbe costituire un atteggiamento antiscientifico ingiustificato. È un dilemma epistemologico molto serio. Se si decide di valutare l'impatto di una certa variabile su un'altra attraverso un algoritmo, bisogna essere consapevoli che si può incorrere in un effetto selezione e che questo può introdurre un *bias* interpretativo sui risultati. Il secondo è legato al valore conoscitivo di un campione di osservazioni. Il paradosso di Berkson ci mostra che se siamo in presenza di un insieme di misure (e null'altro) in linea di principio non ci è possibile capire se quelle misure si riferiscano a un campione, e dunque se si possa verificare un effetto selezione, oppure se le misure possano essere considerate rappresentative dell'intera popolazione. La decisione che ne consegue, come abbiamo visto, può essere di segno opposto e l'unico antidoto che abbiamo per evitare questo *bias* è quello di aumentare il numero di misure prima che una decisione o un giudizio siano assunti.

4. Il problema della discriminazione algoritmica

La possibilità di incorrere in paradossi decisionali, come quelli studiati da Simpson o da Berkson, non è soltanto una bizzarria logica. Assumere una decisione paradossale (ma perfettamente razionale, come abbiamo visto) può anche determinare conseguenze discriminatorie quando essa riguarda singoli o gruppi di persone.

Situazioni di discriminazione possono verificarsi tutte le volte che il fattore che determina l'esito della decisione è un attributo che "segnala" la presenza di una differenza socio-economica, di genere o di età, all'interno del campione di osservazioni che saranno utilizzate dall'algoritmo. Ad esempio, l'essere maschio o femmina, l'essere lavoratore o disoccupato, l'avere una certa nazionalità o l'essere straniero, e così via. Si tratta di esempi di variabili che possono essere il presupposto di decisioni a vantaggio degli appartenenti a una classe a discapito dell'altra. Vediamone un caso. Immaginiamo di avere raccolto le statistiche dei voti di laurea degli studenti e delle studentesse iscritte a due diverse facoltà della stessa università, come rappresentato in tabella 1.9. Come si osserva, siamo in presenza di un fenomeno di inversione: i dati a livello disaggregato mostrano che nelle due facoltà le studentesse hanno percentualmente il numero maggiore di voti alti. Se però osserviamo i dati a livello aggregato, la situazione è ribaltata e sono gli studenti maschi ad avere la maggiore percentuale di voti più alti (80% contro il 74% delle donne). Tenuto conto che il numero complessivo di studenti e studentesse dell'ateneo è lo stesso (100 studenti per entrambi i generi), ci si può domandare se questa differenza sia frutto del caso o se il dato aggregato sia indice di un atteggiamento complessivamente discriminato-

rio nei confronti delle studentesse, che sulla base di questa statistica potrebbero avere maggiori difficoltà a trovare un lavoro per effetto dei voti di laurea più bassi.

	Studenti		**Studentesse**	
	Voti bassi	**Voti alti**	**Voti bassi**	**Voti alti**
Facoltà A	*12/80 (15%)*	*68/80 (85%)*	*2/30 (7%)*	*28/30* ***(93%)***
Facoltà B	*8/20 (40%)*	*12/20 (60%)*	*24/70 (34%)*	*46/70* ***(66%)***
Totale	*20/100 (20%)*	*80/100* ***(80%)***	*26/100 (26%)*	*74/100 (74%)*

Tabella 1.9: Un caso di decisione binaria potenzialmente discriminatoria

Il dato, come si è già avuto modo di evidenziare, non spiega il dato e serve una narrazione che chiarisca se siamo in presenza di una discriminazione da correggere. Anche in questo caso esistono diverse possibili narrazioni compatibili con le osservazioni. Notiamo innanzitutto la presenza di dati "forti" e di dati "deboli" all'interno del campione: la maggior parte dei ragazzi è infatti iscritta alla facoltà A, mentre la maggior parte delle ragazze alla facoltà B, che risulta anche la più severa. L'aspetto centrale della questione, e che può eventualmente qualificare la condotta dell'università come discriminatoria, è la relazione tra il genere degli studenti e la facoltà frequentata. Si possono presentare due situazioni. Da una parte, si potrebbe sostenere la tesi che la scelta delle studentesse e degli studenti con riguardo alla facoltà da frequentare sia stata libera e senza condizionamenti. In questa ipotesi, la differenza osservata sul dato aggregato sarebbe il frutto della diversa distribuzione degli studenti e del diverso livello di severità delle due facoltà, senza dunque una precisa volontà discriminatoria dell'ateneo. Dall'altra invece si potrebbe argomentare che l'università sia attivamente intervenuta (ad esempio con borse di studio distinte per genere, o garantendo maggiori alloggi per ragazzi fuori sede anziché per le studentesse o viceversa, oppure con atteggiamenti deliberatamente discriminatori da parte dei docenti in taluni esami) impedendo che fosse raggiunta una parità nelle statistiche per genere complessivamente osservate. Le due situazioni avrebbero un diverso tipo di "narrazione" logico-matematica e potrebbero produrre differenti interpretazioni. Naturalmente ciascuna richiederebbe di essere sostenute da evidenze in un ipotetico giudizio. Torneremo sulla formalizzazione e sulle conseguenze di questa narrazione nei prossimi capitoli. Al momento è bene osservare che lo scopo di questa narrazione logico-matematica non è di risolvere il problema. Questa formalizzazione non attribuisce né scongiura la connotazione discriminatoria di una certa decisione, ma consente di porre il problema in termini oggettivi, offrendo la possibilità alle parti contrapposte in un giudizio di portare elementi misurabili a sostegno

dell'una o dell'altra tesi e di sottrarre ogni valutazione a un gioco di forza tra le parti. Questa constatazione è estremamente importante tenuto conto dell'asimmetria che spesso si osserva tra chi decide (in modo algoritmico o meno) e chi subisce le conseguenze delle decisioni. In una decisione algoritmica, grazie a una buona formalizzazione logico-matematica, questa asimmetria può essere ridotta. Questo è un punto importante a favore delle decisioni automatizzate: esse possono dare conto di ogni passaggio in modo misurabile e consentire interventi di mitigazione dei rischi. Non sempre questa stessa accountability è garantita dalle decisioni dell'uomo.

5. Verso una "narrazione" dei dati

Questa prima rassegna di esempi ci ha permesso di accorgerci che il passaggio dalla parzialità delle decisioni dell'uomo all'oggettività delle decisioni automatizzate purtroppo non scongiura l'occorrenza di errori e anche di possibili discriminazioni. La ragione di queste discriminazioni è legata alla possibilità che nel decidere attraverso l'impiego di un algoritmo non ci si accorga di essere incorsi in un paradosso. Un algoritmo infatti non trova paradossale ciò che a noi lo appare. Tutti i paradossi mostrati sono in realtà risultati pienamente razionali e solo attraverso l'intervento interpretativo dell'uomo il risultato esce dal paradosso. La macchina tratta dati numerici ed etichette senza esprimere un giudizio morale sulle misure effettuate e sulle classificazioni impiegate. Comprendere se l'impiego di un certo valore o una certa classe sia a beneficio o a danno dei soggetti a cui la decisone o il giudizio si riferiscono è un'attitudine umana: l'uomo riesce a fare questo salto concettuale sulla base delle conoscenze pregresse, dell'esperienza e di un insieme di valori etici e di costrutti giuridici che ispirano le proprie decisioni. Formalizzare questi valori in modo algoritmico non è un'operazione immediata e a questo scopo occorre una certa narrazione dei dati esprimibile con una notazione logico-matematica.

Una prima indicazione, anche progettuale, che si può trarre da questa disamina generale che abbiamo svolto è che la disomogeneità delle misure è un fattore che senz'altro può innescare paradossi. In tutti i casi che abbiamo analizzato erano presenti nella stessa rilevazione statistica sia dati "forti" (ossia campioni costituiti da molte osservazioni) che dati "deboli" (nei quali il numero di osservazioni era assai inferiore). Se tutti i dati fossero ugualmente forti non avremmo paradossi. Ce ne accorgiamo facilmente guardando la struttura che abbiamo introdotto per le decisioni binarie e la stessa notazione numerica della tabella 1.6.

Formalmente, come si può facilmente constatare, si ha un'inversione di risultati tra popolazione e sottogruppi se sono vere le seguenti disuguaglianze:

$$\frac{N_{110}}{N_{100}+N_{110}} > \frac{N_{010}}{N_{000}+N_{010}}$$
$$\frac{N_{111}}{N_{101}+N_{111}} > \frac{N_{011}}{N_{001}+N_{011}}$$
$$\frac{N_{110}+N_{111}}{N_{100}+N_{101}+N_{110}+N_{111}} < \frac{N_{010}+N_{011}}{N_{000}+N_{001}+N_{010}+N_{011}}$$

Se i sottogruppi di osservazioni fossero tutti della stessa ampiezza, ovvero se risultasse

$$N_{000}+N_{010}=N_{001}+N_{011}=N_{100}+N_{110}=N_{101}+N_{111}$$

questa situazione non potrebbe mai verificarsi, dal momento che non esistono due numeri più piccoli di altri due numeri che sommati tra loro danno come risultato un numero maggiore della somma dei due numeri più grandi. In altri termini, con semplici passaggi algebrici dalle disuguaglianze, non potrà mai accadere che

$$N_{110} > N_{010}$$
$$N_{111} > N_{101}$$
$$N_{110}+N_{111} < N_{010}+N_{011}$$

Disporre di rilevazioni basate su campioni della medesima ampiezza è dunque certamente una buona indicazione per chi progetta algoritmi. Tuttavia, con spirito pragmatico, bisogna anche constatare che i dati impiegati nelle decisioni algoritmiche possono provenire da fonti diverse e dobbiamo considerare come verosimile l'ipotesi che essi non abbiano sempre la stessa forza. Dunque serve un impianto concettuale che funzioni anche in presenza di dati "deboli" e dati "forti".

Una fonte di errori interpretativi di un risultato, come visto, è anche la presenza di effetti di selezione non ben identificati tra grandezze indipendenti, le quali possono essere confuse per causa ed effetto di un certo fenomeno. Nell'esempio dello studio su redditi e pressione sanguigna abbiamo osservato come vi fosse un effetto di selezione (il fatto che la rilevazione della pressione fosse effettuata in un centro medico) che creava una relazione apparente tra le due grandezze. Non sempre siamo nelle condizioni di accorgerci con prontezza di questi effetti selezione ed è buona norma, dove possibile, aumentare l'ampiezza del campione e variare il contesto delle misure effettuate.

Tutti questi effetti paradossali, e le eventuali conseguenze discriminatorie che da essi potrebbero discendere, possono essere evitati disponendo di una opportuna rappresentazione che spieghi l'origine dei dati. Questa narrazione logico-matematica dei dati è essenziale per rendere più spedita questa fase interpretativa, in modo da essere effettuata con la velocità richiesta dalla quanti-

tà di dati, in continua crescita, e per un numero di decisioni molto più ampio di quanto consentito all'uomo. Essa consente alla macchina di indagare "con occhio umano" i propri stessi risultati riconducendoli ai valori di equità e causalità adottati dall'uomo nelle proprie decisioni. Il punto di partenza per sviluppare questa narrazione è la conoscenza di specifici strumenti logico-matematici desunti dal calcolo della probabilità e dalla teoria dei grafi, che introduciamo nel prossimo capitolo.