

1. Spesso in passato mi è capitato di deplorare il fatto che il contributo apportato dai grandi scienziati e tecnici al progresso dell'umanità non abbia il risalto che merita, e che nelle nostre scuole la storia che si insegna sia quasi esclusivamente la storia politica, cioè quella che ricorda guerre, conquiste, trattati di pace, spartizioni ecc. Ben raramente vien fatta menzione dei grandi del pensiero filosofico e scientifico, personaggi che prepararono l'ambiente per i cambiamenti politici. Recentemente la nostra scuola ha recepito anche l'utilità di inquadrare storicamente la scienza e le sue scoperte: si ottiene così, tra l'altro, il risultato di dare un significato umano alla scienza, di render conto della passione di scoperta e di ricerca che ha ispirato i grandi, ma che ha anche diretto l'impegno dei meno grandi. Purtroppo in questo parziale risarcimento di interesse verso gli scienziati da parte del pubblico esiste una specie di gerarchia che porta quasi necessariamente in primo piano le persone che hanno fatto scoperte più direttamente collegate con le utilizzazioni. Così, per esempio, in Italia Marconi è certo più conosciuto di Maxwell; ma noi continuiamo a pensare che l'opera scientifica di Maxwell sia stata molto più vasta e profonda di quella dello scienziato italiano. Tuttavia quest'ultima è universalmente ricordata perché sta alla base di certe applicazioni che utilizziamo quotidianamente: radio, televisione ecc. Analogamente, nel campo della medicina, la cronaca quotidiana ricorda con grandi clamori i chirurghi che compiono ardite operazioni di trapianto; ma dimentica di dire che queste porterebbero a morte sicura se non vi fossero le lunghe, pazienti ed intelligentissime ricerche sulla compatibilità e sui meccanismi cellulari, enzimatici ed ormonali che portano l'organismo ad accettare oppure a rifiutare certi elementi estranei trapiantati. Senza questi studi pazienti e profondi l'opera del chirurgo, celebrato e portato sugli altari, porterebbe il paziente a morte sicura, a brevissima scadenza; come avvenne nei secoli scorsi per le trasfusioni di sangue e per i trapianti che furono tentati. Per ritornare all'ambito della fisica e della tecnica, vorrei ricordare che le scoperte di molti fisici non avrebbero potuto trovare mezzi di espressione né possibilità di sviluppo teorico e pratico in assenza di un pensiero matematico che le ha precedute e che ha fornito i mezzi per la formulazione delle ipotesi e per gli sviluppi teorici successivi.

2. Abbiamo detto poco fa che la Statistica si occupa di raccogliere informazioni su grandissimi numeri di individui analoghi, e di elaborarle per la loro ulteriore utilizzazione. Vorremmo ricordare inoltre che la Statistica può essere utilizzata non soltanto per la descrizione di fenomeni collettivi, ma anche per una operazione scientifica che viene spesso chiamata "inferenza". Con questa procedura la Statistica cerca di scoprire le qualità di una grande folla di enti a partire dalle osservazioni eseguite su pochi elementi della folla. Questi enti molto numerosi che si vogliono conoscere possono esistere contemporaneamente, oppure possono esistere in tempi diversi; in questo secondo caso ci troviamo di fronte al tentativo di costruire delle leggi che vengono chiamate "induttive", leggi con le quali si cerca di predire i fenomeni futuri, non conosciuti nella loro essenza, in base alle osservazioni di molti casi avvenuti nel passato. In generale si potrebbe dire che la procedura di inferenza statistica analizza un sottoinsieme dell'insieme numeroso di enti che si vogliono conoscere; tale insieme viene spesso chiamato "universo", ed il sottoinsieme che si analizza viene spesso chiamato "campione" dell'universo stesso.

A questo proposito ricordiamo qui che, nelle applicazioni pratiche delle procedure statistiche, vengono spesso impiegati dei termini come "molti", "numerosi", oppure correlativamente "pochi", e simili. Ora occorre osservare che questi termini, come pure le espressioni nelle quali si fa menzione di "numeri grandi", non hanno un significato matematico preciso e rigoroso. Pertanto la loro validità si esplica soltanto in relazione a determinate situazioni particolari, e spesso il loro significato ha un valore prevalentemente psicologico.

Analoghe considerazioni possono essere fatte a proposito di quella che i cultori di Calcolo delle Probabilità chiamano spesso "Legge empirica dei grandi numeri". Anche negli enunciati di questa legge empirica si fa riferimento a numeri che vengono qualificati come "grandi", sebbene questa qualifica non abbia alcun senso matematico preciso. Ciò non significa tuttavia che le previsioni ed i comportamenti fondati sulla legge empirica dei grandi numeri e sulle osservazioni statistiche siano inattendibili e totalmente privi di significato; tuttavia tali previsioni non hanno lo stesso grado di certezza delle deduzioni logiche e dei calcoli matematici. In sintesi si potrebbe dire che le previsioni fondate sulla legge empirica dei grandi numeri e sulle osservazioni statistiche hanno il loro fondamento su una nostra presunzione la quale, qualora sia stata rilevata una certa regolarità in un numero grande di casi, ci porta a pensare che la Natura, oppure l'uomo, si comportino sempre in modo tale da non provocare sensibili differenze dal comportamento osservato. Per esempio, quando sia stata rilevata da molti decenni che il tasso di nascite maschili, nelle nostre società ed alle nostre latitudini, è di circa il 51 per cento del totale delle nascite, in mancanza della conoscenza precisa delle leggi biologiche che regolano la determinazione del sesso nei neonati, appare ragionevole stringere un contratto aleatorio, o prendere altre decisioni economiche nelle quali la probabilità che un neonato futuro sia di sesso maschile sia valutata in 0,51. Questa decisione è basata su una certa presunzione di costanza delle leggi della Natura; essa sarà da considerarsi ragionevole fino a

che non sia possibile diagnosticare in tempo il sesso di un nascituro molto prima della nascita, o addirittura influire sui meccanismi biologici che determinano il sesso.

Potremmo quindi concludere che, nell'esempio ora presentato, ed in altri numerosissimi che si potrebbero citare, l'estensione della validità delle osservazioni fatte su pochi casi ad altri numerosissimi, avviene presumendo che nel futuro siano valide le modalità con le quali certi fenomeni si sono presentati nel passato. Tuttavia la estensione della validità potrebbe essere fatta anche in maniera diversa, dando luogo ad altri casi di induzione, con diverse modalità e diversi atteggiamenti.

È questo il caso che si presenta nell'impiego della Statistica nei cosiddetti sondaggi di opinione, o nei controlli di qualità. In questi casi si tratta di estendere, nel modo più ragionevole e più efficace possibile, le conoscenze che si hanno su un sottoinsieme dell'universo all'universo intero. Ciò si fa per esempio quando, nei sondaggi di opinione, si interrogano dei cittadini scelti in modo opportuno, ma formanti una minoranza della popolazione, per poter avere delle informazioni attendibili sui comportamenti di un sottoinsieme più vasto della popolazione, o addirittura dell'intera popolazione. Per esempio, quando si fanno dei sondaggi elettorali, per poter predire i risultati delle elezioni; oppure si fanno dei sondaggi commerciali, per poter predire il comportamento dei consumatori.

Analoghe considerazioni si possono fare a proposito dei cosiddetti controlli di qualità: in questi casi si considera per esempio una industria che fabbrica molti pezzi in serie di un medesimo prodotto. Poiché non tutti i pezzi prodotti sono della medesima qualità, per evitare i danni che potrebbero insorgere dalla immissione sul mercato di troppi pezzi difettosi, si fanno dei controlli su campioni, per poter valutare la qualità presunta dei difetti, ed eventualmente per poter provvedere alla correzione delle lavorazioni ed alla riparazione delle macchine.

Questi esempi possono servire per dare un'idea delle moltissime ed utilissime applicazioni della Statistica; si potrebbe dire addirittura che oggi questa scienza è diventata una componente necessaria della procedura scientifica; per esempio delle procedure di ricerca della Biologia, nella quale non è possibile trattare sempre gli esseri viventi come degli apparecchi di laboratorio, nella Medicina, nelle Scienze sociali ecc. In queste scienze, ed in altre numerosissime, la Statistica fornisce il supplemento di tecnica conoscitiva che è necessario per aiutare la nostra insufficiente conoscenza delle cause dei fenomeni e quindi dell'essenza degli enti che vogliamo conoscere.

3. Esporremo qui di seguito alcuni criteri metodologici generali che si seguono per la raccolta delle informazioni, raccolta che è il primo scopo della Statistica.

Sia U un insieme di elementi che per qualche buona ragione possiamo considerare come analoghi; indichiamo con N il numero degli elementi dell'insieme U , che d'ora innanzi sarà chiamato convenzionalmente "universo" o anche "popolazione", senza che a quest'ultimo termine sia dato necessariamente il significato che esso ha nel linguaggio comune: così per esempio potremmo parlare della popolazione delle mosche, oppure di quella degli atomi di una data sostanza, contenuti in un dato recipiente. Siano anche dati certi insiemi, in numero di n :

$$(1) \quad U_1, U_2, \dots, U_n$$

che sono sottoinsiemi di U . Diremo che i sottoinsiemi (1) costituiscono una "partizione" dell'insieme U se sono verificate le seguenti condizioni:

a) ogni insieme (1) contiene almeno un elemento dell'insieme universo U ; b) ogni elemento dell'universo U appartiene ad un sottoinsieme (1); c) due insiemi (1) quali si vogliono non hanno alcun elemento in comune.

Indichiamo con

$$(2) \quad x_1, x_2, \dots, x_n$$

le cardinalità degli insiemi (1). Come conseguenza delle condizioni ora enunciate si avrà:

$$(3) \quad x_1 + x_2 + \dots + x_n = N.$$

In conseguenza della partizione dell'universo U in sottoinsiemi (1) si avrà una corrispondenza:

$$(4) \quad \begin{array}{c} U_1, U_2, \dots, U_n \\ x_1, x_2, \dots, x_n \end{array}$$

tra i sottoinsiemi della partizione (1) e le loro cardinalità. In forza della condizione a) enunciata sopra, i numeri x_i sono tutti interi e positivi.

Spesso, invece di far corrispondere ad ogni sottoinsieme U_i della partizione (1) la sua cardinalità, si associa al sottoinsieme U_i il rapporto:

$$(5) \quad k_i = \frac{x_i}{N}$$

tra il numero degli elementi di un sottoinsieme ed il numero globale degli elementi dell'universo U . In questo caso, invece della tabella (4) si ottiene una tabella del tipo:

$$(4)\text{bis} \quad \begin{array}{c} U_1, U_2, \dots, U_n \\ k_1, k_2, \dots, k_n \end{array}$$

In forza della definizione (5) i numeri k_i soddisfano alla condizione:

$$(6) \quad k_1 + k_2 + \dots + k_n = 1.$$

Come è noto, spesso, invece dei numeri (5), vengono presentati i loro multipli secondo 100, multipli che vengono chiamati "percentuali". Pertanto, ponendo:

$$(7) \quad h_i = 100 k_i$$

si può costruire una tabella analoga alla (4)bis; abitualmente, in questo caso, ogni numero h_i viene fatto seguire dal simbolo "%". In questo caso, invece della (6), vale ovviamente la

$$(8) \quad h_1 + h_2 + \dots + h_n = 100.$$

Osserviamo ora che la partizione dell'universo U in sottoinsiemi può avvenire in vari modi: infatti uno dei primi punti che si debbono stabilire quando si raccolgono le informazioni è la precisazione dei fini a cui sono dirette le informazioni stesse, e l'insieme delle elaborazioni alle quali le informazioni debbono essere sottoposte. Tale precisazione viene fatta abitualmente stabilendo i criteri secondo i quali un elemento dell'insieme universo U viene attribuito ad uno dei sottoinsiemi della partizione (1).

4. Abbiamo detto ripetutamente che uno degli scopi della Statistica è quello della elaborazione delle informazioni raccolte su gruppi numerosi di individui; tale elaborazione viene fatta di solito con la costruzione di certi numeri che rappresentano l'universo, in relazione a certe informazioni che si vogliono avere, ed in vista di determinati fini. Tali numeri vengono chiamati *medie* o anche *numeri indici*, a seconda delle circostanze e degli scopi che si vogliono raggiungere con le informazioni che vengono elaborate e che si vogliono trasmettere. Come è ovvio, la elaborazione delle informazioni che si sono ottenute dipende dalla partizione dell'universo U , e dai caratteri che sono stati scelti per costruire tale partizione. Pertanto, in particolare, la elaborazione delle informazioni è diversa a seconda del fatto che la partizione dell'universo sia stata fatta in base a caratteri qualitativi oppure in base a caratteri quantitativi. È chiaro che questi ultimi si prestano meglio dei primi ed una ulteriore elaborazione, eseguita con i mezzi e gli strumenti della Matematica. Nel presente paragrafo daremo qualche nozione a proposito delle serie statistiche, riservando i prossimi paragrafi alle questioni riguardanti le seriazioni.

Quando sia data una serie statistica, cioè, ripetiamo, quando sia data una partizione dell'universo U considerato in base ad un carattere qualitativo, una informazione importante è fornita dalla numerosità del sottoinsieme più numeroso; tale numerosità viene chiamata *media modale* o anche semplicemente *moda* della serie. L'impiego del termine "moda" è coerente con quello che viene fatto nel linguaggio comune nel quale si dice che qualche cosa (un colore, un atteggiamento e simili) è "di moda" quando il gruppo di coloro che li adottano è più numeroso di ogni altro gruppo, singolarmente preso.

A titolo di esempio, si consideri la serie seguente che riguarda le spese dei cittadini italiani in divertimenti nell'anno 1981: si tratta di una serie nella quale le spese degli italiani per divertirsi durante il 1981 sono ripartite in 4 sottoinsiemi. I valori corrispondenti sono dati in percentuali:

Teatro	10.1
Cinematografo	38.2
Trattenimenti vari	35.4
Manifestazioni sportive	16.3

Quindi, secondo la nomenclatura adottata, si può dire che nel 1981 la media modale, o moda, delle spese degli italiani per divertimenti è stata la spesa per il cinematografo.

Può avvenire che sia possibile introdurre in una serie statistica un ordinamento; in questo caso è possibile prendere in considerazione anche un altro elemento che descrive in qualche modo la serie: si tratta del carattere che, rispetto a quell'ordinamento, divide la serie in due parti approssimativamente uguali tra loro; tale carattere viene chiamato *mediano*, ed il valore corrispondente viene detto *mediana* della serie. Ovviamente, nel caso di una serie statistica, la introduzione di un ordinamento può essere materia notevolmente opinabile; rimandiamo la ulteriore trattazione di questi concetti al caso delle seriazioni statistiche, cioè al caso in cui la partizione di un insieme in sottoinsiemi è fondata su un carattere quantitativo.

5. Consideriamo dunque una seriazione statistica, cioè un universo nel quale l'attribuzione di un determinato elemento ad un sottoinsieme della partizione (1) del paragrafo 3 viene fatta in base ad un carattere quantitativo. Per fissare le idee, e per maggiore chiarezza ed immediatezza di presentazione, consideriamo un esempio concreto tratto da esperimenti effettivamente eseguiti.

In un allevamento di animali da laboratorio, vennero osservate 815 nidiate di ratti; le nidiate furono classificate a seconda del numero di topolini che le componevano. Si ottenne così il seguente insieme di

osservazioni:

Numero dei componenti la nidiate	Numero delle nidiate che hanno quei componenti	Frequenze relative
1	7	1
2	33	4
3	58	7
4	116	14
5	125	15
6	126	15
7	121	15
8	107	13
9	56	7
10	37	5
11	25	3
12	4	1
	Totale delle nidiate = 815	

I numeri della terza colonna sono stati ottenuti dividendo i numeri della seconda per il totale delle osservazioni fatte (815) ed arrotondando i risultati. Volendo inquadrare le osservazioni qui riportate nella cornice della teoria esposta nei paragrafi precedenti, si potrebbe dire che l'universo U delle nidiate di ratti, composto da 815 unità, è stato ripartito in 12 sottoinsiemi, ed il criterio della attribuzione di una nidiate ad un determinato sottoinsieme è quello, quantitativo, del numero dei topolini componenti la nidiate.

Si pone ora il problema di descrivere la seriazione con certi numero parametri, in modo da sintetizzare le informazioni che si hanno per utilizzarle ai fini della conoscenza dei fenomeni che ci interessano, e per ulteriori decisioni. Ovviamente, così facendo si perdono delle informazioni; ma quelle che si ottengono dalla elaborazione dei dati sono quasi sempre più utili al ricercatore per i fini che gli interessano. I parametri che si costruiscono che descrivere una seriazione sono molto numerosi e, ripetiamo, dipendono dagli scopi della ricerca teorica e delle applicazioni; presenteremo qui soltanto quelli più comunemente usati per la descrizione dei fenomeni collettivi. Uno dei parametri di cui parliamo è già stato presentato nel paragrafo precedente: è la "media modale" o "moda". Essa fornisce il carattere corrispondente al sottoinsieme o ai sottoinsiemi a cui compete la massima numerosità; una semplice ispezione della tabella conduce a concludere che la moda in questo caso è 6.

Osserviamo di passaggio che la seriazione che stiamo esaminando presenta un caso abbastanza semplice, cioè il caso della esistenza di un solo carattere che può essere chiamato "moda"; si vuol dire che questa seriazione è *unimodale*. Ma si presentano anche dei casi più complicati, di seriazioni che vengono chiamate *plurimodali*, perché diversi sottoinsiemi della partizione hanno numerosità maggiore di quelli che sono vicini.

Ricordiamo anche, con riferimento alla valutazione di probabilità che può essere dedotta dalle osservazioni, che la seriazione viene spesso chiamata *distribuzione statistica*; e questo modo di esprimersi sarà talvolta adottato anche da noi nel seguito.

Un secondo parametro che serve per descrivere la distribuzione considerata è la *mediana*; questo numero individua il carattere numerico che ha la seguente proprietà: il numero degli elementi dei sottoinsiemi che corrispondono a valori minori del carattere stesso è uguale al numero dei componenti dei sottoinsiemi che hanno carattere numerico superiore.

Così, in pratica, nell'esempio considerato sopra, si ha che la metà del numero delle nidiate è 497.5. Analizzando la tabella, si ha che le nidiate con 5 componenti o meno sono 339, mentre quelle con 6 componenti o più sono 476. Quindi la mediana della seriazione considerata sta tra 5 e 6. Il concetto di mediana può essere ovviamente generalizzato; infatti, si possono considerare quei valori del carattere numerico (che determina la partizione) che suddividono l'universo in parti aventi un determinato rapporto con il totale. In questo ordine di idee sono usati spesso i *quartili*; sono questi dei valori del carattere quantitativo che dividono i componenti dell'universo in quattro parti uguali. Così, nella seriazione che stiamo esaminando, si ha che il quartile inferiore è situato tra 3 e 4, ed il quartile superiore tra 7 ed 8.

Infine uno dei parametri che viene più frequentemente impiegato per descrivere la seriazione è la *media aritmetica*. Questo numero si ottiene semplicemente dividendo la somma delle cardinalità dei componenti i singoli sottoinsiemi della partizione, per il numero N dei componenti dell'universo U . Così nel caso della

distribuzione considerata si deve calcolare il numero:

$$7 \cdot 1 + 33 \cdot 2 + 58 \cdot 3 + \dots + 12 \cdot 4 = 4992,$$

e dividerlo per 815: si ottiene così $\frac{4992}{815} = 6,125$ come *valore medio*, o *media* della distribuzione.

Possiamo osservare che, nelle elaborazioni precedenti, abbiamo ottenuto dei numeri non interi; questa circostanza ci offre il destro di aprire una breve parentesi dedicata alla precisazione del significato e della natura delle informazioni che si ottengono utilizzando gli strumenti matematici nella descrizione della realtà.

Nel caso esaminato i numeri impiegati hanno significato soltanto se sono numeri naturali, perché danno delle informazioni sul numero dei componenti di determinati insiemi, costituiti da elementi tutti distinti tra loro. Se ci limitiamo ad attribuire ai numeri ottenuti soltanto questi significati, appare chiaro che i discorsi fatti in precedenza non hanno sempre un significato preciso. Tuttavia si può pensare che i numeri ottenuti abbiano un significato se sono interpretati come delle valutazioni, o delle informazioni che servono a valutazioni di probabilità.

6. Le considerazioni che abbiamo svolto nei paragrafi precedenti possono essere ulteriormente generalizzate, con riferimento a vari problemi di Matematica applicata; tra questi vorremmo ricordare il problema di sintetizzare le informazioni che vengono fornite da molti numeri, che possono avere diversi significati: per esempio misure di grandezze, o conteggi degli elementi di vari insiemi o altri significati ancora. In molti casi risulta particolarmente scomodo il dare l'elenco di tutti i numeri che si sono ottenuti, elenco che potrebbe anche essere poco significativo per ulteriori applicazioni; si suole allora rappresentare l'insieme numerico con determinati numeri che vengono spesso chiamati *medie*. Per esempio si sente spesso parlare di "reddito medio degli italiani", oppure, considerando un gruppo di persone, si parla di "età media"; oppure di "altezza media" di un gruppo di coscritti ecc. ecc.

Questi concetti potrebbero essere presentati in modo generale con le considerazioni che ora svolgeremo. Si abbia un insieme di numeri:

$$(1) x_1, x_2, \dots, x_n$$

e si voglia calcolare un determinato numero, funzione di questi:

$$(2) z = f(x_1, x_2, \dots, x_n).$$

Si chiama "media" dei numeri (1), con riferimento alla funzione f , un numero M il quale, sostituito nella (2) al posto di ciascuno dei numeri (1), dia alla funzione lo stesso valore; in altre parole un numero M tale che sia

$$(3) z = f(M, M, \dots, M).$$

Per esempio, si considerino i redditi di tutti i cittadini italiani, e si voglia eseguire la somma di tutti questi redditi, somma che, in questo caso, è la funzione f che consideriamo; allora la stessa somma può essere ottenuta con addendi M tutti uguali tra loro, purché si ponga:

$$(4) M = (x_1 + x_2 + \dots + x_n) / n.$$

Il numero definito dalla (4) viene chiamato *media aritmetica* dei numeri (1); esso viene spesso chiamato semplicemente "la media" dei numeri stessi; ma si possono prendere in considerazione anche altri numeri, costruiti a partire dai numeri (1), che sono atti a rappresentare i numeri stessi per certi scopi, che debbono essere precisati e specificati di volta in volta, e che determinano la scelta delle operazioni che si eseguono per costruire la media. Ovviamente, ripetiamo, quando ai numeri (1) si sostituisce un unico numero, opportunamente calcolato, si perdono delle informazioni; ma si ottiene in compenso il vantaggio di avere una informazione sintetica, la quale è spesso utile ai fini dei calcoli ulteriori o delle decisioni che si debbono prendere. Tuttavia, ripetiamo ancora, il calcolo della media deve essere determinato dagli scopi a cui il numero che si ottiene deve essere diretto.

Si consideri a questo proposito il seguente esempio elementare: siano dati due numeri positivi a , b , che vengono interpretati come le lunghezze di due segmenti; ovviamente le informazioni che si forniscono dalla elencazione dei due valori a e b possono essere sintetizzate in una sola, fornita con un unico numero, che si può chiamare la "media" dei due; ma il calcolo di quest'unico numero è determinato dall'uso che si vuole fare delle informazioni ottenute. Supponiamo per esempio che si voglia ottenere un unico numero il quale, raddoppiato, dia la lunghezza del semiperimetro del rettangolo che ha i lati di lunghezze a e b . In questo caso la "media" che fornisce questa informazione è: $M = (a + b) / 2$.

Ma supponiamo invece che si voglia la lunghezza di un unico segmento che è lato di un quadrato il quale ha la stessa area del rettangolo di lati a e b ; in questo caso la media da calcolare è quella che viene chiamata *geometrica* e che è data dalla formula: $G = \sqrt{ab}$.

Ciò che abbiamo detto fin qui ci pare una adeguata giustificazione per le considerazioni che abbiamo svolto a proposito delle informazioni che si possono trarre dalla Statistica per determinare le valutazioni di probabilità di eventi sui quali non si posseggono informazioni complete.

Il fatto che molto spesso nell'Economia o nelle scienze sociali o anche nelle scienze della Natura non si

possano avere delle formulazioni certe e sicure delle leggi che regolano i fenomeni, non impedisce che si possano utilizzare gli strumenti della Matematica per rendere minima possibile la nostra ignoranza e per prendere delle decisioni che hanno la maggior razionalità possibile, compatibilmente con le informazioni che si posseggono.

NdR

Testo reimpaginato nel gennaio 2014.

Si tratta presumibilmente di una conferenza tenuta nei primi anni '80, certamente dopo il 1982.