L'elaborazione statistica dei dati

Roberto Parroni

0. Introduzione

La statistica è la disciplina che studia i fenomeni collettivi allo scopo di metterne in evidenza le regolarità. Il vocabolo *Statistica* deriva dal latino «*status*» perché inizialmente questa scienza si occupava esclusivamente degli avvenimenti dello Stato. La parola *Statistica* viene usata sia al singolare che al plurale.

Usata al singolare sta a significare l'insieme dei metodi e delle teorie che permettono di studiare i fenomeni collettivi, mentre usata al plurale sta ad indicare un insieme di dati numerici relativi a gruppi di persone o fatti in senso lato.

In tutti i problemi di statistica ci si trova di fronte ad un insieme di dati che sono stati raccolti in vista di determinati scopi. Poiché questi possono presentare diversi aspetti, oggi, nella statistica si presentano due rami principali: la ***Statistica descrittiva*** e la ***Statistica inferenziale***.

La statistica descrittiva si ha quando lo studio del fenomeno collettivo si fa osservando interamente la collettività degli individui (cioè l'intera popolazione). Storicamente questa parte corrisponde alla prima fase degli studi statistici, quando lo scopo era proprio quello di raccogliere (e commentare) certi dati relativi allo stato.

La statistica inferenziale si ha quando lo studio del fenomeno collettivo si fa osservando soltanto una sua parte (detta *campione casuale*). Ad esempio, per sapere qual è il numero dei pezzi difettosi prodotti da una certa macchina non è comodo né economico esaminare l'intera produzione. Allora dall'intera produzione si estrae un campione scelto secondo opportuni criteri. La percentuale dei pezzi difettosi presenti nel campione dà un'idea di quella che, probabilmente, è la percentuale dei pezzi difettosi dell'intera produzione. Così anche, se chiedessimo a 1000 persone per chi voteranno il giorno prima delle votazioni, avremmo la possibilità di prevedere quale partito potrebbe vincere.

Questi metodi statistici, esposti qui molto sinteticamente, vengono poi applicati allo studio dei fenomeni nei vari campi, ed per questo che sono nate le statistiche applicate.

1. Fenomeni e metodi statistici

L'indagine statistica è basata sull'osservazione dei fenomeni che, naturalmente, possono manifestarsi nelle forme più varie. Per fenomeno intendiamo tutto ciò che accade intorno a noi o che noi stessi provochiamo. Tutti i fenomeni che si presentano sempre con le stesse caratteristiche sono detti *fenomeni tipici* (un fenomeno tipico è quello che nasce dall'osservare che un corpo abbandonato ad una certa altezza cade verso il basso a causa della forza di gravità).

Vi sono però dei fenomeni che si manifestano, ogni volta, con caratteristiche diverse per i quali è quindi difficile fare delle previsioni. Questi sono definiti *fenomeni atipici* (ne sono un esempio tutti i fenomeni meteorologici).

Quando invece si prendono in esame fenomeni come le nascite, i matrimoni, i fenomeni sociali, economici, ecc. che riguardano, appunto, delle collettività (note più comunemente con il nome di *popolazioni*) vengono detti fenomeni collettivi. In questo caso se si effettuano osservazioni numerose su tali fenomeni, essi rivelano determinate caratteristiche uniformi per cui si può dire che, pur essendo fenomeni atipici, considerati collettivamente, presentano un comportamento regolare che permette di studiarne le leggi che li governano.

I metodi di ricerca delle leggi statistiche sono principalmente due: il metodo *induttivo* e quello *deduttivo.*

Si ha il metodo induttivo quando, partendo dall'osservazione di singoli fatti, generalizzando, si risale alle leggi di carattere generale relative ai fatti studiati.

Nel caso invece che si stabiliscano a priori degli assiomi che si pongono come premessa al processo logico per poi, attraverso un ragionamento logico, ricavarne le conseguenze, si fa uso del metodo deduttivo.

Possiamo quindi dire che nel primo caso si procede dal particolare al generale, nel secondo dal generale al particolare.

Ora, poiché la statistica è uno strumento della ricerca scientifica basata sull'osservazione (viene infatti applicata a fenomeni naturali, economici, fisici, ecc.) non riproducibili per via sperimentale, è conveniente usare il metodo induttivo.

2. I dati statistici e la loro natura

La statistica non è solo necessaria allo studio di quelle scienze che non esisterebbero neppure senza il ricorso ad essa (quali la demografia, la sociologia, ecc.), ma ormai la sua utilità è accertata in tutti i rami delle scienze naturali, sociali, economiche e in particolare in quelle riguardanti

l'istruzione, il commercio, l'industria, l'agricoltura, il turismo, ecc. Ed è dai dati statistici che si apprendono le condizioni di una nazione. Infatti le caratteristiche demografiche di una nazione si compiono attraverso la determinazione del numero dei suoi abitanti riportati per sesso, per età, professione, ecc. Così, ad esempio, lo studio dell'infortunistica stradale viene fatto attraverso il computo del numero degli infortuni ripartiti secondo le modalità del loro verificarsi. È chiaro quindi che per queste indagini occorrono dei metodi appropriati.
L'insieme di questi metodi costituisce la *statistica metodologica.*
Viene definita *unità statistica* ciascuna osservazione fatta sul fenomeno da indagare. Esempi di unità statistiche sono rappresentati da ogni individuo di una certa popolazione, ogni nato, ogni studente, ogni infortunio sul lavoro, ecc. Queste unità statistiche vengono poi esaminate e classificate secondo certe caratteristiche. Ad esempio, nel caso si volesse studiare il fenomeno della mortalità in una certa popolazione, i morti possono essere distribuiti in vari gruppi: o secondo l'età in cui è avvenuto il decesso, o secondo le cause che l'hanno determinato, ecc. Il numero delle unità statistiche rilevate costituisce il cosiddetto *dato statistico.* Il dato statistico, quindi, non è altro che il risultato di diverse osservazioni effettuate su più fenomeni singoli. Nel caso che il dato statistico esprima quante volte si manifestata quella modalità, esso viene detto *frequenza* di quella modalità. Se invece esprime una misura (peso, lunghezza, velocità, volume, ecc.) il dato statistico viene detto *intensità* di quella modalità.
Sono dati statistici di frequenza il numero dei nati (o dei morti) in un certo periodo, il numero delle persone colpite da influenza, il numero di transazioni commerciali in un determinato settore, eccetera.
Sono dati statistici di intensità la quantità di merci oggetto di transazioni in un certo periodo ed in un certo settore, i salari corrisposti a determinate categorie dei lavoratori, eccetera.
L'intensità di un fenomeno collettivo può essere poi *globale* o *media.* È globale quando è la somma delle intensità di diversi fenomeni individuali mentre è media quando è un valore determinato variamente (come vedremo in altro capitolo).
Le unità statistiche vengono inoltre studiate secondo uno o più caratteri comuni che rappresentano gli aspetti che si vogliono mettere in evidenza e, successivamente, sono divisi rispetto alle varie modalità con cui tale carattere si manifesta.
Le modalità secondo cui vengono classificate le unità statistiche possono essere *quantitative* o *qualitative.*
Le modalità quantitative sono espresse da numeri risultanti da misurazioni o da enumerazioni (es. la rilevazione dei redditi di una

popolazione, le altezze dei militari di leva, il numero dei vani degli alloggi in un comune, ecc.).
Le modalità qualitative sono espresse da attributi (es. la rilevazione della popolazione italiana secondo la stato civile, il colore degli occhi della popolazione di un paese, ecc.).
Per chiarire meglio quanto detto osserviamo le seguenti tabelle:

Colore degli occhi	*Numero di studenti*
Celeste	12
Grigio	7
Castano	24
Nero	4
Totale	47

« *Colore degli occhi* » = carattere qualitativo
« *Celeste, grigio,...* » = modalità
« 12, 7, 24, 4 » = dati statistici.
I numeri 12, 7, 24, 4 rappresentano anche le frequenze.

Frutta	*Quantità*
Mele	4000
Pere	2850
Pesche	2500
Ciliege	1050
Uva	5000
Totale	15.400

« *Frutta* » = carattere qualitativo
« *Mele, pere, pesche, ciliegie, uva* » = modalità
« 4000, 2850, 2500, 1050, 5000 » = dati statistici.
I numeri 4000, …. , 5000 rappresentano le intensità.

Numero delle stanze	*Numero delle abitazioni*
1	2015
2	3203
3	4506
4	5324
5	4813
6	2906
7	1341
Totale	24108

«*Numero delle stanze* » = carattere quantitativo
« 1,2, 3, 4, 5, 6, 7 » = modalità
«2015, ….. , 1341 » = dati statistici (che rappresentano le frequenze).

3. Le rilevazioni statistiche

La prima operazione da compiere per analizzare un fenomeno collettivo, è quella della rilevazione, la quale consiste nella raccolta dei dati statistici riguardanti i fenomeni individuali che compongono il fenomeno collettivo oggetto dell'indagine. Una rilevazione statistica può avere caratteristiche diverse e può essere:

saltuaria o *continua*
pubblica o *privata*
parziale o *totale*
diretta o *indiretta*
preliminare o *definitiva*

Ad esempio un censimento è una rilevazione saltuaria, pubblica e totale. Per avere i dati dei nati in una certa popolazione (o la quotazione di alcune merci) si richiedono, invece, rilevazioni continue, pubbliche e complete.
Una rilevazione statistica richiede, innanzitutto, l'esatta definizione del fenomeno da rilevare ed occorre anche stabilire il *modo*, il *tempo*, e lo *spazio* in cui essa deve essere effettuata e quali sono gli organi ed i mezzi interessati alla rilevazione.
Il *modo* secondo cui può essere condotta la rilevazione si distingue in:

automatica quando deriva da dichiarazioni provenienti direttamente dalle persone interessate (es. le rilevazioni dell'ufficio di stato civile per le nascite, i morti, i matrimoni, ecc.);
riflessa quando i dati vengono raccolti da appositi rilevatori (es. il censimento).

Riguardo al tempo la rilevazione può essere:

continua quando le rilevazioni vengono registrate man mano che i fenomeni si verificano;
periodica quando viene effettuata ad intervalli regolari di tempo (es. il censimento);
occasionale quando viene compiuta senza alcuna periodicità (es. la rilevazione dei danni provocati da una guerra, oppure i sondaggi politici).

Gli organi che eseguono le rilevazioni statistiche possono *pubblici* o *privati*. Le rilevazioni compiute dagli organi pubblici riguardano fenomeni di interesse pubblico come, ad esempio, quelle di carattere demografico ed economico. In Italia il principale organo pubblico dedito agli studi di statistica è l'ISTAT. Le rilevazioni private sono compiute da imprese commerciali su determinati fenomeni che rivestono particolare interesse di ricerca per alcuni privati.
Per quanto riguarda i mezzi con i quali possono essere condotte le rilevazioni statistiche diciamo soltanto che per le rilevazioni automatiche si usano registri, ruoli, ecc., mentre per quelle riflesse si usano dei questionari.

4. Lo spoglio dei dati

Una volta ultimata la raccolta delle unità statistiche si riunisce tutto il materiale e si procede a controlli di natura diversa per cercare di eliminare inesattezze ed errori. Una volta eseguiti i controlli si passa allo spoglio ed alla classificazione dei risultati raggruppando gli elementi raccolti secondo i caratteri prestabiliti formando delle tabelle di spoglio. Queste sono costituite da varie colonne o righe che sono riferite ai diversi caratteri del fenomeno collettivo che sono stati oggetto della rilevazione. In ciascuna colonna (o riga) vengono riportati i rispettivi dati di frequenza che sono stati rilevati. Le tabelle statistiche si dividono in *semplici*, *complesse* e *a doppia entrata*.
Le tabelle semplici sono prospetti nei quali sono elencate le modalità qualitative o quantitative del fenomeno in esame ed a fianco le relative frequenze o intensità.
Esempio. Riportiamo una tabella semplice riguardante la distribuzione di una popolazione di 10.000 individui secondo la statura suddivisa in classi di intensità di 10cm in 10cm a partire dall'altezza di 120cm:

Statura	*Frequenza*
120-130	4
130-140	15
140-150	136
150-160	2324
160-170	5604
170-180	1801
180-190	98
190-200	18
TOTALE	10.000

Le tabelle complesse possono ritenersi una composizione di tabelle semplici che presentano dati statistici riguardanti più fenomeni.

Osservatori	*Pressione media*	*Umidità relativa*	*Sereno*	*Misto*	*Coperto*
Torino	1016,2	80	104	112	150
Milano	1017,3	77	103	101	161
Venezia	1015,5	76	112	127	127
Bologna	1016,8	72	106	109	150
Firenze	1015,6	71	97	115	154
Roma	1015,8	72	119	147	100
Pescara	1014,9	75	109	127	130
Napoli	1015,2	69	107	127	132
Palermo	1015,1	60	121	141	104
Cagliari	1016,5	69	125	144	96

Quando lo spoglio delle unità statistiche è stato effettuato secondo due caratteri (ad esempio numero di abitanti e numero di stanze per unità abitative), l'osservazione di ogni unità statistica conduce a due risultati. Allora per la rappresentazione di queste distribuzioni statistiche si fa uso delle tabelle a doppia entrata:

Prof.ne padre	*Licei Cl. e Sci.*	*Ist.Tecnici*	*Ist.Prof.*	*Totale*
Agricoltori	50	160	90	300
Artigiani	20	50	30	100
Commercianti	60	110	30	200
Impiegati	150	440	110	700
Liberi Prof.sti	90	50	10	150
Operai	10	60	330	400
Altre Prof.ni	20	30	100	150
Totale	400	900	700	2000

Le modalità del primo carattere sono rappresentate dal tipo di scuola, mentre quelle del secondo carattere sono rappresentate dalla professione del padre. In questo caso tanto le modalità del primo carattere quanto quelle del secondo sono qualitative.

5. Serie e seriazioni statistiche

Prima di procedere all'elaborazione dei dati bisogna fare una distinzione fra le distribuzioni statistiche provenienti da caratteri qualitativi da quelle provenienti da caratteri quantitativi

Orbene definiamo *serie statistica* una distribuzione statistica a carattere qualitativo. Chiameremo invece *seriazione statistica* una distribuzione avente carattere quantitativo.
Così, ad esempio, una distribuzione di dati statistici riguardanti una popolazione ripartita secondo la professione degli abitanti costituisce una serie statistica. Così è pure una serie statistica la distribuzione degli individui di una collettività secondo il colore degli occhi.
Se invece consideriamo una distribuzione di dati statistici riguardanti la ripartizione dei contribuenti secondo l'ammontare delle imposte cui sono soggetti, si ha una seriazione poiché la modalità assunta a base della ripartizione è di carattere quantitativo. Costituisce anche una seriazione la ripartizione di 1000 conteiner di un cargo secondo classi di peso. Quindi, per distinguere una serie da una seriazione è sufficiente stabilire se il carattere è qualitativo o quantitativo.
Tra le serie statistiche rivestono particolare importanza le *serie temporali* (o *serie storiche*) e quelle di *luogo* (o *territoriali*).
Sono serie temporali quelle in cui viene esposta la distribuzione di un dato fenomeno nel tempo. Tipici esempi ne sono i dati statistici relativi alle produzioni industriali nei vari anni, quelli relativi alla natalità (o mortalità) distinti per giorni, mesi, anni, ecc. A loro volta le serie storiche possono essere *statiche* (quando non vi sono variazioni apprezzabili) e *dinamiche* (quando il fenomeno preso in considerazione tende a diminuire o ad aumentare).
Una serie è di luogo (o territoriale) quando la distribuzione del fenomeno avviene nello spazio. Ad esempio, la serie dei nati in Italia in un dato anno distinti per regione costituisce una serie territoriale.
Per concludere possiamo dire che, rispetto alla disposizione da darsi alle modalità del fenomeno preso in considerazione, le serie statistiche si distinguono in:

a) *Serie rettilinee* che sono quelle le cui modalità vengono disposte secondo un ordine logico o naturale dal principio alla fine (es. è rettilinea la serie temporale dei nati vivi in Italia, di anno in anno, dal 1991 al 2001;
b) *Serie cicliche* che sono quelle le cui modalità si succedono secondo un ordine logico il quale però si ripete ciclicamente. Ne costituiscono un tipico esempio quelle che espongono dati relativi alle stagioni;
c) *Serie sconnesse* che sono quelle le cui modalità non necessitano di alcun ordine. È sconnessa, ad esempio, la serie che rappresenta la distribuzione di una data popolazione secondo la professione o la religione degli individui.