A13

La copertina è stata realizzata dall'arch. Rosalia Mangiacavallo.

Mariantonietta Ruggieri

Conoscere la Statistica per interpretare i dati

Aracne editrice

www.aracneeditrice.it
info@aracneeditrice.it

www.gioacchinoonoratieditore.it
info@gioacchinoonoratieditore.it

via Vittorio Veneto, 20
00020 Canterano (RM)
(06) 45551463

ISBN 978-88-255-2960-9

I edizione: dicembre 2019

A mia madre

Coloro che amiamo e che abbiamo perduto non sono più dove erano, ma sono ovunque noi siamo.

Sant'Agostino

Indice

Prefazione

Il volume nasce dall'esigenza di pubblicare un manuale creato inizialmente per gli studenti di Statistica 1 dei corsi di laurea triennale in Economia e Finanza, Amministrazione ed Economia delle Imprese e Statistica per l'analisi dei dati dell'Ateneo di Palermo, presso cui l'autore insegna. Il manuale ancora oggi è usato dagli studenti di Statistica del corso di laurea in Scienze del Turismo. Il testo presenta numerosi esempi che introducono gli argomenti trattati e ne rendono più facile la lettura e la comprensione. Il testo riporta, altresì, i passaggi dettagliati di dimostrazioni e formule, utili a chi si approccia alla Statistica per la prima volta. Gli argomenti trattati riguardano fondamentalmente la Statistica descrittiva (scale di misura e classificazione di variabili, costruzione di tabelle e grafici, indici di sintesi, relazioni tra variabili); un capitolo introduce anche al Calcolo delle probabilità e all'adattamento di alcuni modelli teorici (binomiale, Poisson, Gauss) a distribuzioni empiriche. Gli esempi e gli esercizi proposti sono stati svolti in aula, con l'ausilio del foglio elettronico Excel, durante le esercitazioni. In Appendice sono inclusi, infine, alcuni testi di compiti lasciati durante le prove di esame.

Capitolo I

Cos'è la Statistica

1.1. Cenni storici

Tracce di ciò che potremmo definire statistiche si riscontrano già dai tempi della preistoria; l'uomo, infatti, da sempre ha sentito l'esigenza di quantificare e registrare avvenimenti importanti della propria vita, come l'ammontare delle nascite, delle morti, del numero dei capi di bestiame posseduti, dei prodotti agricoli raccolti e scambiati, e così via.

Ma la Statistica come disciplina vera e propria nasce in Inghilterra e in Germania intorno al 1600; essa si occupa dello studio dei fenomeni demografici, sociali e dei principali fatti riguardanti la vita di uno Stato.

Nello stesso periodo nasce il Calcolo delle probabilità, branca della Matematica, tuttavia per lungo tempo resta confinato ai giochi d'azzardo; solo successivamente il Calcolo delle probabilità, e più in generale la Matematica, diventano uno strumento fondamentale per la metodologia statistica, in particolare per la *Statistica inferenziale*.

Con lo sviluppo dell'Informatica la metodologia statistica ha fatto un enorme passo avanti; oggi è possibile trattare una gran quantità di dati ed effettuare elaborazioni prima impossibili da eseguire manualmente o comunque in tempi brevi.

Oggi tutti i paesi industrializzati dispongono di SERVIZI STATISTICI NAZIONALI, per cui non solo è aumentata la quantità disponibile dei dati statistici, ma ne è migliorata anche la qualità.

In Italia l'ISTAT nasce come Istituto autonomo nel 1926; dal 1989 è un Istituto di Stato a gestione autonoma, dotato di personalità giuridica, ed è diventato "Istituto nazionale di Statistica", sotto la dipendenza del Consiglio dei Ministri. L'ISTAT ha sede in Roma ed ha il compito di raccogliere, elaborare e diffondere informazioni statistiche riguardanti tutti gli aspetti (demografici, sociali, economici) della vita dello Stato.

L'ISTAT per legge non possiede il monopolio della informazione statistica; esistono anche altri enti, sia pubblici che privati, che producono statistiche di rilevante interesse nazionale, che non hanno però valore ufficiale. Si pensi, ad esempio, ai vari ministeri, ai comuni, alle regioni, alle province, nonché

alla Banca d'Italia. Altri enti sono la Camera di Commercio, la Confindustria, il Censis, la RAI, l'ENEL, l'ENI, la Doxa, la Demoskopea, e così via.

1.2. Fonti di dati

Le pubblicazioni ISTAT hanno carattere periodico; ci sono pubblicazioni annuali, decennali, ma anche occasionali e saltuarie. Citiamo fra le più importanti l'*Annuario*, il *Compendio*, il *Bollettino mensile*, gli *Annuari* specializzati, che costituiscono un'analisi dettagliata dei vari capitoli compresi nell'*Annuario*, oltre alle pubblicazioni dedicate ai Censimenti. Ricordiamo, infatti, che con periodicità decennale l'ISTAT effettua il Censimento della popolazione e delle abitazioni, il Censimento dell'agricoltura e il Censimento dell'industria, commercio, servizi e artigianato.

Oggi ci si può collegare a una *banca dati*, che consente di disporre di dati aggiornati in tempo reale su diversi fenomeni.

Ci sono alcuni paesi, come l'Africa, che non dispongono di un servizio statistico nazionale, per i quali non è mai stato effettuato un censimento e per i quali, dunque, è impossibile valutare i mutamenti e le dimensioni dei fenomeni demografici, economici, sanitari, ecc.

Per quanto riguarda le fonti statistiche internazionali, ricordiamo le pubblicazioni effettuate da alcuni organismi internazionali quali:

— l'ONU (*Statistical yearbook*, *Demographic yearbook*, *Yearbook of national accounts Statistics*, *Monthly bullettin of Statistics*);
— l'UNESCO (*Annuario dell'Istruzione*);
— la FAO (*Production yearbook*, *Trade yearbook*, *Yearbook of forest products*);
— il BIT-ILO (*Yearbook of labour Statistics*);
— l'OMS (*World health Statistics annual*);
— l'OCSE;
— il FMI;

e così via.

1.3. La Statistica come disciplina ausiliaria

La Statistica nasce come "Scienza di Stato", e in questo senso trovano una connotazione i "censimenti", ma col tempo assume un altro significato: «*la Statistica è una disciplina ausiliaria alle altre discipline scientifiche, di cui la disciplina principale è la fisica, e assume un ruolo fondamentale nel processo di*

acquisizione scientifico della conoscenza». Vediamo di capire meglio quanto affermato.

Il *Metodo Sperimentale*, come è noto, fu introdotto da *Galileo Galilei* intorno al 1600. Per molti secoli l'uomo, interrogandosi sul comportamento della natura e sul verificarsi di determinati fenomeni, ha trovato risposta nel ragionamento filosofico e in alcuni teorie, come quella aristotelica, servendosi della sola logica.

Il *Metodo Sperimentale* rivendica la necessità di "un'accurata sperimentazione" e riconosce la caducità di qualsiasi legge o modello, la cui importanza è assolutamente relativa.

Galilei evidenzia il valore del legame esistente fra:

— il mondo simbolico del razionale (TEORIA);
— il mondo empirico del reale (ESPERIENZA).

Secondo il metodo da lui fondato, la conoscenza passata di un fenomeno deve essere arricchita e integrata da nuove informazioni o esperienze, che consentono di formulare nuove ipotesi, le quali possono essere formalizzate mediante modelli o leggi. In questa fase interviene la Matematica, dunque il Calcolo delle probabilità. Le ipotesi vanno continuamente verificate e aggiornate, eventualmente sostituite, dopo aver osservato nuovi dati. In questa fase interviene la Statistica. Qualsiasi teoria, dunque, e di conseguenza qualsiasi scienza, ha carattere assolutamente temporaneo.

In tale processo scientifico induttivo–deduttivo di acquisizione della conoscenza, la Statistica ricopre il ruolo essenziale di "disciplina ausiliaria".

Essa interviene nelle seguenti fasi:

— osservazione dei caratteri che descrivono un fenomeno;
— raccolta delle informazioni sotto forma di dati, loro organizzazione, elaborazione e sintesi;
— verifica di conformità dei modelli teorici alla realtà.

1.4. Fenomeni ripetibili, parzialmente ripetibili, non ripetibili

La Statistica, dopo aver organizzato i dati, li predispone per l'analisi e li elabora per sintetizzare, nel modo migliore, le informazioni in essi contenute. L'obiettivo è quello di ottenere alcuni indici appropriati, che consentano di avere una visione globale del fenomeno oggetto di studio.

La fase dell'elaborazione dei dati, e in particolare quello della sintesi, è un momento molto importante e dipende:

— dal particolare tipo di fenomeno studiato;
— dalla natura del carattere osservato;
— dalla tipologia degli errori che influenzano i dati.

I fenomeni in natura possono essere distinti in:

— ripetibili
— parzialmente ripetibili
— non ripetibili

I *fenomeni ripetibili* sono quei fenomeni del reale per i quali è possibile ripetere più volte e nelle stesse condizioni la misura di una grandezza incognita.

Ciascuna misura x_i è affetta da errori ε_i di natura accidentale:

$$x_i = X + \varepsilon_i$$

Tali errori sono ineliminabili, qualunque sia la cura dei rilevatori e la precisione degli strumenti di misura. È compito della Statistica trovare il modo migliore di combinare le osservazioni, al fine di ottenere la migliore valutazione del vero valore della grandezza incognita X.

I *fenomeni parzialmente ripetibili* sono quei fenomeni del reale legati all'evoluzione delle stagioni. È noto, ad esempio, che in Sicilia a giugno matura il grano, a settembre l'uva, a novembre le olive.

Per questi fenomeni le metodologie statistiche disponibili sono meno informative rispetto a quelle relative ai fenomeni ripetibili.

I fenomeni non ripetibili sono quei fenomeni del reale per i quali interviene la variabilità biologica. Ogni uomo, ad esempio, presenta caratteristiche diverse tali da rendere impossibile la "ripetibilità della prova".

Per questi fenomeni le metodologie statistiche risultano scarsamente informative.

1.5. Scale di misura e classificazione delle variabili statistiche

La qualità e il significato dell'informazione sintetica ricavata, tramite l'analisi statistica, dalle singole osservazioni dipendono fortemente dalla natura del fenomeno, ma dipendono anche dal tipo di carattere che lo descrive e dalla sua misurabilità.

In Statistica distinguiamo diversi tipi di CARATTERI O VARIABILI, in relazione a quattro distinte SCALE DI MISURA:

— nominale;

— ordinale;
— di intervalli;
— di rapporti.

Un carattere è esprimibile su scala nominale o cardinale se fra le modalità del carattere si può stabilire solo una relazione di EQUIVALENZA. In tal caso, il carattere prende il nome di VARIABILE QUALITATIVA SCONNESSA O MUTABILE.

Esempi di variabile qualitativa sconnessa sono:

— il sesso;
— la nazionalità.

Rilevati su n soggetti il sesso e/o la nazionalità, è possibile dire solo se due diversi soggetti hanno uguale sesso/nazionalità oppure no. Questo tipo di dati ha, pertanto, un contenuto informativo molto basso.

Un carattere si dice misurabile su scala ordinale, e in tal caso prende il nome di VARIABILE QUALITATIVA ORDINABILE O GRADUABILE, se fra le modalità del carattere è possibile stabilire, oltre a una relazione di equivalenza, anche una relazione d'ORDINE. In poche parole, fra le modalità è possibile formulare una graduatoria:

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \cdots \leq x_{(n)}$$

Il contenuto informativo di tali variabili è pertanto maggiore rispetto a quello delle variabili considerate in precedenza.

Esempi di variabili qualitative ordinabili sono:

— il titolo di studio;
— la qualifica professionale.

In tal caso, di due soggetti diversi, è possibile dire se hanno lo stesso titolo di studio o la stessa qualifica professionale, ma è anche possibile stabilire chi ha il titolo di studio o la qualifica migliore.

In genere, quando si parla semplicemente di "caratteri", si intendono le "variabili qualitative".

Le VARIABILI QUANTITATIVE, o semplicemente le VARIABILI, a differenza delle variabili qualitative, sono espresse da valori numerici.

Le variabili quantitative si distinguono in:

— discrete;
— continue.

Le variabili quantitative discrete possono anche derivare da enumerazione o conteggio di oggetti o soggetti e assumono valori interi positivi.

Esempi di variabili quantitative discrete sono:

— il numero di figli di una famiglia;
— il numero di vani di un appartamento.

Le variabili quantitative continue sono espresse da "misure" (numeri razionali o, più in generale, reali) e possono assumere infiniti valori all'interno di un intervallo.

Esempi di variabili quantitative continue sono: la statura, il reddito, il tempo.

Un carattere quantitativo continuo si dice misurabile su SCALA A INTERVALLI se fra i valori del carattere è possibile stabilire una relazione di:

— equivalenza;
— ordine;
— uguaglianza ($x_{i+1} - x_i = x_{j+1} - x_j$).

Per i valori di tali caratteri sono lecite le operazioni di addizione e sottrazione; la differenza fra due punti della scala è uguale alla differenza fra altri due punti della scala che hanno la stessa distanza. Ovvero un intervallo, preso in diversi punti della scala, deve rappresentare sempre la stessa quantità.

Un carattere quantitativo si dice misurabile su SCALA DI RAPPORTI se tra i valori del carattere è possibile stabilire una relazione di:

— equivalenza;
— ordine;
— uguaglianza;
— rapporto ($x_{i+1}/x_i = x_{j+1}/x_j$).

Le variabili quantitative continue misurabili su scala di rapporti hanno, dunque, un contenuto informativo molto elevato.

Per i valori di tali caratteri sono lecite, oltre alle operazioni di addizione e sottrazione, anche le operazioni di moltiplicazione e divisione; il rapporto fra due punti della scala è uguale al rapporto fra altri due punti della scala che hanno la stessa distanza.

La temperatura (in gradi Celsius, Fahrenheit, Reamur), il peso, la statura sono variabili misurabili su scala di intervallo; sono misurabili su scale di rapporto se rilevate sempre nelle stesse condizioni fisiche, per esempio nello stesso luogo. Lo zero della scala è, infatti, uno zero convenzionale e

non coincide con lo zero assoluto (zero fisico, reale). La temperatura in gradi Kelvin, invece, è sempre misurabile su scala di rapporti, perché lo zero della scala coincide con lo zero assoluto, che è il punto in cui le molecole di qualsiasi gas non si muovono più. Tali variabili, dunque, non possono assumere valori negativi.

VARIABILI SEMPLICI E MULTIPLE

Raramente in natura i fenomeni sono descritti da un solo carattere.

Quando su uno stesso oggetto o soggetto si rilevano contemporaneamente le modalità o i valori di k caratteri siamo in presenza di una VARIABILE MULTIPLA.

Una variabile multipla è OMOGENEA se le k variabili che la compongono sono tutte rilevate con la stessa scala di misura, è MISTA in tutti gli altri casi.

In Statistica si impiegano metodologie diverse a seconda se i dati sono omogenei o misti.

1.6. Gli errori nei dati

Gli errori modificano la qualità dell'informazione contenuta nei dati.

Si suddividono in:

— grossolani;
— sistematici;
— accidentali.

Gli errori GROSSOLANI sono dovuti, ad esempio, a un rilevatore maldestro o a una immissione errata dei dati. Gli errori SISTEMATICI sono dovuti a strumenti poco precisi o tarati male. Gli errori ACCIDENTALI sono dovuti, invece, ad infinite cause perturbatrici, infinitesime, spesso non note.

In un'indagine statistica seria gli errori grossolani e gli errori sistematici non dovrebbero mai essere presenti. La Statistica ha perciò il compito arduo di eliminare gli errori accidentali o meglio di individuare la migliore combinazione delle osservazioni ai fini di ridurne l'influenza.

1.7. I dati statistici

I dati statistici possono essere suddivisi in dati spaziali, temporali, territoriali.

I *dati spaziali* sono indipendenti dal luogo e dal tempo, per cui non è importante l'ordine con cui sono stati rilevati. Volendo, ad esempio, indagare sul carattere "statura" degli studenti che compongono una classe, è possibile effettuare le rilevazioni in giorni e in ambienti diversi.

I *dati temporali* (serie storiche) dipendono fortemente dal tempo, per cui è importante effettuare un'osservazione in un determinato istante piuttosto che in un altro. Si pensi, ad esempio, se si vuole studiare la legge di accrescimento del peso di una cucciolata durante il primo anno di vita.

I dati territoriali dipendono dal luogo in cui sono stati osservati. Si pensi, ad esempio, se si vogliono effettuare studi sulla natalità o sulla mortalità di una determinata regione geografica.

1.8. Popolazioni e campioni di dati

Non sempre è possibile disporre di tutti i dati necessari per descrivere un fenomeno, cioè di tutta la POPOLAZIONE O UNIVERSO dei dati.

Per motivi di tempo o di costo, o semplicemente per impossibilità, il più delle volte si ricorre a un CAMPIONE sufficientemente rappresentativo della popolazione.

Dalle proprietà sintetiche rilevate sul campione si "inferisce" poi alle proprietà incognite dell'universo dei dati. A disciplinare tale procedura è una branca particolare della Statistica, denominata "Statistica inferenziale".