## Indice-Sommario

APPENDICE *A*

# CAPITOLO 1

# NOZIONI INTRODUTTIVE

## 1.1. Introduzione

È opportuno introdurre la nozione di dati statistici o, in breve, di "statistiche". Con questa espressione si intendono le informazioni espresse numericamente - percentuali, medie, frequenze di accadimento di eventi in un intervallo di tempo, ecc. - riferite ad un insieme di entità omogenee da qualche punto di vista (persone, oggetti, aziende, situazioni), che per ora verrà chiamato "insieme di riferimento".

Sono statistiche, ad esempio: i dati sulla variazione media dei prezzi che l'Istituto Nazionale di Statistica comunica con periodicità mensile; il numero degli occupati e dei disoccupati che lo stesso Istituto fa conoscere periodicamente; i dati sugli incidenti mortali occorsi in una data settimana sulle strade; le percentuali relative agli orientamenti di voto ottenute tramite un sondaggio, ecc. Negli esempi, gli insiemi di riferimento sono, nell'ordine, i prezzi di beni e servizi sul mercato al consumo, la popolazione attiva (persone al di sopra di una certa età che lavorano o intendono lavorare), i sinistri verificatisi sulle strade nella settimana considerata, i cittadini che hanno diritto al voto.

La "produzione" di statistiche è un'attività spesso complessa che segue le regole, le procedure e i metodi propri della disciplina che si chiama Statistica.

Ma la produzione delle statistiche è solo la manifestazione più evidente e immediata dell'applicazione della Statistica: vi sono comportamenti, decisioni, scelte, nell'attività professionale e anche nella vita quotidiana che si basano su informazioni di tipo qualitativo che, però, sottendono un'attività investigativa di tipo statistico. Così, il medico di famiglia che prescrive una nuova medicina al proprio paziente lo fa, generalmente, perché ha letto nel relativo foglio illustrativo (o ha saputo dal proprio informatore scientifico) che essa è adatta allo scopo. Questa informazione non è un dato statistico, ma poggia su un'attività statistica che non appare: una o più sperimentazioni cliniche effettuate su un conveniente numero di pazienti che hanno dimostrato, al di là di ogni ragionevole dubbio, l'efficacia del farmaco, da una parte, e, dall'altra, l'assenza di seri effetti collaterali. Nella situazione descritta, la Statistica non si limita ad indicare

gli strumenti per misurare l'efficacia del farmaco (ad esempio, il confronto tra le percentuali di guarigione tra coloro che hanno assunto il farmaco e quelli a cui è stato somministrato il placebo), ma presiede all'impostazione stessa della ricerca, con la formulazione di quello che si chiama il piano degli esperimenti (selezione del gruppo sperimentale, determinazione delle dosi del farmaco, modalità di accertamento dei risultati, ecc.). L'attività investigativa descritta culmina nella "generalizzazione dell'evidenza osservata". Lo sperimentatore arriva ad affermare che il farmaco, al di là di ogni ragionevole dubbio, è efficace, nel senso che è molto elevata la probabilità di guarigione per il paziente (con caratteristiche simili a quelle del gruppo sperimentale) che l'assumerà.

Come altro esempio di analisi statistica che non ha come fine ultimo la produzione di statistiche, si consideri l'attività dell'addetto al controllo di processo in un'azienda industriale. Quando egli comunica alla direzione che il processo è sotto controllo esprime un giudizio qualitativo sullo stato del processo: il giudizio è, però, basato sull'osservazione delle caratteristiche di un campione di pezzi prodotti e sulla constatazione che la percentuale dei pezzi che si discostano dagli standard più del dovuto è al di sotto dei limiti di tolleranza. Qui la Statistica è chiamata in causa per diversi aspetti: per la fissazione dei limiti di tolleranza, per la scelta del campione e, infine, per l'osservazione dei dati. Anche in questa situazione, si è dinanzi ad una generalizzazione dell'evidenza empirica tratta dal campione.

Alla luce delle precedenti esemplificazioni, si può così definire il contenuto della Statistica.

**Definizione 1.1.** *La Statistica è la disciplina che elabora i principi e le metodologie che presiedono*: *al processo di rilevazione e raccolta dei dati*, *alla rappresentazione sintetica e alla interpretazione dei dati stessi*, *e*, *laddove ve ne siano le condizioni*, *alla generalizzazione delle evidenze osservate*.

La definizione apparirà più chiara via via che si procederà nell'esposizione degli argomenti: se ne consiglia, pertanto, la rilettura quando si avrà contezza dei contenuti della disciplina. Qui è necessario precisare che la locuzione "laddove ve ne siano le condizioni", riferita alla generalizzazione dei risultati, allude al fatto che solo il calcolo delle probabilità consente di effettuare induzioni e che, quindi, si devono verificare i presupposti per l'applicazione appropriata di questo strumento.

## 1.2. Cenni storici sullo sviluppo della Statistica

La statistica è una disciplina relativamente giovane: il suo sviluppo è avvenuto in gran parte nei secoli XIX e XX; le sue origini come disciplina autonoma risalgono al XVII secolo quando in Inghilterra si sviluppò, ad opera di John Graunt (1620–1674) e William Petty (1623–1687), un indirizzo di ricerca, che prese il nome di *Political arithmetic*, caratterizzato dall'uso del metodo empirico induttivo, proprio delle scienze naturali, nell'investigazione dei fenomeni demografici e sociali[1]. Il merito di queste ricerche sta nell'aver messo in risalto, tra i contemporanei, l'importanza della raccolta e dell'appropriato uso dei dati, più che nell'aver introdotto specifiche metodologie di analisi.

Nasceva così una nuova disciplina che, però, non aveva ancora assunto il nome di Statistica. Secondo l'opinione prevalente, la parola statistica venne impiegata per la prima volta per designare l'indirizzo scientifico iniziato in Germania da Hermann Conring (1606–1681), con l'avvio di un insegnamento universitario di scienza politica avente per obiettivo la descrizione delle "cose notevoli di uno Stato".

Dovettero trascorrere circa due secoli prima che l'indirizzo investigativo inaugurato dalla *Political arithmetic* potesse assumere alcuni dei connotati della Statistica moderna: infatti, due secoli occorsero perché si sviluppasse quella branca della Matematica che ha a che fare con la logica dell'incerto, il Calcolo delle probabilità, e perché fosse affrontato il problema della "misura dell'incertezza".

Per misura dell'incertezza si intende qui la quantificazione del grado di credibilità da attribuire ai risultati dell'indagine empirica: Graunt poteva calcolare, sulla base delle registrazioni ecclesiastiche (cfr. Nota 1), che il tasso di sopravvivenza di maschi dai 50 ai 70 anni fosse, ad esempio, pari al 40%, ma non era in grado di effettuare in alcun modo una valuta-

[1] Un esempio di applicazione del nuovo approccio allo studio dei fenomeni demografici riguarda la determinazione del numero delle famiglie di Londra nel 1660 (esposta nell'opera del Graunt dal titolo *Natural and political observations upon the bills of mortality* … pubblicata nel 1662 e basata sulle risultanze dei registri dei battesimi e dei defunti istituite dalla Chiesa d'Inghilterra): osservando che durante un anno si registravano 3 morti ogni 11 famiglie, calcolò che vi dovessero essere circa 84.000 famiglie, dato che il numero dei decessi in quell'anno erano stati circa 23.000.

Dall'osservazione dei dati con metodo scientifico, Graunt pervenne alla scoperta di vere e proprie leggi che governano i fenomeni demografici e sociali, quali l'eccedenza delle nascite maschili su quelle femminili, l'inurbamento delle popolazioni rurali, ecc.

zione del grado di errore o della validità di questa stima, valutazione indispensabile per la generalizzazione del risultato.

Il Calcolo delle probabilità è lo strumento per la soluzione del problema diretto: usando una metafora, consente di calcolare la probabilità di estrarre una pallina di un dato colore, nota la composizione dell'urna; la Statistica ha il compito di risolvere quello che si chiama il "problema inverso": di rispondere, cioè, a domande del tipo "quale è la percentuale di palline bianche nell'urna, avendo osservato il numero di palline bianche apparse in un certo numero di prove?".

Lo sviluppo del Calcolo delle probabilità va ascritto a grandi matematici, tra i quali meritano particolare menzione Blaise Pascal (1623-1662), Pierre Fermat (1601-1665), Jacob Bernoulli (1654-1705), Abraham De Moivre (1667-1754), Thomas Bayes (1702-1761), Pierre Simon Laplace (1749-1827), Adrien Marie Legendre (1752-1833), Carl Friedrich Gauss (1777-1855). In verità, i contributi scientifici degli ultimi quattro studiosi non sono di pura probabilità: viene affrontato, in diverse guise, il problema inverso, e vengono ottenuti risultati essenziali per lo sviluppo della Statistica.

A Bayes va ascritto il merito di aver affrontato, se non risolto (a causa di calcoli matematici troppo complessi per quei tempi), il seguente problema. Sia $\theta$ la probabilità che un evento $E$ si verifichi in una prova, quale è la probabilità che $\theta$ assuma un valore compreso nell'intervallo $(a, b)$, sapendo che in $n$ prove l'evento $E$ si è verificato $X$ volte? Si tratta, evidentemente, di una questione importante: poter dire, alla luce delle osservazioni empiriche, che la quantità incognita $\theta$ (ad esempio la probabilità di sopravvivenza a 70 anni di un cinquantenne, quantità evocata in precedenza) è compresa con un elevato livello di probabilità entro un determinato intervallo numerico è un genuino processo induttivo.

Il campo di ricerca di Laplace fu molto vasto. Limitandosi agli argomenti più attinenti alla Statistica, vanno ricordati: il problema induttivo già affrontato da Bayes[2], la scelta della media per sintetizzare un insieme di misure ripetute di una stessa grandezza allo scopo di stimare al meglio tale grandezza, il cosiddetto teorema del limite centrale, per il quale, sotto certe condizioni, la probabilità che la media di un numero elevato di

[2] Come applicazione del suo metodo, Laplace considerò il fenomeno dell'eccedenza delle nascite maschili (fenomeno già osservato da Graunt; cfr. Nota 1). Indicando con $\theta$ la probabilità di una nascita maschile, trovò che la probabilità che $\theta \leq 0{,}5$ era pari a $1{,}1521 \times 10^{-42}$, ciò alla luce dell'osservazione empirica che a Parigi, nel periodo 1745-1770, erano nati 251.527 maschi e 241.945 femmine. Da ciò trasse la conclusione che doveva essere $\theta > 0{,}5$.

quantità omogenee (ad esempio, il peso medio dei pezzi provenienti da un determinato processo produttivo) sia compresa in un dato intervallo $(a, b)$ può essere approssimata tramite la curva normale.

Un importante avanzamento che avrà un impatto fondamentale sullo sviluppo della Statistica è rappresentato dalla formulazione del metodo dei minimi quadrati. Nella sua applicazione più semplice, il metodo risolve il problema seguente. Date $n$ misure di una stessa grandezza, qual è la sintesi di tali misure che meglio approssima la grandezza incognita? Secondo il metodo dei minimi quadrati, la sintesi migliore è data dalla media aritmetica delle $n$ misure. L'introduzione del metodo viene attribuita congiuntamente a Legendre e a Gauss[3] e l'importanza che ebbe all'epoca era legata ad alcuni problemi scientifici in campo astronomico, come quello di determinare e rappresentare matematicamente il moto lunare.

Anche se l'applicazione della Statistica al campo dei fenomeni sociali era stata adombrata negli scritti di Bernoulli e di Laplace, un passo decisivo in questa direzione fu compiuto grazie ai contributi di Adolphe Quetelet (1796-1874) e, soprattutto, di Francis Galton (1822-1911).

Quetelet fu un uomo dai molteplici interessi: matematico come formazione, si occupò di astronomia, di fisica, di meteorologia e di sociologia. Fu anche un grande organizzatore: organizzò le statistiche ufficiali del Belgio e fondò numerose associazioni scientifiche, in patria e all'estero. Per quanto riguarda l'analisi dei dati sociali, il contributo fondamentale dello studioso fu la formulazione del concetto di uomo medio. Nello studio dei dati sulla popolazione, esaminò una molteplicità di possibili relazioni tra fenomeni attraverso la costruzione di tabelle e grafici. Esaminò tassi di natalità e di mortalità per mese, per città, per temperatura e per ora del giorno. Studiò i caratteri antropometrici, peso statura, ecc., ma anche le qualità morali degli individui, tramite le statistiche su alcolismo, suicidi, malattie mentali, ecc. La finalità di queste ricerche era quella di scoprire le leggi che regolano la società umana, similmente a quanto gli astronomi avevano fatto per le leggi dell'universo nel secolo precedente.

L'idea di uomo medio fu probabilmente concepita nel sintetizzare i dati antropometrici allo scopo di effettuare confronti tra gruppi di persone. Ad esempio, disponendo dei dati sulla statura e il peso di un grande numero di coscritti francesi, la statura media ed il peso medio furono as-

[3] In realtà Gauss rivendicò nei suoi scritti di aver usato il metodo anteriormente al 1805, data della pubblicazione di Legendre contenente l'esposizione del metodo. In ogni caso, Gauss seguì un'impostazione del problema in termini probabilistici che lo portò alla riscoperta della curva normale come distribuzione di probabilità degli errori accidentali.

sunte come statura e peso del "coscritto medio" francese. La stessa cosa fu fatta per un gruppo numeroso di coscritti belgi. Ciò allo scopo di effettuare confronti tra i due gruppi. Questo approccio, potenzialmente applicabile a qualsiasi caratteristica misurabile, doveva servire, nell'idea dello studioso, per effettuare confronti nel tempo, nello spazio e per categorie di persone: l'uomo medio era un espediente per neutralizzare le variazioni casuali presenti a livello individuale e far emergere le regolarità, le leggi che regolano la società.

Un secondo filone di ricerca fu l'applicazione della curva normale allo studio dei dati antropometrici come strumento per stabilire se la popolazione studiata presentasse omogeneità, condizione necessaria per poter effettuare confronti tramite le medie. Il ragionamento era il seguente: se un insieme di misure di una data variabile sono omogenee (nel senso che sono influenzate da fattori comuni dominanti, differendo soltanto per cause accidentali), i dati si devono disporre secondo la curva degli errori accidentali.

Va osservato che l'importanza dell'opera di Quetelet sta, più che nelle indicazioni metodologiche, nell'impulso che diede all'applicazione dei modelli probabilistici nello studio dei fenomeni sociali.

Si deve, senza dubbio, a Galton una vera e propria svolta per la costruzione di una metodologia empirica e concettuale nello studio dei fenomeni sociali, metodologia che presenta i caratteri della Statistica moderna. Galton studiò medicina a Cambridge, ma, grazie ad una ricca eredità, poté abbandonare la carriera di medico. Ebbe molti interessi: esploratore in Africa per due anni, si occupò, successivamente, di meteorologia, psicologia, antropologia e sociologia. Tra i risultati del Quetelet, fece tesoro, in particolare, della tecnica di rappresentare con la curva normale i dati relativi a casi omogenei (come gli individui di uno stesso gruppo etnico), ma le sue analisi furono molto più profonde: cercò, infatti, di spiegare i meccanismi che fanno sì che i dati tendano a disporsi nella forma campanulate tipica della curva normale. Si interrogò, in particolare, su come la trasmissione ereditaria dei caratteri da una generazione all'altra potesse essere compatibile con tale caratteristica dei dati. E fu proprio nello studio della trasmissione ereditaria dei caratteri che Galton diede il contributo più importante, l'introduzione del concetto e della tecnica della regressione. Osservò, infatti, dall'esame di numerosi dati sulle stature di padri e figli, che da padri più alti della media nascevano figli più alti della media, e che da padri più bassi della media nascevano figli più bassi della media; tuttavia, osservò che le stature dei figli si discostavano dalla media meno di quelle dei loro padri. Questo fenomeno fu chiamato "re-

gressione", intesa come tendenza della razza a tornare verso i valori medi.

Con Galton si approda al XX secolo, dove nel corso dei primi quattro decenni, si può dire, la Statistica assume una fisionomia molto prossima a quella di oggi. Due sono le figure preminenti di questo periodo Karl Pearson (1857-1936) e Ronald Alymer Fisher (1890-1962).

Il primo, inglese di nascita, ma fervente ammiratore della cultura tedesca e di Marx, in onore dei quali cambiò il suo nome da Charles a Karl, fu fautore dell'uso della probabilità nella generalizzazione dei risultati empirici. Introdusse una importantissima tecnica (nota come test del $\chi^2$) per la verifica dell'adattamento dei dati osservati ad un curva teorica. Introdusse un'ampia classe di curve asimmetriche, utili nella descrizione dei dati relativi a situazioni in cui non sussistono le condizioni per l'applicazione della curva normale. Pearson fu un grande organizzatore: fondò riviste scientifiche di Statistica e creò una scuola di grande rilievo. Tra i suoi allievi vanno menzionati: Gorge Udny Yule (1871-1951) e Jerzy Neyman (1894-1981).

Fisher è stato, forse, la figura di statistico più eminente del secolo scorso: i suoi contributi permeano gran parte della Statistica odierna che si occupa dell'analisi dei dati di natura sperimentale (i dati provenienti da esperimenti controllati condotti negli studi di medicina, biometria, psicologia, ecc.). Fondamentale è stato anche il suo apporto per la definizione e la costruzione dei piani degli esperimenti.

La rassegna precedente è sicuramente approssimativa e lacunosa: ci si è limitati a tratteggiare le figure degli studiosi che hanno avuto un maggiore influsso sullo sviluppo del metodo statistico, rinunciando alla costruzione di un organico quadro storico, per il quale non si dispone né di spazio né di competenze. Per ogni ulteriore approfondimento si rinvia a Leti (1983) e Stigler (1986).

## 1.3. La Statistica nelle scienze empiriche

Come è emerso nella panoramica sullo sviluppo storico del metodo statistico, la Statistica è nata proprio come mezzo per stabilire se, in una determinata situazione, l'evidenza empirica, i dati osservati, sono in accordo con un'ipotesi o una teoria scientifica. In questo senso Laplace, Gauss ed altri uomini di scienza operarono nell'ambito dell'astronomia. Questo stesso paradigma venne applicato nelle scienze sociali. Come esempio emblematico, si ricordi il fenomeno dell'eccedenza delle nascite maschili (cfr. Nota 2) trattato da Laplace: era stata accertata, come evidenza empi-

rica, un'eccedenza di nascite maschili (il fenomeno era stato messo in luce da Graunt nel 1662; cfr. Nota 1); nell'analisi di Laplace, questa osservazione venne assunta come ipotesi da verificare; l'ipotesi venne confrontata con i dati reali relativi alle nascite registrate a Parigi in un certo periodo; l'ipotesi venne accettata in quanto era altamente probabile, alla luce dei dati, che il tasso delle nascite maschili fosse maggiore di 0,5.

Dunque nelle scienze empiriche, intendendo con ciò sia le scienze naturali (astronomia, fisica, biologia, ecc.) che le scienze sociali (economia, psicologia, sociologia, ecc.), la ricerca scientifica procede, almeno idealmente, secondo le seguenti fasi:

1. osservazione preliminare, con cui il ricercatore, concentrandosi su una sezione specifica del mondo reale, si pone una domanda del tipo "come funziona?"
2. formulazione di un'ipotesi sul comportamento del fenomeno studiato;
3. programmazione e realizzazione dell'attività di acquisizione di dati empirici pertinenti;
4. verifica dell'ipotesi.

La Statistica è uno strumento indispensabile nelle fasi 3 e 4. In particolare, nella fase 3, presiede alla appropriata programmazione degli esperimenti o delle osservazioni; nella fase 4, viene impiegata per mettere a confronto l'ipotesi con i dati osservati, tramite i procedimenti dell'*inferenza statistica*.

Va precisato che nella fase 2, la Statistica si limita a fornire i criteri e gli schemi con cui acquisire i dati: la selezione dei dati utili ai fini della ricerca è compito della disciplina specifica in cui si opera (astronomia, economia, ecc.). Sono pure di pertinenza della disciplina, le conclusioni sostanziali da trarre dalla verifica dell'ipotesi.

## 1.4. La Statistica nelle attività operative e nella vita quotidiana

Seguendo Frosini (1995), per attività operative si intendono tutte le attività pratiche finalizzate al raggiungimento di uno scopo. Quando tali attività presentano una certa complessità, la decisione di svolgimento dell'attività viene presa seguendo, almeno idealmente, le fasi appresso indicate:

1. individuazione e descrizione dello scopo dell'attività;

2. analisi della situazione di partenza;
3. esame dei mezzi e degli strumenti da utilizzare;
4. previsione dei risultati conseguibili;
5. decisione finale.

È facile comprendere come, in certi casi, nelle fasi 2 e 4, possa essere utile l'impiego della Statistica.

Si pensi, ad esempio, alla decisione da parte di una catena di supermercati di aprire un nuovo punto vendita. Nella fase 2, la Statistica si rende necessaria per il reperimento e l'analisi dei dati concernenti gli esercizi concorrenti localizzati nell'area prescelta, per il reperimento e l'analisi delle informazioni sulla clientela potenziale per poter adattare l'offerta alle caratteristiche della clientela stessa, ecc. Nella fase 4, la Statistica può intervenire con le tecniche di analisi di mercato volte a prevedere il comportamento di acquisto del consumatore; a questo fine, può essere utile effettuare un indagine sul campo con apposite interviste ai potenziali clienti.

Come altro esempio, si pensi ad un provvedimento governativo volto ad incentivare il mercato dell'auto. Un comportamento razionale e rigoroso richiede l'intervento della Statistica nelle due fasi anzidette. Nella fase 2, per analizzare le caratteristiche del parco auto potenzialmente interessato al provvedimento. Per il reperimento di questi dati, non si deve ricorrere ad un'indagine *ad hoc* potendo utilizzare i dati del registro automobilistico. Nella fase 4, la Statistica consente di valutare il costo dell'operazione con l'impiego di modelli per prevedere il numero e le caratteristiche dei fruitori del provvedimento.

E ancora, l'operatore di borsa che deve decidere un'operazione finanziaria impiegherà la Statistica, prima di tutto, per la valutazione dello stato dell'arte nel mercato dei titoli, poi, per prevedere il futuro andamento dei titoli che intende trattare.

In definitiva, in ogni attività operativa di qualche rilevanza e complessità, la Statistica è di ausilio per l'assunzione di decisioni coscienti e responsabili.

La Statistica nella ricerca scientifica e nelle attività operative più complesse svolge il suo compito mettendo a disposizione il suo apparato fatto di metodi e procedure talvolta molto complessi. Ma occorre riflettere che anche nella vita di tutti i giorni la Statistica svolge una funzione importante: certo, all'individuo, che pure si trova frequentemente a prendere decisioni in condizioni d'incertezza, non si richiede di impiegare gli

strumenti dell'inferenza statistica, si richiede, invece, di essere in grado di comprendere ed interpretare le informazioni di natura quantitativa che sempre più copiosamente i mezzi di comunicazione diffondono: si tratti di grafici, tabelle, statistiche sintetiche sull'inflazione, sull'occupazione, sugli incidenti stradali, ecc. Non solo: sarebbe auspicabile che l'individuo sia in grado di esprimersi, di comunicare, oralmente e per iscritto, informazioni quantitative ricorrendo agli strumenti della Statistica (grafici, tabelle, indici sintetici, ecc.). Perché quello della statistica è un vero e proprio linguaggio, con la sua sintassi, che amplia fortemente le capacità di comunicare e interpretare; come tale, dovrebbe far parte del bagaglio culturale di base. Con un neologismo inglese, si chiama *numeracy* questa abilità, che come la *literacy* (saper leggere e scrivere), ormai conquistata da tutti, è un requisito indispensabile per il cittadino moderno.

## 1.5. Cenni sulle fonti statistiche

In Italia, alla produzione e diffusione delle statistiche relative a tutti gli aspetti della vita del paese è preposto l'ISTAT, l'Istituto nazionale di statistica. Si tratta di un ente di diritto pubblico con ordinamento autonomo, sottoposto alla vigilanza della Presidenza del Consiglio. È posto al vertice di un sistema organizzativo, il sistema statistico nazionale (SISTAN), di cui fanno parte, oltre all'ISTAT, gli uffici di statistica centrali e periferici delle amministrazioni dello Stato, quelli delle Regioni, delle Province e dei Comuni, delle Unità sanitarie locali, delle Camere di commercio e tutti gli uffici di statistica comunque denominati, di enti ed amministrazioni pubbliche.

Sono molteplici le statistiche che l'ISTAT mette a disposizione sotto forma di pubblicazioni mensili, annuali o con altra periodicità (ad esempio, indagini speciali e censimenti). Fra le pubblicazioni periodiche si segnalano, in particolare:

- il Bollettino Mensile di Statistica;
- l'Annuario Statistico Italiano, pubblicazione che contiene le statistiche relative ai principali fenomeni demografici, sociali ed economici;
- l'annuario di Statistiche Demografiche;
- l'annuario di Statistiche Sanitarie;
- l'annuario di Statistiche del Lavoro;
- l'annuario di Statistiche Industriali;
- l'annuario di Contabilità Nazionale.

L'ISTAT effettua con periodicità decennale il censimento della popolazione, il censimento dell'industria, del commercio, dei servizi e dell'artigianato, e il censimento dell'agricoltura. I risultati delle rilevazioni censuarie sono raccolti in volumi riferiti ai livelli nazionale, regionale e provinciale.

Molti altri enti, pubblici o privati, producono e pubblicano statistiche settoriali. Si pensi alla Banca d'Italia (per i fenomeni economici, finanziari e monetari), ai diversi ministeri, ad alcuni enti pubblici, come l'ACI, l'INPS, l'INAIL, ecc.

A livello internazionale, si segnalano le pubblicazioni dell'ONU (si ricordano, particolare, lo *Statistical Yearbook*, il *Demographic Yearbook*, lo *Yearbook of International Trade Statistics*), quelle dell'UNESCO, della CEE e del FMI (Fondo Monetario Internazionale).

## 1.6. Terminologia essenziale

Come è stato già detto, i numeri assumono la qualità di statistiche quando sottendono un insieme di riferimento; da qui l'affermazione essenziale che "i dati statistici sono numeri in un contesto". In questo senso, la temperatura osservata in una certa stazione meteorologica alle ore 13 del 21 giugno non è un dato statistico; diviene tale se è considerata nel quadro delle temperature rilevate nello stesso istante in altre stazioni meteorologiche della regione o del paese, perché in tal caso i dati consentono di effettuare valutazioni e confronti di qualche interesse. E ancora: il livello di colesterolo nel sangue di una persona, quale risulta dalle analisi cliniche, non è un dato statistico; lo è se fa parte di un insieme di osservazioni su soggetti che si trovano in condizioni di salute simili che sono sottoposte a qualche terapia: potrebbe interessare, infatti, studiare l'effetto della terapia sul colesterolo.

In definitiva, i dati assumono la veste di statistiche se sono il risultato dell'osservazione intenzionale di una molteplicità di casi individuali finalizzata alla conoscenza e/o alla comprensione del fenomeno oggetto di studio. La molteplicità dei casi individuali, o, come anche si è detto, l'insieme di riferimento, va sotto il nome di *collettivo statistico*.

Nello studio della dimensione aziendale (ad esempio, in termini di addetti), il collettivo statistico è l'insieme delle aziende esaminate. Nelle prove di collaudo di una bilancia di precisione, il collettivo statistico è l'insieme delle misure effettuate (pesate di un oggetto di peso noto). Nel

controllo di un processo produttivo, il collettivo statistico è costituito dal campione di pezzi prodotti di cui vengono esaminate le caratteristiche.

**Definizione 1.2.** *Si chiama unità statistica il caso individuale componente del collettivo statistico*.

Negli esempi precedenti, l'unità statistica è, nell'ordine, l'azienda, la singola ripetizione dell'operazione di pesatura, ciascun pezzo osservato.

**Definizione 1.3.** *Si chiama carattere ogni aspetto elementare oggetto di rilevazione nelle unità statistiche del collettivo*.

Così, nell'esempio dello studio della dimensione aziendale, il carattere è il numero di addetti; nel caso del collaudo della bilancia, il carattere è il peso dell'oggetto considerato; nel controllo di processo il carattere è la caratteristica (peso, lunghezza, diametro, ecc.) sulla quale verte il controllo.

Nella realtà, la rilevazione riguarda sempre una pluralità di caratteri che possono essere studiati singolarmente o in un modo congiunto (studio delle relazioni). Nelle pagine che seguono, ciò sarà sottinteso: si parlerà di carattere al singolare per intendere un generico carattere tra quelli considerati in una rilevazione statistica.

**Definizione 1.4.** *Si chiamano modalità di un carattere i diversi modi con cui questo si presenta nelle unità statistiche del collettivo*.

Nella rilevazione sull'occupazione, le modalità del carattere "stato occupazionale" sono: "occupato", "disoccupato", "persona in cerca di prima occupazione". Nel censimento della popolazione le modalità del carattere "professione del capofamiglia" sono: "imprenditore", "libero professionista", "impiegato", ecc. In un'indagine sulle abitazioni, le modalità dei caratteri "numero di vani" e "superficie dell'abitazione" sono, nel primo caso, i numeri: 1, 2, 3, ecc., nel secondo caso, tutti i valori di un intervallo di numeri reali.

I caratteri sono di due tipi: *qualitativi* e *quantitativi*. I primi hanno modalità costituite da espressioni verbali; i secondi hanno modalità rappresentate da numeri. Le modalità del carattere qualitativo possono essere o non essere ordinabili. Sulla base di questa distinzione, si parla di *caratteri rettilinei*, con riferimento a quelli le cui modalità sono ordinabili, e di *caratteri sconnessi*, negli altri casi. È un carattere rettilineo il "grado" de-

gli ufficiali dell'esercito italiano; è sconnesso il carattere "professione" dei lavoratori.

Un carattere quantitativo si dice *discreto* se le sue modalità sono quantità distinte, preventivamente individuabili ed elencabili; si dice *continuo* quando esso può assumere tutti i valori di un certo intervallo di numeri reali. Un esempio di carattere discreto è il numero di vani delle abitazioni; un esempio di carattere continuo è la statura degli individui. Si osservi che quella di carattere continuo è una nozione astratta: nelle rilevazioni reali, poiché la precisione delle misurazioni è sempre limitata, il carattere continuo è trattato come se fosse discreto. Si pensi ad una rilevazione di stature; se l'approssimazione è al cm, l'osservazione produrrà un insieme di risultati preventivamente definibili ed elencabili: sono tutti i numeri interi compresi tra un minimo e un massimo; ciò è vero anche se la misurazione viene effettuata a livelli di precisione più elevati, al millimetro, al decimo di millimetro, ecc.

I caratteri quantitativi si distinguono, inoltre, in *trasferibili* e *non trasferibili* a seconda che abbia o non abbia senso ipotizzare il trasferimento di parte del carattere da una unità all'altra. Sono esempi di caratteri trasferibili il reddito e il patrimonio delle persone.

Una ulteriore distinzione tra i caratteri è determinata dalla loro relazione con il fattore tempo. Per alcuni caratteri, come la statura e il peso delle persone adulte, le opinioni dei cittadini su un determinato servizio, si verificano variazioni più o meno grandi nel tempo. Per essi il decorso del tempo è un *fattore di disturbo*; quindi bisognerebbe misurarli simultaneamente. Si parla al riguardo di *caratteri di stato*, o, con lo stesso significato, di *fenomeni di stato*.

Quando, invece, il decorso del tempo è un elemento indispensabile per la rilevazione del carattere, si parla di *caratteri di movimento* o di *fenomeni di movimento*. I caratteri "numero di nati", "numero di morti", "produzione dell'industria tessile in Italia" sono esempi di caratteri che non possono essere rilevati se non con riferimento ad un intervallo di tempo. Qui il decorso del tempo è elemento indispensabile per l'esistenza stessa del carattere. Tali caratteri non sono altro che il risultato di un conteggio di eventi che si verificano in un arco di tempo.

## 1.7. Misurazione dei caratteri

Con l'osservazione del carattere nella singola unità del collettivo, si effettua una misurazione: per i caratteri qualitativi la "misurazione" è in realtà la de-

scrizione verbale del carattere nell'unità osservata; per i caratteri discreti è, in genere, un conteggio; per i caratteri continui si tratta, invece, di una misurazione in senso proprio che presuppone una scala numerica.

*Caratteri qualitativi*

La misurazione di un carattere qualitativo consiste nel registrare la modalità che si presenta nella singola unità statistica. Le modalità possono essere predefinite ed inserite come tali nel questionario o nella scheda di rilevazione (come avviene, ad esempio, per il carattere "stato civile" nel censimento della popolazione; si veda al riguardo la domanda 3.1 del questionario del censimento 2001 riprodotto nel CD annesso al volume). Si opera in questo modo quando le situazioni individuate dalle varie modalità sono nettamente distinte e, quindi, non vi sono margini di incertezza nell'attribuzione della singola unità all'una o all'altra modalità. Quando questa condizione non sussiste, le modalità vengono desunte *a posteriori* dalla descrizione dettagliata che il rilevatore fa dello stato della singola unità relativamente al carattere in questione.

Si prenda come esempio il caso del carattere "professione" nel censimento della popolazione. Nel modello di rilevazione non vengono elencate le diverse modalità, vi sono, invece, due quesiti (quesiti 7.9 e 7.10), di cui il primo è volto ad inquadrare la posizione nella professione (dirigente, quadro, impiegato, ecc.), ed il secondo è diretto ad accertare il tipo di lavoro effettivamente svolto: le modalità del carattere "professione" vengono desunte *a posteriori* dall'esame congiunto delle risposte ai due quesiti. In ogni caso, le modalità devono rispondere ai requisiti della esaustività e della unicità, nel senso che ad ogni unità della popolazione possa essere attribuita una modalità e che questa modalità sia unica.

La determinazione delle modalità di un carattere qualitativo equivale, in un certo senso, alla costruzione di una scala di riferimento, una scala fatta di nomi o di espressioni verbali. In particolare, si parla di *scala nominale* per i caratteri sconnessi e di *scala ordinale* per i caratteri rettilinei. La scala nominale consente soltanto di classificare le unità del collettivo statistico in tanti gruppi distinti quante sono le modalità del carattere, gruppi che al loro interno presentano omogeneità (le unità appartenenti a ciascun gruppo sono "uguali" nel senso che presentano la stessa modalità). La scala ordinale consente, come nel caso precedente, la classificazione delle unità statistiche in gruppi omogenei; in più permette di "graduare" i gruppi in base all'ordine che le modalità presentano. Ad esempio, la scala del carattere "stato civile" è nominale ed è costituita dalle

modalità “celibe/nubile”, “coniugato/a”, “divorziato/a”, “vedovo/a”. La scala con cui si misura il carattere “grado di soddisfazione del cliente” è, invece, ordinale. Nell’ipotesi che le modalità siano “basso”, “medio”, “alto” e “molto alto”, la scala consente di classificare le unità del collettivo statistico in gruppi omogenei, e di graduare tali gruppi secondo il livello di soddisfazione.

*Caratteri discreti*

La misurazione si risolve, generalmente, in un conteggio. La scala con cui si misura un carattere discreto è la cosiddetta *scala proporzionale*. Le modalità del carattere sono espresse da numeri, talché esse consentono di classificare le unità statistiche del collettivo in gruppi omogenei (fatti di unità “uguali” per quanto riguarda il carattere esaminato), di graduare i gruppi secondo il valore della modalità e di misurare la differenza tra i gruppi tramite la differenza o il rapporto tra le modalità.

Così, se il carattere è il “numero dei componenti la famiglia” e le modalità sono i numeri 1, 2, 3, 4 e 5, non solo si può dire (come con la scala ordinale) che le famiglie di due componenti sono “più grandi” delle famiglie di un componente, ma anche che le prime hanno un componente in più rispetto alle seconde, o, in modo equivalente, che hanno ampiezza doppia.

*Caratteri continui*

La misurazione di un carattere continuo comporta necessariamente un’approssimazione dovuta al troncamento dei numeri: i limiti di precisione dello strumento di misura uniti ai limiti della capacità di lettura dell’utente impongono di usare un numero limitato di cifre decimali.

Se si misura una lunghezza, occorre preliminarmente stabilire se si vuole una precisione al centimetro, oppure al millimetro, oppure al decimillimetro, ecc. Il livello di precisione è, naturalmente, legato all’ordine di grandezza del fenomeno da rilevare; così, per la statura delle persone l’approssimazione al centimetro è ordinariamente sufficiente; non lo è, ovviamente, per la lunghezza dei petali di una specie di fiori, per la quale una precisione adeguata è il millimetro, se non il decimo di millimetro. È altresì evidente che dal livello di precisione che si vuole raggiungere dipende la scelta dello strumento più appropriato: il metro per le stature, il calibro per la lunghezza dei petali, ecc. Un discorso analogo vale per la misura dei pesi: la precisione può essere al chilogrammo, al grammo, ecc., a seconda dell’ordine di grandezza del fenomeno: quanto più piccoli

sono i numeri che esprimono l'intensità del fenomeno tanto più fine deve essere l'unità di misura da assumere.

Una volta scelto il livello di precisione, il rilevatore annoterà il risultato della misurazione, esprimendola nell'unità di misura prescelta ed effettuando l'appropriata approssimazione. Ad esempio, se il livello di precisione è il centimetro, l'annotazione 174 cm indica che è stato osservato un valore compreso nell'intervallo [173,5;174,5) ; l'annotazione comporta un errore rispetto alla misura effettiva. L'errore massimo che si commette è di mezzo centimetro e si realizza quando la misura effettiva è uguale a 173,5 oppure a 174,5. Se il livello di precisione è il millimetro, l'annotazione del rilevatore 174,6 cm indica che la lettura dello strumento ha dato un numero compreso nell'intervallo [174,55;174,65) ; in questo caso l'errore dovuto al troncamento sarà al massimo mezzo millimetro.

Ciò vale ogni volta che si procede alla misurazione di un carattere continuo: la misura che il rilevatore annota sottende sempre un intervallo di valori possibili, detto *intervallo di tolleranza*, e presenta una deviazione rispetto alla misura effettiva non superiore alla metà dell'unità che esprime il livello di precisione prescelto.

È opportuno segnalare un'eccezione alla regola indicata. Si consideri, ad esempio, l'età delle persone espressa in anni compiuti. L'annotazione 52 anni indicherà che l'età osservata è compresa nell'intervallo [52; 53), non già, come in precedenza, nell'intervallo [51,5; 52,5). Ciò perché l'unità di misura "anni compiuti" implica un'approssimazione solo per difetto: si prende, infatti, la parte intera del numero che esprime l'età reale (in anni e frazioni) del soggetto; così alle età effettive 52 e 2 mesi e 52 e 11 mesi corrisponderà lo stesso numero di anni compiuti, 52. Pertanto, l'errore massimo che si commette è vicino ad un anno. Questa osservazione va estesa, ovviamente, a tutti i casi in cui il carattere è il tempo e le modalità sono durate espresse dal numero di intervalli temporali (anni, mesi, ecc.) completamente trascorsi.

La scala con cui si misurano i caratteri continui è quella proporzionale con l'eccezione dei caratteri per i quali lo 0 è un numero convenzionale che non significa assenza del carattere. Ciò accade, ad esempio, per le scale termometriche (Celsius, Réamur e Fahrenheit); così, se le temperature di due stazioni meteorologiche $A$ e $B$, registrate alla stessa ora, sono 5° e 10° non è corretto affermare che in $B$ la temperatura è doppia rispetto a quella in $A$, mentre ha senso affermare che la differenza tra le due temperature è di 5°. In questi casi il confronto tra le modalità va effettuato con le differenze; da qui l'espressione *scala ad intervalli* che viene utilizzata con riferimento alla misurazione di tali caratteri.