Premessa

La *Statistica* è una disciplina che fornisce gli strumenti per analizzare i *fenomeni collettivi*, ovvero i fenomeni che si manifestano su un insieme di casi singoli, come le caratteristiche somatiche (peso, altezza, colore degli occhi e così via) di un gruppo di individui, la preferenza espressa dagli elettori per i candidati di una lista politica, il voto conseguito dagli studenti universitari ad un esame, la produzione, l'ammontare delle vendite, il reddito e tutte le grandezze economico-finanziarie di un Paese. Tale disciplina si suddivide in due aree: quella attinente alla *Statistica descrittiva* e quella relativa alla *Statistica inferenziale*.
La Statistica descrittiva fornisce gli strumenti per descrivere, sintetizzare numericamente, presentare e, quindi, interpretare le osservazioni relative a fenomeni collettivi. D'altra parte, le tecniche di Statistica inferenziale consentono di valutare, o meglio stimare, le caratteristiche di una *popolazione*, ovvero dell'insieme della totalità dei casi, esaminando solo le osservazioni relative ad una parte ridotta (sottoinsieme) della stessa, denominata *campione*.
I fondamenti della Statistica descrittiva saranno oggetto di trattazione del presente volume. Questo testo, infatti, è un valido supporto per il lettore che deve affrontare un corso di Statistica descrittiva e può essere adottato in tutte quelle aree dell'offerta formativa che, nel rispetto delle nuove direttive sui programmi universitari, prevedono l'acquisizione di competenze statistiche per l'analisi esplorativa dei dati. È indubbio, tuttavia, che i contenuti di questo volume possano rappresentare anche una guida utile allo studio autonomo dei fondamenti della Statistica descrittiva.
Nella stesura del testo, è stato prestato il massimo impegno al fine di rendere la disciplina comprensibile al lettore che possieda, come unico prerequisito, la conoscenza dell'Algebra a livello di scuola secondaria di secondo grado. Tutti i termini tecnici sono definiti mediante un linguaggio semplice ed i concetti fondamentali sono introdotti in maniera chiara, utilizzando talvolta illustrazioni esplicative. Inoltre, gli esempi e gli esercizi presentati, da un lato, favoriscono la comprensione dei contenuti teorici e dall'altro, chiariscono i diversi ambiti di applicazione della Statistica.
Il volume è organizzato in due parti, una teorica ed una applicativa. Ciascuna parte si compone di nove capitoli, nei quali vengono trattati aspetti teorici (prima parte) e aspetti pratici (seconda parte) riguardanti

- il formalismo statistico, le tabelle e le rappresentazioni grafiche;
- gli indici di posizione, di variabilità e di forma;
- i rapporti statistici ed i numeri indici;
- l'analisi della dipendenza e della interdipendenza tra due caratteri statistici;

- la curva normale e la disuguaglianza di Bienaymé-Chebyshev.

Nella parte teorica, ciascun capitolo si apre con una breve e significativa introduzione, mediante la quale vengono delineati, spesso ricorrendo ad esempi, gli obiettivi principali del capitolo, la sua organizzazione e gli aspetti più interessanti ed utili ai fini applicativi.
Il parallelismo stabilito tra i capitoli che compongono le due parti attribuisce al lettore interessato la facoltà di selezionare agevolmente gli esercizi su cui testare il livello di apprendimento raggiunto.
In conclusione, l'intero volume è stato concepito, senza alcuna presunzione di completezza, rispettando la convinzione che un corso introduttivo di Statistica debba fondamentalmente evidenziare i concetti essenziali e le potenzialità di tale disciplina, i cui metodi e strumenti trovano applicazione in diversi contesti. In tal modo, si ritiene che il lettore possa trarne la massima utilità, poiché, solo disponendo di un indispensabile bagaglio di base, potrà decidere se intraprendere un corso più avanzato, in relazione ai suoi interessi specifici.
Non resta, quindi, che augurare a tutti buon lavoro!

Ringraziamenti
È opportuno rivolgere un ringraziamento particolare alla Dr.ssa Sabrina Maggio per il prezioso contributo fornito nella stesura della parte teorica del volume, nella formulazione degli esercizi e nella cura dimostrata per gli aspetti grafici.

Teoria

Capitolo 1

Concetti introduttivi e formalismo

Per molto tempo, la *Statistica* è stata considerata una disciplina costituita da strumenti per rilevare e rappresentare dati in forma tabellare e grafica.
Oggi la Statistica si è arricchita di teorie e metodi che consentono di effettuare elaborazioni più complesse, rispetto alla semplice costruzione di tabelle e grafici, al fine di descrivere, sintetizzare ed interpretare le osservazioni relative ai *fenomeni collettivi*, ovvero ai fenomeni che si manifestano su un insieme di casi singoli. Sono esempi di fenomeni collettivi: il livello di disoccupazione di un Paese, l'ammontare delle vendite di un'azienda, il numero di posti letto delle strutture sanitarie presenti in una regione, la quotazione di borsa di un pacchetto azionario, le caratteristiche somatiche (peso, altezza, colore degli occhi e così via) di un gruppo di individui. Lo studio di tali fenomeni richiede il ricorso a metodologie e strumenti statistici più o meno complessi, a seconda degli obiettivi che si intendono raggiungere.
In questo capitolo, dopo una breve presentazione delle origini e dei campi di applicazione della Statistica, saranno descritte le principali fasi con cui si sviluppa un'indagine statistica. Inoltre, saranno definiti i concetti di *unità statistica*, *collettivo statistico*, *carattere*, *modalità* e *frequenza*, indispensabili per la comprensione degli strumenti statistici che saranno descritti nei capitoli successivi.

1.1 Cenni storici

La Statistica venne riconosciuta come disciplina autonoma nella seconda metà del secolo XVII. Nel corso degli anni, numerose sono state le definizioni attribuite al termine "statistica". Si pensi che nel 1935 lo studioso Walter Francis Willcox (1861-1964) pubblicò ben 124 definizioni differenti, proposte da vari autori tra il 1749 ed il 1934. Attualmente, la seguente definizione è ritenuta da molti autori tra le più sintetiche ed efficaci.

Definizione 1.1. Statistica. *La Statistica è una disciplina che si compone di metodologie e strumenti che consentono lo studio dei fenomeni collettivi, ovvero di particolari aspetti della realtà oggetto d'interesse.*

Dal punto di vista etimologico, le derivazioni più diffuse, attribuite al termine *statistica*, sono le seguenti:

- dal latino *statera*, *bilancia*, perché la statistica misurava l'insieme delle risorse economiche e le forze di un Paese;

- dal tedesco *stadt*, *città*, dal momento che i primi utilizzatori delle tecniche statistiche se ne servirono per misurare l'ammontare del reddito prodotto dalle città;

- dal latino *status*, *condizione*, perché la statistica riguardava lo studio di fenomeni che interessavano la collettività.

Tuttavia, è ampiamente documentato che tale termine sia di origine italiana e derivi da *Status*, *Stato*, che nel Medioevo aveva il significato di forma di governo e, successivamente, di società politicamente organizzata. Tale derivazione trova riscontro nella considerazione che la Statistica nasce come attività rivolta alla descrizione dei molteplici aspetti della vita di uno Stato.
Al fine di delineare alcuni passaggi importanti nella storia della Statistica, dalle sue origini fino ai tempi più recenti, è necessario evidenziare che, secondo alcuni studiosi, la Statistica nasce con l'*arte del contare*. Sulla base di tale concezione, è possibile individuare prime e rudimentali statistiche nei simboli grafici incisi, dalle antiche comunità, su pietre o travi di legno. Tali simboli rappresentano le più remote forme di conteggio e di rilevazione delle quantità di beni possedute dai membri di una comunità.
In particolare, risalgono alla civiltà dei Sumeri (IV-III millennio a.C.), in Mesopotamia, alcune tavolette d'argilla che riportano, mediante caratteri cuneiformi, elenchi di uomini e beni.
Nell'antico Egitto (III millennio a.C.), alcune testimonianze scritte rinvenute attestano un'accurata organizzazione amministrativa con regolari rilevazioni, sia del movimento della popolazione, che dell'ammontare dei beni posseduti.
Nell'Estremo Oriente, il primo censimento della popolazione e delle terre, citato in un libro di Confucio, fu ordinato dall'imperatore Yao nel II millennio a.C.
In uno dei primi libri della Bibbia, intitolato *I numeri*, sono documentate notizie relative al primo censimento del popolo di Israele disposto da Mosé nel deserto del Sinai, dopo la fuga dall'Egitto (1290 a.C.).
Tuttavia, i *census*, ordinati dal re Servio Tullio (578-534 a.C.) nell'antica Roma, possono essere considerati le prime ed importanti iniziative finalizzate e programmate di raccolta dei dati. Essi consistevano nella rilevazione quinquennale della numerosità della popolazione, con indicazione del numero delle nascite e del numero dei decessi, e del reddito pro capite dei cittadini.
La Statistica è nata, quindi, come attività di carattere amministrativo e fiscale ed ha continuato ad essere tale anche in epoche più recenti. Risalgono, infatti, al 1300 i primi registri parrocchiali dei matrimoni, delle nascite e dei decessi dell'epoca, attualmente considerati le prime fonti di dati per lo studio del movimento naturale della popolazione.
Nel XVII secolo, ad opera di Hermann Conring (1606-1681), la Statistica venne riconosciuta come una disciplina autonoma rivolta alla raccolta di informazioni riguardanti i molteplici aspetti della vita di uno Stato.
H. Conring, studioso di grande cultura e docente di diritto pubblico presso l'Università di Helmstädt, in Germania, istituì nel 1660 un corso universitario di Scienze Politiche, in cui venivano esposte, in forma descrittiva, le informazioni concernenti le cose notevoli dello Stato (eventi storici, aspetti geografici, fattori socio-economici). Qualche decennio più tardi, Gottfried Achenwall (1719-1772), professore presso l'Università di Gottinga in Germania,

diede al suddetto corso il nome Statistica (*Statistik*, nella terminologia tedesca). Da allora, furono istituite in Germania numerose cattedre di Statistica, per cui fu attribuita la denominazione di *Statistica universitaria*.
Un altro importante indirizzo di studi, che contribuì allo sviluppo della Statistica, ebbe luogo in Inghilterra nella seconda metà del 1600 ad opera di John Graunt (1620-1674). Quest'ultimo si occupò dello studio di fenomeni demografici, come l'indice di mortalità di una popolazione, la numerosità degli abitanti di una regione, il tasso di urbanizzazione delle città, per la ricerca di leggi generali in grado di fornire informazioni più ampie riguardanti il movimento naturale e quello migratorio della popolazione in quel periodo. J. Graunt utilizzò i dati demografici ricavati dai registri parrocchiali battesimali e mortuari istituiti dalla Chiesa d'Inghilterra, per riscontrare regolarità scientifiche, come l'inurbamento delle popolazioni rurali e l'eccedenza delle nascite maschili su quelle femminili. Nel 1662, J. Graunt pubblicò a Londra l'opera *Osservazioni naturali e politiche*, ritenuta da alcuni autori il primo effettivo studio condotto con metodi statistici. Per tale ragione, J. Graunt viene considerato uno dei padri della dottrina denominata *Aritmetica Politica*, ovvero l'arte di ragionare per mezzo di cifre su aspetti attinenti con il governo di uno Stato.
Un'altra radice storica della Statistica è collegata al calcolo delle probabilità, di cui si occuparono nel corso del Seicento diversi celebri studiosi, quali Blaise Pascal (1623-1662), Pierre Fermat (1608-1665) e Jacob Bernoulli (1654-1705). Quest'ultimo, con l'opera dal titolo *Ars conjectandi*, ovvero l'*Arte di fare congetture*, affermò l'importanza della teoria della probabilità nello studio dei fenomeni naturali e sociali.
Nel corso del Settecento, la Statistica proseguì ricalcando gli indirizzi già delineati nel secolo precedente, evidenziando così l'importanza della teoria della probabilità al fine di determinare, mediante l'esame di un numero limitato di eventi elementari, leggi generali che, in termini probabilistici, regolano i suddetti eventi.
Nell'Ottocento, due grandi studiosi, Karl Friedrich Gauss (1777-1855) e Pierre Simon Laplace (1749-1827), fornirono notevoli ed originali contributi allo sviluppo della Statistica. Il primo, teorizzò la distribuzione degli errori accidentali e definì una importante distribuzione di probabilità, nota come *curva di Gauss* o *curva normale*. D'altra parte, P. S. Laplace scrisse il primo importante trattato generale sul calcolo delle probabilità, in cui furono enunciati e dimostrati numerosi teoremi di fondamentale rilevanza per la cosiddetta *Statistica induttiva* o *inferenziale*. Quest'ultima, come sarà descritto successivamente, riguarda l'insieme dei metodi statistici che, a partire dall'informazione contenuta in un campione, consentono di trarre conclusioni sulla popolazione dalla quale il campione è stato estratto, esprimendo in termini probabilistici la validità di tali conclusioni.
Uno dei principali sostenitori della funzione induttiva della Statistica, fu il grande statistico inglese Ronald Alymer Fisher (1890-1962), particolarmente noto per aver sviluppato la tecnica statistica, denominata *analisi della varianza*.
Negli stessi anni, in Italia, la Statistica si concentrava ancora sull'approccio metodico-descrittivo. Fra gli studiosi italiani, è doveroso annoverare Corrado Gini (1884-1965). Quest'ultimo fu il promotore di una Scuola di Statistica; fondò nel 1920 la rivista internazionale di statistica *Metron*; istituì nel 1926 l'Istituto Centrale di Statistica e nel 1936 la prima Facoltà di Scienze Statistiche, Demografiche ed Attuariali. A C. Gini si deve l'introduzione di un indice statistico molto utilizzato come misura di variabilità e lo sviluppo dei concetti di concentrazione, dissomiglianza, connessione e dei relativi indici.

1.2 Campi di applicazione della Statistica

Gli sviluppi storici appena delineati, evidenziano che la Statistica è una disciplina che si compone di teorie, metodi e tecniche per l'analisi dei dati raccolti mediante l'osservazione degli aspetti oggetto d'interesse. I problemi che si affrontano possono riguardare diversi settori, quali l'industria, l'economia, l'ambiente, la medicina, la sociologia, e così via.
Il diffuso utilizzo degli strumenti statistici in molti campi della scienza ha contribuito, da un lato, allo sviluppo di discipline autonome, quali la *Demografia*, per lo studio della popolazione, l'*Analisi delle serie storiche* e l'*Analisi statistica spaziale*, per lo studio di fenomeni che evolvono, rispettivamente, nel tempo e nello spazio, dall'altro, alla nascita delle cosiddette *Statistiche applicate*, tra le quali:

- la *Statistica sociale*, per l'analisi dei fenomeni sociali, quali il livello di scolarizzazione di una comunità, la qualità dei servizi offerti da un'amministrazione pubblica, le cause del disagio giovanile di un quartiere;
- la *Statistica sanitaria*, per l'analisi dei fenomeni sanitari, quali la mortalità per particolari malattie, la spesa per attrezzature e servizi sanitari, i costi per l'assistenza domiciliare;
- la *Statistica aziendale*, per lo studio delle dinamiche aziendali, quali i flussi finanziari ed economici, gli indici di bilancio, gli stipendi degli impiegati;
- la *Statistica giudiziaria*, per lo studio dei fenomeni connessi all'attività della magistratura, quali le condanne per un determinato reato, i divorzi, le adozioni;
- la *Statistica economica*, per l'analisi degli aspetti economici di un Paese, quali prezzi, consumi, produzione, inflazione in un determinato periodo;
- la *Statistica ambientale*, per lo studio dei fenomeni ambientali, quali l'inquinamento, le risorse idriche sotterranee, i giacimenti minerari.

Tuttavia, le metodologie statistiche sviluppate in letteratura si suddividono fondamentalmente in due branche, denominate *Statistica descrittiva* e *Statistica inferenziale*.

Definizione 1.2. Statistica descrittiva. *La Statistica descrittiva è una disciplina che si compone di metodologie e strumenti che consentono di rappresentare, sintetizzare ed interpretare le osservazioni relative ad uno o più aspetti di un determinato fenomeno.*

Spesso, l'osservazione di un fenomeno produce una gran mole di dati, per cui risulta indispensabile l'utilizzo delle tecniche di Statistica descrittiva, che consentono di evidenziare importanti caratteristiche presenti nei dati mediante opportune rappresentazioni tabellari e grafiche ed alcuni indici sintetici.

Definizione 1.3. Statistica inferenziale. *La Statistica inferenziale è una disciplina che si compone di risultati teorici fondamentali ed appropriate metodologie che consentono di utilizzare le osservazioni relative ad un campione, allo scopo di giungere a conclusioni valide per la popolazione di riferimento.*

Negli studi di Statistica inferenziale, svolge un ruolo fondamentale la *teoria della probabilità*, la quale consente di valutare i limiti di validità entro cui è possibile estendere alla popolazione i risultati ottenuti sul campione osservato.
Per gli obiettivi di questo testo, nel seguito saranno fornite esclusivamente le nozioni di campione e popolazione e saranno descritte le tecniche più utilizzate per la selezione di un campione; quindi, per lo studio della Statistica inferenziale si rinvia il lettore a testi specialistici.

1.3 Indagine statistica

Come premesso, la Statistica descrittiva offre metodologie e strumenti per valutare ed interpretare particolari aspetti della realtà oggetto d'interesse, quali, ad esempio, il reddito della popolazione residente in una regione, la temperatura media giornaliera registrata in un mese in una località, la lunghezza delle barre di ferro prodotte da una macchina industriale.
A tale scopo, ciascun aspetto oggetto di studio viene osservato su un insieme di unità elementari, denominate *unità statistiche*.

Definizione 1.4. Unità statistica. *I singoli elementi che sono oggetto di rilevazione sono denominati unità statistiche* (*u.s.*).

Riprendendo gli esempi precedenti, per la rilevazione del reddito della popolazione residente in una regione, l'*u.s.* oggetto di osservazione è rappresentata dal singolo individuo residente in quella regione; oppure, nel caso della lunghezza delle barre di ferro prodotte da una macchina industriale, ogni singola barra prodotta costituisce l'*u.s.* su cui misurare la lunghezza.
In generale, le *u.s.* si distinguono in

- **semplici**, se formate da un unico elemento non scomponibile ulteriormente. Tali unità possono essere rappresentate da
  - persone o esseri viventi, come ad esempio gli individui di un determinato territorio, i laureati di una facoltà universitaria, gli animali di una determinata specie;
  - oggetti, come ad esempio le merci, i libri, i reperti archeologici, i pezzi prodotti da una macchina industriale;
- **composte**, se costituite da più unità semplici; sono esempi di *u.s.* composte le famiglie residenti in un comune, le aziende di un determinato settore economico, i *club* culturali presenti in una regione.

Definizione 1.5. Collettivo statistico. *L'insieme delle u.s. oggetto di osservazione viene denominato collettivo statistico.*

Ai fini dell'applicazione del metodo statistico, è importante individuare correttamente il collettivo statistico, il quale può coincidere alternativamente con

- la *popolazione*, ovvero l'insieme di tutte le *u.s.* su cui la caratteristica in esame si manifesta; ad esempio, la popolazione delle auto immatricolate in Italia in un anno, la popolazione di tutte le aziende registrate ed operanti in una regione.
  Si osservi che il termine popolazione non si riferisce esclusivamente ad una popolazione reale, effettivamente esistente e visibile, ma anche ad una popolazione virtuale; si pensi, ad esempio, alla popolazione di tutte le cinquine che possono essere estratte su una ruota nel gioco del Lotto;

- il *campione*, ovvero un sottoinsieme di *u.s.* appartenenti alla popolazione. La selezione delle *u.s.* che costituiscono il campione avviene attraverso particolari tecniche di campionamento, come sarà descritto nel § 1.5.

Nel seguito si utilizzerà il termine "collettivo statistico" con riferimento all'insieme delle *u.s.* sulle quali sono state osservate una o più caratteristiche, senza specificare ulteriormente se tale insieme rappresenta tutta la popolazione o soltanto un suo campione.

1.3.1 Fasi di un'indagine statistica

La ricerca scientifica, condotta con metodi statistici, si basa su alcune importanti fasi che devono essere opportunamente pianificate, al fine di giungere a risultati attendibili, come descritto di seguito.

Definizione degli obiettivi
In questa prima fase dell'indagine statistica devono essere definite le aree tematiche della ricerca, i mezzi e le risorse umane disponibili, nonché gli obiettivi che si intendono raggiungere con l'indagine. Questi ultimi devono essere opportunamente dettagliati al fine di evitare equivoci o ambiguità. Essi possono riguardare la semplice costruzione di una base di dati che raccoglie informazioni sugli aspetti più importanti di un determinato fenomeno, oppure possono essere più complessi e mirare all'individuazione di eventuali relazioni tra due o più aspetti caratteristici del fenomeno in esame.
Questa fase è la più delicata, perché, se non viene curata nei dettagli, può pregiudicare i risultati dell'intera indagine. Ad esempio, per effettuare un'indagine sui consumi dei giovani occorre definire, in maniera puntuale, la fascia d'età dei giovani da intervistare, la tipologia dei consumi oggetto di studio, l'intervallo temporale di riferimento, il territorio su cui si deve svolgere l'indagine.
In generale, dopo aver delineato gli obiettivi della ricerca, occorre precisare:

- il collettivo statistico su cui effettuare l'indagine e le caratteristiche da rilevare;
- il metodo di rilevazione dei dati;
- l'ambito spaziale e temporale dell'indagine;
- le risorse umane necessarie al reperimento dei dati;
- gli strumenti utili alla rilevazione ed elaborazione dei dati;
- i tempi ed i costi di rilevazione ed elaborazione dei dati;
- i mezzi per la diffusione dei risultati.

Rilevazione
In questa fase, vengono rilevati tutti i dati utili al raggiungimento degli obiettivi precedentemente definiti.
A seconda del settore scientifico in oggetto, i dati possono essere acquisiti in modi differenti, ovvero

- direttamente, attraverso appositi questionari sottoposti alle unità oggetto di rilevazione;
- indirettamente, mediante fonti ufficiali preposte alla raccolta periodica di dati di varia natura (Camere di Commercio, associazioni di categoria, province, comuni, aziende ospedaliere e così via);
- sperimentalmente, attraverso *test* psicologici, analisi cliniche, esperimenti effettuati in laboratorio.

Si osservi che i modi con cui vengono rilevati i dati, così come gli errori causati da imperfezioni nel processo di rilevazione, influenzano l'accuratezza dell'informazione, con evidenti conseguenze sulla qualità dell'indagine.
Uno degli strumenti più utilizzati per la rilevazione dei dati, soprattutto nell'ambito di indagini sociali, economiche, di *customer satisfaction*, è il *questionario*. Esso si compone di domande, solitamente raggruppate in differenti sezioni, ognuna delle quali identifica uno specifico argomento d'interesse per l'indagine che si sta conducendo.
Per gli obiettivi del presente testo, non vengono discusse le numerose problematiche inerenti la stesura di un questionario, per le quali si rinvia il lettore ai testi citati nella bibliografia. Tuttavia, si ritiene indispensabile evidenziare che nella realizzazione di un questionario devono essere dettagliatamente curate le seguenti operazioni:

1. concettualizzazione e definizione degli obiettivi, ovvero indicazione accurata dei temi interessanti per l'indagine e delle informazioni che si intende acquisire mediante la somministrazione del questionario;
2. redazione del questionario, ovvero specificazione delle sezioni in cui lo stesso si articola e della sequenza di domande all'interno di ciascuna sezione, nonché formulazione chiara dei quesiti e delle risposte, nel caso di domande a risposta multipla;
3. verifica del questionario, ovvero valutazione della completezza delle domande e della conformità del questionario alle esigenze dell'indagine.

Le rilevazioni statistiche si distinguono in:

- **totali**, se viene esaminato l'insieme, denominato popolazione, di tutti gli elementi su cui il fenomeno si manifesta;
- **parziali**, se viene esaminato solo un sottoinsieme, denominato campione, della popolazione di riferimento.

Il censimento della popolazione residente in un Paese è un classico esempio di rilevazione totale. In questi casi, considerata la complessità della rilevazione, l'analisi viene condotta congiuntamente su diversi aspetti, in modo da ottenere una molteplicità di informazioni sulla

popolazione di riferimento. D'altra parte, nelle rilevazioni parziali occorre definire opportunamente il campione su cui effettuare l'analisi, in modo che i risultati ottenuti possano essere estesi, con una certa approssimazione, alla popolazione da cui quel campione è stato estratto.

Elaborazione
Questa fase è sicuramente la più complessa. Infatti, conclusa la fase della rilevazione, spesso si dispone di una gran mole di dati. Questi ultimi, denominati *dati grezzi*, devono essere selezionati e predisposti opportunamente per l'elaborazione con metodi e strumenti statistici. Dall'elaborazione di dati grezzi si ottengono i cosiddetti *dati derivati*.
I metodi di analisi utilizzati in questa fase dipendono dagli obiettivi da raggiungere. In generale,

- se occorre sintetizzare alcune importanti caratteristiche presenti nei dati rilevati, la metodologia utilizzata rientra nell'ambito della Statistica descrittiva;
- se, come spesso accade nelle rilevazioni parziali, i dati vengono elaborati al fine di ottenere risultati che, sotto determinate condizioni, possano essere estesi alla popolazione di riferimento, allora le metodologie da adottare rientrano nell'ambito della Statistica inferenziale.

Presentazione dei risultati
I risultati raggiunti mediante l'analisi statistica devono essere opportunamente presentati in maniera organica, mediante tabelle, grafici ed indici sintetici. Si osservi che spesso l'efficacia del metodo statistico, utilizzato nella fase dell'elaborazione, dipende anche dal modo con cui i risultati ottenuti vengono presentati ai destinatari della ricerca, quali i lettori di una rivista, gli specialisti del settore, i ricercatori, gli organi di controllo.

Interpretazione dei risultati
Tale fase consiste nell'esaminare i risultati dell'indagine e verificare se gli obiettivi prefissati sono stati raggiunti. I risultati ottenuti sono interpretati anche alla luce delle conoscenze acquisite indirettamente sul fenomeno in esame oppure delle indicazioni fornite dagli esperti di settore, nonché delle eventuali relazioni che possono esistere con altri fenomeni.

In Fig. 1.1 vengono schematizzate le fasi appena descritte.

1.4 Fonti di rilevazione statistica

Le rilevazioni statistiche possono essere eseguite da

- *privati*, se la raccolta dei dati viene effettuata da una persona fisica o da un ente privato (ad esempio, l'istituto italiano *DOXA* specializzato in sondaggi);
- *enti pubblici*, se il fenomeno oggetto di indagine è di rilevante interesse pubblico (ad esempio, i fenomeni demografici o economici) e richiede notevoli mezzi finanziari ed un'organizzazione capillare.

Gli organismi pubblici, che istituzionalmente raccolgono e diffondono informazioni statistiche, sono innumerevoli ed agiscono secondo una gerarchia di competenze che indivi-

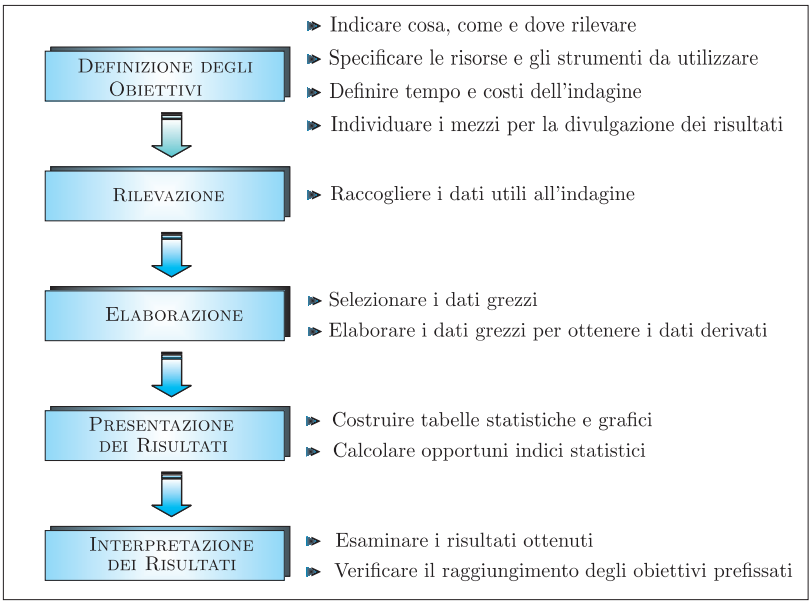

Fig. 1.1: fasi dell'indagine statistica.

dua nell'ente locale la sede prioritaria di raccolta del dato elementare, mentre la verifica, l'aggregazione e la pubblicazione sono di competenza dell'ente centrale che, in Italia, è rappresentato dall'*Istituto Centrale di Statistica*.

L'Istituto Centrale di Statistica

L'*Istituto Centrale di Statistica* (*ISTAT*), fondato nel 1926 da Corrado Gini (1884-1965), è l'organismo statale designato alla rilevazione ed elaborazione di dati statistici di interesse nazionale. Tale istituto è un organo collegiale indipendente e di controllo, con sede a Roma, presso la presidenza del Consiglio dei Ministri ed è parte integrante del *Sistema Statistico Nazionale* (*SISTAN*).
Dal 1989 gli uffici di statistica, istituiti presso i Ministeri, le regioni e le province, comprese quelle autonome, i comuni, le Camere di Commercio e gli altri enti di interesse nazionale, collaborano con l'*ISTAT* per fornire un quadro attendibile sulla realtà del Paese. L'*ISTAT* ha il compito di produrre e diffondere informazione statistica in maniera affidabile, imparziale e trasparente. In particolare, tale istituto provvede alla predisposizione del programma statistico nazionale; all'esecuzione dei censimenti e delle altre rilevazioni statistiche inerenti i vari aspetti dell'amministrazione dello Stato; alla raccolta dei dati direttamente presso le famiglie e le imprese, utilizzando e rielaborando anche le informazioni presenti negli archivi delle amministrazioni pubbliche.
Tra le pubblicazioni periodiche a cura dell'*ISTAT* si possono citare:

- il *Bollettino Mensile di Statistica*, che consiste in una raccolta completa di dati riguardanti l'evoluzione dei fenomeni demografici, sociali, economici e finanziari;

- l'*Annuario Statistico Italiano*, che presenta in forma tabellare, i risultati delle rilevazioni e le elaborazioni statistiche prodotte dall'*ISTAT* e dagli altri enti del *SISTAN*;

- gli *Annali di Statistica*, che rappresentano la sede di dibattiti scientifici e di riflessioni storiche sulla materia.

La produzione editoriale, che raccoglie i risultati dell'*ISTAT*, è attualmente disponibile sia in formato cartaceo che elettronico, ed è consultabile, in parte gratuitamente, anche in rete, visitando il sito *web www.istat.it*. Mediante tale sito, viene fornita la possibilità di accedere direttamente, quindi *online*, a numerose banche dati ed indicatori socio-economici e di trasferire dati e tabelle, pronti per eventuali rielaborazioni. Ciò, evidentemente, garantisce una diffusione in tempi brevi dell'informazione raccolta ed un suo agevole utilizzo.
In ambito europeo, il compito di informazioni statistiche di qualità, indispensabili per la progettazione e lo sviluppo di politiche comunitarie è demandato al *Sistema Statistico Europeo* (*SSE*). L'armonizzazione della raccolta, dell'elaborazione e della presentazione dei dati ufficiali degli Stati membri dell'Unione Europea, è affidata ad un organismo, denominato *EuroStat*, il quale ha il compito di confrontare e verificare l'andamento economico, sociale e culturale degli Stati membri.

1.5 Tecniche di campionamento

Come è stato precedentemente accennato, le rilevazioni statistiche possono essere totali o parziali, a seconda che esse siano rivolte ad ottenere informazioni, rispettivamente, su tutte le *u.s.* della popolazione o su un suo sottoinsieme.
La rilevazione parziale può scaturire dall'esigenza di ridurre tempi e costi dell'indagine statistica, ma può essere anche una scelta obbligata, come nei casi in cui risulta impossibile definire con precisione la popolazione di riferimento: si pensi ad una particolare specie animale presente su tutto il pianeta, oppure al caso in cui l'indagine comporti la distruzione delle *u.s.* osservate, come avviene spesso in campo industriale nel controllo della qualità di un prodotto.
Ogni qualvolta l'indagine statistica venga effettuata sul campione, è opportuno procedere ad un'accurata selezione, denominata *campionamento*, delle *u.s.*, in modo tale che il campione possa rappresentare un'immagine fedele, ma ridotta, della popolazione da cui deriva.
La costruzione di un campione può avvenire in due modi differenti:

- *per scelta casuale*,

- *per scelta ragionata*.

1.5.1 Campionamento per scelta casuale

Nel campionamento per scelta casuale, l'individuazione delle *u.s.* della popolazione che costituiranno il campione avviene in base a criteri che assicurano a tutte le unità della popolazione la stessa possibilità di entrare a fare parte del campione.
In tal caso, le eventuali divergenze tra le caratteristiche del campione e quelle della popolazione sono attribuibili esclusivamente a fattori di natura accidentale, per cui il campione

casuale può essere considerato un'immagine ridotta ma fedele della popolazione. Ciò rappresenta un aspetto molto importante per l'attendibilità dei risultati derivanti da rilevazioni parziali.
L'espressione "per scelta casuale" non deve indurre a pensare, quindi, ad un procedimento di selezione svolto senza regole, ma ad una scelta eseguita sulla base di un metodo rigoroso che, nel selezionare una *u.s.* dalla popolazione, ne assicuri l'imparzialità.
Le procedure di campionamento delle *u.s.* sono ampiamente descritte nell'ambito della *Teoria dei campioni*. Per le finalità del presente volume, si analizzeranno le tecniche di *campionamento da popolazioni finite*, ovvero da popolazioni costituite da un numero finito di *u.s.* Per cui, nel presentare le diverse tecniche di campionamento, si indicherà con M il numero delle *u.s.* di cui è composta la popolazione e con m la numerosità del campione.

Campionamento per estrazione casuale
Tale tecnica di campionamento presuppone che le *u.s.* della popolazione siano preventivamente numerate da 1 ad M. L'estrazione di m *u.s.* dalla popolazione viene così associata all'estrazione di m palline da un'urna contenente M palline, anch'esse numerate progressivamente da 1 ad M, a ciascuna delle quali corrisponde l'*u.s.* contraddistinta dallo stesso numero d'ordine. L'estrazione casuale può essere effettuata

1. *con ripetizione*, se ogni pallina estratta viene rimessa nell'urna prima della successiva estrazione. Questo implica che una stessa pallina possa essere estratta più volte e, quindi, che una stessa *u.s.* della popolazione possa essere presente più di una volta nel campione. Inoltre, dal momento che la composizione dell'urna rimane la stessa dopo ogni estrazione, il numero m di estrazioni che si possono eseguire non è limitato dalla numerosità M della popolazione.
   Tale tecnica di estrazione casuale è denominata anche *estrazione bernoulliana*, dal nome del matematico J. Bernoulli che la propose;

2. *senza ripetizione*, se ogni pallina estratta non viene reinserita nell'urna. In tal caso, una stessa *u.s.* può presentarsi nel campione una sola volta. Inoltre, a differenza dell'estrazione con ripetizione, quella senza ripetizione è di tipo esaustivo, poiché il contenuto dell'urna si riduce progressivamente e, quindi, non è possibile eseguire un numero di estrazioni superiore ad M;

3. *in blocco*, se le m palline, che individuano le unità del campione, vengono estratte dall'urna contemporaneamente, mediante un'unica operazione, oppure vengono estratte in sequenza, senza reimmissione e senza attribuire alcuna importanza all'ordine con cui le palline vengono estratte dall'urna.

Altre tecniche di campionamento

Esistono, tuttavia, altre tecniche di campionamento che rispondono ad esigenze pratiche differenti. Quelle più ricorrenti nelle indagini statistiche sono descritte di seguito.

- **Campionamento stratificato**. Assegnata la popolazione di riferimento, essa viene suddivisa in un determinato numero di gruppi, denominati *strati*, costituiti da elementi il più possibile omogenei fra loro rispetto alla caratteristica che si sta esaminando; da ciascuno strato vengono estratti casualmente (con o senza ripetizione) gli elementi

del campione. Ad esempio, in un'indagine sulle propensioni di spesa, la popolazione residente in una regione può essere suddivisa in classi di reddito e da ciascuna classe possono essere estratte casualmente le *u.s.* del campione.

- **Campionamento a grappoli**. Assegnata la popolazione di riferimento e scelti casualmente i cosiddetti *grappoli*, ovvero insiemi di unità contigue, il campione è composto da tutte le *u.s.* appartenenti ai grappoli selezionati. Ad esempio, in un'indagine sulle caratteristiche degli alunni delle scuole secondarie di primo grado di una provincia, se si assume che ciascuna scuola rappresenti un grappolo, si estrae casualmente un determinato numero di scuole e tutti gli alunni delle scuole estratte costituiranno il campione.

- **Campionamento a due o più stadi**. In presenza di una popolazione molto numerosa, si può procedere con l'individuazione di unità primarie di rilevazione, denominate *unità di primo stadio*, successivamente di unità secondarie, denominate *unità di secondo stadio*, e così via fino ad individuare le *u.s.* del campione. Ad esempio, per la costruzione di un campione di famiglie di una determinata regione, è possibile dapprima scegliere un campione casuale di comuni della regione (unità di primo stadio) e poi selezionare nell'ambito di tali comuni un campione casuale di famiglie (unità di secondo stadio).

La selezione delle unità dalla popolazione può avvenire mediante l'utilizzo delle tavole aleatorie, oppure in maniera sistematica.

Campionamento mediante l'utilizzo di tavole aleatorie
Questa modalità di selezione richiede la preventiva numerazione delle *u.s.* della popolazione da 1 ad M, cosicché l'individuazione delle m *u.s.*, che formeranno il campione, può avvenire selezionando casualmente m numeri, compresi tra 1 ed M. Tali numeri possono essere scelti, leggendo opportunamente, per riga o per colonna, le cosiddette *tavole dei numeri aleatori*. In esse sono riportate delle raccolte di numeri ottenuti mediante un procedimento che ne assicuri la casualità e l'assenza di una legge di successione o di ordinamento. Alcune tavole, ad esempio, sono state costruite registrando le unità dei primi estratti nelle prime ruote nel gioco del Lotto, oppure avvalendosi di un calcolatore elettronico e di idonei algoritmi di generazione di numeri casuali. L'insieme dei numeri, scelti mediante l'utilizzo delle tavole dei numeri casuali, consente di definire un campione casuale bernoulliano.

Campionamento sistematico
Tale modalità di campionamento richiede che venga definito il cosiddetto *intervallo di campionamento*, indicato con k, pari ad M/m, e che sia scelto casualmente un numero i compreso tra 1 e k. In tal modo, posto che le *u.s.* siano state preventivamente numerate da 1 ad M, la prima *u.s.* selezionata sarà quella associata al numero i, le successive *u.s.* saranno individuate aggiungendo al suddetto numero il valore k e tutti i suoi multipli fino ad M. In altri termini, mediante tale tecnica, si procede a selezionare una *u.s.* ogni k *u.s.* della popolazione. Ad esempio, per formare un campione di 25 *u.s.* da una popolazione costituita da 250 elementi, numerati da 1 a 250, si definisce l'intervallo di campionamento $k = 250/25 = 10$ e si sceglie casualmente un numero i compreso tra 1 e 10. Supponendo che il numero i sia pari a 7, la prima *u.s.* del campione sarà quella corrispondente al numero

d'ordine 7; successivamente le altre *u.s.* saranno individuate aggiungendo, al numero $i = 7$, scelto a caso, il valore 10 e tutti i suoi multipli fino a 250; per cui le *u.s.* successive alla prima, saranno quelle corrispondenti ai numeri 17, 27, ..., 237, 247.

1.5.2 Campionamento per scelta ragionata

Tali tecniche di campionamento non si basano su criteri di casualità, ma piuttosto sulla conoscenza del fenomeno posseduta da chi pianifica l'indagine. Le *u.s.* del campione vengono, pertanto, scelte in modo del tutto soggettivo. I campioni costruiti con scelta ragionata possono essere considerati rappresentativi della popolazione soltanto se le informazioni su cui si basa la selezione sono complete e veritiere.

1.6 Caratteri e modalità

Mediante un'indagine statistica, è possibile analizzare uno o più aspetti del collettivo statistico oggetto di interesse. Ad esempio, una ricerca condotta sugli studenti dell'Università del Salento può essere rivolta a conoscere l'età, il tipo di diploma, il voto del diploma, il reddito della famiglia, il luogo di residenza, la professione del capofamiglia e così via. I differenti aspetti delle *u.s.* su cui interessa indagare sono denominati *caratteri statistici*, o più semplicemente *caratteri.*

Definizione 1.6. Carattere statistico. *Si definisce carattere statistico un particolare aspetto del collettivo statistico oggetto d'osservazione.*

Un carattere si manifesta sulle *u.s.* con modalità differenti. Ad esempio, il carattere "età" può presentarsi su uno studente dell'Università del Salento con la modalità 19 anni e su un altro studente con la modalità 21 anni. Analogamente, il carattere "tipo di diploma" può manifestarsi con la modalità "maturità tecnica" su uno studente e con la modalità "licenza liceale" su un altro. Come sarà specificato in seguito, le modalità di un carattere possono essere numeri o intensità, come nel caso del carattere "età", oppure attributi, come nel caso del carattere "tipo di diploma".
Pertanto, è possibile fornire una definizione rigorosa di *modalità.*

Definizione 1.7. Modalità. *Si definiscono modalità i differenti attributi o le differenti intensità che un carattere presenta nel collettivo statistico.*

È possibile definire le seguenti tipologie di caratteri statistici.

- **Caratteri qualitativi**, denominati anche *mutabili statistiche*, se le modalità con cui si presentano sono attributi o espressioni verbali. Esempi di caratteri qualitativi sono: il sesso, la nazionalità, il colore degli occhi, la religione, la professione, il titolo di studio, il gruppo sanguigno.
  Un carattere qualitativo può essere definito

  1. *ordinabile*, se tra le modalità esiste un ordine naturale di successione, come ad esempio il grado di istruzione, il grado di una gerarchia militare;

2. *non ordinabile*, se tra le modalità non esiste alcun ordine di successione, come ad esempio la professione, il luogo di residenza, lo stato civile.

Si evince, quindi, che i caratteri qualitativi ordinabili sono misurabili mediante una *scala ordinale*. In altri termini, assegnato un sistema di classificazione gerarchica delle modalità, si può affermare che un attributo precede o, viceversa, segue un altro, quindi è possibile effettuare operazioni di ordinamento tra gli attributi osservati. Nell'esempio del grado di istruzione, le modalità possono essere ordinate, dalla più bassa a quella più alta, in base alla seguente scala ordinale:

a) analfabeta,
b) alfabeta privo di titolo di studio,
c) licenza elementare,
d) licenza di scuola media,
e) diploma,
f) laurea,
g) master.

D'altra parte, i caratteri non ordinabili sono misurabili a livello di *scala nominale*, poiché è possibile valutare soltanto l'identità tra gli attributi, ovvero l'uguaglianza, o viceversa, la disuguaglianza tra gli stessi.

- **Caratteri quantitativi**, denominati anche *variabili statistiche*, se le modalità con cui si presentano sono espresse da numeri, ovvero sono intensità misurabili a livello di *scala cardinale*. Esempi di caratteri quantitativi sono: l'età, il peso corporeo, la statura, il reddito annuo, il punteggio in una competizione agonistica.
  Un carattere quantitativo può essere definito *discreto*, se può assumere al più un'infinità numerabile di valori, altrimenti, viene denominato *continuo*.
  Si intuisce facilmente che la distinzione tra caratteri continui e discreti è prevalentemente di natura concettuale, poiché, in pratica, anche l'insieme delle misurazioni di un carattere continuo non sarà altro che un insieme discreto di numeri, a causa del limite oggettivo oltre il quale non può spingersi la precisione dello strumento di misurazione utilizzato, o di chi esegue la rilevazione. In altri termini, con riferimento ai caratteri continui, occorre tenere conto che nelle applicazioni è inevitabile una discretizzazione dei valori osservati.

Dopo aver osservato un carattere statistico ed aver individuato le modalità distinte con cui tale carattere si manifesta, può essere particolarmente utile determinare il numero di volte con cui le modalità si presentano nel collettivo statistico e, quindi, calcolare le cosiddette *frequenze assolute.*

Definizione 1.8. Frequenza assoluta. *Si definisce frequenza assoluta di una modalità, il numero di volte con cui tale modalità si è presentata nel collettivo statistico osservato.*

Ad esempio, si supponga che sia stato osservato il carattere "tipo di diploma" su un collettivo statistico costituito da 30 studenti universitari. Se nel collettivo statistico considerato, 10 studenti hanno conseguito il diploma di maturità tecnica, 5 la licenza liceale e 15 il diploma di scuola professionale, allora le modalità del carattere statistico "tipo di diploma" sono 3 ed, in particolare

- la modalità "maturità tecnica" si è presentata con una frequenza assoluta pari a 10,
- la modalità "licenza liceale" si è presentata con una frequenza assoluta pari a 5,
- la modalità "diploma di scuola professionale" si è presentata con una frequenza assoluta pari a 15.

1.7 Formalismo statistico

Al fine di utilizzare un formalismo omogeneo e rigoroso nel corso del presente volume, è opportuno introdurre la seguente notazione generale.

- I caratteri statistici si indicano con le lettere maiuscole dell'alfabeto latino, solitamente X, Y, Z. Ad esempio, con la seguente notazione

  $$X = \text{"età"}$$

  si assegna al carattere statistico "età", oggetto di osservazione, la lettera X.

- La numerosità del collettivo statistico viene indicata con n e le osservazioni di un carattere X su ciascuna *u.s.* sono denotate con le lettere minuscole,

  $$x_1, x_2, \ldots, x_n,$$

  dove l'indice, posto a pedice, è riferito all'ordine con cui sono osservate le *u.s.* Pertanto, x_1 rappresenta l'osservazione del carattere X sulla prima *u.s.*, x_2 rappresenta l'osservazione di X sulla seconda *u.s.*, ed analogamente, x_n rappresenta l'osservazione del carattere X sull'n-esima *u.s.*
  Ad esempio, con riferimento al carattere quantitativo X = "età", osservato su alcuni iscritti ad un *club* sportivo, si supponga di aver rilevato i seguenti dati, in anni compiuti:

  - il valore "20" sulla prima *u.s.*,
  - il valore "20" sulla terza *u.s.*,
  - il valore "22" sulla quinta *u.s.*
  - il valore "19" sulla seconda *u.s.*,
  - il valore "19" sulla quarta *u.s.*,

  Per cui risulta $n = 5$; d'altra parte i valori osservati sono indicati come segue:

  $$x_1 = 20, \qquad x_2 = 19, \qquad x_3 = 20, \qquad x_4 = 19, \qquad x_5 = 22.$$

- Le osservazioni del carattere X, disposte secondo un prefissato criterio di ordinamento, sono indicate mediante la notazione

  $$x_{(1)}, x_{(2)}, \ldots, x_{(n)}.$$

  Nel caso di carattere quantitativo, si fissa un criterio di ordinamento non decrescente, in maniera tale che risulti

  $$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)},$$