

1

La statistica e i suoi metodi

1.1 Note storiche

Sin dalle origini della civiltà l'uomo ha sentito il bisogno di annotare con registrazioni epigrafiche dati relativi a fenomeni che lo coinvolgevano nella vita sociale. Sono noti antichissimi scritti cinesi in cui si fa cenno di misure delle terre e di enumerazioni degli abitanti, con distinzione degli addetti ai mestieri, attribuiti ad un'epoca attorno al 2300 a.c.

Nella Bibbia sono ricordati diversi censimenti degli Ebrei ed anche gli antichi Egizi eseguirono catasti e censimenti demografici talora corredati dello stato civile dei censiti. Per quanto riguarda i Romani con la Costituzione di Servio Tullio venne prescritto un «census» da ripetersi ogni cinque anni; diversi documenti in cui si citano registrazioni di dati relativi al periodo dell'Impero Romano sono giunti fino a noi. Il Medio Evo risulta invece scarso di documentazione.

A partire dal XV secolo cominciarono ad apparire opere aventi come scopo la raccolta di dati demografici ed economici dei vari Stati. Nacque così una scienza descrittiva degli Stati indicata come *Statistica*. Il maggior merito dell'evoluzione di questa scienza, che raggiunse un notevole sviluppo già nel XVIII secolo, può essere attribuito a G. Achenwall (1719-1772) dell'Università di Gottinga. Comparvero in questo periodo le prime tabelle comparative delle risorse geografiche ed economiche dei vari Stati.

Nel XVII secolo comparve in Inghilterra, ad opera di J. Graunt (1620-1674) e di W. Petty (1623-1678), una disciplina che fu in seguito denominata *aritmetica politica*, avente come oggetto di studio l'osservazione numerica delle nascite, delle morti e in generale dei fenomeni demografici. Tali studi, finalizzati al calcolo delle rendite vitalizie, si diffusero in tutti gli Stati ed ebbero come massimo esponente il pastore prussiano J.P. Süßmilch (1707-1767), il quale comprese che le regolarità di un fenomeno potevano essere evidenziate attraverso un numero molto elevato di osservazioni e per primo calcolò il periodo di raddoppiamento di una popolazione.

Nel XIX secolo venne introdotto nelle indagini statistiche, soprattutto ad opera di P.S. Laplace (1749-1827), il calcolo delle probabilità, nato in Francia come branca autonoma della matematica da studi di B. Pascal (1623-1662) e di P. Fermat (1608-1665).

L'importanza di tale introduzione nella metodologia statistica è facilmente intuibile se si pensa che sovente è possibile associare modelli probabilistici allo studio di fenomeni statistici trattandoli con il formalismo proprio del calcolo delle probabilità.

Agli inizi del XX secolo la statistica è ormai una scienza autonoma con una sua precisa fisionomia indipendente dai settori di applicazione.

I contributi maggiori furono opera di: F. Galton (1822-1911), nel settore della *statistica medica* e nella *teoria della correlazione*; K. Pearson (1857-1936) che svolse importanti studi di *statistica matematica* e fece costruire la maggior parte delle tavole statistiche tuttora in uso; G.S. Gosset (1876-1937), che pubblicò i suoi lavori sotto lo pseudonimo di Student, si occupò a fondo della *teoria dei campioni*.

Infine per quanto riguarda il nostro secolo si può ricordare la formalizzazione dell'inferenza statistica, vale a dire di quella branca della statistica che si occupa di trarre conclusioni generali analizzando un piccolo numero di elementi dal punto di vista statistico, soprattutto per opera di R.A. Fisher (1890-1962). Il contributo maggiore della Scuola Italiana fu di C. Gini (1884-1965) che si occupò soprattutto delle misure di variabilità e fondò nel 1926 l'Istituto Centrale di Statistica (ISTAT), mentre nel 1936 fondò la prima *Facoltà di Scienze Statistiche* in Italia e nel mondo.

A conclusione di questa breve panoramica storica si può notare che l'analisi statistica dei dati si è andata sempre più sviluppando e attualmente il ricorso alle metodologie statistiche da parte delle scienze più disparate è di uso quotidiano, soprattutto per la diffusione del calcolo automatico che permette una rapida elaborazione di notevoli quantità di dati.

1.2 Definizione, scopi ed applicazioni della statistica

La definizione storica di statistica come scienza degli Stati è ormai da considerarsi superata. Le definizioni che vengono date oggi di tale scienza sono parecchie; W.F. Willcox (1861-1946) nel 1934 ne raccolse addirittura 124 e da allora ne sono state aggiunte parecchie altre. Tuttavia la definizione che ricorre maggiormente prende lo spunto dal tipo di fenomeni studiati dalla statistica: i *fenomeni statistici* o *collettivi* ovvero quei fenomeni il cui studio richiede una raccolta di numerose osservazioni.

Le osservazioni possono essere eseguite su diversi soggetti una sola volta oppure ripetute sullo stesso soggetto più volte; sono fenomeni collettivi ad esempio il numero di nascite in un anno, il numero di studenti promossi in una scuola e il numero di case costruite in un anno, il prezzo di un bene di consumo in periodi successivi, la radioattività in uno stesso luogo misurata in periodi successivi, la temperatura di una resistenza al variare della corrente, il peso di una persona che segue una dieta particolare.

Possiamo quindi definire:

Statistica quella scienza che ha come fine lo studio quantitativo o qualitativo dei fenomeni collettivi.

I metodi che vengono utilizzati per studiare tali fenomeni sono molteplici e spesso legati al tipo di fenomeno; si possono distinguere due grandi settori: quello della *statistica descrittiva* e quello della *inferenza statistica*.

Compito della statistica descrittiva è quello di raccogliere, ordinare, riassumere, presentare ed analizzare i dati ottenuti dallo studio del fenomeno collettivo.

Compito della inferenza statistica è l'estendere i risultati ottenuti con la statistica descrittiva ad un insieme di elementi più vasto di quello studiato, fornendo strumenti analitici per trarre delle conclusioni e prendere delle decisioni.

Attualmente la statistica trova applicazioni in diversi settori; tra i più importanti si possono citare:

- a) Ricerca scientifica in genere. Studi e ricerche quantitative e qualitative di varia natura dalla medicina all'agricola, dalla chimica alla geologia, dalla fisica all'astronomia e alle scienze dell'informazione.
- b) Scienze demografiche e attuariali. Censimenti demografici, demografia storica, dinamica della popolazione, studi di natalità, mortalità, stato civile, assicurazioni, rendite.
- c) Scienze economiche. Produzione agricola, mineraria, industriale, prezzi delle merci, andamento dell'inflazione, dei salari, dei consumi, delle importazioni ed esportazioni, dei trasporti, redditi, bilanci in genere.
- d) Scienze sociali. Sondaggi di pubbliche opinioni, inchieste sociali, andamento della disoccupazione, dell'assistenza pubblica, del turismo, della criminalità, studi sulla sanità pubblica.
- e) Scienze per la programmazione economica e territoriale. Elaborazione e previsione degli effetti di piani economici e sociali, di piani settoriali di sviluppo.
- f) Ricerca operativa. Gestione delle risorse, delle scorte, degli investimenti, ottimizzazione dei processi di lavorazione, strategie competitive.

1.3 L'indagine statistica e le sue fasi

I fenomeni statistici possono essere studiati mediante le *indagini statistiche*. Per compiere un'indagine statistica è necessario effettuare delle osservazioni; scopo di tali osservazioni è evidenziare eventuali regolarità del fenomeno investigato.

Nello svolgimento di una qualsiasi indagine statistica si possono distinguere varie fasi:

- definizione degli obiettivi;
- individuazione del fenomeno e del collettivo;
- rilevazione dei dati;
- ordinamento o spoglio dei dati;
- elaborazione ed analisi;
- interpretazione dei risultati e divulgazione.

Più specificamente le precedenti fasi si possono dettagliare come segue:

- a) *Definizione degli obiettivi.* Occorre stabilire accuratamente i termini del problema che si deve affrontare e per il quale si pensa di utilizzare la statistica; inoltre in questa fase si deve chiarire quali obiettivi si vogliono raggiungere onde poter applicare le metodologie più appropriate. Per compiere un'indagine sul numero di giornate lavorative perse nell'industria, ad esempio, è necessario stabilire in primo luogo quali sono le possibili cause: malattia, sciopero o altro.
- b) *Individuazione del fenomeno e del collettivo.* Consiste nell'individuare quale tipo di *dato statistico* è utile alla soluzione del problema. Con il termine dato statistico si intende quel valore che è il risultato di più osservazioni e che costituisce la misura del fenomeno considerato. È quindi necessario definire su quali soggetti saranno effettuate le osservazioni (ovvero stabilire i confini del collettivo da osservare). Queste due fasi, che possono sembrare ovvie a prima vista, sono quelle che il più delle volte, se trascurate, portano a notevoli perdite di tempo e di risorse. Nel caso dell'indagine sulle giornate lavorative perse nell'industria, ad esempio, in questa fase è necessario stabilire su quali fasce di lavoratori deve essere effettuata l'indagine, in quali settori, in quali regioni e se i dati devono essere annuali o mensili.
- c) *Rilevazione dei dati.* Consiste nell'individuare le modalità tecniche per effettuare la raccolta dei dati. Questa fase, comunemente detta rilevazione, può utilizzare, a seconda del problema, registri, schede, tavole statistiche, questionari, modelli di rilevazione. Per quel che riguarda l'esempio precedente i dati possono essere rilevati per mezzo di schede compilate dalle diverse industrie.
- d) *Ordinamento o spoglio dei dati.* Questa fase consiste nell'ordinamento o *spoglio dei dati*, vale a dire nella classificazione delle informazioni ottenute e nell'eliminazione di eventuali errori individuabili. A conclusione di questa

e)

f)

stati

elabo

S

men

gine

posso

1.4

P

zioni

te e

L

popo

collet

fase si ottengono tabelle di dati rappresentativi dell'intero fenomeno che si possono meglio visualizzare e sintetizzare mediante rappresentazioni grafiche.

- e) *Elaborazione ed analisi.* Nella quinta fase si usano tecniche matematiche per elaborare i dati ottenuti. Tali tecniche possono spaziare dal semplice calcolo di una media al calcolo di un andamento di un fenomeno nel tempo (*trend*).

In questa fase i dati raccolti vengono trasformati in dati sintetici in modo da permettere l'individuazione di regolarità del fenomeno statistico investigato. Relativamente all'esempio precedente, si potrà calcolare il numero medio di giornate lavorative perse nei diversi settori da ogni lavoratore, confrontando il fenomeno nei diversi settori dell'industria, vedere l'andamento nel tempo se i dati si riferiscono a più anni o se l'indagine viene svolta ogni anno.

- f) *Interpretazione dei risultati e divulgazione.* Ottenuti i parametri riassuntivi del fenomeno si può passare alla sesta fase: interpretazione e divulgazione dei risultati. L'interpretazione dirà se sono stati o meno raggiunti gli obiettivi prefissati e quindi se si può considerare conclusa l'indagine. A questo punto si possono divulgare i risultati. Nell'esempio precedente può essere valutato il costo economico delle giornate lavorative perse per le diverse cause di assenza dal lavoro: possono così essere evidenziati i fenomeni che portano le perdite maggiori per valutare le possibilità di interventi.

La rilevazione, l'ordinamento, la rappresentazione e il calcolo dei parametri statistici dei dati vengono attualmente svolti per lo più con l'ausilio di elaboratori elettronici che rendono veloci ed affidabili tali operazioni.

Si può osservare, a conclusione di questo paragrafo, che le fasi precedentemente elencate sono sostanzialmente una traccia per lo svolgimento di un'indagine statistica; infatti non sempre si susseguono nell'ordine indicato e talvolta possono sovrapporsi.

1.4 Classificazione di una popolazione statistica

Per poter compiere una qualsiasi indagine statistica e ricavare delle informazioni ovvero dei *dati statistici*, è necessario innanzi tutto scegliere adeguatamente e definire rigorosamente gli elementi su cui svolgere l'indagine stessa.

Si definiscono unità statistiche gli elementi su cui viene svolta l'indagine statistica.

L'insieme delle unità statistiche costituisce la *popolazione statistica*; tale popolazione viene anche sovente indicata con i termini *universo statistico* o *collettivo statistico*.

Sono popolazioni statistiche ad esempio:

- l'insieme dei lavoratori di una certa categoria;
- l'insieme delle auto prodotte in una fabbrica;
- l'insieme degli studenti di una città;
- l'insieme delle misure effettuate su di un circuito elettronico.

Nella popolazione statistica costituita da tutti i lavoratori l'unità statistica è rappresentata dal singolo lavoratore; nella popolazione costituita dalle automobili l'unità statistica è la singola automobile; il singolo studente rappresenta l'unità statistica della popolazione degli studenti.

Quando si svolge un'indagine statistica si desidera studiare determinate caratteristiche comuni a tutti gli elementi della popolazione, tali caratteristiche prendono più propriamente il nome di *caratteri statistici della popolazione*.

Sono caratteri statistici ad esempio:

- per una popolazione di lavoratori: il salario, il tipo di orario, il tipo di contratto, il numero di giornate lavorative perse per malattia;
- per una popolazione di auto: il prezzo, la cilindrata, il colore, la targa pari o dispari, la ditta di produzione;
- per una popolazione di studenti: l'età, il numero di promozioni durante la carriera scolastica, il tipo di scuola frequentato.

Il carattere statistico investigato deve essere presente in tutte le unità statistiche della popolazione, ma certamente non si presenta in modo uguale in tutte le unità.

Se si studia il carattere «ditta di produzione» della popolazione statistica costituita da tutte le auto vendute in Italia nel 1985, le auto devono essere raggruppate a seconda della ditta di produzione; se due auto sono state prodotte dalla stessa ditta, allora diciamo che il carattere si presenta, nelle due auto, allo stesso modo, o più esattamente le due auto *hanno la stessa modalità*; viceversa se due auto sono state prodotte da ditte diverse, allora *presentano modalità diverse*.

In generale si può dire che:

le modalità di un carattere statistico sono i modi di presentarsi del carattere.

Nel caso della popolazione di auto dell'esempio precedente le modalità sono le diverse ditte di produzione. Presa dunque una popolazione, per studiare un carattere statistico è necessario *classificare* ogni unità in base alla modalità che presenta.

Al termine di questa operazione la popolazione statistica risulta suddivisa in tanti gruppi omogenei quante sono le modalità del carattere, ovvero *suddivisa in classi*.

Gli elementi che presentano una stessa modalità formano una classe e viceversa tutti gli elementi di una classe presentano almeno una modalità in

comune. Ad esempio se si classifica un gruppo di persone in base al colore dei capelli il carattere è: «colore dei capelli»; le modalità sono: biondo, castano, nero; tutte le persone bionde formano la classe delle persone bionde, tutte le persone castane formano la classe delle persone castane e così via.

Se si sceglie di classificare il precedente insieme di individui secondo il carattere «stato civile», si possono raggruppare le persone secondo le modalità: celibe, coniugato, vedovo e si ottiene la classe dei celibi, la classe dei coniugati, la classe dei vedovi; se invece queste persone devono essere classificate in base all'età, le modalità di classificazione sono costituite dalle varie fasce di età in cui gli individui vengono raggruppati.

Come si può vedere da questi esempi i caratteri secondo cui si può studiare una popolazione statistica possono essere di varia natura; generalmente si parla di *caratteri qualitativi e caratteri quantitativi*.

Si parla di caratteri qualitativi quando le modalità rappresentano delle qualità. Sono caratteri qualitativi: lo stato civile, il colore dei capelli, la nazionalità di una persona, la categoria di un albergo. Si parla invece di caratteri quantitativi quando le modalità esprimono delle quantità misurabili e sono rappresentate con singoli valori oppure con intervalli di valori. Sono caratteri quantitativi: l'età, la statura, il peso, il numero di figli di una persona.

Se si considera la popolazione statistica delle automobili vendute in un certo anno, queste possono essere classificate secondo i caratteri qualitativi: ditta di produzione, colore, modello, tipo di trazione, tipo di motore; o secondo i caratteri quantitativi: cilindrata, prezzo, numero di posti, velocità massima raggiungibile, consumo medio.

I caratteri quantitativi possono suddividersi in *discreti e continui*; si hanno caratteri quantitativi discreti quando le loro modalità assumono solo determinati valori (il numero di figli per ogni coppia di coniugi, il numero di matrimoni celebrato in una settimana in un dato comune); si hanno caratteri quantitativi continui quando le loro modalità possono assumere qualsiasi valore in un certo intervallo.

In questo caso è necessario suddividere l'intervallo che comprende tutti i valori in diversi sottointervalli nei quali viene classificata ogni unità statistica. Se ad esempio si vogliono classificare delle persone secondo la loro statura, si possono raggruppare in un unico sottointervallo tutte le persone con statura compresa tra 150 e 159.99 centimetri, in un altro sottointervallo tutte le persone con statura compresa tra 160 e 169.99 centimetri e così via.

Per classificare le diverse persone si può costruire una tabella come in fig. 1.1 dove nella colonna di sinistra sono riportate le modalità espresse da intervalli di valori e nella colonna di destra è riportato il numero di persone che rientrano in ciascuna classe.

Sovente vengono trattati come continui dati che in realtà sono discreti; questo si verifica quando si parla di grandezze che possono assumere un numero molto elevato di valori quali ad esempio il numero di abitanti di una città, il salario di un lavoratore, il numero dei disoccupati in una nazione; tali grandezze possono assumere solo valori discreti ma sono approssimabili a valori continui

Fig. 1.1

Modalità (cm)	Numero persone
150-159.99	3
160-169.99	12
170-179.99	10
.....	...

in quanto la differenza tra due valori consecutivi è trascurabile rispetto ai valori stessi.

Vi è infine da notare che le modalità di un carattere quantitativo vengono generalmente presentate in un ordine logico: dalla minore alla maggiore.

Per i caratteri qualitativi invece si hanno casi in cui l'ordine delle modalità è definito a priori, ma anche casi in cui l'ordine di presentazione delle diverse modalità è lasciato alla discrezione di chi le enumera; nel primo caso si parla di *modalità ordinate*, nel secondo caso si parla di *modalità non ordinate* o *caratteri sconnessi*.

Sono caratteri qualitativi a modalità ordinate: la categoria degli alberghi, l'ordine di nascita dei fratelli, il titolo di studio conseguito, il ceto sociale.

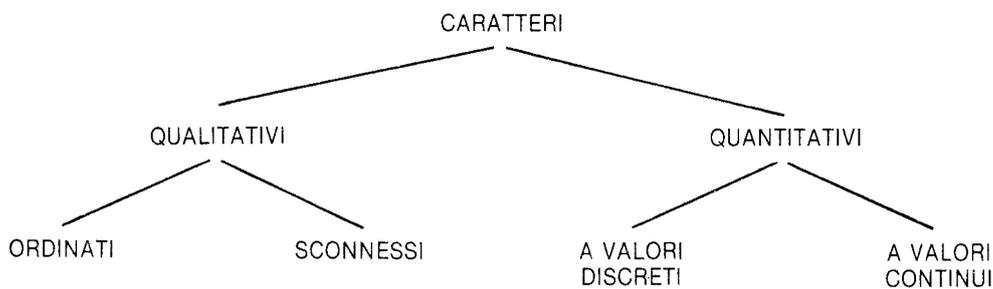
Sono invece caratteri qualitativi a modalità non ordinate o caratteri sconnessi: la nazionalità delle persone, la religione di appartenenza, la professione, lo stato civile. In fig. 1.2 è sintetizzata la classificazione dei caratteri statistici.

1.5 Rilevazione dei dati

Individuata la popolazione statistica, scelti uno o più caratteri da investigare, prima di procedere alla raccolta vera e propria dei dati occorre definire il *piano di rilevazione* dei dati.

Il piano di rilevazione consiste principalmente nello stabilire: le risorse disponibili e il tipo di rilevazione che deve essere effettuato. Le risorse

Fig. 1.2



disponibili consistono negli strumenti, nel personale, nei fondi da destinare all'indagine; prima di iniziare l'indagine devono essere valutati accuratamente i costi al fine di stabilire se le risorse sono sufficienti.

La rilevazione può essere di tipo *totale* o *parziale*; è totale quando viene effettuata su tutte le unità statistiche della popolazione, è parziale quando viene eseguita solamente su una parte delle unità statistiche.

In questo caso la rilevazione è più comunemente detta *campionaria* in quanto analizza solamente un *campione statistico*.

Un campione statistico è un sottoinsieme opportunamente scelto dall'intera popolazione.

Ad esempio, se si vuole compiere un'indagine sul reddito medio degli abitanti di una città, la popolazione statistica è costituita da tutti gli abitanti della città che presentano qualche forma di reddito; se si compie un'indagine totale si devono classificare *tutti* gli abitanti. Se invece si compie un'indagine campionaria, si deve scegliere un campione vale a dire un gruppo ristretto di individui da classificare.

Si noti che risulta indispensabile che il campione sia scelto in modo tale da essere rappresentativo dell'intera popolazione della città: il numero di persone a reddito basso del campione deve essere proporzionale al numero di persone a reddito basso dell'intera città, così come il numero di persone a reddito medio e il numero di persone a reddito elevato.

Infatti, se nel campione figurano troppe persone con reddito medio-alto, si ottiene un valore sovrastimato, se figurano troppe persone con un reddito basso si ottiene una sottostima del reddito medio.

La scelta tra rilevazione totale (effettuata su tutta la popolazione) e rilevazione campionaria (effettuata solo sul campione) è dovuta alle risorse disponibili.

Le rilevazioni totali offrono informazioni complete circa l'intera popolazione, ma, quando la popolazione è numerosa, comportano tempi di realizzazione e costi elevati.

Le rilevazioni parziali invece comportano costi decisamente meno elevati, sono più rapide, presentano però lo svantaggio di dare risultati approssimati e nel caso peggiore, se il campione non è rappresentativo, possono portare a conclusioni errate.

Per minimizzare questi svantaggi sono stati messi a punto diversi metodi di scelta del campione e di analisi dei risultati; tali metodi costituiscono l'inferenza statistica che sarà trattata nei prossimi volumi.

Una rilevazione statistica può inoltre essere *preliminare* o *definitiva*; infatti se l'indagine è complessa oppure totalmente nuova, prima della rilevazione finale può essere svolta una rilevazione preliminare di prova in modo da avere delle indicazioni sulle dimensioni della popolazione e dei mezzi occorrenti per effettuare la rilevazione definitiva.

Infine si deve notare che le rilevazioni possono essere *occasional*, *periodiche*

o *continue*; sono rilevazioni occasionali i sondaggi d'opinione su argomenti specifici; sono rilevazioni periodiche i censimenti; sono rilevazioni continue le registrazioni inerenti le nascite e le morti.

Nel piano di rilevazione dei dati occorre inoltre indicare i mezzi tecnici con cui si intende acquisire i dati (rilevazione postale, intervista, telefonata, esame di documenti) e i metodi di classificazione delle unità statistiche.

Il metodo più semplice per classificare i dati ottenuti da un'indagine e poterli facilmente elaborare è quello di riportarli in *tabelle* che li raggruppano in classi. È opportuno che tali tabelle siano preparate durante l'impostazione dell'indagine in quanto devono riportare tutte le modalità dei caratteri da investigare e, se prima dell'indagine non viene predisposto anche un piano di elaborazione dei dati, difficilmente si possono evitare lacune nell'acquisizione delle informazioni.

A questo proposito è da tener presente il fatto che se un'indagine statistica, una volta effettuata risulta incompleta, difficilmente può essere completata in un tempo successivo, in quanto i nuovi risultati possono risentire di circostanze diverse e fornire informazioni distorte.

1.6 Il trattamento dei dati

Una volta rilevati i dati si deve effettuare lo spoglio (quarta fase dell'indagine statistica) per poter procedere all'analisi e all'elaborazione (quinta fase).

Un tempo queste fasi di spoglio, analisi ed elaborazione, erano lunghissime in quanto svolte manualmente; attualmente invece avvengono prevalentemente con l'ausilio degli elaboratori elettronici.

Prima di procedere all'elaborazione vera e propria i dati devono essere registrati su un supporto magnetico (nastro o *floppy disk*); già durante la fase di registrazione, a seconda del tipo di elaboratore disponibile per tale operazione, viene effettuato un primo controllo parziale della validità dei dati registrati.

Questa registrazione prende il nome di *data entry* e normalmente viene espletata da appositi uffici dotati di attrezzature atte a svolgere esclusivamente simili operazioni (centri di registrazioni dati).

Terminata la fase di *data entry*, i dati registrati su nastro o disco vengono elaborati dal computer con appositi programmi.

Tali programmi dapprima provvedono ad un ulteriore controllo di validità dei dati stessi, segnalando all'utente eventuali anomalie e successivamente, dopo che sono state effettuate tutte le modifiche opportune, sintetizzano i dati acquisiti dal rilevamento e forniscono le analisi statistiche richieste.

Naturalmente i dati per poter essere immessi nel calcolatore devono essere opportunamente *codificati*, vale a dire ridotti in una opportuna forma numerica indipendentemente dalla modalità che rappresentano.

Per meglio chiarire in cosa consiste la codifica dei dati si consideri la fig. 1.3 che rappresenta parte di un questionario ISTAT (mod. ISTAT/P/70) per la rilevazione trimestrale delle forze di lavoro.

La compilazione del questionario avviene barrando le apposite caselle, dove ad ogni casella viene fatto corrispondere un codice (per lo più numerico).



Fig. 1.3



ISTITUTO CENTRALE DI STATISTICA

FAC-SIMILE

RILEVAZIONE TRIMESTRALE
DELLE FORZE DI LAVORO

Numero generale progressivo:

da 0001 al totale dei modelli compilati dal Comune.

1. SESSO

— Maschio 1
 — Femmina 2

2. RELAZIONE CON IL CAPO FAMIGLIA

— Capo famiglia 1
 — Coniuge del capo famiglia 2
 — Figlio del capo famiglia (o del coniuge) 3
 — Ascendente del capo famiglia (o del coniuge) 4
 — Altro parente 5
 — Altro 6

3. TITOLO DI STUDIO

— Analfabeta 1
 — Nessun titolo 2
 — Licenza elementare 3
 — Licenza scuola media inferiore 4
 — Diploma scuola media superiore 5
 — Laurea 6

4. STATO CIVILE

— Celibe o nubile 1
 — Coniugato 2
 — Vedovo 3
 — Separato, divorziato, già coniugato 4

5. CONDIZIONE UNICA O PREVALENTE

Qual è attualmente la sua condizione?

— Occupato 1
 — In cerca di nuova occupazione 2
 — In cerca di prima occupazione 3
 — In servizio di leva 4
 — Casalinga 5
 — Studente 6
 — Inabile al lavoro 7
 — Persona ritirata dal lavoro 8
 — Altra condizione (benestante, anziano, e simili) 9

6. Ha effettuato ore di lavoro nella settimana di riferimento qualunque sia la condizione dichiarata al q. 10.1?

— Sì 1
 — No 2

Tutti i questionari compilati dalle diverse unità statistiche costituiscono l'insieme dei *dati grezzi*. Il *data entry* consiste nella trasposizione dei dati grezzi codificati in una *matrice dei dati grezzi* o *tabella inventario* pronta per essere elaborata automaticamente dall'elaboratore.

In fig. 1.3 i codici sono rappresentati a fianco di ogni singola risposta, accanto alla casella che viene barrata in caso di risposta affermativa. La tabella dei dati grezzi si presenterà come in fig. 1.4 dove sono riportate per brevità solamente 4 unità statistiche.

Fig. 1.4

Caratteri						
Num.	(1)	(2)	(3)	(4)	(5)	(6)
001	1	1	4	2	1	1
002	2	3	6	1	3	2
003	1	3	5	1	4	2
004	2	2	4	2	5	2
...
...

- (1) Sesso
 (2) Relaz. con il capofam.
 (3) Titolo di studio
 (4) Stato civile
 (5) Condizione
 (6) Lavoro nella settimana precedente

Nel caso specifico riportato in fig. 1.4 l'unità statistica contrassegnata con il numero 001 risulta essere una persona di sesso maschile, capofamiglia, con licenza di scuola media inferiore, coniugato, occupato, che ha lavorato nella settimana precedente la compilazione del questionario; analogamente si possono ricavare le informazioni per le altre unità statistiche.

Questa tabella contiene pertanto tutte le informazioni rilevate per tutti gli intervistati e, letta nel senso delle righe, presenta il profilo completo di ogni intervistato, letta nel senso delle colonne, presenta la distribuzione delle diverse modalità di ogni singolo carattere analizzato.

Si noti che le tabelle inventario sono tabelle esclusivamente numeriche e contengono tutte le informazioni richieste all'intervistato, indipendentemente dal fatto che si tratti di caratteri qualitativi o quantitativi.

L'introduzione di questa tabella in un elaboratore non presenta particolari difficoltà proprio perché si tratta di una tabella esclusivamente numerica.

Nella fase di *data entry* possono essere di grande utilità programmi di controllo degli errori di immissione dati; tali programmi, in alcuni casi, prevedono anche la correzione automatica degli errori individuati.

Immessi nell'elaboratore, i dati vengono analizzati, raggruppati, elaborati mediante opportuni programmi di analisi; si ottengono in output tabelle complessive sull'andamento dei caratteri investigati.

In fig. 1.5 è presentata una tabella riassuntiva inerente l'intero modello ISTAT/P/70 (parzialmente riportato in fig. 1.3)

I sistemi informatici sono utilizzati soprattutto nelle grandi masse di dati che possono essere elaborati rapidamente e con alta affidabilità delle informazioni in uscita.

Le difficoltà che si possono presentare in un'indagine non sono più dunque i tempi di elaborazione e spoglio, una volta lunghissimi, ma nella buona preparazione dell'indagine e del questionario che già al momento della stesura deve essere predisposto opportunamente.

1.7 Errori dei dati statistici

In tutte le fasi di un'indagine statistica possono essere commessi degli errori o delle approssimazioni.

Ad esempio obiettivi confusi o mal definiti portano a risultati inutili, perdite di tempo, spreco di risorse, in quanto si traducono in una scelta errata dei caratteri e delle modalità da investigare se non addirittura in una errata individuazione delle unità statistiche.

Spesso è necessario procedere ad una prima indagine campionaria su un numero ridotto di unità statistiche per poter individuare e precisare meglio obiettivi e metodi.

Ad esempio prima di procedere ad una ristrutturazione dei servizi socio-sanitari in una regione può essere utile svolgere una prima indagine orientativa sulle loro carenze per individuare correttamente i campi di intervento; in seguito

Fig. 1.5 - Popolazione presente secondo il sesso, l'età ed il grado di partecipazione al lavoro (migliaia)

TABELLA N. 1	PROVINCIA DI ALESSANDRIA					
	Cifre assolute			Percentuali		
	maschi	femmine	maschi e femmine	maschi	femmine	maschi e femmine
1. Forze di lavoro.....	132	70	202	59.2	29.7	44.0
1.1 Occupati.....	127	65	192	57.0	27.5	41.8
1.2 Persone in cerca di occupazione.....	4	5	10	1.8	2.1	2.2
Di cui disoccupati ed in cerca di 1 ^a occupazione..	3	3	7	1.3	1.3	1.5
2. Non forze di lavoro in età lavorativa (14/70 anni).....	41	102	143	18.4	43.2	31.2
2.1 Che non cercano lavoro ma sono disposte a lavorare a particolari condizioni.....	1	3	5	0.4	1.3	1.1
2.2 Che non cercano lavoro né sono disposte a lavorare.....	39	98	138	17.5	41.5	30.1
3. Non forze di lavoro in età non lavorativa (fino a 13 anni e oltre i 70).....	50	64	115	22.4	27.1	25.1
Totale popolazione presente (1+2+3).....	223	236	459	100.0	100.0	100.0

Fonte ISTAT

verrà effettuata una indagine conclusiva per dimensionare opportunamente l'intervento.

I dati statistici che si ottengono nelle rilevazioni sono tutti soggetti ad *errori di osservazione* qualunque sia il tipo di fenomeno cui si riferiscono (fisico, sociale, economico)

Gli errori dipendono essenzialmente da:

- 1) tipo di fenomeno investigato;
- 2) mezzi di indagine e strumenti di misura;
- 3) osservatore.

Nelle ricerche riguardanti fenomeni sociali possono anche comparire errori dovuti alle reazioni che può manifestare il soggetto osservato.

Quando la rilevazione dei dati viene fatta con strumenti di misura e la rilevazione della stessa grandezza viene ripetuta più volte in condizioni che si presumono identiche, i risultati ottenuti possono essere diversi fra loro. Gli

errori che si verificano in questo caso sono *errori casuali* o *accidentali* e sono dovuti al fatto che fenomeni non controllabili di vario genere si sovrappongono in modo irregolare al fenomeno in esame influenzando in modo casuale i risultati.

Non esiste alcun modo per annullare gli errori casuali, ma ripetendo più volte la stessa misura si può determinare in quale intervallo di valori cade la grandezza misurata.

Gli errori accidentali possono essere dovuti a:

- 1) errori di stima nella lettura;
- 2) condizioni ambientali fluttuanti (temperatura, tensione di rete);
- 3) disturbi meccanici ed elettrici (vibrazioni, scariche, rumore);
- 4) cattiva definizione della grandezza misurata.

Gli *errori casuali* sono trattati e stimati matematicamente dalla *teoria degli errori*. Vi sono altri tipi di errori non trattabili con nessuna teoria generale che costituiscono il vero problema dell'osservatore in quanto possono essere ignoti. Questi errori sono detti *errori sistematici* o *costanti* poiché falsano i risultati agendo algebricamente sempre nello stesso verso.

Gli *errori sistematici* possono essere dovuti a:

- 1) errori di taratura degli strumenti;
- 2) errori personali;
- 3) uso di strumenti in condizioni sperimentali diverse da quelle per cui sono tarati;
- 4) tecniche imperfette;
- 5) teorie imperfette.

Una volta che siano individuati, però, se ne può valutare l'entità e quindi eliminarli. Per determinarli conviene eseguire, se possibile, le misure con metodi diversi; per ridurli si possono usare teorie, strumenti e metodi sempre più raffinati.

Tutti gli errori di cui si è parlato finora sono presenti più o meno in ogni misura ed è compito dell'osservatore minimizzarli.

Esistono poi degli errori imputabili al solo osservatore, sono gli errori di lettura, di progettazione, di approssimazione, di calcolo, uso di formule improprie, errori di trascrizioni nei dati, e così via. Questi errori sono eliminabili cercando di eseguire le misure in modo accurato e l'elaborazione dei dati ottenuti con una certa attenzione.

Qualora la rilevazione dei dati anziché estesa a tutte le unità statistiche della popolazione considerata, venga eseguita solo su un campione, si devono tener presenti gli *errori campionari* ovvero gli errori dovuti alle differenze esistenti fra i risultati di un'indagine campionaria e i risultati che si ottengono con una rilevazione completa.

Tali errori sono inevitabili e diminuiscono all'aumentare delle dimensioni del campione.

Nelle indagini statistiche di tipo sociale sono presenti anche errori dovuti a mancate risposte, risposte incomplete, inesatte o incompatibili, errata interpre-

tazione dei quesiti, errori connessi alle diverse fasi dell'elaborazione.

È generalmente difficile dare una valutazione dell'entità degli errori in quanto ciò richiederebbe la conoscenza di dati di riferimento che non sempre sono disponibili. Per alcuni di questi errori vengono effettuate delle correzioni indicate con il termine *imputazione*.

Tale metodo consiste nell'integrare la risposta mancante o inesatta in un questionario; l'operazione viene eseguita mediante una stima effettuata sulla base di elementi noti.

Durante la rilevazione vi possono essere errori dovuti ad omissioni, duplicazioni, inclusioni improprie di unità statistiche del collettivo investigato; tali errori vengono indicati come *errori di copertura*.

Possibilità di errori sistematici nei dati si hanno anche nelle fasi che seguono la rilevazione come ad esempio nella codificazione dei modelli (mancanza di codificazione, codice errato) oppure durante l'introduzione dei dati nella memoria archivio di un elaboratore (errori di battitura su tastiera, mancata introduzione di dati).

Sono inoltre importanti gli errori che introducono gli intervistatori e gli intervistati; per i primi gli errori sono dovuti a eccessiva fretta nel condurre l'intervista, nell'indisporre l'intervistato, o nel condizionarlo con proprie opinioni; per i secondi gli errori sono dovuti a incomprensioni, pregiudizi, diffidenza verso l'indagine statistica, timori di ripercussioni di natura fiscale.

Nelle analisi statistiche per l'investigazione scientifica e di ricerca operativa si possono avere inoltre errori nella stima dei parametri, nelle assunzioni delle ipotesi, nella costruzione e nella scelta dei modelli e di decisione.

Come si è visto le cause di errore sono molteplici: per alcuni di essi la statistica stessa permette di scoprirli e trattarli, per altri il compito di eliminarli spetta a chi conduce l'indagine statistica e può farlo eseguendo nel modo più accurato possibile le varie fasi dell'indagine.

1.8 Organizzazione della statistica ufficiale in Italia

In tutti i Paesi sviluppati vi è un organismo cui lo Stato con opportune leggi ha attribuito il compito di fornire le statistiche ufficiali.

In Italia tale organismo è un Istituto di Stato a gestione autonoma dotato di personalità giuridica: l'*Istituto Centrale di Statistica* (ISTAT), posto sotto la diretta dipendenza del Consiglio dei Ministri, con sede in Roma.

L'Istituto di Statistica, ha il ruolo fondamentale di raccogliere, elaborare e diffondere informazioni statistiche su tutti gli aspetti della vita dello Stato e della collettività nazionale.

Numerosi sono i compiti dell'ISTAT; in particolare:

- esegue i censimenti e le altre rilevazioni, ricerche e studi necessari per assicurare la disponibilità di dati statistici che interessano il Parlamento, il Governo, le Amministrazioni, le parti sociali, le imprese, gli studiosi e le famiglie;

- fornisce i dati richiesti dagli Organi internazionali;
- predispone le nomenclature di base per la classificazione dei fenomeni demografici ed economici;
- coordina le statistiche degli organi della pubblica amministrazione attraverso l'omogeneizzazione di concetti, definizioni e metodologie.

L'ISTAT prepara e distribuisce i modelli per la raccolta dei dati; nelle fig. 1.6 e 1.7 sono rappresentati due esempi di modelli di rilevazione, relativi ai settori della statistica demografica e della statistica culturale rispettivamente.

Nel 1985 l'organico dell'ISTAT era costituito da poco più di 2800 persone. Le strutture informatiche erano costituite da 4 elaboratori della capacità di memoria complessiva di 32 *Mbyte*; memorie in linea su dischi della capacità globale di 85 *Gbyte*; 250 terminali; 380 tastiere di *data entry*; 700 memorie virtuali.

L'organigramma dell'ISTAT, al cui vertice si colloca il Presidente, si articola in una Direzione Generale e due Direzioni Centrali: la Direzione Centrale dei Servizi Tecnici e la Direzione Centrale degli Studi, Contabilità Nazionale ed Informatica.

Diciotto Uffici Regionali dell'ISTAT assicurano l'assistenza tecnica ai fornitori delle informazioni di base, la revisione e la registrazione dei dati di alcune rilevazioni, la diffusione a livello locale.

Funzionano da organi periferici gli Uffici Provinciali di Statistica presso le Camere di Commercio e gli Uffici Comunali di Statistica dove sono istituiti. Sono 100 mila gli organi della pubblica amministrazione (Comuni, Province, Regioni, Scuole, Ospedali, Dogane, ecc.) che forniscono all'ISTAT dati elementari nell'ambito dei fenomeni dell'area di propria competenza (natalità, nuzialità, mortalità, istruzione, giustizia, commercio estero, ecc.).

A queste informazioni sono da aggiungere quelle ottenute direttamente presso le imprese o le famiglie.

I dati elementari una volta raccolti ed elaborati, vengono diffusi soprattutto tramite le pubblicazioni. Tra esse le più utilizzate sono: l'Annuario, il Compendio, il Bollettino Mensile, i Notiziari, gli Indicatori Mensili, gli Annuari specializzati, oltre naturalmente quelle dedicate ai censimenti.

L'ISTAT potenzia la diffusione facendo tra l'altro ricorso a supporti informatici. È operativo il sistema di interrogazione delle banche dati, cui sono collegati, al momento attuale, oltre ad alcuni uffici regionali dell'ISTAT, le principali istituzioni (Senato, Camera dei Deputati, Presidenza del Consiglio).

È stato messo a punto poi un progetto che prevede in tempi brevi un notevole sviluppo del sistema che consentirà, tra l'altro, il collegamento delle Camere di Commercio e di altri utenti, che potranno disporre delle informazioni anche attraverso società di servizi informatici.

Si richiama ancora una volta l'attenzione sul fatto che l'ISTAT non possiede per legge il monopolio dell'informazione statistica, ma ha il dovere di elaborare la Statistica Ufficiale e garantire la massima obiettività e imparzialità dei risultati.