1. NOZIONI PRELIMINARI

1.1 Introduzione

La Statistica fornisce gli strumenti con cui si acquisisce informazione su un fenomeno di interesse per tradurla in conoscenza. Ovviamente il tipo di informazione da raccogliere dipende dal fenomeno oggetto di studio e dagli obiettivi che si vogliono perseguire. In particolare, la statistica è la disciplina che elabora i principi e le metodologie che presiedono al processo di rilevazione e di raccolta dei dati, alla loro rappresentazione e interpretazione. Un amministratore pubblico, per esempio, può avere bisogno di disporre di informazioni sugli abitanti di una certa zona per decidere gli interventi più opportuni in campo sanitario, scolastico, del trasporto pubblico; una banca può essere interessata a conoscere le caratteristiche dei suoi clienti per migliorare i servizi offerti; un'azienda può aver bisogno di reperire informazioni per produrre articoli che rispondano alle aspettative dei potenziali consumatori e così via. La raccolta di dati può essere anche finalizzata ad esaminare le variazioni di un certo fenomeno in un determinato arco di tempo, come le variazioni nei consumi delle famiglie, l'andamento del PIL o del tasso di inflazione, il tasso di rendimento di un certo titolo in Borsa.

Talvolta la raccolta dei dati avviene in relazione a esperimenti condotti su un insieme di soggetti come ad esempio nel caso in cui si voglia valutare l'efficacia di un farmaco, l'effetto di un fertilizzante, i tempi di conservazione di un alimento in scatola, la quota di articoli difettosi prodotti da un certo macchinario.

Nel caso in cui la raccolta dei dati avvenga sullo stesso insieme di unità a diversi istanti temporali si parla di studi longitudinali, come ad esempio nel caso in cui si consideri la progressione nella carriera di un gruppo di studenti universitari al fine di evidenziare il numero di crediti ottenuti durante il corso di studio.

1.2 Definizione dei termini statistici di uso comune

Si definisce **popolazione** o **collettivo statistico** un insieme di unità omogenee rispetto a una o più caratteristiche. Ciascuna unità del collettivo è anche chiamata **unità statistica** (o più brevemente unità).

Per definire una popolazione è però necessario individuare esattamente le caratteristiche che deve avere un'unità per farne parte. Per esempio, un collettivo di neonati può essere definito indicando il luogo e la data di nascita, specificando se si intende considerare i soli figli legittimi o anche quelli naturali, i soli nati vivi o anche i nati morti e così via.

Alcuni esempi di popolazioni sono: i residenti in un determinato comune italiano il 31 dicembre del 2015, le aziende agricole operanti in Toscana nel mese di giugno del 2001, gli immigrati clandestini arrivati in Italia nel corso dello scorso anno solare, le filiali di uno specifico istituto bancario aperte nel corso di un determinato decennio in Svizzera, i faggi presenti in un parco naturale nell'autunno di un certo anno, i lavoratori in nero dello scorso anno solare, i camosci in Alto Adige nella primavera del 2013.

Va osservato che la singola unità statistica può essere anche costituita da un gruppo di più soggetti, se lo scopo dell'indagine è lo studio di una qualche caratteristica complessiva del gruppo stesso. Se, per esempio, l'oggetto dell'indagine è il reddito delle famiglie italiane, l'unità statistica è la famiglia; se è il numero di dipendenti di un certo gruppo di aziende, l'unità di rilevazione è l'azienda; se è la numerosità dei branchi di una certa specie, l'unità statistica è il branco e così via.

Vi sono popolazioni per le quali è possibile ottenere la lista delle unità che le compongono e altre popolazioni per le quali questo non è possibile. Quest'ultima circostanza si verifica in genere quando le popolazioni oggetto di studio sono di tipo elusivo (immigrati clandestini, lavoratori in nero, turisti che non alloggiano nelle strutture ricettive ufficiali) oppure quando si ha a che fare con popolazioni biologiche (alberi o animali di una certa specie).

Nel caso di popolazioni con lista si può, almeno in linea teorica, effettuare la rilevazione dei dati su ogni unità della popolazione.

Le indagini estese a tutte le unità che compongono la popolazione oggetto di studio vengono indicate generalmente con il termine **censimento** o **rilevazione totale**.

In Italia l'Istituto Nazionale di Statistica (ISTAT) effettua il Censimento generale della popolazione ogni 10 anni. Altre comuni rilevazioni totali, sempre condotte dall'ISTAT, sono il Censimento delle abitazioni (che è generalmente abbinato a quello della popolazione) il Censimento industriale e commerciale ed il Censimento dell'agricoltura.

Tra le cause che possono impedire in pratica l'effettuazione di una rilevazione totale vi sono gli elevati costi dell'indagine, le difficoltà di reperimento delle unità statistiche e di rilevazione dei dati (come, per esempio, in caso di misurazioni su microrganismi) o l'impossibilità di terminare la rilevazione e l'analisi dei dati in tempi brevi.

Altri casi in cui sono possibili solo rilevazioni parziali si presentano quando le unità statistiche per poter essere esaminate devono essere distrutte, come può accadere nei controlli di qualità dei prodotti (durata delle batterie di un telefono cellulare, resistenza alla rottura dei fogli di carta, tempo di ossidazione di una certa sostanza).

Vi sono, infine, situazioni in cui l'indagine non può che basarsi sulle sole unità statistiche che sono effettivamente disponibili, come avviene comunemente nelle ricerche paleontologiche o archeologiche.

Nella quasi totalità delle situazioni concrete, quindi, per tutta una serie di motivi, non è possibile rilevare le informazioni di interesse su tutte le unità che compongono la popolazione, cosicché diventa necessario limitarsi a rilevazioni effettuate su un sottoinsieme delle unità complessive.

Indicata con N la numerosità della popolazione, si definisce **campione** un sottoinsieme di n unità selezionate dalla popolazione ($n \leq N$). Le indagini parziali effettuate su un campione sono dette **indagini campionarie**. In

caso di popolazione con lista, in cui è nota la dimensione *N* della popolazione, il rapporto *n*/*N* è detto **frazione di campionamento**.

Nelle ricerche sperimentali l'interesse può rivolgersi addirittura a collettivi di tipo virtuale, ossia non si considerano più le unità effettivamente presenti in un certo luogo e in certo tempo, ma ci si riferisce piuttosto a tutte le potenziali unità che, per certi aspetti, possono essere considerate di uno stesso tipo. Se, per esempio, si vogliono valutare gli effetti di un farmaco nella cura di una determinata malattia o di un fertilizzante su un certo tipo di pianta, non ha senso pensare di somministrare il farmaco a tutte le unità statistiche che presentano quella particolare malattia o il fertilizzante a tutte le piante di una determinata specie. In questi casi la rilevazione non può che essere effettuata su un campione.

Le informazioni parziali ottenute attraverso un'indagine campionaria, sotto certe condizioni, possono però essere utilizzate per trarre conclusioni circa il fenomeno di interesse nell'intera popolazione.

I metodi statistici vengono infatti distinti in descrittivi e inferenziali dove i primi, che si occupano della raccolta, della presentazione e della sintesi di un insieme di dati, costituiscono l'oggetto della cosiddetta **statistica descrittiva**, mentre i secondi, che consentono di studiare una caratteristica della popolazione o di prendere una decisione che riguarda l'intera popolazione sulla base delle informazioni parziali ottenute su un campione, costituiscono l'oggetto della cosiddetta **inferenza statistica**.

Nelle pagine successive saranno introdotti gli strumenti di statistica descrittiva e quindi si farà genericamente riferimento ad un insieme di *n* unità, senza specificare se costituiscono un campione o la popolazione.

La **variabile** (o **carattere**) X è quel particolare aspetto delle unità statistiche che costituisce l'oggetto dell'indagine.

Esempi di variabili che possono essere rilevate su un insieme di persone sono l'età, lo stato civile, la religione, la professione, il titolo di studio, il reddito, il numero di automobili possedute. Su un insieme di imprese si può voler rilevare il fatturato, gli investimenti, il numero di dipendenti, la quota di mercato. Su un insieme di appezzamenti agricoli si può essere interessati a conoscere la composizione del suolo, l'esposizione al sole, il livello di irrigazione, il clima.

Le possibili manifestazioni che la variabile può assumere sulle singole unità sono dette **determinazioni**.

Così, per esempio, le determinazioni che la variabile "stato civile" può assumere su un insieme di persone di sesso maschile sono: celibe, coniugato, separato, divorziato, vedovo. Per quanto riguarda invece la variabile "titolo di studio" le possibili determinazioni sono: nessun titolo, licenza elementare, licenza media, diploma, laurea.

Analizzando un insieme di *n* unità sulla base di una variabile X, le determinazioni assunte dalla X su ciascuna unità sono dette **osservazioni**. In particolare, indicando con x_i l'*i*-esima osservazione, ovvero la determinazione assunta dalla variabile X sull'*i*-esima unità (con $i = 1, 2, \ldots, n$), la sequenza delle *n* osservazioni $x_1, x_2, \ldots, x_n$ rappresenta l'insieme dei dati osservati.

1.3 Le variabili

Le variabili si suddividono, in relazione alle loro caratteristiche, in variabili **qualitative** (o categoriali) e variabili **quantitative**.

Una variabile si dice qualitativa quando le sue determinazioni (dette anche **modalità** o **categorie**) sono espresse attraverso attributi.

Esempi comuni di variabili qualitative sono il sesso, lo stato civile, il gruppo sanguigno, il colore degli occhi e dei capelli, il titolo di studio, il livello di soddisfazione relativo ad un certo prodotto o servizio.
Va sottolineato che le modalità di una qualsiasi variabile qualitativa risultano sempre incompatibili fra di loro ed esaustive, nel senso che ciascuna di esse non può coesistere con nessuna delle altre e che la lista delle modalità comprende tutti i modi in cui la variabile può manifestarsi.

Le variabili qualitative si suddividono ulteriormente in **ordinabili** e **non ordinabili** (dette anche **sconnesse** o **sparse**) in relazione alla possibilità di stabilire o meno un ordinamento naturale delle modalità.

Per esempio, il titolo di studio è una variabile qualitativa ordinabile in quanto c'è un'ordinamento naturale delle sue modalità, mentre la variabile sesso risulta sparsa.
Altri esempi di caratteri ordinabili sono il ceto sociale, l'anno del corso di studi, la qualifica funzionale degli impiegati, il grado nella gerarchia militare, la "dimensione" delle imprese (piccola, media e grande).
Altri esempi di variabili non ordinabili sono la religione, il colore degli occhi o dei capelli, lo stato civile, il luogo di nascita.

Una variabile si dice **quantitativa** quando le sue determinazioni (dette anche **valori** o **intensità**) sono espresse mediante valori numerici.

Esempi di variabili quantitative sono il reddito, il numero di figli, il rendimento di un titolo azionario, il voto conseguito all'esame di maturità.

Le variabili quantitative si suddividono in **discrete** e **continue**. Una variabile quantitativa si dice **discreta** quando può assumere un insieme finito o numerabile di valori, mentre si dice continua quando può assumere, almeno in teoria, tutti i valori compresi in un intervallo reale o, in altri termini, può assumere una infinità non numerabile di valori diversi.

Le variabili quantitative discrete derivano generalmente da operazioni di conteggio, come ad esempio il numero dei componenti delle famiglie, il numero di veicoli circolanti, il numero di dipendenti di un'azienda e quello degli sportelli bancari.
Le variabili quantitative continue, invece, derivano spesso da operazioni di misurazione, come ad esempio la temperatura, la statura, il peso, l'altitudine, la superficie coltivabile.
In realtà, nelle situazioni concrete, il valore assunto da una variabile continua può essere misurato solo in modo approssimato, con un grado di precisione che dipende dallo strumento di misura utrilizzato e dagli scopi dell'indagine, ed è quindi finito il numero di determinazioni distinte effettivamente rilevate (i valori della temperatura, della statura o del peso di una persona vengono espressi di solito mediante un numero intero di gradi, centimetri o chilogrammi seguito, al più, da poche cifre decimali).

La distinzione fra variabili discrete e continue non dipende dai valori effettivamente rilevati ma piuttosto dalla natura stessa della variabile e questa distinzione risulta importante nell'organizzazione dei dati, nella loro rappresentazione grafica e nelle successive elaborazioni.

Nota
Esistono variabili le cui determinazioni, pur essendo espresse di solito mediante valori numerici, non sono in realtà di tipo quantitativo. L'anno di nascita, l'anno del corso di studi, la categoria degli esercizi alberghieri, la classe di stipendio, per esempio, sono variabili qualitative, dato che non derivano da una misurazione o da un conteggio.

1.4 Scale di misura e cambiamenti di scala

Nelle indagini in cui la variabile è di tipo quantitativo è necessario specificare una qualche "convenzione" che consenta di esprimere i valori rilevati sulle unità statistiche esaminate.
In un'indagine volta a rilevare la variabile "lunghezza" su un insieme di unità statistiche si potrebbe decidere di effettuare le misurazioni in millimetri, centimetri, metri o chilometri, non solo a seconda di quali siano le unità oggetto di indagine, ma anche a seconda del grado di precisione desiderata. La statura potrebbe essere misurata, per esempio, in centimetri oppure in millimetri; il peso corporeo potrebbe essere misurato in chilogrammi, in ettogrammi oppure in grammi; la capienza dei thermos prodotti da un'azienda potrebbe essere misurata in decilitri, centilitri o millilitri.

La convenzione utilizzata per esprimere i valori di una variabile quantitativa è detta **scala di misura**. Per definire la scala di misura è necessario fissare l'**origine** e **l'unità di misura**. L'origine è quel particolare valore che rappresenta lo zero, in modo tale che quantità maggiori e minori di zero abbiano rispettivamente segno positivo e segno negativo. L'unità di misura è la quantità posta essere uguale a uno, in modo che ogni altra quantità (considerata in valore assoluto) possa essere espressa come multiplo dell'unità di misura utilizzata.

In numerosi casi la scelta dell'origine è naturale, come quando si rilevano variabili positive come il peso, la statura o il reddito, ma in altri casi la scelta dell'origine deriva da una convenzione.

Ad esempio, se la variabile di interesse è l'altitudine, per convenzione, lo zero corrisponde al livello del mare (per cui le località al di sotto di tale livello assumeranno valori negativi della variabile), mentre come unità di misura si può scegliere il metro (definito come la distanza che la luce percorre nel vuoto in un tempo pari a 1/299.792.458 di secondo).

Nel caso della temperatura, in Italia si utilizza la scala di misura Celsius che ha l'origine in corrispondenza della temperatura di fusione del ghiaccio in condizioni standard di pressione, mentre l'unità di misura è pari a un centesimo della temperatura dell'acqua bollente. Esistono però scale di misura diverse, come per esempio la scala Fahrenheit, normalmente utilizzata nei paesi anglosassoni (se si volesse confrontare la temperatura rilevata in una località italiana e in una località inglese sarebbe quindi necessario ottenere misurazioni espresse nella stessa scala di misura).

Il cambiamento del sistema di misura si effettua in modo semplice, tenendo conto che ogni scala di misura può essere cambiata in qualsiasi altra mediante una opportuna trasformazione lineare. In particolare, se è X la variabile espressa utilizzando una determinata scala di misura e Y la variabile espressa mediante una diversa scala di misura, allora

$$Y = a + bX, \qquad 1.4.1$$

dove il parametro a (con $-\infty < a < +\infty$) opera una traslazione e produce un cambiamento dell'origine mentre il parametro b (con $b > 0$) è il fattore di scala che modifica l'unità di misura operando una "dilatazione" (quando $b > 1$) oppure una "contrazione" (quando $b < 1$).

Considerato il valore x_i della variabile X rilevato sulla i-esima unità, il corrispondente valore della variabile Y definita nella 1.4 è $y_i = a + bx_i$.

Esempio 1.4.1

La variabile X "temperatura espressa in gradi Fahrenheit" è stata rilevata per quattro giorni in una certa località e sono stati ottenuti i seguenti valori

23 32 50 59

Si calcolino i valori assunti dalla variabile Y "temperatura espressa in gradi Celsius" sapendo che $Y = \frac{5}{9}(X - 32)$.

Effettuando la trasformazione si ottengono immediatamente i seguenti valori

-5 0 10 15

Esempio 1.4.2
Dati i seguenti valori espressi in vecchie lire italiane

1000 5000 100000

si calcolino i valori espressi in euro.

Sapendo che 1 euro equivale a 1936.27 lire italiane, i valori trasformati (arrotondati a due cifre decimali) sono i seguenti

0.52 2.58 51.65

Brevi cenni storici

Secondo alcune interpretazioni, l'etimologia del termine "statistica" deriva dal vocabolo "stato". Le prime rilevazioni statistiche vennero infatti effettuate dagli Stati fin dai tempi più antichi, con lo scopo di ottenere informazioni sulla popolazione nel suo complesso, sul numero di uomini che potevano combattere, sull'estensione del territorio oppure sulla ripartizione delle superfici coltivabili. Gli obiettivi principali di queste rilevazioni erano essenzialmente di natura fiscale, per la stima dei tributi dovuti sulla base dei beni posseduti, e militare, per conoscere il numero di uomini che potevano venire destinati ad attività belliche.

I primi esempi di rilevazioni statistiche sono costituiti dai censimenti (indagini estese a tutta la popolazione) che vennero effettuati già a partire dal IV millennio a.C. dalle antiche civiltà che abitavano in Mesopotamia. Anche nell'antico Egitto vennero effettuate numerose rilevazioni statistiche per ottenere informazioni sulla popolazione per fini fiscali e militari e per il calcolo della manodopera necessaria per effettuare grandi opere (dighe, canali, templi, piramidi).

In Cina e in India già dal III millennio a.C. furono effettuati i primi conteggi della popolazione per valutare le risorse finanziarie delle famiglie. Si hanno notizie certe di un censimento effettuato nel 2.200 a.C. dopo una grave inondazione con lo scopo di conoscere l'estensione del territorio, la ripartizione delle terre coltivabili ed il numero degli abitanti, classificati secondo l'attività ed il mestiere esercitato.

Anche all'interno della Bibbia si trovano notizie di diversi censimenti, a partire da quello effettuato da Mosè nel XIII secolo a.C. dopo l'esodo dall'Egitto, e molte rilevazioni simili vennero effettuate anche dai Greci e dai Romani già alcuni secoli prima della nascita di Cristo (in quello effettuato da Solone ad Atene venne stilata una lista degli elettori basata sul valore della terra in loro possesso ed in quello effettuato da Tucidide si trovano i primi esempi di elaborazioni statistiche, come il calcolo della media aritmetica).

Presso l'antica Roma i cittadini dovevano dichiarare allo Stato il proprio nome, la discendenza paterna, il nome della moglie e dei figli e l'entità dei beni posseduti ed è proprio presso questa civiltà che viene introdotta la periodicità dei censimenti. I romani venivano registrati insieme ai propri beni nelle liste del cosiddetto "census" (da cui deriva il termine censimento) che servivano nell'organizzazione politica ed economica.

La caduta dell'Impero Romano e le invasioni barbariche causarono un'interruzione dei censimenti, anche se alcune rilevazioni statistiche, parziali e incomplete, si registrarono durante tutto il medioevo.

Nel XIII secolo vennero effettuate rilevazioni periodiche della popolazione dai Comuni e dalle Repubbliche italiane. La Repubblica di Venezia istituì il primo censimento verso la metà del 1300, effettuando una rilevazione della popolazione secondo età, professione, sesso, nazionalità e condizione sociale che fu ripetuta periodicamente in tempi successivi mediante questionari simili a quelli attualmente usati.

Il primo censimento ufficiale italiano risale al 1861, subito dopo l'Unità d'Italia. Da quel momento, con cadenza decennale, sono state eseguite tutte le rilevazioni successive (tranne nel 1891 per mancanza di fondi e nel 1941 a causa della seconda guerra mondiale).

Da diversi decenni i censimenti italiani sono effettuati dall'ISTAT (Istituto Nazionale di Statistica).