

RILEVAZIONI STATISTICHE

Per poter entrare nel merito dell'argomento occorre, in primo luogo, riportare alcune definizioni dei termini statistici maggiormente utilizzati nel seguito, molti dei quali coincidono con i vocaboli del linguaggio corrente.

Rilevazione statistica: complesso di operazioni rivolte ad acquisire una o più informazioni su un insieme di elementi oggetto di studio. Le rilevazioni statistiche possono essere classificate:

- rispetto alla complessità delle operazioni: *semplici* (ad esempio, misurare l'altezza di un individuo, chiedere il sesso o la data di nascita ad un impiegato, ...) o *complesse* (ad esempio, codificare un bilancio aziendale, valutare il ritmo di accrescimento di cellule tumorali, ...);
- rispetto alla natura delle informazioni raccolte: *risposte* (ad esempio, opinioni, informazioni personali, gusti, ...) o *misure* (ad esempio, metro, bilancia, orologio, ...);
- rispetto al gruppo di riferimento: *globali* (ad esempio, i censimenti, lo studio di tutti i laureati di un certo Ateneo, ...) o *parziali* (ad esempio, i sondaggi di opinione, le interviste telefoniche, ...).

Popolazione (o Universo): qualsiasi insieme di elementi che forma l'oggetto di uno studio statistico.

La popolazione può essere:

- *reale*, quando essa è effettivamente esistente e visibile (ad esempio, le lampadine prodotte nell'ultimo mese da un'azienda di Milano, le stelle della Via Lattea, ...);
- *virtuale*, quando essa non è osservata né è osservabile perché astratta o connessa al futuro, ma è comunque ben definita (ad esempio, gli acquirenti di un certo modello di automobile che si sta progettando, gli studenti che il prossimo anno supereranno l'esame di matematica, ...)

Campione: qualsiasi sottoinsieme derivato da una certa popolazione e finalizzato ad uno studio statistico. A tal riguardo, però, va chiarito che la dicotomia popolazione-campione può essere risolta solo in funzione degli obiettivi dell'indagine: si parla di *popolazione*, infatti, quando il collettivo di riferimento esaurisce tutte le informazioni che si ritengono utili per l'indagine statistica; si parla, invece, di *campione*, quando tali informazioni sono derivate da un sottoinsieme proprio della popolazione di riferimento (ad esempio, i residenti del comune di Firenze sono un campione degli italiani ma sono anche la popolazione dei residenti a Firenze; l'analisi delle caratteristiche di tali elementi, quindi, sarà svolta con metodologie differenti, a seconda che l'indagine punti a studiare la collettività dei fiorentini o quella degli italiani; del resto, poiché la conoscenza umana non può che

essere parziale o provvisoria, la Statistica privilegia un approccio allo studio dei fenomeni che presuppone sempre una dimensione campionaria).

Unità statistica: elemento di base della popolazione sul quale viene effettuata la rilevazione di uno o più fenomeni oggetto dell'indagine, ovvero il più piccolo elemento su cui è possibile effettuare un'osservazione; le unità statistiche si possono distinguere in *semplici* (una singola persona, una singola abitazione, ...), *composte* (una famiglia intesa come insieme dei suoi componenti, un edificio inteso come insieme di abitazioni, ...), *complesse* (il rapporto coniugale di cui sono unità semplici il marito e la moglie, un determinato lavoro di cui sono unità semplici il datore di lavoro ed i dipendenti, ...).

Dato statistico: risultato di un'operazione compiuta sulle unità statistiche (numero di alunni maschi di una classe, prezzo medio di un determinato bene, ...).

Carattere: un aspetto particolare, che si desidera mettere in evidenza, relativo all'unità statistica in esame.

Variabile statistica: fenomeno oggetto dello studio, rilevato o misurato sulle unità statistiche; le variabili possono essere *qualitative* o *quantitative*.

Modalità: espressione concreta (*numero* o *attributo*) mediante la quale la variabile si manifesta nelle unità statistiche. L'elenco di tutte le possibili modalità di una variabile si dice *esaustivo* se tale elenco è completo; le modalità si dicono *esclusive* o *disgiunte* se un'unità statistica può manifestare la variabile in una ed una sola modalità tra quelle indicate; le modalità si dicono *quantitative* se sono espresse tramite valori risultanti da misurazioni o conteggi (se tali valori sono compresi in un intervallo si parlerà di modalità *continue*; se invece si tratta di determinati valori si parlerà di modalità *discrete*); le modalità si dicono *qualitative* se sono individuate tramite espressioni verbali (ad esempio, gli anni di nascita, il grado delle scuole, le marche di automobili, ...).

Ne segue che:

le variabili qualitative possono essere:

- *nominali*: se le modalità (attributi) non assumono alcun ordine preconstituito (ad esempio, il sesso, la religione, la nazionalità delle persone, ...);
- *ordinali*: se le modalità (attributi) sono logicamente sequenziali in ordine crescente o decrescente, ovvero se si può stabilire tra di esse una relazione d'ordine rispetto alla variabile individuata (ad esempio, il grado militare, la valutazione scolastica, ...).

le variabili quantitative possono essere:

- *discrete*: se le modalità possono essere poste in corrispondenza con l'insieme dei numeri naturali o con un suo sottoinsieme proprio, ovvero le variabili discrete assumono un insieme

numerico di modalità elencabili (ad esempio, il numero di clienti in attesa ad uno sportello bancario, il voto dell'esame di matematica, il numero dei componenti di una famiglia, ...);

- *continue*: se le modalità possono assumere qualsiasi valore contenuto in un intervallo reale (ad esempio, l'età, il peso, l'altezza, la temperatura di una stanza, ...).

In genere, una singola variabile si indica con la lettera X e le modalità che essa assume con x_1, x_2, \dots . Quando le variabili sono più di una si parla di variabili doppie, triple, ..., multiple, indicate rispettivamente con $(X, Y), (X, Y, Z), \dots, (X, Y, Z, \dots, W)$, e per le modalità che esse assumono sulle singole unità statistiche si parla di:

- *coppie ordinate*: $(x_1, y_1), (x_2, y_2), \dots$
- *terne ordinate*: $(x_1, y_1, z_1), (x_2, y_2, z_2), \dots$
- *m-uple ordinate*: $(x_1, y_1, z_1, \dots, w_1), (x_2, y_2, z_2, \dots, w_2), \dots$

Un dato fenomeno, inoltre, si definisce:

- *variabile* (o *variabile quantitativa*) se assume per modalità dei numeri reali;
- *mutabile* (o *variabile qualitativa*) se assume attributi non numerici ma di altra natura.

Nel seguito, quindi, si parlerà di *variabili quantitative* se le modalità sono numeri e di *variabili qualitative* in tutti gli altri casi. Ad esempio, l'età, il peso, la temperatura, il voto di un esame, ... costituiscono fenomeni di natura quantitativa (*fenomeni variabili*); il sesso, la religione, lo stato civile, la professione, ... costituiscono, invece, fenomeni di natura qualitativa (*fenomeni mutabili*).

Frequenza: numero di volte in cui una determinata modalità si verifica nel collettivo di riferimento in un certo periodo di tempo (ad esempio, un contatore che esegue un conteggio). Una frequenza si dice *assoluta* se è data dal numero di volte in cui un evento si verifica, a prescindere dal numero totale di prove eseguite, *relativa* se è data dal rapporto tra la frequenza assoluta ed il numero totale delle unità statistiche ovvero delle prove eseguite, e *percentuale* se è data dalla frequenza relativa moltiplicata per 100.

Serie: insieme delle modalità rilevate sulla popolazione di riferimento organizzate in modo che a ciascuna unità della popolazione corrisponda una ben definita modalità, ovvero se in essa sono espresse le modalità di un carattere *qualitativo*.

Seriazione: insieme delle modalità di una popolazione di riferimento organizzate in modo che a ciascuna unità della popolazione corrisponda la rispettiva frequenza. In pratica, quando si dispone di rilevazioni sulle singole unità statistiche isolatamente considerate si ha una *serie* (territoriale, temporale, ...); quando, invece, queste informazioni sono organizzate in tabelle nelle quali alle modalità della variabile corrispondono le rispettive frequenze, ovvero qualora vengano espresse le modalità di un carattere *quantitativo*, si è in presenza di una *seriazione*.

Esempi.

- 1) La registrazione “Adriano è alto cm. 173” costituisce una rilevazione sull’unità statistica “Adriano” per il quale la variabile qualitativa “Sesso” si manifesta nella modalità (=attributo) “Maschio” e la variabile quantitativa “Altezza” si manifesta nella modalità (=numero) “cm.173”.
- 2) Se si considera la distribuzione di 30 giovani presenti in una sala giochi in base all’età, secondo quanto riportato nella seguente tabella:

ETÀ	NUMERO
15	7
16	8
17	5
18	4
19	6

si può facilmente osservare che l’“età” è il carattere, “15,16,17,18,19” sono le modalità quantitative discrete, “7,8,5,4,6” sono le frequenze.

- 3) Se si considera la distribuzione delle auto consegnate da un concessionario nel 2001, secondo quanto riportato nella seguente tabella:

MARCA/MODELLO	CONSEGNE
Fiat Punto	144017
Lancia Y	57540
Toyota Yaris	42481
Peugeot 206	42141
Volkswagen Golf	16849
Daewoo Matiz	19240
Renault Megane	17258
Opel Corsa	40416

si può facilmente osservare che la “marca/modello” è il carattere, “Fiat Punto, Lancia Y, Toyota Yaris, Peugeot 206, Volkswagen Golf, Daewoo Matiz, Renault Megane, Opel Corsa” sono le modalità qualitative, “144017, 57540, 42481, 42141, 16849, 19240, 17258, 40416” sono le frequenze.

- 4) Se in un’aula ci sono 25 studenti per i quali si registra la variabile qualitativa “Sesso” e se la rilevazione ha fornito i seguenti risultati:

{M F F F M F M F M M F F F F M F F M F M F M F M F}

allora la precedente elencazione costituisce una serie della variabile “Sesso” nella popolazione in esame; se, invece, gli stessi dati vengono organizzati in una tabella ove si fanno corrispondere alle due possibilità (M, F) le relative frequenze, allora ci si trova di fronte ad una seriazione, detta anche *distribuzione di frequenza*:

Modalità	Frequenze assolute	Frequenze relative	Frequenze percentuali
Femmine	15	$\frac{15}{25} = \frac{3}{5}$	$\frac{3}{5} \cdot 100 = 3 \cdot 20 = 60\%$
Maschi	10	$\frac{10}{25} = \frac{2}{5}$	$\frac{2}{5} \cdot 100 = 2 \cdot 20 = 40\%$
Totale	25	$\frac{3}{5} + \frac{2}{5} = 1$	60%+ 40%=100%

- 5) Se si considera la distribuzione degli studenti italiani in base al grado delle scuole, come riportato nella seguente tabella:

GRADI DELLE SCUOLE	NUMERI ALUNNI ISCRITTI
Scuole Materne	1500000
Scuole Elementari	2900000
Scuole Medie	1775000
Scuole Superiori	2545000

allora si può osservare che, essendo i “gradi delle scuole” modalità qualitative, si è in presenza di una serie statistica.

- 6) Se si considera la distribuzione degli elettori nelle elezioni del Parlamento Europeo, per circoscrizione, come riportato nella seguente tabella:

CIRCOSCRIZIONI ELETTORALI	ELETTORI
Nord Occidentale	12702945
Nord Orientale	8879456
Centrale	9445290
Meridionale	11654099
Insulare	5576311

allora si può osservare che, essendo le “circoscrizioni elettorali” modalità qualitative, si è in presenza di una serie territoriale o geografica.

- 7) Se si considera la rilevazione delle abitazioni di nuova costruzione, classificate per stanze, come riportato nella seguente tabella:

NUMERO STANZE	NUMERO ABITAZIONI
1	12702945
2	8879456

3 – 4	9445290
5 e oltre	11654099

allora si può osservare che, essendo il “numero delle stanze” modalità quantitative discrete, si è in presenza di una seriazione discreta.

- 8) Se si considera la distribuzione dei comuni italiani per classi di superficie nel 1991, come riportato nella seguente tabella:

CLASSI DI SUPERFICIE	NUMERO COMUNI
fino a 1	1741
1 – 2	2061
2 – 4	2093
4 – 6	883
6 – 10	742
10 – 25	513

allora si può osservare che, essendo le “classi di superficie” modalità quantitative continue, si è in presenza di una seriazione continua.

Informazione statistica: ogni risultato ottenuto da un’indagine sui collettivi esaminati (popolazione o campione) rispetto ai loro costituenti (unità statistiche) ed in rapporto ad uno o più fenomeni (variabili).

Anche per le informazioni statistiche multiple, si parla di:

- *serie* quando le informazioni vengono presentate distinte per ciascuna unità statistica;
- *seriazione* quando le informazioni vengono raccolte per modalità e corrispondenti frequenze.

In tale ambito, poi, è molto diffusa la cosiddetta *matrice dei dati*, rappresentazione tabellare mediante la quale si schematizzano le informazioni (misure, registrazioni o risposte) raccolte su ciascuna unità statistica in rapporto ad una molteplicità di variabili. La matrice dei dati, quindi, costituisce la serie dei dati su una variabile statistica multipla in cui ogni *colonna* esprime una variabile, qualitativa o quantitativa, rilevata sulle diverse unità statistiche ed ogni *riga* esprime ordinatamente le rilevazioni su tutte le variabili ottenute per una singola unità statistica.

Esempio.

Il questionario rivolto agli studenti frequentanti il corso di Matematica contiene numerose domande di natura qualitativa e quantitativa. Le risposte fornite da ciascuno studente costituiscono una riga della matrice dei dati; in ogni colonna, invece, sono contenute le informazioni sulla specifica variabile fornita da tutti gli studenti.

Una delle principali operazioni statistiche elementari che consente un primo esame dei dati statistici, è rappresentata proprio dalla *differenza* tra le modalità x_1 ed x_2 di una variabile quantitativa X , che può essere:

- *assoluta*, definita da:

$$x_2 - x_1$$

- *relativa*, pari alla differenza assoluta, rapportata ad una delle due modalità, generalmente la prima:

$$\frac{x_2 - x_1}{x_1}$$

- *percentuale*, pari alla differenza relativa moltiplicata per 100:

$$\frac{x_2 - x_1}{x_1} \times 100$$

Dunque, se la *differenza assoluta* è utile per il confronto tra fenomeni simili in circostanze differenti, la *differenza relativa* risulta necessaria per confrontare le variazioni di fenomeni espressi in unità di misure differenti e la *differenza percentuale* è solo un modo di effettuare i confronti relativi in rapporto al numero 100.

Rispetto poi alla mera elencazione di un'indagine, è preferibile presentare, al fine di agevolare i confronti e le analisi successive, le informazioni statistiche, ovvero i dati, in una forma organizzata, precisamente in forma:

- *enumerativa*: quando le informazioni statistiche sono disponibili secondo la sequenza delle modalità relative alle singole unità statistiche (è il caso delle serie);
- *tabellare*: quando le informazioni statistiche sono presentate secondo un raggruppamento di modalità evidenziato in tabelle, semplici o complesse (è il caso delle seriazioni);
- *grafica*: quando le informazioni statistiche sono presentate mediante disegni e schemi grafici, di natura diversa, in funzione della disponibilità dei dati e delle finalità dell'indagine.