LA STATISTICA

Daniela Tondini

dtondini@unite.it

Facoltà di Medicina Veterinaria

C.L.M in Medicina Veterinaria

Università degli Studi di Teramo

FENOMENI COLLETTIVI

I collettivi statistici sono alla base dello studio dei cosiddetti *fenomeni collettivi*, ovvero di quei fenomeni naturali o sociali (ammontare della popolazione, grado di istruzione, produzione agricola, …) la cui conoscenza e misura richiede l'osservazione delle diverse unità che fanno parte del collettivo.

Sulla base di tale concetto, quindi, si può affermare che la statistica è un insieme di metodi per lo studio dei fenomeni collettivi, ovvero delle caratteristiche che nei collettivi statistici si manifestano in modo variabile in seguito all'influenza di varie circostanze.

Il collettivo statistico è, dunque, l'insieme che si studia (le aziende); il fenomeno collettivo è l'aspetto particolare che interessa studiare del collettivo (il numero degli addetti).

Il collettivo statistico osservato può comprendere tutte le unità omogenee rispetto ad una caratteristica comune e lo si può indicare, in tal caso, col termine *popolazione*. Si va ad osservare, pertanto, l'intera popolazione o un *campione* della popolazione qualora ci sia difficoltà ad osservare tutte le unità statistiche.

LE FASI DI UN'INDAGINE STATISTICA

Lo studio di un fenomeno con metodo statistico, ovvero l'*indagine statistica*, si può articolare in quattro fasi:

- *rilevazione*: insieme di operazioni con le quali si perviene alla conoscenza dei dati ossia delle modalità di uno o più caratteri collettivi; la rilevazione è *completa* se si esaminano tutti gli elementi oggetto di studio, e *parziale* se, invece, ci si limita a studiare un sottoinsieme, ovvero un *campione*, dell'insieme di riferimento;
- *elaborazione*: insieme di operazioni attraverso le quali i dati rilevati (dati originari o grezzi) vengono opportunamente classificati e sintetizzati al fine di ottenere dati più espressivi (dati derivati);
- *presentazione*: esposizione dei dati statistici in forma chiara e compatta, con tabelle, grafici, medie, indici, …
- *interpretazione*: spiegazione delle risultanze dell'indagine statistica alla luce delle teorie e delle precedenti conoscenze del fenomeno studiato o di altri fenomeni ad esso connessi.

Si osservi che la seconda e la terza fase hanno caratteri squisitamente tecnico-statistico; la prima e l'ultima, invece, richiedono la conoscenza, non solo del metodo statistico, ma anche del fenomeno studiato.

LA RILEVAZIONE STATISTICA

In particolare la *rilevazione statistica*, ovvero quel complesso di operazioni rivolte ad acquisire una o più informazioni su un insieme di elementi oggetto di studio, può essere classificata:

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

LA RILEVAZIONE STATISTICA

Popolazione (*o Universo*) è un 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 è un qualsiasi sottoinsieme derivato da una certa popolazione e finalizzato ad uno studio statistico. Si parla di *popolazione*, quindi, 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 costituiscono un campione degli italiani ma sono anche la popolazione dei residenti a Firenze; l'analisi delle caratteristiche di tali elementi, pertanto, sarà svolta con metodologie differenti, a seconda che l'indagine punti a studiare la collettività dei fiorentini o quella degli italiani).

La Statistica privilegia un approccio allo studio dei fenomeni che presuppone sempre una dimensione campionaria.

DUE DIVERSE «STATISTICA»

All'interno della disciplina metodologica, inoltre, si possono distinguere due diverse correnti: la statistica *descrittiva* e la statistica *inferenziale*.

Con il termine di *Statistica Descrittiva* si intende un insieme di tecniche e strumenti finalizzati ad assolvere uno dei principali compiti assegnati alla Statistica, ovvero descrivere, rappresentare e sintetizzare in maniera opportuna un insieme o campione di dati relativamente ad un problema; tale branca, che ha come obiettivo quello di organizzare, riassumere e presentare i dati in modo ordinato attraverso strumenti di tipo sia grafico che numerico, si occupa di fotografare una data situazione e di sintetizzarne le caratteristiche salienti, ovvero di descrivere ciò che si osserva o ciò che i dati evidenziano nei loro tratti essenziali. Tale corrente tende ad evidenziare le regolarità presenti nei dati.

La ***Statistica Inferenziale*** o ***Inferenza Statistica***, invece, comprende le tecniche matematiche per quantificare il processo di apprendimento tramite l'esperienza; utilizza i dati statistici, anche opportunamente sintetizzati dalla statistica descrittiva, per fare previsioni di tipo probabilistico su situazioni future o comunque incerte; con la statistica inferenziale, quindi, si cerca di raggiungere conclusioni che si estendono oltre i dati raccolti nel loro immediato e che possono essere valide e riferibili ad un contesto più ampio rispetto a quello dei dati di quel singolo esperimento. Tale corrente tende a giustificare le osservazioni in termini di modelli teorici esplicativi dei fenomeni.

INDICI STATISTICI

Nella ricerca scientifica e tecnologica è importante misurare la reale efficacia di interventi sul sistema oggetto di studio, ovvero valutare gli effetti complessivi indotti da una causa nota, pur nella mutevolezza ed instabilità dei risultati individuali. A tal riguardo, la Statistica ha proposto numerosi *indici statistici*, aventi quale obiettivo proprio la misurazione di due componenti del fenomeno oggetto di studio e di interesse scientifico: la *consistenza della sistematicità*, cioè la *centralità*, ovvero l'attitudine che hanno i fenomeni ad assumere tendenzialmente una certa dimensione all'osservazione, e la *variabilità* o *mutabilità*, cioè la *dispersione*, ovvero l'attitudine che hanno i fenomeni ad assumere dimensioni e tendenze diverse all'osservazione, nel tempo e nello spazio.

In particolare, la *centralità* è misurata dai cosiddetti *indici di posizione* (o *indici di tendenza centrale* o *indicatori di posizione* o *misure di tendenza centrale*) o *medie statistiche* o ancora più semplicemente *medie*, in grado di esprimere e sintetizzare la posizione di una distribuzione di frequenza mediante un valore reale rappresentativo della globalità del fenomeno, riassumendone gli aspetti ritenuti più importanti.

INDICI STATISTICI

Tali indici si possono ricavare effettuando operazioni che coinvolgono:

- tutti i termini della serie; in tal caso gli indici di posizione maggiormente usati, denominati *medie analitiche* o *di calcolo*, sono la media aritmetica M_a, la media geometrica M_g, la media armonica M_h e la media quadratica M_p tra le quali sussiste la seguente relazione:

$$M_h \leq M_g \leq M_a \leq M_p$$

- solo alcuni termini della serie, che si differenziano dagli altri per particolari caratteristiche; in tal caso gli indici di posizione maggiormente usati, denominati *medie posizionali* o *di posizione* o *lasche*, sono la mediana, la moda, i quartili.

INDICI STATISTICI

La ***media aritmetica semplice***, denominata semplicemente *media* ed indicata con M_a, usata per riassumere con un solo numero un insieme di n dati relativi ad un fenomeno misurabile, ovvero in presenza di variabili quantitative qualora la differenza tra un dato ed il precedente risulti costante, è ottenuta dividendo la somma di tutti gli n valori per il numero n di osservazioni; in formule è data da:

$$M_a = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

avendo indicato con n_i le frequenze delle x_i.
La media aritmetica di n numeri, dunque, è quel numero che, sostituito a ciascuno di essi, lascia invariata la somma totale e non può essere maggiore del valore più grande né minore del valore più piccolo.

INDICI STATISTICI

Esempio

La media aritmetica dei seguenti 5 = n numeri:

$$x_1 = 10; \quad x_2 = 13; \quad x_3 = 9; \quad x_4 = 7; \quad x_5 = 12$$

è data da:

$$M_a = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{1}{5}(10+13+9+7+12) = \frac{1}{5}(51) = \frac{51}{5} = 10,2$$

Si osservi che, sostituendo a ciascun x_i (i = 1, …, 5) il valore della media M_a e sommando i risultati, si ottiene;

$$10,2+10,2+10,2+10,2+10,2 = 5 \cdot M_a = 5 \cdot 10,2 = 51$$

che è proprio la somma degli x_i, 10 + 13 + 9 + 7 + 12 = 51.

INDICI STATISTICI

La ***media aritmetica ponderata***, invece, è ottenuta dividendo la somma di tutti gli *n* valori, moltiplicati per le rispettive frequenze, per il numero *n* di osservazioni; in formule è data da:

$$M_a = \frac{1}{n}\sum_{i=1}^{s} x_i n_i = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_n n_s}{n}$$

avendo indicato con n_i le frequenze delle x_i e con n la somma delle n_i.
Tale denominazione deriva dal fatto che, a volte, le n_i non esprimono le frequenze ma opportuni *pesi di ponderazione* che tengono conto di altri aspetti rilevanti: -basti pensare, ad esempio, ai prezzi delle merci che vengono ponderati con cifre che esprimono le quantità vendute di ciascuna merce, allo scopo proprio di tener conto del valore globale (prezzo per quantità) degli scambi effettuati sul mercato considerato.

INDICI STATISTICI

Esempio

Se i voti riportati in matematica da $n = 20$ alunni di una scuola media di secondo grado sono riassunti nella seguente tabella:

Voti x_i	Alunni n_i
3	1
4	2
5	5
6	7
7	4
8	1
Tot.	20

allora la media aritmetica è data da:

$$M_a = \frac{1}{n}\sum_{i=1}^{s} x_i n_i = \frac{3\cdot 1 + 4\cdot 2 + 5\cdot 5 + 6\cdot 7 + 7\cdot 4 + 8\cdot 1}{20} = \frac{114}{20} = 5,7$$

INDICI STATISTICI

Se poi la v.s. *X* è divisa in intervalli, si può fare l'ipotesi che le intensità di *X* di ogni intervallo siano concentrate nel valore centrale della classe, in modo da riportarsi al caso discreto.

Esempio

Calcolare la statura media (aritmetica) dei coscritti italiani nati nel 1955.

Classi di statura (in cm)	**Valori centrali delle classi** x_i	**Frequenze** n_i	**Prodotti** $x_i * n_i$
meno d1 150	145	300	43500
150 ---160	155	12200	1891000
160 ---170	165	120800	19932000
170 ---180	175	160400	28070000
180 e oltre	185	36300	6715500
Tot.		**330000**	**56652000**

$$M_a = \frac{1}{n}\sum_{i=1}^{s} x_i n_i = \frac{56652000}{330000} = 171,67cm$$

La sostituzione delle singole classi con il valore centrale introduce un errore di approssimazione poco rilevante, anche se, tuttavia, si perde informazione.

INDICI STATISTICI

La media aritmetica, quindi, rappresenta quel valore che si può attribuire singolarmente a ciascuna unità statistica del collettivo lasciando invariato l'ammontare complessivo del carattere.
La media aritmetica di *n* numeri, dunque, rappresenta il *baricentro* dei dati e, quindi, propone un valore che equi-ripartisce il fenomeno tra le unità statistiche, pervenendo così a decisioni nelle quali contano, a parità numerica, gli estremi molto più dei valori centrali: la media aritmetica, infatti, costituisce un indice di equilibrio generale. Essendo, inoltre, la media statistica per eccellenza, consente un'ottima correzione degli errori accidentali commessi in una rilevazione statistica, risultando così utile, nonostante la sua scarsissima resistenza ai valori eccezionali, in tutti i campi della scienza e della tecnica in cui vengono effettuate misurazioni di qualunque genere.
Se la media coincide con una delle modalità viene detta *media effettiva* o *reale*; se, invece, non coincide con una delle modalità è detta *media di conto*.

INDICI STATISTICI

La ***media geometrica semplice***, usata quando le variabili quantitative risultano non lineari ma ottenute da un prodotto o da un rapporto di valori lineari non negativi e diversi da zero, si ottiene estraendo la radice *n*-esima del prodotto degli *n* termini; in formule è data da:

$$M_g = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

dove Π è il simbolo di prodotto.

La media geometrica, considerata come quel valore che sostituito a ciascuno degli *n* dati ne lascia inalterato il prodotto, è usata soprattutto quando i dati non sono numerosi, i termini della distribuzione presentano valori molto differenti tra loro ed il rapporto tra un dato ed il precedente risulta costante (ad esempio, la determinazione del tasso di interesse medio equivalente alla sequenza dei tassi variabili, nel regime di capitalizzazione composta).

INDICI STATISTICI

Esempio

Uno studente ha sostenuto 6 = n esami riportando i seguenti voti:

$$x_1 = 21; \quad x_2 = 20; \quad x_3 = 24; \quad x_4 = 30; \quad x_5 = 28; \; x_6 = 25$$

La media geometrica dei voti è data da:

$$M_g = \sqrt[6]{\prod_{i=1}^{6} x_i} = \sqrt[6]{21 \cdot 20 \cdot 24 \cdot 30 \cdot 28 \cdot 25} = \sqrt[6]{211680000} \approx 24,41$$

INDICI STATISTICI

La ***media geometrica ponderata*** è usata, invece, qualora ci si trovi in presenza di una distribuzione costituita da *n* osservazioni e dalle relative frequenze; in formule, è data da:

$$M_g = \sqrt[n]{\prod_{i=1}^{s} x_i^{n_i}} = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_s^{n_s}}$$

dove Π è il simbolo di prodotto ed $n = n_1 + n_2 + \ldots + n_s$.
Ogni termine, dunque, viene ponderato, ad esponente, con la relativa frequenza.
Analogamente, si può utilizzare la seguente formula:

$$M_g = 10^{\left(\frac{\sum_{i=1}^{s} n_i \log x_i}{n} \right)}$$

INDICI STATISTICI

Esempio

La seguente tabella riporta i voti ottenuti da un gruppo di studenti all'esame di Matematica:

Voti x_i	Numeri di studenti n_i
21	5
24	6
26	10
30	4
Tot.	25

La media geometrica ponderata è data da:

$$M_g = \sqrt[25]{21^5 \cdot 24^6 \cdot 26^{10} \cdot 30^4} =$$

$$= \sqrt[25]{4084101 \cdot 191102976 \cdot 141167095653376 \cdot 810000} \approx 25,00479$$

INDICI STATISTICI

Analogamente, utilizzando i logaritmi, si può impostare la seguente tabella:

Voti x_i	Numeri di studenti n_i	Logaritmi dei voti $\log x_i$	Prodotti $n_i \bullet \log x_i$
21	5	1,322219	6,611096
24	6	1,380211	8,281267
26	10	1,414973	14,149733
30	4	1,4771121	5,908485
Tot.	25		34,950582

Essendo, poi,

$$\frac{\sum_{i=1}^{4} n_i \log x_i}{n} = \frac{34,950582}{25} = 1,398023297$$

si ha la seguente media geometrica ponderata:

$$M_g = 10^{1,398023297} = 25,00479$$

INDICI STATISTICI

La ***media armonica semplice***, usata nello studio di variabili quantitative tra loro inversamente proporzionali, ovvero quando si deve trovare il valore medio, non del fenomeno considerato, ma di un fenomeno che è l'inverso del primo (ad esempio, prezzo di un bene e potere di acquisto della moneta, interesse effettivo che cresce al decrescere del costo del titolo, …), è pari al reciproco della media aritmetica dei reciproci dei termini; in formule è data da:

$$M_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

La media armonica, dunque, è quel valore tale che il suo reciproco, sostituito ai dati, che devono essere tutti positivi, fa rimanere invariata la somma dei reciproci dei dati stessi: viene usata, infatti, per mediare rapporti di tempo.

INDICI STATISTICI

Esempio

La media armonica dei seguenti $5 = n$ numeri:

$$x_1 = 10; \quad x_2 = 13; \quad x_3 = 9; \quad x_4 = 7; \quad x_5 = 12$$

è data da:

$$M_h = \frac{5}{\sum_{i=1}^{5} \frac{1}{x_i}} = \frac{5}{\frac{1}{10} + \frac{1}{13} + \frac{1}{9} + \frac{1}{7} + \frac{1}{12}} =$$

$$= \frac{5}{\frac{1638 + 1260 + 1820 + 2340 + 1365}{16380}} = \frac{5}{\frac{8423}{16380}} = 5 \cdot \frac{16380}{8423} \approx 9,72$$

INDICI STATISTICI

La ***media armonica ponderata***, invece, è data da:

$$M_h = \frac{n}{\sum_{i=1}^{s} \frac{n_i}{x_i}} = \frac{n}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \ldots + \frac{n_s}{x_s}}$$

dove $n = n_1 + n_2 + \ldots + n_s$.
La media armonica, dunque, è pari al valore reciproco della media aritmetica dei reciproci dei termini.

INDICI STATISTICI

Esempio

Si consideri la seguente tabella la seguente tabella:

Voti x_i	Numeri di studenti n_i
20	2
21	3
22	6
23	2
24	1
Tot.	14

INDICI STATISTICI

Ne segue, allora, che la media armonica ponderata è data da:

$$M_h = \frac{n}{\sum_{i=1}^{s} \frac{n_i}{x_i}} = \frac{14}{\frac{2}{20} + \frac{3}{21} + \frac{6}{22} + \frac{2}{23} + \frac{1}{24}} = 22$$

INDICI STATISTICI

La ***media quadratica semplice*** si ottiene estraendo la radice quadrata della media aritmetica dei quadrati degli *n* termini; in formule è data da:

$$M_2\left(x_1, x_2, \ldots, x_n\right) = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \sqrt[2]{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}}$$

Tale media, denominata anche *media di precisione*, usata tutte le volte che alle differenze tra i termini ed il valore medio si dà il significato di *deviazione* o *errore del valore esatto*, ovvero nei casi in cui alcuni termini considerati risultano negativi e si desidera quindi eliminare la loro influenza, trova applicazione soprattutto nell'ambito della teoria degli errori.
Generalizzando ora il concetto di media quadratica, si può definire la cosiddetta ***media di potenza di indice t*** data da:

$$M_t\left(x_1, x_2, \ldots, x_n\right) = \sqrt[t]{\frac{1}{n}\sum_{i=1}^{n} x_i^t} = \sqrt[t]{\frac{x_1^t + x_2^t + \ldots + x_n^t}{n}}$$

INDICI STATISTICI

Esempio

La media quadratica dei seguenti 10 = n numeri:

$x_1 = 1; x_2 = 1; x_3 = 2; x_4 = 2; x_5 = 3; x_6 = 3; x_7 = 4; x_8 = 4; x_9 = 5; x_{10} = 5$

è data da:

$$M_2\left(x_1, x_2, ..., x_{10}\right) = \sqrt{\frac{1}{10}\sum_{i=1}^{10} x_i^2} =$$

$$= \sqrt{\frac{1}{10}\left(1^2 + 1^2 + 2^2 + 2^2 + 3^2 + 3^2 + 4^2 + 4^2 + 5^2 + 5^2\right)} =$$

$$= \sqrt{\frac{1}{10}\left(1 + 1 + 4 + 4 + 9 + 9 + 16 + 16 + 25 + 25\right)} =$$

$$= \sqrt{\frac{1}{10}\left(1 + 1 + 4 + 4 + 9 + 9 + 16 + 16 + 25 + 25\right)} =$$

$$= \sqrt{\frac{110}{10}} = \sqrt{11} \approx 3,31$$

INDICI STATISTICI

La ***media quadratica ponderata***, invece, è data da:

$$M_2\left(x_1, x_2, \ldots, x_n\right) = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{s} x_i^2 n_i} = \sqrt[2]{\frac{x_1^2 \cdot n_1 + x_2^2 \cdot n_2 + \ldots + x_n^2 \cdot n_s}{n}}$$

dove n è sempre la somma delle n_i. La precedente espressione, generalizzata alle potenze di indice t, diventa:

$$M_t\left(x_1, x_2, \ldots, x_n\right) = \sqrt[t]{\frac{1}{n}\sum_{i=1}^{s} x_i^t n_i} = \sqrt[t]{\frac{x_1^t \cdot n_1 + x_2^t \cdot n_2 + \ldots + x_n^t \cdot n_s}{n}}$$

dove n è sempre la somma delle n_i.

INDICI STATISTICI

La ***mediana*** o ***valore mediano*** M_e è quell'indice di posizione che, una volta ordinate in senso crescente le osservazioni di un fenomeno, divide la distribuzione in due gruppi di uguale numerosità: al primo gruppo, infatti, appartengono le osservazioni uguali o inferiori alla mediana; al secondo gruppo, invece, quelle superiori o uguali alla mediana. La mediana, dunque, è la modalità dell'unità statistica che occupa il posto centrale nella distribuzione ordinata delle osservazioni. Dato, cioè, un insieme costituito da n intensità (x_1, x_2, …, x_n), la determinazione della mediana è diversa a seconda che n sia pari o dispari, precisamente si ha:

- se n è pari, la mediana è data dalla semisomma delle intensità individuate dalle due posizioni centrali, C_1 e C_2, ovvero dalla loro media aritmetica:

$$C_1 = x_{\frac{n}{2}}, \quad C_2 = x_{\frac{n}{2}+1} \quad \Rightarrow \quad M_e = \frac{C_1 + C_2}{2}$$

- se n è dispari, la mediana è data dal valore che occupa la posizione centrale nella distribuzione dei valori posti in graduatoria:

$$M_e = x_{\frac{n+1}{2}}$$

INDICI STATISTICI

Esempio

La mediana delle seguenti intensità ($n = 7$, dispari):

3; 15; 9; 2; 6; 12; 5

si ottiene ordinando dapprima le intensità in ordine crescente,

$$x_1 = 2;\ x_2 = 3;\ x_3 = 5;\ x_4 = 6;\ x_5 = 9;\ x_6 = 12;\ x_7 = 15$$

e poi considerando l'intensità che occupa il posto centrale, essendo n dispari:

$$M_e = x_4 = 6$$

INDICI STATISTICI

Esempio

La mediana delle seguenti intensità ($n = 8$, pari):

$$7;\ 16;\ 2;\ 3;\ 9;\ 12;\ 15;\ 5$$

si ottiene ordinando dapprima le intensità in ordine crescente,

$$x_1 = 2;\ x_2 = 3;\ x_3 = 5;\ x_4 = 7;\ x_5 = 9;\ x_6 = 12;\ x_7 = 15;\ x_8 = 16$$

e poi considerando le intensità che occupano i due posti centrali, essendo n pari:

$$C_1 = x_{\frac{8}{2}} = x_4 = 7, \quad C_2 = x_{\frac{8}{2}+1} = x_5 = 9 \quad \Rightarrow \quad M_e = \frac{7+9}{2} = \frac{16}{2} = 8$$

INDICI STATISTICI

Se, invece, si ha una distribuzione di frequenze, per calcolare la mediana, occorre determinare le frequenze cumulate: indicando con *n* la somma delle frequenze, se *n* è pari, la mediana è data da

$$\frac{n}{2}$$

Se, invece, *n* è dispari, la mediana è data da:

$$\frac{n+1}{2}$$

INDICI STATISTICI

Esempio

Se si effettua l'indagine su un numero di figli su un campione di famiglie, come riportato nella seguente tabella:

Figli x_i	F.A n_i	F.C.A.
0	3	3
1	8	11
2	7	18
3	4	22
4	1	23
5	1	24
6	1	25
Tot.	25	

essendo n dispari, la mediana è il valore corrispondente a

$$\frac{n+1}{2}=\frac{25+1}{2}=\frac{26}{2}=13$$

ovvero la mediana è 2 poiché 11<13<18.

INDICI STATISTICI

La mediana, pertanto, si può calcolare per tutte quelle variabili le cui modalità possono essere ordinate, ovvero per le variabili qualitative ordinali, e per tutte le variabili quantitative: risulta, infatti, più conveniente usarla qualora si voglia esprimere il valore centrale di distribuzioni di caratteri che non possono essere misurati "esattamente" (ad esempio, i caratteri psicologici graduabili) oppure qualora non si possa far riferimento alla distribuzione normale, proprio grazie alla sua capacità di essere rappresentativa della posizione della distribuzione anche in presenza di valori estremi notevolmente diversi da tutti gli altri.

La mediana, dunque, *minimizza i costi complessivi* ed è soprattutto resistente ai valori estremi: rappresenta, infatti, un indice per decisioni che implicano costi elevati nei casi estremi.

INDICI STATISTICI

La ***moda*** o ***norma*** M_O di una distribuzione di frequenza X, calcolabile per caratteri sia quantitativi sia qualitativi, non risentendo dei valori estremi, rappresenta la modalità, o classe di modalità, caratterizzata dalla massima frequenza (assoluta o relativa) o densità di frequenza, ovvero il valore numerico che, nella distribuzione di frequenza, è maggiormente presente rispetto agli altri. A tal riguardo occorre evidenziare che la moda è una modalità, non una frequenza. Se si rappresenta, pertanto, la distribuzione di frequenza in termini grafici, si può affermare che la moda corrisponde al picco della distribuzione (ad esempio in un grafico a colonne o a nastri, la colonna più alta o il nastro più lungo individua la moda della distribuzione) che, di conseguenza, risulterà *zeromodale* se non ammette alcun valore modale, ovvero nessun picco, *unimodale* se ne ammette uno solo (in tal caso la moda ha significato di sintesi), *bimodale* se ne ammette due, *trimodale* se ne ammette tre, … Per poter determinare, quindi, la classe modale risulta opportuno ricorrere all'istogramma, individuando l'intervallo di altezza massima, ovvero il punto di massimo della curva; la classe con la maggiore densità media, corrispondente proprio all'altezza dell'istogramma, sarà quella modale. La moda, dunque, *minimizza gli scontenti* ed è utilizzata in tutte quelle situazioni ove il consenso ed il numero delle singole unità ha significato per la decisione: la moda, infatti, è un indice utile per individuare la modalità più rappresentativa.

INDICI STATISTICI

Esempio

La moda della seguente successione di termini (n = 13):

$$x_1 = 3;\ x_2 = 5;\ x_3 = 9;\ x_4 = 3;\ x_5 = 5;\ x_6 = 7;\ x_7 = 3;$$
$$x_8 = 2;\ x_9 = 9;\ x_{10} = 3;\ x_{11} = 4;\ x_{12} = 3;\ x_{13} = 6$$

è data dal termine che compare con maggiore frequenza, ovvero è $M_O = 3$ perché compare 5 volte.

Esempio

Data la variabile X = numero di esami sostenuti da sei studenti ed osservati i seguenti valori:

STUDENTI	Nicola	Mary	Eleonora	Beatrice	Davide	Christian
ESAMI	30	19	8	7	27	10

Si può concludere che la variabile X non ha moda, ovvero è *zero modale*, essendo la moda definita come la modalità più frequente: non esiste, infatti, nessuna modalità (numero di esami) ripetuta più delle altre e tutte le modalità hanno la stessa frequenza assoluta pari ad uno studente.

Qual è la modalità più alta? 30

Qual è la modalità più frequente? Nessuna in quanto tutte hanno la stessa frequenza pari ad 1.

Per individuare la moda di una variabile, dunque, bisogna chiedersi in primo luogo qual è la variabile e poi quali sono le modalità e qual è la modalità con la frequenza più alta.

INDICI STATISTICI

Esempi

v.s. discrete

Voti x_i	Numeri di studenti n_i
25	3
26	2
27	8
28	1

v.s. continue di uguale ampiezza

Voti x_i	Numeri di studenti n_i
18---20	3
21---23	5
24---26	10
27---29	4

v.s. continue di diversa ampiezza

Voti x_i	Numeri di studenti n_i	d_i	$H_i = n_i / d_i$
18---21	5	3	5/3 = 1,6
21---23	4	2	4/2 = 2
24---28	6	4	6/4 = 1,5
29---30	3	1	3/1 = 3

INDICI STATISTICI

Si osservi che:

- per *caratteri discreti* la moda si individua facilmente scorrendo lungo la colonna delle frequenze;
- per *caratteri continui*, se le classi di modalità hanno tutte uguale ampiezza, la moda cade nella classe con maggiore frequenza; se le classi di modalità, invece, hanno ampiezza diversa, si divide ogni frequenza per l'ampiezza della rispettiva classe calcolando, così la densità di frequenza; la moda, poi, cade nella classe con maggiore densità di frequenza.

INDICI STATISTICI

I ***quantili*** sono le intensità che dividono, dopo aver ordinato i dati, una distribuzione di frequenza in un certo numero di parti uguali (ad esempio, la mediana è quel valore che divide in due parti uguali l'insieme delle unità ordinate per grandezza, ovvero la distribuzione è divisa, rispetto a tale valore, in due parti ognuna contenente il 50% delle unità). Se si divide la distribuzione in due parti si parla di *terzili* (il primo terzile è quello che lascia alla sua sinistra un terzo delle osservazioni e alla sua destra i rimanenti due terzi; il secondo terzile è quello che lascia alla sua sinistra i due terzi e alla sua destra un terzo rimanente). Se si divide la distribuzione in tre parti si parla di *quartili* (il primo quartile Q_1 lascia alla sua sinistra il 25% dei casi e alla sua destra il rimanente 75%; il secondo quartile Q_2, che coincide con la mediana, lascia alla sua sinistra il 50% dei casi e alla sua destra il rimanente 50%; il terzo quartile Q_3 lascia alla sua sinistra il 75% dei casi e alla sua destra il rimanente 25%). Se si divide la distribuzione in nove parti si parla di *decili*, …, in novantanove parti si parla di *centili*, in cento parti si parla di *percentili*.

INDICI STATISTICI

Se X è un carattere con n modalità ordinate $x_1, x_2, \ldots, x_n$ ($x_1 \leq x_2 \leq \ldots \leq x_n$), per il calcolo dei quartili si procede in maniera analoga a quanto visto in precedenza per la mediana, considerando le posizioni degli elementi:

- se n è pari:

$$Q_1 = \frac{x_{\frac{n}{4}} + x_{\frac{n}{4}+1}}{2}$$

- se n è dispari:

$$Q_1 = x_{\frac{n+1}{4}}$$

I quantili, dunque, si possono calcolare per tutte quelle variabili per le quali risulta possibile ordinarne le modalità, ovvero per variabili qualitative ordinali, oltre che per tutte le variabili quantitative.

INDICI STATISTICI

Esempio

Date le seguenti intensità ($n = 7$, dispari):

$$20;\ 65;\ 2;\ 10;\ 37;\ 15;\ 3$$

il loro quartile Q_1 si ottiene ordinando dapprima le intensità in ordine crescente:

$$x_1 = 2;\ x_2 = 3;\ x_3 = 10;\ x_4 = 15;\ x_5 = 20;\ x_6 = 37;\ x_7 = 65$$

e poi considerando, come primo quartile, l'intensità che occupa il posto:

$$x_{\frac{n+1}{4}} = x_{\frac{7+1}{4}} = x_{\frac{8}{4}} = x_2 = 3 = Q_1$$

Analogamente il terzo quartile Q_3 si ottiene considerando l'intensità che occupa sempre il secondo posto partendo, però, dall'ultima osservazione, ovvero $Q_3 = x_6 = 37$.

INDICI STATISTICI

Esempio

Date le seguenti intensità (n = 8, pari):

20; 65; 83; 10; 37; 15; 3; 2

il loro quartile Q_1 si ottiene ordinando dapprima le intensità in ordine crescente:

$x_1 = 2$; $x_2 = 3$; $x_3 = 10$; $x_4 = 15$; $x_5 = 20$; $x_6 = 37$; $x_7 = 65$; $x_8 = 83$

e poi considerando, come primo quartile, l'intensità che occupa il posto:

$$x_{\frac{n}{4}} = x_{\frac{8}{4}} = x_2 = 3; x_{\frac{n}{4}+1} = x_{\frac{8}{4}+1} = x_{2+1} = x_3 = 10$$

Effettuando, infine, la semisomma tra tali numeri, si ottiene:

$$Q_1 = \frac{3+10}{2} = \frac{13}{2} = 6,5$$

Analogamente il terzo quartile Q_3 si ottiene considerando la semisomma delle intensità che occupano sempre il secondo ed il terzo posto partendo, però, dall'ultima osservazione, ovvero:

$$Q_3 = \frac{37+65}{2} = \frac{102}{2} = 51$$