Il concetto di valore medio in generale

Nella statistica descrittiva si distinguono solitamente due tipi di medie:

- **le medie analitiche**, che soddisfano ad una condizione di invarianza e si calcolano tenendo conto di tutti i valori di una determinata distribuzione;

- **le medie di posizione,** che si calcolano considerando solo alcuni particolari valori.

E' evidente che, considerati certi valori di una certa variabile statistica, e'possibile calcolare molti tipi di medie, ma tutti i valori medi soddisfano *la condizione di internalita' di Cauchy*, ossia la media e'sempre un qualsiasi valore compreso fra il minimo ed il massimo dei dati considerati.

Pertanto si puo' chiamare **media** di una distribuzione x_1, x_2, …., x_n, rispetto ad una funzione quella quantità M che, sostituita alle x_i nella funzione lascia invariato il risultato. Ossia:

$$f(M, M, \ldots\ldots, M) = f(x_1, x_2, \ldots\ldots, x_n)$$

La media aritmetica e le sue proprietà

In questo caso si assume come funzione la somma dei dati. Si ha pertanto la seguente definizione:

Si definisce **media aritmetica** di più unità statistiche, quel valore che sostituito ai dati, lascia invariata la loro somma.

Indicati con x_1, x_2, ..., x_n i numeri dati, per la definizione precedente si ha:

$$M + M + ... + M = x_1 + x_2 + ... + x_n = nM$$

da cui si ricava:

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

che si chiama **media aritmetica semplice**. Se i valori x_i hanno frequenze diverse, cioè compaiono più volte nelle osservazioni, e quindi il valore x_1 ha frequenza y_1, x_2 ha frequenza y_2, ecc., la condizione di invarianza della somma diventa:

$$x_1 y_1 + x_2 y_2 + ... + x_n y_n = My_1 + My_2 + ... + My_n$$

da cui si ricava:

$$M = \frac{x_1 y_1 + x_2 y_2 + ... + x_n y_n}{y_1 + y_2 + ... + y_n} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} y_i}$$

che si chiama **media aritmetica ponderata**, perchè le frequenze vengono dette anche pesi.

Proprietà della media aritmetica

1. La somma algebrica degli scarti dei valori dati dalla loro media aritmetica è uguale a zero. Quindi:

$$(x_1 - M) + (x_2 - M) + ... + (x_n - M) = \sum_{i=1}^{n} (x_i - M) = 0$$

2. La somma dei quadrati degli scarti dei valori di una distribuzione dalla loro media aritmetica è minore della somma dei quadrati degli scarti da un qualsiasi altro valore. Quindi:

$$\sum_{i=1}^{n} (x_i - m)^2 = \min$$ e solo se $m = M$

3. Se tutti i valori di una distribuzione vengono moltiplicati (o divisi) per una stessa quantità k, si ottiene una nuova distribuzione la cui media aritmetica corrisponde alla media aritmetica della distribuzione originaria moltiplicata (o divisa) per k (*proprietà di omogeneità*).

4. Se a tutti i valori di una distribuzione si addiziona (o si sottrae) una stessa quantità h, si ottiene una nuova distribuzione la cui media aritmetica corrisponde alla media aritmetica della distribuzione originaria aumentata (o diminuita) della quantità h (*proprietà traslativa*).

5. Tra la media aritmetica della distribuzione del carattere X e la media aritmetica di una trasformata lineare Y=h+kX, con h e k costanti, per le proprietà traslativa e di omogeneità, vale la seguente relazione:

$$M(Y) = M(h + kX) = h + kM(X)$$

6. Date due distribuzioni, con rispettive medie M_1 e M_2, la distribuzione che si ottiene riunendo le due distribuzioni originarie, avrà per media aritmetica la media aritmetica delle due medie aritmetiche ,M_1 e M_2 (*proprietà associativa).*

La media geometrica e le sue proprietà

Si definisce media geometrica dei valori x_1, x_2, ..., x_n , quel numero G che, sostituito ai valori x_i, lascia invariato il loro prodotto. Ossia:

$$x_1 \cdot x_2 \cdot ..... \cdot x_n = G \cdot G \cdot ..... \cdot G = G^n$$

da cui si ricava:

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

che è la **media geometrica semplice**. Nel caso di distribuzione i cui valori x_i si presentano con le rispettive frequenze o pesi y_i, si ottiene:

$$G = \sqrt[N]{x_1^{y_1} \cdot x_2^{y_2} \cdot ..... \cdot x_n^{y_n}} = \sqrt[N]{\prod_{i=1}^{n} x_i^{y_i}}$$

che è la **media geometrica ponderata**, dove: $N = \sum_{i=1}^{n} y_i$

Proprietà della media geometrica

1. Il logaritmo della media geometrica (semplice o ponderata) è la media aritmetica (semplice o ponderata) dei logaritmi dei valori della variabile statistica.

2. Se tutti i valori di una distribuzione vengono moltiplicati per una stessa quantità $k>0$, si ottiene una nuova distribuzione la cui media geometrica corrisponde alla media geometrica della distribuzione originaria moltiplicata per k (*proprietà di omogeneità*).

3. La media geometrica di più rapporti è uguale al rapporto tra le media geometrica delle grandezze poste al numeratore e la media geometrica di quelle che figurano al denominatore.

4. Il reciproco della media geometrica è uguale alla media geometrica dei reciproci delle quantità date.

La media armonica e le sue proprietà

Si definisce media armonica dei valori x_1, x_2, ...,x_n, quel valore Ma che, sostituito ai valori x_i, lascia invariata la somma dei loro reciproci. Ossia:

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

che è **la media armonica semplice**. Nel caso di distribuzione i cui valori x_i si presentano con le rispettive frequenze o pesi y_i, si ottiene:

$$\frac{y_1}{x_1}+\frac{y_2}{x_2}+.....+\frac{y_n}{x_n}=\frac{y_1}{Ma}+\frac{y_2}{Ma}+.....+\frac{y_n}{Ma}=\frac{N}{\sum_{i=1}^{n}\frac{y_i}{x_i}}$$

che è la **media armonica ponderata**.

Proprietà della media armonica

Tale media annulla la somma algebrica degli scarti relativi di n valori, cioè:

$$\sum_{i=1}^{n}\frac{x_i-m}{x_i}=0$$

; se m=Ma

MODA E MEDIANA

Moda o valore normale

La *moda* o *valore normale* (detta anche *valore modale* o *norma*) è un valore caratteristico di una distribuzione di frequenze la cui determinazione non richiede alcun calcolo.

Si dice **moda** o **valore normale** di una distribuzione di frequenze la modalità o il valore della variabile al quale corrisponde la massima frequenza.

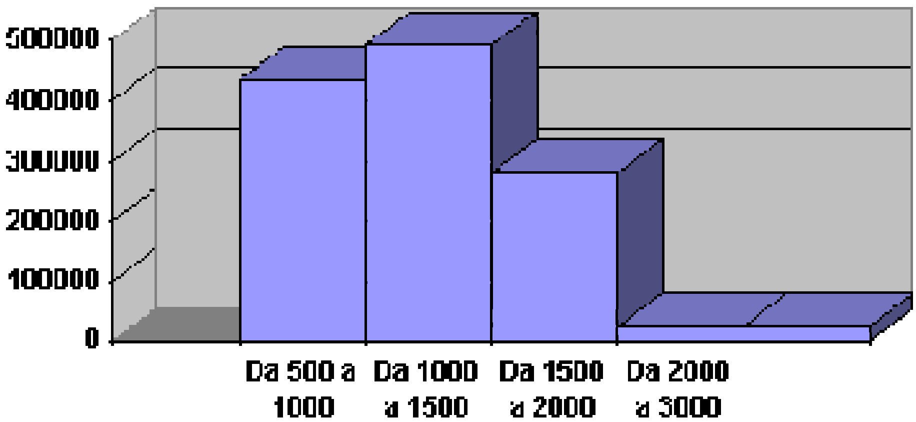
Se si ha una serie o una seriazione con valori discreti, la moda è il valore che ha la massima frequenza. Se i dati sono raggruppati in classi, il calcolo della moda presenta maggiori difficoltà. Se l'ampiezza della classe è costante si dirà classe modale quella che ha la frequenza maggiore. Se le classi hanno ampiezza diversa, come si è già fatto per la rappresentazione grafica, si divide ogni frequenza per l'ampiezza della rispettiva classe e la classe modale è la classe alla quale corrisponde il rapporto maggiore. Talvolta per individuare la classe modale è sufficiente esaminare l'istogramma della distribuzione: la classe modale è quella che è base del rettangolo di altezza massima. Il valore modale è, fra tutti i valori medi, il più significativo in quanto è un dato che esprime il valore di una concreta osservazione sul fenomeno, mentre le medie di calcolo possono o meno coincidere con un valore della distribuzione. Considerando le retribuzioni di un certo insieme di lavoratori, il valore normale è senz'altro il più significativo, in quanto corrisponde alla retribuzione più frequente e non è influenzato dalle retribuzioni limite (o molto basse o molto alte). Esistono delle variabili statistiche che presentano più di un valore modale (distribuzioni plurimodali), in altri casi la moda può essere agli estremi della distribuzione (distribuzioni zeromodali). La moda è molto utilizzata nelle scienze applicate, sia per il suo significato, sia per la facilità di calcolo.

Esempio

Si consideri la seguente tabella: ripartizione delle autovetture italiane prodotte nell'anno 1981 secondo la cilindrata.

Cilindrata (in cm^3)	N. autovetture
500 --------| 1.000	433.963
1.000 ------| 1.500	491.798
1.500------| 2.000	281.239
oltre 2.000	50.340
Totale	1.257.340

Le classi hanno tutte la stessa ampiezza, eccetto l'ultima che è aperta. In caso di classi aperte (la prima o l'ultima), o si chiudono indicando un valore estremo logico, oppure si trascurano, quando il rapporto espresso in percentuale fra la frequenza di tale classe e il valore totale delle frequenze è relativamente piccolo. La precedente dà origine al seguente grafico.



Si usano gli istogrammi per rappresentare seriazioni continue con i dati raggruppati in classi. Fissato un sistema di assi cartesiani ortogonali, sull'asse delle ascisse si riportano tanti intervalli consecutivi quante sono le classi; su questi intervalli si costruiscono dei rettangoli **le cui aree sono proporzionali alle frequenze**. Si noti che se le classi hanno la stessa ampiezza, o modulo, basterà riportare altezze proporzionali alle frequenze; in caso contrario le altezze si ottengono dividendo la relativa frequenza per l'ampiezza della classe, in modo che l'area rappresenti la frequenza. Con gli istogrammi la somma delle aree di tutti i rettangoli è proporzionale alla somma delle frequenze.

L'ultima classe, che è aperta, si può chiudere alla cilindrata di 3000 cm^3 e allora l'altezza del rettangolo risulta:

$$50.340:2 = 25.170$$

perché la base di tale rettangolo è doppia delle basi dei rettangoli precedenti.

Mediana

La mediana è una media di posizione e rappresenta il valore centrale della distribuzione quando i dati sono ordinati. Precisamente:

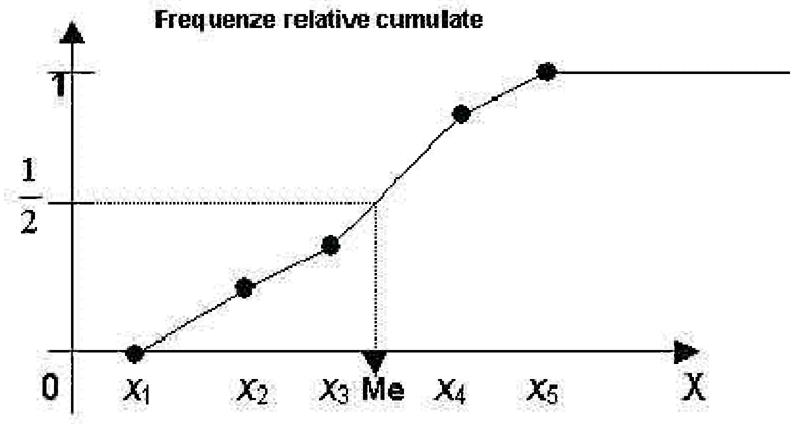
Siano x_1, x_2, …, x_n i valori ordinati in senso non decrescente, si dice **mediana Me** il valore che bipartisce la successione, ossia il valore non inferiore a metà dei valori e non superiore all'altra metà.

Ordinati i valori, se il numero *n* dei termini è *dispari*, la mediana è proprio il valore centrale; se *n* è *pari*, si assume come mediana la semi somma dei due valori centrali. Il procedimento sopra esposto si applica per le serie. Per le distribuzioni di frequenze con valori discreti, i dati sono generalmente già ordinati; occorre allora calcolare le frequenze assolute cumulate, che si ottengono associando a ogni valore la somma della rispettiva frequenza con tutte quelle che la precedono, e determinare *quale valore* corrisponde:

alla frequenza cumulata $\frac{\sum_{i=1}^{i=n} y_i}{2}$, se la somma è pari;

alla frequenza cumulata $\frac{1+\sum_{i=1}^{i=n} y_i}{2}$, se la somma è dispari;

tale valore è la mediana. Se i dati sono raggruppati in classi, si determina la classe mediana mediante le frequenze assolute cumulate. Per ottenere esattamente il valore mediano, si applica un'interpolazione lineare fra i due valori estremi della classe in cui cade la mediana, supponendo che le frequenze siano distribuite nella classe a intervalli regolari. Per il calcolo approssimato in questo caso è utile servirsi del grafico della poligonale delle frequenze relative cumulate. È sufficiente trovare l'ascissa del punto di ordinata $\frac{1}{2}$, come si può rilevare dal grafico seguente:



La mediana non è influenzata dai valori estremi della distribuzione, quindi anche se le classi estreme, in caso di distribuzione continua, sono aperte, non occorre chiuderle. Inoltre, se la distribuzione è molto asimmetrica, il valore mediano è più appropriato della media aritmetica per esprimere un valore sintetico della distribuzione. Una proprietà caratteristica della mediana è la seguente: *la mediana rende minima la somma dei valori assoluti degli scarti*, in altre parole la somma dei valori assoluti degli scarti dalla mediana e *non* è superiore (ossia minore o eguale) alla somma dei valori degli scarti da qualunque altro valore.

La scelta della media più opportuna

In linea di principio la scelta della media dipende dalla natura dei dati, dal loro andamento, dagli obiettivi della ricerca: in definitiva occorre sempre considerare caso per caso.

La **media aritmetica** è la media statistica per eccellenza. Essa permette un'ottima correzione degli errori accidentali commessi in una rilevazione statistica, per cui risulta utile in tutti i campi della scienza e della tecnica in cui vengono effettuate misurazioni di qualunque genere.

Permette, inoltre, di risolvere il problema della equipartizione della somma. Risente, però, della presenza di valori estremi alti o bassi. E' la media più opportuna se la distribuzione tende ad una progressione aritmetica.

La **media geometrica** è impiegata allorchè i dati non sono numerosi ed i termini della distribuzione presentano valori molto differenti fra loro.

Permette di risolvere il problema della equipartizione del prodotto. Non si può utilizzare quando anche uno solo dei termini è nullo; inoltre sorgono inconvenienti quando i termini non sono tutti positivi. E' la media più opportuna se la distribuzione tende ad una progressione geometrica. Se il valore che si discosta sensibilmente dagli altri è un valore alto si può calcolare, al posto della media aritmetica, la media geometrica che non risulta influenzata dai valori alti.

La **media armonica** si utilizza quando i termini sono inversamente proporzionali ad altri termini o grandezze che variano in progressione aritmetica, cioè quando gli inversi dei termini costituiscono una progressione aritmetica; ovvero quando si deve trovare il valore medio non del fenomeno considerato, ma di un fenomeno che è l'inverso del primo (per es. prezzo e potere d'acquisto; interesse effettivo che cresce al decrescere del corso del titolo, ecc.).

E' la media più opportuna se la distribuzione tende ad una progressione armonica.