Media geometrica

La *media geometrica* di una variabile statistica X che assume il valore x_i con frequenza f_i (con i=1,2,...,n) tale che la somma delle f_i sia pari a N osservazioni complessive è definita come la radice di ordine N del prodotto degli n termini x_i ciascuno elevato alla potenza f_i. In formule si avrà:

per la semplice $M_0 = \sqrt[N]{\prod_{i=1}^{i=N} X_i}$, **dove i = 1, 2, ..., N.**

per la ponderata $M_0 = \sqrt[N]{\prod_{i=1}^{i=n} x_i^{f_i}}$, **dove** $\sum_{i=1}^{i=n} f_i$ **= N.**

Esempio Supponiamo di impiegare 1 lira ad interesse composto ai seguenti tassi : i1=0,05 nel primo anno; i2=0,06 nel secondo anno; i3=0,055 nel terzo anno; i4=0,07 nel quarto anno; i5=0,065 nel quinto anno. Il montante alla fine del primo anno sarà dato dalla relazione C1=1+0,05; alla fine del secondo anno sarà C2=(1+0,05)(1+0,06); alla fine del terzo anno sarà C3=(1+0,05)(1+0,06)(1+0,055) e così via.

Ci si chiede qual'è il tasso medio i a cui capitalizzare la nostra lira per ottenere alla fine del quinquennio il montante C5 che rappresenta, evidentemente, l'invariante del problema.

Da quanto detto, si deduce che deve essere:

$$(1+i)^5 = 1{,}05 \times 1{,}06 \times 1{,}055 \times 1{,}07 \times 1{,}065$$

Proprietà della media geometrica

1) Il reciproco della media geometrica è uguale alla media geometrica dei reciproci dei termini.
2) La potenza emmesima della media geometrica è uguale alla media geometrica delle potenze emmesime dei termini.
3) Il logaritmo della media geometrica è uguale alla media aritmetica dei logaritmi dei termini.
4) La media geometrica è utilizzata quando le variabili non sono rappresentate da valori ottenuti come prodotto o rapporto tra valori lineari. Serve per il confronto di superfici o volumi, di tassi di variazione, cioè valori che sono espressi da rapporti.
5) Per il calcolo della media geometrica è condizione necessaria che le quantità siano tutte positive.

da cui si vede che (1+i) è la media geometrica dei prodotti indicati nel secondo membro e non dei singoli tassi annui, per cui, mediante i logaritmi si calcola:

$$\log(1+i) = \frac{1}{5}(\log 1{,}05 + \log 1{,}06 + \log 1{,}055 +$$

$$+ \log 1{,}07 + \log 1{,}065 = 0{,}025296$$

Risalendo al numero, si ha che i = 0,0599, pari a 5,99%.

Media armonica **La media armonica è definita come il reciproco della media aritmetica dei reciproci dei termini. E' quindi anche vero che il reciproco della media armonica è la media aritmetica dei reciproci dei termini. In formule:**

per la semplice $M_{-1} = \frac{N}{\sum_{i=1}^{i=N} \frac{1}{X_i}}$**, dove i = 1, 2, ..., N.**

per la ponderata $M_{-1} = \frac{\sum_{i=1}^{i=n} f_i}{\sum_{i=1}^{i=n} \frac{f_i}{X_i}}$**, dove** $\sum_{i=1}^{i=n} f_i$ **= N.**

La media armonica è la stima più corretta della tendenza centrale per distribuzioni di dati in cui devono essere usati gli *inversi o reciproci*, come nel caso di misure dei tempi di reazione.

<u>Esempio</u> Si voglia conoscere il consumo medio annuo di rasoi usa-e-getta in Italia, mediante una ricerca diretta sui consumatori.

Non sarà opportuno chiedere: "Quanti rasoi consuma in media all'anno?" perché la domanda così formulata richiede una stima relativa ad un ampio intervallo di tempo; si potrà invece chiedere: "Quanti giorni le dura in media un rasoio?". Immaginiamo di esaminare le risposte di cinque persone:

1a persona	**10 giorni in media**
2a persona	**6 giorni in media**
3a persona	**30 giorni in media**
4a persona	**5 giorni in media**
5a persona	**14 giorni in media**
Totale	**65**

La media aritmetica delle durate è 65:5=13 giorni. Ma da questo dato non è corretto ricavare il consumo medio annuo pari a 365:13=28,1 rasoi in media per persona, equivalente per i 5 consumatori considerati a 28,1x5=140,5 rasoi di consumo annuo. Infatti con i dati di partenza possiamo ricavare direttamente il consumo globale

Persone	**Consumo annuo di rasoi**
1a	**365:10=36,5**
2a	**365: 6=60,8**
3a	**365:30=12,2**
4a	**365: 5=73,0**
5a	**<u>365:14=26,1</u>**

In complesso 208,6 rasoi

mentre in precedenza si era ottenuto il risultato di 140,5 rasoi. Con l'ultimo risultato il consumo pro-capite è 208,6:5=41,7 rasoi e la durata media 365:41,7=8,8 giorni. Questo valore si ottiene immediatamente come media armonica dei dati iniziali:

$$M_{-1} = \frac{5}{\frac{1}{10} + \frac{1}{6} + \frac{1}{30} + \frac{1}{5} + \frac{1}{14}} = 8,8.$$

Per comprendere il motivo per il quale si deve adoperare la media armonica e non quella aritmetica delle durate, occorre osservare che il problema riguarda il *consumo*, per cui si deve tenere conto che la prima persona consuma in un giorno 1/10 di rasoio, la seconda consuma 1/6 di rasoio e così via, per cui, nel

complesso, le cinque persone consumano in un giorno la somma delle quantità suddette.

Il valore unico 1/x̄ da sostituire a questi consumi diversi, lasciando invariato il consumo complessivo delle 5 persone è dato, perciò, dalla equazione:

$$5\frac{1}{\bar{x}} = \frac{1}{10} + \frac{1}{6} + \frac{1}{30} + \frac{1}{5} + \frac{1}{14}$$

da cui si ricava la durata media facendo l'inverso della media dei consumi, cioè proprio la media armonica. Avendo quindi rilevato la durata dei rasoi invece dei consumo, bisogna tener conto che tra queste due quantità esiste una relazione inversa.

Formula generale per le medie ottenibili a calcolo

Media di potenze

La media di potenze di indice r di una variabile statistica che si presenta con n modalità differenti, ciascuna avente frequenza fi , è quel valore che si ottiene considerando la radice di ordine r della media aritmetica delle potenze r-esime delle singole determinazioni. In simboli:

$$M_r = \sqrt[r]{\frac{\sum_{i=1}^{i=n} x_i^r f_i}{\sum_{i=1}^{i=n} f_i}}$$

La media di potenze si definirà semplice o ponderata, a seconda che le frequenze siano tutte uguali all'unità oppure tra loro diverse.

Media di potenze di ordine r

Mr è una funzione continua e crescente con r e in statistica definisce la Media di potenza di ordine r della variabile considerata. Essa è pari alla radice r-esima del Momento di ordine r rispetto all'origine.

Per r= -1, la media di potenze è uguale alla media armonica

~~Per r = 0, la media di potenze è uguale alla media geometrica~~

Per r = 1, la media di potenze è uguale alla media aritmetica

Per r=2, la media di potenze è uguale alla media quadratica

media *armonica*<media *geometrica*

media *geometrica*<media *aritmetica*

Medie lasche o di posizione

Mediana **Se le unità statistiche sono in ordine crescente dei valori della variabile, la mediana è quel valore al di sotto ed al di sopra del quale si situa la metà del numero totale dei casi, cioè divide l'insieme delle unità in due parti di uguale frequenza.**

Se il numero dei termini è dispari, la mediana è il valore relativo al termine che occupa il posto di mezzo; se è pari essa è uguale alla media aritmetica dei valori relativi ai due termini che occupano i posti centrali.

Se i dati raggruppati in n classi, la mediana si calcola con la formula:

$$M_e = Cl_i + \frac{\left[\left(\frac{N}{2} - \sum_{k=1}^{k=i-1} f_k\right) \cdot A_i\right]}{f_i}$$ **, con** $\sum_{i=1}^{i=n} f_i = N$**,**

dove Cl_i è il confine inferiore della classe i , A_i è la sua ampiezza e f_i è la frequenza della classe.

Proprietà della mediana

1) Il numero degli scostamenti positivi è uguale al numero degli scostamenti negativi. Quindi, in una distribuzione o serie di dati ogni valore estratto a caso ha la stessa probabilità di essere inferiore o superiore alla mediana.

2) La somma dei valori assoluti degli scostamenti è un minimo in confronto alla somma dei valori assoluti degli scostamenti cui darebbe luogo un altro valore medio qualsiasi diverso dal valore mediano.

3) E' una misura robusta: non è influenzata dalla presenza di dati anomali e in particolare dai valori estremi, ma soltanto dal numero delle osservazioni

4) La mediana è la misura di posizione o di tendenza centrale utilizzata in quasi tutti i tests non parametrici.

Esempio Nella seguente successione di numeri 5, 9, 6, 14, 11, il valore mediano è il 9. Nella seguente 5, 6, 9, 11, 14, 18 formata da un numero pari di termini la mediana è (9+11)/2=10.

Nel caso di variabili statistiche divise in intervalli, il metodo migliore è quello di costruire la distribuzione cumulativa delle frequenze. Ad esempio, consideriamo la distribuzione:

Classi	n_i	$N_i=\Sigma n_k$
50|--100	110	110
100|--200	400	510
200|--300	90	600
	N = 600	

La mediana corrisponde alla modalità del l'unità che occupa il posto 600/2=300, quindi si tratta di un valore interno alla classe 100|--200.

Per individuarlo, si fa l'ipotesi di uniforme distribuzione delle unità all'interno della classe e si considera la proporzione:

$$(M_{-1} - 100): (200-100) = (300-110) : 400$$

dalla quale si ricava la mediana

$$M_{-1} = 190x100/400 + 100 = 147,5$$

che risulta leggermente inferiore al valore centrale della classe 100|--200 alla quale apparteneva il valore che lasciava da una parte e dall'altra lo stesso numero di termini.

Quartili

Il *primo quartile* di una successione di termini non decrescente è quella quantità al di sotto della quale sta 1/4 ed al di sopra della quale stanno i 3/4 dei valori dati.

Il *secondo quartile* coincide con la mediana della distribuzione. Mentre il *terzo quartile* è quella quantità al di sotto della quale stanno i 3/4 ed al di sopra della quale sta 1/4 dei valori dati.

***Il primo quartile è la mediana dei valori inferiori alla mediana* e *il terzo quartile è il valore mediano dei valori superiori alla mediana* della distribuzione.**

Moda o norma

Si chiama *moda* o *norma o valore modale* quella modalità della variabile che si presenta con la frequenza più elevata e la classe in cui essa risulta compresa si chiama *classe modale*. Nel caso di dati raggruppati per classi, se i è la classe modale, si può calcolare la moda in base alla formula

$$M_d = Cl_i + \frac{[(f_i - f_{i-1}).A_i]}{(f_i - f_{i-1}) - (f_i - f_{i+1})},$$

dove Cl_i è il confine inferiore della classe i , A_i è la sua ampiezza e f_i è la frequenza della classe.

Proprietà della moda

1) La moda rende *massimo il numero degli scostamenti nulli*.

2) *Non è influenzata dalla presenza di valori estremi*, tuttavia viene utilizzata solamente per scopi descrittivi, perché è *meno stabile ed oggettiva* di altre misure di tendenza centrale.
Essa differisce sia da campione a campione, sia quando con gli stessi dati si formano classi di distribuzione con ampiezza differente

3) Se si fanno variare tutti i termini di una serie in base ad una certa legge, la moda della serie data corrisponde alla moda della nuova serie.

<u>Esempio</u> Si abbia la seguente distribuzione del numero di frantoi in funzione della capacità annua di produzione di olio:

Capacità produttiva (q.li)	Numero di frantoi
150|--200	60
200|--300	115
300|--500	140
500|--750	75
750|--1000	15
Totale	405

A prima vista si potrebbe ritenere che la moda sia compresa nella classe 300|--500, ma ciò è falso, in quanto le classi hanno ampiezza differente.

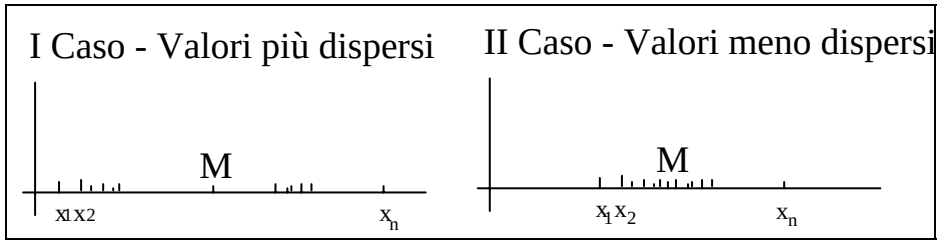
Per trovare la classe modale, quindi occorre anzitutto ridurre le varie classi ad una ampiezza uguale: ad esempio:

Capacità Produttiva	Numero di frantoi		
150|--200	60	700|--750	15
200|--250	57,5	750|--800	3
250|--300	57,5	800|--850	3
300|--350	35	850|--900	3
350|--400	35	900|--950	3
400|--450	35	950|--1000	3
450|--500	35		
500|--550	15	Totale	405
550|--600	15		
600|--650	15		
650|--700	15		

Così facendo si scopre che la classe modale è la prima.

Dispersione o variabilità

Definizione **Si definisce *dispersione* oppure *variabilità* di una distribuzione l'attitudine dei dati a disporsi intorno a un valore medio**



Misure di dispersione o variabilità

Campo di variazione

Intervallo di valori compreso tra il più piccolo ed il più grande dei valori assunti dalla variabile.

Deviazione semplice media o scostamento semplice medio

$$\textit{S. s. m. rispetto a } K = \frac{1}{N}\sum_{i=1}^{i=n}\left|x_i - K\right|.f_i$$

ove $N = \sum_{i=1}^{i=n} f_i$**.**

E' una misura di dispersione che *dipende da tutti i valori della variabile*, ma non è molto utilizzata, specie negli sviluppi teorici, per la *non derivabilità* della funzione dovuta al valore assoluto degli scarti. Ad essa viene generalmente preferita un'altra misura della variabilità basata sul quadrato degli scarti dalla media.

Scostamento quadratico o scarto quadratico medio

$$\textit{Scostamento quadratico da } K = \sqrt{\frac{\sum_{i=1}^{i=n}(x_i - K)^2.f_i}{N}}$$

K può essere la media aritmetica, la mediana o qualsiasi altro valore medio preferito. Se K è la media aritmetica, lo scostamento quadratico medio si chiama anche *deviazione standard* e, in genere, viene sempre indicato con la lettera greca σ (sigma) quando ci si riferisce ad una intera popolazione di valori oppure con la lettera latina s (esse), quando ci si riferisce ad un campione di valori tratto dalla popolazione studiata.

E' la misura di dispersione più utilizzata.

Varianza **E' pari al *quadrato della deviazione standard:***

$$\sigma^2 = \frac{\sum_{i=1}^{i=n}(x_i - M)^2.f_i}{N}$$

Devianza **E' pari al *numeratore della varianza*. Ha una grande importanza in statistica, perché può essere scomposta in porzioni che sono di grande utilità per la teoria dell'analisi della varianza.**

Coefficiente di variazione

E' pari al *rapporto tra scostamento quadratico medio e media* della distribuzione moltiplicato per cento.

Essendo una misura relativa della variabilità, consente di comparare tra loro due o più distribuzioni le cui unità di misura sono molto diverse. In formule:

$$V = \frac{\sigma}{M}.100$$

Altre misure di variabilità si possono avere anche con riferimento agli scarti delle singole modalità da altri tipi di medie.

Nella statistica non parametrica è molto usato lo scostamento quadratico medio dalla mediana della distribuzione.

Momento centrato di ordine k per distribuzioni discrete

Il k-esimo momento rispetto ad una origine arbitraria A è definito dalla espressione:

$$\nu_k = \frac{1}{N}\sum_{i=1}^{i=n}(x_i - A)^k.f_i$$

Se A = M, il k-esimo momento rispetto alla media aritmetica (M) è pari a:

$$\mu_k = \frac{1}{N}\sum_{i=1}^{i=n}(x_i - M)^k.f_i$$

Se A = 0, si definisce il k-esimo momento rispetto all'origine degli assi (zero), cioè:

$$m_k = \frac{1}{N}\sum_{i=1}^{i=n}x_i^k.f_i$$

Momento di ordine k per distribuzioni continue

Nel caso di distribuzioni continue il posto del segno di sommatorio è preso dal segno di integrale. Quindi le tre espressioni date per i momenti di ordine k rispetto all'origine arbitraria, rispetto alla media e rispetto a zero sono:

$$\nu_k = \int_a^b (x - A)^k.f(x).dx,$$

$$\mu_k = \int_a^b (x - M)^k.f(x).dx,$$

$$m_k = \int_a^b x^k.f(x).dx,$$

in cui, al solito l'intervallo (a, b) indica il campo di definizione della funzione o distribuzione.

Da queste definizioni segue immediatamente che:

- ***il primo momento rispetto all'origine non è altro che la media aritmetica***

- ***il secondo momento centrato rispetto alla media è la varianza.***

Il terzo ed il quarto momento rispetto alla media, divisi rispettivamente per il cubo e per la quarta potenza dello scostamento quadratico medio, sono utilizzati in statistica anche per misurare l'asimmetria e l'appiattimento delle distribuzioni statistiche di tipo campanulare.

Correzioni di Sheppard

Supporre che una classe possa essere rappresentata dal suo valore centrale comporta errori nel calcolo dei momenti che possono essere corretti con le seguenti formule di Sheppard:

- per il momento secondo:

$$\mu_2 \text{ corretto} = \mu_2 - \frac{h^2}{12}$$

- per il momento quarto:

$$\mu_4 \text{ corretto} = \mu_4 - \frac{h^2}{2}.\mu_2 + \frac{7}{240}h^4$$

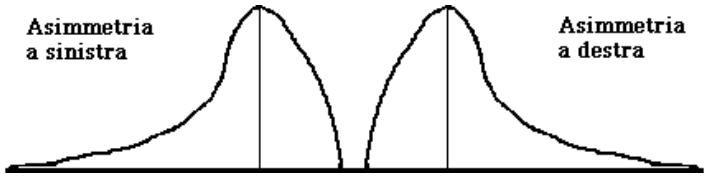
dove: h = ampiezza delle classi
mentre i momenti μ1 e μ3 non hanno bisogno di essere corretti.

Forma della distribuzione

Caratteristiche descrittive della forma: **Asimmetria-Appiattimento**

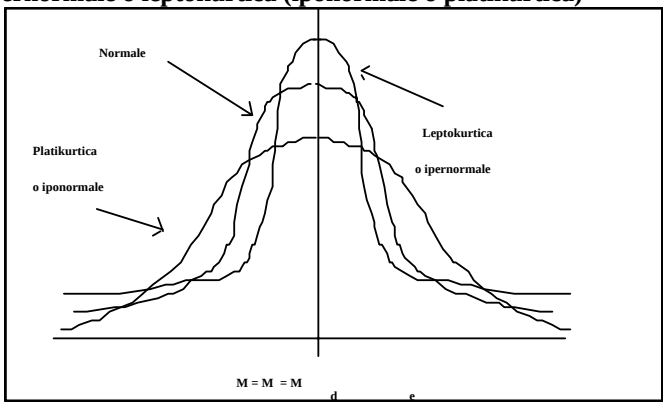
ASIMMETRIA (Skewness)

Una distribuzione si dice *simmetrica* se, con riferimento al valore mediano, le *frequenze* dei valori inferiori a quest'ultimo si distribuiscono così come le frequenze dei valori superiori ad esso ed in maniera *speculare* rispetto al valore centrale. Se ciò non si verifica la distribuzione è asimmetrica: positiva, se la coda si allunga verso i valori positivi, negativa in caso contrario. Una distribuzione simmetrica non è detto che sia unimodale.



APPIATTIMENTO (Kurtosis)

L'appiattimento o disnormalità è quella caratteristica che individua le distribuzioni i cui valori centrali e quelli estremi hanno frequenza più elevata (oppure meno elevata) di quella tipica di una distribuzione gaussiana, mentre i valori intermedi tra quelli estremi e quelli centrali hanno una frequenza meno elevata (oppure più elevata). In tal caso si parla di distribuzione ipernormale o leptokurtica (iponormale o platikurtica)



Misure di asimmetria

Misura basata sulla relazione tra media, moda e scostamento quadratico medio

$$S_k = \frac{M_1 - M_d}{\sigma}$$

Misura basata sulla relazione tra media, mediana e scostamento quadratico medio

$$S_k = \frac{M_1 - M_e}{\sigma}$$

Misura basata sul terzo momento rispetto alla media (dipende dall'unità di misura)

$$S_k = \frac{1}{N}\sum_{i=1}^{i=n}\left(x_i - M_1\right)^3 . f_i$$

Misura basata sul terzo momento rispetto alla media (non dipende dall'unità di misura)

$$\alpha_3 = \frac{\frac{1}{N}\sum_{i=1}^{i=n}\left(x_i - M_1\right)^3 . f_i}{\sigma^3} = \frac{\mu_3}{\sqrt{\mu_2^3}}$$

Coefficiente beta uno del Pearson

$$\beta_1 = \alpha_3^2 = \frac{\mu_3^2}{\mu_2^3}$$

Misure di appiattimento

1) Misura basata sul quarto momento rispetto alla media (dipende dall'unità di misura)

$$K_u = \frac{1}{N}\sum_{i=1}^{i=n}\left(x_i - M_1\right)^4 . f_i$$

2) Misura basata sul quarto momento rispetto alla media (non dipende dall'unità di misura) o coefficiente beta due del Pearson

$$\alpha_4 = \frac{\frac{1}{N}\sum_{i=1}^{i=n}\left(x_i - M_1\right)^4 . f_i}{\sigma^4} = \frac{\mu_4}{\mu_2^2}$$

Nelle distribuzioni gaussiane l'appiattimento misurato dal coefficiente del Pearson è $\beta_2 = 3$