Concetti base della statistica

Con il termine **popolazione** in statistica si intende un insieme finito o infinito di tutte le unità statistiche di cui si vuole indagare una certa caratteristica che le individua come omogenee.

L'**unità statistica** è l'elemento di base del collettivo su cui si rilevano le informazioni oggetto della rilevazione.

Con **variabile statistica** (o carattere statistico) si intende un particolare aspetto o caratteristica delle unità statistiche che si vogliono osservare (es. genere degli individui, reddito).

Le singole variabili statistiche si manifestano attraverso diverse **modalità**. Se la variabile è il genere, sono rilevanti solo le modalità «maschio» e «femmina».

Il fenomeno collettivo viene studiato attraverso l'**osservazione** e la **misurazione** di una o più caratteristiche delle unità statistiche.

Esempio

Consideriamo l'indagine statistica «genere dei residenti in Italia nel 2021».

Popolazione statistica: residenti in Italia nel 2021

Unità statistiche: i singoli individui residenti nel 2021

Variabile statistica: il genere degli italiani residenti

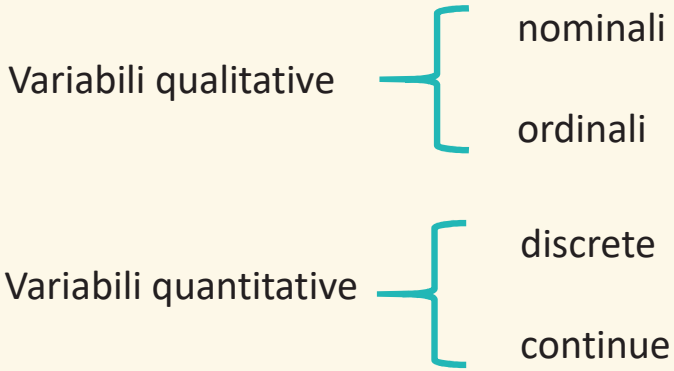
Modalità della variabile: maschio o femmina

Valore sintetico risultato dell'indagine: percentuale di maschi e femmine residenti in Italia nel 2021

I caratteri o variabili statistiche

Variabili qualitative: sono variabili che assumono attributi non numerici (sesso, credo religioso, colore degli occhi, ecc.)

Variabili quantitative: sono variabili i cui valori sono dei numeri reali (età, peso, altezza, temperatura, numero di persone in attesa alla fermata, ecc.)



Variabili qualitative

Variabile nominale (o qualitativa sconnessa)
Si tratta di variabili le cui modalità non sono logicamente ordinabili.

L'unico confronto tra le unità statistiche rispetto alle modalità assunte consiste nello stabilire se le unità posseggono modalità uguali o diverse. → **= o ≠**

Es. sesso, religione, colore della pelle, professione etc.

Variabile ordinale o ordinabile
Si tratta di variabili le cui modalità sono logicamente sequenziali, in ordine crescente o decrescente.
Non è possibile effettuare operazioni aritmetiche, ma si può stabilire una relazione di ordine. → **= o ≠ ma anche > <**

Es. i gradi militari, il giudizio valutativo, un indice di gradimento, ecc.

Variabili quantitative

Variabile discreta
Si tratta di variabili le cui modalità sono un sottoinsieme dei numeri naturali {1,2,3,…}.

Può trattarsi di un ***numero finito*** (es. voto all'esame di statistica {18,19,…,30})

O di ***un'infinità al più numerabile*** (es. numero di lanci di un dado)

Es. numero di figli, numero di automobili, ecc.

Variabile continua
Si tratta di variabili che possono assumere un numero infinito di valori contenuto in un intervallo reale.

Es. età, altezza, peso, distanza, temperatura, tempo ect.

Domande

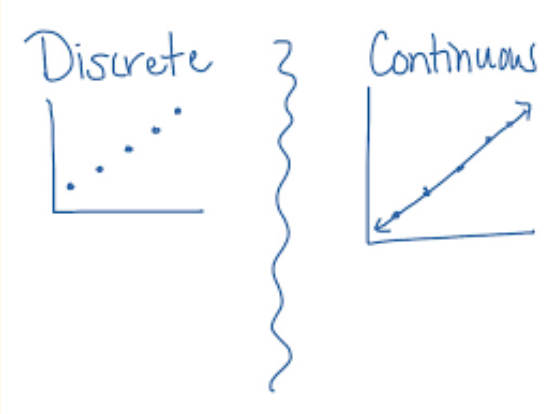
1 - Quali variabili possono essere conteggiate?

2 - I km percorsi da un ciclista sono?

3 - Il volume dell'acqua in una piscina è?

4 - Numero di studenti presenti in aula?

5 - Quali variabili sono misurabili?



Caratteri trasferibili

La **trasferibilità** vale solo per variabili quantitative.

Il carattere che un'unità statistica può cedere, anche parzialmente, ad un'altra è detto carattere trasferibile.

Ne sono un esempio il patrimonio o il reddito, nonché il numero di dipendenti di un'azienda o il numero di autovetture di una famiglia.

Alcuni caratteri quantitativi sono propri di una data unità statistica e non sono cedibili o trasferibili da questa ad altre unità, come per esempio la statura, il peso, l'età o il numero di figli partoriti da una donna.

Misure di tendenza centrale o di posizione

Gli **indici di tendenza centrale** sono numeri che esprimono la sintesi numerica di una distribuzione statistica di una variabile X.

Gli indici più noti sono:
1. la **media aritmetica**
2. la **mediana**
3. la **moda**

Sono utilizzati per fornire un'efficace rappresentazione del fenomeno studiato.
La scelta dell'indice di tendenza centrale idoneo a descrivere una distribuzione di valori dipende dalla natura (qualitativa o quantitativa) dei dati.

Media aritmetica

La **media aritmetica**, indicata con lettera greca μ, è calcolata sommando tutti i valori x_i (da 1 a N) della variabile X in oggetto e dividendo per il numero totale delle osservazioni N.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

ESEMPIO

n = 10
139
140
154
154
154
155
180
192
192
196

$$\mu = \frac{139 + 140 + 154 + 154 + 154 + 155 + 180 + 192 + 192 + 196}{10}$$

$$\mu = 165{,}6$$

Definizione di media di Chisini

Definizione di Chisini: data una distribuzione semplice di valori x1, x2 …,xn, una media è la quantità μ (lettera greca) che, se sostituita a ciascun termine della distribuzione, lascia inalterato il risultato della funzione f.

$$f(x1, x2,\ldots, xn) = f(\mu, \mu, \ldots, \mu)$$

Al variare della funzione f si ottengono tipi diversi di media.
Per esempio se f è la "somma dei numeri» la media relativa a f sarà la media aritmetica.

Criterio di rappresentatività

$$\mu = \frac{139 + 140 + 154 + 154 + 154 + 155 + 180 + 192 + 192 + 196}{10}$$

$$\mu = \frac{165{,}6+165{,}6+165{,}6+165{,}6+165{,}6+165{,}6+165{,}6+165{,}6+165{,}6+165{,}5}{10}$$

Proprietà della media

Internalità: il valore di un indicatore sintetico deve necessariamente trovarsi all'interno del campo di osservazione, ossia la media è sempre compresa tra il valore minimo e il valore massimo della distribuzione.

$$x_{min} \leq \mu \leq x_{max}$$

Esempio: prendiamo l'età di 3 fratelli, 32, 36 e 40.
La loro età media non potrà mai essere inferiore a 32 o maggiore di 40.

Proprietà della media

La somma degli scarti (differenze) dei valori osservati X dalla media aritmetica è zero, per cui la media è il **baricentro** di una distribuzione.

$$\sum_{i=1}^{n}(x_i - M) = 0$$

n = 10	μ	Delta
139	-165,6	-26,6
140	-165,6	-25,6
154	-165,6	-11,6
154	-165,6	-11,6
154	-165,6	-11,6
155	-165,6	-10,6
180	-165,6	14,4
192	-165,6	26,4
192	-165,6	26,4
196	-165,6	30,4
		0,00

Proprietà della media

La somma dei quadrati degli scarti (differenze) dai valori X dalla media è un **MINIMO**.
Non esiste nessun valore diverso dalla media per cui le differenze delle x al quadrato, diano una somma inferiore.

$$\sum_{i=1}^{n}(x_i - M)^2 \leq \sum_{i=1}^{n}(x_i - C)^2$$

per qualsiasi valore di C

Esempio: per i valori 2,3,4 la M è 3

Lo scarto al quadrato: $(2-3)^2+(3-3)^2+(4-3)^2 = 2$ = **MINIMO**

Proprietà della media

Invarianza rispetto a trasformazioni lineari: se ad ogni termine della distribuzione viene sommata o moltiplicata una quantità, la media aritmetica risulterà incrementata o moltiplicata della sessa quantità.

Esempio: 2, 3, 4 con media 3. Se sommiamo 10 a ogni termine 12, 13,14 la media sarà 13.

Esempio: 2, 3, 4 con media 3. Se moltiplichiamo per 10 ogni termine 20, 30, 40 la media sarà 30.

La distribuzione di frequenza

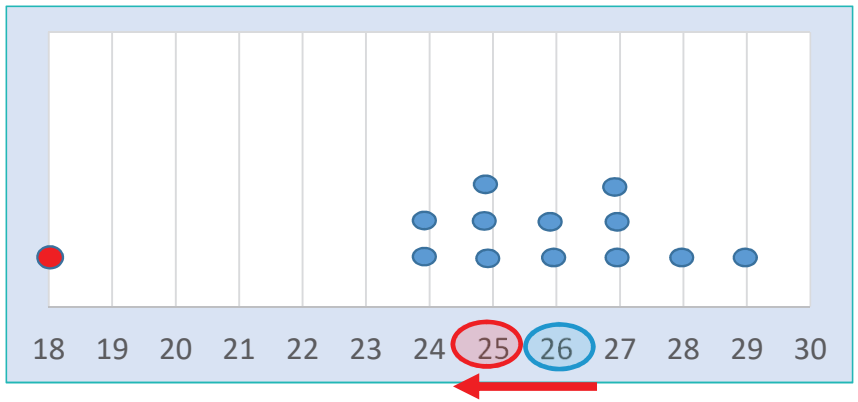
La **frequenza** è un numero che indica quante volte è stato rilevato un certo valore (o modalità) durante l'osservazione.
Se ad ogni valore assunto dalla variabile statistica viene associata la relativa frequenza si ottiene:

Distribuzione di frequenza

Modalità della variabile statistica «voto»	Frequenza (numero di esami)
24	2
25	3
26	2
27	3
28	1
29	1

μ = ~ 26

Diagramma a punti



Limiti della media aritmetica

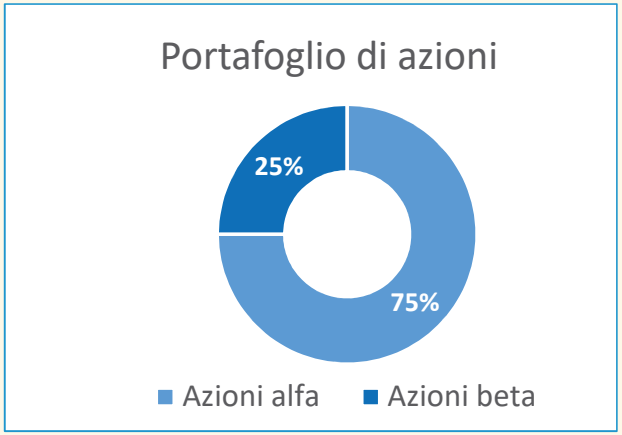
- L'utilizzo è consigliabile solo quando l'insieme dei dati è abbastanza omogeneo.
  E' sconsigliato in presenza di valori estremi o anomali (**outlier**), perché il suo valore potrebbe essere poco aderente alla realtà.
  La media aritmetica costituisce il **baricentro** di una distribuzione. Questo vuol dire che un valore eccezionalmente diverso dagli altri sposta il baricentro riducendo la rappresentatività del valore medio per le unità considerate. Per questo motivo è un **indice poco robusto**.

- Può essere calcolata solo per dati numerici, quantitativi, mentre per le rilevazioni qualitative non ha senso.

- Ogni osservazione pesa come tutte le altre.

Media aritmetica ponderata

Qual è stato il rendimento del portafoglio alla fine dell'anno?



	Alfa	Beta
Valore	7.500	2.500
Rendimento	2%	40%

Se si calcola la media aritmetica il rendimento sarebbe pari 10.000€x21% (42%/2) =2.100€ (12.100€) → **VALORE NON CORRETTO**

Media aritmetica ponderata

In realtà i 2 titoli hanno un **peso diverso**: Alfa rappresenta il 25% del portafoglio e Beta il 75%.
Il calcolo corretto deve tener conto dei diversi pesi:

$$= \frac{\textbf{2.500}*40\%+\textbf{7.500}*2\%}{\textbf{2.500+7.500}} = 11{,}50\% = 1.150€\ (11.150€)$$

La **media aritmetica ponderata** si calcola sommando i valori osservati già corretti per la loro rilevanza, cioè moltiplicati per il loro peso.

$$\mu = \frac{\sum_{i=1}^{N} x_i W_i}{\sum_{i=1}^{N} W_i}$$

peso wi è il **coefficiente di importanza**

Mediana

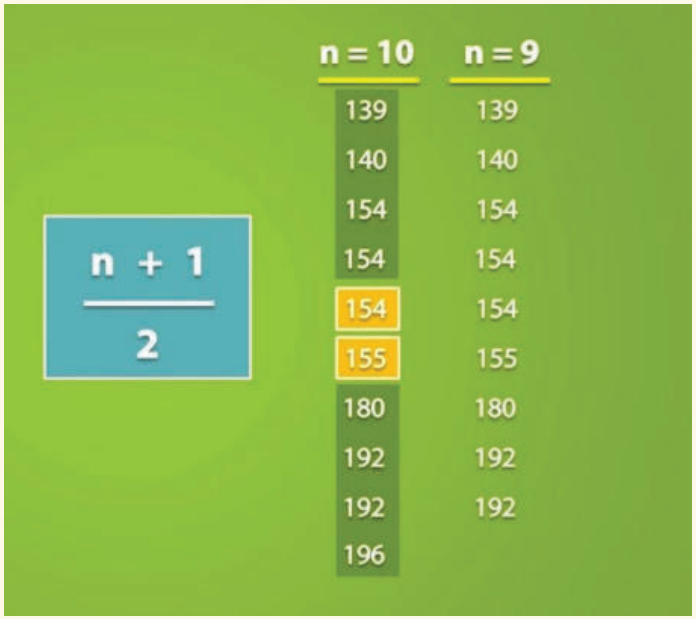
La **mediana** (μe) è quel valore di una variabile X che, rispetto a una serie di dati ordinati (distribuzione ordinata delle osservazioni dal più piccolo al più grande), occupa la posizione centrale, cioè li divide in 2 parti uguali.

Posso calcolare la mediana solo per le variabili per cui sia possibile stabilire un ordinamento tra le modalità.

Metà delle unità posseggono modalità inferiori alla mediana e metà superiori alla mediana.

Mediana

Quando le osservazioni sono pari si hanno 2 valori centrali.
Per identificare la mediana basta calcolare la media aritmetica dei 2 valori mediani.



Mediana

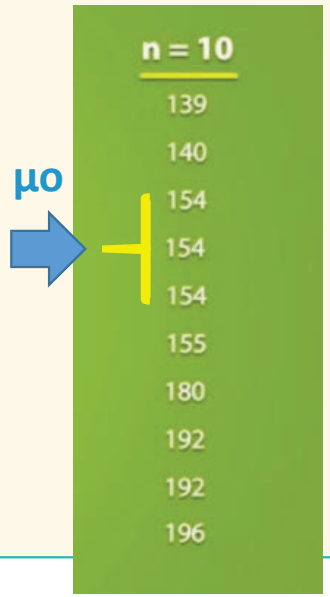
Anche se l’insieme di dati contiene 2 o più osservazioni di pari valore, ciascun osservazione va trattata separatamente quando si dispongono i dati in ordine crescente.

La **mediana** è un indice altamente rappresentativo e supera il problema dei valori anomali, in quanto non tiene conto di come i singoli valori siano grandi o dispersi.

E’ la misura di posizione usata più spesso per i dati sul reddito e sul patrimonio, poiché pochi valori estremamente elevati possono influenzare la media.

Moda

La **moda** (μo) è la modalità del carattere alla quale corrisponde la massima frequenza (assoluta, relativa o percentuale).
In sintesi è il valore che ricorre più volte nel collettivo.
La moda, a differenza della media aritmetica, si calcola anche su un carattere qualitativo (es. il colore).



Moda

Possono verificarsi situazioni in cui la frequenza più elevata si osserva per due o più valori diversi.
Se due modalità posseggono la massima frequenza, allora la distribuzione si dice essere **bimodale** (per n modalità multimodale).

Modalità «colore occhi»	Frequenza
Verde	3
Azzurro	2
Marrone	3
Nero	1

Quando tutte le osservazioni hanno la stessa frequenza, non può essere individuata nessuna moda.

Esercizio

Calcolare la media aritmetica, la mediana e la moda per le seguenti tre distribuzioni:

A: [1,12,3,38,25,26,35,27,36,24,40]

B: [1,12,3,38,25,26,35,27,36,24,400]

C: [-100,12,3,38,25,26,35,27,36,24,40]

	Media aritmetica	Mediana
A		
B		
C		

Esercizio

Calcolare la media aritmetica, la mediana e la moda per le seguenti tre distribuzioni:

A: [1,12,3,38,25,26,35,27,36,24,40]

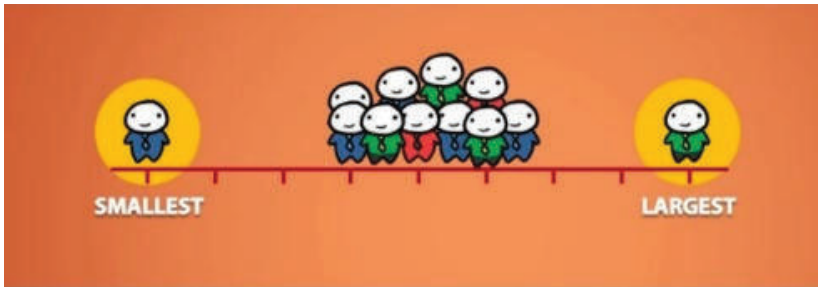
B: [1,12,3,38,25,26,35,27,36,24,400]

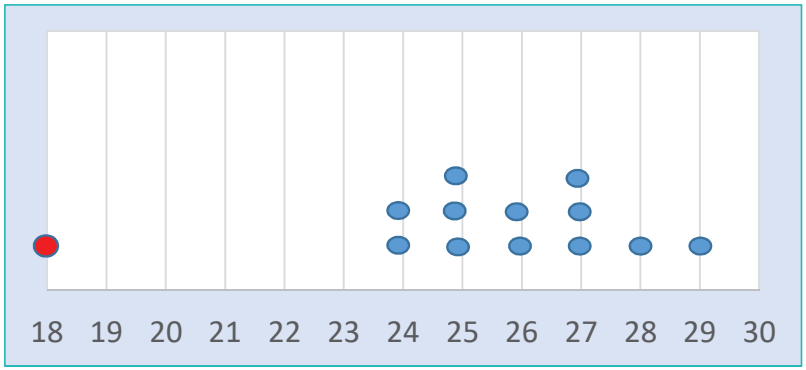
C: [-100,12,3,38,25,26,35,27,36,24,40]

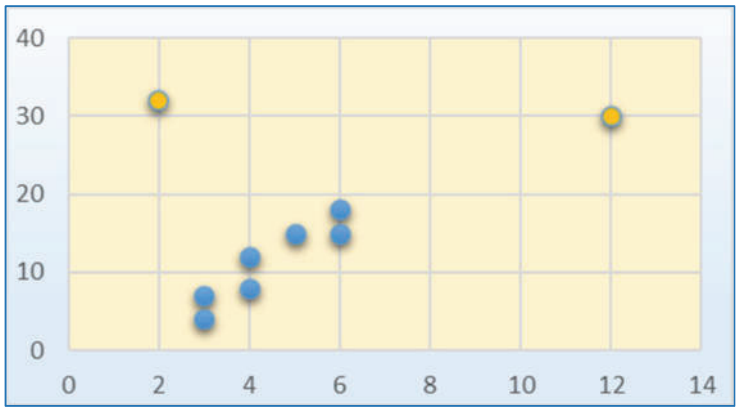
	Media aritmetica	Mediana
A	24,3	26
B	57,0	26
C	15,1	26

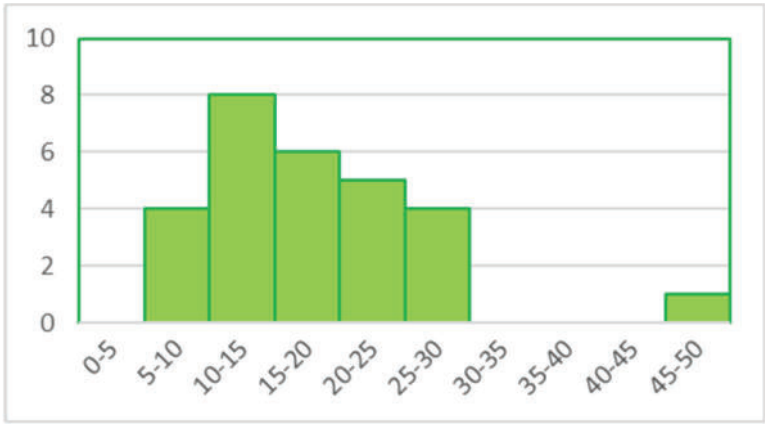
Outlier

Un valore anomalo è un punto dati significativamente diverso dai valori delle restanti osservazioni.
Può essere insolitamente grande o insolitamente piccolo.









Outlier

Dati
- 320°C
15°C
21°C
24°C
26°C
32°C
32°C

Indici	Con outlier	Senza outlier
Media	-24,28	25
Mediana	24	25
Moda	32	32
Range	288	17

Outlier

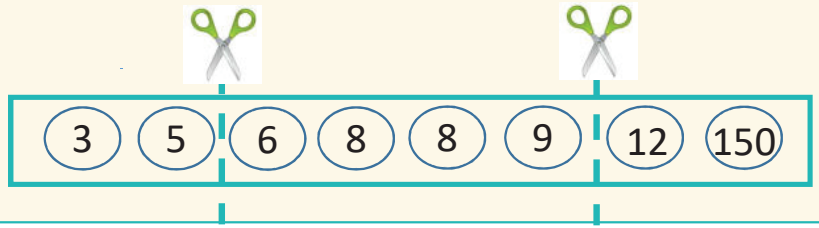
Dati
- 320°C
15°C
21°C
24°C
26°C
32°C
32°C

Indici	Con outlier	Senza outlier
Media	-24,28	25
Mediana	24	25
Moda	32	32
Range	288	17

Media troncata

Per eliminare l'influenza dei valori anomali si può calcolare la **media troncata** (trimmed mean). Si ottiene eliminando una percentuale dei valori più grandi e più piccoli da un insieme di dati ordinati e poi calcolando la media dai valori rimanenti.

Esempio: media troncata al 50% significa che occorre eliminare il 25% dei valori minori e il 25% dei valori maggiori.
Si parte dal numero delle osservazioni. In questo caso sono 8, quindi elimino il 50% di 8, ossia 4 osservazioni (2 minori e 2 maggiori).
Rimangono 4 osservazioni = 6, 8, 8, 9 con M=7,75



Media geometrica

Nella **capitalizzazione composta** l'interesse prodotto in ogni periodo si somma al capitale e produce a sua volta interessi tale che il Montante = Capitale$(1+i)^t$

Anni	Capitale	Tassi di Interessi
1	1.000	4%
2	1.040	2,5%
3	1.066	-3%
	1.034	

Capitale = 1.000€

Gli interessi maturati sono andati ad aumentare il capitale

Fattori di crescita		
1,04	1,025	0,97 (1-0,03)

Media aritmetica = $\frac{1,04+1,025+0,97}{3}$ = 1,0116 (1,16%)

1.000*(1,0116)^^{3} = **1.035,2 > 1.034** ?

Media geometrica

La **media geometrica** è data dal prodotto delle modalità x.

$$\mu g = \sqrt[n]{x_1 * x_2 * \ldots x_n} = (x_1 * x_2 * \ldots x_n)^{(1/n)}$$

Media geometrica = $(1{,}04 * 1{,}025 * 0{,}97)^{\wedge(1/3)} = 1{,}0112$ (1,12%)

$1.000 * (1{,}0112)^{\wedge 3} =$ **1.034**

Viene utilizzata per tener conto degli effetti compositi quando le variabili sono dipendenti l'una dall'altra.

Trova applicazione nel mondo della finanza (tassi d'interesse, tassi di crescita), ma anche in settori come la demografia, la medicina.
Non va usata quando la distribuzione potrebbe contenere valori negativi perché è impossibile calcolare la radice di ordine pari di un valore negativo.

Proprietà della media geometrica

Vale anche per la media geometrica il principio di internalità:

$\min(x_i) \leq$ media geometrica $\leq$ media aritmetica $\leq \max(x_i)$

La media geometrica è sempre inferiore alla media aritmetica. Può essere uguale solo se i numeri delle serie sono tutti uguali.

Esempio: X che presenta le modalità: **1,3,2,5,4,6,8,10,20,10**

Media aritmetica=(1+3+2+…+10)/10=69/10=6,900

Media geometrica=(1*3*2*…*10)^(1/10) =(11520000)^(0,1) = 5,083

1 ≤5,083 ≤6,900 ≤20

Indicatori sintetici

L'idea centrale è di riassumere una distribuzione in una sintesi che ne evidenzi le caratteristiche essenziali.
Le medie sintetizzano in un solo valore la distribuzione.
La sintesi sarà un numero che andrà a sostituire tutte le osservazioni, ciò implica una perdita d'informazione rispetto ai dati elementari e per questo occorre cercare di arrivare agli obiettivi dell'indagine, minimizzando tale perdita di informazioni.

Le medie analitiche (ferme):
media aritmetica
media geometrica
trimmed mean
Si calcolano su variabili quantitative attraverso operazioni algebriche sui valori del carattere.

Le medie di posizione (lasche):
mediana e moda
Si calcolano anche su variabili qualitative. Esprimono la posizione, sulla scala ordinata delle misurazioni, intorno alla quale si addensano le osservazioni.

Domande

1- Viene svolta un'indagine sulle famiglie per rilevare il numero di macchine possedute.
Indicare:
Unità di analisi
Variabile statistica
Tipo di variabile

2-Trovare la mediana e la moda dei seguenti valori 11, 14, 17, 14, 23, 26, 18

3- Quali delle seguenti variabili non sono qualitative?

- ◯ Mesi dell'anno
- ◯ Età di una persona
- ◯ Pressione arteriosa
- ◯ Numero del cellulare

4- In una distribuzione quantitativa esiste sempre una sola mediana?

5- Dato che la moda è la modalità con la massima frequenza, può essere individuata direttamente da una distribuzione di frequenza?

6- Quali di questi indici non sono influenzati dagli outliers?

- ◯ Media
- ◯ Range
- ◯ Mediana
- ◯ Moda

7- Negli ultimi 3 anni l'investimento fatto è cresciuto del 3% annuo. Quale della seguente affermazioni è vera:

- ◯ La media geometrica è inferiore alla media aritmetica
- ◯ La media geometrica è uguale alla media aritmetica
- ◯ La media geometrica è superiore alla media aritmetica

8- Esiste un numero diverso dalla media che minimizza la somma degli scarti dalle x al quadrato?

9- La media aritmetica di 10 osservazioni è 10. Se tutte le osservazioni sono incrementate del 10%, qual è l'incremento della media?

- ◯ 11
- ◯ 10
- ◯ 1

10- Calcolare la mediana per i seguenti valori:

Gruppo sanguigno	A	B	AB	0
N. di soggetti	3	5	1	10

11- Quali dei seguenti indici non si basa su tutte le osservazioni?

- ◯ Media geometrica
- ◯ Media aritmetica
- ◯ Moda