

Preliminari

- **Popolazione statistica** $\square$ **campione statistico** $\square$ **Unità statistica** $\square$ Ogni unità statistica presenta uno o più caratteri e quindi si parla di **carattere di una unità statistica**
- Un carattere si presenta mediante una o più modalità e quindi si parla di **modalità di un carattere** $\square$ Se le modalità di un carattere esprimono delle quantità , cioè se le modalità sono espresse da numeri allora il carattere delle singole unità statistiche della popolazione statistica rappresenta una **variabile statistica** . Simbolo usato **X** $\square$ **Variabile statistica discreta , variabile statistica continua** $\square$ Se le modalità del carattere esprimono delle qualità e quindi non possono essere rappresentate da numeri allora il carattere delle singole unità statistiche rappresenta una **mutabile statistica** . Simbolo usato **M** .
- n_r = **frequenza assoluta** di una modalità = numero di volte in cui compare la modalità
- $C_r = n_1 + n_2 + \dots + n_r$ = **frequenza assoluta cumulata**
- **k** = numero di modalità
- $N = n_1 + n_2 + \dots + n_k$ = somma di tutte le frequenze assolute = numero di unità statistiche dell'intera popolazione statistica
- $f_r = \frac{n_r}{N}$ = **frequenza relativa**
- $p_r = 100 \cdot f_r = \frac{n_r}{N} \cdot 100$ = **frequenza relativa percentuale**
- $F_r = f_1 + f_2 + \dots + f_r = \frac{n_1}{N} + \frac{n_2}{N} + \dots + \frac{n_r}{N} = \frac{n_1 + n_2 + \dots + n_r}{N} = \frac{C_r}{N} =$
= **frequenza relativa cumulata**
- $P_r = 100 \cdot F_r$ = **frequenza relativa percentuale cumulata**
- **Valore centrale** della classe $x_{i-1} \text{ ——— } x_i$ $c_i = \frac{x_{i-1} + x_i}{2}$
- **Ampiezza della classe** $x_{i-1} \text{ ——— } x_i$ $\alpha_i = x_i - x_{i-1}$
- $h_i = \frac{n_i}{\alpha_i} = \frac{\text{frequenza assoluta di ogni singola classe}}{\text{ampiezza della singola classe}} =$ **densità di frequenza**
- Una distribuzione di frequenza è una distribuzione che ha una delle seguenti strutture , cioè le **distribuzioni di frequenza** sono quelle rilevazioni statistiche che hanno modalità e frequenze come quelle sotto elencate :

$X \begin{cases} x_1 & x_2 & \cdots & x_k \\ n_1 & n_2 & \cdots & n_k \end{cases}$ **distribuzione di frequenza assoluta**

$X \begin{cases} x_1 & x_2 & \cdots & x_k \\ f_1 & f_2 & \cdots & f_k \end{cases}$ **distribuzione di frequenza relativa**

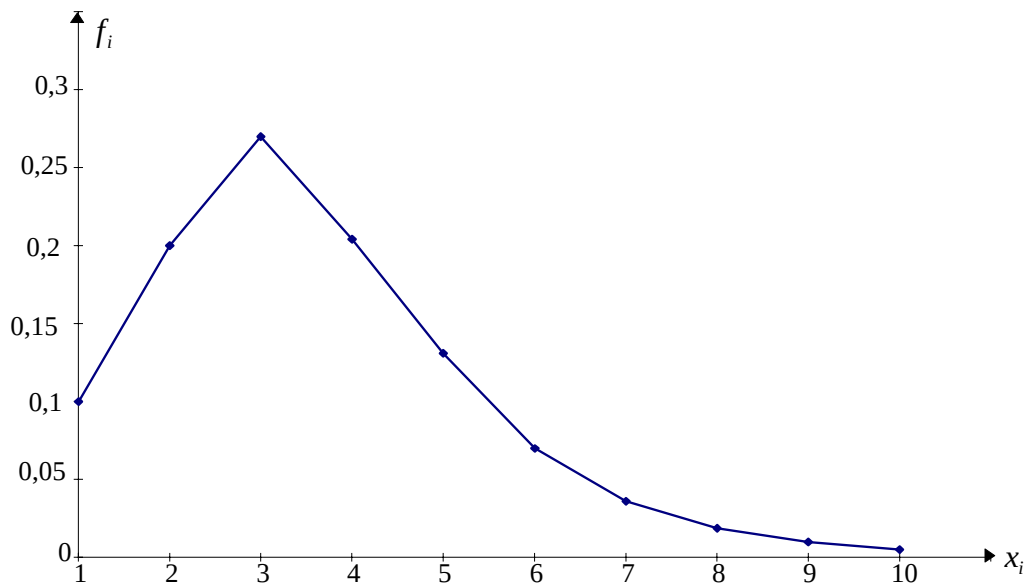
$X \begin{cases} x_1 & x_2 & \cdots & x_k \\ F_1 & F_2 & \cdots & F_k \end{cases}$ **distribuzione di frequenza relativa cumulata**

$X \begin{cases} x_1 & x_2 & \cdots & x_k \\ P_1 & P_2 & \cdots & P_k \end{cases}$ **distribuzione di frequenza percentuale cumulata**

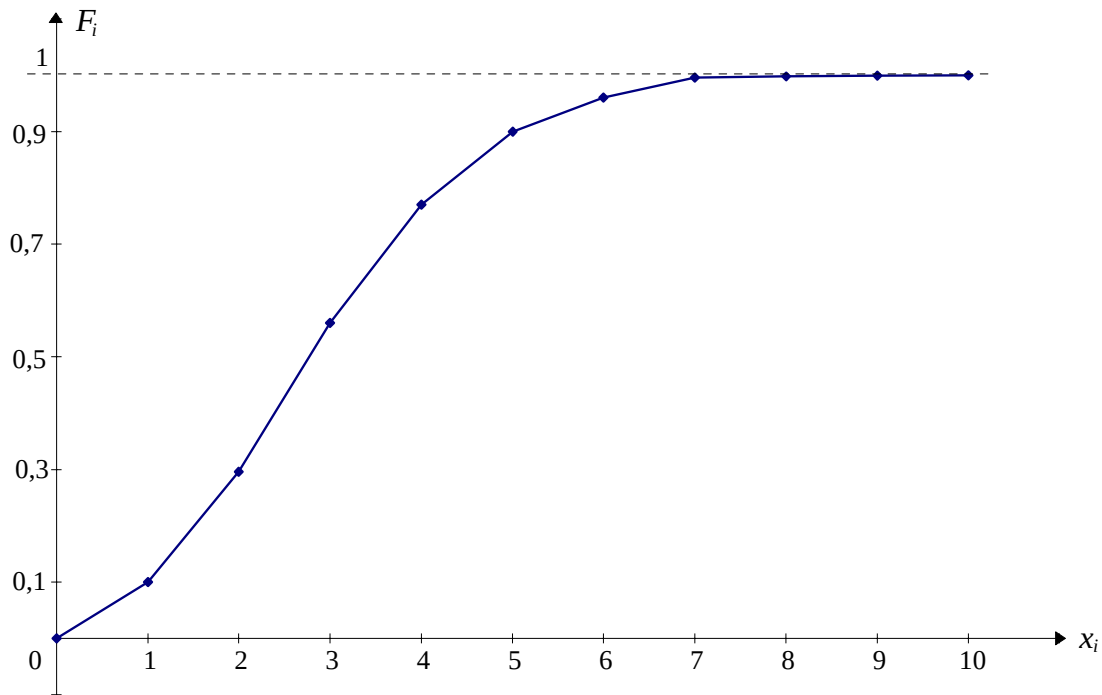
• Per **distribuzioni di frequenze** derivate da variabili statistiche si può ricorrere a due tipi

di rappresentazioni grafiche :

• il **diagramma di frequenza** se in ascissa riportiamo le modalità x_r della distribuzione **X** ed in ordinate le rispettive **frequenze** (assolute o relative) . I punti del piano così individuati si uniscono con una spezzata .



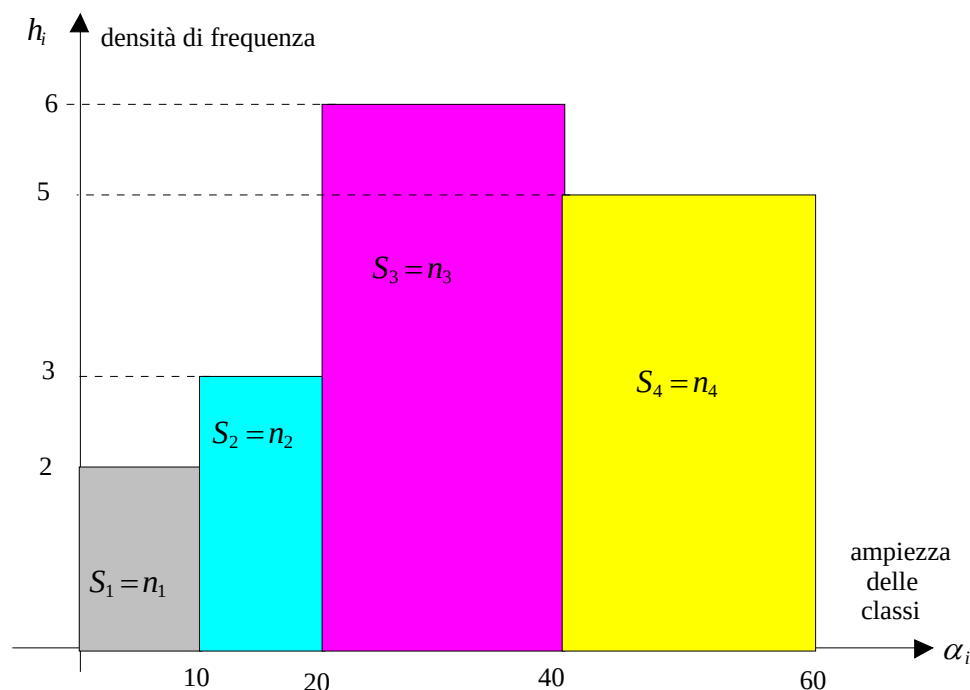
• la **funzione di ripartizione** se in ascissa riportiamo le modalità x_r della distribuzione **X** ed in ordinate le rispettive **frequenze relative cumulate** . I punti del piano così individuati si uniscono con una spezzata .



- Se in ascissa riportiamo l'ampiezza di ciascuna classe ed in ordinata riportiamo le rispettive densità di frequenze otteniamo l' **istogramma di frequenza** .

Disegnare l' **istogramma** della seguente distribuzione di frequenza in classi .

Classi $x_{i-1} \text{ — } x_i$	frequenze assolute n_i	ampiezza della classe $\alpha_i = x_i - x_{i-1}$	densità di frequenza $h_i = \frac{n_i}{\alpha_i}$
0—10	20	10	2
10—20	30	10	3
20—40	120	20	6
40—70	100	20	5



Indici statistici descrittivi

□ Gli **indici statistici descrittivi** riassumono alcune caratteristiche delle distribuzioni di frequenza consentendo il confronto fra le diverse distribuzioni .

Noi faremo riferimento a tre diversi tipi di **indici descrittivi** .

- a) quelli che servono a localizzare la distribuzione e li chiameremo **indici di posizione**
- b) quelli che servono a misurare la **variabilità** o **dispersione** di una distribuzione e li chiameremo **indici di variabilità**
- c) quelli che servono ad individuare la **forma** della distribuzione e li chiameremo **indici sulla forma** .

Indici di posizione

Gli **indici statistici di posizione** (detti anche **indici statistici di locazione** o parametri di posizione) servono per **localizzare** la distribuzione . Tentano di individuare intorno a quale valore del carattere la distribuzione si concentra . In tal senso gli **indici di posizione** servono ad individuare quel valore del carattere che è **rappresentativo** dell'intera distribuzione e quindi la **sintetizza** .

I più usati indici statistici sono le *Medie algebriche* $\left\{ \begin{array}{l} 1) \text{ Media aritmetica} \\ 2) \text{ Media geometrica} \\ 3) \text{ Media armonica} \\ 4) \text{ Media quadratica} \end{array} \right.$
(si calcolano tenendo conto di tutti i valori della distribuzione)

e le *Medie di posizione* $\left\{ \begin{array}{l} 1) \text{ Moda} \\ 2) \text{ Mediana} \end{array} \right.$
(si calcolano tenendo conto solo di particolari valori della distribuzione)

Media aritmetica semplice

$$\mu = \bar{x} = m = \frac{x_1 + x_2 + \dots + x_k}{k} = \frac{\sum_{i=1}^k x_i}{k}$$

se il carattere X assume ciascuna modalità x_i una

sola volta .

Media aritmetica ponderata

$$\mu = A = M_1 = \bar{x} = m = \frac{n_1 x_1 + n_2 x_2 + \dots + n_k x_k}{n_1 + n_2 + \dots + n_k} = \frac{\sum_{i=1}^k n_i x_i}{\sum_{i=1}^k n_i} = \frac{\sum_{i=1}^k n_i x_i}{N}$$

se il carattere X assume ciascuna modalità x_i n_i volte .

Proprietà della media aritmetica

- la media aritmetica rappresenta il **baricentro** della distribuzione
- la media aritmetica è sempre compresa tra il valore più piccolo e quello più grande della distribuzione
- La **somma algebrica degli scarti dalla media aritmetica è nulla** , cioè :

$$\sum_{i=1}^k \hat{x}_i = \sum_{i=1}^k (x_i - \mu) = 0$$

$$\sum_{i=1}^k \hat{x}_i \cdot n_i = \sum_{i=1}^k (x_i - \mu) \cdot n_i = 0$$

$$\hat{x}_i = x_i - \mu$$

- la media aritmetica è l'unico valore che rende minima la somma dei quadrati degli scarti per le

rispettive frequenze :

$$\sum_{i=1}^k (x_i - a)^2 \cdot n_i = \text{minimo}$$

se e solo se $a = \mu$

Media geometrica

Distribuzione per unità semplici , cioè se il carattere X assume ciascuna modalità x_i una sola volta

$$M_g = G = M_o = \sqrt[k]{x_1 \cdot x_2 \cdot \dots \cdot x_k} = \sqrt[k]{\prod_{i=1}^k x_i}$$

Distribuzioni per modalità x_i raggruppate , cioè se il carattere X assume ciascuna modalità x_i n_i volte .

$$M_g = G = M_o = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot \dots \cdot x_k^{n_k}} = \sqrt[N]{\prod_{i=1}^k x_i^{n_i}} = x_1^{\frac{n_1}{N}} \cdot x_2^{\frac{n_2}{N}} \cdot \dots \cdot x_k^{\frac{n_k}{N}} = x_1^{f_1} \cdot x_2^{f_2} \cdot \dots \cdot x_k^{f_k}$$

- Mentre la media aritmetica μ è quel valore che sostituito agli addendi lascia immutata la somma

$$\mu \cdot N = \sum_{i=1}^k x_i \cdot n_i \quad \text{la media geometrica è quel valore che ne lascia immutato il prodotto}$$

$$(M_g)^N = \prod_{i=1}^k x_i \cdot n_i .$$

Media quadratica

Distribuzioni per unità semplici :

$$Q = M_Q = M_2 = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_k^2}{k}} = \sqrt{\frac{\sum_{i=1}^k x_i^2}{k}}$$

Distribuzioni per modalità x_i raggruppate :

$$Q = M_Q = M_2 = \sqrt{\frac{n_1 x_1^2 + n_2 x_2^2 + \dots + n_k x_k^2}{n_1 + n_2 + \dots + n_k}} = \sqrt{\frac{\sum_{i=1}^k n_i x_i^2}{\sum_{i=1}^k n_i}} = \sqrt{\frac{\sum_{i=1}^k n_i x_i^2}{N}}$$

Media armonica semplice :

$$H = M_{-1} = \frac{k}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_k}} = \frac{k}{\sum_{i=1}^k \frac{1}{x_i}}$$

media armonica ponderata :

$$H = M_{-1} = \frac{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \dots + \frac{n_k}{x_k}}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \dots + \frac{n_k}{x_k}} = \frac{\sum_{i=1}^k \frac{n_i}{x_i}}{\sum_{i=1}^k \frac{n_i}{x_i}} = \frac{N}{\sum_{i=1}^k \frac{n_i}{x_i}}$$

N.B. Per le distribuzioni per classi le formule sono uguali a quelle delle distribuzioni per modalità , ma alle classi bisogna sostituire il rispettivo **valore centrale** .

La moda

La **moda** M_d della distribuzione di frequenza X è quella modalità (o classe di modalità) a cui corrisponde la **massima frequenza** (relativa o assoluta) .

A differenza della media aritmetica e di quella geometrica, una distribuzione X può ammettere più di un valore modale. In particolare, una distribuzione è **unimodale** se ammette un solo valore modale, **bimodale** se ne ammette due, **trimodale**, se ne ammette tre.

Mediana

La **mediana** M_e di una distribuzione X è quel valore M_e di X che occupa il posto centrale della successione ordinate di modalità. Con parole diverse possiamo dire che la **mediana** di una distribuzione X è quel valore M_e per cui la **frequenza relativa cumulata** F_r vale 0,5 e quindi **bipartisce** la distribuzione.

Per calcolare la mediana di una distribuzione X si procede come segue:

Si ordinano le modalità x_i in senso crescente $x_1 \leq x_2 \leq \dots \leq x_k$

a) dati non raggruppati: k dispari

La mediana è rappresentata dal termine che nella graduatoria $x_1, x_2, \dots, x_k$ occupa il posto $\frac{k+1}{2}$.

b) dati non raggruppati: k pari Si hanno due posti mediani: $P_1 = \frac{k}{2}$ $P_2 = \frac{k}{2} + 1$

La mediana è rappresentata dal termine che nella graduatoria $x_1, x_2, \dots, x_k$ occupa il posto

$$\frac{P_1 + P_2}{2} = \frac{\frac{k}{2} + \frac{k}{2} + 1}{2} = \frac{k+1}{2} \quad (\text{mediana convenzionale})$$

c) dati raggruppati e modalità del carattere discontinue

Si utilizzano le **frequenze cumulate** $C_i = n_1 + n_2 + \dots + n_i$. Si trova il valore mediano delle frequenze cumulate. (Volendo potremmo utilizzare la frequenza relativa cumulata F_i).

$$\mathbf{N \text{ dispari}} \Rightarrow P = \frac{N+1}{2} \quad \mathbf{N \text{ pari}} \Rightarrow P_1 = \frac{N}{2} \quad P_2 = \frac{N}{2} + 1 \quad P = \frac{P_1 + P_2}{2} = \frac{N+1}{2}$$

Il primo valore maggiore od uguale a $P = \frac{N+1}{2}$ ci consente di calcolare M_e , cioè a

$$P = \frac{N+1}{2} \text{ corrisponde la modalità } x_i \text{ che è il } \mathbf{\text{valore mediano}}.$$

Esempi

N dispari

x_i	n_i	C_i
3	2	2
5	4	6
9	3	9

Il primo numero maggiore di 5 nella colonna C_i è 6 e ad esso corrisponde il numero 5 sotto la colonna delle modalità x_i . Il numero 5 è il **valore mediano**

della distribuzione. $N = 9 \quad \frac{N+1}{2} = \frac{9+1}{2} = 5$

N pari

x_i	n_i	C_i
1	3	3
2	2	5
4	4	9
10	1	10

$$P_1 = \frac{N}{2} = 5 \quad P_2 = \frac{N}{2} + 1 = 6 \quad P = \frac{P_1 + P_2}{2} = \frac{5 + 6}{2} = 5,5$$

a $P_1 = 5$ corrisponde la modalità $\bar{x} = 2$ a $P_2 = 6$ corrisponde la modalità

$$\bar{\bar{x}} = 4 \quad M_2 = \frac{\bar{x} + \bar{\bar{x}}}{2} = \frac{2 + 4}{2} = 3 \quad N = 10$$

d) dati raggruppati in classi ; carattere continuo

Si determina la **classe mediana** dopo avere costruito la colonna della **classe cumulata**. Supponiamo che la classe mediana sia $x_{i-1} \text{ — } x_i$, che corrisponde alla prima frequenza

cumulata $\geq \bar{C} = \frac{N}{2}$. Poi si applica la seguente formula :

$$M_e = x_{i-1} + \frac{\bar{C} - C_{i-1}}{C_i - C_{i-1}} \cdot (x_i - x_{i-1}) = x_{i-1} + \frac{\frac{N}{2} - C_{i-1}}{C_i - C_{i-1}} \cdot (x_i - x_{i-1})$$

che si deduce dalla proporzione : $(M_e - x_{i-1}) : (x_i - x_{i-1}) = (\bar{C} - C_{i-1}) : (C_i - C_{i-1})$

Qualche autore per N dispari sceglie $\bar{C} = \frac{N+1}{2}$.

Al posto delle frequenze cumulate C_i possiamo utilizzare le frequenze relative cumulate F_i .

Se le modalità di X sono **raggruppate in classi** occorre individuare la classe mediana

$x_{i-1} \text{ — } x_i$ cui corrisponde $F_i \geq \frac{1}{2} = 0,5$ ed è tale che alla classe immediatamente precedente

corrisponda $F_i \leq \frac{1}{2} = 0,5$. La formula da utilizzare è la seguente :

$$M_e = x_{i-1} + \frac{0,5 - F_{i-1}}{F_i - F_{i-1}} \cdot (x_i - x_{i-1})$$

che si ricava dalla proporzione : $(M_e - x_{i-1}) : (x_i - x_{i-1}) = (0,5 - F_{i-1}) : (F_i - F_{i-1})$ ottenuta sotto l'ipotesi che le frequenze assolute n_i si distribuiscono uniformemente all'interno della classe mediana .

Esempio

$x_i - x_{i-1}$	n_i	C_i
0 — 4	15	15
4 — 10	30	45
10 — 10	20	65

$$N = 65 \quad \bar{C} = \frac{N + 1}{2} = \frac{65 + 1}{2} = 33 \quad \text{La prima frequenza}$$

cumulata dopo 33 è 45 . La classe mediana è : 4 |— 10

$$x_{i-1} = 4, \quad x_i = 10 \quad C_{i-1} = 15 \quad C_i = 45$$

$$M_e = 4 + \frac{33 - 15}{45 - 15} \cdot (10 - 4) = 4 + \frac{18}{30} \cdot 6 = 4 + 3,6 = 7,6$$

Se poniamo :

- $\ell_i = x_{i-1}$ **estremo inferiore** della classe mediana x_{i-1} — x_i
- $L_i = x_i$ **estremo superiore** della classe mediana x_{i-1} — x_i
- $n_i = C_i - C_{i-1}$ **frequenza assoluta** della classe mediana x_{i-1} — x_i
- $\alpha_i = x_i - x_{i-1}$ **ampiezza della classe mediana** x_{i-1} — x_i
- C_i **frequenza cumulata della classe mediana** x_{i-1} — x_i
- C_{i-1} **frequenza cumulata della classe che precede la classe mediana** x_{i-1} — x_i
- F_i **frequenza relativa cumulata della classe mediana** x_{i-1} — x_i
- F_{i-1} **frequenza relativa cumulata della classe che precede la classe mediana** x_{i-1} — x_i

allora la formula che ci consente di calcolare la mediana è la seguente :

$$M_e = \ell_i + \frac{\frac{N}{2} - C_{i-1}}{n_i} \cdot \alpha_i$$

Indici di variabilità

Una volta calcolato un indice di posizione (**media**), dobbiamo elaborare qualche indice che ci dia una misura della sua **rappresentatività** per l'intera distribuzione. Per raggiungere questo obiettivo studiamo la **variabilità** di una distribuzione X . La **variabilità** si manifesta attraverso la **dispersione** e la **disuguaglianza**. La **dispersione** caratterizza il maggiore o minore addensamento delle frequenze attorno ad un valore medio, la **disuguaglianza** mette in evidenza la diversità esistente tra le frequenze osservate senza riferimento ad alcuna media. Il **campo di variazione** della distribuzione X è la differenza fra il massimo ed il minimo valore delle modalità di X , cioè :

$$\omega = v = \max(x_i) - \min(x_i)$$

$$S_\mu = \frac{\sum_{i=1}^k |x_i - \mu| \cdot n_i}{\sum_{i=1}^k n_i} = \frac{|x_1 - \mu| \cdot n_1 + |x_2 - \mu| \cdot n_2 + \dots + |x_k - \mu| \cdot n_k}{n_1 + n_2 + \dots + n_k = N} =$$

= **scostamento semplice medio dalla media aritmetica**

$$S_{M_e} = \frac{\sum_{i=1}^k |x_i - M_e| \cdot n_i}{\sum_{i=1}^k n_i} = \frac{|x_1 - M_e| \cdot n_1 + |x_2 - M_e| \cdot n_2 + \dots + |x_k - M_e| \cdot n_k}{n_1 + n_2 + \dots + n_k = N} =$$

= **scostamento semplice medio dalla mediana = scostamento semplice mediano**

$$\sigma^2 = \frac{\sum_{i=1}^k (x_i - \mu)^2 \cdot n_i}{\sum_{i=1}^k n_i} = \frac{(x_1 - \mu)^2 \cdot n_1 + (x_2 - \mu)^2 \cdot n_2 + \dots + (x_k - \mu)^2 \cdot n_k}{n_1 + n_2 + \dots + n_k = N} =$$

= **varianza**

Per calcolare la varianza può essere utile utilizzare la seguente formula :

$$\sigma^2 = \overline{x^2} - \bar{x}^2 = \frac{\sum_{i=1}^n n_i \cdot x_i^2}{N} - \left(\frac{\sum_{i=1}^n n_i \cdot x_i}{N} \right)^2 = \mu_2 - \mu^2$$

La **varianza** è uguale alla **media aritmetica dei quadrati dei dati** meno il **quadrato della media aritmetica dei dati**.

$$\mu_2 = \frac{\sum_{i=1}^n x_i^2 \cdot n_i}{N} = \text{media aritmetica dei quadrati delle modalità} = \text{secondo momento di } X$$

$$N = \sum_{i=1}^k n_i$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^k (x_i - \mu)^2 \cdot n_i}{\sum_{i=1}^k n_i}} = \sqrt{\frac{(x_1 - \mu)^2 \cdot n_1 + (x_2 - \mu)^2 \cdot n_2 + \dots + (x_k - \mu)^2 \cdot n_k}{n_1 + n_2 + \dots + n_k = N}} =$$

= **scarto quadratico medio** = **scostamento quadratico medio**

Differenza semplice media

La **differenza media** è un indice di disuguaglianza fondato sui confronti diretti delle osservazioni .
Dati n valori $x_1, x_2, \dots, x_n$ calcoliamo tutte le differenze fra ogni termine e ciascuno degli altri n termini (compreso se stesso) . Si ottiene un quadro come è indicato nella seguente tabella :

x_1	x_2	x_3		x_n	
$x_1 - x_1$	$x_1 - x_2$	$x_1 - x_3$		$x_1 - x_n$	x_1
$x_2 - x_1$	$x_2 - x_2$	$x_2 - x_3$		$x_2 - x_n$	x_2
.....					.
$x_n - x_1$	$x_n - x_2$	$x_n - x_3$		$x_n - x_n$	x_n

Si tratta di un **quadrato numerico** di n righe ed n colonne , i cui termini posti sulla **diagonale principale** (quelli cioè del tipo $x_i - x_i$) sono tutti nulli .

Riferendoci al caso di n dati semplici e distinti , il numero totale delle differenze è n^2 , di cui n sono nulle e quindi le rimanenti $n^2 - n = n(n - 1)$ sono diverse da zero .

Definiamo la **differenza semplice media** Δ la media aritmetica dei valori assoluti di tutte le $n(n - 1)$ possibili differenze $|x_i - x_j|$ per $i \neq j$:

$$\Delta = \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{n(n - 1)} = \frac{|x_1 - x_2| + \dots + |x_1 - x_n| + |x_2 - x_1| + \dots + |x_n - x_{n-1}|}{n(n - 1)} \quad i \neq j$$

Se includiamo le differenze che si ottengono per $i = j$ otteniamo la **differenza media con ripetizione** Δ_R :

$$\Delta_R = \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{n^2} = \frac{|x_1 - x_2| + \dots + |x_1 - x_n| + |x_2 - x_1| + \dots + |x_n - x_{n-1}|}{n^2}$$

Si dimostra che risulta :

$$\Delta_R = \frac{n-1}{n} \Delta$$

L'indice Δ si differenzia dallo **scostamento semplice medio** S_μ e da analoghi indici di variabilità perché questi ultimi sono misure di dispersione rispetto ad un centro medio (media aritmetica , mediana ,...) che , in qualche modo , **sostituisce** l'intera distribuzione mentre la **differenza semplice media** è una misura della variabilità media **interna** alla distribuzione , cioè fra i singoli valori tra di loro .

Volendo un **indice relativo** variabile tra 0 ed 1 occorre dividere Δ (oppure Δ_R) per il suo valore

massimo : Risulta :

$$\frac{\Delta}{\Delta_{\max}} = \frac{\Delta}{\Delta_{R(\max)}} = \frac{\Delta}{2\mu}$$

Tale rapporto $\frac{\Delta}{2\mu}$ coincide con il **rapporto di concentrazione R** .