Università degli Studi di Napoli «Federico II»
Facoltà di Architettura
U*p*ta
Corso di laurea in Urbanistica e Scienze della Pianificazione Territoriale e Ambientale
AA 2017-2018

Corso integrato di Matematica e statistica

Docente: Livia D'Apuzzo.

Statistica Descrittiva Univariata

7. Variabilità

indici di dispersione da un valore medio e misure di variabilità reciproca per un carattere quantitativo

Variabilità

Il risultato immediato un'indagine statistica è l'individuazione delle diverse modalità con cui il carattere indagato si presenta nella popolazione o campione in considerazione: in genere si presentano più modalità distinte e, quando ciò accade, si dice che c'è **variabilità** nei dati.
Per comprendere meglio le caratteristiche di una distribuzione, accanto a un valore medio che sintetizza i dati, occorre avere una *misura della variabilità*. Nel caso di un carattere quantitativo, la variabilità è leggibile sia attraverso la distanza dei valori del carattere dal valore medio prescelto, sia, attraverso le differenze tra i valori stessi.
Consideriamo i seguenti esempi di distribuzioni dei valori di un carattere quantitativo.
Assumendo come valore medio di riferimento la media aritmetica, analizziamo la variabilità dei valori rilevati rispetto alla media; osserviamo che le distribuzioni hanno tutte la stessa media aritmetica ma gli scarti dalla media sono di diversa entità distribuzione per distribuzione.

	valori x_i in ordine crescente					***scarti " x_i- 6 " dalla media aritmetica 6***				
A)	**6**	**6**	**6**	**6**	**6**	**0**	**0**	**0**	**0**	**0**
B)	**5**	**5**	**6**	**7**	**7**	**-1**	**-1**	**0**	**1**	**1**
C)	**2**	**4**	**6**	**8**	**10**	**-4**	**-2**	**0**	**2**	**4**
D)	**1**	**2**	**6**	**10**	**11**	**-5**	**-4**	**0**	**4**	**5**
E)	**0**	**4**	**5**	**7**	**14**	**-6**	**-2**	**-1**	**1**	**8**

La media aritmetica per tutte le distribuzioni è **6** e per le prime quattro coincide con la mediana; tuttavia il valore **6** non è allo stesso modo rappresentativo dei dati delle varie distribuzioni.

Nella distribuzione **A** i dati sono uguali, non c'è variabilità e la media aritmetica (o mediana) coincidente con ognuno dei dati è perfettamente rappresentativa di essi.

Nella distribuzione **B** la distanza dalla media di ogni dato diverso da 6 è **1.**

Nella distribuzione **C** le distanza dalla media aumentano, diventano 2 e 4, c'è allora più variabilità rispetto a 6: possiamo perciò affermare che la media 6 è più rappresentativa dei dati nella distribuzione B che nella distribuzione C.

Nella distribuzione **D** la distanza dalla media dei dati è ancora più grande e la media perde ulteriormente in rappresentatività.

Nella distribuzione **E** ci sono distanze piccole come 1 (due volte) e grandi come 6 o 8: la situazione della variabilità non è immediatamente leggibile come nelle altre.

Tuttavia possiamo osservare che la somma delle distanze $|x_i-6|$ dei valori dalla media è 18 sia nella distribuzione **D** che nella **E**; possiamo affermare perciò che E e D presentano globalmente la stessa variabilità rispetto al valore medio considerato.

Consideriamo ora la variabilità tra i valori: notiamo che, man mano che passiamo dalla prima all'ultima distribuzione, le distanze tra i valori aumentano: in A la distanza tra due qualsiasi dati è 0, in B la distanza tra due dati distinti è 2 o 1, in C la distanza tra due dati distinti è 2, 4 o 6, in D è 1, 4, 5, 8, 9 o 10, in E 1, 2, 4, 5, 7, 9, 10.

Indici di Variabilità per un carattere quantitativo

Gli indici di variabilità sono indicatori numerici di distribuzioni di frequenze di caratteri quantitativi che danno un'idea della variabilità dei dati. Essi si suddividono in due categorie collegate a due aspetti della variabilità del carattere:

- ***Indici di dispersione***, che indicano quanto mediamente i valori del carattere sono addensati intorno a un indice medio prefissato x_M (media aritmetica, moda, mediana…) o sono da essi lontani;
- ***Indici di disuguaglianza o variabilità reciproca***, che indicano la distanza tra valori significativi della distribuzione o quanto i valori tutti sono tra loro mediamente distanti o vicini.

Gli ***indici di dispersione*** sono detti ***scostamenti medi***: essi sintetizzano, con una operazione di media (aritmetica, quadratica..), le differenze o scarti $x_i - x_M$ dei valori x_i del carattere da un valore medio prefissato x_M.
Gli ***indici di variabilità reciproca*** sono dette ***differenze medie***: essi indicano la variabilità attraverso una differenza tra due valori o indici sintetici o sintetizzano la variabilità operando una media delle differenze in valore assoluto, $|x_i - x_j|$.

Gli indici di variabilità si distinguono ulteriormente in

-indici assoluti: sono espressi nella stessa unità di misura dei valori del carattere quantitativo indagato (non permettono confronti tra distribuzioni relativi a caratteri non omogenei come peso, area, lunghezze);

-indici relativi: prescindono da un'unità di misura e permettono di effettuare confronti tra distribuzioni di caratteri non omogenei come, ad esempio, altezza e peso.

Un indice assoluto di variabilità deve:

- **1**. **annullarsi se e solo se non c'è variabilità** (i valori del carattere sono tutti uguali)**,**

- **2. essere espresso nella stessa unità di misura del carattere,**

- **3. riflettere un cambio di unità di misura " $x \rightarrow \alpha\, x$" , con $\alpha > 0$,** (***omogeneità positiva dell' indice***),

- **4. non cambiare in conseguenza d'una traslazione "$x \rightarrow x + k$" operante sui valori del carattere**

- **5. non cambiare se si moltiplicano le frequenze per uno stesso numero reale.**

1. Indici assoluti di dispersione da un valore medio

Assegnato un valor medio, si possono indicare più indici assoluti di dispersione che verificano le proprietà 1, 2, 3, 4, 5, prima menzionate. La scelta dell'indice più opportuno è determinata anche dalle proprietà che il valor medio considerato ha rispetto agli scarti. Considereremo gli indici di dispersione dalla media aritmetica e dalla mediana.

1.1 - Indici assoluti di dispersione dalla media aritmetica

Utilizziamo i simboli $\bar{x}$ o μ ***per indicare la media aritmetica di k dati*** $x_1, x_2, \ldots x_n$.

Ricordiamo che:

$\bar{x}$ può essere determinata a partire dai valori raccolti unità per unità come media semplice

Distribuzione per unità (o unitaria)

unità statistica	*valore del carattere nell'unità*
1^a	x_1
2^a	x_2
$\vdots$	$\vdots$
n^a	x_n

media aritmetica

$$\rightarrow \quad \bar{x} = \mu = \frac{\sum_i x_i}{n}$$

$\bar{x}$ può essere determinata, come media pesata dei valori distinti che compaiono in una tabella di frequenze, utilizzando le frequenze come pesi.

Distribuzioni di frequenze

x_i *valore*	$f_i = f_a(x_i)$	$f_2' = f_r(x_i)$
x_1	f_1	f_1'
x_2	f_2	f_2'
$\vdots$	$\vdots$	$\vdots$
x_k	f_k	f_k'
	$\sum_i f_i = n$	$\sum_i f_i' = 1$

$x_i \neq x_j$ per $i \neq j$

media aritmetica

$$\rightarrow \quad \bar{x} = \mu = \frac{\sum_i f_i \cdot x_i}{\sum_i f_i} = \sum_i f_i' \cdot x_i$$

a) Misura della dispersione dalla media $\bar{x}$ dei valori di un carattere quantitativo discreto

La differenza $x_i - \bar{x} = x_i - \mu$ è lo scarto del valore x_i dalla media aritmetica $\bar{x}$.
La somma degli scarti dalla media aritmetica $\bar{x}$ è 0, indipendentemente dalla maggiore o minore variabilità dei dati della distribuzione (cfr. fascicolo sulle medie di calcolo).

Se $x_1, x_2, \ldots x_n$ sono i valori raccolti unità per unità (dati grezzi) è allora

$$\sum_1^n (x_i - \bar{x}) = 0;$$

se $x_1, x_2, \ldots x_k$ sono i valori distinti che compaiono in una tabella valori/ frequenza è

$$\sum_1^k f_i (x_i - \bar{x}) = 0$$

Pertanto non si può utilizzare la media aritmetica degli scarti per avere una misura della dispersione.

Per misurare la dispersione da $\bar{x}$ possiamo allora ricorrere agli scarti in valore assoluto, $|x_i - \bar{x}|$, cioè le distanze dei dati dalla media. Un indice di dispersione è allora lo ***scarto semplice medio,*** definito come **media degli scarti in valore assoluto:** esso è rappresentato dalla prima delle seguenti formule se consideriamo i valori raccolti unità per unità, è rappresentato dalla seconda formula, se consideriamo i valori distinti che compaiono in una tabella valori/ frequenza

$$S_{\bar{x}} = \frac{\sum_1^n |x_i - \bar{x}|}{n} \qquad S_{\bar{x}} = \frac{\sum_1^k f_i |x_i - \bar{x}|}{\sum_1^k f_i}$$

scarto semplice medio = *media aritmetica degli scarti in valore assoluto*

$S_{\bar{x}}$ verifica tutti e 5 i requisiti richiesti ad un indice assoluto di variabilità ed è quindi:

$S_{\bar{x}}$ **è un indice assoluto di dispersione da** $\bar{x}$.

Per eliminare l'effetto segno degli scarti che rende nulla la somma degli scarti dalla media aritmetica si può ricorrere anche ai quadrati degli scarti $(x_i - \bar{x})^2$. La somma degli scarti al quadrato si chiama **devianza**

$$\sum_1^n (x_i - \bar{x})^2 \qquad \sum_1^k f_i (x_i - \bar{x})^2 \qquad \textit{devianza}$$

Essa non è un indice assoluto di dispersione perché non verifica i requisiti 2., 3., 5., inoltre aumenta all'aumentare del numero di dati e sono pertanto confrontabili solo devianze che derivano dallo stesso numero di dati.

La media aritmetica dei quadrati degli scarti è detta ***varianza***

$$\sigma^2 = \frac{\sum_1^n (x_i - \bar{x})^2}{n} \qquad \sigma^2 = \frac{\sum_1^k f_i (x_i - \bar{x})^2}{\sum_1^k f_i}$$

Varianza = media aritmetica dei quadrati degli scarti

La varianza non verifica i requisiti 2., 3..

Per avere un indice che abbia la stessa unità di misura dei dati e della loro media aritmetica, si considera la radice quadrata della varianza: si ha allora la ***media quadratica degli scarti*** $x_i - \bar{x}$ che è detta **scarto quadratico medio o devianza standard**

$$\sigma = \sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{n}} \qquad \sigma = \sqrt{\frac{\sum_1^k f_i (x_i - \bar{x})^2}{\sum_1^k f_i}}$$

Scarto quadratico medio o devianza standard *=media* ***quadratica*** *degli scarti*

Lo scarto quadratico medio σ *verifica tutti i cinque requisiti chiesti ed è quindi:*
σ *è* **un indice assoluto di dispersione da** $\bar{x}$

Lo scarto quadratico medio σ ***è l'indice di dispersione caratteristico per la media aritmetica*** in conseguenza del seguente fatto*:* la media aritmetica è quel numero che rende minima la somma dei quadrati degli scarti $\sum_i f_i (x_i - y)^2$ da una quantità y; allora $\sigma_y = \sqrt{\frac{\sum_i f_i (x_i - y)^2}{\sum_i f_i}}$ che indica lo scarto quadratico medio da una quantità y è minimo per $y = \bar{x}$.

NOTA: σ è la media quadratica degli scarti e quindi è compreso tra il più piccolo scarto e il più grande scarto. $S_{\bar{x}}$ è la media aritmetica dei valori assoluti degli scarti e quindi è compreso tra il più piccolo e il più grande dei valori assoluti degli scarti

Confronto tra gli indici

$$S_{\bar{x}} = \frac{\sum_1^k f_i |x_i - \bar{x}|}{\sum_1^k f_i} \quad e \quad \sigma = \sqrt{\frac{\sum_1^k f_i (x_i - \bar{x})^2}{\sum_1^k f_i}}$$

Dalla definizione di media di ordine p

$$M_p = \left(\frac{{a_1}^p + {a_2}^p + \ldots + {a_n}^p}{n} \right)^{\frac{1}{p}} = \left(\frac{1}{n} \sum_j {a_j}^p \right)^{\frac{1}{p}}$$

$S_{\bar{x}}$ è **la media di ordine 1** dei valori assoluti degli scarti e σ è **la media di ordine 2** dei valori assoluti degli scarti; allora dalla implicazione (cfr. fascicolo sulle medie di calcolo)

$$p < q \Rightarrow M_p < M_q$$

segue:

$$S_{\bar{x}} = \frac{\sum_1^k f_i |x_i - \bar{x}|}{\sum_1^k f_i} \leq \sigma = \sqrt{\frac{\sum_1^k f_i (x_i - \bar{x})^2}{\sum_1^k f_i}}$$

l'uguaglianza realizzandosi nel caso non ci sia variabilità e cioè nel caso

$x_i = \bar{x} \quad \forall i$.

Esempio di calcolo di $S_{\bar{x}}$ e σ, utilizzando i dati grezzi

Sono state raccolte le età di 9 individui

Tabella unitaria

i unità	x_i= età di *i*	$\lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$
1	81	8	$8^2=64$
2	69	4	$4^2=16$
3	69	4	$4^2=16$
4	72	1	$1^2=1$
5	81	8	$8^2=64$
6	86	13	$13^2=169$
7	69	4	$4^2=16$
8	73	0	$0^2=0$
9	57	16	$16^2=256$
totali	657	58	602

$\bar{x}$ = (657/9)=73 $S_{\bar{x}}$= 58/9 =6,444 $\sigma=\sqrt{602/9}=\sqrt{66,888}=$ 8,18

Esempi di calcolo di $S_{\bar{x}}$ e σ, utilizzando la tabella di frequenze

Tabella di distribuzione di frequenze

x_i	f_i=fr. ass.	$f_i x_i$	$f_i \lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$	$f_i (x_i - \bar{x})^2$
81	2	2•81=162	2•8=16	$8^2=64$	2•64=128
69	3	3•69=207	3•4=-12	$4^2=16$	3•16=48
72	1	1•72=72	1•1=1	$1^2=1$	1
86	1	1•86=86	1•13=13	$13^2=169$	169
73	1	1•73=73	1•0=0	$0^2=0$	0
57	1	1•57=57	1•16=16	$16^2=256$	256
	$\Sigma_i f_i = 9$	$\Sigma_i f_i x_i = 657$	$\Sigma_i f_i \lvert x_i - \bar{x} \rvert = 58$		$\Sigma_i f_i (x_i - \bar{x})^2 = 602$ devianza

↓

$\bar{x} = \Sigma_i f_i x_i / \Sigma_i f_i$ = (657/9)=73 ↓ ↓

$S_{\bar{x}}$= 58/9 =6,444 $\sigma=\sqrt{602/9}=\sqrt{66,888}=$ 8,18

Distribuzione delle partite di calcio dello scorso campionato per numero di goals segnati

N. gol (x_i)	n_i	$(x_i-\mu)$	$(x_i-\mu)^2$	$(x_i-\mu)^2\, n_i$
0	36	-2.65	7.0225	323.035
1	51	-1.65	2.7225	166.073
2	80	-0.65	0.4225	35.0675
3	52	+0.35	0.1225	6.3700
4	36	+1.35	1.8225	65.6100
5	22	+2.35	5.5225	121.495
6	18	+3.35	11.2225	202.005
7	7	+4.35	18.9225	264.915
8	4	+5.35	28.6225	114.490
Totale	306			1299.06

La media aritmetica è

μ =(0•36+51+2•80+3•52+4•36+
5•22+6•18+7•7+8•4)/306 = 2.65

la varianza

σ^2=1299.06/306 =4.245

la devianza standard è

σ = 2.060339777804

Confronto tra variabilità rispetto alla media aritmetica di due distribuzioni sullo stesso insieme di valori

$$A \qquad \begin{matrix} x_i \\ f_i \end{matrix}\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 40 & 40 & 40 & 40 & 40 \end{pmatrix} \qquad \bar{x} = \frac{40(1+2+3+4+5)}{200} = 3$$

$$B \qquad \begin{matrix} x_i \\ f_i \end{matrix}\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 10 & 16 & 148 & 16 & 10 \end{pmatrix} \qquad \bar{x} = \frac{10+32+444+64+50}{200} = 3$$

Quale distribuzione presenta più dispersione dalla media?

40 scarti uguali a -2
40 scarti uguali a –1
40 scarti uguali a 0
40 scarti uguali a 1
40 scarti uguali a 2

10 scarti uguali a -2
16 scarti uguali a -1
148 scarti uguali a 0
16 scarti uguali a -1
10 scarti uguali a 2

$$\sum_i f_i\left(x_i - \bar{x}\right)^2 = 400 \quad \leftarrow \text{devianza} \rightarrow \quad \sum_i f_i\left(x_i - \bar{x}\right)^2 = 112$$

$$\sigma^2 = \frac{\sum_i f_i\left(x_i - \bar{x}\right)^2}{\sum_i f_i} = \frac{400}{200} = 2 \quad \leftarrow \text{varianza} \rightarrow \quad \sigma^2 = \frac{\sum_i f_i\left(x_i - \bar{x}\right)^2}{\sum_i f_i} = \frac{112}{200} = 0{,}56$$

$$\sigma = \sqrt{\frac{\sum_i f_i\left(x_i - \bar{x}\right)^2}{\sum_i f_i}} = \sqrt{2} \leftarrow \begin{matrix} \textit{scarto} \\ \textit{quadratico} \\ \textit{medio} \end{matrix} \rightarrow \sigma = \sqrt{\frac{\sum_i f_i\left(x_i - \bar{x}\right)^2}{\sum_i f_i}} = \sqrt{0{,}56}$$

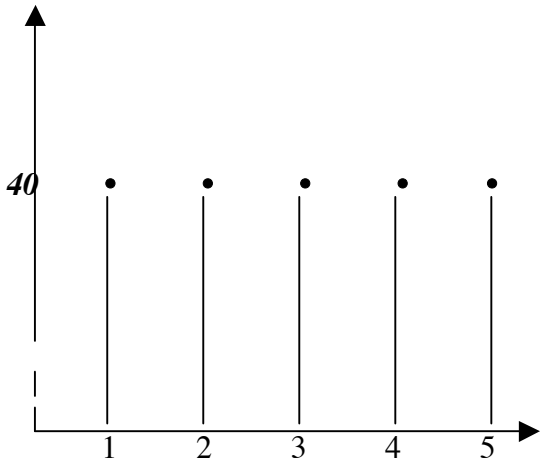
$$\frac{\sum_i f_i\left|x_i - \bar{x}\right|}{\sum_i f_i} = \frac{240}{200} = 1{,}2 \leftarrow \begin{matrix} \textit{scarto} \\ \textit{semplice} \\ \textit{medio} \end{matrix} \rightarrow \frac{\sum_i f_i\left|x_i - \bar{x}\right|}{\sum_i f_i} = \frac{72}{200} = 0{,}36$$

Leggiamo la diversa dispersione dei dati dal grafico

A

$$\begin{matrix} x_i \\ f_i \end{matrix}\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 40 & 40 & 40 & 40 & 40 \end{pmatrix}$$

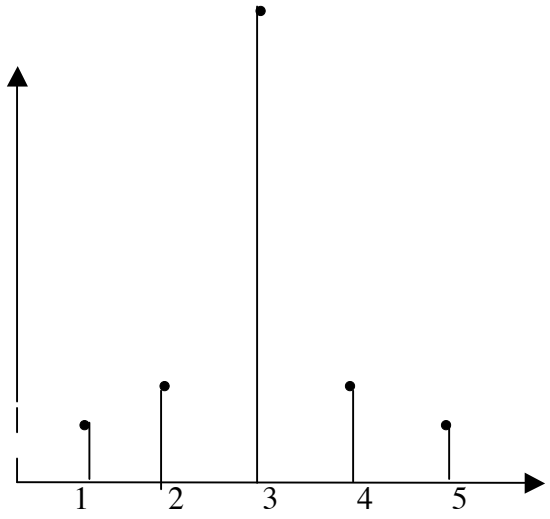
$$\bar{x} = \frac{40(1+2+3+4+5)}{200} = 3$$



B

$$\begin{matrix} x_i \\ f_i \end{matrix}\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 10 & 16 & 148 & 16 & 10 \end{pmatrix}$$

$$\bar{x} = \frac{10+32+444+64+50}{200} = 3$$



b) Stima della dispersione dalla media aritmetica nel caso di una distribuzione con valori del carattere raggruppati in classi

Se i valori del carattere sono raggruppati in classi, per il calcolo dell'indice di dispersione si considerano **gli scarti dei valori centrali della classe dal valore medio** (ottenuto come media dei valori centrali)

Esempio

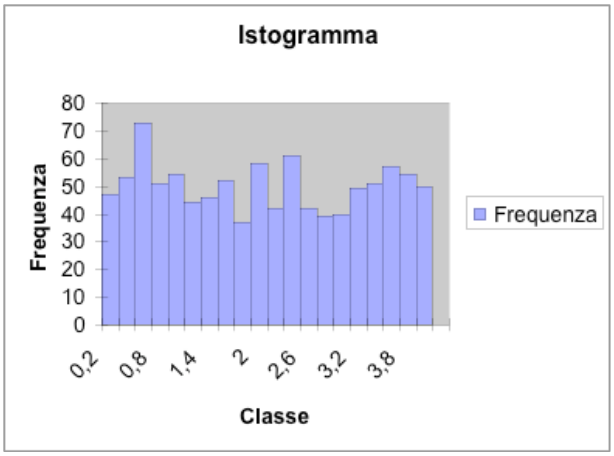
Classi età in anni	c_i **valore centrale**	$f_i=f_a(x_i)$	$f_i\, x_i$	$\lvert c_i$-media$\rvert$	$f_i \lvert c_i$-media$\rvert$	$f_i\,(c_i$-media$)^2$
0--|10	5	2	2•5	13,75	2•13,75=27,5	2•189,0625=378,125
10--|20	15	3	3•15	3,75	3•3,75=11,25	3•14,0625=42,1875
20--|30	25	1	1•25	6,25	1•6,25=6,25	39,0625
30--|40	35	2	2•35	16,25	2•16,25=32,5	2•264,0625=528,125
		8	150		77.5	987,5

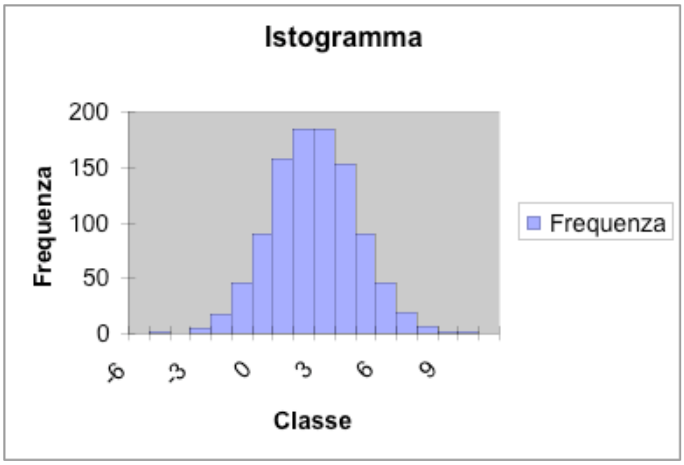
Media aritmetica = 150/8=18,75

$S_{\bar{x}}$**=77, 5 /8 = 9,6875**

σ=√(987,5)/8 =√84,33=11,10243..

Consideriamo un carattere con valori suddivisi in classi e due diverse distribuzioni rappresentate graficamente dai seguenti istogrammi: quale rappresenta più variabilità?





Media=2
Varianza=4

Media=3
Varianza=1.33

Calcolo alternativo della varianza e dello scarto quadratico medio

Posto: $M_2 = \sqrt{\frac{\sum_i f_i x_i^2}{\sum_i f_i}} = \left(\frac{\sum_i f_i x_i^2}{\sum_i f_i}\right)^{1/2}$ ***(media quadratica dei valori)***

e

$M_1 = \bar{x} = \frac{\sum_i f_i x_i}{\sum_i f_i}$ (***media aritmetica dei valori***)

Lo scarto quadratico medio è data da

$$\sigma = \sqrt{M_2^2 - M_1^2} = \sqrt{\frac{\sum_i f_i x_i^2}{\sum_i f_i} - \left(\frac{\sum_i f_i x_i}{\sum_i f_i}\right)^2}$$ ***scarto quadratico medio***

Dimostrazione. Calcoliamo la varianza

$$\sigma^2 = \frac{\sum_i f_i \left(x_i - \bar{x}\right)^2}{\sum_i f_i} = \frac{\sum_i f_i \left(x_i^2 - 2x_i \bar{x} + \bar{x}^2\right)}{\sum_i f_i} =$$

$$= \frac{\sum_i f_i x_i^2 - 2\bar{x}\sum_i f_i x_i + \bar{x}^2 \sum_i f_i}{\sum_i f_i} = \frac{\sum_i f_i x_i^2}{\sum_i f_i} - 2\bar{x}\bar{x} + \bar{x}^2 = \frac{\sum_i f_i x_i^2}{\sum_i f_i} - \bar{x}^2$$

Risulta allora: $\sigma^2 = \frac{\sum_i f_i x_i^2}{\sum_i f_i} - \left(\frac{\sum_i f_i x_i}{\sum_i f_i}\right)^2 = M_2^2 - M_1^2$ ***varianza***

e quindi $\sigma = \sqrt{M_2^2 - M_1^2}$ ***scarto quadratico medio***

Ricalcolare lo scarto quadratico medio nella seguente distribuzione utilizzando la formula alternativa per il calcolo della varianza

x_i	f_i=fr. ass.	$f_i\,x_i$	x_i^2	$f_i\,x_i^2$
81	**2**			
69	**3**			
72	**1**			
86	**1**			
73	**1**			
57	**1**			
	$\Sigma f_i = 9$			

Rispondere alla seguente domanda il signore che pesa 57 Kg di quanti scarti è al disotto della media?

1.1.1 Teorema di Chebyshev

Il teorema di Chebyshev, che di seguito riportiamo, permette di avere alcune informazioni su di una distribuzione di cui si conoscono solo la media aritmetica $\bar{x}$ e lo scarto quadratico medio σ_x

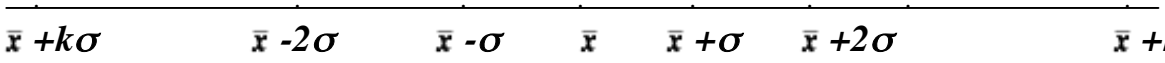
Teorema di Chebyshev

Siano $\bar{x}$ e σ la media e lo scarto quadratico medio relativi alla distribuzione di un carattere quantitativo. Allora, per k > 0, la frequenza relativa dei valori i cui scarti dalla media $\bar{x}$ verificano la disuguaglianza

$$\left|x_i - \bar{x}\right| \geq k\sigma$$

è minore o uguale a $\frac{1}{k^2}$.

In formule $$f_r\left(\left|x_i - \bar{x}\right| \geq k\sigma\right) \leq \frac{1}{k^2} \qquad (Ch)$$



La disuguaglianza (Ch) ci dice, in altre parole che la frequenza relativa dei casi con valori che non cadono internamente all'intervallo $I_k =]\bar{x} - k\sigma, \bar{x} + k\sigma[$ non supera $\frac{1}{k^2}$, ma non ci dice quanto effettivamente vale; ad esempio.

$f_r(|x_i - \bar{x}| \geq \sigma) = f_r(x_i \leq \bar{x} - \sigma) + f_r(x_i \geq \bar{x} + \sigma) \leq 1$, (*ovvio perché le frequenze relative sono tutte ≤ 1*)

$f_r(|x_i - \bar{x}| \geq 2\sigma) = f_r(x_i \leq \bar{x} - 2\sigma) + f_r(x_i \geq \bar{x} + 2\sigma) \leq \frac{1}{4}$

Esercizio.

Supponiamo che le case nel centro storico hanno un prezzo medio di 3000 euro a m^2 e che lo scarto quadratico medio sia di 500 euro. Quante sono le case con un prezzo compreso tra 2000 euro e 4000 euro?

Dati del problema: $\bar{x} = 3000$ ***prezzo medio*** $\sigma = 500$

I prezzi 2000 euro e 4000 euro hanno uno scarto dal prezzo in valore assoluto pari a $2\sigma = 1000$. Per la formula *(Ch)* la frequenza relativa delle case con scarto in valore assoluto maggiore o uguale $2\sigma = 1000$, verifica la disuguaglianza

$$f_r\left(\left|x_i - 3000\right| \geq 2\sigma\right) \leq \frac{1}{4}$$

cioè la percentuale delle case con prezzo <u>non compreso tra</u> 2000 euro e 4000 euro non supera il 25% **quindi ……**

Nota: ***l'informazione che fornisce*** **il teorema di Chebyshev** ***è corretta ma non dà una informazione precisa sulla quantità dei valori che cadono al di fuori dell'intervallo*** $I_k =]\bar{x} - k\sigma, \bar{x} + k\sigma[$**. Per** ***esempio per k= 2, il teorema afferma che i valori che cadono al di fuori di*** I_k ***non*** **superano 1/4** ***di tutti i valori, ma non ci dice se in realtà sono effettivamente*** **1/4 o 1/10 o 1/1000 etc….**

1.2- Indice assoluto di dispersione dalla mediana

a') Misura della dispersione dalla mediana *Me* dei valori di un carattere discreto

La dispersione dalla media *Me* di una distribuzione viene misurata attraverso gli scarti x_i - *Me*. Poiché tavolta la mediana coincide con la media aritmetica non conviene utilizzare un indice di dispersione basato sulla la somma degli scarti dalla mediana, perché tale somma potrebbe essere 0 anche in presenza di variabilità.

L'indice di dispersione dalla mediana più usato è lo **scarto semplice medio dalla mediana**

$$S_{Me} = \frac{\sum_i f_i \left| x_i - Me \right|}{\sum_i f_i}$$

scarto semplice medio dalla mediana

che è considerato indice di dispersione caratteristico della mediana in quanto la mediana è quel numero che rende minima la somma dei valori assoluti degli scarti $\sum_i f_i \left| x_i - y \right|$ da un valore numerico y e quindi se $S_y = \frac{\sum_i f_i \left| x_i - y \right|}{\sum_i f_i}$ indica lo scarto semplice medio da y S_y è minimo per y= Me.

Allora

$$S_{Me} = \frac{\sum_i f_i \left| x_i - Me \right|}{\sum_i f_i} \le S_{\bar{x}} = \frac{\sum_i f_i \left| x_i - \bar{x} \right|}{\sum_i f_i} \le \sigma = \sqrt{\frac{\sum_i f_i \left(x_i - \bar{x} \right)^2}{\sum_i f_i}}$$

Esercizio

Sia data la seguente tabella di frequenze per il carattere età. Calcolare la mediana e lo scostamento semplice dalla mediana e confrontare $S_{Me} = \frac{\sum_i f_i \left| x_i - Me \right|}{\sum_i f_i}$ **con gli indici** $S_{\bar{x}} = \frac{\sum_i f_i \left| x_i - \bar{x} \right|}{\sum_i f_i}$ **e** $\sigma = \sqrt{\frac{\sum_i f_i \left(x_i - \bar{x} \right)^2}{\sum_i f_i}}$ **precedentemente calcolati**

x_i	f_i=fr. ass.
81	2
69	3
72	1
86	1
73	1
57	1
	$\Sigma_i f_i = 9$

Ricordiamo che per calcolare la mediana i valori del carattere vanno ordinati

x_i	f_i=fr. ass.
57	1
69	3
72	1
73	1
81	2
86	1
	$\Sigma_i f_i = 9$

b') Stima della dispersione dalla mediana nel caso di una distribuzione con valori del carattere raggruppati in classi.

Se i valori del carattere sono raggruppati in classi, per il calcolo dell'indice di dispersione dalla mediana si considerano **gli scarti dei valori centrali della classe dalla mediana**

Esempio

Classi età in anni	c_i **valore centrale**	f_i**=fr. ass.**	$f_{a\ cum}$	$\lvert c_i$**-mediana**$\rvert$	$f_i \lvert c_i$**-mediana**$\rvert$
0--|10	**5**	**2**	**2**	**|5-….|**	**2•..**
10--|20	**15**	**3**	**5**		**3•….**
20--|30	**25**	**2**	**7**		**2•….**
30--|40	**35**	**2**	**8**		**2•…..**
		9			

Posto centrale

Classe mediana =

Mediana =

2 Un indice relativo di dispersione da un valore medio: *il coefficiente di variazione*

Gli indici assoluti di dispersione sono influenzati dall'intensità dei valori del carattere: se i valori del carattere indicano i chilometri aerei percorsi in un anno dagli individui di un collettivo il corrispondente indice di dispersione sarà espresso in chilometri; mentre se i valori del carattere esprimono il diametro in millimetri di pietre preziose, il corrispondente indice di dispersione è espresso in millimetri: ciò rende difficile confrontare le variabilità delle due distribuzioni di valori. Analogamente è difficile confrontare le variabilità di distribuzioni di caratteri di tipologie (e unità di misura) diverse come età e peso. Si ovvia a questo inconveniente dividendo l'indice assoluto di dispersione per la media rispetto alla quale si studia la dispersione, ottenendo così un indice relativo, che è un numero puro che prescinde da unità di misura. Tra gli indici relativi di dispersione consideriamo il

Coefficiente di Variazione

$$C_V = \frac{\sigma}{\bar{x}} = \frac{\sqrt{\frac{\sum_i f_i (x_i - \bar{x})^2}{\sum_i f_i}}}{\frac{\sum_i f_i x_i}{\sum_i f_i}} = \frac{\sqrt{\sum_i f_i}\sqrt{\sum_i f_i (x_i - \bar{x})^2}}{\sum_i f_i x_i} = \frac{\sqrt{n}\sqrt{\sum_i f_i (x_i - \bar{x})^2}}{\sum_i f_i x_i}$$

Il coefficiente di variazione rapporta lo scarto quadratico medio alla media della distribuzione. Esso permette di confrontare le variabilità di due distribuzioni di valori di ordine di grandezza molto diversi o distribuzioni di caratteri non omogenei (peso ed altezza, per esempio) e che hanno quindi unita' di misure diverse e non trasformabili l'una nell'altra con un cambio di scala. In alcuni casi il coefficiente di variazione è assunto uguale a $\frac{\sigma}{\bar{x}} 100$

Esempio:

distribuzione dei pesi in un gruppo di adulti

$\bar{x}$ **=75 Kg.** **σ =4 Kg** **C_V =(4/75)=0,053**

distribuzione dei pesi in un gruppo di neonati

$\bar{x}$ **=3,2 Kg.=*3200* g.** **σ =0,6 Kg=*600g*.** **C_V =(0,6/3,2)=0,1875**

Quale distribuzione presenta maggiore variabilità?

Confrontando gli scarti quadratici medi sembrerebbe che la prima distribuzione presenti maggiore variabilità, ma è i realtà la seconda distribuzione ad essere più variabili come si deduce dal confronto degli indici di dispersione.

Esempio: ***stabilire se è più variabile la distribuzione del peso dei bambini alla nascita o la distribuzione del peso delle relative madri al parto. Indichiamo con x il peso dei bambini e con y il peso delle madri.***

x	Y	x-media	y-media	$(x-\mu_x)^2$	$(y-\mu_y)^2$
2.80	62.00	-0.40	-3.60	0.16	12.96
3.00	65.00	-0.20	-0.60	0.04	0.36
3.50	63.00	0.30	-2.60	0.09	6.76
4.20	70.00	1.00	4.40	1.00	19.36
2.50	68.00	-0.70	2.40	0.49	5.76
16.00	328.00	0.00	0.00	1.78	45.20

$\mu_x=3.2$ $\sigma_x=0.5967$ $CVx=\sigma_x/\mu_x=0.1864$

$\mu_y=65.6$ $\sigma y=3.0067$ $CVy=\sigma_y/\mu_y=0.0458$

Esercizio

In una gara di atletica leggera sono stati rilevati i seguenti 5 migliori risultati di salto in alto (in metri) e di corsa sui 100 m (in secondi):

- **Salto in alto**

1,85 1,92 1,95 1,94 1,94

- **100 m**

11,7 11,3 11,4 11,2 11,6.

Indicare quale delle due serie di risultati presenta maggiore variabilità

Qui di seguito sono riportati altri due esercizi sul confronto tra le variabilità rispetto alla media di due distribuzioni: nei due esercizi dovete scegliere gli indici di dispersione più adatti ad operare il confronto.

1. Quale delle seguenti distribuzioni, relative ad una stessa variabile X, mostra minore variabilità? Dedurre prima la risposta dalla rappresentazione grafica, quindi rispondere calcolando l'opportuno indice di dispersione.

x_i	n_i
0	**5**
1	**5**
2	**5**
3	**5**
4	**5**

x_i	n_i
0	**1**
1	**4**
2	**15**
3	**4**
4	**1**

2. Sono qui di seguito riportate le distribuzioni di frequenza secondo il peso di un gruppo di 100 neonati e dell'insieme delle rispettive madri:

Peso	n_i
1,5 – 2, 0	**5**
2,0 – 2,5	**12**
2,5 – 3,0	**25**
3,0 – 3,5	**35**
3,5 – 4,0	**18**
4,0 – 4,5	**5**

Peso	n_i
45 – 50	**4**
50 – 55	**12**
55 – 60	**22**
60 – 65	**40**
65 – 70	**19**
70 – 75	**3**

Dire quale dei due collettivi presenta maggiore variabilità rispetto alla media, **utilizzando un opportuno indicatore di variabilità.**

3 Indici di variabilità reciproca

I seguenti indici rappresentano la variabilità reciproca tra valori del carattere e non la variabilità rispetto ad una media.

- ***Campo di variazione*** $\omega = \max_i x_i - \min_i x_i$

ω prende in considerazione solo la dispersione tra i valori estremi della distribuzione e risente di valori anomali

Esempio: in un gruppo di 7 individui sono state rilevate le seguenti età in anni

35 24 28 46 22 25 27

per ottenere il campo di variazione ordiniamoli in senso crescente

22 24 25 27 28 35 46

il campo di variazione è dato dalla differenza tra i valori estremi dell'allineamento

$$\omega = 46 - 22 = 24$$

e rappresenta l'ampiezza del più piccolo intervallo contenente i valori considerati

Se il valore 46 viene sostituito dal valore 104 (anomalo rispetto ai rimanenti)

22 24 25 27 28 35 104

il nuovo campo di variazione è $\omega = 104 - 22 = 82$

- ***Differenza interquartilica* o *distanza interquartile***

Un indice di dispersione di uso comune è l'**intervallo interquartile o distanza interquartile**, dato dalla **differenza tra 3° e 1° quartile** (cioè tra 75° e 25° centile):

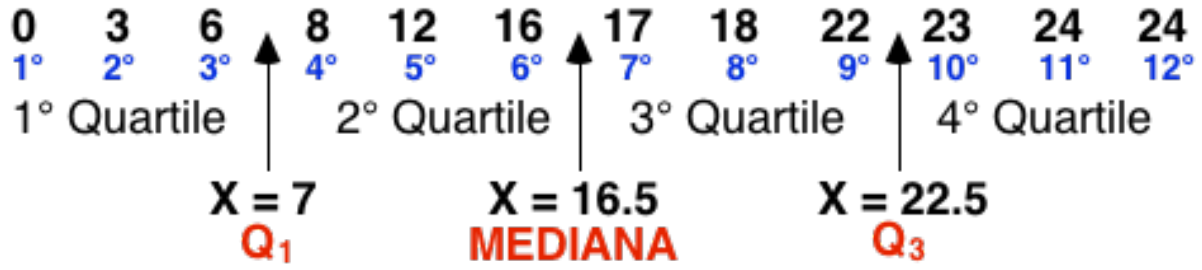
$$W = Q_3 - Q_1$$

Poiché prima del primo quartile c'è il 25% della distribuzione e dopo il terzo quartile c'è un altro 25% la **distanza interquartile W è l'ampiezza dell'intervallo che contiene la metà dei valori inclusi nel campione o polazione.**

Esempio

VARIABILE: **Numero di errori nella lettura di un brano (N=12)**

Distribuzione ordinata dei punteggi (e posizioni):



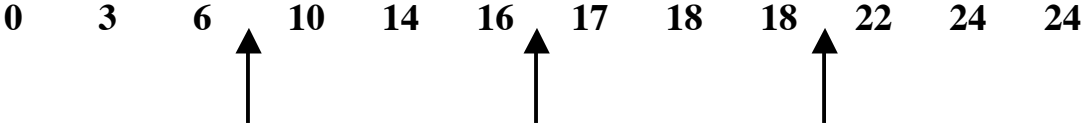
Posti centrali n/2 =12/2 =6 e (n/2)+1 =7 Mediana = (16+17)/2 = 16.5

n/4 =12/4 =3 e (n/4)+1 =4 ↓ **Q_1= (6+8)/2 =7**

3n/4 =36/4 =9 e (3n/4)+1 =10 ↓ **Q_3= (22+23)/2 =22,5**

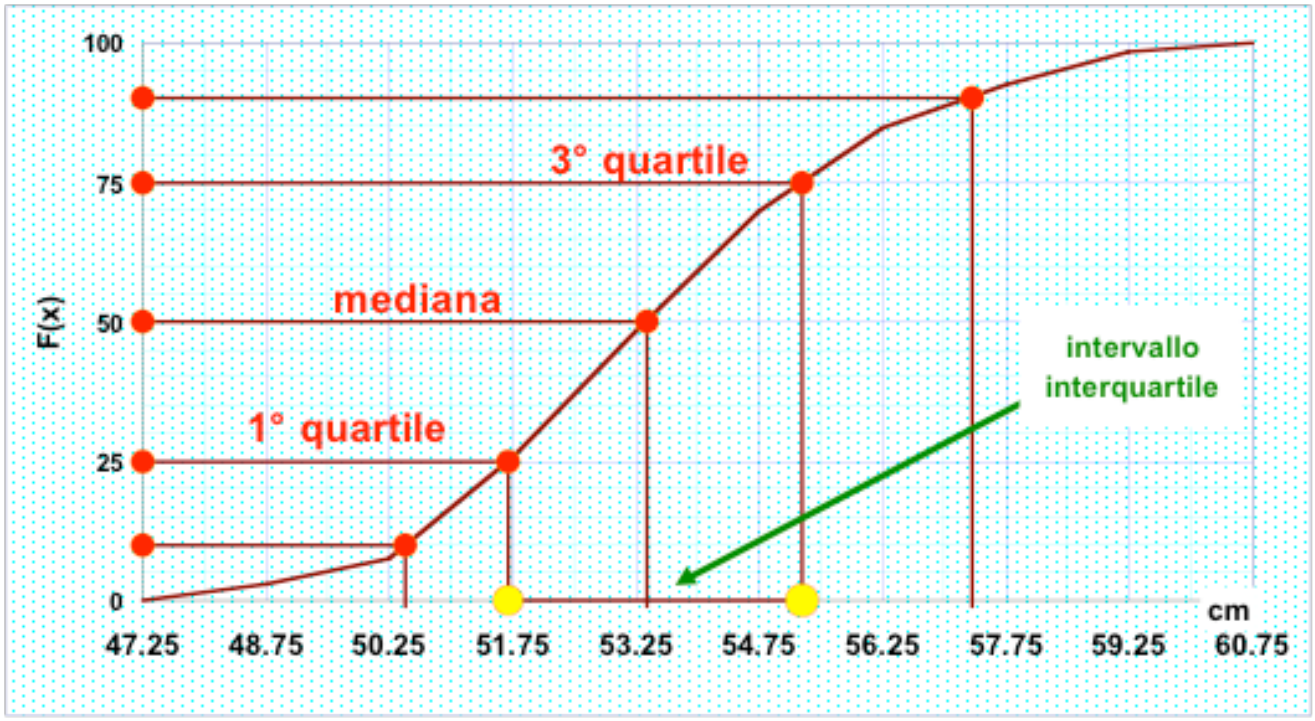
Differenza interquartilica = Q_3 - Q_1 = 22.5 - 7 = 15.5

La differenza interquartilica costituisce anche un indice di dispersione dalla mediana: più piccola è più i dati compresi tra il primo e terzo quartile sono vicini alla mediana



Differenza interquartilica = Q_3 - Q_1 = 20 - 8 = 12

La distanza interquartile dedotta dall'ogiva delle frequenze percentuali cumulate



- ***Differenza media assoluta***

Un indice di variabilità reciproca è il seguente

$$\frac{\sum_i \sum_j |x_i - x_j|}{n(n-1)} \quad i \neq j$$

differenza media assoluta
o
differenza media di Gini

Esso dà una misura di quanto mediamente distano tra di loro i valori.

Sul calcolo della differenza media semplice

Si consideri la seguente matrice delle differenze in valore assoluto tra modalità:

$$\begin{array}{cccccccc}
|x_1-x_1| & |x_1-x_2| & \cdots & |x_1-x_j| & \cdots & |x_1-x_n| & \rightarrow & s_1=\sum_j|x_1-x_j| \rightarrow \frac{s_1}{n-1}\ o\ \frac{s_1}{n} \\
 & \vdots & & & & & & \\
|x_2-x_1| & |x_2-x_2| & \cdots & |x_2-x_j| & \cdots & |x_2-x_n| & \rightarrow & s_2=\sum_j|x_2-x_j| \rightarrow \frac{s_2}{n-1}\ o\ \frac{s_2}{n} \\
\vdots & \vdots & & & & & & \\
|x_i-x_1| & |x_i-x_2| & \cdots & |x_i-x_j| & \cdots & |x_i-x_n| & \rightarrow & s_i=\sum_j|x_i-x_j| \rightarrow \frac{s_i}{n-1}\ o\ \frac{s_i}{n} \\
\vdots & \vdots & & & & & & \\
|x_{n-1}-x_1| & |x_{n-1}-x_2| & \cdots & |x_{n-1}-x_j| & \cdots & |x_{n-1}-x_n| & \rightarrow & s_{n-1}=\sum_j|x_{n-1}-x_j| \rightarrow \frac{s_n}{n-1}\ o\ \frac{s_n}{n} \\
|x_n-x_1| & |x_n-x_2| & \cdots & |x_n-x_j| & \cdots & |x_n-x_n| & \rightarrow & s_n=\sum_j|x_n-x_j| \rightarrow \frac{s_n}{n-1}\ o\ \frac{s_n}{n}
\end{array}$$

Alla fine di ogni riga sono riportate la somma degli elementi di riga e due medie. Per ciò che riguarda la somma $s_i=\sum_j|x_i-x_j|$, notiamo che l'indice ***j*** varia da 1 a n, ma possiamo limitarci a sceglierlo diverso da ***i*** in quanto $|x_i-x_i| = 0$ non dà alcun contributo alla somma: quindi $s_i=\sum_j|x_i-x_j|=\sum_{j\neq i}|x_i-x_j|$.

Delle due medie $\frac{s_i}{n-1}\ o\ \frac{s_i}{n}$, la prima è la media calcolata su n-1 elementi, in quanto viene escluso l'elemento delle diagonale principale $|x_i-x_i|$ che è nullo, l'altra è la media di tutti gli n elementi di riga. La sintesi delle differenze in valore assoluto tra i valori del carattere, non considerando i confronti tra ciascuna modalità e se stessa, si ottiene operando la media la media aritmetica delle n(n-1) differenze $|x_i-x_j|$, $i\neq j$. Per la proprietà associativa della media aritmetica, la media su tutti gli elementi della matrice esclusi quelli della diagonale principale elementi) è la media aritmetica delle n medie "***s_i/(n-1)***"

$$\Delta=\frac{\sum_{i=1}^{n}\frac{s_i}{n-1}}{n}=\frac{\sum_{i=1}^{n}s_i}{n(n-1)}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i-x_j|}{n(n-1)}$$ **differenza semplice media.**

Siccome la matrice è simmetrica, le differenze in valore assoluto al di sopra della diagonale principale sono uguali alle corrispondenti differenze al di sotto della diagonale; per questo la uguaglianza di cui sopra può essere scritta nel seguente modo:

$$\Delta=2\frac{\sum_{i=1}^{n}\sum_{j=1}^{i}|x_i-x_j|}{n(n-1)}$$

Tale indice è una misura della variabilità media interna alla distribuzione, cioè fra i singoli valori tra di loro. Quando il confronto tra tutte le modalità è fatto tenendo conto anche della differenza di una modalità con se stessa allora si parla di **differenza semplice media con ripetizione** definita da:

$$\Delta = \frac{\sum_{i=1}^{n} \frac{s_i}{n}}{n} = \frac{\sum_{i=1}^{n} s_i}{n^2} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} |x_i - x_j|}{n^2}$$ **differenza semplice media con ripetizione**

I due indici sono legati dalla relazione: $\Delta = \Delta_R \frac{n}{n-1}$