

Dal coefficiente di Correlazione lineare tra due grandezze x, y alla covarianza.

La seguente nota è dedicata ad una riorganizzazione del capitolo 9 del Taylor. Si fa notare che viene invertito l'ordine, ovvero prima si deriva il coefficiente di correlazione (r) tra una serie di misure (x_i, y_i) di due grandezze, poi, una volta trovato tale coefficiente, riconosciuto come un buon stimatore della linearità (r), si osserva che tale estimatore è utile per considerazioni sulla covarianza di due grandezze x e y nel caso della propagazione dell'errore per una grandezza g funzione di x e y . In questo senso l'ordine dei paragrafi risulta invertito, prima il 9.3 per il quale si deriverà il coefficiente di correlazione r , poi i paragrafi 9.4 e 9.5 (che risultano sufficientemente chiari), ed infine la giustificazione della propagazione del paragrafo 9.2.

Il paragrafo 9.1 è solo una rilettura di quanto già spiegato.

Ovvero si deriva il coefficiente di correlazione lineare tra due variabili x, y e poi lo si utilizza per giustificare la propagazione degli errori in quadratura o lineare.

Coefficiente di correlazione lineare.

Per introdurre il coefficiente di correlazione partiamo da un esempio semplice: il problema 8.17, che riconsidera i dati del problema 8.1

Caso a: miglior adattamento di una retta $y=A+Bx$ a coppie di dati (x_i, y_i) .

Si trovi il miglior adattamento di una retta mediante il metodo dei minimi quadrati dei dati del problema 8.1, ovvero per (x, y) : (1, 6), (3, 5) e (5, 1) per avere anche un controllo dimensionale poniamo che x sia data in cm e y in s.

Riportiamo in ordine i dati

Variabile indipendente $x[s]$	Variabile dipendente $y[cm]$
1	6
3	5
5	1

Il metodo dei minimi quadrati fornisce, per la retta $y=A+Bx$, che meglio si adatta ai punti, i seguenti coefficienti: $A=7.75 \pm 1.48$ cm e $B=-1.25 \pm 0.43$ cm/s.

Osserviamo che tale retta si deduce dall'aver minimizzato gli scarti al quadrato delle y ovvero trovando il minimo del

$$\chi^2 = \sum_{i=1}^N \frac{(y_i - A - Bx_i)^2}{\sigma_y^2}$$

(Equ. 1)

il minimo degli scarti al quadrato dei dati sperimentali y_i dalla retta teorica $y=A+Bx$.

Miglior adattamento di una retta $y'=A'+B'x'$ a coppie di dati (x'_i, y'_i) .

Ci si pone il problema se invertendo le variabili, ovvero utilizzando y come variabile indipendente e invece x come variabile dipendente, si ottiene lo stesso risultato.

Si osservi che in questo stiamo cercando una relazione del tipo $y'=A'+B'x'$ dei dati precedenti – per cui x' è la precedente y , invece y' risulta la x . Per maggiore evidenza si è utilizzato lo sfondo giallo per associare anche visivamente lo scambio delle variabili.

Variabile indipendente $x'[cm]$	Variabile dipendente $y'[s]$
6	1
5	3
1	5

La retta che meglio si adatta a (x'_i, y'_i) , con il metodo dei minimi quadrati, è data adesso dal minimizzare le differenze al quadrato delle y'_i rispetto a $y' = A' + B'x'$ (dove le y'_i sono le precedenti x_i)

$$\chi'^2 = \sum_{i=1}^N \frac{(y'_i - A' - B'x'_i)^2}{\sigma_{y'}^2}$$

Tanto che per maggiore chiarezza si è utilizzato l'apostrofo anche per il chi-quadro.

Si ottiene $A' = 5.86 \pm 1.12$ s e $B' = -0.71 \pm 0.25$ s/cm.

Da questa equazione, ottenuta applicando i minimi quadrati a $y' = A' + B'x'$ si ha:

$$y' = A' + B' \cdot x' \equiv x = A' + B' \cdot y$$

da cui si ottiene dal relazione di y in funzione di x rispetto ai coefficiente A' e B' :

$$y = -\frac{A'}{B'} + \frac{1}{B'} \cdot x$$

Verifichiamo se sono uguali a quelli ottenuti dall'applicazione del metodo dei minimi quadrati a $y = A + Bx$. Per esplicitare meglio etichettiamo $A^* = -A'/B'$ e $B^* = 1/B'$.

Otteniamo quindi per $A^* = 8.2$ cm $B^* = -1.4$ cm/s, leggermente diversi da quanto ottenuto per i coefficienti A e B .

Tutto questo si conclude nell'osservare che nel riporare sullo stesso grafico le due rette si osserva:

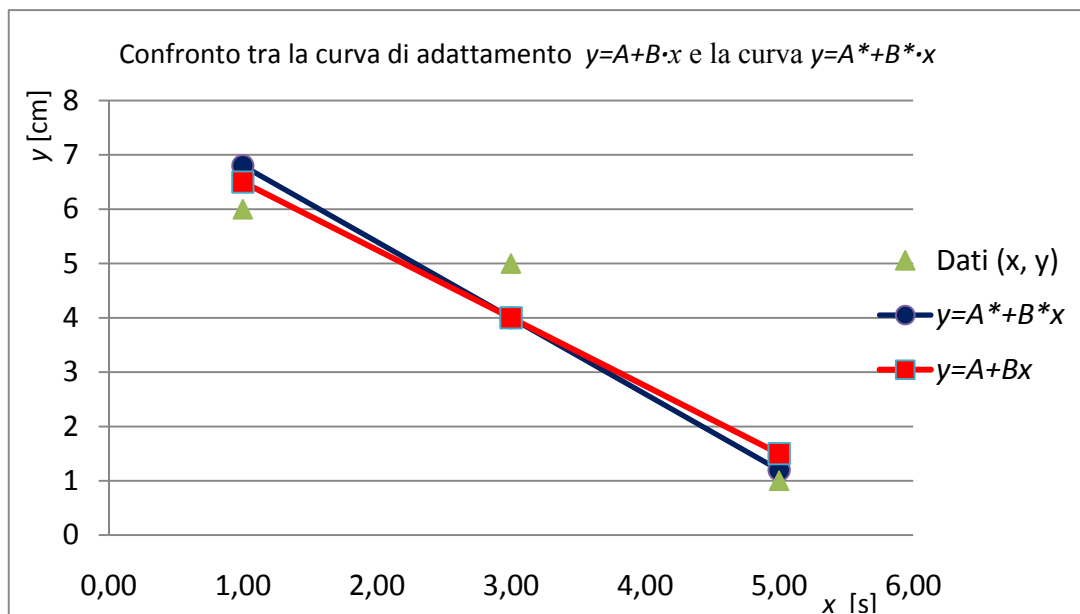


Figura 1. Nel grafico sono riportati i dati (x, y) ($\blacktriangle$), la retta $y = A + Bx$ ($\blacksquare$) ottenuta applicando il metodo dei minimi quadrati alle coppie (x, y) , la retta $y = A^* + B^*x$ ($\bullet$) dedotta applicando il metodo dei minimi quadrati alle coppie (x', y') e ricavando A^* e B^* da A' e B' .

Si può ricavare anche l'errore su A^* e B^* , propagando opportunamente gli errori su A' e B' . Si ottiene pertanto $A^* = 8.2 \pm 3.6$, $B^* = -1.4 \pm 0.5$. Si possono confrontare con A e B e fare considerazione sulla consistenza dei risultati sulla base della discrepanza e degli errori.

Caso b: punti (x, y) disposti su una retta

Quale sarebbe l'andamento del miglior addattamento ad una retta per i due modi diversi, se i tre punti si trovassero precisamente su una retta? Per convenienza prendiamo come y_i quanto ottenuto dalla relazione $y=A+Bx$ per i dati di partenza quindi

Variabile indipendente $x[s]$	Variabile dipendente $y[cm]$
1	6.5
3	4
5	1.5

Se si applica il metodo dei minimi quadrati a tali coppie otteniamo, ovviamente $A= 7.75$ cm, $B= -1.25$ cm/s (gli errori sono tutti nulli).

Cambiamo variabili ed applichiamo il metodo dei minimi quadrati a $y'=A'+B'x'$. Si ottiene che $A'=6,2$ s e $B'=-0.8$ s/cm (incertezze nulle). Questa volta se mi riporto alle variabili iniziale ottengo

$$B^*=1/B'=-1.25 \text{ cm/s}$$

$$A^*=-A'/B'=7.75 \text{ cm}$$

Si osserva che nel caso in cui le coppie (x_i, y_i) sia precisamente su una retta

$$B = \frac{1}{B'}$$

ovvero si ottiene $B \cdot B' = 1$.

Si riconsideri il prodotto suddetto per il **caso a**, si ottiene 0.892.

Caso c: coppie (x_i, y_i) che non in relazione lineare fra loro:

Prendo i seguenti punti

Variabile indipendente $x[s]$	Variabile dipendente $y[cm]$
1	2
3	15
5	3

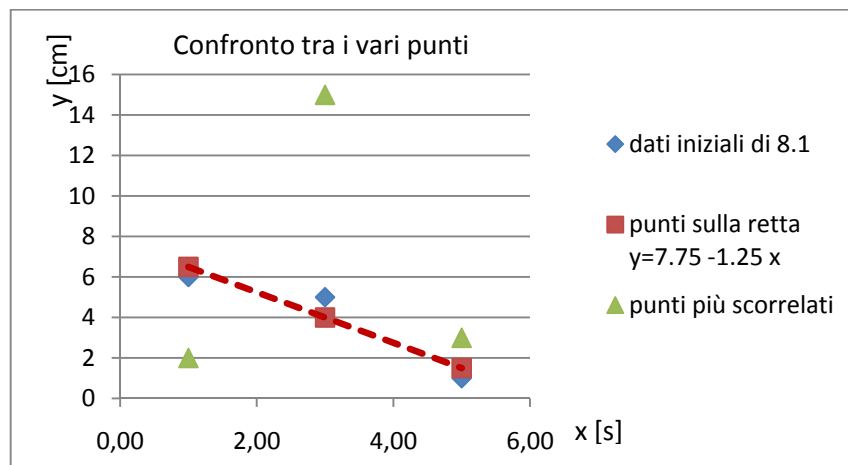


Figura 2. Grafico che riporta i dati iniziali (♦) del caso a, i punti su una retta (■) del caso b, e punti più scorrelati (▲) caso c.

La retta $y = A + Bx$ che si adatti meglio a questi dati, ottenuta dal metodo dei minimi quadrati viene riportata in figura 3.

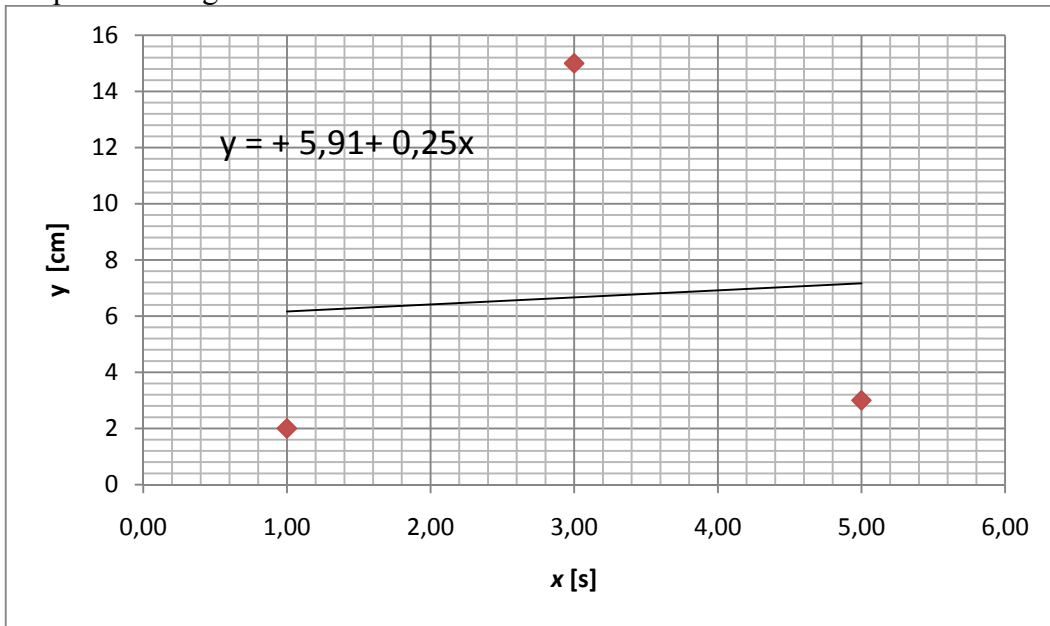


Figura 3. Nel grafico sono riportati i punti (♦) del caso c e la retta di regressione.

Si trovano delle costanti con errori notevoli $A = 5.91 \pm 12.3$ cm e $B = 0.25 \pm 3.61$ cm/s. Si applica il metodo dei minimi quadrati invertendo le variabili x e y ovvero per $y' = A' + B'x'$, con le stesse considerazioni di cui sopra si ha invece per i coefficienti i seguenti valori $A' = 2.87 \pm 2.91$ s, $B' = 0.019 \pm 0.85$ s/cm. Si osservi in entrambi i casi, quanto sono elevate le incertezze sui coefficienti. Se si confrontano i coefficienti delle due equazioni $y = A + Bx$ e $y = A' + B'x$, dedotta da A' e B' , si osserva che sono molto diversi, come può vedersi anche dalla figura 4.

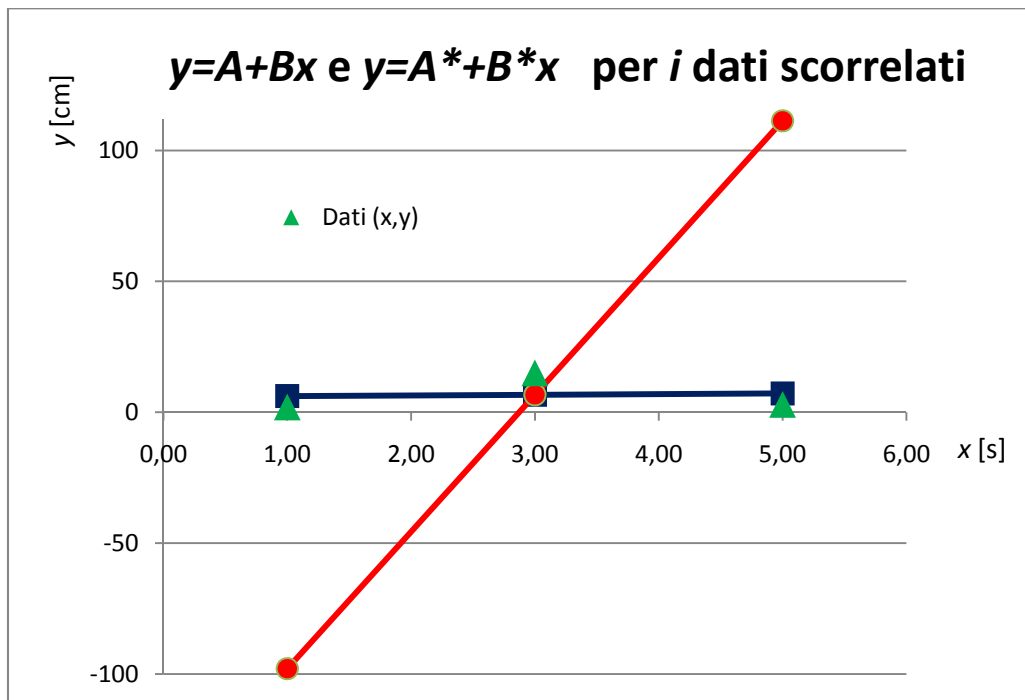


Figura 4. Nel grafico sono riportati i punti scorrelati (x_i, y_i) (▲) del caso c, la retta $y = A + Bx$ (■) ottenuta applicando il metodo dei minimi quadrati alle coppie (x_i, y_i) , la retta $y = A' + B'x$ (●)

dedotta applicando il metodo dei minimi quadrati alle coppie (x'_i, y'_i) e ricavando A^* e B^* da A' e B' .

Si ha che la l' adattamento $y=A+Bx$ tende ad appiattirsi orizzontalmente, mentre $y=A^*+B^*x$, dedotta dall'adattamento $y'=A'+B'x'$ tende ad allinearsi lungo l'asse delle y .

Si osservi che in questo caso sia **B che B' tendono a 0**, quindi quanto più le coppie di punti non seguono un andamento lineare tanto più $B \cdot B'$ tenderà a zero.

Questo ci conferma che il prodotto $B \cdot B'$ è un buon stimatore di quanto siano bene o mal **correlati linearmente** le coppie di punti (x_i, y_i) .

Sicuramente possiamo dire che quanto più tale prodotto si avvicina ad 1 tanto più le variabili x e y seguono una **relazione lineare**, quanto più si avvicina allo 0 tanto meno seguono una relazione lineare.

Sembra che questo prodotto $B \cdot B'$ sia un buon stimatore della relazione lineare tra due variabili x e y . Possiamo sicuramente dire che se la relazione tra le grandezze x e y è lineare il prodotto $B \cdot B'$ tenderà a 1, diversamente quanto meno saranno correlati linearmente tanto più tale prodotto tenderà a zero.

Possiamo generalizzare e formalizzare questa equivalenza e quindi utilizzare poi tale prodotto per verificare immediatamente se alcune variabili sono correlate linearmente fra loro?

*****Derivazione del coefficiente di correlazione lineare*****

Per fare questo dobbiamo esplicitare il prodotto $B \cdot B'$ rispetto alle coppie (x_i, y_i) .

Il coefficiente B è legato a tali coppie dalla relazione dedotta con il metodo dei minimi quadrati:

$$B = \frac{N \sum_{i=1}^N (x_i y_i) - (\sum_{i=1}^N x_i)(\sum_{i=1}^N y_i)}{N \sum_{i=1}^N (x_i^2) - (\sum_{i=1}^N x_i)^2}$$

Nella forma suddetta sono esplicitati i pedici sui quali si fanno le sommatorie, spesso risulta comodo sottintendere gli indici delle sommatorie:

$$B = \frac{N \sum(xy) - \sum(x) \sum(y)}{N \sum(x^2) - (\sum x)^2}$$

(Equ. 1)

dove si utilizzano le parentesi per distinguere opportunamente su cosa agiscono le sommatorie, su alcuni testi non vengono utilizzate (come sul Taylor stesso).

L'equazione 1 è la formula generale per trovare il coefficiente B di una serie di coppie di punti (x_i, y_i) dobbiamo ricavare una descrizione del prodotto $B \cdot B'$ rispetto alle suddette coppie.

L'equazione per B ricavato dall'applicazione dei minimi quadrati a $y=A+Bx$		l'equazione 1, etichettata equazione 1', per B' , ricavato dall'applicazione dei minimi quadrati a $y'=A'+B'x'$
$B = \frac{N \sum(xy) - \sum(x) \sum(y)}{N \sum(x^2) - (\sum x)^2}$ (Equ.1)		$B' = \frac{N \sum(x'y') - \sum(x') \sum(y')}{N \sum(x'^2) - (\sum x')^2}$ (Equ. 1')
Dato che $x'=y$ e $y'=x$ si ha che l'equazione 1' diventa		
		$B' = \frac{N \sum(yx) - \sum(y) \sum(x)}{N \sum(y^2) - (\sum y)^2}$
Si può ricavare quindi lo stimatore di linearità dal prodotto delle seguenti equazione		
$B = \frac{N \sum(xy) - \sum(x) \sum(y)}{N \sum(x^2) - (\sum x)^2}$	•	$B' = \frac{N \sum(yx) - \sum(y) \sum(x)}{N \sum(y^2) - (\sum y)^2}$

$$B \cdot B' = \frac{N \sum(xy) - \sum(x) \sum(y)}{N \sum(x^2) - (\sum x)^2} \cdot \frac{N \sum(yx) - \sum(y) \sum(x)}{N \sum(y^2) - (\sum y)^2}$$

(Equ. 2)

L'equazione 2 risulta si può riscrivere utilizzando alcune uguaglianze $\sum(x) = N\bar{x}$, $\sum(y) = N\bar{y}$,

$$B \cdot B' = \frac{N \sum(xy) - N\bar{x}N\bar{y}}{N \sum(x^2) - (N\bar{x})^2} \cdot \frac{N \sum(yx) - N\bar{y}N\bar{x}}{N \sum(y^2) - (N\bar{y})^2} = \frac{(\sum(xy) - N\bar{x}\bar{y})^2}{(\sum(x^2) - N(\bar{x})^2)(\sum(y^2) - N(\bar{y})^2)}$$

Si ricordi che possiamo esprimere

$\sum(x^2) - N(\bar{x})^2 = \sum(x - \bar{x})^2$, che ovviamente vale anche per la y e che $\sum(xy) - N\bar{x}\bar{y} = \sum[(x - \bar{x})(y - \bar{y})]$.

Il prodotto $B \cdot B'$ viene presentato anche come r^2 (alcuni autori utilizzano anche R^2) e la sua la sua radice quadrata è detta coefficiente di correlazione lineare r

Quindi si ha, riesplicitando gli indici delle sommatorie

$$r = \sqrt{B \cdot B'} = \frac{\sum_{i=1}^N [(x_i - \bar{x}) \cdot (y_i - \bar{y})]}{\left(\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \right) \cdot \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}$$

Si osservi che al denominatore abbiamo la sommatoria degli scarti quadratici di x_i rispetto al valore medio $\bar{x}$. Si faccia attenzione perché si parla delle x_i diverse relative alle N coppie (x_i, y_i) . Si trova anche la sommatoria degli scarti quadratici delle y_i . Al numeratore abbiamo invece la sommatoria dei prodotti misti degli scarti.

Il coefficiente r in valore assoluto è compreso tra 0 e 1, il coefficiente r invece è negativo che la pendenza della linearità è negativa e positivo al contrario. Infatti per il caso **a**) si trova $r = -0,945$, che al quadrato fornisce BB' .

La covarianza e la propagazione degli errori (somma lineare o in quadratura?).

Tale argomento sarà trattato senza approfondire tutti i dettagli, in quanto già trattato nella propagazione degli errori ed inoltre descritto anche nel paragrafo 9.2. La differenza è nel collegare la covarianza di due grandezze x e y al coefficiente di correlazione lineare.

Possiamo approssimare una funzione g dipendente da più variabili $(x, y, z, \dots)$ secondo lo sviluppo al primo ordine del tipo:

$$g(x, y, z, \dots) \approx g(\bar{x}, \bar{y}, \bar{z}, \dots) + \left(\frac{\partial g}{\partial x} \right)_{(\bar{x}, \bar{y}, \bar{z}, \dots)} (x - \bar{x}) + \left(\frac{\partial g}{\partial y} \right)_{(\bar{x}, \bar{y}, \bar{z}, \dots)} (y - \bar{y}) + \left(\frac{\partial g}{\partial z} \right)_{(\bar{x}, \bar{y}, \bar{z}, \dots)} (z - \bar{z}) \dots$$

Per l'analisi degli errori, g è una grandezza fisica misurata indirettamente mediante la misura diretta delle grandezze x e y .

Orientiamo tale sviluppo di g in (x, y, z) molto vicini a $(\bar{x}, \bar{y}, \bar{z})$ a due sole variabili, ed inoltre se tale sviluppo vale per ogni x , varrà anche per ogni x_i , per cui si avrà

$$g(x_i, y_i) \approx g(\bar{x}, \bar{y}) + \left(\frac{\partial g}{\partial x} \right)_{(\bar{x}, \bar{y})} (x_i - \bar{x}) + \left(\frac{\partial g}{\partial y} \right)_{(\bar{x}, \bar{y})} (y_i - \bar{y})$$

Per semplicità in seguito si ometterà il pedice sotto le derivate parziali, che sta ad indicare che una volta effettuate le derivate parziali della funzione g , bisogna calcolarle in $(\bar{x}, \bar{y})$.

Per trovare il valore medio di g :

$$\bar{g} = \frac{1}{N} \sum_{i=1}^N g(x_i, y_i) \approx \frac{1}{N} \sum_{i=1}^N g(\bar{x}, \bar{y}) + \left(\frac{\partial g}{\partial x}\right) \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) + \left(\frac{\partial g}{\partial y}\right) \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})$$

dove al secondo membro il 2° ed il 3° termine sono nulli, si ha quindi $\bar{g} = g(\bar{x}, \bar{y})$.

Si ottiene anche lo scarto quadratico medio e lo si utilizza nel calcolo della varianza. Si ribadisce in questa sede che in statistica si ha la varianza (della popolazione) e la varianza campionaria (del campione)

$$\text{varianza} \quad s_g^2 = \frac{1}{N} \sum_{i=1}^N (g_i - \bar{g})^2$$

$$\text{varianza campionaria} \quad \sigma_g^2 = \frac{1}{N-1} \sum_{i=1}^N (g_i - \bar{g})^2$$

per le quali non ci preoccupiamo neanche di usare simboli diversi. Nel caso dell'utilizzo della statistica per la fisica sperimentale, avremo sempre il campione come riferimento, in quanto potremmo fare un'infinità di misure (popolazione infinita).

Nelle considerazioni successive, la divisione per N o $N-1$ non cambia i termini della questione.

Per la varianza si avrà:

$$s_g^2 = \frac{1}{N} \sum_{i=1}^N (g_i - \bar{g})^2 = \frac{1}{N} \sum_{i=1}^N \left[\left(\frac{\partial g}{\partial x}\right) (x_i - \bar{x}) + \left(\frac{\partial g}{\partial y}\right) (y_i - \bar{y}) \right]^2$$

Si possono esprimere gli scarti con il simbolo Δ e quindi:

$$s_g^2 = \frac{1}{N} \sum_{i=1}^N (\Delta g_i)^2 = \frac{1}{N} \sum_{i=1}^N \left[\left(\frac{\partial g}{\partial x}\right)^2 (\Delta x_i)^2 + \left(\frac{\partial g}{\partial y}\right)^2 (\Delta y_i)^2 + 2 \left(\frac{\partial g}{\partial x}\right) \left(\frac{\partial g}{\partial y}\right) (\Delta x_i)(\Delta y_i) \right]$$

Si osserva che

$$s_g^2 = \left(\frac{\partial g}{\partial x}\right)^2 s_x^2 + \left(\frac{\partial g}{\partial y}\right)^2 s_y^2 + 2 \left(\frac{\partial g}{\partial x}\right) \left(\frac{\partial g}{\partial y}\right) \frac{1}{N} \sum_{i=1}^N (\Delta x_i)(\Delta y_i)$$

in cui compaiono termini già noti, quali le varianze di x e di y e il termine dei prodotti misti.

Abbiamo ora argomenti più solidi per affrontare tale prodotto misto e segnalare che viene definito **covarianza** della popolazione s_{xy}

$$s_{xy} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

o nel caso del campione **covarianza** campionaria

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

Si può osservare che il coefficiente di correlazione tra due grandezze x e y e le varianze di x e y , nonché la covarianza si osserva che possiamo trovare una relazione immediata:

$$r = \frac{\sum_{i=1}^N [(x_i - \bar{x}) \cdot (y_i - \bar{y})]}{\left(\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \right) \cdot \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}} = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

Si osserva immediatamente che se le due grandezze x ed y non sono correlate linearmente tra loro la sommatoria dei prodotti misti tende ad un numero prossimo a zero, che diviso per N risulterà

ancora più trascurabile, per cui nel caso di variabili indipendenti e casuali (che non hanno correlazione lineare) si avrà per la deviazione standard del campione:

$$\sigma_g = \sqrt{\left(\frac{\partial g}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial g}{\partial y}\right)^2 \sigma_y^2}$$

Ovvero gli errori si propagano in quadratura.

Se la correlazione è positiva, ad una sovrastima di x si avrà sempre una sovrastima di y , ad una sottostima di x si avrà sempre una sotto stima di y , pertanto ogni prodotto misto nella sommatoria della covarianza sarà positivo e la loro somma non trascurabile.

Così se si otterrà una correlazione negativa ad una sovrastima di x corrisponderà sempre una sottostima di y e viceversa, quindi la covarianza sarà negativa.

In un approccio immediato, in cui si individua una relazione tra le grandezze in x e y è data la relazione

$$|\sigma_{xy}| \leq \sigma_x \sigma_y$$

(disuguaglianza di Schwarz si pensi a due vettori $\vec{x}$ ed $\vec{y}$, si ha che il valore assoluto del loro prodotto scalare è sempre minore o uguale al prodotto dei moduli dei due vettori: $|\vec{x} \cdot \vec{y}| \leq |\vec{x}| |\vec{y}|$).

Da questo punto in poi si ritiene soddisfacente la trattazione del Taylor per motivare la somma lineare ovvero

$$\sigma_g \leq \left| \frac{\partial g}{\partial x} \right| \sigma_x + \left| \frac{\partial g}{\partial y} \right| \sigma_y$$

in cui si ha un limite superiore all'errore sulla grandezza g , stimato dalla somma lineare, per cui si richiede allo studente di completare l'argomento dal capitolo 9.