



corso di Laurea in Scienze Biologiche

Lucidi Proiettati alle lezioni di

Laboratorio di Fisica

a.a. 1998-99

Livio Narici

LA MISURA E IL SUO ERRORE

una misura non è completa senza la sua incertezza.

esempio

ci pesiamo: 52.5 kg

ci pesiamo il giorno dopo: 53.2 kg

⇒ siamo ingrassati!?

Durante il corso questi lucidi hanno costituito lo spunto per le lezioni. Possono quindi essere utilizzati come riferimento e traccia, devono però essere completati con gli appunti presi a lezione e con il testo.

Segnalazioni di errori od inesattezze saranno gradite.

Avviciniamoci a questo problema con un atteggiamento scientifico.

La curiosità dovrebbe farci porre la seguente domanda: se il nostro "vero peso" non fosse cambiato, la bilancia avrebbe indicato lo stesso peso o no?

Per risponderci dobbiamo semplicemente provare e riprovare a pesarci. Dovremmo tentare, salendo in modi diversi sulla bilancia, posizionandoci in modi simili ma non uguali (peso sulle punte, sui talloni, etc) per vedere se il peso indicato è sempre lo stesso o no.

Fatelo: avrete sorprese!

supponiamo di averlo fatto ottenendo la seguente serie di misure (in kg):

52.7 52.4 52.0 52.5 53.2

53.1 52.2 52.6 52.3 53.0

⇒ da un minimo di 52.0 kg ad un massimo di 53.2 kg!

Intuitivamente pensiamo che la media di queste misure ($\langle P \rangle = 52.6$ kg) sia una ragionevole stima del nostro peso

possiamo scrivere

52.6 ± 0.6 kg

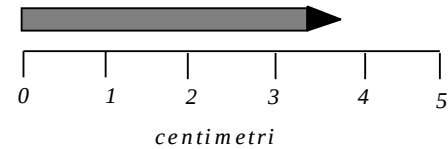
Cosa ci dice tutto ciò in merito al nostro problema originale?

Quali considerazioni possiamo fare in merito allo strumento, al suo errore ed alle nostre misure di peso?

Cosa ci suggerisce quanto detto sulla possibilità di confrontare nostre misure con "valori accettati"?

Errori di lettura e misure ripetute

Consideriamo ora una scala graduata:



E se potessimo ripetere la nostra misura come prima?

I risultati di misure ripetute ci danno importanti informazioni riguardo gli errori.

..... sempre ???

Errori e cifre significative

Consideriamo ora le seguenti misure di tempo (in s):

1.6 1.2 1.4 1.5 1.4 1.5

la cui media è

1.433333333333333333333333... s

ma, quanti "3" (in generale quante cifre) è corretto scrivere??

i valori minimi e massimi sono

1.2 s ed 1.6 s

$\Rightarrow 1.4 \pm 0.2$ s

Discrepanza e propagazione degli errori

Il confronto di due misure (con i loro errori) ci permette di introdurre il concetto di discrepanza e di "propagazione degli errori"

Misuriamo la quantità di moto di due carrelli che scorrono su una rotaia "priva" di attrito prima e dopo un urto fra loro:

$$p = 1.49 \pm 0.04 \text{ kg m/s}$$

$$p' = 1.56 \pm 0.06 \text{ kg m/s}$$

Il nostro risultato è compatibile con il principio di conservazione della quantità di moto?

$$\text{cioè: } p - p' = 0 ?$$

Le nostre misure ci dicono che

$$p - p' = - 0.07 \text{ kg m/s}$$

ma come si propagano gli errori su p e p' sulla differenza $p - p'$?

Si può scrivere:

$$(p - p')_{\min} = [p - \delta_p - (p' + \delta_{p'})]$$

$$(p - p')_{\max} = [p + \delta_p - (p' - \delta_{p'})]$$

$$\Rightarrow \delta_{p-p'} = \frac{(p - p')_{\max} - (p - p')_{\min}}{2}$$

{semidispersione massima}

$$\delta_{p-p'} = \frac{1}{2} [p + \delta_p - (p' - \delta_{p'}) - p + \delta_p + (p' + \delta_{p'})]$$

$$= \frac{1}{2} [2\delta_p + 2\delta_{p'}] = \delta_p + \delta_{p'}$$

$$\Rightarrow p - p' = -0.07 \pm 0.10 \text{ kg m/s}$$

Nota: in generale:

$$q = x - y$$

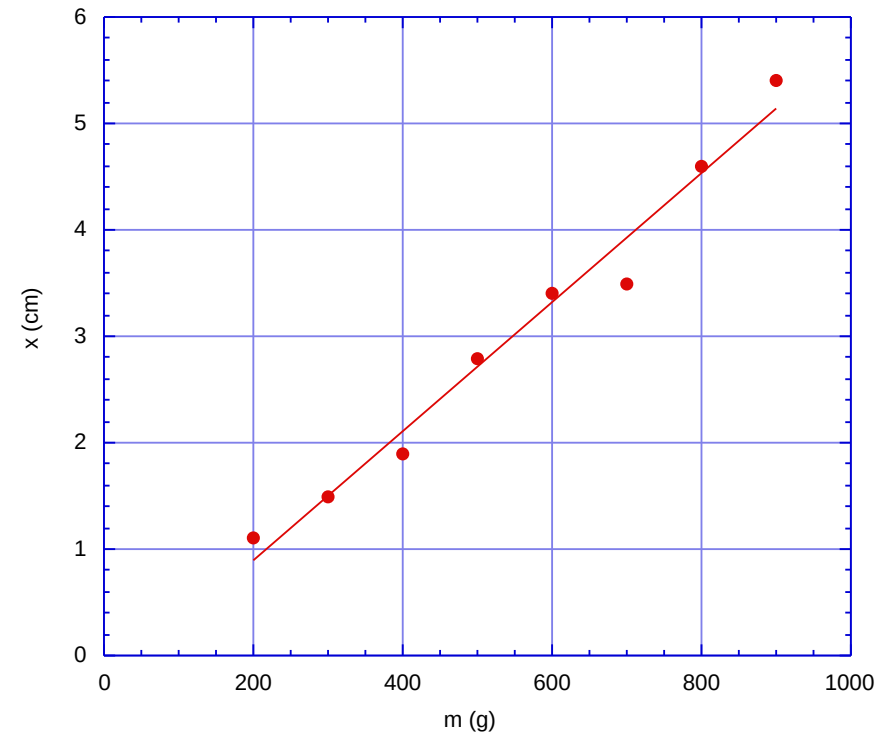
$$\delta_q = \delta_x + \delta_y$$

PROPORZIONALITÀ

Per studiare la relazione tra due grandezze misurate è raccomandato riportare le misure su di un grafico.

Misuriamo l'allungamento x di una molla in funzione della massa m ad essa connessa.

$m(g)$	$x(cm)$
200	1.1
300	1.5
400	1.9
500	2.8
600	3.4
700	3.5
800	4.6
900	5.4



Le nostre misure appaiono non compatibili con la legge di Hooke:

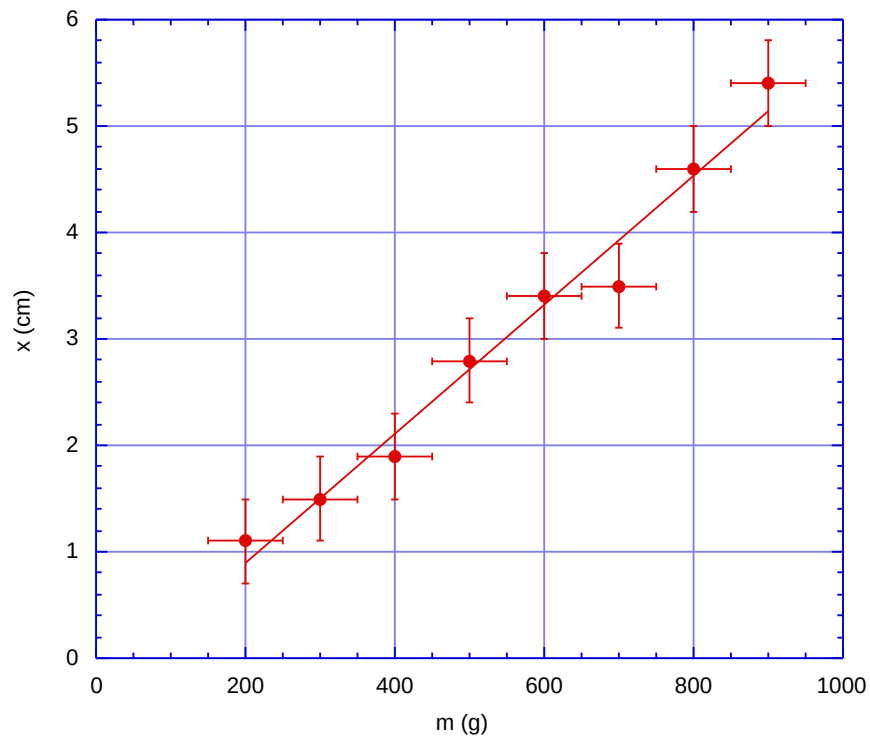
$$x = \frac{g}{k} m$$

Anche in questo caso abbiamo trascritto su grafico delle informazioni parziali, non ci siamo infatti interessati agli errori. Assumiamo che:

$$\delta_m = 50 \text{ g}$$

$$\delta_x = 0.4 \text{ cm}$$

e riportiamo questa informazione sul grafico:



Errori relativi

Il valore un errore in una misura è una informazione parziale anche rispetto ad una valutazione dello stesso errore.

Si pensi all'errore descritto all'inizio (0.6 kg) relativamente a misure del nostro peso. Lo stesso errore se relativo a misure, ad esempio, di peso di alimenti (per esempio legato a pesate di una bilancia per spaghetti) assume un aspetto ben diverso!

Per descrivere questa caratteristica si introduce l'errore relativo.

Se la misura di una certa grandezza è:

$$x \pm \delta_x$$

il suo errore relativo è:

$$\delta_{r,x} = \frac{\delta_x}{|x|}$$

(dove il valore assoluto mantiene positivo il valore dell'errore per qualsiasi x)

È facile comprendere che l'errore relativo è in qualche modo legato al numero delle cifre significative:

N. cifre	Errore relativo corrispondente	
	è fra	è (approssimativamente)
2	5% e 50%	10%
3	0.5% e 5%	1%
4	0.05% e 0.5%	0.1%

Cosa ci suggerisce questo relativamente all'uso delle cifre significative?

PROPAGAZIONE DEGLI ERRORI

La misura di una grandezza è assai spesso indiretta.

(area, velocità, accelerazione di gravità, densità, .. etc)

Abbiamo visto come stimare/calcolare l'errore di una misura diretta, ad esempio su $x, y, \dots, z$.

Ora assumiamo di voler calcolare l'errore su una grandezza derivata:

$$a) \quad q = x \pm y$$

$$b) \quad q = xy \text{ oppure } q = x/y$$

$$c) \quad q = f(x, y, \dots, z)$$

[dove ovviamente l'ultimo caso comprende le prime due]

Somma e sottrazione

La strategia per calcolare l'errore su q è analoga a quanto già visto.

Si calcolano i valori massimi e minimi possibili di q e da questi la semidispersione massima:

$$a) \quad q_{+max} = x_b + \delta x + y_b + \delta y$$

$$q_{+min} = x_b - \delta x + y_b - \delta y$$

$$q_{-max} = x_b + \delta x - (y_b - \delta y)$$

$$q_{-min} = x_b - \delta x - (y_b + \delta y)$$

$$\begin{aligned} \Rightarrow \delta q_+ &= \frac{1}{2} \{x_b + \delta x + y_b + \delta y - \\ &\quad (x_b - \delta x + y_b - \delta y)\} \\ &= \delta x + \delta y \end{aligned}$$

$$\begin{aligned} \Rightarrow \delta q_- &= \frac{1}{2} \{x_b + \delta x - (y_b - \delta y) - \\ &\quad [x_b - \delta x - (y_b + \delta y)]\} \\ &= \delta x + \delta y \end{aligned}$$

Cioè

$$\delta q_{\pm} = \delta x + \delta y$$

Prodotto

$$q = x_b y_b$$

$$q_{max} = (x_b + \delta x)(y_b + \delta y)$$

$$q_{min} = (x_b - \delta x)(y_b - \delta y)$$

$$\begin{aligned} \Rightarrow \delta q &= \frac{1}{2} \{ (x_b + \delta x)(y_b + \delta y) - \\ &\quad (x_b - \delta x)(y_b - \delta y) \} \\ &= \frac{1}{2} \{ x_b y_b + y_b \delta x + x_b \delta y + \delta x \delta y - \\ &\quad x_b y_b + y_b \delta x + x_b \delta y - \delta x \delta y \} \\ &= \frac{1}{2} \{ y_b \delta x + x_b \delta y + y_b \delta x + x_b \delta y \} \\ &= x_b \delta y + y_b \delta x \end{aligned}$$

cioè

$$\delta q = |x_b| \delta y + |y_b| \delta x$$

{dove il valore assoluto serve a considerare sempre positivi i contributi degli errori}

Quoziente

$$q = \frac{x}{y}$$

(da ora in poi, per semplicità, non scriveremo più il sottoscritto "b")

$$q_{max} = \frac{x + \delta x}{y - \delta y}$$

$$q_{min} = \frac{x - \delta x}{y + \delta y}$$

$$\begin{aligned} \Rightarrow \delta q &= \frac{1}{2} \left\{ \frac{x + \delta x}{y - \delta y} - \frac{x - \delta x}{y + \delta y} \right\} \\ &= \frac{1}{2} \left\{ \frac{(x + \delta x)(y + \delta y) - (x - \delta x)(y - \delta y)}{y^2 - \delta y^2} \right\} \\ &= \frac{1}{2} \left\{ \frac{xy + y\delta x + x\delta y + \delta x\delta y - xy + y\delta x + x\delta y - \delta x\delta y}{y^2 - \delta y^2} \right\} \\ &= \frac{1}{2} \left\{ \frac{y\delta x + x\delta y + y\delta x + x\delta y}{y^2 - \delta y^2} \right\} = \left\{ \frac{x\delta y + y\delta x}{y^2 - \delta y^2} \right\} \\ &= \frac{1}{1 - \frac{\delta y^2}{y^2}} \left\{ \frac{x\delta y}{y^2} + \frac{\delta x}{y} \right\} \end{aligned}$$

cioè:

$$\delta q = \frac{1}{1 - \frac{\delta y^2}{y^2}} \left\{ \frac{x\delta y}{y^2} + \frac{\delta x}{y} \right\}$$

 inciso di matematica

Ricordiamo la serie di Mac Laurin

(Taylor in zero):

$$f(x_{\text{piccolo}}) = f(0) + x f'(0) + \frac{1}{2} x^2 f''(0) + \dots$$

nel nostro caso $x = \frac{\delta y^2}{y^2}$

$$f(x) = \frac{1}{1-x}; f'(x) = \frac{1}{(1-x)^2}; f''(x) = \frac{2}{(1-x)^3}$$

$$f(0) = 1; f'(0) = 1; f''(0) = 2$$

$$\frac{1}{1-x} = 1 + x + x^2 + \dots$$

che, in questo caso, corrisponde anche al teorema binomiale.

δy è piccolo, a maggior ragione lo è δy^2 , possiamo quindi trascurare i termini in δy^2 e minori:

$$\frac{1}{1 - \frac{\delta y^2}{y^2}} \approx 1$$

quindi: $\delta q = \frac{x \delta y}{y^2} + \frac{\delta x}{y}$

Propagazione con gli errori relativi

ricordando che

$$q = \frac{x}{y}$$

$$\delta q = q \left\{ \frac{\delta y}{y} + \frac{\delta x}{x} \right\}$$

$$\Rightarrow \frac{\delta q}{q} = \frac{\delta y}{y} + \frac{\delta x}{x}$$

(dove x e y sono sempre i valori assoluti)

cioè l'errore relativo di un quoziente è la somma degli errori relativi del numeratore e del denominatore.

Anche per il prodotto, dove avevamo trovato che

$$q = xy \Rightarrow \delta q = x \delta y + y \delta x$$

possiamo scrivere:

$$\frac{\delta q}{q} = \frac{\delta y}{y} + \frac{\delta x}{x}$$

In generale possiamo dire che l'errore relativo su una grandezza q

$$q = \frac{xy \dots z}{uv \dots w}$$

è dato dalla somma degli errori relativi di tutte le grandezze:

$$\frac{\delta q}{q} = \frac{\delta x}{x} + \frac{\delta y}{y} + \dots + \frac{\delta z}{z} + \frac{\delta u}{u} + \frac{\delta v}{v} + \dots + \frac{\delta w}{w}$$

Un caso particolare di quanto detto si ha quando si vuole propagare l'errore su una grandezza derivata come prodotto (rapporto) di una grandezza con un errore noto ed un "numero" non affetto da errore.

In questo caso l'errore è dato dall'errore noto, moltiplicato (diviso) il numero esatto.

Inoltre nel caso di una potenza si ha, ovviamente:

$$q = x^n \Rightarrow \frac{\delta q}{q} = n \frac{\delta x}{x}$$

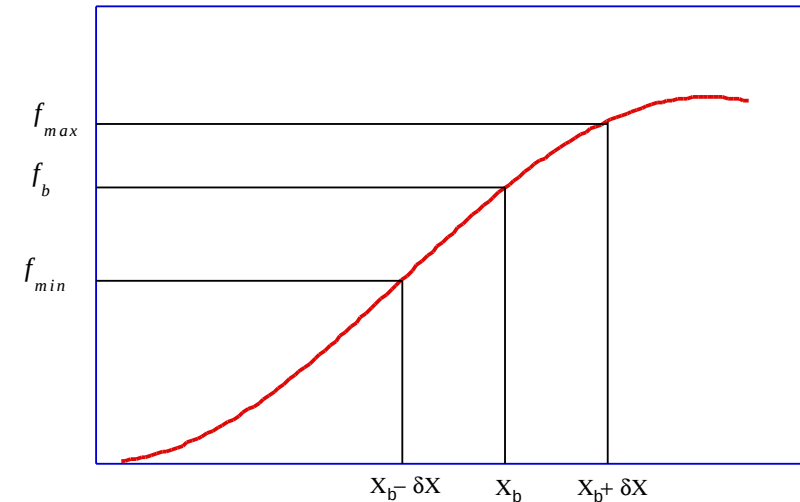
SOMMARIO:

quando le grandezze misurate si sommano o sottraggono: gli errori si sommano

quando le grandezze misurate si moltiplicano o dividono: gli errori relativi si sommano

Funzione di una variabile

Nel caso di una funzione di una variabile $f(x)$ come si propaga l'errore da x a f ?



vediamo che

$$\delta f_b = f(x + \delta x) - f(x_b) \Rightarrow \frac{df}{dx} \delta x$$

in generale:

$$\delta f = \left| \frac{df}{dx} \right| \delta x$$

La propagazione passo passo

Le regole che abbiamo imparato sino ad ora ci consentono di calcolare "passo passo" la propagazione degli errori anche per relazioni piuttosto complicate, ad esempio:

$$q = x(y - z \sin\theta)$$

In alcuni casi, però, questa procedura può sovrastimare l'errore. Si supponga, ad esempio, di misurare tre grandezze x , y e z e calcolare

$$q = \frac{x + y}{x + z}$$

nella quale una variabile compare più di una volta.

Risulta evidente che gli errori su x a numeratore possono cancellare quelli su x a denominatore, mentre la nostra procedura vedrebbe le due x come due variabili diverse e sommerebbe gli errori (si pensi ad una relazione tipo $q = \frac{x}{x}$).

La formula generale

La generalizzazione a più variabili della formula di propagazione per una funzione ad una variabile vista prima è semplice:

$$q(x_b, y_b, \dots, z_b) \pm \delta q(x_b, y_b, \dots, z_b)$$

$$\delta q(x_b, y_b, \dots, z_b) = \left| \frac{\partial q}{\partial x} \right| \delta x + \left| \frac{\partial q}{\partial y} \right| \delta y + \dots + \left| \frac{\partial q}{\partial z} \right| \delta z$$

Errori indipendenti

Questi nostri risultati sono "pessimisti". Se gli errori sulle grandezze misurate sono indipendenti, ci sarà una certa probabilità che l'errore su una variabile possa parzialmente compensare quello sull'altra.

Ciò che abbiamo studiato rappresenta il limite superiore nel caso di errori completamente dipendenti.

Nel caso di errori indipendenti gli errori si sommano "in quadratura".

In generale avremo quindi

$$\delta q = \sqrt{\sum_{i=1}^N \left(\frac{\partial q}{\partial x_i} \delta x_i \right)^2}$$

da cui si ricava, ad esempio, per le somme:

$$\delta q = \sqrt{\sum_{i=1}^N \delta x_i^2}$$

o, per gli errori relativi:

$$\frac{\delta q}{|q|} = \sqrt{\sum_{i=1}^N \left(\frac{\delta x_i}{x_i} \right)^2}$$

Provate ora a calcolare l'errore su g (accelerazione di gravità), misurato indirettamente con misure di tempo e lunghezza su un pendolo:

$$g = 4\pi^2 \frac{L}{T^2}$$

con $L = 92.9 \pm 0.1 \text{ cm}$; $T = 1.936 \pm 0.004 \text{ s}$

ANALISI STATISTICA DEGLI ERRORI CASUALI

Il vantaggio menzionato in precedenza del ripetere più volte una misura dipende dalla casualità degli errori.

Gli errori casuali possono essere trattati statisticamente.

Si distinguono da questi gli errori sistematici, che non è possibile trattare statisticamente.

Attenzione: la separazione tra errore casuale ed errore sistematico è spesso sottile e dipende anche dalla capacità dello sperimentatore di fare le misure.

La media

Abbiamo già parlato, senza soffermarci, di media. È ragionevole, e lo abbiamo già accettato, che la migliore stima " x_b " di una serie di N misure x_i sia proprio la media (la dimostrazione la vedremo in seguito)

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

Ritorniamo alle nostre misure di peso (P , in kg) che abbiamo considerato all'inizio ($N=10$):

52.7 52.4 52.0 52.5 53.2 53.1 52.2 52.6 52.3 53.0

la cui media è

$$\langle P \rangle = 52.6$$

Una quantità interessante è la deviazione " d " di ogni singola misura dalla media:

<i>misura</i>	<i>peso(kg)</i>	<i>deviazione(kg)</i>
1	52.7	0.1
2	52.4	-0.2
3	52.0	-0.6
4	52.5	-0.1
5	53.2	0.6
6	53.1	0.5
7	52.2	-0.4
8	52.6	0
9	52.3	-0.3
10	53.0	0.4
	$\langle x \rangle = 52.6$	$\langle d \rangle = 0$

La deviazione standard

Per stimare l'attendibilità della nostra media potremmo sommare tutte le deviazioni ma, ovviamente tale somma è zero:

$$\sum_{i=1}^N (x_i - \langle x \rangle) = \sum_{i=1}^N x_i - N \langle x \rangle = 0$$

Per evitare questo inconveniente eleviamo ogni deviazione al quadrato:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \langle x \rangle)^2}$$

abbiamo così introdotto la deviazione standard.

misura	peso(kg)	d(kg)	d ² (kg ²)
1	52.7	0.1	0.01
2	52.4	-0.2	0.04
3	52.0	-0.6	0.36
4	52.5	-0.1	0.01
5	53.2	0.6	0.36
6	53.1	0.5	0.25
7	52.2	-0.4	0.16
8	52.6	0	0
9	52.3	-0.3	0.09
10	53.0	0.4	0.16
$\langle x \rangle = 52.6$		$\langle d \rangle = 0$	$\sum_{i=1}^N d^2 = 1.4$

Quindi

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N d^2 = 0.14 \text{ kg}^2 \Rightarrow \sigma = 0.37 \text{ kg}$$

Dove σ^2 è la "varianza"

Ci sono argomenti (ci torneremo più avanti) per sostituire ad N $N-1$ nella definizione della deviazione standard e della varianza, che perciò divengono:

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \langle x \rangle)^2}$$

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \langle x \rangle)^2$$

La differenza con la versione usata prima è minima non appena N è sufficientemente grande. Notate che nel nostro caso σ non varia.

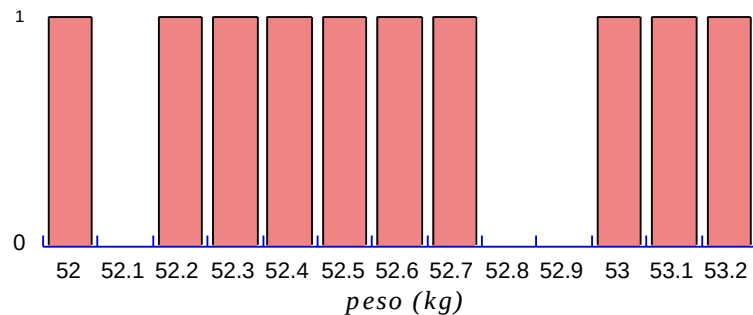
ISTOGRAMMI

Per mettere in evidenza come i risultati di molte misure di una stessa grandezza si "distribuiscono" possiamo costruire un istogramma.

Riconsideriamo le nostre misure di peso (tutte in kg):

52.7 52.4 52.0 52.5 53.2 53.1 52.2 52.6 52.3 53.0

Possiamo costruire il seguente grafico:

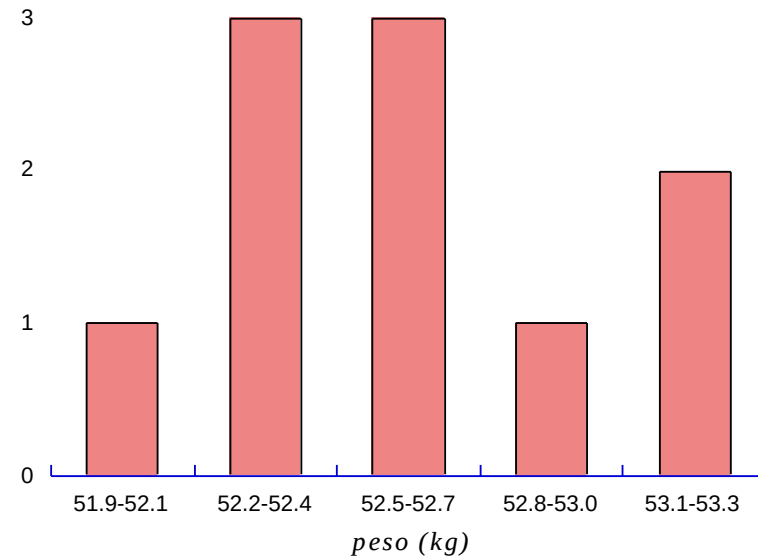


Questo è un istogramma. Costruito così, comunque, ci fornisce poche informazioni. Assai spesso a dieci misure corrisponderebbero 10 colonne alte tutte uno.

Ora raggruppiamo tutte le misure che capitano in un certo intervallo, largo Δ , ad esempio $\Delta = 0.3$ kg. Per comprendere tutte le misure partiamo da 51.9 ed arriviamo a 53.3:

k	intervallo	P_k	n_k
1	51.9-52.1	52.0	1
2	52.2-52.4	52.3	3
3	52.5-52.7	52.6	3
4	52.8-53.0	52.9	1
5	53.1-53.3	53.2	2

e l'istogramma diviene:



e già illustra più chiaramente come i dati si distribuiscono.

Alla luce di quanto visto possiamo riscrivere la media

$$\langle x \rangle = \frac{\sum_{i=1}^N x_i}{N}$$

in funzione del numero di volte n_k in cui si è trovato x_k nell'intervallo k_{mo}

$$\langle x \rangle = \frac{\sum_{k=1}^M x_k n_k}{N}$$

dove M è il numero di intervalli e ovviamente

$$\sum_{k=1}^M n_k = N$$

Un altro modo, a volte più conveniente, per descrivere quanto detto è introdurre la frazione delle nostre misure in ciascun intervallo:

$$F_k = \frac{n_k}{N}$$

quindi

$$\langle x \rangle = \sum_{k=1}^M x_k F_k$$

$$\sum_{k=1}^M F_k = 1$$

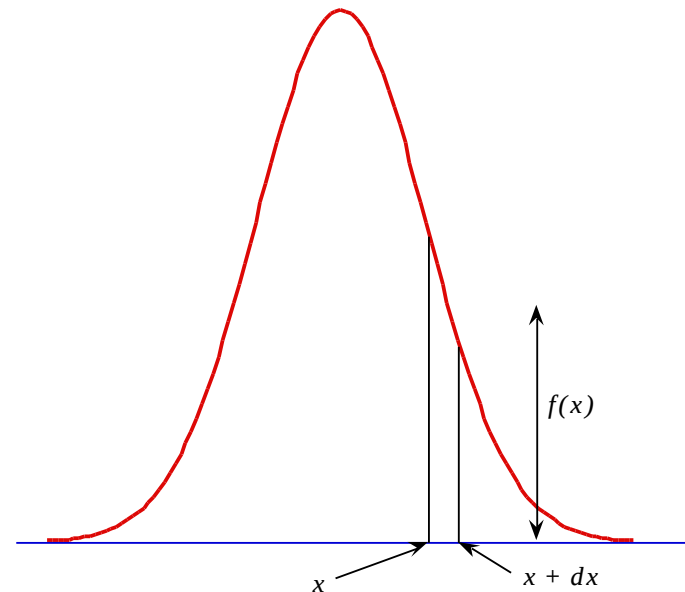
F_k è la distribuzione dei nostri risultati (normalizzata)

k		P_k	n_k	F_k
1	51.9-52.1	52.0	1	0.1
2	52.2-52.4	52.3	3	0.3
3	52.5-52.7	52.6	3	0.3
4	52.8-53.0	52.9	1	0.1
5	53.1-53.3	53.2	2	0.2

Distribuzioni limite

Se si aumenta il numero delle misure, l'istogramma costruito da queste tende quasi sempre ad una forma ben definita.

Al limite questo è una curva continua, ad esempio come:



Dove la frazione di misure che cadono fra x e $x+dx$ è data da $f(x)dx$, e la condizione di normalizzazione è:

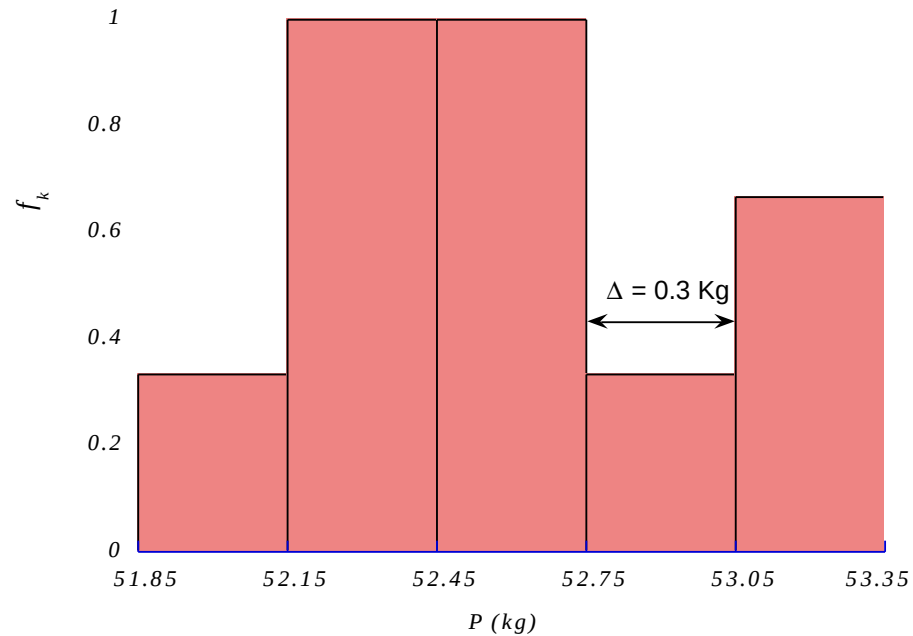
$$\int_{-\infty}^{\infty} f(x) dx = 1$$

In analogia a quanto ora visto è utile introdurre un modo diverso di esprimere la frazione di misure che cadono in un singolo intervallo Δ (vedi sopra). Scrivendo:

$$F_k = f_k \Delta \Rightarrow f_k = \frac{F_k}{\Delta}$$

In tal modo l'istogramma di prima diviene:

1.2



possiamo quindi scrivere

$$\sum_{k=1}^5 f_k \Delta = 1$$

(analogo a $\int_{-\infty}^{\infty} f(x) dx = 1$)

Con questa notazione la media si scrive:

$$\langle P \rangle = \sum_{k=1}^N P_k f_k \Delta \quad \text{oppure} \quad \langle P \rangle = \int_{-\infty}^{\infty} P f(P) dP$$

e, analogamente, essendo la varianza la media dello scarto quadratico:

$$\sigma^2 = \int_{-\infty}^{\infty} (P - \langle P \rangle)^2 f(P) dP$$

LA DISTRIBUZIONE NORMALE

Se una misura è soggetta a molte sorgenti di piccoli errori casuali, la sua distribuzione limite è una campana simmetrica (come quella illustrata prima): la distribuzione normale o di Gauss:

$$f(x) \propto e^{-(x-X)^2/2\sigma^2}$$

Dove X è il valore in cui la campana è centrata.

Ragioniamo: X appare una ottima stima del "valore vero"

Dovendo la $f(x)$ verificare la condizione di normalizzazione, troviamo la costante A per la quale deve essere moltiplicata:

$$\int_{-\infty}^{\infty} A e^{-(x-X)^2/2\sigma^2} dx = 1$$

poniamo $x-X = y$

$$= A \int_{-\infty}^{\infty} e^{-y^2/2\sigma^2} dy$$

ed ora $y/\sigma = z$ ($\Rightarrow dy = \sigma dz$)

$$= \sigma A \int_{-\infty}^{\infty} e^{-z^2/2} dz = \sigma A \sqrt{2\pi} = 1$$

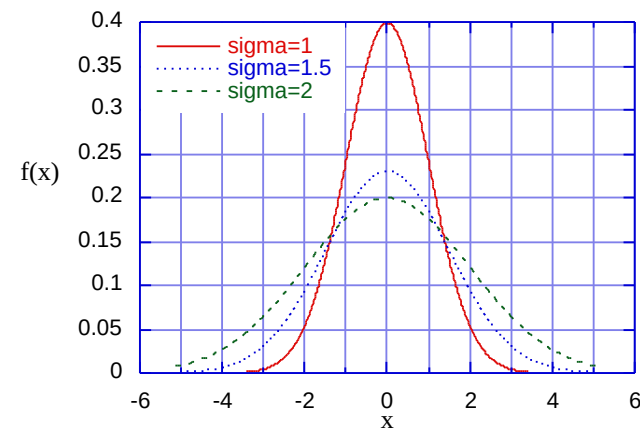
$$\Rightarrow A = \frac{1}{\sigma\sqrt{2\pi}}$$

Quindi:

$$f_{X\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-X)^2/2\sigma^2}$$

Il valore di σ è la semi-larghezza della curva ad $1/\sqrt{e}$ del suo massimo:

$$x = X \pm \sigma \Rightarrow f_{X\sigma}(X \pm \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2} = \frac{1}{\sqrt{e}} \frac{1}{\sigma\sqrt{2\pi}} = \frac{1}{\sqrt{e}} f_{X\sigma}(X)_{[max]}$$



Possiamo ora calcolare il valor medio $\langle x \rangle$ atteso dopo un gran numero di prove secondo la distribuzione di Gauss.

$$\langle x \rangle = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-X)^2/2\sigma^2} dx$$

poniamo $x-X = y$

$$\langle x \rangle = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} y e^{-y^2/2\sigma^2} dy + X \int_{-\infty}^{\infty} e^{-y^2/2\sigma^2} dy$$

$$= \frac{1}{\sigma\sqrt{2\pi}} X \int_{-\infty}^{\infty} e^{-z^2/2} \sigma dz = X \quad (\text{avendo posto } y/\sigma = z)$$

Come ci aspettavamo.

La deviazione standard ed il limite di confidenza del 68%

L'integrale di $f_{X,\sigma}(x)$ tra a e b

$$\int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-X)^2/2\sigma^2} dx$$

è la probabilità che una delle nostre misure dia un risultato

$$a \leq x \leq b.$$

La probabilità che una misura cada entro $t\sigma$ dal suo valor medio è, analogamente

$$\int_{X-t\sigma}^{X+t\sigma} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-X)^2/2\sigma^2} dx$$

che, con la solita sostituzione $(x-X)/\sigma=z$, $dx=\sigma dz$ diventa:

$$\frac{1}{\sqrt{2\pi}} \int_{-t}^t e^{-z^2/2} dz \equiv \text{erf}(t)$$

per $t = 1$ è pari a 0.68, per $t = 2$ a 0.95, per $t = 3$ a 0.997.

Adottando la deviazione standard σ come incertezza di una misura ripetuta più volte:

$$x = x_b \pm \sigma_x$$

Dove x_b è per esempio la media ed allora σ_x è la "sua" deviazione standard (della media).

Possiamo essere confidenti che il 68% delle misure cadano in quell'intervallo.

La media come miglior stima del valor vero

Se si fanno N misure di una grandezza x , normalmente distribuita, la miglior stima del valore vero è $\langle x \rangle$.

Perchè??

Usiamo il principio di massima verosimiglianza:

Siano N misure $x_1, \dots, x_N$, di una grandezza il cui valore vero è X , incognito.

Assumiamo che le N misure si distribuiscano secondo una distribuzione normale $f_{X,\sigma}(x)$. La probabilità di ottenere la i -ma misura è:

$$f_{X,\sigma}(x_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-X)^2/2\sigma^2}$$

dove X e σ sono incognite.

La probabilità di ottenere le tutte N misure sarà il prodotto delle N $f_{X,\sigma}(x_i)$:

$$f_{X,\sigma}(x_N) = \prod_{i=1}^N \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-X)^2/2\sigma^2}$$

quindi:

$$f_{X,s}(x_N) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\sum_{i=1}^N (x_i - X)^2 / 2\sigma^2\right)$$

Le miglior stime di X e σ sono quindi quei i valori che massimizzano la probabilità $f_{X,\sigma}(x_N)$.

Miglior stima di X : è il valore che minimizza il numeratore della funzione esponenziale:

$$\frac{d}{dX} \left(\sum_{i=1}^N (x_i - X)^2 \right) = 0 \quad \Rightarrow \quad -2 \sum_{i=1}^N (x_i - X) = 0$$

quindi

$$\sum_{i=1}^N (x_i - X) = \sum_{i=1}^N x_i - N X = 0$$

$$X = \frac{1}{N} \sum_{i=1}^N x_i = \langle x \rangle$$

c.v.d.

Possiamo procedere analogamente per trovare la miglior stima di σ :

Si dovrà derivare rispetto a σ , ed in questo caso la dipendenza di $f_{X,\sigma}(x_N)$ è più complicata. Da

$$\frac{d}{d\sigma} f_{X,\sigma}(x_N) = 0$$

si ottiene:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - X)^2}$$

Sostituendo a X il valore stimato $\langle x \rangle$ si commette una leggera sottostima del valore di σ .

Si può dimostrare che per ottenere la migliore stima di s usando il valore stimato $\langle x \rangle$ si deve moltiplicare il valore sopra ottenuto per

$$\sqrt{\frac{N}{N-1}} :$$

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \langle x \rangle)^2}$$

Giustificazione della somma in quadratura

Consideriamo solo il caso di due grandezze indipendenti x ed y con incertezze σ_x e σ_y . Vogliamo calcolare l'incertezza (la deviazione standard) associata alla loro somma. Per semplicità assumiamo che $\langle x \rangle = \langle y \rangle = 0$.

Sappiamo che

$$P(x) \propto \exp\left(-\frac{x^2}{2\sigma_x^2}\right)$$

$$P(y) \propto \exp\left(-\frac{y^2}{2\sigma_y^2}\right)$$

Vogliamo trovare qual è la probabilità di $x+y$.

Ovviamente la probabilità di trovare x ed y qualunque sarà il prodotto delle probabilità.

$$P(x,y) \propto \exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right]$$

utilizziamo l'identità (controllate!):

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} = \frac{(x+y)^2}{\sigma_x^2 + \sigma_y^2} + \frac{(\sigma_x^2 x - \sigma_y^2 y)^2}{\sigma_x^2 \sigma_y^2 (\sigma_x^2 + \sigma_y^2)}$$

dove l'ultimo membro non contiene $x+y$ e non è quindi di interesse. Possiamo scrivere:

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} = \frac{(x+y)^2}{\sigma_x^2 + \sigma_y^2} + z^2$$

e quindi:

$$P(x,y) \propto \exp\left[-\frac{1}{2}\left(\frac{(x+y)^2}{\sigma_x^2 + \sigma_y^2} + z^2\right)\right] = \exp\left[-\frac{(x+y)^2}{2(\sigma_x^2 + \sigma_y^2)}\right] \exp\left(-\frac{z^2}{2}\right)$$

ma a noi interessa la probabilità di ottenere $x+y$ indipendentemente dal valore di z . Possiamo quindi integrare su z ottenendo di nuovo il fattore $\sqrt{2\pi}$ che facciamo rientrare nel segno di proporzionalità:

$$P(x+y) \propto \exp\left[-\frac{(x+y)^2}{2(\sigma_x^2 + \sigma_y^2)}\right]$$

Vediamo quindi che la varianza di $x+y$ è

$$\sigma_{x+y}^2 = \sigma_x^2 + \sigma_y^2 \quad \text{c.v.d}$$

In generale

si dimostra che

$$q = x + a \quad \sigma_q = \sigma_x$$

$$q = x + y \quad \sigma_q = \sqrt{\sigma_x^2 + \sigma_y^2}$$

$$q = f(x,y) \quad \sigma_q = \sqrt{\left(\frac{\partial q}{\partial x} \sigma_x\right)^2 + \left(\frac{\partial q}{\partial y} \sigma_y\right)^2}$$

$$q = f(x_1, x_2, \dots, x_N) \quad \sigma_q = \sqrt{\left(\frac{\partial q}{\partial x_1} \sigma_{x_1}\right)^2 + \dots + \left(\frac{\partial q}{\partial x_N} \sigma_{x_N}\right)^2}$$

Errore standard della media

Qual è l'incertezza che si ha nella stima di $\langle x \rangle$?

Immaginiamo di compiere le nostre N misure molte volte, e di calcolare $\langle x \rangle$ ogni volta.

La serie di $\langle x \rangle$ calcolata sarà distribuita normalmente, centrata su X .

Vogliamo calcolare la larghezza di questa distribuzione.

Utilizziamo la propagazione degli errori:

$$\sigma_{\langle x \rangle} = \sqrt{\left(\frac{\partial \langle x \rangle}{\partial x_1} \sigma_{x_1} \right)^2 + \dots + \left(\frac{\partial \langle x \rangle}{\partial x_N} \sigma_{x_N} \right)^2}$$

nel nostro caso, applicandola alla media delle medie

$$\sigma_{x_1} = \sigma_{x_2} = \dots = \sigma_{x_N} \equiv \sigma_x$$

$$\text{ed anche} \quad (\text{da } \langle x \rangle = \frac{x_1 + x_2 + \dots + x_N}{N}):$$

$$\frac{\partial \langle x \rangle}{\partial x_1} = \frac{\partial \langle x \rangle}{\partial x_2} = \dots = \frac{\partial \langle x \rangle}{\partial x_N} = \frac{1}{N}$$

si trova:

$$\sigma_{\langle x \rangle} = \sqrt{N \left(\frac{1}{N} \sigma_x \right)^2} = \frac{\sigma_x}{\sqrt{N}}$$

$$\sigma_{\langle x \rangle} = \frac{\sigma_x}{\sqrt{N}}$$

che spesso è chiamato "errore standard della media".

Un ausilio di calcolo

Il calcolo delle variabili statistiche può essere fatto con un calcolatore tascabile pre-programmato, con un adeguato programma e un computer, ma anche utilizzando un normale calcolatore tascabile non programmabile, purché abbia alcune caratteristiche che ora mai hanno praticamente tutti i calcolatori

Questo deve essere in grado di calcolare man mano che vengono inseriti, la somma e la somma dei quadrati dei dati inseriti e, laddove si inseriscano coppie di valori, somma, somma del quadrato di entrambi i valori, e la somma del prodotto delle coppie.

Introduciamo la seguente notazione:

$$S_x = \sum_{i=1}^N x_i$$

In tal modo la media diviene:

$$\langle x \rangle = \frac{1}{N} S_x$$

Il calcolo della varianza (e quindi della deviazione standard) più sembrare più complesso, in quanto si deve prima calcolare la media e poi ri-inserire tutti i dati per calcolare le deviazioni quadrate da sommare. Con una semplice algebra il tutto si semplifica:

$$\begin{aligned} \sigma^2 &= \frac{1}{N-1} \sum_{i=1}^N (x_i - \langle x \rangle)^2 \\ &= \frac{N}{N-1} \frac{1}{N} \sum_{i=1}^N (x_i^2 + \langle x \rangle^2 - 2x_i \langle x \rangle) = \frac{N}{N-1} (\langle x^2 \rangle - \langle x \rangle^2) \end{aligned}$$

E quindi:

$$\sigma^2 = \frac{N}{N-1} \left[\frac{1}{N} S_{x^2} - \left(\frac{1}{N} S_x \right)^2 \right]$$

Dove ovviamente

$$\frac{N}{N-1} \approx 1$$

se N è abbastanza grande ($N > 10$)

Si noti quindi che è sufficiente inserire i nostri dati (gli $N x_i$) solo una volta. Al termine media e deviazione standard sono calcolati da S_x e S_{x^2} .

Si noti che questa procedura è assai utile anche quando si voglia programmare un calcolatore per calcolare $\langle x \rangle$ e σ_x da una serie di $N x_i$.

Confronto tra medie

Ora siamo bene attrezzati per confrontare quantitativamente due medie.

Assumiamo di misurare molte volte una grandezza di un sistema prima di un evento (o una modifica di un parametro del sistema) e molte volte dopo.

Ad esempio misuriamo, come descritto precedentemente, la quantità di moto prima e dopo un urto. Chiediamoci se le in questo caso la quantità di moto si è conservata.

Misuriamo:

$$\langle p \rangle \pm \sigma_{\langle p \rangle}$$

prima dell'urto e

$$\langle p' \rangle \pm \sigma_{\langle p' \rangle}$$

dopo l'urto.

Ricordiamo che

- le medie delle nostre misure sono la miglior stima del valore vero se tutti gli errori sistematici sono stati ridotti ad un livello trascurabile;
- il valore di σ calcolato è una buona stima della deviazione standard se il numero di misure è grande

Se il numero di misure fosse piccolo, dovremmo seguire una strada diversa, utilizzando la distribuzione del t-student.

Verifichiamo la compatibilità delle nostre misure con:

$$p'_v - p_v = 0$$

(dove p'_v e p_v sono i valori veri di p e p')

Ovviamente $\langle p \rangle - \langle p' \rangle$ non sarà zero, ma la discrepanza potrebbe essere dovuta alle sole fluttuazioni statistiche (casuali) associate alle misure.

Il rapporto "t"

$$t = \frac{\langle p' \rangle - \langle p \rangle}{\sigma_{\langle p' \rangle - \langle p \rangle}}$$

è il numero di deviazioni standard per cui $\langle p' \rangle$ differisce da $\langle p \rangle$.

Avendo assunto una distribuzione normale delle nostre misure, possiamo ora utilizzare la tabella dell'integrale normale degli errori (a ppendice A del Taylor)

Questa tabella ci fornisce, per ogni dato t , la probabilità $P(t)$ che una misura di x cada nell'intervallo $X - t\sigma < X < X + t\sigma$, per motivi "puramente" statistici, legati alle fluttuazioni casuali delle misure. Ovviamente $p = 1 - P(t)$ è la probabilità che una misura cada fuori dall'intervallo $x - t\sigma < x < x + t\sigma$, per motivi "puramente" statistici.

Se, come esempio, $t = 1.26$ la tabella ci dice che

$$P = 0.7923$$

cioè c'è una probabilità pari a

$$p = 1 - P = 0.2077$$

che possa esservi un risultato con una discrepanza maggiore di quella ottenuta, dovuto alle sole fluttuazioni statistiche.

In altri termini le nostre misure hanno prodotto una discrepanza che sole fluttuazioni statistiche avrebbero potuto produrre (di valore uguale o maggiore) con una probabilità del 21%.

Come consideriamo la discrepanza dei due valori di p e p' ?

Significativa (in questo caso la conservazione della quantità di moto non risulterebbe verificata) o no?

Questo è solitamente lasciato allo sperimentatore. Il limite tra l'accettabilità e l'inaccettabilità di una discrepanza come significativa dipende dal livello di probabilità al di sotto del quale giudichiamo che una discrepanza sia decisamente improbabile se dovuta solo a fluttuazioni statistiche.

Una convenzione spesso usata suggerisce una soglia del 5%. Cioè una discrepanza viene considerata "significativa" quando le fluttuazioni statistiche avrebbero potuto produrre la stessa discrepanza od una maggiore con una probabilità p pari al 5% od inferiore.

Accettando questa convenzione potremmo dire in questo caso che i due valori di p e p' sono compatibili con la conservazione della quantità di moto.

Dalla tabella notiamo che il valore di t corrispondente a $P = 95\%$ (cioè $p = 1 - P = 5\%$) è $t = 1.96$. Cioè la discrepanza deve essere maggiore di circa due volte la deviazione standard per essere significativa ($p < 0.05$).

Si noti che è opportuno indicare sempre quale soglia si è usata per definire la significatività.

Media pesata

Si supponga di avere M serie di misure di una stessa grandezza fisica, ciascuna con un valor medio $\langle m \rangle_i$ e deviazione standard dalla media σ_i .

Si definisce media pesata:

$$\langle m \rangle = \frac{\sum_{i=1}^M \frac{\langle m \rangle_i}{\sigma_i^2}}{\sum_{i=1}^M \frac{1}{\sigma_i^2}}$$

Per dimostrare questa formula si usa il principio della massima verosimiglianza.

Consideriamo solo due misure:

$$x_A \pm \sigma_A$$

$$x_B \pm \sigma_B$$

Assumiamo, ovviamente, che la discrepanza tra x_A e x_B non sia significativa (in altre parole che entrambe le misure siano state eseguite correttamente).

Assumiamo inoltre che le misure siano governate dalla distribuzione di Gauss

La probabilità di ottenere x_A o x_B è:

$$P(x_A) = \frac{1}{\sigma_A \sqrt{2\pi}} e^{-(x_A - X)^2 / 2\sigma_A^2}$$

$$P(x_B) = \frac{1}{\sigma_B \sqrt{2\pi}} e^{-(x_B - X)^2 / 2\sigma_B^2}$$

La probabilità che entrambe le misure siano ottenute è:

$$P(x_A, x_B) = \frac{1}{\sigma_A \sigma_B 2\pi} e^{-\chi^2 / 2}$$

dove

$$\chi^2 = \left(\frac{x_A - X}{\sigma_A} \right)^2 + \left(\frac{x_B - X}{\sigma_B} \right)^2$$

Per il principio della massima verosimiglianza la miglior stima di X è il valore per il quale $P(x_A, x_B)$ è massima cioè χ^2 è minimo:

$$\frac{d\chi^2}{dX} = -2 \frac{x_A - X}{\sigma_A^2} - 2 \frac{x_B - X}{\sigma_B^2} = 0$$

$$\begin{aligned} \frac{x_A}{\sigma_A^2} - \frac{X}{\sigma_A^2} + \frac{x_B}{\sigma_B^2} - \frac{X}{\sigma_B^2} &= \\ = \frac{x_A}{\sigma_A^2} + \frac{x_B}{\sigma_B^2} - X \left(\frac{1}{\sigma_B^2} + \frac{1}{\sigma_A^2} \right) &= 0 \end{aligned}$$

e quindi

$$X = \frac{\frac{x_A}{\sigma_A^2} + \frac{x_B}{\sigma_B^2}}{\frac{1}{\sigma_B^2} + \frac{1}{\sigma_A^2}}$$

c.v.d.

La deviazione standard della media pesata si può calcolare dalla propagazione degli errori statistici:

$$q = f(x_1, x_2, \dots, x_N) \rightarrow \sigma_q = \sqrt{\left(\frac{\partial \langle x \rangle}{\partial x_1} \sigma_{x_1}\right)^2 + \dots + \left(\frac{\partial \langle x \rangle}{\partial x_N} \sigma_{x_N}\right)^2}$$

$$\sigma_{\langle m \rangle} = \sqrt{\sum_{i=1}^M \left(\frac{\frac{1}{\sigma_i^2} \sigma_i}{\sum_{i=1}^M \frac{1}{\sigma_i^2}} \right)^2}$$

$$= \sqrt{\frac{\sum_{i=1}^M \frac{1}{\sigma_i^2}}{\left(\sum_{i=1}^M \frac{1}{\sigma_i^2}\right)^2}}$$

$$= \sqrt{\frac{1}{\sum_{i=1}^M \frac{1}{\sigma_i^2}}}$$

Possiamo quindi scrivere:

$$\langle m \rangle = \frac{\sum_{i=1}^M \frac{\langle m \rangle_i}{\sigma_i^2}}{\sum_{i=1}^M \frac{1}{\sigma_i^2}} \pm \sqrt{\frac{1}{\sum_{i=1}^M \frac{1}{\sigma_i^2}}}$$

Regressione lineare

Come vedremo uno dei problemi più interessanti è risalire da misure di due variabili fisiche alla relazione matematica fra le stesse.

La più semplice di queste relazioni è quella lineare.

Se misuriamo, ad esempio, posizione e tempo in un moto rettilineo uniforme, ci aspettiamo di trovare una relazione lineare fra queste due grandezze. Le stesse misure riportate su un grafico X vs t dovrebbero apparire come punti allineati su una retta.

Gli inevitabili errori di misura rendono ciò impossibile ed i punti non saranno "sulla" retta ma si distribuiranno a una distanza da questa compatibile con gli errori di misura.

Sorgono quindi alcune domande:

- qual è la "migliore" retta che "passa" per i punti?
- qual è la probabilità che fra le due variabili misurate sussista una relazione lineare?

Vi sono strumenti semplici e potenti per dare risposte quantitative a queste domande.

Cosideriamo due grandezze x e y .

Assumiamo:

- che queste siano connesse da una relazione lineare:

$$y_i = a + bx_i$$

- che le incertezze su x siano trascurabili.
- che le incertezze su tutte le y_i siano uguali, e che queste si distribuiscano gaussianamente con parametro di larghezza σ_y .

Se conoscessimo le costanti a e b , per ogni valore di x_i (che assumiamo privo di errore), potremmo calcolarci y_i :

$$y_{i(\text{valore vero})} = a + bx_i$$

Avendo assunto che le y_i sono distribuite gaussianamente con parametro di larghezza σ_y , possiamo affermare che la probabilità di ottenere il valore osservato y_i è:

$$P_{a,b}(y_i) \propto \frac{1}{\sigma_y} e^{-(y_i - a - bx_i)^2 / 2\sigma_y^2}$$

Dove i sottoscritti a e b stanno ad indicare che questi sono i nostri parametri incogniti da cui dipende P .

La probabilità di ottenere le N misure è il prodotto delle probabilità e quindi:

$$P_{a,b}(y_1, \dots, y_N) \propto \frac{1}{\sigma_y^N} e^{-\chi^2/2}$$

dove

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

Le migliori stime di a e b sono quelle che massimizzano P (minimizzano χ^2).

Quindi:

$$\frac{\partial \chi^2}{\partial a} = -\frac{2}{\sigma_y^2} \sum_{i=1}^N (y_i - a - bx_i) = 0$$

$$\frac{\partial \chi^2}{\partial b} = -\frac{2}{\sigma_y^2} \sum_{i=1}^N x_i (y_i - a - bx_i) = 0$$

da cui

$$Na + b \sum_{i=1}^N x_i = \sum_{i=1}^N y_i$$

$$a \sum_{i=1}^N x_i + b \sum_{i=1}^N x_i^2 = \sum_{i=1}^N x_i y_i$$

Con la nostra notazione

$$Na + S_x b = S_y$$

$$S_x a + S_{x^2} b = S_{xy}$$

da cui:

$$a = \frac{\begin{vmatrix} S_y & S_x \\ S_{xy} & S_{x^2} \end{vmatrix}}{\begin{vmatrix} N & S_x \\ S_x & S_{x^2} \end{vmatrix}} \quad b = \frac{\begin{vmatrix} N & S_y \\ S_x & S_{xy} \end{vmatrix}}{\begin{vmatrix} N & S_x \\ S_x & S_{x^2} \end{vmatrix}}$$

$$a = \frac{S_y S_{x^2} - S_x S_{xy}}{N S_{x^2} - S_x^2}$$

$$b = \frac{N S_{xy} - S_y S_x}{N S_{x^2} - S_x^2}$$

Queste sono le nostre migliori stime per le costanti a e b .

Si può stimare l'incertezza delle y_i considerando la loro distribuzione attorno al valore vero $a + bx_i$. Le deviazioni

$$y_i - a - bx_i$$

saranno normalmente distribuite con valore medio 0 e larghezza σ_y .

La miglior stima di σ_y è data da:

$$\sigma_y^2 \equiv \sigma^2 = \frac{1}{N-2} \sum_{i=1}^N (y_i - a - bx_i)^2$$

Come si può verificare usando il principio di massima verosimiglianza:

$$\frac{\partial}{\partial \sigma} P_{ab}(y_1 \dots y_N) \Rightarrow \frac{\partial}{\partial \sigma} \left(\frac{1}{\sigma^N} e^{-\chi^2/2} \right) = 0$$

$$\frac{\partial}{\partial \sigma} P_{ab}(y_1 \dots y_N) = -\frac{N}{\sigma^{N+1}} e^{-\chi^2/2} + \frac{1}{\sigma^N} e^{-\chi^2/2} \frac{\partial}{\partial \sigma} \left(-\frac{\chi^2}{2} \right)$$

$$\begin{aligned} \frac{\partial}{\partial \sigma} \left(-\frac{\chi^2}{2} \right) &= \frac{\partial}{\partial \sigma} \frac{-1}{2\sigma_y^2} \left(\sum_{i=1}^N (y_i - a - bx_i)^2 \right) = \\ &= \frac{1}{\sigma_y^3} \left(\sum_{i=1}^N (y_i - a - bx_i)^2 \right) \end{aligned}$$

quindi

$$\frac{\partial}{\partial \sigma} P_{ab}(y_1 \dots y_N) =$$

$$= -\frac{N}{\sigma^{N+1}} e^{-\chi^2/2} + \frac{1}{\sigma^N} e^{-\chi^2/2} \frac{1}{\sigma^3} \left(\sum_{i=1}^N (y_i - a - bx_i)^2 \right) = 0$$

dividendo per $\frac{e^{-\chi^2/2}}{\sigma^{N+1}}$ si ottiene

$$= -N + \frac{1}{\sigma^2} \left(\sum_{i=1}^N (y_i - a - bx_i)^2 \right) = 0$$

e quindi

$$\sigma_y^2 = \frac{1}{N} \sum_{i=1}^N (y_i - a - bx_i)^2$$

A questo punto occorre ricordare che $a + bx_i$ è il valore vero se a e b sono i valori veri dei parametri. Noi, invece, non conosciamo tali valori veri e dobbiamo usare le nostre migliori stime. Inserite nella equazione precedente tali stime riducono il valore di σ_y in quanto sono state calcolate proprio come quei valori che minimizzano la somma delle deviazioni quadrate. Si può dimostrare che questa riduzione è compensata inserendo al posto del fattore N nel denominatore con $N-2$ ottenendo la relazione citata all'inizio:

$$\sigma_y^2 \equiv \sigma^2 = \frac{1}{N-2} \sum_{i=1}^N (y_i - a - bx_i)^2$$

Si noti che $N-2$ corrispondono ai "gradi di libertà", in questo caso ad N meno il numero di "vincoli" cioè di parametri calcolati dai dati stessi. Si noti inoltre che con la stessa logica si spiega la presenza del fattore $N-1$ a denominatore della varianza, dove si è dovuto calcolare un solo parametro dai dati (la media che si è utilizzata al posto del valor vero, incognito)

L'incertezza su a e b si calcola facilmente con la propagazione degli errori usando i valori trovati in precedenza:

$$a = \frac{S_y S_{x^2} - S_x S_{xy}}{N S_{x^2} - S_x^2}$$

$$b = \frac{N S_{xy} - S_y S_x}{N S_{x^2} - S_x^2}$$

$$\sigma_a^2 = \sum_{i=1}^N \left(\frac{\partial a}{\partial y_i} \right)^2 \sigma^2$$

$$\sigma_b^2 = \sum_{i=1}^N \left(\frac{\partial b}{\partial y_i} \right)^2 \sigma^2$$

quindi

$$\begin{aligned} \sigma_a^2 &= \sum_{i=1}^N \left(\frac{S_{x^2} - x_i S_x}{N S_{x^2} - S_x^2} \right)^2 \sigma^2 \\ &= \frac{N S_{x^2}^2 + S_x^2 S_x^2 - 2 S_x^2 S_{x^2}}{(N S_{x^2} - S_x^2)^2} \sigma^2 = \\ &= S_{x^2} \frac{N S_{x^2} + S_x^2 - 2 S_x^2}{(N S_{x^2} - S_x^2)^2} \sigma^2 = \\ &= S_{x^2} \frac{N S_{x^2} - S_x^2}{(N S_{x^2} - S_x^2)^2} \sigma^2 = \\ &= \frac{S_{x^2}}{N S_{x^2} - S_x^2} \sigma^2 \end{aligned}$$

analogamente:

$$\begin{aligned} \sigma_b^2 &= \sum_{i=1}^N \left(\frac{N x_i - S_x}{N S_{x^2} - S_x^2} \right)^2 \sigma^2 = \\ &= \frac{N^2 S_{x^2} + N S_x^2 - 2 N S_x^2}{(N S_{x^2} - S_x^2)^2} \sigma^2 = \end{aligned}$$

$$= N \frac{NS_{x^2} - S_x^2}{(NS_{x^2} - S_x^2)^2} \sigma^2 =$$

$$= \frac{N}{NS_{x^2} - S_x^2} \sigma^2$$

Ricapitolando:

$$y_i = a + bx_i$$

$$a \pm \sigma_a = \frac{S_y S_{x^2} - S_x S_{xy}}{\Delta} \pm \sqrt{\frac{S_{x^2}}{\Delta}} \sigma$$

$$b \pm \sigma_b = \frac{NS_{xy} - S_y S_x}{\Delta} \pm \sqrt{\frac{N}{\Delta}} \sigma$$

$$\Delta = NS_{x^2} - S_x^2$$

$$\sigma = \sqrt{\frac{1}{N-2} \sum_{i=1}^N (y_i - a - bx_i)^2}$$

Altre funzioni con il metodo dei minimi quadrati

Altre relazioni con due variabili due parametri possono essere ricondotte al caso lineare. Relazioni a più parametri (ad esempio relazioni polinomiali) invece, necessitano di formule analoghe a quanto visto sin'ora, ma diverse e, ovviamente, più lunghe, che qui non tratteremo.

Lo stesso dicasi per relazioni con più di due variabili (regressioni multiple), come ad esempio $y = a + bx + cz$.

Menzioniamo solo due delle relazioni che possono essere ricondotte al caso lineare studiato.

relazione di potenza

$$Y = AX^B$$

In questo caso è sufficiente prendere il logaritmo di entrambi i membri:

$$\log Y = \log A + B \log X$$

e quindi possiamo usare le formule trovate prima con

$$\begin{array}{ll} y = \log Y & a = \log A \\ x = \log X & b = B \end{array}$$

relazione esponenziale

$$Y = Ae^{BX}$$

anche in questo caso, prendendo il logaritmo di entrambi i membri:

$$\log Y = \log A + BX$$

ci riconduciamo alla formula lineare con

$$\begin{array}{ll} y = \log Y & a = \log A \\ x = X & b = B \end{array}$$

PROPAGAZIONE DEGLI ERRORI 2

Assumiamo di misurare x ed y e di voler calcolare una grandezza $q(x,y)$. Abbiamo visto che, intuitivamente,

$$i) \langle q \rangle = q(\langle x \rangle, \langle y \rangle)$$

inoltre abbiamo visto che, se x ed y sono gaussiane ed i loro errori indipendenti

$$ii) \sigma_q^2 = \left(\frac{\partial q}{\partial x} \right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y} \right)^2 \sigma_y^2$$

In precedenza avevamo introdotto una propagazione degli errori (che avevamo chiamato "massimi") che si scriveva:

$$iii) \delta q = \left| \frac{\partial q}{\partial x} \right| \delta_x + \left| \frac{\partial q}{\partial y} \right| \delta_y$$

Senza dimostrarlo avevamo detto che la precedente costituisce un limite massimo all'errore propagato.

Ora cerchiamo di dare un fondamento a tali affermazioni senza neanche invocare la normalità di x e y .

Assumiamo solo che le incertezze nelle nostre misure x_i ed y_i siano piccole.

Possiamo quindi utilizzare l'espansione in serie di Taylor per una funzione a due variabili attorno a $\langle x \rangle$, $\langle y \rangle$, al primo ordine:

$$q_i = q(x_i, y_i) \approx q(\langle x \rangle, \langle y \rangle) + \frac{\partial q}{\partial x} (x_i - \langle x \rangle) + \frac{\partial q}{\partial y} (y_i - \langle y \rangle)$$

dove le derivate sono calcolate in $\langle x \rangle$ ed $\langle y \rangle$.

Posso quindi calcolare la media di q :

$$\langle q \rangle = \frac{1}{N} \sum_{i=1}^N q_i$$

$$\langle q \rangle = \frac{1}{N} \sum_{i=1}^N \left[q(\langle x \rangle, \langle y \rangle) + \frac{\partial q}{\partial x} (x_i - \langle x \rangle) + \frac{\partial q}{\partial y} (y_i - \langle y \rangle) \right]$$

Il secondo ed il terzo termine in parentesi sono nulli.

Troviamo quindi il risultato (i):

$$\langle q \rangle = q(\langle x \rangle, \langle y \rangle)$$

La varianza (usiamo il denominatore N , per semplicità) è:

$$\begin{aligned} \sigma_q^2 &= \frac{1}{N} \sum_{i=1}^N (q_i - \langle q \rangle)^2 = \\ &= \frac{1}{N} \sum_{i=1}^N \left[q(\langle x \rangle, \langle y \rangle) + \frac{\partial q}{\partial x} (x_i - \langle x \rangle) + \frac{\partial q}{\partial y} (y_i - \langle y \rangle) - q(\langle x \rangle, \langle y \rangle) \right]^2 = \\ &= \frac{1}{N} \sum_{i=1}^N \left[\frac{\partial q}{\partial x} (x_i - \langle x \rangle) + \frac{\partial q}{\partial y} (y_i - \langle y \rangle) \right]^2 = \\ &= \left(\frac{\partial q}{\partial x} \right)^2 \frac{1}{N} \sum_{i=1}^N (x_i - \langle x \rangle)^2 + \left(\frac{\partial q}{\partial y} \right)^2 \frac{1}{N} \sum_{i=1}^N (y_i - \langle y \rangle)^2 + \\ &\quad + 2 \frac{\partial q}{\partial x} \frac{\partial q}{\partial y} \frac{1}{N} \sum_{i=1}^N (x_i - \langle x \rangle)(y_i - \langle y \rangle) \end{aligned}$$

Che possiamo scrivere:

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x} \right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y} \right)^2 \sigma_y^2 + 2 \frac{\partial q}{\partial x} \frac{\partial q}{\partial y} \sigma_{xy}$$

dove abbiamo introdotto la "covarianza" σ_{xy}

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

Se le due grandezze fisiche x e y non sono indipendenti vi sarà probabilità non nulla che le deviazioni di x da $\langle x \rangle$ abbiano un andamento "co-variante" con le deviazioni di y da $\langle y \rangle$, rendendo la covarianza non nulla. Viceversa se le misure di x ed y sono indipendenti ed N è grande, i contributi delle diverse coppie avranno segni casuali e quindi σ_{xy} tenderà a zero. In questo caso si ritrova il risultato (ii) menzionato precedentemente per grandezze indipendenti.

Si noti che:

σ_{xy} può essere negativa

$\sigma_{xx} = \sigma_x^2$ (e, in questo caso è sempre positiva).

Inoltre si dimostra che la covarianza soddisfa la "disuguaglianza di Schwartz"

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

Inserendo questa disuguaglianza in σ_q :

$$\begin{aligned} \sigma_q^2 &\leq \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y^2 + 2 \left|\frac{\partial q}{\partial x} \frac{\partial q}{\partial y}\right| \sigma_x \sigma_y \\ \sigma_q^2 &\leq \left\{ \left|\frac{\partial q}{\partial x}\right| \sigma_x + \left|\frac{\partial q}{\partial y}\right| \sigma_y \right\}^2 \\ \sigma_q &\leq \left|\frac{\partial q}{\partial x}\right| \sigma_x + \left|\frac{\partial q}{\partial y}\right| \sigma_y \end{aligned}$$

Abbiamo in questo modo dato un preciso significato al nostro risultato (iii)

$$\delta_q \leq \left|\frac{\partial q}{\partial x}\right| \delta_x + \left|\frac{\partial q}{\partial y}\right| \delta_y$$

che rappresenta realmente un limite superiore dell'incertezza propagata.

Coefficiente di correlazione

La covarianza descrive, quindi, quanto un gruppo di misure di due variabili indichi una relazione lineare fra loro.

Il suo valore numerico, comunque, dipende, ovviamente, da quanto i valori misurati sono grandi o piccoli. Per ottenere un descrittore "normalizzato" di relazione lineare, che non dipenda cioè dalla grandezza dei valori misurati, si può dividere per le deviazioni standard delle due variabili, ottenendo il "coefficiente di correlazione lineare r "

$$\begin{aligned} r &= \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \\ &= \frac{\sum_{i=1}^N (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\sum_{i=1}^N (x_i - \langle x \rangle)^2 \sum_{i=1}^N (y_i - \langle y \rangle)^2}} \end{aligned}$$

Il coefficiente di correlazione è un descrittore normalizzato adimensionale di quanto le variazioni di x ed y siano correlate.
 $-1 \leq r \leq 1$

Infatti se x ed y fossero completamente scorrelate (indipendenti), data una deviazione ($x_i - \langle x \rangle$) vi sarà eguale probabilità di una deviazione ($y_i - \langle y \rangle$) positiva o negativa, e viceversa. Se il numero delle misure è elevato la somma delle coppie delle deviazioni tenderà a zero.

Viceversa se x ed y fossero completamente correlate, ad una deviazione positiva (negativa) ($x_i - \langle x \rangle$) corrisponderebbe sempre una deviazione dello stesso segno (o, equivalentemente, sempre del segno opposto) portando il coefficiente r ad un valore pari a 1 (o, equivalentemente, -1).

Aiuti per il calcolo

Per calcolare r è sembrerebbe necessario calcolare le deviazioni individuali $(x_i - \langle x \rangle)$ e $(y_i - \langle y \rangle)$, e quindi, prima di tutto le medie $\langle x \rangle$ e $\langle y \rangle$.

Ciò non è vero, come già visto per la varianza e per i coefficienti della regressione lineare.

$$\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \left(\frac{1}{N} \sum_{i=1}^N x_i \right)^2 = \frac{1}{N} S_{x^2} - \frac{1}{N^2} S_x^2$$

analogamente (si ricorda che $S_a = \sum_{i=1}^N a_i$)

$$\sigma_y^2 = \frac{1}{N} S_{y^2} - \frac{1}{N^2} S_y^2$$

inoltre:

$$\begin{aligned} \sigma_{xy} &= \frac{1}{N} \sum_{i=1}^N (x_i - \langle x \rangle)(y_i - \langle y \rangle) = \\ &= \frac{1}{N} \sum_{i=1}^N (x_i y_i + \langle x \rangle \langle y \rangle - x_i \langle y \rangle - y_i \langle x \rangle) = \\ &= \langle xy \rangle + \langle x \rangle \langle y \rangle - \langle x \rangle \langle y \rangle - \langle x \rangle \langle y \rangle = \langle xy \rangle - \langle x \rangle \langle y \rangle = \\ &= \frac{1}{N} S_{xy} - \frac{1}{N^2} S_x S_y \end{aligned}$$

Da cui

$$\begin{aligned} r &= \frac{\sigma_{xy}}{\sigma_x \sigma_y} \\ &= \frac{\frac{1}{N} S_{xy} - \frac{1}{N^2} S_x S_y}{\sqrt{\left(\frac{1}{N} S_{x^2} - \frac{1}{N^2} S_x^2 \right) \left(\frac{1}{N} S_{y^2} - \frac{1}{N^2} S_y^2 \right)}} \\ &= \frac{NS_{xy} - S_x S_y}{\sqrt{(NS_{x^2} - S_x^2)(NS_{y^2} - S_y^2)}} \end{aligned}$$

che ci permette di calcolare r inserendo i nostri N dati una sola volta e raccogliendo le somme di x , y , x^2 , y^2 e xy .

Il significato statistico di r

Lo studio di r ci fornisce informazioni su quanto le due variabili x ed y siano correlate.

E' possibile formulare una risposta quantitativa in termini statistici in merito a tale problema.

Assumiamo di aver trovato un coefficiente di correlazione r_0 . La probabilità di ottenere "per caso" un r "migliore", cioè tale che $|r| \geq |r_0|$ si trova sulle tabelle di probabilità del coefficiente di correlazione lineare (Appendice C del Taylor). Se questa probabilità è grande le nostre misure sono probabilmente poco correlate, viceversa la loro correlazione sarebbe significativa. Come già visto in precedenza una soglia che viene spesso scelta è $P(|r| \geq |r_0|) \leq 0.05$

Supponiamo di avere ottenuto le seguenti coppie di risultati (unità arbitrarie)

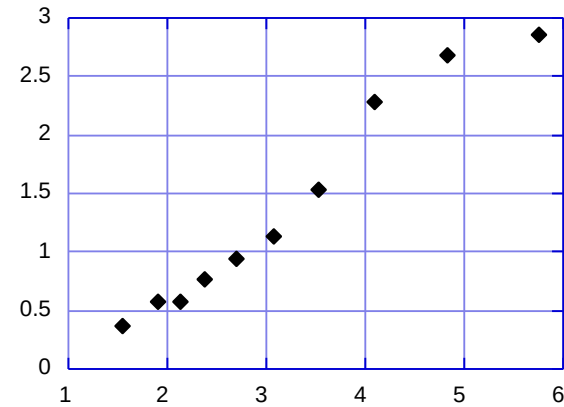
x	y
5.75	2.86
4.83	2.68
4.10	2.28
3.53	1.54
3.07	1.14
2.70	0.943
2.38	0.770
2.13	0.571
1.90	0.571
1.55	0.372

da cui posso ricavare:

N	S_x	S_y	S_{x^2}	S_{y^2}	S_{xy}
10	31.94	13.727	118.59	26.504	54.93

da cui si ricava, dalla formula calcolata sopra: $r = 0.9839$

Il grafico dei dati è il seguente



Che interpretazione si deve dare a questo risultato? Quanto sono correlate x e y ?

La probabilità che, pur essendo le due variabili totalmente indipendenti (scorrelate), $|r|$ sia maggiore di un valore finito pari a r_0 è:

N	r_0								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3	94	87	81	74	67	59	51	41	29
8	81	63	47	33	21	12	5.3	1.7	0.2
10	78	58	40	25	14	7	2	0.5	-
20	67	40	20	8	2	0.5	0.1	-	-
50	49	16	3	0.4	-	-	-	-	-

(- $\equiv \leq 0.05$)

Si vede quindi che nel nostro caso la probabilità che x e y siano scorrelate è decisamente trascurabile, e che possiamo sostenere che sono fra loro correlate ($P < 0.05$)

CENNI SULLA DISTRIBUZIONE BINOMIALE

Se la probabilità di avere un singolo evento "favorevole" è p , la probabilità che non avvenga è $1-p$. La probabilità di avere una qualsiasi sequenza di m eventi "favorevoli" (indipendentemente dall'ordine) su un totale di N è data da:

$$P_{Np}(m) = \frac{N!}{(N-m)!m!} p^m (1-p)^{N-m}$$

(distribuzione binomiale)

La distribuzione binomiale è generalmente non simmetrica (a meno che sia $p = 1/2$) e generalmente il valore medio non coincide col valore più probabile.

Calcoliamo $\langle m \rangle$ e σ^2 .

$$\begin{aligned} \langle m \rangle &= \sum_{m=0}^N m P_{Np}(m) = \sum_{m=0}^N m \frac{N!}{(N-m)!m!} p^m (1-p)^{N-m} \\ &= Np \sum_{m=1}^N \frac{(N-1)!}{(N-m)!(m-1)!} p^{m-1} (1-p)^{N-m} = \end{aligned}$$

poniamo:

$$m-1 = r$$

$$N-1 = N$$

$$\langle m \rangle = Np \sum_{r=0}^N \frac{N!}{(N-r)!r!} p^r (1-p)^{N-r}$$

$$\langle m \rangle = Np \sum_{r=0}^N P_{N,p}(r) = Np$$

$$\Rightarrow \langle m \rangle = Np$$

Dove si è usato la relazione

$$\sum_{r=0}^N P_{N,p}(r) = 1$$

{In modo analogo si dimostra che $\sigma^2 = Np(1-p)$ }

Per N grande la distribuzione binomiale tende a quella di Gauss con stesse deviazione standard e valor medio.

La distribuzione binomiale, di Gauss, e gli errori casuali

Giustificiamo il fatto che la distribuzione di Gauss mi descriva una misura affetta da molti errori casuali.

Assumiamo che molte sorgenti di errore contribuiscano con errori della stessa (piccola) dimensione ϵ , e che questi siano con eguale probabilità positivi o negativi.

Il numero di queste sorgenti sia n .

Se, per una determinata misura x_i , v di queste sorgenti danno contributo positivo (e, conseguentemente $n-v$ negativo), avremo:

$$x_i = X + v\epsilon - (n - v)\epsilon$$

La probabilità di avere v errori positivi su n è la probabilità binomiale $P_{n,1/2}(v)$.

I risultati di questa serie di misure sono distribuiti simmetricamente attorno al valore vero X con probabilità date da $P_{n,1/2}(v)$.

Quando n è grande (ed ϵ tende a zero in modo da avere una distribuzione continua di valori) $P_{n,1/2}(v)$ tende alla distribuzione di Gauss

CENNI SULLA DISTRIBUZIONE DI POISSON

- è il limite dalla distribuzione binomiale per

i) $p \ll 1$

ii) $Np = \langle m \rangle$ finito (cioè $N \gg 1$)

Calcoliamo questo limite:

$$P(N, m) = \frac{N!}{(N-m)!m!} p^m (1-p)^{N-m}$$

considerando (i) ed (ii) [$N \gg m$]:

$$\frac{N!}{(N-m)!} \approx N^m$$

quindi:

$$P(N, m) = \frac{(Np)^m}{m!} (1-p)^{N-m}$$

inoltre

$$(1-p)^{N-m} \approx (1-p)^N = (1-p)^{\langle m \rangle / p}$$

quindi

$$P(m) = \frac{\langle m \rangle^m}{m!} (1-p)^{\langle m \rangle / p}$$

ricordando infine:

$$\lim_{p \rightarrow 0} (1-p)^{1/p} = e^{-1}$$

$$\rightarrow P(m) = \frac{\langle m \rangle^m}{m!} e^{-\langle m \rangle}$$

E' la probabilità di avere m eventi favorevoli (indipendentemente dall'ordine) ciascuno con probabilità p molto piccola) su un totale di un numero molto grande di eventi possibili.

Esempio: il contatore Geiger.

Contiamo il numero di elettroni n emessi in un minuto per decadimento radioattivo. Non ci sarà incertezza su n , tuttavia ripetendo la misura n varierà di volta in volta. Questa variabilità riflette una proprietà intrinseca del processo di decadimento. Ogni nucleo ha infatti una probabilità definita di decadere p in un intervallo di un minuto. La probabilità che avvengano v decadimenti al minuto, con n "eventi possibili" (nuclei) è la probabilità binomiale $b_{np}(v)$. Essendo p molto piccola ed numero di "eventi possibili" (cioè di nuclei) è molto grande ($\approx 10^{20}$) la probabilità è ben descritta dalla distribuzione di Poisson $P_{\langle m \rangle}(m)$.

La condizione di normalizzazione è:

$$\sum_{m=0}^{\infty} P_{\langle m \rangle}(m) = 1$$

la media:

$$\sum_{m=0}^{\infty} m P(m) = \langle m \rangle$$

infine la varianza

$$\sigma^2 = \sum_{m=0}^{\infty} (m - \langle m \rangle)^2 P(m) = \langle m \rangle$$

Ciò significa che ad un conteggio di N eventi in un dato intervallo di tempo, si associa una deviazione standard pari a $\sqrt{N}$.

Infine se $\langle m \rangle = Np$ grande allora $P(m) \rightarrow f_{\langle m \rangle, \sqrt{\langle m \rangle}}(x_i)$

TEST DEL χ^2

Test per una distribuzione

Come possiamo stabilire se i dati in nostro possesso siano consistenti con una distribuzione teorica nota?

Supponiamo di fare 40 misure ($x_1, x_2, \dots, x_{40}$) della lunghezza di un campo (in m):

731 772 771 681 722 688 653 757 733 742
 739 780 709 676 760 748 672 687 766 645
 678 748 689 810 905 778 764 753 709 675
 698 770 754 830 725 710 738 638 787 712

Ritenendo che queste misure seguano una distribuzione normale:

$$\text{miglior stima di } x: \langle x \rangle = \frac{1}{N} \sum_{i=1}^N x_i = 730.1 \text{ m}$$

$$\text{miglior stima di } \sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \langle x \rangle)^2} = 46.8 \text{ m}$$

Quindi (si ricorda che $\sigma_{\langle x \rangle} = \sigma / \sqrt{N}$):

$$x = 730 \pm 7 \text{ m}$$

La nostra ipotesi di normalità è corretta?

Confrontiamo le nostre misure con ciò che avremmo dovuto ottenere.

La funzione $f_{\langle x \rangle, \sigma}(x)$ è continua, suddividiamo quindi le misure, ad esempio, in quattro intervalli. O_k siano i valori osservati nel k mo intervallo P_k la probabilità teorica che una misura cada nel k mo intervallo e quindi NP_k il "valore aspettato" (E_k) cioè il numero "teorico" di misure che dovrebbe cadere nel k mo intervallo.

k	1	2	3	4
	$x < \langle x \rangle - \sigma$;	$\langle x \rangle - \sigma < x < \langle x \rangle$;	$\langle x \rangle < x < \langle x \rangle + \sigma$;	$x > \langle x \rangle + \sigma$
O_k	8	10	16	6
P_k	0.16	0.34	0.34	0.16
NP_k	6.4	13.6	13.6	6.4
$O_k - NP_k$	1.6	-3.6	2.4	0.4

Come si può dire queste deviazioni siano "piccole" o "grandi"?

Ripetendo molte volte le 40 misure il numero O_k di misure nel k mo intervallo è equivalente ad un esperimento di conteggio. Possiamo quindi considerare le fluttuazioni "teoriche" di O_k , alle quali paragonare le nostre deviazioni osservate, pari alla deviazione standard di un processo poissoniano: $E_k = \sqrt{NP_k}$:

k	1	2	3	4
$\sqrt{E_k}$	2.5	3.7	3.7	2.5

Si definisce χ^2 :

$$\chi^2 = \sum_{k=1}^n \frac{(O_k - NP_k)^2}{NP_k}$$

Nel nostro caso $\chi^2 = 1.8$.

Si potrebbe ragionevolmente pensare che, se $\chi^2/n \leq 1$, non si dovrebbero avere motivi per mettere in dubbio la nostra ipotesi ma

.... è invece corretto confrontare il χ^2 con il numero di gradi di libertà "d".

Nel caso precedente i "vincoli" sono:

$$N (= \sum_{k=1}^n O_k) \quad \langle x \rangle \quad \sigma$$

cioè:

$$d = n - 3 = 1$$

(questo ci dice che nel nostro caso gli intervalli devono essere almeno quattro)

NOTA: È importante la cura nello scegliere il numero degli intervalli

Si può dimostrare che il valore medio aspettato di χ^2 è d

Il nostro valore di χ^2 (=1.8) è sufficientemente maggiore di uno da escludere che la nostra distribuzione sia gaussiana?

Si trova dalle tavole che, se la distribuzione fosse gaussiana, la probabilità di avere un disaccordo uguale o peggiore di quello trovato per le sole fluttuazioni statistiche è:

$$P_d(\chi^2 / d \geq 1.8) = 0.18$$

In generale spesso si dice che c'è un disaccordo significativo se

$$P_d(\chi^2 / d > \chi_0^2 / d) < 0.05$$

Dove la presenza del sottoscritto d mette in evidenza la dipendenza di P dal numero dei gradi di libertà.

Nel nostro caso quindi non abbiamo motivo sulla base di quest'analisi di rigettare l'ipotesi che la distribuzione sia normale.

Per il computo delle probabilità si usano le tavole , ad esempio:

d	χ_0^2 / d					
	0	0.5	1	1.5	2	4
1	1	0.48	0.32	0.22	0.16	0.05
2	1	0.61	0.37	0.22	0.14	0.02
5	1	0.78	0.42	0.19	0.08	0.001

Test per un modello

Possiamo utilizzare il test del χ^2 anche per stabilire se dei dati sperimentali siano compatibili con un modello o meno.

Un metodo di procedere è confrontare le occorrenze sperimentali con quelle predette dal modello basandosi su una variabilità gaussiana descritta da una distribuzione con media nulla e con σ pari a quello calcolato dalle misure.

Misuriamo un evento che ha una dinamica temporale non nota. Ad esempio la posizione di un oggetto ad istanti diversi:

t (s)	x (m)
3	30
4	37
5	45
6	53
7	64
8	70
9	78
10	88
11	96
12	103
13	110
14	121
15	129
16	135
17	142
18	150
19	156
20	165
21	172
22	178
23	186
24	192
25	201
26	206
27	211

Supponiamo di voler confrontare queste misure con due modelli:

moto rettilineo uniforme:

$$x = x_0 + vt$$

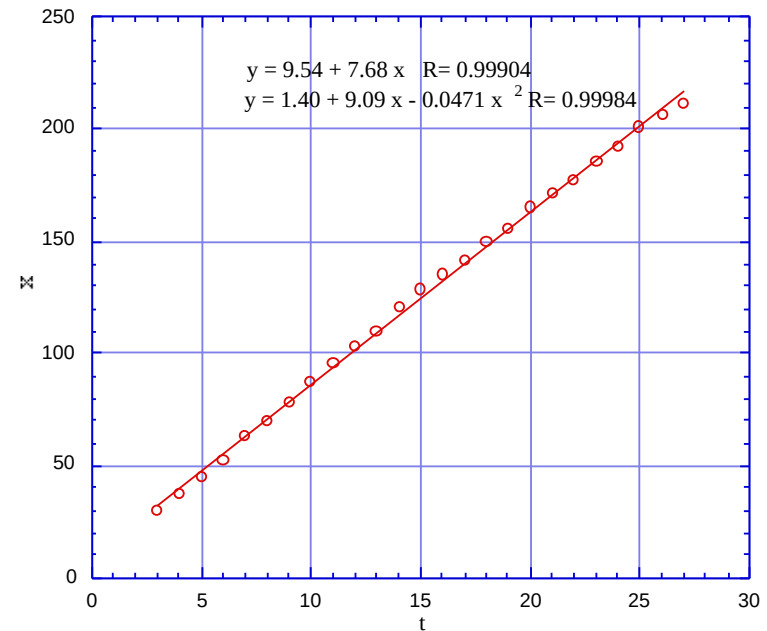
moto uniformemente accelerato: $x = x_0 + vt + \frac{1}{2} a t^2$

La domanda che ci poniamo è la seguente:

- la discrepanza osservata tra i valori sperimentali e quelli predetti dal modello è significativa o meno? Oppure, in altre parole, tale discrepanza è spiegabile con le sole fluttuazioni statistiche?

Riportiamo tutto su un grafico e procediamo al calcolo dei coefficienti " x_0 " e " v " nei due casi (che non daranno risultati uguali) ed " a ":

(Nota: per il calcolo dei coefficienti del moto accelerato servono le formule per la regressione polinomiale che non abbiamo visto, ma che si trovano nel Taylor cap. 8.6)



dove nel grafico si è riportata solo la migliore retta e non la migliore parabola, per chiarezza.

Si possono quindi calcolare le deviazioni e le deviazioni quadrate tra le misure sperimentali $x(t_i)$ e i corrispondenti valori forniti dai due modelli:

t (s)	x (m)	lin (m)	quad (m)	(lin-sp) (m)	(quad-sp) (m)	(lin-sp) ² (m ²)	(quad-sp) ² (m ²)
3	30	32.579	28.247	2.5786	-1.7533	6.6492	3.0742
4	37	40.257	37.008	3.2571	0.0079727	10.609	6.3564e-05
5	45	47.936	45.675	2.9356	0.67513	8.6177	0.45579
6	53	55.614	54.248	2.6141	1.2481	6.8335	1.5578
7	64	63.293	62.727	-0.70740	-1.2731	0.50042	1.6207
8	70	70.971	71.112	0.97110	1.1116	0.94303	1.2356
9	78	78.650	79.402	0.64960	1.4021	0.42198	1.9658
10	88	86.328	87.598	-1.6719	-0.40160	2.7952	0.16129
11	96	94.007	95.701	-1.9934	-0.29945	3.9736	0.089668
12	103	101.69	103.71	-1.3149	0.70855	1.7290	0.50204
13	110	109.36	111.62	-0.63640	1.6224	0.40500	2.6321
14	121	117.04	119.44	-3.9579	-1.5580	15.665	2.4273
15	129	124.72	127.17	-4.2794	-1.8325	18.313	3.3580
16	135	132.40	134.80	-2.6009	-0.20114	6.7647	0.040458
17	142	140.08	142.34	-1.9224	0.33601	3.6956	0.11291
18	150	147.76	149.78	-2.2439	-0.22099	5.0351	0.048838
19	156	155.43	157.13	-0.56540	1.1278	0.31968	1.2720
20	165	163.11	164.38	-1.8869	-0.61749	3.5604	0.38130
21	172	170.79	171.54	-1.2084	-0.45700	1.4602	0.20885
22	178	178.47	178.61	0.47009	0.60933	0.22099	0.37128
23	186	186.15	185.58	0.14861	-0.41850	0.022084	0.17514
24	192	193.83	192.46	1.8271	0.45949	3.3383	0.21113
25	201	201.51	199.24	0.50560	-1.7567	0.25563	3.0859
26	206	209.18	205.93	3.1841	-0.067001	10.138	0.0044892
27	211	216.86	212.53	5.8626	1.5285	34.370	2.3363
		SOMMA				146.64	27.329

La distribuzione delle deviazioni si può riportare su un grafico scegliendo opportunamente l'intervallo Δ .

Scegliamo

$$\Delta(\text{lin}) = 1 \text{ m}$$

$$\Delta(\text{quad}) = 0.5 \text{ m:}$$

modello lineare		modello quadratico	
D_k (m)	$O(D_k)$	D_k (m)	$O(D_k)$
-7	0	-3.5	0
-6	0	-3.0	0
-5	0	-2.5	0
-4	2	-2.0	3
-3	1	-1.5	2
-2	5	-1.0	0
-1	5	-0.5	5
0	2	0.0	4
1	3	0.5	5
2	1	1.0	3
3	5	1.5	3
4	0	2.0	0
5	0	2.5	0
6	1	3.0	0
7	0	3.5	0

Dove le D_k sono i valori centrali relativi all'intervallo k_{mo} delle deviazioni dal modello ($\pm 0.5 \text{ m}$ nel caso lineare, $\pm 0.25 \text{ m}$ nel caso quadratico), e le $O(D_k)$ il numero osservato di deviazioni nell'intervallo k_{mo} .

La gaussiana "teorica" secondo la quale i valori delle deviazioni dovrebbero distribuirsi ha media nulla e varianza data dalla varianza della distribuzione delle deviazioni:

$$\sigma^2(\text{lin}) = \frac{1}{N-1} \sum_{i=1}^N (\text{lin} - \text{sp})^2 = 6.1$$

$$\sigma^2(\text{quad}) = \frac{1}{N-1} \sum_{i=1}^N (\text{quad} - \text{sp})^2 = 1.1$$

Quindi la "Gaussiana modello" è data da

$$f_{\langle D \rangle \sigma}(D) = \frac{1}{\sigma \sqrt{2\pi}} e^{(-D^2/2\sigma^2)}$$

Il valore aspettato E_k per ogni intervallo è quindi:

$$E_k = N \int_{\text{intervallo } k \text{ mo}} f_{\langle D \rangle \sigma}(D) dD$$

Possiamo comunque calcolare tale valore in modo approssimato nel seguente modo:

$$= N f_{\langle D \rangle \sigma}(D_k) \Delta$$

Ci si può ora può calcolare il contributo di ogni singola D_i al χ^2 :

$$\chi_i^2 = \frac{[N f_{\langle D \rangle \sigma}(D_i) \Delta - F(D_i)]^2}{N f_{\langle D \rangle \sigma}(D_i) \Delta}$$

dove a denominatore, abbiamo utilizzato, come nel caso del test per una distribuzione, E_k .

modello lineare:

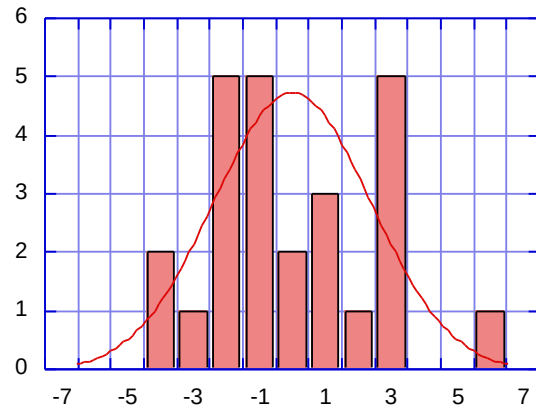
D_k (m)	$O(D_k)$	$N f_{\langle D \rangle \sigma}(D_k) \Delta$	χ_k^2
-7	0	0.072759	0.072759
-6	0	0.21118	0.21118
-5	0	0.52028	0.52028
-4	2	1.0880	0.76453
-3	1	1.9311	0.44894
-2	5	2.9093	1.5024
-1	5	3.7204	0.44012
0	2	4.0382	1.0287
1	3	3.7204	0.13949
2	1	2.9093	1.2531
3	5	1.9311	4.8771
4	0	1.0880	1.0880
5	0	0.52028	0.52028
6	1	0.21118	2.9464
7	0	0.072759	0.072759

modello quadratico:

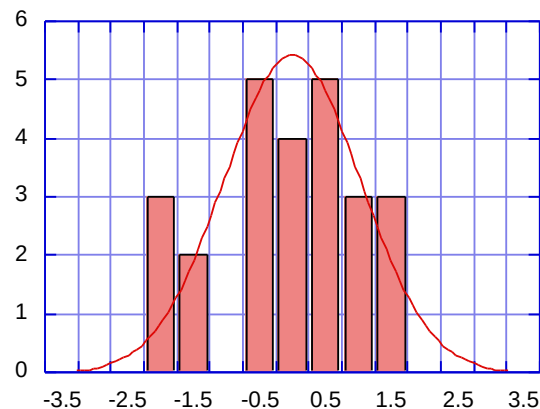
D_k (m)	$F(D_k)$	$N f_{\langle D \rangle \sigma}(D_k) \Delta$	χ_k^2
-3.5	0	0.020761	0.020761
-3.0	0	0.087456	0.087456
-2.5	0	0.29529	0.29529
-2.0	3	0.79913	6.0614
-1.5	2	1.7334	0.040991
-1.0	0	3.0138	3.0138
-0.5	5	4.1999	0.15242
0.0	4	4.6912	0.10183
0.5	5	4.1999	0.15242
1.0	3	3.0138	6.3236e-05
1.5	3	1.7334	0.92543
2.0	0	0.79913	0.79913
2.5	0	0.29529	0.29529
3.0	0	0.087456	0.087456
3.5	0	0.020761	0.020761

Lo stesso risultato si può riportare su grafico:

modello lineare:



modello quadratico:



Sommando tutti i contributi χ_k^2 si ottiene χ^2 :

$$\chi^2(\text{lin}) = 15.9 \quad (c = 2 \rightarrow d = n - c = 11)$$

$$\chi^2(\text{quad}) = 12.1 \quad (c = 3 \rightarrow d = n - c = 10)$$

Dove il numero dei vincoli "c" è rispettivamente 4 e 5

(N, σ ed i parametri calcolati per il modello lineare e quadratico)

Dalle tabelle risulta quindi che la probabilità che sole fluttuazioni statistiche spieghino la discrepanza fra le misure e i modelli è rispettivamente:

$$P(\text{lin}) \approx 16\%$$

$$P(\text{quad}) \approx 29\%$$

Questo significa che in entrambi i casi non possiamo escludere che i modelli siano corretti, in quanto per entrambi la soglia del 5% è ampiamente superata.

Si noti che dire questo NON significa dire che il modello corretto è quello lineare (o quadratico). Entrambi sono compatibili con i dati. Per entrambi le discrepanze con i dati non sono significative.

Infine si noti che, pur suggerendo una maggiore consistenza col modello quadratico (moto uniformemente accelerato), questi risultati nulla dicono a proposito di un "confronto" in quanto la nostra domanda iniziale non concerneva un "paragone" fra i modelli.

Per questa domanda ci si deve indirizzare su altri test (test F per le varianze).

CENNI SUL TEST *t*-student

Supponiamo di avere un numero di dati non grande: l'applicazione della distribuzione della probabilità di Gauss non è lecita.

Si può introdurre un correttivo, per mezzo della distribuzione di *t* (*t* di Student).

Ad esempio, date 10 misure con un certo σ , qual è la probabilità di trovare le nostre misure entro l'intervallo $\langle x \rangle \pm t\sigma$?

I valori di *t* per una determinata probabilità si ricavano le tabelle conoscendo il numero di gradi di libertà *N*-1. Ad esempio, per *N* = 10:

90% $t = 1.83$

95% $t = 2.26$

99% $t = 3.25$

99.9% $t = 4.78$

Cioè, ad esempio, ho il 90 % di probabilità di trovare le mie misure in un intervallo $\langle x \rangle \pm 1.83 \sigma$.

Il test di Student ci permette anche di paragonare medie ricavate da un numero piccolo di misure, in generale:

$$t = \frac{\text{differenza tra le medie}}{\text{errore standard della differenza tra le medie}}$$

Le tabelle della statistica del *t* di Student forniscono la probabilità di avere una differenza fra le medie, dovuta alle sole fluttuazioni statistiche, pari o peggiore di quella trovata. Tale probabilità è funzione dei gradi di libertà $d = N-c$ (c =vincoli), e tende a quella fornita dall'integrale normale degli errori per *N* grande:

<i>d</i>	<i>P</i> =0.90	0.50	0.20	0.10	0.05	0.02	0.01	0.001
1	0.158	1.000	3.078	6.314	12.706	31.821	63.657	636.619
2	0.142	0.816	1.886	2.920	4.303	6.965	9.925	31.598
3	0.137	0.465	1.638	2.353	3.182	4.541	5.841	12.924
4	0.134	0.741	1.533	2.132	2.776	3.747	4.604	8.610
5	0.132	0.727	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.718	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.711	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.706	1.397	1.860	2.306	2.896	3.355	5.041
9	0.129	0.703	1.383	1.833	2.262	2.821	3.250	4.781
10	0.129	0.700	1.372	1.812	2.228	2.764	3.169	4.587
12	0.128	0.695	1.356	1.782	2.179	2.681	3.055	4.318
15	0.128	0.691	1.341	1.753	2.131	2.602	2.947	4.073
20	0.127	0.687	1.325	1.725	2.086	2.528	2.845	3.850
25	0.127	0.684	1.316	1.708	2.060	2.485	2.787	3.725
30	0.127	0.683	1.310	1.697	2.042	2.457	2.750	3.646
40	0.126	0.681	1.303	1.684	2.021	2.423	2.704	3.551
60	0.126	0.679	1.296	1.671	2.000	2.390	2.660	3.460
120	0.126	0.677	1.289	1.658	1.980	2.358	2.617	3.373
∞	0.126	0.674	1.282	1.645	1.960	2.326	2.576	3.291

Si noti:

- per $n \Rightarrow \infty$ la distribuzione tende a quella di Gauss (vedi integrale degli errori, appendice A del; Taylor);
- oltre ≈ 25 misure si commette un errore piccolo usando l'integrale degli errori; l'approssimazione è tanto peggiore quanto maggiore è il valore di *t*.