

Stage PhIScA: Analisi dei dati

Luca Arnaboldi*, Marco Malandrone†

7 febbraio 2018

1 Introduzione

Scopo di questa lezione è:

- illustrare i concetti di errore e incertezza e il loro utilizzo in una prova sperimentale,
- mostrare i principali metodi di fit,
- illustrare alcune funzioni della calcolatrice scientifica particolarmente utili in una prova,
- spiegare le modalità d'uso di alcuni strumenti di laboratorio comuni,
- offrire alcuni consigli generali sul come svolgere una prova sperimentale di laboratorio.

2 Errore: cosa è, che tipi ci sono, come si usa come funziona

2.1 Errore e incertezza

Il metodo sperimentale si basa sulla registrazione di misure sperimentali. Queste sono caratterizzate da un errore e da un'incertezza. Precisamente, l'errore è la differenza tra la misura registrata e il valore "vero" di una misura, mentre l'incertezza identifica un intervallo all'interno del quale si teorizza si

*luca.arnaboldi@sns.it

†marco.malandrone@sns.it

trovi il valore "vero" della misura, che è spesso - se non sempre - ignoto. In particolare è possibile definire l'errore e l'incertezza sia in maniera assoluta che relativa alla grandezza misurata.

2.2 Errori sistematici e casuali. Stima dell'incertezza

L'errore (e, conseguentemente, l'incertezza) è in parte dovuto ad errori umani ed in parte ai limiti degli strumenti, ovvero alla risoluzione strumentale. Esistono errori sistematici e casuali. Idealmente quelli sistematici dovrebbero essere "eliminabili", se non in fase di misurazione almeno con una elaborazione dei dati a posteriori. Quelli casuali sono invece "fisiologici".

Prendiamo in considerazione la misura ripetuta di un particolare dato sperimentale costante x .

Qualora le misure prese siano meno di 5 o 6, un modo efficiente e coerente per stimare l'incertezza è quello della semidisposizione:

$$\Delta x = \frac{1}{2}(x_{max} - x_{min})$$

Se invece il numero di misure è adeguato (5, 6 o più) è possibile invece ipotizzare che i dati si distribuiscano approssimativamente secondo una "curva a campana"; è quindi lecito e raccomandabile in tal caso utilizzare la deviazione standard come incertezza sulla misura (vedi sezione 3).

2.3 Propagazione degli errori massimi

In una prova sperimentale è spesso necessario utilizzare le misure raccolte per stimare altri parametri. In tal caso si pone il problema di studiare come si propagano l'errore e l'incertezza attraverso i calcoli utilizzati per giungere alla stima del parametro. In particolare, è utile differenziare la propagazione degli errori e quella delle incertezze.

Gli errori si propagano in modo "banale": sia $r = f(x)$ un parametro di cui vorremmo avere una stima a partire da x . Sia μ_x il valore "vero" di x , e sia x_{exp} la misura sperimentale raccolta, cosicché l'errore risulti essere $E_x = x_{exp} - \mu_x$. Sia μ_r il valore "vero" di r e sia E_r l'errore su r ottenuto a partire dalla misura sperimentale x_{exp} :

$$E_r = r_{exp} - \mu_r = f(x_{exp}) - f(\mu_x)$$

L'incertezza si propaga in questo modo qualora si consideri l'errore massimo come incertezza (ovvero l'incertezza è stata ottenuta con il metodo della semidisposizione). Per errori sufficientemente piccoli, ciò è equivalente a

considerare l'incertezza come somma delle derivate parziali moltiplicate per gli errori corrispondenti:

$$\Delta f(x_1, x_2, \dots, x_n) = \left| \frac{\partial f}{\partial x_1} \right| \Delta x_1 + \left| \frac{\partial f}{\partial x_2} \right| \Delta x_2 + \dots + \left| \frac{\partial f}{\partial x_n} \right| \Delta x_n \quad (1)$$

Tuttavia, qualora l'incertezza fosse stata ottenuta tramite deviazione standard si tratterebbe allora di un'incertezza "probabilistica", e andrebbe propagata opportunamente (vedi sezione 3)

2.4 Esercizi

Esercizio 1. Larghezza di un campo di calcetto. Si considerino tre misure della larghezza di un campo di calcetto: 29.2 m, 27.7 m e 28.2 m.

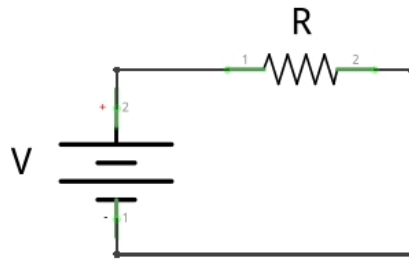
1. Calcolare il valor medio e l'incertezza di questa misura.
2. Effettuiamo ora la misura non più contando i nostri passi ma con un distanziometro laser, che fornisce come misura 28.214 m. Calcolare l'errore massimo compiuto nelle misurazioni. Ricorda di esprimere il risultato con il giusto numero di cifre significative

Esercizio 2. Gara di macchinine. Lanciamo tre macchinine ideali lungo un piano, inclinato di un angolo ignoto α . Il punto di partenza si trova $1 \text{ m} \pm 1 \text{ cm}$ sopra l'arrivo. Gli attriti sono trascurabili, le ruote delle macchinine hanno massa molto minore rispetto a quella di un'intera macchinina. Le macchinine partono ferme. Vengono effettuate le seguenti misure di velocità al traguardo:

Macchinina	VERDE	BIANCA	ROSSA
Velocità [m/s]	4.47	4.42	4.35

1. Stimare l'accelerazione di gravità g e la sua incertezza.

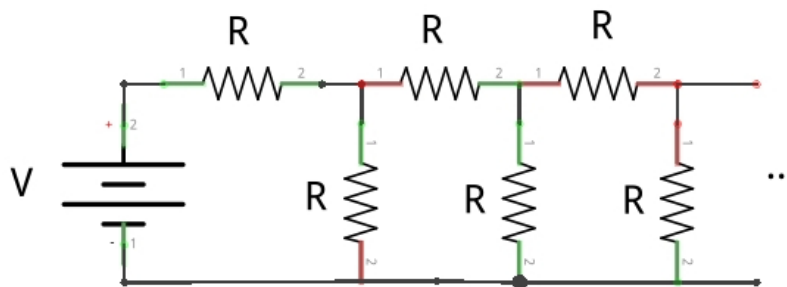
Esercizio 3. Prima legge di Ohm: occhio alle derivate! Si consideri il circuito sottostante, in cui il generatore eroga una tensione $V = 24V \pm 0.2V$ e la resistenza ha valore $R = 2.2k\Omega \pm 10\%$.



1. Quale è il valore della corrente che scorre nel circuito? Quale è l'incertezza relativa?
2. Calcolare ora, utilizzando la prima legge di Ohm, il valore massimo e quello minimo della corrente che può scorrere nel circuito.
3. Confrontare i risultati del punto 1. e del punto 2. e spiegare l'origine delle differenze alla luce della (1). Ricavare algebricamente tali differenze.

HINT: Scrivere $I_{MAX} \approx f(R_0, V_0) + \frac{\partial f}{\partial R}(R_{MIN} - R_0) + \frac{\partial f}{\partial V}(V_{MAX} - V_0)$ e confrontarla con ciò che si deriva dalla prima legge di Ohm: $I_{MAX} = f(R_{MIN}, V_{MAX})$.

Esercizio 4. Calcolare la corrente che scorre nel circuito e la relativa incertezza sapendo che le resistenze sono tutte uguali e hanno resistenza pari a R , dove $R = 100\Omega \pm 5\Omega$ e che il generatore di tensione offre una d.d.p. di $10V \pm 0.1V$.



3 Distribuzione normale ed errore statistico

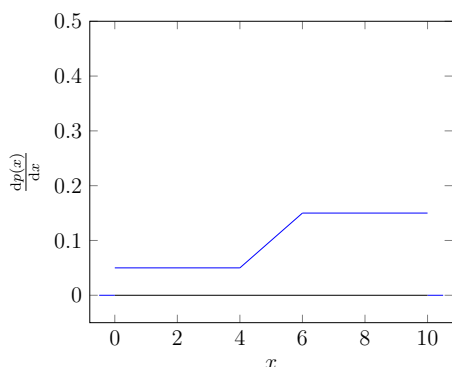
3.1 Distribuzioni di probabilità

In questa sezione si approfondisce il discorso relativo all'incertezza "probabilistica". Introduciamo innanzitutto il concetto di distribuzione continua di probabilità.

Definiamo variabili casuali continue quelle variabili che possono assumere un insieme continuo di valori (al contrario delle distribuzioni discrete, per le quali l'insieme dei possibili valori ha cardinalità minore o uguale a $\mathbb{N}$). Una distribuzione continua di probabilità descrive la probabilità che, presa una variabile casuale continua, questa si trovi in un qualsiasi range possibile di valori. Formalmente, una distribuzione continua di probabilità è definita da una densità di probabilità:

$$f(x) = \frac{dp(x)}{dx}$$

Preso un intorno di x di larghezza dx , tale funzione è il rapporto tra la probabilità che la variabile si trovi in quell'intorno e la larghezza dell'intorno stesso, nel limite in cui tale larghezza tende a 0.



Si consideri, a scopo esemplificativo, la distribuzione di probabilità al lato, definita nell'intervallo continuo $[0, 10]$. La variabile descritta da questa distribuzione può assumere solo valori in questo intervallo. Si noti il fatto l'integrale della funzione è pari a 1, ovvero, considerato l'intervallo $[0, 10]$ come somma di infinitesimi intervalli di larghezza dx e preso un valore casuale per la variabile, la probabilità che questo valore si trovi nell'intervallo $[0, 10]$ è pari a 1. La variabile è in effetti, definita solo

ed esclusivamente nell'intervallo $[0, 10]$.

La funzione della densità di probabilità, per come è definita, descrive la probabilità $p(x_0, x_1)$ che, preso un dato a caso della distribuzione, questo si trovi tra x_0 e x_1 :

$$p(x_0, x_1) = \int_{x_0}^{x_1} f(x) dx$$

3.2 La distribuzione normale e la curva Gaussiana

Nella teoria della probabilità la distribuzione normale, o di Gauss (o gaussiana) è una distribuzione di probabilità continua che è spesso usata come prima approssimazione per descrivere variabili casuali a valori reali che tendono a concentrarsi attorno a un singolo valor medio. Il grafico della funzione di densità di probabilità associata è simmetrico e ha una forma a campana. In altri termini, la distribuzione normale è una curva che ben approssima la distribuzione della maggior parte dei dati sperimentali che raccoglierete in sede di gara. La curva è descritta dalla seguente densità di probabilità:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2} \quad (2)$$

Come si evince dalla definizione della densità di probabilità della gaussiana, la curva è caratterizzata da due parametri, μ e σ , rispettivamente il valor medio della distribuzione e la radice quadrata della varianza, che coincide con lo scarto quadratico medio dei dati nel limite del numero di dati che tende a infinito.

3.3 I dati raccolti e la curva gaussiana

Introduciamo i concetti di popolazione e di campione: chiamiamo popolazione dei dati l'insieme dei dati raccogliabili potenzialmente e campione dei dati l'insieme dei dati raccolti. Ad esempio, si consideri un eventuale studio sugli sport praticati dagli italiani, effettuato intervistando 1000 italiani, scelti casualmente. In tal caso la popolazione dei dati corrisponderebbe ai dati di tutti i 60 milioni di italiani. I dati dei mille intervistati corrisponderebbero invece al campione.

In un esperimento fisico il campione dei dati continua a corrispondere all'insieme dei dati raccolti, mentre la popolazione è virtualmente infinita, in quanto non vi è limite al numero di dati raccogliabili da un determinato esperimento. La curva gaussiana descrive proprio come si distribuiscono i dati relativamente a tutta la popolazione. Nel limite in cui il numero di dati raccolti (costituenti il campione) tende a infinito, la gaussiana ben descrive anche i dati del campione.

3.4 Dai dati alla curva

Introduciamo informalmente il concetto di stimatore: uno stimatore s è una funzione che, a partire dai dati sperimentali, ovvero dal campione dei

dati, è in grado di fornire una *buona stima* di un parametro m della curva che descrive i dati della popolazione. Inoltre uno stimatore, nel limite in cui il numero dei dati del campione tende a infinito, deve tendere in probabilità al valore "vero" del parametro di cui è stimatore¹.

Una volta raccolti i dati sperimentali ci si ritrova davanti a un campione della popolazione dei dati. Da ciò è necessario, di norma, trovare una stima del valor medio della distribuzione e ottenere una stima della deviazione standard della media campionaria.

La stima del valor medio si ottiene calcolando il valor medio del campione:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3)$$

Introduciamo ora lo stimatore della varianza σ^2 (ovvero del quadrato della deviazione standard) della popolazione:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Il fatto che al denominatore si usa $n-1$ e non n deriva dal fatto che $\bar{x}$ non è determinato a priori, e quindi, fissato $\bar{x}$, le variabili "libere" sono $n-1$, non n .

Supponiamo ora di avere a disposizione n osservazioni di una popolazione: $X_1, X_2, \dots, X_n$. Ogni dato sperimentale fra $X_1, X_2, \dots, X_n$ è in realtà una variabile casuale estratta da una distribuzione, quella della popolazione, con media μ e varianza σ^2 . La stima di tale varianza può essere ottenuta con lo stimatore S^2 .

Riportiamo il fatto che nel caso di due distribuzioni normali indipendenti, non correlate, X e Y è possibile dimostrare che:

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) \quad (4)$$

Cerchiamo ora di stimare la varianza di $\bar{x}$:

$$\begin{aligned} S_{\bar{x}}^2 &= \text{var}\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n^2} \text{var}\left(\sum_{i=1}^n x_i\right) \stackrel{(4)}{=} \frac{1}{n^2} \sum_{i=1}^n \text{var}(x_i) \Rightarrow \\ &\Rightarrow S_{\bar{x}}^2 = \frac{1}{n^2} \sum_{i=1}^n S^2 = \frac{1}{n^2} (nS^2) = \frac{S^2}{n} \end{aligned}$$

¹Con l'espressione "tendere in probabilità" si intende che la probabilità che $|s - m| > \varepsilon$ sia vero tende a 0 al tendere all'infinito del numero di dati raccolti, per qualunque $\varepsilon \in \mathbb{R}^+$

Otteniamo quindi come stima della deviazione standard della stima del valor medio:

$$S_{\bar{x}} = \frac{S}{\sqrt{n}} \quad (5)$$

3.5 Interpretare la curva

Dato un intervallo (a, b) , possiamo calcolare la probabilità che x cada nell'intervallo (a, b) , che per definizione è

$$P(x \in (a, b)) = \int_a^b p(x) dx.$$

Se $p(x)$ è la gaussiana di media μ e deviazione standard σ , questa tabella riporta il valore della probabilità per certi intervalli:

$P(x \in (a, b))$	(a, b)
68.3%	$(\mu - \sigma, \mu + \sigma)$
95.0%	$(\mu - 1.96\sigma, \mu + 1.96\sigma)$
95.5%	$(\mu - 2\sigma, \mu + 2\sigma)$
99.7%	$(\mu - 3\sigma, \mu + 3\sigma)$

Sia $\mathbf{X} = (x_1, x_2, \dots, x_n)$ un set di dati con distribuzione normale, media campionaria $\bar{x}$ e deviazione standard della media campionaria $S_{\bar{x}}$. È possibile definire il concetto di intervallo di confidenza: dato un intervallo che dipende dai dati $(x_L(\mathbf{X}), x_R(\mathbf{X}))$ è definito un relativo livello di confidenza come la probabilità che μ si trovi effettivamente nell'intervallo considerato.

Nel limite in cui n , il numero di misurazioni, tende a infinito, $S_{\bar{x}}$ *ben approssima* $\sigma_{\bar{x}}$. Ciò consente di valutare quantitativamente la probabilità che μ si trovi in un determinato intervallo. Per fare ciò, è sufficiente considerare la gaussiana di media $\bar{x}$ e deviazione standard $S_{\bar{x}}$, e valutare l'integrale della densità di probabilità fra x_L e x_R . Scegliendo opportunamente i due estremi e ricordandosi alcuni valori riportati nella tabella precedente, è possibile inoltre evitare di dover ripetere il calcolo dell'integrale definito (che, in questi casi, si effettua con metodi numerici).

3.6 Propagare l'errore statistico

Spesso è sufficiente propagare le incertezze con il metodo delle derivate parziali illustrato nel paragrafo 2.3. Tuttavia sarebbe più rigoroso, in caso

di incertezza ottenuta tramite deviazione standard, utilizzare la somma del prodotto di covarianze e derivate parziali.

La covarianza tra un campione di due variabili x, y , di cui sono state prese n osservazioni congiunte, con medie campionarie rispettive $\bar{x}$ e $\bar{y}$, è definita come:

$$S_{x,y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} \left(\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i \right)$$

Qualitativamente, la covarianza è un numero che fornisce una misura di quanto x e y varino assieme, ovvero della loro correlazione.

Uno stimatore della deviazione standard della funzione $f(x_1, \dots, x_n)$ è:

$$S_f = \sqrt{\sum_{i=1}^n \sum_{j=1}^n \left(\frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} S_{x_i, x_j} \right)} \quad (6)$$

Che nel caso $n = 2$ prende la forma:

$$S_f = \sqrt{\left(\frac{\partial f}{\partial x_1} S_{x_1} \right)^2 + 2 \frac{\partial f}{\partial x_1} \frac{\partial f}{\partial x_2} S_{x_1, x_2} + \left(\frac{\partial f}{\partial x_2} S_{x_2} \right)^2}$$

Nel caso di variabili $x_1, \dots, x_n$ non correlate (quindi con covarianza nulla fra x_i e x_j per $i \neq j$) la (6) assume la forma:

$$S_f = \sqrt{\sum_{i=1}^n \left(\frac{\partial f}{\partial x_i} S_{x_i} \right)^2} \quad (7)$$

Che nel caso $n = 2$ prende la forma:

$$S_f = \sqrt{\left(\frac{\partial f}{\partial x_1} S_{x_1} \right)^2 + \left(\frac{\partial f}{\partial x_2} S_{x_2} \right)^2}$$

Esercizio 5. Patate. Fabio prende un piatto di patate al forno e le cataloga una a una su una scala da 0 a 100 in base alla cottura, ovvero al colore, che va dal bianco (0-20), al giallo (20-40), all'arancio (40-60), al marrone (60-80) e al nero (80-100). Catalogati tutti i 30 pezzi di patate, Fabio constata che la loro cottura può essere ben approssimata con una gaussiana di media 40 e deviazione standard 20. Sapendo che il numero di presenti nel locale è pari a 200 persone, e tutti hanno preso un piatto di patate, stimare il numero di pezzi di patate neri serviti.

Esercizio 6. Annullare la covarianza. Siano dati i due seguenti set di dati:

$$X = \{x_1, \dots, x_n\}$$

$$Y = \{y_1, \dots, y_n\}$$

Si tratta di misure congiunte (x_i, y_i) di due parametri di un sistema. In tabella si riportano le misure:

x_i	y_i
21,34	33,07
22,22	34,23
21,70	33,72
21,83	33,65
22,05	33,73
21,98	34,00
21,80	33,59
21,97	33,87

1. Stimare l'incertezza sul parametro:

$$\bar{t} = \bar{x} + \bar{y}$$

2. X e Y mostrano una correlazione non nulla, e dunque una covarianza non nulla. In particolare, la teoria afferma che dovrebbero essere legate dalla relazione $y_i = mx_i + q$. L'analisi dei dati mostra che $S_m = 1.2$ è una buona stima di m . Costruire una nuova variabile $z_i = f(x_i, y_i)$ tale che $\text{cov}(X, Z)=0$. Scrivere t come funzione di x e z . Trovare l'incertezza associata a $\bar{t}$.

4 Metodi di fit

4.1 Cos'è il fit

Una volta raccolto un set di dati costituito da coppie di misurazioni (x_i, y_i) , nota una relazione teorica tra x e y , può essere necessario ottenere sperimentalmente i valori dei parametri della funzione $y = f(x, p_1, \dots, p_n)$ che lega le due variabili. La funzione $f(x)$ può avere più parametri, e può essere necessario stimarli contemporaneamente. Il caso più comune e utilizzato è quello della retta, in cui la funzione prende la forma $f(x) = mx + q$. Ci sono in questo caso diversi modi di procedere per ottenere una stima di m e di q .

4.2 Fit grafico

Il modo più elementare è quello di riportare su un grafico x, y le misurazioni e tracciare una retta che "a occhio" meglio descrive l'andamento dei punti.

Qualora siano note le incertezze delle misure, è possibile inoltre tracciare le cosiddette barre di errore, e utilizzarle per tracciare le rette di massima e minima pendenza. Dai relativi valori di m e q è poi possibile stimare l'errore massimo tramite semidisersione. In tal caso si ha che:

$$m = \frac{m_+ + m_-}{2}$$

$$q = \frac{q_+ + q_-}{2}$$

$$\Delta m = \frac{m_+ - m_-}{2}$$

$$\Delta q = \frac{q_- - q_+}{2}$$

Dove m_+, m_-, q_+, q_- descrivono la retta di massima ($y = m_+x + q_+$) e di minima pendenza ($y = m_-x + q_-$).

4.3 Metodo delle coppie di punti

In caso si abbia a disposizione un numero pari di punti, è possibile utilizzare il metodo delle coppie di punti. In particolare, da $2n$ misurazioni è possibile ottenere un set di n stime di m e q :

$$m_i = \frac{y_{n+i} - y_i}{x_{n+i} - x_i}$$

$$q_i = y_i - m_i x_i$$

Il criterio con cui vengono formate le coppie di punti ha fondamentalmente due pregi: il primo è quello di non utilizzare due volte uno stesso punto (il quale avrebbe quindi un "peso" maggiore nel determinare le stime dei parametri), il secondo è quello di mantenere massima la distanza tra due punti, cosicché gli errori sulle misure siano piccoli rispetto alle distanze fra i due punti.

Dai vari m_i e q_i è poi possibile ricavare delle stime di m e q :

$$m_s = \frac{1}{n} \sum_{i=1}^n m_i$$

$$q_s = \frac{1}{n} \sum_{i=1}^n q_i$$

L'incertezza su m_s e q_s può essere ottenuta sia tramite semi-dispersione che tramite calcolo della deviazione standard del campione, in funzione del numero di misurazioni effettuate.

4.4 Metodo dei minimi quadrati ordinari

Un altro metodo di fit è quello dei minimi quadrati ordinari, che consiste nel minimizzare la quantità:

$$\sum_i (f(x_i) - y_i)^2$$

Il grande pregio di questo metodo è quello di poter essere spesso risolto analiticamente. Tuttavia, è utilizzabile efficacemente solo ove le incertezze lungo l'asse x siano trascurabili rispetto a quelle sull'asse y (formalmente, $\Delta y \gg m \Delta x$). Inoltre, le incertezze dei vari dati y_i dovrebbero essere uguali fra loro. In particolare la soluzione nel caso di fit lineare è:

$$m = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}$$

$$q = \frac{\sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}$$

$$\sigma_m = \sigma_y \sqrt{\frac{n}{n \sum_i x_i^2 - (\sum_i x_i)^2}}$$

$$\sigma_q = \sigma_y \sqrt{\frac{\sum_i x_i^2}{n \sum_i x_i^2 - (\sum_i x_i)^2}}$$

$$\sigma_{mq} = \sigma_y^2 \frac{-\sum_i x_i}{n \sum_i x_i^2 - (\sum_i x_i)^2}$$

ove σ_y è l'incertezza su un singolo dato y_i .

4.5 Metodo del minimo χ^2

Qualora le incertezze dei vari y_i non siano uguali fra loro, è opportuno far "pesare" di più la distanza tra la retta e i punti con incertezza minore;

viceversa, la distanza dalle misure meno precise dovrebbe "influenzare meno" l'andamento della funzione. A tale scopo, è definita la variabile χ^2 :

$$\chi^2 = \sum_i \left(\frac{f(x_i) - y_i}{\sigma_{y_i}} \right)^2 \quad (8)$$

Minimizzando la variabile χ^2 si ottiene proprio lo scopo prima descritto, utilizzando gli inversi dei quadrati delle incertezze delle singole misurazioni come pesi della somma delle differenze $f(x_i) - y_i$.

5 Usare la calcolatrice

La calcolatrice è uno strumento potentissimo che vi è concesso usare durante la gara, e saperla usare veramente vi permette di molto tempo, e quindi punti, durante le prove (anche nella teorica è importante saper usare la calcolatrice).

Di seguito verranno descritte alcune delle principali feature delle calcolatrici, anche se la procedura esatta per eseguirle dipende da modello a modello; leggete il manuale della vostra (o cercate video-tutorial su YouTube).

Il nostro consiglio è di provare a casa a "smanettare" con la calcolatrice, per acquisire dimestichezza e abilità, e non avere problemi durante la gara.

5.1 Modalità statistica

Le calcolatrici scientifiche sono state programmate per eseguire automaticamente tutti i conti visti nelle sezioni precedenti, così da non doversi ricordare a memoria tutte le formule.

5.1.1 Modalità SD

All'interno di questa modalità si può gestire la statistica di una singola variabile. Si inseriscono i valori raccolti e successivamente premendo pochi tasti si possono avere valore medio $\bar{x}$, deviazione standard sx e σx , e un sacco di altri valori meno utili come $\sum x$ e $\sum x^2$.

5.1.2 Modalità regressione-lineare

Il funzionamento è simile alla modalità precedente: si inseriscono le coppie di valori (x, y) misurati, e successivamente si possono ottenere i coefficienti della retta di best fit calcolati direttamente dalla calcolatrice.

La maggior parte delle calcolatrici utilizza come metodo di fit i minimi quadrati

ordinari, ma ne esistono alcune (in particolare le SHARP) che invece usano il metodo dei minimi quadrati totali. Se i dati sono sufficientemente vicini all'essere una retta non ci sono differenze pratiche tra le due interpolazioni, mentre in altre situazioni in cui i dati non sono così buoni ci possono essere grandi differenze. È bene ricordare che il metodo dei minimi quadrati ordinari ha l'ipotesi che l'errore sulle x sia trascurabile, mentre quello dei minimi quadrati totali no, e in base alla situazione può cambiare il metodo da usare; nella pratica durante una prova sperimentale delle Olimpiadi di Fisica non è necessario curarsi particolarmente questa differenza. Oltre ai parametri

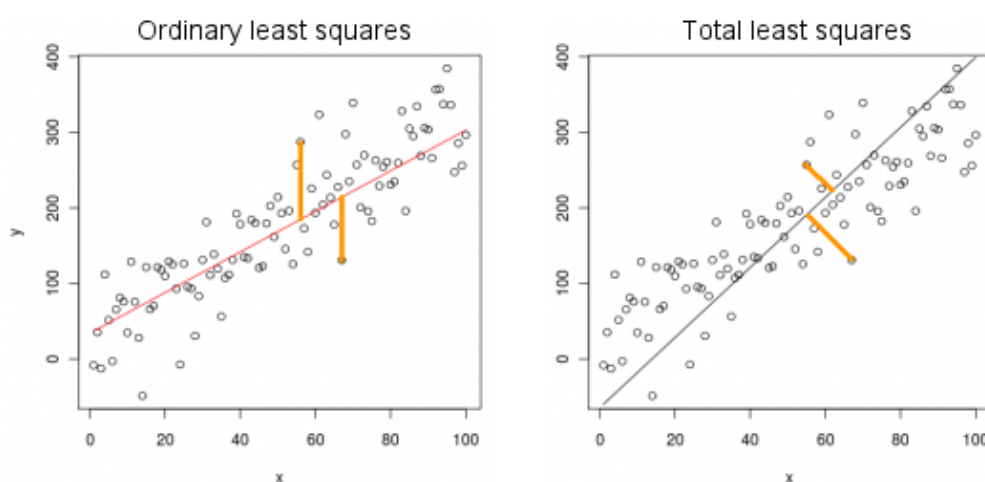


Figura 1: Differenza tra minimi quadrati totali e minimi quadrati ordinari. Con gli stessi punti si possono ottenere risultati molto diversi.

della retta la calcolatrice fornisce il *coefficiente di correlazione di Pearson*, che misura la correlazione dei inseriti

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}.$$

Dalla disuguaglianza di Cauchy-Schwarz segue che $-1 \leq r \leq 1$, e in particolare si ha che $r = 1$ quando i punti sono su una retta di coefficiente angolare positivi, viceversa $r = -1$ quando i punti sono su una retta di coefficiente angolare negativo. Si può utilizzare quindi $-r$ per dare una stima di quanto i punti siano vicini ad essere una retta: più il valore è vicino ad 1, più i punti rappresentano veramente una relazione lineare. Non è possibile dare una regola che valga in ogni situazione per valutare i vostri dati, ma sicuramente se $|r| < 0.90$ c'è qualcosa di profondamente sbagliato in ciò che avete fatto; se $0.90 < |r| < 0.96$ è il caso di farsi qualche domanda se tutto ciò che avete fatto è corretto ed eventualmente ripetere qualche misura; se $0.96 < |r| < 0.99$

allora la vostra regressione *dovrebbe* essere accettabile; infine se $|r| > 0.99$ allora sicuramente avete fatto un buon lavoro.

Come detto esistono eccezioni alle regole precedenti. In una regressione come quella della prima parte di Senigallia 2017 ottenere $r = 0.97$ era un buon risultato, dato che ci potevano essere diversi errori, mentre in altre situazioni è da ritenere cattivo anche un $r = 0.98$. In generale usate il coefficiente di Pearson come double-check e non come criterio assoluto per decidere se i vostri dati vanno bene oppure no.

Per acquisire esperienza su come funziona coefficiente di Pearson può essere interessante il sito <http://www.istics.net/Correlations/> dove si può giocare a *Guessing correlation*, un gioco dove vince chi indovina più volte il valore di r^2 dato un set di punti.

Esercizio 7. Alla mensa della SNS si vuole stimare la quantità di patate da cucinare ad ogni pasto. Viene supposto che la quantità di patate servite dipenda linearmente dal numero di persone che mangiano nella mensa. Nella Tabella 1 sono riportati i dati raccolti negli ultimi 8 pasti: Trovare i coefficienti

Persone	Quintali di patate servite
180	39
135	29
90	9
10	7
50	8
220	35
140	36
120	22

Tabella 1

della relazione tra patate e persone. Il modello scelto può funzionare?

Le stime vengono riviste dopo che le addette si sono accorte che non tutti hanno consegnato lo scontrino prima di ricevere i piatti e inoltre parte delle patate servite risalgono ai giorni precedenti. Controllando gli accessi ai tornelli e rifacendo i conti su ciò che è uscito dalla cucina vengono aggiustati i dati come in Tabella 2. Calcola i nuovi valori del fit. Il modello è più attendibile ora?

Esercizio 8. Calcolare il coefficiente di Pearson dei dati (x, x^2) con $x \in \{0, 1, 2, 3, 4, 5\}$. Trarre le dovute conclusioni.

Ripetere l'esercizio con $x \in \{-3, -2, -1, 0, 1, 2, 3\}$ e commentare il risultato

Persone	Quintali di patate servite
183	56
135	39
87	28
55	17
59	23
218	65
140	42
241	71

Tabella 2

considerando che tanto più $|r|$ è vicino a 0 meno i dati dovrebbero essere correlati.

5.1.3 Regressioni non lineari

Le calcolatrici offrono anche la possibilità di fittare funzioni non lineari, come quadratiche, esponenziali e logaritmi. Il funzionamento è molto simile al caso lineare, cioè si inseriscono i valori e la calcolatrice restituisce i parametri della funzione; alle olimpiadi è *molto sconsigliato* usare la calcolatrice per trovare i parametri di funzioni non lineari, solitamente si preferisce ricondursi al caso di una funzione lineare, come spiegato in seguito. Nonostante ciò i fit non lineari vi possono essere utili indirettamente per controllare che le vostre misure effettivamente siano sensate, prima di linearizzarle, che in genere è un processo abbastanza lungo.

5.2 Memorie interne

Ogni calcolatrice scientifica possiede almeno 4-5 variabili che posso essere inizializzate a piacere. Questo può far molto comodo durante la prova, permette di risparmiare tempo e evita che ci siano errori di trascrizione nella calcolatrice.

Solitamente per utilizzare questa funzionalità si preme il tasto **STO** e successivamente il tasto della variabile che si vuole utilizzare; a questo punto è possibile utilizzare la variabile nei propri conti.

5.3 Altre funzionalità

Di seguito vengono elencate delle funzionalità vagamente utili (per la prova Teorica in realtà) che alcuni modelli di calcolatrici possiedono:

- sommatorie;
- derivate puntuali;
- integrali definiti;
- risoluzione numerica di equazioni;
- risoluzione di sistemi lineari, anche se è comunque utile imparare risolverli in maniera efficiente (vedi *Algoritmo di Gauss*).

Se avete bisogno acquistare una calcolatrice tenete presente che le calcolatrici programmabili *non* sono consentite. Il nostro consiglio è *CASIO FX-570 ES PLUS*, offre tutte le funzionalità che sono state presentate e il prezzo è veramente basso.

6 Come affrontare le prove sperimentali di gara

6.1 Prendere le misure

Durante una prova sperimentale è fondamentale saper organizzare il tempo al meglio.

Prima di iniziare a prendere misure è molto importante aver letto il testo di tutta la sezione che si sta affrontando, per essere sicuri che non stia dimenticando nulla. Inoltre prima di registrare dati è consigliabile fare qualche prova per essere sicuri che tutto sta funzionando come deve e non ritrovarsi a metà dell'esperimento a dover ripetere tutto perché qualcosa è andato storto. Quando si prendono misure è importante di ricordarsi di:

- *distribuire le misure su tutto l'intervallo possibile*: se avete la possibilità di far variare una grandezza su un intervallo copritelo tutto oppure perderete punti;
- *raffinare le misure intorno ad una zona interessante*: se state effettuando un esperimento in cui si prevede che una zona di particolare interesse (ad esempio un picco) è molto apprezzato prendere misure più fitte in questo intervallo;
- *prendere il numero adeguato di misure*: effettuare troppe misure potrebbe essere controproducente perché vi richiede tempo in più sia per farle che per gestire più dati. Fare invece poche misure vi può costare punti preziosi. In generale è consigliabile fare poche misure di più di quelle richieste nel testo per riuscire a prendere tutti i punti.

Una volta ottenuti i dati inizia la parte di analisi, che sicuramente richiederà un fit.

6.2 Linearizzazione

Il modello che sta alla base dell'esperimento della prova molto probabilmente non è lineare, cioè mettendo in un grafico i dati sperimentali che avete misurato quello che vi aspettate di ottenere non è una retta. Dato che gli strumenti presentati prevedono solamente fit lineari è necessario ricondursi a casi di questo tipo, attraverso manipolazioni algebriche.

Supponiamo di aver a disposizione un campione di dati (x_i, y_i) legati dalla relazione

$$f(x, y) = 0, \quad (9)$$

dove f è una funzione che dipende anche da parametri incogniti (e l'obiettivo dell'esperimento è trovarli), più eventualmente altri parametri noti. Quello che si vuole fare è riscrivere i dati sperimentali trasformando le coppie

$$(x_i, y_i) \rightarrow (X(x_i, y_i), Y(x_i, y_i))$$

dove X e Y sono funzioni tali che la relazione (9) si possa scrivere come:

$$Y(x, y) = M \cdot X(x, y) + Q$$

dove M e Q dipendono solamente dai parametri e non dalle variabili. A questo punto è possibile applicare i noti metodi per i fit lineari alle coppie (X_i, Y_i) e ricavare i valori di M e Q , da cui poi si trovano i parametri cercati originariamente.

Esempio. Supponiamo di misurare coppie (x, y) legate dalla relazione

$$y = A\sqrt{x^3 + B}$$

dove A e B sono due valori da determinare a partire dai dati sperimentali. Ovviamente dobbiamo ricondurci a una situazione dove si può applicare un fit lineare. Eleviamo entrambi i membri al quadrato:

$$y^2 = A^2 (x^3 + B)$$

$$y^2 = A^2 x^3 + A^2 B$$

A questo punto chiamando $X = x^3$ e $Y = y^2$ la relazione appena scritta è lineare in X e Y , e particolare $M = A^2$ e $Q = A^2 B$.

y	x	y^2	x^3
13.2	0	174.24	0
14.1	1	198.81	1
19.4	2	376.36	8
29.2	3	852.64	27

Tabella 3: Dati esempio di linearizzazione della relazione $y = A\sqrt{x^3 + B}$.

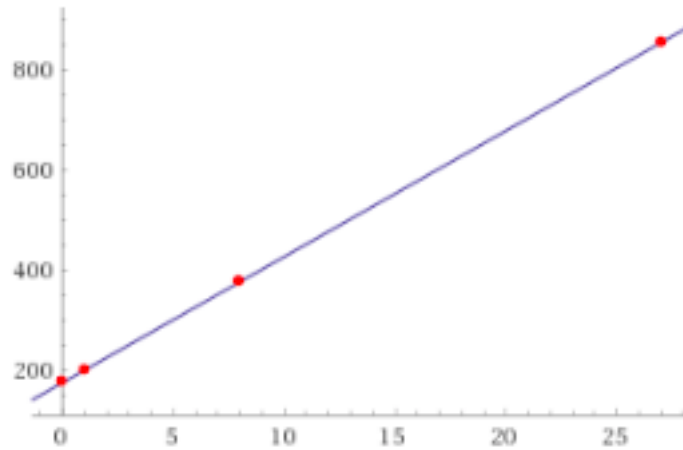


Figura 2: y^2 in funzione di x^3 , con la retta di best fit.

Nella Tabella 3 degli ipotetici dati sperimentali, con le relative trasformazioni.

Mettendo in un grafico i punti trasformati si ottiene la Figura 2. I coefficienti della retta di best-fit permettono di trovare che $A^2 = 25.133$ e $A^2B = 174.315$, da cui si ricava $A = 5.013$ e $B = 6.935$.

Perché la linearizzazione non è usata al di fuori delle olimpiadi

Negli esperimenti di fisica al di fuori delle olimpiadi la linearizzazione non è praticamente mai usata. Questo è dovuto al fatto che nel trasformare i valori da (x, y) a (X, Y) si propagano anche gli errori, e spesso questo non accade in maniera controllata. Linearizzare una funzione nella realtà spesso vuol dire trasformare i propri dati sperimentali in qualcosa di inutile per fare un'analisi seria. Ad esempio si pensi ad un punto che si trova nei pressi di un asintoto verticale della funzione: linearizzando il punto diventa totalmente inutile, dato che l'incertezza ad esso associato è troppo grande.

Quando si fanno esperimenti veri si hanno anche a disposizione strumenti di calcolo molto più potente rispetto alla calcolatrice (i.e. computer con Python)

e quindi viene anche meno la necessità di limitarsi a fit lineari. Nonostante i limiti visti, la linearizzazione resta una tecnica accettata (e apprezzata) nelle gare delle Olimpiadi di Fisica, dato che non si hanno a disposizione metodi migliori di questo. Si consiglia quindi di prendere familiarità nel suo utilizzo, perchè molto utile durante la competizione.

Esercizio 9 (Tratto dalla prova sperimentale Senigallia 2017). Conoscendo il valore di m_m e avendo misurato coppie di valori (v_i, θ_i) trovare la linearizzazione appropriata per poter calcolare i coefficienti k e μ_m a partire dalla seguente relazione:

$$m_m g \sin \theta - \mu_m m_m g \cos \theta - kv = 0.$$

Supponendo di poter misurare il campo magnetico B in un determinato punto al variare della distanza z del magnete da esso, determinare i parametri A e x legati dalla relazione:

$$B = A \left(\sqrt{R^2 + z^2} \right)^x,$$

dove R è un valore noto.

Esercizio 10 (Tratto da Senigallia 1 - 2013). Nei liquidi newtoniani la viscosità diminuisce all'aumentare della temperatura secondo la relazione di Arrhenius

$$\mu = \mu_0 e^{E/(RT)}$$

dove μ_0 è una costante legata al peso molecolare e al volume molare del liquido, E è una costante caratteristica del liquido chiamata energia di attivazione ed è riferita a una mole, R la costante dei gas e T la temperatura assoluta. I dati in Tabella 4 riportano l'andamento della viscosità del mercurio in funzione della temperatura.

Si vogliono determinare i valori delle costanti μ_0 ed E per il mercurio.

Esercizio 11 (Tratto da IPhO 1999 - Padova). Noto che $M_2 = (15.0 \pm 0.4)$ g, $k = (0.055 \pm 0.001)$ N · m · rad⁻¹ e la relazione

$$\frac{k}{4\pi^2 M_2} T^2 = x^2 - lx + \frac{l^2}{3} + \frac{I_1}{M_2},$$

ricavare il valore di I_1 e l a partire dai dati riportati in Tabella 5.
Soluzione: $I_1 = (1.7 \pm 0.7) \cdot 10^{-4}$ kg · m² e $l = (230 \pm 20)$ mm.

Temperatura [K]	Viscosità μ [$10^{-3}\text{Pa} \cdot \text{s}$]
273	1.681
283	1.621
293	1.552
303	1.499
313	1.450
323	1.407
333	1.367
343	1.327
373	1.232

Tabella 4: Dati per l'esercizio 10.

x [mm]	T [s]
204 ± 1	0.502 ± 0.002
215 ± 1	0.528 ± 0.002
231 ± 1	0.562 ± 0.002
258 ± 1	0.628 ± 0.002
290 ± 1	0.708 ± 0.002
321 ± 1	0.790 ± 0.002

Tabella 5: Dati per l'esercizio 11.

6.3 Che fit usare?

I fit grafici richiedono che sia fatto un disegno molto preciso, e questo spesso richiede moltissimo tempo, troppo per una gara: meglio evitarli.

Il metodo delle coppie piace molto ai professori delle olimpiadi e non richiede che ci sia un grafico fatto per bene. Nonostante ciò ha due svantaggi non indifferenti: affinché abbia senso la sua applicazione bisogna avere un numero sufficiente di punti sperimentali (che spesso significa almeno 10), e per fare tutti i conti che richiede bisogna lavorare non poco con la calcolatrice. Il mio consiglio è di usarlo solo in casi di emergenza, ovvero quando è chiaro che l'esperimento non sta dando i risultati attesi (che nelle gare si traduce in: *quello che dovrebbe essere una retta non lo è per niente*) in modo da ben disporre il correttore con un metodo da lui apprezzato.

Il metodo dei minimi quadrati, applicato utilizzando la calcolatrice, è probabilmente il migliore da utilizzare durante la gara, per quantità di tempo richiesta e per qualità dei risultati. Dopo aver riportato i dati in una tabella sulla prova è sufficiente scrivere una frase del tipo "applicando una regressione lineare con la calcolatrice si ottiene..." e poi riportare i parametri che avete

calcolato. Se viene richiesta l'incertezza sui parametri è apprezzato utilizzare il metodo delle rette di massima e minima pendenza, ma questo richiede tempo e quindi è consigliato di farlo solo alla fine, per non perdere tempo utile nel proseguire.

Spesso le sperimentali contengono alcune parti risolvibili senza aver preso misure: è una buona strategia risolverle tutte anche se non si è arrivati così avanti nella prova, magari quando mancano pochi minuti alla fine e non c'è più tempo di fare altro.

6.4 Grafico

Disegnare un grafico che rappresenti i dati raccolti è una richiesta presente in ogni prova sperimentale delle olimpiadi, che di solito copre una parte considerevole del punteggio. Per fare un grafico velocemente e senza perdere punti bisogna:

- **non** disegnare un grafico che non sia una retta a meno che non sia esplicitamente richiesto. Assolutamente **non** disegnare spezzate²;
- mettere grandezze e unità di misura sugli assi;
- occupare tutto o quasi lo spazio sulla carta millimetrata. Questo non vuol dire solo che gli assi devono essere effettivamente lunghi quanto i lati del foglio, ma anche che i punti che inserite occupino tutto il grafico;
- scegliere unità comode per non impazzire (e perdere tempo) convertendo i dati da riportare nel grafico;
- riportare la retta di best-fit ed eventualmente le barre di errore.

6.5 Altri consigli

Le prove sperimentali richiedono molto esperienza per essere affrontate al meglio, per aver imparato tutti i trucchetti per ridurre gli errori; purtroppo spesso non c'è la possibilità di allenarsi. Quello che si può fare è leggere le prove e le soluzioni degli anni passati per "rubare" i trucchetti che vengono adottati, anche se talvolta può risultare difficile capire ciò che c'è scritto senza avere niente del materiale davanti. Leggere le griglie di valutazione può aiutare a capire cosa viene valutato di più e cosa di meno in una prova, per poi comportarsi di conseguenza.

²Esistono rarissime eccezioni, dove però è specificato sul testo cosa fare (i.e. IPhO 2012 E1).

Nelle prove sperimentali italiane (cioè Senigallia e PreIPhO) viene dato molto peso agli accorgimenti che si prendono per migliorare l'esperimento e ridurre l'errore: è importante rispondere per bene a tutte le domande a riguardo e segnalare qualsiasi accorgimento aggiuntivo che prendete, perché potrebbe valere punti. Alle IPhO questa cosa non è presa molto in considerazione e la prova spesso è una lunghissima serie di analisi dati e grafici, e tutti i punti vengono assegnati con una fittissima griglia di valutazione. Si possono quindi portare a casa punti semplicemente consegnando grafici con solo gli assi disegnati o con altri stratagemmi simili.

7 Strumenti di laboratorio

Come avvertimento generale è bene fare in modo che nessun fattore esterno interferisca con gli strumenti. Non sempre è facile farlo, ma è importante se non si vuole rovinare l'esperimento. Di seguito una lista di cose che dovrete imparare a fare/usare:

- *Leggere da una scala graduata*: in particolare saper scegliere la tacca giusta e prendere l'incertezza in maniera sensata;
- Evitare l'errore di parallasse;
- *Calibro*;
- *Massiera*;
- *Multimetro*: sarebbe utile aver visto anche l'uso di uno analogico;
- *Componenti elettrici*: ecco una lista di componenti con cui bisognerebbe aver dimestichezza (cioè aver visto almeno una volta un circuito che li usa):
 - Resistori: è molto utile aver visto
 - Potenzziometro;
 - Condensatori: anche con polarità;
 - Diodi e LED;
 - Batteria, generatori di tensione e di corrente.

Riferimenti bibliografici

- [1] Luca Baldini, *Introduzione all'analisi dei dati*