Analisi dei dati

Luca Arnaboldi*, Marco Malandrone†, Fabio Zoratti‡

6 febbraio 2019

Sommario

Lo scopo della 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.

1 Errore nelle misurazioni

1.1 Incertezza

Nelle scienze sperimentali il concetto di *incertezza* o *errore* ricopre un ruolo di fondamentale importanza. Dato che non è concettualmente possibile misurare una grandezza fisica con accuratezza infinita, il risultato di una misurazione deve essere sempre corredato dalla relativa incertezza, che esprime un intervallo in cui possiamo ragionevolmente pensare di trovare il *valore vero*[1] della grandezza in questione. Ogni grandezza fisica misurata x si scriverà dunque come:

$$x = \hat{x} \pm \delta x,$$

dove x rappresenta il cosiddetto valore vero, $\hat{x}$ è la migliore stima di x che abbiamo ottenuto dalla misurazione e δx è l'incertezza associata ad $\hat{x}$. Il rapporto $\frac{\delta x}{\hat{x}}$ viene indicato come *errore relativo*.

*luca.arnaboldi@sns.it

†marco.malandrone@sns.it

‡fabio.zoratti@sns.it

[1] L'espressione "valore vero" è spesso evitata nei testi di fisica viene sostituita con il termine *misurando*, dato che è impossibile conoscere quale sia effettivamente questo valore. Siccome però è utile a chiarire di cosa si sta parlando nel seguito verrà utilizzata ugualmente.

L'incertezza di una misura è dovuta a diversi fattori. In primo luogo c'è la *risoluzione* dello strumento, ovvero la minima variazione di una certa quantità che lo strumento è capace di misurare. Ogni dispositivo atto alla misurazione di una qualche grandezza ha ovviamente dei limiti fisici che impediscono di raggiugengere una precisione infinita[2]. Una seconda tipologia di errore sono i cosiddette *errori casuali* che, come si può dedurre dal nome, intervengono in maniera imprevedibile e randomica nella singola misura. Essi sono fisiologici e non eliminabili, ma dato il loro compertamento casuale possono essere trattati con opportuni strumenti matematici, come si vedrà in seguito nella Sezione 2.

Infine la terza categoria di errori sono gli *errori sistematici*, che racchiundono al loro volta all'interno una vasta gamma di fattori che influenzano la misura. Esempi tipici sono gli errori di calibrazione degli strumenti (che a loro volta si dividono in errori di scala e di zero) oppure l'*effetto parallasse*. Gli errori sistematici sono idealmente eliminabili preparando un apparato sperimentale sempre più preciso e curato, ma nella realtà è impossibile farli sparire del tutto. Identificare e trattare opportunamente tutte le possibili fonti di errori sistematici è il compito più complesso di un fisico sperimentale.

Grandezze derivate

In un esperimento fisico spesso si vuole ottenere una stima di una grandezza che non è misurabile direttamente. Supponiamo di voler misurare la densità di un materiale: non esiste alcuno strumento (o quantomeno nessuno strumento semplice che possa essere ragionevolmente utilizzato in una prova delle Olimpiadi di Fisica) in grado di quantificare la densità in maniera diretta. È invece relativamente più semplice ottenere la massa e il volume dell'oggetto in questione, e quindi attraverso esse è comunque possibile avere una stima della densità. Ovviamente, in quanto valori misurati, massa e volume hanno le relative incertezze, che ci aspettiamo producano quindi un'incertezza sul valore della densità. La quantificazione dell'incertezza sulle grandezze derivate è detta *propagazione dell'errore*, ed è fondamentale eseguirla in maniera sensata all'interno di un'analisi sperimentale.

1.2 Errore massimo

L'errore massimo è definito come l'incertezza il cui intervallo associato contiene *sicuramente* il valore vero. Dato un insieme di diverse misurazioni

[2]Questo non implica che l'incertezza non possa essere piú bassa della risoluzione, come vedremo in seguito.

$(x_1, x_2, \ldots, x_n)$, esso si stima tramite la *semidispersione*, definita come

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

È la stima più grossolana che si può avere di incertezza e soprattutto porta con sè una contraddizione intrinseca: non possiamo essere certi che ripetendo la misurazione un'altra volta il valore ottenuto sia nell'intervallo. Si conclude quindi che la ripetizione di misurazioni non può che peggiorare la nostra incertezza, il che è abbastanza assurdo. Tuttavia, in situazioni in cui non è richiesta un'analisi degli errori particolarmente complessa e strutturata (come in una prova sperimentale olimpica), esso fornisce un ottimo modo per stimare l'incertezza di misure dirette e, attraverso la propagazione, indirette.

La regola della nonna per il calcolo dell'errore massimo su una misura diretta, effettuata una volta sola, è la tacca o la mezza tacca dello strumento se abbiamo in mano un righello o un calibro, mentre **è assolutamente necessario leggere il manuale** se avete in mano uno strumento di misura come un tester, sia analogico che digitale. Dietro la misurazione di tensioni e correnti c'è logica non banale che porta a non rendere completamente affidabili tutte le cifre che leggete sul display. Un tester sensato ha come ordine di grandezza dell'incertezza l'1%, se lo state usando con la scala giusta.

Propagazione dell'errore massimo

Sia $y = f(x_1, \ldots, x_n)$ una grandezza che dipende dalle grandezze $x_1, \ldots, x_n$ delle quali conosciamo una stima $\hat{x}_i$ e l'errore massimo Δx_i. Una prima rozza stima dell'errore massimo su y è data dalla semidispersione sul valore massimo e il valore minimo che possono essere assunti da y

$$\Delta y = \left| \frac{f(\hat{x_1} + \Delta x_1, \ldots, \hat{x_n} + \Delta x_n) - f(\hat{x_1} - \Delta x_1, \ldots, \hat{x_n} - \Delta x_n)}{2} \right|$$

Nota che la formula è vera soltanto se la y cresce al crescere di tutte le x_i; se così non è, i termini al numeratore nell'equazione non sono i valori massimi e minimi di y e di conseguenza questa formula perde il significato che volevamo darle.

Per ottenere una stima più ragionevole di come si propaga l'errore massimo possiamo sfruttare l'andamento locale della funzione f, nell'ipotesi, più che ragionevole nella maggior parte dei casi, che i vari Δx_i siano sufficientemente piccoli. Sviluppando in serie di Taylor al primo ordine è facile convincersi che l'errore massimo associato alla grandezza y è

$$\Delta y = \left|\frac{\partial f}{\partial x_1}\right| \Delta x_1 + \left|\frac{\partial f}{\partial x_2}\right| \Delta x_2 + \cdots + \left|\frac{\partial f}{\partial x_n}\right| \Delta x_n \tag{2}$$

La formula può sembrare laboriosa, ma nella maggior parte dei casi si ha che una delle quantità x_i ha associata un'incertezza molto più alta delle altre e quindi quando si fa il calcolo si può dimenticare tutte le altre in quanto comunque ricadrebbero dentro la barra di errore di quella dominante. Usate sempre il buonsenso per capire cosa fare.

Per fissare le idee facciamo un esempio. Supponiamo di aver misurato un angolo e di aver ottenuto $\theta = \hat{\theta} \pm \Delta\theta = (0.74 \pm 0.09)\,\text{rad}$ e di voler calcolare il seno con relativo errore massimo. Applicando la formula

$$\Delta(\sin\theta) = \left|\frac{\mathrm{d}\sin\theta}{\mathrm{d}\theta}(\hat{\theta})\right| \Delta\theta = \left|\cos\hat{\theta}\right| \Delta\theta$$

Sostituendo ora si ottiene $\sin\theta = 0.67 \pm 0.07$. Discutiamo brevemente cosa succede nel caso in cui sia invece $\theta = (1.57 \pm 0.01)\,\text{rad}$. Ho scelto proprio questo valore in quanto $\cos\hat{\theta} \approx 0$, proprio perché $\hat{\theta} \approx \pi/2$. Possiamo dire che l'errore sul seno dell'angolo è zero? Ovviamente no, ma ci aspettiamo che sia il minimo errore che possiamo ottenere. Se vogliamo dare una stima del massimo errore, ricorriamo di nuovo alla serie di Taylor e approssimiamo al primo ordine non nullo.

$$\Delta(\sin\theta) \approx \left|\frac{\mathrm{d}\sin\theta}{\mathrm{d}\theta}(\hat{\theta})\right| \Delta\theta + \frac{1}{2}\left|\frac{\mathrm{d}^2\sin\theta}{\mathrm{d}\theta^2}(\hat{\theta})\right| (\Delta\theta)^2 \approx \frac{(\Delta\theta)^2}{2}$$

Questo era un esempio puramente illustrativo sul cosa fare nel caso in cui il primo ordine sia esattamente zero. Nel caso concreto, difficilmente si giunge ad una situazione simile. Inoltre, dato che il primo ordine non è mai *esattamente* zero, proprio perché difficilmente una misura in laboratorio ha come esito π, è sempre opportuno confrontare il primo ordine con il secondo.

1.3 Espressione delle misure nelle sperimentali delle Olimpiadi

Nella maggior parte delle situazioni che si possono incontrare nelle sperimentali olimpiche la miglior stima è fornita dalla media e l'incertezza dalla semidispersione. Se invece una misura è stata ripetuta un numero staticamente sigificativo (diciamo > 7) allora si può prendere in considerazione l'*errore statistico*, come verrà spiegato nella sezione 2.6.

Cifre significative

Cosa vuol dire "cifre significative"? È una stima molto preliminare della bontà di una misura. In particolare è un numero intero che ci dovrebbe dire

"con quante cifre conosciamo un numero". Questa è una definizione un po' vaga in quanto, con l'esempio che faremo subito, vedremo che è opportuno definirla in modo più rigoroso. Infatti, in matematica è evidente che i numeri 300×10^2, 30000 e 3×10^4 sono tutti uguali, mentre nell'ambito della fisica sperimentale e nell'analisi dei dati hanno tutti significati diversi. Quello che si sottintende infatti è che tutte le cifre che vengono scritte che non fanno parte della notazione esponenziale, ovvero quello che viene prima del $\times 10^{\text{qualcosa}}$, *si è sicuri di saperle*. Moralmente, il significato che il fisico dà alle espressioni precedenti è $(30 \pm 1) \times 10^2$, $30\,000 \pm 1$, $(3 \pm 1) \times 10^4$, che sono effettivamente stime molto diverse.

In particolare, la prima stima ha 2 cifre significative, la seconda 5, la terza una sola. Il discorso si estende in modo uguale alle cifre dopo la virgola, infatti 10.0 e 10 hanno rispettivamente 3 e due cifre significative, portando quindi a stime diverse. A questo punto possiamo quindi formalizzare leggermente meglio il concetto di cifra significativa:

1. La cifra più significativa è quella più a sinistra e diversa da 0;
2. La cifra meno significativa è quella più a destra (fatta eccezione per gli zeri nei numeri interi);
3. tutte le cifre comprese tra la più significativa e la meno significativa sono significative.

Cifre significative ed errore. Dovete sempre ricordare che una misura deve avere il corretto numero di cifre significative per non essere presa per scherzo da chi la legge. Una misura senza errore non è credibile, una misura con il numero sbagliato di cifre significative è insensata e viene punita. Ricordate che voi non conoscete l'errore sulla misura, in quanto se lo conosceste potreste sottrarlo ed ottenere il valore vero, mentre voi siete solo in grado di dare una stima di questo errore, per cui anche l'errore dovrebbe avere un suo errore associato. Ovviamente questo nel 99% dei casi non si fa e semplicemente si indica l'errore con il numero giusto di cifre significative, intendendo che "l'errore sull'errore è sull'ultima cifra".

In pratica, per fare un esempio concreto, se indico $l = (43 \pm 2)\,\text{cm}$, vuol dire che l ha due cifre significative, mentre l'errore su l ha una sola cifra significativa. In generale, come regola della nonna, non indicate mai più di due cifre sull'errore di una misura. Alle Olimpiadi in particolare è difficile ottenere una misura che abbia davvero due cifre significative, indicarne due o più sull'errore è prendersi in giro nella maggior parte dei casi.

Lista di cose da non scrivere

- Una misura senza incertezza, a meno che non sia espressamente richiesto.
- Una misura con incertezza ed errore con un numero non coerente di cifre significative. Un paio di esempi da **non** fare:
    1. 30.2 ± 0.11 m Questo è sbagliato perché sto indicando l'errore con due cifre dopo la virgola mentre indico la misura con una sola. Una delle due affermazioni è falsa
    2. 30.22 ± 0.1 m Questo è falso perché adesso la misura ha più cifre dell'errore. L'unico caso in cui si ammette una scrittura simile e nel caso della "mezza tacca", ovvero quando si ha 30.25 ± 0.1 m. Io di solito indico la cifra 5 fra parentesi in quanto vorrebbe dire "sono a metà, questa cifra è falsa ma non mi sbilancio né da una parte né dall'altra perché non so dove andare", tuttavia è una convenzione che usano in pochi e spesso viene fraintesa, quindi in sostanza la sconsiglio.
    3. 30.333 ± 0.123 m. Qui semplicemente ho indicato un errore con 3 cifre, cosa assolutamente inverosimile.

1.4 Esercizi

Problema 1.1 (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 fatte precedentemente. Ricorda di esprimere il risultato con il giusto numero di cifre significative.

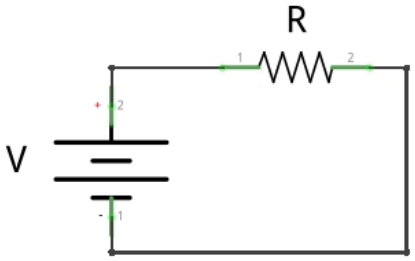
Problema 1.2 (Gara di macchinine)**.** Lanciamo tre macchinine ideali lungo un piano, inclinato di un angolo ignoto α. Il punto di partenza si trova ad $h = (100 \pm 1)$ cm sopra l'arrivo (ovvero il dislivello è di h). Gli attriti sono trascurabili, le ruote delle macchinine hanno massa molto minore rispetto a quella di un'intera macchinina. Le macchinine partono ferme. Vengono effettuate le seguenti misure di velocità al traguardo:

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

Macchinina	Verde	Bianca	Rossa
Velocità [m/s]	4.47	4.42	4.35

Tabella 1: dati del Problema 1.2.

Problema 1.3 (Prima legge di Ohm: occhio alle derivate!)**.** Si consideri il circuito sottostante, in cui il generatore eroga una tensione $V = (24.0 \pm 0.2)\,\mathrm{V}$ e la resistenza ha valore $R = 2.2\,\mathrm{k\Omega} \pm 10\%$.



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 Equazione 2. Ricavare algebricamente tali differenze.

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

2 Distribuzione normale ed errore statistico

2.1 Distribuzioni di probabilità, versione supercondensata

La teoria della probabilità gioca un ruolo fondamentale nell'analisi dei dati, in particolare quando bisogna fare dei fit, che vedremo in Sezione 3. In questa

breve nota non saremo rigorosi e cercheremo di dare una visione intuitiva di alcuni concetti chiave, dando delle definizioni concrete e non partendo dagli assiomi di Kolmogorov, cosa che prima o poi va fatta, ma non è questo il momento adatto. Questa sezione, fino ai metodi di fit, sarà molto teorica e dovrebbe dare delle giustificazioni per quello che si farà in seguito. Il lettore non interessato può tranquillamente saltare tutta questa zona e rimandarla a mai più per concentrarsi sui metodi di fit.

In queste note la probabilità sarà vista come probabilità *frequentista*, ovvero semplicemente

$$P(\text{Evento}) = \frac{\#\text{Casi favorevoli}}{\#\text{Casi totali}}$$

Questa nozione è abbastanza intuitiva per probabilità discrete. Per esempio, la probabilità che esca un 2 lanciando una volta un dado a sei facce bilanciato è ovviamente 1/6. È opportuno estendere questa nozione anche a degli oggetti che possono assumere un insieme continuo di valori. Per fare un esempio, potremmo considerare delle estrazioni di un numero casuale compreso fra 0 e $m > 0$. Supponiamo che la distribuzione sia uniforme (fra poco formalizzeremo meglio il concetto) e facciamo un paio di considerazioni. La probabilità di ottenere un numero reale $x \in [0, 1]$, strettamente parlando è zero, in quanto "abbiamo un caso solo e stiamo dividendo per un numero infinito di casi possibili".

Ha più senso porsi la domanda "quant'è la probabilità che un numero estratto sia compreso in un certo intervallo?". La probabilità di questo evento, fissato l'intervallo $I = [x_1, x_2]$, con $x_1, x_2 \in [0, m]$, ci aspettiamo sia finita e in generale diversa da 0. In particolare è abbastanza ovvio pensare che la probabilità sia proporzionale a $x_2 - x_1$. Dato che abbiamo supposto la probabilità uniforme e dato che la somma delle probabilità deve fare 1, varrà $P(x \in I) = \frac{1}{m}(x_2 - x_1)$.

Come formalizziamo il caso di probabilità non uniforme? Beh, qualsiasi funzione continua è costante se la guardiamo abbastanza da vicino. Se ora consideriamo quindi un intervallo infinitesimo, ovvero consideriamo $x_2 = x_1 + \mathrm{d}x$, è ragionevole pensare che $P(x \in I) = p(x_1)\,\mathrm{d}x$, dove $p(x)$ è una funzione che dipende per l'appunto da come è distribuita la variabile x.

Quali sono le proprietà che deve rispettare $p(x)$? In qualche modo deve descrivere una probabilità, quindi dobbiamo chiederci che assiomi deve rispettare la probabilità. In questo caso ci basta imporre che la probabilità sia sempre ≥ 0 e che la somma di tutte le probabilità faccia 1. Il che si tramuta nel richiedere

$$p(x) \geq 0 \quad \forall x \qquad \int_{\mathcal{X}} p(x)\,\mathrm{d}x = 1 \tag{3}$$

Dove con $\mathcal{X}$ si indica l'insieme su cui è definita $p(x)$, di solito l'intero asse reale. Di conseguenza, la probabilità che la variabile x sia compresa fra x_1 e x_2 si potrà esprimere in questo caso come

$$P(x \in [x_1, x_2]) = \int_{x_1}^{x_2} p(x)\,\mathrm{d}x$$

Nel caso di distribuzione uniforme come prima, ci riconduciamo a

$$p(x) = \begin{cases} \frac{1}{m} & \text{se } x \in [0, m] \\ 0 & \text{altrimenti} \end{cases}$$

Definizione 2.1 (Distribuzione di probabilità continua)**.** Una qualsiasi funzione che rispetti le proprietà in Equazione 3 è una distribuzione di probabilità continua, chiamata anche *densità di probabilità*.

A questo punto è opportuno definire cos'è una *variabile aleatoria*. La definizione formale di questo concetto presuppone la conoscenza della teoria della misura e permette di estenderlo anche ad oggetti molto generali, ma noi ci focalizzeremo su qualcosa di concreto per rimanere concentrati sull'obiettivo della lezione. Informalmente, una variabile aleatoria è un evento il cui esito non è prevedibile a priori ma di cui è definita una distribuzione di probabilità discreta o continua. Ogni volta che si realizza questo evento, si dice che si fa una *estrazione* di questa variabile, che indicheremo con X.

Per esempio, una variabile aleatoria è il risultato del lancio di un dado a sei facce. Fare un'estrazione vuol dire lanciare il dado e vedere cosa esce. I vari eventi sono: "è uscito il numero 1", "è uscito il numero 2", eccetera, e di ognuno di loro si sa dire quanto vale la probabilità.

Solitamente delle variabili aleatorie si fa una *parametrizzazione*, ovvero una funzione $f : \{\text{Spazio degli eventi}\} \to \mathbb{R}$. Per molte variabili in un certo senso la parametrizzazione è naturale, ovvero la corrispondenza "è uscito il numero 2" $\to 2$ è quasi ovvia ed è la più sensata da fare, ma è ovviamente non unica, in quanto nessuno ci vieta di far corrispondere "è uscito il numero x" a $x + 1$. Il punto della questione è che **la probabilità deve dipendere dall'evento e non dalla sua parametrizzazione.**

2.2 I cambi di variabile

Come si comportano le distribuzioni di probabilità continue per cambio di variabile? Facciamo un esercizio a titolo di esempio per capire come il

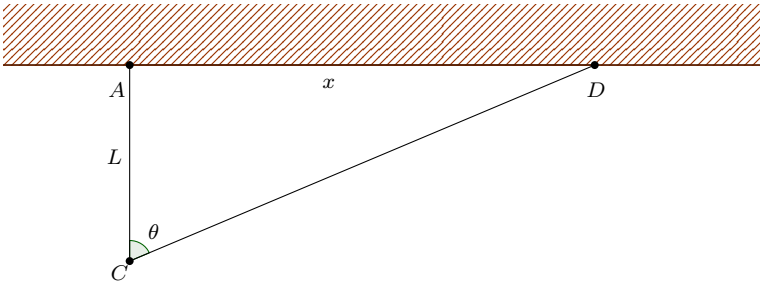


Figura 1: cannone che ruota.

cambio di variabile possa non essere completamente banale. Consideriamo un cannone fisso nel punto C che può ruotare solo nel piano orizzontale e sparare solo in un semipiano. Il cannone sparerà randomicamente con distribuzione uniforme nell'angolo $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. Si veda la Figura 1 per capire meglio la situazione.

Ad una distanza L dal cannone c'è un muro piatto. I colpi sul muro non arriveranno in modo uniforme, è lecito aspettarsi che nella zona vicina al cannone ne arrivino molti di più che nelle zone molto lontane. Si consideri la variabile x come in Figura 1, ovvero la lunghezza del segmento AD. La domanda di questo problema è: qual è la probabilità di essere colpiti da un proiettile occupando l'intervallo $[x, x + \mathrm{d}x]$? In altre parole, qual è la distribuzione di probabilità nella variabile x?

Per rispondere a questa domanda e capire cosa sta succedendo è importante capire che la probabilità dipende dall'evento e non dalla parametrizzazione dello stesso. Di conseguenza, se noi diamo una relazione bigettiva fra θ e x (fra pochissimo la scriveremo, è puramente trigonometrica), dovrà accadere

$$p(\theta)\,\mathrm{d}\theta = p(x(\theta))\,\mathrm{d}x(\theta)$$

È importante anche notare che in questa formula ci sono i differenziali $\mathrm{d}x$ e $\mathrm{d}\theta$, ovvero gli intervallini che rappresentano nelle due parametrizzazioni la stessa porzione dello spazio degli eventi, in quanto noi stiamo dicendo che sono uguali le *probabilità*, non le *funzioni che le rappresentano*. A questo punto abbiamo tutti gli ingredienti e possiamo fare il conto, una volta data la relazione bigettiva fra θ e x. Questa è semplicemente $x = L\tan\theta$. Dobbiamo inserire questa relazione nell'equazione precedente. Per chiarezza, calcoliamoci prima la trasformazione dei differenziali

$$\mathrm{d}x(\theta) = \frac{\mathrm{d}x(\theta)}{\mathrm{d}\theta}\,\mathrm{d}\theta = L\frac{\mathrm{d}}{\mathrm{d}\theta}\tan\theta\,\mathrm{d}\theta = L(1+\tan^2\theta)\,\mathrm{d}\theta$$

Inoltre, dato che la distribuzione di probabilità in θ è uniforme, $p(\theta) = 1/\pi$. Inserendo il tutto nell'equazione precedente,

$$\frac{1}{\pi}\,\mathrm{d}\theta = Lp(x)(1+\tan^2\theta(x))\,\mathrm{d}\theta$$
$$\Downarrow$$
$$p(x) = \frac{1}{\pi L}\frac{1}{1+\tan^2\theta} = \frac{1}{\pi L}\frac{1}{1+(\frac{x}{L})^2}$$

Per cui la probabilità di essere colpiti da un proiettile stando nell'intervallo $[x, x+\mathrm{d}x]$ è semplicemente

$$P(\text{Evento}) = \frac{1}{\pi}\frac{1}{1+(\frac{x}{L})^2}\frac{1}{L}\,\mathrm{d}x = p(x)\,\mathrm{d}x$$

Consiglio come esercizio di riflettere su cosa succede quando la corrispondenza $x(\theta)$ non è univoca. Puramente a titolo di esempio, pensate cosa succede quando si va a considerare $\theta(x) = x^2$ e riflettete su cosa cambia nel nostro ragionamento, anche senza effettuare i calcoli.

2.3 Media e varianza

Per definizione, noi non conosciamo a priori il risultato di un'estrazione di una variabile casuale, sappiamo solo dare delle informazioni "in media" su quello che può accadere, a partire dalla distribuzione di probabilità che supponiamo di conoscere.

Per esempio, se lanciamo sei milioni di volte un dado a sei facce, ci aspettiamo mediamente di aver ottenuto in media un milione di volte la faccia con sopra l'uno e via dicendo. Di conseguenza, il "numero medio"[3] che prevediamo di ottenere è 3.5. Notare che questa previsione dipende solo dalla distribuzione di probabilità e non dai dati ottenuti, in quanto, tautologicamente, è una previsione e non una misura.

Dato che l'accordo fra previsione e misura è un cardine della Fisica, è opportuno formalizzare un po' meglio queste definizioni vaghe in modo da poter elaborarci sopra.

Consideriamo quindi una variabile aleatoria X, una sua parametrizzazione x con distribuzione di probabilità $p(x)$. Consideriamo una generica funzione $f(x)$ e definiamo il *valore atteso di* $f(x)$, (indicato con $\mathbb{E}_X[f(x)]$) nel seguente modo:

[3] Ovvero la media di tutti i numeri che abbiamo ottenuto.

$$\mathbb{E}_X[f(x)] = \int_{\mathcal{X}} p(x) f(x)\, \mathrm{d}x$$

Il valore atteso è quindi un semplice numero reale che si ottiene a partire da una funzione $f(x)$ e da una distribuzione di probabilità $f(x)$. Alcune proprietà elementari che useremo fra poco sono riportate in appendice. Il *valor medio* che abbiamo definito informalmente prima con l'esempio dei dadi è quindi $\mu_1(X) = \mathbb{E}_X[x]$. In qualche modo possiamo vedere questo oggetto appena definito come una "media pesata" di ogni valore con la sua probabilità e di conseguenza ci aspettiamo che rappresenti la media dei valori che andremo ad ottenere se estraiamo tante volte la variabile X. Questo è un oggetto che è definito indipendentemente dalla distribuzione di probabilità $p(x)$ e assumerà valori diversi proprio in funzione di questa probabilità $p(x)$. In particolare, $p(x)$ è l'unica cosa da cui dipende $\mu_1(X)$.

Un'altra informazione a cui possiamo essere interessati è "quanto sono dispersi i nostri dati". Consideriamo il seguente esempio, molto banale: due studenti hanno preso i seguenti voti nelle verifiche di Fisica: $\{6, 6, 6\}$, $\{6, 3, 9\}$. È evidente che i due insiemi hanno la stessa media, ma è altrettanto evidente che uno dei due studenti è molto costante, mentre l'altro sembra imprevedibile. Come stimiamo quindi con un numero solo questa dispersione? La risposta è ovviamente estremamente arbitraria e ci sono milioni di espressioni che possono fare questo lavoro. Tuttavia ci sono dei motivi per cui una di queste è particolarmente significativa e viene chiamata *varianza*, $\mathrm{Var}(X)$, che è definita nel seguente modo

$$\mathrm{Var}(X) = \mathbb{E}_X[(x - \mu_1(X))^2]$$

Notare che il quadrato dentro il valore atteso è fondamentale. Se considerassimo infatti la stessa quantità senza il quadrato, $\mathbb{E}_X[x - \mu_1(X)] = \mathbb{E}_X[x] - E[\mu_1(X)] = \mu_1(X) - \mu_1(X)E[1] = 0$. Una cosa che fa zero sempre, indipendentemente dalla distribuzione di probabilità, è evidentemente poco utile in termine di informazioni. Avremmo potuto scegliere un'espressione diversa, mettendo per esempio un valore assoluto al posto del quadrato. Questa è in effetti una scelta ragionevole, ma per diversi motivi, fra cui la facilità nel fare i conti, si preferisce l'oggetto che ho definito. Notiamo che possiamo scrivere la varianza in modo diverso, utilizzando le proprietà di $\mathbb{E}_X$:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}_X[(x-\mu_1(X))^2] = \mathbb{E}_X[x^2+\mu_1(X)^2-2x\mu_1(X)] && (4)\\
&= \mathbb{E}_X[x^2]-2\mu_1(X)\mathbb{E}_X[x]+\mu_1(X)^2 && (5)\\
&= \mathbb{E}_X[x^2]-2\mu_1(X)^2+\mu_1(X)^2 && (6)\\
&= \mathbb{E}_X[x^2]-\mu_1(X)^2 = \mathbb{E}_X[x^2]-(\mathbb{E}_X[x])^2 && (7)
\end{aligned}$$

Notare che, per come è definita la varianza, questa è sicuramente maggiore o uguale a zero. In particolare, è uguale a zero se e solo se X può assumere solo il suo valore medio, ovvero è una variabile casuale molto poco casuale.

2.4 Covarianza

Cosa succede quando andiamo a fare delle misure simultanee di oggetti che immaginiamo essere *correlati*? Proviamo a descrivere la situazione utilizzando il linguaggio della teoria della probabilità e delle variabili aleatorie. Supponiamo quindi di avere due osservabili, che saranno quindi rappresentabili da due variabili aleatorie X e Y. Dire che le due variabili sono correlate vuol dire che quando estraiamo un risultato, questo dipende da tutte e due le misure e non da una sola, ovvero che le due variabili saranno rappresentabili da distribuzioni di probabilità che dipendono da x e da y simultaneamente e in modo non fattorizzabile, ovvero $p(x,y) \neq p(x)p(y)$. Rimaniamo con i piedi per terra e facciamo un esempio concreto. Consideriamo una partita a Monopoli: fissata la plancia (ovvero la posizione delle pedine e le proprietà dei giocatori), due variabili estremamente correlate sono per esempio X, l'estrazione del lancio dei due dadi a sei facce e Y, i soldi che devo pagare alla persona che possiede il terreno su cui finirò dopo essermi spostato di X passi in avanti.

Questo è un caso limite in cui le due variabili sono direttamente e completamente correlate, ovvero conoscendo X, so immediatamente il valore di Y, in quanto è semplicemente determinato dalle regole del gioco. Due variabili meno correlate sono invece la quantità di cereali che mangio a colazione e il mio peso corporeo. È evidente che se ogni mattino mangio un kilogrammo di cereali il mio peso corporeo aumenterà di conseguenza, ma la colazione non è l'unico pasto della giornata, quindi il mio peso non è completamente determinato dalla prima variabile.

Ora che abbiamo fatto un esempio stupido, possiamo cercare di fare delle affermazioni quantitative. Date quindi due variabili aleatorie X e Y correlate, ovvero (data una parametrizzazione) fornita una probailità $p(x,y)$, quale sarà la probabilità che facendo un'estrazione si ottenga un valore compreso nell'intervallo $[x, x+\mathrm{d}x] \times [y, y+\mathrm{d}y]$? Ovviamente questo varrà $p(x,y)\,\mathrm{d}x\,\mathrm{d}y$, semplicemente dalla definizione.

Di conseguenza, se non mi interesso di Y e guardo solo x, quale sarà la probabilità di ottenere un valore di X nell'intervallo $[x, x + \mathrm{d}x]$? Questo significa che va bene qualsiasi valore di y, ovvero possiamo *marginalizzare*, ovvero integrare sulla variabile non interessante

$$P(x \in [x, x + \mathrm{d}x]) = p(x)\,\mathrm{d}x = \int_{\mathcal{Y}} p(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

Dove con la notazione precedente si intende che l'integrale si fa solo su y, mentre il $\mathrm{d}x$ serve perché stiamo considerando un intervallo su x infinitesimo. In questo modo, abbiamo una probabilità solo su X e acquista quindi senso il concetto di valore atteso di qualcosa $\mathbb{E}_X[\cdot]$. Ovviamente il discorso che ho fatto è perfettamente simmetrico e si applica tale e quale a Y.

Cerchiamo ora di dare una misura di quanto due variabili sono correlate. Un oggetto che viene usato spesso, ma ha i suoi grossi difetti, è la *covarianza*, definita in modo semplice, molto simile alla varianza, come

$$\mathrm{Cov}(X, Y) = \mathbb{E}_{X,Y}[(X - \mathbb{E}_X[X])(Y - \mathbb{E}_Y[Y])] \tag{8}$$

Assomiglia ad una varianza ma è fatta con due variabili diverse. Con facili manipolazioni si può trovare una forma equivalente che ricavo qui sotto, che renderà evidente quale sarà la scelta appropriata dello stimatore, che vedremo fra un paio di paragrafi.

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathbb{E}_{X,Y}[XY - \mathbb{E}_X[X]Y - \mathbb{E}_Y[Y]X + \mathbb{E}_X[X]\mathbb{E}_Y[Y]] \\
&= \mathbb{E}_{X,Y}[XY] - \mathbb{E}_{X,Y}[\mathbb{E}_X[X]Y] - \mathbb{E}_{X,Y}[\mathbb{E}_Y[Y]X] + \mathbb{E}_X[X]\mathbb{E}_Y[Y] \\
&= \mathbb{E}_{X,Y}[XY] - \mathbb{E}_X[X]\mathbb{E}_{X,Y}[Y] - \mathbb{E}_Y[Y]\mathbb{E}_{X,Y}[X] + \mathbb{E}_X[X]\mathbb{E}_Y[Y] \\
&= \mathbb{E}_{X,Y}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y] - \mathbb{E}_Y[Y]\mathbb{E}_X[X] + \mathbb{E}_X[X]\mathbb{E}_Y[Y] \\
&= \mathbb{E}_{X,Y}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y]
\end{aligned}$$

Che è una sorta di "media del prodotto meno il prodotto delle medie". Se le due variabili non sono scorrelate, in generale questo oggetto non sarà zero. A differenza della varianza, però, questo può tranquillamente essere negativo.

Attenzione! È vero che due variabili scorrelate hanno correlazione zero, ma non è vero il viceversa. Dire quindi che la covarianza è una misura della correlazione non è sempre corretto e va sempre preso con le pinze in quanto in qualche situazione strana possono comparire risultati poco intuitivi. Prendiamo come esempio per mostrare quello che abbiamo appena detto una variabile aleatoria discreta X che vale -1 o $+1$ con probabilità $\frac{1}{2}$. Prendiamo poi un'altra variabile aleatoria Y tale che $Y = 0$ se $X = -1$, e Y sia

randomicamente -1 o $+1$ con probabilità $\frac{1}{2}$ se $X = 1$. Chiaramente X e Y sono estremamente dipendenti, in quanto conoscendo Y sappiamo con certezza quanto vale X, ma la loro covarianza è zero: tutti e due hanno media zero e

$$\begin{aligned}\mathbb{E}[XY] &= (-1)\cdot 0 \cdot P(X=-1) \\ &\quad + 1 \cdot 1 \cdot P(X=1, Y=1) \\ &\quad + 1 \cdot (-1) \cdot P(X=1, Y=-1) \\ &= 0.\end{aligned}$$

per cui, dalla definizione di covarianza si vede che anche questa è zero.

2.5 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[4]. 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[5]. La curva è descritta dalla seguente densità di probabilità:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{9}$$

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.

[4]Ci sono degli ottimi motivi che discendono da principi sensati per cui questa curva è così importante. A rigor di logica, con le affermazioni precedenti qualsiasi curva a campana sembrerebbe andare bene, come per esempio una distribuzione lorenziana $p(x) = \frac{1}{\pi}\frac{1}{1+x^2}$. Tuttavia, vi sorprenderò con la seguente supercazzola: la gaussiana è una curva che compare così spesso in teoria della probabilità perché è la funzione che, a media e varianza finite e fissate, massimizza il funzionale entropia $S[p(x)] = -\int p(x)\log p(x)\,\mathrm{d}x$. Inoltre, è pure un'autofunzione della trasformata di Fourier.

[5]Si intende la distribuzione del singolo dato attorno al "valore vero".

2.6 Stimatori

Affrontiamo ora il seguente problema concreto: supponiamo di sapere ed essere in qualche modo convinti che i dati che abbiamo raccolto seguano una certa distribuzione di probabilità $p(x, \vec{\theta})$, che dipende da dei parametri, che chiameremo $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$. Supponiamo di aver raccolto una serie di dati sperimentali della stessa quantità, avendo ottenuto degli esiti $\vec{x} = (x_1, x_2, \ldots, x_n)$. Una domanda legittima è chiedersi come ottenere una *stima consistente* dei vari parametri a partire dai nostri dati sperimentali. Inoltre, tutti sanno che una misura fornita senza errore è una misura a cui non crede nessuno, quindi in un certo senso dobbiamo anche trovare una *stima sensata* degli errori da attribuire alla nostra stima. Diamo quindi la definizione informale di stimatore.

Definizione 2.2 (Stimatore)**.** Dato un modello, ovvero supposto che i nostri dati sperimentali seguano una distribuzione di probabilità $p(x, \vec{\theta})$ che dipende da dei parametri $\vec{\theta}$, uno stimatore è una funzione dalle $n-$uple di dati sperimentali allo spazio dei parametri $\vec{\theta}$. Un buon stimatore è uno stimatore che nel limite di tante misure tende in probabilità al "valore vero" dei nostri parametri. In sostanza, stiamo dicendo che uno stimatore è una espressione che a partire dai dati sperimentali fornisce una stima sensata dei parametri ignoti di una curva.

Lo stimatore, essendo una funzione di variabili aleatorie, è a sua volta una variabile aleatoria, per cui possiamo domandarci quanto vale per ogni stimatore il suo valore atteso e la sua varianza. Sarà meglio che il valore atteso dello stimatore tenda, almeno per il limite di tante misure, al valore vero dello stimatore.

Esempio 2.1. Consideriamo un esempio concreto. Supponiamo di prendere una scatola piena di chiodini, che dovrebbero essere tutti uguali, ma il macchinario che li ha prodotti è difettoso e la lunghezza dei vari chiodini non è uniforme. Supponiamo che i nostri dati sperimentali siano distribuiti secondo una curva gaussiana, ovvero che la nostra $p(x)$ sia $p(x, \mu, \sigma)$, con $\vec{\theta} = (\mu, \sigma)$ e supponiamo anche che la σ sia maggiore della nostra risoluzione sperimentale, ovvero che con il righello o il calibro che andremo ad utilizzare sia chiaramente distinguibile un chiodino fuori dalla media. Noi andremo a raccogliere dei dati, che chiameremo $\vec{x} = (x_1, x_2, \ldots, x_n)$ e supporremo che le misure siano completamente scorrelate. Di conseguenza, la probabilità di ottenere il vettore $\vec{x}$ sarà semplicemente il prodotto delle probabilità di ottenere ogni singola componente, ovvero

$$p(\vec{x}, \mu, \sigma) = \prod_{i=1}^{n} p(x_i, \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right]$$
$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=0}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2\right]$$

Abbiamo a questo punto molti modi per scegliere uno stimatore dei parametri μ e σ, ma una soluzione sensata può essere la coppia di valori $(\hat{\mu}, \hat{\sigma})$ tale che la probabilità precedente sia massima. In un certo senso è una scelta quasi naturale, in quanto se ci aspettiamo che il modello, ovvero la scelta di $p(x)$, sia quello corretto, ci aspetteremmo che la probabilità sia alta. Scegliere i parametri $\hat{\mu}, \hat{\sigma}$ in modo che questa sia massima, sembra la cosa più sensata da fare. Questo modo di approcciare il problema si chiama *principio di massima verosimiglianza* ed è generalmente un metodo (quasi) costruttivo che permette a partire da un'idea molto generale di trovare il valore *migliore* dei parametri.

A questo punto abbiamo quindi una funzione di due variabili e non più di $n + 2$ e dobbiamo massimizzarla[6]. Di solito questa funzione si chiama "likelihood" oppure "verosimiglianza" e si indica con $\mathcal{L}_{\vec{x}}(\mu, \sigma)$, in cui $\vec{x}$ è a tutti gli effetti un parametro che ora è fisso.

Come si fa a massimizzare il tutto? La scelta migliore è fare delle derivate parziali e il conto è fatto in appendice. Qui riportiamo i risultati e li commentiamo nel dettaglio.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{10}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2 \tag{11}$$

Vediamo che in Equazione 10 c'è semplicemente la definizione di media aritmetica, il che un po' ci rassicura, in quanto sembra la cosa più ragionevole da fare e in questo caso corrisponde anche allo stimatore di massima verosimiglianza, che ha di solito delle buone proprietà. Ricordiamo che gli stimatori sono delle variabili casuali, quindi avranno anche loro una distribuzione di probabilità che in qualche modo deve dipendere dai valori veri μ e σ.

[6]Perché abbiamo una funzione in meno variabili? Semplicemente perché i vari x_i sono i risultati dell'esperimento e quindi li scegliamo come valori fissati. A partire da questi sceglieremo μ e σ appropriati.

Possiamo quindi chiederci qual è il valore atteso di $\hat{\mu}$, che dovrà quindi essere una funzione solo di μ, σ, n

$$E[\hat{\mu}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Questo è rassicurante, in quanto ci fa sperare che nel limite di tante misure questo stimatore ci dia effettivamente una buona stima del parametro vero. Chiediamoci ora l'altra cosa fondamentale, ovvero l'errore da attribuire a questa misura. Per farlo, possiamo calcolare la varianza dello stimatore $\hat{\mu}$

$$\begin{aligned}\mathrm{Var}[\hat{\mu}] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\,\mathrm{Var}\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[x_i] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}\end{aligned}$$

Per cui uno stimatore sensato dell'errore statistico da attribuire a $\hat{\mu}$ è $\hat{s}_\mu = \frac{1}{\sqrt{n}}\hat{\sigma}$, in quanto se $\hat{\sigma} \to \sigma$, allora $\hat{s}_\mu^2 \to \mathrm{Var}[\hat{\mu}]$. Qui bisogna fare **molta** attenzione al significato che si attribuisce alla formula precedente. Qui stiamo dicendo che l'errore che attribuiamo ad una misura in sostanza cala come $1/\sqrt{n}$. Questo è vero per l'errore statistico, ma non per l'errore sistematico. In un esempio concreto, distinguiamo le due seguenti situazioni:

- Nel primo caso supponiamo di misurare lo stesso tavolo da cucina con un metro da sarta diverse volte. È evidente che a meno di grossi problemi di vista il risultato della misura sarà sempre lo stesso, a meno di un paio di millimetri. È altrettanto evidente che quindi non ha senso pensare di fare un milione di misure per affermare che conosciamo la lunghezza del tavolo al micrometro, in quanto evidentemente questo ci è impedito dai limiti dello strumento. È per questo che negli articoli in cui si fa un'analisi dei dati seria e pensata, di solito si indica il risultato di una misura come $l = 56 \pm 2(\text{stat.}) \pm 3(\text{syst.})\,\text{cm}$, dove le parole fra parentesi indicano rispettivamente statistico e sistematico. Nel nostro caso, è vero che l'errore statistico tenderà a zero facendo molte misure, ma quello sistematico rimarrà dettato dalla risoluzione dello strumento, per cui affermare di conoscere la dimensione del tavolo al micrometro è una balla. In questo caso quindi l'affermazione "l'errore cala con $1/\sqrt{n}$" è proprio dare aria alla bocca.

- Supponiamo di misurare il tempo di caduta di una pallina dal bordo del tavolo. Facciamolo con un cronometro da allenatore di atletica. È evidente che in questo caso il tempo di reazione umano conta molto, in quanto lo strumento che utilizziamo non è il più adatto. Su un paio di secondi di tempo di caduta, i 2/3 decimi di secondo che impiego a premere il pulsante contano davvero e i dati che ottengo facendo molte misure si potranno mettere su un'istogramma che si spera abbia una distribuzione a campana intorno ad un valore sensato. In tal caso, probabilmente l'errore umano è dominante rispetto a quello dello strumento, che potenzialmente misura il centesimo se non il millesimo di secondo, e in tal caso conta davvero fare un sacco di misure per avere una stima più precisa.

A questo punto abbiamo spolpato abbastanza lo studio dello stimatore $\hat{\mu}$. Vediamo di guardare un paio di proprietà di $\hat{\sigma}$, per esempio il suo valore atteso.

$$E[\hat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right] = \frac{1}{n}\sum_{i=1}^{n}E[(x_i - \hat{\mu})^2] = \frac{1}{n}\sum_{i=1}^{n}E[x_i^2 + \hat{\mu}^2 - 2x_i\hat{\mu}]$$
$$= \frac{1}{n}\sum_{i=1}^{n}\left(E[x_i^2] + E[\hat{\mu}^2] - 2E[x_i\hat{\mu}]\right) = \frac{n-1}{n}\sigma^2$$

il dettaglio del calcolo è riportato in appendice. Può non essere così intuitivo il fatto che ci sia quel fattore correttivo, ma possiamo giustificarlo informalmente dicendo che "i gradi di libertà sono uno in meno", ovvero che nell'espressione dello stimatore $\hat{\sigma}^2$ c'è lo stimatore $\hat{\mu}$ e non il valore μ. Questo è in un certo senso un vincolo che imponiamo noi in quanto μ non lo conosciamo, quindi in un certo senso dovremmo dividere per $n-1$ invece che per n. Come commento personale, queste giustificazioni hanno senso solo a posteriori, ovvero non sono molto valide. La risposta è sempre: fai il conto e vedi quanto viene.

Notiamo che stavolta $\hat{\sigma} \to \sigma$ solo davvero nel limite di tante misure, mentre prima il valore atteso di $\hat{\mu}$ era quello corretto anche per un numero finito di misure! Per questo molta gente preferisce lo stimatore

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \tag{12}$$

chiamato *varianza campione*, che è semplicemente lo stesso di prima moltiplicato per il fattore correttivo che sistema questo *bias*. Quale utilizzare

dei due? Meglio il secondo, ma in realtà alle Olimpiadi non cambia davvero molto.

2.7 Propagare l'errore statistico

Spesso è sufficiente propagare le incertezze con il metodo delle derivate parziali illustrato nella Sezione 1.2. Tuttavia sarebbe più rigoroso, in caso di incertezza ottenuta tramite deviazione standard, utilizzare la somma del prodotto di covarianze e derivate parziali.

La *covarianza campione* 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:

$$\hat{S}_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right) \quad (13)$$

Qualitativamente la covarianza campione è uno stimatore della covarianza delle variabili X e Y, ma come abbiamo visto in precedenza ci sono dei casi in cui la covarianza non è un oggetto che rappresenta davvero la correlazione di due variabili, per cui ogni affermazione fatta a partire da questo oggetto va presa un po' con le pinze e pesata opportunamente. *Correlazione senza causalità* è una cosa che può accadere semplicemente per caso e portare a risultati sorprendenti. Consigliamo una visita alla pagina http://www.tylervigen.com/spurious-correlations per farsi un paio di risate come in Figura 2

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

$$S_f = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}S_{x_i,x_j}\right)} \quad (14)$$

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

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

Come si ottiene il risultato in Equazione 14? L'idea è ragionevole ed è la seguente: si calcola il valore atteso del quadrato della variazione di f, ovvero

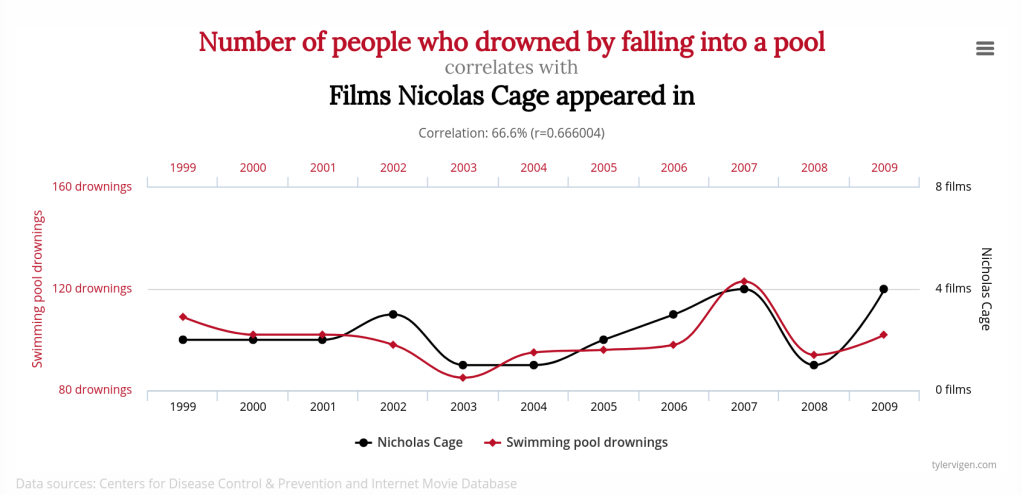


Figura 2: correlazione non implica causalità.

$$
\begin{aligned}
S_f^2 = \mathbb{E}[(\Delta f)^2] &= \mathbb{E}\left[\left(\sum_i \frac{\partial f}{\partial x_i}(\vec{x})\Delta x_i\right)^2\right] = \\
&= \sum_{ij} \mathbb{E}\left[\frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}(\vec{x})\Delta x_i \Delta x_j\right] = \\
&= \sum_{ij} \frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}(\vec{x})\mathbb{E}[\Delta x_i \Delta x_j] = \\
&= \sum_{ij} \frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}(\vec{x}) S_{x_i x_j} =
\end{aligned}
$$

Nel caso di variabili $x_1, \ldots, x_n$ non correlate (quindi con covarianza nulla fra x_i e x_j per $i \neq j$) l'Equazione 14 assume la forma:

$$
S_f = \sqrt{\sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} S_{x_i}\right)^2} \tag{15}
$$

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

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

Problema 2.1 (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.

Problema 2.2 (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

Tabella 2: misure per l'esercizio.

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 $\mathrm{Cov}(X, Z) = 0$. Scrivere t come funzione di x e z. Trovare l'incertezza associata a $\bar{t}$.

3 Metodi di fit

3.1 Cos'è il fit

Una grandezza fisica può essere misurata indirettamente se compare all'interno di un modello fisico che la lega a grandezze che sono misurabili. Supponiamo quindi che x e y siano le due grandezze misurabili legate dalla relazione

$$y = f(x, p_1, \ldots, p_n),$$

dove $p_1, \ldots, p_n$ sono dei parametri incogniti, ma fissi, cioè che non cambiano al variare di x e y. Il *fit* è il processo che consente di risalire a partire da un set di dati (x_i, y_i) ai valori dei parametri che meglio rispettano l'andamento del set.

Passando ad un esempio concreto, si supponga di voler misurare la densità di un materiale avendo a disposizione diversi oggetti tutti costituiti del materiale in questione. Per ognuno di essi si misurano massa m_i e volume V_i. Chiaramente la relazione che lega le grandezze è $m_i = \rho V_i$, con ρ la densità in questione. Volendo ricondurci alla notazione generale chiamiamo x il volume, y la massa e la funzione f sarà semplicemente la relazione di proporzionalità tra le due quantità, dipendente dall'unico parametro ρ. Il problema è quindi trovare la migliore densità che è in accordo con i dati raccolti.

Ovviamente, essendo i parametri dellla funzione f dei valori ottenuti sperimentalmente, anche ad essi è associata un'incertezza, che va opportunamente stimata nel fit.

Il caso più comune e utilizzato[7] 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. Nelle successive sezioni si discuteranno alcuni metodi che si utilizzano per eseguire i fit di rette.

3.2 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 rappresentare 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 semidispersione. In tal caso si ha che:

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

[7] E anche l'unico utile all'interno delle Olimpiadi, dato che, come si vedrà nella Sezione 5.2, tutti i modelli possono essere ricondotti a questo caso.

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

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

3.3 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.

3.4 Metodo dei minimi quadrati ordinari

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

$$h(m, q) = \sum_i (f(x_i) - y_i)^2$$

Il grande pregio di questo metodo è quello di poter essere spesso risolto analiticamente, ovvero il sistema che si ottiene imponendo la condizione ha una soluzione esprimibile nei termini del set di dati[8]. 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 devono essere uguali fra loro affinchè abbia senso la quantità minimizzata.

Per risolvere questo problema l'idea è molto semplice, mentre i calcoli sono laboriosi. L'idea è la seguente: noi abbiamo a disposizione una funzione $h(m, q)$ di due variabili. Per trovarne il massimo, sotto ipotesi di regolarità che sono quasi sempre rispettate, bisogna imporre

$$\begin{cases} \frac{\partial h}{\partial m}(m, q) = 0 \\ \frac{\partial h}{\partial q}(m, q) = 0 \end{cases}$$

Questo è un sistema di due equazioni a due incognite e la soluzione è data da:

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

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

A partire dalle precedenti, propagando gli errori si trova anche una stima sugli errori da attribuire ai parametri; chiamando σ_y l'incertezza su un singolo parametro y si ottiene

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

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

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

dove σ_{mq} è la covarianza sui parametri m e q.

[8]Questo è vero in realtà solamente per fit di modelli lineari. In genearle non è possibile trovare una soluzione analitica per un modello generico, e si ricorre quindi a soluzione numeriche calcolate al computer

3.5 Metodo del minimo χ^2

Qualora le incertezze dei vari y_i non siano uguali fra loro, è intuitavamente evidente che è opportuno 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:

$$\chi^2 = \sum_i \left(\frac{f(x_i) - y_i}{\sigma_{y_i}} \right)^2 \tag{16}$$

Minimizzando la variabile χ^2 si raggiunge proprio lo scopo prima descritto, utilizzando i reciproci dei quadrati delle incertezze delle singole misurazioni come pesi della somma delle differenze $f(x_i) - y_i$. La giustificazione teorica della scelta di questo oggetto come target da minimizzare è data in appendice.

4 Usare la calcolatrice

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

Di seguito verranno descritte alcune delle principali caratteristiche 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.

4.1 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.

Modalità SD oppure 1-VAR

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 $\sum x^2$.

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 di questa differenza. Oltre ai parametri

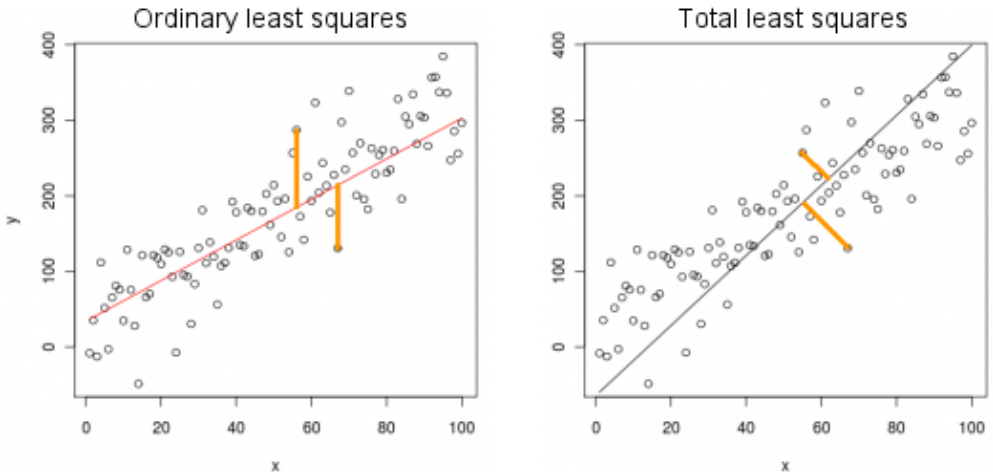


Figura 3: differenza tra minimi quadrati totali e minimi quadrati ordinari. Con gli stessi punti si possono ottenere risultati molto diversi.

della retta la calcolatrice fornisce il *coefficiente di correlazione di Pearson*, un particolare numero dato dalla Equazione 17, che valuta la bontà della vostra regressione lineare misurando la correlazione dei dati inseriti

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \tag{17}$$

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 il 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.

Problema 4.1. 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 3 sono riportati i dati raccolti negli ultimi 8 pasti: Trovare i coefficienti

Persone	Patate servite [quintali]
180	39
135	29
90	9
10	7
50	8
220	35
140	36
120	22

Tabella 3: porzioni di patate.

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 4.

Calcola i nuovi valori del fit con i dati in Tabella 4. Il modello è più attendibile ora?

Persone	Patate servite [quintali]
183	56
135	39
87	28
55	17
59	23
218	65
140	42
241	71

Tabella 4: altre patate.

Problema 4.2. 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 considerando che tanto più $|r|$ è vicino a 0 meno i dati dovrebbero essere correlati.

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.

4.2 Memorie interne

Ogni calcolatrice scientifica possiede almeno 4-5 variabili che possono 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.

4.3 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 a risolverli in maniera efficiente (vedi *Algoritmo di Gauss*).

Alcune calcolatrici offrono anche tabulatori di funzioni: permettono di inserire un'espressione analitica di una funzione ed ottenerne i valori corrispondenti ai numeri compresi fra il primo e l'ultimo valore inserito, a passo costate.

Se avete bisogno di 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.

5 Come affrontare le prove sperimentali di gara

5.1 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 si 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 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 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.

5.2 Linearizzazione

Molto probabilmente il modello che sta alla base dell'esperimento della prova non è lineare, cioè mettendo in un grafico i dati sperimentali che avete misurato vi aspettate di ottenere qualcosa che 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

$$f(x, y) = 0, \tag{18}$$

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

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

dove X e Y sono funzioni tali che la relazione (18) 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.

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

Esempio 5.1. 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:

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

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

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

Nella Tabella 5 sono riportati degli ipotetici dati sperimentali, con le relative trasformazioni. In questo caso siamo interessati solo ai valori prima e dopo la trasformazione e non alle incertezze, per cui non le riportiamo, anche se in linea di principio è sbagliato.

Mettendo in un grafico i punti trasformati si ottiene la Figura 4. I coefficienti della retta di best-fit permettono di trovare che $A^2 = 25.133$ e $A^2 B = 174.315$, da cui si ricava $A = 5.013$ e $B = 6.935$.

Perché la linearizzazone 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, come visto in precedenza. 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. Si veda l'esempio della Figura 5

Quando si fanno esperimenti veri si hanno anche a disposizione strumenti

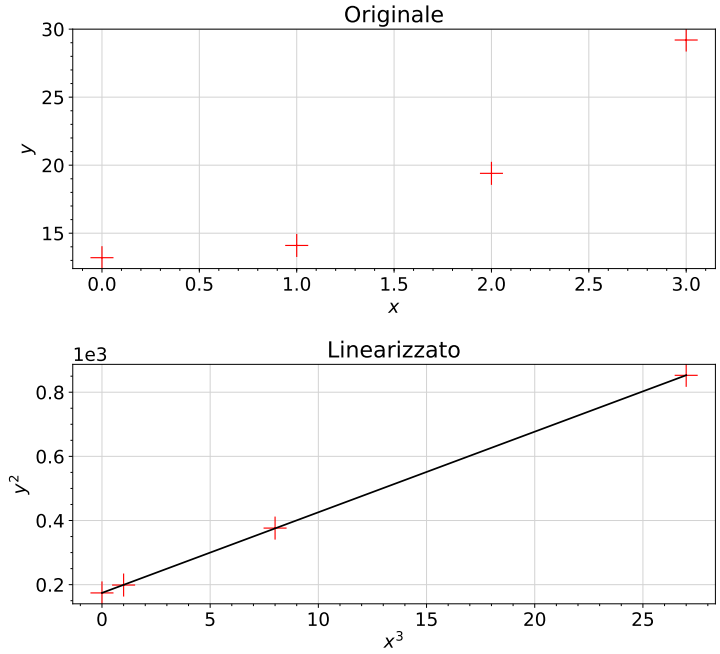


Figura 4: y^2 in funzione di x^3, con la retta di best fit.

di calcolo molto più potenti della 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 famigliarità nel suo utilizzo, perchè molto utile durante la competizione.

Problema 5.1 (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.

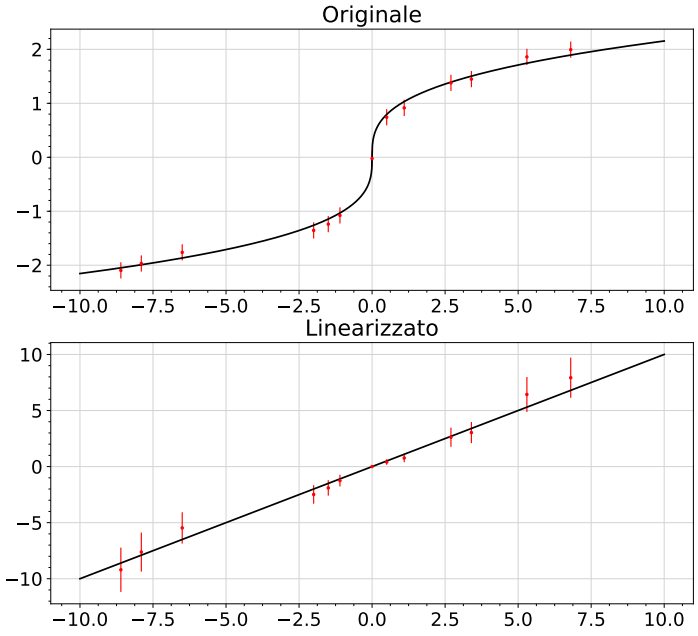


Figura 5: punti sperimentali e relativa curva di best fit nella parte soprastante, e i corrispondenti linearizzati sotto nel caso di relazione $y = Ax^{\frac{1}{3}}$. Le barre di errore non sono più tutte identiche e si espandono notevolmente ai bordi; fittando con i punti linearizzati si otterrebbe un risultato peggiore.

Problema 5.2 (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 6 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.

Problema 5.3 (Tratto da IPhO 1999 - Padova)**.** Noto che $M_2 = (15.0 \pm 0.4)\,\mathrm{g}$, $k = (0.055 \pm 0.001)\,\mathrm{N\,m\,rad^{-1}}$ e la relazione

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

Temperatura [K]	Viscosità μ [mPa $\cdot$ 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 6: dati per l'esercizio 5.2. Notare che questi dati sono senza incertezza, chiaro segno che si tratti di un problema messo in una prova teorica e non frutto di un vero esperimento.

ricavare il valore di I_1 e l a partire dai dati riportati in Tabella 7. I dati sono rappresentati in un grafico in Figura 6, attenzione però a non farsi ingannare! Soluzione: $I_1 = (1.7 \pm 0.7) \times 10^{-4}\,\mathrm{kg \cdot m^2}$ e $l = (230 \pm 20)\,\mathrm{mm}$.

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 7: dati per l'esercizio 5.3.

5.3 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

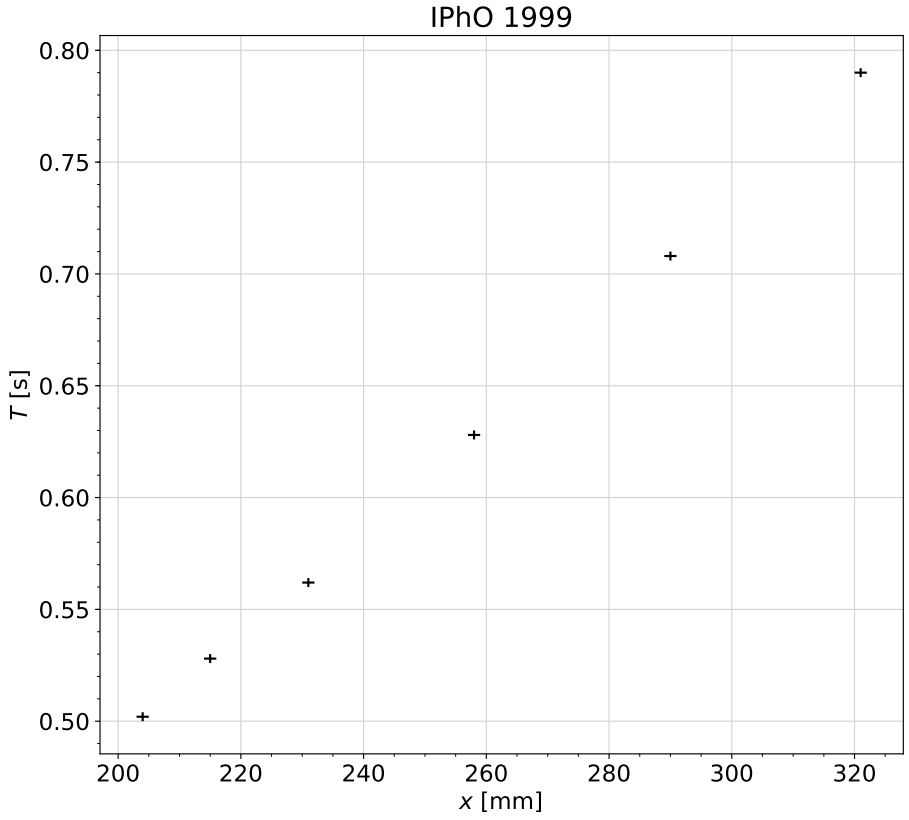


Figura 6: rappresentazione dei dati in Tabella 7. Attenzione: i dati non seguono un modello lineare.

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 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.

5.4 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 spezzate[9];
- 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.

Nella Figura 7 viene riportao un esempio di grafico da consegnare in una gara.

5.5 Altri consigli

Le prove sperimentali richiedono molta 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.
Nelle prove sperimentali italiane (cioè Senigallia e PreIPhO) viene dato molto

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

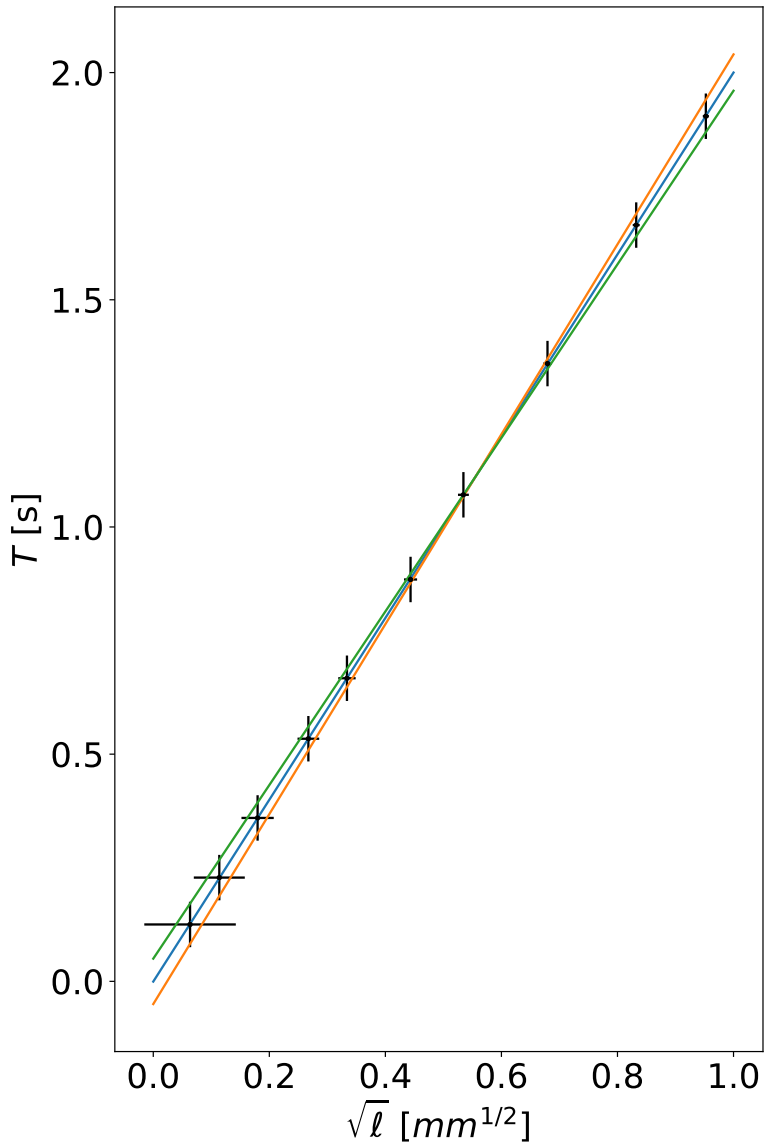


Figura 7: esempio di grafico ben fatto. I dati rappresentano il periodo di un pendolo semplice al variare della lunghezza, quindi seguono la legge $T = 2\pi\sqrt{\frac{\ell}{g}}$. In blu la retta di best fit calcolate con il metodo dei minimi quadrati, in arancione e in verde le rette di massima e minima pendenza rispettivamente. Questo grafico dovrebbe occupare l'intero foglio A4 in una gara vera.

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.

6 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 dovreste 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;
  - Potenziometro;
  - Condensatori: anche con polarità;
  - Diodi e LED;
  - Batteria, generatori di tensione e di corrente.

7 Soluzioni ai problemi

Soluzione 7.1 (Soluzione al Problema 1.1)**.** Le risposte sono abbastanza banali, si tratta di un problema di riscaldamento.

1. La media aritmetica vale $\ell = 28.4\,\mathrm{m}$, che è importante indicare con questo numero di cifre significative e andava pure approssimata per eccesso. L'incertezza che possiamo associare, dato che abbiamo poche misure, può essere la semidispersione $\Delta\ell = 0.8\,\mathrm{m}$, per cui la misura si scrive $\ell = (28.4 \pm 0.8)\,\mathrm{m}$.

2. L'unica cosa a cui bisogna stare attenti in questo caso è indicare il risultato con il numero giusto di cifre significative. L'errore massimo compiuto è $1.0\,\mathrm{m}$ e non si può indicare con più cifre semplicemente perché le misure con senza distanziometro hanno questo numero di cifre.

Soluzione 7.2 (Soluzione al Problema 1.2)**.** Essendoci un dislivello di h, potremo imporre la conservazione dell'energia

$$\frac{1}{2}mv^2 = mgh \Rightarrow g = \frac{v^2}{2h}$$

Il punto della questione è ora come calcolare la miglior stima e l'errore su di essa. Di conseguenza, intanto troviamo una miglior stima della velocità, di nuovo con media e semidispersione: $v = (4.41 \pm 0.06)\,\mathrm{m/s}$. Ora usiamo le formule di propagazione degli errori:

$$\begin{aligned}\Delta g &= \left|\frac{\partial g}{\partial v}\right| \Delta v + \left|\frac{\partial g}{\partial h}\right| \Delta h = \frac{v}{h}\Delta v + \frac{v^2}{2h^2}\Delta h = \\ &= \frac{v^2}{2h}\left(\frac{2\Delta v}{v} + \frac{\Delta h}{h}\right)\end{aligned}$$

Per cui otteniamo una stima di $g = (9.7 \pm 0.4)\,\mathrm{m/s^2}$.

Soluzione 7.3 (Soluzione al Problema 2.1)**.** La risposta è molto semplice e si tratta semplicemente di fare un integrale. Le patate nere, in base ai dati del problema, sono nell'intervallo $[\mu + 2\sigma, \mu + 3\sigma]$. Potendo approssimare la distribuzione con una gaussiana, per avere il numero di patate nere servite ci basta moltiplicare la probabilità che una patata sia nera per il numero totale di patate, ovvero

$$N = 200 \int_{\mu+2\sigma}^{\mu+3\sigma} \mathrm{gaus}(x, \mu, \sigma)\,\mathrm{d}x \approx 200(0.9987 - 0.9772) = 4.3$$

L'integrale non si può fare in modo analitico, ma esistono delle tabelle oppure si può fare numericamente con la calcolatrice. Ovviamente a questa stima bisogna anche dare un'incertezza in base ai dati del problema. Dato che tutti i dati numerici hanno una sola cifra significativa indicata, la stima sarà 4 ± 1.

Soluzione 7.4 (Soluzione al Problema 2.2)**.** L'unica cosa a cui fare attenzione è calcolare anche la covarianza nella propagazione delle incertezze. Usando la Equazione 13 e l'Equazione 14 dobbiamo calcolare separatamente la varianza campione di X, che chiameremo S_x, quella di Y, S_y e la covarianza campione di X, Y, chiamata $S_{x,y}$. La stima dell'errore sarà quindi

$$\sqrt{S_x^2 + S_y^2 + 2S_{x,y}} = 0.59$$

Notare che il risultato è abbastanza diverso da quello che si otterrebbe non considerando la correlazione, sommando solo in quadratura le deviazioni standard, ottenendo 0.43.

Per la seconda parte bisogna essere un po' più furbi. Dobbiamo trovare una variabile Z che abbia covarianza campione nulla con X, ovvero tale che valga

$$\frac{1}{n-1}\left(\overline{xz} - \overline{x}\,\overline{z}\right) = 0$$

Dato che dobbiamo trovare z in modo che sia funzione di x e y e questi due sono legati da una funzione lineare, la cosa più naturale da fare è cercare z nella forma $Ax + By$.

$$\begin{gathered}
\frac{1}{n-1}\left(\overline{x(Ax+By)} - \overline{x}\,\overline{Ax+By}\right) = 0 \\
\frac{1}{n-1}\left(A\overline{x^2} + B\overline{xy} - A(\overline{x})^2 - B\overline{x}\,\overline{y}\right) = 0 \\
A\operatorname{Var} X + B\operatorname{Cov}(X,Y) = 0
\end{gathered}$$

Quindi ci basta calcolare la covarianza di X e Y supponendo di sapere $y = mx + q$ per avere un'equazione che leghi A e B.

$$\operatorname{Cov}(X, mX+q) = \operatorname{Cov}(X, mX) + \operatorname{Cov}(X, q) = m\operatorname{Cov}(X,X) + 0 = m\operatorname{Var}(X)$$

Basta leggere le definizioni per convincersi della validità dei passaggi. Di conseguenza, dobbiamo avere

$$(A + mB)\operatorname{Var}(X) = 0$$

Dato che la varianza campione di X non è qualcosa che misuriamo, dobbiamo scegliere A e B in modo intelligente. La scelta naturale è di prendere $A = -mB$ e $B = 1$, ma ovviamente nessuno ci vieta di prendere dei multipli di questo. Notare che con questa scelta si ha $z = y - mx = q$, la cosa più naturale che si poteva pensare.

A questo punto possiamo costruire una nuova tabella con la variabile $z = y - 1.2x$. Notiamo che la covarianza campione di X e Z non è esattamente zero[10] ma è molto più vicina a zero di quanto non lo fosse quella fra X e Y, come desiderato (0.08 contro -0.001). Calcolando la varianza di $\bar{t}$ con queste altre variabili otteniamo un risultato compatibile con il precedente.

Soluzione 7.5 (Soluzione al Problema 5.1)**.** Rimanipolando l'espressione (dividendo per $k \cos\theta$ e spostando qualche termine) si ottiene

$$\frac{v}{\cos\theta} = -\frac{m_m g}{k}\tan\theta + \frac{\mu_m m_m g}{k},$$

che è un'espressione linearizzata ponendo $(X, Y) = (\tan\theta, \frac{v}{\cos\theta})$.

É sufficiene fare il logartimo dell'espresisone

$$\log B = \log A + \frac{x}{2}\left(R^2 + z^2\right).$$

Soluzione 7.6 (Soluzione al Problema 5.2)**.** Per risolvere il problema basta linearizzare l'espressione e successivamente fare un qualsiasi fit lineare. Facciamo il logaritmo dell'espressione:

$$\log\mu = \log\mu_0 + E/(RT),$$

dove adesso basta porre $(X, Y) = (\frac{1}{T}, \log\mu)$. Per maggiori dellatgli si rimanda alla soluzione ufficiale presente sul sito olifis.it.

Soluzione 7.7 (Soluzione al Problema 5.3)**.** Anche in questo caso si tratta solamente di trovare un modo per linearizzare l'espressione, per poi effettuare un fit lineare e trovare i parametri richiesti. Spostando il termine x^2

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

e sostituendo $X = x$, $Y = \frac{k}{4\pi^2 M_2}T^2 - x^2$, $M = -l$ e $Q = \frac{l^2}{3} + \frac{I_1}{M_2}$ si trova la relazione lineare cercata. Notare che le costanti che compaiono in Y sono tutte note a priori (e se così non fosse stato non sarebbe una buona linearizzazione).

[10] Questo semplicemente perché la relazione $y = 1.2x + q$ non è esatta.

A Proprietà utilizzate del valore atteso

- Il valore atteso è lineare, ovvero $\forall a, b \in \mathbb{R}, \forall f(x), g(x)$

$$\mathbb{E}_X[af(x) + bg(x)] = a\mathbb{E}_X[f(x)] + b\mathbb{E}_X[g(x)]$$

  La dimostrazione è semplicissima e si basa sulla linearità dell'integrale.

- Il valore atteso di un prodotto di variabili indipendenti fattorizza, ovvero

$$\mathbb{E}_{(X,Y)}[f(x)g(y)] = \mathbb{E}_X[f(x)]\mathbb{E}_Y[g(y)]$$

  Questa è un'affermazione banale se si capisce cosa vuol dire e si guardano i passaggi della dimostrazione. Innanzitutto qui abbiamo due variabili aleatorie, X, Y, che in generale sono descritte da una distribuzione di probabilità *congiunta* $p(x, y)$ che in generale **non** si può scrivere come $p(x, y) = a(x)b(y)$. Tuttavia, l'ipotesi di variabili indipendenti vuole dire esattamente che la distribuzione di probabilità si può scrivere come prodotto, ovvero nel nostro caso, proprio grazie all'ipotesi di indipendenza, possiamo scrivere $p(x, y) = a(x)b(y)$. Di conseguenza,

$$\begin{aligned}
\mathbb{E}_{(X,Y)}[f(x)g(y)] &= \int_{\mathcal{XY}} p(x,y)f(x)g(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{\mathcal{XY}} a(x)b(y)f(x)g(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{\mathcal{X}} a(x)f(x)\,\mathrm{d}x \int_{\mathcal{Y}} b(y)g(y)\,\mathrm{d}y \\
&= \mathbb{E}_X[f(x)]\mathbb{E}_Y[g(y)]
\end{aligned}$$

  Il primo passaggio è vero per definizione di valore atteso, il secondo è vero per l'ipotesi di indipendenza delle due variabili, il terzo è semplicemente un raggruppamento di termini che si può fare per le proprietà degli integrali $n-$dimensionali, mentre l'ultimo semplicemente riconosce la definizione di valore atteso.

B Calcolo degli stimatori per la gaussiana

Dobbiamo massimizzare la funzione

$$\mathcal{L}(\mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=0}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2\right]$$

rispetto ai parametri μ, σ. I valori ottimali verranno indicati con $\hat{\mu}, \hat{\sigma}$. Il primo trucco per semplificarsi la vita è notare che il logaritmo è una funzione monotona crescente, per cui se $f(x)$ è una funzione strettamente positiva ovunque, massimizzare $f(x)$ o $\log f(x)$ è assolutamente equivalente.

$$\log \mathcal{L}(\mu, \sigma) = -\frac{n}{2} \log(2\pi) - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=0}^{n} (x_i - \mu)^2$$

Questo è un po' più semplice da sistemare. Per trovare i punti di massimo si impone che tutte le derivate parziali siano nulle nel punto.

$$\begin{cases} \dfrac{\partial}{\partial \mu} \log \mathcal{L}(\mu, \sigma) \Big|_{\mu=\hat{\mu}, \sigma=\hat{\sigma}} = 0 \\ \dfrac{\partial}{\partial \sigma} \log \mathcal{L}(\mu, \sigma) \Big|_{\mu=\hat{\mu}, \sigma=\hat{\sigma}} = 0 \end{cases}$$

Questo è a tutti gli effetti un sistema di due equazioni a due incognite, che ora scriveremo. La soluzione sarà effettivamente la coppia di stimatori che massimizza il tutto.

$$\begin{cases} -\frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (x_i - \hat{\mu}) = 0 \\ -\dfrac{n}{\hat{\sigma}} + \dfrac{1}{\hat{\sigma}^3} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 = 0 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \dfrac{1}{n} \sum_{i=1}^{n} x_i \\ \hat{\sigma} = \dfrac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \end{cases}$$

Valore atteso di $\hat{\sigma}^2$

Eravamo arrivati all'espressione

$$\mathbb{E}[\hat{\sigma}^2] = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}[x_i^2] + \mathbb{E}[\hat{\mu}^2] - 2\mathbb{E}[x_i \hat{\mu}] \right) \tag{19}$$

Non facciamoci prendere la mano inventando proprietà inesistenti del valore atteso e andiamo a scrivere la definizione di $\hat{\mu}$ per poi computare il tutto, pezzo per pezzo. Il più semplice è quello con $\mathbb{E}[x_i^2] = \sigma^2 + \mu^2$. Consideriamo quello subito dopo

$$\mathbb{E}[\hat{\mu}^2] = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right] = E\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j\right] =$$
$$= \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{E}[x_i x_j] = \frac{1}{n^2}\left(\sum_{i=j=1}^{n}\mathbb{E}[x_i x_j] + \sum_{i\neq j}\mathbb{E}[x_i x_j]\right)$$
$$= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}[x_i^2] + \sum_{i\neq j}\mathbb{E}[x_i]\mathbb{E}[x_j]\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n}\sigma^2 + \sum_{i\neq j}\mu^2\right)$$
$$= \frac{1}{n^2}(n\sigma^2 + 2n(n-1)\mu^2) = \frac{\sigma^2}{n} + 2\frac{n-1}{n}\mu^2$$

Qui abbiamo spezzato la somma in due pezzi in quanto se $i \neq j$, x_i e x_j sono scorrelati, in quanto immaginiamo di fare misure scorrelate, mentre se $i = j$, allora $x_i = x_j$, per cui dobbiamo tenerne conto. Vediamo ora cosa succede all'altro pezzo.

$$\mathbb{E}[x_i\hat{\mu}] = E\left[\frac{1}{n}\sum_{j=1}^{n} x_i x_j\right] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[x_i x_j] = \frac{1}{n}\left(\sum_{j\neq i}\mathbb{E}[x_i x_j] + \mathbb{E}[x_i^2]\right)$$
$$= \frac{1}{n}\left(\sum_{j\neq i}\mathbb{E}[x_i]\mathbb{E}[x_j] + \sigma^2\right) = \frac{n-1}{n}\mu^2 + \frac{1}{n}\sigma^2$$

Ora che abbiamo tutti i pezzi, riprendiamo l'Equazione 19 riscrivendo ogni pezzo al suo posto

$$\mathbb{E}[\hat{\sigma}^2] = \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2 + \frac{\sigma^2}{n} + \cancel{2\frac{n-1}{n}\mu^2} - \cancel{2\frac{n-1}{n}\mu^2} \to 0 - \frac{2}{n}\sigma^2\right) =$$
$$= \frac{1}{n}\sum_{i=1}^{n}\frac{n-1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2$$

C Giustificazione del minimo χ^2

Supponiamo di essere in una tipica situazione di fit, ovvero di misurare coppie di punti (x_i, y_i) e di avere un modello che ci dice che le grandezze x, y sono legate da una certa funzione che dipende da dei parametri, ovvero ci

aspettiamo $y = f(x, \vec{\theta})$, dove $\vec{\theta}$ sono tutti i parametri. L'obiettivo è, a partire dalle coppie di dati sperimentali, ottenere una stima sensata dei parametri $\vec{\theta}$. Per fare questa cosa, useremo di nuovo il principio di massima verosimiglianza. Se infatti ogni misura y_i ha associato un errore σ_i, ci aspettiamo che sia

$$p(y_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\left(\frac{x_i - f(x_i, \vec{\theta})}{\sigma_i}\right)^2\right]$$

Ovviamente questa assunzione è assolutamente arbitraria e non è detto che sia sempre quella migliore da fare, in quanto non è assolutamente detto che nel caso generale la distribuzione degli errori sia gaussiana. Tuttavia, per vari teoremi come il teorema centrale del limite, se non ci sono delle sistematiche grosse è lecito pensare che la gaussiana sia, fra tutte, se proprio dobbiamo scegliere una funzione senza fare un modello specifico per il nostro caso, la migliore se dobbiamo tirarne fuori una dal cappello.

A partire da questa assunzione, la verosimiglianza è quindi

$$\mathcal{L}_{(\vec{x},\vec{y})}(\vec{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\left(\frac{x_i - f(x_i, \vec{\theta})}{\sigma_i}\right)^2\right]$$

Dovremo trovare il massimo di questa funzione rispetto ai parametri $\vec{\theta}$, tenendo fisso i vari x_i, y_i, che sono le nostre misure. Come abbiamo fatto prima, dato che $\mathcal{L} > 0$, massimizzare $\mathcal{L}$ oppure $\log \mathcal{L}$ è assolutamente equivalente, per cui

$$\begin{aligned}\log \mathcal{L}_{(\vec{x},\vec{y})}(\vec{\theta}) &= -\frac{1}{2}\sum_{i=1}^{n} \log(2\pi\sigma_i^2) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - f(x_i, \vec{\theta})}{\sigma_i}\right)^2 \\ &= -\frac{1}{2}\sum_{i=1}^{n} \log(2\pi\sigma_i^2) - \frac{1}{2}\chi^2\end{aligned}$$

Tuttavia il primo pezzo è costante, perché dipende dagli errori sulle y, mentre il secondo pezzo lo riconosciamo nel χ^2. Vediamo quindi dal segno $-$ davanti al χ^2 che massimizzare la verosimiglianza è, nel caso gaussiano, equivalente a minimizzare il χ^2. Nel caso in cui tutti i σ_i siano uguali ad un certo valore, questo diventa un semplice metodo dei minimi quadrati.

A partire quindi da un principio generale e sensato siamo riusciti a ricavare due metodi che abbiamo visto in precedenza.

Riferimenti bibliografici

[Bal17] Luca Baldini. Introduzione all'analisi dei dati. 2017. Reperibile qui.

[Pet18] Giacomo Petrillo. Dispense del corso di Analisi e Statistica dei dati. 2018. Reperibile qui.