Al tendere del numero N di misure effettuate all'infinito, risulta

$$\lim_{N\to\infty} \frac{n_i}{N} = \Pr\left(x_i - \frac{\Delta x_i}{2} \le x < x_i + \frac{\Delta x_i}{2}\right)$$
$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{x_i-\frac{\Delta x_i}{2}}^{x_i+\frac{\Delta x_i}{2}} e^{-\frac{1}{2}\left(\frac{x-x^*}{\sigma}\right)^2} \mathrm{d}x$$

e dunque

$$y_i \xrightarrow[N\to\infty]{} \frac{NA}{\Delta x_i} \frac{1}{\sigma\sqrt{2\pi}} \int_{x_i-\frac{\Delta x_i}{2}}^{x_i+\frac{\Delta x_i}{2}} e^{-\frac{1}{2}\left(\frac{x-x^*}{\sigma}\right)^2} \mathrm{d}x \ .$$

Cioè l'altezza dell'istogramma, in ognuna delle classi di frequenza, tende al valore medio sull'intervallo corrispondente della funzione di Gauss moltiplicato per un fattore costante NA. Allora la curva da sovrapporre all'istogramma sperimentale deve essere quella che corrisponde alla funzione

$$f(x) = \frac{NA}{s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{s}\right)^2}$$

(in luogo del valore vero x^* e dell'errore quadratico medio σ, generalmente ignoti, si pongono le loro stime, $\bar{x}$ e s rispettivamente, ottenute dal campione stesso); osserviamo che $f(x)$ sottende la stessa area NA dell'istogramma.

Se gli intervalli hanno tutti la medesima ampiezza Δx, l'area del rettangolo elementare vale $A = \Delta x$, assumendo l'arbitraria unità di misura per le ordinate pari all'altezza costante del rettangolo elementare, e la funzione diviene

$$f(x) = \frac{N\,\Delta x}{s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{s}\right)^2} \ .$$

Sempre per quel che riguarda le implicazioni "pratiche" della legge normale di distribuzione degli errori, un altro punto sul quale gli studenti hanno frequentemente dei dubbi riguarda l'applicazione della funzione di Gauss a grandezze misurate sì commettendo errori casuali, ma che siano per loro natura *limitate*. Ad esempio, una lunghezza è una grandezza fisica implicitamente non negativa: quindi la densità di probabilità associata ai particolari valori ottenibili x dovrebbe essere identicamente nulla quando $x < 0$, mentre la funzione normale si annulla soltanto quando $x = \pm\infty$. Affermare che i risultati della misura seguono la legge di Gauss sembra dunque una contraddizione.

La risposta a questa obiezione è che la funzione di distribuzione della x effettivamente *non può essere normale*: ma che la reale differenza tra la vera funzione di distribuzione e quella di Gauss è assolutamente trascurabile.

Facciamo un esempio pratico: supponiamo di voler misurare la dimensione del lato di un quaderno (di valore vero 20 cm) con un regolo graduato, e di commettere un errore di misura $\sigma = 1$ mm; la probabilità di trovare un risultato in un intervallo ampio 1 mm appena alla sinistra dello zero secondo la legge normale vale

$$p \simeq \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}200^2} \approx 0.4\, e^{-2\times 10^4} \;;$$

quindi $\ln(p) \sim -2 \times 10^4$, e $\log_{10}(p) = \ln(p) \cdot \log_{10}(e) \sim -10^4$, mentre dovrebbe essere rigorosamente $p \equiv 0$.

Per valutare le reali implicazioni di un valore di p come quello che stiamo considerando, attualmente il numero di atomi presenti nell'universo si stima essere dell'ordine di 10^{79}; mentre l'età dell'universo stesso si stima in circa 10^{10} anni, ovvero dell'ordine di 10^{18} secondi; se pensiamo ad un gruppo di misuratori in numero pari al numero di atomi nell'universo, ed ognuno dei quali esegua una misura al secondo, dovrebbe passare un tempo pari circa a 7 volte l'età dell'universo stesso per ottenere un valore illegale qualora le misure seguissero veramente la legge di Gauss: quindi la differenza tra la funzione di distribuzione reale e quella ipotizzata è effettivamente trascurabile.

9.6 Esame dei dati

Talvolta nella misura si compiono errori non classificabili né come casuali né come sistematici: ad esempio, dopo aver misurato un angolo di un triangolo con un goniometro, si può riportare come valore un numero diverso scambiando tra di loro due cifre contigue. La conseguenza sarà quella di ottenere per il risultato finale della misura (la somma degli angoli interni di un triangolo) un dato molto differente dagli altri, che si impone quindi alla nostra attenzione come *sospetto*.

Nasce quindi il desiderio di avere un criterio preciso in base al quale decidere se un dato possa o meno considerarsi *sospetto*, ed essere in conseguenza eliminato.

Normalmente la procedura consigliata è la seguente: dopo aver calcolato media e scarto quadratico medio, si eliminano dal campione i dati che differiscano da $\bar{x}$ per più di tre volte s. Sappiamo infatti che valori che si trovino nella regione oltre 3σ hanno probabilità molto bassa di presentarsi (del tre per mille circa); bisogna comunque osservare che questo modo di procedere è giustificato *solo* in presenza di un *numero piccolo* di dati.

Se le misure effettuate sono in numero ad esempio di 60, ci si attende che (per fluttuazioni dovute esclusivamente al caso) solo 0.18 misure (praticamente: nessuna) differiscano dal valore medio, in modulo, per più di 3σ; se troviamo una (o più) misure di questo tipo, possiamo attribuire la loro presenza, piuttosto che ad una fluttuazione casuale, a cause d'errore del tipo di quelle considerate, quindi etichettarle come *sospette* ed infine scartarle.

Le cose cambiano se ci troviamo di fronte invece ad un milione di misure, per le quali ci aspettiamo che ben 3000 cadano (per motivi perfettamente normali) al di fuori dell'intervallo di 3σ, e non possiamo quindi permetterci di scartare alcun dato particolare.

9.7 Sommario delle misure dirette

Per concludere, dovendo effettuare delle misure dirette:

- *Bisogna considerare criticamente le modalità della misura e le formule usate, e controllare le caratteristiche di costruzione e d'uso degli strumenti per mettere in evidenza la possibilità di errori sistematici; se questi sono presenti bisogna eliminarli: o cambiando gli strumenti, o modificando le modalità delle operazioni da compiere, o correggendo opportunamente i risultati.*
- *Potendo, bisogna effettuare misure ripetute: perché in questo caso sappiamo stimare ragionevolmente l'errore commesso a partire dalle misure stesse (se non è possibile effettuare misure ripetute, si assumerà convenzionalmente come errore l'inverso della sensibilità dello strumento, ovverosia la più piccola variazione della grandezza indicata sulla scala di lettura); e bisogna effettuarne quante più possibile per aumentare in corrispondenza la validità statistica dei nostri risultati.*
- *Se il numero di misure effettuate è basso*[4] *si scartano quei dati che differiscano dal valore medio per più di 3 volte lo scarto quadratico medio s. Effettuata questa operazione si ricalcolano la media $\bar{x}$ e lo scarto quadratico medio s, e si ricava da quest'ultimo la stima dell'errore della media $\sigma_{\bar{x}}$ costituita da $s_{\bar{x}} = s/\sqrt{N}$.*

[4] "Basso" si può ad esempio considerare un numero di misure tale che il numero atteso di eventi da scartare in base alla distribuzione normale sia inferiore all'unità.

- *Come valore più verosimile per la grandezza misurata si assume $\bar{x}$, e come errore di questo valore $s_{\bar{x}}$; se le misure sono in numero sufficiente e non si sono commessi errori sistematici, il significato dell'errore è quello di semiampiezza dell'intervallo di confidenza centrato sulla media e avente probabilità di includere il valore vero pari al 68%.*

9.8 Il teorema del limite centrale

Fino ad ora abbiamo più volte sottolineato il fatto che un preciso significato (quello statistico) dell'errore quadratico medio può essere enunciato solo se la distribuzione delle misure effettuate è quella normale.

Con riguardo alla media aritmetica delle misure, se queste seguono la legge normale e se, inoltre, sono statisticamente indipendenti tra loro, il teorema di pagina 103 ci assicura che qualunque loro combinazione lineare (ed in particolare la media aritmetica) è ancora distribuita secondo la legge normale; ed all'errore della media $\sigma_{\bar{x}}$ si può quindi attribuire lo stesso significato statistico.

Vogliamo ora ampliare questo discorso dimostrando un importantissimo teorema della statistica e discutendone le implicazioni:

> TEOREMA (DEL LIMITE CENTRALE): *siano N variabili casuali x_i, statisticamente indipendenti tra loro e provenienti da una distribuzione avente densità di probabilità ignota, della quale esistano finite sia la media μ che la varianza σ^2; sotto questa ipotesi, la distribuzione della media aritmetica del campione, $\bar{x}$, tende asintoticamente alla distribuzione normale con media μ e varianza σ^2/N al crescere di N.*

Dimostreremo questo teorema facendo l'ipotesi, più restrittiva, che esistano i momenti della funzione di frequenza delle x_i di qualunque ordine k (esso può essere dimostrato, come si vede dall'enunciato, anche se esistono solamente i primi due); e partiamo dal fatto che, sotto le ipotesi su dette, la somma S delle N variabili casuali

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

ha valore medio e varianza date dalle

$$E(S) = N\mu \qquad \text{e} \qquad {\sigma_S}^2 = N\sigma^2 \ .$$

Inoltre, visto che i valori x_i sono tra loro statisticamente indipendenti, possiamo applicare l'equazione (6.11) per trovare la funzione caratteristica della S, che vale

$$\begin{aligned}\phi_S(t) &= \prod_{i=1}^{N} \phi_{x_i}(t) \\ &= \left[\phi_x(t)\right]^N\end{aligned}$$

visto che le x_i hanno tutte la stessa distribuzione (e quindi la stessa funzione caratteristica). Se consideriamo invece gli scarti $z_i = x_i - \mu$ delle x_i dalla media, dalla (6.17) possiamo ricavare la funzione caratteristica della z:

$$\phi_z(t) = e^{-i\mu t}\phi_x(t) \tag{9.7}$$

e, se esistono tutti i momenti fino a qualsiasi ordine della x (e in conseguenza anche della z), la (6.8) implica

$$\phi_z(t) \;=\; \sum_{k=0}^{\infty} \frac{(it)^k}{k!}\lambda_k \;=\; 1 + 0 - \frac{1}{2}t^2\sigma^2 + \mathcal{O}(t^3) \tag{9.8}$$

in cui i λ_k sono i momenti della funzione di frequenza della z, i primi due dei quali valgono 0 e σ^2.

Introduciamo infine la nuova variabile

$$y \;=\; \frac{S - E(S)}{\sigma_S} \;=\; \frac{1}{\sigma\sqrt{N}}S - \frac{N\mu}{\sigma\sqrt{N}} \;=\; \frac{1}{\sigma\sqrt{N}}\sum_{i=1}^{N}(x_i - \mu)$$

e indichiamo poi con $\phi_y(t)$ la funzione caratteristica della y; essendo quest'ultima lineare in S abbiamo dalla (6.17) che

$$\begin{aligned}\phi_y(t) &= e^{-it\frac{N\mu}{\sigma\sqrt{N}}} \cdot \phi_S\left(\frac{t}{\sigma\sqrt{N}}\right) \\ &= e^{-it\frac{N\mu}{\sigma\sqrt{N}}}\left[\phi_x\left(\frac{t}{\sigma\sqrt{N}}\right)\right]^N \\ &= \left[e^{-i\mu\frac{t}{\sigma\sqrt{N}}} \cdot \phi_x\left(\frac{t}{\sigma\sqrt{N}}\right)\right]^N \\ &= \left[\phi_z\left(\frac{t}{\sigma\sqrt{N}}\right)\right]^N\end{aligned}$$

ricordando la (9.7). Da qui, introducendo l'espressione (9.8) prima ottenuta per lo sviluppo di $\phi_z(t)$,

$$\begin{aligned}\phi_y(t) &= \left[1 - \frac{1}{2}\frac{t^2}{N\sigma^2}\sigma^2 + \mathcal{O}\left(\frac{t^3}{N^{\frac{3}{2}}}\right)\right]^N \\ &= \left[1 - \frac{t^2}{2N} + \mathcal{O}\left(N^{-\frac{3}{2}}\right)\right]^N\end{aligned}$$

e quando N tende all'infinito

$$\lim_{N\to\infty} \phi_y(t) = e^{-\frac{t^2}{2}}$$

sfruttando il limite notevole

$$\lim_{x\to+\infty}\left(1 + \frac{k}{x}\right)^x = e^k \tag{9.9}$$

(qui, appunto, $k = -t^2/2$). Insomma la funzione caratteristica della y tende a quella di una distribuzione normale di media zero e varianza 1: quindi la S tende asintoticamente ad una distribuzione normale di media $N\mu$ e varianza $N\sigma^2$; e $\bar{x} = S/N$ tende asintoticamente ad una distribuzione normale di media μ e varianza σ^2/N.

Il teorema è di fondamentale importanza perché non fa alcuna ipotesi sulla distribuzione delle variabili che compongono il campione (all'infuori del requisito dell'esistenza di media e varianza). Con riguardo alle misure ripetute di una stessa grandezza fisica esso ci dice che, se anche la loro distribuzione non segue la legge di Gauss, *purché se ne abbia un numero sufficiente* il nostro risultato finale (la media aritmetica) tuttavia la segue ugualmente in modo approssimato: così che l'errore della media conserva il consueto significato statistico (di semiampiezza dell'intervallo, centrato su $\bar{x}$, che contiene il valore vero con probabilità costante prefissata del 68%) anche se questo non è verificato per le singole misure.

Da notare che il teorema del limite centrale implica una convergenza asintoticamente normale del valore medio del campione al valore medio della popolazione delle misure; per attribuire a quest'ultimo, come si è fatto nell'ultima frase, il significato di valore vero della grandezza misurata, si sottintende che le misure abbiano distribuzione, ancorché di forma non specificata, simmetrica rispetto al valore vero x^*; insomma che errori per difetto e per eccesso siano ugualmente probabili.

Incidentalmente, notiamo qui come il *prodotto* di molte variabili casuali indipendenti debba avere un comportamento, indipendentemente dal tipo di distribuzione, asintoticamente tendente a quello di una distribuzione *log-normale*.

9.8.1 Applicazione: numeri casuali normali

Siano gli u_i (con $i = 1, \ldots, N$) dei numeri provenienti da una popolazione u distribuita uniformemente nell'intervallo $[0, 1]$; abbiamo visto nel paragrafo 8.1 che $E(u) = \frac{1}{2}$ e $\mathrm{Var}(u) = \frac{1}{12}$. La loro media aritmetica $\bar{u}$, in conseguenza del teorema del limite centrale, tende asintoticamente (al crescere di N) alla distribuzione normale con media $\frac{1}{2}$ e varianza $\frac{1}{12 \cdot N}$; quindi la loro somma $N\bar{u}$ è asintoticamente normale con media $\frac{N}{2}$ e varianza $\frac{N}{12}$; e, infine, la variabile casuale

$$x = \frac{\sum_{i=1}^{N} u_i - \frac{N}{2}}{\sqrt{\frac{N}{12}}} \tag{9.10}$$

è asintoticamente normale con media 0 e varianza 1.

Di questa proprietà si può far uso per ottenere da un computer dei numeri pseudo-casuali con distribuzione (approssimativamente) normale, a partire da altri numeri pseudo-casuali con distribuzione uniforme; in pratica l'approssimazione è già buona quando $N \gtrsim 10$, e scegliendo $N = 12$ possiamo, ad esempio, porre semplicemente

$$x = \sum_{i=1}^{12} u_i - 6 \ .$$

È da notare, comunque, che **non è** buona pratica servirsi di questo metodo: anche se la parte centrale della distribuzione normale è approssimata abbastanza bene, le code mancano totalmente (essendo impossibile che risulti $|x| > \sqrt{3N}$); l'effetto di questa mancanza, quando (come nelle analisi fisiche basate su metodi di Montecarlo) vengano richiesti numeri pseudo casuali per generare eventi simulati in quantità dell'ordine di milioni almeno, è tale da invalidare completamente i risultati.

Soprattutto, poi, generare numeri pseudo-casuali normali usando il teorema del limite centrale non è solo sbagliato, ma inutile: esistono altri metodi (come ad esempio quello di Box–Muller che discuteremo ora) che sono in grado di generare numeri pseudo-casuali con una *vera* distribuzione normale usando, per il calcolo, un tempo non molto superiore a quello richiesto dalla (9.10).

Siano x ed y due variabili casuali *statisticamente indipendenti*, ed aventi distribuzione uniforme nell'intervallo $[0, 1]$; consideriamo le altre due variabili casuali u e v definite attraverso le

$$u = \sqrt{-2 \ln x} \cdot \cos(2\pi y) \qquad \text{e} \qquad v = \sqrt{-2 \ln x} \cdot \sin(2\pi y) \ . \tag{9.11}$$

Queste funzioni si possono invertire, e risulta

$$x = e^{-\frac{1}{2}(u^2+v^2)} \qquad \text{e} \qquad y = \frac{1}{2\pi}\arctan\left(\frac{v}{u}\right)$$

con derivate parziali prime date da

$$\begin{cases} \dfrac{\partial x}{\partial u} = -u \cdot e^{-\frac{1}{2}(u^2+v^2)} \\ \dfrac{\partial x}{\partial v} = -v \cdot e^{-\frac{1}{2}(u^2+v^2)} \end{cases}$$

e da

$$\begin{cases} \dfrac{\partial y}{\partial u} = \dfrac{1}{2\pi}\,\dfrac{1}{1+\frac{v^2}{u^2}}\left(-\dfrac{v}{u^2}\right) \\ \dfrac{\partial y}{\partial v} = \dfrac{1}{2\pi}\,\dfrac{1}{1+\frac{v^2}{u^2}}\,\dfrac{1}{u} \end{cases}$$

Il determinante Jacobiano delle (x, y) rispetto alle (u, v) vale

$$\begin{aligned} \frac{\partial(x,y)}{\partial(u,v)} &= \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u} \\ &= -\frac{1}{2\pi}\,e^{-\frac{1}{2}(u^2+v^2)} \end{aligned}$$

per cui, essendo la densità di probabilità congiunta delle due variabili casuali x ed y data dalla

$$f(x, y) = 1$$

e applicando la (7.9), la densità di probabilità congiunta della u e della v è data da

$$f(u, v) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2} \cdot \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}v^2}$$

e quindi, in conseguenza della (7.8), la u e la v sono due variabili casuali *statisticamente indipendenti* tra loro ed entrambe aventi funzione di frequenza data dalla *distribuzione normale standardizzata*; questo è appunto il metodo cosiddetto "di Box–Muller" per la generazione di numeri pseudo-casuali con distribuzione normale, a partire da numeri pseudo-casuali con distribuzione uniforme.

Una variante che consente di sveltire questo metodo (lento, perché l'esecuzione delle funzioni logaritmo, seno e coseno consuma molto tempo

di CPU) consiste nel generare dapprima due numeri pseudo-casuali x' e y' distribuiti uniformemente tra i limiti -1 e $+1$; e nell'accettarli se $S = R^2 = x'^2 + y'^2 \leq 1$, in modo che il punto P le cui coordinate essi rappresentano nel piano $\{x', y'\}$ sia uniformemente distribuito entro il cerchio avente centro nell'origine O e raggio unitario — o nel rigettarli in caso contrario, ripetendo il passo precedente.

Questa prima condizione in realtà *rallenta* il procedimento, perché la coppia di numeri a caso viene accettata con probabilità $\frac{\pi}{4} \approx 78.5\%$; ma se, a questo punto, si usa al posto della x nella (9.11) il valore di S (che, come non è difficile dimostrare, è anch'esso distribuito uniformemente nell'intervallo $[0, 1]$); e se si prende poi in luogo dell'angolo $2\pi y$ l'angolo polare θ tra $\overline{OP}$ e l'asse delle x', il calcolo risulta in definitiva molto più rapido: perché il seno ed il coseno di θ si possono valutare come y'/R ed x'/R rispettivamente, eseguendo il calcolo di una radice quadrata e due divisioni soltanto.

Capitolo 10

Le misure indirette

Misure indirette, come sappiamo, sono quelle eseguite non sulla grandezza fisica che interessa determinare ma su altre grandezze che siano a quest'ultima legate da una qualche relazione funzionale; quest'ultima ci permetterà poi di ricavarne il valore mediante il calcolo.

Supponiamo per semplicità che queste altre grandezze vengano misurate direttamente: gli inevitabili errori commessi per ognuna di esse si ripercuoteranno poi attraverso i calcoli effettuati, e si *propagheranno* fino al risultato finale; l'entità dell'errore nelle misure indirette dipenderà dunque sia dai valori di quelli commessi nelle misure dirette, sia dalla forma analitica della funzione usata per il calcolo.

Consideriamo il caso del tutto generale di una qualsiasi funzione F di più variabili $x, y, z, \ldots$: ammettendo che i valori di queste variabili si possano ottenere da misure di tipo diretto, vogliamo determinare un algoritmo per ricavare da tali misure una stima del valore vero di F; infine, nell'ipotesi che le variabili siano anche tra loro statisticamente indipendenti, vedremo come si può valutare l'errore collegato a tale stima.

10.1 Risultato della misura

Innanzi tutto, è chiaro che il valore vero F^* della grandezza F è quello che corrisponde ai valori veri delle variabili indipendenti da cui F dipende:

$$F^* = F(x^*, y^*, z^*, \ldots) \quad . \tag{10.1}$$

Non avendo però a disposizione tali valori veri, tutto quello che possiamo fare è usare le migliori stime di cui disponiamo: cioè, supponiamo, i valori medi di campioni di determinazioni ripetute di tutte queste grandezze; insomma, calcolare il valore $\overline{F}$ assunto dalla funzione in corrispondenza dei valori $\bar{x}, \bar{y}, \bar{z}, \ldots$ delle variabili.

Ricordiamo che il valore di una funzione di più variabili F in un qualsiasi punto si può ricavare dal valore assunto dalla F e dalle sue derivate successive in un punto diverso, attraverso la formula dello sviluppo in serie di Taylor:

$$\begin{aligned}
F(x, y, z, \ldots) = F\left(x_0, y_0, z_0, \ldots\right) + & \qquad \text{(ordine zero)} \\
+ \frac{\partial F}{\partial x}\left(x - x_0\right) + \frac{\partial F}{\partial y}\left(y - y_0\right) + \cdots & \qquad \text{(primo ordine)} \\
+ \frac{\partial^2 F}{\partial x^2} \frac{\left(x - x_0\right)^2}{2!} + \cdots & \qquad \text{(secondo ordine)} \\
+ \mathcal{O}(3) &
\end{aligned}$$

(in cui per brevità si è omesso di indicare che le derivate parziali vanno calcolate per i valori delle variabili $x = x_0$, $y = y_0$ e così via).

Se è possibile trascurare i termini di ordine superiore al primo, possiamo in particolare ricavare:

$$\begin{aligned}
\overline{F} &= F\left(\bar{x}, \bar{y}, \bar{z}, \ldots\right) \\
&\approx F\left(x^*, y^*, z^*, \ldots\right) + \frac{\partial F}{\partial x}\left(\bar{x} - x^*\right) + \frac{\partial F}{\partial y}\left(\bar{y} - y^*\right) + \cdots \quad .
\end{aligned}$$

Prendendo poi il valore medio di entrambi i membri e tenendo presente nei passaggi sia la (10.1), sia che risulta $E(\bar{x} - x^*) = E(\bar{x}) - x^* \equiv 0$ in assenza di errori sistematici (e similmente per le altre variabili), si ottiene

$$\begin{aligned}
E\left\{F\left(\bar{x}, \bar{y}, \bar{z}, \ldots\right)\right\} &\approx F^* + \frac{\partial F}{\partial x} E\left(\bar{x} - x^*\right) + \frac{\partial F}{\partial y} E\left(\bar{y} - y^*\right) + \cdots \\
&= F^*
\end{aligned}$$

cioè

$$E\left(\overline{F}\right) \approx F^*$$

e, in definitiva:

> *In media, il valore di una funzione F calcolato per le medie misurate delle variabili coincide col valore vero.*

(ossia $\overline{F}$ è una stima *imparziale* di F^*).

Ricordiamo che questa conclusione è valida *solo approssimativamente*, perché nello sviluppo in serie di Taylor abbiamo trascurato tutti i termini di ordine superiore al primo; ma quali sono i limiti della validità della conclusione? In quali casi si possono cioè effettivamente considerare trascurabili i termini del second'ordine e degli ordini superiori?

Ognuno dei termini di ordine i nello sviluppo di F contiene una delle derivate i-esime della funzione, moltiplicata per un fattore del tipo $(\bar{x} - x^*)$ elevato alla i-esima potenza e divisa per il fattoriale di i; sarà senz'altro lecito trascurare questi termini se le differenze tra i valori medi stimati per le variabili indipendenti ed i loro valori veri sono piccole, in altre parole *se gli errori commessi nelle misure dirette sono piccoli*.

Un caso particolare è poi quello in cui la F è una funzione lineare in ognuna delle variabili da cui dipende; in questo caso, ovviamente, tutte le derivate di ordine successivo al primo sono identicamente nulle, e le conclusioni precedenti sono valide *esattamente*.

10.2 Combinazioni lineari di misure dirette

Supponiamo che le misure dirette delle variabili indipendenti da cui dipende la F siano esenti da errori sistematici, e che siano pertanto distribuite secondo la legge normale; consideriamo dapprima quelle particolari funzioni che sono le combinazioni lineari di più variabili:

$$F \;=\; k_1\, x_1 + k_2\, x_2 + k_3\, x_3 + \cdots \;=\; \sum\nolimits_i k_i\, x_i \quad .$$

Abbiamo già visto, nell'equazione (5.2) a pagina 51, che il valore medio di una tale funzione è la combinazione lineare, con gli stessi coefficienti, delle medie delle variabili; e, se supponiamo inoltre che le variabili siano tutte *statisticamente indipendenti* tra loro, sappiamo anche che la varianza di F è poi data (equazione (5.5) a pagina 54) dalla combinazione lineare delle loro varianze con coefficienti pari ai quadrati dei rispettivi coefficienti:

$$E(F) \;=\; \sum\nolimits_i k_i\, E(x_i) \;=\; \sum\nolimits_i k_i\, x_i^* \;\equiv\; F^*$$

e

$$\sigma_F{}^2 = \sum\nolimits_i k_i{}^2 \sigma_i{}^2 \quad .$$

Abbiamo inoltre dimostrato, a pagina 103, un teorema secondo il quale una qualsiasi combinazione lineare a coefficienti costanti di variabili casuali

aventi distribuzione normale, ed inoltre tra loro statisticamente indipendenti, è anch'essa distribuita secondo la legge normale.

Ora, sapendo che la distribuzione della F è data dalla funzione di Gauss, siamo anche in grado di attribuire un significato più preciso al suo errore quadratico medio σ_F: quello cioè di semiampiezza dell'intervallo, avente centro sul valore medio $E(F) = F^*$, che contiene un qualsiasi valore della F (ed in particolare la nostra miglior stima $\overline{F}$) con una probabilità del 68%; o, di converso, la semiampiezza di un intervallo centrato sulla nostra migliore stima $\overline{F}$ e che contiene l'ignoto valore vero F^* con una probabilità del 68%.

In definitiva le formule precedenti risolvono il problema delle misure indirette per quelle particolari funzioni che sono le combinazioni lineari, permettendoci di calcolare per esse sia il valore stimato più verosimile che l'errore, e dandoci inoltre l'interpretazione probabilistica di questo errore.

10.3 La formula di propagazione degli errori

Ora, qualsiasi funzione di più variabili si può considerare in prima approssimazione lineare; questo se ci limitiamo a considerarla in un dominio di definizione abbastanza ristretto da poter trascurare i termini di ordine superiore al primo in uno sviluppo in serie di Taylor. In definitiva possiamo estendere le conclusioni del paragrafo precedente ad una qualsiasi funzione di più variabili

$$F(x, y, z, \ldots) \approx F(\bar{x}, \bar{y}, \bar{z}, \ldots) + \frac{\partial F}{\partial x}(x - \bar{x}) + \frac{\partial F}{\partial y}(y - \bar{y}) + \frac{\partial F}{\partial z}(z - \bar{z}) + \cdots$$

per la cui varianza avremo

$$\boxed{{\sigma_F}^2 \approx \left(\frac{\partial F}{\partial x}\right)^2 {\sigma_x}^2 + \left(\frac{\partial F}{\partial y}\right)^2 {\sigma_y}^2 + \left(\frac{\partial F}{\partial z}\right)^2 {\sigma_z}^2 + \cdots} \tag{10.2}$$

(le derivate vanno calcolate per i valori $x = \bar{x}$, $y = \bar{y}$, $z = \bar{z}$,... delle variabili indipendenti).

Questa formula è nota sotto il nome di *formula di propagazione degli errori*: ripetiamo che si tratta di una formula *approssimata*; che è valida solo se non si commettono errori troppo grandi nelle misure dirette delle variabili; e che presuppone che le variabili stesse siano tra loro statisticamente

indipendenti[1]. La formula di propagazione è invece *esatta* nel caso particolare (esaminato nel paragrafo precedente) di una combinazione lineare di variabili casuali indipendenti, caso questo nel quale tutte le derivate parziali di ordine superiore al primo sono identicamente nulle.

10.4 Errore dei prodotti di potenze

Applichiamo ora la formula (10.2) di propagazione degli errori a quella particolare classe di funzioni costituita dai prodotti di potenze delle variabili indipendenti: cioè alle funzioni del tipo

$$F(x, y, z, \ldots) = K \cdot x^{\alpha} \cdot y^{\beta} \cdot z^{\gamma} \cdots \quad .$$

Calcoliamo innanzi tutto le derivate parziali di F; risulta

$$\begin{cases} \dfrac{\partial F}{\partial x} &= K \cdot \alpha\, x^{\alpha-1} \cdot y^{\beta} \cdot z^{\gamma} \cdots &= \alpha \dfrac{F}{x} \\[2ex] \dfrac{\partial F}{\partial y} &= K \cdot x^{\alpha} \cdot \beta\, y^{\beta-1} \cdot z^{\gamma} \cdots &= \beta \dfrac{F}{y} \\[2ex] \cdots \end{cases}$$

(ammettendo che nessuna delle variabili sia nulla; questo implica che anche la F abbia valore diverso da zero). Introducendo questi valori delle derivate nella formula di propagazione degli errori, avremo

$$\begin{aligned} {\sigma_F}^2 &\approx \left(\frac{\partial F}{\partial x}\right)^2 {\sigma_x}^2 + \left(\frac{\partial F}{\partial y}\right)^2 {\sigma_y}^2 + \cdots \\ &= \alpha^2 \frac{F^2}{x^2} {\sigma_x}^2 + \beta^2 \frac{F^2}{y^2} {\sigma_y}^2 + \cdots \end{aligned}$$

ed in definitiva

$$\left(\frac{\sigma_F}{F}\right)^2 \approx \alpha^2 \left(\frac{\sigma_x}{x}\right)^2 + \beta^2 \left(\frac{\sigma_y}{y}\right)^2 + \cdots \quad ;$$

relazione che permette di ricavare con semplici calcoli l'errore relativo di F dagli errori relativi commessi nella misura delle variabili indipendenti. Per quanto detto in precedenza, questa relazione è solo una prima approssimazione; e possiamo ritenerla valida se le variabili indipendenti sono misurate con errori piccoli.

[1] Una formula di propagazione degli errori per variabili qualsiasi (che ossia non ne presupponga l'indipendenza statistica) verrà ricavata più avanti, nel paragrafo C.3.

10.5 Errori massimi

Quando si parla di errori di misura senza specificare null'altro, si sottintende di norma che i numeri riportati si riferiscono ad errori quadratici medi; talvolta però si è in grado di indicare un intervallo all'interno del quale si sa con assoluta certezza che deve trovarsi il valore vero della grandezza misurata: in questi casi si può ovviamente attribuire un errore in termini *assoluti* (sia in difetto che in eccesso) al valore indicato. Supponendo per semplicità che i valori limite siamo simmetrici rispetto al risultato trovato, che si potrà quindi ancora esprimere nella forma

$$x = x_0 \pm \Delta x \quad ,$$

vogliamo ora determinare la legge secondo la quale si propagano questi *errori massimi* nelle misure indirette.

È immediato riconoscere che, se $F = Kx$ e $x \in [x_0 - \Delta x, x_0 + \Delta x]$, necessariamente F deve appartenere ad un intervallo di semiampiezza ΔF, con $\Delta F = |K|\,\Delta x$. Similmente, sommando (o sottraendo) due grandezze indipendenti di cui si conoscano gli errori massimi, il risultato $F = x \pm y$ dovrà essere compreso in un intervallo di semiampiezza $\Delta F = \Delta x + \Delta y$. Usando entrambe queste conclusioni, nel caso di una combinazione lineare

$$F = \sum_{i=1}^{N} a_i\, x_i$$

l'errore massimo su F vale

$$\Delta F = \sum_{i=1}^{N} |a_i|\, \Delta x_i \quad .$$

Per una relazione funzionale qualsiasi

$$F = F(x_1, x_2, \ldots, x_N) \quad ,$$

e nei limiti in cui si possano trascurare i termini di ordine superiore al primo in uno sviluppo in serie di Taylor, la formula di propagazione per gli errori massimi è dunque

$$\Delta F \approx \sum_{i=1}^{N} \left| \frac{\partial F}{\partial x_i} \right| \Delta x_i \quad ;$$

e, per i prodotti di potenze del tipo $F = K \cdot x^{\alpha} \cdot y^{\beta} \cdots$,

$$\frac{\Delta F}{|F|} \approx |\alpha| \frac{\Delta x}{|x|} + |\beta| \frac{\Delta y}{|y|} + \cdots \quad .$$

Capitolo 11

Stime di parametri

In questo capitolo prenderemo in considerazione due speciali tecniche di elaborazione dei dati che sono utilizzate per stimare il valore di parametri ignoti dai quali le distribuzioni teoriche dipendono: la media pesata di determinazioni sperimentali aventi diversa precisione; e la valutazione dei parametri da cui dipende l'equazione di una curva che deve descrivere una relazione tra più variabili interconnesse e misurate indipendentemente (curva interpolante i dati sperimentali).

Il metodo usato per la soluzione è, in entrambi i casi, quello della *massima verosimiglianza* (introdotto originariamente da Fisher[1] nel 1921); la prima parte del capitolo riguarderà appunto il problema della stima del valore dei parametri in generale, e questo metodo in particolare.

11.1 Stime e loro caratteristiche

Supponiamo che la densità di probabilità $f(x;\theta)$ di una variabile casuale continua x (che possa assumere tutti i valori dell'asse reale) dipenda da un parametro θ, il cui valore vero θ^* ci sia ignoto; se si hanno a disposizione N determinazioni sperimentali indipendenti x_i della grandezza x, vogliamo

[1] Sir Ronald Fisher nacque a Londra nel 1890 e morì ad Adelaide (in Australia) nel 1962. È considerato, per l'importanza dei suoi lavori, uno dei fondatori della moderna statistica: oltre al concetto di verosimiglianza (*likelihood* in inglese), introdusse per primo l'analisi delle varianze e scoperse la forma analitica delle funzioni di distribuzione di molte importanti variabili casuali; dette poi importanti contributi ai metodi per i piccoli campioni ed a quelli per la verifica delle ipotesi.

trovare una funzione $\bar{\theta} = \bar{\theta}(x_1, x_2, \ldots, x_N)$ che, a partire da esse, ci permetta di ricavare, nella maniera migliore possibile, un valore numerico da attribuire a θ^*: le funzioni $\bar{\theta}$ di questo tipo si chiamano appunto *stime*.

Una stima è dunque una funzione di variabili casuali, e, pertanto, una variabile casuale essa stessa; potremo in conseguenza parlare del valore medio o della varianza di una particolare stima, intendendo così riferirci alle caratteristiche della popolazione dei possibili valori restituiti dalla stima stessa in corrispondenza di tutti i possibili campioni che possono essere usati per calcolarla.

Nella statistica, alle stime si possono associare svariate caratteristiche; la prima di esse (e la più importante) è la **consistenza**. Una stima si dice *consistente* quando converge (probabilisticamente) al valore vero del parametro, ossia quando

$$\lim_{N\to\infty} \bar{\theta}(x_1, x_2, \ldots, x_N) \;=\; \theta^* \quad .$$

Ad esempio, il teorema di Čebyšef si può enunciare sinteticamente affermando che "il valore medio di un campione è una stima consistente del valore medio della popolazione".

Una seconda caratteristica delle stime è la **distorsione**: una stima si dice *indistorta*, o *imparziale*, se mediamente coincide col valore vero del parametro; insomma se

$$E(\bar{\theta}) = \theta^* \quad .$$

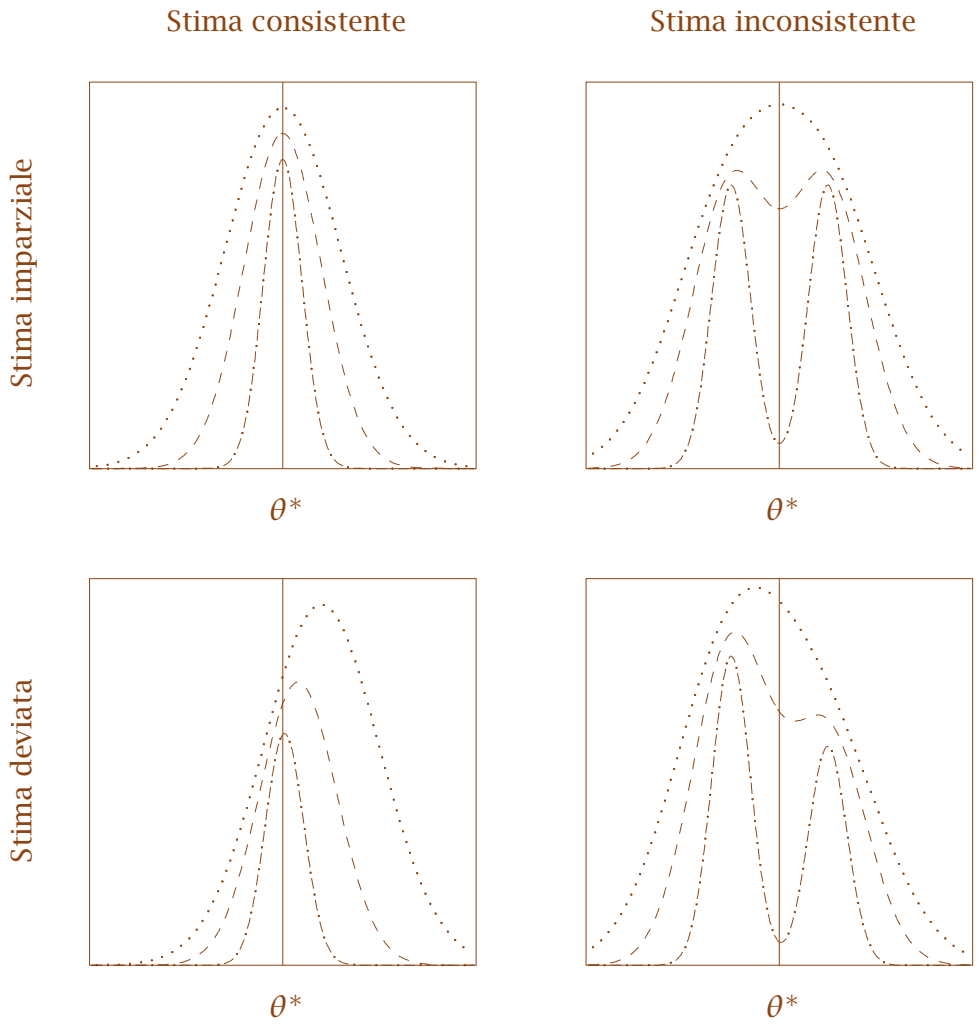
Già sappiamo, dai paragrafi 5.3 e 5.8 rispettivamente, che la media dei campioni è una stima indistorta del valore medio della popolazione; mentre la varianza del campione è una stima distorta, ancorché consistente, della varianza della popolazione (a meno che non sia opportunamente corretta moltiplicandola per un fattore $N/(N-1)$).

Nella figura 11a sono riportati esempi di stime consistenti ed inconsistenti, distorte ed indistorte, per dare un'idea dell'andamento della densità di probabilità delle stime stesse all'aumentare delle dimensioni del campione.

Una terza caratteristica delle stime è l'**efficienza**: diremo che una prima stima è più *efficiente* di una seconda se la sua varianza è inferiore, e, quindi, se mediamente essa è più vicina al valore centrale $E(\bar{\theta})$; che coincide con θ^* se la stima è anche imparziale. Esiste un teorema (teorema di Cramér–Rao) del quale ci occuperemo sia più avanti nel corso di questo capitolo, sia in particolare nell'appendice E; questo teorema dimostra l'esistenza di un limite inferiore per la varianza delle stime, e quindi di un limite superiore per la loro efficienza.

Se abbiamo a disposizione, poi, M stime differenti θ_j dello stesso parametro θ, ogni campione di N valori x_i produrrà, attraverso l'applicazione

Figura 11a - Stime consistenti ed inconsistenti, imparziali e deviate.



di ognuna di tali stime, M diversi valori per θ. Se ora indichiamo con f la densità di probabilità congiunta di questi M valori, risulterà in generale

$$f(\theta_1, \theta_2, \ldots, \theta_M; \theta^*) = f^M(\theta_1; \theta^*) \cdot \varphi(\theta_2, \theta_3, \ldots, \theta_M; \theta^* | \theta_1)$$

dove con $f^M(\theta_1; \theta^*)$ abbiamo, al solito, indicato la funzione densità di probabilità marginale della sola θ_1 (ovvero la densità di probabilità collegata al presentarsi di un certo valore per θ_1 indipendentemente da quello ottenuto per le altre stime); mentre $\varphi(\theta_2, \theta_3 \ldots, \theta_M; \theta^* | \theta_1)$ è la densità di probabilità di queste ulteriori $M - 1$ stime condizionata dal valore della prima.

Nel caso che φ risulti *indipendente* da θ^*, la conseguenza che da questo fatto si deduce è che, una volta calcolata θ_1, le altre stime sarebbero distribuite comunque nello stesso modo per *qualunque* valore di θ^*; esse non potrebbero quindi aggiungere nulla alla conoscenza già ottenuta sul valore del parametro θ: ovverosia θ_1 *sfrutta tutta l'informazione* sul parametro ignoto che è contenuta nei dati, ed in questo caso la stima θ_1 si dice **sufficiente**. Non è detto che una stima sufficiente per un certo parametro θ esista; ma se ne esiste una, $\bar{\theta}$, allora ne esistono infinite: si può dimostrare infatti che ogni funzione monotona in senso stretto di $\bar{\theta}$ gode della stessa proprietà.

11.2 La stima di massima verosimiglianza

Dato un campione di N determinazioni *indipendenti* x_i, l'espressione

$$\prod_{i=1}^{N} f(x_i; \theta^*)$$

rappresenta la densità di probabilità da associare all'evento casuale consistente nell'ottenere una determinata N-pla di valori, essendo θ^* il valore del parametro da cui la f dipende.

Se in questa espressione si sostituisce al valore vero (che avevamo supposto noto) θ^* il generico valore θ; e se le x_i non vengono considerate più variabili casuali, ma costanti che sono state determinate dalle nostre operazioni di misura, la funzione

$$\mathcal{L}(x_1, x_2, \ldots, x_N; \theta) = \prod_{i=1}^{N} f(x_i; \theta) \tag{11.1}$$

(*funzione di verosimiglianza*) rappresenta la densità di probabilità da associare all'evento casuale consistente nell'essere un certo θ il valore vero del

nostro parametro, nell'ipotesi di avere già ottenuto la particolare N-pla di valori sperimentali $x_1, x_2, \ldots, x_N$.

Il metodo della massima verosimiglianza consiste nell'adottare, come stima del parametro θ, quel valore $\hat{\theta}$ *che rende massima* la funzione di verosimiglianza (11.1); ovvero la soluzione delle

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\theta} = 0 \qquad\qquad \frac{\mathrm{d}^2\mathcal{L}}{\mathrm{d}\theta^2} < 0 \tag{11.2}$$

(nel caso che le (11.2) abbiano più di una soluzione, si sceglie quella che corrisponde al massimo assoluto).

Visto che il logaritmo naturale è (essendo la base, e, maggiore di uno) una funzione monotona strettamente crescente dell'argomento, trovare il massimo di $\ln \mathcal{L}$ condurrebbe ancora a tutti e soli i valori che rendono massima $\mathcal{L}$; questo corrisponde al sostituire (essendo $\mathcal{L} > 0$), alla prima delle (11.2), l'equivalente

$$\frac{1}{\mathcal{L}}\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\theta} = \frac{\mathrm{d}\,(\ln \mathcal{L})}{\mathrm{d}\theta} = 0 \ .$$

Enunciamo qui, senza dimostrarle, alcune proprietà fondamentali della stima di massima verosimiglianza:

1. La stima di massima verosimiglianza è una stima *asintoticamente consistente* al crescere della dimensione del campione.

2. La stima di massima verosimiglianza ha una densità di probabilità *asintoticamente normale* al crescere della dimensione del campione.

3. La stima di massima verosimiglianza è asintoticamente, al crescere della dimensione del campione, anche *la stima più efficiente possibile* (ossia quella di minima varianza).

4. Se esiste una stima sufficiente di θ, essa può sempre essere espressa come funzione della sola stima di massima verosimiglianza $\hat{\theta}$.

Le ipotesi sotto le quali si riesce a dimostrare che la stima di massima verosimiglianza gode asintoticamente delle proprietà su dette sono estremamente generali: per la normalità basta che esistano i primi due momenti della $f(x;\theta)$; per la consistenza e la massima efficienza basta che $f(x;\theta)$ sia continua, dotata di derivata prima e seconda rispetto al parametro, e che l'operazione di integrazione rispetto a x commuti con quella di derivazione rispetto a θ (ovvero, in pratica, che il dominio di definizione della x non dipenda dal parametro).

Il teorema di Cramér–Rao (cui si è prima accennato) permette di dimostrare, sotto ipotesi del tutto generali, che esiste *un estremo inferiore* per le

varianze delle stime *imparziali* di una qualsiasi grandezza dipendente dal parametro θ; non solo, ma che, se una stima di varianza minima esiste, essa *rende massima la funzione di verosimiglianza*.

Più in dettaglio: nell'ipotesi che la densità di probabilità $f(x;\theta)$ sia una funzione definita in una regione dell'asse x avente estremi indipendenti dal parametro θ; che esista ovunque la derivata rispetto a θ di $\ln f(x;\theta)$; e, infine, che esista finito il valore medio del quadrato di questa derivata

$$E\left\{\left[\frac{\partial}{\partial\theta}\ln f(x;\theta)\right]^2\right\} = \frac{1}{N}\cdot E\left\{\left[\frac{\partial(\ln\mathcal{L})}{\partial\theta}\right]^2\right\}$$

il teorema di Cramér–Rao afferma che una *qualsiasi* stima *imparziale* $\bar{\theta}$ di θ ha una varianza che non può essere inferiore ad un valore (*limite di Cramér–Rao*) dato dalla

$$\mathrm{Var}(\bar{\theta}) \geq \frac{1}{N\cdot E\left\{\left[\frac{\partial}{\partial\theta}\ln f(x;\theta)\right]^2\right\}} \,. \tag{11.3}$$

Inoltre questo estremo inferiore viene raggiunto, e vale il segno di uguaglianza nella (11.3), *se e solo se* esiste una funzione $R(\theta)$ per la quale risulti

$$\frac{\partial(\ln\mathcal{L})}{\partial\theta} = \sum_{i=1}^{N}\frac{\partial}{\partial\theta}\ln f(x_i;\theta) = \frac{\bar{\theta}(x_1,x_2,\ldots,x_N)-\theta}{R(\theta)} \tag{11.4}$$

e, in tal caso, la stima di minima varianza *rende anche massima la funzione di verosimiglianza*.

La condizione (11.4) è assai restrittiva, potendosi tra l'altro dimostrare che essa implica che la densità di probabilità $f(x;\theta)$ deve essere una funzione di tipo esponenziale: nel caso generale non è quindi affatto certo che una stima di varianza minima esista, essendo questo subordinato alla validità della (11.4).

In ogni caso la stima di massima verosimiglianza deve, come prima detto, tendere *asintoticamente* a questo comportamento al crescere di N; però nulla si può dire sulla rapidità di tale convergenza. Così, per un numero di misure finito, non c'è alcuna garanzia che la funzione di verosimiglianza abbia un solo massimo; e, se essa ne ammette più d'uno, non esiste modo di sapere quale di essi corrisponde (asintoticamente) alla stima di minima varianza, né esiste modo di sapere quale di questi massimi rappresenti la stima corretta del valore vero.

Come abbiamo detto, la funzione di verosimiglianza (11.1) può essere interpretata come densità di probabilità del parametro una volta che si

sia ottenuto un certo insieme di valori misurati; sfruttando la seconda delle proprietà su elencate, la densità di probabilità di θ deve anche essere (asintoticamente) data da

$$\mathcal{L}(\theta) = \frac{1}{\sigma_\theta \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\theta-\hat{\theta}}{\sigma_\theta}\right)^2}$$

quindi, nell'intorno di $\hat{\theta}$, deve essere

$$\ln \mathcal{L} = -\ln\left(\sigma_\theta \sqrt{2\pi}\right) - \frac{1}{2}\left(\frac{\theta-\hat{\theta}}{\sigma_\theta}\right)^2$$

e, derivando due volte rispetto al parametro,

$$\frac{\mathrm{d}^2(\ln \mathcal{L})}{\mathrm{d}\theta^2} = -\frac{1}{{\sigma_\theta}^2}$$

ed infine si giunge alla

$$\mathrm{Var}(\hat{\theta}) \equiv {\sigma_\theta}^2 = -\frac{1}{\left.\frac{\mathrm{d}^2(\ln \mathcal{L})}{\mathrm{d}\theta^2}\right|_{\theta=\hat{\theta}}} \tag{11.5}$$

frequentemente usata per il calcolo dell'errore della stima di massima verosimiglianza.

11.2.1 Un esempio di stima sufficiente

Supponiamo di avere un campione di N determinazioni indipendenti x_k di una variabile che segua la distribuzione di Poisson; le probabilità dei differenti valori sono date dalla (8.14), e dipendono da un unico parametro: il valore medio della distribuzione, α. La funzione di verosimiglianza è la

$$\begin{aligned}
\Pr(x_1, \ldots, x_N; \alpha) &= \frac{\alpha^{x_1}}{x_1!} e^{-\alpha} \cdot \frac{\alpha^{x_2}}{x_2!} e^{-\alpha} \cdots \frac{\alpha^{x_N}}{x_N!} e^{-\alpha} \\
&= \frac{\alpha^{\sum_k x_k} \cdot e^{-N\alpha}}{x_1! x_2! \cdots x_N!} \cdot \frac{(N\bar{x})!}{(N\bar{x})!} \\
&= \frac{\alpha^{N\bar{x}}}{(N\bar{x})!} e^{-N\alpha} \cdot \frac{(N\bar{x})!}{x_1! x_2! \cdots x_N!} \cdot \frac{N^{N\bar{x}}}{N^{N\bar{x}}} \\
&= \left\{\frac{(N\alpha)^{N\bar{x}}}{(N\bar{x})!} e^{-N\alpha}\right\} \left\{\frac{(N\bar{x})!}{x_1! x_2! \cdots x_N!} \frac{1}{N^{N\bar{x}}}\right\}
\end{aligned} \tag{11.6}$$

Nei passaggi, per due volte si è moltiplicato e diviso per una stessa quantità non nulla: prima per $(N\bar{x})!$ e poi per $N^{N\bar{x}}$.

La stima di massima verosimiglianza si trova annullando la derivata della (11.6); che, a meno di un fattore costante, è della forma

$$f(\alpha) = \alpha^{N\bar{x}} e^{-N\alpha}$$

per cui

$$\frac{\mathrm{d}f}{\mathrm{d}\alpha} = N\,\bar{x}\,\alpha^{N\bar{x}-1}e^{-N\alpha} - N\,\alpha^{N\bar{x}}e^{-N\alpha} = N\,\alpha^{N\bar{x}-1}e^{-N\alpha}(\bar{x}-\alpha)$$

e quindi la stima cercata è

$$\hat{\alpha} = \bar{x}$$

Il primo termine dell'espressione finale (11.6) per la funzione di verosimiglianza è la probabilità $\Pr(S)$ che la variabile casuale

$$S = \sum_{k=1}^{N} x_k = N\bar{x}$$

abbia un determinato valore: $\Pr(S)$ infatti, come già sappiamo dal paragrafo 8.5, segue la distribuzione di Poisson con valore medio $N\alpha$. Notiamo anche che, avendo N un valore costante noto a priori, $\Pr(S)$ coincide con $\Pr(\bar{x})$: il secondo termine *deve* quindi essere la probabilità che i dati osservati valgano $x_1, x_2, \ldots, x_N$ *condizionata* dal fatto che la loro somma vale $N\bar{x}$; ma, non dipendendo questo termine da α, tale probabilità è la stessa qualunque sia il parametro.

Qualunque sia $\bar{x}$, una volta noto il suo valore le x_k sono distribuite allo stesso modo: $\bar{x}$ riassume insomma tutta l'informazione contenuta nei dati, ed è quindi per definizione una stima *sufficiente* del parametro. In effetti, se la probabilità dei valori x_k una volta nota $\bar{x}$ non dipende dal parametro, questo implica che *qualunque* funzione dei dati ha probabilità (condizionata) che gode della stessa proprietà. Citiamo senza dimostrarlo, in proposito, il seguente

> TEOREMA: *$\bar{\theta}$ è una stima sufficiente di θ se e solo se la funzione di verosimiglianza è fattorizzabile nella forma*
>
> $$\mathcal{L}(x_1, x_2, \ldots, x_N; \theta) = f(\bar{\theta}, \theta) \cdot \phi(x_1, x_2, \ldots, x_N)$$

11.3 Media pesata

Quando si abbiano a disposizione più determinazioni ripetute di una stessa grandezza fisica, sappiamo che da esse si può ricavare un valore unico da usare come risultato finale attraverso il calcolo della media aritmetica; questa (come già anticipato senza dimostrazione nel paragrafo 4.4) è la funzione dei dati con la distribuzione più stretta attorno al valore vero, e ci fornisce quindi la stima più verosimile di esso. Però questo presuppone che i dati, essendo considerati tutti allo stesso modo nella formula, posseggano la stessa precisione sperimentale: ad esempio che siano stati valutati dallo stesso sperimentatore, con lo stesso strumento e nelle stesse condizioni; in altre parole, che le misure provengano da un'unica popolazione.

Può capitare invece di disporre di più determinazioni della stessa grandezza fisica fatte da sperimentatori diversi, od in condizioni sperimentali differenti: e di voler ugualmente estrarre da queste valutazioni, affette da differenti errori, un valore unico da usare come risultato complessivo.

Facendo le ipotesi che tutte le misure x_i siano tra loro statisticamente indipendenti, ed inoltre affette da errori casuali distribuiti secondo la legge di Gauss, la densità di probabilità corrispondente all'evento casuale costituito dall'osservazione degli N valori $x_1, x_2, \ldots, x_N$ si può scrivere (applicando il teorema della probabilità composta)

$$\prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x^* - x_i}{\sigma_i}\right)^2}$$

dove x^* è il valore vero (ignoto) di x, e le σ_i sono gli errori quadratici medi (supposti noti) delle diverse determinazioni.

La funzione di verosimiglianza è la

$$\mathcal{L}(x_1, x_2, \ldots, x_N; x) \;=\; \prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x - x_i}{\sigma_i}\right)^2}$$

(cioè la densità di probabilità di cui sopra, nella quale il valore vero x^* è sostituito dal parametro variabile x); e ricordiamo che essa rappresenta la densità di probabilità associata all'evento casuale consistente nell'essere il numero x il valore vero della grandezza misurata, qualora di essa si siano ottenute le N stime indipendenti x_i, di errori rispettivi σ_i, supposte seguire la legge normale.

La stima più verosimile è quella che, rendendo massima $\mathcal{L}$, individua quel numero che, sulla base delle osservazioni disponibili, possiede la *massima probabilità* di coincidere con il valore vero: vedremo tra poco che la

soluzione è unica. Prendendo il logaritmo naturale di $\mathcal{L}$,

$$-2\,\ln\mathcal{L} \;=\; \sum_{i=1}^{N}\left(\frac{x-x_i}{\sigma_i}\right)^2 \;+\; 2\sum_{i=1}^{N}\ln\sigma_i \;+\; 2N\ln\sqrt{2\pi}$$

e ricordando, come prima detto, che il logaritmo naturale è una funzione monotona strettamente crescente dell'argomento, si vede che il massimo di $\mathcal{L}$ corrisponde al minimo di $-2\,\ln\mathcal{L}$; la determinazione del valore più verosimile di x (nel caso di errori normali) si riduce allora al problema analitico di trovare il minimo della funzione

$$f(x) \;=\; \sum_{i=1}^{N}\left(\frac{x-x_i}{\sigma_i}\right)^2$$

(infatti nessuno degli altri termini dipende dall'incognita x). Risolviamo il problema facendo uso del calcolo infinitesimale:

$$\frac{\mathrm{d}f}{\mathrm{d}x} \;=\; 2\sum_{i=1}^{N}\left(\frac{x-x_i}{\sigma_i}\right)\frac{1}{\sigma_i} \;=\; 2\left(x\sum_{i=1}^{N}\frac{1}{{\sigma_i}^2} \;-\; \sum_{i=1}^{N}\frac{x_i}{{\sigma_i}^2}\right) \;;$$

$$\frac{\mathrm{d}^2f}{\mathrm{d}x^2} \;=\; 2\sum_{i=1}^{N}\frac{1}{{\sigma_i}^2} \;>\; 0 \;.$$

Se per brevità poniamo

$$K \;=\; \sum_{i=1}^{N}\frac{1}{{\sigma_i}^2}$$

la condizione per l'estremante di $f(x)$ si scrive

$$\frac{\mathrm{d}f}{\mathrm{d}x} \;=\; 2\left(Kx - \sum_{i=1}^{N}\frac{x_i}{{\sigma_i}^2}\right) \;=\; 0$$

e la derivata prima di f si annulla quando la variabile x assume il valore

$$\bar{x} \;=\; \frac{1}{K}\sum_{i=1}^{N}\frac{x_i}{{\sigma_i}^2} \;. \tag{11.7}$$

Il fatto che la derivata seconda sia positiva assicura poi che si tratta effettivamente di un punto di minimo; si vede come $\bar{x}$ sia una *media pesata* dei valori misurati x_i, ottenuta assegnando ad ognuno di essi peso relativo inversamente proporzionale al *quadrato* dell'errore rispettivo.

Per determinare poi l'errore del risultato $\bar{x}$, è in questo caso possibile usare in tutta generalità la formula della propagazione degli errori: infatti $\bar{x}$ è una particolare funzione delle variabili x_i, di ognuna delle quali conosciamo per ipotesi l'errore quadratico medio σ_i; ed inoltre dipende *linearmente* da ognuna di queste N variabili, e questo fa sì che la formula di propagazione (10.2) sia in questo caso *esatta* e non approssimata (dando insomma risultati sempre validi, indipendentemente dall'entità degli errori commessi).

Applichiamo direttamente l'equazione (5.5) per la varianza delle combinazioni lineari di variabili tra loro indipendenti, invece della più complicata (10.2): $\bar{x}$ è calcolata come combinazione lineare delle x_i con coefficienti $1/\left(K\,\sigma_i{}^2\right)$, e quindi avremo

$$\sigma_{\bar{x}}{}^2 \;=\; \sum_{i=1}^{N}\left(\frac{1}{K\,\sigma_i{}^2}\right)^2 \sigma_i{}^2 \;=\; \frac{1}{K^2}\sum_{i=1}^{N}\frac{1}{\sigma_i{}^2} \;=\; \frac{1}{K}$$

cioè

$$\sigma_{\bar{x}}{}^2 = \frac{1}{\sum\limits_{i=1}^{N}\dfrac{1}{\sigma_i{}^2}} \quad . \tag{11.8}$$

Per la osservata linearità della formula, la media pesata $\bar{x}$ (nelle ipotesi ammesse) è una variabile casuale normale come le singole x_i; ed il suo errore quadratico medio $\sigma_{\bar{x}}$ ha dunque l'analoga interpretazione di semiampiezza dell'intervallo con centro in $\bar{x}$ avente probabilità pari al 68% di contenere il valore vero x^*.

Per quanto concerne le proprietà della media pesata $\bar{x}$ come stima del valore vero, la derivata del logaritmo della funzione di verosimiglianza rispetto al parametro incognito (che è x) vale

$$\frac{\mathrm{d}(\ln \mathcal{L})}{\mathrm{d}x} \;=\; \sum_{i=1}^{N}\frac{x_i}{\sigma_i{}^2} - x\sum_{i=1}^{N}\frac{1}{\sigma_i{}^2} \;=\; K(\bar{x}-x)$$

ed è soddisfatta la condizione (11.4) sotto la quale il teorema di Cramér–Rao (che esamineremo in dettaglio nell'appendice E) ci permette di affermare che la stima di massima verosimiglianza *è anche quella di varianza minima*: ovvero, tra tutte le possibili funzioni dei dati che si potrebbero definire per stimare il valore vero x^* dal campione, quella mediamente più vicina ad esso.

È da notare come, prima di comporre tra loro determinazioni indipendenti della stessa grandezza, sia opportuno controllare che queste siano (entro

i rispettivi errori) tra loro *compatibili*; analogamente a quanto si fa per le misure ripetute, è preferibile non considerare dati che non vadano d'accordo con gli altri entro i limiti della pura casualità.

Il caso di N misure ripetute effettuate nelle medesime condizioni sperimentali non è altro che il caso particolare in cui tutti gli errori quadratici medi σ_i sono uguali tra di loro: la media pesata (11.7) si riduce allora alla media aritmetica (4.1) (ed il suo errore (11.8) alla già nota espressione (5.6)).

Questo prova l'asserto del paragrafo 4.4 (giustificazione della media); abbiamo finalmente *dimostrato* che la media aritmetica è il valore *più verosimile* della grandezza misurata: cioè quello che ha la massima probabilità di coincidere con il valore vero sulla base del nostro campione di misure, e che rappresenta la stima di minima varianza.

11.4 Interpolazione dei dati con una curva

Può in alcuni casi capitare di conoscere la forma analitica della legge fisica che mette in relazione tra loro due variabili, e di dover stimare dai dati misurati il valore di uno o più parametri da cui tale funzione dipende.

Ad esempio, nel moto dei corpi soggetti all'azione di una forza costante le velocità assunte in istanti successivi dal corpo crescono linearmente rispetto ai tempi trascorsi, secondo la nota formula $v = v_0 + at$; misurando in istanti successivi del moto tempi e velocità, i punti aventi per coordinate cartesiane i valori determinati per queste due grandezze devono disporsi approssimativamente lungo una linea retta: e sarebbero tutti quanti esattamente allineati se fosse possibile misurare senza commettere errori.

In questo ultimo caso sarebbe possibile ricavare immediatamente dal grafico il valore dell'accelerazione del moto, che corrisponderebbe al coefficiente angolare (o pendenza) della retta tracciata; vedremo ora come, pur commettendo errori, sia comunque possibile ricavare una stima sia dei valori dei parametri da cui l'equazione del moto dipende, sia degli errori inerenti a tale valutazione.

C'è una qualche analogia fra questo problema e quello delle misure indirette, nel senso che in entrambi i casi si presuppone esistente una relazione funzionale tra più grandezze fisiche; tuttavia, mentre in quel caso la funzione era completamente nota e veniva usata per trovare il valore di una di quelle grandezze una volta misurati quelli di tutte le altre, qui si suppone di conoscere soltanto la forma della funzione: ma sono ignoti uno o più parametri da cui pure essa dipende, e si usano i valori osservati di tutte le grandezze per stimare quelli dei parametri stessi.

11.4.1 Interpolazione lineare per due variabili

Cominciamo col supporre che le variabili oggetto della misura siano due sole, e che la legge che le mette in relazione reciproca sia di tipo lineare:

$$y = a + bx \quad .$$

Supponiamo poi che siano state effettuate misure del valore della x e di quello corrispondente assunto dalla y in diverse condizioni, così che si disponga in definitiva di N coppie di valori tra loro corrispondenti (x_i, y_i); abbiamo già detto che, una volta riportati sul piano cartesiano $\{x, y\}$ punti con queste coordinate, essi si dovranno disporre approssimativamente lungo una linea retta.

Ora, si può dimostrare che vale, sul piano, qualcosa di analogo a quanto abbiamo già asserito riguardo alla media aritmetica di misure ripetute di una stessa grandezza fisica (cioè, geometricamente, su di una retta, visto che quelle determinazioni potevano essere univocamente rappresentate da punti su di una retta orientata); infatti

- Sulla base delle misure effettuate, non si può escludere con certezza che alcuna delle infinite rette del piano corrisponda a quella vera su cui le nostre osservazioni si disporrebbero in assenza di errori; tuttavia esse non appaiono tutte quante ugualmente verosimili, e la verosimiglianza sarà in qualche modo in relazione con la *distanza complessiva* tra i nostri punti sperimentali e la retta stessa.

- Nel caso particolare che siano verificate le seguenti ipotesi:

  1. una sola delle variabili coinvolte (ad esempio la y) è affetta da errori;
  2. gli errori quadratici medi delle misure dei diversi valori di y sono tutti uguali (o comunque non molto differenti);
  3. questi errori seguono la legge normale di distribuzione;
  4. le N determinazioni effettuate sono tra loro statisticamente indipendenti;

  dimostreremo ora che per "distanza complessiva" si deve intendere *la somma dei quadrati delle lunghezze dei segmenti di retta parallela all'asse y compresi tra i punti misurati e la retta esaminata.*

Infatti, detto σ_y l'errore quadratico medio delle y_i, la funzione di verosimiglianza è

$$\mathcal{L}(x_1, y_1, x_2, y_2, \ldots, x_N, y_N; a, b) \;=\; \prod_{i=1}^{N} \frac{1}{\sigma_y \sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{a+bx_i-y_i}{\sigma_y}\right)^2} \quad .$$

Per scrivere questa espressione si è fatto uso di tutte le ipotesi postulate: in particolare, il fatto che le x_i siano misurate senza errore ci permette di affermare che il valore vero assunto in corrispondenza dalla y è $a + bx_i$; visto che è $y = a + bx$ la legge fisica che lega le due variabili tra loro.

Questa funzione di verosimiglianza rappresenta allora la densità di probabilità collegata all'evento casuale consistente nell'essere la legge fisica che lega x ad y rappresentata dall'equazione $y = a + bx$, qualora si siano ottenuti gli N valori misurati (x_i, y_i), e sotto le quattro ipotesi su elencate.

I valori più verosimili del parametro saranno quelli che rendono massima $\mathcal{L}$: vedremo ora che la soluzione è unica; e, ancora, il teorema di Cramér–Rao ci permetterebbe di dimostrare che la stima, appunto, più verosimile (la retta che corrisponde al massimo della probabilità) è anche la stima di minima varianza (ovvero la più precisa possibile). Prendendo il logaritmo naturale di entrambi i membri, risulta

$$-2 \ln \mathcal{L} \;=\; \frac{1}{{\sigma_y}^2} \sum_{i=1}^{N} (a + bx_i - y_i)^2 \;+\; 2N \ln \sigma_y \;+\; 2N \ln \sqrt{2\pi} \;\; .$$

I valori più verosimili dei parametri a e b sono quelli per cui è massima $\mathcal{L}$, ovvero è minima $-2 \ln \mathcal{L}$: il problema dell'interpolazione lineare dunque si riduce (se sono soddisfatte le ipotesi citate) a quello di trovare tra le infinite rette del piano quella che rende minima la funzione

$$f(a, b) \;=\; \sum_{i=1}^{N} \Big[(a + bx_i) - y_i \Big]^2$$

(essendo tutti gli altri termini indipendenti dalle due incognite a e b).

L'interpretazione geometrica è evidente: la retta soluzione del nostro problema è (come già preannunciato) quella che rende minima la somma dei quadrati delle distanze, misurate però *parallelamente all'asse* y, dall'insieme dei punti misurati; queste "distanze" sono anche comunemente chiamate "residui". Per trovare il valore dei coefficienti dell'equazione di tale retta, calcoliamo ora le derivate prime della funzione f:

$$\frac{\partial f}{\partial a} \;=\; 2 \sum_{i=1}^{N} (a + bx_i - y_i) \;=\; 2 \left(Na \;+\; b \sum_{i=1}^{N} x_i \;-\; \sum_{i=1}^{N} y_i \right) \;\; ;$$

$$\frac{\partial f}{\partial b} \;=\; 2 \sum_{i=1}^{N} (a + bx_i - y_i)\, x_i \;=\; 2 \left(a \sum_{i=1}^{N} x_i \;+\; b \sum_{i=1}^{N} {x_i}^2 \;-\; \sum_{i=1}^{N} x_i y_i \right) \;\; .$$

Imponendo che le due derivate prime siano contemporaneamente nulle,

dovranno essere verificate le

$$\begin{cases} a \cdot N & + & b \cdot \sum_i x_i & = & \sum_i y_i \\ a \cdot \sum_i x_i & + & b \cdot \sum_i x_i^2 & = & \sum_i x_i y_i \end{cases} \tag{11.9}$$

e questo sistema di due equazioni in due incognite ammette, come si può verificare, sempre una ed una sola soluzione, purché vi siano almeno due punti sperimentali non coincidenti; esaminando poi le derivate seconde si troverebbe che essa corrisponde in effetti ad un minimo. La soluzione è

$$\begin{cases} a & = & \frac{1}{\Delta}\left[\left(\sum_i x_i^2\right) \cdot \left(\sum_i y_i\right) - \left(\sum_i x_i\right) \cdot \left(\sum_i x_i y_i\right)\right] \\ b & = & \frac{1}{\Delta}\left[N \cdot \left(\sum_i x_i y_i\right) - \left(\sum_i x_i\right) \cdot \left(\sum_i y_i\right)\right] \end{cases} \tag{11.10}$$

in cui si è posto per brevità

$$\Delta = N \sum_i x_i^2 - \left(\sum_i x_i\right)^2$$

(le formule (11.10) sono note sotto il nome di *formule dei minimi quadrati*).

Per quanto attiene al calcolo degli errori commessi nella valutazione di a e b in base ai dati, osserviamo che entrambi si ricavano da relazioni lineari in ognuna delle variabili affette da errore che, nelle nostre ipotesi, sono le sole y_i: possiamo dunque adoperare la formula della propagazione degli errori (10.2), che è in questo caso esatta; oppure la più semplice (5.5). Possiamo esprimere a e b in funzione delle y_i come

$$a = \sum_{i=1}^{N} a_i y_i \qquad \text{e} \qquad b = \sum_{i=1}^{N} b_i y_i$$

una volta posto

$$\begin{cases} a_i & = & \frac{1}{\Delta}\left(\sum_j x_j^2 - x_i \sum_j x_j\right) \\ b_i & = & \frac{1}{\Delta}\left(N x_i - \sum_j x_j\right) \end{cases}$$

e, se indichiamo con σ_y^2 la varianza comune a tutte le y_i, si ottiene, per l'errore di a:

$$
\begin{aligned}
{\sigma_a}^2 &= \textstyle\sum_i {a_i}^2 \, {\sigma_y}^2 \\
&= {\sigma_y}^2 \, \textstyle\sum_i \left[\frac{1}{\Delta} \left(\sum_j {x_j}^2 - x_i \sum_j x_j \right) \right]^2 \\
&= \frac{{\sigma_y}^2}{\Delta^2} \, \textstyle\sum_i \left[\left(\sum_j {x_j}^2 \right)^2 + {x_i}^2 \left(\sum_j x_j \right)^2 - 2 \, x_i \left(\sum_j {x_j}^2 \right) \left(\sum_j x_j \right) \right] \\
&= \frac{{\sigma_y}^2}{\Delta^2} \left[N \left(\sum_j {x_j}^2 \right)^2 + \left(\sum_i {x_i}^2 \right) \left(\sum_j x_j \right)^2 - 2 \left(\sum_j {x_j}^2 \right) \left(\sum_j x_j \right)^2 \right] \\
&= \frac{{\sigma_y}^2}{\Delta^2} \left[N \left(\sum_j {x_j}^2 \right)^2 - \left(\sum_j x_j \right)^2 \left(\sum_j {x_j}^2 \right) \right] \\
&= \frac{{\sigma_y}^2}{\Delta^2} \left(\sum_j {x_j}^2 \right) \left[N \left(\sum_j {x_j}^2 \right) - \left(\sum_j x_j \right)^2 \right] \\
&= {\sigma_y}^2 \, \frac{\sum_j {x_j}^2}{\Delta}
\end{aligned}
$$

e, similmente, per b:

$$
\begin{aligned}
{\sigma_b}^2 &= \textstyle\sum_i {b_i}^2 \, {\sigma_y}^2 \\
&= {\sigma_y}^2 \, \textstyle\sum_i \left[\frac{1}{\Delta} \left(N \, x_i - \sum_j x_j \right) \right]^2 \\
&= \frac{{\sigma_y}^2}{\Delta^2} \, \textstyle\sum_i \left[N^2 {x_i}^2 + \left(\sum_j x_j \right)^2 - 2 \, N \, x_i \sum_j x_j \right] \\
&= \frac{{\sigma_y}^2}{\Delta^2} \left[N^2 \left(\sum_i {x_i}^2 \right) + N \left(\sum_j x_j \right)^2 - 2 \, N \left(\sum_j x_j \right)^2 \right] \\
&= \frac{N \, {\sigma_y}^2}{\Delta^2} \left[N \left(\sum_i {x_i}^2 \right) - \left(\sum_j x_j \right)^2 \right] \\
&= {\sigma_y}^2 \, \frac{N}{\Delta} \; .
\end{aligned}
$$

In definitiva, a e b hanno errori quadratici medi dati dalle

$$\begin{cases} \sigma_a = \sigma_y \sqrt{\dfrac{\sum_i x_i^2}{\Delta}} \\ \\ \sigma_b = \sigma_y \sqrt{\dfrac{N}{\Delta}} \end{cases} \tag{11.11}$$

ed il fatto poi che a e b siano funzioni lineari di variabili che seguono la legge di Gauss ci permette ancora di affermare che anch'esse sono distribuite secondo la legge normale; e di attribuire così ai loro errori il consueto significato statistico.

11.4.2 Stima a posteriori degli errori di misura

È da osservare come nelle formule (11.10) dei minimi quadrati non compaia il valore di σ_y: la soluzione del problema dell'interpolazione lineare è indipendente dall'entità degli errori di misura, nel senso che i coefficienti della retta interpolante possono essere calcolati anche se gli errori sulle y non sono noti (purché naturalmente si assuma che siano tutti uguali tra loro).

Se non è a priori nota la varianza delle y, essa può però essere stimata a partire dai dati stessi una volta eseguita l'interpolazione lineare; infatti gli stessi ragionamenti fatti per le variabili casuali unidimensionali potrebbero essere ripetuti (con le opportune modifiche) sul piano, per giungere a risultati analoghi.

In una dimensione abbiamo a suo tempo potuto collegare l'errore commesso alla dispersione dei dati rispetto al *valore stimato* della grandezza misurata; sul piano è in effetti ancora possibile calcolare l'errore commesso, partendo dalla dispersione dei dati misurata rispetto alla *retta stimata* che passa attraverso di essi: dati disposti mediamente lontano da questa retta indicheranno errori maggiori rispetto a dati ben allineati (e quindi vicini alla retta interpolante).

In una dimensione abbiamo visto che la dispersione dei dati, misurata dal valore medio del quadrato degli scarti rispetto alla loro media aritmetica (nostra migliore stima per la grandezza misurata), era sistematicamente in difetto rispetto alla corrispondente grandezza riferita all'intera popolazione delle misure. Sul piano si può, analogamente, dimostrare che il valore medio del quadrato delle distanze dei punti misurati dalla retta nostra migliore stima è ancora sistematicamente in difetto rispetto alla varianza riferita alla popolazione delle misure ed alla retta vera che corrisponde alla legge fisica reale che collega le due variabili.

Così come abbiamo dimostrato che, al fine di correggere questa sottostima (in media) per le misure ripetute, occorre dividere la somma dei quadrati degli scarti per $N-1$ invece che per N, si potrebbe analogamente dimostrare che una corretta stima dell'errore dei punti misurati si ha, in media, dividendo l'analoga somma per $N-2$; in definitiva, che la corretta stima di σ_y è data dalla formula

$$\sigma_y{}^2 = \frac{\sum_{i=1}^{N}\Big[(a+bx_i)-y_i\Big]^2}{N-2} \quad .$$

In essa a numeratore compare la somma dei quadrati dei residui, cioè delle "distanze" dei punti misurati (x_i, y_i) dalla retta interpolante di equazione $a+bx$ calcolate secondo la direzione parallela all'asse delle ordinate. Questa formula[2] permette una corretta stima dell'errore dei dati interpolati, qualora sia impossibile (o scomodo) determinarli per altra via; l'errore è stimato dai residui dei dati sperimentali, ed è quindi scientificamente affidabile.

Il fatto che la corretta stima dell'errore si ottenga dividendo per $N-2$ invece che per N deve essere messo in relazione con il fatto che gli scarti, invece che rispetto al valore vero, sono calcolati rispetto ad un valore stimato che dipende da *due* parametri, che sono a loro volta stati preventivamente determinati sulla base dei dati sperimentali: cioè i due coefficienti a e b dell'equazione della retta.

Nell'analogo caso della stima dell'errore quadratico medio di una variabile casuale unidimensionale, gli scarti erano calcolati rispetto ad un valore che, unica grandezza necessaria, veniva preventivamente determinato sulla base delle misure: appunto la media aritmetica.

In generale, disponendo di N dati sperimentali dai quali possiamo determinare un valore dell'errore quadratico medio che dipende da M parametri che debbano essere preventivamente derivati dai dati stessi, la modifica da apportare alla formula per ottenere una corretta valutazione dell'errore della popolazione consiste nel dividere la somma dei quadrati degli scarti per un fattore $N-M$.

11.4.3 Interpolazione con una retta per l'origine

Se conosciamo altri vincoli cui debba soddisfare la legge che mette in relazione i valori delle variabili misurate x e y, possiamo imporre che la retta corrispondente appartenga ad un particolare sottoinsieme delle rette del piano; ad esempio, un caso che si può presentare è che la retta sia vincolata

[2] Una formula equivalente (ma più semplice) per il calcolo di σ_y si può trovare nell'equazione (C.8) alla pagina 264.

a passare per una posizione particolare, che supporremo qui essere l'origine degli assi coordinati.

Una generica retta per l'origine ha equazione $y = mx$; ammesso ancora che gli errori commessi riguardino soltanto la misura della y e non quella della x, che tutti i vari y_i abbiano errori distribuiti secondo la legge normale e tra loro uguali, e che le misure siano tra loro indipendenti, il problema dell'interpolazione lineare si riduce a trovare tra le infinite rette passanti per l'origine quella che rende massima la funzione di verosimiglianza

$$\mathcal{L}(x_1, y_1, x_2, y_2, \ldots, x_N, y_N; m) = \prod_{i=1}^{N} \frac{1}{\sigma_y \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{mx_i - y_i}{\sigma_y}\right)^2} .$$

Passando al logaritmo naturale di $\mathcal{L}$, è facile vedere che la soluzione ricercata è sempre quella che rende minima la somma dei quadrati dei residui dai punti misurati: che cioè minimizza la funzione

$$f(m) = \sum_{i=1}^{N} (mx_i - y_i)^2 .$$

La derivata prima di f vale

$$\frac{\mathrm{d}f}{\mathrm{d}m} = 2\sum_{i=1}^{N} (mx_i - y_i)\, x_i = 2\left(m\sum_{i=1}^{N} {x_i}^2 - \sum_{i=1}^{N} x_i y_i\right)$$

e, imponendo che essa sia nulla, l'estremante si ha per

$$m = \frac{\sum_i x_i y_i}{\sum_i {x_i}^2}$$

e corrisponde in effetti ad un minimo. La legge di propagazione degli errori è esatta anche in questo caso, perché m è una combinazione lineare delle variabili affette da errore (le y_i); il coefficiente di y_i nella combinazione vale

$$\frac{x_i}{\sum_k {x_k}^2}$$

e quindi

$${\sigma_m}^2 = \sum_{i=1}^{N} \left(\frac{x_i}{\sum_k {x_k}^2}\right)^2 {\sigma_y}^2 = \frac{{\sigma_y}^2}{\left(\sum_k {x_k}^2\right)^2} \sum_{i=1}^{N} {x_i}^2 = \frac{{\sigma_y}^2}{\sum_k {x_k}^2}$$

e la formula per il calcolo degli errori a posteriori diventa

$${\sigma_y}^2 = \frac{\sum_i (mx_i - y_i)^2}{N-1}$$

visto che il parametro da cui l'errore quadratico medio dipende e che deve essere stimato sulla base dei dati è uno soltanto: m.

11.4.4 Interpolazione lineare nel caso generale

Le condizioni 1) e 2) sugli errori delle grandezze misurate x e y date nel paragrafo 11.4.1 non potranno ovviamente mai essere verificate esattamente; come ci si deve comportare quando nemmeno in prima approssimazione le possiamo considerare vere?

Se gli errori quadratici medi delle y_i sono tra loro diversi, non è più possibile raccogliere a fattore comune $1/\sigma^2$ nell'espressione del logaritmo della verosimiglianza; e ciascun addendo sarà diviso per il corrispondente errore σ_i. In definitiva la retta più verosimile si trova cercando il minimo della funzione

$$f(a,b) = \sum_{i=1}^{N} \left[\frac{(a + bx_i) - y_i}{\sigma_i} \right]^2 \quad .$$

Questo avviene quando

$$\begin{cases} a = \dfrac{1}{\Delta} \left[\left(\sum_i \dfrac{{x_i}^2}{{\sigma_i}^2} \right) \cdot \left(\sum_i \dfrac{y_i}{{\sigma_i}^2} \right) - \left(\sum_i \dfrac{x_i}{{\sigma_i}^2} \right) \cdot \left(\sum_i \dfrac{x_i y_i}{{\sigma_i}^2} \right) \right] \\ b = \dfrac{1}{\Delta} \left[\left(\sum_i \dfrac{1}{{\sigma_i}^2} \right) \cdot \left(\sum_i \dfrac{x_i y_i}{{\sigma_i}^2} \right) - \left(\sum_i \dfrac{x_i}{{\sigma_i}^2} \right) \cdot \left(\sum_i \dfrac{y_i}{{\sigma_i}^2} \right) \right] \end{cases}$$

in cui si è posto

$$\Delta = \left(\sum_i \frac{1}{{\sigma_i}^2} \right) \cdot \left(\sum_i \frac{{x_i}^2}{{\sigma_i}^2} \right) - \left(\sum_i \frac{x_i}{{\sigma_i}^2} \right)^2 \quad .$$

Le varianze di a e di b saranno poi date dalle

$$\begin{cases} {\sigma_a}^2 = \dfrac{1}{\Delta} \sum_i \dfrac{{x_i}^2}{{\sigma_i}^2} \\ {\sigma_b}^2 = \dfrac{1}{\Delta} \sum_i \dfrac{1}{{\sigma_i}^2} \end{cases}$$

Si deve tuttavia osservare che per applicare questo metodo è necessario conoscere, per altra via e preventivamente, tutte le N varianze ${\sigma_i}^2$. Ciò può essere molto laborioso o addirittura impossibile, e non risulta conveniente rinunciare ad una stima unica e ragionevole ${\sigma_y}^2$ di queste varianze per tener conto di una variazione, generalmente debole, delle σ_i in un intervallo limitato di valori della x.

Volendo tener conto dell'errore su entrambe le variabili x ed y, non è generalmente possibile usare un metodo, descritto in alcuni testi, consistente nel cercare la retta che rende minima la somma dei quadrati delle distanze

dai punti, misurate però *ortogonalmente* alla retta stessa: a prescindere dalla complicazione della soluzione di un sistema di equazioni non lineari, resta il fatto che se x ed y sono due grandezze fisiche diverse, o anche soltanto misurate con strumenti e metodi diversi, i loro errori quadratici medi sono generalmente differenti; mentre la distanza sul piano attribuisce lo stesso peso agli scarti in x ed a quelli in y.

Per applicare questo metodo si dovrebbe conoscere, per via indipendente, almeno il rapporto tra σ_x e σ_y; e rappresentare i valori misurati (x_i, y_i) non già sul piano $\{x, y\}$, bensì su quello delle variabili ridotte $\{x/\sigma_x\,,\, y/\sigma_y\}$.

Per solito nella pratica si preferisce considerare affetta da errore una soltanto delle variabili, ad esempio la y, la scelta cadendo generalmente su quella determinata in maniera *più indiretta*, e che risente perciò degli errori di tutte le altre grandezze misurate direttamente; così, in un diagramma velocità-tempo trascorso o velocità-spazio percorso, si assumerà affetta da errore la sola velocità.

Un eventuale errore sulla x si *propagherà* attraverso la relazione funzionale anche alla y, e, se l'errore quadratico medio σ_y è stimato dai dati sperimentali, esso congloberà anche l'indeterminazione dovuta alla x.

Per meglio chiarire il concetto, consideriamo la legge $v = v(t) = v_0 + gt$ che descrive la caduta di un grave, e pensiamo di misurare la sua velocità in un certo istante: nell'ipotesi originale t sarebbe determinabile esattamente, ma l'imprecisione nella misura delle velocità ci darebbe valori di v compresi in un intervallo di ampiezza non nulla (dipendente dall'errore quadratico medio σ_v).

Se delle due grandezze, al contrario, fosse la velocità ad essere conoscibile esattamente, l'impossibilità di determinare con precisione l'istante t in cui essa deve essere misurata ci darebbe ugualmente valori di v distribuiti in un intervallo di ampiezza non nulla (legata stavolta a $g \cdot \sigma_t$).

Indicando, insomma, con σ_x e σ_y gli errori (sempre supposti costanti) di ognuna delle determinazioni (sempre supposte indipendenti) x_i e y_i, la formula dell'errore a posteriori ci permette di ricavare dai dati una ragionevole stima non tanto del solo σ_y quanto, piuttosto, della combinazione (quadratica) dell'errore intrinseco delle ordinate e di quello intrinseco delle ascisse propagato sulle ordinate:

$$\sigma^2 \approx {\sigma_y}^2 + (b^* \, \sigma_x)^2$$

(ove b^* è il valore vero della pendenza della retta).

11.4.5 Interpolazione non lineare

Formule analoghe a quelle trovate si possono ricavare per risolvere il problema dell'interpolazione di curve di ordine superiore al primo (parabole, cubiche, polinomiali in genere) ad un insieme di dati sperimentali, sempre usando il metodo della massima verosimiglianza.

Nel caso poi ci si trovasse di fronte ad una curva di equazione diversa da un polinomio, in parecchi casi è possibile *linearizzare* la relazione cambiando variabile: così, ad esempio, se due grandezze hanno tra loro una relazione di tipo esponenziale, il logaritmo naturale ne avrà una di tipo lineare:

$$y = Ke^{-bx} \qquad \Longleftrightarrow \qquad \ln y \;=\; \ln K - bx \;=\; a - bx \;\; .$$

11.5 Altre applicazioni della stima di massima verosimiglianza

Per concludere il capitolo, presentiamo altre tre applicazioni del metodo della massima verosimiglianza: la stima delle probabilità ignote di un insieme di modalità esclusive ed esaurienti cui può dar luogo un fenomeno casuale; la stima sia della media che della varianza di una popolazione normale; e la stima del range di una popolazione uniforme.

11.5.1 Stima di probabilità

Supponiamo che un fenomeno casuale possa dare origine ad un numero finito M di eventualità, ognuna delle quali sia associata ad un valore p_i ignoto della probabilità; se, eseguite N prove indipendenti, indichiamo con n_i la frequenza assoluta con cui ognuna delle M eventualità si è presentata nel corso di esse, quale è la stima di massima verosimiglianza, $\hat{p}_i$, per le incognite probabilità p_i?

La funzione di verosimiglianza è, visto che la generica delle M eventualità, di probabilità p_i, si è presentata n_i volte, data[3] da

$$\mathcal{L}(\boldsymbol{n};\boldsymbol{p}) = \prod_{i=1}^{M} {p_i}^{n_i}$$

[3] A meno di un fattore moltiplicativo costante, corrispondente al numero di modi in cui N oggetti si possono ripartire tra M gruppi in modo che ogni gruppo sia composto da n_i oggetti; numero delle *partizioni ordinate* (vedi in proposito il paragrafo A.7).

(in cui abbiamo indicato sinteticamente con due vettori, $\boldsymbol{n}$ e $\boldsymbol{p}$, entrambi di dimensione M, l'insieme degli M valori n_i e quello degli M valori p_i rispettivamente); ed il suo logaritmo da

$$\ln \mathcal{L}(\boldsymbol{n};\boldsymbol{p}) = \sum_{i=1}^{M} n_i \ln p_i \ . \tag{11.12}$$

Il problema della ricerca del massimo della (11.12) è complicato dal fatto che i valori delle p_i non sono liberi, ma *vincolati* dalla condizione

$$\sum_{i=1}^{M} p_i = 1 \ . \tag{11.13}$$

Usiamo quindi il metodo dei moltiplicatori di Lagrange, costruendo la funzione

$$f(\boldsymbol{n};\boldsymbol{p}) \ = \ \sum_{i=1}^{M} n_i \ln p_i - \lambda \left(\sum_{i=1}^{M} p_i - 1 \right) \tag{11.14}$$

e risolvendo il sistema delle $M+1$ equazioni, nelle $M+1$ incognite p_i e λ, composto dalla (11.13) e dalle altre M ottenute derivando la (11.14) rispetto ad ognuna delle p_k:

$$\frac{\partial f}{\partial p_k} \ = \ n_k \frac{1}{p_k} - \lambda \ = \ 0 \qquad\qquad (k = 1, 2, \ldots, M) \ .$$

Da quest'ultima si ricava

$$\hat{p}_k = \frac{n_k}{\lambda}$$

e, sostituendo nella (11.13),

$$\sum_{i=1}^{M} \hat{p}_i \ = \ \frac{1}{\lambda} \sum_{i=1}^{M} n_i \ = \ \frac{N}{\lambda} \ = \ 1$$

si ottiene

$$\lambda = N$$

per cui in definitiva la soluzione di massima verosimiglianza è (cosa non sorprendente) data dalle

$$\hat{p}_i = \frac{n_i}{N} \ .$$

11.5.2 Media e varianza di una popolazione normale

Abbiamo già visto nel paragrafo 11.3 che, *ammessa nota la varianza* σ^2 di una popolazione normale, il suo valore medio μ ha come stima di massima verosimiglianza la media aritmetica $\bar{x}$ di un campione di stime indipendenti; vogliamo ora stimare *contemporaneamente sia* μ *che* σ dai dati, usando sempre il metodo della massima verosimiglianza.

La densità di probabilità vale

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

ed il suo logaritmo

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

Il logaritmo della funzione di verosimiglianza è

$$\ln \mathcal{L}(\boldsymbol{x};\mu,\sigma) \;=\; \sum_{i=1}^{N} \ln f(x_i;\mu,\sigma)$$

e dunque

$$\ln \mathcal{L}(\boldsymbol{x};\mu,\sigma) \;=\; -N\ln\sigma - N\,\ln\sqrt{2\pi} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2 \quad ;$$

e le sue derivate parziali prime sono

$$\frac{\partial}{\partial\mu}\ln\mathcal{L} \;=\; \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu) \;=\; \frac{1}{\sigma^2}\left(\sum_{i=1}^{N} x_i - N\mu\right)$$

e

$$\frac{\partial}{\partial\sigma}\ln\mathcal{L} \;=\; -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^2 \;=\; \frac{1}{\sigma^3}\left[\sum_{i=1}^{N}(x_i-\mu)^2 - N\sigma^2\right] \quad .$$

Il sistema ottenuto annullando le due derivate parziali prime ha l'unica soluzione (in effetti un massimo) data da

$$\hat{\mu} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{e} \qquad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu})^2 \quad .$$

Questo era già noto: entrambe le stime, come sappiamo, sono consistenti; però la seconda non è imparziale (ma può essere resa tale moltiplicandola per un opportuno fattore di correzione).

In sostanza il fatto che la varianza della popolazione abbia un determinato valore (come assunto nel paragrafo 11.3) non cambia il fatto che la nostra migliore stima del valore medio della popolazione sia comunque data dalla media aritmetica del campione: vedremo poi nel paragrafo 12.1 che il valore medio del campione e la sua varianza sono variabili casuali *statisticamente indipendenti tra loro*.

11.5.3 Range di una popolazione uniforme

Sia una variabile casuale x distribuita uniformemente tra un estremo inferiore noto, che senza perdere in generalità possiamo supporre sia lo zero, ed un estremo superiore ignoto A; in questo caso dobbiamo innanzi tutto osservare sia che il dominio di definizione della funzione di frequenza $f(x)$ della x *dipende dal parametro* che dobbiamo stimare, sia che $f(x)$ e la sua derivata prima hanno dei punti di discontinuità: e non possiamo in conseguenza garantire *a priori* né la consistenza, né la massima efficienza asintotica del metodo usato.

Comunque, introducendo la cosiddetta *funzione gradino* (o *step function*) $S(x)$, definita attraverso la

$$\begin{cases} S(x) = 0 & (x < 0) \\ S(x) = 1 & (x \geq 0) \end{cases}$$

la funzione di frequenza $f(x)$ si può anche scrivere

$$f(x) \;=\; \frac{1}{A}\, S(x)\, S(A - x)$$

e la funzione di verosimiglianza

$$\mathcal{L}(x_1, x_2, \ldots, x_N; A) = \frac{1}{A^N}\, S(x_{\min})\, S(A - x_{\max}) \quad .$$

Come sappiamo, ammesso noto il valore del parametro A essa rappresenta la densità di probabilità di ottenere gli N valori $x_i \in [-\infty, +\infty]$; se invece si considera A come l'unica variabile e si ammettono noti gli N valori x_i, rappresenta la densità di probabilità che un dato A abbia prodotto i dati osservati. Ma in quest'ultimo caso $S(x_{\min}) \equiv 1$, e la funzione di verosimiglianza si riduce alla

$$\mathcal{L}(A) = \frac{1}{A^N}\, S(A - x_{\max}) \tag{11.15}$$

che è nulla per $A < x_{\max}$ e monotona strettamente decrescente per $A \geq x_{\max}$; ed ammette quindi un unico massimo all'estremo del dominio, che vale

$$\hat{A} = x_{\max} \quad . \tag{11.16}$$

Valore medio e varianza della stima valgono, come già sappiamo dal paragrafo 8.1.3,

$$E(\hat{A}) \;=\; A^* - \frac{A^*}{N+1} \;=\; \frac{N}{N+1} A^*$$

e

$$\mathrm{Var}(\hat{A}) \;=\; \frac{N}{(N+1)^2\,(N+2)}\,(A^*)^2$$

e quindi la stima è consistente, ma non imparziale; una stima imparziale è invece

$$\bar{A} \;=\; \frac{N+1}{N} x_{\max} \;=\; x_{\max}\left(1 + \frac{1}{N}\right) \quad ,$$

di varianza ovviamente superiore per un fattore $(1 + 1/N)^2$. È anche ovvio, dalla forma sia della (11.15) che della (11.16), che $\hat{A}$ è una stima sufficiente di A^*.

11.5.4 Stima della vita media di una particella

Nel processo di decadimento di una particella instabile, indichiamo con τ l'incognita vita media e con t i tempi (propri) di decadimento osservati; tempi che (come sappiamo) seguono la distribuzione esponenziale:

$$f(t;\tau) = \frac{1}{\tau} e^{-\frac{t}{\tau}} \qquad \Longrightarrow \qquad \begin{cases} E(t) = \tau \\ \mathrm{Var}(t) = \tau^2 \end{cases}$$

Ammettendo per semplicità che l'osservazione avvenga con una efficienza unitaria, o, in altre parole, che tutti i decadimenti vengano osservati, la funzione di verosimiglianza si scrive

$$\mathcal{L} \;=\; \prod_{k=1}^{N} f(t_k;\tau) \;=\; \frac{1}{\tau^N} e^{-\frac{1}{\tau}\sum_k t_k} \quad ,$$

ed il suo logaritmo vale

$$\ln(\mathcal{L}) \;=\; -N\ln\tau - \frac{1}{\tau}\sum_{k=1}^{N} t_k \;=\; -N\left(\frac{\bar{t}}{\tau} + \ln\tau\right) \quad .$$

Derivando rispetto al parametro e cercando gli estremanti,

$$\frac{\mathrm{d}\,\ln(\mathcal{L})}{\mathrm{d}\tau} \;=\; \frac{N}{\tau^2}(\bar{t} - \tau) \;=\; 0 \;\;;$$

e quindi l'unico estremante della funzione di verosimiglianza si ha per

$$\hat{\tau} = \bar{t} \quad .$$

Se calcoliamo la derivata seconda,

$$\frac{\mathrm{d}^2\,\ln(\mathcal{L})}{\mathrm{d}\tau^2} = -\frac{N}{\tau^3}\,(2\bar{t} - \tau)$$

essa, calcolata per $t = \bar{t}$ è negativa; quindi l'unico estremante è effettivamente un punto di massimo.

La soluzione di massima verosimiglianza $\hat{\tau} = \bar{t}$ è *consistente ed imparziale* (essendo il valore medio del campione); *di varianza minima* (per il teorema di Cramér–Rao); inoltre la stima è *sufficiente* (riassume insomma tutta l'informazione del campione).

Normalmente l'efficienza non è però unitaria; ad esempio il nostro rivelatore può avere dimensioni confrontabili col cammino medio delle particelle, che possono quindi uscirne prima di decadere. In questo caso, visto che i decadimenti possono essere stati osservati solo essendo avvenuti all'interno di un intervallo compreso tra un valore temporale minimo (eventualmente nullo) ed uno massimo (ad esempio dipendente dalla posizione del decadimento, dalla direzione di emissione dei suoi prodotti, dalle dimensioni del rivelatore, ...) — intervallo differente per ognuno dei decadimenti — dovremo costruire la funzione di verosimiglianza considerando le probabilità di osservazione *condizionate* dall'essere il decadimento i-esimo avvenuto tra un certo $(t_{\min})_i$ ed un certo $(t_{\max})_i$:

$$\mathcal{L} = \prod_{i=1}^{N}\left[\frac{\frac{1}{\tau}e^{-\frac{t_i}{\tau}}}{e^{-\frac{(t_{\min})_i}{\tau}} - e^{-\frac{(t_{\max})_i}{\tau}}}\right] \tag{11.17}$$

Il denominatore della (11.17) rappresenta infatti la probabilità di decadere tra il tempo $t_{\min}$ e quello $t_{\max}$, come è immediato ricavare dalla funzione di distribuzione della densità di probabilità esponenziale, che vale

$$F(t) \;=\; \int_0^t \frac{1}{\tau}\, e^{-\frac{x}{\tau}}\, \mathrm{d}x \;=\; 1 - e^{-\frac{t}{\tau}} \;\;;$$

dalla (11.17) si ricava poi

$$\ln(\mathcal{L}) = -N \ln \tau + \sum_{i=1}^{N} \left\{ -\frac{t_i}{\tau} - \ln\left[e^{-\frac{(t_{\min})_i}{\tau}} - e^{-\frac{(t_{\max})_i}{\tau}} \right] \right\}$$

e, posto per brevità

$$\varphi_i(\tau) = \frac{(t_{\min})_i^2 \cdot e^{-\frac{(t_{\min})_i}{\tau}} - (t_{\max})_i^2 \cdot e^{-\frac{(t_{\max})_i}{\tau}}}{e^{-\frac{(t_{\min})_i}{\tau}} - e^{-\frac{(t_{\max})_i}{\tau}}}$$

si arriva alla

$$\frac{\mathrm{d}\,\ln(\mathcal{L})}{\mathrm{d}\tau} \;=\; \frac{1}{\tau^2} \left\{ \sum_{i=1}^{N} \left[t_i - \varphi_i(\tau) \right] - N\tau \right\} \;=\; 0$$

che bisogna risolvere in modo numerico. Non si può inoltre in questo caso garantire che le proprietà precedentemente delineate per $\hat{\tau}$ (consistenza, normalità, efficienza, ...) siano ancora valide, almeno per N finito. Può darsi che la funzione di verosimiglianza ammetta più di un massimo, e non si sa a priori quale di essi convergerà verso τ^*; e, per finire, l'errore della stima deve essere ricavato dalla concavità della funzione di verosimiglianza, supposta approssimativamente normale.

Capitolo 12

La verifica delle ipotesi (I)

Una volta eseguita una misura, si può voler controllare se i nostri risultati possono confermare o rigettare una determinata ipotesi riguardante il fenomeno fisico che li ha prodotti; naturalmente, visto che risultati di una misura comunque lontani dal valore vero sono sempre possibili (anche se con probabilità sempre più piccole al crescere dello scarto), una qualunque ipotesi sulla grandezza fisica misurata potrà essere confermata o rigettata dai dati solo ad un certo livello di probabilità.

Qui ci occuperemo inoltre di alcune funzioni di frequenza collegate a quella di Gauss, ossia della distribuzione del χ^2, di quella di Student[1] e di quella di Fisher; e dell'uso che di esse si può fare per la verifica di ipotesi statistiche: quali ad esempio quella che un campione di dati sperimentali provenga da una popolazione descritta da una densità di probabilità nota a priori; o quella che il valore vero della grandezza misurata coincida con un valore determinato, noto anch'esso a priori.

12.1 La distribuzione del χ^2

Se le N variabili casuali x_i, tra loro statisticamente indipendenti, sono variabili normali standardizzate (ovverosia distribuite secondo la legge normale con media 0 e varianza 1), si può dimostrare che la nuova variabile

[1] "Student" è lo pseudonimo con cui vennero pubblicati i lavori statistici di William Gosset, scienziato inglese vissuto dal 1876 al 1937. Uno dei pionieri di questo ramo della matematica, svolse le sue ricerche essendo dipendente (prima come chimico, poi come dirigente) della Guinness Brewery di Dublino.

casuale

$$X = \sum_{i=1}^{N} {x_i}^2$$

(ovviamente non negativa) è distribuita con una densità di probabilità data dalla

$$\frac{\mathrm{d}p}{\mathrm{d}X} = f(X;N) = K_N X^{\left(\frac{N}{2}-1\right)} e^{-\frac{X}{2}} \tag{12.1}$$

(*distribuzione del chi quadro*); la costante K_N viene fissata dalla condizione di normalizzazione, ed il parametro N prende il nome di *numero di gradi di libertà* della distribuzione.

La funzione caratteristica della X si può trovare facilmente considerando che, se la x è una variabile normale standardizzata, il suo quadrato $y = x^2$ ha una funzione caratteristica

$$\begin{aligned}
\phi_y(t) &= E\left(e^{ity}\right) \\
&= E\left(e^{itx^2}\right) \\
&= \int_{-\infty}^{+\infty} e^{itx^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \,\mathrm{d}x \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}(1-2it)} \,\mathrm{d}x \\
&= \frac{1}{\sqrt{1-2it}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \,\mathrm{d}u \\
&= (1-2it)^{-\frac{1}{2}}
\end{aligned}$$

(si è eseguita la sostituzione di variabile $u = x\sqrt{1-2it}$; l'integrale definito è quello di una distribuzione normale $N(u;0,1)$ e vale dunque 1). Di conseguenza, applicando l'equazione (6.11), la funzione caratteristica della X vale

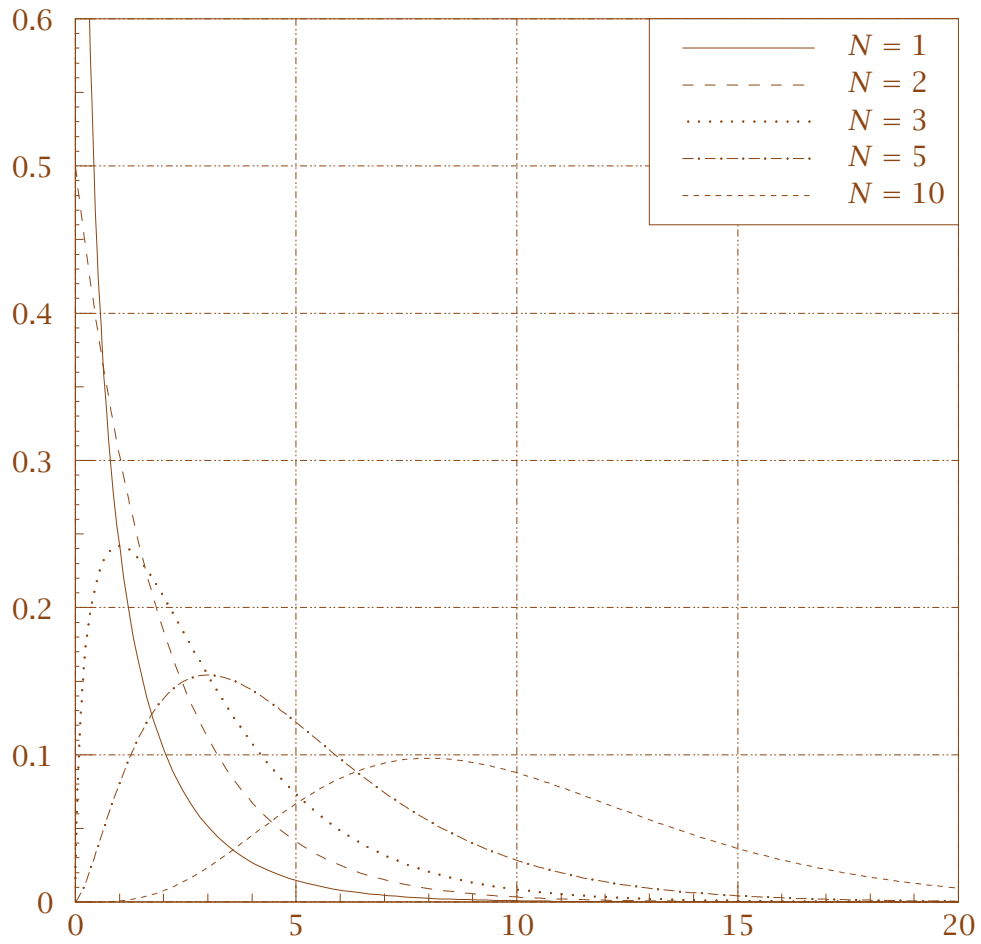
$$\phi_X(t) = (1-2it)^{-\frac{N}{2}} \ . \tag{12.2}$$

Per dimostrare che la funzione di frequenza della X è effettivamente la (12.1), si parte poi dall'espressione (12.2) della funzione caratteristica e le si applica la trasformazione inversa di Fourier già definita nella (6.10).

Con simili passaggi si potrebbe ricavare la funzione generatrice dei momenti, che vale

$$M_X(t) = (1-2t)^{-\frac{N}{2}}$$

Figura 12a - La distribuzione del χ^2 per alcuni valori del parametro N.



e, da queste, si ottiene infine che il valore medio e la varianza di una variabile casuale distribuita come il χ^2 a N gradi di libertà sono

$$E(X) = N \qquad \text{e} \qquad \mathrm{Var}(X) = 2N$$

mentre i coefficienti di asimmetria e di curtosi valgono

$$\gamma_1 = 2\sqrt{\frac{2}{N}} \qquad \text{e} \qquad \gamma_2 = \frac{12}{N} \ .$$

La distribuzione del χ^2 tende asintoticamente ad una distribuzione normale con la stessa media N e la stessa varianza $2N$; infatti la funzione caratteristica della variabile standardizzata

$$y = \frac{X - N}{\sqrt{2N}} = \frac{X}{\sqrt{2N}} - \frac{N}{\sqrt{2N}}$$

vale, ricordando la (6.17),

$$\phi_y(t) = e^{-\frac{iNt}{\sqrt{2N}}} \left[1 - \frac{2it}{\sqrt{2N}}\right]^{-\frac{N}{2}} \ .$$

Passando ai logaritmi naturali,

$$\ln \phi_y(t) = -\frac{iNt}{\sqrt{2N}} - \frac{N}{2} \ln\left(1 - \frac{2it}{\sqrt{2N}}\right)$$

e, sviluppando in serie di McLaurin il logaritmo,

$$\begin{aligned} \ln \phi_y(t) &= -\frac{iNt}{\sqrt{2N}} - \frac{N}{2}\left[-\frac{2it}{\sqrt{2N}} - \frac{1}{2}\left(\frac{2it}{\sqrt{2N}}\right)^2 + \mathcal{O}\left(N^{-\frac{3}{2}}\right)\right] \\ &= -\frac{t^2}{2} + \mathcal{O}\left(N^{-\frac{1}{2}}\right) \end{aligned}$$

da cui

$$\lim_{N\to\infty} \phi_y(t) = e^{-\frac{t^2}{2}}$$

che è appunto la funzione caratteristica di una distribuzione normale standardizzata.

In definitiva:

- Quando N assume valori sufficientemente grandi, la distribuzione del χ^2 è ben approssimata da una distribuzione normale avente la stessa media N e la stessa varianza $2N$; tale approssimazione si può ritenere in pratica già buona quando N è superiore a 30.

- Inoltre si potrebbe analogamente dimostrare che la variabile casuale $\sqrt{2X}$, anche per valori relativamente piccoli di N, ha una distribuzione che è assai bene approssimata da una funzione normale con media $\sqrt{2N-1}$ e varianza 1; l'approssimazione è già buona per $N \gtrsim 8$.

Dalla definizione (o dalla funzione caratteristica (12.2)) discende immediatamente la cosiddetta *regola di somma del* χ^2: ossia che, se X ed Y sono due variabili casuali statisticamente indipendenti entrambe distribuite come il χ^2, con N ed M gradi di libertà rispettivamente, la loro somma $Z = X + Y$ è una variabile casuale ancora distribuita come il χ^2; però con $N + M$ gradi di libertà.

Ovviamente, se le x_i (con $i = 1, \ldots, N$) sono N variabili casuali statisticamente indipendenti tra loro e provenienti da una stessa distribuzione normale con media μ e varianza σ^2, discende da quanto detto che la nuova variabile casuale

$$X' = \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma} \right)^2$$

è distribuita come il χ^2 a N gradi di libertà. Indichiamo ora, al solito, con $\bar{x}$ la media aritmetica delle x_i: vogliamo dimostrare che la variabile casuale

$$X'' = \sum_{i=1}^{N} \left(\frac{x_i - \bar{x}}{\sigma} \right)^2$$

è distribuita *ancora come il* χ^2*, ma con* $N - 1$ *gradi di libertà.*

A questo scopo facciamo dapprima alcune considerazioni, indipendenti dalle ipotesi prima fatte sulle x_i e che risultano quindi valide per variabili casuali qualunque: supponiamo di definire N nuove variabili y_i come generiche combinazioni lineari delle x_j, con coefficienti che indicheremo col simbolo A_{ij}; in modo insomma che risulti

$$y_i = \sum_{j=1}^{N} A_{ij}\, x_j \quad .$$

La somma dei quadrati delle y_i è data da

$$\sum_{i=1}^{N} {y_i}^2 = \sum_{i=1}^{N} \left(\sum_{j=1}^{N} A_{ij}\, x_j \right) \left(\sum_{k=1}^{N} A_{ik}\, x_k \right) = \sum_{jk} x_j\, x_k \sum_i A_{ij}\, A_{ik} \quad ;$$

è possibile che questa somma risulti uguale alla somma dei quadrati delle x_i *qualunque* sia il valore di queste ultime? Ovviamente questo avviene se e

solo se vale la

$$\sum_i A_{ij} A_{ik} = \delta_{jk} = \begin{cases} 0 & \text{per } j \neq k \\ 1 & \text{per } j = k \end{cases} \tag{12.3}$$

(il simbolo δ_{jk}, che assume il valore 1 quando gli indici sono uguali e 0 quando sono invece diversi, si chiama *simbolo di Kronecker* o *delta di Kronecker*).

Consideriamo gli A_{ij} come gli elementi di una matrice quadrata $\boldsymbol{A}$ di ordine N; gli x_j e le y_i si possono invece considerare come le componenti di due *vettori* $\boldsymbol{X}$ ed $\boldsymbol{Y}$ definiti in uno spazio N-dimensionale — ossia come gli elementi di due matrici rettangolari con N righe ed 1 colonna.

La trasformazione che muta $\boldsymbol{X}$ in $\boldsymbol{Y}$ si può scrivere, in forma matriciale, come $\boldsymbol{Y} = \boldsymbol{AX}$; la somma dei quadrati delle x_j o delle y_i altro non è se non il prodotto scalare, di $\boldsymbol{X}$ ed $\boldsymbol{Y}$ rispettivamente, per loro stessi: ovverosia la loro *norma*, il quadrato della loro lunghezza nello spazio a N dimensioni. Quella che abbiamo ricavato adesso è la condizione perché una *trasformazione lineare* applicata ad un vettore ne conservi la lunghezza: occorre e basta che la matrice $\boldsymbol{A}$ sia *ortogonale*. Infatti la (12.3) si può scrivere

$$\tilde{\boldsymbol{A}}\boldsymbol{A} = \mathbf{1} \qquad \text{ossia} \qquad \tilde{\boldsymbol{A}} = \boldsymbol{A}^{-1}$$

($\tilde{\boldsymbol{A}}$ è la matrice trasposta di $\boldsymbol{A}$, di elementi $\tilde{\boldsymbol{A}}_{ij} = A_{ji}$; $\mathbf{1}$ è la matrice unità, di elementi $\mathbf{1}_{ij} = \delta_{ij}$; $\boldsymbol{A}^{-1}$ è la matrice inversa di $\boldsymbol{A}$; ed una matrice per cui $\tilde{\boldsymbol{A}} = \boldsymbol{A}^{-1}$ si dice, appunto, ortogonale).

Consideriamo adesso una trasformazione lineare definita dalle seguenti relazioni:

$$\begin{cases} y_1 = \dfrac{1}{\sqrt{N}}(x_1 + x_2 + \cdots + x_N) \\ y_2 = \dfrac{1}{\sqrt{2}}(x_1 - x_2) \\ y_3 = \dfrac{1}{\sqrt{6}}(x_1 + x_2 - 2x_3) \\ \cdots \\ y_N = \dfrac{1}{\sqrt{N(N-1)}}\left[x_1 + x_2 + \cdots + x_{N-1} - (N-1)x_N\right] \end{cases} \tag{12.4}$$

e per la quale la matrice di trasformazione abbia, insomma, elementi A_{ij}