Capitolo 8

Stima dei Parametri Metodo di Massima Verosimiglianza

Lo scopo dello studio dei fenomeni fisici è quello di scoprire le leggi che legano le grandezze studiate e di misurare il valore delle costanti che compaiono della formulazione matematica delle leggi fisiche. Il primo passo in questa analisi consiste nell'ipotizzare una relazione matematica tra le grandezze fisiche che caratterizzano il fenomeno osservato. Questa funzione matematica costituisce il cosiddetto "modello matematico", brevemente "modello" del fenomeno studiato. Con modello si intende una forma matematica (lineare, quadratica, esponenziale,... o una loro combinazione) nella quale compaiono un certo numero di parametri incogniti $\lambda_1, \lambda_2, \ldots, \lambda_k$ (il cui insieme sarà indicato con Λ per brevità). Ottenere, o come si suole dire, stimare il valore di questi parametri è, in molti casi, l'obiettivo principale dello studio del fenomeno fisico.

La stima del valore dei parametri che caratterizzano una popolazione statistica è un'importante capitolo della statistica detto *teoria degli stimatori*. La teoria degli stimatori si divide in *puntuale*, quando valuta il valore di un parametro, e in stima di *intervallo* o intervallare, quando assegna un intervallo di valori che includa il parametro cercato con un prefissato livello di confidenza[1]

Come esempio si consideri un esperimento in cui per stimare l'accelerazione di gravità locale si studia il moto di un grave in caduta libera; le grandezze fisiche osservate sono un certo numero di distanze percorse dal grave in corrispondenza di tempi fissati (s_i, t_i), $i = 1, \ldots, N$. La teoria, in particolare il secondo principio della dinamica, stabilisce che queste grandezze soddisfano la relazione[2] $s = gt^2/2$ dove g è l'accelerazione di gravità; il modello matematico da adattare ai dati sarà quindi una forma quadratica: $s = \lambda_1 + \lambda_2 t + \lambda_3 t^2$ dove il parametro λ_1 tiene conto di una eventuale valore non nullo della reale coordinata di inizio del moto, λ_2 tiene conto di un'eventuale velocità iniziale non nulla e infine λ_3 è il parametro che stima l'accelerazione di gravità diviso 2. Se i dati sperimentali sono in numero sufficiente (possibilmente in numero molto maggiore dei parametri Λ) la teoria degli estimatori fornisce gli strumenti per stimare i parametri dai dati sperimentali.

Nei paragrafi successivi saranno esposti i due metodi più usati per la stima puntuale dei parametri: il metodo di *massima verosimiglianza* e il metodo dei *minimi quadrati*.

---

[1] Si parla di *confidenza* o *fiducia* e non di *probabilità* perché nell'interpretazione frequentista della probabilità i parametri delle teorie hanno valori fissi per quanto non noti e quindi *non ammettono una distribuzione di probabilità*. L'interpretazione soggettivista della probabilità al contrario, ammette che i parametri incogniti in quanto tali abbiano una distribuzione di probabilità.

[2] Supponendo velocità iniziale e coordinata di inizio del moto nulle

8.1 Il metodo di Massima Verosimiglianza

Esempio introduttivo. Per introdurre il concetto della *verosimiglianza* (ing. *Likelihood*) in statistica e del suo uso nella valutazione dei parametri consideriamo un semplice esperimento consistente in N lanci di un dado e definiamo come evento favorevole l'uscita del numero "6". Come è noto, la distribuzione del numero di eventi favorevoli segue la distribuzione binomiale[3]

$$\mathcal{B}(k|p) = \binom{N}{k} p^k (1-p)^{N-k} \tag{8.1}$$

dove p è la probabilità del singolo evento e, fissato il numero N degli eventi, p è l'unico parametro della distribuzione. Se il dado è ben equilibrato possiamo supporre $p = 1/6$ e la forma della distribuzione di probabilità (8.1) è completamente determinata. Ad esempio, tramite la (8.1), possiamo calcolare la probabilità che in 10 lanci del dado esca un solo 6 (il 32%), oppure che ne escano tre (il 16%). La distribuzione di probabilià binomiale (8.1) con $N = 10$ e $p = 1/6$ è mostrata nella figura 8.1. In generale potremo affermare che

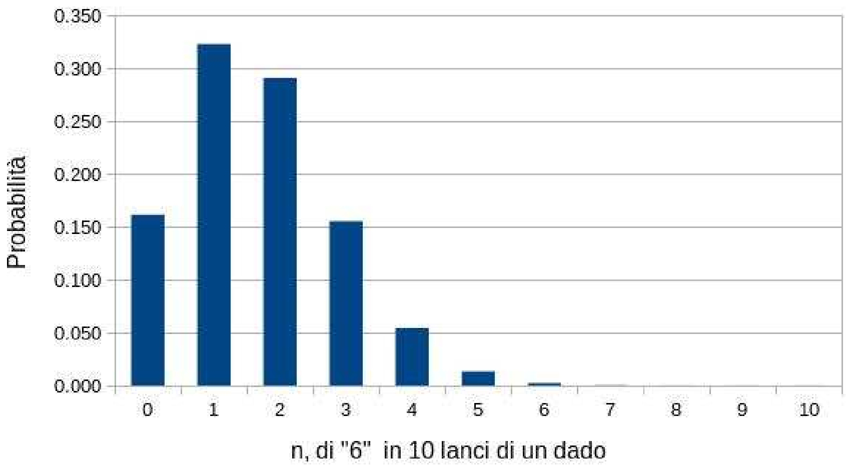

Figura 8.1: Distribuzione binomiale di probabilità per $N = 10$ e $p = 1/6$

se conosciamo la distribuzione di probabilità e il valore dei parametri che la caratterizzano siamo in grado di conoscere la probabilità con cui si realizzano gli eventi.

Tuttavia il problema che si deve affrontare nell'attività sperimentale è quello inverso, e cioè *noti i dati si vuole ottenere la stima dei parametri*. Continuando con l'esempio del dado, supponiamo che si abbia il sospetto che il dado sia "truccato" (ad esempio appesantendo una faccia) e che quindi la probabilità del "6" sia p, differente da 1/6. Eseguiamo quindi un esperimento lanciando il dado N volte e sia $k = k_o$ il numero di volte che è uscito il "6". Un modo di stimare il valore di p, consiste nel calcolare la probabilità assegnata all'evento osservato come una funzione del *parametro* p avendo fissato il valore della variabile aleatoria a quello osservato ($k = k_o$):

$$L(p|k_0) = \binom{N}{k_0} p^{k_0} (1-p)^{N-k_0} \tag{8.2}$$

la funzione $L(p|k_0)$, detta *funzione di verosimiglianza*, descrive l'informazione guadagnata eseguendo l'esperimento. La funzione di verosimiglianza ha la stessa forma algebrica della

[3] Sulla distribuzione binomiale vedi il paragrafo 5.8.2

distribuzione di probabilità (8.1) ma ha un significato totalmente differente, infatti la funzione L dipende unicamente dai parametri (in questo esempio p) mentre le variabili aleatorie prendono un valore fisso ($k = k_0$). Nella figura 8.2 è mostrato l'andamento della funzione di verosimiglianza (8.2) per due valori differenti di k_o.

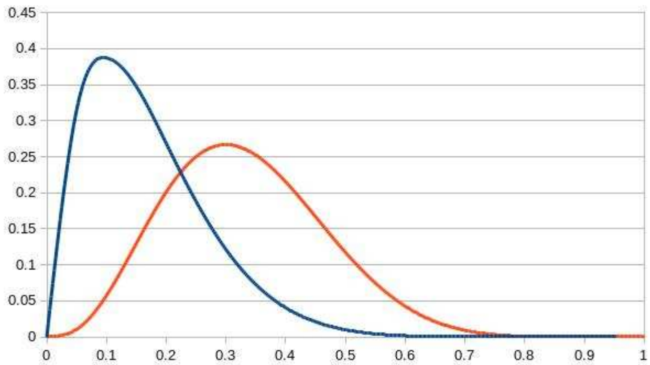

Figura 8.2: Funzione di verosimiglianza (8.2) in funzione del parametro p. Le due curve di verosimiglianza mostrate corrispondono a due serie di 10 lanci di un dado in cui si sono avuti $k_0 = 1$ eventi favorevoli (curva con il massimo a $p = 0.1$) e $k_0 = 3$ eventi favorevoli (curva con il massimo a $p = 0.3$).

Il criterio per stimare p, noto come il *criterio di massima verosimiglianza*, afferma che l'evento che si è realizzato (in questo caso esemplificato da $k = k_0$) ha la massima probabilità di accadere (infatti è accaduto) e la stima del parametro p (con il metodo di massima verosimiglianza) si ottiene *massimizzando la funzione di verosimiglianza* (8.2) in funzione di p. Tornando all'esempio, calcoliamo il valore di p per cui la funzione di verosimiglianza (8.2) ha il suo massimo. Questo valore sarà la stima di massima verosimiglianza ($\hat{p}$) del parametro p. Per semplificare i calcoli è consuetudine utilizzare al posto della funzione L il suo logaritmo[4]: $\mathcal{L} = \ln L$. Nel nostro esempio:

$$\mathcal{L} = \ln\left[\binom{N}{k_0} p^{k_0}(1-p)^{N-k_0}\right] = k_o \ln p + (N - k_0)\ln(1-p) + \ln\binom{N}{k_0}$$

Per trovare il massimo di $\mathcal{L}$ annulliamo la sua derivata prima:

$$\frac{d\mathcal{L}}{dp} = \frac{k_0}{p} - \frac{N-k_0}{1-p} = 0, \quad \text{da cui} \quad \hat{p} = \frac{k_0}{N}$$

Il valore $\hat{p}$ così ottenuto è la *stima di massima verosimiglianza del parametro* p.

Caso generale. Generalizziamo l'esempio precedente considerando un evento definito da N grandezze aleatorie $X_1, X_2, \ldots, X_N$ indipendenti[5] tra loro e siano $f_i(X_i|\Lambda)$ le distribuzioni di probabilità che danno la probabilità di osservare X_i dato Λ; $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$ rappresenta l'insieme degli m parametri da cui dipendo le f_i.

Data l'indipendenza delle realizzazioni X_i la probabilità di osservare una particolare realizzazione delle X_i, che indichiamo con $\{x_1, x_2, \ldots, x_N\}$ è proporzionale a:

$$f_1(x_1|\Lambda) \cdot f_2(x_2|\Lambda) \cdot \ldots f_n(x_N|\Lambda) = \prod_{i=1}^{N} f_i(x_i|\Lambda) \tag{8.3}$$

[4]Si può facilmente dimostrare che L e $\mathcal{L} = \ln L$ hanno massimi (e minimi) per gli stessi valori delle variabili da cui dipendono

[5]La condizione di indipendenza non è essenziale per nella definizione del metodo di massima verosimiglianza ma ne semplifica la formulazione matematica

se in questa relazione le x_i sono gli esiti di un esperimento, ovvero uno specifico campione statistico, le x_i hanno un valore numerico definito e la relazione (8.3) diviene una funzione dei soli parametri Λ. Questa funzione, in analogia con l'esempio precedente, è detta funzione di *verosimiglianza*:

$$L = \prod_{i=1}^{N} f(x_i|\Lambda) \tag{8.4}$$

Come già anticipato nell'esempio introduttivo, in luogo della (8.4) si usa definire come funzione di verosimiglianza il logaritmo della (8.4):

$$\mathcal{L} = \ln L = \ln \prod_{i=1}^{N} f(x_i|\Lambda) = \sum_{i=1}^{N} \ln f(x_i|\Lambda) \tag{8.5}$$

Il criterio di massima verosimiglianza. Questo criterio, originariamente sviluppato da R. A. Fisher nel 1922, stabilisce che

> i valori preferiti dei parametri di una funzione di verosimiglianza sono quelli che rendono massima la probabilità di ottenere i dati osservati.

quindi le stime dei parametri con il metodo di massima verosimiglianza sono le soluzioni del seguente sistema:

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \lambda_1} = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda_2} = 0 \\ \dots \\ \frac{\partial \mathcal{L}}{\partial \lambda_m} = 0 \end{cases}$$

La stima dei parametri che si ottengono massimizzando la funzione (8.5) è detta *stima di massima verosimiglianza* dei parametri.

Esempi.

1. *Stima del valore medio della gaussiana.* Consideriamo n misurazioni ripetute di una grandezza $\{x_1, x_2, \dots, x_n\}$ che sappiamo avere una distribuzione normale di varianza nota σ e valore medio da determinare. Supponiamo inoltre che le misurazioni ripetute della grandezza siano indipendenti. Determiniamo la stima di massima verosimiglianza del valore medio μ. Da quanto detto la funzione di verosimiglianza in questo caso è

   $$\mathcal{L}(\mu) = \ln \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x_i-\mu)^2/2\sigma^2} = \sum_{i}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2} + \frac{n}{2} \ln 2\pi\sigma^2 \tag{8.6}$$

   Derivando la funzione di verosimiglianza rispetto a μ e annullando la derivata si ha:

   $$\frac{d\mathcal{L}}{d\mu} = \sum_{i}^{n} \frac{(x_i-\mu)}{\sigma^2} = 0, \quad \text{da cui} \quad \hat{\mu} = \frac{1}{n} \sum_{i}^{n} x_i$$

   La stima di massima verosimiglianza del valore medio di una gaussiana $\hat{\mu}$ è data dalla media campionaria del campione di dati acquisiti.

2. *Stima della deviazione standard della gaussiana.* Riprendiamo l'esempio precedente supponendo che sia la varianza σ il parametro da determinare. Derivando rispetto a σ l'espressione della funzione di verosimiglianza (8.6) e annullando la derivata si ha:

$$\frac{d\mathcal{L}}{d\sigma} = \sum_i^n \frac{(x_i - \mu)^2}{\sigma^3} + \frac{n}{\sigma} = 0, \qquad \text{da cui} \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_i^n (x_i - \mu)^2}$$

Si noti che in questo caso la stima di massima verosimiglianza della deviazione standard è *distorta* se, come succede nella maggior parte dei casi, il valore medio μ è ignoto ed è sostituito con la sua stima $\hat{\mu}$.

3. *Media pesata.* La formula della media pesata già usata precedentemente può essere ottenuta dal metodo di massima verosimiglianza. Per semplicità consideriamo solo due misure indipendenti $x_1 \pm u_1, x_2 \pm u_2$ di una grandezza supponendo inoltre che le due misure siano distribuite in modo normale con parametri σ dati rispettivamente da u_1 e u_2.

Indichiamo con μ il valore "vero" della grandezza, che è il parametro da stimare, le pdf delle variabili aleatorie x_1 e x_2 sono:

$$\frac{1}{\sqrt{2\pi}u_1}\mathrm{e}^{-(x_1-\mu)^2/2u_1^2}; \qquad \frac{1}{\sqrt{2\pi}u_2}\mathrm{e}^{-(x_2-\mu)^2/2u_2^2}$$

La funzione di verosimiglianza $\mathcal{L}$ è

$$\mathcal{L}(\mu) = \ln\left[\frac{1}{2\pi}\frac{1}{u_1u_2}\mathrm{e}^{-\frac{(x_1-\mu)^2}{2u_1^2}}\mathrm{e}^{-\frac{(x_2-\mu)^2}{2u_2^2}}\right] = -\left[\frac{(x_1-\mu)^2}{2u_1^2} + \frac{(x_2-\mu)^2}{2u_2^2}\right] + \text{cost.}$$

Per trovare il massimo di $\mathcal{L}$ uguagliamo a zero la derivata rispetto a μ

$$\frac{d}{d\mu}\mathcal{L}(\mu) = -\frac{d}{d\mu}\left(\frac{(x_1-\mu)^2}{2u_1^2} + \frac{(x_2-\mu)^2}{2u_2^2}\right) = 0$$

da cui, ponendo $w_i = 1/u_i^2$, $(i = 1,2)$ si ottiene $w_1(x_1 - \mu) + w_2(x_2 - \mu) = 0$, e risolvendo per μ, si ha

$$\hat{\mu} = \frac{w_1x_1 + w_2x_2}{w1 + w2}$$

che è proprio la formula della media pesata.

ESERCIZIO. Calcolare la stima di massima verosimiglianza di due misure $x_1 \pm u_1, x_2 \pm u_2$ entrambe distribuite in modo normale di una stessa grandezza supponendo che siano correlate con coefficiente di correlazione ρ. Per il calcolo utilizzare la relazione (5.52).

4. *Processo di Poisson.* Consideriamo un processo governato dalla statistica di Poisson come ad esempio il numero di conteggi di un contatore di raggi cosmici in un fissato intervallo temporale. Siano $k_1, k_2, \ldots, k_n$ n realizzazioni del processo di Poisson considerato; indicando con μ il parametro della distribuzione di Poisson, la funzione di verosimiglianza si scrive:

$$\mathcal{L}(\mu) = \ln\left(\prod_i^n \frac{\mu^{k_i}}{k_i!}\mathrm{e}^{-\mu}\right) = \left(\sum_i k_i\right)\ln\mu - n\mu - \sum_i \ln k_i!$$

Derivando rispetto a μ e annullando la derivata si ottiene:

$$\frac{d\mathcal{L}}{d\mu} = \frac{\sum_i^n k_i}{\mu} - n = 0, \qquad \text{da cui} \quad \hat{\mu} = \frac{1}{n}\sum_i^n k_i \tag{8.7}$$

La stima di massima verosimiglianza del parametro μ della distribuzione di Poisson coincide con la media campionaria.