Stima puntuale

Problema 1: Un'azienda produce pacchi di pasta che hanno un peso dichiarato di 500 grammi. Il processo di confezionamento della pasta è tale che il peso di ciascun pacco non sia determinabile con esattezza a priori. Il peso reale di ciascun pacco, quindi, può non essere *esattamente* pari a 500 grammi. Un peso compreso tra 480 e 520 grammi è considerato accettabile, al di fuori di questa fascia, invece, il pacco è considerato difettoso. L'addetto al controllo di qualità dell'azienda vuole farsi un'idea di quale sia la proporzione θ di pacchi difettosi. Si decide quindi di selezionare un campione casuale (estratto con ripetizione) dei pacchi prodotti un una data settimana e dall'esame di questi cercare di risalire alla vera proporzione θ di pacchi difettosi. Vengono selezionati $n = 40$ pacchi di pasta, 5 dei quali risultano essere difettosi.

Che informazione questi dati possono dare riguardo alla proporzione incognita θ? $\diamond\diamond$

Soluzione 1: In generale θ può essere un qualsiasi numero compreso tra 0 (tutti i pacchi di pasta prodotti sono accettabili) e 1 (i pacchi prodotti sono tutti difettosi).

La prima conseguenza della lettura dei dati è che θ non può essere 0 né 1. Questo perché nel caso $\theta = 0$ sarebbe stato impossibile osservare un campione con 5 pacchi difettosi e, analogamente, in caso di $\theta = 1$ sarebbe stato impossibile osservare 35 pacchi accettabili.

Con minore certezza, si può comunque ritenere che dopo aver osservato il campione nessuno si aspetti un valore di θ pari a 0.99. Questo perché se il 99% dei pacchi prodotti fosse difettoso, nel campione estratto ci si sarebbe aspettato di avere ben più di 5 pacchi difettosi.

Alla base di queste considerazioni di buon senso, c'è un implicito ragionamento probabilistico: in caso di $\theta = 0.99$, il campione effettivamente osservato sarebbe un campione piuttosto anomalo che nessuno si sarebbe immaginato potesse essere generato da quella popolazione. Insomma, alla base di quel ragionamento c'è una valutazione di quanto (im)probabile

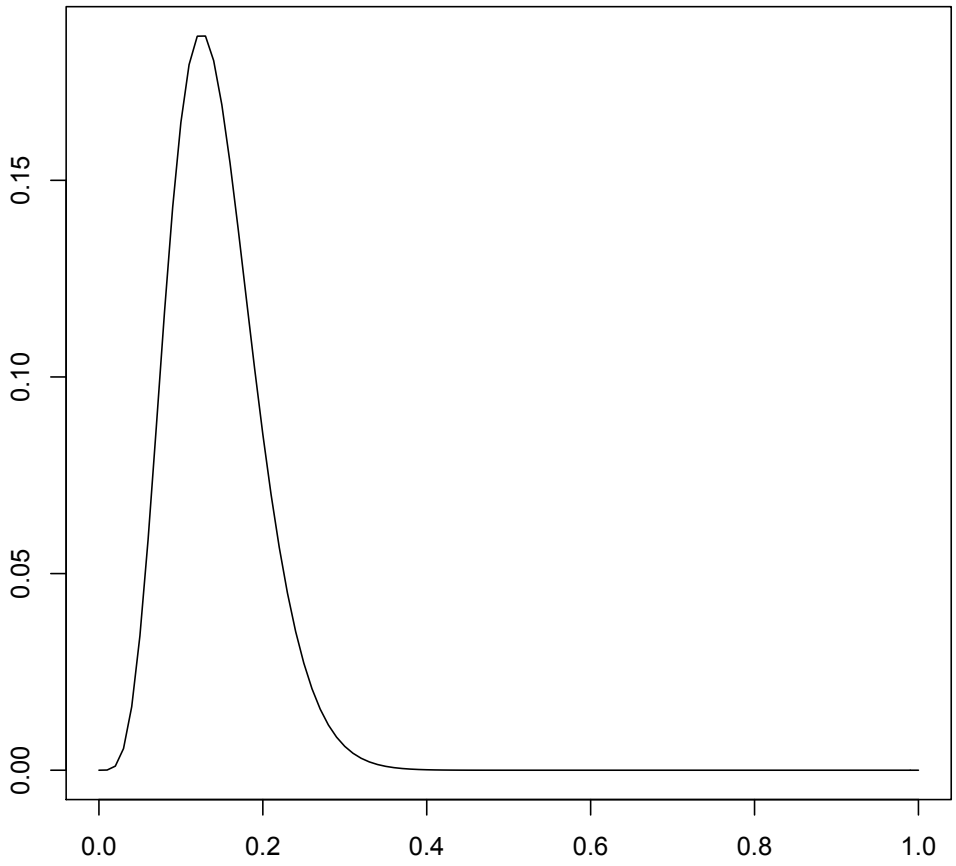

Figura 1: Probabilità del campione osservato in funzione di θ.

sia il campione osservato in presenza di una proporzione di pacchi difettosi $\theta = 0.99$. Al valore $\theta = 0.99$ si associa una probabilità estremamente bassa di ottenere un campione con 5 difettosi e 35 accettabili e per questo –alla luce del campione osservato– il valore $\theta = 0.99$ è decisamente poco plausibile. Lo stesso discorso vale per tutti i θ "alti" quindi, ad esempio per tutti i $\theta \geq 0.90$.

Cercando di estendere il ragionamento a tutti i possibili θ: la plausibilità di ciascun possibile valore di θ è valutata in base alla corrispondente probabilità del campione osservato. In particolare, nell'esempio in questione,

$$P(\text{5 difettosi su 40 osservati}) = \binom{40}{5}\theta^5(1-\theta)^{35} \qquad (1)$$

che, vista come funzione della proporzione incognita θ, ha l'andamento riportato in Figura 1. Dal grafico emerge che valori di θ maggiori o uguali a 0.4 generano un campione come quello osservato con probabilità pres-

soché nulla e pertanto sono decisamente poco plausibili. Lo stesso vale per valori di θ vicini a 0. Il valori più attendibili sembrano essere quelli compresi tra 0.1 e 0.2.

Dovendo alla fine produrre un unico valore, quello che i dati osservati fanno emergere come più plausibile, sembra a questo punto naturale cercare di individuare il $\hat{\theta}$ che massimizza la probabilità in (1). Determiniamo quindi il punto di massimo studiando il segno della derivata della probabilità in (1):

$$\begin{aligned}\frac{d}{d\theta}P(\text{5 difettosi su 40 osservati}) &= \binom{40}{5}\left[5\theta^4(1-\theta)^{35}+35\theta^5(1-\theta)^{34}\right] \\ &= \binom{40}{5}\theta^4(1-\theta)^{34}\left[5(1-\theta)+35\theta\right]\end{aligned}$$

e vediamo facilmente che il punto di massimo è in $\hat{\theta}=5/40$.

La conclusione quindi è che, alla luce dei dati campionari osservati, il valore θ più plausibile è $\hat{\theta}=5/40$. $\diamond\diamond$

Osservazione: Verrebbe da dire che abbiamo scoperto l'acqua calda. In effetti la risposta finale è quella più ovvia, che chiunque avrebbe dato senza aver seguito un corso di Statistica. A parte il fatto che non sempre succede che la risposta finale è quella più ovvia, questo esempio volutamente semplice dovrebbe aver chiarito il processo logico da seguire: i diversi valori che il parametro può assumere sono considerati più o meno plausibili a seconda di quanto vale – in corrispondenza di ciascun valore – la probabilità del campione effettivamente osservato. Quanto più alta è la probabilità del campione osservato tanto più deve essere considerato plausibile il corrispondente valore del parametro. $\diamond\diamond$

Problema 2: Nel 1998 a Berkeley il numero di incidenti stradali in 10 giornate senza pioggia scelte a caso è stato: 4, 0, 6, 5, 2, 1, 2, 0, 4, 3.

Usare i dati campionari per cercare di dedurre la frazione di giornate senza pioggia del 1998 in cui non si verifica al più un incidente.

Siccome gli automobilisti sono tantissimi, ciascuno con una probabilità bassa di essere coinvolto in un incidente, è ragionevole assumere che il numero di incidenti in una giornata senza pioggia segua una distribuzione di Poisson con parametro $\lambda > 0$ incognito. $\diamond\diamond$

Problema 3: Si dispone di un algoritmo che dovrebbe generare sequenze di numeri (pseudo)casuali indipendenti secondo una distribuzione uniforme sull'intervallo $[0, 1]$. Si dubita però del fatto che i numeri generati coprano tutto l'intervallo unitario. In particolare si sospetta che le sequenze siano indipendenti, ma con distribuzione uniforme su $[0, \theta]$, dove $\theta < 1$. Da una lunga sequenza $x_1, \cdots, x_n$ di dati generati si vuol cercare di dedurre il valore di θ. $\diamond\diamond$

Un campione casuale $X_1, \cdots, X_n$ è generato da una popolazione la cui distribuzione dipende da un parametro θ (poi vedremo anche qualche caso in cui sono presenti due o più parametri).

Indichiamo con $x_1, \cdots, x_n$ i valori osservati. Tanto per chiarire, nell'esempio degli incidenti di Berkeley

$$x_1 = 4 \quad x_2 = 0 \quad x_3 = 6 \quad x_4 = 5 \quad x_5 = 2$$
$$x_6 = 1 \quad x_7 = 2 \quad x_8 = 0 \quad x_9 = 4 \quad x_{10} = 3$$

Abbiamo in precedenza argomentato che valutiamo la plausibilità di ciascun valore del parametro attraverso la corrispondente probabilità del campione osservato e consideriamo come più plausibile il valore $\hat{\theta}$ che massimizza questa probabilità. Sempre per chiarire, nell'esempio degli incidenti di Berkeley (in cui abbiamo chiamato λ il parametro) la funzione da massimizzare è

$$\begin{aligned} P(X_1 = 4, X_2 = 0, \cdots, X_{10} = 3) &= P(X_1 = 4) \cdot P(X_2 = 0) \cdots P(X_{10} = 3) \\ &= \frac{\lambda^4 \mathrm{e}^{-\lambda}}{4!} \cdot \frac{\lambda^0 \mathrm{e}^{-\lambda}}{0!} \cdots \frac{\lambda^3 \mathrm{e}^{-\lambda}}{3!}. \end{aligned}$$

Quando la distribuzione di riferimento è continua, invece di massimizzare la probabilità del campione osservato (che è nulla così come quella di qualsiasi altro campione) se ne massimizza la densità.

Quest'oggetto (probabilità o densità visti come funzione del parametro incognito) prende il nome di *funzione di verosimiglianza*:

Sia $X_1, \cdots X_n$ *un campione casuale da una popolazione con distribuzione dipendente da un parametro* θ*.*
La funzione di verosimiglianza relativa al campione osservato $X_1 = x_1, \cdots X_n = x_n$ *è*

$$L(\theta) = \begin{cases} P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n|\theta) & \text{se } X_i \text{ discrete} \\ f_{X_1,\cdots X_n}(x_1, x_2, \cdots, x_n|\theta) & \text{se } X_i \text{ continue.} \end{cases}$$

La stima di massima verosimiglianza (in inglese maximum likelihood estimate) è il punto di massimo

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta).$$

Dal momento che in un campione casuale le osservazioni sono indipendenti, possiamo scrivere la verosimiglianza come

$$L(\theta) = \begin{cases} \prod_{i=1}^{n} P(X_i = x_i|\theta) & \text{se } X_i \text{ discrete} \\ \prod_{i=1}^{n} f_{X_i}(x_i|\theta) & \text{se } X_i \text{ continue} \end{cases}$$

Nota: La dipendenza della probabilità e della densità dal parametro, indicata con la notazione $|\theta$, non va interpretata come un condizionamento perché in questo caso non c'è nessun evento aleatorio condizionante. $\diamond\diamond$

Nota: La funzione di verosimiglianza è definita attraverso un prodotto (di densità o di probabilità). Dal momento che derivare dei prodotti è notoriamente fastidioso, spesso conviene lavorare con il logaritmo della funzione verosimiglianza. Attraverso il passaggio al logaritmo ci si trova a lavorare con somme invece che prodotti, che sono più comode da derivare, senza alterare il punto di massimo (perché il logaritmo è funzione monotona). $\diamond\diamond$

Soluzione 2: Risolviamo il quesito relativo agli incidenti di Berkeley.

Se la variabile aleatoria X indica il numero di incidenti in un generico giorno senza pioggia (ricordiamo che stiamo supponendo che X segua una distribuzione di Poisson con parametro incognito λ), la frazione di giornate con al massimo un incidente è

$$P(X=0)+P(X=1)=\frac{\lambda^0 \mathrm{e}^{-\lambda}}{0!}+\frac{\lambda^1 \mathrm{e}^{-\lambda}}{1!}=\mathrm{e}^{-\lambda}(1+\lambda).$$

Se il valore di λ fosse noto, allora avremmo risolto il problema: se ad esempio sapessimo che $\lambda=2$, allora la frazione di giornate con al massimo un incidente sarebbe $\mathrm{e}^{-2}(1+2)\simeq 0.40$.

Non conoscendo il valore del parametro λ, cerchiamo di stimarlo dai dati osservati e riportati nel quesito.

La funzione di verosimiglianza relativa ad un campione $x_1,\cdots,x_n$ ottenuto da una distribuzione di Poisson è

$$L(\lambda)=\prod_{i=1}^{n} P(X_i=x_i|\lambda)=\prod_{i=1}^{n}\frac{\lambda^{x_i}\mathrm{e}^{-\lambda}}{x_i!}=\frac{\lambda^{\sum_{i=1}^n x_i}\mathrm{e}^{-n\lambda}}{\prod_{i=1}^n (x_i!)}$$

Per cui

$$\begin{aligned}\log L(\lambda) &= \left(\sum_{i=1}^{n} x_i\right)\log\lambda - n\lambda - \sum_{i=1}^{n}\log(x_1!)\\ \frac{d}{d\lambda}\log L(\lambda) &= \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n\end{aligned}$$

e la stima di massima verosimiglianza è l'unica radice dell'equazione $\frac{d}{d\lambda}\log L(\lambda)=0$, quindi $\hat{\lambda}=\frac{\sum_{i=1}^n x_i}{n}=\bar{x}$.

Alla luce dei dati osservati, il valore più plausibile per λ coincide con la media del campione. Nel nostro esempio la media osservata è $\hat{\lambda}=\bar{x}=2.7$ e il corrispondente valore per la frazione di giornate senza pioggia è $\mathrm{e}^{-2.7}(1+2.7)\simeq 0.25$. $\diamond\diamond$

Osservazione: Sia nell'esempio dei pacchi di pasta (campione proveniente da una distribuzione di Bernoulli) che in quello degli incidenti di Berkeley

(in cui abbiamo ipotizzato che il campione provenga da una distribuzione di Poisson) la stima di massima verosimiglianza del parametro coincide con la media del campione osservato. Naturalmente non sempre è così, come ora vedremo nella soluzione del problema del generatore di numeri casuali. $\diamond\diamond$

Soluzione 3: Nel problema legato al generatore di numeri casuali, si vuol valutare quale sia l'estremo superiore θ dell'intervallo all'interno del quale il generatore opera. In sostanza, i dati $x_1, \cdots, x_n$ generati provengono da una distribuzione uniforme su $[0, \theta]$ e vogliamo cercare di dedurre il valore di θ (che comunque è necessariamente positivo).

Prima di iniziare notiamo che, necessariamente, le x_i osservate sono tutte non negative e il vero valore del parametro θ è maggiore o uguale a tutte le x_i.

Dal momento che la distribuzione di riferimento è continua, la funzione di verosimiglianza si ottiene attraverso la densità del campione

$$L(\theta) \;=\; f_{X_1,\cdots X_n}(x_1, x_2, \cdots, x_n|\theta) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta)$$

e, poiché la densità di ciascuna osservazione campionaria è

$$f_{X_i}(x_i|\theta) = \begin{cases} 1/\theta & \text{se } x_i \in [0, \theta] \\ 0 & \text{se } x_i \notin [0, \theta], \end{cases}$$

la verosimiglianza diventa

$$L(\theta) = \prod_{i=1}^{n} \begin{cases} 1/\theta & \text{se } x_i \in [0, \theta] \\ 0 & \text{se } x_i \notin]0, \theta] \end{cases} = \begin{cases} 1/\theta^n & \text{se tutte le } x_i \text{ sono in } [0, \theta] \\ 0 & \text{se almeno una delle } x_i \text{ non è in } [0, \theta]. \end{cases}$$

La massimizzazione della funzione di verosimiglianza $L(\theta)$ può sembrare complicata, ma diventa molto facile dopo aver fatto poche semplici considerazioni.

Indicando con $x_{\min}$ e $x_{\max}$ rispettivamente la più piccola e la più grande delle osservazioni campionarie,

$$\text{se almeno una delle } x_i \text{ non è in } [0, \theta] \quad \Longleftrightarrow \quad x_{\min} < 0 \text{ oppure } \theta < x_{\max}.$$

Siccome le x_i sono necessariamente tutte non negative, la condizione $x_{\min} < 0$ non può mai verificarsi. In conclusione, quindi

$$\text{se almeno una delle } x_i \text{ non è in } [0, \theta] \quad \Longleftrightarrow \quad \theta < x_{\max}.$$

Possiamo riscrivere la funzione di verosimiglianza come

$$L(\theta) = \begin{cases} 1/\theta^n & \text{se } \theta \geq x_{\max} \\ 0 & \text{se } \theta < x_{\max} \end{cases}$$

ed è ora immediato constatare che assume il valore massimo in $\hat{\theta} = x_{\max}$, per cui per questo problema la stima di massima verosimiglianza coincide con la massima osservazione campionaria. $\diamond\diamond$

Nota: Nella soluzione del problema del generatore di numeri casuali abbiamo prima dato per scontato che le le x_i osservate siano tutte non negative e che il vero valore del parametro θ sia maggiore o uguale a tutte le x_i. Come conseguenza di questo, abbiamo affermato che $x_{\min} < 0$ ma non abbiamo escluso la possibilità che $\theta < x_{\max}$.

Non è una dimenticanza, ma dipende dal fatto che la condizione "il vero valore del parametro θ è maggiore o uguale a tutte le x_i" riguarda solo *il vero valore* del parametro, mentre nello studio della funzione di verosimiglianza L bisogna prendere in considerazione tutti i possibili valori di θ. $\diamond\diamond$

Problema 4: (stima dei parametri di una normale) Nel pesare un oggetto, il valore fornito dalla bilancia è pari al peso reale dell'oggetto più un errore casuale. Ha spesso senso supporre che l'errore abbia distribuzione normale con valore atteso pari a 0 e varianza pari a σ^2 (varianza può, a seconda dei casi, essere nota oppure no). I risultati di n pesate successive dello stesso oggetto hanno dato i valori $x_1, x_2, \cdots, x_n$.

Stimare il peso reale dell'oggetto e, nel caso non sia nota, la varianza dell'errore. $\diamond\diamond$

Soluzione 4: Possiamo rappresentare il risultato X_i di ogni misurazione come

$$X_i = \mu + \epsilon_i$$

dove μ è il peso reale dell'oggetto e ϵ_i è l'errore casuale. Naturalmente le X_i hanno distribuzione normale di valore atteso μ e varianza σ^2. Nell'ipotesi che gli errori relativi alle diverse pesate siano indipendenti, abbiamo che $X_1, \cdots, X_n$ è un campione casuale estratto da una popolazione normale con valore atteso μ e varianza σ^2.

Considereremo tre casi diversi:

1. stima di μ supponendo σ^2 noto;
2. stima di σ^2 supponendo μ noto;
3. stima simultanea di μ e σ^2, entrambi non noti.

Nel primo caso (stima di μ supponendo σ^2 noto) la verosimiglianza relativa al campione $x_1, \cdots, x_n$ osservato è

$$\begin{aligned} L(\mu) &= f_{X_1,\cdots X_n}(x_1, x_2, \cdots, x_n|\mu) = \prod_{i=1}^{n} f_{X_i}(x_i|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}. \end{aligned}$$

Il logaritmo della funzione di verosimiglianza è

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

la cui derivata rispetto a μ è

$$\begin{aligned} \frac{d}{d\mu}\log L(\mu) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i - \mu)\cdot(-1) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) \\ &= -\frac{n}{\sigma^2}\left(\mu - \frac{\sum_{i=1}^{n} x_i}{n}\right) \end{aligned}$$

per cui il massimo si ottiene in corrispondenza della media campionaria

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}.$$

Nel secondo caso (stima di σ^2 supponendo μ noto) la verosimiglianza relativa al campione $x_1, \cdots, x_n$ osservato è

$$\begin{aligned} L(\sigma^2) &= f_{X_1,\cdots X_n}(x_1, x_2, \cdots, x_n|\mu) = \prod_{i=1}^{n} f_{X_i}(x_i|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}. \end{aligned}$$

Il logaritmo della funzione di verosimiglianza è

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

la cui derivata rispetto a σ^2 è

$$\begin{aligned} \frac{d}{d\sigma^2}\log L(\sigma^2) &= -\frac{n}{2}\cdot\frac{1}{\sigma^2} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 \cdot (-1)(\sigma^2)^{-2} \\ &= -\frac{n}{2(\sigma^2)^2}\cdot\left(\sigma^2 - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}\right) \end{aligned}$$

per cui il massimo si ottiene in

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

ATTENZIONE: non è la varianza campionaria!!!

Nel terzo caso (stima simultanea di μ e σ^2) la verosimiglianza relativa al campione $x_1, \cdots, x_n$ osservato è

$$L(\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

e il suo logaritmo

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

Per massimizzare simultaneamente rispetto alle due variabili bisogna risolvere il sistema di equazioni ottenute eguagliando a 0 le due derivate parziali

$$\begin{cases} \frac{\partial}{\partial\mu}\log L(\mu, \sigma^2) = 0 \\ \frac{\partial}{\partial\sigma^2}\log L(\mu, \sigma^2) = 0 \end{cases}$$

Le derivate parziali sono

$$\begin{aligned}\frac{\partial}{\partial\mu}\log L(\mu,\sigma^2) &= -\frac{n}{\sigma^2}(\mu-\bar{x})\\ \frac{\partial}{\partial\sigma^2}\log L(\mu,\sigma^2) &= -\frac{n}{2(\sigma^2)^2}\cdot\left(\sigma^2-\frac{\sum_{i=1}^n(x_i-\mu)^2}{n}\right)\end{aligned}$$

per cui il sistema

$$\begin{cases}-\frac{n}{\sigma^2}(\mu-\bar{x})=0\\ -\frac{n}{2(\sigma^2)^2}\cdot\left(\sigma^2-\frac{\sum_{i=1}^n(x_i-\mu)^2}{n}\right)=0\end{cases}$$

ammette come unica soluzione $(\hat{\mu}=\bar{x},\hat{\sigma}^2=\frac{\sum_{i=1}^n(x_i-\bar{x})^2}{n})$.
ATTENZIONE: anche in questo caso la stima della varianza è diversa dalla varianza campionaria!!! ◇◇

In tutti gli esempi visti, le stime di massima verosimiglianza sono funzioni dei dati campionati osservati. Dal momento che i dati osservati sono una delle possibili realizzazioni del campione casuale, possiamo pensare alla stima ottenuta come ad una possibile realizzazione di una variabile aleatoria, che prende il nome di *stimatore di massima verosimiglianza* (detto anche *stimatore MLE*).

Tanto per capirci, dai dati osservati nell'esempio degli incidenti a Berkeley abbiamo ottenuto una stima $\hat{\lambda}=\bar{x}=2.7$. Se avessimo registrato il numero di incidenti in giorni diversi, avremmo ottenuto altri dati, avremmo seguito lo stesso metodo e la stima risultante sarebbe stata ancora uguale alla media campionaria, ma il numero finale sarebbe stato diverso. Per quel problema si può quindi concludere che *lo stimatore di massima verosimiglianza per il parametro λ coincide con la media campionaria, quindi* $\hat{\lambda}=\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i$.
(NOTARE CHE ADESSO LE LETTERE SONO MAIUSCOLE).

L'espressione dello stimatore di massima verosimiglianza dipende dal modello della distribuzione dei dati. Dagli esempi visti finora possiamo concludere:

- **Modello Binomiale**: lo stimatore di massima verosimiglianza del parametro p di una binomiale è

$$\hat{p} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{(frazione di successi sul numero totale di prove)}$$

- **Modello di Poisson**: lo stimatore di massima verosimiglianza del parametro λ di una Poisson è

$$\hat{\lambda} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{(media del campione)}$$

- **Modello Uniforme su** $(0, \theta)$: lo stimatore di massima verosimiglianza dell'estremo superiore θ dell'insieme di valori ottenibili da una uniforme è

$$\hat{\theta} = X_{(n)} = \max\{X_1, \cdots, X_n\} \quad \text{(valore massimo nel campione)}$$

- **Modello Normale**: lo stimatore di massima verosimiglianza della media μ di una Normale è

$$\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{sia con varianza nota che incognita}$$

lo stimatore di massima verosimiglianza della varianza σ^2 di una Normale è

$$\hat{\sigma}^2 = \begin{cases} \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n} & \text{se la media } \mu \text{ è nota} \\ \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n} & \text{se la media } \mu \text{ è incognita} \end{cases}$$

Il metodo della massima verosimiglianza non è l'unico modo di determinare uno stimatore, anche se quello largamente più utilizzato. Quando la massimizzazione della verosimiglianza è complicata, spesso viene utilizzato il *metodo dei momenti*. Lo stimatore dei momenti è in genere soluzione (rispetto al parametro incognito) dell'equazione ottenuta imponendo l'uguaglianza tra media campionaria e valore atteso della popolazione.

In alcune situazioni lo stimatore dei momenti coincide con quello di massima verosimiglianza, ma talvolta è sostanzialmente diverso. Nel caso del modello Uniforme su $(0, \theta)$, ad esempio, abbiamo visto che lo stimatore di massima verosimiglianza è $\hat{\theta}_{MLE} = X_{(n)}$, la massima osservazione campionaria.

Lo stimatore dei momenti è soluzione dell'equazione l'equazione $\bar{X} = E(X)$. Ricordando che per questo modello $E(X) = \theta/2$, abbiamo

$$\bar{X} = \frac{\theta}{2} \qquad \text{per cui} \qquad \hat{\theta}_{MOM} = 2 \cdot \bar{X}.$$

Quando si vogliono stimare più parametri, lo stimatore dei momenti viene ottenuto risolvendo un sistema costituito da tante equazioni quanti sono i parametri da stimare. Le equazioni del sistema impongono l'uguaglianza tra i primi momenti della popolazione e le corrispondenti medie campionarie. A titolo di esempio, consideriamo il modello normale con entrambi i parametri μ e σ^2 sono incogniti. I primi due momenti della popolazione sono

$$E(X) = \mu \qquad E(X^2) = \sigma^2 + \mu^2$$

per cui il sistema di equazioni diventa

$$\begin{cases} \bar{X} = \mu \\ \frac{1}{n} \sum_{i=1}^{n} X_i^2 = \sigma^2 + \mu^2 \end{cases}$$

e ha come soluzioni $\hat{\mu}_{MOM} = \bar{X}$ e $\hat{\sigma}^2_{MOM} = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2$ (anche in questo caso coincidono con gli stimatori di massima verosimiglianza).

Esercizio: Si consideri un campione casuale di n osservazioni generato da una popolazione discreta X con funzione di massa di probabilità

$$f_X(x|\theta) = \theta(1-\theta)^{x-1} \qquad x = 1, 2, \cdots \quad \theta \in (0, 1).$$

(a) Dimostrare che $E(X) = 1/\theta$ e da questo risultato derivare l'espressione dello stimatore dei momenti;

(b) in corrispondenza di un generico campione $(x_1, \cdots, x_n)$ scrivere la funzione di verosimiglianza L;

(c) determinare la stima di massima verosimiglianza di θ e quindi scrivere l'espressione dello stimatore MLE;

(d) avendo osservato il seguente campione $x_1 = 3, x_2 = 5, , x_3 = 1, x_4 = 2, x_5 = 4$ stimare la probabilità che la v.a. X assuma valori maggiori di 3.

$\diamond\diamond$

Esercizio: Sia $(X_1, ..., X_n)$ un campione casuale generato dalla popolazione X con funzione di densità

$$f_X(x|\theta) = \frac{1}{\theta^2} x \mathrm{e}^{-\frac{x}{\theta}} \qquad x > 0 \quad \theta > 0.$$

(a) determinare lo stimatore di massima verosimiglianza $\hat{\theta}_{MLE}$;

(b) dopo aver riconosciuto che la popolazione è distribuita secondo una distribuzione Gamma di parametri $(\alpha = 2, \lambda = 1/\theta)$, determinare lo stimatore dei momenti $\hat{\theta}_{MOM}$.

$\diamond\diamond$

Sia T uno stimatore per un parametro θ (quindi T è una variabile aleatoria funzione del campione aleatorio $X_1, \cdots, X_n$).

L'errore quadratico medio di T (in inglese mean square error) è $MSE_\theta(T) = E[(T-\theta)^2]$.

Nota: L'errore quadratico medio dipende da θ. Non solo per la presenza del parametro nella differenza ma anche perché la distribuzione di T, rispetto a cui si calcola il valore atteso, dipende da θ. L'errore quadratico medio di uno stimatore $MSE_\theta(T)$ è quindi una funzione del parametro θ. $\diamond\diamond$

L'errore quadratico medio è il valore atteso del quadrato della distanza tra T e il parametro da stimare, quindi è un indicatore di quanto lo stimatore tende ad assumere valori vicini o lontani al valore incognito del parametro: valori piccoli di $MSE_\theta(T)$ indicano che T tende ad assumere valori vicini a θ. Quanto più piccolo è $MSE_\theta(T)$ tanto migliore è considerato lo stimatore.

Siano T_1 e T_2 due stimatori di uno stesso parametro θ. T_1 è più efficiente di T_2 se

$$MSE_\theta(T_1) \leq MSE_\theta(T_2) \qquad \forall\theta.$$

In un problema di stima di un parametro sarebbe auspicabile riuscire ad individuare lo stimatore più efficiente di tutti (quindi tale che il suo MSE sia minimo per ogni possibile θ), ma quest'obiettivo in genere non è realizzabile: per quanto ci si possa sforzare ad individuare un ottimo stimatore, ne esiste sempre qualcun altro che, per almeno un valore di θ ha MSE più piccolo.

Esempio: Sia $X_1, \cdots X_n$ un campione casuale generato da una distribuzione dipendente da un parametro incognito θ.

Si consideri lo stimatore $T = 3$: quale che sia il campione osservato, la stima del parametro è 3. Questo stimatore sarebbe eccellente nel caso estremamente fortunato in cui il vero valore del parametro θ fosse proprio uguale a 3, altrimenti sarebbe a dir poco discutibile. Insomma, è uno stimatore che non ha nessun senso, ci serve solo come esempio.

Dal momento che T assume sempre valore 3 indipendentemente da quali siano i valori campionari, abbiamo

$$MSE_\theta(T) = E[(T-\theta)^2] = (3-\theta)^2.$$

Per $\theta = 3$ si ha $MSE_{\theta=3}(T) = 0$. Dal momento che qualsiasi altro stimatore in $\theta = 3$ ha errore quadratico medio positivo, non può esistere uno stimatore che sia più efficiente di $T = 3$. Ma a sua volta (e per gli stessi motivi) $T = 3$ non è più efficiente dello stimatore $T' = 4$. $\diamond\diamond$

Quello che talvolta è possibile è l'individuazione di uno stimatore che risulta essere il più efficiente tra tutti gli stimatori che soddisfano una determinata proprietà, detta *non distorsione*:

> *Uno stimatore T è <u>non distorto</u> se il suo valore atteso coincide con il parametro θ da stimare.*

Quindi T è non distorto se $E(T) = \theta$ o, analogamente, se

$$B_\theta(T) = E(T) - \theta = 0$$

dove la differenza $B_\theta(T)$ viene generalmente indicata come *distorsione* oppure *bias* dello stimatore T. Anche la distorsione $B_\theta(T)$ è una funzione del parametro incognito θ (per gli stessi motivi per cui lo è l'MSE).

Riassumendo quindi, in un problema di stima di un parametro:

(a) Ci piacerebbe individuare lo stimatore più efficiente di qualsiasi altro (quindi che minimizzi MSE_θ per ogni θ), ma questo in genere non è possibile.

(b) È spesso possibile individuare quello più efficiente tra tutti gli stimatori non distorti (viene detto stimatore UMVUE).

(c) Tuttavia potrebbe esistere qualche stimatore distorto più efficiente dello stimatore UMVUE.

Abbiamo introdotto la stima di massima verosimiglianza $\hat{\theta}_{MLE}$, ottenuta come valore più plausibile alla luce del campione osservato, che per questo motivo ha un ruolo privilegiato. Dal punto di vista delle proprietà campionarie (efficienza e non distorsione) è in genere il migliore possibile quando si ha a che fare con campioni di numerosità elevata. Si può infatti dimostrare (ma non lo facciamo) che

- Lo stimatore di massima verosimiglianza è quello *asintoticamente più efficiente*, nel senso che al tendere all'infinito della numerosità campionaria il suo errore quadratico medio tende ad essere minimo per ogni possibile θ.

- Lo stimatore di massima verosimiglianza è *asintoticamente non distorto*, nel senso che al tendere all'infinito della numerosità campionaria la distorsione tende a 0.

Per campioni di numerosità moderata, lo stimatore di massima verosimiglianza è spesso distorto, ma in ogni caso abbastanza efficiente.

Esempio: Abbiamo già visto che per il modello uniforme su $(0, \theta)$ lo stimatore di massima verosimiglianza $\hat{\theta}_{MLE} = X_{(n)}$ (il massimo campionario), pur essendo distorto è più efficiente dello stimatore dei momenti $\hat{\theta}_{MOM} = 2\bar{X}$. $\diamond\diamond$

Vediamo ora che un discorso molto simile vale per la stima della varianza di un campione generato da distribuzione normale.

Esempio: Sia $X_1, \cdots, X_n$ un campione casuale generato da una distribuzione normale di valore atteso μ e varianza σ^2.

Vogliamo valutare le proprietà campionarie (efficienza e distorsione) di due possibili stimatori della varianza:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

(il primo, S^2, è la varianza campionaria mentre il secondo, $\hat{\sigma}^2$, lo stimatore di massima verosimiglianza).

Qualche lezione fa abbiamo argomentato e in parte dimostrato che se il campione è generato da una distribuzione $N(\mu, \sigma^2)$ allora

$$C = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}$$

ha una distribuzione χ^2_{n-1} (chi quadrato con $n-1$ gdl).

Dal momento che valore atteso e varianza di una distribuzione chi quadro sono pari rispettivamente al numero di gdl e al doppio dei gdl, vediamo immediatamente che $E(C) = n-1$ e $V(C) = 2(n-1)$.

A questo punto esprimiamo $S^2 = \frac{\sigma^2}{n-1}C$ e $\hat{\sigma}^2 = \frac{\sigma^2}{n}C$ e otteniamo

$$\begin{aligned} E(S^2) &= \frac{\sigma^2}{n-1}E(C) = \sigma^2 \\ E(\hat{\sigma}^2) &= \frac{\sigma^2}{n}E(C) = \frac{n-1}{n}\sigma^2 \end{aligned}$$

da cui emerge che la varianza campionaria è non distorta mentre lo stimatore di massima verosimiglianza tende a sottostimare la varianza (la distorsione è $B(\hat{\sigma}^2) = -\sigma^2/n$).

Il risultato non ci sorprende, avevamo infatti detto che la varianza campionaria con $n-1$ al denominatore è stata "inventata" proprio per far sì che il suo valore atteso coincida con la varianza della popolazione.

Per il calcolo dell'errore quadratico medio dei due stimatori sfruttiamo ancora la loro relazione con C:

$$\begin{aligned} MSE(S^2) &= V(S^2) = \left(\frac{\sigma^2}{n-1}\right)^2 V(C) = \frac{2}{n-1}\sigma^4 \\ MSE(\hat{\sigma}^2) &= V(\hat{\sigma}^2) + B(\hat{\sigma}^2) = \left(\frac{\sigma^2}{n}\right)^2 V(C) + \frac{1}{n^2}\sigma^4 \\ &= \frac{2(n-1)}{n^2}\sigma^4 + \frac{1}{n^2}\sigma^4 = \frac{2n-1}{n^2}\sigma^4 \end{aligned}$$

Confrontando i due MSE otteniamo

$$MSE(S^2) - MSE(\hat{\sigma}^2) = \frac{2}{n-1}\sigma^4 - \frac{2n-1}{n^2}\sigma^4 = \frac{3n-1}{n^2(n-1)}\sigma^4$$

da cui emerge che S^2 ha un errore quadratico medio più alto ed è quindi meno efficiente. $\diamond\diamond$