Data Science per psicologi

Corrado Caudek

2021-09-13

Indice

Verosimiglianza

Capitolo 1

La funzione di verosimiglianza

Per introdurre la funzione di verosimiglianza utilizzeremo un esempio proposto da McElreath (2020).[1] Supponiamo di tenere in mano un mappamondo e di chiederci: "qual'è la proporzione della superficie terreste ricoperta d'acqua?" Sembra una domanda a cui è difficile rispondere. Ma McElreath (2020) propone questa idea brillante: lanciamo in aria il mappamondo e, quando lo riprendiamo, osserviamo se la superfice del mappamondo sotto il nostro dito indice destro rappresenta acqua o terra. Possiamo ripetere questa procedura più volte, così da ottenere un campione causale di diverse porzioni della superficie dal mappamondo. Eseguiamo l'esperimento casuale nove volte e osserviamo i seguenti risultati: A, T, A, A, A, T, A, T, A, dove "A" indica acqua e "T" indica terra.

1.1 Dati e parametri

Per l'esempio del mappamondo possiamo dire quanto segue:

1. la proporzione del pianeta Terra ricoperta d'acqua è θ;
2. un singolo lancio del mappamondo ha una probabilità θ di produrre l'osservazione "acqua" (A);
3. i lanci del mappamondo sono indipendenti (nel senso che il risultato di un lancio non influenza i risultati degli altri lanci).

Nella descrizione precedente, i *dati* sono le frequenze degli eventi A ("acqua") e T ("terra") prodotte da una serie di osservazioni. La somma delle frequenze di A e T è il numero totale dei lanci del mappamondo: $N = A + T$. Oltre ai dati, abbiamo anche fatto riferimento alla proporzione (incognita) di acqua sul globo terrestre. Tale proporzione corrisponde al *parametro* $\theta \in \Theta$, dove Θ rappresenta lo spazio parametrico in cui può variare il parametro θ.

[1] Per una trattazione più formale, si consulti il tutorial di Etz (2018).

Nel caso presente, $\theta \in [0, 1]$, in quanto θ è una proporzione. La descrizione del parametro θ rappresenta l'obiettivo dell'inferenza.

Anche se il parametro θ non può essere direttamente osservato è possibile inferire il suo valore a partire dai dati. Avendo specificato ciò che abbiamo detto sopra, possiamo formulare un *modello statistico*: abbiamo una sequenza di prove Bernoulliane indipendenti e, dunque, il modello statistico sarà quello binomiale. Prima di descrivere questo modello in dettaglio, è utile visualizzare il suo comportamento. Dopo aver visto come questo modello apprende dai dati ci porremo il problema di capire come funziona.

1.1.1 Come impara un modello statistico?

Prima di lanciare in aria il mappamondo e di ottenere il primo dato, non sappiamo nulla del parametro θ. Dato che θ è una proporzione, i suoi valori possibili vanno da 0 a 1. Se non possediamo alcuna informazione su θ, allora riteniamo che tutti i valori θ siano egualmente credibili. Rappresentiamo dunque la nostra incertezza a proposito del parametro θ mediante una distribuzione uniforme su tutti i valori θ, come indicato dalla linea tratteggiata nel pannello $n = 1$ della figura 1.1.

Lanciamo in aria il mappamondo una prima volta e, quando lo riprendiamo, notiamo che sotto il nostro indice destro c'è "acqua". Dopo avere osservato il risultato del primo lancio, ovvero "A", il modello aggiorna le credibilità dei valori del parametro θ che ora sono rappresentate dalla linea continua nel pannello $n = 1$ della figura 1.1. La credibilità associata all'evento $\theta = 0$ è scesa esattamente a zero, l'equivalente di "impossibile". Infatti, avendo osservato almeno un luogo sul mappamondo in cui c'è dell'acqua, possiamo dire che l'evento "non c'è acqua" (ovvero $\theta = 0$) è impossibile. Allo stesso modo, la credibilità di $\theta > 0.5$ è aumentata. Non abbiamo ancora evidenze che ci sia terra sul mappamondo, quindi le credibilità iniziali sono state modificate per essere coerenti con questa informazione: le credibilità associate a θ aumentano passando dal valore $\theta = 0$ a valore $\theta = 1$, in maniera coerente con i dati che abbiamo. Il punto importante è che le evidenze disponibili fino a questo momento vengono incorporata nelle credibilità attribuite a ciascun possibile valore θ. Il modello implementa questa logica in maniera *automatica*. Non è necessario fornire al modello alcuna istruzione per ottenere questo risultato. La teoria della probabilità svolge tutti i calcoli necessari per noi.

Lanciamo in aria il mappamondo una seconda volta e osserviamo "T". Consideriamo dunque il pannello $n = 2$ della figura 1.1. La linea tratteggiata in questo pannello ricopia semplicemente la descrizione del livello di credibilità di ciascun valore θ che era disponibile nel caso di un solo lancio del mappamondo. La linea continua, invece, aggiorna tali valori di credibilità incorporando l'informazione secondo la quale in due lanci abbiamo ottenuto "acqua" una volta e "terra" una volta. Vediamo che ora il valore di credibilità di θ è uguale a zero per l'evento $\theta = 0$; infatti, abbiamo osservato "acqua" nel primo lancio. In maniera corrispondente, il valore di credibilità di θ è uguale a zero per l'evento

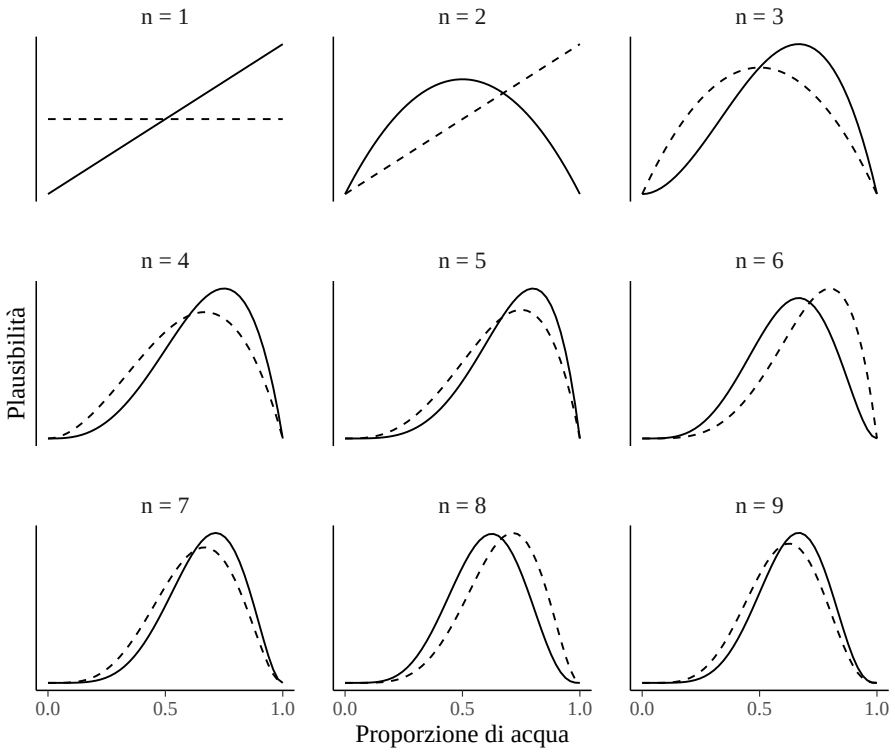


Figura 1.1: *Come un modello statistico impara? Ciascun lancio del mappamondo produce un'osservazione: acqua (A) o terra (T). La stima della proporzione di acqua (θ) sulla superficie terreste prodotta dal modello è espressa nei termini del grado di credibilità (o Plausibilità, nella figura) di ciascun possibile valore θ. Le linee e le curve nella figura rappresentano il grado di credibilità dei valori θ per diversi set di dati. In ogni pannello, le credibilità (curva tratteggiata) calcolate in base alle informazioni fornite dai lanci $1, \ldots, k$ vengono aggiornate a seguito dell'informazione fornita dal lancio $k + 1$. I nuovi valori di credibilità di θ sono rappresentati dalla curva solida.*

$\theta = 1$ (c'è solo acqua); infatti, abbiamo osservato "terra" nel secondo lancio. Avendo osservato "acqua" nel 50% dei casi, il valore più verosimile per θ sarà 0.5, come indicato dalla linea continua in questo pannello.

Nei pannelli rimanenti della figura 1.1 i nuovi dati prodotti dai successivi lanci del mappamondo vengono analizzati dal modello, uno alla volta. La curva tratteggiata in ciascun pannello corrisponde alla curva solida del pannello precedente, spostandosi da sinistra a destra e dall'alto verso il basso. Ogni volta che si ottiene un dato A il picco della curva di credibilità si sposta a destra, verso valori più grandi di θ. Ogni volta si ottiene T ci si sposta nella direzione opposta. L'altezza massima della curva aumenta con ogni campione, il che significa che, all'aumentare della quantità di prove, viene associato un livello di credibilità maggiore ad un minor numero di valori di θ. Man mano che viene aggiunta una nuova osservazione, la curva che rappresenta la credibilità dei valori θ viene aggiornata in maniera coerente con tutte le osservazioni precedenti.

1.2 La funzione di verosimiglianza

Nella figura 1.1 abbiamo descritto con una curva il grado di credibilità associato a ciascun valore del parametro θ. Una curva è il grafico di una funzione matematica. In statistica, tale funzione si chiama *verosimiglianza* di θ basata sui dati $y = (y_1, ..., y_n)$.

1.2.1 La verosimiglianza del modello binomiale

Vediamo ora come si costruisce la funzione di verosimiglianza per l'esperimento casuale corrispondente al lancio del mappamondo. Gli eventi possibili che possono essere osservati nell'esperimento casuale sono acqua (A) e terra (T). Avendo lanciato in aria il mappamondo 9 volte, abbiamo osservato una serie di eventi A e T. Ora ci chiediamo: qual è la probabilità di osservare questo *specifico* campione (6 volte "acqua" in 9 lanci del mappamondo) nell'universo di tutte le possibili sequenze risultanti da 9 lanci del mappamondo? Sembra una domanda a cui è molto difficile rispondere, ma in realtà non è vero. Se specifichiamo le caratteristiche dell'esperimento casuale come abbiamo fatto sopra, ovvero: (1) ogni lancio è indipendente dagli altri e (2) la probabilità di osservare "acqua" è la stessa in ciascun lancio, allora la teoria della probabilità ci consente di trovare facilmente una risposta alla nostra domanda. Le caratteristiche dell'esperimento casuale che abbiamo descritto specificano infatti una variabile casuale binomiale. La funzione che stiamo cercando, dunque, è la distribuzione binomiale. In precedenza abbiamo discusso tale distribuzione facendo riferimento al lancio di una moneta. Ma l'esperimento casuale corrispondente a n lanci di una moneta è strutturalmente identico a quello del mappamondo (gli unici esiti possibili sono "acqua" e "terra", i lanci sono indipendenti gli uni dagli altri e la probabilità di osservare "acqua" rimane costante in ciascun lancio). Possiamo dunque usare la distribuzione binomiale per descrivere la probabilità di osservare A = "numero di volte in cui abbiamo osservato acqua" e T = "nu-

mero di volte in cui abbiamo osservato terra", quando il nostro mappamondo è stato lanciato in aria per $n = A + T$ volte. Tale probabilità è data dalla distribuzione binomiale di parametro θ:

$$P(A, T \mid \theta) = \frac{(A+T)!}{A!T!}\theta^A + (1-\theta)^T. \tag{1.1}$$

In altre parole, la frequenza degli eventi "numero di volte in cui abbiamo osservato acqua" e "numero di volte in cui abbiamo osservato terra" segue la distribuzione binomiale nella quale la probabilità di osservare "acqua" in ciascun lancio è uguale a θ.

1.2.2 La verosimiglianza vista da vicino

Ma cosa dobbiamo fare, in pratica, per generare le funzioni di verosimiglianza che sono rappresentate nei diversi pannelli della figura 1.1? Iniziamo con una definizione formale.

Definizione 1.1. La *funzione di verosimiglianza* $\mathcal{L}(\theta \mid y) = f(y \mid \theta), \theta \in \Theta$, è la funzione di massa o di densità di probabilità dei dati y vista come una funzione del parametro sconosciuto θ.

Spesso per indicare la verosimiglianza si scrive $\mathcal{L}(\theta)$ se è chiaro a quali valori y ci si riferisce. La verosimiglianza $\mathcal{L}$ è una curva (in generale, una superficie) nello spazio Θ del parametro (in generale, dei parametri θ) che riflette la credibilità relativa dei valori θ alla luce dei dati osservati. Notiamo un punto importante: la funzione $\mathcal{L}(\theta \mid y)$ non è una funzione di densità. Infatti, essa non racchiude un'area unitaria.

Nel caso presente, la funzione di verosimiglianza è descritta dalla (1.1), ovvero, corrisponde alla funzione binomiale di parametro $\theta \in (0, 1)$ sconosciuto. Nell'esempio che stiamo discutendo, abbiamo osservato "acqua" sei volte in nove lanci del mappamondo. Dunque, $y = 6$ successi in $n = 9$ prove. Per i dati del campione considerato, la funzione di verosimiglianza diventa

$$\mathcal{L}(\theta \mid y) = \frac{(6+3)!}{6!3!}\theta^6 + (1-\theta)^3. \tag{1.2}$$

La definizione precedente ci dice che, per costruire la funzione di verosimiglianza, dobbiamo applicare tante volte la (1.2) *tenendo costanti i dati* e cambiando ogni volta il valore θ.

Per esempio, se poniamo $\theta = 0.1$

$$\mathcal{L}(\theta \mid y) = \frac{(6+3)!}{6!3!}0.1^6 + (1-0.1)^3$$

otteniamo il valore 0.0446. Se poniamo $\theta = 0.2$

$$\mathcal{L}(\theta \mid y) = \frac{(6+3)!}{6!3!}0.2^6 + (1-0.2)^3$$

otteniamo 0.1762; e così via.

La tabella seguente riportata alcuni valori rappresentativi della funzione di verosimiglianza definita da 6 successi in 9 prove Bernoulliane.

θ	$\mathcal{L}(\theta \mid y)$
0.0	0.0000
0.1	0.0001
0.2	0.0028
0.3	0.0210
0.4	0.0743
0.5	0.1641
0.6	0.2508
0.7	0.2668
0.8	0.1762
0.9	0.0446
1.0	0.0000

La figura 1.2 fornisce una rappresentazione grafica della funzione di verosimiglianza – la figura è stata costruita utilizzando 100 valori equispaziati $\theta \in [0, 1]$.

```
n <- 9
y <- 6
theta <- seq(0, 1, length.out=100)
like <- choose(n, y) * theta^y * (1 - theta)^(n - y)
plot(
  theta, like,
  type = 'l',
  xaxt = "n",
  bty = 'l',
  main = "Funzione di verosimiglianza",
  ylab = expression(L(theta)),
  xlab = expression('Valori possibili di' ~ theta)
)
axis(side = 1, at = seq(0, 1, length.out = 11))
segments(
  0.67, 0, 0.67,
  choose(n, y) * 0.67^y * (1 - 0.67)^(n - y),
  lty = 2
)
```

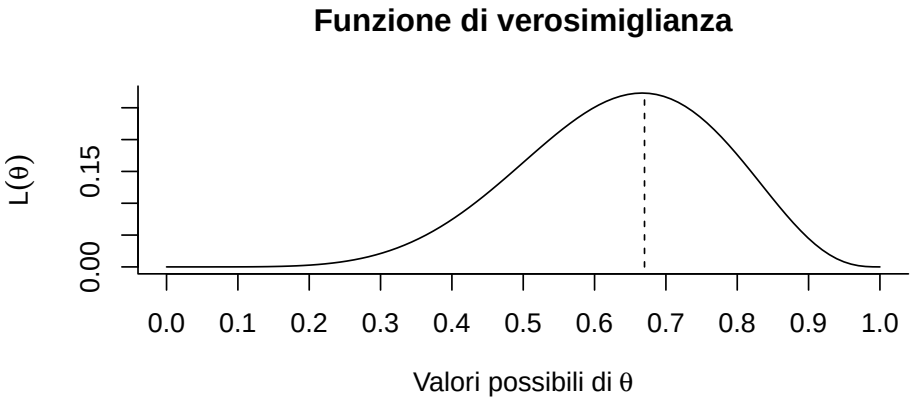


Figura 1.2: *Funzione di verosimiglianza nel caso in cui l'esito acqua sia stato osservato 6 volte in 9 lanci del mappamondo.*

1.2.2.1 Interpretazione

Come possiamo interpretare la curva che abbiamo ottenuto? Per alcuni valori θ la funzione di verosimiglianza assume valori piccoli; per altri valori θ la funzione di verosimiglianza assume valori più grandi. Questi ultimi sono i valori di θ "più credibili" e il valore 0.67 è il valore più credibile di tutti. In termini più formali possiamo dire che la funzione di verosimiglianza ha la seguente interpretazione: sulla base dei dati, $\theta_1 \in \Theta$ è più credibile di $\theta_2 \in \Theta$ come indice del modello probabilistico generatore delle osservazioni se $\mathcal{L}(\theta_1) > \mathcal{L}(\theta_1)$.

In conclusione, la funzione di verosimiglianza descrive in termini relativi il sostegno empirico che $\theta \in \Theta$ riceve da y. La figura 1.1, infatti, mostra come la funzione di verosimiglianza assume una forma diversa quando y varia: le curve nei pannelli della figura 1.1 sono tutte state ottenute usando la (1.1), ma sono tra loro diverse perché i dati sono diversi: 1 successo in 1 prova (abbiamo lanciato il mappamondo una volta e abbiamo osservato "acqua"); 1 successo in 2 prove (abbiamo lanciato il mappamondo due volte e abbiamo osservato "acqua" e "terra"); 2 successi in 3 prove (abbiamo lanciato il mappamondo tre volte e abbiamo osservato "acqua", "terra" e "acqua"); eccetera.

1.2.3 La stima di massima verosimiglianza

La funzione di verosimiglianza rappresenta la "credibilità relativa" dei valori del parametro di interesse. Ma qual è il valore più credibile? Se utilizziamo soltanto la funzione di verosimiglianza, allora la risposta è data dalla stima di massima verosimiglinza.

Definizione 1.2. Un valore di θ che massimizza $\mathcal{L}(\theta \mid y)$ sullo spazio parametrico Θ è detto *stima di massima verosimiglinza* (s.m.v.) di θ ed è indicato con $\hat{\theta}$:

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} \mathcal{L}(\theta). \tag{1.3}$$

Il paradigma frequentista utilizza la funzione di verosimiglianza quale unico strumento per giungere alla stima del valore più credibile del parametro sconosciuto θ. Tale stima corrisponde al punto di massimo della funzione di verosimiglianza. Nell'esempio presente, $\hat{\theta} = 0.6667$. Il massimo della funzione di verosimiglianza, ovvero $\hat{\theta}$, si può ottenere con metodi numerici o grafici.

In base all'approccio bayesiano, invece, il valore più credibile del parametro sconosciuto θ non corrisponde alla s.m.v.. Per l'approccio bayesiano, invece, il valore più credibile del parametro sconosciuto θ è dato dalla moda (o media, o mediana) della distribuzione a posteriori $p(\theta \mid y)$ che si ottiene combinando la verosimiglianza $p(y \mid \theta)$ con la distribuzione a priori $p(\theta)$.

1.2.4 La log-verosimiglianza

Dal punto di vista pratico risulta più conveniente utilizzare, al posto della funzione di verosimiglianza, il suo logaritmo naturale, ovvero la funzione di log-verosimiglianza

$$\ell(\theta) = \log \mathcal{L}(\theta).$$

Poiché il logaritmo è una funzione strettamente crescente (usualmente si considera il logaritmo naturale), allora $\mathcal{L}(\theta)$ e $\ell(\theta)$ assumono il massimo (o i punti di massimo) in corrispondenza degli stessi valori di θ:

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} \ell(\theta).$$

Per le proprietà del logaritmo, si ha

$$\ell(\theta) = \log \left(\prod_{i=1}^{n} f(y \mid \theta) \right) = \sum_{i=1}^{n} \log f(y \mid \theta). \tag{1.4}$$

Si noti che non è necessario lavorare con i logaritmi, anche se è fortemente consigliato, e questo perché i valori della verosimiglianza, in cui si moltiplicano valori di probabilità molto piccoli, possono diventare estremamente piccoli (qualcosa come 10^{-34}). In tali circostanze, non è sorprendente che i programmi dei computer mostrino problemi di arrotondamento numerico. Le trasformazioni logaritmiche risolvono questo problema.

1.2.5 Derivazione della s.m.v. per una proporzione

Nel Paragrafo 1.2.5 abbiamo trovato che la s.m.v. di θ è uguale alla proporzione di successi campionari. Questo risultato può essere dimostrato come segue.

Dimostrazione. Per n prove Bernoulliane indipendenti, le quali producono y successi e $(n-y)$ insuccessi, la funzione nucleo (ovvero, la funzione di verosimiglianza da cui sono state escluse tutte le costanti moltiplicative che non hanno alcun effetto su $\hat{\theta}$) è

$$\mathcal{L}(p \mid y) = \theta^y (1-\theta)^{n-y}.$$

La funzione nucleo di log-verosimiglianza è

$$\begin{aligned}\ell(\theta \mid y) &= \log \mathcal{L}(\theta \mid y) \\ &= \log\left(\theta^y (1-\theta)^{n-y}\right) \\ &= \log \theta^y + \log\left((1-\theta)^{n-y}\right) \\ &= y \log \theta + (n-y)\log(1-\theta).\end{aligned}$$

Per calcolare il massimo della funzione di log-verosimiglianza è necessario differenziare $\ell(\theta \mid y)$ rispetto a θ, porre la derivata a zero e risolvere. La derivata di $\ell(\theta \mid y)$ è:

$$\ell'(\theta \mid y) = \frac{y}{\theta} - \frac{n-y}{1-\theta}.$$

Ponendo l'equazione uguale a zero e risolvendo otteniamo la s.m.v.:

$$\hat{\theta} = \frac{y}{n}, \tag{1.5}$$

ovvero la frequenza relativa dei successi nel campione. □

1.2.5.1 Calcolo numerico

In maniera più semplice, il risultato descritto nel Paragrafo 1.2.5 può essere ottenuto mediante una simulazione in R. Iniziamo a definire un insieme di valori possibili per il parametro incognito θ:

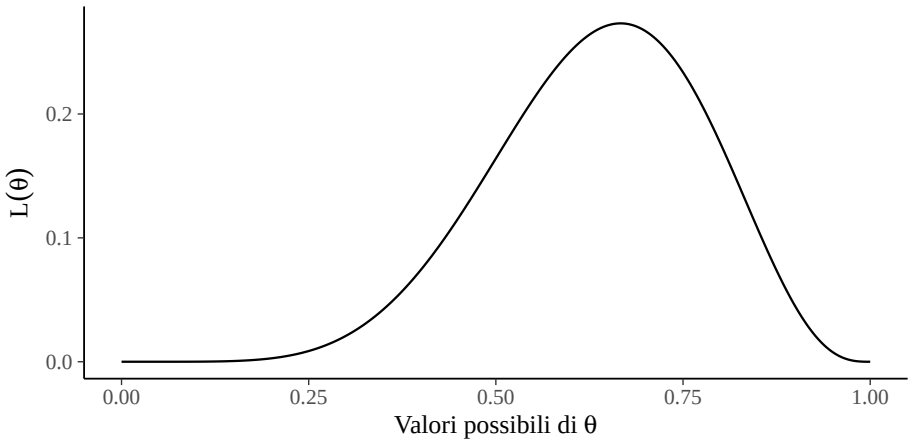
```
theta <- seq(0, 1, length.out=1e3)
```

Sappiamo che la funzione di verosimiglianza è la funzione di massa di probabilità espressa in funzione del parametro sconosciuto θ assumendo come noti i dati. Questo si può esprimere in R nel modo seguente:

```
like <- dbinom(x = 6, size = 9, prob = theta)
```

Si noti che, nell'istruzione precedente, abbiamo passato alla funzione `dbinom()` i dati, ovvero `x = 6` successi in `size = 9` prove. Inoltre, abbiamo passato alla funzione il vettore `prob = theta` che contiene 1000 valori possibili per il parametro $\theta \in [0, 1]$. Per ciascuno dei valori θ, la funzione `dbinom()` ritorna un valore che corrisopnde all'ordinata della funzione di verosimiglianza, tenendo sempre costanti i dati (ovvero, 6 successi in 9 prove). Un grafico della funzione di verosimiglianza è dato da:

```
tibble(theta, like) %>%
  ggplot(aes(x = theta, y = like)) +
  geom_line() +
  labs(
    y = expression(L(theta)),
    x = expression('Valori possibili di' ~ theta)
  )
```



Nella simulazione, il valore θ che massimizza la funzione di verosimiglianza può essere trovato nel modo seguente:

```
theta[which.max(like)]
#> [1] 0.6666667
```

Il valore così trovato è uguale al valore definito dalla (1.5).

1.3 La verosimiglianza del modello Normale

Ora che abbiamo capito come costruire la funzione verosimiglianza di una binomiale è relativamente semplice fare un passo ulteriore e considerare la verosimiglianza del caso di una funzione di densità, ovvero nel caso di una variabile casuale continua. Consideriamo qui il caso della Normale. La densità di una distribuzione Normale di parametri μ e σ è

$$f(y \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}.$$

Nel caso in cui le n osservazioni $y = (y_1, \dots, y_n)$ siano realizzazioni indipendenti ed identicamente distribuite (di seguito, i.i.d.) della medesima variabile casuale

Y con densità Normale di parametri μ e σ, poniamoci il problema di trovare la s.m.v. dei parametri sconosciuti μ e σ. Per semplicità, scriviamo $\theta = \{\mu, \sigma\}$.

In precedenza abbiamo utilizzato la nozione di probabilità congiunta per fare riferimento alla probabilità del verificarsi di un insieme di eventi. Estendiamo questo ragionamento al caso presente. Consideriamo il campione osservato come un insieme di eventi. Ciascuno di tali eventi è la realizzazione di una variabile casuale — possiamo pensarla come l'estrazione casuale di un valore dalla "popolazione" $\mathcal{N}(\mu, \sigma)$. Se tali variabili casuali sono i.i.d. la loro densità congiunta è data da:

$$\begin{aligned} f(y \mid \theta) &= f(y_1 \mid \theta) \cdot f(y_2 \mid \theta) \cdot \ \ldots \ \cdot f(y_n \mid \theta) \\ &= \prod_{i=1}^{n} f(y_i \mid \theta), \end{aligned}$$

laddove la funzione $f(\cdot)$ è data dalla (1.3). La funzione di verosimiglianza è dunque:

$$\mathcal{L}(\theta \mid y) = \prod_{i=1}^{n} f(y_i \mid \theta). \tag{1.6}$$

L'obiettivo è quello di massimizzare la funzione di verosimiglianza per trovare i valori θ ottimali. Usando la notazione matematica questo si esprime dicendo che cerchiamo l'argmax della (1.6) rispetto a θ, ovvero

$$\hat{\theta} = \text{argmax}_\theta \prod_{i=1}^{n} f(y_i \mid \theta).$$

Questo problema si risolve calcolando le derivate della funzione rispetto a θ, ponendo le derivate uguali a zero e risolvendo. Saltando tutti i passaggi algebrici di questo procedimento, per μ abbiamo

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{1.7}$$

e per σ abbiamo

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{n} \frac{1}{n} (y_i - \mu)^2}. \tag{1.8}$$

In altri termini, la s.m.v. del parametro μ è la media del campione e la s.m.v. del parametro σ è la deviazione standard del campione.

1.3.1 Simulazione

Consideriamo ora un esempio relativo al campione di valori BDI-II dei trenta soggetti del campione clinico di Zetsche et al. (2019), ovvero

```
d <- tibble(
  y = c(26, 35, 30, 25, 44, 30, 33, 43, 22, 43, 24,
        19, 39, 31, 25, 28, 35, 30, 26, 31, 41, 36,
        26, 35, 33, 28, 27, 34, 27, 22)
  )
```

Ci poniamo lo scopo di generare la funzione di verosimiglianza per questi dati. Supponiamo che ricerche precedenti ci dicano che il BDI-II si distribuisce secondo una legge Normale.

Ci concentreremo qui sul parametro μ della distribuzione Normale: per semplificare il problema assumiamo di conoscere σ (lo porremo uguale alla deviazione standard del campione) in modo da avere un solo parametro sconosciuto. Il nostro problema è dunque quello di trovare la funzione di verosimiglianza per il parametro μ, date le 30 osservazioni che abbiamo a disposizione e dato $\sigma = s$.

Abbiamo visto sopra che, per una singola osservazione, la funzione di verosimiglianza è la densità Normale espressa in funzione dei parametri. Per un *campione* di osservazioni $y = (y_1, y_2, \dots, y_n)$ dobbiamo utilizzare la funzione di densità congiunta $f(y \mid \mu, \sigma)$ espressa in funzione dei parametri, ovvero $\mathcal{L}(\mu, \sigma \mid y)$. Se le osservazioni sono i.i.d., la densità congiunta è data dal prodotto delle densità delle singole osservazioni. Per una singola osservazione y_i abbiamo

$$f(y_i \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y_i - \mu)^2}{2\sigma^2}\right\},$$

dove il pedice i specifica l'osservazione y_i tra le molteplici osservazioni y, e μ e σ sono i parametri sconosciuti che devono essere determinati. La densità congiunta è dunque

$$f(y \mid \mu, \sigma) = \prod_{i=1}^{n} f(y_i \mid \mu, \sigma)$$

e, alla luce dei dati osservati, la verosimiglianza diventa

$$\begin{aligned}\mathcal{L}(\mu, \sigma \mid y) = \prod_{i=1}^{n} f(y_i \mid \mu, \sigma) = \\ &\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(26 - \mu)^2}{2\sigma^2}\right\} \times \\ &\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(35 - \mu)^2}{2\sigma^2}\right\} \times \\ &\vdots \\ &\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(22 - \mu)^2}{2\sigma^2}\right\}.\end{aligned}$$

Poniamoci ora il problema di rappresentare graficamente la funzione di verosimiglianza per il parametro μ — per semplicità, assumiamo σ noto e uguale alla deviazione standard del campione.

```
true_sigma <- sd(d$y)
```

Avendo un solo parametro sconosciuto da stimare possiamo rappresentare la verosimiglianza con una curva, anziché con una superficie. In R, possiamo definire la funzione di log-verosimiglianza nel modo seguente:

```
log_likelihood <- function(y, mu, sigma=true_sigma) {
  sum(dnorm(y, mu, sigma, log=TRUE))
}
```

Si noti che, nella funzione `log_likelihood`, `y` è un vettore che, nel caso presente conterrà $n = 30$ valori. Per ciascuno di questi valori, la funzione `dnorm()` troverà la densità Normale (l'ordinata della funzione) utilizzando il valore μ che viene passato a `log_likelihood` e un valore σ sempre uguale, dato che, nell'esempio, questo parametro verrà mantenuto costante. L'argomento `log = TRUE` specifica che deve essere preso il logaritmo. La funzione `dnorm()` è un argomento della funzione `sum()`. Ciò significa che i 30 valori così trovati, espressi su scala logaritmica, verranno sommati — sommare logaritmi è equivalente a fare il prodotto dei valori sulla scala originaria.

Se applichiamo questa funzione ad un solo valore μ otteniamo un singolo valore della funzione di log-verosimiglianza (ovvero, l'ordinata di un singolo punto della funzione rappresentata nella figura (1.3.1)). Tale singolo valore viene trovato utilizzando tutti i 30 dati del campione, il valore $\sigma = s$ che viene tenuto fisso e il singolo valore μ che abbiamo passato alla funzione `log_likelihood()`. Dobbiamo, tuttavia, applicare la funzione a tutti i possibili valori che μ può assumere. Per cui il procedimento che abbiamo descritto sopra per un singolo valore μ viene ripetuto moltissime volte.

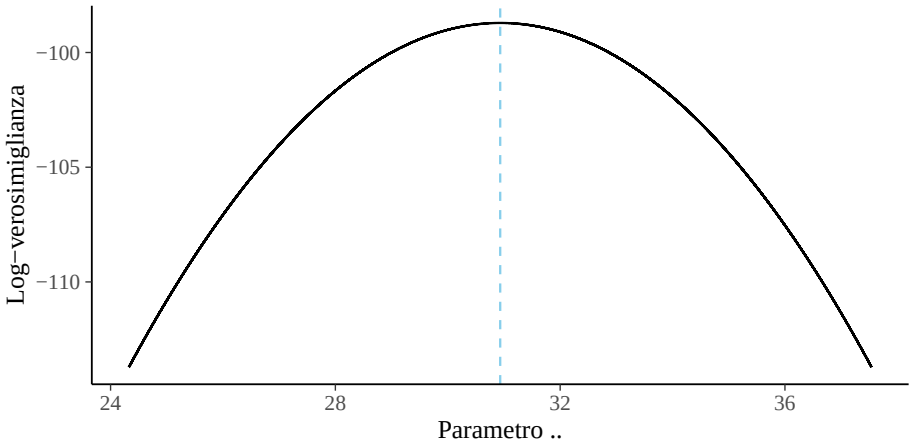
Nel seguente ciclo `for()` usato nelle istruzioni seguenti viene calcolata la log-verosimiglianza di 100,000 possibili valori del parametro μ:

```
nrep <- 1e5
mu <- seq(
  mean(d$y) - sd(d$y),
  mean(d$y) + sd(d$y),
  length.out = nrep
)

ll <- rep(NA, nrep)
for (i in 1:nrep) {
  ll[i] <- log_likelihood(d$y, mu[i], true_sigma)
}
```

Il vettore `mu` contiene 100,000 possibili valori del parametro μ. Tali valori sono stati scelti in modo tale da essere compresi nell'intervallo $\bar{y} \pm s$. Per ciascuno dei possibili valori del parametro μ la funzione `log_likelihood()` calcola la log-verosimiglianza seguendo la procedura descritta sopra. All'interno del ciclo `for()` i 100,000 risultati così ottenuti vengono salvati nel vettore `ll`. Possiamo ora utilizzare i valori contenuti nei vettori `mu` e `ll` per disegnare il grafico della funzione di log-verosimiglianza per il parametro μ:

```
tibble(mu, ll) %>%
ggplot(aes(x = mu, y = ll)) +
  geom_line() +
  vline_at(mean(d$y), color = "sky blue", linetype = "dashed") +
  labs(
    y = "Log-verosimiglianza",
    x = c("Parametro \u03BC")
  )
```



Dalla figura notiamo che, per questi dati, il massimo della funzione di log-verosimiglianza calcolata per via numerica è uguale a 30.93. Tale valore è identico alla media dei dati campionari e corrisponde al risultato teorico della (1.3.1).

Considerazioni conclusive

La verosimiglianza viene utilizzata sia nell'inferenza bayesiana che in quella frequentista. In entrambi i paradigmi di inferenza, il suo ruolo è quantificare la forza con la quale i dati osservati supportano i possibili valori dei parametri sconosciuti.

Nella funzione di verosimiglianza i dati (osservati) vengono trattati come fissi, mentre i valori del parametro (o dei parametri) θ vengono variati: la verosimiglianza è una funzione di θ per il dato fisso y. Pertanto, la funzione di verosimiglianza riassume i seguenti elementi: un modello statistico che genera stocasticamente i dati (in questo capitolo abbiamo esaminato due modelli statistici: quello binomiale e quello Normale), un intervallo di valori possibili per θ e i dati osservati y.

Nella statistica frequentista l'inferenza si basa solo sui dati a disposizione e qualunque informazione fornita dalle conoscenze precedenti non viene presa in considerazione. Nello specifico, nella statistica frequentista l'inferenza viene condotta massimizzando la funzione di (log) verosimiglianza, condizionatamente ai valori assunti dalle variabili casuali campionarie. Nella statistica bayesiana, invece, l'inferenza statistica viene condotta combinando la funzione di verosimiglianza con le distribuzioni a priori dei parametri incogniti θ.

La differenza fondamentale tra inferenza bayesiana e frequentista è dunque che i frequentisti non ritengono utile descrivere in termini probabilistici i parametri: i parametri dei modelli statistici vengono concepiti come fissi ma sconosciuti. Nell'inferenza bayesiana, invece, i parametri sconosciuti sono intesi come delle variabili casuali e ciò consente di quantificare in termini probabilistici il nostro grado di intertezza relativamente al loro valore.

Bibliografia

Etz, A. (2018). Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*, 1(1):60–69.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC Press, Boca Raton, Florida, 2nd edition edition.

Zetsche, U., Bürkner, P.-C., and Renneberg, B. (2019). Future expectations in clinical depression: Biased or realistic? *Journal of Abnormal Psychology*, 128(7):678–688.