Verosimiglianza: funzione e dintorni...

Domenico Vistocco

Corso di Inferenza Statistica

La carte

- Notazione essenziale
- La funzione di probabilità congiunta (duplice utilizzo)
- La funzione di verosimiglianza per il modello di Bernoulli
- MLE, funzione Score e informazione osservata di Fisher
- Verosimiglianza per il modello normale
- Approssimazione del secondo ordine per la funzione di verosimiglianza
- Alcune (prime) proprietà della funzione di verosimiglianza

Lessico di base

Notazione essenziale

Simbolo	Legenda
Y	variabile (casuale) - *v.c.*
y	valore della v.c.
$\mathbf{y}$	un insieme di valori della *v.c.*
$f(y)$	distribuzione di probabilità
	funzione di densità (*pmf* / *pdf*)
$F(y)$	funzione di ripartizione (*cdf*)
$Y \sim f(y, \theta)$	modello della popolazione
θ	parametro(i) della popolazione
n	ampiezza campionaria
$N(\rightarrow \infty)$	ampiezza della popolazione

Verosimiglianza

La funzione di verosimiglianza

- Ruolo centrale per l'inferenza e la modellistica parametrica
- Utilizzata per condurre le tre tradizionali procedure circa un parametro: stima puntuale, stima intervallare, verifica di ipotesi
- Strumento per ragionare sui e dai dati (Fisher): "the whole likelihood function, and not only its maximum, plays a crucial role for reasoning about data"

Il problema

Il modello della popolazione: $Y \sim f(y, \theta)$

L'informazione disponibile:

- il modello (processo) che genera i dati - DGP
- il campione osservato

$$y_1, y_2, \cdots, y_n$$

La funzione di verosimiglianza è in grado di raccogliere entrambi i tipi di informazione:

- il DGP è utilizzata tramite il modello probabilistico
- il campione osservato è utilizzato per stabilire il grado di plausibilità (o verosimiglianza) dei singoli valori che θ può assumere

Il campione casuale

Modello della popolazione: $Y \sim f(y, \theta)$, dove $\theta \in \Theta$ (spazio parametrico)

Sia $(Y_1, Y_2, \dots, Y_n)$ un campione casuale, dove:

$$Y_i \sim f(y, \theta), i = (1, \dots, n)$$

Nel caso di campionamento con reimmissione (o senza ripetizione con $N \to \infty$), la funzione di probabilità congiunta del campione casuale:

$$\Psi(\mathbf{y}; \theta) = \Psi(y_1, y_2, \dots, y_n, \theta) = \prod_{i=1}^{n} f(y_i; \theta)$$

Due usi della funzione di probabilità congiunta

uso probabilistico (prima di osservare il campione)

- $\Psi(\mathbf{y};\theta)$ è la probabilità di osservare il campione $(y_1, y_2, \dots, y_n)$ se Y è una v.c. discreta

- è proporzionale alla probabilità che il campione casuale $(Y_1, Y_2, \dots, Y_n)$ assuma valori "vicino" $(y_1, y_2, \dots, y_n)$ se Y è una v.c. continua

uso inferenziale (dopo aver osservato il campione)

- $\Psi(\mathbf{y};\theta)$ è una funzione di θ sullo spazio parametrico Θ

- non è una distribuzione di probabilità anche se fornisce (o è proporzionale) alla probabilità per ciascun valore fissato di θ

Un semplice esperimento:

Consideriamo il semplice esperimento consistente nel lancio di una moneta. Si ha:

$$Y \sim Ber(\theta)$$

$$f(y;\theta) = \theta^y \left(1-\theta\right)^{(1-y)}$$

dove: $y = 0, 1$ e $0 \leq \theta \leq 1$

Esercizio (matematico)

Derivare $\Psi(\mathbf{y};\theta) = \prod_{i=1}^{n} f(y_i;\theta)$

Esercizio (R)

Scrivere una funzione R per $\Psi(\mathbf{y};\theta)$

1. Usare il nome `lik_bernoulli`
2. Usare due argomenti:
    a. il campione (`campione`)
    b. il parametro (`theta`) con valore di default fissato a 0.5

L'uso probabilistico di $\Psi(\mathbf{y}; \theta)$

L'obiettivo è di ragionare circa i possibili campioni che possiamo osservare (ci possiamo aspettare)

```
lik_bernoulli <- function(campione, theta = 0.5){
    n <- length(campione)
    theta^sum(campione) *
      (1 - theta)^(n - sum(campione))
}
```

```
flips <- c(0, 1, 0, 0, 0, 0, 0, 0, 1, 0)
lik_bernoulli(flips)
```

[1] 0.0009765625

Un punto di vista frequentista su $\Psi(\mathbf{y}; \theta)$

```
sample(c(0, 1), size = 10,
       prob = c(0.5, 0.5), replace = TRUE)
```

```
##  [1] 1 0 1 0 1 1 0 0 1 1
```

```
set.seed(1)
rep_samples <- replicate(n = 100000,
                         expr = sample(c(0, 1),
                                       size = 10,
                                       prob = c(0.5, 0.5),
                                       replace = TRUE))
sum(apply(rep_samples, 2, identical, flips)) / 100000
```

```
## [1] 0.00098
```

L'uso probabilistico di $\Psi(\mathbf{y}; \theta)$

- La probabilità è la stessa per qualsiasi campioni con 2 successi e 8 insuccessi
- Se siamo interessati alla probabilità di osservare $n = 2$ successi abbiamo:

$$\binom{n}{y} \theta^{\sum y_i} (1 - \theta)^{(n - \sum y_i)}$$

```
choose(10, 2)
dbinom(x = 2, size = 10, prob = 0.5)
sum(colSums(rep_samples) == 2) / 100000
```

L'uso inferenziale di $\Psi(\mathbf{y};\theta) \rightarrow \mathcal{L}(\theta)$

- L'obiettivo è di ragionare circa il modello della popolazione, ed in particolare su θ
- Per un dato valore di θ, fornisce la probabilità (caso discreto) o è proporzionale alla probabilità (caso continuo) di osservare il campione disponibile
- Se θ è incognito è uno strumento per ragionare su di esso a partire dall'informazione disponibile:
  1. modello della popolazione: $Y \sim Ber(\theta)$
  2. campione osservato

```
flips
```

```
##  [1] 0 1 0 0 0 0 0 0 1 0
```

La funzione di verosimiglianza per il modello di Bernoulli

$$\mathcal{L}(\theta; \mathbf{y}) = \mathcal{L}(\theta) = \theta^{\sum y_i}(1-\theta)^{(n-\sum y_i)}$$

```
flips
```

```
##  [1] 0 1 0 0 0 0 0 0 1 0
```

- Qual è la probabilità di ottenere il campione osservato per i differenti valori di θ?
- Qual è la migliore "scommessa" su θ?
- Quali valori di θ possiamo escludere?

La funzione di verosimiglianza per il modello di Bernoulli

La funzione di verosimiglianza fornisce una sorta di ordinamento dei valori di θ:

```
lik_bernoulli(flips, theta = 0.2)
```

```
## [1] 0.006710886
```

```
lik_bernoulli(flips, theta = 0.5)
```

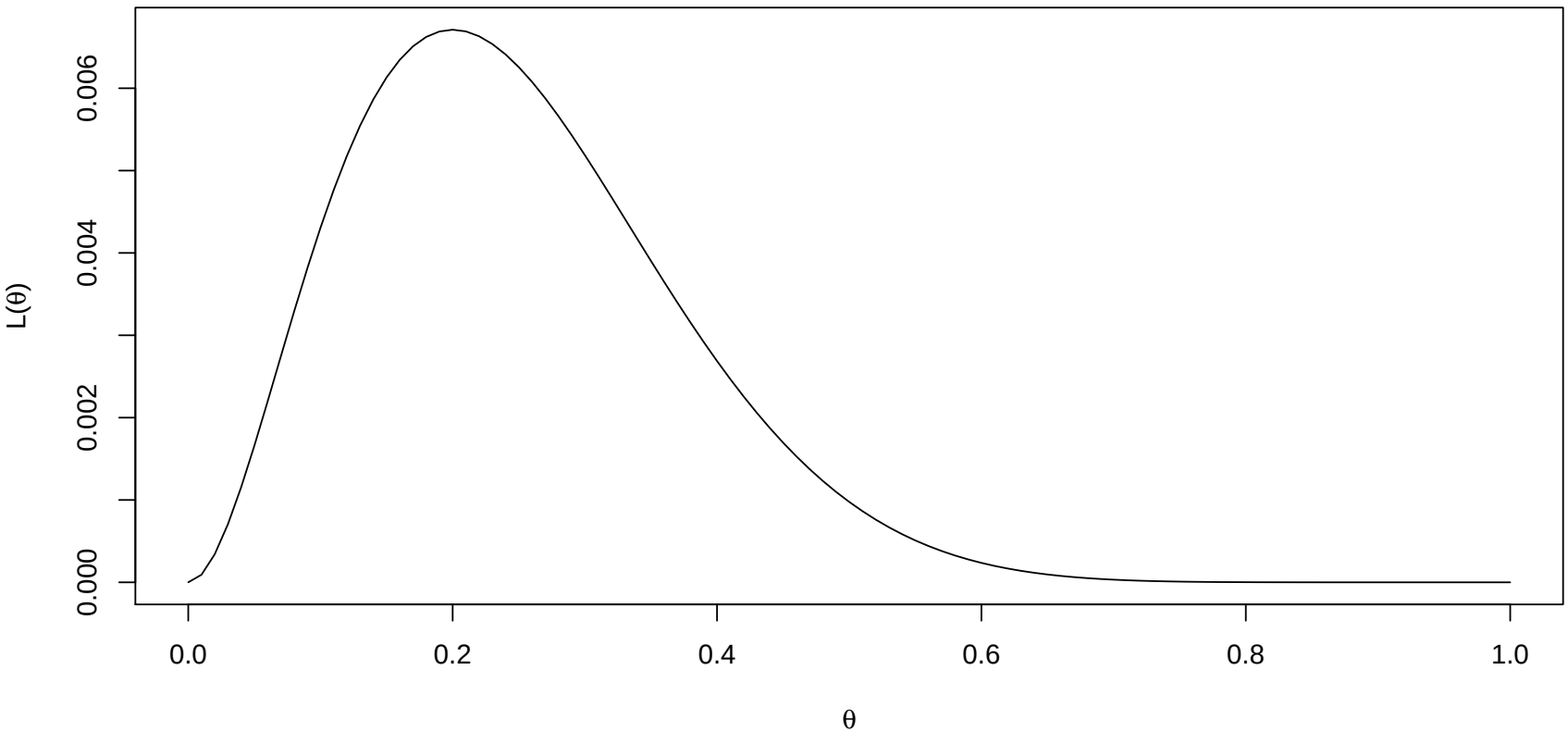
```
## [1] 0.0009765625
```

```
lik_bernoulli(flips, theta = 0.8)
```

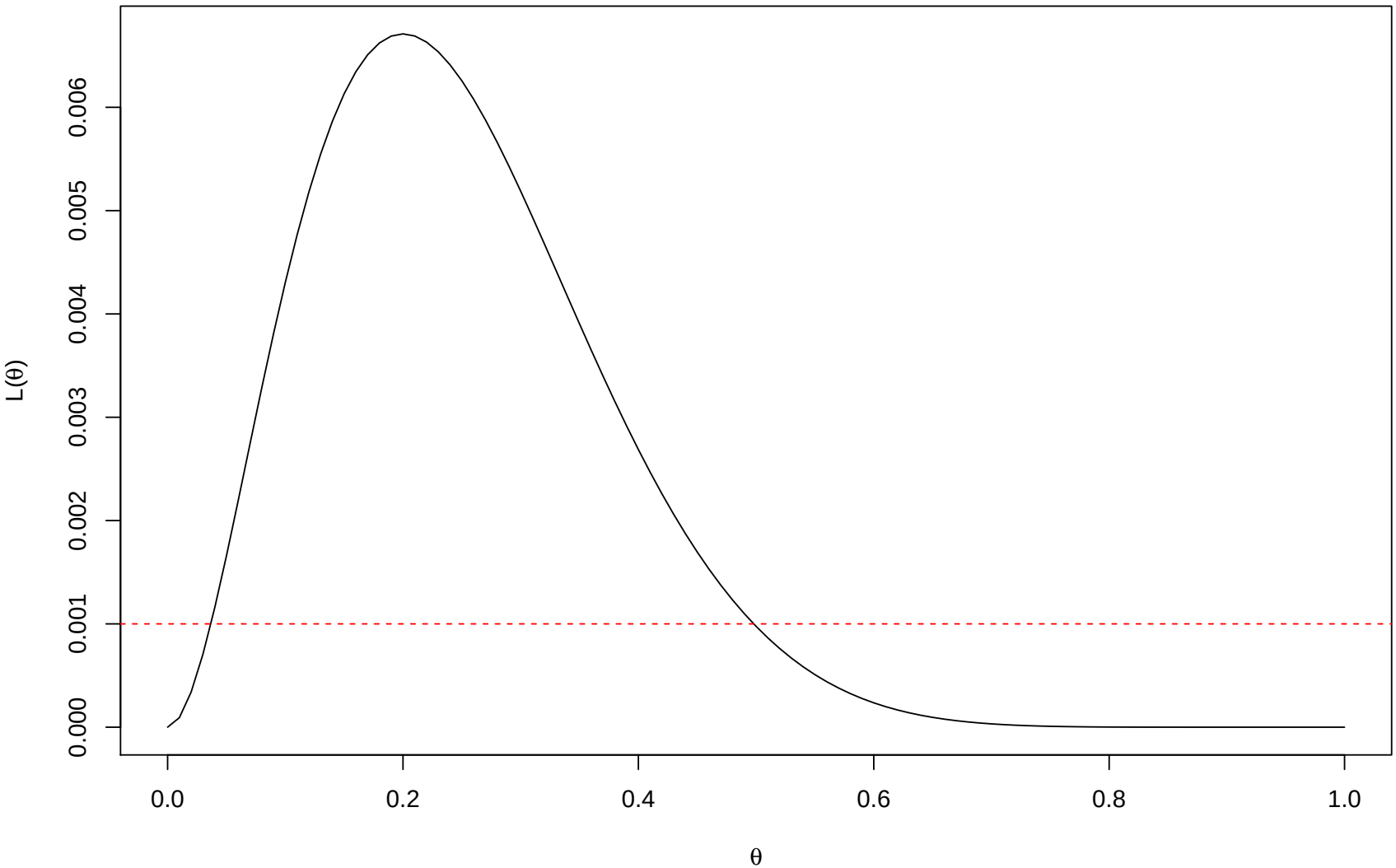
```
## [1] 1.6384e-06
```

La funzione di verosimiglianza per il modello di Bernoulli

```r
theta <- seq(0, 1, by = 0.01)
plot(theta, lik_bernoulli(flips, theta), type = "l",
     xlab = expression(theta),
     ylab = expression(paste("L(", theta, ")")))
```



Un unico valore o più valori verosimili?



Una necessaria premessa

- L'incertezza nei dati è veicolata dalla/nella funzione di verosimiglianza
- La verosimiglianza fornisce una misura di preferenza relativa tra differenti valori di θ
- La funzione di verosimiglianza è **sufficiente minimale** per θ
- Ruolo cruciale del modello di probabilità della popolazione $Y \sim f(y; \theta)$: la manipolazione matematica della verosimiglianza è (tipicamente) precisa ma la scelta del modello è (il più delle volte) incerta
- Per sfruttare la teoria della verosimiglianza è necessario un modello ma un modello sbagliato può condurre a conclusioni sbagliate (alternativa non parametrica: verosimiglianza empirica)

Il rapporto di verosimiglianza

- La funzine di verosimiglianza ordina i possibili valori di θ in base alle corrispondenti probabilità di generare i dati osservati
- La funzione di verosimiglianza è uguale o proporzionale a $\Psi(\theta; y_1, \dots, y_n)$
- Il nostro interesse non è (solo) sui singoli valori di $\mathcal{L}(\theta)$ ma nel confronto tra due valori:

$$\frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_2)}$$

- Il rapporto di verosimiglianza fornisce una misura di verosimiglianza di θ_1 rispetto a θ_2

La legge di verosimiglianza

Il rapporto di verosimiglianza è una misura della forza dell'evidenza campionaria a supporto di un'ipotesi (valore di un parametro) contro un'altra:

$$\Lambda = \frac{L(\theta_1|Y=\mathbf{y})}{L(\theta_2|Y=\mathbf{y})} = \frac{P(Y=\mathbf{y}|\theta_1)}{P(Y=\mathbf{y}|\theta_2)}$$

è il grado per cui l'evidenza $\mathbf{y}$ supporta il valore del parametro (ipotesi) θ_1 contro θ_2

- Se il rapporto è 1, le due ipotesi sono indifferenti
- Se il rapporto è maggiore (minore) di 1, l'evidenza campionaria supporta θ_1 contro θ_2 (o viceversa)
- L'uso dei fattori di Bayes può estendere questo ragionamento permettendo di tenere in considerazione la complessità delle differenti ipotesi

La definizione formale

Sia $(y_1, y_2, \dots, y_n)$ il campione osservato dalla popolazione $Y \sim f(y; \theta)$, la funzione di verosimiglianza è:

$$\mathcal{L}(\theta) = \mathcal{L}(\theta; \mathbf{y}) = \mathcal{L}(\theta; y_1, \dots, y_n) = \Psi(\theta; y_1, \dots, y_n) = \prod_{i=1}^{n} f(y_i; \theta)$$

dove $(y_1, y_2, \dots, y_n)$ è dato e θ è variabile nello spazio parametrico Θ.

Una definizione più formale

Sia $(y_1, y_2, \dots, y_n)$ il campione osservato dalla popolazione $Y \sim f(y; \theta)$, la funzione di verosimiglianza è:

$$\mathcal{L}(\theta) = \mathcal{L}(\theta; \mathbf{y}) \propto \prod_{i=1}^{n} f(y_i; \theta)$$

- Possiamo trascurare termini moltiplicativi che non dipendono da θ

La verosimiglianza relativa (normalizzata)

Poichè solo i rapporti di verosimiglianza hanno senso ai fini di conclusioni su θ, è conveniente normalizzare la verosimiglianza rispetto al suo massimo:

$$\mathcal{R}(\theta) = \mathcal{R}(\theta; \mathbf{y}) = \frac{\mathcal{L}(\theta; \mathbf{y})}{\mathcal{L}(\hat{\theta}; \mathbf{y})}$$

- La MLE $\hat{\theta}$ è il valore più verosimile per θ poichè rende il campione osservato più "probabile"
- La verosimiglianza relativa misura la plausibilità di qualsiasi valori di θ rispetto a $\hat{\theta}$
- La verosimiglianza relativa varia tra 0 e 1

La log-verosimiglianza

- Per molte applicazioni, il logaritmo naturale della funzione di verosimiglianza (log-verosimiglianza) risulta più conveniente da trattare
- Poichè il logaritmo è una funzione monoticamente crescente, il logaritmo di una funzione ha il suo massimo negli stessi punti della funzione stessa, per cui la log-verosimiglianza può essere usata al posto della verosimiglianza nella MLE (e in generale per le altre procedure inferenziali)

La log-verosimiglianza relativa (normalizzata)

La log-verosimiglianza relativa:

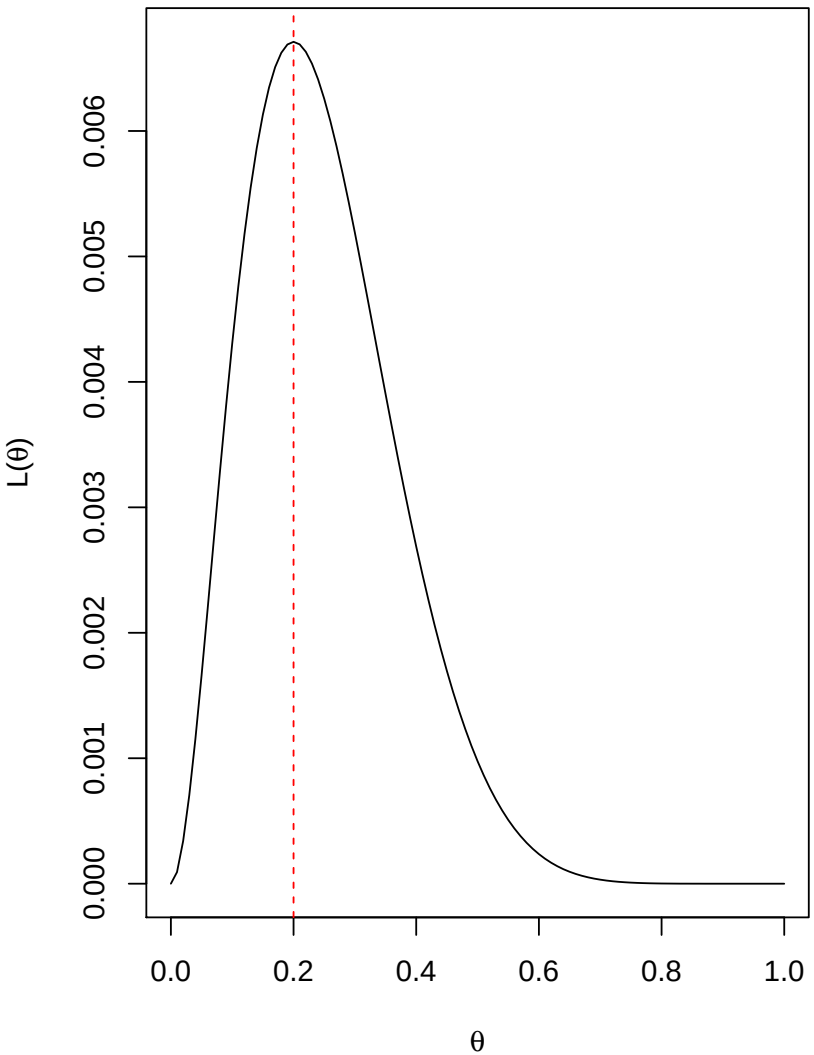
$$r(\theta) = log \frac{\mathcal{L}(\theta; \mathbf{y})}{\mathcal{L}(\theta; \widehat{\mathbf{y}})} = log\ \mathcal{L}(\theta; \mathbf{y}) - log\ \mathcal{L}(\hat{\theta}; \mathbf{y})$$

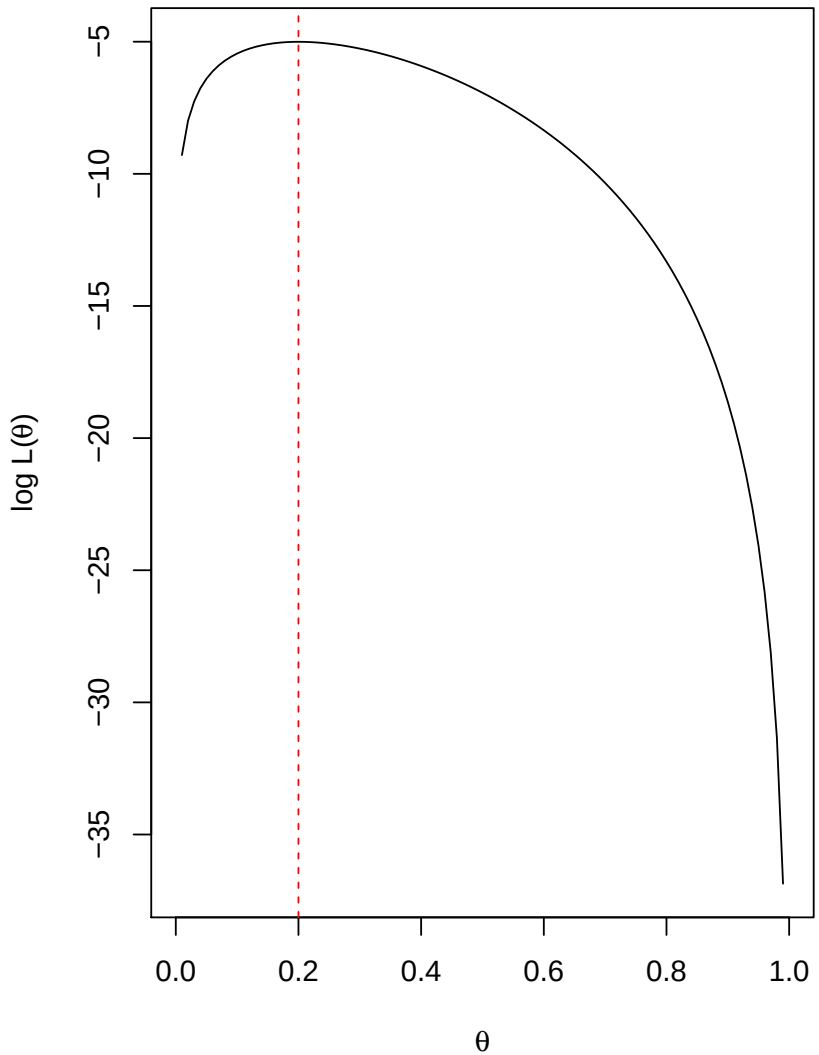
- La log-verosimiglianza relativa varia da $-\infty$ a 0
- Possiamo trascurare termini additivi che non dipendono da θ

Una funzione R per la log-verosimiglianza di Bernoulli

```r
llik_bernoulli <- function(observed_sample, theta = 0.5){
    n <- length(observed_sample)
    sum(observed_sample) * log(theta) +
      (n - sum(observed_sample)) * log(1 - theta)
}
```

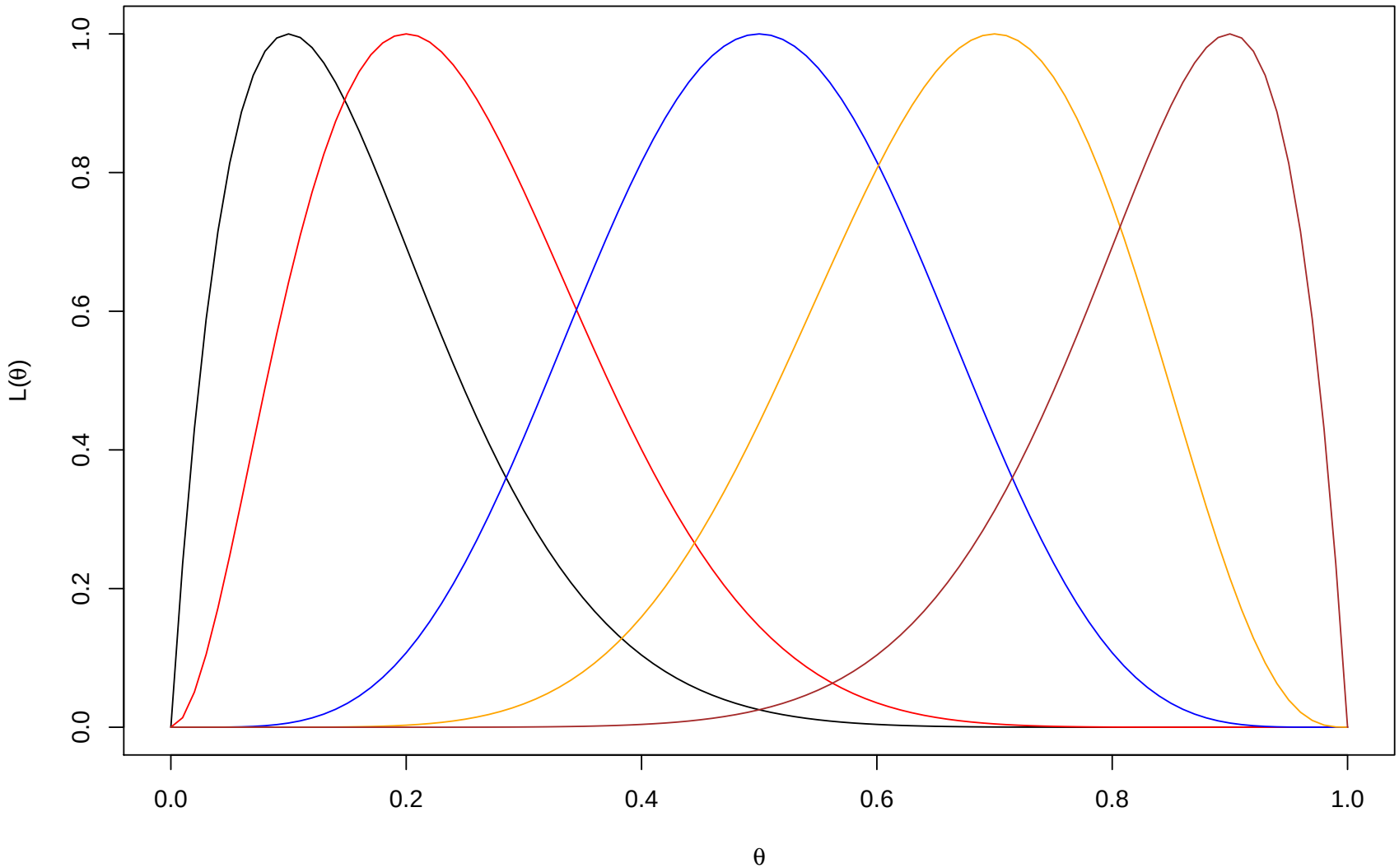
Verosimiglianza e log-verosimiglianza: confronto





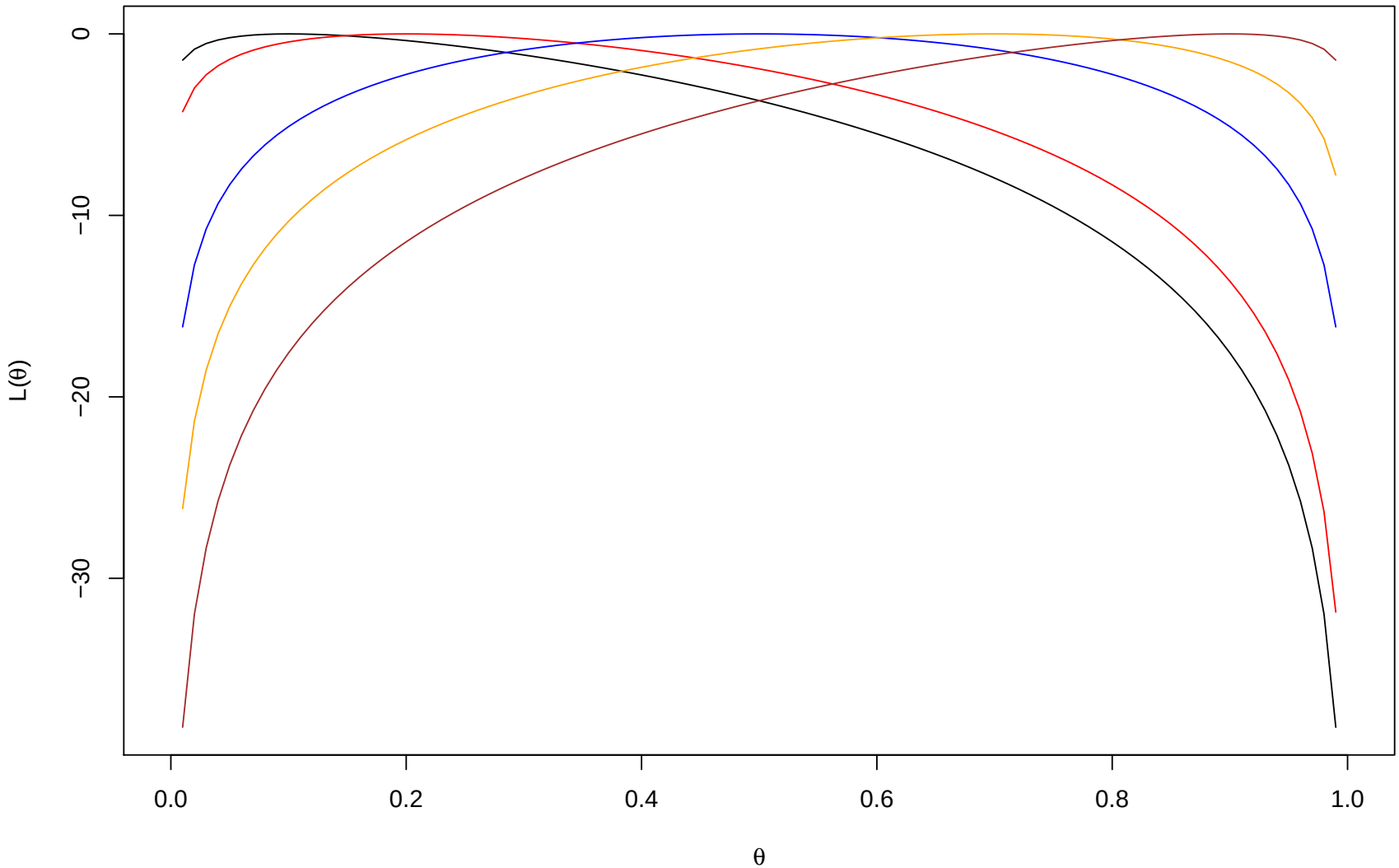
L'effetto del campione osservato

Cambiando il $\#(successi)$: $\sum y_i = \{1, 3, 5, 7, 9\}$, con $n = 10$



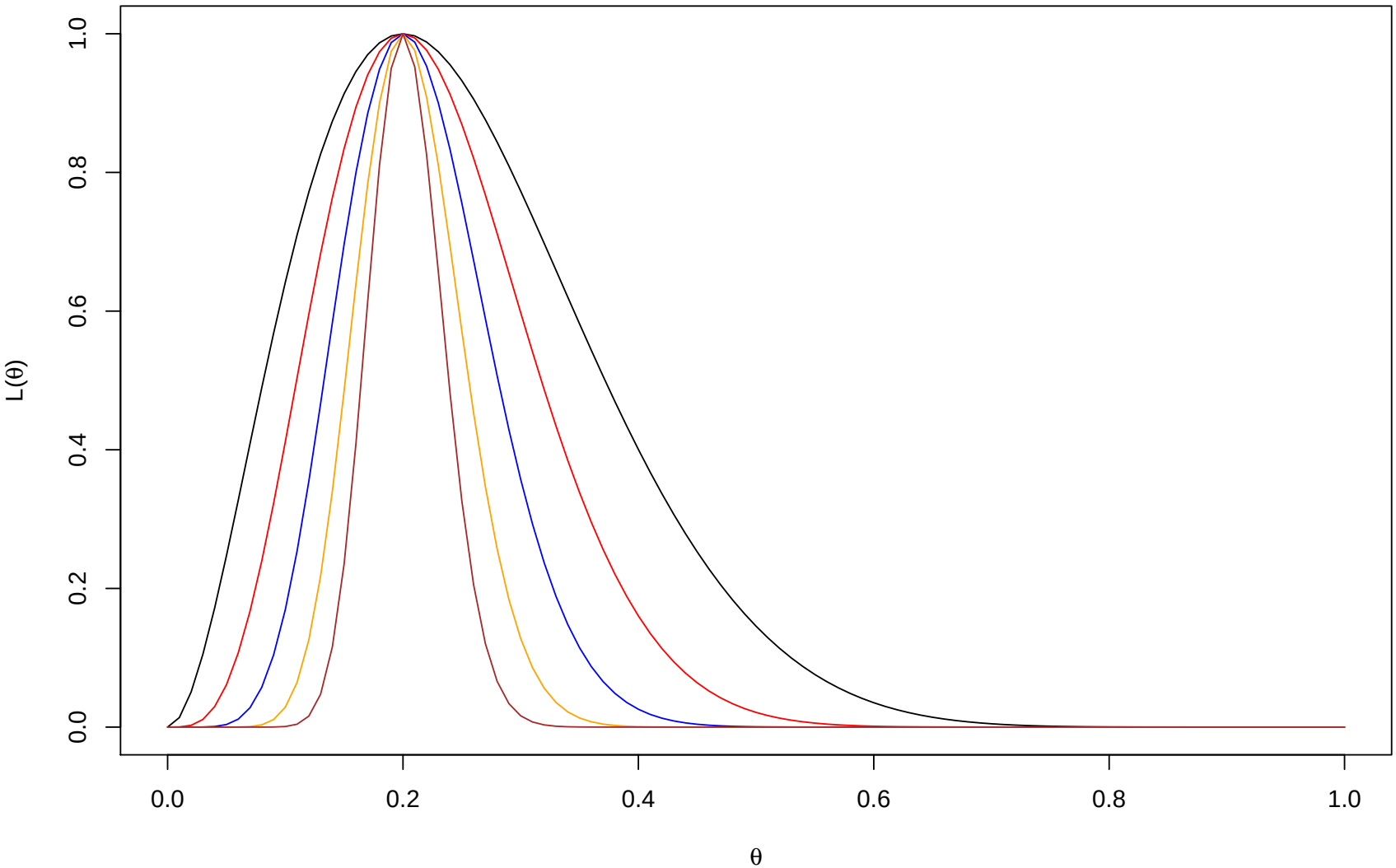
L'effetto del campione osservato

Cambiando il $\#(successi)$: $\sum y_i = \{1, 3, 5, 7, 9\}$, con $n = 10$



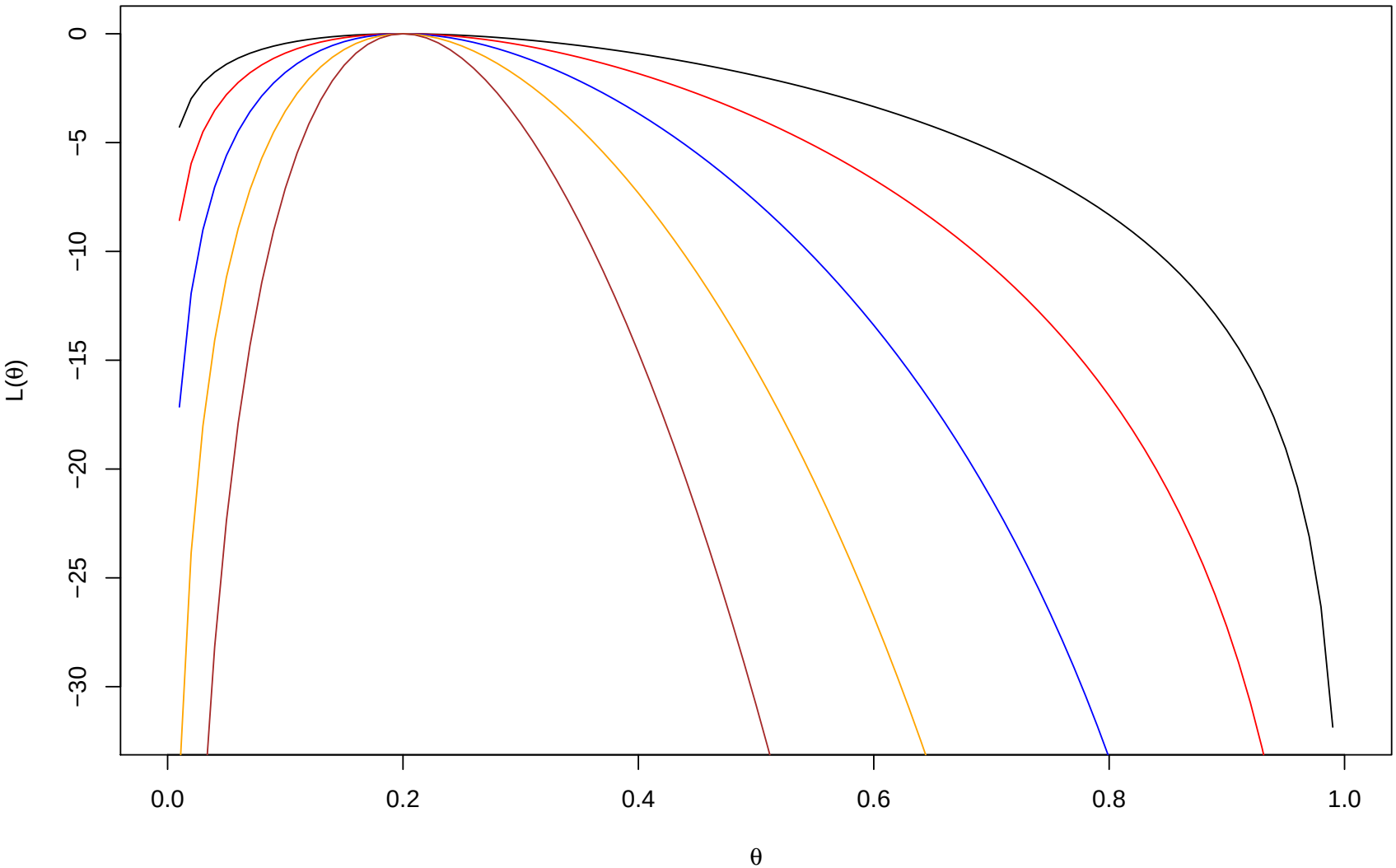
L'effetto del campione osservato

Aumentando l'ampiezza campionaria da $n = 10$ a $n = 160$, con $\hat{\theta} = 0.2$



L'effetto del campione osservato

Aumentando l'ampiezza campionaria da $n = 10$ a $n = 160$, con $\hat{\theta} = 0.2$



Tecnicalità

La funzione di log-verosimiglianza:

$$l(\theta) = log\ \mathcal{L}(\theta)$$

La funzione Score:

$$S(\theta) = \frac{\partial l(\theta)}{\partial \theta}$$

$\hat{\theta}$, la stima di massima verosimiglianza (MLE), è la soluzione dell'equazione:

$$S(\theta) = 0$$

In corrispondenza del massimo, avremo:

$$\frac{\partial^2 l(\theta)}{\partial \theta^2} < 0$$

—

Tecnicalità

L'informazione osservata di Fisher:

$$I(\hat{\theta}) = - \left. \frac{\partial^2 l(\theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}}$$

Un'ampia curvatura è associata con un picco stretto (o forte), ad indicare minore incertezza su θ

Il ruolo della MLE

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}(\theta) \equiv \underset{\theta \in \Theta}{\operatorname{argmax}}\, l(\theta)$$

- “It is the entire likelihood that captures all the information in the data about a certain parameter, not just its maximizer”” (Pawittan, 2001)
- La MLE permette di semplificare la presentazione della funzione di verosimiglianza ma è raramente “sufficiente” per rappresentarla adeguatamente
- Se la log-verosimiglianza è ben approssimata da una funzione quadratica (problemi regolari), abbiamo bisogno di due quantità per rappresentarla: la posizione del massimo e la curvatura nel punto di massimo

Log-verosimiglianza per il modello di Bernoulli

La funzione di verosimiglianza:

$$\mathcal{L}(\theta) = \theta^{\sum y_i} (1-\theta)^{(n-\sum y_i)}$$

La funzione di log-verosimiglianza:

$$l(\theta) = \log \mathcal{L}(\theta) = \sum y_i \log(\theta) + (n - \sum y_i) log (1-\theta)$$

—

Homework 1 (verosimiglianza)

Sia $Y \sim Ber(\pi)$, dove:

$$f(y; \pi) = \pi^y (1 - \pi)^{(1-y)}, \qquad y = \{0, 1\}$$

Esercizio (matematico)

- Derivare la MLE per il parametro π
- Derivare l'informazione osservata di Fisher

Homework 1 (soluzione)

$$\hat{\theta} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$I(\hat{\theta}) = \frac{n}{\theta(1-\theta)}$$

Qualcosa di familiare?

Homework 2 (verosimiglianza)

Sia $Y \sim Poi(\lambda)$, dove $f(y; \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$

Esercizio (matematico)

- Derivare la log-verosimiglianza per il modello di probabilità di Poisson
- Derivare la MLE per il parametro λ
- Derivare l'informazione osservata di Fisher

Homework (verosimiglianza)

R exercise

Il numero di infortuni sul lavoro in uno stabilimento di produzione segue una distribuzione di Poisson.

- Scrivere una funzione R per la funzione di log-verosimiglianza di Poisson

- Considerare il seguente campione di $n = 10$ mesi:
  1 3 0 0 5 3 1 6 2 4
  - Rappresentare graficamente la log-verosimiglianza
  - Calcolare la stima di massima verosimiglianza l'informazione osservata di Fisher

- Considerare il seguente campione di $n = 20$ mesi:
  0 0 2 1 4 4 2 0 1 5 3 1 5 1 3 5 1 4 3 5
  - Rappresentare graficamente la log-verosimiglianza
  - Calcolare la stima di massima verosimiglianza

- discutere i risultati ottenuti

La funzione di verosimiglianza per il modello normale

Sia $Y \sim N(\mu, \sigma^2)$, dove:

$$f(\mathbf{y}; \mu, \sigma^2) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \exp\left[-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right]$$

In termini di una normale standard:

$$f(\mathbf{y}; \mu, \sigma^2) = \frac{1}{\sigma} \underbrace{\frac{1}{\sqrt{2\,\pi}} \exp\left[-\frac{1}{2}(\frac{y-\mu}{\sigma})^2\right]}_{\phi(\frac{y-\mu}{\sigma})}$$

Percioò la funzione di verosimiglianza è:

$$\mathcal{L}(\theta; \mathbf{y}) = \prod_{i=1}^{n} f(y_i; \mu, \sigma^2) = \left(\frac{1}{2\,\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\left[-\frac{\sum (y_i-\mu)^2}{2\,\sigma^2}\right]$$

La funzione di log-verosimiglianza per il modello normale

Sia $Y \sim N(\mu, \sigma^2)$, la log-verosimiglianza è:

$$l(\theta; \mathbf{y}) = \sum_{i=1}^{n} \log f(y_i; \mu, \sigma^2) = -\frac{n}{2} \log(2\,\pi) - n \log \sigma - \frac{\sum (y_i - \mu)^2}{2\,\sigma^2}$$

Approssimazione quadratica della funzione di log-verosimiglianza

Consideriamo un'approssimazione di Taylor del secondo ordine intorno alla stima di massima verosimiglianza $\hat{\theta}$:

$$l(\theta) \approx l(\hat{\theta}) + \underbrace{\left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}}}_{S(\hat{\theta})=0} (\theta - \hat{\theta}) + \frac{1}{2} \underbrace{\left.\frac{\partial^2 l(\theta)}{\partial^2 \theta}\right|_{\theta=\hat{\theta}}}_{-I(\hat{\theta})} (\theta - \hat{\theta})^2$$

L'approssimazione quadratica della funzione di log-verosimiglianza relativa è:

$$r(\theta) = l(\theta) - l(\hat{\theta}) \approx -\frac{1}{2} I(\hat{\theta}) (\theta - \hat{\theta})^2$$

Alcune proprietà della funzione di verosimiglianza

Non additività

- Distinzione netta tra probabilità e verosimiglianza
- La verosimiglianza è una funzione di punto, la probabilità è una funzione di insieme
- Le probabilità sono associate agli eventi, le verosimiglianza sono associate con ipotesi semplici

La mancanza di additività implica anche le verosimiglianze non possono essere ottenute per intervalli: la verosimiglianza è una funzone di punto che assegna plausibilità relative a valori singoli del parametro all'interno dell'intervallo

Alcune proprietà della funzione di verosimiglianza

Combinazione di osservazioni

- semplicità per combinare dati proveniente da differente esperimenti: poichè la probabilità congiunta di eventi indipendenti è il prodotto delle loro probabilità individuali, la verosimiglianza di θ basata su dataset indipendenti è il prodotto delle singole verosimiglianze basata su ogni dataset separatamente
- Si applica a dati provenienti da differenti esperimenti, purchè le verosimiglianze sono riferite allo stesso parametro

Alcune proprietà della funzione di verosimiglianza

Combinazione di osservazioni (continua)

- A causa della semplicità di combinare le log-verosimiglianze attraverso l'addizione, la log-verosimiglianza è utilizzata più della verosimiglianza stessa: le log-verosimiglianze basate su dataset indipendenti sono combinate attraverso l'addizione per ottenere la log-verosimiglianza combinata su tutti i dati

Alcune proprietà della funzione di verosimiglianza

Invarianza funzionale

La verosimiglianza è invariante in termini funzionali (come le probabilità, ma a differenza delle densità di probabilità)

- qualsiasi affermazione quantitativa su θ implica una corrispondente affermazione circa qualsiasi funzione 1 to 1 $\delta = \delta(\theta)$ attraverso la sostituzione algebrica diretta $\theta = \theta(\delta)$

Alcune proprietà della funzione di verosimiglianza

Invarianza funzionale (continua)

- questo è abbastanza utile da un punto di vista applicativo: in molti casi qualche altro parametro potrebbe essere di maggiore interesse di θ
- spesso un cambiamento di parametro potrebbe semplificare la forma della funzione di verosimiglianza: $R(\delta)$ potrebbe essere più simmetrica, o approssimativamente normale in forma, rispetto a $R(\theta)$
- Le inferenze in termini di δ saranno strutturalmente più semplici, ma matematicamente equivalenti, a quelle in termini di θ