La verosimiglianza

Capitolo 20

Analisi statistica dei dati biologici

La verosimiglianza

- Come possiamo scoprire il valore esatto di un parametro della popolazione tra i molti possibili?

Se siamo in possesso di un campione estratto dalla popolazione, dovremmo riuscire a misurare quanto i possibili valori alternativi del parametro si adattano ai dati e confrontare la qualità dell'adattamento tra loro.

La verosimiglianza

La verosimiglianza misura la bontà con cui un insieme di dati «sostiene» un particolare valore di un parametro.

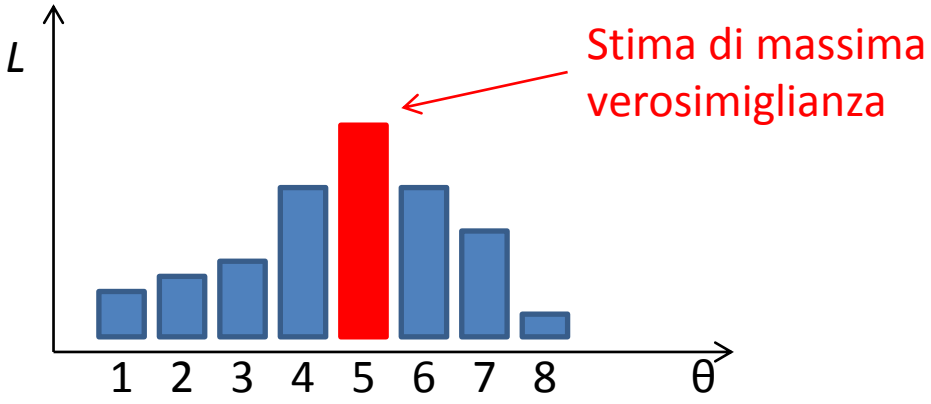
La verosimiglianza (*L*, detta anche *likelihood*) di uno specifico valore del parametro (θ_{oss}) è la probabilità di ottenere i dati osservati se il parametro fosse uguale proprio a quello specifico valore:

$$L(\theta_{oss} \mid \text{dati}) = \Pr(\text{dati} \mid \theta = \theta_{oss})$$

La verosimiglianza

La verosimiglianza (*L*) di un particolare valore dice poco in se per se, ma acquista significato quando viene confrontata con le verosimiglianze degli altri valori possibili.

Il valore del parametro che riceve il massimo sostegno fra tutti i valori possibili è la **stima di massima verosimiglianza**: il valore per il quale è massima la probabilità di ottenere i dati osservati.



La stima di massima verosimiglianza

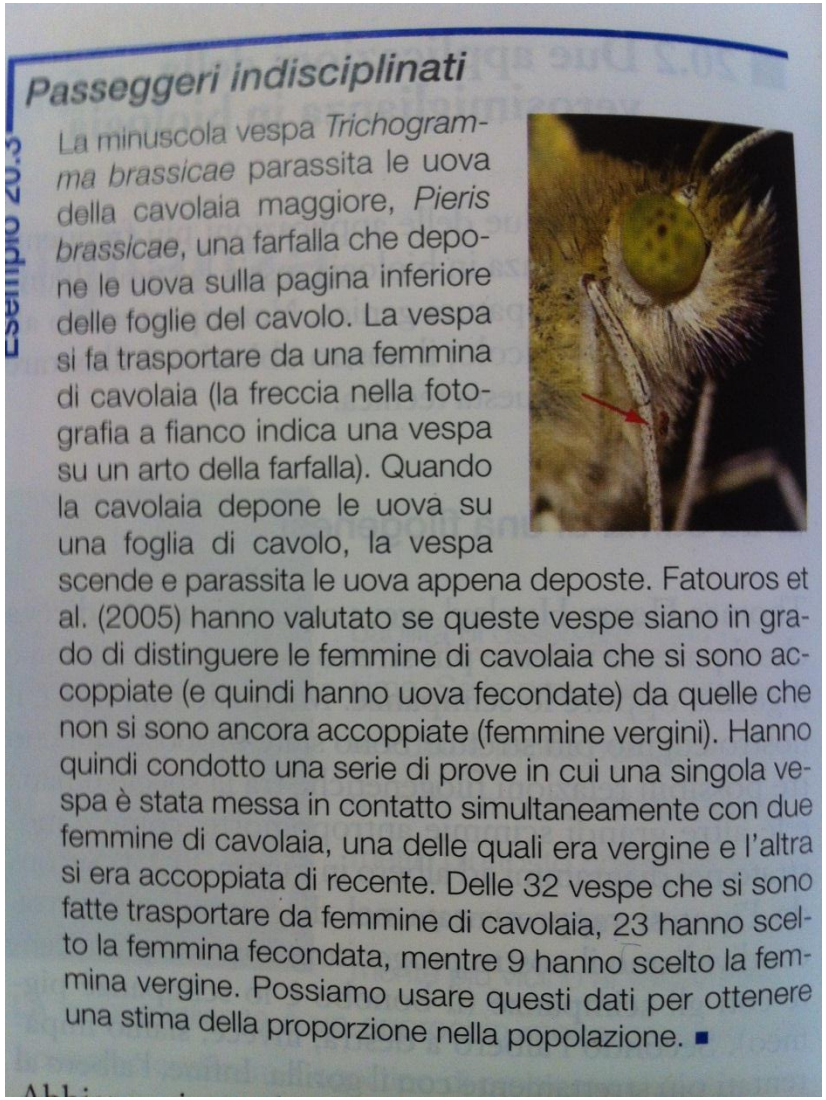
$$L(\theta_{oss} \mid dati) = Pr(dati \mid \theta = \theta_{oss})$$

La verosimiglianza **è una probabilità**.

La stima di massima verosimiglianza richiede di trovare **il valore del parametro che ha la L maggiore**.

Abbiamo bisogno di un modello probabilistico che permetta di calcolare la probabilità che diversi valori del parametro abbiano generato i dati osservati

La stima di massima verosimiglianza - Esempio



Esempio 20.3

Passeggeri indisciplinati

La minuscola vespa *Trichogramma brassicae* parassita le uova della cavolaia maggiore, *Pieris brassicae*, una farfalla che depone le uova sulla pagina inferiore delle foglie del cavolo. La vespa si fa trasportare da una femmina di cavolaia (la freccia nella fotografia a fianco indica una vespa su un arto della farfalla). Quando la cavolaia depone le uova su una foglia di cavolo, la vespa scende e parassita le uova appena deposte. Fatouros et al. (2005) hanno valutato se queste vespe siano in grado di distinguere le femmine di cavolaia che si sono accoppiate (e quindi hanno uova fecondate) da quelle che non si sono ancora accoppiate (femmine vergini). Hanno quindi condotto una serie di prove in cui una singola vespa è stata messa in contatto simultaneamente con due femmine di cavolaia, una delle quali era vergine e l'altra si era accoppiata di recente. Delle 32 vespe che si sono fatte trasportare da femmine di cavolaia, 23 hanno scelto la femmina fecondata, mentre 9 hanno scelto la femmina vergine. Possiamo usare questi dati per ottenere una stima della proporzione nella popolazione. ■

In questo esempio siamo interessati ad effettuare un'inferenza sul parametro *p* nella popolazione (proporzione di vespe che scelgono la femmina fecondata)

Per rispondere alla domanda potremmo seguire due approcci:

- Eseguire un test binomiale dove:
  H0: il parametro *p* nella popolazione è 0.5
  HA: il parametro p nella popolazione è diverso da 0.5

- **Effettuare una stima di massima verosimiglianza di *p***

La stima di massima verosimiglianza - Esempio

- **Il modello probabilistico**

La stima di massima verosimiglianza richiede un modello probabilistico che specifichi le probabilità di differenti risultati del processo di campionamento in funzione del parametro che viene stimato.

Nel nostro caso, se ciascuna delle *n* vespe nel campione rappresenta una prova casuale e le diverse prove sono indipendenti, allora **il numero di vespe che scelgono la femmina fecondata** dovrebbe adattarsi a una **distribuzione binomiale**:

$$\Pr[Y\ scelgono\ la\ femmina\ fecondata \mid p] = \binom{n}{Y} p^{Y}(1-p)^{n-Y}$$

Questa formula ci permette di calcolare la probabilità che **esattamente** Y vespe scelgano femmine fecondate.

Questa formula ci permette di variare p per un valore fisso di Y (23 vespe hanno scelto la femmina fecondata) e vedere come la variazione influenzi la probabilità di ottenere i dati osservati.

La stima di massima verosimiglianza - Esempio

- **La formula della verosimiglianza**

Il modello probabilistico (distribuzione binomiale) ci permette di calcolare la verosimiglianza di un particolare valore di p dato che Y vespe scelgono la femmina fecondata:

$$L[p|Y\ scelgono\ la\ femmina\ fecondata] = \binom{n}{Y} p^{Y}(1-p)^{n-Y}$$

Per calcolare la verosimiglianza di p poniamo Y=23 (numero di vespe che scelgono la femmina fecondata osservato) e n=32 (numero totale di prove)

$$L[p|23\ scelgono\ la\ femmina\ fecondata] = \binom{32}{23} p^{23}(1-p)^{32-23}$$

La stima di massima verosimiglianza - Esempio

- **La formula della verosimiglianza**

La verosimiglianza di p=0.5 :

$$L[0.5|23\ scelgono\ la\ femmina\ fecondata] = \binom{32}{23} 0.5^{23}(1-0.5)^{9}$$
$$= 0.00653$$

Questa quantità *L* rappresenta il sostegno all'ipotesi che esattamente la metà delle vespe nella popolazione scelgano la femmina fecondata, dato che l'hanno fatto 23 delle 32 vespe nel campione.

La stima di massima verosimiglianza - Esempio

- **La formula della verosimiglianza**

Generalmente è più facile lavorare con il logaritmo naturale della verosimiglianza (log-verosimiglianza):

$$lnL[p|Y\ scelgono\ la\ femmina\ fecondata] = ln\left[\binom{n}{Y}\right] + Yln[p] + (\mathrm{n} - \mathrm{Y})\ln[1 - \mathrm{p}]$$

per cui sostituendo alla formula i dati come in precedenza otteniamo:

$$lnL[0.5|23\ scelgono\ la\ femmina\ fecondata] = ln\left[\binom{32}{23}\right] + 23ln[0.5] + (9)\ln[0.5]$$
$$= -5.03125$$

Questo valore è uguale al ln di 0.00653 calcolato in precedenza senza l'uso dei logaritmi.

La stima di massima verosimiglianza - Esempio

- **La stima di massima verosimiglianza**

La stima di massima verosimiglianza di un parametro è quel valore del parametro che ha la più alta verosimiglianza considerati i dati osservati.

Il valore per parametro che massimizza la verosimiglianza è anche quello che massimizza la funzione di log-verosimiglianza.

Un metodo semplice per trovare la stima di massima verosimiglianza **è calcolare con un computer il logaritmo della verosimiglianza nell'intervallo dei valori possibili** del parametro e scegliere quello più alto.

Alternativamente, quando la funzione non è troppo complessa è possibile **trovare analiticamente il massimo della funzione**.

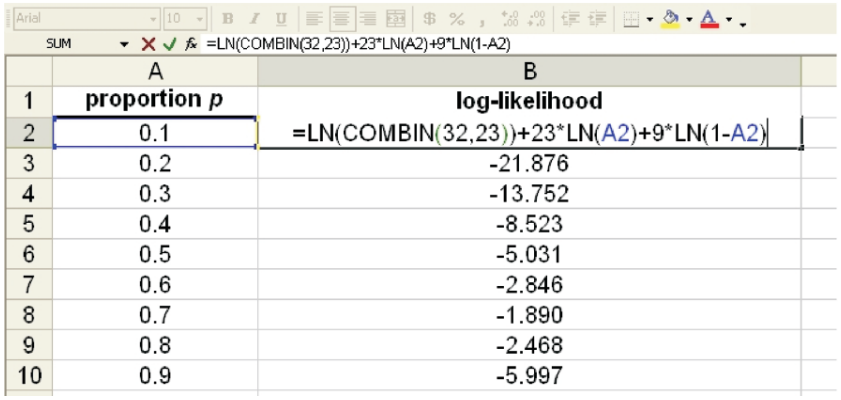
La stima di massima verosimiglianza - Esempio

- **La stima di massima verosimiglianza**

Usando un computer è molto semplice trovare il massimo: basta definire un intervallo ampio di valori e calcolare il logaritmo della verosimiglianza per ognuno di essi.

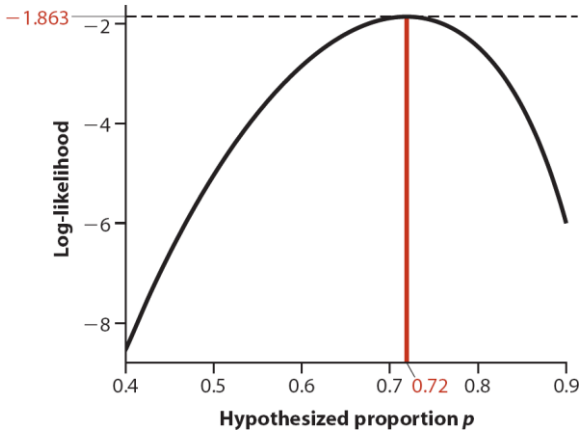
Nell'esempio potremmo calcolare la log-verosimiglianza nell'intervallo 0.1-0.9 e **costruire la curva di log-verosimiglianza**:

(*L* di 9 punti nell'intervallo)



	A	B
1	**proportion *p***	**log-likelihood**
2	0.1	=LN(COMBIN(32,23))+23*LN(A2)+9*LN(1-A2)
3	0.2	-21.876
4	0.3	-13.752
5	0.4	-8.523
6	0.5	-5.031
7	0.6	-2.846
8	0.7	-1.890
9	0.8	-2.468
10	0.9	-5.997

(Curva di log-verosimiglianza con 39 punti)



Analizzando la curva di log-verosimiglianza otteniamo una stima di massima verosimiglianza di p uguale a 0.72. Questo valore stima la proporzione nella popolazione.

La stima di massima verosimiglianza – Esempio con R

```
#salviamo in alcune variabili i dati a nostra disposizone:
n<-32
y<-23
#calcoliamo per esempio la verosimiglianza (L) di un particolare
#valore di p
#calcoliamo le varie parti della distribuzione binomiale per
#p=0.5
p<-0.5
L<-choose(n,y)*(p^y)*((1-p)^(n-y))
#proviamo a calcolare la log-verosimiglianza
logL<-log(choose(n,y))+y*log(p)+(n-y)*log(1-p)
#possiamo anche scrivere
logL<-lchoose(n,y)+y*log(p)+(n-y)*log(1-p)
```

La stima di massima verosimiglianza – Esempio con R

```
#proviamo adesso a calcolare logL per un intervallo di valori di p e creare la curva di log-verosimiglianza:
#definiamo 90 punti nell'intervallo di variazione di p da 0.1 a 0.9
int_p<-seq(from=0.1, to=0.9, length.out=90)
#calcolo la logL per tutti i punti nell'intervallo
logL<-lchoose(n,y)+y*log(int_p)+(n-y)*log(1-int_p)
#trovo il massimo di logL nell'intervallo
logLmax<-max(logL)
stima<-int_p[logL==logLmax]
#creo la curva di massima verosimiglianza
matplot(int_p, logL, type="l")
#aggiungo una riga verticale in corrispondenza del valore massimo di logL
abline(v= stima, col=2)
```

La stima di massima verosimiglianza - Esempio

- **Intervalli di confidenza basati sulla verosimiglianza**

La curva di verosimiglianza permette di calcolare, direttamente da essa, un intervallo di confidenza per il parametro nella popolazione.

L'intervallo di valori di p la cui log-verosimiglianza è situata entro $\chi^2_{1,\alpha}/2$ unità dal valore massimo costituisce **l'intervallo di confidenza basato sulla verosimiglianza** (con confidenza 1-α).

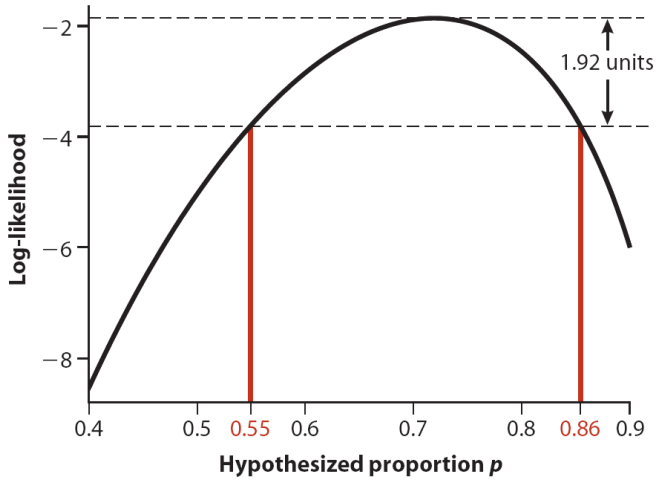
La stima di massima verosimiglianza - Esempio

- **Intervalli di confidenza basati sulla verosimiglianza**

Un intervallo di confidenza al 95% per p è l'intervallo di valori la cui log-verosimiglianza è situata entro 1.92 unità dal massimo ($\chi^2_{1,0.05}/2=3.84/2=1.92$) e sarà uguale a:

$$IC95\% = 0.55 < p < 0.86$$

La proporzione di vespe nella popolazione che scelgono la femmina fecondata è 0.55-0.86 e include valori di preferenza molto forti (0.86) e moderatamente deboli (0.55) ma comunque maggiori di 0.5 che indicherebbe nessuna preferenza.



La stima di massima verosimiglianza – Esempio con R – Intervalli di confidenza

```
#proviamo adesso a calcolare logL per un intervallo di valori di p e creare la curva di log-verosimiglianza:
#definiamo 90 punti nell'intervallo di variazione di p da 0.1 a 0.9
int_p<-seq(from=0.1, to=0.9, length.out=90)
#calcolo la logL per tutti i punti nell'intervallo
logL<-lchoose(n,y)+y*log(int_p)+(n-y)*log(1-int_p)
#trovo il massimo di logL nell'intervallo
logLmax<-max(logL)
stima<-int_p[logL==logLmax]
#calcolo la distanza di ogni punto dal valore massimo di logL
difL<-logL-max(logL)
#creo una matrice per il controllo visivo della distanza
cbind(int_p,difL,difL<=-1.92)
#controllo visivamente quali valori stanno a più di 1.92 unità dal massimo
#e scelgo i primi a destra e a sinistra del massimo
#0.549-0.855
```

Il test del rapporto di verosimiglianza

La stima di massima verosimiglianza fornisce una stima di un parametro nella popolazione, ma non permette di confrontare ipotesi.

IL **test del rapporto di verosimiglianza**, detto anche Likelihood ratio test, utilizza la verosimiglianza per confrontare la bontà con cui due modelli probabilistici (ipotesi) si adattano ai dati.

Il test del rapporto di verosimiglianza

In questo test vengono confrontati due modelli probabilistici:

1) Il parametro di interesse (o i parametri) sono vincolati a essere uguali ai valori specificati dall'ipotesi nulla
2) Il parametro (o i parametri) assume il valore stimato che massimizza la verosimiglianza. Questo modello definisce l'ipotesi alternativa.

Se la verosimiglianza del secondo modello è significativamente più alta di quella del primo modello, allora rifiutiamo l'ipotesi nulla.

Il test del rapporto di verosimiglianza

La statistica test per il test del rapporto di verosimiglianza, G, è uguale a:

$$G = 2ln\left(\frac{L[valore\ di\ massima\ verosimiglianza\ del\ parametro|dati]}{L[valore\ del\ parametro\ sotto\ H_0|dati]}\right)$$

Se H_0 è vera, allora G ha approssimativamente una distribuzione χ^2 , ed è possibile dunque calcolare un P-value per decidere se rifiutare o meno l'ipotesi nulla.

Questa approssimazione è valida per campioni grandi, altrimenti è possibile usare la simulazione per trovare la distribuzione nulla di G.

Il numero di gradi di libertà per G è uguale alla differenza del numero di parametri stimati a partire dai dati nei due modelli.

Nel caso di un singolo parametro abbiamo sempre 1 gradi di libertà (0 parametri stimati in H_0; 1 parametro stimato in H_A)

Il test del rapporto di verosimiglianza - Esempio

Utilizziamo l'esempio delle vespe fatto in precedenza per capire come eseguire il test del rapporto di verosimiglianza.

In questo esempio il numero di vespe che sceglievano la femmina di cavolaia fecondata erano 23 su 32 mentre quelle che sceglievano la cavolaia vergine erano 9 su 32.

Questo risultato prova che le vespe nella popolazione preferiscono femmine di cavolaia fecondate?

Il test del rapporto di verosimiglianza - Esempio

Per rispondere alla domanda dobbiamo testare la seguente ipotesi nulla:

H0: Le vespe scelgono le femmine fecondate e quelle vergini con uguale probabilità (p=0.5)

HA: Le vespe preferiscono uno dei due tipi di femmina ($p \neq 0.5$)

Il test del rapporto di verosimiglianza - Esempio

La statistica test G è pari a:

$$G = 2ln\left(\frac{L[\hat{p}|Y\ vespe\ scelgono\ la\ femmina\ fecondata]}{L[p_0|Y\ scelgono\ la\ femmina\ fecondata]}\right)$$

dove $\hat{p}$ è la stima di massima verosimiglianza e p_0 è la proporzione secondo l'ipotesi nulla.

E' in generale più semplice lavorare con il logaritmo della verosimiglianza per cui possiamo scrivere:

$$G = 2(\ln L[\hat{p}|Y\ scelgono\ la\ femmina\ fecondata] - lnL[p_0|Y\ scelgono\ la\ femmina\ fecondata])$$

Il test del rapporto di verosimiglianza - Esempio

Sostituendo alla formula i dati otteniamo:

$Ln\ L[p\_0|23\ scelgono\ la\ femmina\ fecondata]=$
$ln[(32¦23)]+23ln[0.5]+(9)\ln[0.5]=-5.03125$

Il valore di verosimiglianza, associato alla stima di massima verosimiglianza p=0.72 stimata in precedenza, è pari a -1.863.

A questo punto possiamo calcolare la statistica test :
G=2(ln$[\![L[\hat{p}\,|Y\ scelgono\ la\ femmina\ fecondata]-lnL[p_0\,|Y\ scelgono\ la\ femmina\ fecondata]]\!]$)
=2(-1.863-(-5.031))=6.33

Se H_0 è vera, allora G segue una distribuzione χ^2 con 1 grado di libertà. Il valore critico della statistica con un α=0.05 è dunque uguale a 3.84.

Poiché G > 3.84 rifiutiamo l'ipotesi nulla: le vesoe preferiscono effettivamente le femmine di cavolaia fecondate rispetto alle femmine vergini.

Il test del rapporto di verosimiglianza – Esempio con R

#H0: Le vespe scelgono le femmine fecondate e quelle vergini con uguale probabilità (p=0.5)
#HA: Le vespe preferiscono uno dei due tipi di femmina (p≠0.5)

```
#stimo la log-verosimiglianza per p=0.5
n<-32
y<-23
p<-0.5
logLH0<-log(choose(n,y))+y*log(p)+(n-y)*log(1-p)
#effettuo la stima di massima likelihood
int_p<-seq(from=0.1, to=0.9, length.out=90)
logL<-log(choose(n,y))+y*log(int_p)+(n-y)*log(1-int_p)
logLmax<-max(logL)
stima<-int_p[logL==logLmax]
#calcolo la statistica test
G<-2*(logLmax-logLH0)
#calcolo chi-quadro critico con alpha=5% e 1 gdl
chicrit<-qchisq(0.05,df=1,lower.tail=F)
#calcolo il p-value
pchisq(G,df=1,lower.tail=F)
```

Un altro esempio dell'utilizzo della verosimiglianza

Esempio 20.4

Il conteggio degli elefanti

Stimare il numero di elefanti è molto più problematico di quanto si possa pensare, almeno quando essi vivono in foreste fitte e si alimentano di notte. Eggert et al. (2003) hanno usato il metodo di «cattura-ricattura» per stimare il numero totale di elefanti africani che vivono nelle foreste del Kakum National Park in Ghana, senza averne visto neppure uno. I ricercatori hanno trascorso circa due settimane nel parco raccogliendo feci di elefanti (vedi l'immagine a destra), dalle quali sono riusciti a estrarre il DNA degli animali. Usando 5 geni, i ricercatori hanno ottenuto un DNA fingerprint unico per ogni elefante «incontrato». Con questo metodo, hanno identificato 27 individui nei primi 7 giorni di campionamento. Chiamiamo «primo campione» questo campione e indichiamo come «marcati» questi 27 elefanti. Negli 8 giorni successivi i ricercatori hanno campionato sempre attraverso le feci 74 individui, 15 dei quali erano già inclusi nel primo campione (avevano un profilo genetico già riscontrato nel primo campione). Indichiamo come «ricatturati» questi 15 elefanti. Sulla base del numero di ricatture nel secondo campione, qual è la dimensione totale della popolazione di elefanti nel parco? ■

Un altro esempio dell'utilizzo della verosimiglianza

Per usare l'approccio della verosimiglianza dobbiamo conoscere la probabilità di ottenere il numero osservato di ricatture per differenti valori possibili di N (dimensione della popolazione di elefanti)

IL modello probabilistico in questo caso è conosciuto (Gazey & Stanley 1986). Se indichiamo con n1 il numero di individui catturati e marcati nei primi 7 giorni (nel nostro caso 27) e con n2 la dimensione del secondo campione di individui monitorati nel corso dei successivi 8 giorni (nel nostro caso 74 individui), allora la probabilità di Y ricatture data la dimensione di popolazione N è:

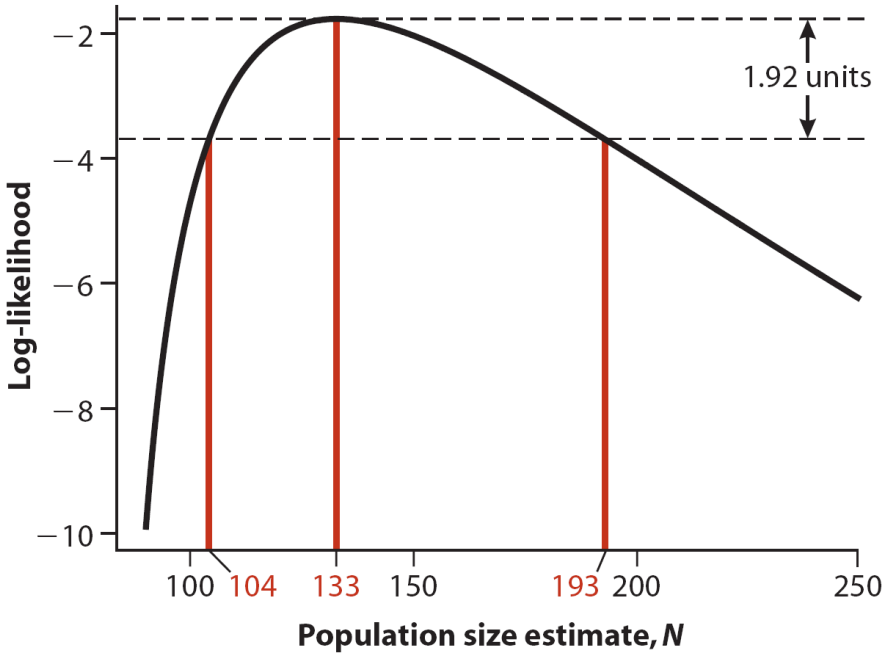
$$\Pr[numero\ di\ ricatture = Y|N] = \frac{\binom{n1}{Y}\binom{N-n1}{n2-Y}}{\binom{N}{n2}}$$

Un altro esempio dell'utilizzo della verosimiglianza

Utilizzando il logaritmo della verosimiglianza otteniamo che:

$$lnL[N|Y\ ricatture] = \ln[\binom{n1}{Y}]+\ln[\binom{N-n1}{n2-Y}]-\ln[\binom{N}{n2}]$$

Sostituendo Y=15, n1=27 e n2=74 e risolvendo la funzione nell'intervallo di N tra 90 e 250 otteniamo:



La stima di massima verosimiglianza della dimensione della popolazione di elefanti è 133 con un intervallo di confidenza al 95% tra 104 e 193 individui.

Un altro esempio dell'utilizzo della verosimiglianza con R

```
#salvo i dati di partenza in variabili
Y<-15
n1<-27
n2<-74
#imposto un intervallo di variazione di N dove calcolare la verosimiglianza
N<-90:250
#calcolo la log-verosimiglianza in tutti i valori nell'intervallo
logL<- log(choose(n1,Y))+log(choose(N-n1,n2-Y))-log(choose(N,n2))
#disegno la curva di verosimiglianza
matplot(N,logL,type="l")
#identifico il valore di N che massimizza la log-verosimiglianza
N[logL==max(logL)]
#calcolo l'intervallo di confidenza guardando i valori dei parametri ad almeno 1.92 unità di log-verosimiglianza dal massimo
difL<-logL-max(logL)
cbind(N,difL,difL<=-1.92)
#IC95%: 104-193
```