STATISTICA 1, metodi matematici e statistici

Introduzione al linguaggio R

Esercitazione3: 18-03-2005

Luca Monno

Università degli studi di Pavia

luca.monno@unipv.it
http://www.lucamonno.it

Intervalli di confidenza

Spesso in statistica si è intenzionati a conoscere non una stima puntuale di un certo parametro θ ma una stima per intervalli, chiamati intervalli di confidenza, infatti se θ è il parametro di una variabile aleatoria continua e $\hat{\theta}$ è un suo stimatore si ha che

$$P(\hat{\theta} = \theta) = 0.$$

Vediamo quindi un semplice esempio: generiamo M campioni di numerosità n da una normale con media 0 e varianza 1 e poniamo ogni campione in una matrice di M righe:

```
> M = 1000
> n = 100
> X = matrix(rnorm(M * n), nrow = M, ncol = n)
```

Avete visto a lezione che l'intervallo di confidenza di livello 1-2α per la media di una variabile aleatoria normale con varianza incognita è del tipo

$$(\hat{\mu} - t_{n-1}(1-\alpha) * s/\sqrt{n}, \hat{\mu} - t_{n-1}(1-\alpha) * s/\sqrt{n})$$

calcoliamo quindi le quantità in gioco:

```
> a = 0.05
> mu.hat = apply(X, MAR = 1, FUN = mean)
> sd.hat = apply(X, MAR = 1, FUN = sd)
> t = qt(1 - a, df = n - 1)
> interval = matrix(nrow = M, ncol = 2)
> interval[, 1] = mu.hat - t * sd.hat/sqrt(n)
> interval[, 2] = mu.hat + t * sd.hat/sqrt(n)
```

e facciamo un grafico di questi intervalli:

```
> plot(1:M, rep(mu, M), type = "l",
+     ylim = c(min(interval[, 1]), max(interval[, 2])))
> segments(x0 = 1:M, y0 = interval[, 1], x1 = 1:M, y1 = interva
+     2])
```

Vediamo che la quantità di intervalli che contengono il vero valore della media, 0 nel nostro caso, è abbastanza vicino al vero valore (1-2α):

```
> for (i in 1:M) ok[i] = (interval[i, 1] < mu) && (interval[i,
+     2] > mu)
> mean(ok)

[1] 0.896
```

Provate, per esercizio, a colorare gli intervalli che non comprendono lo 0.

Legame tra distribuzioni binomiale e Poisson

Sia R una v.a. binomiale con parametri m e π. Quando $m \to \infty$ e $\pi \to 0$ in modo tale che $m\pi \to \lambda$ allora

$$Pr(R = r) \to \frac{\lambda^r}{r!} e^{-\lambda}$$

Per una verifica numerica proviamo

```
>y<-0:10
>lambda<-1
>m<-10
>p<-lambda/m
>cbind(y,pbinom(y,size=m,prob=p),ppois(y,lambda))
>round(cbind(y,pbinom(y,size=m,prob=p),ppois(y,lambda)),
+digits=3)
```

Provate anche con altri valori di m e λ.

Confronto tra distribuzione t e normale

Creiamo ora varie funzioni che restituiscono la densità di una t di student con diversi gradi di liberta.

```
dt1<-function(x) dt(x,df=1)
dt5<-function(x) dt(x,df=5)
dt30<-function(x) dt(x,df=30)
```

Consideriamo ora l'istogramma del campione di osservazioni da una normale (generato precedentemente) modificando il range degli assi

```
hist(campione,prob=T,xlim=c(-5,5),ylim=c(0,0.6))
```

Andiamo ora a disegnare i grafici delle densità t proprio sopra l'istogramma

```
curve(dt1,from=-5,to=5,add=T,col=1)
curve(dt5,from=-5,to=5,add=T,col=2)
curve(dt30,from=-5,to=5,add=T,col=3)
```

Si può notare che la curva relativa alla densità t con 30 gradi di libertà interpola bene l'istogramma e che all'aumentare dei gradi di libertà l'area sottostante le code diventa sempre più piccola.

Legame tra la somma di v.a. esponenziali e la gamma

E' noto che la somma di n $v.a.$ esponenziali con densità $f_X(x) = \alpha e^{-\alpha x}$ si distribuisce come una gamma con densità $f_{S_n}(s) \propto e^{-\alpha x} x^{n-1}$.

Proviamo a verificarlo per $n = 10$ e $\alpha = 2$ costruendo 1000 realizzazioni da una $v.a.$ ottenuta come somma di 10 esponenziali con $\alpha = 2$ e confrontando, attraverso i quantili, la distribuzione empirica di queste 1000 realizzazioni con una distribuzione gamma con i parametri ipotizzati.

```
>se<-c()
>for (i in 1:1000 ) se[i]<-sum(rexp(10,2))
>quantile(se,c(0.05,0.1,0.25,0.5,0.75,0.9,0.95))
>qgamma(c(0.05,0.1,0.25,0.5,0.75,0.9,0.95),shape=10,rate=2)
>x<-qgamma(seq(0.01,0.99,0.01),shape=10,rate=2)
>y<-quantile(se,seq(0.01,0.99,0.01))
>plot(x,y)
>abline(c(0,1))
```

Analisi dei Dati III

Dopo aver visto come verificare con R alcuni risultati di teoria delle probabilità ritorniamo ad analizzare un data set reale.

```
>data(rivers)
```

Vogliamo costruire delle procedure grafiche per vedere se i dati in questione provengono da una particolare d.d.p.
Una possibilità è costruire un grafico i cui punti hanno come coordinate i quantili campionari e i quantili della d.d.p. dello stesso livello.
Tale grafico viene detto qqplot e se la d.d.p. è quella che ha generato i dati i punti dovrebbero disporsi come una retta

I quantili campionari sono

```
>qc<-sort(rivers)
```

Abbiamo quindi che la proporzione di osservazioni più piccole dell'i-esimo elemento di qc è $(i-1)/n$, ovvero qc[i] dovrebbe stimare il quantile di livello $(i-1)/n$ della vera d.d.p. che ha generato i dati.

Vediamo allora se il nostro campione proviene da un'esponenziale con media pari a quella campionaria

```
>n<-length(rivers)
>s<-(0:(n-1))/n
>qt<-qexp(s,rate=1/mean(rivers))
>plot(qt,qc)
```

In realtà il valore atteso della $i-esima$ statistica d'ordine di un campione è il quantile di livello $i/(n+1)$ della d.d.p. che ha generato il campione, per cui è più corretto considerare

```
>s<-(1:n)/(n+1)
>qt<-qexp(s,rate=1/mean(rivers))
>plot(qt,qc)
```

I quantili empirici elevati sono molto più alti di quelli teorici: la coda della distribuzione empirica è più pesante di quella esponeziale.

Vediamo se la situazione migliora ipotizzando per i logaritmi delle lunghezze dei fiumi una distribuzione normale con media uguale alla media campionaria dei logaritmi e varianza uguale alla varianza campionaria dei logaritmi

```
>par(mfrow=c(1,2))
>qlc<-sort(log(rivers))
>qlt<-qnorm(s,mean(log(rivers)),sd(log(rivers)))
>plot(qt,qc)
>plot(qlt,qlc)
```

L'adattamento dei dati di coda sembra migliorare, ma peggiora l'adattamento nella parte centrale dei dati.
L'informazione data nell'ultimo grafico si possono ottenere direttamente in R con

```
>qqnorm(log(rivers))
>qqline(log(rivers))
```

Verosimiglianza

Sia $y = (y_1, \ldots, y_n)$ un campione casuale estratto da una variabile esponenziale con media $1/\theta$, $Y \sim Exp(\theta)$. La densità di Y è

$$p(y; \theta) = \theta e^{-y\theta}, \quad y, \theta > 0,$$

quindi la verosimiglianza associata al campione y è

$$L(\theta) = L(\theta; y) = \theta^n e^{-\theta \sum_{i=1}^{n} y_i}.$$

In R possiamo scrivere la verosimiglianza nel modo seguente

```
> exp.lik <- function(theta, data) {
+     n <- length(data)
+     sum.data <- sum(data)
+     return(theta^n * exp(-theta * sum.data))
+ }
```

Fissiamo il generatore di numeri casuali (con `set.seed()`), generiamo un campione di numerosità 10 generato da un'esponenziale di parametro $\theta = 2$ e facciamo il grafico della verosimiglianza attraverso il comando `curve()`

```
> set.seed(1)
> data.exp <- rexp(10, rate = 2)
> curve(exp.lik(x, data = data.exp), from = 0.1, to = 10, xlab
+     ylab = "verosimiglianza")
```

Il punto di massimo della verosimiglianza si può trovare analiticamente tramite derivazione ed è pari a $\left(\sum_{i=1}^{n} y_i/n\right)^{-1}$.

```
> 1/mean(data.exp)

[1] 2.373543
```

Possiamo ricavare imprecisamente le coordinate del punto di massimo attivando il comando `locator(n=1)` e cliccando con il tasto sinistro del mouse sul punto più alto della curva

```
>locator(n=1)
```

Un modo più preciso per ottenere il massimo della verosimiglianza si basa sul comando `nlm` che esegue una minimizzazione numerica

```
> mexp.lik <- function(theta, data) -exp.lik(theta, data)
> punto.iniziale <- c(2)
> nlm(f = mexp.lik, p = punto.iniziale, data = data.exp)
```

Nel comando `nlm` il parametro `f` è la funzione che deve essere minimizzata (rispetto al suo primo argomento), il parametro `p` è il punto iniziale per effettuare la minimizzazione numerica; osserviamo inoltre che possono essere *passati* ulteriori argomenti specifici della funzione `f`. Infatti, nel nostro caso, i dati su cui calcolare la verosimiglianza vengono specificati all'interno del comando `nlm` con `data=data.exp`

ESERCIZI PER CASA

1 Si dimostra che se $Z \sim \mathcal{N}(0,1)$ e $W \sim \chi^2(\nu)$ sono indipendenti allora $T = Z/(W/\nu)^{1/2}$ si distribuisce come una T di Student con ν g.d.l.
Verificare tale risultato attraverso i seguenti passi

1. generate un campione di 1000 osservazioni da $\mathcal{N}(0,1)$ e un campione di 1000 osservazioni da $\chi^2(\nu)$
2. costruite sulla base dei campioni appena generati 1000 osservazioni da una T di student con ν g.d.l.
3. generate un campione di numerosità 1000 da una T direttamente con il comando `rt`
4. confrontate i campioni ottenuti nei punti 3 e 4 con il comando `qqplot`

2 A che cosa serve il comando `qqplot`?

3 Ripetere la verifica della legge debole dei grandi numeri e del teorema del limite centrale utilizzando il comando `apply` invece del ciclo `for`. Inoltre porre i grafici in un unica finestra grafica composta da 4 righe e due colonne (nella prima colonna porre gli istogrammi relativi alla verifica della legge debole)

4 Trovare i primi due momenti di una distribuzione gamma. Relativamente al data set rivers, ipotizzare che tali dati provengono da una distribuzione gamma con i primi due momenti uguali a quelli campionari e verificarne l'adattamento attraverso un qqplot.

5 Dato il campione $y = (3, 0, 1, 1, 1, 0, 2, 4)$, e supponendo che siano osservazioni iid da una v.a. esponenziale, disegnare il grafico della verosimiglianza e trovare analiticamente e attraverso il comando `nlm` il punto di massimo della verosimiglianza.