STATISTICA 1, metodi matematici e statistici

Introduzione al linguaggio R

Esercitazione2: 15-03-2004

Andrea Tancredi

Università di Roma “La Sapienza”, Rome, Italy

andrea.tancredi@uniroma1.it
http://3w.eco.uniroma1.it/utenti/tancredi

Analisi dei dati I

In R sono disponibili svariati insiemi di dati. Un elenco lo possiamo avere attraverso il comando

```
>data()
```

Tali dati vanno però caricati nel nostro workspace specificando, sempre attraverso il comando `data()`, quelli a cui siamo interessati

```
>data(chickwts)
```

Per vedere i dati basta digitare

```
>chickwts
```

Vediamo che le colonne hanno un nome: weight e feed. Possiamo vedere il loro contenuto selezionandole come in una matrice

```
>chickwts[,1]
```

oppure inserendole direttamente nel workspace

```
>attach(chickwts)
>weight
```

Informazioni base sul peso dei polli del campione si possono trovare con

```
>summary(weight)
```

L'istogramma ci permette di rappresentare graficamente la distribuzione del peso

```
>hist(weight,prob=T,nclass=20)
```

Per ottenere la funzione di ripartizione empirica possiamo richiamare la libreria stepfun e poi utilizzare il comando `ecdf`

```
>library(stepfun)
>plot(ecdf(weight),main="funzione di ripartizione")
```

Per vedere come varia la distribuzione del peso dei polli al variare della dieta utilizzata possiamo confrontare i boxplot realtivi al peso per le varie diete

```
>boxplot(weight feed)
```

Ma che cosa è il boxplot? Il box plot non è altro che il disegno di una scatola

- tagliata in due da una linea che è la mediana, Q_2;
- delimitata in alto e in basso dai quartili Q_3, Q_1;
- con dei baffi, (le linee orizzonatali esterne e più piccole) rappresentanti il minimo e il massimo se non ci sono punti all'esterno di $Q_1 - 1.5(Q_3 - Q1); Q_3 + 1.5(Q_3 - Q1)$;
- e con dei pallini rappresentanti i valori esterni al range $Q_1 - 1.5(Q_3 - Q1); Q_3 + 1.5(Q_3 - Q1)$, ovvero gli outliers; in questo caso i baffi coincidono con le ultime osservazioni prima degli outliers.

Simulazioni

Per molte delle distribuzioni di probabilità note (ad es. normale, esponenziale, Poisson, t di Student....), R può generare delle realizzazioni e calcolarne densità, funzione di ripartizione e quantili.

Per esempio, una realizzazione da una $\mathcal{N}(0,1)$ si ottiene con il comando `rnorm`

```
>rnorm(n=1,mean=0,sd=1)
```

Come abbiamo già detto, non è importante dare i nomi degli argomenti (però dobbiamo sempre ricordarci l'ordine, altrimenti possiamo trovarci nei guai)

```
>rnorm(1,-100,1)
>rnorm(1,1,-100)
```

Generiamo allora un campione di numerosità 1000 da una $\mathcal{N}(0,1)$ e ne calcoliamo media e varianza

```
>campione<-rnorm(n=1000,0,1)
>mean(campione)
>var(campione)
```

Sempre per la normale, per ottenere il valore della funzione di densità, della funzione di ripartizione e i quantili si usano i comandi `dnorm`, `pnorm`, `qnorm`.

Volendo lavorare con distribuzione diverse dobbiamo solo ricordarci del nome che la distribuzione ha per R . Ad esempio i comandi `rgamma`, `dgamma`, `pgamma`, `qgamma` ci daranno realizzazioni, densità, ripartizione e quantili di una gamma.

A questo punto basta ricordarsi i nomi utilizzati da R

`pois, binom, geom, unif, t, gamma, exp, chisq`

Legge debole dei grandi numeri

Sia $Y_1, \ldots, Y_n$ una successione i.i.d. di v.a. con media μ, allora $\bar{Y} = n^{-1}(Y_1 + \ldots + Y_n) \xrightarrow{p} \mu$.

Illustriamo la legge debole dei grandi numeri attraverso gli istogrammi di 10000 medie ottenute con altrettanti campioni esponenziali di numerosità 1,5,10,20.

```
>s<-c()
>for (i in 1:10000) s[i]<-mean(rexp(n=1,rate=1))
>hist(s,prob=T,xlim=c(0,4),ylim=c(0,2))
>for (i in 1:10000) s[i]<-mean(rexp(n=5,rate=1))
>hist(s,prob=T,xlim=c(0,4),ylim=c(0,2))
>for (i in 1:10000) s[i]<-mean(rexp(n=10,rate=1))
>hist(s,prob=T,xlim=c(0,4),ylim=c(0,2))
>for (i in 1:10000) s[i]<-mean(rexp(n=20,rate=1))
>hist(s,prob=T,xlim=c(0,4),ylim=c(0,2))
```

Teorema del limite centrale

Sia $Y_1, \ldots, Y_n$ una successione i.i.d. di v.a. con media finita μ e varianza finita σ^2, allora

$$Z_n = n^{1/2} \frac{(\bar{Y} - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$$

Verifichiamo il teorema nel caso in cui Y_i è esponenziale con media $\mu = 1$ (e di conseguenza varianza $\sigma^2 = 1$) con i seguenti comandi di R

```
>s<-c()
>n<-1
>for(i in 1:10000)
s[i]<-sqrt(n)*(mean(rexp(n,rate=1))-1)
>hist(s,prob=T,xlim=c(-3,3),ylim=c(0,1),main="n=1")
>curve(dnorm(x),from=-3,to=3,add=T)
```

e ripetendoli con $n = 5, 10, 20$.

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)
```

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 densita 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
>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))
```

Distribuzione del massimo di un campione esponenziale

Dato un campione di osservazioni $i.i.d$ provenienti da una distribuzione esponenziale $f_X(x) = e^{-x}$ vogliamo studiare il comportamento del massimo del campione. Consideriamo la seguente funzione

```
>sim<-function(n,M) matrix(rexp(n*M,1),nrow=M)
```

La funzione `sim` restituisce quindi una matrice con M campioni (messi nelle righe della matrice) $i.i.d.$ di numerosità n da una esponenziale di parametro 1

Modifichiamo ora la funzione in maniera tale da avere direttamente un vettore con i massimi degli n campioni

```
>sim.max<-function(n,M) {
>apply(matrix(rexp(n*M,1),nrow=M),FUN=max,MAR=1)}
```

La funzione `apply` applicata ad una matrice C permette di calcolare una funzione (l'argomento di `FUN`) ai vettori riga (se `MAR=1`) o colonna (se `MAR=2`) di C

Possiamo visualizzare le realizzazioni ottenute con il comando `plot`. Ad esempio per vedere i 500 massimi corrisspondenti a 500 campioni di numerosità 10 possiamo ricorrere a

```
>plot(sim.max(10,500))
```

Il comando `plot(x,y)` con x e y vettori di numerosità uguale (pari ad n produce il grafico con i punti $(x[i], y[i])$. Quando invece diamo un solo vettore v come argomento viene prodotto un grafico con i pinti $(i, v[i])$.

Vediamo ora come visualizzare la forma della distribuzione del massimo attraverso degli istogrammi.

```
>hist(sim.max(10,500)
>hist(sim.max(50,500)
```

I valori della seconda simulazione sono mediamente più grandi di quelli della prima. (Il massimo di 50 osservazioni è probabilmente più grande del massimo di 10 osservazioni). Possiamo vederlo meglio osservando i due grafici contemporeanamente.

```
>par(mfrow=c(1,2))
>hist(sim.max(10,500)
>hist(sim.max(50,500)
>par(mfrow=c(1,1))
```

E’ possibile dimostrare analiticamente che
$Z_n = \max(X_1, \ldots, X_n) - \log n$ converge in distribuzione ad una *v.a.* Z con funzione di densità

$$f_Z(z) = e^{-z} e^{-e^{-z}}.$$

(Tale distribuzione viene detta Gumbel)

Per verificare tale risultato simuliamo allora 5000 massimi da altrettanti campioni di 50 osservazioni esponenziali, sottraendo ad ogni massimo ottenuto $\log(50)$.

```
>massimi<-sim.max(50,5000)-log(50)
```

A questo punto calcoliamo una stima non parametrica della densità della *v.a.* che ha prodotto queste 5000 realizzazioni e ne facciamo il grafico.

```
>plot(density(massimi),type="l")
```

Scriviamo la funzione di densità della *v.a.* limite

```
>f<-function(x) exp(-x)*exp(-exp(-x))
```

e la disegniamo nel grafico precedente attraverso il comando `curve`

```
>curve(f,add=T,col=2)
```

Le due curve si sovrappongono quasi perfettamente, il che indica che con campioni esponenziali di numerosità 50, la distribuzione del massimo viene ben approssimata da una distribuzione $\log 50 + Z$ dove Z ha distribuzione Gumbel.

Analisi dei Dati II

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 disposrsi come una retta

I quantili campionari sono

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

Abbiamo quindi che la proporzione di osservazioni più piccole dell'iesimo 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<-(1:n)/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))
```

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.