STATISTICA 1, metodi matematici e statistici

Introduzione al linguaggio R

Esercitazione2: 04-03-2005

Luca Monno

Università degli studi di Pavia

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

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 (se la versione di R non è recente) 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.

Analisi dei dati II

Vedremo ora un'altro data-set relativo ai gol della serie A segnati a partire dal campionato 1996/1997.

Prima di iniziare salvate nella working directory il dataset `goal.dat` già utilizzato a lezione. Aprendo il file possiamo osservare che sulla prima riga ci sono dei caratteri corrispondenti al nome delle colonne (in questo caso gli anni dei campionati di calcio) mentre sulla prima colonna i nomi delle righe (ovvero le giornate della serie A).

Per importare un file di questo tipo si utilizza il comando `read.table`:

```
> goal <- read.table("goal.dat")
```

Possiamo calcolare la media e la varianza dei goal disputati ogni anno attraverso i comandi `mean` e `var`.

```
> mean(goal)
> var(goal)
```

Notiamo che in questo modo otteniamo la media e la varianza suddivisa per anno. Questo perché l'oggetto considerato è un data frame.

Per ottenere una semplice matrice utilizziamo il comando `as.matrix`, potrebbe essere utile anche il vettore di tutti i goal:

```
> goal.mx <- as.matrix(goal)
> goal.vec <- c(goal.mx)

> mean(goal.vec)
> var(goal.vec)
> sd(goal.vec)
```

Nel caso fossimo interessati ai quantili (empirici) del campione possiamo utilizzare due funzioni: `quantile` e `summary`:

```
> summary(goal.vec)
> quantile(goal.vec)
> quantile(goal.vec, 0.36)
```

Vediamo ora come si distribuiscono i dati tramite l'istogramma:

```
> hist(goal.vec)
```

Notiamo che l'istogramma ha il tipico andamento a campana caratteristico della distribuzione normale. Per verificare ciò standardizziamo il vettore `goal.vec` e sovrapponiamo all'istogramma del vettore standardizzato la densità di una normale (`dnorm`) con media 0 e varianza 1:

```
> z = (goal.vec - mean(goal.vec))/sd(goal.vec)
> hist(z, prob = T)
> curve(dnorm(x), add = T, col = 2)
```

Osserviamo che la curva approssima abbastanza bene l'istogramma. Un'altra verifica può essere fatta confrantando la funzione di ripartizione empirica (`ecdf`) con quella teorica (`pnorm`):

```
> plot(ecdf(z))
> curve(pnorm(x), add = T, col = 2)
```

Le funzioni di R che iniziano con una `d,p,q` seguite dal nome di una distribuzione (`norm,beta,gamma,...`) servono per calcolare rispettivamente la densità, la funzione di ripartizione, e i quantili di tali distribuzioni.

Le funzioni di R che iniziano con una `r` invece servono per generare un campione di variabili iid da quella distribuzione:

```
> dgamma(3, shape = 2, rate = 1)
> pnorm(0)
> qnorm(0.5)
> hist(rbeta(n = 1000, shape1 = 2, shape2 = 3))
```

Nelle prossime slides vedremo meglio come generare campioni aleatori.

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 $n = 5, 10, 20$.