INTRODUZIONE A R
Lezione 3

Silvia Bacci*e Silvia Pandolfi†

1 Creare grafici in R

R consente di realizzare, con semplicità, grafici di qualità professionale. Questi sono poi esportabili come file in numerosi formato direttamente copiabili per l'utilizzo in altri programmi. Per la realizzazione dei grafici si utilizzano delle funzioni specifiche nelle quali occorre specificare sia i dati da utilizzare sia eventuali parametri opzionali.

1.0.1 Istogrammi

La funzione di R per creare un istogramma di frequenza di un vettore numerico è `hist()`.

```
> dati.studenti = read.csv("DATI.csv", header=T, sep=\,")
> hist(dati.studenti$Altezza)

> #per specificare le classi da utilizzare
> hist(dati.studenti$Altezza,breaks=15)
```

Altre possibili opzioni sono:

```
> hist(dati.studenti$Altezza,xlab="altezza",ylab="freq",
+ main="Istogramma di Altezza studenti", col="green", label = TRUE)
```

1.0.2 Diagrammi a barre

Per le variabili di tipo qualitativo, una rappresentazione possibile è il grafico a barre. Il comando `barplot()` richiede come argomento un oggetto contenente le modalità della variabile e le relative frequenze.

```
> #partendo dal dataset dati-ita ricavare i diagrammi a barre
> #delle variabili sesso, zona, figli e genitori

> install.packages("gdata")
> library(gdata)
> dati.ita = read.xls("dati-ita.xls", sheet=1)
> head(dati.ita)
> attach(dati.ita)
```

*silvia.bacci@unipg.it
†pandolfi@stat.unipg.it

```
> par(mfrow=c(2,2)) #inserisce più grafici in una stessa finestra
> barplot(table(sesso),col=c("pink","blue"),main="barplot di Sesso")
> barplot(table(zona),col=c("green","gray"),main="barplot di zona")
> barplot(table(figli),col=c("red","green","black","yellow"),main="barplot di figli")
> barplot(table(genitori),main="barplot di genitori")
```

1.0.3 Diagrammi a torta

Un'altra rappresentazione classica di distribuzioni di frequenza è il diagramma a torta:

```
> pie(table(dati.studenti$Diploma))
> pie(table(dati.studenti$Diploma),col=1:4)
> pie(table(dati.studenti$Diploma),col=c("pink","red","yellow","green"),
>+main="diagramma a torta di diploma")
```

1.0.4 Diagrammi a scatola (Box-plot)

I diagrammi a scatola consentono di visualizzare i principali indici di posizione di una variabile. In particolare, il grafico descrive la tendenza centrale, tramite la mediana, la dispersione dei dati, tramite i quartili e l'estensione complessiva del campione. In R è possibile costruire anche boxplot di una variabile quantitativa, condizionatamente alle modalità di un variabile qualitativa.

```
> boxplot(reddito)
> boxplot(reddito ~ sesso)
> boxplot(reddito ~ figli)
```

1.0.5 La funzione generica `plot()`

La funzione `plot()` è una funzione generica che produce un grafico adeguato al tipo di oggetto a cui viene applicato.

- `plot(x)`: se `x` è un vettore numerico si produce un grafico dei valori ordinati rispetto all'indice di posizione
- `plot(x,y)`: si produce il grafico a dispersione (o scatterplot) dei due vettori numerici
- `plot(f)`: se `f` è una variabile qualitativa si produce il grafico a barre
- `plot(f,x)`: se `f` è una variabile qualitativa (fattore) si produce il boxplot di `x` per ogni livello di `f`

Carichiamo il dataset `iris` e facciamo alcuni esempi:

```
> data(iris)
> str(iris)
> x = iris$Petal.Length
> y = iris$Sepal.Length
> f = iris$Species
> plot(x)
```

```
> plot(x,y)
> plot(f)
> plot(f, y)
```

In questa funzione, ma in generale in tutte quelle per eseguire rappresentazioni grafiche dei dati, è presente una molteplicità di parametri grafici che permettono una notevole personalizzazione dell'output. La maggior parte sono visualizzabili dall'help di `plot` o di `par`, mentre altri sono specifici della singola tipologia di grafico. I più diffusi sono:

- `type`: argomento di plot che definisce la tipologia di rappresentazione: punti, linee etc.
- `xlim, ylim`: definiscono gli estremi degli assi
- `xlab, ylab`: assegnano delle etichette agli assi
- `main, sub`: titolo e sottotitolo del grafico
- `col`: specifica i colori da utilizzare nel grafico
- `lty`: tipo di linea del grafico: tratteggiata, continua,con punti etc.
- `lwd`: gestisce lo spessore delle linee e degli assi

1.0.6 Comandi grafici

In R è possibile utilizzare alcuni comandi per aggiungere informazioni o modificare un grafico esistente. I più comuni sono:

```
points(x, y) # aggiunge un punto nelle coordinate x y
lines(x, y) # connette linee
text(x, y, labels, ...) # aggiunge una etichetta al punto di coordinate x y
abline(a, b) # aggiunge una retta di pendenza b e intercetta a
polygon(x, y, ...) # disegna un poligono
legend(x, y, legend, ...) # aggiunge una legenda
title(main, sub) # aggiunge un titolo
axis(side, ...) # aggiunge un asse
```

1.0.7 Esercizio

Applicare le principali statistiche descrittive al dataset var-qualitative.txt e costruire i grafici appropriati.

2 Distribuzioni statistiche

Per molte delle distribuzioni statistiche usate comunemente in R è possibile calcolare la funzione di densità, o di probabilità nel caso discreto, la funzione di ripartizione o di probabilità cumulata, i quantili, e generare campioni di una data ampiezza:

- Normale
  - `dnorm(x, mean=0, sd=1)`: calcola densità in x
  - `pnorm(q, mean=0, sd=1)`: calcola probabilità cumulata fino a q
  - `qnorm(p, mean=0, sd=1)`: calcola quantile corrispondente alla probabilità p
  - `rnorm(n, mean=0, sd=1)`: genera un campione di dimensione n

    Se i parametri opzionali non sono impostati si ottiene una distribuzione normale standardizzata, altrimenti `mean` corrisponde alla media e `sd` alla deviazione standard.

- Chi-quadrato
  - `dchisq(x, df)`: calcola densità in x
  - `pchisq(q, df)`: calcola probabilità cumulata fino a q
  - `qchisq(p, df)`: calcola quantile corrispondente alla probabilità p
  - `rchisq(n, df)`: genera un campione di dimensione n

    Il parametro `df` corrisponde ai gradi di libertà della variabile chi-quadrato.

- Binomiale
  - `dbinom(x, size, prob)`: calcola densità in x
  - `pbinom(q, size, prob)`: calcola probabilità cumulata fino a q
  - `qbinom(p, size, prob)`: calcola quantile corrispondente alla probabilità p
  - `rbinom(n, size, prob)`: genera un campione di dimensione n

    I parametri `size` e `prob` si riferiscono al numero di prove e alla probabilità di successo di ciascuna prova. La variabile casuale di Bernoulli è ottenuta come caso particolare specificando `size=1`

- Poisson
  - `dpois(x, lambda)`: calcola densità in x
  - `ppois(q, lambda)`: calcola probabilità cumulata fino a q
  - `qpois(p, lambda)`: calcola quantile corrispondente alla probabilità p
  - `rpois(n, lambda)`: genera un campione di dimensione n

2.1 Esempi

2.1.1 Variabile Casuale Normale

Si calcolino i seguenti quantili della V.C. Normale Standard:

- $p(Z < z) = 0.95$

```
> qnorm(0.95, mean = 0, sd = 1)  # P(X<x)=0.95
[1] 1.644854
> qnorm(0.95)
[1] 1.644854
```

- $p(Z > z) = 0.07$

```
> qnorm(1-0.07,0,1)  # P(X>x)=0.07
[1] 1.475791
> qnorm(0.07,0,1, lower.tail = FALSE)
[1] 1.475791
```

- $p(Z < z) = 0.10$

```
> qnorm(0.10,0,1, lower.tail = TRUE)
[1] -1.281552
> -qnorm(1-0.10,0,1)
[1] -1.281552
> -qnorm(0.10,0,1,FALSE)
[1] -1.281552
```

Si consideri adesso una V.C. Normale con media 5 e varianza 9 e si calcolino:

- $p(X < x) = 0.95$

```
> qnorm(0.95, mean = 5, sd = sqrt(9))
[1] 9.934561
```

- $p(X > x) = 0.07$

```
> qnorm(0.07,5,sqrt(9),FALSE)
[1] 9.427373
```

Con riferimento ad una V.C. Normale Standard si calcolino le seguenti probabilità:

- $\phi(1.64)$:

```
> dnorm(x=1.64, mean=0, sd=1)
[1] 0.1039611
```

- $\Phi(2.3)$:

```
> pnorm(q = 2.3, mean = 0, sd = 1, lower.tail = TRUE)
[1] 0.9892759
```

- $p(Z > 1.5)$:

```
> pnorm(q = 1.5, mean = 0, sd = 1, lower.tail = FALSE)
[1] 0.0668072
> 1-pnorm(q = 1.5, mean = 0, sd = 1)
[1] 0.0668072
```

- $\Phi(-2.7)$:

```
> pnorm(q = -2.7, mean = 0, sd = 1)
[1] 0.003466974
```

- $p(Z > -1)$:

```
> pnorm(q = -1.0, mean = 0, sd = 1, FALSE)
[1] 0.8413447
```

- $p(-2.4 < Z < 3.2)$:

```
pnorm(q = 3.2, mean = 0, sd = 1)  - pnorm(q = -2.4, mean = 0, sd = 1)
[1] 0.9911153
```

Rappresentare graficamente la densità di una V.C. Normale Standard:

```
> x = sort(rnorm(1000)) # si generano 1000 num. casuali da una Normale Standard
                        # e si ordinano in senso non decrescente
> f = dnorm(x)          # si calcolano le corrispondenti densità
> plot(x, f, type="l")
> rug(x) # aggiunge delle linee verticali in corrispondenza dei valori osservati
```

Si provi a ripetere l'esperimento cambiando la dimensione del campione. Si generino anche osservazioni da una distribuzione normale non standardizzata (suggerimento: vedi `help(rnorm)`).

Supponiamo di voler disegnare il grafico della distribuzione Normale con media uguale e diversa varianza o varianza uguale e diversa media

```
> curve(dnorm(x,0,0.5),-6,6)
> curve(dnorm(x,0,1),-6,6,add=TRUE)
> curve(dnorm(x,2,1),-6,6,add=TRUE)
```

2.1.2 Approssimazione di una v.c. Chi-quadrato con una Normale

Ricordando che se $y \sim \chi^2_k$ allora $E(y) = k$ e $Var(y) = 2k$, possiamo rappresentare graficamente una v.c. χ^2 e, contemporaneamente, una v.c. normale con stessa media e varianza:

```
> y = sort(rchisq(1000,10)) #genera 1000 num.casuali da una chi-quadro con 10 g.d.l.
> plot(y,dchisq(y,10), type="l")    #disegna la funzione di densità
> x <- seq(10-4*sqrt(20),10+4*sqrt(20),by=1) #crea una sequenza di valori
> lines(x,dnorm(x,10,sqrt(20)),col=2,lty=2) #aggiunge la funzione di densità normale
```

2.1.3 Distribuzione Binomiale

Con riferimento ad una V.C. Binomiale con $n = 20$ e $p = 0.30$ si calcolino le seguenti probabilità:

- $p(X = 6)$

```
> dbinom(x = 6, size = 20, prob = 0.30)
[1] 0.191639
```

- $p(X \leqslant 5)$

```
> pbinom(q = 5, size = 20, prob = 0.30, TRUE)
[1] 0.4163708
```

- $p(X \geqslant 13)$

```
> pbinom(q = 13, size = 20, prob = 0.30, FALSE)
[1] 0.000261047
```

- $p(X > 13)$

```
> pbinom(q = 12, size = 20, prob = 0.30, FALSE)
[1] 0.00127888
```

- $p(2 \leqslant X \leqslant 7)$

```
> pbinom(q = 7, size = 20, prob = 0.30)- pbinom(q = 2, size = 20, prob = 0.30)
[1] 0.7367887
```

Si rappresenti graficamente la distribuzione di probabilità di una Binomiale con $n = 10$ prove indipendenti e $p = 0.5$.

```
> k = c(0:10)
> k
 [1]  0  1  2  3  4  5  6  7  8  9 10
>  p= dbinom(k,10,0.5)
> p
 [1] 0.0009765625 0.0097656250 0.0439453125 0.1171875000 0.2050781250
 [6] 0.2460937500 0.2050781250 0.1171875000 0.0439453125 0.0097656250
[11] 0.0009765625
> plot(k,p,type="h")
```

Al crescere del numero di prove la distribuzione Binomiale tende a quella Normale

```
> n = c(5, 10, 50, 100, 1000, 10000)
> x = matrix(0, nrow = 1000, ncol = 6)
> for (i in 1:6)
 {
  x[,i] <- rbinom(1000, size = n[i], prob = 0.8)
```

```
}
> 0.8*n
[1]    4    8   40   80  800 8000
> colMeans(x)
[1]    3.990    7.963   40.134   79.938  799.991 8001.107
```

Provare la seguente funzione:

```
> p=0.8
> media = n*p
> varianza=n*p*(1-p)
> par(mfrow=c(2,3))
> for(i in 1:6){
hist(x[,i],freq=F)
curve(dnorm(x,media[i],sqrt(varianza[i])),col="red",add=TRUE)
}
```

2.1.4 Esercizi

1. Con riferimento ad una V.C. V del tipo Chi quadrato con 5 gradi di libertà, calcolare:
   (a) il quantile v tale che $p(V < v) = 0.95$;
   (b) il quantile v tale che $p(V > v) = 0.07$;
   (c) il quantile v tale che $p(V < v) = 0.10$;
   (d) $p(V > 3)$;
   (e) $p(V < 1) \cup p(V > 10)$.
2. Con riferimento ad una V.C. T del tipo t di Student con 12 gradi di libertà, calcolare:
   (a) $p(T < 2.3)$;
   (b) $p(T > 1.5)$;
   (c) $p(T < -2.7)$;
   (d) $p(T > -1)$;
   (e) $p(-2.4 < T < 3.2)$;
   (f) il quantile t tale che $p(T < t) = 0.80$;
   (g) il quantile t tale che $p(T > t) = 2.20$.
3. Con riferimento ad una V.C. Y di tipo Poisson con parametro caratteristico uguale a 8, calcolare:
   (a) $p(Y = 6)$;
   (b) $p(Y \leqslant 5)$;
   (c) $p(Y \geqslant 13)$;
   (d) $p(Y > 13)$;
   (e) $p(2 \leqslant Y \leqslant 7)$.