Introduzione ad R
Versione 6

Roberto Boggiani

24 ottobre 2004

1 Introduzione

1.1 Introduzione ad R

R è un ambiente statistico per la manipolazione, l'analisi e la rappresentazione grafica dei dati. E' un ambiente interattivo, ossia i comandi producono una risposta immediata, e prevedono una programmazione orientata agli oggetti. Avviato il programma R, apparirà una nuova finestra con ilsimbolo
>
che indica che l'ambiente è pronto per ricevere delle istruzioni.

1.2 Help

L'ambiente R dispone di un help in linea molto efficiente che consente di ottenere delle informazioni sulle funzioni in esso implementate. Vi sono tre diversi sitemi di help:

- help in formato windows si ottiene con la seguente sequenza di comandi
  > options(winhelp=T)
  > help()
- help in formato html compilato si ottiene con
  > options(chmhelp=T)
  > help()
- help in formato html si ottiene con
  > help.start()

In ogni caso se si vogliono delle informazioni su un particolare comando basterà utilizzare la sintassi
> ?comando
comparirà la descrizione del comando e la sua sintassi d'uso. Provare per esempio a digitare
> ?mean
Nel caso in cui si voglia la ricerca di un help relativo ad una parola chiave possiamo utilizzare il comando:
`>help.search`("parolachiave")
in questo modo compariranno tutti i comandi in cui la parola chiave è contenuta.

1.3 Gli esempi e i demo

Ogni help relativo ad un comando di R termina con degli esempi che illustrano l'utilizzo del comando stesso. Per richiamare solamente gli esempi relativi ad un determinato comando utilizziamo la sintassi:
`>example(nomecomando)`

Per vedere alcune delle potenzialità di R può anche essere utile digitare il comando:
`>demo()`
che permette di visualizzare una serie di dimostrazioni su alcuni aspetti particolari di R.

1.4 Personalizzazione

Il programma R permette anche di effettuare una personalizzazione di certe opzioni come il browser in cui sarà visualizzato l'help, le cifre decimali da visualizzare, come deve apparire come prompt dei comandi. Una lista completa di tali personalizzazioni si può vedere dando in R il comando:

`>?options`
Se vogliamo utilizzare nel corso della sessione avviata i numeri con 4 cifre decimali dobbiamo dare il comando:
`>options( digits=4)`
se desideriamo invece che in ogni avvio di R i numeri siano sempre visualizzati con 4 cifre decimali dobbiamo modificare la variabile di avvio .First inserendo in essa il comando appena digitato. Ciò e fatto con:
`>.First<-fix(.First)`
ed inserendo tra le parentesi graffe la riga:
`>options( digits=4)`
Tale metodo resta valido per ogni opzione che puo' essere introdotta con il comando options, così se vogliamo modificare cifre ed editor basterà date il comando:
`>options( digits=4,editor=`"myeditor")
o inserirlo all'interno della varibile .First a seconda dei casi.

1.5 Fissare il numero delle cifre da visualizzare

Per fissare il numero di cifre da visualizzare si usa il seguente comando:
`>option(digits=n)`
in questo modo tutte le cifre successivamente ottenute saranno visualizzate con n cifre decimali. Se si vuole mantenere fisso questo numero in ogni sessione di avvio del programma si veda quanto scritto nel paragrafo 1.4.

2 Ia gestione dei pacchetti in R

2.1 I pacchetti in R

Il programma presenta una serie di pacchetti aggiuntivi che ne aumentano notevolmente la potenzialità. I pacchetti che possono essere utilizzati possono essere di tre tipologie diverse:

- pacchetti automaticamente installati e automaticamente avviati
- pacchetti automaticamente installati ma non avviati
- pacchetti reperibili in rete

mentre i pacchetti del primo tipo non presentano problemi, in quanto disponibili e utilizzabili sempre all'avvio del programma, i pacchetti degli altri due tipi richiedono alcune considerazioni circa il loro uso.

2.2 Pacchetti automaticamente installati ma non avviati

Quando il programma è installato oltre alla sua configurazione base sono automaticamente installati una serie di pacchetti aggiuntivi che sono molto utili nelle analisi statistiche. L'elenco di tali pacchetti si ottiene con il comando:
`>library()`
Alcuni di questi pacchetti non sono immediatamente utilizzabili ma devono essere caricati in R con il comando:
`>library`("nomepacchetto")
se usiamo il sistema operativo linux e con l'uso dell'apposito menu:
`>Packages - load packages`
se usiamo il sistema operativo windows.

Per far avviare i automaticamente i pacchetti dobbiamo con un editor di testo modificare il file `.First` inserendo in esso con una opportuna sintassi il nome dei pacchetti che vogliamo caricati automaticamente all'avvio di R. Ad esempio il seguente file `.First` così composto:

```
function() {
require("ctest", quietly = TRUE)
require("grid", quietly = TRUE)
require("lattice", quietly = TRUE)
require("stepfun", quietly = TRUE)
require("nls", quietly = TRUE)
```

```
require("nlme", quietly = TRUE)
require("mvtnorm", quietly = TRUE)
require("multcomp", quietly = TRUE)
require("scatterplot3d", quietly = TRUE)
require("mva", quietly = TRUE)
require("cluster", quietly = TRUE)
require("MASS", quietly = TRUE)
require("mgcv", quietly = TRUE)
options(editor="kwrite",browser="mozilla")}
```

permette di risolvere innumerevoli problemi legati all'analisi statistica dei dati. Si ricordi in ogni caso che i pacchetti `mvtnorm`, `multcomp`,`scatterplot3d` sono da installare come pacchetti aggiuntivi come precisato nel paragrafo successivo.

2.3 Pacchetti reperibili in rete

In rete possiamo trovare molti pacchetti che permettono di svolgere molteplici analisi statistiche dei dati. Per caricare tali pacchetti se usiamo il sistema operativo linux dobbiamo digitare da shell il seguente comando:
`R CMD INSTALL` "percorso/pacchetto"
se invece usiamo il sistema operativo windows dobbiamo fare riferimento alla voce di menu apposita nel seguente modo:

- prima di tutto aggiungere il pacchetto ad R e ciò viene fatto con la sequenza di comandi Packages - install package from local zip file - selezionare la directory dove è memorizzato il file - doppio click
- caricare il pacchetto in R è ciò avviene con la sequenza di comandi Packages - load packages e scegliere il pacchetto da installare dalla lista presentata

Vedremo in un capitolo successivo come creare propri pacchetti aggiungibili ad R.

2.4 Funzioni e script in codice sorgente

Un metodo molto pratico per aggiungere funzioni e script appositamente creati è quello che consiste nella creazione di appositi file detti source nei quali vengono inserite più funzioni scritte in linguaggio R come vedremo nel paragrafo 20. Per caricare tali funzioni basterà digitare la sequenza file - source e indicare il percorso e il nome del file dove il pacchetto codice è memorizzato. In questo modo possiamo facilmente caricare funzioni personali create ad hoc per l'analisi di dati.

3 L'uso delle directory e il salvataggio dei dati

3.1 Introduzione

L'uso delle directory in R è di fondamentale importanza al fine di poter salvare e utilizzare successivamente il lavoro svolto. Ogni sessione di R crea due file `RData` ed `Rhistory` che memorizzano i dati ed i comandi che sono stati utilizzati nella sessione. Esistono alcuni comandi che permettono di gestire in modo appropriato le directory che sono:

- `setwd(mydir)` che permette di spostare la directory di lavoro corrente in mydir
- `getwd()` che visualizza la directory di lavoro corrente
- `dir()` che visualizza il contenuto della directory di lavoror corrente

3.2 Il salvataggio del workspace

Il workspace di R è dato dal file `RData` e `Rhistory` uscendo da R con il comando `q()` e dando l'opzione `y`, i file `RData` ed `Rhistory` saranno memorizzati sempre nella directory di lavoro corrente. Se vogliamo però memorizzare il file `RData` in altra directory possiamo o modificare la directory di lavoro corrente come visto nel paragrafo precedente o utilizzare i seguenti comandi:

- `save.image(file=mydirectory/.RData)` che permette di salvare il file `RData` in mydirectory

- `save.image()` che permette di salvare il file `RData` nella directory corrente
- `save.image(file=mydirectory/mynome)` che permette di salvare il file `RData` in mydirectory ma con un nome diverso
- `save.image(file=mynome)` che permette di salvare il file `RData` nella directory corrente ma con un nome diverso

Per caricare un workspace precedentemente salvato possiamo usare il comando
`sys.load.image(mydirectory/mynome,quiet=F)`
che aggiungerà le varibili del file mynome a quelle del file `RData` presente nella directory corrente.

3.3 Il salvataggio dell'history

L'hystory di R è dato da tutti i comandi digitati in una sessione e nelle precedenti se abbiamo effettuato regolarmente il salvataggio del workspace uscendo da R. La history è memorizzata nel file `Rhistory` automaticamente generato o aggiornato uscendo da R nella directory di lavoro corrente. Indipendentemente da ciò possiamo anche salvare l'history in uno dei seguenti modi:

- `savehistory()` salva l'history nel file `Rhistory` nella directory di lavoro corrente
- `savehistory(file=mydirectory/myfile)` salva l'history nel file myfile della directory mydirectory

Per caricare l'history precedentemente salvata si può usare uno dei seguenti comandi:

- `loadhistory(file=mydirectory/myfile)` carica l'history memorizzata in mydirectory e in myfile
- `loadhistory(file=myfile)` carica l'history memorizzata nel file myfile della directory corrente

Si noti che sarà anche possibile visualizzare l'history memorizzata con uno dei seguenti comandi:

- `history()` visualizza gli ultimi 25 comandi
- `history(n)` visualizza gli ultimi n comandi

3.4 Salvataggio delle variabili in file di dati

Possiamo anche salvare le variabili presenti nel workspace in file al fine di poterli archiviare o richiamare all'occorenza. Ciò può essere effettuato con uno dei seguenti comandi:

- `save(nomevariabile,file=mydirectory/myfile)` salva la variabile nomevariabile in myfile della mydirectory
- `save(list=ls(),file=mydirectory/myfile)` salva tutte le variabili del workspace in myfile della mydirectory

Per aggiungere le varibili precedentemente memorizzate in tali file al workspace corrente si utilizza il comando:
`load(mydirectory/myfile`
che aggiunge appunto le varibili presenti in myfile al workspace corrente.

3.5 Salvataggio dell'output di R

E' possibile salvare l'output prodotto da R in una particolare sessione in modo da poterlo esportare in un formato di testo nel seguente modo:

- `sink(myoutput.txt`
- comandi vari
- `sink()`

Tutto l'output tra i due comandi `sink` non viene visualizzato sullo schermo ma reindirizzato sul file `myoutput.txt`.

4 Gli operatori

4.1 Introduzione

Prima di proseguire nell'analisi dei principali comandi del programma R, sarà opportuno conoscere i principali operatori che possono essere utilizzati in tale programma. Tale analisi sarà svolta nei paragrafi successivi.

4.2 Operatori aritmetici

Gli operatori aritmetici sono gli operatori fondamentali che servono per svolgere le principali operazioni matematiche. Essi sono:

- + addizione
- − sottrazione
- ∗ moltiplicazione
- / divisione
- ^ elevamento a potenza

Il loro uso è analogo a quello che comunemente si fa su una calcolatrice ad esempio:

- `> (7+2)*3`
- `> 7^2`

4.3 Gli operatori relazionali

Gli operatori relazionali sono gli operatori fondamentali per effettuare confronti tra numeri o lettere. Essi sono:

- minore: `<`
- maggiore: `>`
- minore o uguale: `<=`
- maggiore o uguale: `>=`
- uguale: `==`
- diverso: `!=`

Un esempio del loro uso può essere il seguente:

```
> 4>2
>c(1,5,3)<c(3,1,0)
```

Si noti che tali operatori possono essere applicati anche ad un vettore o una matrice restituendo in questo caso o un vettore logico o una matrice logica ossia vettori o matrici formati solamente da True o False.

4.4 Gli operatori logici

Gli operatori logici sono gli operatori fondamentali che servono generalmente per collegare tra di loro espressioni contenenti operatori relazionali. Essi sono:

- and: `&`
- or: `|`

il valore da loro restituito segue le normali regole delle tabelle di verità relative a tali operatori. Un esempio del loro uso può essere il seguente:

```
> 1>2 & 2<3
```

tali operatori possono anche essere applicati ad un vettore o una matrice è restituiscono un vettore o una matrice logica ossia vettori o matrici formati solamente da True o False.

Possiamo anche applicare gli operatori logici complessivi restituiscono un unico valore. Essi sono:

- and complessivo `&&`
- or complessivo `||`

tali operatori se applicati ad un vettore o una matrice restituiscono solamente un valore di True o False.

La differenza tra gli operatori `& |` e `&& ||` è che mentre la prima coppia esegue un controllo logico termine a termine tra gli elementi di due vettori, la seconda coppia esegue un controllo in sequenza da sinistra a destra e si arresta fornendo il primo risultato valido.

Si noti inoltre che:

- `NA & TRUE` da come risultato `NA`
- `NA & FALSE` da come risultato `FALSE`

5 Le funzioni matematiche elementari e i numeri complessi

5.1 Le funzioni elementari matematiche

Il programma mette a disposizione molte funzioni matematiche tra cui segnaliamo:

- sqrt() radice quadrata
- abs() valore assoluto
- sin() cos() tan() funzioni trigonometriche
- asin(), acos(), atan() funzioni trigonometriche inverse
- sinh(), cosh() tanh() funzioni iperboliche
- asinh(), acosh(), atanh() funzioni iperboliche inverse
- exp() la funzione e
- log() logaritmo base e
- log10() logaritmo in base 10
- ceiling() arrotonda all'intero più alto
- floor() arrotonda all'intero più basso
- trunc() tronca la parte decimale
- round(x,n) arrotonda al numero di cifre specificato con n
- signif(x,n) arrotonda al numero di cifre significative specificate con n
- pi restituisce il valore di pi greco

5.2 I numeri complessi

I numeri complessi sono utilizzati nel seguente modo:

- `>a+ib` restituisce un numero complesso
- `>a+1i` restituisce un numero complesso con parte immaginaria avente coefficiente reale 1
- `>Re(a+ib)` restituisce la parte reale del numero complesso
- `>Im(a+ib)` restituisce la parte immaginaria del numero complesso
- `>Mod(a+ib)` restituisce il modulo della forma polare del numero complesso
- `>Arg(a+ib)` restituisce l'argomento della forma polare del numero complesso
- `>Conj(a+ib)` restituisce il coniugato di un numero complesso

I calcoli con i numeri complessi posso essere eseguiti normalmente con i comuni operatori aritmetici visti sopra.

6 Le variabili

6.1 Assegnazione di un valore ad una variabile

Per assegnare un valore numerico alla variabile x si usa il comando:
`> x<-2`
mentre per assegnare un carattere alla variabile y si usa il comando:
`> x<-`“casa“
Assegnando un nuovo valore ad una variabile, verrà automaticamente cancellato il valore precedentemente assunto dalla stessa.

Un altro modo per assegnare un valore ad una variabile è quello evidenziato dai seguenti esempi:

- `>assign(`“x“,10)
- `>assign(`“x“,“casa“)

6.2 Visualizzare il contenuto di una variabile

Per visualizzare il contenuto di una variabile basterà digitare il nome della variabile stessa: `> x`
`>y`

6.3 Visualizzazione di tutte le variabili esistenti

Per visualizzare il nome di tutte le variabili esistenti nell’area di lavoro si usa uno dei seguenti comandi:
`>objects()`
`>ls()`

6.4 Cancellazione di variabili

Per cancellare una variabile di nome `x` basterà digitare il comando:
`> rm(x)` Per calcellare tutte le variabili esistenti nell’area di lavoro comprese le funzioni create o aggiunte si usa invece il comando:
`>rm(list=ls())`
attenzione in questo caso ad essere sicuri di quello che si sta facendo.

7 Oggetti base di R

7.1 Gli oggetti base

Gli oggetti base usati dal programma R sono i seguenti:

- vettori
- matrici
- array
- liste
- fattori
- serie storiche
- data frame

I prossimi paragrafi saranno dedicati allo studio di tali oggetti.

7.2 Gli attributi degli oggetti base

Tutti gli oggetti base di R possiedono degli attributi che permettono loro di essere più flessibili ed utilizzabili in modo più veloce. Nella singola trattazione degli oggetti verrano indicati in dettaglio oggetto per oggetto gli attributi posseduti.

8 I vettori

8.1 Introduzione

I vettori sono dati dello stesso tipo che sono raggruppati in una unica variabile. Essi sono molto importanti nell'analisi statistica dei dati in quanto dati raggruppati in vettori sono generalmente facili da essere studiati e manipolati.

8.2 La funzione *c*

Con il comando `c(elemento1,elemento2,.....)` viene creato un vettore tramite la concatenazione degli elementi specificati tra parentesi e separati da virgole. Gli elementi non possono essere contemporaneamente caratteri e numeri. Ad esempio con:
`>x<-c(1.5,2,2.5)`
viene creato un vettore numerico che contiene i valori specificati.

Sempre tramite in comando `c()` si possono aggiungere nuovi elementi ad un vettore precedentemente creato. Ad esempio con:
`>x<-c(x,3)`
viene aggiunto alla fine del vettore `x` il numero 3.

Per ottenere un vettore di stringhe sarà sufficiente inserire ogni stringa tra apici, come nel seguente esempio:
`>x<-c`("questo","è","un esempio")

8.3 Tipologie di vettori

I vettori in R possono essere di varie tipologie:

- numerico
- logico
- complesso
- caratteri

a secondo della tipologia di elementi che li compongono.

8.4 Come creare un vettore

Oltre alla funzione `c` per creare un vettore possiamo utilizzare uno dei seguenti metodi:

- `>x<-scan()` crea un vettore numerico
- `>x<-scan(what=`"character") crea un vettore di caratteri
- `>x<-scan(what=`"complex") crea un vettore di numeri complessi
- `x<-n1:n2` crea un vettore composto da numeri che vanno da `n1` a `n2` con passo pari ad `uno`

il vantaggio di tale metodo è quello di poter inserire gli elementi uno ad uno direttamente da tastiera, il semplice invio senza aver digitato nulla equivale alal conclusione dell'immissione dei dati nel vettore.

8.5 Inizializzazione di un vettore

In molti casi invece di creare un vettore abbiamo solamente bisogno di inizializzarlo. Questa operazione viene fatta con uno dei seguenti comandi:

- `>x<-vector`("numeric",5)
- `>x<-logical(5)`
- `>x<-numeric(5)`
- `>x<-complex(5)`
- `>x<-character(5)`

Si noti che il primo modo di inizializzare un vettore potrà essere usato anche per vettori di tipo logico, complessi o di caratteri.

8.6 Creazione di un vettore con la funzione *fix*

Tramite l'utilizzo della funzione `fix` è possibile procedere alla creazione di un vettore utilizzando la seguente sintassi:
`>x<-0`
`>x<-fix(x)`
e scrivendo i numeri utilizzando la sintassi:
`c(n1,n2,...)`
in cui `n1,n2` sono i numeri che compongono il vettore stesso.

8.7 Attributi di un vettore

Se `x` è un vettore qualsiasi, gli attributi di cui esso gode saranno i seguenti:

- `>length(x)` lunghezza del vettore
- `>mode(x)` modo del vettore
- `>names(x)` nomi del vettore

8.8 Nomi degli elementi di un vettore

Una volta creato un vettore che indicheremo con `x` sarà possibile assegnare a ciascun elemento del vettore un nome o una etichetta. Per fare ciò dobbiamo pero' avere a disposizione un vettore che indicheremo con `nomi` della stessa lunghezza di `x` che potrà essere sia numerico che composto di caratteri. L'attribuzione del nome avviene nel seguente modo:
`>names(x)<-c(y)`

8.9 Richiamare i singoli elementi di un vettore

Per richiamare i singoli elementi di un vettore `x` possiamo usare uno dei seguenti modi:

- `>x` richiama l'intero vettore
- `>x[n]` richiama l'elemento di posto `n` del vettore
- `>x[c(n1,n2,n3)]` richiama gli elementi di posto `n1,n2,n3` del vettore
- `>x[n1:n2]` richiama gli elementi di posto da `n1` a `n2` del vettore
- `>x[-(n1:n2)]` richiama tutti gli elementi del vettore tranne quelli da `n1` a `n2`
- `>x[-c(n1,n2,n3)]` richiama tutti gli elementi del vettore tranne quelli di posto `n1,n2,n3`
- `>x[x>n1]` richiama gli elementi del vettore maggiori di `n1`
- `>x[x>n1 | x<n2]` richiama gli elementi del vettore maggiori di `n1` o minori di `n2`
- `>x[x>n1 & x<n2]` richiama gli elementi del vettore maggiori di `n1` e minori di `n2`
- `>x["a"]` restituisce l'elemento del vettore con etichetta a"

8.10 Creare un vettore con seq

La funzione `seq` viene usata per creare delle sequenze di numeri. La sintassi è la seguente
`seq(primo,ultimo,incremento)`
in cui:

- `primo` è il primo elemento della sequenza
- `ultimo` è l'ultimo elemento della sequenza
- `incremento` è il passo con cui si va da `primo` ad `ultimo`

Si noti anche che:

- `incremento` può anche essere un numero negativo, in tale caso la sequeza al posto di essere crescente sarà decrescente, in questo caso si faccia attenzione ai valori dati a `primo` e `ultimo`
- se `incremento` non viene specificato viene utilizzato come valore `uno`

8.11 Creare un vettore con rep

La funzione `rep` serve per creare un vettore con dati ripetuti. La sua sintassi è la seguente:

- `>x<-rep(2,5)` crea 2, 2, 2, 2, 2
- `>x<-rep(c(1,2,3),2)` crea 1, 2, 3, 1, 2, 3
- `>x<-rep(c(1,2,3), each=2)` crea 1, 1, 2, 2, 3, 3
- `>x<-rep(c(1,2,3),c(2,3,4))` crea 1, 1, 2, 2, 2, 3, 3, 3, 3

8.12 Creare un vettore con cut

La funzione `cut` serve per assegnare i valori di un vettore, di solito derivante da dati continui, a classi di ampiezza prefissata o no. Si tratta in generale di una operazione di dicretizzazione del vettore. In certi casi, infatti, dato un vettore derivante da una distribuzione continua è necessario assegnare le sue componenti a classi appositamente prefissate. Supponendo di avere a disposizione un vettore `x` per compiere tale operazione possiamo usare i seguenti comandi:

- `>x<-cut(y,3)` suddivide y in tre gruppi
- `>x<-cut(y,breaks=c(0,2,4,6))` suddivide i dati nei gruppi 0-2, 2-4, 4,6
- `>x<-cut(y,breaks=c(0,2,4,6),labels=c`(“0-2“,“2-4“,“4-6“)) li suddivide nei gruppi e attribuisce loro i nomi scritti in labels
- `>x<-cut(y,breaks=seq(2,6,2),labels=c`(“0-2“,“2-4“,“4-6“))come sopra ma con uso di seq

Le calssi così ottenute risulteranno essere chiuse a destra, ma usando il comando `cut` con l'opzione `right=F` la classe che viene creata risulta essere chiusa a sinistra.

Si noti che possiamo anche limitare le classi ad esempio se x è un vettore con i primi 100 numeri naturali con in comando
`<cut(x,c(0,5,10,15,100),labels=c`(“0-5“,“5-10“,“10-15“,“15+))
otteniamo una suddivisione di x in 4 classi.

8.13 Creare un vettore logico

In certi casi è necessario stabilre quanti e quali elementi che compongono un vettore soddisfino ad una determinata condizione. Ciò può essere fatto con la creazione di un vettore logico, ossia un vettore che contiene solamente `TRUE o FALSE`. Se `x` è un vettore composto da numeri naturali con:
`>x<=5`
otteniamo un vettore logico composto da `T o F` a seconda che l'elemento del vettore soddisfi o meno la condizione scritta.

8.14 Ordinare un vettore

Per ordinare elementi di un vettore `x` utilizziamo il seguente comando:
`>sort(x)`
con il quale gli elementi di `x` sono ordinati in modo crescente, se invece vogliamo ottenere l'ordinamento decrescente dobbiamo utilizzare il comando:
`>sort(x,decreasing=T)`

8.15 Le funzioni elementari statistiche

Dato un vettore di tipo numerico `x`, le principali funzioni elementari statistiche applicabili a tale vettore sono le seguenti:

- `>min(x)` restituisce il minimo di x
- `>max(x)` restituisce il massimo di x
- `>range(x)` restituisce il range di x
- `>mean(x)` restituisce la media aritmetica semplice di x

- `>median(x)` restituisce la mediana di x
- `>quantile(x,y)` con `y` vettore di numero compresi tra zero ed uno, restituisce i quantili di x in base ai valori contenuti in `y`
- `>var(x)` restituisce la varianza di x
- `>sd(x)` restituisce la deviazione standard di x
- `>mad(x)` calcola la median absolute deviation
- `cor(x,y)` con `y` vettore numerico restuisce la correlazione tra x ed y
- `>cumsum(x)` restituisce la somma progressiva di x
- `>cumprod(x)` restituisce il prodotto progressivo di x
- `>sum(x)` restituisce la somma di x
- `>prod(x)` restituisce il prodotto di x

8.16 Operazioni che coinvolgono i vettori

Quando applichiamo una funzione aritmetica o di confronto ad un vettore, tale funzione sarà applicata a ciascun elemento che compone il vettore stesso. Quindi se `x` e `y` sono due vettori qualunque con:

- `>a+b` otteniamo un vettore che ha come elementi la somma degli elementi corrispondenti dei vettori `x` e `y`
- `>a*b` otteniamo un vettore che ha come elementi il prodotto degli elementi corrispondenti dei vettori `x` e `y`
- `>log(x)` otteniamo un vettore i cui elementi sono i logaritmi degli elementi di `x`

8.17 Estrarre valori da un vettore condizionati da altro vettore

Supponiamo di disporre di due vettori di uguale lunghezza:

- x di dati di tipo numerico
- y di tipo factor

Se vogliamo ottenere i valori numerici di x ma relativi alla modalità y1 del fattore y dobbiamo dare il seguente comando:
`x[y==y1]`
e così per ogni modalità di y.

9 Le matrici

9.1 Introduzione

Le matrici sono costotuite da dati dello stesso tipo che sono graggruppati in tabelle a doppia entrata. Anche le matrici come i vettori sono fondamentali nell'analisi statistica dei dati in quanto permettono un'analisi congiunta di due aspetti di una stessa unità statistica.

9.2 Come creare una matrice

Le matrici possono essere di tipo numerico, di stringhe di caratteri e logiche. Se il vettore `x` contiene i dati da inserire nella matrice ordinati secondo le colonne, per creare una matrice possiamo utilizzare la funzione `matrix` nel seguente modo:
`matrix(x,nrow=numerorighe,ncol=numerocolonne,byrow=F)`
Si noti che la lunghezza del vettore deve essere pari al valore di `nrow*ncol` in caso contrario si ottiene un messaggio di errore. Precisiamo inoltre quanto segue:

- l'opzione `byrow=F` è stabilita di default e fa sì che la creazione della matrice avvenga procedendo seguendo l'ordine delle colonne. Essa può essere quindi omessa se si desidera che i dati del vettore `x` siano ordinati secondo le colonne
- specificando l'opzione `byrow=T` vengono assegnati alla matrice i dati del vettore `x` seguendo l'ordine delle righe
- è sufficiente fissare uno solo dei parametri `nrow` e `ncol`in quanto il mancante sarà automaticamente determinato dalla lunghezza del vettore `x`

Altri modi in cui può essere creata una matrice sono i seguenti:

- `>x<-matrix(1:10,5,2)`
- `>x<-matrix(c(1,2,3,4),nr=2,dimnames=list(AAA=c(`“a“,“b“),BBB=c(“A“,“B“)))
- `>x<-matrix(c(1,0,0,1),2)`
- `>x<-1:10`
  `>dim(x)<-c(5,2)`
- `>x<-matrix(scan(datafile),ncol=10)`
- `x>x<-matrix(scan(datafile,what=`“character“),ncol=10)
- `>x<-matrix(scan(),ncol=5)`

Se si vuole creare una matrice avente tutti gli elementi uguali ad un numero prefissato `a` possiamo dare il seguente comando:
`>matrix(a,2,10)`
crea una matrice 2 x 10 composta di numeri `a`

9.3 Come creare una matrice diagonale

Per creare una matrice diagonale da un vettore `x` di dati noti possiamo dare la seguente sequenza di comandi:
`diag(x)`
Si ottiene una matrice diagonale di ordine pari alla lunghezza del vettore `x`.

9.4 Creazione di una matrice la funzione fix

Tramite l'utilizzo della funzione `fix` è possibile procedere alla creazione di una matrice utilizzando la seguente sintassi:
`>x<-matrix(0)`
`>fix(x)`
e scrivendo i numeri all'interno delle singole celle.

9.5 Attributi di una matrice

Se `x` è una data matrice, applicando i seguenti comandi possiamo ottenere alcune proprietà della matrice stessa:

- `>length(x)` restituisce la lunghezza di x
- `>mode(x)` restiduisce il modo di x
- `>dimnames(x)` restituisce i nomi di x
- `>dim(x)` restituisce le dimensioni di x

9.6 Dare un nome alle righe e colonne di una matrice

Se `x` è una matrice di dimensioni mxn è possibile dare un nome alle sue righe ed alle sue colonne con il seguente comando:

```
dimnames(x)<-list(c("nomerig1","nomerig2",...,"nomerigm"),
c("nomecol1","nomecol2",...,"nomecoln"))\\
```

Se vogliamo dare un nome solamente alle sue righe usiamo il seguente comando:

```
dimnames(x)[[1]]<-list("nomerig1","nomerig2",...,"nomerigm")
```

Se invece vogliamo dare un noime solamente alle sue colonne usiamo il seguente comando:

```
dimnames(x)[[2]]<-list("nomecol11","nomecol2",...,"nomecoln")
```

9.7 Estrarre dati da una matrice

Data una matrice mxn `x` di cui sappiamo che:

- le righe sono state chiamate nomerig1, nomerig2,...,nomerigm
- le colonne sono state chiamate nomecol1, nomecol2, ..., nomecoln

per richiamare i suoi elementi possiamo utilizzare i seguenti comandi:

- `>x[,`"nomecol1"]=x[,1] richiama la prima colonna di `x`
- `>x[`"nomerig1",]=x[1,] richiama la prima riga di `x`
- `>x[,c(`"nomecol1","nomecol2")]=x[,c(1:2)] richiama le prime due colonne di `x`
- `>x[,c(1,3)]` richiama la prima e la terza colonna di `x`
- `>x[c(1,3),c(1,3)]` richiama la prima e la terza riga e colonna di `x`
- `>x[-1,c(1,3)]` richiama la prima e la terza colonne di `x` tranne la riga 1
- `>x[1:2,3]` richiama i primi 2 elementi della colonna 3
- `>x[x[,1]>=2,1:2]` richiama le colonne da 1 a 2 di `x` i cui elementi della prima colonna sono maggiori o uguali a 2
- `>x[x[,1]>=2,]` richiama tutte le colonne di `x` i cui elementi della prima colonna sono maggiori o uguali a 2
- x[1,1] richiama l'elemento di posto $(1; 1)$

Da questi esempi si possono facilmente ricavare altri casi di estrazione non contemplati.

9.8 Richiamare i nomi delle righe e delle colonne una matrice

Se `x` è una matrice allora con il comando:
`>dimnames(x)`
richiamiamo i nomi delle righe e delle colonne di `x`.

Per richiamare solo i nomi delle righe della matrice dobbiamo usare il comando:
`>dimnames(x)[[1]]`
mentre per richiamare solo i nomi delle colonne dobbiamo eseguire:
`>dimnames(x)[[2]]`

9.9 Le funzioni cbind e rbind

Per aggiungere righe o colonne ad una matrice possiamo utilizzare le funzioni:

- `cbind` con il quale si aggiungono una o più colonne ad una matrice esistente
- `rbind` con il quale si aggiungono una o più righe ad una matrice esistente.

Se `x` è una matrice ed `y,z` sono due vettori che hanno lo stesso numero di righe di `x` con il comando:
`>w<-cbind(x,yyy=y,zzz=z)` si crea la matrice `w` che contiene la matrice `xxx` aumentata di due colonne denominate `xxx,yyy`

Se `x` è una matrice ed `y,z` sono due vettori che hanno come lunghezza lo stesso numero di colonne di di `x` con il comando:
`>w<-rbind(x,yyy=y,zzz=z)` si crea la matrice `w` che contiene la matrice `x` aumentato di due righe denominate `xxx,yyy`

9.10 Trasformare una matrice in un vettore

Può essere utile a volte trasformare una matrice in un vettore. Se `x` è una matrice con il comando:
`>y <- as.vector(x)`
si trasforma la matrice `x` nel vettore `y` seguendo l'ordine delle colonne.

9.11 Operazioni sulle matrici

Come accade per i vettori, anche per le matrici le operazioni aritmetiche e di confronto vengono eseguite elemento per elemento. Se `a e b` sono due matrici che permettono di eseguire le operazioni elementari si avrà che:

- per eseguire la somma
  `>a+b`
- per eseguire il prodotto
  `>a%*%b`
- per calcolare l'inversa
  `>solve(a)`
- per trovare la trasposta `>t(a)`

Se `x` è una matrice mxn ed `y` è un vettore $mx1$ allora con il comando
`x-y`
da ogni colonna di `x` il corrispondente valore del vettore `y`.

9.12 Trovare il determinante di una matrice

Data una matrice `x` il determinante di tale matrice si trova con il comando:
`>det(x)`

9.13 Autovalori ed autovettori di una matrice

Data una matrice `x` gli autovalori e gli autovettori di tale matrice si trova con il comando:
`>eigen(A)`
gli autovettori ottenuti in questo modo sono normalizzati. Si veda l'help in linea del comando `eigen` per vedere il significato delle numerosi opzioni che si possono ottenere

9.14 Aggiungere righe e colonne ad una matrice

Per creare una matrice o aggiungere righe o colonne ad una matrice possiamo utilizzare i comandi

- `cbind` il quale crea una matrice per colonne o aggiunge una o più colonne ad una matrice esistente
- `rbind` il quale crea una matrice per righe o aggiunge una o più righe ad una matrice esistente

Se `x` ed `y` sono vettori di uguale lunghezza con:

- `>w<-cbind(x,y)` si crea una matrice avente come colonne gli elementi dei vettori `x` e `y` e nomi delle colonne `x` ed `y`
- `>w<-rbind(x,y)` si crea una matrice avente come righe gli elementi dei vettori `x` e `y` e nomi delle righe `x` ed `y`
- `>w<-cbind(A,x)` se `A` è una matrice compatibile con il vettore `x` si ottiene una matrice data dalla matrice `A` alla quale è stata aggiunta una nuova colonna composta dagli elementi di `x` e chiamata proprio `x`
- `>w<-rbind(A,x)` se `A` è una matrice compatibile con il vettore `x` si ottiene una matrice data dalla matrice `A` alla quale è stata aggiunta una nuova riga composta dagli elementi di `x` e chiamata proprio `x`

10 Gli array

10.1 Introduzione

Gli array sono costituiti da dati dello stesso tipo che sono graggruppati in tabelle ad entrata multipla ordiante in relazione alle variabili che lo compongono. Gli array come i vettori e le matrici sono fondamentali nell'analisi statistica dei dati in quanto permettono un'analisi congiunta di più di due aspetti di una stessa unità statistica.

10.2 Come creare un'array

Se il vettore `x` contiene i dati da inserire nell'array ordinati secondo le colonne e secondo l'ordine delle matrici da ottenere, per creare un'array possiamo utilizzare la funzione `array` nel seguente modo:
`matrix(x,dim=c(n1, n2, ... nk)`
in cui `n1, n2, ... nk` sono numeri naturali che rappresentano la numerosità delle modalità di ogni singola variabile che compone l'array. Si noti che la lunghezza del vettore deve essere pari al valore di `n1*n2*...*nk` in caso contrario si ottiene un messaggio di errore. Un alto modo per creare un array è il seguente:
`>dim(x)<-c(n1, n2, ..., nk)`
in cui `n1, n2, ... nk` come appena detto, sono numeri naturali che rappresentano la numerosità delle modalità di ogni singola variabile che compone l'array.

10.3 Attributi di un'array

Se `x` è un array applicando i seguenti comandi possiamo ottenere alcune proprietà dell'array stesso:

- `>length(x)` restituisce la lunghezza di x
- `>mode(x)` restituisce il modo di x
- `>dimnames(x)` restituisce i nomi di x
- `>dim(x)` restituisce le dimensioni di x

10.4 Dare un nome agli elementi dell'array

Se `x` è un'array formato dalle variabili `n1 x n2 x ... x nk` è possibile dare un nome alle modalità con cui si presenta ogni singola variabile con il seguente comando:

```
dimnames(x)<-list(c("n11","n12",...,"n1m"),
c("n21","n22",...,"n2n"),
c("....................."),
c("nk1","nk2",...,"nkq"))\\
```

in cui gli elementi fra virgolette sono i nomi che vengono dati alle singole modalità di ogni variabile dell'array.

10.5 Richiamare un array in base ad una modalità

Dato un array `x` formato dalle variabili `n1 x n2 x n3` per richiamare l'array in base allo stato che assume una singola modalità di una variabile dell'array possiamo utilizzare i seguenti comandi:

- `>x[,1,]` stato primo assunto dalal variabile seconda
- `>x[,`"n2i",] stato i assunto dalla variabile seconda

Si può facilmente estendere questi esempi a casi non contemplati.

10.6 Richiamare i nomi delle modalità delle varibili di un'array

Se `x` è un'array allora con il comando:
`>dimnames(x)`
richiamiamo i nomi delle modalità delle singole variabili di `x`.

Per richiamare i nomi delle modalità della i-ma variabile dell'array dobbiamo usare il comando:
`>dimnames(x)[[i]]`

10.7 Trasformare un array in un vettore

Può essere utile a volte trasformare un array in un vettore. Se `x` è un array con il comando:
`>y <- as.vector(x)`
si trasforma l'array `x` nel vettore `y` seguendo l'ordine delle colonne.

10.8 Lavorare con gli array

Vogliamo ne l seguente paragrafo mostrare come con l'uso degli array sia possibile semplificare notevolmente l'analisi statistica dei dati di una certa analisi. Supponiamo di avere a disposizione il seguente array:

```
Titanic
   , , Age = Child, Survived = No

       Sex
   Class  Male Female
   1st     0      0
   2nd     0      0
   3rd    35     17
   Crew    0      0

   , , Age = Adult, Survived = No

       Sex
   Class  Male Female
   1st   118      4
   2nd   154     13
   3rd   387     89
   Crew  670      3

   , , Age = Child, Survived = Yes

       Sex
   Class  Male Female
   1st     5      1
   2nd    11     13
   3rd    13     14
   Crew    0      0

   , , Age = Adult, Survived = Yes
```

```
      Sex
Class  Male Female
1st    57    140
2nd    14     80
3rd    75     76
Crew  192     20
```

Notiamo che l'array Titanic presenta le variabili

- class
- sex
- age
- survived

aventi le seguenti modalità:

- class = 1 2 3 4
- sex = m f
- age = child adult
- survived = no Yes

Se vogliamo ottenere un array in cui perdiamo la variabile `age` dobbiamo dare i seguenti comandi:
`>xxx<-Titanic[,,1,]+Titanic[,,2,]`
così facendo otteniamo un array denominato `xxx` avente variabili e modalità date da:

- class = 1 2 3 4
- sex = m f
- survived = no Yes

Digitando ora il seguente comando:
`yyy<-xxx[,1,]+xxx[,2,]`
otteniamo l'array denominato `yyy` in cui perdiamo la variabile sex e che sarà composto dalle seguenti variabili aventi modalità:

- class = 1 2 3 4
- survived = no Yes

Da questo semplice esempio è possibile rendersi immediatamente conto dell'importanza dell'uso degli array nell'analisi dei dati.

11 List

11.1 Introduzione

Le liste sono degli oggetti destinati a contenere più dati anche di diversa natura. Esse sono di notevole importanza quando un comando deve produrre dati sia di tipo numerico che di tipo carattere.

11.2 Come creare una lista

Una lista può essere creata nel seguente modo:
`>x<-rnorm(100) >y<-c`(“stringa1“,“stringa2“,“stringa3“)
`>z<-matrix(rnorm(100),ncol=10)`
`>mylist<-list(nome1=x,nome2=y,nome3=z)`
si ottiene in questo modo una lista di nome mylist e contenente oggetti diversi tra di loro il primo del quale si chiama nome1, il secondo nome2 e il terzo nome3.

11.3 Attributi di una lista

Gli attributi di una lista sono i seguenti:

- `>length(x)` restituisce la lunghezza di x
- `>mode(x)` restituisce il modo di x
- `>names(x)` restituisce i nomi di x

11.4 Dare un nome agli elementi di una lista

Se `x` è una lista composta da tre oggetti per dare dei nomi a tali oggetti o per cambiare i nomi eventualmente già asseganti possiamo usare il seguente comando:
`>names(x)<-c(`“nome1“,“nome2“,“nome3“)

11.5 Richiamare gli elementi di una lista

Se `x` è una lista composta da tre oggetti aventi rispettivamente `nome1, nome2, nome3` per richiamarne i singoli elementi possono usare operare nel seguente modo:

- `>x[[1]]` restituisce il primo elemento della lista
- `x[[1]][n1:n2]` restituisce gli elementi dal n1 al n2 del primo elemento della lista
- `>x$a` richiama il primo elemento della lista
- `x$a[9:12]` richiama gli elementi da n1 ad n2 del primo elemento della lista

Analogamente si opera per gli altri elementi della lista.

12 Factor

12.1 Introduzione

In una indagine statistica può accadere alle volte che i dati di una certa variabile siano raccolti interamente in un unico vettore e solamente tramite un altro vettore della stessa lunghezza del primo è possibile stabilire da quale realtà essi provengono. Ad esempio se volessimo fare un'analisi dei voti riportati in italiano dagli alunni di una scuola sarebbe più agevole raccogliere in una variabile di nome `italiano` i voti riportati dai singoli alunni e creare una variabile `classe` nella quale inseriremo in corrispondenza del voto la classe dell'alunno che ha riportato quel voto. Le variabili che permetto di fare cio' sono dette variabili di tipo `factor`. Gli elementi distinti che compongono l'oggetto factor sono detti `livelli` del fattore. Le variabili di tipo factor permettono quindi di attribuire un dato ad un deteminato livello di un fattore considerato.

12.2 Come creare una variabile factor

Supponiamo di avere a disposizione due variabili `x` e `y` così formate:
`>x<-c(`“a,“b“,“c“,“a“,“a“)
`>y<-c(1,2,3,4,3,2,1,2,3,3,3,4,4,3,2,1,1,1,2,3,4,5,5,5,4,3,2,3)`
per trasformarle in oggetti di tipo factor possiamo operare nel seguente modo:

- `>x<-factor(x)`
  si crea un factor da caratteri che saranno ordinati secondo ordine alfabetico
- `>y<-factor(y)`
  si crea un factor da numeri che saranno ordinati secondo l'ordine naturale
- `>x<-ordered(x),levels=c(`“c“,“b“,“a“))
  si crea un factor da caratteri che saranno ordinati secondo l'ordine dato dall'opzione `levels`
- `>y<-ordered(y),levels=c(5,4,3,2,1))`
  si crea un factor da numeri che saranno ordinati secondo l'ordine dato dall'opzione `levels`

- `>x<-factor(x,levels=c(`a“,“b“))
  gli elementi del vettore con la lettera c evidenzieranno un NA
- `>x<-factor(x,levels=c(`a“,“b“,“c“),labels=c(“CorsoA“,“CorsoB“,“Corso C“))
  si crea un factor con i livelli nominati come scritto nell'opzione `labels`
- `>x<-factor(x,levels=c(`a“,“b“),labels=c(“CorsoA“,“CorsoB“))
  si crea un factor con i livelli nominati come scritto il `labels` tranne quelli con c che appariranno con NA
- `>x<-factor(x,exclude=c(`c“))
  in questo modo escudiamo dalla creazione dei fattori il corso c e dove ci sarà comparirà un NA
- `>x<-factor(x)` ed `levels(x)<-c(a,b,c,d,e)` si ottiene un oggetto factor in cui ad 1 è associato il livello a, a 2 il livello b e così via di seguito
- `>x<-factor(x)` ed `levels(x)<-c(x,y,z,t,v)` si ottiene un oggetto factor in cui ad 1 è associato il livello x, a 2 il livello y e così via di seguito

12.3 Creare factor da dati continui

E' possibile creare degli oggetti factor da dati continui. Se creiamo un vettore x contenente 100 numeri casuali provenienti da una variabile aleatoria normale di media 10 e varianza 1:
`>x<-rnorm(100,10,1)`
con il comando
`>y<-cut(x,breaks=c(0,2,4,6,10))`
`>y<-factor(y)`
E' anche possibile ordinare i dati in modo diverso o escluderne qualcuno utilizzando i comandi visti precedentemente.

12.4 Come creare un factor dicotomico

Se `x` è un vettore che contiene i numeri naturali da 1 a 5 con il comando:
`>y<-factor(x==3)` si crea un factor avente tutti F tranne un T al posto 3. Utilizzando questo procedimento è sempre possibile creare da ogni qualsivoglia vettore dei factor di tipo dicotomico.

12.5 Attributi di un factor

Gli attributi di un factor `x` sono i seguenti:

- `>length(x)` restituisce la lunghezza di x
- `>mode(x)` restituisce il modo di x
- `>names(x)` restituisce i nomi di x
- `>levels(x)` restituisce i livelli di x
- `>class(x)` restituisce la classe di x

12.6 La funzione gl

La funzione `gl` è una importante funzione che permette di creare oggetti factor in modo molto semplice. La sua sintassi è la seguente:
`>gl(n, k, length = n*k, labels = 1:n, ordered = FALSE)`
in cui:

- n è un intero che rappresenta il numero dei livelli
- k è un intero che indica il numero delle replicazioni
- length è un intero che restituisce la lunghezza del risultato
- labels è un vettore opzionale di lunghezza n che da i nomi dei livelli del fattore

- ordered è un oggetto logico per stabilire se i fattori devono essere ordinati o no

Utilizzando questa funzione sarà possibile creare replicazioni di livelli nell'ordine che vogliamo per poter risparmiare tempo nell'introduzione di tali oggetti soprattutto in problemi di analisi della varianza.

13 Data Frame

13.1 Introduzione

I `data frame` sono oggetti simili alle matrici ma con colonne che possono contenere dati di diverso tipo. Nel `data frame` sarà allora possibile combinare colonne contenenti numeri con colonne di stringhe o di oggetti factor. E' molto comodo rappresentare i dati provenienti da una indagine statistica in `data fra me` in cui le righe rappresentano l'unità statistica di rilevazione mentre le colonne rappresentano le variabili rilevate nell'indagine stessa.

13.2 Come creare un data frame

Supponiamo di avere a disposizione tre variabili con lo stesso numero di elementi date da:
`>x<-rnorm(6)`
`>y<-c`(“a“,“a“,“b“,“c“,“a“,“c“)
`>z<-1:6`
il data frame `xxx` sarà creato in uno dei seguenti modi:

1. `>xxx<-data.frame(x,y,z)` crea un data frame con tre varibili aventi per nomi `x,y,z`
2. `>xxx<-data.frame(nome1=x,nome2=y,nome3=z)` crea un data frame con tre variabili aventi per nomi `nome1,nome2,nome3`

Se una delle varibile del data frame è di tipo `character` viene automaticamente trasformata in variabile di tipo factor.

13.3 Creazione di un data.frame con la funzione fix

Tramite l'utilizzo della funzione `fix` è possibile procedere alla creazione di un data frame utilizzando la seguente sintassi:
`>xxx<-data.frame() >fix(x)`

13.4 Attributi di un data frame

Se `xxx` è un data frame i suoi attributi sono ottenuti con:

- `>length(xxx)` restituisce la lunghezza di xxx
- `>mode(xxx)` restituisce il modo di xxx
- `>row.names(xxx)` restituisce i nomi delle righe di xxx
- `>dimnames(xxx)` restituisce i nomi sia delle righe che delle colonne di di xxx
- `>dim(xxx)` restituisce le dimensioni di xxx
- `>nrow(xxx)` restituisce il numero delle righe di xxx
- `>ncol(xxx)` restituisce il numero delle colonne di xxx
- `>names(xxx)` restituisce i nomi delle colonne di xxx
- `>class(xxx)` restituisce la classe di xxx

13.5 Dare un nome alle righe e colonne di un data frame

Se `xxx` è una data frame di dimensioni mxn è possibile dare un nome alle sue righe ed alle sue colonne con il seguente comando:

```
dimnames(xxx)<-list(c("nomerig1","nomerig2",...,"nomerigm"),
c("nomecol1","nomecol2",...,"nomecoln"))\\
```

Se vogliamo dare un nome solamente alle sue righe usiamo il seguente comando:

```
dimnames(xxx)[[1]]<-list(c("nomerig1","nomerig2",...,"nomerigm"))
```

Se invece vogliamo dare un noime solamente alle sue colonne usiamo il seguente comando:

```
dimnames(xxx)[[2]]<-list(c("nomecol11","nomecol2",...,"nomecoln"))
```

13.6 Estrarre dati da un data frame

Se `xxx` è un data frame di dimensioni mxn di cui sappiamo che:

- le righe sono state chiamate nomerig1, nomerig2,...,nomerigm
- le colonne sono state chiamate nomecol1, nomecol2, ..., nomecoln

per richiamare i suoi elementi possiamo utilizzare i seguenti comandi:

- `>xxx[,`"nomecol1"]=xxx[,1] richiama la prima colonna di `xxx`
- `>xxx[`"nomerig1",]=xxx[1,] richiama la prima riga di `xxx`
- `>xxx[,c(`"nomecol1","nomecol2")]=xxx[,c(1:2)] richiama le prime due colonne di `xxx`
- `>xxx[,c(1,3)]` richiama la prima e la terza colonna di `xxx`
- `>xxx[c(1,3),c(1,3)]` richiama la prima e la terza riga e colonna di `xxx`
- `>xxx[-1,c(1,3)]` richiama la prima e la terza colonne di `xxx` tranne la riga 1
- `>xxx[1:2,3]` richiama i primi 2 elementi della colonna 3
- `>xxx[x[,1]>=2,1:2]` richiama le colonne da 1 a 2 di `xxx` i cui elementi della prima colonna sono maggiori o uguali a 2
- `>xxx[x[,1]>=2,]` richiama tutte le colonne di `xxx` i cui elementi della prima colonna sono maggiori o uguali a 2
- xxx[1,1] richiama l'elemento di posto $(1; 1)$

Da questi esempi si possono facilmente ricavare altri casi di estrazione non contemplati. Si noterà certamente che questo modo di procedere è identico a quello usato nelle matrici. In un data frame è possibile anche usare una sintassi differente per richimare le singole colonne del data frame stesso. Infatti con:

```
>xxx$nomecol1 si richiama la prima colonna di xxx
>xxx$nomecol5 richiama la quinta colonna di xxx
```

13.7 Estrarre dati da un data frame: casi di notevole interesse

Vogliamo in questo paragrafo esaminare alcuni casi interessanti di estrazione di dati da un data frame. Se `xxx` è un data frame con variabili `var1,var2,var3,var4`, in cui `var1,var2` sono di tipo numerico, mentre `var3` e di tipo factor con livelli rappresentati da caratteri e denominati `var3liv1,var3liv2,var3liv3` e `var4` è ancora di tipo factor ma con livelli rappresentati da numeri `var4liv1,var4liv2,var4liv3` con i seguenti comandi:

```
>xxx[xxx$var4==var4liv1,]      estraiamo dal data frame solo gli elementi
                               che presentano nella variabile 4il livello 1
>xxx[xxx$var3=="var3liv1",]    estraiamo dal data frame solo gli elementi
                               che presentano nella variabile 3 il livello 1
>xxx[xxx$var3=="var3liv2",]    estraiamo dal data frame solo gli elementi
                               che presentano nella variabile 3 il livello 2
>xxx[xxx$var1>numero,]         estraiamo dal data frame solo gli elementi
                               che presentano nella variabile 1 numeri maggiori
                               di numero
>xxx[xxx$var2>=numero,"var1"]  estraiamo dal data frame solo gli elementi
      che presentano nella variabile 2 numeri maggiori
      di numero ma estraiamo solamente i valori della
                               variabile 2
>xxx[xxx$var2>=numero,1]       idem come sopra
>xxx[xxx$var2<numero,c("var1","var3")] estraiamo dal data frame solo
      gli elementi che presentano nella variabile 2
                               numeri maggiori o uguali di numero  ma
      estraiamo solamente i valori della variabile
      1 e variabile 3
>xxx[xxx$var2<,c(1,3)]         idem come sopra
>xxx[,"var1"]                  si ottiene un vettore che riporta solamente
      i valori della var1
>xxx[,1]                       idem come sopra
```

Naturalmente sarà anche possibile applicare gli operatori logici `&` ed `|` per combinare tra di loro più condizioni.

In ogni caso si noti che se nella colonna che stiamo analizzando sono presenti dei valori `NA` con i comandi dati compariranno anche tutti i valori in cui risulta presente anche `NA`. Per ovviare a tale fatto sarà sufficiente nella condizione che stiamo trattando inserire `& is.na(variabile)==F`, in questo modo compariranno solamente i valori diversi da `NA` con la condizione inserita.

13.8 Richiamare i nomi delle righe e delle colonne un data frame

Se `xxx` è un data frame allora con il comando:
`>dimnames(xxxx)`
richiamiamo i nomi delle righe e delle colonne di `xxx`.

Per richiamare solo i nomi delle righe del data frame dobbiamo usare il comando:
`>dimnames(xxx)[[1]]`
mentre per richiamare solo i nomi delle colonne dobbiamo eseguire:
`>dimnames(xxx)[[2]]`

13.9 Le funzioni cbind e rbind

Per aggiungere righe o colonne ad un data frame possiamo utilizzare le funzioni:

- `cbind` con il quale si aggiungono una o più colonne ad un data frame esistente
- `rbind` con il quale si aggiungono una o più righe ad un data frame esistente.

Se `xxx` è un data frame ed `y,z` sono due vettori che hanno la stessa lunghezza di `xxx` con il comando:
`>www<-cbind(xxx,yyy=y,zzz=z)` si crea il data frame `www` che contiene in data frame `xxx` aumentato di due variabili denominate `xxx,yyy`

Se `xxx` è un data frame ed `y,z` sono due vettori che hanno come lunghezza il numero delle variabili di `xxx` con il comando:
`>www<-rbind(xxx,yyy=y,zzz=z)` si crea il data frame `www` che contiene in data frame `xxx` aumentato di due righe denominate `xxx,yyy`

Si noti in ogni caso che che per aggiungere una colonna ad un data frame `xxx` vi è anche la possibilità di usare la sintassi:
`w<-data.frame(xxx,yyy=y)`
con la quele si aggiunge la varibile `y` con nome `yyy` al data frame `xxx`

13.10 Le funzioni attach e detach

Queste due funzioni sono molto importanti in un data frame in quanto permetto di accedere alle singole variabili che compongono il data frame con una sintassi più semplice e più immediata. Se `xxx` è un data frame con colonne denominate `var1, var2` dopo aver eseguito il comando:
`>attach(xxx)`
potremmo accedere alle variabili `var1` e `var2` senza dover utilizzare la sintassi

```
xxx$var1
xxx$var2
```

ma semplicemente digitando

```
var1
var2
```

Con il comando `>detach(xxxx)` si ripristina la condizione originaria.

13.11 Ordinamento di un data frame

Molte volte cui troviamo nella necessità di dover ordinare un intero data frame facendo riferimento all'ordinamento di una sua variabile. Se `xxx` è un data frame con colonne `var1,var2,var3,var4` dopo aver eseguito il comando:
`>attach(xxx)`
se desideriamo ordinare l'intero data frame rispetto alla variabile `var1` dobbiamo eseguire il comando:
`>xxx[order(xxx[,1]),]`
oppure l'equivalente
`>xxx[order(xxx[,`var1]),]
analogamente si ragiona se l'ordinamento deve essere fatto utilizzando le altre variabili.

13.12 Dati provenienti da una tabella semplice

Può accadere, soprattuto per analisi statistiche avanzate, di dover inserire in un data frame i dati relativi ad una tabella semplice in modo da riportare per ogni dato sia la riga che la colonna da cui questi dati provengono. Supponiamo di disporre di una tabella del tipo:

Blocco \ Trattamento	A	B	C	D
1	89	88	97	94
2	84	77	92	79
3	81	87	87	85
4	87	92	89	84
5	79	81	80	88

Per inserire tali dati nel modo precisato in premessa in un data frame denominato `xxx` possiamo operare nel seguente modo:

- prima di tutto inseriamo i dati numerici in ordine di colonna nella variabile dati
- quindi creiamo la variabile trattamento con
  `>trattamento<-rep(paste(`"T",LETTERS[1:4]),c(5,5,5,5))
  oppure con
  `>trattamento<-rep(paste(`"T",LETTERS[1:4]),rep(5,4))
- quindi creiamo la variabile blocco con
  `>blocco<-rep(paste(`"blocco ",1:5),4)
- quindi creiamo il data.frame con
  `>xxx<-data.frame(dati=dati,trattamenti=trattamento,blocchi=blocco)`

13.13 Dati provenieti da una tabella con più entrate di valori

Può accadere che i dati provenienti da una stessa riga e una stessa colonna siano più di uno. Ci troviamo allora nella necessità di dover inserire in un data frame che chiameremo `xxx` più dati per ogni riga e ogni colonna. Supponiamo di disporre di una tabella del tipo:

Blocco \ Trattamento	A	B	C	D
1	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
2	0.32	0.72	0.53	0.15
	0.15	1.16	0.15	0.11
	0.26	0.08	0.03	1.66
	0.93	0.72	0.16	2.62
3	0.21	0.42	0.03	0.85
	0.05	1.30	0.15	0.41
	0.06	0.28	0.93	0.16
	0.03	0.12	0.16	0.02

Per inserire tali dati in un data frame chiamato `xxx` nel modo precisato in premessa possiamo operare nel seguente modo:

- prima di tutto inseriamo i dati numerici in ordine di colonna nella variabile dati
- quindi creiamo la variabile trattamento con
  `>trattamento<-rep(paste(`“T“,LETTERS[1:4]),rep(12,4))
- quindi creiamo la variabile blocco con
  `>xxx<-rep(paste(`“blocco “,1:3),rep(3,3))
  e quindi
  `>blocco<-c(xxx,xxx,xxx,xxx)`
  oppure con
  `>blocco<-rep(xxx,4)`
- quindi creiamo il data.frame con
  `>xxx<-data.frame(dati=dati,trattamenti=trattamento,blocchi=blocco)`

14 Testing e coercing data

14.1 Introduzione

Molte volte ci si trova nella necessità di dover verificare se un determinato oggetto appartiene ad un tipo specificato oppure di convertire il tipo di oggetto in un’altro. Per compiere queste operazioni possiamo utilizzare i comandi descritti nei paragrafi successivi.

14.2 Testare e convertire oggetti

Per testare e convertire gli oggetti utilizzati dal programma possiamo utilizzare i seguenti comandi:

- is.array(x)
- as.array(x)
- is.complex(x)
- as.complex(x)
- is.data.frame(x)
- as.data.frame(x)
- is.double(x)

- as.double(x)
- is.factor(x)
- as.factor(x)
- is.integer(x)
- as.integer(x)
- is.list(x)
- as.list(x)
- is.logical(x)
- as.logical(x)
- is.matrix(x)
- as.matrix(x)
- is.na(x)
- is.null(x)
- as.null(x)
- is.numeric(x)
- as.numeric(x)
- is.ts(x)
- as.ts(x)
- is.vector(x)
- as.vector(x)

Si noti in particolare che con l'uso di:

- `is.` vogliamo verificare se un certo oggetto sia di un certo tipo
- `as.` forziamo un oggetto ad essere di un tipo specificato

15 Uso di alcune funzioni notevoli

15.1 Introduzione

Vogliamo in questi paragrafi analizzare alcune funzioni di notevole importanza soprattutto nell'analisi dei dati contenuti in un data frame.

15.2 Richiesta di una funzione in un comando

In certi casi sarà necessario costruire una funzione per applicarla ai dati che si hanno a disposizione. Si possono presentare due casi:

- la funzione che vogliamo applicare ai dati è già presente in R o perchè predefinita o perchè già creata in precedenza: in questo caso basterà digitare il nome della funzione stessa quando necessario
- la funzione non è presente in R e quindi dovrà essere inserita tramite, ad esempio, con la sintassi: `function(x) x^2`

Inoltre ricordiamo che tramite il comando:
`na.rm=T`
facciamo in modo che la funzione ignori i valori `NA`, in caso contrario in presenza di tali valori errebe generato un errore.

15.3 La funzione aggregate

Con `aggregate` possiamo applicare una funzione contemporaneamente a più varibili inserite in un un data frame e di applicare contemporaneamente tale funzione anche a sottoinsiemi in cui eventualmente tali varibili sono suddivise. Se `xxx` è un data frame con colonne ordinate `var1,var2,var3,var4` con il seguente comando:

```
>aggregate(xxx[,c("var1","var2")],by=list(x$var3,x$var4),FUN=mean)
```

otteniamo la media delle variabili `var1, var2` suddivise in base ai valori assunti dalle variabili `var3, var4`. Di solito le variabili `var3, var4` sono di tipo factor.

15.4 La funzione apply

Con `apply` possiamo applicare una funzione alle righe o alle colonne di un data frame o di una matrice che indicheremo con `xxx`. Il suo uso è il seguente:

- `>apply(xxx,1,mean)` calcola la media di tutte le righe
- `>apply(xxx,2,mean)` calcola la media di tutte le colonne
- `>apply(xxx,2,mean,na.rm=T)` escude dalla media i valori NA

Nell'uso di tale funzione bisognerà fare attenzione, nel caso di un data frame, della natura delle singole varibili che lo compongono. Non è infatti possibile sommare o fare medie tra numeri e caratteri.

15.5 La funzione tapply

Con `tapply` possiamo applicare una funzione ad una variabile di un data frame facendo riferimento a eventuali sottoinsiemi in cui tale varibile e suddivisa e rappresentati da variabili di tipo `factor`. La sua sintassi è la seguente:

- `>tapply(x,factor,mean)`
- `>tapply(x,list(factor1,factor2),mean)`
- `>tapply(x,list(factor1,factor2),var)`

Si noti che factor può anche essere un oggetto numeric ed in questo caso sarà automaticamente convertito in oggetto factor. Si può usare dove necessario l'opzione `na.rm` come visto precedentemente. Tale funzione è analoga alla funzione aggregate solamente che è applicata ad una sola varibile a differenza di aggregate che può essere applicato a più variabili contemporaneamente.

15.6 La funzione lapply

Con `lapply` possiamo applicare una funzione ad un vettore, lista o data frame di dati. Il risultato sarà `una lista` che conterrà il valore della funzione applicato ad ogni elemento dell'oggetto considerato. Ad esempio con:
`>x<-seq(-3,3,0.1)`
`>lapply(x, function(x) pnorm(x))`
si ottiene una lista composta da 61 elementi.

15.7 La funzione sapply

Con `sapply` possiamo applicare una funzione ad un vettore, lista o data frame di dati. Il risultato sarà `un vettore` o `un array` che conterrà il valore della funzione applicato ad ogni elemento dell'oggetto considerato. Ad esempio con:
`>x<-seq(-3,3,0.1)`
`>lapply(x, function(x) pnorm(x))`
si ottiene `un vettore` composta da 61 elementi.

15.8 La funzione paste

Con `paste` possaimo unire stringhe con numeri o altro. Ad esempio con :
`>paste(letters[1:5],1:5,sep=“ “)`
otteniamo come risultato:
`>[1] a 1  b 2  c 3  d 4  e 5`
Se non si vuole la separazione tra le lettere e i numeri basterà semplicemente nell'opzione `sep` introdurre le virgolette senza lo spazio.

15.9 La funzione split

Con `split` possiamo ottenere dati suddivisi per gruppi a seconda di uno o più oggetti factor specificati. Se `xxx` è un data frame con variabili `var1,var2,var3,var4` e le variabili `var3, var4` sono di tipo factor, con il comando:
`>split(xxx$var1,list(xxx$var3))`
otteniamo i vettori della `var1` suddivisi in base ai valori della `var3`. Se invece usiamo:
`>split(xxx,list(xxx$var3,xxx$var4))`
otteniamo più data frame suddivisi sia per `var3` che per `var4`. I risultati sono ottenuti sotto forma di oggetti di tipo list.

15.10 La funzione stem

Con `stem` si ottiene un grafico che permette di fare delle considerazioni preliminari sui dati di una analisi contenuti in un vettore. Se `x` è un vettore che contiene i dati da analissare possiamo utilizzare la seguetne sintassi:
`>stem(x)`
`>stem(x,scale=2)`

15.11 La funzione summary

Con `summary` otteniamo dei dati di riepilogo su molti oggetti utilizzati dal programma stesso. La sua sintassi è la seguente:
`>summary(x)`

15.12 La funzione tabulate

Con `tabulate` possiamo ottenere le frequenze assolute con cui una certa modalità di una certa variabile compare nel vettore in cui abbiamo raccolto i dati della nostra analisi. Se abbiamo a disposizione un vettore `x` di dimensione `n` con la funzione:
`tabulate(x)`
otteniamo un vettore che raccoglie le frequenze assolute delle modalità con cui la variabile `x` si presenta.

15.13 La funzione fivenum

Con `fivenum` otteniamo i cinque numeri fondamentali relativi ad una singola variabile di tipo quantitativo. Se abbiamo a disposizione un vettore `x` di dimensione `n` con la funzione:
`fivenum(x)`
otteniamo come risultato un vettore formato dal minimo, primo quartile, mediana, terzo quartile e massimo calcolati sui valori del vettore argomento della funzione stessa.

15.14 La funzione which

Con `which` possiamo conoscere gli indici degli elementi di un vettore che soddisfano ad una certa condizione. Se `x` è un vettore numerico, con il comando:
`which(x>7)`
vengono individuati gli indice degli elementi del vettore che soddisfano la condizione data.

15.15 La funzione unique

Con `unique` vengono individuate le modalità con cui una certa varibile si presenta. Se `x` è un vettore la funzione verrà eseguita con il seguente comando:
`unique(x)`

15.16 Uso della funzione subset

Dato un data frame `xxx` se vogliamo ottenere il data frame condizionato al valore assunto da una sua variabile dobbiamo utilizzare il comando
`subset(nomedataframe,variabile==valore)` In cui nomedataframe è il nome del data frame da considerare, variabile è il nome della variabile e valore è il valore che deve assumenre la variabile. Si noti che l'operatore relazionale `==` può assumenre qualunque valore relazionale desiderato.

15.17 La funzione []

Dato un oggetto di R denominato `xxx` con `xxx[ ]` possiamo vederne tutte le sue componenti

15.18 La funzione by

La funzione `by` serve ad applicare una funzione ad un insieme di dati per ottenere risultati suddivisi per fattori di applicazione. Il suo uso è il seguente:
`by(data,indicies,fun)`
in cui si ha che:

- data è un data frame
- indicies è un fattore o una lista di fattori
- fun è la funzione da applicare

15.19 La funzione xtabs

Con il comando `xtabs` è possibile creare delle tabelle di contingenza utilizzando la modalità formula. Quindi per creare delle tabelle di cintingenza possiamo dare anche il seguente comando:
`xtabs(y x1+x2,data=....)`
In cui x1 ed x2 sono delel variabili di tipo fattoriale.

16 Importare i dati in R

16.1 Introduzione

Per poter compiere una analisi statistica dei dati è necessario caricare questi dati in R. Se i dati sono in numero ridotto ciò può essere fatto utilizzando le funzioni:

- `scan` per i vettori
- `fix` per i data frame

Molte volte i dati da analizzare sono già disponibili in file generati da programmi esterni ad R e quindi disponibili in formati di diverso tipo. L'importazione dei dati, in questo caso, può avvenire in diversi modi come documentato nell'help delle funzioni dedicate allo scopo ossia:

```
>read.table
>read.csv
>read.csv2
>read.delim
>read.delim2
>scan
```

Si ricorda in ogni caso che i file sono importabili in R se sono in formato testo, esistono però innumerevoli tipi di formato testo a seconda del modo in cui i record e i campi sono tra di loro separati. Nei prossimi paragrafi analizzeremo solamente l'importazione di alcune tipologie tra le più diffuse di file testo. Se i dati a disposizione sono in formato diverso da quelli trattati, sarà possibile importarli in R consultando l'help delle funzioni sopra descritte.

16.2 La preparazione dei dati

Se i file provengono da programmi della famiglia dei prodotti office come office, staroffice o openoffice, prima di procedere alla loro importazione sarà necessario operare una preparazione preliminare consistente in:

- eliminazione di ogni tipo di formattazione dei dati in particolare alla suddivisione in migliaia delle cifre dei numeri da importare, questa operazione è della massima importanza infatti se lasciamo i punti di separazione i dati non saranno importati in R correttamente
- una volta eliminata la formattazione per evitare problemi legati ad una diversa interpretazione dell'area dei dati da importare sarà consigliabile copiare in un nuovo foglio di lavoro i dati puliti ottenuti eliminando la formattazione e lavorare con questo nuovo foglio di lavoro
- nella prima riga di ogni colonna dovrà essere indicato, preferibilmente in minuscolo, il nome della variabile
- per i dati di tipo numerico procedere alla loro riformattazione come numeri

16.3 La funzione read.table

La funzione read table consente di importare i dati specificando tra l'altro sia il tipo di separatore di campo che il carattere utilizzato per il separatore decimale. Per utilizzare questa funzione sarà quindi necessario conoscere il separatore di campo utilizzato e il formato per la separazione dei numeri decimali utilizzato. Se i dati sono disponibili in file generati dai prodotti della famiglia office, effettuata l'operazione preliminare sui dati, per portare gli stessi un formato testo, leggibile da read table possiamo:

- se si usa il programma office i dati dovranno essere salvati nel formato testo delimitato da tabulazione.
- se si usano i programmi staroffice o openoffice i dati dovranno essere salvati nel formato `csv` usando come separatore di campo `tab`.

In questo modo viene generato un file formato testo delimitato da tabulazione e in cui il separatore decimale è la virgola. Si noti che i programmi office permettono di salvare dati in formato testo anche con modalità di tipo diverso come vedremo successivamente.

La funzione `read.table`, come detto, permette di importare un file salvato nel formato testo delimitato da tabulazione in un `data.frame`. Supponendo che il file da importare già convertito in tale formato e sia denominato `xxx.txt` per importarlo si usa la seguente sintassi:

```
\verb">xxx<-read.table("c:\\mydir\\xxx.txt",header=T,sep="\t",dec=",")
```

in cui:

- `header=T` significa che la prima riga del file sarà destinata ai nomi delle colonne del data frame
- `dec=“,“` significa che i dati numerici da importare hanno come separatore decimale la virgola
- \ significa che il separatore di colonne è il tabulatore

Si noti che è anche possibile aggiungere l'opzione:
`row.names=n`
che indica che la colonna `n` contiene le etichette delle righe del data frame, se è impostata l'opzione `header=T`, la colonna che conterrà i nomi delle righe dovrà necessariamente iniziare dalla seconda posizione ossia non si dovrà considerare la prima della della colonna.

16.4 Uso della funzione read.csv

Effettuata l'operazione preliminare sui dati, per portare gli stessi in formato testo, possiamo se si utilizza office salvarli nel formato `csv`. La funzione `read.csv` permette di importare un file salvato nel formato `csv` delimitato dal separatore di elenco in un `data.frame`. Supponendo che il file da importare già convertito in csv delimiatato dal separatore di elenco e sia denominato `xxx.csv` per importare i dati si usa la seguente sintassi:

```
>xxx<-read.csv("c:\\mydir\\xxx.csv")
```

16.5 Uso della funzione read.csv2

Effettuata l'operazione preliminare sui dati, per portare gli stessi in formato testo, possiamo se si utilizza office salvarli nel formato `csv msdos`. La funzione `read.csv2` permette di importare un file salvato nel formato `csv msdos` in un `data.frame`. Supponendo che il file da importare già convertito in csv msdos e sia denominato `xxx.csv` per importare i dati si usa la seguente sintassi:

```
>xxx<-read.csv2("c:\\mydir\\xxx.csv")
```

16.6 Conclusioni

L'importazione dei dati è una operazione alquanto complessa. Quando si ha a disposizione un file in formato testo dobbiamo conoscere e sattamente in che modo sono stati separati i record e i campi e il formato utilizzato per la virgola decimale. Conoscendo tali informazioni e con lievi modifiche a quanto scritto ei paragrafi precedenti sarà sempre possibile importare un file in formato testo contenente dati da analizzare in R.

17 Esportare i dati da R

17.1 Introduzione

Molte volte dopo aver eseguito delle analisi statistiche sarà necessario esportare i dati così ottenuti da R in un altro programma di solito un prodotto della famiglia office. Per fare questo abbiamo bisogno di una funzione che ci permetta di esportare i dati in un formato leggibile da tali prodotti. La funzione che possiamo utilizzare è `write.table`.

17.2 La funzione write.table

La funzione write table converte un qualunque oggetto di R in un file di testo importabile in programmi esterni o in qualche editor di dati. La sintassi di tale comando è la seguente:

```
write.table(x, file = "c:\\miofile.txt", row.names = TRUE, col.names = TRUE,
sep="",quote=T,dec=",")
```

in cui:

- con tale comando sono esportati anche i nomi delle righe e delle colonne, per non esportarli combiare in F nelle opzioni row.names e col.names
- tutti i dati sono esportati tra virgolette, per evitare questo fatto basterà cambiare il F l'opzione relativa a quote
- se vogliamo che i dati siano salvati nel formato testo delimitato da tabulazione bisognerà utilizzare

  ```
  "sep="\t"
  ```

- se vogliamo che i dati presentino come separatore decimale la virgola bisognerà inserire l'opzione dec=,

18 Le tavole in R

18.1 Introduzione

Molto spesso se i dati sono disponibili in vettori sarà necessario creare delle tabelle che ci permettano di ottenere le frequenze assolute o relative delle modalità con cui i dati si presentano nel singolo vettore o in vettori congiuntamente considerati. Per fare ciò si usa la funzione `table` o la funzione `ftable`.

18.2 Uso delle funzioni table e ftable

La funzione table è una funzione molto potente di R e permette di calcolare:

- se applicata ad un vettore le frequenze assolute delle modalità presenti nel vettore
- se applicata a due vettori una tabella di contingenza con le frequenze assolute delle modalità congiunte dei due vettori
- se applicata a più vettori un array con le frequenze assolute congiunte di tutti i vettori

La funzione si usa con il seguente comando:
`>table(x)`
in cui `x` è un vettore di dati.

Si noti che la funzione table si presta bene ad essere utilizzate con la funzione cut. Supponiamo di creare con il comando:
`>x<-runif(1000,1,100)`
mille numeri casuali tra 1 e 1000 e di voler vederne una tabella degli stessi ma raggruppati di 5 in 5. Ciò puo' essere effettuato con:
`>table(cut(x,breaks=(0+5*(1:20))))` in questo modo otteniamo una tabella delle frequenze di 5 in 5 dei numeri inseriti nel vettore x.

Se `x,y` sono due vettori di dati, per ottenere una tavola di contingenza dobbiamo operare nel seguente modo:
`>table(x,y)`

Per utilizzare la funzione `table` con più vettori possiamo usare la seguente sintassi:
`>table(x,y,z,....)`
in cui `x,y,z,....` ottenendo come già detto un array.

Molto usata in R è anche la funzione `ftable` la quale ha la stessa sintassi di `table` ma fornisce le tabelle in modo diverso. La sua sintassi è la seguente:
`>ftable(x)`
`>ftable(x,y)`
`>ftable(x~y)`
`>ftable(x,y,z,....)` in cui `x,y,z,...` sono dei vettori.

18.3 Uso della funzione prop.table

La funzione `prop.table` prende come argomento o una matrice o un oggetto creato con table e ne restituisce le frequenze relative. La sua sintassi è la seguente:
`>prop.table(xxx,n)`
in cui

- `xxx` è una matrice o un oggetto table
- n=1 viene calcolata la frequenza relativa sul totale generale
- n=2 viene calcolata la frequenza relativa per riga
- n=3 viene calcolata la frequenza relativa per colonna

18.4 Uso della funzione margin.table

La funzione `margin.table` prende come argomento o una matrice o un oggetto creato con table e ne restituisce le frequenze assoluta. La sua sintassi è la seguente:
`>margin.table(xxx,n)`
in cui

- `xxx` è una matrice o un oggetto table
- n non è specificato viene calcolata la somma di tutti gli elementi della tabella
- n=1 viene calcolata la frequenza assoluta per riga
- n=2 viene calcolata la frequenza assoluta per colonna

18.5 La funzione plot con un oggetto table

La funzione `plot` applicata ad un oggetto table o una matrice consente di ottenere un grafico del tipo `mosaicplot`. La sua sintassi è la seguente:
`>plot(xxx)`
in cui `xxx` è un oggetto table o una matrice

18.6 La funzione summary con un oggetto table

Se applichiamo la funzione summary su un oggetto table
`>summary(xxx)`
con `xxx` oggetto table:
si ottiene il test `chi quadrato` di contingenza sulla tabella data.

18.7 La funzione barplot con un oggetto table

Se `xxx` è un oggetto table, possiamo appliocare ad esso il comando grafico `barplot` in uno dei seguenti comandi:
`>barplot(x)`
`>barplot(x,beside=T)`
`>barplot(x,legend=T)`
ottenendo un grafico di tipo `barplot`.

19 Elementi di programmazione in R

19.1 Introduzione

Pur presentando innumerevoli funzioni, il programma R ci consente sia di creare funzioni personalizzate che degli script ossia delle successioni di istruzioniche permettono di eseguire in modo rapido una serie di comandi senza bisogno di digitarli dalla tastiera. Per fare ciò dobbiamo conoscere alcuni elementi di programmazione in R. Si noti che ogni istruzione o ogni comando visto precedentemente può essere inserito in un listato di programmazione nello stesso modo in cui noi lo abbiamo trattato. Per questo motivo vedremo successivametne solo alcuni aspetti particolari della programmazione rimandando alla documentazione in linea offerta da R gli ulteriori approfondimenti.

19.2 Come scrivere una funzione

Per scrivere e quindi memorizzare unafunzione conviene utilizzare la seguente successione di comandi:
`>nomefunzione<-function(var1=val1,var2=val2,var3=val3,...,varn=valn){}`
`>nomefunzione<-edit(nomefunzione)`
in cui:

- `nomefunzione` è il nome assegnato alla funzione
- `(var1=val1,var2=val2,var3=val3,...,varn=valn)` sono le variabili indipendenti con il relativo valore die default che la funzione dovrà utilizzare. Non è necessario impostare il valore predefinito anche se in certi casi ciò è molto utile
- `{ }` è il corpo della funzione ossia le istruzioni che devono essere eseguite al richiamo del nome delal funzione.

Un modo alternativo per creare una funzione prevede l'utilizzo della funzione `fix(x)` nel seguente modo:
`>nomefunzione<-fix(nomefunzione)`
in questo nodo si aprirà automaticamente l'editor di testo e sarà possibile inserire sia le variabili che il corpo della funzione.

19.3 Come scrivere gli script

Gli sript sono scritti in modo analogo alle funzioni solamente che non prevedono l'inserimentodi alcuna variabile. Per scriverli utilizzeremo allora la seguente successione di comandi:

```
>nomescript<-function(){ }
>nomescript<-edit(nomescript)
```

19.4 Come eseguire una funzione

Una volta creata la funzione per eseguirla dobbiamo semplicemente richiamarla con il nome seguito dal nome e dal valore delle variabili nel seguente modo:

```
>nomefunzione(var1=val1,var2=val2,var3=val3,...,varn=valn)
```

Se non inseriamo nome e valore di una variabile ad essa automaticamente sarà passato il valore di default.

19.5 Come eseguire uno script

Essendo gli script delel funzioni senza varibili, la loro esequzione avviene nel seguente modo:

```
>nomescript()
```

19.6 Le strutture di controllo

Qualunque linguaggio di programmazione prevede l'utilizzo delle strutture di controllo. Tali strutture sono le seguenti:

- if
- repeat
- while
- for
- switch

Vedremo nei paragrafi successivi il loro uso.

19.7 La struttura di confrollo if

Il suo uso è il seguente:

```
if(condizione){ }
else{ }
```

19.8 La struttura di confrollo repeat

Il suo uso è il seguente:

```
repeat{ }
```

19.9 La struttura di confrollo while

Il suo uso è il seguente:

```
while(condizione){ }
```

19.10 La struttura di confrollo for

Il suo uso è il seguente:

```
for(i in 1:n){ }
```

19.11 La struttura di confrollo switch

Il suo uso è il seguente:

```
switch(variabilecontrollo,istr1,istr2,istr3,...istrn)
```

19.12 Esempio di programmazione: equazione di secondo grado

Una semplice funzione per ottenere le soluzioni di una equazione di secondo grado è la seguente:

```
secgrad<-function(a,b,c)
           {
              xx<-0
              delta<-b^2-4*a*c
              if(delta<0)
               return("Equazione impossibile")
              else{
xx[1]<--b+sqrt(delta))/(2*a)
               xx[2]<--b-sqrt(delta))/(2*a)}
           }
```

Una versione alternativa potrebbe essere la seguente:

```
secgrad<-function(a,b,c)
          {
              delta<-b^2-4*a*c
              if(delta<0)
               return("Equazione impossibile")
              else{
x1<--b+sqrt(delta))/(2*a)
                x2<--b-sqrt(delta))/(2*a)
                list(x1=x1,x2=x2)}
          }
```

19.13 Esempio di programmazione: indici di connessione

Una semplice funzione per ottenere gli indici di conessione dato un vettore di dati è la seguente:

```
connessione<-function (x)
{
      options(warn=-1)
      dimensione<-dim(x)
      chis<-chisq.test(x)$statistic
      names(chis)<-c()
      n<-sum(x)
      chiquadrato<-1/(n*min(dimensione-1))*chis
      y<-chisq.test(x)$expected
      mortara<-1/(2*n)*sum(abs(x-y))
      options(warn=0)
     list(chiquadrato=chiquadrato,mortara=mortara)
         }
```

19.14 Esempio di programmazione: indici di mutabilità

Una semplice funzione per ottenere gli indici di mutabilità dato un vettore di dati è la seguente:

```
mutability<-function (x)
{
        h<-length(x)
        hh<-sum(x)
        p<-x/hh
        som<-sum(p*(1-p))
        gini<-h/(h-1)*som
        shannon<--sum(p*log(p))/log(h)
        list(gini=gini,shannon=shannon)
}
```

19.15 Esempio di programmazione: le medie

Una semplice funzione per ottenere le medie di un dato vettore di dati è la seguente:

```
medie<-function (x, r)
{
       if (r == 0) {
                                 prod(x)^(1/length(x))
                                }
       else {
                         pd<-r-trunc(r/2)
                         if(pd==0){
                                   mean(x^r)^(1/r)
                                 }
                         else{
                               segno<-sign(sum(x^r))
                               valore<-abs(mean(x^r))^(1/r)
                               valore*segno
                             }
          }
}
```

20 Come creare propri file di funzioni in R

20.1 Introduzione

Una volta create funzioni e script, avremmo la necessità di creare anche dei file per memorizzarle e che siano leggibili da R in modo tale da poterle anche importare ed utilizzare successivamente o di permettere anche ad altre persone di utilizzarle. Vefremo nei paragrafi successivi come creare questi file.

20.2 I file source

I file necessari per compiere queste operazioni sono i file `source`. Tali file raccolgono semplicemente una dietro l'altra le funzioni e gli script che abbiamo creato. Ad esempio il file `personal.r` contiene le funzioni e gli script:

- connessione per l'analisi della connessione
- mutability per l'analisi della mutabilità
- limite.centrale per l'analisi grafica del limite centrale
- intervalli per l'analisi grafica degli intervalli di confidenza
- medie per il calcolo delle medie potenziate
- lotto per l'estrazione casuale di numeri al lotto
- plotline per la visualizzazione di un grafico particolare

e sarà così composto:

```
  connessione<-
function (x)
{
    options(warn=-1)
    dimensione<-dim(x)
    chis<-chisq.test(x)$statistic
    names(chis)<-c()
    n<-sum(x)
    chiquadrato<-1/(n*min(dimensione-1))*chis
    y<-chisq.test(x)$expected
```

```
    mortara<-1/(2*n)*sum(abs(x-y))
    options(warn=0)
    list(chiquadrato=chiquadrato,mortara=mortara)
    }

mutability<-
function (x)
{
      h<-length(x)
      hh<-sum(x)
      p<-x/hh
      som<-sum(p*(1-p))
      gini<-h/(h-1)*som
      shannon<--sum(p*log(p))/log(h)
      list(gini=gini,shannon=shannon)
}

medie<-
function (x, r)
{
     if (r == 0) {
         prod(x)^(1/length(x))
     }
     else {
            pd<-r-trunc(r/2)
            if(pd==0){
                mean(x^r)^(1/r)
               }
            else{
               segno<-sign(sum(x^r))
               valore<-abs(mean(x^r))^(1/r)
               valore*segno
                }
          }
}


intervalli<-
function(media=4,sqm=1,alfa=0.05,nN=100,nn=20)
{
    m.sim<-0
    for(i in 1:nN)
       {
        sim<-rnorm(nn,media,sqm)
        m.sim[i]<-mean(sim)
       }
    zeta<-qnorm(1-alfa/2,0,1)
    lim<-matrix(0,2,nN)
    for(i in 1:nN)
       {
         lim[1,i]<-m.sim[i]-zeta*sqm/sqrt(nn)
         lim[2,i]<-m.sim[i]+zeta*sqm/sqrt(nn)
       }
    y<-seq(1:nN)
    y<-matrix(y,nN,2)
    lim2<-t(lim)
    plot(y,lim2)
    for(i in 1:nN)
      lines(y[i,],lim2[i,])
```

```
    lines(c(0,100),c(media,media))
    vero<-matrix(0,nN,2)
    vero<-c(lim2[,1]>media,lim2[,2]<media)
    sum(vero)

}

limite.centrale<-
function(x,n=100,min=0,max=1)
{
dd<-0
for(i in 1:x)
dd[i]<-mean(runif(n,min,max))
dd<-(dd-(min+max)/2)/((max-min)/sqrt(12*n))
aaa<-hist(dd,plot=F)$breaks
par(mfrow=c(2,2))
hist(dd,main="Istogramma delle frequenze assolute")
hist(dd,prob=T,main="Istogramma delle frequenza relative")
irq<-summary(dd)[5]-summary(dd)[1]
lines(density(dd,width=irq))
hist(dd,prob=T,main="Distribuzione media campionaria")
bbb<-seq(min(aaa),max(aaa),0.01)
lines(bbb,dnorm(bbb,0,1))
plot(ecdf(dd),verticals=T,do.p=F,main="Fz distr. emp. e rip. N(0,1)")
lines(sort(dd),pnorm(sort(dd),mean=0,sd=1))
par(mfrow=c(1,1))
}

lotto<-
function (n)
{
     x<-trunc(runif(1,0,10))+1
     y<-trunc(runif(n,0,90))+1
     z<-list(ruota=x,numeri=y)
     z
}


plotline<-
function(x,y)
{
     yy<-unique(y)
     l<-length(yy)
     media<-tapply(x,y,mean)
     sqm<-sqrt(tapply(x,y,var))
     z<-sort(media)
     z<-sort(sqm)
     plot(yy,media,xlab="Fattori",ylab="Valori",ylim=c(media[1]-2*sqm[l],
     media[l]+2*sqm[l]),
     main="variabilita' delle medie")
     rug(y)
     for (i in 1:l)
{
          segments(yy[i],media[i]-sqm[i],yy[i],media[i]+sqm[i])
       }

}
```

20.3 Importare i file source

Per importare i file source basterà digitare il seguente comando:
`>source`(directory/filesource.r)
Le funzioni in esso inserite saranno richiamabili nel modo descritto in precedenza. In windows sarà possibile richiamare tali file direttamente dal menu.

21 I grafici tradizionali

21.1 Introduzione all'uso dei grafici in R

In R abbiamo varie tipologie di grafici che verranno trattati nei paragrafi seguenti.

21.2 Il comando plot

Il comando plot è il comando base per l'analisi grafica dei dati. Se x ed y sono due vettori di uguale dimensione, con

- `>plot(x)` otteniamo un grafico avente per ordinate gli elementi del vettore x e per ascissa un vettore numerato da 1 a length(x)
- `>plot(x,y)` otteniamo un grafico formato dai punti avente come ascissa gli elementi di x e ordinata gli elementi di y

Si noti in particolare che con:

- `>plot(table(x))` si ottiene un grafico a bastoncini
- `>plot(x)` con x factor si ottiene un grafico barplot

21.3 Opzioni del comando plot

Possiamo usare il comando plot con una serie di opzioni che permettono di personalizzare il grafico in base a specifiche esigenze. Ad esempio con le seguenti opzioni possiamo:

- `main=“ “` crea un titolo nel grafico
- `sub=“ “` crea un sottotitolo nel grafico
- `xlab=“ “` crea un nome dell'asse delle x
- `ylab=“ “` crea nome dell'asse delle y
- `xlim=c(valore1, valore2)` traccia l'asse del x tra valore1 e valore2
- `ylim=c(valore1, valore2)` traccia l'asse dell y tra valore1 e valore2
- `log=“x“,log=“y“,log=“xy“` traccia i grafici con scala logaritmica come da opzioni

Con le seguenti opzioni possiamo decidere se ottenere grafici per punti o per linee o con altre forme specificate:

- `type=“p“` traccia il grafico per punti
- `type=“l“` traccia il grafico per linee
- `type=“b“` traccia il grafico per linee e punti
- `type=“o“` traccia il grafico per linnee e punti
- `type=“h“` traccia il grafico con un High-density plot
- `type=“s“` traccia il grafico con un stairstep
- `type=“n“` non traccia nulla

Se vogliamo un grafico per linee possiamo decidere anche che tipo di linea desideriamo utilizzando le seguenti opzioni:

- `lty=1` traccia il grafico con linea continua
- `lty=2` traccia il grafico con linea con punti
- `lty=3` traccia il grafico con linea con punti e tratteggiata
- `lty=4` traccia il grafico con linea con tratteggio
- `lty=5` traccia il grafico con linea con tratteggio e tre punti ogni tanto
- `lty=6` traccia il grafico con linea con due tipi di tratteggio
- `lty=7` traccia il grafico con linea con tratteggio a punti
- `lty=8` traccia il grafico con linea con tratteggio piccolo

Se vogliamo un grafico per punti, possiamo specificare il tipo di punto che vogliamo ottenere tramite l'opzione:
`pch=carattere`
in cui carattere sarà ciò che verrà stampato dal grafico. Tale opzione può anche essere eseguita nel seguente modo:
`pch=numero`
in cui numero è un numero naturale variabile da 1 a 25 e che produrrà una serie di punti come evidenziato dalla figura 1 Altre opzioni che possono essere usate sono:

- `col=numero` con numero un natuare da 1 a 8 permette di ottenere colori diversi per punti e linee del grafico
- `axes= T o F` permette di decidere se stampare o meno gli assi coordinati
- `lwd=1,....` gestisce la grossezza delle linee e dei punti

Se vogliamo ottenere solamente la formattazione del grafico e non quella degli assi dovremmo usare tali opzioni con i comandi lines e points

Si noti che in ogni caso si potrà utilizzare il comando plot senza le opzioni precedentemente scritte e successivamente con il comando:
`>title(main="...",sub="...",xlab="...",ylab="...")`
possiamo aggiungere successivamente il titolo, il sottotitolo e le etichette degli assi x ed y.

21.4 I comandi points e lines

I comandi points e lines consentono di ottenere rispettivamente punti e linee su un grafico già precedentemente tracciato. Tutte le opzioni che valgono per plot valgono anche per queste due funzioni. Sono molto utili per evidenziare nello stesso grafico linee o punti di colore, dimensioni forme o con caratteri diversi. Vediamo in che modo possiamo utilizzarli:

- ottenere punti di un grafico di colore diverso:
  ```
  >plot(x,y)
  >points(x,z,col=2)
  ```
- ottenere punti di un grafico di grandezza diversa:
  ```
  >plot(x,y)
  >points(x,z,lwd=6)
  ```
- ottenere linee di un grafico di colore diverso:
  ```
  >plot(x,y)
  >lines(x,y,col=3)
  ```
- ottenere linee di un grafico di grandezza diversa:
  ```
  >plot(x,y,type="n")
  >lines(x,y,lwd=6)
  ```

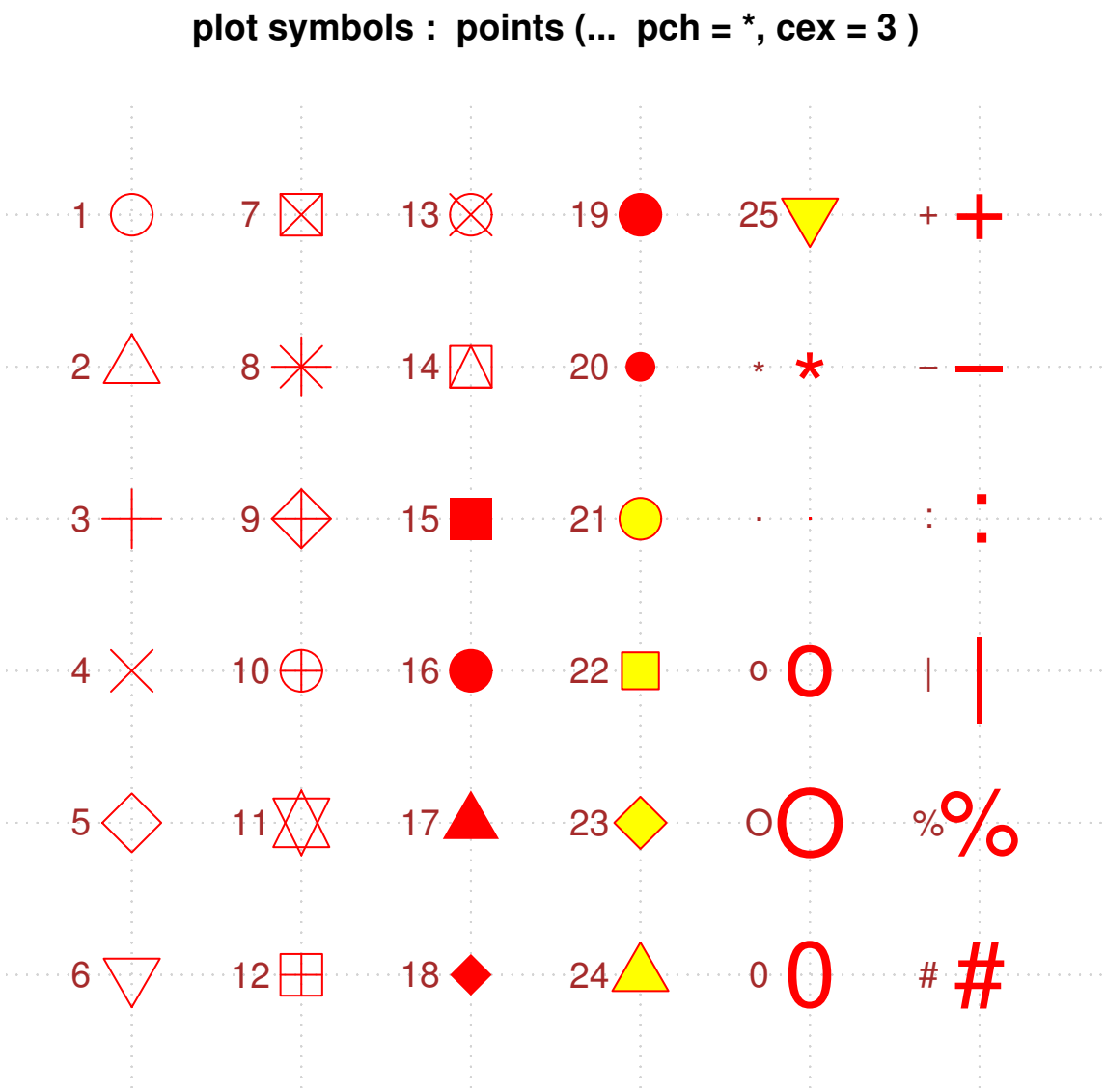

Figura 1: Simboli con pch

- ottenere linee diverse di un grafico:
  `>plot(x,y)`
  `>lines(x,z,lty=6)`

In questo caso per l'uso della legenda bisognerà fare riferimento a quanto specificato nelle funzioni points e lines esposto nel paragrafo successivo.

21.5 Aggiungere una legenda ad un grafico

Se in un grafico abbiamo introdotto attraverso i comandi lines e points punti o lineee formattati in modo doverso, sarà necessario per distinguerne il tipo inserire nel grafico stesso una legenda Questo può essere fatto in molti modi legati alla costruzione del grafico stesso come evidenziato successivamente:

- grafico creato con linee usando l'opzione lty. In questo caso potremmo usare la seguente successione di comandi:
  `>plot(x,y,type=l,lty=1)`
  `>lines(x,z,lty=2)`
  `>legend(locator(1),legend=c(primo,secondo),lty=1:2)`
  e cliccando su qualunque punto del grafico la legenda sarà li posizionata.
- grafico creato con linee usando l'opzione lwd ci si comporta in modo analogo a quanto visto con lty
- grafico creato con linee usando l'opzioni col ci si comporta in modo analogo a quanto visto con lty
- grafico creato con punti usando l'opzione pch. In questo caso potremmo usare la seguente successione di comandi:
  `>plot(x,y,type=p,pch=primopch)`
  `>points(x,z,pch=secondopch)`
  `>legend(locator(1),legend=c(primo,secondo),pch=c(primopch,secondopch)`
  e cliccando su qualunque punto del grafico la legenda sarà li posizionata.
- grafico creato con punti usando l'opzione col. In questo caso potremmo usare la seguente successione di comandi:
  `>plot(x,y,type=p,col=1`
  `>points(x,z,col=2)`
  `>legend(locator(1),legend=c(primo,secondo),pch=1,col=1:2`
  e cliccando su qualunque punto del grafico la legenda sarà li posizionata.

21.6 Scatterplot con fattori evidenziati separatamente

Supponiamo di avere a disposizione un data frame che chiameremo `aaa` che contiene tre variabili che indicheremo con `x`, `y` e `z`. Le variabili `x` ed `y` sono numeriche mentre la variabile `z` è una variabili di tipo fattore con tre livelli. Possiamo tracciare lo scatterplot della variabile `x` contro la variabile `y` ma colorato diversamente a seconda dell'appartenenza ad un fattore rispetto ad un altro con la seguente sequenza di comandi:

```
>plot(aaa$x,aaa$y,col=unclass(aaa$c))
```

Se vogliamo dei caratteri diversi invece del colore basterà semplicemente fare:

```
>plot(aaa$x,aaa$y,pch=unclass(aaa$c))
```

Un modo alternativo potrebbe anche essere il seguente, applicabile anche se z non è di tipo factor:

```
>plot(x,y,type="n")
>text(x,y,as.character(z))
```

21.7 Plot di vettori con fattori evidenziati separatamente

Se si hanno a disposizione due vettori ciascuno di lunghezza n, x numerico e y di tipo fattore, può essere necessario a volte rappresentare in uno scatter plot il vettore x in modo però da evidenziare il fattore di provenienza. Ciò può essere fatto con la sequenza di comandi:

```
>plot(1:n,x,type="n")
>text(1:n,x,as.character(y))
```

eventualmente per rendere meglio il grafico possiamo anche utilizzare l'opzione `col=.....`

21.8 Identificare il numero di posizione della coppia

Per identificare il numero di posizione della coppia si usa il la seguente successione di comandi:
`>identify(x,y,n=n)`
con la quale viene identificata la posizione di n punti tracciati nel grafico. Se l'opzione n non è inserita si continuerà a identificare numeri fino a che non si premerà il tasto esc.

Potrà anche essere inserita l'opzione: `labels=names(...)`
in cui al posto dei puntini va il nome di una variabile che potrà essere o x o y. In questo caso quando si clicca su un punto apparirà il nome della variabile scelta.

21.9 Plot e boxplot

Si noti anche che con il comando:
`>plot(x ~ y )`
in cui:

- x e un vettore di dati
- y è la corrispondente variabile di fattori

si ottiene un grafico formato da tanti boxplot quanti sono i fattori di y e naturalmente fatti con idati dei fattori corrispondenti.

21.10 Uso del comando plot due grafici

Quando devo tracciare due grafici contemporaneamente sarà necessario per tracciare il primo grafico utilizzare il comando plot e successivamente per tracciare i grafi successivi utilizzare i comandi points e lines. Il loro uso è abbastanza semplice, in ogni caso si vedano i seguenti esempi in cui x, y e z sono vettori aventi analoga lunghezza:

- per tracciare contemporaneamente due grafici con colori diversi possiamo utilizzare i comandi:
  `>plot(x,y,type="l",col=1)`
  `>lines(x,z,type="l",col=2)`
  `>legend(locator(1),legend=c("y","z"),col=c(1,2),lwd=c(1,1)`
- per tracciare contemporaneamente due grafici con linee di diversa forma possiamo utilizzare i comandi:
  `>plot(x,y,type="l",lty=1)`
  `>lines(x,z,type="l",lty=2)`
  `>legend(locator(1),legend=c("y","z"),lty=c(1,2)`
- per tracciare contemporaneamente due grafici con linee di diversa grandezza possiamo utilizzare i comandi:
  `>plot(x,y,type="l",lwd=1)`
  `>lines(x,z,type="l",lwd=2)`
  `>legend(locator(1),legend=c("y","z"),lwd=c(1,2)`
- per tracciare contemporaneamente due grafici con caratteri diversi possiamo utilizzare i comandi:
  `>plot(x,y,type="p",pch="a")`
  `>lines(x,z,type="p",pch="b")`
  `>legend(locator(1),legend=c("y","z"),pch=c("ab")`
- per tracciare contemporaneamente due grafici con caratteri e linee possiamo utilizzare i seguenti comandi:
  `>plot(x,y,type="p",pch="a")`
  `>lines(x,z,type="l",lty=1)`
  `>legend(locator(1),legend=c("y","z"),pch="a",lty=c(0,1)`

21.11 Grafico della funzione ad una variabile

Per tracciare il grafico delle funzioni ad una variabile dobbiamo operare nel seguente modo:
`>curve(sin(x),a,b)`
`>axis(1,pos=0)`

```
>axis(2,pos=0))
```
naturalmente la funzione può essere definita come visto in programmazione o scritta semplicemente in termini di x.

Si noti che se in curve aggiungiamo l'opzione add=Tla curva sarà aggiunta a lgrafico corrente.

21.12 Aggiungere una linea ai minimi quadrati

Tracciato uno scatter plot tra due vettori x ed y, per aggiungere una linea ai minimi quadrati si usa il comando:
```
>abline(lm(y~x),lty=2)
```
Si noti che la linea ai minimi quadrati può essere anche usata con l'opzione:
```
lty=...
```
molto utile per tracciare linee ai minimi quadrati provenienti da modelli differenti sullo stesso grafico.

21.13 Il grafico assocplot

Se `x` è una tabella di contingenza bidimensionale è possibile tracciare un grafico del tipo assocplot con il comando:
```
>assocplot(x)
```

21.14 Il grafico barplot

Dato un vettore numerico x e un vettore di caratteri y contenente i nomi degli elementi di x e avente la stessa lunghezza di x per per tracciare un diagramma a barre possiamo usare uno dei seguenti modi:

- `>barplot(x)` crea solamente il digramma a barre
- `>barplot(x,names=y)` crea il diagramma a barre e assegna ad ogni barra un nome
- `>barplot(x,col=1:5,legend=y)` crea il diagramma ed una legenda con il nome di ogni barra colorata
- `>barplot(x,col=1:5)`
  `>legend(locator(1),legend=y,fill=c(1:5))`
  lo stesso del punto precedente.

si noti che il vettore y al posto di essere generato prima può essere creato contestualmente alla creazione del diagramma.

Se invece x è una matrice di 2 righe e tre colonne,y un vettore contenente i nomi delle colonne e z un vettore contenente i nomi delle righe sempre di x, possiamo creare il diagramma a barra in uno dei seguenti modi:

- `>barplot(x,beside=T)` crea solamente il diagramma a barre
- `>barplot(x,names=y,legend=z,beside=T)` crea il diagramma a barre, assegna i nomi alle barre e crea una legenda
- `>barplot(x,col=c(1,2),beside=T)`
  `>legend(locator(1),legend=z,fill=c(1:2))` lo stesso del punto precedente

E' interessante usare questa opzione popo aver usato il comando tapply con più fattori, il grafico che si ottiene è molto efficace come si evince dal seguente esempio:
```
>xxx<-tapply(ragionieri$italiano,list(ragionieri$classe,ragionieri$corso)
,mean,na.rm=T)
>barplot(xxx,beside=T,names=dimnames(xxx)[[2]],legend=dimnames(xxx)[[1]],
ylim=c(0,10))
```
Il comando barplot si usa anche per tracciare dei grafici a barre, facenti funzioni di istogramma quando i dati sono raggruppati in classi non aventi la stessa dimensione. Supponiamo di dover rappresentare graficamente la seguente situazione:

classi	frequenze
120 -| 135	10
135 -| 145	20
145 -| 150	60
150 -| 165	10

Si puo' operare nel seguente modo:
`>lunghezza<-c(15,10,5,15)`
`>frequenze<-c(10,20,60,10)`
possiamo allora utilizzare il comando:
`>barplot(frequenza,lunghezza)`
così facendo però abbiamo una visione sfalsata della realtà in quanto sono le aree dei rettangoli e non le altezze che devono essere proporzionali. Per ovviare al problema conviene allora tracciare il barplot con il seguente comando:
`>fc<-frequenza/(sum(frequenza)*lunghezza)`
`>barplot(fc,lunghezza,names=c(“120-|135“,“135-|145“,“145-|150“,“150-|165|“)“`
Un modo alternativo di procedere è quello che consiste nell'operare nel seguente modo:
`>dati<-c(rep(“120-|135“,10),rep(“135-|145“,20),rep(“145-|150“,60),rep(“150-|165|“,10))“`
`>lunghezza<-c(15,10,5,15)`
`>frequenze<-table(lunghezza)`
`>fc<-frequenza/(sum(frequenza)*lunghezza)`
`>barplot(fc,lunghezza,names=c(“120-|135“,“135-|145“,“145-|150“,“150-|165|“)“`

Nell'uso del comando barplot vi sono alcune opzioni di particolare importanza. Esse sono:

- `>barplot(x,names=y,horiz=T)` le barre sono tracciate in modo orrizontale
- `>barplot(x,names=y,beside=T)` le barre non sono tracciate una sopra l'altra ma avvicinate, si usa con x matrice

21.15 Il grafico dotchart

Dato un vettore numerico x e un vettore di caratteri y contenente i nomi degli elementi di x e avente la stessa lunghezza di x per per tracciare un diagramma dotchart possiamo utilizzare il seguente comando:
`>dotchart(x,labels=y)`
Se invece z e un oggetto factor della stessa lunghezza di x possiamo utilizzare il comando:
`>dotchart(x,groups=z)`
Si noti che in questo caso `x` può essere anche matrice.

21.16 Il grafico pie

Dato un vettore numerico x e un vettore di caratteri y contenente i nomi degli elementi di x e avente la stessa lunghezza di x per per tracciare un diagramma piechart possiamo utilizzare il seguente comando:
`>pie(x,labels=y,col=1:n)`
in cui n sta a significare la lunghezza di x.

21.17 Il grafico boxplot

Dati i vettori numerici x,y e z e un vettore di caratteri w contenente i nomi dei vettori x, y e z per tracciare un diagramma boxplot possiamo utilizzare uno dei seguenti comandi:

- `>boxplot(x,names=`nomedix)
- `>boxplot(x,y,z,names=w)`

Si noti che se x è un vettore numerico ed y un vettore factor della stessa lunghezza con il comando:
`>boxplot(split(x,y))`
otteniamo tanti grafici boxplot quanti sono i livelli di y.

Un modo alternativo per usare questa funzione è quello di usarla con una formula nel seguente modo:
`>boxplot(x~y)`
in cui x è un vettore numerico ed y il corrispondente vettore di fattori.

E' anche possibile far stampare il grafico in modo orrizzontale utilizzando il seguente comando:
`>boxplot(x,horizontal=T)`

21.18 IL grafico coplot

Tramite il comando coplot è possibile ottenere uno scatter plot di un vettore numeri x contro y condizionato al valore assunto dalla variabile a . L'uso del comando è il seguente:

`coplot(y~x|a*b,data=....)`
in cui .... è il nome del data frame che contiene le tra variabili precedentemente usate.Si noti che:

- se a è factor avremo i due scatterplot esattamente divisi per i livelli di a
- si possono condizionare i valori anche a due variabili a*b

Si vedano dall'help in linea le numerosissime opzioni di tale comando.

21.19 Il grafico fourfoldplot

Se x è un array del tipo 2x2x2x....x2xk tabelle di contingenza è possibile usare il comando:
`>fourfoldplot(x)`

21.20 Il grafico hist

Dato un vettore numerico x per tracciare un diagramma hist utilizzando le frequenze assolute possiamo utilizzare uno dei seguenti comandi:

- `>hist(x)` traccia l'istogramma in modo automatico
- `>hist(x,breaks=seq(a,b,c))` traccia l'istogramma seguendo le classi indicate dal comando seq
- `>hist(x,breaks=c(a,b,c,d,e)` traccia l'istogramma considerando le classi indicate dal comando breaks
- `>hist(x,breaks=n)` traccia l'istogramma considerando n classi indicate nel comando breaks
- `>hist(x,nclass=n)` traccia l'istogramma considerando `n` classi indicate con il comando nclass

Una opzione particolarmente importante del comando hist è l'opzione probability equivalente alla opzione freq. Se infatti inseriamo in hist l'opzione:
`probability=T`
oppure `freq=F`
otteniamo l'istogramma del vettore x in base alle frequenze relative. Si usi con particolare attenzione queste opzioni nel caso in cui le ampiezze delle classi con cui si vuole costruire l'istrogramma hanno ampiezza diversa.

21.21 Il grafico density

Molto utile nell'analisi grafica delle distribuzioni dei dati è il diagramma density. Esso si ottiene con la seguente successione di comandi: `>idq<-summary(x)[5]-summary(x)[2]`
`>lines(density(x,width=2*idq))`
Tale comando inserisce in un grafico già esistente una linea che interpola graficamente una densità. Si usa molto spesso dopo aver creato il grafico hist con l'opzione probability=T.

21.22 Il grafico qqPlot

Per verificare se la la distribuzione da cui si pensa possano provenire i dati raccolti nel vettore x è una normale possiamo usare le seguenti funzioni:

- `>qqnorm(x)`
- `>qqline(x)`

Se pensiamo invece che la distribuzione da cui provengono i dati raccolti nel vettore x sia una uniforme continua possiamo usare la seguente sequenza di comandi:
`>plot(qunif(ppoints(x)),sort(x))`
Naturalmente la funzione quinif potrà essere sostituita da qualunque funzione tra le seguenti:

- qbeta con argomenti obbligatori shape1 e shape2
- qcauchy con argomenti facoltativi location scale
- qchisq con argomenti obbligatori df

- qexp con argomenti facoltativi rate
- qf con argomenti obbligatori df1 e df2
- qgamma con argomenti obbligatori shape
- qlnorm con argomenti facoltativi mean sd
- qnorm con argomenti facoltativi mean sd
- qt con argomenti obbligatori df
- qunif con argomenti facoltativi min max

Per verificare invece se due vettori x ed y provengono dalla medesima distribuzione di probabilità possiamo utilizzare il comando:
`>qqplot(x,y)`
e naturalmente x ed y devono avere lo stesso numero di dati. Se hanno numeri di dati differenti il grafico è tracciato ugualmente su dati interpolati.

21.23 Il grafico pairs

Se x è un data frame o una matrice possiamo ottenere il grafico dello scatter plot di ogni variabile verso le altre con il comando:
`>pairs(x)` Possiamo usare anche con pairs e con la stessa sintassi quanto detto nel paragrafo 21.6.

21.24 Il grafico mosaiplot

Se x è una tabella di contingenza possiamo ottenere il grafico di tipo mosaiplot con il seguente comando:
`>mosaicplot(x)`

21.25 Il grafico matplot

Se x ed y sono due matrici delle stesse dimensioni, supponiamo aventi m righe ed n colonne, con il comando maplot otteniamo n scatter plot tutti sullo stesso grafico ottenuti contrapponendo ordinatamente le colonne della prima matrice con le colonne della seconda matrice. Il comando da utilizzare sarà allora il seguente:
`>matplot(x, y, pch=“c(.....)“)`
in cui nell'opzione pch sono inseriti gli n caratteri che saranno stampati nel grafico. Se tale opzione non viene inserita la numerazione inizierà da 1 fino a raggiungere il numero delle colonne delle due matrici.

Un altro uso molto interessante del comando matplot è il seguente, se abbiamo una matrice x di m righe ed n colonne con il comando:
`>matplot(1:m,x)`
`>matplot(1:m,x,pch=“c(.....)“)`
otteniamo n scatter plot tutti sullo stesso grafico aventi tutti per per ascisse la successione 1:m e per ordinata ogni colonna di x.

Notiamo che esistono anche altri comandi della classe matplot e sono dati da:

- `>matpoints(x, y, pch=“c(.....)“)`
- `>matlines(x, y, pch=“c(.....)“)`

i quali aggiungono punti o linee ad un grafico già esistente.

Tali tipi di grafici sono usati prevalentmente per analisi statistiche multivariate per vedere i legami tra le varie distribuzioni delle variabili

21.26 Il grafico stars

E' anche possibile generare degli star plot. Se abbiamo una matrice 20x5 con il comando:
`>stars(x)`
otteniamo 20 stelle a 5 punte che permettono di vedere i legami tra le 5 varibili colonna all'interno di ciascuna riga. E' usato prevalentemente per analisi multivariate in cui le righe della matrici sono le osservazioni mentre le colonne sono le variabili

21.27 Il grafico stripchart

Possiamo ottenere un grafico del tipo stripchart in uno dei seguenti modi:
`>stripchart(list(x,y))`
`>stripchart(x~a)`
in cui x ed y sono vettori numerici, mentre a è un vettore factor.

21.28 I grafici ldahist

Dopo aver caricato il pacchetto aggiuntivo MASS è possibile ottenere degli istrogrammi per gruppi di fattori con il seguente comando:
`ldahist(x, g)`
in cui `x` è il vettore dei dati e g il corrispondente vettore dei fattori.

21.29 Il grafico scatterplot3d

Dopo aver caricato il pacchetto aggiuntivo scatterplot3d è possibile ottenere il grafico scatterplot3d con il seguente comando:
`>scatterplot3d(x,y,z)`
in cui `x,y,z` sono tre vettori di dati della stessa numerosità.

21.30 I grafici tridimensionali

Un modo molto utile per utilizzare le funzioni tridimensionali è il procedimento che consente di rappresentare graficamente una tabella di contingenza. Possiamo allora utilizzare il seguente procediemnto:

```
>w<-table(xxx,yyy)
>contour(w)
>filled.contourw)
>image(w)
>persp(w)
```

Se x ed y sono due vettori numerici che danno il range di x e di y per una funzione a due variabili, e myfunction è una funzione generata come visto nel paragrafo relativo alla programmazione per generare una grafico tridimensionale possiamo allora utilizzare la seguente sintassi:

```
>z<-outer(x,y,function)
>contour(x,y,z)
>filled.contour(x,y,z)
>image(x,y,z)
>persp(x,y,z)
```

21.31 Parti comuni

Si noti che in ogni caso dopo aver generato un grafico possiamo usare i comandi

- `>title(main, sub, xlab, ylab, axes=F)`
- `>axes(main, sub, xlab, ylab, axes=T)`

per inserire titolo, sottotitolo, nome degli assi. Si noti che se usiamo
`>title(.....)`
senza alcuna opzione è predefinita l'opzione `main`. Naturalmente dopo il nome dell'opzione ci sarà il segno di uguale e quindi sarà necessario inserire i nomi desiderati racchiusa tra virgolette.

Si noti inoltre che se non si vuole che compaiano gli assi in un grafico in qualunque modo generato si può usare l'opzione:
`>plot(1:4, rnorm(4), axes=FALSE)`
Per aggiungere invece tick in intervalli voluti possiamo operare nel seguente modo:

```
>plot(seq(20,75,5),cumsum(fi))
>axis(1, seq(20,75,5), as.character(seq(20,75,5)))
```

in questo modo le etichette volute compariranno nell'asse delle `x`. Per farle comparire nell'asse delle `y` basterà sostituire con il `axis` l'opzione 2. Tale comando può essere usato anche per inserire etichette non numeriche.

21.32 Il comando rug

Dopo aver tracciato un grafico, molto utile è anche lanciare il seguente comando:
`>rug(x)`
il quale mette delle tacchette sull'asse delle x del grafico in corrispondenza dei valori presenti nel vettore x. Per inserire il corrispondente anche nell'asse delle y si dovrà usare il comando:
`>rug(x,side=2)`

21.33 Il comando locator

Una volta creato un grafico possiamo agire in modo interattivo su di esso con il comando:
`>locator(n=n)` con il quale è possibile cliccando n volte in qualunque punto del grafico ottenere una lista di due elementi contenente le coordinate degli n punti selezionati.

Il comando locator può anche essere utilizzato con la seguente sintassi:
`>locator(n=n,type=“ “)`
in cui

- se tra le virgolette vi è l viene aggiunta nel grafico una linea congiungente i punti in cui si è cliccato con il mouse
- se tra le virgolette è inserito p vengono aggiunti nel grafico n punti in corrispondenza a dove si è cliccato con il mouse
- se tra le virgolette è inserito o vengono inseriti nel grafico sia gli n punti dove si clicca con il mouse che una linea che li congiunge
- se tra le virgolette è inserito b si ottiene com il parametro o ma la linea non tocca esattamente il punto creato

si noti che questo comando è utilissimo per evidenziare i punti già tracciati nel grafico e si usa spesso dopo il comando plot.

21.34 Aggiungere un testo in un grafico

In alcuni casi è necessario dopo aver costruito un grafico, aggiungere allo stesso un testo in punti particolari. Questo può essere effettuato con il comando:
`>text(locator(1),“....“)`
con il quale è possibile aggiungere il testo “....“ nel punto dove si fa click con il mouse.

E' possibile anche aggiungere più testi in punti specificati, si veda in particolare l'esempio seguente:
`>plot(1:n,x,type=“n“)`
`>text(1:55,x,as.character(y))` con x vettore numerico ed y vettore di caratteri della stessa lunghezza.

21.35 I grafici multipli nella stessa pagina

E' anche possibile far comparire più grafici su di una stessa pagina. Per fare ciò possiamo utilizzare due distinti procedimenti. Il primo consiste nell'utilizzare il comando:
verb»par(mfrow=c(m,n))
il quale permette di realizzare un numero d imxn grafici su di una pagina sola. Una volta dato questo comando ogni volta che si da un comando grafico questo sarà automaticamente aggiunto alla pagina secondo la scansione proposta. Il secondo comando fa riferimento alla funzione screen la quale prevede il seguente gruppo di comandi:
`>split.screen(c(m,n))`
in quale permette di realizzare una pagina con un numero di mxn grafici.
`>screen(n)`
con il quale si decide in quale parte dello schermo deve essere disegnato il grafico prescelto.

22 Grafici Trellis

22.1 Uso dei grafici trellis

I grafici trellis sono grafici utilissimi per tracciare diagrammi a multipannello. Inserendo nel comando per generare il grafico trellis una delle seguente opzioni

- `|a`
- `|a*b*...`

in cui a, b, ecc possono essere sia fattori che vettori numerici, si otterranno dei grafici multipannello con le condizioni specificate. Per poter essere utilizzati necessitano del caricamento dei pacchetti `grid` e `lattice`. In questo paragrafo prenderà il nome di condizione un vettore che potrà essere o numerico o factor.

Se a,b,c, sono numeri discreti ossia valori che si ripetono sempre uguali il condizionamento è fatto per ogni singolo valore considerando come dati il numeri di volte in cui esso si è ripetuto.

Se a,b,c sono numeri reali ma non discreti nel senso che non hanno valori che si ripetono uguali tra di loro ma sono costruiti da valori diversi dovremmo considerare il condizionamento rispetto ad un intervallo in quanto in caso contrario il grafico che otterremo sarebbe privo di significato. In questo caso prima di effettuare il condizionamento dobbiamo creare gli intervalli con:
`k <-equal.count(x,number=n,overlap=m)`
in cui x è la variabile condizionante, n è il numero delle classi ed m è il numero dei punti condivisi con intervalli successivi. Si noti che con tale comando il programma tenta di ripartire in n classi lo stesso numero di osservazioni. Fatto ciò sarà la variabile k che dovremmo usare nel condizionamento. Molto importanti sono le seguenti istruzioni:

- `>range(k)`
- `>plot(k)`
- `>levels(k)`

In tale tipo di grafici le variabili possono essere chiamate solamente con il nome senza inserire davanti il nome del data frame o matrice in cui esse sono inserite. Sarà però necessario usare l'opzione:
`>data=......`
all'interno del comando trellis utilizzato per creare il grafico. Naturalmente `......` è il nome del data frame o matrice che contiene le variabili usate per generare il grafico.

Usando i grafici Trellis, viene generato come detto un notevole numero di grafici in base al condizionamento proposto. Per evitare di ottenere grafici illeggibili possiamo con l'opzione:
`>layout=c(n,m,k)`
verranno inserite in k pagine $n * m$ grafici. Naturalmente questa opzione deve essere inserita all'interno del comando usato per generare il grafico trellis desiderato.

22.2 I pacchetti necessari

Per poter utilizzare i grafici trellis in R sarà necessario installare i pacchetti:
`grid`
`lattice`
solo in questo modo essi potranno essere tracciati

22.3 Il grafico xyplot

Per usare il comando xyplot possiamo usare uno dei seguenti modi:

- `>xyplot(numeric1 ~ numeric2|condizione)`
- `>xyplot( ~ numeric|condizione)`

Si noti anche che è possibile ottenere il grafico di una sola parte dei dati numerici considerati nel seguente modo:

- `>xyplot(y[x<1] ~ x[x<1]|condizione,data=z)`
- `>xyplot(y ~ x|condizione ,data=z,subset=x<1)`

22.4 Il grafico bwplot

Per usare il comando bwplot possiamo usare uno dei seguenti modi:

- `>bwplot(factor ~ numeric|condizione)`
- `>bwplot( ~ numeric|condizione)` ma in questo caso la condizione non deve essere un factor

Se la variabile factor usata non è un fattore viene automaticamente convertita in fattore dal programma.

22.5 Il grafico stripplot

Per usare il comando stripplot possiamo usare uno dei seguenti modi:

- `>stripplot(factor ~ numeric|condizione)`
- `>stripplot( ~ numeric|condizione)` ma in questo caso la condizione non può essere un factor

Se la variabile factor usata non è un fattore viene automaticamente convertita in fattore dal programma.

22.6 Il grafico qq

Per usare il comando qq possiamo usare il seguente modo:
`>qq(factor ~ numeric|condizione)`
E si noti che la variabile factor deve essere fattore o numerica con esattamente due livelli. Nel caso in cui la variabile factor avesse più livelli per usare qq dovremmo indicarne solo due operando allora nel seguente modo:
`>qq(factor ~ numeric|condizione  ,subset(factor==`“a“)|factor==“b“))

22.7 Il grafico dotplot

Per usare il comando dotplot possiamo usare uno dei seguenti modi:

- `>dotplot(factor ~ numeric|condizione)`
- `>dotplot( ~ numeric|condizione)` ma in questo caso la condizione non può essere un factor

Se la variabile factor usata non è un fattore viene automaticamente convertita in fattore dal programma.

22.8 Il grafico qqmath

Per usare il comando qqmath possiamo usare uno dei seguenti modi:
`>qqmath( ~ numeric|condizione,distribution= function(p)qt(p,df=7)))`
`>qqmath(~numeric|condizione,subset=(y==`“a“),distribution=function(p)qt(p,df=7)))
in cui y è una variabile factor esistente nel data frame originario.

Si noti che il valore di distribution che finora è stato supposto qt, è quello visto parlando dei grafici tradizionali ossia:

- qbeta con argomenti obbligatori shape1 e shape2
- qcauchy con argomenti facoltativi location scale
- qchisq con argomenti obbligatori df
- qexp con argomenti facoltativi rate
- qf con argomenti obbligatori df1 e df2
- qgamma con argomenti obbligatori shape
- qlnorm con argomenti facoltativi mean sd
- qnorm con argomenti facoltativi mean sd
- qt con argomenti obbligatori df
- qunif con argomenti facoltativi min max

22.9 Il grafico barchart

Per usare il comando barchart possiamo usare uno dei seguenti modi:

- `>barchart(factor ~ numeric|condizione)`
- verb»barchart(numeric|condizione) ma in questo caso la condizione non può essere un factor

Se la variabile factor usata non è un fattore viene automaticamente convertita in fattore dal programma.

22.10 Il grafico histogram

Per usare il comando histogram possiamo usare il seguente modo:
h>`histogram( ~  numeric|condizione)`
Possiamo naturalemnte inserire anche le seguenti opzioni

- `breaks` come nel caso nei grafici non trellis
- `nint` per specificare il numero delle classi
- `type`="percent" o "count" se si vuole il grafico percentualizzato o di frequenza assolute.

22.11 Il grafico densityplot

Per usare il comando densityplot possiamo usare il seguente modo:
`>densityplot( ~  numeric|condizione)`

22.12 Il grafico splom

Per usare il comando splom possiamo usare il seguente modo:
`>splom( ~ dataframe|condizione)`

22.13 Il grafico parallel

Per usare il comando parallel possiamo usare il seguente modo:
`>parallel( ~ dataframe|condizione)`

22.14 Il grafico rfs

Il grafico rfs consente di ottenere due grafici per visualizzare i valori residui e quelli stimati con il modello. La sintassi da usare è la seguente:
`>rfs(model)`
in cui model è un oggetto ottenuto con la regressione o l'analisi della varianza.

22.15 Grafici a più dimensioni

Avendo a disposizione tre vettori con la stessa numerosità x, y e z possiamo rappresentare una relazione tra queste tre variabili tramite grafici di tipo tridimensionali di tipo trellis. Un primo modo di rappresentare questi dati è quello di usare la seguente sintassi:

```
>cloud(z ~ x*y|condizione)
>levelplot(z~x*y|condizione)
```

si ricorda inoltre sempre la possibilità se x, y e z sono variabili di un data frame di utilizzare i nomi delle stesse variabiabili con l'opzione `data=.....`.

Anche i trellis grafici permettono di ottenere grafici di funzioni a due variabili. Dobbiamo prima di tutto inizializzare le varibili nel seguente modo:

```
>x<-rep(seq(-1.5,1.5,length=50),50)
>y<-rep(seq(-1.5,1.5,length=50),each=50)
>z<-exp(-(x^2+y^2+x*y))
>w<-data.frame(x,y,z)
```

Possiamo allora usare i grafici trellis tridimensionali nel seguente modo:

```
>cloud(z ~ x*y|condizione,data=w)
```

in cui condizione come al solito è un vettore numeric o factor.

23 Aggiungere del testo ad un grafico

23.1 Introduzione

In molti casi è necessario inserire in un grafico un testo qualsiasi per evidenziare un punto, una linea o per altri motivi. Questo può essere fatto in modo molto semplice in R attraverso il comando *text*.

23.2 Il comando text

Creato un grafico con il comando *text* possiamo aggiungere un testo qualsiasi in un punto qualsiasi del grafico con la seguente sintassi:
`>text(x, y, labels =“.....“ )`
`>text(locator(n),labels =“.....“ )`
in cui x ed y sono vettori o scalari ,nel caso di inserimento di un solo testo, che mi danno le coordinate dei punti dove inserire il testo.

24 Aggiungere del testo matematico ad un grafico

24.1 Introduzione

In molti casi è necessario aggiungere legende e testi matematici ad un grafico precedentemente creato o aggiungerne ad uno che stimo creando. Tutte le volte che ciò è richiesto da comandi tipo *main*, *xlab*, *ylab* e altri possiamo con un comando apposito introdurre formule matemtiche a ltesto così creato.

24.2 L'opzione expression

Il comando expression si usa ad esempio ne lseguente modo:
`>text(locator(1),expression(pi))`
`>title(main=expression(pi^2))`
se i grafici sono già stati precedentemente creati o con ad esempio:
`>plot(x,y,main=expression(pi^2)`
in sede di creazione del grafico.

24.3 Quali sono i simboli matematici inseribili

Un elenco dei simboli matematici inseribili si ottiene lanciando il demo di R con il comando:
`>demo`(“plotmath“)

I simboli matematici inseribili sono evidenziati nei grafici 2, 3, 4, 5 e 6.

25 Le Variabili aleatorie fondamentali dell'inferenza statistica

25.1 Elenco delle variabili aleatorie fondamentali

L'inferenza statisitca fa continuamente uso di variabili aleatorie, fondamentali in tale tipo di analisi. Le varibili che possono essere utilizzate sono le seguenti:

- normale indicata con `norm`
- beta indicata con `beta`
- cauchy indicata con `cauchy`
- di chiquadrato indicata con `chisq`
- esponenziale indicata con `exp`
- f indicata con `f`
- gamma indicata con `gamma`
- lognormale indicata con `lnorm`
- logistica indicata con `logis`
- t di studentindicata con `t`
- uniforme continua indicata con `unif`
- weibull indicata con `weibull`
- binomiale indicata con `binom`

Arithmetic Operators		Radicals	
x + y	$x+y$	sqrt(x)	$\sqrt{x}$
x – y	$x-y$	sqrt(x, y)	$\sqrt[y]{x}$
x * y	xy	Relations	
x/y	x/y	x == y	$x=y$
x %+–% y	$x\pm y$	x != y	$x\neq y$
x%/%y	$x\div y$	x < y	$x<y$
x %*% y	$x\times y$	x <= y	$x\leq y$
–x	$-x$	x > y	$x>y$
+x	$+x$	x >= y	$x\geq y$
Sub/Superscripts		x %~~% y	$x\approx y$
x[i]	x_i	x %=~% y	$x\cong y$
x^2	x^2	x %==% y	$x\equiv y$
Juxtaposition		x %prop% y	$x\propto y$
x * y	xy	Typeface	
paste(x, y, z)	xyz	plain(x)	x
Lists		italic(x)	*x*
list(x, y, z)	x, y, z	bold(x)	**x**
		bolditalic(x)	***x***

Figura 2: Prima lista

Ellipsis		Arrows	
list(x[1], ..., x[n])	$x_1, \ldots, x_n$	x %<–>% y	$x \leftrightarrow y$
x[1] + ... + x[n]	$x_1 + \cdots + x_n$	x %–>% y	$x \rightarrow y$
list(x[1], cdots, x[n])	$x_1, \cdots, x_n$	x %<–% y	$x \leftarrow y$
x[1] + ldots + x[n]	$x_1 + \ldots + x_n$	x %up% y	$x \uparrow y$
Set Relations		x %down% y	$x \downarrow y$
x %subset% y	$x \subset y$	x %<=>% y	$x \Leftrightarrow y$
x %subseteq% y	$x \subseteq y$	x %=>% y	$x \Rightarrow y$
x %supset% y	$x \supset y$	x %<=% y	$x \Leftarrow y$
x %supseteq% y	$x \supseteq y$	x %dblup% y	$x \Uparrow y$
x %notsubset% y	$x \not\subset y$	x %dbldown% y	$x \Downarrow y$
x %in% y	$x \in y$	**Symbolic Names**	
x %notin% y	$x \notin y$	Alpha – Omega	$A - \Omega$
Accents		alpha – omega	$\alpha - \omega$
hat(x)	$\hat{x}$	infinity	∞
tilde(x)	$\tilde{x}$	32 * degree	32°
ring(x)	$\mathring{x}$	60 * minute	60′
bar(xy)	$\overline{xy}$	30 * second	30″
widehat(xy)	$\widehat{xy}$		
widetilde(xy)	$\widetilde{xy}$		

Figura 3: Seconda lista

Style	
displaystyle(x)	$\displaystyle x$
textstyle(x)	$\textstyle x$
scriptstyle(x)	$\scriptstyle x$
scriptscriptstyle(x)	$\scriptscriptstyle x$
Spacing	
x ~ ~y	$x\ \ y$

x + phantom(0) + y	$x++y$
x + over(1, phantom(0))	$x+\overset{1}{}$
Fractions	
frac(x, y)	$\frac{x}{y}$
over(x, y)	$\over{x}{y}$
atop(x, y)	$\atop{x}{y}$

Figura 4: Terza lista

Big Operators	
sum(x[i], i = 1, n)	$\sum_{1}^{n} x_i$
prod(plain(P)(X == x), x)	$\prod_{x} P(X = x)$
integral(f(x) * dx, a, b)	$\int_{a}^{b} f(x)dx$
union(A[i], i == 1, n)	$\bigcup_{i=1}^{n} A_i$
intersect(A[i], i == 1, n)	$\bigcap_{i=1}^{n} A_i$
lim(f(x), x %–>% 0)	$\lim_{x \to 0} f(x)$
min(g(x), x >= 0)	$\min_{x \geq 0} g(x)$
inf(S)	$\inf S$
sup(S)	$\sup S$

Figura 5: Quarta lista

Grouping	
(x + y) * z	$(x+y)z$
x^y + z	$x^{y}+z$
x^(y + z)	$x^{(y+z)}$
x^{y + z}	x^{y+z}
group("(", list(a, b), "]")	$(a, b]$
bgroup("(", atop(x, y), ")")	$\binom{x}{y}$
group(lceil, x, rceil)	$\lceil x \rceil$
group(lfloor, x, rfloor)	$\lfloor x \rfloor$
group("|", x, "|")	$\|x\|$

Figura 6: Quinta lista

- geometrica indicata con `geom`
- ipergeometrica indicata con `hyper`
- binomiale negativa indicata con `nbinom`
- poisson indicata con `pois`
- wilxon indicata con `wilcox`

Esistono altri pacchetti aggiuntivi come il pacchetto `SuppDists` i quali forniscono altre varibili aleatorie utilizzabili per analisi statistiche particolari.

25.2 Uso delle variabili aleatorie fondamentali

Le variabili aleatori fondmanetali non possono essere usate semplicemente digitando i loro nome ma tale nome deve essere fatto precedere da un prefisso che potrà essere:

- `p` per ottenere la funzione di distribuzione della variabile
- `d` per ottenere la funzione di probabilità della variabile
- `q` per ottenere i quantili della variabile
- `r` per ottenere numeri casuali derivanti dalla variabile

Naturalmente a seconda del prefisso inserito, per ottenere il valore desiderato, dovremmo indicare alcuni parametri dati da:

- se il prefisso inserito è `p` dovremmo indicare il quantile o il vettore dei quantili dei quali desideriamo ottenere la funzione di ripartizione
- se il prefisso inserito è `d` dovremmo indicare un numero o un vettore di numeri del quale ottenere la funzione di densità
- se il prefisso inserito è `q` indicare la coda inderiore di probabilità o il vettore delle code superiori di probabilità dei quali desideriamo ottnere i quantili
- se il prefisso inserito è `r` dovremmo indicare un numero intero indicante la numerosità dei numeri casuali che vogliamo ottnerei

25.3 Argomenti addizionali

Oltre al prefisso e ai parametri indicati nei paragrafi precedenti, per ottenere un valore dalle variabili aleatorie dovremmo indicare alcuni argomenti addizionali diversi da variabile aleatoria a variabile aleatoria. L'elenco completo di tali argomenti addizionali è il seguente:

- `norm` richiede mean=, sd=
- `beta` richiede shape1=, shape2=, ncp=
- `cauchy` richiede location=, scale=
- `chisq` richiede df=, ncp=
- `exp` richiede rate=
- `f` richiede df1=, df2=, ncp=
- `gamma` richiede shape=, scale=
- `lnorm` richiede meanlog=, sdlog=
- `logis` richiede location=, scale=
- `t` richiede df=, ncp=
- `unif` richiede min=, max=

- `weibull` richiede shape=, scale=
- `binom` richiede size=, prob=
- `geom` richiede prob=
- `hyper` richiede m=, n=, k=
- `nbinom` richiede size=, prob=
- `pois` richiede lambda=
- `wilcox` richiede m=,n=

indipendentemente dal prefisso utilizzato. Per ulteriori informazioni si consiglia di utilizzare per le varibili viste l'help in linea.

26 Utilizzo dei grafici nell'inferenza statistica

26.1 Grafici per la verifica della normalità dei campioni

Molti dei test inferenziali che vedremo scessivamente si basano sull'assunzione della normalità della popolazione da cui proviene il campione. Per la verifica di tale ipotesi, oltre a quanto visto nel capitolo riguardante l'adattamento dei dati ad una particolare distribuzione possiamo utilizzare una serie di grafici dati da:

- `>hist(x)`
- `>boxplot(x)`
- `>idq<-summary(x)[5]-summary(x)[3]`
  `>plot(density(x,width=2*idq),xlab=`“x“,ylab=““,type=“l“)
- `>qqnorm(x)`
- `>qqline(x)`

Tali grafici possono facilmente essere ottenuti in modo automatico con la creazione di una opportuna funzione.

26.2 Grafici per la verifica di correlazione nel campione

I test inferenzaili che utilizzeremo presuppongono sempre che il campione sia un campione di tipo bernulliano e che quindi i dati campionari non siano tra di loro correlati. Per la verifica di tale fatto dopo aver caricato il pacchetto aggiuntivo:
`ts`
possiamo utilizzare i seguenti grafici:

- `>ts.plot(x)`
- `>acf(x)`

Tali grafici possono facilmente essere ottenuti in modo automatico con la creazione di una opportuna funzione.

26.3 Grafici per la verifica della correlazione tra due campioni

I test inferenzali che implicano due campioni presuppongono sempre che tra i due campioni non vi sia traccia di correlazione. Per la verifica di tale fatto dopo aver caricato il pacchetto aggiuntivo:
`ts`
possiamo utilizzare i seguenti grafici:

- `>plot(x,y)`
- `>identify(x,y,n=...)` per identificare i valori anomali

- `>acf(x)`
- `>acf(y)`

Tali grafici possono facilmente essere ottenuti in modo automatico con la creazione di una opportuna funzione.

27 Test inferenziali

27.1 I pacchetti necessari

Per poter utilizzare i test inferenziali descritti in questo paragrafo sarà necessario avere installato il pacchetto aggiuntivo:
`ctest`
che di solito è già presente e operante in qualunque distribuzione del programma.

28 Verifica di ipotesi su una variabile casuale normale

28.1 Introduzione

In tutto il presente paragrafo faremo esclusivamente riferimento a campioni bernulliani estratti da popolazioni normalmente distribuite. Ci troviamo quindi nell'ambito della statistica parametrica.

28.2 Verifica di ipotesi sulla media

La verifica di ipotesi sulla media della popolazione da cui proviene un campione nel caso in cui la varianza di tale popolazione sia sconosciuta è effettuato tramite il `test t di student`. Tale test consente di verificare le seguenti ipotesi:

- se la media della popolazione da cui proviene il campione è pari ad un valore prefissato
- se la media delle popolazioni da cui sono estratti due campioni sono uguali tra di loro
- se un certo trattamento effettuato sulle unità campionarie dopo l'estrazione ne ha modificato la media

Tale test ha la seguente sintassi:
>t.test(x,y=null,alt="two.sided",mu=0,paired=F, var.equal=F,conf.level=0.95)
Notiamo allora che:

- `x` ed `y` sono vettori numerici e rappresentano i dati raccolti con l'analisi campionaria. Se è introdotto un unico vettore il test riguarderà la verifica della media della popolazione da cui il campione è estratto, se invece sono introdotti due vettori, il test potrà riguardare la verifica dell'ipotesi dell'uguaglinza delle medie delle due popolazioni da cui sono estratti i campioni. In questo caso possiamo anche considerare l'analisi per dati appaiati modificando il valore di `paired` e portandolo a `TRUE`
- nel caso in cui si voglia cambiare il valore dell'ipotesi nulla da testare basterà cambiare il valore di `mu` al valore desiderato
- nel caso si voglia cambiare il livello di significatività del test sarà necessario cambiare il valore di `conf.level` portandolo al valore desiderato
- l'opzione `alt` può essere cambiata anche in `greater` e `less` a seconda che interessi la coda superiore o quella inferiore
- nel caso di due campioni il test ipotizza sempre che la varianza da cui provengono i campioni sia diversa, e quindi per il corretto uso del test sarà necessario modificare l'opzione in `var.equal=T`. Eè buona norma in questo caso far precedere tale test dal test `var.test` che testa se le varianze delle popolazioni da cui provengono i campioni sono uguali oppure diverse.

28.3 Verifica dell'ipotesi di uguaglianza della media: caso dei confronti multipli

Avendo a disposizione `k` campioni provenienti da `k` popolazioni diverse, a volte è necessario effettuare un test per verificare l'uguaglianza delle medie mediante confronti multipli effettuati tra i `k` campioni prendendone sempre due a due in modo da considerare tutte i possibili accoppiamenti. Si tratta allora di effettuare un test `t di student` ripetuto più volte. Il comando per eseguire tali confronti nultipli è il seguente:
`>pairwise.t.test(x, g, p.adjust.method=p.adjust.methods, pool.sd=TRUE,...)`
in cui abbiamo che:

- `x` è il vettore dei dati
- `g` rappresenta il vettore dei fattori ossia un oggetto factor che indicherà con numeri o lettere l'appartenenza dei dati di x ad un determinato campione
- `p.adjust` indica il metodo da utilizzare che può essere dato da
  - holm"
  - hochberg"
  - bonferroni"
  - none"

Nel paragrafo 36.6 vedremo un altro modo per effettuare tali confronti multipli.

28.4 Verifica dell'ipotesi di uguaglianza della media: caso del confronto globale

Avendo a disposizione `k` campioni provenienti da `k` popolazioni diverse, a volte è necessario effettuare un test per verificare l'uguaglianza delle medie mediante un confronto globale effettuato simultaneamente sui `k` campioni. Si tratta di effettuare un test noto come analisi della varianza che verrà studiato nel paragrafo 36. Senza entrare in dettagli vediamo che l'analisi della varianza viene eseguita con il seguente comando:
`>anavar<-aov(x~g)` oppure
`>anavar<-aov(x~g,data=...)`
`>summary(anavar)`
in cui si avrà che:

- `x` il vettore dei dati
- `g` rappresenta il vettore dei fattori ossia un oggetto factor che indicherà con numeri o lettere l'appartenenza dei dati di x ad un campione

28.5 Il test sulla uguaglianza delle varianze di due popolazioni

Avendo a disposizione 2 campioni provenienti da 2 popolazioni diverse, ci si pone il problema di verificare se le varianze delle popolazioni da cui tali campioni provengono sono tra di loro uguali. Tale verifica, come detto, è preliminare all'applicazione del test `t di student` nel caso di due campioni. Il test per eseguire questa verifica viene eseguito con il seguente comando:
`>var.test(x,y,alt=`"two.sided",conf.level=0.95)
in cui

- `x` ed `y` sono vettori numerici e rappresentano i dati raccolti con l'analisi campionaria nei due campioni
- nel caso si voglia cambiare il livello di significatività del test sarà necessario cambiare il valore di `conf.level` portandolo al valore desiderato
- l'opzione `alt` può essere cambiata anche in `greater` e `less` a seconda che interessi la coda superiore o quella inferiore

Si ricordi che se tale test non permette di accettare l'ipotesi nulla, non potremmo eseguire il test t di student almeno in presenza di piccoli campioni.

28.6 Il test sulla uguaglianza delle varianza di più di due popolazioni

Avendo a disposizione `k` campioni provenienti da `k` popolazioni diverse ci si pone il problema di verificare se le varianze delle popolazioni da cui provengono tali campioni siano tra di loro uguali. Ciò può esserre effettuato con il `test di  barlett`. Tale test viene eseguito con il seguente comando:
`>bartlett.test(x, g, ...)`
oppure con il seguente comando:
`>bartlett.test(formula, data, subset,na.action,...)`
in cui:

- `x` è il vettore dei dati
- `g` rappresenta il vettore dei fattori ossia un oggetto factor che indicherà con numeri o lettere l'appartenenza dei dati di x ad un determinato campione
- `formula` rappresenta una formula intesa nel seguente modo `y~g`

Si ricordi che se tale test non permette di accettare l'ipotesi nulla, non potremmo eseguire l'analisi della varianza sui campioni in precedenza raccolti.

28.7 Il test sulla correlazione

Avendo a disposizione 2 campioni provenienti da 2 popolazioni diverse, ci si pone il problema di verificare se le popolazioni da cui provengono tali campioni sono tra di loro correlate. Ciò può esserre effettuato con il `test di correlazione`. Tale test viene effettuato con il seguente comando:
`>cor.test(x,y,alt=`“two.sided“,method=“pearson“,conf.level = 0.95)
Notiamo allora che:

- `x` ed `y` sono vettori numerici e rappresentano i dati raccolti con l'analisi campionaria nei due campioni
- nel caso si voglia cambiare il livello di significatività del test sarà necessario cambiare il valore di `conf.level` portandolo al valore desiderato
- l'opzione `alt` può essere cambiata anche in `greater` e `less` a seconda che
- il valore di `method` utilizzato è quello di `Pearson` e quindi per effettuare il test di correlazione non deve essere modificato

Con lo stesso comando ma cambiando l'opzione `method` possiamo eseguire i test non parametrici o con distribuzione libera come vedremo nel paragrafo 29.6.

28.8 Funzione potenza per il test t

La funzione potenza indica la probabilità di rifiutare l'ipotesi nulla al variare del parametro ignoto oggetto del test. La funzione potenza per il test `t di student` può essere eseguita con il comando:

```
power.t.test(n=NULL, delta=NULL, sd=1, sig.level=0.05, power=NULL, type=
c("two.sample","one.sample", "paired"),alternative=c("two.sided", "one.sided"))
```

Si avrà che:

- `n` indica la numerosità del campione o dei campioni nel caso di test con due campioni
- delta indica la differenza tra il valore di `mu` utilizzato nel test t di student e quello variabile in base al quale si vuole calcolare il valore della funzione potenza
- `sd` indica lo scarto quadratico medio del campione o dei due campioni
- `sig.level` indica il livello di significatività ossia l'errore di prima specie
- `power` indica il valore della funzione potenza
- `type` indica se la funzione potenza è calcolata su test t di student calcolato su unico campione, su due campioni o su due campioni con dati appaiati
- `alternative` indica se la funzione potenza è calcolata su un test t di student in cui l'ipotesi alternativa era two.sided o one.sided

si noterà certamente che vei valori `n,delta,sd,sig.level,power` ne dovrà mancare solamente uno e il test restituirà proprio il valore del paramentro non introdotto in funzione degli altri.

Per tracciare il grafico della funzione potenza dobbiamo distinguere i seguenti casi:

- se `alternative=`"one" e se il sistema di ipotesi è il seguente:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

per ottenere il grafico dobbiamo operare nel seguente modo:

```
xxx<-power.t.test(delta=seq(-1,1,0.1),sd=1, n=25,type="one",
alternative="one")$power
plot(seq(-1,1,0.1),xxx)
```

- se `alternative=`"one" e se il sistema di ipotesi è il seguente:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$$

per ottenere il grafico dobbiamo operare nel seguente modo:

```
xxx<-power.t.test(delta=seq(-1,1,0.1),sd=1, n=25,type="one",
alternative="one")$power
plot(seq(-1,1,0.1),sort(xxx,decreasing=T)
```

- se `alternative=`"two" per ottenere il grafico dobbiamo operare nel seguente modo:

```
xxxpos<-power.t.test(delta=seq(0,1,0.1),sd=1, n=25,type="one",
alternative="""two""")$power
xxxneg<-power.t.test(delta=seq(0,1,0.1),sd=1, n=25,type="one",
alternative="two")$power
plot(c(seq(-1,0,0.1),seq(0,1,0.1)),c(sort(xxxneg,decreasing=T),xxxpos))
```

Analogamente si ragione nel caso in cui `type=`"two".

29 Verifica di ipotesi su variabili con distribuzione libera

29.1 Introduzione

In tutto questo capitolo si presuppone di non conoscere o di non avere alcuna informazione circa la popolazione da cui i campioni sono estratti. Ci troviamo quindi nell'abito della statistica non parametrica.

29.2 Verifica di ipotesi sulla mediana

La verifica di ipotesi sulla mediana della popolazione da cui proviene un campione in ambito non parametrico è effettuata tramite il `test di wilcoxon`. Tale test consente di verificare le seguenti ipotesi:

- se la mediana della popolazione da cui proviene il campione è pari ad un valore prefissato
- se la mediana delle popolazioni da cui sono estratti due campioni sono uguali tra di loro
- se un certo trattamento effettuato sulle unità campionarie dopo l'estrazione ne ha modificato la mediana

Tale test ha la seguente sintassi:
>wilcox.test(x,y=null,alt="two.sided",mu=0,paired=F, var.equal=F,conf.level=0.95)
Notiamo allora che:

- `x` ed `y` sono vettori numerici e rappresentano i dati raccolti con l'analisi campionaria. Se è introdotto un unico vettore il test riguarderà la verifica della mediana della popolazione da cui il campione è estratto, se invece sono introdotti due vettori, il test potrà riguardare la verifica dell'ipotesi dell'uguaglianza delle mediane delle due popolazioni da cui sono estratti i campioni. In questo caso possiamo anche considerare l'analisi per dati appaiati modificando il valore di `paired` e portandolo a `TRUE`

- nel caso in cui si voglia cambiare il valore dell'ipotesi nulla da testare basterà cambiare il valore di `mu` al valore desiderato
- nel caso si voglia cambiare il livello di significatività del test sarà necessario cambiare il valore di `conf.level` portandolo al valore desiderato
- l'opzione `alt` può essere cambiata anche in `greater` e `less` a seconda che interessi la coda superiore o quella inferiore
- nel caso di due campioni il test ipotizza sempre che la varianza da cui provengono i campioni sia diversa, e quindi per il corretto uso del test sarà necessario modificare l'opzione in `var.equal=T`. Eè buona norma in questo caso far precedere tale test dal test `var.test` che testa se le varianze delle popolazioni da cui provengono i campioni sono uguali oppure diverse.

29.3 Verifica dell'ipotesi di uguaglianza della mediana: caso dei confronti multipli

Avendo a disposizione `k` campioni provenienti da `k` popolazioni diverse, a volte è necessario effettuare un test per verificare l'uguaglianza delle mediane mediante confronti multipli effettuati tra i `k` campioni prendendone sempre due a due in modo da considerare tutte i possibili accoppiamenti. Si tratta allora di effettuare un `test di wilcoxon` ripetuto più volte. Il comando per eseguire tali confronti multipli è il seguente:
`>pairwise.wilcox.test(x, g, p.adjust.method=p.adjust.methods, pool.sd=TRUE,...)`
in cui abbiamo che:

- `x` è il vettore dei dati
- `g` rappresenta il vettore dei fattori ossia un oggetto factor che indicherà con numeri o lettere l'appartenenza dei dati di x ad un determinato campione
- `p.adjust` indica il metodo da utilizzare che può essere dato da
  - holm"
  - hochberg"
  - bonferroni"
  - none"

29.4 Verifica dell'ipotesi di uguaglianza della mediana: caso del confronto globale

Avendo a disposizione `k` campioni provenienti da `k` popolazioni diverse, a volte è necessario effettuare un test per verificare l'uguaglianza delle mediane mediante un confronto globale effettuato simultaneamente sui `k` campioni. Si tratta di effettuare un test noto come analisi della varianza non parametrica che verrà studiato nel paragrafo 36. Senza entrare in dettagli vediamo che l'analisi della varianza viene eseguita con il seguente comando:
`>kruskal.test(x, g, ...)`
oppure `>kruskal.test(formula, data, subset,na.action,...)`
in cui si avrà che:

- `x` il vettore dei dati
- `g` rappresenta il vettore dei fattori ossia un oggetto factor che indicherà con numeri o lettere l'appartenenza dei dati di x ad un campione
- il valore di `formula` rappresenta la formula intesa nel seguente modo `y~g`

29.5 Il test sulla uguaglianza delle varianze di più popolazioni

Avendo a disposizione `k` campioni provenienti da `k` popolazioni diverse ci si pone il problema di verificare se le varianze delle popolazioni da cui provengono tali campioni siano tra di loro uguali. Ciò può esserre effettuato con il `test di fligner`. Tale test viene eseguito con il seguente comando:
`>fligner.test(x, g, ...)`
oppure il seguente comando:
`>fligner.test(formula, data, subset,na.action,...)`
in cui:

- `x` è il vettore dei dati
- `g` rappresenta il vettore dei fattori ossia un oggetto factor che indicherà con numeri o lettere l'appartenenza dei dati di x ad un determinato campione
- `formula` rappresenta una formula intesa nel seguente modo `x~g`

29.6 Il test sulla correlazione

Avendo a disposizione 2 campioni provenienti da 2 popolazioni diverse, ci si pone il problema di verificare se le popolazioni da cui provengono tali campioni sono tra di loro correlate. Ciò può esserre effettuato con il `test di correlazione`. Tale test viene effettuato con il seguente comando:
`>cor.test(x,y,alt=`"two.sided",method=" ",conf.level = 0.95)
Notiamo allora che:

- `x` ed `y` sono vettori numerici e rappresentano i dati raccolti con l'analisi campionaria nei due campioni
- nel caso si voglia cambiare il livello di significatività del test sarà necessario cambiare il valore di `conf.level` portandolo al valore desiderato
- l'opzione `alt` può essere cambiata anche in `greater` e `less` a seconda che
- il valore di `method` utilizzato per l'analisi non parametrica della correlazione è dato da:
  - kendall
  - sperman

30 I test per le proporzioni

30.1 Introduzione

I test per proporzioni sono quei test che riguardano inferenze su una certa modalità di un carattere quantitativo presente in una certa popolazione. Essi si articolano in una serie di test che studieremo nei paragrafi successivi.

30.2 Il test binomiale

Si supponga che una certa modalità di un carattere quantitativo sia presente in una certa popolazione in una percentuale che indicheremo con `p`. Sulla base dei risultati derivanti da un campione bernulliano estratto da tale popolazione vogliamo sottoporre a test l'ipotesi che la vera proporzione con cui il carattere è ripartito nella popolazione sia pari ad un certo valore. Tale inferenza viene effettuata con il `test binomiale esatto`. Tale test si esegue con il comando:
`>binom.test(x, n, p = 0.5,alt = c(`"two.sided", "less", "greater"),conf.level=0.95)
in cui si avrà che:

- `x` indica la numerosità dei successi che si sono presentati nel campione
- `n` indica la numerosità campionaria
- `p` indica la vera percentuale che supponiamo essere presente nella popolazione

Possiamo anche notare che:

- nel caso in cui si desidera modificare il livello di significatività del test sarà necessario cambiare il valore di `conf.level` sostituendolo con il valore desiderato
- il valore di `alt` può essere cambiato in `greater` e `less` a seconda che interessi la coda superiore o quella inferiore
- è possibile inserire in `x` un vettore con due componenti i cui elementi sono dati dal numero di successi e da quello degli insuccessi, in questo caso non si deve inserire alcun valore per `n`

Il test binomiale è importante in quanto fornisce anche gli intervalli di confidenza della vera proporzione con cui il carattere quantitativo è presente nella popolazione. Per ottenere tale intervallo sarà sufficiente introdurre come valore di `p` la percentuale di successi presente nel campione. Ad esempio se vogliamo ottenere l'intervallo di confidenza della vera proporzione dei votanti un determinato condidato in una certa popolazione avendo rilevato su un campione di 100 persone che 55 hanno votato per esso basterà utilizzare il seguente comando:
`>binom.test(55, 100, p = 0.55)`

30.3 Il test per due o più le proporzioni

Si supponga che una certa modalità di un carattere quantitativo sia presente in `k` popolazioni in una certa percentuale che indicheremo con `p1,p2,...pk`. Sulla base dei risultati derivanti da `k` campioni bernulliani estratti dalle `k` popolazioni vogliamo sottoporre a test l'ipotesi che la proporzione con cui il carattere è presente nelle `k` popolazioni sia uguale ai valori `p1,p2,...pk`. Tale inferenza viene effettuata con il `test prop.test`. Tale test si esegue con il comando:
`>prop.test(x, n, p = NULL, alt = c`(“two.sided“, “less“, “greater“),conf.level = 0.95,
`correct=TRUE)`
in cui si avrà che:

- `x` indica un vettore contenente i successi ottenuti in ognuno dei `k` campioni o una matrice con due colonne contenente nella prima i successi ottenuti in ogni campione e nella seconda colonna gli insuccessi ottenuti in ogni campione
- `n` indica un vettore contenente la numerosità dei `k` campioni e dovrà avere lo stesso numero di righe di x. Tale parametro è ignorato se `x` è una matrice
- `p` indica un vettore che contiene le vere percentali che supponiamo essere presenti nelle popolazioni da cui i campioni sono estratti . Tale vettore deve avere la stessa lunghezza di `x` e se `p` non è inserito si considera che la proporzioni di successi siano uguali in ogni popolazione.
- `correct` indica se usare o meno la correzione di continuità

Si noti inoltre che il `prop test` nel caso di due proporzioni fornisce anche gli intervalli fiduciari per le differenze delle vere proporzioni fra le due popolazioni.

Il prop.test può anche essere usato valori scalari e non con vettori ossia in questo caso

- x scalare di successi
- n scalare di tentativi
- p scalare indicante una probabilità se p non è inserito viene considerata pari a 0.5
- correct indica se usare o meno la correzione di continuità

Si noti che in questo caso si ottiene anche un intervallo di confidenza per la vera proporzione dei successi nella popolazione.

Si noti che il `prop test` è l'equivalente del test di chi quadrato per l'indipendenza di una tabella di contingenza solamente che in questo caso si ragiona solamente su due risposte si o no e quindi si desume che le popolazioni di origine siano delle binomiali.

30.4 Funzione potenza nel caso di due proporzioni

Il test per calcolare la funzione potenza nel caso di due proporzioni è il seguente:

```
power.prop.test(n=NULL, p1=NULL, p2=NULL, sig.level=0.05,
power=NULL,alternative=c("two.sided", "one.sided"))
```

in cui abbiamo che:

- n numero delle osservazioni per gruppo
- p1 probabilità in un gruppo
- p2 probabilità nell'altro gruppo
- sig.level livello di significatività o errore di prima specie
- power potenze del test o 1 meno probabilità dell'errore di seconda specie
- alternative alternativa a due code o ad una sola cosa

Si noti che dei valori predefiniti con `NULL` tre devono essere dati ed in funzione di essi il comando calcola il quarto mancante.

31 I test per tabelle di contingenza

31.1 I pacchetti necessari

Per utilizzare i test previsti in questo capitolo sarà necessario aver installato il pacchetto:
`ctest`

31.2 Le tabelle di contingenza

La tabella di contingenza è una tabella a doppia entrata che raccoglie le osservazioni effettuate su un campione analizzato attraverso l'impiego di due variabili statistiche congiuntamente considerate. Il fatto più importante che interessa in una tabella di contingenza è quello di stabilire, facendo delle inferenze sul campione estratto, se nella popolazione d'origine le due variabili sono tra di loro indipendenti. Vi sono sono molti test che a tal fine possono essere utilizzati e che verranno specificati nei paragrafi successivi.

31.3 Il test di Chi quadrato

Il test di Chi quadrato per l'indipendenza viene effettuato con uno dei seguenti comandi:
`>chisq.test(z)`
`>chisq.test(x,y)`
in cui si avrà che:

- `z` è una tabella di contingenza
- `x` è la variabile risposta mentre `y` è una variabile di tipo factor

è possibile usare l'opzione `simulate.p.value=T` mediante la quale si passa dalla distribuzione esatta a quella asintotica

31.4 Test di indipendenza completo per tutti i fattori

Con il comando `xtabs` è possibile creare delle tabelle di contingenza utilizzando la modalità formula. Quindi per creare delle tabelle di cintingenza possiamo dare anche il seguente comando:
`xtabs(y x1+x2,data=....)`
In cui x1 ed x2 sono delel variabili di tipo fattoriale. Con il comando:
`summary(xtabs(y x1+x2,data=....))`
si ottiene il test di indipendenza di tutti i fattori considerati. Si noti infatti che il test di chi quadrato vale solamente nel caso in cui si abbiano due fattori. E' anche possibile usare l'opzione `subset`.

31.5 Il test di Fisher

Il test di Fisher per l'indipendenza viene utilizzato soprattutto per analizzare dati discontinui, sia nominali che ordinali, quando i due campioni indipendenti sono molto piccoli. Esso viene effettuato con uno dei seguenti comandi:
`>fisher.test(z)`
`>fisher.test(x,y)`
in cui si avrà che:

- `z` è una tabella di contingenza
- `x` e la variabile risposta mentre `y` è una variabile di tipo factor

31.6 Il test di Mantelhaen

Il test di Mantelhaen per l'indipendenza viene effettuato con il seguente comando:
`>mantelhaen.test(z,correct=T)`
in cui si avrà che `z` è un array di n matrici di dimensione 2 x 2.

31.7 Il test di McNemar

Il test di McNemar per l'indipendenza viene utilizzato prevalentemente per verificare l'esistenza di differenze prima e dopo un certo trattamento qualora siano disponibili dati sotto forma di frequenze. Esso viene effettuato con il seguente comando:
`>mcnemar.test(z,correct=T)`
in cui si avrà che `z` è una tavola di contingenza di dimensione 2 x 2

32 Adattamento dei dati ad una distribuzione

32.1 I pacchetti necessari

Per utilizzare i grafici previsti in questa sezione sarà necessario aver installato il pacchetto:
`stepfun`

32.2 Una prima analisi dei dati

Quando si hanno a disposizione dei dati provenienti da una indagine statistica e si tenta di adattare questi dati ad una distribuzione teorica, la prima analisi da effettuare e quella di tentare di verificare se i dati provengono da una distribuzione discreta o da una distribuzione continua. Fatto ciò dobbiamo tentare di ipotizzare tramite una analisi grafica il tipo di popolazione da cui i dati provengono. Mentre per quanto riguarda il primo punto, in generale non si presentano problemi rilevanti, l'identificazione di una distribuzione teorica tramite l'analisi grafica si presenta molte volte difficile e ardua. Tale tipo di analisi può essere effettuata utilizzando i seguenti comandi:

- `plot(table(x))`
- `hist(x)`
- `qqnorm(x)  e qqline(x)`
- `qqplot(x)`

32.3 La funzione cumulata di distribuzione

Un altro utile strumento grafico di analisi è quello legato alla funzione cumulata di distribuzione. Tale funzione ci aiuta a verificare:

- in un campione l'adattamento ad una distribuzione specifica
- in due campioni l'uguaglianza della distribuzione di provenienza degli stessi

Nel caso avessimo a disposizione un solo campione i cui dati sono memorizzati nella variabile `x`, per verificare se tale campione proviene da una distribuzione specificata si può utilizzare la seguente sintassi:
```
>plot(ecdf(x),verticals=T,do.p=F)
>y<-sort(x)
>lines(y,pfunc(y,....))
```
in cui si avrà che:

- il parametro `pfunc` sarà sostituito con i nomi di una delle distribuzioni base viste in precedenza ossia: pnorm, pbeta, pcauchy, pchisq, pexp, pf, pgamma, plnorm, plogis, pt, punif, pweibull, pbinom, pgeom, phyper, pnbinom, ppois, pwilcox
- al posto dei puntini andranno inseriti i parametri richiesti dalla funzione `pfunc` di solito stimati con i dati derivanti dal campione `x` stesso.

Ad esempio se supponiamo che `x` possa provenire da una distribuzione normale potremmo operare nel seguente modo:
```
>plot(ecdf(x),verticals=T,do.p=F)
>y<-sort(x)
>lines(y,pnorm(y,mean(y),sd(y)))
```
Si noti che avremmo potuto anche operare nel seguente modo:
```
>plot(ecdf(x),verticals=T,do.p=F)
>curve(pnorm(x,mean(x),sd(x)),min(x),max(x),add=T)
```

Alcuni suggeriscono di non considerare per tracciare il grafico delal continua i valori del vettore di dati x ma di costruirne uno che varia dal minimo di x al massimo di x con il comando seq. Si noti che operando nel secondo modo visto sopra noi praticamente adottiamo questa scelta.

Nel caso avessimo a disposizione due campioni, per verificare se essi provengono dalla stessa distribuzione si può utilizzare la seguente sintassi:
```
>plot(ecdf(x),verticals=T,do.p=F)
>plot(ecdf(y),verticals=T,do.p=F,add=T)
```
in cui `x ed y` sono vettori che possono anche assumere diversa lunghezza.

32.4 Le funzioni qqnorm, qqline e qqplot

Un metodo alternativo per verificare se un vettore proviene da una distribuzione di probabilità o se due vettori provengono dalla stessa distribuzione di probabilità è quello che fa riferimento all'uso delle funzioni `qq`. Si possono allora distinguere vari casi.

Se abbiamo a disposizione un vettore di osservazioni `x` è vogliamo verificare se esso proviene da una distribuzione normale possiamo utilizzare i comandi:
```
>qqnorm(x)
>qqline(x)
```
Per verificare invece se i dati provengono da altre distribuzioni possiamo usare la seguente sintassi:
```
>plot(qfunc(ppoints(x),...),sort(x))
>aaa<-quantile(x,c(0.25,0.75))
>bbb<-qfunc(c(0.25,0.75),...)
>coeff<-(aaa[2]-aaa[1])/(bbb[2]-bbb[1])
>inter<-aaa[1]-coeff*bbb[1]
>abline(inter,coeff)
```
in cui si avrà che:

- il parametro `qfunc` sarà sostituito con i nomi di una delle distribuzioni base viste in precedenza ossia: pnorm, pbeta, pcauchy, pchisq, pexp, pf, pgamma, plnorm, plogis, pt, punif, pweibull, pbinom, pgeom, phyper, pnbinom, ppois, pwilcox
- al posto dei puntini andranno inseriti i parametri richiesti dalla funzione `pfunc` di solito stimati con i dati derivanti dal campione `x` stesso.

Per verificare invece se i dati di due vettori `x` ed `y` provengono dalla stessa distribuzione possiamo usare la seguente sintassi:
```
>qqplot(x,y)
>aaa<-quantile(x,c(0.25,0.75))
>bbb<-quantile(y,c(0.25,0.75))
>coeff<-(aaa[2]-aaa[1])/(bbb[2]-bbb[1])
>inter<-aaa[1]-coeff*bbb[1]
```

`>abline(inter,coeff)`
e verificare se i dati risultano essere allineati lungo la retta così ottenuta.

32.5 Il test chi quadrato la bontà dell'adattamento nel caso di distribuzioni discrete

Il test chi quadrato per valutare la bontà dell'adattamento viene effettuato in modo diverso a seconda della distribuzione oggetto di adattamento. In ogni caso si deve ricordare che per usare i test che vedremmo successivamente il confronto è fatto sempre tra due vettori, il primo che raccoglie le frequenze osservate e che indicheremo con `fi` mentre il secondo raccoglie le probabilità stimate e che abbiamo inserito nella variabile `p`. Si avrà allora che:

- se la distribuzione a cui si vogliono adattare i dati è una `distribuzione binomiale` possiamo utilizzare la seguente sintassi:
  `>fi<-c(15,45,76,51,13)`
  `>pro<-dbinom(0:4,4,0.5)`
  `>chisq.test(fi,p=pro)`
  in cui il valore di 0.5 è stato scelto come esempio ma è proprio il parametro più importante da prendere in considerazione. Il valore di `p` di solito viene stimato dai dati campionari a disposizione.
- se la distribuzione a cui si vogliono adattare i dati è una `distribuzione uniforme` possiamo utilizzare la seguente sintassi:
  `>fi<-c(15,45,76,51,13)`
  `>chisq.test(fi)`
- se la distribuzione a cui si vogliono adattare i dati è una `distribuzione multinomiale` possiamo utilizzare la seguente sintassi:
  `>fi<-c(15,45,76,51,13)`
  `>pro<-c(a,b,c,d,e)`
  `>chisq.test(fi,p=pro)`
  in cui i valore di `a,b,c,d,e` sono le ipotizzate probabilità per ogni frequenza ottenuta di solito stimate attraverso i dati campionari otenuti. Si noti che se si vuole provare che le probabilità effettive sono uguali basterà usare il comando: `>chisq.test(fi)`
  in quanto sarà automaticamente ipotizzato che la probabilità delle frequenze teoriche siano per ogni frequenza assoluta pari ad `1/5`.

32.6 Il test chi quadrato la bontà dell'adattamento nel caso di distribuzioni continue

Se la distribuzione a cui si vogliono adattare i dati è una distribuzione continua, ad esempio una normale la procedura da seguire è più articolata. Dobbiamo prima di tutto ragruppare le frequenze assolute ottenute dall'analisi campionaria in classi ed ottenere quindi una tabella tipo la seguente:

<55	55-60	60-65	65-70	70-75	75-80	80-85	>85
8	68	272	680	617	283	57	15

Se vogliamo verificare se il campione ottenuto proviene da una `N(70,25)`. La sintassi da usare è la seguente:

```
>fi<-c(8,68,272,680,617,283,57,15)
>y1<-seq(55,85,5)
y2<-seq(50,80,5)
>y1<-(y1-70)/5
y2<-(y2-70)/5
>xxx<-pnorm(y1)-pnorm(y2)
>xxx[8]<-1-pnorm(3)
>chisq.test(fi,p=xxx)
```

32.7 Il test di Kolmogorov-Smirnov per la bontà dell'adattamento

Il test di Kolmogorov-Smirnov viene effettuato con il seguente comando:
`>ks.test(x,y=null,`"pfunc",...,alt="two.sided")
in cui si ha che:

- il parametro `pfunc` sarà sostituito con i nomi di una delle distribuzioni base viste in precedenza ossia: pnorm, pbeta, pcauchy, pchisq, pexp, pf, pgamma, plnorm, plogis, pt, punif, pweibull, pbinom, pgeom, phyper, pnbinom, ppois, pwilcox
- al posto dei puntini andranno inseriti i parametri richiesti dalla funzione `pfunc` di solito stimati con i dati derivanti dal campione `x` stesso.
- alt può essere greater e less

32.8 Il test di Shapiro

Per verificare la normalità dei dati su cui vogliamo fare un'analisi è anche possibile usare il test di Shapiro che ha la seguente sintassi:
`>shapiro.test(x)`
in cui x è il vettore dei dati da analizzare.

33 Ricerca degli zeri ed ottimizzazione di una funzione

33.1 Introduzione

In R sono presenti alcune funzioni di particolare importanza che sono adatte per la ricerca degli zeri e per l'ottimizzazione di funzioni ad una o più variabili. Tali funzioni saranno descritte dettagliatamente nei paragrafi che seguiranno. In tutti i casi in cui è richiesto l'uso di una funzione questa potrà essere assegnata in due modi distinti:

- definendola direttamente come visto nel paragrafo relativo alla programmazione ossia nel seguente modo:
  `>ff<-function(x){ }`
  e quindi nel comando utilizzeremo per richiamare la funzione il suo nome ossia ff
- definendola direttamente nel comando con la seguente sintassi:
  `>comando(function(x) { }.....)`

Si noti che i due modi di procedere sono alternativi ma equivalenti.

33.2 La funzione uniroot

La funzione uniroot serve per trovare gli zeri di una funzione ad una variabile. Il suo uso è il seguente:
`>uniroot(f, interval)`
Si noti il seguente esempio:
```
>f <- function (x,a) x - a
>uniroot(f, c(0, 1), tol = 0.0001, a = 1/3)
```
Con tale comando abbiamo ottenuto gli zeri di f nell'intervallo [0,1] e si noti che è possibile stabilire anche la tolleranza per il calcolo delle radici stesse.

33.3 La funzione optimize

La funzione optimize serve per trovare il massimo o il minimo di una funzione in un intervallo specificato. Il suo uso è il seguente:
`>optimize(f, interval , maximum = FALSE)`
Si noti il seguente esempio:
```
>f <- function (x,a) (x-a)^2
>optimize(f, c(0, 1), tol = 0.0001, a = 1/3)
```
Si noti che in tale modo abbiamo ottenuto il minimo della funzione f nell'intervallo [0,1], per ottenere il massimo avremmo dovuto specificare l'opzione maxim=T.

33.4 La funzione optim

La funzione optim è usata prevalentemente per trovare il minimo di un sistema di equazioni non lineari. Tale ricerca è effettuata utilizzando vari metodi di ottimizzazione ossia Nelder-Mead, quasi-Newton e conjugate-gradient algorithms. Il suo uso è il seguente:

```
>optim(par, fn, gr = NULL, method = c("Nelder-Mead", "BFGS", "CG", "L-BFGS-B",
 SANN"), lower = -Inf, upper = Inf, control = list(), hessian = FALSE,...)
```
Si noti il seguente esempio:

```
fr <- function(x) {   ## Rosenbrock Banana function
                    x1 <- x[1]
                    x2 <- x[2]
                    100 * (x2 - x1 * x1)^2 + (1 - x1)^2
                    }
grr <- function(x) { ## Gradient of 'fr'
                    x1 <- x[1]
                    x2 <- x[2]
                    c(-400 * x1 * (x2 - x1 * x1) - 2 * (1 - x1),
                    200 *      (x2 - x1 * x1))
                    }
optim(c(-1.2,1), fr)
optim(c(-1.2,1), fr, grr, method = "BFGS")
optim(c(-1.2,1), fr, NULL, method = "BFGS", hessian = TRUE)
optim(c(-1.2,1), fr, grr, method = "CG")
optim(c(-1.2,1), fr, grr, method = "CG", control=list(type=2))
optim(c(-1.2,1), fr, grr, method = "L-BFGS-B")
```

Tale funzione può anche essere usata per ottimizzare una funzione ad una variabile.

33.5 La funzione simplex

La funzione simplex serve per ottimizzare una funzione lineare soggetta a vincoli lineari utilizzando il metodo del simplesso. Tale funzione richiede l'uso del pacchetto boot. Si noti che i vincoli sono espressi nel seguente modo:

- a rappresenta il vettore dei coefficienti della funzione obiettivo
- $A1 * x <= b1$
- $A2 * x >= b2$
- $A3 * x = b3$

in cui ovviamente vale il vincolo di segno che $x >= 0$. La sintassi dell'uso di questa funzione è quindi la seguente:
```
>simplex(a, A1=NULL, b1=NULL, A2=NULL, b2=NULL,A3=NULL,
b3=NULL, maxi=FALSE, n.iter=n + 2 *m,eps=1e-10)
```
Un esempio di uso di tale funzione è il seguente:
```
>enj <- c(200, 6000, 3000, -200)
>fat <- c(800, 6000, 1000, 400)
>vitx <- c(50, 3, 150, 100)
>vity <- c(10, 10, 75, 100)
>vitz <-c(150, 35, 75, 5)
>simplex(a=enj, A1=fat, b1=13800, A2=rbind(vitx,vity,vitz),
b2=c(600, 300, 550), maxi=TRUE).
```

34 Formule di quadratura di una funzione

35 La regressione

35.1 Introduzione

La regressione lineare consiste in una serie di tecniche volte principalmente alla ricerca di un modello con cui una variabile dipende linearmente da un'altra serie di variabili. I dati per lo studio della regressione lineare sono composti principalmente da un data frame che nel corso del paragrafo chiameremo sempre

`anar` composto da più variabili, la prima che chiameremo `y` di tipo numerico, rappresenta la variabile dipendente mentre le altre che indicheremo con `x1,x2,...`, dette variabili indipendenti o espicative, saranno di tipo numerico. Il caso in cui le variabili esplicative siano di tipo factor verrà affrontato in un apposto paragrafo. Per indicare che `y` dipende da `x1,x2,...` scriveremo:
`y~x1+x2+x3+....`

35.2 Analisi grafica preliminare

Un primo approccio relativo all'analisi della regressione che serve per valutare quali variabili sono tra loro legate da un legame di tipo lineare è l'approccio grafico. Tale analisi viene compiuta con il comando:
`>pairs(anar)`
In tale modo siamo in grado di poter visualizzare i possibili legami di tipo linearetra le variabili.

Nel caso in cui fossimo in presenza di due varibili x ed y con y dipendente da x tale tipo di ricerca potrà essere più facilemente effettuata disegnando uno scatterplot di y contro x con il comando:
`>plot(x,y)`
In questo caso potremmo anche utilizzare la funzione `lowess` la quale restituisce una serie di coordinate di x ed y che possono essere utilizzate per lisciare lo scatterplot ottenuto. Il suo uso è il seguente:

```
>plot(x,y)
>smooth<-lowess(x,y)
>lines(smooth)
```

se vogliamo accentuare il lisciamento dovremmo provvedere a introdurre nella funzione `lowess` un terzo paramentro ossia un numero compreso tra zero ed 1.

35.3 Analisi della correlazione lineare

Prima di procedere alla costruzione del modello può anche essere utile effettuare una analisi sul grado di correlazione lineare che esiste tra le varibili e tra la variabile dipendente e quelle esplicative. Ciò può essere effettuato con il comando:
`>cor(anar)`
Essendo molto utile per la regressione di tipo step-wise vogliamo introdurre anche il concetto di coefficiente di correlazione parziale. Tale coefficiente misura la dipendenza lineare tra due variabili al netto di una terza. Se si desidera quindi calcolare la correlazione parziale tra la varibile `x1` e la variabile `x2` al netto della influenza della variabile `x3` è sufficiente calcolare la correlazione lineare tra i residui della regressione di `x1` su `x3` e quelli della regressione di `x2` su `x3`. Quindi se ipotizziamo che il nostro modello sia del tipo:
`y~x1+x2+x3+x4`
calcolando i residui del modello:
`y~x4`
e quelli dei modelli:
`x1~x4`
`x2~x4`
`x3~x4`
possiamo calcolare tre coefficienti di correlazione parziale che eliminano l'influenza di `x4` ossia:

- la correlazione tra i residui del modello `y~x4` e quelli del modello `x1~x4`
- la correlazione tra i residui del modello `y~x4` e quelli del modello `x2~x4`
- la correlazione tra i residui del modello `y~x4` e quelli del modello `x3~x4`

Naturalmente il procedimento può essere generalizzato per eliminare l'influenza di due o più variabili. Se infatti ipotizziamo che il nostro modello sia del tipo:
`y~x1+x4`
calcolando i residui del modello:
`y~x1+x4`
e quelli dei modelli:
`x2~x1+x4`
`x3~x1+x4`
possiamo calcolare due coefficienti di correlazione parziale che eliminano l'influenza di `x1 e x4` ossia:

- la correlazione tra i residui del modello `y~x1+x4` e quelli del modello `x2~x1+x4`
- la correlazione tra i residui del modello `y~x1+x4` e quelli del modello `x3~x1+x4`

Si noti che un modo alternativo al precedente per ottenere il coefficiente di correlazione parziale è il seguente:

- si calcola il valore di R^2 del modello `y~x4` detto modello A
- si calcola il valore di R^2 del modello `y~x4+x1` detto modello B

il coefficiente di correlazione parziale di `y` con `x1` al netto dell'influenza di `x4` è dato da:

$$\sqrt{\frac{R_B^2 - R_A^2}{1 - R_A^2}}$$

il modello B dovrà contenere sempre una variabile im più rispetto al modello A.

35.4 La regressione lineare ai minimi quadrati

La regressione lineare ai minimi quadrati viene effettuata con il comando:
`>anar.lm<-lm(y~x1+x2+...+xn,data=anar)`
se siamo interessati ad ottenere la regressione senza il termine noto il comando diviene:
`>anar.lm<-lm(y~-1+x1+x2+...+xn,data=anar)`
Se anar è un data frame contenente come prima variabile la varibile dipendente e come successive variabili quelle indipendenti, per effettuare la regressione può essere usata la seguente sintassi semplificata:
`>anar.lm<-lm(anar)`

Se l'analisi non deve essere compiuta su tutti i dati a disposizione ma solamente su un sottoinsieme di dati che può essere ottenuto anche da un'altra variabile presente nel data frame `anav`, possiamo effettuare l'analisi con il comando:
`>anar.lm<-lm(y~x1+x2+...xn,subset=(variabile=='valore')`
in cui

- `variabile` indica in nome della variabile attraverso la quale estraiamo il sottoinsieme di dati che ci interessano
- l'operatore relazionale `==` può essere sostituito a seconda dei casi da un qualunque altro operatore relazione
- `valore` indica il criterio di estrazione dei dati

L'opzione subset può anche essere usata in questo modo, supponiamo che da una analisi grafica o altro abbiamo deciso che alcuni valori del data frame non debbano essere usati nel calcolo. Stabiliti tali valori ad esempio quelli della riga1,riga5 e riga10 del data frame `anar` dobbiamo costruire un vettore:
`outliers<-c(1,5,10)`
e dare il comando:
`>ana<-lm(y~x1,subset=-outliers`
in tale modo vengono utilizzate solamente le righe del data frame `anar` escluse duelle indicate in `outliers`.

Una volta effettuata tale regressione per visualizzare vedere le informazioni ottenute possiamo usare i seguenti comandi:
`>anar.lm` per vedere i dati salienti della regressione
`>summary(anar.lm)` per vedere informazioni più dettagliate sulla regressione
`>summary(anar.lm,correlation=T)` otteniamo anche la matrice di correlazione tra le varibili
`>coefficients(anar.lm)` per vedere aolamente i coefficienti stimati tramite la regressionei
`>residuals(anar.lm)` per vedere i residui della regressione
`>fitted.values(xxx.lm)` per vedere i valori stimati della regressione

Da tali comandi individuiamo una serie di informazioni utili per una corretta stima del modello ai minimi quadrati, si noti che tra l'altro che il `residual standard error` fornito da tali comandi è la stima di σ e non di σ^2 come tra l'altro era facile intuire dal nome utilizzato.

Si noti che con il comando `lm` possiamo anche utilizzare l'opzione `x=T` che permette di vedere utilizzando `anar.lm$x` di vedere come sono effettivamente codificate le variabili predittive da R, cosa molto importante da conoscere nel caso in cui le variabili predittive siano di tipo factor.

35.5 La multicollinearità

Come già sappiamo dalla teoria, la multicollinearità presenta uno dei problemi più gravi legato all'analisi di regressione. La multicollinearità può essere individuata con uno dei seguenti metodi:

- analisi dei coefficienti di correlazione tra coppie di variabili esplicative
- determinante della matrice `(X'X)` prossimo allo zero in cui con `X` si indica la matrice delle variabili esplicative
- autovalori della matrice `(X'X)` prossimi allo zero in cui con `X` si indica la matrice delle variabili esplicative

Per porre rimedio alla presenza di multicollinearità nei dati, si può fare ricorso alla tecnica degli stimatori `ridge`. Per utilizzare tali stimatori dobbiamo preventivamente caricare in R il pacchetto `MASS` e la stima sarà effettuata tramite il comando:
`>anar.lm<-lm.ridge(y~x1+x2+...+xn,data=anar)`
e valgono tutte le cose che abbiamo detto finora. Per utlteriori precisazioni si faccia ricorso all'help in linea della funzione `lm.ridge`.

35.6 Analisi della varianza di regressione

Una volta creato il modello come visto nei paragrafi precedenti possiamo effettuare un'analisi della varianza del modello stesso attraverso l'uso dei seguente comando: `>anova(anar.lm)`
si ottiene una tavola della analisi della varianza relativa alla regressione in oggetto da cui è facile calcolare il valore di R^2 dato da:

$$R^2 = 1 - \frac{devianzaresidua}{devianzatotale}$$

mentre il test F sarà dato da:

$$F = \frac{devianzaspiegata/(k-1)}{devianzatotale/(n-k)}$$

in cui k sono i parametri stimati della regressione mentre n rappresenta la numerosità del campione.

35.7 Regressione su un insieme limitato di dati

Per effettuare la regressione su un insieme limitato di dati dovremmo utilizzare il comando:
`>xxx.ln<-lm(y~x1+x2+...+xn,subset=.....)`
in cui i puntini vanno sostituiti con un vettore soddisfante una condizione particolare. Se ad esempio disponiamo di un data frame con variabili vento, temperatura e mesi e vogliamo effettuare una regressione della temperatura sul vento ma considernado solamente i dati del mese 5 dobbiamo utilizzare la sintassi:
`>xxx.ln<-lm(temperatura~vento,subset=mese==5)`

35.8 L'analisi dei residui

Supponendo di aver costruito un modello di regressione lineare e averlo memorizzato nella variabile `anar.lm` lo studio grafico dei residui di tale modello ci permette di acquisire una serie di informazioni circa il modello creato. Sappiamo infatti che se valgono tutte le assunzioni base del metodo dei minimi quadrati allora:

- $E(Y) = X\beta$
- $var(Y) = \sigma^2 I_n$
- il vettore degli errori di regressione $U = Y - X\beta$ ha media uguale a zero e varianza uguale a $\sigma^2 I_n$, e cioè gli errori di regressione sono omoschedastici ed incorrelati tra d iloro.

Ora il vettore dei residui delle stime ai minimi quadrati ordinari viene definito come:

$$\hat{U} = Y - X\hat{\beta}$$

in cui $\hat{\beta}$ è la stima ai minimi quadrati. Sappiamo inoltre che la proprietà fondamentale di $\hat{U}$ è di essere ortogonale alla matrice delle varibili esplicative ossia $X'\hat{U} = 0$. Senza entrare in ulteriori questioni d icarattere teorico posto

$$M = I_n - X(X'X)^{-1}X'$$

il comportamento di U e quello di $\hat{U}$ sarà somigliante se

$$M \approx I_n$$

è ciò si avrà quando $\max h_{ii} \to 0$ indicando con h_{ii} l'elemento della diagonale della matrice di proiezione

$$X(X'X)^{-1}X'$$

Si noti che h_{ii} è una misura dell'effetto di leva esercitato dal caso i-esimo sulla stima dei minimi quadrati. Per studiare quindi l'andamento di $\hat{U}$ come stima d iquello di U sarà necessario chè nessun caso eserciti un effetto di leva. Come primo approccia all'analisi grafica dei residui dovremmo allora evidenziare se esistono eventuali punti di leva. Memorizzata la matrice di proiezione $X(X'X)^{-1}X'$ nella variabile leva, i punti di leva possono essere evidenziati graficamente con il seguente comando:
`dotchart(diag(leva),label=1..n)`
un metodo alternativo è quello che fa riferimento al grafico della distanza di cook che vedremo successivamente.

Fatta questa analisi, possiamo utilizzare una serie di grafici legati ai residui per verificare se l ipotesi base dei minimi quadrati ordinari sono soddisfatte. Tali grafici sono i seguenti:

- il grafico dei residui contro le osservazioni ottenuto con il comando
  `>plot(1:n,residuals(anar.lm))`
  ci permette di avere una prima idea di come si dispongono i residui del modello
- il grafico dei residui contro i valori stimati ottenuto con il comando:
  `>plot(fitted(anar.lm),residuals(anar.lm))`
  ci permette di verificare se le ipotesi di omoscedasticità , media nulla e incorrelazione dei residui sono verificate.
- il grafico dei residui rispetto alle variabili esplicative ottenuto con il comando:
  `>plot(xi,residuals(anar.lm))`
  fatto per ogni variabile esplicativa x1, x2, xn, ci permette di individuare scorrette specificazioni della dipendenza dalle variabili esplicative, come ad esempio dipendenze non lineari
- il grafico di `y` contro i valori stimati dal modello ottenuto con:
  `>plot(y,fitted(anar.lm))`
  `>abline(0,1)`
  in presenza di corretta specificazione i punti di tale scatterplot dovrebbero essere allineati sulla bisettrice del primo e terzo quadrante.
- il grafico di $\hat{U}_t$ contro $\hat{U_{t-1}}$ ci permette di evidenziare la presenza di autocorrelazione negli errori di regressione

In generale, sotto le ipotesi del modello lineare dei minimi quadrati, tali diagrammi non presentano alcun andamento particolare o sistematico. Ogni scostamento da tali ipotesi indicherà la necessità di migliore il modello procedendo ad una sua riformulazione anche con curve di tipo non lineare o ad una trasformazione delle variabili esplicative ottenuibili attraverso la famiglia $x^{\frac{1}{m}}$.

Un altro aspetto molto importante riguarda la verifica della normalità dei residui. Per verificare tale fatto possiamo utilizzare anche congiuntamente due strumenti già visti precedentemente:

- l'analisi grafica ottenibile con l'utizzo delle funzioni `qqnorm` e `qqline`
- i test di adattamento e quelli di normalità

Alcuni grafici essenziali per l'analisi della bontà del modello possono essere ottenuti con il comando:

- `plot(anar.lm)` in cui vengono identificati i 3 più estremi valori
- `plot(anar.lm,id.n=n)` in cui vengono identificati gli n più estremi valori

35.9 Regressione con varibili di tipo factor

Come abbiamo precisato all'inzio, le variabili espicative possono essere variabili di tipo factor. Ad esempio se disponiamo di una variabile risposta y e di tre variabili esplicative x1 e x2 di tipo continuo e x3 di tipo factor con 3 livelli a,b,c possiamo stimare il modello con il comando:
`lm(y x1+x2+x3)`
In questo modo R stima 4 coefficienti relativi alle variabili x1, x2,x32 e x33 in cui la variabile x32 assume il valore 1 se l'osservazione in oggetto aveva come fattore 2 e il valore 0 se l'osservazione in oggetto non aveva come fattore 2. La variabile x33 si comporta esattamente nello stesso modo. Se utilizziamo il comando:
`lm(y -1+x1+x2+x3`
l'intercetta non è presente nella regressione e quindi possiamo ottenere un numero di coefficienti pari ad 5 assia pari alla numerosità del livelli del fattore x più 2. Notiamo in questo caso che il coefficiente del primo livello è uguale all'intercetta del primo modello stimato, quella del secondo livello sarà l'intercetta del primo modello sommata al coefficiente dello stesso livello nel primo modello e così via di seguito.

Può essere utile visualizzare le variabili che R ha utilizzato per effettuare il calcolo e ciò è possibile effettuando la regressione nel seguente modo:
`xxx<-lm(y x,x=T)`
`xxx<-lm(y -1+x,x=T)`
in modo tale che con il comando:
`xxx$x`
sarà possibile evidenziare le codifiche che R ha assegnato alle variabili di tipo factor.

In presenza di un modello in cui oltre alle varibili factor ci troviamo in presenza di variabili di tipo quantitativo il modello potrebbe prevedere di verificare eventuali interazioni tra i due tipi di variabili. Ad esempio se il modello considera la variabile `y` come varibile dipendente mentre la variabile `x` come variabile quantitativa e la variabile `z` come variabile factor, un modello che permette di studiare anche le interazioni tra x e z potrebbe essere il seguente: `y~x*z`

Il modo predefinito utilizzato da R per codificare le variabili di tipo factor prende il nome di contrasto di trattamenti. Si noti però che è possibile utilizzare varie tipologie di contrasto specificandole con il comando:
`lm(y x1+x2+x3,contrast=list(x3=tipo di contrasto))`
in cui i tipi di contrasto sono i seguenti:

- contr.treatment (predefinito)
- contr.helmert (usato da statgraphics)
- contr.poly
- contr.sum

E' anche possibile la creazione di contrasti personalizzati. Tali contrasti predefiniti seapplicati ad una variabile fattoriale a tre livelli 1,2,3 creano sempre due nuove variabile che conterranno i seguenti valori a seconda del tipo di contrasto utilizzato:

```
> contr.treatment(3)
  2 3
1 0 0
2 1 0
3 0 1
> contr.helmert(3)
  [,1] [,2]
1   -1   -1
2    1   -1
3    0    2
> contr.poly(3)
                .L         .Q
[1,] -7.071068e-01  0.4082483
[2,] -9.073264e-17 -0.8164966
[3,]  7.071068e-01  0.4082483
> contr.sum(3)
```

```
  [,1] [,2]
1    1    0
2    0    1
3   -1   -1
```

E' ovvia la generalizzazione per variabili con un numero più elevato di livelli.

35.10 Procedimento manuale di creazione del modello

Quando si hanno a disposizione più varibili esplicative il problema che ci si pone è quello di stabilire se tutte le variabili devono essere considerate nella regressione e in caso negativo decidere quali considerare e quali invece escludere. Un procedimento semplicistico e non esaustivo che può essere utilizzato come primo approccio può essere il seguente:

- si crea il primo modello con la variabile che presenta il coefficiente di correlazione con y più elevato
- creato il modello si deciderà di escludere dal modello quelle varibili che presentano un p-value basso
- si inseriranno nel modello quelle variabili che presentano un coefficiente di correlazione parziale con y al netto delle varibili già presenti nel modello più elevato. Ad esempio se il modello ipotizato è:
  `y~x4`
  calcoliamo i residui di tale modello e quelli dei modelli:
  `x1~x4`
  `x2~x4`
  `x3~x4`
  come visto i tre coefficienti di correlazione parziale che n risultano saranno i seguenti: l'influenza di `x4` ossia:
  - la correlazione tra i residui del modello `y~x4` e quelli del modello `x1~x4`
  - la correlazione tra i residui del modello `y~x4` e quelli del modello `x2~x4`
  - la correlazione tra i residui del modello `y~x4` e quelli del modello `x3~x4`

  potremmo quindi inserire nel modello la variabile che presenta il coefficiente di correlazione lineare parziale più elevato.

e ci si arresta quando l'inserimento di una nuova varibile non ha alcun effetto su R^2.

Ribadisco che tale procedimento è puramente indicativo e non pretende di esaurire in modo meccanicistico il problema della scelta delle variabili legate alla regressione.

Naturalmente la scelta delle variabili a considerare nella regressione potrebbe avvenire anche considerando l'analisi grafica tra le variabili esplicative con la varibile dipendente ottenibile con il comando:
`>pairs(anar)`
in cui anar è il data frame che contiene sia la variabile dipendente che quella indipendente. Dall'analisi grafica è possibile quindi evidenziare le relazioni di tipo lineare che legano ciascuna variabile indipendente a quella dipendente e stabilire così in prima approssimazione le variabili indipendenti da utilizzare.

35.11 Procedimento automatico di creazione del modello

Il programma R prevede anche la possibilità di utilizzare un procedimento automatico di selezione automatica delle varibili che di volta in volta saranno inserite o esclude dal modello. Per capire bene tale procedimento vediamo prima di tutto manualmente il suo utilizzo.

Se abbiamo già costruito un modello con il comando:
`>anar.lm<-lm(y~x1+x2+...+xn)`
è possibile decidere quali varibili eliminare dalla regressione tramite l'utilizzo del comando `drop1` il quale indica la variabile da escludere dal modello in base al metodo dell'AIC. In pratica tramite il comando:
`>drop1(xxx.lm)`
viene evidenzata per ogni variabile presente nella regressione il suo AIC. Dovrà allora uscire dalla regressione la variabile che presenta l'AIC più basso. Decisa quindi la varibile da far uscire dalla regressione che supponiamo essere x2 l'aggiornamento del modello avviene con il comando:
`>update(xxx.lm,.~.-x2)`

Se stiamo partendo dall'inizio e non abbiamo nessuna idea di quali varibili aggiungere al modello possiamo utilizzare il comando `add1` il quale indica la variabile da aggiungere alla regressione in base al

metodo dell'AIC. Il suo utilizzo è il seguente, prima di tutto viene costruito il modello comprendente solo l'intercetta con
`>xxx.ln<-lm(y~1)`
e tramite il comando
`>add1(xxx.lm,~x1+x2+x3+...+xn)`
viene evidenzata per ogni variabile il sui AIC. Dovrà allora entrare nella regressione la variabile che presenta l'AIC più basso. Decisa quindi la variabile da far entrare nella regressione che supponiamo essere x3 l'aggiornamento del modello avviene con il comando:
`>update(xxx.lm,.~.+x3)`
Si noti chè è possibile utilizzare il comando `add1` anche con un modello già precedentmente stimato in questo caso il comando dovrà essere utilizzato nel seguente modo:
`>add1(xxx.lm,~x1+x2+x3+...+xn)`
in cui però x1, x2 x3 ,....,xn sono sia le variabili della regressione già stimata sia quelle che vogliamo aggiungere.

Per ottenere il miglior modello sempre in base all'AIC ma in modo completamente automatico si può utilizzare il comando
`step`
il quale fornisce il miglior modello stimabile in base al metodo dell'AIC. Il suo uso è il seguente, prima di tutto si procede alla costruzione del modello avente solo l'intercetta con :
`>xxx.ln<-lm(y~1)`
successivamente con il comando: `>step(xxx.lm,~x1+x2+x3+...+xn)`
viene automaticamente creato il modello mostrando tutti i passaggi che sono stati effettuati. Se non si desidera vedere i passaggi ma ottenere il miglio modello stimabile si deve utilizzare la seguente sintassi:
`>step(xxx.lm,~x1+x2+x3+...+xn,trace=F)`
In questo modo il comando `step` utilizza una procedura stepwise di dipo backward ossia all'indietro consistente nella eliminazione delle variabili. E' possibile utilizzare il comando con la procedura forward che consiste nell'aggiunta delle variabili, ciò può essere effettuato utilizzando nel comando `step` l'opzione `direction=forward`. Si noti in ogni caso che l'uso di tali strumenti automatici deve sempre essere mediato dall'esperienza e dalle informazioni di chi effettua l'analisi.

I comandi:

- `add1`
- `drop1`
- `step`

possono essere utilizzati con l'opzione `k=numero`. Il valore predefinito di tale valore di k è uguale a 2. Tale volore di k rappresenta il valore di penalità necessario al R per stabilere quali variabili far entrare od uscire dalla regressione.

E' possibile anche usare il comando:
`stepAIC`
disponibile dopo aver caricato il pacchetto `MASS`. Si veda l'help per maggiori dettagli.

35.12 Regressione polinomiale o con altra funzione predefinita

Nel caso in cui dovessimo effettuare una regressione ai minimi quadrati utilizzando una funzione predefinita ad esempio la funzione

$$f(x) = a + bx + cx^2$$

oppure con la funzione

$$f(x) = \frac{a}{x} + bx^2$$

il comando da utilizzare nel primo caso sarà
`lm(y~I(x)+I(x^2))'`
mentre per il secondo caso sarà
`lm(y~I(1/x)+I(x^2)` Se la regressione è di tipo polinomiale per indicare $f(x) = a + bx + cx^2$ possiamo anche utilizzare `poly(x,3)` questo soprattutto per ottenere precisioni maggiori nei calcoli. Si noti però che i coefficienti sono diversi da quelli ottenuti con il procedeiemtno diretto.

35.13 La regressione pesata

Nell'ipotesi che il grafico dei residui contro una variabile esplicativa tenda ad allargarsi, o nell'ipotesi che il grafico dei valori teorici contro i residui abbia lo stesso andamento dobbiamo considerare come già detto l'ipotesi di eteroschedasticità e quindi una delle ipotesi base del modello lineare non può più essere considerata valida. Oltre alla soluzione della scelta degli stimatori ridge possiamo in questo caso cercare di eliminare la variabilità introducendo la regressione pesata. Tale regressione viene effettuata con il solito comando `lm` introducendo però il parametro:
`weights=`
ossia un vettore di nulghezza pari ad `n`. Con la presenza di tale parametro verrà minimizzata la funzione:

$$\sum(we^2)$$

I valori degli elementi del vettore dipendono da varie considerazioni. Nel caso di valori di variabile esplicativa ripetuti potrebbero essere il reciproco della varianza della variabile dipendente ripetuto per la numerosità di ogni modalità della variabile ma si potrebbe anche usare il reciproco della numerosità di ciascun valore della variabile esplicativa.

35.14 La previsione

Una volta stimato il modello di regressione e memorizzato nella variabile `anar.lm` è possibile calcolare i valori predetti nel seguente modo:

- `>predict(anar.lm)` se vogliamo calcolare i valore predetti per i dati del data frame utilizzato per stimare i coefficienti della regressione
- `>predict(anar.lm,nuovodataframe)` se vogliamo stimare i valori predetti per nuovi valori delle variabili contenuti nella varaibile `nuovodataframe`
- `>predict(anar.lm,se.fit=T)` se vogliamo anche ottenere lo standard error dei valori predetti per il primo caso
- `>predict(anar.lm■nuovodataframe se.fit=T)` se vogliamo anche ottenere lo standard error dei valori predetti nel secondo caso riportato

A questo punto occorre aprire una piccola parentesi. Supponiamo di effettuare una regressione lineare con due variabili, una dipendnete la y e una indipendente la x. Se x_0 è un particolare valore della variabile indipendente, la funzione di regressione lineare assume in tal punto il valore di

$$M(Y_0|x_0) = \beta_0 + \beta_1 x_0$$

possiamo quindi ottenere un intervallo di confidenza per

$$M(Y_0|x_0$$

detto intervallo di confidenza del valore predetto. In R ciò è fatto con il comando:
`>predict(anar.lm,interval=confidence)`

Nelle applicazioni risulta però più interessante fornire indicazioni di tipo intervallare non sulla media ma su una nuova possibile osservazione. Parliamo allora in questo caso di intervallo di previsione che in R può essere ottenuto con il comando:
`>predict(anar.lm,interval=prediction)`

Si noti che:

- gli intervalli di confidenza dei valori predetti forniscono una indicazione sulla qualità della regressione stessa
- gli intervalli di predizione sono generalmente sono più ampi ed esprimono l'attendibilità previsiva della regressione stessa
- mentre gli intervalli di confidenza si schiacciano sulla retta di regressione al crescere delle osservazioni campionarie, le bande di previsione si mantengono sempre ad una certa distanza dalla retta stessa

Ottenuti gli intervalli di confidenza e di predizione e memorizzati nella variabile `xxx` una rappresentazione grafica degli stessi può avvenire nel seguente modo:

- `>matplot(varindipendente,xxx)`
- `>points(varindipendente,vardipendente)`

Un altro modo per rappresentare gli intervalli di confidenza e predizione è il seguente:

- `>x<-rnorm(15)`
- `>y<-rnorm(15)+x`
- `>xxx<-lm(y x)`
- `>new<-data.frame(x=seq(-3,3,0.5))`
- `>aaa<-predict(xxx, new,interval=prediction)`
- `>bbb<-predict(xxx, new,interval=confidence)`
- `>matplot(new$x,cbind(aaa,bbb[,-1])lty=c(1,2,2,3,3),type=l)`

35.15 Lo stimatore ai minimi quadrati generalizzati

Se dall'analisi grafica effettuata con lo scatter plot dei valori stimati contro i valori residui otteniamo che tale grafico al posto di essere contenuto in una strisica regolare di piano ha un andamento diverso, l'ipotesi di omoschedasticità non può più essere ritenuta valida e quindi il modello dovrà essere stimato considerando lo stimatore ai minimi quadrati generalizzato. In generale lo stimatore ai minimi quadrati generalizzati indicato con `gls` viene utilizzato quando i residui non hanno struttura della covarianza standard ma vengono di fatto a cadere le ipotesi base del modello lineare ossia:

- mancanza di omoschedasticità
- mancanza di incorrelazione

Per utilizzare in R tale stimatore si dovranno installare due pacchetti aggiuntivi:

- `nls`
- `nlme`

Caricati tali pacchetti lo stimatore potrà essere ottenuto con i seguenti comandi:

- `gls(model,data,correlation)`
- `gls(model,data,weights)`
- `gls(model,data,correlation,weights)`

in cui per vedere:

- i tipi di `correlation` si veda l'help di `corClasses`
- i tipi di `weights` si veda l'help di `varClasses`

Se l'eteroschedasticità è dovuta al fatto che la matrice Ω presenta sulla diagonale principale i reciproci dei numeri`12,6,11,10,11` possiamo dare il seguente comando:
`>n<-c(12,6,11,10,11)`
`>gls(y~x1+x2+....,weights=varFunc(~1/n))`
oppure `>gls(y~x1+x2+....,weights=varFixed(~1/n))`
naturalmente si suppone che l'analisi è fatta su cinque osservazioni. Si noti che è possibile ottenere più formulazioni della funzione di varianza.

Se gli errori non sono tra di loro indipendenti ma sono correlati con tipo `Ar1`, nell'ipoesi che $\phi = 0.10''$ lo stimatore ai minimi quadrati generalizzato potrà essere ottenuto con il seguente comando:
`>gls(y~x1+x2+....,correlation=corAR1(0.10,form=~1,T))`
Tale tipo di correlazione può essere visto con l'analisi grafica dei residui che otteniamo dai grafici delle serie storiche. Utili sono i grafici di

- `>plot(xxx.gls)`
- `>plot(acf(xxx.gls)`

Si ricordi inoltre che se abbiamo a disposizione due modelli è anche possibie effettuare una anlisi della varianza congiunta su tali modelli con il comando:
`>anova(modello1,modello2)`

35.16 Modelli lineari generalizzati

I modelli lineari generalizzati sono ottenibili in R con il comando
`glm(formula,family,data,weight)`
si veda l'help in linea di tale comando per gli ulteriori approfondimenti.

35.17 La regressione robusta

La regressione robusta può essere ottenuta in R precaricando il pacchetto
`lqs` si veda l'help in linea del comando `lqs`.

36 L'analisi della varianza

36.1 Introduzione

L'analisi della varianza consiste in una serie di tecniche volte principalmente alla verifica dell'uguaglianza delle medie tra popolazioni sottoposte a trattamenti differenti. I dati per lo studio dell'analisi della varianza sono composti principalmente da un data frame che nel corso del paragrafo chiameremo sempre `anav` composto da più variabili, la prima che chiameremo sempre `y` ,sempre di tipo numerico, rappresenta il vettore delle osservazioni mentre le altre che indicheremo con `x1,x2,...`, sempre di tipo factor, rappresentano i fattori a cui sono state sottoposte le singole osservazioni. Ogni fattore sarà composto da più modalità. Si avrà allora che:

- nel caso di un solo fattore per trattamenti si intendono le modalità con cui si manifesta il fattore medesimo
- nel caso di due fattori i trattamenti per trattamenti si intendono tutte le coppie di modalità con cui si manifestano i due fattori in gioco
- nel caso di tre o più fattori l'estenzione è naturale

E' molto importante che i dati delle variabili `x1,x2,...` siano di tipo factor, in caso contrario i risultati potrebbero essere errati. Si noti in ogni caso che se i dati sono inseriti in un file ed importati con il comando `read.table` essi vengono, in caso non siano numerici, convertiti automaticamente in un oggetto di tipo factor. Bisognerà quindi prestare attenzione al caso in cui i dati delle varibili suddette siano numerici e quindi una volta importati dovranno essere convertiti manualmente in oggetti di tipo factor.

36.2 Analisi con un solo fattore

Ci occuperemo in questo paragrafo del caso più semplice dell'analisi della varianza ossia quella relativa al fatto di avere a disposizione il data frame `anav` composto dalla variabile risposta `y` e da un'unica varuiabile fattore `x1` avente come modalità `a,b,c,d`.

Una prima analisi dei dati del data frame `anav` finalizzata ad avere un primo tipo di risposta relativa alla uguaglianza delle medie relative ai diversi trattamenti a cui è stata sottoposta la variabile fattore `x1` consiste nella costruzione di un grafico che evidenzia le medie e le mediane delle variabili dei singoli trattamenti. Ciò può essere effettuato tramite la seguente sequenza di comandi:
`>attach(anav)`
`>par(mfrow=c(1,2))`
`>plot(c(1,1,1,1),tapply(y,x1,mean),type=`"n") il numero di 1 dipende dal numero di modalità di `x1`
`>text(c(1,1,1,1),tapply(y,x1,mean),c(`"a","b","c","d"))
`>plot(c(1,1,1,1),tapply(y,x1,median),type=`"n") il numero di 1 dipende dal numero di modalità di `x1`
`>text(c(1,1,1,1),tapply(y,x1,median),c(`"a","b","c","d"))
Un altro grafico molto interessante da analizzare è il boxplot relativo alla della variabile `y` suddiviso per le singole modalità di `x1`. Tale grafico si ottiene con il comando:
`>plot(y~x1)`
E possibile costruire una funzione la quale automatizza la procedura per la creazione dei grafici.

L'analisi della varianza ad un fattore viene effettuata con il seguente comando:
`>ana<-aov(y~x1)`
`>summary(ana)`

Molto interessante è anche il caso in cui l'analisi non deve essere compiuta su tutti i dati a disposizione ma solamente su un sottoinsieme che può essere ricavato anche da an'altra variabile presente nel data frame `anav`. In questo caso dovremo dare il seguente comando:
`>ana<-aov(y~x1,subset=(variabile==valore)`
`>summary(ana)`
in cui variabile indica in nome della variabile attraverso la quele estraiamo i dati desiderati, il simbolo `==` può essere sostituito a seconda dei casi da un qualunque operatore relazione, in nome valore indica il criterio di estrazione dei dati.

L'opzione subset può anche essere usata in questo modo, supponiamo che da una analisi grafica o altro abbiamo deciso che alcuni valori del data frame non debbano essere usati nel calcolo. Stabiliti tali valori ad esempio quelli della riga1,riga5 e riga10 dobbiamo costruire un vettore:
`outliers<-c(1,5,10)`
e dare il comando:
`>ana<-aov(y~x1,subset=-outliers`
`>summary(ana)`

Si noti che l'analisi della varianza ad un fattore equivale ad una analisi di regressione con i minimi quadrati in cui le variabili indipendenti sono tante variabili di tipo variabili dummy che valgono 1 se la risposta proviene dalla modalità del fattore e zero in caso contrario quante sono le modalità con cui si presenta il fattore `x1`. A questo punto per compiere una più approfondita analisi possiamo utilizzare i seguenti comandi: `>plot(ana)` per visualizzare i grafici relativi all'analisi della varianza
`>coefficients(ana)` per vedere i coefficienti della regressione ai minimi quadrati
`>residuals(ana)` per vedere i residui della regressione ai minimi quadrati
`>fitted.values(ana)` per vedere i valori stimati della regressione
`>model.tables(ana)`
`>model.tables(ana,type=`"means")

Compiuta l'analisi della varianza, possiamo passare all'analisi dei residui volta ad identificare se i residui derivanti dalla analisi fatta abbiano una distribuzione di tipo normale oppure no. Infatti affinchè l'analisi della varianza sia soddisfacente i residui dovrebbero disporsi in forma di una normale. Tale analisi può essere effettuata con le tecniche viste nel capitolo relativo all'adattamento dei dati ad una distribuzione prefissata, in ogni caso è anche molto utile un'analisi grafica dei residui che si ottiene attraverso i seguenti comandi:
`>hist(resid(ana))`
`>qqnorm(resid(ana))`.

Ricordiamo inoltre che per effettuare l'analisi della varianza è essenziale che le varianze delle popolazioni da cui i campioni provengono siano uguali. Ciò può essere verificato utilizzando il test di bartlett con il seguente comando:
`>bartlett.test(y~x1)`
dal quale si evince se si deve accettare o respingere l'uguaglianza delle varianze tra le varie popolazioni.

Nel caso in cui l'analisi della varianza porti a rifiutare l'ipotesi della uguaglianza delle medie una prima indicazione realtiva a quali possono essere le coppie di medie tra loro diverse potrà essere effettuata utilizzando i test legati alla t di student ed in particolare si potrà utilizzare il seguente comando
`>pairwise.t.test(y,x1)`

36.3 Analisi con due fattori non replicati

L'analisi della varianza con due fattori non replicati si ha quando si prendono in considerazione due fattori e per ogni coppia di modalità relativa ai due fattori si effettua una sola osservazione. Sia il data frame `anav` composto dalla variabile risposta `y` , da una variabile fattore `x1` avente come modalità `a,b,c,d` e da una variabile fattore `x2` avente come modalità `A,B,C,D`. In questo caso per ogni combinazione di modalità dei fattori `x1` e `x2` supponiamo di disporre di una sola osservazione `y`.

Una prima analisi dei dati del data frame `anav` finalizzata ad avere un primo tipo di risposta relativa alla uguaglianza delle medie relative ai diversi trattamenti a cui è stata sottoposta la variabile fattore `x1` consiste nella costruzione di un grafico che evidenzia le medie e le mediane delle variabili dei singoli trattamenti. Ciò può essere effettuato tramite la seguente sequenza di comandi:
`>attach(anav)`
`>par(mfrow=c(1,2))`
`>plot(c(1,1,1,1),tapply(y,x1,mean),type=`"n") il numero di 1 dipende dal numero di modalità di `x1`
`>text(c(1,1,1,1),tapply(y,x1,mean),c(`"a","b","c","d"))
`>plot(c(1,1,1,1),tapply(y,x1,median),type=`"n") il numero di 1 dipende dal numero di modalità di

x1
>text(c(1,1,1,1),tapply(y,x1,median),c("a","b","c","d"))
Un altro grafico molto interessante da analizzare è il boxplot relativo alla della variabile y suddiviso per le singole modalità di x1. Tale grafico si ottiene con il comando:
>plot(y~x1)
I comandi per il fattore x2 sono analoghi.

Nel caso di due fattori è anche necessario analizzare le interazioni tra le modalità dei due fattori considerati e tale analisi può essere compiuta con il comando:
>interaction.plot(y,x1,x2)
>interaction.plot(y,x1,x2,median)

L'analisi della varianza a due fattori viene effettuata con il seguente comando:
>ana<-aov(y~x1+x2)
>summary(ana)

A questo punto per compiere una più approfondita analisi possiamo utilizzare i seguenti comandi:
>plot(ana) per visualizzare i grafici relativi all'analisi della varianza
>coefficients(ana) per vedere i coefficienti della regressione ai minimi quadrati
>residuals(ana) per vedere i residui della regressione ai minimi quadrati
>fitted.values(ana) per vedere i valori stimati della regressione
>model.tables(ana)
>model.tables(ana,type="means")

Compiuta l'analisi della varianza, possiamo passare all'analisi dei residui volta ad identificare se i residui derivanti dalla analisi fatta abbiano una distribuzione di tipo normale oppure no. Infatti affinchè l'analisi della varianza sia soddisfacente i residui dovrebbero disporsi in forma di una normale. Tale analisi può essere effettuata con le tecniche viste nel capitolo relativo all'adattamento dei dati ad una distribuzione prefissata, in ogni caso è anche molto utile un'analisi grafica dei residui che si ottiene attraverso i seguenti comandi:
>hist(resid(ana))
>qqnorm(resid(ana)).

Ricordiamo inoltre che per effettuare l'analisi della varianza è essenziale che le varianze delle popolazioni da cui i campioni provengono siano uguali. Ciò può essere verificato utilizzando il test di bartlett con il seguente comando:
>bartlett.test(y~x1)
>bartlett.test(y~x2)
dal quale si evince se si deve accettare o respingere l'uguaglianza delle varianze tra le varie popolazioni.

Nel caso in cui l'analisi della varianza porti a rifiutare l'ipotesi della uguaglianza delle medie una prima indicazione realtiva a quali possono essere le coppie di medie tra loro diverse potrà essere effettuata utilizzando i test legati alla t di student ed in particolare si potrà utilizzare il seguente comando
>pairwise.t.test(y,x1) >pairwise.t.test(y,x2)

36.4 Analisi con due fattori replicati

L'analisi della varianza con due fattori non replicati si ha quando si prendono in considerazione due fattori e per ogni coppia di modalità relativa ai due fattori si effettuano più osservazioni. Sia il data frame anav composto dalla variabile risposta y , da una variabile fattore x1 avente come modalità a,b,c,d e da una variabile fattore x2 avente come modalità A,B,C,D. In questo caso per ogni combinazione di modalità dei fattori x1 e x2 supponiamo di disporre di più osservazioni di y.

A differenza del caso precedente, in questa situazione possiamo anche effettuare la costruzione di un modello non solo additivo ma che tenga conto delle interrelazioni tra i due fattori. In questo caso dunque oltre il modello:
>ana<-aov(y~x1+x2)
potremmo considerare anche il modello:
>ana<-aov(y~x1*x2)

36.5 Analisi con più di due fattori

In questo caso per ottenere l'analisi della varianza basterà generalizzare quanto detto nei paragrafi precedenti.

36.6 Confronti multipli

Se l'analisi della varianza porta a rifiutare l'ipotesi dell'uguaglianza delle medie tra le varie modalità di un unico fattore potremmo essere interessati a rilevare quale coppia di modalità ha creato il rigetto di tale ipotesi. Questa analisi consiste nel calcolo dei confronti multipli. Se siamo interessati solamente al metodo dei confronti multipli di Tukey possiamo eseguire i seguenti comandi:

```
 > ana<-aov( y~x1)
> tHSD<-TukeyHSD(ana, "x1",ordered=FALSE)
> tHSD
> plot(tHSD)
```

Se l'analisi invece è a due fattori dobbiamo utilizzare i seguenti comandi:

```
> ana<-aov( y~x1+x2)
> tHSD<-TukeyHSD(ana, "x1 o x2",ordered=FALSE)
> tHSD
> plot(tHSD)
```

Se invece siamo interessati ad altri metodi di confronto dobbiamo installare i pacchetti:

- mvtnorm
- multcomp

ed eseguire i seguenti comandi:

```
> mca<-simint(y~x1, data=anav,method="Tukey")
> mca
```

Se l'analisi invece è a due fattori dobbiamo utilizzare i seguenti comandi:

```
> mca<-simint(y~x1+x2, data=anav,whichf="x1",method="Tukey")
> mca
```

I metodi che possono essere utilizzati sono i seguenti:

- Tukey
- Dunnett
- Sequen
- AVE
- Changepoint
- Williams
- Marcus
- McDermott
- Tetrade

Il comando `simint` presenta una serie di opzioni molto interessanti si veda l'help in linea per il loro utilizzo.

Si noti inoltre che con l'uso di `simint` è anche possibile ottenere altri valori molto utili per l'analisi statisitca dati da:

- il punto critico che si ottiene con `mca\$calpha`
- gli standar error che si ottengono con `\$sd`

36.7 Maggiori dettagli nell'analisi della varianza

Da un'analisi statistica effettuata abbiamo ricavato i dati che sono evidenziati nella tabella 1 L'analisi della varianza come visto si ottiene con il comando:

```
>survival.aov<-aov(dati~poison*treatment)
```

in quanto si tratta di una analisi della varianza a due fattori replicati. Per ottenere le informazioni sull'analisi della varianza possiamo eseguire il comando:

```
summary(survival.aov)
```

ma possiamo anche ottenere informazioni più dettagliate utilizzando il seguente comando:

```
summary(survival.aov,split=list(poison=list(1"=1,"2"=2,"3"=3,"4"=4),
treatment=list(a"=1,"b"=2,"c"=3))
```

dati	poison	treatment
0,31	1	a
0,45	1	a
0,46	1	a
0,43	1	a
0,36	2	a
0,29	2	a
0,4	2	a
0,23	2	a
0,22	3	a
0,21	3	a
0,18	3	a
0,23	3	a
0,82	1	b
1,1	1	b
0,88	1	b
0,72	1	b
0,92	2	b
0,61	2	b
0,49	2	b
1,24	2	b
0,3	3	b
0,37	3	b
0,38	3	b
0,29	3	b
0,43	1	c
0,45	1	c
0,63	1	c
0,76	1	c
0,44	2	c
0,35	2	c
0,31	2	c
0,4	2	c
0,23	3	c
0,25	3	c
0,24	3	c
0,22	3	c
0,45	1	d
0,71	1	d
0,66	1	d
0,62	1	d
0,56	2	d
1,02	2	d
0,71	2	d

Tabella 1: varianza

36.8 Analisi della varianza multivariata

L'analisi dellavarianza multivariata viene eseguita con il comando `manova` nel seguente modo. Supponiamo di disporre di un data frame chiamato `Y` ottenuto nel seguente modo:

```
tear <- c(6.5, 6.2, 5.8, 6.5, 6.5, 6.9, 7.2, 6.9, 6.1, 6.3,
                6.7, 6.6, 7.2, 7.1, 6.8, 7.1, 7.0, 7.2, 7.5, 7.6)
gloss <- c(9.5, 9.9, 9.6, 9.6, 9.2, 9.1, 10.0, 9.9, 9.5, 9.4,
                9.1, 9.3, 8.3, 8.4, 8.5, 9.2, 8.8, 9.7, 10.1, 9.2)
opacity <- c(4.4, 6.4, 3.0, 4.1, 0.8, 5.7, 2.0, 3.9, 1.9, 5.7,
                2.8, 4.1, 3.8, 1.6, 3.4, 8.4, 5.2, 6.9, 2.7, 1.9)
Y <- cbind(tear, gloss, opacity)
```

Creiamo successivamente due oggetti factor nel seguente modo:

```
rate <- factor(gl(2,10), labels=c("Low", "High"))
additive <- factor(gl(2, 5, len=20), labels=c("Low", "High"))
```

L'analisi della varianza multivariata viene eseguita nel seguente modo:

```
fit <- manova(Y ~ rate * additive)
summary.aov(fit)
```

in cui si noti Y deve essere necessariamente una matrice.

Un altro modo per ottenere l'analisi della varianza multivariata consiste nel dare il seguente comando:
`>summary(fit, test=`"Wilks")
In cui il valore per la variabile test può essere uno dei seguenti:

- wilks
- Pillai
- Hotelling-Lawley
- Roy

36.9 La funzione anova

Tale funzione permette di effettuare una analisi della varianza e della devianza incrementando le variabili risposta nell'ordine in cui sono state inserite nei vari modelli da stimare. Si applica quindi alla costruzione di modelli lineari, di analisi della varianza, di modelli ols e gls. Ha come argomento fondamentale il modello stimato e come parametro opzionale il test che può essere chisq, F o cp. Per certi modelli viene scelto automaticamente dal programma come per i modelli lm e aov.

37 Analisi delle componenti principali

38 Analisi fattoriale

39 Analisi discriminante lineare

40 Analisi discriminante quadratica

41 Correlazione canonica

42 Cluster analysis

42.1 Introduzione

Vogliamo analizzare in questo paragrafo le principali funzioni che R mette a disposizione per effettuare la cluster analysis.

42.2 I pacchetti necessari

Per effettuare la cluster analysis è necessario installare il pacchetto `cluster`.

42.3 La funzione daisy

Una delle funzioni che utilizzere maggiormente nella cluster analysis è la funzione `daisy`. Tale funzione serve per calcolare a due a due tutte le distanze tra le osservazioni in un data frame. Il suo uso è il seguente:
`daisy(x, metric = c(euclidean,manhattan), stand = FALSE, type = list())`
Le varibili che si possono usare con la funzione `daisy` sono di qualsiasi tipo.

42.4 La funzione dist

Altra funzione molto importante è la funzione `dist`. Tale funzione calcola la distanza tra le righe di una matrice utilizzando uno dei metodi che ora indicheremo. Tale funzione si usa nel seguente modo:
`dist(x, method = euclidean, diag = FALSE, upper = FALSE, p = 2)`
in cui `method` può essere una delle seguenti opzioni:

- euclidean
- maximum
- manhattan
- canberra
- binary
- minkowki

Si veda l'help del comando per ulteriori approfondimenti

42.5 Metodi di analisi

Come sappiamo dalla teoria i metodi che possiamo utilizzare per effettuare la cluster analysis possono essere

- metodi non gerarchici quali: kmeans, pam, clara, fanny
- metodi gerarchici quali: hclust, agnes, diana, mona

Analizzeremo successivamente e in modo sintetico tali metodi facendo vedere i principali che implementano tali metodi con R. Per una descrizione approfondita del loro contenuto si rimanda all'help in linea.

42.6 Kmeans

Il metodo si ottiene con il seguente comando:
`kmeans(x,n)`
in cui:

- x è il data frame da analizzare
- n è il numero dei cluster ipotizzati

42.7 Pam

Tale metodo risulta essere più robusto di `kmeans` e siottiene con il seguente comando:
`pam(x,n)`
in cui:

- x è il data frame da analizzare
- n è il numero dei cluster ipotizzati

Con il comando:
`plot(pam(x,n))`
otteniamo anche una analisi grafica composta da due grafici:

- il cusplot che è il grafico che si ottiene con le due prime componenti principali dei dati considerati
- il silhouetteplot che è il grafico che per ogni oggetto di ogni gruppo produce una linea di lunghezza con range [-1;+1]. Se la lunghezza è vicina a +1 l'item risulta essere ben calssificato, se è vicino a -1 male classificato, se è vicino a zero potrebbe essere di altri gruppi.

42.8 Clara

Tale metodo, simile al metodo `pam` si utilizza quando il numero delle osservazioni è molto elevato e si ottiene con il seguente comando:
`clara(x,n)`
in cui:

- x è il data frame da analizzare
- n è il numero dei cluster ipotizzati

Con il comando:
`plot(clara(x,n))`
otteniamo la solita analisi grafica come per la funzione `pam`.

42.9 Fanny

Tale metodo, simile al metodo `pam` si utilizza quando il numero delle osservazioni è molto elevato e si ottiene con il seguente comando:
`fanny(x,n)`
in cui:

- x è il data frame da analizzare
- n è il numero dei cluster ipotizzati

Con il comando:
`plot(fanny(x,n))`
otteniamo la solita analisi grafica come per la funzione `pam`.

42.10 Hclust

Tale metodo si ottiene con il seguente comando:
`hclust(dist(x),method=complete)`
o con il comando:
`hclust(daisy(x),method=complete)`
in cui x è il data frame da analizzare. L'opzione `method` può essere una delle seguenti:

- ward
- single
- complete
- average
- mcquitty
- median
- centroid

Con il comando:
`plot(dist(x),method=complete)`
otteniamo la solita analisi grafica come per la funzione `pam`.

42.11 Agnes

Tale metodo si ottiene con il seguente comando:
`agnes(x)`
in cui x è il data frame da analizzare. Con il comando:
`plot(dist(x),method=complete)`
otteniamo la solita analisi grafica come per la funzione `pam`.

42.12 Diana

Tale metodo si ottiene con il seguente comando:
`diana(x)`
in cui x è il data frame da analizzare. Con il comando:
`plot(diana(x),method=complete)`
otteniamo la solita analisi grafica come per la funzione `pam`.

42.13 Mona

Tale metodo si ottiene con il seguente comando:
`mona(x)`
in cui x è il data frame da analizzare. Con il comando:
`plot(mona(x))`
otteniamo la solita analisi grafica come per la funzione `pam`.

43 Alcuni pacchetti aggiuntivi

43.1 Introduzione

Oltre ai pacchetti che sono automaticamente forniti con R, possiamo installare altri pacchetti per svolgere analisi statistiche molto complesse. Io propongo di installare altre al programma base i seguenti pacchetti aggiuntivi:

- suppdist che fornisce distribuzioni di probabilità aggiuntive rispetto a quelle base
- rcmdr che permette di relizzare una interfaccia grafica con R e richiede i pacchetti lattice, foreign, tcltk, abind, effects,car
- ineq per ottenere funzioni aggiuntive di statistica descrittiva
- tree per ottenere una analisi statistica utilizzando la tecnica degli alberi
- deal per ottenere e creare adai dati il bbn
- lpsolve e linprog per implementare problemi di ottimizzazione lineare

Importante per l'analisi statistica è anche l'utilizzo dei pacchetti che verrano descritti successivamente.

43.2 Il pacchetto rodbc

Tale pacchetto consente di utilizzare il sistema odbc di windows per inportare direttamente file excel, access e di altri tipologie. Senza entrare in ulteriori dettagli, ottenibili attraverso l'help in linea dei comandi che lo compongono, vediamo con un semplice esempio come importare in R un file di Access. Dobbiamo operare nel seguente modo:

- caricare in `R` il pacchetto `rodbc`
- in windows lanciare il programma `odbc` e in `user dns` creare il collegamento del file da importare e attrribuirgli un nome ad esempio `tabellone`
- per effettuare il collegamento tra access ed r dobbiamo dare il comando:
  `>channel<-odbcConnect(tabellone)`
- per vedere il nome delle tabelle dobbiamo dare il comando:
  `>sqlTables(channel)`

- e per importare tale tabella in R diamo il comando:
  `tabellone<-sqlFetch(channel,Foglio1)`

Si noti che se il file da utilizzare è in formato Excel è più conveniente utilizzare la funzione `odbcConnectExcel` nel seguente modo:

- caricare in R il pacchetto `rodbc`
- per effettuare il collegamento tra excel ed r dobbiamo dare il comando:
  `>channel<-odbcConnectExcel(directory)`
- per vedere il nome delle tabelle dobbiamo dare il comando:
  `>sqlTables(channel)`
- e per importare tale tabella in R diamo il comando:
  `tabellone<-sqlFetch(channel,Foglio1)`

43.3 Il pacchetto xtable

Molte volte accade di dover inserire in documenti scritti in LaTeXo in pagine web oggetti ottenuti con R. In prevalenza si tratta di inserire in tali documenti tabelle che sono state ottenute con elaborazioni effettuate da R. Ciò può essere effettuato facilmente con l'ausilio di un pacchetto aggiuntivo il pacchetto `xtable`. Tutto quanto scritto in questo paragrafo presuppone che tale pacchetto sia stato già precaricato in R.

Supponiamo di aver già creato in R una tabella e di averla memorizzata nella variabile `mytabella`. Per potere esportare tale tabella in un file da inserire in un documento LaTeXdobbiamo dare i seguenti comandi:
`>print(xtable(mytabella),file=`"/percorso/nomefile.tex")
Verrà creato un file dal nome `nomefile.tex` che potrà essere modificato come previsto da LaTeXe inserito in un qualunque documento LaTeX.

Se l'oggetto di cui vogliamo creare il formato html non è una tabella, basterà convertirlo con l'opzione `as.matrix` o `as.data.frame`.

Supponiamo di aver già creato in R una tabella e di averla memorizzata nella variabile `mytabella`. Per potere esportare tale tabella in un file html dobbiamo dare i seguenti comandi:
`>print(xtable(mytabella),type=`"html",file="/percorso/nomefile.html")
Verrà creato un file dal nome `nomefile.html` che potrà essere modificato come previsto dal linguaggio html e inserito in una qualunque pagina web.

Se l'oggetto di cui vogliamo creare il formato html non è una tabella, basterà convertirlo con l'opzione `as.matrix` o `as.data.frame`.

Il comando `xtable` prevede numerose opzioni per formattare l'output. Si veda a proposito l'help in linea di tale comando.

43.4 Il pacchetto r2html

Potrebbe essere molto utile ottenere delle pagine web direttamente da R. Ciò può essere fatto con l'uso del pacchetto `R2HTML` che dovrà quindi essere caricato preventivamente in R prima del suo utilizzo.

Per creare pagine web dobbiamo dare la seguente sequenza di comandi:
`library(R2HTML)`
`HTMLStart(outdir=`"/home/roberto",filename="nomedelfile")
si noti che deve essere inserita la barra prima di home e non dopo roberto. A questo punto tutto ciò che digiteremo nella console di R compresi irisultati saranno inseriti in una pagine html e consultabili con un normale browser.

Per approfondimenti si faccia riferimento all'ottimo help dei pacchetti `R2HTML`.

44 R e linux

44.1 Introduzione

Se utilizziamo R nel sistema operativo linux abbiamo alcune particolarità da ricordare. Nei prossimi paragrafi vedremo quali sono

44.2 Intallazione del pacchetti aggiuntivi

Per aggiungere dei pacchetti aggiunti dobbiamo eserguire i seguenti comandi:
`R CMD INSTALL nome pacchetto`
in questo modo il pacchetto sarà automaticamente installato in R.

Notiamo che per poterlo usare dobbiamo richimarlo dopo aver avviato R con il comando:
`library(nome pacchetto)`
e questo per ogni pacchetto presente in R

44.3 Utilizzo del pacchetto RMySQL

Il pacchetto `RMySQL` permette di interfacciare direttamente R con il programma `MySQL` consentendo quindi di reperire direttamente dal data base di `MySQL` i dati da utilizzare in R. Per poter utilizzare tale pacchetto donbbiamo prima di tutto installare i pacchetti aggiuntivi:

- `DBI`
- `RMySQL`

Una volta installati essi dovranno essere richiamati per l'uso con:

- `library(DBI)`
- `library(RMySQL)`

Quindi per importare i dati in R dobbiamo utilizzare la seguente sequenza di comandi:

- drv<-dbDriver(MySQL)
- con<-dbConnect(drv,nome data base,user,password)
- dbListTables(con) per ottenere la lista delle tabelle
- variabile<-dbReadTable(con,nome tabella) per memorizzare in R il nome della tabella

Sono possibili molti altri comandi ai quali si rimanda con gli aiuti forniti dal programma stesso.

44.4 Ottenere un grafico in eps

Può accadere molte volte di dover inserire grafici creati con R in altri documenti scritti il linguaggio LaTeX. E' quindi necessario conoscere un metodo che ci consenta di salvare i grafici ottenuti in un faormato leggibile da tale programma ossia il formato eps. Per ottenere un grafico in formato `eps` possiamo usare la seguente sequenza di comandi:
`>postscript`("percorso/"nomefile.eps")
`>hist(x)` ad esempio
`>dev.off()` Il grafico così ottenuto potrà essere utilizzato con la consueta sintassi in ogni documento LaTeX.

Elenco delle figure

Indice