Introduzione a R

Andrea Benazzo

E-mail: andrea.benazzo@unife.it

Dip. Scienze della vita e Biotecnologie

Caratteristiche di R

- Linguaggio ad alto livello *interpretato*
- Dotato di insiemi di operatori ad alto livello per *calcoli su array e matrici*
- Fornisce un ambiente per la *elaborazione interattiva* dei dati
- *Ambiente integrato di risorse software* per la gestione ed elaborazione di dati e la visualizzazione di grafici
- Dispone di *interfacce* verso programmi e moduli software scritti con altri linguaggi
- Ambiente di sviluppo e package open source disponibili liberamente in internet.

Breve storia di R

- Deriva da *S*, un linguaggio ed un sistema sviluppati da *John Chambers* e collaboratori negli anni ‘80 presso i Laboratori Bell.
- *R è un progetto Open Source* conforme per la maggior parte ad S:
  - Sviluppato inizialmente da *Ross Ihaka and Robert Gentleman* all’ Università di Auckland (Nuova Zelanda)
  - Attualmente sviluppato da una comunità internazionale di ricercatori e sviluppatori in ambito sia accademico sia industriale
  - Opera attraverso il web: www.r-project.org
  - Archivi software e documentazione: cran.r-project.org/

R

Advantages

- Fast and free.
- State of the art: Statistical researchers provide their methods as R packages. SPSS and SAS are years behind R!
- 2nd only to MATLAB for graphics.
- Active user community
- Excellent for simulation, programming, computer intensive analyses, etc.
- Forces you to *think* about your analysis.
- Interfaces with database storage software (SQL)

Disadvantages

- Not user friendly @ start - steep learning curve, minimal GUI.
- No commercial support; figuring out correct methods or how to use a function on your own can be frustrating.
- Easy to make mistakes and not know.
- Working with large datasets is limited by RAM
- Data prep & cleaning can be messier & more mistake prone in R vs. SPSS or SAS

Nozioni di sintassi

- Una volta che R è stato lanciato, tutte le istruzioni sono eseguibili dalla linea di comando dell'ambiente.

Nel suo utilizzo più semplice, R può essere utilizzato come calcolatrice:

> 3+5*3.5

[1] 20.5

- > 3+5*(3.5/15)+5-(2/6*4)
- [1] 7.833333

Operatori

- Aritmetici: +,-,*,/,^

  (2+2)^2

- Relazionali: >,<,>=,<=,==,!=

  4>2

- Logici: &,|

  4>2 & 5<6

Variabili

- Per assegnare un valore numerico alla variabile x si usa il comando:

  > x<-2

mentre per assegnare una stringa alla variabile y si usa il comando:

> y<-“casa“

Per visualizzare il contenuto di una variabile basterà digitare il nome della variabile stessa:

> x

Assegnando un nuovo valore ad una variabile, verrà automaticamente cancellato il valore precedentemente assunto dalla stessa.

Gli oggetti base

- Vettori

  7,8,9,5,6,4,7,8,9

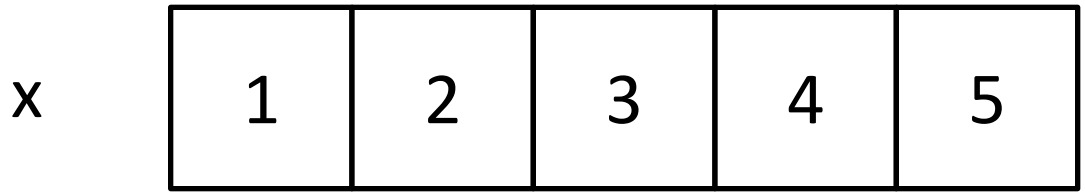
- Matrici e data-frame

	Col1	Col2	Col3
Rig1	25	12	7
Rig2	78	14	1

- Liste (vettori di oggetti)

Vettori

I vettori sono dati dello stesso tipo che sono raggruppati in una unica variabile:



Creare un vettore

- La funzione **c**:

  >x<-c(1.5,2,2.5)

  >x<-c(“questo“,“è“,“un esempio“)

  >x<-1:10

- Prova:
- >x<-c(1, “casa“)

Richiamare gli elementi di un vettore

- >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

Funzioni di base

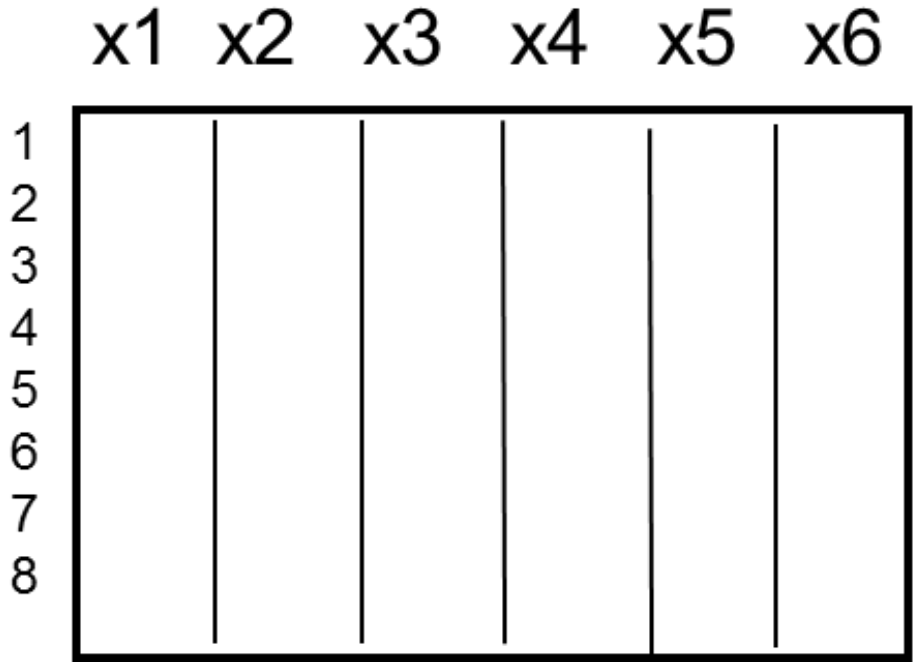
- Dato un vettore di tipo numerico x, le principali funzioni elementari statistiche applicabili a tale vettore sono le seguenti:
- >**length(x)** restituisce il numero di elementi in x
- >**min(x)** restituisce il minimo di x
- >**max(x)** restituisce il massimo di x
- >**range(x)** restituisce il range di x
- >**mean(x)** restituisce la media aritmetica semplice di x
- >**median(x)** restituisce la mediana di x
- >**quantile(x,y)** con y numero compreso 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
- >**sum(x)** restituisce la somma degli elementi in x

Esercizio 1

- Costruire un vettore di 10 elementi numerici interi e calcolare:
  - Minimo e massimo,
  - La media e la mediana,
  - La varianza (attraverso il calcolo manuale e attraverso la formula)

Matrici

- Le matrici sono costituite da **dati dello stesso tipo** che sono raggruppati in tabelle a doppia entrata.



Esempi di matrici

Matrice numerica

```
     [,1] [,2]
[1,]    1   10
[2,]  100 1000
```

Matrice di caratteri

```
     [,1] [,2]
[1,] "a"  "c"
[2,] "b"  "d"
```

Creare una matrice

La funzione **matrix**:

>help(matrix)

>x<-matrix(1:10,ncol=5,nrow=2)

>x<-matrix(c(1:5,8:12,22:26),ncol=5,nrow=3,byrow=T)

Nomi di riga e colonna:

row.names(x)<-c("rig1", "rig2", "rig3")

col.names(x)<-c("c1", "c2", "c3«, "c4", "c5"))

Accedere ai dati in una matrice

x[*id-riga*,*id-colonna*]

- x[1,1] richiama l'elemento di posto (1; 1)
- >x[,"nomecol1"] equivalente a x[,1] richiama la prima colonna di x
- >x["nomerig1",] equivalente a 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[-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

Esercizio 2

- Creare una matrice 10x10 con i numeri da 1 a 100 e calcolare:
  - Assegnare i numeri da 1 a 10 come nomi di riga e colonna
  - La media della prima colonna
  - La somma della prima riga
  - La somma degli elementi della prima e ultima colonna

Data frame

- I data frame sono oggetti simili alle matrici ma con colonne che **possono contenere dati di diverso tipo**.

Esempio:

	a	sesso
1	1	M
2	2	F
3	3	F
4	4	M

Creare un data frame

- La funzione data.frame():

```
x<-data.frame(a=1:4, sesso=c("M","F","F","M"))
```

- Aggiungere una variabile:

```
x$eta<-c(2.5,3,5,6.2)
```

Selezionare gli elementi

- In maniera analoga alle matrici possiamo selezionare righe e colonne tramite l'id, oppure tramite il nome riga/colonna:

X[1,] (prima riga)

X[,1] (prima colonna)

X$nome (visualizza la colonna «nome»)

Esercizio 3

- Creare il data frame di esempio con tre variabili, a, sesso ed età:

```
x<-data.frame(a=1:4,
sesso=c("M","F","F","M"),eta=c(15,18,36,65))
```

- seleziona i valori della variabile eta per i maschi
- seleziona i valori della variabile sesso per cui eta<=30

Importazione dati da file

La funzione read.table() è il modo più semplice di importare dati in R presenti su files.

```
> X<-read.table(file="c:/documenti/dati.txt",
+ header=TRUE,
+ sep="\t",
+ na.strings = "NA",
+ dec=".")
```

Esercizio 4

- Creare una matrice 3x3 in un file di testo, e caricarlo in R con il seguente comando:

x<-read.table(choose.files())

- Specificare i nomi di riga e di colonna nel file e ricaricarlo

- Calcolare la media in ogni colonna

Esempi di funzioni statistiche

- Test T:

t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"),

mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95, ...)

x	a (non-empty) numeric vector of data values.
y	an optional (non-empty) numeric vector of data values.
alternative	a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less". You can specify just the initial letter.
mu	a number indicating the true value of the mean (or difference in means if you are performing a two sample test).
paired	a logical indicating whether you want a paired t-test.
var.equal	a logical variable indicating whether to treat the two variances as being equal. If TRUE then the pooled variance is used to estimate the variance otherwise the Welch (or Satterthwaite) approximation to the degrees of freedom is used.
conf.level	confidence level of the interval.
formula	a formula of the form lhs ~ rhs where lhs is a numeric variable giving the data values and rhs a factor with two levels giving the corresponding groups.
data	an optional matrix or data frame (or similar: see model.frame) containing the variables in the formula formula. By default the variables are taken from environment(formula).
subset	an optional vector specifying a subset of observations to be used.

Esempi di funzioni statistiche

- Chi-quadro

chisq.test(x, y = NULL, correct = TRUE, p = rep(1/length(x), length(x)), rescale.p = FALSE, simulate.p.value = FALSE, B = 2000)

x	a vector or matrix.
y	a vector; ignored if x is a matrix.
correct	a logical indicating whether to apply continuity correction when computing the test statistic for 2x2 tables: one half is subtracted from all $\|O - E\|$ differences. No correction is done if simulate.p.value = TRUE.
p	a vector of probabilities of the same length of x. An error is given if any entry of p is negative.
rescale.p	a logical scalar; if TRUE then p is rescaled (if necessary) to sum to 1. If rescale.p is FALSE, and p does not sum to 1, an error is given.
simulate.p.value	a logical indicating whether to compute p-values by Monte Carlo simulation.
B	an integer specifying the number of replicates used in the Monte Carlo test.

Esempi di funzioni statistiche

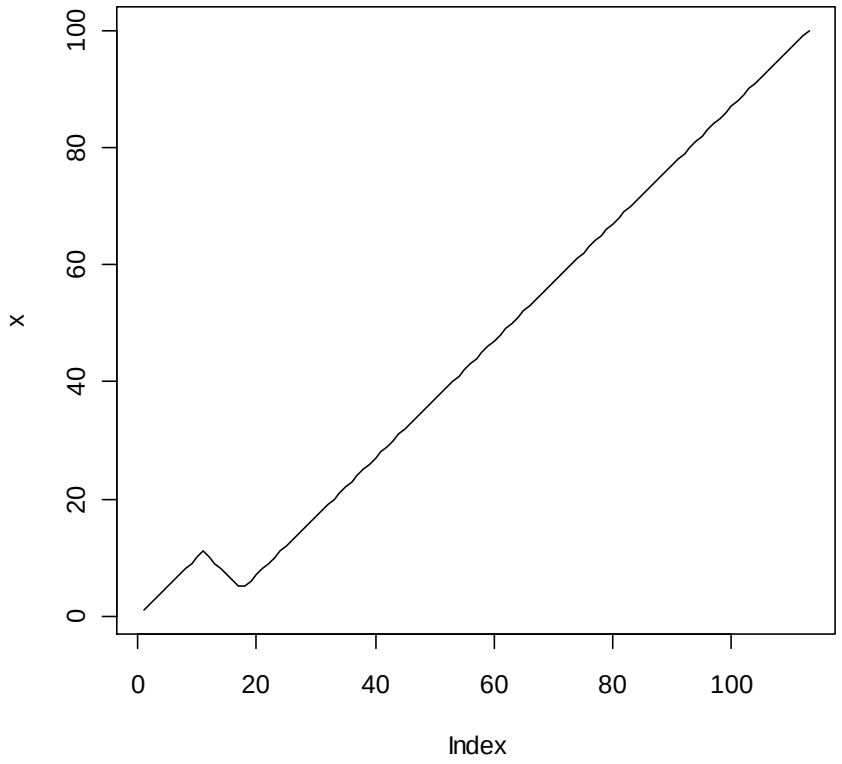
- Simulazione da una distribuzione normale:

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

Grafici

- Grafico a linee/punti

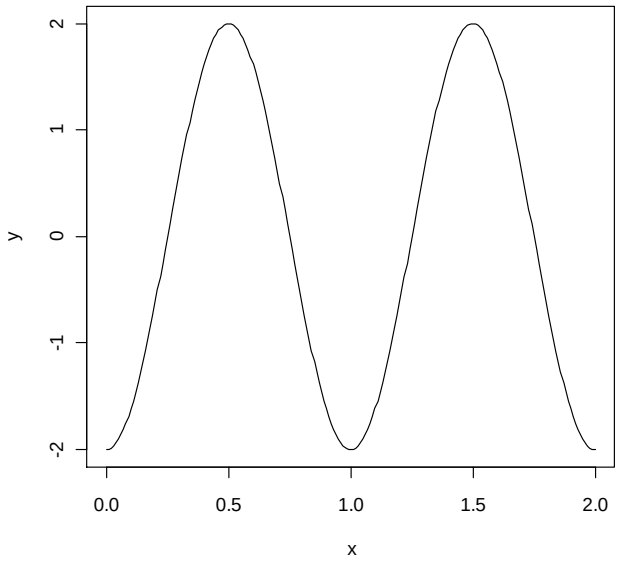
```
x<-c(1:10,11:5,5:100)

plot(x,type="l")
```



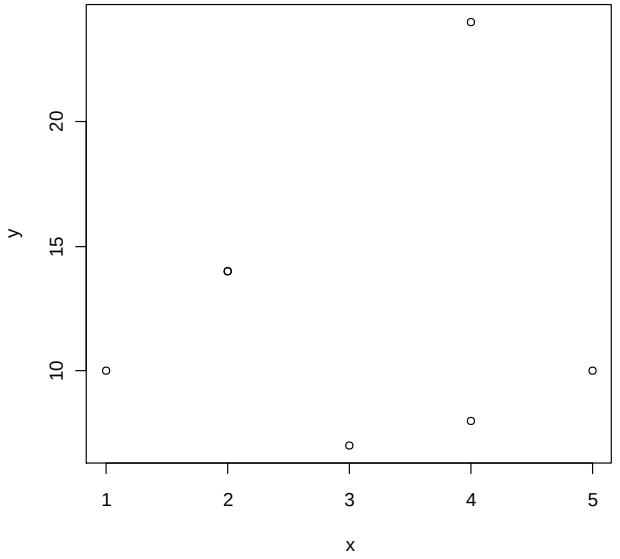
Grafici

- Grafico a linee/punti

```
x<-seq(0,2,by=0.01)
y<-2*sin(2*pi*(x-1/4))
matplot(x,y,type="l")
```



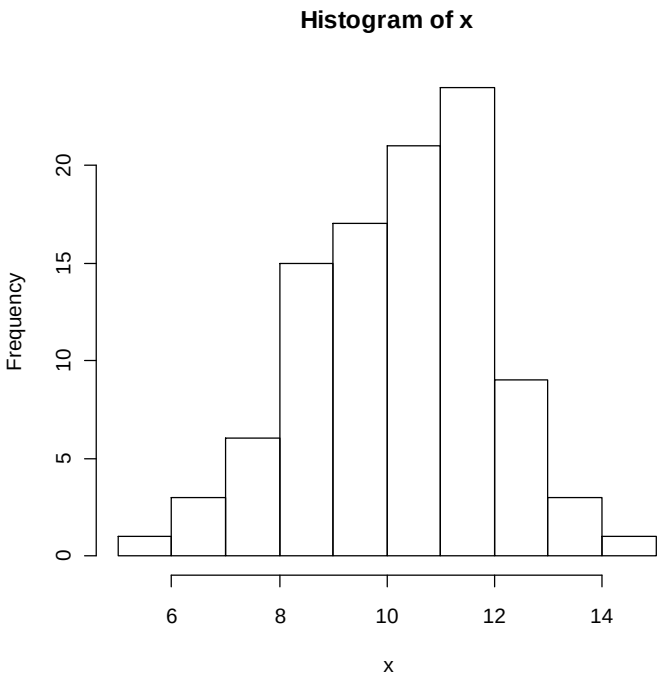

```
x<-c(1,5,2,4,2,3,4)
y<-c(10,10,14,24,14,7,8)
matplot(x,y,type="p",pch=1)
```



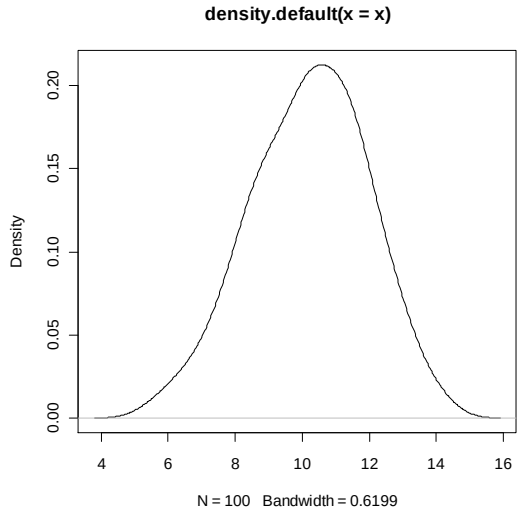
Grafici

- Istogramma/densità

```
x<-rnorm(100,10,2)
hist(x,breaks=10)
```




```
x<-rnorm(100,10,2)
plot(density(x))
```



Esercizio su T test

Nel mondo anglosassone, in cui le temperature vengono frequentemente espresse in gradi Fahrenheit, si attribuisce alla temperatura corporea normale il valore di 98,6 °F.

Ma questo valore trova realmente conferma nei dati?

Alcuni ricercatori hanno misurato la temperatura corporea, espressa in gradi Fahrenheit, in un campione casuale di soggetti sani.

Di seguito sono riportati i dati relativi a 25 soggetti:

98.40	98.6	97.8	98.8	97.9
99	98.2	98.8	98.8	99
98	99.2	99.5	99.4	98.4
99.1	98.4	97.6	97.4	97.5
97.5	98.8	98.6	100	98.4

Le misure sono compatibili con una media della popolazione di 98,6 °F?

Soluzione 1

```
#carico i dati nel vettore x
x<-c(98.40,98.6,97.8,98.8,97.9,99,98.2,98.8,98.8,99,98,99.2,
99.5,99.4,98.4,99.1,98.4,97.6,97.4,97.5,97.5,98.8,98.6,100,98.4)
#calcolo la media campionaria
med<-mean(x) #oppure med<-sum(x)/length(x)
#calcolo la deviazione standard
devst<-sd(x) #oppure devst<-sqrt(sum((x-med)^2/(length(x)-1)))
#calcolo la statistica T osservata
tc<-(med-98.6)/(devst/sqrt(n))
#calcolo i limiti della regione di accettazione
qt(p=0.975,df=24)
```

Soluzione 2

```
t.test(x=x, mu=98.6, alternative="two.sided")
```

```
#risultato
One Sample t-test
data:  x
t = -0.5606, df = 24, p-value = 0.5802
alternative hypothesis: true mean is not equal to 98.6
95 percent confidence interval:
 98.24422 98.80378
sample estimates:
mean of x
   98.524
```

Esercizio su test binomiale

In un campione di 30 studenti liceali, 5 scelgono di studiare Biologia.

Se la proporzione totale di liceali che scelgono biologia è del 30%, determinare se il campione studiato può considerarsi estratto dalla popolazione generale.

Soluzione 1

```
n<-30
p<-0,3
x<-0:30
#calcolo la probabilità di ogni evento se H0 vera
PB<-choose(30,x)*(p^x)*((1-p)^(n-x))
#grafico della distribuzione ottenuta
barplot(PB,names.arg=as.character(0:30),cex.names=0.5)
#calcolo il P-value approssimato
sum(PB[1:6])*2
```

Soluzione 2

```
binom.test(x=5,n=30,p=0.3,alternative="two.sided",
conf.level=0.95)
#risultato
Exact binomial test

data:  5 and 30
number of successes = 5, number of trials = 30, p-value = 0.1611
alternative hypothesis: true probability of success is not equal to 0.3
95 percent confidence interval:
 0.0564217 0.3472117
sample estimates:
probability of success
             0.1666667
```