# Corso di Statistica per Biotecnologi

Saverio Vicario: saverio.vicario@ba.itb.cnr.it

# La statistica

- I problemi conoscitivi si possono dividere:
  1. in quelli risolvibili nel chiuso di una stanza
  2. in quelli in cui e' necessario uscire e sperimentare

Nel primo caso ci troviamo tipicamente in un problema di tipo matematico

Nel secondo caso e' necessario, qualunque sia il campo di indagine, purché che interagisca con il reale, di un approccio statistico

# La statistica

- La statistica e' quella disciplina che affronta la nostra difficoltà e limitatezza nella  osservazione della realtà.
- E' una implementazione/risposta pratica al lavoro degli epistemologi in filosofia
  - I corvi sono neri? (Hume) -> si, ma non in maniera assoluta ( alcuni sono affetti di albinismo)

  Ma anche:

  1. Misurare la concentrazione di alcool in una bottiglia di vino. (come conciliare misure leggermente diverse ottenute da misurazioni ripetute)
  2. Il farmaco X migliora la salute di un paziente affetto dalla patologia Y (non tutti i pazienti rispondono allo stesso modo, le misure della variabile salute sono incerte, altri parametri possono influenzare la variabile salute)

# La statistica

- I tipi di incertezze delle osservazioni:
  - Incertezza delle misurazioni
  - Presenza di fattori non controllati che perturbano le osservazioni

# La statistica

- Statistica descrittiva
  - Approcci necessari per riassumere le caratteristiche di un insieme di dati in maniera sintetica e facilmente intellegibile per la mente umana.
- Statistica inferenziale
  - Approcci volti alla sostegno o alla falsificazione di una teoria. La La statistica descrittiva consente di definire gli approcci ottimali da usare nella statistica inferenziale. All'interno della statistica inferenziale si definiscono sia le modalità sperimentali che le procedure analitiche volte ad analizzare i risultati e raggiungere la conclusione.

# Obbiettivi del corso

- Sapere usare un ambiente di lavoro (R) di tipo statistico per gestire, osservare un largo insieme di dati, applicando le metodologie di statistica descrittiva
- Usare, anche in R, le metodologie di statistiche inferenziale di base in situazioni predefinite come uno strumento per raggiungere delle conclusioni in esperimenti biologici
- Riconoscere come inquadrare un esperimento biologico in una delle tipologie di situazioni predefinite studiate sulla base delle domande che uno vuole porre e sulla natura dei dati

# Cosa non aspettarsi dal corso

1. Una comprensione approfondita della teoria che sottende la definizione delle metodologie statistiche presentate
2. La capacita di sviluppare approcci statistici idonei per dati  non facilmente riportabili a situazioni predefinite

- La conoscenza di un ambiente di lavoro come R unita ad una curiosità ed impegno  personale può facilmente consentire di riempire le due precedenti mancanze fino ad un livello soddisfacente per la maggior parte delle problematiche biotecnologiche.

# Gestione del corso

- le lezioni frontali spiegheranno i metodi statistici dando anche i comandi R per implementarli
- Mi aspetto che gli studenti accedano al laboratorio informatica per esercitarsi nelle applicazioni dei metodi spiegati.
- Eventuali domande negli orari di ricevimento dovranno essere sempre corredate di un rapporto in formato elettronico dove lo studente mostra, anche in maniera molto sintetica, i suoi tentativi di implementare il concetto in R.
- Orari di ricevimento:
  - giovedì 15:00 alle 16:00
  - Via Amendola 122D, 4° piano presso CNR-ITB)
- Email: saverio.vicario@ba.itb.cnr.it
- Usare come oggetto della mail: Corso di Statistica Biotec

# Testi consigliati

- WayneW. Daniel, Biostatistica: concetti di base per l'analisi statistica delle scienze dell'area medico-sanitaria, EdiSESs.r.l. – Napoli.
- StantonA. Glantz, Statistica per discipline biomediche, 5°Edizione, McGrawHill.
- M. Pagano, K.Gauvreau, Biostatistica, II Edizione, IdelsonGnocchi.
- K. Seefeld, E. Linder. Statistics Using R with Biological Examples. Dispensa disponibile sul sito

# Come scrivere un Rapporto

- Aprire un editor di testo (es. Word) e R. Il rapporto va scritto mentre si eseguono le operazioni.
- Il rapporto si compone:
  - Introduzione che definisce lo scopo che si vuole raggiungere
  - Descrizione dei dati: numero e tipo di variabili, copiare da R la rappresentazione di parte dei  dati come esempio
  - Risultati:
    - Copiare e incollare da R i comandi e le risposte riportate, eventuali grafici
    - Cancellare dal rapporto gli errori risolti o le procedure scartate
  - Interpretazioni dei risultati ed eventuali commenti
  - Conclusioni: una risposta alle domande poste nell'introduzione

# Calendario

1. Introduzione ad R
2. Importare dati in R, rappresentazione grafiche di dati (istogrammi, grafici a barre, boxplot, scatterplot, matrice di grafici, …)
3. Concetto di campione statistico. Descrizione numerica di dati categorici (tabelle) e numerici (quantili e momenti)
4. Come trarre conclusione dai dati: il metodo scientifico e gli approcci statistici. Teoria delle probabilità, interpretazione frequentista e bayesiana.
5. Teoria del test da un punto di vista frequentista e bayesiano. Problemi delle variabili nascoste.
6. Test di adeguatezza: test di chi quadro, test di normalità
7. Distribuzioni di riferimento: densità di probabilità, distribuzioni cumulativa, funzione dei quantili per distribuzioni di Bernoulli, binomiale, Poisson, uniforme, Gaussiana.
8. Test con parametri categorici: Test T di student, ANOVA
9. Test di associazione fra parametri numerici : Pearson ; Test di paragone fra modelli: LRT
10. Influenza delle condizioni sperimentali sui test (problemi di test multipli e gerarchici, selezione dei dati)
11. Test non parametrici (signTest, Sperman rank test)
12. Test non parametrici (KS-test, permutazioni)

# R come strumento per le analisi statistiche

# R

- R e' un progetto, un linguaggio e un ambiente di lavoro
  - Progetto open source sotteso da una comunità di utenti esperti e non
  - Un linguaggio di programmazione semplice
  - Un ambiente di lavoro entro cui gestire i dati e usare il linguaggio  di programmazione per implementare analisi statistiche

# R come progetto

- il progetto (cran.r-project.org) essendo open source non garantisce il corretto funzionamento di tutte le funzioni o la correttezza di tutta la documentazione. La garanzia e' data esclusivamente dalla comunità degli utenti che segnala e propone correzioni nel forum del sito.
- L'essere open source garantisce però una vasta libreria di funzioni che consentono la facile implementazioni anche dei più recenti protocolli statistici.
- Il programma che include l'ambiente di lavoro e le librerie di funzioni e' scaricabile dal sito sia per Linux, MacOSx che per Windows XP e Vista. L'ambiente di lavoro e' leggermente diverse tra le diverse versioni anche se il linguaggio e' lo stesso

# R come linguaggio: dati atomici

- I dati in R hanno tre aspetti:
  - tipologia: numerico, carattere, o logico
  - Un valore (es. 3 ,"Pippo" oppure True)
  - Un nome
- Operatore "<-" assegna un valore ad un nome. Il modo o tipologia del valore e' dedotto dal tipo di valore

```
> x<-3
> x
[1] 3
> mode(x)
[1] "numeric"
```

```
> x<-"pippo"
> x
[1] "pippo"
> "pippo"=='pippo'
[1] TRUE
> mode(x)
[1] "character"
```

```
> x<-True
Errore: oggetto "True" non trovato
> x<-TRUE
> X<-T
> mode(X)
[1] "logical"
> mode(x)
[1] "logical"
```

# Valori speciali

- I valori dei dati logici possono essere T, F e NA
- NA vuol dire Non Applicabile e viene usato in R per descrivere i dati mancanti o operazione senza risposta ( anche per la presenza di dati mancanti)

```
>2>0/0
NA
> NA+3
[1] NA
```

- I valori numerici reali sono integrati con Inf e NaN. NaN indica un operazione numerica indeterminata

```
 > 2/0
[1] Inf
> 0/0
[1] NaN
```

# R come linguaggio: strutture dei dati

I singoli dati sono organizzati in oggetti ordinati:

- Vettori o "vector" (serie ordinate unidimensionali)
- Matrici (serie ordinate bi-dimensionali, "matrix", e n-dimensionali, "array" )

Vettori e matrici sono composti di elementi tutti con la stessa tipologia

Le funzioni as.vector(X), is.vector(X), as.matrix(X), …

```
> s<-c(1,2,3)
> s
[1] 1 2 3
> s<-c(1,2,3)
> s
[1] 1 2 3
> is.vector(s)
[1] TRUE
> S<-1:3
> S
[1] 1 2 3
> is.vector(S)
[1] TRUE
```

# R come linguaggio: strutture dei dati

- Liste "list" (serie ordinata unidimensionale di oggetti di diverse tipologie)
- "data.frame" : lista di vettori di stessa lunghezza ma potenzialmente diversa tipologia
- Fattori, "factor", sono dei vettori di caratteristiche qualitative e sono archiviate come dei vettori di numeri interi con etichette testuali per ogni valore

```
> AspettoSemi<-factor(c(1,1,1,2,2,3,1))
> AspettoSemi
[1] 1 1 1 2 2 3 1
Levels: 1 2 3
> levels(AspettoSemi)<-c('liscio','rugoso','spinoso')
> AspettoSemi
[1] liscio  liscio  liscio  rugoso  rugoso  spinoso liscio
Levels: liscio rugoso spinoso
```

Useremo nel nostro corso solo vettori, vettori di fattori e data.frame

# Le operazioni

Operazioni su oggetti atomici

```
> x<-5
> y<-2
z<-x+y
> z
[1] 7
> x-y
[1] 3
> x*y
   +2
[1] 12
```

Operazioni su oggetti strutture di dati

```
> v1<-c(1,2,3)
> v2<-c(4,5,6)
> z<-v1+v2
 > z
[1] 5 7 9
```

```
> x1<-c(1,2,3)
> x2<-c(3,4)
> x3<-x1+x2
Warning message:
In x1 + x2 :
  longer object length is not a multiple of shorter object length
> x3
[1] 4 6 6
```

# Le funzioni

- Una serie di operazioni possono essere riassunte in un solo commando con una funzione

```
> MEDIA<-function(X,Y,Z)
 {
S<-X+Y+Z
 S/3
 }
```

Il corpo della funzione e' delimitato da parentesi graffe. L'ultimo rigo e' il risultato della funzione

La funzione e' chiamata con le parentesi tonde mettendo l'argomento/i della funzione

```
> MEDIA(4,5,6)
[1] 5
```

Si può generalizzare la funzione

```
> MEDIA<-function(X)
{
 S<-sum(X)
 S/length(X)
 }
> MEDIA(c(4,5,6))
[1] 5
```

# Le funzioni

- E' possibile costruire in R funzioni anche piuttosto complesse simili a quelle di veri e propri linguaggi di programmazione. Non e' argomento di questo corso ma vi invito a documentarvi in caso pensiate di dover gestire grandi quantità di dati.
- Essenziali per fare un salto di qualità e' l'uso di operatori che regolano le iterazioni delle operazioni come :

  For, while,

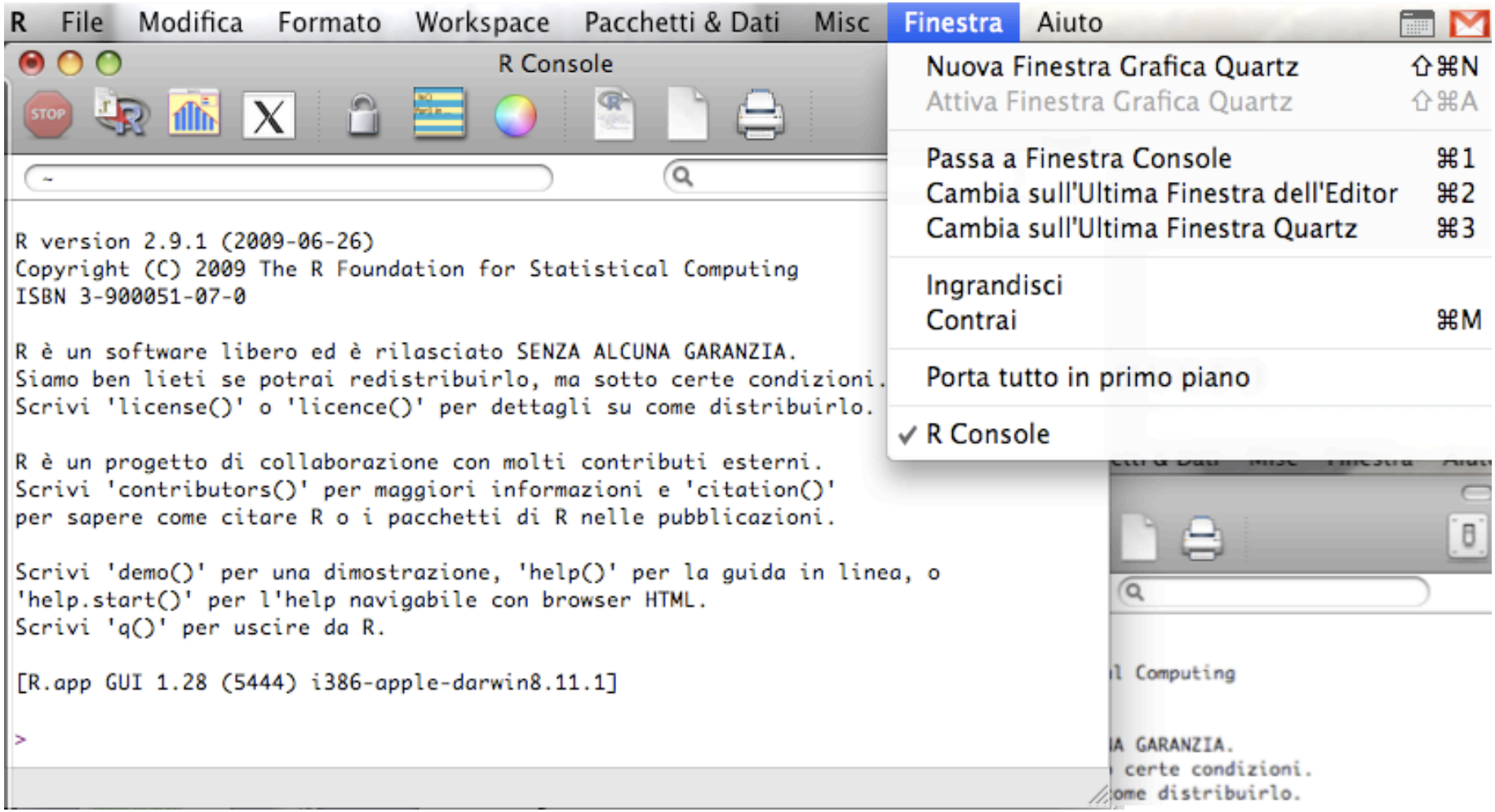
Oppure operatori di condizioni:

  If, else, switch, break, next

Per accedere ad una prima scheda di aiuto digitare:

```
>??"control-flow"
```

# R come Ambiente di Lavoro




# Anatomia di una scheda di documentazione

**Titolo**

Tra parentesi quadre il nome del pacchetto

**Descrizione**

**Uso**

Sintassi e valori predefiniti degli argomenti

**Argomenti**

Descrizione dettagliata degli argomenti della funzione, sia come tipo di informazione che come tipo di oggetto da usare

**Risultati**

Formato dei risultati e tipo di informazioni che porta

Dettagli : eventuali dettagli tecnici

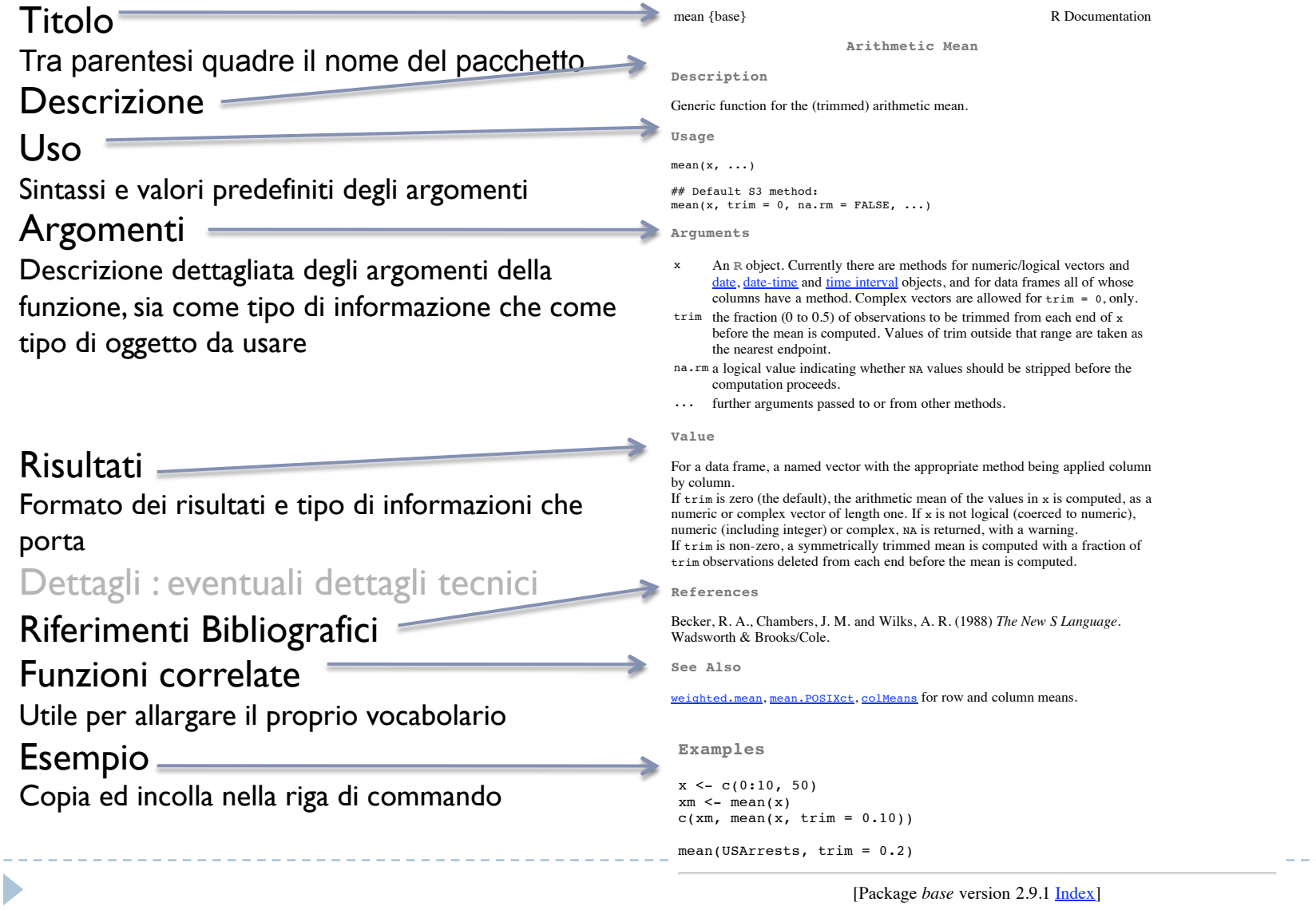
**Riferimenti Bibliografici**

**Funzioni correlate**

Utile per allargare il proprio vocabolario

**Esempio**

Copia ed incolla nella riga di commando



mean {base} R Documentation

**Arithmetic Mean**

**Description**

Generic function for the (trimmed) arithmetic mean.

**Usage**

```
mean(x, ...)

## Default S3 method:
mean(x, trim = 0, na.rm = FALSE, ...)
```

**Arguments**

| | |
|---|---|
| x | An R object. Currently there are methods for numeric/logical vectors and date, date-time and time interval objects, and for data frames all of whose columns have a method. Complex vectors are allowed for trim = 0, only. |
| trim | the fraction (0 to 0.5) of observations to be trimmed from each end of x before the mean is computed. Values of trim outside that range are taken as the nearest endpoint. |
| na.rm | a logical value indicating whether NA values should be stripped before the computation proceeds. |
| ... | further arguments passed to or from other methods. |

**Value**

For a data frame, a named vector with the appropriate method being applied column by column.
If trim is zero (the default), the arithmetic mean of the values in x is computed, as a numeric or complex vector of length one. If x is not logical (coerced to numeric), numeric (including integer) or complex, NA is returned, with a warning.
If trim is non-zero, a symmetrically trimmed mean is computed with a fraction of trim observations deleted from each end before the mean is computed.

**References**

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Wadsworth & Brooks/Cole.

**See Also**

weighted.mean, mean.POSIXct, colMeans for row and column means.

**Examples**

```
x <- c(0:10, 50)
xm <- mean(x)
c(xm, mean(x, trim = 0.10))

mean(USArrests, trim = 0.2)
```

[Package *base* version 2.9.1 Index]

# Sommario della lezione

- Statistica descrittiva:
  - Aiuta a esplorare i dati per generare ipotesi
- Teoria delle probabilità:
  - definisce una misura con cui rigettare o corroborare un ipotesi

# Definizioni degli elementi che compongono un'osservazione

- Unità statistica

- Caratteristica o attributo dell'unita statistica
- Modalità o valori dell'attributo
- Variabile aleatoria descrive la variazione dei valori dell'attributo nelle varie unita statistiche

- Il campione come insieme di unità statistiche
- La popolazione come insieme da cui e' tratto il campione

# Definizioni degli elementi che compongono un'osservazione

- Una unità statistica è un singolo soggetto preso in considerazione nell'analisi.

Es.: se studiamo l'altezza media degli studenti di questa aula, ogni studente è una unità statistica.

# Definizioni degli elementi che compongono un'osservazione

- Una caratteristica, detta anche feature, è un attributo o aspetto di una unità statistica.

Es.: se studiamo l'altezza media degli studenti di questa aula, l'altezza è una caratteristica

# Definizioni degli elementi che compongono un'osservazione

- Variabile aleatoria o casuale

In generale una caratteristica (feature) è rappresentata da una variabile aleatoria che può assumere valori (modalità) diversi su unità statistiche differenti.

Es .pressione diastolica del sangue, la frequenza cardiaca, la statura di maschi adulti, il livello di espressione di un gene in un tessuto,… .

Variabile descritta con una lettera maiuscola tipicamente X

Mentre singolo valore la corrispondente minuscola x con pedice i che identifica il valore esatto $x_i$

$X=\{x_1,x_2, \ldots, x_i, \ldots, x_n\}$ sono gli n valori della variabile X

# Definizioni degli elementi che compongono un'osservazione

A seconda delle modalità o valori possibili dell'attributo le variabili possono essere definite come

- Qualitative:
  - Ordinate (es. seme grande piccolo, medio, o ricercatore, prof. Associato, prof. Ordinario)
  - Non Ordinate (es. gruppo sanguigno)
- Quantitative:
  - Continue (es. lunghezza in cm di un seme, peso)
  - Discrete (es. numero di semi in una spiga)

# Definizioni degli elementi che compongono un'osservazione

- Se esiste una sola variabile per unita statistica si parla di statistica univariata
  - Le domande sono:
    - Descrizione della distribuzione
    - Paragone della variabile tra due campioni
    - Paragone tra un campione e una popolazione teorica
- Diverse variabili per campione
  - Le domande sono:
    - Descrizione della distribuzione multidimensionale
    - Relazione tra variabili(es. predizione di un carattere sulla base di uno o più altri caratteri