

# METODI STATISTICI PER LA RICERCA SOCIALE


## INTRODUZIONE A

Alessandra Mattei

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)  
Università degli Studi di Firenze  
[mattei@disia.unifi.it](mailto:mattei@disia.unifi.it)

LM-88 SOCIOLOGIA E RICERCA SOCIALE


# Che cosa è

-  è un software statistico, ossia un linguaggio / ambiente (insieme di macro, librerie, oggetti) che possono essere utilizzati per la gestione e l'analisi dei dati e la produzione e la visualizzazione di grafici
- È un *GNU-Software*, ossia è disponibile gratuitamente sotto i vincoli della GPL (General Public Licence)
- È estendibile: Esiste un'ampia disponibilità di moduli e librerie distribuiti con licenza GPL che possono essere scaricati dal sito CRAN (Comprehensive R Archive Network)
- Si noti che poichè si tratta di ambiente, per un determinato problema possono esistere diverse soluzioni
- È disponibile per diversi sistemi operativi: GNU/Linux, Mac, Windows

# Che cosa rende speciale ?

- È gratuito
- È facile da installare e non è un software pesante
- È open-source (non protetto da copyright e liberamente modificabile dagli utenti)
- Esistono più di 4000 pacchetti in cui sono implementate funzioni per analisi statistiche e rappresentazione grafiche di vario livello
- Permette di produrre e visualizzare (direttamente sul pc oppure su carta) grafici di ottime qualità

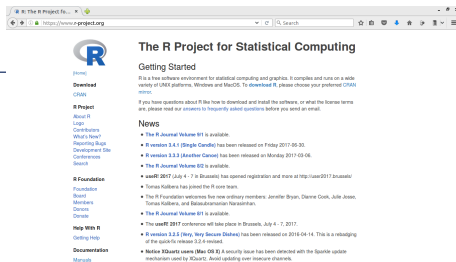
# ha una curva di apprendimento

- Gli utenti di ambienti statistici tendono a lanciare un comando per ogni tipo di analisi
  - ✓ Viene lanciato il comando e controllati i risultati estraendo l'informazione di interesse
-  ha alla base una filosofia diversa: la procedura è interattiva
  - ✓ Un'analisi statistica viene in genere fatta attraverso una serie di passi, con risultati intermedi che sono immagazzinati in oggetti che possono essere richiamati da altre funzioni
  - ✓ Passi: (1) lanciare un comando; (2) Salvare i risultati (3) Utilizzare i risultati salvati come argomento di un altro comando (4) Salvare i risultati del nuovo comando (5) Utilizzare i risultati salvati come argomento di un altro comando etc.
  - ✓ Il ciclo termina quando si conclude l'analisi

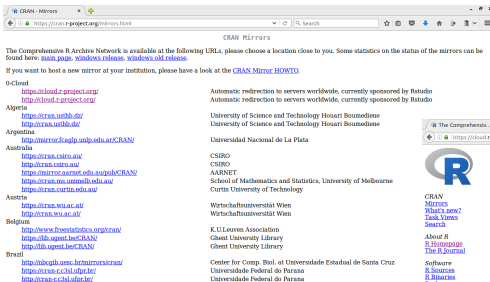
# Ottenere e installare

- <http://cran.r-project.org/>  
(o scrivere su google la lettera “R”)

- ✓ Download R – Scegliere un “mirror” (ad esempio 0-Cloud) – Click su “Download R for Linux” oppure “Download R for (Mac) OS X” oppure “Download R for Windows”
- ✓ Seguire le istruzioni per l’installazione



# Ottenere e installare



The Comprehensive R Archive Network is available at the following URL, please choose a location close to you. Some statistics on the status of the mirrors can be found here: [main page](#), [windows releases](#), [windows old releases](#).

If you want to host a new mirror at your institution, please have a look at the [CRAN Mirror HOWTO](#).

**O-Cloud**  
<https://cloud.r-project.org/>  
<http://cloud.r-project.org/>  
Automatic redirection to servers worldwide, currently sponsored by RStudio

**Algeria**  
<https://cran.usthb.dz/>  
<http://cran.usthb.dz/>  
University of Science and Technology Houari Boumediene  
University of Science and Technology Houari Boumediene

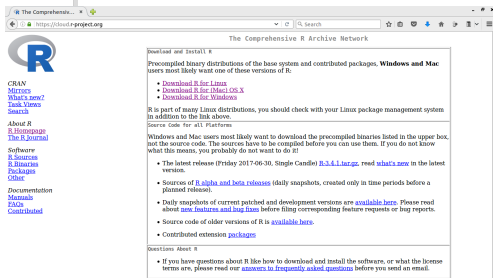
**Argentina**  
<http://mirror.fcaglp.unlp.edu.ar/CRAN/>  
Universidad Nacional de La Plata

**Australia**  
<https://cran.csiro.au/>  
<http://cran.csiro.au/>  
<https://mirror.aarnet.edu.au/pub/CRAN/>  
<https://cran.ms.unimelb.edu.au/>  
<https://cran.curtin.edu.au/>  
CSIRO  
CSIRO  
AARNET  
School of Mathematics and Statistics, University of Melbourne  
Curtin University of Technology

**Austria**  
<https://cran.wu.ac.at/>  
<http://cran.wu.ac.at/>  
Wirtschaftsuniversität Wien  
Wirtschaftsuniversität Wien

**Belgium**  
<http://www.fvestatgscs.org/cran/>  
<https://lib.ugent.be/CRAN/>  
<http://lib.ugent.be/CRAN/>  
K.U.Leuven Association  
Ghent University Library  
Ghent University Library

**Brazil**  
<http://bibogit.usc.br/mirrors/cran/>  
<https://cran.rcsl.ufpr.br/>  
<http://cran.rcsl.ufpr.br/>  
Center for Comp. Biol. at Universidade Estadual de Santa Cruz  
Universidade Federal do Paraná  
Universidade Federal do Paraná



The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for MacOS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

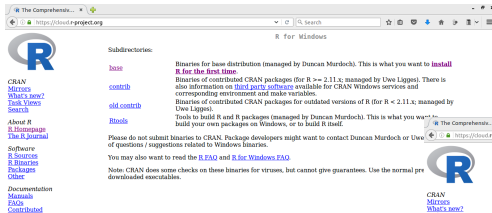
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (Friday 2017-06-30, Single Candle) [R 3.4.1 tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha](#) and [beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features](#) and [bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

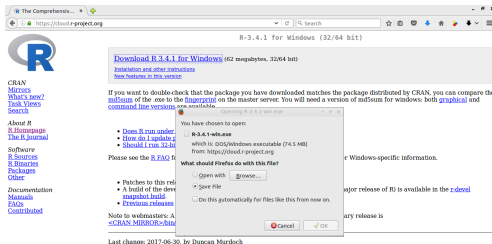
Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

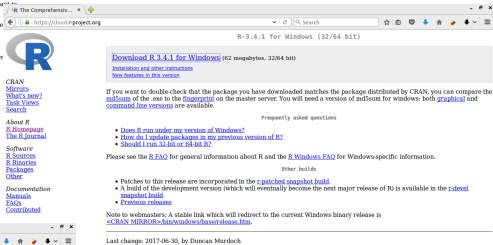
# Ottenere R per Windows



The screenshot shows the 'R for Windows' page on the CRAN website. The page has a sidebar with links: CRAN, Mirrors, What's new?, Task Views, Search, About R, R Homepage, The R Journal, Software, R Sources, R Binaries, Packages, Other, Documentation, Manuals, FAQs, and Contributed. The main content area is titled 'Subdirectories:' and lists: 'base' (Binaries for base distribution), 'contrib' (Binaries of contributed CRAN packages), 'old.contrib' (Binaries of contributed CRAN packages for outdated versions), and 'Rtools' (Tools to build R and R packages). Below this, it states that the page does not submit binaries to CRAN and provides links to the 'R FAQ' and 'R for Windows FAQ'. A note mentions that CRAN does some checks but cannot give guarantees.

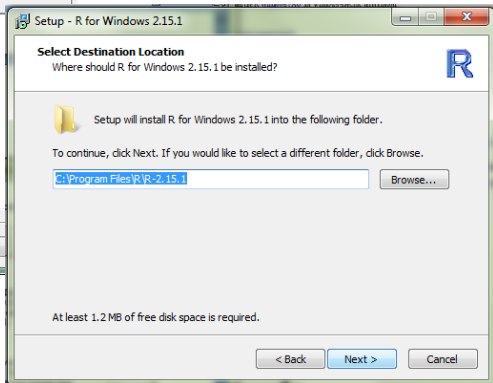
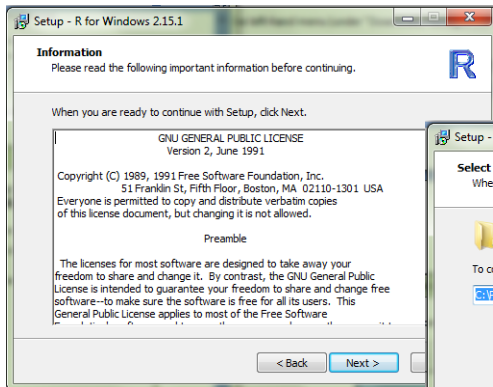


The screenshot shows the 'Download R 3.4.1 for Windows' page. It includes a download link for the 62 megabyte, 32/64 bit installer. Below the link, it provides instructions on how to verify the package against the CRAN master server using the 'md5sum' command. A list of frequently asked questions is provided, including 'Does R run under my version of Windows?', 'How do I update packages in my previous version of R?', and 'Should I run 32-bit or 64-bit R?'. A note to webmasters is also present. At the bottom, it states the last change was on 2017-06-30 by Duncan Murdoch.



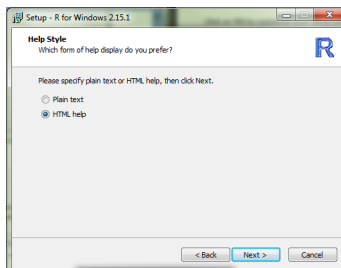
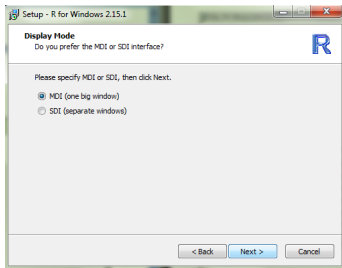
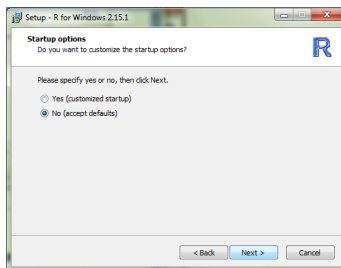
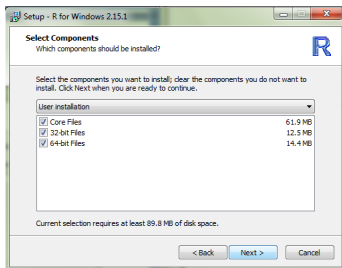
This screenshot shows the same 'Download R 3.4.1 for Windows' page, but with a file dialog box open over it. The dialog box is titled 'Opening R-3.4.1-win.exe' and shows the file 'R-3.4.1-win.exe' (74.5 MB) selected. The file is located in 'D:\OS\Windows executable'. The dialog box has options to 'Open with...', 'Save file', or 'Do this automatically for files like this from now on'. The 'Open with...' button is highlighted.

# Installare R in Windows

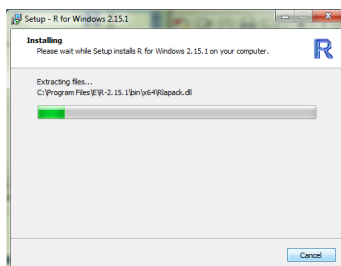
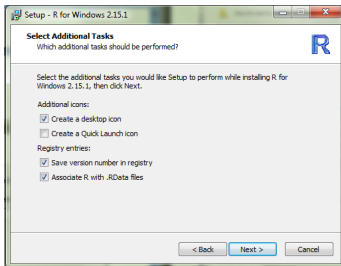
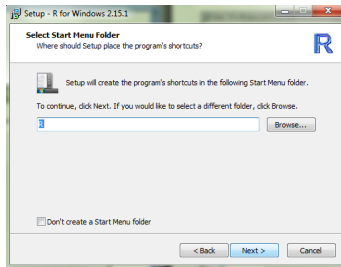
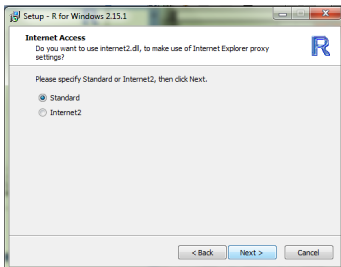






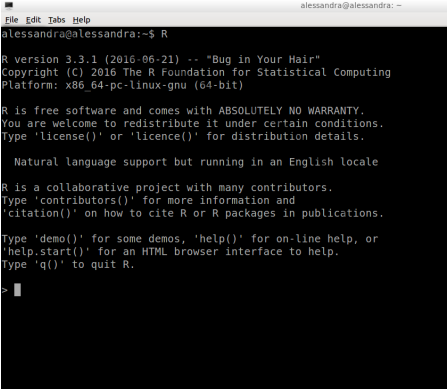
# Installare in Windows





# Installare R in Windows



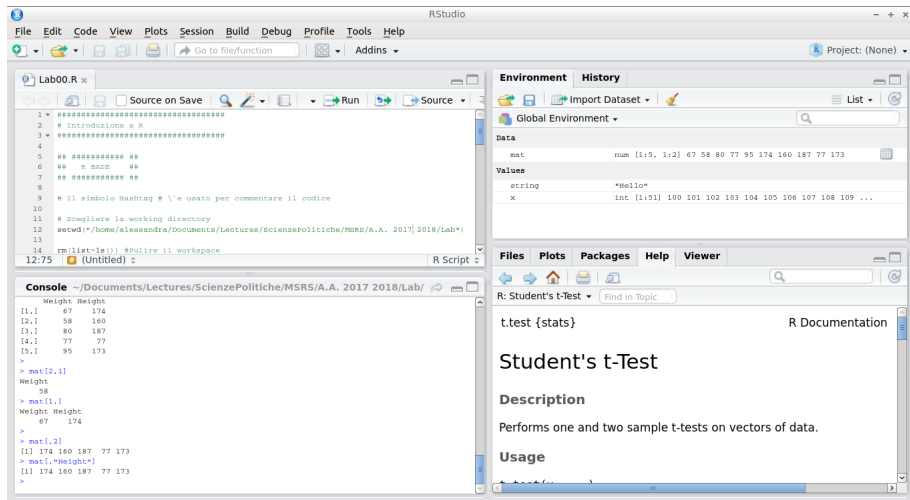
-  si presenta mediante una interfaccia grafica (RGui - R Graphical User Interface) a linea di comando
  - ✓ Essenziale nelle funzionalità utente attivabili tramite la barra dei menù e le icone pulsante
- In  tutte le istruzioni sono eseguibili dalla linea di comando dell'ambiente, tipicamente identificata mediante il carattere prompt ">"
  - ✓ La linea di comando viene eseguita semplicemente premendo il tasto "Invio" della tastiera






```
alessandra@alessandra: ~  
File Edit Tabs Help  
alessandra@alessandra:~$ R  
  
R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"  
Copyright (C) 2016 The R Foundation for Statistical Computing  
Platform: x86_64-pc-linux-gnu (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
Natural language support but running in an English locale  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.  
  
> |
```

- La maggior parte delle analisi richiedono di scrivere uno **script** (file di testo con estensione .R), che viene passato nella R-console
- R-Studio rende questo processo più semplice:  
<http://www.rstudio.com/>
- R-Studio è un ambiente di sviluppo integrato (IDE: Integrated development environment) libero per  ossia un software che aiuta nello sviluppo del codice sorgente per le analisi in 

# Layout di R-Studio




# Usare in R-Studio

- Se  e R-Studio sono installati, è sufficiente lavorare in R-Studio
- Aprire R-Studio (non è necessario aprire )
- Aprire uno script R – file con estensione .R (File -> Open File -> ...) oppure aprire un nuovo script R (File -> New File -> R Script)
- Iniziare le analisi in !

# è una calcolatrice estremamente potente

```
> rm(list=ls()) #Pulire il workspace
>
> #Usare R come una calcolatrice
> 1+1
[1] 2
>
> 1+2*7
[1] 15
>
> (1+2)*7
[1] 21
```

- Il carattere `#` è utilizzato per indicare l'inizio di un commento; tutti i caratteri che seguono vengono da  ignorati
- Principali funzioni matematiche: `+`, `-`, `*`, `/`, `^`

- Un oggetto è un nome simbolico usato per salvare delle informazioni
- Oggetti primitivi: Variabili numeriche (numeri interi, reali, complessi), Variabili non numeriche (Stringhe), Variabili logiche (TRUE o FALSE)
- Un oggetto è creato attraverso un'assegnazione che si ottiene usando indistintamente i caratteri: "<-" oppure "="

```
> x <- 12
```

```
> x
```

```
[1] 12
```

```
>
```

```
> string <- "Hello"
```

```
> string
```

```
[1] "Hello"
```



# Gli oggetti di

- Usando gli oggetti primitivi si possono costruire oggetti più complessi: vettori, matrici, array, Liste
- Un vettore è una collezione ordinata di dati dello stesso tipo
- Un vettore numerico è composto da numeri e può essere creato in vari modi:

✓ Usando la funzione `c()` (“Concatenate”)

```
> x <- c(3.5,7,9,11)
```

```
> x
```

```
[1] 3.5 7 9 11
```

✓ Usando la funzione `rep()` (“replicate”)

```
> x <- rep(10,3)
```

```
> x
```

```
[1] 10 10 10
```

✓ Usando l'operatore `:` per una serie di numeri interi

```
> x <- 4:15
```

```
> x
```

```
[1] 4 5 6 7 8 9 10 11 12 13 14 15
```

- Vettore non-numerico (Le stringhe devono essere poste tra virgolette)

```
> x=c("A","B","B","D","E")  
> x  
[1] "A" "B" "B" "D" "E"
```

- Vettore logico

```
> x<-1:4  
> x  
[1] 1 2 3 4  
> x<3  
[1] TRUE TRUE FALSE FALSE
```

- ✓ Operatori per il confronto

Uguale: ==

Non uguale: !=

Minore / Maggiore: < / >

Minore / Maggiore o uguale: <= / >=

- ✓ Operatori logici: E: & Oppure: | Negazione: !

- Operazioni elemento per elemento

```
> x<- c(100, 105, 110, 115, 120, 125, 130, 135)
> x
[1] 100 105 110 115 120 125 130 135
>
> #Aggiungere 1 a tutti gli elementi di x
> x+1
[1] 101 106 111 116 121 126 131 136
>
> #Moltiplicare tutti gli elementi di x per 2
> x*2
[1] 200 210 220 230 240 250 260 270
```

# Manipolazione di vettori in

- Accedere agli elementi di un vettore

```
> #Estrarre l'elemento in posizione 2
> x[2]
[1] 105
>
> #Estrarre gli elementi in posizione 3 e 4
> x[3:4]
[1] 110 115
>
> #Elementi di x maggiori di 125
> x[x>125]
[1] 130 135
>
> #Elementi di x maggiori di 115 e minori di 137
> x[x>120 & x<137]
[1] 125 130 135
```

# Le matrici in

- Una matrice è una tabella ordinata che include dati dello stesso tipo

```
> mat <- matrix(c(67,58,80,77,95,174,160,187,173, 185), nrow=6,ncol=2)
> colnames(mat)<-c("Weight","Height")
> mat
Weight Height
[1,]      67    174
[2,]      58    160
[3,]      80    187
[4,]      77    173
[5,]      95    185
>
> mat[2,1]
58
>
> mat[1,]
Weight Height
67      174
>
> mat[,2]
[1] 174 160 187 173 185
>
> mat[, "Height"]
[1] 174 160 187 173 185
```

- Una lista è una collezione ordinata di dati di qualsiasi tipo

```
> unit <- list(name="Joe", age=28, age_children=c(1,4), married=T)
> unit
$name
[1] "Joe"

$age
[1] 28

$age_children
[1] 1 4

$married
[1] TRUE
```

- Accedere agli elementi di una lista

```
> unit$name
```

```
[1] "Joe"
```

```
>
```

```
> unit[3]
```

```
$age_children
```

```
[1] 1 4
```

```
>
```

```
> unit[[3]]
```

```
[1] 1 4
```

```
>
```

```
> unit[[3]][2]
```

```
[1] 4
```

- Una lista è una collezione ordinata di dati di qualsiasi tipo

```
> unit <- list(name="Joe", age=28, age_children=c(1,4), married=T)
> unit
$name
[1] "Joe"

$age
[1] 28

$age_children
[1] 1 4

$married
[1] TRUE
```



- Un data frame è una matrice di dati
- È una tabella rettangolare con un certo numero di righe e un certo numero di colonne
  - ✓ I dati all'interno a ogni colonna sono dello stesso tipo (numerici, caratteri, logici) ma colonne diverse possono includere dati di tipo diverso
- Un data frame combina caratteristiche delle matrici e delle liste
- Esempio

Scuola	Voto di maturità	Borsa di Studio
Scienze Politiche	100	Si
Economia	95	Si
Scienze Politiche	80	No

# Data frame in

```
> scuola <- c("Scienze Politiche", "Economia", "Scienze Politiche")
> voto <- c(100, 95, 80)
> borsa <- c(TRUE, TRUE, FALSE)
> dati <- data.frame(scuola, voto, borsa)
> dati
  scuola voto borsa
1 Scienze Politiche 100  TRUE
2      Economia    95  TRUE
3 Scienze Politiche  80 FALSE
>
> dati$scuola
[1] Scienze Politiche Economia      Scienze Politiche
Levels: Economia Scienze Politiche
>
> dati[,2]
[1] 100  95  80
>
> dati[1,]
  scuola voto borsa
1 Scienze Politiche 100  TRUE
```

- File di dati: .Rdata

```
> load("Dati/BirthWeights.Rdata")
> ls()
[1] "birth.weights"
> str(birth.weights)
'data.frame': 24 obs. of  1 variable:
 $ bweight: num  2008 2023 2043 2056 2006 ...
> head(birth.weights)
bweight
1      2008
2      2023
3      2043
4      2056
5      2006
6      1989
```

# Importare dati in

- File di dati: .txt

```
> asma<-read.table("Dati/Asma.txt", header=TRUE)
> str(asma)
'data.frame': 215 obs. of  1 variable:
 $ asma: int  0 0 0 0 0 1 0 0 0 0 ...
> head(asma)
asma
1    0
2    0
3    0
4    0
5    0
6    1
```

- File di dati: .csv

```
> asma<-read.csv("Dati/Asma.csv", header=TRUE, row.names=1)
```

- Library 'foreign' per importare file di dati 'stranieri'

```
> library(foreign)
> asma<-read.dta("Dati/Asma.dta")
```


- Per avere informazioni su una particolare funzione, inclusa la sintassi, si può scrivere `?Nome_della_funzione` or `help(Nome_della_funzione)`

```
> ?t.test  
> help(t.test)
```

- `help.search()` per capire quale funzione usare in un particolare contesto

```
> help.search("standard deviation")
```

- Se non è noto il nome della funzione:

- ✓ Da google “R + t test”
- ✓  reference card

# Inferenza per la media in

- Problema: Studio del peso di neonati da madri fumatrici
  - ✓ Neonati di peso inferiore a 2500 grammi sono considerati bambini nati sottopeso

```
> ####Lettura dei dati: Peso di neonati da madri fumatrici
> rm(list=ls())
> load("Dati/BirthWeights.Rdata")
>
> str(birth.weights)
'data.frame': 24 obs. of  1 variable:
 $ bweight: num  2008 2023 2043 2056 2006 ...
>
> n <- nrow(birth.weights)
> n
[1] 24
>
> attach(birth.weights)
```

# Inferenza sulla media con $\sigma^2$ nota in

```
> #Supponiamo sigma2=198025
> sigma2<- 198025
>
> ybar <- mean(bweight)
> ybar
[1] 2290.083
>
> se<- sqrt(sigma2)/sqrt(n)
> se
[1] 90.83524
>
> alpha<- 0.05 #1-alpha=0.95
> z<- qnorm(1-alpha/2)
> z
[1] 1.959964
>
> ybar - z*se
[1] 2112.05
> ybar + z*se
[1] 2468.117
```

```
> #H0: mu=2500 versus Ha:mu<2500
> zoss <- (ybar-2500)/se
> zoss
[1] -2.310961
>
> ##Regione critica:
> qnorm(alpha)
[1] -1.644854
> ##Regione critica: Z <= -1.644854
>
> #p-valore
> pnorm(zoss)
[1] 0.01041752
```

```
###Stima
> ybar
[1] 2290.083
> var(bweight)
[1] 239057.1
> sd(bweight)
[1] 488.9347
>
> #Stima dell'errore standard della media campionaria
> se.hat <- sd(bweight)/sqrt(n)
> se.hat
[1] 99.80337
```



# Intervalli di confidenza per la media in



```
> alpha <- 0.05
> tt<- qt(1-alpha/2, n-1)
> tt
[1] 2.068658
>
> ybar - tt*se.hat
[1] 2083.624
> ybar + tt*se.hat
[1] 2496.542
>
>#La funzione t.test
> t.test(bweight)
```

One Sample t-test

```
data:  bweight
t = 22.946, df = 23, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
2083.624 2496.542
sample estimates:
mean of x
2290.083
```

# Intervalli di confidenza per la media in



```
> t.test(bweight, conf.level=0.99)
```

One Sample t-test

```
data:  bweight
```

```
t = 22.946, df = 23, p-value < 2.2e-16
```

```
alternative hypothesis: true mean is not equal to 0
```

```
99 percent confidence interval:
```

```
2009.902 2570.265
```

```
sample estimates:
```

```
mean of x
```

```
2290.083
```

# Test per la media in

```
> #H0: mu=2500 versus Ha:mu<2500
>
> toss <- (ybar-2500)/se.hat
> toss
[1] -2.103302
>
> alpha <- 0.05
>
> ##Regione critica:
> qt(alpha, n-1)
[1] -1.713872
>##Regione critica: T <= -1.713872
>
> #p-valore
> pt(toss, n-1)
[1] 0.02329044
```

# Test per la media in

```
> #Usando la funzione t.test ...  
> t.test(bweight, alternative = "less", mu = 2500)
```

One Sample t-test

```
data:  bweight  
t = -2.1033, df = 23, p-value = 0.02329  
alternative hypothesis: true mean is less than 2500  
95 percent confidence interval:  
-Inf 2461.133  
sample estimates:  
mean of x  
2290.083  
>  
> detach(birth.weights)
```

- Problema: Proporzione di asmatici in una popolazione omogenea di pazienti

```
> #####Lettura dei dati: campione casuale di 215 pazienti estratti
> #####da una popolazione omogenea.
> Asma<-read.table("Dati/Asma.txt", header=TRUE)
> str(Asma)
'data.frame': 215 obs. of  1 variable:
 $ asma: int  0 0 0 0 0 1 0 0 0 0 ...
>
> n <- nrow(Asma)
> n
[1] 215
>
> attach(Asma)
```

## Proporzione campionaria e suo errore standard della proporzione campionaria in



```
> nsucc<- sum(asma)
> nsucc
[1] 39
>
> nsucc/n
[1] 0.1813953
>
> phat <- mean(asma)
> phat
[1] 0.1813953
>
> ##Varianza
> phat*(1-phat)
[1] 0.1484911
>
> ##Varianza della proporzione campionaria
> phat*(1-phat)/n
[1] 0.0006906562
>
> ##Errore standard della proporzione campionaria
> se.hat <- sqrt(phat*(1-phat)/n)
> se.hat
[1] 0.02628034
```

# Intervalli di confidenza per la proporzione in



```
> alpha <- 0.05
> z<- qnorm(1-alpha/2)
> z
[1] 1.959964
>
> phat - z*se.hat
[1] 0.1298868
> phat + z*se.hat
[1] 0.2329039
```

# Intervalli di confidenza per la proporzione in

```
> #Usando la funzione prop.test con correct = FALSE  
> #si ottiene un intervallo di confidenza molto simile  
>  
> prop.test(nsucc, n, correct = FALSE)
```

1-sample proportions test without continuity correction

```
data:  nsucc out of n, null probability 0.5  
X-squared = 87.298, df = 1, p-value < 2.2e-16  
alternative hypothesis: true p is not equal to 0.5  
95 percent confidence interval:  
0.1356282 0.2383479  
sample estimates:  
p  
0.1813953
```



# Test per la proporzione in

```
> ##H0: p=0.1 versus Ha: p!=0.1
> p0<-0.1
> se0 <- sqrt(p0*(1-p0)/n)
> se0
[1] 0.02045983
>
> zoss<- (phat -p0)/se0
> zoss
[1] 3.9783
>
> alpha <- 0.05
> qnorm(1-alpha/2)
[1] 1.959964
> ##Regione critica: Z<= -1.96 oppure >= 1.96
>
> ###p-valore
> 2*{1-pnorm(abs(zoss))}
[1] 6.940967e-05
```

# Test per la proporzione in

```
> ##Usando la funzione prop.test  
> p.test<-prop.test(nsucc, n, p=p0, correct = FALSE)  
> p.test
```

1-sample proportions test without continuity correction

```
data:  nsucc out of n, null probability p0  
X-squared = 15.827, df = 1, p-value = 6.941e-05  
alternative hypothesis: true p is not equal to 0.1  
95 percent confidence interval:  
0.1356282 0.2383479  
sample estimates:  
p  
0.1813953  
>  
> p.test$statistic  
X-squared  
15.82687  
> sqrt(p.test$statistic)  
X-squared  
3.9783
```

# Test per la proporzione in

```
> ##H0: p=0.1 versus Ha: p>0.10
>
> ##Regione critica:
> qnorm(1-alpha)
[1] 1.644854
> ##Regione critica: Z>1.644854
>
> ###p-valore
> {1-pnorm(zoss)}
[1] 3.470483e-05
>
> prop.test(nsucc, n, p=p0, alternative = "greater", correct = FALSE)
```

1-sample proportions test without continuity correction

```
data:  nsucc out of n, null probability p0
X-squared = 15.827, df = 1, p-value = 3.47e-05
alternative hypothesis: true p is greater than 0.1
95 percent confidence interval:
0.1422149 1.0000000
sample estimates:
p
0.1813953
```