Uso elementare di R in Statistica

G. Marchetti

Lezione 1

Introduzione

R è un ambiente statistico di pubblico dominio. Il software è libero e può essere scaricato dal sito https://cran.r-project.org/.

Una volta installato sul proprio sistema provate ad eseguire semplici calcoli:

```
1+3
```

```
[1] 4
```

```
1*3
```

```
[1] 3
```

```
1/3
```

```
[1] 0.3333
```

Definizione di un insieme di dati

Va definito un vettore nel modo seguente. Per esempio se i dati su una variabile X sono

$$(73, 80, 84, 78, 90, 87, 72, 70, 75)$$

si dà il comando

```
x = c(73, 80, 84, 78, 90, 87, 72, 70, 75)
x
```

```
[1] 73 80 84 78 90 87 72 70 75
```

che definisce il vettore `x`.

Dimensione campionaria e sommatoria

La dimensione campionaria n è la lunghezza del vettore `x`

```
length(x)
```

```
[1] 9
```

```
n = length(x)
n
```

```
[1] 9
```

Notate che abbiamo assegnato al simbolo `n` la dimensione campionaria.

La somma di tuddi i dati campionari $\sum_{i=1}^{n} x_i$ si ottiene con il comando `sum`

```
sum(x)
```

```
[1] 709
```

```
tot = sum(x)
tot
```

```
[1] 709
```

Indici statistici

La media del campione

La media aritmetica del campione è $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Si può calcolare direttamente col comando `mean` o usando la formula:

```
mean(x)
```

```
[1] 78.78
```

```
tot/n
```

```
[1] 78.78
```

```
xbar = mean(x)
xbar
```

```
[1] 78.78
```

Varianza del campione

La varianza è

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}(\sum_{i=1}^{n}(x_i^2 - n\bar{x}^2).$$

Si calcola direttamente o usando la formula

```
var(x)
```

```
[1] 49.19
```

```
sum((x - xbar)^2)/(n - 1)
```

```
[1] 49.19
```

```
(sum(x^2) - n*xbar^2)/(n-1)
```

```
[1] 49.19
```

```
s2 = var(x)
```

La deviazione standard è la radice della varianza

```
sqrt(s2)
```

```
[1] 7.014
```

```
sd(x)
```

```
[1] 7.014
```

Statistiche ordinate

Le statistiche ordinate sono i dati scritti dal più piccolo al più grande. Si calcolano con il comando `sort`.

```
x
```

```
[1] 73 80 84 78 90 87 72 70 75
```

```
xord = sort(x)
xord
```

```
[1] 70 72 73 75 78 80 84 87 90
```

La prima statistica ordinata è il minimo di `x` e l'ultima è il massimo di `x`:

```
xord[1]
```

```
[1] 70
```

```
min(x)
```

```
[1] 70
```

```
xord[n]
```

```
[1] 90
```

```
max(x)
```

```
[1] 90
```

Mediana e quantili

La mediana è la statistica ordinata che ha prima di sé e dopo di sé almeno metà delle osservazioni. Se `n` è dispari è `xord[(n+1)/2]`

```
median(x)
```

```
[1] 78
```

```
xord[(n+1)/2]
```

```
[1] 78
```

Il campo di variazione

Il campo di variazione è la differenza tra il massimo e il minimo.

```
max(x) - min(x)
```

```
[1] 20
```

Quantili

I quantili di ordine 0.25 e 0.75 si calcolano con il comando `quantile`.

```
quantile(x, 0.25)
```

```
25%
 73
```

```
quantile(x, 0.75)
```

```
75%
 84
```

I 5 numeri fondamentali di un insieme di dati sono

```
min(x)
```

```
[1] 70
```

```
quantile(x, 0.25)
```

```
25%
 73
```

```
median(x)
```

```
[1] 78
```

```
quantile(x, 0.75)
```

```
75%
 84
```

```
max(x)
```

```
[1] 90
```

che si ottengono anche col comando `fivenum`:

```
fivenum(x)
```

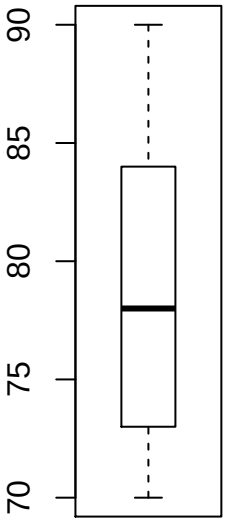
```
[1] 70 73 78 84 90
```

Rappresentazioni grafiche

Box-plot

Un box-plot (grafico a scatola) si ottiene con il comando `boxplot`.

```
boxplot(x)
```



```
fivenum(x)
```

```
[1] 70 73 78 84 90
```

Notate che è costruito usando i 5 numeri magici. Il box-plot è utile anche per identificare gli outlier, cioè i dati che sono molto lontani dal centro della distribuzione.

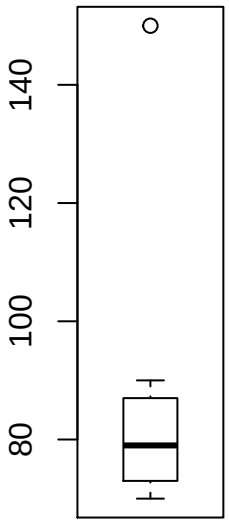
Per esempio se definiamo una nuova variabie `y` che ha gli stessi valori di `x` più un nuovo dato con un valore 150, che è molto sorprendente rispetto agli altri.

```
y = c(x, 150)
y
```

```
 [1]  73  80  84  78  90  87  72  70  75 150
```

Il box-plot evidenzia il dato sorprendente con un punto singolo:

```
boxplot(y)
```



Lezione 2

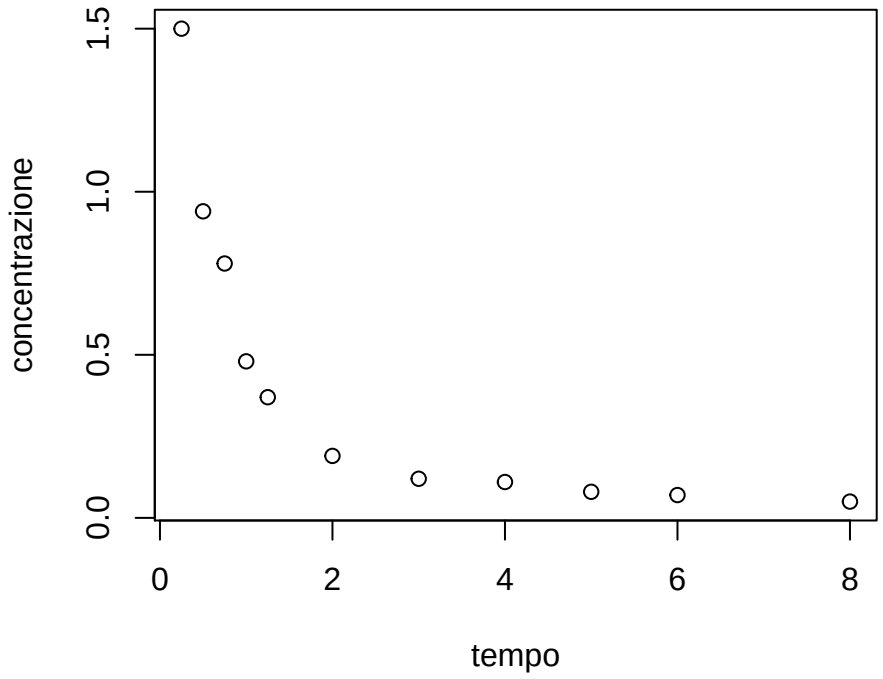
Un grafico

A un paziente viene fatta un endovena di Indometacina e quindi viene misurata al tempo t (in ore) la concentrazione $x(t)$ di Indometacina nel plasma (in mcg/ml). Ecco i dati:

```
t =c(.25, .5, .75, 1, 1.25, 2, 3, 4, 5, 6, 8)
x = c(1.50, .94, .78, .48, .37, .19, .12, .11, .08, .07, .05)
```

Uno scatter dei dati è il seguente

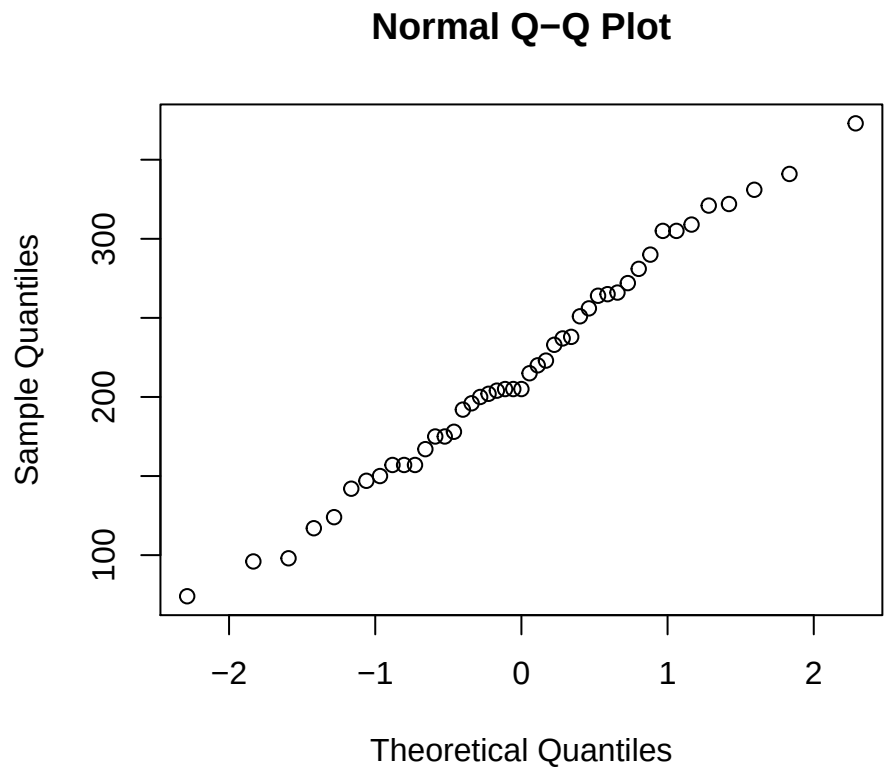
```
plot(t, x, xlab = 'tempo', ylab = 'concentrazione')
```



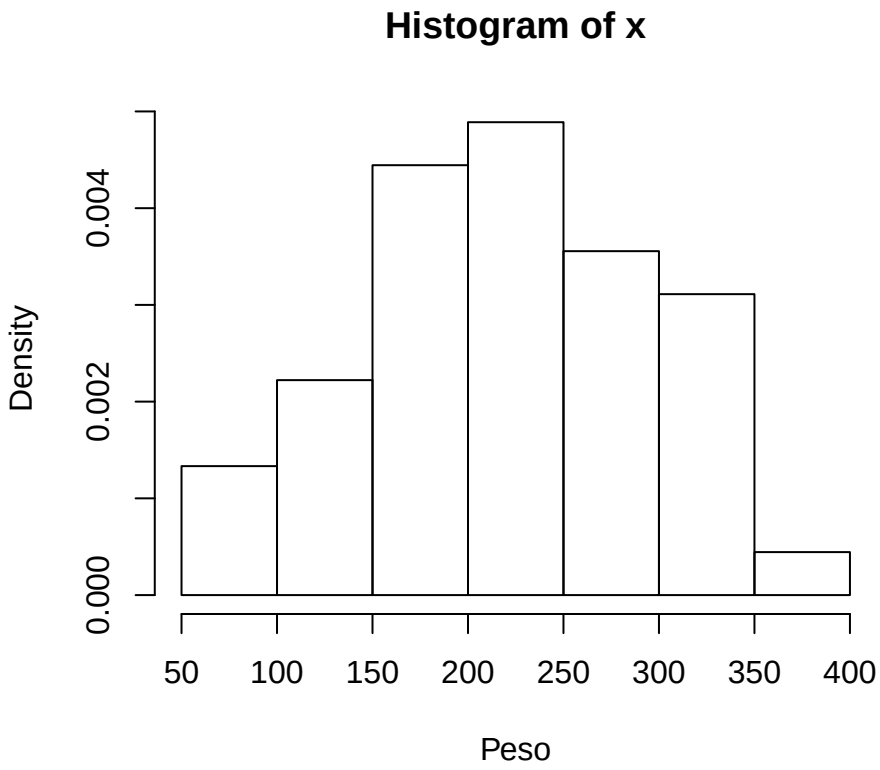
Dati sul peso di polli dopo 21 giorni

Lettura dei dati. Peso in grammi.

```
data("ChickWeight")
sel= ChickWeight$Time == 21
x  = ChickWeight$weight[sel]
qqnorm(x)
```




```r
hist(x, xlab = "Peso", freq=FALSE)
```



Stima della media della popolazione e del suo errore quadratico medio

- Stima $= \bar{x}$
- Errore quadratico medio $= \sigma^2/n$
- Errore standard stimato $SE = s/\sqrt{n}$ dove

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}(x_i - \bar{x})^2}$$

```
xbar = mean(x)
xbar
```

```
[1] 218.7
```

```
s = sd(x)
s
```

```
[1] 71.51
```

```
n = length(x)
n
```

```
[1] 45
```

```
SE = s/sqrt(n)
SE
```

```
[1] 10.66
```

Nota:

- s è la stima della variabilità del peso X nella popolazione.
- SE è la stima della variabilità del peso MEDIO $\bar{X}$ nell'universo dei campioni

Lunghezze dei fiumi degli Stati Uniti

Le lunghezze sono in miglia. Sommario degli indici

```
data(rivers)
summary(rivers)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    135     310     425     591     680    3710
```

Varianza e deviazione standard

```
var(rivers)
```

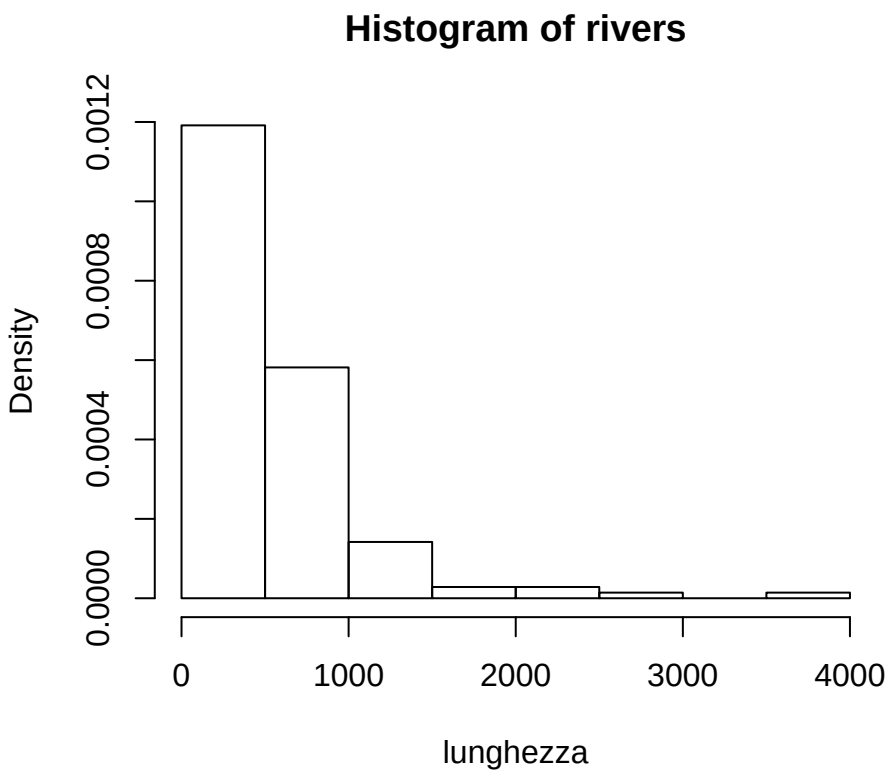
```
[1] 243908
```

```
sd(rivers)
```

```
[1] 493.9
```

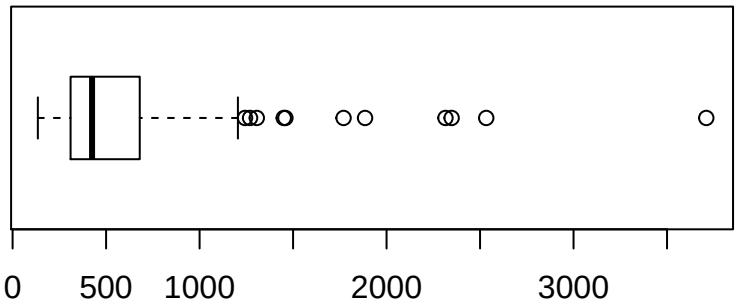
Istogramma

```
hist(rivers, xlab = 'lunghezza', freq = FALSE)
```



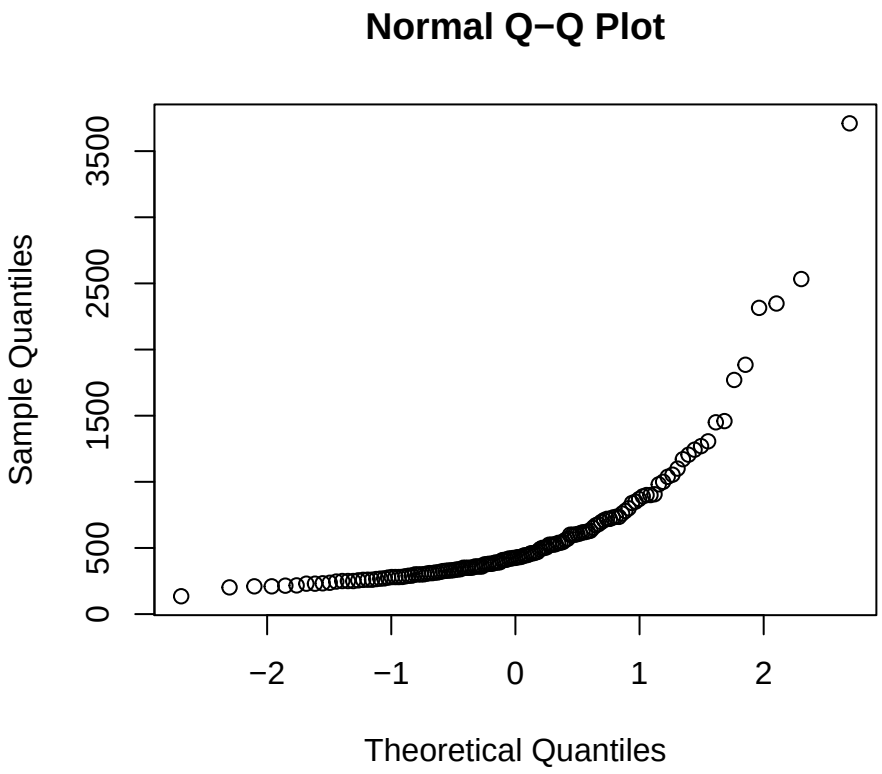
Non appare distribuita normalmente. Questo è evidente anche con un boxplot.

```
boxplot(rivers, horizontal = TRUE)
```



O con un grafico dei quantili

```
qqnorm(rivers)
```



Stima della lunghezza media dei fiumi da un campione casuale

Estrazione di un campione con ripetizione di dimensione 25.

```
x = sample(rivers, size = 25, replace = TRUE)
x
```

```
 [1]  800 1038  500  291  250  735  600  735  431  529  260  470  377  329  315  255  325
[18]  625  720  217  652  407  350  360  392
```

Stima della media

```
xbar = mean(x)
xbar
```

```
[1] 478.5
```

Errore standard

```
SE = sd(x)/5
SE
```

```
[1] 41.98
```

Lezione 3

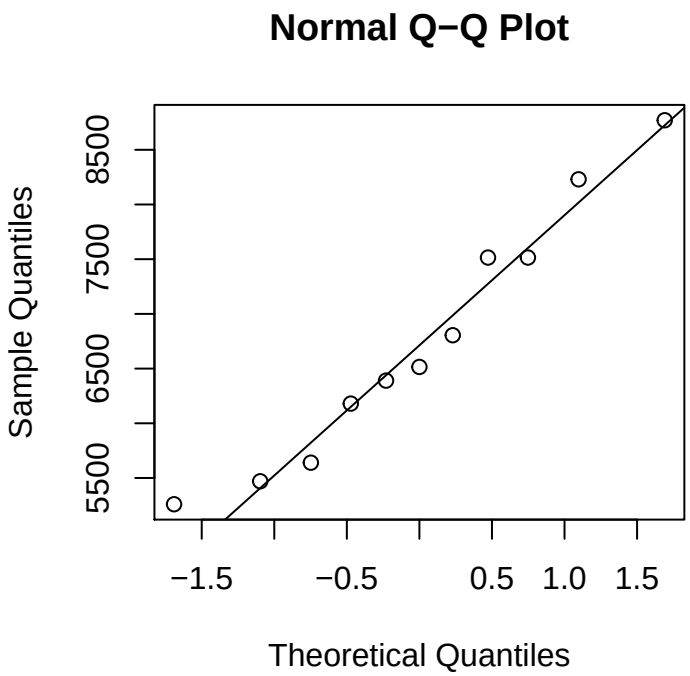
Intervalli di confidenza (e test)

Ecco un campione casuale di $n = 11$ donne scelte da una popolazione omogenea dove si è misurata per la quantità giornaliera di energia assunta (in kJoule).

```
energia = c(5260,5470,5640,6180,6390,6515,6805,7515,7515,8230,8770)
```

Vediamo dal grafico di probabilità normale se si può considerare una popolazione generatrice gaussiana.

```
qqnorm(energia)
qqline(energia)
```



Facendo l'assunzione di probabilità normale possiamo costruire un intervallo di confidenza per il quantitativo medio di energia assunta nella popolazione.

Supponiamo di sapere che la deviazione standard della popolazione sia $\sigma = 1000$ kj. Allora l'intervallo di confidenza al 95% per μ è

$$\bar{x} \pm 1.96\sigma/\sqrt{n}$$

cioè un intorno della media campionaria

```
mean(energia)
```

```
[1] 6754
```

di raggio (il cosiddetto margine di errore) $1.96 \cdot 1000/\sqrt{11} = 619.81$.

Quindi l'intervallo di confidenza per μ ha estremi

```
mean(energia) - 1.96 * 1000/sqrt(length(energia))
```

```
[1] 6163
```

```
mean(energia) + 1.96 * 1000/sqrt(length(energia))
```

```
[1] 7345
```

Caso in cui σ non è noto

In questo caso l'intervallo di confidenza al 95% è

$$\bar{x} \pm t \cdot s/\sqrt{n} \quad \text{ossia} \quad \bar{x} \pm t \cdot SE$$

dove

- s è la deviazione standard del campione (non quella supposta nota della popolazione)
- t è un coefficiente maggiore di 1.96 dipendente dalla dimensione del campione

Il coefficiente t si ottiene non dalla tavola della normale ma da quella della distribuzione t di Student. Si può sostituire la tavola della t con la funzione R seguente

```
LC = 0.95 # livello di confidenza
n = 11    # numerosità campionaria
t = qt(LC + (1-LC)/2, n-1)
t
```

```
[1] 2.228
```

Quindi l'intervallo di confidenza al 95% ha estremi

```
SE = sd(energia)/sqrt(11)
xbar = mean(energia)
A = xbar - t * SE
B = xbar + t * SE
c(A,B)
```

```
[1] 5986 7521
```

Come si vede è un po' più ampio del precedente.

Funzione per calcolare direttamente l'intervallo di confidenza

```
t.test(energia)
```

```

    One Sample t-test

data:  energia
t = 20, df = 10, p-value = 3e-09
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 5986 7521
sample estimates:
mean of x
     6754
```

NOTA Per il momento non considerate le prime 4 righe di output (che riguardano il test sulla media).

L'intervallo di confidenza viene dato subito dopo.

La funzione consente di rifare i calcoli velocemente se si vuole cambiare il livello di confidenza. Per esempio se si vuole un intervallo al 99% basta dare il comando:

```
t.test(energia, conf.level = 0.99)
```

```

    One Sample t-test

data:  energia
t = 20, df = 10, p-value = 3e-09
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
```

```
 5662 7845
sample estimates:
mean of x
     6754
```

Intervallo di confidenza per una proporzione.

Dato un campione casuale con ripetizione di n osservazioni da una popolazione dicotomica composta da una proporzione p di 1, l'intervallo di confidenza (approssimato) al 95% per p è

$$\hat{p} \pm 1.96 \cdot SE$$

dove

- $\hat{p} = \frac{r}{n} = \frac{\#successi}{n}$ è lo stimatore di p
- $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ è l'errore standard dello stimatore $\hat{p}$.

Esempio

In un campione casuale di 215 pazienti estratti da una popolazione omogenea si sono osservati 39 pazienti con l'asma. Trovare un intervallo di confidenza per la vera proporzione di asmatici nella popolazione.

```
stima = 39 / 215
stima
```

```
[1] 0.1814
```

```
SE = sqrt(stima * (1 - stima)/215)
SE
```

```
[1] 0.02628
```

```
A = stima - 1.96 * SE
B = stima + 1.96 * SE
c(A, B)
```

```
[1] 0.1299 0.2329
```

Anche in questo caso c'è una funzione apposita in R che evita di inserire tutte le istruzioni

```
prop.test(39, 215)
```

```
    1-sample proportions test with continuity correction

data:  39 out of 215, null probability 0.5
X-squared = 86, df = 1, p-value <2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.1336 0.2409
sample estimates:
     p
0.1814
```

che tuttavia usa un metodo diverso.

Intervallo di confidenza per la varianza

```
data = energia
LC = 0.95
n = length(data)
chilower = qchisq((1 - LC)/2, n-1)
chiupper = qchisq((1 - LC)/2, n-1, lower.tail = FALSE)
v = var(data)
c((n-1) * v/chiupper, (n-1) * v/chilower)
```

```
[1]  636837 4017420
```