Università degli Studi di Salerno
Dipartimento di Informatica

Corso di Metodi e tecniche per l'analisi dei dati

Dati ambientali nella città

Indicatori sulla raccolta differenziata per le regioni

Docente
Prof.ssa Amelia Giuseppina Nobile

Studente
Maria Domenica Guida
Matr. 0522500236

Anno Accademico 2014/15

Indice

1 Introduzione

In statistica una variabile casuale (detta anche aleatoria) è una variabile che può assumere valori differenti in dipendenza di un certo evento casuale.
Le variabili aleatorie vengono utilizzate per descrivere fenomeni dei quali non è possibile prevedere il risultato con certezza come, ad esempio, il lancio di un dado.
Questi tipi di variabili si dividono in due categorie:

1. **Variabili discrete:** i valori possibili sono rappresentati da un numero finito o da un infinito numerabile, ovvero possono essere descritti attraverso l'insieme dei numeri naturali.
   Tra le variabili aleatorie discrete abbiamo:
   - Distribuzione di Bernoulli.
   - Distribuzione binomiale.
   - Distribuzione geometrica e di Pascal.
   - Distribuzione ipergeometrica.
   - Distribuzione di Poisson.
2. **Variabili continue:** i valori possono essere rappresentati da tutti i numeri reali.
   Tra le variabili aleatorie continue abbiamo:
   - Distribuzione uniforme.
   - Distribuzione esponenziale.
   - Distribuzione normale.
   - Distribuzione chi quadrato.
   - Distribuzione di Student.

In questo lavoro verranno presentate la distribuzione normale, la chi-quadrato e la Student poichè giocano un ruolo rilevante nell'inferenza statistica.
L'inferenza statistica è un insieme di metodi con cui si cerca di trarre una conclusione su una popolazione(insieme che raccoglie tutte le osservazioni possibili relativamente ad un certo fenomeno) in base ai dati conosciuti relativi ad un campione (raccolta finita di elementi estratti da una popolazione con il fine di estrarre e di ottenere informazioni sulla popolazione). Uno dei modi in cui è possibile fare ciò è con gli intervalli di confidenza, che verranno descritti in dettaglio nei capitoli successivi.
Per ogni argomento, inoltre, verranno anche descritti gli strumenti messi a disposizione da R.

2 Distribuzione normale, chi-quadrato e Student

2.1 Distribuzione Normale

La distribuzione normale o gaussiana per variabili continue è considerata una delle più importanti in quanto essa costituisce un limite al quale tendono anche le altre funzioni di distribuzione sotto opportune ipotesi.
Data una variabile aleatoria X la formula della distribuzione gaussiana è la seguente:

$$f_X = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad x \in \mathbb{R}, \sigma > 0$$

Dove μ rappresenta la media, σ la deviazione standard e σ^2 la varianza, si dice che X ha distribuzione normale di parametri μ e σ.
Nel caso monodimensionale a livello geometrico la gaussiana è rappresentata da una funzione a campana.
In R sono disponibili differenti funzioni per lavorare con le gaussiane che descriveremo di seguito.

Densità normale

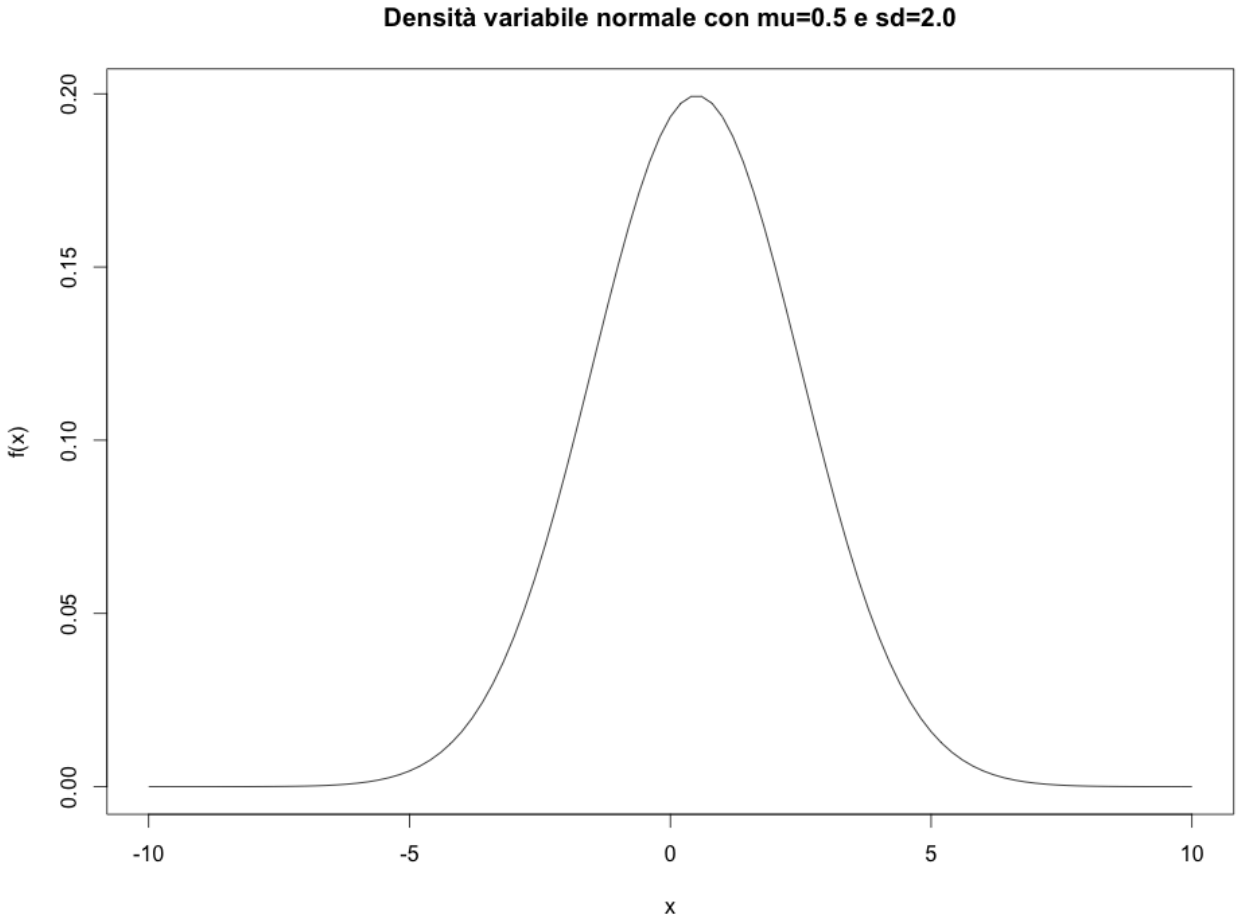
```
dnorm (x, mean = mu, sd = sigma, log = FALSE)
```

Dove:

- x è il valore assunto (o i valori assunti) dalla variabile aleatoria normale.
- mean e sd rappresentano il valore medio e la deviazione standard della densità normale.
- log se tale parametro è TRUE le probabilità sono espresse in forma logaritmica come $log(p)$.

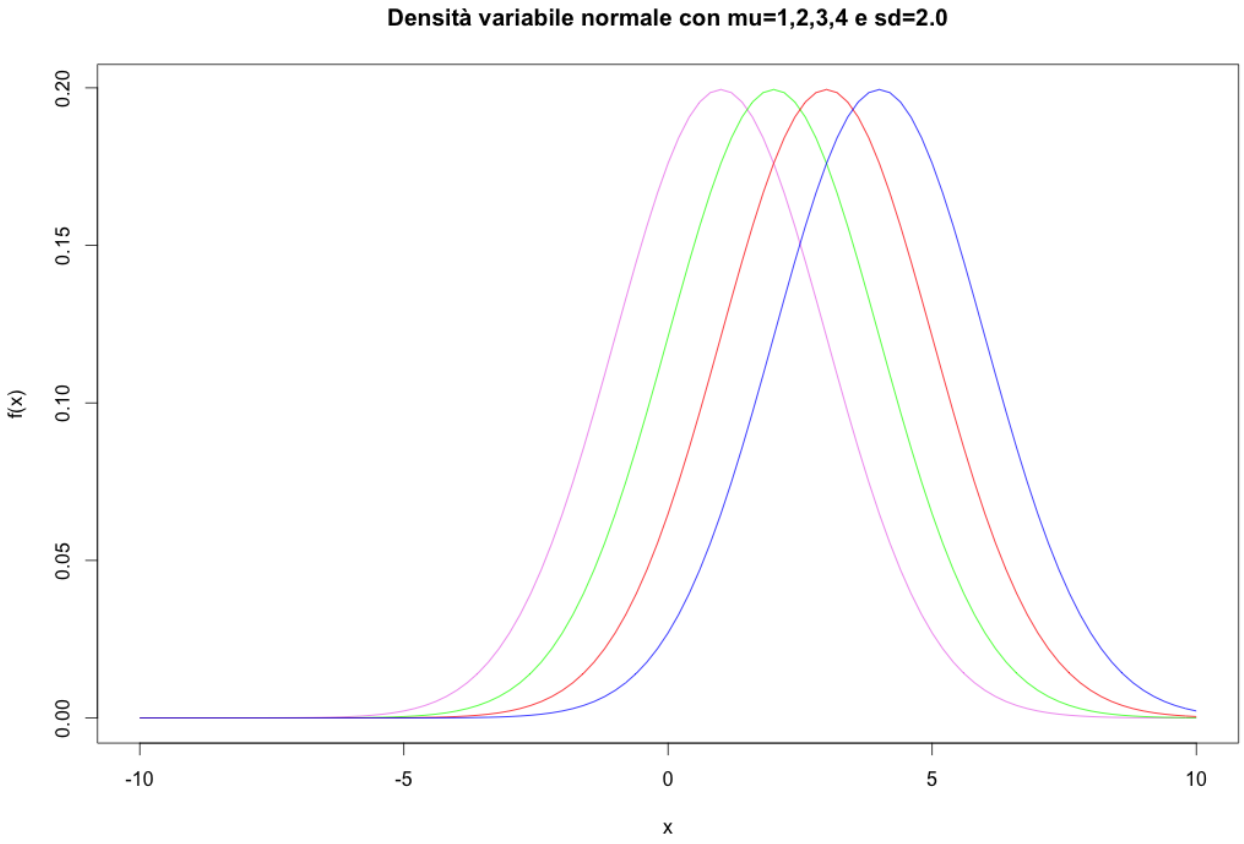
Viene riportato di seguito un esempio con il relativo codice R.

```
> curve(dnorm(x,mean=0.5,sd=2.0),from=-10,to=10,ylab="f(x)",
  xlab="x" ,main="Densita' variabile normale con mu=0.5 e sd=2.0")
```



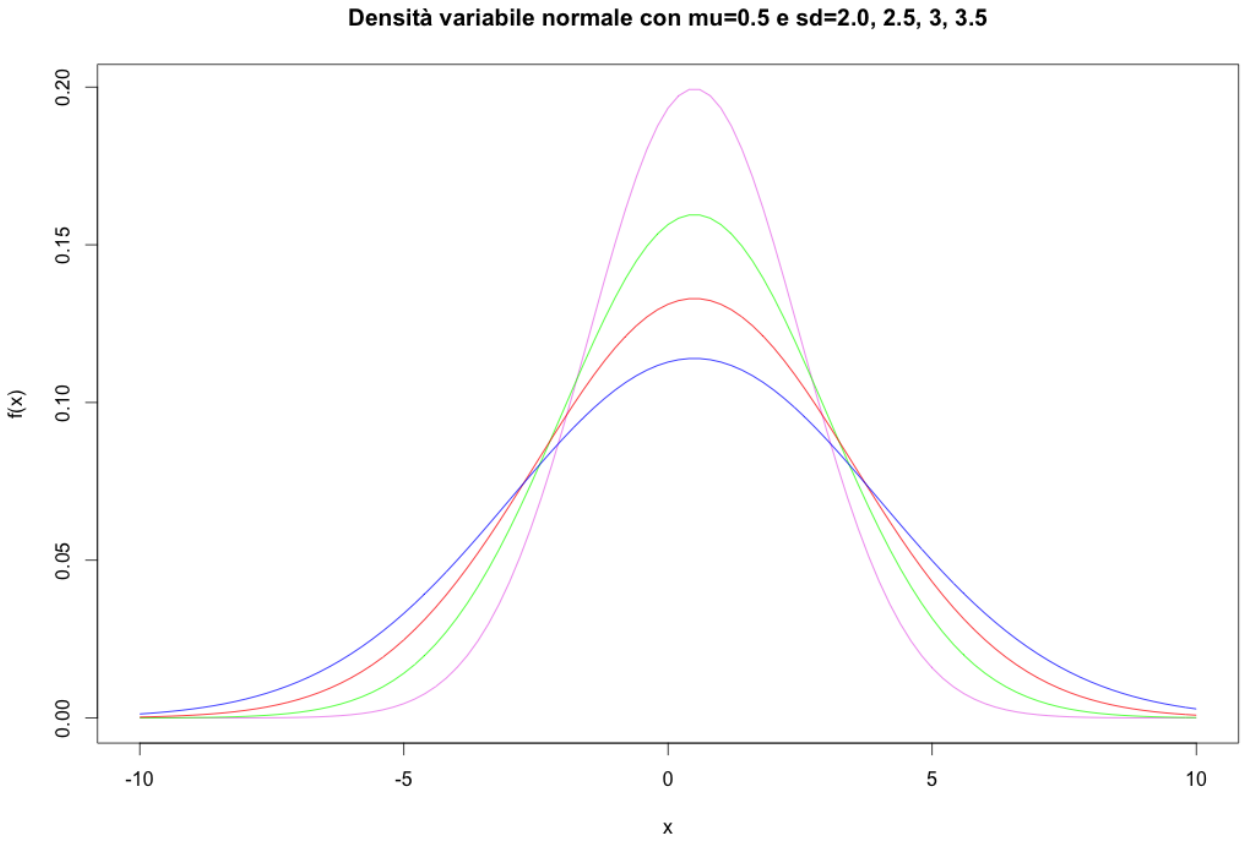
È possibile notare come variando μ si ha una traslazione della gaussiana lungo l'asse delle ascisse, mentre variando σ si ha una traslazione lungo l'asse delle ordinate.
Illustriamo questi due casi graficamente iniziando con la variazione di μ

```
> curve(dnorm(x,mean=1,sd=2.0),from=-10,to=10,ylab="f(x)",col="violet",
  xlab="x", main="Densita' variabile normale con mu=1,2,3,4 e sd=2.0")
> curve(dnorm(x,mean=2,sd=2.0),from=-10,to=10,ylab="f(x)",col="green"
  ,xlab="x",add=TRUE)
> curve(dnorm(x,mean=3,sd=2.0),from=-10,to=10,ylab="f(x)",xlab="x",
   col="red" ,add=TRUE)
> curve(dnorm(x,mean=4,sd=2.0),from=-10,to=10,ylab="f(x)",xlab="x",
   col="blue",add=TRUE)
```



Osserviamo i cambiamenti variando σ

```
> curve(dnorm(x,mean=0.5,sd=2.0),from=-10,to=10,ylab="f(x)",col="violet",
  xlab="x", main="Densita' variabile normale con mu=0.5 e sd=2.0, 2.5, 3, 3.5")
> curve(dnorm(x,mean=0.5,sd=2.5),from=-10,to=10,ylab="f(x)",col="green",
  xlab="x",add=TRUE)
> curve(dnorm(x,mean=0.5,sd=3.0),from=-10,to=10,ylab="f(x)",xlab="x",
  col="red" ,add=TRUE)
> curve(dnorm(x,mean=0.5,sd=3.5),from=-10,to=10,ylab="f(x)",xlab="x",
  col="blue",add=TRUE)
```



Quantili della distribuzione

```
qnorm (x, mean = mu, sd = sigma, lower.tail = TRUE, log.z = FALSE )
```

Dove:

- lower.tail se tale parametro è TRUE (caso di default) calcola $P(X \leqslant q)$ mentre se tale parametro è FALSE calcola $P(X > q)$.

Funzione di distribuzione

```
pnorm(x, mean = mu, sd = sigma , lower.tail = TRUE, log.p = FALSE)
```

Simulazione di una variabile aleatoria gaussiana

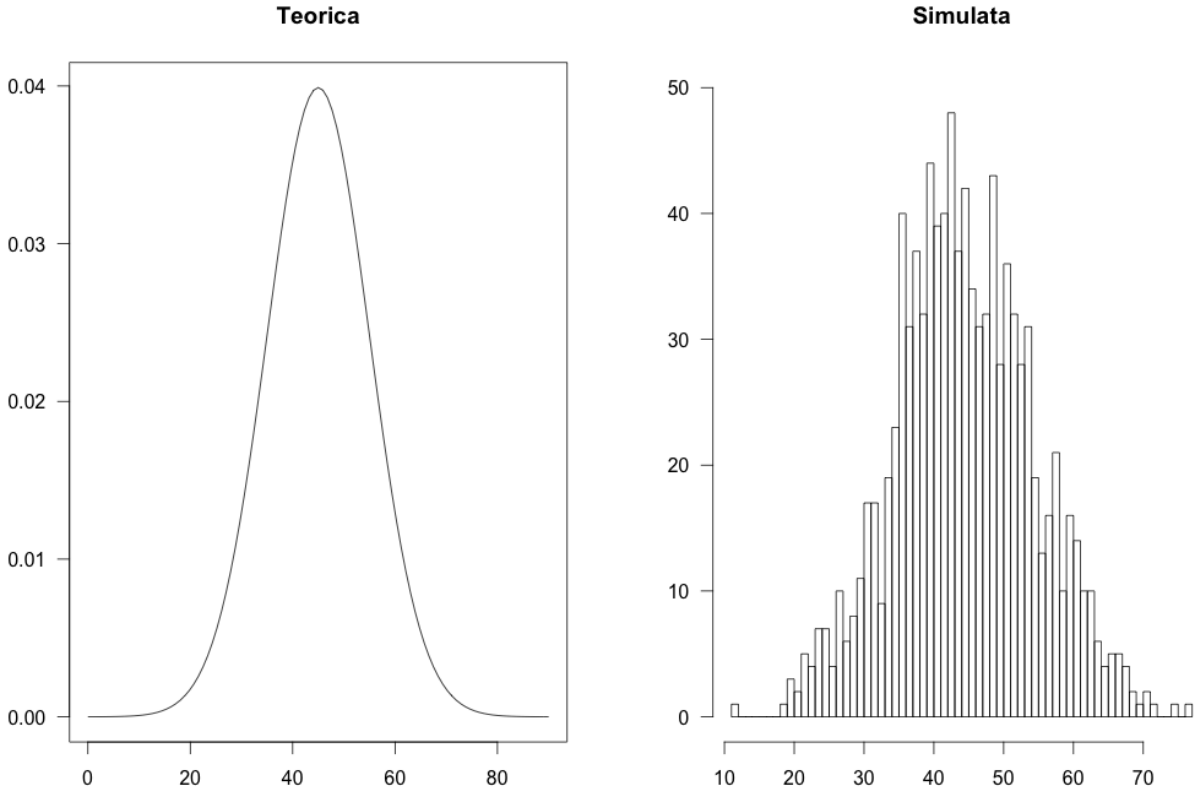
```
rnorm(N, mean = mu, sd = sigma)
```

dove:

- N è la lunghezza della sequenza da generare;
- Mean e sd sono il valore medio e la deviazione standard della densità normale.

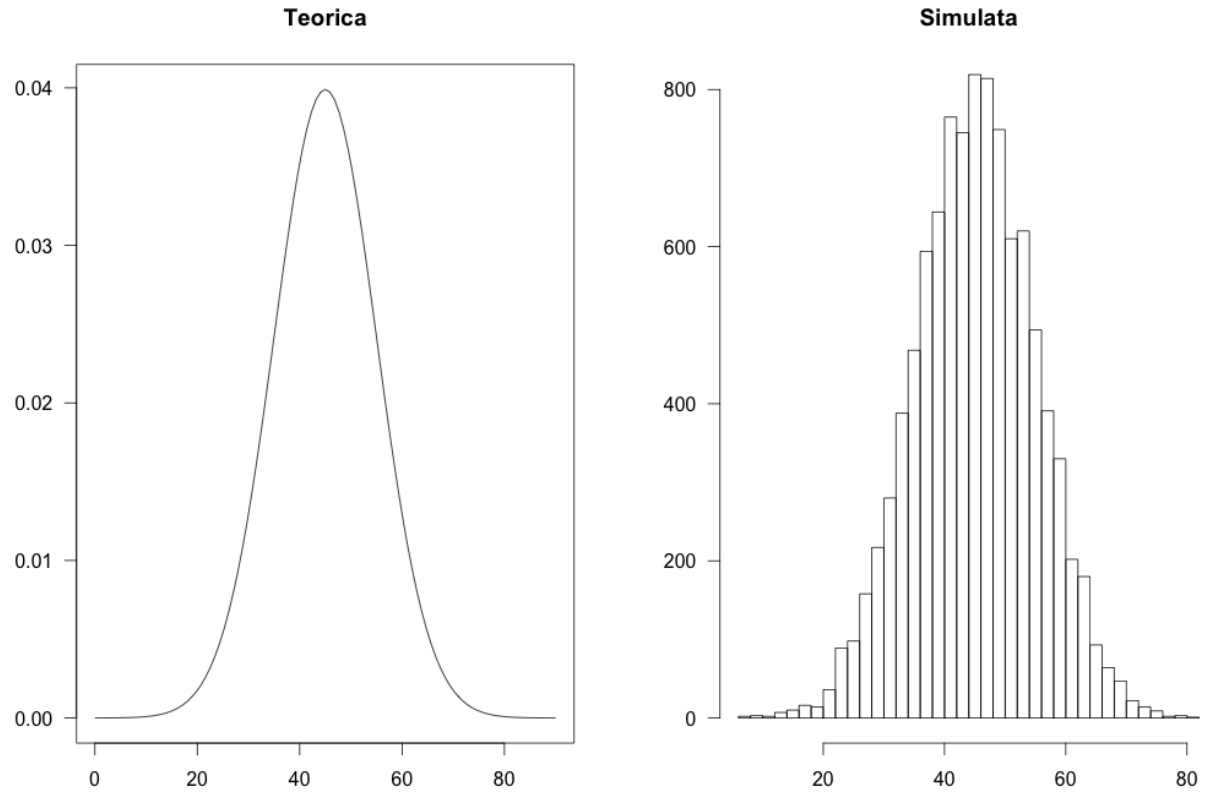
Tale funzione ritornerà utile per il calcolo degli intervalli di confidenza e per questo motivo di seguito verranno generate due popolazioni per mostrarne il funzionamento.
Nel primo caso generiamo una popolazione di 1000 unità, con valore medio 45 e deviazione standard 7.

```
> popolazione1 <- rnorm(1000, mean=45, sd=10)
> par(mfrow = c(1,2))
> curve(dnorm(x, mean=45, sd=10), from=0, to=90,xlab="", ylab="",
  main="Teorica", las=1)
> hist(popolazione1 , breaks=50, xlab="", ylab="",main = "Simulata",
  las=1,ylim=c(0,50))
```



Generiamo ora una popolazione di 10000 individui.

```
> popolazione2 <- rnorm(10000, mean=45, sd=10)
> par(mfrow = c(1,2))
> curve(dnorm(x, mean=45, sd=10), from=0, to=90,xlab="", ylab="",
main="Teorica", las=1)
> hist(popolazione2 , breaks=50, xlab="", ylab="",main = "Simulata",
las=1,ylim=c(0,50))
```



Si può notare come al crescere della popolazione la variabile simulata si avvicina a quella teorica.

2.1.1 Distribuzione chi-quadrato

La distribuzione chi-quadrato descrive la somma di quadrati di alcune variabili aleatorie indipendenti aventi distribuzione normale standard.
In statistica, viene particolarmente utilizzata per l'omonimo test di verifica d'ipotesi $test - X^2$ che permette di confrontare una serie di dati osservarti sperimentalmente con la serie dei dati attesi in base a unipotesi teorica e di stimare la bontà di questa ipotesi.
Una variabile aleatoria X di densità di probabilità:

$$f_X(x) = \begin{cases} \dfrac{1}{\Gamma(n/2)} \left(\dfrac{1}{2}\right)^{n/2} x^{(n/2)-1}\, e^{-x/2}, & se\ x > 0 \\ 0, & se\ x \leqslant 0 \end{cases}$$

con n intero positivo e $\Gamma(\nu)$ definita nel seguente modo:

$$\Gamma(\nu) = \int_0^{+\infty} x^{\nu-1}\, e^{-x}\, dx, \quad con\ \nu > 0$$

si dice distribuzione chi-quadrato con n gradi di libertà.
R permette di calcolare la densità di probabilità, la funzione di distribuzione e i quantili di una variabile aleatoria chi–quadrato e anche di simulare tale variabile.

Densità di propabilità

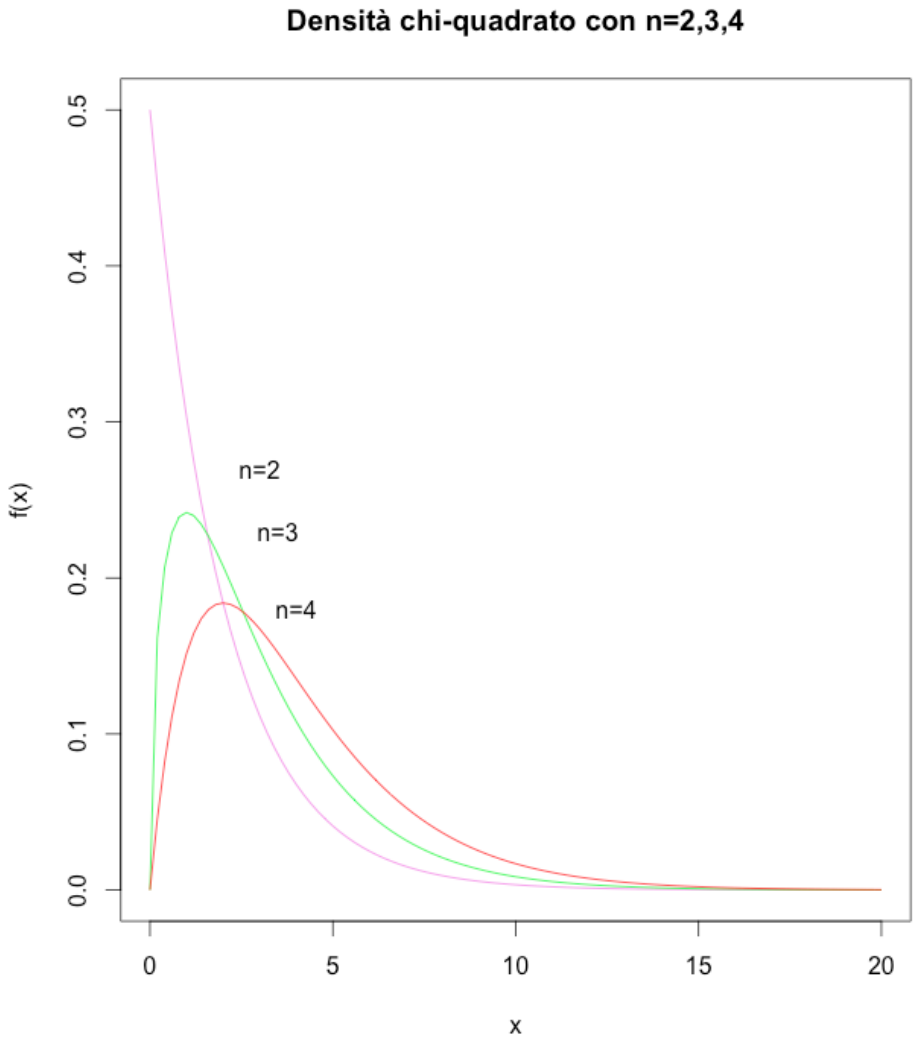
```
dchisq (x, df , log = FALSE )
```

Dove:

- x è il valore assunto dalla variabile aleatoria chi-quadrato;
- *df* è il numero di gradi di libertà;
- *log* se tale parametro è TRUE le probabilità sono espresse in forma logaritmica.

Viene riportato di seguito un esempio con R.

```
> curve(dchisq(x,df=2),xlab="x", col = "violet", ylab="f(x)",
  main="Densita' chi-quadrato con n=2,3,4",from=0,to=20,ylim=c(0,0.50))
> curve(dchisq(x,df=3), add = TRUE , col = "green")
> curve(dchisq(x,df=4), add = TRUE , col = "red")
> text(4,0.18,"n=4")
> text(3,0.27,"n=2")
> text(3.5,0.23,"n=3")
```



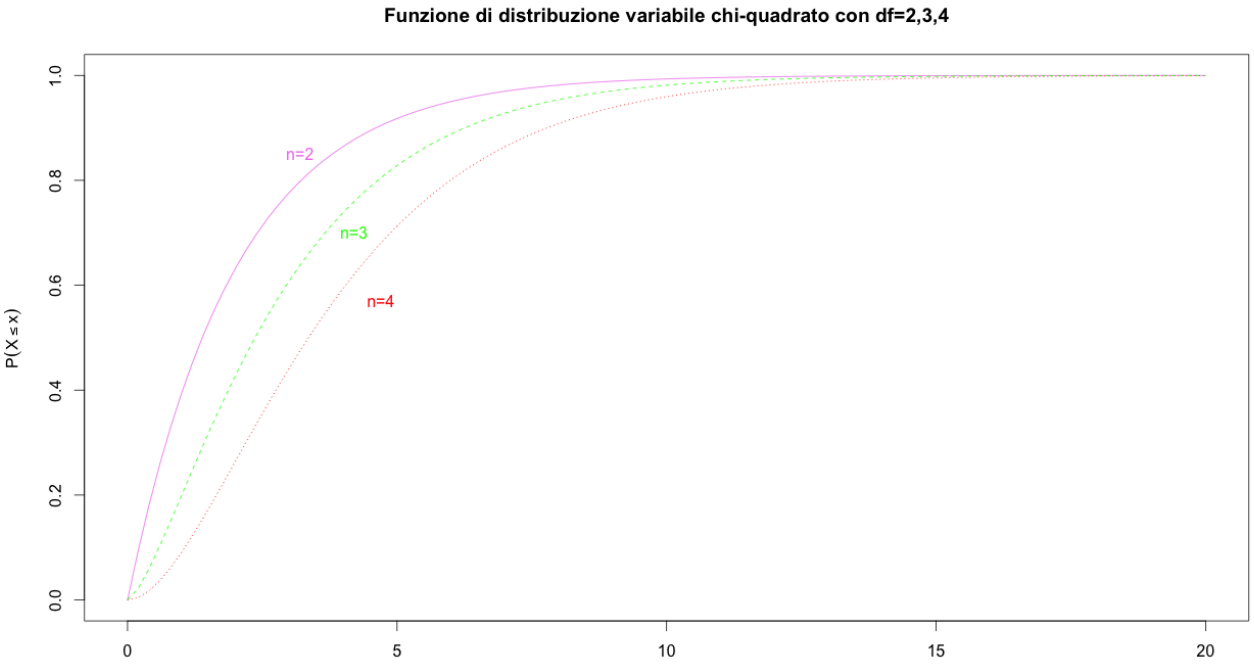
Funzione di distribuzione

```
> pchisq (q, df , lower .tail = TRUE , log .p = FALSE )
```

Dove:

- *lower.tail* se tale parametro è TRUE(caso di default) calcola $P(X \leqslant q)$ mentre se tale parametro è FALSE calcola $P(X > q)$.

```
> curve(pchisq(x,df=2),xlab="",col ="violet",ylab=expression(P(X<=x)),
  main="Funzione di distribuzione variabile chi-quadrato con df=2,3,4",
  from=0,to=20,ylim=c(0,1))
> curve(pchisq(x,df=4),add=TRUE , lty=3, col = "red")
> curve(pchisq(x,df=3),add=TRUE , lty=2, col = "green")
> text(4.2,0.7,"n=3",col="green")
> text(3.2,0.85,"n=2", col="violet")
> text(4.7,0.57,"n=4", col="red")
```



Calcolo dei quantili

```
> qchisq (p, df , lower .tail = TRUE , log .p = FALSE )
```

Simulazione di una variabile chi-quadrato

```
rchisq (N, df)
```

2.1.2 Distribuzione Student

In teoria delle probabilità, la distribuzione di Student, o t di Student, è una distribuzione di probabilità continua che governa il rapporto tra due variabili aleatorie, la prima con distribuzione normale

e la seconda il cui quadrato ha distribuzione chi quadrato.
È impiegata nel *t-test* che serve per confrontare le medie di due campioni che seguono la distribuzione normale e viene utilizzato o quando non è nota la varianza della popolazione o quando il campione è molto piccolo.
In termini formali una variabile aleatoria X di densità di probabilità:

$$f_X(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} \qquad x \in \mathbb{R}$$

Con n intero positivo e $\Gamma(\nu)$ definita come per chi-quadrato, si dice *distribuzione di Student*, o avere *distribuzione t di Student*, con n gradi di libertà.

Densità

Per il calcolo della densità si può utilizzare il comando:

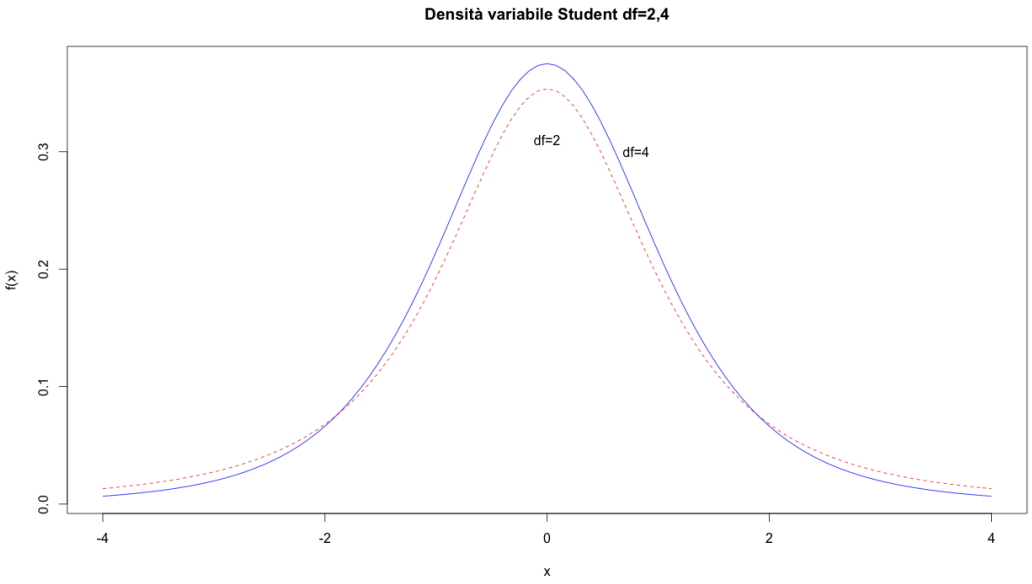
```
dt(x, df , log = FALSE )
```

dove:

- x 'e il valore assunto (o i valori assunti) dalla variabile aleatoria di Student;
- *df* numero di gradi di libertà;
- *log* se tale parametro è TRUE le probabilità sono espresse in forma logaritmica come $log(p)$.

Di seguito è riportato un esempio con R:

```
> curve(dt(x, df=4), from=-4, to=4, xlab="x", col = "blue", ylab="f(x)",
  main="Densita' variabile Student df=2,4")
> curve(dt(x, df=2), from=-4, to=4, xlab="x", col = "red", ylab="f(x)",
  add=TRUE, lty=2)
> text(0.8,0.30,"df=4")
> text(0,0.31,"df=2")
```



Funzione di distribuzione

```
pt(q, df , lower .tail = TRUE , log .p = FALSE )
```

Gli argomenti di tale funzione sono:

- q è il valore assunto (o i valori assunti) dalla variabile aleatoria di Student;
- df è il numero di gradi di libertà;
- *lower.tail* se tale parametro è TRUE calcola $P(X \leqslant q)$, se è FALSE calcola $P(X > q)$;
- *log.p* se TRUE le probabilità sono espresse in forma logaritmica come $log(p)$.

Calcolo dei quantili

```
qt(p, df , lower .tail = TRUE , log .p = FALSE )
```

Calcolo dei quantili

Simulazione di una variabile di Student

```
rt(N, df)
```