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 $\mu$ rappresenta la media, $\sigma$ la deviazione standard e $\sigma^2$ la varianza, si dice che $X$ ha distribuzione normale di parametri $\mu$ e $\sigma$.
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

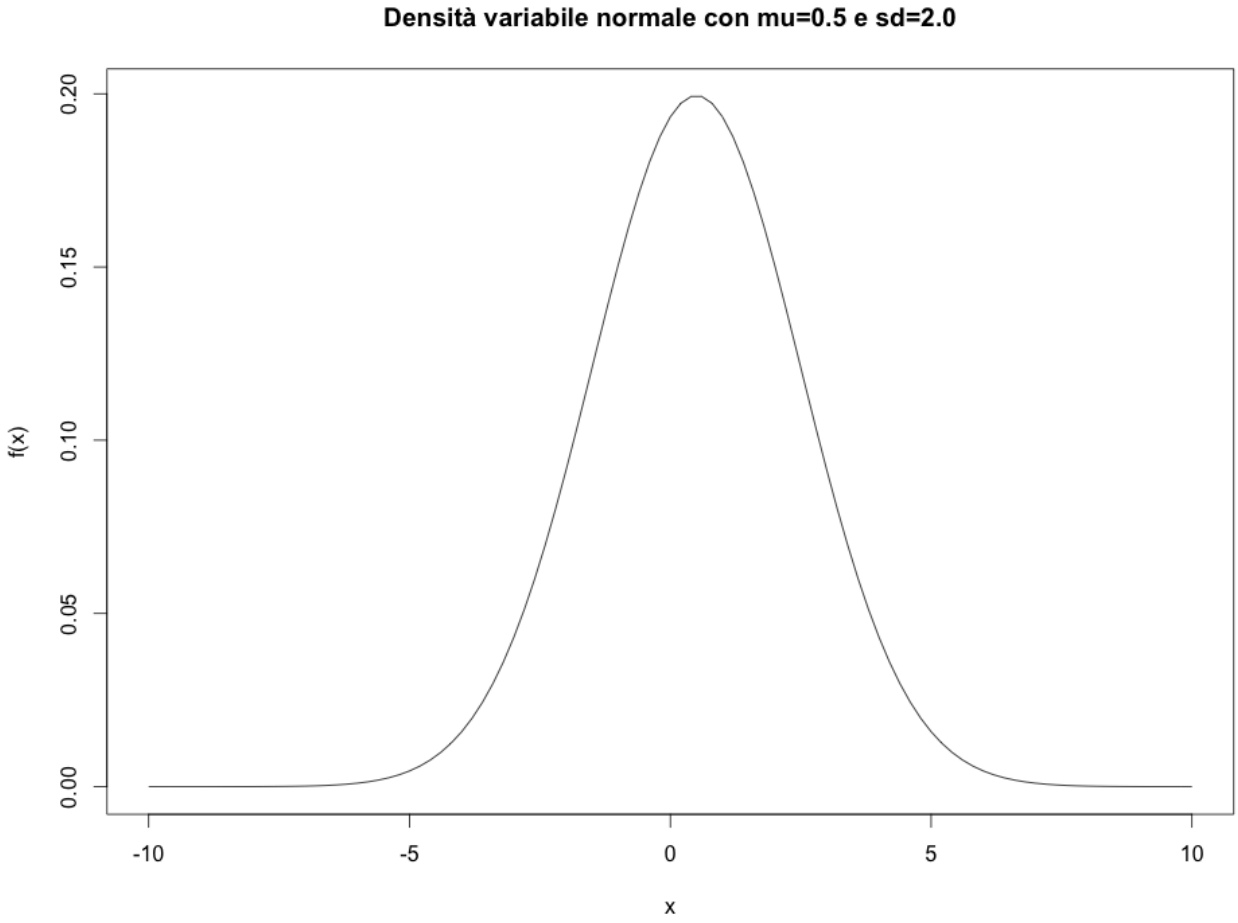
```
dnorm (x, mean = mu, sd = sigma, log = FALSE)
```

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 $log(p)$.

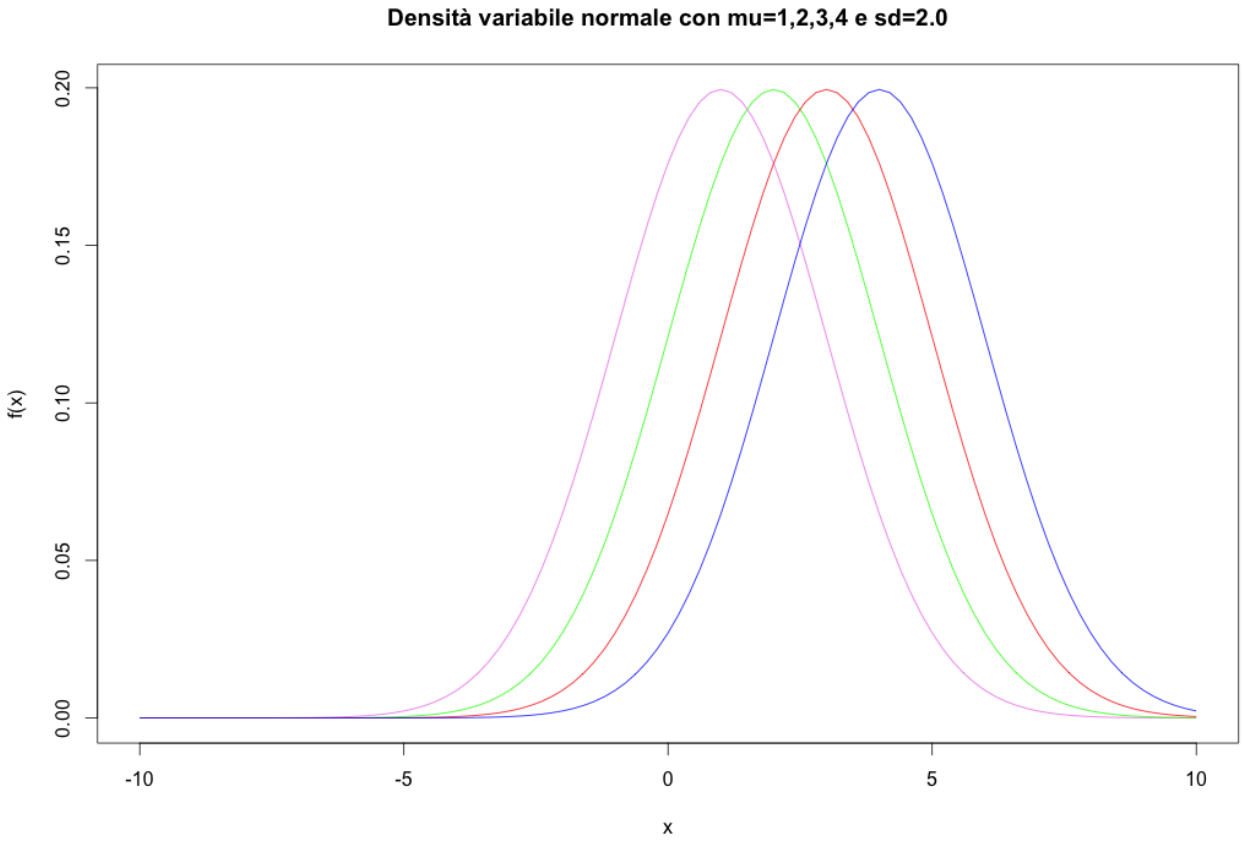
Viene riportato di seguito un esempio con il relativo codice R.

```
> 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")
```



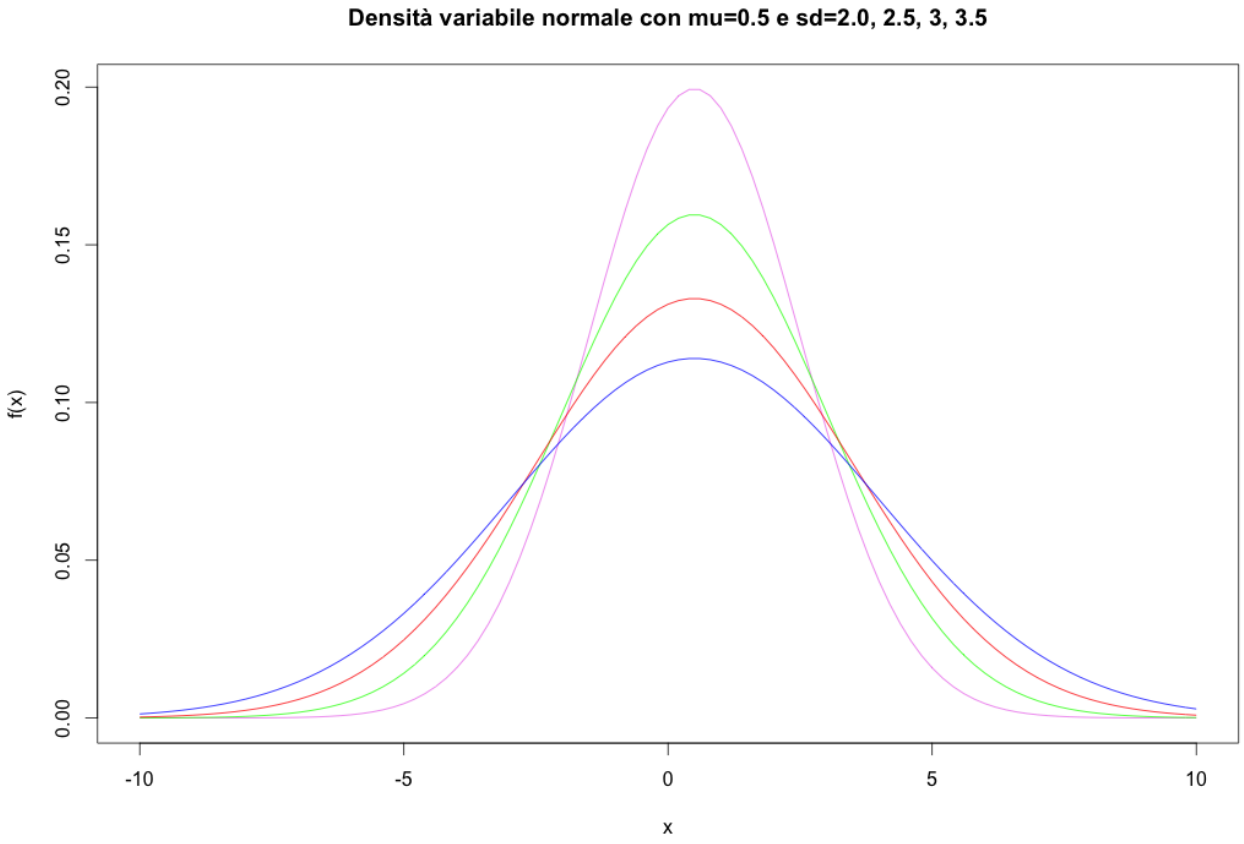

È possibile notare come variando $\mu$ si ha una traslazione della gaussiana lungo l'asse delle ascisse, mentre variando $\sigma$ si ha una traslazione lungo l'asse delle ordinate.
Illustriamo questi due casi graficamente iniziando con la variazione di $\mu$

```
> 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 $\sigma$

```
> 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

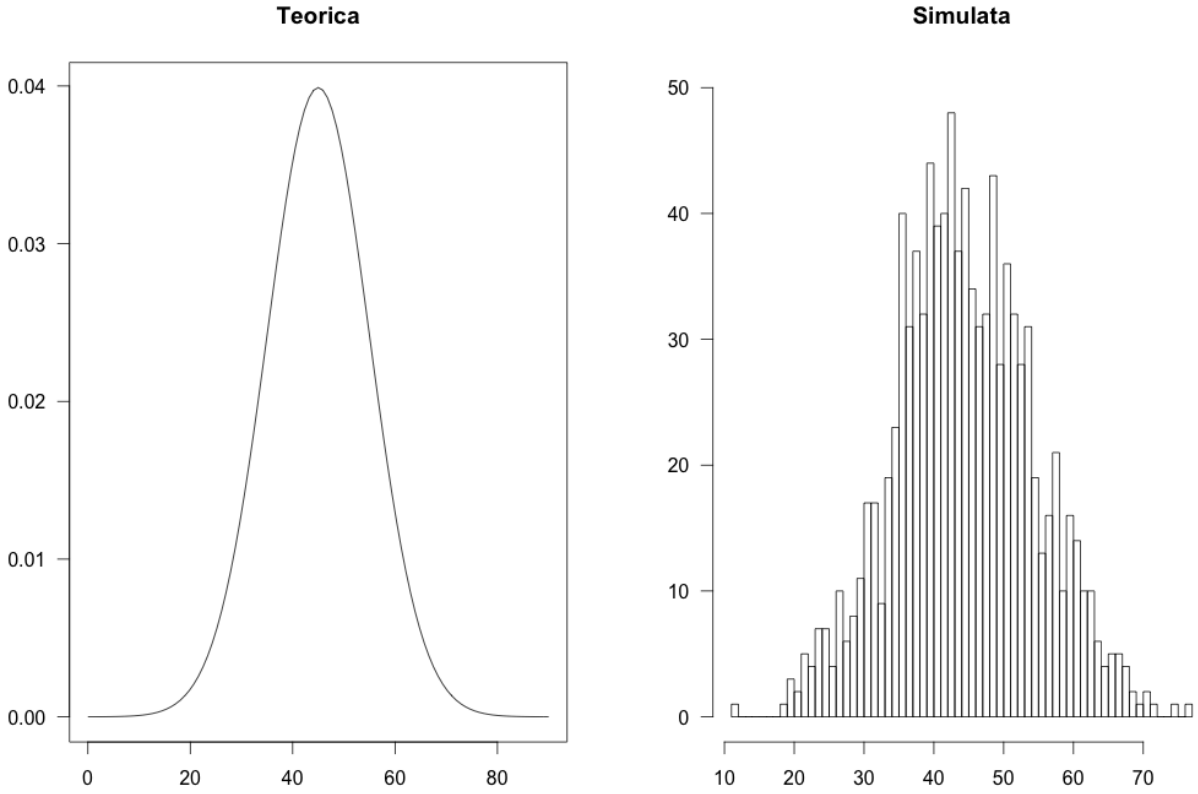
```
rnorm(N, mean = mu, sd = sigma)
```

dove:

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

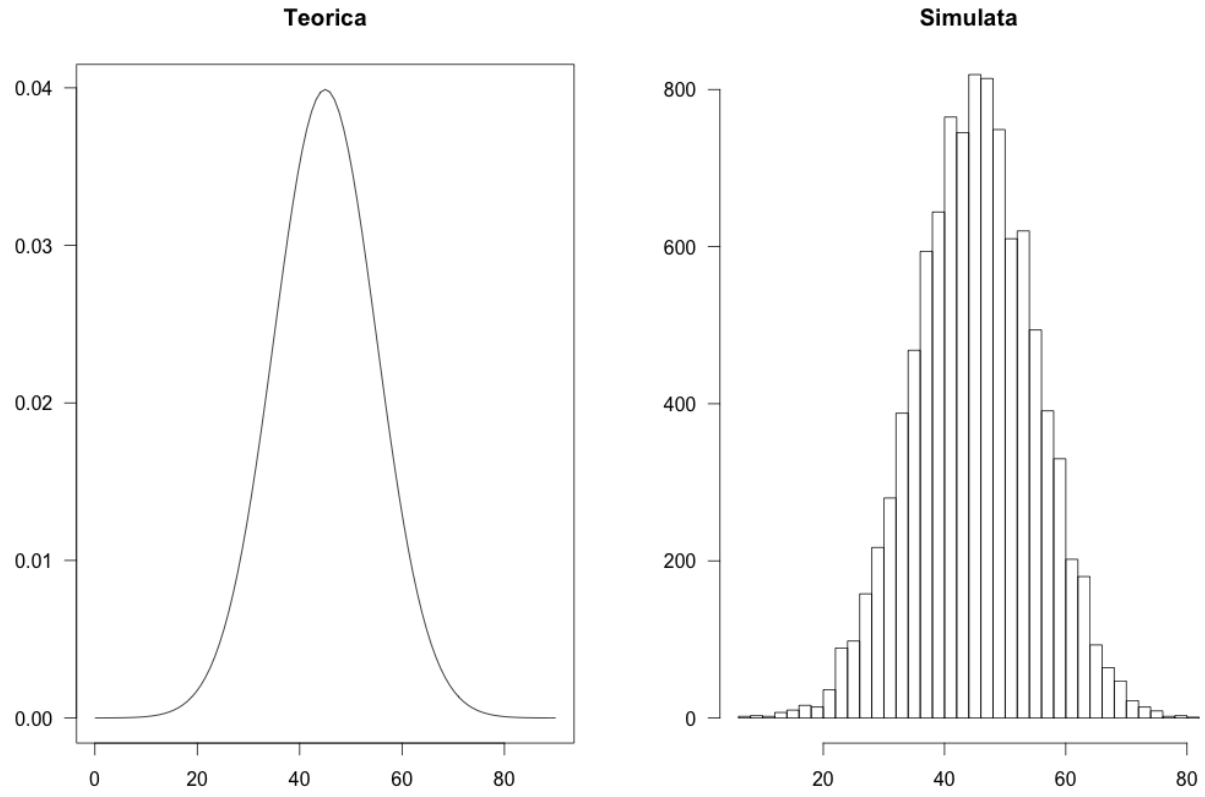
Tale funzione ritornerà utile per il calcolo degli intervalli di confidenza e per questo motivo di seguito verranno generate due popolazioni per mostrarne il funzionamento.
Nel primo caso generiamo una popolazione di 1000 unità, con valore medio 45 e deviazione standard 7.

```
> 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

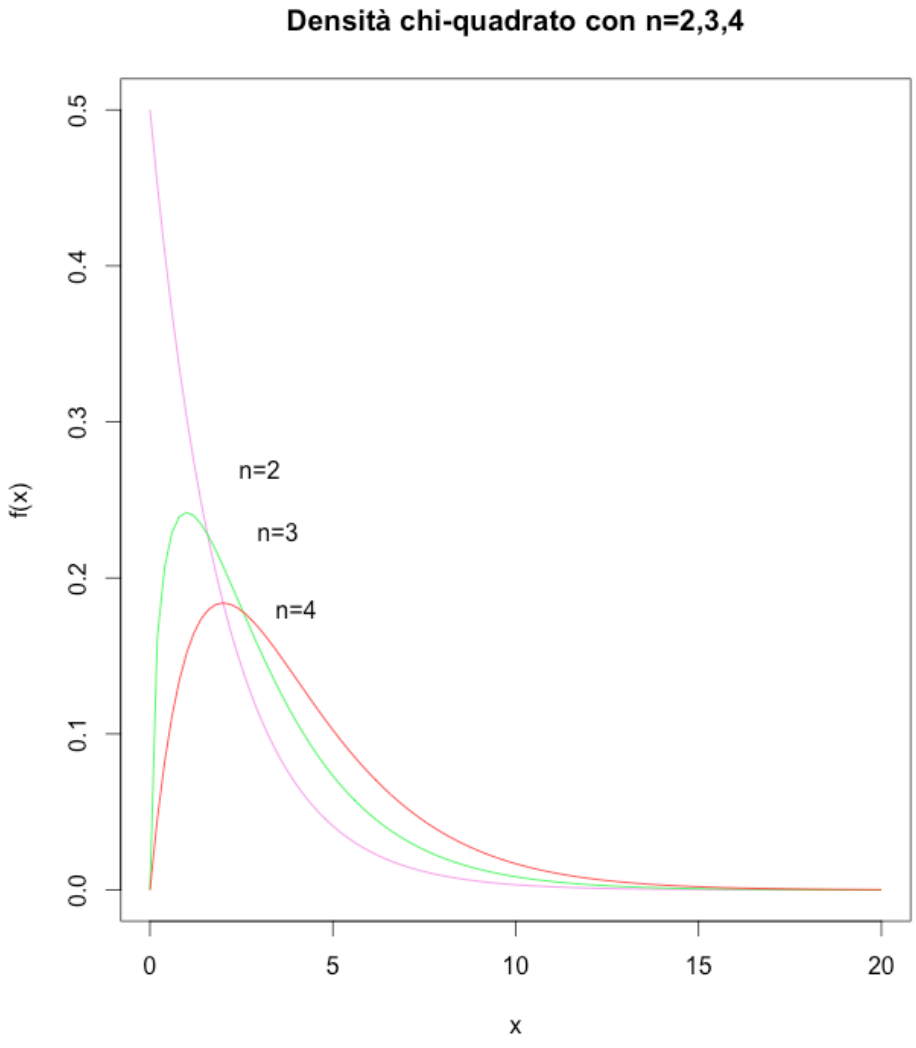
```
dchisq (x, df , log = FALSE )
```

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")
```



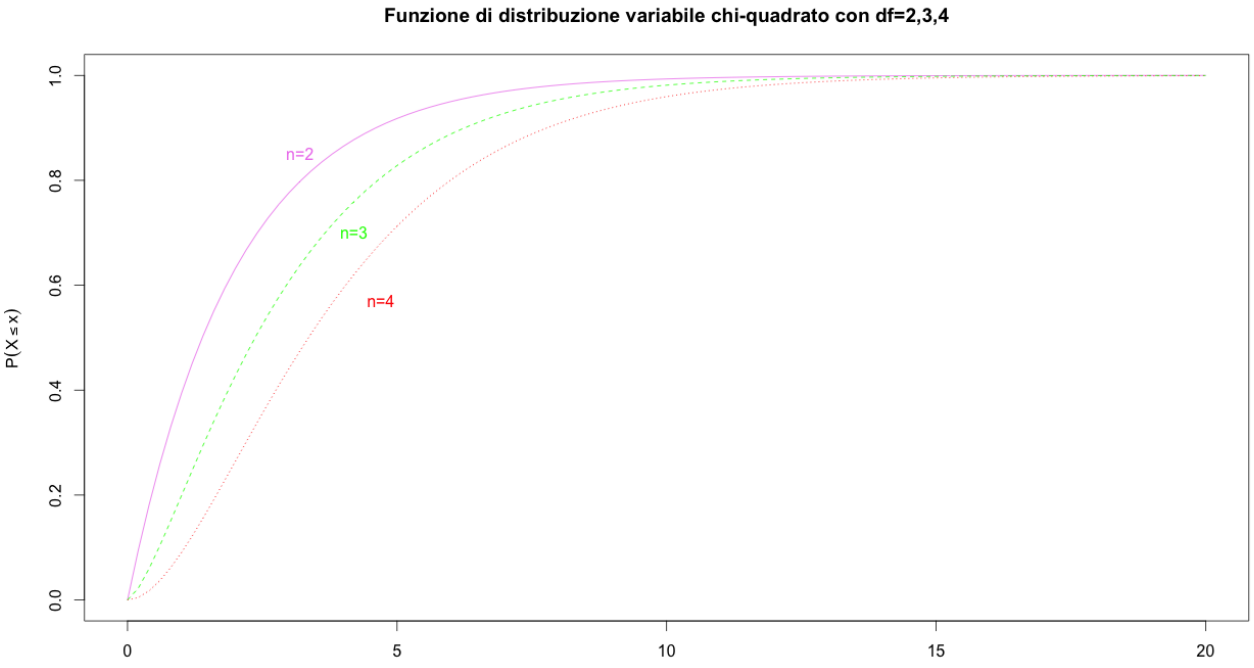
### Funzione di distribuzione

```
> 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:

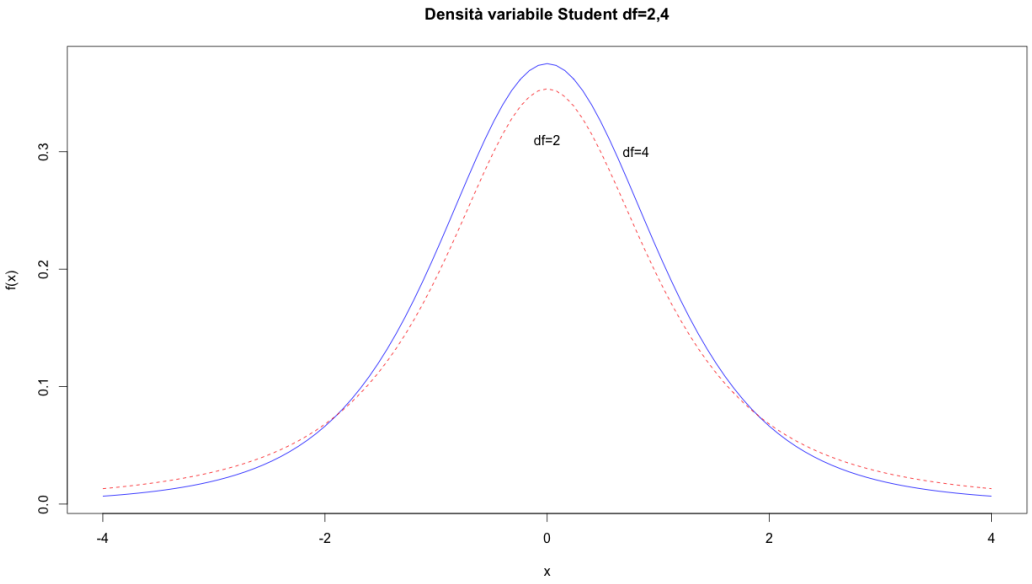
```
dt(x, df , log = FALSE )
```

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 )
```

**Simulazione di una variabile di Student**

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

# 3 Intervalli di confidenza

I metodi di stima puntuali anche se sembrano ipoteticamente sempre desiderabili, difficilmente potranno fornire delle stime che coincidono con un parametro incognito, poiché bisogna sempre tener conto di un certo errore di campionamento. Per questo motivo, alla stima puntuale di un parametro non noto di una popolazione (costituita da un singolo valore reale) spesso si preferisce sostituire un intervallo di valori, detto *intervallo di confidenza.*
In pratica, si cerca di determinare in base ai dati in possesso due limiti (uno superiore ed uno inferiore) entro i quali sia compreso il parametro non noto con un certo *coefficiente di confidenza* (detto anche grado di fiducia) $1-\alpha$ che rappresenta la probabilità di contenere il vero parametro della popolazione.
Definiamo formalmente quanto appena detto. Sia $X_1, X_2, ..., X_n$ un campione casuale di ampiezza $n$ estratto da una popolazione con funzione di probabilità (nel caso discreto) oppure densità di probabilità (nel caso assolutamente continuo) $f(x;\vartheta)$, dove $\vartheta$ denota il parametro non noto della popolazione. Denotiamo con $\underline{C}_n = g_1(X_1, X_2, ..., X_n)$ e con $\overline{C}_n = g_2(X_1, X_2, ..., X_n)$ due statistiche (funzioni osservabili del campione casuale) che soddisfino la condizione $\underline{C}_n < \overline{C}$, cioè che godono della proprietà che per ogni possibile fisasato campione osservato $x = (x_1, x_2, ..., x_n$ risulti $g_1(x) < g_2(x)$.
Fissato un coefficiente di confidenza $1-\alpha$ con $(0 < \alpha < 1)$, se possibile scegliere le statistiche $\underline{C}_n$ e $\overline{C}_n$ in modo tale che:

$$P(\underline{C}_n < \alpha < \overline{C}_n) = 1 - \alpha$$

allora si dice che $(\underline{C}_n; \overline{C}_n)$ è un intervallo di confidenza di grado $1-\alpha$ per $\vartheta$.
Inoltre, le statistiche $\underline{C}_n$ e $\overline{C}_n$ sono dette limite superiore e limite inferiore dell'intervallo di confidenza.
Se $g_1(x)$ e $g_2(x)$ sono i valori assunti dalle statistiche $\underline{C}_n$ e $\overline{C}_n$ per il campione osservato $x = (x_1, x_2, ..., x_n)$, allora l'intervallo $(g_1(x); g_2(x))$ è detto stima dell'intervallo di confidenza di grado $1-\alpha$ per $\vartheta$ ed i punti finali $g_1(x) e g_2(x)$ di tale intervallo sono detti rispettivamente stima del limite inferiore e stima del limite superiore dell'intervallo di confidenza.
Un metodo per la costruzione degli intervalli di confidenza è il metodo *pivoitale.* Tale metodo consiste essenzialmente nel determinare una variabile aleatoria di pivot $\lambda(X_1, X_2, ..., X_n; \vartheta)$ che presenta le seguenti caratteristiche:

- Dipende dal campione casuale $X_1, X_2, ..., X_n$.
- Dipende dal parametro non noto $\vartheta$.
- La funzione di distribuzione non contiene il parametro da stimare.

Tale variabile aleatoria non è una statistica poichè non è osservabile.

### Intervalli di confidenza per una popolazione normale

Sia $X_1, X_2, ..., X_n$ un campione casuale di ampiezza $n$ estratto da una popolazione normale con valore medio $\mu$ e varianza $\sigma^2$.
Si possono analizzare i seguenti problemi:

- Determinare un intervallo di confidenza di grado $1-\vartheta$ per il valore medio $\mu$ nel caso in cui la varianza $\sigma^2$ della popolazione normale è nota;
- Determinare un intervallo di confidenza di grado $1-\vartheta$ per il valore medio $\mu$ nel caso in cui la varianza $\sigma^2$ della popolazione normale non è nota;

- Determinare un intervallo di confidenza di grado $1-\vartheta$ per la varianza $\sigma^2$ nel caso in cui il valore medio della popolazione normale è noto;
- Determinare un intervallo di confidenza di grado $1-\vartheta$ per la varianza $\sigma^2$ nel caso in cui il valore medio della popolazione normale non è noto;

Di seguito illustreremo i casi sopra descritti.

## 3.1 Intervallo di confidenza per $\mu$ con $\sigma^2$ nota

Per determinare un intervallo di confidenza di grado $1-\vartheta$ per valore medio $\mu$ in cui la varianza $\sigma^2$ è fissata si utilizza il metodo pivoitale considerando la variabile aleatoria di pivot così definita:

$$Z_n = \frac{X_n - \mu}{\sigma/\sqrt{n}}$$

Tale variabile aleatoria è distribuita normalmente con valore medio nullo e varianza unitaria, dipende dal campione casuale e dal parametro non noto $\mu$ e quindi può essere interpretata come una variabile aleatoria di pivot.
Scegliendo nel metodo pivoitale $\alpha_1 = -z_\alpha/2$ e $\alpha_2 = +z_\alpha/2$, dove $z_\alpha$ è tale che:

$$P(Z_n < -z_{\alpha/2}) = P(Z_n > z_{\alpha/2}) = \frac{\alpha}{2}$$

si ha:

$$P(-z_{\alpha/2} < Z_n < z_{\alpha/2}) = 1 - \alpha$$

Sostituendo il valore di $Z_n$ otteniamo:

$$P\left(\overline{X}_n - z_{\alpha/2}\frac{\alpha}{\sqrt{n}} < \mu < \overline{X}_n + z_{\alpha/2}\frac{\alpha}{\sqrt{n}}\right)$$

Se poniamo

$$\underline{C}_n = \overline{X}_n - z_{\alpha/2}\frac{\alpha}{\sqrt{n}}, \quad \overline{C}_n = \overline{X}_n + z_{\alpha/2}\frac{\alpha}{\sqrt{n}}$$

allora $(\underline{C}_n; \overline{C}_n)$ è un intervallo di confidenza di grado $1-\alpha$ per $\mu$ e le statistiche $(\underline{C}_n e \overline{C}_n)$ rappresentano rispettivamente il limite inferiore e superiore di tale intervallo.
Da ciò possiamo dimostrare il seguente enunciato:
sia $x_1, x_2, ..., x_n$ un campione osservato di ampiezza $n$ estratto da una popolazione normale con varianza nota $\sigma^2$. Una stima dell'intervallo di confidenza di grado $1-\alpha$ per il valore medio $\mu$ è:

$$\overline{x}_n - z_{\alpha/2}\frac{\alpha}{\sqrt{n}} < \mu < \overline{x}_n + z_{\alpha/2}\frac{\alpha}{\sqrt{n}}$$

dove

$$\overline{x}_n = \frac{x_1, x_2, ..., x_n}{n}$$

denota la media campionaria delle n osservazioni.

**Esempio di applicazione con** $1 - \alpha = 0.99$

Di seguito viene mostrato un esempio prendendo un campione di 100 elementi della popolazione casuale generata in precedenza di 10000 unità.

```
> n <- 100
> x <- sample(popolazione2 , n, replace=TRUE)
```

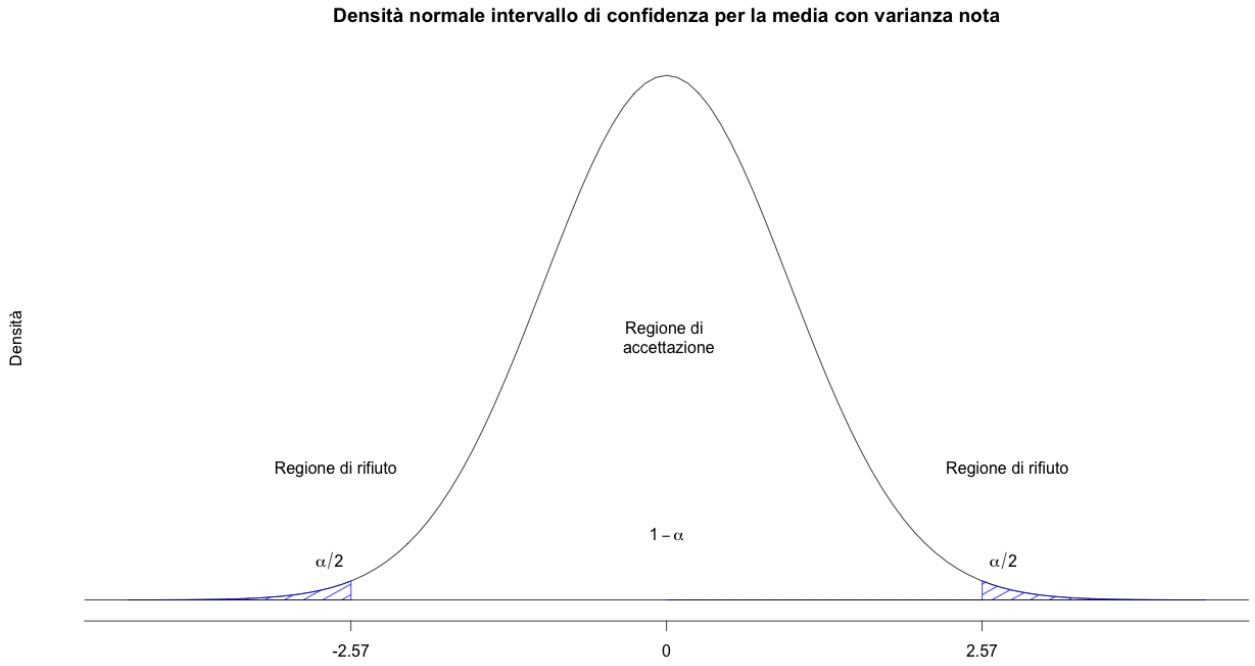
Supponiamo che la popolazione dalla quale proviene il campione sia normale e calcoliamo un intervallo di confidenza di grado $1 - \alpha = 0.99$ per la media $\mu$ degli elementi. In questo caso abbiamo che $\alpha = 0.01 e \alpha/2 = 0.005$. Inseriamo i dati in R e procediamo al calcolo degli intervalli di confidenza:

```
> alpha <- 0.01
> alphaMezzi <- alpha/2
> mediaCampionaria <- mean(x)
> z <- qnorm(1-alphaMezzi, mean=0, sd=1)
> b <- sd(popolazione2)/sqrt(n)
> c_1 <- mediaCampionaria - z * b
> c_2 <- mediaCampionaria + z * b
> mediaCampionaria
[1] 46.23546
> z
[1] 2.575829
> c_1
[1] 43.67419
> c_2
[1] 48.79672
```

La stima dell'intervallo di confidenza per $\mu$ con $1-\alpha = 0.99$ è pari a $(43.67419, 48.79672)$, ciò implica che il valore della media si trova in quell'intervallo con una probabilità $\mu$ con $1 - \alpha = 0.99$ o che alternativamente la media non si trova nell'intervallo che abbiamo stimato con probabilità $\sigma = 0.01$.
La lunghezza dell'intervallo di confidenza è pari a $48.79672 - 43.67419 = 5.122528$.
Possiamo anche rappresentare graficamente quanto ottenuto implementando i seguenti comandi:

```
> curve(dnorm(x,mean=0,sd=1),from=-4.4, to=4.4, axes=FALSE , xlab="",
  ylab="Densita'", main="Densita' normale intervallo di confidenza
  per la media con varianza nota")
> text(0,0.05,expression(1-alpha))
> text(0,0.2,"Regione di \n accettazione")
> axis(1,c(-5,-z,0,z,5), c("",abbreviate(-z,5) ,0,abbreviate(z,4),""))
> vals<-seq(-4.4,-z, length =100)
> x1<-c(-4.4,vals ,-z,-4.4)
> y1<-c(0,dnorm(vals) ,0,0)
> polygon(x1,y1,density=20,angle=45, col="blue")
> vals<-seq(z,4.4, length=100)
> x1<-c(z,vals ,4,0)
> y1<-c(0,dnorm(vals) ,0,0)
> polygon(x1,y1,density=20,angle=45, col="blue")
> abline(h=0)
> text(-2.75,0.03, expression (alpha/2))
> text(-2.70,0.10,"Regione di rifiuto")
> text(2.75,0.03, expression (alpha/2))
> text(2.80,0.10,"Regione di rifiuto ")
```

Con il seguente risultato:



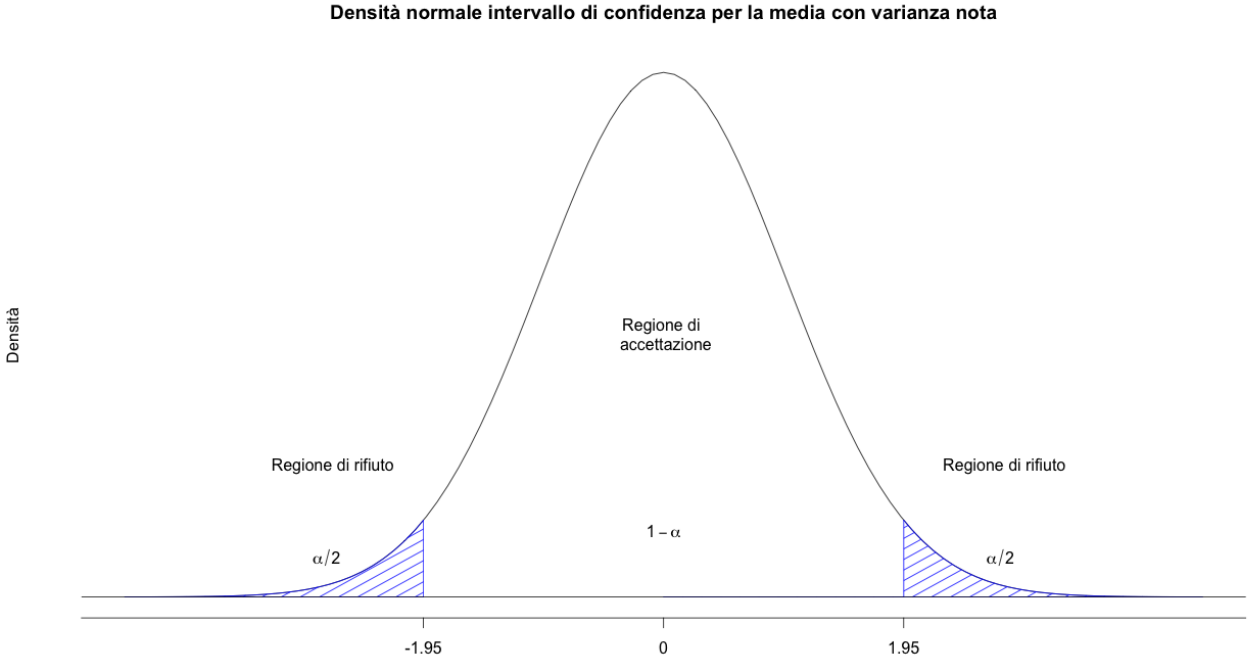

Per riempire il grafico (regione di rifiuto) è stato utilizzato il comando *polygon* che necessita in input di due vettori di coordinate x e y.
Il vettore *vals* contiene una successione di 100 valori tra i due estremi e il vettore x è costruito in modo tale che la prima e l'ultima coordinata x dei punti del poligono coincidano. Il vettore y deve invece contenere tutti i punti di ordinata pari alla densità normale standard, esclusi i valori estremi in cui si pone y uguale a 0. Infine, la funzione *polygon* traccia finalmente il poligono riempendolo di linee inclinate di 45 gradi e equispaziate.

### Esempio di applicazione con $1 - \alpha = 0.95$

In questo secondo esempio determiniamo la stima dell'intervallo di confidenza per $1 - \alpha = 0.95$ del campione precedente per osservarne i risultati. In questo caso abbiamo che $\alpha = 0.05$ e $\alpha/2 = 0.025$ . Inseriamo i dati opportuni in R e procediamo al calcolo di $z_{\alpha/2} = z_{0.025}$

```
> alpha <- 0.05
> alphaMezzi <- alpha/2
> mediaCampionaria <- mean(x)
> z <- qnorm(1-alphaMezzi, mean=0, sd=1)
> b <- sd(popolazione2)/sqrt(n)
> c1 <- mediaCampionaria - z * b
> c2 <- mediaCampionaria + z * b
> mediaCampionaria
[1] 44.79331
> z
[1] 1.959964
> c1
[1] 42.83057
> c2
[1] 46.75604
```

La stima dell'intervallo di confidenza per $\mu$ con $1-\alpha = 0.95$ è pari a $(42.83057, 46.75604)$, ciò implica che il valore della media si trova in quell'intervallo con una probabilità $\mu$ con $1-\alpha = 0.95$ o che alternativamente la media non si trova nell'intervallo che abbiamo stimato con probabilità $\sigma = 0.05$. La lunghezza dell'intervallo di confidenza è pari a $46.75604 - 42.83057 = 3.925469$, rispetto all'esempio precedente la lunghezza dell'intervallo di confidenza è minore in quanto nel caso precedente viene richiesta una maggiore sicurezza di trovare il valore non noto e quindi ampliamo il range dei possibili valori disponibili.
Possiamo anche rappresentare graficamente i risultati ottenuti riutilizzando il codice precedente, il risultato è il seguente:




## 3.2 Intervallo di confidenza $\mu$ con $\sigma^2$ non nota

Per determinare un intervallo di confidenza di grado $1-\alpha$ per valore medio $\mu$ in cui la varianza $\sigma^2$ della popolazione normale non è nota, utilizziamo il metodo pivoitale e consideriamo la seguente variabile aletoria di pivot:

$$T_n = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$$

Dove $X_n$ ed $S_n$ rappresentano rispettivamente la media e la deviazione standard. $T_n$ dipende dal campione casuale e dal parametro non noto $\sigma^2$ e quindi può essere interpretata come variabile di pivot e si può dedurre che è distribuita con legge di Student con $n-1$ gradi di libertà.
Scegliendo nel metodo pivoitale $\alpha_1 = -t_{\alpha/2,\ n-1}$ e $\alpha_2 = +t_{\alpha/2,\ n-1}$ dove $t_{\alpha/2,\ n-1}$ è tale che:

$$P\left(T_n < -t_{\alpha/2,\ n-1}\right) = P\left(T_n > +t_{\alpha/2,\ n-1}\right) = \frac{\alpha}{2}$$

si ha

$$P\left(-t_{\alpha/2,\ n-1} < T_n < +t_{\alpha/2,\ n-1}\right) = 1-\alpha$$

Sostituendo il valore di $T_n$ otteniamo:

$$P\left(\overline{X}_n - t_{\alpha/2,\ n-1}\frac{S_n}{\sqrt{n}} < \mu < \overline{X}_n + t_{\alpha/2,\ n-1}\frac{S_n}{\sqrt{n}}\right) = 1-\alpha$$

Se poniamo

$$\underline{C}_n = \overline{X}_n - t_{\alpha/2,\ n-1}\frac{S_n}{\sqrt{n}}, \qquad \overline{C}_n = \overline{X}_n + t_{\alpha/2,\ n-1}\frac{S_n}{\sqrt{n}}$$

allora $(\underline{C}_n, \overline{C}_n)$ è un intervallo di confidenza di grado $1-\alpha$ per $\mu$ e le statistiche $\underline{C}_n$ e $\overline{C}_n$ rappresentano rispettivamente il limite inferiore e superiore di tale intervallo.

**Esempio di applicazione con** $1-\alpha = 0.95$

Prendiamo in considerazione un campione di 100 elementi da una popolazione di 1000 elementi:

```
> n <- 100
> x <- sample(popolazione1 , n, replace=TRUE)
```

Supponendo che la popolazione da cui proviene il campione risulta essere normale, calcoliamo un intervallo di confidenza di grado $1-\alpha = 0.95$ per la media $\mu$ senza avere informazioni sulla varianza. In questo caso abbiamo $\alpha = 0.05$ e $\alpha/2 = 0.025$. Inseriamo i dati in R e determiniamo il valore di $t_{\alpha/2,n-1} = t_{0.025,99}$:

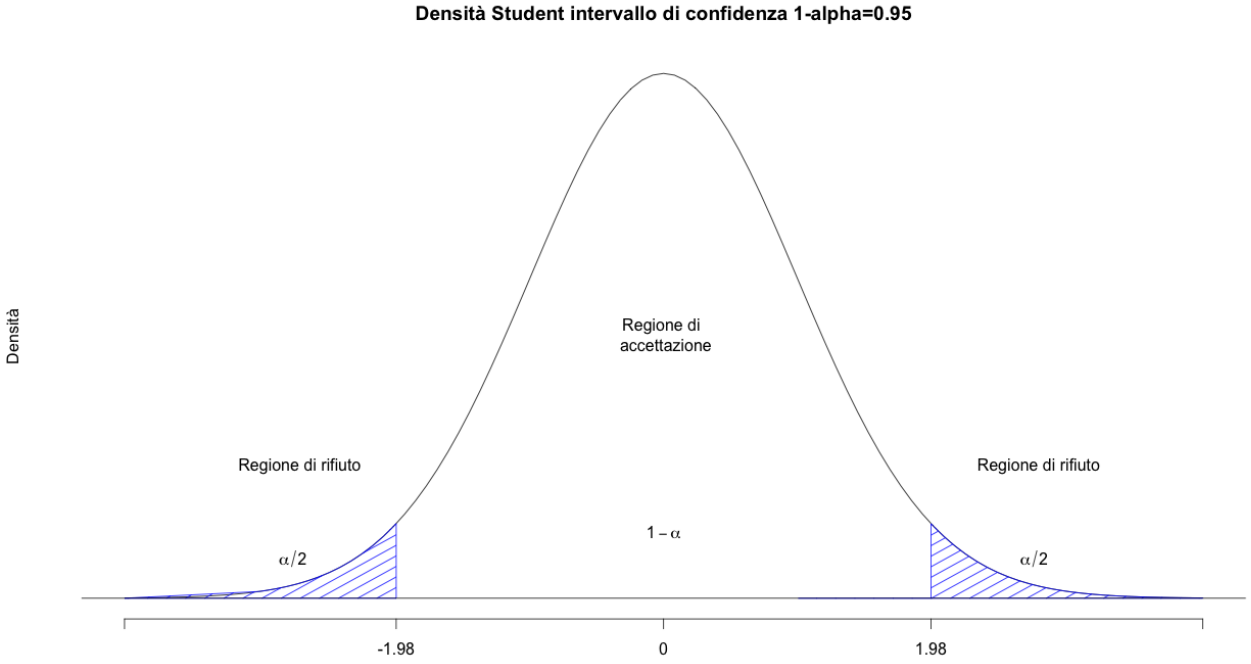
```
> alpha <- 0.05
> alphaMezzi <- alpha/2
> mediaCampionaria <- mean(x)
> ta <- qt(1-alphaMezzi , df=n-1)
> ds <- sd(x)    #deviazione standard (varianza al quadrato)
> lowC <- mediaCampionaria - ta*ds/sqrt(n)
> highC <- mediaCampionaria + ta*ds/sqrt(n)
> mediaCampionaria
[1] 44.43933
> ta
[1] 1.984217
> ds
[1] 10.66209
> lowC
[1] 42.32374
> highC
[1] 46.55492
```

La stima dell'intervallo di confidenza di grado $1-\alpha = 0.95$ è $(42.32374, 46.55492)$ e la lunghezza di tale intervallo è pari a $(46.55492 - 42.32374) = 4.231182$. Il codice per eseguire il grafico è il seguente:

```
> curve(dt(x,df=n-1),from=-4, to=4, axes=FALSE , xlab="", ylab="
Densita'",
  main="Densita' Student intervallo di confidenza 1-alpha=0.95")
> axis(1,c(-4,-ta,0,ta,4),c("",abbreviate(-ta ,5),0,abbreviate(ta,4),""))
> text(0,0.05,expression(1-alpha))
> text(0,0.2,"Regione di \n accettazione")
> vals<-seq(-3,-ta, length =100)
> x1<-c(-4,vals ,-ta,-4)
> y1<-c(0,dt(vals ,n-1) ,0,0)
> polygon(x1,y1,density=20,angle=45, col="blue")
```

```
> vals<-seq(ta,4, length=100)
> x1<-c(ta,vals ,4,1)
> y1<-c(0,dt(vals ,n-1) ,0,0)
> polygon(x1,y1,density=20,angle=45, col="blue")
> abline(h=0)
```

Con il seguente risultato



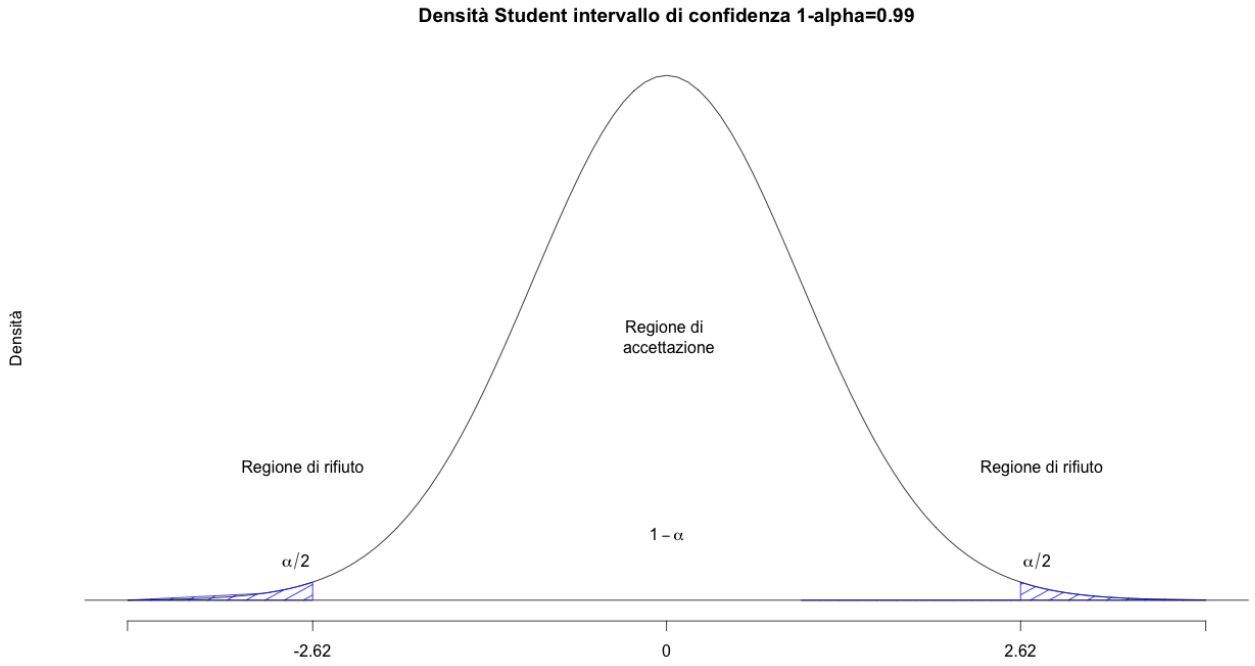

**Esempio di applicazione con** $1 - \alpha = 0.99$

Riprendiamo in considerazione il campione esaminato in precedenza e calcoliamo un intervallo di confidenza di grado $1 - \alpha = 0.00$ per $\mu$ senza avere informazioni sulla varianza.
In questo caso abbiamo che $\alpha = 0.01$ e $\alpha/2 = 0.005$.
Inseriamo i dati in R e procediamo al calcolo degli intervalli di confidenza:

```
alpha <- 0.01
alphaMezzi <- alpha/2
mediaCampionaria <- mean(x)
mediaCampionaria <- mean(x)
ta <- qt(1-alphaMezzi, df=n-1)
ds <- sd(x) #deviazione standard (varianza al quadrato)
lowC <- mediaCampionaria - ta*ds/sqrt(n)
highC <- mediaCampionaria + ta*ds/sqrt(n)
> mediaCampionaria
[1] 46.0899
> ta
[1] 2.626405
> ds
[1] 9.991823
> lowC
```

```
[1] 43.46564
> highC
[1] 48.71416
```

La stima dell'intervallo di confidenza di grado $1-\alpha = 0.99$ è $(43.46564, 48.71416)$ e la lunghezza di tale intervallo è pari a $48.71416 - 43.46564 = 5.248516$ che risulta maggiore di quello calcolato in precedenza. Di seguito vengono rappresentati graficamente i risultati ottenuti, il codice è equivalente a quello utilizzato in precedenza.




## 3.3 Intervallo di confidenza per $\sigma^2$ con $\mu$ nota

Per costruire un intervallo di confidenza di grado $1-\alpha$ per la varianza $\sigma^2$ nel caso in cui il valore medio $\mu$ della popolazione normale è noto viene utilizzato il metodo pivotale considerando la seguente variabile aleatoria di pivot:

$$V_n = \sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right)^2$$

Tale variabile dipende dal campione casuale e dal parametro non noto $\sigma^2$ ed è distribuita con legge chi-quadrato con $n$ gradi di libertà, essendo costituita dalla somma dei quadrati di $n$ variabili aleatorie normali standard. Possiamo riscrivere $V_n$ in termini di media e varianza:

$$V_n = \frac{(n-1)S^2}{\sigma^2} + \left( \frac{\overline{X}_n - \mu}{\mu/\sqrt{n}} \right)^2$$

Tale variabile aleatoria dipende dal campione casuale e dal parametro non noto $\sigma^2$ ed è distribuita con legge di chi-quadrato con n gradi di libertà, essendo costituita dalla somma dei quadrati di $n$ variabili aleatorie normali standard. Scegliendo nel metodo pivoitale $\alpha_1 = X^2_{1-\alpha/2,n}$ e $\alpha_2 = X^2_{\alpha/2,n}$ in maniera tale che:

$$P(0 < V_n < X^2_{1-\alpha/2,n}) = P(V_n > X^2_{\alpha/2,n}) = \frac{\alpha}{2}$$

si ha:

$$P(X^2_{1-\alpha/2,n} < V_n < X^2_{\alpha/2,n} = 1-\alpha)$$

Sostituendo il valore di $V_n$ otteniamo:

$$P\left(\frac{(n-1)S_n^2 + n(\overline{X}_n - \mu)^2}{X^2_{\alpha/2,n}} < \sigma^2 < \frac{(n-1)S_n^2 + n(\overline{X}_n) - \mu^2}{X^2_{1-\alpha/2,n}}\right) = 1-\alpha$$

Se poniamo:

$$\underline{C}_n = \frac{(n-1)S_n^2 + n(\overline{X}_n - \mu)^2}{X^2_{\alpha/2,n}}; \qquad \overline{C}_n = \frac{(n-1)S_n^2 + n(\overline{X}_n) - \mu^2}{X^2_{1-\alpha/2,n}}$$

allora $(\underline{C}_n, \overline{C}_n)$ è un intervallo di confidenza di grado $1-\alpha$ per $\sigma^2$ e le statistiche $\underline{C}_n$ e $\overline{C}_n$ rappresentano rispettivamente il limite inferiore e superiore di tale intervallo.

**Esempio di applicazione con** $1-\alpha = 0.99$

Prendiamo in considerazione il seguente intervallo:

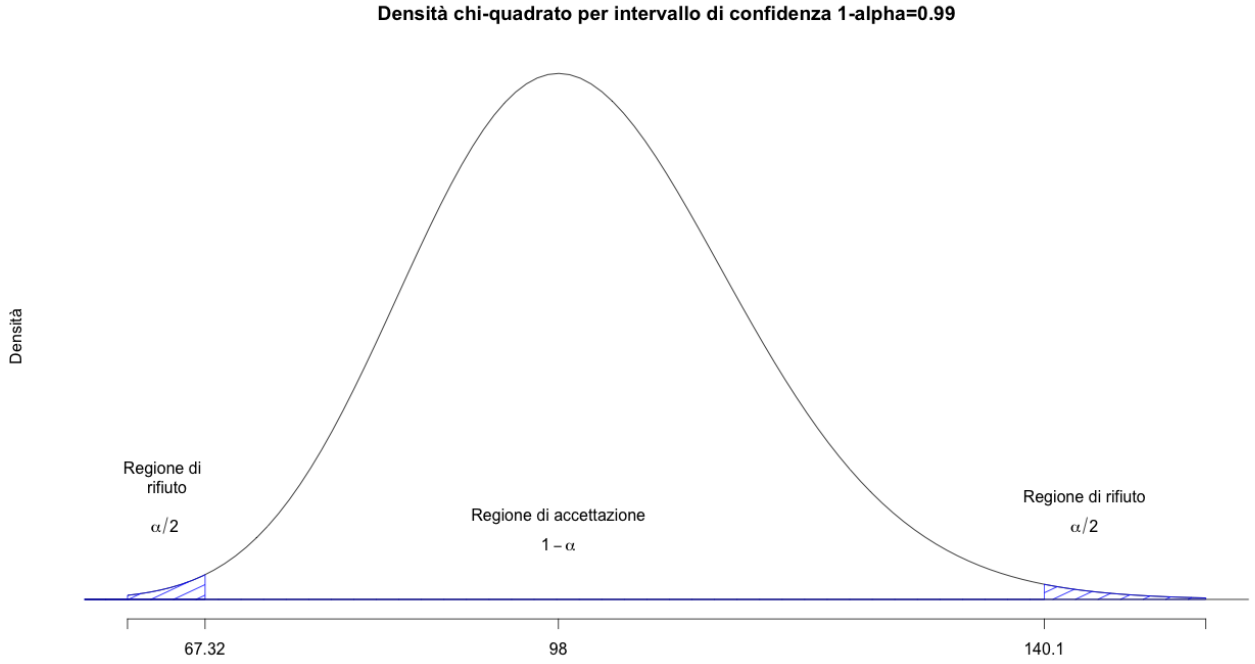
```
> n <- 100
> x <- sample(popolazione1 , n, replace=TRUE)
```

Supponendo che la popolazione da cui proviene il campione preso in considerazione è normale con $\mu = 41.70$ determiniamo una stima per l'intervallo di confidenza $1-\alpha = 0.99$ per la varianza $\sigma^2$. In questo caso $\alpha = 0.01$, $\alpha/2 = 0.005$ e $1-\alpha/2 = 0.995$. I valori di $X^2_{1-\alpha/2,n} = X^2_{0.995,100}$ e $X^2_{\alpha/2,n} = X^2_{0.005,100}$ possono essere calcolati grazie ad R.
I comandi sono i seguenti:

```
> alpha <- 0.01
> alphaMezzi <- alpha/2
> varianzaC <- var(x)
> z1 <- qchisq(1-alphaMezzi , df=n)
> z2 <- qchisq(alphaMezzi , df=n)
> lowC <- ((n-1)*varianzaC+n*((mean(x)-41.70)**2))/qchisq(1-alphaMezzi,df=n)
> highC <- ((n-1)*varianzaC+n*((mean(x)-41.70)**2))/qchisq(alphaMezzi,df=n)
> alpha
[1] 0.01
> alphaMezzi
[1] 0.005
> varianzaC
[1] 119.6361
> z1
[1] 140.1695
> z2
[1] 67.32756
> lowC
[1] 85.09089
> highC
[1] 177.151
```

La stima dell'intervallo di confidenza di grado $1-\alpha = 0.99$ è $(85.09089, 177.151)$ e la lunghezza dell'intervallo è pari a$177.151 - 85.09089 = 92.06011$. I risultati ottenuti possono essere rappresentati graficamente attraverso il seguente codice R.

```
> curve(dchisq(x, df=n), from=z2-z2/10, to=z1+z1/10, axes=FALSE ,xlab ="",
  ylab="Densita'",main="Densita' chi-quadrato per intervallo di confidenza
  1-alpha=0.99")
> axis(1,c(z2-z2/10,z2,n-2,z1,z1+z1/10),c("",abbreviate(z2,minl=5),n-2,
  abbreviate(z1,minl=5),""))
> vals<-seq(z2-z2/10,z2, length =100)
> x1<-c(z2-z2/10,vals ,z2 ,0)
> y1<-c(0,dchisq(vals ,n) ,0,0)
> polygon (x1,y1,density =20,angle=45, col="blue")
> vals<-seq(z1,z1+z1/10, length =100)
> x1<-c(z1,vals ,z1+z1/10,1)
> y1<-c(0,dchisq(vals ,n) ,0,0)
> polygon (x1,y1,density =20,angle=45, col="blue")
> abline(h=0)
> text(z1+3.5,0.004,expression(alpha/2))
> text(z1+3.5,0.0055,"Regione di rifiuto")
> text(98,0.003,expression(1-alpha))
> text(98,0.0045,"Regione di accettazione")
> text(z2 -3.5,0.0066,"Regione di \n rifiuto")
> text(z2 -3.5,0.004,expression(alpha/2))
```



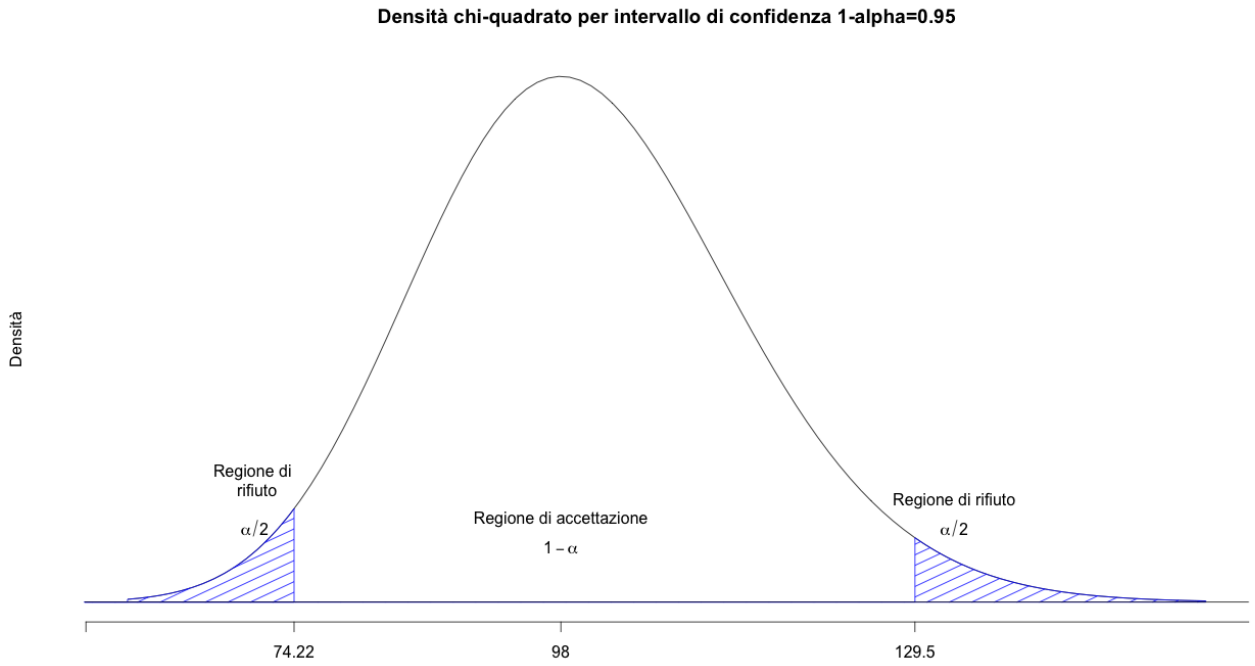

**Esempio di applicazione con** $1-\alpha = 0.95$

Riprendendo l'intervallo considerato in precedenza determiniamo una stima per l'intervallo di confidenza $1-\alpha = 0.99$ per la varianza $\sigma^2$. In questo caso $\alpha = 0.05$, $\alpha/2 = 0.025$ e $1-\alpha/2 = 0.975$. I valori di $X^2_{1-\alpha/2,n} = X^2_{0.975,100}$ e $X^2_{\alpha/2,n} = X^2_{0.025,100}$ possono essere calcolati grazie ad R.

```
> alpha <- 0.05
```

```
> alphaMezzi <- alpha/2
> varianzaC <- var(x)
> z1 <- qchisq(1-alphaMezzi , df=n)
> z2 <- qchisq(alphaMezzi , df=n)
> lowC <- ((n-1)*varianzaC+n*((mean(x) -41.70)**2))/qchisq(1-alphaMezzi,df=n)
> highC <- ((n-1)*varianzaC+n*((mean(x)-41.70)**2))/qchisq(alphaMezzi,df=n)
> alpha
[1] 0.05
> alphaMezzi
[1] 0.025
> varianzaC
[1] 83.76991
> z1
[1] 129.5612
> z2
[1] 74.22193
> lowC
[1] 71.94914
> highC
[1] 125.5938
```

La stima dell'intervallo di confidenza di grado $1-\alpha = 0.95$ è $(71.94914, 125.5938)$ e la lunghezza dell'intervallo è pari a$125.5938 - 71.94914 = 53.64466$. I risultati ottenuti possono essere rappresentati graficamente attraverso il codice R eseguito nell'esempio precedente.




### 3.3.1 Intervallo di confidenza per $\sigma^2$ con $\mu$ non nota

Per costruire un intervallo di confidenza di grado $1-\alpha$ per la varianza nel caso in cui il valore medio $\mu$ della popolazione normale non è noto viene utilizzato il metodo pivotale considerando la seguente

variabile aleatoria di pivot:

$$Q_n = \frac{(n-1)S_n^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Tale variabile aleatoria dipende dal campione casuale e dal parametro non noto $\sigma^2$ ed è distribuita con legge di chi-quadrato con n gradi di libertà. Scegliendo nel metodo pivoitale $\alpha_1 = X^2_{1-\alpha/2,n-1}$ e $\alpha_2 = X^2_{\alpha/2,n-1}$ in maniera tale che:

$$P(0 < Q_n < X^2_{1-\alpha/2,n-1}) = P(Q_n > X^2_{\alpha/2,n-1}) = \frac{\alpha}{2}$$

Si ha:

$$P(X^2_{1-\alpha/2,n-1} < Q_n < X^2_{\alpha/2,n-1} = 1-\alpha)$$

Sostituendo il valore di $Q_n$ otteniamo:

$$P\left(\frac{(n-1)S_n^2}{X^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S_n^2}{X^2_{1-\alpha/2,n-1}}\right) = 1-\alpha$$

Se poniamo:

$$\underline{C}_n = \frac{(n-1)S_n^2}{X^2_{\alpha/2,n-1}}; \qquad \overline{C}_n = \frac{(n-1)S_n^2}{X^2_{1-\alpha/2,n-1}}$$

allora $(\underline{C}_n, \overline{C}_n)$ è un intervallo di confidenza di grado $1-\alpha$ per $\sigma^2$ e le statistiche $\underline{C}_n$ e $\overline{C}_n$ rappresentano rispettivamente il limite inferiore e superiore di tale intervallo.

**Esempio di applicazione con** $1-\alpha = 0.95$

Prendiamo in considerazione il seguente intervallo:

```
> n <- 100
> x <- sample(popolazione1 , n, replace=TRUE)
```

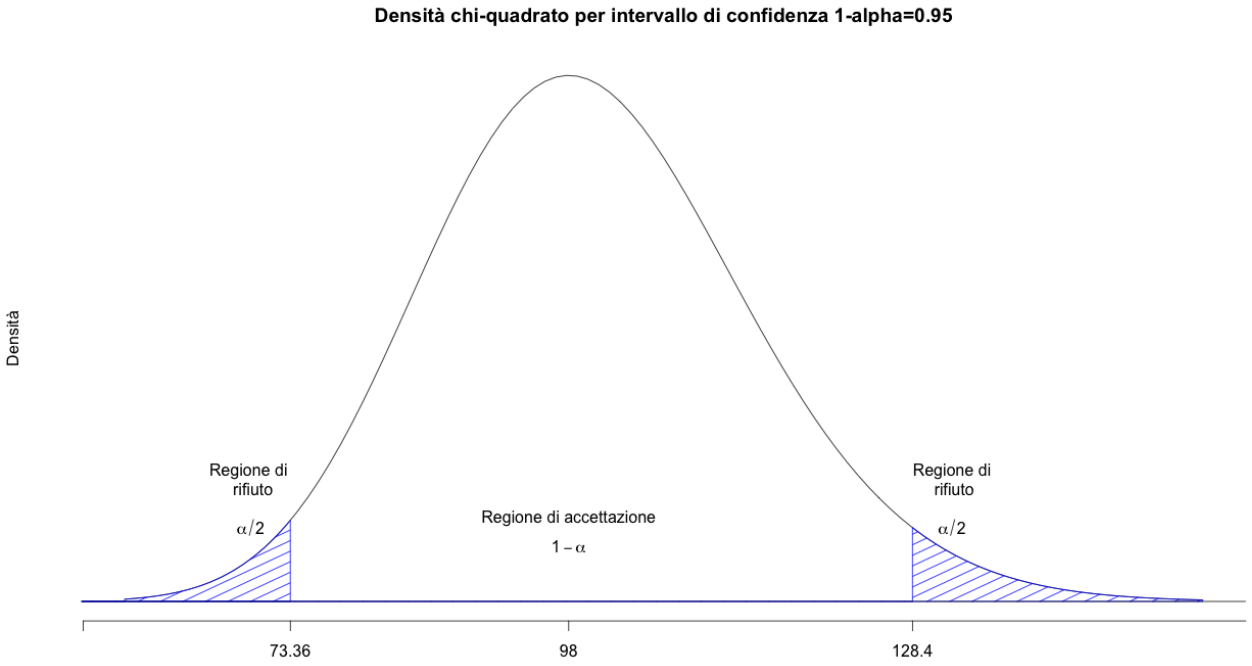
Supponendo che la popolazione da cui proviene il campione sia normale, determiniamo una stima dell'intervallo di confidenza di grado $1-\alpha = 0.95$ per la varianza $\sigma^2$ degli elementi senza possedere informazioni su $\mu$. In questo caso $\alpha = 0.05$, $\alpha/2 = 0.025$ e $1-\alpha/2 = 0.975$. I valori di $X^1_{1-\alpha/2,n-1} = X^2_{0.975,99}$ e $X^2_{\alpha/2,n-1} = X^2_{0.025,99}$ possono essere calcolati grazie ad R.

```
> alpha <- 0.05
> alpha_mezzi <- alpha/2
> varianza <- var(x)
> z1<-qchisq(1-alpha_mezzi ,df=n-1)
> z2<-qchisq(alpha_mezzi ,df=n-1)
> lowC <- ((n-1)*varianza)/qchisq(1-alpha_mezzi ,df=n-1)
> highC <- ((n-1)*varianza)/qchisq(alpha_mezzi ,df=n-1)
> alpha
[1] 0.05
> alpha_mezzi
[1] 0.025
> varianza
[1] 109.7469
> z1
```

```
[1] 128.422
> z2
[1] 73.36108
> lowC
[1] 84.60344
> highC
[1] 148.1023
```

La stima dell'intervallo di confidenza di grado $1-\alpha = 0.95$ è $(84.60344, 148.1023)$ e la lunghezza dell'intervallo è pari a$148.1023 - 84.60344 = 63.49886$. I risultati ottenuti possono essere rappresentati graficamente attraverso il seguente codice R:

```
> n <- 100
> x <- sample(popolazione1 , n, replace=TRUE)
> alpha <- 0.05
> alpha_mezzi <- alpha/2
> varianza <- var(x)
> z1<-qchisq(1-alpha_mezzi ,df=n-1)
> z2<-qchisq(alpha_mezzi ,df=n-1)
> lowC <- ((n-1)*varianza)/qchisq(1-alpha_mezzi ,df=n-1)
> highC <- ((n-1)*varianza)/qchisq(alpha_mezzi ,df=n-1)
> curve(dchisq(x, df=n), from=z2-z2/5, to=z1+z1/5, axes=FALSE,
  xlab="",ylab="Densita'",main="Densita' chi-quadrato per intervallo
  di confidenza 1-alpha=0.95")
> axis(1,c(z2-z2/4,z2,n-2,z1,z1+z1/4),c("",abbreviate(z2,minl=5),n-2,
  abbreviate(z1,minl=5),""))
> vals<-seq(z2-z2/5,z2, length =100)
> x1<-c(z2-z2/5,vals ,z2 ,0)
> y1<-c(0,dchisq(vals ,n) ,0,0)
> polygon (x1,y1,density =20,angle=45, col="blue")
> vals<-seq(z1,z1+z1/5, length =100)
> x1<-c(z1,vals ,z1+z1/5,1)
> y1<-c(0,dchisq(vals ,n) ,0,0)
> polygon (x1,y1,density =20,angle=45,col="blue")
> abline(h=0)
> text(z1+3.5,0.004,expression(alpha/2))
> text(z1+3.5,0.0066,"Regione di\n rifiuto")
> text(98,0.003,expression(1-alpha))
> text(98,0.0045,"Regione di accettazione")
> text(z2 -3.5,0.0066,"Regione di \n rifiuto")
> text(z2 -3.5,0.004,expression(alpha/2))
```




**Esempio di applicazione con** $1 - \alpha = 0.99$

Riprendiamo in considerazione il campione esaminato in precedenza e calcoliamo un intervallo di confidenza di grado $1 - \alpha = 0.99$. In questo caso $\alpha = 0.01$, $\alpha/2 = 0.005$ e $1 - \alpha/2 = 0.995$. I valori di $X^2_{1-\alpha/2,n-1} = X^2_{0.995,99}$ e $X^2_{\alpha/2,n-1} = X^2_{0.005,99}$ possono essere calcolati grazie ad R.

```
> alpha <- 0.01
> alpha_mezzi <- alpha/2
> varianzaCampionaria <- var(x)
> varianzaCampionaria
> z1<-qchisq(1-alpha_mezzi ,df=n-1)
> z2<-qchisq(alpha_mezzi ,df=n-1)
> lowC <- ((n-1)*varianzaCampionaria)/qchisq(1-alpha_mezzi ,df=n-1)
> highC <- ((n-1)*varianzaCampionaria)/qchisq(alpha_mezzi ,df=n-1)
> alpha
[1] 0.01
> alpha_mezzi
[1] 0.005
> varianzaCampionaria
[1] 109.0861
> z1
[1] 138.9868
> z2
[1] 66.51011
> lowC
[1] 77.70179
> highC
[1] 162.3742
```

La stima dell'intervallo di confidenza di grado $1 - \alpha = 0.95$ è $(77.70179, 162.3742)$ e la lunghezza di tale intervallo è pari a $162.3742 - 77.70179 = 84.67241$ che è maggiore rispetto a quello calcolato in precedenza. I risultati ottenuti possono essere rappresentati graficamente attraverso il codice R eseguito nell'esempio precedente.