Statistica e analisi dei dati A.A. 2018/19

Analisi di un campione di una distribuzione normale

Antonio Vivone

Distribuzione normale

La **funzione di distribuzione normale**, detta anche di Gauss o gaussiana, è molto importante nella statistica in quanto rappresenta una distribuzione limite alla quale tendono varie altre funzioni di distribuzioni utilizzando opportune ipotesi.

Una variabile aleatoria X di densità di probabilità

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

si dice avere distribuzione normale di parametri μ e σ.

La densità normale:

- risulta essere simmetrica rispetto all'asse $x = \mu$ in quanto per ogni $x \in \mathbb{R}$ risulta che $f_X(\mu - x) = f_X(\mu + x)$;
- presenta il massimo $(\sigma\sqrt{2\pi})^{-1}$ nel punto di ascissa $x = \mu$;
- presenta due flessi nei punti di ascisse $\mu - \sigma$ e $\mu + \sigma$.

Viene utilizzata la notazione $X \sim N(\mu, \sigma)$ per indicare che la variabile X ha distribuzione normale dei parametri μ e σ ed è chiamata *variabile normale*.

In R utilizziamo la funzione:

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

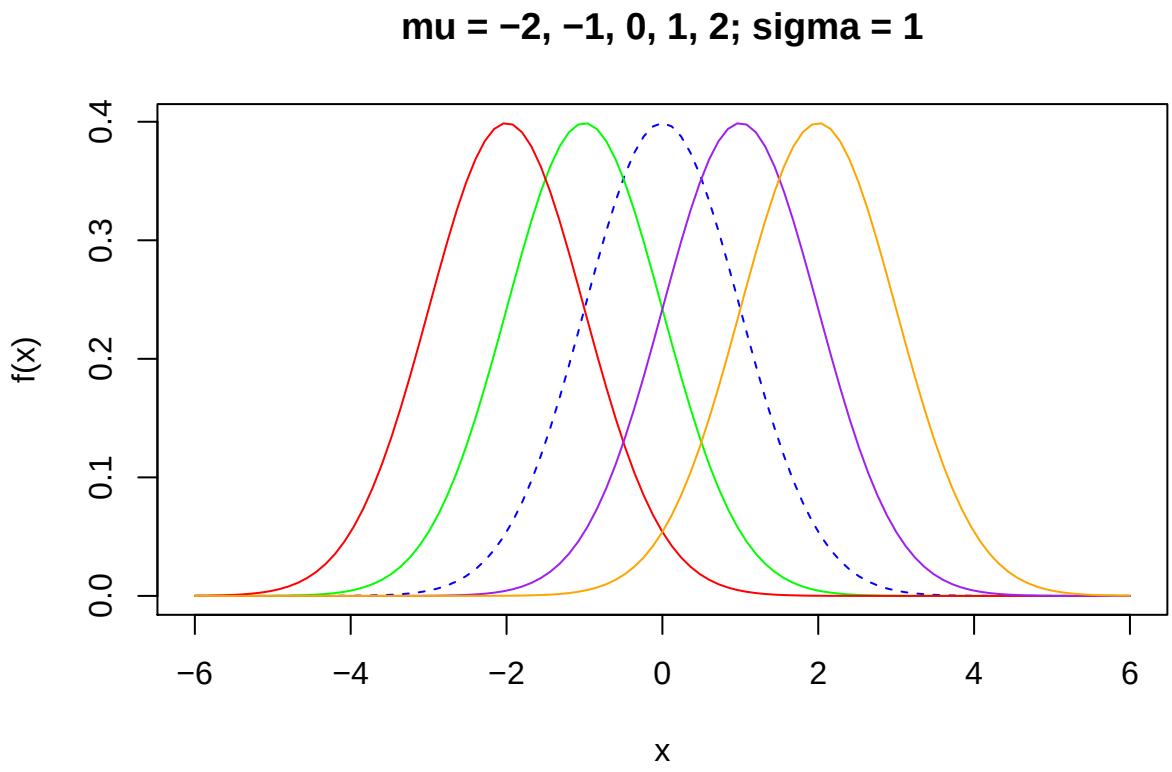
per calcolare la densità normale.

Quello che facciamo adesso è mostrare quello che accade quando i valori μ e σ vengono modificati. Procediamo quindi facendo variare μ e mantenendo σ fissato.

```
x <- seq(from = -7.5, to = 7.55, by = 0.1)

curve(dnorm(x, mean= 0, sd = 1), from = -6, to= 6, xlab = "x", ylab="f(x)",
      main="mu = -2, -1, 0, 1, 2; sigma = 1", col="blue", lty=2)
curve(dnorm(x, mean= -1, sd = 1), from = -6, to= 6, xlab = "x", ylab="f(x)",
      add = TRUE, col="green")
```

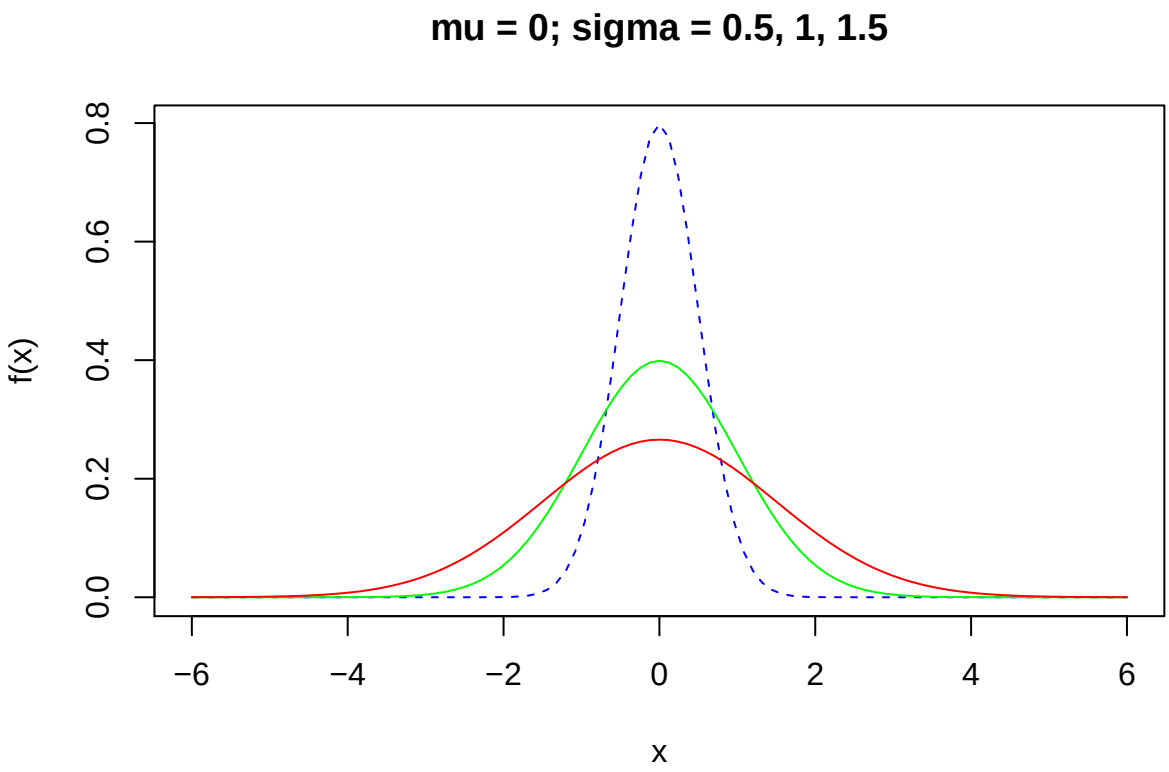
```
curve(dnorm(x, mean= -2, sd = 1), from = -6, to= 6, xlab = "x", ylab="f(x)",
      add = TRUE, col="red")
curve(dnorm(x, mean= 1, sd = 1), from = -6, to= 6, xlab = "x", ylab="f(x)",
      add = TRUE, col="purple")
curve(dnorm(x, mean= 2, sd = 1), from = -6, to= 6, xlab = "x", ylab="f(x)",
      add = TRUE, col="orange")
```



La curva di colore blu tratteggiata rappresenta la curva con $\mu = 0$. Facendo variare *mean*, la curva si sposta sull'asse delle ascisse mantenendo inalterato il suo valore.

Vediamo cosa succede modificando il valore *sd*.

```
curve(dnorm(x, mean= 0, sd = 0.5), from = -6, to= 6, xlab = "x", ylab="f(x)",
      main = "mu = 0; sigma = 0.5, 1, 1.5", col="blue", lty=2)
curve(dnorm(x, mean= 0, sd = 1), from = -6, to= 6, xlab = "x", ylab="f(x)",
      add = TRUE, col="green")
curve(dnorm(x, mean= 0, sd = 1.5), from = -6, to= 6, xlab = "x", ylab="f(x)",
      add = TRUE, col="red")
```



Facendo variare il parametro σ viene influenzata la larghezza della funzione: se infatti il parametro σ cresce allora l'ordinata massima decresce e la curva diventa sempre più piatta; se invece σ decresce allora l'ordinata massima cresce. Sono inversamente proporzionali in pratica.
Notasi che l'area al di sotto continuerà sempre ad avere valore unitario.

Funzione di distribuzione

La funzione di distribuzione di una variabile aleatoria $X \sim N(\mu, \sigma)$ è uguale a:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(y)dy = \Phi\Big(\frac{x-\mu}{\sigma}\Big) \qquad x \in \mathbb{R}$$

dove

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} exp\Big\{ -\frac{y^2}{2} dy \Big\}, \qquad z \in \mathbb{R}$$

è la funzione di distribuzione di una variabile aleatoria $Z \sim N(0, 1)$ detta *normale standard*.
Per questo, se $X \sim N(\mu, \sigma)$ si ha che:

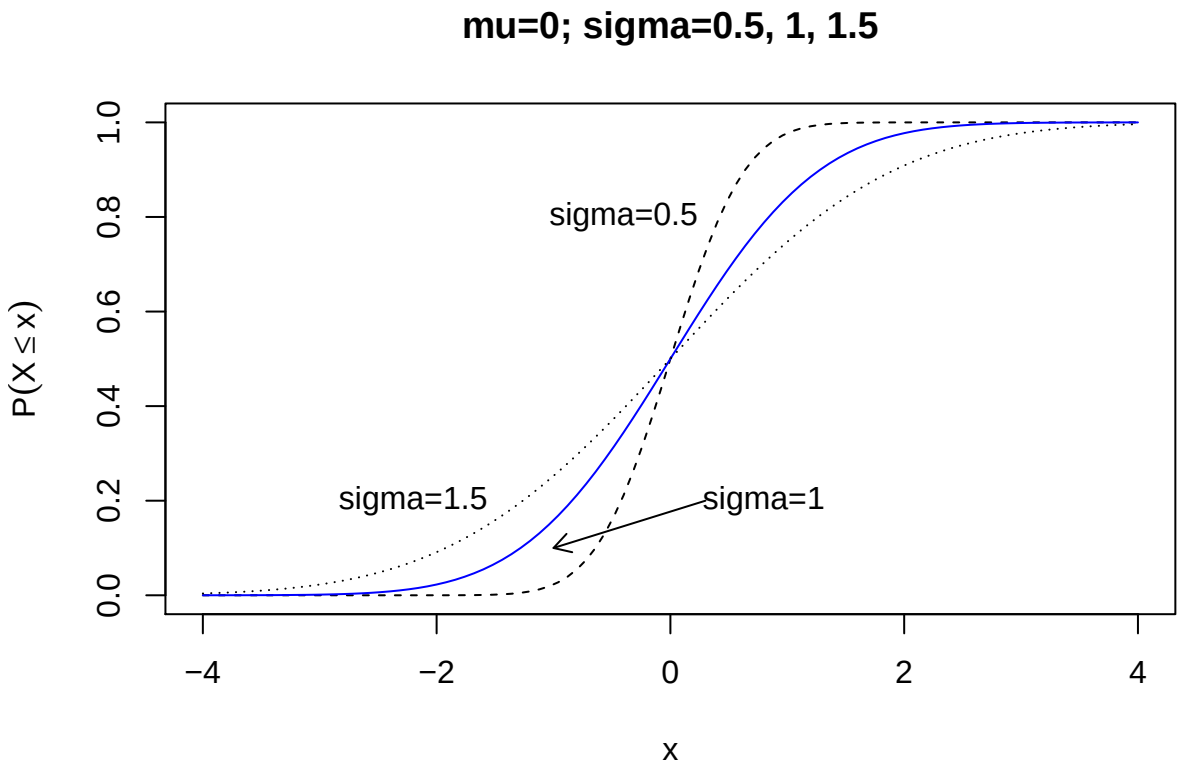
$$P(a < X < b) = F_X(b) - F_X(a) = \phi\Big(\frac{b-\mu}{\sigma}\Big) - \phi\Big(\frac{a-\mu}{\sigma}\Big)$$

In R calcoliamo la funzione di distribuzione di una variabile $X \sim N(\mu, \sigma)$ tramite la funzione:

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

Come fatto in precedenza con la densità, procediamo al confronto fra le funzioni di distribuzioni ottenute facendo variare il parametro σ.

```
curve(pnorm(x, mean = 0, sd = 0.5), from=-4, to=4, xlab = "x",
      ylab = expression(P(X<=x)),
      main="mu=0; sigma=0.5, 1, 1.5", lty=2)
text(-0.4, 0.8, "sigma=0.5")
curve(pnorm(x, mean=0, sd=1), add = TRUE, col="blue")
arrows(-1,0.1,0.3,0.2,code = 1,length = 0.10)
text(0.8, 0.2, "sigma=1")
curve(pnorm(x, mean=0, sd=1.5), add = TRUE, lty=3)
text(-2.2,0.2,"sigma=1.5")
```



Regola del 3-σ

Per una qualsiasi variabile aleatoria normale $X \sim N(\mu, \sigma)$ risulta che:

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = P(-3 < \frac{X - \mu}{\sigma} < 3) = P(-3 < Z < 3) = 0.9973002$$

Quello che la regola vuol dire è che la probabilità che una variabile aleatoria $X \sim N(\mu, \sigma)$ assuma valori in un intervallo avente come centro μ e semiampiezza 3σ è mlto vicino all'unità, ovvero ad 1. Questa proprietà delle variabili aleatorie normali è detta **regola del 3σ**. Utilizzando la funzione *pnorm()* possiamo mostrare quanto detto:

```
pnorm(3,mean = 0, sd = 1) - pnorm(-3, mean = 0, sd = 1)
```

```
## [1] 0.9973002
```

Quantili

È possibile anche calcolare i quantili (percentili) della distribuzione normale attraverso la funzione:

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

La funzione restituisce il percentile $z \cdot 100 - esimo$, ovvero il più piccolo numero x assunto dalla variabile aleatoria normale X tale che $P(X \leq x) \geq z$.

Considerando ad esempio una variabile normale standard $Z \sim N(0, 1)$, è possibile ottenere i quartili nella seguente maniera:

```
z<-c(0,0.25,0.5,0.75,1)
qnorm(z,mean=0, sd=1)
```

```
## [1]       -Inf -0.6744898  0.0000000  0.6744898        Inf
```

Fatto interessanta da notare è la simmetria fra Q_1 e Q_3 dovuto alla simmetria intorno all'origine della densità normale standard.

Approssimare la distribuzione binomiale con la distribuzione normale

Siccome il calcolo delle probabilità binomiali risulta aumentare di complessità al crescere di n, sono state ricercate formule in grado di approssimare tale distribuzione con quella normale. Vediamo i due metodi proposti:

Teorema di De Moivre-Laplace

Sia $X_1, X_2, ...$ una successione di variabili aleatorie indipendenti distribuite alla Bernoulli con parametro p $(0 < p < 1)$, e sia $Y_n = X_1 + X_2 + ... + X_n$. Allora per ogni $x \in \mathbb{R}$ risulta che:

$$\lim_{n\to\infty} P\left(\frac{Y_n - np}{\sqrt{np(1-p)}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$$

ovvero

$$\frac{Y_n - np}{\sqrt{np(1-p)}} \to Z$$

converge in distribuzione alla variabile aleatoria Z normale standard.

Se le variabili $X_1 + X_2 + ...$ sono variabili aleatorie di Bernoulli di parametro p, allora $Y_n = X_1 + X_2 + ... + X_n$ è una variabile aleatoria binomiale di valore medio np e varianza $np(1-p)$. Quello che il teorema fa è mostrare che sottraendo a Y_n la sua media np e dividendo la differenza per la deviazione standard $\sqrt{np(1-p)}$, si ottiene una variabile aleatoria standardizzata la cui funzione di distribuzione è per n grandi una normale standard approssimata. L'approssimazione della binomiale alla normale è la seguente:

$$Y_n \simeq np + \sqrt{np(1-p)}Z$$

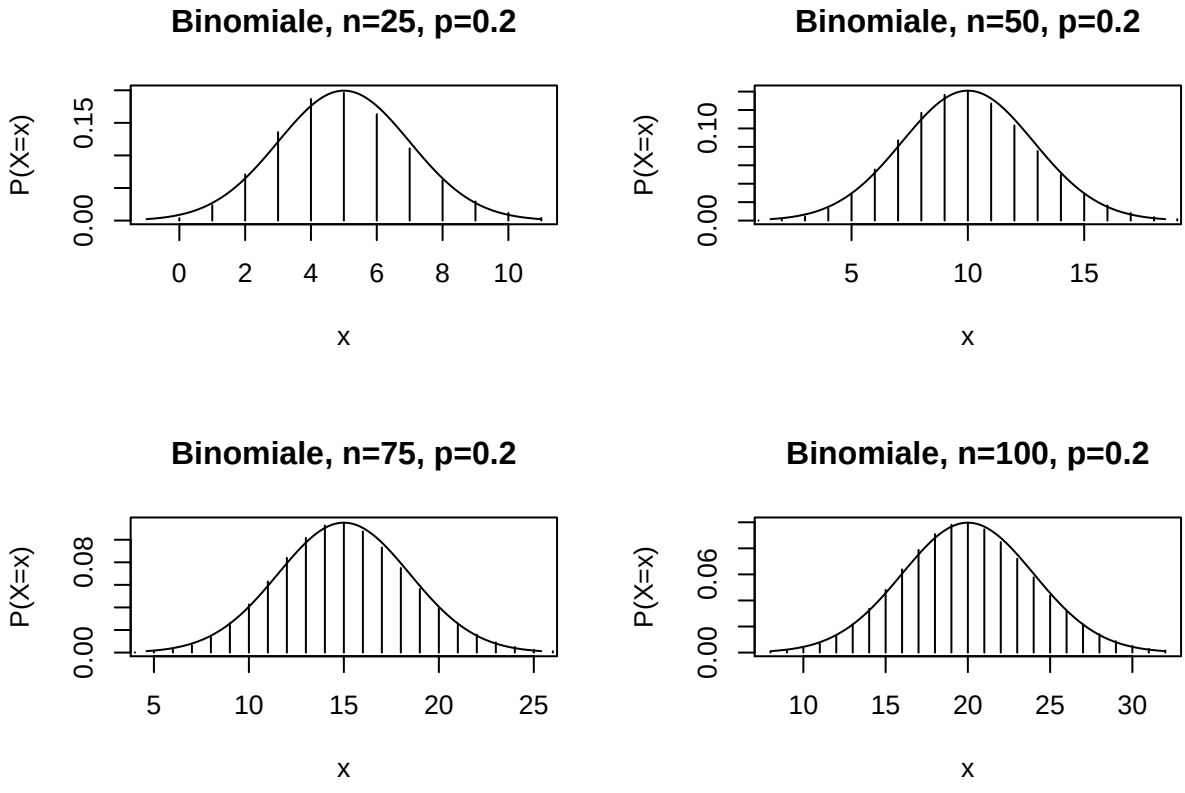
al variare di n con p fissato. La variabile aleatoria con densità normale ottenuta ha valore medio np e varianza $np(1-p)$.

È possibile valutare l'approssimazione della binomiale ottenuta confrontandola con la densità normale di valore np e varianza$np(1-p)$ per $n = 25, 50, 75, 100$ e $p = 0.2$.

```
par(mfrow=c(2,2))
p<-0.2
q<-1-p
x<-0:25
n<-25
curve(dnorm(x, n*p, sqrt(n*p*q)), from=n*p-3*sqrt(n*p*q),
      to=n*p+3*sqrt(n*p*q), xlab="x", ylab="P(X=x)",
      main="Binomiale, n=25, p=0.2")
lines(x, dbinom(x,n,0.2), type="h")
x<-0:50
n<-50
curve(dnorm(x, n*p, sqrt(n*p*q)), from=n*p-3*sqrt(n*p*q),
      to=n*p+3*sqrt(n*p*q), xlab="x", ylab="P(X=x)",
      main="Binomiale, n=50, p=0.2")
lines(x, dbinom(x,n,0.2), type="h")
```

```r
x<-0:75
n<-75
curve(dnorm(x, n*p, sqrt(n*p*q)), from=n*p-3*sqrt(n*p*q),
      to=n*p+3*sqrt(n*p*q), xlab="x", ylab="P(X=x)",
      main="Binomiale, n=75, p=0.2")
lines(x, dbinom(x,n,0.2), type="h")

x<-0:100
n<-100
curve(dnorm(x, n*p, sqrt(n*p*q)), from=n*p-3*sqrt(n*p*q),
      to=n*p+3*sqrt(n*p*q),xlab="x", ylab="P(X=x)",
      main="Binomiale, n=100, p=0.2")
lines(x, dbinom(x,n,0.2), type="h")
```



Come è possibile vedere, l'approssimazione migliora quando p tende a $\frac{1}{2}$ e diventa eccellente quando p=$\frac{1}{2}$

Teorema centrale di convergenza

Il **teorema centrale di convergenza** fornisce un'approssimazione alla distribuzione della somma di variabili aleatorie indipendenti, evidenziando allo stesso tempo l'importanza della distribuzione normale.

Sia $X_1, X_2, ...$ una successione di variabili aleatorie, definite nello stesso spazio di probabilità, indipendenti e identicamente distribuite con valore medio μ finito e varianza σ^2 finita e positiva. Posto per ogni intero n positivo $Y_n = X_1 + X_2 + ... + X_n$, per ogni $x \in \mathbb{R}$ risulta:

$$\lim_{n\to\infty} P\left(\frac{Y_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy = \Phi(x)$$

ovvero

$$\frac{Y_n - E(Y_n)}{\sqrt{Var(Y_n)}} = \frac{Y_n - n\mu}{\sigma\sqrt{n}} \to Z$$

Quello che il teorema mostra è che sottraendo a Y_n la sua media e dividendo il tutto per la sua deviazione standard, ovvero $\sigma\sqrt{n}$ si ottiene una variabile aleatoria standardizzata la cui funzione di distribuzione per **n sufficientemente grandi** è approssimativamente una normale standard con valore $n\mu$ e varianza $n\sigma^2$. Quindi la bontà dell'approssimazione dipende dal tipo di distribuzione delle variabili $X_1, X_2, ..., X_n$ e dalle dimensioni di n. Spesso l'approssimazione risulta soddisfacente già con $n \geq 30$.

Simulare una variabile in R

È possibile simulare in R una variabile aleatoria normale generando una sequenza di numeri pseudocasuali mediante la funzione:

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

Confrontiamo quindi la densità normale teorica con la densità simulata. Vediamo la densità normale con $\mu = 2$ e $\sigma = 1$ con la densità simulata con $N = 500, 5000, 50000$

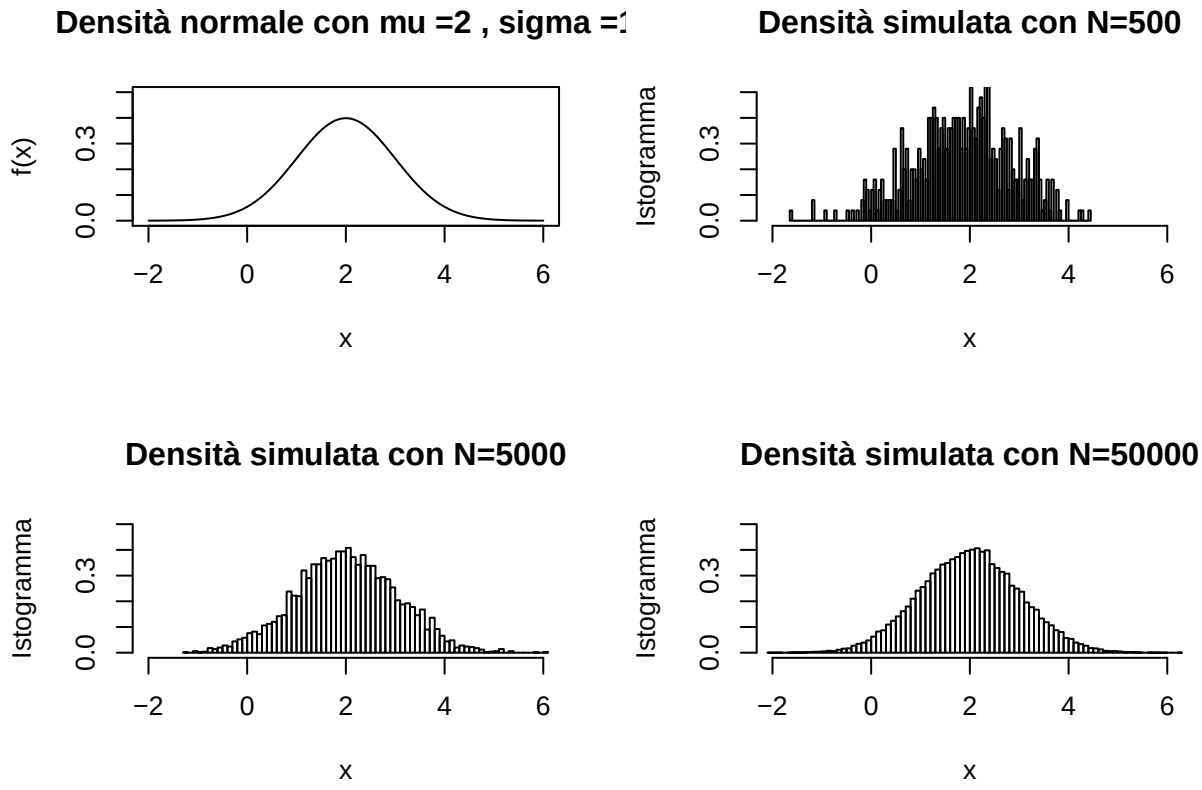
```
par ( mfrow = c (2 ,2) )
curve(dnorm(x,mean=2,sd =1),from=-2, to=6 , xlab="x", ylab="f(x)",
ylim = c (0 ,0.5) , main = " Densità normale con mu =2 , sigma =1 " )

sim1<-rnorm(500,mean=2,sd =1)
hist ( sim1 , freq =F , xlim = c ( -2 ,6) , ylim = c (0 ,0.5) ,
       breaks =100 , xlab = " x " ,
       ylab = " Istogramma " , main = " Densità simulata con N=500 " )

sim2 <- rnorm (5000 , mean =2 , sd =1)
hist ( sim2 , freq =F , xlim = c ( -2 ,6) , ylim = c (0 ,0.5) ,
       breaks =100 , xlab = " x " ,
       ylab = " Istogramma " , main = " Densità simulata con N=5000 " )

sim3 <- rnorm (50000 , mean =2 , sd =1)
```

```
hist ( sim3 , freq =F , xlim = c ( -2 ,6) , ylim = c (0 ,0.5) ,
       breaks =100 , xlab = " x " ,
       ylab = "Istogramma" , main = " Densità simulata con N=50000 " )
```



All'aumentare della numerosità, l'istogramma della densità simultata si avvicina sempre di più alla curva teorica.

Analisi di un campione normale

Procediamo adesso a generare un campione facente parte di una popolazione normale in modo tale da effettuare diverse analisi:

```
campione <- rnorm(10000, mean = 110, sd = 1)
```

Riportiamo di seguito la media campionaria, la varianza campionaria e la deviazione standard campionaria del campione:

```
## [1] 110.01
```

```
## [1] 0.9837746
```

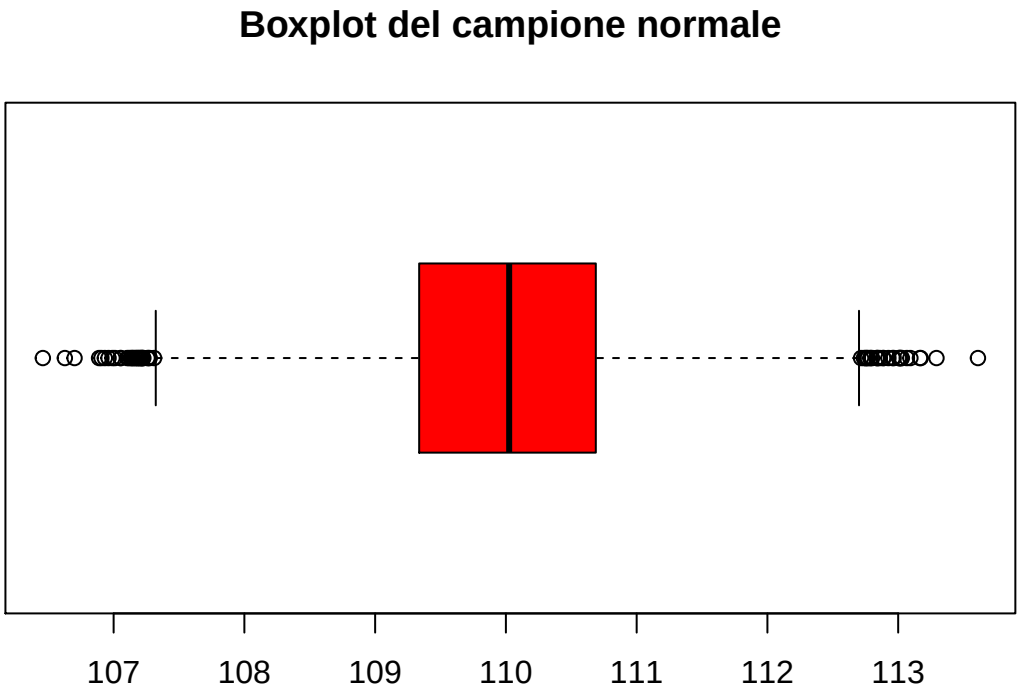
```
## [1] 0.9918541
```

Calcoliamo i quantili del campione e generiamo il boxplot in modo da capire come sono distribuiti i valori:

```
quantile(campione)
```

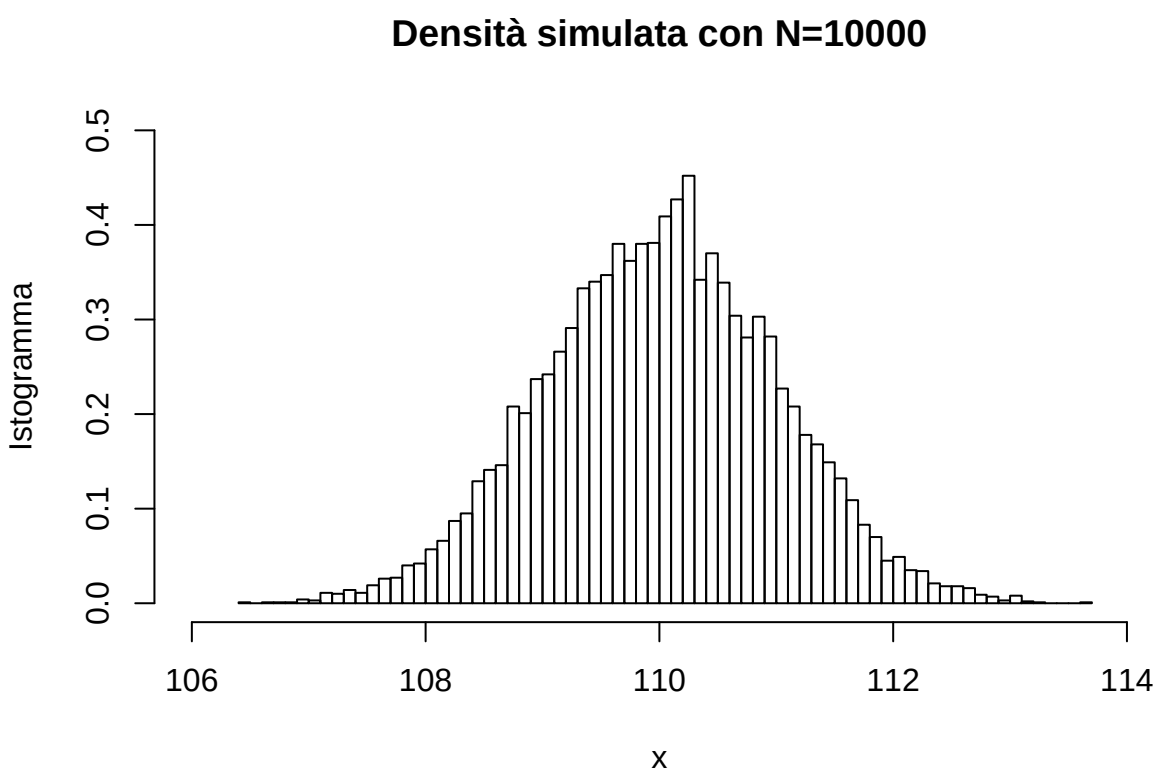
```
##       0%      25%      50%      75%     100%
## 106.4587 109.3368 110.0250 110.6875 113.6101
```

```
boxplot(campione, horizontal = T, main="Boxplot del campione normale", col="red")
```



Dal boxplot capiamo che il campione risulta simmetrico e centrato intorno al valore 2, come stabilito. Effettuiamo un'ultima analisi mostrando la densità tramite istogramma:

```
hist (campione , freq =F , xlim = c (106,114) , ylim = c (0 ,0.5) , breaks =100 ,
      xlab = " x " ,ylab = "Istogramma" , main = " Densità simulata con N=10000 ")
```



La forma della densità del campione è decisamente simile a quella della normale teorica, grazie soprattutto alla scelta del campione molto grande.

Passiamo adesso alla stima dei parametri.

Stima puntuale

Campioni casuali e stimatori

Uno dei problemi della inferenza statistica è quello di ottenere informazioni su parametri non noti di una popolazioni di cui si conosce però la forma della funzione di distribuzione.

Per ottenere informazioni sui parametri non noti è possibile fare uso dell'inferenza statistica e considerare un campione estratto dalla popolazione. Su questo campione poi utilizziamo alcune variabili aleatorie, ovvero funzioni misurabili del campione casuale, dette **statistiche** e **stimatori**. La definizione di stimatore è la seguente:

Uno stimatore $\hat{\theta} = t(X_1, X_2, ..., X_n)$ è una funzione misurabile e osservabile del campione casuale $(X_1, X_2, ..., X_n)$ i cui valori vengono utilizzati per stimare un parametro non noto ϑ della popolazione. I valori $\hat{\vartheta}$ assunti da tale stimatore sono detti stime del parametro non noto ϑ.

Stimatori tipici sono *media campionaria* e *varianza campionaria*.

Proposizione: Sia $X_1, X_2, ..., X_n$ un campione casuale estratto da una popolazione descritta da una variabile aleatoria ossevabile X caratterizzata da valore medio $E(X) = \mu$ finito e varianza $Var(X) = \sigma^2$ finita. Risulta:

$$E(\bar{X}) = \mu, \qquad Var(\bar{X}) = \frac{\sigma^2}{n}$$

Per la proprietà di linearità del valore medio e l'identica distribuzione delle variabili aleatorie che costituiscono il campione della proposizione sopra descritta, si ha:

$$E(\bar{X}) = E\Big[\frac{1}{n}\sum_{i=1}^{n} X_i\Big] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu$$

e per la varianza:

$$Var(\bar{X}) = Var\Big[\frac{1}{n}\sum_{i=1}^{n} X_i\Big] = \frac{1}{n^2 2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}$$

Questa proposizione ci dice che al crescere dell'ampiezza del campione, la media campionaria fornisce una stima sempre più accurata del valore medio della popolazione. Dal teorema centrale di convergenza sappiamo che per n sufficientemente grandi, la funzione di distribuzione della media campionaria $\bar{X}$ è approssimativamente normale con valore medio μ e varianza σ^2/n.

Metodo per la ricerca di stimatori

Esistono due metodi maggiormente utilizzati per la ricerca di stimatori: il **metodo dei momenti** e il **metodo della massima verosomiglianza**.

Metodo dei momenti

Prima di parlare del metodo dei momenti è necessario introdurre i **momenti campionari**.

Si definisce **momento campionario r-esimo** relativo ai valori osservati (x_1, x_2,..., x_n) del campione casuale il valore

$$M_r(x_1, x_2, ..., x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^r \qquad (r = 1, 2, ...)$$

Il momento campionario r-esimo risulta essere quindi la media aritmetica delle potenze r-esime delle n osservazioni effettuate sulla popolazione. In particolare, se $r = 1$ otteniamo la media campionaria $\bar{X}$.

Se esistono k parametri da stimare, il **metodo dei momenti** consiste nell'uguagliare i primi k momenti della popolazione in esame con quelli del campione casuale. Se i k momenti esistono e sono finiti il metodo consiste nel risolvere il sistema di equazioni:

$$E(X^r) = M_r(x_1, x_2, ..., x_n) \qquad (r = 1, 2, ..., k)$$

Le incognite sono i k parametri non noti $\vartheta_1, \vartheta_2, ..., \vartheta_k$. Per poter utilizzare questo metodo, il sistema deve ammettere un'unica soluzione. Le stime dipendono dal campione osservato e quindi, al variare del campione, cambiano anche queste ultime. Gli stimatori prendono il nome di *stimatori del metodo dei momenti*.

Procediamo quindi ad utilizzare il metodo alla nostra analisi della popolazione normale. Siamo quindi interessati a fornire degli stimatori dei parametri μ e σ^2. Siccome $E(X) = \mu$ e $E(X^2) = \sigma^2 + \mu^2$ si ha un sistema di due equazioni, in quanto due sono i parametri da stimare. Si ha che:

$$\hat{\mu} = \frac{x_1 + x_2 + ... + x_n}{n}; \qquad \hat{\sigma}^2 + \hat{\mu}^2 = \frac{(x_1 + x_2 + ... + x_n)^2}{n}$$

dalla seconda equazione ricaviamo:

$$\hat{\sigma}^2 = \frac{(x_1 + x_2 + ... + x_n)^2}{n} - \frac{(x_1 + x_2 + ... + x_n)^2}{n^2} = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Il metodo dei momenti ci fornisce quindi come stimatore del valore medio $/mu$ la media campionaria $\bar{X}$ e come stimatore della varianza σ^2 la variabile aleatoria $\frac{(n-1)}{n}S^2$.

Per quanto riguarda il nostro campione quindi abbiamo che:

```
stimaMeanMomenti <- mean(campione)
stimaMeanMomenti
```

```
## [1] 110.01
```

```
stimaVarMomenti <- (length(campione)-1) * var(campione)/length(campione)
stimaVarMomenti
```

```
## [1] 0.9836762
```

Metodo della massima verosomiglianza

Il **metodo della massima verosomiglianza** è il metodo più importante fra i due ed è anche preferito a quello dei momenti. Prima di illustrare il metodo va introdotta la **funzione di verosomiglianza**.

Sia $X_1, X_2, ..., X_n$ un campione casuale estratto dalla popolazione. La **funzione di verosomiglianza** $L(\vartheta_1, \vartheta_2, ..., \vartheta_k) = L(\vartheta_1, \vartheta_2, ..., \vartheta_k; x_1, x_2, ..., x_n)$ del campione osservato è la funzione di densità di probabilità congiunta del campione casuale $X_1, X_2, ..., X_n$ ovvero:

$$L(\vartheta_1, \vartheta_2, ..., \vartheta_k) = L(\vartheta_1, \vartheta_2, ..., \vartheta_k; x_1, x_2, ..., x_n) = f(x_1 : \vartheta_1, \vartheta_2, ..., \vartheta_k) f(x_2 : \vartheta_1, \vartheta_2, ..., \vartheta_k) \quad ... \quad f(x_n : \vartheta_1, \vartheta$$

Il metodo consiste quindi nel massimizzare la funzione di verosomiglianza rispetto ai parametri $\vartheta_1, \vartheta_2, ..., \vartheta_k$ cercando di determinare da quale funzione di densità di probabilità congiunta è **più verosimile** che provenga il campione osservato. Si cercando quindi i valori $\vartheta_1, \vartheta_2, ..., \vartheta_k$ che massimizzino la funzione di verosomiglianza e, una volta trovati, vengono indicati nella seguente maniera $\hat{\vartheta}_1, \hat{\vartheta}_2, ..., \hat{\vartheta}_k$ e costituiscono le **stime di massima verosomiglianza**. Siccome queste stime dipendono dal campione, facendo variare il campione osservato si ottengono gli stimatori di massima verosomiglianza $\bar{\Theta}_1, \bar{\Theta}_2, ..., \bar{\Theta}_k$ dei parametri non noti $\vartheta_1, \vartheta_2, ..., \vartheta_k$ della popolazione, detti **stimatori di massima verosomiglianza**.

Per quanto riguarda la popolazione normale da noi presa in esame abbiamo:

$$\hat{\mu} = \bar{x}, \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Quindi lo stimatore di massima verosomiglianza di μ risulta essere la media campionaria di $\bar{X}$, mentre lo stimatore di σ^2 risulta essere pari a $\frac{(n-1)}{n} S^2$. Gli stimatori ottenuti coincidono con quelli ottenuti con il metodo dei momenti, quindi eviteremo di ripetere i calcoli.

Proprietà degli stimatori

Data la presenza di diversi stimatori in grado di stimare un parametro non noto di una popolazione occorre definire alcune proprietà che li caratterizzano. Queste proprietà sono:

- corretto,
- più efficiente di un altro,
- corretto e con varianza uniformamente minima,
- asintoticamente corretto
- consistente.

Uno simatore è **corretto** se il suo valore medio è uguale al corrispondente parametro non nodo della popolazione. È possibile avere più stimatori corretti per stimare un parametro. Per scegliere quale utilizzare, si vede quale risulta essere più efficiente confrontando la varianza degli stimatori e scegliendo quello con varianza più piccola. Altro metodo è quello della ricerca dello stimatore con **errore quadratico uniformemente minimo** per la classe degli stimatori corretti.

Come visto prima, sia dal metodo dei momenti che da quello della massima verosomiglianza sono stati ricavati gli stessi stimatori per i parametri μ e σ^2. Per μ lo stimatore ricavato è la media campionaria, che risulta essere **corretto con varianza minima** e **consistente**, mentre per σ^2 abbiamo $\frac{n-1}{n} S^2$ che è **asintoticmanete corretto** e **consistente**. $\underline{C}_n$ e $\bar{C}_n$

Stima intervallare

Spesso si preferisce sostituire alla stima puntale di un parametro non noto un intervallo di valori, chiamato **intervallo di confidenza**, entro il quale sia compreso il valore del parametro non noto con un certo **grado di fiducia** o **coefficiente di confidenza**. Supponiamo di avere un campione casuale con una densità di probabilità o funzione di probabilità uguale a $f(x;\vartheta)$ dove ϑ rappresenta il parametro non noto della popolazione. Per definire questo intervallo, ci occorre denotare due statistiche, $\underline{C}_n$ e $\bar{C}_n$ che soddisfino la seguente condizione $\underline{C}_n < \bar{C}_n$. Fissato un coefficiente di confidenza $1-\alpha$ $(0<\alpha<1)$, se è possibile scegliere le statistiche $\underline{C}_n$ e $\bar{C}_n$ in modo tale che

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

allora si dice che $(\underline{C}_n, \bar{C}_n)$ è un **intervallo di confidenza** di grado $1-\alpha$ per ϑ. Le statistiche $\underline{C}_n$ e $\bar{C}_n$ sono dette rispettivamente *limite inferiore* e *limite superiore*. La stima puntale detta in precedenza deve ricadere in questo intervallo.

Vediamo adesso i metodi per la costruzione di questo intervallo di confidenza.

Metodo pivotale

Il **metodo pivotale** è uno dei metodi a disposizione per la costruzione degli intervalli di confidenza. Esso consiste nel determinare una variabile aleatoria $\gamma(X_1, X_2, ..., X_n; \vartheta)$ che dipende dal campione casuale e dal parametro non noto ϑ la cui **funzione di distribuzione non contiene il parametro da stimare**. Questa variabile aleatoria non è statistica in quanto dipende dal parametro non noto ed è quindi non osservabile.

Per ogni fissato coefficiente α $(0<\alpha<1)$ siano α_1 e α_2 $(\alpha_1<\alpha_2)$ due valori dipendenti soltanto dal coefficiente fissato α tali che per ogni $\vartheta \in \Theta$ si abbia:

$$P(\alpha_1 < \gamma(X_1, X_2, ..., X_n; \vartheta) < \alpha_2) = 1-\alpha$$

Se per ogni campione osservato $(x_1, x_2, ..., x_n)$ e per ogni parametro non noto $\vartheta \in \Theta$ si riesce a dimostrare che:

$$\alpha_1 < \gamma(x:\vartheta) < \alpha_2 \iff g_1(x) < \vartheta < g_2(x)$$

con g_1 e g_2 dipendenti dal campione osservato allora è possibile riscrivere la probabilità sopra descritta come:

$$P(g_1(X_1, X_2, ..., X_n) < \vartheta < g_2(X_1, X_2, ..., X_n)) = 1-\alpha$$

denotando con $\underline{C}_n = g_1(X_1, X_2, ..., X_n)$ e $\bar{C}_n = g_2(X_1, X_2, ..., X_n)$ deduciamo che $(\underline{C}_n, \bar{C}_n)$ è un intervallo di confidenza di grado $1-\alpha$ per ϑ.

Per una popolazione normale è possibile analizzare i seguenti problemi:

1. Determinare un intervallo di confidenza di grado 1 - α per il valore medio μ nel caso in cui la varianza σ^2 della popolazione normale è nota.

Intervallo di confidenza per μ con σ^2 nota

Per determinare un intervallo di confidenza $\alpha - 1$ per il valore medio μ nel caso in cui la varianza σ^2 della popolazione normale è nota, utilizziamo il metodo pivotale visto primo e consideriamo la variabile aleatoria standardizzata di valore medio nullo e varianza unitaria

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

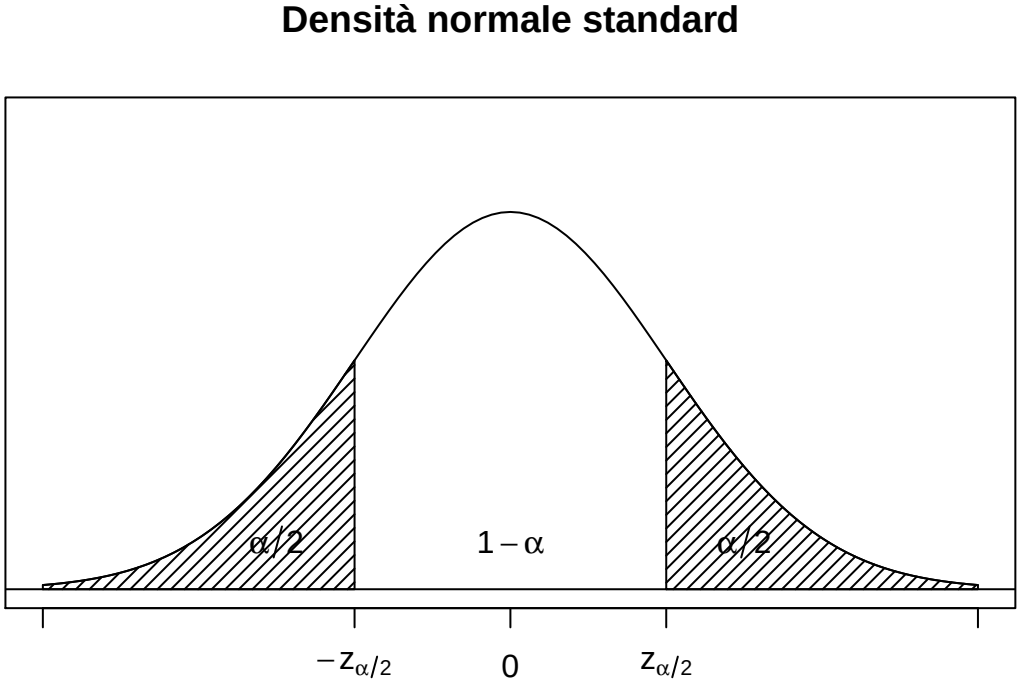
Questa variabile aleatoria risulta essere una normale standard che dipende dal campione casuale e dal parametro non noto μ e quindi è possibile utilizzarla come variabile di pivot. Scegliendo nel metodo pivotale $\alpha_1 = -z_{a/2}$ e $\alpha_2 = z_{a/2}$ è tale che

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

si ha

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

Graficamente abbiamo:



L'intervallo di confidenza con grado di fiducia $1-\alpha$ per il valore medio μ è pari a

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

Per quanto riguarda la nostra analisi, vogliamo ottenere un grado di fiducia pari a 0.95 quindi poniamo $\alpha = 0.05$ e supponiamo che la varianza

```
alpha <- 1-0.95

n<-length(campione)

#limite inferiore
mean(campione)-qnorm(1-alpha/2, mean=0, sd=1)*8/sqrt(n)
```

```
## [1] 109.8532
```

```
#limite superiore
mean(campione)+qnorm(1-alpha/2, mean=0, sd=1)*8/sqrt(n)
```

```
## [1] 110.1668
```

2. Determinare un intervallo di confidenza di grado $1-\alpha$ per il valore medio μ nel caso in cui la varianza della popolazione normale è non nota

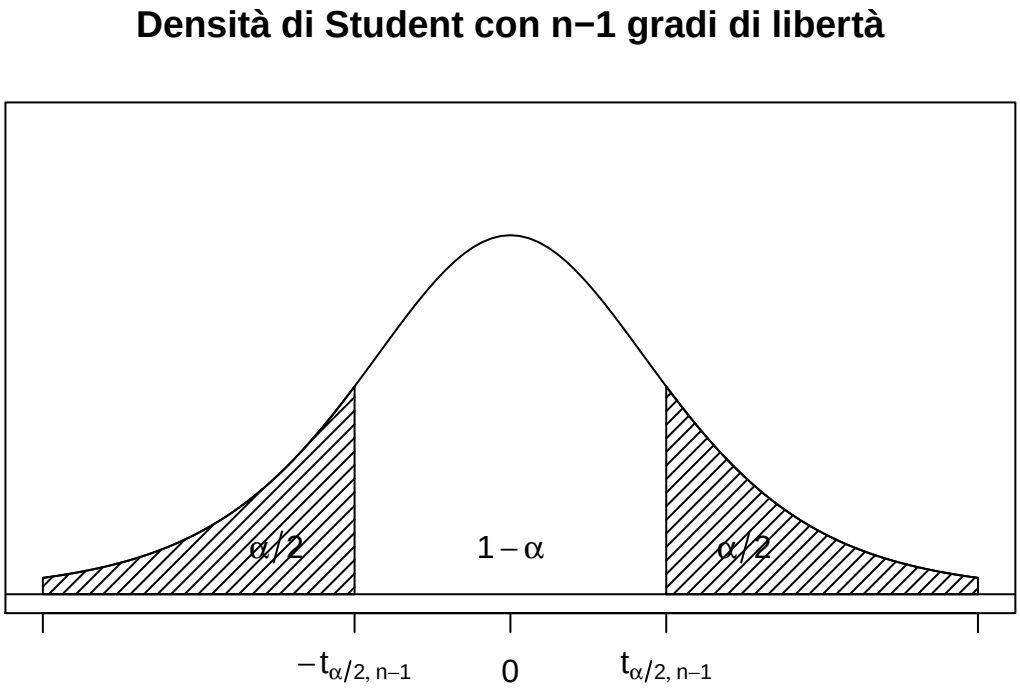
Intervallo di confidenza per μ con varianza σ^2 non nota

Per determinare un intervallo di confidenza in questo caso, quello che facciamo è utilizzare il metodo pivotale e considerare la variabile aleatoria di pivot pari a:

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

Questa variabile dipende dal campione casuale e dal parametro non noto μ e può essere quindi interpretata come una variabile aleatoria di pivot. La variabile in questione risulta anche essere distribuita con la legge di Student con $n-1$ gradi di libertà. Scegliendo nel metodo pivotale $\alpha_1 = -t_{\alpha/2,n-1}$ e $\alpha_2 = t_{\alpha/2,n-1}$ dove $t_{\alpha/2,n-1}$ abbiamo che:

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



abbiamo:

Una stima dell'intervallo di confidenza $1-\alpha$ per il valore medio μ risulta essere:

$$\bar{x_n} - t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}} < \mu < \bar{x_n} + t_{\alpha/2,n-1}\frac{s_n}{\sqrt{n}}$$

Ponendo come nel caso precedente vogliamo ottenere un grado di fiducia pari a 95 quindi poniamo $\alpha = 0.05$ e andiamo a stimare l'intervallo di confidenza:

```
alpha <- 1-0.95

ds <- sd(campione)
n <- length(campione)

#stima limite inferiore
mean(campione)-qt(1-alpha/2, df=n-1)*ds/sqrt(n)
```

```
## [1] 109.9906
```

```
#stima limite superiore
mean(campione)+qt(1-alpha/2, df=n-1)*ds/sqrt(n)
```

```
## [1] 110.0295
```

3. Determinare un intervallo di confidenza di grado $1-\alpha$ per la varianza nota σ^2 nel caso in cui il valore medio della popolazione normale μ risulti essere noto.

Intervallo di confidenza per σ^2 con μ noto

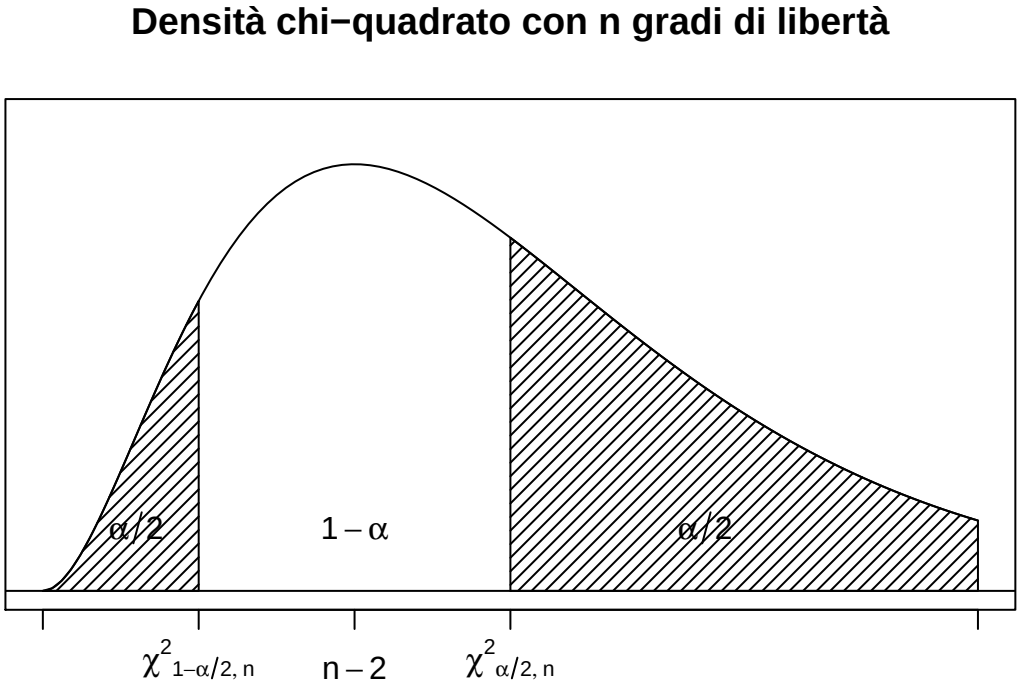
Come nei casi precedenti, anche qui utilizziamo il metodo pivotale e consideriamo come variabile di pivot la seguente:

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

Questa variabile dipende dal campione casuale e dal parametro non noto σ^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. Nel metodo pivotale, scegliamo $\alpha_1 = \chi^2_{1-\alpha/2,n}$ e $\alpha_2 = \chi^2_{\alpha/2,n}$ avendo così:

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

graficamente:



Una stima dell'intervallo di confidenza $1-\alpha$ per σ^2 è:

$$\frac{(n-1)s_n^2 + n(\overline{X}_n - \mu)^2}{\chi^2_{\alpha/2,n}} < \sigma^2 < \frac{(n-1)s_n^2 + n(\overline{X}_n - \mu)^2}{\chi^2_{1-\alpha/2,n}}$$

Il grado di fiducia che vogliamo ottenere è pari a 95, quindi poniamo $\alpha = 0.05$ e supponiamo che la media nota sia $\mu = 2$, stimiamo adesso il parametro:

```
alpha <- 1-0.95
n <-length(campione)
mu <- 2

#limite inferiore
((n-1)*var (campione)+n*(mean(campione)-mu)**2)/qchisq(1- alpha/2,df=n)
```

```
## [1] 11350.39
```

```
#limite superiore
((n-1)*var (campione)+n*(mean(campione)-mu)**2)/qchisq(alpha/2,df=n)
```

```
## [1] 11997.42
```

4. Determinare un intervallo di confidenza di grado $1-\alpha$ per la varianza nota σ^2 nel caso in cui il valore medio della popolazione normale non è noto.

Intervallo di confidenza per σ^2 con valore medio μ non noto

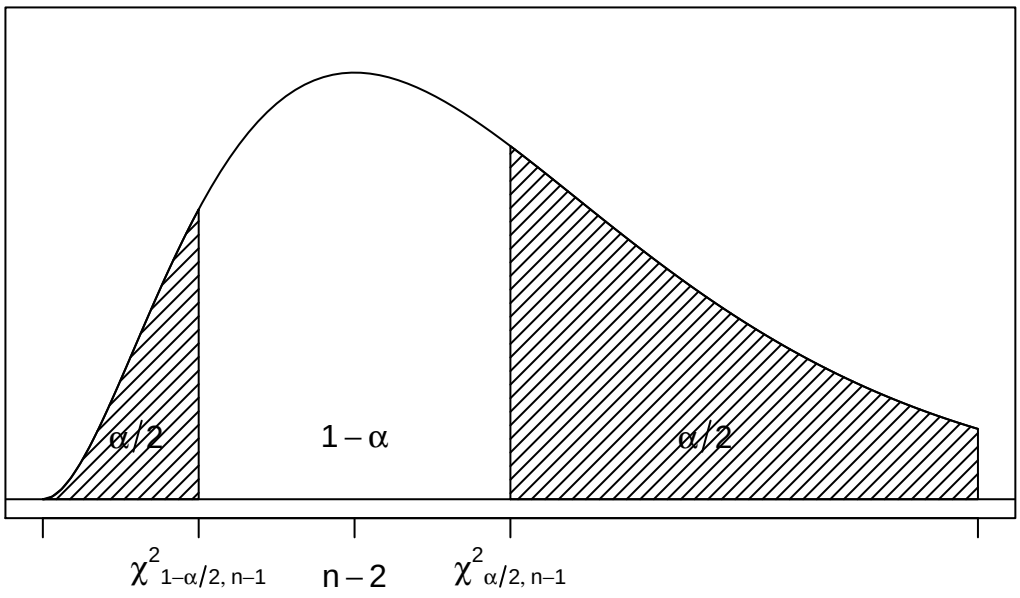
Siccome qui non abbiamo a disposizione la media, utilizziamo la media campionaria. La variabile aleatoria di pivot utilizzata è la seguente:

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

Questa variabile dipende dal campione casuale e dal parametro non noto σ^2 ed è distribuita con legge chi-quadrato con n-1 gradi di libertà. Poniamo $\alpha_1 = \chi^2_{1-\alpha/2,n}$ e $\alpha_2 = \chi^2_{\alpha/2,n}$ e abbiamo che:

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

graficamente:



Una stima dell'intervallo di confidenza per $1-\alpha$ per σ^2 è:

$$\frac{(n-1)s_n^2}{\chi^2_{1-\alpha/2,n}} < \sigma^2 < \frac{(n-1)s_n^2}{\chi^2_{\alpha/2,n}}$$

Il grado di fiducia che vogliamo ottenere è pari a 0.95 quindi poniamo $\alpha = 0.05$ e stimiamo così σ^2:

```
alpha <- 1-0.95
n <- length(campione)

#stima del limite inferiore
(n-1)*var(campione)/qchisq (1- alpha/2,df=n-1)
```

```
## [1] 0.9570644
```

```
#stima del limite superiore
(n-1)*var(campione)/qchisq (alpha/2,df=n-1)
```

```
## [1] 1.011624
```

In conclusione sulla stima, per la popolazione normale le stime per intervallo del valore medio μ e della varianza σ^2 possono essere effettuate **qualsiasi sia la dimensione del campione casuale osservato**, questo grazie al fatto che conosciamo la distribuzione esatta della variabile pivotale usata: normale, di Student e chi-quadrato.

I casi interessanti che rappresentano davvero situazioni reali sono il secondo e il quarto, dove nel secondo ricordiamo che vogliamo stimare il valore medio μ con varianza σ^2 non nota e nel quarto dove vogliamo stimare la varianza σ^2 con μ non noto.

Differenza tra valori medi

Diverse problematiche richiedono di confrontare i valori medi di due popolazioni quindi vediamo come determinare gli intervalli di confidenza per la differenza tra i valori medi di due popolazioni normali.

Introduciamo un nuovo campione per il confronto

```
campione2 <- rnorm(11000, mean = 100, sd = 1.5)
```

Riportiamo di seguito la media campionaria, la varianza campionaria e la deviazione standard campionaria del campione:

```
## [1] 99.99413
```

```
## [1] 2.284522
```

```
## [1] 1.511464
```

I quantili:

```
##        0%       25%       50%       75%      100%
##  93.49992  98.97647  99.99080 101.01517 106.12078
```

Adesso che abbiamo due campioni casuali indipendenti $X_1, X_2, ..., X_{n_1}$ e $Y_1, Y_2, ..., Y_{n_2}$ di ampiezza n_1 e n_2 estratti da due popolazioni normali $N(mu_1, \sigma_1^2)$ e $N(mu_2, \sigma_2^2)$, vogliamo analizzare i seguenti problemi:

1. Determinare un intervallo di confidenza di grado $1 - \alpha$ per $\mu_1 - \mu_2$ quando entrambe le varianze σ_1^2 e σ_2^2 sono note;

Denotiamo con $\bar{X_{n_1}}$ e $\bar{X_{n_2}}$ le medie campionarie delle due popolazioni normali. Per ipotesi i campioni risultano essere **indipendenti** e la statistica $\bar{X}_{n_1}$ e $\bar{Y}_{n_2}$ risulta essere distribuita normalmente con valore medio $\mu_1 - \mu_2$ e varianza $\sigma_1^2/n_1 + \sigma_2^2/n_2$.

Per determinare un intervallo di confidenza di grado $1-\alpha$ con le varianze note consideriamo la seguente variabile aleatoria di pivot:

$$Z_n = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

la variabile appena descritta dipende dal campione casuale e dal parametro non noto $\mu_1 - \mu_2$ ed è caratterizzata da una *densità normale standard*.

L'intervallo ricavato per una stima dell'intervallo di confidenza $1-\alpha$ per la differenza $\mu_1 - \mu 2$ è:

$$\overline{X}_{n1} - \overline{Y}_{n2} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \overline{X}_{n1} - \overline{Y}_{n2} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Poniamo il grado di fiducia pari a ottenendo quindi $\alpha = 0.05$ e stimiamo $\mu_1 - \mu_2$ per i due campioni sotto analisi.

```
alpha <- 1 - 0.95

ncamp1 <- length(campione)
ncamp2 <- length(campione2)

mc1 <- mean(campione)
mc2 <- mean(campione2)

sigmacamp1 <- 1
sigmacamp2 <- 2.25

#limite inferiore

mc1-mc2-qnorm(1-alpha/2,mean=0,sd=1)*sqrt(sigmacamp1^2/ncamp1+sigmacamp2^2/ncamp2)
```

```
## [1] 9.969501
```

```
#limite superiore
mc1-mc2+qnorm(1-alpha/2,mean=0,sd=1)*sqrt(sigmacamp1 ^2/ncamp1+sigmacamp2^2/ncamp2)
```

```
## [1] 10.06228
```

2. Determinare un intervallo di confidenza di grado $1-\alpha$ per $\mu_1 - \mu_2$ per campioni numerosi con entrambe le varianze σ_1^2 e σ_2^2 non note.

Durante l'analisi abbiamo visto che le varianze campionarie $S_{n_1}^2$ e $S_{n_1}^2$ risultano essere stimatori di σ_1^2 e σ_2^2 quando le ampiezze tendono ad essere grandi. È quindi possibile considerare la variabile aleatoria

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_{n_1}^2}{n_1} + \frac{S_{n_2}^2}{n_2}}}$$

applicando il metodo pivotale in forma apporssimata abbiamo che

$$\overline{X}_{n1} - \overline{Y}_{n2} - z_{\alpha/2}\sqrt{\frac{S_{n_1}^2}{n_1} + \frac{S_{n_2}^2}{n_2}} < \mu_1 - \mu_2 < \overline{X}_{n1} - \overline{Y}_{n2} + z_{\alpha/2}\sqrt{\frac{S_{n_1}^2}{n_1} + \frac{S_{n_2}^2}{n_2}}$$

Come prima, poniamo il grado di fiducia pari a 0.95 e da questo ne risulta che $\alpha = 0.05$. Procediamo quindi alla stima di $\mu_1 - \mu_2$:

```
alpha <- 1 - 0.95

ncamp1 <- length(campione)
ncamp2 <- length(campione2)

mc1 <- mean(campione)
mc2 <- mean(campione2)

sigmacamp1 <- sd(campione)
sigmacamp2 <- sd(campione2)

#limite inferiore
mc1-mc2-qnorm(1-alpha/2,mean=0,sd=1)*sqrt(sigmacamp1^2/ncamp1+sigmacamp2^2/ncamp2)
```

```
## [1] 9.981602
```

```
#limite superiore
mc1-mc2+qnorm(1-alpha/2,mean=0,sd=1)*sqrt(sigmacamp1^2/ncamp1+sigmacamp2^2/ncamp2)
```

```
## [1] 10.05018
```

Verifica delle ipotesi in R

La verifica delle ipotesi è, insieme alla stima dei parametri, una delle aree più importanti dell'inferenza statistica. Essa viene utilizzata in molti ambiti reali, soprattutto quando vi sono da fare indagini di mercato o indagini sperimentali e industriali. Introduciamo quindi il concetto di ipotesi statistica. Una **ipotesi statistica** è un'affermazione o una congettura su un parametro non noto ϑ. Se l'ipotesi statistica specifica completamente $f(x : \vartheta)$ è detta **ipotesi semplice**, altrimenti è chiamata **ipotesi composta**.

L'ipotesi soggetta a verifica viene denotata con H_0 e viene chiamata **ipotesi nulla**. Il procedimento o regola attraverso il quale si decide, sulla base dei dati del campione, se *accettare* o *rifiutare* H_0 è chiamato **test di ipotesi**. La costruzione del test di ipotesi prevede la formulazione di una ipotesi in contrapposizione all'ipotesi nulla. Questa ipotesi prende il nome di **ipotesi alternativa** e viene indicata con H_1.

Quindi, abbiamo l'ipotesi nulla se $\vartheta \in \Theta_0$ e l'ipotesi alternativa se $\vartheta \in \Theta_1$, dove Θ_0 e Θ_1 sono due sottoinsiemi disgiunti dello spazio Θ dei parametri.

Il problema si riduce quindi a determinare un test che ci permetta di suddividere l'insieme dei possibili campioni in due sottoinsiemi detti **regione di accettazione** A dell'ipotesi nulla ed una **regione di rifiuto** R dell'ipotesi nulla. In caso l'ipotesi nulla risulti essere falsa, l'ipotesi alternativa sarà vera e viceversa. Si dice che l'ipotesi H_0 va verificata in alternativa all'ipotesi H_1.

È possibile incorrere in due tipi di errore: - **rifiutare** *l'ipotesi nulla* H_0 *nel caso in cui essa risulti vera*: viene commesso un **errore di tipo I** e la probabilità di commettere questo errore è denotata con α.

- **accettare** *l'ipotesi nulla* H_0 *nel caso in cui essa risulti falsa*: viene commesso un **errore di tipo II** e la probablità di commettere questo errore è denotata con β.

Per campioni casuali di fissata ampiezza, se si diminuisce la probabilità di commettere un errore del tipo I aumenta quella di commettere un errore di tipo II e viceversa. Quindi, siccome non è possibile minimizzare entrambe le probabilità, quello che si fa è **fissare la probabilità di comemttere un errore di tipo I** (si sceglie piccolo) e cercare un test ψ che minimizzi la probabilità di commettere un errore di tipo II. Viene fissata l'errore di tipo I perchè solitamente, quando vengono formulate le ipotesi, commettere questo tipo di errore risulta essere più grave della tipologia II in quanto corrisponde a rifiutare il vero. Si sceglie α uguale a 0.05, 0.01 o 0.001 ed il test viene rispettivamente detto **statisticamente significativo**,**statisticamente molto significativo** e **statisticamente estremamente significativo**. Più piccolo è il valore di α tanto maggiore è la credibilità di un eventuale rifiuto dell'ipotesi nulla.

I test statistici sono di due tipi: - **unilaterali (o unidirezionali)**:

$$H_0 : \vartheta \leq \vartheta_0 \qquad H_1 : \vartheta > \vartheta_0$$

oppure

$$H_0 : \vartheta \geq \vartheta_0 \qquad H_1 : \vartheta < \vartheta_0$$

- **bilaterali (o bidirezionali)**:

$$H_0 : \vartheta = \vartheta_0 \qquad H_1 : \vartheta \neq \vartheta_0$$

Passiamo alla verifica delle ipotesi sul campione da noi generato.

Verifica delle ipotesi su popolazione normale

Analizziamo i seguenti problemi utilizzando il campione da noi generato in precedenza:

Verifica ipotesi su μ con varianza σ^2 nota

Test bilaterale

Dato il nostro campione, andiamo ad considerare le seguenti ipotesi:

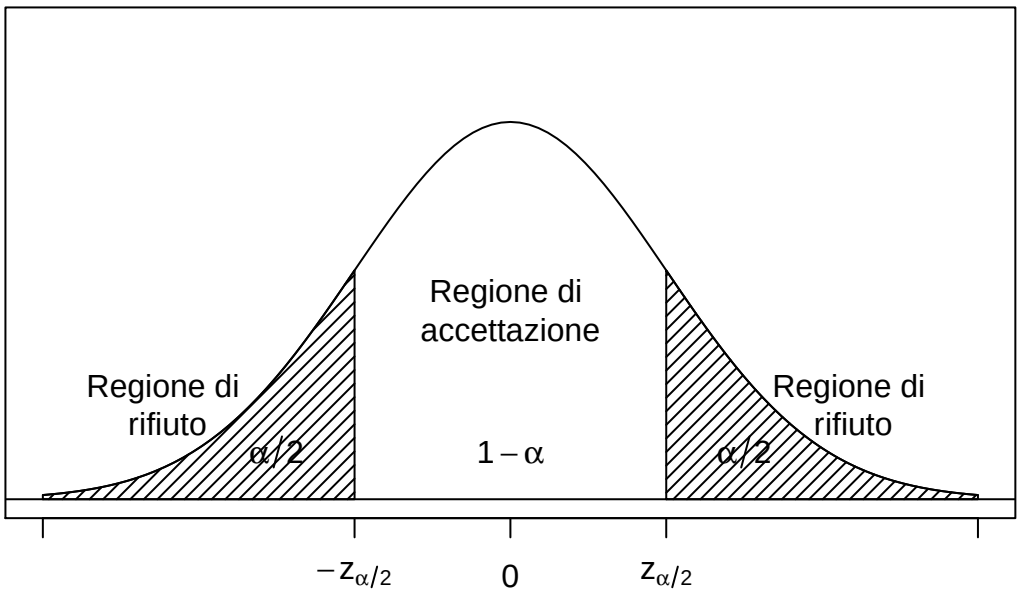
$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$$

Siccome la varianza è nota, l'ipotesi H_0 risulta essere **semplice** mentre H_1 è **composta**. Quando l'ipotesi nulla è vera, la variabile aleatoria

$$Z_n = \frac{\overline{X_n} - \mu_0}{\sigma/\sqrt{n}}$$

ci aiuta moltissimo nella verifica delle ipotesi. Ricordiamo che questa variabile aleatoria è distribuita normalmente e ha valore medio nullo e varianza pari a 1. Effettuiamo il test bilaterale utilizzando le regioni di accettazione:

- accettiamo H_0 se $-z_{\alpha/2} < \frac{\overline{X_n}-\mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}$
- rifiutiamo H_0 se $\frac{\overline{X_n}-\mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2}$ oppure se $z_{\alpha/2} < \frac{\overline{X_n}-\mu_0}{\sigma/\sqrt{n}}$



Effettuiamo la verifica sul nostro campione

```
alpha <- 0.05
mu0 <- 110.20
sigma <- 2

qnorm(1-alpha/2, mean=0, sd=1)
```

```
## [1] 1.959964
```

```
- qnorm(1- alpha/2, mean=0, sd=1)
```

```
## [1] -1.959964
```

```
n <- length(campione)

meancamp <- mean(campione)

(meancamp-mu0)/(sigma/sqrt(n))
```

```
## [1] -9.49891
```

Siccome z cade al di fuori della zona di accettazione, andiamo a rifiutare l'ipotesi nulla con un livello di significatività del 5%.

Test unilaterale sinistro

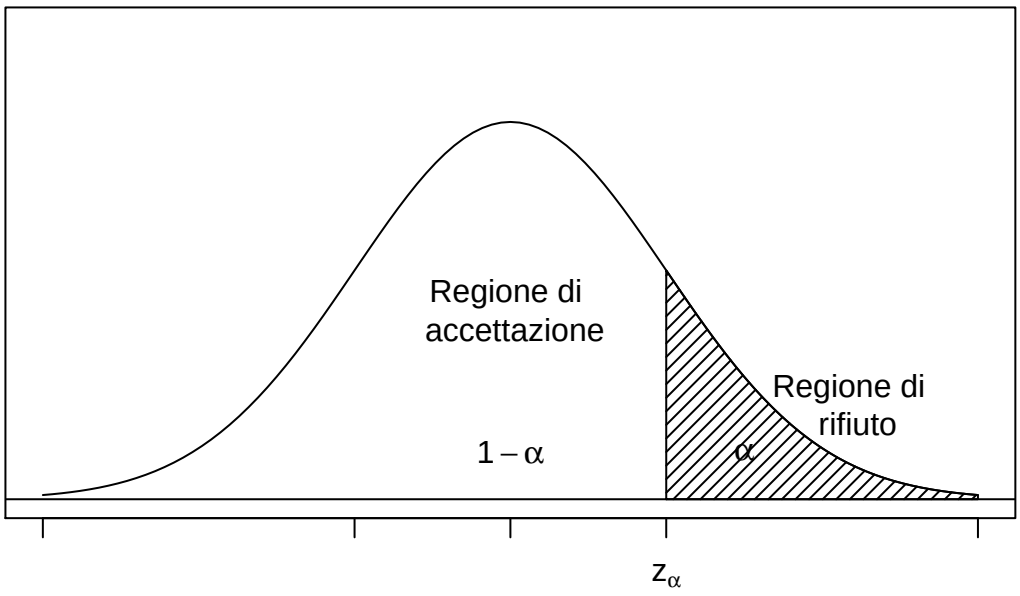
Andiamo a considerare le seguenti ipotesi:

$$H_0 : \mu \leq \mu_0, \qquad H_1 : \mu > \mu_0$$

Sia H_0 e H_1 sono ipotesi composte.

- **accettiamo** H_0 se $\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < z_\alpha$
- **rifiutiamo** H_0 se $\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$

```
curve (dnorm (x,mean=0,sd=1) ,from=-3, to=3, axes=FALSE ,ylim=c(0 ,0.5)
,xlab="",
ylab="",main="Densità normale standard")
text (0,0.05, expression (1- alpha))
text (0,0.2,"Regione di \n accettazione")
axis(1,c(-3,-1,0,1,3) ,c("","" ","" ,expression (z[alpha ]),""))
vals <-seq(1,3, length =100)
x<-c(1,vals ,3,1)
y<-c(0, dnorm(vals) ,0,0)
polygon (x,y,density =20, angle =45)
abline (h=0)
text (1.5 ,0.05 , expression (alpha))
text (2.2 ,0.1 , "Regione di \n rifiuto")
box ()
```



Test unilaterale destro

Andiamo a considerare le seguenti ipotesi:

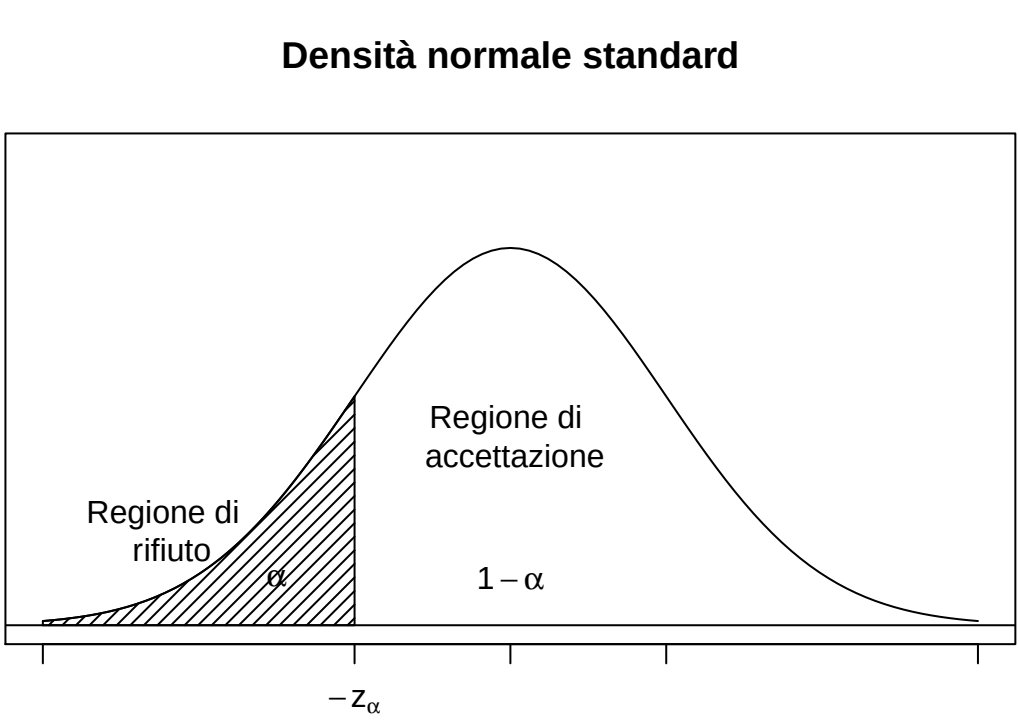
$$H_0 : \mu \geq \mu_0, \qquad H_1 : \mu < \mu_0$$

Sia H_0 e H_1 sono ipotesi composte.

- **accettiamo** H_0 se $\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} > -z_\alpha$
- **rifiutiamo** H_0 se $\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$

```
curve (dnorm(x,mean=0,sd =1) ,from=-3, to=3, axes=FALSE ,ylim=c(0 ,0.5)
,xlab="",
ylab="",main="Densità normale standard")
text (0,0.05, expression (1- alpha))
text (0,0.2,"Regione di \n accettazione")
axis(1,c(-3,-1,0,1,3) ,c("",expression (-z[alpha ]),"","",""))
vals <-seq(-3,-1, length =100)
x<-c(-3,vals ,-1,-3)
```

```
y<-c(0, dnorm(vals) ,0,0)
polygon (x,y,density =20, angle =45)
abline (h=0)
text (-1.5,0.05, expression (alpha ))
text (-2.2 ,0.1 , "Regione di \n rifiuto")
box ()
```



Verifica ipotesi su μ con varianza σ^2 NON nota

Test bilaterale

Dato il nostro campione, andiamo ad considerare le seguenti ipotesi:

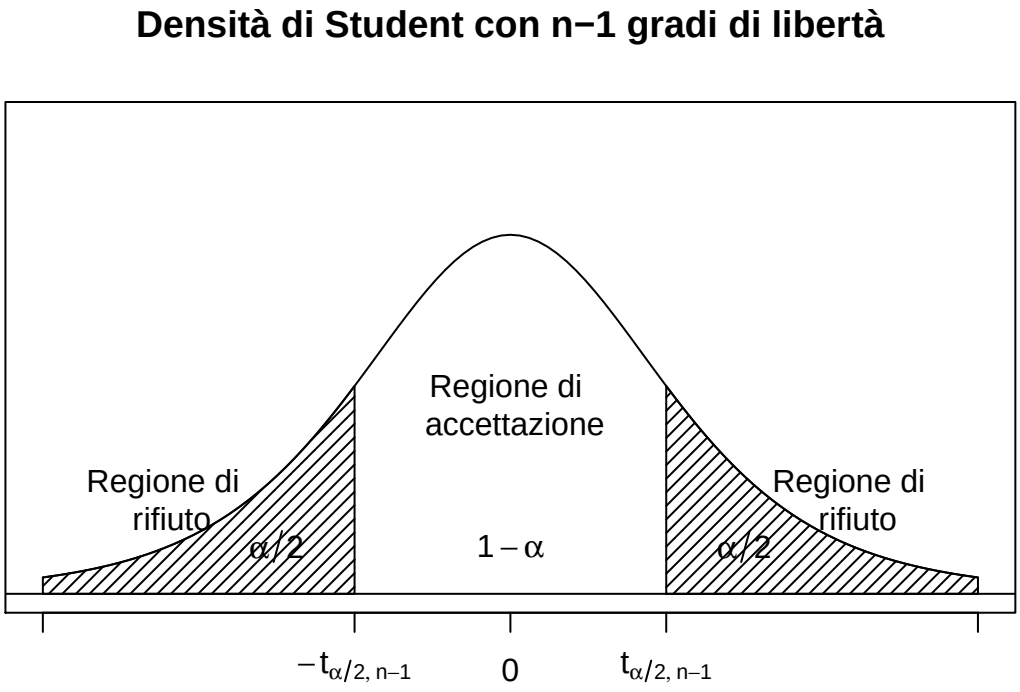
$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$$

Siccome la varianza non è nota, le ipotesi H_0 e H_1 risultano essere composte. Quando l'ipotesi nulla è vera, la variabile aleatoria

$$T_n = \frac{\overline{X_n} - \mu_0}{S_n/\sqrt{n}}$$

distribuita con la legge di Student con *n-1* gradi di libertà ci aiuta moltissimo nella verifica delle ipotesi. Effettuiamo il test bilaterale utilizzando le regioni di accettazione:

- accettiamo H_0 se $-t_{\alpha/2,n-1} < \frac{\overline{x_n}-\mu_0}{s_n/\sqrt{n}} < t_{\alpha/2,n-1}$
- rifiutiamo H_0 se $\frac{\overline{x_n}-\mu_0}{s_n/\sqrt{n}} < -t_{\alpha/2,n-1}$ oppure se $t_{\alpha/2,n-1} < \frac{\overline{x_n}-\mu_0}{s_n/\sqrt{n}}$



Effettuiamo la verifica sul nostro campione

```
alpha <- 0.05
mu0 <- 110.20
sigma <- 2

qt(1-alpha/2, df=n-1)
```

```
## [1] 1.960201
```

```
-qt(1- alpha/2, df=n-1)
```

```
## [1] -1.960201
```

```r
n <- length(campione)

meancamp <- mean(campione)
sdCamp <- sd(campione)
(meancamp-mu0)/(sdCamp/sqrt(n))
```

```
## [1] -19.15384
```

Siccome z cade al di fuori della zona di accettazione, andiamo a rifiutare l'ipotesi nulla con un livello di significatività del 5%.

Test unilaterale sinistro

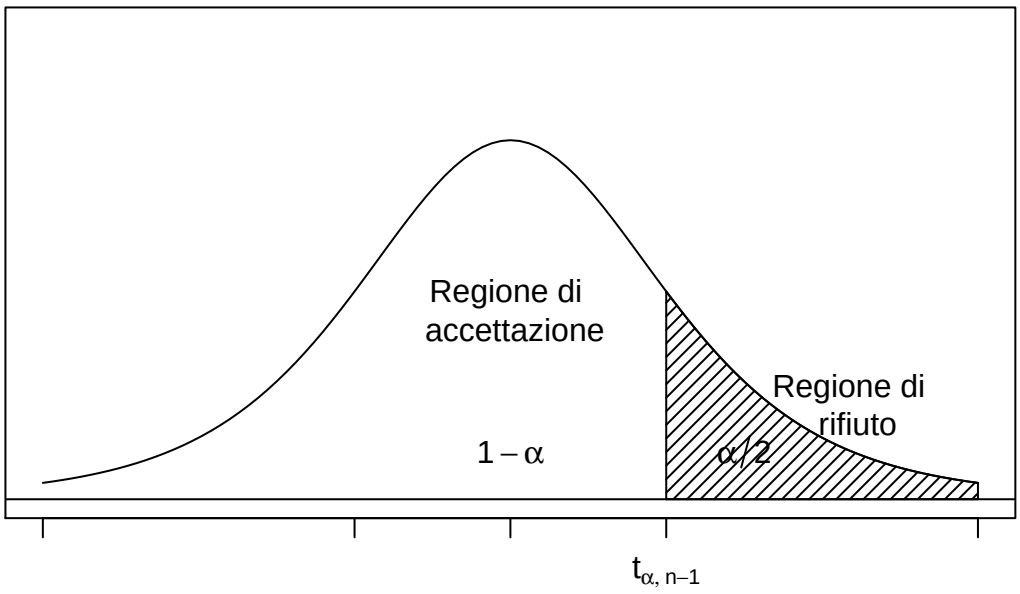
Andiamo a considerare le seguenti ipotesi:

$$H_0 : \mu \leq \mu_0, \qquad H_1 : \mu > \mu_0$$

Sia H_0 e H_1 sono ipotesi composte.

- **accettiamo** H_0 se $\frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} < t_{\alpha,n-1}$
- **rifiutiamo** H_0 se $\frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} > t_{\alpha,n-1}$

```r
curve (dt(x,df=5) ,from=-3, to=3, axes=FALSE ,ylim=c(0 ,0.5) ,xlab="",
ylab="",main="Densità di Student con n-1 gradi di libertà")
text (0,0.05, expression (1- alpha))
text (0,0.2,"Regione di \n accettazione")
axis(1,c(-3,-1,0,1,3) ,c("","," ","," ",expression (t[list(alpha ,n-1) ]),""))
vals <-seq(1,3, length =100)
x<-c(1, vals ,3,1)
y<-c(0, dt(vals ,df =5) ,0,0)
polygon (x,y,density =20, angle =45)
abline (h=0)
text (1.5 ,0.05 , expression (alpha/2))
text (2.2,0.1, "Regione di \n rifiuto")
box ()
```



Test unilaterale destro

Andiamo a considerare le seguenti ipotesi:

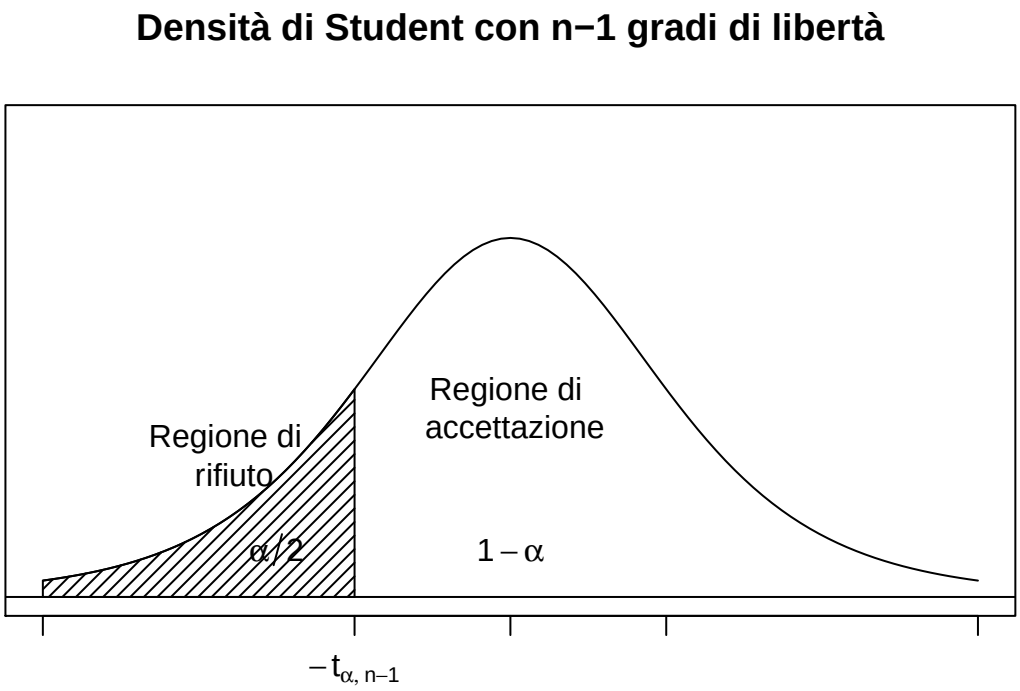
$$H_0 : \mu \geq \mu_0, \qquad H_1 : \mu < \mu_0$$

Sia H_0 e H_1 sono ipotesi composte.

- **accettiamo** H_0 se $\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} > -z_\alpha$
- **rifiutiamo** H_0 se $\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$

```
curve (dt(x,df=5) ,from=-3, to=3, axes=FALSE ,ylim=c(0 ,0.5) ,xlab="",
ylab="",main="Densità di Student con n-1 gradi di libertà")
text (0,0.05, expression (1- alpha))
text (0,0.2,"Regione di \n accettazione")
axis(1,c(-3,-1,0,1,3) ,c("",expression (-t[list(alpha ,n-1) ]),"","",""))
vals <-seq(-3,-1, length =100)
x<-c(-3,vals ,-1,-3)
y<-c(0, dt(vals ,df =5) ,0,0)
```

```
polygon (x,y,density =20, angle =45)
abline (h=0)
text (-1.5,0.05, expression (alpha /2))
text (-1.8 ,0.15,"Regione di \n rifiuto")
box ()
```



Verifica ipotesi sulla varianza σ^2 con valore medio μ noto

Verifica ipotesi sulla varianza σ^2 con valore medio μ NON noto

Criterio chi-quadrato

Con il **criterio del chi-quadrato** è possibile verificare che un certo campione, descritto da una variabile aleatoria X, sia caratterizzato da una funzione di distribuzione $F_X(x)$ con k parametri non noti da stimare. Denotiamo con H_0 l'ipotesi **nulla**, soggetta a verifica, e con H_1 l'ipotesi **alternativa** e le definiamo come di seguito: - H_0: X ha una funzione di distribuzione $F_X(x)$ (avendo stimato k parametri non noti in base al campione) - H_1: X non ha una funzione di distribuzione $F_X(x)$

Il test del chi-quadrato di misura α vuole verificare l'ipotesi nulla dove α è la probabilità massima di rifiutare l'ipotesi nulla quando essa è vera.

Quello che dobbiamo fare è determinare un test ψ di misura α che ci permetta di determinare una regione di accettazione e una di rifiuto. Suddividiamo l'insieme dei valori che la variabile aleatoria X può assumere in r sottoinsiemi $I_1, I_2, ..., I_r$ in modo che, a seconda della distribuzione ipotizzata, p_i rappresenti la probabilità che la variabile aleatoria assuma un valore appartenente a I_i. Il criterio del chi-quadrato si basa sulla seguente statistica

$$Q = \sum_{i=1}^{r} \left(\frac{N_i - np_i}{\sqrt{np_i}} \right)^2$$

dove N_i è la variabile aleatoria che descrive il numero degli elementi del campione casuale $X_1, X_2, ..., X_n$ che cadono nell'intervallo I_i con $i = 1, 2, ..., r$. Se la variabile aleatoria X ha una funzione di distribuzione $F_X(x)$ con k parametri non noti, si può dimostrare che per n abbastanza grande la funzione di distribuzione della statistica Q è approssimabile con la funzione di distribuzione chi-quadrato con *r-k-1* gradi di libertà. Per garantire che ogni classe contenga in media almeno 5 elementi, riteniamo valida l'approssiamzione se

$$min(np_1, np_2, ..., np_r) \geq 5$$

Vediamo la definizione del test chi-quadrato bilaterale:

per un campione sufficientemente numeroso di ampiezza n, il test chi-quadrato bilaterale di misura α è il seguente:

- si rifiuti l'ipotesi H_0 se $X^2 < X^2_{1-\alpha/2,r-k-1}$ oppure $X^2 > X^2_{\alpha/2,r-k-1}$
- si accetti l'ipotesi se H_0 se $X^2_{1-\alpha/2,r-k-1} < X^2 < X^2_{\alpha/2,r-k-1}$

dove $X^2_{1-\alpha/2,r-k-1}$ e $X^2_{\alpha/2,r-k-1}$ sono soluzioni delle equazioni:

$$P(Q < X^2_{1-\alpha/2,r-k-1}) = \frac{\alpha}{2}, \qquad P(Q < X^2_{\alpha/2,r-k-1}) = 1 - \frac{\alpha}{2}$$

Eseguiamolo per il nostro campione iniziando a determinare i sottoinsiemi

```
m <- mean(campione)
d <- sd(campione)

a<-numeric(4)
for(i in 1:4)
  a[i] <- qnorm(0.2 * i, mean=m, sd = d)
a
```

```
## [1] 109.1753 109.7587 110.2613 110.8448
```

Calcoliamo il numero di elementi che cadono negli intervalli $I_1, I_2, ..., I_5$

```
r <- 5
nint <- numeric(r)
nint[1] <- length(which(campione < a[1]))
nint[2] <- length(which((campione >= a[1]) & (campione < a[2])))
nint[3] <- length(which((campione >= a[2]) & (campione < a[3])))
nint[4] <- length(which((campione >= a[3]) & (campione < a[4])))
nint[5] <- length(which(campione > a[4]))
nint
```

```
## [1] 2015 1972 2020 1959 2034
```

Calcoliamo X^2

```
chi2 <- sum(((nint-n*0.2) / sqrt(n*0.2))^2)
chi2
```

```
## [1] 2.123
```

La distribuzione normale ha due parametri non noti $(\mu, \sigma$š) e quindi il $k = 2$. La funzione Q è quindi approssimabile con la funzione di distribuzione chi-quadrato con $r - k - 1 = 2$ gradi di libertà. Procediamo dunque nel calcolo di $X^2_{1-\alpha/2,r-k-1}$ e $X^2_{\alpha/2,r-k-1}$ con $\alpha = 0.05$.

```
k <- 2
alpha <- 0.05
qchisq(alpha/2, df=r-k-1)
```

```
## [1] 0.05063562
```

```
qchisq(1-alpha/2, df=r-k-1)
```

```
## [1] 7.377759
```

Siccome Xš $= 0.422$ risulta essere compresa fra $X^2_{1-\alpha/2,r-k-1} = 7.377759$ e $X^2_{\alpha/2,r-k-1} = 0.05063562$, l'ipotesi H_0 di popolazione normale può essere accettata.