UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI INFORMATICA

CORSO DI LAUREA MAGISTRALE IN INFORMATICA

PROGETTO INTERNO PER
LAUREA MAGISTRALE IN
INFORMATICA

Progetto Statistica e Analisi dei Dati

Docente	Studenti	
Prof.ssa	**Marrazzo Vincenzo**	**Zizzari Antonio**
Amelia Giuseppina Nobile	**0522501325**	**0522501309**

Anno Accademico 2021/2022

Indice

1. Introduzione

Il seguente documento ha come scopo quello di fornire le nozioni e le conoscenze di base sulla distribuzione di probabilità continua normale e di mostrare tramite quest'ultima le applicazioni delle tecniche dell'inferenza statistica.

Lo scopo della statistica inferenziale è quello di derivare le caratteristiche di una popolazione tramite un campione estratto da essa.

L'utilizzo che faremo dunque dell'inferenza statistica è quello di studiare una popolazione descritta da una variabile aleatoria avente distribuzione normale e ottenere delle stime sui parametri non noti e verificare delle ipotesi. La variabile aleatoria è definita osservabile poiché si possono osservare i valori assunti dalla variabile: il parametro non è noto solo nella legge di probabilità (funzione di distribuzione). Il campione inoltre deve essere scelto in modo da essere rappresentativo della popolazione.

Nella relazione da noi presentata si è deciso di approfondire una variabile aleatoria continua, in particolare la variabile aleatoria normale.

2. Variabile aleatoria normale

Ricordiamo la definizione di variabile aleatoria: una variabile aleatoria è una funzione che fa corrispondere un numero reale a ogni esito di un esperimento. Se l'insieme dei valori assunti dalla variabile aleatoria non è numerabile, la variabile si definisce continua: non è possibile elencare tutti i valori essendo un'infinità e non è possibile attribuire una probabilità ai singoli valori. Mentre per una variabile discreta è possibile elencare tutti i valori che essa può assumere, per una variabile continua è necessario definire delle classi, cioè degli intervalli in cui suddividere i possibili valori della variabile.

Introduciamo adesso la funzione di distribuzione normale.

L'importanza della distribuzione normale è dovuta alla sua caratteristica di poter efficacemente approssimare molte distribuzioni (lo vedremo in seguito) di numerosi fenomeni, basti pensare che non sono poche le distribuzioni che sono normalizzabili tramite delle trasformazioni.

Vediamone le caratteristiche.

2.1 Densità di probabilità

Una variabile aleatoria X di densità di probabilità

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

si dice avere distribuzione normale di **parametri** μ e σ.

La densità è simmetrica rispetto all'asse x=μ, risulta infatti fX(μ−x)=fX(μ+x).
La densità ha le seguenti caratteristiche:

- La **forma a campana** rispetto a x=μ
- Il **massimo** è in corrispondenza del punto x=μ ed è pari a $\frac{1}{\sigma\sqrt{2\pi}}$

– Ha due **flessi** in corrispondenza di $\mu-\sigma$ e $\mu+\sigma$

Per indicare una variabile aleatoria **X** che ha distribuzione normale di parametri **μ** e **σ** useremo la notazione $X \sim N(\mu, \sigma)$ (X è una variabile normale).

Prendiamo in esame il seguente array di dati che contiene i voti di 100 studenti.
Per ottenere questo risultato in R utilizziamo la funzione `set.seed()` per rendere l'esperimento riproducibile e la funzione `sample()` per generare i valori desiderati:

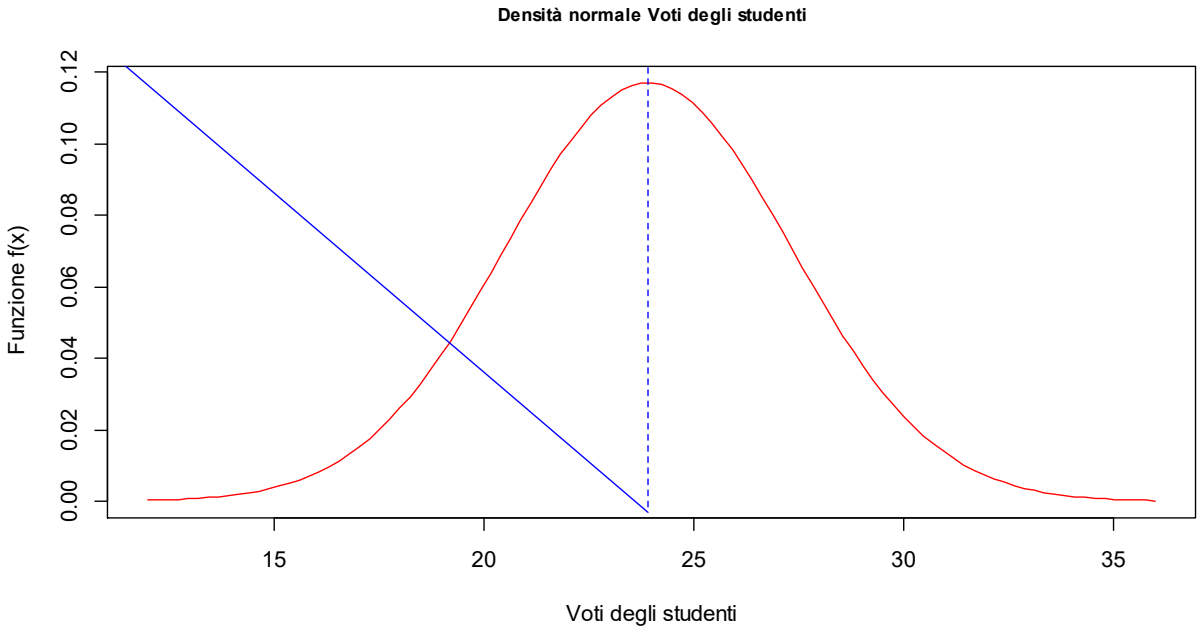
```
> set.seed(3)
> sample(c(18:30), size=100, replace=TRUE)
  [1] 22 27 29 24 21 27 25 28 25 21 27 24 25 25 25 22 19 30 29 30 22 25 22
 [24] 26 26 25 23 19 26 29 25 23 29 20 23 25 27 28 30 24 20 22 24 23 25 27
 [47] 21 18 20 23 21 26 27 30 24 27 18 26 18 26 22 19 23 30 25 28 27 21 26
 [70] 20 21 29 19 18 21 23 18 30 19 22 19 25 24 24 27 23 20 18 27 19 22 28
 [93] 27 21 26 22 23 24 24 20
```

Adesso calcoliamo la densità normale dell'array considerato:

```
> media<-mean(voti)
> media
[1] 23.92
> dev.std<-sd(voti)
> dev.std
[1] 3.410264
```

Successivamente disegniamo attraverso la funzione `curve()` la funzione di densità normale:

```
x<-voti
curve(dnorm(x,media,dev.std),from=12, to=36,xlab="Voti degli studenti",
      ylab="Funzione f(x)",col="red",
      main="Densità normale Voti degli studenti",cex.main=0.8)
abline(v=media,lty=2,col="blue")
```



L'intervallo to-from è stato calcolato utilizzando la **regola del 3σ(sigma)**, in maniera tale che l'area sottesa dalla curva sia circa unitaria e l'area delle code destra e sinistra sia trascurabile.

Regola del 3σ

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

Quindi la probabilità che una variabile aleatoria X ~ N (μ, σ) assuma valori in un intervallo avente come centro $\boldsymbol{\mu}$ e semiampiezza $\boldsymbol{3\sigma}$ è prossima all'unità

Andiamo a verificare nel nostro caso tramite il comando `pnorm()`:

```
> pnorm(36,media,dev.std)-pnorm(12,media,dev.std)
[1] 0.9995649
```

Tale valore è prossimo all'unità e perciò possiamo dire che il nostro intervallo è corretto.

Il punto più alto della curva si trova proprio in corrispondenza del valore medio, ossia 23.92. La deviazione standard è 3.410264, se la deviazione standard fosse stata più bassa la curva sarebbe stata più stretta.

2.2 Funzione di distribuzione

La funzione di distribuzione di una variabile aleatoria X ~ N(μ, σ) è:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} fx(y)dy = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

dove

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{y^2}{2}} dy$$

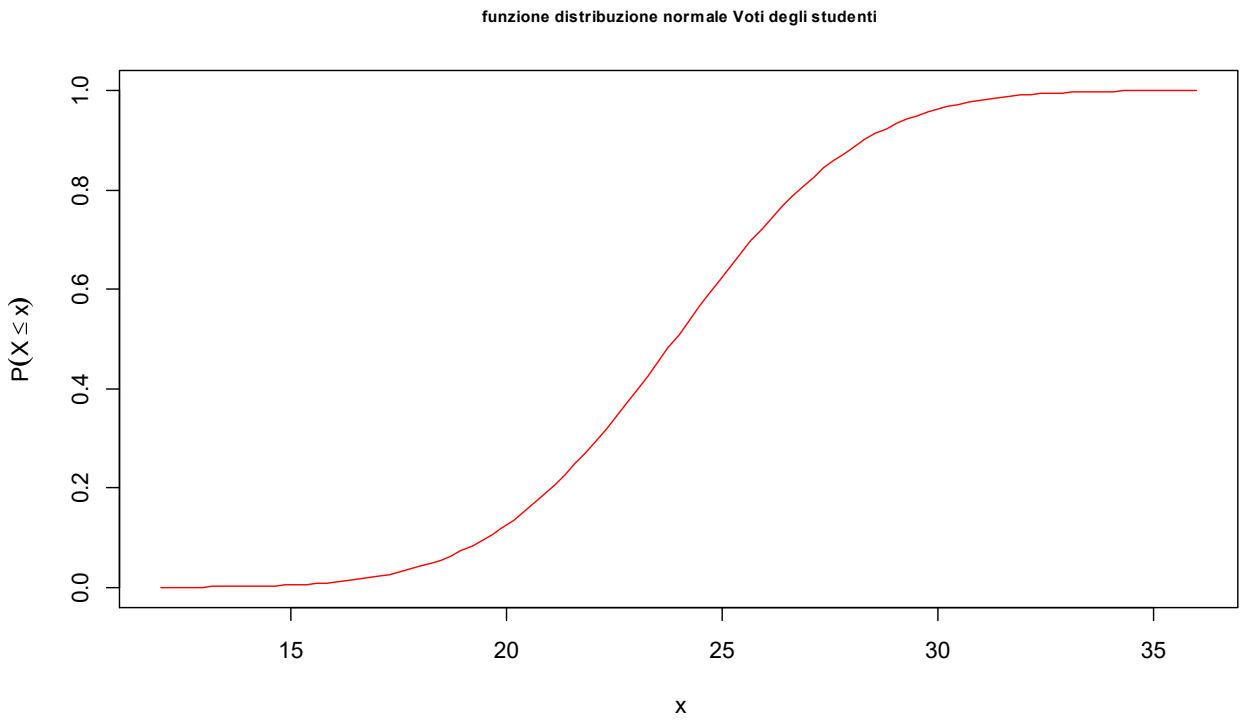
è la funzione di distribuzione di una variabile aleatoria Z ~ N(0, 1), detta **normale standard**.

Quindi se X ~ N(μ, σ) si ha:

$$P(a < X < b) = F_X(b) - F_X(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

In R si calcola con `pnorm()`:

```
curve(pnorm(x, media, dev.std), from=12, to=36,
      xlab="x", ylab=expression (P(X<=x)),
      main="funzione distribuzione normale Voti degli studenti",
      cex.main=0.65,col="red")
```



Notiamo che per i voti che vanno da 20 a 26 circa, la probabilità aumenta (la curva tende ad alzarsi). In R è possibile calcolare i quantili per una distribuzione normale, tramite `qnorm()`:

```
> scelta<-c(0,0.25,0.5,0.75,1)
> qnorm(scelta,media, dev.std)
[1]     -Inf 21.61981 23.92000 26.22019      Inf
```

Il voto 21 corrisponde al 25% dei voti, il voto 26 corrisponde al 75% e la media (ossia il 50% dei voti) corrisponde proprio a 23.92 Q0= -inf e Q4= +inf.

3. Stima puntuale

Quando parliamo di stime puntuali quello che vogliamo fare è ottenere informazioni su un parametro non noto della popolazione effettuando su un campione estratto da quest'ultima delle opportune misure.

Introduciamo quindi gli stimatori.
Quando parliamo di uno stimatore si intende una funzione che associa ad ogni possibile campione un valore del parametro che si vuole stimare.
Abbiamo dunque una variabile casuale funzione del campione che assume valore tra i possibili valori del parametro che si vuole stimare.

Nell'inferenza statistica si fa uso degli stimatori, detti anche statistiche, per ricavare da un campione di *n* osservazioni un valore per un parametro non noto della funzione di distribuzione statistica.

Vediamo la definizione formale di **stimatore**:
Uno stimatore $\hat{\Theta} = t(X_1, X_2, \dots, X_n)$ è una **funzione misurabile e osservabile** del campione $(X_1, X_2, \dots, X_n)$ i cui valori sono usati per stimare un parametro non noto ϑ della popolazione. I valori $\hat{\vartheta}$ assunti dallo stimatore sono dette stime del parametro ϑ.

Tra gli stimatori tipici ci sono:

- **media campionaria**,
- **varianza campionaria**.

Vediamo la seguente proposizione:
Sia $(X_1, X_2, \dots, X_n)$ un campione casuale estratto da una popolazione descritta da una variabile aleatoria osservabile X caratterizzata da valore medio E(X)=μ e varianza $Var(x) = \sigma^2$ entrambi finiti, risulta:

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

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

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

e anche:

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

Questo significa che tanto più è numeroso il campione, migliore è la stima del valore medio della popolazione.

3.2 Metodi di ricerca di stimatori

I principali metodi di stima puntuale dei parametri sono il **metodo dei momenti** e il **metodo della massima verosimiglianza**

3.2.1 Metodo dei momenti

Per descrivere il metodo bisogna introdurre il concetto di momento campionario.

Si definisce momento campionario r-esimo relativo ai valori osservati $(x_1, x_2, \ldots, x_n)$ del campione casuale il valore:

$$M_r(x_1, x_2, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^r, \qquad (r = 1,2,3,\ldots)$$

si tratta dunque della media aritmetica delle potenze r-esime delle *n* osservazioni effettuate sulla popolazione. Se r = 1 otteniamo la media campionaria.

Il metodo dei momenti prevede **nell'uguagliare i primi k momenti della popolazione con i corrispondenti momenti del campione casuale**.
Bisogna cioè risolvere il sistema di k equazioni:

$$E(X^r) = M_r(x_1, x_2, \ldots, x_n) \qquad (r = 1,2,3,\ldots,k)$$

La stima dipende dal campione osservato.

Vediamo dunque l'applicazione del metodo su una popolazione normale. Quello che vogliamo è stimare i parametri $\boldsymbol{\mu}$ e $\boldsymbol{\sigma^2}$.

Ricordando che $\sigma^2 = E[x_i^2] - (E[x_i])^2$, il momento di ordine 2: $E[x_i^2] = \sigma^2 + (E[x_i])^2 = \sigma^2 + \mu^2$.

Quindi avremo:

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

Dalla seconda equazione si ricava:

$$\widehat{\sigma}^2 = \frac{(x_1, x_2, \ldots, x_n)^2}{n} - \frac{(x_1, x_2, \ldots, x_n)^2}{n^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Quindi per μ indichiamo lo stimatore della media campionaria, mentre per σ lo stimatore della varianza campionaria, ovvero $\frac{n-1}{n}S^2$

Consideriamo dunque il nostro campione e stimiamo i parametri in R:

```
> stimaMediaMomenti<-mean(voti)
> stimaMediaMomenti
[1] 23.92
> stimaVarianzaMomenti<-(length(voti)-1)*var(voti)/length(voti)
> stimaVarianzaMomenti
[1] 11.5136
```

Abbiamo ottenuto quindi le stime per $\widehat{\mu} = 23.92$ e per $\widehat{\sigma}^2 = 11.5136$.
Dato che $\widehat{\sigma}^2$ è stato calcolato su una dimensione del campione abbastanza grande allora esso è asintoticamente corretto. Dunque, la dimensione del campione è importante poiché quando si va a divedere "(length(voti)-1)" con "length(voti)" il risultato deve essere circa 1 e non proprio pari ad 1, altrimenti non otterremmo più quella che è la stima della varianza, ma la varianza stessa.

3.2.2 Metodo della massima verosimiglianza

Il metodo della massima verosimiglianza di solito è preferito a quello dei metodi, ed è infatti considerato il metodo migliore per la stima dei parametri non noti. Dobbiamo introdurre il concetto di funzione di verosimiglianza per descriverlo: Sia $(x_1, x_2, \ldots, x_n)$ un campione casuale estratto dalla popolazione. La funzione di verosimiglianza L($\vartheta 1, \vartheta 2, \ldots, \vartheta k$) = L($\vartheta 1$, $\vartheta 2$, ..., ϑk; $x_1, x_2, \ldots, x_n$) è la **funzione di probabilità** (**densità nel caso continuo**) **congiunta del campione casuale** $(x_1, x_2, \ldots, x_n)$, cioè:

$$L(\vartheta 1, \vartheta 2, \ldots, \vartheta k) = f(x_1;\, \vartheta 1, \vartheta 2, \ldots, \vartheta k) f(x_2;\, \vartheta 1, \vartheta 2, \ldots, \vartheta k) \ldots f(x_n;\, \vartheta 1, \vartheta 2, \ldots, \vartheta k)$$

Il metodo consiste nel **massimizzare questa funzione** rispetto ai parametri $\vartheta 1$, $\vartheta 2$, ..., ϑk: si cerca di determinare da quale funzione di probabilità (densità) congiunta è più **verosimile** (per questo verosimiglianza) che provenga il campione osservato.
Si cercano i vari ϑ_i in modo tale che spieghino meglio il campione osservato.

I valori stimati, indicati con $\hat{\vartheta}_i$ sono detti **stime di massima verosimiglianza**.
Anche in questo caso le stime dipendono dal campione.

Vediamo per stimare i parametri di una popolazione normale cosa dobbiamo fare.

La funzione di densità della normale è la seguente:

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

Abbiamo dunque:

$$L(\mu, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^2 \; exp\left\{-\sum_{i=1}^{n} \frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Dai calcoli si ricavano rispettivamente $\widehat{\mu}$ lo stimatore è $\frac{1}{n}\sum_{i=1}^{n} x_i$, mentre per σ^2 lo stimatore è $\frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2$

Quindi per μ lo stimatore è la media campionaria, mentre per σ la variabile aleatoria $\frac{n-1}{n}S^2$: entrambi gli stimatori coincidono con quelli calcolati col metodo dei momenti; dunque, risulta inutile ricalcolarli col metodo della massima verosimiglianza.

3.3 Proprietà degli stimatori

Dato che per stimare un parametro di una popolazione ci possono essere diversi stimatori, sono definite alcune proprietà.

Gli stimatori sono quindi classificati in:

- corretto,
- più efficiente,
- corretto e con varianza uniforme minima,
- asintoticamente corretto,
- consistente.

Uno stimatore si dice corretto se il **valore medio dello stimatore è uguale al corrispondente parametro** non noto della popolazione.

Bisogna dire che ci possono essere **più stimatori corretti**, quindi qualche volta va considerato quale conviene: ci sono dei criteri che permettano di confrontare stimatori dello stesso parametro. Ad esempio, viene usata la ricerca dello stimatore con **errore quadratico uniformemente minimo** per la classe degli stimatori corretti.

Riguardo la popolazione normale ricaviamo che la media campionaria è uno **stimatore corretto** del parametro μ di una popolazione normale con varianza minima, mentre lo stimatore $\frac{n-1}{n}S^2$ della varianza σ^2 individuato sia con il metodo dei momenti che con il metodo della massima verosimiglianza, risulta **asintoticamente corretto**: il valore medio dello stimatore con n grande tende al corrispondente parametro non noto della popolazione.
Inoltre, entrambi gli stimatori sono **consistenti**.

4. Stima intervallare

Anziché determinare un singolo valore per un parametro non noto come si fa nel caso delle stime puntuali, spesso si preferisce trovare un **intervallo di valori** nel quale il parametro non noto sia compreso in modo tale che questo intervallo abbia un **buon coefficiente di confidenza**.

Diamo dunque una definizione di **intervallo di confidenza**: Fissato un coefficiente di confidenza $1-\alpha$ ($0<\alpha<1$) se è possibile scegliere due statistiche $\underline{C}_n$ e $\overline{C}_n$ in modo tale che:

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

allora $(\underline{C}_n, \overline{C}_n)$ è un **intervallo di confidenza** di grado $1-\alpha$ per il parametro ϑ.

Le statistiche $(\underline{C}_n, \overline{C}_n)$ sono dette **limite inferiore e superiore** dell'intervallo di confidenza.

L'intervallo ottenuto è detto **stima dell'intervallo di confidenza**, e i punti forniti dalle statistiche sono detti rispettivamente **stima del limite inferiore** e **stima del limite superiore dell'intervallo di confidenza**.

Dato che gli intervalli di confidenza di grado $1-\alpha$ possono essere più di uno, di solito si sceglie l'intervallo, fissato il grado di confidenza, che abbia la **lunghezza assoluta o media più piccola possibile** (restringiamo l'intervallo il più possibile).

Va detto che ovviamente la **stima puntuale deve cadere nell'intervallo**.

4.2 Metodo pivotale

Il metodo prevede la determinazione di una variabile aleatoria di pivot $\gamma(X_1, X_2, \dots, X_n; \vartheta)$ che **dipende dal campione casuale estratto** e **dal parametro non noto** ϑ e la cui funzione di distribuzione non contiene il parametro che si vuole stimare.

Si noti che la variabile aleatoria definita non è osservabile in quanto dipende dal parametro non noto ϑ, quindi **non è statistica**.

Vediamo dunque nel dettaglio in cosa consiste il metodo: Per ogni coefficiente α, siano α_1 e α_2, con $\alpha_1 < \alpha_2$, due valori dipendenti soltanto dal coefficiente fissato α tali che per qualunque parametro non noto ϑ si abbia:

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

Se per ogni campione $(X_1, X_2, \dots, X_n)$ e per ogni ϑ e qualunque campione si riesce a dimostrare:

$$\alpha_1 < \gamma(x; \vartheta) < \alpha_2 \Leftrightarrow g_1(x) < \vartheta < g_2(x)$$

con $g_1(x)$ e $g_2(x)$ dipendenti soltanto dal campione osservato allora la probabilità precedente è esprimibile come:

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

Denotiamo $\underline{C}_n = g_1(X_1, X_2, \dots, X_n)$ e $\overline{C}_n = g_2(X_1, X_2, \dots, X_n)$, allora $(\underline{C}_n, \overline{C}_n)$ è un intervallo di confidenza di grado $1-\alpha$ per ϑ.

Analizziamo quindi di seguito diversi problemi relativi a un campione normale.

1. Determinare un intervallo di confidenza di grado $1-\alpha$ 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

Abbiamo visto che $E(\overline{X}_n = \mu)$ e $Var(\overline{X}_n) = \frac{\sigma^2}{n}$.

Vogliamo determinare un intervallo di confidenza α − 1 per il parametro μ avendo nota la varianza.
Usiamo il metodo pivotale e consideriamo la variabile aleatoria standardizzata

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

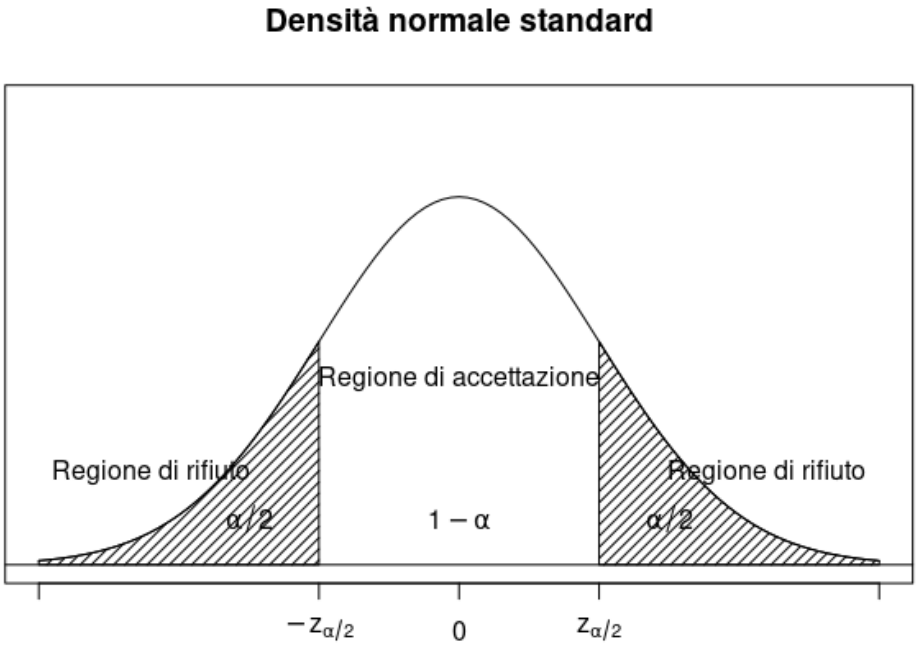
Questa variabile è una normale standard che dipende dal campione e dal parametro non noto, quindi posso applicare il metodo pivotale.

Dato che la distribuzione è normale, sappiamo che la curva è simmetrica quindi ci conviene scegliere $\alpha_1 = -\alpha_2$. Scegliamo quindi $\alpha_1 = -z_{a/2}\ e\ \alpha_2 = z_{a/2}$ in modo che

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

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

Graficamente quanto detto si traduce nel seguente modo:



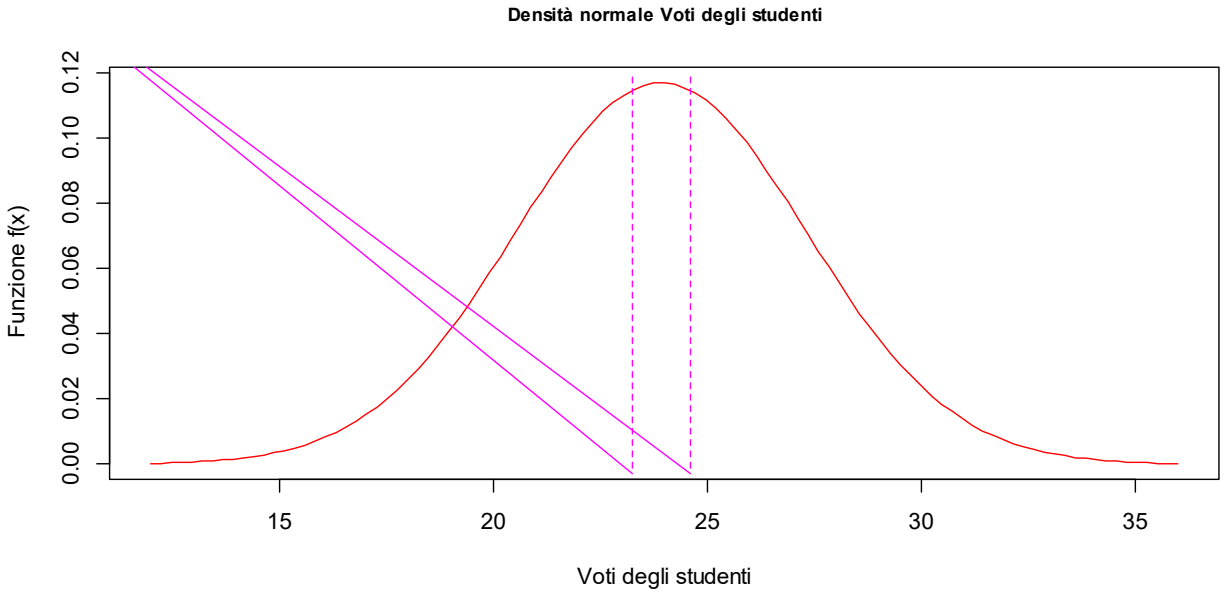
Una stima dell'intervallo di confidenza 1−α per il valore medio μ è:

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

Si considera un grado di fiducia pari a $1 - \alpha = 0.95$, con conseguente α pari a 0.05 si ottiene in R:

```
> alpha<-1-0.95
```

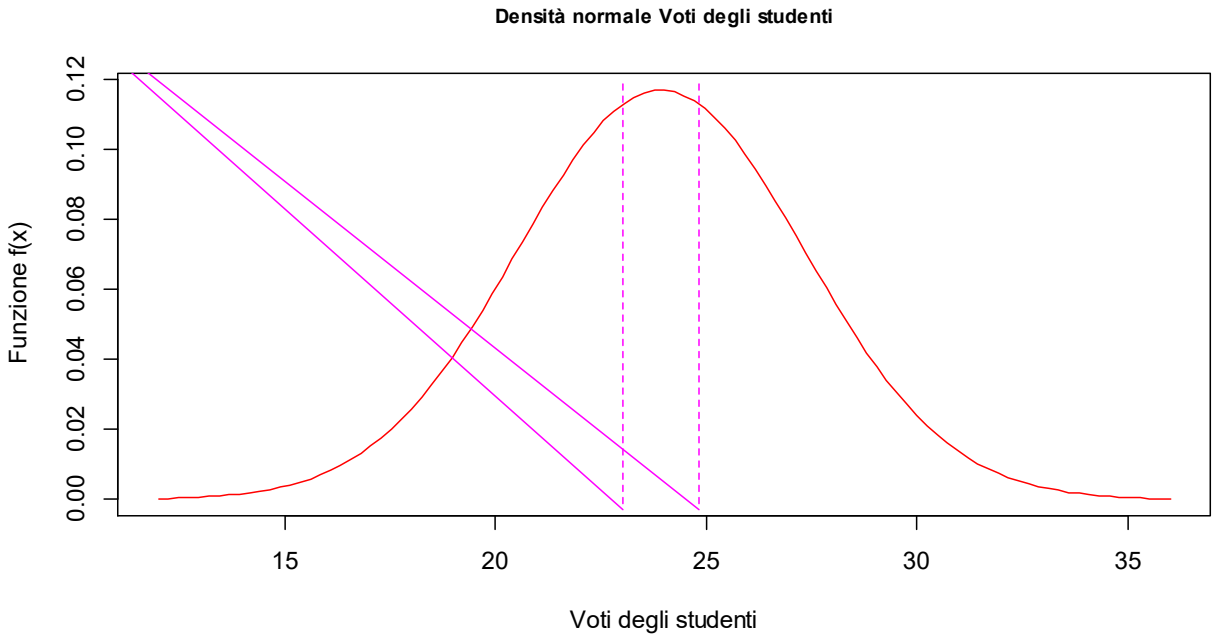
```
>
> deviazioneStandard<-3.4
>
> n<-length (voti)
> #stima del limite inferiore
> mean(voti)-qt(1- alpha /2,df=n-1)*deviazioneStandard/sqrt(n)
[1] 23.24537
>
> #stima del limite superiorez
> mean(voti)+qt(1- alpha /2,df=n-1)*deviazioneStandard/sqrt(n)
[1] 24.59463
```



Risulta che il nostro intervallo di confidenza è (23.24537, 24.59463). Essendo il valore medio uguale a 23.92, essa risulta compresa nella stima dell'intervallo ossia risulta accettato.

Si considera un grado di fiducia pari a $1 - \alpha = 0.99$, con conseguente α pari a 0.01 si ottiene in R:

```
> alpha<-1-0.99
>
> deviazioneStandard<-3.4
>
> n<-length (voti)
> #stima del limite inferiore
> mean(voti)-qt(1- alpha /2,df=n-1)*deviazioneStandard/sqrt(n)
[1] 23.02702
>
> #stima del limite superiorez
> mean(voti)+qt(1- alpha /2,df=n-1)*deviazioneStandard/sqrt(n)
[1] 24.81298
```



Risulta che il nostro intervallo di confidenza è (23.02702, 24.81298). Essendo il valore medio uguale a 23.92, essa risulta compresa nella stima dell'intervallo ossia risulta accettato.

Ciò dimostra che all'aumentare del grado di fiducia, aumenta anche l'ampiezza dell'intervallo di confidenza per il parametro lambda. In ogni caso però è importante notare come la stima puntuale per il parametro ricada all'interno degli intervalli di confidenza calcolati tramite stime intervallari.

4.3 Differenza tra valori medi di una popolazione normale

Alcuni problemi richiedono il confronto tra i valori medi di due popolazioni, vediamo come costruire dunque degli intervalli di confidenza per la differenza tra i valori medi di due popolazioni normali.

Come operazione preliminare introduciamo un nuovo campione che ci servirà per il confronto:

```
> set.seed(1)
> voti2<-sample(c(18:30), size=50, replace=TRUE)
> voti2
 [1] 26 21 24 18 19 30 24 28 19 28 20 18 22 22 27 23 27 24 26 22 22 26 26 22 22 19
[27] 27 26 29 18 21 20 23 27 27 23 21 29 21 27 29 26 24 23 26 25 29 26 24 25
```

Di seguito riportiamo i valori ottenuti della media, varianza e deviazione standard sul campione ricavato:

	Indici di sintesi del campione
Media	24.02
Varianza	3.304234
Deviazione Standard	10.91796

Di seguito i **quantili** del campione appena creato:

```
> quantile(voti2)
   0%   25%   50%   75%  100%
18.00 22.00 24.00 26.75 30.00
```

Possiamo dunque ora considerare i problemi veri e propri.

Consideriamo due campioni, $X_1, X_2, \ldots, X_{n1}$ e $Y_1, Y_2, \ldots, Y_{n2}$, casuali e indipendenti, di ampiezza n_1 e n_2 estratti rispettivamente da due **popolazioni normali** $N(\mu_1, \sigma_1^2)$ e $N(\mu_2, \sigma_2^2)$.

I problemi che vogliamo affrontare sono i seguenti:

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. **Intervalli di confidenza** per $\mu_1 - \mu_2$ con σ_1^2 *e* σ_2^2 note

Innanzitutto, consideriamo le medie campionarie dei due campioni, poiché per ipotesi abbiamo detto che i campioni sono **indipendenti**, $\bar{X}_{n1} - \bar{X}_{n2}$ è distribuita normalmente con valore medio $\mu_1 - \mu_2$ e varianze $\frac{\sigma_1^2}{n_1}$ *e* $\frac{\sigma_2^2}{n_2}$.

Per determinare l'intervallo di confidenza $1 - \alpha$ (conoscendo le varianze), consideriamo la **variabile aleatoria di pivot**:

$$Zn = \frac{\overline{X}_{n1} - \overline{Y}_{n2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Questa variabile è di pivot:

- Dipende dal parametro non noto
- Dipende dal campione
- È caratterizzata da una densità normale

Ricaviamo dunque che una stima dell'intervallo di confidenza 1 − α per la differenza tra le medie $\mu_1 - \mu_2$ è:

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

Poniamo α=0.01 e stimiamo $\mu_1 - \mu_2$ per i due campioni che abbiamo a disposizione sapendo che le varianze note sono $\sigma_1^2 = 3.410264$ (sigma1) e $\sigma_2^2 = 3.304234$ (sigma2), mentre la numerosità del primo campione è pari a 100, mentre quella del secondo 50.

Procediamo con la stima:

```
> alpha <-1 -0.95
>
> n1 <- length(voti)
> n2 <- length(voti2)
>
> m1 <- mean(voti)
> m2 <- mean(voti2)
>
> s1 <- 3.4
> s2 <- 3.3
>
> #stima del limite inferiore
> m1-m2-qnorm(1-alpha/2,mean=0,sd=1)*sqrt(s1^2/n1+s2^2/n2)
[1] -1.231699
>
> #stima del limite superiore
> m1-m2+qnorm(1-alpha/2,mean=0,sd=1)*sqrt(s1^2/n1+s2^2/n2)
[1] 1.031699
```

Passiamo ora al problema successivo.

2. Determinare un intervallo di confidenza di grado 1 − α per $\mu_1 - \mu_2$ quando entrambe le varianze σ_1^2, σ_2^2 **non sono note** per campioni numerosi estratti dalla popolazione. **Intervalli di confidenza per $\mu_1 - \mu_2$ con σ_1^2, σ_2^2 non note**

Consideriamo, quindi, ora il caso in cui non abbiamo nessun parametro noto (**caso reale**).

Abbiamo visto in precedenza che le varianze campionarie S_{n1}^2 e S_{n2}^2 sono stimatori di σ_1^2, σ_2^2 quando le ampiezze dei campioni sono abbastanza grandi. Possiamo quindi considerare la variabile aleatoria ricavata dal caso precedente:

$$\frac{\overline{X}_{n1} - \overline{Y}_{n2} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_{n1}^2}{n_1} + \frac{S_{n2}^2}{n_2}}}$$

Ovviamente anche qui abbiamo una variabile aleatoria pivotale, possiamo dunque applicare il metodo pivotale in **forma approssimata** e ricavare che :

$$P\left(\overline{X}_{n1} - \overline{Y}_{n2} - z_{\frac{a}{2}}\sqrt{\frac{S_{n1}^2}{n_1} + \frac{S_{n2}^2}{n_2}} < \mu_1 - \mu_2 < \overline{X}_{n1} - \overline{Y}_{n2} + z_{\frac{a}{2}}\sqrt{\frac{S_{n1}^2}{n_1} + \frac{S_{n2}^2}{n_2}}\right) \simeq 1 - \alpha$$

Ricaviamo dunque che una stima dell'intervallo di confidenza 1 − α per la differenza tra le medie $\mu_1 - \mu_2$ è:

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

Poniamo α=0.01 e stimiamo $\mu_1 - \mu_2$ per i due campioni che abbiamo a disposizione. Procediamo con la stima:

```
> alpha <-1 -0.99
>
> n1 <- length(voti)
> n2 <- length(voti2)
>
> m1 <- mean(voti)
> m2 <- mean(voti2)
>
> s1 <- sd(voti)
> s2 <- sd(voti2)
>
> #stima del limite inferiore
> m1-m2-qnorm(1-alpha/2,mean=0,sd=1)*sqrt(s1^2/n1+s2^2/n2)
[1] -1.590108
>
> #stima del limite superiore
> m1-m2+qnorm(1-alpha/2,mean=0,sd=1)*sqrt(s1^2/n1+s2^2/n2)
[1] 1.390108
```

Quindi per stimare la differenza tra le medie su una popolazione normale questi sono i metodi, ricordiamo che la **numerosità** dei campioni è importante nel secondo caso in quanto la varianza campionaria è asintoticamente corretta.

Essendo che i due intervalli **(-1.231699, 1.031699)** e **(-1.590108, 1.390108)** hanno un estremo negativo e l'altro positivo, allora in entrambi i casi, sia che prendo grado di fiducia 0.99 e sia che prendo grado di fiducia 0.95, risulta sempre contenere lo 0, quindi non posso dire che i voti di una classe sono migliori di un'altra.

5. *Verifica delle ipotesi con R*

Definiamo innanzitutto il concetto di **ipotesi statistica**: Un'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** (se si specifica o meno completamente la legge della popolazione).

L'ipotesi che si vuole verificare è denotata con H_0 e viene chiamata **ipotesi nulla**.
Il procedimento con il quale decidiamo, sulla base del campione, se accertare o meno H_0 si chiama **test di ipotesi**. Il test prevede che venga specificata un'ipotesi alternativa a quella sotto verifica, definita appunto **ipotesi alternativa**, ed è indicata con H_1.

Il problema consiste dunque nell'individuare un test capace di suddividere l'insieme dei possibili campioni in due sottoinsiemi che rappresentano la **regione di accettazione** e la **regione di rifiuto** dell'ipotesi nulla.
Se l'ipotesi nulla risulta falsa, quella alternativa risulta vera: l'ipotesi H_0 va verificata in alternativa all'ipotesi H_1.

Ci sono ovviamente dei margini di errore di cui tenere conto. La seguente immagine ci aiuta a capire:

	Rifiutare H_0	**Accettare H_0**
H_0 vera	Errore del I tipo Probabilità a	Decisione esatta Probabità 1 - a
H_0 falsa	Decisione esatta Probabilità 1 - β	Errore del II tipo Probabilità β

Ci sono due possibilità di errore quindi:

- **Rifiutare** l'ipotesi nulla H_0 nel caso in cui tale ipotesi sia **vera**; si dice allora che si commette un errore di **tipo I**, prob α
- **Accettare** l'ipotesi nulla H_0 nel caso in cui tale ipotesi sia **falsa**; si dice allora che si commette un errore di **tipo II**, prob β.

Dato che non è possibile rendere piccole entrambe le probabilità (se non in casi banali), la strategia che si usa è quella di fissare una delle due probabilità (α) e minimizzare l'altra.
Fissiamo l'errore più grave che in statistica corrisponde a **rifiutare il vero**.

Fissiamo α e costruiamo un test per minimizzare β.

Solitamente la probabilità di commettere un errore di tipo I si sceglie uguale a 0.05, 0.01, 0.001 ed il test viene rispettivamente detto **statisticamente significativo**, **statisticamente molto significativo** e **statisticamente estremamente significativo**. Infatti, quanto minore è il valore di α tanto **maggiore è la credibilità di un eventuale rifiuto dell'ipotesi nulla**.

I test statistici si dividono in due categorie:

- **Unilaterali** del tipo → $H_0 : \vartheta \leq \vartheta_0$, $H_1 : \vartheta > \vartheta_0$;
- **Bilaterali** del tipo → $H_0 : \vartheta = \vartheta_0$, $H_1 : \vartheta \neq \vartheta_0$.

Vediamo dunque la verifica delle ipotesi su una popolazione normale usando il nostro campione.

5.1 Popolazione normale

I Problemi esistenti:

1. Verifica di ipotesi sul valore medio μ nel caso in cui la varianza $\sigma 2$ della popolazione normale è nota;
2. Verifica di ipotesi sul valore medio μ nel caso in cui la varianza della popolazione normale è non nota;
3. Verifica di ipotesi sulla varianza $\sigma 2$ nel caso in cui il valore medio μ della popolazione normale è noto;
4. Verifica di ipotesi sulla varianza $\sigma 2$ nel caso in cui il valore medio della popolazione normale è non noto.

Il nostro caso di studio riguarda il secondo problema:

5.1.1 Test su μ con varianza σ2 non nota

Consideriamo le ipotesi:

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

La varianza non è nota quindi **entrambe le ipotesi sono composte**.

In analogia a quanto visto per gli intervalli di confidenza gioca un ruolo fondamentale la variabile aleatoria:

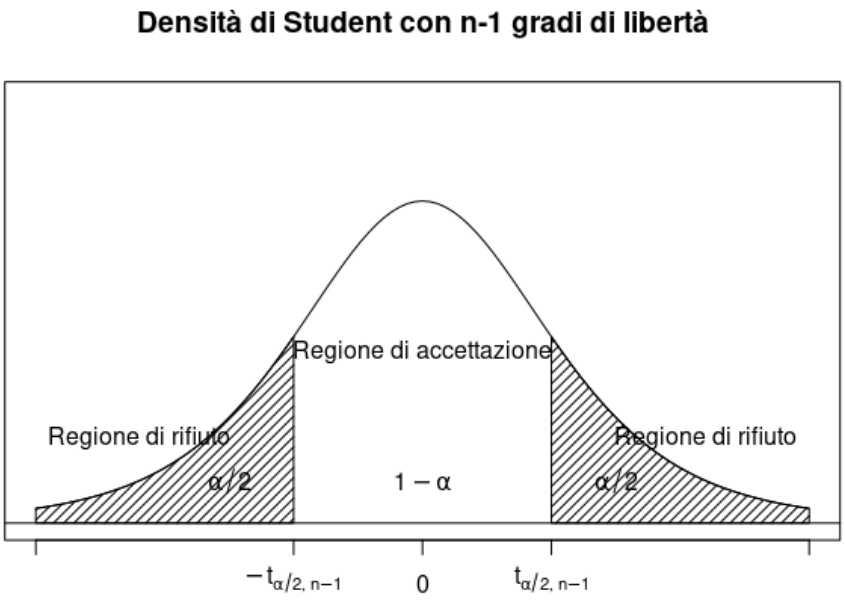
$$T_n = \frac{\bar{X}_n - \mu_0}{\frac{S_n}{\sqrt{n}}}$$

distribuita con legge di Student con n - 1 gradi di libertà

Il test bilaterale è dunque il seguente:

- si **accetta** H_0 se $-t_{\alpha/2,n-1} < \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} < t_{\alpha/2,n-1}$
- si **rifiuta** H_0 se :
  - $\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} < -t_{\alpha/2,n-1}$
  - $\frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} > t_{\alpha/2,n-1}$

Graficamente è rappresentata la densità normale standard e sono riportate le zone di accettazione e di rifiuto dell'ipotesi nulla per il test bilaterale:



Applichiamo dunque la verifica bilaterale al nostro campione.
Nel caso considerato un ateneo sostiene che la media dei voti dei suoi studenti è 24.
Prendendo un campione di 100 studenti si è verificata che la media è di 23.92 .
Si desidera utilizzare il test di misura con $a = 0.01$ per verificare l'ipotesi nulla $H_0 : \mu = 24$ in alternativa all'ipotesi $H_1 : \mu \neq 24$

```
> alpha <- 0.01
> mu0 <- 24
> sigma <- 3.4
>
> n<-length(voti)
>
> #z alpha/2
> qt(1- alpha/2,df=n-1)
[1] 2.626405
> #-z alpha/2
> -qt(1- alpha/2,df=n-1)
[1] -2.626405
>
>
> meancamp <-mean(voti)
> sdCamp <- sd(voti)
> (meancamp -mu0)/(sdCamp /sqrt(n))
[1] -0.234586
```

Quindi, essendo il valore risultate pari a -0.234586 è nella regione di accettazione. Quindi occorre accettare l'ipotesi nulla.

5.2 Criterio chi-quadrato

Il criterio del chi-quadrato ci permette di verificare se un dato campione osservato può essere stato estratto da una popolazione descritta da una variabile aleatoria X con una funzione di distribuzione $F_X(x)$.

Denotiamo con H_0 l'ipotesi nulla soggetta a verifica e con H_1 l'ipotesi alternativa tali che:

- 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 chi-quadrato di misura α mira a verificare l'ipotesi nulla, dove α è la probabilità massima di rifiutare l'ipotesi nulla quando essa è vera.

Bisogna determinare un test per determinare la regione di accettazione e di rifiuto dell'ipotesi nulla.
Suddividiamo dunque l'insieme dei valori che può assumere la variabile aleatoria in r sottoinsiemi, in modo tale che la probabilità p_i rappresenti la probabilità secondo la distribuzione ipotizzata che la variabile aleatoria assuma un valore appartenente al sottoinsieme I_i. Dal campione osserviamo le frequenze assolute $n_1, n_2, \dots, n_r$ in cui gli elementi del campione si distribuiscono rispettivamente in $I_1, I_2, \dots, I_r$.

Il criterio del chi-quadrato si basa sulla 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 di elementi del campione casuale che cadono nell'intervallo** I_i.

Se la variabile aleatoria X ha una funzione di distribuzione $F_x(x)$ con *k* parametri non noti, si dimostra che con n sufficientemente grande la funzione di distribuzione della statistica Q è **approssimabile con la funzione di distribuzione chi–quadrato con** $r - k - 1$ **gradi di libertà**.

È importante, inoltre, che **ogni classe abbia almeno 5 elementi**.

La definizione del test del chi-quadrato bilaterale è la seguente: Per un campione sufficientemente numeroso di ampiezza n, il test chi-quadrato di misura α è il seguente:

- Si accetta H_0 se $X^2_{1-a/2,r-k-1} < X^2 < X^2_{a/2,r-k-1}$
- Si rifiuta H_0 se
  - $X^2 < X^2_{1-a/2,r-k-1}$
  - $X^2 > X^2_{a/2,r-k-1}$

Con $X^2_{1-a/2,r-k-1}$ soluzione dell'equazione:

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

E con $X^2_{a/2,r-k-1}$ soluzione dell'equazione:

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

Vediamo quindi l'esempio per la normale.
Suddividiamo l'insieme in **5 sottoinsiemi**, determiniamo quindi i sottoinsiemi utilizzando i quantili:

```
> m<-mean(voti)
> d<-sd(voti)
>
> a<-numeric(4)
> for(i in 1:4)
+   a[i]<-qnorm(0.2*i,mean=m,sd=d)
> a
[1] 21.04985 23.05602 24.78398 26.79015
```

Dai valori uscenti si possono ricavare gli intervalli dei 5 sottoinsiemi:

1. (-infinito, 21.04985)
2. (21.04985, 23.05602)
3. (23.05602, 24.78398)
4. (24.78398, 26.79015)
5. (26.79015, +infinito)

Determiniamo dunque il **numero di elementi che cadono in ogni insieme**:

```
> r<-5
> nint <-numeric(r)
> nint[1]<-length(which(voti < a[1]))
> nint[2]<-length(which((voti >= a[1])&(voti <a[2])))
> nint[3]<-length(which((voti >= a[2])&(voti <a[3])))
> nint[4]<-length(which((voti >= a[3])&(voti <a[4])))
> nint[5]<-length(which(voti >= a[4]))
> nint
[1] 27 18  9 20 26
```

Calcoliamo dunque χ^2:

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

```
[1] 10.5
```

Per la distribuzione normale abbiamo due parametri non noti e quindi k=2. Quindi la statistica Q è approssimabile con la funzione di distribuzione chi-quadrato con $r - k - 1 = 2$ gradi di libertà. Usiamo $\alpha=0.05$:

```
> 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
```

Quindi abbiamo che la funzione di distribuzione del chi-quadrato **ha 2 gradi di libertà**.

I **limiti inferiori e superiore** sono rispettivamente **0.05063562 e 7.377759**.

Il valore del chi-quadrato 10.5 **non è compreso** in tale intervallo quindi per questo campione la distribuzione normale non può essere accettata.