Università degli Studi di Salerno

Dipartimento di Informatica
Corso di Laurea Magistrale in Informatica

Progetto di Statistica e Analisi dei Dati - Parte 2

Distribuzione normale

Costante Marco 0522501330

Anno Accademico 2021-2022

Indice

Capitolo 1

Inferenza statistica: distribuzione normale

L'indagine statistica è effettuata su un insieme di entità su cui si manifesta il fenomeno che si studia. Questo insieme è detto popolazione e può essere costituito da un numero finito oppure infinito di unità. La conoscenza delle caratteristiche di una popolazione finita può essere ottenuta osservando la totalità delle entità della popolazione oppure un sottoinsieme di questa. Una popolazione illimitata può invece essere studiata soltanto tramite un campione estratto dalla popolazione.

L'**inferenza statistica** ha lo scopo di estendere le misure ricavate dall'esame di un campione alla popolazione da cui il campione è stato estratto. Uno dei problemi centrali dell'inferenza statistica è il seguente:
si desidera studiare una popolazione descritta da una variabile aleatoria osservabile X la cui funzione di distribuzione ha una forma nota ma contiene un parametro non noto (o più parametri non noti).

L'inferenza statistica si basa su due metodi fondamentali di indagine: la **stima dei parametri** e la **verifica delle ipotesi**.
La stima dei parametri ha lo scopo di determinare i valori non noti dei parametri di una popolazione per mezzo dei corrispondenti parametri derivati dal campione estratto dalla popolazione.
La verifica delle ipotesi è un procedimento che consiste nel fare una congettura o un'ipotesi sul parametro non noto o sulla distribuzione di probabilità e nel decidere, sulla base del campione estratto se essa è accettabile.

L'analisi in questo lavoro verterà sullo studio di un campione di una popolazio-

ne avente distribuzione normale.
Tale distribuzione verrà analizzata e approfondita nel dettaglio attraverso l'uso del linguaggio e ambiente integrato R.

1.1 Distribuzione normale

La funzione di **distribuzione normale**, detta anche di **Gauss** o gaussiana, riveste estrema importanza nel calcolo delle probabilità e nella statistica poiché essa costituisce una distribuzione limite alla quale tendono varie altre funzioni di distribuzioni sotto opportune ipotesi.

Una variabile aleatoria X di **densità di probabilità**

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

si dice avere distribuzione normale di parametri μ e σ.

La densità è **simmetrica** rispetto all'asse $x = \mu$, risulta infatti $f_X(\mu - x) = f_X(\mu + x)$.
Ha una **forma a campana** rispetto a $x = \mu$.
Il **massimo** è in corrispondenza del punto $x =\mu$ ed è pari a $\frac{1}{\sigma\sqrt{2\pi}}$.
Ha due **flessi** in corrispondenza di μ - σ e $\mu + \sigma$.

Per una variabile aleatoria normale il valore medio e la varianza sono: $E(X) = \mu$ $VAR(X) = \sigma^2$, da cui segue che σ rappresenta la deviazione standard.
Per indicare una variabile aleatoria X che ha distribuzione normale di parametri μ e σ useremo la notazione $X \tilde{N}(\mu, \sigma)$.

Per calcolare in R la densità normale si usa la funzione **dnorm(x, mean=mu, sd=sigma)**.

Vediamo come varia la curva al variare dei parametri.
Notiamo che al variare del parametro μ quello che accade è che la curva viene traslata lungo l'asse delle ascisse, ma la forma non cambia.
Notiamo come dal parametro σ dipenda la larghezza della funzione: se aumenta σ la curva è sempre più piatta, al contrario invece si allunga verso l'alto. Questo succede in quanto il punto massimo è inversamente proporzionale a σ.

L'aria sottesa rimane sempre unitaria.

```
#DENSITA NORMALE VARIANDO MU

curve(dnorm(x, mean=0, sd=1), from=-6, to=6,
   xlab="x",ylab="f(x)",
   main="mu = -1,0,1; sigma = 1",col = "red")

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

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

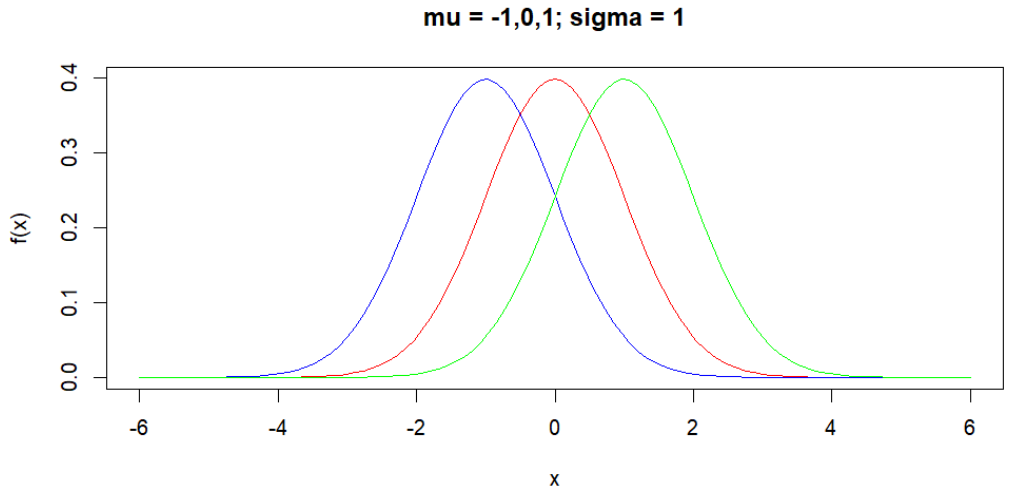


Figura 1.1: Densità normale variando mu

```
#DENSITA NORMALE VARIANDO SIGMA

curve(dnorm(x,mean=0, sd =0.5) ,from=-4, to=4,
   xlab="x", ylab="f(x)",main="mu=0;
   sigma = 0.5 ,1 ,1.5 ",col = "red")

curve(dnorm(x,mean=0, sd=1) ,from=-4, to=4,
   xlab="x",ylab="f(x)",
   add=TRUE,col="blue")

```

```
11  curve(dnorm(x,mean=0, sd =1.5) ,from=-4, to=4,
12    xlab="x",ylab="f(x)",
13    add=TRUE,col="green")
```

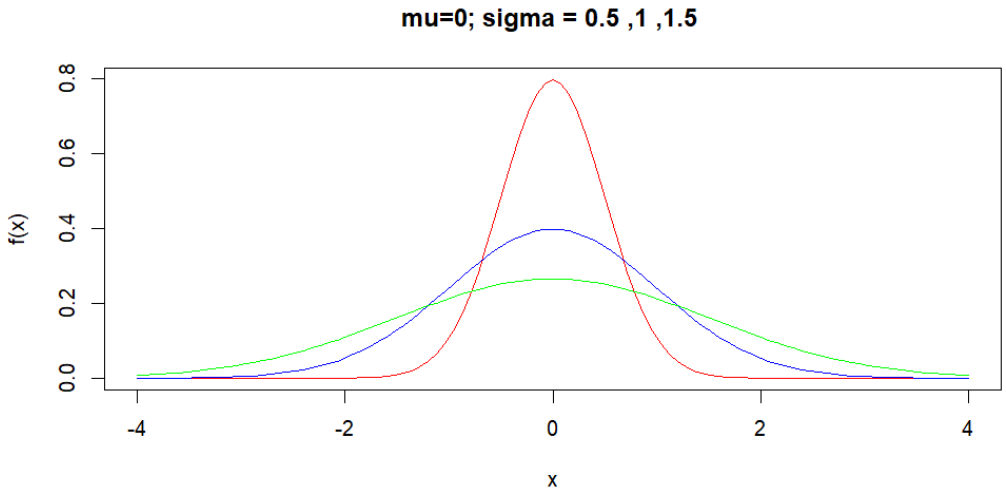


Figura 1.2: Densità normale variando sigma

Una **variabile aleatoria normale standard**, indicata solitamente con Z, può essere ottenuta da una variabile aleatoria normale non standard **standardizzando**, ovvero sottraendo il valore medio e dividendo per la deviazione standard.
Da cui si ottiene $X = \mu + \sigma Z$.

1.1.1 Funzione di distribuzione

La funzione di distribuzione di una variabile aleatoria $X \tilde{} N(\mu, \sigma)$ è:

$$F_X(X) = P(X <= x) = \int_{-\infty}^{x} f_X(y)dy = \Phi(\frac{x-\mu}{\sigma})$$

dove

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

è la funzione di distribuzione di una variabile aleatoria $Z \tilde{} N(0, 1)$, detta **normale standard**, ossia normale con valore medio nullo e varianza unitaria.
Pertanto, se $X \tilde{} N(\mu, \sigma)$ si ha:

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

In R la funzione di distribuzione di una variabile $X \tilde{} N(\mu, \sigma)$ si calcola tramite la funzione **pnorm(x, mean, sd, lower.tail=TRUE)**:

```
#FUNZIONE DISTRIBUZIONE

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="red")
arrows (-1,0.1,0.21,0.18, 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")
```

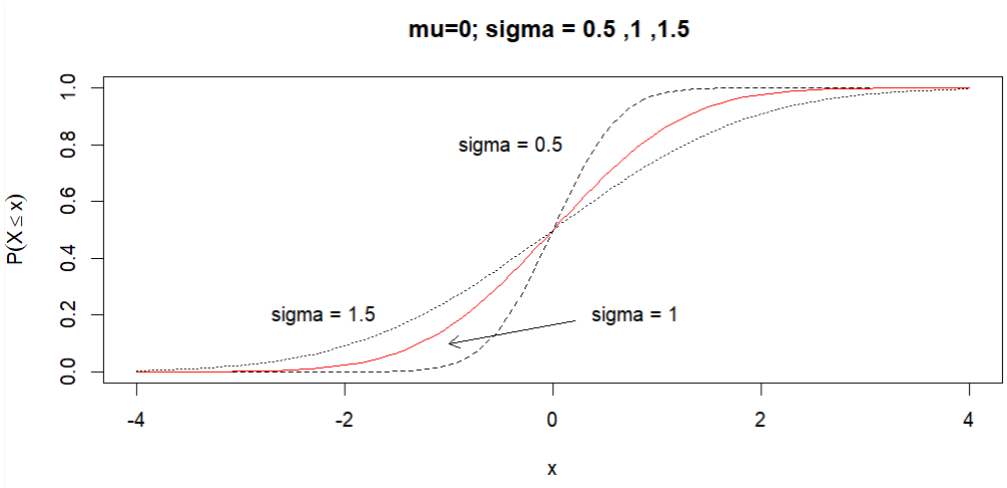


Figura 1.3: Funzione di distribuzione normale

1.1.2 Regola del 3σ

Per una qualsiasi variabile aleatoria normale $X \tilde{} N(\mu, \sigma)$ risulta:

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

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

La regola del 3σ permette di individuare l'intervallo $(\mu - 3\sigma, \mu + 3\sigma)$ in cui rappresentare la funzione densità di una variabile normale di valore medio μ e varianza σ^2 in maniera tale che l'area sottesa dalla curva sia circa unitaria e l'area delle code destra e sinistra sia trascurabile.

1.1.3 Simulazione della distribuzione normale in R

Sarebbe interessante confrontare la normale teorica con la densità simulata vedendo come si comporta l'istogramma del campione. Noteremmo come aumentando la sequenza in output, l'istogramma delle frequenze relative si avvicina sempre maggiormente alla curva teorica.
Simuliamo ora il comportamento di una distribuzione normale in R attraverso la funzione **rnorm(N, mean, sigma)**, che permette di generare una sequenza di numeri pseudocasuali:

```
#SIMULAZIONE

campione_normale<-rnorm(10000, mean=1000, sd=1.5)

mu<-round(mean(campione_normale), digits=4)
mu
[1] 999.9737

var<-round(var(campione_normale), digits=4)
var
[1] 2.2294

sigma<-round(sd(campione_normale), digits=4)
sigma
[1] 1.4931
```

Vediamo quindi i valori di media, varianza e deviazione standard sul campione simulato.

Ora è possibile disegnare la variabile aleatoria normale che approssima il nostro campione attraverso la funzione dnorm:

```
hist(campione_normale, freq=F, xlim=c(994,1006), ylim=c(0,0.5),
  breaks=100, xlab="x", yla="Istogramma",
  main="Densita' simulata N=10000")
```

Genera il grafico seguente dal quale notiamo quindi come, per n abbastanza grande, l'istogramma assuma una forma a campana che ricorda fortemente la distribuzione normale (Gaussiana).

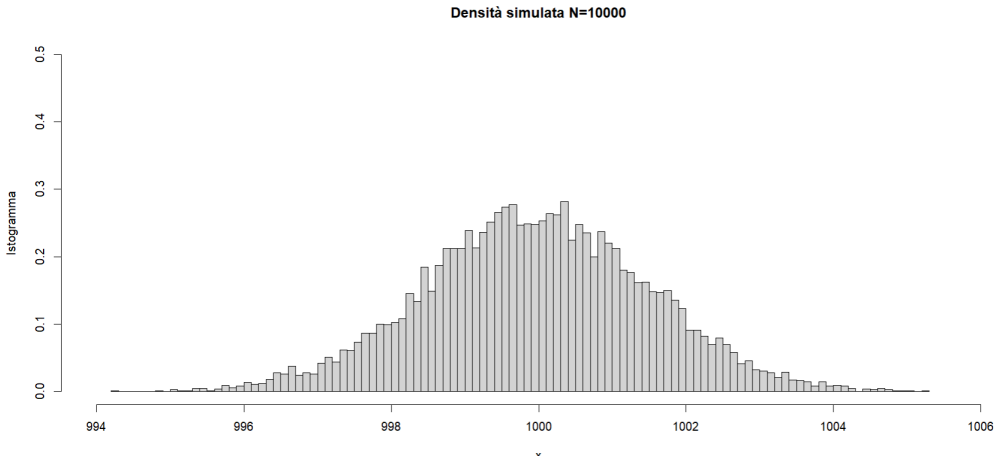


Figura 1.4: Simulazione distribuzione normale

Capitolo 2

Stima puntuale

Uno dei problemi centrali dell'inferenza statistica è il seguente: si desidera studiare una popolazione descritta da una variabile aleatoria osservabile X la cui funzione di distribuzione ha una forma nota ma contiene un parametro $\vartheta \in \theta$ non noto (o più parametri non noti).

2.1 Stimatori

Nei metodi di indagine dell'inferenza statistica si considera un campione casuale $X_1, ..., X_n$ di ampiezza n estratto dalla popolazione e si cerca di ottenere informazioni sui parametri non noti facendo uso di alcune variabili aleatorie, che sono funzioni misurabili del campione casuale, dette **statistiche** e **stimatori**.

Una **statistica** $t(X_1, ..., X_n)$ è una funzione misurabile e osservabile del campione casuale $X_1, ..., X_n$. Essendo la statistica osservabile, i valori da essa assunti dipendono soltanto dal campione osservato $x_1, ..., x_n$ estratto dalla popolazione e i parametri non noti sono presenti soltanto nella funzione di distribuzione della statistica.

Uno **stimatore** $\hat{\theta} = t(X_1, ..., X_n)$ è una funzione misurabile e osservabile del campione casuale $X_1, ..., X_n$ i cui valori possono essere usati 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.

Statistiche tipiche sono la **media campionaria** e la **varianza campionaria**.

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

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

Quindi al crescere dell'ampiezza del campione la media campionaria fornisce una stima sempre più accurata del valore medio della popolazione. Inoltre, dal teorema centrale di convergenza scaturisce che per n sufficientemente grande (ossia per campioni di grande ampiezza) la funzione di distribuzione della media campionaria X è approssimativamente normale con valore medio μ e varianza $\frac{\sigma^2}{n}$.

2.2 Metodi per la ricerca di stimatori

Lo scopo di un decisore, dopo aver osservato i valori del campione casuale, è quello di stimare i parametri non noti della popolazione. I principali metodi di stima puntuale dei parametri sono il **metodo dei momenti** e il **metodo della massima verosimiglianza**.

2.2.1 Metodo dei momenti

Definiamo innanzitutto il concetto di **momenti campionari**.
Si definisce momento campionario $r - esimo$ relativo ai valori osservati $(x_1, ..., x_n)$ del campione casuale il valore:

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

Si nota quindi che il momento campionario $r–esimo$ è la media aritmetica delle potenze $r–esime$ delle n osservazioni effettuate sulla popolazione.
Automaticamente se $r = 1$, il momento campionario coincide con il valore osservato dalla media campionaria $\bar{X}$.
Se esistono k parametri da stimare, uguagliamo i primi k momenti della popolazione con i corrispondenti momenti del campione casuale. Se i primi k momenti esistono e sono finiti, il metodo dei momenti si riduce in una soluzione di sistema di k equazioni:

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

Le incognite del sistema sono i parametri $\vartheta_1, ..., \vartheta_k$ e sono presenti a sinistra del sistema.

Popolazione normale

Nel caso di popolazione normale si è interessati a stimare i parametri μ e σ^2 .

Sapendo che $E(x) = \mu$ e $E(X^2) = \sigma^2 + \mu^2$, si ha:

$\hat{\mu} = \frac{x_1+...+x_n}{n}$

$\hat{\sigma^2} + \hat{\mu}^2 = \frac{x_1^2+...+x_n^2}{n}$

Utilizzando la prima nella seconda si ricava:

$\hat{\sigma^2} = \frac{(n-1)s^2}{n}$

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

Vediamo quindi in R:

```
#STIMA METODO DEI MOMENTI

stima_mu_momenti<-mean(campione_normale)
stima_mu_momenti
[1] 999.9653

stima_sigma2_momenti<-(length(campione_normale)-1)*
                var(campione_normale)/length(campione_normale)
stima_sigma2_momenti
[1] 2.274796
```

2.2.2 Metodo della massima verosimiglianza

Vediamo ora il metodo della massima verosimiglianza, che risulta essere molto diffuso e preferito a quello dei momenti. Introduciamo il concetto di **funzione di verosimiglianza**.

Sia $X_1, ..., X_n$ un campione casuale di ampiezza n estratto dalla popolazione. La funzione di verosimiglianza $L(\vartheta_1, ..., \vartheta_k) = L(\vartheta_1, ..., \vartheta_k; x_1, ..., x_n)$ del campione osservato $(x_1, ..., x_k)$ è la funzione di probabilità congiunta (nel caso discreto), o la funzione di densità di probabilità congiunta(nel caso continuo) del campione casuale $X_1, ..., X_n$, ossia:

$$L(\vartheta_1, ..., \vartheta_k) = L(\vartheta_1, ..., \vartheta_k; x_1, ..., x_n) = f(x_1; \vartheta_1, \vartheta_2, \vartheta_k)...f(x_n; \vartheta_1, \vartheta_2, \vartheta_k)$$

Il metodo consiste nel **massimizzare** questa funzione rispetto ai parametri $\vartheta_1, ..., \vartheta_k$ si cerca di determinare da quale funzione di probabilità (densità) congiunta è più verosimile che provenga il campione osservato.
Si cercano i vari ϑ_k in modo tale che spieghino meglio il campione osservato. I valori stimati, indicati con $\hat{\vartheta_k}$ sono detti stime di massima verosimiglianza e anche in questo caso dipendono dal campione.

Popolazione normale

Si desidera determinare lo stimatore di massima verosimiglianza dei parametri μ e σ^2 di una popolazione normale caratterizzata da funzione densità di probabilità:

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

Si ha:

$$L(\mu, \sigma^2) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2})\ (\mu \in R, \sigma > 0)$$

Le stime di massima verosimiglianza dei parametri μ e σ^2 sono rispettivamente:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

Lo stimatore di massima verosimiglianza e dei momenti del valore medio μ è la me-

dia campionaria $\bar{X}$. Invece lo stimatore di massima verosimiglianza e dei momenti della varianza σ^2 è $(n-1)S^2/n$.

2.3 Proprietà e stimatore asintoticamente corretto

Dato che per stimare un parametro di una popolazione ci possono essere diversi stimatori, sono definite alcune proprietà.
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 $(n-1)S^2/n$ 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**.

Capitolo 3

Intervalli di confidenza

Alla stima puntuale di un parametro non noto di una popolazione spesso si preferisce sostituire un intervallo di valori, detto **intervallo di confidenza**, ossia si cerca di determinare in base ai dati del campione, due limiti entro i quali sia compreso il parametro non noto con un certo coefficiente di confidenza.

Intervallo di confidenza: Fissato un coefficiente di confidenza $1-\alpha(0<\alpha<1)$ se è possibile scegliere due statistiche $\underline{C}_n$ e $\bar{C}_n$ in modo tale che:

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

allora $(\underline{C}_n, \bar{C}_n)$ è un intervallo di confidenza di grado $1-\alpha$ per il parametro ϑ. Le statistiche $(\underline{C}_n, \bar{C}_n)$ sono dette limite inferiore e superiore dell'intervallo di confidenza.

Se $g1(x)$ e $g2(x)$ sono i valori assunti dalle statistiche $\underline{C}_n$ e $\bar{C}_n$ per il campione osservato allora l'intervallo $(g1(x), g2(x))$ è detto **stima dell'intervallo di confidenza** di grado $1-\alpha$ per ϑ ed i punti finali $g1(x)$ e $g2(x)$ di tale intervallo sono detti **stima del limite inferiore** e **stima del limite superiore** dell'intervallo di confidenza.
In generale esistono numerosi intervalli di confidenza e la scelta deve essere effettuata in base ad alcune proprietà statistiche. Ad esempio, fissato un coefficiente di confidenza di $1-\alpha$, alcune proprietà desiderabili sono che la lunghezza dell'intervallo di confidenza sia la più piccola possibile oppure che la lunghezza media di tale intervallo sia la più piccola possibile.

3.1 Metodo pivotale

Questo metodo consiste nel determinare una variabile aleatoria di pivot $\delta(X1, ..., X_n; \vartheta)$ che:

- dipende dal campione casuale $X1, ..., X_n$;
- dipende dal parametro non noto ϑ;
- la sua funzione di distribuzione non contiene il parametro ϑ da stimare;

La variabile aleatoria di pivot non è una statistica poiché dipende dal parametro non noto ϑ e quindi non è osservabile.

3.2 Popolazione normale

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

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

3.2.1 Intervallo di confidenza per μ con σ^2 nota

Utilizzando il metodo pivotale e consideriamo la variabile aleatoria:

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

che è distribuita normalmente con valore medio nullo e varianza unitaria, ossia è una normale standard. Tale variabile aleatoria dipende dal campione casuale e dal parametro non noto μ (la varianza è nota) e la sua legge di probabilità non dipende dal parametro non noto. Quindi, Z_n può essere interpretata come una variabile

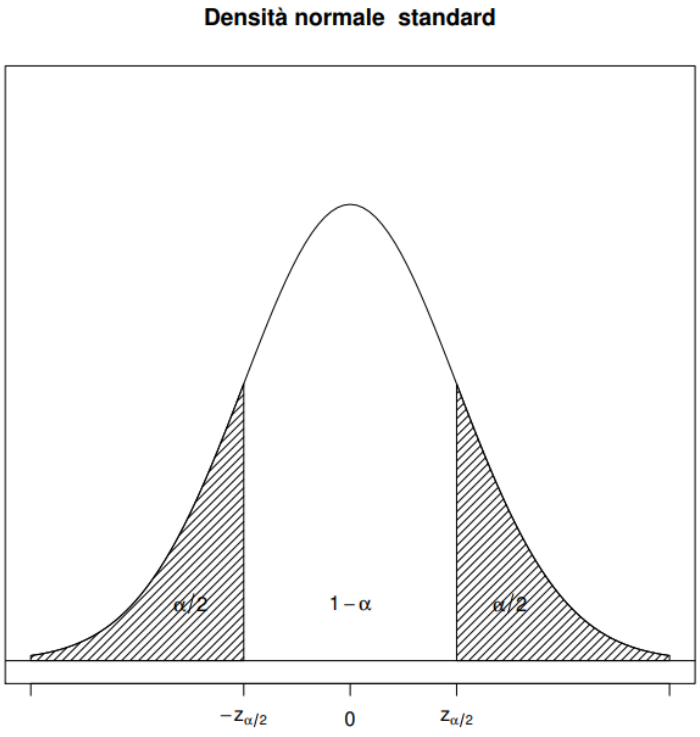
aleatoria di pivot.

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

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

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

Graficamente:



A valori sempre più piccoli di α, corrispondono lunghezze di intervalli di confidenza sempre più ampi.

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

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

dove $\hat{x}_n = \frac{x_1+...+x_n}{n}$ denota la media campionaria delle n osservazioni.

Vediamo in R, poniamo $\alpha = 0.05$ e supponiamo che la varianza nota sia $\sigma^2 = 2.25$ e quindi $\sigma = 1.5$, stimiamo il parametro con l'intervallo di confidenza che abbiamo trovato:

```
#INTERVALLI PER MU E SIGMA2 NOTA

   alpha <-1-0.95
deviazione_standard<-1.5
n<-length(campione_normale)

#stima del limite inferiore
mean(campione_normale)-qnorm(1-alpha/2,mean=0,sd=1)*
              deviazione_standard/sqrt(n)
[1] 999.9359

#stima del limite superiore
mean(campione_normale)+qnorm(1-alpha/2,mean=0,sd=1)*
              deviazione_standard/sqrt(n)
[1] 999.9947
```

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

Per determinare un intervallo di confidenza di grado $1-$ per il valore medio μ nel caso in cui la varianza della popolazione normale non è nota, utilizziamo il metodo pivotale e consideriamo la variabile aleatoria di pivot:

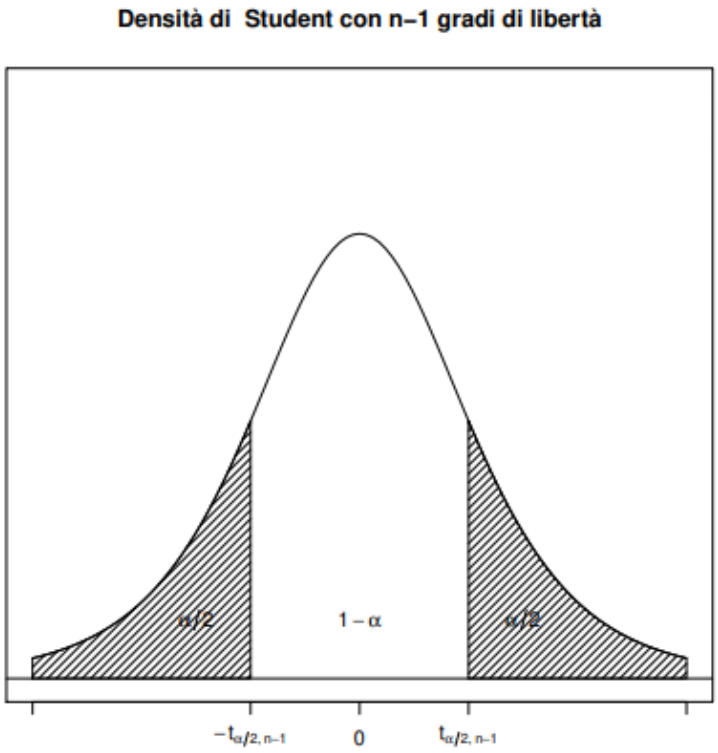
$T_n = \frac{\bar{X}_n-\mu}{S_n/\sqrt{n}}$

dove S_n^2 rappresenta la varianza campionaria.
La variabile aleatoria T_n dipende dal campione casuale e dal parametro non noto μ e la sua legge di probabilità non dipende dal parametro non noto. Quindi, T_n può essere interpretata come una variabile aleatoria di pivot, inoltre è distribuita con legge di Student con $n-1$ gradi di libertà
Scegliamo quindi $\alpha_1 = -t_{\alpha/2,n-1}$ e $\alpha_2 = t_{\alpha,n-1}$ tale che:

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

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

Graficamente:



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

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

dove $\hat{x}_n = \frac{x_1+\ldots+x_n}{n}$ denota la media campionaria delle n osservazioni e s_n la deviazione standard.

Vediamo in R, poniamo $\alpha = 0.05$ e stimiamo il parametro con l'intervallo di confidenza che abbiamo trovato:

```
#INTERVALLI PER MU E SIGMA2 NON NOTA

alpha<-1-0.95
deviazione_standard<-sd(campione_normale)
```

```
5  deviazione_standard
6  [1] 1.508318
7
8  n<-length(campione_normale)
9
10 #stima del limite inferiore
11 mean(campione_normale)-qt(1-alpha/2,df=n-1)*
12             deviazione_standard/sqrt(n)
13 [1] 999.9357
14
15 #stima del limite superiore
16 mean(campione_normale)+qt(1-alpha/2,df=n-1)*
17             deviazione_standard/sqrt(n)
18 [1] 999.9948
```

3.2.3 Intervallo di confidenza per σ^2 con μ nota

Per determinare un intervallo di confidenza di grado $1-\alpha$ per la varianza nel caso in cui il valore medio della popolazione normale è noto, utilizziamo nuovamente il metodo pivotale e consideriamo la variabile aleatoria di pivot:

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

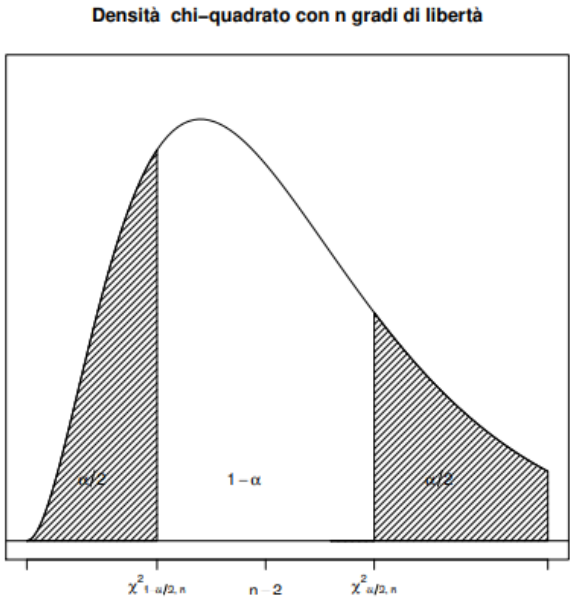
Tale variabile dipende dal campione casuale e dal parametro non noto σ^2 ed è distribuita con legge chi–quadrato con n gradi di libertà, essendo costituita dalla somma dei quadrati di n variabili aleatorie normali standard.
Inoltre, dipende dal campione casuale e dal parametro non noto e la sua legge di probabilità non contiene il parametro non noto. Quindi, può essere interpretata come una variabile di pivot.

Scegliendo nel metodo pivotale $\alpha_1 = \chi^2_{1-\alpha/2,n}$ e $\alpha_2 = \chi^2_{\alpha/2,n}$, tale che:

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

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

Graficamente:



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

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

dove $\bar{x}_n$ e s_n^2 rappresentano media e varianza campionaria.

Poniamo $\alpha = 0.05$ e supponiamo che la media nota sia $\mu = 1000$, stimiamo il parametro con l'intervallo di confidenza che abbiamo trovato:

```
#INTERVALLI PER SIGMA2 CON MU NOTA

n<-length(campione_normale)
mu<-1000
alpha<-1-0.95

#stima del limite inferiore
((n-1)*var(campione_normale)+n*
     (mean(campione_normale)-mu)**2)/
     qchisq(1-alpha/2,df=n)
[1] 2.21421


#stima del limite superiore
((n-1)*var(campione_normale)+n*
```

```
16          (mean(campione_normale)-mu)**2)/
17          qchisq(alpha/2,df=n)
18      [1] 2.34043
```

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

Per determinare un intervallo di confidenza di grado $1-\alpha$ per la varianza nel caso in cui il valore medio della popolazione normale non è noto, consideriamo la variabile aleatoria di pivot:

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

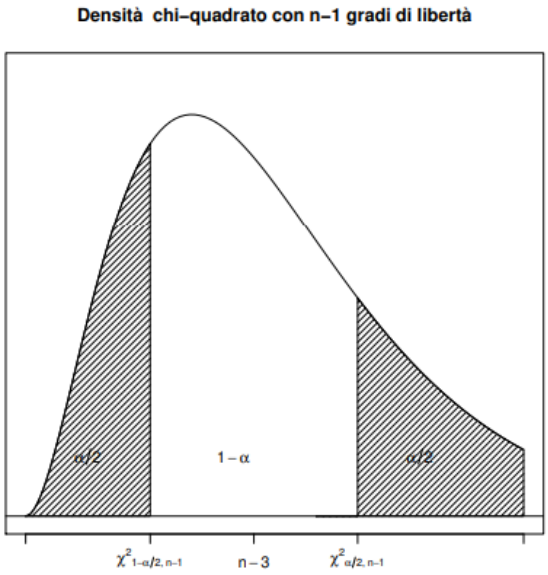
Tale variabile dipende dal campione casuale e dal parametro non noto σ^2 ed è distribuita con legge chi–quadrato con $n-1$ gradi di libertà.

Scegliendo nel metodo pivotale $\alpha_1 = \chi^2_{1-\alpha/2,n-1}$ e $\alpha_2 = \chi^2_{\alpha/2,n-1}$, tale che:

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

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

Graficamente:



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

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

Dove s_n^2 rappresenta la varianza campionaria delle n osservazioni.

Poniamo $\alpha = 0.05$ stimiamo il parametro con l'intervallo di confidenza che abbiamo trovato:

```
#INTERVALLI PER SIGMA2 CON MU NON NOTA

n<-length(campione_normale)
alpha <-1-0.95

#stima limite inferiore
(n-1)*var(campione_normale)/qchisq(1-alpha/2,df=n -1)
[1] 2.213255

#stima limite superiore
(n-1)*var(campione_normale)/qchisq(alpha/2,df=n-1)
[1] 2.339427
```

3.3 Confronto tra popolazioni normali

Consideriamo due campioni, $X_1, ..., X_{n1}$ e $Y_1, ..., Y_{n2}$, casuali e indipendenti, di ampiezza $n1$ e $n2$ estratti rispettivamente da due popolazioni normali.
Vogliamo analizzare due problemi:

- determinare un intervallo di confidenza di grado $1-\alpha$ per $\mu_1 - \mu_2$ quando entrambe le varianze sono note;

- determinare un intervallo di confidenza di grado $1-\alpha$ per $\mu_1 - \mu_2$ quando entrambe le varianze non sono note;

3.3.1 Intervallo di confidenza per μ_1 - μ_2 con σ_1^2 e σ_2^2 note

Siano $\bar{X}_{n1}$ e $\bar{Y}_{n2}$ rispettivamente le medie campionarie delle due popolazioni normali. Poiché per ipotesi i campioni casuali sono indipendenti la statistica $\bar{X}_{n1} - \bar{Y}_{n2}$ è distribuita normalmente con valore medio e varianza:

$$E(\bar{X}_{n1} - \bar{Y}_{n2}) = \mu_1 - \mu_2$$

$$var(\bar{X}_{n1} - \bar{Y}_{n2}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

ottenute con la proprietà di linearità del valore medio e le proprietà della varianza per combinazioni lineari di variabili aleatorie indipendenti.

Per determinare un intervallo di confidenza di grado $1 - \alpha$ per $\mu_1 - \mu_2$ quando entrambe le varianze sono note usiamo la variabile aleatoria di pivot:

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

Tale variabile è di pivot in quanto dipende dal parametro non noto, dipende dal campione è caratterizzata da una densità normale standard.
Una stima dell'intervallo di confidenza $1 - \alpha$ per la differenza tra le medie è:

$$\bar{x}_{n1} - \bar{y}_{n2} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \bar{x}_{n1} - \bar{y}_{n2} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

dove $\bar{x}_n$ e $\bar{y}_n$ sono le medie campionarie delle due osservazioni.

Poniamo $\alpha = 0.05$ e stimiamo $\mu_1 - \mu_2$ per i due campioni che abbiamo a disposizione sapendo che le varianze note sono $\sigma_1^2 = 2.25$ e $\sigma_2^2 = 4$, mentre la numerosità del primo campione è pari a 10000, mentre quella del secondo 9000:

```
#CONFRONO TRA POPOLAZIONI NORMALI CON VARIANZE NOTE

#generiamo un secondo campione normale
campione_normale2<-rnorm(9000, mean = 980, sd = 2)

alpha<-1-0.95

n1<-length(campione_normale)
n2<-length(campione_normale2)
```

```
10
11  m1<-mean(campione_normale)
12  m2<-mean(campione_normale2)
13
14  sigma1<-2.25
15  sigma2<-4
16
17  #stima del limite inferiore
18  m1-m2-qnorm(1-alpha/2,mean=0, sd=1)*sqrt(sigma1^2/n1+sigma2^2/n2)
19  [1] 19.91981
20
21  #stima del limite superiore
22  m1-m2+qnorm(1-alpha/2,mean=0, sd=1)*sqrt(sigma1^2/n1+sigma2^2/n2)
23  [1] 20.10715
```

3.3.2 Intervallo di confidenza per μ_1 - μ_2 con σ_1^2 e σ_2^2 non note

Vogliamo determinare un intervallo di confidenza di grado $1-\alpha$ per la differenza $\mu_1 - \mu_2$ delle due popolazioni per grandi valori di $n1$ e $n2$. Denotiamo con S_{n1}^2 e S_{n2}^2 le varianze campionarie delle due popolazioni normali.
Le varianze campionarie delle due popolazioni normali sono stimatori corretti e consistenti delle varianze delle due popolazioni. Quindi quando le ampiezze dei campioni sono grandi, applicando il metodo pivotale in forma approssimata so ha:

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

Una stima dell'intervallo di confidenza $1-\alpha$ per la differenza tra le medie è:

$$\bar{x}_{n1} - \bar{y}_{n2} - z_{\alpha/2}\sqrt{\frac{S_{n1}^2}{n_1} + \frac{\tilde{S_{n2}^2}}{n_2}} < \mu_1 - \mu_2 < \bar{x}_{n1} - \bar{y}_{n2} + z_{\alpha/2}\sqrt{\frac{S_{n1}^2}{n_1} + \frac{\tilde{S_{n2}^2}}{n_2}}$$

dove $\bar{x}_n$ e $\bar{y}_n$ sono le medie campionarie delle due osservazioni e S_{n1}^2, $\tilde{S_{n2}^2}$ le varianze campionarie delle due osservazioni.

Poniamo $\alpha = 0.05$ e stimiamo $\mu_1 - \mu_2$ per i due campioni che abbiamo a disposizione:

```
#CONFRONO TRA POPOLAZIONI NORMALI CON VARIANZE NON NOTE

alpha<-1-0.95

n1<-length(campione_normale)
n2<-length(campione_normale2)

m1<-mean(campione_normale)
m2<-mean(campione_normale2)

sd1<-sd(campione_normale)
sd2<-sd(campione_normale2)

#stima del limite inferiore
m1-m2-qnorm(1-alpha/2,mean=0, sd=1)*sqrt(sd1^2/n1+sd2^2/n2)
[1] 19.88816

#stima del limite superiore
m1-m2+qnorm(1-alpha/2,mean=0, sd=1)*sqrt(sd1^2/n1+sd2^2/n2)
[1] 19.98998
```

Capitolo 4

Verifica delle ipotesi

Dopo la stima dei parametri il passo successivo è la **verifica delle ipotesi**. La verifica delle ipotesi interviene ogni volta che si ha il bisogno di predire qualcosa.
In generale gli elementi che costituiscono il punto di partenza del procedimento di verifica delle ipotesi sono una popolazione descritta da una variabile aleatoria X caratterizzata da una funzione di probabilità o densità di probabilità $f(x;\vartheta)$, un'ipotesi su di un parametro non noto della popolazione ed un campione casuale estratto dalla popolazione. Si può ora definire il concetto di ipotesi statistica.

Ipotesi statistica: è un'affermazione o una congettura sul 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 in genere denotata con H_0 e viene chiamata **ipotesi nulla**. Si chiama test di ipotesi il procedimento o regola con cui si decide, sulla base dei dati del campione, se accettare o rifiutare H_0. La costruzione del test richiede la formulazione, in contrapposizione all'ipotesi nulla, di una proposizione alternativa. Questa proposizione prende il nome di **ipotesi alternativa** e viene indicata con H_1.
L'ipotesi nulla, cioè ipotesi soggetta a verifica, si ha quando $\vartheta \in \theta_0$ e l'ipotesi alternativa si ha quando $\vartheta \in \theta_1$ e si scrive:

$H_0 : \vartheta \in \theta_0$
$H_1 : \vartheta \in \theta_1$

Avendo denotato con θ_0 e θ_1 i due sottoinsiemi disgiunti dello spazio dei parametri.

L'obiettivo è determinare un test ψ che permetta di suddividere, l'insieme dei possibili campioni in due sottoinsiemi:

- Una regione di accettazione A dell'ipotesi nulla;
- Una regione di rifiuto R dell'ipotesi nulla;

Si potrebbe incorrere in due tipi di errori:

- Rifiutare l'ipotesi nulla H_0 nel caso in cui tale ipotesi sia vera;
- Accettare l'ipotesi nulla H_0 nel caso in cui tale ipotesi sia falsa;

	Rifiutare $\mathbf{H_0}$	Accettare $\mathbf{H_0}$
$\mathbf{H_0}$ vera	Errore del I tipo Probabilità α	Decisione esatta Probabilità $1-\alpha$
$\mathbf{H_0}$ falsa	Decisione esatta Probabilità $1-\beta$	Errore del II tipo Probabilità β

Altro concetto rilevante è quello di **misura della regione critica**, ovvero la probabilità massima di rifiutare l'ipotesi nulla quando essa è vera.

I test statistici sono di due tipi:

- Test bilaterali del tipo:
  $H_0 : \vartheta = \vartheta_0$
  $H_1 : \vartheta \neq \vartheta_0$
- Test unilaterali, divisi in:
  - Test unilaterale sinistro:
    $H_0 : \vartheta \leq \vartheta_0$
    $H_1 : \vartheta > \vartheta_0$
  - Test unilaterale destro:
    $H_0 : \vartheta \geq \vartheta_0$
    $H_1 : \vartheta < \vartheta_0$

4.1 p-value

Il p-value è definito come la probabilità, supposta vera l'ipotesi nulla, che la statistica del test assuma un valore uguale o più estremo da quello effettivamente osservato. Il p-value si basa su una statistica del test, che dipende dal campione osservato e

dal test statistico considerato. Nell'effettuare un test statistico è importante fissare il livello di significatività α prima di calcolare il p-value. Se si calcola prima il p-value, il decisore potrebbe scegliere α in funzione del risultato desiderato in modo da rigettare l'ipotesi nulla H_0.
Calcolando il p-value è possibile comportarsi come segue:

- se $p > \alpha$ l'ipotesi H_0 non può essere rifiutata;
- se $p <= \alpha$ l'ipotesi H_0 deve essere rifiutata;

4.2 Popolazione normale

Utilizzando test bilaterali e unilaterali, desideriamo affrontare i seguenti problemi:

- Verifica di ipotesi sul valore medio μ nel caso in cui la varianza σ^2 popolazione normale è nota;
- Verifica di ipotesi sul valore medio μ nel caso in cui la varianza della popolazione normale è non nota;
- Verifica di ipotesi sulla varianza σ^2 nel caso in cui il valore medio μ della popolazione normale è noto;
- Verifica di ipotesi sulla varianza σ^2 nel caso in cui il valore medio μ della popolazione normale non è noto;

4.2.1 Test su μ con varianza σ^2 nota

TEST BILATERALE

Consideriamo le ipotesi:

$H_0 : \mu = \mu 0$
$H_1 : \mu \neq \mu 0$

Essendo la varianza nota, H_0 è semplice, mentre l'ipotesi H_1 è composita.
Quando H_0 è vera, gioca un ruolo fondamentale la variabile aleatoria:

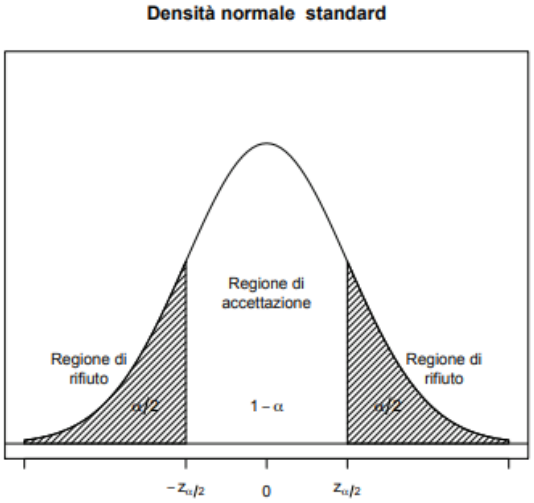
$Z_n = \frac{\bar{X}_n - \mu_0}{\sigma\sqrt{n}}$

distribuita secondo una normale standard. Z è una statistica, e non una variabile di pivot, poiché dipende esclusivamente dal campione casuale.

Il test bilaterale ψ di misura α per le ipotesi considerate è:

- si accetti H_0 se
  $-z_{\alpha/2} < \frac{\bar{x}_n - \mu_0}{\sigma\sqrt{n}} < z_{\alpha/2}$
- si rifiuti H_0 se
  $\frac{\bar{x}_n - \mu_0}{\sigma\sqrt{n}} < -z_{\alpha/2}$ oppure $\frac{\bar{x}_n - \mu_0}{\sigma\sqrt{n}} > z_{\alpha/2}$

Graficamente:



p-value: $p - value = 2[1 - P(Z_n <= |\frac{\bar{x}_n - \mu_0}{\sigma\sqrt{n}}|)]$

TEST UNILATERALE SINISTRO

Consideriamo le ipotesi:

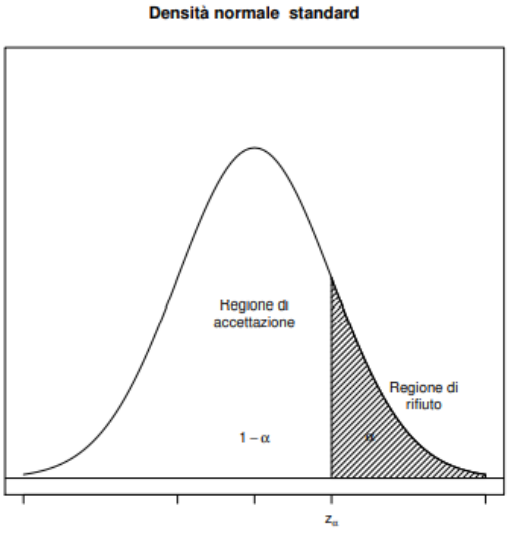
$H_0 : \mu <= \mu 0$
$H_1 : \mu > \mu 0$

Le ipotesi sono entrambe composite.
Il test unilaterale sinistro ψ di misura α per le ipotesi considerate è:

- si accetti H_0 se
  $\frac{\bar{x}_n - \mu_0}{\sigma\sqrt{n}} < z_\alpha$

- si rifiuti H_0 se
  $\frac{\bar{x}_n - \mu_0}{\sigma\sqrt{n}} > z_\alpha$

Graficamente:



p-value:$1 - P(Z_n <= \frac{\bar{x}_n - \mu_0}{\sigma\sqrt{n}})$

TEST UNILATERALE DESTRO

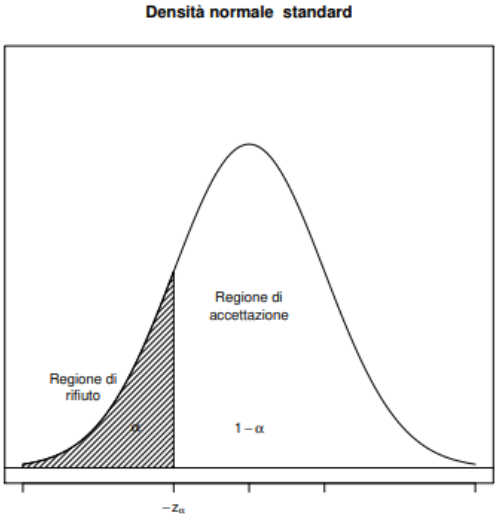
Consideriamo le ipotesi:

$H_0 : \mu >= \mu 0$
$H_1 : \mu < \mu 0$

Le ipotesi sono entrambe composite.
Il test unilaterale destro ψ di misura α per le ipotesi considerate è:

- si accetti H_0 se
  $\frac{\bar{x}_n - \mu_0}{\sigma\sqrt{n}} > -z_\alpha$
- si rifiuti H_0 se
  $\frac{\bar{x}_n - \mu_0}{\sigma\sqrt{n}} < -z_\alpha$

Graficamente:



p-value:$P(Z_n <= \frac{\bar{x}_n - \mu_0}{\sigma\sqrt{n}})$

Vediamo in R i vari test:

```
#IPOTESI BILATERALE

alpha<-0.05
mu0<-999.98
sigma<-2.27

qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.959964

n<-length(campione_normale)
meancamp<-mean(campione_normale)
z_os<-(meancamp-mu0)/(sigma/sqrt(n))
z_os
[1] 0.4279424

p_value<-2*(1-pnorm(z_os, mean=0, sd=1))
p_value
[1] 0.6686931
```

Si nota che z_{os} cade all'interno della zona di accettazione il il $p - value > \alpha$ ci consiglia di accettare l'ipotesi.

```
#IPOTESI UNILATERALE DX

alpha<-0.05
mu0<-999.98
sigma<-2.27

qnorm(alpha/2,mean=0,sd=1)
[1] -1.959964

n<-length(campione_normale)
meancamp<-mean(campione_normale)
z_os<-(meancamp-mu0)/(sigma/sqrt(n))
z_os
[1] 0.4279424

p_value<-pnorm(z_os, mean=0, sd=1)
p_value
[1] 0.6656535
```

Si nota che z_{os} cade all'interno della zona di accettazione il il $p - value > \alpha$ ci consiglia di accettare l'ipotesi.

```
#IPOTESI UNILATERALE SX

alpha<-0.05
mu0<-999.98
sigma<-2.27

qnorm(1-alpha/2,mean=0,sd=1)
[1] 1.959964

n<-length(campione_normale)
meancamp<-mean(campione_normale)
z_os<-(meancamp-mu0)/(sigma/sqrt(n))
z_os
[1] 0.4279424

p_value<-1-pnorm(z_os, mean=0, sd=1)
p_value
[1] 0.3343465
```

Si nota che z_{os} cade all'interno della zona di accettazione il il $p - value > \alpha$ ci consiglia di accettare l'ipotesi.

4.2.2 Test su μ con varianza σ^2 non nota

TEST BILATERALE

Consideriamo le ipotesi:

$H_0 : \mu = \mu 0$
$H_1 : \mu \neq \mu 0$

Essendo la varianza non nota entrambe le ipotesi sono composite.
Quando H_0 è vera, gioca un ruolo fondamentale la variabile aleatoria:

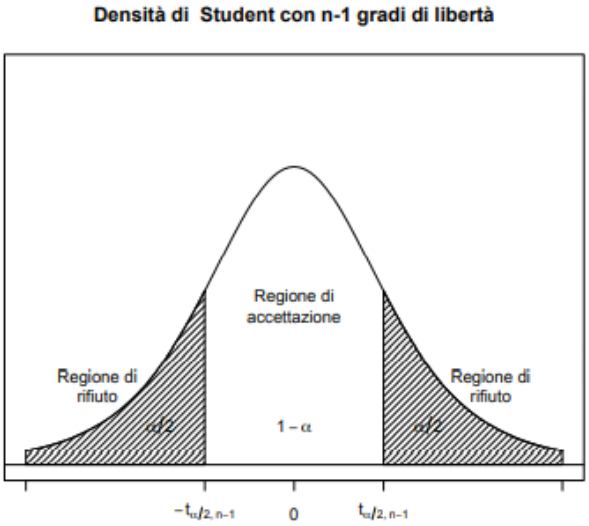
$T_n = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$

distribuita secondo una legge di Student con n-1 gradi di libertà, e non una variabile di pivot, poiché dipende esclusivamente dal campione casuale.

Il test bilaterale ψ di misura α per le ipotesi considerate è:

- si accetti H_0 se
  $-t_{\alpha/2,n-1} < \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}} < t_{\alpha/2,n-1}$
- si rifiuti H_0 se
  $\frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}} < -t_{\alpha/2,n-1}$ oppure $\frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}} > t_{\alpha/2,n-1}$

Graficamente:



p-value: $p - value = 2[1 - P(T_n <= |\frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}|)]$

TEST UNILATERALE SINISTRO

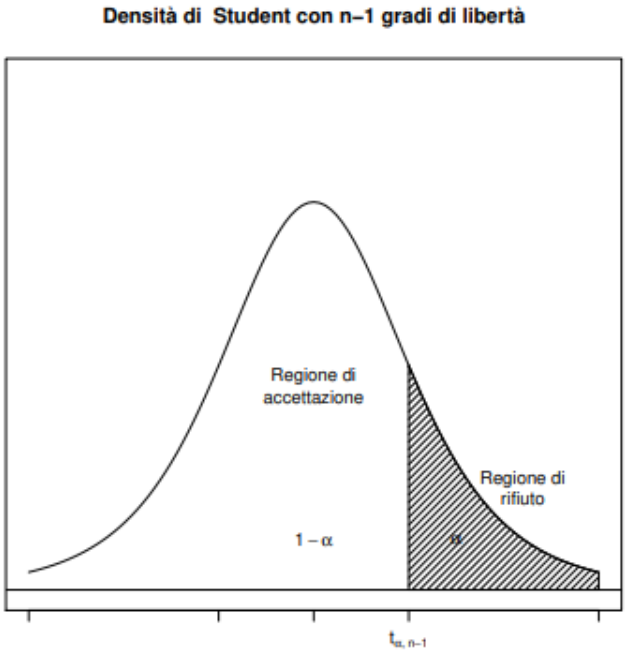
Consideriamo le ipotesi:

$H_0 : \mu <= \mu 0$
$H_1 : \mu > \mu 0$

Le ipotesi sono entrambe composite.
Il test unilaterale sinistro ψ di misura α per le ipotesi considerate è:

- si accetti H_0 se
  $\frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}} < t_{\alpha,n-1}$
- si rifiuti H_0 se
  $\frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}} > t_{\alpha,n-1}$

Graficamente:



p-value:$1 - P(T_n <= \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}})$

TEST UNILATERALE DESTRO

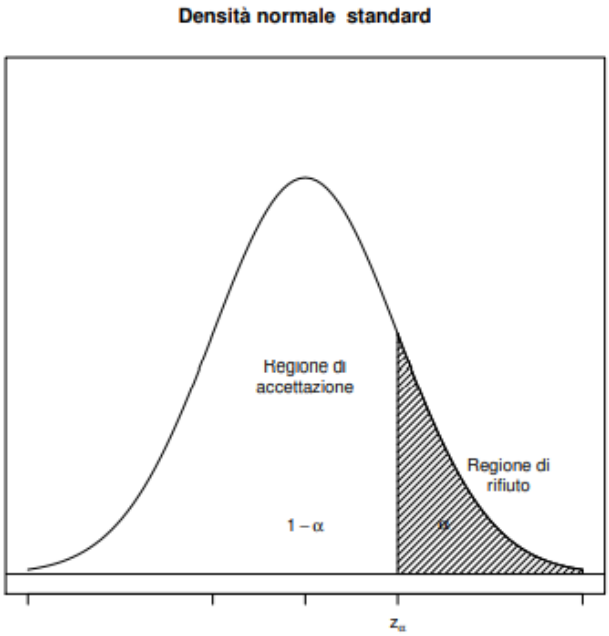
Consideriamo le ipotesi:

$H_0 : \mu >= \mu 0$
$H_1 : \mu < \mu 0$

Le ipotesi sono entrambe composite.
Il test unilaterale destro ψ di misura α per le ipotesi considerate è:

- si accetti H_0 se
  $\frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}} > -t_{\alpha,n-1}$
- si rifiuti H_0 se
  $\frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}} < -t_{\alpha,n-1}$

Graficamente:



p-value: $P(T_n <= \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}})$

Vediamo in R i vari test:

```
#IPOTESI BILATERALE

alpha<-0.01
mu0<-999.98
n<-length(campione_normale)
qt(1-alpha/2, df=n-1)
[1] 2.576321

meancamp<-mean(campione_normale)
devcamp<-sd(campione_normale)
t_os<-(meancamp-mu0)/(devcamp/sqrt(n))
t_os
[1] 0.6446597

pvalue<-2*(1-pt(t_os, df=n-1))
pvalue
[1] 0.5191625
```

Si nota che $t_{\alpha/2,n-1} = 2.57$ e $t_{os} = 0.64$ cade dentro la regione di accettazione, con-

fermato anche da un p-value maggiore di α.

```
#IPOTESI UNILATERALE SX

alpha<-0.01
mu0<-999.98
n<-length(campione_normale)
qt(1-alpha, df=n-1)
[1] 2.326721

meancamp<-mean(campione_normale)
devcamp<-sd(campione_normale)
t_os<-(meancamp-mu0)/(devcamp/sqrt(n))
t_os
[1] 0.6446597

pvalue<-1-pt(t_os,df=n-1)
pvalue
[1] 0.2595813
```

Si nota che $t_{\alpha/2,n-1} = 2.57$ e $t_{os} = 0.64$ cade dentro la regione di accettazione, confermato anche da un p-value maggiore di α.

```
#IPOTESI UNILATERALE DX

alpha<-0.05
mu0<-999.98
n<-length(campione_normale)
qt(alpha, df=n-1)
[1] -1.645006

meancamp<-mean(campione_normale)
devcamp<-sd(campione_normale)
t_os<-(meancamp-mu0)/(devcamp/sqrt(n))
t_os
[1] 0.6446597

pvalue<-pt(t_os, df=n-1)
pvalue
[1] 0.7404187
```

Si nota che $t_{\alpha/2,n-1} = 2.57$ e $t_{os} = 0.64$ cade dentro la regione di accettazione, confermato anche da un p-value maggiore di α.

4.2.3 Test su σ^2 con valore medio μ noto

TEST BILATERALE

Consideriamo le ipotesi:

$H_0 : \sigma^2 = \sigma_0^2$
$H_1 : \sigma^2 \neq \sigma_0^2$

Essendo la varianza nota, H_0 è semplice, mentre l'ipotesi H_1 è composita.
Quando H_0 è vera, gioca un ruolo fondamentale la variabile aleatoria:

$V_n = \sum_{i=1}^{n} (\frac{X_i - \mu}{\sigma_0})^2 = \frac{(n-1)S_n^2}{\sigma_0^2} + (\frac{\bar{X}_n - \mu}{\sigma_0 \sqrt{n}})^2$

distribuita con legge del chi-quadrato con n-1 gradi di libertà.

Il test bilaterale ψ di misura α per le ipotesi considerate è:

- si accetti H_0 se
  $\chi^2_{1-\alpha/2,n} < \sum_{i=1}^{n} (\frac{X_i - \mu}{\sigma_0})^2 < \chi^2_{\alpha/2,n}$
- si rifiuti H_0 se
  $\sum_{i=1}^{n} (\frac{X_i - \mu}{\sigma_0})^2 < \chi^2_{1-\alpha/2,n}$ oppure $\sum_{i=1}^{n} (\frac{X_i - \mu}{\sigma_0})^2 > \chi^2_{\alpha/2,n}$

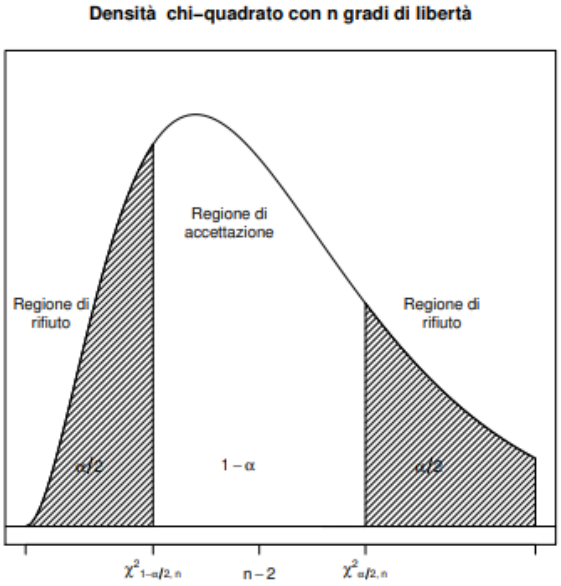
Graficamente:

TEST UNILATERALE SINISTRO

Consideriamo le ipotesi:

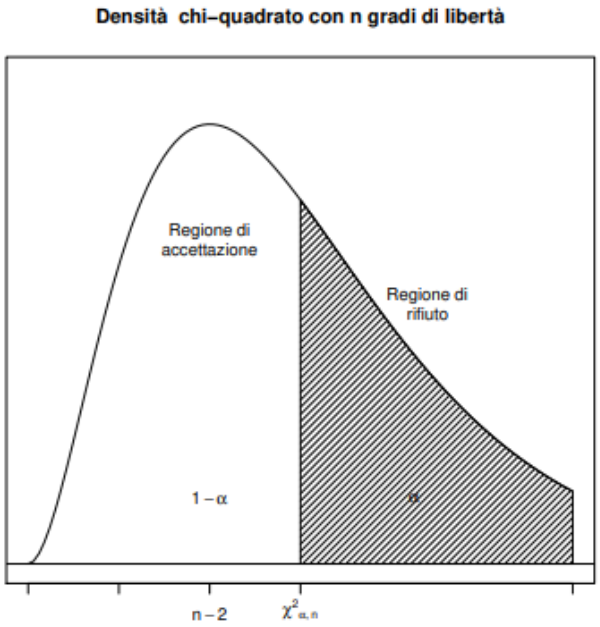
$H_0 : \sigma^2 <= \sigma_0^2$
$H_1 : \sigma^2 > \sigma_0^2$

Le ipotesi sono entrambe composite.
Il test unilaterale sinistro ψ di misura α per le ipotesi considerate è:



- si accetti H_0 se
  $\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma_0})^2 < \chi^2_{\alpha,n}$
- si rifiuti H_0 se
  $\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma_0})^2 > \chi^2_{\alpha,n}$

Graficamente:



TEST UNILATERALE DESTRO

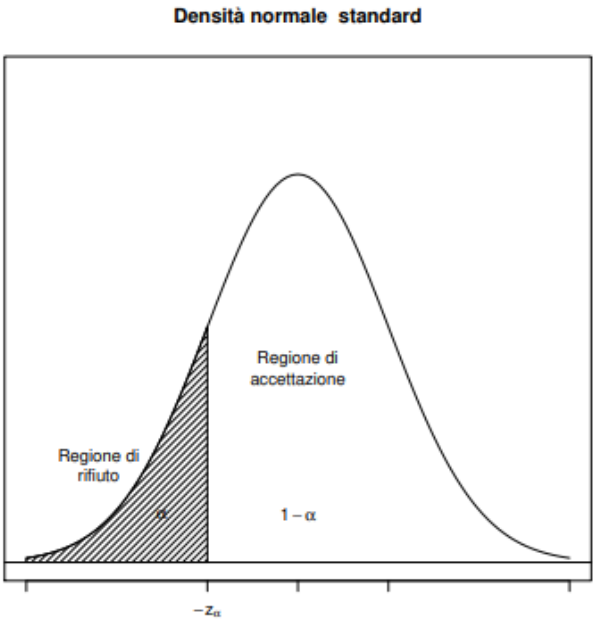
Consideriamo le ipotesi:

$H_0 : \sigma^2 >= \sigma_0^2$
$H_1 : \sigma^2 < \sigma_0^2$

Le ipotesi sono entrambe composite.
Il test unilaterale destro ψ di misura α per le ipotesi considerate è:

- si accetti H_0 se
  $\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma_0})^2 > \chi^2_{1-\alpha,n}$
- si rifiuti H_0 se
  $\sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma_0})^2 < \chi^2_{1-\alpha,n}$

Graficamente:



Vediamo in R:

```
#IPOTESI BILATERALE

alpha<-0.05
mu<-999.98
sigma02<-var(campione_normale)
n<-length(campione_normale)
medcamp<-mean(campione_normale)
varcamp<-var(campione_normale)

qchisq(alpha/2, df=n)
[1] 9724.718

qchisq(1-alpha/2, df= n)
[1] 10279.07

(n-1)*varcamp/sigma02+n*(medcamp-mu)**2/sigma02
[1] 9999.416
```

Notiamo che χ^2 è compreso nella regione d'accettazione, quindi si accetta l'ipotesi.

4.2.4 Test su σ^2 con valore medio μ non noto

TEST BILATERALE

Consideriamo le ipotesi:

$H_0 : \sigma^2 = \sigma_0^2$
$H_1 : \sigma^2 \neq \sigma_0^2$

Essendo la varianza non nota entrambe sono composite.
Quando H_0 è vera, gioca un ruolo fondamentale la variabile aleatoria:

$Q_n = \frac{(n-1)S_n^2}{\sigma_0^2}$

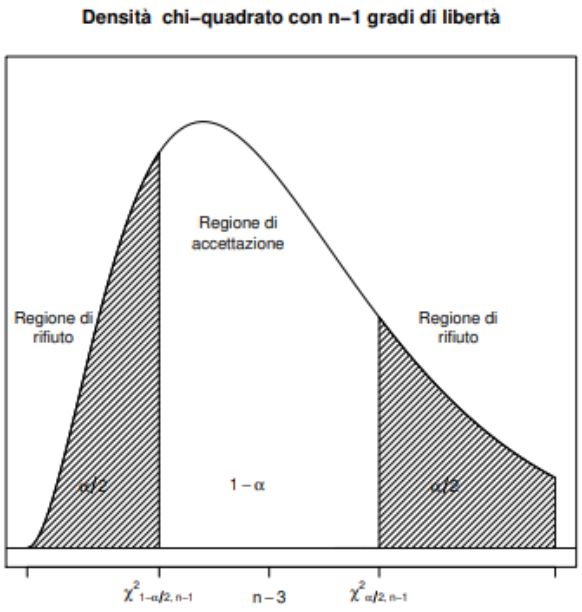
distribuita con legge del chi-quadrato con n-1 gradi di libertà.

Il test bilaterale ψ di misura α per le ipotesi considerate è:

- si accetti H_0 se
  $\chi_{1-\alpha/2,n-1} < \frac{(n-1)S_n^2}{\sigma_0^2} < \chi^2_{\alpha/2,n-1}$

- si rifiuti H_0 se
  $\frac{(n-1)S_n^2}{\sigma_0^2} < \chi_{1-\alpha/2,n-1}$ oppure $\frac{(n-1)S_n^2}{\sigma_0^2} > \chi^2_{\alpha/2,n-1}$

Graficamente:



TEST UNILATERALE SINISTRO

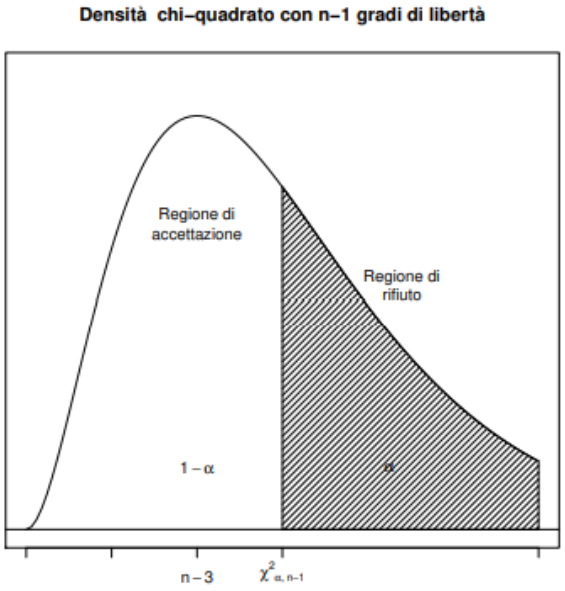
Consideriamo le ipotesi:

$H_0 : \sigma^2 <= \sigma_0^2$
$H_1 : \sigma^2 > \sigma_0^2$

Le ipotesi sono entrambe composite.
Il test unilaterale sinistro ψ di misura α per le ipotesi considerate è:

- si accetti H_0 se
  $\frac{(n-1)S_n^2}{\sigma_0^2} < \chi^2_{\alpha,n-1}$
- si rifiuti H_0 se
  $\frac{(n-1)S_n^2}{\sigma_0^2} > \chi^2_{\alpha,n-1}$

Graficamente:



TEST UNILATERALE DESTRO

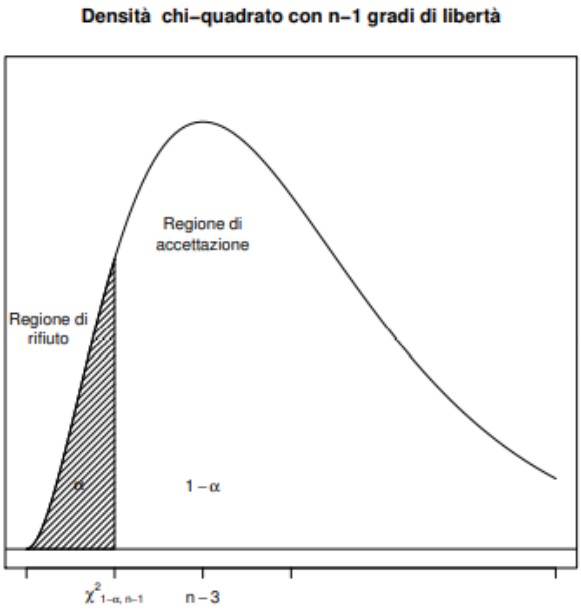
Consideriamo le ipotesi:

$H_0 : \sigma^2 >= \sigma_0^2$
$H_1 : \sigma^2 < \sigma_0^2$

Le ipotesi sono entrambe composite.
Il test unilaterale destro ψ di misura α per le ipotesi considerate è:

- si accetti H_0 se
  $\frac{(n-1)S_n^2}{\sigma_0^2} > \chi^2_{1-\alpha,n-1}$
- si rifiuti H_0 se
  $\frac{(n-1)S_n^2}{\sigma_0^2} < \chi^2_{1-\alpha,n-1}$

Graficamente:



Vediamo in R:

```
#IPOTESI BILATERALE

alpha<-0.05
sigma02<-2.5
n<-length(campione_normale)
varcamp<-var(campione_normale)

qchisq(alpha/2,df=n-1)
[1] 9723.732

qchisq(1-alpha/2,df=n-1)
[1] 10278.06

(n-1)*varcamp/sigma02
[1] 9081.925
```

Notiamo che χ^2 è compreso nell'area di accettazione quindi si accetta l'ipotesi nulla di un $\sigma^2 = 2.5$

```
#IPOTESI UNILATERALE SX

alpha<-0.05
sigma02<-2.5
n<-length(campione_normale)
varcamp<-var(campione_normale)

qchisq(1-alpha,df=n-1)
[1] 10279.07

(n-1)*varcamp/sigma02
[1] 9081.925
```

Notiamo che 9081.92 è nell'area di accettazione (a sinistra di 10279.07).

```
#IPOTESI UNILATERALE DX

alpha<-0.05
sigma02<-2.5
n<-length(campione_normale)
varcamp<-var(campione_normale)
```

```
7
8   qchisq(alpha, df=n-1)
9   [1] 9767.537
10
11  (n-1)*varcamp/sigma02
12  [1] 9081.925
```

Che non è nell'area di accettazione.

Capitolo 5

Criterio del chi-quadrato

Con il criterio del chi–quadrato si desidera verificare l'ipotesi che un certa popolazione, descritta da una variabile aleatoria X, sia caratterizzata da una funzione di distribuzione $F_X(x)$, con k parametri non noti da stimare.

Denotando con H_0 l'ipotesi soggetta a verifica (ipotesi nulla) e con H_1 l'ipotesi alternativa, il test chi–quadrato con livello di significatività α mira a verificare l'ipotesi nulla:

- 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 $F_X(x)$;

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 X 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.($p_i = P(X \in I_i)$)
Dal campione osserviamo le frequenze assolute $n_1, n_2, ..., n_r$ in cui gli elementi del campione si distribuiscono rispettivamente in $I_1, I_2, ..., I_r$.
Quindi n_i rappresenta il numero degli elementi del campione che cadono nell'intervallo I_i $(i = 1, 2, ..., r)$.
Vale chiaramente che la somma dei p_i è unitaria, mentre la somma degli $n_i = n$.

Il criterio del chi-quadrato si basa sulla seguente statistica:

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

dove N_i è la variabile aleatoria che descrive il numero degli elementi del campione casuale $X_1, ..., X_n$ 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.

Per un campione sufficientemente numeroso di ampiezza n, il **test chi–quadrato bilaterale di misura** α è il seguente:

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

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

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

Vediamo in R per la distribuzione normale.

Dividiamo l'insieme in 5 in sottoinsiemi utilizzando i quantili della distribuzione normale:

```
#divido in 5 sottoinsimi il campione

m<-mean(campione_normale)
d<-sd(campione_normale)
a<-numeric(4)

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

>a
[1]  998.7215  999.6079 1000.3715 1001.2579
```

Gli intervalli sono $I_1 = (-\infty, 998.7215)$, $I_2 = [998.7215, 999.6079)$,
$I_3 = [999.6079, 1000.3715)$, $I_4 = [1000.3715, 1001.2579)$, $I_5 = [1001.2579, +\infty)$.

Occorre ora determinare il numero di elementi del campione che cadono negli intervalli:

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

nint
[1] 1997 2030 1989 1941 2043

sum(nint)
[1] 10000
```

Allora $n_1 = 1997$, $n_2 = 2030$, $n_3 = 1989$, $n_4 = 1941$, $n_5 = 2043$.

Calcoliamo ora χ^2:

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

chi2
[1] 3.18
```

ossia $\chi^2 = 3.18$.

La distribuzione normale ha due parametri non noti (μ, σ^2) e quindi $k = 2$.
Pertanto, la funzione di distribuzione della statistica Q è approssimabbile con la funzione di distribuzione chi–quadrato con$r - k - 1 = 2$ gradi di libertà.
Ora occorre calcolare $\chi^2_{1-\alpha/2,r-k-1}$ e $\chi^2_{\alpha/2,r-k-1}$ 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
```

Essendo $0.0506 < \chi^2 = 3.18 < 7.3777$ l'ipotesi H_0 della popolazione normale può essere accettata.