**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**

LABANCA PAOLO 0522501161

*Anno accademico 2022-2023*

**INDICE**

[CAPITOLO 1 3](#_Toc122010337)

[Inferenza statistica: distribuzione normale 3](#_Toc122010338)

[1.1 Distribuzione normale 4](#_Toc122010339)

[1.1.1 FUNZIONE DI DISTRIBUZIONE 7](#_Toc122010340)

[1.1.2 REGOLA DEL 3*σ* 8](#_Toc122010341)

[1.1.3 SIMULAZIONE DELLA DISTRIBUZIONE NORMALE IN R 9](#_Toc122010342)

[CAPITOLO 2 11](#_Toc122010343)

[STIMA PUNTUALE 11](#_Toc122010344)

[2.1 STIMATORI 11](#_Toc122010345)

[2.2 METODI PER LA RICERCA DI STIMATORI 12](#_Toc122010346)

[2.2.1 METODO DEI MOMENTI 13](#_Toc122010347)

[2.2.2 METODO DELLA MASSIMA VEROSIMIGLIANZA 14](#_Toc122010348)

[2.3 PROPRIETÀ DEGLI STIMATORI E STIMATORE ASINCRONO CORRETTO 16](#_Toc122010349)

[CAPITOLO 3 17](#_Toc122010350)

[Intervalli di confidenza 17](#_Toc122010351)

[3.1 METODO PIVOTALE 18](#_Toc122010352)

[3.2 POPOLAZONE NORMALE 18](#_Toc122010353)

[3.2.1 INTERVALLO DI CONFIDENZA PER µ CON σ2 NOTA 18](#_Toc122010354)

[3.2.2 INTERVALLO DI CONFIDENZA PER µ CON σ2 NON NOTA 20](#_Toc122010355)

[3.2.3 INTERVALLO DI CONFIDENZA PER σ2 CON µ NOTA 22](#_Toc122010356)

[3.2.4 INTERVALLO DI CONFIDENZA PER σ2 CON µ NON NOTA 24](#_Toc122010357)

[3.3 CONFRONTO TRA POPOLAZIONI NORMALI 25](#_Toc122010358)

[3.3.1 INTERVALLO DI CONFIDENZA PER *µ1 - µ2* CON E NOTE 25](#_Toc122010359)

[3.3.2 INTERVALLO DI CONFIDENZA PER µ1 - µ2 CON E NON NOTE 27](#_Toc122010360)

[4 CAPITOLO 29](#_Toc122010361)

[Verifica delle ipotesi 29](#_Toc122010362)

[4.1 P-VALUE 30](#_Toc122010363)

[4.2 POPOLAZIONE NORMALE 30](#_Toc122010364)

[4.2.1 TEST SU *µ* CON VARIANZA σ2 NOTA 31](#_Toc122010365)

[4.2.2 TEST SU *µ* CON VARIANZA *σ2* NON NOTA 36](#_Toc122010366)

[4.2.3 TEST SU *σ*2 CON VALORE MEDIO *µ* NOTO 41](#_Toc122010367)

[4.2.4 TEST SU *σ*2 CON VALORE MEDIO *µ* NON NOTO 44](#_Toc122010368)

[CAPITOLO 5 48](#_Toc122010369)

[CRITERIO DEL CHI-QUADRATO 48](#_Toc122010370)

# CAPITOLO 1

# Inferenza statistica: distribuzione normale

L’indagine statistica viene svolta su un insieme, che viene definito come popolazione ed è compreso in un numero finito o infinito. 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.

Di particolare importanza in statistica è l’inferenza statistica. Essa 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 ϑ ∈ Θ non noto (o più parametri non noti). Per ottenere informazioni sul parametro non noto ϑ della popolazione, si può fare uso dell’inferenza statistica considerando un campione (x1, x2, ..., xn) estratto dalla popolazione e poi effettuando delle misure. Ovviamente il campione deve essere preso in modo tale da essere rappresentativo della popolazione. 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 ϑ 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 popolazione avente distribuzione normale. Tale distribuzione viene analizzata in R.

## Distribuzione normale

La distribuzione normale o detta anche gaussiana, ha una forte centralità nella statistica e nel calcolo della probabilità, in quanto costituisce una distribuzione limite alla quale tendono varie altre funzioni di distribuzioni sotto opportune ipotesi. La distribuzione normale è spesso considerata un buon modello per variabili fisiche come peso, altezza, livello di inquinamento, per analizzare gli errori di misurazione ed anche per descrivere il reddito familiare.

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

*Immagine che contiene testo, orologio, calibro

Descrizione generata automaticamente*

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

La densità è simmetricarispetto all’asse *x* = *µ*, risulta infatti *fX(µ - x) = fX(µ + x).*

Il grafico di *fX(x)* esibisce una caratteristica **forma a campana**, simmetrica rispetto a x = µ.

Il **massimo** è in corrispondenza del punto *x* =*µ* ed è pari a

Ha due **flessi** in corrispondenza di *µ - σ* e *µ + σ.*

Per una variabile aleatoria normale il valore medio e la varianza sono:

*E(X) = µ, VAR(X) = σ2*

da cui segue che *σ* rappresenta la deviazione standard.

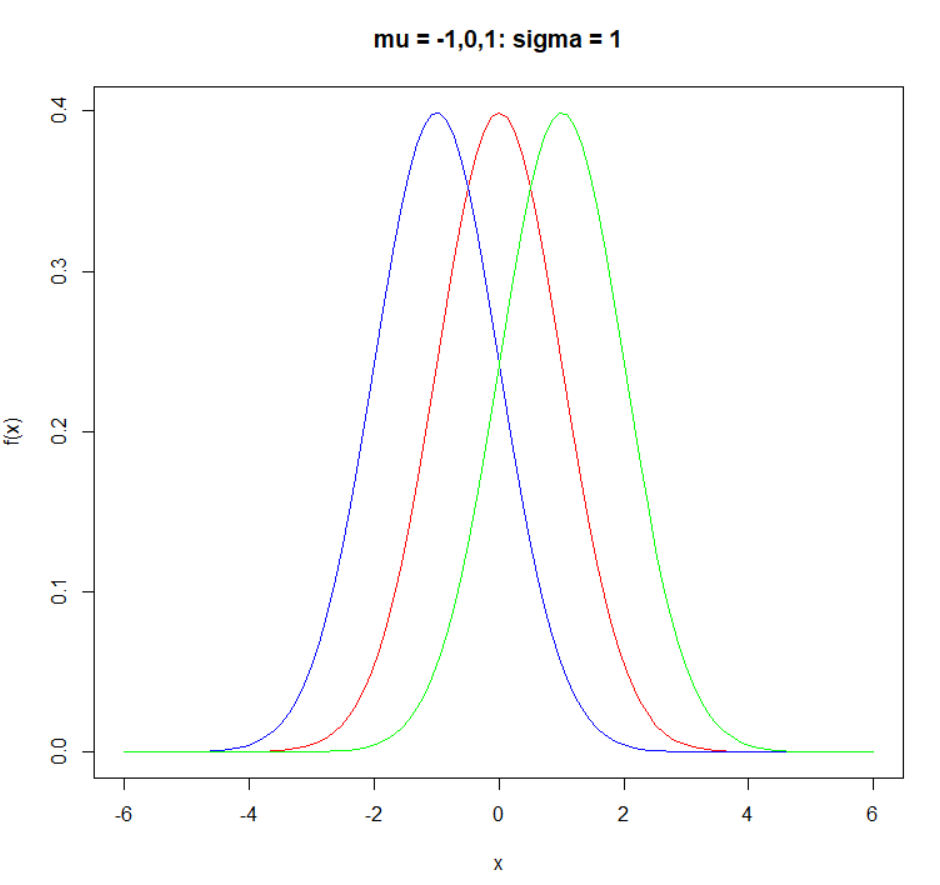
Per calcolare in R la densità normale si usa la funzione **dnorm(x, mean=mu, sd=sigma)**.

Dove:

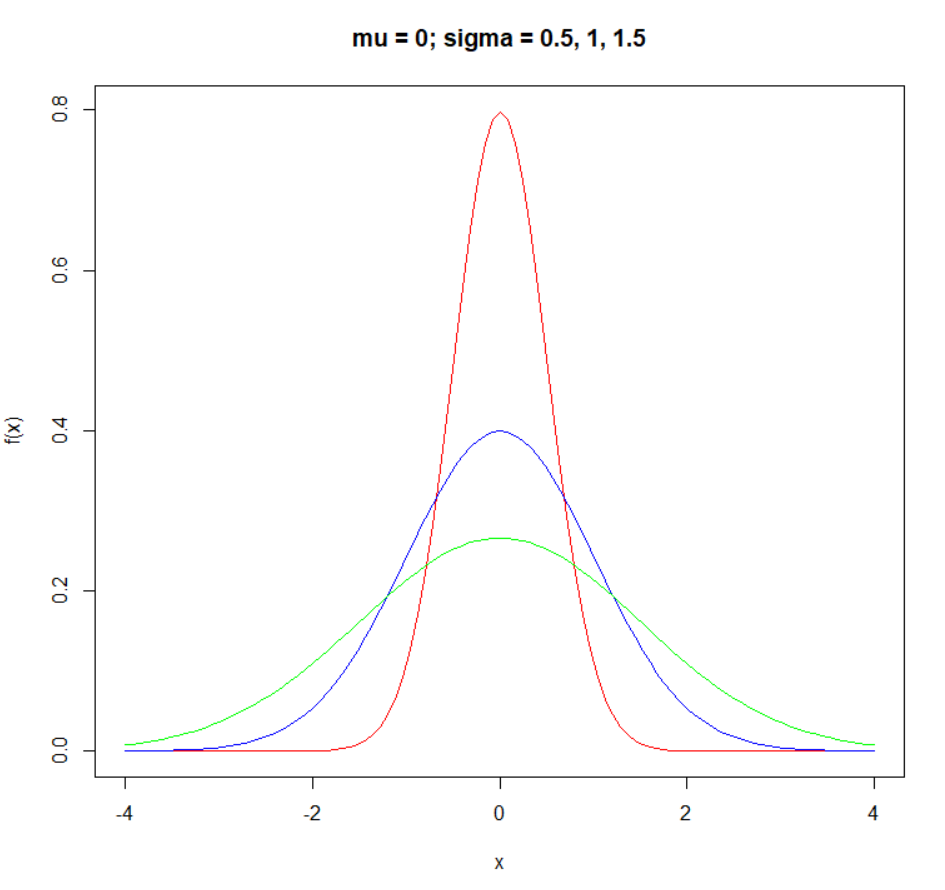
* x è il valore assunto (o i valori assunti) dalla variabile aleatoria normale
* **mean** e **sd** sono il valore medio e la deviazione standard della densità normale

Immagine che contiene testo

Descrizione generata automaticamenteNotiamo 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’area sottesa rimane sempre unitaria.



*Figura 1.1: Densità normale variando mu.*

Immagine che contiene testo

Descrizione generata automaticamente

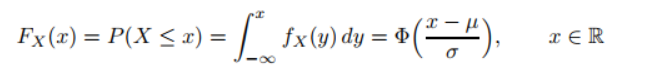
Figura 1.2: Densità normale variando sigma.

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* = *µ* + *σZ*.

### 1.1.1 FUNZIONE DI DISTRIBUZIONE

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

**

Dove

Immagine che contiene testo, orologio

Descrizione generata automaticamente

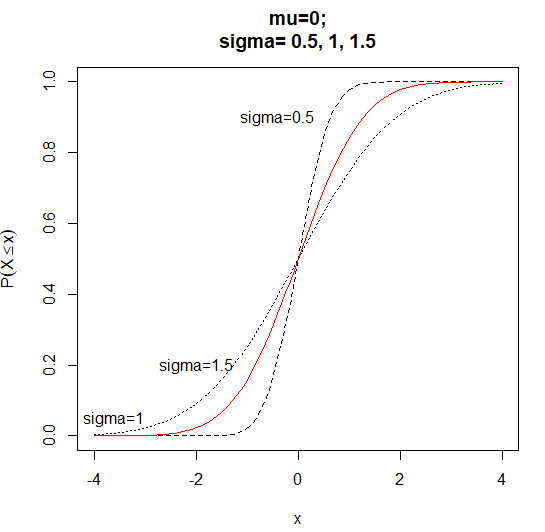
È la funzione di distribuzione di una variabile aleatoria *Z ∼ N(0, 1)*, detta **normale standard**, ossia normale con valore medio nullo e varianza unitaria. Pertanto, se *X ∼ N(µ, σ)* si ha:



In R la funzione di distribuzione di una variabile *X ∼* *N*(*µ, σ*) si calcola tramite la funzione **pnorm(x, mean, sd, lower.tail=TRUE)**

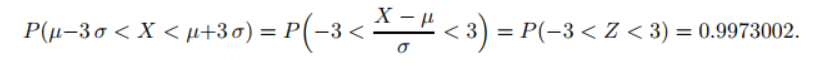
Immagine che contiene testo

Descrizione generata automaticamente

Figura 1.3: Funzione di distribuzione normale.

### 1.1.2 REGOLA DEL 3*σ*

Per una qualsiasi variabile aleatoria normale X ∼ N(µ, σ) risulta



Quindi la probabilità che una variabile aleatoria X ∼ N(µ, σ) assuma valori in un intervallo avente come centro *µ* e semiampiezza *3σ* è prossima all’unità. Questa proprietà delle variabili aleatorie normali è nota come regola del *3σ*. Tale regola permette di individuare l’intervallo (*µ - 3σ, µ + 3σ*) in cui rappresentare la funzione densità di una variabile normale di valore medio *µ* e varianza *σ2* 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

Si và a confrontare la normale teorica con la densità simulata e si determina il cambiamento dell’istogramma del campione, in quanto, aumentiamo la sequenza in output, l’istogramma delle frequenze relative si avvicina 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:

Immagine che contiene testo

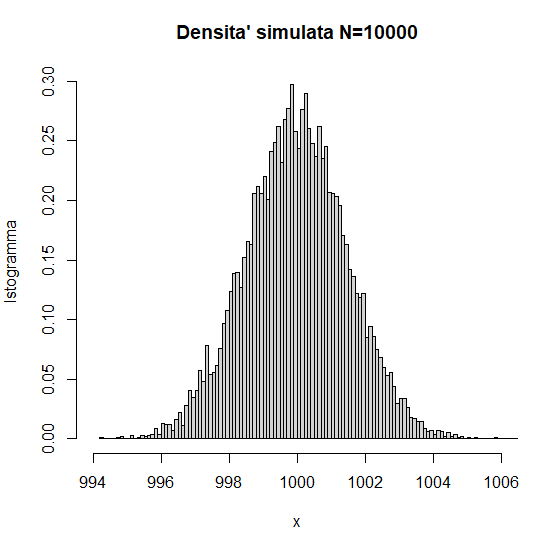
Descrizione generata automaticamente

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:



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 1.3: 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 ϑ ∈ Θ non noto (o più parametri non noti). Il termine osservabile significa che si possono osservare i valori assunti dalla variabile aleatoria X, quindi il parametro non noto è presente soltanto nella legge di probabilità. Ovviamente se i parametri sono noti la legge di probabilità è completamente specificata. Affinché le conclusioni dell’inferenza statistica siano valide il campione deve essere scelto in modo da essere rappresentativo della popolazione. Molti metodi dell’inferenza statistica sono basati sull’ipotesi di campioni casuali.

## 2.1 STIMATORI

Nei metodi di indagine dell’inferenza statistica si considera un campione *casuale X1, X2, . . . , Xn* 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 (X1, X2, . . . , Xn)* è una funzione misurabile e osservabile del campione casuale *X1, X2, . . . , Xn*. Essendo la statistica osservabile, i valori da essa assunti dipendono soltanto dal campione osservato (x1, x2, . . . , xn) estratto dalla popolazione e i parametri non noti sono presenti soltanto nella funzione di distribuzione della statistica.

Uno **stimatore**  *= t (X1, X2, . . . , Xn)* è una funzione misurabile e osservabile del campione casuale *X1, X2, . . . , Xn* i cui valori possono essere usati per stimare un parametro non noto ϑ della popolazione. I valori assunti da tale stimatore sono detti stime del parametro non noto ϑ.

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

Sia X1, X2, . . . , Xn un campione casuale. La statistica

Immagine che contiene testo, orologio

Descrizione generata automaticamente

e detta media campionaria, mentre la statistica

Immagine che contiene orologio

Descrizione generata automaticamente

e detta varianza campionaria.

Sia X1, X2, . . . , Xn un campione casuale estratto da una popolazione descritta da una variabile aleatoria osservabile X caratterizzata da valore medio E(X) = µ finito e varianza Var (X) = σ2 finita. Risulta:

*E() = µ, Var (X) =*

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

## 2.2 METODI PER LA RICERCA DI STIMATORI

Sia X1, X2, . . . , Xn un campione casuale estratto da una popolazione con funzione di probabilità (nel caso discreto) oppure densità di probabilità (nel caso assolutamente continuo) f(x; ϑ1, ϑ2, . . . , ϑk) dove ϑ1, ϑ2, . . . , ϑk denotano i parametri non noti della popolazione. Lo scopo del 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

Per illustrarlo occorre in primo luogo definire i momenti campionari

Si definisce momento campionario r–esimo relativo ai valori osservati (x1, x2, . . . , xn) del campione casuale il valore

Immagine che contiene orologio

Descrizione generata automaticamente

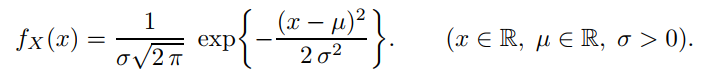
Si nota quindi che il momento campionario r–esimo è la media aritmetica delle potenze r–esime delle n osservazioni effettuate sulla popolazione. In particolare, se r = 1 il momento campionario M1(x1, x2, . . . , xn) coincide con il valore osservato della media campionaria , ossia M1 = (x1 + x2 + . . . + xn)/n. Se esistono k parametri da stimare, il metodo dei momenti consiste nell’uguagliare i primi k momenti della popolazione in esame con i corrispondenti momenti del campione casuale. Quindi, se i primi k momenti esistono e sono finiti, tale metodo consiste nel risolvere il sistema di k equazioni



Le incognite del sistema sono i parametri *ϑ*1*, ..., ϑk* e sono presenti a sinistra del sistema.

**POPOLAZIONE NORMALE**

Si è interessati a determinare con il metodo dei momenti gli stimatori dei parametri µ e σ2 di una popolazione normale descritta da una variabile aleatoria X ∼ N(*µ, σ*) di densità di probabilità

****

Occorre quindi stimare due parametri µ e σ2. Poiché E(x) = µ e E(X2) = σ2 + µ2,

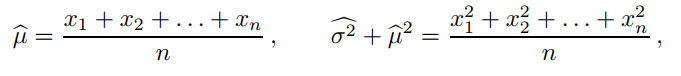
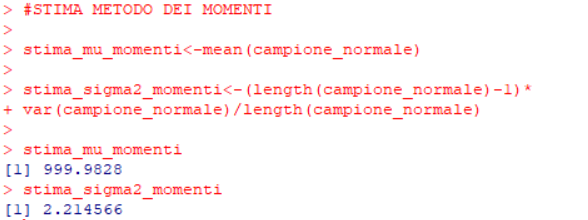
****

Immagine che contiene testo

Descrizione generata automaticamente

Utilizzando la prima nella seconda si ricava:

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.



### 2.2.2 METODO DELLA MASSIMA VEROSIMIGLIANZA

Il metodo della massima verosimiglianza è il più importante metodo per la stima dei parametri non noti di una popolazione e solitamente è preferito al metodo dei momenti. Per illustrare il metodo della massima verosimiglianza occorre introdurre in primo luogo la **funzione di verosimiglianza**.

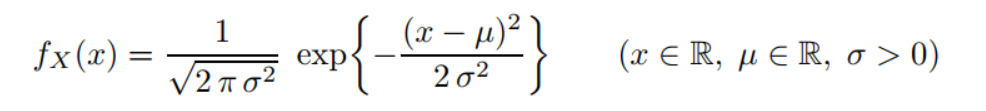
Sia X1, X2, . . . , Xn un campione casuale di ampiezza n estratto dalla popolazione. La funzione di verosimiglianza *L(ϑ1, ϑ2, . . . , ϑk) = L(ϑ1, ϑ2, . . . , ϑk; x1, x2, . . . , xn)* del campione osservato (x1, x2, . . . , xn) è la funzione di probabilità congiunta (nel caso di popolazione discreta) oppure la funzione densità di probabilità congiunta (nel caso di popolazione assolutamente continua) del campione casuale X1, X2, . . . , Xn, ossia

*L(ϑ1, ϑ2, . . . , ϑk) = L(ϑ1, ϑ2, . . . , ϑk; x1, x2, . . . , xn)= f(x1; ϑ1, ϑ2, . . . , ϑk) f(x2; ϑ1, ϑ2, . . . , ϑk) · · · f(xn; ϑ1, ϑ2, . . . , ϑk).*

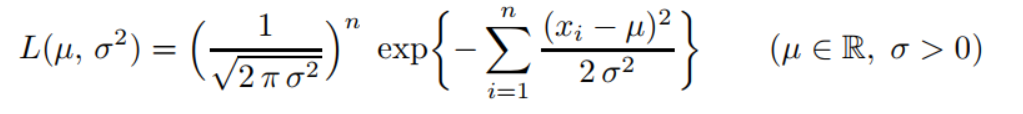
Il metodo della massima verosimiglianza consiste nel massimizzare la funzione di verosimiglianza rispetto ai parametri ϑ1, ϑ2, . . . , ϑk. Tale metodo cerca quindi di determinare da quale funzione di probabilità congiunta (nel caso di popolazione discreta) oppure di densità di probabilità congiunta (nel caso di popolazione assolutamente continua) è più verosimile, che provenga il campione osservato. Pertanto si cercano di determinare i valori ϑ1, ϑ2, . . . , ϑk che rendono massima la funzione di verosimiglianza. I valori di ϑ1, ϑ2, . . . , ϑk che massimizzano la funzione di verosimiglianza sono indicati con 1, 2, . . . , k; essi costituiscono le stime di massima verosimiglianza dei parametri non noti ϑ1, ϑ2, . . . , ϑk della popolazione. Ovviamente i valori stimati, indicati con *k* **dipendono** dal campione osservato.

**POPOLAZIONE NORMALE**

Si desidera determinare lo stimatore di massima verosimiglianza dei parametri *µ* e *σ2*di una popolazione normale caratterizzata da funzione densità di probabilità:

****

Si ha:



Le stime di massima verosimiglianza dei parametri µ e σ2 sono rispettivamente

Immagine che contiene orologio, calibro

Descrizione generata automaticamente

Lo stimatore di massima verosimiglianza e dei momenti del valore medio µ è la media campionaria X. Invece lo stimatore di massima verosimiglianza e dei momenti della varianza σ2 è (n - 1) S2/n.

## 2.3 PROPRIETÀ DEGLI STIMATORI E STIMATORE ASINTOTICAMENTE CORRETTO

In generale esistono molti stimatori che possono essere utilizzati per stimare il parametro non noto di una popolazione. Occorre quindi definire delle proprietà di cui può o meno godere uno stimatore. Uno stimatore si dice **corretto** se il valore medio dello stimatore è uguale al corrispondente parametro non noto della popolazione. Occorre sottolineare che possono esistere differenti stimatori corretti di un parametro non noto di una popolazione. 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 (costituita da un singolo valore reale) spesso si preferisce sostituire un intervallo di valori, detto **intervallo di confidenza** (o intervallo di fiducia), ossia si cerca di determinare in base ai dati del campione, due limiti (uno inferiore ed uno superiore) entro i quali sia compreso il parametro non noto con un certo coefficiente di confidenza (detto anche grado di fiducia).

Fissato un coefficiente di confidenza 1 - α (0 < α < 1), se è possibile scegliere le statistiche n e n in modo tale che

*P( n < ϑ < n) = 1 – α*

allora si dice che (n, n) è un intervallo di confidenza (intervallo di fiducia) di grado 1 - α per ϑ. Inoltre, le statistiche n e n sono dette, limite inferiore e superiore dell’intervallo di confidenza. Se g1(x) e g2(x) sono i valori assunti dalle statistiche n e n per il campione osservato x = (x1, x2, . . . , xn), allora l’intervallo g1(x), g2(x) è detto stima dell’intervallo di confidenza di grado 1 - α per ϑ ed i punti finali g1(x) e g2(x) di tale intervallo sono detti rispettivamente **stima del limite inferiore** e **stima del limite superiore dell’intervallo di confidenza.** In generale, esistono numerosi intervalli di confidenza dello stesso grado 1-α per un parametro non noto ϑ della popolazione. La scelta dell’intervallo di confidenza deve essere effettuata dal decisore in base ad alcune proprietà statistiche. Ad esempio, fissato un coefficiente di confidenza 1 - α, 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.**

*L(X1, X2, . . . , Xn; 1 - α) = n – n*

## 3.1 METODO PIVOTALE

Questo metodo consiste nel determinare una variabile aleatoria di pivot *δ (X1, ..., Xn; ϑ)* che:

* dipende dal campione casuale *X*1*, ..., Xn*;
* 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 POPOLAZONE NORMALE

Sia X1, X2, . . . , Xn un campione casuale di ampiezza n estratto da una popolazione normale descritta da una variabile aleatoria *X = N(µ, σ)* con valore medio *µ* e varianza *σ2*. Si possono analizzare i seguenti problemi:

* determinare un intervallo di confidenza di grado *1 - α* per il valore medio *µ* nel caso in cui la varianza *σ2* della popolazione normale è nota;
* determinare un intervallo di confidenza di grado *1 - α* per il valore medio *µ* nel caso in cui la varianza della popolazione normale è non nota;
* determinare un intervallo di confidenza di grado *1 - α* per la varianza *σ2* nel caso in cui il valore medio *µ* della popolazione normale è noto;
* determinare un intervallo di confidenza di grado *1 - α* per la varianza *σ2* nel caso in cui il valore medio della popolazione normale è non noto.

### 3.2.1 INTERVALLO DI CONFIDENZA PER µ CON σ2 NOTA

Per determinare un intervallo di confidenza di grado 1-α per il valore medio µ nel caso in cui la varianza σ2 della popolazione normale è nota, utilizziamo il metodo pivotale e consideriamo la variabile aleatoria

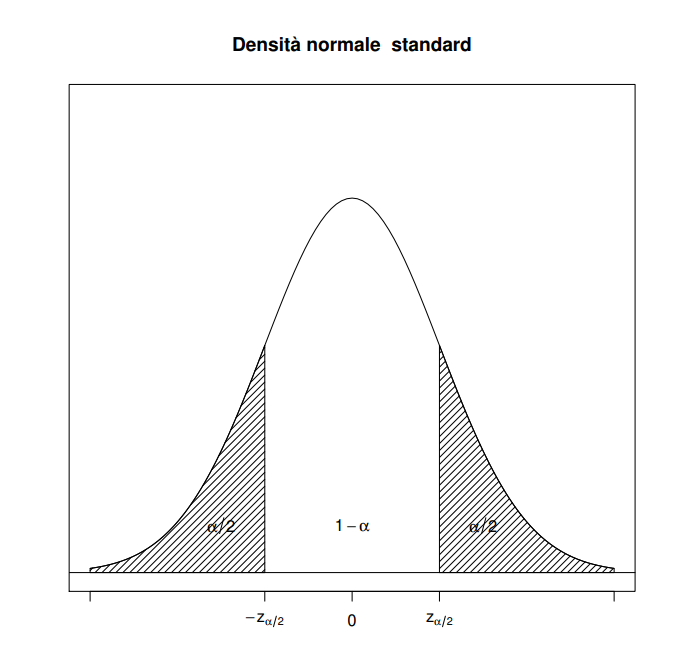
*Zn*=

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, *Zn* 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 *α*1 = *-α*2. Scegliamo quindi *α*1 = *-zα/*2 e *α*2 = *zα* tale che:

*P (Zn < -zα/2) = P(Zn > zα/2) =*

*P (-zα/2 < Zn < zα/2) = 1 – α*

Graficamente avremo che:



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 *- α* per il valore medio *µ* è:

Immagine che contiene antenna

Descrizione generata automaticamente

Dove

Immagine che contiene testo, orologio

Descrizione generata automaticamente

Denota la media campionaria delle n osservazioni.

Vediamo in R, poniamo *α* = 0*.*05 e supponiamo che la varianza nota sia *σ*2 = 2*.*25 e quindi *σ* = 1*.*5, stimiamo il parametro con l’intervallo di confidenza che abbiamo trovato:

**Immagine che contiene testo

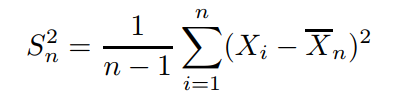
Descrizione generata automaticamente**

### 3.2.2 INTERVALLO DI CONFIDENZA PER µ CON σ2 NON NOTA

Per determinare un intervallo di confidenza di grado 1-α per il valore medio µ nel caso in cui la varianza σ2 della popolazione normale non è nota, utilizziamo il metodo pivotale e consideriamo la variabile aleatoria di pivot:

*Tn=*

Dove S2n rappresenta la variazione campionaria.

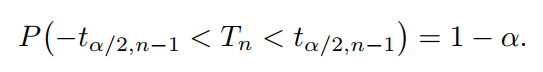


La variabile aleatoria *Tn* dipende dal campione casuale e dal parametro non noto *µ* e la sua legge di probabilità non dipende dal parametro non noto. Quindi, *Tn* può essere interpretata come una variabile aleatoria di pivot, inoltre è distribuita con legge di Student con *n -* 1 gradi di libertà. Scegliendo nel metodo pivotale α1 = - tα/2,n-1 e α2 = tα/2,n-1, dove tα/2,n-1 è tale che:

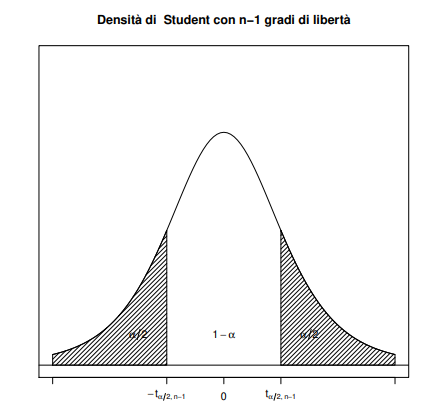
Immagine che contiene testo

Descrizione generata automaticamente

Si ha:



Graficamente:



Una stima dell’intervallo di confidenza 1 *- α* per il valore medio *µ* è:

Poniamo *α* = 0*.*05 e supponiamo che la media nota sia *µ* = 1000, stimiamo il parametro con l’intervallo di confidenza che abbiamo trovato:

Immagine che contiene antenna

Descrizione generata automaticamente

Dove n=denotato con la media campionaria delle *n* osservazioni e *sn* la deviazione standard. Vediamo in R, poniamo *α* = 0*.*05 e stimiamo il parametro con l’intervallo di confidenza che abbiamo trovato:

Immagine che contiene testo

Descrizione generata automaticamente

### 3.2.3 INTERVALLO DI CONFIDENZA PER σ2 CON µ NOTA

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

Immagine che contiene orologio

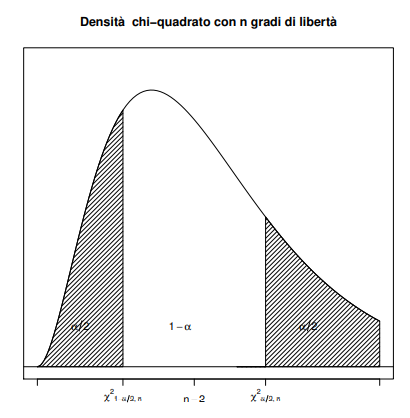
Descrizione generata automaticamente

Tale variabile aleatoria è distribuita con legge chi–quadrato con *n* gradi di libertà, essendo costituita dalla somma dei quadrati di *n* variabili aleatorie normali standard. Inoltre, Vn dipende dal campione casuale e dal parametro non noto σ2 (essendo il valore medio µ noto) e la sua legge di probabilità non contiene il parametro non noto. Quindi, Vn può essere interpretata come una variabile di pivot. Scegliendo nel metodo pivotale α1 = X21-α/2,n e α2 = X2α/2,n in maniera tale che

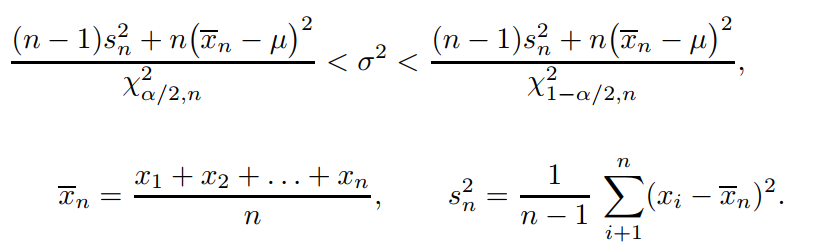
Immagine che contiene testo

Descrizione generata automaticamente

Graficamente



Una stima dell’intervallo di confidenza 1 *- α* per varianza *σ*2 è: dove xn e s2n rappresentano mediana e varianza campionaria



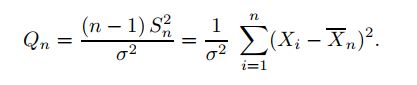
Poniamo *α* = 0*.*05 e supponiamo che la media nota sia *µ* = 1000, stimiamo il parametro con l’intervallo di confidenza che abbiamo trovato con R:

Immagine che contiene testo

Descrizione generata automaticamente

### 3.2.4 INTERVALLO DI CONFIDENZA PER σ2 CON µ NON NOTA

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

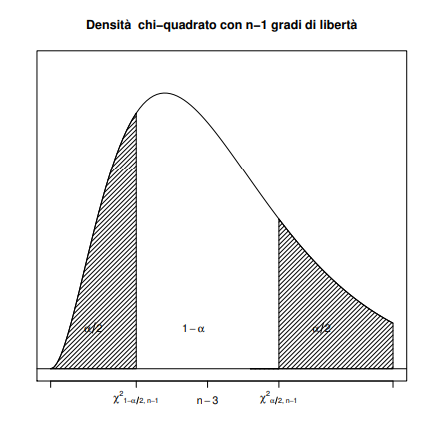


Tale variabile dipende dal campione casuale e dal parametro non noto *σ*2 ed è distribuita con legge chi–quadrato con *n -* 1 gradi di libertà. Scegliendo nel metodo pivotale *α*1 = *X*21*-α/*2*,n-*1 e *α*2 = *X*2*α/*2*,n-*1, tale che:

Immagine che contiene testo

Descrizione generata automaticamente

Graficamente



Una stima dell’intervallo di confidenza di grado 1 - α per la varianza σ2 è:

Immagine che contiene testo, orologio

Descrizione generata automaticamente

Dove si denota la varianza campionaria delle *n* osservazioni

Immagine che contiene orologio

Descrizione generata automaticamente

Poniamo *α* = 0*.*05 stimiamo il parametro con l’intervallo di confidenza che abbiamo trovato:

Immagine che contiene testo

Descrizione generata automaticamente

## 3.3 CONFRONTO TRA POPOLAZIONI NORMALI

Consideriamo due campioni, *X*1*, ..., Xn*1 e *Y*1*, ..., Yn*2, casuali e indipendenti, di ampiezza *n*1 e *n*2 estratti rispettivamente da due popolazioni normali. Vogliamo analizzare due problemi:

* determinare un intervallo di confidenza di grado 1 *- α* per *µ*1 *- µ*2 quando entrambe le varianze sono note;
* determinare un intervallo di confidenza di grado 1 *- α* per *µ*1 *- µ*2 quando entrambe le varianze non sono note;

### 3.3.1 INTERVALLO DI CONFIDENZA PER *µ1 - µ2* CON E NOTE

Denotiamo con

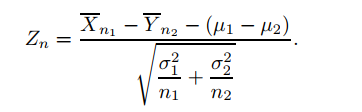
Immagine che contiene testo

Descrizione generata automaticamente

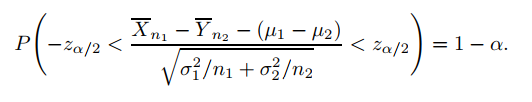
rispettivamente le medie campionarie delle due popolazioni normali. Poiché per ipotesi i campioni casuali X1, X2, . . . , Xn1 e Y1, Y2, . . . , Yn2 sono indipendenti, la statistica Xn1 - Y n2 è distribuita normalmente con valore medio e varianza



Ottenute ricordando 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 - α* per *µ1 - µ2* quando entrambe le varianze e delle due popolazioni normali sono note, consideriamo la variabile aleatoria di pivot



Tale variabile aleatoria dipende dal campione casuale e dal parametro non noto µ1 - µ2 (le varianze campionariee delle due popolazioni sono note) ed è caratterizzata da una **densità normale standard**. Pertanto, utilizzando il metodo pivotale si ha



Sussiste quindi la seguente proposizione. Una stima dell’intervallo di confidenza 1 *- α* per la differenza tra le medie è:

Immagine che contiene testo

Descrizione generata automaticamente

dove *n* e *n* sono le medie campionarie delle due osservazioni.

Poniamo *α* = 0*.*05 e stimiamo *µ*1 *- µ*2 per i due campioni che abbiamo a disposizione sapendo che le varianze note sono = 2*.*25 e = 4, mentre la numerosità del primo campione è pari a 10000, mentre quella del secondo 9000:

Immagine che contiene testo

Descrizione generata automaticamente

Essendo il limite superiore e inferiore entrambi positivi, possiamo dire che la media della prima popolazione è maggiore della media della seconda popolazione.

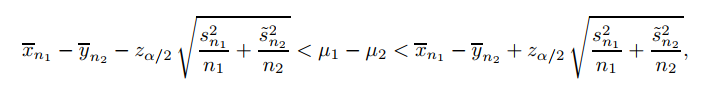
### 3.3.2 INTERVALLO DI CONFIDENZA PER µ1 - µ2 CON E NON NOTE

Siano X1, X2, . . . , Xn1 e Y1, Y2, . . . , Yn2 due campioni casuali indipendenti di ampiezza n1 e n2 estratti da due popolazioni normali X ∼ N(*µ1, σ1*) e Y ∼ N(*µ2, σ2*) con tutti i parametri non noti. Vogliamo determinare un intervallo di confidenza di grado 1 - α per la differenza µ1 - µ2 delle due popolazioni per grandi valori di n1 e n2. Denotiamo con e 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 si ha

Immagine che contiene testo

Descrizione generata automaticamente

Una stima dell’intervallo di confidenza di grado 1 - α per la differenza tra le due medie µ1 - µ2 è



dove xn1 e yn2 denotano rispettivamente le medie campionarie delle due osservazioni e dove e denotano rispettivamente le varianze campionarie delle due osservazioni. Poniamo *α* = 0*.*05 e stimiamo *µ*1 *- µ*2 per i due campioni che abbiamo a disposizione:

Immagine che contiene testo

Descrizione generata automaticamente

Essendo il limite superiore e inferiore entrambi positivi, possiamo dire che la media della prima popolazione è maggiore della media della seconda popolazione.

# 4 CAPITOLO

# Verifica delle ipotesi

Dopo la stima dei parametri il passo successivo è la **verifica delle ipotesi**. La verifica delle ipotesi è un procedimento che consiste nel fare una congettura o un’ipotesi sul parametro non noto ϑ o sulla distribuzione di probabilità e nel decidere, sulla base del campione estratto se essa è accettabile. 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*; *ϑ*), 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*; *ϑ*) è detta **ipotesi semplice**, altrimenti è chiamata **ipotesi composta**.

Spesso lo spazio Θ dei parametri, ossia l’insieme in cui può variare il parametro non noto della popolazione, si suddivide in due sottoinsiemi disgiunti Θ0 e Θ1 tali che Θ = Θ0 ∪ Θ1. L’ipotesi H0 soggetta a verifica su ϑ consiste nell’affermare che ϑ ∈ Θ0 ed è detta **ipotesi nulla**, mentre **nell’ipotesi alternativa** H1 si assume invece che ϑ ∈ Θ1. Il problema della verifica delle ipotesi consiste allora nel suddividere, mediante opportuni criteri, l’insieme dei possibili campioni in due sottoinsiemi, un sottoinsieme A di **accettazione dell’ipotesi nulla** e un sottoinsieme R di **rifiuto 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;

Immagine che contiene tavolo

Descrizione generata automaticamenteTabella 1: Errori di tipo I e II in ambito giudiziario

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 : *ϑ* = *ϑ*0 e *H*1 : *ϑ ϑ*0
* Test unilaterali, divisi in:
* Test unilaterale sinistro: *H*0 : *ϑ ≤ ϑ*0 e *H*1 : *ϑ > ϑ*0
* Test unilaterale destro: *H*0 : *ϑ ≥ ϑ*0 e *H*1 : *ϑ < ϑ*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 pvalue, 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 > α* l’ipotesi *H*0 non può essere rifiutata;
* se *p <*= *α* 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 σ2 NOTA

Sia X1, X2, . . . , Xn un campione casuale estratto da una popolazione normaledescritta da una variabile aleatoria X ∼ N(µ, σ) con varianza nota σ2. **Nel test bilaterale si considerano le ipotesi:**

* H0 : µ = µ0,
* H1 : µ ≠µ0

Essendo la varianza nota, l’ipotesi H0 è semplice, mentre l’ipotesi H1 è composita. Quando H0 è vera, gioca un ruolo fondamentale la variabile aleatoria

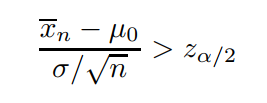
Immagine che contiene testo

Descrizione generata automaticamente

che è distribuita secondo una normale standard. Occorre osservare che Zn è una statistica (non è una variabile di pivot) poiché dipende esclusivamente dal campione casuale essendo µ0 e σ2 noti. Il test bilaterale ψ di misura α per le ipotesi considerate è il seguente:

* Immagine che contiene testo, orologio

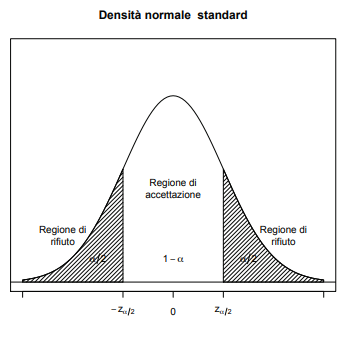
  Descrizione generata automaticamentesi accetti H0 se

Immagine che contiene testo, orologio

Descrizione generata automaticamente

* Si rifiuti H0 se oppure

Graficamente:

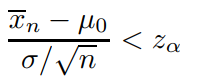




**TEST UNILATERALE SINISTRO**

Consideriamo le ipotesi:

* *H*0 : *µ <*= *µ*0
* *H*1 : *µ > µ*0

Le ipotesi H0 e H1 sono entrambe composite. Scegliamo il più grande valore di µ tale che l’ipotesi H0 sia vera. Il test unilaterale sinistro ψ di misura α per le ipotesi considerate è il seguente:

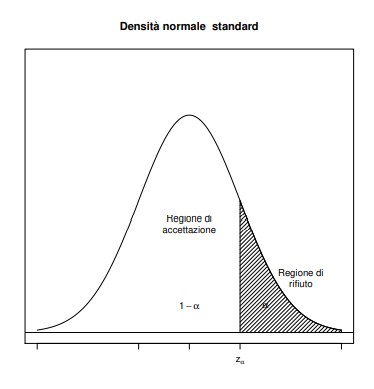
* Si accetti H0 se

Immagine che contiene testo

Descrizione generata automaticamente

* Si rifiuti H0 se

Graficamente:

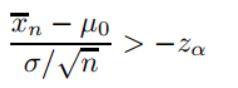




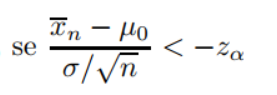
**TEST UNILATERALE DESTRO**

Consideriamo le ipotesi:

* *H*0 : *µ >*= *µ*0
* *H*1 : *µ < µ*0

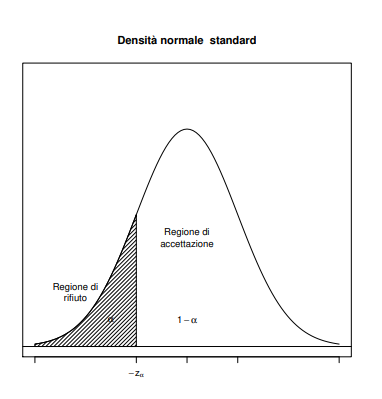
Le ipotesi H0 e H1 sono entrambe composite. Scegliamo il più piccolo valore di µ tale che l’ipotesi H0 è vera. Il test unilaterale destro ψ di misura α per le ipotesi considerate è il seguente:

* Si accetti H0 se



* Si rifiuti H0 se

Graficamente:





Considerando che il campione normale creato in precedenza con **rnorm** (10000, mean=1000, sd =1.5) ha un numero di componenti molto elevato e che la deviazione standard è piccola rispetto alla media, si mostra un campione non generato in modo randomico e con un numero di componenti pari a 50, in cui i test sono più concreti.

Si mostrano i test prima peri il campione piccolo e poi per quello grande.

Si mostra il codice in R dove andiamo a definire il nuovo campione e i valori caratterizzanti del campione.

Immagine che contiene testo

Descrizione generata automaticamente

Si mostrano in R i vari risultati dei test

**Nel test bilaterale si considerano le ipotesi:**

* H0 : µ = µ0,
* H1 : µ ≠µ0

Test per il campione piccolo.

Immagine che contiene testo

Descrizione generata automaticamente

Si nota che *zos* cade all’interno della zona di accettazione e il *p-value > α* ci consiglia di accedere all’ipotesi.

Test per il campione grande.

Immagine che contiene testo

Descrizione generata automaticamente

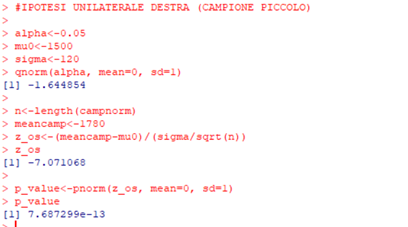
Si nota che *zos* cade all’interno della zona di accettazione e il *p-value > α* ci consiglia di accedere all’ipotesi.

**Nel test unilaterale destro**

Consideriamo le ipotesi:

* *H*0 : *µ <*= *µ*0
* *H*1 : *µ > µ*0

Test per il campione piccolo.



Si nota che *zos***non** cade all’interno della zona di accettazione e il *p-value < α* ci consiglia di non accedere all’ipotesi.

Test per il campione grande.

Immagine che contiene testo

Descrizione generata automaticamente

Si nota che *zos* cade all’interno della zona di accettazione e il *p-value > α* ci consiglia di accedere all’ipotesi.

**Nel test unilaterale sinistro**

Consideriamo le ipotesi:

* *H*0 : *µ >*= *µ*0
* *H*1 : *µ < µ*0

Test per il campione piccolo.

Immagine che contiene testo

Descrizione generata automaticamente

Si nota che *zos* cade all’interno della zona di accettazione e il *p-value > α* ci consiglia di accedere all’ipotesi.

Test per il campione grande.

Immagine che contiene testo

Descrizione generata automaticamente

Si nota che *zos* cade all’interno della zona di accettazione e il *p-value > α* ci consiglia di accedere all’ipotesi.

### 4.2.2 TEST SU *µ* CON VARIANZA *σ2* NON NOTA

**TEST BILATERALE**

Sia X1, X2, . . . , Xn un campione casuale estratto da una popolazione normale con varianza non nota σ2.

**Consideriamo le ipotesi:**

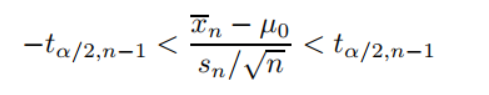
* *H*0 : *µ* = *µ*0
* *H*1 : *µ ≠* *µ*0

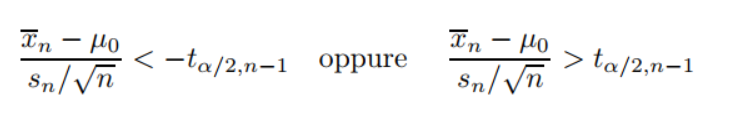
Essendo la varianza non nota entrambe le ipotesi sono composite. Quando *H*0 è vera, gioca un ruolo fondamentale la variabile aleatoria:

Immagine che contiene testo

Descrizione generata automaticamente

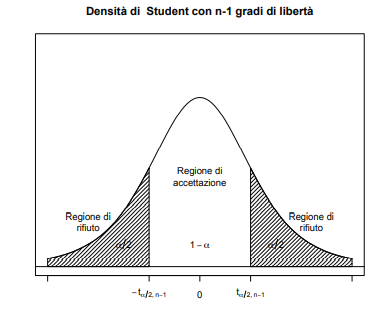
Che è distribuita con legge di Student con n-1 gradi di libertà. Notiamo che Tn è una statistica poiché dipende esclusivamente dal campione casuale considerato. Il test bilaterale ψ di misura α per le ipotesi considerate è il seguente:

* Si accetti H0 se



* Si rifiuti H0 se

Graficamente:

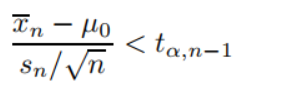




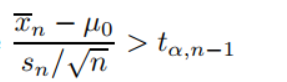
**TEST UNILATERALE SINISTRO**

**Consideriamo le ipotesi:**

* *H*0 : *µ <*= *µ*0
* *H*1 : *µ > µ*0

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

* Si accetti H0 se



* Si rifiuti H0 se

Graficamente:

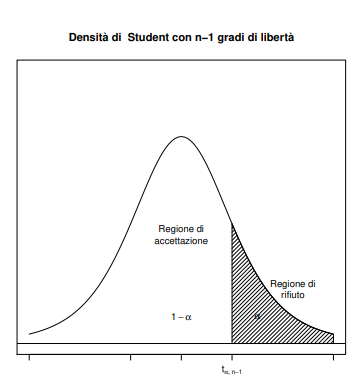


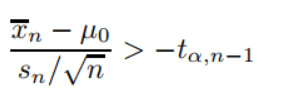
Immagine che contiene testo

Descrizione generata automaticamente

**TEST UNILATERALE DESTRO**

Consideriamo le ipotesi:

* H0 : µ ≥ µ0
* H1 : µ < µ0

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

* Si accetti H0 se

Immagine che contiene testo, orologio

Descrizione generata automaticamente

* Si rifiuti H0 se

Graficamente:

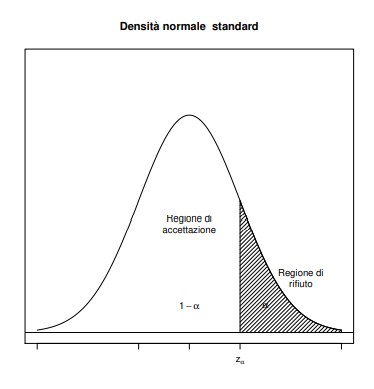


Immagine che contiene testo

Descrizione generata automaticamente

Si mostrano in R i vari risultati dei test

**Nel test bilaterale**

**Consideriamo le ipotesi:**

* *H*0 : *µ* = *µ*0
* *H*1 : *µ ≠* *µ*0

**Test per il campione piccolo**

**Immagine che contiene testo

Descrizione generata automaticamente**

Si nota che *tα/*2*,n-*1 = 2 e *tos* = 1.2 cade dentro la regione di accettazione, con fermato anche da un p-value maggiore di *α*.

**Test per il campione grande**

Immagine che contiene testo

Descrizione generata automaticamente

Si nota che *tα/*2*,n-*1 = 1.96 e *tos* = -1.16 cade dentro la regione di accettazione, con fermato anche da un p-value maggiore di *α*.

**Nel test unilaterale sinistro**

**Consideriamo le ipotesi:**

* *H*0 : *µ <*= *µ*0
* *H*1 : *µ > µ*0

**Test per il campione piccolo**

**Immagine che contiene testo

Descrizione generata automaticamente**

Si nota che *tα/*2*,n-*1 = 1.67 e *tos* = 0.2 cade dentro la regione di accettazione, confermato anche da un p-value maggiore di *α*.

**Test per il campione grande**

Immagine che contiene testo

Descrizione generata automaticamente

Si nota che *tα/*2*,n-*1 = 1.64 e *tos* = -1.16 cade dentro la regione di accettazione, confermato anche da un p-value maggiore di *α*.

**Nel test unilaterale destro**

**Consideriamo le ipotesi:**

* H0 : µ ≥ µ0
* H1 : µ < µ0

**Test per il campione grande**

Immagine che contiene testo

Descrizione generata automaticamente

Si nota che *tα/*2*,n-*1 = -1.67 e *tos* = -6*.*9 non cade dentro la regione di accettazione, confermato anche da un p-value < di *α*.

**Test per il campione grande**

Immagine che contiene testo

Descrizione generata automaticamente

Si nota che *tα/*2*,n-*1 = -1.64 e *tos* = -1*.*16 cade dentro la regione di accettazione in quanto < - *tα/*2*,n-*1, confermato anche da un p-value maggiore di *α*.

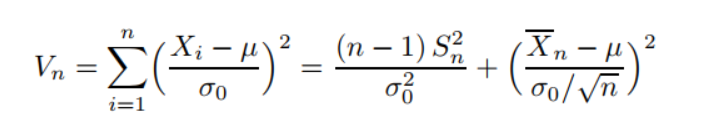
### 4.2.3 TEST SU *σ*2 CON VALORE MEDIO *µ* NOTO

**TEST BILATERALE**

Consideriamo le ipotesi:

* *H*0 : *σ*2 =
* *H*1 : *σ*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:

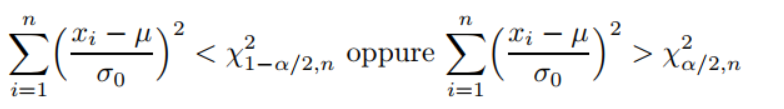


distribuita con legge del chi-quadrato con n-1 gradi di libertà.

Immagine che contiene testo, orologio

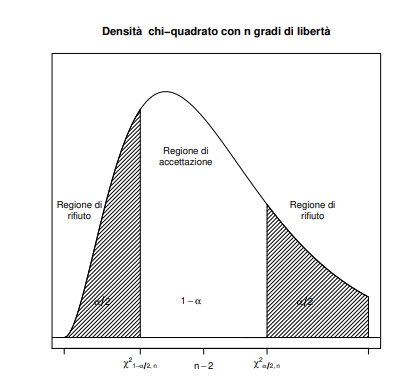
Descrizione generata automaticamenteIl test bilaterale *ψ* di misura *α* per le ipotesi considerate è:

* Si accetti H0 se



* Si rifiuti H0 se

Graficamente:



**TEST UNILATERALE SINISTRO**

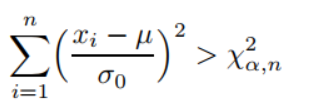
Consideriamo le ipotesi:

* *H*0 : *σ*2 <=
* *H*1 : *σ*2 >

Immagine che contiene testo, orologio

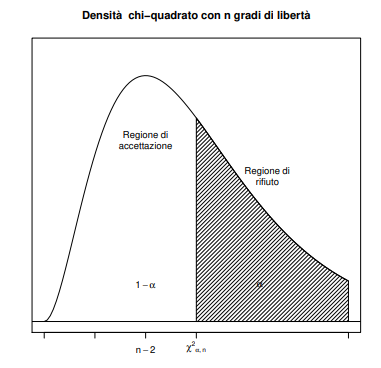
Descrizione generata automaticamenteLe ipotesi sono entrambe composite. Il test unilaterale sinistro *ψ* di misura *α* per le ipotesi considerate è:

* Si accetti H0 se



* Si rifiuti H0 se

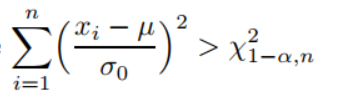
Graficamente:



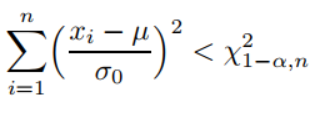
**TEST UNILATERALE DESTRO**

Consideriamo le ipotesi:

* *H*0 : *σ*2 >=
* *H*1 : *σ*2 <

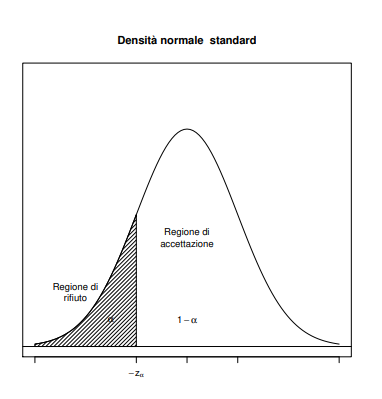
Le ipotesi sono entrambe composite. Il test unilaterale destro *ψ* di misura *α* per le ipotesi considerate è:

* Si accetti H0 se



* Si rifiuti H0 se

Graficamente:



Si mostrano in R i vari risultati dei test

Consideriamo le ipotesi:

* *H*0 : *σ*2 =
* *H*1 : *σ*2 ≠

Immagine che contiene testo

Descrizione generata automaticamente

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:

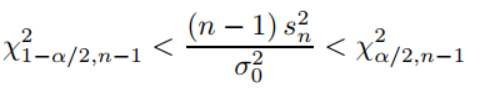
* *H*0 : *σ*2 =
* *H*1 : *σ*2 ≠

Essendo la varianza non nota entrambe sono composite. Quando *H*0 è vera, gioca un ruolo fondamentale la variabile aleatoria:

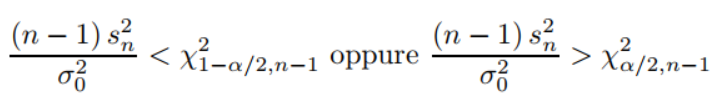
Immagine che contiene testo, orologio

Descrizione generata automaticamente

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

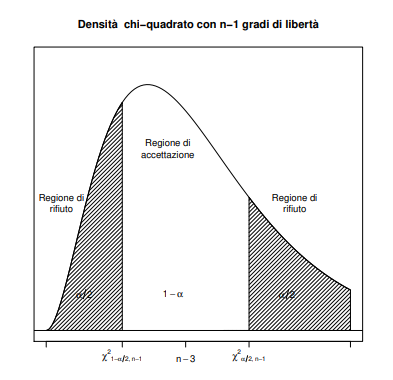
Il test bilaterale *ψ* di misura *α* per le ipotesi considerate è:

* Si accetti H0 se



* Si rifiuti H0 se

Graficamente:



**TEST UNILATERALE SINISTRO**

Consideriamo le ipotesi:

* *H*0 : *σ*2 <=
* *H*1 : *σ*2 >

Immagine che contiene testo, orologio

Descrizione generata automaticamenteLe ipotesi sono entrambe composite. Il test unilaterale sinistro *ψ* di misura *α* per le ipotesi considerate è:

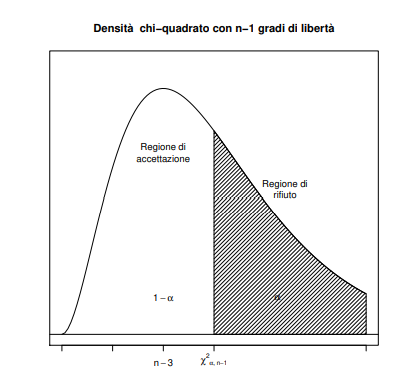
* Si accetti H0 se

Immagine che contiene testo

Descrizione generata automaticamente

* Si rifiuti H0 se

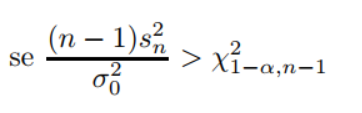
Graficamente:



**TEST UNILATERALE DESTRO**

Consideriamo le ipotesi:

* *H*0 : *σ*2 >=
* *H*1 : *σ*2 <

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

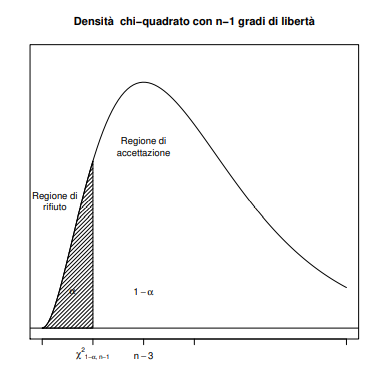
* Si accetti H0 se

Immagine che contiene testo

Descrizione generata automaticamente

* Si rifiuti H0 se

Graficamente:



Si mostrano in R i vari risultati dei test:

Consideriamo le ipotesi:

* *H*0 : *σ*2 =
* *H*1 : *σ*2 ≠

Immagine che contiene testo

Descrizione generata automaticamente

Notiamo che *χ*2 è compreso nell’area di accettazione quindi si accetta l’ipotesi nulla di un *σ*2 = 2*.*5

Consideriamo le ipotesi:

* *H*0 : *σ*2 <=
* *H*1 : *σ*2 >

Immagine che contiene testo

Descrizione generata automaticamente

Si nota che il valore 8858 è nell’area di accettazione in quanto minore di 10232.

Consideriamo le ipotesi:

* *H*0 : *σ*2 >=
* *H*1 : *σ*2 <

Immagine che contiene testo

Descrizione generata automaticamente

Si nota che il valore 8858 non è nell’area di accettazione in quanto minore di 9767.

# 

# CAPITOLO 5

# CRITERIO DEL CHI-QUADRATO

In molti problemi reali, si desidera verificare se il campione osservato può essere stato estratto da una popolazione descritta da una variabile aleatoria X con funzione di distribuzione FX(x), con k parametri non noti da stimare. A questo scopo, utilizzeremo il criterio di verifica delle ipotesi del chi–quadrato, detto anche test del chi–quadrato o test del buon adattamento.

Denotando con H0 l’ipotesi soggetta a verifica (**ipotesi nulla**) e con H1 **l’ipotesi alternativa**, il test chi–quadrato con livello di significatività *α* mira a verificare l’ipotesi nulla:

* H0: X ha una funzione di distribuzione FX(x) (avendo stimato k parametri non noti in base al campione)

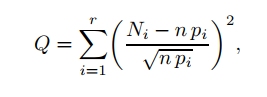
In alternativa all’ipotesi:

* H1: X non ha una funzione di distribuzione FX(x),

dove α è la probabilità massima di rifiutare l’ipotesi nulla quando essa è vera.

Occorre determinare un test ψ con livello di significatività α che permetta di determinare una regione di accettazione e di rifiuto dell’ipotesi nulla. Il test di verifica delle ipotesi considerato è bilaterale (o a due code). Suddividiamo dunque l’insieme dei valori che può assumere la variabile aleatoria *X* in *r* sottoinsiemi, in modo tale che la probabilità *pi* rappresenti la probabilità secondo la distribuzione ipotizzata che la variabile aleatoria assuma un valore appartenente al sottoinsieme *Ii*.(*pi* = *P*(*X ∈ Ii*))

Dal campione osserviamo le frequenze assolute *n*1*, n*2*, ..., nr* in cui gli elementi del  
campione si distribuiscono rispettivamente in *I*1*, I*2*, ..., Ir*. Quindi *ni* rappresenta il numero degli elementi del campione che cadono nell’intervallo *Ii* (*i* = 1*,* 2*, ..., r*). Vale chiaramente che la somma dei *pi* è unitaria, mentre la somma degli *ni* = *n*. Il criterio del chi-quadrato si basa sulla seguente statistica:



dove Ni è la variabile aleatoria che descrive il numero degli elementi del campione casuale X1, X2, . . . , Xn (costituito da n variabili aleatorie osservabili, indipendenti e identicamente distribuite con la stessa legge di probabilità FX(x) della popolazione) che cadono nell’intervallo Ii (i = 1, 2, . . . , r). Se la variabile aleatoria *X* ha una funzione di distribuzione *Fx*(*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.

Immagine che contiene antenna

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

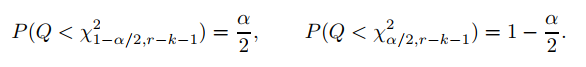
* Si accetti l’ipotesi H0 se



* Si rifiuti l’ipotesi H0 se



Dove sono soluzioni delle equazioni:



Vediamo in R per la distribuzione normale.

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

Immagine che contiene testo

Descrizione generata automaticamente

Gli intervalli sono *I*1 = (*-∞,* 998*.*7302), *I*2 = [998*.*7302*,* 999*.*6057), *I*3 = [999*.*6057*,* 1000*.*3598), *I*4 = [1001*.*3598*,* 1001*.*2353), *I*5 = [1001*.*2353*,* +*∞*).

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

Immagine che contiene testo

Descrizione generata automaticamente

Allora *n*1 = 1991, *n*2 = 1972, *n*3 = 2036, *n*4 = 2033, *n*5 = 1968.

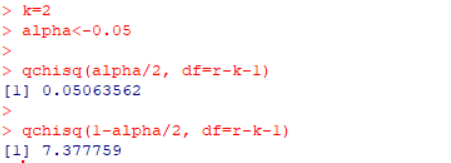
Calcoliamo ora *χ*2:

Immagine che contiene testo

Descrizione generata automaticamente

La distribuzione normale ha due parametri non noti (*µ, σ*2) e quindi *k* = 2. Pertanto, la funzione di distribuzione della statistica *Q* è approssimabile con la funzione di distribuzione chi–quadrato con *r - k -* 1 = 2 gradi di libertà.

Ora occorre calcolare con *α* = 0.05:



Essendo 0*.*0506 *< χ*2 = 2.137 *<* 7*.*3777 l’ipotesi *H*0 della popolazione normale può essere accettata.