# Università degli Studi di Salerno

**Corso di Statistica e Analisi dei Dati**


# Studio di un problema fisico con una variabile aleatoria normale

Progetto di Selice Andrea e Nisivoccia Giuseppe

**1 - INTRODUZIONE** **3**

**2 - DISTRIBUZIONE NORMALE** **4**
2.1 Densità di probabilità 4
2.2 Funzione di distribuzione 7
2.2.1 Regola 3σ 8
2.3 Quantili 8
2.4 Simulazione della variabile 8

**3 - STIMA PUNTUALE** **11**
3.1 Stimatori 11
3.2 Metodi per la ricerca di stimatori 11
3.2.1 Metodo dei momenti 11
3.2.2 Metodo della massima verosimiglianza 13
3.3 Proprietà degli stimatori 14

**4 - STIMA INTERVALLARE** **15**
4.1 Metodo pivotale 15
4.1.1 Intervallo di confidenza per μ con σ2 non noto 15
4.1.2 Intervallo di confidenza per σ2 con μ non noto 17
4.2 Confronto tra due popolazioni normali 17
4.2.1 Intervallo di confidenza per μ1 - μ2 con varianze non note 18

**5 - VERIFICA DELLE IPOTESI CON R** **20**
5.1 Test su μ con σ2 non nota 22
5.1.1 Test bilaterale 22
5.1.2 Test unilaterale sinistro 24
5.1.3 Test unilaterale destro 25
5.2 Test su σ2 con μ non noto 26
5.2.1 Test bilaterale 26
5.2.2 Test unilaterale sinistro 28
5.2.3 Test unilaterale destro 28

**6 - CRITERIO DEL CHI-QUADRATO** **29**

# 1 - INTRODUZIONE

La seguente analisi esamina i dati forniti da una variabile aleatoria continua (temperatura di un forno).

L'inferenza statistica ha lo scopo di estendere le misure ricavate da un campione alla popolazione da cui esso è stato estratto.
Nella prima parte del documento si tratta della distribuzione normale descrivendo le caratteristiche e le proprietà.
Successivamente si tratta di stime puntuali, stime intervallari, verifica delle ipotesi e criterio del chi-quadrato.
La stima di un parametro può essere:

- Puntuale
    - che restituisce un valore esatto per il parametro della popolazione con una precisione prefissata
- Intervallare
    - che restituisce un campo di valori che, con un fissato margine di errore, contiene il valore vero del parametro

La verifica delle ipotesi consiste nel determinare un test che permetta di suddividere l'insieme dei possibili campioni in due sottoinsiemi, ovvero una regione di accettazione o rifiuto di una data ipotesi, definita ipotesi nulla.
Il criterio del chi-quadrato verifica l'ipotesi che una certa popolazione, descritta da una variabile aleatoria, sia caratterizzata da una certa funzione di distribuzione con k parametri non noti da stimare.

# 2 - DISTRIBUZIONE NORMALE

Una variabile aleatoria è una variabile che può assumere valori diversi in dipendenza da qualche fenomeno aleatorio.
Il termine “aleatorio” fa riferimento al fatto che ci occupiamo degli esiti possibili di un esperimento aleatorio, ovvero, di un esperimento il cui esito è incerto prima che l’esperimento venga eseguito.
Una variabile aleatoria si dice discreta se può assumere solo un numero finito o infinito numerabile di valori.
Una variabile aleatoria si dice continua se può assumere tutti gli infiniti valori dei numeri reali o di un loro intervallo [a,b].

Una variabile aleatoria normale è una distribuzione di probabilità continua che è spesso usata per descrivere variabili casuali a valori reali che tendono a concentrarsi attorno a un singolo valor medio, viene spesso considerata un buon modello per variabile fisiche come la temperatura di un forno.

## 2.1 Densità di probabilità

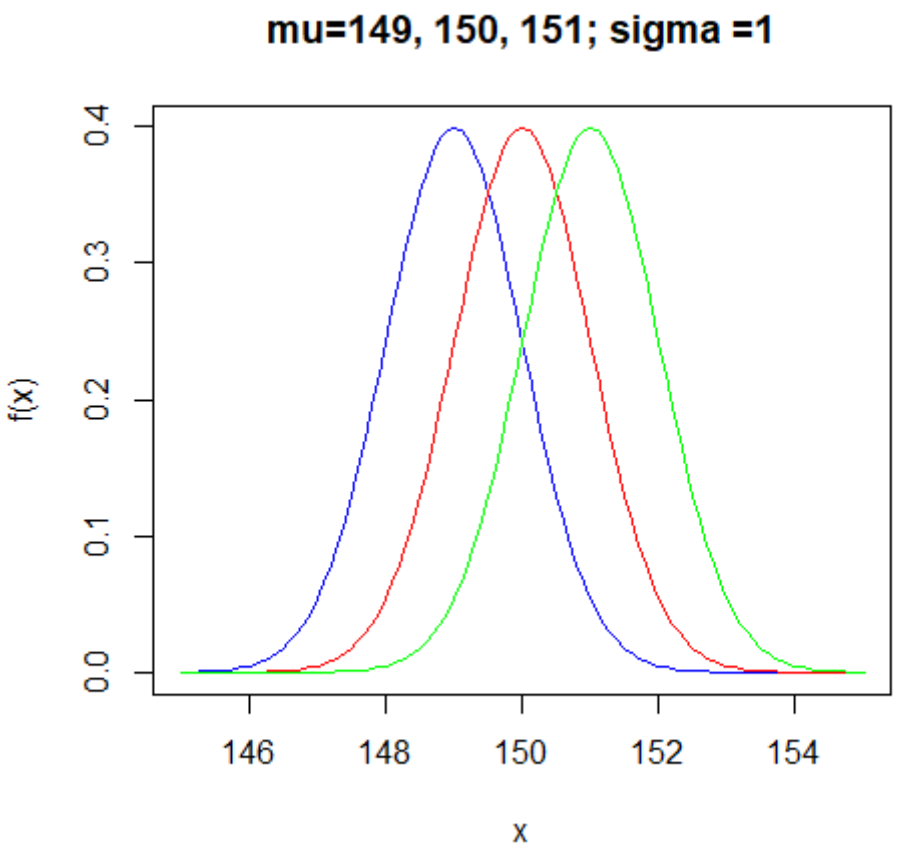
Una variabile aleatoria X di densità di probabilità

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

si dice avere distribuzione normale di parametri μ e σ.
Per ogni x ∈ R risulta quindi $f_x(\mu - x) = f_x(\mu + x)$, la densità normale è simmetrica rispetto all’asse x = μ.
La densità $f_x(x)$ presenta il massimo $(\sigma\sqrt{2\pi})^{-1}$ nel punto di ascissa x = μ e due flessi nei punti di ascisse μ - σ e μ+ σ.
Il grafico della densità presenta una caratteristica forma a campana, simmetrica rispetto a x = μ.
Per indicare una variabile aleatoria X che ha distribuzione normale di parametri μ e σ si usa la notazione X~N(μ,σ).
Di seguito viene riportato il calcolo della densità normale modificando il parametro μ:

```
curve(dnorm (x,mean =149, sd =1) ,from =145, to=155,
xlab="x",ylab="f(x)",main="mu=149, 150, 151; sigma =1", col="blue")

curve(dnorm (x,mean =150, sd =1) ,from =145, to=155, xlab="x",ylab="f(x)",
add=TRUE, col="red")
```

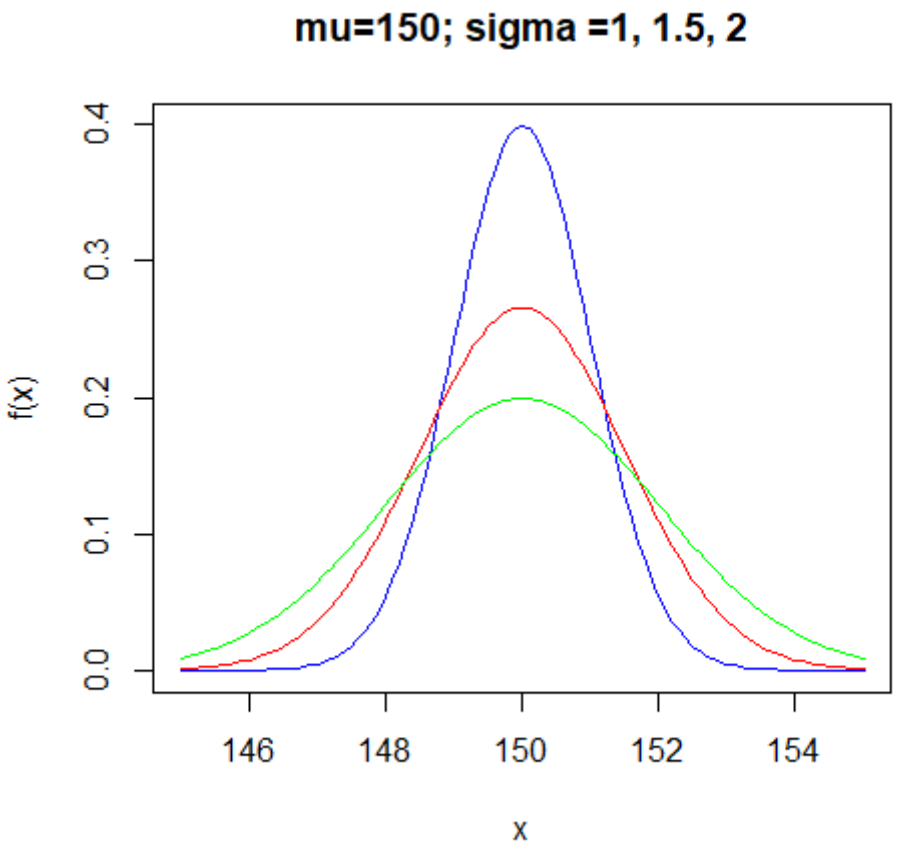
```
curve(dnorm (x,mean =151, sd =1) ,from =145, to=155, xlab="x",ylab="f(x)",
add=TRUE, col="green")
```




Le tre curve descrivono la funzione di densità normale con media pari a 149, 150 e 151. Al variare del parametro μ la curva viene traslata lungo l'asse delle ascisse, ma la forma non cambia.

Di seguito viene riportato il calcolo della densità normale modificando il parametro σ:

```
curve(dnorm (x,mean =150, sd =1) ,from =145, to=155,
xlab="x",ylab="f(x)",main="mu=150; sigma =1, 1.5, 2", col="blue")

curve(dnorm (x,mean =150, sd =1.5) ,from =145, to=155, xlab="x",ylab="f(x)",
add=TRUE, col="red")

curve(dnorm (x,mean =150, sd =2) ,from =145, to=155, xlab="x",ylab="f(x)",
add=TRUE, col="green")
```




Le tre curve descrivono la funzione di densità normale con deviazione standard pari a 1, 1.5 e 2. La larghezza della funzione dipende dal parametro σ, se quest'ultimo aumenta allora la curva risulta sempre più piatta, se diminuisce allora la curva si allunga verso l'alto. Questo è dovuto in quanto l'ordinata massima è inversamente proporzionale a σ, inoltre l'area sottesa della densità rimane sempre unitaria.

## 2.2 Funzione di distribuzione

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

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

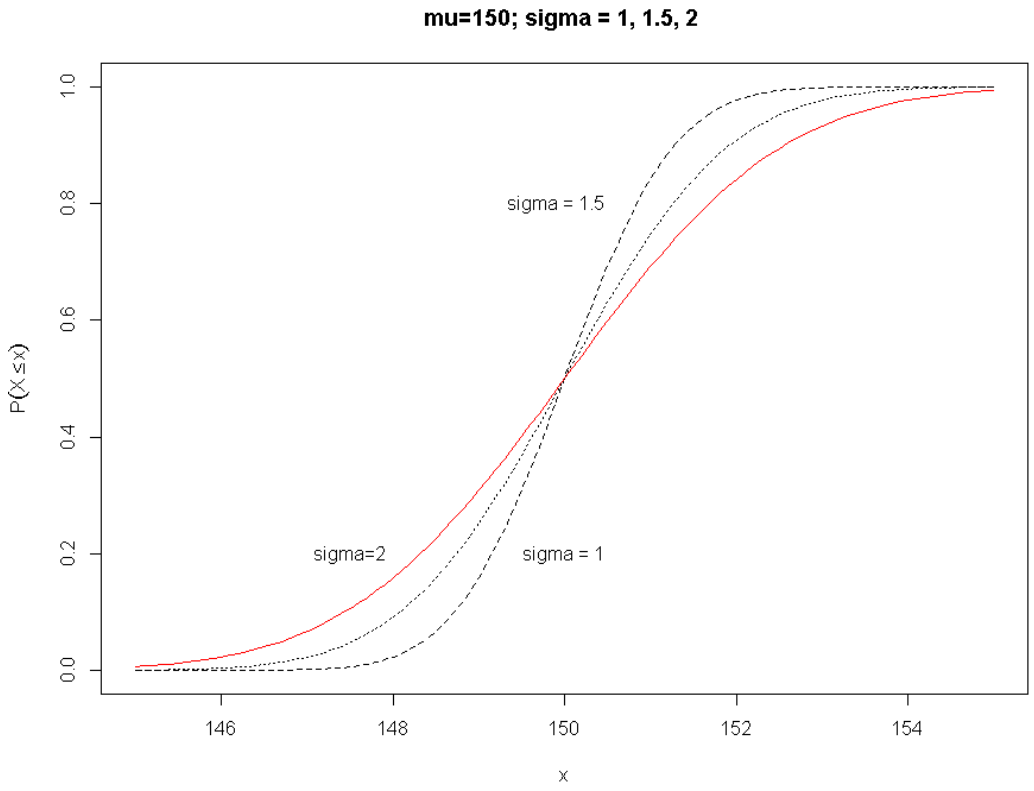
dove

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

è la funzione di distribuzione di una variabile aleatoria $Z \sim N(0, 1)$, detta normale standard.

Di seguito viene mostrata la funzione per il calcolo della funzione di distribuzione:

```
curve(pnorm (x,mean=150,sd = 1) ,from=145, to=155, xlab="x",ylab=expression
(P(X<=x)),main="mu=150; sigma = 1, 1.5, 2 ",lty =2)
text (150,0.2, "sigma = 1")

curve(pnorm(x,mean=150,1.5),add=TRUE,lty=3)
text(149.9,0.8,"sigma = 1.5")

curve(pnorm(x,mean=150,sd=2), add=TRUE, col=”red”)
text(147.5,0.2,"sigma=2")
```



### 2.2.1 Regola 3σ

La regola del 3σ dice che per una qualsiasi variabile aleatoria X ~ N(μ, σ) risulta:

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

Ovvero che 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 regola permette di individuare l'intervallo (μ - 3σ, μ + 3σ) in cui rappresentare la funzione di densità di una variabile normale di valore medio μ e varianza $\sigma^2$ in modo tale che l'area sottesa della curva sia circa unitaria e l'area delle code destra e sinistra sia trascurabile.
Di seguito viene riportata la regola del 3σ per una variabile aleatoria Z ~ N(150, 2):

```
pnorm(156, mean = 150, sd = 2) - pnorm(144, mean = 150, sd = 2)
[1] 0.9973002
```

## 2.3 Quantili

Di seguito viene mostrata la funzione per calcolare i quantili della distribuzione normale:

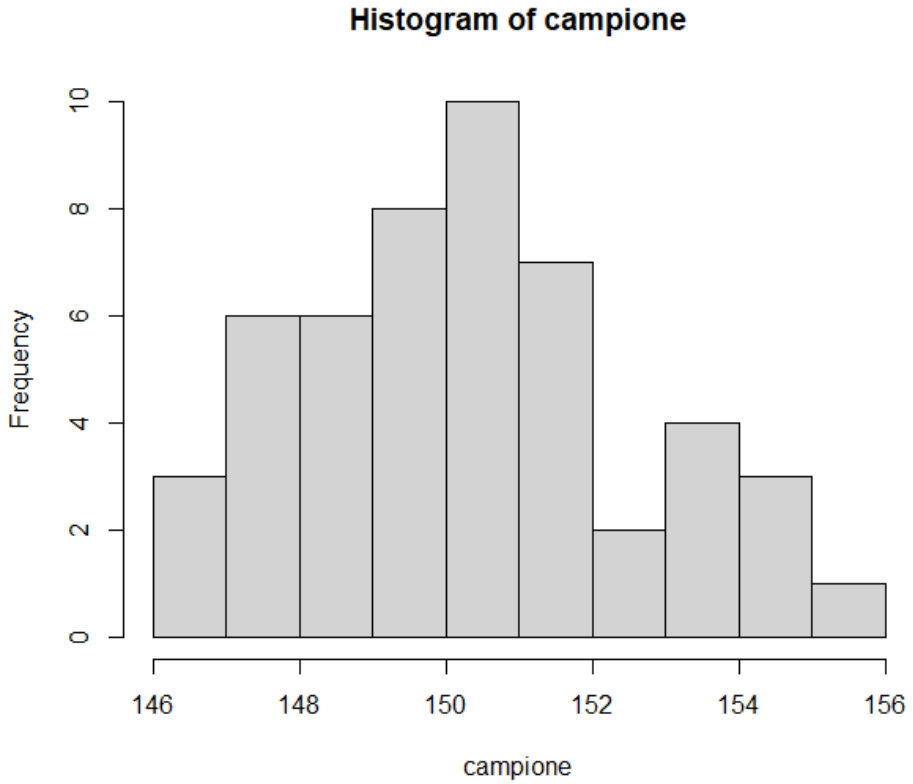
```
quantili<-c(0,0.25,0.5,0.75,1)
qnorm(quantili, mean=150, sd=2)

[1]    -Inf 148.651 150.000 151.349     Inf
```

La funzione qnorm restituisce il più piccolo numero x assunto dalla variabile aleatoria normale X tale che $P(X \leq x) \geq quantili$

## 2.4 Simulazione della variabile

Di seguito viene riportato il comando per simulare in R una variabile normale:

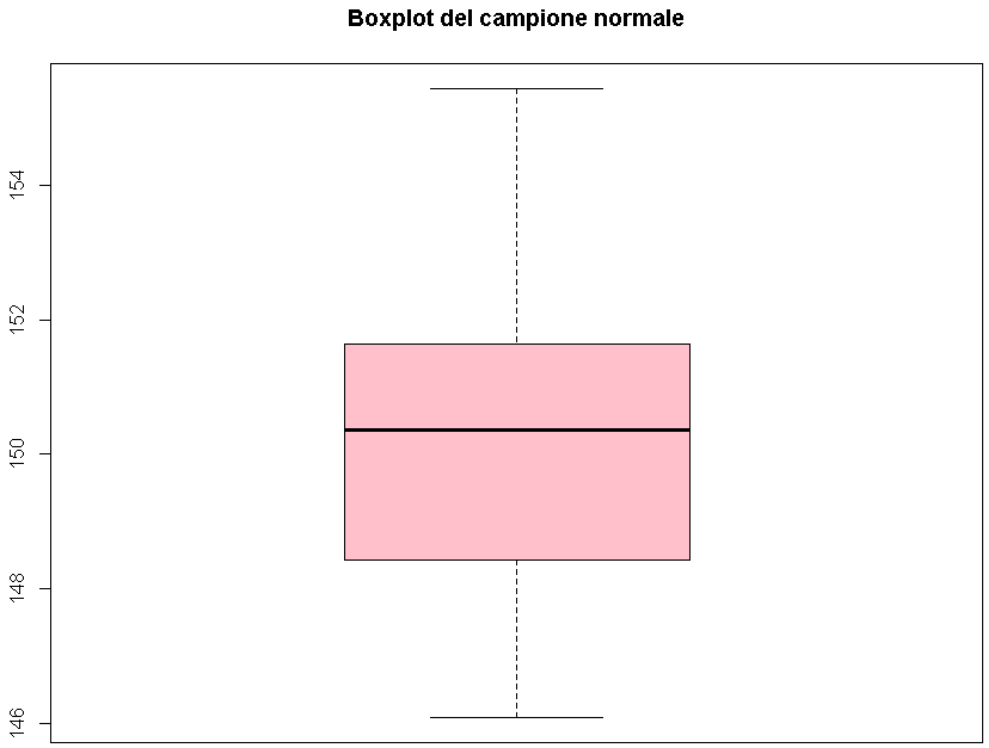
```
rnorm(50, mean = 150, sd = 2)
```




| Indici di sintesi del campione | |
|---|---|
| Media | 150.285 |
| Varianza | 5.296842 |
| Deviazione Standard | 2.301487 |

Di seguito vengono riportati i quantili del campione e il boxplot:

| 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|
| 146.0908 | 148.4542 | 150.3658 | 151.5651 | 155.4346 |


Boxplot del campione normale
154
152
150
148
146

# 3 - STIMA PUNTUALE

## 3.1 Stimatori

Nei metodi di indagine dell'inferenza statistica si considera un campione casuale $X_1$, $X_2$, …, $X_n$ di ampiezza n che viene estratto dalla popolazione e si cerca di ottenere le informazioni sui parametri non noti usando variabili aleatorie dette stimatori.
Uno stimatore è una funzione misurabile e osservabile del campione casuale i cui valori possono essere usati per stimare un parametro non noto della popolazione.
I valori che assumono questi stimatori sono detti stime del parametro non noto.
Da un campione casuale $X_1$, $X_2$, …, $X_n$ estratto da una popolazione descritta da una variabile aleatoria osservabile X caratterizzata da valore medio E(X) = μ finito e varianza Var(X) = $\sigma^2$ finita risulta:

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

Questa proposizione mostra che al crescere dell'ampiezza del campione la media campionaria fornisce una stima sempre più accurata del valore medio della popolazione. Dal teorema centrale di convergenza si può ricavare che per n sufficientemente grande la funzione di distribuzione della media campionaria è approssimativamente normale con valore medio μ e varianza $\sigma^2/n$.

## 3.2 Metodi per la ricerca di stimatori

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

### 3.2.1 Metodo dei momenti

Per descrivere il metodo dei momenti occorre definire il momento campionario.
Il momento campionario r-esimo relativo ai valori osservati del campione casuale è definito come:

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

ovvero è la media aritmetica delle potenze r-esime delle n osservazioni che sono state effettuate sulla popolazione.
Se r=1 si ottiene la media campionaria.
Il metodo dei momenti consiste nell'uguagliare i primi k momenti della popolazione con i corrispondenti momenti del campione casuale.

Tale metodo consiste nel risolvere il seguente sistema di k equazioni:

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

I termini a sinistra del sistema dipendono dalla legge di probabilità e contengono i parametri non noti della popolazione. I termini a destra possono essere calcolati a partire dai dati osservati del campione.
Il metodo dei momenti è utilizzabile nel momento in cui il sistema ammette un'unica soluzione.
Le stime dei parametri ottenuti da tale metodo dipendono dal campione osservato e al variare dei possibili campioni si ottengono stimatori dei parametri non noti della popolazione, detti stimatori del metodo dei momenti.
Per una popolazione normale occorre stimare μ e $\sigma^2$.
Poiché E(X) = μ e $E(X^2) = \sigma^2 + \mu^2$ si ottiene:

$$\widehat{\mu} = \frac{x_1 + x_2 + \ldots + x_n}{n}, \qquad \widehat{\sigma^2} + \widehat{\mu}^2 = \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}$$

Dalla seconda equazione si ricava:

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

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{(n-1)\,s^2}{n}.$$

Il metodo dei momenti fornisce come stimatore del valore medio μ la media campionaria e come stimatore della varianza $\sigma^2$ la variabile aleatoria $(n-1)s^2/n$.
Di seguito viene riportata la stima dei parametri del campione considerato:

```
stima_medio <- mean(campione)
stima_medio
[1] 150.285

stima_varianza <- (length(campione)-1) * var(campione)/length(campione)
stima_varianza
[1] 5.190905
```

La stima del del valore medio μ corrisponde a 150.285; la stima della varianza $\sigma^2$ corrisponde a 5.190905.

### 3.2.2 Metodo della massima verosimiglianza

Il metodo della massima verosimiglianza è un procedimento matematico per determinare uno stimatore, ed è il più importante metodo per la stima dei parametri non noti di una popolazione.
Di seguito viene riportata la definizione di "Funzione di verosimiglianza":
Sia $X_1,X_2,...,X_n$ un campione casuale di ampiezza n estratto dalla popolazione.
La funzione di verosimiglianza $L(\vartheta_1,\vartheta_2,...,\vartheta_k; x_1,x_2,...,x_n)$ del campione osservato $(x_1,x_2,...,x_n)$ 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 $X_1,X_2,...,X_n$, ossia

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

Il metodo della massima verosimiglianza consiste nel massimizzare la funzione di verosimiglianza rispetto ai parametri $\vartheta_1,\vartheta_2,...,\vartheta_k$, cioè si cerca di determinare da quale funzione di probabilità congiunta è più verosimile che provenga il campione osservato.
I valori di $\vartheta_1,\vartheta_2,...,\vartheta_k$ che massimizzano la funzione di verosimiglianza sono indicati con $\widehat{\vartheta}_1,\widehat{\vartheta}_2,...,\widehat{\vartheta}_k$; essi costituiscono le stime di massima verosimiglianza dei parametri non noti $\vartheta_1,\vartheta_2,...,\vartheta_k$

Di seguito viene riportato il procedimento per stimare i parametri di una popolazione normale:
La funzione di densità della normale è

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

Si ha

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

e quindi si ha che le stime di massima verosimiglianza dei parametri $\mu$ e $\sigma^2$ sono rispettivamente

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

Lo stimatore di massima verosimiglianza e dei momenti del valore medio $\mu$ è la media campionaria $\overline{X}$.

Lo stimatore di massima verosimiglianza e dei momenti della varianza $\sigma^2$ è $(n-1)S^2/n$

## 3.3 Proprietà degli stimatori

Uno stimatore può essere:

- Corretto (o equivalentemente non distorto)
- Più efficiente di un altro
- Corretto e con varianza uniformemente minima
- Asintoticamente corretto
- Consistente

Uno stimatore $\widehat{\Theta} = t(X_1,X_2,...,X_n)$ del parametro non noto $\vartheta$ della popolazione è detto corretto se e solo se per ogni $\vartheta \in \Theta$ si ha

$$E(\widehat{\Theta}) = \vartheta,$$

Ossia se il valore medio dello stimatore $\widehat{\Theta}$ è uguale al corrispondente parametro non noto della popolazione.
Ci possono essere più stimatori corretti e si utilizzano dei criteri per confrontare stimatori dello stesso parametro.

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

# 4 - STIMA INTERVALLARE

## 4.1 Metodo pivotale

Il metodo pivotale consiste essenzialmente nel determinare una variabile aleatoria di pivot $\gamma(X_1, X_2,..., X_n; \vartheta)$ che dipende dal campione casuale $X_1, X_2,..., X_n$ e dal parametro non noto $\vartheta$ e la sua funzione di distribuzione non contiene il parametro $\vartheta$ da stimare. La variabile aleatoria di pivot non è una statistica poiché dipende dal parametro non noto $\vartheta$ e quindi non è osservabile.

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

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

Se per ogni possibile campione osservato x = ($x_1$, $x_2$, . . . , $x_n$) e per ogni $\vartheta \in \Theta$, si riesce a dimostrare che

$$\alpha_1 < \gamma(\mathbf{x}; \vartheta) < \alpha_2 \quad \Longleftrightarrow \quad g_1(\mathbf{x}) < \vartheta < g_2(\mathbf{x})$$

con $g_1(x)$ e $g_2(x)$ dipendenti soltanto dal campione osservato, quindi la probabilità precedente è equivalente a

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

Denotando con $\underline{C}_n = g_1(X_1, X_2,..., X_n)$ e $\overline{C}_n = g_2(X_1, X_2,..., X_n)$, allora $(\underline{C}_n, \overline{C}_n)$ è un intervallo di confidenza di grado 1 - α per il parametro $\vartheta$ non noto della popolazione

### 4.1.1 Intervallo di confidenza per μ con $\sigma^2$ non noto

Per determinare un intervallo di confidenza di grado 1−α per il valore medio μ nel caso in cui la varianza $\sigma^2$ della popolazione normale non è nota con il metodo pivotale, si utilizza la variabile aleatoria di pivot

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

dove

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

denota la varianza campionaria.
Si dimostra che questa variabile è distribuita con legge di Student con n-1 gradi di libertà.

Di seguito si utilizzano $a_1 = -\ t_{a/2,n-1}\ e\ \alpha_2 = t_{a/2,n-1}$ quindi:

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

quindi una stima dell'intervallo di confidenza 1-α per il valore medio μ è:

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

Di seguito si stima il parametro con l'intervallo di confidenza trovato e con $\alpha = 0.05$

```
alpha<-1-0.95
devSd<-sd(campione)
devSd
[1] 2.301487

n<-length(campione)

mean(campione)-qt(1-alpha/2,df=n-1)*devSd/sqrt(n)
[1] 149.6309

mean(campione)+qt(1-alpha/2,df=n-1)*devSd/sqrt(n)
[1] 150.939
```

### 4.1.2 Intervallo di confidenza per $\sigma^2$ con $\mu$ non noto

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, si utilizza la variabile aleatoria di pivot

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

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

Di seguito si utilizzano $a_1 = \chi^2_{1-\alpha/2,n-1}$ $e$ $\alpha_2 = \chi^2_{\alpha/2,n-1}$ quindi:

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

quindi una stima dell'intervallo di confidenza 1-α per $\sigma^2$ è

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

Di seguito si stima il parametro con l'intervallo di confidenza trovato e con $\alpha = 0.05$

```
(n-1)*var(campione)/qchisq(1-alpha/2,df=n-1)
[1] 3.696046

(n-1)*var(campione)/qchisq(alpha/2,df=n-1)
[1] 8.225193
```

## 4.2 Confronto tra due popolazioni normali

Spesso si è interessati a stimare la differenza tra le medie di due distinte popolazioni, in questo caso si considerano due campioni $X_1, X_2, \ldots, X_{n1}$ e $Y_1, Y_2, \ldots, Y_{n2}$, casuali indipendenti di ampiezza $n_1$ ed $n_2$ estratti da due popolazioni normali $X \sim N(\mu_1, \sigma_1)$ e $Y \sim N(\mu_2, \sigma_2)$.
Di seguito viene riportato un nuovo campione per eseguire il confronto:

```
campione2 <- rnorm(30, mean = 100, sd = 2.5)
```

| Indici di sintesi del campione | |
|---|---|
| Media | 100.1147 |
| Varianza | 4.523432 |
| Deviazione Standard | 2.126836 |

```
#quantili del nuovo campione

     0%      25%      50%      75%      100%
94.95045 98.62839 100.66104 101.42695 103.83885
```

### 4.2.1 Intervallo di confidenza per $\mu_1 - \mu_2$ con varianze non note

L'obiettivo è determinare un intervallo di confidenza di grado 1 - α per la differenza $\mu_1 - \mu_2$ delle due popolazioni.
Sia $S^2_{n1}$ e $S^2_{n2}$ le varianze campionarie delle due popolazioni normali.
Le varianze campionarie delle due popolazioni normali sono stimatori corretti e consistenti delle varianze delle due popolazioni.
Per determinare l'intervallo di confidenza 1 - α si considera la seguente variabile aleatoria di pivot:

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

Applicando il metodo pivotale in forma approssimata si ottiene:

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

Da cui segue che la stima dell'intervallo di confidenza di grado 1 - α per la differenza tra le due medie $\mu_1$ - $\mu_2$ è la seguente:

$$\overline{x}_{n_1} - \overline{y}_{n_2} - z_{\alpha/2}\sqrt{\frac{s^2_{n_1}}{n_1} + \frac{\tilde{s}^2_{n_2}}{n_2}} < \mu_1 - \mu_2 < \overline{x}_{n_1} - \overline{y}_{n_2} + z_{\alpha/2}\sqrt{\frac{s^2_{n_1}}{n_1} + \frac{\tilde{s}^2_{n_2}}{n_2}}$$

Sia α = 0.05, di seguito viene riportata la stima dell'intervallo di confidenza di grado 1 - α = 0.95 per la differenza tra le due medie $\mu_1$ - $\mu_2$:

```
alpha <- 1-0.95
qnorm (1- alpha /2,mean =0, sd =1)
[1] 1.959964

n1 <- length(campione)
n2 <- length(campione2)

m1 <- mean(campione)
m2 <- mean(campione2)

s1 <- sd(campione)
s2 <- sd(campione2)

#stima del limite inferiore
m1-m2-qnorm(1-alpha/2,mean=0,sd=1)*sqrt(s1^2/n1+s2^2/n2)
[1] 49.17719

#stima del limite superiore
m1-m2+qnorm(1-alpha/2,mean=0,sd=1)*sqrt(s1^2/n1+s2^2/n2)
[1] 51.16331
```

Essendo positivi sia limite inferiore che superiore, è possibile dire che la media della prima popolazione è superiore alla media della seconda popolazione con grado di confidenza 1 - α = 0.95

# 5 - VERIFICA DELLE IPOTESI CON R

Gli elementi che costituiscono il procedimento relativo alla verifica delle ipotesi sono:

- una popolazione descritta da variabile aleatoria X, caratterizzata da una funzione di probabilità o densità di probabilità f(x; $\vartheta$)
- un'ipotesi su un parametro non noto $\vartheta$ della popolazione
- un campione casuale $X_1, X_2, \ldots, X_n$ estratto dalla popolazione

Un'ipotesi statistica è un'affermazione o una congettura sul parametro non noto $\vartheta$. Se l'ipotesi statistica specifica completamente f(x; $\vartheta$) è definita ipotesi semplice, altrimenti è definita ipotesi composita.

L'ipotesi soggetta a verifica ($H_0$) è definita ipotesi nulla. Il procedimento o la regola con cui si decide se accettare o rifiutare $H_0$ è definito come test di ipotesi. Per la costruzione del test è richiesta anche la formulazione di un'ipotesi alternativa ($H_1$), ovvero un'ipotesi che vada in contrapposizione all'ipotesi nulla.

Il problema della verifica delle ipotesi è dunque determinare un test $\psi$ 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 accetta come valida l'ipotesi nulla se il campione osservato appartiene ad A, si rifiuta se appartiene ad R.

Si possono verificare due tipi di errori:

- rifiutare l'ipotesi $H_0$ nel caso in cui tale ipotesi sia vera, questo errore è definito errore di tipo I

$$\alpha(\vartheta) = P(\text{rifiutare } \mathbf{H}_0|\vartheta), \qquad \vartheta \in \Theta_0;$$

- accettare l'ipotesi nulla $H_0$ nel caso in cui tale ipotesi sia falsa, questo errore è definito errore di tipo II

$$\beta(\vartheta) = P(\text{accettare } \mathbf{H}_0|\vartheta), \qquad \vartheta \in \Theta_1.$$

| | Rifiutare $\mathbf{H}_0$ | Accettare $\mathbf{H}_0$ |
|---|---|---|
| $\mathbf{H}_0$ vera | Errore del I tipo<br>Probabilità $\alpha$ | Decisione esatta<br>Probabilità $1-\alpha$ |
| $\mathbf{H}_0$ falsa | Decisione esatta<br>Probabilità $1-\beta$ | Errore del II tipo<br>Probabilità $\beta$ |

La misura della regione critica, o livello di significatività α, di un test fornisce la probabilità massima di rifiutare l'ipotesi nulla quando essa è vera, ovvero:

$$\alpha = \sup_{\vartheta \in \Theta_0} \alpha(\vartheta).$$

Nella costruzione di un test si va a fissare la probabilità di commettere un errore di tipo I e si cerca un test ψ che vada a minimizzare la probabilità di commettere un errore di tipo II. Per la probabilità di commettere un errore di tipo I si sceglie solitamente tra:

- 0.05, il test viene detto statisticamente significativo
- 0.01, il test viene detto statisticamente molto significativo
- 0.001, il test viene detto statisticamente estremamente significativo

I test statistici si suddividono in:

- test bilaterali

$$\mathbf{H}_0 : \vartheta = \vartheta_0$$
$$\mathbf{H}_1 : \vartheta \neq \vartheta_0,$$

- test unilaterali, che a sua volta si suddividono in test unilaterale sinistro e destro

$$\mathbf{H}_0 : \vartheta \leq \vartheta_0 \qquad \mathbf{H}_0 : \vartheta \geq \vartheta_0$$
$$\mathbf{H}_1 : \vartheta > \vartheta_0 \qquad \mathbf{H}_1 : \vartheta < \vartheta_0,$$

Le conclusioni di questi test dipendono dal livello di significatività α.

Nei test statistici si calcola anche il livello di significatività osservato, definito come p-value. Il p-value è definito come la probabilità, supposta vera l'ipotesi $H_0$, che la statistica del test assuma un valore uguale o più estremo di quello effettivamente osservato.

Il criterio del p-value è il seguente:

- se p>α, l'ipotesi $H_0$ non può essere rifiutata
- se p≤α, l'ipotesi $H_0$ deve essere rifiutata

In un test statistico è importante fissare il livello di significatività α prima di calcolare il p-value.

## 5.1 Test su μ con $\sigma^2$ non nota

### 5.1.1 Test bilaterale

Sia $X_1, X_2, \ldots, X_n$ un campione casuale estratto da una popolazione normale con varianza $\sigma^2$ non nota.
Si prendano in considerazione le seguenti ipotesi:

$$\mathbf{H}_0 : \mu = \mu_0 \qquad \mathbf{H}_1 : \mu \neq \mu_0$$

Entrambe sono composite, essendo la varianza non nota.
In analogo a quanto visto per gli intervalli di confidenza, si considera la seguente variabile aleatoria, distribuita con legge di Student con n-1 gradi di libertà:

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

Il test bilaterale ψ di misura α con le ipotesi sopra citate è il seguente:

- si accetta $H_0$ se:

$$-t_{\alpha/2,n-1} < \frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} < t_{\alpha/2,n-1}$$

- si rifiuta $H_0$ se:

$$\frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} < -t_{\alpha/2,n-1}$$

$$\frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} > t_{\alpha/2,n-1}$$

Per il test bilaterale si può anche calcolare il p-value, partendo dalla stima della statistica del test, denotata come segue:

$$t_{os} = \frac{\overline{x}_n - \mu_0}{(s_n/\sqrt{n})}$$

da cui si ricava il p-value:

$$\begin{aligned} pvalue &= P(T_n < -|t_{os}|) + P(T_n > |t_{os}|) = 2\,P(T_n > |t_{os}|) \\ &= 2\Big[1 - P(T_n \leq |t_{os}|)\Big] \end{aligned}$$

Di seguito viene riportato il test bilaterale sul campione considerato, per verificare l'ipotesi nulla $H_0$ : μ = 149 in alternativa all'ipotesi $H_1$ : μ ≠ 149:

```
alpha <- 0.05
mu0 <- 149
n <- 50
qt(1-alpha/2, df=n-1)
[1] 2.009575

m1 <- mean(campione)
s1 <- sd(campione)
(m1-mu0)/(s1/sqrt(n))
[1] 3.947905

#calcolo del p-value
2*(1-pt(3.947905, df=n-1))
[1] 0.0002518318
```

$t_{\alpha/2,\ n-1}$ = 2.009575 e $t_{os}$ = 3.947905 dunque cade al di fuori della regione di accettazione. Si rifiuta l'ipotesi nulla con livello di significatività del 5%.
Essendo il 0.0002518318 < α anche per il criterio del p-value si rifiuta l'ipotesi nulla.

### 5.1.2 Test unilaterale sinistro

Si prendano in considerazione le seguenti ipotesi:

$$\mathbf{H}_0 : \mu \leq \mu_0 \qquad \mathbf{H}_1 : \mu > \mu_0$$

Entrambe sono composite, essendo la varianza non nota.
Il test unilaterale sinistro ψ di misura α con le ipotesi sopra citate è il seguente:

- si accetta $H_0$ se:

$$\frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} < t_{\alpha,n-1}$$

- si rifiuta $H_0$ se:

$$\frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} > t_{\alpha,n-1}$$

Come per il test bilaterale, anche per il test unilaterale sinistro si può calcolare il p-value, partendo dalla stima della statistica del test, denotata come segue:

$$t_{os} = (\overline{x}_n - \mu_0)/(s_n/\sqrt{n})$$

da cui si ricava il p-value:

$$pvalue = P(T_n > t_{os}) = 1 - P(T_n \leq t_{os})$$

Di seguito viene riportato il test unilaterale sinistro sul campione considerato, per verificare l'ipotesi nulla $H_0$ : μ ≤ 151 in alternativa all'ipotesi $H_1$ : μ > 151

```
alpha <- 0.05
mu0 <- 151
qt(1-alpha, df=n-1)
[1] 1.676551

(m1-mu0)/(s1/sqrt(n))
[1] -2.196877

#calcolo del p-value
1-pt(-14.48644, df=n-1)
[1] 1
```

$t_{\alpha, n-1}$ = 1.676551 e $t_{os}$ = -2.196877 dunque cade nella regione di accettazione. Si accetta l'ipotesi nulla con livello di significatività del 5%.
Essendo il 1 > α anche per il criterio del p-value si accetta l'ipotesi nulla.

### 5.1.3 Test unilaterale destro

Si prendano in considerazione le seguenti ipotesi:

$$\mathbf{H}_0 : \mu \geq \mu_0 \qquad \mathbf{H}_1 : \mu < \mu_0$$

Entrambe sono composite, essendo la varianza non nota.
Il test unilaterale destro ψ di misura α con le ipotesi sopra citate è il seguente:

- si accetta $H_0$ se:

$$\frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} > -t_{\alpha,n-1}$$

- si rifiuta $H_0$ se:

$$\frac{\overline{x}_n - \mu_0}{s_n/\sqrt{n}} < -t_{\alpha,n-1}$$

Anche per il test unilaterale destro si può calcolare il p-value, partendo dalla stima della statistica del test, denotata come segue:

$$t_{os} = (\overline{x}_n - \mu_0)/(s_n/\sqrt{n})$$

da cui si ricava il p-value:

$$pvalue = P(T_n \leq t_{os})$$

Di seguito viene riportato il test unilaterale destro sul campione considerato, per verificare l'ipotesi nulla $H_0$ : μ > 151 in alternativa all'ipotesi $H_1$ : μ ≤ 151

```
alpha <- 0.05
mu0 <- 151
qt(alpha, df=n-1)
[1] -1.676551
```

```
(m1-mu0)/(s1/sqrt(n))
[1] -2.196877

#calcolo del p-value
pt(-2.196877, df=n-1)
[1] 0.01639449
```

$-t_{\alpha, n-1}$ = -1.676551 e $t_{os}$ = -2.196877 dunque cade al di fuori della regione di accettazione. Si rifiuta l'ipotesi nulla con livello di significatività del 5%.
Essendo il 0.01639449 < α anche per il criterio del p-value si rifiuta l'ipotesi nulla.

## 5.2 Test su $\sigma^2$ con μ non noto

### 5.2.1 Test bilaterale

Sia $X_1, X_2,...,X_n$ un campione casuale estratto da una popolazione normale con valore medio noto μ
Si prendano in considerazione le seguenti ipotesi:

$$\mathbf{H}_0 : \sigma^2 = \sigma_0^2, \qquad \mathbf{H}_1 : \sigma^2 \neq \sigma_0^2$$

Entrambe le ipotesi sono composite.
In analogo a quanto visto per gli intervalli di confidenza, si considera la seguente variabile aleatoria, distribuita con legge di Student con n-1 gradi di libertà:

$$Q_n = \frac{(n-1)\,S_n^2}{\sigma_0^2} = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Il test bilaterale ψ di misura α con le ipotesi sopra citate è il seguente:

- Si accetta $H_0$ se:

$$\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)\,s_n^2}{\sigma_0^2} < \chi^2_{\alpha/2,n-1}$$

- Si rifiuta $H_0$ se:

$$\frac{(n-1)\,s_n^2}{\sigma_0^2} < \chi^2_{1-\alpha/2,n-1}$$

oppure

$$\frac{(n-1)\,s_n^2}{\sigma_0^2} > \chi^2_{\alpha/2,n-1}$$

Di seguito si esegue il test bilaterale sul campione

```
alpha<-0.05
sigma2<-2
varcamp<-var(campione)

qchisq(alpha/2,df=n-1)
[1] 31.55492

qchisq(1-alpha/2,df=n-1)
[1] 70.22241

(n-1)*varcamp/sigma2
[1] 129.7726
```

Si nota che $\chi^2_{1-\alpha/2,49} = 31.55$ e $\chi^2_{\alpha/2,49} = 70.22$ e $\chi^2 = 129.77$.
Poiché il valore osservato non è compreso nella regione di accettazione, si rifiuta l'ipotesi

### 5.2.2 Test unilaterale sinistro

Si prendano in considerazione le seguenti ipotesi:

$$\mathbf{H}_0 : \sigma^2 \leq \sigma_0^2, \qquad \mathbf{H}_1 : \sigma^2 > \sigma_0^2.$$

Il test bilaterale $\psi$ di misura $\alpha$ on le ipotesi sopra citate è il seguente:

- Si accetta $H_0$ se:

$$\frac{(n-1)s_n^2}{\sigma_0^2} < \chi^2_{\alpha,n-1}$$

- Si rifiuta $H_0$ se:

$$\frac{(n-1)s_n^2}{\sigma_0^2} > \chi^2_{\alpha,n-1}$$

### 5.2.3 Test unilaterale destro

Si prendano in considerazione le seguenti ipotesi:

$$\mathbf{H}_0 : \sigma^2 \geq \sigma_0^2 \qquad \mathbf{H}_1 : \sigma^2 < \sigma_0^2.$$

Il test bilaterale $\psi$ di misura $\alpha$ on le ipotesi sopra citate è il seguente:

- Si accetta $H_0$ se:

$$\frac{(n-1)s_n^2}{\sigma_0^2} > \chi^2_{1-\alpha,n-1}$$

- Si rifiuta $H_0$ se:

$$\frac{(n-1)s_n^2}{\sigma_0^2} < \chi^2_{1-\alpha,n-1}$$

# 6 - CRITERIO DEL CHI-QUADRATO

Il criterio del chi-quadrato verifica che l'ipotesi che una 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.
Come per la verifica delle ipotesi, anche con il test del chi-quadrato viene definita con $H_0$ l'ipotesi nulla e con $H_1$ l'ipotesi alternativa.

- $H_0$: X ha una funzione di distribuzione $F_x(x)$
- $H_1$: X non ha una funzione di distribuzione $F_x(x)$

Il test del chi-quadrato, con livello di significatività α, ha come obiettivo quello di verificare l'ipotesi nulla, dove α è la probabilità massima di rifiutare l'ipotesi nulla quando essa è vera.
Bisogna determinare un test ψ con livello di significatività α che permetta di determinare una regione di accettazione e di rifiuto dell'ipotesi nulla.
Si suddivide l'insieme dei valori che la variabile aleatoria X può assumere in r sottoinsiemi $I_1, I_2, \ldots, I_r$ in modo tale che la probabilità che la variabile aleatoria assuma un valore appartenente ad $I_i$ sia uguale a $p_i$:

$$p_i = P(X \in I_i) \qquad (i = 1, 2, \ldots, r).$$

Successivamente si va ad estrarre un campione $x_1, x_2, \ldots, x_n$ di ampiezza n e si osservano le frequenze assolute $n_1, n_2, \ldots, n_r$ con cui gli n elementi si distribuiscono nei rispettivi insiemi $I_1, I_2, \ldots, I_r$, $n_i$ rappresenta il numero degli elementi del campione che cadono nell'intervallo $I_i$ (i = 1,2, …, r).
Si ottiene quindi:

$$p_i \geq 0 \quad (i = 1, 2, \ldots, r), \qquad \sum_{i=1}^{r} p_i = 1;$$

$$n_i \geq 0 \quad (i = 1, 2, \ldots, r), \qquad \sum_{i=1}^{r} n_i = n.$$

La probabilità che esattamente $n_1$ elementi appartengano ad $I_1$, $n_2$ ad $I_2$, …, $n_r$ ad $I_r$ è proprio uguale ad una funzione di probabilità multinomiale, ossia:

$$p(n_1, n_2, \ldots, n_r) = \frac{n!}{n_1!\, n_2! \ldots n_r!}\, p_1^{n_1} p_2^{n_2} \ldots p_r^{n_r}$$

Segue quindi che il numero medio di elementi che cadono nell'intervallo $I_i$ è $np_i$.

Successivamente si calcola la quantità:

$$\chi^2 = \sum_{i=1}^{r} \left( \frac{n_i - n\,p_i}{\sqrt{n\,p_i}} \right)^2 .$$

Il criterio chi-quadrato si basa sulla statistica

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

dove $N_i$ è la variabile aleatoria che descrive il numero degli elementi del campione casuale $X_1, X_2, ..., X_n$ che cadono nell'intervallo $I_i$.

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

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

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

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

Di seguito viene riportato il test del chi-quadrato con 5 sottoinsiemi:

```
media<-mean(campione)
devSd <- sd(campione)
a <- numeric (4)

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

[1] 148.3480 149.7019 150.8680 152.2219


#Numero di elementi per ogni insieme
r<-5
nint <-numeric (r)

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

[1] 12  8 11  9 10


#Calcolo di χ^2
chi2 <-sum ((( nint -n*0.2)/sqrt(n*0.2))^2)

[1] 1
```

La distribuzione normale ha due parametri non noti ($\mu,\sigma^2$) 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.

```
#Calcolo di χ^2_{α/2,2} e χ^2_{1−α/2,2} con α = 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
```

L'ipotesi $H_0$ può essere accettata.