UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

CORSO DI LAUREA MAGISTRALE IN INFORMATICA

DATA SCIENCE E MACHINE LEARNING

PROGETTO DI STATISTICA E ANALISI DEI DATI

Stime e verifica delle ipotesi su una popolazione esponenziale

DOCENTE

Prof. Amelia Giuseppina Nobile

STUDENTI

Maria Natale, matricola: 0522500967

Gaetano Casillo, matricola: 0522501057

ANNO ACCADEMICO 2020-2021

SOMMARIO

1 Variabile aleatoria esponenziale 3

1.1 Stima del parametro non noto 6

1.1.1 Stima puntuale 7

1.1.2 Stima intervallare 9

1.1.3 Confronto tra due popolazioni esponenziali 14

1.2 Verifica delle ipotesi 16

1.2.1 Ipotesi zero e test di ipotesi 17

1.2.2 Test statici 18

1.2.3 Test statistici su grandi campioni 19

1.2.4 Criterio del chi-quadrato 21

1 VARIABILE ALEATORIA ESPONENZIALE

La **distribuzione esponenziale** è una distribuzione di probabilità continua che descrive la "durata di vita" di un fenomeno che *non invecchia*. Un esempio è la *durata di vita* di una particella radioattiva prima decadere. Si dice che X ha distribuzione esponenziale di parametro $\lambda>0$ e si indica con $X\sim EXP(\lambda)$, se la sua funzione di distribuzione è:

$$F_x(x) = P(X \leq x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \geq 0 \end{cases}$$

e corrispondente densità di probabilità:

$$f_x(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & altrimenti \end{cases}$$

Per una variabile esponenziale si ha che:

$$E(X) = \frac{1}{\lambda} \quad Var(X) = \frac{1}{\lambda^2}$$

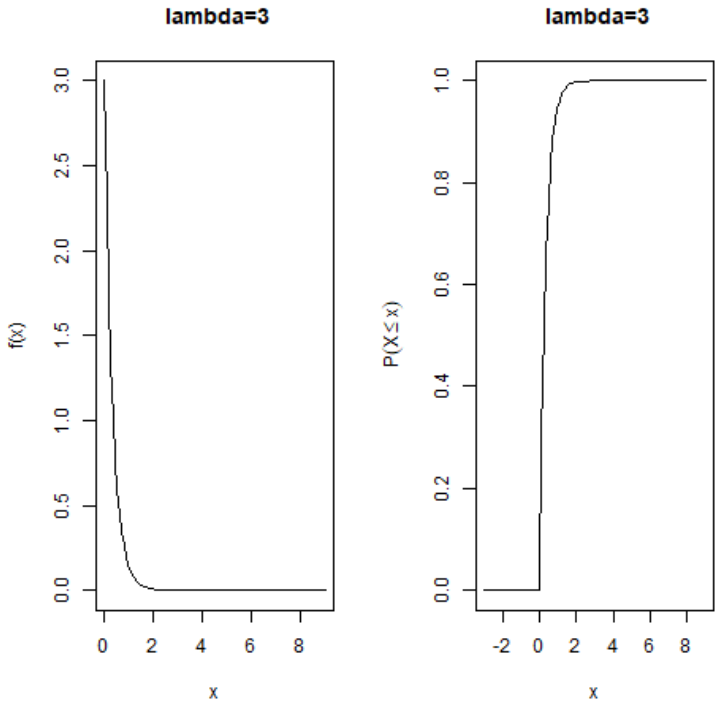
Osservando che $E(X) = \frac{1}{\lambda}$, se X rappresenta un tempo allora λ rappresenta una frequenza. Quindi se la variabile aleatoria descrive, ad esempio, la durata di vita di un componente elettronico si intuisce che i tempi di vita maggiori corrispondono ai parametri λ più piccoli. Infatti, la funzione densità, al diminuire di λ, si schiaccia sull'asse delle ascisse. Di conseguenza la media si sposta verso valori più elevati e il componente si guasta mediamente più tardi. Pertanto, λ risulta essere inversamente proporzionale al tempo di vita medio del componente.

Come specificato precedentemente, tale variabile aleatoria descrive un fenomeno che non invecchia, ciò significa che è privo di memoria. Gode infatti della seguente proprietà di "assenza di memoria", per ogni s, t reali positivi risulta:

$$P(X > s + t \mid X > s) = P(X > t)$$

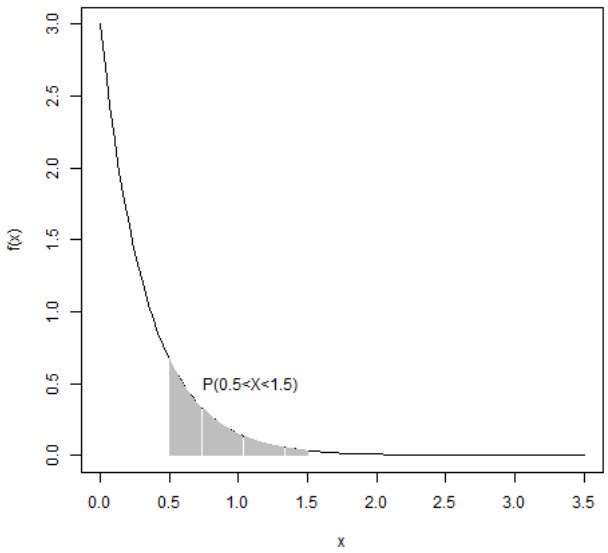
Se si interpreta X come un tempo di attesa, la precedente equazione mostra che la probabilità condizionata che il tempo di attesa X sia maggiore di t+s dato che essa è maggiore di s non dipende da quanto si è già atteso, ossia da s.

Nel seguente grafico è rappresentata la densità di probabilità e la funzione di distribuzione di una variabile aleatoria con distribuzione esponenziale e parametro $\lambda=3$.



La probabilità che la variabile aleatoria esponenziale con λ=3 assuma valori nell'intervallo (0.5, 1.5) corrisponde all'area sottesa dalla densità esponenziale ottenuta tramite il seguente codice:

```
curve ( dexp(x, rate=3) ,from =0, to =2.5 , xlab="x",ylab="f(x)")
x<-seq (0.5 ,1.5 ,0.01)
lines (x, dexp(x, rate=3) ,type="h",col =" grey")
text (1.1 ,0.5 , "P(0.5 <X <1.5)")
```



Questa probabilità può essere così valutata in R:

```
> prob<-pexp (1.5 ,3) -pexp (0.5 ,3)
> prob
[1] 0.2120212
```

La funzione qexp(z, rate) permette di calcolare i quantili:

- Z è il valore assunto (o i valori assunti) dalle probabilità relative al percentile z·100-esimo.
- Rate è il valore del parametro λ.

Il risultato della funzione è il percentile z ·100–esimo, ossia il più piccolo numero x assunto dalla variabile aleatoria esponenziale X tale che

$$F_x(x) = P(X \leq x) \geq z$$

Le seguenti linee di codice permettono di calcolare i quantili di una variabile aleatoria esponenziale con frequenza λ=3.

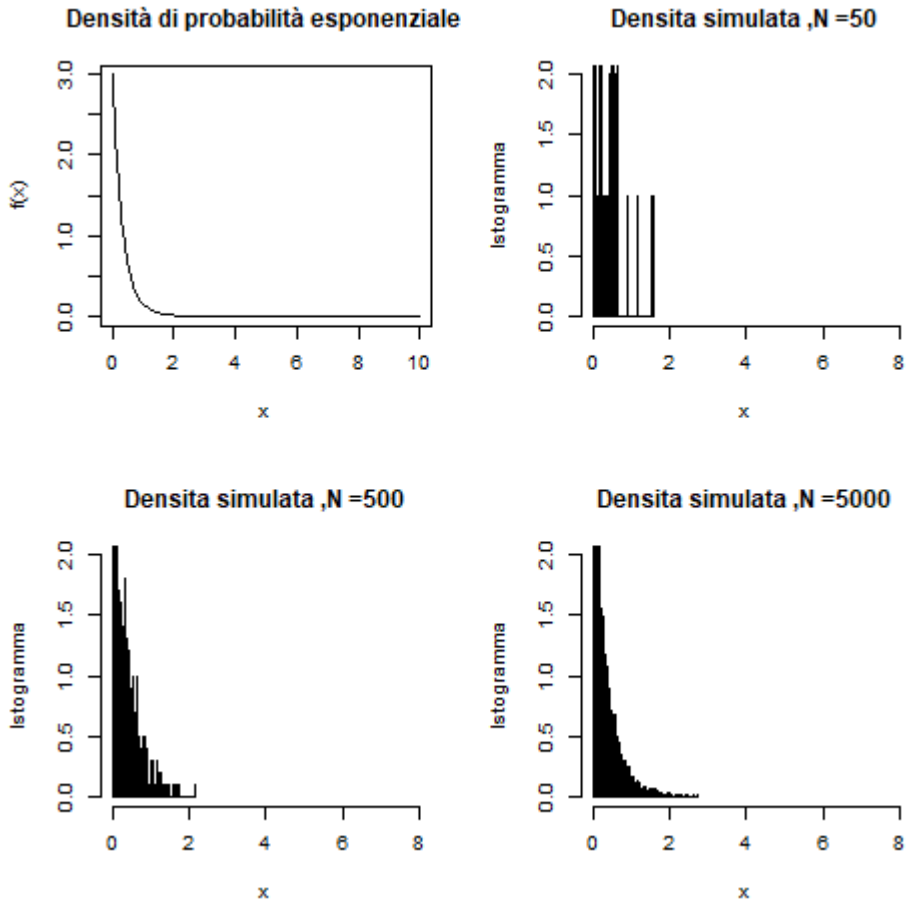
```
> z <-c (0 ,0.25 ,0.5 ,0.75 ,1)
> qexp(z, rate =3)
[1] 0.00000000 0.09589402 0.23104906 0.46209812        Inf
```

In R è possibile generare dei campioni casuali utilizzando la funzione rexp(N, rate=lambda) dove:

- N corrisponde all'ampiezza del campione da generare.
- Rate è il valore del parametro λ.

Il seguente codice permette di confrontare la densità teorica esponenziale di parametro λ=3 con la densità simulata generando tre campioni di ampiezza, rispettivamente, 50, 500 e 5000.

```
par ( mfrow =c(2 ,2))
curve ( dexp(x,rate=3) ,from =0, to=10, xlab="x", ylab="f(x)",main="Densità di
probabilità geometrica")
sim<-rexp(50, rate =3)
hist(sim,freq=F,xlim =c(0 ,8) ,ylim =c(0 ,2) ,breaks =100 , xlab ="x", ylab="
Istogramma ",main=" Densita simulata ,N =50 ")
sim<-rexp(500, rate =3)
hist(sim,freq=F,xlim =c(0 ,8) ,ylim =c(0 ,2) ,breaks =100 , xlab ="x", ylab="
Istogramma ",main=" Densita simulata ,N =500 ")
sim<-rexp(5000, rate =3)
hist(sim,freq=F,xlim =c(0 ,8) ,ylim =c(0 ,2) ,breaks =100 , xlab ="x", ylab="
Istogramma ",main=" Densita simulata ,N =5000 ")
```



Si può notare che all'aumentare dell'ampiezza del campione, l'istogramma delle frequenze relative si avvicina alla densità esponenziale teorica.

1.1 STIMA DEL PARAMETRO NON NOTO

Uno dei principali problemi dell'inferenza statistica consiste nello studiare una popolazione descritta da una variabile aleatoria osservabile X di cui si conosce la forma della funzione di distribuzione ma che contiene il valore di un parametro non noto ϑ. Per ottenere informazioni sui parametri non noti, si considera un campione $X_1, X_2, \dots, X_n$ di ampiezza n estratto dalla popolazione e si fa uso di alcune variabili aleatorie che sono funzioni misurabili del campione, dette statistiche o stimatori.

Una **statistica** $t(X_1, X_2, \dots, X_n)$ è una funzione misurabile e osservabile del campione casuale $X_1, X_2, \dots, X_n$. Essendo la statistica osservabile, i valori da essa assunti dipendono soltanto dal

campione osservato $(x_1, x_2, \dots, x_n)$ estratto dalla popolazione e i parametri non noti sono presenti soltanto nella funzione di distribuzione della statistica.

Uno **stimatore** $\theta = t(X_1, X_2, \dots, X_n)$ è una funzione misurabile e osservabile del campione casuale $X_1, X_2, \dots, X_n$ i cui valori possono essere usati per stimare un parametro non noto ϑ della popolazione. I valori $\hat{\vartheta}$ assunti da tale stimatore sono detti stime del parametro non noto ϑ. Statistiche tipiche sono la media campionaria e la varianza campionaria.

1.1.1 Stima puntuale

I principali metodi per la stima puntuale sono il metodo dei momenti e il metodo della massima verosimiglianza. Se si hanno 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. Si tratta quindi di risolvere un sistema di k equazioni in cui i termini a sinistra dipendono dalla legge di probabilità considerata e contengono i parametri non noti, mentre quelli a destra possono essere calcolati a partire dal campione casuale considerato. In particolare, il momento campionario 1-esimo corrisponde alla media campionaria. Il metodo dei momenti fornisce come stimatore del parametro non noto in una variabile esponenziale la media campionaria. Quindi risulta, $\lambda = \frac{1}{\bar{x}}$.

Sia $X_1, X_2, \dots, X_n$ un campione casuale di ampiezza n estratto dalla popolazione. La funzione di verosimiglianza $L(\vartheta_1, \vartheta_2, \dots, \vartheta_k) = L(\vartheta_1, \vartheta_2, \dots, \vartheta_k; x_1, x_2, \dots, x_n)$ del campione osservato $(x_1, x_2, \dots, x_n)$ 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 $X_1, X_2, \dots, X_n$, ossia:

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

Il metodo della **massima verosimiglianza** consiste nel massimizzare la funzione di verosimiglianza rispetto ai parametri $\vartheta_1, \vartheta_2, \dots, \vartheta_k$. Anche il metodo della massima verosimiglianza, applicato ad una popolazione esponenziale, fornisce come stimatore del parametro λ la media campionaria.

Per una popolazione esponenziale la media campionaria è uno stimatore corretto con varianza minima e consistente per $1/\lambda$.

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

$$E(\hat{\theta}) = \vartheta,$$

ossia se il valore medio dello stimatore $\hat{\theta}$ è uguale al corrispondente parametro non noto della popolazione.

Uno stimatore $\hat{\theta}$ si dice corretto e con **varianza uniformemente minima** per il parametro non noto ϑ se e solo se per ogni $\vartheta \in \theta$ risulta

i. $E(\hat{\theta}) = \vartheta,$

ii. $Var(\hat{\theta}) \leq Var(\hat{\theta}^*)\ per\ ogni\ altro\ stimatore\ corretto\ \hat{\theta}^*\ del\ parametro\ \vartheta$

Uno stimatore $\widehat{\theta_n} = t(X_1, X_2, \dots, X_n)$ del parametro non noto ϑ della popolazione è **consistente** se

i. $\lim_{n\to\infty} E(\hat{\theta}) = \vartheta,$

ii. $\lim_{n\to\infty} Var(\hat{\theta}) = 0,$

Esempio: Si desidera studiare una popolazione descritta da una variabile aleatoria X con funzione di distribuzione esponenziale. In particolare, il campione ha ampiezza 50 e denota i tempi di interarrivo delle chiamate ad un centralino telefonico. Si vuole stimare il parametro non noto λ.

Il campione generato con la funzione rexp risulta:

```
12,55257077     3,557898565     0,066850691     2,412892317
5,634602257     1,380391889     1,946266899     11,42878354
0,480258011     11,3162152      13,66522758     2,618328119
2,881713209     3,709615855     13,45168029     2,040098796
1,843199623     8,041389435     0,675733802     17,35395769
11,3476636      2,758856067     5,484402969     0,625407379
5,265434207     2,992770325     2,284765127     7,652301611
7,803344977     1,518178883     4,278282546     21,26650536
1,026306043     4,415230164     0,762105291     8,245327286
4,940576004     2,101264785     6,98473455      11,59716966
0,070772843     6,792073343     5,675955373     3,50324823
5,427746911     5,56008216      2,798289899     3,364845295
2,041921504     0,87780955
```

La **stima puntuale** del parametro non noto con il metodo dei momenti e della massima verosimiglianza forniscono come stimatore la media campionaria.

```
stimatheta <-1.0 /mean (camp)
```

```
> stimatheta
[1] 0.1876024
```

Risulta quindi λ = 0.1876024.

1.1.2 Stima intervallare

La **stima intervallare** si propone, a differenza della stima puntuale, di determinare in base ai dati del campione un limite superiore e un limite inferiore entro il quale sia compreso il parametro non noto ϑ con un certo coefficiente di confidenza (o grado di fiducia) 1-α.

Un metodo per la costruzione degli intervalli di confidenza è il **metodo pivotale** che consiste nel determinare una variabile aleatoria di pivot $\gamma(X_1 + X_2 + \cdots + X_n; \vartheta)$ che:

- Dipende dal campione casuale $X_1 + X_2 + \cdots + X_n$;
- Dipende dal parametro non noto ϑ;
- La sua funzione di distribuzione non contiene il parametro non noto ϑ.

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

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

Se per ogni possibile campione osservato $x = (x_1 + x_2 + \cdots + x_n)$ e per ogni $\vartheta \in \theta$ si riesce a dimostrare:

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

con $g_1(x)$ e $g_2(x)$ dipendenti soltanto dal campione osservato allora la relazione

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

è equivalente a:

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

Denotando con $\underline{C_n} = g_1(X_1 + X_2 + \cdots + X_n)$ e con $\overline{C_n} = g_2(X_1 + X_2 + \cdots + X_n)$ segue che $(\underline{C_n};\ \overline{C_n})$ è un intervallo di confidenza di grado $1 - \alpha$ per il parametro non noto ϑ della popolazione.

Per effettuare la stima intervallare su un campione con distribuzione esponenziale viene utilizzato il teorema centrale di convergenza.

Teorema centrale di convergenza

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

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

ossia la successione delle variabili aleatorie standardizzate

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

converge in distribuzione alla variabile aleatoria normale standard.

Il teorema mostra inoltre che sottraendo a $X_1 + X_2 + \cdots + X_n$ la sua media $n\mu$ e dividendo la differenza per la deviazione standard di Y_n, si ottiene una variabile aleatoria standardizzata la cui funzione di distribuzione è per n sufficientemente grande approssimativamente normale standard. Quindi, per n grande la distribuzione della somma

$$Y_n = X_1 + X_2 + \cdots + X_n$$

È approssimativamente normale con valore medio $n\mu$ e varianza $n\sigma^2$, ossia

$$Y_n \cong n\mu + \sigma\sqrt{n}Z$$

Inoltre, se denotiamo con

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

la media campionaria, allora

$$\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \to Z$$

converge in distribuzione alla variabile aleatoria normale standard. Quindi per n grande la distribuzione della media campionaria $\overline{X_n}$ è approssimativamente normale con valore medio μ e varianza σ^2/n, ossia

$$X_n \cong \mu + \frac{\sigma}{\sqrt{n}} Z$$

L'approssimazione migliora al crescere di n, nelle applicazioni spesso è già soddisfacente $n \geq 30$.

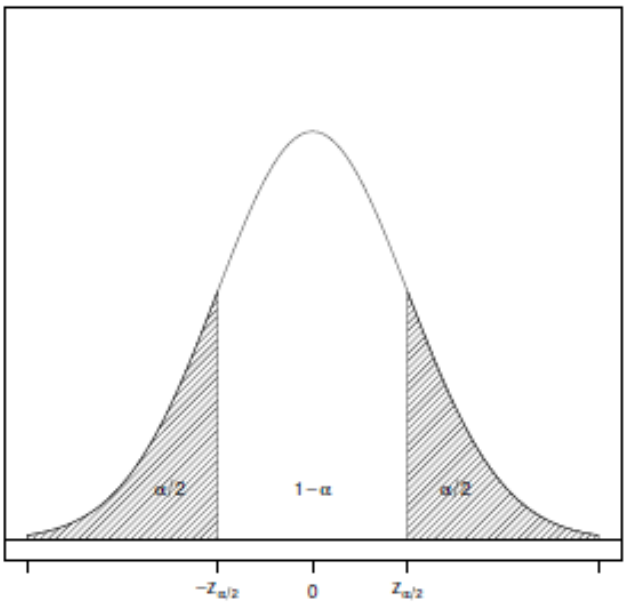
Stima approssimata del parametro non noto di una popolazione esponenziale

Se X denota la variabile aleatoria che descrive la popolazione con $E(X) = \mu$ e $Var(X) = \sigma^2$ e con $(X_1 + X_2 + \cdots + X_n)$ il campione casuale, il teorema centrale di convergenza afferma che la variabile aleatoria

$$Z_n = \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \to Z$$

converge in distribuzione ad una variabile aleatoria normale standard. Tale variabile può essere interpretata come una variabile aleatoria di pivot. Pertanto, per campioni di ampiezza elevata è possibile applicare il metodo pivotale in forma approssimata, cioè:

$$P\left(-z_{\alpha/2} < \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) \cong 1 - \alpha$$



Il valore corrispondente a $-z_{\alpha/2}$ viene ottenuto come il valore numerico più piccolo $-z_{\alpha/2}$ tale che $P(X \leq -z_{\alpha/2}) \geq 1 - \alpha/2$, quindi in R si utilizza `qnorm(1- α/2, mean=0, var=1)`. Discorso analogo viene fatto per calcolare $z_{\alpha/2}$.

Consideriamo una popolazione esponenziale descritta da una variabile aleatoria X∼EXP(λ) con funzione densità di probabilità:

$$f_x(x) = \lambda e^{-\lambda x}, \qquad x > 0 \quad (\lambda > 0)$$

Il valore medio e la varianza sono $E(X) = \frac{1}{\lambda}$, $Var(X) = \frac{1}{\lambda^2}$ e dipendono entrambe dal parametro non noto λ. Si può ricavare che:

$$E(\overline{X_n}) = \frac{1}{\lambda} \ , \qquad Var(\overline{X_n}) = \frac{1}{n\lambda^2}$$

Applicando il teorema centrale di convergenza si ha che

$$\frac{\overline{X_n} - \frac{1}{\lambda}}{1/\left(\frac{\lambda}{\sqrt{n}}\right)} = \sqrt{n}\ \frac{\overline{X_n} - \frac{1}{\lambda}}{\frac{1}{\lambda}} = \sqrt{n}\ (\lambda\overline{X_n} - 1)$$

converge in distribuzione ad una variabile aleatoria normale standard. Per campioni sufficientemente grandi l'intervallo di confidenza di grado 1-α per il parametro $\frac{1}{\lambda}$ corrisponde a:

$$P\left(-\mathrm{z}_{\alpha/2} < \sqrt{n}\ (\lambda\overline{X_n} - 1) < \mathrm{z}_{\alpha/2}\right) \cong 1 - \alpha$$

Ossia:

$$P\left(\overline{X_n}\left(1 + \frac{z_{\alpha/2}}{\sqrt{n}}\right)^{-1} < \frac{1}{\lambda} < \overline{X_n}\left(1 - \frac{z_{\alpha/2}}{\sqrt{n}}\right)^{-1}\right) \cong 1 - \alpha$$

Sussiste quindi la proposizione:

Sia $(x_1, x_2, \dots, x_n)$ un campione osservato di ampiezza n estratto da una popolazione esponenziale di parametro λ. Se la dimensione del campione è elevata, una stima approssimata dell'intervallo di confidenza di grado 1-α per 1/λ è:

$$\overline{x_n}\left(1 + \frac{z_{\alpha/2}}{\sqrt{n}}\right)^{-1} < \frac{1}{\lambda} < \overline{x_n}\left(1 - \frac{z_{\alpha/2}}{\sqrt{n}}\right)^{-1}, \textit{dove } \overline{x_n} \textit{ denota la media campionaria}$$

Esempio: Si supponga che il tempo che intercorre tra l'arrivo di due chiamate successive ad un centralino telefonico sia distribuito esponenzialmente con valore medio non noto 1/λ. Se in 50 osservazioni si riscontra che il tempo medio che intercorre tra due chiamate successive è 5.330421 minuti, determinare una stima dell'intervallo di confidenza di grado 1-α = 0.99 e una stima

dell'intervallo di confidenza di grado 1-α = 0.95 per i tempi medi che intercorrono tra due chiamate successive.

Stima dell'intervallo di confidenza di grado 1 − α = 0.99.

```
alpha <-1 -0.99
n<-50
m<-5.330421
cb<-m/(1+ qnorm (1- alpha/2,mean=0, sd =1) / sqrt(n))
ca<-m/(1-qnorm (1- alpha/2,mean=0, sd =1) / sqrt(n))
```

Il limite inferiore risulta cb=3.907139, mentre il limite superiore risulta ca=8.384482.

```
> cb
[1] 3.907139
> ca
[1] 8.38482
```

Risulta quindi:

$$P(3.907139 < 1/\lambda < 8.384482\ ) = 0.99$$

Stima dell'intervallo di confidenza di grado 1 − α = 0.95.

```
alpha <-1 -0.95
n<-50
m<-5.330421
cb<-m/(1+ qnorm (1- alpha/2,mean=0, sd =1) / sqrt(n))
ca<-m/(1-qnorm (1- alpha/2,mean=0, sd =1) / sqrt(n))
```

Il limite inferiore risulta cb=4.173584, mentre il limite superiore risulta ca=7.374487.

```
> cb
[1] 4.173584
> ca
[1] 7.374487
```

Risulta quindi:

$$P(4.173584 < 1/\lambda < 7.374487\ ) = 0.95$$

Si nota che all'aumentare del grado di confidenza 1 – α l'intervallo diventa più grande.

1.1.3 Confronto tra due popolazioni esponenziali

Consideriamo una prima popolazione esponenziale descritta da una variabile X~EXP(λ_1) con densità di probabilità:

$$f_x(x) = \begin{cases} \lambda_1 e^{-\lambda_1 x}, & x > 0 \\ 0, & altrimenti \end{cases}$$

ed una seconda popolazione esponenziale descritta da una variabile Y~EXP(λ_2) con densità di probabilità:

$$f_y(y) = \begin{cases} \lambda_2 e^{-\lambda_2 y}, & y > 0 \\ 0, & altrimenti \end{cases}$$

E siano $X_1, X_2, \dots, X_n$ e $Y_1, Y_2, \dots, Y_n$ due campioni casuali di ampiezza n_1 e n_2 estratti dalle popolazioni esponenziali. Si vuole determinare un intervallo di confidenza di grado 1-α per la differenza $1/\lambda_1 - 1/\lambda_2$ per grandi valori di n_1 e n_2. Dal teorema centrale di convergenza segue che la variabile aleatoria:

$$\frac{\overline{X_{n1}} - \overline{Y_{n2}} - \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right)}{\sqrt{\frac{1}{n_1\lambda_1^2} + \frac{1}{n_2\lambda_2^2}}} \to Z$$

Converge in distribuzione ad una variabile aleatoria normale standard, quindi le medie campionarie $\overline{X_{n1}}$ e $\overline{Y_{n2}}$ sono stimatori corretti e consistenti di $1/\lambda_1$ e $1/\lambda_2$, per campioni sufficientemente numerosi l'intervallo di confidenza di grado 1 – α per la differenza $1/\lambda_1$ e $1/\lambda_2$ può essere determinato supponendo che:

$$P\left(-z_{\alpha/2} < \frac{\overline{X_{n1}} - \overline{Y_{n2}} - \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right)}{\sqrt{\frac{1}{n_1\lambda_1^2} + \frac{1}{n_2\lambda_2^2}}} < z_{\alpha/2} \right) \cong 1 - \alpha$$

Da cui:

$$\overline{x_{n1}} - \overline{y_{n2}} - z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{x_{n1}}^2}{n_1} + \frac{\overline{y_{n2}}^2}{n_2}} < \frac{1}{\lambda_1} - \frac{1}{\lambda_2} < \overline{x_{n1}} - \overline{y_{n2}} + z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{x_{n1}}^2}{n_1} + \frac{\overline{y_{n2}}^2}{n_2}}$$

Esempio: Si desidera confrontare il tempo che intercorre tra l'arrivo di due chiamate successive a due centralini denotati con A e B. Si supponga che i tempi sono distribuiti come una variabile esponenziale. Il centralino A è descritto da una variabile esponenziale $X \sim EXP(\lambda_A)$ e si osservano 50 chiamate, mentre il centralino B è descritto da una variabile esponenziale $Y \sim EXP(\lambda_B)$ e si osservano 80 chiamate con i seguenti risultati sulle medie e sulle deviazioni standard dei tempi tra due chiamate successive: $media_A=5.330421$, $sd_A=4.737098$, $media_B=10.11495$, $sd_B=10.82139$. Si vuole determinare una stima dell'intervallo di confidenza di grado $1 - \alpha = 0.99$ per la differenza $1/\lambda_A - 1/\lambda_B$ tra i tempi che intercorrono tra l'arrivo di due chiamate successive ai due centralini telefonici.

I valori del campione del centralino A sono quelli elencati precedentemente nel paragrafo 1.1.2. I valori del campione del centralino B sono i seguenti:

12,00437877	30,11981651	36,2732764	5,511409827
3,860684957	33,77324015	40,53770995	6,012060968
1,11152689	0,045442674	3,465351928	0,739132319
36,4652101	53,50107478	4,896664666	3,213116731
7,286642122	4,149451354	11,69270193	3,110677744
15,70766203	23,99411746	1,737911697	6,630810588
10,60050974	3,982419469	1,119041913	6,42280757
14,14385622	3,677816126	2,293871073	7,575114443
2,942939568	9,037694252	27,28819171	0,170129152
5,498995329	7,949043484	9,314060925	1,885408817
4,429491172	2,104222441	20,31249123	28,00607553
5,180053495	3,657424515	7,43016744	14,07186251
1,69316378	1,527658809	3,890972747	4,580235416
4,194758954	9,573003519	3,245679485	13,81225596
1,343266426	1,502164239	6,506694076	1,146984082
13,21266153	13,7555	8,740801522	15,38673783
10,20344139	29,81519982	1,831024233	1,185700023
5,20923879	8,535682289	29,34136988	0,383451576
13,77708594	17,08908137	3,193944894	6,79066407
2,387433276	12,63221709	9,347000354	8,427319665

Il seguente codice permette di calcolare la differenza $1/\lambda_A - 1/\lambda_B$ tra i tempi che intercorrono tra l'arrivo di due chiamate successive ai due centralini telefonici.

```
alpha <-1 -0.99
n2<-80
n<-50
media1<-5.330421
media2<-10.11495
rad<-sqrt(media1^2*(1/n)+media2^2*(1/n2))
```

```
cb<-media1-media2-qnorm (1-alpha/2, mean=0, sd=1)*rad
ca<-media1-media2+qnorm (1-alpha/2, mean=0, sd=1)*rad
```

```
> cb
[1] -8.285358
> ca
[1] -1.283704
```

Risulta quindi:

$$P\left(-8.285358 < \frac{1}{\lambda_A} - \frac{1}{\lambda_B} < -1.283704\right) = 0.99$$

Siccome ca e cb sono entrambi negativi, la differenza $1/\lambda_A - 1/\lambda_B$ risulta essere negativa, pertanto $\lambda_A > \lambda_B$. Siccome in una variabile aleatoria esponenziale λ può essere visto come una frequenza, al centralino A arrivano più chiamate rispetto al centralino B in un determinato intervallo di tempo.

1.2 VERIFICA DELLE IPOTESI

La stima dei parametri e la verifica delle ipotesi sono i campi più importanti dell'inferenza statistica. In termini comuni, gli effetti di questi due campi li possiamo osservare nei sondaggi politici che ci bombardano sui social o su quanto un prodotto sia migliore degli altri nel campo della pubblicità. Ma come si fa a stabilire se effettivamente un prodotto è migliore degli altri? Come si può creare un sondaggio d'opinione valido?

Si fornisce un'ipotesi e la si verifica.

Di cosa abbiamo bisogno?

- Un'ipotesi da verificare su di un parametro non noto ϑ. Un'ipotesi è un'affermazione che ha come oggetto accadimenti del mondo reale. In termini matematici, un'ipotesi statistica è un'ipotesi o congettura sul parametro ϑ. Se l'ipotesi specifica completamente $f(x;\vartheta)$ è detta ipotesi semplice, altrimenti è chiamata ipotesi composita.
  Dire: Ho un campione X,…..,Xn di una popolazione di Bernoulli con $B(1,p)$ dove p è la probabilità di successo. Se dicessi che la mia ipotesi è H: p=0,5 avrei un'ipotesi semplice in quanto l'ipotesi specifica completamente la funzione di probabilità. Se invece dicessi H:

p≠0.5, questa sarebbe composita poiché non specifica completamente la funzione di probabilità.

- Una popolazione descritta da una variabile aleatoria X caratterizzata da una funzione di probabilità o densità di probabilità $f(x;\vartheta)$
- Campione casuale estratto dalla popolazione

1.2.1 Ipotesi zero e test di ipotesi

Quando ipotizziamo, l'ipotesi soggetta a verifica viene chiamata ipotesi nulla, anche se può essere denotata anche con "ipotesi zero" in quanto, in statistica, viene indicata con H_0. Il test d'ipotesi è la regola con cui si decide se preso un campione X….Xn, questo campione appartiene o meno ad H_0.

Se non appartiene ad H_0, appartiene ad H_1. Il test d'ipotesi quindi prevede anche la creazione di un'ipotesi alternativa all'ipotesi zero che viene chiamata, appunto, "ipotesi alternativa".

Le ipotesi H_0 e H_1 sono vere quando

- H_0: $\vartheta \in \theta_0$ (è corretto ipotizzare che il parametro non noto appartiene allo spazio θ_0);
- H_1: $\vartheta \in \theta_1$(è corretto ipotizzare che il parametro non noto appartiene allo spazio θ_1).

I parametri θ_0 e θ_1 sono spazi disgiunti dello spazio θ, lo spazio dei parametri.

Per realizzare il test d'ipotesi occorre suddividere, mediante opportuni criteri, l'insieme di tutti i possibili campioni di popolazione in due regioni:

- Regione A di accettazione dell'ipotesi zero;
- Regione R di rifiuto dell'ipotesi zero.

Il test ω viene quindi così formulato:

- Se il campione osservato appartiene alla regione A, l'ipotesi zero è verificata;
- Se il campione osservato appartiene alla regione R, l'ipotesi zero è rifiutata.

Nel caso in cui l'ipotesi zero viene verificata, l'ipotesi alternativa non viene accettata e viceversa.

Di norma, H_0 va verificata in alternativa ad H_1.

Ma questo può portare a degli errori quali:

- Errore di tipo 1: Si rifiuta l'ipotesi nulla quando essa è vera, come un allarme antiincendio che suona quando non c'è nessun fuoco. La probabilità di commettere questo errore viene espressa come: $\alpha(\vartheta) = P(rifiutare\ , H_0|\vartheta), \vartheta \in \theta_0$

- Errore di tipo 2: Si accetta l'ipotesi nulla quando essa è falsa, come un allarme antiincendio che non suona quando c'è un fuoco. La probabilità di commettere questo errore viene espressa come: $\beta(\vartheta) = P(accettare\ , H_0|\vartheta), \vartheta \in\ \theta_1$

La misurazione della possibilità di commettere uno dei due errori viene espresso dal livello di significatività del test d'ipotesi. Sia ω un test che verifica l'ipotesi nulla H_0: $\vartheta \in\ \theta_0$ in alternativa ad H_1: $\vartheta \in\ \theta_1$. Il livello di significatività del test d'ipotesi è espresso dalla seguente probabilità

$$\alpha = sup_{\vartheta \in \theta_0} \alpha(\vartheta)$$

In questo caso, la misurazione fornisce la possibilità massima di commettere un errore di tipo 1, quindi la probabilità massima di rifiutare l'ipotesi nulla quando essa è vera. Quindi, la possibilità di accettare l'ipotesi zero quando essa è vera è 1- α.

Di norma, quando si costruisce un'ipotesi, si dovrebbe costruire in modo tale che sia più grave commettere un errore di tipo 1 che di tipo 2; per campioni casuali di fissata lunghezza se diminuisce la possibilità di commettere un errore di tipo 1, aumenta quella di commettere un errore di tipo 2 ed è per questo motivo che conviene fissare la probabilità di commettere un errore di tipo 1 e poi formulare un test d'ipotesi che minimizzi la possibilità di commettere un errore di tipo 2.

La probabilità di commettere un errore di tipo 1 viene scelta normalmente tra le seguenti possibilità:

- Se la possibilità di commettere un errore di tipo 1 è 0.05, il test viene detto statisticamente significativo;
- Se la possibilità di commettere un errore di tipo 1 è 0.01, il test viene detto statisticamente molto significativo;
- Se la possibilità di commettere un errore di tipo 1 è 0.001, il test viene detto statisticamente estremamente significativo.

Tanto minore è il valore di α, tanto è più affidabile il rifiuto dell'ipotesi nulla.

1.2.2 Test statici

Esistono due tipi di test statistici:

- Test bilaterali: la regione di rifiuto è costituita da due intervalli. Esempio:

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

- Test unilaterale: regione di rifiuto costituita da un intervallo. Es. (test unilaterale sinistro)

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

Oppure per il test unilaterale destro

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

1.2.3 Test statistici su grandi campioni

Nel caso in cui il campione in esame sia molto ampio per una popolazione descritta da una variabile aleatoria X con valore medio μ *e varianza* σ^2 finiti, si può utilizzare il teorema centrale di convergenza ricordando che:

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

Converge in una variabile normale standard.

Test bilaterale approssimato

Il test bilaterale ω di misura α per le ipotesi $H_0: \mu = \mu_0$ e $H_1: \mu \neq \mu_0$ considera la variabile aleatoria $\frac{\overline{X}_n - \mu_0}{\sigma_0/\sqrt{n}}$, dove σ_0 è la deviazione standard della popolazione quando $\mu = \mu_0$.

Si accetta H_0 se $-z_{\frac{\alpha}{2}} < \frac{\overline{x_n} - \mu_0}{\sigma/\sqrt{n}} < z_{\frac{\alpha}{2}}$

Si rifiuta se $-z_{\frac{\alpha}{2}} > \frac{\overline{x_n} - \mu_0}{\sigma/\sqrt{n}}$ o $\frac{\overline{x_n} - \mu_0}{\sigma/\sqrt{n}} > z_{\frac{\alpha}{2}}$

Esempio: Si supponga che il tempo che intercorre tra l'arrivo di due chiamate successive al centralino A sia distribuito esponenzialmente con valore medio non noto 1/λ. Se in 50 osservazioni si riscontra che il tempo medio che intercorre è di 5.330421 minuti, è stato mostrato che una stima

dell'intervallo di confidenza di grado 1-α = 0.99 per il parametro 1/λ è $(3.907139, 8.384482)$. Si si propone di verificare l'ipotesi $H_0: \frac{1}{\lambda} = 5$ in alternativa a $H_1: \frac{1}{\lambda} \neq 5$.

```
lamba0=1/5
alfa=0.01
qnorm(1-alfa/2,mean=0,sd=1)
n=50
meancap=5.330421
sqrt(n)*(lamba0*meancap-1)
```

$$z_{\frac{\alpha}{2}} = 2.575829$$

$$z_{os} = 0.4672859$$

Poiché z_{os} è compreso fra $z_{\frac{\alpha}{2}}$ e $-z_{\frac{\alpha}{2}}$, accettiamo l'ipotesi H_0 con un livello di significatività del 1%.

Test unilaterale sinistro approssimato

Il test unilaterale sinistro ω di misura α per le ipotesi $H_0: \mu \leq \mu_0$ e $H_1: \mu > \mu_0$

Si accetti H_0 se $z_\alpha > \frac{\overline{x_n} - \mu_0}{\sigma/\sqrt{n}}$

Si rifiuti H_0 se $z_\alpha < \frac{\overline{x_n} - \mu_0}{\sigma/\sqrt{n}}$

Esempio: Si supponga che il tempo che intercorre tra l'arrivo di due chiamate successive al centralino A sia distribuito esponenzialmente con valore medio non noto 1/λ. Se in 50 osservazioni si riscontra che il tempo medio che intercorre è di 5.330421 minuti, è stato mostrato che una stima dell'intervallo di confidenza di grado 1-α = 0.99 per il parametro 1/λ è $(3.907139, 8.384482)$. Si si propone di verificare l'ipotesi $H_0: \frac{1}{\lambda} \leq 3.5$ in alternativa a $H_1: \frac{1}{\lambda} > 3.5$.

```
lamba0=1/3.5
alfa=0.01
qnorm(1-alfa,mean=0,sd=1)
n=50
meancap=5.330421
sqrt(n)*(lamba0*meancap-1)
```

$$z_\alpha = 2.326348$$

$$z_{os} = 3.698009$$

In questo caso l'ipotesi H_0 viene rifiutata in quanto z_{os} non cade nell'intervallo di accettazione.

Test unilaterale destro approssimato

Il test unilaterale destro ω di misura α per le ipotesi $H_0: \mu \geq \mu_0$ e $H_1: \mu < \mu_0$

Si accetti H_0 se $-z_\alpha < \frac{\overline{x_n} - \mu_0}{\sigma/\sqrt{n}}$

Si rifiuti H_0 se $-z_\alpha > \frac{\overline{x_n} - \mu_0}{\sigma/\sqrt{n}}$

Esempio: Si supponga che il tempo che intercorre tra l'arrivo di due chiamate successive al centralino A sia distribuito esponenzialmente con valore medio non noto 1/λ. Se in 50 osservazioni si riscontra che il tempo medio che intercorre è di 5.330421 minuti, è stato mostrato che una stima dell'intervallo di confidenza di grado 1-α = 0.99 per il parametro 1/λ è $(3.907139, 8.384482)$. Si si propone di verificare l'ipotesi $H_0: \frac{1}{\lambda} \geq 3.5$ in alternativa a $H_1: \frac{1}{\lambda} < 3.5$.

```
lamba0=1/3.5
alfa=0.01
qnorm(alfa,mean=0,sd=1)
n=50
meancap=5.330421
sqrt(n)*(lamba0*meancap-1)
```

$$z_\alpha = -2.326348$$

$$z_{os} = 3.698009$$

In questo caso l'ipotesi H_0 viene accettata in quanto z_{os} è più grande di $-z_\alpha$ e quindi cade nella regione di accettazione.

1.2.4 Criterio del chi-quadrato

Con il criterio del chi-quadrato si verifica 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.

Denotando con H_0 l'ipotesi nulla e con H_1 l'ipotesi alternativa, il test chi–quadrato di misura α mira a verificare l'ipotesi nulla:

H_0: *X ha una funzione di distribuzione* $F_X(x)$

mentre

H_1 *: X non ha una funzione di distribuzione* $F_X(x)$

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

Occorre determinare un test ψ di misura α che permetta di determinare una regione di accettazione e di rifiuto dell'ipotesi nulla. Il test di verifica delle ipotesi considerato è bilaterale.

Bisogna suddividere l'insieme dei valori che la variabile aleatoria X possa assumere in r sottoinsiemi: $I_1, I_2, \dots, I_r$in modo che risulti essere uguale a p_i la probabilità che la variabile aleatoria assuma un valore appartenente a I_i , ossia:

$$p_i = P(X \in I_i)\ (i = 1,2, \dots, r)$$

Si estrae poi un campione di ampiezza n e si osservano le frequenze assolute con cui i rispettivi n elementi si distribuiscono nei rispettivi insiemi.

Quindi n_i rappresenta il numero degli elementi del campione che cadono nell'intervallo I_i (i = 1, 2, . . . , r). Quindi:

$$p_i \geq 0\ (i = 1,2 \dots, r) \sum_{i=1}^{r} p_i = 1$$

$$n_i \geq 0\ (i = 1,2 \dots, r) \sum_{i=1}^{r} n_i = n$$

Si nota che la probabilità che esattamente n_r elementi appartengano ad I_r è:

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

Che è una funzione di probabilità multinomiale e quindi il numero medio di elementi che si trovano nell'intervallo I_iè np_i.

Si calcola poi la quantità

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

Il criterio chi–quadrato si basa sulla statistica:

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

Con N_i che è la variabile aleatoria che descrivere il numero degli elementi del campione casuale.

Se la variabile aleatoria X ha una funzione di distribuzione $F_X(x)$ con k parametri non noti, si può dimostrare che per 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. Si sottrae 1 da r a causa della prima delle condizioni secondo la quale se conosciamo r − 1 delle probabilità p_i, la rimanente probabilità può essere univocamente determinata, e si sottrae k poiché si suppone che siano k i parametri indipendenti non noti sostituiti da stime. Per garantire che ogni classe contenga in media almeno 5 elementi, si ritiene valida l'approssimazione se risulta:

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

La definizione del chi quadrato è così data:

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

- si accetti l'ipotesi H_0 se ${x^2}_{1-\frac{\alpha}{2},r-k-1} < x^2 < {x^2}_{\frac{\alpha}{2},r-k-1}$
- si rifiuti l'ipotesi H_0 se ${x^2}_{1-\frac{\alpha}{2},r-k-1} > x^2$ o $x^2 > {x^2}_{\frac{\alpha}{2},r-k-1}$

dove ${x^2}_{1-\frac{\alpha}{2},r-k-1}$ e $x^2 > {x^2}_{\frac{\alpha}{2},r-k-1}$ sono soluzioni alle equazioni:

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

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

Esempio: In 50 osservazioni si riscontra che il tempo che intercorre tra l'arrivo di due chiamate successive ad un centralino telefonico è di 5.330421 minuti. Si desidera verificare utilizzando il test del chi-quadrato se il tempo che intercorre tra le due chiamate successive sia esprimibile con una variabile aleatoria X esponenziale di parametro λ, ossia:

$$f_x(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & altrimenti \end{cases}$$

```
test = read_xlsx("campioneEsponenziale (1).xlsx",sheet = "sheet1")
test=as.matrix(test[,-1])
media=mean(test)
media
a=numeric(4)
for (i in 1:4) {
  a[i]=qexp(0.2*i, rate=1/media)
}
a
r=5
nint=numeric(r)
nint[1]=length(which(test<a[1]))
nint[2]=length(which((test>=a[1])&(test<a[2])))
nint[3]=length(which((test>=a[2])&(test<a[3])))
nint[4]=length(which((test>=a[3])&(test<a[4])))
nint[5]=length(which(test>=a[4]))
nint
sum(nint)
chiquadro=sum(((nint-50*0.2)/sqrt(50*0.2))^2)
chiquadro
#distribuzione esponenziale 1 non noto
k=1
#grado di libertà=3
alfa=0.05
qchisq(alfa/2,df=r-k-1)
#0.2157953
qchisq(1-alfa/2,df=r-k-1)
#9.348404
```

Poiché $x^2 = 1.4$ che è compreso tra ${x^2}_{1-\frac{\alpha}{2},r-k-1}$ (0.215793) e ${x^2}_{\frac{\alpha}{2},r-k-1}$(9.348404) il tempo medio di gestione che intercorre tra due chiamate successive è esprimibile come una popolazione di variabile esponenziale.