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:????????? |

ANNO ACCADEMICO 2020-2021

Sommario

[1 Variabile aleatoria esponenziale 3](#_Toc58935105)

[1.1 Stima del parametro non noto 6](#_Toc58935106)

[1.1.1 Stima puntuale 7](#_Toc58935107)

[1.1.2 Stima intervallare 9](#_Toc58935108)

[1.1.3 Confronto tra due popolazioni esponenziali 13](#_Toc58935109)

[1.2 Verifica delle ipotesi 16](#_Toc58935110)

[1.2.1 Test statistici su grandi campioni 16](#_Toc58935111)

[1.2.2 Criterio del chi-quadrato 17](#_Toc58935112)

[2 Bibliografia 18](#_Toc58935113)

# 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 λ>0 e si indica con X∼EXP(λ), se la sua funzione di distribuzione è:

e corrispondente densità di probabilità:

Per una variabile esponenziale si ha che:

Osservando che , 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:

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 λ=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:



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

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



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.

## 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 . Per ottenere informazioni sui parametri non noti, si considera un campione 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** è una funzione misurabile e osservabile del campione casuale . Essendo la statistica osservabile, i valori da essa assunti dipendono soltanto dal campione osservato estratto dalla popolazione e i parametri non noti sono presenti soltanto nella funzione di distribuzione della statistica.

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 assunti da tale stimatore sono detti stime del parametro non noto . Statistiche tipiche sono la media campionaria e la varianza campionaria.

### 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, .

Sia un campione casuale di ampiezza n estratto dalla popolazione. La funzione di verosimiglianza L(𝜗1, 𝜗2, ..., 𝜗k) = L(𝜗1,𝜗2, ..., 𝜗 k; ) del campione osservato () 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 , ossia:

Il metodo della **massima verosimiglianza** consiste nel massimizzare la funzione di verosimiglianza rispetto ai parametri 𝜗1,𝜗2, ...,𝜗k. Anche il metodo della massima verosimiglianza, applicato ad una popolazione esponenziale, fornisce come stimatore del parametro la media campionaria.

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

Uno stimatore del parametro non noto della popolazione è detto **corretto** se e solo se per ogni si ha

ossia se il valore medio dello stimatore è uguale al corrispondente parametro non noto della popolazione.

Uno stimatore si dice corretto e con **varianza uniformemente minima** per il parametro non noto se e solo se per ogni risulta

Uno stimatore del parametro non noto della popolazione è **consistente** se

**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 gestione in minuti di una richiesta da parte di un servizio A. 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)



Risulta quindi λ = 0.1876024.

### 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 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 che:

* Dipende dal campione casuale ;
* Dipende dal parametro non noto ;
* La sua funzione di distribuzione non contiene il parametro non noto .

Per ogni fissato coefficiente α (0 < α < 1) siano α1 e α2 (α1 < α2) due valori dipendenti soltanto dal coefficiente fissato α e tali che per ogni si abbia:

Se per ogni possibile campione osservato e per ogni si riesce a dimostrare:

con e dipendenti soltanto dal campione osservato allora la relazione

è equivalente a:

Denotando con e con segue che è un intervallo di confidenza di grado per il parametro non noto 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 una successione di variabili aleatorie, definite nello stesso spazio di probabilità, indipendenti ed identicamente distribuite con valore medio µ e varianza σ2 finita e positiva. Posto per ogni intero n positivo , per ogni risulta:

ossia la successione delle variabili aleatorie standardizzate

converge in distribuzione alla variabile aleatoria normale standard.

Il teorema mostra inoltre che sottraendo a la sua media e dividendo la differenza per la deviazione standard di , 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

È approssimativamente normale con valore medio e varianza σ2, ossia

Inoltre, se denotiamo con

la media campionaria, allora

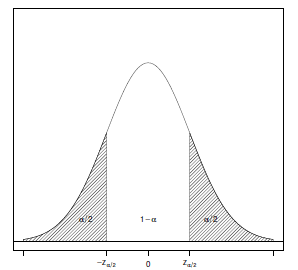
converge in distribuzione alla variabile aleatoria normale standard. Quindi per n grande la distribuzione della media campionaria è approssimativamente normale con valore medio e varianza , ossia

L’approssimazione migliora al crescere di n, nelle applicazioni spesso è già soddisfacente .

**Stima approssimata del parametro non noto di una popolazione esponenziale**

Se X denota la variabile aleatoria che descrive la popolazione con e e con il campione casuale, il teorema centrale di convergenza afferma che la variabile aleatoria

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:



Il valore corrispondente a viene ottenuto come il valore numerico più piccolo tale che , quindi in R si utilizza qnorm(1-, mean=0, var=1). Discorso analogo viene fatto per calcolare .

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

Il valore medio e la varianza sono , e dipendono entrambe dal parametro non noto . Si può ricavare che:

Applicando il teorema centrale di convergenza si ha che

converge in distribuzione ad una variabile aleatoria normale standard. Per campioni sufficientemente grandi l’intervallo di confidenza di grado 1-α per il parametro corrisponde a:

Ossia:

Sussiste quindi la proposizione:

*Sia ) 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:*

**Esempio:** Si supponga che il tempo di gestione di una richiesta ad un servizio A sia distribuito esponenzialmente con valore medio non noto 1/λ. Se in 50 osservazioni si riscontra che i tempi medi di gestione della richiesta in minuti sono di 5.330421, 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 di gestione di una richiesta.

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

alpha <-1 -0.99

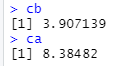
n<-50

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.



Risulta quindi:

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

alpha <-1 -0.95

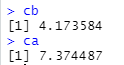
n<-50

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.



Risulta quindi:

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

### Confronto tra due popolazioni esponenziali

Consideriamo una prima popolazione esponenziale descritta da una variabile X∼EXP(λ1) con densità di probabilità:

ed una seconda popolazione esponenziale descritta da una variabile Y∼EXP(λ2) con densità di probabilità:

E siano e due campioni casuali di ampiezza n1 e n2 estratti dalle popolazioni esponenziali. Si vuole determinare un intervallo di confidenza di grado 1-α per la differenza 1/ λ 1− 1/ λ 2 per grandi valori di n1 e n2. Dal teorema centrale di convergenza segue che la variabile aleatoria:

Converge in distribuzione ad una variabile aleatoria normale standard, quindi le medie campionarie e sono stimatori corretti e consistenti di 1/ λ 1 e 1/ λ 2, per campioni sufficientemente numerosi l’intervallo di confidenza di grado 1 – α per la differenza 1/ λ 1 e 1/ λ 2 può essere determinato supponendo che:

Da cui:

**Esempio:** Si desidera confrontare i tempi per gestire una richiesta da parte di due servizi A e B. Si supponga che i tempi sono distribuiti come una variabile esponenziale. Il servizio A è descritto da una variabile esponenziale gestisce X∼EXP(λA) 50 richieste, mentre il servizio B è descritto da una variabile esponenziale gestisce Y∼EXP(λB) e gestisce 80 richieste con i seguenti risultati sulle medie e sulle deviazioni standard dei tempi impiegati per ciascuna richiesta: mediaA=5.330421, sdA=4.737098, mediaB=10.11495, sdB=10.82139. Si vuole determinare una stima dell’intervallo di confidenza di grado 1 − α = 0.99 per la differenza 1/ λ A− 1/ λ B tra i tempi medi impiegati per gestire una richiesta da parte dei due servizi.

I valori del campione del servizio A sono quelli elencati precedentemente nel paragrafo 5.1.2. I valori del campione del servizio 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/ λ A− 1/ λ B tra i tempi medi impiegati per gestire una richiesta da parte dei due servizi.

alpha <-1 -0.99

n2<-80

n<-50

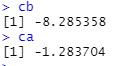
media1<-5.330421

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



Risulta quindi:

Siccome ca e cb sono entrambi negativi, la differenza 1/ λ A − 1/ λ B risulta essere negativa, pertanto λ A > λ B . Siccome in una variabile aleatoria esponenziale λpuò essere visto come una frequenza, il servizio A è in grado di servire più richieste per minuto rispetto al servizio B.

## Verifica delle ipotesi

introduzione su ipotesi nulle ecc.

### Test statistici su grandi campioni

Teoria

Esempi sui 3 tipi di test

**Esempio**: Si supponga che il tempo di gestione di una richiesta ad un servizio A sia distribuito esponenzialmente con valore medio non noto 1/λ. Se in 50 osservazioni si riscontra che i tempi medi di gestione della richiesta in minuti sono di 5.330421, è stato mostrato che una stima dell’intervallo di confidenza di grado 1-α = 0.99 per il parametro 1/λ è . Si si propone di verificare l’ipotesi in alternativa a .

**Esempio**: Si supponga che il tempo di gestione di una richiesta ad un servizio A sia distribuito esponenzialmente con valore medio non noto 1/λ. Se in 50 osservazioni si riscontra che i tempi medi di gestione della richiesta in minuti sono di 5.330421, è stato mostrato che una stima dell’intervallo di confidenza di grado 1-α = 0.99 per il parametro 1/λ è . Si si propone di verificare l’ipotesi in alternativa a .

**Esempio**: Si supponga che il tempo di gestione di una richiesta ad un servizio A sia distribuito esponenzialmente con valore medio non noto 1/λ. Se in 50 osservazioni si riscontra che i tempi medi di gestione della richiesta in minuti sono di 5.330421, è stato mostrato che una stima dell’intervallo di confidenza di grado 1-α = 0.99 per il parametro 1/λ è . Si si propone di verificare l’ipotesi in alternativa a .

### Criterio del chi-quadrato

Teoria

**Esempio**: In 50 osservazioni si riscontra che i tempi medi di gestione della richiesta, da parte di un servizio, espressi in minuti sono di 5.330421. Si desidera verificare utilizzando il test del chi-quadrato se il tempo medio per gestire una richiesta da parte del servizio sia esprimibile con una variabile aleatoria X esponenziale di parametro λ, ossia: