**Università degli Studi di Salerno**



**Dipartimento di INFORMATICA**

**Progetto di Statistica e Analisi dei Dati**

***Statistica Inferenziale applicata al monitoraggio di automobilisti sottoposti ad alcol test***

**Docente: Studente:**

***Prof.ssa. Amelia G. Nobile******Ferrara Carmine***

***Matr.05225/00990***

**ANNO ACCADEMICO 2020/2021**

Sommario

[Introduzione 2](#_Toc57286229)

[Problematica in esame e variabili aleatorie 2](#_Toc57286230)

[Passo 1 – La variabile aleatoria di Poisson 3](#_Toc57286231)

[Passo 2 – accenni sulla variabile aleatoria esponenziale 4](#_Toc57286232)

[Applicazione della variabile aleatoria di Poisson al problema in analisi 6](#_Toc57286233)

[Stime puntuali dei parametri non noti 10](#_Toc57286234)

[Passo 1 – Stima del parametro non noto λ con il metodo dei momenti 11](#_Toc57286235)

[Passo 2 – Stima del parametro non noto λ con il metodo della massima verosimiglianza 13](#_Toc57286236)

# Introduzione

Con lo studio di statistica descrittiva condotto sulla problematica di analisi del consumo alcolemico in Italia del 2019, è emersa la presenza di valori anomali oppure fortemente elevati per molte regioni d’Italia, in particolare poi, è proprio l’Istituto Superiore di Statistica a ribadire che in Italia, relativamente all’anno 2019, oltre il 66% della popolazione sopra gli 11 anni ha consumato almeno una bevanda alcolica nell’anno.

A fronte di tali stime, è facile evincere quanto sia doveroso, tenere sotto controllo i parametri relativi all’alcolismo al fine di non incorrere in situazioni problematiche. In particolare, questa tematica oltre che da un punto di vista puramente descrittivo, può essere anche studiata e settorializzata in vari contesti della quotidianità nazionale, ed in particolar modo è senz’altro interessante capire come il consumo di alcool può influenzare un determinato fenomeno come l’automobilismo o la medicina.

Considerare però un’indagine statistica su una popolazione molto elevata come ad esempio il mondo degli automobilisti italiani o anche pensare di simulare una realtà così grande, per vedere gli effetti che l’alcool provoca su di essa, è un’operazione molto complessa. Per questo motivo in statistica si preferisce utilizzare quella che viene definita come *Inferenza Statistica*: che si pone l’obbiettivo di analizzare un particolare fenomeno su un determinato campione estratto da una popolazione, e cerca di rapportare i risultati ottenuti dal campione su una popolazione statistica infinita o estremamente grande.

Uno degli obbiettivi principali dell’inferenza statistica, è quello studiare una popolazione descritta da una variabile aleatoria osservabile X, la cui funzione di distribuzione ha una forma nota e riconducibile a standard teorici, ma contiene uno o più parametri non noti che ne caratterizzano l’andamento. Stimare questi parametri non noti, al fine di ottenere una buona analisi, è compito di quella che in statistica inferenziale, prende il nome di stima dei parametri, la quale può essere:

* Puntuale: se per ogni termine non noto si decide di utilizzare un valore fisso;
* Intervallare: se ogni valore non noto, viene invece rilevato da un certo intervallo, detto intervallo di confidenza (da cui appunto si estrae poi un determinato parametro non noto, con un certo grado di confidenza).

Queste stime verranno poi successivamente validate, con un’ulteriore procedura che prende il nome di verifica delle ipotesi, dove appunto secondo alcune congetture e metodologie, si andrà a determinare se il parametro rilevato è adatto o meno per il campione studiato.

# Problematica in esame e variabili aleatorie

Il tema del consumo di alcol in Italia, così come in altri paesi, è vincolato da stringenti leggi, le quali però molto spesso vengono infrante, mettendo anche in serio pericolo l’incolumità di chi è a stretto contatto con il trasgressore. In particolare, il mondo dell’automobilismo è strettamente vincolato da questa problematica, visto che non pochi automobilisti che vengono sottoposti a controllo risultano avere un tasso alcolemico al di fuori del limite consentito dalla scala giuridica che vieta o consente la guida in relazione ai livelli di i livelli di alcol nel sangue. In particolare, data questa premessa, viene spontaneo chiedersi:

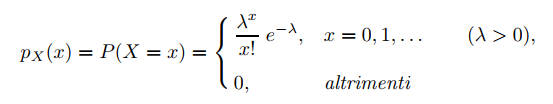
*“Dato una certa popolazione, o un certo campione di automobilisti fermati dalle forze dell’ordine per controlli di routine ed in particolare ad alcol test, quanti di essi risultano avere un tasso alcolemico superiore ai limiti consentiti dalla legge?”.*

Questa problematica, per ogni elemento del campione, risulta essere caratterizzata da un controllo indipendente rispetto agli altri individui, in altri termini, ogni conducente d’auto che viene sottoposto all’alcol test, può risultare in regola o meno rispetto alle norme vigenti, indipendentemente dall’esito di altri conducenti d’auto. In statistica queste tipologie di fenomeni che possono avere soltanto due possibili esiti prendono il nome di prove di Bernulli, ed in particolare quando si considera più di una prova verrebbe da pensare che la variabile più adatta a questa tipologia di fenomeni, sia la variabile discreta Binomiale.

## Passo 1 – La variabile aleatoria di Poisson

Però come si può effettivamente considerare dalla la problematica in esame, sarebbe poco realistico o molto confusionario utilizzare la distribuzione Binomiale, in quanto pur trattandosi di fenomeni di conteggio, il numero di persone che risulta avere un tasso alcolemico fuori limite, tra le tante entità di un campione che risultano essere sottoposte ad alcol test, sicuramente è un numero molto esiguo. Per effettuare delle stime più corrette rispetto a questa tipologia di dati, è senz’altro più significativo considerare la variabile aleatoria di Poisson, che appunto si presta molto meglio a quelli che in statistica vengono definiti come eventi Rari.

La variabile aleatoria di Poisson è una variabile aleatoria di tipo discreto, che è vincolata dalla seguente funzione di probabilità:



Dove appunto x è appunto il valore che la variabile aleatoria assume nel range di valori che la funzione di probabilità consente (in questo caso da 0 ad infinito), mentre il parametro λ, è un parametro non noto che determina al meglio con che probabilità, la variabile aleatoria X, assume un certo valore x, secondo la legge riportata.

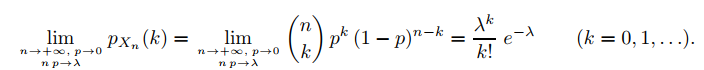
In particolare, è noto che il valore medio che la variabile Aleatoria assume e la varianza tra gli elementi di una data distribuzione Poissoniana (che in probabilità si ricava come momento del secondo ordine – momento del primo ordine al quadrato), risultano essere proprio uguali al parametro non noto λ.



Quest’osservazione, sarà molto importante ai fini dell’indagine da condurre, perché anche successivamente, tramite il parametro λ, si potrà avere una stima quantitativa del valore medio assunto dalla variabile aleatoria.

Come osservato, la variabile aleatoria di Poisson, è indicatrice di una distribuzione di probabilità, che consente di stimare quante volte un evento raro (come quello in esame), rispetto ad un campione molto grande di individui, ed in particolare essa consente di avere una stima intrinseca del valore medio e della dispersione di questi dati, dati appunto dal parametro non noto λ.

È matematicamente dimostrato che la distribuzione di Poisson, è perfettamente utilizzabile al posto di una variabile binomiale nel seguente modo:



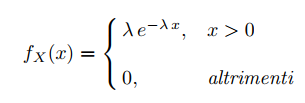
Qui appunto si evince come la variabile aleatoria di Poisson, sia eguagliata al limite per N (numero di individui del campione), per p (probabilità) tendente a 0 (quindi probabilità molto bassa) di una generica variabile binomiale, ad una funzione di probabilità di Poisson dove appunto si eguaglia il valore medio λ, proprio al prodotto di n e p.

Da questa deduzione si giustifica matematicamente come per eventi con probabilità molto rara di verificarsi su un numero molto elevato di individui, fattispecie il numero di automobilisti positivi all’alcool test su un numero molto grande di individui posti a controllo, possano essere rappresentati tramite una distribuzione di Poisson, visto che è una buona approssimazione della variabile binomiale nelle suddette condizioni.

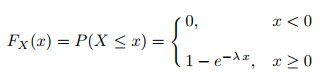
## Passo 2 – accenni sulla variabile aleatoria esponenziale

Dato che il problema così formulato, è intrinseco in una natura principalmente discreta, si è optato per la distribuzione di Poisson. Però nella statistica inferenziale, si può facilmente fare deduzioni tramite questa distribuzione, anche nel contesto continuo (ad esempio con l’utilizzo di altre distribuzioni come la Normale o Gaussiana oppure tramite la variabile aleatoria Esponenziale).

In particolare, le cui funzione di densità di probabilità e di distribuzione di una variabile aleatoria esponenziale sono così definite per un intervallo di valori compreso tra 0 e + infinito:



*Funzione di densità*



*Distribuzione di probabilità*

Da queste funzioni, si evince come anche questa variabile aleatoria, sia strettamente legata all’utilizzo di un parametro non noto λ, che assume lo stesso identico significato per cui è stato definito con la distribuzione di Poisson.

In particolare, per la distribuzione esponenziale, il valore medio e la varianza vengono così definiti:



Come facilmente si osserva, la distribuzione esponenziale ha come valor medio il reciproco del fattore lambda, e questo è un ottimo indicatore della frequenza media con cui due eventi descritti da una variabile Poissoniana si verificano.

La variabile esponenziale, è stata riportata, dato che anche il problema in esame può essere visto anche nel seguente modo:

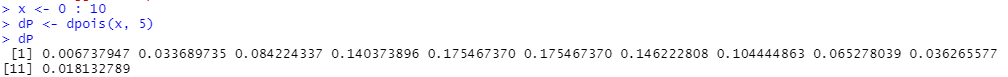
*“Dato una certa popolazione, o un certo campione di automobilisti fermati dalle forze dell’ordine per controlli di routine ed in particolare ad alcol test, con quale frequenza, si risulta avere un tasso alcolemico superiore ai limiti consentiti dalla legge, tra individui del campione?”.*

La variabile Esponenziale, può essere un buon metro di analisi per questa deduzione, quindi una buona ipotesi di verifica progettuale che potrebbe essere intrapresa nel seguito dell’analisi, sarebbe proprio quella di utilizzare lo stesso fattore λ, ottenuto con le tecniche di stima effettuate con la variabile di Poisson, al fine di verificare questo parametro anche in termini di frequenze.

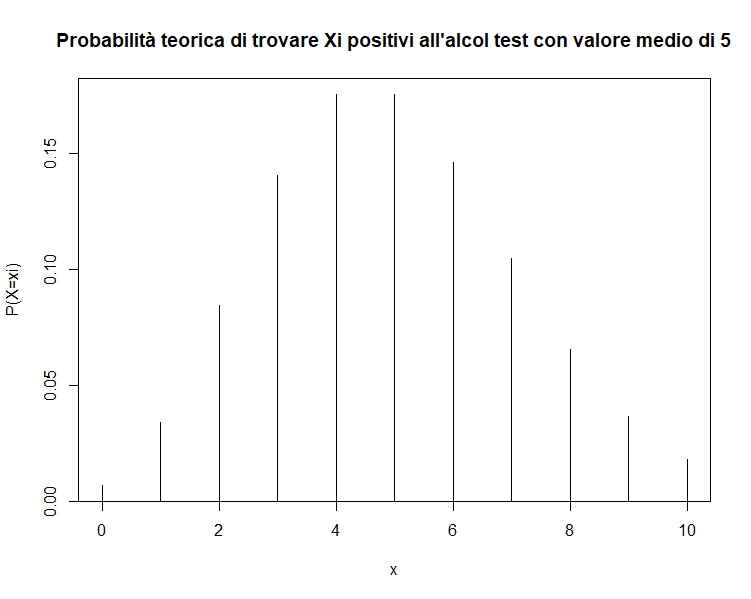
# Applicazione della variabile aleatoria di Poisson al problema in analisi

Al fine di capire con che probabilità un certo numero xi di automobilisti sottoposti ad alcol test, superi i livelli di guardia, è possibile utilizzare il sistema informatico R, applicando la funzione di probabilità per una variabile aleatoria di Poisson, dpois (x, lambda), per la quale sarà necessario indicare:

* Un certo vettore di valori x, il quale conterrà un certo intervallo di valori per il parametro xi, precedentemente definito, tale vettore può essere anche espresso per valori compresi tra 0 e + infinito, così come definito dalla funzione di probabilità di Poisson, ai fini dell’indagine, lo si porrà tra 0 e 10;
* Un determinato coefficiente medio Lambda, tale parametro non è noto a prescindere, e successivamente andrà calcolato tramite le tecniche di stima, per questa prima simulazione, tale parametro verrà posto ad un valore medio indicativo di 5 positivi all’alcol test per ogni prova di Poisson effettuata;

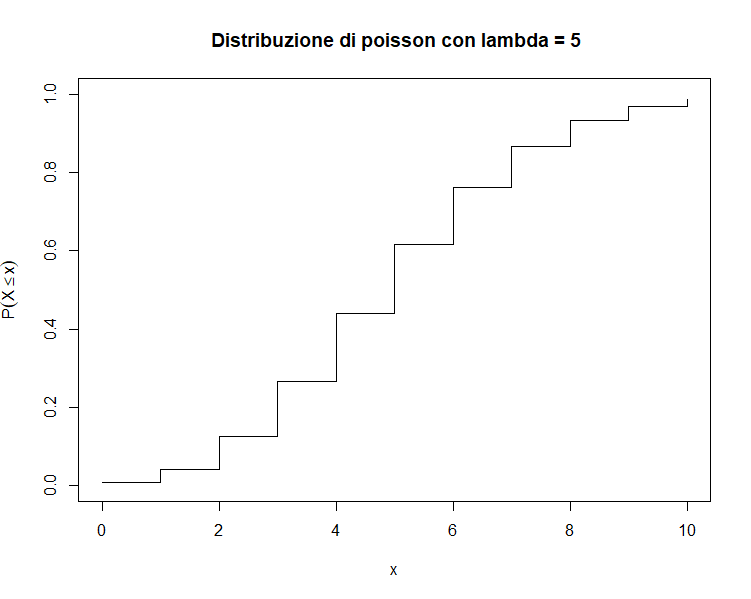


Con questa semplice formulazione, si nota come la probabilità di trovare un xi positivi all’alcol test, intervalla tra 0 e al massimo 0.18 in due punti specifici, dove raggiunge un massimo.



Graficamente sì osserva ancora meglio come la variabile aleatoria di Poisson, distribuisce le probabilità di trovare xi positivi all’alcol test (con Xi, compreso tra 0 e 10), con dei picchi di massimo tra 4 e 5 individui positivi.

Questa prima deduzione può essere anche confermata dalla visione della funzione di distribuzione di Poisson, applicata con i parametri utilizzati con la funzione di probabilità, tramite la funzione di R PPOIS:



Come si nota i picchi massimi di Avanzamento, si hanno proprio quando si considera che la probabilità che X sia <= xi, quando appunto xi è proprio uguale a 4 o 5 individui.

Da questa deduzione, dato un certo numero campione teorico, si può facilmente evincere come il numero di trasgressori alla prova di alcool test intervalli tra 4 e 5 con una probabilità di circa 0.18.

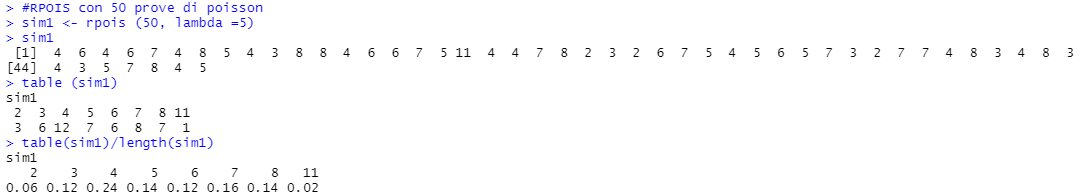


In particolare, considerando anche i quantili per la distribuzione di probabilità, è facile osservare anche come il 50% dei dati sia posto proprio al valore medio lambda = 5 fissato a priori. Ciò da appunto l’idea di come lambda sia il valore più probabile per rappresentare il numero di trasgressori all’alcol test. Ed in particolare, quando si raggiungerà una buona stima di questo parametro si potrà definire al meglio con che probabilità Xi automobilisti, risultino positivi al test alcolemico, sapendo appunto che il numero di individui medio lambda, risulterà appunto essere il più probabile rispetto ad altri valori, anche se da come si è notato, con la stima ipotetica che fissa lambda = 5, anche la probabilità più alta assumerà un valore non molto elevato.

*Ma cosa succede quando si confrontano la probabilità teoriche con i risultati ottenuto con un campione ben definito?*

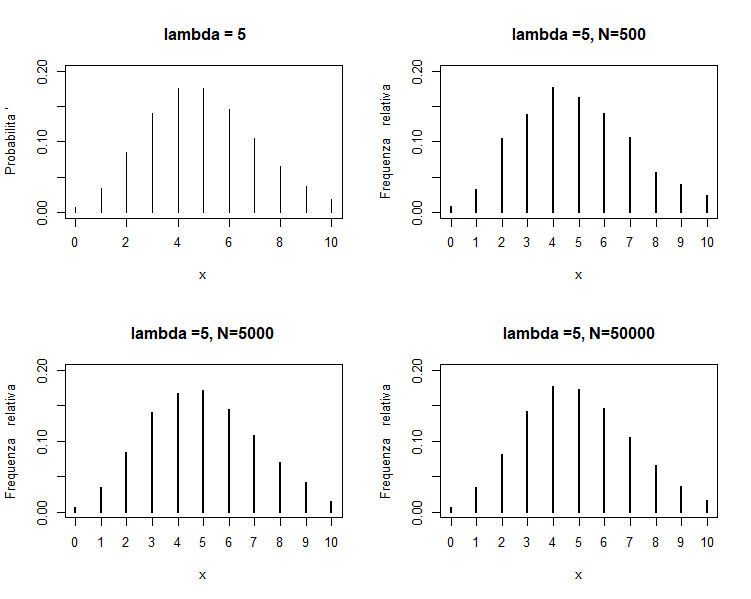
Il sistema R mette a disposizione la funzione RPOIS, che permette facilmente di un numero N di prove Poisson, la quale riporterà un vettore di N valori, che appunto indicherà per ogni prova di Poisson, il numero di persone positive all’evento che la variabile descrive nella singola prova di Poisson.

Se ad esempio si considera un numero di prove di Poisson pari a 50, ed un fattore medio sempre posto a 5, la funzione RPois, restituisce un vettore del tipo;



Con 50 prove simulate di Poisson, di nota anche in questo caso come le probabilità teoriche siano rispettate, con un picco massimo posto sempre tra 4 e 5 individui, anche se il valore non rispetta esattamente il valore teorico.

*Ma cosa succede se il numero di prove aumenta ad esempio a 500, 5000, o 50000?*



Confrontando la funzione di probabilità teorica con i grafici inerenti alle frequenze relative ai valori di xi, si osserva molto facilmente come con 500, 5000 e 50000 prove simulate con un ipotetico campione di individui ben definito, si nota come la distribuzione di frequenze, sia sempre più simile alla funzione teorica con l’aumentare delle prove. Questo è senz’altro un risultato molto positivo, ai fini dello studio condotto, perché permette già di affermare che una volta che il fattore medio lambda non noto, sarà correttamente stimato in relazione al problema, si potrà avere un’attendibile report grafico di probabilità per Xi individui, e soprattutto si potrà determinare con chiarezza, con che probabilità il numero di positivi all’alcol test per un dato campione sia pari al valore medio lambda.

# Stime puntuali dei parametri non noti

Uno dei principali problemi legato allo studio su una popolazione statistica, descritta da una variabile aleatoria), per una specifica problematica (di cui si conoscono forme teoriche come funzione di probabilità e distribuzione, è proprio quello di determinare con esattezza i parametri non noti da cui la variabile aleatoria dipende strettamente.

Al fine di estrarre queste informazioni dalla popolazione, si può far uso dell’inferenza statistica considerando un campione rappresentativo della popolazione, su cui sono state effettuate delle specifiche misurazioni in relazione al problema in analisi.

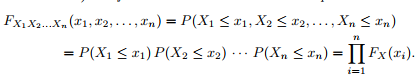
Molti metodi di inferenza statistica per la stima di parametri non noti, come il Metodo dei Momenti e il Metodo della Massima Verosimiglianza, per stimare puntualmente i parametri non noti di una variabile aleatoria fanno utilizzo di quelli che in statistica inferenziale vengono chiamati come Campioni Casuali.

In statistica inferenziale un campione casuale viene definito considerando una popolazione descritta da una variabile aleatoria osservabile X caratterizzata da funzione di distribuzione FX(x).

Un determinato vettore Il vettore aleatorio X1, X2, . . . , Xn `e detto campione casuale di ampiezza n se le variabili aleatorie del vettore sono:

* osservabili,
* indipendenti
* Identicamente distribuite con la stessa legge di probabilità dell’intera popolazione (ossia costituiscono delle osservazioni di X).

La funzione di distribuzione del campione casuale è definita come:



Sulla base di un campione casuale di ordine N, nell’inferenza statistica si cerca di ottenere informazioni sui parametri non noti di una variabile aleatoria in riferimento ad una popolazione, tramite delle funzioni misurabili sul campione casuale.

Queste misure si dividono in:

* Statistiche - t(X1, X2, . . . , Xn): funzioni misurabili e osservabili del campione casuali che dipendono soltanto dal campione osservato, mentre i parametri non noti sono presenti soltanto nella funzione di distribuzione della statistica.
* Uno stimatore Θ = t(X1, X2, . . . , Xn): è invece una funzione misurabile su un campione casuale, i cui valori possono essere utilizzati per stimare un parametro non noto della popolazione ( tali valori sono infatti detti “stime del parametro non noto”).

Stimatori molto utilizzati in statistica per ottenere dei plausibili valori puntuali per parametri non noti di una variabile aleatoria, sono la media campionaria e la varianza campionaria.

Però è importante sottolineare che tali stimatori sono ragionevoli se e solo se il campione casuale in analisi è sufficientemente grande (Nella statistica inferenziale, tale ipotesi è soddisfatta quando si ha a disposizione un campione maggiore di 30 unità).

## Passo 1 – Stima del parametro non noto λ con il metodo dei momenti

Uno dei più antichi metodi di statistica inferenziale per la stima puntale dei parametri non noti, è il cosiddetto metodo dei momenti campionari.

In particolare, un momento di ordine R in riferimento ad un campione casuale viene definito come:



Da ciò ne deriva anche che il momento di primo ordine è strettamente equivalente allo stimatore di media campionaria del campione.

Il metodo dei momenti, eguaglia i primi k momenti della popolazione in esame (teorici descritti dalla variabile aleatoria), con i corrispondenti momenti campionari:



Tale uguaglianza è molto importante dato che nel calcolo dei momenti della popolazione, rientrano sempre i parametri non noti di una variabile aleatoria e questi ultimi possono essere facilmente ricavati appunto tramite l’uguaglianza.

Nel caso specifico del problema di conteggio di automobilisti sottoposti a controlli di alcooltest, è possibile considerare tele campione casuale di 50 elementi:

X = (4, 6 , 4, 6, 7, 4, 8, 5, 4, 3, 8, 8, 4, 6, 6, 7, 5, 11, 4, 4, 7, 8, 2, 3, 2, 6, 7, 5, 4, 5, 6, 5, 7, 3, 2, 7, 7, 4, 8, 3, 4, 8, 3, 4, 3, 5, 7, 8, 4, 5 )

Dove ogni xi è il numero di automobilisti trovati positivi ad un controllo di alcol test in una sera da un posto di blocco dei carabinieri in un centro urbano affollato.

Con riferimento a tale campione, è possibile calcolare i vari momenti di ordine N e stimarne i parametri non noti della funzione di probabilità che che descrive il fenomeno, rispetto ad un’ampia popolazione di automobilisti che ogni sera frequenta il dato centro urbano (Per applicabilità del metodo per si considera ovviamente può essere fermato anche più volte dalle forze dell’ordine, per essere sottoposto ad alcol test).

Come già osservato in precedenza, tale problematica, è al meglio descritta tramite una distribuzione di Poisson. Quindi al fine di applicare al meglio il teorema dei momenti, è possibile anche considerare le singole xi del campione casuale come i valori aleatori assunti da variabili di Poisson X1,…, X50 indipendenti e equivalentemente disturbate.

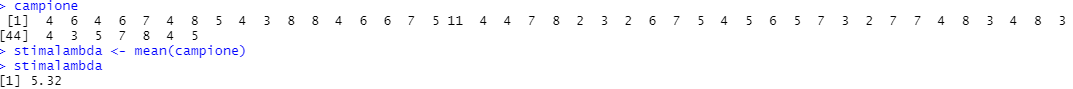
In tal caso, il metodo dei momenti permetterà direttamente di stimare il parametro non noto λ, tramite il calcolo del momento campionario di primo ordine (media campionaria), al fine di ottenere una variabile aleatoria di Poisson X, che descrive al meglio l’intera popolazione per il fenomeno in analisi.

In particolare, per il calcolo del valore medio di λ per una variabile di Poisson il metodo dei momenti dice che:



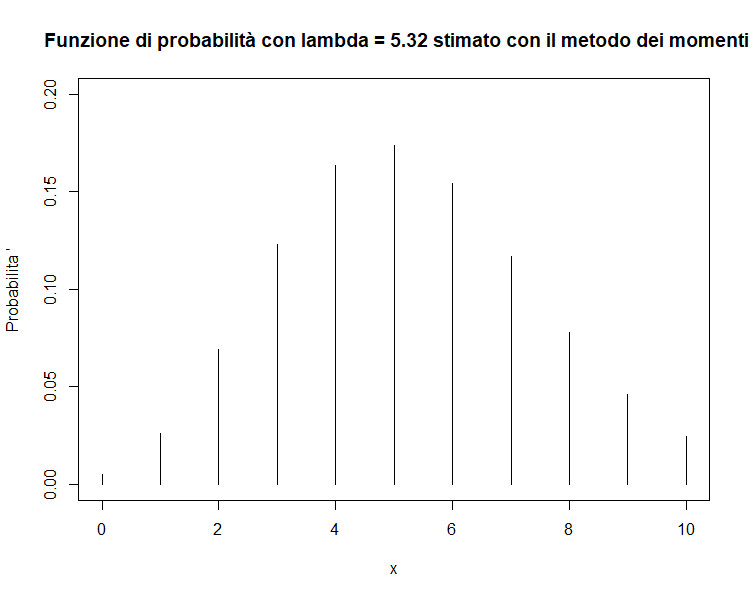
Il fattore non noto della variabile X per l’intera popolazione, è esprimibile come la media campionaria de campione casuale dei risultati ottenuti dalle singole prove di indagine riportate.

Applicando al campione soprariportato il metodo dei momenti, si ottiene un valore medio λ di pari a: 5.32



Trattandosi di un valore medio con due cifre decimali, tale parametro stimato, potrebbe essere già accettabile così, ma solitamente essendo che il metodo dei momenti è basato su un’uguaglianza molto semplice, tale risultato va validato o sostituito (compito che in statistica spetta ai decisori di parametri), con altre stime fatte con dei metodi di stima più complessi, come il metodo della massima verosimiglianza.

In ogni caso, è possibile osservare graficamente la funzione di probabilità teorica di una variabile di Poisson con probabilità pari a 5.32, come descrive al meglio la popolazione in esame da cui sono stati estratti i dati campionari per la stima:



Come è possibile osservare da questo diagramma a barre, è facile intuire come la probabilità di trovare positivi all’alcool test su un numero grande di controlli resti comunque molto bassa, ma comunque ha un massimo che indica che il numero medio di positivi all’alcol test è di 5 individui con una probabilità pari a:

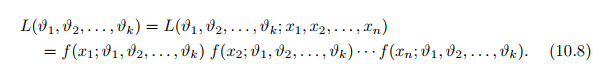


## Passo 2 – Stima del parametro non noto λ con il metodo della massima verosimiglianza

Effettuare una stima puntuale di un parametro non noto in statistica inferenziale, molto spesso non è un’operazione che si può ridurre ad un unico criterio applicativo, infatti molto spesso il metodo dei momenti, viene affiancato a quello che in statistica viene definito il metodo della massima Verosimiglianza.

Per un campione casuale di ampiezza N estratto da una popolazione. La funzione di verosimiglianza L(ϑ1, ϑ2, . . . , ϑk; x1, x2, . . . , xn), è la funzione di probabilità congiunta o di densità di probabilità del campione casuale X1,…,Xn.

In particolare, si è visto che per variabili aleatorie indipendenti ed identicamente distribuite è possibile calcolare la funzione di verosimiglianza (di probabilità congiunta) come il prodotto delle singole funzioni di probabilità rispetto ai parametri non noti.



Ovviamente, se si considera la funzione di probabilità congiunta, essa sarà strettamente dipendente dai parametri non noti della variabile che rappresenta la popolazione sotto analisi.

Ed in particolare per una popolazione di Poisson, dove appunto la funzione di probabilità sarà pari a:



la funzione di Verosimiglianza sarà quindi uguale a:



Volendo ora stimare il valore medio λ, estraendolo da questa funzione, è necessario calcolare la derivata prima rispetto al parametro non noto ed eguagliare tale calcolo a 0 in modo tale da estrarre il valore di λ più grande possibile (quindi il valore di massima verosimiglianza). In particolare, visto che la funzione è composta da elementi del campione all’esponente si preferisce considerare la funzione che ne deriva applicando il logaritmo ad entrambi i membri.

Seguendo questo ragionamento, per il metodo di massima verosimiglianza, il valore medio λ, viene stimato nel seguente modo:



In particolare, si osserva che anche in questo caso lo stimatore che ne deriva è proprio la media campionaria del campione casuale. Volendo quindi verificare che cosa accade tramite il metodo di massima verosimiglianza per il fenomeno in analisi, si otterrebbero gli stessi identici risultati ottenuti con il metodo dei momenti.