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:

Prof.ssa. Amelia G. Nobile

Studente:

Ferrara Carmine

Matr.05225/00990

ANNO ACCADEMICO 2020/2021

Sommario

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:

$$p_X(x) = P(X = x) = \begin{cases} \dfrac{\lambda^x}{x!}\, e^{-\lambda}, & x = 0, 1, \ldots \qquad (\lambda > 0), \\ 0, & \textit{altrimenti} \end{cases}$$

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.

$$E(X) = \lambda, \qquad \mathrm{Var}(X) = \lambda.$$

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:

$$\lim_{\substack{n\to+\infty,\ p\to 0\\ n\,p\to\lambda}} p_{X_n}(k) = \lim_{\substack{n\to+\infty,\ p\to 0\\ n\,p\to\lambda}} \binom{n}{k} p^k\,(1-p)^{n-k} = \frac{\lambda^k}{k!}\,e^{-\lambda} \qquad (k=0,1,\ldots).$$

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:

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

Funzione di densità

$$F_X(x) = P(X \le x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \ge 0 \end{cases}$$

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:

$$E(X) = \frac{1}{\lambda}, \qquad E(X^2) = 2\left(\frac{1}{\lambda}\right)^2, \qquad \mathrm{Var}(X) = E(X^2) - [E(X)]^2 = \frac{1}{\lambda^2}.$$

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.

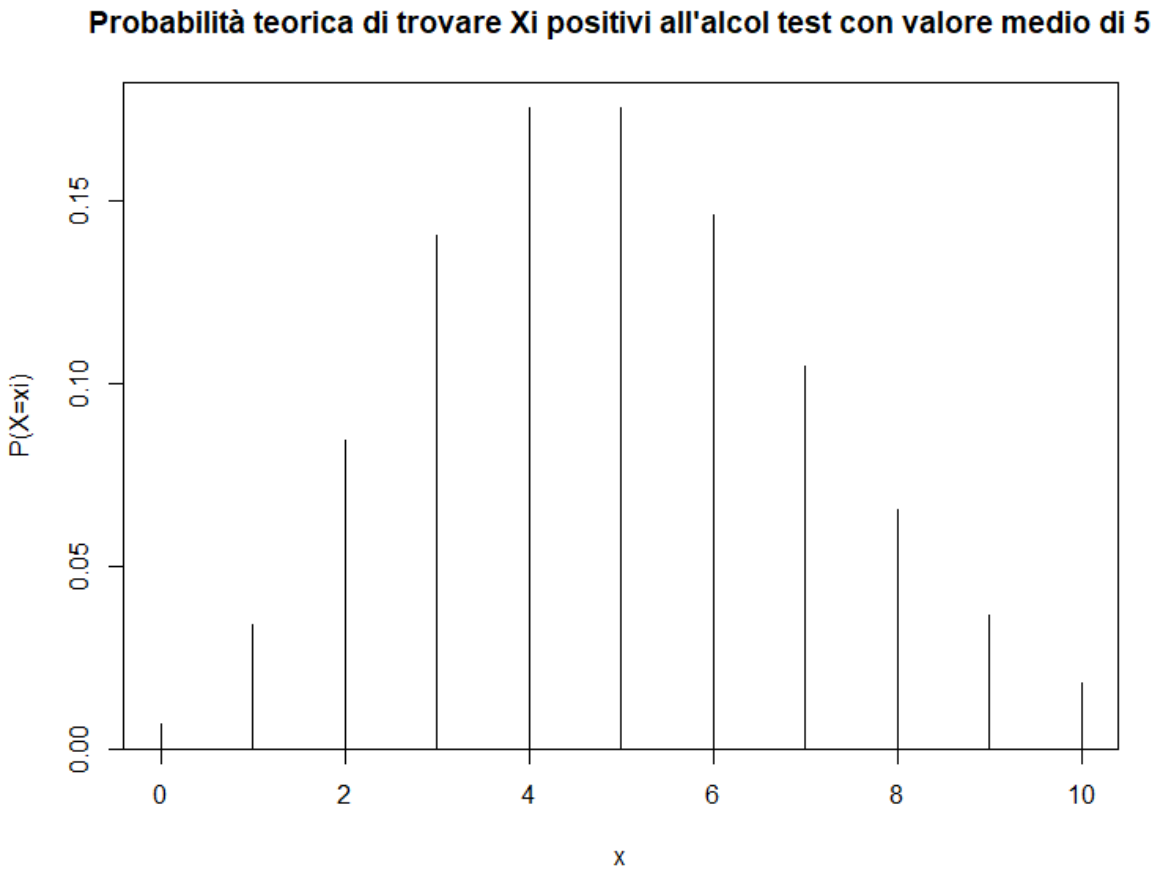
Applicazione della variabile aleatoria di Poisson al problema in analisi

Al fine di capire con che probabilità un certo numero x_i 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 x_i, 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;

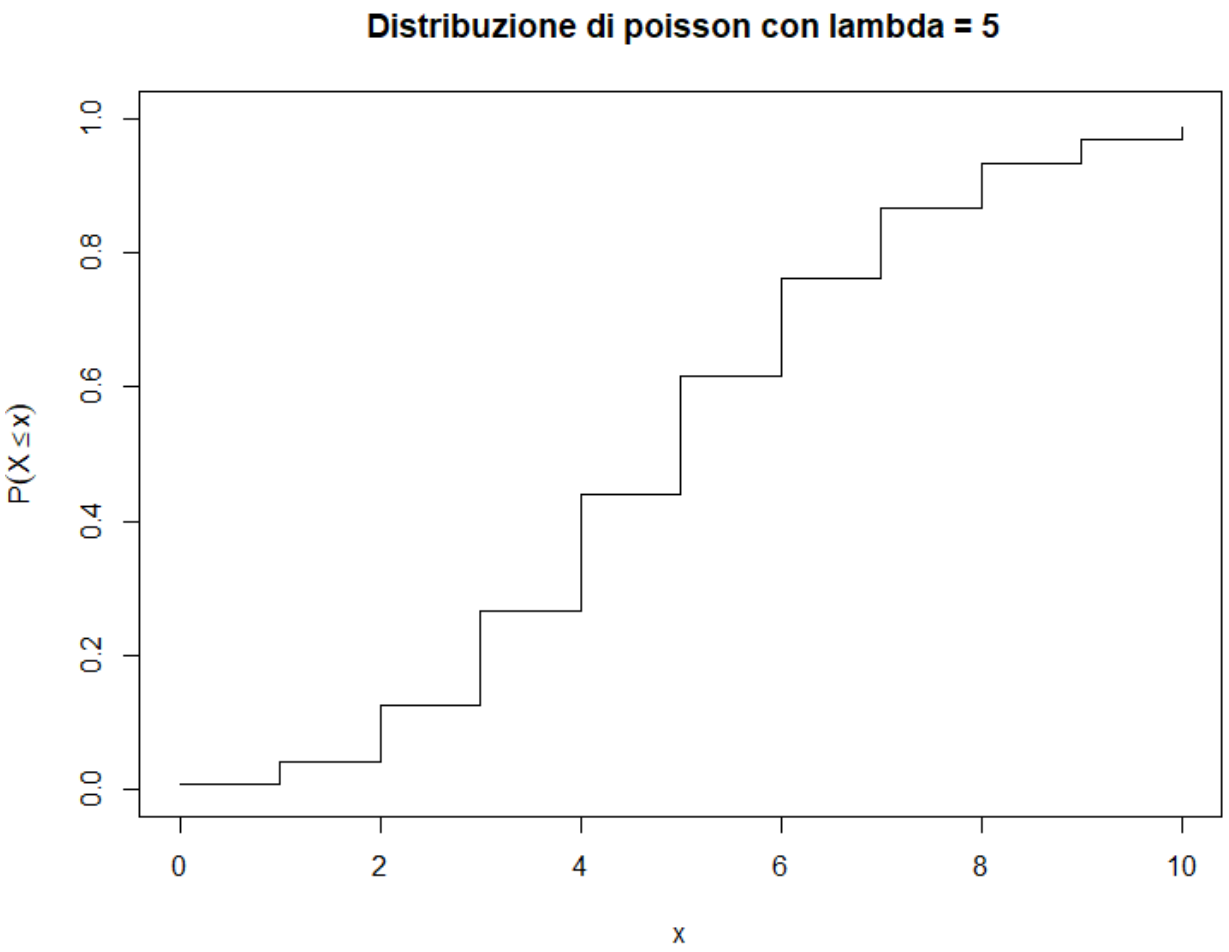
```
> x <- 0 : 10
> dP <- dpois(x, 5)
> dP
 [1] 0.006737947 0.033689735 0.084224337 0.140373896 0.175467370 0.175467370 0.146222808 0.104444863 0.065278039 0.036265577
[11] 0.018132789
```

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.

```
> z<-c(0 ,0.25 ,0.5 ,0.75 ,1)
> qpois(z, lambda = 5)
[1]   0   3   5   6 Inf
```

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?

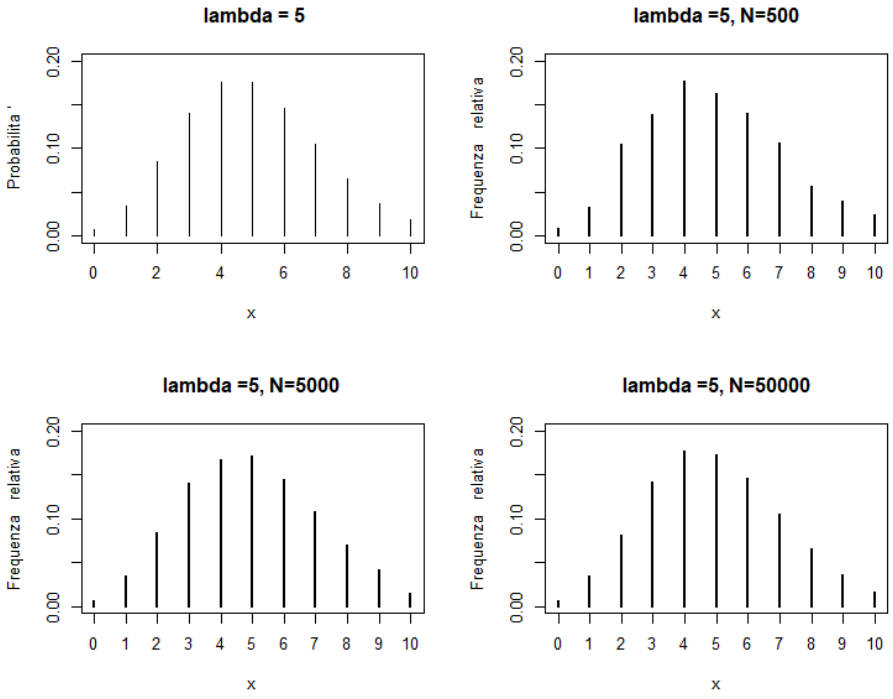
Il sistema R mette a disposizione la funzione RPOIS, che permette facilmente di un numero N di prove Poisson, la quale riporterà un vettore di N valori, che appunto indicherà per ogni prova di Poisson, il numero di persone positive all'evento che la variabile descrive nella singola prova di Poisson.

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;

```
> #RPOIS con 50 prove di poisson
> sim1 <- rpois (50, lambda =5)
> sim1
 [1]  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
[44]  4  3  5  7  8  4  5
> table (sim1)
sim1
 2  3  4  5  6  7  8 11
 3  6 12  7  6  8  7  1
> table(sim1)/length(sim1)
sim1
   2    3    4    5    6    7    8   11
0.06 0.12 0.24 0.14 0.12 0.16 0.14 0.02
```

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:

$$F_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$
$$= P(X_1 \leq x_1)\, P(X_2 \leq x_2) \cdots P(X_n \leq x_n) = \prod_{i=1}^{n} F_X(x_i).$$

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:

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

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:

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

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 x_i 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 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 x_i del campione casuale come i valori aleatori assunti da variabili di Poisson X_1,…, X_{50} 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:

$$\widehat{\lambda} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \overline{x}.$$

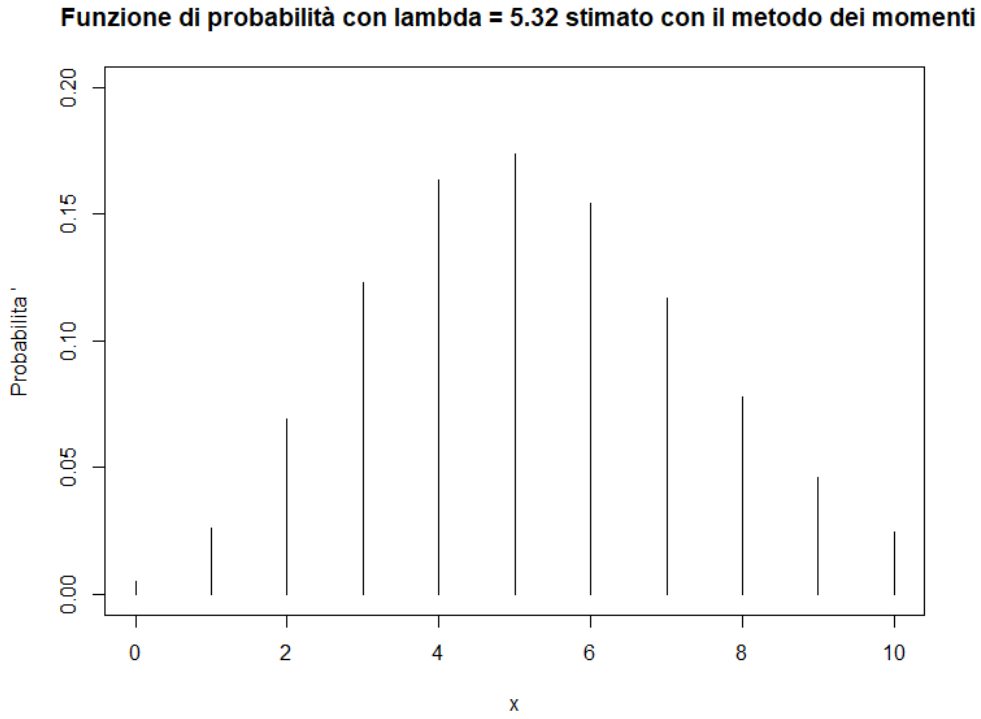
Il fattore non noto della variabile X per l'intera popolazione, è esprimibile come la media campionaria de campione casuale dei risultati ottenuti dalle singole prove di indagine riportate.

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

```
> campione
 [1]  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
[44]  4  3  5  7  8  4  5
> stimalambda <- mean(campione)
> stimalambda
[1] 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:

```
> dpois (5, lambda = stimalambda)
[1] 0.1737522
```

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 $X_1,\ldots,X_n$.

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.

$$\begin{aligned} L(\vartheta_1, \vartheta_2, \ldots, \vartheta_k) &= L(\vartheta_1, \vartheta_2, \ldots, \vartheta_k; x_1, x_2, \ldots, x_n) \\ &= f(x_1; \vartheta_1, \vartheta_2, \ldots, \vartheta_k)\, f(x_2; \vartheta_1, \vartheta_2, \ldots, \vartheta_k) \cdots f(x_n; \vartheta_1, \vartheta_2, \ldots, \vartheta_k). \end{aligned} \quad (10.8)$$

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:

$$P(X = x; \lambda) = \frac{\lambda^x}{x!}\, e^{-\lambda} \qquad (x = 0, 1, \ldots).$$

la funzione di Verosimiglianza sarà quindi uguale a:

$$L(\lambda) = \frac{\lambda^{x_1 + x_2 + \ldots + x_n}}{x_1!\, x_2! \cdots x_n!}\, e^{-n\,\lambda}$$

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:

$$\widehat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}.$$

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.

Passo 3 – Valutazioni del parametro stimato

La stima del valore medio lambda effettuata tramite il metodo dei momenti e validata anche dal metodo di massima verosimiglianza, è stata effettuata sul principio che il parametro non noto è proprio identico al valore medio della popolazione.

$$\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}.$$

Quindi, lo stimatore utilizzato per descrivere al meglio la popolazione in analisi, è stata la media campionaria del valore casuale.

Ma tale stimatore è il migliore per descrivere al meglio la popolazione in esame?

Per verificare la bontà di questo campione è necessario ragionare sulle proprietà di cui esso gode. In particolare, in statistica uno stimatore di un parametro non noto si definisce corretto se il valore medio dello stimatore stesso corrisponde proprio al parametro non noto con esso stimato.

Definizione 10.6 *Uno stimatore* $\widehat{\Theta} = t(X_1, X_2, \ldots, X_n)$ *del parametro non noto* ϑ *della popolazione è detto corretto (non distorto) se e solo se per ogni* $\vartheta \in \Theta$ *si ha*

$$E(\widehat{\Theta}) = \vartheta, \tag{10.9}$$

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

Ma tra le proprietà della media campionaria di un campione casuale, si osserva proprio che il valore medio della variabile è proprio uguale al valore medio della variabile media del campione casuale che si sta considerando.

$$E(\overline{X}) = E\Big[\frac{1}{n}\sum_{i=1}^{n} X_i\Big] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu.$$

Ma essendo che per la variabile di Poisson è proprio uguale a lambda, ne deriva automaticamente che il valore medio dello stimatore media campionaria, applicato al campione casuale è proprio uguale a lambda. Essendo quindi che la proposizione è pienamente rispettata si può affermare che la media campionaria è uno stimatore corretto e quindi effettivamente il valore stimato è pienamente rappresentativo del fenomeno relativo ai controlli di alcol test in analisi.

*si denota che se non si fosse giunti alla conclusione di correttezza per lo stimatore adoperato, si sarebbe dovuto procedere ad effettuare alcune deduzioni in termini della proprietà di correttezza asintotica dello stimatore che per grandi campioni, indica che al limite per n tendente ad infinito il valore medio dello stimatore corrisponde al valore stimato così come la stima della varianza per la variabile aleatoria normale.

Ma lo stimatore utilizzato è effettivamente il migliore possibile?

Per rispondere a tale quesito è necessario confrontare vari stimatori in termini di dispersione, ed un buon indice per questa problematica è quello che viene definito in statistica come errore quadratico medio MSE:

Definizione 10.7 *Sia $\widehat{\Theta} = t(X_1, X_2, \ldots, X_n)$ uno stimatore del parametro non noto ϑ della popolazione. Si chiama errore quadratico medio la quantità*

$$MSE(\widehat{\Theta}) = E\left[(\widehat{\Theta} - \vartheta)^2\right]. \qquad (10.10)$$

Che dipende strettamente dal valore assunto dallo stimatore in funzione del campione e dal valore non noto. In particolare, se si è nella classe degli stimatori corretti si dimostra che l'MSE è proprio equivalente alla varianza dello stimatore.

$$MSE(\widehat{\Theta}) = E\{[\widehat{\Theta} - E(\widehat{\Theta})]^2\} = \mathrm{Var}(\widehat{\Theta}).$$

Si può quindi dire che uno stimatore è migliore rispetto agli altri se la sua varianza (MSE) è minore o uguale di ogni altro stimatore per lo stesso parametro non noto.

(ii) $\mathrm{Var}(\widehat{\Theta}) \leq \mathrm{Var}(\widehat{\Theta}^\star)$ *per ogni altro stimatore* $\widehat{\Theta}^\star$ *corretto del parametro* ϑ.

Erroneamente si potrebbe pensare che per trovare lo stimatore migliore, sia necessario trovare la varianza più piccola tra tutti quelli possibili. Ciò però non è direttamente possibile, perché in statistica viene dimostrato tramite la disuguaglianza di Cramèr-Rao che per ogni stimatore vale la disuguaglianza:

$$\mathrm{Var}(\widehat{\Theta}) \geq \frac{1}{nE\left\{\left[\frac{\partial}{\partial\vartheta}\log f(X;\vartheta)\right]^2\right\}}.$$

Ed in particolare, se lo stimatore corretto, soddisfa addirittura all'uguaglianza questa relazione, esso viene poi definito come stimatore corretto e di varianza uniformemente minima per il parametro stimato.

$$\mathrm{Var}(\widehat{\Theta}) = \frac{1}{nE\left\{\left[\frac{\partial}{\partial\vartheta}\log f(X;\vartheta)\right]^2\right\}},$$

Nel caso della variabile aleatoria di Poisson, si dimostra che la disuguaglianza di Cramèr-Rao è soddisfatta proprio all' uguaglianza per la stima del valore medio lambda con lo stimatore di media campionaria.

$$\mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{\lambda}{n}, \qquad \frac{1}{nE\left\{\left[\frac{\partial}{\partial\lambda}\log p(X;\lambda)\right]^2\right\}} = \frac{\lambda}{n}.$$

Quindi si può dedurre che non solo il parametro lambda = 5.32 ottenuto sia corretto per rappresentare al meglio la funzione di probabilità in funzione del campione casuale su cui è stata effettuata l'indagine, ma per quanto si è appena ottenuto, si può anche dire che non è possibile effettuare una stima migliore (nella classe degli stimatori corretti), perché lo stimatore Media Campionaria rispetta la proprietà di varianza uniformemente minima.

Al fine di determinare la piena affidabilità dello stimatore per il parametro non noto lambda per la popolazione in esame, resta ora soltanto da valutare la proprietà di consistenza, in particolare in statistica:

Definizione 10.10 *Uno stimatore* $\widehat{\Theta}_n = t(X_1, X_2, \ldots, X_n)$ *del parametro non noto* ϑ *della popolazione è detto consistente se e solo se per ogni* $\varepsilon > 0$ *si ha*

$$\lim_{n\to+\infty} P\left(|\widehat{\Theta}_n - \vartheta| < \varepsilon\right) = 1,$$

ossia se e solo se $\widehat{\Theta}_n$ *converge in probabilità a* ϑ.

Ed in particolare si dimostra che per uno stimatore corretto o asintoticamente corretto, la proprietà di consistenza è rispettata se, valgono queste due uguaglianze:

i) $\lim_{n\to\infty} E(\widehat{\Theta}_n) = \vartheta,$

ii) $\lim_{n\to+\infty} \mathrm{Var}(\widehat{\Theta}_n) = 0.$

Ma nel caso della popolazione di Poisson rappresentata, la prima uguaglianza è rispettata proprio all'uguaglianza e quindi lo è anche in termini asintotici.

Mentre, sapendo che la varianza della media campionaria per stimare lambda è uguale a:

$$\mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{\lambda}{n},$$

È facilmente deducibile che al crescere di N il valore di varianza diventi sempre più piccolo, e se si ragioni in termini asintotici per n tendente ad infinito, il valore riportato è proprio uguale a 0.

Riassumendo quindi, lo stimatore Media Campionaria utilizzato per stimare il parametro lambda è corretto, consistente e di varianza minima.

Stime Intervallari dei parametri non noti

Aver stimato tramite uno stimatore corretto il valore non noto lambda che descrive il valore medio della distribuzione di Poisson, che date le ipotesi iniziali possa descrivere al meglio la popolazione del fenomeno di monitoraggio degli alcol test ad un posto di blocco, è senz'altro un buon risultato, tra l'altro anche validato dalla bassa presenza di cifre decimali presenti all'interno della stima riportata.

In Statistica inferenziale però è preferibile fornire per un parametro non noto, un determinato intervallo in cui questo parametro possa ricadere, al posto di un singolo valore una stima. Tale intervallo viene denominato intervallo di confidenza ed è caratterizzato da due estremi C_1 e C_2 tra i quali ricade il parametro non noto (lambda nel caso di Poisson) da stimare.

Tale intervallo di confidenza è legato al parametro da stimare secondo una data legge di probabilità:

$$P(\underline{C}_n < \vartheta < \overline{C}_n) = 1 - \alpha$$

Dove il coefficiente 1-α è detto grado di fiducia della stima, ed è un parametro noto fornito dal decisore, che indica il grado di attendibilità della stima fornita.

Metodo Pivotale

Per effettuare una stima intervallare di un parametro non noto, è possibile far ricorso a quello che viene definito metodo Pivotale. Tale metodo permette di stimare un parametro non noto di una popolazione distribuita secondo una particolare legge di probabilità, tramite l'utilizzo di una particolare variabile aleatoria, che prende il nome di variabile di Pivot, la quale deve godere di alcune caratteristiche particolari:

- La variabile di Pivot deve dipendere dal campione casuale X1, X2, . . . , Xn;
- La variabile di Pivot deve dipendere dal parametro non noto ϑ;
- la sua funzione di distribuzione non contiene il parametro ϑ da stimare.

Tale variabile quindi sarà della forma: γ(X1, X2, . . . , Xn; ϑ).

Sapendo che la variabile di Pivoting è legata da una legge di probabilità, vale la legge:

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

Se dalla disequazione interna si va ad isolare al centro soltanto il parametro non noto:

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

Si ottiene appunto la relazione che lega la stima intervallare al grado di fiducia 1 – α, dove appunto i coefficienti C1 e C2 sono proprio i valori calcolati dalle funzioni g1 e g2 derivanti dalla trasformazione della variabile di Pivot.

Per quanto potente, il metodo di Pivot, da risultati molto attendibili per campioni di qualsiasi dimensione, qualora esso sia applicato per stimare parametri non noti che intervengono nello stimare la distribuzione di probabilità, per Popolazioni che risultano essere distribuite normalmente.

Considerando però il problema di monitoraggio degli alcol test in analisi però, non è possibile ricorrere al metodo Pivotale diretto, dato che fin dall'inizio, si è supposto che la popolazione legata al fenomeno sia descrivibile tramite una popolazione di Poisson. Per determinare quindi una stima intervallare corretta del parametro non noto lambda è necessario ricorrere ad un'approssimazione del metodo pivotale, il quale prende appunto il nome di *Metodo Pivotale Approssimato*.

Passo 1 – Stima intervallare del parametro λ tramite metodo pivotale approssimato

Riconsiderando il precedente campione casuale, usato per la stima puntuale:

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).

che descrive il numero di persone che hanno riportato un valore superiore ai livelli di guardia all'alcol test per 50 prove indipendenti descrivibili ognuna con una distribuzione di Poisson. È necessario adesso dare un intervallo di confidenza per il parametro non noto λ (strettamente dipendente dal campione casuale).

Per raggiungere questo scopo, si vuole far riferimento al metodo pivotale approssimato, che si basa su un risultato fondamentale della statistica inferenziale, che è il teorema centrale di convergenza, il quale afferma che, per grandi campioni (in statistica per campioni con un numero di individui maggiore o uguale a 30), risulta che:

Se X denota la variabile aleatoria che descrive la popolazione con E(X) = μ e Var(X) = σ^2 (supposti entrambi finiti) e con (X1, X2, . . ., Xn) il campione casuale con n >= 30, definendo la media campionaria, il suo valore medio e la sua varianza come:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad E(\overline{X}) = \mu, \qquad \mathrm{Var}\,(\overline{X}) = \frac{\sigma^2}{n}$$

Vale che:

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

Ovvero che la variabile media, può essere standardizzata (sottraendo il valore medio e dividendo per la deviazione Standard) ad una variabile Z_n, converge in distribuzione ad un'altra variabile aleatoria Z – N(0,1), che risulta essere normale standard.

Tale variabile dipende strettamente dal campione casuale, sicuramente dipende dal parametro non noto ϑ, dato che esso per una qualsiasi distribuzione di probabilità è implicato nel calcolo del valor medio e della varianza teorica della distribuzione, la funzione di distribuzione di questa variabile non contiene nessun riferimento al valore ϑ, dato che per il teorema centrale di convergenza, è distribuita allo stesso modo di una Variabile Normale Standard.

Tale variabile può essere quindi vista come variabile di Pivot usabile per il calcolo dell'intervallo [C1 – C2] secondo un dato grado di fiducia 1 – α.

Dalla teoria che disciplina la statistica inferenziale, è possibile applicare il teorema centrale di convergenza e quindi il metodo pivotale approssimato anche ad una popolazione che risulta essere distribuita secondo una legge di probabilità di Poisson.

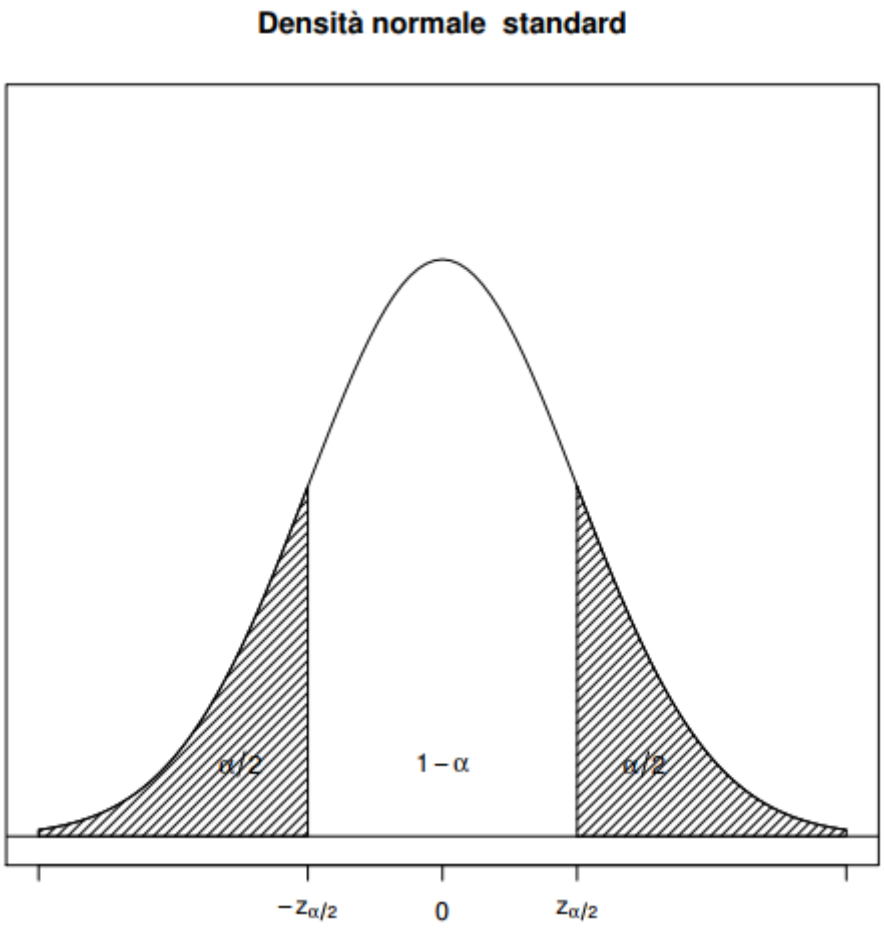
Nel caso specifico di una variabile X di Poisson, dove il valor medio e la varianza, risultano essere uguali a:

$$E(\overline{X}_n) = \lambda, \qquad \mathrm{Var}(\overline{X}_n) = \frac{\lambda}{n}$$

Considerando il campione casuale iniziale, è possibile utilizzare il teorema centrale di convergenza, da cui si ottiene una variabile aleatoria Zn del tipo:

$$\frac{\overline{X}_n - \lambda}{\sqrt{\lambda/n}} = \sqrt{n}\,\frac{\overline{X}_n - \lambda}{\sqrt{\lambda}}$$

Che dipende strettamente dal campione e dal parametro non noto λ, e dato che il campione è di 50 elementi converge in distribuzione ad una variabile Normale Standard Z-N(0,1), distribuita nel seguente modo:



Nel diagramma riportato, si citano due valori simmetrici rispetto allo 0, $-z_{\alpha/2}$ e $z_{\alpha/2}$ che racchiudono una densità di probabilità, proprio pari al grado di fiducia 1-α, che si vuole dare alla stima intervallare del parametro lambda. Dal diagramma riportato, ci si può chiedere con che probabilità la variabile standardizzata può assumere valori compresi tra $-z_{\alpha/2}$ e $z_{\alpha/2}$:

Da cui deriva la legge di densità:

$$P\left(-z_{\alpha/2} < \frac{\overline{X}_n - \lambda}{\sqrt{\lambda/n}} < z_{\alpha/2}\right) \simeq 1 - \alpha.$$

Sostituendo alla variabile media, la media campionaria (stimatore corretto e consistente), la disuguaglianza interna alla legge può essere espressa come:

$$\left[\sqrt{\frac{n}{\lambda}}\,(\overline{x}_n - \lambda)\right]^2 < z^2_{\alpha/2},$$

Ulteriormente trasformabile nella forma di equazione di secondo grado:

$$n\lambda^2 - \lambda(2n\overline{x}_n + z^2_{\alpha/2}) + n\overline{x}^2_n < 0.$$

Le cui soluzioni C1 e C2 sono nient'altro che gli estremi dell'intervallo di confidenza del parametro non noto λ, secondo un dato grado di fiducia 1 – α.

Con R è possibile andare a determinare i valori dell'intervallo di confidenza:

1. Calcolare il valore medio a partire dal campione casuale;
2. Calcolando i valori simmetrici $-z_{\alpha/2}$ e $z_{\alpha/2}$ della distribuzione normale standard tramite la funzione qnorm di R, in particolare basterà rilevare con la funzione qnorm, per che valore $z_{\alpha/2}$, vale la probabilità che ($P(X <= z_{\alpha/2}) = 1 – \alpha/2$)*;
3. Calcolare i coefficienti noti della disequazione soprariportata dati i valori precedenti e il numero di elementi.
4. Risolvere l'equazione sottesa dalla disequazione tramite la funzione polyroot di R.

1 – α/2 è il valore che assume la densità di probabilità normale standard da 0 fino a $z_{\alpha/2}$

In particolare, se si considera un grado di fiducia 1 – α pari a 0.95, con conseguente α pari a 0.5 si verifica che l'intervallo di confidenza per il parametro λ, secondo il metodo pivotale approssimato:

```
> campione <- c(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 )
>
> alpha <- 1 - 0.95
> #calcolo di z alpha/2
> qnorm (1 - alpha/2, mean =0, sd =1)
[1] 1.959964
>
> zalpha <-qnorm (1- alpha /2, mean =0, sd =1)
>
> n <- length(campione)
>
> #media campionaria (stima puntuale)
> medcamp <-  mean(campione)
> medcamp
[1] 5.32
>
> #calcolo dei coefficienti di secondo grado dalla
> #disequazione derivanti dal metodo pivotale approssiamto
> #applicato alla variabile di poisson
>
> a2 <- n
>
> a1 <- - (2 * n * medcamp + zalpha ^2)
>
> a0 <- n * medcamp^2
>
> #estremi dell'intervallo di confidenza per il parametro non noto lambda con
> #grado di fiducia del 95%
> polyroot (c(a0 ,a1 ,a2))
[1] 4.717941-0i 5.998889+0i
```

Si ha che il parametro non noto Lambda ricade nell'intervallo di confidenza [4.7 – 6.00].

Se invece si considera un grado di fiducia dello 0.99, con un relativo coefficiente α = 0.1:

```
> campione <- c(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 )
>
> alpha <- 1 - 0.99
> #calcolo di z alpha/2
> qnorm (1 - alpha/2, mean =0, sd =1)
[1] 2.575829
>
> zalpha <-qnorm (1- alpha /2, mean =0, sd =1)
>
> n <- length(campione)
>
> #media campionaria (stima puntuale)
> medcamp <-  mean(campione)
> medcamp
[1] 5.32
>
> #calcolo dei coefficienti di secondo grado dalla
> #disequazione derivanti dal metodo pivotale approssiamto
> #applicato alla variabile di poisson
>
> a2 <- n
>
> a1 <- - (2 * n * medcamp + zalpha ^2)
>
> a0 <- n * medcamp^2
>
> #estremi dell'intervallo di confidenza per il parametro non noto lambda con
> #grado di fiducia del 99%
> polyroot (c(a0 ,a1 ,a2))
[1] 4.543523-0i 6.229175+0i
```

Si ha un intervallo leggermente più grande che va da 4.54 a 6.22.

Ciò dimostra che all'aumentare del grado di fiducia, aumenta anche l'ampiezza dell'intervallo di confidenza per il parametro lambda. In ogni caso però è importante notare come la stima puntuale per il parametro ricada all'interno degli intervalli di confidenza calcolati tramite stime intervallari. Tali intervalli di confidenza validano ulteriormente quanto detto in precedenza per il problema in analisi.

Infatti non solo è possibile affermare che grazie agli studi effettuati sia tramite stime puntuali che intervallari, il numero medio di persone che risultano superare i livelli di guardia del test alcolemico per una grande popolazione di automobilisti è circa 5, ma dati gli intervalli di confidenza è possibile affermare che per grandi campioni di dati (estratti dalla popolazione di automobilisti sottoposti a test alcolemico), questo valore intervalla tra 4 e 6.

Applicazione della variabile aleatoria esponenziale al problema in analisi

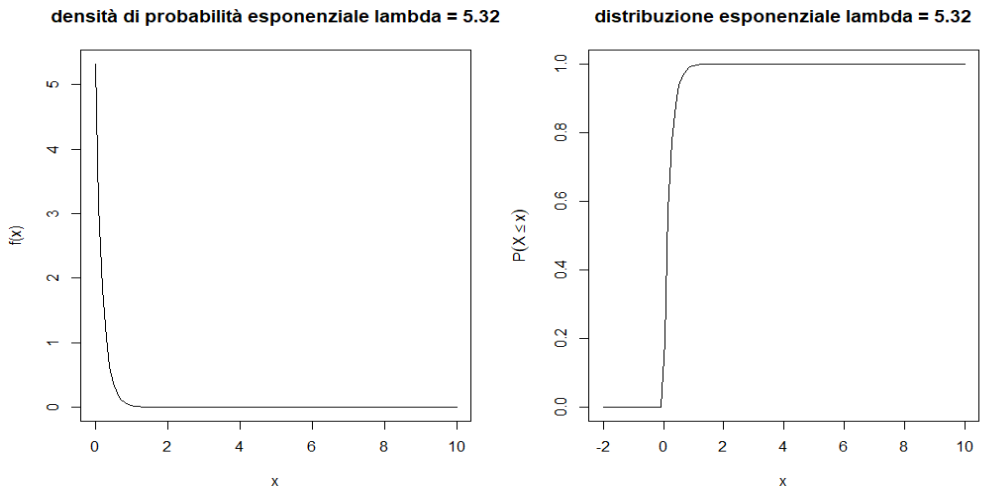
All'inizio del documento, si è voluto considerare il problema di verifica del test alcolemico in una chiave di lettura differente:

"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?"

Come già osservato in precedenza, questo problema può essere espresso per una grande popolazione tramite una variabile aleatoria esponenziale.

Avendo stimato un parametro lambda per la variabile aleatoria di Poisson, compreso nell'intervallo [4.7 – 6.00], è possibile ad esempio determinare con frequenza si verifichi la presenza in un determinato intervallo di tempo.

Supponendo infatti di avere un numero medio di persone che superano i valori di guardia al test alcolemico pari a 5.32 (stima puntuale) tra un numero molto grande di automobilisti sottoposti ad alcol test, sapendo che il valor medio per la variabile esponenziale è uguale ad 1/λ, si avranno 1/5.32 = 0.19 positivi al test alcolemico per unità di tempo, se ad esempio si stesse monitorando la frequenza in ore (altra ipotesi che dovrebbe essere validata da un'indagine statistica differente a quella proposta fino ad ora), si avrebbe quindi la certezza che statisticamente ad un posto di blocco si avrebbero meno di 1 persona all'ora che supera i valori limite della scala alcolemica alla guida.



Come si nota infatti anche dal diagramma di densità, trovare una persona per unità di tempo che supera i limiti consentiti dalla legge che disciplinano l'idoneità alla guida con l'assunzione di alcol, e fortunatamente molto bassa e circa pari allo 0, se si considera un valore stimato lambda compreso nell'intervallo di fiducia [4.7 – 6.00].

Confronti tra popolazioni di Poisson indipendenti

Oltre a rispondere a quesiti come la piccola digressione sulla frequenza con cui una persona risulta superare i livelli limite, sarebbe curioso provare a rispondere anche ad altri piccoli quesiti inerenti al problema in esame, magari mettendo a confronto popolazioni differenti.

Ad esempio, se si considerano i seguenti campioni casuali:

- 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)

Inerente al numero di persone che hanno superato i livelli di controllo al test alcolemico in 50 indipendenti posti di blocco stradali nell'area urbana e suburbana di Salerno.

- Y = (3, 2, 9, 6, 5, 0, 1, 2, 8, 2, 3, 3, 2, 5, 3, 4, 4, 2, 2, 3, 3, 5, 2, 7, 4, 1, 0, 8, 5, 3, 2, 2, 5, 2, 4, 3, 6, 7, 2, 6, 4, 6, 4, 3, 1, 5, 5, 6, 5, 4)

Inerente al numero di persone che hanno superato i livelli di controllo al test alcolemico in 50 indipendenti posti di blocco stradali nell'area urbana e suburbana di Napoli.

Partendo dall'ipotesi che i due campioni siano stati estratti da due popolazioni indipendenti, e supponendo che per il fenomeno in analisi, entrambe le popolazioni di automobilisti (sia quella relativa alla area metropolitana di Napoli che quella di Salerno) siano distribuite tramite Distribuzione di Poisson, ci si può chiedere ad esempio, quale delle due popolazioni, riporti un tasso di trasgressori al test alcolemico medio superiore.

Procedendo tramite stima puntuale, per il valore medio (tramite lo stimatore media campionaria):

```
> #Salerno
> mean(campione)
[1] 5.32
```

```
> #Napoli
> mean(campione2)
[1] 3.78
```

La popolazione di Salerno sembrerebbe riportare un tasso medio di automobilisti che transigono alle norme sui controlli alcolemici alla guida, superiore rispetto alla popolazione relativa all'area di Napoli, ma è effettivamente così?

Come già detto in statistica inferenziale, è preferibile stimare un parametro non noto, tramite un intervallo di confidenza, e ciò vale anche per confronti di questo tipo.

Per procedere con una stima intervallare che risponda a questo nuovo quesito, è possibile far riferimento ad una stima per la differenza dei due parametri non noti: $\lambda x - \lambda y$.

Considerando la variabile aleatoria X – Y (differenza tra la variabile X di Poisson, che descrive l'area di Salerno e Y l'equivalente per Napoli) che risulta a sua volta essere a sua volta una variabile aleatoria di Poisson.

È possibile stimare il parametro non noto λx – λy tramite stima intervallare così come è stato fatto in precedenza:

Sapendo che per via della linearità della media campionaria, è possibile esprimere la media campionaria della differenza tramite differenza delle singole medie campionarie e che è possibile esprimere la deviazione standard della differenza come somma delle singole deviazioni standard (data l'indipendenza delle due popolazioni), è possibile estrarre la variabile aleatoria:

$$\frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\lambda_1 - \lambda_2)}{\sqrt{\lambda_1/n_1 + \lambda_2/n_2}}$$

che come in precedenza, per il teorema centrale di convergenza, risulta convergere in distribuzione ad una Normale Standard (N(0,1)) risulta essere una variabile Aleatoria di Pivoting per il parametro non noto.

fissando ad esempio un grado di fiducia 1 – α pari a 0,99 e considerando tale legge di probabilità:

$$P\left(-z_{\alpha/2} < \frac{\overline{X}_{n_1} - \overline{Y}_{n_2} - (\lambda_1 - \lambda_2)}{\sqrt{\overline{X}_{n_1}/n_1 + \overline{Y}_{n_2}/n_2}} < z_{\alpha/2}\right) \simeq 1 - \alpha,$$

E calcolando i valori – $z_{a/2}$ e $z_{a/2}$ tramite la funzione QNorm, è possibile derivare che gli estremi dell'intervallo di confidenza per il parametro non noto λx – λy, risultano essere:

$$\overline{x}_{n_1} - \overline{y}_{n_2} - z_{\alpha/2}\sqrt{\frac{\overline{x}_{n_1}}{n_1} + \frac{\overline{y}_{n_2}}{n_2}} < \lambda_1 - \lambda_2 < \overline{x}_{n_1} - \overline{y}_{n_2} + z_{\alpha/2}\sqrt{\frac{\overline{x}_{n_1}}{n_1} + \frac{\overline{y}_{n_2}}{n_2}},$$

Con R è quindi possibile calcolare tale intervallo nel seguente modo:

```
> #stima intervallare per differenza popolazioni
>
> alpha <- 1 - 0.99
>
> #numero di individui campione Salerno
> n1 <- length ( campione )
> #numero di individui campione Napoli
> n2 <- length ( campione2 )
>
> #media campionaria campione Salerno # Stima puntuale lambda1
> m1 <- mean( campione )
>
> #media campionaria campione Napoli # Stima puntuale lambda2
> m2 <-mean( campione2 )
>
> rad <-sqrt (m1/n1 + m2/n2)
>
> #stima C1 limite superiore lambda1 - lambda 2
> m1 -m2 + qnorm(1-alpha/2, mean = 0, sd = 1) * rad
[1] 2.638886
>
> #stima C2 limite inferiore lambda1 - lambda 2
> m1 -m2 - qnorm (1- alpha /2, mean =0, sd =1) * rad
[1] 0.4411137
```

Da cui si ottiene un intervallo di confidenza per il parametro non noto λ1 – λ2 che è compreso tra gli estremi 0.44 e 2.64. È vero che tale risultato valida ciò che è stato anticipato dalle stime puntuali di lambda1 e lambda2 (dato che l'intervallo non è frapposto tra un estremo negativo e uno positivo), cioè che per i campioni estratti dalle due popolazioni, la popolazione di Napoli risulta avere un numero medio minore di

trasgressori al test alcolemico rispetto all'equivalente indice per la popolazione Salernitana, però questa stima permette anche di sottolineare con un grado di fiducia dello 0.99 che le due popolazioni sono differenti per il fenomeno in esame.

Verifica delle Ipotesi per una variabile di Poisson

Fino a questo momento, è stata data attenzione allo studio del fenomeno relativo ai test alcolemici in esame, supponendo nelle varie applicazioni che la popolazione di automobilisti fosse effettivamente distribuita secondo una Variabile aleatoria di Poisson, e ne sono state osservate le conseguenze tramite l'applicazione delle tecniche di stima puntuali o intervallari.

La statistica inferenziale però è anche un ottimo strumento per rispondere a domande del tipo:

- La popolazione in analisi è effettivamente distribuita secondo una data variabile, come ad esempio la variabile di Poisson?
- Il parametro non noto della variabile a quanto equivale? È maggiore o uguale di un certo valore reale, magari proprio uguale al valore che si sta considerando?

La branca della statistica inferenziale, che si occupa di rispondere a queste problematiche prende il nome di Verifica delle Ipotesi.

Tali quesiti vengono formalizzati partendo ovviamente da una popolazione che rappresenta un determinato fenomeno, da un campione casuale estratto da una popolazione, e da una variabile aleatoria che descrive tale popolazione, legata ovviamente da un parametro non noto.

Tali quesiti, prendono il nome di ipotesi statistiche, e per un parametro non noto, vengono così definite:

Un'ipotesi statistica `e un'affermazione o una congettura sul parametro non noto θ. Se l'ipotesi statistica specifica completamente f(x; θ) – funzione di distribuzione detta ipotesi semplice, altrimenti `e chiamata ipotesi composita.

Le ipotesi in generale vengono denotate con la lettera H, ed in particolare se un'ipotesi viene sottoposta a verifica essa prenderà il nome di Ipotesi Nulla H_0.

Tornando al problema in analisi, è possibile considerare nuovamente il campione casuale:

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)

Relativo a 50 prove di Poisson, che descrivono il numero di automobilisti con tasso alcolemico superiore al limite per la guida, rilevati in 50 posti di blocco per l'area urbana di Salerno.

Da questo campione casuale e la sua popolazione, è possibile porsi il seguente problema di verifica:

Supponendo che la popolazione Poissoniana inerente al fenomeno, sia descritta da una variabile aleatoria X con valore metodo non noto, si vuole verificare che il parametro non noto lambda sia effettivamente uguale a 5.

$$H_0: \lambda = 5$$

Tale ipotesi statistica, è per il problema in analisi, un'ipotesi semplice, dato che volendo verificare un'uguaglianza per il parametro non noto si descrive completamente la funzione f(x: λ) per la variabile aleatoria X relativa alla popolazione.

Volendo sottoporre a verifica quest'ipotesi, è necessario formulare un'ipotesi alternativa che appunto racchiude il dominio del parametro nel caso l'ipotesi posta a verifica si debba rivelare errata dopo la verifica.

In particolare, per il problema in analisi l'ipotesi alternativa sarà:

$$H_1 : \lambda \; != 5$$

Ed in particolare nel caso l'ipotesi dopo verifica si dovesse confermare vera allora si dirà che $H_0 : \lambda \in \Theta_0$ (dominio ipotesi nulla), altrimenti $H_1 : \vartheta \in \Theta_1$.

Nel verificare la validità di un'ipotesi posta a verifica si possono commettere due tipi di errore, detti:

- Errore di tipo I: rifiutare l'ipotesi nulla quando essa è vera.
- Errore di tipo II: accettare l'ipoteso nulla quando essa è falsa.

	Rifiutare H_0	Accettare H_0
H_0 vera	Errore del I tipo Probabilità α	Decisione esatta Probabilità $1-\alpha$
H_0 falsa	Decisione esatta Probabilità $1-\beta$	Errore del II tipo Probabilità β

Gli errori di tipo uno e di tipo due per un test statistico sono vincolati da proprietà di verificarsi alfa e beta così come riportato, ed al fine di dare validità alle ipotesi, bisogna assicurarsi di non incorrere in nessuno delle due tipologie di errore. Mantenere le probabilità al minimo in statistica inferenziale però non è possibile, quindi al fine di evitare di commettere errori, si fissa una delle due probabilità e si cerca di mantenere al minimo l'altra. In particolare, dato che è considerato più grave commettere un errore di tipo 1, si preferisce fissare la probabilità alfa, cercando di mantenere al minimo la probabilità beta.

Quello che si fa solitamente è fissare la probabilità di non commettere un errore di tipo I, fissando la probabilità alfa a dei valori standard, che appunto garantiscono 3 livelli standard per un test statistico:

- α= 0.05 // test di verifica statisticamente significatico;
- α= 0.01 // test di verifica statisticamente molto significativo;
- α= 0.001 // test di verifica statisticamente estremamente significativo;

Passo 1 – test statistico bilaterale approssimato del parametro λ

In statistica, fissando a priori il parametro alfa per l'errore di tipo 1, si possono considerare differenti tipi di test:

I test statistici sono di due tipi:

- *test bilaterali* (detti anche *test bidirezionali*);
- *test unilaterali* (detti anche *test unidirezionali*).

Un test bilaterale è il seguente

$$\begin{aligned} \mathbf{H}_0 &: \vartheta = \vartheta_0 \\ \mathbf{H}_1 &: \vartheta \neq \vartheta_0, \end{aligned}$$

mentre il *test unilaterale sinistro* e *test unilaterale destro* sono rispettivamente i seguenti

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

avendo fissato a priori un *livello di significatività* α.

Da come si può notare il primo test statistico considerato in questo capitolo, è appunto un test statistico del primo tipo, appunto test statistico bilaterale.

Come per le stime intervallari, anche i test di verifiche delle ipotesi hanno alla base la variabile aleatoria normale standard. Ed è possibile applicare tali test a variabili che non sono normali, tramite il teorema centrale di convergenza, ponendosi nella base che il campione sia numeroso (con ampiezza maggiore o uguale a 30 individui).

Considerando, una generica variabile aleatoria mean(X), è possibile standardizzare questa variabile nel seguente modo:

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

Come già detto più volte questa variabile converge in distribuzione con una variabile normale standard.

Per il parametro λ, questa variabile aleatoria assume la seguente forma:

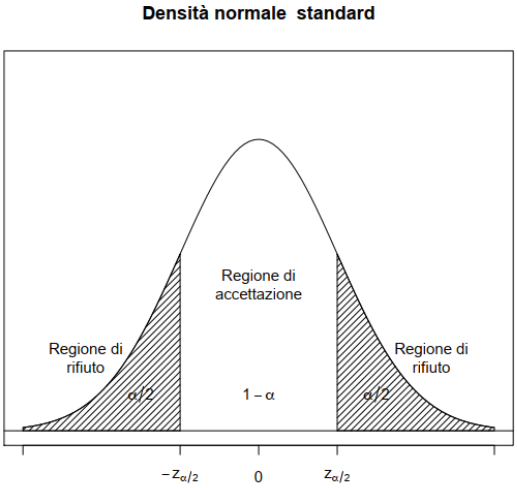
$$\frac{\overline{X}_n - \lambda}{\sqrt{\lambda/n}}$$

Data questa variabile, è possibile sostituire il valore lambda con il valore stimato nell'ipotesi nulla, ottenendo un valore statistico del tipo:

$$\frac{mean(X) - \lambda 0}{\sqrt{\lambda 0/n}}, \qquad \lambda 0 = 5$$

Tale stima è un valore osservabile, dato che dipende strettamente dal campione casuale che è stato estratto dalla popolazione.

Considerando la variabile normale standard, è possibile trovare due valori $-z_{\alpha/2}$ e $z_{\alpha/2}$ simmetrici rispetto allo 0 nella distribuzione normale standard.



In particolare, è possibile accettare o rifiutare l'ipotesi nulla tramite test bilaterale se:

- si accetti $\mathbf{H_0}$ se $\quad -z_{\alpha/2} < \dfrac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} < z_{\alpha/2}$

- si rifiuti $\mathbf{H_0}$ se $\quad \dfrac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} < -z_{\alpha/2} \quad$ oppure $\quad \dfrac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} > z_{\alpha/2}$

Considerando quindi una probabilità α di non commettere un errore di tipo I dello 0,001, è possibile determinare gli estremi della di accettazione e verificare se la stima di verifica dipendente dal campione casuale è compresa nell'intervallo;

```
> campione
 [1]  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
[33]  7  3  2  7  7  4  8  3  4  8  3  4  3  5  7  8  4  5
>
> #media campionaria
> mean <- mean(campione)
> mean
[1] 5.32
>
> #length N
> n <- length(campione)
> n
[1] 50
>
> lambda0 <- 5
>
> #stima di verifica di poisson lambda = 5
> stima <- (mean - lambda0) / (sqrt(lambda0) / sqrt(n))
> stima
[1] 1.011929
>
> #intervallo
> aplha = 0.001
>
> #Estremo inferiore regione di accettazione
> qnorm (alpha/2 ,mean =0, sd =1)
[1] -2.575829
>
> #Estremo superiore regione di accettazione
> qnorm (1-(alpha/2) ,mean =0, sd =1)
[1] 2.575829
```

Effettuando tale verifica, risulta che l'ipotesi H_0 è vera per il test bilaterale, dato che appunto la statistica risulta rientrare nell'intervallo che delimita la regione di accettazione. Essendo che la probabilità di commettere un errore di tipo uno è dello 0.001, ed il test ha avuto esito positivo è possibile affermare che il test statistico effettuato è estremamente significativo.

Spesso, però capita anche che i test statistici vengano validati anche tramite quello che viene definito come il livello di significatività osservato, noto come p-value. Il p-value, come il metodo precedente, che dipende dal campione osservato e dal test statistico considerato si basa su una statistica del test.

Il p-value è definito come la probabilità, supposta vera l'ipotesi H0, che la statistica del test assuma un valore uguale o più estremo di quello effettivamente osservato.

Essendo una probabilità il p-value è un numero compreso tra 0 e 1.

Calcolando il p-value è possibile comportarsi come segue:

- se $p > \alpha$, l'ipotesi H_0 non può essere rifiutata;

- se $p \leq \alpha$, l'ipotesi H_0 deve essere rifiutata.

Avendo già fissato il valore osservato della statistica, è possibile calcolare il p-value nel seguente modo:

$$pvalue = P(Z_n < -|z_{os}|) + P(Z_n > |z_{os}|) = 2\,P(Z_n > |z_{os}|)$$
$$= 2\Big[1 - P(Z_n \leq |z_{os}|)\Big],$$

Che nel caso del test precedentemente effettuato assume valore:

```
> #pvalue
> 2 * (1 - pnorm(abs(stima), mean = 0, sd = 1))
[1] 0.3115721
```

0.31, che è strettamente maggiore della probabilità α di errore di tipo 1 precedentemente stabilita, pari allo 0.001. Quindi anche per il criterio del p-value, l'ipotesi nulla del test bilaterale condotto $H_0 : \lambda = 5$, è da ritenersi vera sulla base del campione casuale di 50 individui estratti dalla popolazione poissoniana di automobilisti sottoposti ad alcol test.

Passo 2 – test statistici unilaterali approssimati per il parametro λ

Dal test statistico bilaterale, è emerso che l'ipotesi semplice è da ritenersi reale con una probabilità di commettere un errore di tipo 1 pari allo 0.001 quindi tale risultato è da ritenersi estremamente significativo in termini statistici per la popolazione di Poisson sotto analisi.

Tale risultato però certifica all'uguaglianza che il valore lambda è uguale a 5, in alcuni casi però è preferibile cercare di trovare un estremo superiore ed uno inferiore per il parametro stesso, e questo è anche possibile (oltre che con le stime intervallari, che ci hanno dato dei risultati certi) formulando alcune ipotesi nulle che al posto di esprimere un legame di uguaglianza tra parametro non noto, cercano di verificare un legame di maggiorazione o minorazione per un dato problema in analisi.

Supponendo di porsi un quesito diverso da quello che ha portato all'esecuzione del test bilaterale, si supponga di voler rispondere al seguente enunciato:

"Dato il problema di studio del numero di automobilisti medio, sulla base di un campione casuale estratto dalla popolazione, si ipotizza che il valore medio sia compreso tra 4 e 6.

Questo enunciato richiede di verificare la seguente relazione matematica per il parametro lambda:

$$4 \leq \lambda \leq 6$$

Dato che si vuole dimostrare la valenza di due legami di disuguaglianza, è possibile ricorrere alla formulazione di due test statistici per maggiorazioni separati che analizzino un singolo legame di disuguaglianza alla volta.

	Test 1	Test 2
Ipotesi nulla	$H0: \lambda \geq 4$	$H0: \lambda \leq 6$
Ipotesi alternativa	$H1: \lambda < 4$	$H1: \lambda > 6$

//Si noti che le ipotesi nulle formulate in questi due casi non descrivono a pieno la funzione di distribuzione della variabile aleatoria X per la popolazione, quindi si parla di ipotesi composite.

Per validare questi test statistici, viene suggerito in statistica di utilizzare dei test che prendono il nome di test unilaterali, che per distribuzioni non normali necessitano di essere applicati a grandi campioni (con N > 30 elementi), creando nuovamente così una stima che converge in distribuzione ad una normale standard (teorema centrale di convergenza).

In particolare, volendo verificare un'ipotesi come la prima formulata:

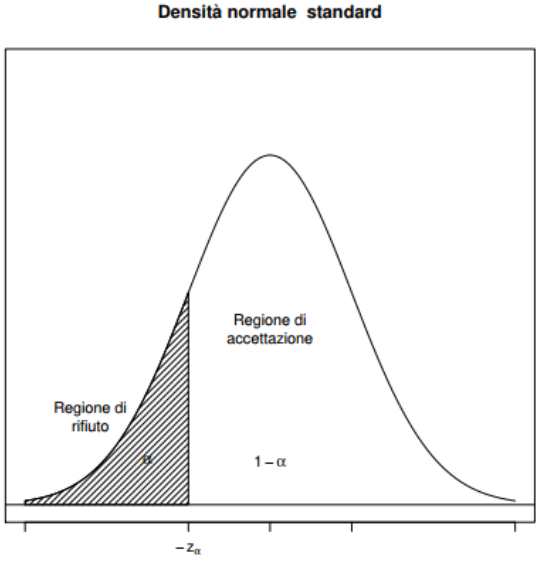
$$\mathbf{H}_0 : \mu \geq \mu_0, \qquad \mathbf{H}_1 : \mu < \mu_0$$

Si parla di test unilaterale destro, il quale è sottoposto alla seguente leggi di accettazioni e di rifiuto:

$$\text{- si accetti } \mathbf{H_0} \text{ se } \frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} < z_\alpha$$

$$\text{- si rifiuti } \mathbf{H_0} \text{ se } \frac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} > z_\alpha$$

Dove appunto la stima risulta essere la stessa che si è utilizzata per il test bilaterale (dove in questo caso il valore osservato è proprio il valore dell'estremo dell'ipotesi – 4), mentre il valore di riferimento Zα, è questa volta il limite inferiore da calcolare tramite la funzione qnorm, graficamente divide la distribuzione di una variabile normale standard nel seguente modo:



Si verifica quindi che il parametro Zα, corrisponde proprio al valore che assume la funzione di densità di una normale standard, nel caso si voglia sapere quando il valore di Z – N(0, 1) sia proprio uguale alla probabilità α di commettere un errore di tipo I.

Per questa tipologia di test risulta possibile anche applicare il criterio del PValue, il quale per il test unilaterale destro risulta essere:

$$pvalue = P(Z_n \leq z_{os}),$$

dove $z_{os} = (\overline{x}_n - \mu_0)/(\sigma/\sqrt{n})$ è la stima della statistica del test.

Volendo applicare con R il test unilaterale destro approssimato per l'ipotesi nulla $H0: \lambda \geq 4$, con una probabilità minima alfa dello 0.001 si ottiene il seguente risultato:

```
> campione
 [1]  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
[33]  7  3  2  7  7  4  8  3  4  8  3  4  3  5  7  8  4  5
>
> #media campionaria
> mean <- mean(campione)
> mean
[1] 5.32
>
> #length N
> n <- length(campione)
> n
[1] 50
>
> lambda0 <- 4
>
> #stima di verifica di poisson lambda >= 4
> stima <- (mean - lambda0) / (sqrt(lambda0) / sqrt(n))
> stima
[1] 4.666905
>
> #calcolo regione di accettazione
> aplha = 0.001
>
> #Estremo superiore regione di rifiuto
> qnorm ((alpha) ,mean =0, sd =1)
[1] -2.326348
>
> #PValue
> pnorm(stima, mean = 0, sd = 1)
[1] 0.9999985
```

Essendo che il valore di stima è di molto superiore all'estremo della regione di accettazione, ed anche il valore di PValue è molto più grande della probabilità alfa di 0.001 di commettere un errore di tipo 1. L'ipotesi $H0: \lambda \geq 4$ è senz'altro da ritenersi esatta, sottolineando che è stata validata da un test statistico estremamente attendibile.

In maniera speculare è possibile procedere anche per la validazione dell'ipotesi:

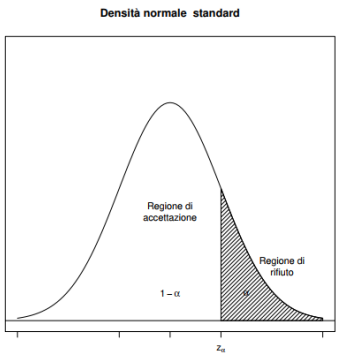
$$H0: \lambda \leq 6$$

La quale può essere validata tramite un test statistico totalmente speculare al test unilaterale destro, chiamato appunto test unilaterale sinistro approssimato:

Questo test statistico è soggetto alle seguenti leggi di accettazione e di rifiuto:

- si accetti $\mathbf{H_0}$ se $\dfrac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} < z_\alpha$

- si rifiuti $\mathbf{H_0}$ se $\dfrac{\overline{x}_n - \mu_0}{\sigma_0/\sqrt{n}} > z_\alpha$



Si noti che date le relazioni, la regione di accettazione e di rifiuto risultano simmetricamente invertite rispetto al caso precedente, quindi in questo caso il valore estremo Zα, assume il ruolo di estremo inferiore della

regione di rifiuto, e dovrà essere calcolato tramite la funzione qnorm come il valore che la densità di probabilità assume quando appunto il valore di una normale standard Z – N(0, 1), risulta essere <= di 1 – α.

Per questa tipologia di test cambia anche in modo speculare che il calcolo per il p-value che in questo caso sarà del tipo:

$$pvalue = P(Z_n > z_{os}) = 1 - P(Z_n \leq z_{os})$$

Con Z osservato valore di stima (calcolato sempre sulla base del valore estremo dell'ipotesi $\lambda_0 = 6$).

Effettuando il test unilaterale sinistro con R per l'ipotesi nulla

$$H0: \lambda \leq 6$$

con una probabilità alfa di commettere un errore di tipo 1 dello 0.001, sulla base del campione X estratto dalla popolazione, si osserva che:

```
> #test unilaterale sinistro approssimato H0 = lambda <= 6
> campione
 [1]  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
[33]  7  3  2  7  7  4  8  3  4  8  3  4  3  5  7  8  4  5
>
> #media campionaria
> mean <- mean(campione)
> mean
[1] 5.32
>
> #length N
> n <- length(campione)
> n
[1] 50
>
> lambda0 <- 6
>
> #stima di verifica di poisson lambda <= 6
> stima <- (mean - lambda0) / (sqrt(lambda0) / sqrt(n))
> stima
[1] -1.962991
>
> #calcolo
> aplha = 0.001
>
> #Estremo inferiore regione di rifiuto
> qnorm (1-(alpha) ,mean =0, sd =1)
[1] 2.326348
>
> #PValue
> 1 - pnorm(stima, mean = 0, sd = 1)
[1] 0.9751764
```

Il valore di stima ricade all'interno della regione di accettazione, visto che esso è minore rispetto al limite inferiore della regione di rifiuto, ed anche il valore di P Value è di molto superiore alla probabilità di commettere un errore di tipo I. Anche in questo caso quindi l'ipotesi H0 è da ritenersi vera sulla base di un test unilaterale sinistro estremamente attendibile.

Concludendo quindi è possibile affermare che sulla base dei 3 test statistici eseguiti, il valore medio lambda per la popolazione Poissoniana di automobilisti sottoposti a test alcolemico, si attesta in un intervallo compreso tra 4 e 6 con un valore idealmente preciso pari a 5, sulla base del campione X di 50 posti di blocco indipendenti estratto dalla popolazione.

Passo 3 – test statistico tramite criterio del chi-quadrato bilaterale

Con la verifica delle ipotesi, è possibile affermare senz'altro che le ipotesi fin ora condotte sono state ampliamente validate, però i test effettuati, come d'altronde anche le stime puntuali e intervallari che sono state effettuate, hanno sempre avuto come ipotesi basilare che il campione o i campioni analizzati per il fenomeno analizzato, fossero stati estratti da una popolazione che risulta essere descritta da una variabile aleatoria discreta di Poisson.

Però solitamente quando si ha a che fare con un problema di inferenza statistica, una delle prime ipotesi che deve essere validata è appunto accertarsi che la popolazione da cui è stato estratto un campione, sia effettivamente rappresentata da una variabile aleatoria con annessa funzione di probabilità che il decisore ritiene più adatta per il fenomeno che si sta analizzando.

La statistica inferenziale mette a disposizione diversi criteri per verificare questo tipo di ipotesi e uno dei più utilizzati è chiamato appunto criterio di verifica delle ipotesi del chi-quadrato, detto anche test del buon adattamento.

Al fine di applicare questo test per il documento in analisi, è necessario ricapitolare brevemente quello che è stato fatto fino a questo punto, in modo tale da avere ben in mente cosa si vuole verificare.

All'inizio di questo documento, è stato specificata l'intenzione di studiare tramite alcuni strumenti di interferenza statistica questa problematica legata al mondo dell'alcolismo:

"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?".

Fin da subito è stato supposto che la variabile aleatoria di Poisson, fosse lo strumento migliore per descrivere la popolazione di automobilisti coinvolta nell'analisi, data la natura di conteggio "rara" intrinseca al fenomeno.

Sulla base di quest'ipotesi, sono stati affrontati alcuni problemi come stimare il parametro lambda tramite stimatori o intervalli di confidenza, dati diversi campioni casuali si è visto quanto il valor medio stimato impattasse nelle probabilità di trovare un dato numero di trasgressori al test alcolemico analizzando appunto i grafici creati con il sistema R. Sono state effettuati poi confronti su campioni estratti da popolazioni diverse e non ultimi i test di verifica delle ipotesi sul parametro lambda per accurare ancora di più i parametri osservati con un'ottica leggermente differente.

Ma cosa effettivamente da la conferma che il fenomeno relativo ai test alcolemici sia effettivamente descritto tramite una variabile aleatoria di Poisson?

Sicuramente questo è un punto a cui è doveroso dare una risposta, altrimenti chiunque potrebbe ribattere su tutte le deduzioni effettuate dicendo che un'informazione così importante per tutto il lavoro fatto non è nient'altro che un'ipotesi non verificata.

Al fine di validare tutto il lavoro effettuato allora, è necessario verificare un'ultima ipotesi:

Considerando una popolazione di automobilisti sottoposti a controllo, dalla quale si estrae tale campione casuale X di 75 elementi, che descrivono il numero di persone che hanno superato il limite di controllo al test alcolemico per l'idoneità alla guida ad un singolo posto di blocco non correlato agli altri posti di blocco coinvolti nella creazione del campione:

Camp = (4, 2, 3, 9, 3, 0, 5, 2, 6 11, 7, 5 12, 1, 6, 5, 6, 5, 4, 4, 4, 3, 8, 8, 4, 4, 4, 5, 2, 6, 5, 7, 4, 3, 4, 7, 4, 4, 3, 4 12, 5, 6, 4, 3, 6, 6, 5 11, 5, 8, 1, 5, 2, 5, 4, 5, 3, 6, 7, 5, 3, 6, 7 6, 6, 4, 6, 4, 7, 5, 4, 4, 4, 6)

Per la variabile aleatoria X, che descrive la popolazione si vuole verificare la seguente ipotesi:

H_0: X è una variabile aleatoria di Poisson, con funzione di Poisson di parametro λ, con funzione di probabilità:

$$p_X(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad (x = 0, 1, \ldots)$$

L'ipotesi alternativa sarà di conseguenza:

H_1: X non è una variabile aleatoria di Poisson di parametro λ.

La prima cosa da fare per verificare l'ipotesi H0 tramite il criterio del chi-quadrato, è suddividere l'insieme dei valori che la variabile aleatoria X può assumere in r sottoinsiemi I1, I2, . . . , Ir (siano esse classi o categorie) in modo che risulti essere uguale a pi la probabilità che, secondo la distribuzione ipotizzata, la variabile aleatoria assuma un valore appartenente a Ii:

in altri termini:

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

I valori che una variabile aleatoria di Poisson può assumere intervallano da 0 a + infinito, e osservando il campione estratto dalla popolazione, si è deciso di dividere i possibili valori in 4 intervalli così formati:

```
#I1 = [0, 2]
#I2 = (2, 5]
#I3 = (5, 8]
#I4 = (8, +inf)
```

Per ogni intervallo queste probabilità pi assumono valori:

```
> p <- numeric(4)
> p[1] <- dpois(0, stimalambda) + dpois(1 , stimalambda) + dpois(2, stimalambda)
> p[2] <- dpois(3, stimalambda) + dpois(4 , stimalambda) + dpois(5, stimalambda)
> p[3] <- dpois(6, stimalambda) + dpois(7 , stimalambda) + dpois(8, stimalambda)
> p[4] <- 1 - p[1] - p[2] - p[3]
>
> p
[1] 0.12023133 0.48637195 0.32176583 0.07163089
```

Ed inoltre considerando il campione casuale si ottiene che per ogni intervallo ricadono i seguenti valori:

```
> r<-4
> nint <- numeric (r)
> nint [1] <-length (which((campTestChiQuadr >= 0) & (campTestChiQuadr <= 2)))
> nint [2] <-length (which((campTestChiQuadr > 2) & (campTestChiQuadr <= 5)))
> nint [3] <-length (which((campTestChiQuadr > 5) & (campTestChiQuadr <= 8)))
> nint [4] <-length (which((campTestChiQuadr > 8)))
> nint
[1]  7 41 22  5
```

Sulla base di questi valori per il criterio del chi quadrato, viene creata la seguente stima:

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

Dove appunto N è la variabile aleatoria che descrive il numero di elementi del campione casuale che cadono in un dato intervallo i, ed n e pi, risultano essere i valori appena calcolati. Si dimostra che la variabile risultante è distribuita con una legge chi-quadrato r - k – 1, dove r e k sono effettivamente il numero di intervalli fissato dal decisore e k è il numero di parametri non noti a cui è legata la variabile aleatoria citata nell'ipotesi opportunamente sostituiti da stime.

Per il test che si sta per condurre, r e k sono rispettivamente 4 e 1.

Per applicare il test chi-quadrato, è necessario che sia vera questa relazione:

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

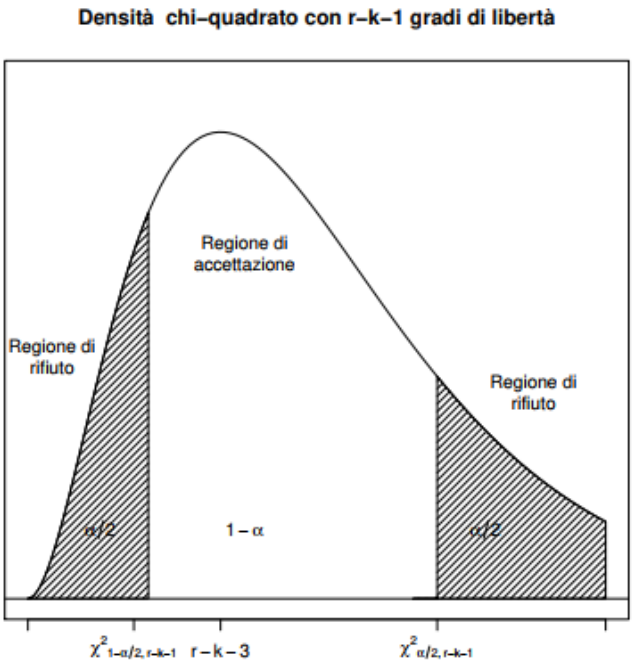
Al fine di ottenere una buona approssimazione per il test effettuato.

Per il campione in analisi quest'ipotesi è rispettata:

```
> min (n*p[1], n*p[2],n*p[3],n*p[4])
[1] 5.372317
```

Si può quindi procedere con l'effettuare il test chi-quadrato per l'ipotesi H_0 formulata.

Considerando quindi una probabilità α di non commettere un errore di tipo I dello 0,05, è possibile calcolare 2 valori χ2 α/2,r-k-1 e χ2 1-α/2,r-k-1, che andranno a dividere (come per il test bilaterale con la distribuzione normale), la distribuzione chi-quadrato con r-k-1 gradi di libertà in una regione di rifiuto e in una di accettazione:



Tali valori sono soluzioni delle equazioni:

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

Che con R possono essere calcolati con la funzione qchsq per il calcolo dei quantili per una distribuzione di frequenza chi-quadrato con r-k-1 gradi di libertà.

```
> #intervalli regione di accettazione
> r <- 4
> k <- 1
>
> alpha <- 0.05
> qchisq ( alpha /2,df= r -k -1)
[1] 0.05063562
> qchisq (1- alpha /2,df= r -k -1)
[1] 7.377759
```

Ora se (come per il test bilaterale approssimato), la stima definita per il test chi-quadrato risulta all'interno della regione di accettazione l'ipotesi formulata H_0, deve essere accettata.

Proposizione 14.1 *Per un campione sufficientemente numeroso di ampiezza n, il test chi–quadrato bilaterale di misura α è il seguente:*

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

Per il test attuale risulta che, la stima calcolata sulla base del campione casuale risulta essere pari a:

```
> #stima chiquadro
>
> chi2 <- sum ((( nint -n*p)/sqrt (n*p)) ^2)
> chi2
[1] 1.22615
```

Quindi all'interno della regione di accettazione definita con grado di significatività dello 0.05, quindi tramite un test di verifica significativo. Per tale ragione l'ipotesi H0 deve essere accettata.

Si noti anche che la stima ricade anche nella regione di accettazione calcolata con alpha pari a 0.001.

```
> #intervalli regione di accettazione
> r <- 4
> k <- 1
>
> alpha <- 0.001
> qchisq ( alpha /2,df= r -k -1)
[1] 0.00100025
> qchisq (1- alpha /2,df= r -k -1)
[1] 15.2018
```

Quindi anche per un test chi-quadrato estremamente significativo è possibile validare che la popolazione che descrive il numero di automobilisti che vengono trovati alla guida con un tasso alcolemico superiore alla norma, sulla base di un campione di 75 unità estratte, è descritta al meglio tramite variabile aleatoria di Poisson di parametro λ, validando così anche tutte le precedenti ipotesi e stime validate in questo progetto.

Un ringraziamento speciale alla professoressa

Amelia Giuseppina Nobile che con tanta dedizione

mi ha guidato in ogni passo di questo progetto

e che ha fatto del corso di Statistica e Analisi dei Dati 2020-2021

uno dei momenti più belli che ricorderò

della mia carriera accademica.

Nella speranza che questo mio lavoro, se pur piccolo,

sia significativo per il messaggio che a voluto lanciare,

mi auguro che questa sia stata solo la prima

di tante belle applicazioni

che il mondo della statistica
ha da offrirmi.

Carmine Ferrara

mat.0 522500990

Baronissi, 11/12/2020