Prove ripetute, distribuzione binomiale e di Poisson

Gruppo M04: Alberto Malvezzi, Matteo Bosio, Alberto Appoloni*
Università di Trento - Dipartimento di Fisica
(Dated: 8 aprile 2025)

Dopo aver generato un simulatore per il lancio dei dadi ci poniamo l'obiettivo di effettuare una misura (frequentista) della probabilità con cui certi eventi accadono, con un'incertezza predeterminata $\delta_p = 0.01$ dove $\delta_p = k\sigma_p$ dove k è il fattore di copertura scelto da noi ed uguale a 2. Abbiamo scelto 3 coppie di variabili aleatorie e abbiamo studiato la loro relazione tramite la covarianza campionaria confrontandola con quella teorica. I grafici di dispersione hanno confermato risultati coerenti tra di essi.

Successivamente il confronto tra istogramma campionario e distribuzione binomiale di un nuovo evento, con probabilità fissata a priori, fornisce dei risultati in accordo tra di loro, con un'ulteriore verifica di questo fatto grazie al test del chi quadrato.

L'ultima parte pone a confronto una distribuzione gaussiana con una distribuzione poissoniana e verifichiamo che per un certo valore M (valore medio) si ha una ipotetica uguaglianza tra i due andamenti, nel nostro caso M>40.

* alberto.malvezzi@studenti.unitn.it, matteo.bosio@studenti.unitn.it, alberto.appoloni@studenti.unitn.it

I. INTRODUZIONE

L'esperienza ha come obiettivo l'analisi, sotto vari aspetti, degli esiti di prove ripetute indipendenti tra loro, ponendoli a confronto con i valori teorici forniti dalla teoria delle probabilità e dalla statistica. Otteniamo i dati tramite simulazioni virtuali.

Nella prima sezione abbiamo calcolato il numero minimo di lanci di campioni da 50 dadi, divisi in gruppi di forma diversa, da effettuare per calcolare la probabilità frequentistica di due eventi con una tolleranza fissata.
Nella seconda sezione costruiamo coppie di variabili aleatorie utilizzando i conteggi degli esiti della prima sezione e ne studiamo la correlazione sia a livello grafico, con i grafici a dispersione, sia quantitativamente, calcolando i rispettivi indici di Pearson.
Nella terza sezione consideriamo i valori assunti dalla variabile aleatoria che rappresenta il numero di conteggi di esiti, legati al lancio di tre dadi cubici. Scegliamo il numero di lanci in modo da avere almeno 10 conteggi per ogni esito. Eseguendo il test del χ^2 sugli istogrammi su cui raccoglieremo i dati, verificheremo se i conteggi ottenuti dai lanci sono risultati compatibili con quelli teorici.
Nella quarta sezione abbiamo generato una distribuzione di valori casuali con media nota per verificare alcune proprietà della distribuzione di Poisson come, ad esempio, il confronto con una distribuzione gaussiana.

II. MISURA DELLA PROBABILITÀ DI EVENTI

Dopo aver generato un simulatore per il lancio dei dadi ci poniamo l'obiettivo di effettuare una misura (frequentista) della probabilità con cui certi eventi accadono, con una incertezza predeterminata $\delta_p = 0.01$ dove $\delta_p = k\sigma_p$ e k=2 è il fattore di copertura scelto, quindi il 93% delle nostre misure cade nell'intervallo.
Gli eventi presi in considerazione sono:

- E_1: "esce il numero 1 in un lancio di un dado";
- E_2: "esce un numero pari in un lancio di un dado".

Dopo aver calcolato le probabilità teoriche $p(E_1)$ e $p(E_2)$ ce ne serviamo per calcolare il numero minimo di N lanci da effettuare per avere una tolleranza $\delta_p = 0.01$. Per calcolarle usiamo

$$\begin{aligned} p(E_i) &= p(E_i \cap E_{d12}) + p(E_i \cap E_{d20}) = \\ &= p(E_i|E_{d12})p(E_{d12}) + p(E_i|E_{d20})p(E_{d20}) \end{aligned} \tag{1}$$

dove E_{d12} e E_{d20} indicano gli esiti rispettivamente del lancio del dado a 12 facce e del dado a 20 facce. Ricaviamo le probabilità $p(E_1) = 0.07$ e $p(E_2) = 0.5$.
Usando la probabilità degli eventi complementari, riusciamo ad ottenere N dalla formula

$$N_i = p(E_i)p(\overline{E_i})\left(\frac{k}{\delta_p}\right)^2 \tag{2}$$

da cui $N_1 = 2604$ e $N_2 = 10000$ Siccome facciamo 50 lanci alla volta ci basterà lanciare i dadi contemporaneamente $N = 200$ volte. I valori di $p(E_i)$ e σ_i sono dati da

$$p(E_i) = \frac{n_i}{N_2}$$

$$\sigma_i = \sqrt{\frac{p(E_i)q(E_i)}{N_2}}$$

otteniamo

$$p(E_1) = 0.0670 \quad \sigma[p(E_1)] = 0.003$$
$$p(E_2) = 0.5073 \quad \sigma[p(E_2)] = 0.005$$

e volendo verificare il valore di tolleranza δ_p=0.01 consideriamo la differenza in modulo tra valore calcolato e valore previsto:

$$|R_1| = |p(E_1) - 0.070| = 3 \times 10^{-3} \leq 0.01 \tag{3}$$
$$|R_2| = |p(E_2) - 0.500| = 7.3 \times 10^{-3} \leq 0.01 \tag{4}$$

Quindi entrambi i valori sono compatibili con i valori teorici scelti a priori.

III. COVARIANZA CAMPIONARIA DI CONTEGGI ALEATORI

In questa parte dell'esperienza ci siamo occupati della stima della covarianza campionaria di 3 coppie di variabili aleatorie, le quali fanno riferimento ai risultati degli N lanci di 50 dadi raccolti in precedenza. Studiamo la correlazione di ciascuna coppia: facciamo riferimento al modello teorico di distribuzione di probabilità binomiale insieme all'utilizzo dei grafici di dispersione. Per la probabilità dei singoli eventi è stata utilizzata la definizione classica di probabilità.
Su un totale di 10000 conteggi ottenuti in precedenza abbiamo costruito 3 coppie di variabili aleatorie:
- (κ, υ)=("esce pari nei d12" ,"esce pari nei d20").
-$(\kappa + \upsilon, \kappa)$=("esce pari in qualunque dado" ,"esce pari nei d12").
-$(\kappa + \upsilon, \upsilon)$=("esce pari in qualunque dado" ,"esce pari nei d20").

Otteniamo prima di tutto una valutazione quantitativa sul grado di correlazione delle nostre coppie di variabili aleatorie tramite la covarianza campionaria

Introduciamo l'indice di Pearson che pone in rapporto il nostro valore campionario della covarianza con il rispettivo valore massimo teorico max$[\sigma_{xy}] = \sigma[x]\sigma[y]$.
Se le variabili sono correlate l'una con l'altra allora si avvicineranno al valore limite 1 (in modulo), invece se sono indipendenti tra di loro l'indice di Pearson tenderà a 0 (in modulo).

Coppia(x,y)	Covarianza σ_{xy}	$max[\sigma_{xy}]$	indice di Pearson p_{xy}
κ, υ	0.10	6.1	0.02
$\kappa + \upsilon, \kappa$	8.3	9.7	0.86
$\kappa + \upsilon, \upsilon$	5.2	7.9	0.66

Abbiamo poi utilizzato un metodo qualitativo per la verifica della correlazione delle 3 variabili aleatorie in esame. Tale metodo utilizza i grafici a dispersione ("Scatter plots") che mostrano la dispersione dei dati sulla base del conteggio delle variabili aleatorie. Un punto del grafico individua una coppia di esiti per i due eventi ma non da informazioni sul numero di esiti identici, per questo abbiamo impostato i grafici in modo tale che i valori più frequenti presentino un colore più scuro.
Essendo le speranze matematiche ricavate dal modello teorico della distribuzione binomiale, l'evento "esce pari" possiede eguale probabilità di 1/2 in tutti e 3 gli eventi, di conseguenza:

- m[κ]=30 $\cdot$ p("esce pari")=15
- m[υ]=20 $\cdot$ p("esce pari")=10
- m[$\kappa + \upsilon$]=50 $\cdot$ p("esce pari")=25

In [figura 1] possiamo osservare che la dispersione è uniforme attorno alle rette y=10 e x=15. Il grafico suggerisce che non è presente alcuna relazione tra le coppie di variabili aleatorie, in accordo con i valori dell'indice di Pearson.

In [figura 2] osserviamo che la dispersione, attorno ai punti delimitati dalle rette y=25 e x=15, non è omogenea infatti ad un valore minore o maggiore di m[κ] ne corrisponde uno minore o maggiore di m[$\kappa+\upsilon$]. Il grafico suggerisce una relazione lineare tra le coppie di variabili aleatorie , in accordo con i valori dell'indice di Pearson.

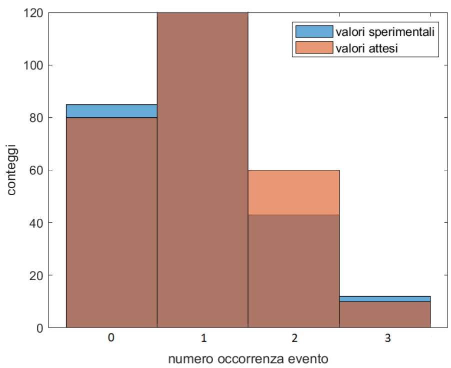
In [figura 3] osserviamo, come in precedenza, che la dispersione attorno ai punti delimitati dalle rette y=25 e x=10 non è omogenea infatti ad un valore minore o maggiore di m[υ] ne corrisponde uno minore o maggiore di m[$\kappa + \upsilon$.Il grafico suggerisce una relazione lineare tra le coppie di variabili aleatorie , in accordo con i valori dell'indice di Pearson.

IV. CONFRONTO TRA ISTOGRAMMA MISURATO E DISTRIBUZIONE BINOMIALE

Si vuole configurare una serie di prove ripetute con un set di tre dadi d6 in modo che la probabilità dell'evento considerato sia $p_k = 1/3$ e i valori di conteggio siano 0, 1, 2 o 3: si sceglie come evento "Quanti numeri divisibili per 3 escono nel lancio di tre dadi". Si deve anche tenere conto del fatto che si vuole avere per ogni bin un conteggio di almeno $< n_k >= 10$ elementi. Consideriamo la distribuzione binomiale

$$P(n_k, N) = \binom{N}{n_k} \left(\frac{1}{3}\right)^{n_k} \left(\frac{2}{3}\right)^{N-n_k} \tag{5}$$

Dove si usa $N = 3$ e distinguiamo i vari casi n_k con $k = 0, ..., 3$. Allora



$$P(0,3) = \left(\frac{8}{27}\right) \quad P(1,3) = \left(\frac{4}{9}\right)$$

$$P(2,3) = \left(\frac{2}{9}\right) \quad P(3,3) = \left(\frac{1}{27}\right)$$

Siccome vogliamo per ogni k $< n_k >= 10$ e $< n_k >= Np_k$, usando la probabilità minore , $P(3,3)$,otteniamo $N = 270$. Quindi dobbiamo effettuare almeno 270 lanci per avere una popolazione per bin di almeno 10 elementi. Confrontiamo i conteggi ottenuti dai lanci casuali con la previsione teorica nell'istogramma qui a fianco.

IV.1. Test del chi quadrato

Si vuole verificare la concordanza tra i valori ottenuti sperimentalmente e i valori attesi dal modello, usando il test del chi quadrato. Ci aspetteremmo un chi-quadrato di 3 a causa della condizione di normalizzazione, che vincola il numero di conteggi di uno dei quattro bin, tuttavia otteniamo

$$\chi^2 = \sum_{k=1}^{4} \frac{(\frac{<n_k>}{N} - p_k)^2}{(\sigma[< p_k >])^2} = \sum_{k=1}^{4} \frac{(< n_k > -Np_k)^2}{(\sigma[< n_k >])^2} = \sum_{k=1}^{4} \frac{(< n_k > -m[< p_k >])^2}{(\sigma[< n_k >])^2} = 1.1 \qquad (6)$$

$$\chi^2/3 = 1.1/3 \qquad (7)$$

Notiamo che il χ^2 è prossimo al valore di aspettazione uguale 3, quindi possiamo considerare l'istogramma compatibile con la predizione teorica fornita dalla distribuzione binomiale.

V. STATISTICA POISSONIANA E LIMITE GAUSSIANO

V.1. Procedura operativa

Attraverso la funzione Matlab "poissrnd(mean)" si vuole verificare la convergenza alla media e al limite gaussiano all'aumentare del valore medio della distribuzione di Poisson.

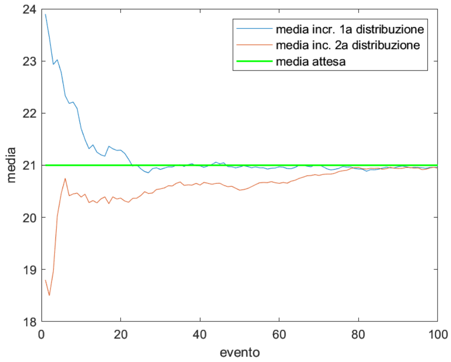
- Per generare un campione C di 1000 valori si è creata una matrice formata da 10 righe e 100 colonne contenente valori con media 21.
- Successivamente si è generato un nuovo vettore A 1x100 formato dalle medie μ_j dei 10 valori contenuti nella i-esima colonna della matrice generata in precedenza.
- Per confrontare i valori ottenuti μ_j con il valor medio vero della distribuzione si è utilizzato un istogramma.
- Attraverso un codice si è creato un nuovo vettore V 1x100 contenente la media incrementale dei valori contenuti nel vettore A.

In un secondo momento si è ripetuto l'esercizio descritto in precedenza utilizzando una nuova distribuzione di Poisson con medesima media e uguale numero di valori.
Si sono poi riportate in un nuovo grafico le medie incrementali delle due distribuzioni al fine di confrontarne la convergenza al valore medio.

Allo scopo di verificare il limite gaussiano della distribuzione di Poisson, all'aumentare del valore medio M di quest'ultima, si è prodotto un campione molto consistente di eventi poissoniani (una matrice 1x10^4) e si sono riportati i valori in un istogramma.
Tale grafico è stato poi confrontato con una distribuzione gaussiana di media M e varianza M.
Infine, si è osservata la differenza tra i due χ^2 delle due distribuzioni.

V.2. Analisi dati

Dall'analisi del vettore A, composto dalle medie di 10 eventi poissoniani estratti da un campione di 1000 valori, si è dedotto che le oscillazioni variano tra un valore massimo di 24.4 e un valore minimo di 17.6.
Tale risultato corrobora le aspettative teoriche, ovvero la convergenza al valor medio M di sottogruppi della distribuzione poissoniana è più veloce. In appendice si riporta un istogramma [Figura 4] rappresentante il confronto tra le medie ottenute e il valor medio M.
Si è poi generata una nuova distribuzione poissoniana con medesima grandezza e valor medio M e se n'è calcolata la media incrementale generando un nuovo vettore V 1x100.



Dal grafico proposto a lato si può notare come i vari elementi del vettore V convergono al valor medio di aspettazione M e alla media incrementale di una nuova distribuzione poissoniana formata dai medesimi dati della prima. Come ultima analisi si vuole osservare come una distribuzione di media M è approssimata da una distribuzione gaussiana di media M e varianza M. Dal grafico riportato in appendice si può notare come la poissoniana converga ad una curva gaussiana al crescere del valor medio M [Figura 5]. Invece al diminuire di M la differenza tra le due curve diventa sempre più rilevante, lo si può notare dal grafico [Figura 6]. Infine, per valutare a livello quantitativo la discrepanza tra le due distribuzioni si è usato il test del χ^2. Si è scelto un grado di accettazione valutato a priori e si è misurato la differenza tra i due χ^2. Il χ^2 della poissoniana ,quando il numero del campione è uguale a 10000 e il valore medio è M=10, risulta uguale a 27 mentre quello della gaussiana risulta 250. A parità del numero del campione ma all'aumentare di M, ad esempio uguale a 100, la differenza tra i due χ^2 tende a ridursi: $\chi^2_{gauss} = 87$, $\chi^2_{poiss} = 72$. Si vuole trovare il valore M per il quale la descrizione poissoniana possa essere considerata equivalente a quella gaussiana. Per farlo si può imporre che $\frac{\chi^2_{gauss}}{\chi^2_{poiss}} < 2$, ovvero un χ^2 non sia multiplo di un altro. Perchè questo accada si è trovato sperimentalmente che M deve essere maggiore di 40 [Figura 7].

VI. CONCLUSIONI

Nella prima parte dell'esperienza le probabilità teoriche calcolate per mezzo dell'interpretazione classica sono state confrontate con i parametri statistici riassuntivi delle misurazioni raccolte. I due valori per entrambi gli eventi sono risultati compatibili entro un fattore di copertura K = 2.

Successivamente, utilizzando le misurazioni effettuate nella prima parte, abbiamo studiato eventuali correlazioni tra coppie di variabili aleatorie, formate sulla base degli eventi: "esce pari nei d12", "esce pari nei d20" e "esce pari nei 50 dadi". Il grado di correlazione è stato valutato quantitativamente attraverso l'indice di Pearson e qualitativamente attraverso l'utilizzo di grafici a dispersione. Abbiamo quindi riscontrato una correlazione debole tra gli eventi "esce pari nei d12" e "esce pari nei d20", con indice di Pearson $p_{\kappa,\upsilon} = 0.02$. Al contrario gli eventi "esce pari nei d12" e "esce pari nei d20", accoppiati con "esce pari nei 50 dadi", sono risultati essere eventi caratterizzati rispettivamente da una correlazione moderata e una correlazione forte. Gli indici di Pearson associati alle due coppie di variabili aleatorie sono rispettivamente: $p_{\kappa+\upsilon,\upsilon} = 0.66$ e $p_{\kappa+\upsilon,\kappa} = 0.86$.

La terza parte dell'esperienza è stata dedicata al confronto tra un istogramma sperimentale e la corrispondente distribuzione binomiale teorica. Da un punto di vista qualitativo l'istogramma sperimentale sembra essere ben approssimato dalla distribuzione binomiale teorica. Un'analisi quantitativa, per mezzo del test del χ^2, ci ha permesso di concludere che i due andamenti sono tra loro compatibili, ovvero che il modello teorico utilizzato approssima bene i dati raccolti con un tale numero di ripetizioni.

Infine l'ultima parte della relazione è stata dedicata all'analisi di una distribuzione poissoniana e al confronto con la rispettiva distribuzione gaussiana. Si è verificato che la distribuzione poissoniana all'aumentare del valor medio M converge a una distribuzione gaussiana di media M e varianza M. Per dare una descrizione quantitativa di tale discrepanza si è utilizzato il test del χ^2. La differenza tra le due curve si può considerare irrilevante quando M è maggiore di 40.

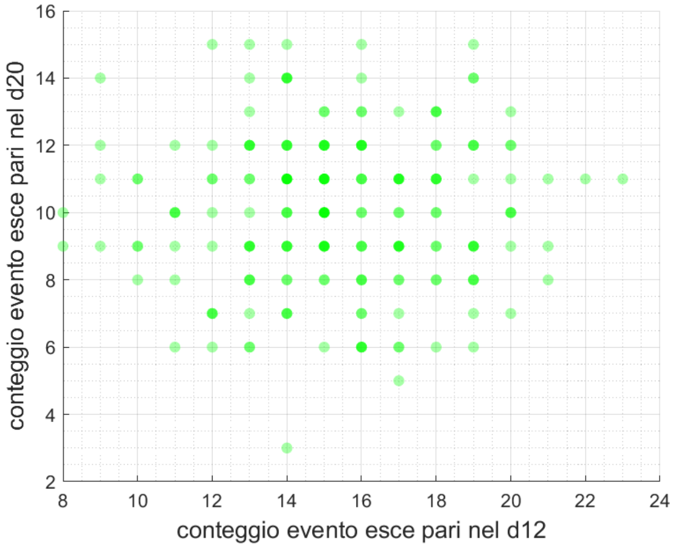


Figura 1: scatter tra d20 e d12

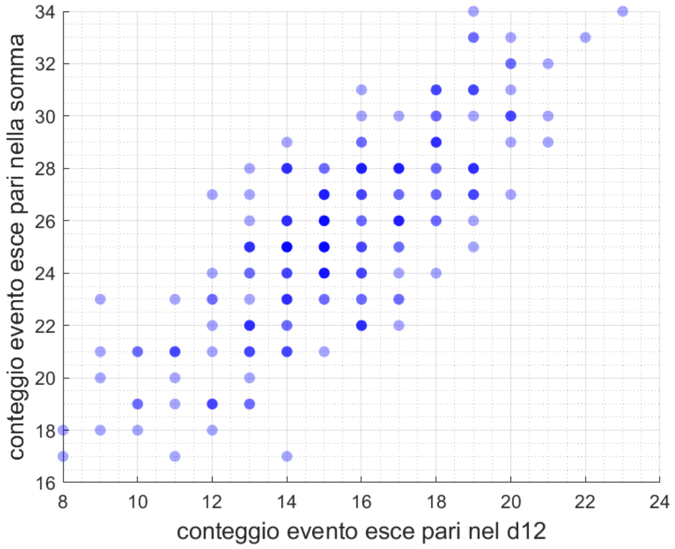


Figura 2: scatter tra somma dadi e d12

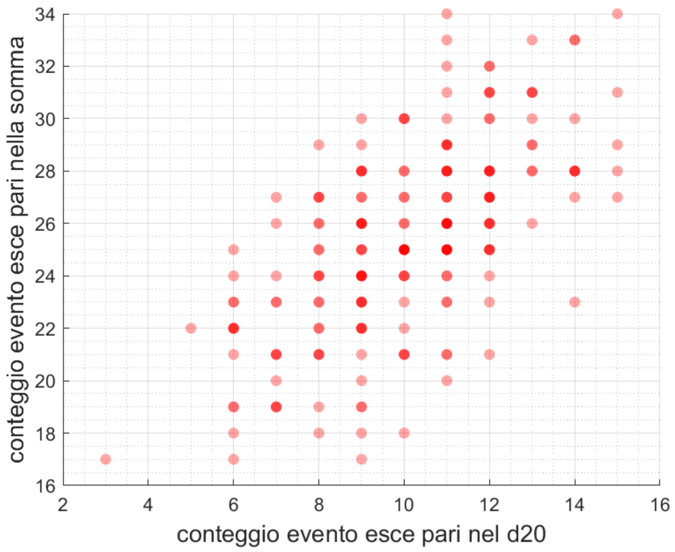
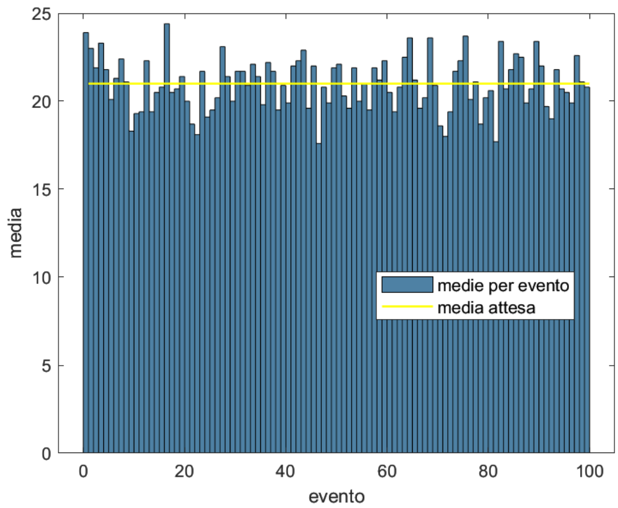


Figura 3: scatter tra somma dadi e d20

Figura 4



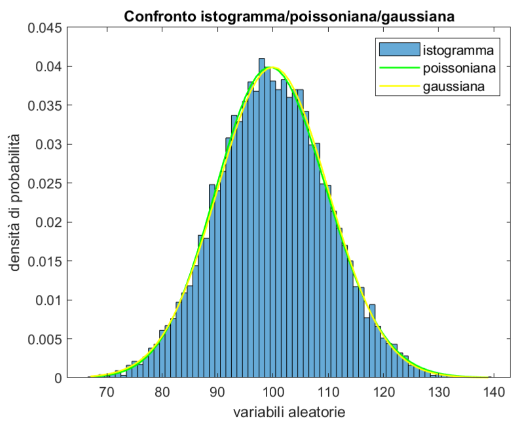


Figura 5: M=100

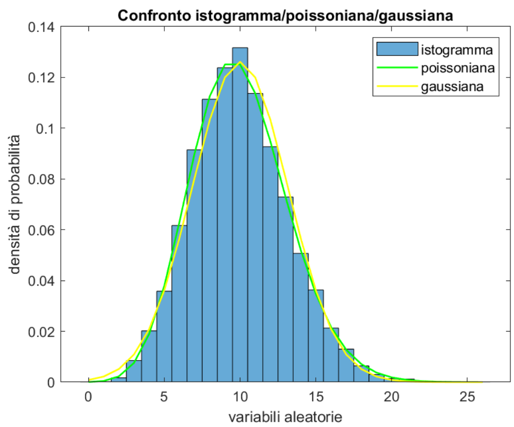


Figura 6: M=10

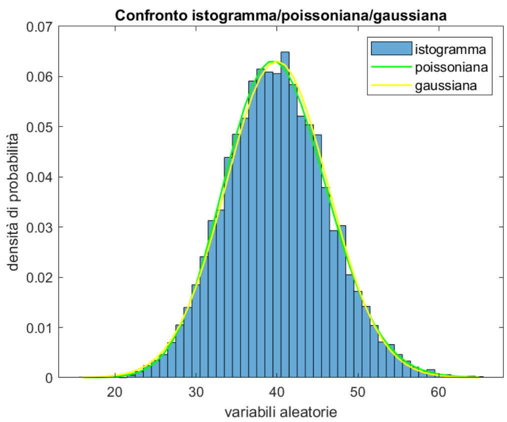


Figura 7: M=40