APPENDICE A

ESERCIZI ULTERIORI

In queste pagine ci sono esercizi che potrebbero essere in un esame. Chiaramente l'esame non sarà necessariamente di questa forma, con questo numero di esercizi eccetera. In particolare val la pena notare che le prime proposte non contengono esercizi di Statistica.

A.1. UN POSSIBILE ESAME (PROBABILITÀ)

Problema 1. Siano X e Y due variabili aleatorie esponenziali indipendenti di medie rispettivamente uguali a 1 e 0.2857143.

1. Determinare la seguente probabilità: $P(X < 0.57 \leq Y)$.
2. Determinare la seguente probabilità: $P(3.5 \cdot Y \leq X)$.
3. Determinare la legge di Y/X.

Solution. La prima cosa che possiamo osservare è che ci sono date due variabili aleatoria esponenziali indipendenti con *parametri* pari rispettivamente a 1 e 3.5, poiché la media è il reciproco del parametro.

Partiamo dalla prima domanda. Possiamo riformularla nel seguente modo, sfruttando l'indipendenza delle due variabili aleatorie: $P(X < 0.57 \leq Y) = P(\{X < 0.57\} \cap \{Y \geq 0.57\}) = P(Y \geq 0.57) \cdot P(X < 0.57)$. Notare l'intersezione, poiché vogliamo che entrambe le disuguaglianze siano valide contemporaneamente.

Ora possiamo richiamare la definizione di funzione di ripartizione (e le sue proprietà) per scrivere $P(Y \geq 0.57) = 1 - F_Y(0.57)$, $P(X < 1) = F_X(0.57)$, utilizzando due volte il fatto che la distribuzione esponenziale è (assolutamente) continua, quindi la probabilità di singoli punti è 0. Per calcolare i valori effettivi, possiamo fare affidamento sulla funzione `pexp` di R, con parametri appropriati.

Passiamo alla seconda domanda e iniziamo osservando la legge di $3.5 \cdot Y$. Poiché Y è una esponenziale, allora αY è anch'essa una esponenziale per ogni $\alpha > 0$, con parametro $\alpha^{-1}\lambda_Y$. In questo caso, $\alpha = 3.5$ e $\lambda_Y = 3.5$, quindi $3.5 \cdot Y$ ha parametro 1.

Avremmo potuto ottenere lo stesso risultato anche dalla definizione:

$$P(3.5 \cdot Y \leq t) = P\left(Y \leq \frac{t}{3.5}\right) = 1 - \mathrm{e}^{-3.5 \cdot \frac{t}{3.5}} = 1 - \mathrm{e}^{-t}.$$

Per rispondere alla domanda, dobbiamo calcolare la densità di probabilità congiunta e integrarla sulla porzione del piano che ci interessa (ovvero, dove $3.5 \cdot Y \leq X$). Poiché le due variabili aleatorie X e Y e quindi X e $3.5 \cdot Y$ sono indipendenti, la densità congiunta è

$$f_{X,3.5\cdot Y}(s,t) = f_X(s) \cdot f_{3.5\cdot Y}(t) = \mathrm{e}^{-s} \cdot \mathrm{e}^{-t}.$$

Integriamo sopra la bisettrice del primo quadrante, cioè

$$\int_0^{+\infty} \int_s^{+\infty} \mathrm{e}^{-t}\mathrm{e}^{-s}\,\mathrm{d}t\,\mathrm{d}s = \int_0^{+\infty} \mathrm{e}^{-2s}\,\mathrm{d}s = 0.5$$

Avremmo potuto integrare anche in orizzontale: $\int_0^{\infty} \int_0^t \mathrm{e}^{-s}\,\mathrm{d}s\,\mathrm{e}^{-t}\,\mathrm{d}t$, ma in questo modo avremmo integrali leggermente più complicati, a causa della presenza di un termine aggiuntivo.

Un'altra possibilità sarebbe stata riscrivere la domanda come somma di variabili casuali:

$$P(3.5 \cdot Y \leq X) = P(3.5 \cdot Y - X \leq 0)$$

e quindi usare le proprietà delle trasformazioni lineari delle variabili casuali e la somma di variabili casuali indipendenti. È utile sottolineare, a questo punto, che la somma di due esponenziali non è un'esponenziale.

Infine, la risposta numerica ci può suggerire un altro punto di vista ancora: siccome le due variabili aleatorie X e $3.5\,Y$ sono indipendenti e identicamente distribuite, nonché continue, la probabilità che una sia minore o uguale dell'altra è esattamente $\frac{1}{2}$.

Arriviamo all'ultima domanda. Possiamo partire dalle definizioni e vedere dove ci portano:

$$P\left(\frac{Y}{X} \leqslant t\right) = P(Y \leqslant t\,X) = \iint_{0 \leqslant y \leqslant tx} 3.5\,\mathrm{e}^{-3.5y}\mathrm{e}^{-x}\,\mathrm{d}y\,\mathrm{d}x = 3.5\int_0^{+\infty}\int_0^{tx}\mathrm{e}^{-3.5y}\,\mathrm{d}y\,\mathrm{e}^{-x}\,\mathrm{d}x$$

dove gli integrali sono abbastanza simili a quelli della domanda precedente. Non ci piace molto che abbiamo due estremi in cui la primitiva è diversa da zero nell'integrale più interno, poiché dovremmo integrare $e^{-3.5y}$ nell'intervallo $[0, t\,x]$.

Possiamo integrare in orizzontale nell'integrale più interno, oppure possiamo provare a passare al complementare: $P(Y/X \leq t) = 1 - P(Y/X \geq t) = 1 - P(Y \geq t \cdot X)$ e

$$P(Y \geq t\,X) = \int_0^{+\infty}\int_{tx}^{+\infty} 3.5\,\mathrm{e}^{-3.5y}\mathrm{e}^{-x}\,\mathrm{d}y\,\mathrm{d}x = \int_0^{+\infty}\mathrm{e}^{-(3.5t+1)x}\,\mathrm{d}x = \frac{(3.5)^{-1}}{t + (3.5)^{-1}}.$$

Otteniamo quindi $P(Y/X \leq t) = 1 - P(Y/X \geq t) = 1 - \frac{(3.5)^{-1}}{t+(3.5)^{-1}} = 1 - \frac{t}{t+(3.5)^{-1}}$.

Problema 2. In quanti modi diversi è possibile scrivere 23 come somma di 8 interi positivi (tenendo conto dell'ordine: $2+1$ e $1+2$ sono due modi diversi di scrivere 3 come somma di due interi positivi)?

Solution. Possiamo visualizzare il numero 23 come una sequenza di punti. Riscriverlo come somma di 8 interi positivi può quindi essere visto come raggruppare questi punti in 8 sequenze. Possiamo farlo mettendo 7 separatori tra i punti. Ci sono 22 spazi tra i punti dove possiamo mettere quei separatori (uno spazio prima di ciascun punto, ad eccezione del primo punto).

Abbiamo quindi $\binom{22}{7} = 1.70544 \times 10^5$ modi di farlo.

Problema 3. Due amici giocano al seguente gioco: lanciano una moneta 11 volte e contano il numero di teste. Per ogni testa, estraggono (senza sostituzione) da un'urna contenente biglie numerate da 1 a 90.

1. Se la probabilità che la moneta esca testa è 0.55, qual è la probabilità che venga estratto il numero 41?
2. Se vengono estratte 3 biglie, qual è il valore più probabile per la probabilità che la moneta esca testa?
3. Se vengono estratte 5 biglie, qual è la probabilità che vengano estratte in ordine decrescente?

Problema 4. Il server di proprietà di un gruppo di ricerca impiega un tempo medio di 3.53 giorni (con deviazione standard 1.38) per eseguire una simulazione particolarmente difficile.

1. Se il gruppo di ricerca vuole eseguire simulazioni per 41 differenti dati iniziali, quanto tempo ci vorrà (in media) per eseguirle tutte?
2. Qual è la probabilità che siano necessari più di 144 giorni per eseguire tutte le 41 simulazioni?

Solution. Notiamo che non ci è data una distribuzione, ma abbiamo solamente media e deviazione standard. Per la prima domanda, poiché ci viene chiesta solo la media, possiamo usare la Legge dei Grandi Numeri e ottenere che la media è semplicemente la somma di tutte le medie, cioè $41 \cdot 3.53 = 144.73$.

Passiamo alla seconda domanda. Nella domanda precedente non avevamo bisogno di una distribuzione, ora sì, poiché vogliamo calcolare una probabilità. Non conosciamo la distribuzione dei tempi, ma possiamo usare il Teorema del Limite Centrale per approssimare questa somma con una Gaussiana di media 144.73 e deviazione standard $1.38 \cdot \sqrt{41} = 8.8363114$.

Stiamo chiedendo che la somma di tutti i tempi delle simulazioni sia maggiore o uguale a 144, quindi calcoliamo la probabilità che una Gaussiana con i parametri sopra sia maggiore o uguale a 144, ottenendo 0.5329206.

A.2. UN POSSIBILE ESAME

Problema 5. Un'urna contiene 10 biglie nere e 7 bianche.

1. Se estraiamo 15 biglie *senza reinserimento*, qual è la probabilità di vedere 9 biglie nere e 6 bianche, sapendo che la nona biglia estratta è nera?
2. Se estraiamo 15 biglie *con reinserimento*, qual è la probabilità di vedere 9 biglie nere e 6 bianche, sapendo che la nona biglia estratta è nera?

Solution. La prima cosa da notare è che la probabilità che ci viene richiesta è difficile da calcolare direttamente, ma la formulazione condizionale suggerisce l'uso del teorema di Bayes.

Chiamiamo S l'evento "la sequenza estratta ha 9 biglie nere e 6 bianche", B_9 l'evento "la nona biglia estratta è nera". Quindi ci viene chiesto

$$P(S|B_9) = \frac{P(B_9|S)\,P(S)}{P(B_9)}.$$

Esaminiamo gli oggetti differenti sul lato destro dell'equazione.

Iniziamo con $P(B_9|S)$: abbiamo una sequenza di 15 biglie, 9 delle quali nere e le restanti 6 bianche, e vogliamo sapere la probabilità che la nona biglia sia nera. Possiamo vedere il problema in modo diverso: abbiamo una parola composta da 9 lettere "N" e 6 lettere "B". I casi favorevoli sono quelli in cui la nona lettera è "N", cioè gli anagrammi di 14 lettere di cui 8 sono "N" e 6 sono "B". Questi sono $\frac{14!}{8!6!}$. I casi possibili sono tutte le parole composte da 9 lettere "N" e 6 lettere "B", che sono tutti gli anagrammi, ossia $\frac{15!}{9!6!}$. Se prendiamo il loro rapporto, si semplificano e otteniamo $\frac{9}{15}$. Questo infatti non dovrebbe sorprenderci, poiché non abbiamo altre informazioni oltre al numero totale di lettere e al numero di lettere "N".

Analogamente, possiamo affrontare il termine $P(B_9)$: l'idea è la stessa, ma siamo nel caso generale, con tutte le 10 biglie nere e le 7 bianche nell'urna. Il rapporto che stiamo cercando è quindi $\frac{10}{17}$.

Resta solo un termine, il più difficile: $P(S)$. Qui possiamo notare, scrivendo alcuni casi espliciti, che ogni sequenza della forma corretta ha la stessa probabilità: $\frac{10\cdots2\cdot7\cdots2}{17\cdots3} = 1.0283834 \times 10^{-4}$.

Dobbiamo però moltiplicare questa probabilità per il numero di sequenze possibili di questa forma (fondamentalmente stiamo sommando la loro probabilità, poiché sono a due a due disgiunte). Le sequenze sono $\binom{15}{9}$. Mettendo tutto insieme $P(S) = 0.5147059$.

La probabilità che stiamo cercando è quindi

$$P(S|B_9) = \frac{P(B_9|S)\,P(S)}{P(B_9)} = \frac{0.6 \cdot 0.5147059}{0.5882353} = 0.525.$$

Passiamo alla seconda domanda. Questo è un caso più semplice, rispetto al precedente: a causa del reinserimento abbiamo una distribuzione binomiale. Notiamo che sapere che la nona biglia estratta è nera ci dice solo che vogliamo estrarre 14 biglie e puntare su 8 successi (se consideriamo successo l'estrazione di una biglia nera). La risposta è quindi 0.2098257.

Problema 6. Sono date 8 variabili aleatorie indipendenti e identicamente distribuite $X_1, \ldots, X_8$, con funzione densità di probabilità $f(x) = 1$ in $[0,1]$ e 0 altrimenti. Sia $M = \min\{X_1, \ldots, X_8\}$.

1. Qual è la probabilità che $M \geq 0.672$?

2. Qual è il valore atteso di M?

Solution. Poiché M è il minimo degli X_i, la probabilità $P(M \geq 0.672)$ è uguale alla probabilità che tutti gli $X_i \geq 0.672$, cioè

$$P(M \geq 0.672) = P(\{X_1 \geq 0.672\} \cap \cdots \cap \{X_8 \geq 0.672\}).$$

Grazie all'indipendenza degli X_i, possiamo riscrivere la probabilità dell'intersezione come il prodotto delle probabilità:

$$P(M \geq 0.672) = \prod_{i=1}^{8} P(X_i \geq 0.672).$$

Gli X_i sono identicamente distribuiti come uniformi in $[0,1]$, e in particolare sono variabili aleatorie continue, quindi

$$P(M \geq 0.672) = \prod_{i=1}^{8} (1 - P(X_i \leq 0.672)) = \prod_{i=1}^{8} (1 - F_{X_i}(0.672)) = (1 - 0.672)^8.$$

Possiamo calcolare questo valore usando R, ottenendo 1.3396482×10^{-4}.

Per calcolare il valore atteso di M, una variabile aleatoria continua, utilizzeremo la definizione, cioè

$$E[M] = \int_0^1 x \cdot f_M(x)\,\mathrm{d}x$$

Tuttavia, per farlo, dobbiamo calcolare la funzione di

densità per M, poiché abbiamo solo la funzione di distribuzione cumulativa dalla domanda precedente. Prendiamo la derivata della cdf:

$$f_M(x) = F'_M(x) = \frac{\mathrm{d}}{\mathrm{d}x} 1 - (1-x)^8 = 8(1-x)^7.$$

Quindi

$$E[M] = \int_0^1 x \cdot 8(1-x)^7 \mathrm{d}x = \left[-\frac{(1-x)^8 (8x+1)}{9} \right]_0^1 = \frac{1}{9}$$

Problema 7. La funzione di densità congiunta di due variabili aleatorie (assolutamente continue) X e Y è $f_{X,Y}(x,y) = \alpha x y$ per $0 \leq x \leq 1$ e $0 \leq y \leq \sqrt{x}$ e 0 altrimenti.

1. Determina il valore di α.
2. Determina se X e Y sono indipendenti o meno.
3. Calcola la funzione di densità condizionata per X dato $Y = y$.

Solution. Nella prima domanda dobbiamo solo calcolare un integrale doppio e impostarlo uguale a 1:

$$1 = \int_0^1 \int_0^{\sqrt{x}} \alpha x y \mathrm{d}y \mathrm{d}x = \frac{\alpha}{2} \int_0^1 x^2 \mathrm{d}x = \frac{\alpha}{2 \cdot 3}$$

quindi $\alpha = 6$.

Nella seconda dobbiamo calcolare le marginali e verificare se il loro prodotto è uguale alla funzione di densità congiunta.

L'unica cosa a cui prestare attenzione sono i limiti di integrazione, per X otteniamo

$$f_X(x) = \int_0^{\sqrt{x}} 6 x y \mathrm{d}y = 3x^2$$

per $x \in [0,1]$ e 0 altrimenti.

Per Y dovremmo notare che x varia in $[0,1]$ ma deve essere tale che $y \leq \sqrt{x}$, ossia (poiché $y \geq 0$) $x \geq y^2$, quindi

$$f_Y(y) = \int_{y^2}^1 6 x y \mathrm{d}x = 3y(1 - y^4)$$

per $y \in [0,1]$ e 0 altrimenti.

Ora possiamo verificare che $f_X(x) \cdot f_Y(y) \neq f_{X,Y}(x,y)$, quindi le due variabili aleatorie non sono indipendenti.

Nella terza dobbiamo ricordare che

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

se $f_Y(y) \neq 0$ e 0 altrimenti. Abbiamo tutti gli elementi necessari e ancora una volta dobbiamo solo prestare attenzione agli estremi di definizione.

$$f_{X|Y}(x|y) = \begin{cases} \frac{2x}{1-y^4} & y^2 \leq x \leq 1, y \neq 1 \\ 0 & \text{altrimenti} \end{cases}.$$

A.3. PROVA IN ITINERE (2024/04/15)

Problema 8. La variabile aleatoria X ha come funzione di densità

$$f_X(x) = \begin{cases} c\,x\,(1-x) & 0 \leqslant x \leqslant 1 \\ 0 & \text{altrimenti.} \end{cases}$$

Determina c. Calcola media, mediana, moda e varianza di X e traccia il grafico (qualitativo) della funzione di ripartizione. Sia $Y = 2X - 1$. Calcola media e varianza di Y. Scrivi la funzione di densità di Y e tracciane il grafico (qualitativo).

Solution. Dal momento che f_X deve essere una densità, il suo integrale su $\mathbb{R}$ deve essere uguale a 1. Dal momento che l'integrale in $[0,1]$, che è il supporto di f_X, della funzione polinomiale $x\,(1-x)$ è $\frac{1}{6}$, abbiamo $c = 6$.

La moda è per definizione l'argomento del massimo della densità, dal momento che stiamo trattando una variabile assolutamente continua. La densità è una parabola con concavità verso il basso e ha massimo nel suo punto medio, ossia in 0.5.

La funzione di ripartizione è $F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$, ossia

$$F_X(x) = \begin{cases} 0 & x \leqslant 0 \\ (3-2x)\,x^2 & 0 \leqslant x \leqslant 1 \\ 1 & x \geqslant 1 \end{cases}.$$

Per quanto riguarda la media e la mediana, per simmetria possiamo individuarle nel punto medio del supporto, ossia 0.5. Possiamo confermare che questo è il caso per la mediana andando a controllare che la funzione di ripartizione in 0.5 valga 0.5. Per la media possiamo calcolare

$$E[X] = \int_0^1 x\,6\,x\,(1-x)\,\mathrm{d}x = -6\left[\frac{x^4}{4} - \frac{x^3}{3}\right]_0^1 = 6\left(\frac{1}{12} - 0\right) = 0.5.$$

Avendo congetturato che la media fosse 0.5, possiamo anche pensare di sfruttare la linearità della speranza, la speranza di una trasformazione e un cambio di variabile

$$\begin{aligned} E[X] &= E[X - 0.5] + 0.5 \\ &= 6\int_0^1 (x-0.5)\,x(1-x)\,\mathrm{d}x + 0.5 \\ &= 6\int_{-0.5}^{0.5} x\,(x+0.5)\,(0.5-x)\,\mathrm{d}x + 0.5 \\ &= 6\int_{-0.5}^{0.5} -x^3 + 0.25\,x\,\mathrm{d}x + 0.5 \\ &= 0.5 \end{aligned}$$

perché la funzione integranda è dispari e integrata su un intervallo simmetrico.

Per la varianza abbiamo bisogno di calcolare il momento secondo

$$E[X^2] = \int_0^1 x^2 6x(1-x)\,dx = 6\left[\frac{x^4}{4} - \frac{x^5}{5}\right]_0^1 = \frac{3}{10} = 0.3$$

che dà $\mathrm{Var}[X] = E[X^2] - E[X]^2 = 0.3 - 0.25 = 0.05$.

Per quanto riguarda il grafico qualitativo di F_X dobbiamo usare le proprietà che abbiamo raccolto. In 0.5 (mediana) $F_X(0.5) = 0.5$, quindi il grafico deve passare da $(0.5, 0.5)$. Tale punto è anche moda, quindi ci sarà un cambio di concavità per F_X. In più $F_X \equiv 0$ a sinistra di 0 e $F_X \equiv 1$ a destra di 1.

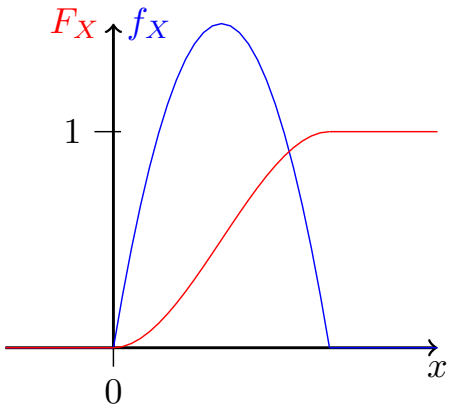

Figura A.1. Funzione di ripartizione e di densità di X.

Passiamo allora a $Y = 2X - 1$. È una trasformazione lineare di X, quindi, invece di considerarla come una nuova variabile aleatoria e ripartire "da zero", cerchiamo di usare quanto già ricavato per X.

La media di Y è 0 per linearità,

$$E[Y] = E[2X - 1] = 2E[X] - 1 = 0,$$

la varianza è 0.2

$$\mathrm{Var}[Y] = \mathrm{Var}[2X - 1] = 4\,\mathrm{Var}[X] = 0.2.$$

La densità di Y può essere ricavata usando il risultato visto per le trasformazioni lineari,

$$f_Y(x) = \frac{1}{2} f_X\left(\frac{x+1}{2}\right) = 3\left(\frac{x+1}{2}\right)\left(\frac{1-x}{2}\right)$$

per $x \in [-1, 1]$ (e nulla altrimenti), dal momento che $\frac{x+1}{2} \in [0, 1]$.

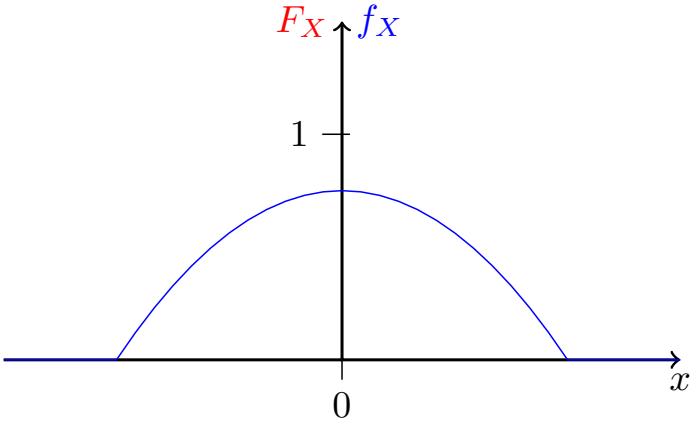

Figura A.2. Funzione di densità di $Y = 2X - 1$.

Commento: più che un problema questo è un esercizio, molto 'scolastico'.

Problema 9. Quello che segue è un problema tratto da un esame passato, con la risposta data da una persona che ha preso parte. Controlla se la risposta è corretta e, nel caso non fosse così, indica tutti gli errori che hai individuato nel ragionamento. Fornisci anche una soluzione corretta che superi i limiti che hai trovato. I valori calcolati sono corretti (nel senso che gli errori, se presenti, non sono nei conti).

> *Problema:* In una pasticceria ci sono 7 cannoncini alla crema, 6 allo zabaglione e 5 alla nutella. Bisogna preparare un vassoio con 7 paste. Se i cannoncini sono scelti uniformemente a caso, qual è la probabilità che nel vassoio ci siano almeno 2 cannoncini per tipo?
>
> *Soluzione:* Ci sono 7 cannoncini alla crema C, 6 allo zabaglione Z, e 5 alla nutella N, quindi in tutto abbiamo 18 cannoncini. Per lo spazio di probabilità scegliamo come Ω l'insieme di tutte le possibili cinquine di cannoncini, come tribù l'insieme delle parti, visto che l'insieme Ω è finito, come probabilità quella uniforme discreta, classica che dà la medesima probabilità a ogni quintupla.
>
> Cominciamo con il conteggio dei casi totali: abbiamo $\binom{18}{7} = 31824$ modi di scegliere 7 cannoncini tra 18.
>
> Per i casi favorevoli, abbiamo $\binom{6}{2} = 15$ modi di scegliere le due Z, $\binom{7}{2} = 21$ modi di scegliere le due C, $\binom{5}{2} = 10$ modi di scegliere le due N e $\binom{12}{1} = 12$ modi di scegliere il rimanente cannoncino, dal momento che può essere di uno qualunque dei tre tipi e che ne sono rimasti $(6-2)+(7-2)+(5-2) = 12$.
>
> La probabilità richiesta è allora
>
> $$\frac{\#F}{\#T} = \frac{\binom{7}{2}\cdot\binom{6}{2}\cdot\binom{5}{2}\cdot\binom{12}{1}}{\binom{18}{7}} = \frac{37800}{31824}.$$

Solution. La soluzione proposta chiaramente non può essere corretta, dal momento che il risultato numerico finale è strettamente maggiore di 1, pur essendo una probabilità. Dal momento che il testo ci garantisce che non ci sono errori di calcolo, il punto critico deve essere da qualche parte nel ragionamento.

A un primo sguardo il ragionamento appare corretto, anche se è indicato *cinquine* invece di *settuple*, quindi se il risultato numerico non fosse palesemente errato, potremmo avere la tentazione di accettarlo. A priori non c'è un solo modo valido di contare casi favorevoli e casi totali, quindi proviamo a tenere quello scelto da chi ha risolto il problema. Osserviamo che gli oggetti in Ω sono i diversi cannoncini (possiamo pensarli numerati), non i cannoncini dei vari gusti, quindi sono effettivamente equiprobabili. Passiamo al numeratore, che dobbiamo a questo punto calcolare in maniera analoga. La scelta fatta da chi ha risolto il problema ricorda il multinomiale, ossia la generalizzazione del binomiale a più tipi di oggetti (che abbiamo coinvolto nell'ipergeometrica, per esempio). Tuttavia, come suggerisce il fatto che il numeratore sia maggiore del denominatore, ci sono casi che stiamo contando più volte.

In particolare l'ultimo termine non sembra dipendere dalle scelte fatte in precedenza, quindi necessariamente alcuni casi vengono tirati in ballo più volte. Una possibilità di risolvere la questione è osservare che per ottenere 7 paste prendendone almeno 2 per tipo abbiamo solamente 3 modi: prenderne 3 di un tipo e 2 per ciascuno degli altri tipi. Quindi

$$\binom{7}{3}\binom{6}{2}\binom{5}{2} + \binom{7}{2}\binom{6}{3}\binom{5}{2} + \binom{7}{2}\binom{6}{2}\binom{5}{3} = 12600$$

e la probabilità richiesta è quindi circa 40%. Volendo essere sicuri a questo punto possiamo fare un controllo di quali siano gli altri casi e verificare di non aver dimenticato nulla.

Problema 10. In una versione semplificata del Monopoli, ogni turno si tirano due dadi e si avanza di un numero di caselle uguale alla somma dei risultati dei due dadi. Inoltre, se i dadi hanno lo stesso risultato, si tirano nuovamente i dadi e si avanza ancora (continuando a tirare e avanzare senza limiti a priori, finché i dadi non danno punteggi diversi). Di quanto si avanza mediamente in un turno?

Solution. Sia X il numero di caselle di cui si avanza. Potremmo pensare di procedere per esaustione dei casi:

$$\varphi_X(k) = P(X=k) = \begin{cases} 0 & k=2 \\ \frac{2}{36} & k=3 \\ \frac{2}{36} & k=4 \\ \dots & \dots \end{cases}$$

in cui abbiamo prestato attenzione al fatto che non possiamo avanzare di 2, perché lo possiamo ottenere solamente come somma di due 1, cosa che causa un nuovo lancio dei dadi, mentre non abbiamo problemi per $k=3$. Per $k=4$ consideriamo solamente l'opzione che escano 1 e 3, mentre per $k=5$ inizierebbero ad esserci le prime aggiunte: le coppie $(1,4)$ o $(2,3)$, ma anche la coppia $(1,1)$ seguita da una delle coppie $(1,2)$. Non sembra una strada veramente percorribile.

Prendiamo un approccio un po' differente. Indichiamo con Y il punteggio ottenuto dalla prima coppia di dadi e sia H l'evento "la prima coppia di dadi ha dato due valori uguali". Allora possiamo osservare che:

$$E[X|H^c] = E[Y|H^c]$$

abbastanza chiaramente, ma anche (sfruttando l'assenza di memoria dei dadi)

$$E[X|H] = E[Y|H] + E[X]$$

perché nel momento in cui sappiamo che i dadi verranno lanciati di nuovo siamo al punto di partenza. A questo punto possiamo usare la formula di fattorizzazione:

$$\begin{aligned} E[X] &= E[X|H^c]\,P(H^c) + E[X|H]\,P(H) \\ &= E[Y|H^c]\,P(H^c) + E[Y|H]\,P(H) + E[X]\,P(H) \\ &= E[Y] + E[X]\,P(H). \end{aligned}$$

Ora possiamo risolvere in $E[X]$:

$$E[X] = \frac{E[Y]}{1-P(H)} = \frac{7}{1-\frac{1}{6}} = \frac{42}{5} = 8.4.$$

Soluzione alternativa. Chiamiamo Y_i per $i \in \mathbb{N}^+$ i lanci successivi di coppie di dadi (tutti i.i.d.) e H_i, $i \in \mathbb{N}^+$ gli eventi "al lancio i-simo ci sono due dadi uguali". Allora

$$X = Y_1 + \mathbb{1}_{H_1} Y_2 + \mathbb{1}_{H_1}\mathbb{1}_{H_2} Y_3 + \dots = \sum_{i\in\mathbb{N}^+} Y_i \prod_{j<i} \mathbb{1}_{H_j}$$

e facendo il valore atteso (osservando che Y_i è indipendente da tutti gli H_j con $j<i$):

$$\begin{aligned} E[X] &= E[Y_1] + E[\mathbb{1}_{H_1} Y_2] + E[\mathbb{1}_{H_1}\mathbb{1}_{H_2} Y_3] + \cdots \\ &= E[Y_1] + E[\mathbb{1}_{H_1}]\,E[Y_2] + E[\mathbb{1}_{H_1}]\,E[\mathbb{1}_{H_2}]\,E[Y_3] + \cdots \\ &= E[Y_1] + P(H_1)\,E[Y_2] + P(H_1)\,P(H_2)\,E[Y_3] + \cdots \\ &= E[Y_1] + P(H_1)\,E[Y_1] + P(H_1)^2\,E[Y_1] + \cdots \\ &= \sum_{k=0}^{+\infty} E[Y_1]\,P(H_1)^k = 7\sum_{k=0}^{+\infty}\left(\frac{1}{6}\right)^k = \frac{7}{1-\frac{1}{6}} = 8.4. \end{aligned}$$

Commento: La prima soluzione richiede un po' di creatività, la seconda è simile alle costruzioni viste in classe per geometrica e ipergeometrica.

Problema 11. In un recente festival di tecnologia e intrattenimento, sono state installate due diverse stazioni che distribuiscono automaticamente gadget. La prima stazione, *Alpha*, dà gadget seguendo una distribuzione di Poisson di media 5 gadget per visitatore. La seconda stazione, *Beta*, offre premi "esclusivi", e dà gadget secondo lo stesso tipo di distribuzione, ma con media 2 gadget per visitatore. I visitatori scelgono la stazione *Alpha* con probabilità $p = 0.8$, altrimenti la stazione *Beta*.

1. Determina la distribuzione del numero di gadget dati per visitatore e calcolarne la media.
2. Sapendo che un visitatore ha ricevuto esattamente 3 gadget, qual è la probabilità che essi siano stati distribuiti dalla stazione *Beta*?

Solution. Indichiamo con G la variabile aleatoria che conta il numero di gadget dati a un visitatore. Indichiamo inoltre con A e B gli eventi "scelta stazione Alpha" e "Beta" rispettivamente. I dati del problema sono

$$P(G=k|A) = \frac{5^k}{k!}\mathrm{e}^{-5} \qquad\qquad P(G=k|B) = \frac{2^k}{k!}\mathrm{e}^{-2}$$

e $P(A) = 0.8$ e $P(B) = 0.2$.

Allora la densità discreta di G è

$$\begin{aligned} P(G=k) &= P(G=k|A)\,P(A) + P(G=k|B)\,P(B) \\ &= \frac{1}{k!}\,(5^k\mathrm{e}^{-5}\cdot 0.8 + 2^k\mathrm{e}^{-2}\cdot 0.2). \end{aligned}$$

Possiamo osservare in particolare che non si tratta di una distribuzione di Poisson. Questo non collide con la proprietà di riproducibilità delle Poisson, dal momento che non stiamo sommando due Poisson indipendenti, ma ne stiamo facendo una media pesata.

Per calcolare la media potremmo usare la definizione nel caso di variabili aleatorie discrete, ma non sembra troppo facile. Per fortuna abbiamo la linearità della speranza, quindi

$$E[G] = 0.8\,E[G|A] + 0.2\,E[G|B] = 0.8\cdot 5 + 0.2\cdot 2 = 4.4.$$

Per la seconda domanda ci appoggiamo al teorema di Bayes per riscrivere in termini di probabilità che abbiamo la probabilità richiesta,

$$\begin{aligned} P(B|G=3) &= \frac{P(G=3|B)\,P(B)}{P(G=3)} \\ &= \frac{\frac{1}{3!}2^3\mathrm{e}^{-2}\cdot 0.2}{\frac{1}{3!}\,(5^3\mathrm{e}^{-5}\cdot 0.8 + 2^3\mathrm{e}^{-2}\cdot 0.2)} \\ &\approx 0.25. \end{aligned}$$

A.4. PROVA IN ITINERE (2024/06/06)

Problema 12. Un quadrato Q è inscritto in una circonferenza C. Sia p la probabilità che un punto preso uniformemente a caso nel cerchio delimitato da C sia strettamente interno a Q.

1. Qual è il valore di p?
2. Qual è la probabilità che, scegliendo uniformemente a caso e indipendentemente tra loro 10 punti nel cerchio delimitato da C almeno 3 non giacciano all'interno di Q? (Nel caso il risultato dipenda da p e tu non lo abbia calcolato in 1., puoi lasciare indicato il parametro p.)
3. Sia x_0 un punto sulla *circonferenza* C fissato. Scegliamo uniformemente a caso un altro punto sulla circonferenza C. Qual è la lunghezza media della corda che unisce questi due punti? (Nel caso il risultato dipenda da p e tu non lo abbia calcolato in 1., puoi lasciare indicato il parametro p. Può essere utile passare alle coordinate polari e ricordare l'identità $\frac{1-\cos x}{2} = \sin^2(\frac{x}{2})$.)

Solution. Per rispondere alla prima domanda possiamo modellizzare il problema con delle distribuzioni uniformi sulle figure di interesse, in particolare ci basta fare il rapporto tra le aree. Osserviamo che la richiesta che il punto sia "strettamente" interno è irrilevante, perché la probabilità che un punto giaccia su un segmento (o un'unione finita di segmenti) è nulla.

La risposta cercata è dunque (chiamato D il disco delimitato da C)

$$\frac{S_Q}{S_D} = \frac{(r\sqrt{2})^2}{\pi r^2} = \frac{2}{\pi}.$$

Possiamo vedere ogni punto come una variabile aleatoria Bernoulliana di parametro p: consideriamo come successo il fatto che un punto sia strettamente interno a Q e come insuccesso che sia esterno a Q (si potrebbe rappresentare il problema in maniera speculare, prendendo come parametro della Bernoulliana $p' = 1-p$). Siccome abbiamo 10 punti presi indipendentemente, la variabile da considerare è la somma di 10 Bernoulliane di medesimo parametro, ossia una binomiale. Chiedere che almeno 3 giacciano all'esterno equivale a chiedere che al più 7 giacciano all'interno.

La risposta cercata è pertanto (chiamata $S_{10} \sim \text{bin}(10, p)$

$$P(S_{10} \leqslant 7) = \sum_{i=0}^{7} \binom{10}{i} p^i (1-p)^{10-i} = 1 - \sum_{i=8}^{10} \binom{10}{i} p^i (1-p)^{10-i}$$

che possiamo calcolare con R come `pbinom(7, size = 10, prob = 2/pi)`, circa 76%.

Come prima cosa osserviamo che possiamo assumere senza perdita di generalità che $x_0 = (r, 0)$ dove r è il raggio della circonferenza C. Passando in coordinate polari, x_0 ha la medesima rappresentazione, mentre X_1 sarà della forma $X_1 = (r, \vartheta)$, con $\vartheta \sim \text{unif}[0, 2\pi]$. La distanza D tra i due punti è

$$\begin{aligned} D &= |X_1 - x_0| \\ &= \sqrt{r^2(\cos(\vartheta)-1)^2 + r^2\sin(\vartheta)^2} \\ &= r\sqrt{\cos(\vartheta)^2 - 2\cos(\vartheta) + 1 + \sin(\vartheta)^2} \\ &= 2r\sqrt{\frac{1-\cos(\vartheta)}{2}} \\ &= 2r\sqrt{\sin\left(\frac{\vartheta}{2}\right)^2} \\ &= 2r\left|\sin\left(\frac{\vartheta}{2}\right)\right|. \end{aligned}$$

Il valore atteso di D è

$$E[D] = \int_0^{2\pi} 2r\left|\sin\left(\frac{\vartheta}{2}\right)\right| \frac{1}{2\pi}\,\mathrm{d}\vartheta = \frac{4r}{2\pi}\int_0^{\pi} \sin\left(\frac{\vartheta}{2}\right)\mathrm{d}\vartheta = \frac{2r}{\pi}\int_0^{\frac{\pi}{2}} \sin(\vartheta)\,2\,\mathrm{d}\vartheta = \frac{4r}{\pi}.$$

Problema 13. Quello che segue è un problema tratto da un esame passato, con la risposta data da una persona che ha preso parte. Controlla se la risposta è corretta e, nel caso non fosse così, indica tutti gli errori che hai individuato nel ragionamento. Fornisci anche una soluzione corretta che superi i limiti che hai trovato.

Problema: È data una sequenza di variabili aleatorie iid di media 1 e varianza 1.

1. Calcolare la probabilità che la somma delle prime 12 variabili della sequenza sia strettamente maggiore di 15, supponendo che siano **discrete**.
2. Calcolare la probabilità che la somma delle prime 12 variabili della sequenza sia strettamente maggiore di 15, supponendo che siano tutte esponenziali.

3. Vengono osservati i seguenti dati: $(12,16,19,12,20,5,11,10,8,13)$, di cui il primo è la somma delle prime 12 variabili (realizzate) della sequenza, il secondo la somma delle successive 12 e così via. Calcolare un intervallo di confidenza al 90% per la media.

Soluzione: Nella prima domanda dobbiamo prestare attenzione al fatto che le variabili sono discrete, quindi il fatto che sia maggiore e non maggiore o uguale è rilevante. Ciò premesso, non avendo altre informazioni usiamo il Teorema Centrale del Limite, sfruttando le informazioni sulla media e sulla varianza

$$P(S_{12}>15)=1-P(S_{12}\leqslant 15)\simeq 1-\Phi\left(\frac{15-12\cdot 1}{\sqrt{12\cdot 1}}\right)\approx 19.3\%.$$

Per quanto riguarda la seconda domanda non abbiamo bisogno di usare il TLC dal momento che le variabili esponenziali sono riproducibili, di parametro uguale alla somma dei parametri quindi la probabilità richiesta può essere calcolata in R (ricordando che le esponenziali sono continue, quindi $\geqslant 15$ o >15 è uguale) come segue: `pexp(15, rate = 12, lower.tail= FALSE)` che dà un risultato piccolissimo, prossimo a 0.

Per l'intervallo di confidenza usiamo di nuovo il TLC per approssimare con una Gaussiana. Usiamo una statistica normale perché è nota la varianza, anche se il campione è piccolo (10 osservazioni). La media campionaria puntuale è 12.6, ottenuta con R `sum(c(12,16,19,12,20,5,11,10,8,13))/10`. L'intervallo di confidenza è $12.6 \pm \Phi^{-1}(0.95)\frac{1}{10} = (12.44, 12.76)$

Solution. Iniziamo dalla prima domanda. Il fatto che si tratti di variabili aleatorie discrete è rilevante, ma non nel modo indicato, quanto nel fatto che è opportuno, data la limitatezza del numero di variabili considerate, appoggiarsi alla correzione di continuità. Abbiamo pertanto

$$P(S_{12}>15)=1-P(S_{12}\leqslant 15)\simeq 1-\Phi\left(\frac{15+0.5-12\cdot 1}{\sqrt{12\cdot 1}}\right)\approx 15.6\%$$

un risultato abbastanza diverso da quello ottenuto senza correzione.

Passiamo alla seconda domanda. Ora abbiamo una collezione di variabili assolutamente continue, quindi se usiamo il teorema limite centrale possiamo ricopiare quanto scritto nella soluzione proposta per il caso discreto

$$P(S_{12}>15)=1-P(S_{12}\leqslant 15)\simeq 1-\Phi\left(\frac{15-12\cdot 1}{\sqrt{12\cdot 1}}\right)\approx 19.3\%.$$

Per quanto riguarda invece la soluzione proposta per il caso esponenziale, osserviamo che ci sono almeno due problemi: il primo è che la distribuzione esponenziale non è riproducibile e in ogni caso la media sarebbe la somma delle medie, quindi il parametro (che è il reciproco della media) non sarebbe la somma dei reciproci, ma il reciproco della somma.

Passiamo alla terza domanda. Possiamo interpretare i dati che ci sono stati assegnati come 120 osservazioni della medesima popolazione (la legge della variabile aleatoria di partenza), di cui non ci viene data ogni singola osservazione, ma la somma a blocchi di 12. Questo non è un limite per il calcolo della stima della media (grazie alla proprietà associativa della somma), che è quindi data da `sum(c(12,16,19,12,20,5,11,10,8,13))/120`, uguale a 1.05.

Per quanto riguarda la varianza, possiamo assumere vera l'ipotesi di partenza che essa sia pari a 1 (come osservato anche nello svolgimento proposto). Se non facciamo questo, non abbiamo un modo di stimare la varianza, dal momento che non sappiamo calcolare lo scarto quadratico medio, avendo i dati aggregati. Se supponiamo nota la varianza, possiamo usare la statistica normale, come approssimazione (non abbiamo idea se la popolazione di partenza sia distribuita in maniera gaussiana o meno), ottenendo come intervallo $1.05 \pm \Phi^{-1}(0.95)\frac{1}{\sqrt{120}} \approx [0.90, 1.20]$.

Possiamo anche decidere di considerare come nuova variabile di interesse la somma di 12 delle variabili originali. In questo caso ha senso la stima puntuale ottenuta nella soluzione proposta (12.6) che è proprio $12 \cdot 1.05$. La numerosità del campione è decisamente limitata. Possiamo supporre nota la varianza (pari a $12 \cdot 1 = 12$) e usare una statistica normale approssimata (nemmeno in questo caso sappiamo come sia distribuita la variabile aleatoria, anche se essendo somma di variabili aleatorie iid possiamo aspettarci che abbia distribuzione approssimativamente normale, grazie al TLC citato anche nella soluzione proposta), ottenendo $12.6 \pm \Phi^{-1}(0.95)\sqrt{\frac{12}{10}} \approx [10.80, 14.40]$, che è esattamente il riscalamento di un fattore 12 di quanto ottenuto prima (come mai?).

Osserviamo che la soluzione proposta fa confusione tra questo approccio e il precedente, usando la media (e la dimensione del campione, dimenticando anche una radice) della dozzina ma la varianza della singola. In effetti l'intervallo di confidenza ottenuto avrebbe dovuto lasciare qualche sospetto, in quanto molto piccolo, specie per un campione così limitato.

Nel caso in cui consideriamo le nuove variabili aleatorie, somma di 12 realizzazioni delle precedenti, possiamo anche supporre che la varianza non sia nota e usare una statistica t, per via della scarsa numerosità del campione, come approssimazione. Calcoliamo lo stimatore puntuale della varianza `var(c(12,16,19,12,20,5,11,10,8,13))`, ottenendo 21.82, da cui abbiamo una stima intervallare pari a $12.6 \pm F_{t_9}^{-1}(0.95)\sqrt{\frac{21.82}{10}} \approx [9.89, 15.31]$. Se riscaliamo questo intervallo di un fattore 12 otteniamo $[0.82, 1.28]$, ancora centrato in 1.05, ma diverso dal precedente (come mai?).

Problema 14. Il dataset `iris` presente in ogni installazione di R contiene misurazioni di diverse specie di fiori. L'obiettivo dell'esercizio è quello di analizzare le lunghezze del sepalo per due specie di iris, `setosa` e `versicolor`, delle tre presenti nel dataset.

1. Confrontare (anche graficamente) le distribuzioni della variabile `Sepal.Length` per le due specie, spiegando quanto fatto e commentando i risultati ottenuti.
2. C'è evidenza statistica che le lunghezze medie dei sepali nelle due specie considerate siano differenti? Spiegare il procedimento seguito e commentare eventuali ipotesi fatte.

A.5. APPELLO D'ESAME (2024/06/24)

Problema 15. Un laboratorio deve valutare il contenuto di caffeina nel tè verde di una certa marca. Per fare questo, è stato considerato un campione di 20 bustine di tè e misurato il contenuto di caffeina in ciascuna bustina. I valori risultanti dalle analisi, in milligrammi per bustina, sono i seguenti:

$$35, 40, 38, 37, 39, 36, 35, 40, 38, 41, 34, 35, 37, 36, 40, 39, 37, 33, 36, 42.$$

Le leggi in vigore vietano la commercializzazione del tè per valori medi superiori a 39 milligrammi per bustina, quindi il dipartimento legale dell'azienda chiede un'analisi statisica sui dati sopra individuati che risponda ad alcuni quesiti.

1. Qual è una stima (con il 95% di fiducia) della probabilità che una bustina scelta a caso nell'intera produzione abbia contenuto di caffeina oltre la soglia legale?
2. C'è evidenza statistica che il prodotto sia a norma di legge?
3. Osservando altrettanti dati (estratti dalla medesima popolazione) e considerando il campione complessivo ottenuto unendo i dati nuovi ai precedenti, come cambieranno le risposte alle due domande sopra?

Solution. La prima domanda chiede di fare un intervallo di confidenza per il parametro di una bernoulli. Abbiamo 5 successi su 20 tentativi. Possiamo calcolare l'intervallo di fiducia al 95% con R usando la funzione `binom.test(5,20)`, da cui abbiamo (0.08657147, 0.49104587). La stima puntuale è 0.25.

Per discutere la legalità del prodotto dobbiamo invece discutere la media. Non abbiamo informazioni sulla distribuzione dei valori, in particolare non sappiamo se sia una popolazione normale (possiamo fare qualche controllo qualitativo), ma possiamo comunque pensare di usare un'approssimazione normale, nella consapevolezza che stiamo un po' forzando la mano. Dobbiamo impostare un test d'ipotesi in cui l'ipotesi alternativa è che sia legale, ossia che la media sia strettamente minore di 39. Se invece usassimo minore o uguale di 39 come ipotesi nulla, non rigettandola non daremmo evidenza statistica forte in suo favore.

Con R abbiamo allora `t.test(x, mu = 39, alternative = ``less")`, avendo salvato i dati nel vettore `x`. La risposta è

```
One Sample t-test

data:  x
t = -2.911, df = 19, p-value = 0.00448
alternative hypothesis: true mean is less than 39
95 percent confidence interval:
    -Inf 38.3504
sample estimates:
mean of x
     37.4
```

da cui ricaviamo che c'è evidenza statistica sufficiente a rigettare l'ipotesi nulla a ogni ragionevole livello di significatività (soprattutto considerando la dimensione ridotta del campione).

Per quanto riguarda la terza domanda, al crescere della numerosità del campione ci possiamo aspettare che, supponendo non varino troppo media e varianza campionarie, diminuisca la larghezza dell'intervallo di fiducia (possiamo anche dire esattamente di quanto, sapendo come entra in gioco la taglia del campione nella forma degli estremi dell'intervallo…). Per il test d'ipotesi sappiamo che aumenteranno i gradi di libertà della t e che (sotto le stesse ipotesi viste sopra) aumenterà di un fattore $\sqrt{2}$ il valore test della statistica. Siccome il valore della statistica era negativo, diventerà ancora più negativo e siccome al crescere del numero di gradi di libertà le code della t dimagriscono, diminuirà anche il p-dei-dati. Chiaramente siccome il numero dei dati è ancora ridotto, se i nuovi dati avessero media e varianza campionaria abbastanza diverse, potremmo dire molto di meno: il numero di gradi di libertà aumenta, così come $\sqrt{n}$ aumenta di un fattore $\sqrt{2}$, ma variando anche gli indicatori in gioco non possiamo più fare ipotesi così specifiche sui cambiamenti.

Problema 16. Sia X una variabile aleatoria continua con densità $f(t) = c$ per $4 < t < 8$.

1. Dopo aver determinato c, calcolare media, deviazione standard, moda e mediana di X e la probabilità $P(X > 5 | X < 6.5)$.
2. Ricavare legge, media e mediana della variabile aleatoria $Y := |X - 5|$.
3. Trovare una trasformazione $g\colon \mathbb{R} \to \mathbb{R}$ tale che $Z = g(X)$ sia una normale standard.

Solution. La variabile aleatoria proposta è una uniforme sull'intervallo proposto. Il valore di c è univocamente identificato dalla condizione di integrale uguale a 1, quindi $c = \frac{1}{4}$. La media di una uniforme continua è il punto medio del suo supporto, ossia 6, e coincide con la mediana. "La" moda è ogni punto interno all'intervallo di supporto. Per quanto riguarda la deviazione standard, ricordiamo che è per definizione la radice quadrata della varianza, che per una uniforme continua su $[a, b]$ ha valore $\frac{(b-a)^2}{12}$, ossia in questo caso $\frac{4}{3}$, da cui la deviazione standard è $\frac{2\sqrt{3}}{3}$.

Per calcolare la probabilità condizionata richiesta osserviamo che ci basta misurare le lunghezze dei segmenti coinvolti e moltiplicarle per $c = \frac{1}{4}$ (ma…).

Abbiamo $P(X > 5 | X < 6.5) = \frac{P(X \in [5, 6.5])}{P(X \in [4, 6.5])} = \frac{c \cdot 1.5}{c \cdot 2.5} = \frac{3}{5}$.

Possiamo affrontare la seconda domanda in diversi modi. Qui ne consideriamo uno che si appoggia alla legge di Y, ossia la funzione di ripartizione, dal momento che viene esplicitamente chiesta. Per definizione abbiamo

$$F_Y(y) = P(|X-5| \leqslant y) = P(-y \leqslant X-5 \leqslant y) = P(5-y \leqslant X \leqslant 5+y) = F_X(5+y) - F_X(5-y).$$

La funzione di ripartizione di X uniforme continua su $[4,8]$ è non banale solamente sull'intervallo, in cui vale $F_X(x) = \frac{x-4}{8-4} = \frac{x-4}{4}$. Allora

$$\begin{aligned} F_Y(y) &= \frac{5+y-4}{4}\mathbb{1}_{\{y\in[-1,3]\}} + \mathbb{1}_{\{y\in(3,+\infty)\}} - \frac{5-y-4}{4}\mathbb{1}_{\{y\in[-3,1]\}} - \mathbb{1}_{\{y\in(1,+\infty)\}} \\ &= \begin{cases} 0 & y<0 \\ \frac{y+1}{4} - \frac{1-y}{4} = \frac{y}{2} & 0 \leqslant y < 1 \\ \frac{y+1}{4} & 1 \leqslant y < 3 \\ 1 & y \geqslant 3 \end{cases} \end{aligned}$$

La mediana è 1, punto in cui F_Y assume il valore $\frac{1}{2}$. Per la media dobbiamo ricavare la densità di probabilità, che è costantemente $\frac{1}{2}$ in $[0,1]$, costantemente $\frac{1}{4}$ in $[1,3]$ e nulla altrove. Quindi la media di Y è

$$E[Y] = \int_{-\infty}^{+\infty} y f_Y(y)\,\mathrm{d}y = \int_0^1 \frac{y}{2}\mathrm{d}y + \int_1^3 \frac{y}{4}\mathrm{d}y = \left[\frac{y^2}{4}\right]_0^1 + \left[\frac{y^2}{8}\right]_1^3 = \frac{5}{4} > 1.$$

Avremmo anche potuto usare un'altra proprietà e calcolare $E[Y]$ come integrale di $f(x)\,f_X(x)$ con $f(x) = |x=5|$.

Per la trasformazione innanzitutto ci riconduciamo a una uniforme su $[0,1]$, quindi $W = \frac{X-4}{4}$, e poi la usiamo come argomento della funzione quantile di una normale standard. In altre parole $Z = \Phi^{-1}\left(\frac{X-4}{4}\right)$.

Problema 17. Un edicolante vende un quotidiano a due tipi di clienti: quelli con abbonamento e gli altri. Ogni giorno il numero di copie che vengono chieste dai primi è una variabile aleatoria X che può assumere, con la medesima probabilità, i valori 0, 1, 2, 3 e 4. Il numero Y di copie chieste dai secondi ha invece distribuzione di Poisson di media 2.5. È possibile assumere le due variabili indipendenti.

1. Sia $S = X + Y$ il numero totale di copie vendute. Calcolare media e varianza di S e il valore della funzione di massa di probabilità di S nei punti 0, 2 e 5.
2. Questa mattina l'edicolante ha ricevuto solamente 7 copie. Con che probabilità non saranno sufficienti?
3. Venerdì scorso ha venduto 5 copie. Com'è la distribuzione della coppia X, Y, conoscendo questa informazione?

Solution. La media di S è, per linearità, la somma delle medie di X e Y. Inoltre, dal momento che X e Y sono indipendenti, la varianza di S è la somma delle varianze di X e Y.

Allora $E[S] = 2 + 2.5 = 4.5$, mentre per la varianza dobbiamo fare qualche calcolo in più nel caso di X:

$$\mathrm{Var}[X] = \sum_{i=0}^{4} i^2 \frac{1}{5} - E[X]^2 = \frac{30}{5} - 4 = 2.$$

Abbiamo quindi che la varianza di S è $\mathrm{Var}[S] = 2 + 2.5 = 4.5$. Questo non implica che S abbia una legge Poissoniana.

Per calcolare le probabilità richieste, osserviamo che per avere somma 0 occorre che siano entrambe 0, cosa che avviene con probabilità $\frac{1}{5}\cdot \mathrm{e}^{-2.5} \approx 0.0164$. Per avere somma 2 abbiamo 3 casi possibili: $(2,0)$, $(1,1)$ e $(0,2)$. Il primo ha probabilità $\frac{1}{5}\cdot p_Y(0)$, il secondo $\frac{1}{5}p_Y(1)$, il terzo $\frac{1}{5}p_Y(2)$, cioè possiamo trovare la probabilità richiesta come $\frac{1}{5}P(Y \leqslant 2) \approx 0.1088$ (possiamo calcolare il valore numerico con R, usando `1/5*ppois(2,2.5)`). In pratica ci vanno bene i primi 3 valori della Poisson e per ciascuno di essi c'è un solo valore di X che permette di avere la somma cercata.

Possiamo chiaramente ottenere il medesimo risultato considerando la convoluzione delle due marginali, sfruttando l'indipendenza di X e Y, facendo attenzione agli estremi di sommazione: $p_S(s) = \sum_{x=0}^{4} p_x(x)\, p_Y(s-x) = \frac{1}{5}\sum_{x=0}^{4} p_Y(s-x)$.

Passiamo al terzo punto in cui valutare la massa di S: qui i casi sono solamente 5, ossia $(4,1)$, $(3,2)$, $(2,3)$, $(1,4)$, $(0,5)$. Con lo stesso trucco ricavato prima possiamo osservare che la probabilità cercata è $\frac{1}{5}(P(Y \leqslant 5) - P(Y \leqslant 0)) \approx 0.1752$, calcolabile in R come `1/5*(ppois(5,2.5)-ppois(0,2.5))`.

Per la seconda domanda vogliamo calcolare $P(S > 7)$. Non avendo calcolato la legge di S possiamo pensare a come si può presentare questa circostanza: a abbonati (tra 0 e 4 inclusi) e almeno $8-a$ non abbonati. Quindi abbiamo

$$\sum_{a=0}^{4} \frac{1}{5} \cdot P(Y > 7-a) \approx 0.0823$$

calcolabile come `1/5*sum(ppois(7-(0:4), 2.5, lower.tail=FALSE))`.

Arriviamo alla terza e ultima domanda. È chiesta $p_{X,Y}((x,y) \mid S=5)$, che possiamo vedere anche come $\frac{P(X=5-y, Y=y)}{P(S=5)} = \frac{1/5 \cdot p_Y(y)}{P(S=5)}$ per $y \in \{1,\dots,5\}$ e nulla altrimenti.

Combinando con quanto visto per la prima domanda, possiamo osservare che questa probabilità si riduce (grazie alla semplificazione dei termini $\frac{1}{5}$) al rapporto tra $p_Y(y)$ e $(F_Y(5) - F_Y(0))$, di nuovo, in maniera non sorprendente. Numericamente le probabilità si possono calcolare in R come `(1/5*dpois(1:5,2.5))/(1/5*(ppois(5,2.5)-ppois(0,2.5)))`, ricavando

$$p_{X,Y}((x,y) \mid S=5) \approx \begin{cases} 0.2343 & (x,y) = (4,1) \\ 0.2929 & (x,y) = (3,2) \\ 0.2440 & (x,y) = (2,3) \\ 0.1525 & (x,y) = (1,4) \\ 0.0763 & (x,y) = (0,5) \end{cases}$$

A.6. APPELLO D'ESAME (2024/07/15)

Problema 18. Di una variabile aleatoria X sappiamo che la funzione di densità è nulla sui reali negativi e ha la forma $f(x) = \frac{c}{(1+x)^4}$ per $x \geqslant 0$.

1. Determinare i possibili valori di c.
2. Calcolare moda e mediana di X.
3. Determinare la legge di $Y = X+1$ e di $Z = aX$. Le variabili aleatorie Y e Z sono tra loro indipendenti?

Solution. La prima domanda fa riferimento a una proprietà ben nota delle densità di probabilita: l'integrale della funzione di densità su $\mathbb{R}$ deve valere 1. In altre parole

$$\begin{aligned} \frac{1}{c} &= \int_0^{+\infty} (1+x)^{-4}\,\mathrm{d}x \\ &= \left[-\frac{1}{3}(1+x)^{-3}\right]_0^{+\infty} = 0 - \left(-\frac{1}{3}\right), \end{aligned}$$

da cui $c = 3$.

Passiamo alla seconda domanda. Per la mediana si tratta di trovare m_X tale che

$$\int_0^{m_X} \frac{3}{(1+x)^4}\,\mathrm{d}x = \int_{m_X}^{+\infty} \frac{3}{(1+x)^4}\,\mathrm{d}x = \frac{1}{2}.$$

Possiamo risolvere partendo dal primo integrale (per il quale dovremo calcolare la primitiva in entrambi gli estremi) o dal secondo. Il risultato, come auspicabile, non cambia. Abbiamo

$$\left[-\frac{1}{(1+x)^3}\right]_0^{m_X} = \frac{1}{2} \qquad \left[-\frac{1}{(1+x)^3}\right]_{m_X}^{+\infty} = \frac{1}{2}$$

da cui otteniamo

$$1-\frac{1}{(1+m_X)^3}=\frac{1}{2} \qquad\qquad \frac{1}{(1+m_X)^3}=\frac{1}{2}$$

che ha come soluzione $m_X=\sqrt[3]{2}-1$. Per la moda potremmo pensare di derivare e porre la derivata uguale a zero, ma questa strada è destinata al fallimento, infatti la funzione è monotona decrescente in $[0,+\infty)$ e il suo massimo è nell'estremo $x=0$.

Arriviamo alla terza domanda. Sia Y sia Z sono trasformazioni lineari della variabile aleatoria X, quindi possiamo individuare le loro densità come

$$f_Y(y)=f_X(y-1)=\frac{3}{y^4}, y\geqslant 1 \qquad\qquad f_Z(z)=\frac{1}{|a|}f_X\left(\frac{z}{a}\right)=\frac{3}{|a|\left(1+\frac{z}{a}\right)^4}, z\,a\geqslant 0$$

supponendo $a\neq 0$, altrimenti $Z\equiv 0$. Per quanto riguarda la loro indipendenza, osserviamo che, per $a\neq 0$ esse sono l'una trasformazione lineare dell'altra, quindi non possono essere indipendenti. Per $a=0$ sono indipendenti.

Problema 19. La dirigenza di una fabbrica di cioccolato, nonostante le pressioni del temibile sindacato OLU, non vuole introdurre la settimana lavorativa ridotta, sostenendo che la maggior parte di chi lavora non desidera un tale cambiamento. Per risolvere la questione viene effettuato un sondaggio su una piccola parte della forza lavoro: a 380 dipendenti viene chiesto se siano favorevoli o meno alla settimana corta. Delle risposte, 191 sono favorevoli alla nuova politica, le altre contrarie.

Sia la dirigenza, sia il sindacato convocano una conferenza stampa ed entrambe le parti dichiarano che i dati mostrano, a un livello di significatività del 5%, che la maggior parte di chi lavora nella fabbrica è dalla loro parte.

1. Quale delle due parti ha ragione da un punto di vista statistico?
2. Determinare il numero minimo di dipendenti favorevoli affinché ci sia evidenza statistica che la forza lavoro sia favorevole alla settimana corta.
3. Determinare il numero massimo di dipendenti favorevoli affinché ci sia evidenza statistica che la forza lavoro non sia favorevole alla settimana corta.
4. Commentare i risultati ottenuti.

Solution. Siamo nella classica situazione di test statistici a una coda con diversa scelta dell'ipotesi nulla da parte delle due fazioni. Per quanto riguarda la prima domanda, la risposta è "nessuna delle due/entrambe le parti", perché entrambe hanno scelto come ipotesi nulla quella a proprio favore: contrarietà alla settimana corta per la dirigenza e situazione opposta per il sindacato.

Il modo in cui possiamo rappresentare matematicamente il problema è come test d'ipotesi sul parametro di una popolazione Bernoulliana, chiedendoci se sia maggiore o minore di 0.5.

Cominciamo a impostare il codice R:

```
# Dati del sondaggio
n <- 380
successes <- 191

# Parametri
p_hat <- successes / n
p0 <- 0.5
# Livello di significatività assegnato
alpha <- 0.05
```

Ora iniziamo la parte di impostazione dei test veri e propri. Iniziamo con il punto di vista della dirigenza

```
# Punto di vista della dirigenza
#
```

```
# Test di ipotesi unilaterale (destra)
# H0: p <= 0.5
# H1: p > 0.5
#
# In questo modo in assenza di evidenza in senso contrario si
#tiene la settimana lunga


# Calcolo del test di proporzione
prop_test <- binom.test(successes, n, p = p0,
                            alternative = "greater")
# Possibile alternativa con test approssimato e correzione
# prop_test <- prop.test(successes, n, p = p0,
                       alternative = "greater", correct = TRUE)

print(prop_test)

p_value <- prop_test$p.value

# Risultato del test
if (p_value < alpha) {
  result <- "Rigettiamo l'ipotesi nulla."
} else {
  result <- "Accettiamo l'ipotesi nulla."
}

print(result)
```

Passiamo al punto di vista del sindacato.

```
# Punto di vista del sindacato
#
# Test di ipotesi unilaterale (sinistro)
# H0: p >= 0.5
# H1: p < 0.5
#
# In questo modo in assenza di evidenza in senso contrario si
#tiene la settimana lunga

# Calcolo del test di proporzione
prop_test <- binom.test(successes, n, p = p0,
                            alternative = "less")

# Possibile alternativa con test approssimato e correzione
# prop_test <- prop.test(successes, n, p = p0,
                         alternative = "less", correct = TRUE)

print(prop_test)

p_value <- prop_test$p.value

# Risultato del test
if (p_value < alpha) {
```

```
  result <- "Rigettiamo l'ipotesi nulla."
} else {
  result <- "Accettiamo l'ipotesi nulla."
}

print(result)
```

Cosa succede? I dati raccolti sono talmente prossimi alla soglia che non abbiamo evidenza statistica né in un senso né nell'altro: non riusciremo mai, con questi dati, a rigettare l'ipotesi nulla, qualunque sia il verso.

Passiamo così alle domande successive: ci chiediamo quanti dovrebbero essere i favorevoli come minimo/massimo per avere evidenza statistica che c'è una maggioranza per/contro la settimana corta. Per farlo iteriamo sui possibili favorevoli (crescendo/decrescendo) finché non otteniamo un p-dei-dati sotto la soglia richiesta. Cominciamo dal numero minimo che garantirebbe evidenza statistica a favore della settimana corta:

```
find_critical_successes <- function(n, p0, alpha) {
  for (x in 0:n) {
    p_value <- binom.test(x, n, p = p0,
                          alternative = "greater")$p.value
    if (p_value < alpha) {
      return(x)
    }
  }
  return(NULL)
}

critical_successes <- find_critical_successes(n, p0, alpha)
cat("\nIl numero minimo di successi è:", critical_successes,
"\n")
```

Infine facciamo il conto del massimo di favorevoli compatibile con l'evidenza statistica contro la settimana corta

```
find_critical_successes2 <- function(n, p0, alpha) {
  for (x in n:0) {
    p_value <- binom.test(x, n, p = p0,
                          alternative = "less")$p.value
    if (p_value < alpha) {
      return(x)
    }
  }
  return(NULL)
}

critical_successes2 <- find_critical_successes2(n, p0, alpha)
cat("\nIl numero massimo di successi è:", critical_successes2,
"\n")
```

Problema 20. La professoressa Mari riceve, in media, 4 email al giorno. Indicando con i numeri interi positivi $k \in \{1,\ldots,7\}$ i giorni della settimana, la variabile aleatoria X_k rappresenta il numero di email ricevute nel giorno k. Assumiamo che le variabili X_k siano indipendenti e identicamente distribuite secondo una legge di Poisson.

1. Le ipotesi fatte sulle X_k sono ragionevoli?

2. Sapendo che il terzo giorno della settimana ($k=3$) ha ricevuto almeno una mail, qual è la probabilità che ne riceva esattamente altre 2 nello stesso giorno?
3. Qual è il valore atteso di email ricevute nel corso di una settimana?
4. Sapendo che nel weekend (giorni 6 e 7) ha ricevuto complessivamente $n \in \mathbb{N}$ email, qual è il valore atteso del numero di email ricevute la domenica?
5. Qual è il coefficiente di correlazione tra le variabili aleatorie che contano le mail ricevute sabato e domenica e quelle ricevute domenica e lunedì?

Solution. Procediamo per punti.

1. Le ipotesi sono ragionevoli. Osserviamo innanzitutto che conosciamo solamente il numero medio di mail ricevute al giorno, quindi se riusciamo a non fare altre ipotesi sui parametri siamo più felici. Inoltre cerchiamo una variabile discreta che abbia valori non negativi non limitati dall'alto. Dei modelli studiati la Poisson sembra la più promettente. Per quanto riguarda l'ipotesi di indipendenza, non abbiamo motivo per supporre che le mail ricevute in un giorno siano legate a quelle ricevute in un altro.
2. Dal momento che per ogni k $X_k \sim \text{Pois}(4)$, possiamo scrivere
$$\begin{aligned} P(X_k = 1+2 \mid X_k \geqslant 1) &= \frac{P(X_k = 3 \cap X_k \geqslant 1)}{P(X_k \geqslant 1)} \\ &= \frac{P(X_k = 3)}{1 - P(X_k < 1)} = \frac{P(X_k = 3)}{1 - P(X_k = 0)} \\ &= \frac{4^3 \mathrm{e}^{-4}}{3!} \left(1 - \frac{4^0 \mathrm{e}^{-4}}{0!}\right)^{-1} \\ &= \frac{4^3 \mathrm{e}^{-4}}{3!\,(1-\mathrm{e}^{-4})} \approx 0.199. \end{aligned}$$
3. Il numero N di email ricevute in una settimana è $N = \sum_{k=1}^{7} X_k$. Siccome le X_k sono Poissoniane indipendenti e identicamente distribuite di parametro 4, $N \sim \text{Pois}(7 \cdot 4)$ e siccome il parametro di una Poissoniana coincide con la sua media, la risposta cercata è 28. Osserviamo che non era necessario sfruttare indipendenza e riproducibilità, per calcolare la media: basta osservare che la media di una somma è la somma delle medie e che ciascuna delle X_k ha media 4 (per ipotesi).
4. Il problema chiede il valore atteso di X_7 condizionato a $X_6 + X_7 = n$. In altre parole dobbiamo ricavare la densità di X_7 condizionata a $X_6 + X_7$. Abbiamo
$$f_{X_7 \mid X_6 + X_7}(k \mid n) = \frac{f_{X_7, X_6 + X_7}(k, n)}{f_{X_6 + X_7}(n)}$$
e per quanto osservato al punto precedente, $X_6 + X_7 \sim \text{Pois}(8)$, quindi abbiamo il termine al denominatore. Per il numeratore usiamo la definizione di densità congiunta discreta
$$f_{X_7, X_6 + X_7}(k, n) = P(X_7 = k, X_6 + X_7 = n) = P(X_7 = k, X_6 = n - k) = P(X_7 = k)\, P(X_6 = n - k)$$
in cui abbiamo sfruttato, nell'ultimo passaggio, l'ipotesi di indipendenza tra le X_k. Dobbiamo però prestare attenzione nel momento in cui andiamo a usare la forma esplicita delle densità di X_6 e X_7 che entrambi gli argomenti siano non negativi: deve essere $k \geqslant 0$ e $n - k \geqslant 0$, ossia $0 \leqslant k \leqslant n$, con $k \in \mathbb{N}$. Esplicitando in questo caso
$$f_{X_7, X_6 + X_7}(k, n) = \frac{4^k \mathrm{e}^{-4}}{k!} \frac{4^{n-k} \mathrm{e}^{-4}}{(n-k)!} = \frac{4^n \mathrm{e}^{-8}}{k!\,(n-k)!}.$$
Allora
$$f_{X_7 \mid X_6 + X_7}(k \mid n) = \frac{4^n \mathrm{e}^{-8}}{k!\,(n-k)!} \frac{n!}{8^n \mathrm{e}^{-8}} = \binom{n}{k} \left(\frac{4}{8}\right)^n = \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k},$$

cioè è una variabile aleatoria binomiale di parametri n e $\frac{1}{2}$. (Nota: questo fatto era stato sviluppato nelle note del corso, ma si poteva anche recuperare qui come parte dell'esercizio.) In particolare il valore atteso è $\frac{n}{2}$.

5. Cominciamo usando la definizione di covarianza e le sue proprietà, ricordando che le X_k sono indipendenti e Poissoniane di parametro 4

$$\begin{aligned}\operatorname{Cov}(X_6+X_7, X_7+X_1) &= \operatorname{Cov}(X_6, X_7+X_1)+\operatorname{Cov}(X_7, X_7+X_1)\\ &= \operatorname{Cov}(X_6, X_7)+\operatorname{Cov}(X_6, X_1)+\operatorname{Cov}(X_7, X_7)+\operatorname{Cov}(X_7, X_1)\\ &= 0+0+\operatorname{Var}(X_7)+0=4.\end{aligned}$$

Siccome $\rho(X_6+X_7, X_7+X_1)=\frac{\operatorname{Cov}(X_6, X_7+X_1)+\operatorname{Cov}(X_7, X_7+X_1)}{\sqrt{\operatorname{Var}(X_6+X_7)\operatorname{Var}(X_7+X_1)}}$, non ci resta che calcolare le varianze al denominatore. Ma siccome X_k+X_h con $k\neq h$ sono Poissoniane di media (e quindi parametro) 8, anche le varianze sono entrambe 8, pertanto

$$\rho(X_6+X_7, X_7+X_1)=\frac{4}{8}=\frac{1}{2}.$$

A.7. APPELLO D'ESAME (2024/08/29)

Problema 21. Una strada in uscita dalla Città Senza Nome (CSN) è tristemente nota per la frequenza degli incendi che si sviluppano ai suoi margini (anche a causa dell'incuria dei guidatori fumatori). La CSN decide pertanto di costruire una stazione dei vigili del fuoco sulla strada, ma vuole farlo in maniera ottimale.

Dai dati raccolti negli anni si ipotizza che nel tratto di competenza della CSN della strada (lungo L leghe) la densità di probabilità di incendi sia minima all'inizio della strada e cresca linearmente con coefficiente angolare 0.5 fino al termine.

1. Quanto vale il minimo h della densità di probabilità?
2. Quanto è lungo al più il tratto di competenza della CSN della strada?
3. A che punto deve essere costruita la stazione se vogliamo che ci sia la medesima probabilità che i vigili debbano intervenire in ciascuna delle due direzioni (verso la città e verso i confini comunali)?
4. Uscita dal territorio comunale la strada continua all'infinito nel territorio del Paese Senza Nome (PSN). Anche il PSN decide di costruire una stazione dei vigili del fuoco lungo la parte di propria competenza (ossia la parte non "coperta" dalla CSN). Dai dati raccolti la densità di probabilità degli incendi è, lungo la strada a partire dai confini comunali della CSN) un'esponenziale di parametro $\lambda=\frac{1}{4}$. Dove deve costruire il PSN la centrale dei vigili del fuoco se vuole minimizzare la distanza media percorsa dai mezzi per intervenire a spegnere un incendio?
5. Credi che sia meglio utilizzare la distanza media o la misura usata al punto 3 nella scelta della posizione della stazione? (Giustifica matematicamente la tua scelta di modello.)

Solution. È possibile ricavare h dalle proprietà della densità di probabilità. L'area sottesa da un lato deve essere pari a 1, dall'altro è l'area di un trapezio di basi h e $0.5L+h$ e altezza L, quindi $h=\frac{1}{L}-\frac{L}{4}$.

Osserviamo che siccome $h\geqslant 0$ abbiamo un vincolo su L (che a sua volta è positiva):

$$\frac{1}{L}-\frac{L}{4}\geqslant 0 \qquad\Longrightarrow\qquad L^2\leqslant 4 \qquad\Longrightarrow\qquad L\leqslant 2$$

Passiamo ora al punto 3. Se matematizziamo la richiesta, stiamo cercando l tale che (indicando con X il punto in cui si sviluppa un incendio) $P(X\leqslant l)=P(X\geqslant l)=0.5$, sfruttando il fatto che X è assolutamente continua. Dobbiamo quindi calcolare la mediana di X. Possiamo farlo in molti modi. Uno relativamente semplice è osservare che "tagliando" in l otteniamo un trapezio simile a quello di partenza di cui vogliamo che l'area sia la metà (ossia 0.5). L'area di questo trapezio è

$$0.5 \overset{!}{=} \left(h + \frac{l}{2} + h\right)\frac{l}{2} \qquad \Longrightarrow \qquad 2 = l\,(l+4h) = l^2 + \left(\frac{4}{L} - L\right)l$$

da cui, risolvendo l'equazione di secondo grado e tenendo solamente la radice che giace nell'intervallo $[0, L]$ abbiamo $l = \frac{L^2 - 4 + \sqrt{L^4+16}}{2L}$.

Per il punto 4 dobbiamo invece minimizzare la media della distanza. La media della distanza se la stazione è a l leghe dai confini comunali della CSN è

$$\begin{aligned}
E[|X-l|] &= \int_0^{+\infty} |x-l| \frac{1}{4} \mathrm{e}^{-\frac{1}{4}x} \mathrm{d}x \\
&= \int_0^{l} (l-x) \frac{1}{4} \mathrm{e}^{-\frac{1}{4}x} \mathrm{d}x + \int_l^{+\infty} (x-l) \frac{1}{4} \mathrm{e}^{-\frac{1}{4}x} \mathrm{d}x \\
&= \frac{l}{4} \int_0^{l} \mathrm{e}^{-\frac{1}{4}x} \mathrm{d}x - \frac{1}{4} \int_0^{l} x \mathrm{e}^{-\frac{1}{4}x} \mathrm{d}x + \frac{1}{4} \int_l^{+\infty} x \mathrm{e}^{-\frac{1}{4}x} \mathrm{d}x - \frac{l}{4} \int_l^{+\infty} \mathrm{e}^{-\frac{1}{4}x} \mathrm{d}x \\
&= l - l\mathrm{e}^{-\frac{l}{4}} - 4 + (4+l)\mathrm{e}^{-\frac{l}{4}} + (4+l)\mathrm{e}^{-\frac{l}{4}} - l\mathrm{e}^{-\frac{l}{4}} \\
&= l + 8\mathrm{e}^{-\frac{l}{4}} - 4.
\end{aligned}$$

Prendiamo la derivata di questa funzione (che è convessa) e poniamola uguale a 0 per trovarne il minimo:

$$g'(l) = 1 - 2\mathrm{e}^{-\frac{l}{4}}$$

che si annulla in $l = -4\log(0.5) = 4\log(2) \approx 2.77$. Questa è la mediana della distribuzione cercata (sappiamo che la mediana può essere caratterizzata come l'argomento che minimizza la distanza euclidea media, mentre la media è l'argomento che minimizza la distanza quadratica media)e le due strategie sono del tutto equivalenti. Potrebbe essere sensato mettere la stazione dei vigili in corrispondenza della media della distribuzione, se pensiamo che un incendio molto lontano (che richiede quindi più tempo a essere raggiunto) sia significativamente più grave (magari perché superato un certo tempo diventa sempre meno gestibile, mentre intervenire in tempi brevi anche se leggermente diversi non cambia la situazione).

Problema 22. Il corso di Fisica 1 tenuto dal temuto professor Tassi è frequentato da 70 studenti. Ogni volta che c'è un ricevimento, ciascuno studente e ciascuna studentessa decide di approfittare di questa occasione con una probabilità uguale a 0.01, indipendentemente dalle altre persone.

1. In media quante persone si presentano a un dato ricevimento?
2. Con che probabilità non si presenta nessuno a un dato ricevimento?
3. Durante l'anno il professor Tassi offre 50 ricevimenti. Se indichiamo con V i ricevimenti in cui non si presenta alcuna persona, quali sono valore atteso e varianza di V?
4. Stimare la probabilità che in almeno 30 ricevimenti non si presenti alcuna persona spiegando le ipotesi fatte e confrontando con il risultato "esatto" ottenuto in R.

Solution. Per ogni indice $k \in \{1, \dots, 70\}$ chiamiamo Y_k la variabile aleatoria indicatrice dell'evento "lo/la studente k è andato/a a ricevimento". Dai dati forniti nel testo sappiamo che le Y_k sono i.i.d. con $Y_k \sim \mathrm{bin}(1, 0.01)$. Allora il numero di studenti che prendono parte al ricevimento è

$$Y = \sum_{k=1}^{70} Y_k \sim \mathrm{bin}(70, 0.01),$$

quindi $E[Y] = 0.7$. La probabilità che non si presenti alcuna persona è $P(Y=0) =: p$ che può essere calcolata (approssimativamente) usando R: `dbinom(x=0,size=70, prob=0.01)` che restituisce 0.4948387.

Anche la variabile aleatoria V ha una distribuzione binomiale, questa volta di parametri 50 (numero dei tentativi) e $p \approx 0.4948387$ (probabilità di successo, dove il "successo" è il non avere persone a ricevimento). Allora $E[V] = 50p \approx 24.74$ e $\mathrm{Var}[V] = 50p(1-p) \approx 12.5$.

Se calcoliamo la probabilità richiesta in R, possiamo usare la funzione `pbinom`, scegliendo la coda giusta: `pbinom(q = 29, size = 50, prob = dbinom(x=0,size=70, prob=0.01), lower.tail = FALSE)` che restituisce 0.08895088. Siccome la richiesta è di stimarla, possiamo usare il Teorema Limite Centrale e un'approssimazione normale della variabile aleatoria binomiale V. Dal momento che V è una variabile aleatoria discreta sarà opportuno usare la correzione di continuità:

$$\begin{aligned} P(V \geqslant 30) &= 1 - P(V \leqslant 29) = 1 - P(V \leqslant 29.5) \\ &= 1 - P\left(\sum_{j=1}^{50} V_j \leqslant 29.5\right) \\ [\text{TLC}] &\approx 1 - \Phi\left(\frac{29.5 - 50\,E[V_1]}{\sqrt{50\,\text{Var}[V_1]}}\right) = 1 - \Phi\left(\frac{29.5 - E[V]}{\sqrt{\text{Var}[V]}}\right) \end{aligned}$$

e possiamo calcolarlo in R `1 - pnorm(q = (29.5-50 * dbinom(x=0,size=70, prob=0.01))/sqrt(50 * dbinom(x=0,size=70, prob=0.01) * (1 - dbinom(x=0,size=70, prob=0.01))))` oppure con `1 - pnorm(q = 29.5, mean = 50 * dbinom(x=0,size=70, prob=0.01), sd = sqrt(50 * dbinom(x=0,size=70, prob=0.01) * (1 - dbinom(x=0,size=70, prob=0.01))))` che danno approssimativamente 0.08917443. Osserviamo che non tenere conto della correzione di continuità causa in questo caso errori abbastanza significativi (fare la prova per credere).

Problema 23. Un ingegnere automobilistico sospetta che il consumo medio di carburante di certe auto sia diverso da quello dichiarato (pari a 17 mpg, miglia per gallone). I dati sono raccolti nel dataset `mtcars` disponibile in R. Rispondere alle seguenti richieste, esplicitando eventuali ipotesi necessarie e discutendo se sono soddisfatte o meno.

1. Dare una stima puntuale del consumo medio.
2. Dare una stima intervallare a un livello di fiducia del 98% del consumo medio.
3. Impostare un opportuno test statistico al 2% di significatività, commentando le scelte fatte e i risultati ottenuti.

Solution. Per quanto riguarda il primo punto la richiesta è semplicemente di dare una stima puntuale della media, che non richiede ipotesi sulla popolazione, se non l'indipendenza degli elementi del campione. Abbiamo

```
data(mtcars)
summary(mtcars)
mean(mtcars$mpg)
```

in cui già nell'output della funzione summary abbiamo il valore della media, 20.09062.

Per calcolare l'intervallo di fiducia dobbiamo controllare che siano soddisfatte (almeno approssimativamente) le ipotesi di normalità della popolazione:

```
qqnorm(mtcars$mpg)
qqline(mtcars$mpg, col = "red")
hist(mtcars$mpg)
```

La popolazione non è propriamente normale, anche per il numero relativamente ridotto di elementi che costituiscono il campione, tuttavia la possiamo considerare approssimativamente normale. Per il calcolo dell'intervallo di fiducia possiamo scrivere del codice dedicato, usando la statistica t dal momento che non è nota la varianza, di cui possiamo solamente avere una stima campionaria. Siccome nel punto successivo vogliamo fare un test d'ipotesi a sua volta basato sulla statistica t, possiamo fare entrambe le cose con un'unica funzione: `t.test(mtcars$mpg, mu = 17, conf.level = 0.98, alternative = "two.sided")` che ci restituisce quanto segue

```
One Sample t-test

data:  mtcars$mpg
t = 2.9008, df = 31, p-value = 0.006788
alternative hypothesis: true mean is not equal to 17
98 percent confidence interval:
 17.47733 22.70392
sample estimates:
mean of x
 20.09062
```

Nel leggere i risultati vediamo che l'intervallo richiesto è $(17.47733, 22.70392)$ e che dato il p-dei-dati pari a 0.006788 possiamo rifiutare l'ipotesi nulla che la media sia uguale a 17 e accettare l'alternativa che sia differente.

In questo test d'ipotesi l'alternativa è simmetrica, ma avendo visto che la media campionaria è circa 20 avremmo forse potuto impostare un test che mostrasse come la media sia maggiore di quanto dichiarato, ossia con ipotesi alternativa $\mu > \mu_0 = 17$. In questo caso il codice non è molto diverso, `t.test(mtcars$mpg, mu = 17, conf.level = 0.98, alternative = "greater")` e come p-dei-dati abbiamo 0.003394 che ci spinge anche in questo caso a rigettare l'ipotesi nulla e a sostenere l'alternativa, ossia che la media sia significativamente maggiore di 17.

A.8. PROVA IN ITINERE (2025/04/17)

Problema 24. Ogni persona ha, per quanto riguarda il colore degli occhi, un aspetto genetico (genotipo) e un aspetto fenomenologico (fenotipo). I due sono legati tra loro.

I genotipi associati al colore degli occhi sono $\{MM, AA, MA, AM\}$. Si sviluppa il fenotipo A (i.e. si hanno gli occhi azzurri) se e solo se si ha il genotipo AA. Si sviluppa il fenotipo M (occhi marroni) se e solo se si ha uno tra i genotipi $\{MA, AM, MM\}$.

Secondo le leggi di Mendel, il genotipo della prole (biologica) di due individui è equamente distribuito tra le possibili quattro combinazioni $(X_i Y_j)$, dove X_i e Y_i rappresentano, rispettivamente, il primo e il secondo elemento del genotipo del genitore G_i $(i \in \{1,2\})$. A titolo di esempio, si considerino le seguenti tabelle.

	M	M
A	AM	AM
M	MM	MM

	M	A
A	AM	AA
M	MM	MA

Tabella A.1. Sinistra: possibili esiti per genitori con genotipi AM e MM. Il figlio avrà genotipo AM con probabilità $\frac{2}{4}$, mentre avrà genotipo MM con probabilità $\frac{2}{4}$. Destra: possibili esiti per genitori con genotipi AM e MA. Il figlio avrà genotipo AM, AA, MM, MA con probabilità $\frac{1}{4}$.

Supponiamo di vivere in una popolazione omogenea, ovvero che $P(AA) = P(AM) = P(MA) = P(MM) = 1/4$, per ogni individuo della popolazione.

Denotiamo G_1 e G_2 i genitori.

1. Se un individuo ha gli occhi azzurri e G_1 ha gli occhi azzurri, qual è la probabilità che anche G_2 abbia gli occhi azzurri?
2. Se un individuo ha gli occhi marroni e G_1 ha gli occhi azzurri, qual è la probabilità che G_2 abbia gli occhi marroni?
3. Se un individuo ha gli occhi azzurri, qual è la probabilità che almeno uno dei due genitori abbia gli occhi marroni?

Solution. Per semplicità, chiamiamo $\mathcal{G} := \{MM, MA, AM, AA\}$ l'insieme dei genotipi e con $\mathcal{M} := \{MM, MA, AM\}$ e $\mathcal{A} := \{AA\}$ gli insiemi dei genotipi con fenotipo marrone e azzurro, rispettivamente. Notiamo che $P(G_i \in \mathcal{A}) = \frac{1}{4}$ e $P(G_i \in \mathcal{M}) = \frac{3}{4}$, per ogni $i = 1, 2$. Inoltre, poiché il genotipo nei genitori G_1 e G_2 è indipendente, abbiamo che il condizionamento al genotipo dell'altro genitore è ininfluente. Quindi, per esempio, $P(G_1 \in \mathcal{A} \mid G_2 \in \mathcal{M}) = P(G_1 \in \mathcal{A})$ e $P(G_2 \in \mathcal{A} \mid G_1 \in \mathcal{A}) = P(G_2 \in \mathcal{A})$.

Partiamo dalla prima domanda. Denotiamo con T il fenotipo dell'individuo. Calcoliamo la probabilità che $G_2 \in \mathcal{A}$ ossia che abbia fenotipo A, sapendo che sia l'individuo sia G_1 sono in $\mathcal{A}$, ossia hanno fenotipo A.

Abbiamo

$$\begin{aligned} P(G_2 \in \mathcal{A} \mid G_1 \in \mathcal{A}, T \in \mathcal{A}) &= \frac{P(G_2 \in \mathcal{A}, G_1 \in \mathcal{A}, T \in \mathcal{A})}{P(G_1 \in \mathcal{A}, T \in \mathcal{A})} \\ &= \frac{P(T \in \mathcal{A} \mid G_2 \in \mathcal{A}, G_1 \in \mathcal{A})\, P(G_2 \in \mathcal{A}, G_1 \in \mathcal{A})}{P(G_1 \in \mathcal{A}, T \in \mathcal{A})} \\ &= \frac{P(T \in \mathcal{A} \mid G_2 \in \mathcal{A}, G_1 \in \mathcal{A})\, P(G_2 \in \mathcal{A} \mid G_1 \in \mathcal{A})\, P(G_1 \in \mathcal{A})}{P(T \in \mathcal{A} \mid G_1 \in \mathcal{A})\, P(G_1 \in \mathcal{A})} \\ &= \frac{P(T \in \mathcal{A} \mid G_2 \in \mathcal{A}, G_1 \in \mathcal{A})\, P(G_2 \in \mathcal{A} \mid G_1 \in \mathcal{A})}{\sum_{g \in \mathcal{G}} P(T \in \mathcal{A} \mid G_2 = g, G_1 \in \mathcal{A})\, P(G_2 = g \mid G_1 \in \mathcal{A})} \\ &= \frac{P(T \in \mathcal{A} \mid G_2 \in \mathcal{A}, G_1 \in \mathcal{A})\, P(G_2 \in \mathcal{A})}{\sum_{g \in \mathcal{G}} P(T \in \mathcal{A} \mid G_2 = g, G_1 \in \mathcal{A})\, P(G_2 = g)} \end{aligned}$$

in cui abbiamo usato la definizione di probabilità condizionata, la fattorizzazione di $P(\cdot \mid G_1 \in \mathcal{A})$ rispetto alla partizione $G_2 = g$ e infine l'ipotesi di indipendenza tra G_1 e G_2. La scelta della fattorizzazione a denominatore è suggerita dalla forma del denominatore, perché per calcolare entrambe le quantità usiamo i medesimi ingredienti.

Ora ricordiamo che i valori che può assumere $g \in \mathcal{G}$ sono $\{MM, MA, AM, AA\}$ e per ciascuno di essi abbiamo (ricordando che $G_1 \in \mathcal{A}$ significa $G_1 = AA$)

g	$P(T \in \mathcal{A} \mid G_2 = g, G_1 \in \mathcal{A})$	$P(G_2 = g)$
MM	0	$\frac{1}{4}$
MA	$\frac{1}{2}$	$\frac{1}{4}$
AM	$\frac{1}{2}$	$\frac{1}{4}$
AA	1	$\frac{1}{4}$

Tabella A.2. La partizione espressa in forma tabulare.

Allora

$$P(G_2 \in \mathcal{A} \mid G_1 \in \mathcal{A}, T \in \mathcal{A}) = \frac{\frac{1}{4} \cdot 1}{0 \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{4} + 1 \cdot \frac{1}{4}} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}.$$

Passiamo alla seconda domanda. In questo caso la domanda è, una volta che la formalizziamo, molto banale: se entrambi i genitori hanno gli occhi azzurri, anche il figlio/a avrà occhi azzurri. Avendo l'individuo in questione gli occhi marroni, sicuramente almeno uno dei genitori ha gli occhi marroni, quindi $P(G_2 \in \mathcal{M} \mid T \in \mathcal{M}, G_1 \in \mathcal{A}) = 1$.

La terza domanda è un'applicazione immediata del teorema di Bayes:

$$\begin{aligned} P(G_1 \in \mathcal{M} \cup G_2 \in \mathcal{M} \mid T \in \mathcal{A}) &= 1 - P(G_1 \in \mathcal{A}, G_2 \in \mathcal{A} \mid T \in \mathcal{A}) \\ &= 1 - \frac{P(T \in \mathcal{A} \mid G_2 \in \mathcal{A}, G_1 \in \mathcal{A})\, P(G_1 \in \mathcal{A}, G_2 \in \mathcal{A})}{P(T \in \mathcal{A})} \\ &= 1 - \frac{1 \cdot \left(\frac{1}{4} \cdot \frac{1}{4}\right)}{\frac{1}{4}} = \frac{3}{4}. \end{aligned}$$

Qualsiasi risposta equivalente, ottenuta con ragionamenti analoghi, è corretta. In particolare ci sono molte altre possibili rappresentazioni, anche diverse da quella proposta qui.

Problema 25. Sia $f_X\colon \mathbb{R} \to \mathbb{R}$ definita come

$$f_X(x) := \begin{cases} c\,(x^2 + \alpha x) & \text{se } x \in [0,1], \\ 0 & \text{altrimenti,} \end{cases}$$

ove $c \geq 0$ e $\alpha \in \mathbb{R}$.

1. Per quali valori dei parametri α, c risulta che f_X è una densità di probabilità?
2. Sia X la variabile aleatoria con densità f_X. Determina, se esistono, i valori dei parametri α e c per cui la media di X vale 0.7.
3. Utilizzando i valori ottenuti al punto 2., definiamo $Y := X^2$. Determina la funzione densità f_Y e la funzione di ripartizione F_Y di Y.
4. Quanto valgono $P\big(Y \geq \frac{\pi}{3}\big)$ e $P\big(Y \in [0, \frac{1}{2}]\big)$?

Solution. Anzitutto, affinché f_X sia una densità di probabilità, si deve avere $f_X(x) \geq 0$ $(\forall x \in \mathbb{R})$ e che $\int_{\mathbb{R}} f_X(x)\,\mathrm{d}x = 1$. Dalla seconda condizione, otteniamo subito che $c \neq 0$, quindi $c > 0$. Notiamo, inoltre, che se $\alpha < 0$, allora $f_X'(0) < 0$, quindi, essendo f_X differenziabile, esiste un intorno destro di 0 in cui $f_X(x) < 0$. Di conseguenza, $\alpha \geq 0$. Notiamo ora che, se $\alpha \geq 0$, $f_X(x) \geq 0$ in $[0,1]$, in quanto somma di funzioni non-negative. Quindi la prima condizione è soddisfatta per ogni valore di $\alpha \geq 0$. Consideriamo ora la seconda condizione:

$$1 \overset{!}{=} \int_{\mathbb{R}} f_X(x)\,\mathrm{d}x = \int_0^1 c\,(x^2 + \alpha x)\,\mathrm{d}x = c\left[\frac{x^3}{3} + \alpha\frac{x^2}{2}\right]_0^1 = c\left(\frac{1}{3} + \frac{\alpha}{2}\right),$$

da cui si ottiene $c = \frac{6}{2+3\alpha}$. Notiamo che, per ogni valore di $\alpha \geq 0$, abbiamo $c > 0$. Pertanto, f_X è una densità di probabilità se e solo se valgono entrambe le condizioni

- $\alpha \geq 0$,
- $c = \frac{6}{2+3\alpha}$.

Calcoliamo la media di X:

$$\begin{aligned} E[X] &= \int_0^1 x \cdot \frac{6}{2+3\alpha}\,(x^2 + \alpha x)\,\mathrm{d}x \\ &= \frac{6}{2+3\alpha}\int_0^1 x^3 + \alpha x^2\,\mathrm{d}x \\ &= \frac{6}{2+3\alpha}\left[\frac{x^4}{4} + \alpha\frac{x^3}{3}\right]_0^1 \\ &= \frac{6}{2+3\alpha}\left(\frac{1}{4} + \frac{\alpha}{3}\right) \\ &= \frac{3+4\alpha}{4+6\alpha}. \end{aligned}$$

Imponendo ora la condizione $E[X] \overset{!}{=} 0.7$, otteniamo facilmente che l'unica soluzione è $\alpha = 1$. Di conseguenza, abbiamo

$$f_X(x) = \frac{6}{5}\,(x^2 + x)\,\mathbb{1}_{[0,1]}(x).$$

Per calcolare la densità di $Y = X^2$ usiamo il teorema di trasformazione. Possiamo farlo perché $g(x) = x^2$ è differenziabile ed è strettamente monotona su $[0,1]$. Notiamo che $g^{-1}(y) = \sqrt{y}$. Quindi:

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \cdot \left|\frac{\mathrm{d}}{\mathrm{d}y}\,g^{-1}(y)\right| \\ &= \frac{6}{5}\,(\sqrt{y}^2 + \sqrt{y})\,\mathbb{1}_{[0,1]}(\sqrt{y})\left|\frac{1}{2\sqrt{y}}\right| \\ &= \frac{3}{5}\,(\sqrt{y} + 1)\,\mathbb{1}_{[0,1]}(y) \end{aligned}$$

Ora possiamo calcolare la funzione di ripartizione di Y:

$$\begin{aligned} F_Y(y) &= \int_{-\infty}^{y} f_Y(t)\,\mathrm{d}t \\ &= \int_{-\infty}^{y} \frac{3}{5}\left(\sqrt{t}+1\right)\mathbb{1}_{[0,1]}(t)\,\mathrm{d}t \\ &= \frac{3}{5}\int_0^y \sqrt{t}+1\,\mathrm{d}t \\ &= \frac{3}{5}\left[\frac{2}{3}t^{\frac{3}{2}}+t\right]_0^y \\ &= \frac{3}{5}\left(\frac{2}{3}y^{\frac{3}{2}}+y\right) \\ &= \begin{cases} 0 & \text{se } y \le 0 \\ \frac{3}{5}\left(\frac{2}{3}y^{\frac{3}{2}}+y\right) & \text{se } 0 \le y \le 1 \\ 1 & \text{se } y \le 1. \end{cases} \end{aligned}$$

È sufficiente calcolare: $P\left(Y \ge \frac{\pi}{3}\right) = 1 - F_Y\left(\frac{\pi}{3}\right) = 0$ e

$$P\left(Y \in \left[0, \frac{1}{2}\right]\right) = F_Y\left(\frac{1}{2}\right) - F_Y(0) = \frac{3}{5}\left(\frac{2}{3}2^{-\frac{3}{2}} + \frac{1}{2}\right) = \frac{1}{5\sqrt{2}} + \frac{3}{10}.$$

Problema 26. In una fabbrica di graffette vengono usati due macchinari diversi. Il primo produce 1000 graffette all'ora, mentre il secondo le produce a un ritmo variabile. Entrambi i macchinari producono in media ogni ora 50 graffette difettose. Chiamiamo X e Y le variabili aleatorie che descrivono il numero di graffette difettose prodotte dal primo e dal secondo macchinario, rispettivamente, nell'arco di una certa ora.

1. Che distribuzione possiamo ipotizzare per X e Y? Come mai?
2. Sia Z il numero totale di graffette difettose prodotte in un'ora. Calcolane il momento primo e momento centrato secondo.
3. Ogni giorno un'ora è dedicata al controllo qualità: tutte le graffette prodotte da entrambi i macchinari vengono controllate. Il conteggio determina che ci sono 160 graffette difettose. Quanto è sorprendente questo numero?

Solution. Cominciamo con il determinare il modello. Di entrambe sappiamo che sono modelli discreti e che hanno media 50. Di una delle due variabili sappiamo anche il numero di tentativi, quindi possiamo supporre che sia binomiale di parametri 1000 e 0.05. Della seconda invece abbiamo solamente la media, anche se cattura un fenomeno simile, pur non limitato a priori. È ragionevole quindi pensarlo come una Poisson di parametro 50 che ricordiamo può essere vista come limite di una famiglia di binomiali per cui sia costante la media. Riassumendo:

$$X \sim \mathrm{bin}(1000, 0.05) \qquad Y \sim \mathrm{Pois}(50).$$

Possiamo anche, ragionevolmente, supporle indipendenti: anche se potrebbero esserci problemi o altre cose comuni tra le due, non abbiamo abbastanza dati per sostenere una dipendenza.

Passando a Z, il momento primo o valore atteso o speranza o media della somma è la somma delle medie per la linearità dell'operatore valore atteso, quindi $E[Z] = 100$. Per la varianza (o momento secondo centrato) usiamo invece l'ipotesi di indipendenza e abbiamo $\mathrm{Var}[Z] = \mathrm{Var}[X] + \mathrm{Var}[Y] = 97.5$, usando il fatto che $\mathrm{Var}[X] = np(1-p)$ e $\mathrm{Var}[Y] = \lambda$.

Un possibile approccio alternativo è impostare come nel caso visto sopra e considerare che per i valori che abbiamo dei parametri n e p della binomiale X possiamo approssimare X con una Poisson di parametro (e media) 50. In questo caso assumendo l'indipendenza di X e Y abbiamo che per la riproducibilità delle Poisson la loro somma è una Poisson di parametro 100 (ossia sia media sia varianza uguali a 100).

Per l'ultimo punto possiamo ritenerci sorpresi se la probabilità di vedere un valore almeno così grande (i.e. maggiore o uguale di 160) è molto piccola. Infatti non ci basta guardare il singolo valore, perché se guardiamo la distribuzione molti valori hanno probabilità molto bassa, ma potremmo avere tanti valori e quindi ciascuno con probabilità bassa. Per calcolare $P(X \geqslant 160)$ possiamo adottare almeno due strategie differenti: possiamo stimare la probabilità con le disuguaglianze di Markov o di Chebychev, oppure possiamo fare i conti esplicitamente con R.

Cominciamo con la disuguaglianza di Markov. Abbiamo

$$P(X \geqslant 160) \leqslant \frac{E[X]}{160} = \frac{100}{160} = 62.5\%$$

che possiamo interpretare come "non è un risultato troppo sorprendente". Tuttavia la disuguaglianza di Markov ci dà una stima molto rozza dall'alto e in particolare non tiene conto della larghezza della distribuzione, come misurata dalla varianza o deviazione standard, che pure abbiamo calcolato nei punti precedenti.

Proviamo allora a usare la disuguaglianza di Chebychev, assumendo che la distribuzione sia relativamente simmetrica (cosa che possiamo verificare empiricamente rappresentandola in R, come vediamo più sotto). Abbiamo

$$P(|X - E[X]| \geqslant a) \leqslant \frac{\mathrm{Var}[X]}{a^2},$$

che nel nostro caso diventa

$$P(|X - 100| \geqslant 60) \leqslant \frac{97.5}{3600} \approx 0.02708333.$$

Una stima ben diversa dalla precedente. In realtà se usiamo la simmetria della distribuzione possiamo anche dimezzarla, senza che cambi la sostanza.

Se volessimo fare i conti accurati con R, possiamo calcolarci la densità discreta della somma. Un possibile codice è riportato qui sotto:

```
z <- rep(0,160)
x <- dbinom(0:159, 1000, 0.05)
y <- dpois(0:159, 50)
for(i in 0:159){
  z[i] <- sum(x[0:i]*y[i:0])
}
1-sum(z)
```

che dà come risultato 2.116438e-08.

Nel caso in cui avessimo approssimato X con una Poisson il conto è più semplice: calcoliamo $1 - F_Z(159)$ che calcolata con `ppois(159, 100, FALSE)` dà 2.050698e-08, molto vicino al valore calcolato esplicitamente, a testimonianza della bontà dell'approssimazione.