19/09/2024 — Sesto appello

Nome e Cognome: _____ Matricola: _____

**Indicazioni Importati:** Siamo liberi di utilizzare appunti scritti o stampati. **Non possiamo però utilizzare dispositivi elettronici, comunicare tra di noi o con l'esterno, né passarci del materiale tra di noi**.
**A meno che non venga esplicitamente detto il contrario,** non è necessario calcolare il valore numerico delle soluzioni. A parte questo fatto, **è sempre importante mostrare il ragionamento e le formule usate, ed arrivare ad una risposta esatta**, anche se espressa in funzione di altri operatori (per esempio coefficienti binomiali, radici quadrate, esponenti, ecc). Un esempio: se arriviamo a una espressione del tipo $\binom{10}{5}$, possiamo lasciare questa come risposta senza ulteriori semplificazioni, oppure possiamo fare i calcoli ed arrivare al valore numerico $252$. Invece scrivere solo "252" senza che sia chiaro da dove viene quel numero, **sarà considerata una risposta invalida.**

**NOTA BENE: In alcune di queste soluzioni si riporta il valore numerico delle risposte solo per completezza. Il calcolo di tali valori infatti non è stato oggetto di valutazione in sede di correzione, a meno che non sia stato detto esplicitamente il contrario sulla traccia dell'esercizio corrispondente.**

## 1 Database

Un team di sviluppo deve scegliere due tipi di database diversi tra cinque opzioni disponibili per un nuovo progetto: MongoDB, PostgreSQL, MySQL, Redis e Cassandra. La scelta viene effettuata in base all'expertise dei componenti del team sui vari database. Non conosciamo l'expertise specifica del team, quindi assumiamo che ogni database abbia uguali probabilità di essere scelto.

1. Definire lo spazio campionario $\Omega$ dell'esperimento e calcolare la sua cardinalità utilizzando i concetti del calcolo combinatorio. (1pt)
2. PostgreSQL e MySQL sono database di tipo SQL, mentre gli altri tre non lo sono. Descrivere i seguenti eventi come sottoinsiemi di $\Omega$ e calcolare la loro cardinalità:
   - $\mathcal{A}$: "Viene scelto almeno un database SQL" (1pt)
   - $\mathcal{B}$: "Viene scelto esattamente un database NoSQL e un database SQL" (1pt)
   - $\mathcal{C}$: "Viene scelto MongoDB o Redis (o entrambi)" (1pt)
3. Calcolare le seguenti probabilità, utilizzando i risultati del calcolo combinatorio:
   - $P(\mathcal{A})$ (1pt)
   - $P(\mathcal{B})$ (1pt)
   - $P(\mathcal{C})$ (1pt)
4. Calcolare la probabilità dell'evento $\mathcal{D} = (\mathcal{A} \cap \mathcal{C})$. (3pt)

### *1 Soluzione*

1. Lo spazio campionario $\Omega$ è l'insieme di tutte le possibili coppie non ordinate di database diversi. La sua cardinalità è il numero di combinazioni di classe 2, estratte da un insieme di 5 elementi:
   $|\Omega| = \binom{5}{2} = \frac{5!}{2!(5-2)!} = 10$

2. Descrizione degli eventi e calcolo delle cardinalità:

   - $\mathcal{A}$: "Viene scelto almeno un database SQL" (PostgreSQL e MySQL sono SQL)

     $$|\mathcal{A}| = |\Omega| - (\text{numero di scelte di due database no-SQL}) = 10 - \binom{3}{2} = 10 - 3 = 7$$

   - $\mathcal{B}$: "Viene scelto esattamente un database NoSQL e un database SQL" $|\mathcal{B}| = 3 \cdot 2 = 6$
   - $\mathcal{C}$: "Viene scelto MongoDB o Redis (o entrambi)" $|\mathcal{C}| = |\Omega| -$ (numero di scelte senza MongoDB e Redis) $= 10 - \binom{3}{2} = 10 - 3 = 7$

3. Calcolo delle probabilità:

   - $P(\mathcal{A}) = \frac{|\mathcal{A}|}{|\Omega|} = \frac{7}{10} = 0,7$
   - $P(\mathcal{B}) = \frac{|\mathcal{B}|}{|\Omega|} = \frac{6}{10} = \frac{3}{5} = 0,6$
   - $P(\mathcal{C}) = \frac{|\mathcal{C}|}{|\Omega|} = \frac{7}{10} = 0,7$

4. $\mathcal{D} = (\mathcal{A} \cap \mathcal{C})$: Cioè $\mathcal{D}$ rappresenta l'evento "Viene scelto almeno un database SQL **e** viene scelto MongoDB o Redis".

   Per calcolare $P(\mathcal{D})$, dobbiamo prima trovare $|\mathcal{D}|$, ovvero il numero di elementi in $\mathcal{A} \cap \mathcal{C}$. Dal primo quesito, abbiamo capito che $|\mathcal{A}| = 7$ e anche $|\mathcal{C}| = 7$, anche se gli elementi degli insiemi sono diversi: l'uno continene sempre almeno un dataset di tipo SQL, mentre l'altro potrebbe anche non contenerlo.

   Ci sono due modi di trovare l'intersezione di $\mathcal{A}$ e $\mathcal{C}$. Il primo modo è il conteggio diretto, dato che ci rendiamo conto che i casi in cui si compiono tutte e due le condizioni simultaneamente è quando si sceglie uno dei seguenti casi:

   (a) PostgreSQL e MongoDB

   (b) PostgreSQL e Redis

   (c) MySQL e MongoDB

   (d) MySQL e Redis

   Quindi $|\mathcal{D}| = 4$, perciò: $P(\mathcal{D}) = P(\mathcal{A} \cap \mathcal{C}) = \frac{|\mathcal{D}|}{|\Omega|} = \frac{4}{10} = \frac{2}{5} = 0,4$

   Un'altro modo di risolverlo è cercando di trovare la cardinalità di $\mathcal{D}$ in funzione del suo evento complementare: $\mathcal{D} = |\Omega| - |\bar{\mathcal{D}}|$.

   Questo perché sapiamo, dalle leggi di De Morgan, che: $\bar{\mathcal{D}} = \overline{\mathcal{A} \cap \mathcal{C}} = \bar{\mathcal{A}} \cup \bar{\mathcal{C}}$.

   Sempre dalla soluzione del primo quesito sappiamo che $|\bar{\mathcal{A}}| = 3$ e $|\bar{\mathcal{C}}| = 3$. E, dato che questi insiemi sono disgiunti -il primo contiene scelte che per forza includono MongoDB oppure Redis oppure entrambi, mentre l'altro, per definizione, non le contiene- la loro unione ha cardinalità 6. Da cui giungiamo allo stesso risultato.

## 2 CrowdStrike

Un evento informatico con conseguenze gravi su scala mondiale, sia esso un bug oppure un attacco mirato, verrà chiamato GBoA (*Global Bug or Attack*) per gli scopi di questo esercizio.

Il tempo che trascorre tra l'occorrenza di due GBoA si può considerare come un fenomeno che segue una distribuzione esponenziale con parametro $\lambda = 0.1$ anni$^{-1}$.

1. In media, ogni quanti anni succede un GBoA? (1 pt.)

2. Qual è la probabilità che un GBoA si verifichi entro i prossimi 5 anni? (2 pt.)

3. Il GBoA *Crowdstrike* del 2024 è successo dopo 7 anni dall'ultimo GBoA, *WannaCry*, del 2017.

   Supponiamo di essere nel 2020, e di chiederci, visto che sono passati 3 anni senza nessun GBoA, quale è la probabilità che passino altri 4 anni senza nessun GBoA? (2 pt)

4. Qual è la probabilità di osservare esattamente 2 GBoA in un periodo di 15 anni? (3pt.)

5. Quale è invece la probabilità di osservare al più un GBoA in un periodo di 10 anni? (2pt.)

***2 Soluzione***

Dai dati del problema, sappiamo che il tempo di inter-arrivo tra due bug o attacchi su scala mondiale è descritto da una variabile aleatoria $\mathcal{T}$, tale che la probabilità che un evento di questo tipo si verifichi entro un certo tempo $t$ è data da:

$$P(\mathcal{T} \leq t) = 1 - e^{-\lambda t}$$

Dove $\lambda = 0,1$ anni$^{-1}$.

1. Sappiamo che se $\mathcal{T}$ è una variabile aleatoria esponenziale, allora il suo valore atteso è pari a:

   $$\mathbb{E}[\mathcal{T}] = \frac{1}{\lambda}$$

   E cioè:

   $$\mathbb{E}[\mathcal{T}] = \frac{1}{0.1} \approx 10$$

   In media, ci dobbiamo aspettare un bug o attacco su scala mondiale ogni 10 anni. (Andava bene anche $\frac{1}{0.1}$ come risposta :/ )

2. Il quesito ci chiede $P(\mathcal{T} \leq 5)$, quindi sostituendo $t = 5$ anni abbiamo:

   $P(T \leq 5) = 1 - e^{-0,1 \cdot 5} = 1 - e^{-0,5} \approx 0,39$

   Quindi, la probabilità richiesta è circa il 39%.

   (La risposta si poteva lasciare scritta anche come $1 - e^{-0,5}$ )

3. Il quesito ci chiede

   $$P(\mathcal{T} \geq 3 + 4 | \mathcal{T} \geq 3)$$

   Utilizziamo la proprietà di assenza di memoria della distribuzione esponenziale. La probabilità di dover attendere almeno altri 4 anni per un attacco o bug su scala mondiale è:

   $P(T \geq 3 + 4 | \mathcal{T} \geq 3) = P(\mathcal{T} > 4) = e^{-\lambda t} = e^{-0,1 \cdot 4} \approx 0,67$

   Quindi, la probabilità è circa il 67%.

4. Sappiamo che se una variabile aleatoria $\mathcal{T}$ segue la distribuzione esponenziale, essa descrive il tempo di interarrivo di eventi generati da un processo di Poisson. Questo fatto ci permette di collegare la v.a. $\mathcal{T}$ con un'altra v.a. $\mathcal{X}$ che conta il numero di occorrenze dell'evento il cui tempo di interarrivo è descritto da $\mathcal{T}$.

   Se $\mathcal{X}$ conta il numero di bug o attacchi su scala mondiale in un intervallo di 15 anni, allora avremmo che $\mathcal{X}$ è distribuita secondo poisson, e il suo valor medio sarà $\mu = \lambda \times 15 = 1.5$.

   $$\mathcal{X} \sim \text{Poisson}(1.5)$$

   Sappiamo inoltre che la probabilità di osservare esattamente $k$ eventi in un processo di Poisson è:

$$P(\mathcal{X} = k) = \frac{1.5^k e^{-1.5}}{k!}$$

E quindi, per $k = 2$ avremo:

$$P(\mathcal{X} = 2) = \frac{1.5^2 e^{-1.5}}{2!} \approx 0,25$$

Quindi, la probabilità è circa il 25%.

Nota: In questo esercizio non ci è stato detto di assumere che gli arrivi dei GBoA siano indipendenti tra di loro, (c'è una chiara dipendenza invece, per definizione di tempo di interarrivo) perciò non era corretto modellare una variabile aleatoria Binomiale che rappresenti il numero di occorrenze dei GBoA.

5. Seguendo il ragionamento del punto precedente, sappiamo che il numero di attacchi osservati in un periodo di 10 anni è descritto da una v.a. $\mathcal{Y}$ distribuita secondo Poisson e di valore atteso $\mu = 1$:

$$\mathcal{Y} \sim \text{Poisson}(1)$$

Perciò la probabilità di osservare al più 1 attacco nei prossimi 10 anni sarà pari a:

$$P(\mathcal{Y} <= 1) = P(\mathcal{Y} = 0) + P(\mathcal{Y} = 1) = \frac{1^0 e^{-1}}{0!} + \frac{1^1 e^{-1}}{1!} = 2e^{-1} \approx 0.75$$

## 3 Linux

Si stima che il 20% degli studenti di informatica in Italia utilizzi Linux come sistema operativo principale. Per verificare questa stima, viene condotto un sondaggio su un campione casuale di 200 studenti provenienti da diverse università italiane.

1. Supponendo che la stima sopra citata sia corretta, definire una variabile aleatoria Binomiale $\mathcal{X}$ che rappresenti il numero di studenti del campione che utilizzano Linux come sistema operativo principale. Quali sono i parametri della distribuzione di $\mathcal{X}$? (1 pt.)

2. Definire la variabile aleatoria $\mathcal{Y}$ che approssima la v.a. del punto precedente, ma che ha una distribuzione Normale. Quali sono i parametri di questa nuova distribuzione? (1 pt.)

3. Utilizzando la standardizzazione di $\mathcal{Y}$, calcolare la probabilità di ottenere *esattamente* 32 studenti che utilizzano Linux nel campione. Trovare un valore numerico. (3 pt.)

   *Suggerimento*. Si utilizzino le seguenti approssimazioni: $\frac{31.5-40}{\sqrt{32}} \approx -1.5$ e $\frac{32.5-40}{\sqrt{32}} \approx -1.33$

4. Si calcoli ora la probabilità che l'esito del sondaggio corrisponda ad un valore tra i 30 e i 50 studenti. Trovare un valore numerico. (2 pt.)

   *Suggerimento*. Si utilizzino le seguenti approssimazioni: $\frac{29.5-40}{\sqrt{32}} \approx -1.86$ e $\frac{50.5-40}{\sqrt{32}} \approx 1.86$

5. Supponiamo invece che il valore atteso di $\mathcal{X}$ sia $\mu = 40$ e la sua deviazione standard sia $\sigma = 10$. Senza ipotizzare null'altro riguardo la distribuzione di probabilità $\mathcal{X}$, trovare la probabilità minima che l'esito del sondaggio corrisponda ad un valore tra i 15 e i 65 studenti (estremi **esclusi**, cioè $15 < \mathcal{X} < 65$)? (3 pt.)

### *3 Soluzione*

1. La variabile aleatoria $\mathcal{X}$ segue una distribuzione Binomiale con parametri:

   - $n = 200$ (numero di studenti nel campione)
   - $p = 0.20$ (probabilità di successo, ovvero la probabilità che uno studente usi Linux)

   Quindi, $\mathcal{X} \sim$ Binomiale$(200, 0.20)$

2. La variabile aleatoria $Y$ che approssima $X$ con una distribuzione Normale ha i seguenti parametri:

   - $\mu = \mathbb{E}[\mathcal{Y}] = \mathbb{E}[\mathcal{X}] = np = 200 \cdot 0.20 = 40$
   - $\sigma^2 = Std[\mathcal{Y}] = Std[\mathcal{X}] = np(1-p) = 200 \cdot 0.20 \cdot 0.80 = 32$
   - $\sigma = \sqrt{32}$

   Quindi, $\mathcal{Y} \sim \mathcal{N}(\mu = 40, \sigma^2 = 32)$ oppure: $\mathcal{Y} \sim \mathcal{N}(\mu = 40, \sigma = \sqrt{32})$

3. Per calcolare la probabilità di ottenere esattamente 32 studenti, dobbiamo usare la correzione di continuità:

   $$P(\mathcal{Y} = 32) \approx P(31.5 < \mathcal{Y} < 32.5)$$

   Nota: una risposta completa include la esplicita specificazione di questo fatto: Non basta con usare il suggerimento, ma indicare che si sta usando la correzione di continuità.

   Standardizziamo $\mathcal{Y}$ per ottenere una variabile aleatoria Normale Standard, $\mathcal{Z}$:

   $$\mathcal{Z} = \frac{\mathcal{Y} - \mu}{\sigma} = \frac{\mathcal{Y} - 40}{\sqrt{32}}$$

   Con questa nuova variabile possiamo scrivere:

   $$P(31.5 < \mathcal{Y} < 32.5) = P\left(\frac{31.5 - 40}{\sqrt{32}} < \mathcal{Z} < \frac{32.5 - 40}{\sqrt{32}}\right)$$

   E, usando le approssimazioni del suggerimento, possiamo scrivere:

   $$P(31.5 < \mathcal{Y} < 32.5) = P(-1.5 < Z < -1.33) = \Phi(-1.33) - \Phi(-1.5)$$

   Facendo un lookup sulla tabella $\mathcal{Z}$ abbiamo poi:

   $$\Phi(-1.33) - \Phi(-1.5) = 0.09 - 0.07 = 0.03$$

   Quindi, la probabilità richiesta è di circa il 3%.

4. Il quesito ci chiede:

   $$P(30 \leq \mathcal{X} \leq 50)$$

   Di nuovo, con la correzione di continuità avremo:

   $$P(30 \leq \mathcal{X} \leq 50) = P(29.5 \leq \mathcal{Y} \leq 50.5)$$

   Aiutandoci dalla standardizzazione, sappiamo che:

   $$P(29.5 \leq \mathcal{Y} \leq 50.5) = P\left(\frac{29.5 - 40}{\sqrt{32}} \leq Z \leq \frac{50.5 - 40}{\sqrt{32}}\right)$$

E, utilizzando l'approssimazione del suggerimento avremo:

$$= P(-1.86 \leq Z \leq 1.86) = \Phi(1.86) - \Phi(-1.86) = 2\Phi(1.86) - 1 = 0.94$$

Quindi, la probabilità è di circa il 94%.

5. Ci viene chiesta $P(15 < \mathcal{X} < 65)$. Sapendo che il valore atteso è: $\mu = 40$, e la deviazione standard è $\sigma = 10$.

   Osserviamo che ci viene chiesta una probabilità minima riguardo ad un intervallo simmetrico attorno alla media. Essendo poi $\mu$ e $\sigma$ positivi, abbiamo la possibilità di usare la disuguaglianza di Chebyshev.

   Tale disuguaglianza afferma che:

   $$P(|\mathcal{X} - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

   Oppure, equivalentemente:

   $$P(|\mathcal{X} - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

   Dato che ci viene chiesta la probabilità minima del fatto che l'esito del sondaggio sia *compreso* all'interno dell'intervalo, usiamo la seconda formulazione.

   Notiamo che:

   $$P(16 \leq \mathcal{X} \leq 65) = P(15 < \mathcal{X} < 65) = P(|\mathcal{X} - 40| < 25)$$

   Se poniamo $k\sigma = 25$, avremo che $k = \frac{25}{\sigma} = \frac{25}{10} = 2.5$.

   Quindi, avremo:

   $$P(|X - 40| < 10 \times 2.5) \geq 1 - \frac{1}{2.5^2} = 0.16$$

   Quindi la probabilità minima è del 16%.