**Contrassegnare il proprio canale di afferenza:** [A] [B]
**Cognome e nome:** _____
**Numero di matricola:** _____
**Postazione:** _____
**Nome del file consegnato:** _____

# SOLUZIONE Simulazione d'esame 13/01/2025 - Prova MATLAB

La prova di MATLAB consiste in **4 esercizi** da svolgere in **60 minuti** e conta per **12 trentesimi** (additivi) ai fini della valutazione finale.

Ciascun esercizio richiede di **riportare sui fogli del testo d'esame le risposte** ad alcune domande. Risposte eventualmente prodotte dal codice MATLAB ma non riportate nel foglio saranno valutate 0 punti.

È obbligatorio (nonché oggetto di valutazione) **consegnare, oltre ai fogli del testo d'esame,** lo svolgimento degli esercizi in **un unico script MATLAB** (file .m) da intitolare

`letteraMaiuscolaDelCanale_Cognome_Nome_simulazione.m`

Lo script dovrà iniziare con un commento riportante

- Nome e cognome
- Numero di matricola
- Canale di afferenza

**ALL'ESAME**: Periodicamente (in modo da creare un backup nel server in cui i docenti ritroveranno la prova d'esame svolta) e, soprattutto, alla fine dell'esame, **utilizzare il commando `consegna('nome_file.m')`** nella command window di MATLAB.

Quindi, l'ipotetico studente Carlo Azeglio Ciampi, afferente al canale B e con matricola 9929778, dovrà produrre e consegnare un unico script MATLAB dal nome `B_Ciampi_Carlo_Azeglio_simulazione.m` le cui prime righe saranno

```
% Carlo Azeglio Ciampi
% 9929778
% Canale B
```

Carlo Azeglio Ciampi, poi, dovrà consegnare con il comando `consegna('B_Ciampi_Carlo_Azeglio_simulazione.m')`

**Attenzione:**

- Il nome del file deve essere passato come stringa, completa di estensione .m, senza spazi.
- La mancata consegna del file .m o la sua consegna in bianco risulteranno in una valutazione della prova di MATLAB pari a 0 (su possibili 12) trentesimi.
- La presenza di eventuali errori MATLAB che impediscano la completa esecuzione dello script risulterà altresì in una valutazione della prova di MATLAB pari a 0 (su possibili 12) trentesimi.

**Caricamento dati**

Scaricare il file `simulazione_20250113.mat` e incollarlo nella cartella di lavoro.

Il file `simulazione_20250113.mat` contiene le seguenti variabili

- `data`: matrice [1000 soggetti x 7 variabili] con le seguenti 7 colonne
    1. Età in anni (`age`)
    2. Peso in kg (`weight`)
    3. Altezza in cm (`height`)
    4. Frequenza cardiaca in bpm (`heart_rate`)
    5. Pressione sistolica alla baseline in mmHg (`systolic_blood_pressure_baseline`)
    6. Assunzione di un farmaco (`medication`) codificata come un indicatore sì/no tale che "1" significa che il paziente assume il farmaco e "0" che non lo assume.
    7. Pressione sistolica al follow-up in mmHg (`systolic_blood_pressure_followup`)
- `labels`: cell array di 7 elementi, ciascuno contenente l'etichetta della colonna corrispondente nella matrice data (ovvero le stringhe tra parentesi e in font monospace sopra riportate).
- `units`: cell array di 7 elementi, ciascuno contenente l'unità di misura della colonna corrispondente nella matrice data.
- `signal_data`: matrice [250 segnali x 240 istanti di tempo] contenente 250 segnali di concentrazione di un farmaco nel sangue in mg/dL.
- `signal_time`: vettore di lunghezza 240 contenente gli istanti di tempo a cui sono stati campionati i segnali in `signal_data` (unità di misura: minuti).

**1. Pulizia dati.** Data la matrice di dati `data`, portare a termine la procedura di pulizia dati sotto descritta.

1. Considerando l'intera matrice, individuare e riportare il numero di
    a. valori mancanti: 464
    b. valori negativi: 102

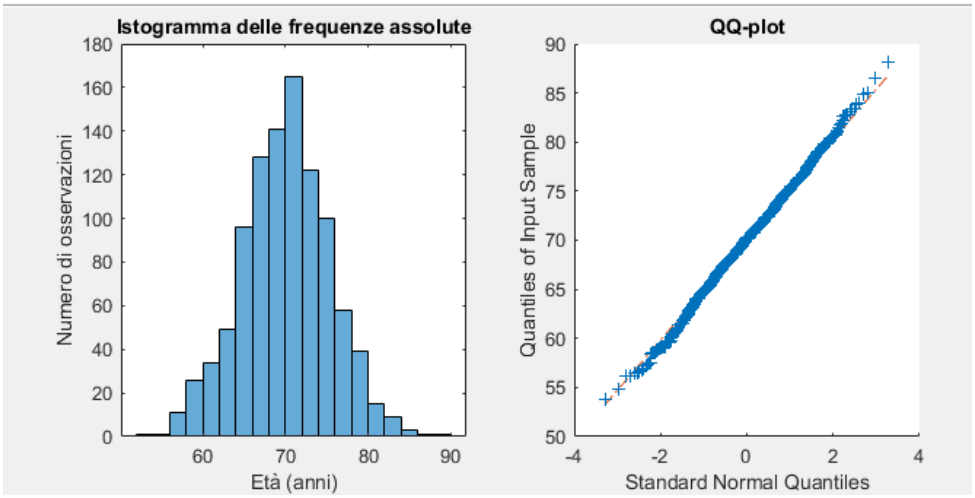
2. Eliminare le colonne con più del 20% di valori mancanti o non fisiologici e salvare la matrice così ottenuta in una variabile chiamata `data_reduced`.
   v. codice

3. Eliminare da `data_reduced` le righe che hanno almeno un valore mancante o non fisiologico. Salvare la matrice così ottenuta in una variabile chiamata `data_reduced_filtered`.
   v. codice

4. Riportare le dimensioni di `data_reduced_filtered`: 938x5

**2. Test statistici.** Data la matrice di dati `data`, considerarne la prima colonna (age).

1. Sulla base degli indicatori visti a lezione, la sua distribuzione può ragionevolmente considerarsi gaussiana? Giustificare sinteticamente la risposta:




Il test di gaussianità di Lilliefors è inconclusivo (non rifiuta l'ipotesi nulla, dunque non ci permette di concludere nulla).
L'istogramma delle frequenze assolute non evidenzia particolari asimmetrie; il QQ-plot giace quasi perfettamente su una retta; la skewness è circa 0; la curtosi è circa 3.
Sulla base di questi indicatori, la distribuzione della variabile age può considerarsi gaussiana.

2. A prescindere da quanto trovato nel punto 1, applicare l'opportuno test statistico per confrontare la sua media con il valore medio dell'età in Italia, ovvero 46.6 anni (usare un livello di significatività pari α = 0.05) e riportare le conclusioni che è possibile trarre:
   Il p-value restituito da MATLAB è così piccolo da essere stato approssimato a 0. Poiché 0 < 0.05, si può concludere che la media della variabile age è significativamente diversa dal valore 46.6 anni.

**3. Regressione lineare.** Data la matrice di dati `data`, considerare le colonne 1, 5, 6 (`age`, `systolic_blood_pressure_baseline`, `medication`) come variabili indipendenti e la colonna 7 (`systolic_blood_pressure_followup`) come variabile dipendente.

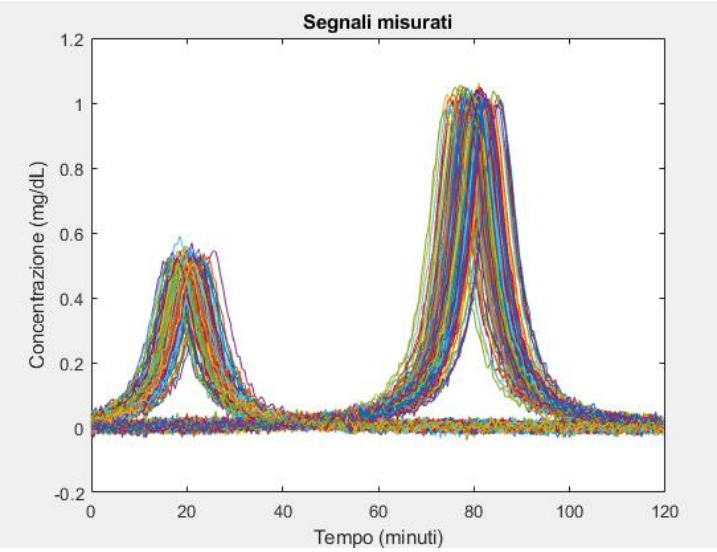
1. Scrivere l'equazione del modello di regressione lineare (intercetta inclusa, termine relativo all'errore escluso) che lega la variabile dipendente alle variabili indipendenti, specificando il significato dei simboli utilizzati:
   $$SBP_{fup} = \beta_1 \cdot age + \beta_2 \cdot SBP_b + \beta_3 \cdot medication + \beta_0$$
   $\beta_0$: intercetta
   $\beta_1, \beta_2, \beta_3$coefficienti di regressione corrispondenti alle variabili `age`, `systolic_blood_pressure_baseline`, e `medication`
   $SBP_{fup}$: variabile dipendente `systolic_blood_pressure_followup`
   $age$: variabile indipendente `age`
   $SBP_b$: variabile indipendente `systolic_blood_pressure_baseline`
   $medication$: variabile indipendente `medication`

2. Stimare, implementando "a mano" la formula risolutiva per la stima dei parametri dei minimi quadrati lineari non pesati, il valore dei coefficienti del modello. Riportare quello relativo alla variabile `medication`: -0.87

3. Calcolare e riportare di seguito l'intervallo di confidenza al 95% intorno alla stima del parametro relativo alla variabile `medication`: [-1.47, -0.26] (utilizzando il termine moltiplicativo 1.96; va bene anche utilizzare 2)

**4. Clustering.** Considerando la matrice di segnali signal_data (unità di misura: mg/dL), acquisiti ai tempi signal_time (unità di misura: minuti), si vuole determinare se 2 oppure 3 sia il numero di cluster K migliore con cui implementare l'algoritmo k-means.

1. Disegnare, in uno stesso grafico, tutte sovrapposte, le tracce contenute nella matrice signal_data.




2. Utilizzare due volte l'algoritmo k-means: la prima con k fissato pari a 2; la seconda con K fissato pari a 3. Entrambe le volte, fare in modo che
    a. La distanza di riferimento sia la distanza euclidea
    b. Il numero di repliche (inizializzazioni diverse) sia 10
    c. Il numero massimo di iterazioni sia 100
    d. L'istruzione subito precedente alla chiamata della funzione MATLAB che implementa k-means sia `rng(42);` (per agevolare la correzione)

    v. codice

3. Riportare i valori di silhouette medi ottenuti per
    a. K fissato pari a 2: 0.895
    b. K fissato pari a 3: 0.787

4. Qual è il numero di cluster migliore tra K=2 e K=3? 2, perché il valore di silhouette medio è più alto.