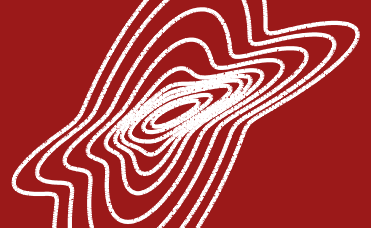


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

METODI STATISTICI PER LA BIOINGEGNERIA

Laboratorio 4

A.A. 2024-2025

Enrico Longato

Grazie alla Prof.ssa Alessandra Bertoldo e al Dott. Marco Castellaro per la versione iniziale delle slide.

Dal lab 3: Indicizzazione booleana senza il find

1. Costruisco una matrice qualunque

```
>> v = [1, 2, 3, 4; 5, 6, 7, 8]

v =

     1     2     3     4
     5     6     7     8
```

```
>> (v >= 3) & (v <= 6)

ans =

  2×4 logical array

   0   0   1   1
   1   1   0   0
```

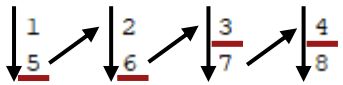
2. Imponendo una condizione logica, creo una maschera booleana che ha 1 dove la condizione è vera e 0 altrove

3. Filtrando con la condizione logica direttamente (senza il find), trovo quello che mi aspetto, ovvero i valori che corrispondono alla condizione logica

```
>> v((v >= 3) & (v <= 6))

ans =

     5
     6
     3
     4
```

4. NB: l'ordine sembra "strano", ma è quello del find lineare!

```
>> v = [1, 2, 3, 4; 5, 6, 7, 8]

v =
```



Dal lab 3: Negazione di una maschera

```
mask_on_v =

  2×4 logical array

   0   0   1   1
   1   1   0   0

>> ~mask_on_v

ans =

  2×4 logical array

   1   1   0   0
   0   0   1   1
```

Banalmente, la negazione elemento per elemento si fa con il simbolo tilde ~ (alt+125)

Dal lab 3: Funzioni con struttura f(qualcosa, dim)

Tantissime funzioni MATLAB di default agiscono colonna per colonna.

- Esempio: `sum(A)` dà un vettore riga contenente la somma di ciascuna colonna di `A`; se vogliamo un vettore colonna contenente la somma di ciascuna riga di `A` dobbiamo fare `sum(A, 2)`

Convenzionalmente, per una matrice 2D, "lavorare colonna per colonna" si dice "lavorare lungo la prima dimensione" (che, però, sarebbero le righe), **mentre "lavorare riga per riga" si dice "lavorare lungo la seconda dimensione"** (che, però, sarebbero le colonne).

- Da cui, `sum(A)` è come dire `sum(A, 1)`; mentre `sum(A, 2)` si può dire solo così

L'ambiguità in termini fa, in effetti, confusione; dobbiamo inventarci uno stratagemma mnemonico. Eccone un paio (ma usate quello che volete):

- È come quando facciamo l'integrale: lo facciamo lungo x (la prima dimensione), ma il numero che esce dipende dal valore della funzione che leggiamo su y (la seconda dimensione).
- La dimensione da indicare a MATLAB è quella lungo cui un ideale ciclo for dovrebbe scorrere (per fare la somma di una colonna, il ciclo for scorrerebbe le righe, quindi la dimensione 1).

Dal lab 3: Eliminare i soggetti (= righe) con almeno un dato mancante

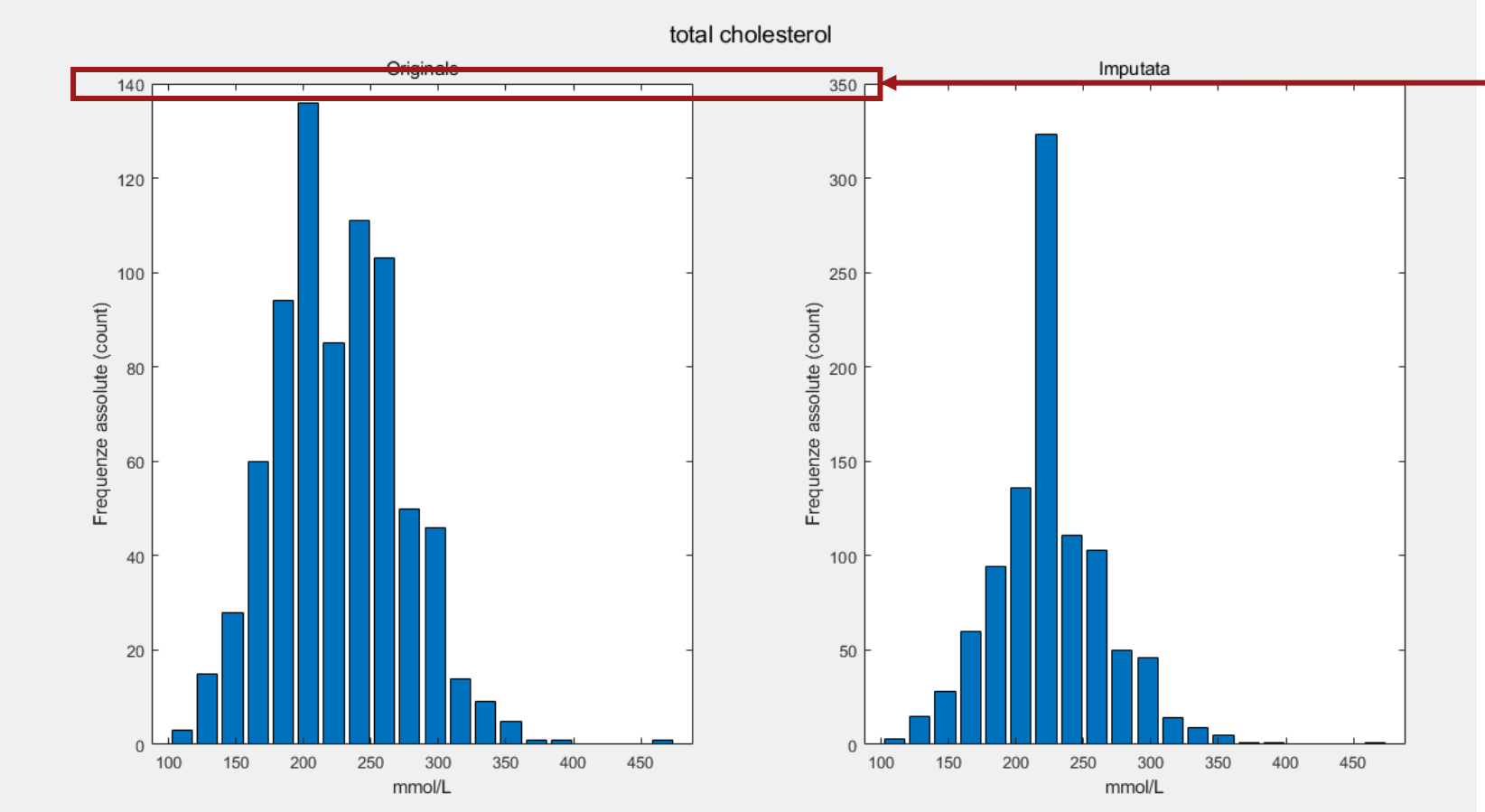
```
%% Calcolare la percentuale di nan per ogni variabile
% Calcoliamo il numero di nan per colonna
nan_counts = sum(isnan(elsa));

% Trasformiamo in percentuale
nan_percentages = 100 * nan_counts / size(elsa, 1);

%% Rimuovere le colonne con >20% dati mancanti
elsa_reduced = elsa(:, nan_percentages <= 20);
```

```
%% Rimuovere i soggetti che hanno, in almeno una delle variabili rimaste,
%% un valore mancante
% Sommo per righe con sum(..., 2) e guardo quando i nan sono più di 0
i_at_least_one_nan = sum(isnan(elsa_reduced), 2) > 0;
% Alternativamente, potevo fare any(isnan(elsa_reduced), 2)

% Filtro
elsa_reduced_filtered = elsa_reduced(~i_at_least_one_nan, :);
```

Esattamente speculare, con due ovvie differenze

1. **Per le colonne ho usato `sum` senza parametri, quindi come se fosse `sum(isnan(elsa), 1)`; mentre per le righe ho usato `sum` con l'argomento `dim = 2`, cioè `sum(isnan(elsa_reduced), 2)`**
2. **Tengo tutte le righe `(:, …)` quando filtro le colonne, mentre tengo tutte le colonne `(..., :)` quando filtro le righe.**

Dal lab 3: Esito dell'imputazione con la mediana



Notare la scala!

Avendo imputato con la mediana, i dati che erano mancanti (per la variable total cholesterol erano il 23.8%) ora sono tutti pari alla mediana.

Ecco, quindi, che vediamo comparire un picco dell'istogramma sull'ora frequentissimo valore 224 mmol/L.

Laboratorio 4: Contenuti e obiettivi

1. Esercitazione "alla lavagna"
   - Caricamento ed esplorazione dati
   - t-test a un campione a due code
2. Esercizi da svolgere in autonomia (per superare la "paura del file bianco")
   - Statistiche descrittive e istogrammi
   - t-test a un campione a una coda
   - Variazione del livello di significatività
   - Riflessioni sull'ipotesi nulla
3. Ripasso di teoria (**parte integrante del programma d'esame di teoria!**)
   - Test statistici in pratica

Verifica di ipotesi

I test statici sono tentativi di confutare un'ipotesi (l'ipotesi nulla) che, di conseguenza, corrisponde al "contrario" di quello che vorremmo dimostrare.

- In altre parole, se ci interessa dimostrare che due fenomeni siano diversi, procediamo ipotizzando che siano uguali e cercando di confutare questo fatto.
- La confutazione passa dal ragionamento "i dati che ho a disposizione sono così incompatibili con l'ipotesi nulla che essa non può che essere falsa e, dunque, con un certo livello di confidenza, posso dire che l'ipotesi alternativa (quella che volevo dimostrare) è vera".

Statistica del test

Matematicamente, i test statistici si basano sulla costruzione di una statistica, cioè di una variabile aleatoria "intelligentemente congegnata" che ha le seguenti tre caratteristiche

1. Ha senso solo se prendiamo per vera l'ipotesi nulla.
2. Si può formulare e se ne possono calcolare i valori a partire da informazioni facilmente reperibili sui campioni statistici a disposizione (es.: media, varianza, numerosità, …).
3. È matematicamente "comoda", cioè, sappiamo "tutto" della sua distribuzione, che è completamente definita date l'ipotesi nulla presa per vera al punto 1 e le poche informazioni di cui al punto 2.

- In pratica, conviene pensare alla **statistica del test** come un marchingegno matematico che noi **usiamo "così com'è"**, senza farci troppe domande.
- Sempre in pratica, la **forma matematica complicata** della statistica del test è principalmente "di comodo": all'aumentare della complessità del test, essa **perde progressivamente riscontro intuitivo** e diventa progressivamente sempre più difficile (fino a praticamente impossibile) interpretarla in termini di fenomeni reali.
- Più formalmente, vale che, in generale, **una statistica di test non è una buona statistica descrittiva** della popolazione (e le statistiche descrittive di una popolazione raramente sono buone statistiche di test).

Test statistici in pratica

1. Individuo la proprietà che voglio dimostrare e scelgo l'ipotesi nulla che, qualora dovesse essere rifiutata, mi darebbe la risposta che cerco.
2. Scelgo una regola di decisione, cioè un livello di significatività (lo posso scegliere "quando voglio", basta che sia **prima** di fare i calcoli).
3. Verifico alcune assunzioni di partenza sui dati (v. prossimo laboratorio; oggi ci fidiamo).
4. Assumo che l'ipotesi nulla sia vera (importante!).
5. Individuo la statistica corrispondente all'ipotesi nulla.
6. Calcolo i parametri necessari a partire dai dati a disposizione.
7. Calcolo il valore della statistica usando i numeri del punto sopra.
8. Confronto i p-value corrispondente con il livello di significatività e:
    1. Se il p-value è minore del livello di significatività, rifiuto l'ipotesi nulla.
    2. Se il p-value è maggiore del livello di significatività, mi chiudo in rispettoso silenzio, non potendo concludere niente sull'esperimento, (in questo caso, si dice che """"accetto"""" l'ipotesi nulla).

Test statistici in pratica (1 di 8)

- Individuo la proprietà che voglio dimostrare e scelgo l'ipotesi nulla che, qualora dovesse essere rifiutata, mi darebbe la risposta che cerco.

  Voglio dimostrare che la media della popolazione da cui ho estratto il mio campione non è 100 mg/dL.
  - Mi chiedo: "Esiste un'ipotesi nulla che mi aiuterebbe, se negata, in tal senso?"
  - Risposta: "Sì, posso assumere $H_0: \mu = 100$ mg/dL"

Test statistici in pratica (2 di 8)

- Scelgo una regola di decisione, cioè un livello di significatività (lo posso scegliere "quando voglio", basta che sia **prima** di fare i calcoli).

  Scelgo $\alpha = 0.05$

  Non c'è molto da dire: lo **scelgo "per protocollo"** sulla base di quanto severo voglio essere con il test:

  - Se mi basta un ragionevole dubbio che l'ipotesi nulla sia falsa, sceglierò α più grande
  - Se mi serve essere molto più sicuro, sceglierò α più piccolo.

  L'importante è sceglierlo **prima (!!!) di fare i calcoli**.

Test statistici in pratica (3 di 8)

- Verifico alcune assunzioni di partenza sui dati (v. prossimo laboratorio; oggi ci fidiamo).

La statistica del test si può costruire correttamente **solo** se sono valide contemporaneamente:

1. L'ipotesi nulla.
2. Alcune ipotesi su tutte le variabili aleatorie coinvolte (es. quelle di cui i dati sono realizzazione).

Sull'ipotesi nulla non si può sindacare.

Le altre ipotesi vanno verificate e bisogna essere consapevoli di quanto si stia deviando dal "caso ideale"; a volte può non essere un problema grave (v. stimatori robusti).

=> Prossimo laboratorio.

Test statistici in pratica (4 di 8)

- Assumo che l'ipotesi nulla sia vera (importante!).

Devo interiorizzare e ricordami a ogni passaggio che l'ipotesi nulla è un'ipotesi e, dunque, è **data per vera da ora in avanti**.

Sembra controintuitivo, perché stiamo cercando di confutarla, ma dobbiamo darla per vera fino al momento in cui la confutiamo (*).

Altrimenti, la matematica perde di significato e, con essa, tutti i numeri che calcoliamo.

(*) Il ragionamento torna perché, tecnicamente, non confutiamo mai H_0 categoricamente; bensì solo con una certa "confidenza", corrispondente al livello di significatività α.

Test statistici in pratica (5 di 8)

- Individuo la statistica corrispondente all'ipotesi nulla.

All'atto pratico, non ci è mai richiesto di inventare la statistica del test.

- Semplicemente, cerchiamo in letteratura la statistica che meglio corrisponde all'ipotesi nulla e alla nostra situazione sperimentale.

Nel nostro caso, l'ipotesi nulla era $H_0: \mu = 100$ mg/dL e nulla sapevamo della varianza del campione (e abbiamo glissato sulle altre ipotesi; v. prossimo lab).

- Dunque, usiamo un t test a un campione con varianza incognita.
- Cerchiamo la funzione MATLAB corrispondente: `ttest` chiamata con un vettore di dati e una costante.

Test statistici in pratica (6 e 7 di 8)

- Calcolo i parametri necessari a partire dai dati a disposizione.
- Calcolo il valore della statistica usando i numeri del punto sopra.

MATLAB fa entrambe le cose per noi. In particolare, ci restituisce quanto segue

Test the null hypothesis that the sample data comes from a population with mean equal to zero.

```
[h,p,ci,stats] = ttest(x)
```

h = 1 se e solo se l'ipotesi nulla è rifiutata con il livello di significatività specificato (qui non abbiamo scritto nulla, quindi 0.05)

Il p-value

L'intervallo di confidenza per la media di popolazione (non ci interessa, ai fini del test)

Una `struct` di campi

- `stats.tstat` = valore assunto dalla statistica del test
- `stats.df` = i gradi di libertà del test
- `stats.sd` = la deviazione standard stimata

Test statistici in pratica (8 di 8)

- Confronto i p-value corrispondente con il livello di significatività e:
  1. Se il p-value è minore del livello di significatività, rifiuto l'ipotesi nulla.
  2. Se il p-value è maggiore del livello di significatività, mi chiudo in rispettoso silenzio, non potendo concludere niente sull'esperimento, (in questo caso, si dice che """"accetto"""" l'ipotesi nulla).

Si tratta di confrontare il valore di p restituito dalla function con $\alpha = 0.05$ che avevamo deciso all'inizio.

- Possiamo addirittura leggere l'esito del confronto nel parametro di uscita `h`.

Il calcolo non è un problema, l'interpretazione potrebbe esserlo (v. slide successiva)

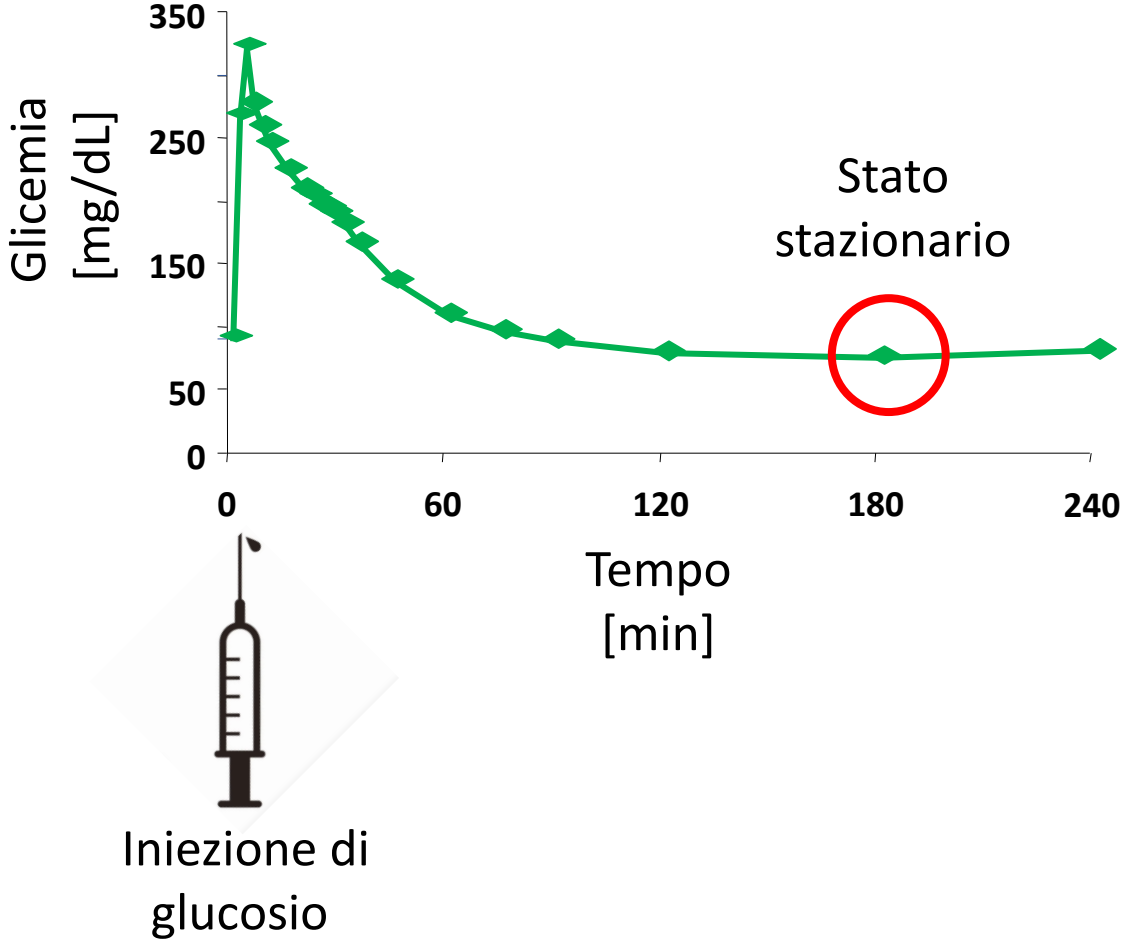
Test statistici in pratica (per la vita)

L'ipotesi nulla è un'ipotesi

- **Se il p-value è alto, ovvero i dati sembrano concordare con l'ipotesi nulla, questo NON MI DICE ASSOLUTAMENTE NULLA rispetto al fatto che l'ipotesi nulla sia vera oppure no.**
- **Se non assumo (cioè do per assodato, cioè faccio finta!) che l'ipotesi nulla sia vera, non posso neanche costruire la statistica del test e, quindi, non posso calcolare il p-value (che si basa sulla distribuzione della statistica del test).**
  - Sarebbe come dire "Se questo (l'ipotesi nulla) è vero, allora è vero con una certa probabilità uguale al p-value".
  - Ma che senso ha? Se è vero, è vero e basta; cosa c'entra la probabilità?
  - Ecco, dunque, il cortocircuito logico che ci indica che l'ipotesi nulla non si può dimostrare, ma va presa per buona e, eventualmente, confutata.

ESERCIZIO – Prima parte (punti da A a C) da effettuare a casa (codice a disposizione)

Utilizziamo una variante del dataset Intravenous Glucose Tollerance Test (IVGTT) dello laboratorio 2 (il nome file sarà diverso, perché non sono gli stessi dati).



Abbiamo a disposizione i dati di concentrazione di glucosio (glicemia) allo stato stazionario raccolti in individui appartenenti a due popolazioni:

1) **Anziani (`glucose_ss_elderly`)**
2) **Giovani (`glucose_ss_young`)**

I dati sono contenuti nel file **`IVGTT_SS.mat`**

- **`data`**: matrice double 2D (un gruppo di pazienti per colonna).
- **`labels`**: cell array con le etichette delle colonne (= nomi delle variabili / gruppi).
- **`units`**: cell array con le unità di misura.

NB: avere i dati nella matrice in questo modo è piuttosto inconsueto (perché dati sulla stessa riga non corrispondono allo stesso individuo), ma può capitare. Ci aiuta a fare esercizio di preprocessing.

Prima di svolgere l'esercitazione (oppure al bisogno), utilizzare il comando help di MATLAB seguito dal nome delle seguenti function, utili allo svolgimento degli esercizi.

Parte 1 (svolto)

- --

Parte 2 (proposto, per casa, da fare alla fine solo se avanza tempo)

- skewness, kurtosis, iqr

Parte 3 (svolto)

- [h, p, ci, stats] = ttest(dati, media_target)

Parte 4 (proposto)

- [h, p, ci, stats] = ttest(dati, media_target, 'Tail', 'right') oppure [h, p, ci, stats] = ttest(dati, media_target, 'Tail', 'left')

Parte 5 (proposto)

- --

Parte 6 (proposto)

- [h, p, ci, stats] = ttest(dati, media_target, 'Alpha', livello_di_significativita)

PARTE 1: CARICAMENTO E PULIZIA DATI (svolto)

- Caricare il file `IVGTT_SS.mat`

- Effettuare un'analisi naive sull'intera tabella, individuando e contando
  - I valori mancanti (`NaN`), con la funzione `isnan`
  - I valori infiniti (`Inf`), con la funzione `isinf`
  - I valori negativi, con la consueta indicizzazione logica

- Sostituire eventuali valori infiniti e valori negativi con NaN (v. laboratorio 3 per il razionale)

PARTE 2: CARICAMENTO E PULIZIA DATI (proposto, per casa, da fare alla fine solo se avanza tempo)

- Per ciascuna colonna della matrice `data`
  - Individuare e rimuovere i dati mancanti
  - Calcolare media e standard deviation
  - Calcolare mediana e IQR (funzione `iqr`)
  - Calcolare skewness e curtosi (`skewness`, `kurtosis`)
  - Mostrare a schermo tutte queste informazioni

- Rappresentare in una figura con 4 pannelli:
  - In alto a sinistra, il boxplot della variabile **`glucose_ss_elderly`**
  - In alto a destra, il boxplot della variabile **`glucose_ss_young`**
  - In basso a sinistra, l'istogramma delle frequenze assolute della variabile **`glucose_ss_elderly`**
  - In alto a destra, l'istogramma delle frequenze assolute della variabile **`glucose_ss_young`**

- Che idea ci possiamo fare sulla gaussianità delle due variabili?

PARTE 3: TEST STATICI (svolto)

Considerando solamente la variabile **`glucose_ss_elderly`**, privata dei suoi eventuali dati mancanti:

- Impostare un test statistico a un campione a due code per confrontare la media con il valore 100 mg/dL (livello di significatività $\alpha = 0.05$).
- Impostare un test statistico a un campione a due code per confrontare la media con il valore 82 mg/dL (livello di significatività $\alpha = 0.05$).
- Commentare entrambi i risultati ottenuti (con moltissima attenzione).

PARTE 4: TEST STATICI (proposto)

Considerando solamente la variabile **`glucose_ss_elderly`**, privata dei suoi eventuali dati mancanti:

- Impostare un test statistico a un campione e **a una coda** per confrontare la media con il valore 100 mg/dL (livello di significatività $\alpha = 0.05$; ipotesi alternativa di interesse: $\mu > 100$ mg/dL).
  - Commentare il risultato (facile).
  - Darsene una spiegazione (difficile, lo rivedremo al prossimo laboratorio).

PARTE 5: L'IPOTESI NULLA NON SI PUO' DIMOSTRARE (proposto)

Considerando solamente la variabile **`glucose_ss_elderly`**, privata dei suoi eventuali dati mancanti:

- Impostare due test statistici a un campione e a due code per confrontare la media con i valori 81.5 e 82.5 mg/dL (livello di significatività $\alpha = 0.05$).
  - Commentare il risultato di ciascun test individualmente (facile).
  - Considerare cosa si può dire dopo averli fatti entrambi ("difficile", ma istruttivo).

PARTE 6: ESERCIZIO DI SINTASSI (proposto)

Considerando solamente la variabile **`glucose_ss_elderly`**, privata dei suoi eventuali dati mancanti:

- Ripetere l'esercizio PARTE 3 punto 1 (confronto con 100 mg/dL) usando un livello di significatività $\alpha = 0.01$ (consultare autonomamente l'help di MATLAB).
- Commentare il risultato ottenuto