METODI STATISTICI PER LA BIOINGEGNERIA (B)

PARTE 6: VERIFICA DELL'ASSUNZIONE DI NORMALITA' E TEST STATISTICI NON PARAMETRICI

A.A. 2025-2026

Prof. Martina Vettoretti

L'ASSUNZIONE DI NORMALITA'

- I test statistici che abbiamo visto finora assumono che il campione aleatorio sia normale → applicabili solo se i dati presentano una distribuzione approssimativamente normale.
  - Per fortuna le quantità che analizziamo negli esperimenti presentano spesso una distribuzione approssimativamente normale.
- Come possiamo valutare se i dati hanno distribuzione approssimativamente normale?
- Non esiste un metodo che possa darci una risposta certa. Tuttavia possiamo avvalerci di una serie di strumenti statistici per cercare di capire se l'approssimazione normale possa essere ragionevole per descrivere i nostri dati.

STRUMENTI PER VERIFICARE L'ASSUNZIONE DI NORMALITA'

1. Istogramma
2. Test di normalità
3. Q-Q plot
4. Indici di forma campionari

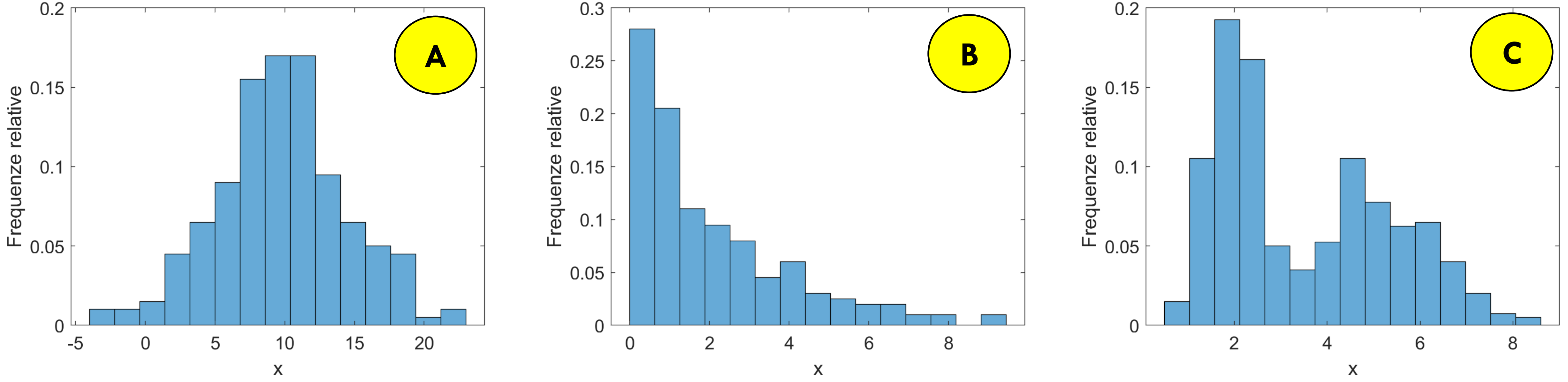
1. CHECK VISIVO DELL’ISTOGRAMMA

<u>Istogramma delle frequenze relative:</u> stima approssimata della ddp della variabile aleatoria X che definisce la distribuzione del campione.

- Se l’istogramma ha una forma che ricorda quella di una ddp normale (forma a campana, simmetrica, unimodale) → il campione potrebbe essere normale.
- Se l’istogramma ha una forma chiaramente lontana da quella della ddp normale (multi-modale, asimmetrica, …) → il campione con alta probabilità non è normale.

QUESTI CAMPIONI SONO APPROSSIMATIVAMENTE NORMALI?

Come valuteresti questi 3 campioni sulla base dell'analisi visiva dell'istogramma? Normali o non normali?



2. TEST DI NORMALITA'

- Esistono diversi test statistici per verificare l'ipotesi che il campione presenti una distribuzione normale.
- Dato un campione aleatorio composto dalle variabili aleatorie, $X_1, X_2, \dots, X_n$, di cui noi osserviamo una realizzazione, $x_1, x_2, \dots, x_n$, i test di normalità cercano di confutare l'ipotesi nulla:

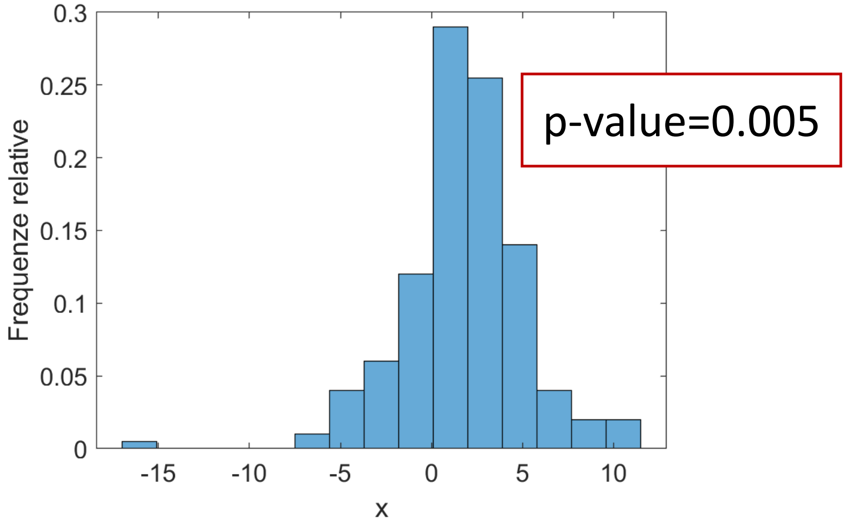
H_0: Il campione $X_1, X_2, \dots, X_n$ presenta distribuzione normale.

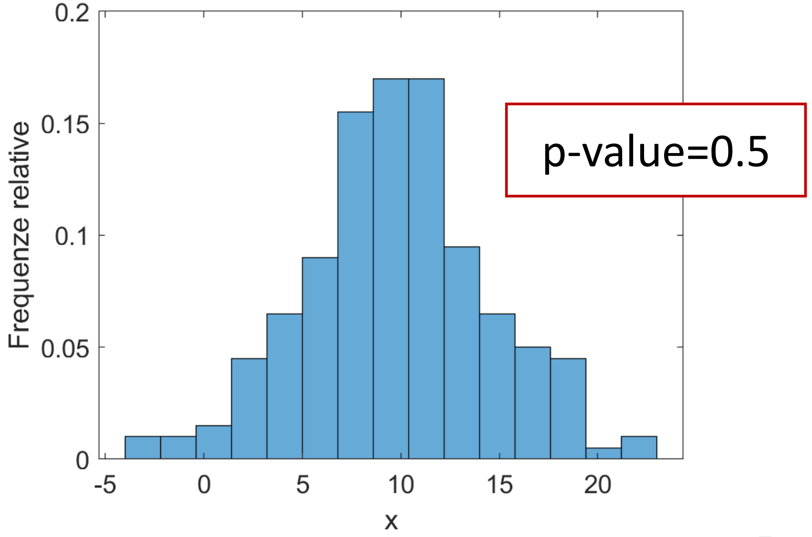
- I più noti test di normalità:
  - Test di Shapiro-Wilk (preferibile per campioni di piccole dimensioni)
    - Considera una statistica che è il rapporto tra due diversi stimatori della devianza.
  - Test di Kolmogorov-Smirnov
  - Test di Lilliefors

Basati sul confronto tra una stima campionaria della funzione di ripartizione e quella teorica.

UTILIZZO DEI TEST DI NORMALITA'

- Confronto del p-value restituito dal test di normalità con il livello di significatività scelto, α.
  - p-value $< \alpha$ → il test rifiuta H_0 → quindi cosa possiamo dire?
  - p-value $\geq \alpha$ → il test non rifiuta H_0 → quindi cosa possiamo dire?
- Nota: un test di normalità ci consente di dire quando un campione non è normale, ma **non ci consente di dire se un campione è normale.**





3. QUANTILE-QUANTILE (Q-Q) PLOT

Un metodo molto utilizzato per valutare la normalità è il **quantile-quantile plot** o **Q-Q plot**.

Idea:

- Si visualizzano su un grafico i quantili campionari del campione (asse y) vs. i corrispondenti quantili teorici della distribuzione normale standard (asse x).
- Se i punti si distribuiscono ragionevolmente lungo una retta → il campione ha distribuzione approssimativamente normale.

COSTRUZIONE DEL Q-Q PLOT

- F_X: funzione di ripartizione del campione $\mathbf{X} = (X_1, X_2, \dots, X_n)$.
- **Quantili campionari:** i dati x_i rappresentano gli n quantili campionari della distribuzione F_X di probabilità α_i, dove α_i è la frazione di elementi di $\mathbf{X}$ maggiori di x_i.
- **Quantili teorici della normale standard** di probabilità α_i:

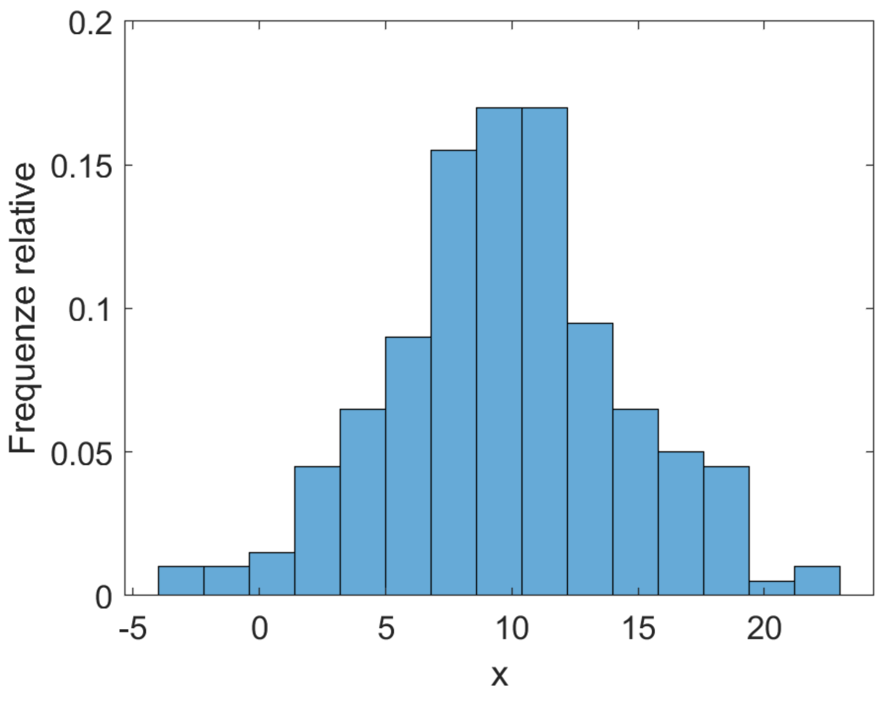
$$z_{\alpha_i} := \Phi^{-1}(1 - \alpha_i)$$

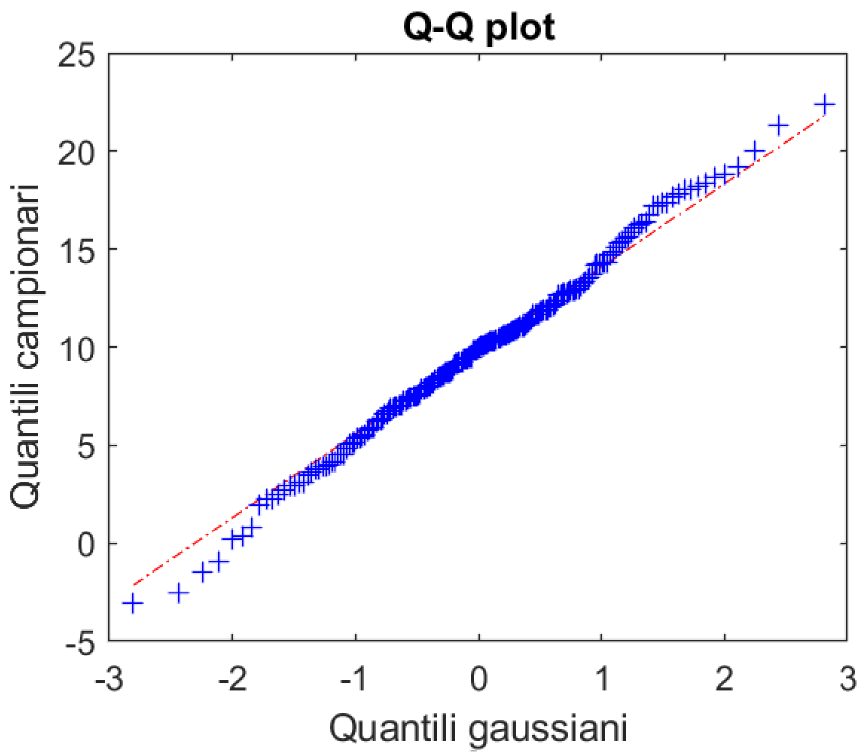
⬇

Q-Q plot: rappresentazione dei punti (z_{α_i}, x_i) sul piano cartesiano.

INTERPRETAZIONE DEL Q-Q PLOT (1/2)

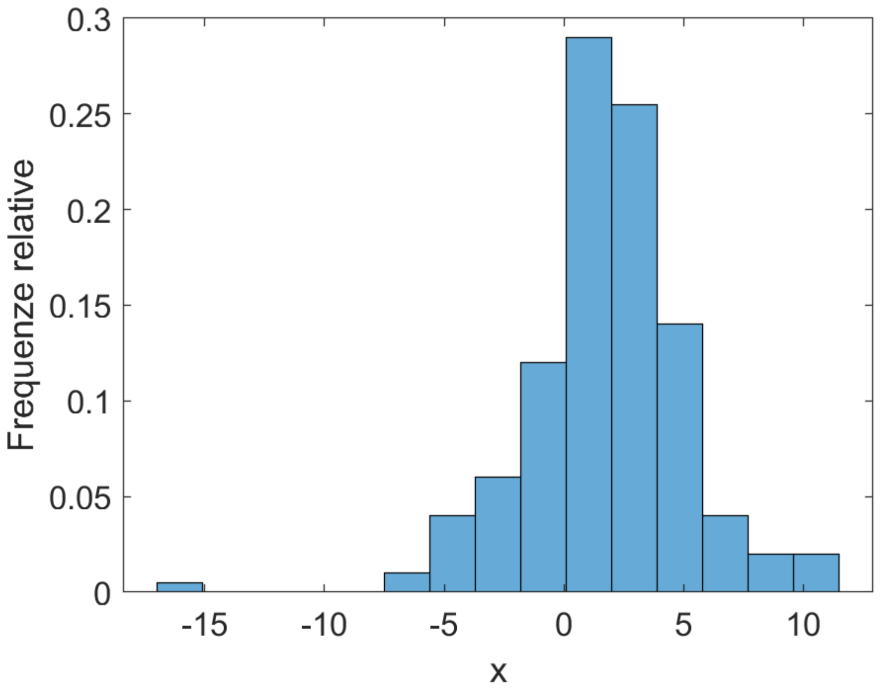
- Se i punti del q-q plot si dispongono lungo una retta → il campione ha distribuzione approssimativamente normale

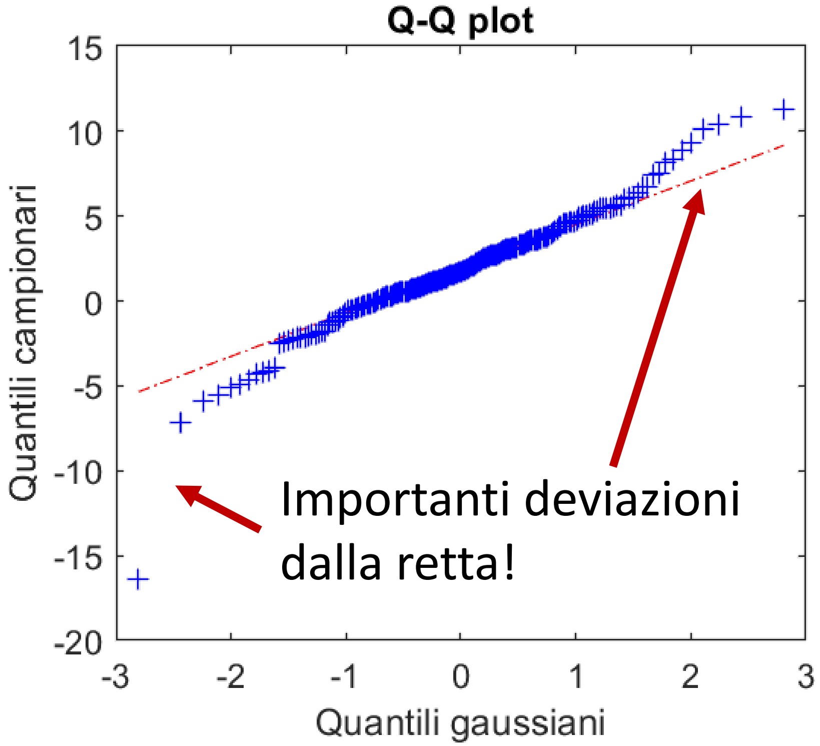




INTERPRETAZIONE DEL Q-Q PLOT (2/2)

- Se i punti del q-q plot non stanno su una retta → il campione non ha distribuzione approssimativamente normale



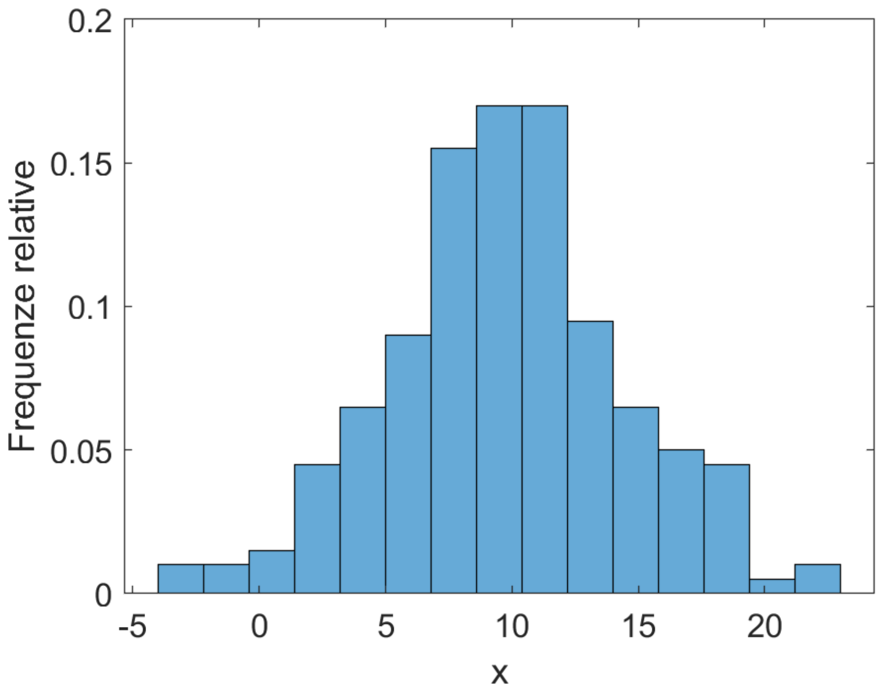


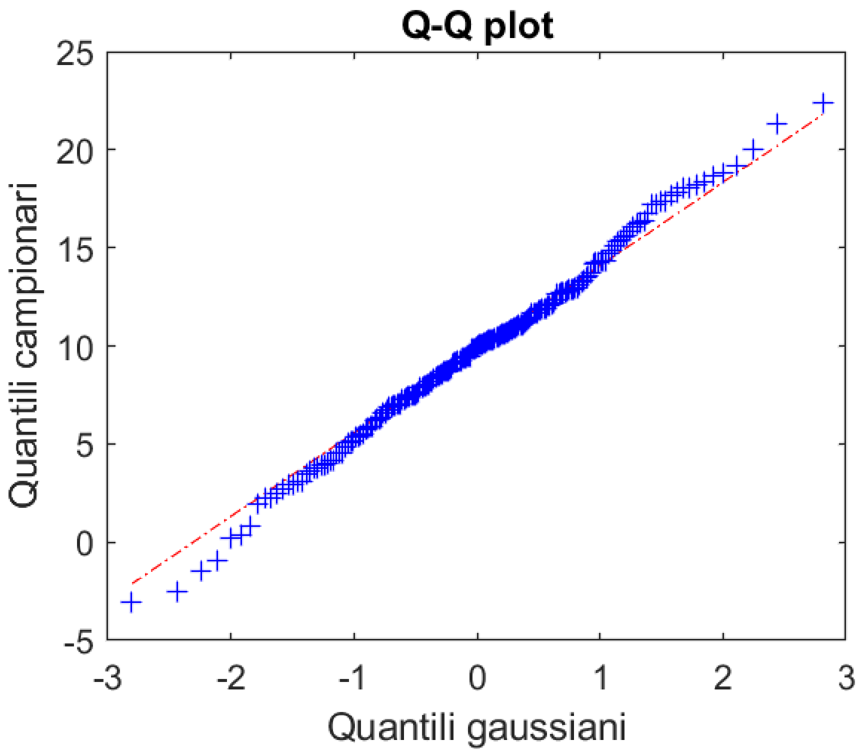
4. INDICI DI FORMA CAMPIONARI

Gli indici di forma campionari possono darci un'indicazione quantitativa sulla forma della distribuzione dei dati e quanto essa si discosti da quella di una distribuzione normale.

- **Indice di asimmetria o skewness campionaria:**
  - Se vicino a 0 → il campione ha distribuzione approssimativamente simmetrica → fatto compatibile con l'assunzione di normalità.
  - Se significativamente ≠ 0 → il campione ha distribuzione asimmetrica → con alta probabilità non è normale.
- **Indice di curtosi campionaria:**
  - Se vicino a 3 → il campione ha distribuzione «appuntita» come quella di una normale → fatto compatibile con l'assunzione di normalità.
  - Se significativamente ≠ 3 → poco verosimile che il campione sia normale.

ESEMPIO 1

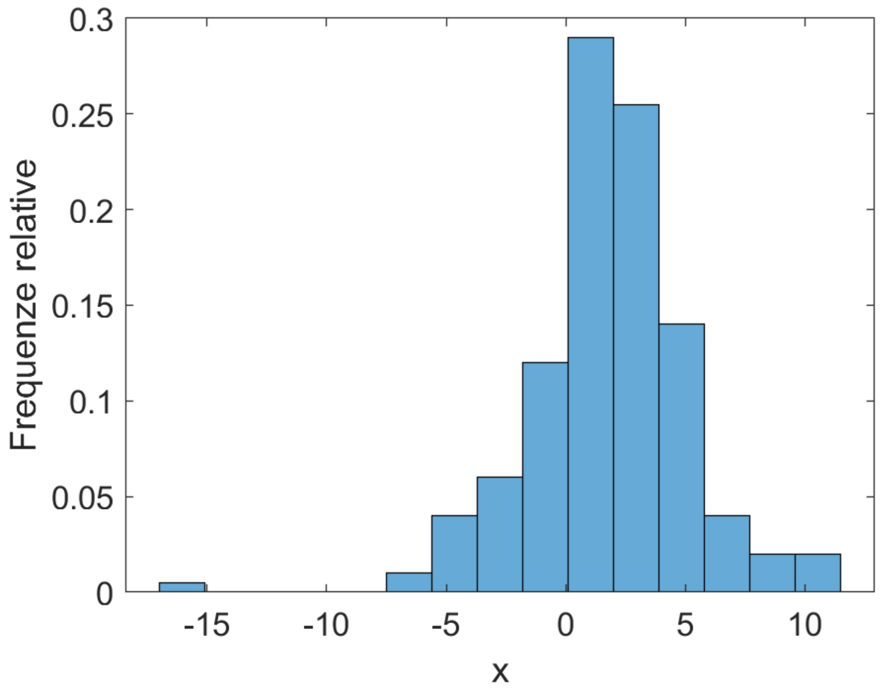


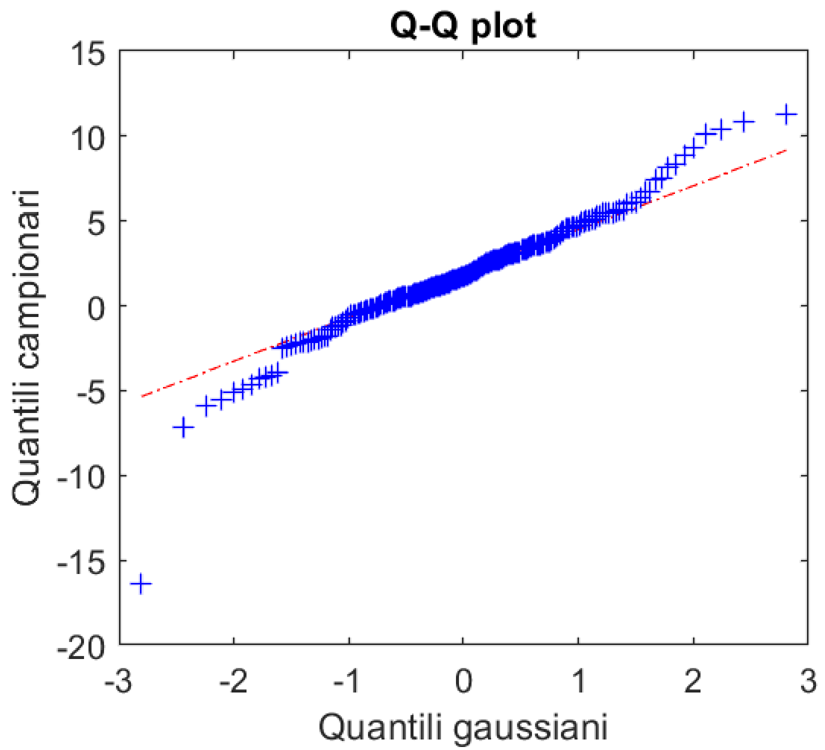


- Indice di skewness campionaria: -0.01
- Indice di curtosi campionaria: 3.09

Possiamo concludere che il campione è normale?

ESEMPIO 2





- Indice di skewness campionaria: -0.62
- Indice di curtosi campionaria: 7.17

Possiamo concludere che il campione è normale?

TEST STATISTICI NON PARAMETRICI

- **Test statistici non parametrici:** tecniche per verificare ipotesi su distribuzioni la cui forma o classe di appartenenza non è nota.
- Quando vi sono buone ragioni per supporre che il campione sia normale, conviene applicare i test parametrici che assumono che il campione sia normale, in quanto risultano più potenti.
- Se però l'assunzione di normalità non è verosimile per il campione in esame, e non possiamo neanche fare altre assunzioni sul tipo di distribuzione, conviene avvalersi di test statistici non parametrici.

TEST DEI SEGNI

Test dei segni: test non parametrico sulla <u>mediana</u> di un campione.

- Sia $X_1, X_2, \dots, X_n$ un campione avente distribuzione continua con funzione di ripartizione F_X incognita.
- Sistema di ipotesi:
  - $H_0\!: \mathrm{m} = m_0$
  - $H_1\!: \mathrm{m} \neq m_0$

  Dove m è la mediana (quantile di probabilità 0.5) delle variabili X_i ed m_0 è una costante.
- Il test restituisce un p-value che confrontato con il livello di significatività consente di rifiutare o meno H_0.
- Esiste anche il test dei segni unilaterale per testare i sistemi di ipotesi:
  - $H_0\!: \mathrm{m} \leq m_0$
  - $H_1\!: \mathrm{m} > m_0$

  - $H_0\!: \mathrm{m} \geq m_0$
  - $H_1\!: \mathrm{m} < m_0$

In Matlab: signtest

ESEMPIO

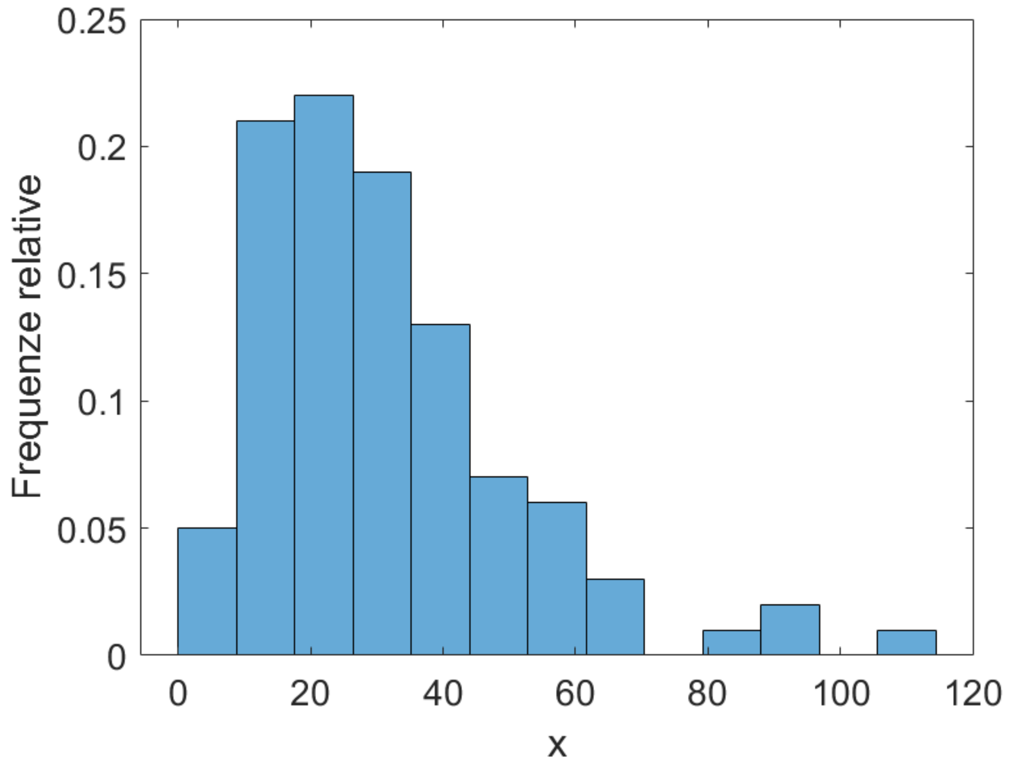
Analizziamo un campione di 100 dati avente mediana campionaria pari a 27.21 il cui istogramma è illustrato in figura. Ci chiediamo se la mediana della popolazione da cui è estratto il campione sia diversa da 30.

Applichiamo il test dei segni per verificare il sistema di ipotesi:

- $H_0: \mathrm{m} = 30$
- $H_1: \mathrm{m} \neq 30$

Il p-value risulta 0.1936.

Con un livello di significatività pari al 5%, cosa possiamo concludere?



TEST DI WILCOXON MANN-WHITNEY

Test di Wilcoxon Mann-Whitney (test di Wilcoxon, test U di Mann-Whitney o test della somma dei ranghi): test non parametrico per testare se due campioni aleatori <u>indipendenti</u> presentano la stessa distribuzione.

- Siano $X_1, X_2, \dots, X_n$ ed $Y_1, Y_2, \dots, Y_m$ due campioni aleatori indipendenti aventi funzioni di ripartizione F_X e F_Y, incognite.
- Assunzioni: le variabili aleatorie X_i e Y_j assumono valori ordinali.
- Sistema di ipotesi:
  - $H_0 \colon F_X = F_Y$
  - $H_1 \colon F_X \neq F_Y$
- In pratica il sistema di ipotesi può essere formulato sulla base delle variabili aleatorie che compongono i due campioni aleatori, nel seguente modo:
  - $H_0 \colon \mathrm{P}(X_i < Y_j) + \frac{\mathrm{P}(X_i = Y_j)}{2} = 0.5$
  - $H_1 \colon \mathrm{P}(X_i < Y_j) + \frac{\mathrm{P}(X_i = Y_j)}{2} \neq 0.5$

LA STATISTICA DEL TEST

- Mettiamo nello stesso insieme i dati $X_1, X_2, \dots, X_n$ e $Y_1, Y_2, \dots, Y_m$ e li ordiniamo dal più piccolo al più grande. La posizione che ciascun dato assume nell'ordinamento è il suo rango.
  - Esempio:

    X = {10.2, 2.7, 7.0, 8.9, 14.0, 1.5}   Y = {2.4, 12.7, 3.5, 4.0, 9.8}

Ranghi	1	2	3	4	5	6	7	8	9	10	11
Dati	1.5	2.4	2.7	3.5	4.0	7.0	8.9	9.8	10.2	12.7	14.0

- Si calcolano poi la somma dei ranghi dei valori di X, R_X, e la somma dei ranghi dei valori di Y, R_Y. Sapendo che:

$$\sum_{i=1}^{n} i \leq R_X \leq \sum_{i=m+1}^{n+m} i\,, \qquad \sum_{i=1}^{m} i \leq R_Y \leq \sum_{i=n+1}^{n+m} i$$

INTUIZIONE DIETRO LA DEFINIZIONE DELLA REGIONE CRITICA

- Se i dati provengono dalla stessa distribuzione, essi saranno ben mescolati ed R_X ed R_Y tenderanno ad un valore intermedio.
- **Idea:** si rifiuta H_0 quando R_X (o R_Y) presenta un valore troppo piccolo (più vicino al minimo) o troppo grande (più vicino al massimo).
- In pratica, il test deriva una statistica U, di definizione piuttosto complicata, che per campioni abbastanza grandi ha distribuzione approssimativamente normale.
- La statistica U viene utilizzata per definire il p-value sulla base del quale si decide se rifiutare o meno H_0 (al solito si rifiuta H_0 se p-value$<\alpha$).
- Esistono anche le versioni unilaterali del test.

In Matlab: ranksum

OSSERVAZIONE

- Quando le distribuzioni dei due campioni aleatori hanno la stessa forma e scala e differiscono solo per uno shift, ovvero:

$$F_Y(v) = F_X(v + \Delta), \text{ con } \Delta \text{ costante}$$

  Allora si può dimostrare che il test di Wilcoxon Mann-Whitney verifica il sistema di ipotesi:
  - $H_0 : m_X = m_Y$
  - $H_1 : m_X \neq m_Y$

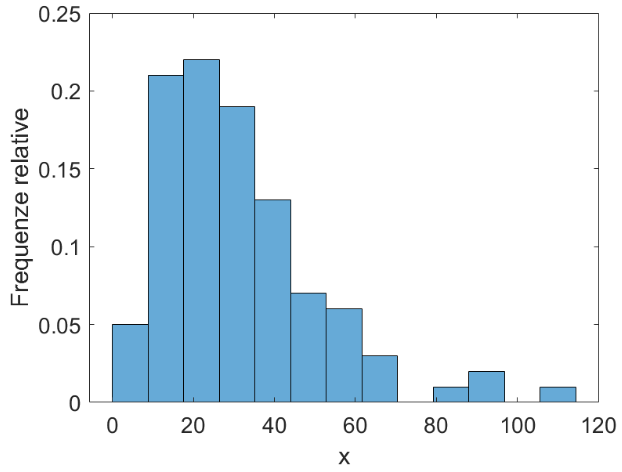
  Dove m_X e m_Y sono le mediane dei due campioni.

- **Attenzione**: nonostante il test di Wilcoxon Mann-Whitney sia spesso menzionato (anche in alcuni libri di statistica!) come test per verificare la differenza tra mediane, questo non è vero in generale. E' vero solo in un caso molto particolare, ovvero quando le due distribuzioni sono uguali eccetto che per uno shift sull'asse delle x.

George W. Divine, H. James Norton, Anna E. Barón & Elizabeth Juarez-Colunga (2018) The Wilcoxon–Mann–Whitney Procedure Fails as a Test of Medians, The American Statistician, 72:3, 278-286.

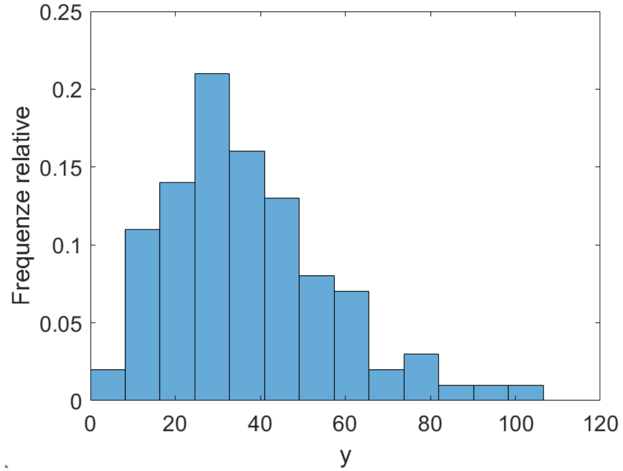
ESEMPIO

Analizziamo due campioni di dati indipendenti di dimensione 100 i cui istogrammi sono illustrati in figura. Ci chiediamo se i due campioni provengono da popolazioni con la stessa distribuzione.



Applichiamo il test di Wilcoxon Mann-Whitney per verificare il sistema di ipotesi:

- H_0: i dati provengono da popolazioni con stessa distribuzione
- H_1: i dati provengono da popolazioni con distribuzione diversa



Il p-value risulta 0.0158.

Con un livello di significatività pari al 5%, cosa possiamo concludere?

TEST DEI RANGHI CON SEGNO

Test dei ranghi con segno (in inglese Wilcoxon signed rank test): test non parametrico per il confronto di due campioni appaiati (dipendenti).

- Date le coppie di dati $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, i cui valori sono ordinali, il test dei ranghi con segno assume che le differenze $X_i - Y_i$ abbiano distribuzione simmetrica rispetto alla mediana e consente di testare il sistema di ipotesi:
  - H_0: le variabili aleatorie $X_i - Y_i$ hanno mediana 0.
  - H_1: le variabili aleatorie $X_i - Y_i$ hanno mediana $m \neq 0$
- Statistica del test: basata sulla somma dei ranghi delle differenze $X_i - Y_i$.
- Fissato il livello di significatività, si calcola il p-value sulla base del quale si decide se rifiutare o meno H_0 (al solito si rifiuta H_0 se p-value$<\alpha$).
- Esistono anche le versioni unilaterali del test.

In Matlab: signrank

ESEMPIO

Analizziamo due campioni appaiati di dimensione 150. Ci chiediamo se la differenza tra i due campioni, che assumiamo avere distribuzione simmetrica rispetto alla mediana, abbia mediana 0. La mediana campionaria è -6.36.

Applichiamo il test dei ranghi con segno per verificare il sistema di ipotesi:

- H_0: la mediana è 0
- H_1: la mediana è m$\neq 0$

Il p-value risulta 0.0261.

Con un livello di significatività pari al 5%, cosa possiamo concludere?