Analisi Dati con
Excel

Giovanni Della Lunga

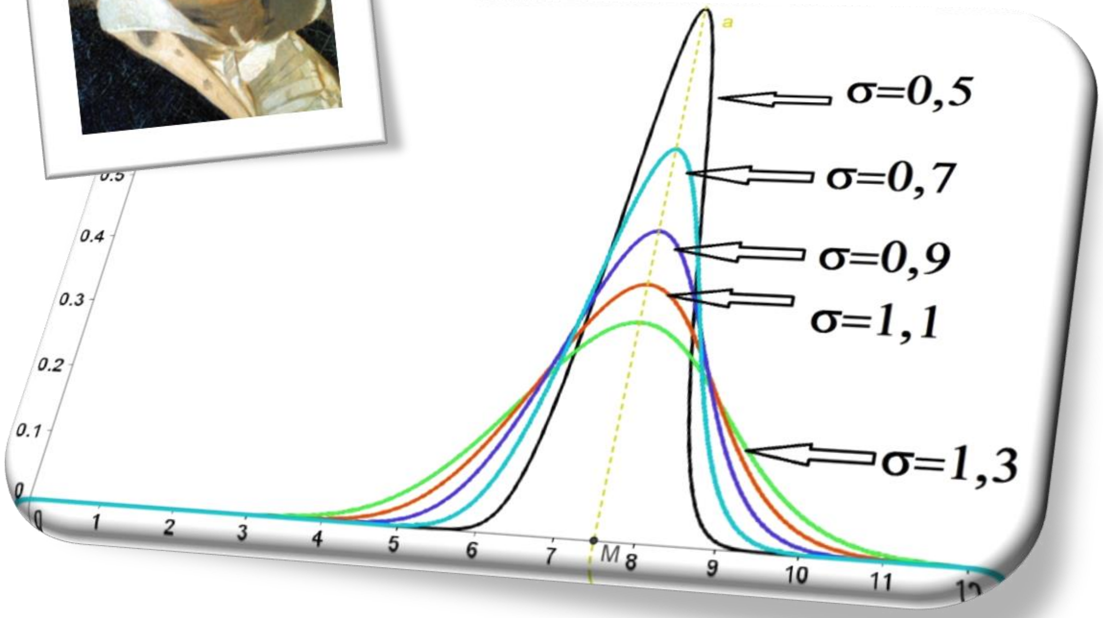
giovanni.dellalunga@gmail.com

La prima regola di ogni tecnologia è che l'automazione applicata ad un'operazione efficiente ne aumenterà l'efficienza. La seconda è che l'automazione applicata ad un'operazione inefficiente ne aumenterà l'inefficienza.

Bill Gates

La Distribuzione di Gauss

esempi e procedure per excel

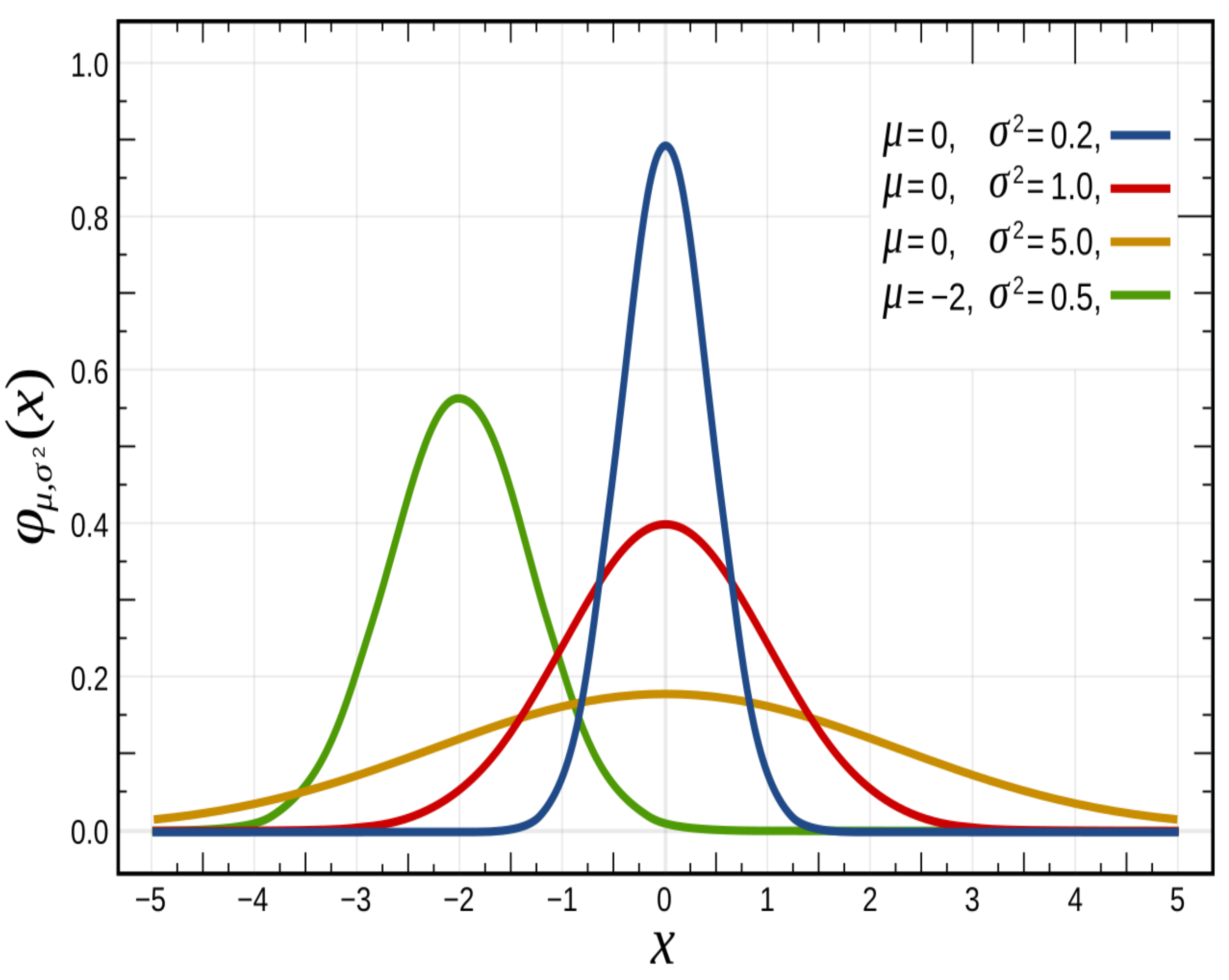


Introduzione alla Distribuzione di Gauss

» La distribuzione di Gauss, o **distribuzione normale**, è una distribuzione continua che descrive molti fenomeni naturali. È caratterizzata dalla forma a campana simmetrica.

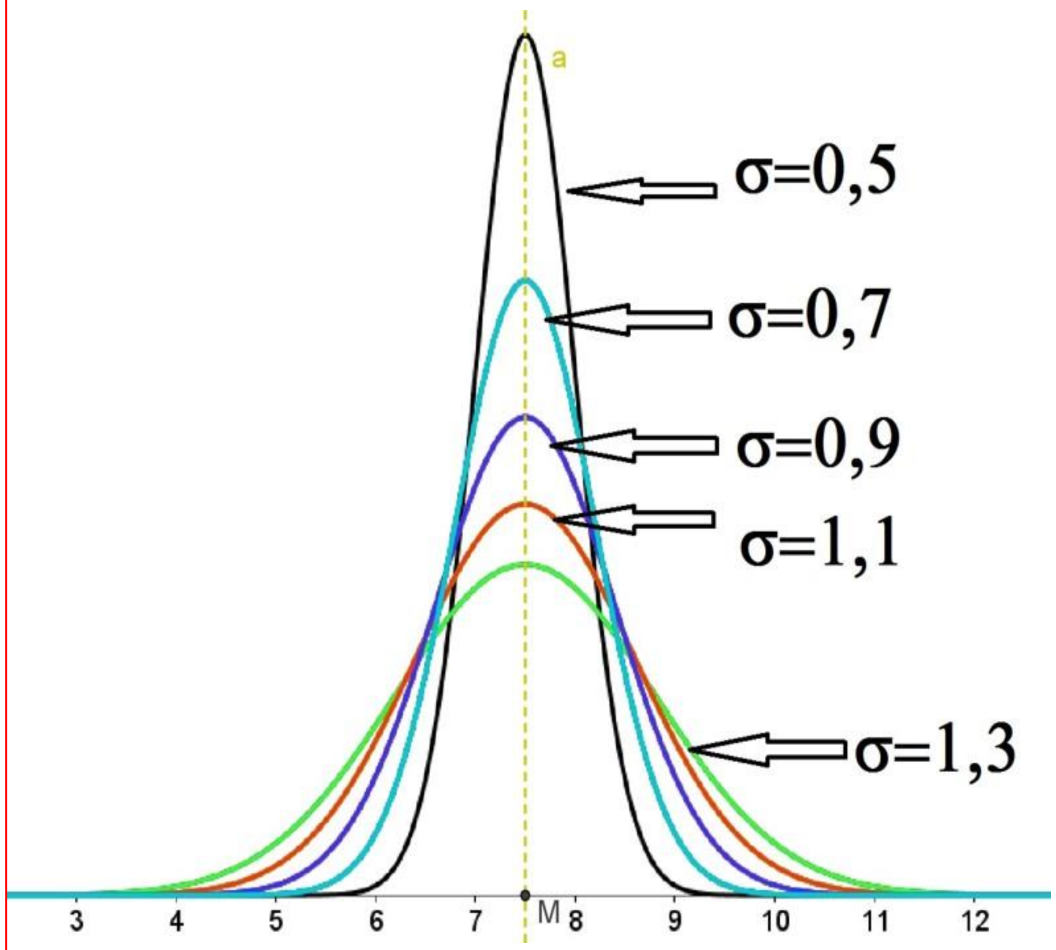
» La distribuzione normale è importante in statistica per tre motivi fondamentali:

1. Diversi fenomeni continui sembrano seguire, almeno approssimativamente, una distribuzione normale.
2. La distribuzione normale può essere utilizzata per approssimare numerose distribuzioni di probabilità discrete.
3. La distribuzione normale è alla base dell'inferenza statistica classica in virtù del teorema del limite centrale.



Caratteristiche della Distribuzione Normale

- 1. Media (μ): il valore centrale della distribuzione.
- 2. Deviazione Standard (σ): misura la dispersione dei dati.
- 3. Simmetria: la distribuzione è simmetrica rispetto alla media.
- 4. Asintoticità: le code della distribuzione si avvicinano all'asse x ma non lo toccano mai.



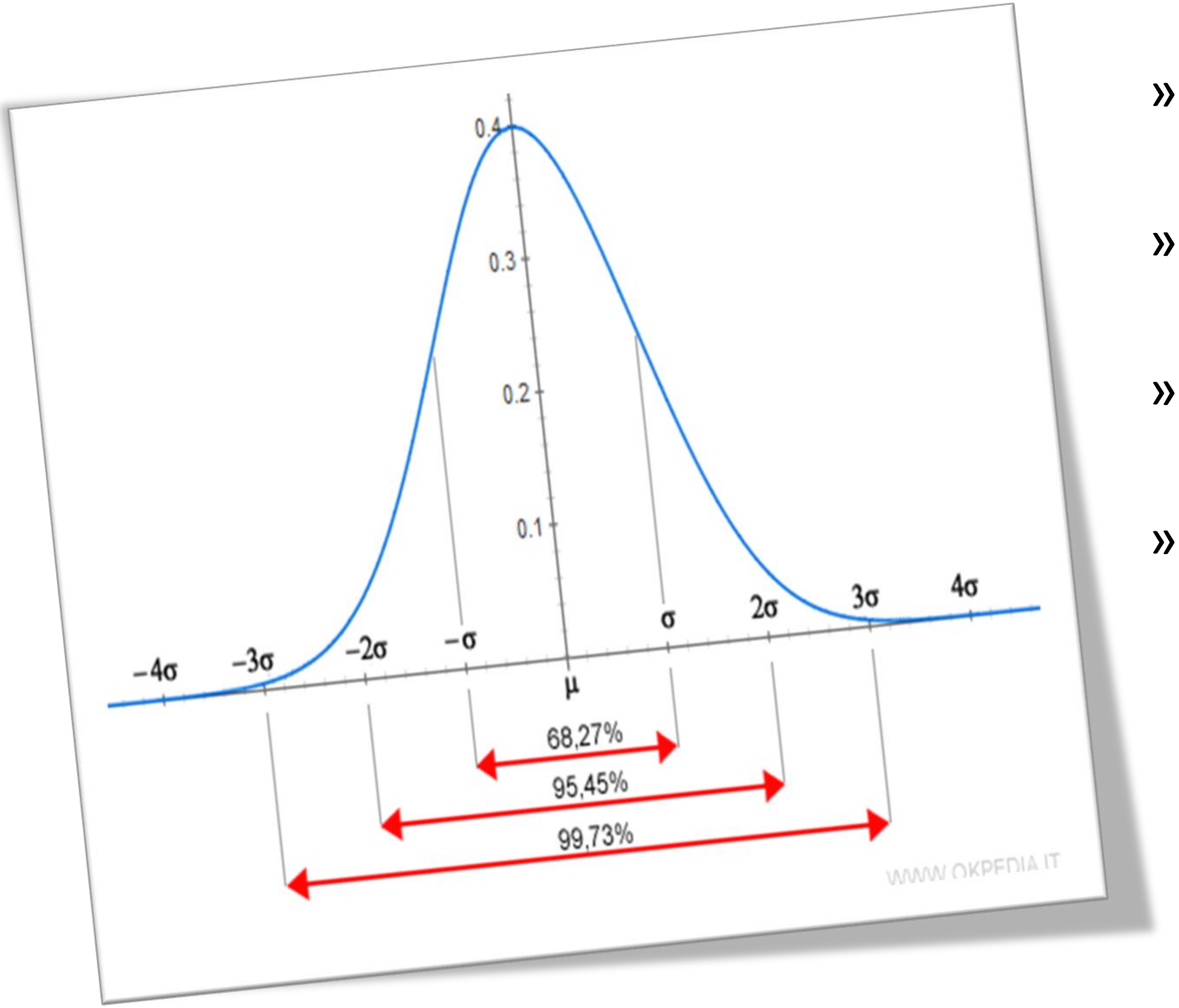
La Funzione di Densità di Probabilità

» La funzione di densità di probabilità (PDF) della distribuzione normale è data da:

$$\boxed{f_X(x) = f_X(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

» Dove μ è la media e σ è la deviazione standard.

Proprietà della Distribuzione Normale



» 1. La curva è simmetrica rispetto alla media.

» 2. Circa il 68% dei dati cade entro una deviazione standard dalla media.

» 3. Circa il 95% dei dati cade entro due deviazioni standard dalla media.

» 4. Circa il 99.7% dei dati cade entro tre deviazioni standard dalla media.

Esempio di Applicazione in Excel

» Per calcolare i valori di una distribuzione normale in Excel, puoi utilizzare la funzione NORM.DIST.

» Esempio:

» =NORM.DIST(x, media, deviazione_standard, cumulativo)

» Dove:

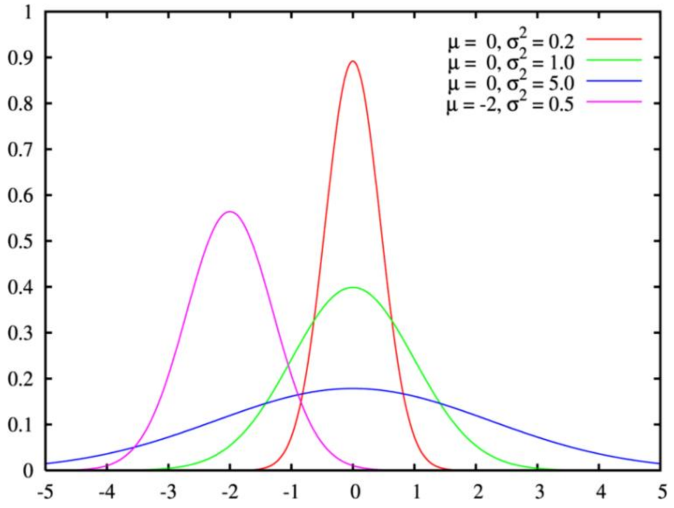
» - x è il valore per il quale si desidera calcolare la distribuzione.

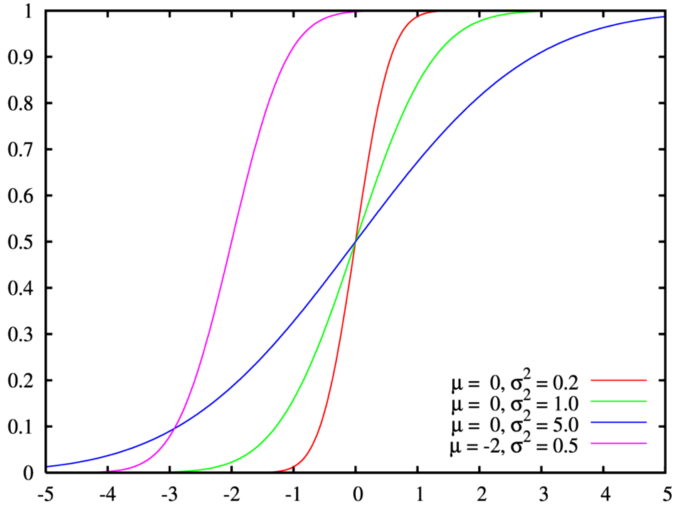
» - media è la media della distribuzione.

» - deviazione_standard è la deviazione standard della distribuzione.

» - cumulativo è un valore logico (TRUE per la funzione di distribuzione cumulativa, FALSE per la funzione di densità di probabilità).

PROBABILITA' | Distribuzioni Continue



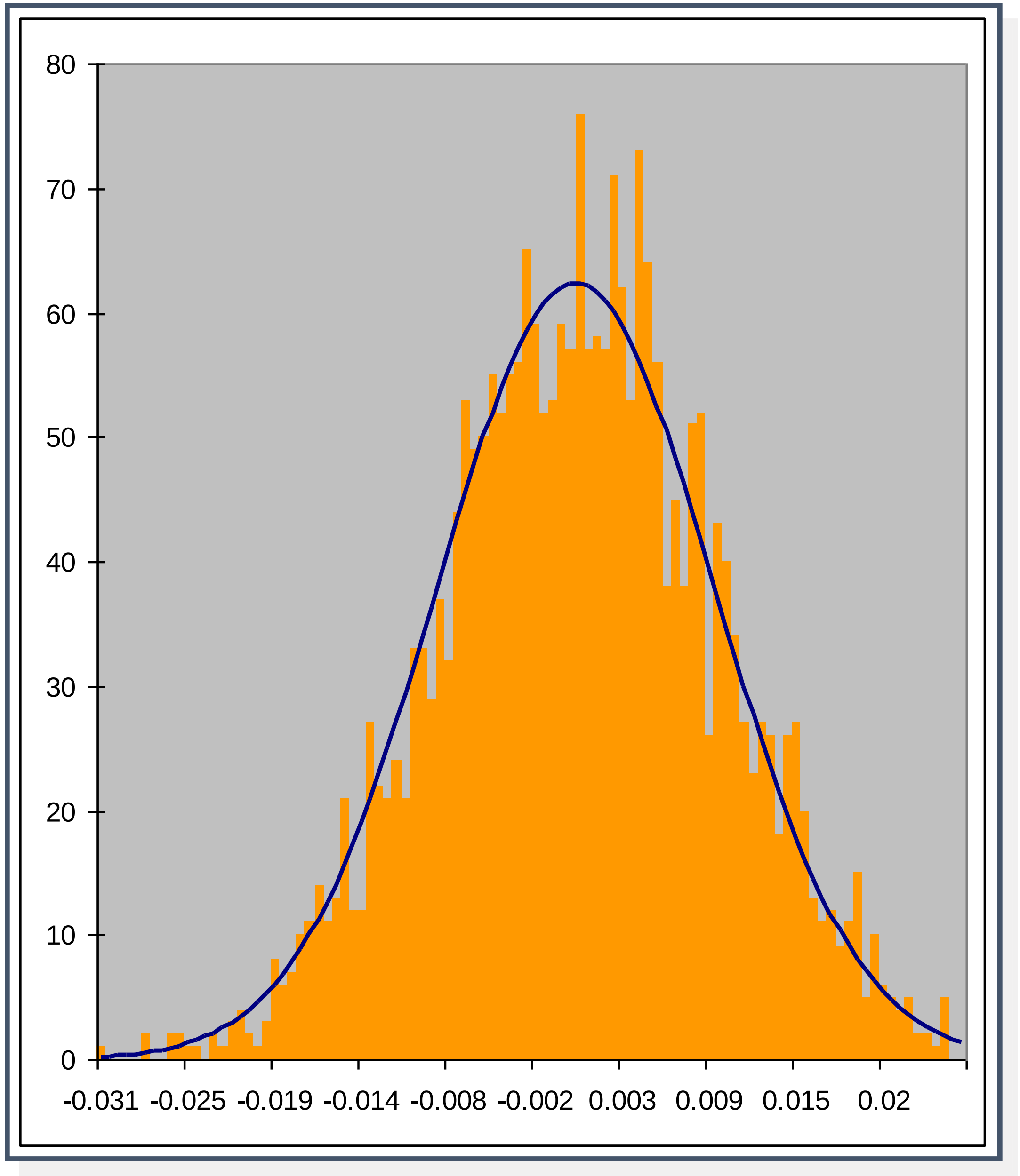


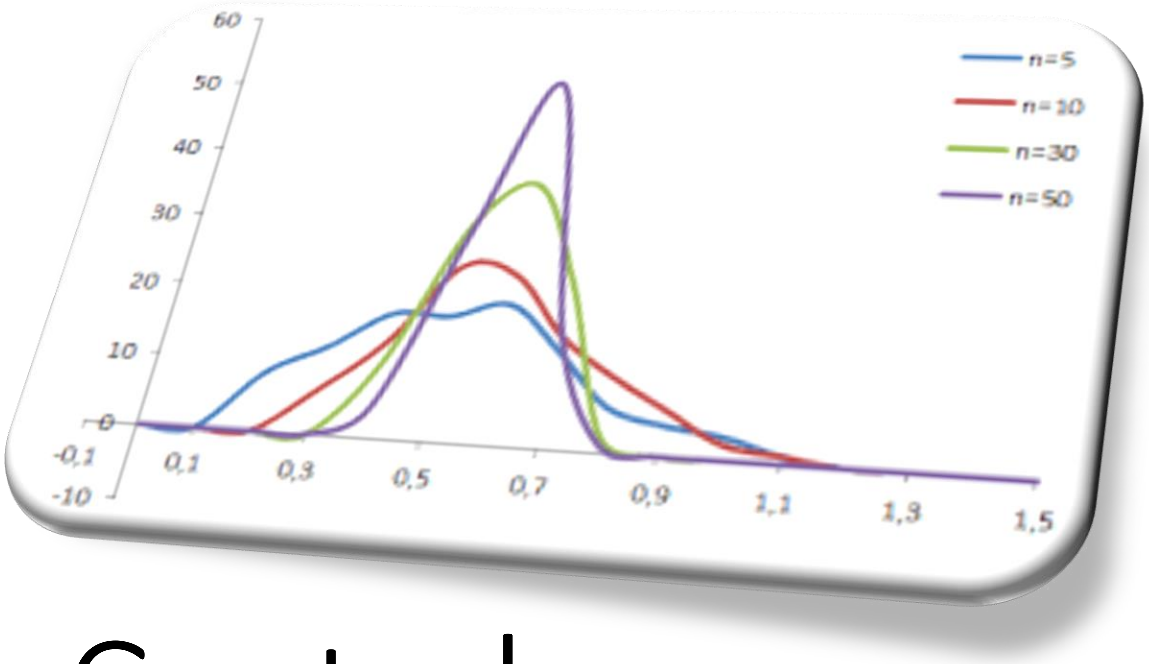
Parametri	$\mu \in \mathbb{R}, \sigma^2 \in (0,\infty)$
Supporto	$\mathbb{R}$
Funzione di densità	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$
Funzione di ripartizione	$\frac{1}{2}\left(1+\operatorname{erf}\frac{x-\mu}{\sigma\sqrt{2}}\right)$
Valore atteso	μ
Mediana	μ
Moda	μ
Varianza	σ^2
Skewness	0
Curtosi	0
Entropia	$\ln\left(\sigma\sqrt{2\,\pi\,e}\right)$
Funz. Gen. dei Momenti	$M_X(x) = \exp\left(\mu\,x + \frac{\sigma^2 x^2}{2}\right)$
Funz. Caratteristica	$\varphi_X(x) = \exp\left(\mu\,i\,x - \frac{\sigma^2 x^2}{2}\right)$

» Non dimenticate che la densità di probabilità rappresenta la frazione di valori che cadono all'interno di un certo intervallo della variabile aleatoria:

$$\frac{N}{N_{tot}} = f(x)\Delta x$$

$$N = f(x)\Delta x \cdot N_{tot}$$





Il Teorema Limite Centrale

Perché la distribuzione di Gauss è così importante

Il Teorema Limite Centrale

» Il teorema del Limite centrale (CLT) è un principio fondamentale nel campo della statistica che spiega perché molte distribuzioni tendono ad apparire normali in determinate condizioni.

» È la pietra angolare che ci permette di fare inferenze sulle popolazioni dai campioni.

» Il teorema afferma che, data una dimensione campionaria sufficientemente ampia, la distribuzione delle medie campionarie sarà approssimativamente distribuita normalmente, indipendentemente dalla distribuzione originale dei dati.

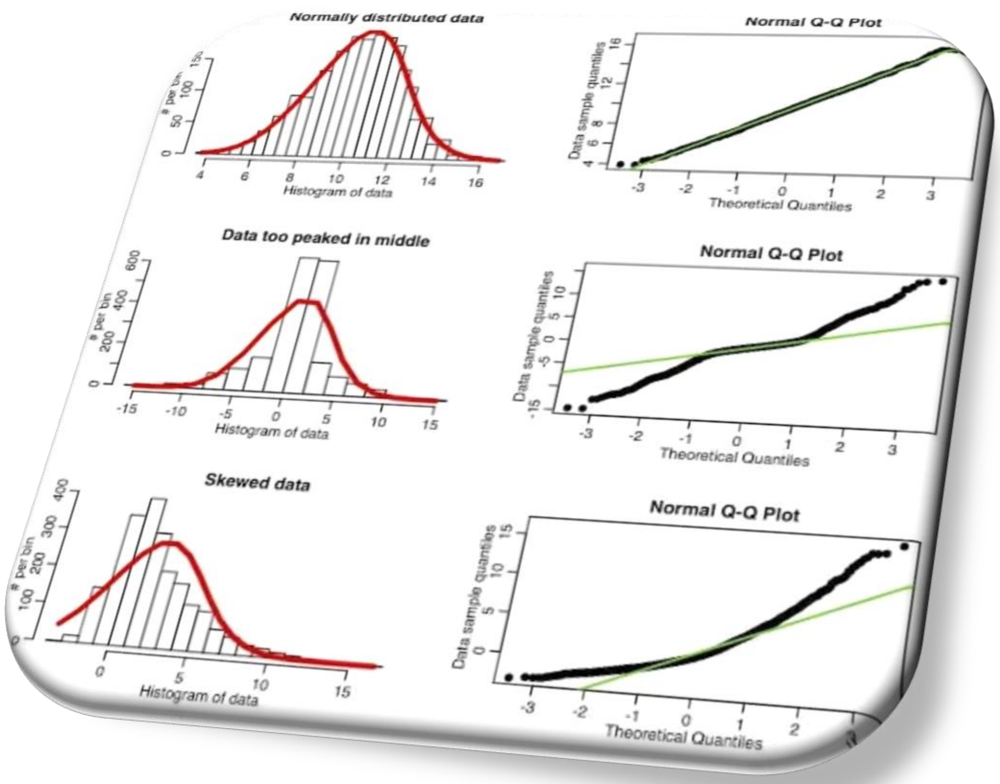
» Questo è un concetto potente perché si applica a un'ampia gamma di problemi e consente l'uso di tecniche di distribuzione normale anche quando i dati inizialmente non sembrano conformi a una distribuzione normale.

Il Teorema Limite Centrale

- Da un punto di vista pratico, il CLT ha un valore inestimabile.
- Ad esempio, nei processi di controllo qualità, aiuta a comprendere la distribuzione delle medie del campione, che può essere fondamentale per prendere decisioni sull'accettabilità del prodotto.
- In finanza, è alla base di molti modelli che presuppongono una distribuzione normale dei rendimenti.
- Da un punto di vista teorico, colma il divario tra la teoria della probabilità e l’inferenza statistica, fornendo un quadro robusto per la verifica delle ipotesi.

Il Teorema Limite Centrale

- La dimensione del campione gioca un ruolo cruciale nel CLT.
- Il teorema richiede tipicamente una dimensione del campione di almeno 30 affinché la distribuzione delle medie campionarie sia approssimativamente normale.
- Tuttavia, se la popolazione originaria è già distribuita normalmente, saranno sufficienti campioni anche più piccoli.
- Asimmetria e curtosi: sebbene il CLT affermi che la distribuzione delle medie campionarie sarà normale, ciò non implica che la distribuzione originale sia priva di asimmetria o curtosi.
- Queste misure di forma possono ancora essere presenti e possono influenzare l'accuratezza dell'approssimazione normale, soprattutto per campioni di dimensioni inferiori.



Q-Q plot ed uso in normality test

esempi e procedure per excel

Q-Q plot ed uso in normality test

- Alcune semplici tecniche grafiche possono essere molto utili per confrontare la distribuzione dei dati di un campione con una distribuzione teorica (utilizzo q-q plot in normality test vero e proprio) o con quella di un secondo campione.
- Tali rappresentazioni grafiche forniscono un approccio visivo e quindi più intuitivo nel confronto di due distribuzioni.

Q-Q plot ed uso in normality test

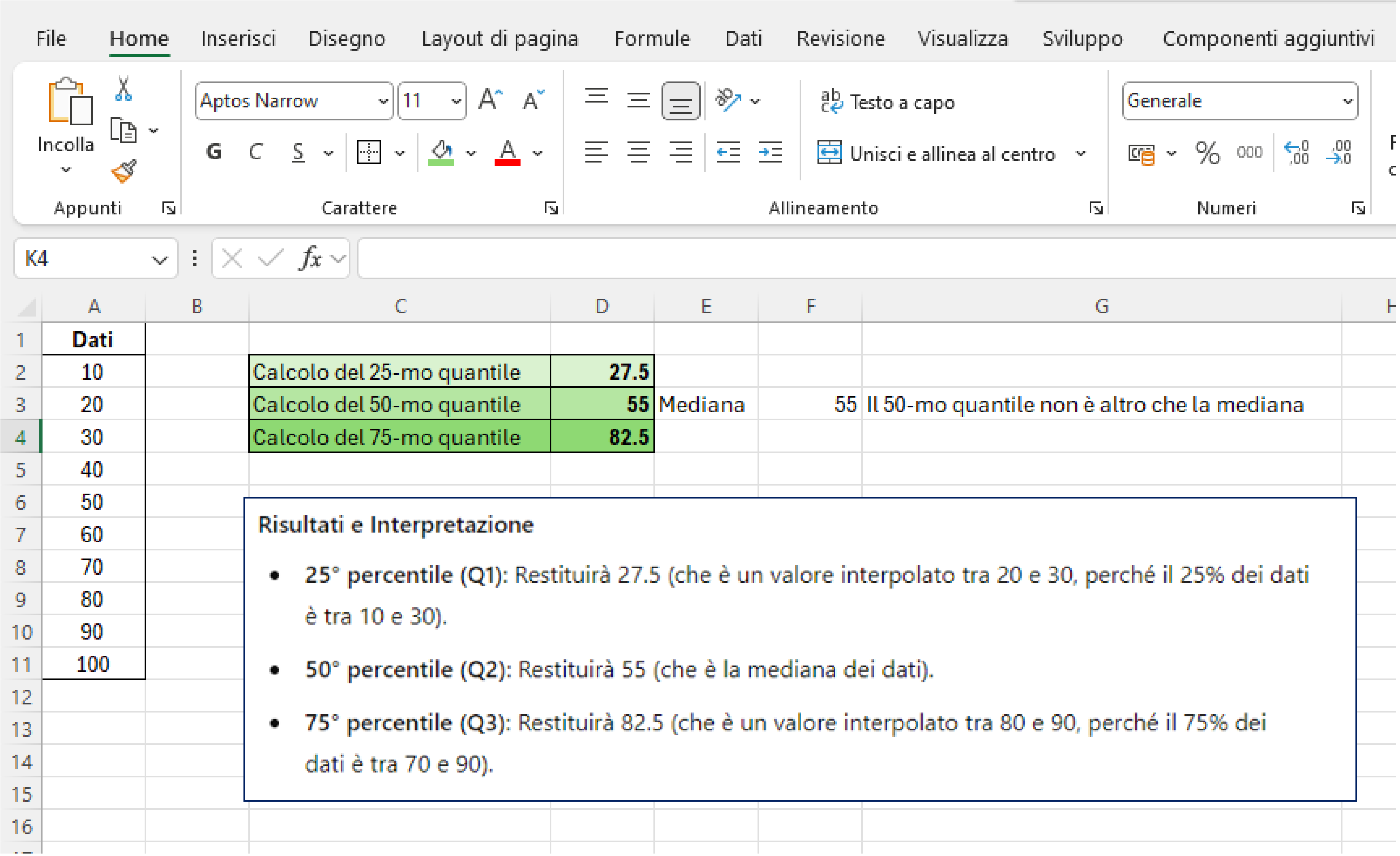
In particolare tali rappresentazioni consentono di:

- Verificare che i dati sperimentali seguano l’andamento di una distribuzione teorica o di un secondo campione
- Fornire informazioni su deviazioni da tale andamento (es. presenza di outlier o di code più larghe)
- Nel caso che si confronti il campione con una distribuzione teorica, consente di stimare i parametri che la distribuzione teorica deve possedere per descrivere meglio l’andamento del campione (es. deviazione standard e media)

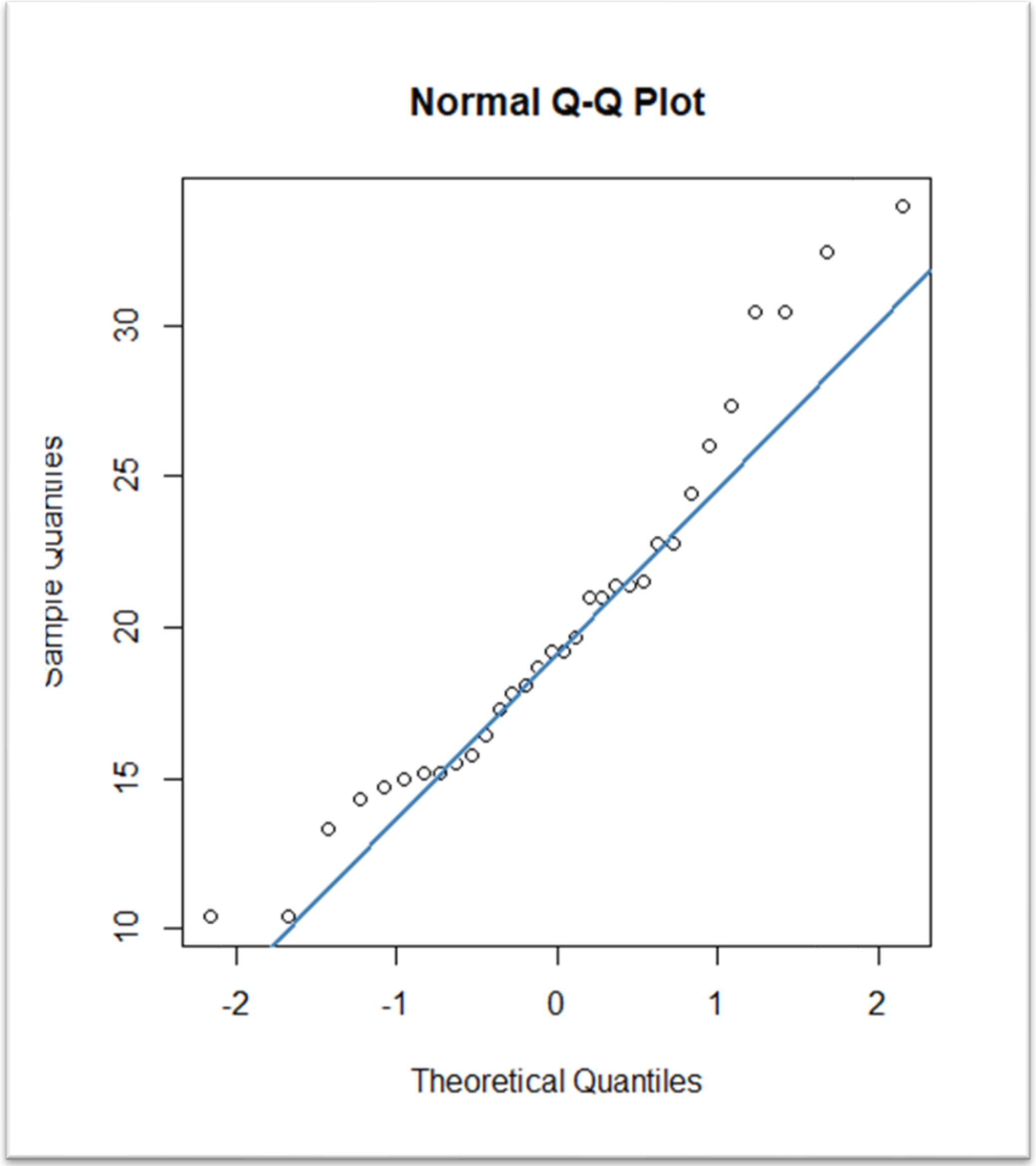
Q-Q plot o quantile-quantile plot

- Il Q-Q plot è la rappresentazione grafica dei quantili di una distribuzione (generalmente il campione) versus i quantili di una seconda distribuzione (distribuzione teorica o secondo campione).
- Per ennesimo quantile si intende il valore della distribuzione tale che l'ennesima percentuale dei suoi dati cade al di sotto di tale valore e la restante percentuale al di sopra di tale valore.
- Ad esempio per calcolare il primo decile di una distribuzione occorre ordinare i dati in modo crescente e individuare il valore per il quale il 10% dei dati giace al di sotto di esso.

Quantili: Un Esempio Pratico con Excel



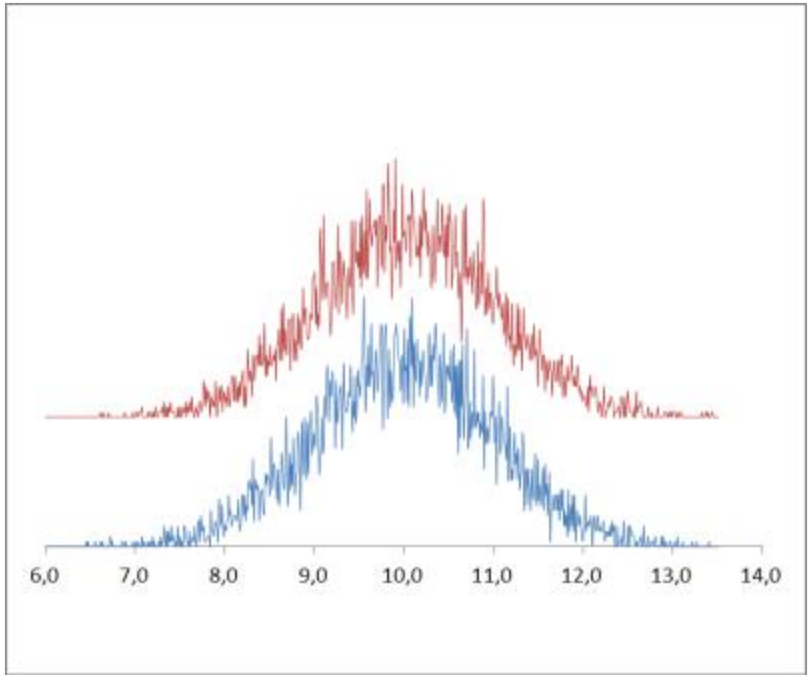
Q-Q plot o quantile quantile plot

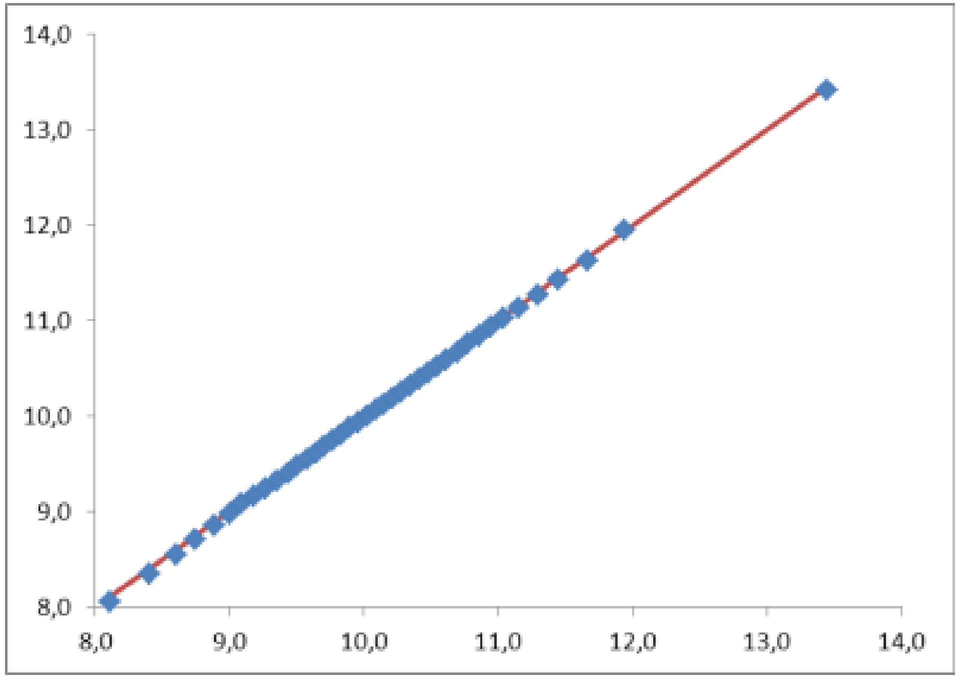


- Insieme al Q-Q plot viene disegnata anche la retta y = x che identifica il caso ideale di due distribuzioni identiche per le quali i quantili sono tutti identici.
- Le deviazioni del Q-Q plot rispetto tale retta permettono di identificare le deviazioni della prima distribuzione rispetto la seconda.
- Di seguito consideriamo alcuni esempi di Q- Q plot in modo tale da fornire con una panoramica dei diversi casi possibili una logica di interpretazione dei dati.

Caso 1: campioni con distribuzioni identiche

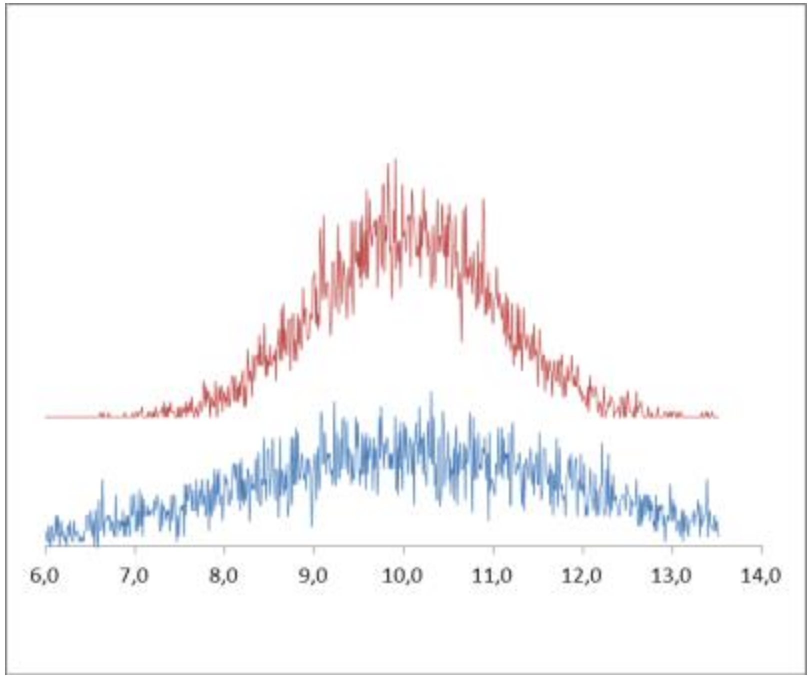
- Se due campioni provengono dalla medesima popolazione, essi saranno descritti dalla medesima distribuzione di probabilità.
- In questo caso il q-qplot (quadrati blu) si dispone esattamente lungo la retta y=x dimostrando che le due distribuzioni sono praticamente identiche.

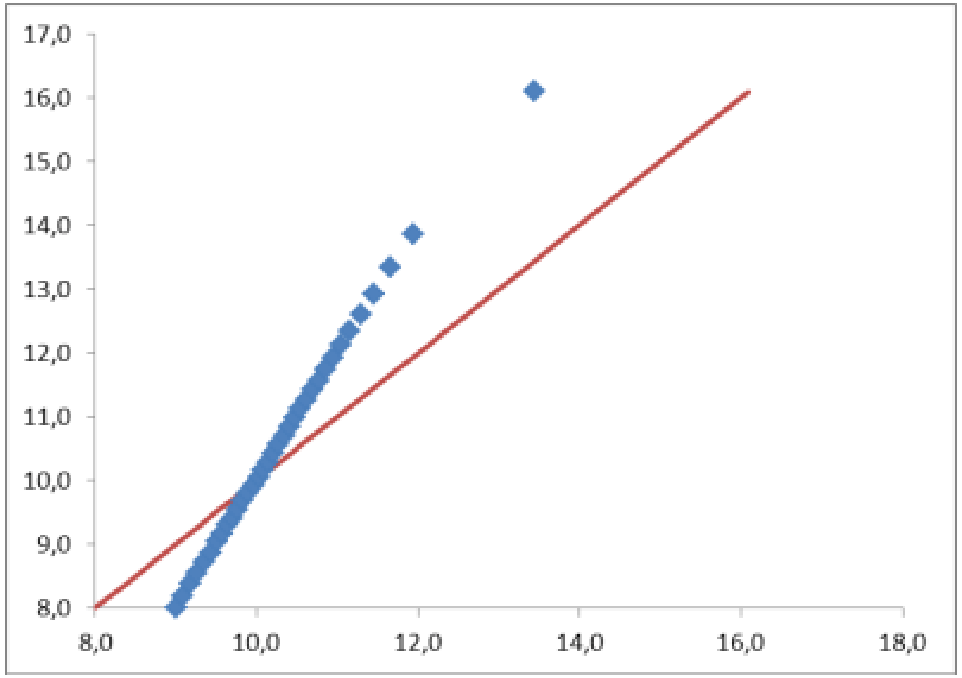




Caso 2: un campione ha una dispersione maggiore

- Come secondo caso si considera quello di due campioni aventi entrambi una distribuzione gaussiana ma con dispersione differente.
- Nel grafico sotto, infatti, la curva blu ha una standard deviation maggiore di quella mostrata dalla curva in rosso:



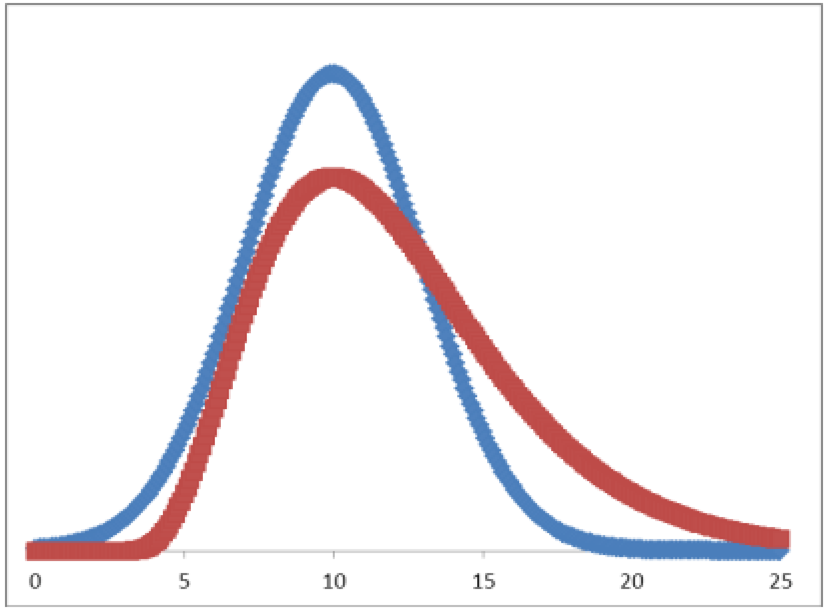


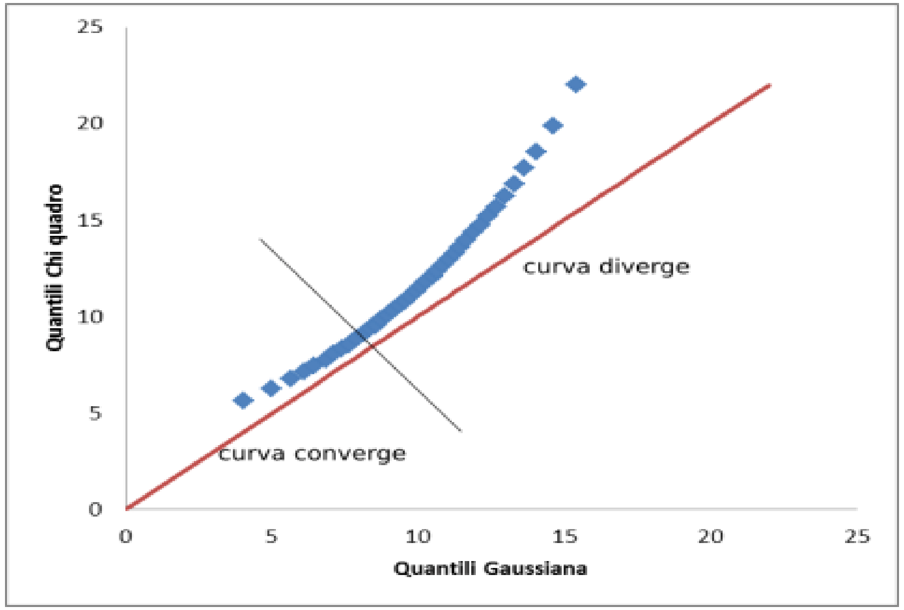
Caso 2: un campione ha una dispersione maggiore

- Il plot in questo caso non coincide con la retta. Le due distribuzioni, pur essendo della stessa famiglia (gaussiane), sono diverse.
- La pendenza elevata del plot indica che la prima distribuzione è più larga della seconda.
- I suoi percentili si estendono infatti da 8 a 16 mentre i percentili della seconda distribuzione si estendono in un range più corto, da 9 a 14.
- In generale per individuare quale delle due distribuzioni ha una o entrambe le code più larghe si guarda la pendenza del grafico alle due estremità.
  - Se ad esempio la pendenza dell'estremità a destra è più elevata della retta, allora sarà la distribuzione delle ordinate ad avere una coda più larga.
  - Stessa cosa per l'estremità sinistra.

Caso 3: distribuzioni diverse

- Si veda l'esempio di un q-q plot in cui si confrontano due distribuzioni molto diverse tra loro (le due distribuzioni sono una gaussiana ed una distribuzione chi quadro traslata)
- La distribuzione gaussiana (blu) è sempre in anticipo rispetto alla distribuzione chi quadro ed ha una coda leggermente più larga per valori più bassi di x e molto più stretta per valori molto alti.





Considerazioni generali sull'interpretazione di un q-q plot

Alla luce di quanto visto sopra si possono riassumere le seguenti considerazioni:

- Se il q-q plot giace sulla retta y=x (pendenza 45°) allora le due distribuzioni sono esattamente le stesse
- Se la pendenza è di 45 ° ma il plot si trova traslato su o giù rispetto alla retta y = x allora le due distribuzioni sono uguali ma con media diversa
- Qualsiasi convergenza/divergenza nei punti estremi va interpretata come diversità nella skewness delle distribuzioni
- Se il q-q plot è una retta con pendenza diversa da 45° allora le due distribuzioni hanno deviazione standard diversa
- Se il grafico non è una retta allora le due distribuzioni sono diverse

Come creare un grafico q-qplot in excel

Nel caso non si avessero a disposizione dei sofware dedicati per la creazione di un grafico q-q plot, è possibile realizzarlo in excel in pochi semplici passaggi. Di seguito l'elenco delle operazioni da eseguire per confrontare due set di dati:

- Scegliere il numero di percentili (punti sul grafico) da mostrare.
- Nel nostro esempio abbiamo scelto 50 punti riportando i percentili con un passo del 2%. Accanto alle colonne con i dati dei due campioni si riporta quindi una colonna con i valori 0,02; 0,04 ; 0,06; 0,08; 0,2 ….. 0,98
- Associare ad ogni % del punto 1 i percentili delle due distribuzioni (utilizzare funzione percentile)
- Eseguire uno scatterplot plottando i valori dei quantili del campione 1 vs quelli del campione 2
- Inserire nel grafico la retta y=x

Q-Q plot per normality test in excel

- In quest’ultima parte si vedrà quali sono i passaggi da eseguire in excel per poter ottenere un q-q plot in excel con l’intento di verificare che la distribuzione dei dati possa essere descritta da una gaussiana teorica.
- Come detto nei paragrafi precedenti si tratta del confronto del nostro campione con una distribuzione teorica.
- Ma cosa succede se non sappiamo con quale tipo di gaussiana confrontare i nostri dati?
- La risposta è semplice: si utilizza la distribuzione normale (gaussiana con media 0 e varianza 1).

Q-Q plot per normality test in excel

Di seguito le operazioni per eseguire il q-q plot per normality test:

- Distribuire i dati del campione in ordine crescente
- Inserire una colonna detta rango in cui si riporta il numero d'ordine del dato rispetto alla totalità del campione
- Calcolare media e deviazione standard del campione
- Normalizzare la distribuzione del campione mediante la formula:

$$Z = \frac{X - \mu}{\sigma}$$

Q-Q plot per normality test in excel

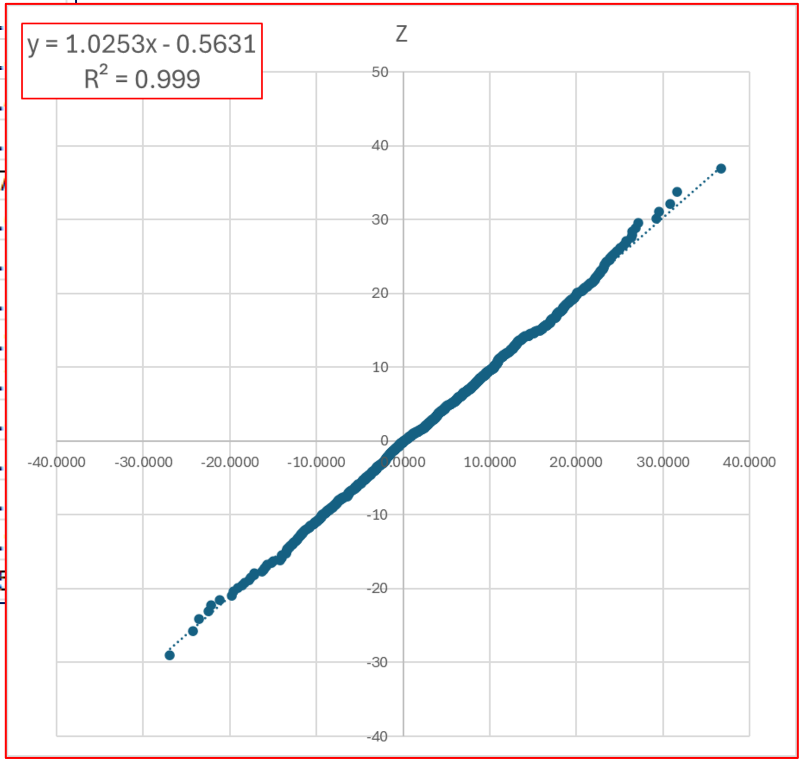
- Diversamente da quanto fatto detto nel paragrafo precedente, invece di assegnare delle % dalle quali calcolare i percentili del campione, si calcolano i valori delle percentuale ai quali corrispondono percentili pari ai valore del campione normalizzato.
- In altri termini, considerato il primo valore del nostro campione ci si chiede: a quale percentile corrisponde tale valore?
- Nel caso di un campione con 100 dati al dato più piccolo dovrebbe essere associato il valore 1%.
- Solitamente si utilizza la formula di Hazen:

$$\boxed{percentuale = \frac{rango\ del\ dato\ - 0.5}{numero\ di\ dati}}$$

Q-Q plot per normality test in excel

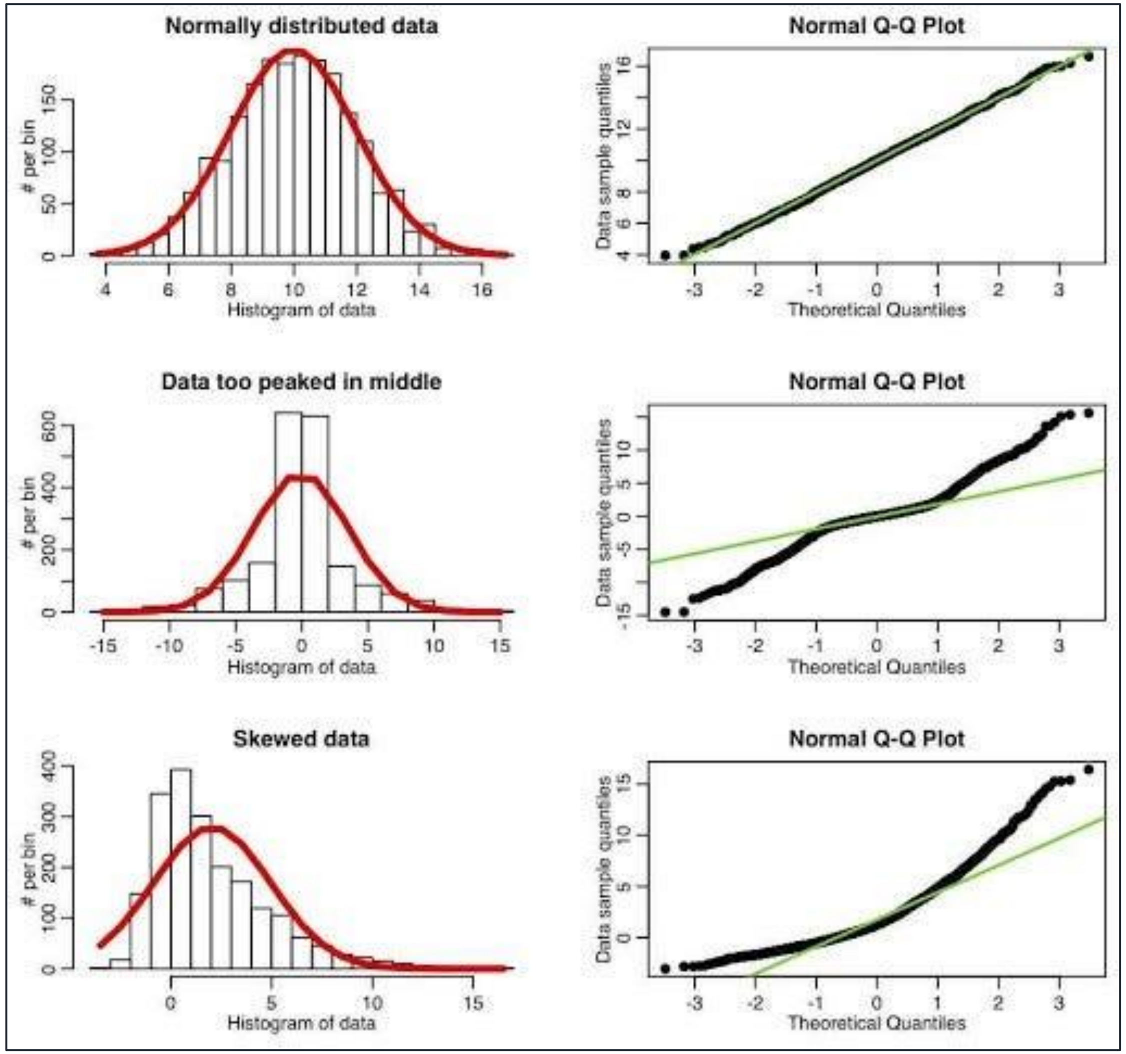
- Il punto precedente ci dice che i percentili del campione sono proprio i valori del nostro campione normalizzato.
- A questo punto andrebbero calcolati i percentili della distribuzione normale standard (detti anche z-score).
- Questo viene eseguito mediante la formula NORM.S.INV

dato	rango	plot pos	Z
-26.9450	1.0000	0.0005	-28.9582
-24.2526	2.0000	0.0015	-25.7352
-23.5784	3.0000	0.0025	-24.1309
-22.4702	4.0000	0.0034	-23.0311
-22.2012	5.0000	0.0044	-22.1848
-21.2153	6.0000	0.0054	-21.4927
-19.8284	7.0000	0.0064	-20.9047
-19.5709	8.0000	0.0074	-20.3919
-19.1015	9.0000	0.0083	-19.9362
-18.5667	10.0000	0.0093	-19.5254
-18.3245	11.0000	0.0103	-19.
-17.7969	12.0000	0.0113	-18.
-17.5440	13.0000	0.0123	-18.
-17.2116	14.0000	0.0132	-18.
-17.2017	15.0000	0.0142	-17
-16.3410	16.0000	0.0152	-17.
-16.1186	17.0000	0.0162	-17.
-15.9496	18.0000	0.0172	-17.
-15.7628	19.0000	0.0182	-16.
-15.7390	20.0000	0.0191	-16.
-15.2381	21.0000	0.0201	-16.
-14.9921	22.0000	0.0211	-16.
-14.2055	23.0000	0.0221	-16.
-14.0732	24.0000	0.0231	-15.
-13.9958	25.0000	0.0240	-15



Q-Q plot per normality test in excel

- Il punto precedente ci dice che i percentili del campione sono proprio i valori del nostro campione normalizzato.
- A questo punto andrebbero calcolati i percentili della distribuzione normale standard (detti anche z-score).
- Questo viene eseguito mediante la formula NORM.S.INV



Riassunto: Ipotesti di Normalità

Creazione di un Normal Probability Plot in Excel

» Ordina i Dati: Metti i tuoi dati in una colonna e ordinali dal più piccolo al più grande.

» Calcola i Quantili Teorici: Usa la funzione NORM.INV per calcolare i quantili teorici.

» Per ogni dato ordinato, calcola il quantile usando la formula:

=NORM.INV((RANGO - 0.5) / N, MEDIA, DEV.STANDARD)

dove:

RANGO è la posizione del dato nell'ordine (1 per il più piccolo, 2 per il secondo più piccolo, ecc.).

N è il numero totale di dati.

MEDIA e DEV.STANDARD sono la media e la deviazione standard dei dati osservati.