




# DATA LAB

## GUARDA AVANTI

**Big Data**, nuove competenze  
per nuove professioni.



"Anticipare la crescita con le nuove competenze sui Big Data" Operazione Rif. PA 2023-19167/RER approvata con DGR n° 843 del 29 maggio 2023 e co-finanziata dal Fondo Sociale Europeo Plus 2021-2027 Regione Emilia-Romagna

The background of the entire slide is an underwater photograph. On the left, a scuba diver in a black wetsuit and yellow fins is visible, swimming towards the right. On the right side, a massive, dense school of silver fish, likely sardines, fills the water column. The water is a deep blue, and bubbles are visible rising from the diver.

Operazione Rif. PA 2023-19167/RER/10/1, "ANTICIPARE LA CRESCITA CON LE NUOVE COMPETENZE SUI BIG DATA", approvata dalla Regione Emilia-Romagna con DGR n° 843 del 29/05/2023 e co-finanziata dal Fondo Sociale Europeo Plus 2021-2027

**DATA LAB**

The logo for Data Lab, featuring a stylized white icon of three human figures standing together, enclosed within a white circle.



# DISTRIBUZIONE NORMALE (GAUSS)

Quando si esegue un esperimento e si descrivono i risultati si costruisce spesso un grafico (istogramma) per mostrare l'andamento del fenomeno in esame.

In un istogramma:

- sull'asse delle ascisse ( $x$ ) poniamo i valori della variabile
- sull'asse delle ordinate ( $y$ ) poniamo le frequenze con le quali un determinato valore, un intervallo di valori in caso di variabili continue, si presenta.

L'area delle colonnine di un istogramma rappresenta la frequenza con cui i valori  $x_1$  e  $x_2$  che delimitano la base della colonnina si presentano nel nostro esperimento

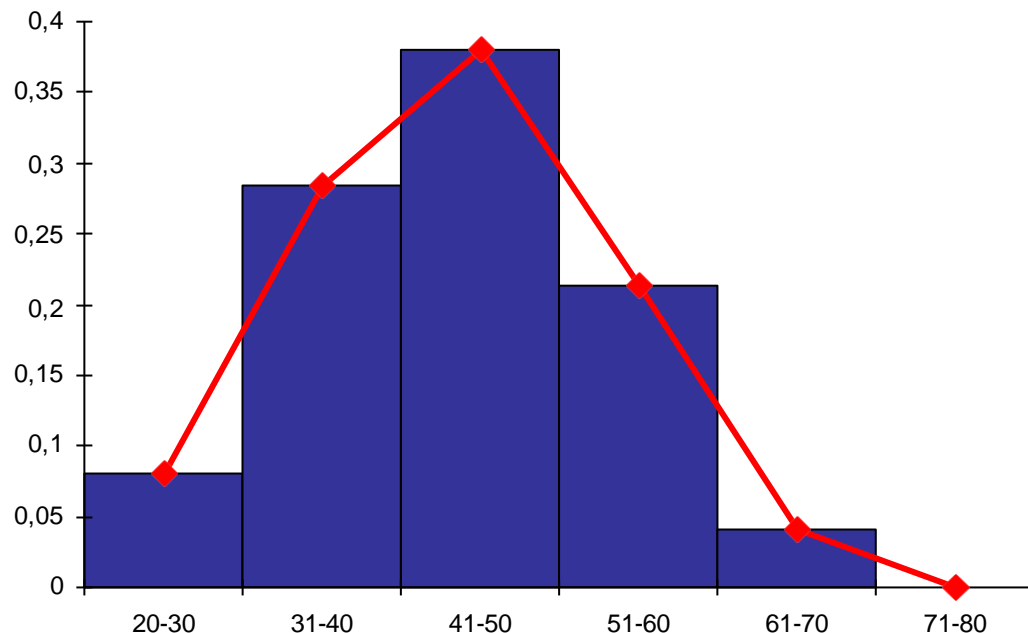
Unendo i punti medi delle classi con una spezzata si costruisce un poligono di frequenza.

Se scegliamo l'intervallo tra i valori sempre più piccolo e con una quantità di prove molto grandi la forma del poligono di frequenza si avvicinerà sempre più ad una linea continua.

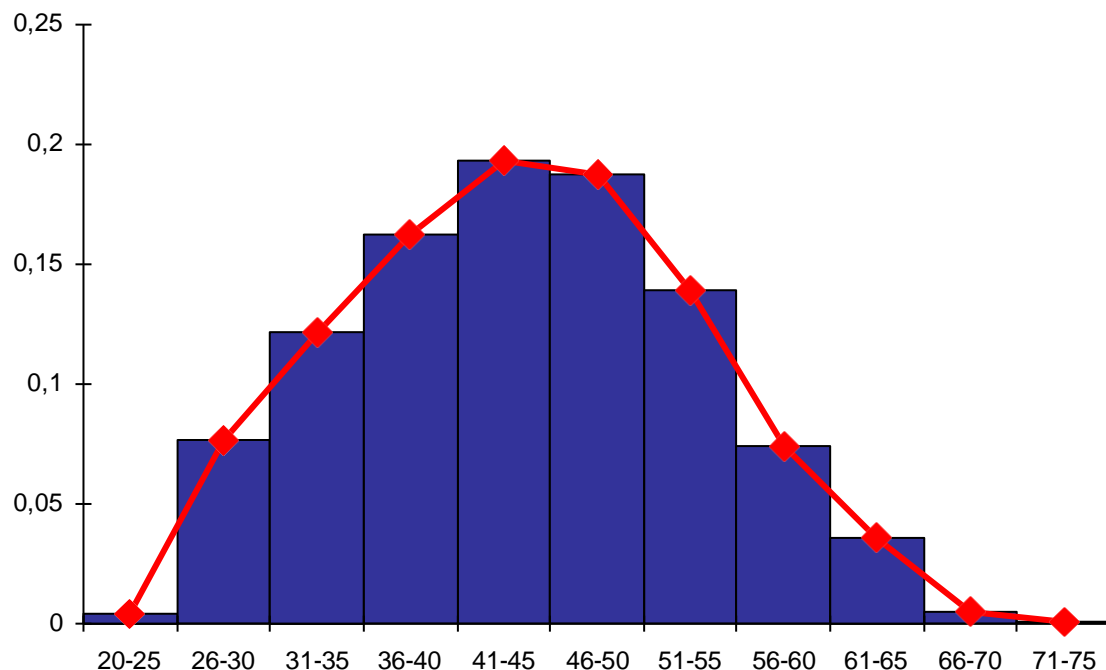
Questa è la rappresentazione grafica della distribuzione di una variabile continua e la funzione che ne è l'espressione matematica è la funzione densità di probabilità

Supponiamo di valutare la distribuzione dei soggetti di una popolazione per età

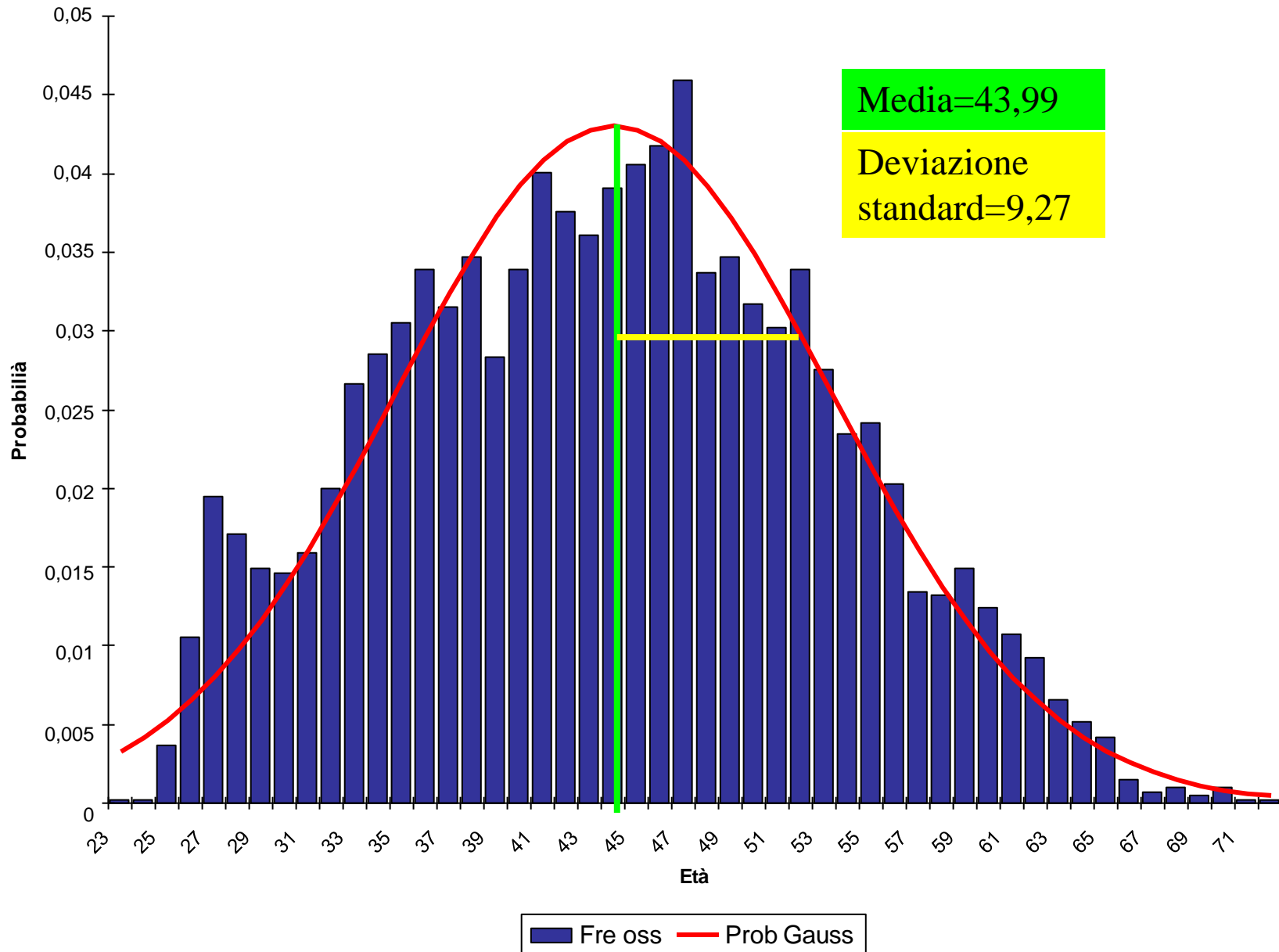
Classe di età	Frequenza assoluta	Frequenza relativa	Frequenza cumulativa	Frequenza cumulativa relativa
20-30	331	8,08%	331	8,08%
31-40	1163	28,38%	1494	36,46%
41-50	1561	38,09%	3055	74,55%
51-60	875	21,35%	3930	95,90%
61-70	166	4,05%	4096	99,95%
71-80	2	0,05%	4098	100,00%
	4098	100,00%		



Possiamo riprodurre la distribuzione di frequenza utilizzando un istogramma. Possiamo inoltre unire i punti medi di ciascuna classe con una linea spezzata per rappresentare il fenomeno con un poligono di frequenza. Se rendiamo l'intervallo di classe progressivamente più piccolo.....



...l'ampiezza può ridursi al punto che il poligono di frequenza possa essere approssimato ad una curva continua



La curva di distribuzione di probabilità di una variabile continua che presenta un andamento “a campana” prende il nome di curva normale o gaussiana. La sua espressione matematica è

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right]$$

$$-\infty \leq x \leq +\infty$$

con

exp = funzione esponenziale

$$\pi = 3,14$$

e i due parametri della distribuzione:

$\mu$  = media

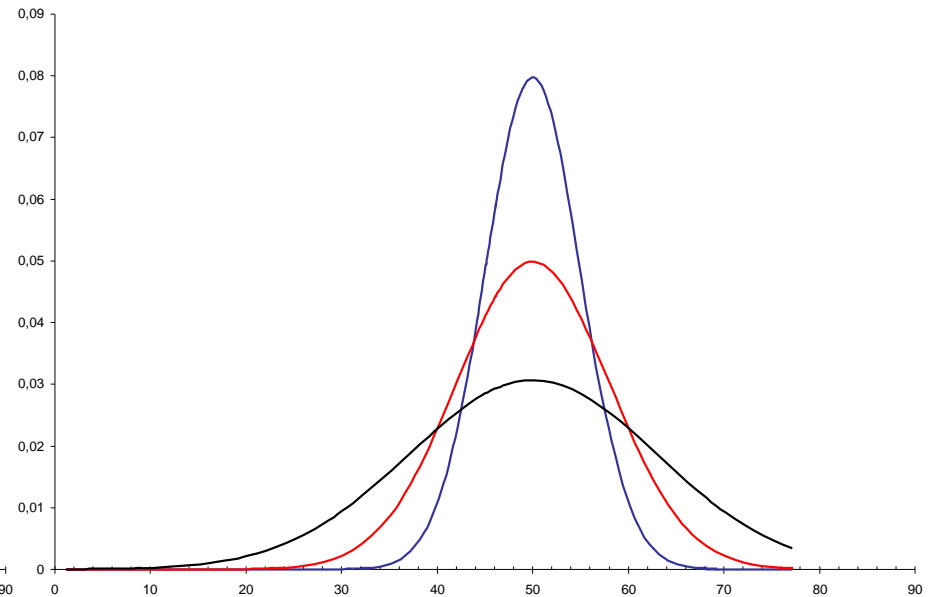
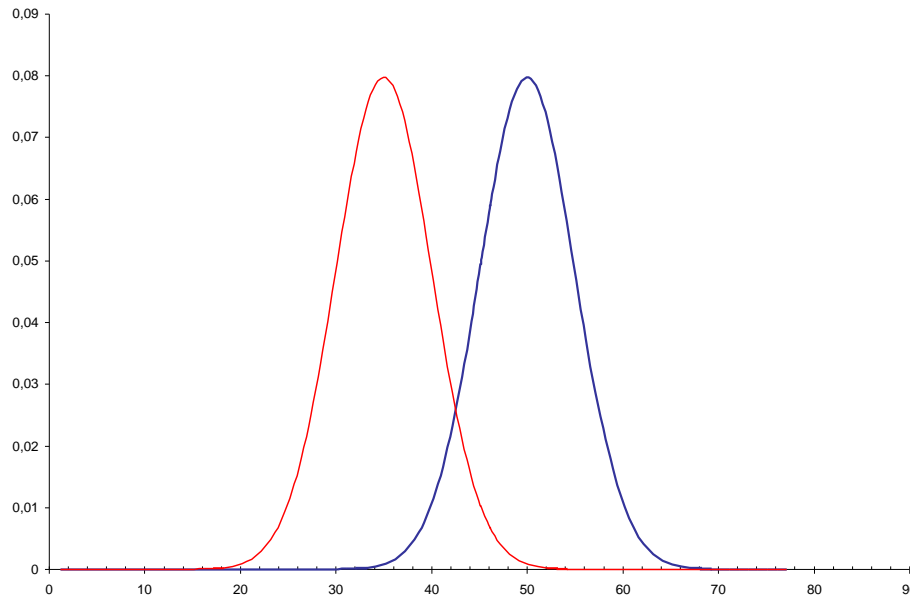
$\sigma$  = deviazione standard



la distribuzione di Gauss è completamente definita dai valori di  $\mu$  e  $\sigma$ :

differenti valori di  $\mu$  spostano la posizione della curva lungo l'asse delle ascisse

differenti valori di  $\sigma$  modificano l'altezza della curva.



## La distribuzione di Gauss ha alcune caratteristiche tipiche:

- è simmetrica intorno alla sua media
- la media, la mediana e la moda coincidono
- l'area sotto la curva è uguale ad 1 (100%)
- l'area sotto la curva compresa nell'intervallo
  - $\mu - \sigma$  ed  $\mu + \sigma$  è pari al 68% dell'area totale
  - $\mu - 2\sigma$  e  $\mu + 2\sigma$  è pari al 95% del totale
  - $\mu - 3\sigma$  ed  $\mu + 3\sigma$  è pari al 99,7% del totale

Esistono degli indici per misurare la normalità della curva di Gauss:

asimmetria:

asimmetria = 0 curva normale

asimmetria < 0 coda sinistra più lunga

asimmetria > 0 coda destra più lunga

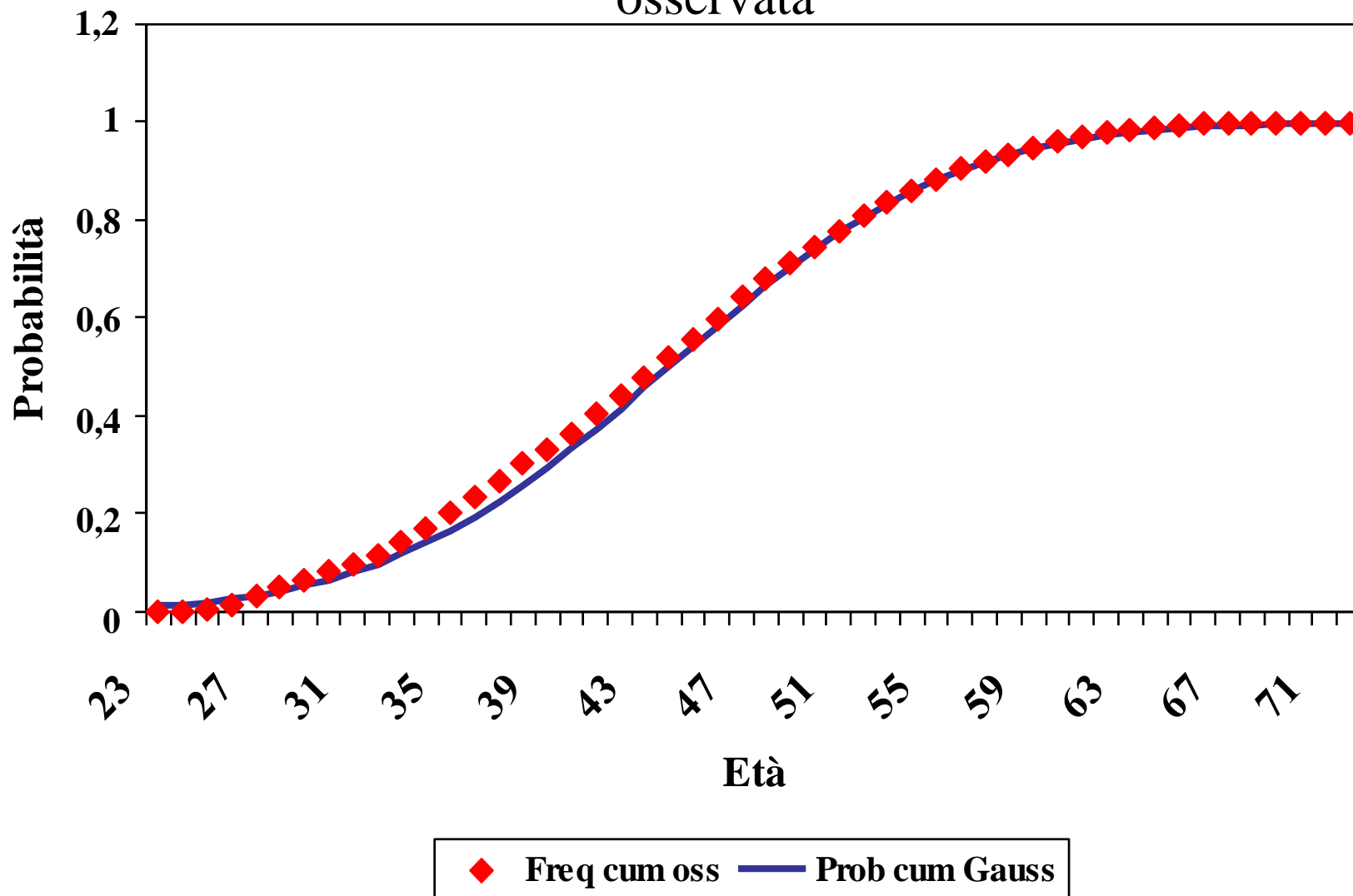
curtosi:

curtosi = 3 curva normale

curtosi < 3 code leggere, distribuzione appuntita (ipernormale o leptocurtica)

curtosi > 3 code pesanti, distribuzione piatta (iponormale o platicurtica).

# Distribuzione di Gauss cumulativa teorica e distribuzione cumulativa osservata



L'ultima caratteristica enunciata ci dice che esiste una famiglia di distribuzioni di gaussiane ed ogni membro è distinto in base ai valori di  $\mu$  e  $\sigma$ .

Tra le varie curve di Gauss la più importante è la distribuzione di Gauss standard che ha

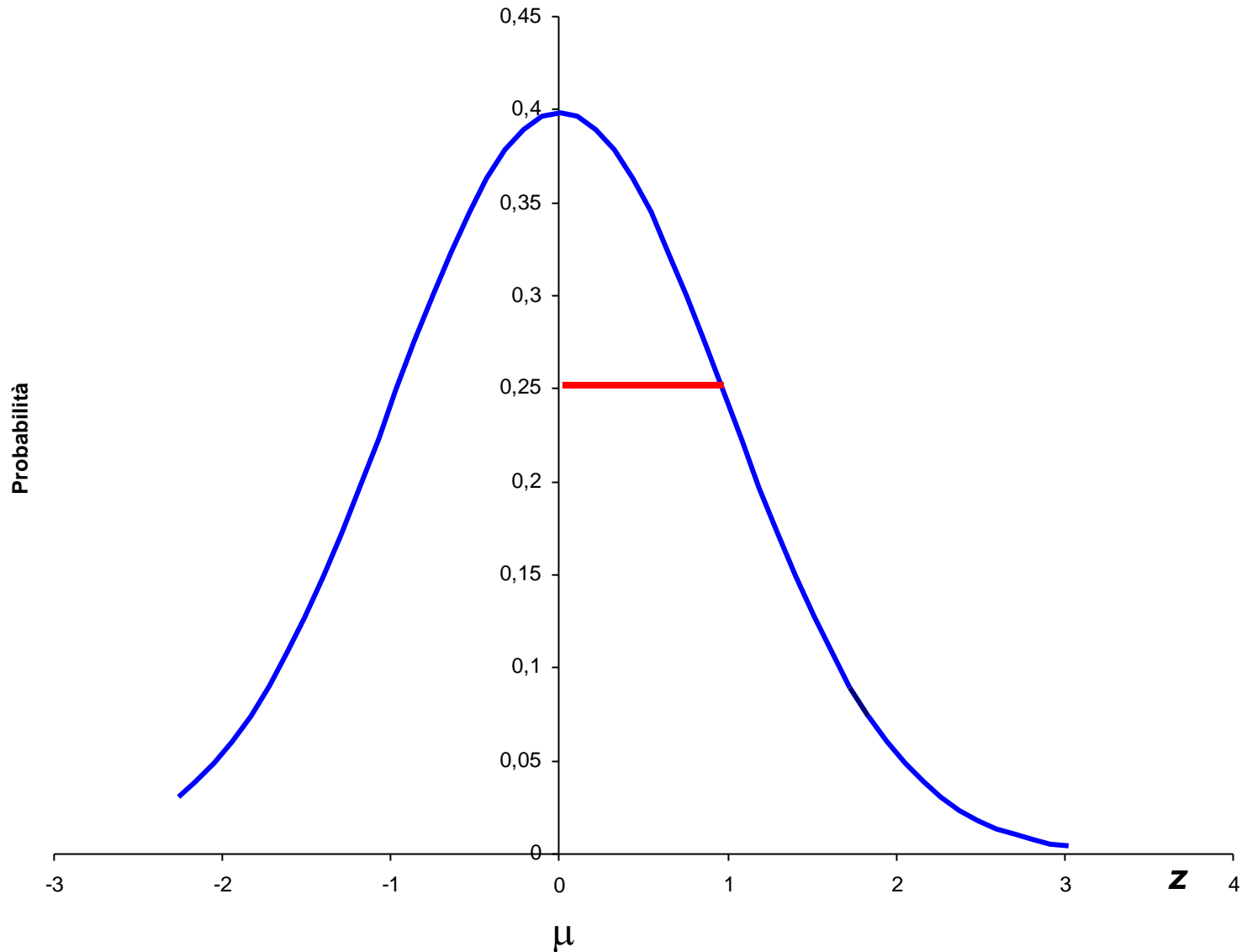
media = 0

deviazione standard = 1

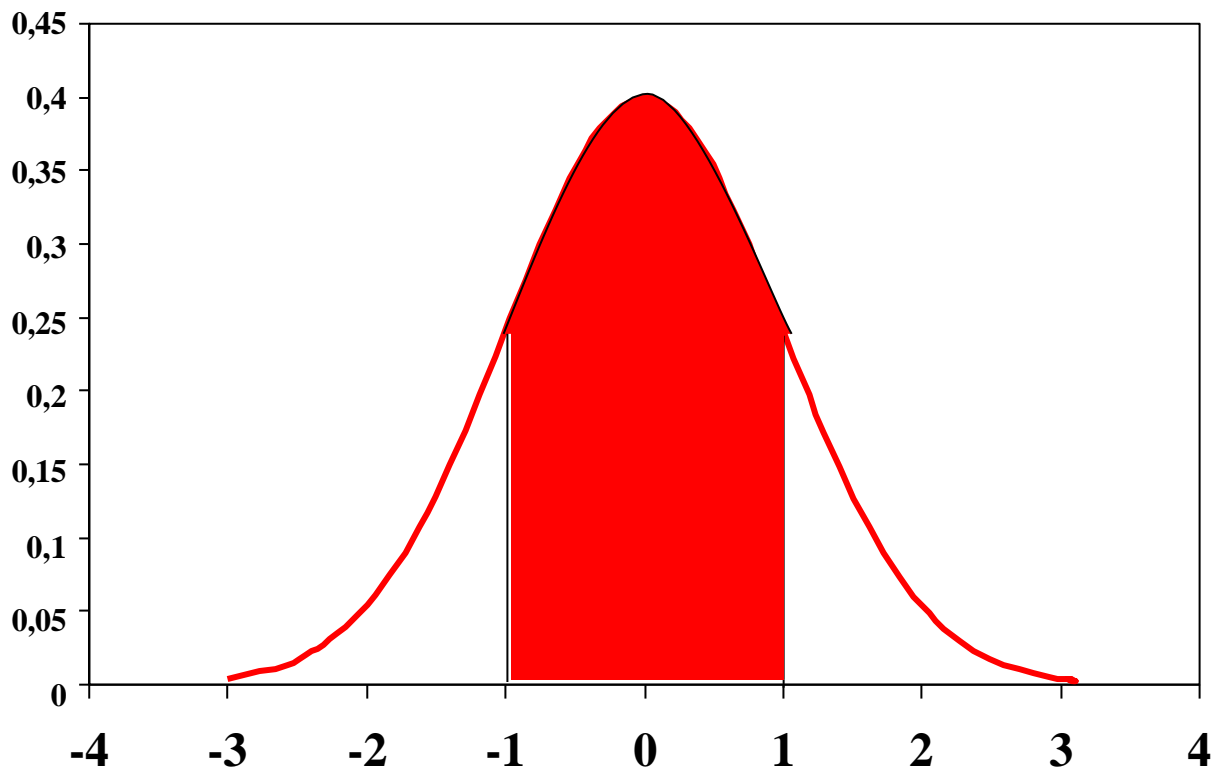
L'espressione matematica della distribuzione di Gauss Standard è:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}$$

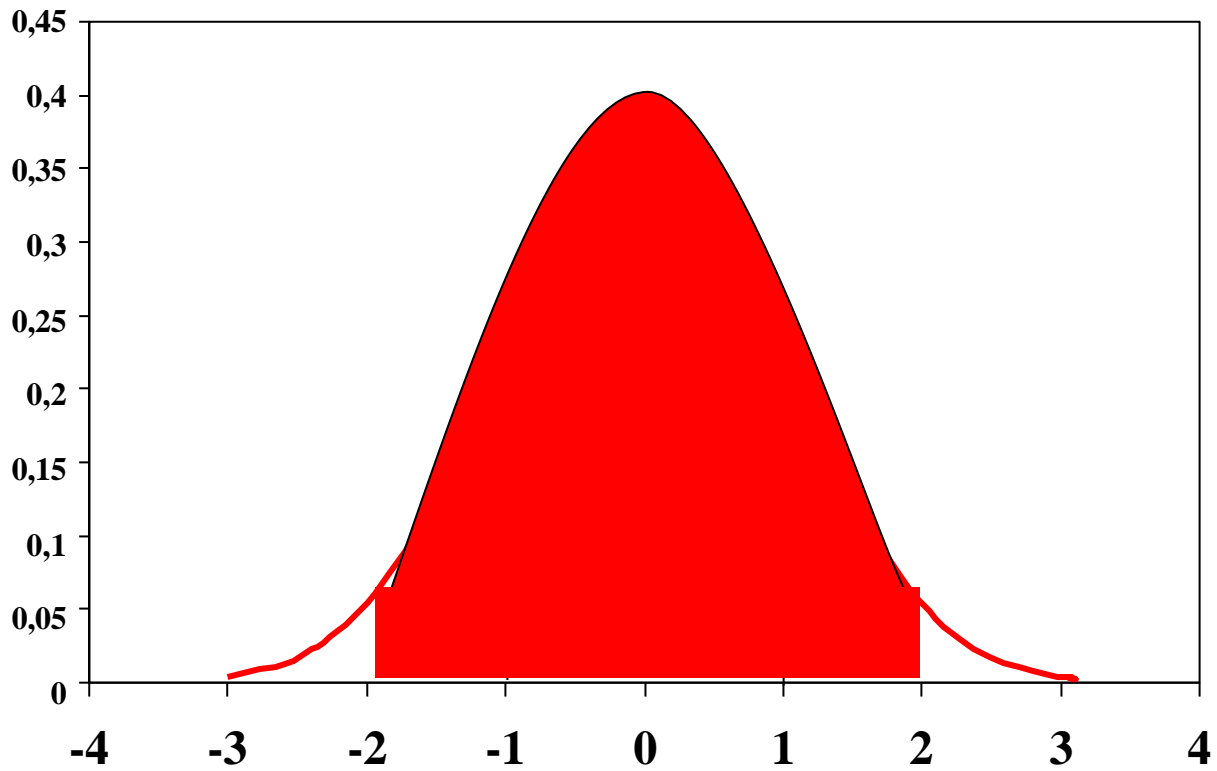
Distribuzione di gauss standard: media=0 e deviazione standard=1







Area in rosso = 68%



Area in rosso = 95%

E' importante sottolineare che ogni variabile  $X$  può essere standardizzata mediante la trasformazione

$$z = \frac{x - \mu}{\sigma}$$

Indicando la distribuzione di Gauss con il simbolo  $N(\mu, \sigma)$  è possibile fare le seguenti approssimazioni:

$$B(n, p) \approx N(\mu, \sigma) \text{ se } n \rightarrow \infty \quad p = 0,5$$

$$P(\lambda) \approx N(\mu, \sigma) \text{ se } n \rightarrow \infty$$