# R e Python per la data science

LUISS BUSINESS SCHOOL

EFMD EQUIS ACCREDITED

# Agenda

- Strutture dati
- Analisi descrittiva esplorativa
  - Sintesi analitiche univariate
  - Sintesi analitiche multivariate
- Data manipulation
- Data visualization

# Analisi statistiche e programmazione

- Efficienza
  - Le interfacce punta e clicca non sono efficienti in termini di tempo
  - Automatizzare significa velocizzare le operazioni

- Riproducibilità
  - Crescente necessità di fornire dati, materiali ed analisi insieme ad i risultati
  - Assicura la possibilità di controllare i risultati e le procedure
  - Rende possibile effettuare analisi in produzione

# R vs Python

- Data analytics, statistica
- Usato da statistici e dalla ricerca
- 12000 package on CRAN
- Semplice comunicazione (visualization, reporting and dashboard)

- Deployment and production
- Usato da programmatori e sviluppatori
- Integrazione con diversi sistemi operativi
- Algoritmi complessi e struttura ad oggetti

# Download

Linguaggio

IDE

(integrated development environment)


Motore

Cruscotto

# Python

## Linguaggio

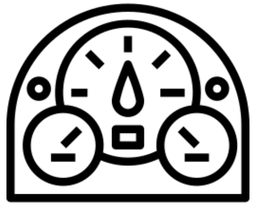
Motore

## IDE

(integrated development environment)





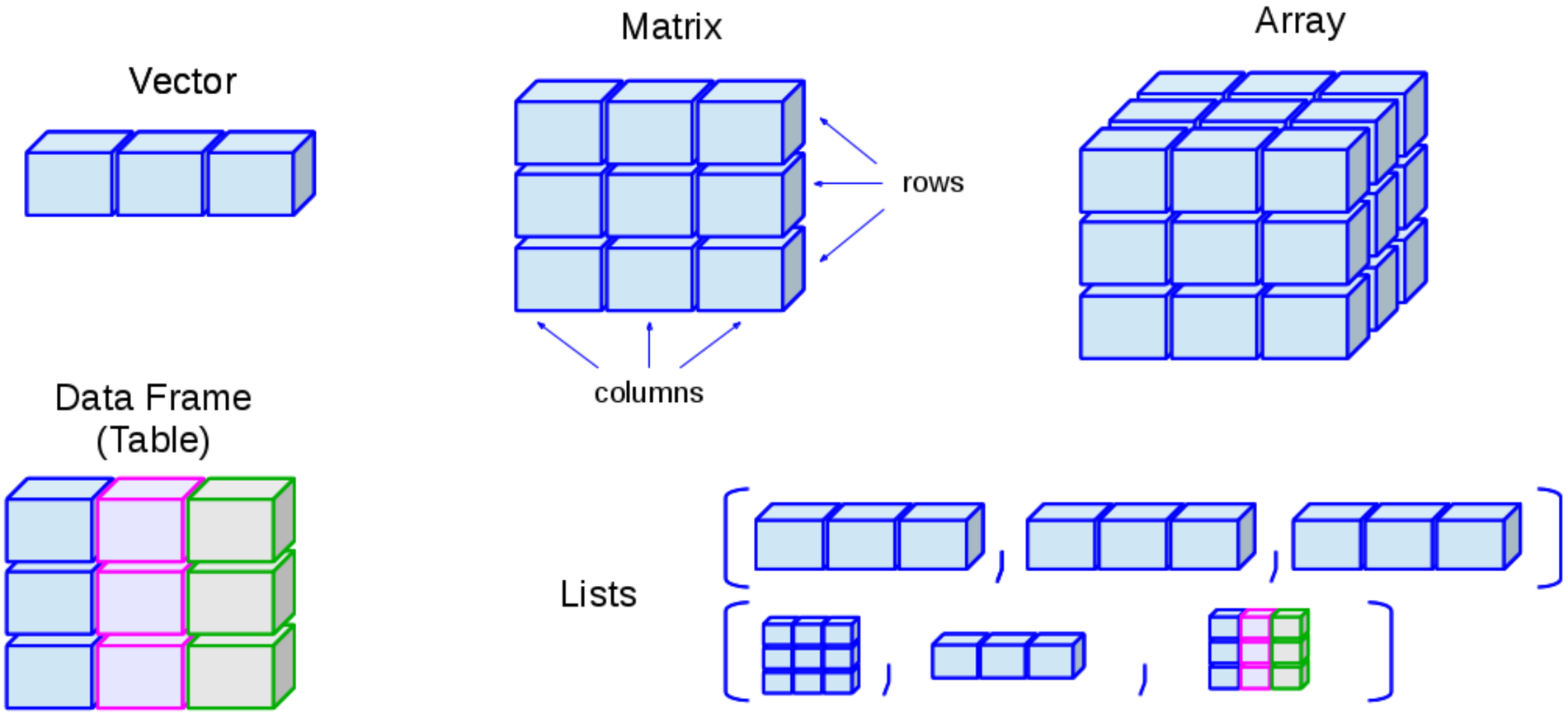


Cruscotto

# R & Python

- Open source
  - Gratuito e liberamente utilizzabile
- Strumenti avanzati
  - Pacchetti e librerie per ogni tipo di analisi
- Documentazione e comunità
  - Nessun supporto cliente a pagamento ma comunità!

# Strutture di dati in R



# Analisi descrittiva/esplorativa

## Sintesi analitiche

### Sintesi univariate

Tendenza centrale
Quantili
Dispersione
Distribuzione

### Sintesi bivariate

Covarianza
Correlazione

# Sintesi analitiche univariate

## Tendenza centrale

### Media

Somma di tutti i valori di un campione o popolazione diviso per il numero di unità

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

### Mediana

Valore che occupa la posizione centrale in un insieme ordinato di dati

### Moda

Valore più frequente di una distribuzione, la modalità più ricorrente della variabile

# Sintesi analitiche univariate

## Quantili

I quantili sono una famiglia di misure, a cui appartiene anche la mediana, che si distinguono a seconda del numero di parti uguali in cui suddividono una distribuzione

Quartili

Decili

Percentili

1Q = 25%
2Q = 50% (MEDIANA)
3Q = 75%

# Sintesi analitiche univariate

## Dispersione

Gli indici di dispersione servono a valutare la diversità esistente tra le osservazioni.

**Range**

**min**
1Q = 25%
2Q = 50% (MEDIANA)
3Q = 75%
**max**

**Scarto interquartilico**

# Sintesi analitiche univariate

## Dispersione

### Varianza

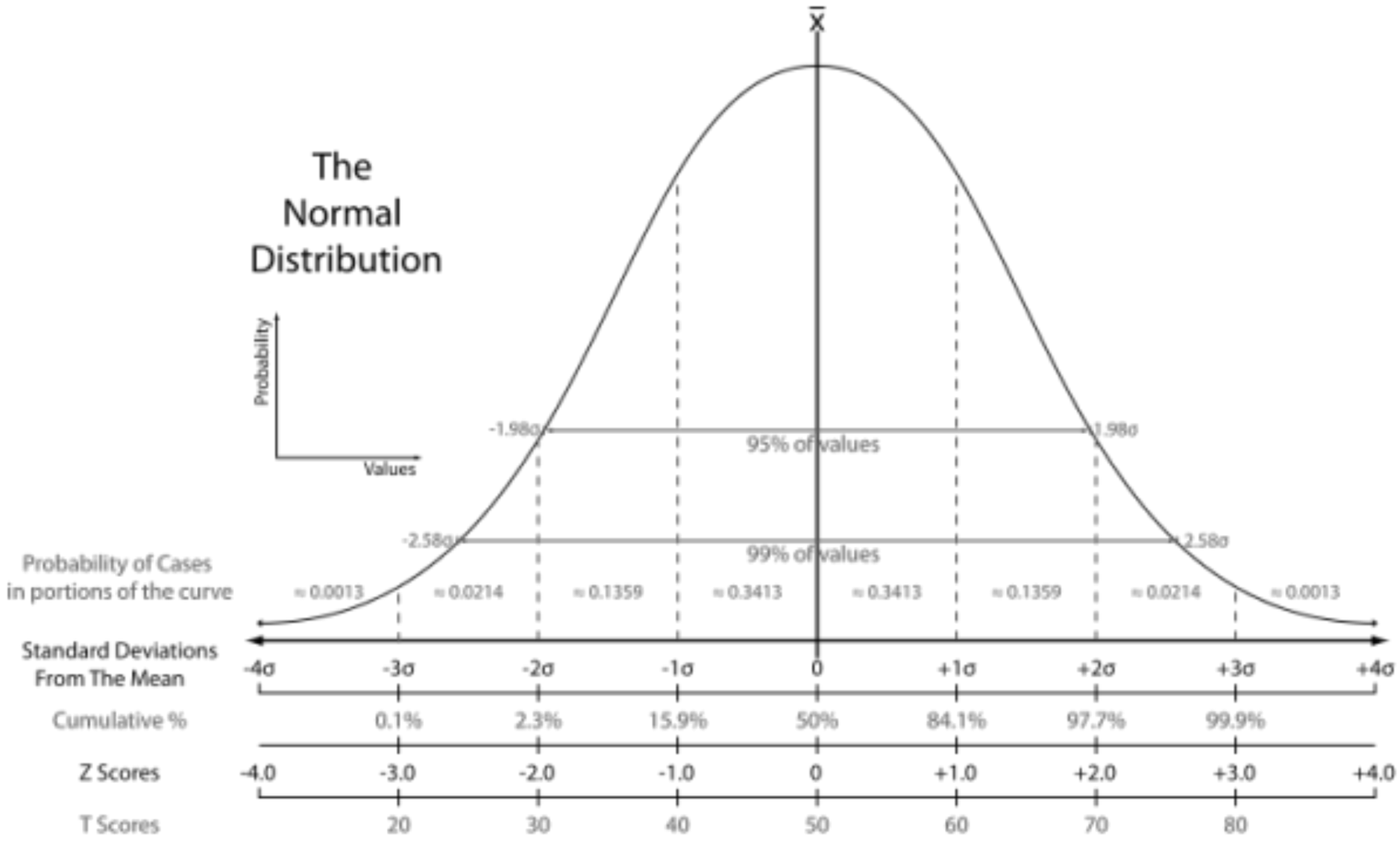
$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

### Deviazione standard

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$$

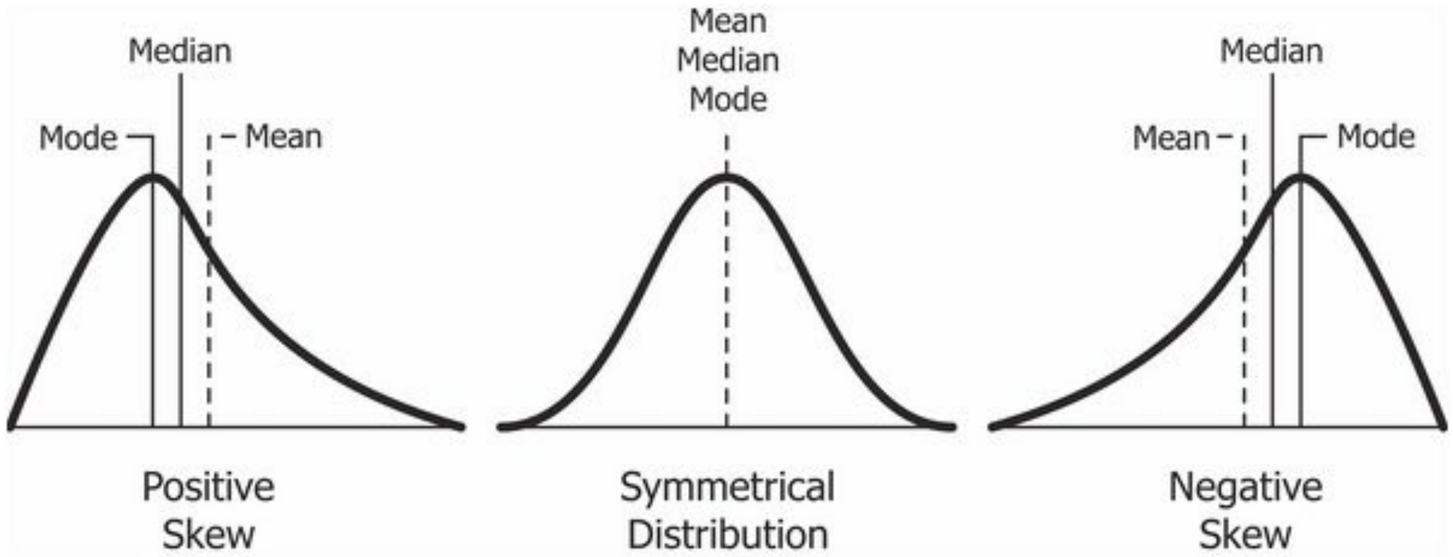
# Sintesi analitiche univariate

## Distribuzione



# Sintesi analitiche univariate

## Distribuzione

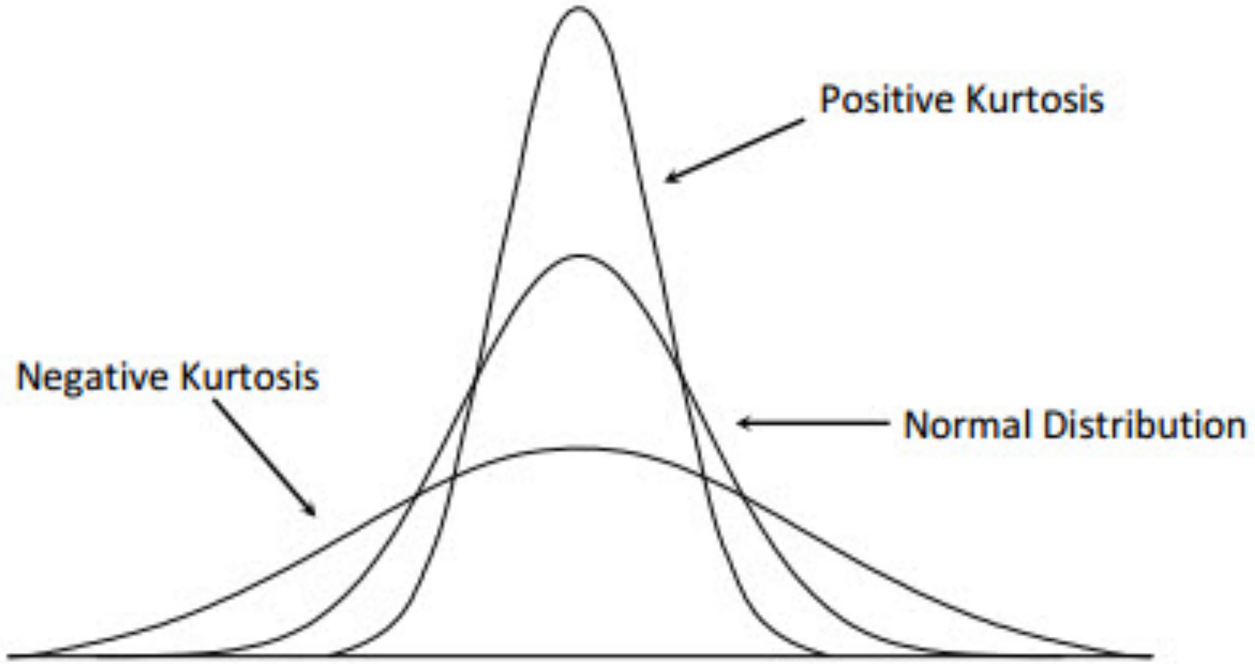



## Asimmetria

Una distribuzione di dati si dice simmetrica se esiste un valore che divide la distribuzione stessa in due parti, con gli elementi di ciascuna parte simmetrici dei corrispondenti elementi dell'altra parte. Se non esiste tale valore, la distribuzione è asimmetrica.

# Sintesi analitiche univariate

## Distribuzione




## Curtosi

L'indice *K* di curtosi misura il maggiore o minore appuntimento di una distribuzione di dati, rispetto alla distribuzione normale. Di conseguenza esso indica il maggiore o minore peso dei valori posti agli estremi della distribuzione (code), rispetto a quelli della parte centrale.

# Sintesi analitiche bivariate

## Covarianza

Indice che consente di verificare se fra due variabili statistiche esiste un legame lineare

$$Cov(X,Y) = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Positiva, Negativa, Nulla

## Correlazione

Grado di intensità del legame lineare tra coppie di variabili

$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{(VarX)(VarY)}}$$

Coefficiente di Pearson
Varia da -1 a 1