# ELEMENTI DI STATISTICA

## CORSO ENGIM IFTS
## MODULO STRUMENTI

top ix torino piemonte internet exchange

# STEFANIA DELPRETE

**astrastefania**


# AGENDA

## i. Prerequisiti e introduzione alla Statistica

## ii. Statistica descrittiva

## iii. Statistica inferenziale

# PREREQUISITI

# MATEMATICI

# PREREQUISITI FONDAMENTALI

- Sommatorie, produttorie
- Fattoriale, modulo
- Logaritmi, esponenziali
- Piano cartesiano, studio di funzioni
- Curiosità e pazienza

# INTRODUZIONE
# ALLA STATISTICA

# STATISTICA

**La Statistica è la scienza che studia in modo quantitativo e qualitativo un insieme di dati e osservazioni al fine di descrivere e prevedere un fenomeno.**

# STATISTICA IN DATA SCIENCE

- Acquisizione dei dati
- Statistica descrittiva
- Esplorazione e analisi dei dati, EDA
- Stime su campioni
- Statistica inferenziale e studio di ipotesi su campione
- Applicazione di modelli di Machine Learning

# STATISTICA DESCRITTIVA

# STATISTICA DESCRITTIVA

**Studio dell'acquisizione, classificazione, e sintesi dei dati tramite rappresentazioni grafiche e indici.**

**I risultati di tale analisi sono pressoché certi a meno di errori di misurazione.**

# INDICI DI POSIZIONE

- Media aritmetica
- Mediana
- Moda
- Quantili

# MEDIA ARITMETICA

**Media**

$$M_a = \frac{1}{n} \sum_{i=1}^{n} x_i$$

**Media ponderata con pesi $f_i$**

$$M_{a,pond} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$

# MEDIANA

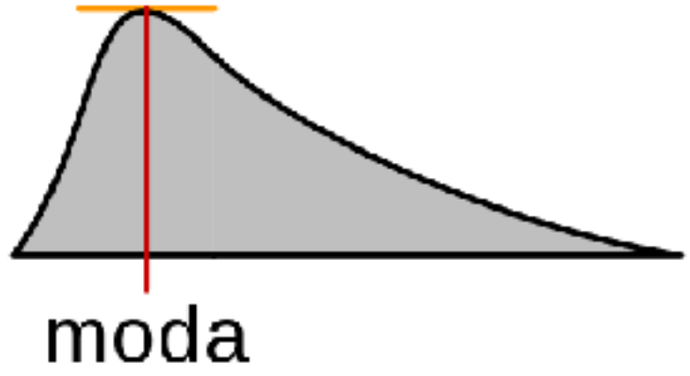
**Il valore mediano è il valore assunto dalle unità statistiche che si trovano nel mezzo della distribuzione, lasciando alla sua destra e sinistra il 50% della distribuzione.**

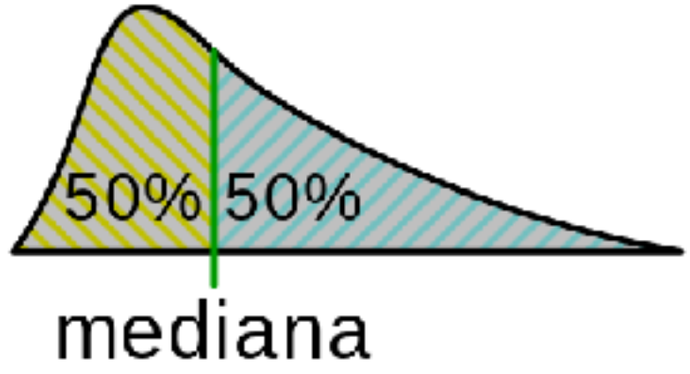
MODA

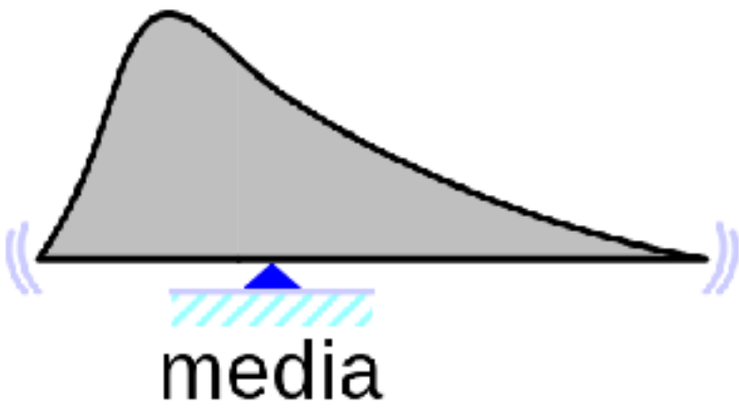
# La moda è il valore che compare più frequentemente fra i dati della distribuzione.

# TENDENZA CENTRALE

**Media, mediana e moda sono tre misure statistiche che analizzano la tendenza centrale.**

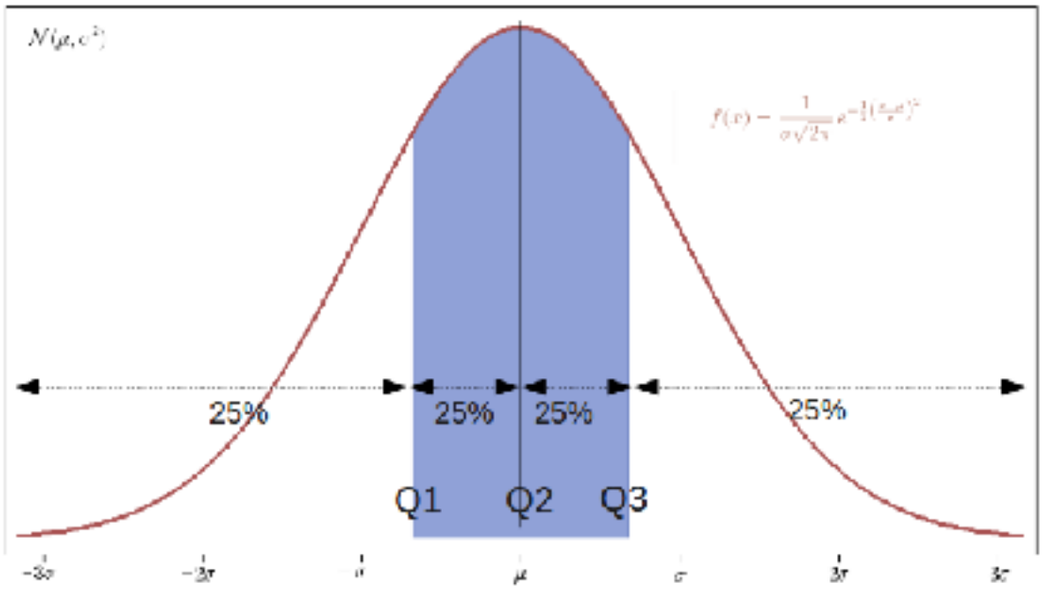


# QUANTILI

**Quantili sono valori che dividono la popolazione in parti uguali.**



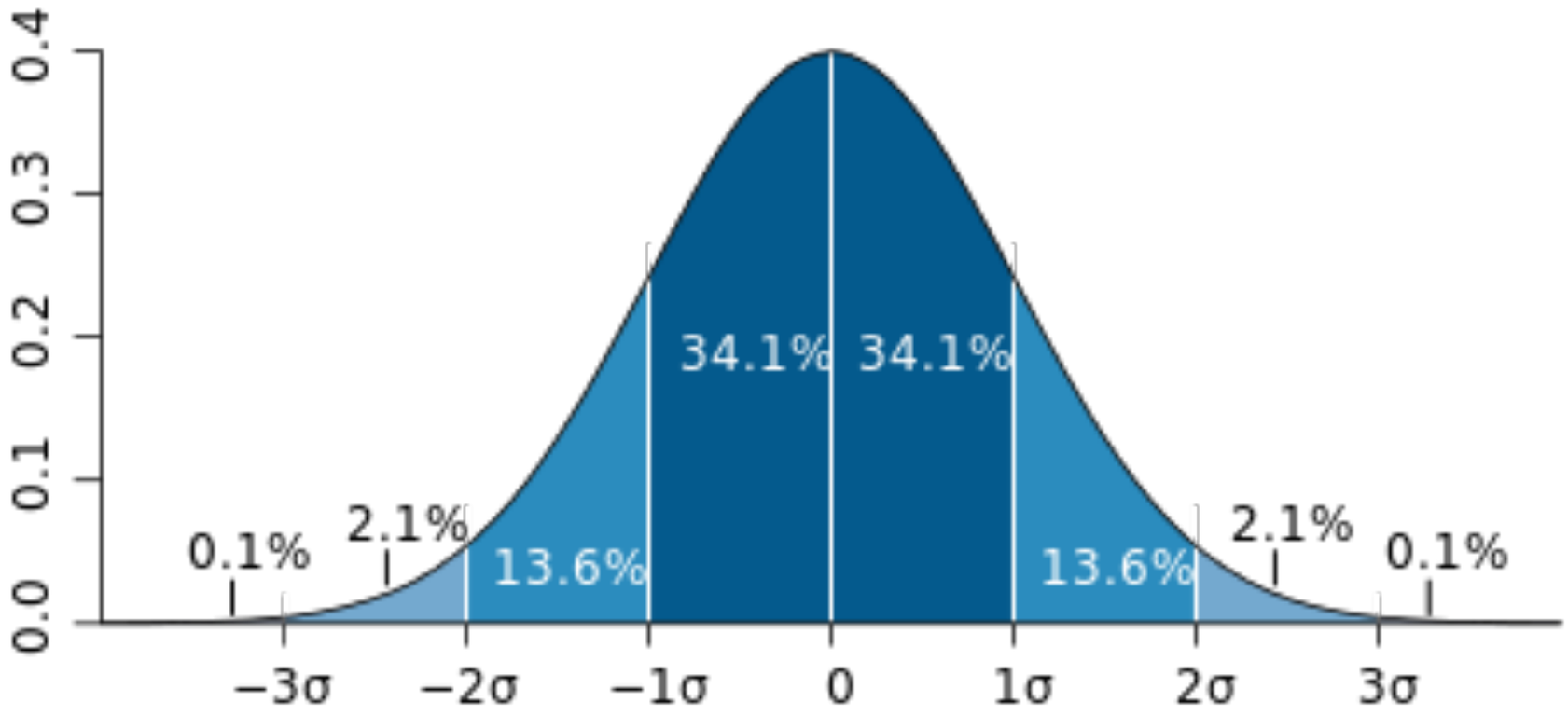

- **La mediana è un quantile di ordine 1/2**
- **I quartili, ordine m/4, dividono la popolazione in 4 parti uguali**

# DEVIAZIONE STANDARD

**Lo scarto quadratico medio, o deviazione standard, è un indice di dispersione che stima la variabilità di una popolazione**

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}}$$

# DEVIAZIONE STANDARD E QUANTILI



# INDICI DI VARIABILITÀ

**Gli indici di variabilità esprimono l'allontanamento dei dati da un'indice della tendenza centrale.**

**Varianza:**

$$\sigma_X^2 = \frac{\sum_i (x_i - \mu_X)^2}{n}$$

# CORRELAZIONE E CAUSALITÀ

**La correlazione fra due variabili indica una connessione di interdipendenza, in cui una è proporzionale o inversamente proporzionale all'altra.**

**La causalità sussiste quando:**

- **le variabili sono correlate**
- **una variabile precede l'altra**
- **non esiste una terza variabile che provochi cambiamenti nelle variabili considerate**

# INDICE DI CORRELAZIONE

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

## Coeffiente di correlazione di Pearson misura la correlazione lineare fra X e Y

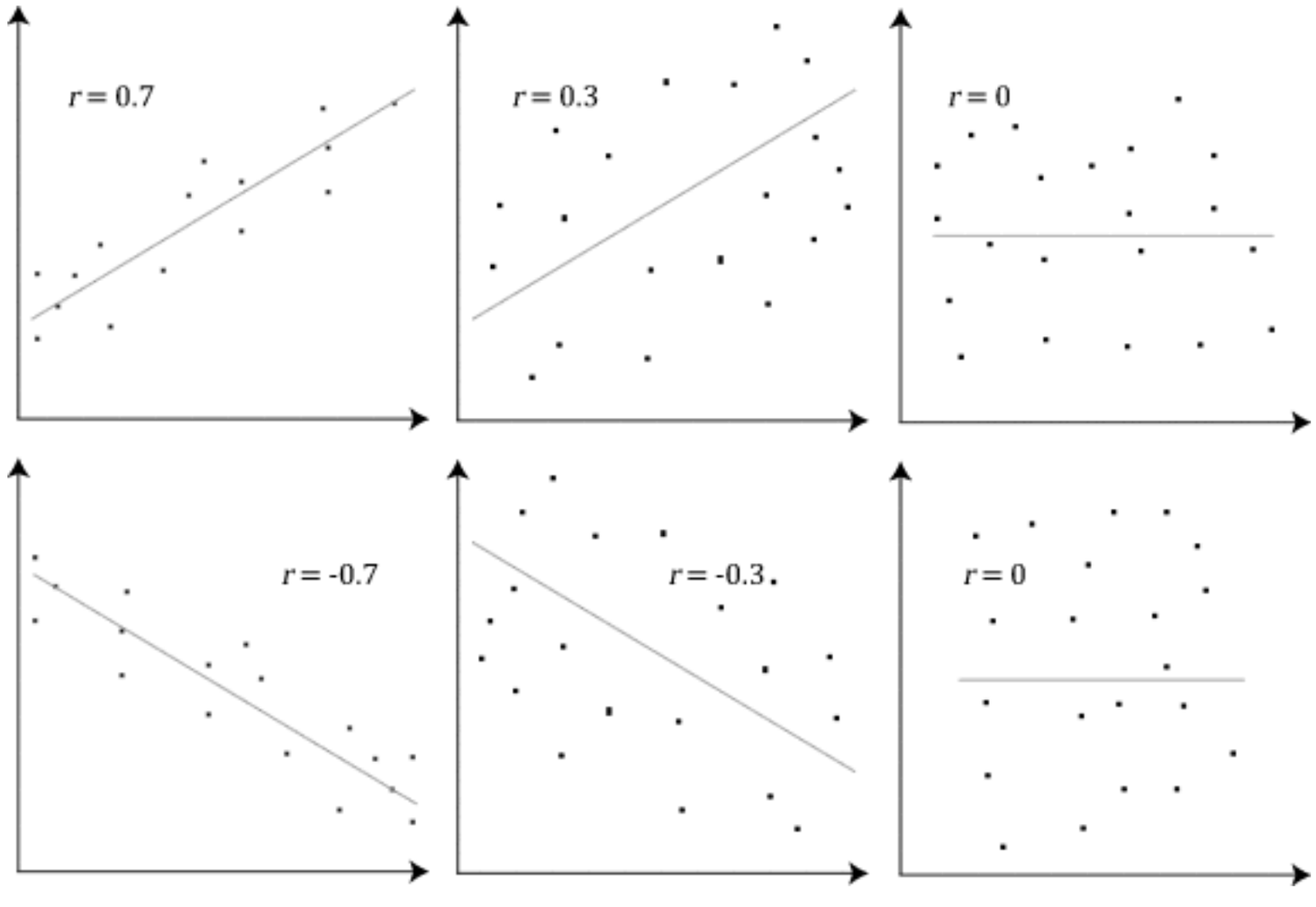
$$\sigma_{X,Y} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)$$

# INDICE DI CORRELAZIONE

/ 1 correlazione lineare positiva

/ 0 nessuna correlazione lineare

/ −1 correlazione lineare negativa

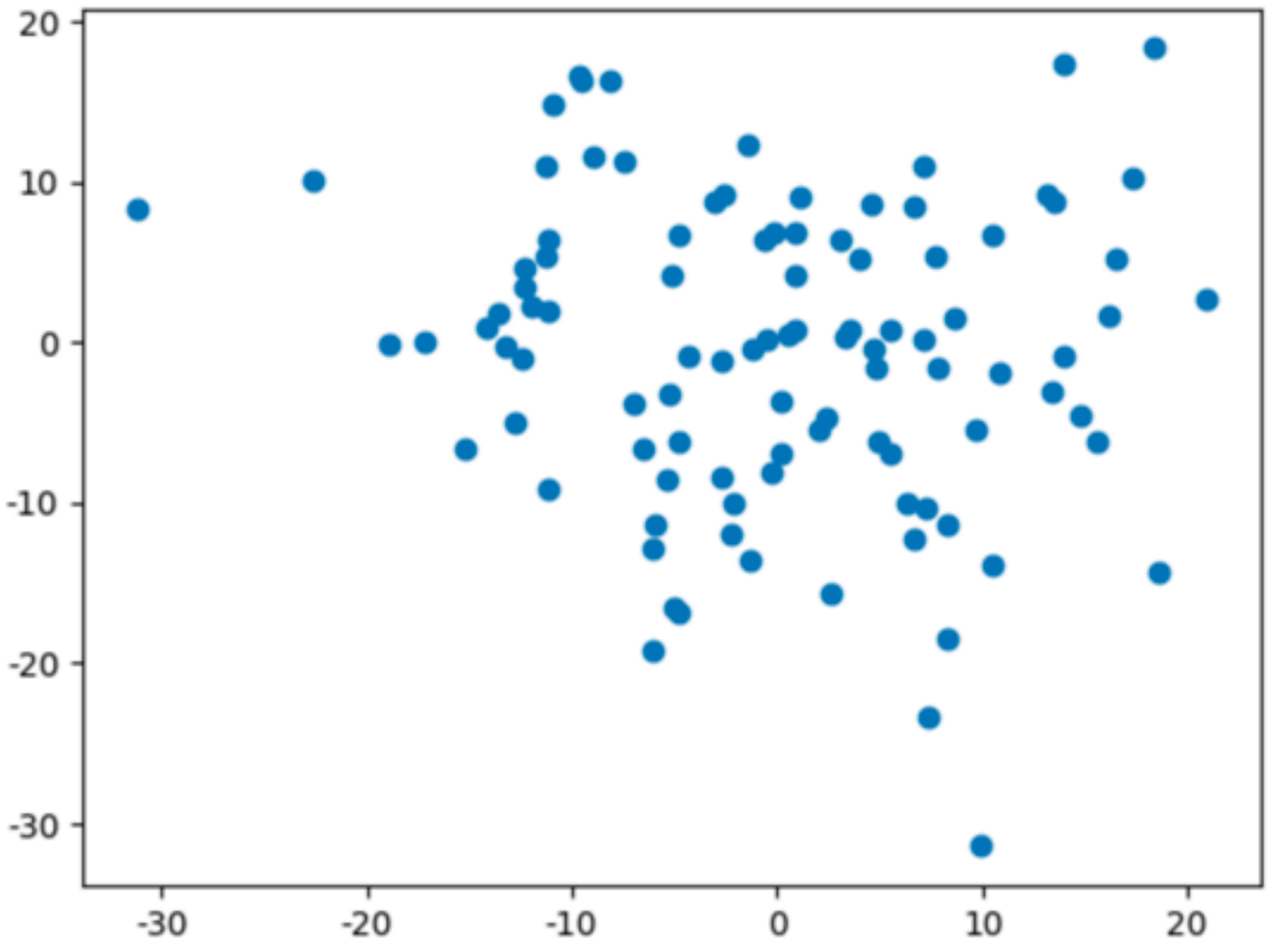


# RAPPRESENTAZIONI GRAFICHE

- Scatterplot
- Boxplot
- Grafico a barre
- Istogramma

# SCATTERPLOT

**Un grafico a dispersione è la rappresentazione dei dati nello spazio.**



# BOXPLOT

**Il diagramma a scatola e baffi è una rappresentazione statistica che sintetizza i dati tramite i valori di:**

- **Mediana**
- **Variazione interquartile, IQR, fra primo e terzo quartile, Q1 e Q3**
- **Minimo e massimo ‘non anomali’**
- **Valori anomali o outliers**

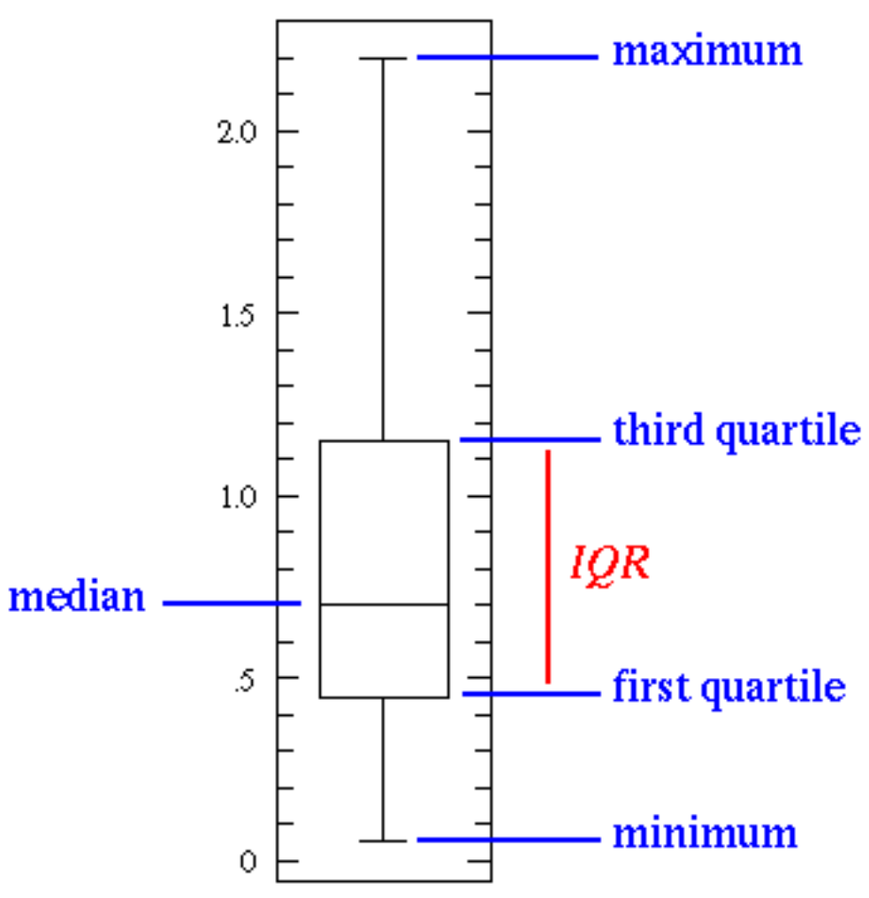
# OUTLIER

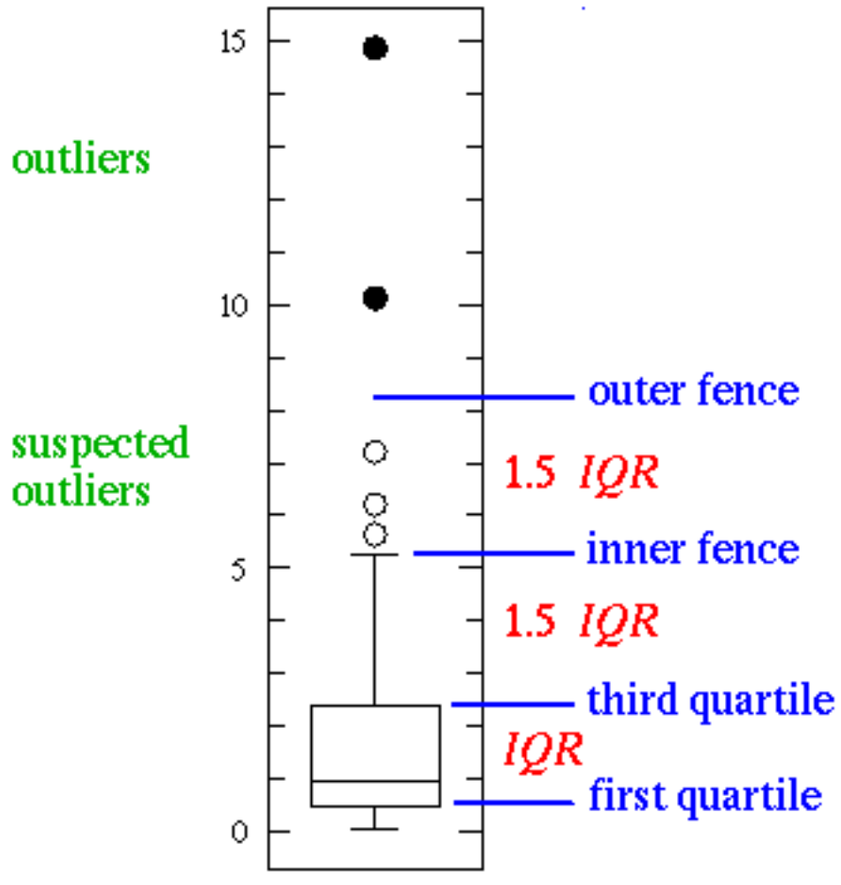
**Outlier è un valore anomalo e aberrante, chiaramente distante dalle altre osservazioni disponibili.**

- **Outliers superiori, valori maggiori di Q3 + 1.5 * IQR**

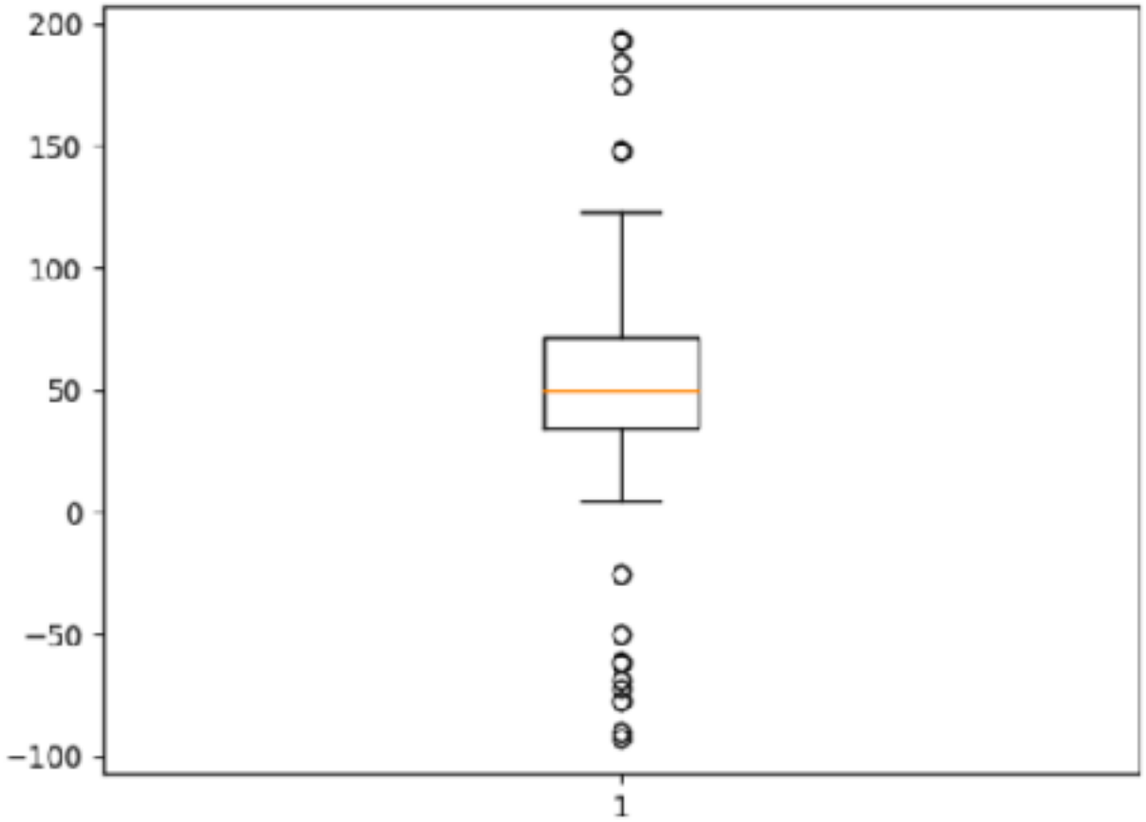
- **Outliers inferiori, valori minori di Q1 - 1.5 * IQR**

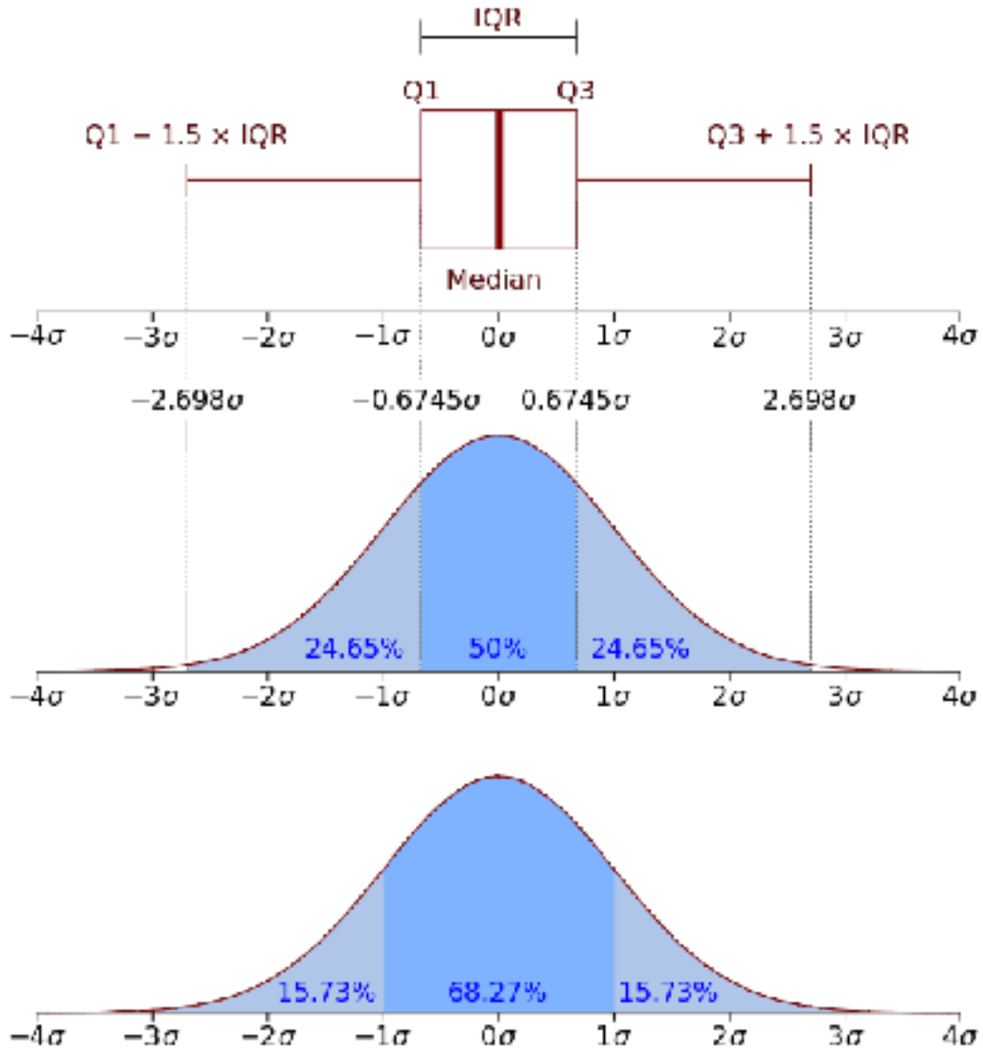
# OUTLIER


maximum
2.0
1.5
1.0
.5
0
third quartile
IQR
median
first quartile
minimum

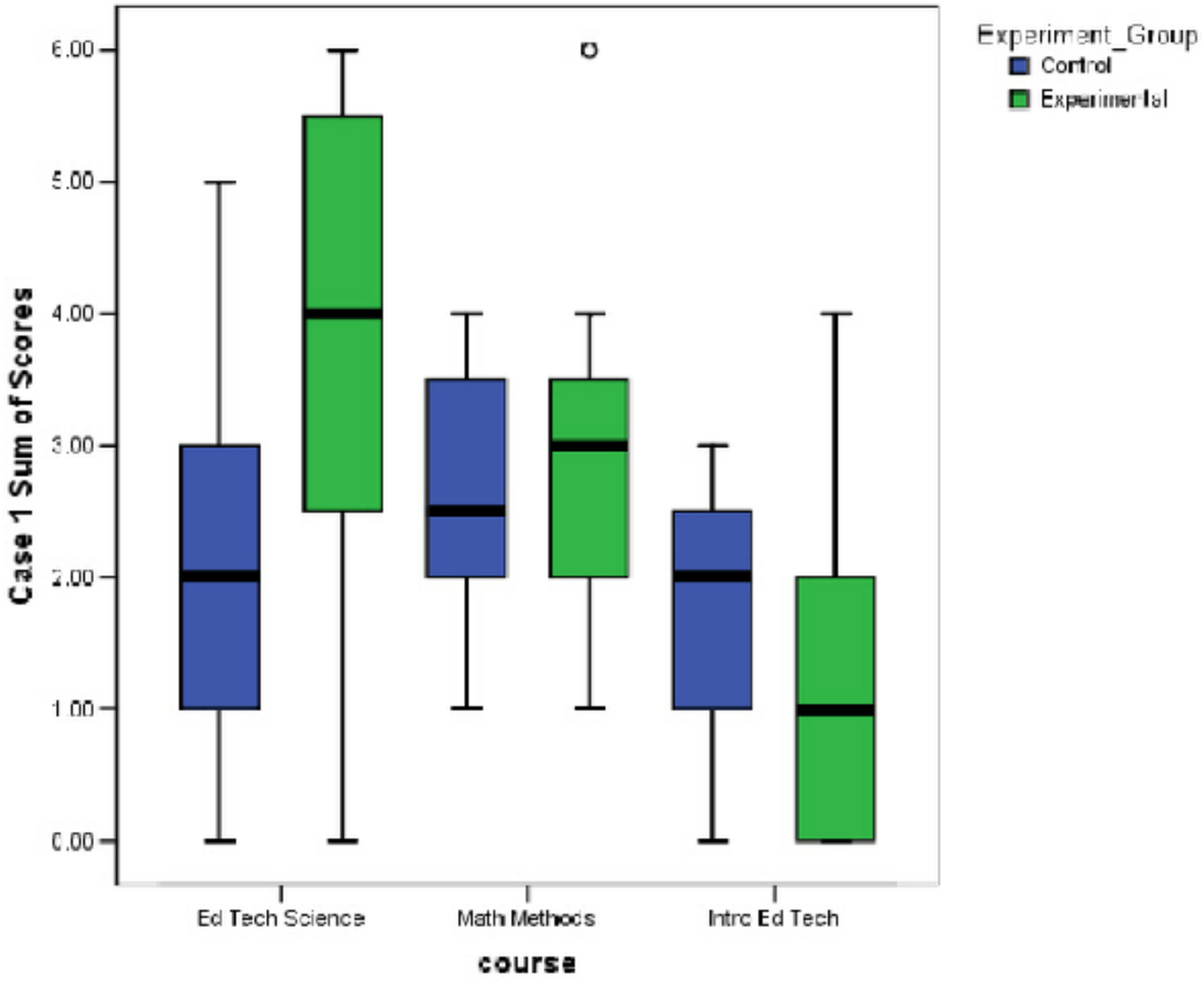

outliers
15
10
5
0
outer fence
1.5 IQR
suspected outliers
inner fence
1.5 IQR
third quartile
IQR
first quartile

# BOXPLOT E DEVIAZIONE STANDARD



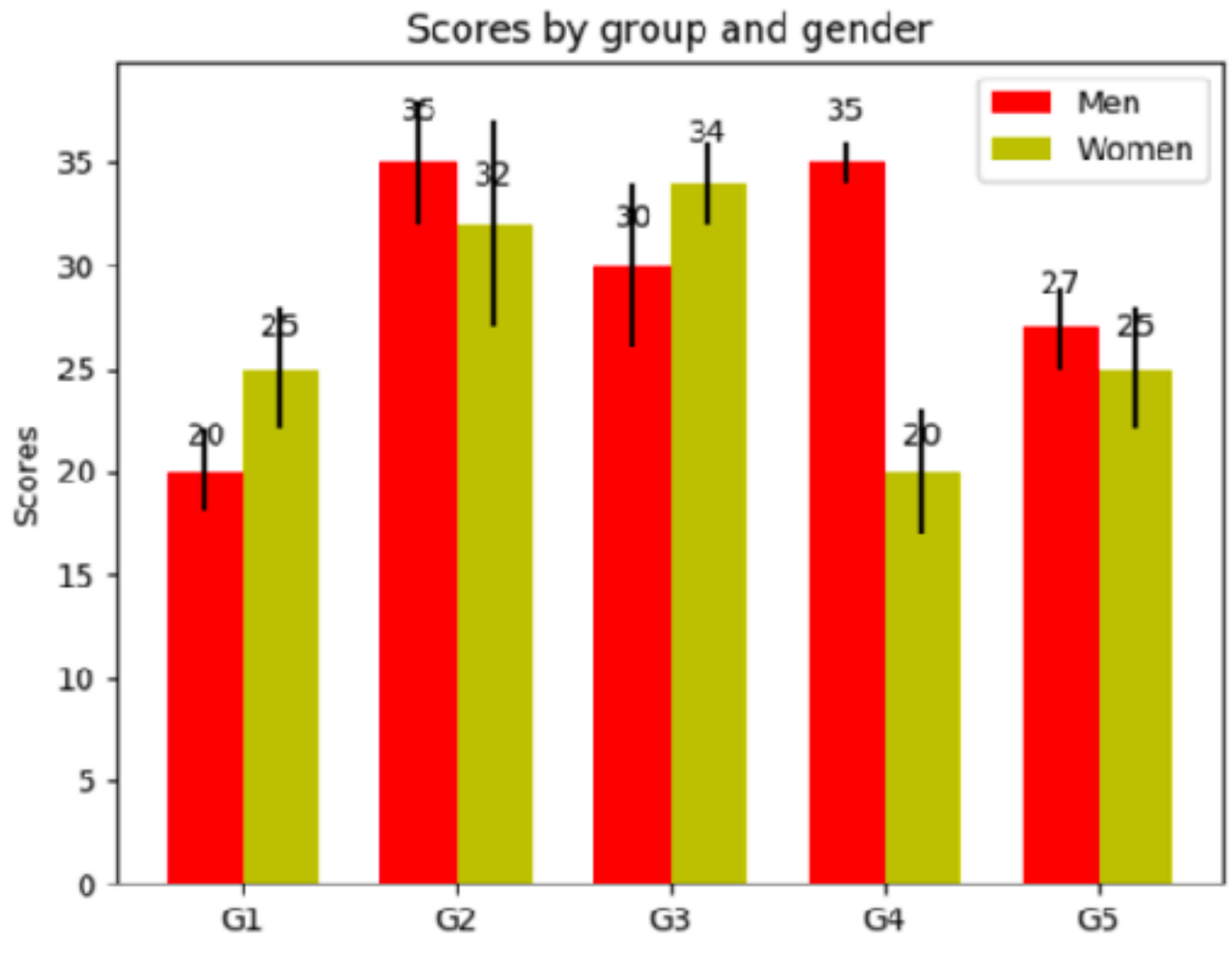



# BOXPLOT ESEMPIO



# BARCHART

**Un grafico a barre è la rappresentazione della frequenza di un determinato valore nei dati della popolazione.**



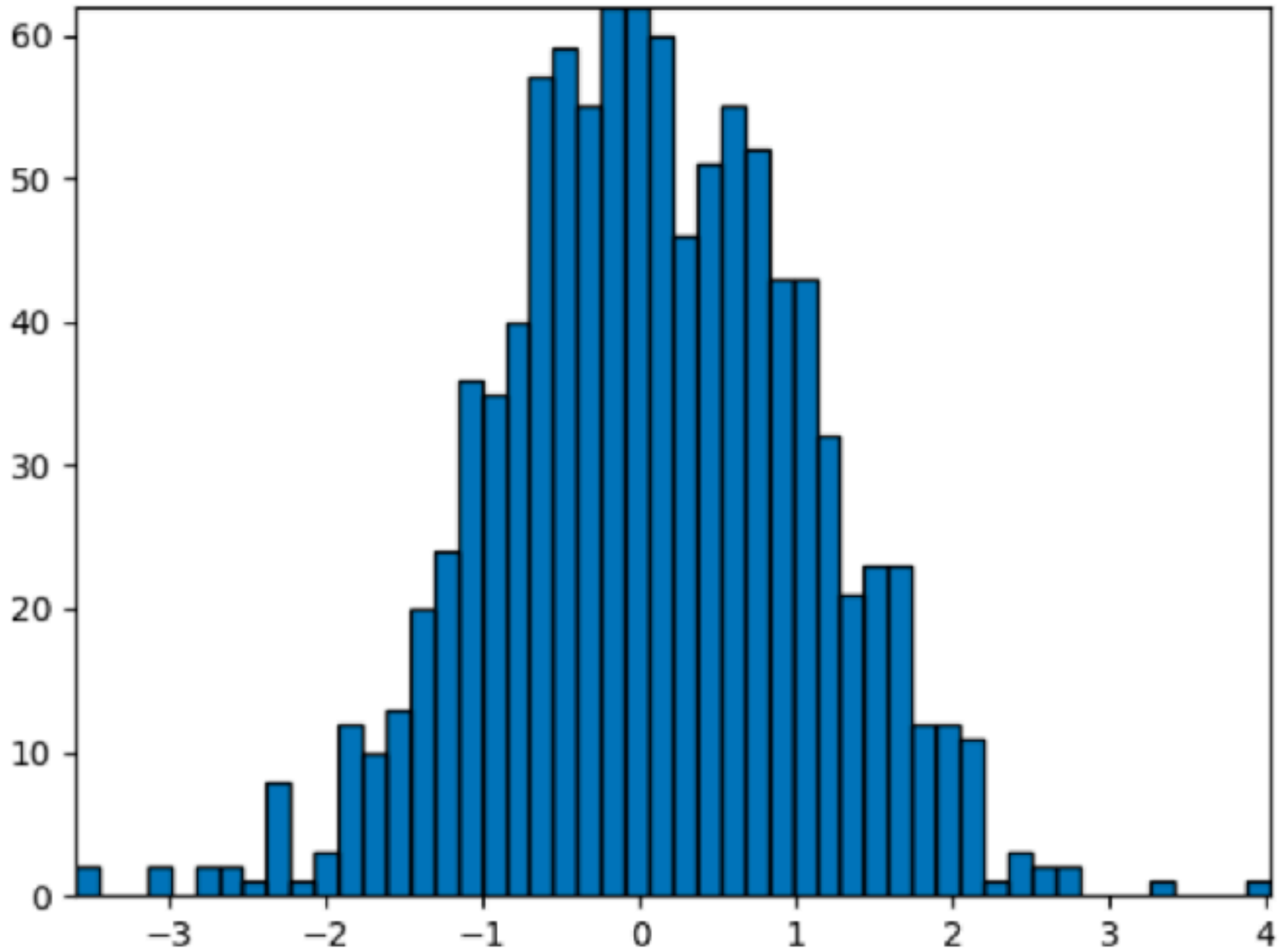
# ISTOGRAMMA

**L'istogramma è la rappresentazione della densità di frequenza di un range di valori di una classe nei dati della popolazione.**

**Tali intervalli di valori possono essere omogenei o disomogenei.**

**La densità di frequenza è il rapporto fra la frequenza assoluta e l'ampiezza di una classe.**

# ISTOGRAMMA ESEMPIO



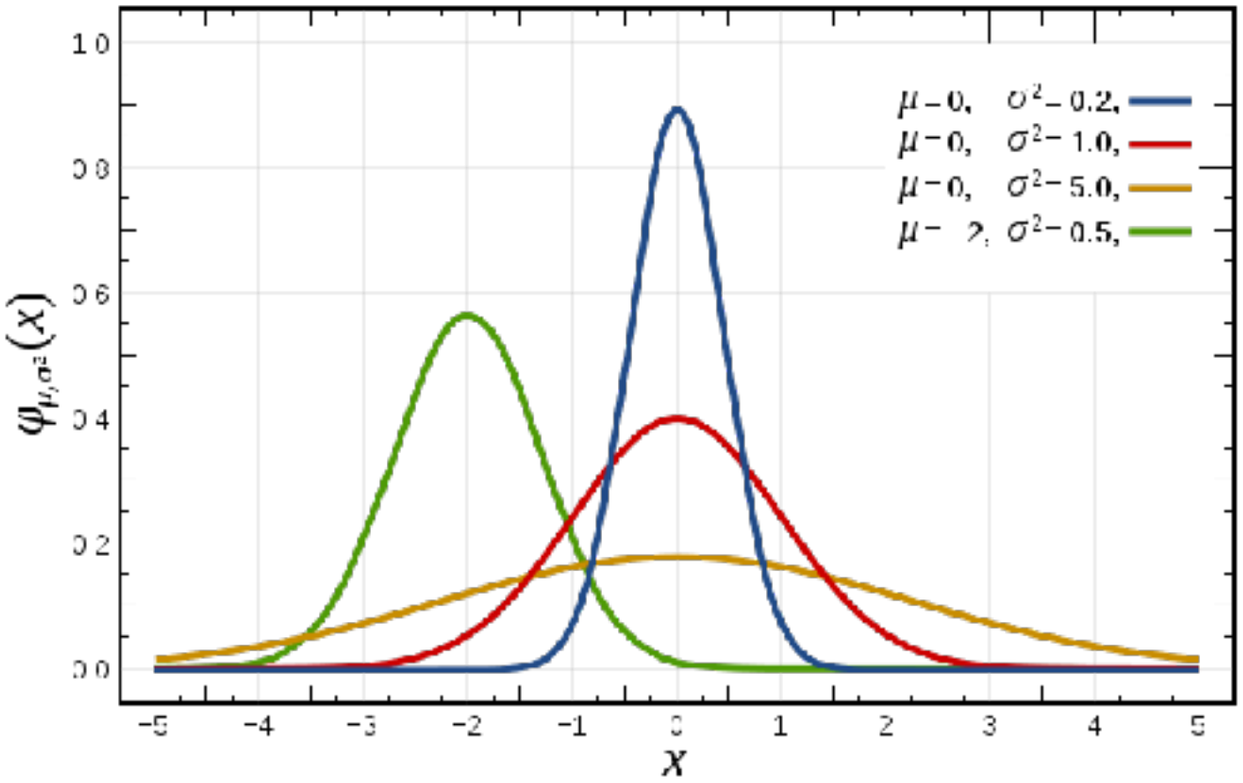
# DISTRIBUZIONI STATISTICHE

/ Distribuzione normale o gaussiana

/ Distribuzione binomiale

/ Distribuzione di Poisson

# DISTRIBUZIONE NORMALE

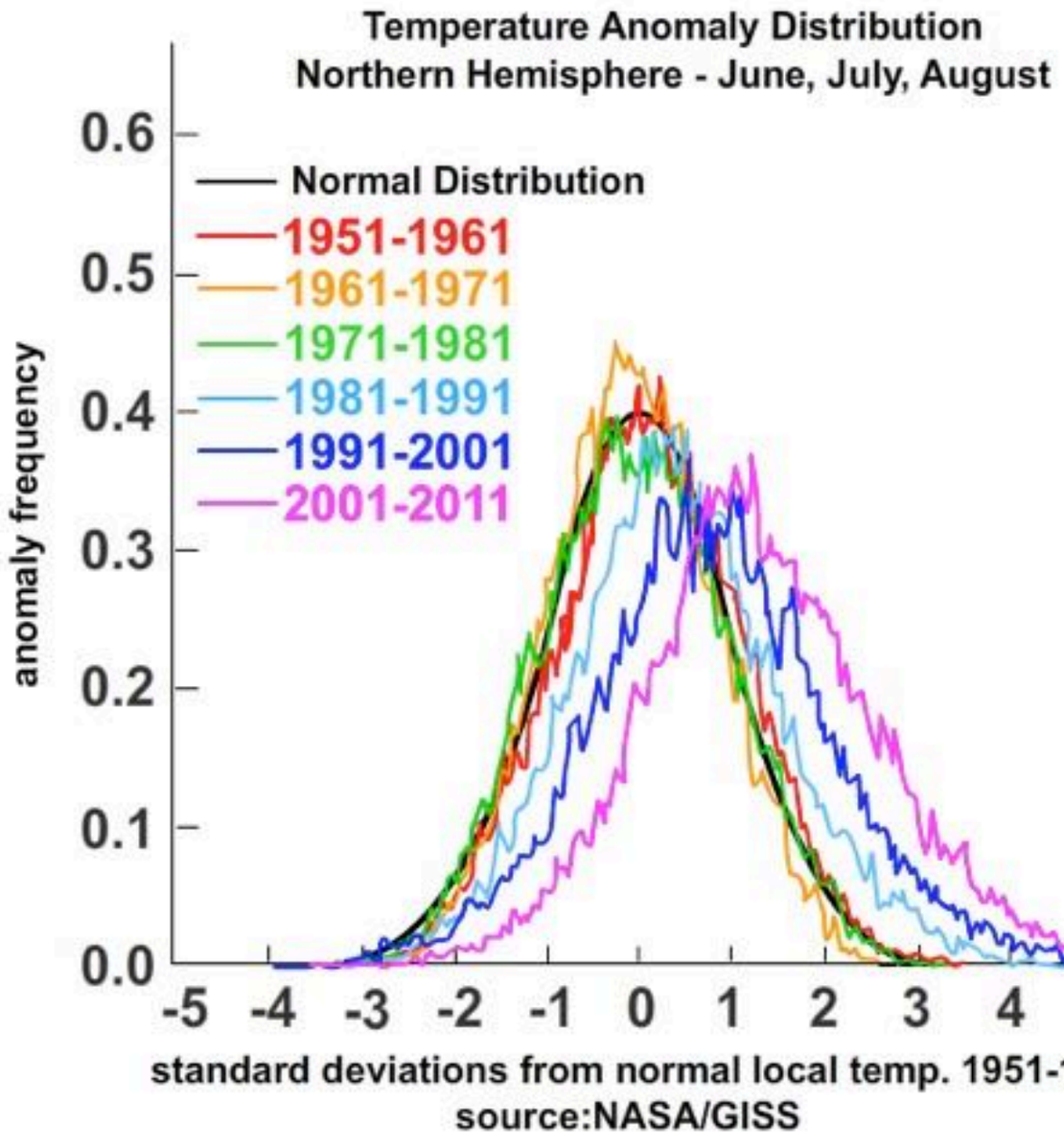



/ **Nella distribuzione normale la media, moda e mediana coincidono**

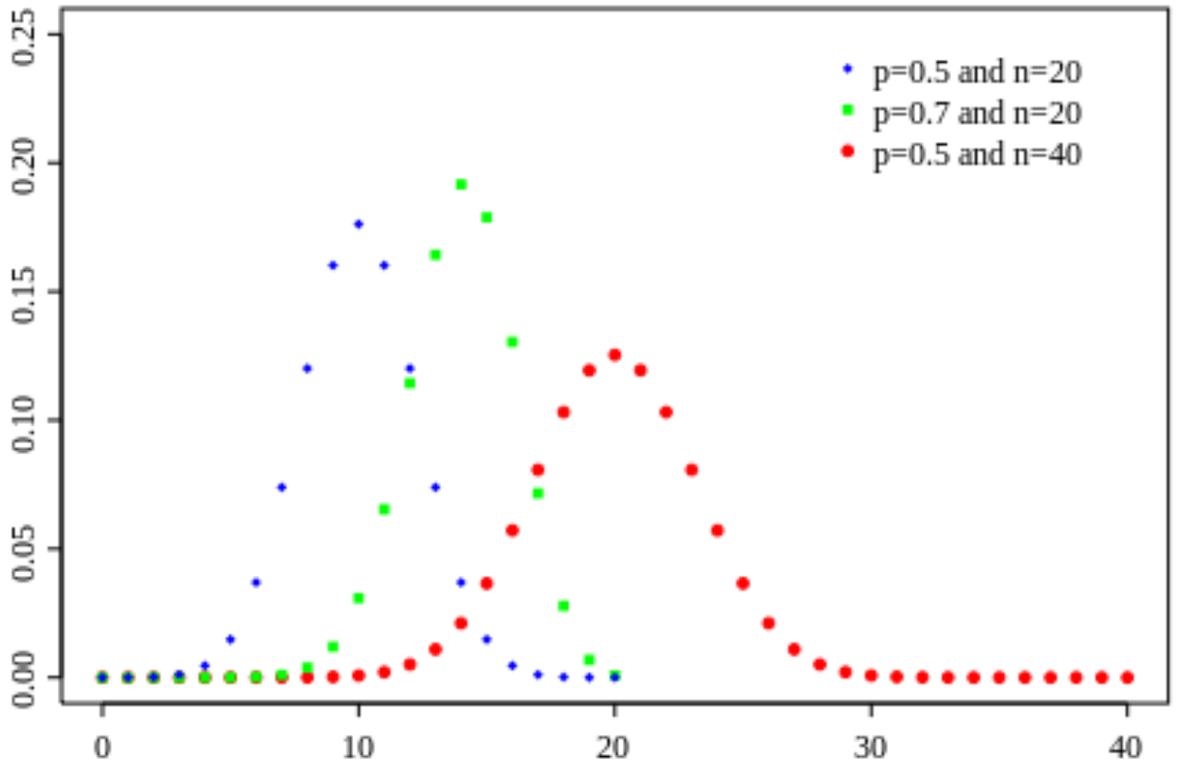
/ **Densità di probabilità (detta PDF, Probability Density Function)**

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

# DISTRIBUZIONE NORMALE
# ESEMPIO



# DISTRIBUZIONE BINOMIALE



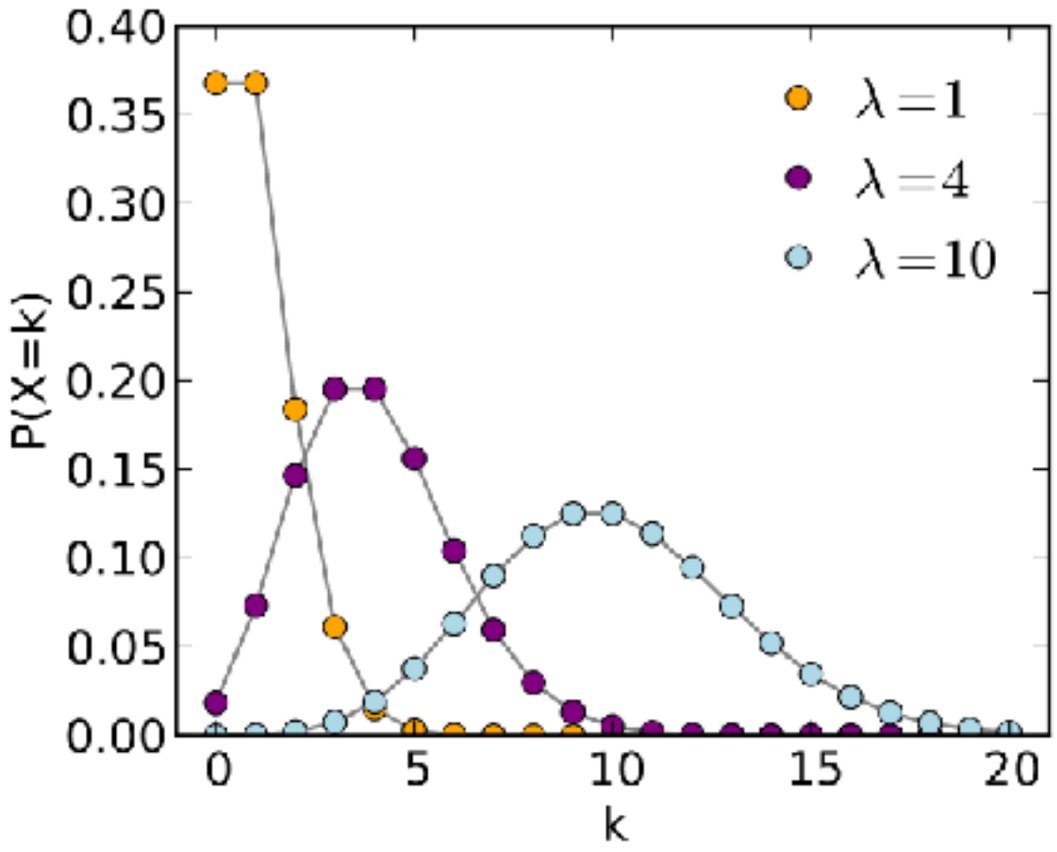

- **n, numero di prove effettuate**
- **k, probabilità di successo**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- **Distribuzione discreta con densità di probabilità (PDF, Probability Density Function):**

$$P(k) = P(X_1 + X_2 + \cdots + X_n = k) = \binom{n}{k} p^k q^{n-k}$$

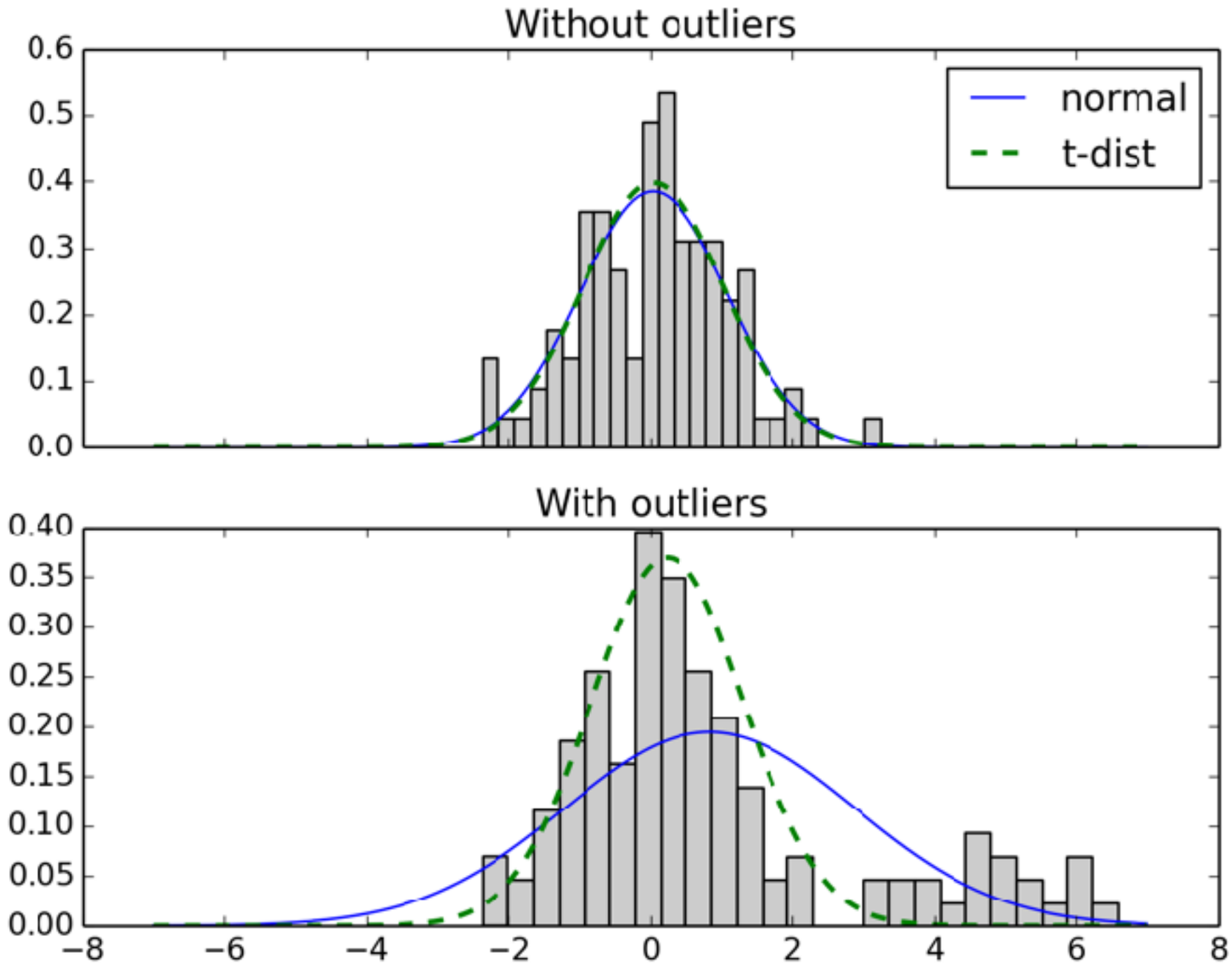
# DISTRIBUZIONE DI POISSON




- λ numero medio di eventi per intervallo di tempo
- k numero di eventi per intervallo di tempo di cui vogliamo la probabilità
- Eventi indipendenti e compresi in un intervallo di tempo fissato
- Probabilità di osservare k eventi nell'intervallo dato

$$f(k;\lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

# DISTRIBUTIONI E OUTLIERS

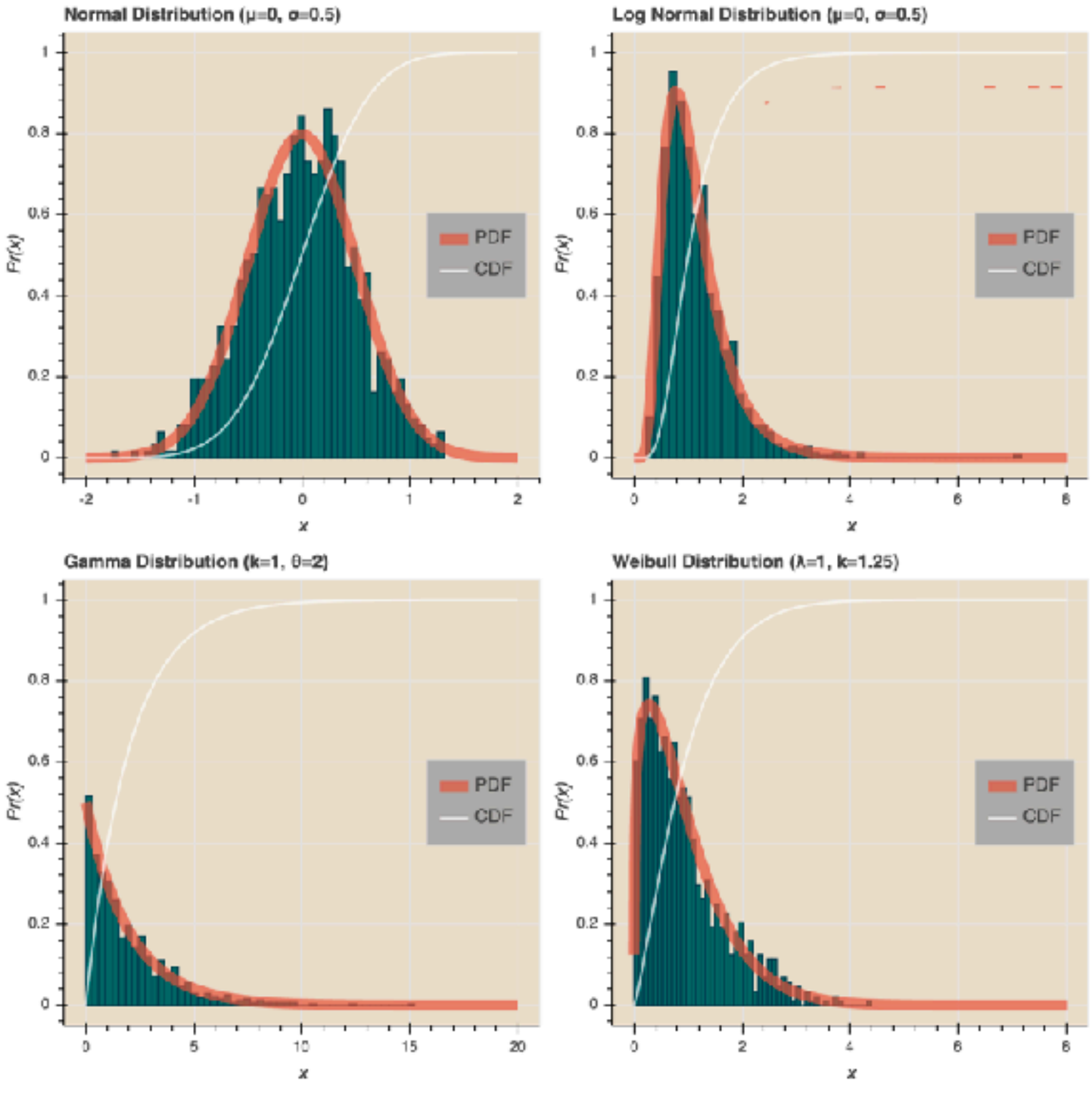


# FUNZIONE DI RIPARTIZIONE

**CDF (Cumulative Distribution Function) è una funzione che rappresenta se un fenomeno X è presente prima o dopo un certo valore x:**

$$F(x) = P(X \leq x)$$

# DISTRIBUZIONI E CDF



# STATISTICA
# INFERENZIALE

# STATISTICA INFERENZIALE PREDITTIVA

**Predizione delle caratteristiche della popolazione e valutazione di ipotesi dei fenomeni tramite l'osservazione di un campione selezionato.**

**Per tale motivo la conoscenza sarà parziale e possono sussistere errori di valutazione.**

# STATISTICA INFERENZIALE PREDITTIVA
## ESEMPIO

**Calcolo della differenza media in percentuale dell'altezza media fra uomini e donne nel mondo.**

**Non avendo i dati dell'intero campione mondiale, si prendono campioni di sottoinsiemi e si fa inferenza, ossia si traggono deduzioni sul campione mondiale.**

# FREQUENTISTI E BAYESIANI

**Statistica inferenziale frequentista:**
**i parametri sono fissati e i dati possono essere rappresentati con un campione random statistico.**

**Statistica inferenziale bayesiana:**
**i dati sono fissati e i parametri che non sono conosciuti possono essere descritti probabilisticamente (ad esempio la media delle altezze).**

# FREQUENTISTI E BAYESIANI

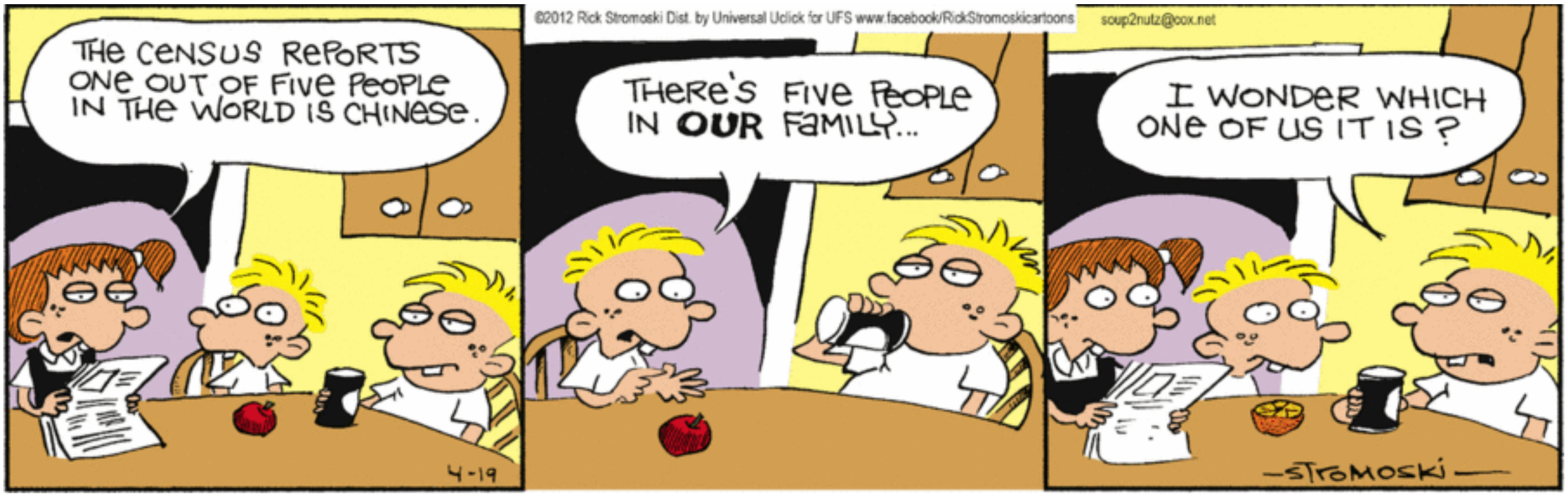
## PROBABILITÀ

**Approccio alla probabilità frequentista: solo eventi casuali ripetibili (ad esempio il lancio di una moneta) possono avere probabilità.**

**Approccio alla probabilità bayesiana: si possono usare probabilità per rappresentare l'incertezza di ogni evento o ipotesi.**

# STATISTICA INFERENZIALE PREDITTIVA
# CAMPIONE E INTERPRETAZIONE

```
int getRandomNumber()
{
    return 4;  // chosen by fair dice roll.
               // guaranteed to be random.
}
```



# TEST DI IPOTESI STATISTICA
# VALORE DI PROBABILITÀ O P-VALUE

**P-value è la probabilità per un dato modello statistico che, quando l'ipotesi nulla è vera, la sintesi statistica scelta sarebbe la stessa o di una magnitudine maggiore dell'attuale risultato osservato.**

# P-VALUE
# IPOTESI NULLA

**In ogni esperimento ci si aspettano alcune differenze e osservazioni nel testare l'ipotesi di un nuovo fenomeno.**

**C'è però sempre da tener conto la possibilità che tale fenomeno non si verifichi e non si osservino differenze di misura.**

**Quest'ultima possibilità e detta ipotesi nulla.**

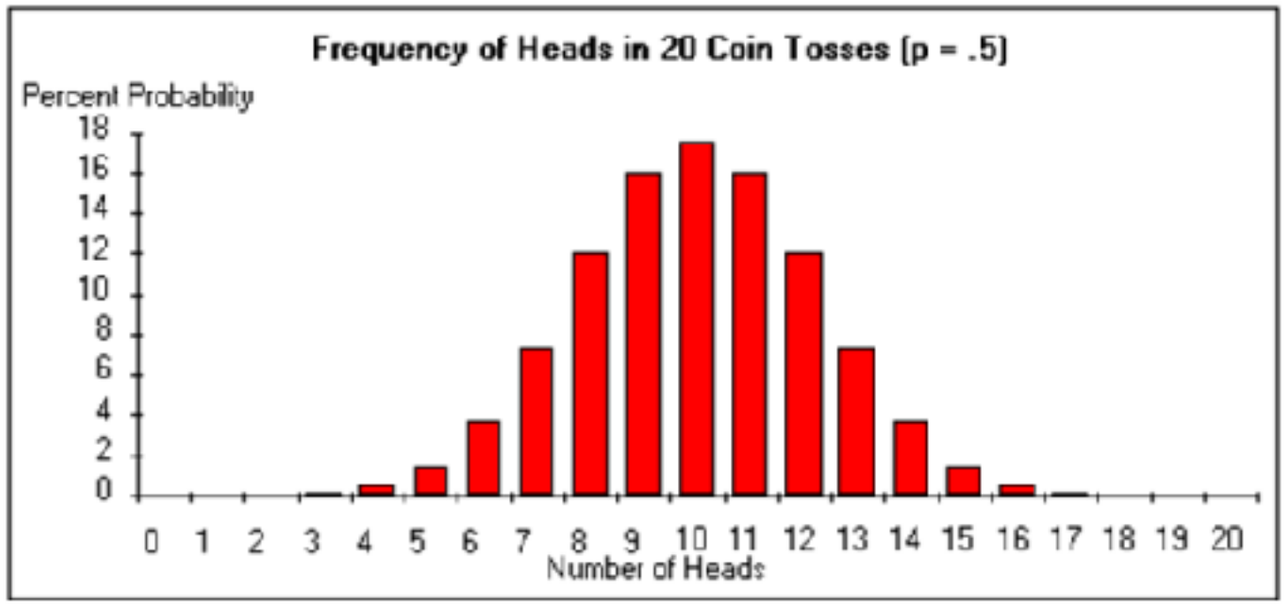
# P-VALUE ESEMPIO


**Clark dice di avere una moneta truccata.**

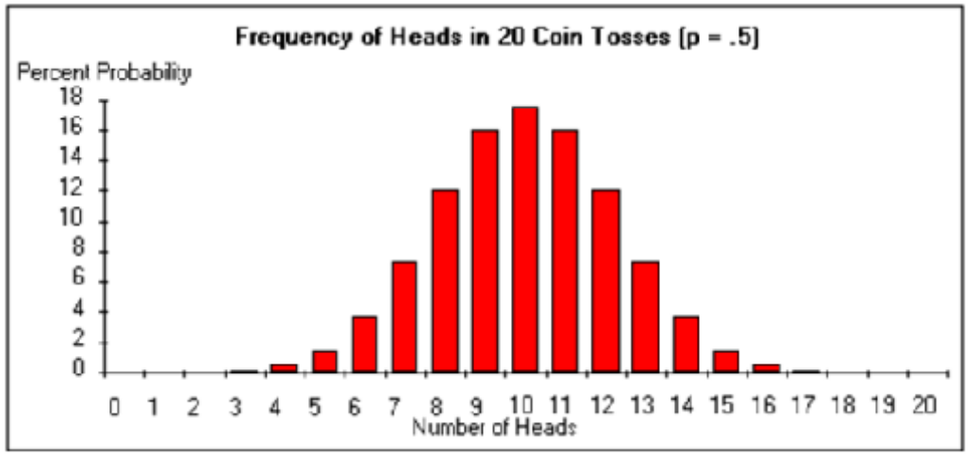
**Decidiamo di condurre un esperimento per provare che non è vero.**

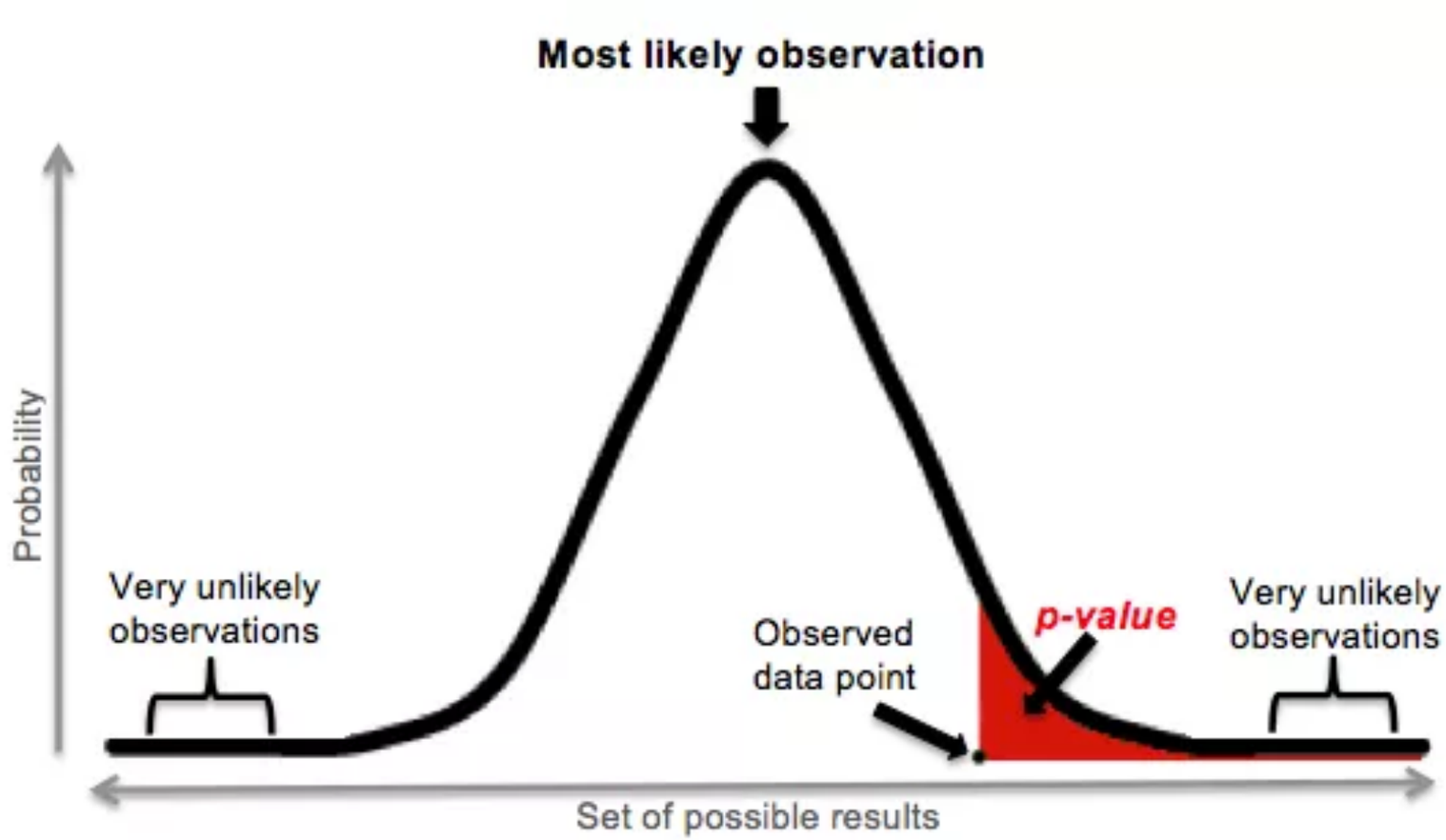
**L'ipotesi nulla $H_0$ è quindi: "La moneta è normale come tutte le altre"**

**Distribuzione delle osservazioni se $H_0$ fosse vera** →



# P-VALUE ESEMPIO




**A *p-value* (shaded red area) is the probability of an observed (or more extreme) result arising by chance**

# P-VALUE ESEMPIO

**Il p-value esprime il livello di significatività osservato, ossia una misura dell'evidenza contro l'ipotesi nulla $H_0$.**

**Assumendo un valore limite tipico di $\alpha = 0{,}05$ possiamo dire che:**

- **se $p \leq \alpha \rightarrow H_0$ può essere scartata, e i dati osservati sono statisticamente significativi,**

- **se $p > \alpha \rightarrow H_0$ non può essere scartata.**

# STATISTICAL HYPOTHESIS TESTING
# P-VALUE IN RESEARCH


"Data don't make any sense,
we will have to resort to statistics."

# STATISTICAL HYPOTHESIS TESTING
# P-VALUE IN RESEARCH

**Un uso erroneo del p-value può portare a:**

- **Dare un'apparente validità a risultati dubbi**
- **Sottostimare importanti risultati che non sembrano abbastanza significativi**

**https://www.vox.com/2016/3/15/11225162/p-value-simple-definition-hacking**

# STATISTICAL HYPOTHESIS TESTING
# P-VALUE NELLA RICERCA

## Nel 2016 l'American Statistical Association ha rilasciato sei principe per l'interpretazione del p-value:

- **P-value può indicare quanto incompatibili I dati possano essere all'interno di un specifico modello statistico**
- **P-value non assicura che un'ipotesi sia vera**
- **Decisioni di business e policy non devono essere prese solo sulla valutazione del p-value**
- **Porre attenzione alla trasparenza e riproducibilità della ricerca**
- **P-value non misura la reale ampiezza di un effetto o l'importanza di un risultato**
- **Il solo valore del P-value non provvede ad una buona misura dell'evidenza di un certo modello di ipotesi**

# RISORSE

# TESTI DI STATISTICA

## / “Think Stats” di Allen B. Downey

## Contiene esempi di analisi statistica in Python

# ULTERIORI RISORSE CITATE

- **Portale di Wikipedia di Statistica**
  **https://it.wikipedia.org/wiki/Portale:Statistica**

- **Grafici generati con Matplotlib (libreria Python)**
  **https://matplotlib.org/gallery/index.html**

- **Grafici generati con Bokeh (libreria Python)**
  **https://bokeh.pydata.org/en/latest/docs/gallery.html**

stefania.delprete@top-ix.org

www.top-ix.org

@top_ix

# GRAZIE!