# Probabilità e Statistica[1]

Isadora Antoniano-Villalobos
isadora.antoniano@unive.it

Laurea in Informatica
(Data science/ Tecnologie e scienze dell'informazione)
Università Ca' Foscari di Venezia

Anno academico 2023/2024

[1] Materiale didattico redatto da: Isadora Antoniano-Villalobos & Federica Giummolè

# Introduzione

# L'era dell'informazione?

- Si dice che viviamo nell'era dell'informazione. Ma ch'è anche tanta disinformazione e *fake news*
- Forse è più corretto dire che viviano nell'era dei **dati**
- La quantità di dati generati dall'uomo è talmente grande che diventa difficile memorizzarle, gestirle, verificarle ed interpretarle.

- Per estrarre l'informazione dai dati, dobbiamo processarli adeguatamente

# La statistica

La statistica ci fornisce strumenti che consentono di estrarre dai dati l'informazione che nascondono.
Uno statistico sa:

- combinare informazioni di tipo differente
- valutarne l'affidabilità
- sintetizzare e presentare molti dati
- costruire modelli
- calcolare previsioni e formulare ipotesi di decisione.

👉 La statistica non è l'unico strumento per estrarre informazione dai dati, ma è quello più adatto in presenza di incertezza

# Data Science

Naturalmente, anche l'informatica svolge un ruolo fondamentale nel salvataggio e nella gestione dei dati:

**Il lavoro più sexy del XXI secolo (chiedete a Google!)**

**Statistics + Computer Science = Data Science**

# Alcuni termini statistici

- La **popolazione di riferimento**, è un insieme di elementi chiamati **unità statistiche** (individui, animali, immagini,...). Rappresenta il fenomeno o la parte del mondo che ci interessa studiare per produrre conoscenza o prendere decisioni.
- I **dati** sono il risultato di rilevare o misurare alcune caratteristiche di tutte o di una parte della popolazione, ottenendo valori potenzialmente diversi per ogni unità statisticha.
- La **statistica** ci permete di estrarre l'informazione dai dati, generando nuove conoscenze o ipotesi di decisione. Ovvero, di trasformare i dati in affermazioni sul mondo (sulla popolazione di riferimento).
- Ogni caratteristica rilevata sulle unità statistiche si chiama **variabile** e i dati corrispondenti a ogni variabile sono le **realizzazioni**.
- Se le variabili non sono rilevate su tutte le unità statistiche, il sottoinsieme della popolazione oggetto della rilevazione è chiamato il **campione**.

# Tipi di variabili

1. Una variabile è **qualitativa** o **categorica** quando i suoi possibili valori o **modalità** prendono la forma di aggettivi o di altre espressioni verbali oppure (anche se sostituite da etichette numeriche). Le variabili qualitative possono essere:
   - **Sconnesse** se non esiste nessun ordinamento naturale tra le modalità, ad esempio il sesso, o il tipo di servizio offerto da un albergo
   - **Ordinali** nel caso in cui un ordinamento naturale esiste, ad esempio il massimo titolo di studio, o il parere di un intervistato (classificato come "mediocre", "discreto", "buono")

   👉 Quando le modalità sono solamente due (ad esempio maschio vs. femmina, vivo vs. morto o buono vs. difettoso) si parla di variabili **dicotomiche** o **binarie**.

# Tipi di variabili

2. Una variabile è **numerica** o **quantitativa** quando le **modalità** sono espresse da numeri. Dal punto di vista dei modelli e delle tecniche utilizzate le variabili numeriche si suddividono in:
   - **Discrete** o **intere** quando le modalità sono esprimibili da numeri interi, ad esempio il numero di clienti di un negozio o il numero di visualizzazioni di un video
   - **Continue** o **reali** quando le modalità sono esprimibili da numeri reali, ad esempio il tempo d'attesa ad uno sportello, il peso di un manufatto

# Piccolo esempio

Vogliamo avere un'idea sul numero di *account* e sul volume delle vendite di 20 siti di *e-commerce* classificati secondo tre categorie ritenute simili. Le **unità statistiche** sono i diversi siti. I **dati** si presentano in questa forma:

| sito | account | vendite | categoria |
|---|---|---|---|
| 1 | 907 | 11.2 | A |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 10 | 420 | 6.12 | B |
| 11 | 679 | 7.63 | B |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 19 | 1010 | 11.77 | C |
| 20 | 621 | 7.41 | A |

Le **variabili** considerate nello studio sono tre:

1. **account:** numericha discreta
2. **vendite:** numericha continua
3. **categoria:** qualitativa sconnessa

# La Statistica

1. **Campionamento e disegno degli esperimenti:** si occupano delle problematiche connesse con la raccolta dei dati:
   - Esperimenti in laboratorio
   - Interviste telefoniche
   - Selezione di dati disponibili attraverso i social network, carte fedeltà, telefonia mobile, etc

   ⋮
2. **Statistica Descrittiva:** metodi per rappresentare, sintetizzare ed evidenziare le caratteristiche più significative di un insieme di dati. Speso si dispone di dati su tutta la popolazione di riferimento.
3. **Inferenza statistica:** i dati disponibili sono stati rilevati solamente su una parte delle unità statistiche, il campione (da cui il termine *indagini campionarie*). Si vogliono utilizzare le informazioni del campione per fare delle affermazioni sulle caratteristiche generali di tutta la popolazione.

# La Statistica

Nelle applicazioni, **Statistica Descrittiva** ed **Inferenza Statistica** non sono facilmente separabili. Infatti i problemi di inferenza vengono normalmente affrontati in accordo con lo schema:


La statistica descrittiva viene dunque utilizzata per un'analisi preliminare delle caratteristiche del campion

# Il Calcolo delle probabilità

- Perché l'inferenza porti a risultati sensati, bisogna che sia noto il legame fra popolazione e campione.
- Il **calcolo delle probabilità** fornisce i **modelli matematici** utili per descrivere la relazione fra campione e popolazione.
- Il **calcolo delle probabilità** è lo strumento necessario per l'inferenza. Permette di **quantificare gli errori** che commettiamo nel passaggio dal particolare (campione) al generale (popolazione).

**Il calcolo delle probabilità si occupa dello studio di fenomeni in condizioni di incertezza:**

- come valutare la probabilità o la possibilità di diversi esiti
- come scegliere un modello per un fenomeno incerto e usarlo per prendere delle decisioni
- come valutare le caratteristiche o i parametri di uno certo fenomeno

⋮

# Il corso di Probabilità e Statistica...

...è dedicato al **calcolo delle probabilità**.

👉 Introdurremo i principali strumenti matematici indispensabili per l'inferenza, ovvero per affrontare l'incertezza di ogni giorno, fare delle previsioni e prendere le decisioni appropriate!

1. Probabilità elementare
2. Distribuzioni di probabilità
3. Distribuzioni congiunte e teoremi limite
4. Catene di Markov

# Il resto della storia...

... ve la raccontiamo nei corsi del curriculum di **Data Science** (https://www.unive.it/data/it/1632/data-science):

- Analisi dei dati
- Analisi predittiva
- Data and web mining
- Social network analysis