

# Statistica II (750AA)

## Lezione 1

Dario Trevisan – *<https://web.dm.unipi.it/trevisan>*

22/09/2025

# **Introduzione al corso**

# Argomenti

0. Richiami di Statistica I (descrittiva, probabilità, test e intervalli)

# Argomenti

- 0. Richiami di Statistica I (descrittiva, probabilità, test e intervalli)
- 1. Clustering

# Argomenti

- 0. Richiami di Statistica I (descrittiva, probabilità, test e intervalli)
- 1. Clustering
- 2. Classificazione

# Argomenti

- 0. Richiami di Statistica I (descrittiva, probabilità, test e intervalli)
- 1. Clustering
- 2. Classificazione
- 3. Regressione

# Argomenti

0. Richiami di Statistica I (descrittiva, probabilità, test e intervalli)
1. Clustering
2. Classificazione
3. Regressione
4. Serie storiche

# Argomenti

- 0. Richiami di Statistica I (descrittiva, probabilità, test e intervalli)
  - 1. Clustering
  - 2. Classificazione
  - 3. Regressione
  - 4. Serie storiche
- L'apprendimento di R è centrale nel corso: installate R ed RStudio <https://posit.co/download/rstudio-desktop/>

# Materiale didattico

- ▶ MS Team: **750AA 25/26 - STATISTICA II [WIG-LM]**

Troverete:

# Materiale didattico

- ▶ MS Team: **750AA 25/26 - STATISTICA II [WIG-LM]**
- ▶ Pagina web del corso  
*<https://web.dm.unipi.it/trevisan/it/courses/750AA.html>*

Troverete:

# Materiale didattico

- ▶ MS Team: **750AA 25/26 - STATISTICA II [WIG-LM]**
- ▶ Pagina web del corso  
*<https://web.dm.unipi.it/trevisan/it/courses/750AA.html>*

Troverete:

- ▶ Appunti (slides annotate)

# Materiale didattico

- ▶ MS Team: **750AA 25/26 - STATISTICA II [WIG-LM]**
- ▶ Pagina web del corso  
*<https://web.dm.unipi.it/trevisan/it/courses/750AA.html>*

Troverete:

- ▶ Appunti (slides annotate)
- ▶ Demo interattive

# Materiale didattico

- ▶ MS Team: **750AA 25/26 - STATISTICA II [WIG-LM]**
- ▶ Pagina web del corso  
*<https://web.dm.unipi.it/trevisan/it/courses/750AA.html>*

Troverete:

- ▶ Appunti (slides annotate)
- ▶ Demo interattive
- ▶ R notebooks (codice e commenti)

# Materiale didattico



▶ MS Team: **750AA 25/26 - STATISTICA II [WIG-LM]** }

▶ Pagina web del corso

*<https://web.dm.unipi.it/trevisan/it/courses/750AA.html>*

Troverete:

▶ Appunti (slides annotate)

▶ Demo interattive

▶ R notebooks (codice e commenti)

▶ RegISTRAZIONI delle lezioni

# Materiale didattico

- ▶ MS Team: **750AA 25/26 - STATISTICA II [WIG-LM]**
- ▶ Pagina web del corso  
*<https://web.dm.unipi.it/trevisan/it/courses/750AA.html>*

Troverete:

- ▶ Appunti (slides annotate)
- ▶ Demo interattive
- ▶ R notebooks (codice e commenti)
- ▶ RegISTRAZIONI delle lezioni
- ▶ Test d'esame

# Orario lezioni

- ▶ Lunedì F4 14:30 - 16:30

# Orario lezioni

- inizio 14:30  
pausa 10 fine 16:10
- ▶ Lunedì F4 14:30 - 16:30
  - ▶ Mercoledì B25 8:30 - 11:30 (portate computer/tablet)

✓  
inizio 8:40  
↓  
2 pause 10 min  
fine 11:20

# Ricevimento

► **Giorno:**

# Ricevimento

- ▶ Giorno:
- ▶ Orario:

# Ricevimento

- ▶ Giorno: MARTEDÌ
- ▶ Orario: 18-20
- ▶ Comunque su appuntamento, anche su Teams

# Ricevimento

- ▶ Giorno:
- ▶ Orario:
- ▶ Comunque su appuntamento, anche su Teams
- ▶ Contattatemi **sempre** via mail o messaggio su Teams.

# Modalità di esame

- ▶ **prova teorica:** test scritto con domande sugli argomenti del corso (vero/falso da argomentare brevemente) durata 1 ora.

# Modalità di esame

- ▶ **prova teorica:** test scritto con domande sugli argomenti del corso (vero/falso da argomentare brevemente) durata 1 ora.
- ▶ **prova pratica:** progetti autonomi di analisi statistica di dati reali, su **almeno due dei tre argomenti principali** (classificazione, regressione, serie storiche). Vedere foglio di indicazioni generali sulle relazioni.

# Modalità di esame

*risultati appunti*

- ▶ **prova teorica**: test scritto con domande sugli argomenti del corso (vero/falso da argomentare brevemente) durata 1 ora.
- ▶ **prova pratica**: progetti autonomi di analisi statistica di dati reali, su **almeno due dei tre argomenti principali** (classificazione, regressione, serie storiche). Vedere foglio di indicazioni generali sulle relazioni.
- ▶ Il voto finale tiene conto di entrambe le prove.

## Altre cose:

- si può ~~sottrarre~~  
nuove relazioni su nuovi dati  
e valgono sempre le due  
più recenti
- pre-appello a dicembre

## Statistica descrittiva

Ross

# Indicatori di centralità e variabilità

Date osservazioni (un campione statistico)  $\{x_i\}_{i=1}^n$  definiamo gli indicatori di centralità:

► moda  $\rightarrow$  il valore che osserviamo con maggiore frequenza

e gli indicatori di variabilità (o dispersione):

taglia n

# Indicatori di centralità e variabilità

Date osservazioni (un campione statistico)  $\{x_i\}_{i=1}^n$  definiamo gli indicatori di centralità:

► moda

► mediana : si ordinano i dati in ordine crescente

$$x_{(1)} < x_{(2)} < x_{(3)} < \dots < x_{(n)}$$

e gli indicatori di variabilità (o dispersione):

# Indicatori di centralità e variabilità

Date osservazioni (un campione statistico)  $\{x_i\}_{i=1}^n$  definiamo gli indicatori di centralità:

► moda

► mediana

► media campionaria

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

e gli indicatori di variabilità (o dispersione):

# Indicatori di centralità e variabilità

Date osservazioni (un campione statistico)  $\{x_i\}_{i=1}^n$  definiamo gli indicatori di centralità:

- ▶ moda
- ▶ mediana
- ▶ media

e gli indicatori di variabilità (o dispersione):

- ▶ quantili (quartili, percentili)

livello  $\alpha \in (0,1)$



# Indicatori di centralità e variabilità

Date osservazioni (un campione statistico)  $\{x_i\}_{i=1}^n$  definiamo gli indicatori di centralità:

▶ moda

▶ mediana

▶ media  $\bar{x}$

e gli indicatori di variabilità (o dispersione):

▶ quantili (quartili, percentili)

▶ varianza (deviazione standard)

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \sigma = \sqrt{\sigma^2}$$

# Moda

# Mediana

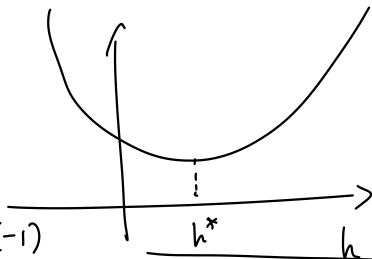
# Media

## Varianza

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$= \min_{h \in \mathbb{R}} \frac{1}{n-1} \sum_{i=1}^n (x_i - h)^2 \quad \text{e } h^* = \bar{x}$$

$$h \mapsto \frac{1}{n-1} \sum_{i=1}^n (x_i - h)^2$$



$$0 = \frac{d}{dh} \sum_{i=1}^n (x_i - h)^2 = \sum_{i=1}^n 2(x_i - h)(-1)$$

$$\sum_{i=1}^n x_i - n \cdot h = 0 \Rightarrow \boxed{h^* = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}}$$

# Funzioni di Perdita/Costo/Errore

La media  $\bar{x}_n$  minimizza il rischio quadratico (MSE)

$$\bar{x}_n \in \arg \min_{h \in \mathbb{R}} \sum_{i=1}^n (x_i - h)^2$$

- Introduciamo nuove funzioni di perdita/costo/errore  $\ell(x; h)$  e **minimizziamo il rischio empirico** (ERM):

$$\arg \min_{h \in \mathbb{R}} \sum_{i=1}^n \ell(x_i; h).$$



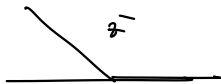
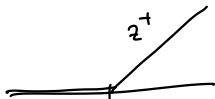
## Errore assoluto asimmetrico

$$\ell(x, h) = (x-h)^+ + p (x-h)^-$$

$$p > 0$$

$$z^+ = \max\{0, z\}$$

$$z^- = \max\{-z, 0\}$$



→ quantili

## Errore 0-1

$$\ell(x, h) = \begin{cases} 1 & \text{se } h \neq x \\ 0 & \text{se } h = x \end{cases}$$

$$h^* \in \operatorname{argmin}_h \sum_{i=1}^n \ell(x_i, h) = \operatorname{argmin}_h \sum_{i=1}^n \mathbb{1}_{\{h \neq x_i\}}$$

$$h^* \equiv (n+1) \bmod 2$$

## Errore esponenziale assoluto

$$\ell(x, h) = \exp(|x - h|)$$



$h^*$

$x$   
 $\uparrow$   
outlier

# WERMS



► Giocate voi! <https://dario-trevisan.shinyapps.io/WERMS/>

# WERMS



- ▶ Giocate voi! <https://dario-trevisan.shinyapps.io/WERMS/>
- ▶ Codice (R shiny) disponibile nel team e nella pagina del corso

## **Dati multivariati**

# Distanze tra punti nello spazio

Nella statistica multivariata si osserva

- un campione di  $n$  individui (taglia):  $x_i$ , per  $i = 1, \dots, n$

Possiamo organizzare i dati in una tabella (*data frame* in R).



A hand-drawn diagram of a data frame table. It consists of a grid with 4 rows and 3 columns. To the left of the grid, the rows are numbered 1, 2, 3, and 4. Above the grid, the columns are labeled 'Area', 'Per', and 'Alma'.

|   | Area | Per | Alma |
|---|------|-----|------|
| 1 |      |     |      |
| 2 |      |     |      |
| 3 |      |     |      |
| 4 |      |     |      |

$n$  righe  
 $d$  colonne

# Distanze tra punti nello spazio

$$(x_i)_{i=1, \dots, d} \quad (h_i)_{i=1, \dots, d}$$

Nella statistica multivariata si osserva

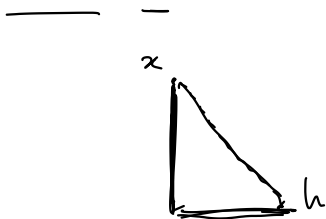
- ▶ un campione di  $n$  individui (tagliati)  $x_i$ , per  $i = 1, \dots, n$   
 $\|x - h\| = \sqrt{\sum (x_i - h_i)^2}$
- ▶ e  $d$  caratteristiche (features) per ciascun individuo  $x_{i,j}$ , per  $1, \dots, d$ .

Possiamo organizzare i dati in una tabella (*data frame* in R).

# Distanze tra punti nello spazio

Confrontare punti nello spazio  $\mathbb{R}^d$  è cruciale per l'analisi di dati multivariati. Distanze comuni:

► Euclidea:



$$\|x-h\|_1 = \sum_{i=1}^d |x_i - h_i|$$

# Distanze tra punti nello spazio

Confrontare punti nello spazio  $\mathbb{R}^d$  è cruciale per l'analisi di dati multivariati. Distanze comuni:

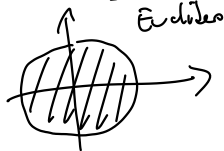
- ▶ Euclidea:
- ▶ Manhattan ( $\ell^1$ , taxicab):

# Distanze tra punti nello spazio

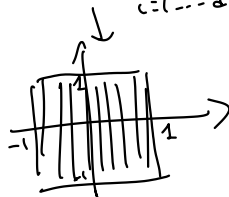
Confrontare punti nello spazio  $\mathbb{R}^d$  è cruciale per l'analisi di dati multivariati. Distanze comuni:

- ▶ Euclidea:
- ▶ Manhattan ( $\ell^1$ , taxicab):
- ▶  $\ell^p$  (Minkowski):

$$\{ \|x\| \leq 1 \}$$



$$\|x - h\|_{\infty} = \max_{i=1 \dots d} |x_i - h_i|$$



# Distanze tra punti nello spazio

Confrontare punti nello spazio  $\mathbb{R}^d$  è cruciale per l'analisi di dati multivariati. Distanze comuni:

- ▶ Euclidea:
- ▶ Manhattan ( $\ell^1$ , taxicab):
- ▶  $\ell^p$  (Minkowski):
- ▶  $\ell^\infty$  (del massimo):

# Centroide e Medoide

# Medoids!



- ▶ Giocate voi! <https://dario-trevisan.shinyapps.io/medoids/>