

# STATISTICHE RIASSUNTIVE E GRAFICI

## Statistica Descrittiva

Dati analizzati senza assunzioni esterne  
 => scopo: organizzare dati, evidenziarne la struttura

## Statistica Inferenziale

Studio dei dati tramite un modello probabilistico  
 => scopo: usare il modello di tale stima per fare previsioni

## Popolazione

Insieme di oggetti o fenomeni da studiare  
 es. studenti università di Pisa → Popolazione reale  
 tutti i possibili lanci di un dado → Popolazione ideale

## Carattere

Caratteristica degli individui della popolazione, ottenuta con stesso procedimento  
 es. altezza studenti

## Modalità

Possibili valori che può assumere il carattere  
 es. 180 cm

## Campione

Sottoinsieme della popolazione  
 es. 10 italiani

## Dati

Esiti delle misure

## Problema

Capire quando un campione è rappresentativo (problema complesso)

## CARATTERI

### Quantitativi

Se gli esiti delle misure (modalità) sono paragonabili fra loro  
 es. 180 cm, 173 cm, 165 cm

### Qualitativi

Se le modalità non sono paragonabili fra loro  
 es. associare un numero a un nome e confrontarli:

### Discreto

Assume una quantità finita (e piccola) di valori  
 es. esiti dei lanci di un dado (1, 2, 3, 4, 5, 6)

### Continuo

Assume una grande quantità di valori  
 es. altezza degli italiani

## FREQUENZA

### Frequenza Assoluta

Numero di volte che una modalità compare nei dati

### Frequenza Relativa

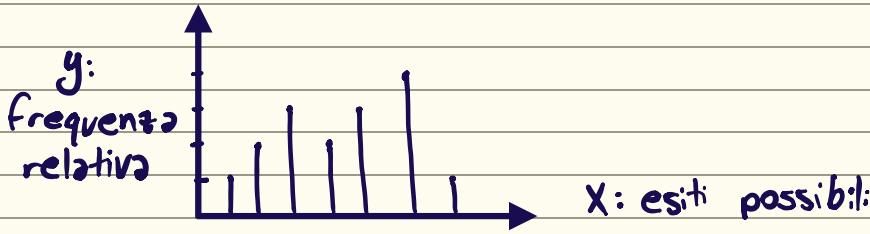
Frazione di volte che una modalità compare nei dati

es. 8 Lanci TCTTCTTT

Frequenza assoluta di T = 6

Frequenza relativa di T = 6/8 = 75%

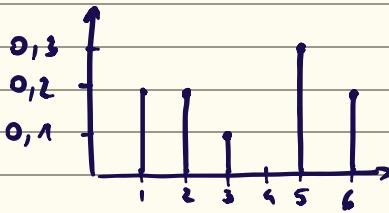
## DIAGRAMMA A BARRE



### Esempio di Esercizio

10 Lanci di dado, esito: 6, 2, 3, 1, 2, 5, 6, 5, 5, 1

- popolazione (ideale): tutti i possibili lanci del dado ( $\infty$ )
- campione: i 10 lanci effettuati
- carattere: esito del lancio di un dado
- modalità: possibili valori che puo' prendere il carattere (1, 2, ..., 6)



### INDICI

**Indice Statistico**

Quantitativo numerico che riassume qualche proprietà significativa di una distribuzione di dati  
 $\Rightarrow$  servono per effettuare sintesi dei dati

**Di centralità**

Indica dove si trova il centro di distribuzione dati

**Media**

Media aritmetica dei dati  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

**Mediana**

Il dato  $x_i$  tale che metà degli altri valori è  $\leq x_i$  e l'altra metà è  $\geq x_i$   
 Per calcolarla:

1) Ordino i dati  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

2) Divido il vettore

- Se  $(n)$  è dispari  $\Rightarrow$  mediana =  $x_{(\frac{n+1}{2})}$

- Se  $(n)$  è pari  $\Rightarrow$  mediana =  $\frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$

es.  $x = (8, 3, 5, 12, 10) \in \mathbb{R}^5$

$\Rightarrow$  media:  $\bar{x} = (8+3+5+12+10)/5 = 38/5 = 7,6$

$\Rightarrow$  mediana: ordino  $x = (3, 5, 8, 10, 12) \Rightarrow n=5$  è dispari  $\Rightarrow \frac{5+1}{2} = 3$

$x_{(3)} = 8$

↑

**Moda**

Dato più comune nel vettore  $x = (x_1, \dots, x_n)$

## Media Sfondata

Media effettuata sui dati che rimangono dopo aver tolto una certa frazione dei dati più grandi e una certa frazione di quelli più piccoli  
es. si elimina il voto più alto e il più basso

## Indici di variabilità

Misurano la dispersione dei dati intorno ai valori "tipici" individuati dagli indici di posizione (media, mediana, moda)

## Varianza campionaria

Si calcola quando si ha un campione della popolazione

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

## Varianza empirica

Si calcola quando si ha l'intera popolazione

$$\text{var}_e(x) = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 \xrightarrow{\text{se si conosce f. relativa}} \text{var}_e(x) = \sum_{i=1}^n x_i^2 \cdot \frac{f_i}{n} - \bar{x}^2$$

## Rapporto tra $\text{var}(x)$ e $\text{var}_e(x)$

$$\text{var}(x) = \frac{n}{n-1} (\text{var}_e(x))$$

## Esempio varianza campionaria ed empirica

10 Lanci di dado, esito: 6, 2, 3, 1, 2, 5, 6, 5, 5, 1

$$\begin{aligned} \text{var}_e(x) &= 1^2 \cdot \frac{2}{10} + 2^2 \cdot \frac{2}{10} + 3^2 \cdot \frac{1}{10} + 4^2 \cdot \frac{0}{10} + 5^2 \cdot \frac{3}{10} + 6^2 \cdot \frac{1}{10} - (3,6)^2 \\ &= 16,6 - 12,96 = 3,64 \end{aligned}$$

$$\text{var}(x) = \frac{10}{9} \cdot \text{var}_e(x) = \frac{10}{9} \cdot 3,64 \approx 4,04$$

## Deviazione Standard

Chiamata anche Scarto quadratico medio, si divide in 2 tipi:

1) Campionario:  $\sigma(x) = \sqrt{\text{var}(x)}$

2) Empirico:  $\sigma_e(x) = \sqrt{\text{var}_e(x)}$

## Sample Skewness

Indica quanto è asimmetrica la distribuzione

$$b = \frac{1}{\sigma^3} \cdot \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^3$$

• Esalta con segno positivo gli  $x_i >> \bar{x}$

• Esalta con segno negativo gli  $x_i << \bar{x}$

## Funzione di ripartizione empirica (c.d.f.)

Restituisce il numero di dati presenti sull'asse

$$F_e: \mathbb{R} \rightarrow \mathbb{R} \implies F_e(t) = \frac{\#\{i \mid x_i \leq t\}}{n} \quad \text{f. relativa dei valori} \leq t$$

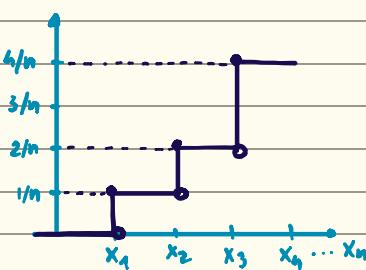
• Si parte da  $(x_1, \dots, x_n)$  e si ordinano i dati

• Per  $t < x_{(1)}$ ,  $F_e(t) = 0$

$$F_e(x_{(1)}) = \frac{\#\{i \mid x_i \leq x_{(1)}\}}{n} = \frac{1}{n}$$

$F_e(t) = F_e(x_{(1)}) \quad \forall x < t < x_{(2)}$  e poi fa un salto di  $\frac{1}{n}$

$$\text{Se } x_{(3)} = x_{(4)} \Rightarrow F_e(x_{(3)}) = \frac{\#\{1, 2, 3, 4\}}{n} = \frac{4}{n}$$



## Esempio Funzione di ripartizione empirica

Consideriamo gli esiti di 10 lanci di un dado: 2, 3, 1, 2, 5, 5, 1, 5

Ordiniamo i dati: 1, 1, 2, 2, 3, 5, 5, 5

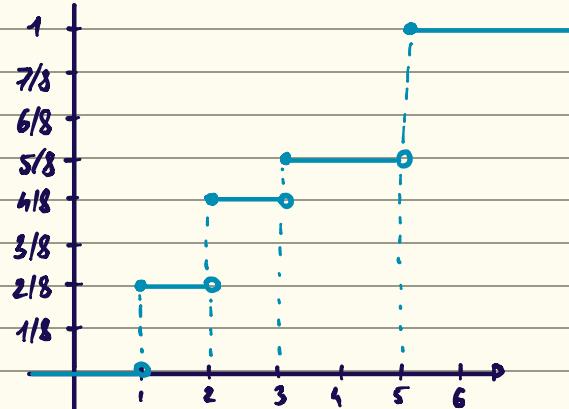
. Per  $t < 1$   $F_E(t) = 0 \quad \left. \begin{array}{l} \\ \end{array} \right\}$  salto di  
 $F_E(1) = 2/8 \quad \left. \begin{array}{l} \\ \end{array} \right\} 1/8 \cdot 2$

. Per  $1 \leq t < 2$   $F_E(t) = 2/8 \quad \left. \begin{array}{l} \\ \end{array} \right\}$  salto di  
 $F_E(2) = 4/8 \quad \left. \begin{array}{l} \\ \end{array} \right\} 1/8 \cdot 2$

. Per  $2 \leq t < 3$   $F_E(t) = 4/8 \quad \left. \begin{array}{l} \\ \end{array} \right\}$  salto di  
 $F_E(3) = 5/8 \quad \left. \begin{array}{l} \\ \end{array} \right\} 1/8$

. Per  $3 \leq t < 5$   $F_E(t) = 5/8 \quad \left. \begin{array}{l} \\ \end{array} \right\}$  salto di  
 $F_E(5) = 8/8 = 1 \quad \left. \begin{array}{l} \\ \end{array} \right\} 1/8 \cdot 3$   
 Salta il 4 perché non c'è

. Per  $t \geq 5$   $F_E(t) = 1$



K-simo percentile  
o  
**B-quantile**

Dato K numero naturale con  $0 < k < 100$ ,  $B = \frac{k}{100}$   
 allora il B-quantile è il dato  $x_i$  tale che:

- almeno il  $k\%$  dei dati è  $\leq x_i$
- almeno il  $(1-k)\%$  dei dati è  $\geq x_i$

In caso due dati soddisfino questa condizione, si prende la media.  
 La mediana è quindi il 50simo percentile ( $\delta \frac{50}{100} = \frac{1}{2}$  quantile)

**Notazione**  
**Percentili e Quantili:**

$$K = 25 \Rightarrow 25^{\circ} \text{ percentile} \leftrightarrow B = 25/100 \rightarrow \text{I Quartile}$$

$$K = 50 \Rightarrow 50^{\circ} \text{ percentile} \leftrightarrow B = 50/100 \rightarrow \text{mediana} \rightarrow \text{II Quartile}$$

$$K = 75 \Rightarrow 75^{\circ} \text{ percentile} \leftrightarrow B = 75/100 \rightarrow \text{III Quartile}$$

**Calcolare B-quantile**

$B \cdot n$  ( $n = \text{numero dati}$ ) può essere

- NON INTERO: arrotondando al numero successivo  $\Rightarrow x_{BN}$
- INTERO: prendo la media tra  $x_{BN}$  e  $x_{BN+1}$

## Esempio B-quantili:

Considero la tabella dei voti degli esami di 14 studenti:

A	25	Calcola I, II e III quartile			$n = 14$
B	23				
C	25				
D	25	20	22	22	
		23	25	25	
E	28	25	25	25	
F	28	26	26	26	
G	26	I quartile	II quartile	III quartile	
H	25				
I	26				
L	22				
M	2				
N	22				
O	20				
P	27				
		III quartile: $\frac{25}{100} \cdot 14 = [10, 50] = 11 \Rightarrow x_{11} = 26$			

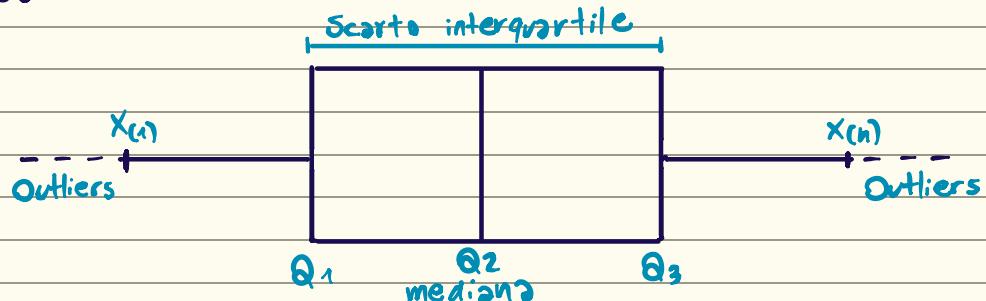
## Box Plot

Grafico che fornisce informazioni su: posizione, variabilità e forme di una distribuzione dati

- Una linea in corrispondenza della mediana, rappresenta centro della distribuzione
- Box che va dal I al III quantile, la cui altezza indica la variabilità
- Segmenti che indicano gli estremi dei dati

Valore Anomalo (Outlier)

Valore che differisce in modo significativo dalla grande maggioranza dei dati



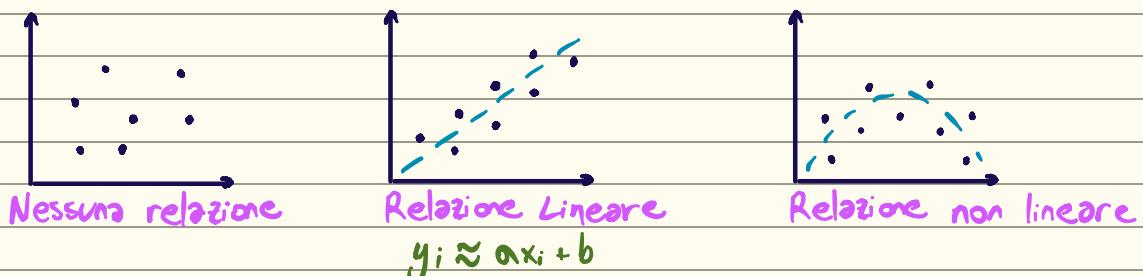
## DATI MULTIVARIATI

K-uple di vettori:

Capita di avere più dati misurati per lo stesso campione. In questo caso i dati sono K-uple di vettori:  $(x_i, y_i, z_i, \dots) \in \mathbb{R}^{k \times n}$   
es.  $(x_i, y_i)$  = altezza e peso iscritti: informatica

Diagramma a dispersione

Rappresenta con un punto ciascuna coppia di dati:



Problema della regressione

Studio dei dati mirato a misurare la dispersione, tramite:

- Componente qualitativa: quale curva approssima meglio i dati  $(x_i, y_i)$
- Componente quantitativa: quanto è buona l'approssimazione con tale curva

Covarianza

Di quanto si distaccano (differiscono) le coppie di dati

Covarianza campionaria

$$\text{cov}(x, y) = \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\text{o ss: } \text{cov}(x, y) = \text{cov}(y, x)$$

Covarianza empirica

$$\text{cov}_e(x, y) = \sum_{i=1}^n \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Relazione tra  $\text{cov}(x, y)$  e  $\text{cov}_e(x, y)$

$$\text{cov}(x, y) = \frac{n}{n-1} \cdot \text{cov}_e(x, y)$$

$$\text{cov}(x, y) \cdot (n-1) = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = n \cdot \text{cov}_e(x, y) \Rightarrow \text{cov}(x, y) = \frac{n}{n-1} \cdot \text{cov}_e(x, y)$$

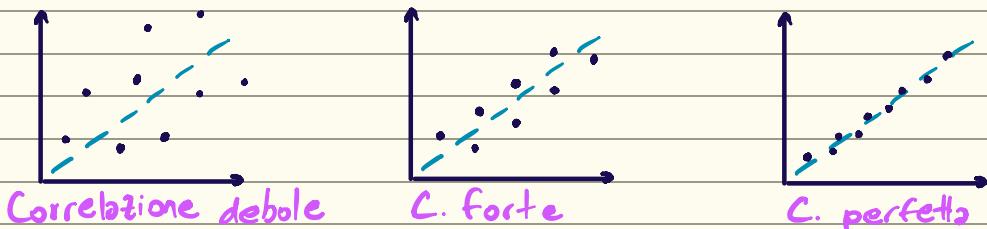
## Coefficiente di correlazione

Si può calcolare tramite  $r(x,y) \doteq \frac{\text{cov}(x,y)}{\sigma(x) \cdot \sigma(y)}$

OSS: coefficiente non cambia sostituendo con  $\text{cov}_e(\dots)$  e  $\sigma_e(\dots)$

La correlazione può essere:

- **Debole**:  $r$  vicino a 0
- **Forte**:  $r$  vicino a -1, oppure vicino a 1
- **Perfecto**:  $r=1$  o  $r=-1$



## Funzione retta

La funzione equivalente alla retta è  $y = q + mx$

- I punti della retta  $y = q + mx$  verificano perfettamente l'equazione.
- I punti della distribuzione, di coordinate  $(x_i, y_i)$ , non la verificano perfettamente a priori, perché non perfettamente allineati. Ci sarà uno scarto tra  $y_i$  e il punto della retta corrispondente a  $x_i$ , ovvero  $q + mx_i$ .
- L'idea è quindi scegliere  $q$  e  $m$  in modo tale da minimizzare gli scarti:  
 $\Rightarrow$  cerco  $\min_{q, m \in \mathbb{R}} \sum_{i=1}^n (y_i - q - mx_i)^2$

## Teorema

Se  $\sigma(x) \neq 0$  e  $\sigma(y) \neq 0$ , allora la  $\sum_{i=1}^n (y_i - q - mx_i)^2$  ha uno e un solo minimo, che si ottiene prendendo:

$$m^* = \frac{\text{cov}(x,y)}{\text{var}(x)} \quad \text{e} \quad q^* = \bar{y} - m^* \cdot \bar{x}$$

Inoltre  $\min_{q, m \in \mathbb{R}} \sum_{i=1}^n (y_i - q - mx_i)^2$  è uguale a  $\sum_{i=1}^n (y_i - \bar{y})^2 \cdot [1 - 2(\bar{x}, \bar{y})^2]$

## Retta di regressione

La retta  $y = m^*x + q^*$  con  $m^*$  e  $q^*$  come sopra si dice retta di regressione tra  $x$  e  $y$ , per cui:

- è la retta che meglio approssima i dati  $(x_i, y_i)$   $i = 1, \dots, n$
- più il coefficiente di correlazione  $r(x,y)$  è vicino a 1, migliore è l'approssimazione
- se  $r(x,y) = \pm 1$  allora  $m = 0$ , ovvero tutti i punti  $(x_i, y_i)$  sono sulla retta
- il segno del coefficiente angolare coincide con quello di  $r(x,y)$