

Corso di Statistica Facoltà di Economia

a.a. 2010-2011

francesco mola

Lezione n° 3

Sommario

- Indici statistici di posizione o locazione o di tendenza centrale
- Medie
- Moda
- Mediana
- Quantile

Perché le medie?

- Individuazione di una misura sintetica che tenda ad esprimere la tipicità della distribuzione
- Vengono definite misure di **tendenza centrale** perché “solitamente” si posizionano al centro della distribuzione (ma non sempre rappresentano esattamente il centro)
- Esistono molti indici di tendenza centrale alcuni che possono essere calcolati solo per fenomeni **quantitativi** altri **sia** per quantitativi che **qualitativi**.
- La scelta del tipo di media da utilizzare riguarda aspetti diversi, per lo più relativi alla conoscenza del fenomeno in studio (o la relazione che lega i dati o li ha generati), oppure ad una preliminare analisi grafica della distribuzione.
- Molti indici di tendenza centrale risentono infatti dei *dati anomali (outliers)*. Questi a prima vista possono sembrare errori nei dati, ma non sempre è così.

Medie secondo Chisini

X = carattere quantitativo
con

$$x_1 \leq x_2 \leq \dots \leq M \leq \dots \leq x_{n-1} \leq x_n$$

La quantità M è una media se...

$$f(x_1, x_2, \dots, x_n) = f(M, M, \dots, M)$$

Cioè:

considerare la funzione $f()$ sulle n modalità della X
o considerarla su n volte la costante M è equivalente.

Seguendo questo approccio, al variare della funzione $f()$
si ottengono differenti tipi di medie.

Media *aritmetica*

(semplice)

Consideriamo come criterio
la somma degli elementi

$$f(x_1, x_2, \dots, x_n) = \sum_{i=1}^n x_i \Rightarrow$$

Nella espressione del Chisini sostituisco a M il simbolo μ che indica la media aritmetica

$$f(x_1, x_2, \dots, x_n) = f(\mu, \mu, \dots, \mu) \Leftrightarrow$$

$$\sum_{i=1}^n x_i = \sum_{i=1}^n \mu \Leftrightarrow \sum_{i=1}^n x_i = n\mu \Rightarrow$$

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

lez 3_2010-2011

statistica-francesco mola

5

Un esempio su tredici studenti ...

Voto in Matematica (X)	Voto In Statistica (Y)
28	30
30	18
27	21
28	28
26	30
28	27
27	30
29	21
28	22
29	26
27	23
29	30
29	18
365	324

$$n=13$$

$$\sum x_i = 365$$

$$\sum y_i = 324$$

$$\mu_x = \frac{\sum x_i}{n} = \frac{365}{13} = 28,08$$

$$\mu_y = \frac{\sum y_i}{n} = \frac{324}{13} = 24,92$$

lez 3_2010-2011

statistica-francesco mola

6

media aritmetica (ponderata) per...
distribuzioni di **frequenze assolute**

Consideriamo una variabile X nella forma:

$$(x_i, n_i) \quad i=1,2,\dots,k$$

...la media aritmetica sarà:

$$\mu = \frac{\sum_{i=1}^k x_i n_i}{\sum_{i=1}^k n_i}$$

Ogni x_i ha un'influenza
dettata da n_i

lez 3_2010-2011

statistica-francesco mola

7

Un esempio su tredici studenti ... con distribuzione di frequenza

Voto in Matematica (X)	Frequenze assolute (n)
26	1
27	3
28	4
29	4
30	1
	13

$$n = n_1 + n_2 + n_3 + n_4 + n_5 = 13$$

$$\sum x_i n_i = 26 * 1 + 27 * 3 + 28 * 4 + 29 * 4 + 30 * 1 = 365$$

$$\mu_x = \frac{\sum x_i n_i}{n} = \frac{365}{13} = 28,08$$

lez 3_2010-2011

statistica-francesco mola

8

Un esempio su tredici studenti ... con distribuzione di frequenza

Voto in Statistica (Y)	Frequenze assolute (n)
18	2
21	2
22	1
23	1
26	1
27	1
28	1
30	4
	13

$$n = n_1 + n_2 + n_3 + n_4 + n_5 + n_6 + n_7 + n_8 = 13$$

$$\sum y_i n_i = 18 * 2 + 21 * 2 + 22 * 1 + \dots + 30 * 4 = 324$$

$$\mu_y = \frac{\sum y_i n_i}{n} = \frac{324}{13} = 24,92$$

lez 3_2010-2011 statistica-francesco mola

9

media aritmetica (ponderata) per...
distribuzioni di **frequenze relative**

Consideriamo una variabile X nella forma:

$$(x_i, f_i) \quad i=1,2,\dots,k$$

...la media aritmetica sarà:

$$\mu = \sum_{i=1}^k x_i f_i$$

Ogni x_i ha un'influenza
dettata da f_i

lez 3_2010-2011

statistica-francesco mola

10

Un esempio su tredici studenti ... con distribuzione di frequenze relative

Voto in Matematica (X)	Frequenze assolute (n)	Frequenze relative (f)
26	1	1/13
27	3	3/13
28	4	4/13
29	4	4/13
30	1	1/13
	13	1

$$\sum x_i f_i = 26 * (1/13) + 27 * (3/13) + 28 * (4/13) +$$

$$+ 29 * (4/13) + 30 * (1/13) =$$

$$= 2 + 6,230769 + 8,615385 + 8,923077 + 2,307692 = 28,08$$

$$\mu_x = \sum x_i f_i = 28,08$$

lez 3_2010-2011

statistica-francesco mola

11

Un esempio su tredici studenti ... con distribuzione di frequenze relative

Voto in Statistica (Y)	Frequenze assolute (n)	Frequenze relative (f)
18	2	2/13
21	2	2/13
22	1	1/13
23	1	1/13
26	1	1/13
27	1	1/13
28	1	1/13
30	4	4/13
	13	1

$$\sum y_i f_i = 18 * (2/13) + 21 * (2/13) +$$

$$+ 22 * (1/13) + \dots + 30 * (4/13) = 24,92 =$$

$$= \mu_y$$

lez 3_2010-2011

statistica-francesco mola

12

media aritmetica (semplice e ponderata) per...
distribuzioni **in classi**

Si considerano i valori centrali di ogni classe

$$c_i = \frac{x_{i-1} + x_i}{2}$$

...la media aritmetica sarà:

$$\mu = \frac{\sum_{i=1}^k c_i}{n}$$

$$\mu = \frac{\sum_{i=1}^k c_i f_i}{n}$$

Si ipotizza uniformità nelle classi!!!!

con alcuni esempi vediamo perché diciamo...
aritmetica

Dati i valori 1, 2 e 3, la media è

$$\mu = \frac{1 + 2 + 3}{3} = 2$$

Progressioni
aritmetiche di
ragione 1 con n
dispari

Dati i valori 1, 2, 3, 4 e 5, la media è

$$\mu = \frac{1 + 2 + 3 + 4 + 5}{5} = 3$$

La media in questo
caso è il VALORE
CENTRALE
della distribuzione

con alcuni esempi vediamo perché diciamo...
aritmetica

Dati i valori 3, 8 e 13, la media è

$$\mu = \frac{3 + 8 + 13}{3} = 8$$

Progressioni
aritmetiche di
ragione 5 con n
dispari

Dati i valori 3, 8, 13, 18 e 23, la media è

$$\mu = \frac{3 + 8 + 13 + 18 + 23}{5} = 13$$

La media in
questo caso è il
VALORE
CENTRALE
della
distribuzione

1^a Proprietà della media aritmetica
l'internalità o criterio di Cauchy

$$x_1 \leq \mu \leq x_n$$

la media è il baricentro della distribuzione

Dire che la media è il baricentro della distribuzione
equivale a dire che:

$$\sum_i (x_i - \mu) = 0$$

dove $(x_i - \mu) \quad \forall i=1,2,\dots,n$

Sono gli **scarti dalla media**

Un esempio su tredici studenti ...

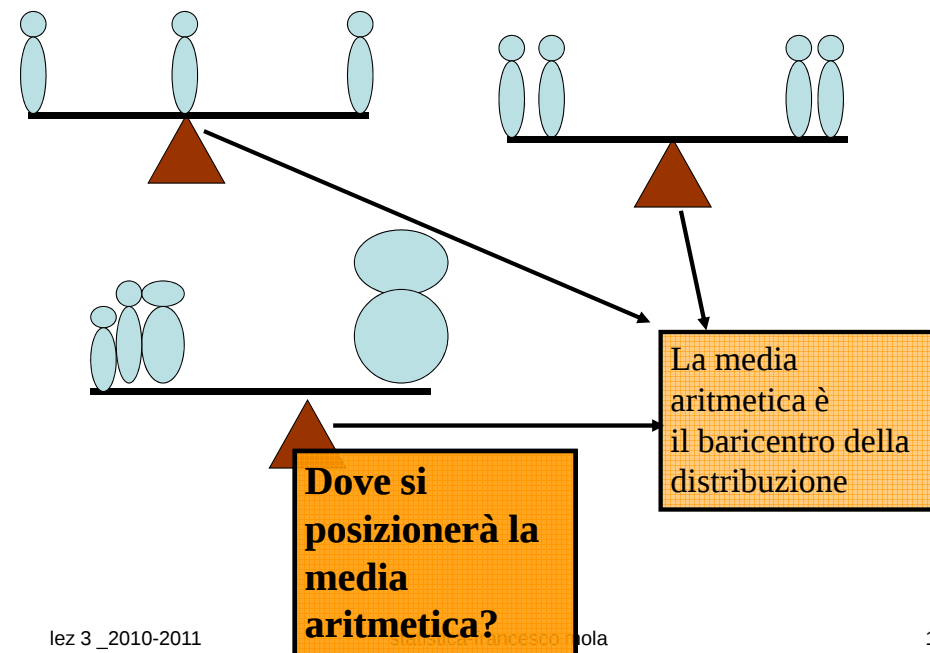
Voto in Matematica (X)	Voto In Statistica (Y)	X-mx	Y-m
28	30	-0,08	5,08
30	18	1,92	-6,92
27	21	-1,08	-3,92
28	28	-0,08	3,08
26	30	-2,08	5,08
28	27	-0,08	2,08
27	30	-1,08	5,08
29	21	0,92	-3,92
28	22	-0,08	-2,92
29	26	0,92	1,08
27	23	-1,08	-1,92
29	30	0,92	5,08
29	18	0,92	-6,92
365	324	0,00	0,00

lez 3 _2010-2011

statistica-francesco mola

17

...graficamente abbiamo....



lez 3 _2010-2011

statistica-francesco mola

18

2^a proprietà:

la media è l'unico valore che rende minima la somma degli scarti al quadrato

$$\sum (x_i - \mu)^2 = \min$$

lez 3 _2010-2011

statistica-francesco mola

19

Un esempio su tredici studenti ...

Voto in Matematica (X)	Voto In Statistica (Y)	X-mx	(X-mx)^2	Y-my	(Y-my)^2
28	30	-0,08	0,01	5,08	25,78
30	18	1,92	3,70	-6,92	47,93
27	21	-1,08	1,16	-3,92	15,39
28	28	-0,08	0,01	3,08	9,47
26	30	-2,08	4,31	5,08	25,78
28	27	-0,08	0,01	2,08	4,31
27	30	-1,08	1,16	5,08	25,78
29	21	0,92	0,85	-3,92	15,39
28	22	-0,08	0,01	-2,92	8,54
29	26	0,92	0,85	1,08	1,16
27	23	-1,08	1,16	-1,92	3,70
29	30	0,92	0,85	5,08	25,78
29	18	0,92	0,85	-6,92	47,93
365	324	0,00	14,92	0,00	256,92

lez 3 _2010-2011

statistica-francesco mola

Minimo

20

3^a Proprietà associativa

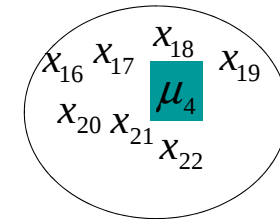
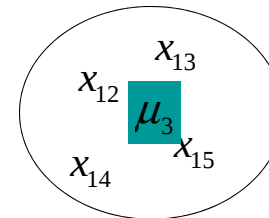
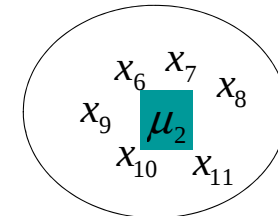
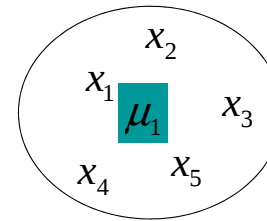
Consideriamo una variabile X ed H gruppi di numerosità non necessariamente uguale e con ognuno una propria media

Consideriamo quindi una variabile

$$X = x_1, x_2, \dots, x_n$$

raggruppata in H gruppi secondo l'esempio seguente:

Es: $H=4$; $n=22$



...allora abbiamo...

$$\mu = \frac{\sum_{h=1}^H \mu_h n_h}{\sum_{h=1}^H n_h}$$

Media ponderata
delle medie dei
singoli gruppi

cioè



.....

$$\frac{\mu_1 n_1 + \mu_2 n_2 + \dots + \mu_H n_H}{n_1 + n_2 + \dots + n_H} =$$

$$\mu_1 \frac{n_1}{n} + \mu_2 \frac{n_2}{n} + \dots + \mu_H \frac{n_H}{n}$$

con

$$n_1 + n_2 + \dots + n_h = n$$

Aggiungiamo come informazione l'anno di corso ai nostri tredici studenti

Studenti	Voto in Matematica (X)	Voto In Statistica (Y)	Anno di corso
	28	30	Primo
	30	18	Secondo
	27	21	Secondo
	28	28	Primo
	26	30	Secondo
	28	27	Terzo
	27	30	Secondo
	29	21	Primo
	28	22	Secondo
	29	26	Secondo
	27	23	Terzo
	29	30	Secondo
	29	18	Secondo
	365	324	

Cosa accade se vogliamo conoscere la media dei voti
in funzione dell'anno di iscrizione?

Per ciascun gruppo
calcoliamo la media $\mu_h = \frac{\sum_{i=1}^{n_h} x_i n_i}{n_h}$ per avere la media del gruppo h

La formula da utilizzare è $\mu = \frac{\sum_{h=1}^H \mu_h n_h}{\sum_{h=1}^H n_h}$ per avere la media generale

Numero di gruppi H = 3

Abbiamo quindi per i voti in **MATEMATICA**:

$$\mu_1 = \frac{\sum_{i=1}^3 x_i n_i}{3} = \frac{28*2 + 29*1}{3} = 28,333$$

Media del
primo anno

$$\mu_2 = \frac{\sum_{i=1}^8 x_i n_i}{8} = \frac{26*1 + 27*2 + 28*1 + 29*3 + 30*1}{8} = 28,125$$

Media del
secondo anno

$$\mu_3 = \frac{\sum_{i=1}^2 x_i n_i}{2} = \frac{27*1 + 28*1}{2} = 27,5$$

Media del
terzo anno

$$\mu = \frac{\sum_{h=1}^H \mu_h n_h}{n} = \frac{28,333*3 + 28,125*8 + 27,5*2}{13} = 28,08$$

Media generale

E' come se avessimo costruito tre distribuzioni in funzione della provenienza

Voto in Matematica (X)	Frequenze assolute (n)
28	2
29	1
	3

Primo anno

Voto in Matematica (X)	Frequenze assolute (n)
27	1
28	1
	2

Secondo anno

Voto in Matematica (X)	Frequenze assolute (n)
26	1
27	2
28	1
29	3
30	1
	8

Terzo anno

4^a Proprietà della media aritmetica linearità

Consideriamo una variabile X con media $M(X)=\mu$ e consideriamo la seguente combinazione lineare:

$$Y = \alpha \pm \beta X$$

Si può facilmente dimostrare che:

$$M(Y) = \alpha \pm \beta M(X) = \alpha \pm \beta \mu$$

Ancora sulla 4^a Proprietà della media aritmetica linearità

Consideriamo una variabile X con media $M(X)=\mu$ e consideriamo la seguente combinazione lineare:

$$Y = \alpha \pm X$$

Si può facilmente vedere che:

$$M(Y) = \alpha \pm M(X) = \alpha \pm \mu$$

Ancora sulla 4^a Proprietà della media aritmetica linearità

Consideriamo una variabile X con media $M(X)=\mu$ e consideriamo la seguente combinazione lineare:

$$Y = \beta X$$

Si può facilmente vedere che:

$$M(Y) = \beta M(X) = \beta \mu$$

Media *geometrica*

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^n x_i \Rightarrow$$

$$f(x_1, x_2, \dots, x_n) = f(G, G, \dots, G) \Leftrightarrow$$

$$\prod_{i=1}^n x_i = \prod_{i=1}^n G \Leftrightarrow \prod_{i=1}^n x_i = G^n \Leftrightarrow$$

Media *geometrica*

$$\left(\prod_{i=1}^n x_i\right)^{\frac{1}{n}} = G \Rightarrow G = \exp\left(\frac{1}{n} \sum \log x_i\right)$$

Consideriamo come criterio
Il prodotto degli elementi

E' geometrica perché è legata ad
una progressione geometrica

$$(x_1 q)(x_1 q^2)(x_1 q^3) \dots (x_1 q^{n-1})$$

dim.

$$\begin{aligned} G = \left(\prod x_i\right)^{\frac{1}{n}} &\Leftrightarrow \lg(G) = \lg\left[\left(\prod x_i\right)^{\frac{1}{n}}\right] = \\ &\frac{1}{n} \lg \prod x_i = \frac{1}{n} \sum \lg x_i \Rightarrow \\ \exp[\lg(G)] &= \exp\left[\frac{1}{n} \sum \lg x_i\right] \Rightarrow \\ G &= \exp\left[\frac{1}{n} \sum \lg x_i\right] \quad \text{C.V.D} \end{aligned}$$

media geometrica (ponderata) per...
distribuzioni di **frequenze assolute**

Consideriamo una variabile X nella forma:

$$(x_i, n_i) \quad i=1,2,\dots,k$$

...la media geometrica sarà:

$$G = \exp\left[\frac{1}{\sum n_i} \left(\sum \lg(x_i) n_i\right)\right]$$

Ogni x_i ha una influenza
dettata da n_i

media geometrica (ponderata) per...
distribuzioni di **frequenze relative**

Consideriamo una variabile X nella forma:

$$(x_i, f_i) \quad i=1,2,\dots,k$$

...la media geometrica sarà:

$$G = \exp\left(\sum f_i \lg x_i\right)$$

Ogni x_i ha una influenza
dettata da f_i

La moda

Per distribuzioni di frequenza assolute

X	n_i
x_1	n_1
x_2	n_2
•	•
x_i	n_i
•	•
x_k	n_k
	n

$$Mo = x_i : n_i = \max$$

La moda

Per distribuzioni di frequenza relative

X	$f_i = n_i/n$
x_1	f_1
x_2	f_2
•	•
x_i	f_i
•	•
x_k	f_k
	1

$$Mo = x_i : f_i = \max$$

La moda

Per distribuzioni in classi

Si parla in questo caso di classe modale

X	n_i	f_i
$x_0 - x_1$	n_1	f_1
$x_1 - x_2$	n_2	f_2
•	•	•
$x_{i-1} - x_i$	n_i	f_i
•	•	•
$x_{k-1} - x_k$	n_k	f_k

$$Mo = x_{i-1} - x_i : n_i = \max$$

$$Mo = x_{i-1} - x_i : f_i = \max$$

Pregi e difetti della moda

- Pregi
 - Semplice da individuare
- Difetti
 - Troppo influenzata da singole osservazioni
 - Non è monotona
 - Per distribuzioni in classi, va analizzata l'ampiezza delle classi e la uniformità della distribuzione nelle stesse

Individuiamo la moda per i nostri studenti

Voto in Matematica (X)	Frequenze assolute (n)	Frequenze relative (f)
26	1	1/13
27	3	3/13
28	4	4/13
29	4	4/13
30	1	1/13
	13	1

Voto in Statistica (Y)	Frequenze assolute (n)	Frequenze relative (f)
18	2	2/13
21	2	2/13
22	1	1/13
23	1	1/13
26	1	1/13
27	1	1/13
28	1	1/13
30	4	4/13
	13	1

In matematica abbiamo 2 mode
28 e 29

In Statistica abbiamo 1 moda **30**

Alla luce dei risultati ottenuti riconsiderare le valutazioni su "pregi e difetti della moda"!!

La *mediana*

È la modalità o la classe della variabile statistica che occupa la posizione centrale e che ha nel punto mediano la funzione di ripartizione empirica pari a 0.5

$$F(Me) = \frac{1}{2} = 0.5$$

La mediana per variabili discrete

$$X = x_1 \leq x_2 \leq x_3 \leq \dots \leq x_n$$

$$Me = \begin{cases} \frac{X_{(n/2)} + X_{(n/2)+1}}{2} & n = \text{pari} \\ X_{\left(\frac{n+1}{2}\right)} & n = \text{dispari} \end{cases}$$

Esempio

1, 2, **3**, 4, 5, 5

$$Me = \frac{x_3 + x_4}{2} = \frac{3 + 4}{2} = 3.5$$

Esempio

1, 2, 3, 4, 5

$$Me = x_{\frac{6}{2}} = x_3 = 3$$

La mediana per variabili discrete

Per distribuzioni di frequenza

X	f_i
x_1	f_1
x_2	f_2
•	•
x_i	f_i
•	•
x_k	f_k
	1

$$Me = X : F(Me) = 0.5$$

esempio

X	f_i	F_i
1	0.2	0.2
3	0.1	0.3
5	0.4	0.7
6	0.3	1
	1	

$$Me = X : F(Me) = 0.5$$

$$Me = 5$$

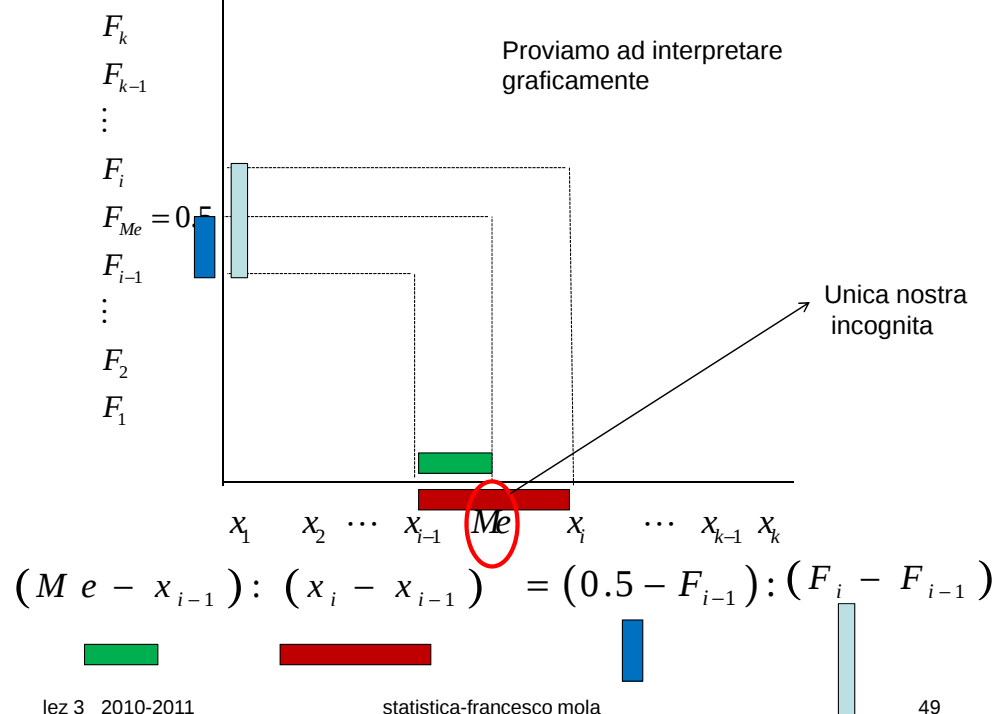
Viene considerata come mediana il valore della x tale che la F arrivi a 0.5

Spesso si usa considerare come valore mediano la semisomma delle x comprendenti la mediana. Nel nostro caso $(3+5)/2=4$

La mediana per variabili continue

X	f_i	F_i
x_0-x_1	f_1	F_1
x_1-x_2	f_2	F_2
•	•	•
$x_{i-1}-x_i$	f_i	F_i
•	•	•
$x_{k-1}-x_k$	f_k	1





$$(Me - x_{i-1}) : (x_i - x_{i-1}) = (0.5 - F_{i-1}) : (F_i - F_{i-1})$$

$$Me \cong x_{i-1} + (x_i - x_{i-1}) \frac{0.5 - F_{i-1}}{F_i - F_{i-1}}$$

Relativamente ai 13 studenti abbiamo che la mediana in **MATEMATICA** è **28** mentre in **STATISTICA** è **26**

Confrontando i risultati con media e moda notiamo sensibili differenze: perché?

La moda andrebbe utilizzata solo per distribuzioni monotone

La media è il baricentro di una distribuzione e non il centro (come la mediana) risentendo quindi del valore effettivo delle modalità.

Cosa accade se nel calcolo della Mediana non ordiniamo i dati?

I quantili

Decili: dividono la distribuzione in 10 parti uguali

1° decile => F(X) = 0.1

2° decile => F(X) = 0.2

.....

9° decile => F(X) = 0.9

Percentili: dividono la distribuzione in 100 parti uguali

Quartili: dividono la distribuzione in 4 parti uguali

Nota che il secondo quartile, il quinto decile, il 50 percentile corrispondono alla mediana!