COME SI STUDIANO

MOLTISSIME MISURE ?

Pb1 In una indagine sanitaria si riscontrano i seguenti dati, relativamente ad un certo tipo di infezione batterica:

Area	immuni	a rischio	infetti
Nord	46	12	25
Centro	12	2	31
Sud	21	4	26

1. In quale area c'è la maggior percentuale di infetti?

Risp. Al Nord i dati sono 46+12+25=83, 25/83≈0.30
Al Centro i dati sono 12+2+31=45 31/45≈0.69
Al Sud i dati sono 21+4+26=51 26/51≈0.51

L'area in cui la percentuale di infetti è maggiore é il Centro

Area	immuni	a rischio	infetti
Nord	46	12	25
Centro	12	2	31
Sud	21	4	26

2. Per quel che riguarda l'immunità, i soggetti del Centro sono più simili a quelli del Nord o a quelli del Sud?

Risp. Calcoliamo "la distanza" tra i soggetti immuni

$$|I_N - I_C| = 46-12 = 34 \quad |I_C - I_S| = 21-12 = 9:$$

I soggetti del Centro sono più simili a quelli del Sud

Area	immuni	a rischio	infetti
Nord	46	12	25
Centro	12	2	31
Sud	21	4	26

3. Se si tiene conto di tutte le caratteristiche (immunità, rischio, infez) quali soggetti sono più diversi tra loro?

Risp. Se N=(46,12,25) C=(12,2,31) S=(21,4,26)

$|NC|=[(46-12)^2+(12-2)^2+(25-31)^2]^{1/2}\approx 36$

$|NS|=[(46-21)^2+(12-4)^2+(25-26)^2]^{1/2}\approx 26$

$|CS|=[(21-12)^2+(4-2)^2+(26-31)^2]^{1/2}\approx 10.5$:

i più distanti (diversi) sono i soggetti del nord e Centro

RAPPRESENTARE LE MISURE

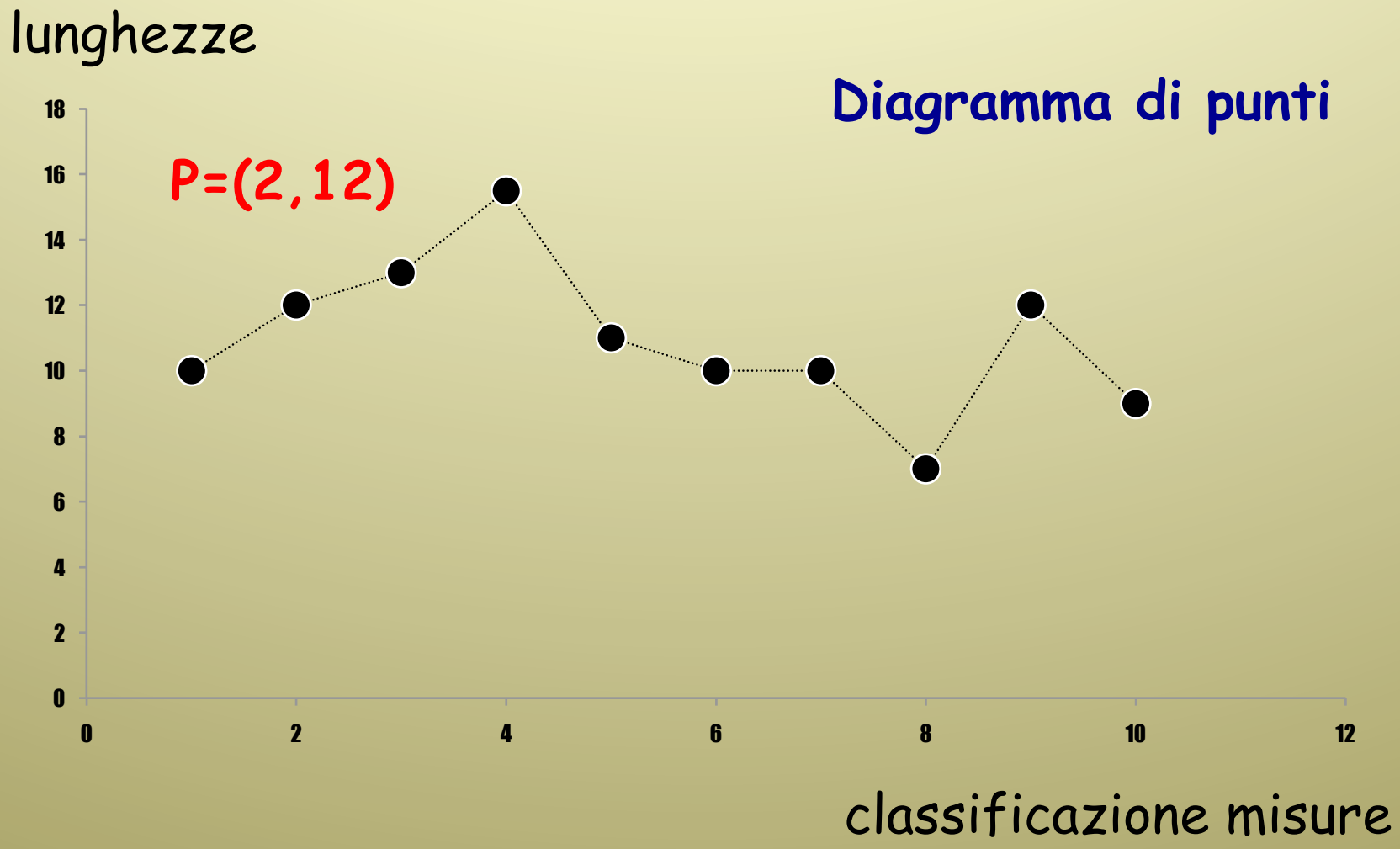
Lista di 10 misure di lunghezza in cm:

(10, 12, 13, 15.5, 11, 10, 10, 7, 12, 9)

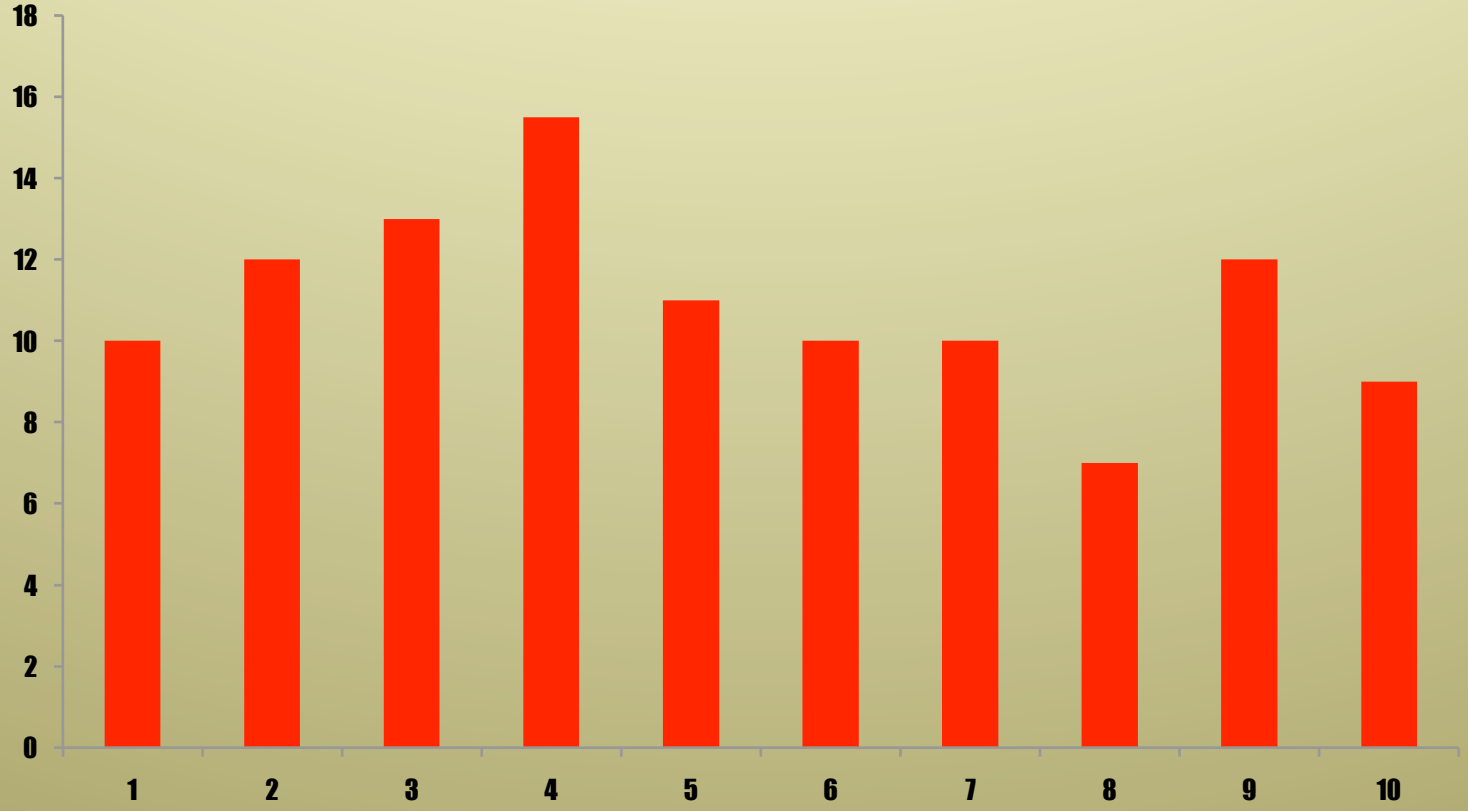
(1 2 3 4 5 6 7 8 9 10)

(Posizione nella lista di ogni misura)

Rappresentazione di misure nel piano cartesiano



ISTOGRAMMA



FREQUENZA delle MISURE = numero di volte che compare la stessa misura

Lista di n=10 misure di lunghezza in cm:

(10, 12, 13, 15.5, 11, 10, 10, 7, 12, 9)

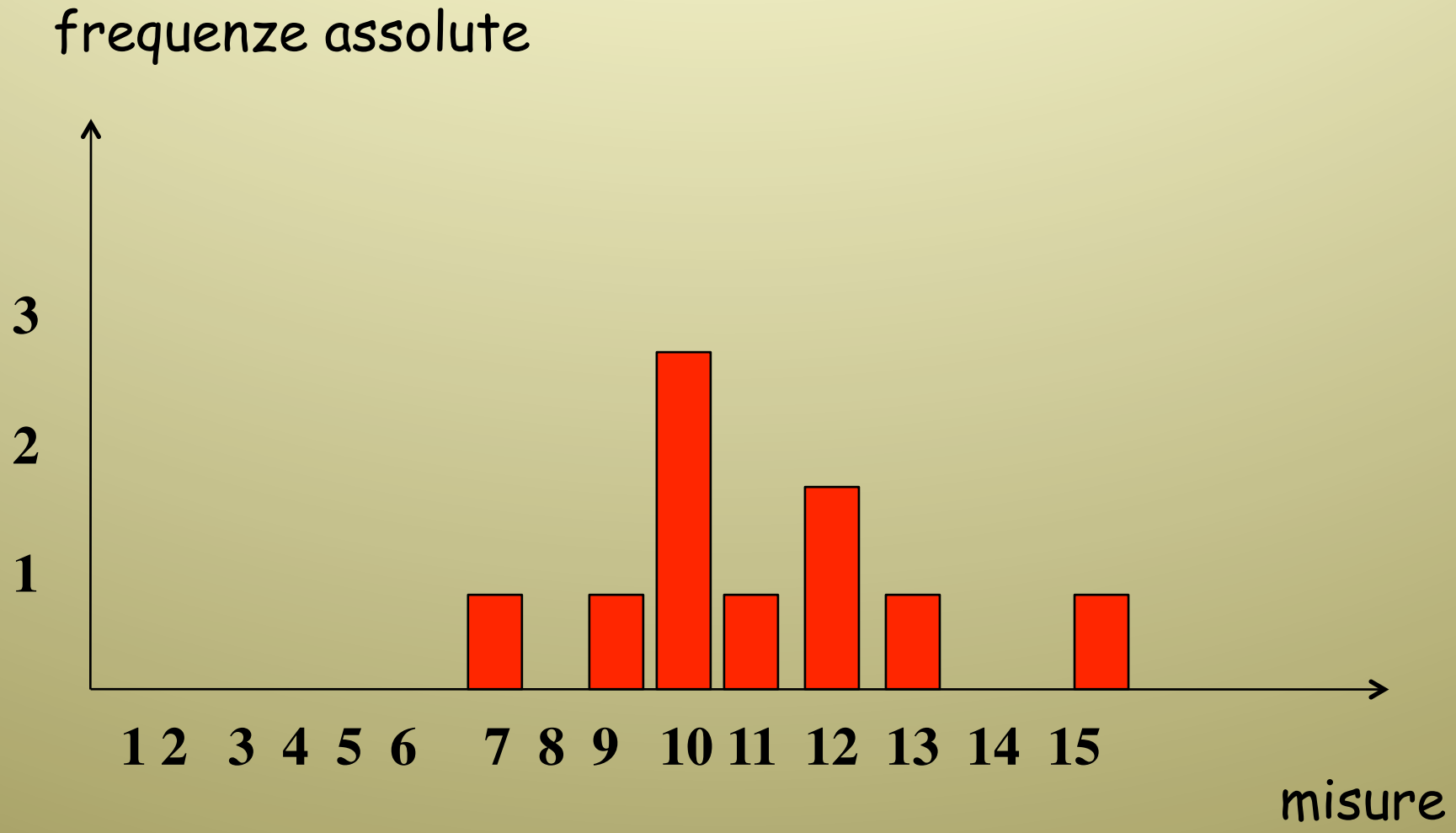
F(10)=3 F(12)=2

F(13)=F(15.5)=F(11)=F(7)=F(9)=1

FREQUENZE ASSOLUTE dei DATI

(N.B. 3+2+5=10=n)

Istogramma delle frequenze



FREQUENZE RELATIVE (f=F/n)

Lista di n=10 misure di lunghezza in cm:

(10, 12, 13, 15.5, 11, 10, 10, 7, 12, 9)

f(10)=3/10=0.3 f(12)=2/10=0.2

f(13)=f(15.5)=f(11)=f(7)=f(9)=1/10=0.1

(N.B. 0.3+0.2+5(0.1)=1)

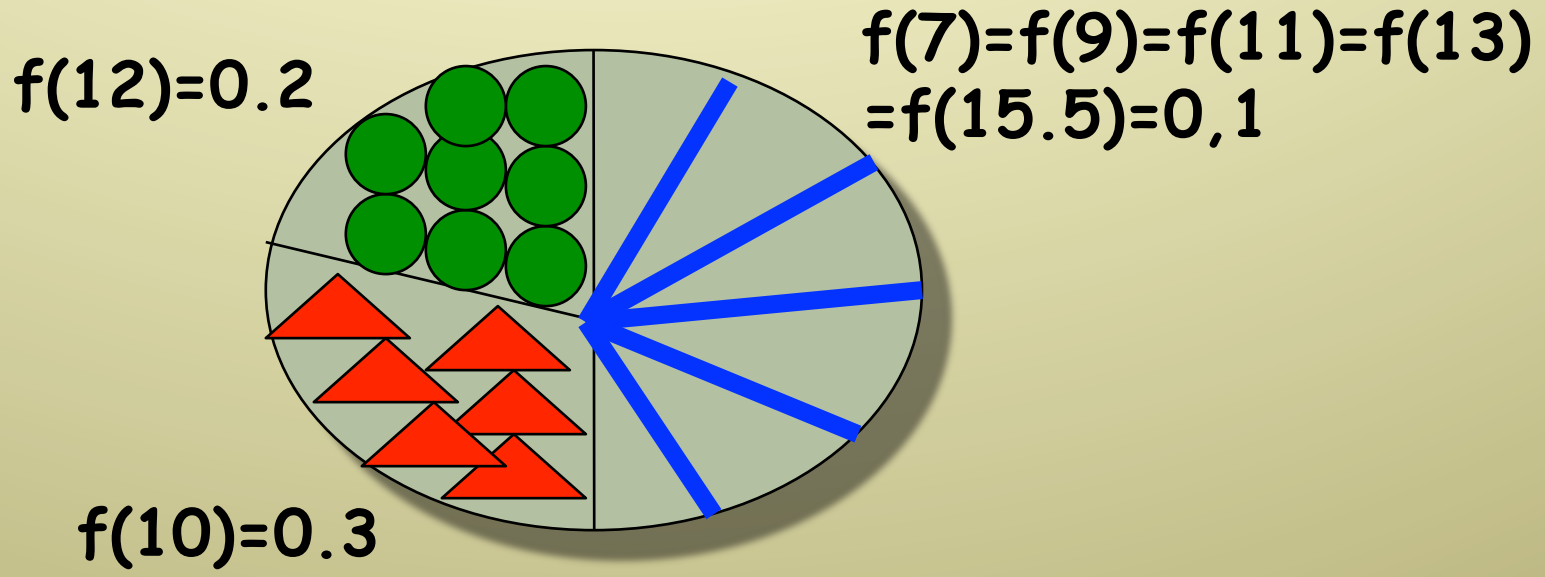
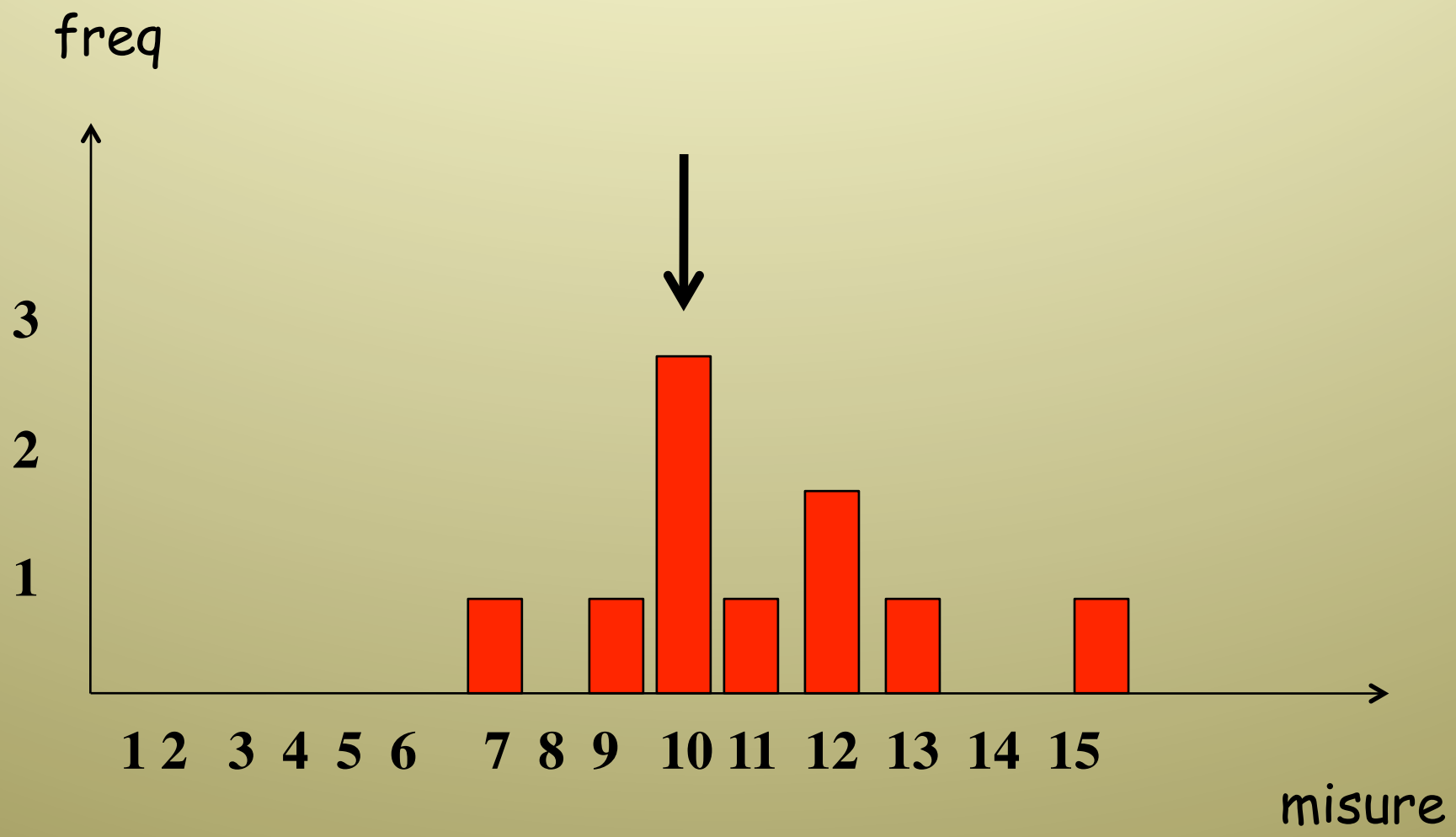


Diagramma a torta delle freq. relative

Poi si calcolano gli "indici di tendenza centrale" che riassumono le caratteristiche delle misure

(- indice=numero
 - tendenza centrale= che si trova quasi al centro)

La moda (il dato più frequente)



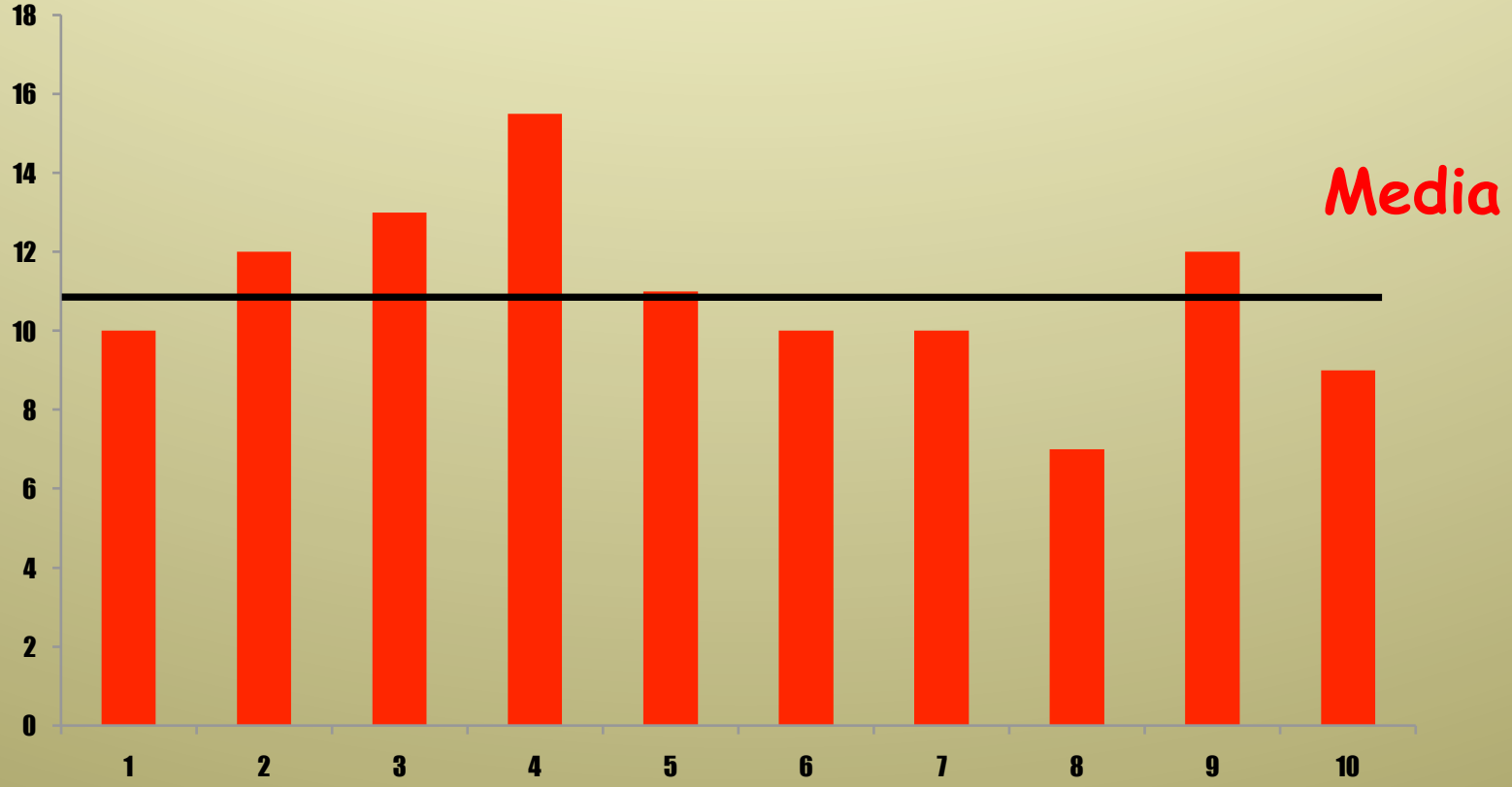
La media aritmetica

Media

Misure:	7	9	10	11	12	13	15.5
Frequenza:	1	1	3	1	2	1	1

$$\text{Media} = \frac{(1)\,[7+\ 9+\ 11+13\ +\ 15.5] + (3)10 + (2)12}{10}$$

$$= 10.95$$


18
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
Media

Per calcolare la mediana

le misure

-si ordinano in ordine crescente

-se sono in numero dispari, la mediana è la misura che si trova "al centro"

-se sono in numero pari, la mediana è la semisomma delle due centrali

La mediana

Misure:	7	9	10	11	12	13	15.5
Frequenza:	1	1	3	1	2	1	1

(10 misure, al centro 10 e 11)

Mediana =(10+11)/2=10.5

La mediana

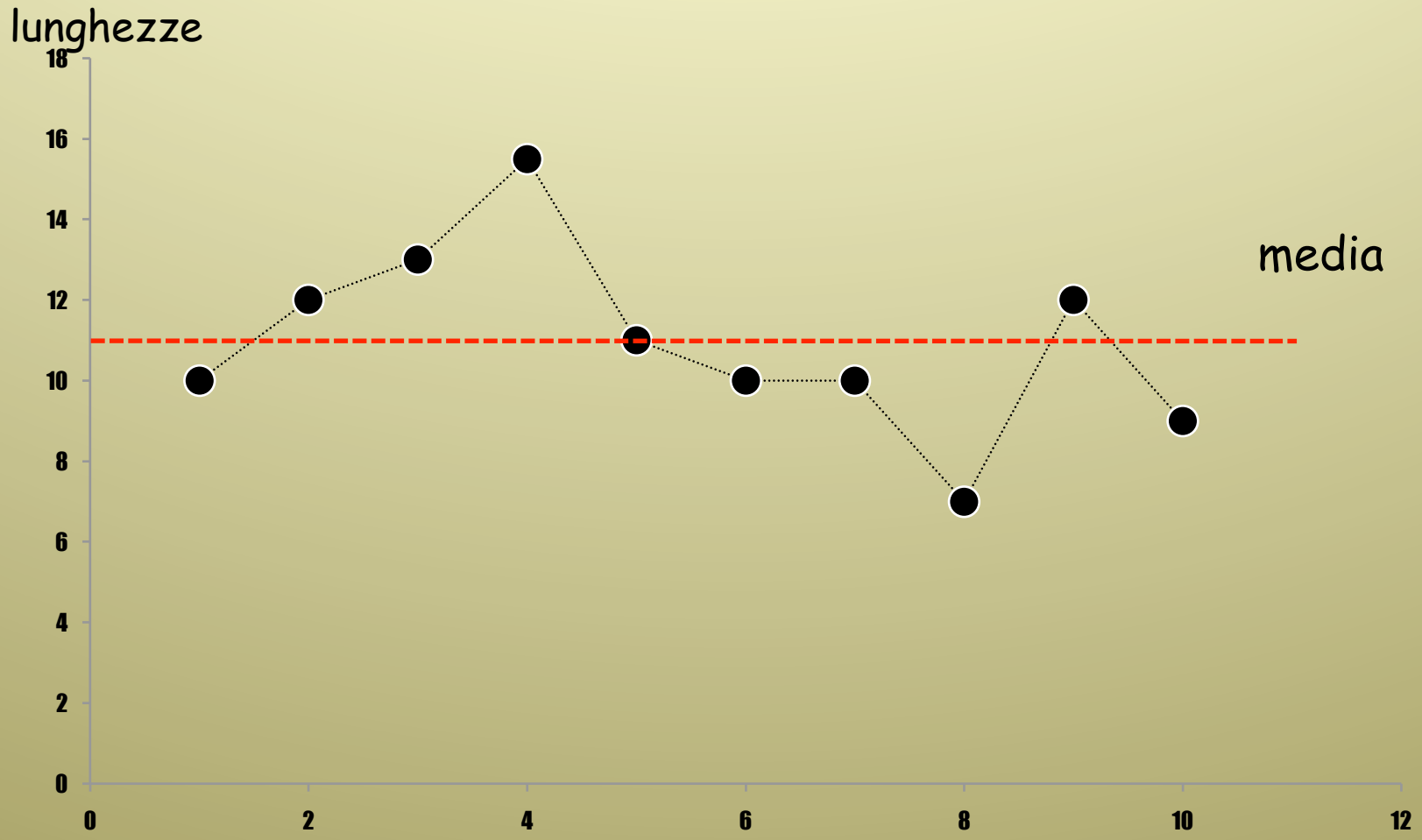
Misure:	1.2	9	10	11	12	13	45
Frequenza:	1	1	3	1	2	1	1

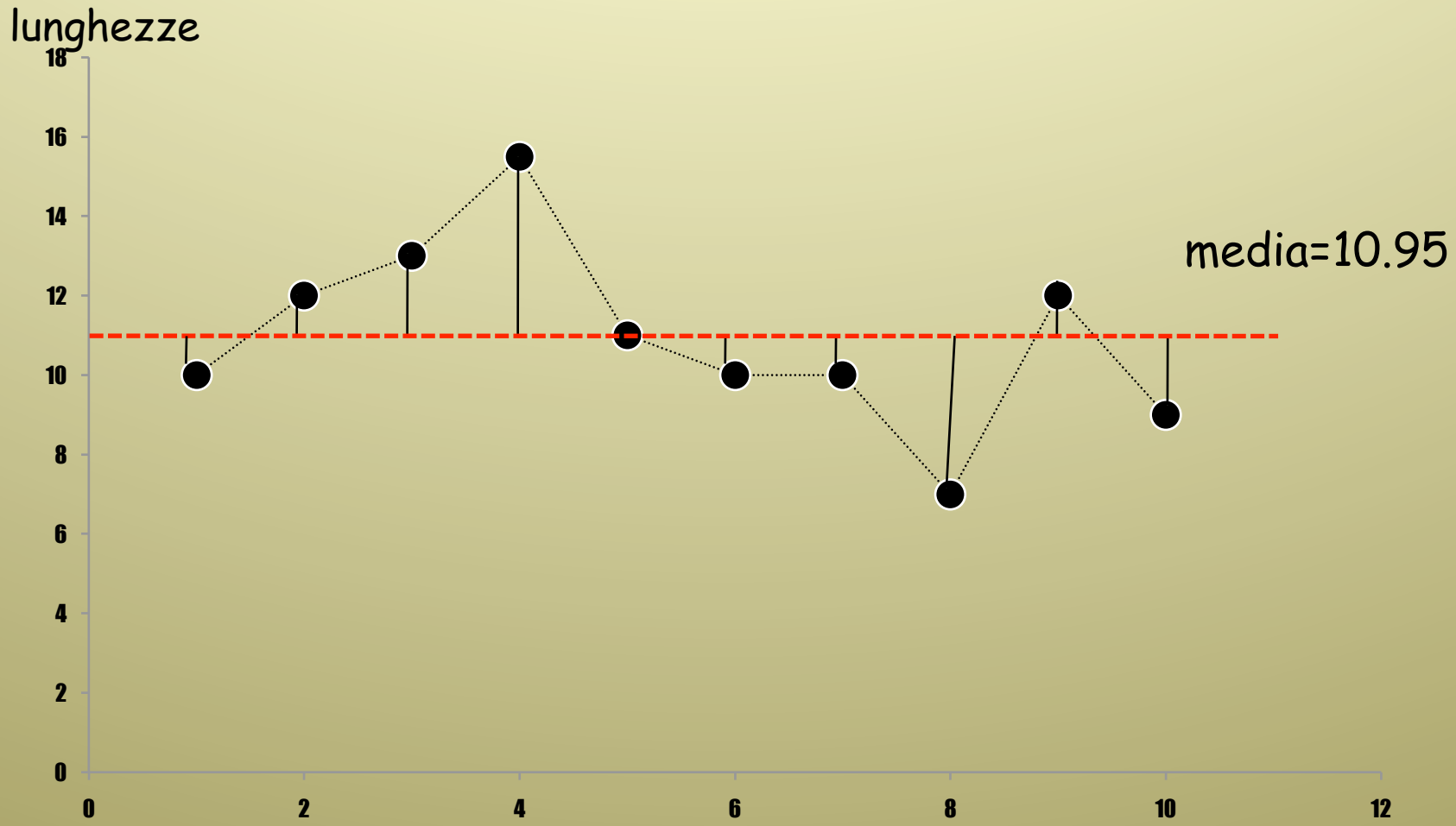
(10 misure, al centro 10 e 11)

Mediana =(10+11)/2=10.5 Media=13.32

Se i dati più piccoli e più grandi sono molto diversi dagli altri, la mediana è una misura di tendenza centrale migliore della media

Pb. La media è una buona misura di tendenza centrale?




lunghezze
18
16
14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
media=10.95

Le misure

10, 12, 13, 15.5, 11, 10, 10, 7, 12, 9

La media 10.95

Distanza di ogni misura dalla media

|10-10.95|=0.95
|12-10.95|=1.05
|13-10.95|=2.05
|15.5-10.95|=4.55
|11-10.95|=0.05
|7-10.95|=3.95
|9.10.95|=1.95

La distanza media dalla media (**scarto medio**) è

$$\frac{(3)0.95+(2)1.05+2.05+4.55+0.05+3.95+1.95}{10} = 1.75$$

(In media i dati sono distanti 1.75 dalla media)

A causa della definizione (c'e' il modulo) lo scarto medio non e' usato. Si preferisce **LA VARIANZA**

La varianza

$$V = \frac{(3)(0.95)^2+(2)(1.05)^2+(2.05)^2+(4.55)^2+(0.05)^2+(3.95)^2+(1.95)^2}{10}$$

$$\approx 4.92$$

La deviazione standard

$$\sigma = (V)^{1/2} \approx 2.22$$

Conclusione:

le misure

10, 12, 13, 15.5, 11, 10, 10, 7, 12, 9

Possono essere stimate dalla media m= 10.95 e sono, in media, comprese nell'intervallo

$[\sigma-m,\sigma+m] = [10.95-2.22, 10.95+2.22] =$

$= [8.73, 13.17]$

Le informazioni che abbiamo ricavato costituiscono

UNA STATISTICA

DESCRITTIVA DEI DATI

GRANDEZZE DIPENDENTI TRA LORO

Nei fenomeni naturali è di grande interesse capire "le cause" del comportamento delle grandezze.

Ad es. nello studio dello sviluppo di una pianta, ci si può chiedere:

- la crescita dello stelo e delle foglie, dipende dalla quantità di luce che la pianta riceve?
- La crescita della pianta dipende dalla quantità di acqua che viene data alla pianta?
- La crescita della pianta dipende dalla quantità di concime che viene somministrato?

COME SI FA A CAPIRE SE due (o piu') elementi, grandezze, . . . riguardanti un certo fenomeno sono dipendenti tra loro?

Si puo' iniziare chiedendosi se quando una delle due varia, varia anche l'altra.
In particolare se una delle grandezze aumenta, l'altra corrispondentemente aumenta o diminuisce?

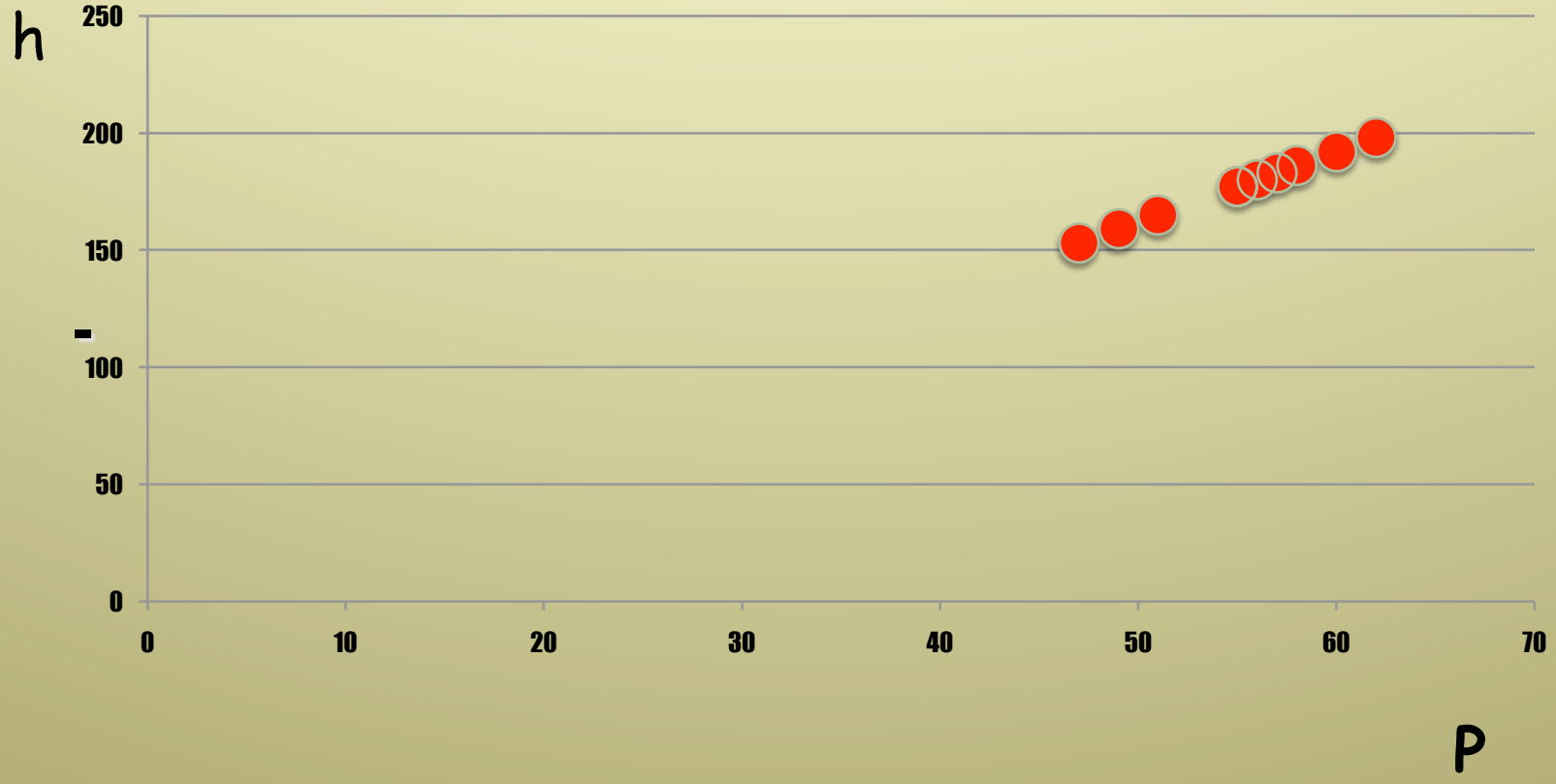
Ad es. vorremmo sapere se, nello sviluppo degli adolescenti, vi sia una relazione tra peso corporeo ed altezza

Il primo passo e' quello di effettuare delle misure.

Campione di 10 adolescenti femmine

h(cm):	153	159	177	186	192	177	198	183	180	165
P(kg):	47	49	55	58	60	55	62	57	56	51

RAPPRESENTAZIONE delle MISURE (come diagramma di punti)



Cosa possiamo dire?

I punti di coordinate (P,h) sembrano allineati, cioè appartenenti ad una retta, quindi le coordinate non possono essere arbitrarie, ma devono obbedire alla relazione

$$P \longrightarrow h(P)=mP+c$$

(m e c costanti opportune)

(h dipende linearmente da P:
h e' un multiplo di P a meno della costante c)

QUALE VALORE PER m e c?

Le misure sono

h(cm): **153 159 177 186 192 177 198 183 180 165**

P(kg): **47 49 55 58 60 55 62 57 56 51**

Consideriamo i punti **A=(47,153)** e **B=(60, 192)**.
Se appartengono alla retta h=mP+c deve essere

$$153=47m+c \quad e \quad 192=60m+c$$

Quindi $c=192-60m$ e $153=47m+(192-60m)=192-13m$
cioè

$$\mathbf{39 = 13m} \quad \text{e quindi} \quad \mathbf{m=3} \text{ e } \mathbf{c=192-180=12}$$

Quindi la relazione e'

$$h(P)=h=3P+12$$

tutte le altre misure obbediscono a questa relazione (ad es. Q=(58, 186) appartiene alla retta perchè 58x3 + 12=186, anche K=(56,180) appartiene alla retta perché 56x3+12= 180 ecc.)

Il fatto che m=3>0 ci dice che se P aumenta, aumenta anche h.

(Se fosse m<0 all'aumento di P corrisponderebbe una diminuzione di h)

CONSEGUENZA IMPORTANTE:

Se conosciamo la relazione tra le grandezze, possiamo "FARE PREVISIONI" (senza misurare direttamente).

In particolare se, vale la legge $h=3P+12$ e se il peso di un'adolescente e' $P=42Kg$, l'altezza deve essere

$$h(P)=3x42+12=138cm$$

Mentre se il peso e' $P=45.5Kg$, l'altezza deve essere $h(P)=3x45.5+12=148.5cm$

ecc.

La scoperta che due grandezze sono dipendenti tra loro in modo lineare non dice solo che quando una delle due grandezze aumenta, aumenta (o diminuisce) anche l'altra.

Ci dice anche che una delle due grandezze è un multiplo dell'altra a meno di una costante: quindi l'informazione è molto precisa.

Negli esperimenti reali è però piuttosto raro che due grandezze siano linearmente dipendenti tra loro.

Può capitare che i punti siano quasi "allineati"

La relazione non e' evidente
(e se esistesse potrebbe essere molto complicata):
che fare per sapere se all'aumentare di P
aumenta anche h (e quanto)?

Si studia un indice detto COVARIANZA

Def. Si chiama covarianza tra due classi di N dati (x,y) che hanno medie M_x e M_y il numero

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - M_x)(y_i - M_y)$$

COVARIANZA

h(cm):	160	160	171	165	168	162	168	168	162	169
P(Kg) :	47	49	55	58	60	55	62	57	56	51

$M_h = 163.3$ $M_P = 55$

$$\sigma_{hP} = \frac{1}{10}[(160-163.3)(47-55)+\ldots+(169-163.3)(51-55)] = 8.96 \approx 9$$

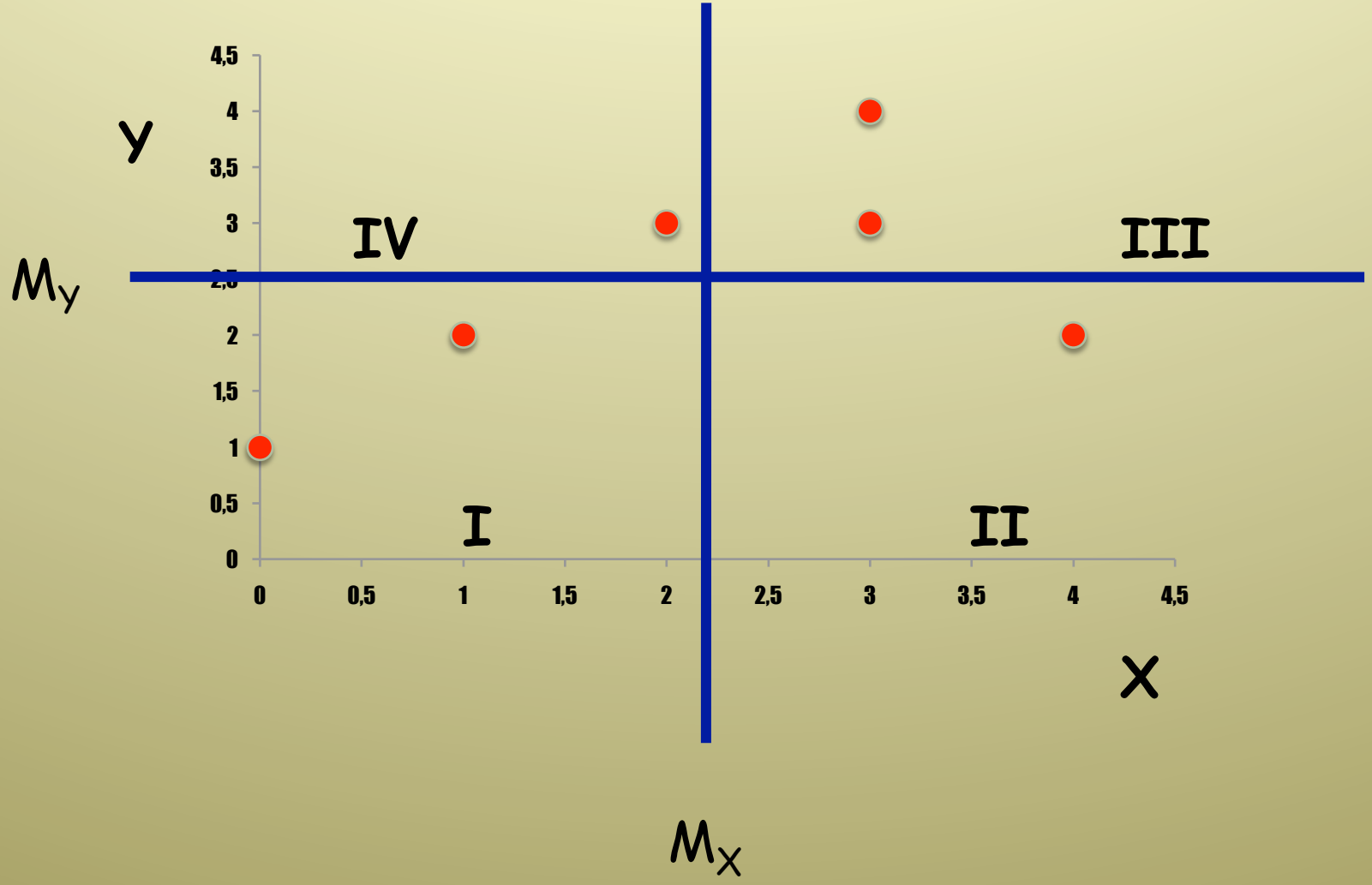
COVARIANZA POSITIVA MA NON MOLTO GRANDE, CHE SIGNIFICA?

Che cosa e' la covarianza?

Campione di 6 dati

X:	1	2	3	0	4	3	$M_X=13/6$
Y:	2	3	3	1	2	4	$M_Y=5/2$

RAPPRESENTAZIONE DATI E MEDIE



La maggior parte dei dati si trova nel I e III quadrante

suggerendo che all'aumentare di X anche Y aumenta.

La covarianza vale

$$\sigma_{XY} = \frac{3}{4} > 0$$

(Se i dati si trovano in gran parte nel II e IV quadrate, la covarianza e' negativa)

La covarianza ha la proprieta'

$$-\sigma_X \sigma_Y \leq \sigma_{XY} \leq \sigma_X \sigma_Y \qquad (*)$$

Se definiamo $\rho = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$ **(Coeff. di correlazione)**

La (*) e' equivalente a

$$-1 \leq \rho \leq 1$$

Per il campione di 6 dati

X:	**1**	**2**	**3**	**0**	**4**	**3**	$M_X = 13/6$
Y:	**2**	**3**	**3**	**1**	**2**	**4**	$M_Y = 5/2$

$$\sigma_{XY} = \frac{3}{4} \quad \sigma_X \approx 1.34 \quad \sigma_Y \approx 0.96 \quad \sigma_X \sigma_Y \approx 1.29$$

$$\Rightarrow \quad \rho \approx \frac{(0.75)}{1.29} \approx 0.58 > 0$$

$\rho = 1$

Massima correlazione positiva= allineamento dei dati su retta con m>0 (I e III quadrante)

$\rho = -1$

Massima correlazione negativa= allineamento dei dati su retta con m<0 (II e IV quadrante)

Pozze di acqua

Dafnia

La numerosità delle Dafnie nelle pozze d'acqua varia molto.

Quali fattori ambientali determinano queste variazioni?

Tra i più accreditati fattori ci sono la quantità d'acqua contenuta nelle pozze e la temperatura dell'acqua.

Studiamo le eventuali dipendenze

N	**3845**	**8673**	**18587**	**11499**	**8502**	**6827**	**9764**	**9315**	**4511**
V	**37,2**	**94**	**112,1**	**112,1**	**137,9**	**141,8**	**126**	**119,5**	**113**
T	**18**	**9,8**	**8,4**	**11**	**7,2**	**6**	**7,2**	**12**	**10**

N = numerosità di Dafnie, V=volume d'acqua della pozza in litri, T= temperatura dell'acqua in gradi centigradi

MEDIE e DEVIAZIONI STANDARD:

$M_N \approx 9058$ $M_V = 110,4$ $M_T \approx 9,96$

$\sigma_N \approx 4022$ $\sigma_V \approx 27,8$ $\sigma_T \approx 3,39$

$\sigma_N\sigma_V \approx 111811.6$ $\sigma_N\sigma_T \approx 13634,58$

COVARIANZE E CORRELAZIONI

$\sigma_{VN} \approx 42493,92$ $\sigma_{TV} \approx -5046,04$

$\rho_{VN} \approx 42493,92/111811,6 \approx 0,38 > 0$

$\rho_{TN} \approx -5046,04/136345,8 \approx -0,37 < 0$

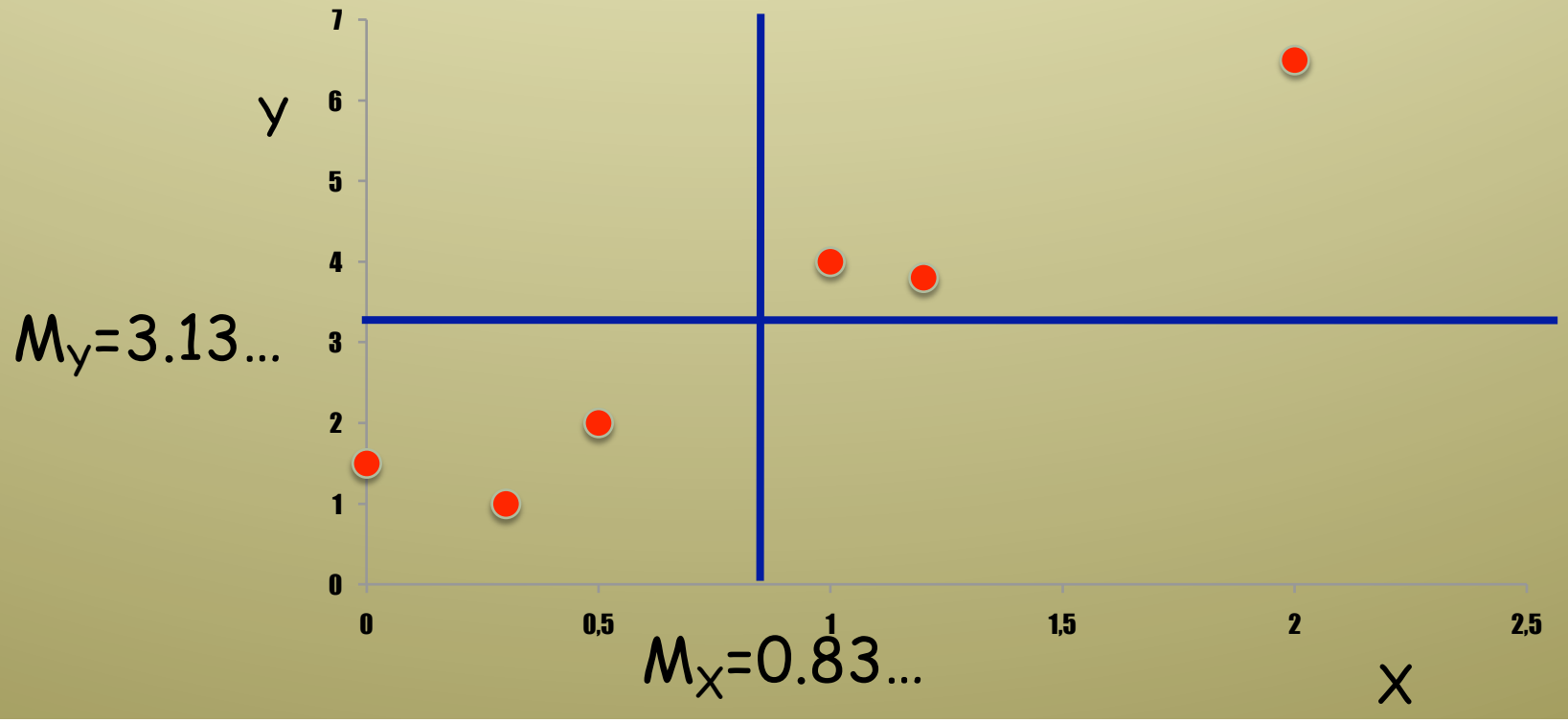
CONCLUSIONE: la relazione positiva tra numerosità e volume d'acqua supera quella (negativa) tra numerosità e temperatura. Questo vuol dire che all'aumentare del volume di acqua aumenta anche la numerosità della popolazione, mentre al diminuire della temperatura aumenta la numerosità.

ATTENZIONE: l'esempio precedente è solo indicativo di un procedimento per stabilire eventuali relazioni tra grandezze che variano.

Un campione di solo 9 elementi è infatti troppo piccolo per fornire indicazioni attendibili della realtà osservata.

ES: campione 6 dati

X:	1	0	1.2	0.3	2	0.5
Y:	4	1.5	3.8	1	6.5	2



M_y=3.13…

M_X=0.83…

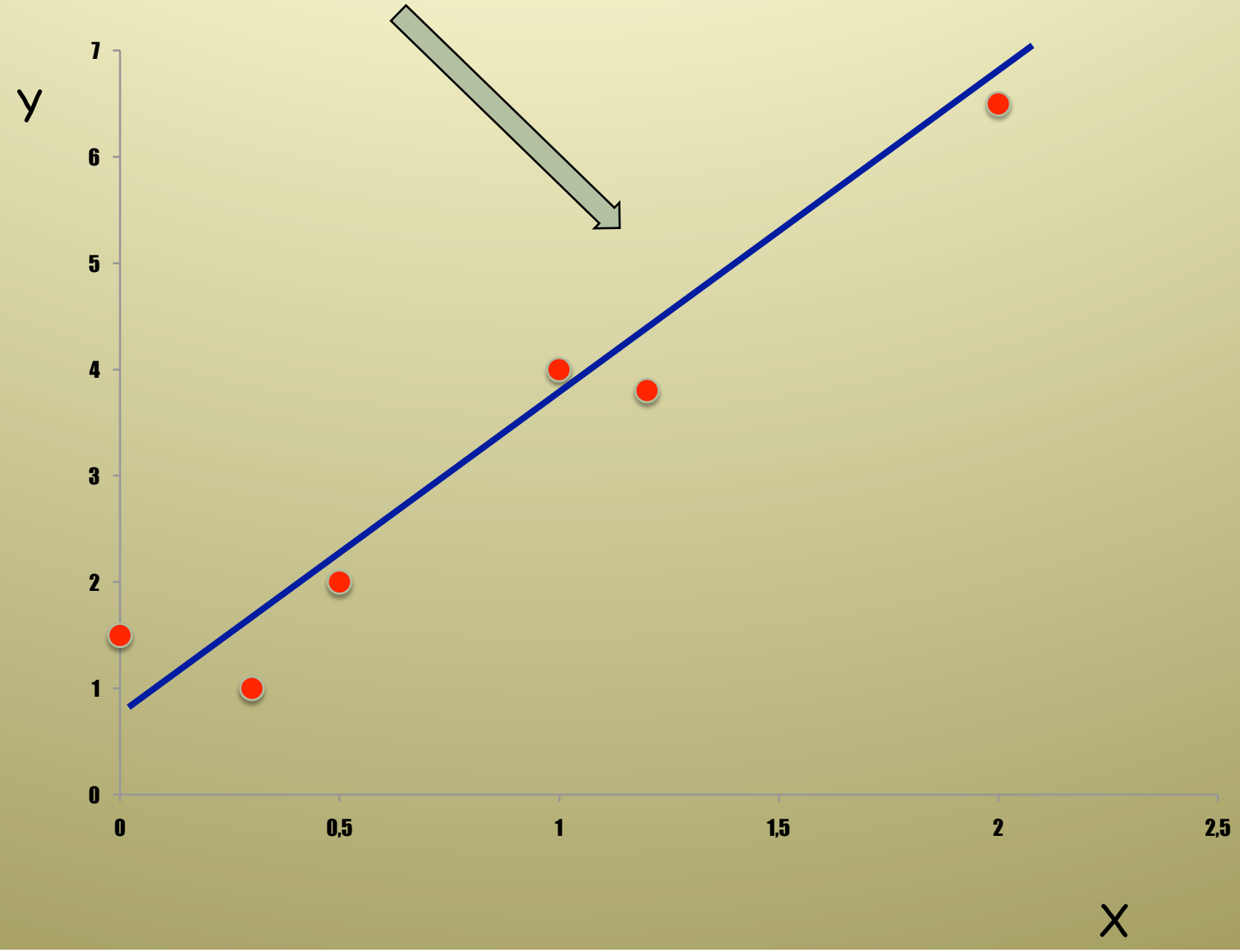
$$\sigma_X \approx 0.66 \quad \sigma_Y \approx 1.87 \quad \sigma_X \sigma_Y \approx 1.23 \quad \sigma_{XY} \approx 1.2$$

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \approx 0.972 (\approx 1)$$

(FORTE CORRELAZIONE POSITIVA= DATI "QUASI ALLINEATI")

Se i dati sono quasi allineati, qual è la relazione lineare Y=aX+b che approssima i dati?

PER FARE PREVISIONI SI USA UNA RETTA "APPROSSIMANTE" (di REGRESSIONE)



Si puo' dimostrare che la retta di regressione ha equazione

$$y = \frac{\sigma_{XY}}{\sigma_X^2} x + \left(M_Y - \frac{\sigma_{XY}}{\sigma_X^2} M_X \right) \approx 1.82x + 1.51$$

Il campione di 6 dati puo' essere ampliato con le coppie

X: 1 0 1.2 0.3 2 0.5 3 1.5 . . .

Y: 4 1.5 3.8 1 6.5 2 6.97 4.24 . . .

che sono "buone" approssimazioni dei dati

(MODULO II SEMESTRE….)