STATISTICA DESCRITTIVA
VARIABILITA’

SAPIENZA
UNIVERSITÀ DI ROMA

annarita.vestri@uniroma1.it

STATISTICA DESCRITTIVA

Sintesi Statistica

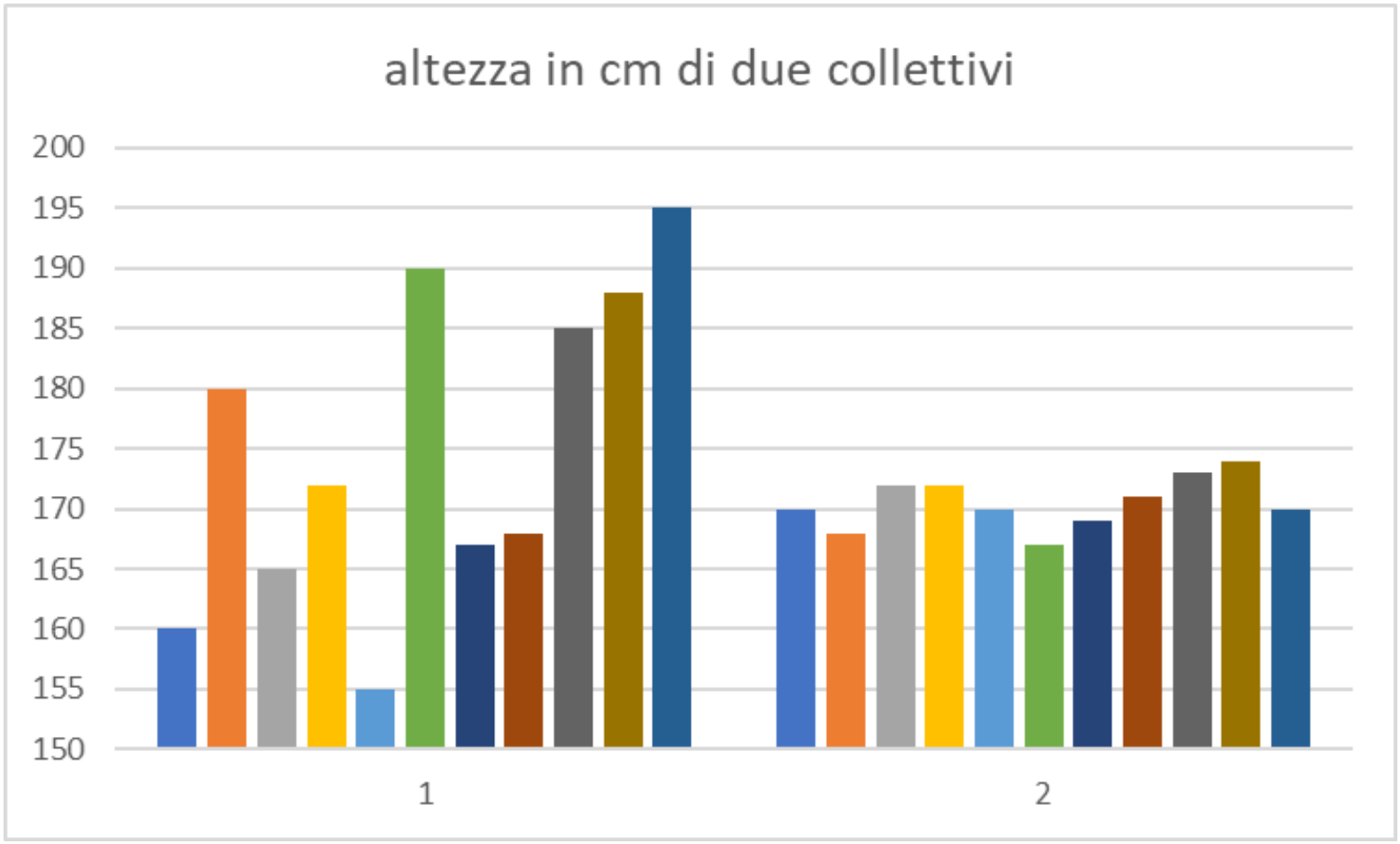
Una serie di dati numerici è compiutamente descritta da tre proprietà principali:

- La **tendenza centrale** o **posizione**
- La **dispersione** o **variabilità**
- La **forma**

Queste misure descrittive sintetiche, riassuntive dei dati tabellari, sono chiamate:

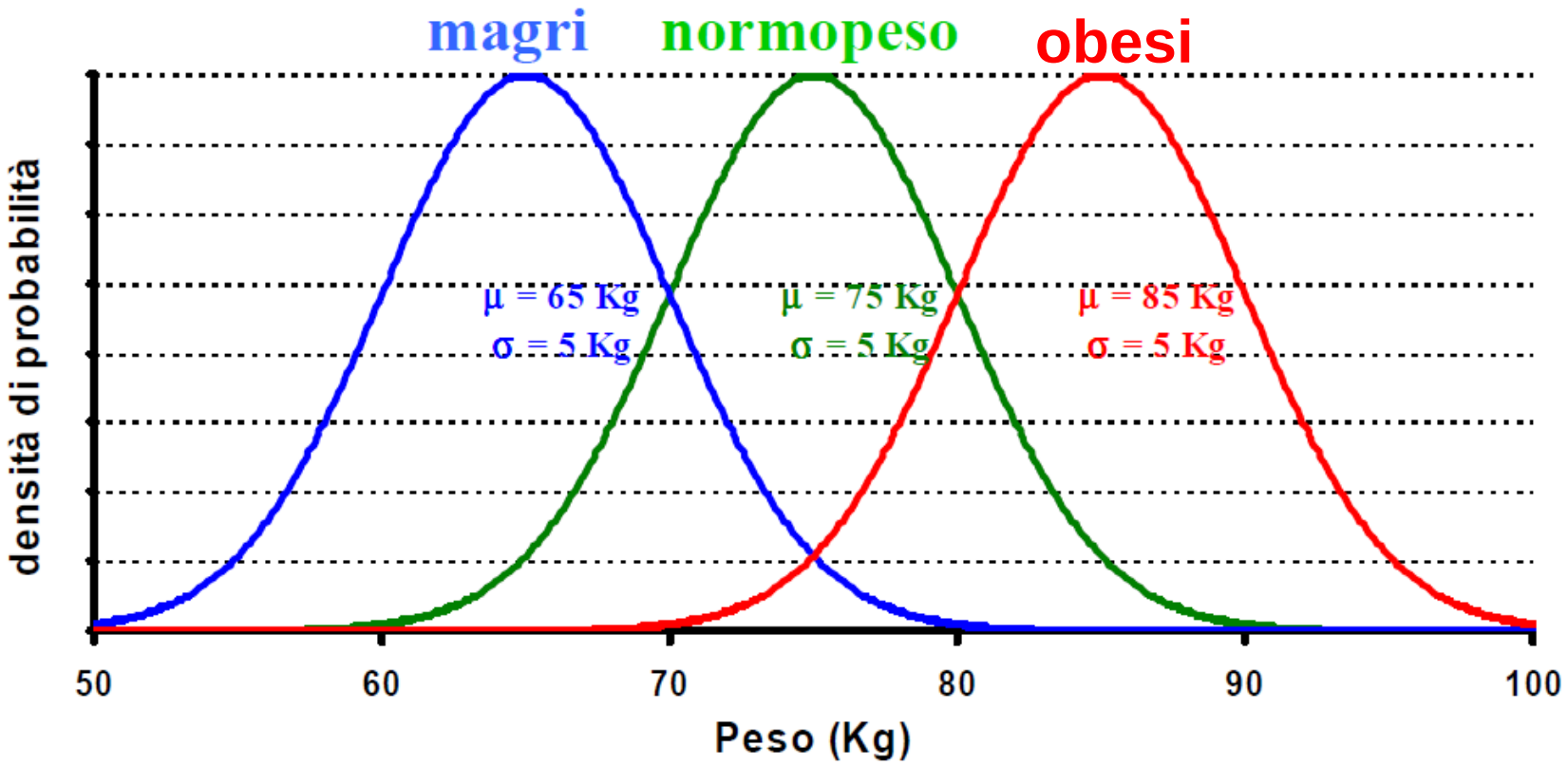
- **statistiche**, quando sono calcolate su un campione di dati (si esprimono con lettere dell'alfabeto latino)
- **parametri**, quando descrivono la popolazione od universo dei dati (si esprimono con lettere dell'alfabeto greco

la variabile d'interesse è l'ALTEZZA

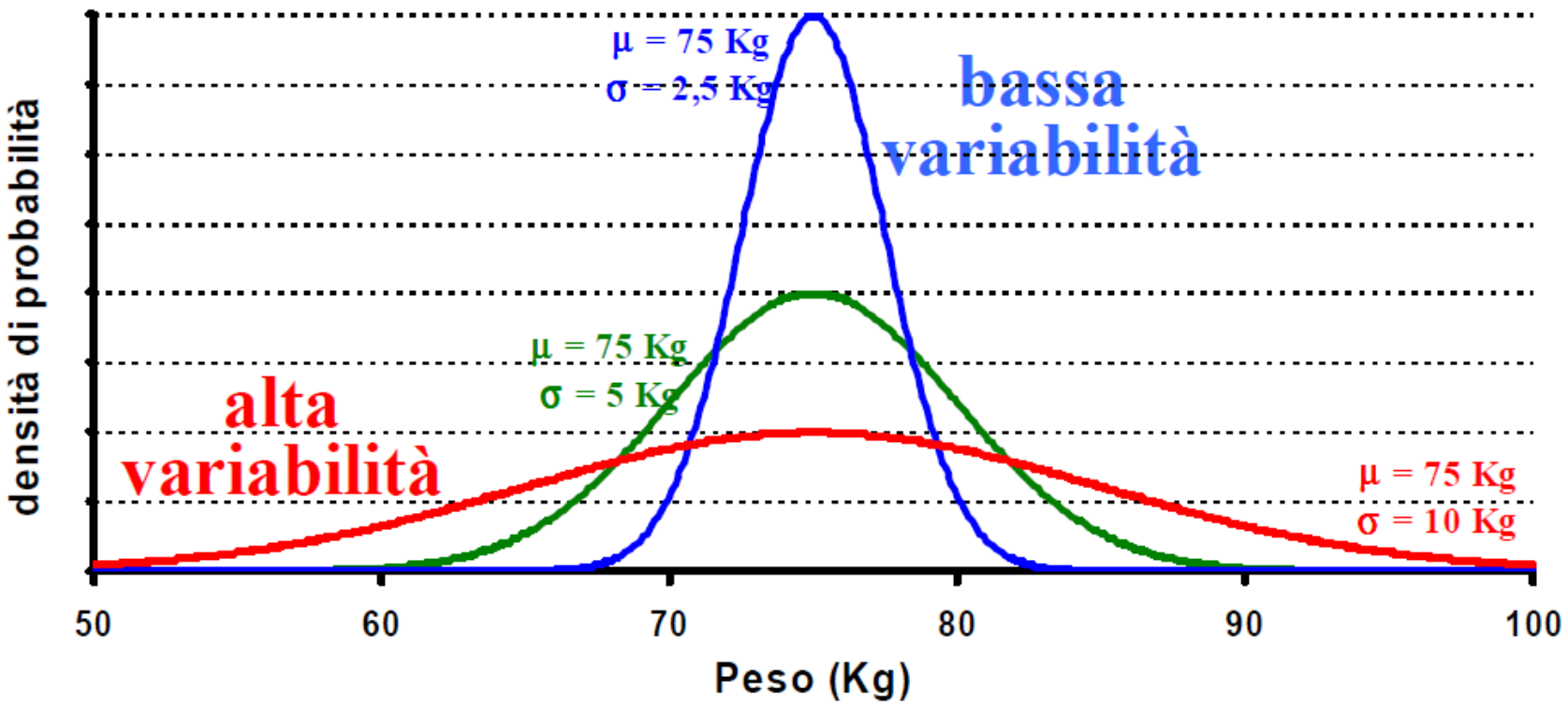


Quale dei due collettivi vi sembra più variabile?

Posizione differente, uguale dispersione



Dispersione differente, uguale posizione



INDICI DI DISPERSIONE

CAMPO DI VARIAZIONE (range)

DISTANZA INTERQUARTILE

DEVIANZA

VARIANZA

DEVIAZIONE STANDARD

COEFFICIENTE DI VARIAZIONE
(indice variabilità relativa)

STATISTICA DESCRITTIVA *dispersione di una distribuzione*

Ore di sonno	Maschi	Femmine
1	1	3
2	3	6
3	3	7
4	7	8
5	11	5
6	8	3
7	4	1
8	2	1
9	1	1
10	-	-
11	-	1
12	-	1
13	-	1
14	-	1
15	-	1

Usando SOLO le medie possiamo ingannarci nel confrontare i caratteri di due gruppi di individui.

Queste sono le distribuzioni di frequenza della **durata di sonno indotto da un anestetico** in un campione di **40+40** pazienti.

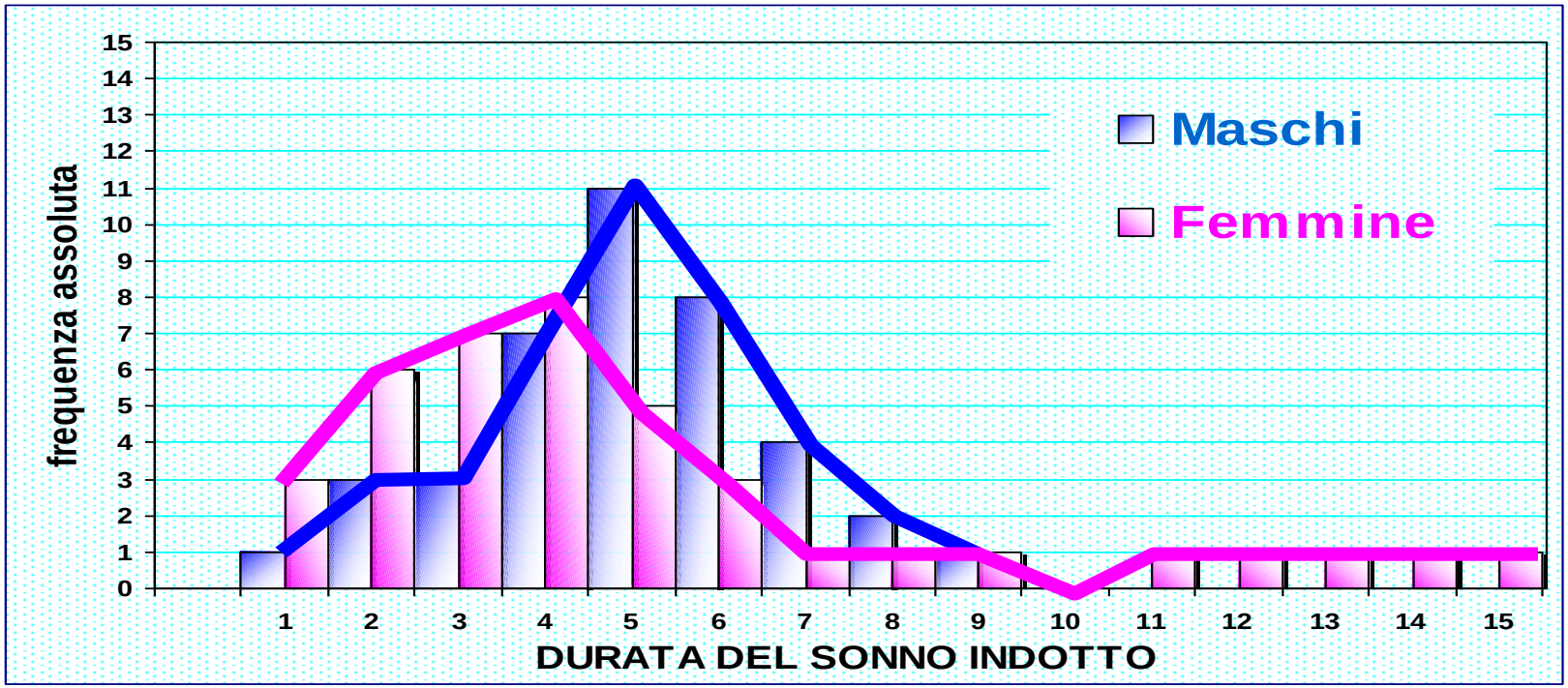
Ad esempio , sappiamo che le donne sono notoriamente diverse dagli uomini sotto molti aspetti

STATISTICA DESCRITTIVA *dispersione di una distribuzione*

Il periodo medio di ore di sonno per le donne risulta di 5 ore così come per gli uomini

Se ci soffermiamo solo sulle medie potremmo concludere che

le donne hanno una durata di sonno uguale a quello dei maschi.



STATISTICA DESCRITTIVA Misure di Variabilità

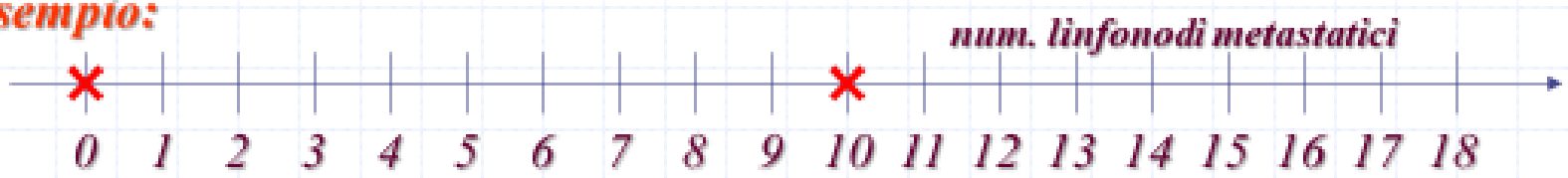
- **Range (campo di variazione) = Xmax - Xmin**

(differenza tra il valore massimo e il valore minimo di una distribuzione)

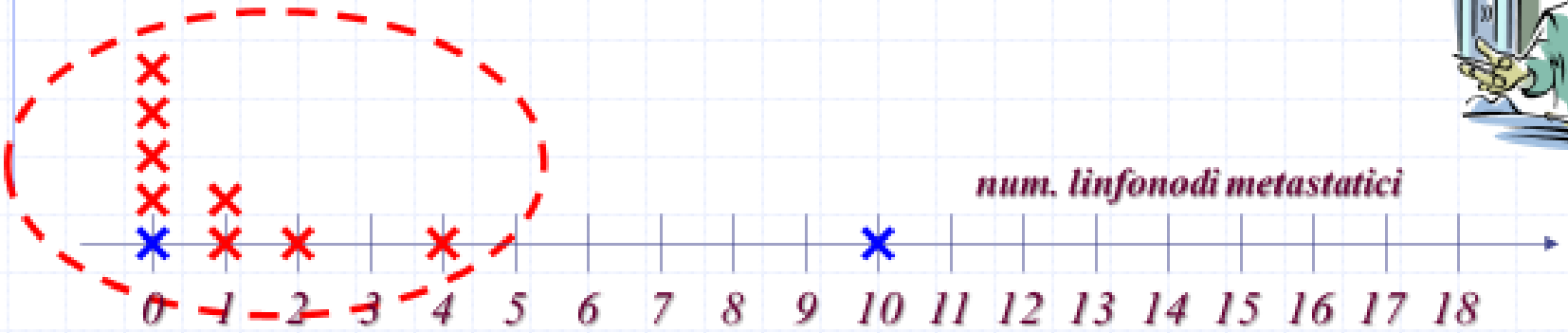
Svantaggi

- Si basa soltanto sui valori estremi della distribuzione e non tiene conto dei valori intermedi
- Tende ad aumentare al crescere del numero delle osservazioni
- E' molto influenzato da osservazioni anomale (*outliers*)

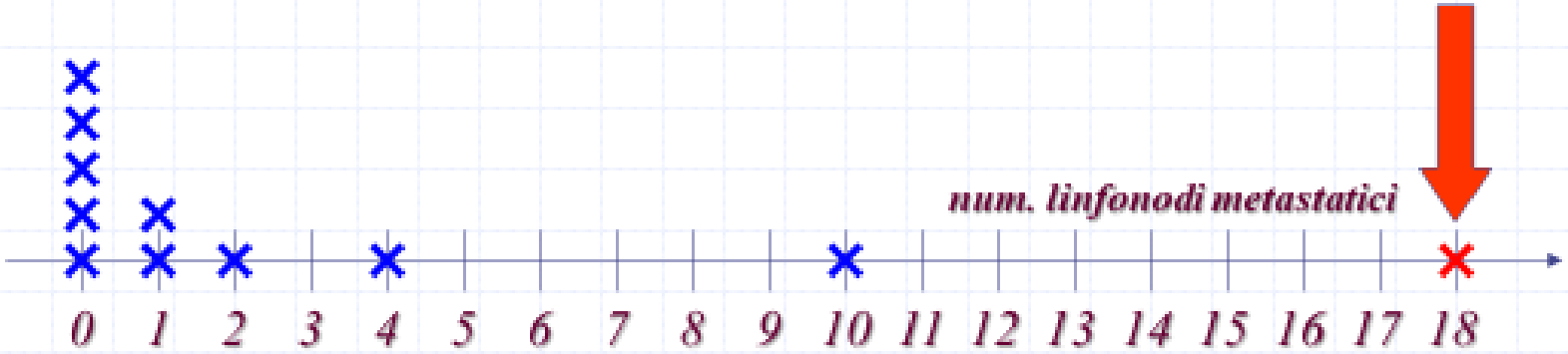
esempio:



$n = 2 \rightarrow Range = x_{max} - x_{min} = 10 - 0 = 10$



$n = 10 \rightarrow Range = x_{max} - x_{min} = 10 - 0 = 10$



$n = 11 \rightarrow Range = x_{max} - x_{min} = 18 - 0 = 18$

campo di variazione

Esempio:

Gli insiemi di valori di VES

{A}:{8,5,7,6,35,5,4}

{B}:{11,8,10,9,17,8,7}

in {A} i valori sono inclusi tra 4 e 35

in {B} i valori sono inclusi tra 7 e 17

La differenza tra il massimo e il minimo valore di un insieme di dati è detto **intervallo di variazione** (o **range**).

il **range** di {A} è RA = 35 - 4 = 31

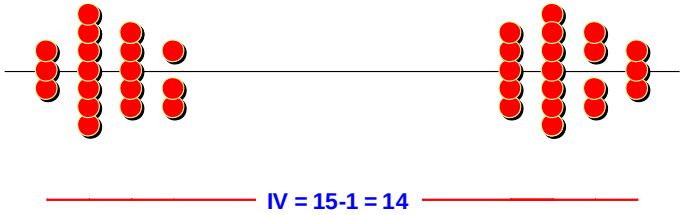
il **range** di {B} è RB = 17 - 7 = 10

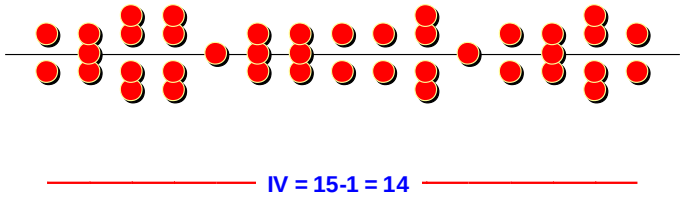
Il **range** è il più ***intuitivo*** fra gli indici di dispersione, ha però il difetto di basarsi solo sui due valori estremi, nei quali si manifesta maggiormente la variabilità di campionamento e l'errore di misura

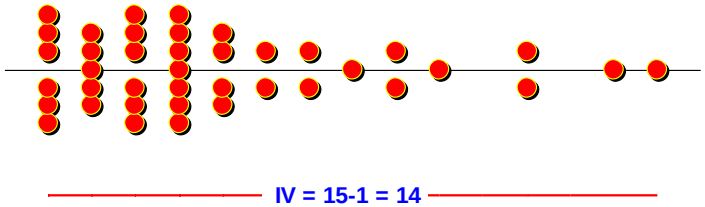
Campo di variazione

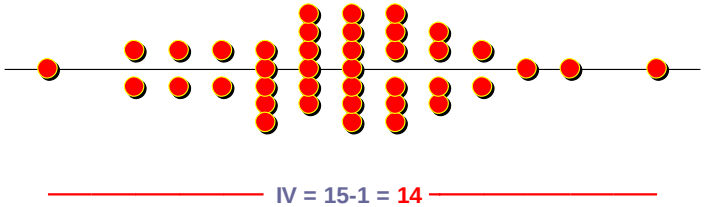
I dati possono …

- essere uniformemente distribuiti,
- concentrarsi ai due estremi della scala
- concentrarsi a un capo della scala
- o disporsi in altro modo









Range o distanza interquartile

IQR = Q3 - Q1

differenza tra il terzo quartile *(75° percentile*)e il 1° quartile *(25° percentile*)

Osservazioni

- In questo intervallo ricade la metà dei valori, posta esattamente al centro della distribuzione
- Non è molto influenzata da osservazioni anomale o estreme **(statistica robusta**)
- E' adatta a esprimere la variabilità di distribuzioni asimmetriche

Range o distanza interquartile
es: Statura matricole della Facoltà di Medicina

Range = x_{max} - x_{min} = 193 - 162 = 31 cm

MASCHI

Statura	Freq.	Cumul.
162	1	1
168	1	2
169	1	3
170	3	6
172	2	8
174	2	10
175	5	15
176	3	18
177	3	21
178	3	24
179	1	25
181	1	26
182	2	28
183	2	30
184	1	31
188	1	32
192	1	33
193	1	34
Totale	34	

Calcolo del I° quartile:

(rango percentilico = 25)

1. rango = (34+1) * 25 / 100
   = 35 / 4 ≈ 9

2. I° quartile = 174 cm

Calcolo del III° quartile:

(rango percentilico = 75)

1. rango = (34+1) * 75 / 100
   = 35 * 3 / 4 ≈ 26

2. III° quartile = 181 cm

IQR = Q3 - Q1 = 181 - 174 = 7 cm

generalmente si riporta: 174,181

Table 3. Allergy parameters in subjects without self-reported allergic rhinitis and in subjects with perennial, seasonal and perennial+seasonal rhinitis. Absolute frequencies with percentage in brackets are reported for all variables but total IgE, which is expressed as median (interquartile range).

	No rhinitis	Subjects with self-reported allergic rhinitis			
		Perennial	Seasonal	Perennial + seasonal	P value
	(n=745)	(n=19)	(n=50)	(n=87)	
Parental allergy	120/736 (16)	5/19 (26)	21/48 (44)	30/87 (34)	<0.001
Pos. specific IgE					
D.pteronyssinus	56/623 (9)	6/15 (40)	7/43 (16)	19/70 (27)	<0.001
Cat	17/623 (3)	2/15 (13)	4/43 (9)	12/70 (17)	---
Timothygrass	57/623 (9)	3/15 (20)	26/43 (60.5)	39/70 (56)	<0.001
Cl.herbarum	3/623 (0.5)	1/15 (7)	1/43 (2)	3/70 (4)	---
Pariet. judaica	29/623 (5)	1/15 (7)	16/43 (37)	32/70 (46)	<0.001
Total IgE	36.1 (13.2-101)	110.5 (11.6-217.5)	87 (38-214.5)	106 (50.5-240)	<0.001

Significance of differences was evaluated by chi-squared test for categorical variables and by one-way ANOVA for total IgE after logarithmic transformation. Significance was notevaluated by chi-squared test (---) when cells with expected value<5 exceeded 25%. NS =not significant

Olivieri M, Verlato G, Corsico A, Lo Cascio V, Bugiani M, Marinoni A, de Marco R, for the ItalianECRHS group (2002) Prevalence and features of allergic rhinitis in Italy. Allergy, 57:600-606

DEVIANZA

Nella popolazione

dimensione della popolazione

$$\sum_{i=1}^{N} (x_i - \mu)^2$$

media nella popolazione (parametro)

Nel campione

dimensione del campione

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

media nel campione (statistica)

- ✓ E' un indice di dispersione definito sulla base del concetto di **scarto rispetto ad un punto centrale della distribuzione**.
- ✓ E' la base delle misure di dispersione per variabili quantitative (da essa discendono la Varianza e la Deviazione Standard).

DEVIANZA

	media	scarto	scarto^2
5		-1	1
6	6	0	0
7		1	1
			Dev=2
5		-1	1
6		0	0
7	6	1	1
5		-1	1
6		0	0
7		1	1
			Dev=4

La devianza raddoppia anche se la variabilità è costante, perché aumenta il numero delle osservazioni!

VARIANZA

- E’ una devianza media ossia la devianza rapportata al numero delle osservazioni campionarie (n) o della popolazione (N).
- E’ la media aritmetica dei quadrati degli scarti delle singole osservazioni dalla loro media.

Nella popolazione

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

parametro

Nel campione

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

statistica

Gradi di Libertà

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

VARIANZA:
formula per il calcolo

I GRADI DI LIBERTÀ rappresentano il numero di osservazioni indipendenti del campione, dal momento che sui dati disponibili è già stata calcolata una statistica (la media campionaria).

Quando le osservazioni sono raggruppate in una distribuzione di frequenza (*in k classi*):

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

Varianza Osservazioni

- E’ adatta per distribuzioni simmetriche
- Tiene conto di tutte le osservazioni ed è dunque influenzata da eventuali osservazioni anomale ***(outliers***)
- Non è direttamente confrontabile con la media o altri indici di posizione in quanto le unità di misura sono elevate al quadrato **(valore teorico**)

Varianza esempio stature matricole a medicina

CLASSE	PUNTO CENTRALE (x_i)	FREQUENZA ASSOLUTA	$n_i * x_i$	$n_i * x_i^2$
[150-155)	152.5	1	152.5* 1 = 152.5	$(152.5)^2$* 1 = 23256.25
[155-160)	157.5	8	157.5* 8 = 1260.0	$(157.5)^2$* 8 = 198450.00
[160-165)	162.5	24	162.5*24 = 3900.0	$(162.5)^2$*24 = 633750.00
[165-170)	167.5	34	5695.0	953912.50
[170-175)	172.5	27	4657.5	803418.75
[175-180)	177.5	19	3372.5	598618.75
[180-185)	182.5	9	1642.5	299756.25
[185-190)	187.5	1	187.5	35156.25
[190-195]	192.5	2	385.0	74112.50
TOTALE		125	21252.5	3620431.25

$$s^2 = \frac{\sum_{i=1}^{n} n_i x_i^2 - (\sum_{i=1}^{n} n_i x_i)^2 / n}{n-1} = \frac{362043125 - (212525)^2 / 125}{124} = 57.1 cm^2$$

DEVIAZIONE STANDARD

Nella popolazione

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Nel campione
(d.s. corretta)

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

- Ha sempre valore positivo
- E' una misura della **dispersione della variabile intorno alla media**
- E' una misura di **distanza dalla media**, direttamente confrontabile con le misure di posizione, essendo calcolata con la stessa unità di misura.

Deviazione Standard

Rappresenta una misura della deviazione dei valori dalla media. Esso ci dice come i valori tendano a disperdersi intorno alla loro media:

se la deviazione standard è piccola, indica un fitto addensamento dei valori intorno alla loro media;

se è grande indica la presenza di valori molto lontani dalla media.

Deviazione standard esempio stature matricole a medicina

CLASSE	PUNTO CENTRALE (x_i)	FREQUENZA ASSOLUTA	$n_i * x_i$	$n_i * x_i^2$
[150-155)	152.5	1	152.5* 1 = 152.5	$(152.5)^2$* 1 = 23256.25
[155-160)	157.5	8	157.5* 8 = 1260.0	$(157.5)^2$* 8 = 198450.00
[160-165)	162.5	24	162.5*24 = 3900.0	$(162.5)^2$*24 = 633750.00
[165-170)	167.5	34	5695.0	953912.50
[170-175)	172.5	27	4657.5	803418.75
[175-180)	177.5	19	3372.5	598618.75
[180-185)	182.5	9	1642.5	299756.25
[185-190)	187.5	1	187.5	35156.25
[190-195]	192.5	2	385.0	74112.50
TOTALE		125	21252.5	3620431.25

$$s = \sqrt{\frac{\sum_{i=1}^{n} n_i x_i^{\,2} - \left(\sum_{i=1}^{n} n_i x_i\right)^2 / n}{n-1}} = \sqrt{\frac{3620431.25 - (21252.5)^2 / 125}{124}} = \sqrt{57.1} = 7.6cm$$

Coefficiente di variazione (CV)

Due gruppi con valori medi molto distanti

Tre neonati pesano rispettivamente **3, 4 e 5 Kg** (media ± DS: 4 ± **1 K**g).

Tre bambini di 1 anno pesano **10, 11 e 12 Kg** (media ± DS: 11 ± **1 K**g).

La deviazione standard è uguale nei due gruppi, ma il buon senso suggerisce che la variabilità del peso sia …….

Due variabili diverse

In 91 ragazze matricole di Medicina a Roma nell'a.a. 2018/2019,

il peso era pari a 55,1 ± **5,7 Kg** (media ± DS) con un range di **45-70 Kg,**

la statura era 166,1 ± **6,1 cm** (media±DS) con un range di **150-182 cm.**

E' maggiore la variabilità del peso o la variabilità della statura?

Coefficiente di variazione (1)

Il coefficiente di variazione di un carattere X di media $\bar{X}_n$ diversa da zero e deviazione standard s è dato dal rapporto tra la deviazione standard e la media aritmetica

$$CV = \frac{s}{\bar{X}_n} \cdot 100 = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{X}}{\bar{X}} \right)^2} \cdot 100$$

L'ultima espressione mostra come il coefficiente di variazione può anche essere interpretato come una media quadratica degli **scarti relativi** rispetto alla media aritmetica.

Coefficiente di variazione (CV)

Per rispondere a queste domande è necessario calcolare il **coefficiente di variazione:**

CV = (deviazione standard / media) * 100.

La deviazione standard viene espressa in percentuale della media.

	Media	Dev. standard	**CV**
Neonati	4 Kg	1 Kg	**25.0 %**
Bambini 1 anno	11 Kg	1 Kg	**9.1 %**

La variabilità del peso è maggiore nei neonati.

	Media	Dev. standard	**CV**
Peso	55.1 Kg	5.7 Kg	**10.3 %**
Statura	166.1 cm	6.1 cm	**3.7 %**

La variabilità del peso è maggiore della variabilità della statura.

Coefficiente di variazione esempio

Esempio: consideriamo 5 altezze misurate in metri

$$1.78 \quad ; \quad 1.99 \quad ; \quad 1.73 \quad ; \quad 1.81 \quad ; \quad 1.58 \quad \Rightarrow \quad s^2 = 0.017496$$

consideriamo ora le **stesse 5 altezze** misurate in centimetri

$$178 \quad ; \quad 199 \quad ; \quad 173 \quad ; \quad 181 \quad ; \quad 158 \quad \Rightarrow \quad s^2 = 174.96$$

Guardando i valori delle varianza sembrerebbe essere molto più variabile la seconda serie dei dati ma i dati confrontati son gli stessi cambia solo l'unità di misura!!

Coefficiente di variazione: Esempio

5 altezze misurate in metri

$$1.78 \quad ; \quad 1.99 \quad ; \quad 1.73 \quad ; \quad 1.81 \quad ; \quad 1.58 \quad \Rightarrow$$

$$\bar{x}_5 = 1.778 \quad ; \quad s^2 = 0.017496 \quad ; \quad s = 0.1322724 \Rightarrow$$

$$CV = \frac{0.1322724}{1.778} 100 = 7.439393$$

stesse 5 altezze misurate in centimetri

$$178 \quad ; \quad 199 \quad ; \quad 173 \quad ; \quad 181 \quad ; \quad 158 \quad \Rightarrow$$

$$\bar{x}_5 = 177.8 \quad ; \quad s^2 = 174.96 \quad ; \quad s = 13.22724 \Rightarrow$$

$$CV = \frac{13.22724}{177.8} 100 = 7.439393$$

Stesso valore del coefficiente di variazione!!

Esempio di calcolo degli indici di dispersione

esempio di due insiemi di valori di VES si ha:

{A}: { 8, 5, 7, 6, 35, 5, 4}

$s^2 = 740/6 = 123.33$ $s = \sqrt{123.3.} = 11.1$ = (-1.1,21.1)

$CV\% = 100 \times (11.1/10) = 111\%$

{B}: { 11, 8, 10, 9, 17, 8, 7}

$s^2 = 68 / 6 = 11.33$ $s = \sqrt{11.33} = 3.4$ = (6.6, 13.4)

$CV\% = 100 \times (3.4/10) = 34\%$

In {A} l'intervallo ± s include anche valori negativi di VES, che ovviamente non sono possibili. L'uso di s per esprimere la dispersione dovrebbe essere quindi limitato alle distribuzioni simmetriche

Coefficiente di variazione: Osservazioni

Il coefficiente di variazione è utilizzato solo quando tutti i valori della distribuzione sono positivi come ad esempio nel caso di variabili come altezza, peso, età etc.

Infatti, per caratteri che assumono sia caratteri negativi o positivi la media aritmetica non rappresenta l'ordine di grandezza effettivo (esempio di variabili con media prossima a zero anche se in realtà presentano valori molto grandi e molto piccoli che però si compensano).