

Statistica descrittiva

I dati vanno raccolti, analizzati ed elaborati con le tecniche appropriate (organizzazione dei dati).

I dati vanno poi interpretati e valutati con le tecniche statistiche appropriate.

Dati raccolti + statistica = informazioni

Rank order distribution

(distribuzione ordinata per classifica)

Quando i dati sono pochi (≤ 20) la semplice loro ordinazione dal più piccolo al più grande può fornire più informazioni.

Rank order distribution

(distribuzione ordinata per classifica)

N = numero di dati

H = valore più alto

L = valore più basso

R (range) = $H - L$

$N=15$
 $H = 18$
 $L = 2$
 $\text{Range} = 18 - 2 = 16$

x
18
17
16
15
14
13
12
12
10
9
9
8
8
5
2

Distribuzione di frequenza semplice

Quando i dati sono molti può essere conveniente organizzare i dati in tabelle in cui in una colonna venga indicato il numero di volte che è stato rilevato un determinato dato (colonna della frequenza)

$N=368$
 $H = 19$
 $L = 0$
 $\text{Range} = 19 - 0 = 19$

x	f
19	1
18	2
17	3
16	4
15	5
14	7
13	10
12	12
11	15
10	20
9	21
8	20
7	15
6	12
5	10
4	7
3	5
2	4
1	3
0	2

Distribuzione di frequenza per classi

I dati vengono raggruppati in classi di frequenze

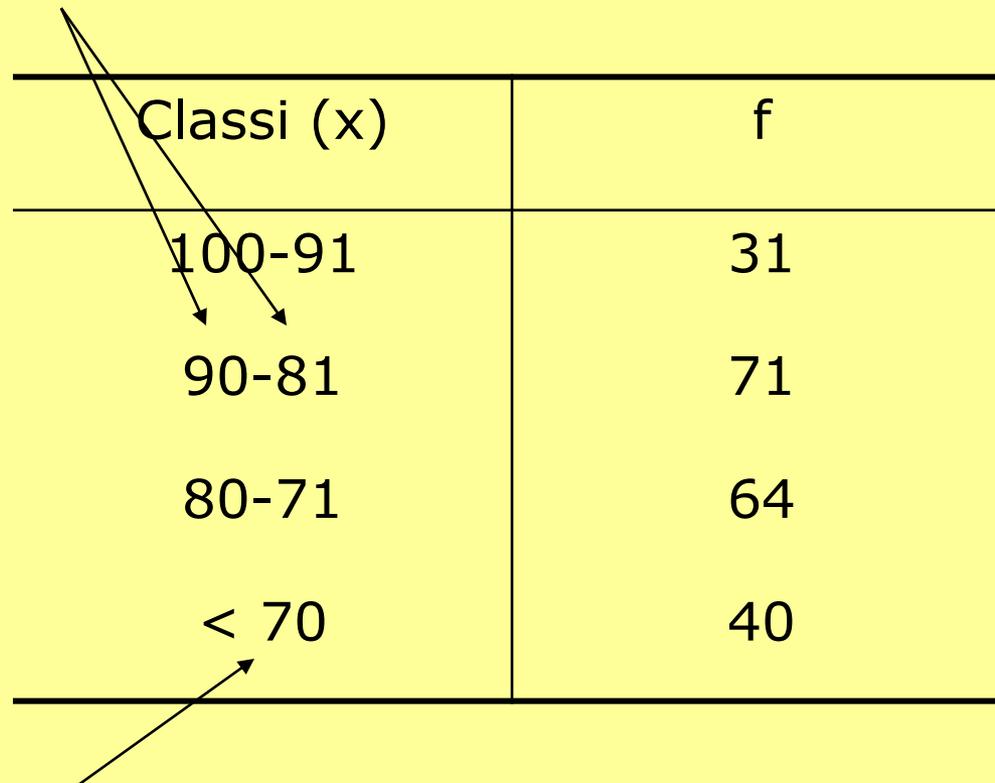
Classi (x)	<i>f</i>
100-91	31
90-81	71
80-71	64
< 70	40

$N = 206$

Distribuzione di frequenza per classi

Ogni classe è definita dall'*intervallo della classe*.

Limiti della classe



The table below shows class intervals and their frequencies. Arrows from the text 'Limiti della classe' point to the boundaries of the first class (100 and 91). An arrow from the text 'Intervallo aperto' points to the boundary of the last class (< 70).

Classi (x)	f
100-91	31
90-81	71
80-71	64
< 70	40

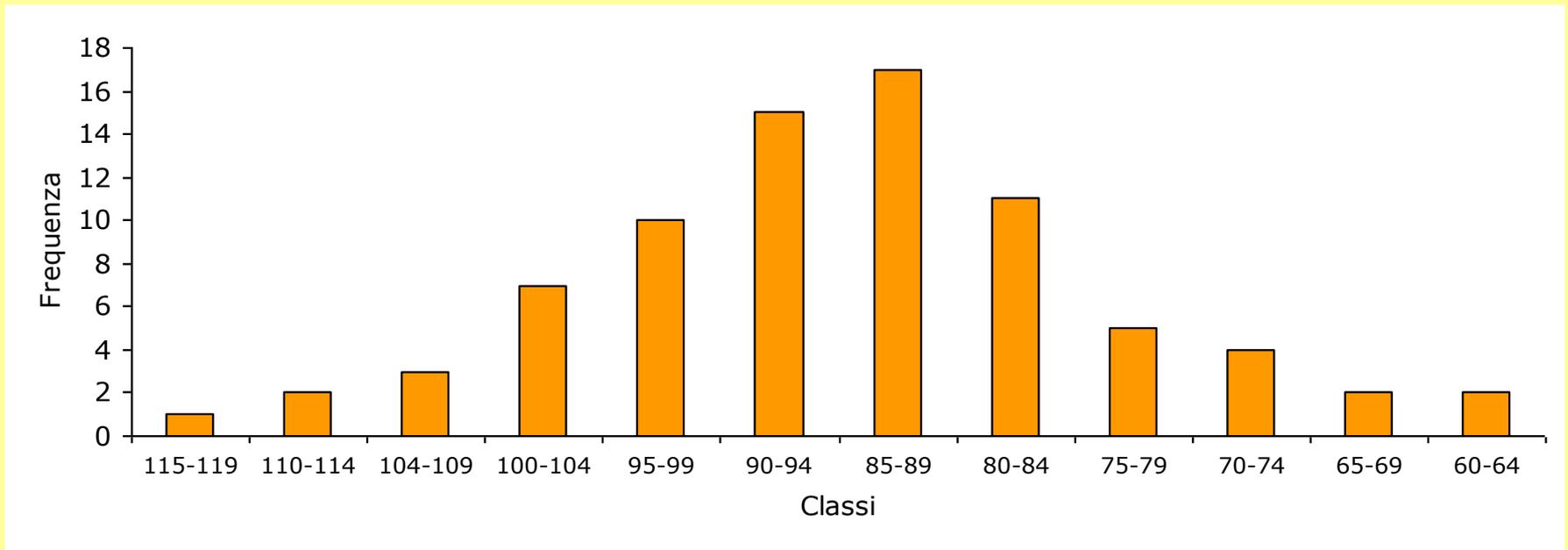
Intervallo aperto

Distribuzione di frequenza cumulativa

La frequenza di un gruppo viene sommata a quella di tutti gli altri gruppi al di sotto di esso

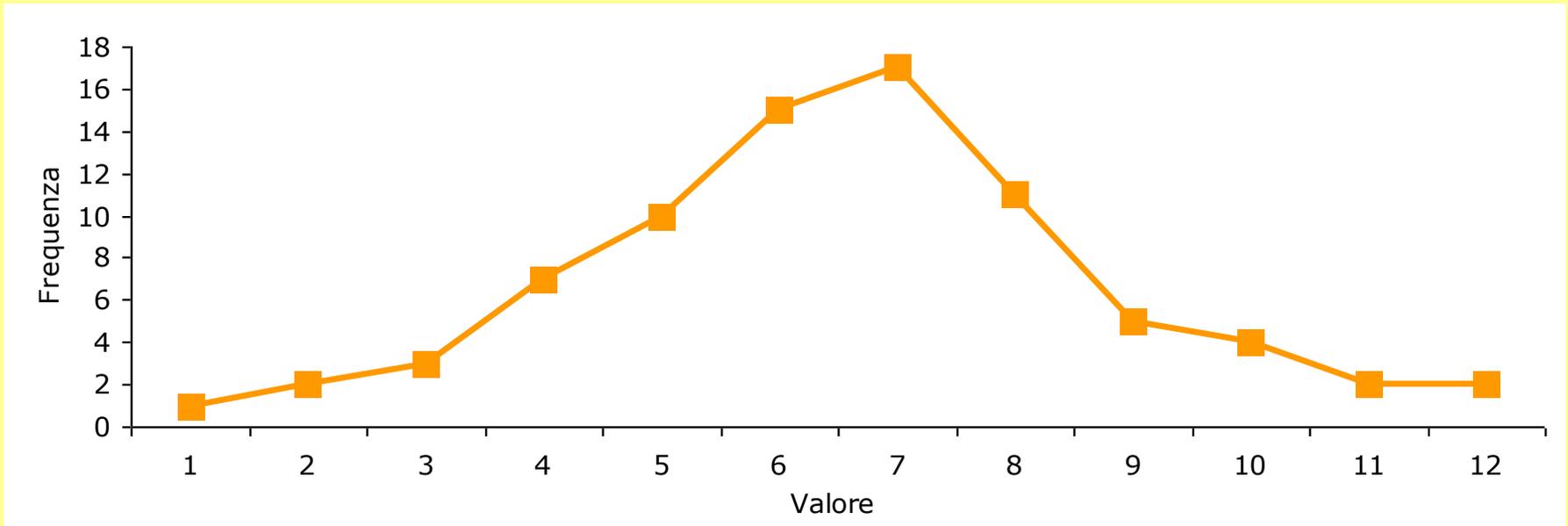
Classi (x)	f	Cum f
100-91	31	206
90-81	71	175
80-71	64	104
< 70	40	40

$N = 206$



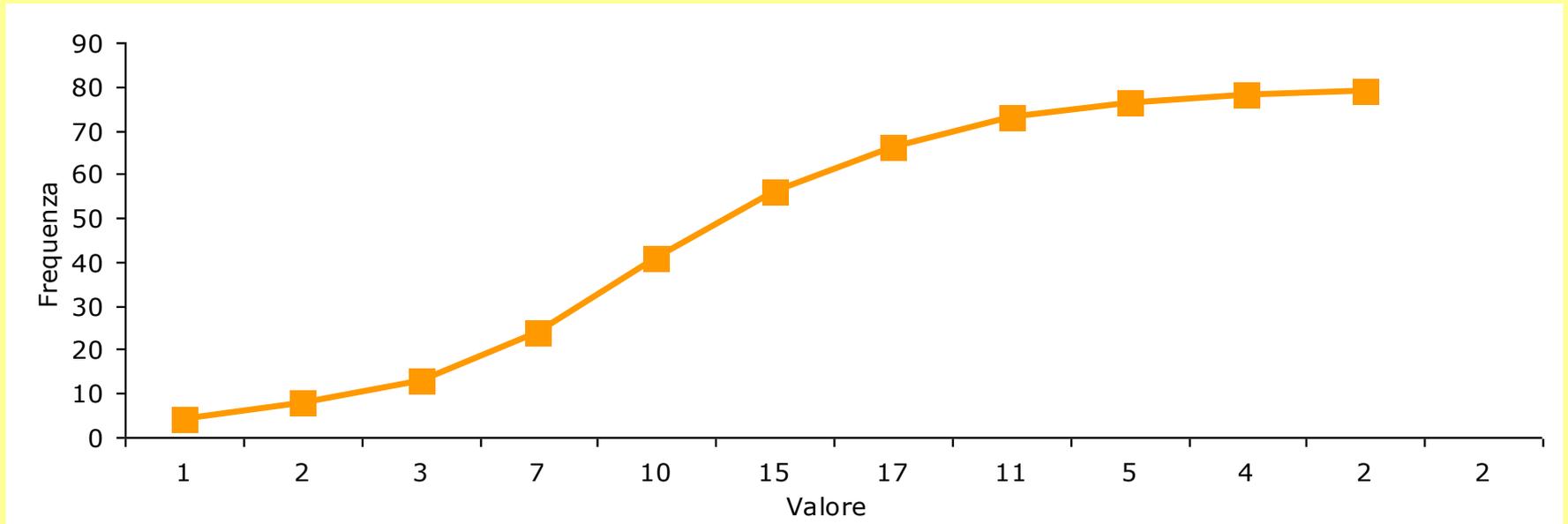
Istogramma

Utilizzato solitamente per distribuzione di frequenza in classi



Poligono di frequenza

Utilizzato solitamente per distribuzione di frequenza semplice

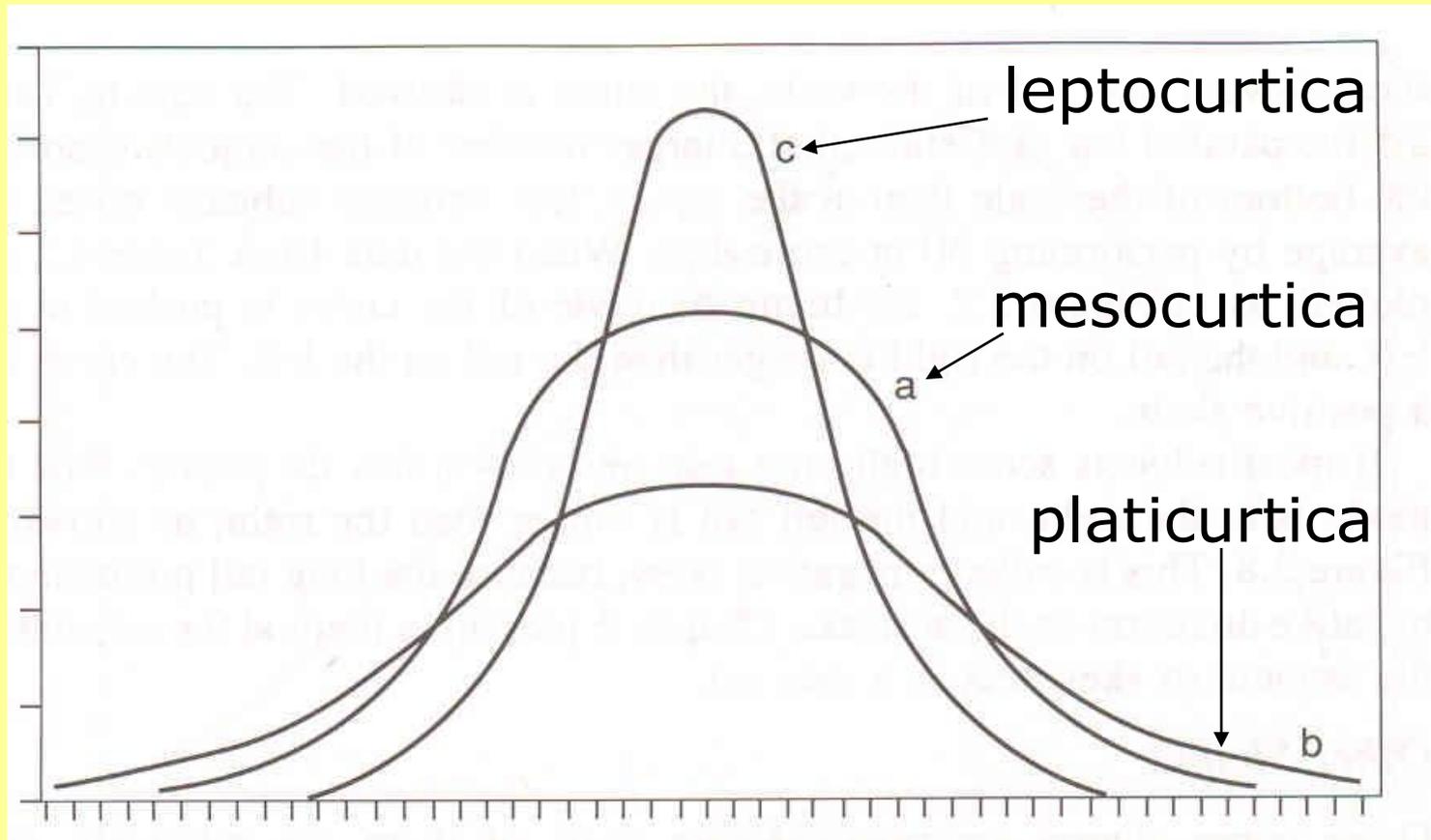


Poligono di frequenza cumulativa

La curva di distribuzione normale

E' caratterizzata dalla distribuzione simmetrica dei dati intorno al centro della curva.

Descritta per la prima volta dal matematico Karl Gauss (1777-1855): curva Gaussiana.



MODA

Valore più frequente

MEDIA

Media aritmetica - Somma dei valori (X) diviso la somma del numero dei valori (N):

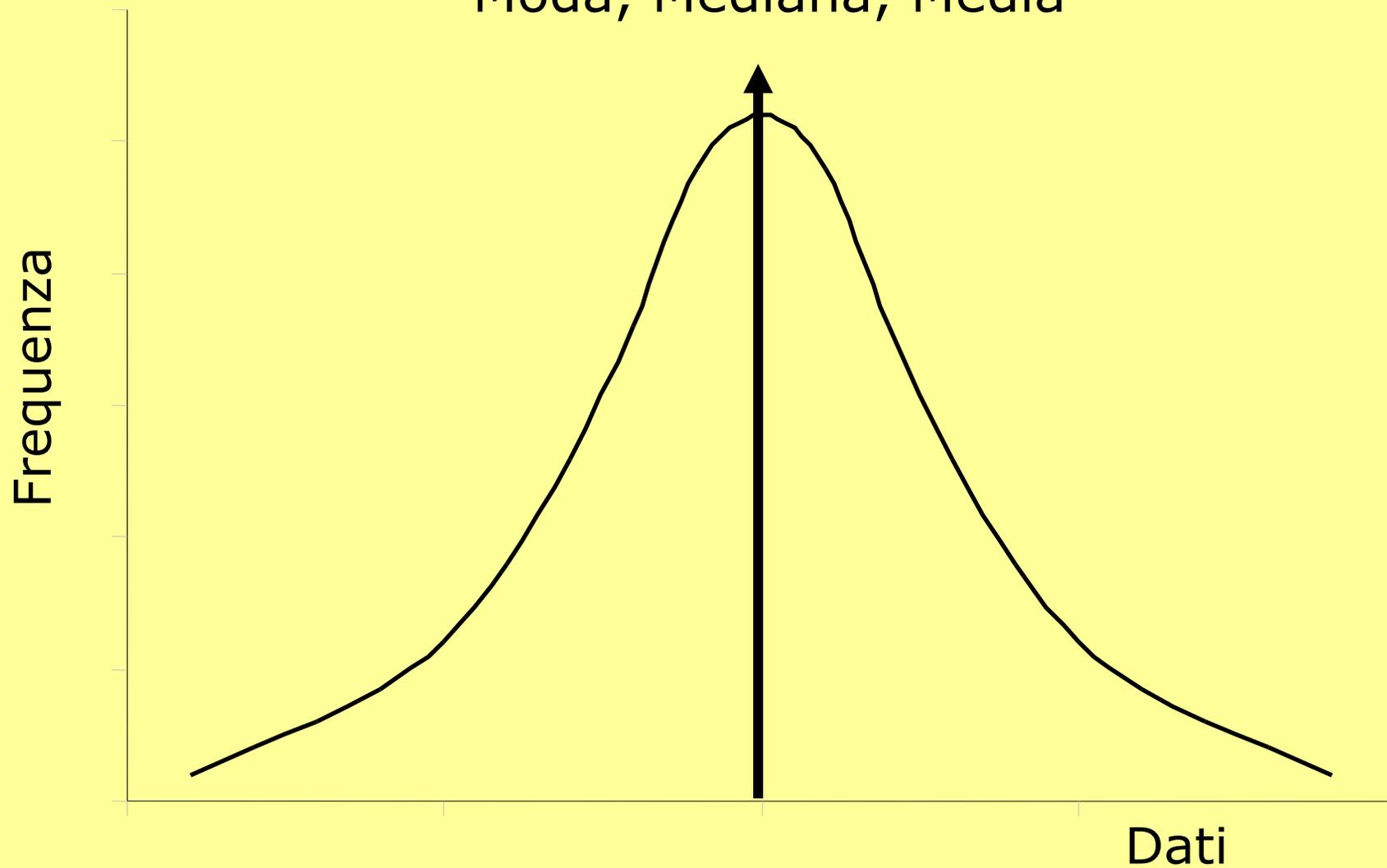
$$\Sigma X/N$$

MEDIANA

Valore associato al 50° percentile

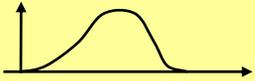
Se la distribuzione è normale moda, media e mediana coincidono

Moda, Mediana, Media





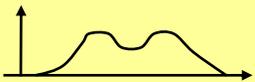
Curva obliqua a sinistra



Curva obliqua a destra



Curva a U



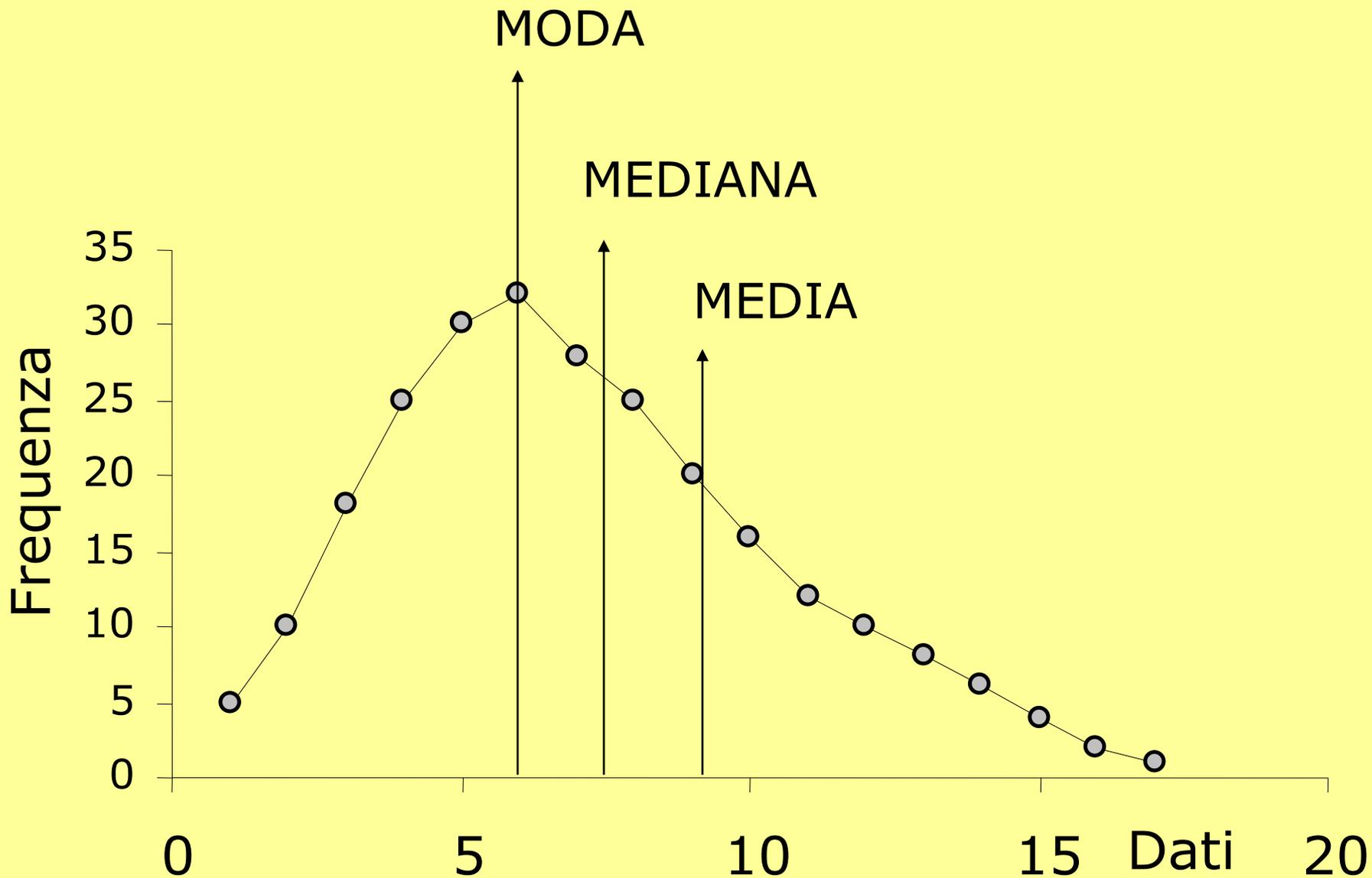
Curva bimodale



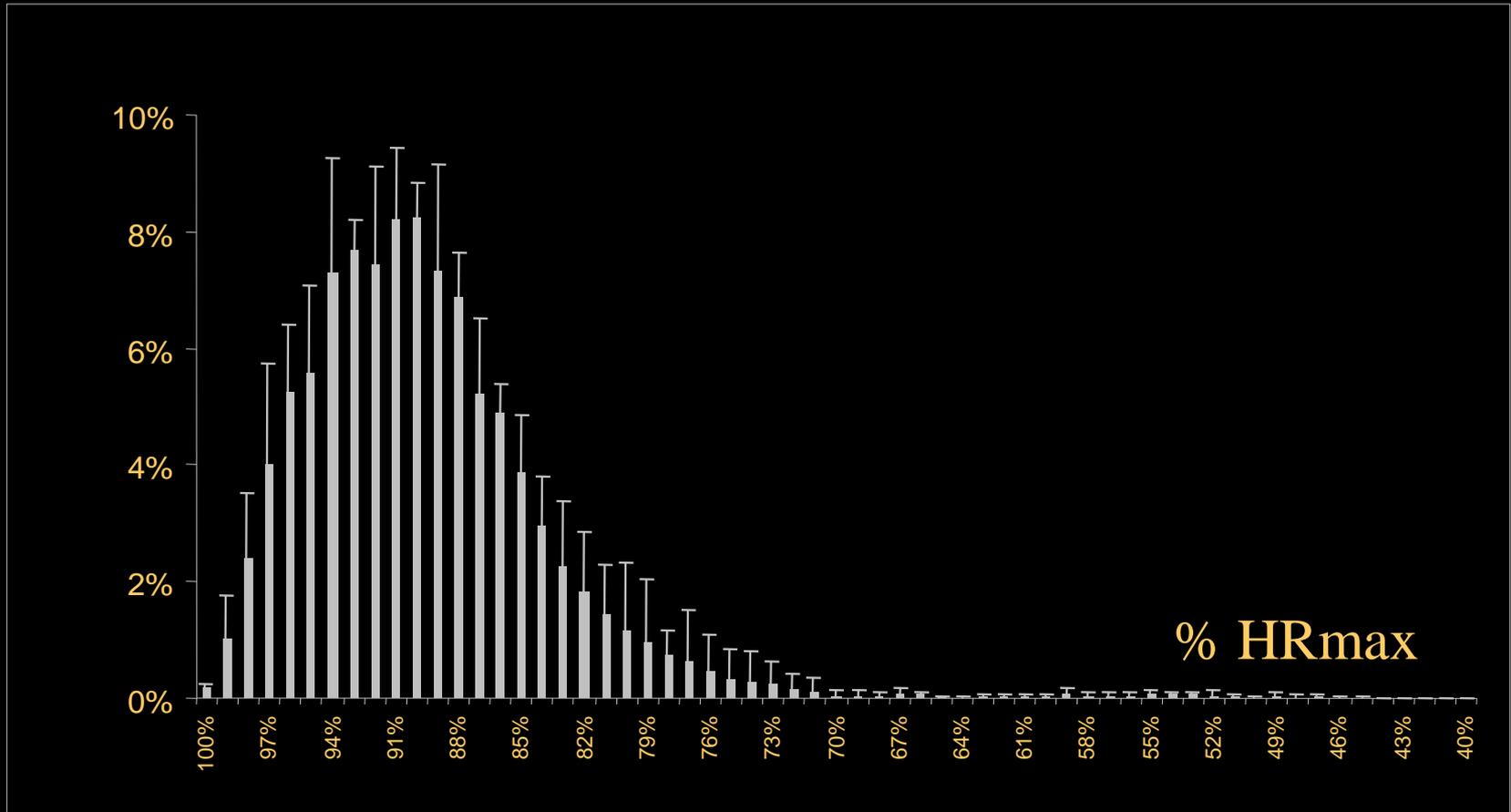
Curva discendente



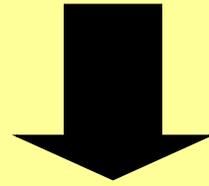
Curva ascendente



Intensità di una gara di mountain bike



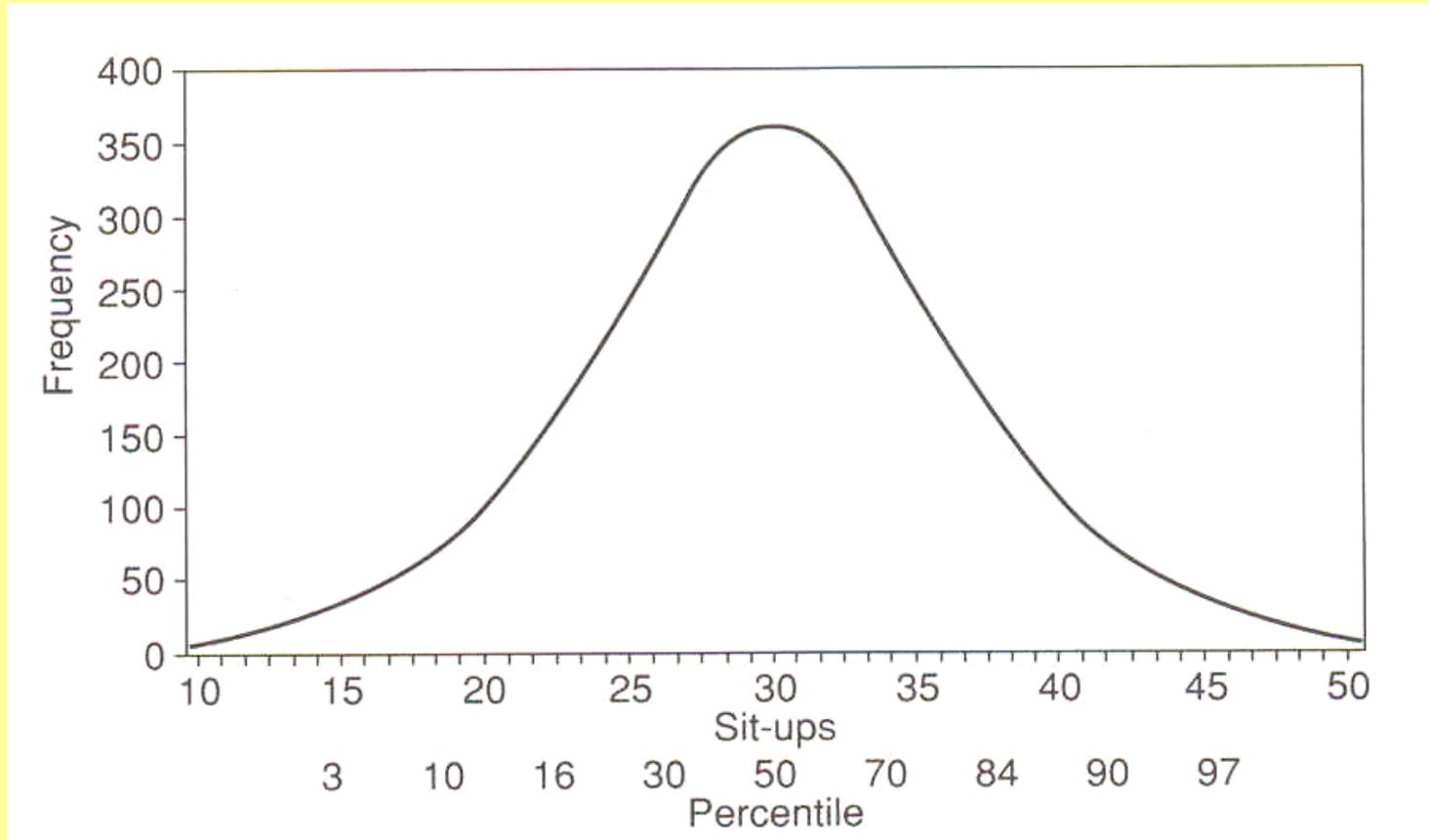
Statistica parametrica



Dati normalmente
distribuiti

Percentili

Il percentile rappresenta un punto su una scala di 100 divisioni teoriche in modo tale che una certa frazione della popolazione risulta uguale o inferiore a quel punto.



X

10
9
9
7
7
7
6
5
5
5
4
4
3
3
1

Numero di canestri su 10 tiri in 15 giocatori.

Qual è il percentile di 6?

Numero di valori uguali o inferiori a 6 = 9

$9/15 = 0.60 = 60\%$

6 è al 60° percentile

N=15

X

10
9
9
7
7
7
6
5
5
5
4
4
3
3
1

Quando una persona riceve lo stesso punteggio, il percentile è calcolato per il più alto valore dei dati ripetuti.

Percentile per 5 = $8/15 = 0.53 = 53\%$



N=15

Calcolo dei percentili con distribuzione cumulativa

x	f	Cum f
90	31	206
85	71	175
80	64	104
75	40	40

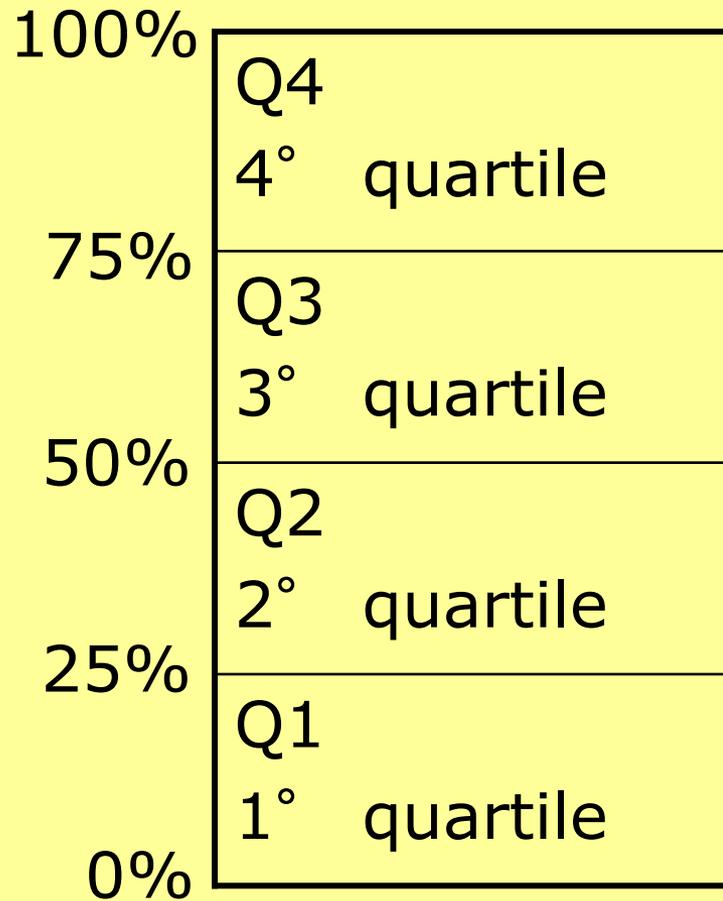
Percentile di 85

$$175/206(N) = 0.85 = 85\%$$

Un uomo ha $40 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ di $\text{VO}_{2\text{max}}$.

Il dato da solo non fornisce indicazioni particolari.

Se però sappiamo che questo valore si inserisce al 65° percentile sappiamo che $40 \text{ ml} \cdot \text{kg}^{-1} \cdot \text{min}^{-1}$ di $\text{VO}_{2\text{max}}$ è superiore al 65% della popolazione di riferimento.



100%	Q4 4° quartile
75%	Q3 3° quartile
50%	Q2 2° quartile
25%	Q1 1° quartile
0%	

100%	D10	10°	decile
90%	D9	9°	decile
80%	D8	8°	decile
70%	D7	7°	decile
60%	D6	6°	decile
50%	D5	5°	decile
40%	D4	4°	decile
30%	D3	3°	decile
20%	D2	2°	decile
10%	D1	1°	decile
0%			

Misure di tendenza centrale

Le misure di tendenza centrale sono valori che descrivono le caratteristiche centrali di un set di dati.

MODA

La moda è il valore che si presenta più frequentemente.

Svantaggi

1. E' instabile (cambia ad esempio con le classi)
2. E' un dato statistico terminale (non può essere usato per altri calcoli).
3. Toglie completamente peso ai valori estremi.

MEDIANA

Valore associato al 50° percentile. Divide il set di dati in due parti uguali.

Non è influenzata dall'entità dei valori estremi, ma da **quanti** ce ne sono.

MEDIA

Media aritmetica ($\Sigma X/N$).

La media è sensibile al numero e all'entità dei valori estremi.

I valori estremi (*outlier*) tirano nella loro direzione la media (possibile svantaggio).

Si usa la MODA quando:

1. Serve solo una grossolana indicazione di tendenza centrale e i dati sono normali.

Si usa la MEDIANA quando:

1. I dati sono su una scala ordinale.
2. Sono necessarie informazioni sul valore mediano del set di dati.
3. La curva di distribuzione è molto schiacciata.

Si usa la MEDIANA quando:

1. I dati sono su una scala ordinale.
2. Sono necessarie informazioni sul valore mediano del set di dati.
3. La curva di distribuzione è molto schiacciata.

Si usa la MEDIANA quando:

1. I dati sono su una scala ordinale.
2. Sono necessarie informazioni sul valore mediano del set di dati.
3. La curva di distribuzione è molto schiacciata.

Si usa la **MEDIA** quando:

1. I dati sono normali o vicino alla normalità.
2. Sono necessarie informazioni su tutti i dati (ordine e loro valore).
3. Serve calcolare altri parametri come la deviazione standard.

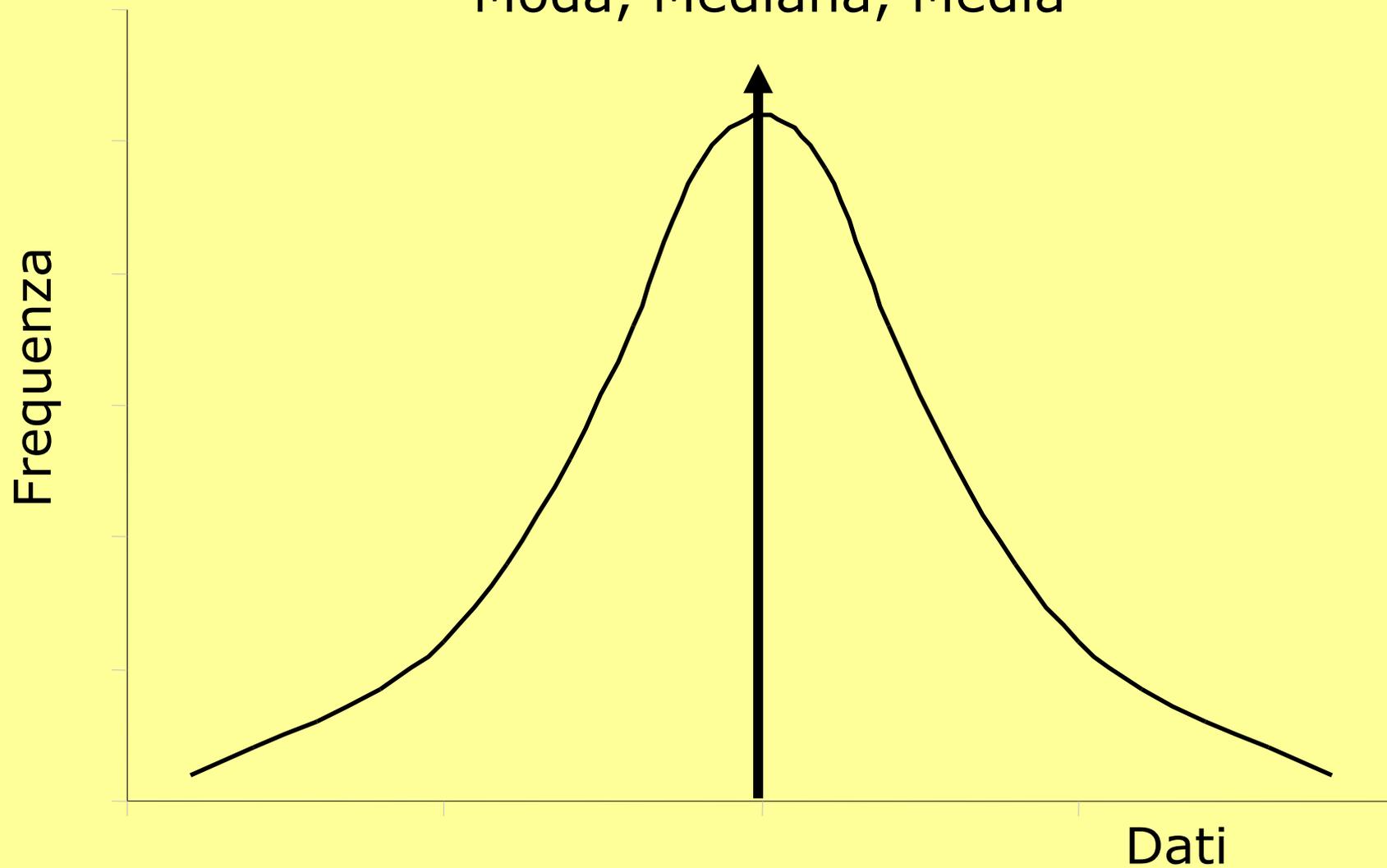
Si usa la **MEDIA** quando:

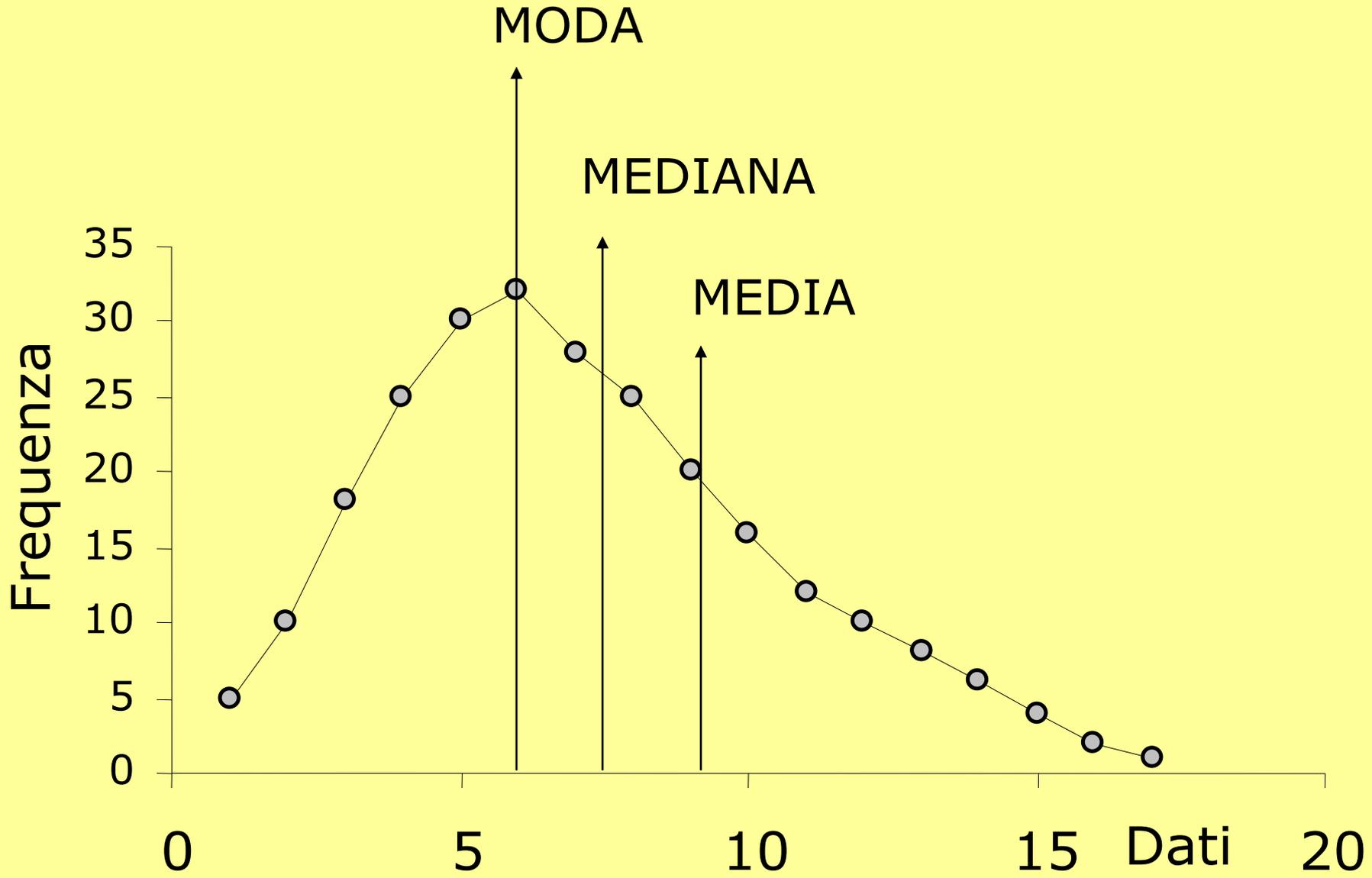
1. I dati sono normali o vicino alla normalità.
2. Sono necessarie informazioni su tutti i dati (ordine e loro valore).
3. Serve calcolare altri parametri come la deviazione standard.

Si usa la MEDIA quando:

1. I dati sono normali o vicino alla normalità.
2. Sono necessarie informazioni su tutti i dati (ordine e loro valore).
3. Serve calcolare altri parametri come la deviazione standard.

Moda, Mediana, Media





Misure di variabilità

1. Range
2. Range interquartile (IQR)
3. Varianza
- 4. Deviazione standard**

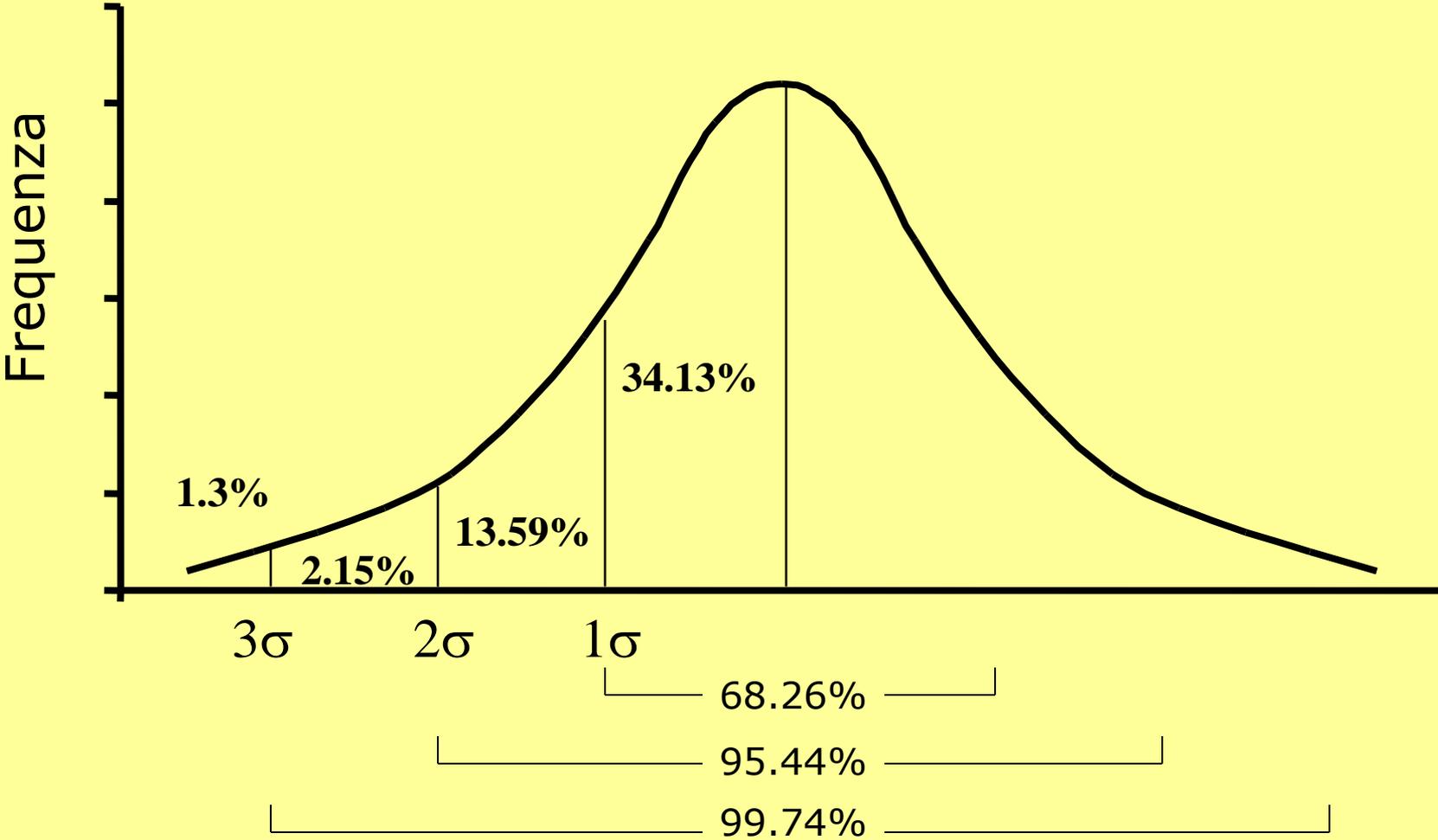
Deviazione standard (σ)

Una serie di dati può essere concentrata intorno alla misura centrale (i.e. distribuzione leptocurtica) oppure può essere molto dispersa (i.e. dispersione platicurtica).

La deviazione standard fornisce indicazioni sulla dispersione di una serie di misure.

$$\sigma = \sqrt{\frac{\sum (x_i - m)^2}{n-1}}$$

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$



Esempio

Età di 10 ± 3 (1σ) anni indica che il 68.26% dei dati si trova in un range compreso tra 7 ($X - \sigma$) e 13 ($X + \sigma$).

Età di 10 ± 6 (2σ) anni indica che il 95.44% dei dati si trova in un range compreso tra 4 ($X - 2\sigma$) e 16 ($X + 2\sigma$).

Coefficiente di variazione (CV)

CV = deviazione standard / MEDIA • 100

Media	σ	CV
120	15	13 %
200	26	13 %
1240	161	13 %

Probabilità

Se tiro un dado ho 1 probabilità su 6 che esca il numero 4

Probabilità (P) = $1/6 = 0.17 = 17\%$

Ho il 17% di probabilità che esca 4 (P=0.17)

