CDL in MEDICINA & CHIRURGIA

Statistica Medica

gbarbati@units.it

A.A. 2024-25

UNITÀ DI BIOSTATISTICA

Dipartimento Universitario Clinico di Scienze Mediche Chirurgiche e della Salute

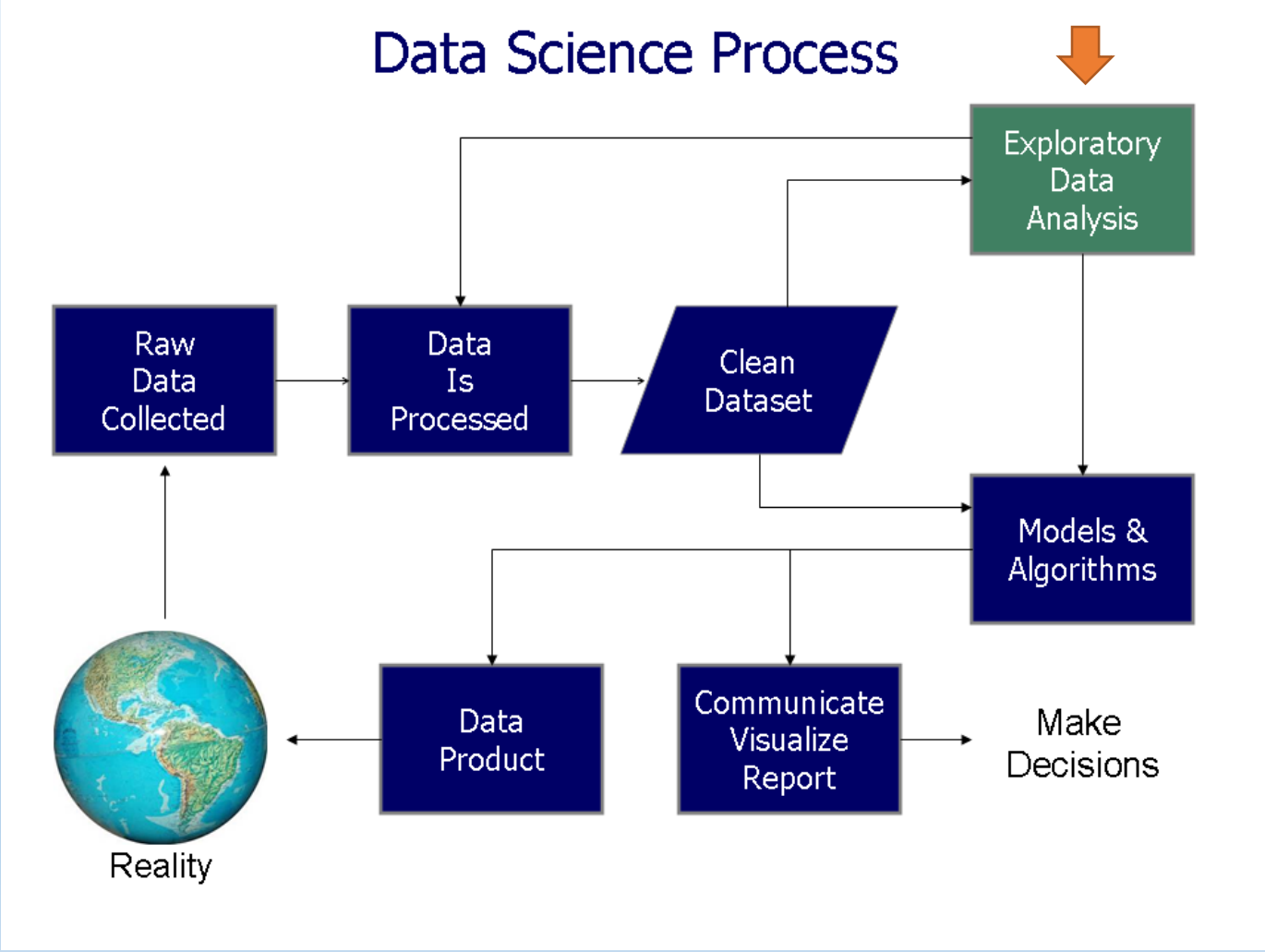
- Indici di Posizione

- Indici di Dispersione

Statistics is the grammar of science.

Karl Pearson (1857-1936).

WE ARE HERE



Gli indici di posizione delle distribuzioni

Per rappresentare in modo obiettivo una massa di informazioni un indice sintetico deve essere facilmente comprensibile, relativamente semplice da calcolare e soprattutto confrontabile con indici ricavati in tempi e luoghi diversi, sullo stesso tipo di dati.

Il dato di sintesi deve essere compreso tra il valore piu' piccolo e quello piu' grande tra quelli osservati (se è possibile ordinarli); deve identificarsi, in qualche modo, con i valori piu' frequenti, i quali corrispondono spesso a quelli localizzati *al centro* delle misure *ordinate*.

Si definiscono quindi gli **'indici di tendenza centrale'** o di **'indici di posizione'**. Un corretto approccio al problema richiede un'analisi preventiva della scala di misura con cui sono espresse le modalità del carattere al fine di scegliere **il tipo di indice** di posizione opportuno da utilizzare.

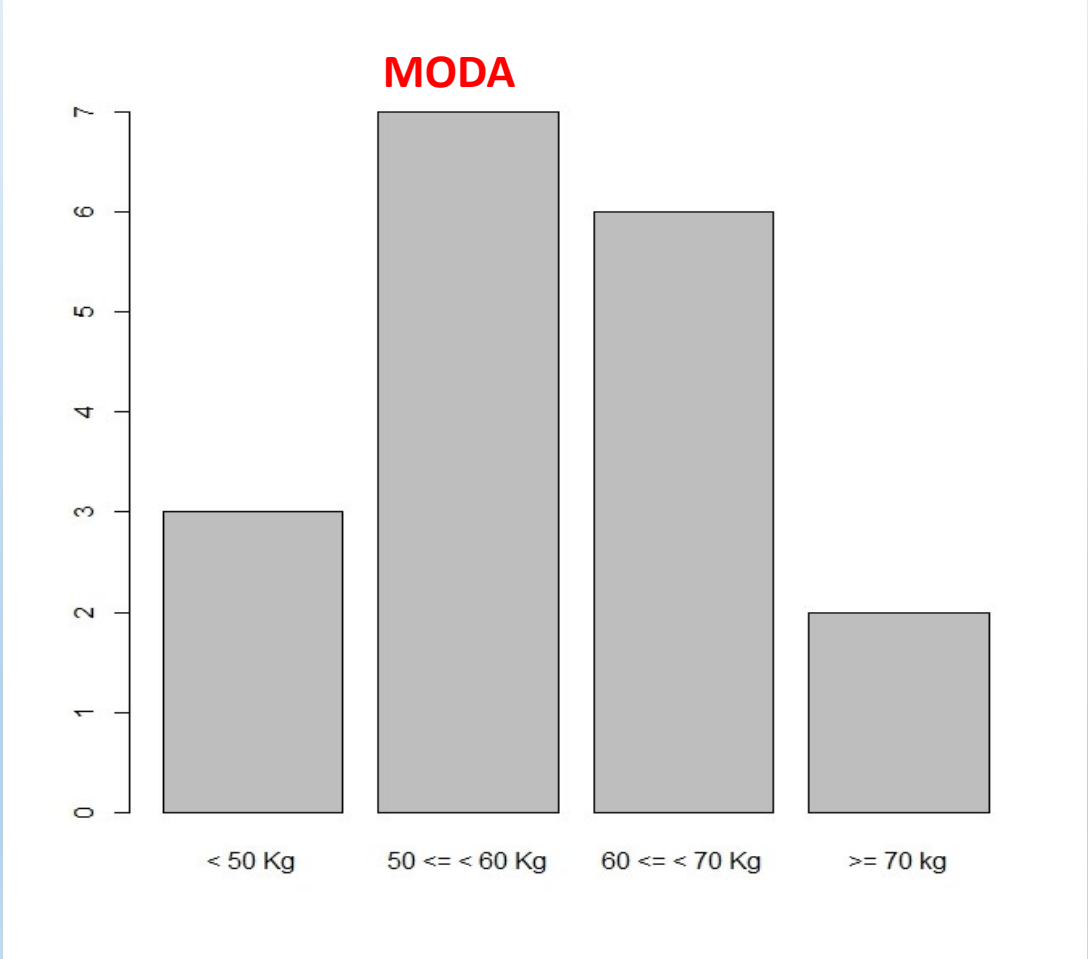
Gli indici di posizione delle distribuzioni (I): MODA

MODA:

Dato un qualsiasi tipo di carattere (qualitativo o quantitativo) la **MODA** della popolazione distribuita secondo quel carattere è la **modalità prevalente** del carattere, ossia quella a cui è associata la massima frequenza.

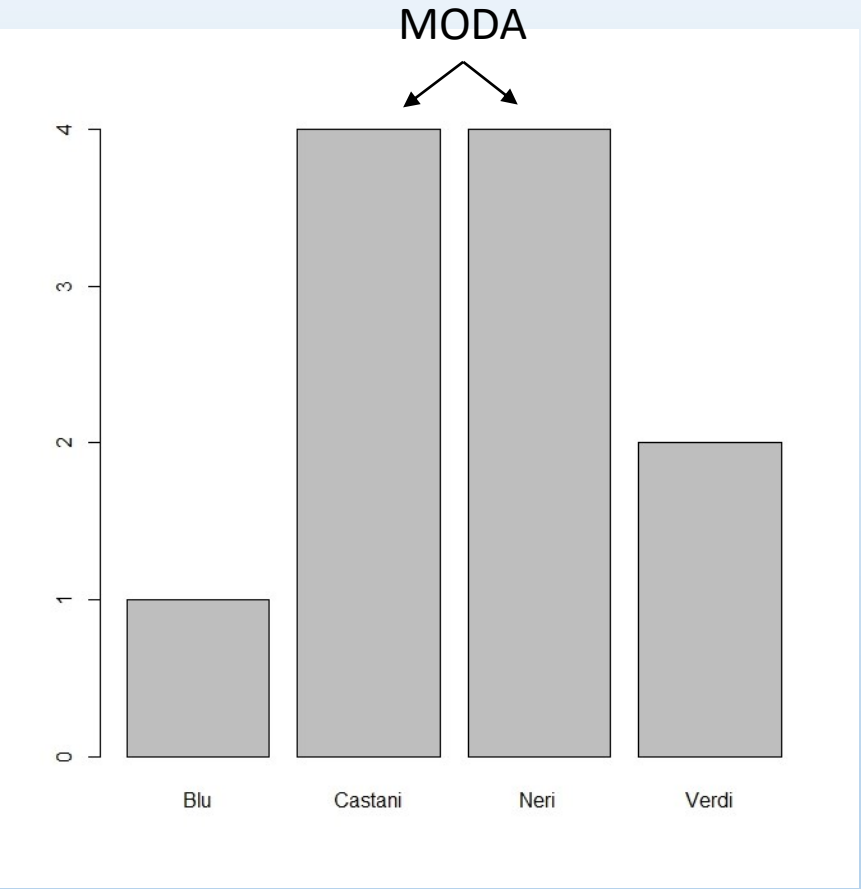
Classe di peso (kg)	**Frequenza di studenti nella classe di peso**
< 50	3
50 ≤ < 60 (Moda o Classe Modale)	**7**
60 ≤ <70	6
≥ 70	2
Tot	18

Non è detto che vi sia un'unica moda in una distribuzione!



Gli indici di posizione delle distribuzioni (I): MODA

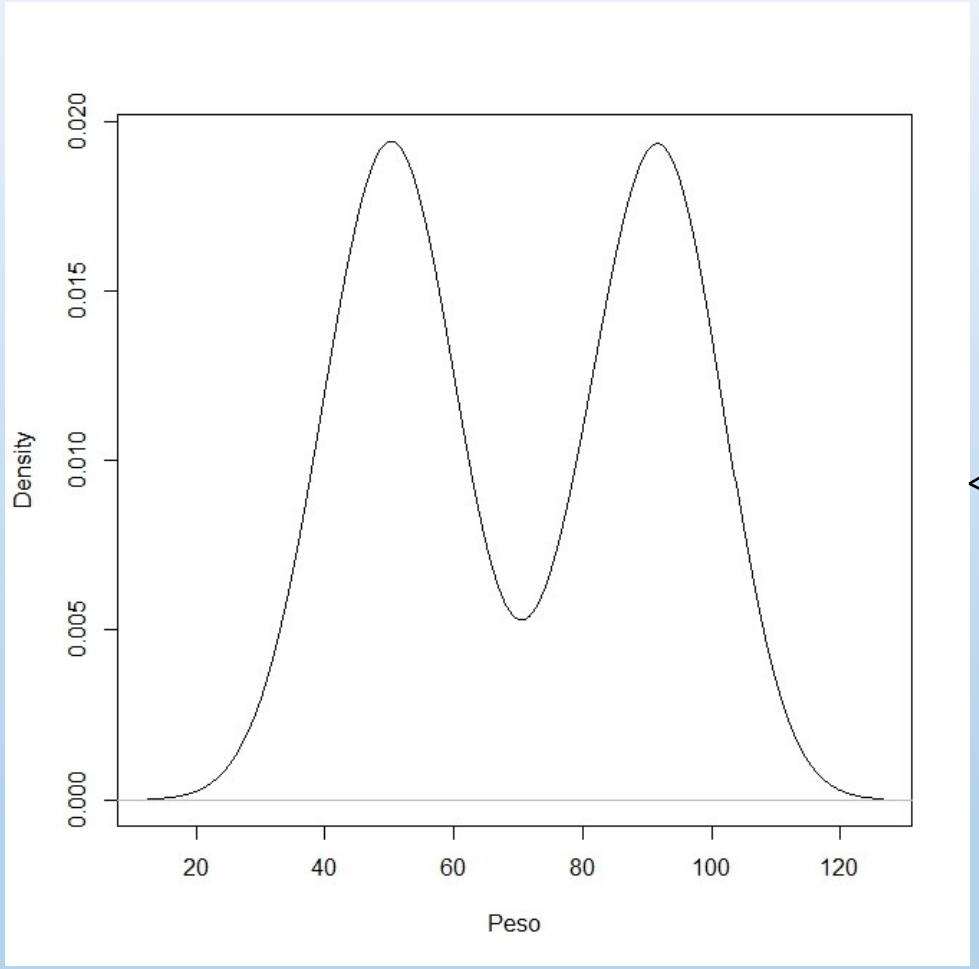
Colore degli occhi	Frequenza di studenti con quel colore di occhi
Blu	1
Castano (Moda o Classe Modale)	**4**
Nero (Moda o Classe Modale)	**4**
Verde	2
totale	11

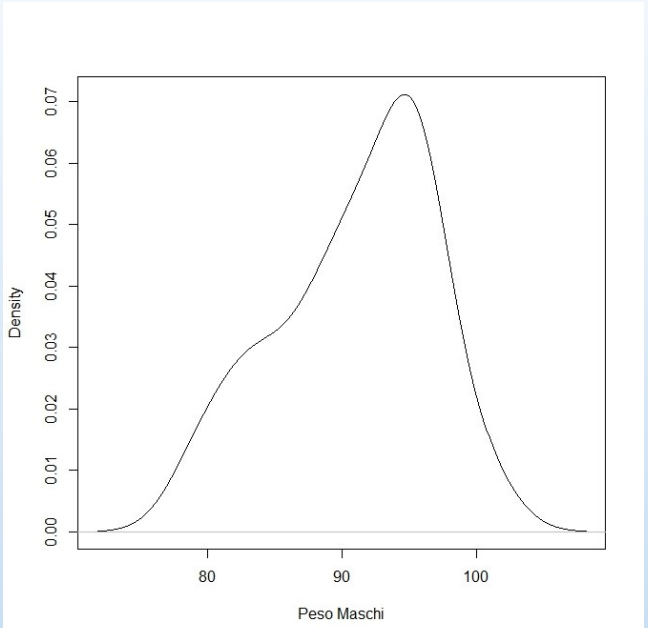


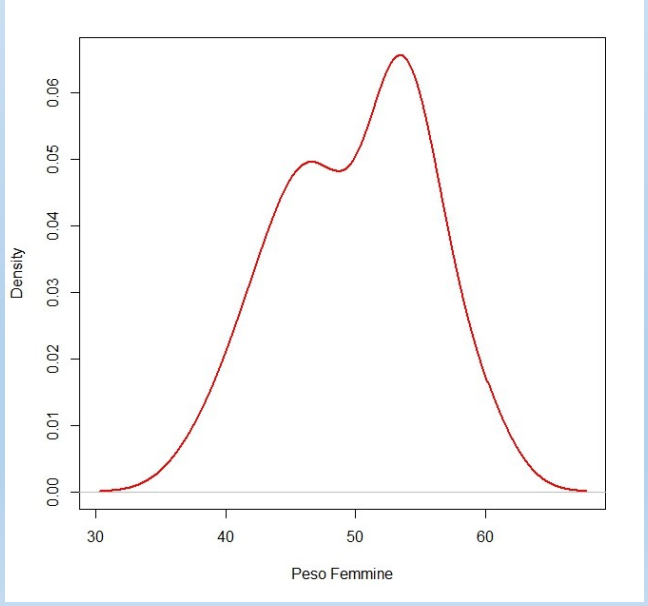
…in questo caso la distribuzione viene definita *bi-modale*, cioè ha due mode.

Per quanto riguarda i **caratteri qualitativi sconnessi** la moda è l'**unico** indice
di posizione che si puo' calcolare.

Gli indici di posizione delle distribuzioni (I): MODA



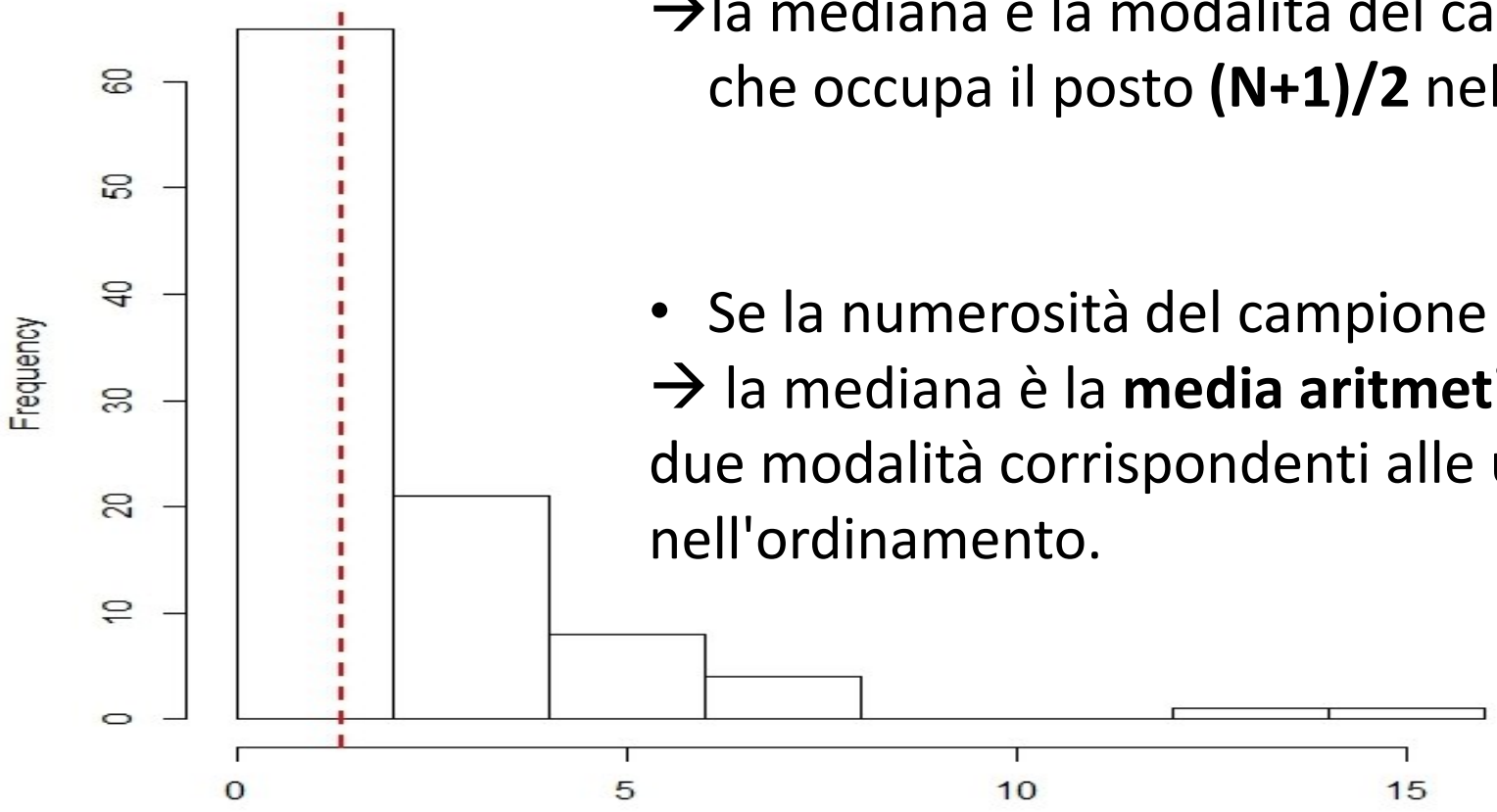




La moda ci aiuta a capire se la distribuzione è **omogenea** oppure no!

Gli indici di posizione delle distribuzioni (II): MEDIANA

Dato un **carattere ordinabile** la **MEDIANA** della distribuzione è la modalità del carattere che bi-partisce (=divide in *due parti uguali*) la distribuzione.



- Se la numerosità del campione **N** è dispari,

→la mediana è la modalità del carattere associata all'unità che occupa il posto **(N+1)/2** nell'ordinamento;

- Se la numerosità del campione **N** è pari,

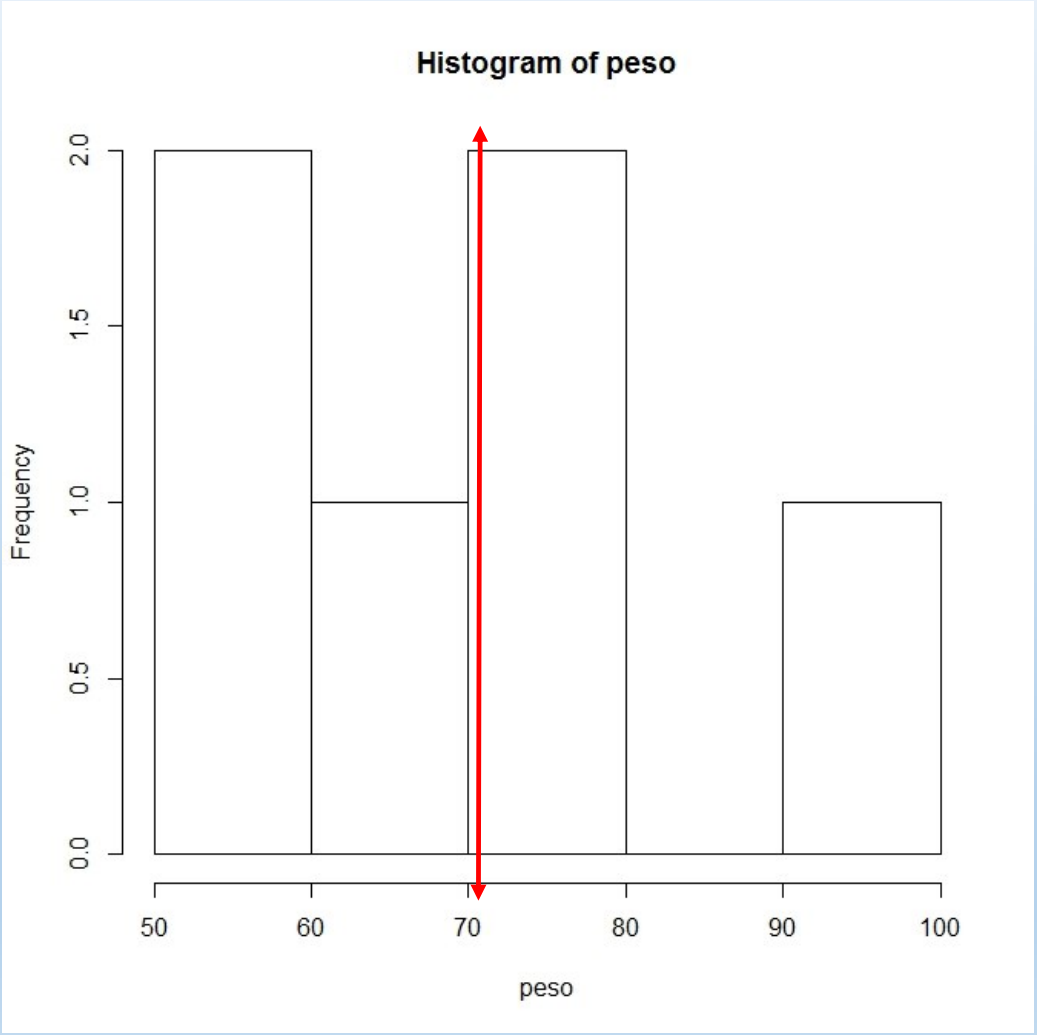
→ la mediana è la **media aritmetica** dei valori assunti dalle due modalità corrispondenti alle unità centrali nell'ordinamento.

Gli indici di posizione delle distribuzioni (II): MEDIANA

SOGGETTO	PESO (kg)	Rango
SOGGETTO 1	55	2
SOGGETTO 2	78	5
SOGGETTO 3	52	1
SOGGETTO 4	**67**	**3**
SOGGETTO 5	91	6
SOGGETTO 6	**76**	**4**

→ la mediana è: (67+76)/2=71,5 Kg

Il 50% del campione esaminato ha un peso corporeo inferiore o uguale a 71,5 Kg (*sintesi* dei dati).

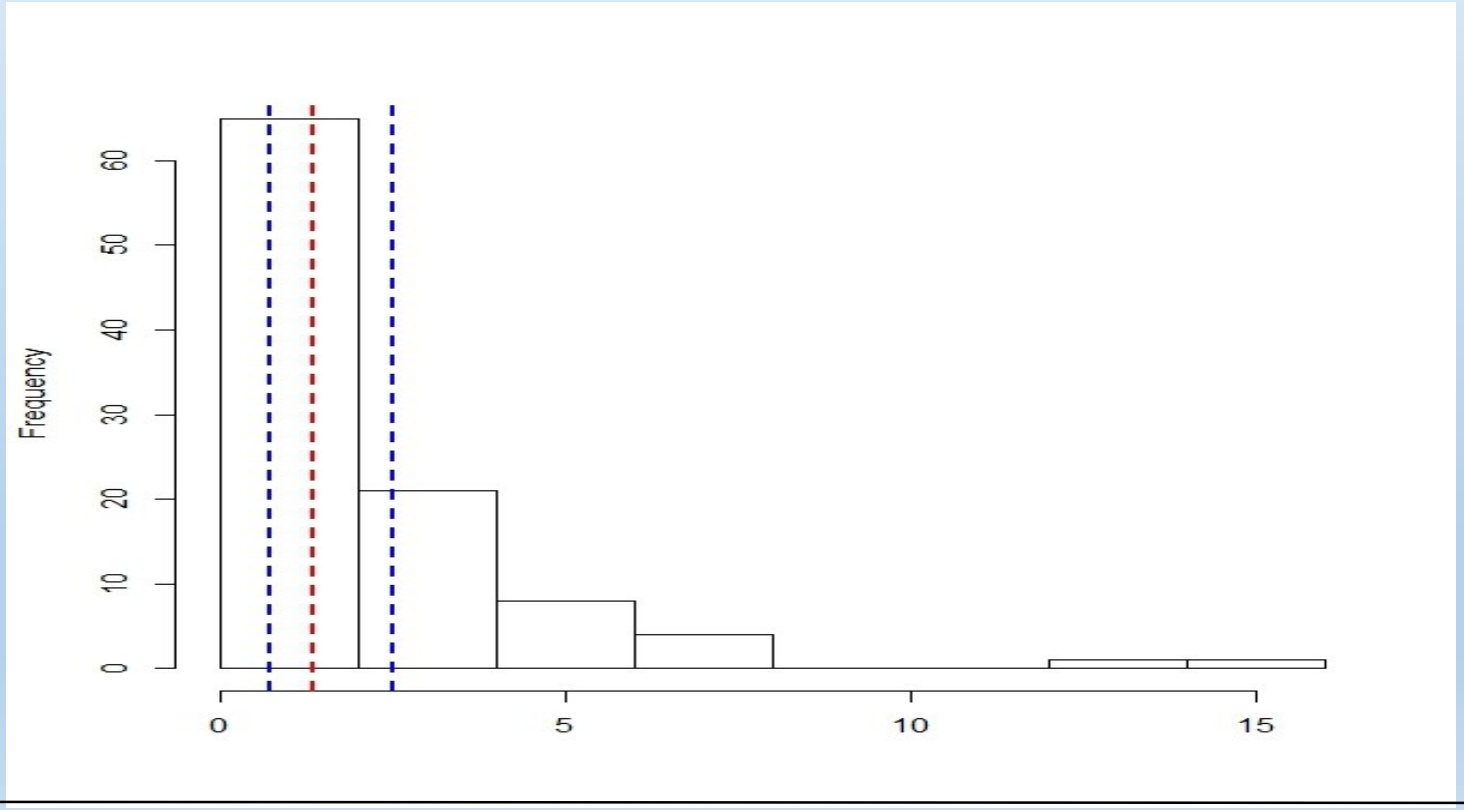


Gli indici di posizione delle distribuzioni (III): Quantili/Percentili

I '**quantili**' o '**percentili**' identificano alcuni indici di posizione che altro non sono che un'estensione del concetto di mediana: suddividono in parti uguali una serie ordinata di dati.

I '**quartili**' sono gli indici di posizione che dividono una serie ordinata di dati in **4** parti uguali:

Q1= primo quartile:
25% della distribuzione
Q3= terzo quartile:
75% della distribuzione



Nel caso di caratteri qualitativi ordinabili o quantitativi è possibile definire accanto alle frequenze assolute e relative (e percentuali), le **frequenze cumulate assolute e relative** (e percentuali).

Considerata una distribuzione di frequenze, siano $x_1, x_2, \dots, x_K$ le **modalità** assunte (ordinate in ordine crescente) da un carattere qualitativo ordinabile o quantitativo sulle N unità di un collettivo:

X	n_i	f_i	N_i	F_i
x_1	n_1	f_1	$N_1 = n_1$	$F_1 = f_1$
x_2	n_2	f_2	$N_2 = n_1+n_2$	$F_2 = f_1+f_2$
x_i	n_i	f_i	$N_i = n_1+n_2+\dots+n_i$	$F_i = f_1+f_2+\dots+f_i$
x_K	n_K	f_K	$N_K = n_1+n_2+n_i+\dots n_k = N$	$F_K = f_1+f_2+f_i+\dots f_k = 1$
Totale	**N**	**1**		

La frequenza assoluta cumulata N_i misura quante unità del collettivo osservato possiedono o la modalità x_1 o la modalità x_2 … o «fino alla» modalità x_i.

Per i=1 abbiamo che N_1 è esattamente uguale alla frequenza assoluta della prima modalità.

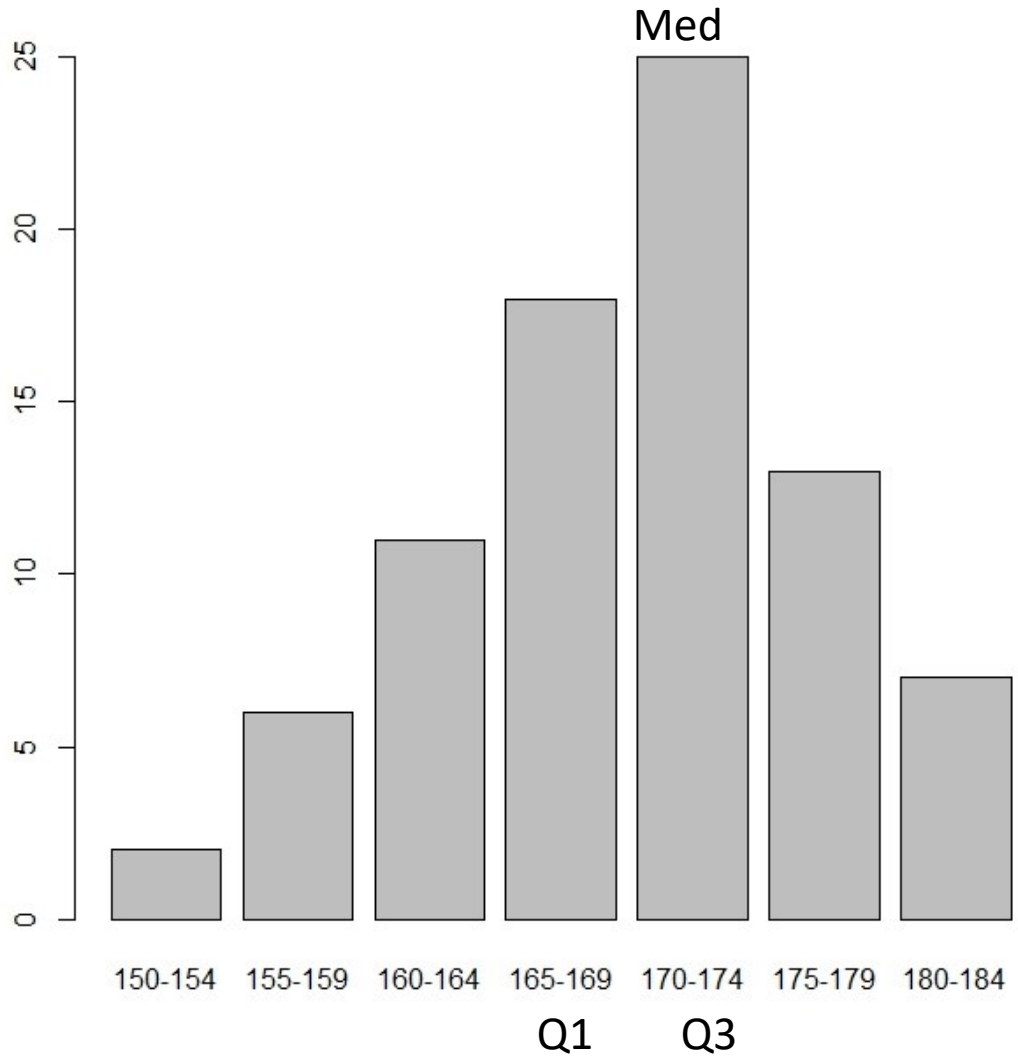
Per i=K abbiamo che N_K è uguale a tutta la numerosità del collettivo (N).

FREQUENZE CUMULATE ASSOLUTE

FREQUENZE CUMULATE RELATIVE: N_i/N

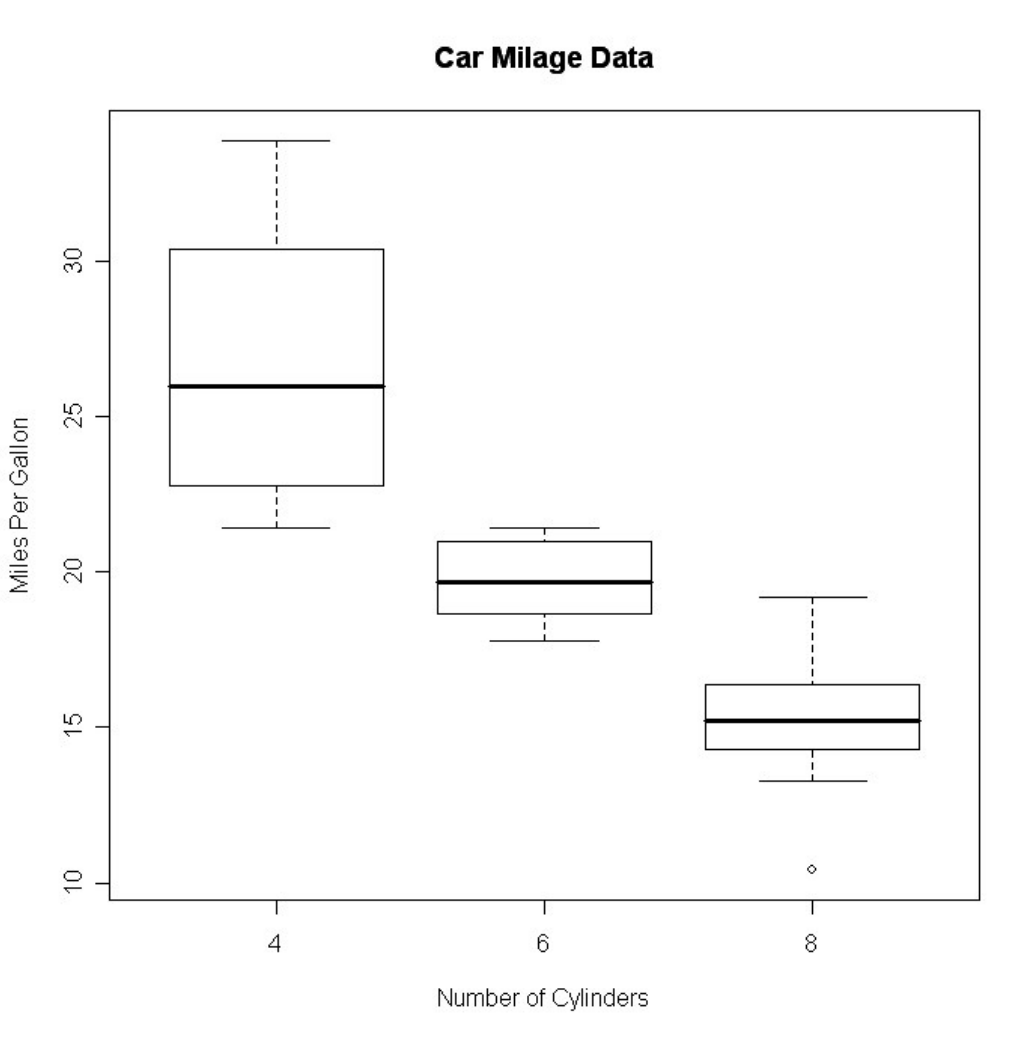
Gli indici di posizione delle distribuzioni (III): Quantili/Percentili

Classi di altezza (cm)	Frequenze assolute	Frequenze cumulate assolute	Frequenze cumulate percentuali (%)
150 – 154	2	2	2
155 – 159	6	8	10
160 – 164	11	19	23
165 – 169 **Q_1**	**18**	**37**	**45** **(contiene il 25%)**
170 – 174 **Q_2** **Q_3**	**25**	**62**	**76** **(contiene il 50%)** **(contiene il 75%)**
175 – 179	13	75	91
180 - 184	7	82	100
totale	82		



Rappresentazioni grafiche dei caratteri su scala numerica:

«Box plot»: una rappresentazione **sintetica** della distribuzione



Il box plot o *diagramma a scatola e baffi*, è un grafico ottenuto a partire dai 5 numeri di sintesi :

- «Minimo»
- 1° quartile (Q1)
- Mediana
- 3° quartile (Q3)
- «Massimo»

che descrivono le caratteristiche salienti della distribuzione.

N.B: Il box plot offre **una rappresentazione univoca** della distribuzione, a differenza dell'istogramma che può offrire rappresentazioni grafiche diverse a seconda degli estremi delle classi scelte.

I baffi : Q3 + 1.5*(IQR); Q1 – 1.5*(IQR) [IQR= Q3-Q1.] Dovrebbero includere il 99.3% dei dati, se provenissero da una distribuzione **gaussiana…**

Gli indici di posizione delle distribuzioni (IV): Media aritmetica

Per i caratteri quantitativi è possibile calcolare anche un altro indice di posizione: **la media aritmetica**.
Rilevate su *n* unità di una popolazione le modalità di un certo carattere quantitativo X:
$x_1, x_2, \ldots, x_n$

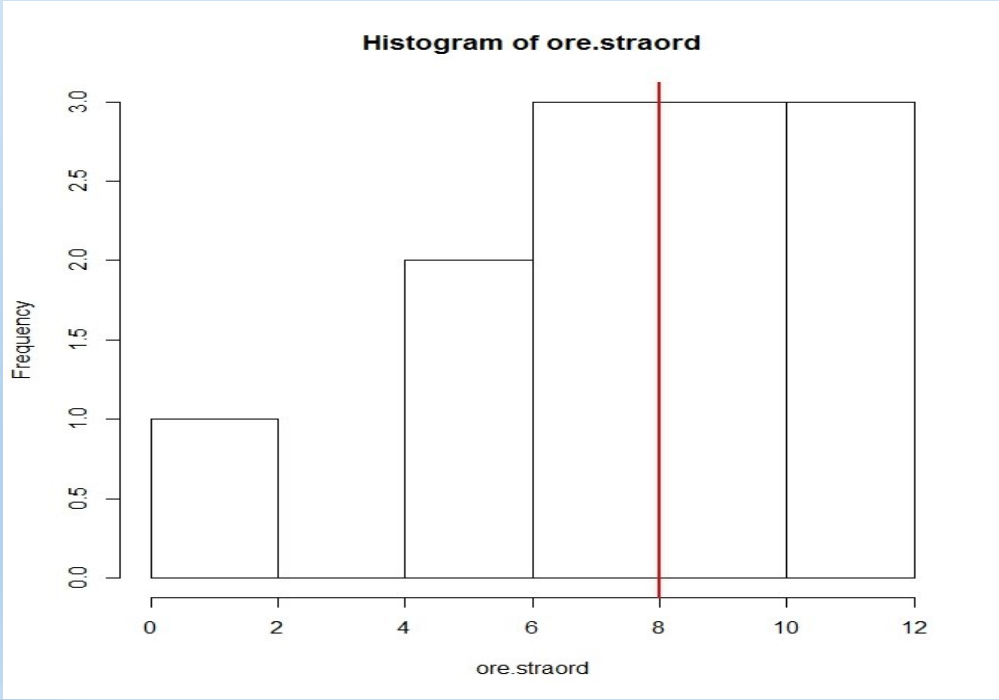
$$\bar{x} = \frac{(x_1 + x_2 + \ldots + x_n)}{n} \quad \Longleftrightarrow \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Un infermiere ha fatto delle ore di lavoro straordinario da gennaio a dicembre:

10, 12, 11, 5, 7, 10, 5, 0, 7, 10, 7, 12.

Qual è il numero medio mensile di ore di straordinario?

$$\bar{x} = \frac{(10+12+11+5+7+10+5+0+7+12)}{12} = \frac{96}{12} = 8$$



Calcolo della media aritmetica da una distribuzione di frequenze

Distribuzioni di frequenza

Variabile x	Frequenze assolute	Frequenze cumulate	Frequenze relative	Frequenze %
x_1	n_1	n_1	n_1/N	$n_1/N*100$
x_2	n_2	n_1+n_2	n_2/N	$n_2/N*100$
…	…	…	…	…
x_k	n_k	$n_1+ \ldots.+n_k=N$	n_k/N	$n_k/N*100$
totale	N		1	100

$$M = \frac{x_1 * n_1 + \cdots x_k * n_k}{N} = \frac{\sum_{i=1}^{k} x_i * n_i}{N}$$

Variabile x	Frequenze assolute
20	5
21	2
totale	7

$$M = \frac{20 \cdot 5 + 21 \cdot 2}{7}$$

M= 20.29

Calcolo della media aritmetica da una distribuzione in classi di frequenza

Valori centrali delle classi	Frequenze assolute
c_1	n_1
c_2	n_2
...	...
c_k	n_k
totale	N

$$M = \frac{c_1 * n_1 + \cdots c_k * n_k}{N} = \frac{\sum_{i=1}^{k} c_i * n_i}{N}$$

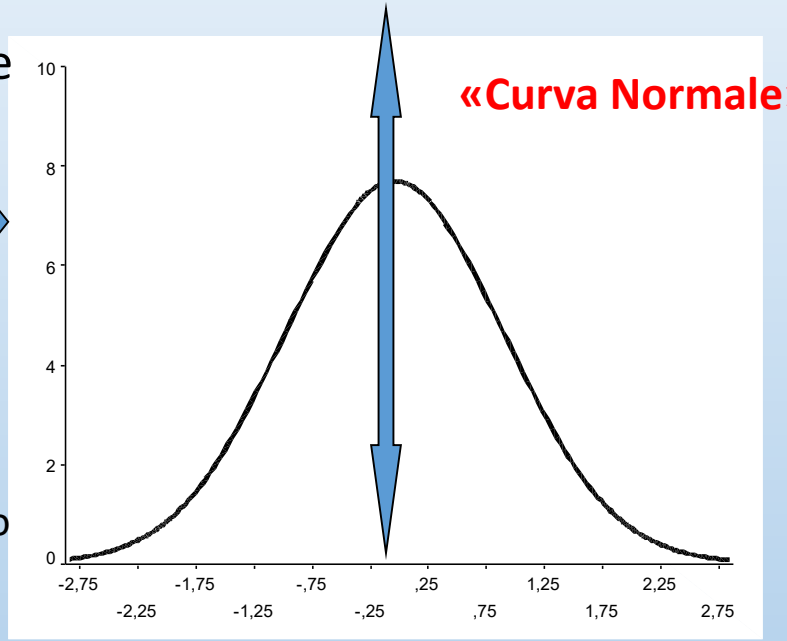
classi	Frequenze assolute	Valori centrali delle classi
[18; 22]	20	20
(22; 26]	30	24.5
(26; 30]	50	28.5
totale	100	

$$M = \frac{20 * 20 + 24.5 * 30 + 28.5 * 50}{100}$$

M=25.6

Moda, **mediana**, **quartili** e **media** sono gli indici di posizione di piu' frequente impiego.
Distribuzioni simmetriche unimodali : moda=mediana=media.
'**Simmetrico**' : una distribuzione simmetrica rispetto al valor medio: ha (circa) la stessa quantità di osservazioni inferiori alla media e superiori alla media:

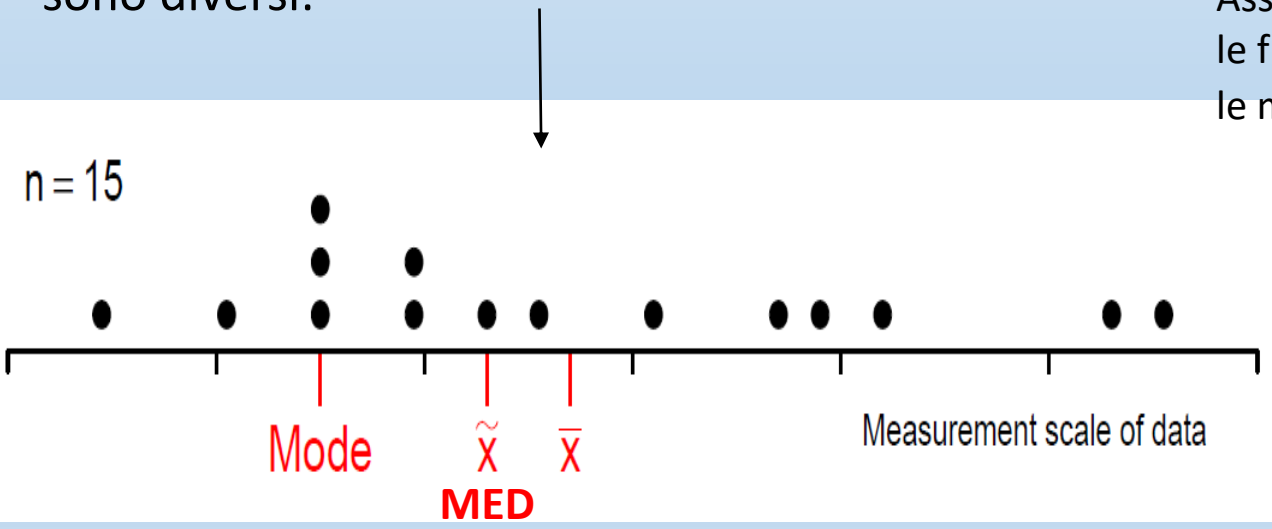
Ex: distribuzione simmetrica e unimodale
Moda=Mediana=Media



Asse verticale (y):
le frequenze con cui
le modalità si presentano

asse orizzontale (x): i valori della distribuzione
(cioè le modalità del carattere quantitativo in studio)

In caso di **distribuzione asimmetrica** invece i tre indici sono diversi:



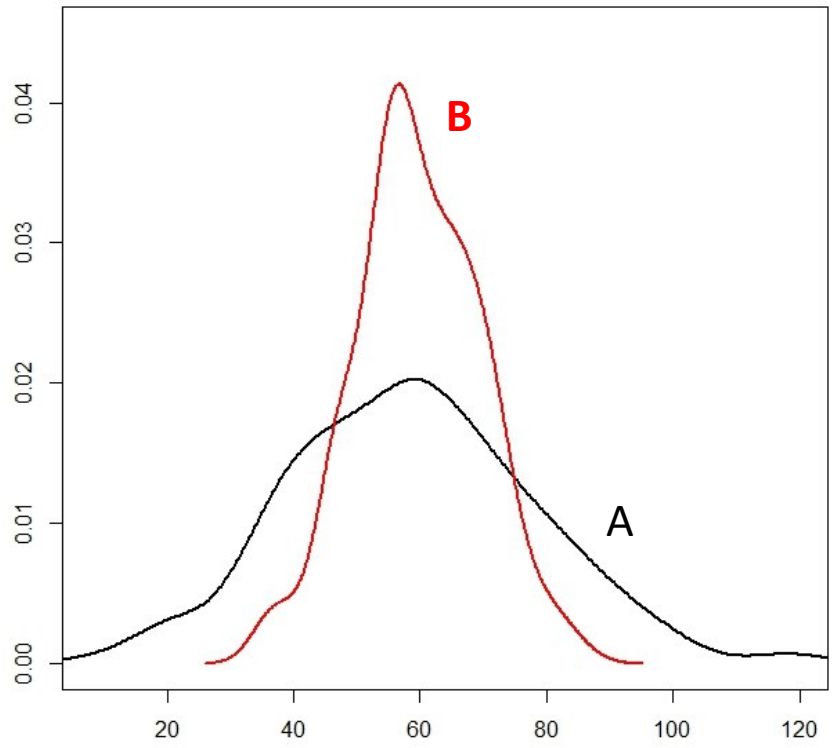
Gli indici di dispersione

La posizione è *rappresentativa* di un fenomeno; tuttavia da sola *non basta* per definire la distribuzione.
Occorrono criteri aggiuntivi per quantificare la **variabilità** delle misure rispetto ad un termine di riferimento.

La ***variabilità*** delle misure puo' essere valutata:

(a) in base alla loro ***oscillazione*** o ***dispersione*** rispetto, per esempio, al valore medio;

(b) come **distanza** tra due particolari modalità della distribuzione.



…A e B non sembrano simili…

(I): Dispersione intorno al valor medio

Variabilità intorno al valore medio di una distribuzione di **valori quantitativi**: $x_1, \ldots x_n$ di media $\bar{x}$

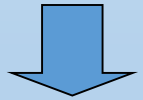
consideriamo gli *scarti* (cioè le differenze) di tutte le misure dalla media:

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

Per proprietà della media aritmetica, la sommatoria degli scarti dalla media è sempre nulla:

(a causa della compensazione tra scarti positivi e scarti negativi).

$$\sum_{i=1}^{n} \left(x_i - \bar{x} \right) = 0$$



Si definisce '**devianza**' la somma dei quadrati degli scarti dalla media:

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

(I): Dispersione intorno al valor medio

La devianza non contiene pero' l'informazione del numero di osservazioni utilizzate nel calcolo.

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

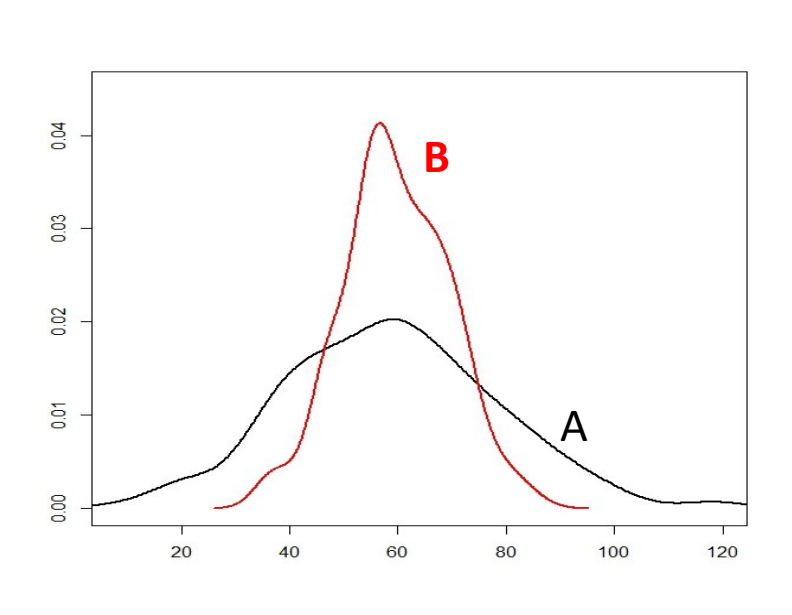
Le varianze diventano cosi' *confrontabili* tra diverse distribuzioni e si puo' stabilire quale, tra le due o piu' serie di misure considerate, presenta una maggiore dispersione rispetto alla media, **indipendentemente da N**.

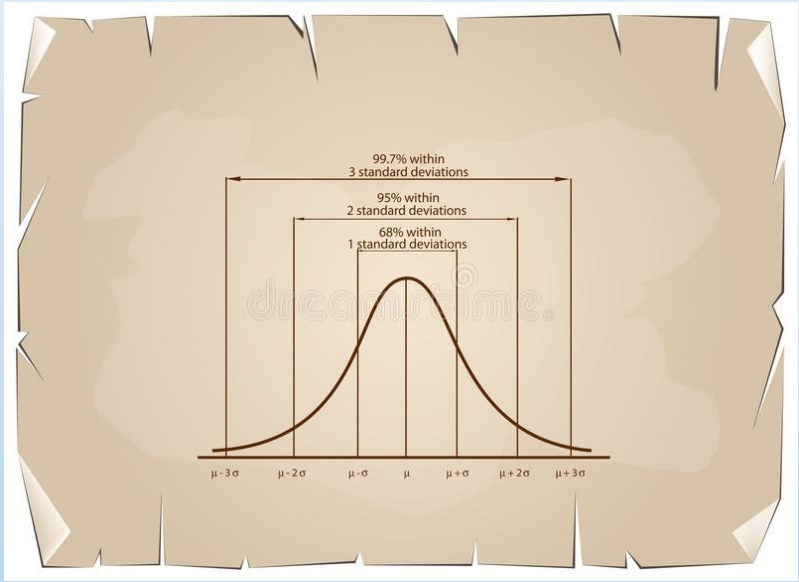
$$dev\ st = s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

Per tornare all'unità di misura del fenomeno si usa *la radice quadrata* => deviazione standard

(I): Dispersione intorno al valor medio

Deviazione standard: errore che si commette *mediamente* considerando il valore medio al posto di ogni singolo valore della distribuzione.





E' **<u>MOLTO IMPORTANTE</u>** calcolare la deviazione standard dalla media perchè è un indice fondamentale per capire se i dati che stiamo osservando siano ***ben sintetizzati*** dalla media oppure no.

Quanto piu' è alta infatti la deviazione standard, tanto meno è informativa la media, perchè i dati si allontanano molto da essa.

Excursus: perché N-1 ?

In *piccoli* studi si riduce la probabilità di includere dati particolarmente "*dispersi*" e quindi si può *sottostimare* la variabilità reale della popolazione.

Queste considerazioni giustificano, per N «piccolo» (<50) di modificare la formula della varianza:

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

Motivo statistico-matematico: se di una distribuzione di dati è nota la media, la conoscenza di uno degli N valori è superflua in quanto ricavabile utilizzando la formula del calcolo della media stessa.

Ne deriva che questa misura, pur essendo indispensabile per determinare la media, non è '**indipendente**' dalle altre, in quanto è implicita in esse e non fornisce ulteriore informazione; per tale motivo viene trascurata nel calcolo della varianza.

Le N-1 osservazioni *indipendenti*, costituiscono un caso particolare di ***gradi di libertà***: un concetto statistico importante, soprattutto in inferenza, per indicare, nelle varie situazioni, il **numero delle *informazioni indipendenti***.

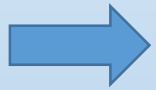
Esempio di calcolo della deviazione standard (distribuzione di frequenze)

Variabile x	Frequenze assolute
x_1	n_1
x_2	n_2
…	…
x_k	n_k
totale	N

$$dev\ st = s = \sqrt{\frac{\sum_{i=1}^{k} n_i * (x_i - M)^2}{N-1}}$$

Variabile x	Frequenze assolute
20	3
22	4
26	5
totale	12

$$M = \frac{20 * 3 + 22 * 4 + 26 * 5}{12}$$



M=23.17

Variabile X	Frequenze assolute	(x-M)	(x-M)**2	freq*(x-m)**2
20	3	-3,17	10,05	30,15
22	4	-1,17	1,37	5,48
26	5	2,83	8,01	40,04
Totale	12		Varianza:	6,88
			Dev std:	**2,62**

Gli indici di dispersione (II): distanza

La dispersione come **distanza** tra due particolari modalità della distribuzione:

a) Distanza tra il valore minimo ed il valore massimo

b) Distanza tra il *primo* ed il *terzo* quartile della distribuzione

a) Data una distribuzione di valori $x_1, x_2, x_3 \ldots, x_n$ di un **carattere qualitativo ordinabile** o **quantitativo**, ordinata in senso crescente: $x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}$ si definisce:

INTERVALLO DI VARIAZIONE (**RANGE**): distanza tra il valore minimo $x_{(1)}$ ed il valore massimo $x_{(n)}$ della distribuzione ordinata:

$$\textbf{RANGE} = x_{(n)} - x_{(1)}$$

b) Data una distribuzione di valori $x_1, x_2, x_3 \ldots, x_n$ di un **carattere qualitativo ordinabile** o **quantitativo**, si definisce:

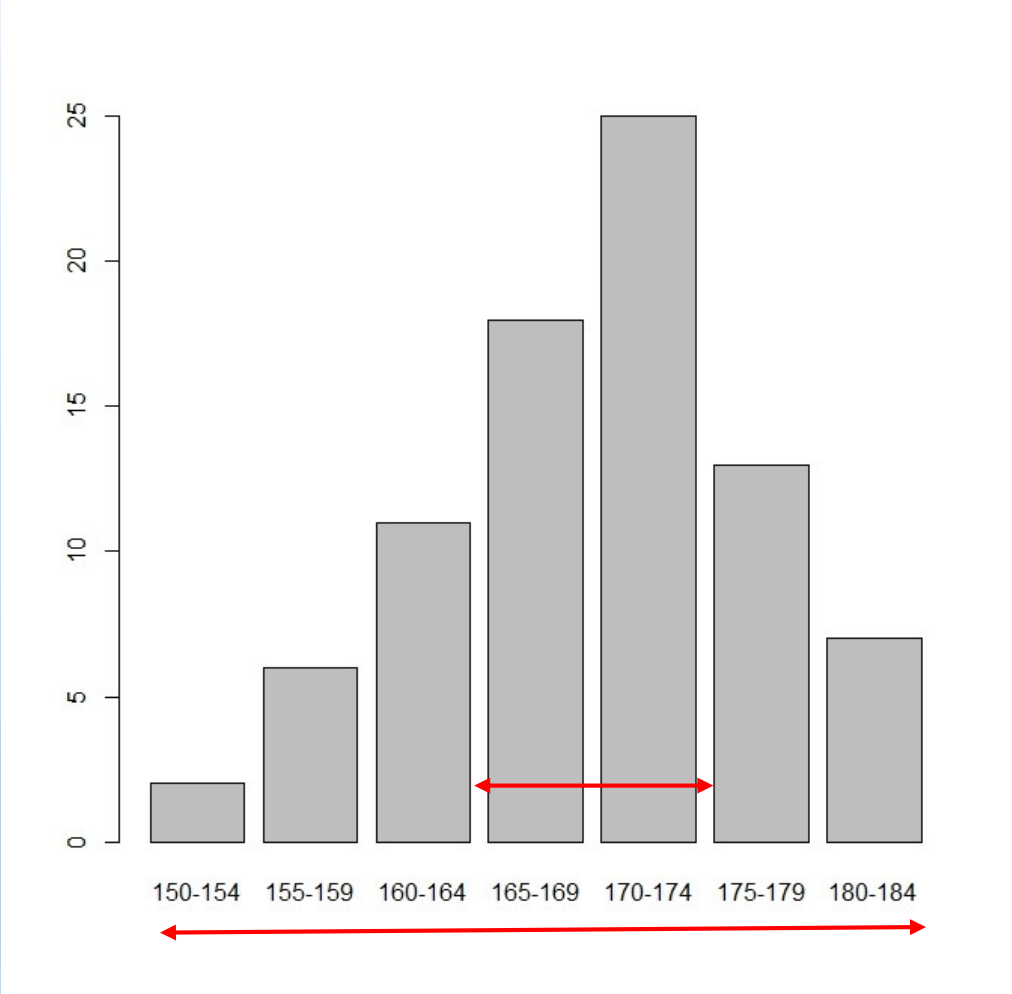
DIFFERENZA INTERQUARTILE (**RANGE INTERQUARTILE, IQR**) la distanza tra il primo ed il terzo quartile:

$$\textbf{RANGE INTERQUARTILE} = Q_3 - Q_1$$

Gli indici di dispersione (II): distanza

Questi indici sono detti ***ASSOLUTI*** perché sono espressi nella stessa unità di misura del carattere.

Classi di altezza (cm)	Frequenze assolute	Frequenze cumulate assolute	Frequenze cumulate percentuali (%)
150 – 154	2	2	2
155 – 159	6	8	10
160 – 164	11	19	23
165 – 169	18	37	45
170 – 174	25	62	76
175 – 179	13	75	91
180 - 184	7	82	100
totale	82		



RANGE = 184 – 150 = 34 cm

Q1=165-169 e Q3=170-174

IQR= 172 – 167 = 5 cm (per convenzione si può utilizzare il valor medio delle classi)

Gli indici di dispersione (III): il coefficiente di variazione

Per confrontare la dispersione **di due o piu'** distribuzioni si può ricorrere al '**coefficiente di variazione**':

$$CV\% = \frac{s}{\bar{x}} * 100$$

CV%= Rapporto tra la deviazione standard e la media in %.

'Numero puro' (non espresso in una determinata unità di misura) e quindi confrontabile con altri.

Utile anche per fenomeni che abbiano ***una media molto diversa*** (pur avendo la stessa unità di misura).

Ex: verificare la precisione di analisi di laboratorio, chimico-cliniche:

- variabilità **intra** operatore
- **inter** operatori
- **inter/intra** laboratori

Gli indici di dispersione (III): il coefficiente di variazione

Quale tra glicemia e calcemia è piu' dispersa rispetto alla media?

$m_{glicemia}$ = 85 mg/100 ml
$s_{glicemia}$ = 11 mg/100 ml
$m_{calcemia}$ = 9 mg/100 ml
$s_{calcemia}$ = 1,5 mg/100 ml

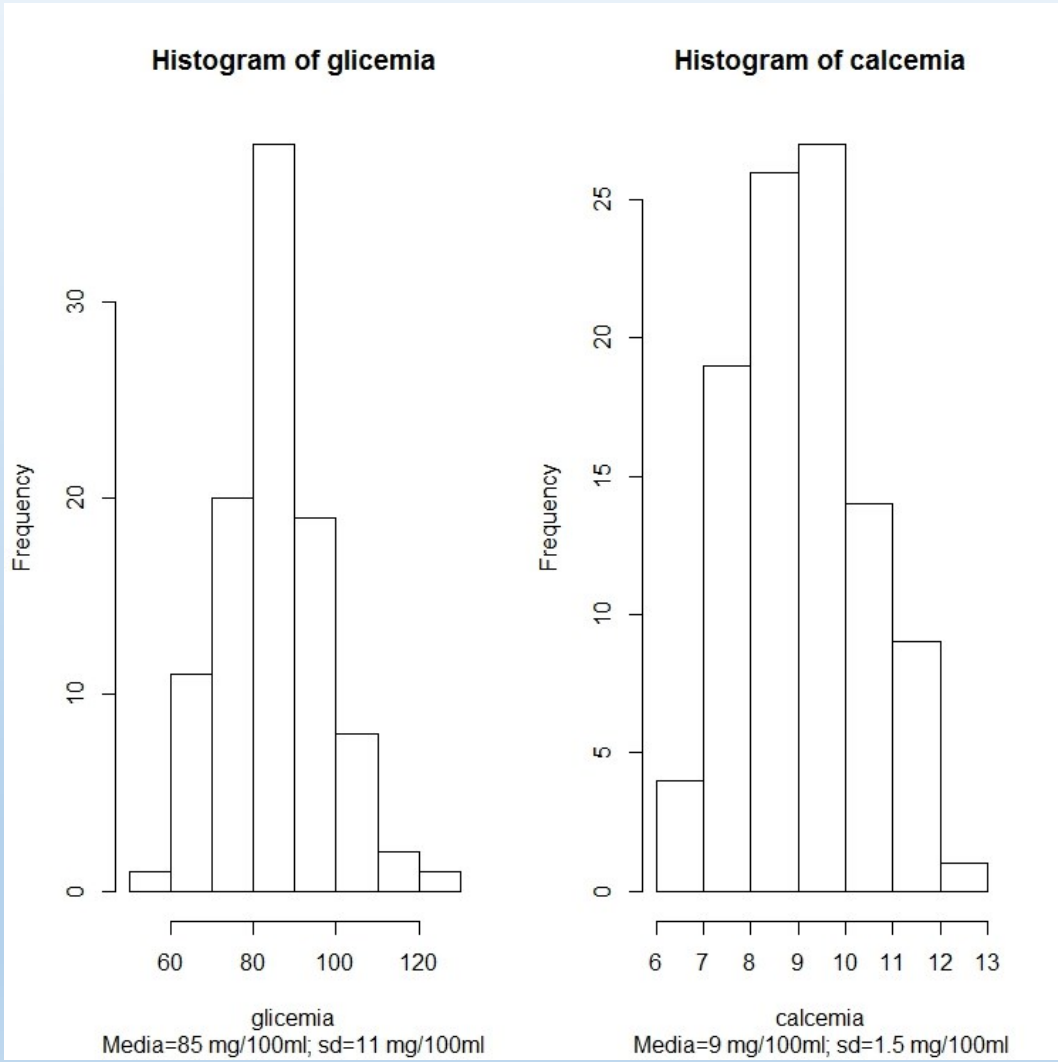
CV di Glicemia

$$\frac{11\,mg/100\,ml}{85\,mg/100\,ml} * 100 = 12.9\%$$

CV di Calcemia

$$\frac{1.5\,mg/100\,ml}{9\,mg/100\,ml} * 100 = 16.7\%$$

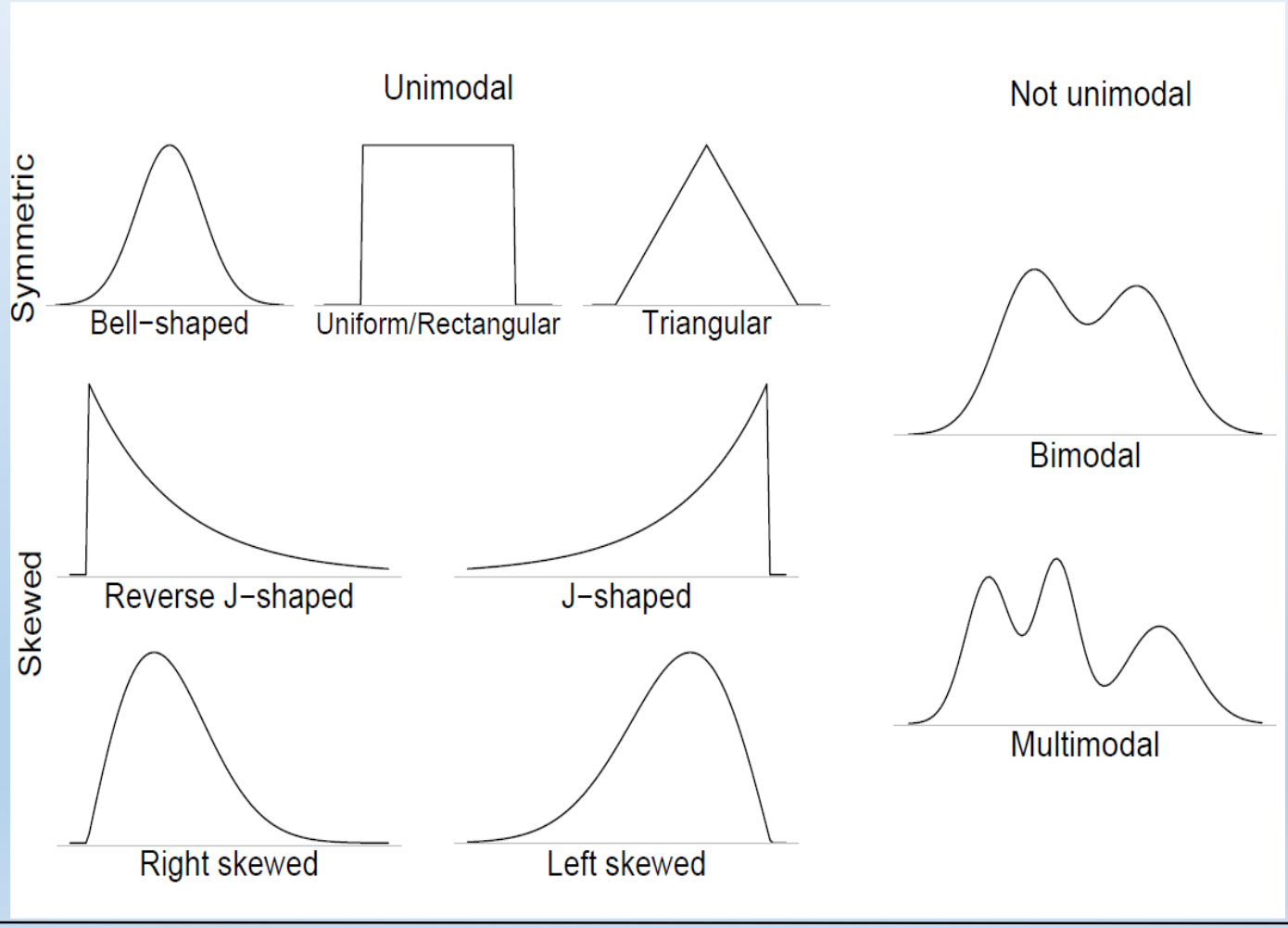
La calcemia ha un grado di dispersione maggiore della glicemia.



E' necessario visualizzare i dati per scegliere un opportuno indice di posizione.
Per distribuzioni asimmetriche è preferibile usare la MEDIANA perché la MEDIA risente eccessivamente di valori *anomali* (estremamente grandi o estremamente piccoli).

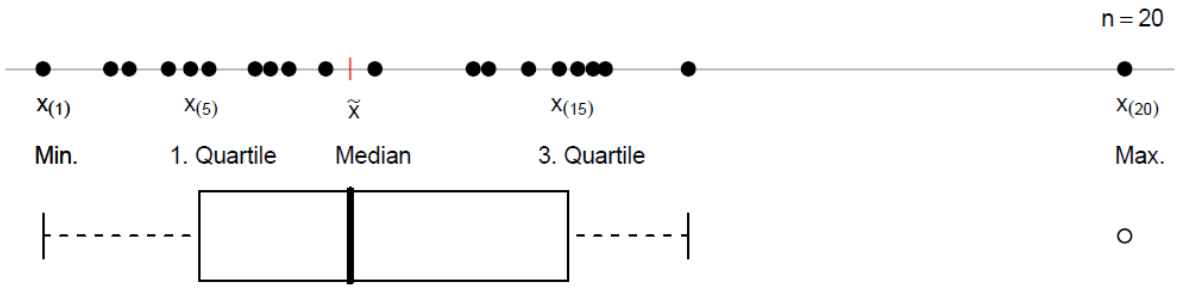
Quali indici di posizione sono calcolabili per i vari tipi di caratteri:

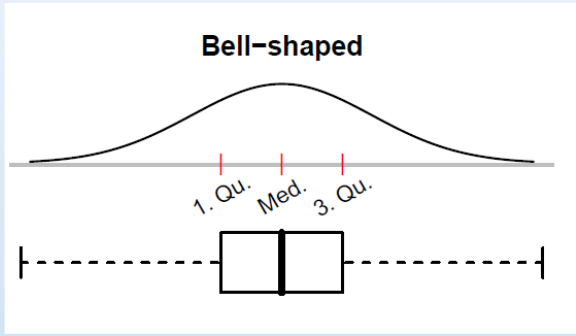
CARATTERE		INDICE DI POSIZIONE
qualitativo	Sconnesso	Moda
	Ordinabile	Moda, mediana, quartili
quantitativo		Moda, mediana, quartili, media

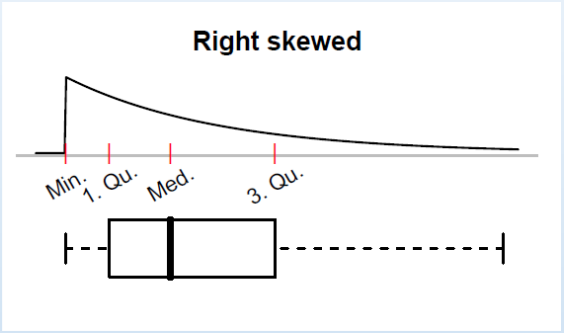


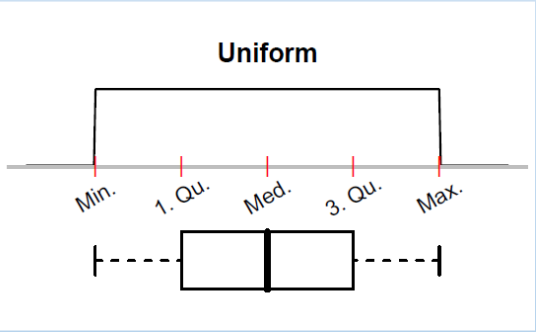
E' **importante** ricordare l'associazione tra il tipo di carattere ed i relativi indici di dispersione che possono essere calcolati:

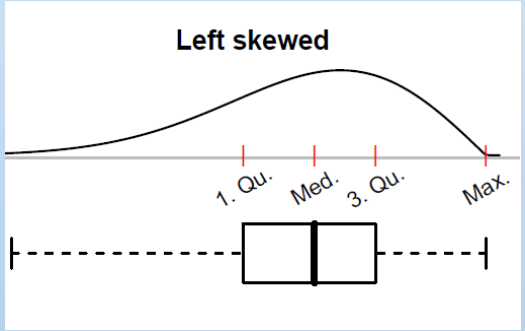
CARATTERE		INDICE DI DISPERSIONE
qualitativo	Sconnesso	nessuno
	Ordinabile	RANGE / IQR *[modalità inferiore/superiore oppure modalità relative a Q1 e Q3 del carattere qualitativo ordinabile, non certo come differenza tra esse !!]*
quantitativo		RANGE / IQR DEVIANZA, VARIANZA, DEVIAZIONE STANDARD e CV%









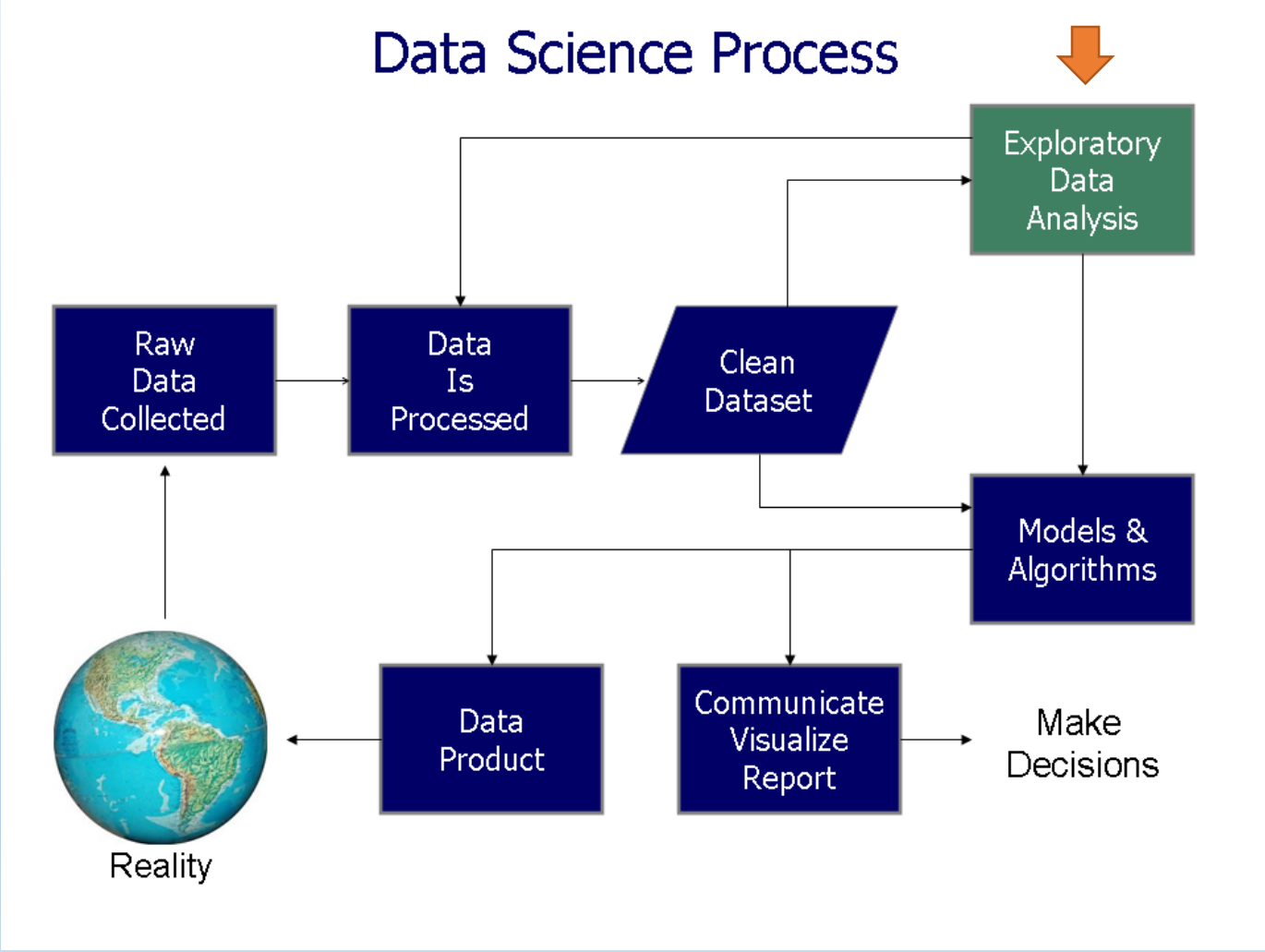


Per una corretta e completa *sintesi* dei dati all'indicazione della tendenza centrale va associata necessariamente l'informazione della loro dispersione.

Indici di Forma

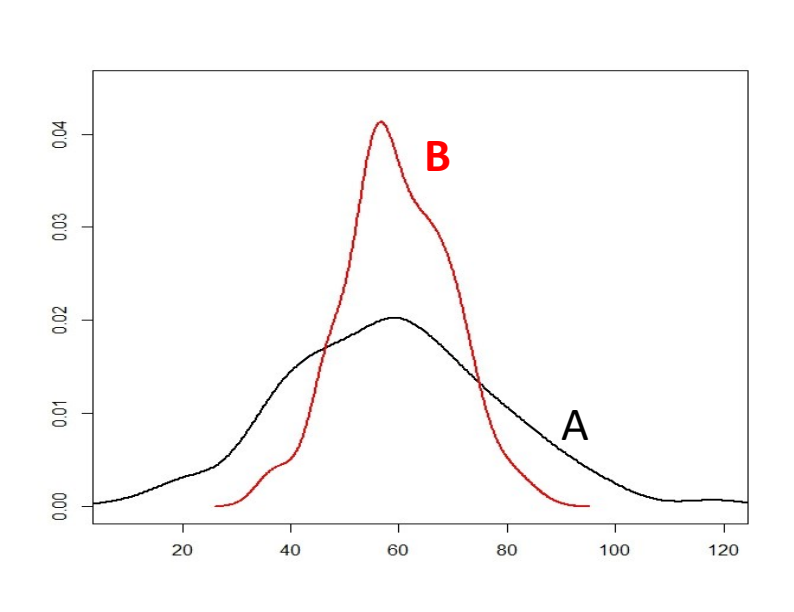
There are lies, damned lies and statistics.

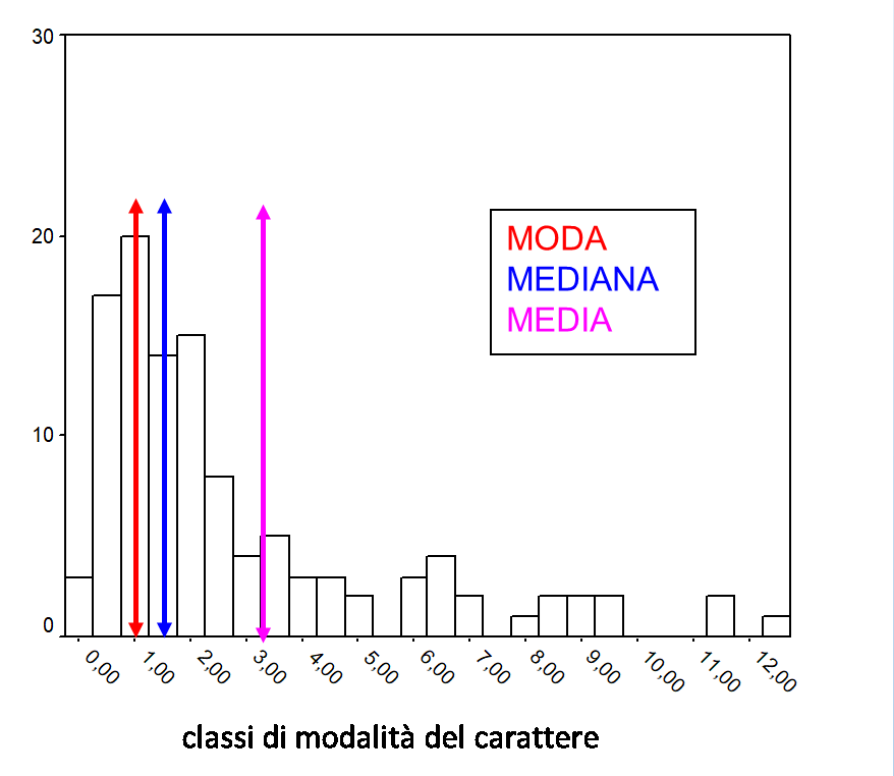
Mark Twain



Media, moda e mediana forniscono indicazioni utili sulla **forma** delle distribuzioni: le loro rispettive posizioni e le differenze tra i loro valori indicano se vi è uno **sbilanciamento** della distribuzione verso destra o verso sinistra.

Però, anche a parità di media e mediana, le distribuzioni possono avere diversa **variabilità**, presentando quindi *forme* differenti.

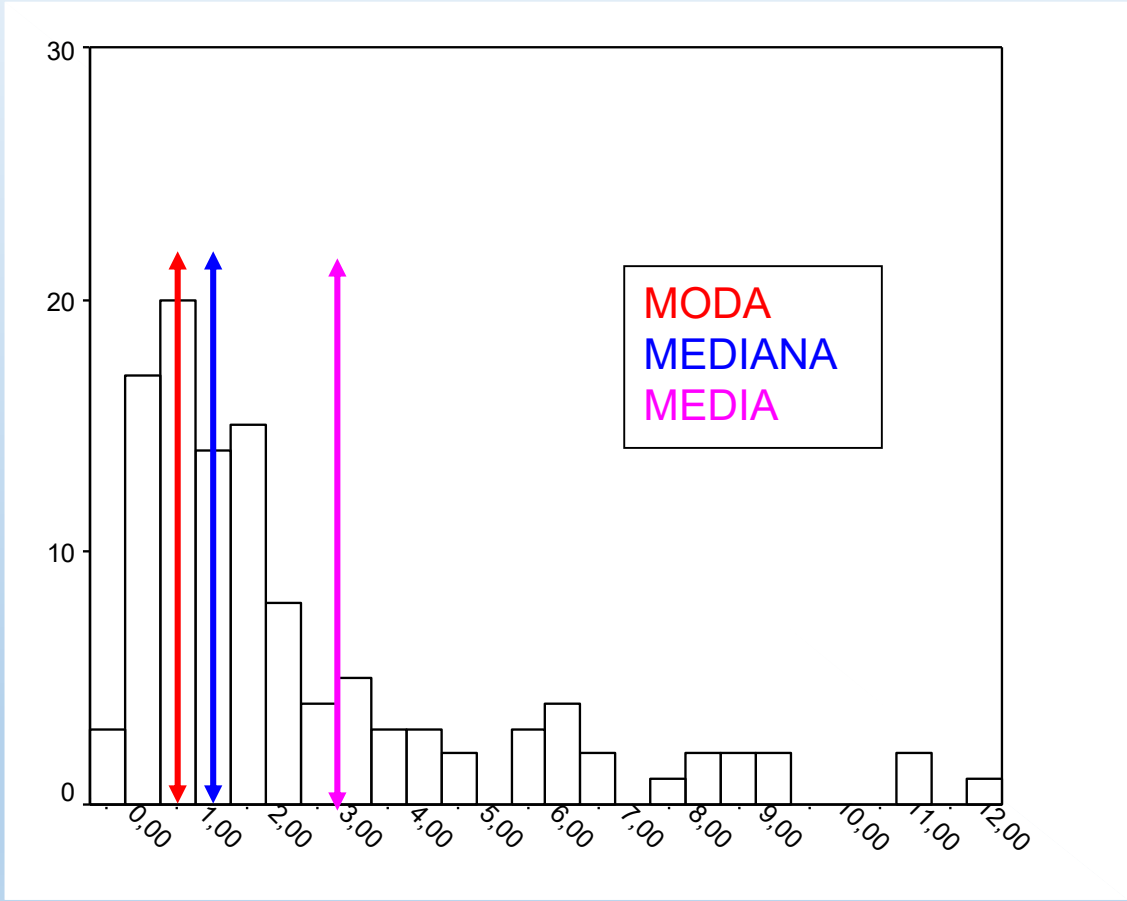




Esistono diversi indici che misurano quanto una distribuzione sia «appiattita» o «appuntita» (si parla in questo caso di curtosi) e l'entità degli eventuali sbilanciamenti verso destra o sinistra (asimmetria).

Gli indici di «forma» delle distribuzioni (I): asimmetria (skewness)

Quanto piu' moda, mediana e media si differenziano,
tanto piu' la distribuzione è asimmetrica.



classi di modalità del carattere

Indice di **skewness** (=asimmetria):

$$\frac{1}{n}\sum\left(\frac{x_i-\bar{x}}{s}\right)^3=\frac{1}{n}\sum z_i^3 \longrightarrow z_i=\frac{x_i-\bar{x}}{s}$$

'z-score' di x_i

La distribuzione è
simmetrica se Skewness=0 (circa)

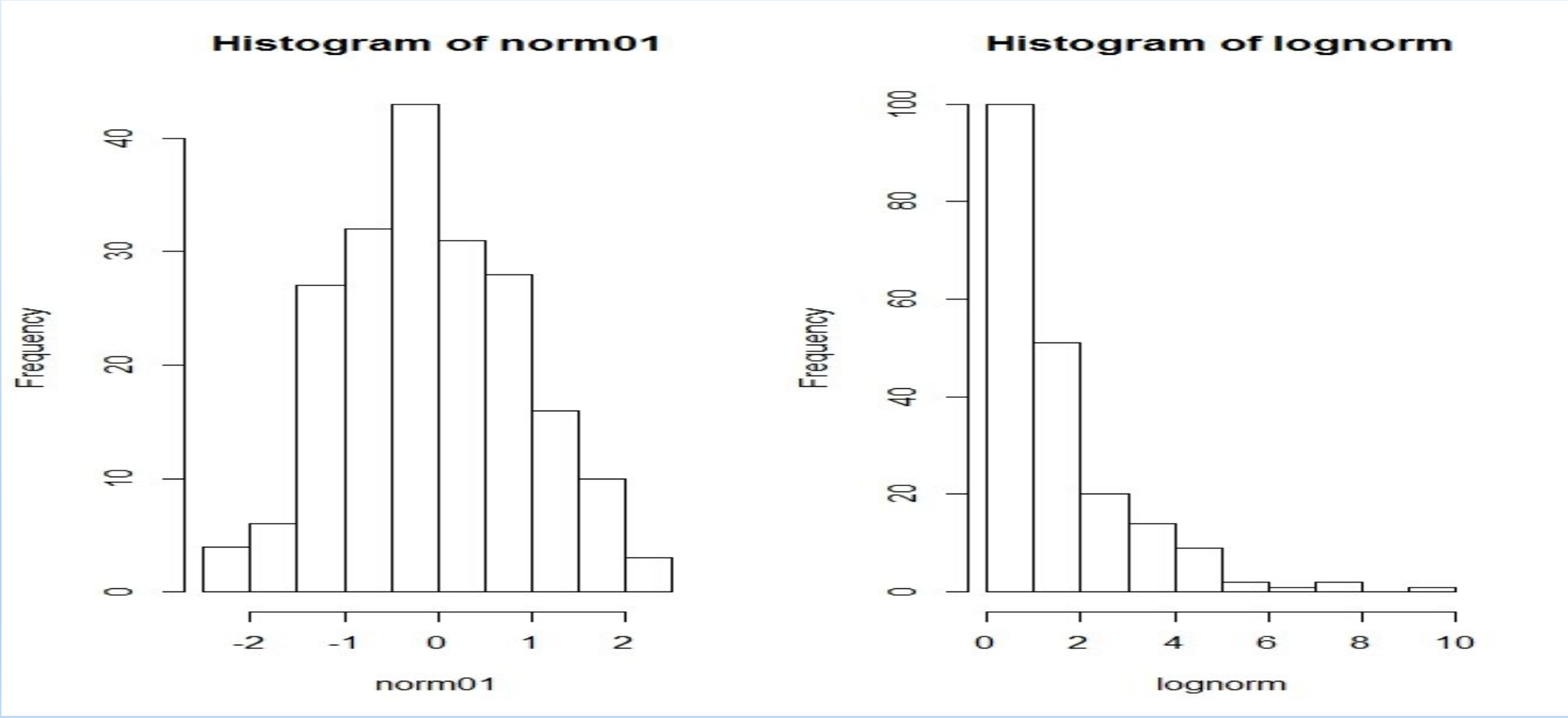
è asimmetrica a destra (positiva) se Skewness > 0

è asimmetrica a sinistra (negativa) se Skewness < 0

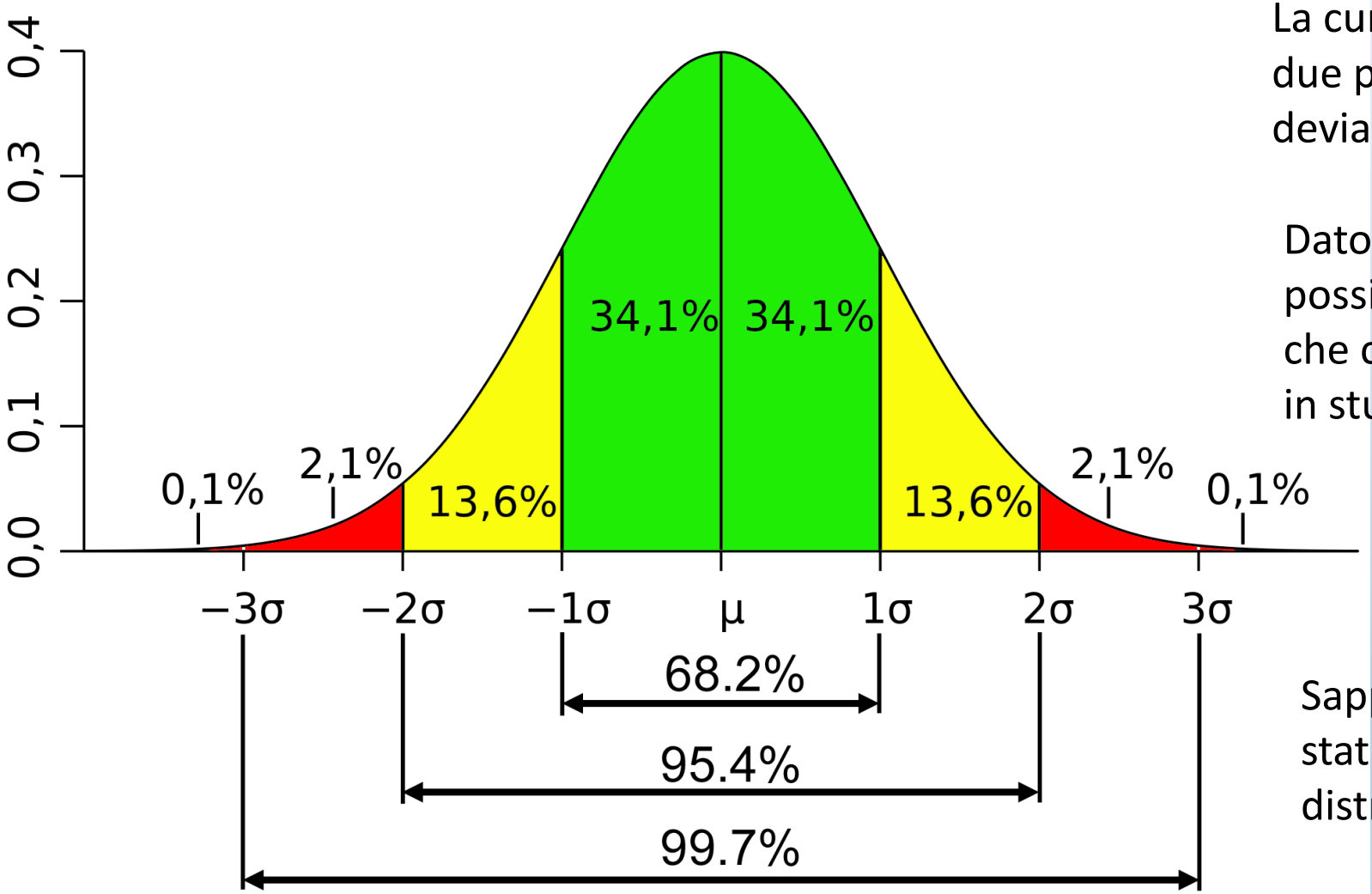
Gli indici di «forma» delle distribuzioni (I): asimmetria (skewness)

```
skewness(norm01)      # 0.17
```

```
skewness(lognorm)     # 2.11
```



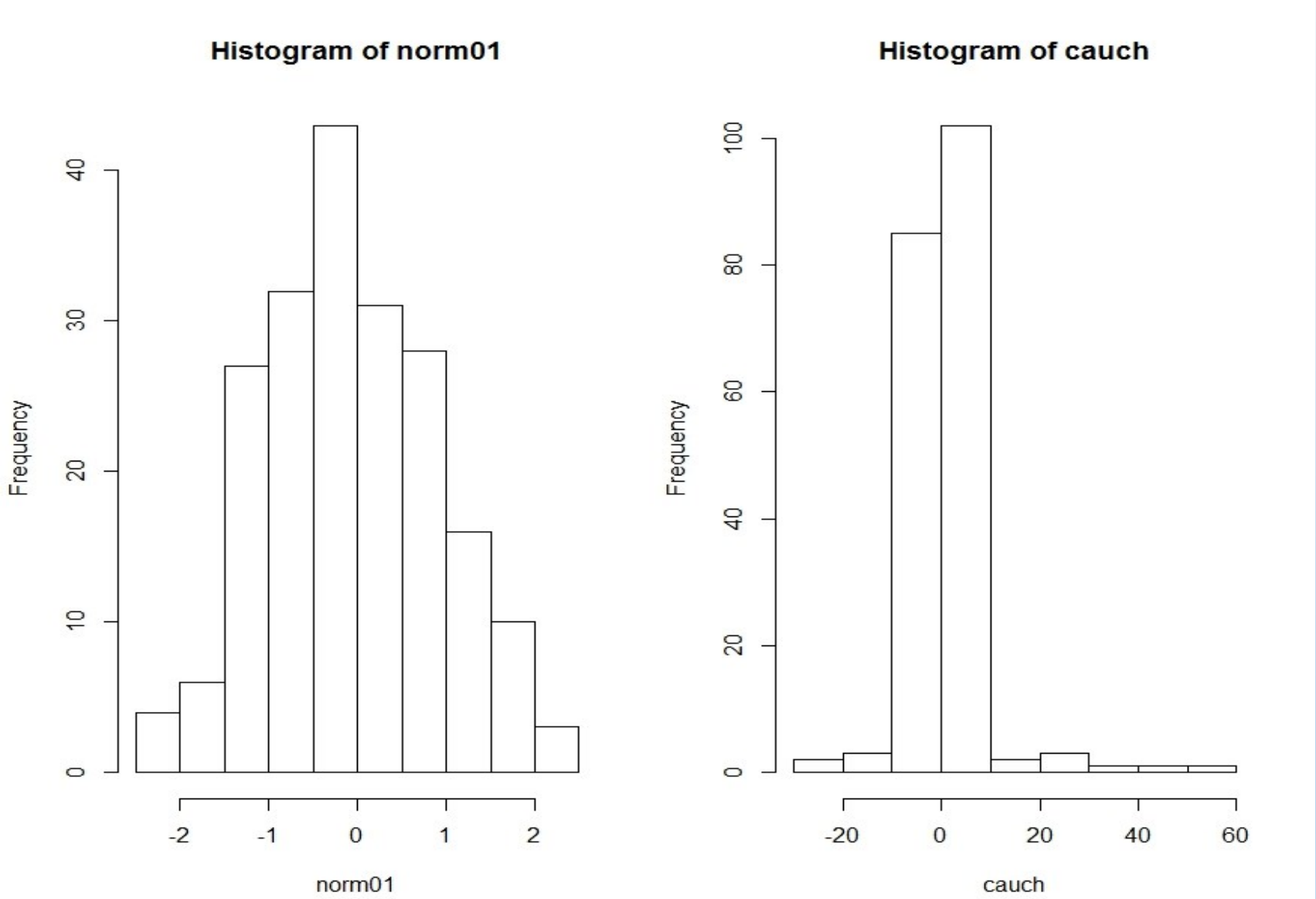
La «forma» della distribuzione normale (curva gaussiana)



La curva normale è caratterizzata da due parametri: la media μ e la deviazione standard σ.

Dato il valore di questi due parametri, possiamo calcolare la % di unità statistiche che cadono in certe regioni della variabile in studio.

Sappiamo quindi anche la % di unità statistiche che cadono nelle «code» della distribuzione.



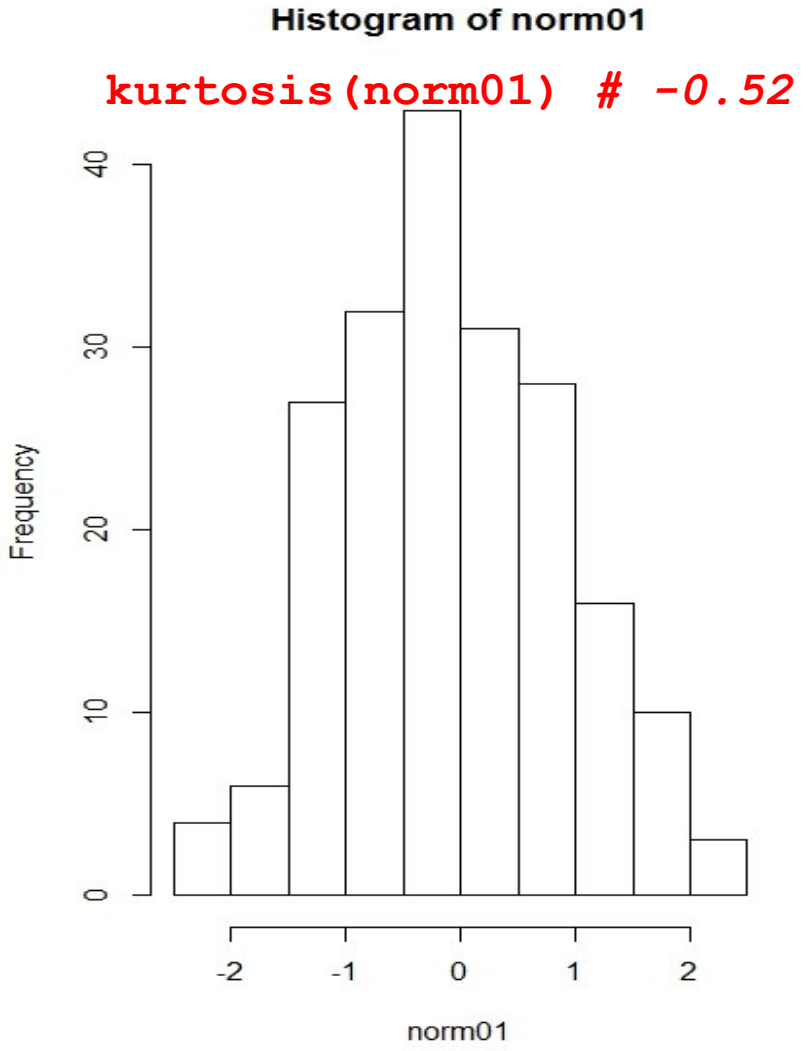
La **curtosi** è un indice che valuta la «pesantezza» delle code relativa al resto della distribuzione (cioè la % di unità statistiche presenti nelle code della distribuzione rispetto al totale).

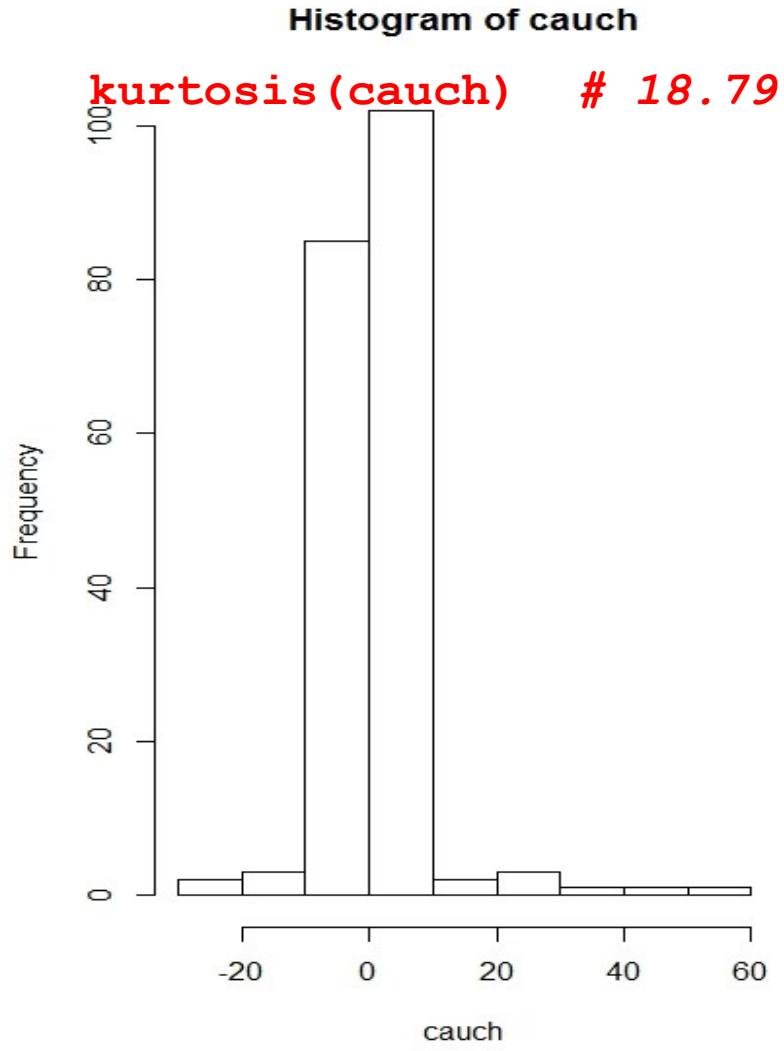
L' indice viene calcolato rispetto alla % teorica che dovrebbe essere presente se la distribuzione fosse normale (con uguale media e varianza).

Gli indici di «forma» delle distribuzioni (II): curtosi

$$\frac{\sum \frac{(x_i - \bar{x})^4}{\sigma^4}}{n} - 3 = \frac{1}{n}\sum z_i^4 - 3$$

Indice di **curtosi**: «3» è il valore di curtosi della curva gaussiana

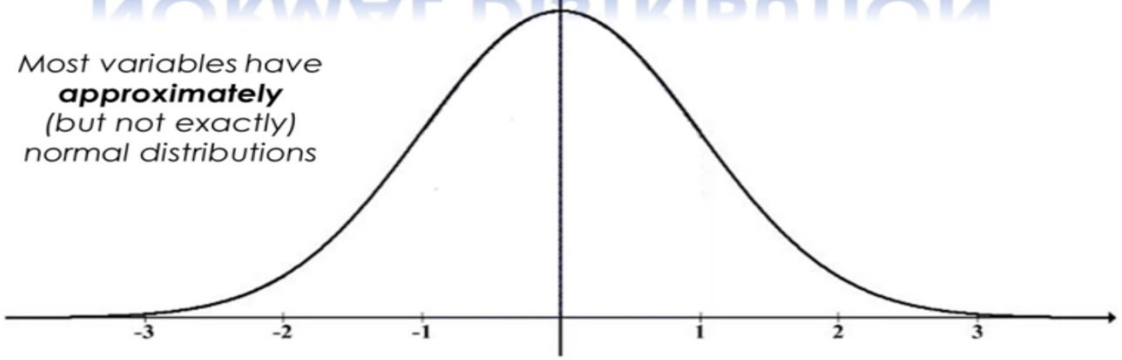




*Il calcolo del coefficiente di curtosi ha senso solo nelle distribuzioni monomodali

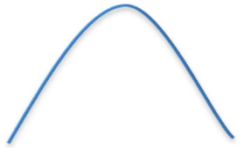
NORMAL DISTRIBUTION

Most variables have **approximately** (but not exactly) normal distributions

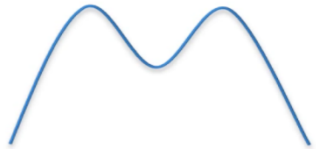


Modality

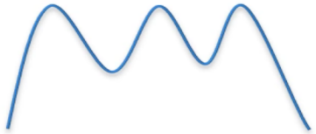
Unimodal
one-peak



Bimodal
two-peaks

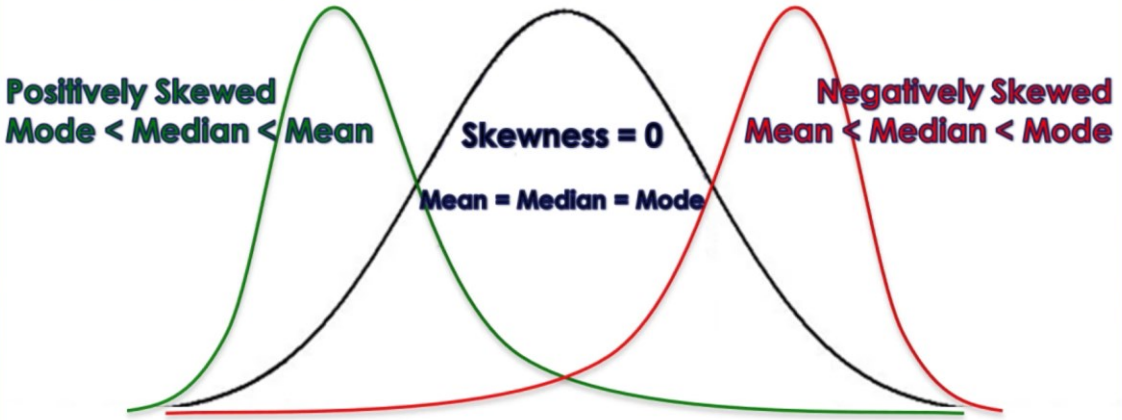


Multimodal
two or more peaks

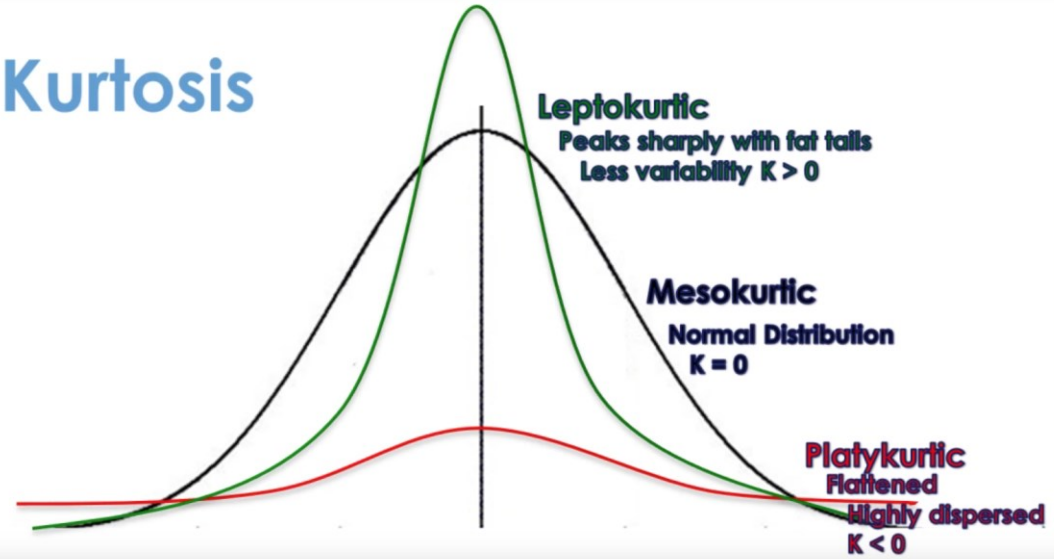


Skewness

$$Skewness = \frac{(Mean - Median)}{Standard\ Deviation}$$



Kurtosis



OPZIONALE

Trasformazioni delle scale [cenni!]

La **trasformazione di scala** di una variabile (su scala quantitativa) può cambiare la sua distribuzione da asimmetrica a una distribuzione più simmetrica, più *simile* alla distribuzione normale.

Questo passaggio è utile per vari scopi: molti metodi statistici (inferenziali) assumono la normalità delle variabili in esame.

Inoltre, se stiamo studiando l'associazione tra una coppia di variabili, opportune trasformazioni della loro scala possono migliorare la *linearità* della associazione.

Inoltre, se stiamo confrontando più distribuzioni, una volta soddisfatta la «*normalità*», possiamo procedere al calcolo dei loro **valori standardizzati**, che possono risultare utili in molte procedure.

Bisogna però poi fare attenzione a come vengono riportati i risultati delle analisi con le variabili trasformate.

Per presentare i valori degli indici di posizione o dispersione, è a volte consigliabile *ri-trasformare* sulla scala originale per evitare ambiguità nell'interpretazione.

Standardizzazione delle variabili (su scala quantitativa)

In statistica, per "**standardizzazione**" si intende la trasformazione di una variabile quantitativa per renderla più facilmente confrontabile con le altre.

Una variabile standardizzata è una variabile quantitativa a cui è stata cambiata la scala di misurazione ottenendo dei *numeri puri* (detti anche **punteggi z** o punteggi standard). Questi nuovi valori sono detti anche *adimensionali*, in quanto sono svincolati dall'unità di misura della variabile di partenza.

La caratteristica principale di una variabile standardizzata è poi che ha sempre media=0 e deviazione standard=1.

La standardizzazione permette quindi di **confrontare** variabili che hanno medie e deviazioni standard misurate su diversa unità di misura o ordine di grandezza. Ad esempio, per capire se c'è più variabilità tra il peso (in kg) o l'altezza (in cm) di un gruppo di individui.

OPZIONALE

Come si ottiene una variabile standardizzata:

- Come prima cosa di calcola la media e la deviazione standard della variabile [se la distribuzione è simmetrica…].
- Successivamente, ai singoli valori della variabile viene sottratta la media.
- Il risultato di tale differenza viene diviso per la deviazione standard della variabile stessa.

In questo modo, si ottiene una variabile standardizzata che ha sempre media=0 e deviazione standard=1.

Calcolo di un punteggio standardizzato: due esempi

Ipotizziamo che un campione di individui abbia un peso medio di 70 kg ed una deviazione standard di 10 kg. Standardizzare la variabile peso significa prendere i singoli pesi degli individui e per ognuno di essi sottrarre 70 e poi dividere il risultato per 10.

Ad esempio, il valore standardizzato per Giovanni, che pesa 85 kg, sarà pari a (85-70)/10= **+1.5**. Il valore standardizzato di Maria, che invece di chili ne pesa 50, sarà (50-70)/10= -2.

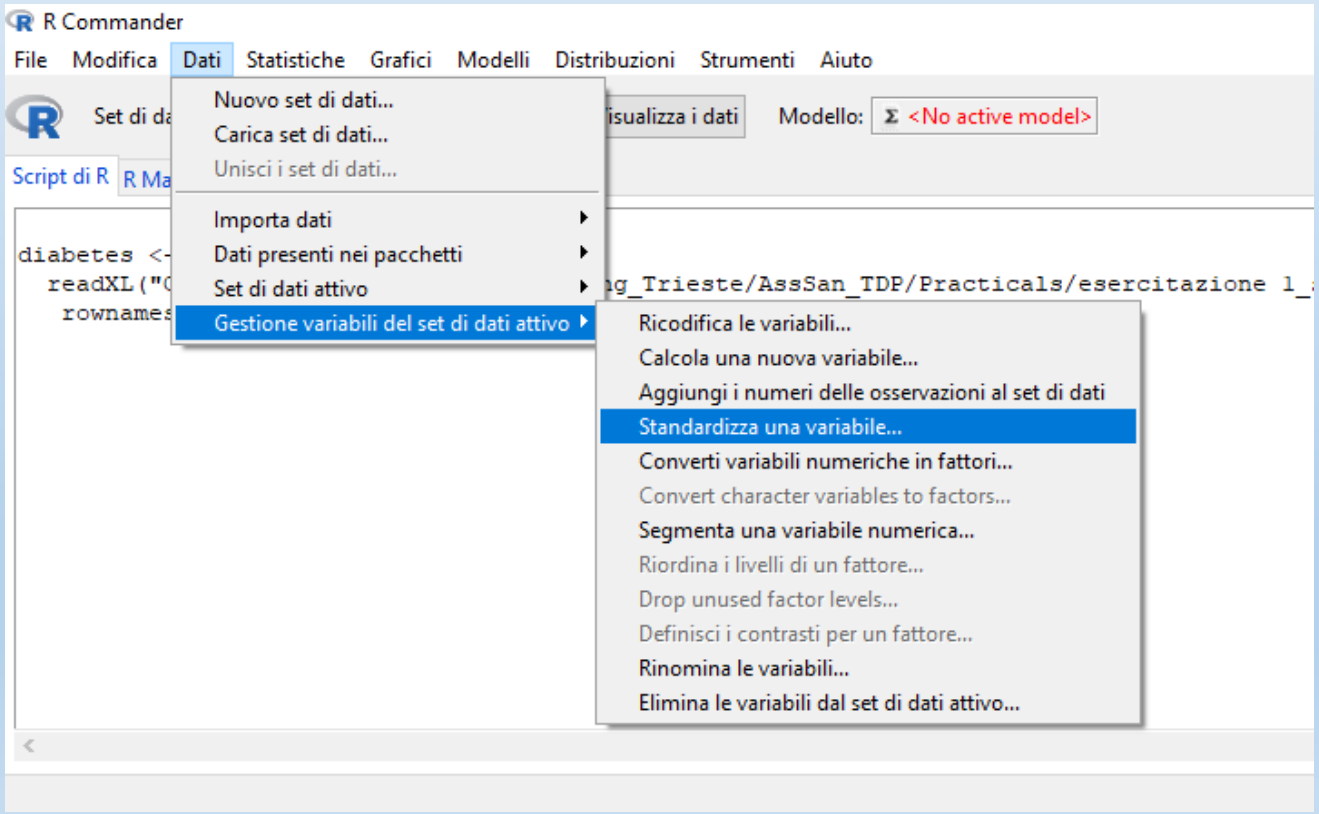
OPZIONALE

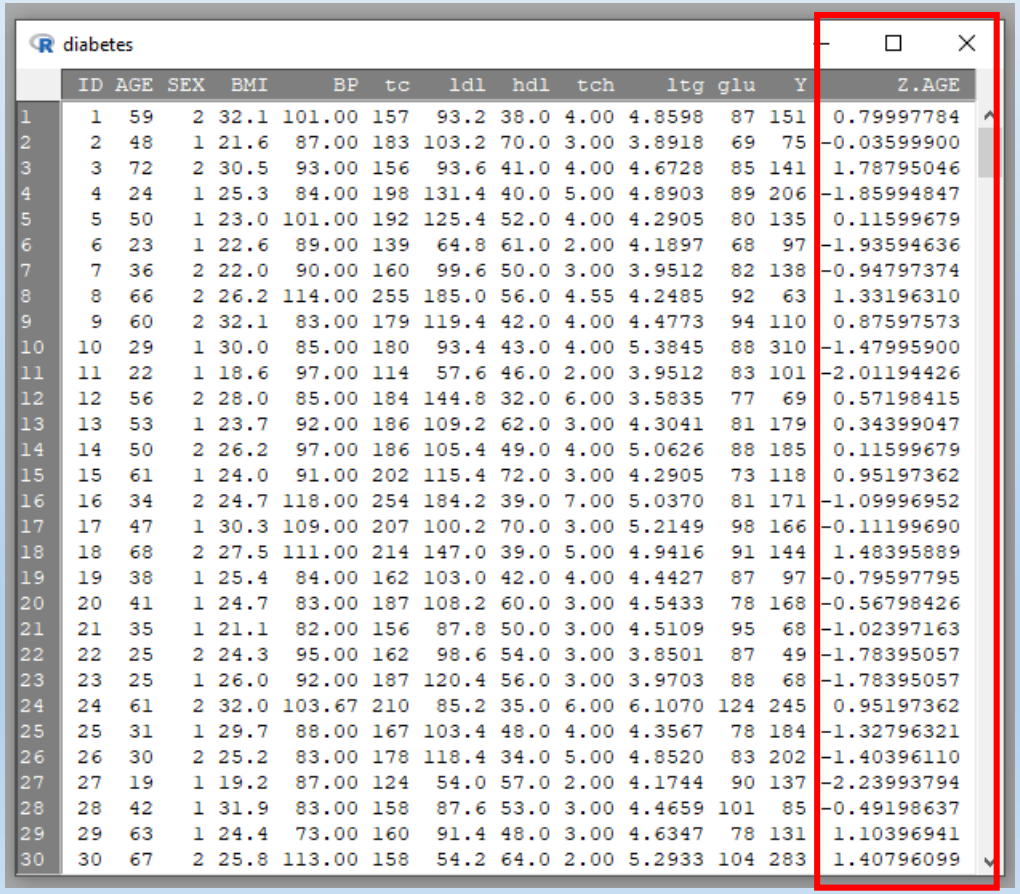
In R Commander:

Una volta caricato il dataset, nel menù principale di R Commander cliccare su:

Dati | Gestione variabili del set di dati attivo | Standardizza una variabile

R aggiungerà in fondo al dataset una nuova variabile denominata Z.*nomevariabiledipartenza*. Ad esempio, se la variabile di partenza si chiama Peso, la nuova variabile si chiamerà Z.Peso.



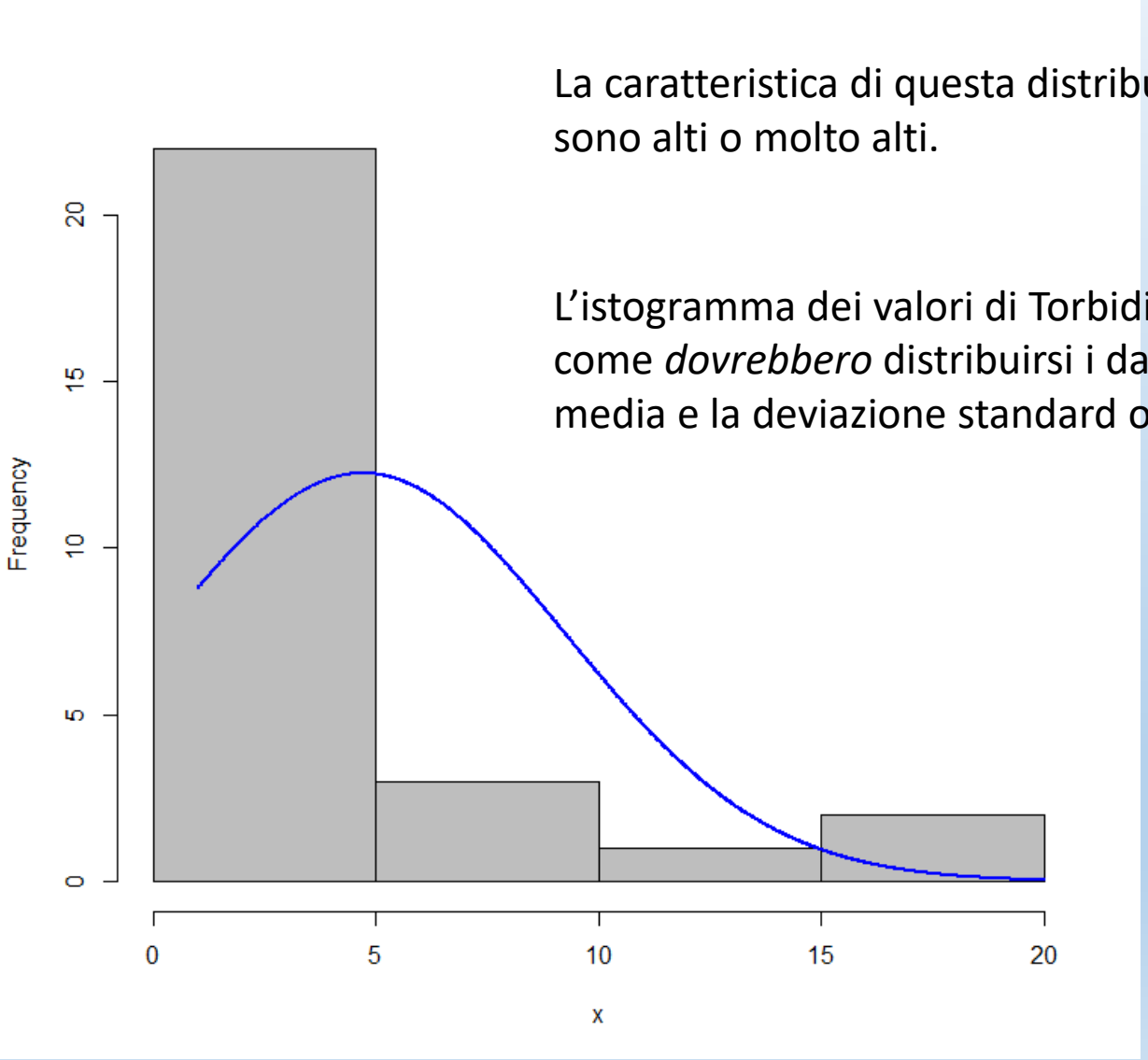

diabetes

	ID	AGE	SEX	BMI	BP	tc	ldl	hdl	tch	ltg	glu	Y	Z.AGE
1	1	59	2	32.1	101.00	157	93.2	38.0	4.00	4.8598	87	151	0.79997784
2	2	48	1	21.6	87.00	183	103.2	70.0	3.00	3.8918	69	75	-0.03599900
3	3	72	2	30.5	93.00	156	93.6	41.0	4.00	4.6728	85	141	1.78795046
4	4	24	1	25.3	84.00	198	131.4	40.0	5.00	4.8903	89	206	-1.85994847
5	5	50	1	23.0	101.00	192	125.4	52.0	4.00	4.2905	80	135	0.11599679
6	6	23	1	22.6	89.00	139	64.8	61.0	2.00	4.1897	68	97	-1.93594636
7	7	36	2	22.0	90.00	160	99.6	50.0	3.00	3.9512	82	138	-0.94797374
8	8	66	2	26.2	114.00	255	185.0	56.0	4.55	4.2485	92	63	1.33196310
9	9	60	2	32.1	83.00	179	119.4	42.0	4.00	4.4773	94	110	0.87597573
10	10	29	1	30.0	85.00	180	93.4	43.0	4.00	5.3845	88	310	-1.47995900
11	11	22	1	18.6	97.00	114	57.6	46.0	2.00	3.9512	83	101	-2.01194426
12	12	56	2	28.0	85.00	184	144.8	32.0	6.00	3.5835	77	69	0.57198415
13	13	53	1	23.7	92.00	186	109.2	62.0	3.00	4.3041	81	179	0.34399047
14	14	50	2	26.2	97.00	186	105.4	49.0	4.00	5.0626	88	185	0.11599679
15	15	61	1	24.0	91.00	202	115.4	72.0	3.00	4.2905	73	118	0.95197362
16	16	34	2	24.7	118.00	254	184.2	39.0	7.00	5.0370	81	171	-1.09996952
17	17	47	1	30.3	109.00	207	100.2	70.0	3.00	5.2149	98	166	-0.11199690
18	18	68	2	27.5	111.00	214	147.0	39.0	5.00	4.9416	91	144	1.48395889
19	19	38	1	25.4	84.00	162	103.0	42.0	4.00	4.4427	87	97	-0.79597795
20	20	41	1	24.7	83.00	187	108.2	60.0	3.00	4.5433	78	168	-0.56798426
21	21	35	1	21.1	82.00	156	87.8	50.0	3.00	4.5109	95	68	-1.02397163
22	22	25	2	24.3	95.00	162	98.6	54.0	3.00	3.8501	87	49	-1.78395057
23	23	25	1	26.0	92.00	187	120.4	56.0	3.00	3.9703	88	68	-1.78395057
24	24	61	2	32.0	103.67	210	85.2	35.0	6.00	6.1070	124	245	0.95197362
25	25	31	1	29.7	88.00	167	103.4	48.0	4.00	4.3567	78	184	-1.32796321
26	26	30	2	25.2	83.00	178	118.4	34.0	5.00	4.8520	83	202	-1.40396110
27	27	19	1	19.2	87.00	124	54.0	57.0	2.00	4.1744	90	137	-2.23993794
28	28	42	1	31.9	83.00	158	87.6	53.0	3.00	4.4659	101	85	-0.49198637
29	29	63	1	24.4	73.00	160	91.4	48.0	3.00	4.6347	78	131	1.10396941
30	30	67	2	25.8	113.00	158	54.2	64.0	2.00	5.2933	104	283	1.40796099

OPZIONALE

TRASFORMAZIONE DELLA SCALA DI MISURA

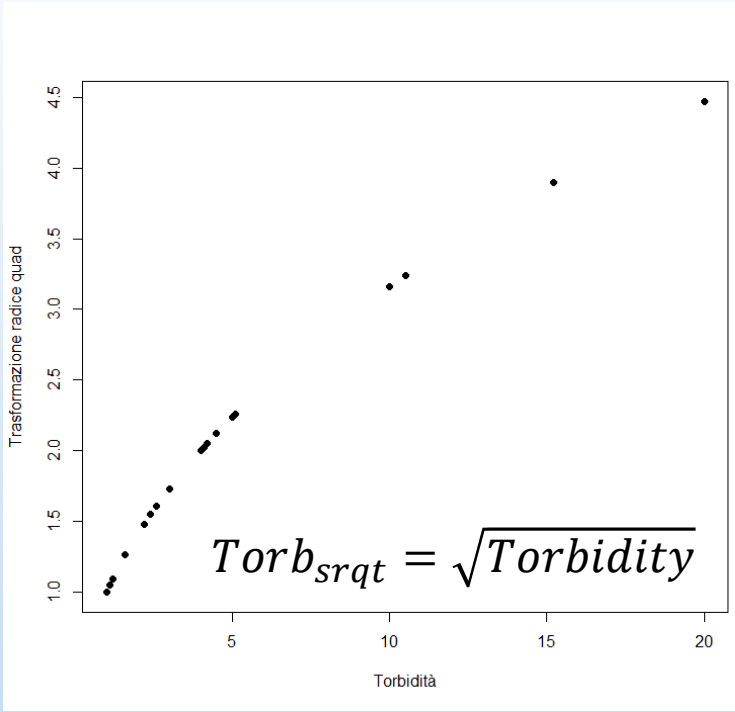
Questo esempio utilizza dei dati ipotetici della torbidità (x) dell'acqua del fiume.
La torbidità è una misura di quanto l'acqua possa diventare torbida a causa del materiale sospeso (inquinanti per lo più).

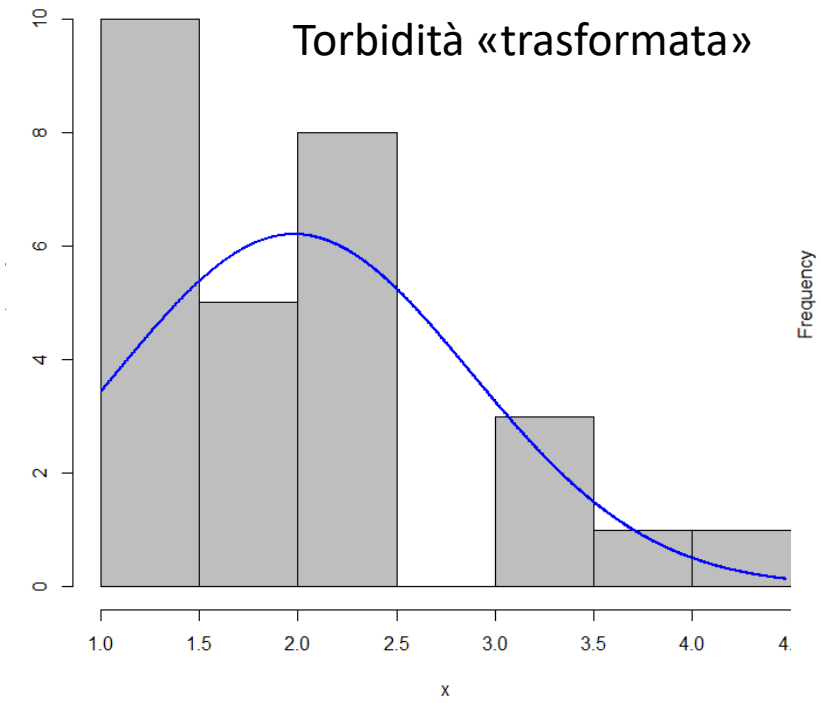


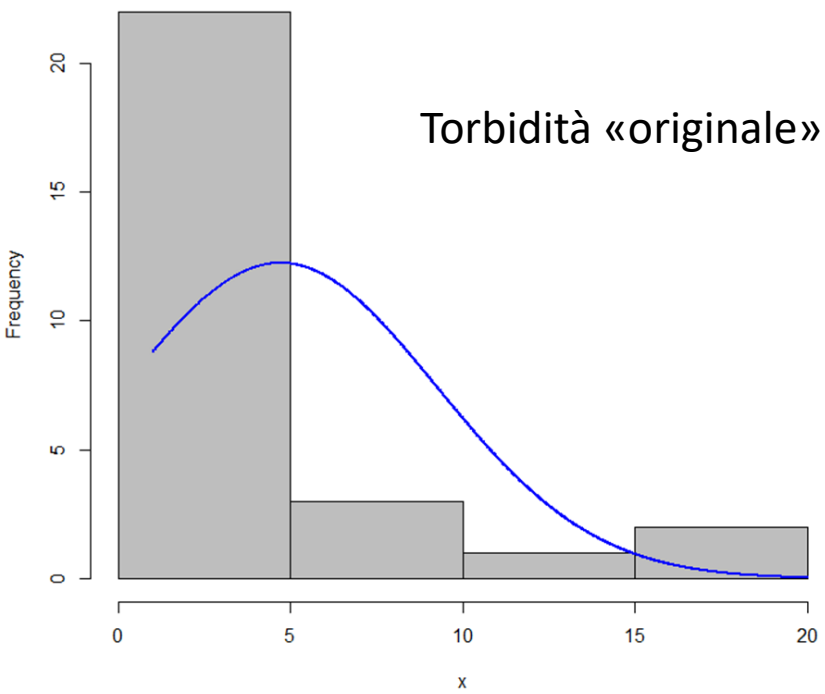
La caratteristica di questa distribuzione è che i valori sono per lo più bassi, ma occasionalmente sono alti o molto alti.

L'istogramma dei valori di Torbidità, con una **curva normale (blu)** sovrapposta [ovvero come *dovrebbero* distribuirsi i dati, se provenissero da una distribuzione normale, data la media e la deviazione standard osservate] ci mostra la forte asimmetria a destra.

Proviamo adesso ad applicare alcune delle trasformazioni più comuni per i dati asimmetrici: la radice quadrata, la radice cubica ed il logaritmo.







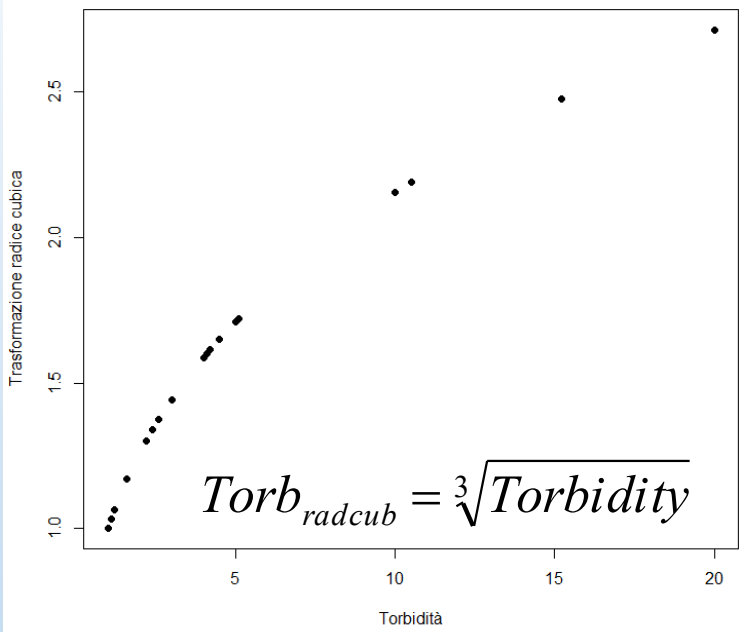
C'è un lieve miglioramento,
ma non soddisfacente

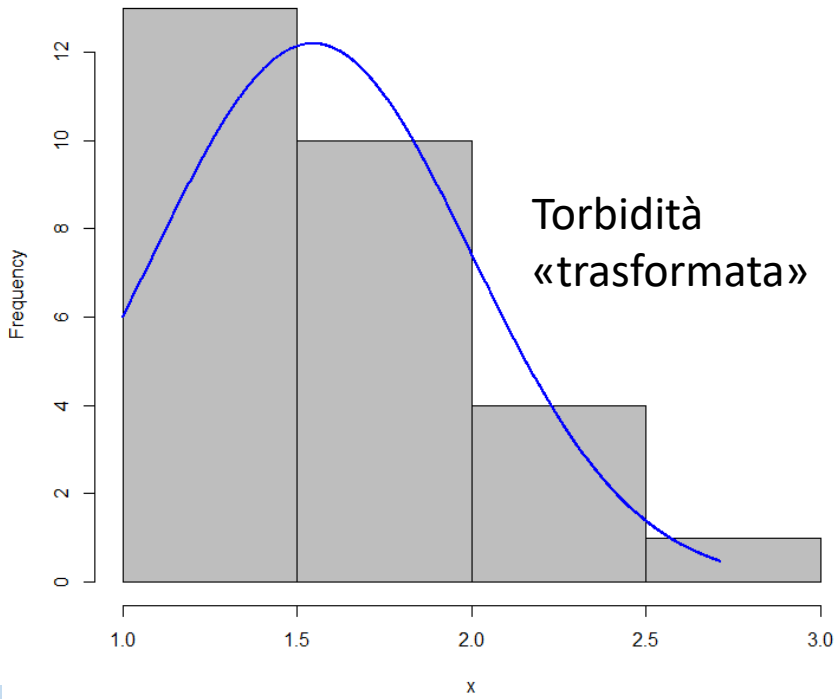
	Media	Dev Std	Mediana	Q1	Q3	Min	Max	Asimmetria	Curtosi
Turbidity	4.7	4.56	4	1.5	5	1	20	1.81	2.87
Sqrt(Turbidity)	1.98	0.9	2	1.22	2.24	1	4.47	1.06	0.49

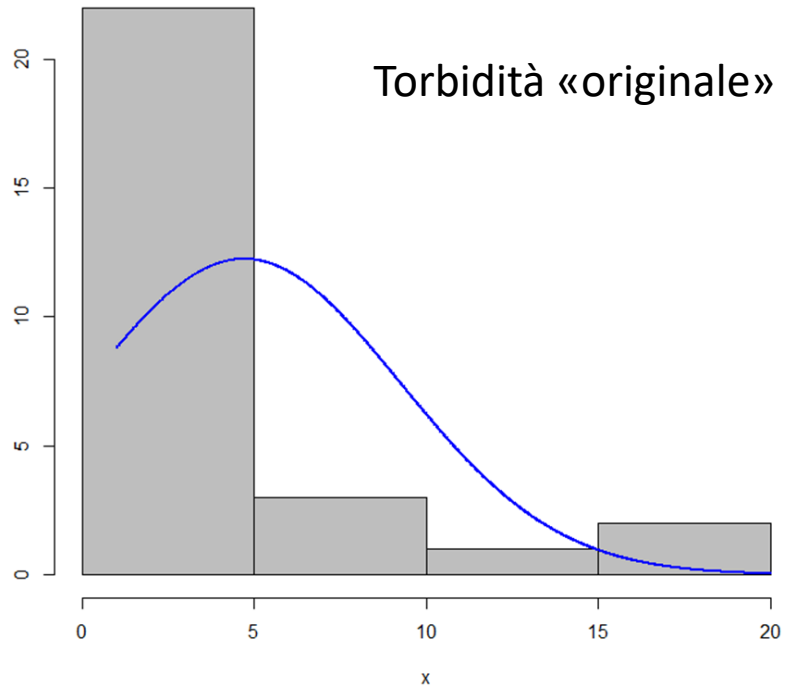
Può essere calcolata su dati >= 0.

OPZIONALE

OPZIONALE





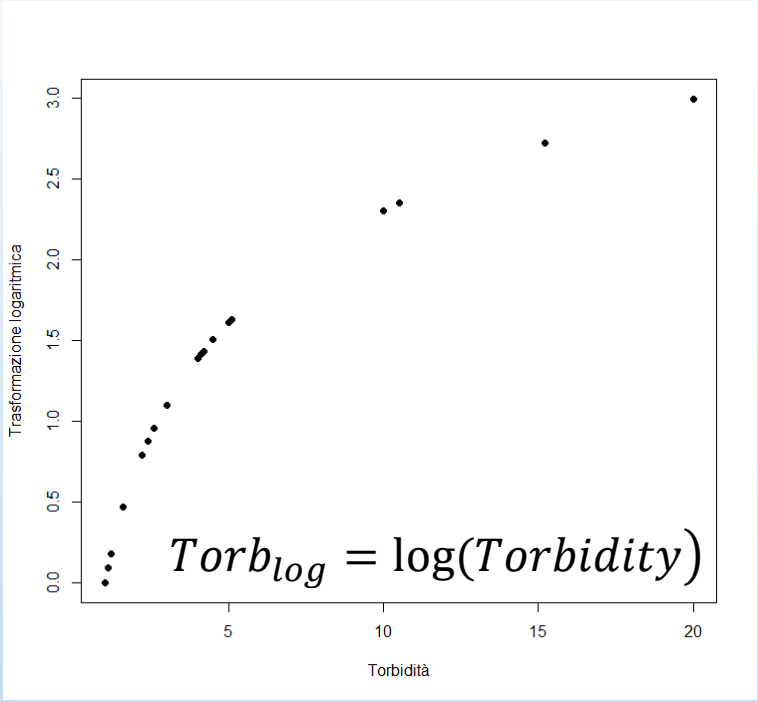


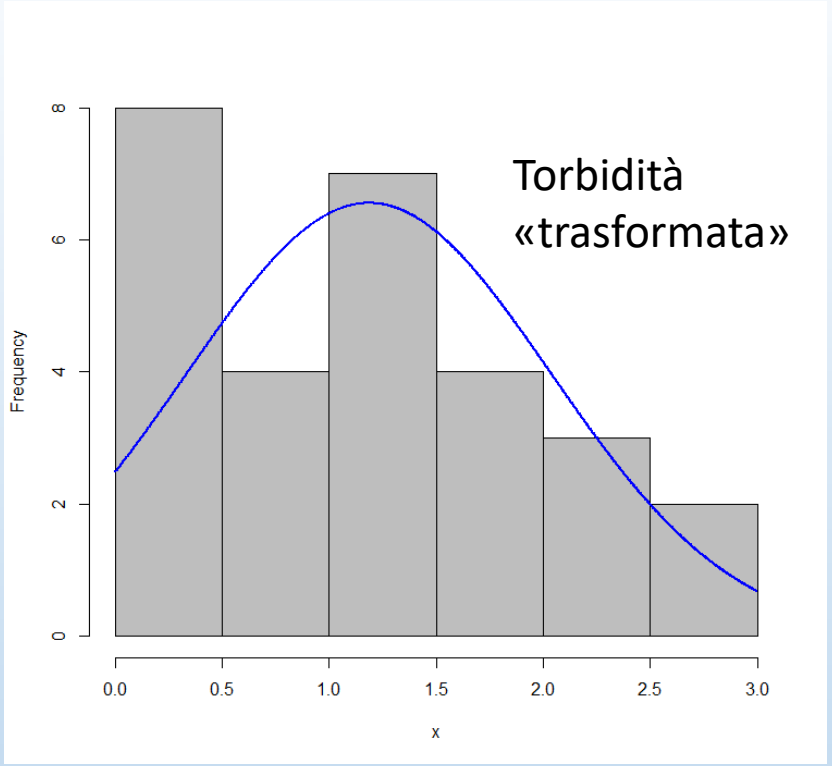
C'è un discreto miglioramento,
ma ancora non del tutto soddisfacente

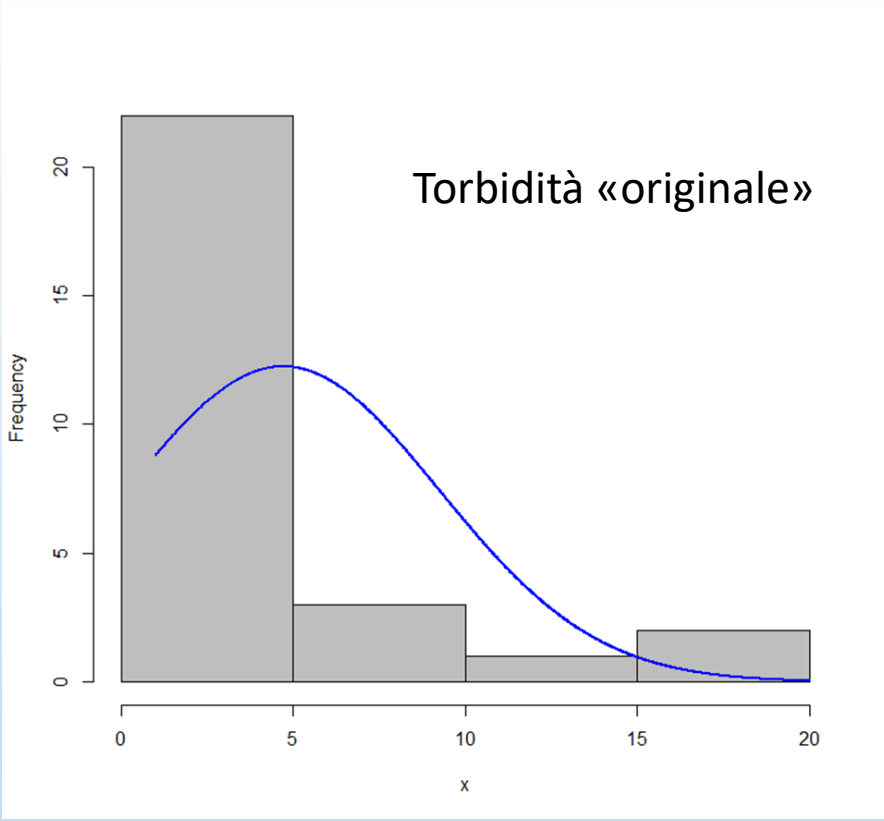
	Media	Dev Std	Mediana	Q1	Q3	Min	Max	Asimmetria	Curtosi
Turbidity	4.7	4.56	4	1.5	5	1	20	**1.81**	**2.87**
Radcub(Turbidity)	1.54	0.46	1.59	1.14	1.71	1	2.71	**0.79**	**-0.10**

Può essere calcolata anche su dati negativi e che contengono il valore zero.

OPZIONALE







Decisamente meglio: visto il tipo di dati è probabilmente il risultato migliore che possiamo ottenere.

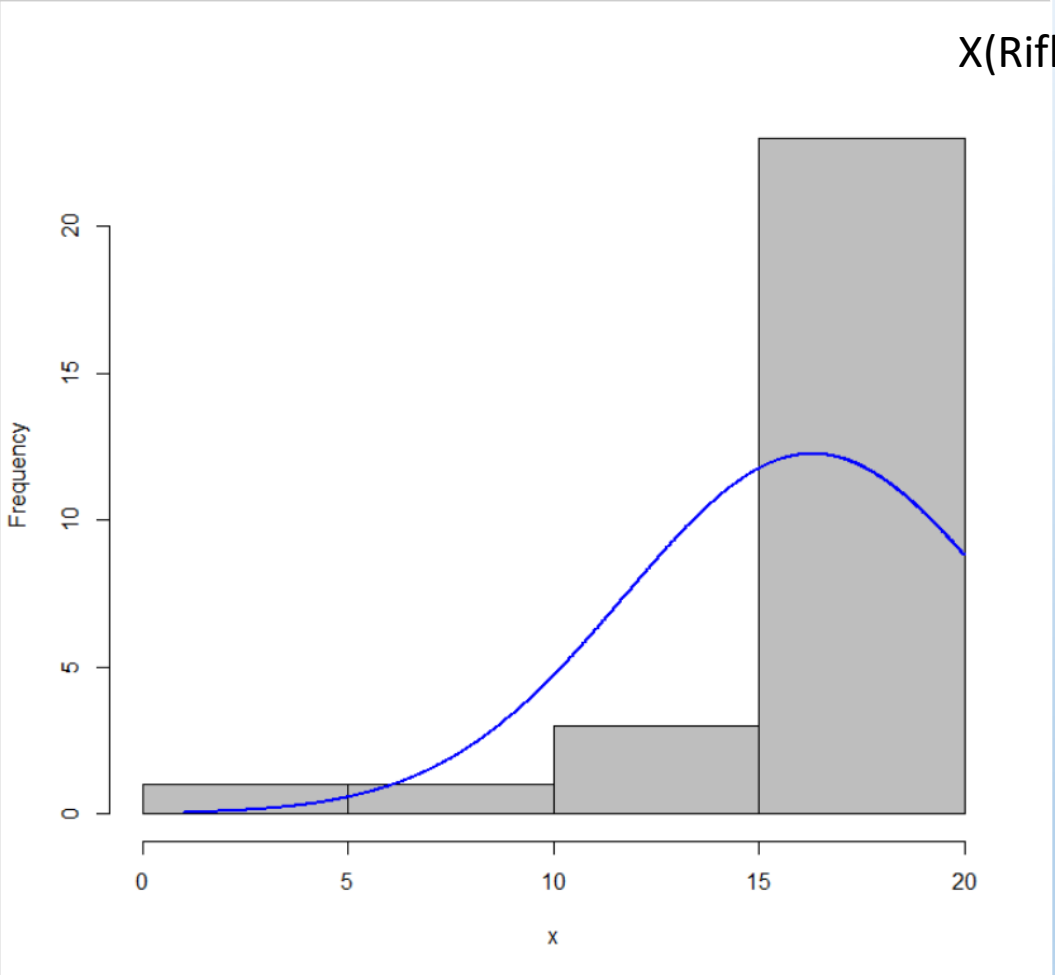
	Media	Dev Std	Mediana	Q1	Q3	Min	Max	Asimmetria	Curtosi
Turbidity	4.7	4.56	4	1.5	5	1	20	**1.81**	**2.87**
log(Turbidity)	1.19	0.85	1.39	0.4	1.61	0	3	**0.28**	**-0.87**

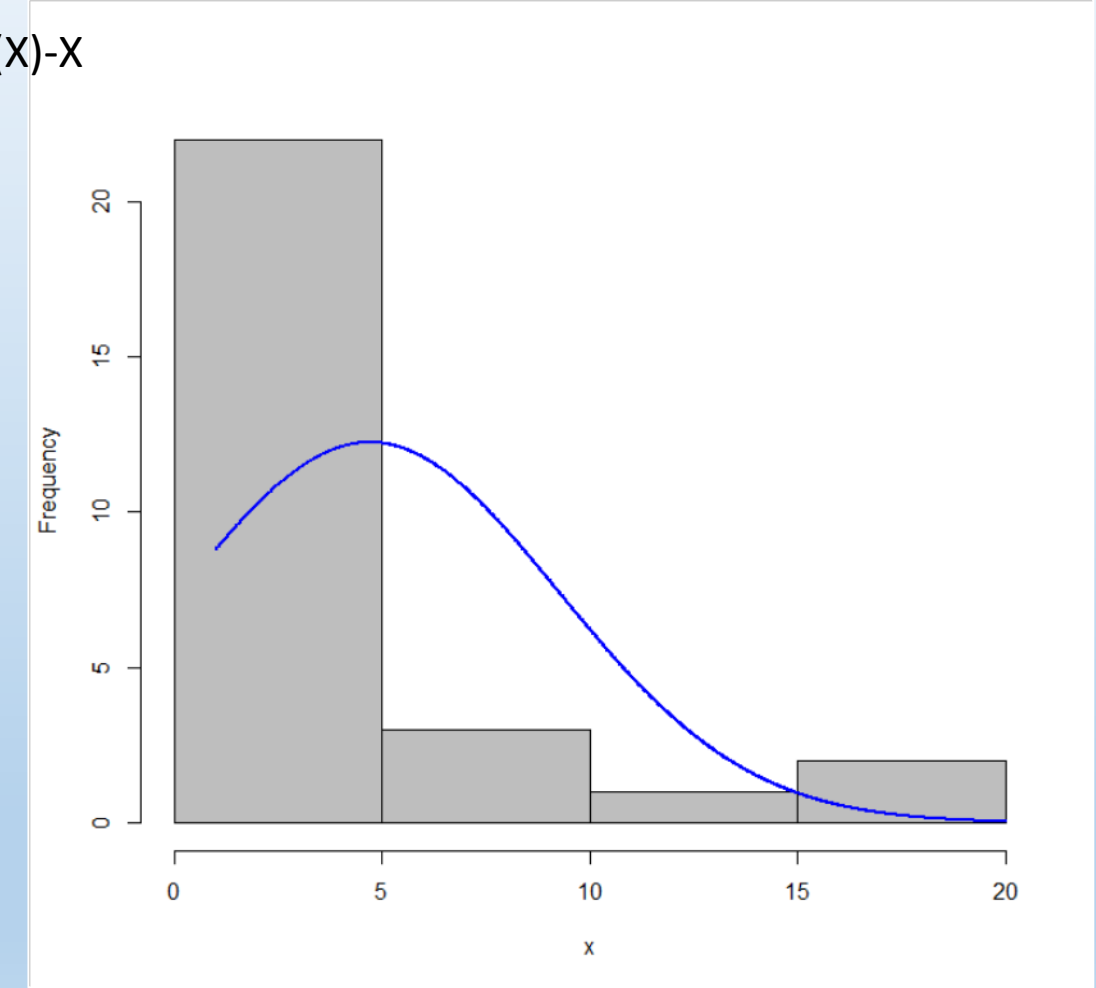
Logaritmo naturale: base è il numero di Nepero (circa 2.7); solo per numeri >0, se la distribuzione contiene 0 va aggiunta una costante

OPZIONALE

Teniamo presente che per le distribuzioni asimmetriche a sinistra, è opportuno prima di tutto creare la variabile «riflessa» e poi si possono applicare le trasformazioni viste (radice quadrata, cubica o logaritmo).

X(Riflessa)=1+Max(X)-X





Ci poi sono alcune funzioni che possiamo applicare sui dati per ottenere «*la migliore trasformazione possibile*», ma non entriamo in questo dettaglio tecnico.