ASSISTENZA SANITARIA & TECNICHE DELLA PREVENZIONE NELL'AMBIENTE E NEI LUOGHI DI LAVORO

STATISTICA MEDICA

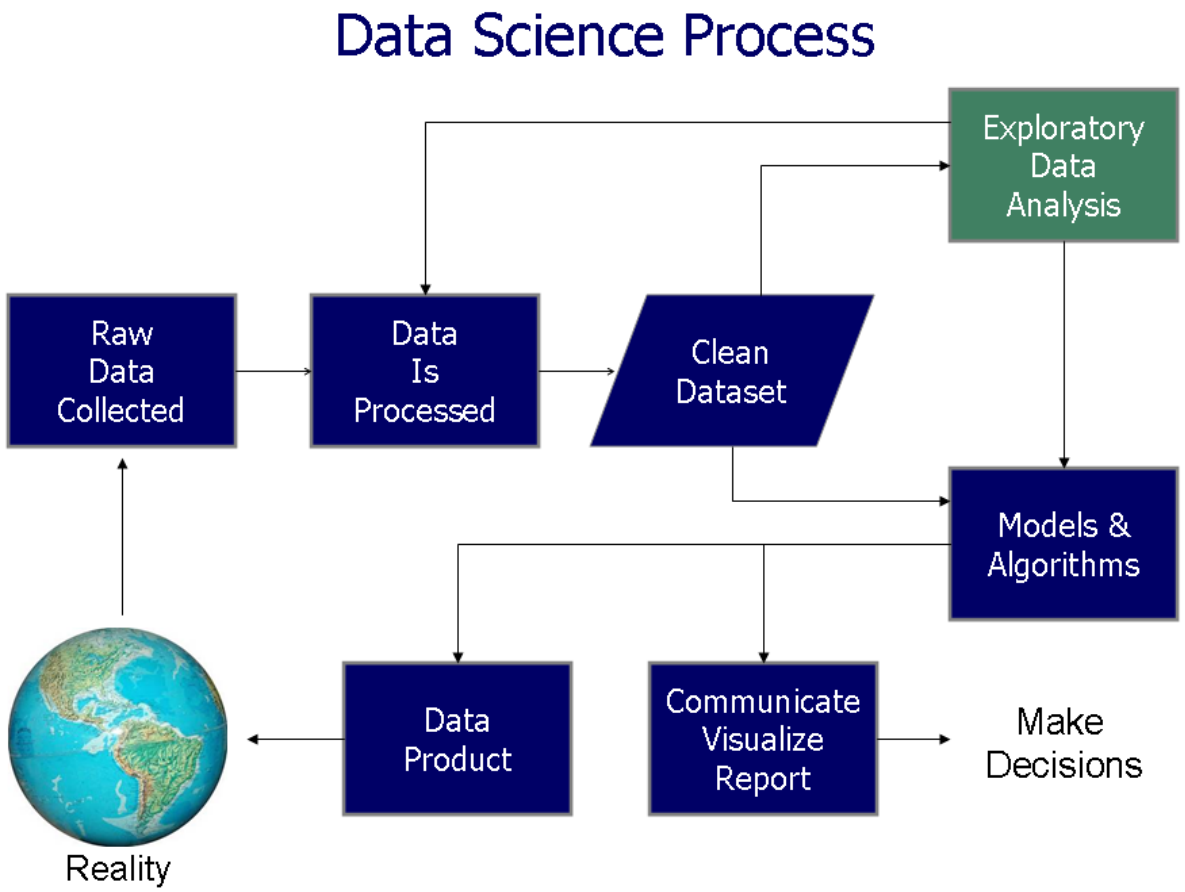
gbarbati@units.it

A.A. 2024-25

Sommario:

WE ARE HERE

- Classificazione dei caratteri
- Distribuzioni & Grafici…
- Indici di Posizione
- Indici di Dispersione



Statistics is the grammar of science.

Karl Pearson (1857-1936)

Ancora un po’ di terminologia…

La statistica è la tecnica che ha come scopo la conoscenza (interpretazione…) **quantitativa** dei **fenomeni collettivi**.

I fenomeni che la nostra mente non può conoscere con una sola osservazione, ma che apprende tramite la sintesi delle osservazioni di fenomeni più semplici, si possono definire

fenomeni di massa* o *fenomeni collettivi

Essi sono composti da una collettività di osservazioni di fenomeni più semplici

fenomeni individuali* o *singoli

«Statistichese»…

L'oggetto dell'osservazione di ogni *fenomeno individuale* che costituisce il *fenomeno collettivo* in studio è detto **'unità statistica'**

Ex: un individuo, una A.S.L., un ospedale...

Ciascuna *unità statistica* presenta delle caratteristiche definite **‘caratteri/variabili’**

Ex: il colore degli occhi in un individuo, il numero di medici che lavorano in una A.S.L., il numero di posti letto in un ospedale....

Ciascun *carattere* è presente in ogni *unità* con una determinata **‘modalità‘**

Ex: colore blu degli occhi, 10 medici in una certa ASL, 300 posti letto in un dato ospedale,...

La classificazione dei caratteri

I caratteri possono essere classificati in:

-Caratteri **qualitativi** (colore degli occhi, tipo di diagnosi...), a loro volta distinti in:

- ***ordinabili*:** è possibile *ordinare* le modalità del carattere in senso crescente o decrescente (es: titolo di studio, livello di gravità della diagnosi...);
- ***sconnessi*:** non c'è alcun ordinamento intrinseco tra le modalità (es: colore degli occhi, sesso...);

-Caratteri **quantitativi** (es: età, peso, numero di medici,...) distinguibili in:

- ***discreti*:** le modalità del carattere sono numeri interi (es: numero di medici...)

- ***continui***: le modalità del carattere sono misurate su una scala continua (es: peso, altezza...).

Alla base di tale classificazione dei caratteri vi è la **'scala di misura'** con cui sono espresse le modalità: se attraverso dei **numeri** o delle '**etichette**'.

Nominale (qualitativi sconnessi)

Ordinale (qualitativi ordinabili)

«factor data»

!! SCALA DI MISURA !!

«date/time data»

Discreta (quantitativo)

Continua (quantitativo)

«numeric data»

Esempio di matrice dei dati:

Abbiamo chiesto ad un campione di studenti se avessero visto l'ultimo Festival di Sanremo e se lo avessero gradito. Inoltre, abbiamo chiesto il genere (maschio/femmina) e il numero di ore giornaliere dedicate al sonno e ad attività di studio.

variabile ↓

Sappiamo definire la scala di misura di queste variabili?

Stu.	sesso	Sanremo	⋯	sonno	studio
1	maschio	Non l'ho visto	⋯	8	2
2	femmina	L'ho visto e mi è piaciuto	⋯	6	30
3	maschio	Non l'ho visto	⋯	9	5
4	femmina	Non l'ho visto	⋯	8	25
⋮	⋮	⋮	⋮	⋮	⋮
52	femmina	Non l'ho visto	⋯	8	20

← *unità statistica*

Classificazione dei caratteri e scala di misura

CARATTERE		SCALA
qualitativo	Sconnesso	Nominale
	Ordinabile	Ordinale
quantitativo		Ad intervalli (scala numerica discreta o continua)

Operazioni che è possibile fare sui caratteri in base alla loro scala di misura:

Operazioni sulle modalità del carattere	**Carattere**		
	qualitativi		**Quantitativi**
	sconnessi	*ordinabili*	*(discreti/continui)*
= ; ≠	si	si	si
> ; <	no	si	si
+ ; -	no	no	si

Distribuzioni

Quando si rilevano le **modalità** con cui uno (o più) **caratteri** si presentano in ciascuna **unità** di una popolazione (campione) si ottiene una **'distribuzione'** della popolazione/campione secondo il carattere considerato.

Ex: distribuzione 'unitaria' del peso in una classe:

Studente	**Peso (kg)**
Marco	62
Chiara	59
Luca	56
...	...

distribuzione «**unitaria**» = ad ogni unità del gruppo è associato il suo peso, cioè la corrispondente modalità del carattere considerato.

Possiamo suddividere in **'classi'** la popolazione secondo il carattere considerato, allora le modalità del carattere vengono raggruppate in classi ed otteniamo una distribuzione di **'frequenze' ->** *frequenza della classe* = numero di unità statistiche che appartengono alla classe.

Distribuzioni di frequenze

Tipo di scuola

		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	Altro	11	34,4	34,4	34,4
	Istituto Professionale	4	12,5	12,5	46,9
	Liceo Classico	3	9,4	9,4	56,3
	Liceo Scientifico	14	43,8	43,8	100,0
	Totale	32	100,0	100,0	

Il numero di unità che appartengono ad una classe= **frequenza assoluta** della classe.

E' possibile calcolare anche le distribuzioni di frequenze relative e di frequenze percentuali

Alcuni simboli:

Abbiamo *k* classi di un dato carattere (nell'ex: 6 classi) cioè *k*=6; le possiamo indicare tramite un indice *i* che varia da 1 a 6: *i*=1,2,3,4,5,6. Abbiamo rilevato le modalità del carattere su *N* unità (nell'ex: *N*=30).

- Frequenza **assoluta** della classe *i* -> n_i
- Frequenza **relativa** della classe *i* -> $f_i=n_i/N$
- Frequenza **percentuale** della classe *i* -> $p_i=f_i*100=(n_i/N)*100$

Esiste poi un simbolo matematico indica l'operazione di **somma** su un gruppo di oggetti ed è chiamato '**sommatoria**' :

$$\sum_{i=1}^{k} n_i = n_1 + n_2 + n_3 + n_4 + n_5 + n_6 = 30$$

$$\sum_{i=1}^{k} n_i = n_1 + n_2 + n_3 + n_4 + n_5 + n_6$$

$$\sum_{i=1}^{k} n_i = N \; ; \quad \sum_{i=1}^{k} f_i = 1 \; ; \quad \sum_{i=1}^{k} p_i = 100$$

(per *k*=6)

Distribuzioni di frequenze da un carattere su scala numerica

Modalità	Frequenza assoluta
160	2
164	2
165	1
166	1
168	1
170	5
172	1
173	4
174	1
175	4
176	4
177	1
178	2
180	6
181	2
182	2
183	3
184	2
185	3
186	1
187	2
188	1
190	2

Abbiamo raccolto dei dati sulla distribuzione delle altezze in un campione di 53 studenti. Vogliamo suddividere questa distribuzione in classi di frequenza. Con che criterio scegliamo le classi?

Modalità	Frequenza assoluta
[160,165]	5
(165,170]	7
(170,175]	10
(175,180]	13
(180,185]	12
(185,190]	6

Possiamo creare ad esempio **6** classi di frequenza:
[: la classe comprende il valore;
(: la classe NON comprende il valore;

➡ Classi [circa] di uguale ampiezza

Modalità	Frequenza assoluta
[160,185]	47
(185,190]	6

Possiamo creare anche solo **2** classi di frequenza:
[: la classe comprende il valore;
(: la classe NON comprende il valore;

La scelta delle classi **è arbitraria**, ma va fatta in maniera ragionevole….

Può capitare, per scelta (si vuole fornire informazioni più dettagliate su una parte della distribuzione) o per necessità (i dati ci sono stati forniti raggruppati in classi e non disponiamo della distribuzione unitaria) di costruire delle classi utilizzando *intervalli di lunghezza diversa*.

In questo caso è conveniente definire il concetto di **densità** di frequenza della classe:

$$\left(\begin{array}{c}\text{densità}\\ \text{di una classe}\end{array}\right) = \frac{\text{frequenza assoluta di } Y \text{ sull'intervallo}}{\text{lunghezza dell'intervallo}}$$

La densità ci dice il *numero atteso [medio] di unità statistiche* per ogni unità di misura della variabile.

Modalità	Freq. ass.
0	1
11	1
79	1
100	1
112	1
119	1
130	2
140	1
150	1
162	1
169	1
176	1
200	2
231	1
254	1
257	1
277	1
300	1
349	1
350	1
356	2
370	1
400	1
438	1
439	1
450	1
463	1
469	1
470	1
500	2
520	1
543	2

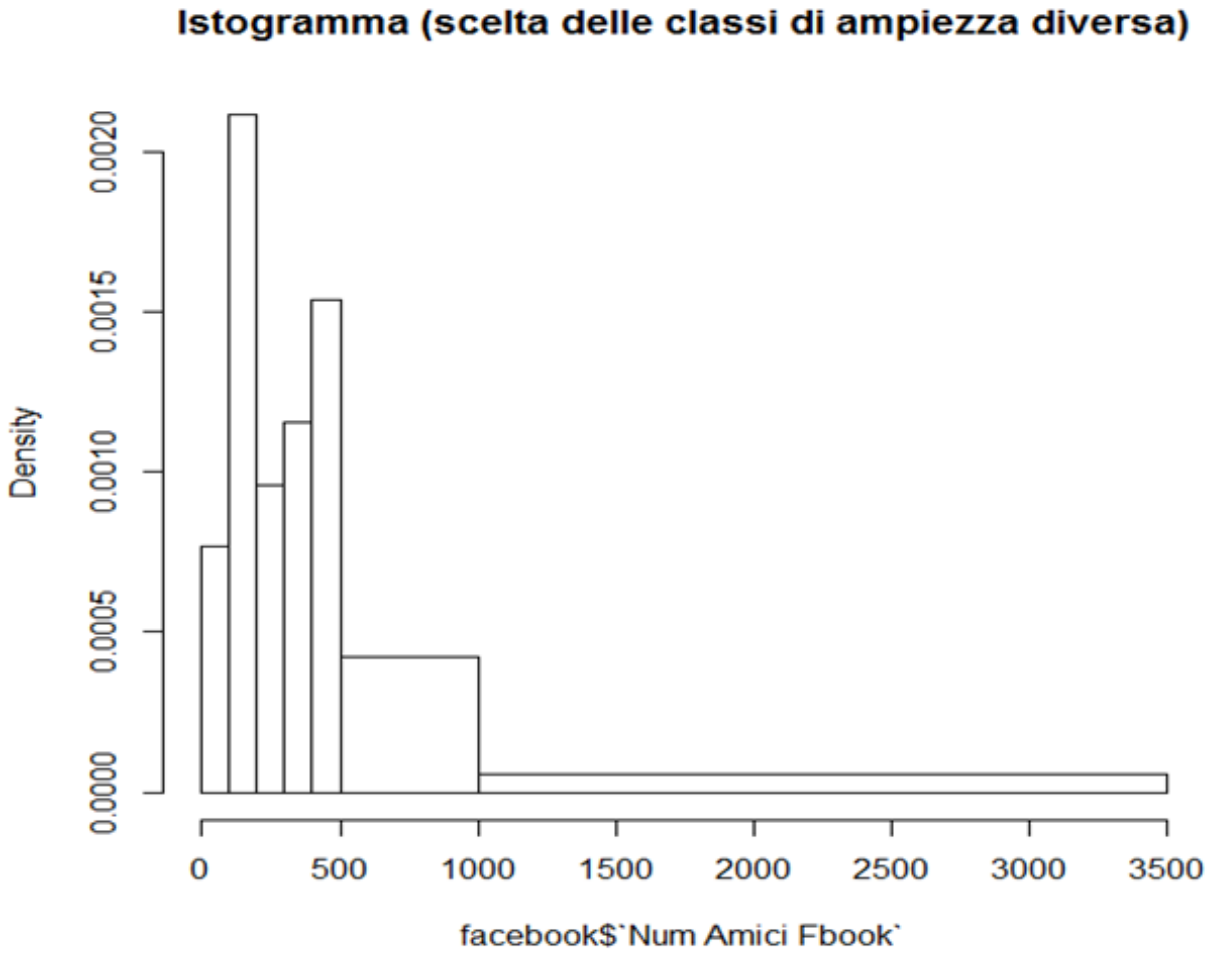
Esempio: Numero di amici su Facebook, indagine fatta su un campione di 52 studenti.

Modalità	Frequenza assoluta	Ampiezza Classe	Densità
[0,100]	4	100	0,04
(100,200]	11	100	0,11
(200,300]	5	100	0,05
(300,400]	6	100	0,06
(400,500]	8	100	0,08
(500,1000]	11	500	0,022
(1000, 3500]	7	2500	0,0028

La densità ci dice il *numero atteso [medio] di unità statistiche* per ogni unità di misura della variabile.

Nella classe [0,100] ci aspettiamo di osservare 4 persone in un intervallo di 100 unità.

Nella classe (500,1000] ci aspettiamo di vedere 2.2 persone ogni 100 unità (cioè 2.2 persone tra 500 e 600, altrettante tra 600 e 700, …etc).



"Altezza" della densità (asse Y istogramma) Freq relativa/ampiezza
(4/52)/100=0.0008
0.0021
0.0010
0.0012
0.0015
0.0004
0.0001

Modalità	Frequenza assoluta	Ampiezza Classe	Densità
[0,100]	4	100	0,04
(100,200]	11	100	0,11
(200,300]	5	100	0,05
(300,400]	6	100	0,06
(400,500]	8	100	0,08
(500,1000]	11	500	0,022
(1000, 3500]	7	2500	0,0028

Distribuzioni doppie (dati categorici/in classi)

Se si rilevano **due** caratteri in una popolazione si ottiene una **distribuzione *doppia***. Per rappresentare le distribuzioni doppie delle modalità di due caratteri (**categorici o espressi in classi**) si utilizza la '**tabella di contingenza**' o '**tabella a doppia entrata**':

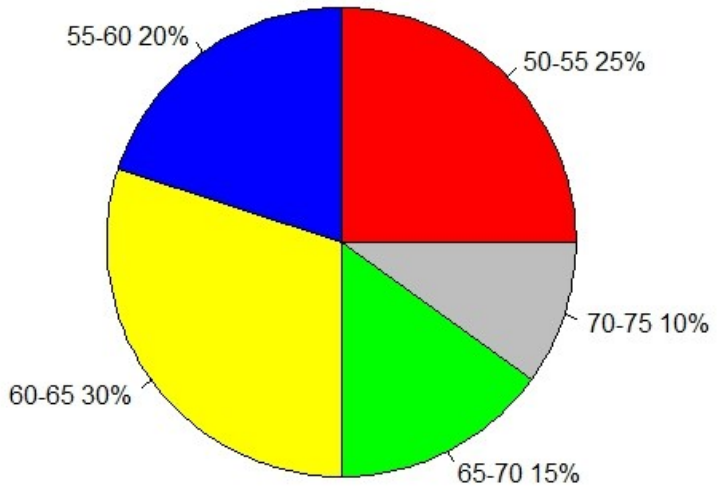
CLASSI DI PESO	M	F	TOT
45-50	0	2	**2**
50-55	5	2	**7**
55-60	4	3	**7**
60-65	6	2	**8**
65-70	3	1	**4**
70-75	2	0	**2**
TOT	**20**	**10**	**30**

Distribuzioni doppie (dati categorici/in classi)

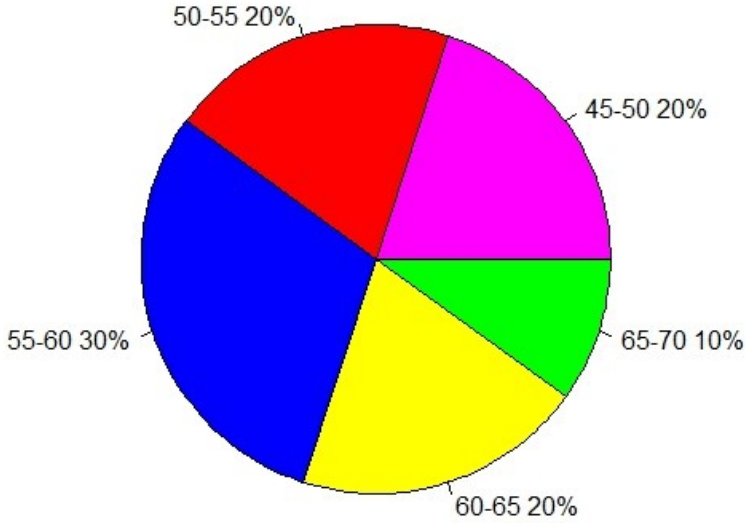
Per confrontare correttamente la distribuzione delle classi di peso rispetto al sesso occorre calcolare le **frequenze relative** all'interno della tabella di contingenza:

CLASSI DI PESO	M	F
45-50	0	20
50-55	25	20
55-60	20	30
60-65	30	20
65-70	15	10
70-75	10	0
TOT	100	100

MASCHI



FEMMINE



La tabella di contingenza permette inoltre di studiare **l'associazione** fra i caratteri analizzati

→ **test di ipotesi** che vedremo in seguito → **Statistica Inferenziale**

Tabella di Contingenza (a doppia entrata):

$n_{a_i b_j}$ = numero di soggetti che presentano *contemporaneamente* la modalità *i* del carattere A e la modalità *j* del carattere B.

		B				Totale
		$\mathbf{B_1}$	$\mathbf{B_2}$	**...**	$\mathbf{B_m}$	
A	$\mathbf{A_1}$	$n a_1 b_1$	$n a_1 b_2$	...	$n a_1 b_m$	$\mathbf{NA_1}$
	$\mathbf{A_2}$	$n a_2 b_1$	$n a_2 b_2$	...	$n a_2 b_m$	$\mathbf{NA_2}$
	...	...	...	$n a_i b_j$	...	...
	...	...	...	...	...	...
	$\mathbf{A_k}$	$n a_k b_1$	$n a_k b_2$	...	$n a_k b_m$	$\mathbf{NA_k}$
Totale		$\mathbf{NB_1}$	$\mathbf{NB_2}$	**...**	$\mathbf{NB_m}$	**N**

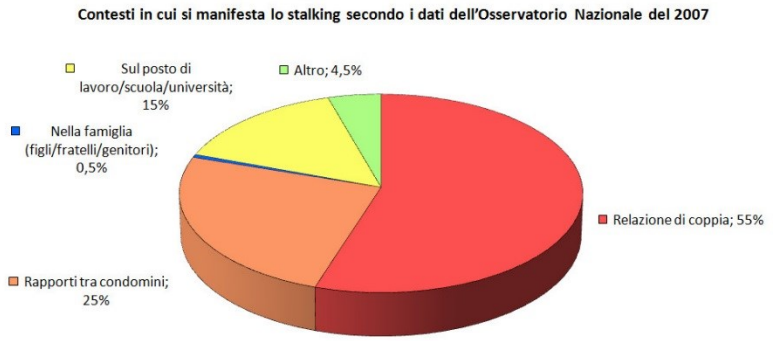
Marginali di Riga

Marginali di colonna

Rappresentazioni grafiche dei dati

Una parte molto importante della statistica descrittiva è la **rappresentazione grafica** dei dati

1. Scoprire la **struttura** dei dati;
2. Identificare **i valori più rilevanti** assunti dal fenomeno;
3. Identificare eventuali **valori anomali** (*outliers*)
4. **Comunicare** i risultati dello studio in modo efficace



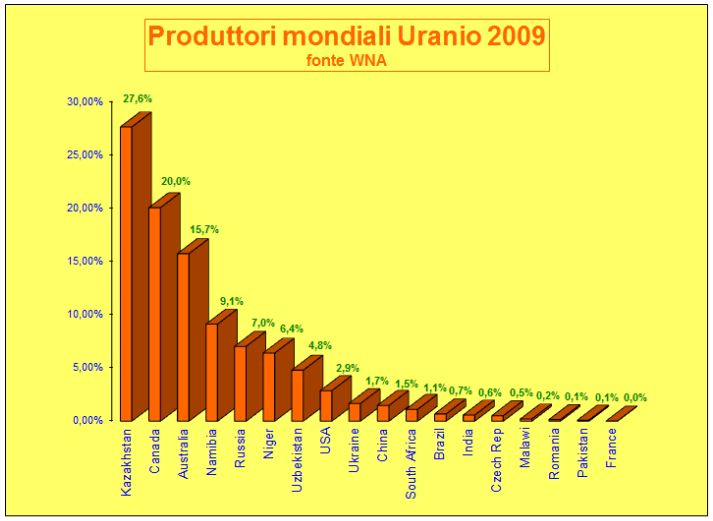
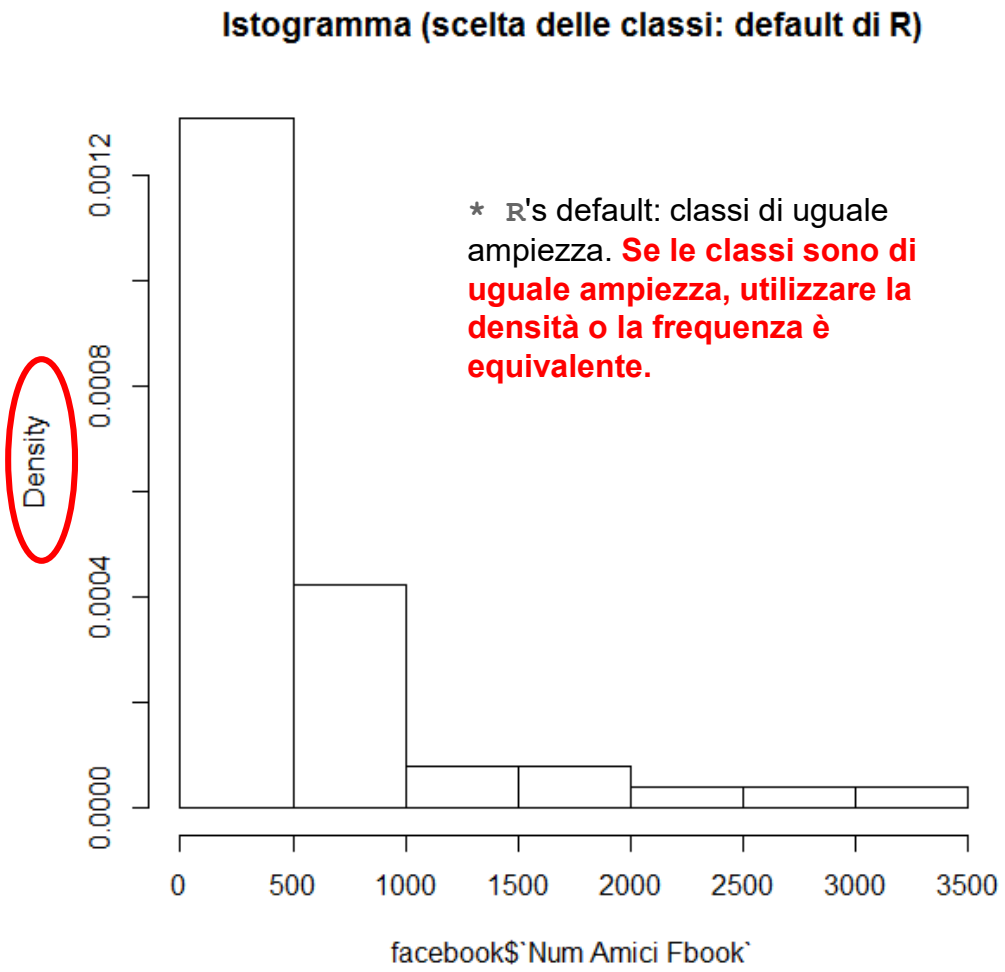


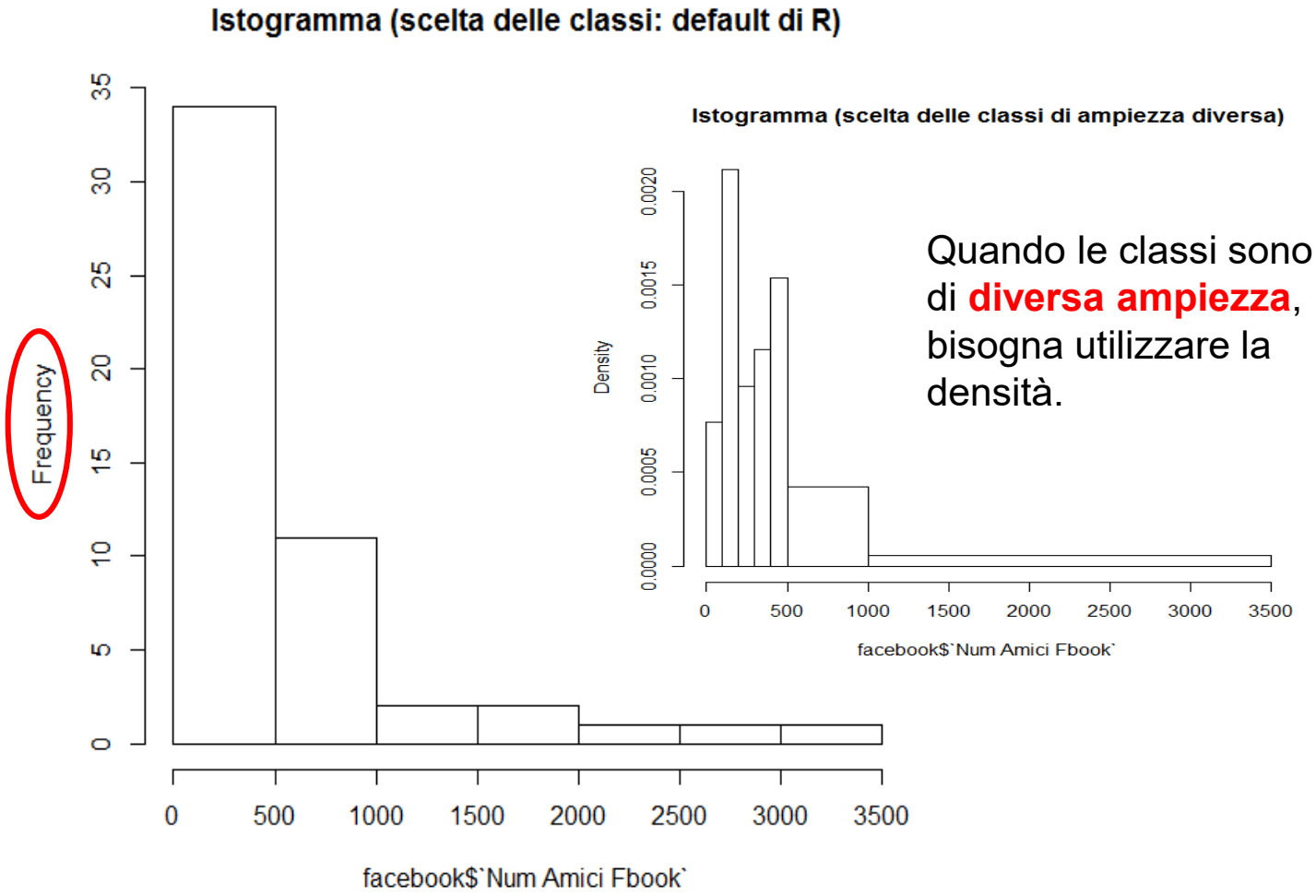
Grafico elaborato da Terenzio Longobardi a partire dai dati WNA

Caratteri su scala numerica:

Per rappresentare graficamente una distribuzione secondo un carattere di tipo quantitativo possiamo utilizzare l'istogramma:



* R's default: classi di uguale ampiezza. **Se le classi sono di uguale ampiezza, utilizzare la densità o la frequenza è equivalente.**

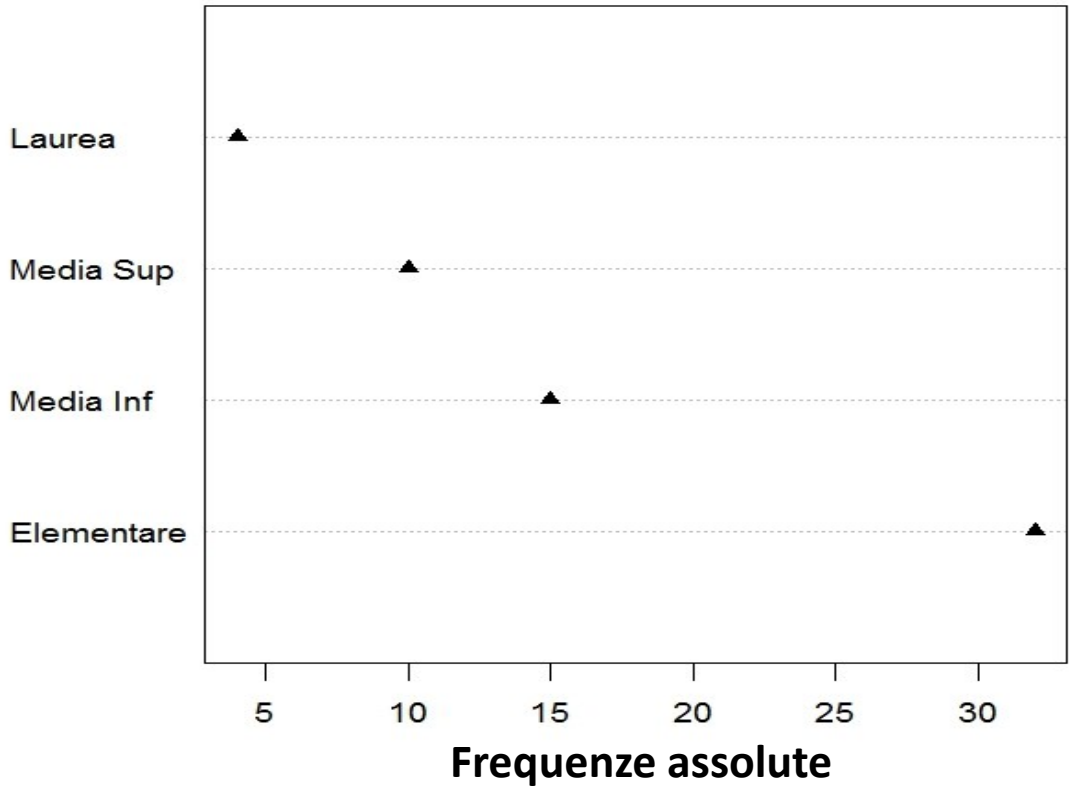


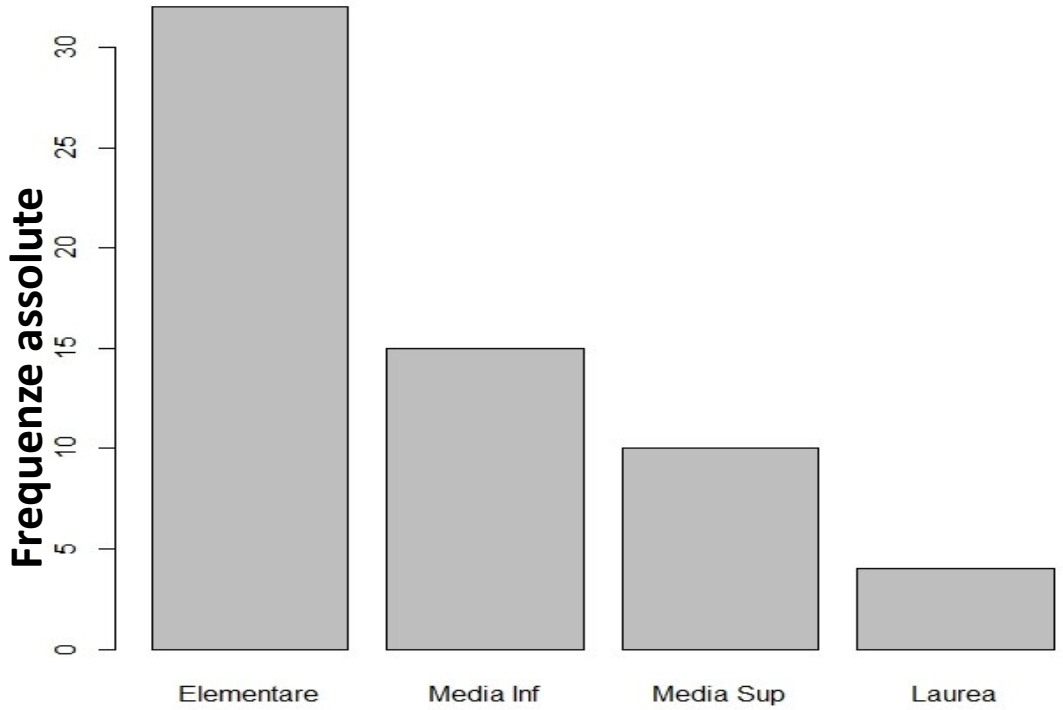
Quando le classi sono di **diversa ampiezza**, bisogna utilizzare la densità.

Caratteri su scala nominale/ordinale (I):

Per rappresentare graficamente una distribuzione secondo un carattere di tipo nominale o ordinale possiamo utilizzare il grafico «dotchart» oppure il «barplot»:

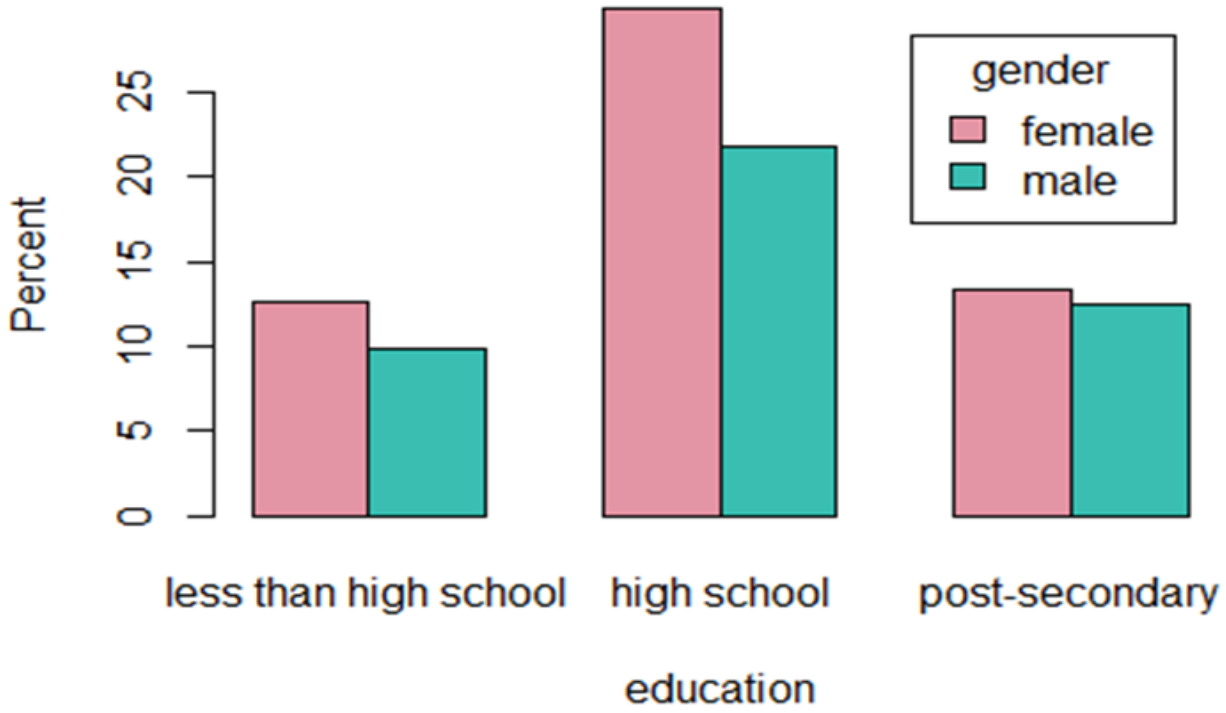
Elementare	Media Inf	Media Sup	Laurea
32	15	10	4





Caratteri su scala nominale/ordinale (II):

Per rappresentare graficamente una distribuzione doppia di due caratteri nominali/ordinali possiamo utilizzare il grafico «barplot» [in questo esempio utilizzo le frequenze percentuali]:

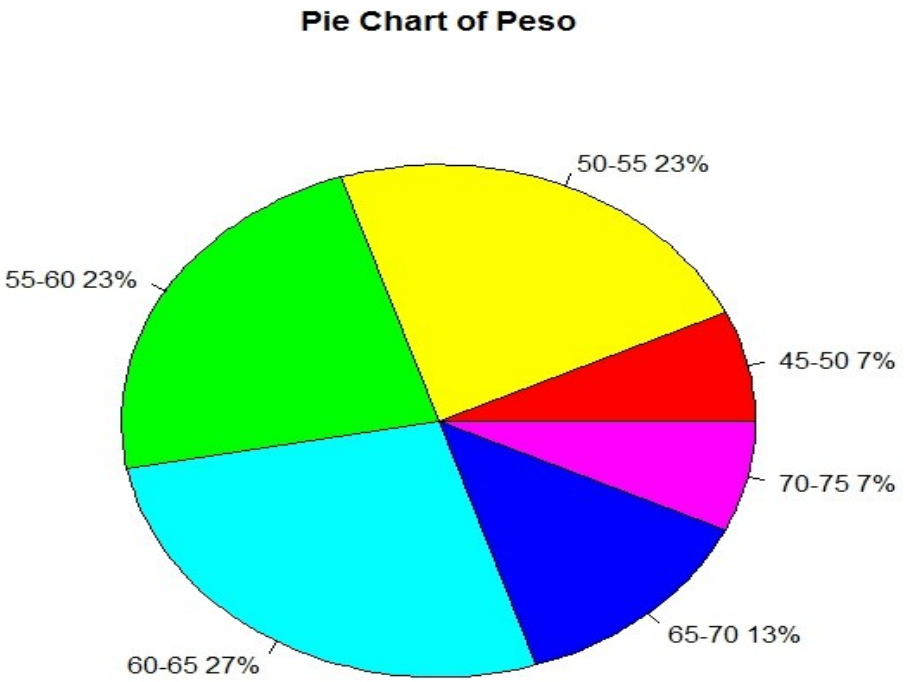


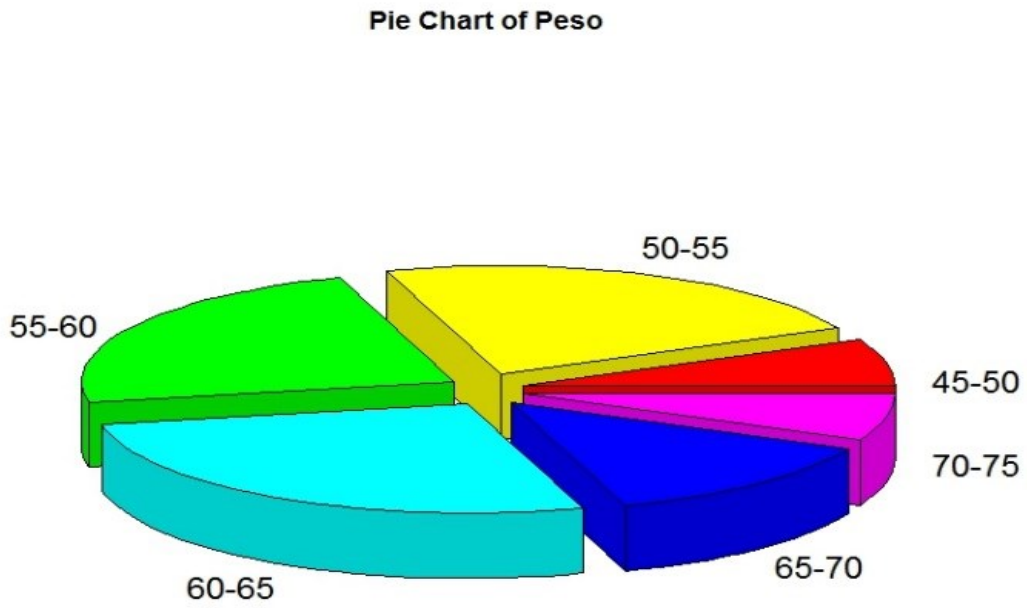

```
Frequency table:
                          gender
education                  female male
  high school                9987 7264
  less than high school      4228 3280
  post-secondary             4436 4159

Total percentages:
                          female male Total
high school                 29.9 21.8  51.7
less than high school       12.7  9.8  22.5
post-secondary              13.3 12.5  25.8
Total                       55.9 44.1 100.0
```

Rappresentazione grafica di distribuzioni in %

Per rappresentare una distribuzione in percentuale secondo un carattere di tipo categorico/in classi possiamo anche utilizzare il grafico a torta (pie chart):



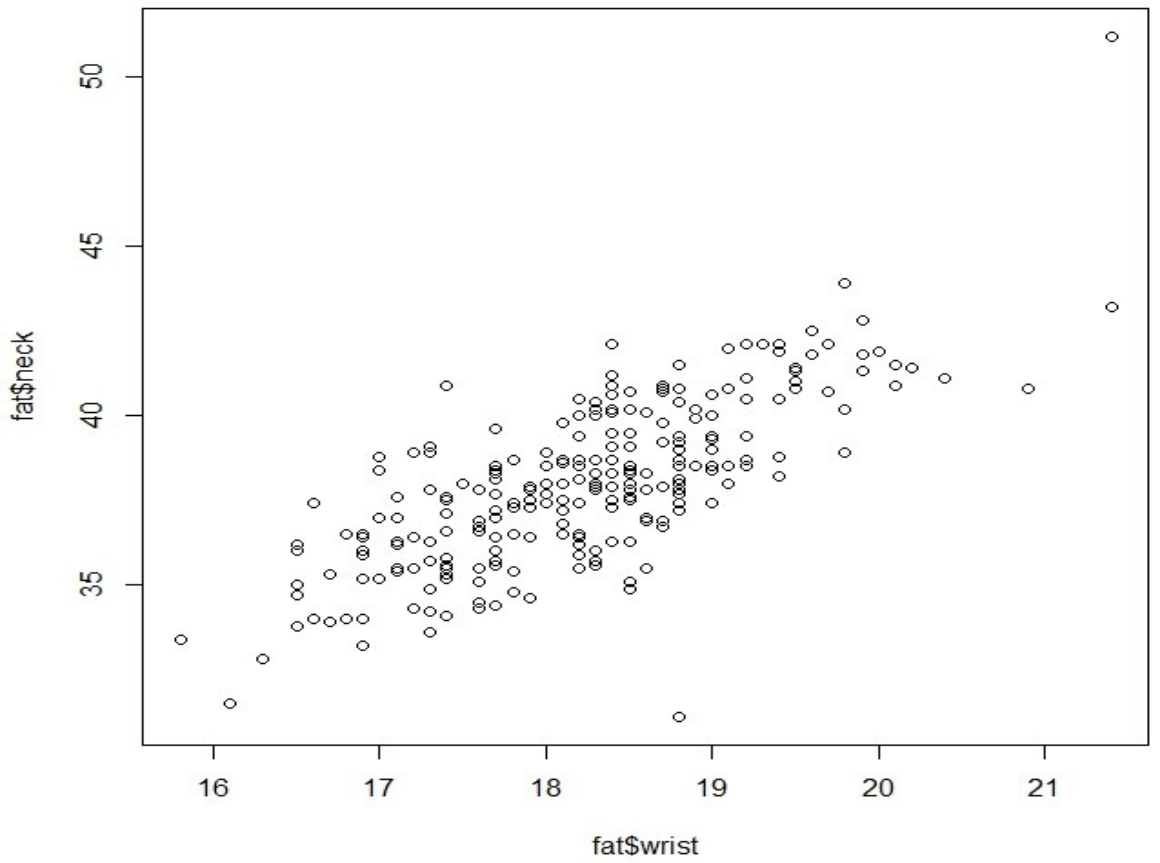


Distribuzioni doppie (dati su scala numerica)

Se si rilevano **due** caratteri in una popolazione su scala numerica e si è interessati a studiare l'associazione tra i due caratteri, la rappresentazione grafica usuale è lo **scatter plot (diagramma cartesiano)**:

Distribuzione su un campione di 252 maschi delle dimensioni del polso e del collo: si ipotizza che l'ampiezza del collo sia circa due volte quella del polso.

Vedremo in seguito come sintetizzare da un punto di vista statistico queste relazioni (**correlazione**/**regressione lineare…**)



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 più piccolo e quello più grande tra quelli osservati (se è possibile ordinarli); deve identificarsi, in qualche modo, con i valori più 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:

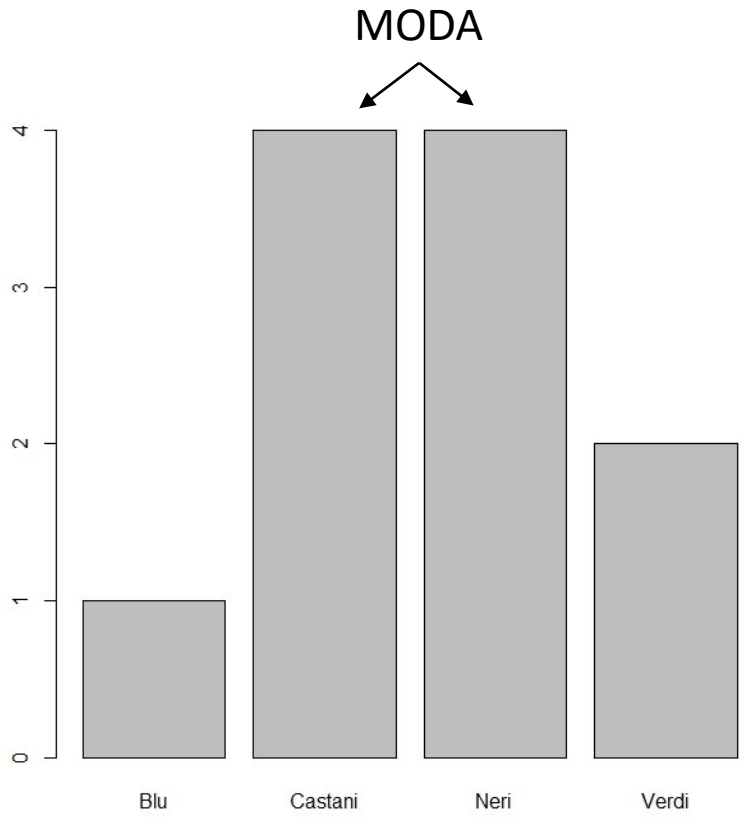
Dato un qualsiasi tipo di carattere (qualitativo o quantitativo) la **MODA** della popolazione distribuita secondo quel carattere è la **modalità prevalente** del carattere, ossia quella a cui è associata la massima frequenza.

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

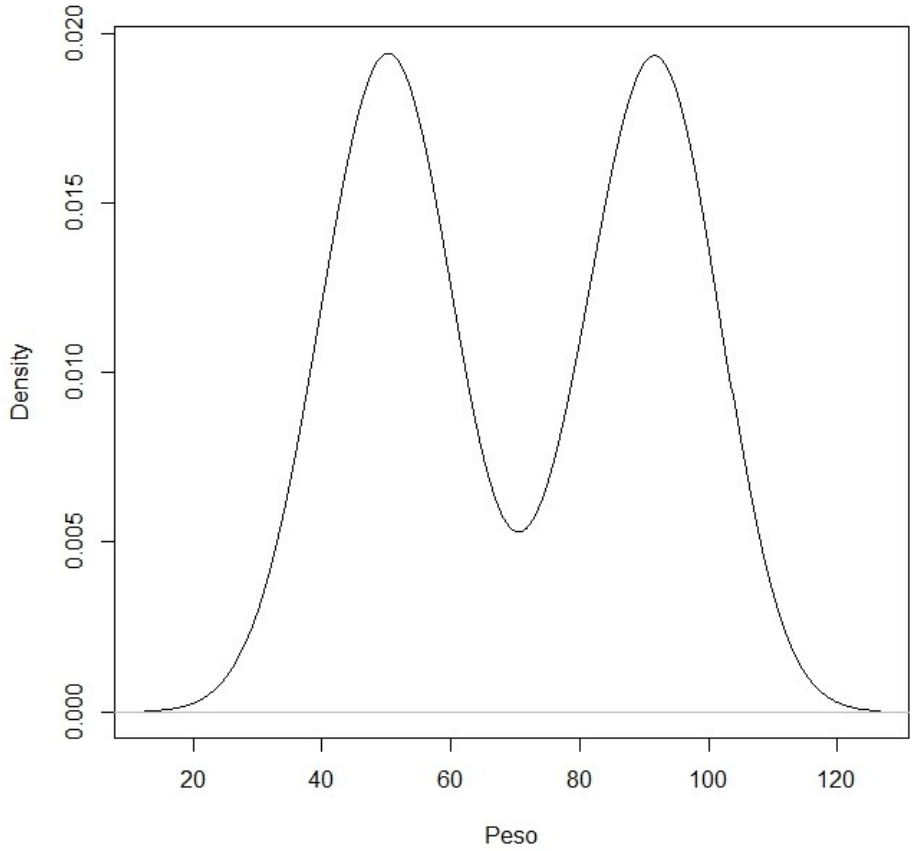
Colore degli occhi	Frequenza di studenti con quel colore di occhi
Blu	1
Castano (Moda o Classe Modale)	**4**
Nero (Moda o Classe Modale)	**4**
Verde	2
totale	11

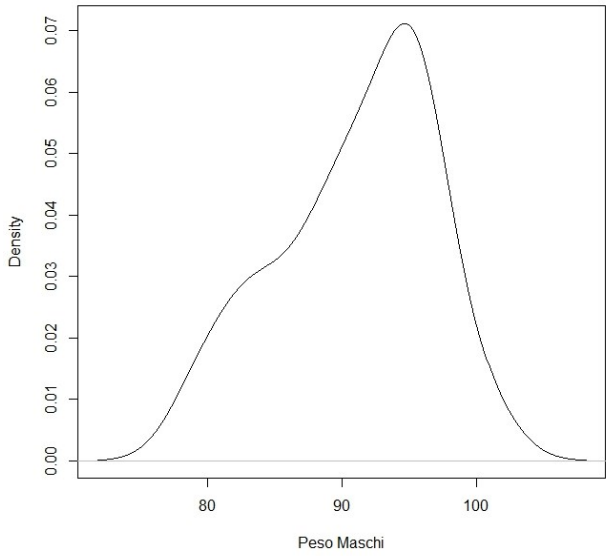


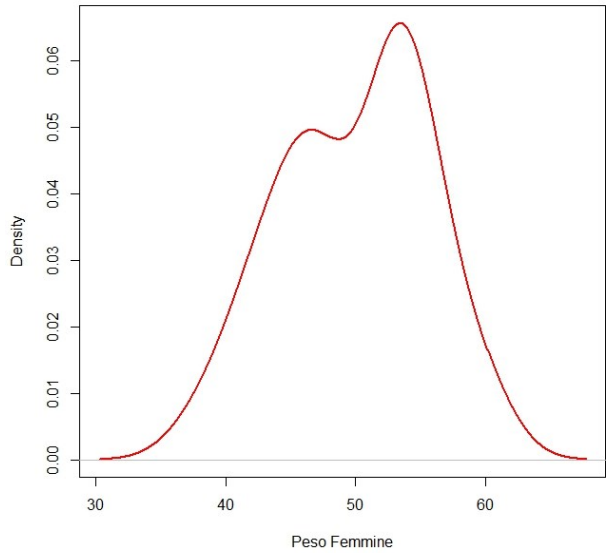
…in questo caso la distribuzione viene definita *bi-modale*, cioè ha due mode.

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



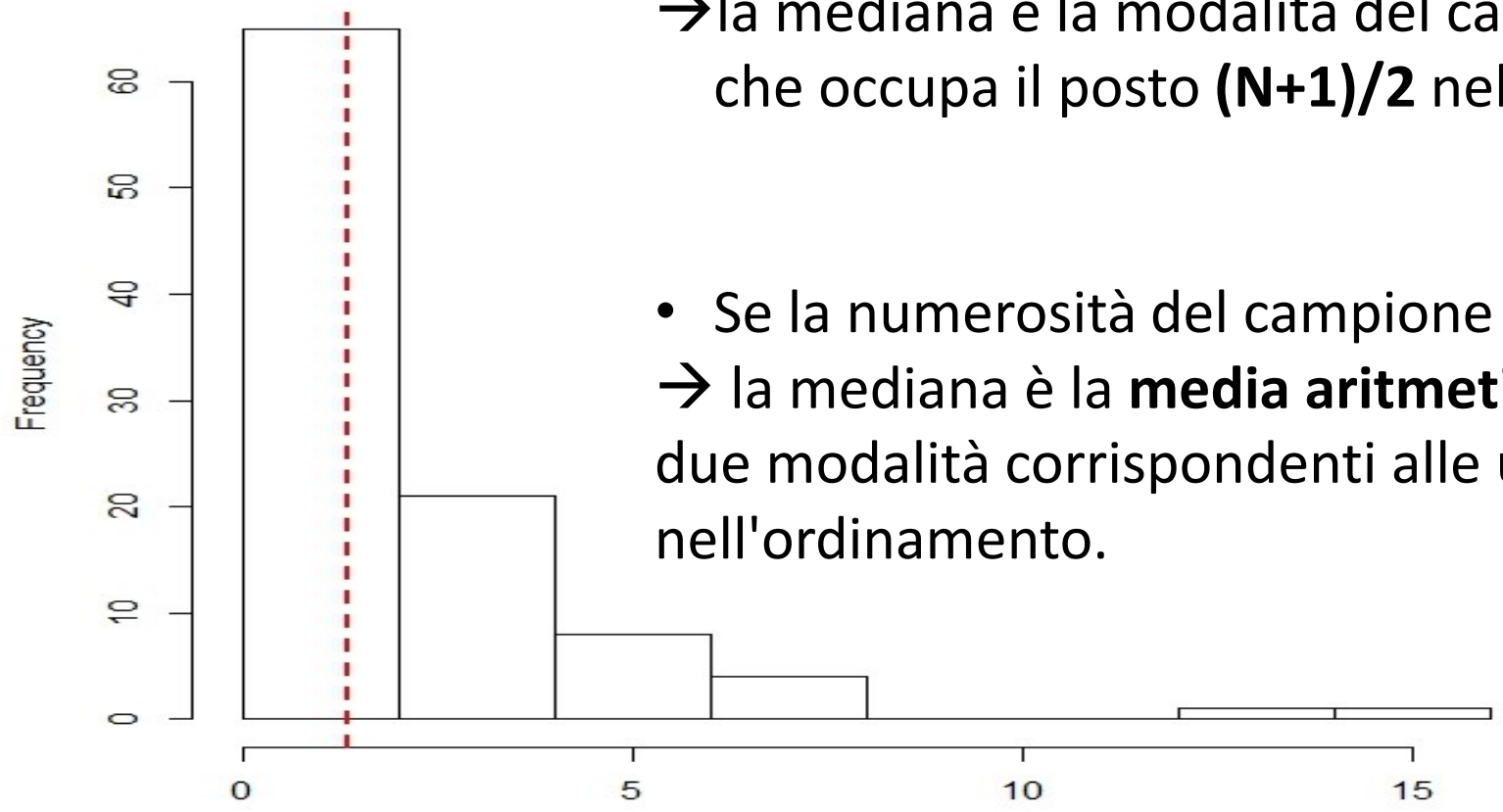




La moda ci aiuta a capire se la distribuzione è **omogenea** oppure no!

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;

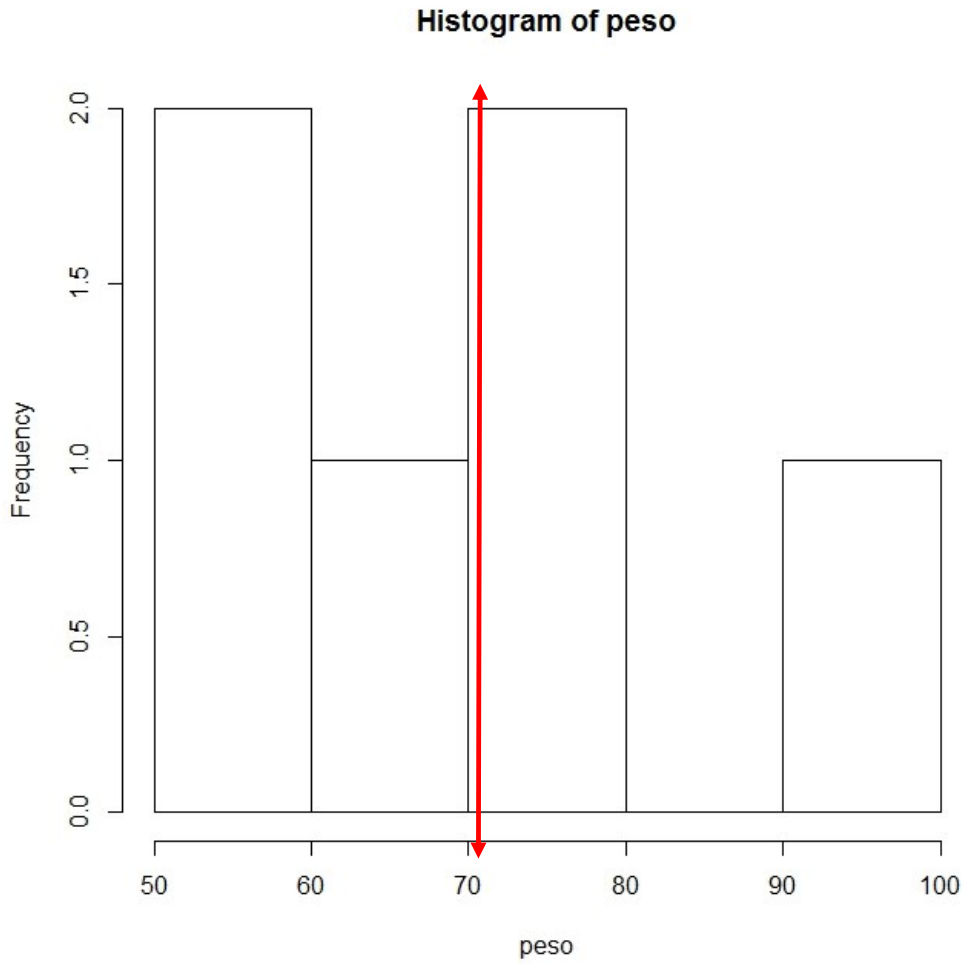
- Se la numerosità del campione **N** è pari,

→ 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



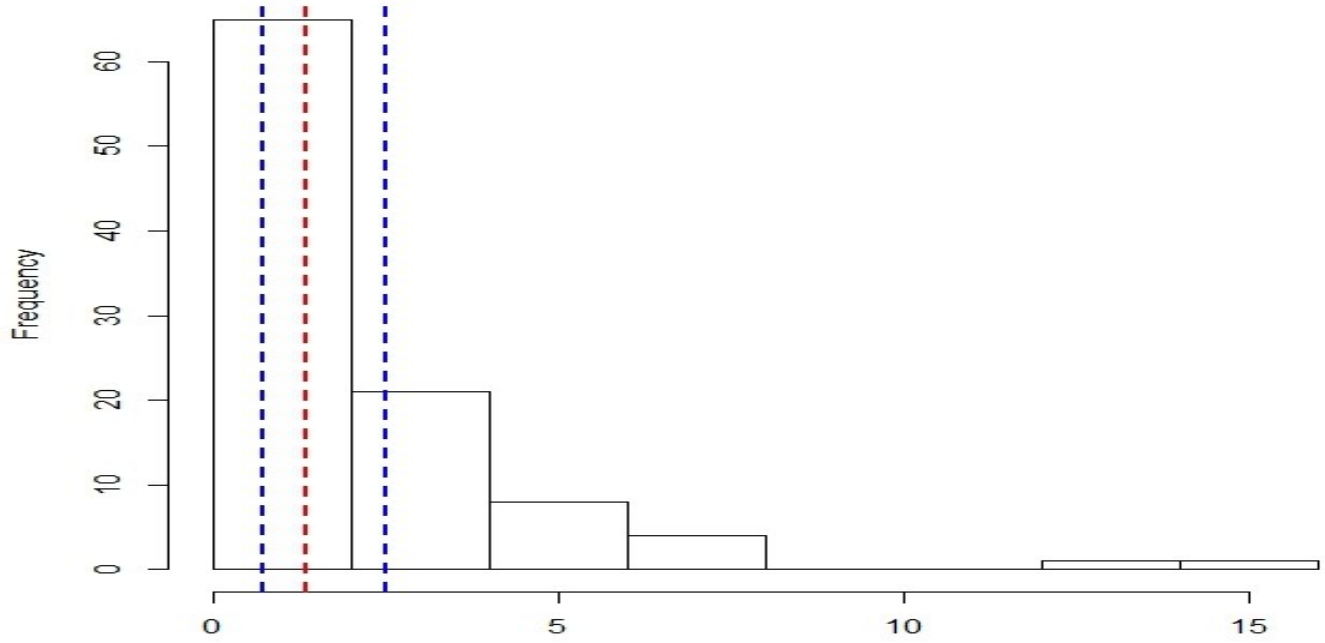
Il 50% del campione esaminato ha un peso corporeo inferiore o uguale a 71,5 Kg (*sintesi* dei dati).

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.

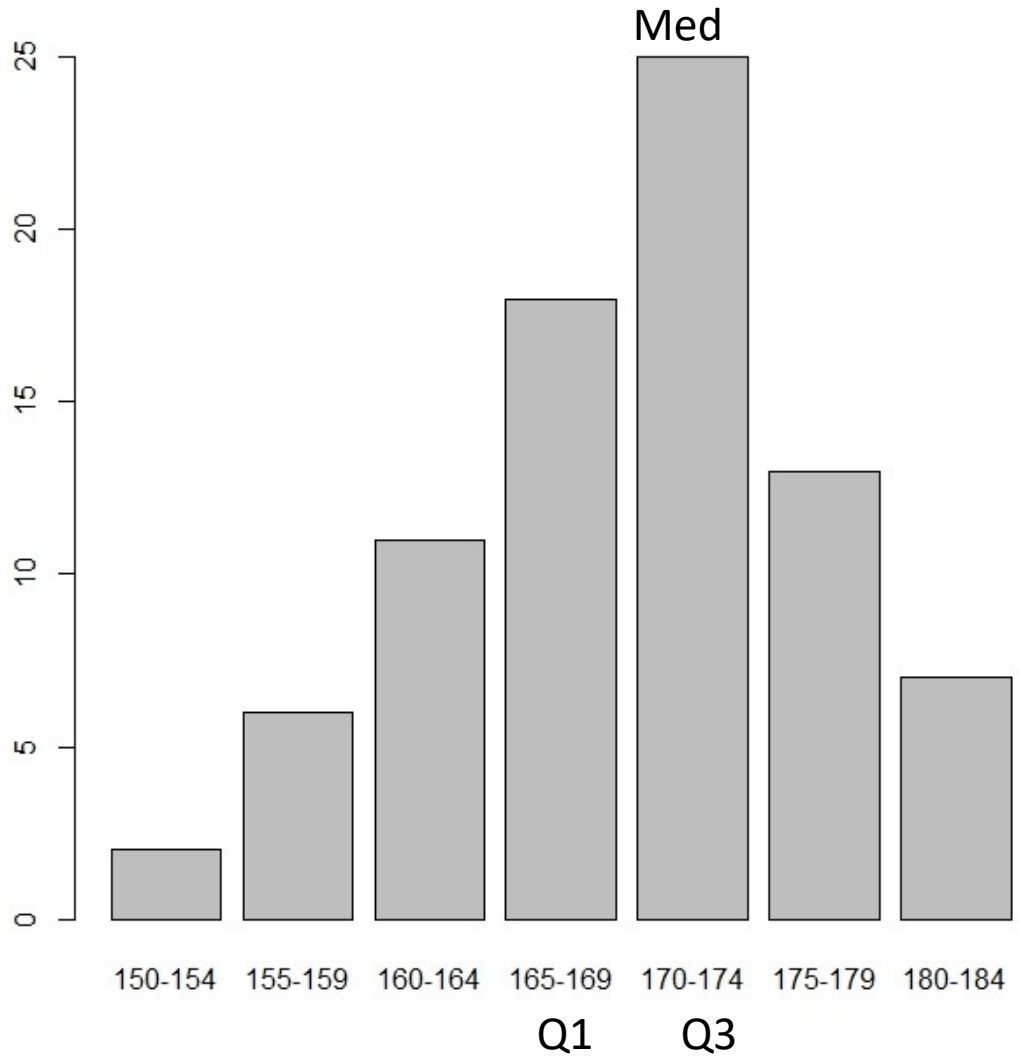
I '**<u>quartili</u>**' sono gli indici di posizione che dividono una serie ordinata di dati in **4** parti uguali:

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



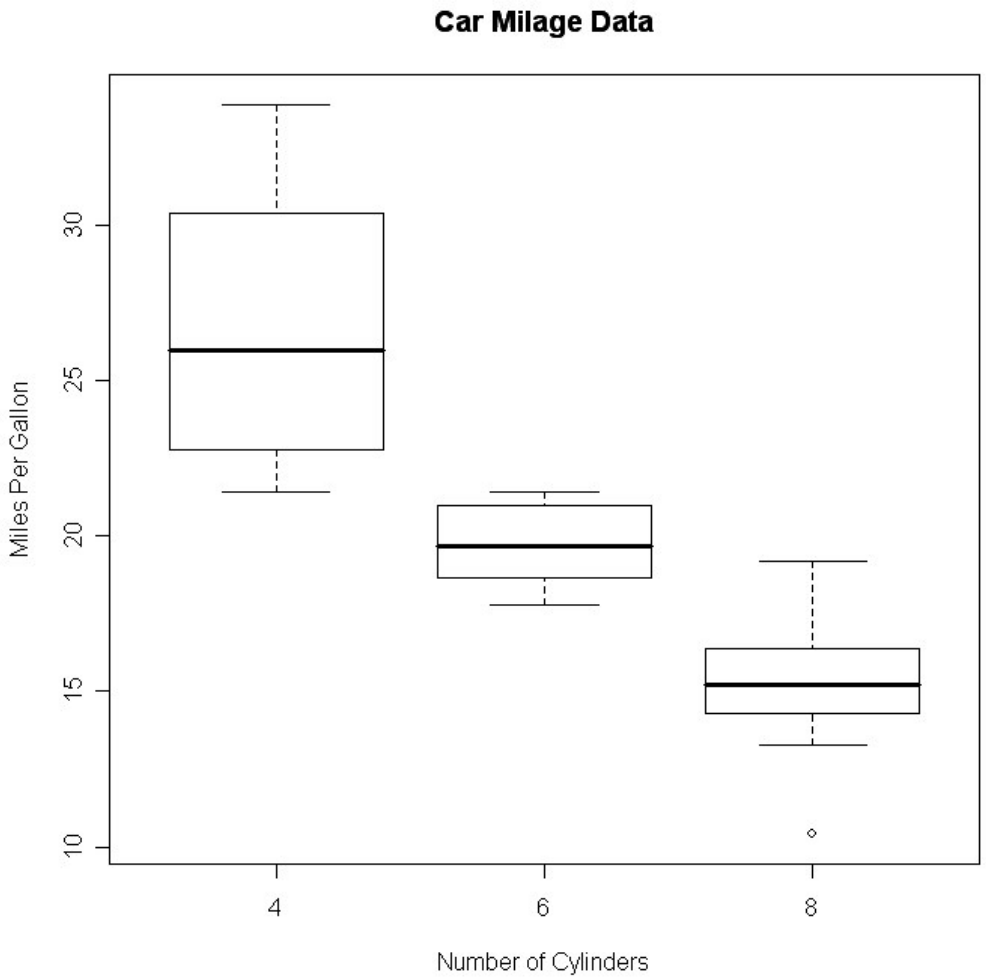
Gli indici di posizione delle distribuzioni (III): Quantili/Percentili

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:
$x_1, x_2, \ldots, x_n$

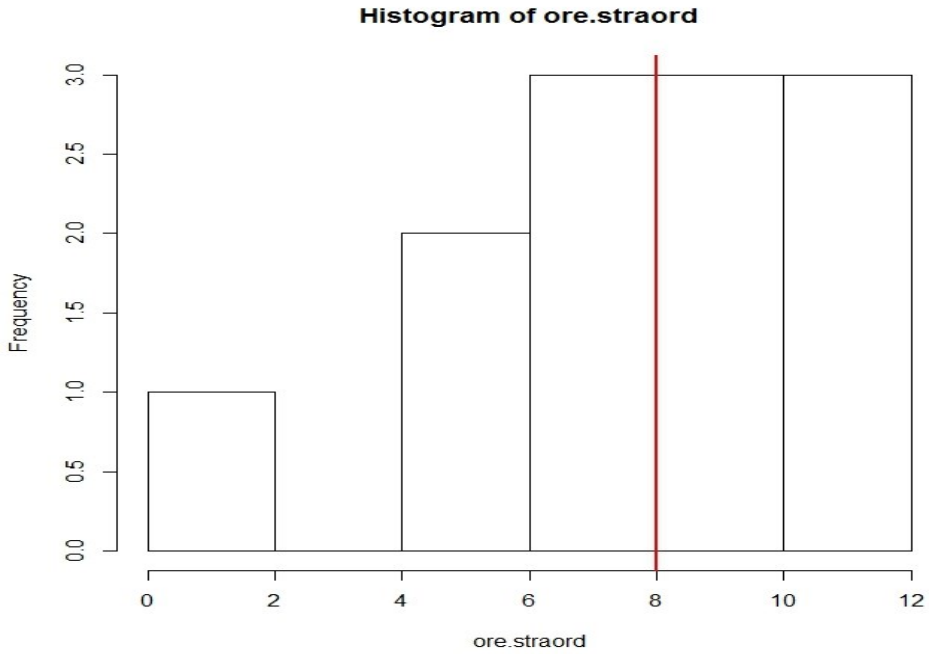
$$\bar{x} = \frac{(x_1 + x_2 + \ldots + x_n)}{n} \quad \Longleftrightarrow \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Un tecnico di laboratorio 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?

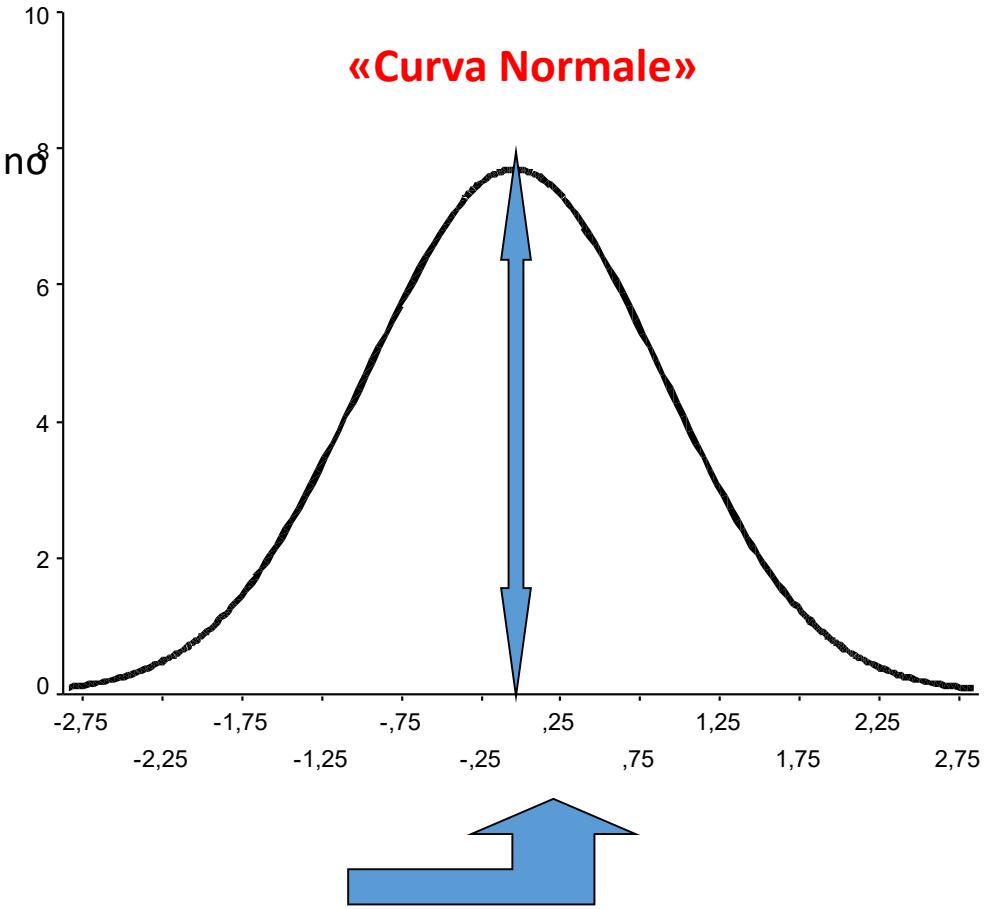
$$\bar{x} = \frac{(10+12+11+5+7+10+5+0+7+12)}{12} = \frac{96}{12} = 8$$



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:

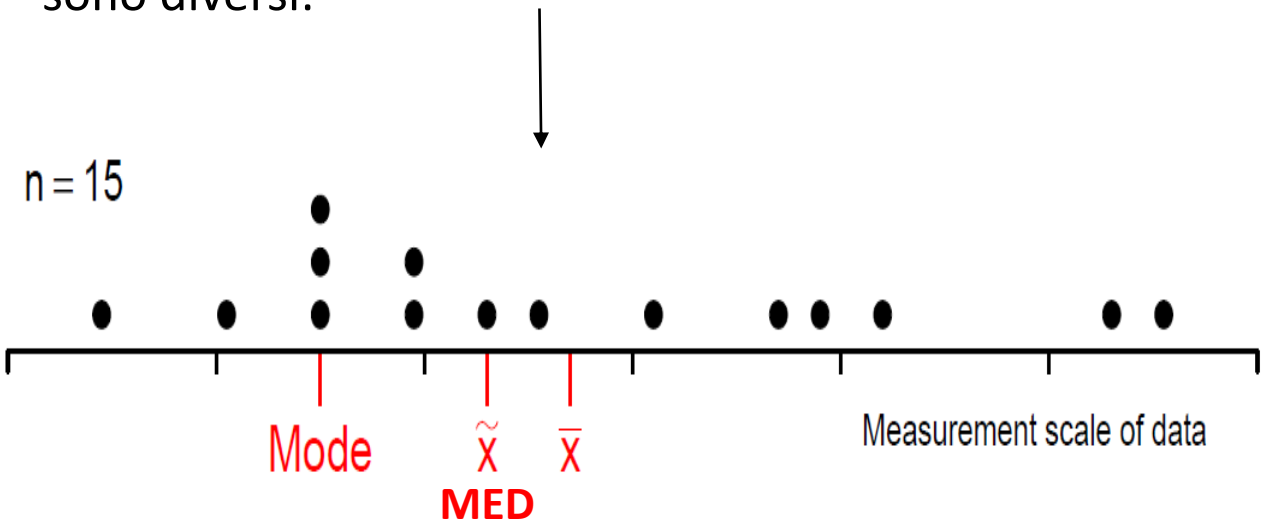
Ex: distribuzione simmetrica e unimodale
Moda=Mediana=Media

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:



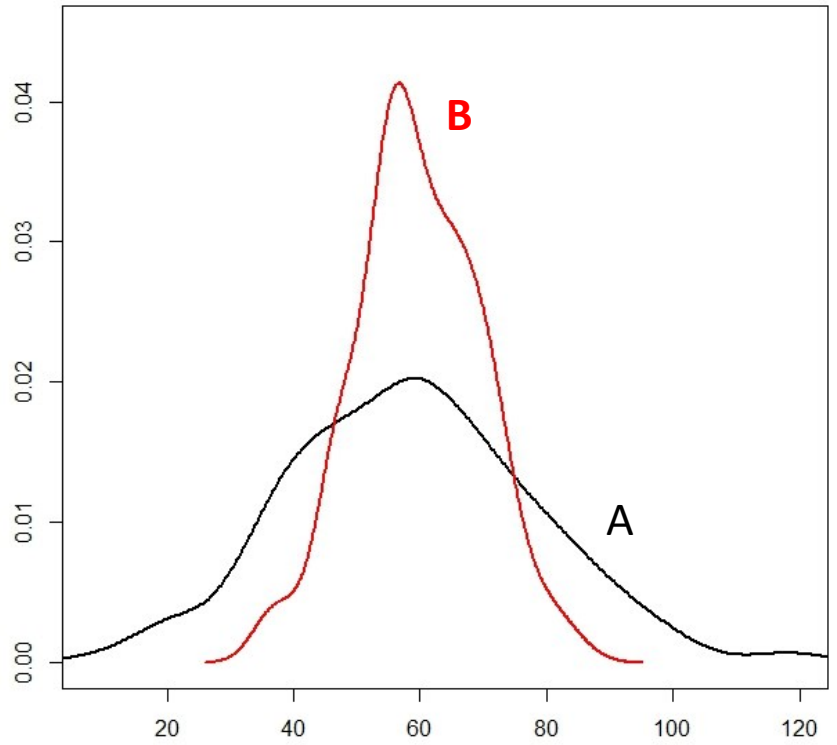
Gli indici di dispersione

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.

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

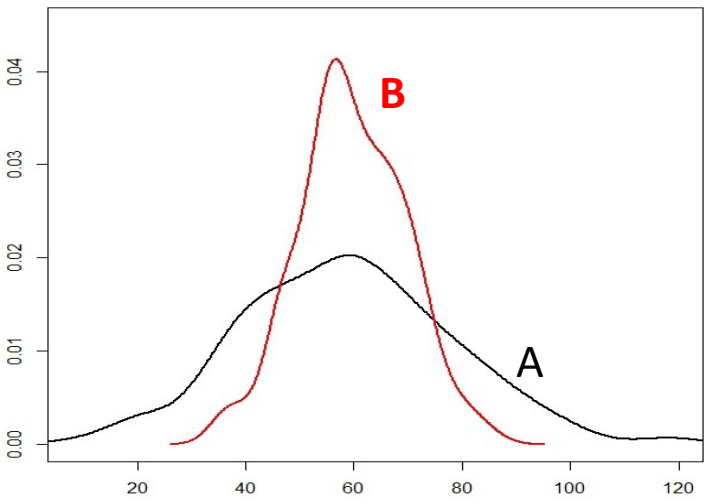
Le varianze diventano cosi' *confrontabili* tra diverse distribuzioni e si puo' stabilire quale, tra le due o piu' serie di misure considerate, presenta una maggiore dispersione rispetto alla media, **indipendentemente da N**.

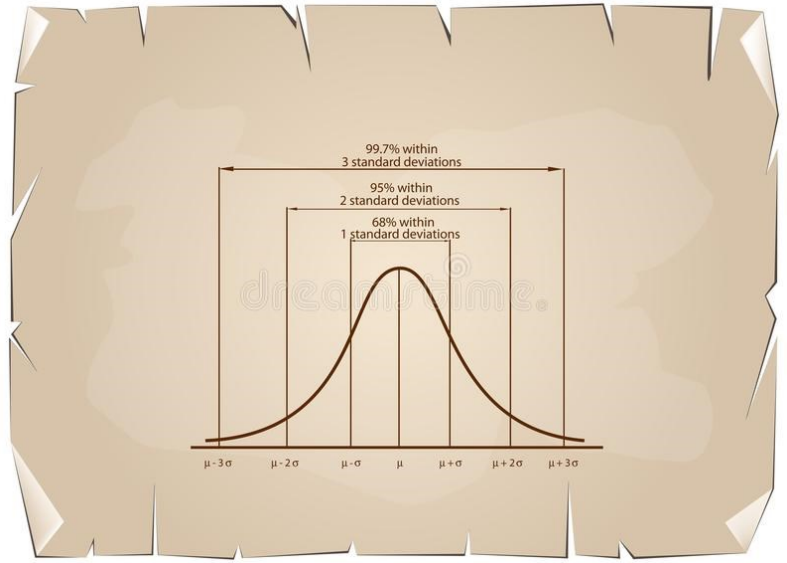
$$dev\ st = s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

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.

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:

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:

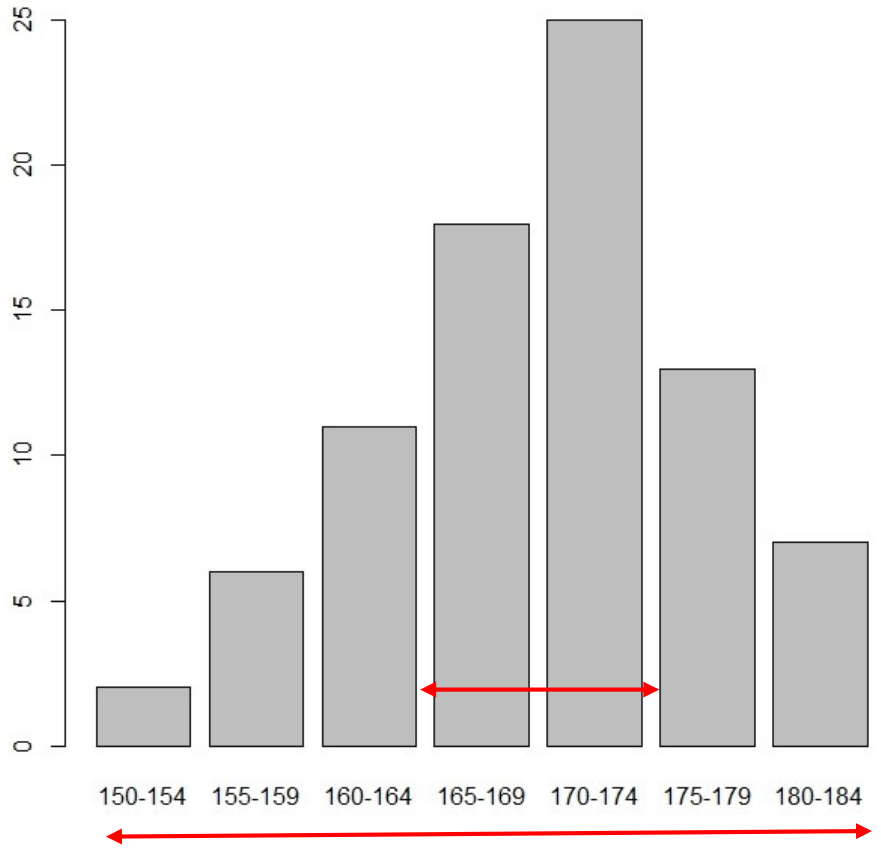
DIFFERENZA INTERQUARTILE (**RANGE INTERQUARTILE, IQR**) la distanza tra il primo ed il terzo quartile:

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

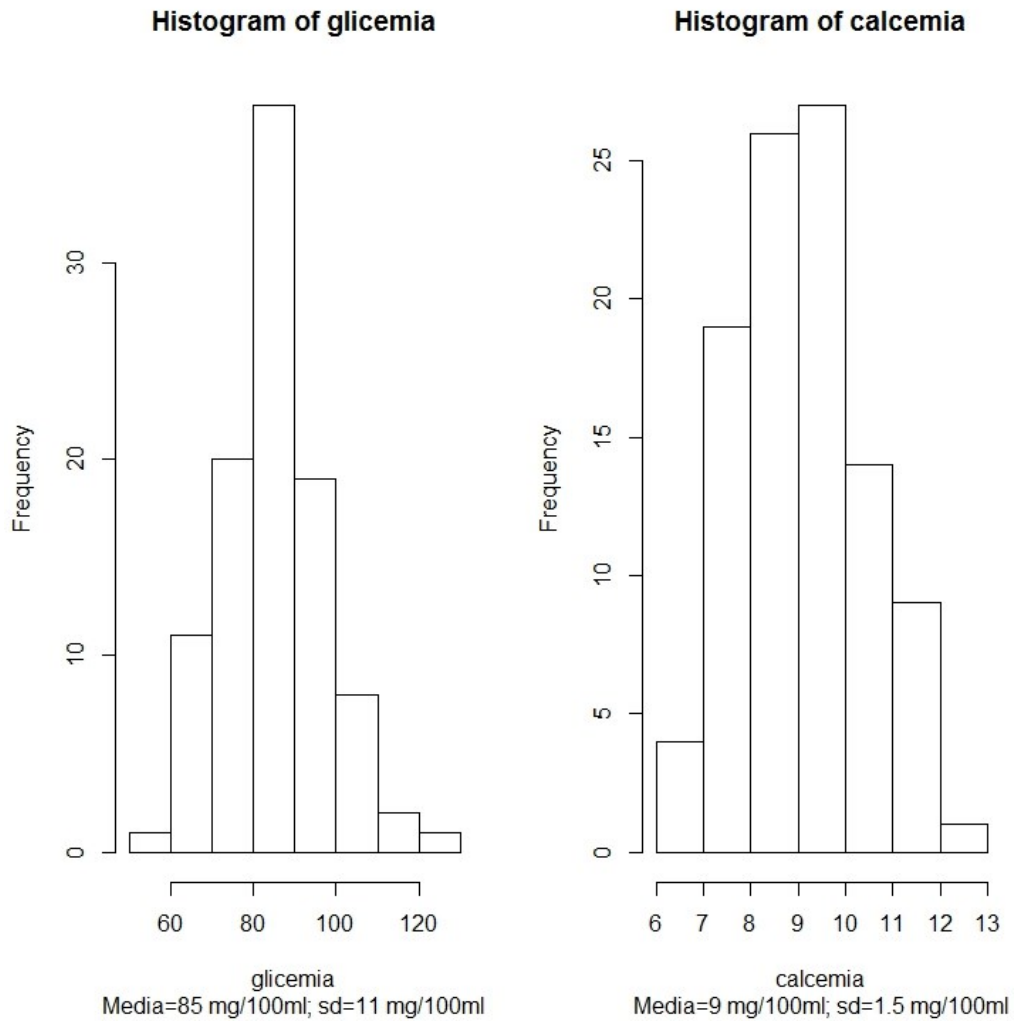
CV di Glicemia

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

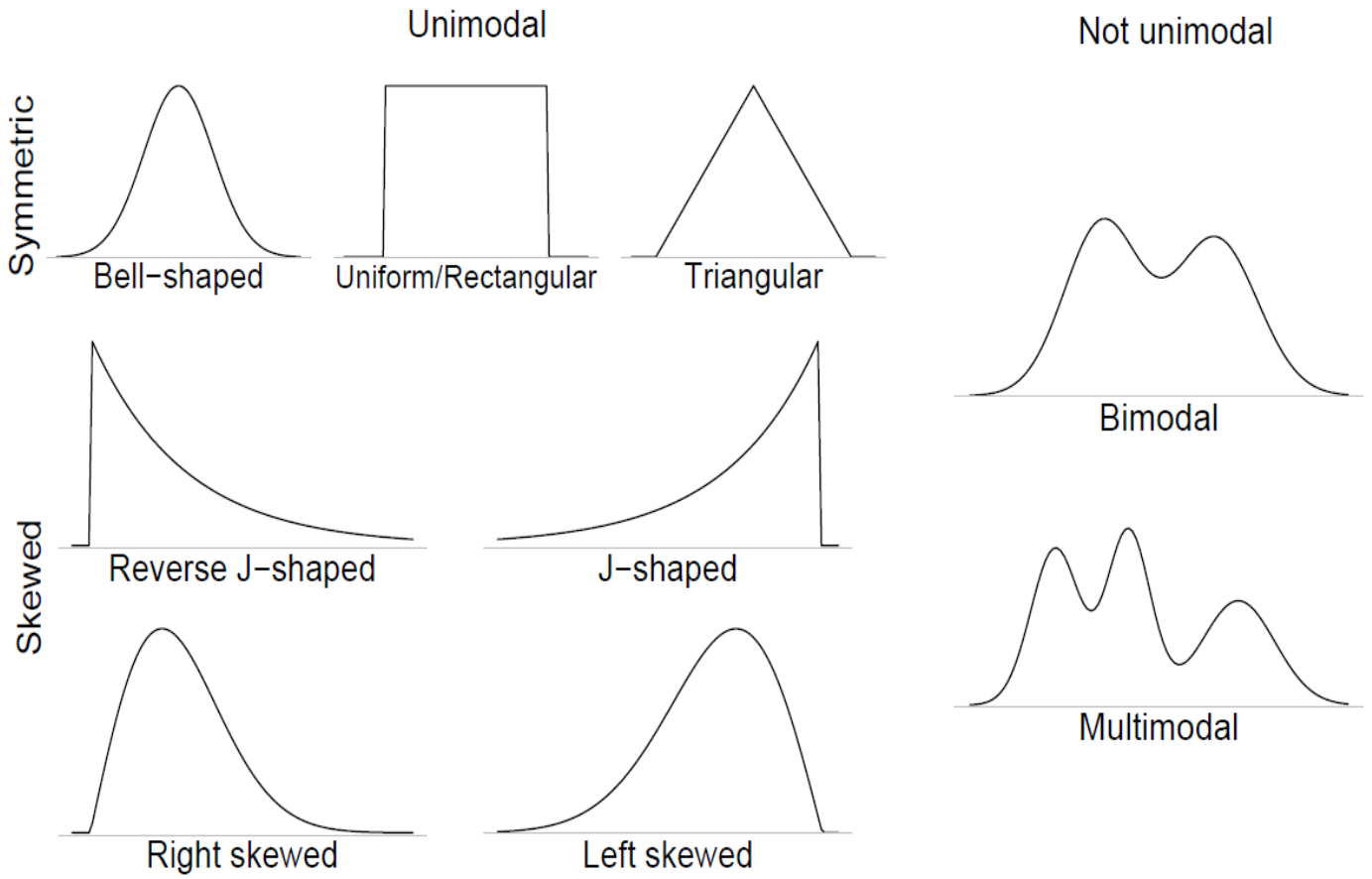
CV di Calcemia

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

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).

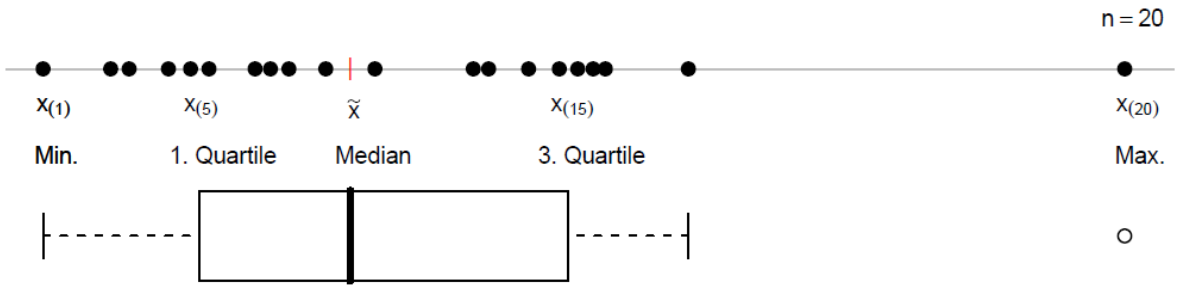


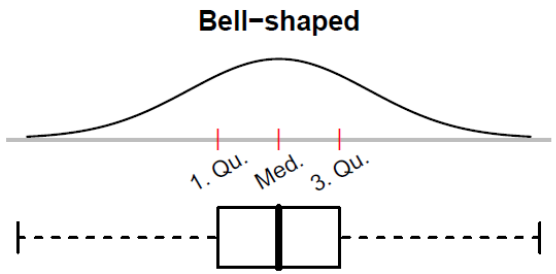
Quali indici di posizione sono calcolabili per i vari tipi di caratteri:

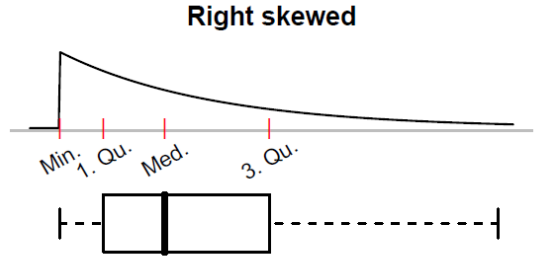
CARATTERE		INDICE DI POSIZIONE
qualitativo	Sconnesso	Moda
	Ordinabile	Moda, mediana, quartili
quantitativo		Moda, mediana, quartili, media

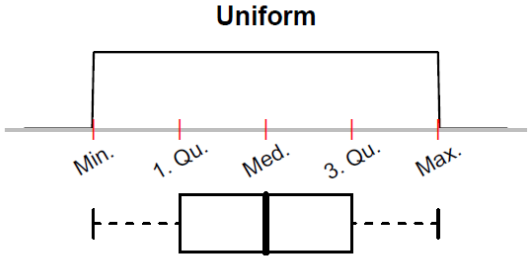
E’ **importante** ricordare l’associazione tra il tipo di carattere ed i relativi indici di dispersione che possono essere calcolati:

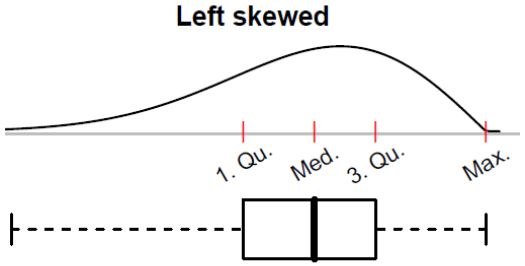
CARATTERE		INDICE DI DISPERSIONE
qualitativo	Sconnesso	nessuno
	Ordinabile	RANGE / IQR *[modalità inferiore/superiore oppure modalità relative a Q1 e Q3 del carattere qualitativo ordinabile, non certo come differenza tra esse !!]*
quantitativo		RANGE / IQR DEVIANZA, VARIANZA, DEVIAZIONE STANDARD e CV%











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