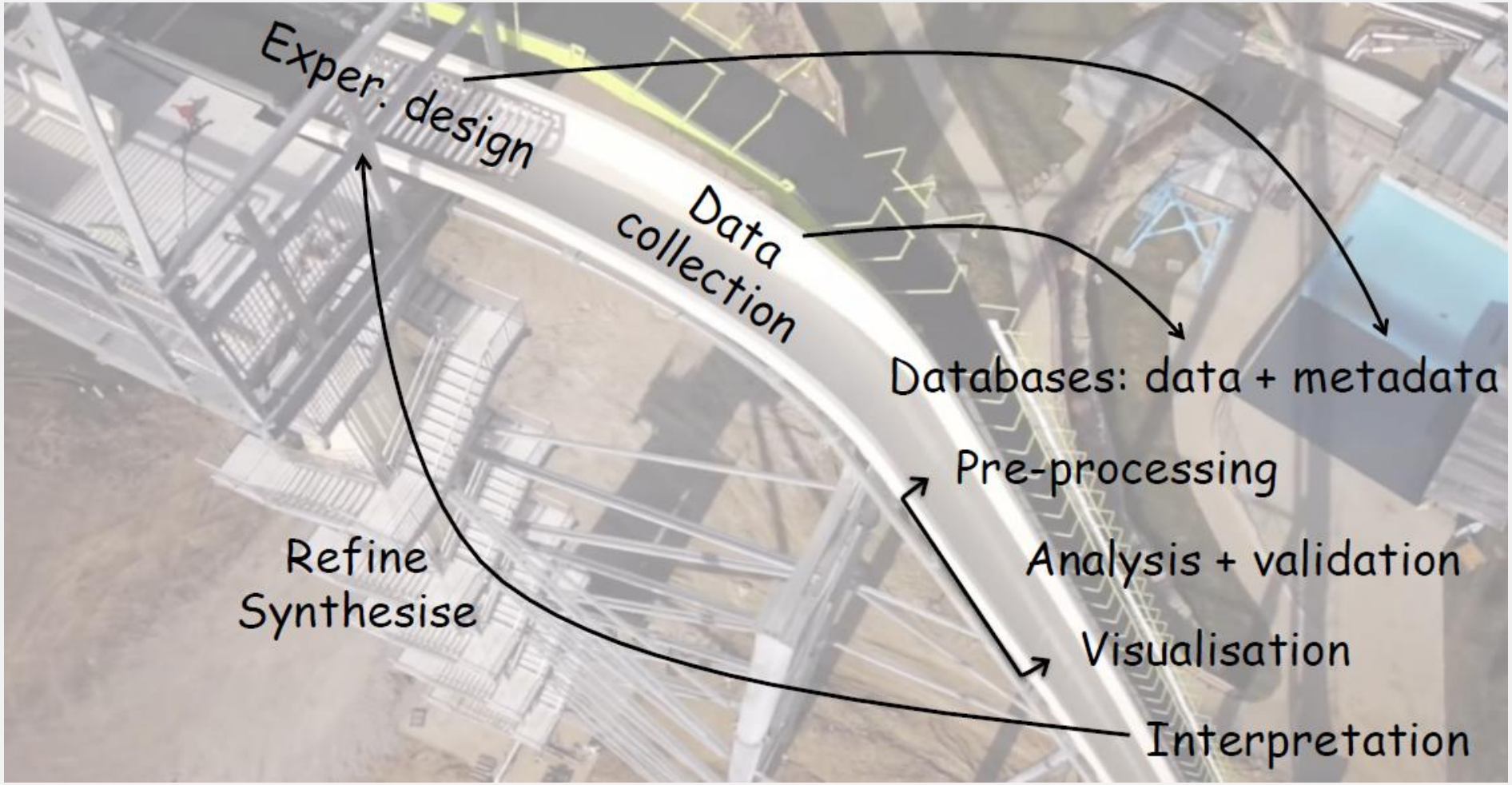
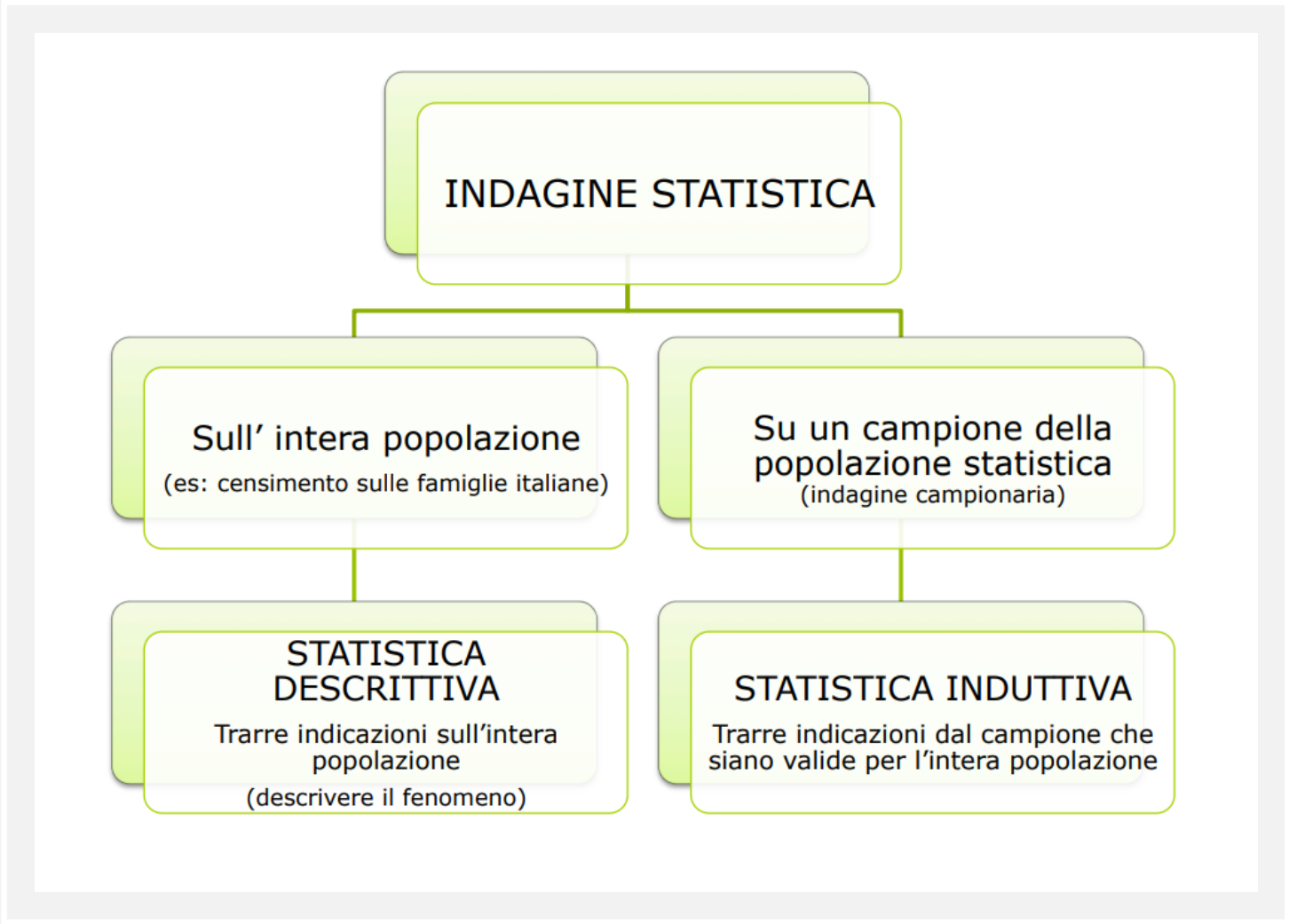
- **ELEMENTI DI STATISTICA**

DATA ANALYSIS 'SLIDE'/'PIPELINE'



LA STATISTICA

- Scienza che ha per oggetto lo studio dei fenomeni collettivi che possono essere misurati e descritti qualitativamente e quantitativamente. Mediante l'applicazione di metodi matematici vengono formulate le cosiddette leggi statistiche che governano questi fenomeni.

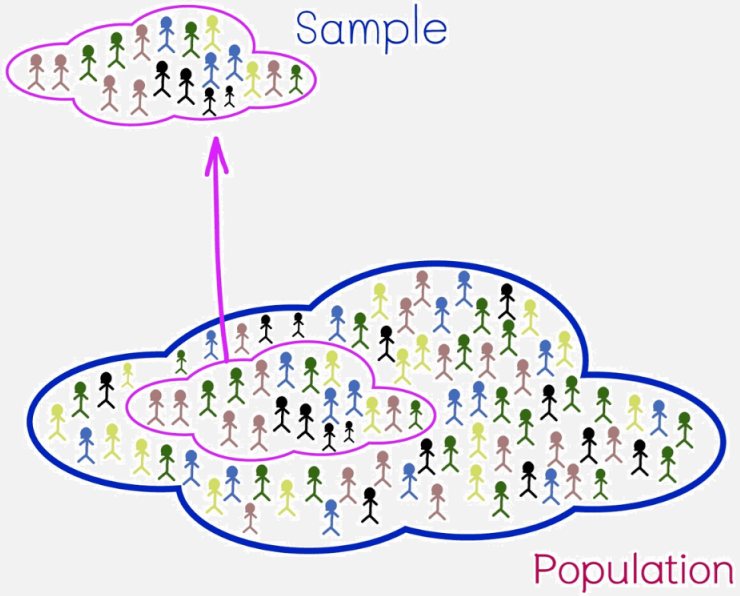


Popolazione, unità, campione statistico

- **Popolazione statistica**: insieme degli elementi a cui si riferisce l'indagine statistica:
  - opinione degli americani riguardo una nuova elezione presidenziale: tutti i cittadini USA
  - geni sovra-espressi nelle persone che soffrono di obesità: tutte le persone obese
  - …

- **Unità statistica**: ogni elemento della popolazione statistica, la minima unità della quale si raccolgono i dati:
  - Un cittadino, una persona obesa….

- **Campione statistico (sample)**: un qualsiasi insieme di unità statistiche prese da tutta la popolazione. Un campione è dunque un sottoinsieme di misurazioni selezionate dalla popolazione
  - 50 persone con problemi di obesità (estratte a caso).



Variabile casuale

- Il *fenomeno collettivo* si presenta secondo modalità diverse nelle varie unità statistiche, perciò lo chiameremo variabile casuale.

- Il valore assunto dalla variabile casuale in una data unità statistica lo chiameremo **osservazione**.

  - Esempio:
  - **variabile casuale**: livello di espressione del gene AAA
  - **osservazione**: il gene AAA della persona X ha un livello di espressione pari a 12.3, il gene AAA della persona Y ha un livello di espressione di 10.2, il gene AAA della persona Z….

Variabile quantitativa o qualitativa

- **Variabile quantitativa**: quando assume valori numerici:
  - **Continua**: assume valori continui in un intervallo (peso e statura di una persona, livelli di intensità dei campioni su microarray, livello di espressione genica, etc.)
  - **Discreta**: assume valori discreti come numero di campioni, numero di geni sovra-espresso, numero di pazienti, etc.

- **Variabile qualitativa**: quando assume valori non numerici
  - **Ordinale**: i dati sono in un ordine (buono-medio-cattivo, freddo-tiepido-caldo…)
  - **Categorica**: uomo/donna, fenotipo, gruppi di pazienti malati/sani, etc.

Types of data

Numerical data or quantitative data

* numerical values

Discrete data

If you can count it, then it is discrete.

Continuous data

If you can measure it, then it is continuous.

* length
* weight
* temperature

Categorical data or qualitative data

* categories

Nominal data

If you can brand it, then it is nominal.

Gender
- [x] Female
- [ ] Male

Colour
- [ ] blue
- [x] red
- [ ] green
- [ ] orange

Ordinal data

If you can rank or order it, then it is ordinal.

- ○ Always
- ✓ Usualy
- ○ Sometimes
- ○ Rarely
- ○ Never

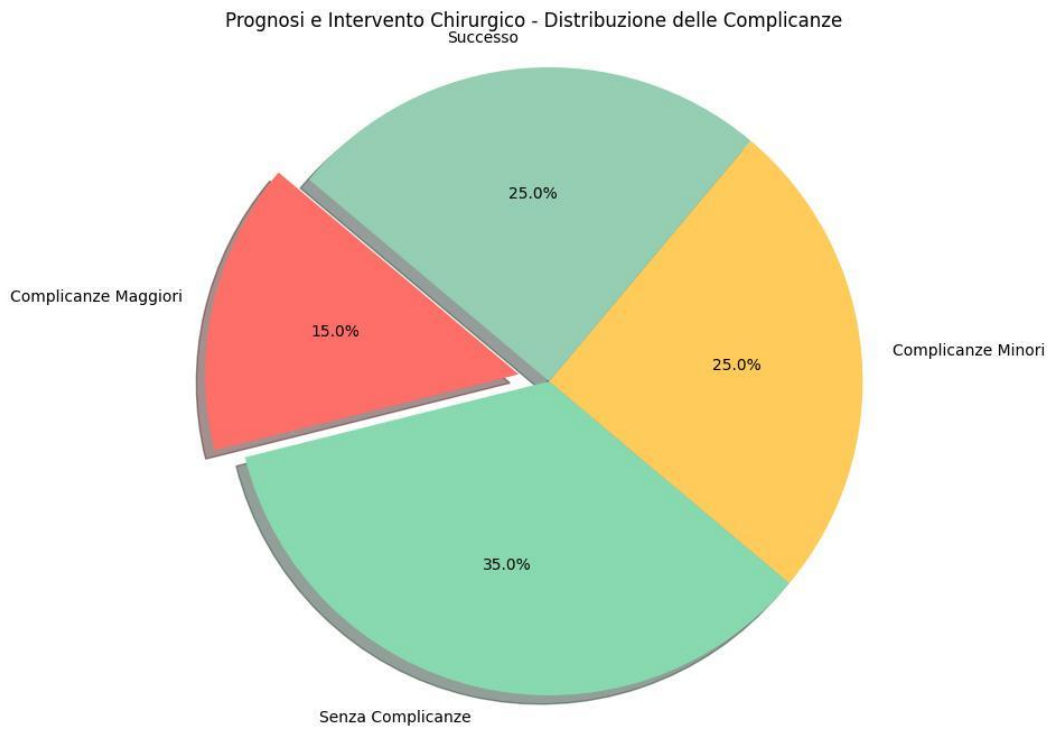
La data table

- I dati codificati di una rilevazione statistica effettuata su n **unità statistiche** con riferimento a p **variabili**, vengono raccolti in una tabella che viene chiamata "**matrice dei dati**"

N.	Sesso	Titolo di studio	Età	Peso	N. Ricoveri
1	M	Licenza media inferiore	36	65	3
2	F	Laurea	45	70	1
…	…	…	…	…	…
N	F	Diploma	60	55	6

Rappresentazione grafica delle variabili

Diagramma a settori o circolare



Esisto intervento	N pazienti	%	Gradi°
Complicanze maggiori	108	15%	54°
Complicanze minori	180	25%	90°
Senza complicanze	252	35%	126°
Successo	180	25%	90°
Totale	**720**	**100%**	**360°**

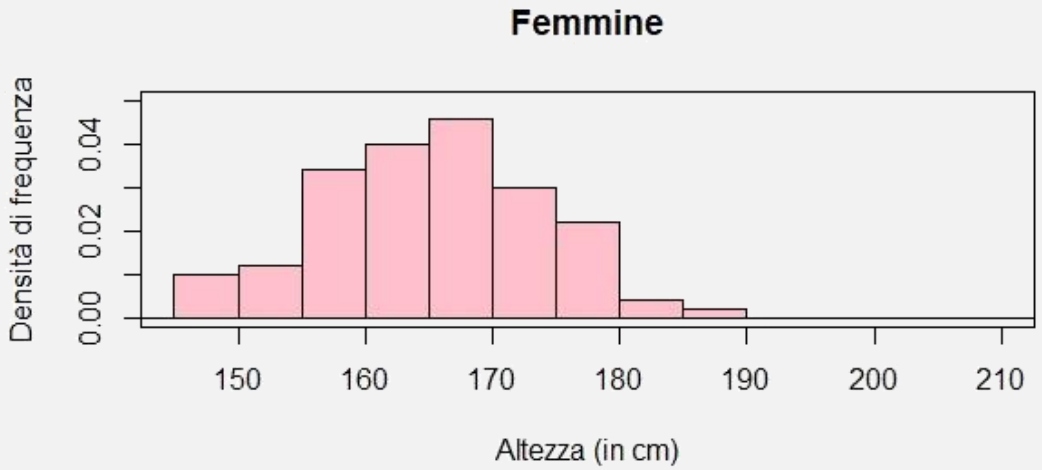
Settore = intervallo di valori

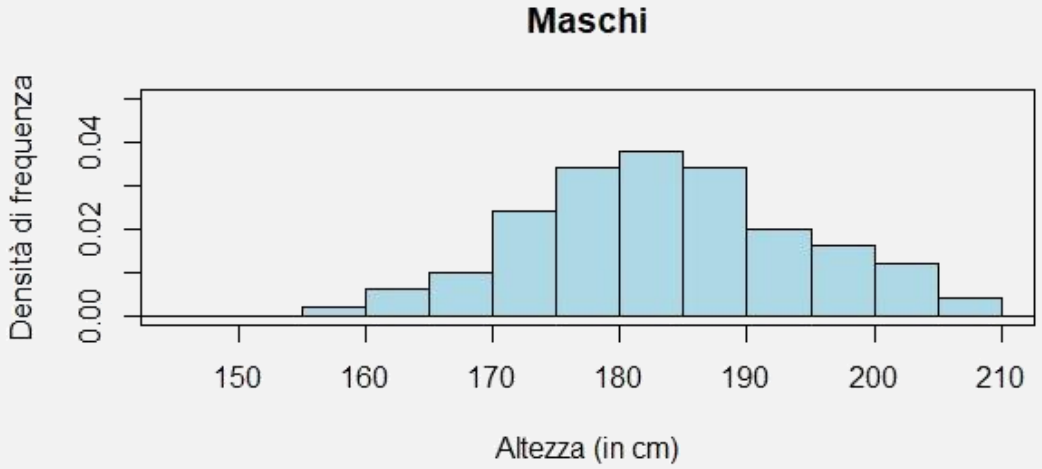
N pazienti : Totale = x : 360° → angolazione di ciascun intervallo

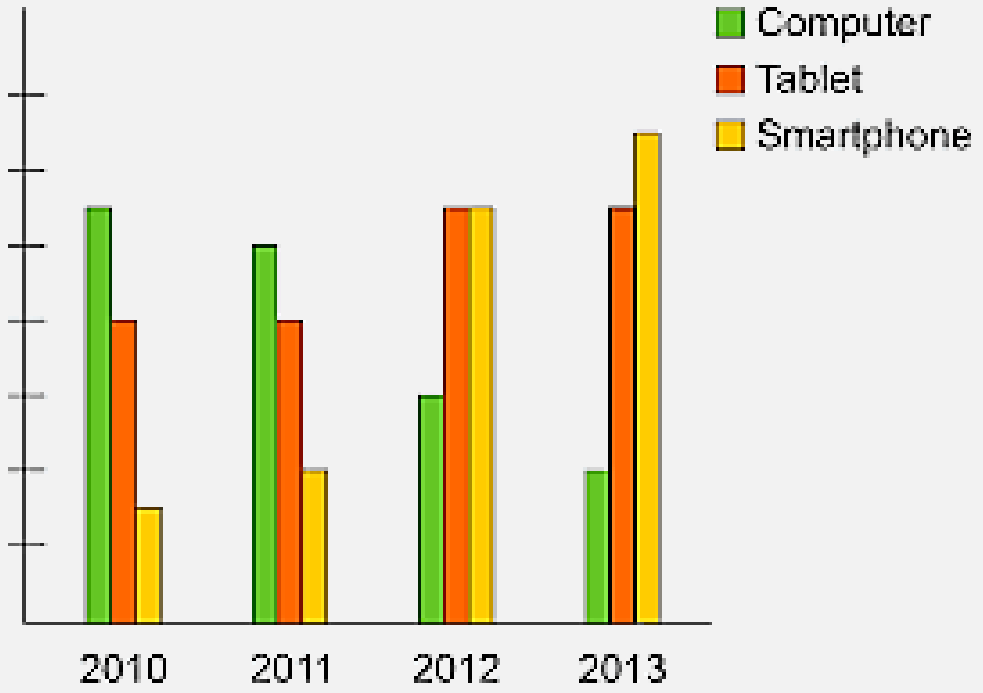
Rappresentazione grafica delle variabili

Istogramma

- Un istogramma descrive la frequenza relativa dei dati compresi in un intervallo (a, b) ed è utilizzato per visualizzare la distribuzione dati.

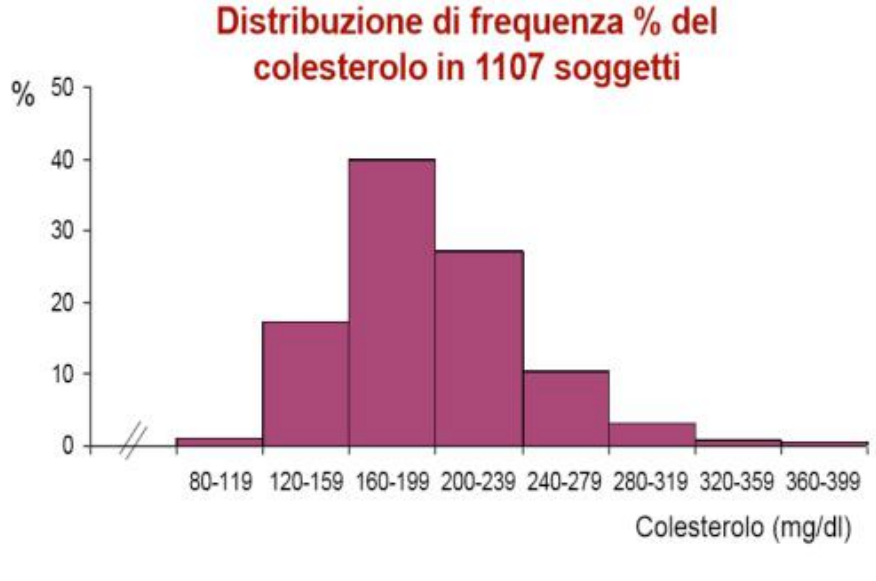


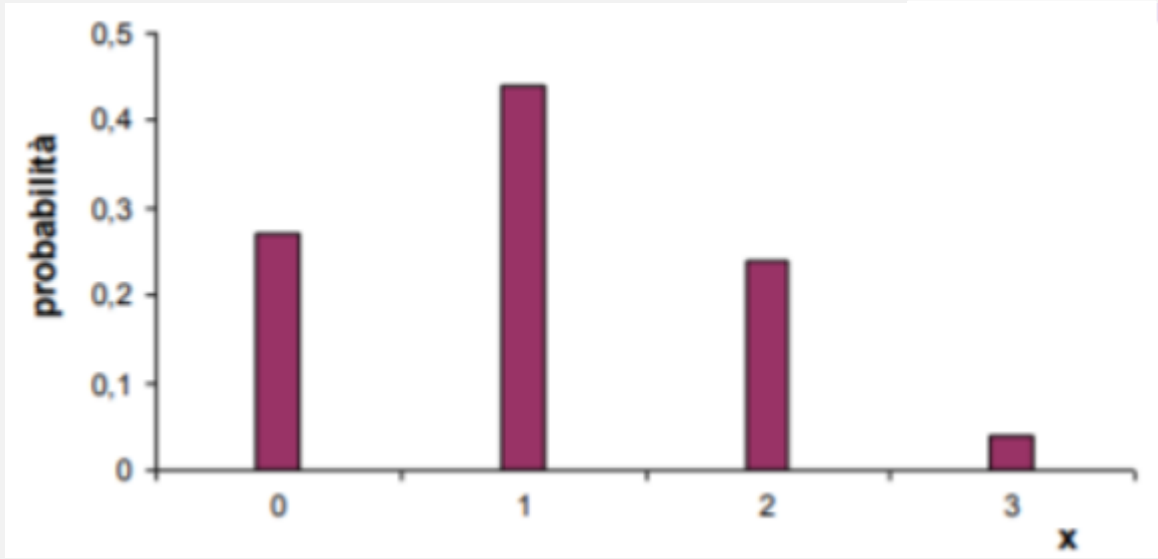




Rappresentazione grafica delle variabili

Istogramma vs. Grafico a barre





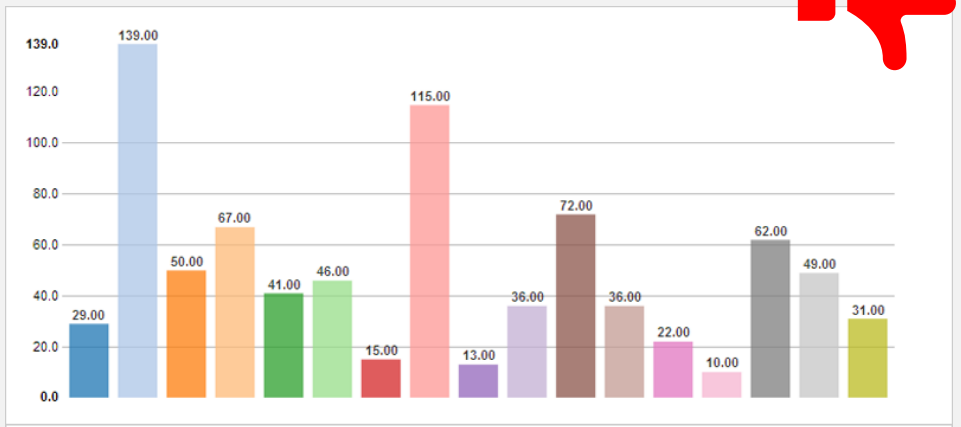
Frequenza = area
Area = ampiezza x altezza
Altezza = frequenza/ampiezza
Altezza = densità di frequenza

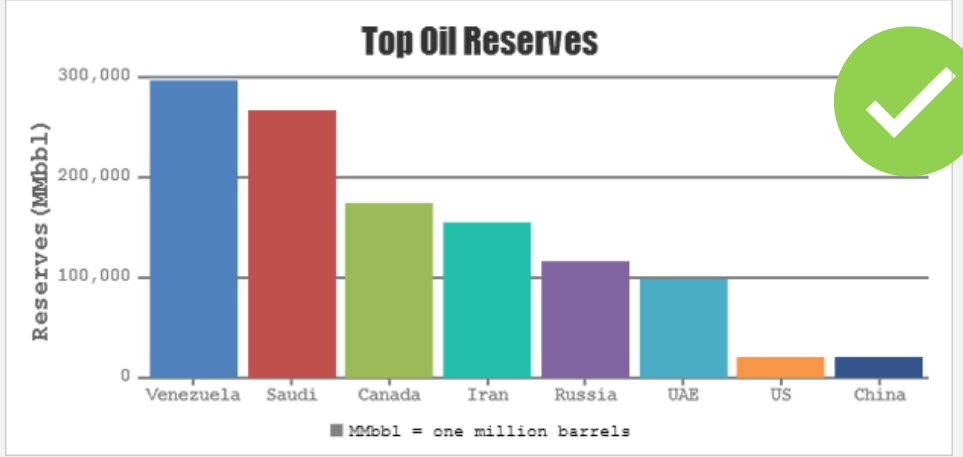
Frequenza = altezza

- ✓ Ciascun rettangolo rappresenta delle possibili categorie (ascissa), in cui l'altezza fa riferimento alla frequenza di ognuna. Si usa soprattutto per le variabili qualitative.

GOOD AND BAD VISUALIZATION

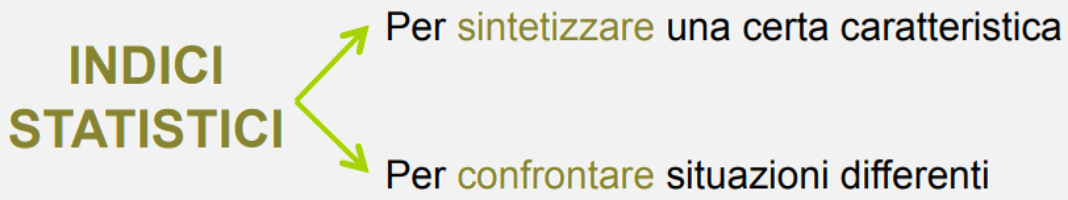
Attenzione alle <u>scale</u> usate per gli assi!!!

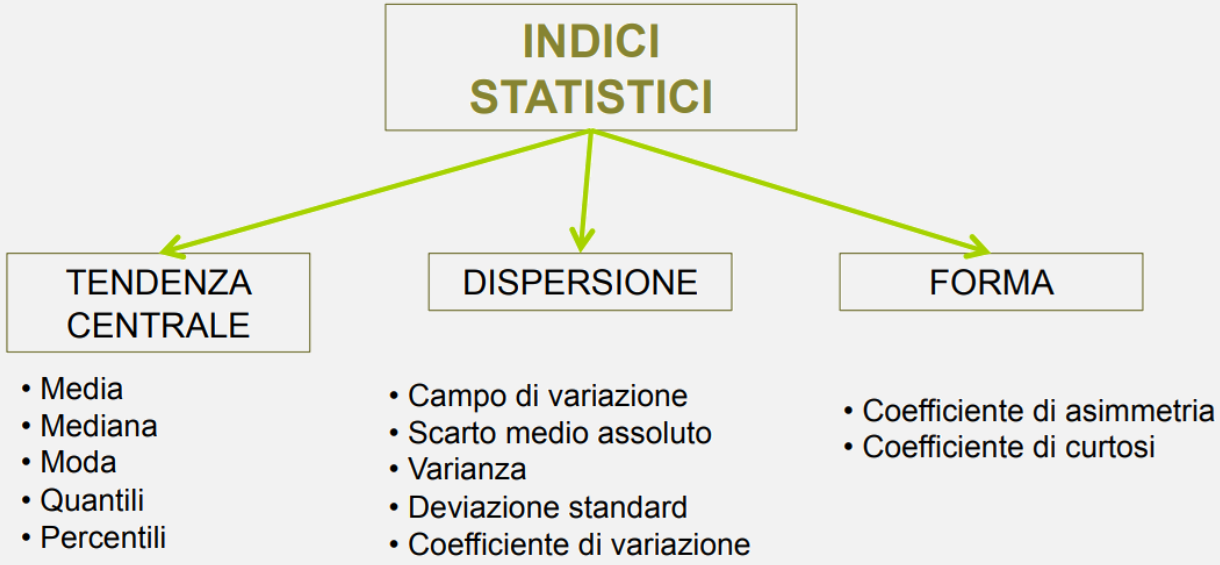




Analisi dei dati

- Quando i dati sono molti, l'analisi diretta della matrice non consente di cogliere in via immediata gli aspetti salienti del fenomeno. Occorre perciò ottenere una sintesi attraverso un'**elaborazione statistica dei dati**





Frequenze assolute e percentuali

- Quando il campione di cui vogliamo descrivere le variabili statistiche è molto grande, anziché considerare tutti i valori, si possono scrivere solo valori distinti e riportare, per ogni valore, quante volte compare.

Sia $Y = (y_1, y_2, \dots, y_N)$ una variabile statistica discreta. Definiamo **modalità** i valori distinti tra $y_1, \dots, y_N$ e **frequenza assoluta** di una modalità il numero di volte che viene osservata nell'espressione della variabile statistica.

$$Y = (60, 80, 92, 100, 83, 84, 96, 74, 63, 80, 100, 90, 75, 74, 92)$$

in termini di modalità e frequenze assolute, attraverso la seguente tabella:

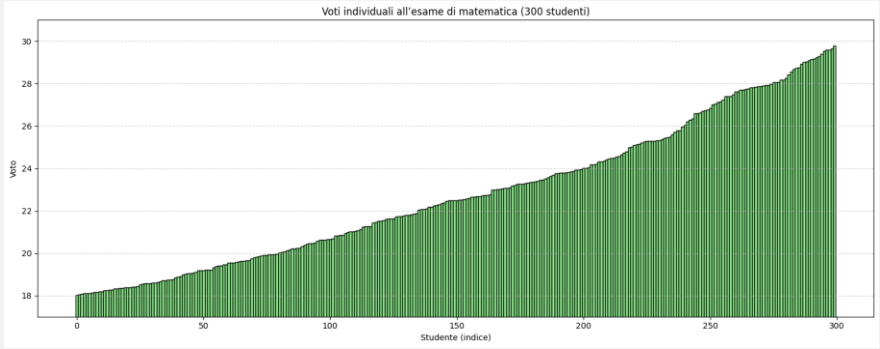
Modalità	60	80	92	100	83	84	96	74	63	90	75
Frequenza assoluta	1	2	2	2	1	1	1	2	1	1	1

Ovviamente la somma di tutte le frequenze assolute dà la numerosità N del campione:

$$N = 1 + 2 + 2 + 2 + 1 + 1 + 1 + 2 + 1 + 1 + 1 = 15$$

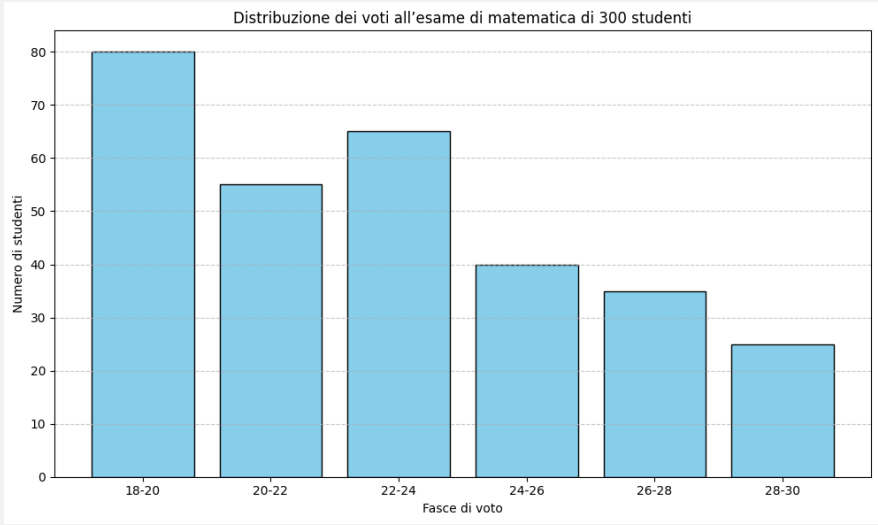
Frequenze assolute e percentuali

Siano Y una variabile statistica e f la frequenza assoluta della modalità z. Definiamo **frequenza relativa** della modalità z il rapporto f/N, ove N è il numero di elementi del campione. La **frequenza percentuale** è data dalla frequenza relativa moltiplicata per 100.



Supponiamo di avere la variabile statistica corrispondente al voto dell'esame di Matematica e Statistica per un campione di 300 studenti. Anziché rappresentare la variabile statistica attraverso i suoi 300 valori (cioè i voti conseguiti dagli studenti), come visto in precedenza, utilizziamo una separazione in classi, dove abbiamo stabilito, arbitrariamente, di considerare intervalli di 2 punti. Esprimiamo dunque i dati nella seguente tabella:

Classi	$[18, 20)$	$[20, 22)$	$[22, 24)$	$[24, 26)$	$[26, 28)$	$[28, 30]$
F. A.	80	55	65	40	35	25
F. R.	$0.2\overline{6}$	$0.18\overline{3}$	$0.21\overline{6}$	$0.1\overline{3}$	$0.11\overline{6}$	$0.08\overline{3}$

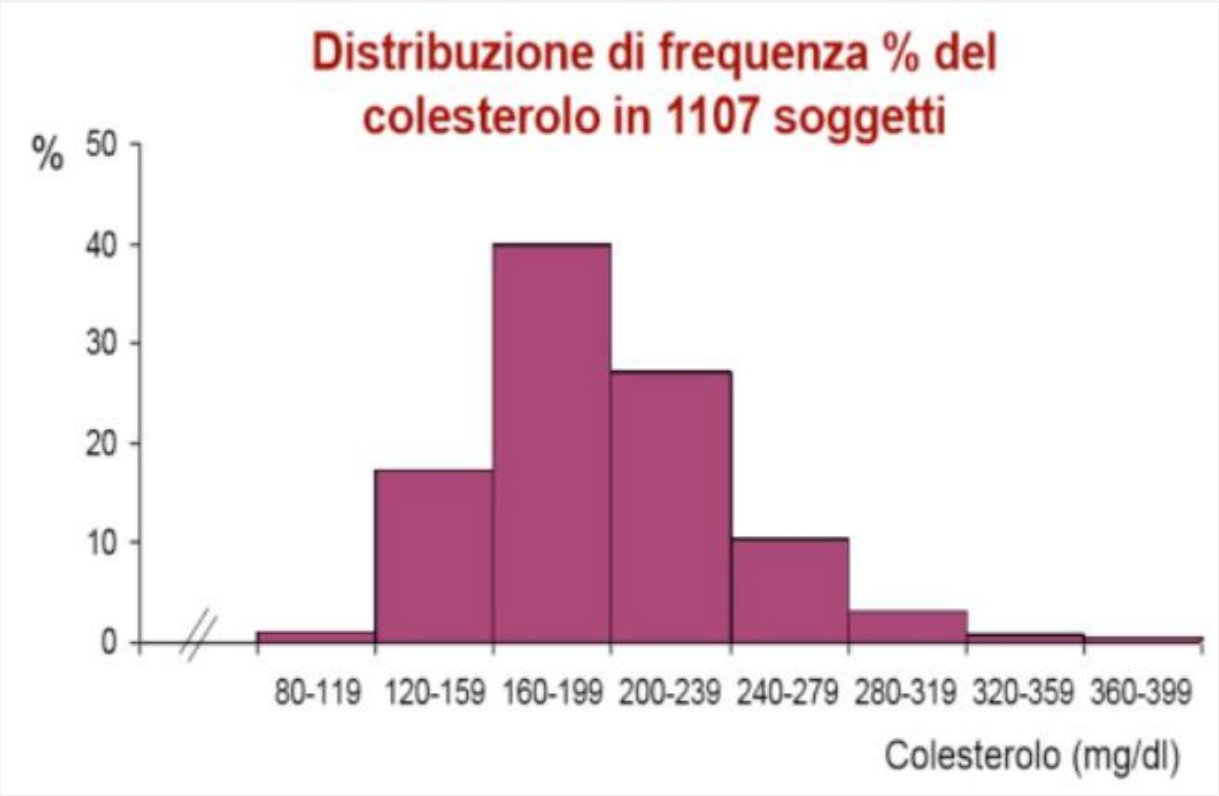


Dalla tabella possiamo leggere subito, per esempio, che 40 studenti hanno ottenuto un voto compreso tra 24 e 26, e rappresentano il 13% circa del campione considerato.

Possiamo anche vedere che 100 studenti, cioè il 33% circa del campione, ha conseguito un voto uguale o superiore a 24/30. Tale dato si può ottenere rapidamente facendo la somma delle frequenze assolute, rispettivamente, relative per le classi individuate dagli intervalli: $[24, 26)$, $[26, 28)$, $[28, 30]$.

Frequenze assolute e percentuali

Colesterolo (mg/dL)	N° soggetti (n)		%
80-119	13	0,012	1,2
120-159	190	0,172	17,2
160-199	442	0,399	39,9
200-239	299	0,270	27,0
240-279	115	0,104	10,4
280-319	34	0,031	3,1
320-359	9	0,008	0,8
360-399	5	0,005	0,5
Totale	**1107**	1,000	100,0
	Frequenza Assoluta	**Frequenza Relativa**	**Frequenza Percentuale**



Misure di tendenza centrale

Informazioni sulla modalità di raggruppamento dei vari valori, in relazione agli individui che formano il campione. Punto centrale della distribuzione.

Media aritmetica

- **Media di una popolazione**: somma di tutti i valori delle variabili della popolazione diviso il numero di unità della popolazione (N)

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Dove N è il numero di elementi della popolazione, X_i è la i-esima osservazione della variabile X_i

- **Media di un campione**: somma di tutti i valori delle variabili di un sottoinsieme della popolazione diviso il numero di unità di tale campione (n)

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Dove n è il numero di elementi del campione della popolazione, X_i è la i-esima osservazione della variabile X_i

Dato il seguente set di misurazioni di livello di espressione dei geni:

55.20	18.06	28.16	44.14	61.61	4.88	180.29	399.11	97.47	56.89	271.95	365.29	807.80

Media della popolazione:

$$\mu = \frac{\sum_{i=1}^{13} 55.20 + 18.06 + 28.16 + 44.14 + 61.61 + \ldots + \ldots + 807.80}{13} = \frac{2390{,}85}{13} = 183.9115$$

Media del campione (55.20; 18.06; 28.16; 44.14):

$$\overline{X} = \frac{55.20 + 18.06 + 28.16 + 44.14}{4} = \frac{145.56}{4} = 36.39$$

La media di qualsiasi campione $\bar{X}$ può essere molto diversa da quella dell'intera popolazione μ . Più è numeroso il campione, più la media del campione sarà vicina a quella della popolazione

Media ponderata

- **Media ponderata di una popolazione**: si assegna ad ogni variabile un peso; si sommano tutti i valori delle variabili, moltiplicate per il peso, e si divide il numero ottenuto per la somma dei pesi

$$\mu = \frac{\sum_{i=1}^{N} p_i X_i}{\sum_{i=1}^{N} p_i}$$

- Il valore atteso di una variabile X, indicato con $E[X]$ è definito come la media di X calcolata su un grande numero di esperimenti

Esempio

Esame	*Voto*	*Crediti (cfu)*
Economia politica	21	7
Ragioneria	25	10
Diritto commerciale	26	6
Matematica	24	5
...	...	...

$$media.p = \frac{(voto \times cfu) + (voto \times cfu) + (...)}{(cfu + cfu + ...)}$$

$$media.p = \frac{(21 \times 7) + (25 \times 10) + (26 \times 6) + (24 \times 5)}{(7 + 10 + 6 + 5)} = \boxed{24.04}$$

Misure di tendenza centrale

Informazioni sulla modalità di raggruppamento dei vari valori, in relazione agli individui che formano il campione. Punto centrale della distribuzione.

Moda

- La **moda** è il valore più frequente di una distribuzione, o meglio, la modalità più ricorrente della variabile (cioè quelle a cui corrisponde la frequenza più elevata).

962	1005	1003	768	980	965	1030	1005	975	989	955	783	1005

La moda di questo campione è 1005, in quanto compare 3 volte.

Caratteristiche della moda:

- viene utilizzata solamente a scopi descrittivi, perché è meno stabile e meno oggettiva delle altre misure di tendenza centrale
- per individuare la moda di una distribuzione si possono usare metodi grafici, come istogrammi
- può differire nella stessa serie di dati, quando si formano classi di distribuzione (intervalli) con ampiezza differente
- per individuare la moda entro una classe di frequenza, non conoscendo come i dati sono distribuiti, si ricorre all'ipotesi della ripartizione uniforme.

Misure di tendenza centrale

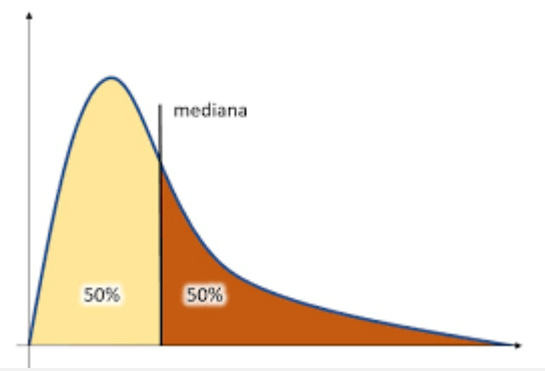
Informazioni sulla modalità di raggruppamento dei vari valori, in relazione agli individui che formano il campione. Punto centrale della distribuzione.

Mediana

- La **mediana** è il valore che occupa la posizione centrale in un insieme ordinato di dati.

- È una misura robusta, in quanto poco influenzata dalla presenza di dati anomali.

- Caratteristiche:
  - si ricorre al suo uso quando si vuole attenuare l'effetto di valori estremi;
  - in una distribuzione o serie di dati, ogni valore estratto a caso ha la stessa probabilità di essere inferiore o superiore alla mediana.



Calcolo della Mediana

Per calcolare la mediana di un gruppo di dati, bisogna:

1. disporre i valori in ordine crescente oppure decrescente e contare il numero totale n di dati;

2. se il numero (n) di dati è **dispari**, la mediana corrisponde al valore numerico del dato centrale, quello che occupa la posizione $(n+1)/2$;

3. se il numero (n) di dati è **pari**, la mediana è stimata utilizzando i due valori centrali che occupano le posizioni $n/2$ e $n/2+1$:
   a. con poche osservazioni, come mediana viene assunta la media aritmetica di queste due osservazioni intermedie;
   b. con molte osservazioni raggruppate in classi, si ricorre talvolta alle proporzioni.

Esempio.
Consideriamo il seguente campione:

96	78	90	62	73	89	92	84	76	86

1. Ordiniamo i campioni in ordine crescente:

62	73	76	78	**84**	**86**	89	90	92	95

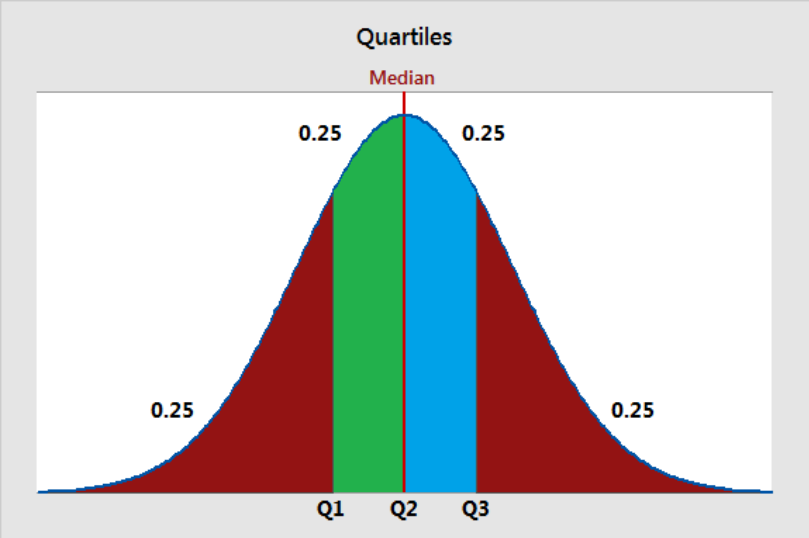
2. Dal momento che il numero di campioni è pari ($n = 10$), la mediana è calcolata come la media dei due elementi centrali:

$$mediana = \frac{84 + 86}{2} = 85$$

Quantili

- I **quantili** sono una famiglia di misure, a cui appartiene anche la mediana, che si distinguono a seconda del numero di parti uguali in cui suddividono una distribuzione.

- I **quartili** ripartiscono la distribuzione in <u>4 parti di pari frequenza</u>, dove ogni parte contiene la stessa frazione di osservazioni:
  - Il **primo quartile** è definito come il numero $q1$ per il quale il 25% dei dati statistici è minore o uguale a $q1$.
  - Il **secondo quartile** è definito come il numero $q2$ per il quale il 50% dei dati statistici è minore o uguale a $q2$. Il secondo quartile corrisponde alla mediana.
  - Il **terzo quartile** è definito come un numero $q3$ per il quale il 75% dei dati statistici è minore o uguale a $q3$.



Misure di posizione (localizzazione)
Si basano sull'ordinamento delle osservazioni da minore a maggiore e sulla successiva divisione della distribuzione ottenuta in gruppi contenenti lo stesso gruppo di osservazioni.

Calcolo dei Quartili

1. Ordiniamo i dati in senso crescente
2. Posizione dei primo quartile (Q1) = $\frac{n+1}{4}$
3. Posizione dei secondo quartile (Q2) = $\frac{n+1}{2}$
4. Posizione dei terzo quartile (Q3) = $\frac{3}{4}(n+1)$
5. Se il risultato è un numero decimale (es. 1,2), bisogna arrotondare sempre per eccesso (es. Considerare la posizione 2).

Esempio.
Consideriamo uno studio che esamina i tempi d'attesa al ristorante in un campione di 10 clienti:

Dati ordinati:

Q2 = Mediana

58.6 59.0 59,3 59,4 62,7 62,8 63,7 65,4 67,3 68,1

$$Q1 = \frac{n+1}{4} = \frac{10+1}{4} = 2.75 \sim 3$$

Q1

58.6 59.0 59,3 59,4 62,7

$$Q2 = \frac{n+1}{2} = \frac{10+1}{2} = 5.5 \sim \textit{media tra posizione } 5 \textit{ e } 6; \frac{62.7+62.8}{2} = 62.75$$

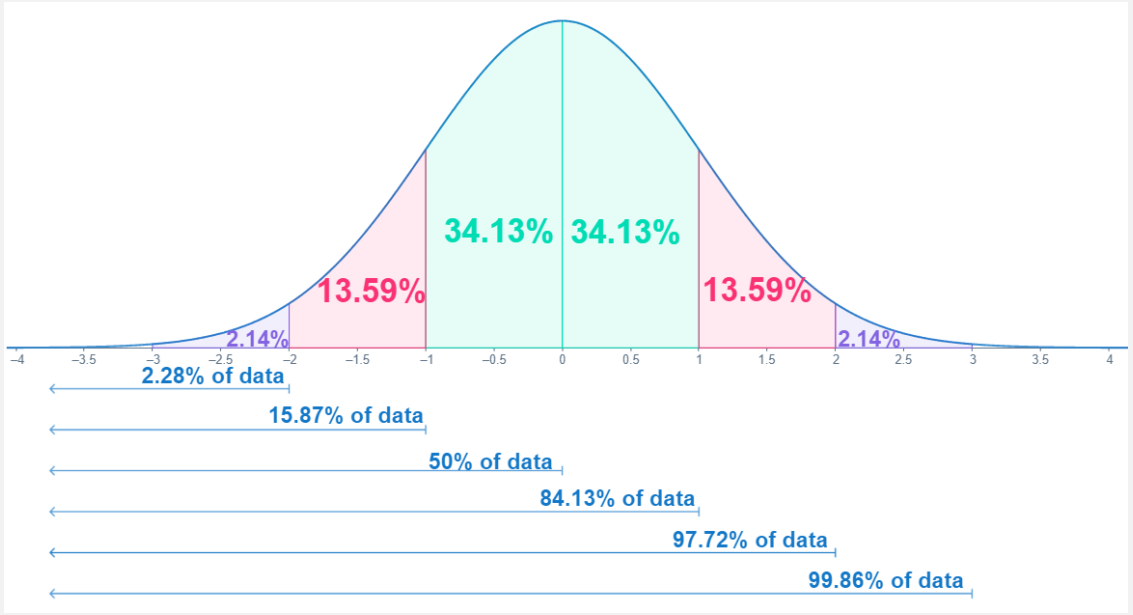
$$Q3 = \frac{3}{4}(n+1) = \frac{3}{4}(10+1) = 8.25 \sim 8$$

62,8 63,7 65,4 67,3 68,1

Q3

Decili e percentili

- In modo analogo, si definiscono:
  - **Decili**: 9 punti che dividono la distribuzione ordinata in 10 parti uguali
  - **Percentili**: 99 punti che dividono la distribuzione ordinata in 100 parti uguali



Misure di dispersione

Forniscono informazioni in merito alla «vicinanza» o «lontananza» delle osservazioni dal centro della distribuzione.

Campo di variazione (o «range», o «intervallo di variabilità»)

- Il **campo di variazione** di una distribuzione è la differenza tra il dato più grande e quello più piccolo della distribuzione:

$$C = x_{max} - x_{min}$$

- Questo indice è abbastanza grossolano non dicendo nulla sulla variabilità dei dati intermedi.

  - Esempio: il campo di variazione della seguente distribuzione:

$$25 - 26 - 28 - 29 - 30 - 32 \rightarrow C = 32 - 25 = 7$$

Misure di dispersione

Forniscono informazioni in merito alla «vicinanza» o «lontananza» delle osservazioni dal centro della distribuzione.

Scarto

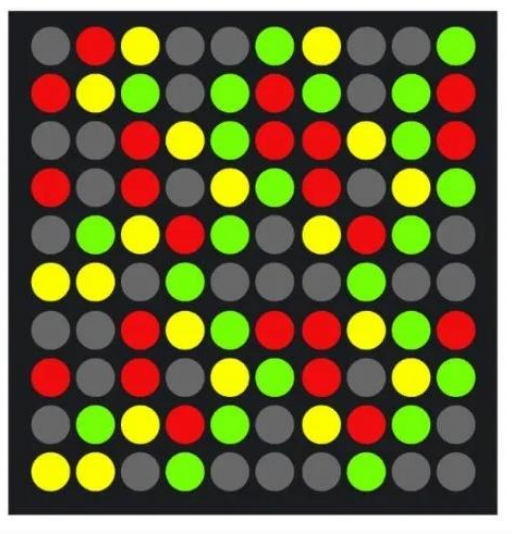
- Lo **scarto** misura quanto ciascun dato x_i si discosta dal valor medio, ovvero $s = x_i - \bar{X}$

  - Esempio: Consideriamo le seguenti intensità rilevate dagli spot dei microarray:

$$435.02, 678.14, 235.35, 956.12, \dots, 1127.82, 456.43$$

  - La media di questi valori è: 515.13; i loro scarti sono:

$$435.02 - 515.13 = -80.11$$
$$678.14 - 515.13 = 163.01$$
$$235.35 - 515.13 = -279.78$$
$$956.12 - 515.13 = 440.99$$



Misure di dispersione

Forniscono informazioni in merito alla «vicinanza» o «lontananza» delle osservazioni dal centro della distribuzione.

Scarto assoluto

Usando s possono essere ricavati diversi altri indici di variabilità

- Si chiama **scarto medio assoluto** e si indica con s_m la media aritmetica dei valori assoluti degli scarti

$$s_m = \frac{\sum_{i=1}^{n} |x_i - \bar{X}|}{n}$$

Devianza

Nel calcolo di alcune statistiche, si ricorre alla devianza, data dal numeratore della varianza. È una misura della dispersione. Si ottiene calcolando la somma dei quadrati delle deviazioni dei dati di una distribuzione rispetto alla media.

$$Dev = \sum_{i=1}^{N} (X_i - \bar{X})^2$$

La varianza è la media della devianza, ovvero indica quanto i valori della distribuzione varia o rispetto alla media

Varianza (o scarto quadratico medio)

- **Varianza della popolazione**: misura che caratterizza molto bene la variabilità di una popolazione. Indicatore della dispersione di una variabile o distribuzione statistica che ottengo calcolando la media dei quadrati degli scarti della media aritmetica (μ).

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

Dove N è il numero di osservazioni dell'intera popolazione; μ è la media della popolazione; x_i è l'i-esimo dato statistico osservato

- **Varianza di un campione**:

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

Dove n è il numero di osservazioni del campione; $\bar{X}$ è la media del campione; x_i è l'i-esimo dato statistico osservato

Quando n è grande, le differenze fra le due formule sono minime (la media riassume bene i dati); quando n è piccolo, le differenze sono sensibili (la media non è rappresentativa dei dati).

Esempio.

Consideriamo il seguente campione di osservazioni:

$$\{2,3,6,9,15\}$$

Calcolo della media:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{2+3+6+9+15}{5} = 7$$

Calcolo della varianza campionaria:

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{(2-7)^2 + (3-7)^2 + (6-7)^2 + (9-7)^2 + (15-7)^2}{4} = 27.5$$

Deviazione standard o scarto quadratico medio

- La varianza ha lo svantaggio di essere una grandezza quadratica e quindi non direttamente confrontabile con la media o con gli altri valori della distribuzione.

- Per trovare una misura espressa nella stessa unità di misura della variabile di partenza è sufficiente estrarre la radice quadrata della varianza.

- La **deviazione standard** è una misura di distanza dalla media e quindi ha sempre un valore positivo.

- È una misura della dispersione della variabile casuale intorno alla media.

Deviazione standard o scarto quadratico medio

- **Deviazione standard della popolazione**:

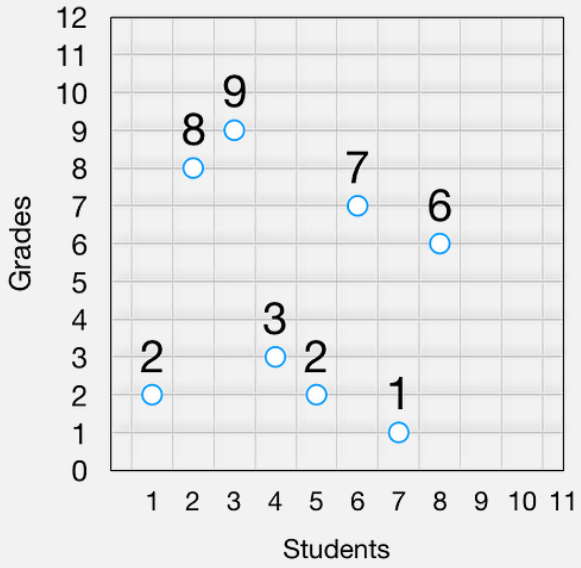
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

Dove N è il numero di osservazioni dell'intera popolazione; μ è la media della popolazione; X_i è l'i-esimo dato statistico osservato.

- **Deviazione standard di un campione**:

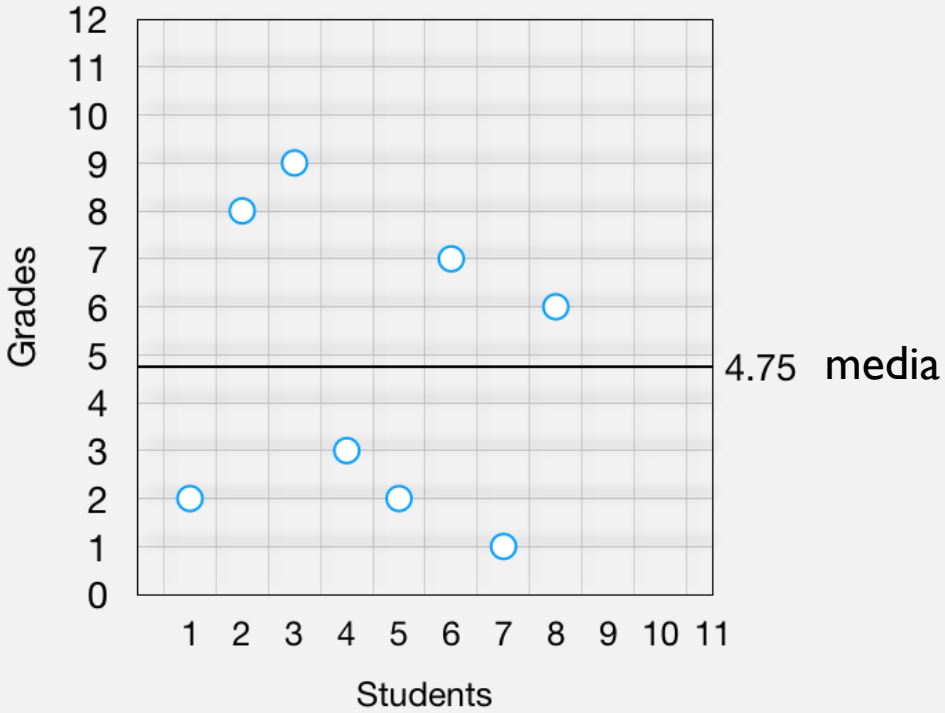
$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

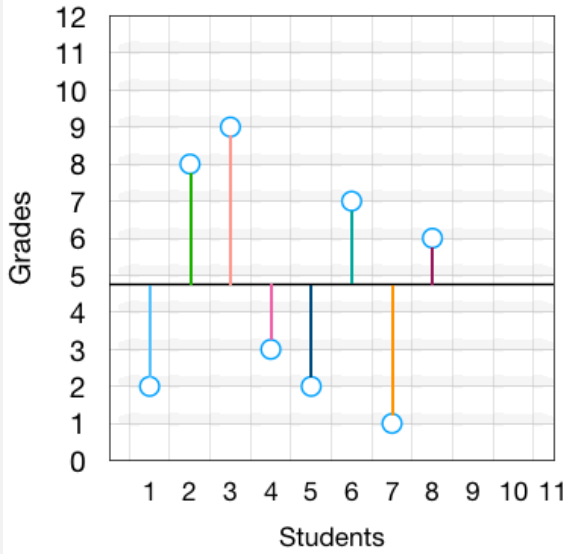
Dove n è il numero di osservazioni del campione; $\bar{X}$ è la media del campione; x_i è l'i-esimo dato statistico osservato



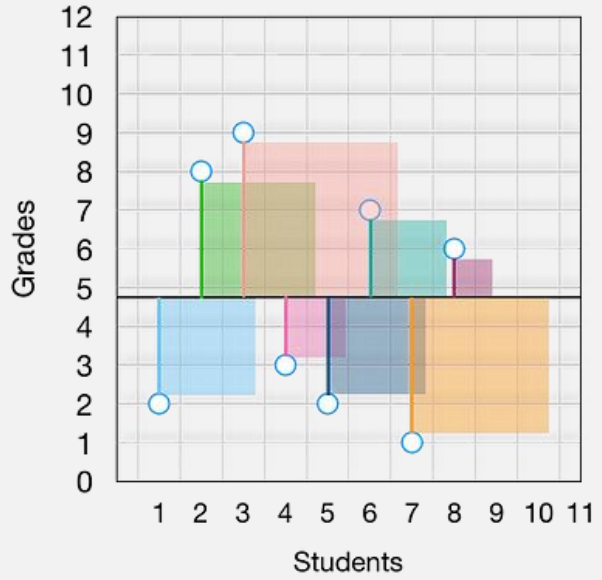
Student	Grade
1	2
2	8
3	9
4	3
5	2
6	7
7	1
8	6

$$\bar{x} = \frac{\sum_{n=1}^{N} x_n}{N} = \frac{2+8+9+3+2+7+1+6}{8} = \frac{38}{8} = 4.75$$





$$x - \bar{x} =$$
$$(2\text{-}4.75) + (8\text{-}4.75) + (9\text{-}4.75) + (3\text{-}4.75) + (2\text{-}4.75) + (7\text{-}4.75) + (1\text{-}4.75) + (6\text{-}4.75)$$



$$\sum (x_n - \bar{x})^2 =$$

$$7.5625 + 10.5625 + 18.0625 + 3.0625 + 7.5625 + 5.0625 + 14.0625 + 1.5625 = 67.5$$

Varianza

$$\frac{\sum (x_n - \bar{x})^2}{N} = \frac{67.5}{8} = 8.45 \text{ points}^2$$

Deviazione standard

$$\sqrt{\frac{\sum (x_n - \bar{x})^2}{N}}$$

Mean ( ) = 8.45

$$\sqrt{\quad 8.45 \quad} \; 2.91 = \; 2.91$$

Esempio.
Consideriamo i voti di due studenti:
- Anna: $30, 30, 28, 27, 26$
- Stefano: $21, 30, 30, 30, 30$

Entrambi hanno la stessa media dei voti (28.2)

Calcoliamo la deviazione standard:
- $\sigma_{Anna} = 1.78$
- $\sigma_{Stefano} = 4.02$

Cosa significa? I voti di Anna sono più concentrati (vicini) rispetto a quelli di Stefano

Tanto più piccola è la deviazione standard rispetto alla media, tanto più i dati sono concentrati attorno alla media (cioè tanto meglio la media riassume i dati).

Distribuzioni unimodali/bimodali

- Una distribuzione può presentare più mode:
  - Distribuzioni **unimodali**: distribuzioni di frequenza che hanno una sola moda, ossia un solo un punto di massimo (che rappresenta sia il massimo relativo che il massimo assoluto)
  - Distribuzioni **bimodali o k-modali**: distribuzioni di frequenza che presentano due o più mode, ossia che hanno due (o k) massimi relativi
    - Esempio: misurando le altezze di un gruppo di giovani in cui la parte maggiore sia formata da femmine e la minore da maschi si ottiene una distribuzione bimodale, con una moda principale ed una secondaria.

Distribuzione zeromodale

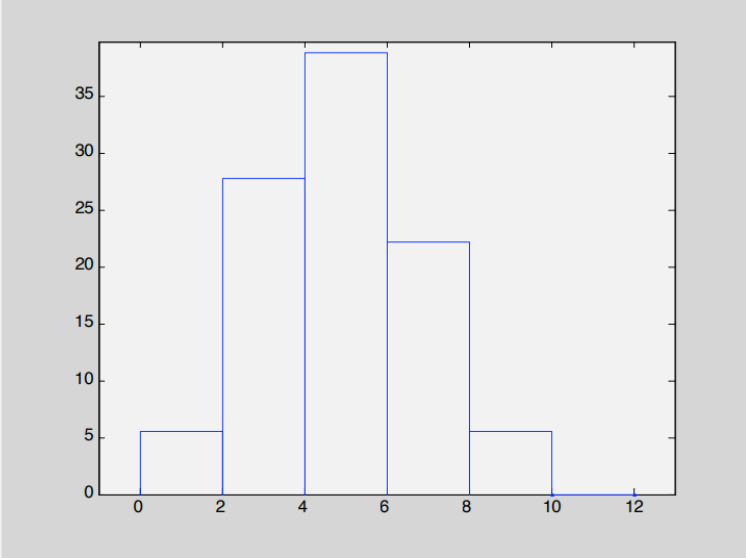
- Nessun valore ha una frequenza più elevata degli altri:

$$A = \{1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6\}$$

Distribuzione unimodale

C’è un solo valore con una frequenza più elevata degli altri

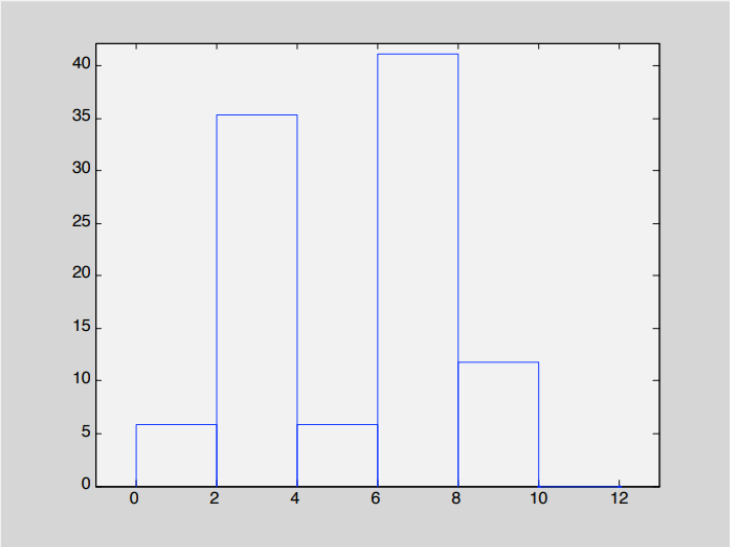
A = {1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7, 7, 8}

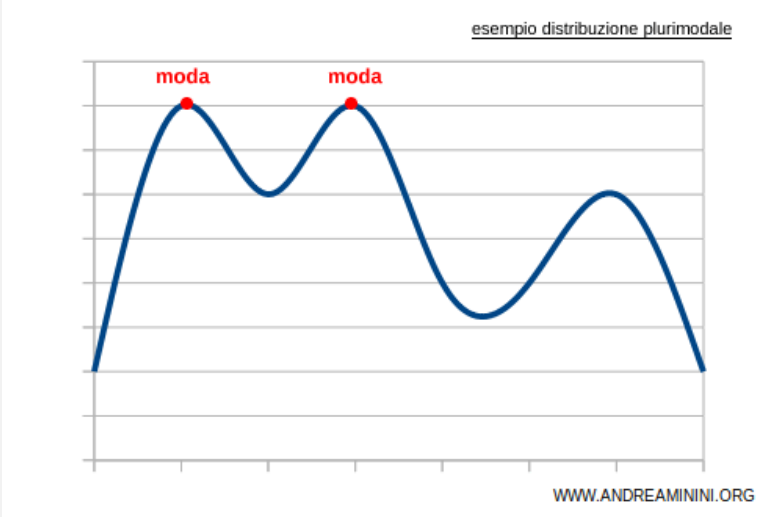


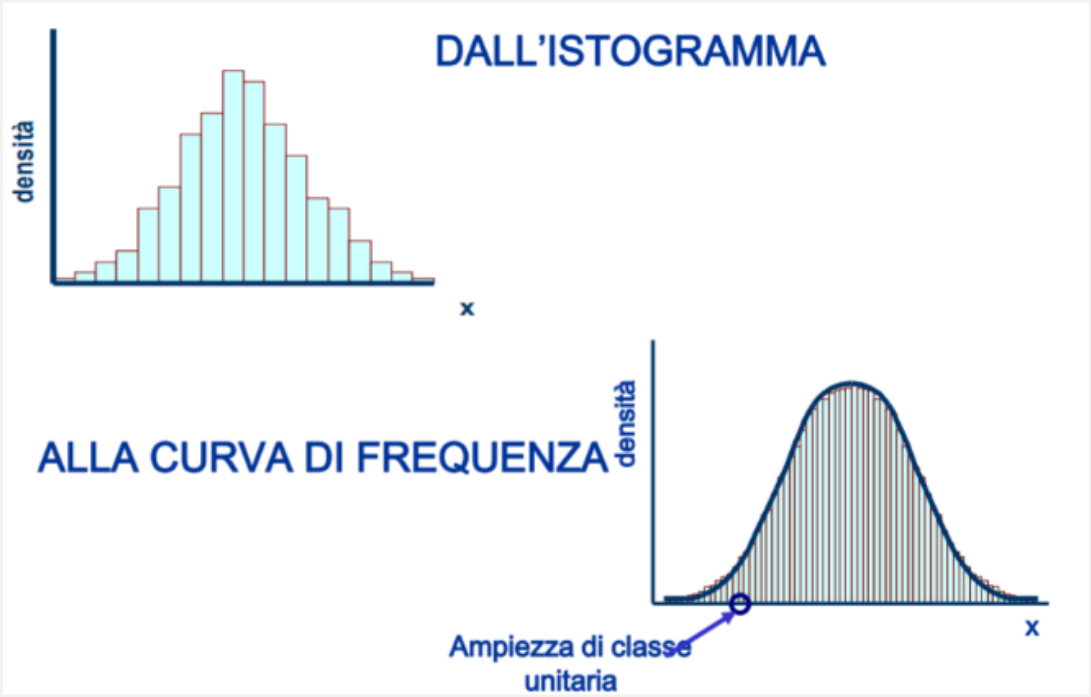
Distribuzione bimodale

Ci sono due valori con una frequenza più elevata degli altri.

A = {1, 2, 2, 3, 3, 3, 3, 5, 6, 6, 6, 6, 6, 7, 7, 8, 8}

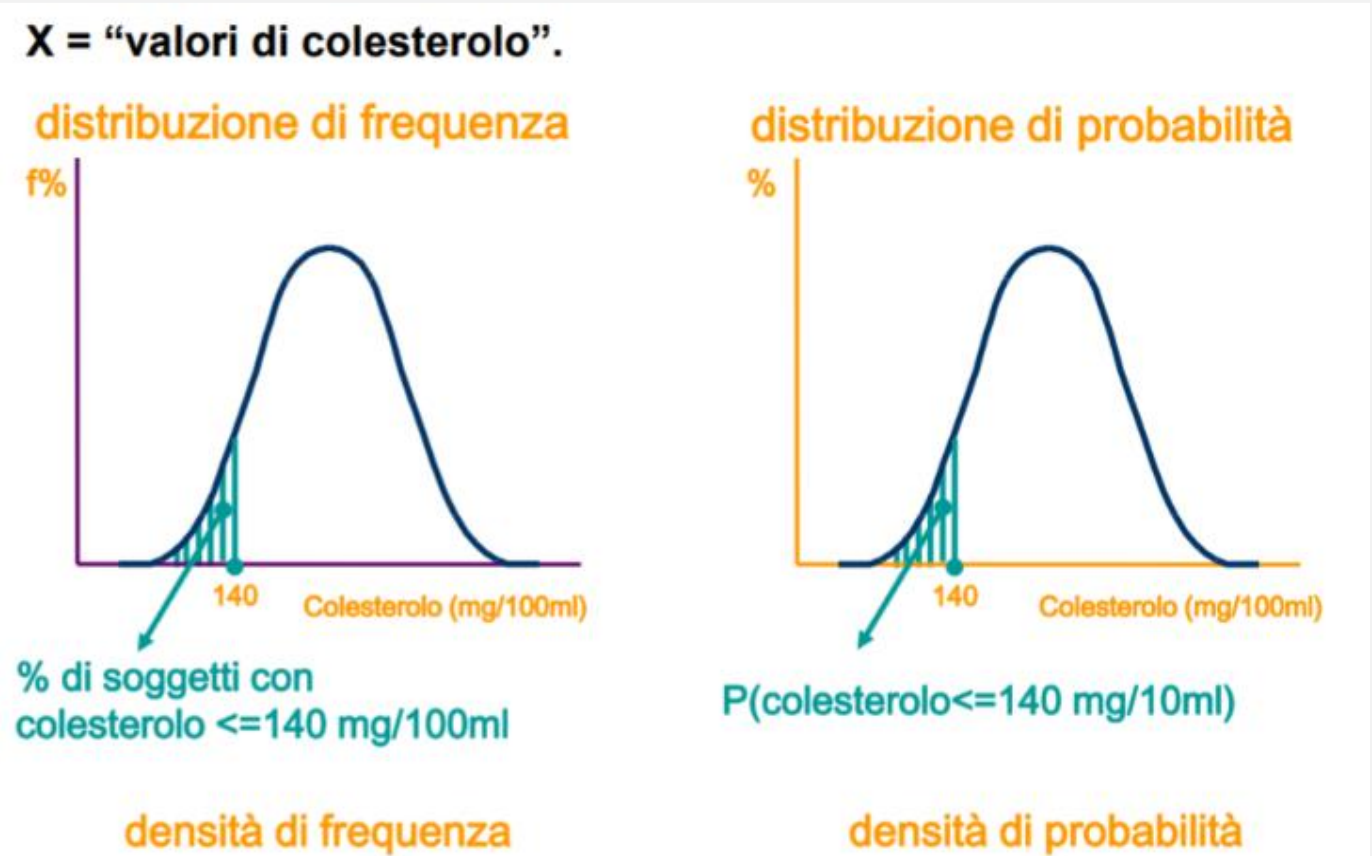






L'istogramma rappresenta la distribuzione di frequenza di una variabile riportando sull'asse delle ascisse i valori della variabile e sull'asse delle ordinate la densità di frequenza.

Proviamo a ridurre l'ampiezza di classe fino a farla diventare unitaria = <u>Curva di Frequenza</u>

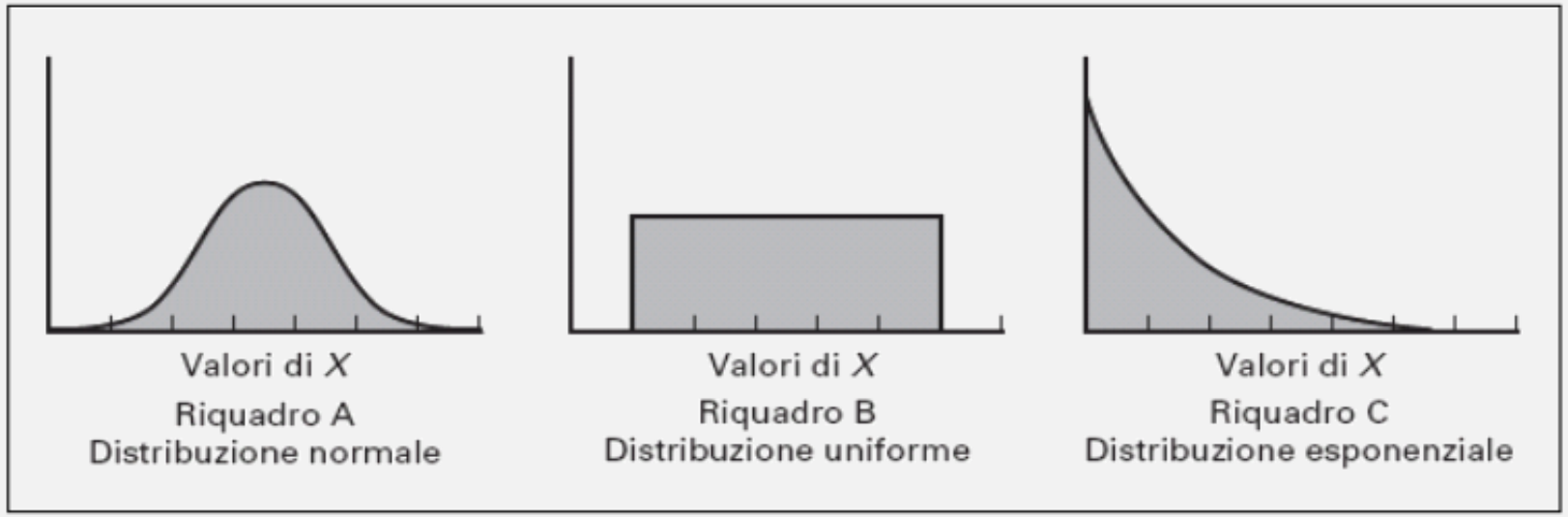


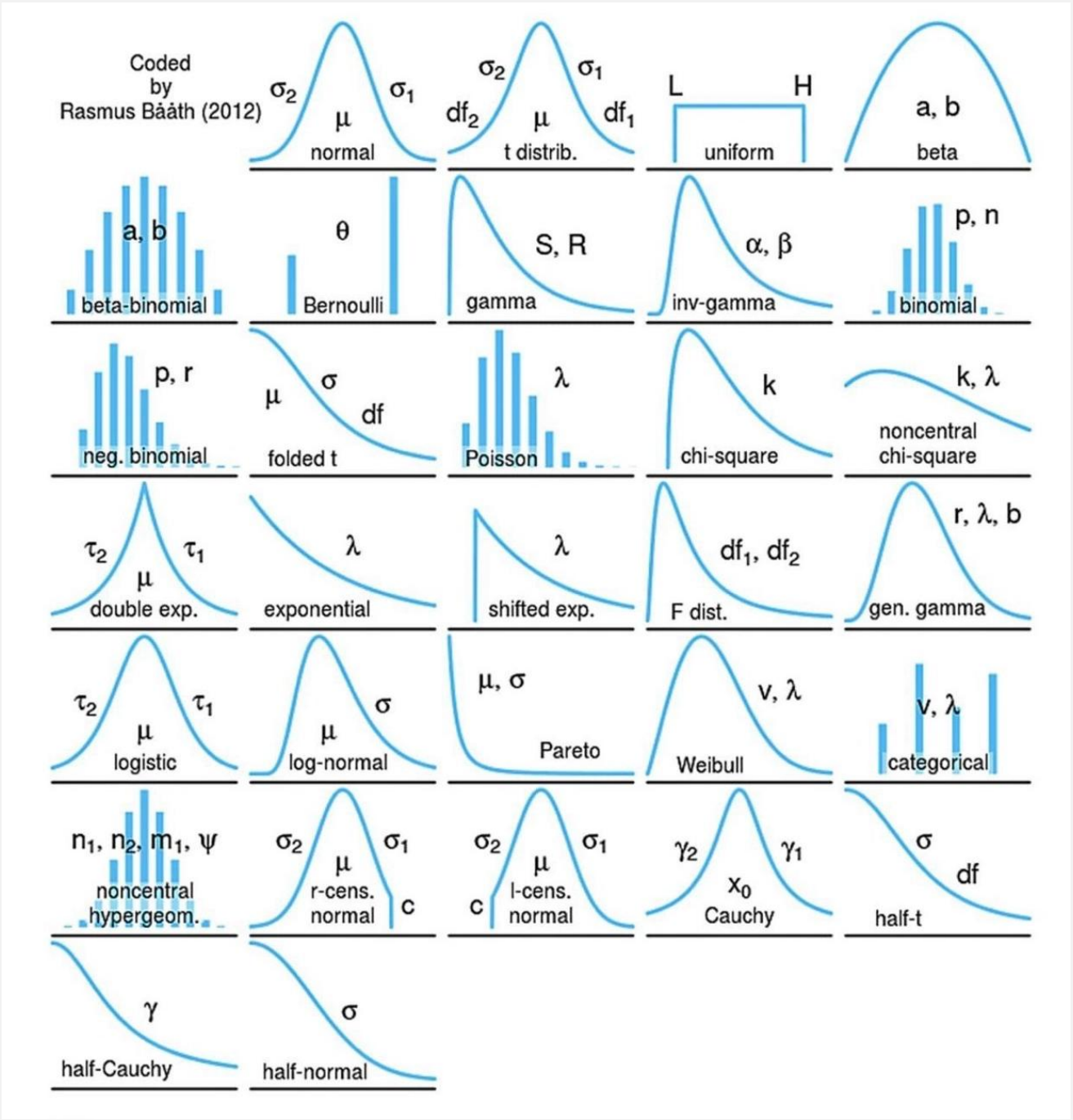
Distribuzioni di probabilità continue

- Una funzione di densità di probabilità continua è un modello che definisce analiticamente come si distribuiscono i valori assunti da una variabile aleatoria continua

- Quando si dispone di un'espressione matematica adatta alla rappresentazione di un fenomeno continuo, siamo in grado di calcolare la probabilità che la variabile aleatoria assuma valori compresi in intervalli

- La figura rappresenta graficamente tre funzioni di densità di probabilità: normale, uniforme ed esponenziale

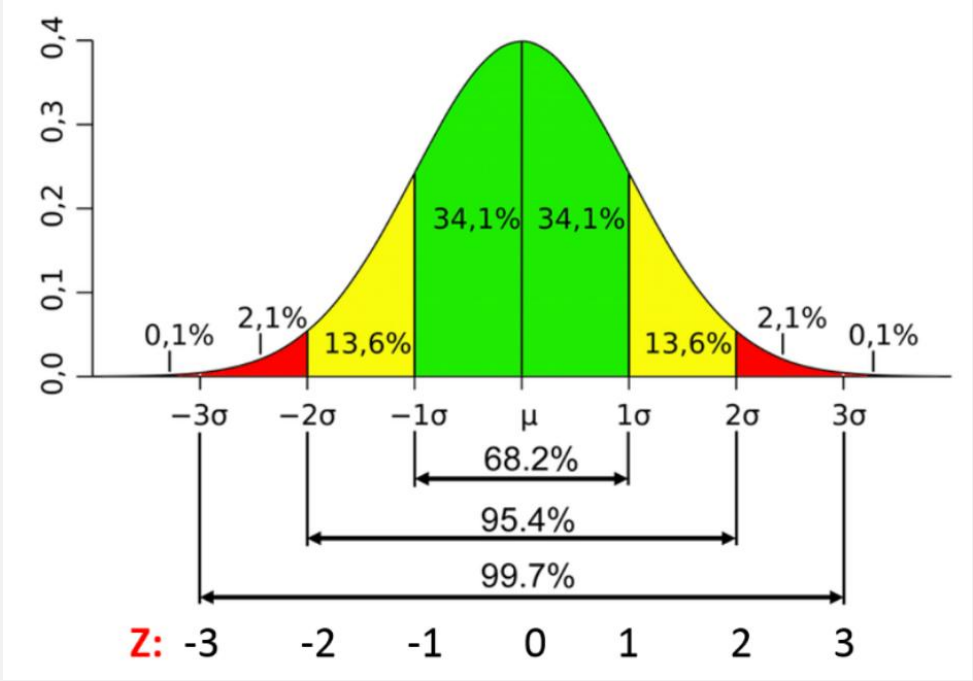


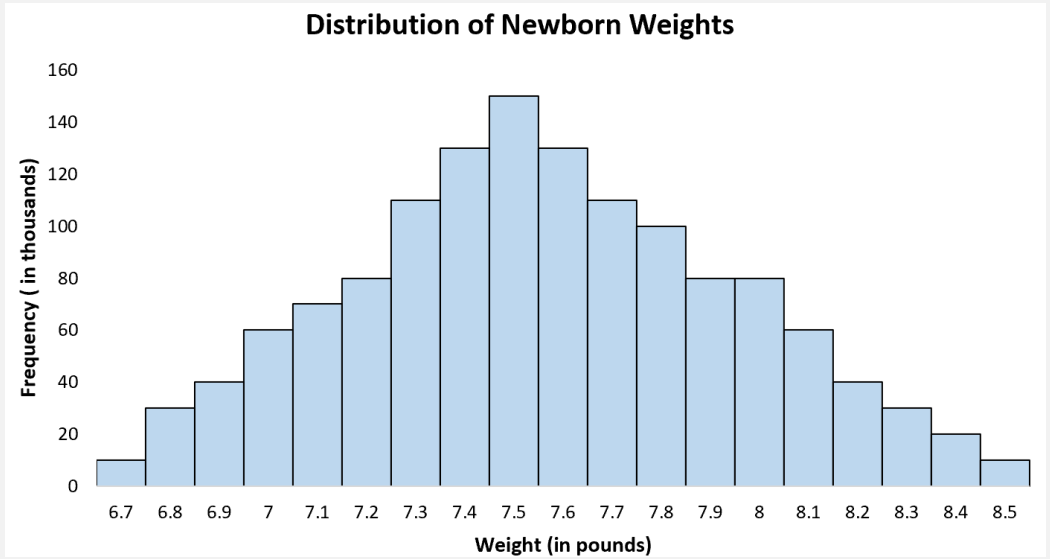

Coded
by
Rasmus Bååth (2012)
σ2
σ1
μ
normal
σ2
σ1
df2
μ
df1
t distrib.
L
H
uniform
a, b
beta
a, b
beta-binomial
θ
Bernoulli
S, R
gamma
α, β
inv-gamma
p, n
binomial
p, r
neg. binomial
μ
σ
df
folded t
λ
Poisson
k
chi-square
k, λ
noncentral
chi-square
τ2
τ1
μ
double exp.
λ
exponential
λ
shifted exp.
df1, df2
F dist.
r, λ, b
gen. gamma
τ2
τ1
μ
logistic
μ
σ
log-normal
μ, σ
Pareto
v, λ
Weibull
v, λ
categorical
n1, n2, m1, ψ
noncentral
hypergeom.
σ2
σ1
μ
r-cens.
normal
c
σ2
σ1
μ
c
l-cens.
normal
γ2
γ1
x0
Cauchy
σ
df
half-t
γ
half-Cauchy
σ
half-normal

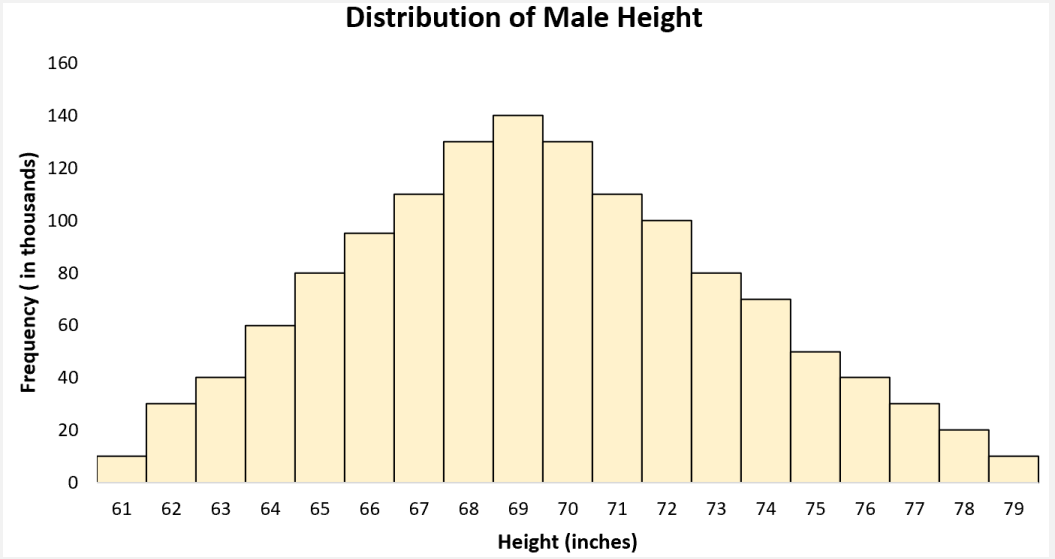
Distribuzione normale (o Gaussiana)

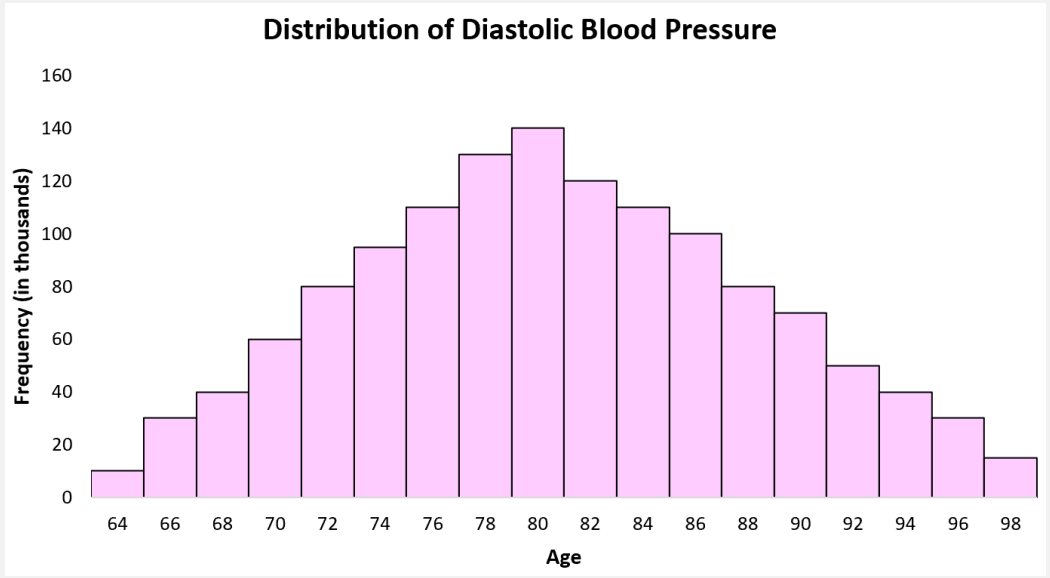
- La distribuzione normale (o distribuzione Gaussiana) è la <u>distribuzione continua più utilizzata in statistica</u> e che permette di descrivere molti fenomeni biologici.

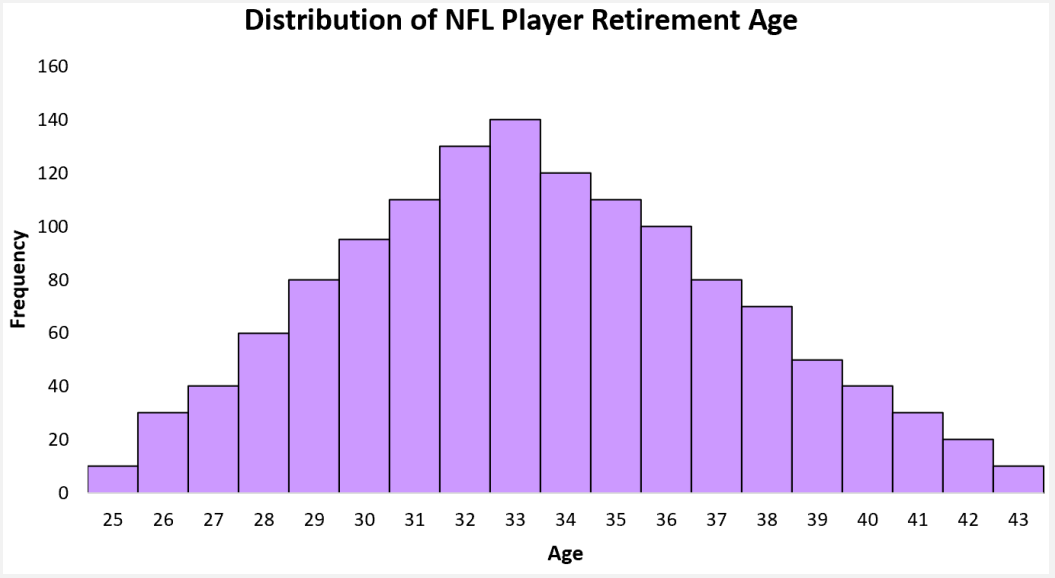
- La distribuzione normale è importante in statistica per tre motivi fondamentali:
  1. Diversi fenomeni continui sembrano seguire, almeno approssimativamente, una distribuzione normale.
  2. La distribuzione normale può essere utilizzata per approssimare numerose distribuzioni di probabilità discrete.




Distribution of Newborn Weights
Frequency (in thousands)
0
20
40
60
80
100
120
140
160
6.7
6.8
6.9
7
7.1
7.2
7.3
7.4
7.5
7.6
7.7
7.8
7.9
8
8.1
8.2
8.3
8.4
8.5
Weight (in pounds)


Distribution of Male Height
Frequency (in thousands)
0
20
40
60
80
100
120
140
160
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
Height (inches)


Distribution of Diastolic Blood Pressure
Frequency (in thousands)
0
20
40
60
80
100
120
140
160
64
66
68
70
72
74
76
78
80
82
84
86
88
90
92
94
96
98
Age

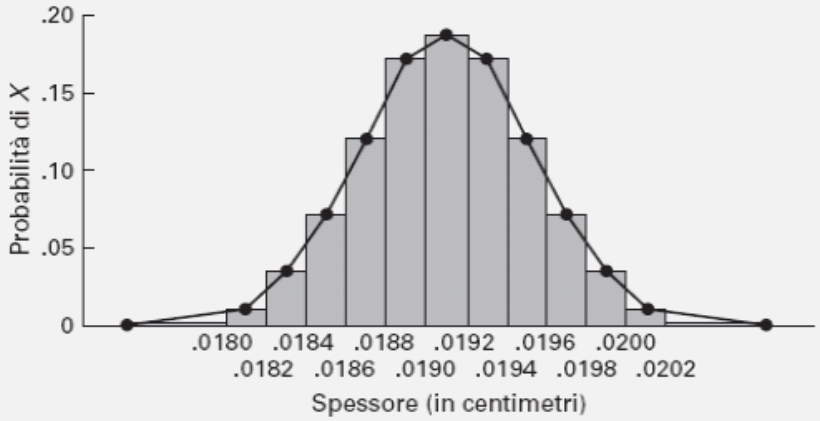

Distribution of NFL Player Retirement Age
Frequency
0
20
40
60
80
100
120
140
160
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
Age

Distribuzione normale

La distribuzione normale ha alcune importanti caratteristiche:

- La distribuzione normale ha una forma campanulare e simmetrica
- Le sue misure di posizione centrale (valore atteso, mediana) coincidono
- Il suo range interquartile è pari a 1.33 volte lo scarto quadratico medio, cioè copre un intervallo compreso tra $\mu - 2/3\,\sigma$ e $\mu + 2/3\,\sigma$
- La variabile aleatoria con distribuzione normale assume valori compresi tra $-\infty$ e $+\infty$

Consideriamo lo spessore misurato in centimetri di 10000 rondelle di ottone prodotte da una grande società metallurgica. Il fenomeno aleatorio continuo di interesse, lo spessore delle rondelle, si distribuisce approssimativamente come una normale



Distribuzione normale

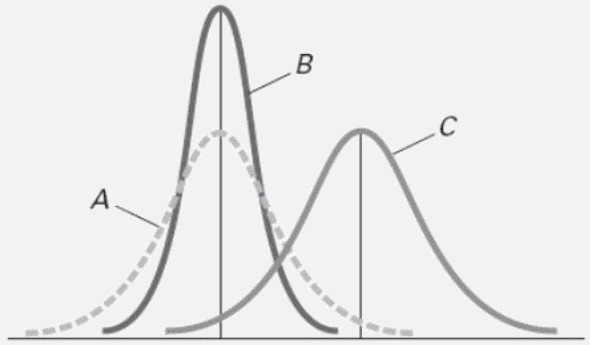
Utilizzeremo il simbolo $f(X)$ per denotare l'espressione matematica di una funzione di densità di probabilità. Nel caso della distribuzione normale la funzione di densità di probabilità normale è data dalla seguente espressione:

Funzione di densità di probabilità normale

$$f(X) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\left(\frac{1}{2}\right)\left[\frac{X-\mu}{\sigma}\right]^2}$$

Dove μ è il valore atteso della popolazione; σ è lo scarto quadratico medio della popolazione; X rappresenta i valori assunti dalla variabile aleatoria, $-\infty < X < +\infty$

Notiamo che, essendo e e π delle costanti matematiche, le probabilità di una distribuzione normale dipendono soltanto dai valori assunti dai due parametri μ e σ. Specificando particolari combinazioni di μ e σ, otteniamo differenti distribuzioni di probabilità normali.

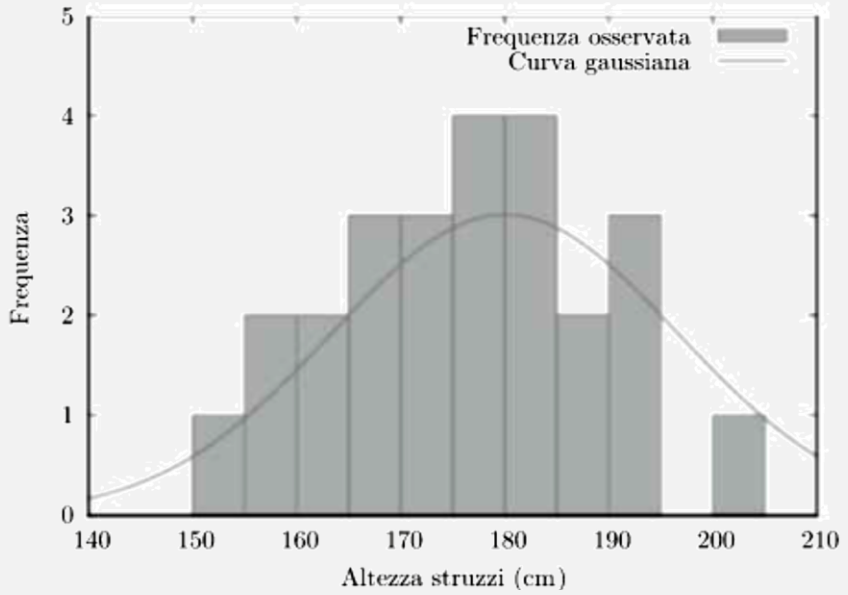


Supponiamo di avere una tabella che riporta i dati dell'altezza in cm (arrotondata) degli struzzi in un allevamento in termini di modalità e frequenze assolute:

Altezza	205	200	195	190	185	180	175	170	165	160	155
F. A.	1	0	3	2	4	4	3	3	2	1	1

Calcoliamo sia varianza sia valor medio ottenendo:

$$\sigma^2 = 16.58, \qquad \mu = 180$$



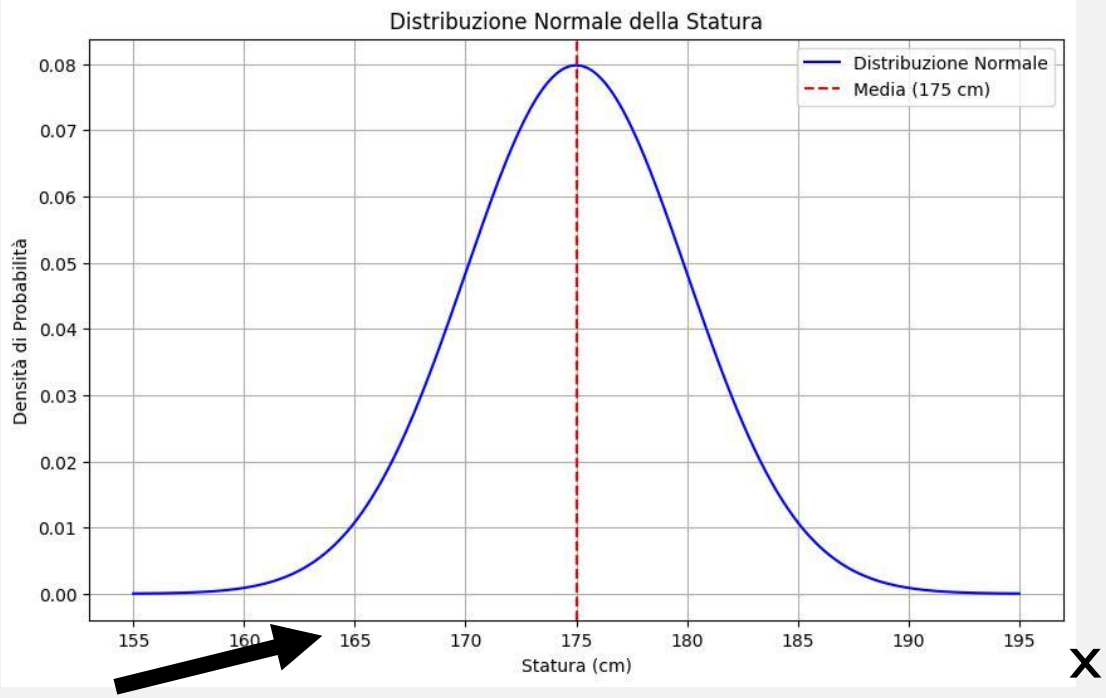
Distribuzione normale standardizzata

La standardizzazione è una trasformazione dei dati che consiste nel:

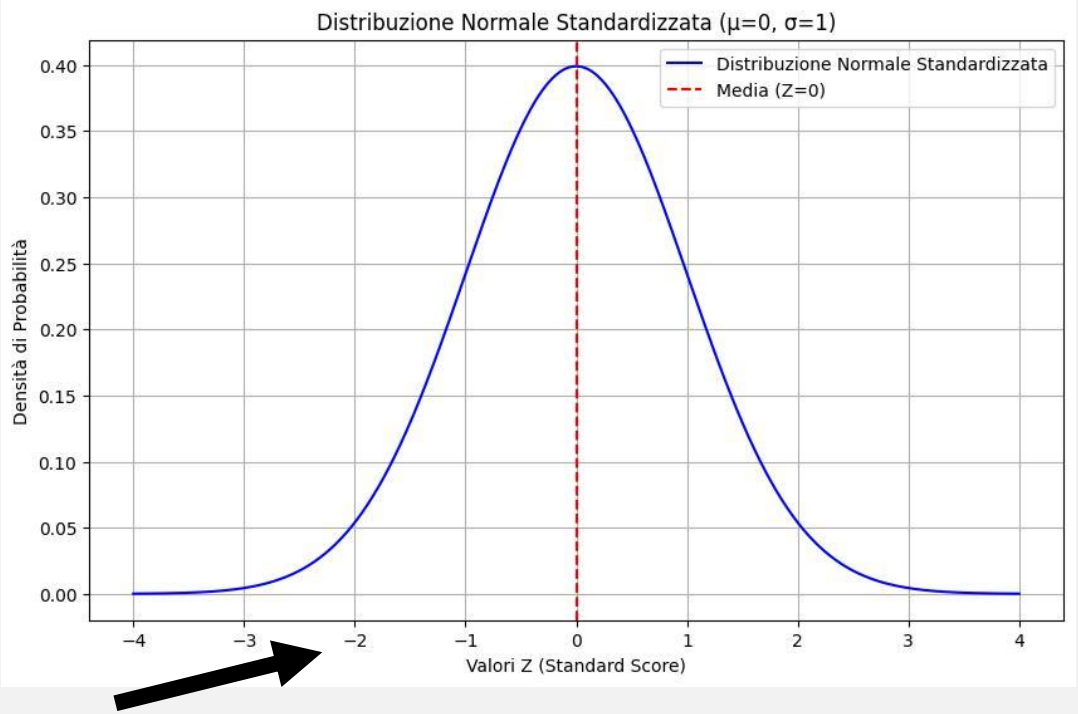
- ✓ Rendere la media nulla (μ = 0): a ogni valore della variabile originaria viene sottratta la media della variabile stessa;
- ✓ Rendere la varianza = 1.
- ✓ La funzione della curva normale standardizzata è la seguente ($-\infty < z < +\infty$):

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

- ✓ La variabile indipendente non sarà più *x*, ma *z* (depurata dell'unità di misura)così definita: $\longrightarrow Z = \frac{X - \mu_x}{\sigma_x}$
- ✓ Permette di individuare facilmente le probabilità relative agli intervalli di valori, utilizzando opportune *tavole statistiche*. Le probabilità corrispondono alle aree sottese dalla curva normale ed possono essere facilmente calcolate.



x



μ_x = 175 cm

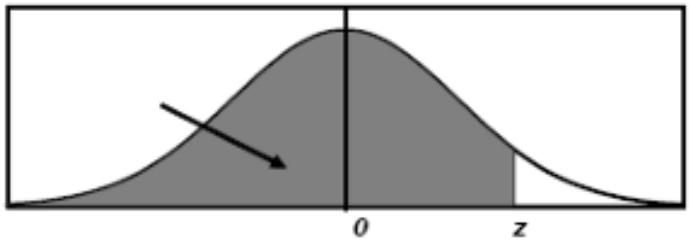
σ_x= 5 cm

Se considero la statura un oggetto pari a 165 cm, posso convertire questo dato in un valore z.

$$Z = \frac{X - \mu_x}{\sigma_x} = \frac{165 - 175}{5} = -2$$

Il valore standardizzato (Z) pari a -2 indica che 165 cm è inferiore alla media di 2 volte la deviazione standard.

Normale Standardizzata



$$-\infty < \Pr(Z) \le z = \int_{-\infty}^{z} f(t)dt$$

(a) Integrali della variabile casuale normale standardizzata

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,0	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361

Se un fenomeno si distribuisce secondo una curva normale, l'area sotto la curva della normale standardizzata è pari a 1, cioè al 100% delle osservazioni.

L'area sotto la curva fa riferimento alla percentuale delle osservazioni tra due valori di x.

Tabelle con valori di z < 0, >0 o valori compresi tra 0 e un certo valore di z.

Le tabelle indicano la **Funzione di Ripartizione,** ovvero l'area da - ∞ a z.

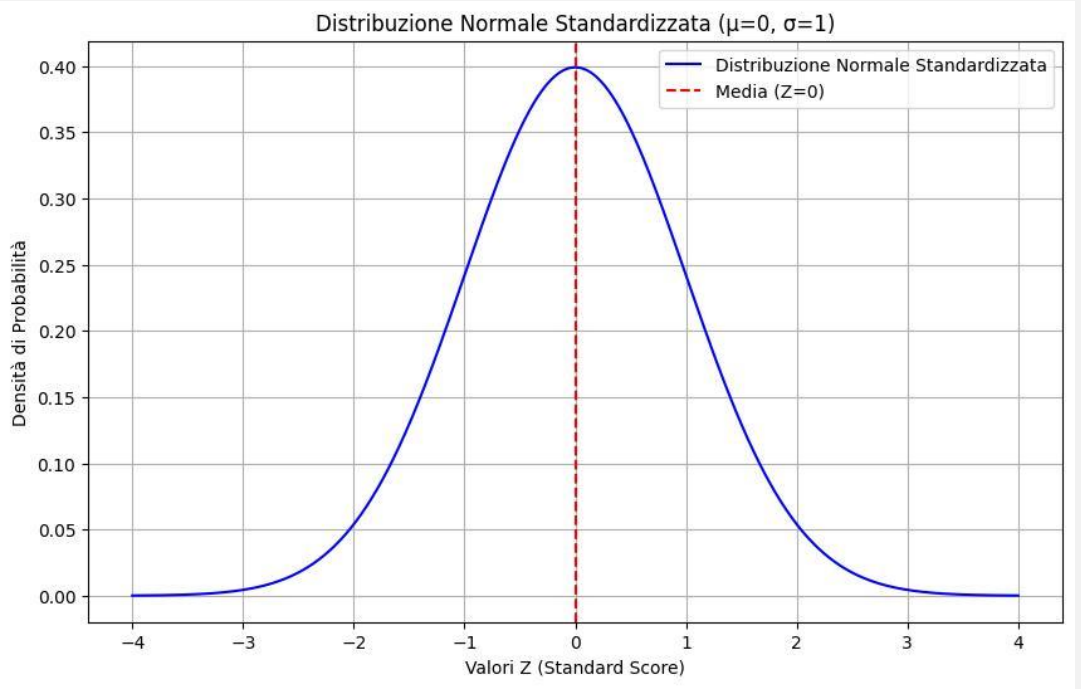
Distribuzione normale standardizzata

$$\int_{x1}^{x2} f(x)dx = \int_{z1}^{z2} f(z)dz$$

$$\int_{-\infty}^{0} f(z)dz = \int_{0}^{+\infty} f(z)dz$$

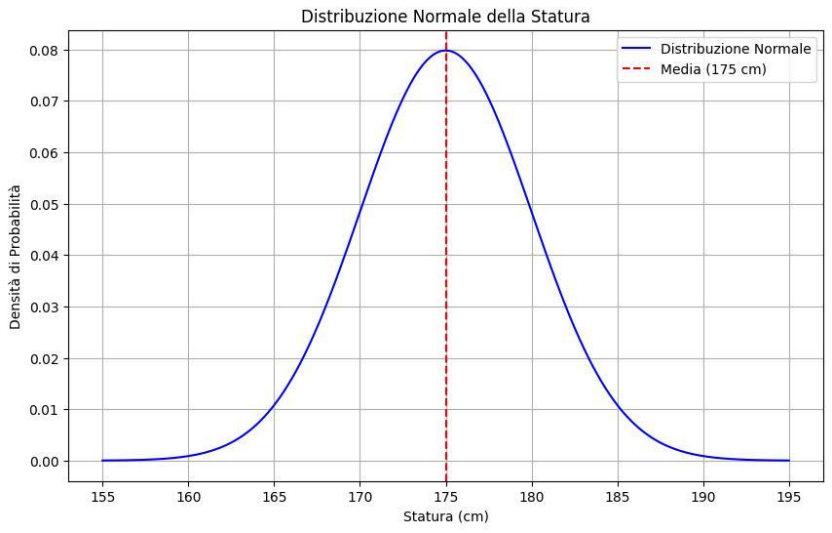
$$\int_{1}^{+\infty} f(z)dz = \int_{-\infty}^{-1} f(z)dz$$

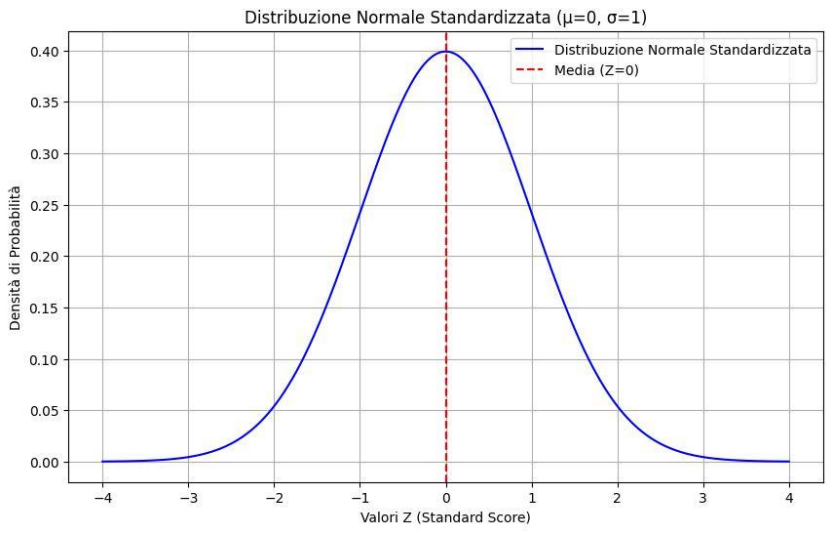
$$\int_{1}^{2} f(z)dz = \int_{0}^{2} f(z)dz - \int_{0}^{1} f(z)$$



Distribuzione normale standardizzata

Come utilizzare le tabelle delle Funzioni di Ripartizione?





Quanti soggetti hanno una statura inferiore a 183.3 cm?

μ_x = 175 cm

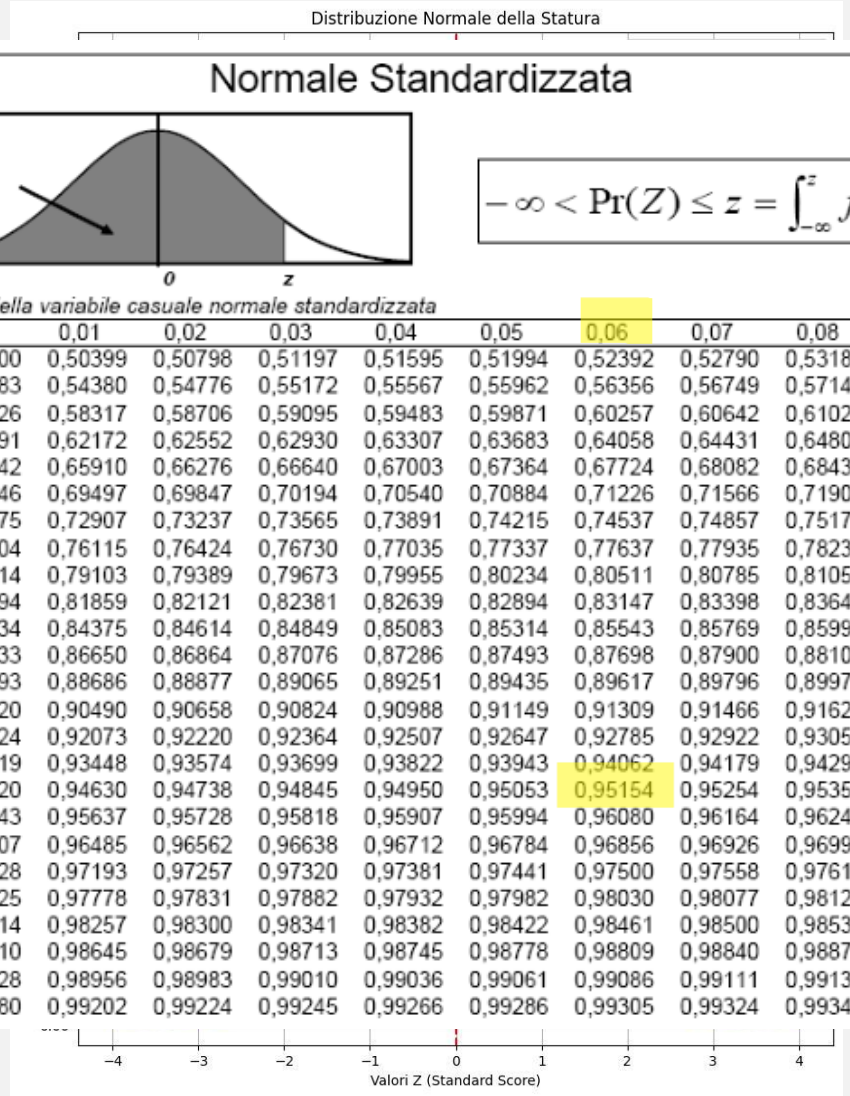
σ_x= 5 cm

N = 100

1. Trasformo X in Z : $$Z = \frac{X - \mu_x}{\sigma_x} = \frac{183 - 175}{5} = 1.6$$
2. Trovo l'area standardizzata tra -∞ e 1.6, tramite le tabelle

Distribuzione normale standardizzata

Come utilizzare le tabelle delle Funzioni di Ripartizione?



Normale Standardizzata

$$-\infty < \Pr(Z) \le z = \int_{-\infty}^{z} f(t)dt$$

(a) Integrali della variabile casuale normale standardizzata

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,0	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361

Quanti soggetti hanno una statura inferiore a 183.3 cm?

μ_x = 175 cm

σ_x= 5 cm

n = 100

1. Trasformo X in Z : $Z = \frac{X - \mu_x}{\sigma_x} = \frac{183.3 - 175}{5} = 1.66$
2. Trovo l'area standardizzata tra -∞ e 1.6, tramite le tabelle

$$\int_{-\infty}^{1.66} f(z)dz = 0.95154 \rightarrow 95.15\%$$

3. Calcolo quanti sono i soggetti con statura < 183.3 cm

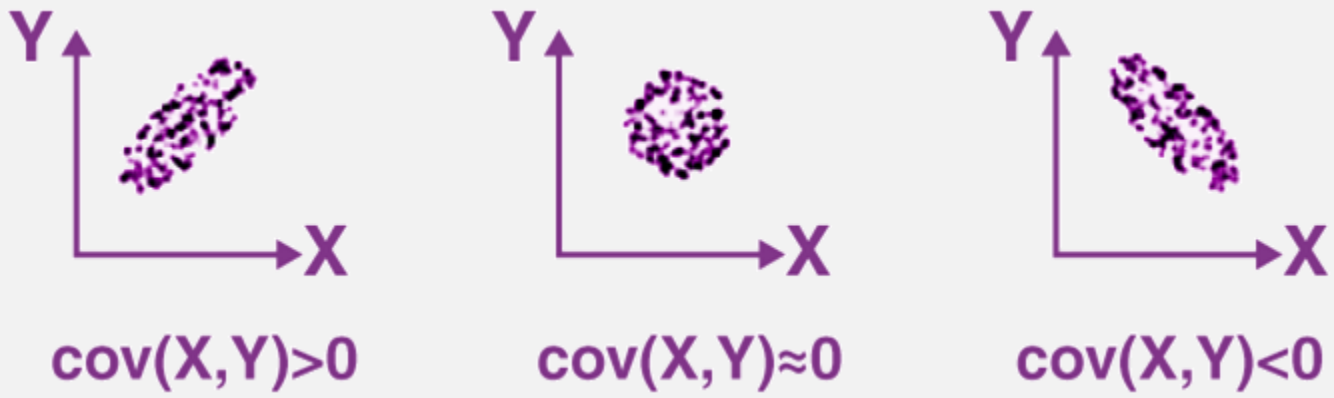
$$100 \cdot 0.95154 = \sim 95\ soggetti$$

Covarianza

- Indice che consente di verificare se fra due variabili statistiche esiste un legame lineare.

- Considerando due serie $\{x_i\}$ e $\{y_i\}, i = 1,2, \dots, n$, pone a confronto le coppie di scarti $(x_i - \bar{x})$ e $(y_i - \bar{y})$:

$$Cov(X, Y) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$



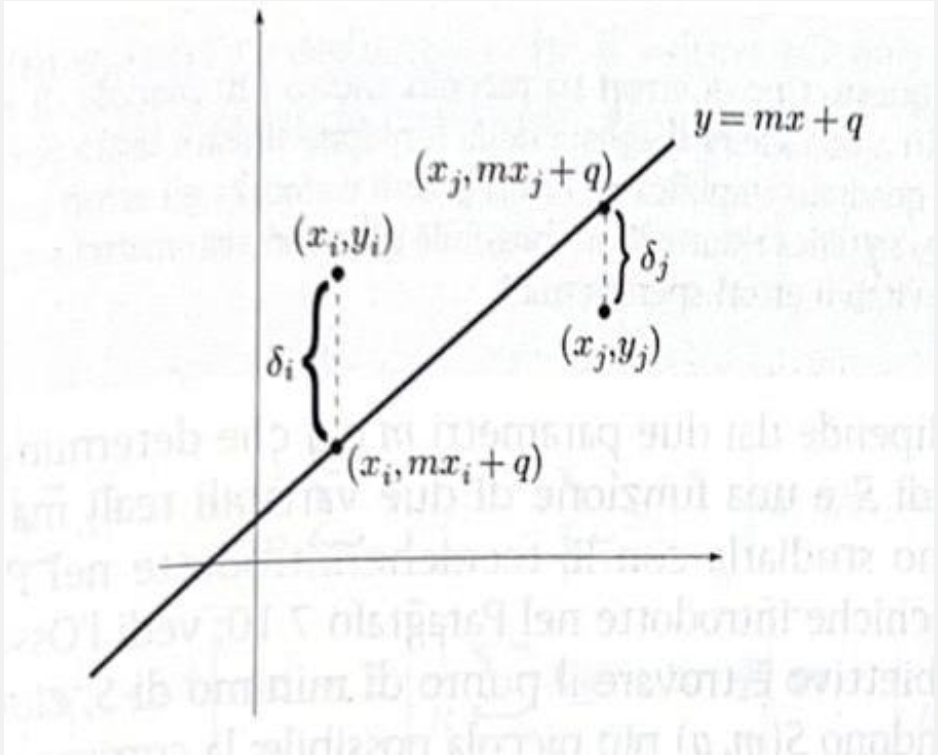
La Covarianza può essere:

- **POSITIVA**: quando X e Y variano tendenzialmente nella stessa direzione, cioè al crescere della X tende a crescere anche Y e al diminuire della X tende a diminuire anche Y
- **NEGATIVA**: quando le due variabili variano tendenzialmente in direzione opposta, cioè quando al crescere di una variabile l'altra variabile tende a diminuire (e viceversa)
- **NULLA**: quando non vi è alcuna tendenza delle 2 variabili a variare nella stessa direzione o in direzione opposta. Quando $Cov(X,Y) = 0$ si dice anche che X ed Y sono non correlate o linearmente indipendenti.

Analisi di regressione lineare

Metodo dei minimi quadrati

- ✓ Regressione: trovare la funzione lineare che meglio interpola un insieme di coppie di dati sperimentali.
- ✓ Lineare: disponiamo di *n* coppie di dati (x1, y1),… (xn, yn) e le ordinate dipendono in modo lineare dalle ascisse. La dipendenza tra queste due variabili viene fornita dalla *retta di regressione lineare*.
- ✓ La funzione lineare «migliore» è quella che fornisce l'errore minore se la usiamo per predire i risultati di un esperimento.



Correlazione

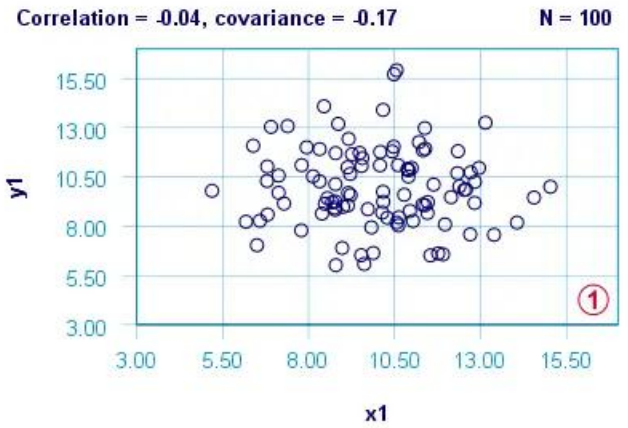
- È una modalità più rigorosa che consente di studiare il **grado di intensità** del legame lineare tra coppie di variabili

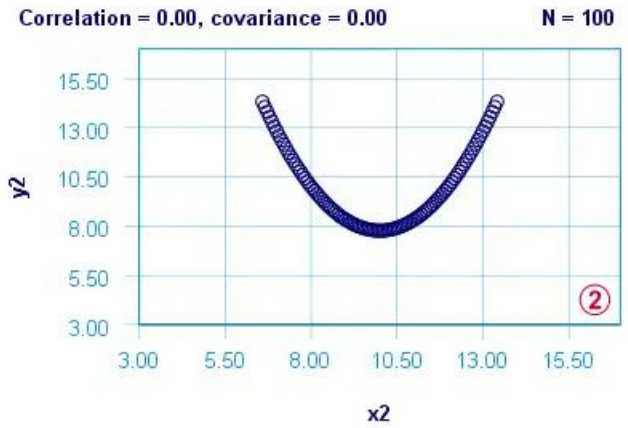
$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{(VarX)(VarY)}}$$

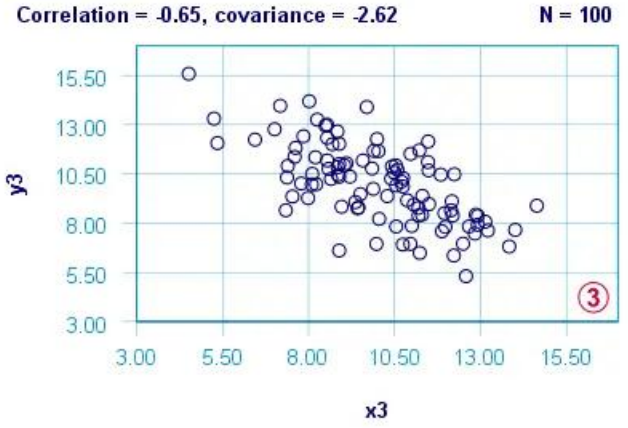
- Coefficiente di Pearson

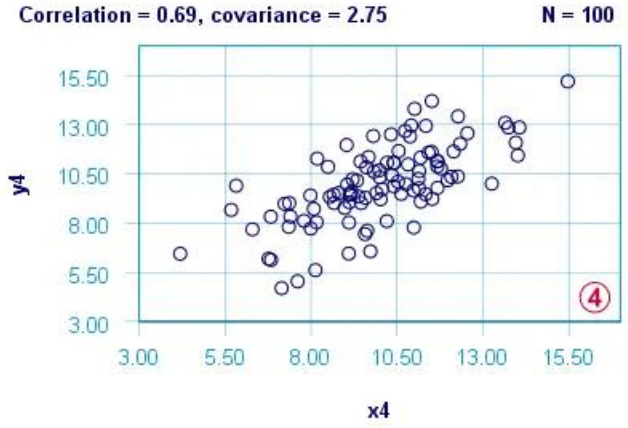
Il coefficiente di correlazione ci permette di:
- riassumere la forza della relazione **lineare** fra le variabili
- verificare l'apparente associazione fra le variabili

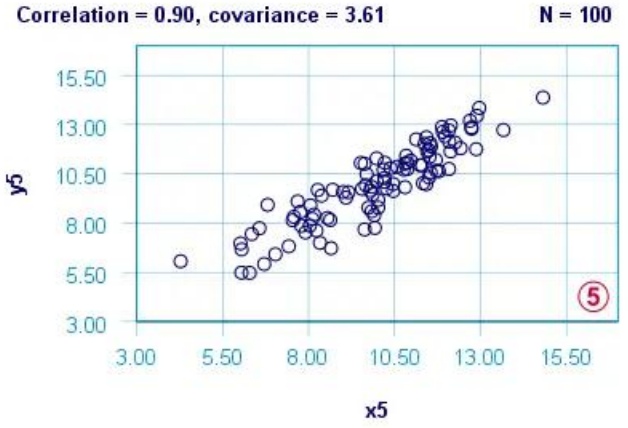
Il coefficiente di correlazione:
- varia da -1 a 1 (se uguale a 1 o a -1: perfettamente correlate)
- è positivo quando i valori delle variabili crescono insieme
- è negativo quando i valori di una variabile crescono al decrescere dei valori dell'altra
- non è influenzato dalle unità di misura

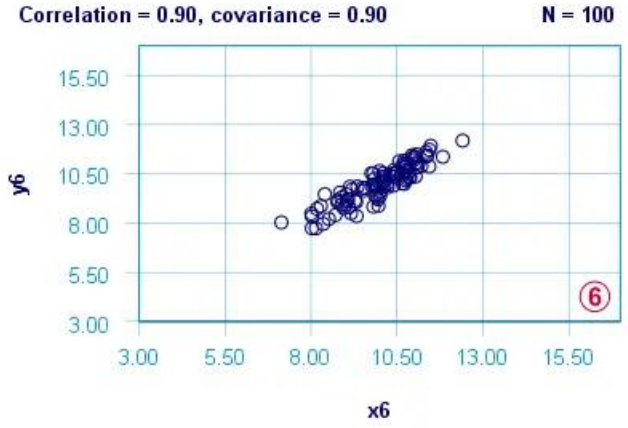

Correlation = -0.04, covariance = -0.17
N = 100
y1
x1
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
①

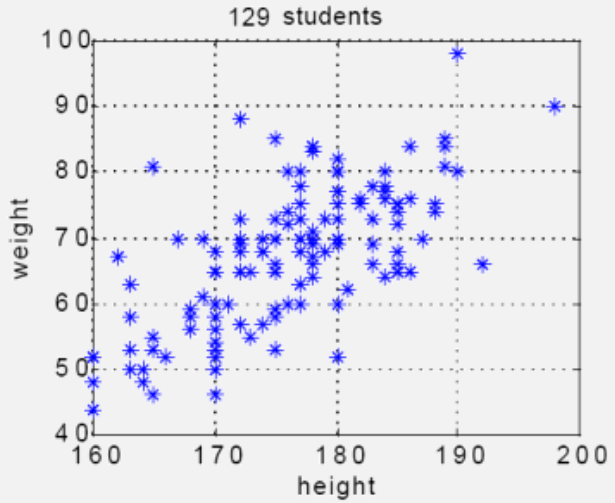

Correlation = 0.00, covariance = 0.00
N = 100
y2
x2
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
②

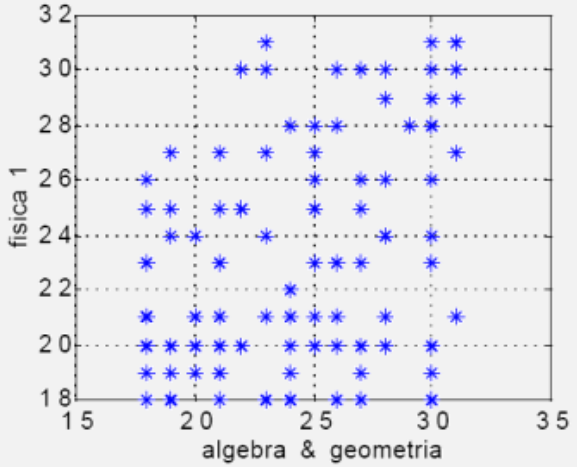

Correlation = -0.65, covariance = -2.62
N = 100
y3
x3
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
③

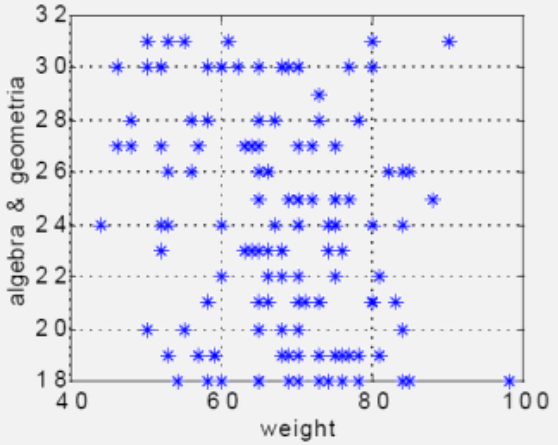

Correlation = 0.69, covariance = 2.75
N = 100
y4
x4
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
④

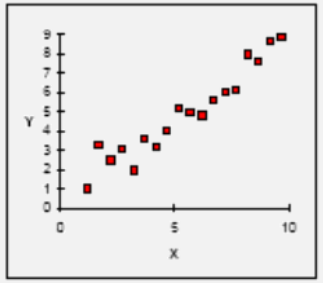

Correlation = 0.90, covariance = 3.61
N = 100
y5
x5
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
⑤


Correlation = 0.90, covariance = 0.90
N = 100
y6
x6
15.50
13.00
10.50
8.00
5.50
3.00
3.00
5.50
8.00
10.50
13.00
15.50
⑥



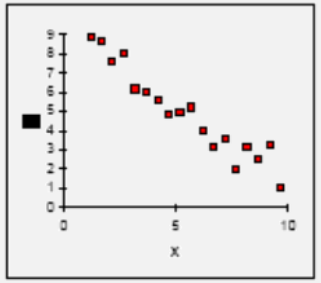


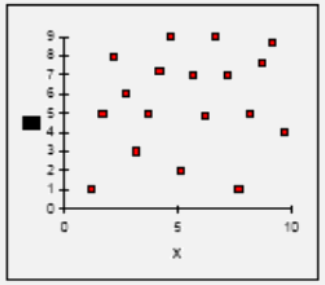




r=0,96

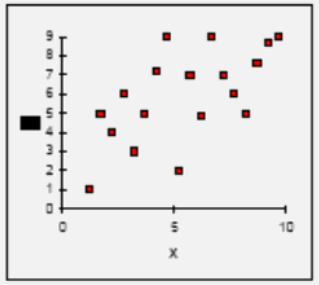
r=-0,96





r=0,12

r=0,62



Salvataggio automatico • esercitazione statis... • Salvataggio completato

Cerca

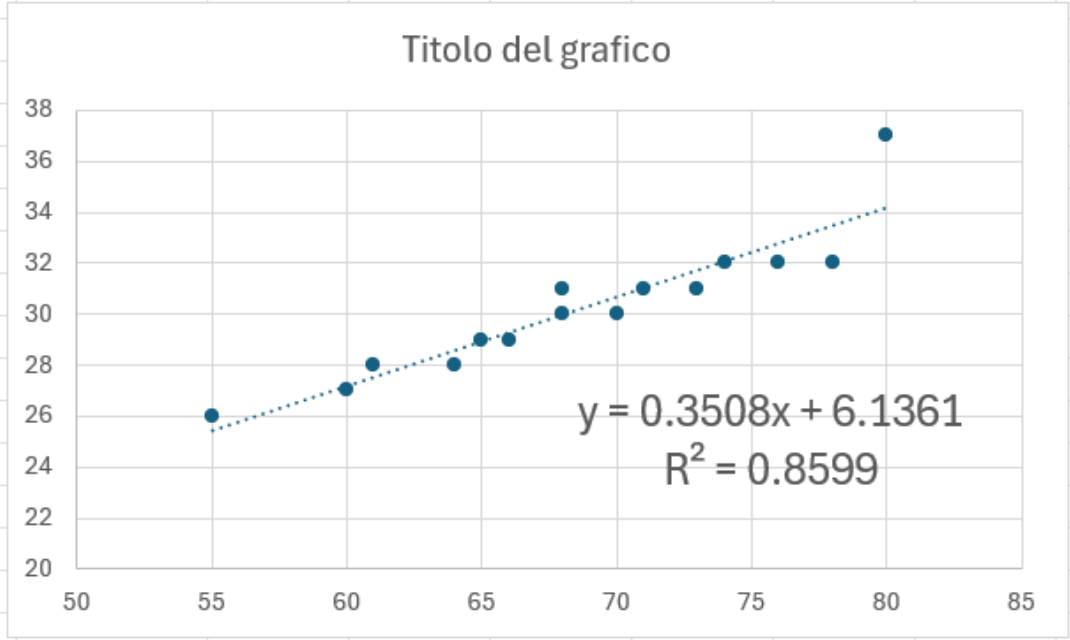
File Home Inserisci Disegno Layout di pagina Formule Dati Revisione Visualizza Automatizza Guida Acrobat

Incolla | Appunti | Carattere | Allineamento | Numeri | Formattazione condizionale | Formatta come tabella | Stili cella | Stili

SOMMA

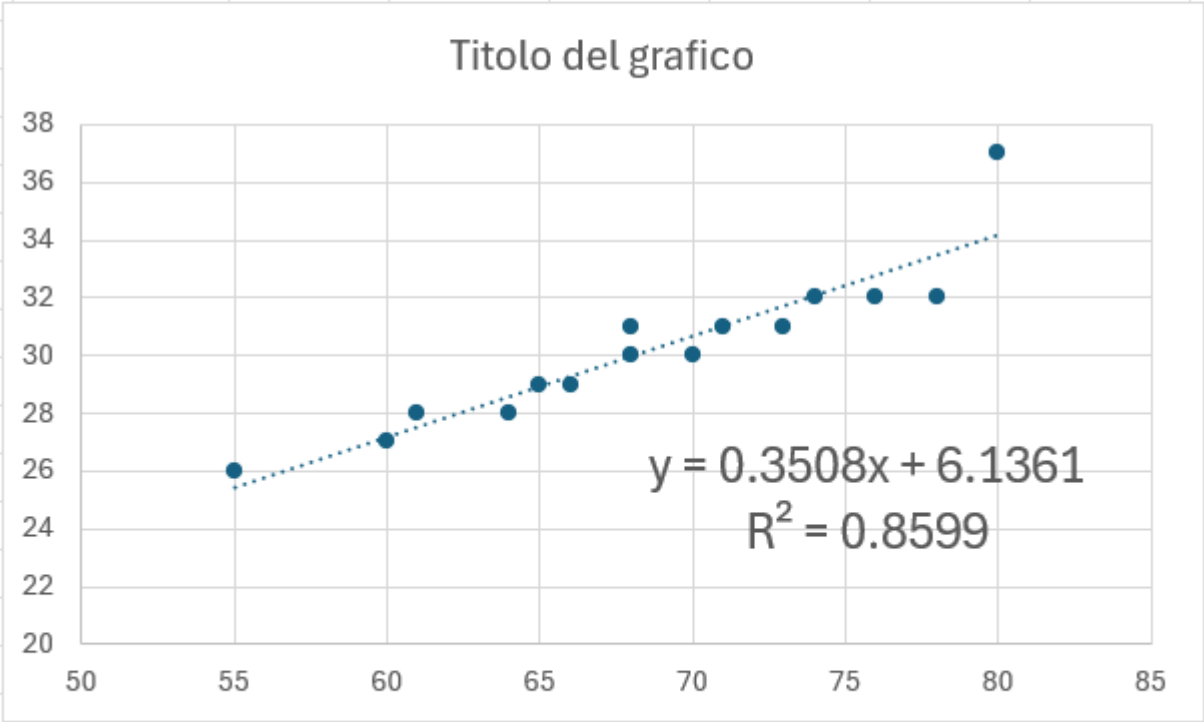
=(C17-A17*B17)/(D17-A17^2)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	**Età (x)**	**peso (y)**	**xy**	**x^2**	**y^2**		**m**	**q**						
2	61	28	1708	3721	784		=(C17-A17*B17)/(D17-A17^2)			**y = 0.351x + 6.13**				
3	76	32	2432	5776	1024									
4	80	37	2960	6400	1369									
5	66	29	1914	4356	841									
6	71	31	2201	5041	961									
7	68	30	2040	4624	900									
8	78	32	2496	6084	1024									
9	55	26	1430	3025	676									
10	74	32	2368	5476	1024									
11	60	27	1620	3600	729									
12	65	29	1885	4225	841									
13	70	30	2100	4900	900									
14	64	28	1792	4096	784									
15	73	31	2263	5329	961									
16	68	31	2108	4624	961									
17	**68.6**	**30.2**	**2087.8**	**4751.8**	**918.6**									
18	**x medio**	**y medio**	**xy medio**	**x^2 medio**	**y^2 medio**									
19														



O13

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Età (x)**	**peso (y)**	**xy**	**x^2**	**y^2**		**m**	**q**					
2	61	28	1708	3721	784		0.35078534	6.136126		**y = 0.351x + 6.13**			
3	76	32	2432	5776	1024								
4	80	37	2960	6400	1369								
5	66	29	1914	4356	841								
6	71	31	2201	5041	961								
7	68	30	2040	4624	900								
8	78	32	2496	6084	1024								
9	55	26	1430	3025	676								
10	74	32	2368	5476	1024								
11	60	27	1620	3600	729								
12	65	29	1885	4225	841								
13	70	30	2100	4900	900								
14	64	28	1792	4096	784								
15	73	31	2263	5329	961								
16	68	31	2108	4624	961								
17	**68.6**	**30.2**	**2087.8**	**4751.8**	**918.6**								
18	**x medio**	**y medio**	**xy medio**	**x^2 medio**	**y^2 medio**								
19													



Esercizio.

Supponiamo di misurare le altezze in centimetri al garrese di dieci cani partecipanti a un concorso canino, ottenendo la variabile statistica:

$$Y = (40, 42, 38, 41, 40, 45, 46, 42, 42, 41)$$

Si determini la mediana, la media aritmetica, la varianza.

Svolgimento

Per determinare la mediana dobbiamo disporre i dati in ordine crescente:

$$38, 40, 40, 41, 41, 42, 42, 42, 45, 46$$

Media: somma dei dati diviso 10 → 41.7 cm

Mediana: se il numero (n) di dati è pari, la mediana è stimata utilizzando i due valori centrali che occupano le posizioni $n/2$ e $n/2 + 1$→ media del quinto e del sesto valore → 41.5 cm

Varianza: $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$ → 5.01

Esercizio.

Uno studente di biologia ha conseguito i seguenti voti:

- Biologia 1 (6 CFU): 27
- Matematica (8 CFU): 24
- Genetica (6 CFU): 30
- Chimica (6 CFU): 25

Calcolare la media pesata secondo i CFU. Che cosa accade alla media pesata se lo studente consegue il voto 26 nell'esame di Istologia (4 CFU)?

Svolgimento

Calcoliamo la media pesata:

$$(27 \times 6 + 24 \times 8 + 30 \times 6 + 25 \times 6)/26 = 26.31$$

Anche senza calcoli possiamo affermare che la media pesata certamente diminuirà, anche se di una quantità molto piccola, dato che l'ultimo voto conseguito è vicino a tale media. Infatti:

$$(27 \times 6 + 24 \times 8 + 30 \times 6 + 25 \times 6 + 26 \times 4)/30 = 26.27$$

Media pesata: $\mu = \frac{\sum_{i=1}^{N} p_i X_i}{\sum_{i=1}^{N} p_i}$ → (27x6 + 24x8 + 30x6 + 25x6)/26=26.31

Dal momento che l'ultimo conseguito è vicino alla media pesata degli altri esami, anche senza calcoli possiamo intuire che la media pesata finale diminuirà di una quantità molto piccola:

(27x6 + 24x8 + 30x6 + 25x6 + 26x4)/26=26.27

Esercizio.

Il peso in tonnellate di alcuni capi di bestiame è dato dalla seguente variabile statistica:

$$Y = (1.4, 1.2, 1.3, 1.4, 1.4, 1.5, 1.6, 1.2, 1.2, 1.1, 0.9)$$

Calcolare media, varianza, moda e frequenze assolute e relative.

Svolgimento

La media è 1.29, mentre la varianza è 0.039. Abbiamo due valori modali: 1.2 e 1.4. La frequenza assoluta e quella relativa sono mostrate nella seguente tabella:

	0.9	1.1	1.2	1.3	1.4	1.5	1.6
Frequenze assolute	1	1	3	1	3	1	1
Frequenze relative	1/11	1/11	3/11	1/11	3/11	1/11	1/11

Esercizio.

Data la seguente variabile statistica

$$Y = (5, 4, 2, 2, 1, 7, 4, 6, 7, 3, 3, 2, 7, 4, 2, 3, 3, 1, 5, 6, 9, 7, 5, 6, 4)$$

calcolare le frequenze assolute e relative dei dati.

Calcolare poi la mediana, la media aritmetica, la varianza e lo scarto quadratico medio.

Mediana 4
Media aritmetica 4.32
Varianza 4.4576
Scarto quadratic medio 2.11

Esercizio.

Data la seguente variabile statistica

$$Y = (7, 6, 7, 4, 5, 8, 7, 7, 8, 6)$$

che corrisponde ai voti di un campione di studenti in una prima liceo. calcolare le frequenze assolute e relative dei dati.
Calcolare poi la mediana, la media aritmetica, la media geometrica, la varianza e lo scarto quadratico medio.

Mediana 7
Media aritmetica 6.5
Varianza 1.45
Scarto quadratico medio 1.2

Esercizio.

I voti ottenuti da un ragazzo nelle verifiche scritte di matematica sono stati: 7, 6, 4, 6, 7. La media aritmetica dei voti ottenuti è

A) 6

B) 5

C) 5.5

D) 6.5

Barrare la risposta corretta

La media dei voti ottenuti nelle prime tre verifiche scritte di matematica da un ragazzo è 5.5. Nella quarta verifica ottiene 6.5. Qual è la media dopo la quarta verifica?

A) 5.75

B) 5.5

C) 6

D) i dati sono insufficienti per calcolarla

Barrare la risposta corretta.

Esercizio.

I voti ottenuti da un ragazzo nelle verifiche scritte di matematica s ono stati: 7, 6, 4, 6, 7. La mediana dei voti ottenuti è:

A) 4

B) 6

C) 6.5

D) 7

Barrare la risposta corretta.

La mediana dei voti ottenuti nelle prime tre verifiche scritte di matematica da uno studente è 6.5. Nella quarta verifica prende 7.5. Cosa si può dedurre della mediana dei quattro voti?

A) Sicuramente aumenta.

B) Rimane 6.5.

C) Sicuramente non diminuisce.

D) Diventa 7.

Barrare la risposta corretta.