Statistica

Introduzione

Il termine statistica deriva da Stato perché è lo Stato che conduce i "censimenti" cioè delle indagini per conoscere il numero degli abitanti, la composizione della popolazione per età, sesso,condizioni economiche (il "censo") .
Si sono poi sviluppate indagini statistiche di vario genere oltre ai "censimenti" dello Stato.

La statistica nel suo sorgere ed evolversi non si discosta di molto dal percorso attuato da altre scienze poiché inizia come *attività pratica*, tesa alla soluzione di problemi concreti, e viene poi sviluppata come disciplina scientifica.
Da attività di conteggio, enumerazione ed anche di calcolo di semplici medie attuate su rilevazioni effettuate per scopi diversi si passa all'osservazione di *proprietà di un insieme di dati*, del quale si cerca di studiarne i seguenti aspetti:

- la **variabilità** degli stessi
- la **sintesi** attraverso varie medie
- la **dipendenza o indipendenza** di due caratteri.

Lo sviluppo della statistica moderna è associato ai nomi di F. Galton (1822-1911) , di K. Pearson (1857-1936) e di Fisher (1890-1962).

Lo studio statistico dei fenomeni riveste oggi grande importanza per poter risolvere e studiare molti problemi: ad esempio uno studio sulla vita media di una popolazione può influenzare le decisioni prese dal governo in campo pensionistico; lo studio degli effetti di un farmaco in via di sperimentazione su un campione di pazienti può far decidere se metterlo in commercio oppure no; in campo medico uno studio statistico può servire a individuare le cause dell'insorgenza di alcune patologie.

Inizieremo con il richiamare quanto già visto nell'introduzione alla statistica svolta nella prima classe, per poi continuare con lo studio delle tabelle a doppia entrata e la definizione di correlazione tra i dati.

Tabella statistica e sua rappresentazione

Quando si compie un'indagine statistica viene indagata la presenza di un certa caratteristica detta "carattere" all'interno di una certa "popolazione".

Il carattere considerato può manifestarsi con modalità diverse e può essere:

- **un carattere quantitativo** se le sue modalità sono espresse da numeri (discreto se può assumere un numero finito di valori o al più un'infinità numerabile o continuo se può assumere tutti i valori di un intervallo reale);
- **un carattere qualitativo** se le sue modalità non sono espresse da numeri.

Esempio 1

Supponiamo di chiedere agli studenti della 3B dell'anno scolastico in corso quale sport praticano di più tra calcio, nuoto, basket, pallavolo,danza e tennis .

La nostra "popolazione statistica" è costituita dagli studenti della 3B dell'anno scolastico in corso.

Il carattere indagato (sport praticato maggiormente) è di tipo qualitativo e le modalità considerate sono calcio, nuoto ecc.

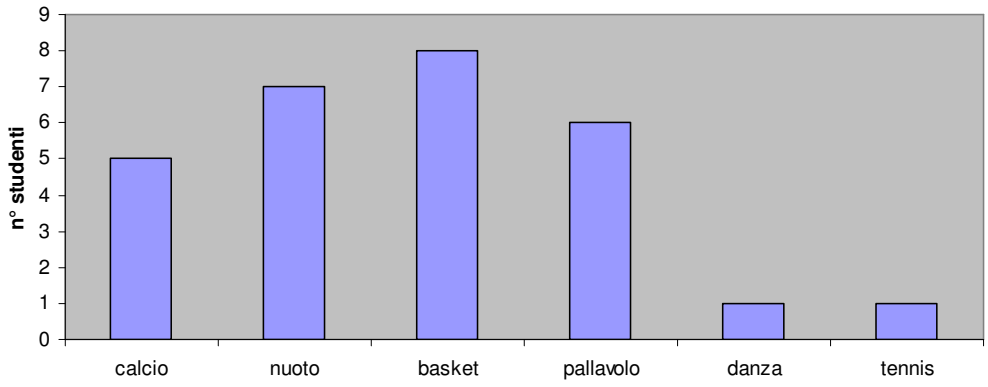
Per ciascuna modalità indichiamo il n° degli studenti che hanno indicato quella modalità come sport maggiormente praticato: la **frequenza** (assoluta) di una modalità è il numero delle volte che quella data modalità si presenta, mentre la **frequenza relativa** è il rapporto tra la frequenza assoluta e il numero delle unità statistiche, cioè degli studenti della 3B.

Supponiamo di avere ottenuto la seguente tabella:

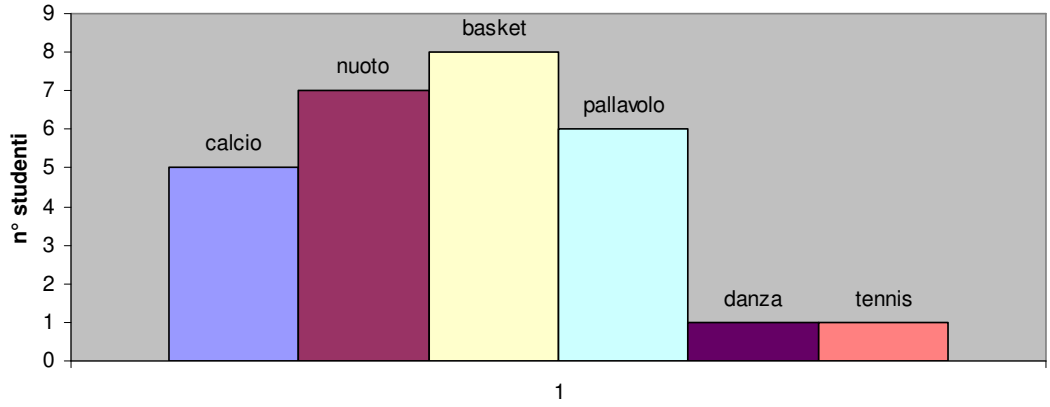
Sport praticato	**n° studenti della 3B (frequenza)**	**Frequenza relativa**	**Freq. Rel %**
calcio	5	5/28	….
nuoto	7	7/28=0,25	25%
basket	8	8/28	…
pallavolo	6	6/28	….
danza	1	1/28	…..
tennis	1	1/28	….

Possiamo rappresentare questi dati con:

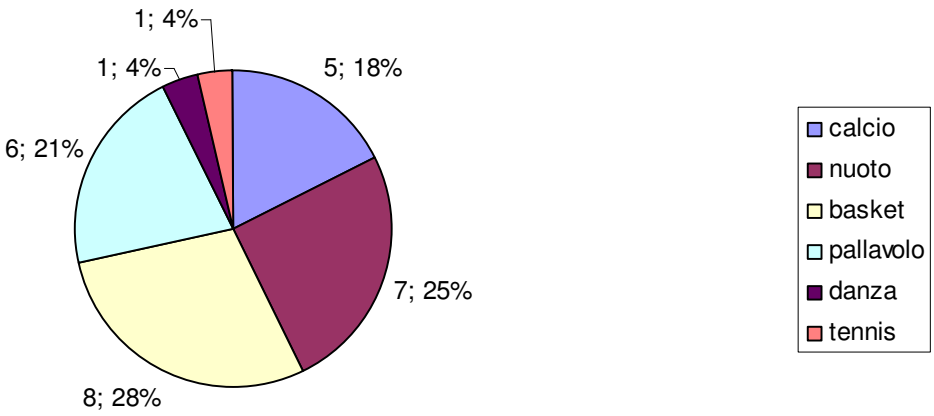
1) un **diagramma a barre** in cui le basi dei rettangoli distanziati corrispondono alle varie modalità e le altezze sono proporzionali alle frequenze



2) un **istogramma** in cui i rettangoli sono affiancati



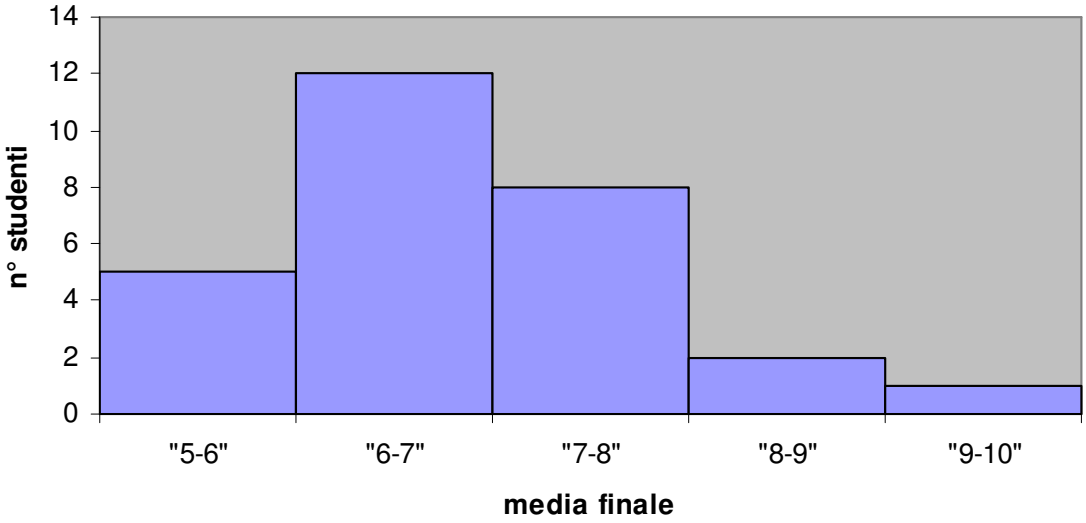
3) un **aerogramma** in cui un cerchio viene suddiviso in settori circolari corrispondenti alle varie modalità e ampiezza proporzionale alla frequenza relativa (o percentuale):

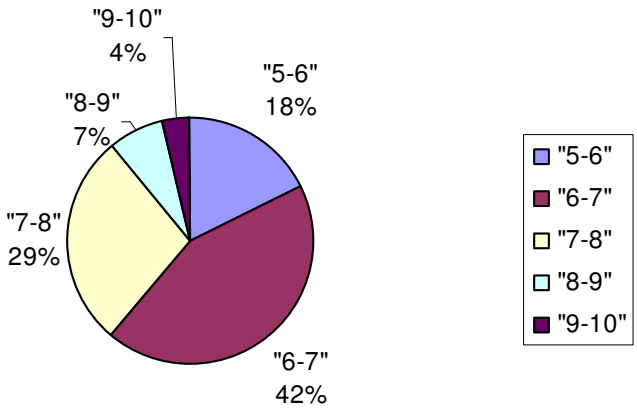


Nota: per determinare l'ampiezza α del settore corrispondente ad una data frequenza relativa percentuale f basta impostare la proporzione α : 360°=f : 100.
Se per esempio f=25% otteniamo α =90°.

Esempio 2

Consideriamo la media ottenuta nello scrutinio finale degli alunni della 3B del liceo scientifico in un dato anno scolastico e supponiamo di avere i seguenti dati raggruppati per "**classi di frequenza**" (la classe "5-6" comprende gli studenti che hanno una media finale m con $5 \le m < 6$): questa volta il carattere esaminato è un carattere quantitativo (discreto).

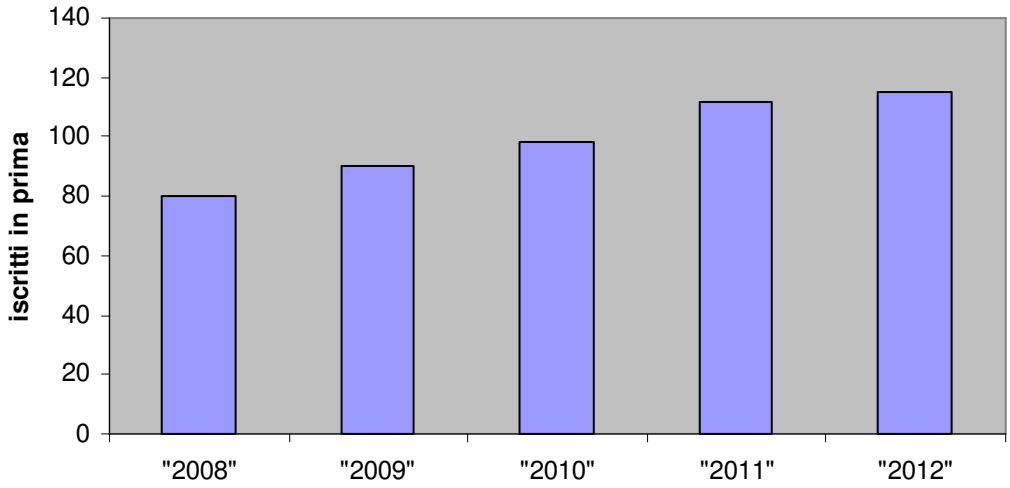




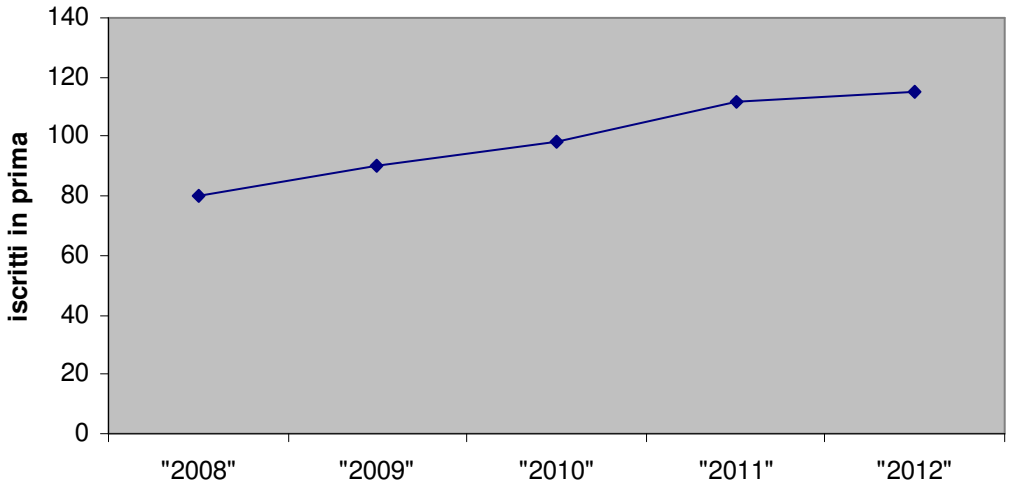
Esempio 3

Consideriamo il n° di studenti iscritti al primo anno del liceo scientifico nel nostro istituto negli anni dal 2008 al 2012 come riportato dalla seguente tabella:

anno scol.	n° studenti
"2008"	80
"2009"	90
"2010"	98
"2011"	112
"2012"	115



In questo caso si parla di **serie "storica**" e può essere utile rappresentarla con un diagramma cartesiano per visualizzare l'andamento del fenomeno.



Nota
La rappresentazione che si sceglie per visualizzare una tabella di dati dipende naturalmente anche dal tipo di tabella: in questo caso il diagramma cartesiano è sicuramente la rappresentazione più significativa.

Media aritmetica, moda, deviazione standard

Per sintetizzare i dati di una tabella statistica possiamo usare 3 "indici": la media aritmetica, la moda e la deviazione standard (o scarto quadratico medio).
Vediamo un esempio.

Esempio 4

Supponiamo di aver rilevato le seguenti temperature massime nei vari giorni dei mesi di marzo e luglio di un dato anno:

Giorno	Temp. Max. Marzo	Temp. Max Luglio
1	16	28
2	18	29
3	20	29
4	22	27
5	21	26
6	22	24
7	22	26
8	24	26
9	20	28
10	20	28
11	21	30
12	18	30
13	16	31
14	16	32
15	14	32
16	19	30
17	20	31
18	18	29
19	19	28
20	22	32
21	24	33
22	24	32
23	20	30
24	24	30
25	25	29
26	25	32
27	24	33
28	22	33
29	21	30
30	17	30
31	16	30

Definiamo i seguenti "indici":

- la **media aritmetica** $\bar{x}$ è la somma di tutti i dati $x_1,...x_n$ divisa per il numero dei dati cioè

$$\bar{x} = \frac{x_1 + ... + x_n}{n}$$

Nel nostro caso per calcolarla possiamo sommare tutte le temperature oppure determinare la frequenza di ciascuna temperatura : se per esempio nel mese di Luglio la temperatura 24 ha frequenza 1, la temperatura 26 ha frequenza 3 , la temperatura 27 frequenza 1, la temperatura 28 frequenza 4…possiamo scrivere

$$media\_aritmetica = \frac{24 \cdot 1 + 26 \cdot 3 + 27 \cdot 1 + 28 \cdot 4 + ...}{31}$$

Otteniamo:

Temp max media Marzo	Temp max media Luglio
20,3	29,6

- la **moda** è il dato che ha la massima frequenza

Temp moda Marzo	Temp moda Luglio
20	30

- **la deviazione standard** σ (o scarto quadratico medio) è definita così

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + ...(x_n - \bar{x})^2}{n}}$$

e ci dà un'idea di quanto i dati sono vicini alla loro media: se σ è piccolo i dati sono tutti vicini alla media aritmetica, se invece è grande sono "dispersi".
Nel nostro esempio abbiamo:

dev. Standard Marzo	dev. Standard Luglio
3,0	2,3

Osserviamo quindi che a Marzo i dati sono più "dispersi" rispetto alla loro media aritmetica rispetto al mese di Luglio.

Nota: se i dati vengono riportati in un foglio Excel, abbiamo a disposizione le tre funzioni

MEDIA, MODA DEV.ST.POP (dev. Stand. su tutta la popolazione)

che permettono di calcolarle automaticamente inserendo l'intervallo dei dati da considerare cioè, relativamente per esempio a Marzo, basterà scrivere (se i dati di marzo sono nelle celle b2..b32):

=media(b2:b32)
=moda(b2:b32)
=dev.st.pop(b2:b32)

Esercizio : considera le seguenti misurazioni del diametro di due tubi capillari effettuate con un calibro ed espresse in millimetri

misura	diametro tubo1	diametro tubo 2
1	5,25	4,8
2	5,26	4,7
3	5,24	4,75
4	5,2	4,74
5	5,21	4,87
6	5,23	4,73
7	5,25	4,72
8	5,24	4,82
9	5,24	4,76
10	5,19	4,78
11	5,2	4,83
12	5,23	4,76
13	5,22	4,8
14	5,26	4,75
15	5,24	4,8
16	5,24	4,75
17	5,22	4,82
18	5,2	4,81
19	5,23	4,73
20	5,26	4,7

Calcola la media aritmetica delle due serie misure, la moda e la deviazione standard.
Quale delle due serie di misure è stata effettuata in modo migliore cioè ha la minore deviazione standard?

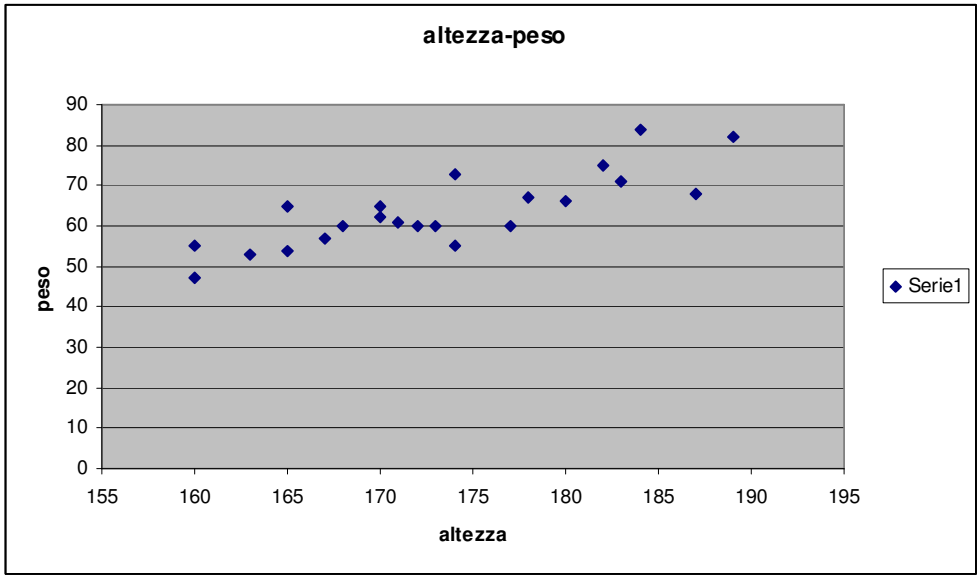
Tabelle a doppia entrata

Finora abbiamo analizzato tabelle in cui viene considerato un solo carattere di una data popolazione (statistica univariata): vediamo un esempio in cui vengono rilevati due caratteri, quantitativi X e Y (statistica bivariata).

Esempio

Consideriamo la tabella seguente relativa ad altezze e pesi degli studenti di una classe, riportiamo i dati in un foglio elettronico e disegniamo il grafico (x, y) corrispondente : possiamo osservare che i punti seguono un andamento "lineare" cioè risultano grosso modo disposti lungo una retta.

altezza	peso
167	57
163	53
170	62
160	47
178	67
189	82
160	55
184	84
187	68
168	60
183	71
180	66
172	60
173	60
171	61
182	75
170	65
165	65
174	73
165	54
174	55
177	60

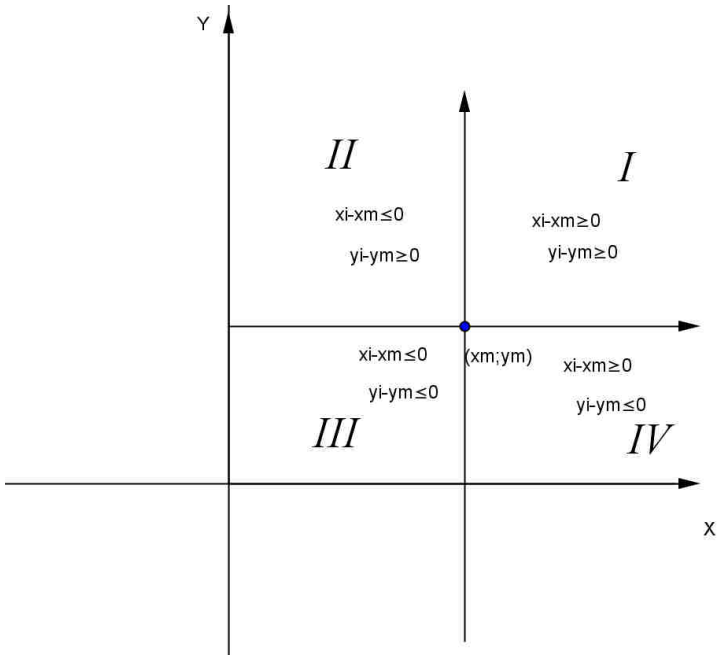


In generale se due caratteri X e Y di tipo quantitativo prendono i valori $x_1,\dots x_n$ e $y_1,\dots y_n$ possiamo calcolare la "**covarianza**" σ_{XY} definita nel modo seguente

$$\sigma_{XY} = \frac{(x_1 - \bar{x}) \cdot (y_1 - \bar{y}) + \dots (x_n - \bar{x}) \cdot (y_n - \bar{y})}{n}$$

dove $\bar{x}$ e $\bar{y}$ rappresentano la media aritmetica delle modalità con cui si presentano i due caratteri e n è il numero dei dati.

Osserviamo che il punto di coordinate $(\bar{x}, \bar{y})$ divide il piano cartesiano in quattro zone in cui i prodotti $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ sono positivi o negativi (vedi figura in cui $\bar{x}$ è stato indicato con xm e analogamente $\bar{y}$ con ym).



- Se $\sigma_{X,Y} > 0$ significa che la maggior parte dei prodotti $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ sono positivi e quindi i punti (x_i, y_i) si trovano soprattutto nelle zone I e III e la "nuvola" di punti ha una forma di tipo lineare crescente;
- Se $\sigma_{X,Y} < 0$ invece la maggior parte dei prodotti $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ sono negativi e quindi i punti (x_i, y_i) cadono soprattutto nelle zone II e IV e la forma della "nuvola" di punti è di tipo lineare decrescente;
- Se $\sigma_{X,Y}$ è vicina a zero significa che i punti sono sparpagliati senza alcuna regolarità oppure sono disposti secondo relazioni diverse da quella lineare.

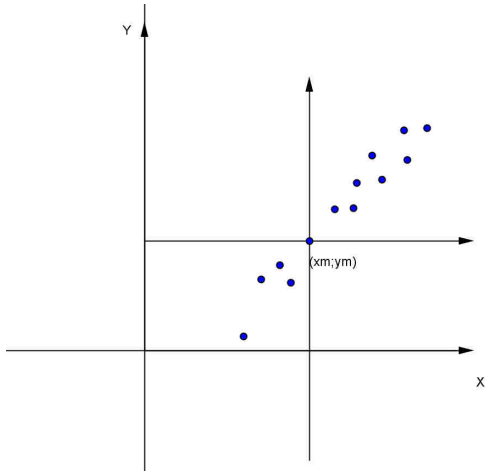
Poiché si può dimostrare che

$$-\sigma_X \cdot \sigma_Y \leq \sigma_{XY} \leq \sigma_X \cdot \sigma_Y$$

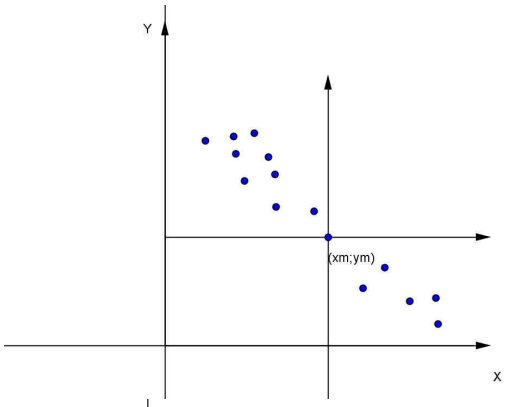
(dove σ_X e σ_Y sono le deviazioni standard di X e Y)

si definisce il **coefficiente di correlazione lineare** $r = \dfrac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$ che risulta compreso tra -1 e 1.

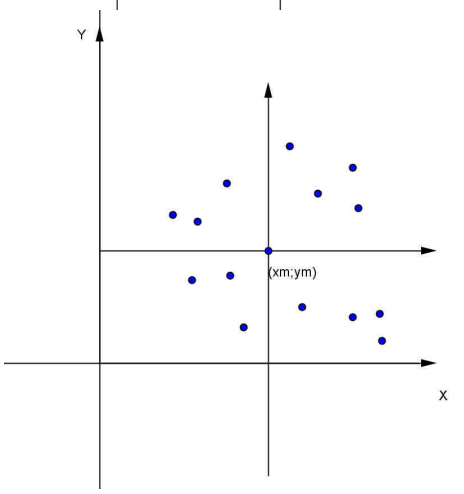
Quindi:



- Se $r > 0$ si tratterà di una **correlazione "diretta"** cioè i caratteri X e Y avranno una relazione lineare crescente, ma solo se r è vicino a 1 il modello lineare interpreterà bene la relazione tra X e Y



- Se $r < 0$ si tratterà di una **correlazione "inversa"** cioè i caratteri X e Y avranno una relazione lineare decrescente, ma solo se r è vicino a -1 il modello lineare interpreterà bene la relazione tra X e Y

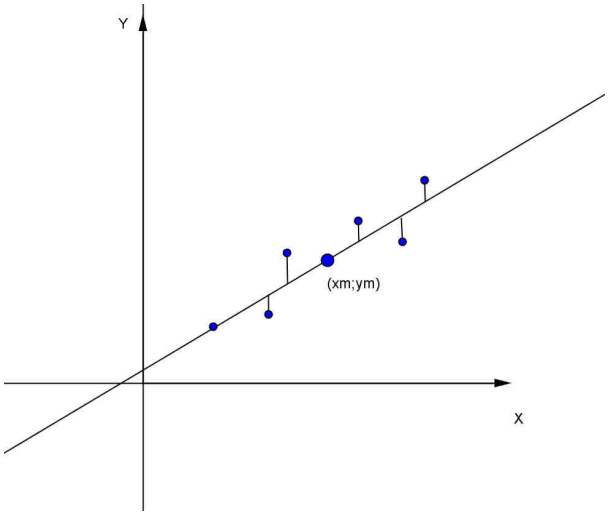


- Se r è vicino a zero il legame tra X e Y è lontano da quello lineare

Se r è vicino a 1 oppure a -1 si cerca la "**retta di regressione**" o retta interpolante che più si avvicina ai nostri dati.

Nota: la retta di regressione $f(x) = m \cdot x + q$ sarà quella retta che rende minima la somma dei quadrati delle differenze tra il valore del dato y_i che si è ottenuto in relazione alla modalità x_i e il valore $f(x_i)$ che si avrebbe sulla retta.

Si può dimostrare che la retta di regressione passa per $(\overline{x}, \overline{y})$ e $m = \dfrac{\sigma_{XY}}{{\sigma_X}^2}$.



Torniamo al nostro esempio.

Calcoliamo **il coefficiente di correlazione lineare** utilizzando la funzione di Excel

=CORRELAZIONE(a2:a23;b2:b23)

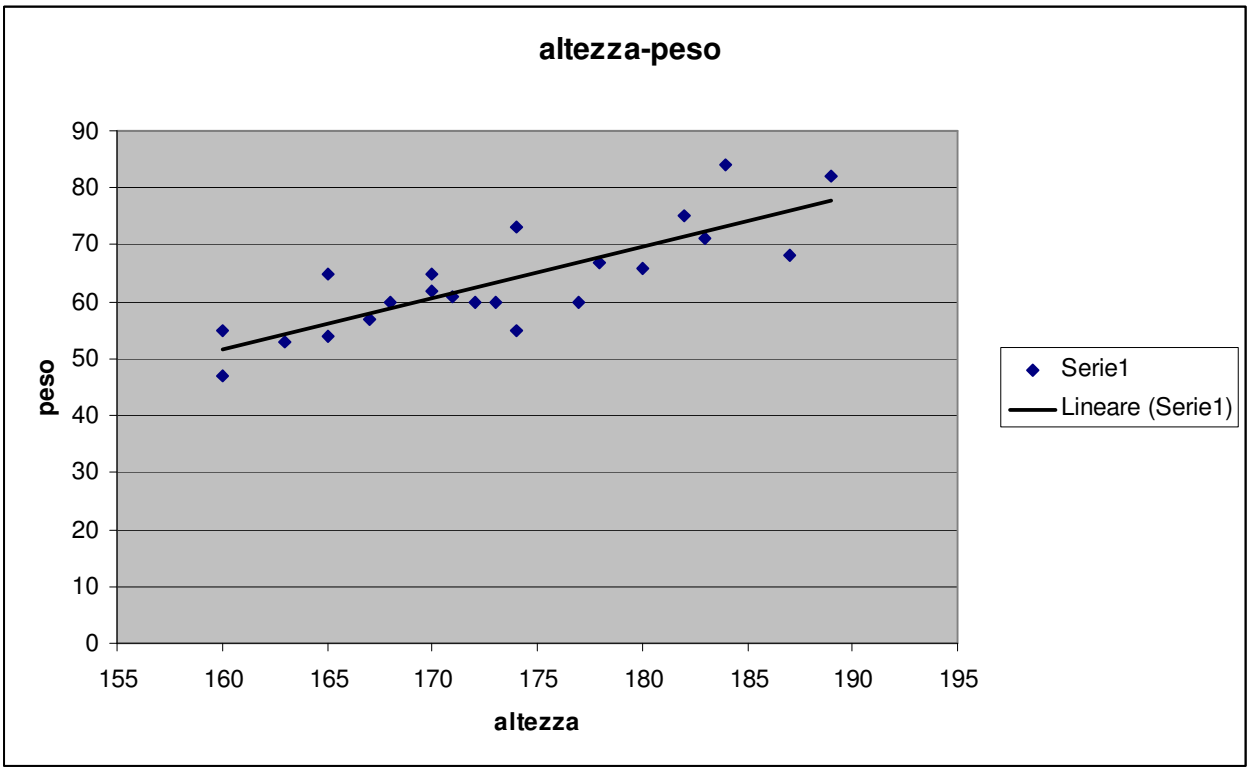
e otteniamo (approssimato a due cifre decimali) 0,82

Quindi **si tratta di una correlazione lineare diretta** e quindi possiamo cercare la retta "interpolante" (detta anche retta di regressione) cioè la retta che più si avvicina ai punti del nostro grafico.
Possiamo usare una funzione presente nel foglio di calcolo Excel: se scriviamo

=REGR.LIN (b2:b23;a2:a23)

otteniamo 0,90 , cioè l'inclinazione della retta che interpola i nostri punti.

Per disegnare la retta di regressione basta cliccare con il tasto destro su un qualsiasi punto del grafico e scegliere "**Aggiungi linea di tendenza**" e scegliere (in questo caso) "lineare".



Possiamo verificare che la retta interpolante passa per la media delle altezze (173,27) e la media dei pesi (63,64).

Approfondimento

1) Se tra i caratteri X e Y sospettiamo (osservando la distribuzione dei punti (x_i, y_i))che vi sia una dipendenza quadratica cioè del tipo $Y = a \cdot X^2$ possiamo studiare la dipendenza tra X^2 e Y: se troviamo una correlazione lineare tra X^2 e Y è chiaro che abbiamo dimostrato che c'è una dipendenza quadratica tra X e Y.

Esercizio
Considera i dati della seguente tabella riferiti ai tempi e agli spazi rilevati da un sensore di moto disposto come nell'esempio 6 ma nel caso che la rotaia sia inclinata rispetto all'orizzontale e non venga impressa nessuna velocità iniziale al carrello.

tempo (x)	Spazio (y)
0,1	0,02
0,2	0,07
0,3	0,19
0,4	0,3
0,5	0,48
0,6	0,7
0,7	1
0,8	1,3
0,9	1,58
1	1,95

Elabora i dati impostando un foglio Excel.Quale dipendenza trovi tra tempo (X) e spazio (Y)?

2) Se invece tra X e Y pensiamo che ci sia un legame esponenziale cioè che sia $Y = b \cdot a^{mX}$ possiamo vedere se c'è correlazione lineare tra X e $\log_a Y$ oppure tra X e lnY (logaritmo in base e) o tra X e logY (logaritmo in base 10). Infatti se, per esempio, $\log_a Y = mX + q$ allora avremo che $Y = a^{mX+q} \Rightarrow Y = a^{mX} \cdot a^q$ e quindi ponendo $a^q = b$ avremo una relazione del tipo $Y = b \cdot a^{mX}$.

Esercizio
Considera i dati della seguente tabella che indicano il numero di batteri rilevati al passare delle ore in una cultura:

tempo(ore) (X)	n°batteri (Y)
1	2
2	5
3	7
4	15
5	30
6	60
7	120
8	250
9	500
10	1010

Trasferisci i dati in un foglio di Excel e cerca di capire che legame c'è tra X e Y.