UNIVERSITA' DEGLI STUDI DI PERUGIA
DIPARTIMENTO DI FILOSOFIA SCIENZE SOCIALI UMANE E DELLA FORMAZIONE
Corso di Laurea in Scienze per l'Investigazione e la Sicurezza

# 8. REGRESSIONE E CORRELAZIONE

Prof. Maurizio Pertichetti

Statistica sociale

# 8. REGRESSIONE E CORRELAZIONE

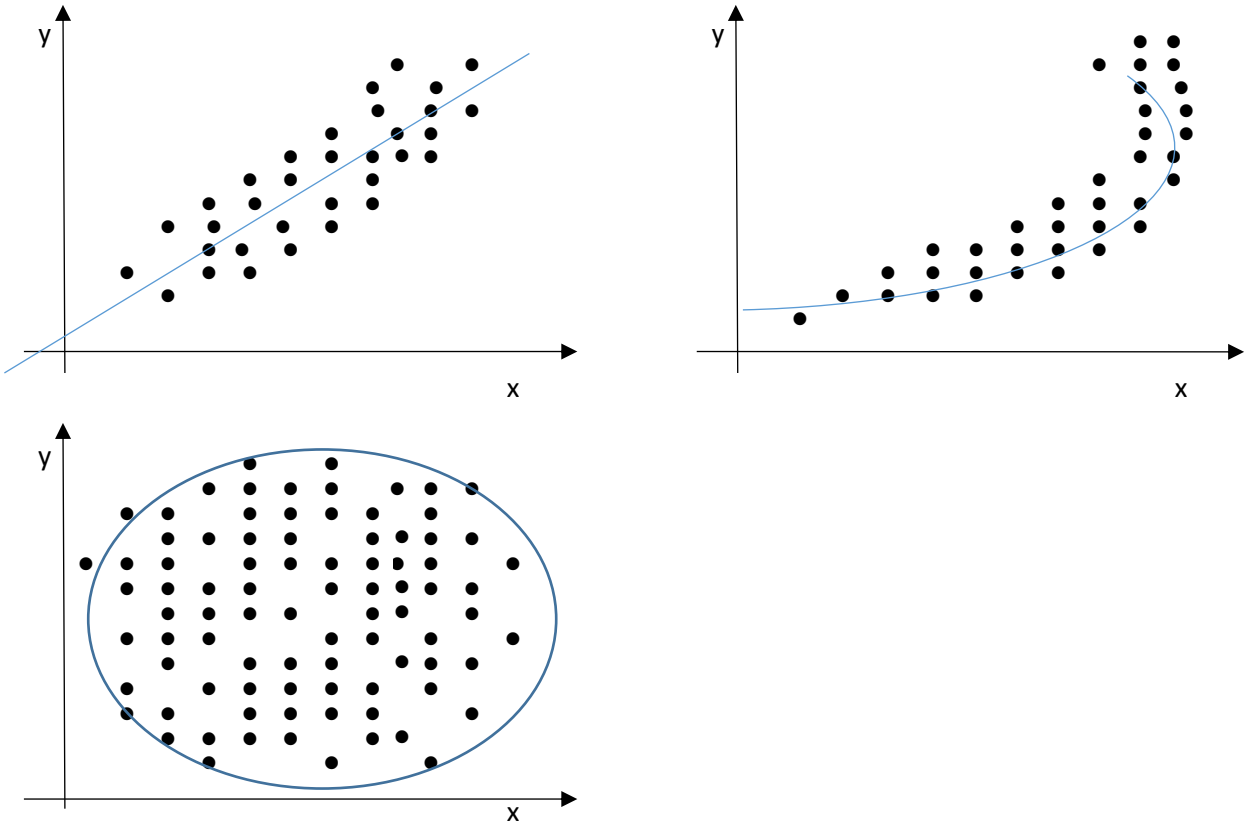
Come abbiamo già detto, nell'analisi dei dati si è sempre più interessati a studiare se tra due o più caratteri, congiuntamente considerati sulle unità statistiche di una popolazione, vi possa essere una qualche relazione ed eventualmente quale ne possa essere la misura. E abbiamo anche detto che in particolare vi è interesse a studiare l'esistenza di forme di dipendenza (o indipendenza) attraverso l'esplicitazione di una funzione analitica.

Nell'analisi statistica per **regressione** si intende la ricerca di un modello atto a descrivere la relazione esistente tra una variabile dipendente e una o più variabili indipendenti o esplicative.

La scelta dell'una o dell'altra variabile come indipendente non è arbitraria ma legata alla natura del fenomeno, nel senso che si sceglie come indipendente la variabile che sia *logicamente antecedente* rispetto all'altra.

Per effettuare una regressione si fa riferimento a modelli teorici di vario tipo: lineare, parabolico, esponenziale, logaritmico, etc. Per cui una volta accertata l'esistenza di una relazione tra due variabili, si deve cercare di trovare la **funzione statistica**, ovvero l'espressione analitica di tale relazione sotto forma di equazione che leghi fra loro le variabili.

Per evidenziare il tipo di legame tra le variabili è di notevole ausilio il diagramma in coordinate cartesiane, o a dispersione, o scatter plot, ossia il diagramma empirico costituito dalle **n coppie di osservazioni sulle variabili** e **rappresentate da una nuvola di punti**.




Generalmente una funzione statistica è rappresentata in termini grafici da una spezzata, in cui si assumono come **variabili indipendenti** le modalità del carattere X, poste sull'asse delle ascisse, e come **variabili dipendenti** le corrispondenti modalità di Y, poste sull'asse delle ordinate.

Dall'analisi del diagramma a dispersione è spesso possibile avere una rappresentazione intuitiva del tipo di relazione e di conseguenza di quale modello teorico (lineare, parabolico, esponenziale, logaritmico, etc come detto) adottare.

Con **interpolazione** si intende l'individuazione di una funzione matematica che passi **per** tutti i punti (x,y) dati o **fra** di essi. La funzione così individuata dovrà rappresentare al meglio l'andamento espresso dai punti.

Il procedimento si attua sia analiticamente sia graficamente:

- la **rappresentazione analitica** consiste nel trovare una funzione matematica che rappresenti nel miglior modo possibile la distribuzione osservata del fenomeno;
- la **rappresentazione grafica** consiste nel sostituire al diagramma rappresentativo dei dati osservati una *curva teorica* associata ad una funzione matematica.

Per realizzare una corretta rappresentazione analitica in un processo di interpolazione, lo statistico deve:

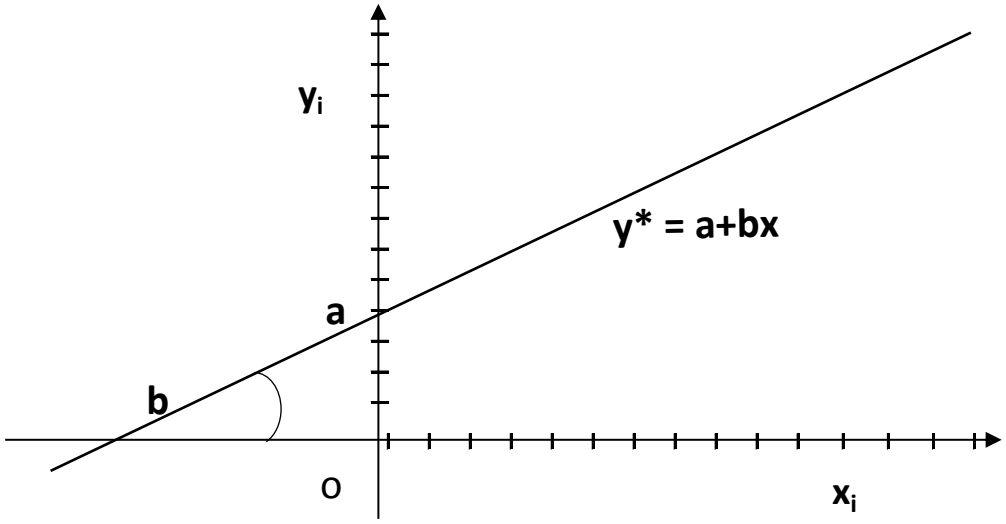
- mutuare dalla matematica una **funzione teorica** in grado di rappresentare con una legge matematica la distribuzione empirica, ovvero una volta trovata, la legge matematica sostituirà nelle diverse applicazioni la legge statistica;
- determinare numericamente i **parametri** che compaiono nella funzione matematica;
- verificare il **grado di accostamento** tra i valori empirici (o osservati) delle frequenze o delle intensità e i valori teorici ottenuti attraverso la funzione matematica.

Limiteremo l'analisi all'ipotesi in cui la relazione tra variabili (causa - effetto) sia di tipo lineare e pertanto che la funzione teorica atta a rappresentare tale relazione sia un'equazione di primo grado, ovvero che ad interpolare efficacemente la nuvola di punti sia una retta. La retta sarà detta **retta di regressione** e la sua equazione sarà chiamata **equazione di regressione di Y su X**.

Posta in forma esplicita, la generica equazione canonica di primo grado in due incognite della retta di regressione è data da: $\mathbf{y^* = a+bx}$ . Ad ogni equazione di questo tipo, una volta assegnati i valori ad **a** e **b** , corrisponde una e una sola retta del piano cartesiano.

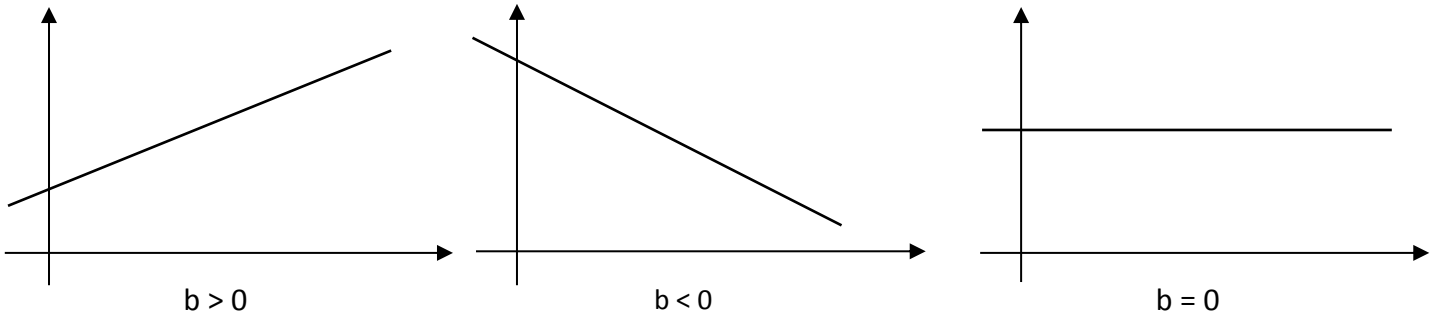
Assunta **x** come variabile indipendente e **y** come variabile dipendente, dalla geometria analitica sappiamo che **a** e **b** sono numeri reali fissati non contemporaneamente nulli:

- **a** si chiama **intercetta** della retta sull'asse delle Y, ovvero il valore della y quando x = 0 ;
- **b** si chiama coeficiente angolare della retta e da la sua pendenza, ovvero l'angolo che essa forma con l'asse delle ascisse.




A seconda del valore assunto dal coefficiente **b** si desume l'associazione tra X e Y, infatti se:

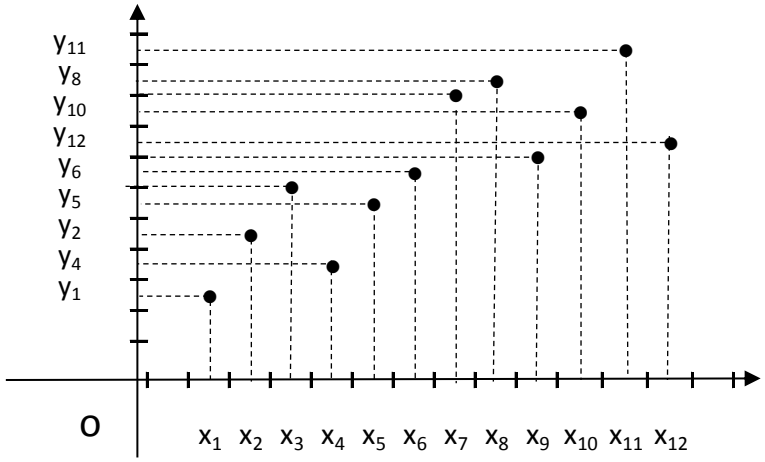
- $b > 0$ , l'associazione tra le variabili x e y è positiva, nel senso che al crescere di x anche y cresce;
- $b < 0$ , l'associazione tra le variabili x e y è negativa, nel senso che al crescere di x la variabile y decresce;
- $b = 0$ , non esiste associazione lineare tra x e y.



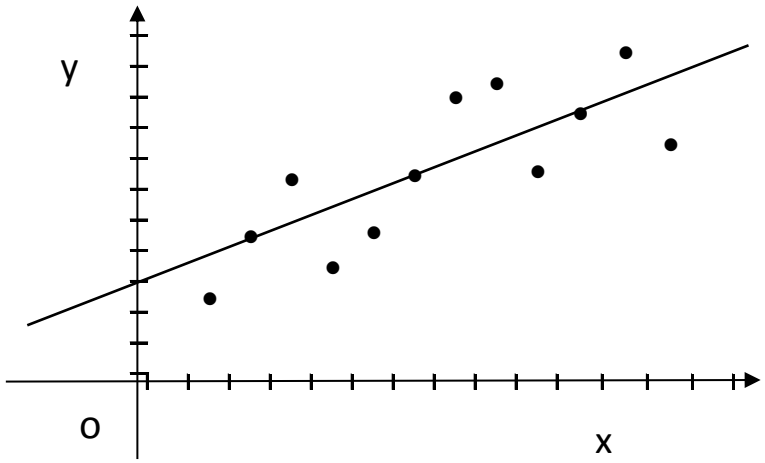
Se dunque è una retta, **retta di regressione**, il modello più appropriato in grado di descrivere la relazione tra le variabili il problema che si pone è quello di individuare in maniera analitica la migliore retta interpolante, ossia la migliore coppia di parametri **a** e **b** da utilizzare.

Esistono diversi metodi per determinare i parametri di una funzione matematica in un procedimento di interpolazione, tuttavia quello più utilizzato è **il metodo dei minimi quadrati** che si definisce come quel metodo che consente di determinare valori dei parametri tali per cui la retta teorica che ne risulta ha la proprietà di *rendere minima la somma dei quadrati degli scarti tra valori teorici e valori osservati* .

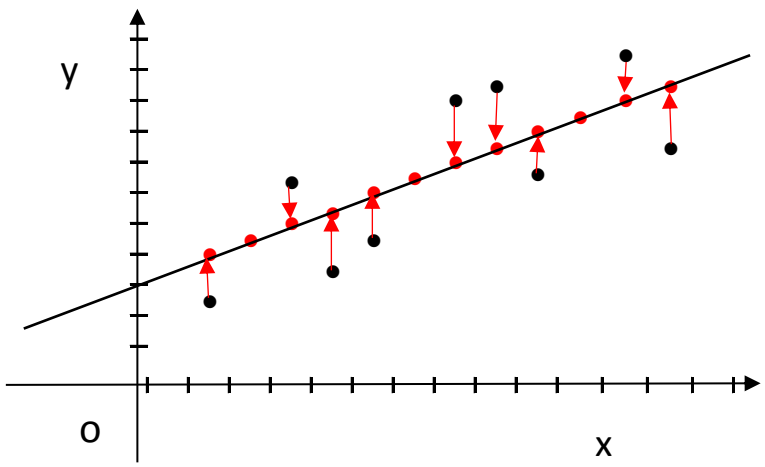
Immaginiamo di aver effettuato alcune osservazioni e di aver riportato i risultati sul un diagramma in coordinate cartesiane.



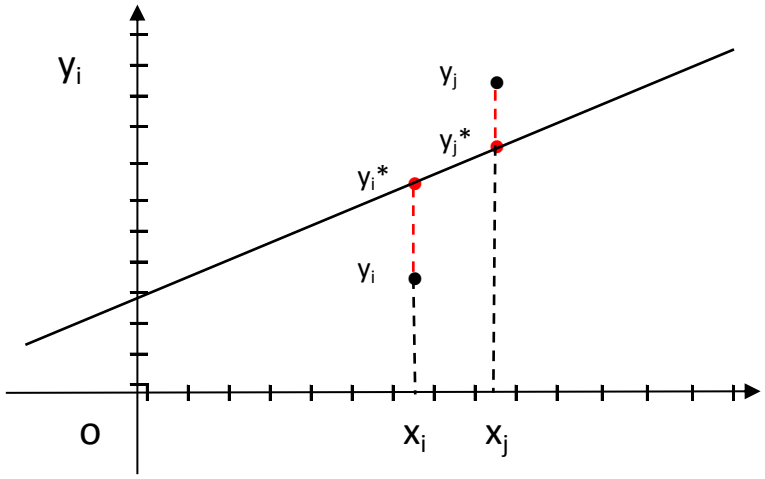

Ipotizzando l'esistenza di una relazione lineare, il problema, per descrivere tale relazione tra le variabili, è quello di individuare in maniera analitica la migliore retta interpolante.




la retta, una volta trovata, diverrà la regolarità matematica che sostituirà, o meglio dire approssimerà, l'esperienza statistica, cosicché ciascuno dei valori $y_i$ delle osservazioni, ovvero della distribuzione empirica, in corrispondenza di ciascun valore di $x_i$ sarà sostituito da quello teorico $y_i^*$ del modello che verrà ad incrociarsi con la retta.



Se dunque si stabilisce che per ciascun valore di $x_i$ , i valori teorici sono dati dalle $y_i^*$, mentre i valori osservati sono dati dalle $y_i$ , e altresì che la funzione interpolatrice è $Y^*= f(x; a,b)$, quello dei minimi quadrati è il metodo che consente di determinare i valori dei parametri di quella retta teorica in grado di rendere la $\sum (y^*-y)^2 = \min$, ovvero di rendere minima la somma dei quadrati degli scarti (nel grafico che segue, dove ne sono stati presi due a caso come esempio, gli scarti sono quelli evidenziati in rosso) tra valori teorici e valori osservati.




Date due variabili X e Y, se la funzione teorica è lineare, cioè del tipo $Y^* = a + bX$ , la teoria dimostra che i parametri **a** e **b** determinati con il metodo dei minimi quadrati corrispondono alle seguenti espressioni:

$$b = \frac{n\sum xy-\sum x\sum y}{n\sum x^2-(\sum x)^2} \qquad a = \mu_y-b\mu_x$$

E si dimostra altresì che la retta dei minimi quadrati ha la caratteristica di passare per il baricentro della nuvola dei punti identificato dalle coordinate $(\mu_y, \mu_x)$, ovvero le medie delle distribuzioni dei due caratteri.

Esempio di determinazione dell'equazione canonica della retta di regressione di Y su X.

| x | y | xy | $x^2$ |
|---|---|---|---|
| 4 | 1 | 4 | 16 |
| 7 | 3 | 21 | 49 |
| 10 | 5 | 50 | 100 |
| 11 | 6 | 66 | 121 |
| 14 | 8 | 112 | 196 |
| 46 | 23 | 253 | 482 |

$\mu_x = 9,2$

$\mu_y = 4,6$

$$b = \frac{n\sum xy-\sum x\sum y}{n\sum x^2-(\sum x)^2} = \frac{5*253-46*23}{5*482-(46)^2} = \frac{207,0}{294,0} = \mathbf{0,7041}$$

oppure $$b = \frac{\sum xy-\sum x\sum y/n}{\sum x^2-(\sum x)^2/n} = \frac{253-(46*23)/5}{482-(46)^2/n} = \frac{253,0 - 211,6}{482,0 - 423,2} = \frac{41,4}{58,8} = \mathbf{0,7041}$$

$a = \mu_y-b\mu_x = 4,6-0,7041*9,2 = \mathbf{-1,8776}$

$Y^* = a + bX$ **$Y^* = -1,8776 + 0,7041X$**

$Y^* = -1,8776 + 0,7041X = -1,8776 + 0,7041*9,2 = 4,6$

Ulteriori espressioni per il calcolo del parametro b.

| x | y | xy | $x^2$ |
|---|---|---|---|
| 4 | 1 | 4 | 16 |
| 7 | 3 | 21 | 49 |
| 10 | 5 | 50 | 100 |
| 11 | 6 | 66 | 121 |
| 14 | 8 | 112 | 196 |
| 46 | 23 | 253 | 482 |

$\mu_x = 9,2$

$\mu_y = 4,6$

| (X-μ) | (Y-μ) | (X-μ)*(Y-μ) | $(X-\mu)^2$ |
|---|---|---|---|
| -5,2 | -3,6 | 18,72 | 27,04 |
| -2,2 | -1,6 | 3,52 | 4,84 |
| 0,8 | 0,4 | 0,32 | 0,64 |
| 1,8 | 1,4 | 2,52 | 3,24 |
| 4,8 | 3,4 | 16,32 | 23,04 |
| 0,0 | 0,0 | 41,4 | 58,8 |

$$\mathbf{b} = \frac{\sum(X-\mu)*(Y-\mu)}{\sum(X-\mu)^2} = \frac{41,4000}{58,8000} = \mathbf{0,7041}$$

$$\mathbf{b} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \frac{\sum(XY)/n-(\mu_x*\mu_y)}{\sum X^2/n-(\mu_x)^2} = \frac{253 \;/\; 5 - 9,20 * 4,60}{482 \;/\; 5 - 84,64} = \frac{8,280}{11,760} = \mathbf{0,7041}$$

Una volta scelta la funzione da adattare alla distribuzione empirica e i relativi parametri, l'esigenza che si pone è quella di valutare il **grado di affidabilità del modello.** Si rende opportuno cioè misurare la **dispersione** dei dati osservati intorno alla retta prescelta.

Tra i diversi indici elaborati assume particolare rilievo l'**indice di determinazione lineare**. Si tratta di un indice della **bontà di accostamento** della retta di regressione alla nuvola di punti osservati.

In simboli $$R^2 = 1 - \frac{\sum (y- y^* )^2}{\sum (y- \mu_y )^2} \qquad 0 < R^2 < 1$$

L'indice di determinazione lineare è in grado di fornire la forza della relazione rappresentata dalla retta di regressione. Se vale 0 significa che la variabilità dei valori di Y non risulta spiegata dalla regressione. Quando vale 1 tutti i punti sperimentali giacciono sulla retta di regressione, per cui la regressione spiega una gran parte della variabilità dei valori di Y e quindi il modello di regressione è appropriato per descrivere l'associazione tra le variabili.

Esempio di calcolo dell'indice di determinazione $R^2$

| x | y |
|---|---|
| 0 | 800 |
| 1 | 980 |
| 2 | 1.040 |
| 3 | 1.200 |
| 4 | 1.240 |
| 5 | 1.550 |
| | 6.810 |

$y^*=800+134x$

$\mu_y = 1.135$

| $y^*$ | $y-y^*$ | $(y-y^*)^2$ | $(y-\mu_y)^2$ |
|---|---|---|---|
| 800 | 0 | 0 | 112.225 |
| 934 | 46 | 2.116 | 24.025 |
| 1.068 | -28 | 784 | 9.025 |
| 1.202 | -2 | 4 | 4.225 |
| 1.336 | -96 | 9.216 | 11.025 |
| 1.470 | 80 | 6.400 | 172.225 |
| 6.810 | 0 | 18.520 | 332.750 |

$$R^2 = 1 - \frac{\sum (y- y^* )^2}{\sum (y- \mu_y )^2} = 1 - \frac{18.520}{332.750} = 1 - 0,0557 = 0,9443$$

Il risultato evidenzia un ottimo accostamento.

Va altresì sottolineato che nel metodo dei minimi quadrati applicato al modello di regressione lineare semplice, la somma dei dati osservati è sempre uguale a quella dei dati teorici.

Nell'analisi statistica di una distribuzione doppia di caratteri entrambi quantitativi, una trattazione a parte è dedicata allo studio di una particolare relazione: l'**interdipendenza**.

Per misurare la correlazione tra due variabili è necessario fare riferimento alla **covarianza**, la cui espressione è:

$$Cov(X,Y) = \sigma_{xy} = \frac{\sum(x - \mu_x)(y - \mu_y)}{n}$$

La **covarianza** è una misura della contemporanea variazione di due caratteri X e Y, che oltre a descrivere la dispersione delle variabili, esprime anche la relazione tra loro.

Il suo segno, a differenza di quello della varianza che è sempre positivo, può essere positivo o negativo, a seconda che la relazione tra le due variabili sia, rispettivamente, **diretta** (ci sia cioè **concordanza**), o **inversa** (se vi è **discordanza**). Il numeratore della covarianza, indicato con Cod(X,Y), è denominato **codevianza**.

La **covarianza** costituisce il numeratore di un'importante misura del grado di dipendenza lineare tra le due variabili: il **coefficiente di correlazione lineare di Bravais-Pearson**, la cui espressione è:

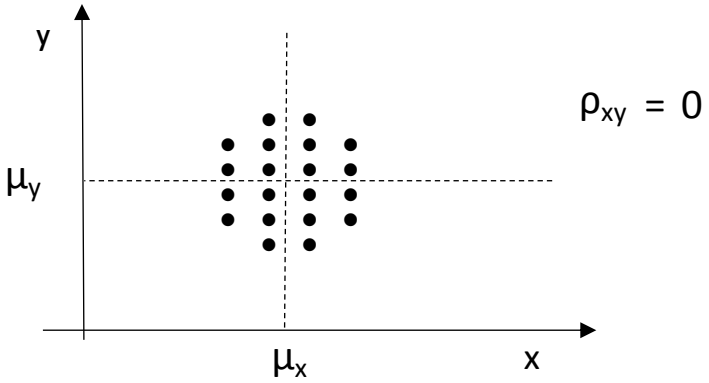
$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum(x - \mu_x)(y - \mu_y)}{\sqrt{\sum(x - \mu_x)^2} * \sqrt{\sum(y - \mu_y)^2}}$$

dove $\sigma_x \sigma_y$ sono lo scarto quadratico medio, rispettivamente della variabile X e della variabile Y.

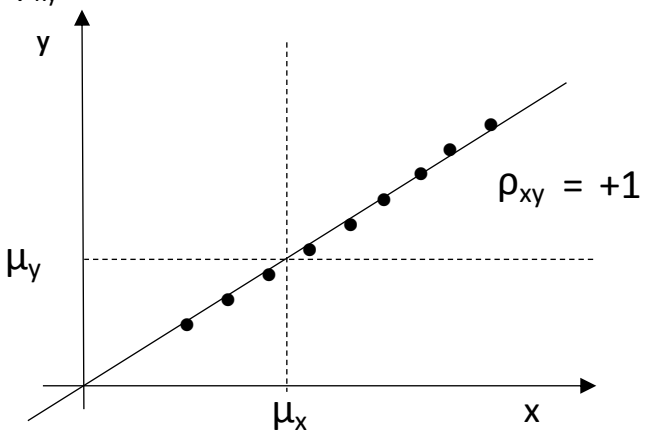
Il **coefficiente di correlazione** assume valori compresi tra $-1$ e $+1$ :

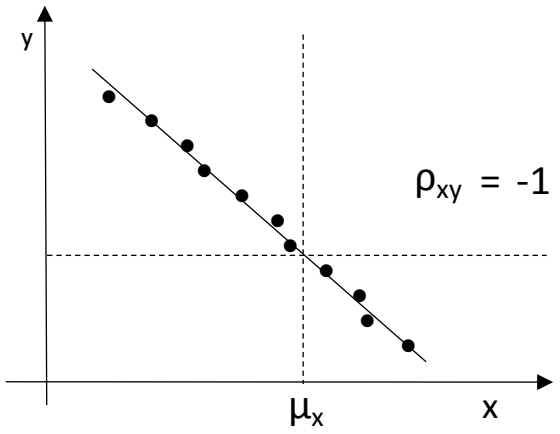
$$-1 \leq \rho_{xy} \leq +1$$

- se $\rho_{xy} = 0$ non vi è relazione di tipo lineare tra i due caratteri (sono **linearmente incorrelati)**.




- se $\rho_{xy} = \pm 1$ esiste, tra i due caratteri, un **legame lineare perfetto di tipo concorde** ($\rho_{xy} = +1$) o **discorde** ($\rho_{xy} = -1$).




- talvolta $\rho_{xy}$ può assumere un valore elevato pur non sussistendo alcuna relazione tra le variabili, ma per l'influenza esercitata sulle stesse da uno o più fattori comuni, in tal caso si dice che esiste una **correlazione spuria**.

Il **coefficiente di correlazione lineare di Bravais-Pearson**, può essere espresso anche nel seguente modo:

$$\rho_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} * \sqrt{n\sum y^2 - (\sum y)^2}}$$

Riassumendo: il **coefficiente di correlazione lineare di Bravais-Pearson $\rho_{xy}$** è un indice della linearità della relazione fra le variabili X e Y. Valori di $\rho_{xy}$ vicini a +1 o -1 indicano un'elevata linearità della relazione, quindi l'interpolazione lineare fornisce un'ottima approssimazione. Viceversa, valori di $\rho_{xy}$ vicini allo 0 indicano indipendenza tra X e Y oppure una relazione non lineare. Inoltre, come si è visto, se il coefficiente è positivo, Y tende ad aumentare con X e l'inclinazione della retta dei minimi quadrati è positiva, mentre se il coefficiente è negativo Y tende a diminuire all'aumentare di X e l'inclinazione della retta dei minimi quadrati è negativa.

Si dimostra che **coefficiente di correlazione lineare di Bravais-Pearson**, è pari alla radice quadrata dell'indice di determanazione lineare, in simboli:

$$\rho_{xy} = \pm \sqrt{R^2}$$

Poiché $\rho_{xy}$ assume valori fra +1 e -1 e $R^2$ assume valori fra 0 e +1, tanto più $R^2$ è prossimo a +1, tanto migliore sarà la rappresentazione di Y tramite le retta di regressione. E' evidente che se $R^2 = 1$ ( cioè $\rho_{xy} = \pm 1$), allora Y è linearmente dipendente da X ed esiste una regressione lineare perfetta ( o correlazione lineare perfetta). La retta di regressione è quindi in grado di rappresentare perfettamente Y.

Esempio di calcolo dell'indice di determinazione $R^2$ e di $\rho_{xy}$

| x | y | y* | y-y* | $(y-y^*)^2$ | $(y-\mu_y)^2$ | $x^2$ | $y^2$ | xy |
|---|---|---|---|---|---|---|---|---|
| 0 | 800 | 800 | 0 | 0 | 112.225 | 0 | 640.000 | 0 |
| 1 | 980 | 934 | 46 | 2.116 | 24.025 | 1 | 960.400 | 980 |
| 2 | 1.040 | 1.068 | -28 | 784 | 9.025 | 4 | 1.081.600 | 2.080 |
| 3 | 1.200 | 1.202 | -2 | 4 | 4.225 | 9 | 1.440.000 | 3.600 |
| 4 | 1.240 | 1.336 | -96 | 9.216 | 11.025 | 16 | 1.537.600 | 4.960 |
| 5 | 1.550 | 1.470 | 80 | 6.400 | 172.225 | 25 | 2.402.500 | 7.750 |
| 15 | 6.810 | 6.810 | 0 | 18.520 | 332.750 | 55 | 8.062.100 | 19.370 |

$\mu_y = 1.135$

$y^* = 800 + 134x$

$$R^2 = 1 - \frac{\sum (y - y^*)^2}{\sum (y - \mu_y)^2} = 1 - \frac{18.520}{332.750} = 1 - 0{,}0557 = 0{,}9443$$

$$\rho_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} * \sqrt{n\sum y^2 - (\sum y)^2}} = \frac{6*19.370 - 15*6.810}{\sqrt{6*55 - (15)^2} * \sqrt{6*8.062.100 - (6.810)^2}} =$$

$$= \frac{116.220 - 102.150}{\sqrt{330 - 225} * \sqrt{48.372.600 - 46.376.100}} = \frac{14.070}{14.478{,}691} = 0{,}9718$$

Il risultato fa rilevare una correlazione diretta tra le due variabili.

Si dimostra anche che:

$$\rho_{xy} = \pm \sqrt{R^2} = \pm \sqrt{0{,}9443} = 0{,}9718$$

Altri esempi di calcolo dell'indice di determinazione di $\rho_{xy}$

| x | y | $x^2$ | $y^2$ | xy |
|---|---|---|---|---|
| 0 | 0 | 0 | - | 0 |
| 1 | 20 | 1 | 400 | 20 |
| 2 | 40 | 4 | 1.600 | 80 |
| 3 | 60 | 9 | 3.600 | 180 |
| 4 | 80 | 16 | 6.400 | 320 |
| 5 | 100 | 25 | 10.000 | 500 |
| 15 | 300 | 55 | 22.000 | 1.100 |

$$\rho_{xy}= \frac{n\sum xy - \sum x\sum y}{\sqrt{n\sum x^2 - (\sum x)^2} * \sqrt{n\sum y^2 - (\sum y)^2}} = \frac{6*1.100 - 15*300}{\sqrt{6*55 - (15)^2} * \sqrt{6*22.000 - (300)^2}} =$$

$$= \frac{6.600 - 4.500}{\sqrt{330 - 225} * \sqrt{132.000 - 90.000}} = \frac{2.100}{2.100} = \mathbf{1,000}$$

| x | y | $x^2$ | $y^2$ | xy |
|---|---|---|---|---|
| 0 | 250 | 0 | 62.500 | 0 |
| 1 | 800 | 1 | 640.000 | 800 |
| 2 | 5 | 4 | 25 | 10 |
| 3 | 0 | 9 | 0 | 0 |
| 4 | 60 | 16 | 3.600 | 240 |
| 5 | 0 | 25 | 0 | 0 |
| 15 | 1.115 | 55 | 706.125 | 1.050 |

$$\rho_{xy}= \frac{n\sum xy - \sum x\sum y}{\sqrt{n\sum x^2 - (\sum x)^2} * \sqrt{n\sum y^2 - (\sum y)^2}} = \frac{6 * 1.050 - 15 * 1.115}{\sqrt{6 * 55 - 225} * \sqrt{6 * 706.125 - 1.243.225}}$$

$$= \frac{6.300 - 16.725}{\sqrt{330 - 225} * \sqrt{4.236.750 - 1.243.225}} = \frac{-10.425}{17.729,1} = -0,588$$