# ANALISI DELLA DIPENDENZA

## La correlazione



## ANALISI DELLA DIPENDENZA

- Supponiamo di aver rilevato su un insieme di soggetti due variabili quantitative, ad es. altezza e peso
- Si parla di dipendenza tra due variabili quando è ipotizzabile una relazione, di qualsiasi tipo, tra i valori (più in generale le modalità) assunti dalle due variabili:
  - la relazione può essere di tipo **causale**, cioè una variazione del valore assunto da una delle due variabili provoca una conseguente modificazione del valore dell'altra variabile

$$X \rightarrow Y$$

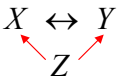
  - oppure non è identificabile una causa e un effetto, ma si osserva solo una associazione tra i comportamenti cioè i valori assunti dalle due variabili

$$X \leftrightarrow Y$$

- In assenza di qualunque relazione, due variabili si dicono **indipendenti**
- E' necessario sottolineare che una relazione di causa-effetto non è un dato di fatto, ma una *ipotesi*: l'attribuzione di causalità ad una relazione è una operazione concettuale che va al di la dei dati osservati, e anche dell'analisi statistica:
  - la causalità è una *categoria* del pensiero umano: *non si osserva …*
  - la possibilità di fare una affermazione di tipo causale che abbia una validità scientifica dipende essenzialmente dal metodo di ricerca (sperimentale) e dal grado di controllo esercitato sulle variabili esplicative e di disturbo

# ANALISI DELLA DIPENDENZA

- L'attribuzione di un nesso causale è comunque sempre una nostra proiezione: a volte ci risulta naturale (a torto o a ragione), altre volte molto problematica. Non è raro confondere la causa con l'effetto.
- Di per se stessi, i dati osservati ci permettono di rilevare solo una *associazione* o *correlazione,* più o meno forte, tra fenomeni, comportamenti, eventi
- Osservare una forte correlazione tra due variabili rappresenta un indizio, ma non significa che la prima variabile abbia un effetto causale sulla seconda: cioè che aumentandone il valore (intensità) si possa indurre un aumento (o diminuzione) nel valore assunto dall'altra
- Molte volte due variabili sono strettamente associate, cioè presentano andamenti molto simili, ma nessuno si sognerebbe di sostenere che esiste una relazione causale, né in una direzione e nemmeno nell'altra
- Altre volte invece avviene che, sulla base di una correlazione, si ritenga di avere "scoperto" la causa di qualcosa (es. HIV -> AIDS)
- Una forte associazione tra due variabili può essere dovuta ad es. a un terzo fattore, che agisce su entrambe:




O essere il risultato di una ben più complessa catena di relazioni e interazioni tra molti fattori …

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

# ANALISI DELLA DIPENDENZA

- Per poter essere di tipo causale, una relazione osservata deve avere due requisiti:
  - una direzione determinata, dedotta di solito dalla sequenza temporale
  - forte associazione tra presunto fattore causale e risposta/effetto
- Queste condizioni sono necessarie ma non sufficienti: ci possono far ipotizzare una relazione causale, ma non sono sufficienti a "dimostrarla"…
  Cosa vuol dire "dimostrato", "provato" scientificamente ?
- Per poter sostenere validamente una ipotesi di causa-effetto, è idealmente necessario garantire:
  - controllo diretto del presunto fattore causale
  - assenza di fattori di confusione: che possono compromettere la correttezza dell'inferenza

  => **metodo sperimentale**
- Se una ricerca che ha misurato una associazione tra due variabili non è stata condotta con metodo sperimentale, cioè non si è realizzata:
  - controllando l'assegnazione del presunto fattore causale (prima): randomizzazione -> gruppo/condizione di controllo
  - misurando poi le eventuali variazioni della variabile dipendente

  l'associazione, per quanto forte, non consente una interpretazione in senso causale

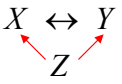
# ANALISI DELLA DIPENDENZA

- Sbagliare nell'identificazione della causa, o addirittura confondere la causa con l'effetto, non è così raro. La storia della scienza è ricca di esempi di teorie che hanno dominato la scena per molto tempo, e che oggi ci sembrano incredibili o addirittura ridicole
- L'ipotesi di causalità deve essere sottoposta a verifica empirica (indefinitamente): se le previsioni della teoria vengono contraddette dall'osservazione, l'ipotesi risulta falsificata e dovrebbe essere abbandonata
- Non sempre una ipotesi falsificata viene subito abbandonata dalla comunità scientifica:
  - occorre avere una ipotesi alternativa, possibilmente migliore
  - il consenso, anche all'interno della comunità scientifica, è una dinamica sociale
  - sono talvolta in gioco anche interessi economici, politici, ...
- Solo molto tempo dopo che l'ipotesi è stata falsificata, e solo dopo che la comunità scientifica l'ha definitivamente abbandonata, essa *diventa* ai nostri occhi così assurda o ridicola ...



# ANALISI DELLA DIPENDENZA

- Se ipotizziamo una relazione di causa-effetto, il valore assunto dalla variabile X (v. indipendente , v. esplicativa, fattore, trattamento) può essere utilizzato per prevedere quello della Y (v. dipendente , risposta, esito, effetto)
- Può anche darsi il caso che una variabile possa essere utilizzata per prevedere il livello assunto da un'altra, pur non essendone la causa, ma solo un *indicatore*, come nel caso:

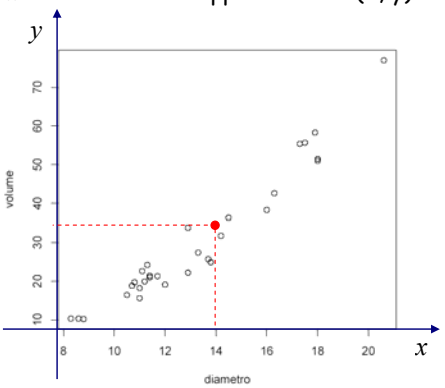



- In effetti, è *sempre* così:
  noi non sappiamo mai con certezza se la nostra ipotesi di causalità è vera...
- Comunque sulla base della nostra ipotesi costruiamo un modello: finché il modello funziona, cioè dà buoni risultati per il campo di applicazione in cui viene impiegato (es. previsioni corrette, terapie efficaci, ...), noi confidiamo nell'ipotesi, cioè la riteniamo valida e confermata, e la utilizziamo (ci comportiamo) come se fosse vera.
- La previsione di una variabile dipendente Y in funzione di un insieme di predittori X viene condotta costruendo un *modello* (es. matematico), cioè una rappresentazione in grado di descrivere il fenomeno, ovvero la relazione osservata:
  - modelli di regressione (lineare/non lineare, generalizzato, ...)
  - alberi di classificazione/regressione (segmentazione)
  - reti neurali artificiali

# ANALISI DELLA DIPENDENZA

- **Covariazione o Correlazione**
- Nel caso di variabili quantitative, si definisce un particolare tipo di associazione detta **covariazione** o **correlazione**: quando in corrispondenza di valori alti di una variabile si osservano prevalentemente valori alti (o bassi) dell'altra e viceversa, si dice che esiste una correlazione (positiva o negativa) tra le variabili.
- Quando si vuole analizzare la relazione tra due variabili quantitative, è sempre opportuno partire dalla rappresentazione grafica della situazione: a partire dalla serie doppia dei dati osservati, si costruisce un grafico rappresentando i dati come punti sul piano cartesiano, aventi come coordinate la coppia di valori (x, y)

| diametro | volume | diametro | volume | diametro | volume |
|---|---|---|---|---|---|
| 8,3 | 10,3 | 11,3 | 24,2 | 14,0 | 34,5 |
| 8,6 | 10,3 | 11,4 | 21,0 | 14,2 | 31,7 |
| 8,8 | 10,2 | 11,4 | 21,4 | 14,5 | 36,3 |
| 10,5 | 16,4 | 11,7 | 21,3 | 16,0 | 38,3 |
| 10,7 | 18,8 | 12,0 | 19,1 | 16,3 | 42,6 |
| 10,8 | 19,7 | 12,9 | 22,2 | 17,3 | 55,4 |
| 11,0 | 15,6 | 12,9 | 33,8 | 17,5 | 55,7 |
| 11,0 | 18,2 | 13,3 | 27,4 | 17,9 | 58,3 |
| 11,1 | 22,6 | 13,7 | 25,7 | 18,0 | 51,5 |
| 11,2 | 19,9 | 13,8 | 24,9 | 18,0 | 51,0 |
| 20,6 | 77,0 | | | | |

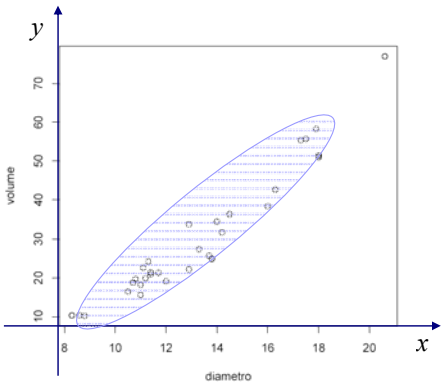



- Dalla rappresentazione grafica si può cominciare a capire il tipo di relazione



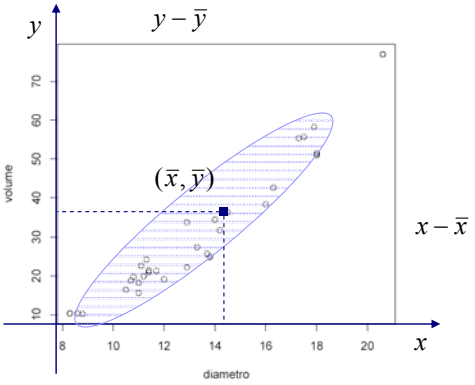
# ANALISI DELLA DIPENDENZA

- L'insieme dei dati osservati è rappresentato sul piano cartesiano come una *nuvola di punti*
- Nell'esempio a fianco, il grafico mostra una correlazione *positiva* (o diretta) tra le due variabili:
  - a valori elevati della X si associano prevalentemente valori alti della Y ...
  - per brevità, si dice che al crescere di X cresce Y
- La relazione sembra piuttosto *forte* : la nuvola di punti è abbastanza "stretta", cioè i punti sono piuttosto concentrati lungo una linea ideale, e non molto dispersi sul piano
- Quanto è forte la relazione ? Vogliamo quantificare la forza della relazione ovvero arrivare ad una misura della dipendenza tra le due variabili




| diametro | volume | diametro | volume | diametro | volume |
|---|---|---|---|---|---|
| 8,3 | 10,3 | 11,3 | 24,2 | 14,0 | 34,5 |
| 8,6 | 10,3 | 11,4 | 21,0 | 14,2 | 31,7 |
| 8,8 | 10,2 | 11,4 | 21,4 | 14,5 | 36,3 |
| 10,5 | 16,4 | 11,7 | 21,3 | 16,0 | 38,3 |
| 10,7 | 18,8 | 12,0 | 19,1 | 16,3 | 42,6 |
| 10,8 | 19,7 | 12,9 | 22,2 | 17,3 | 55,4 |
| 11,0 | 15,6 | 12,9 | 33,8 | 17,5 | 55,7 |
| 11,0 | 18,2 | 13,3 | 27,4 | 17,9 | 58,3 |
| 11,1 | 22,6 | 13,7 | 25,7 | 18,0 | 51,5 |
| 11,2 | 19,9 | 13,8 | 24,9 | 18,0 | 51,0 |
| 20,6 | 77,0 | | | | |

# ANALISI DELLA DIPENDENZA

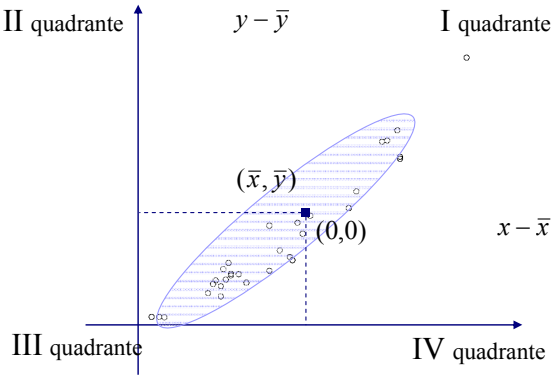



- Operiamo la doppia trasformazione di variabili: $\begin{cases} x' = x - \bar{x} \\ y' = y - \bar{y} \end{cases}$
- E' equivalente a traslare gli assi cartesiani nel baricentro della nuvola di punti
- Nel nuovo sistema di coordinate traslato, ciascun punto che prima aveva coordinate (x, y), ora avrà coordinate: $(x - \bar{x}, y - \bar{y})$



# ANALISI DELLA DIPENDENZA

$\begin{cases} x' = x - \bar{x} \\ y' = y - \bar{y} \end{cases}$




- Osserviamo che quando i punti si trovano nel:
  - I quadrante: entrambe le coordinate sono positive
  - III quadrante: entrambe le coordinate sono negative

  } il prodotto è positivo
  - II quadrante: ascissa negativa e ordinata positiva
  - IV quadrante: ascissa positiva e ordinata negativa

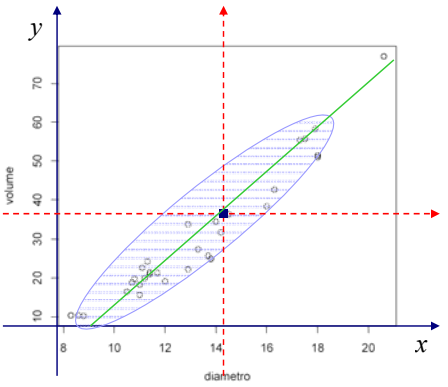
  } il prodotto è negativo
- Sommando i prodotti delle coordinate (trasformate) per tutti i punti, abbiamo un indicatore di come i punti si distribuiscono nel piano = un indice di covariazione

# ANALISI DELLA DIPENDENZA

- **La Covarianza**
  E' la media dei prodotti degli scarti di ciascuna variabile dalla propria media

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

- E' un indice di covariazione: il suo segno descrive l'andamento della relazione:
  - è > 0 quando la relazione è positiva
  - è < 0 quando è negativa
  - è = 0 quando non c'è correlazione
- La covarianza cresce in valore assoluto quanto più i punti tendono ad allinearsi lungo una retta, e diminuisce invece fino al suo minimo (0) quanto più se ne allontanano
- Non è una misura standardizzata: dipende dalle unità di misura delle variabili, e dalle variabilità delle due serie di dati x e y




| diametro | volume | diametro | volume | diametro | volume |
|---|---|---|---|---|---|
| 8,3 | 10,3 | 11,3 | 24,2 | 14,0 | 34,5 |
| 8,6 | 10,3 | 11,4 | 21,0 | 14,2 | 31,7 |
| 8,8 | 10,2 | 11,4 | 21,4 | 14,5 | 36,3 |
| 10,5 | 16,4 | 11,7 | 21,3 | 16,0 | 38,3 |
| 10,7 | 18,8 | 12,0 | 19,1 | 16,3 | 42,6 |
| 10,8 | 19,7 | 12,9 | 22,2 | 17,3 | 55,4 |
| 11,0 | 15,6 | 12,9 | 33,8 | 17,5 | 55,7 |
| 11,0 | 18,2 | 13,3 | 27,4 | 17,9 | 58,3 |
| 11,1 | 22,6 | 13,7 | 25,7 | 18,0 | 51,5 |
| 11,2 | 19,9 | 13,8 | 24,9 | 18,0 | 51,0 |
| 20,6 | 77,0 | | | | |



# ANALISI DELLA DIPENDENZA

- **Metodo di calcolo indiretto della Covarianza**
- La covarianza è calcolabile anche come differenza tra la media del prodotto delle due variabili e il prodotto delle medie:

$$Cov(x,y) = M(x\,y) - M(x)M(y)$$

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

- Infatti:

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

$$= \frac{\sum x_i\,y_i - \sum x_i\,\bar{y} - \sum \bar{x}y_i + \sum \bar{x}\bar{y}}{n} = \frac{\sum x_i\,y_i}{n} - \bar{y}\frac{\sum x_i}{n} - \bar{x}\frac{\sum y_i}{n} + \bar{x}\,\bar{y} =$$

$$= \frac{\sum x_i\,y_i}{n} - \bar{x}\,\bar{y} - \bar{x}\,\bar{y} + \bar{x}\,\bar{y} = \frac{\sum x_i\,y_i}{n} - \bar{x}\,\bar{y}$$

# ANALISI DELLA DIPENDENZA

- **Proprietà della Covarianza**
- E' una quantità *simmetrica*, nel senso che: Cov (x, y) = Cov (y, x)
  La covariazione è una proprietà reciproca delle due variabili: è una proprietà della loro *relazione*
- Se una delle due variabili è costante, la covarianza è nulla, infatti se x(i) = k ∀i:

$$Cov(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum (k-k)(y_i - \bar{y})}{n} = \frac{\sum 0\,(y_i - \bar{y})}{n} = 0$$

- Se x e y sono *statisticamente indipendenti* (in distribuzione), la loro covarianza è nulla. La covarianza può però essere nulla anche senza che x e y siano indipendenti
- La covarianza è in effetti una misura della *linearità* della relazione tra due variabili: altre forme di relazione (non lineari) non vengono rilevate da questo indice
- La covarianza non è una misura standardizzata, dipende dalle unità di misura delle variabili, e dalle variabilità delle due serie di dati x e y: per superare questo difetto, viene introdotto un altro indice di correlazione, ricorrendo ancora una volta alla standardizzazione
- E' sempre inferiore o uguale, in valore assoluto, al prodotto degli scarti quadratici medi delle due variabili :

$$|Cov(x,y)| \leq \sigma(x)\,\sigma(y)$$

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

# ANALISI DELLA DIPENDENZA

- **Il coefficiente di correlazione (lineare)**
- Il coefficiente di correlazione lineare r(x,y) di Pearson è dato dal rapporto tra la covarianza e il prodotto degli scarti quadratici medi delle due variabili:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y}$$

- E' una misura dell'intensità della dipendenza (relazione lineare) tra le due variabili
- E' un indice relativizzato (o normalizzato): è costruito come rapporto tra la covarianza e il valore massimo che essa può assumere
  - è un numero puro (adimensionale) cioè senza unità di misura
  - è compreso tra -1 e +1
- Interpretazione del valore di r(x,y):
  - è > 0 quando la correlazione è positiva, < 0 quando è negativa
  - vale +1 o -1 quando la forza della correlazione lineare (positiva o negativa) è massima, cioè quando tutti i punti sono perfettamente allineati lungo una retta
  - vale 0 quando la correlazione è nulla: le due variabili si dicono **incorrelate**

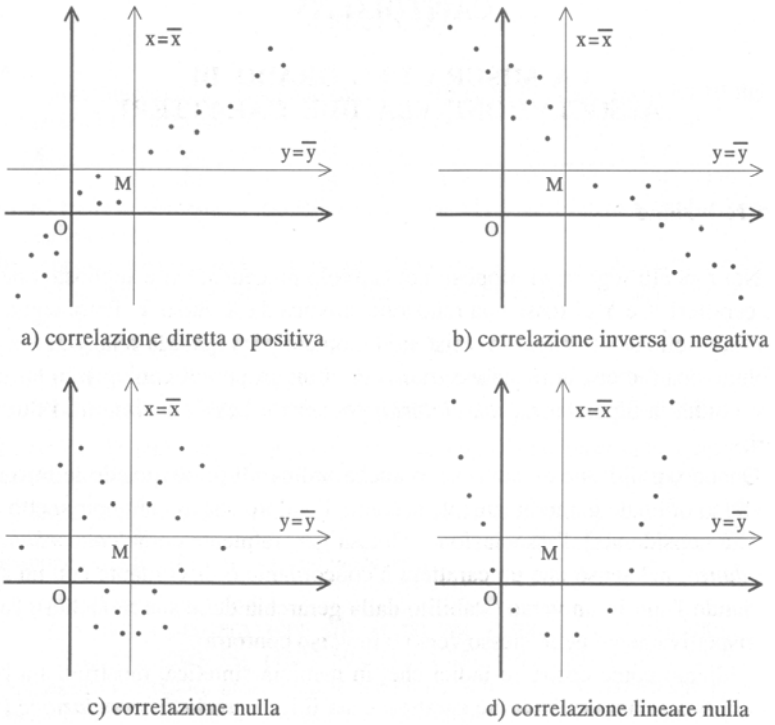
# ANALISI DELLA DIPENDENZA

- **Proprietà del coefficiente di correlazione lineare**
- E' una quantità simmetrica, nel senso che: r(x, y) = r(y, x)
  La co-relazione è reciproca
- Se x e y sono *indipendenti*, allora r è uguale a zero.
  Tuttavia r può essere uguale a zero anche senza che x e y siano indipendenti
- Il coefficiente di correlazione è una misura della relazione lineare tra due variabili: non mette in evidenza altre forme di relazione (parabolica, curvilinee, ...)
- Pertanto quando r è prossimo a zero può significare che non c'è relazione tra le due variabili, ma si potrebbe anche essere invece in presenza di una relazione di tipo non lineare
- Il coefficiente di correlazione lineare è una *covarianza standardizzata*,
  cioè è ottenibile come covarianza tra le due variabili x e y standardizzate:

$$Cov\left(\frac{x-\bar{x}}{\sigma_x},\frac{y-\bar{y}}{\sigma_y}\right)=\frac{\sum_i\left(\frac{x_i-\bar{x}}{\sigma_x}-0\right)\left(\frac{y_i-\bar{y}}{\sigma_y}-0\right)}{n}=\frac{\frac{1}{\sigma_x\sigma_y}\sum_i(x_i-\bar{x})(y_i-\bar{y})}{n}=$$

$$=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n\,\sigma_x\sigma_y}=\frac{\sigma_{xy}}{\sigma_x\sigma_y}=\rho_{xy}$$

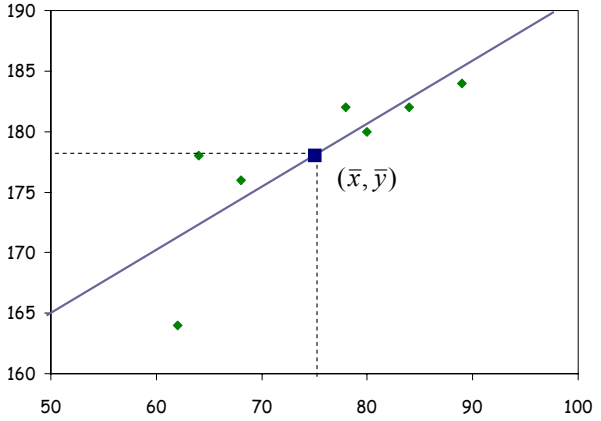




# ANALISI DELLA DIPENDENZA

- Esercizio.
  Studiamo la dipendenza tra le variabili quantitative Y=altezza e X=peso:

| i | x(i) | y(i) |
|---|---|---|
| 1 | 62 | 164 |
| 2 | 64 | 178 |
| 3 | 68 | 176 |
| 4 | 75 | 178 |
| 5 | 78 | 182 |
| 6 | 80 | 180 |
| 7 | 84 | 182 |
| 8 | 89 | 184 |
| Totale | 600 | 1424 |
| Media | 75,00 | 178,00 |




- Dal grafico si può notare una relazione positiva abbastanza forte:
  solo una osservazione è piuttosto distante dalla *linea ideale* intorno alla quale si concentrano i dati

Prof. Claudio Capiluppi - Facoltà di Scienze della Formazione - A.A. 2007/08

# ANALISI DELLA DIPENDENZA

- Prospetto di calcolo del coefficiente di correlazione:

| i | x(i) | y(i) | x(i)-Mx | y(i)-My | $[x(i)-Mx]^2$ | $[y(i)-My]^2$ | (x(i)-Mx)(y(i)-My) |
|---|---|---|---|---|---|---|---|
| 1 | 62 | 164 | -13,00 | -14,00 | 169,00 | 196,00 | 182,00 |
| 2 | 64 | 178 | -11,00 | 0,00 | 121,00 | 0,00 | 0,00 |
| 3 | 68 | 176 | -7,00 | -2,00 | 49,00 | 4,00 | 14,00 |
| 4 | 75 | 178 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| 5 | 78 | 182 | 3,00 | 4,00 | 9,00 | 16,00 | 12,00 |
| 6 | 80 | 180 | 5,00 | 2,00 | 25,00 | 4,00 | 10,00 |
| 7 | 84 | 182 | 9,00 | 4,00 | 81,00 | 16,00 | 36,00 |
| 8 | 89 | 184 | 14,00 | 6,00 | 196,00 | 36,00 | 84,00 |
| Totale | 600 | 1424 | 0 | 0 | 650,00 | 272,00 | 338,00 |
| Media | 75,00 | 178,00 | | | 81,25 | 34,00 | 42,25 |

$$\sigma_x^2 = 81{,}25 \quad \sigma_y^2 = 34 \quad \sigma_{xy} = 42{,}25$$

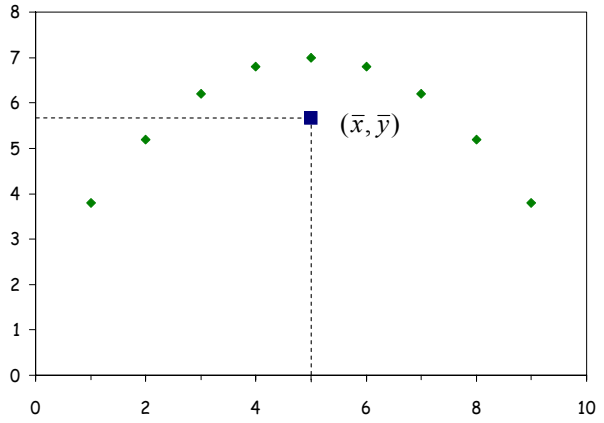
$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \, \sigma_y} = \frac{42{,}25}{\sqrt{81{,}25 \cdot 34}} = 0{,}8039$$

- r è positivo e piuttosto elevato, considerato che il suo massimo è 1 :
  quindi misuriamo una forte dipendenza lineare tra le due variabili

# ANALISI DELLA DIPENDENZA

- Esercizio.
  Studiamo la dipendenza tra due variabili X e Y, legate dalla relazione di dipendenza *esatta* : $Y = 2 + 2\,x - 0{,}2\,x^2$

| i | x(i) | y(i) |
|---|---|---|
| 1 | 1 | 3,8 |
| 2 | 2 | 5,2 |
| 3 | 3 | 6,2 |
| 4 | 4 | 6,8 |
| 5 | 5 | 7 |
| 6 | 6 | 6,8 |
| 7 | 7 | 6,2 |
| 8 | 8 | 5,2 |
| 9 | 9 | 3,8 |
| Totale | 45 | 51 |
| Media | 5,00 | 5,67 |




- Dal grafico si evidenzia chiaramente la relazione quadratica tra le due variabili, che disegna una parabola con la concavità rivolta verso il basso
- Vediamo cosa ci dice il coefficiente di correlazione ...



# ANALISI DELLA DIPENDENZA

- Prospetto di calcolo: questa volta proviamo il metodo indiretto, molto più veloce :

$$\sigma_x^2 = M(x^2) - \bar{x}^2 =$$
$$= 31{,}67 - 5^2 = 6{,}67$$

$$\sigma_y^2 = 33{,}48 - 5{,}67^2 = 1{,}33$$

$$\sigma_{xy} = M(x\,y) - \bar{x}\,\bar{y} =$$
$$= 28{,}33 - 5 \cdot 5{,}67 \cong 0$$

| i | x(i) | y(i) | $[x(i)]^2$ | $[y(i)]^2$ | x(i) y(i) |
|---|---|---|---|---|---|
| 1 | 1 | 3,8 | 1,00 | 14,44 | 3,80 |
| 2 | 2 | 5,2 | 4,00 | 27,04 | 10,40 |
| 3 | 3 | 6,2 | 9,00 | 38,44 | 18,60 |
| 4 | 4 | 6,8 | 16,00 | 46,24 | 27,20 |
| 5 | 5 | 7 | 25,00 | 49,00 | 35,00 |
| 6 | 6 | 6,8 | 36,00 | 46,24 | 40,80 |
| 7 | 7 | 6,2 | 49,00 | 38,44 | 43,40 |
| 8 | 8 | 5,2 | 64,00 | 27,04 | 41,60 |
| 9 | 9 | 3,8 | 81,00 | 14,44 | 34,20 |
| Totale | 45 | 51 | 285 | 301,32 | 255 |
| Media | 5,00 | 5,67 | 31,67 | 33,48 | 28,33 |

- A meno di errori di arrotondamento, r risulta praticamente pari a zero, segnalando assenza di relazione *lineare*
- Le due variabili sono invece in strettissima relazione, essendo y, per costruzione, funzione (quadratica) esatta di x, ma:
- **il coefficiente di correlazione non è in grado di rilevare relazioni non lineari**