Da un lavoro della prof.ssa Donatella Martini
IIS Nullo Baldoni di Ravenna

La correlazione tra due variabili

Scuola o sport?

In una scuola ci sono diversi studenti che praticano sport a livello agonistico e dedicano molte ore agli allenamenti, trascurando un po' le attività di studio nelle diverse materie. Si suppone che esista una dipendenza tra i risultati scolastici e il tempo dedicato alle attività sportive, come individuarla?

Soluzione

Organizzazione dei dati

Prendiamo un campione di 20 studenti e supponiamo che i risultati medi di fine anno scolastico e il numero di ore settimanali che essi dedicano alla loro attività sportiva siano i seguenti:

studenti	media	ore di allenamento
1	6.5	15
2	5.8	18
3	7	10
4	5.5	20
5	6	12
6	6	15
7	6.3	12
8	6.2	12
9	6.8	18
10	5.7	20
11	8	6
12	6.2	15
13	7.3	8
14	8.5	3
15	5.2	20
16	5.7	15
17	5	20
18	7.8	8
19	6	12
20	6.3	12

Per poterli utilizzare, inseriamo i dati in una matrice:

restart
voti := Matrix(⟨⟨6.5⟩, ⟨5.8⟩, ⟨7⟩, ⟨5.5⟩, ⟨6⟩, ⟨6⟩, ⟨6.3⟩, ⟨6.2⟩, ⟨6.8⟩, ⟨5.7⟩, ⟨8⟩, ⟨6.2⟩, ⟨7.3⟩, ⟨8.5⟩, ⟨5.2⟩, ⟨5.7⟩, ⟨5⟩, ⟨7.8⟩, ⟨6⟩, ⟨6.3⟩⟩) :

ore := *Matrix*(⟨⟨15⟩, ⟨18⟩, ⟨10⟩, ⟨20⟩, ⟨12⟩, ⟨15⟩, ⟨12⟩, ⟨12⟩, ⟨18⟩, ⟨20⟩, ⟨6⟩, ⟨15⟩, ⟨8⟩, ⟨3⟩, ⟨20⟩, ⟨15⟩, ⟨20⟩, ⟨8⟩, ⟨12⟩, ⟨12⟩⟩) :

Calcolo delle medie aritmetiche

Calcoliamo le medie aritmetiche delle due serie di valori.
La media dei voti sarà:

$mediavoti := Statistics[Mean](voti)[1]$

$$6.39000000000000 \quad \textbf{(1.2.1)}$$

e la media delle ore di allenamento:

$mediaore := Statistics[Mean](ore)[1]$

$$13.5500000000000 \quad \textbf{(1.2.2)}$$

Osservazioni

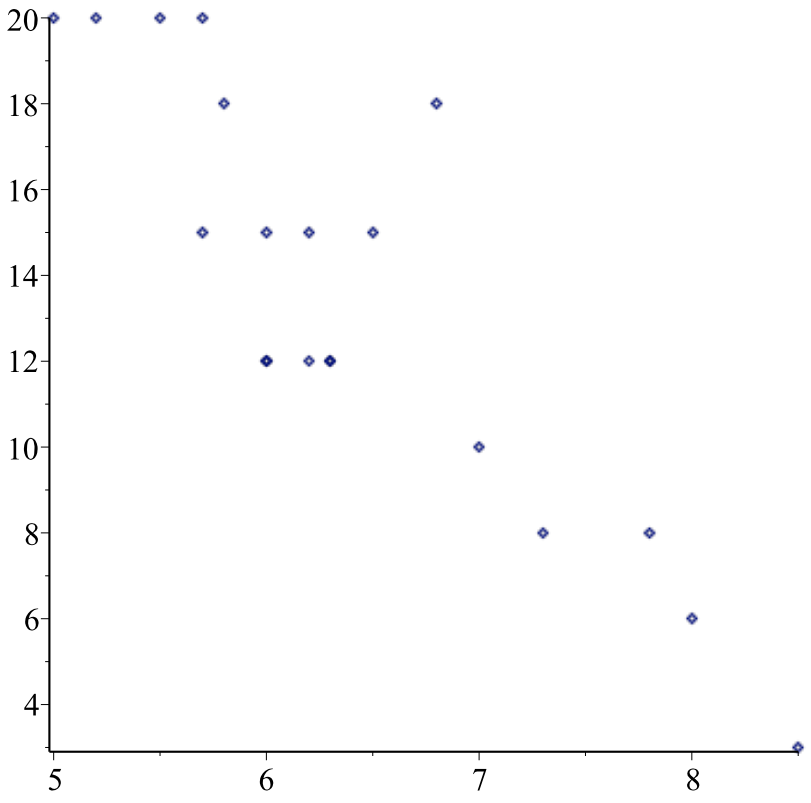
Possiamo dai valori delle due medie dedurre se c'è dipendenza tra le due serie di dati? Certamente no, quindi occorre trovare un altro mezzo per avere un'idea del legame . Proviamo a rappresentarli graficamente. Quale rappresentazione può essere utile in questo caso?

Rappresentazione grafica

La rappresentazione grafica che può essere più indicata in questo caso è il diagramma a dispersione.
Tracciamo allora il grafico a dispersione delle due serie di valori.

Grafico

$Statistics[ScatterPlot](voti, ore)$



Osservazioni

Osserviamo la forma della "nuvola" di punti ottenuta: questo ci permette di ipotizzare un certo tipo di dipendenza tra le due serie di valori.
Nel nostro caso la forma è allungata e quindi si può pensare che segua la linea di una retta.
Osserviamo allora anche la direzione in cui si sviluppa l'allungamento (in direzione crescente o decrescente): sembra decrescente.
Il grafico ancora non chiarisce molto sulla dipendenza, può darci solo un'idea.
Procediamo, allora, inserendo nel grafico le medie aritmetiche precedentemente calcolate.

Grafico con indicazione delle medie

Le osservazioni fatte sul grafico possono essere meglio precisate se tracciamo anche le rette relative alle medie.
I valori dei voti degli studenti sono riportati sull'asse x. Riportiamo su questo asse anche il valore della media dei voti e tracciamo la retta parallela all'asse y passante per il punto trovato (*x= mediavoti)*. Analogamente, sull'asse y sono riportate le ore di allenamento, quindi tracciamo la retta parallella all'asse x di equazione *y=mediaore.*

Grafico

$d1 := Statistics[ScatterPlot](voti, ore)$

$$PLOT(\dots) \qquad \textbf{(1.4.1.1)}$$

$d2 := plots[implicitplot](x = mediavoti, x = 3..10, y = 2..25)$

$$PLOT(\dots) \qquad \textbf{(1.4.1.2)}$$

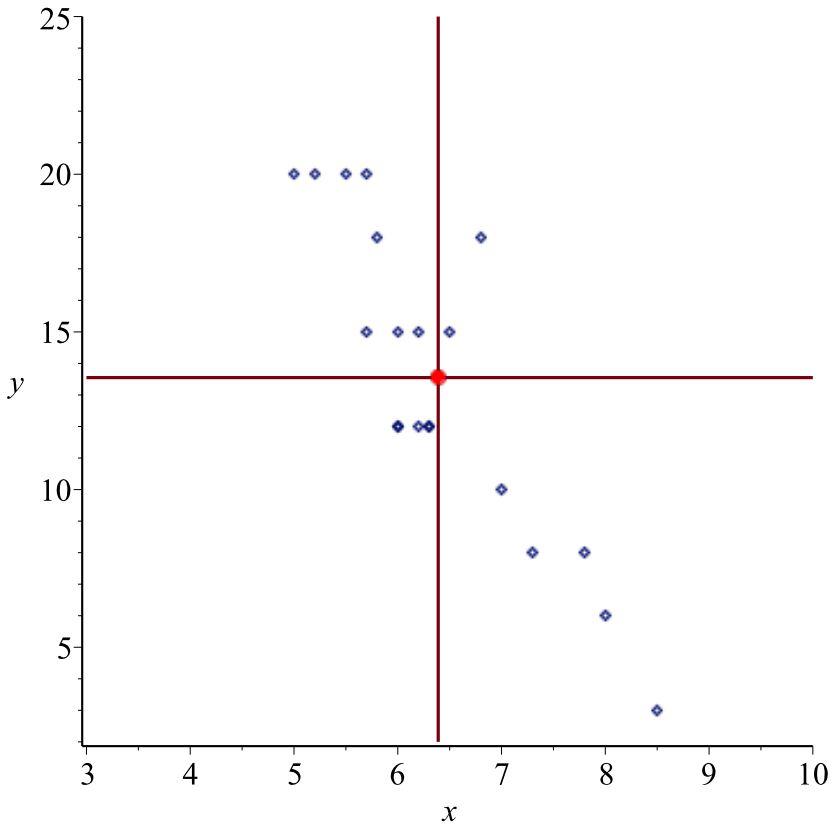
$d3 := plots[implicitplot](y = mediaore, x = 3..10, y = 2..25)$

$$PLOT(\dots) \qquad \textbf{(1.4.1.3)}$$

$d4 := plots[pointplot]([mediavoti, mediaore], color = red, symbol = solidcircle, symbolsize = 15)$

$$PLOT(\dots) \qquad \textbf{(1.4.1.4)}$$

$plots[display](d1, d2, d3, d4)$



Osservazioni

Il piano verrà diviso da queste due rette in quattro quadranti. Il punto di intersezione tra le due rette è detto **baricentro della distribuzione** ed ha come come ascissa il valore medio della variabile X (voti) e come ordinata il valore medio della variabile Y (ore).
Osserviamo la suddivisione dei punti nei quattro quadranti.

- In quale quadrante compare il maggior numero di punti?
- Che cosa possiamo dire del rapporto tra voti e ore di allenamento per gli studenti i cui dati sono riportati in questo quadrante?
- In quale quadrante compare il minor numero di punti?
- Che cosa possiamo dire del rapporto tra voti e ore di allenamento per gli studenti i cui dati sono riportati in questo quadrante?

Calcolo degli indici

Passiamo alla formalizzazione del problema con la definizione e il calcolo degli indici di correlazione.

La covarianza

Osservazione 1

Facciamo riferimento al grafico appena tracciato. Le due rette, corrispondenti alle medie delle due serie di valori, dividono il piano in quattro quadranti.
Numeriamoli in senso antiorario, come vuole la convenzione, e per semplicità chiamiamo x_i i voti riportati in asse x e con $\overline{x}$ la media dei voti, mentre indicheremo con y_i le ore riportate sull'asse y e con $\overline{y}$ la loro media.
Osserviamo le differenze tra i valori dei voti e la loro media $(x_i - \overline{x})$ e tra i valori delle ore e la loro media $(y_i - \overline{y})$. Noteremo che:

Primo quadrante

$x_i - \overline{x}$ è positiva perché $x_i > \overline{x}$
$y_i - \overline{y}$ è positiva perché $y_i > \overline{y}$

Secondo quadrante

$x_i - \overline{x}$ è negativa perché $x_i < \overline{x}$
$y_i - \overline{y}$ è positiva perché $y_i > \overline{y}$

Terzo quadrante

$x_i - \overline{x}$ è negativa perché $x_i < \overline{x}$
$y_i - \overline{y}$ è negativa perché $y_i < \overline{y}$

Quarto quadrante

$x_i - \overline{x}$ è positiva perché $x_i > \overline{x}$
$y_i - \overline{y}$ è negativa perché $y_i < \overline{y}$

Osservazione 2

Osserviamo ora il segno del valore del prodotto $(x_i - \overline{x})(y_i - \overline{y})$:

Primo e terzo quadrante

$(x_i - \overline{x})(y_i - \overline{y}) > 0$

Secondo e quarto quadrante

$(x_i - \overline{x})(y_i - \overline{y}) < 0$

Definizione di covarianza

Chiamiamo **COVARIANZA** la media di questi prodotti. La covarianza sarà quindi:

$$\mathbf{COV(X, Y) = \sigma_{XY}} = \frac{\sum_{i=1}^{n} (x_i - x_{media})(y_i - y_{media})}{n}$$

Se la covarianza è negativa, significa che prevalgono i punti nel secondo e quarto quadrante, quindi esiste una dipendenza: all'aumentare delle ore di allenamento, diminuisce la media.
Se invece la covarianza è positiva, prevalgono i punti del primo e terzo quadrante, quindi esiste ancora una dipendenza, ma di tipo diverso: all'aumentare delle ore di allenamento aumenta anche la media.
Se la covarianza è zero le due serie di dati sono indipendenti perché non esiste una prevalenza dell'una sull'altra.

Calcolo della covarianza

Nel nostro caso la covarianza è:

$cov := Statistics[Covariance]((voti, ore), ignore = false)$

-4.05210526315789 **(1.5.1.4.1)**

E' negativa, quindi c'è dipendenza e in particolare all'aumentare delle ore di allenamento, la media dei voti tende a diminuire.

Conclusioni

Esaminiamo il risultato ottenuto e cerchiamo di trarre le opportune conclusioni:
• le variabili sono dipendenti o no?
• la dipendenza è positiva o negativa?
• questo significa che ...

Il coefficiente di correlazione lineare

Per misurare la dipendenza tra due serie di variabili X e Y si preferisce usare anche un altro indice, di più semplice interpretazione, detto **coefficiente di correlazione lineare** o **indice di Bravais- Pearson** e che viene di solito indicato con il simbolo $\boldsymbol{\rho_{XY}}$

Significato intuitivo del coefficiente di correlazione lineare

Intuitivamente possiamo dire che il coefficiente di correlazione lineare ci dà la misura di quanto un modello di tipo lineare si adatti a descrivere l'interdipendenza tra due variabili, cioè ρ esprime una misura di quanto i punti osservati si addensano attorno ad una retta.

Definizione del coefficiente di correlazione lineare

Il **COEFFICIENTE DI CORRELAZIONE LINEARE** è definito come:

$$\boldsymbol{\rho_{XY} = \frac{\sigma_{XY}}{\sigma_x \sigma_y}}$$

è cioè il rapporto tra la covarianza e il prodotto degli scarti quadratici medi:

Proprietà dell'indice di correlazione lineare

E' possibile dimostrare che l'indice di correlazione $\sigma_{XY} < \sigma_x \cdot \sigma_y$, quindi l'indice di correlazione lineare è sempre un numero compre tra -1 e 1:

$$-1 \leq \rho \leq 1$$

o anche

$$|\rho| \leq 1$$

In particolare:
• se $\boldsymbol{|\rho| = 1}$ esiste una perfetta dipendenza lineare tra le due variabili, cioè i punti si distribuiscono lungo una retta che sarà inclinata positivamente rispetto all'asse delle ascisse se $\rho = 1$, mentre sarà inclinata negativamente rispetto allo stesso asse se $\rho = -1$. Diremo che le variabili sono **perfettamente correlate**.
• se $\boldsymbol{0 < \rho < 1}$ tra le variabili non c'è una perfetta dipendenza lineare tra le due variabili, ma più ρ si avvicina a 1, più la dipendenza è forte. Diremo che le variabili sono **correlate positivamente.**
• **se** $\boldsymbol{-1 < \rho < 0}$ anche in questo caso tra le variabili non c'è una perfetta dipendenza lineare, ma più ρ si avvicina a -1, più la dipendenza è forte. Diremo che le variabili sono **correlate negativamente**.
• se $\boldsymbol{\rho = 0}$ le variabili **non sono correlate**

Calcolo del coefficiente di correlazione lineare

Calcoliamo prima di tutto gli scarti quadratici medi delle due serie di dati.
Indicheremo con:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - x_{media}\right)^2}{n}}$$

lo scarto quadratico medio della serie di voti. Nel nostro caso avremo:

$\sigma_x := Statistics[StandardDeviation](voti)$

$$0.929006260247447 \qquad \textbf{(1.5.2.4.1)}$$

mentre con

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - y_{media}\right)^2}{n}}$$

lo scarto quadratico medio della serie delle ore di allenamento. Nel nostro caso avremo:

$\sigma_y := Statistics[StandardDeviation](ore)$

$$4.96805585187691 \qquad \textbf{(1.5.2.4.2)}$$

L'indice di correlazione lineare nel nostro caso sarà:

$\rho := Statistics[Correlation](voti, ore)$

$$-0.877961776287724 \qquad \textbf{(1.5.2.4.3)}$$

Conclusioni

Cerchiamo di trarre le opportune conclusioni:

- Le due variabili sono correlate o no?
- di che tipo di correlazione si tratta, positiva o negativa?
- possiamo dire che i punti di distribuiscono abbastanza vicini ad una retta?

Sorge quindi il problema di individuare la retta.