Istituto Nazionale di Documentazione, Innovazione e Ricerca Educativa
MIUR

Cosa dicono … le rette?

di G. Baruzzo, A. D'Arpino, P. Ranzani

Area tematica

Dati e previsioni

Autori

G. Baruzzo, A. D'Arpino, P. Ranzani

Ordine di scuola

Scuola secondaria di secondo grado – secondo biennio – Classe IV

Tempo medio per svolgere il percorso

8 -10 ore

©INDIRE 2013

Sommario

Scheda generale 3
Riferimenti curricolari 4
Introduzione 11
Descrizione dell'attività 13
Fase 1 - Analisi dei dati raccolti e loro rappresentazioni 13
Fase 2 - Correlazione lineare 20
Fase 3 - Alla ricerca della retta migliore 24
Indicazioni metodologiche 35
Spunti per approfondire 36
Elementi per prove di verifica 43

©INDIRE 2013

Scheda generale

Nucleo

Dati e previsioni

Autori

G. Baruzzo, A. D'Arpino, P. Ranzani

Tematica affrontata

Regressione

Descrizione

L'attività si inserisce in un percorso pluridisciplinare (italiano, matematica e discipline tecniche scientifiche). Nell'ambito scientifico matematico, l'insegnante si propone di condurre gli studenti ad evidenziare l'importanza dello studio della relazione tra due caratteristiche antropometriche degli studenti stessi, scelte tra quelle di loro interesse. L'attività consente di arrivare alla descrizione della relazione fra due variabili attraverso la costruzione grafica e analitica di un modello lineare di sintesi e di calcolare ed interpretare gli indici collegati.

Ordine di scuola

Secondaria di secondo grado: secondo biennio (preferibilmente classe quarta)

Tempo medio per svolgere l'attività in classe

Indicativamente 8-10 ore

Nodi concettuali

- Correlazione fra caratteri quantitativi: punto di vista grafico e misura della correlazione lineare.
- Relazione fra due caratteri quantitativi: punto di vista grafico e rette di regressione come modello statistico da sostituire ai dati rilevati.

©INDIRE 2013

Riferimenti curricolari

Indicazioni curricolari

Le attività M@t.abel hanno precisi obiettivi di apprendimento che rientrano tra quelli inseriti nelle Indicazioni nazionali attualmente in vigore (D.M. n. 211 del 07/10/2010, Direttiva n. 4 del 16/01/2012, Direttiva n. 5 del 16/01/2012) e nelle Prove INVALSI. All'inizio di ciascuna attività sono riportati, perciò, i relativi riferimenti presenti nelle Indicazioni nazionali e alcuni quesiti delle Prove Invalsi che ripropongono la situazione stimolo dell'attività considerata. Una domanda Invalsi può aiutare a valutare se gli allievi hanno sviluppato, attraverso lo svolgimento dell'attività, la capacità di utilizzare la matematica per rispondere a domande in una situazione specifica. Le domande sono tratte tra quelle presenti nei vari livelli scolastici, in quanto le attività M@t.abel sono pensate in un'ottica di verticalità.

Indicazioni Nazionali per i Licei

Linee generali e competenze

Concetti e metodi che saranno obiettivo dello studio:

- gli elementi del calcolo algeRelazioni e funzionibrico, gli elementi della geometria analitica cartesiana, una buona conoscenza delle funzioni elementari dell'analisi, […];
- la conoscenza elementare di alcuni sviluppi della matematica moderna, in particolare […] dell'analisi statistica.

Obiettivi specifici di apprendimento

Relazioni e funzioni

Lo studio delle funzioni del tipo f(x) = ax + b, f(x) = ax2 + bx + c e la rappresentazione delle rette e delle parabole nel piano cartesiano consentiranno di acquisire i concetti di soluzione delle equazioni di primo e secondo grado in una incognita,[…].

©INDIRE 2013

Dati e previsioni

Lo studente, in ambiti via via più complessi, il cui studio sarà sviluppato il più possibile in collegamento con le altre discipline e in cui i dati potranno essere raccolti direttamente dagli studenti, apprenderà a far uso delle distribuzioni doppie condizionate e marginali, dei concetti di deviazione standard, dipendenza, correlazione e regressione, […].

[…]. In relazione con le nuove conoscenze acquisite approfondirà il concetto di modello matematico.

Linee Guida per gli Istituti Professionali e Tecnici

Le competenze **acquisite** consentiranno allo studente di:

- utilizzare il linguaggio e i metodi della matematica per organizzare e valutare adeguatamente informazioni qualitative e quantitative;
- utilizzare i concetti e i modelli delle scienze sperimentali per investigare fenomeni sociali […] per interpretare dati.

Conoscenze

- Rappresentazione nel piano cartesiano […] e della parabola.
- Proprietà locali e globali delle funzioni.
- Concetti di dipendenza, correlazione, regressione.

Abilità

- Analizzare funzioni continue e discontinue.
- Risolvere problemi di massimo e di minimo.
- Calcolare, anche con l'uso del computer, e interpretare misure di correlazione e parametri di regressione.
- Costruire modelli, continui e discreti, di crescita lineare, esponenziale o ad andamento periodico a partire dai dati statistici.

©INDIRE 2013

Prove INVALSI

a.s. 2012/2013 - Domanda D8

Scuola secondaria di II grado – Classe II

Il quesito può essere considerato propedeutico alle tematiche svolte nell'unità.

D8. Considera la funzione definita da: $y = 3x + 1$.

a. Quale dei seguenti grafici può rappresentare questa funzione?

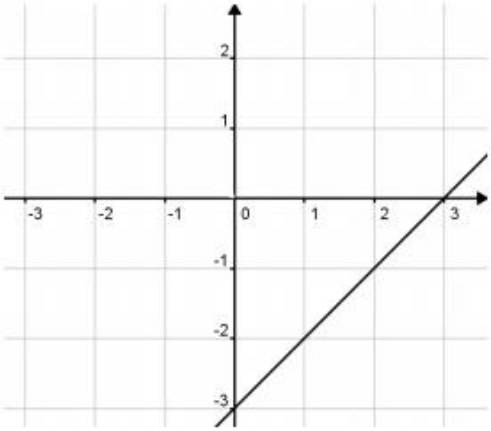


Grafico 1

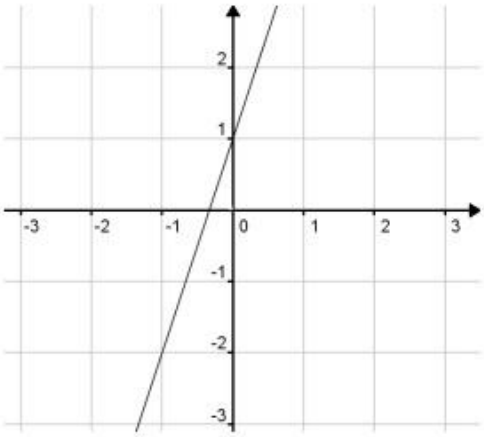


Grafico 2

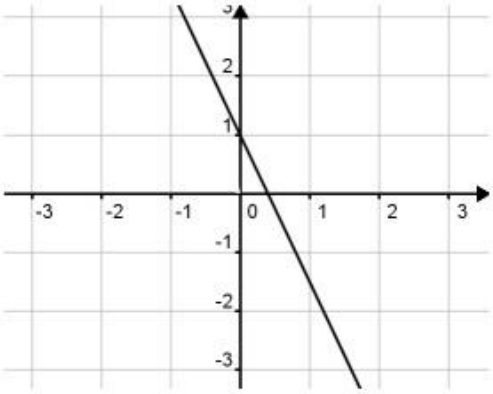


Grafico 3

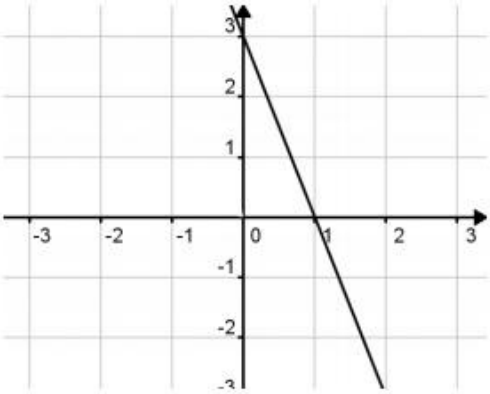


Grafico 4

©INDIRE 2013

A. ☐ Il grafico 1

B. ☐ Il grafico 2

C. ☐ Il grafico 3

D. ☐ Il grafico 4

b. **Quale valore di y si ottiene per $x = 0$?** .....

c. **Per quale valore di x si ottiene $y = 0$?** .....

d. **Per quali valori di x la y assume valori positivi?** .....

Soluzione INVALSI:

D8_a: B

D8_b: 1

D8_c: -1/3

D8_d: x > -1/3 oppure qualunque rappresentazione dell'intervallo aperto]-⅓, ∞[

Commento

Scopo della domanda:

a) Identificare pendenza (coefficiente. angolare) e intercetta (ordinata all'origine) in un grafico.
b) Calcolare il valore di una funzione in un punto.
c) Calcolare il valore di una funzione in un punto (come processo inverso attraverso la soluzione di un'equazione) – determinare lo zero di una funzione lineare.
d) Determinare il segno di una funzione lineare.

Il processo prevalente:

a, d: conoscere diverse forme di rappresentazione e passare da una all'altra;

b, c: conoscere e utilizzare algoritmi e procedure

©INDIRE 2013

Per rispondere correttamente al primo item è sufficiente che gli studenti riconoscano gli effetti che hanno la pendenza e l'intercetta di una funzione lineare sul suo grafico. Osservare che la pendenza è positiva porta immediatamente a escludere che la risposta esatta possa trovarsi fra le opzioni C e D. Osservare che l'ordinata all'origine è 1 esclude la possibilità che il grafico corretto sia il grafico 1. Quindi la riposta corretta è B. Per rispondere correttamente all'item b è sufficiente possedere il concetto di funzione come macchina INPUT-OUTPUT e calcolare y(0), oppure riconoscere immediatamente in y(0) l'ordinata all'origine. La risposta all'item c richiede il procedimento inverso, cioè il procedimento che porta, noto l'OUTPUT, a determinare l'INPUT di una funzione lineare. Riconoscere che il valore richiesto è lo zero della funzione, cioè l'ascissa del punto di ordinata uguale a 0, consente agli studenti di avere uno strumento di controllo sul risultato ottenuto (deve essere negativo e maggiore di – 1). Per rispondere all'item d è sufficiente che gli studenti sappiano leggere graficamente il segno di una funzione lineare di cui hanno già determinato lo zero

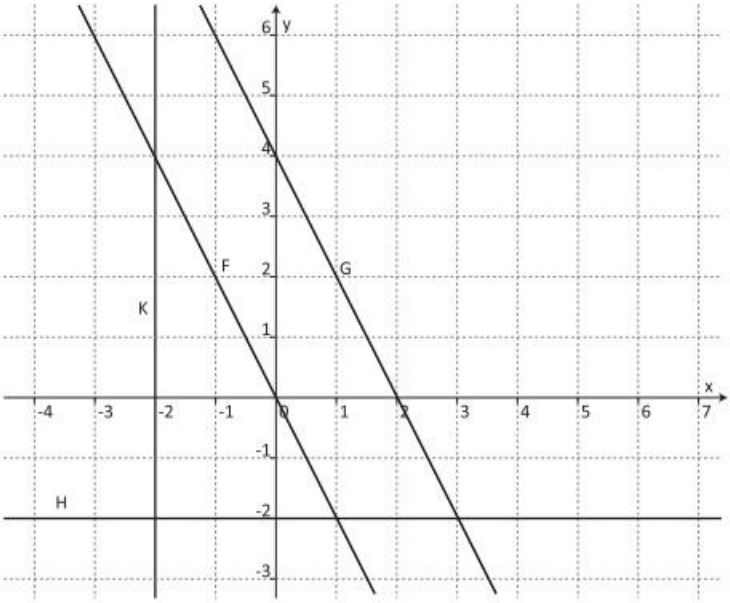
©INDIRE 2013

a.s. 2013/2014 - Domanda D14

Scuola secondaria di II grado – Classe II

Il quesito può essere considerato propedeutico alle tematiche svolte nell'unità.

D14. Sul seguente piano cartesiano sono rappresentate le rette *F, G, H, K*.



Associa a ciascuna delle equazioni in tabella la retta corrispondente.

Metti una crocetta per ogni riga.

	Equazione	Retta *F*	Retta *G*	Retta *H*	Retta *K*
a.	$y = -2x + 4$	☐	☐	☐	☐
b.	$y = -2x$	☐	☐	☐	☐
c.	$y = -2$	☐	☐	☐	☐

Soluzione INVALSI:

D14_a: retta G

D14_b: retta F

D14_c: retta H

©INDIRE 2013

Commento

Lo scopo della domanda è quello di passare dal grafico di una retta ad una sua equazione o viceversa, mentre il processo prevalente è quello di saper riconoscere diverse forme di rappresentazione e passare da una all'altra. L'item consente di verificare la competenza di conversione tra diversi registri di rappresentazione, nello specifico tra il registro di rappresentazione grafica (i grafici) e quello di rappresentazione simbolica (le formule). Per associare in modo corretto alle formule i rispettivi grafici è sufficiente conoscere come possono essere individuate, in un grafico e in una formula del tipo y= ax+ b la pendenza e l'intercetta all'origine.

©INDIRE 2013

Introduzione

Questa attività ha lo scopo di individuare un possibile legame tra due caratteri di tipo quantitativo osservati su *n* unità statistiche, una volta stabilito, in base al fenomeno che si sta studiando, quale tra le due è la variabile dipendente e quale quella indipendente.

Individuato un carattere come antecedente casuale, o carattere esplicativo, rispetto all'altro carattere che si considera come dipendente, la rappresentazione grafica degli *n* punti dà modo di:

- cogliere il possibile andamento dei dati;
- scegliere la funzione matematica più adatta ad esprimere il loro legame;
- calcolare, utilizzando le informazioni disponibili, i valori dei coefficienti incogniti che compaiono nella funzione scelta;
- verificare la qualità dei risultati ottenuti.

Il passaggio dalla rappresentazione grafica della relazione empirica - tramite i punti osservati - ad una funzione risulta utile per esprimere matematicamente il legame tra due caratteri di una distribuzione doppia, in tal modo la funzione matematica diviene il modello teorico per rappresentare sinteticamente la realtà empirica osservata. Questa attività propone la retta come modello di interpolazione dei dati osservati e il metodo di interpolazione dei minimi quadrati come strumento che fa passare dai dati reali al modello. Il metodo viene sviluppato senza ricorrere al calcolo differenziale in modo da poter essere proposto già dal primo anno del secondo biennio. L'insegnante per sviluppare questa attività può recuperare dati da sperimentazioni che riguardano la fisica, da insiemi di dati che riguardano grandezze economiche in ambiti territoriali definiti, da valutazioni degli studenti su materie diverse, da misure antropometriche rilevate sugli studenti della classe, eccetera. In questa attività vengono utilizzati i dati raccolti nel 2009-10 in un'indagine statistica svolta presso l'I. C. Valmorea (CO) attraverso la quale si voleva verificare la veridicità di alcune affermazioni tratte dal disegno di Leonardo da Vinci noto come l'uomo di Vitruvio. Nella Tabella 1 è riportato uno stralcio dei dati raccolti.

©INDIRE 2013

Dati rilevati sugli alunni nati nel 1996 delle classi 3A, 3B, 3C dell'Istituto Comprensivo Valore (CO) nell'anno scolastico 2009/2010					
u.s	genere	altezza cm	apertura braccia cm	larghezza spalle cm	Lunghezza gomito-punta mano cm
1	F	162	158	40	39
2	F	155	153	37	38
3	M	163	159	39	42
4	M	161	167	42	40
5	M	156	153	40	38
6	F	156	153	40	37
7	M	159	157	34	40
8	M	164	172	39	42
9	F	167	168	35	42
10	F	157	156	36	39
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
60	M	151	151	33	38
61	M	171	174	45	42
62	F	171	163	40	43
63	M	150	158	36	38
64	F	145	143	37	36
65	M	161	164	42	41
66	M	168	172	43	43

Fonte: indagine svolta tra gli alunni nati nel 1996 dell'I.C. "Valmorea" di Como a. s. 2009/10

Tabella 1

Scarica il file .xls

©INDIRE 2013

Descrizione dell'attività

Fase 1 - Analisi dei dati raccolti e loro rappresentazioni

Dopo una discussione in classe sul disegno di Leonardo, visualizzato eventualmente attraverso la lavagna interattiva multimediale ed aver letto alcuni passaggi chiave del testo di Leonardo che riprende l'idea riportata nel *De Architectura* di Vitruvio, sulle proprietà presenti nel disegno e verificate utilizzando gli strumenti di misura presenti nel software a disposizione con la lavagna stessa (compasso, righello, ecc), l'insegnante chiede alla classe se sia possibile verificare, con i dati a disposizione la seguente affermazione :

"***Dal gomito alla punta della mano fia la quarta parte dell'omo***."

Nota per l'insegnante

Per lo sviluppo dell'attività è conveniente avere a disposizione un laboratorio informatico per l'elaborazione dei dati attraverso l'uso di un foglio di calcolo. Tale scelta accelera la fase di calcolo lasciando spazio alla riflessione e alla formalizzazione dei concetti che via via emergono nel corso di questa proposta didattica.

L'insegnante presenta alla classe i dati sperimentali di Tabella 1 raccolti dai loro compagni dell'IC Valmorea. Si tratta dei dati rilevati su 66 studenti della stessa età, maschi e femmine, dei quali si sono misurati rispettivamente altezza, apertura braccia, larghezza spalle, lunghezza gomito-punta mano; le misure sono espresse in cm ed arrotondate tutte all'unità. Per verificare l'affermazione di Leonardo che è stata scelta occorre usare i caratteri altezza e lunghezza gomito-punta mano di cui si riporta, in Tabella 2, uno stralcio dei dati.

©INDIRE 2013

Dati rilevati sugli alunni nati nel 1996 delle classi 3A, 3B, 3C dell'Istituto Comprensivo Valore (CO) nell'anno scolastico 2009/2010		
u.s.	**altezza cm**	**Lunghezza gomito-punta mano cm**
1	162	39
2	155	38
3	163	42
4	161	40
5	156	38
6	156	37
7	159	40
8	164	42
9	167	42
10	157	39
.....	.....	.....
.....	.....	.....
60	151	38
61	171	42
62	171	43
63	150	38
64	145	36
65	161	41
66	168	43

Tabella 2

©INDIRE 2013

Nota per l'insegnante

La tabella contiene la rilevazione contemporanea delle modalità di due caratteri rispetto a ciascuna unità statistica, si tratta dunque di una distribuzione doppia in forma unitaria. La coppia di informazioni (x_i; y_i) è fornita infatti per ciascuna unità statistica.

Gli studenti, avendo a disposizione i dati della distribuzione doppia di tabella 2, calcolano il rapporto tra la lunghezza "gomito-punta mano" e l'altezza, per alcune unità statistiche. L'insegnante chiede di usare sempre 3 cifre decimali. Così per la prima unità statistica il rapporto vale 0,241, per la terza unità statistica tale rapporto vale 0,258, ecc. Gli studenti comprendono subito che in realtà si ottengono risultati diversi. A questo punto l'insegnante richiama l'attenzione sulla necessità di verificare il risultato del rapporto su tutte le unità statistiche, per poi cercare di esprimere il legame attraverso una funzione matematica. Chiede come sia possibile rappresentare la distribuzione congiunta delle grandezze oggetto di studio per individuare la relazione esistente tra "l'altezza" di uno studente e la sua "lunghezza gomito - punta mano".

Fa osservare che i dati riguardanti un fenomeno statistico, rilevati mediante coppie ordinate (*xi; yi*) di valori, possono essere rappresentati in un riferimento cartesiano con un diagramma a dispersione o "nuvola di punti" e chiede alla classe quale potrebbe essere considerata come variabile indipendente e quale come variabile dipendente.

[Vedi il percorso I grafici…questi sconosciuti]

Per provare a rispondere al quesito di Leonardo da Vinci il docente suggerisce di scegliere come variabile indipendente l'altezza (*X*) e come variabile dipendente la lunghezza gomito-punta mano (*Y*).

©INDIRE 2013

Nota per l'insegnante

Per una informazione sulle dimensioni assolute del corpo umano e i rapporti dimensionali tra le sue varie parti, alcune delle quali sono utilizzate in questa unità, è possibile fare riferimento al sito

http://www.treccani.it/enciclopedia/antropometria_(Universo_del_Corpo)/

altro sito interessante è:

http://www.design.unifi.it/docenti/corsi1011/pro/tosi1011/07-ANTROPOMETRIA-misure%20e%20applicazioni-2010-11.pdf

Gli studenti, eventualmente suddivisi in gruppi, costruiscono la nuvola dei punti pervenendo alla Figura 1.

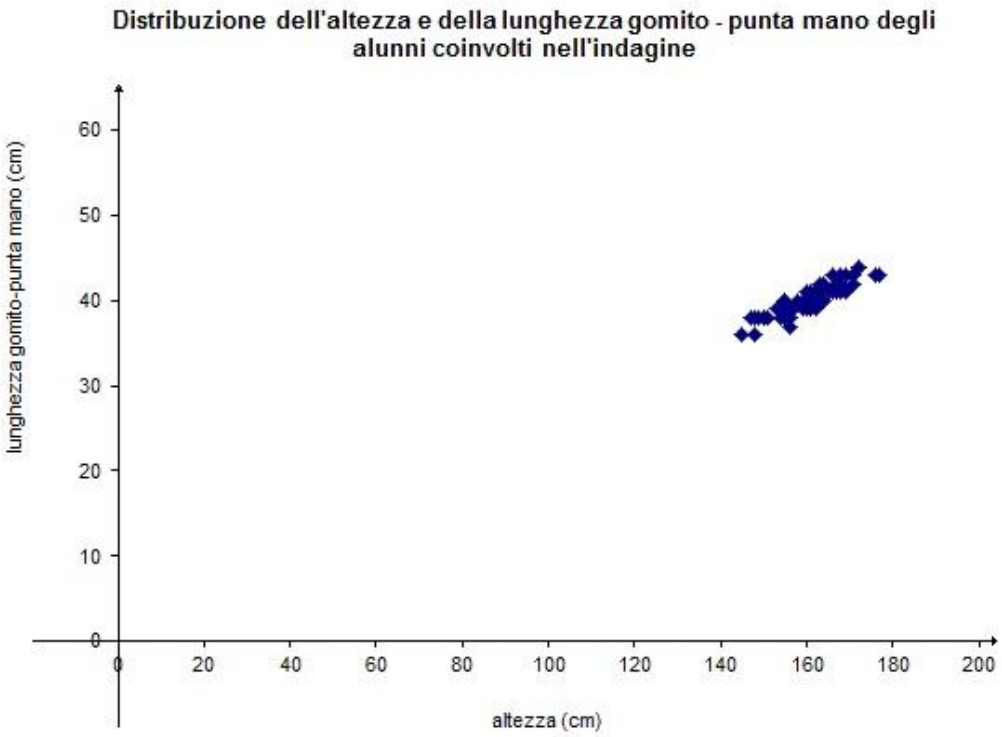


Figura 1

©INDIRE 2013

Ottenuto il grafico di Figura 1 l'insegnante chiede agli studenti di formulare, prima verbalmente e poi in forma scritta, alcune osservazioni inerenti il grafico stesso.

Da una prima osservazione della distribuzione dei punti, gli studenti possono giungere quasi immediatamente ad affermare che, in generale:

a) "più aumenta l'altezza più aumenta la lunghezza gomito-punta mano" e viceversa;

b) "in linea di massima, i ragazzi più alti sono anche quelli con maggior lunghezza gomito-punta mano".

Dopo aver fatto riflettere sulla distribuzione della nuvola dei punti, sollecitando gli studenti a formulare un'ipotesi su un'eventuale relazione fra le due variabili, l'insegnante fa osservare, che i punti sono concentrati in un intervallo piuttosto ristretto sia sull'asse delle ascisse che sull'asse delle ordinate e mostra il grafico seguente ottenuto applicando una "traslazione in (140,35)" e rappresentando la zona dei dati modificando in modo opportuno la scala sia delle ascisse che delle ordinate.

Nota per l'insegnante

È stato scritto traslazione tra virgolette perché non è una vera e propria traslazione ma uno spostamento del grafico nell'area dati per una visione migliore dello stesso. Se la classe non ricordasse la procedura per il cambiamento di scala nei grafici è possibile recuperarla attraverso la lettura dell'approfondimento "Per cambiare scala nei grafici in Excel".

Scarica l'approfondimento: Per cambiare scala nei grafici in Excel

©INDIRE 2013

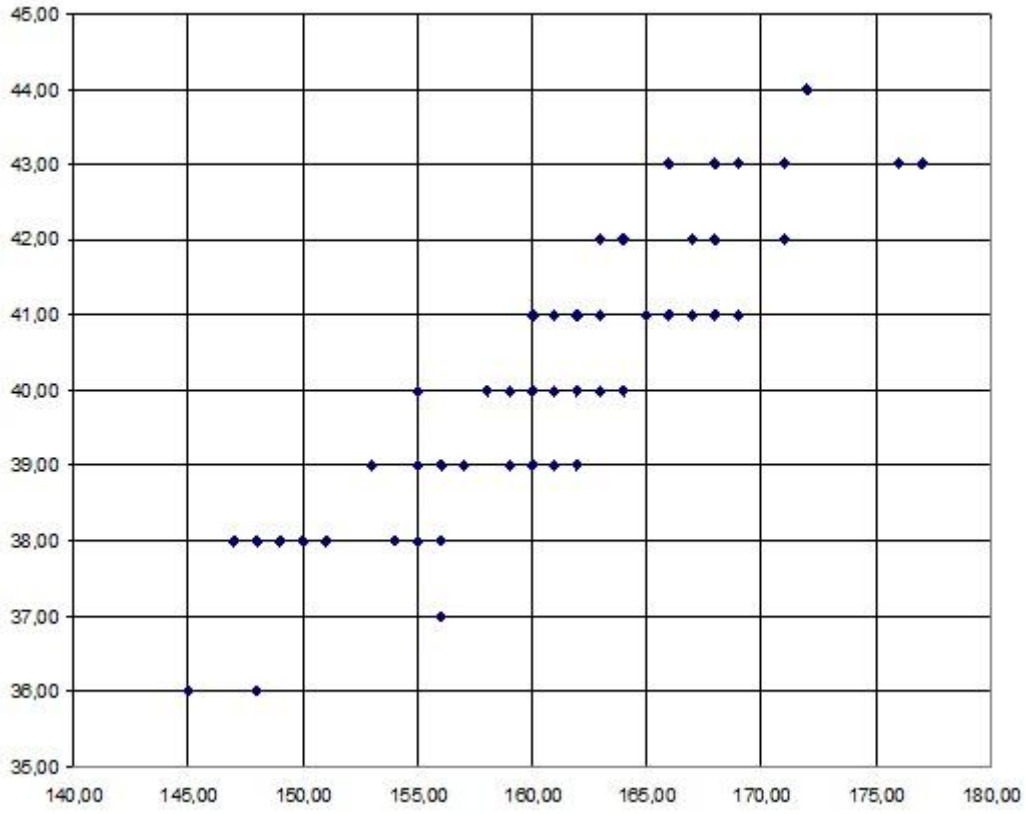


Figura 2

L'insegnante fa notare che la nuvola di punti si è dilatata e sembra disporsi attorno ad una retta che potrebbe essere espressa quindi da una funzione lineare. Chiede se è possibile che la funzione lineare ipotizzata passi per tutti i punti osservati.

Fa osservare che ad un valore di x corrispondono più valori di y pertanto non è possibile costruire una funzione che passi "attraverso" i punti osservati.

Quale strategia utilizzare per esprimere con una funzione il legame fra i due caratteri statistici? Dalla discussione potrebbe emergere la proposta di far passare la funzione scelta fra i dati rilevati; in tale caso l'insegnante chiederà alla classe per quale punto potrebbe essere conveniente farla passare.

L'insegnante, per facilitare la risposta, richiama le proprietà della media aritmetica ed in particolare quella che afferma che la somma degli scarti dei

©INDIRE 2013

valori dalla media è pari a zero dalla quale si può affermare che la media aritmetica rappresenta il punto di equilibrio della distribuzione o baricentro.

Nota per l'insegnante

Per le proprietà della media aritmetica vedi il percorso "Di media ce n'è una sola II "

L'insegnante a questo punto invita gli studenti a riflettere sul significato che il baricentro può assumere nell'ambito di una distribuzione doppia, quando entrambi i caratteri sono quantitativi.

Il risultato di tale riflessione dovrebbe portare gli studenti a concludere che:

a) "il baricentro di una nuvola di punti è senz'altro un punto" che si trova all'interno della nuvola stessa;

b) "il baricentro si colloca più o meno al centro della nostra nuvola di punti".

L'insegnante chiede come determinare con esattezza tale punto.

La discussione può portare a proporre di:

a) "indicare sul grafico un punto che, approssimativamente, si colloca in una posizione centrale rispetto alla nuvola dei punti"

b) "ordinare i dati e prendere il punto centrale della distribuzione".

Si invitano gli studenti a riflettere sul fatto che il primo procedimento si mostra poco adeguato in quanto ogni studente potrebbe individuare un punto diverso, anche di poco. Il secondo procedimento, è ambiguo, intanto perché non è possibile ordinare contemporaneamente i punti rispetto a due caratteri e poi perché il valore centrale delle unità ordinate è la mediana.

Tirando le fila della discussione, l'insegnante potrebbe fare osservare che così come in una distribuzione semplice il baricentro è la sua media aritmetica, in una distribuzione doppia il baricentro è il punto che ha per coordinate i valori medi dei due caratteri. Il baricentro è dunque il punto B($\bar{x}$; $\bar{y}$), in questo caso (161,11; 40,26). L'insegnante invita allora gli studenti a inserirlo nella "nuvola di punti" e, per esso, tracciare le due rette perpendicolari agli assi.

©INDIRE 2013

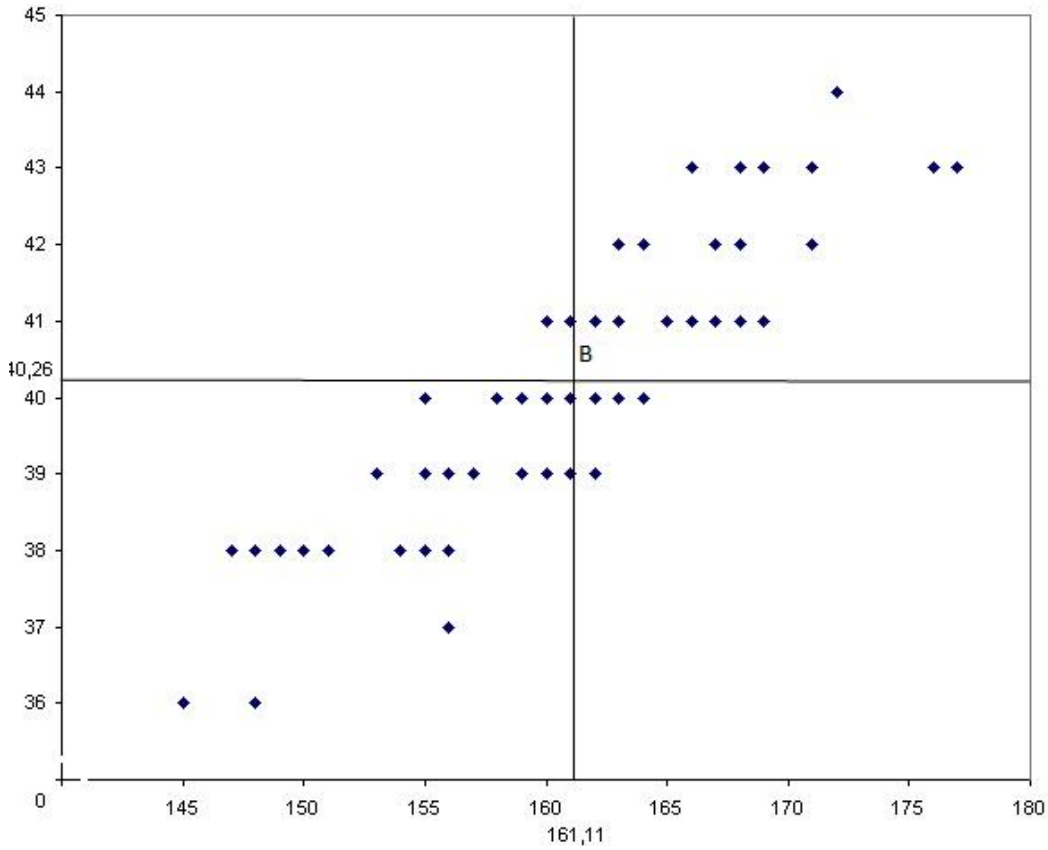


Figura3

Fase 2 - Correlazione lineare

L'insegnante, dopo aver fatto lavorare la classe sulla nuvola dei punti e sul suo baricentro, chiede se la retta, passante per il baricentro che si ipotizzava essere la funzione che meglio può spiegare il fenomeno, esprime una relazione di tipo diretto o inverso.L'insegnante aiuta gli studenti a rendersi conto che se nel I e III quadrante, individuato dalle due rette perpendicolari agli assi e condotte per B, vi sono più punti che negli altri due, questo significa che all'aumentare di una variabile aumenta, in media, anche l'altra; se abbiamo più punti nel II e IV quadrante all'aumentare di una variabile diminuisce, in media, l'altra; se tutti i punti stanno sugli assi, cioè sulle rette x = $\bar{x}$, y = $\bar{y}$, o si distribuiscono in modo uniforme fra i quattro quadranti, tra le due variabili non c'è dipendenza di tipo lineare. Per convalidare quanto emerso dalla discussione l'insegnante fa calcolare, per ogni coppia (x,y), la coppia di scarti [(x_i − $\bar{x}$),(y_i − $\bar{y}$)] e ne fa trovare il prodotto.

©INDIRE 2013

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})*(y_i - \bar{y})$
0,8939	-1,2576	-1,1242
-6,1061	-2,2576	13,7849
1,8939	1,7424	3,3000
-0,1061	-0,2576	0,0273
-5,1061	-2,2576	11,5273
-5,1061	-3,2576	16,6334
-2,1061	-0,2576	0,5425
2,8939	1,7424	5,0425
5,8939	1,7424	10,2697
-4,1061	-1,2576	5,1637
0,8939	-0,2576	-0,2303
6,8939	1,7424	12,0122
0,8939	0,7424	0,6637
-1,1061	-0,2576	0,2849
2,8939	1,7424	5,0425
14,8939	2,7424	40,8455
4,8939	2,7424	13,4213
…	..	…
..	..	…
9,8939	2,7424	27,1334
-11,1061	-2,2576	25,0728
-16,1061	-4,2576	68,5728
-0,1061	0,7424	-0,0787
6,8939	2,7424	18,9061

Tabella 3

©INDIRE 2013

Nella distribuzione in esame, gli studenti evidenziano una prevalenza di scarti di segno concorde (58 su 66) e concludono che nella distribuzione in esame c'è, mediamente, concordanza fra i caratteri. Ovviamente in generale ci possono essere anche situazioni in cui prevalgono coppie con segni discordi , in questi casi mediamente fra i due caratteri c'è discordanza. L'insegnante propone allora un indice che misura tale prevalenza e lo chiama covarianza:

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

L'insegnante, per verificare che gli studenti abbiano ben chiaro il significato della concordanza e della discordanza chiede loro se ritengono che la covarianza possa essere solo positiva, e ribadisce che la covarianza fornisce sia la misura sia il verso in cui le variabili *X* e *Y* tendono a variare, e pertanto: se *cov(x,y)* è maggiore di zero significa che al crescere di una delle variabili cresce mediamente anche l'altra, (vi saranno più punti nel primo e terzo quadrante); in tale situazione si dice che fra *X* e *Y* c'è concordanza o "correlazione positiva"; se *cov(x,y)* è negativa vi è una prevalenza di punti nel secondo e quarto quadrante, e quindi al crescere di una delle variabili l'altra mediamente decresce e, in tal caso, si parla di discordanza o "correlazione negativa".

Con l'aiuto di un foglio elettronico, l'insegnante fa calcolare il valore della covarianza sui dati di Tabella 2 ottenendo cov(*x,y*)=11,9575 con ciò concludendo che fra *X* ed *Y* vi è concordanza o correlazione positiva .

L'insegnante chiede se la cov(*x,y*) è un indice adimensionale, oppure se è riferita ad una unità di misura. Gli studenti dovrebbero riconoscere che cov(*x,y*) in questo caso è espressa in cm^2, mentre in generale è espressa dal prodotto dell'unità di misura di *X* per l'unità di misura di *Y.* L'insegnante osserva inoltre che la covarianza fornisce una misura assoluta della correlazione (concordanza o discordanza), che dipende dalle medie $\bar{x}$ e $\bar{y}$. In conclusione chiede se la cov(*x,y*) consente confronti fra distribuzioni diverse; alla risposta negativa degli studenti, osserva che occorre una misura relativa e che per ottenerla è

©INDIRE 2013

necessario conoscere il massimo assunto dalla cov(*x,y*). Fornisce il valore massimo della covarianza:

$$\text{Max}[\text{cov}(x,y)] = \sigma_X * \sigma_Y$$

che fa calcolare per i dati di Tabella 2. Gli studenti trovano Max[cov(*X,Y*)]=13,335 e commentano che il valore della covarianza trovata, pari a 11,9575, non è"tanto lontano" dal suo massimo. L'insegnante dice che per misurare la correlazione fra due variabili X e Y in modo che l'indice sia adimensionale e che sia possibile effettuare confronti con altre distribuzioni doppie è necessario costruire un indice dato dal rapporto fra il valore della covarianza e il suo massimo:

$$r = \frac{\text{cov}(x,y)}{\sigma_x * \sigma_y}$$

r è noto come **coefficiente di correlazione lineare** di Bravais-Pearson ed il suo valore è compreso fra –1 ed 1 estremi inclusi. Il segno di r dipende dalla covarianza, perciò r è positivo per concordanza ed è negativo per discordanza fra i caratteri. Nel caso analizzato il valore di r è:

$$r = \frac{11.9575}{1.861 * 7.165} = 0.8968$$

L'insegnante fa osservare che fra Y (lunghezza gomito-punta mano) ed X (altezza) vi è, nel collettivo esaminato, correlazione positiva ed essa è pari all'89,68% di quella massima. La concordanza è elevata.

Scarica il file .xls

©INDIRE 2013

Nota per l'insegnante

La relazione sopra scritta deriva dalla disuguaglianza di Cauchy-Schwarz che stabilisce, per due variabili, la seguente relazione fra la media del prodotto dei loro valori, elevata al quadrato, ed il prodotto dei valori medi di ciascuna variabile elevata al quadrato: $[M(X*Y)]^2 \leq M(X^2)*M(Y^2)$

Se tale relazione viene applicata agli scarti $(x_i - \bar{x})$ *e* $(y_i - \bar{y})$ *si ottiene la relazione sopra* indicata *per la covarianza. I membri della diseguaglianza non sono di facile lettura per gli studenti, si suggerisce di proporre loro di lavorare con una manciata di numeri "piccoli" e maneggevoli per verificare la diseguaglianza.*

Fase 3 - Alla ricerca della retta migliore

L'insegnante invita la classe a trarre le conseguenze dalle informazioni raccolte:

- i punti che rappresentano la distribuzione si distribuiscono prevalentemente fra il primo e il terzo quadrante,
- il valore trovato della covarianza è maggiore di zero e di conseguenza anche il coefficiente di correlazione r è positivo.

Gli studenti dovrebbero concludere che fra le due variabili è ipotizzabile una relazione di tipo lineare che segnali che all'aumentare della statura, aumenta, in media, anche la lunghezza gomito-punta mano.

L'insegnante, dopo aver ricordato l'equazione del fascio proprio di rette passanti per il punto $(x_b; y_b)$: $\frac{y - y_b}{x - x_b} = m$, chiede alla classe di trovare l'equazione del fascio di rette passanti per il punto B di coordinate $(\bar{x}; \bar{y})$.

Nel caso studiato il fascio è espresso dall'equazione: $y_t = m(x - \bar{x}) + \bar{y}$ .y_t sta a indicare il valore teorico del carattere Y, fornito dall'equazione per un valore osservato *x*. y_t è in generale diverso dal valore di Y osservato.

©INDIRE 2013

Le rette passanti per un punto sono infinite, ma dalla discussione dovrebbe emergere che i valori che esprimono le pendenze negative sono da escludere nel problema che si sta cercando di risolvere, visto che l'andamento dei punti consiglia un relazione diretta. Dunque *m* deve essere positivo. L'insegnante invita la classe, con l'aiuto di un supporto informatico, a disegnare alcune rette del fascio scegliendo diversi valori di *m* che abbiano senso per la soluzione del problema. Un possibile risultato potrebbe essere quello riportato in Figura 4.

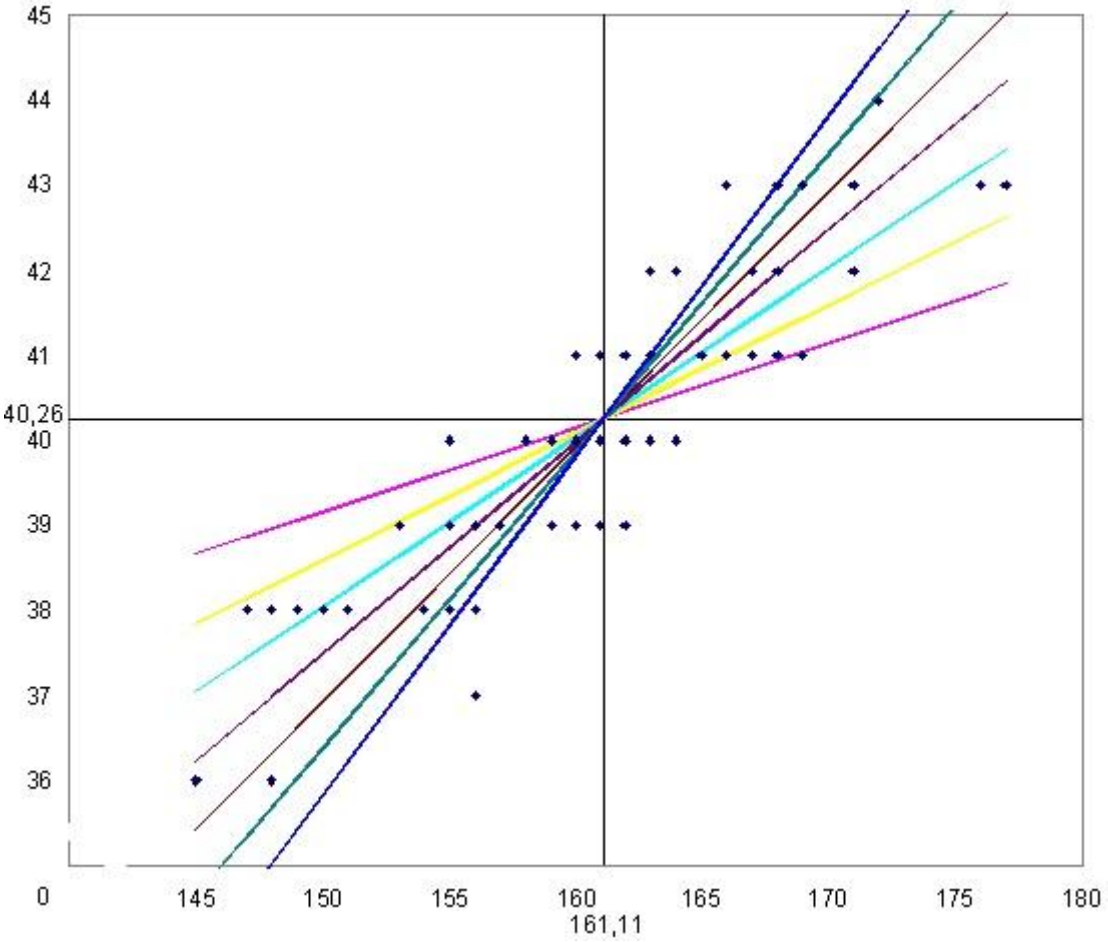


Figura 4

Rimane però ancora da stabilire come scegliere, tra le rette che hanno pendenza positiva quella che descrive "meglio" la relazione ipotizzata tra i dati. Dalla discussione potrebbe emergere la proposta di tenere "sotto controllo" la "distanza" dei punti osservati dai corrispondenti punti appartenenti alle rette del

©INDIRE 2013

fascio e quindi di calcolare, per ciascuna retta disegnata, la differenza *(y_i - y_{ti})* dove y_i sono i valori osservati e y_{ti} sono i valori calcolati sostituendo nell'equazione del fascio a x i valori osservati x_i.

Nota per l'insegnante

Nella frase precedente è stato scritto distanza fra virgolette perché quella definita non è una vera distanza in senso matematico. Ad esempio nel seguito si troveranno anche "distanze" negative, proprietà che una distanza di tipo matematico, non possiede mai.

La domanda che verrà posta alla classe risulta ora la seguente:

Rispetto a ciascuna retta del fascio, la somma delle differenze $\sum_{i=1}^{n}(y_i - y_{ti})$ *porta a* qualche informazione sulla "capacità" della retta scelta di interpolare i punti osservati? Per trovare risposta alla domanda, l'insegnante invita la classe, divisa in gruppi ciascuno dei quali lavora con una retta diversa,a trovare sperimentalmente, con l'uso del foglio di calcolo, tale somma. *Che* risultato *si ottiene*?

Nota per l'insegnante

Se la classe ha sviluppato dei procedimenti di calcolo con l'uso del simbolo sommatoria e se gli studenti conoscono le regole ad essa connesse, è possibile dimostrare la proprietà che la somma delle differenze di ciascun valore osservato dal corrispondente valore teorico è zero

$$\sum_{i=1}^{n}(y_i - y_{ti}) = 0$$

dove

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

$$\sum_{i=1}^{n}(y_i - y_{ti}) = \sum_{i=1}^{n}\left[y_i - m(x_i - \bar{x}) - \bar{y}\right] = \sum_{i=1}^{n} y_i - m \cdot \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} m \cdot \bar{x} - \sum_{i=1}^{n} \bar{y} = n \cdot \bar{y} - m \cdot n \cdot \bar{x} + m \cdot n \cdot \bar{x} - n \cdot \bar{y} = 0$$

©INDIRE 2013

Dalla discussione emerge che gli studenti hanno trovato che il risultato della somma delle differenze calcolate per ciascuna retta risulta pari a zero, e quindi la via intrapresa non dà risultati utili. Sollevato il problema, l'insegnante chiede alla classe se già ci si è trovati in una situazione simile e come è stato affrontato e risolto il problema. Se gli studenti hanno affrontato in modo critico lo studio della variabilità e di σ dovrebbero essere in grado di trovare la soluzione che consiste nell'elevare al quadrato ciascuna differenza $(y_i - y_{ti})$ e nel calcolare $\sum_{i=1}^{n}(y_i - y_{ti})^2$

[Vedi il percorso Siamo vincoli o… sparpagliati]

L'insegnante invita i gruppi a calcolare la quantità $\sum_{i=1}^{n}(y_i - y_{ti})^2$ rispetto alle rette già considerate, in modo da riuscire ad individuare la retta che rende tale somma la più piccola possibile. Per semplificare le successive scritture e per rendere evidente che $\sum_{i=1}^{n}(y_i - y_{ti})^2$ dipende dal parametro m invita la classe a indicarlo con la scrittura

$$g(m) = \sum_{i=1}^{n}(y_i - y_{ti})^2 = \sum_{i=1}^{n}\left(y_i - \left[m(x_i - \bar{x}) - \bar{y}\right]\right)^2$$

e fa notare che, appunto, l'unica variabile nella funzione è m, poiché tutti gli altri elementi sono numeri noti osservati o calcolati sulla distribuzione doppia.

Nota per l'insegnante

L'insegnante non sottovaluti il fatto che l'abitudine degli studenti ad usare per le variabili le lettere x ed y, può creare sconcerto alla interpretazione di m, che sono abituati a leggere come costante, mentre qui invece è variabile.

Per agevolare i calcoli di g(*m*) suggerisce di utilizzare il foglio di calcolo che facilita tali operazioni e di far variare il coefficiente angolare della retta teorica

©INDIRE 2013

entro un intervallo suggerito dall'osservazione della Figura 4 e concordato con la classe.

Successivamente l'insegnante chiede alla classe di collocare in una tabella i valori che legano il coefficiente delle diverse rette del fascio con la corrispondente somma dei quadrati delle differenze, evidenziando il valore del coefficiente angolare per cui g(*m*) risulta avere il valore minore.

m	g(*m*)= $\sum_{i=1}^{n}(y_i - y_{ti})^2$
0,1	104,664
0,15	68,098
0,2	48,473
0,25	45,789
0,3	60,046
0,35	91,245
0,4	139,385
0,45	204,466

Tabella 4

L'insegnante chiede:

siamo sicuri che non ci siano rette che rendono ancora più piccolo g(*m*)? Fa osservare che l'individuazione di *m*=0,25 è avvenuto in modo empirico, usando un processo di simulazione, utilizzando un incremento per *m* pari a 0,05. Suggerisce allora di ripetere la simulazione con valori del coefficiente angolare fra 0, 2 e 0,3 e con un incremento inferiore a 0,05, nell'esempio proposto pari a 0,01. Gli studenti in gruppo ripercorrono l'esperienza precedente con l'aiuto del foglio di calcolo e, successivamente, riportano le nuove coppie [*m*, g(*m*)] nella Tabella 5, che include anche i dati già trovati e presenti in Tabella 4.

©INDIRE 2013

m	g(*m*)= $\sum_{i=1}^{n}(y_i - y_{t_i})^2$
0,1	104,664
0,15	68,098
0,2	48,473
0,21	46,581
0,22	45,366
0,23	44,829
0,24	44,970
0,25	45,789
0,26	47,285
0,27	49,459
0,28	52,310
0,29	55,839
0,3	60,046
0,35	91,245
0,4	139,385
0,45	204,466

Tabella 5

L'insegnante, a questo punto dell'attività, chiede alla classe di analizzare la Tabella 5, di rappresentarla graficamente e, inoltre, di fare delle considerazioni sul grafico ottenuto.

©INDIRE 2013

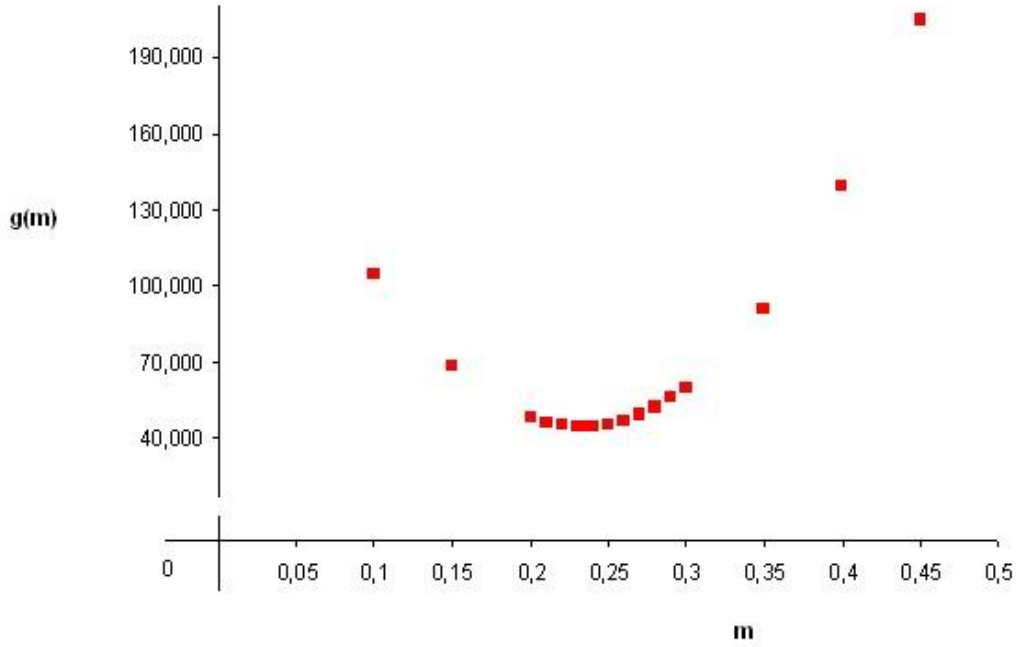


Figura 5

Dalla discussione dovrebbe emergere che i dati si dispongono in modo parabolico con la concavità rivolta verso l'alto e che c'è un minimo per *m* compreso fra 0,23 e 0,24. L'insegnante ricorda la proprietà delle parabole di questo tipo, nelle quali il minimo corrisponde al valore che la funzione assume nell'ascissa del vertice.

L'insegnante chiede:

"*In* questo *caso, che significato hanno le coordinate del vertice?"* Gli studenti, ripensando al percorso logico seguito, dovrebbero rispondere che l'ascissa del vertice rappresenta il coefficiente angolare di una delle rette del fascio e che l'ordinata rappresenta la somma dei quadrati della differenza tra i valori calcolati rispetto a questa retta e quelli osservati. Il metodo che è stato utilizzato ha permesso, quindi, di individuare, fra le rette del fascio passanti per il baricentro B($\bar{x}$; $\bar{y}$) dei dati, quella che rende **minima** la somma dei quadrati delle differenze fra i valori osservati e quelli teorici, cioè $\sum_{i=1}^{n}(y_i - y_{t_i})^2$ = Min

©INDIRE 2013

L'insegnante, con l'ausilio del foglio di calcolo, inserisce nel grafico di Figura 6 la linea di tendenza polinomiale del secondo ordine offerta dalle opzioni del grafico e la sua equazione:

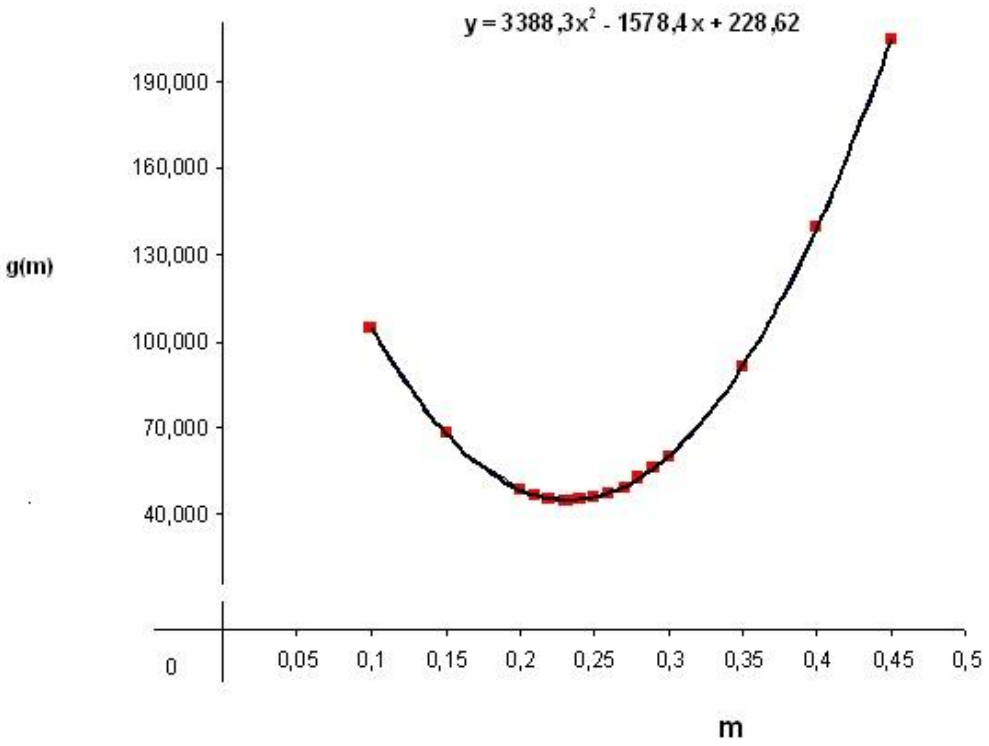


Figura 6

Per confrontare il minimo trovato sperimentalmente dagli studenti e quello della parabola indicata dal foglio di calcolo, l'insegnante chiede di trovare le coordinate del vertice della parabola.

a	3388,3
b	-1578,4
c	228,62
x_v	0,2329
y_v	44,8002

Tabella 6

©INDIRE 2013

Il valore di *m* (ossia, in Tabella 6, di x_v) fornito dalla parabola è effettivamente compreso fra 0,23 e 0,24, come previsto in base alla Tabella 5, ed è *m* = 0,23293 a cui corrisponde un valore di g(m)= $\sum_{i=1}^{n}(y_i - y_{ti})^2$ = 44,8002, ossia le coordinate del vertice della parabola rappresentata in Figura 6.

Trovato, grazie alla parabola, il valore di *m* che minimizza $\sum_{i=1}^{n}(y_i - y_{ti})^2$, rimane da individuare l'ordinata all'origine della retta interpolante.

L'insegnante fa osservare agli studenti che ora della retta interpolante si conoscono due elementi: la retta ha coefficiente angolare 0,233 e deve passare per il punto B($\bar{x}$; $\bar{y}$) ; l'informazione dà la possibilità di scrivere l'equazione della retta:

$$y_t = 0{,}233\, x + 2{,}7326$$

Tale retta è chiamata **retta dei minimi quadrati** perché rende minima la somma dei quadrati delle differenze fra ciascun valore y osservato e il corrispondente valore teorico interpolato.

La differenza $(y_i - y_{ti})$ prende il nome di residuo, è ciò che resta di y_i dopo che lo si è depurato del suo valore teorico sulla retta di regressione.

L'insegnante fa riflettere sul significato del parametro *m* della retta interpolante: 0,233 significa che all'aumentare di un centimetro in altezza di un individuo la lunghezza gomito-punta mano aumenta in media di 0,233 centimetri. *m* è noto in statistica come **coefficiente di regressione.**

L'insegnante ritorna al problema dal quale si erano prese le mosse e domanda agli studenti se la retta trovata, rispetto ad un insieme di studenti quattordicenni, verifica l'affermazione di Leonardo da Vinci. Gli studenti, avendo a disposizione, a questo punto, i valori calcolati in base alla retta y_t = 0,233 *x* + 2,7326 abbastanza agevolmente possono verificare che se rapportano due valori distinti della y_t (lunghezza gomito – punta mano teorica) ai corrispondenti valori della *x* (altezza) ottengono il valore di 0,25 e questo vale per tutte le unità

©INDIRE 2013

statistiche del collettivo. Quindi l'affermazione di Leonardo risulta essere confermata anche con il modello matematico individuato sperimentalmente.

Nota per l'insegnante

L'insegnante in base alla sua valutazione sulla capacità della classe di sviluppare il calcolo algebrico, può proporre la dimostrazione delle formule per il calcolo dei coefficienti m e q, che sono stati trovati per via sperimentale, oppure fornire direttamente le formule.

I coefficienti m e q della retta interpolante $y_t = m \times x + q$, *possono essere individuati per via algebrica a partire dalle seguenti condizioni:*

a) $\sum_{i=1}^{n}(y_i - y_{ti}) = 0$

sostituendo *a* $y_{ti} = m \cdot x_i + q$ si ha:

$$\sum_{i=1}^{n}(y_i - mx_i - q) = 0$$

che esprime *la proprietà che la somma delle differenze tra i valori osservati e i valori teorici vale zero.*

b) *La seconda condizione riprende la relazione:*

$$g(m) = \sum_{i=1}^{n}(y_i - y_{ti})^2 = \text{minimo o anche } \sum_{i=1}^{n}(y_i - m \cdot x_i - q)^2 = \min$$

Si tratta di trovare *cioè quella funzione che rende minima la somma dei quadrati delle differenze tra i valori osservati e i valori teorici.*

Dalla prima equazione *si ha:*

$$n(\bar{y} - q - m \cdot \bar{x}) = 0,$$

ovvero*:*

©INDIRE 2013

$$q = \bar{y} - m \cdot \bar{x}$$

che si interpreta dicendo *che la retta passa per il punto che ha come coordinate* ($\bar{x}$; $\bar{y}$), *baricentro della nuvola di punti* (x_i; y_i).
Sostituendo *q nella seconda condizione si ottiene:*

$$g(m) = \sum_{i=1}^{n} (y_i - mx_i - \bar{y} + m \cdot \bar{x})^2 = \min$$

che può essere *indicata come:*

$$g(m) = \sum_{i=1}^{n} (y_i - \bar{y} - m(x_i - \bar{x}))^2 = \min$$

Sviluppando *i calcoli si ottiene l'equazione di una parabola nel piano (m, g(m)):*

$$g(m) = (\sum_{i=1}^{n} (x_i - \bar{x})^2) m^2 - 2(\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})) m + \sum_{i=1}^{n} (y_i - \bar{y})^2$$

La concavità *della parabola è verso l'alto, il minimo è nel vertice, pertanto, applicando la nota formula, si ottiene:*

$$m = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Sostituendo *di nuovo m nella q =* $\bar{y}$ *- m · x si ottiene il valore dell'intercetta, senza bisogno di utilizzare calcoli complicati.*

I valori trovati consentono di scrivere l'equazione della retta* interpolatrice *detta anche retta dei minimi quadrati.

©INDIRE 2013

Indicazioni metodologiche

Questa attività può essere introdotta in una classe quarta della scuola secondaria di secondo grado dopo che l'insegnante ha verificato le conoscenze di base di statistica descrittiva.

L'introduzione delle tematiche proposte va fatto mediante un approccio di tipo induttivo, ai fini di condurre gli studenti, a partire dall'osservazione di uno o più casi particolari, a ricavare i principi generali che ne sono alla base. Gli studenti vanno supportati, nel processo di apprendimento, tramite discussioni guidate, puntualizzazioni e osservazioni mirate a far acquisire conoscenza e padronanza dei concetti di volta in volta introdotti. Inoltre, ai fini del conseguimento di tale obiettivo gli studenti vanno stimolati ad intervenire manifestando eventuali dubbi e difficoltà che possano fornire spunto per discussioni costruttive atte ad agevolare l'assimilazione dei concetti e a stimolare la riflessione sulle problematiche affrontate.

Nella trattazione delle tematiche proposte nell'unità va posta particolare attenzione ai calcoli e all'interpretazione dei dati all'interno di un contesto e ai concetti statistici sottostanti. Per facilitare il raggiungimento di tale obiettivo è suggerito il ricorso al foglio elettronico che, consentendo attraverso semplici comandi, di ottenere i risultati e i grafici ricercati, permette di poter dedicare più tempo all'analisi e alla discussione.

©INDIRE 2013

Spunti per approfondire

Spunti per un approfondimento disciplinare

L'adattamento della retta interpolata ai dati

L'insegnante, avendo a disposizione le 66 coppie di valori (y_i; y_{ti}) chiede agli studenti di calcolare il valore medio e la varianza delle due distribuzioni. Gli studenti raccolgono i risultati ottenuti nella Tabella 1.

	Y	***Y$_{ti}$***
Media aritmetica	$\bar{y}$ = 40,26 cm	$\bar{y}_t$ = 40,26 cm
Varianza	σ^2_y = 3,46396	σ^2_{yt} = 2,78516

Tabella 1

Scarica il file Analisi correlazione.xls

e osservano che la media della variabile dei dati teorici coincide con la media della variabile osservata *Y*, cioè $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} yt_i}{n}$ come già si sapeva, mentre al contrario la varianza dei dati teorici (y_{ti}) non coincide con la varianza di *Y*. Osservano che la varianza dei dati teorici, pari a 2,78516, è minore del corrispondente valore della varianza di *Y* pari a 3,46396. L'insegnante chiarisce che il dato teorico è meno variabile del dato osservato, d'altronde il dato teorico rappresenta una parte del dato osservato: la parte descritta dal legame lineare di *Y* da *X.* La differenza fra la varianza dei dati osservati e la varianza dei dati teorici in questo caso è 0,6788.

L'insegnante fa notare che tale differenza, moltiplicata per il numero di dati osservati, 66, è uguale al valore minimo di g(*m*), ossia è pari a 44,80 e scrive per la varianza di *Y* la seguente relazione:

$$\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n} = \frac{\sum_{i=1}^{n}(yt_i-\bar{y})^2}{n} + \frac{\sum_{i=1}^{n}(y_i-yt_i)^2}{n}$$

©INDIRE 2013

dove il primo termine del secondo membro essendo la varianza dei valori teorici si chiama varianza di regressione ed il secondo termine essendo la varianza dei residui si chiama varianza residua. In altri termini vale la relazione:

Varianza totale = Varianza di regressione + Varianza residua

Ossia la retta di regressione decompone la varianza totale in due parti, una spiegata dalla forza del legame lineare introdotto e l'altra residuale, in quanto non spiegata.

Nota per l'insegnante

Se la classe lo consente si può chiedere perché il secondo membro è anch'esso una varianza. In effetti ciò è vero perché la media aritmetica dei residui è nulla.

L'insegnante fa osservare che tale relazione è ottenuta scomponendo lo scarto dei dati osservati dalla loro media nel modo che segue:

$$(y_i - \bar{y}) = (y_i - y_{ti}) + (y_{ti} - \bar{y})$$

L'insegnante propone agli studenti tre grafici. Clicca sui segmenti a, c, b per visualizzarli.

©INDIRE 2013

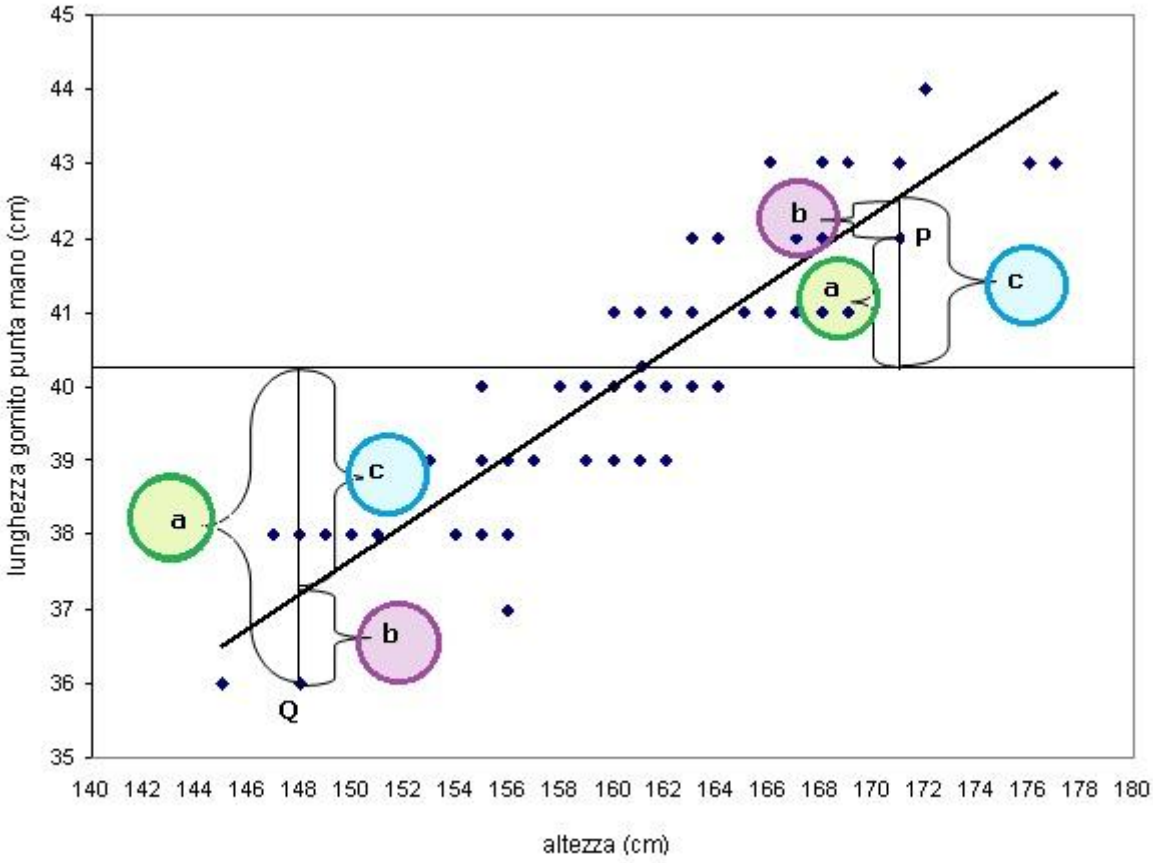


Figura 1

Gli studenti guardando la figura e leggendo alla lavagna la formula $(y_i - \bar{y}) = (y_i - y_{t_i}) + (y_{t_i} - \bar{y})$ si confrontano fra loro e giungono ad individuare tali contributi rispettivamente nei segmenti: a, c e b, scrivendo in generale:

$$\frac{\sum_{i=1}^{n} a_i^2}{n} = \frac{\sum_{i=1}^{n} c_i^2}{n} + \frac{\sum_{i=1}^{n} b_i^2}{n}$$

©INDIRE 2013

L'insegnante pone alla classe le domande seguenti, invitando gli studenti a rispondere servendosi dei grafici e della relazione:

Varianza totale = Varianza di regressione + Varianza residua

1. a parità di varianza totale, che succede alla varianza di regressione se aumenta la varianza residua?
2. e se avviene il contrario?
3. può la varianza di regressione essere nulla e quale è la retta interpolante in questo caso?
4. può la varianza residua essere nulla e come si colloca in tal caso la retta interpolante tra i punti?
5. come si può costruire un indice che consenta di misurare quanto il legame lineare introdotto è in grado di spiegare la variabilità totale rilevata?

Per le prime due domande gli studenti osservano che per l'eguaglianza:

Varianza totale = Varianza di regressione + Varianza residua

se il totale non cambia, all'aumentare di uno dei due addendi l'altro diminuisce. Pertanto, quanto più la varianza residua diminuisce, tanto più la varianza di regressione cresce e i dati osservati sono più "vicini" alla retta interpolata y_t; al contrario più il valore della varianza residua è grande più i dati sono "lontani" dalla retta y_t.

Dall'osservazione della formula $\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}=\frac{\sum_{i=1}^{n}(y_{t_i}-\bar{y})^2}{n}+\frac{\sum_{i=1}^{n}(y_i-y_{t_i})^2}{n}$ si dovrebbe facilmente concludere che:

- se la varianza di regressione $\frac{\sum_{i=1}^{n}(y_{t_i}-\bar{y})^2}{n}$ è nulla, la retta interpolante coincide con la retta parallela all'asse delle ascisse di equazione $y_t = \bar{y}$ e vale la relazione:

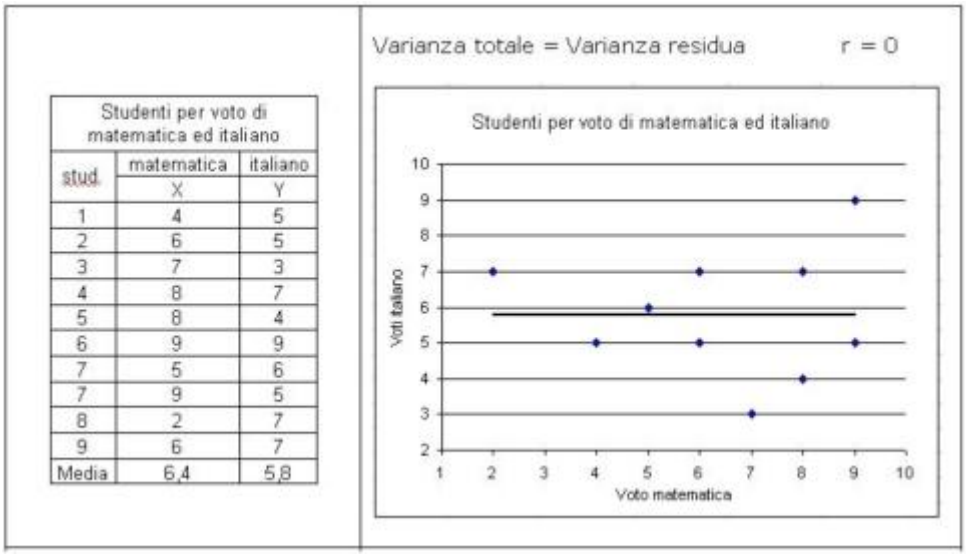
varianza totale = varianza residua;

©INDIRE 2013

- se la varianza residua $\frac{\sum_{i=1}^{n}(y_i - y_{i_i})^2}{n}$ è nulla, tutti i punti osservati appartengono alla retta di regressione individuata (retta dei minimi quadrati) e vale la relazione varianza totale = varianza di regressione.

A questo punto l'insegnante propone due esempi e i relativi grafici che visualizzano le osservazioni fatte dagli alunni per fissare anche visivamente quanto scoperto durante la discussione.

In Figura 5 propone un esempio in cui i dati presentano correlazione nulla e quindi la Varianza Totale è uguale alla Varianza Residua e il coefficiente di correlazione lineare è zero.



Studenti per voto di matematica ed italiano		
stud.	matematica X	italiano Y
1	4	5
2	6	5
3	7	3
4	8	7
5	8	4
6	9	9
7	5	6
7	9	5
8	2	7
9	6	7
Media	6,4	5,8

Figura 5

In Figura 6 propone un esempio in cui i dati osservati giacciono su di una retta e quindi la Varianza Totale è uguale alla Varianza di Regressione e il coefficiente di correlazione lineare è uno.

©INDIRE 2013

Varianza totale = Varianza di regressione r = 1

Profitto per spese di investimento di sei ditte		
Ditte	Profitto (in migliaia di euro)	Spese (in migliaia di euro)
A	50	28
B	60	35
C	30	14
D	85	53
E	95	60
F	40	21

Figura 6

Dopo una attenta discussione che tiene conto delle osservazioni emerse, l'insegnante chiede alla classe:

come si può costruire un indice che consenta di misurare quanto il legame lineare introdotto è in grado di spiegare la variabilità totale rilevata?

Propone un indice noto come **indice di determinazione lineare** dato dal rapporto fra la varianza di regressione e la varianza totale e indicato con $\mathbf{r^2}$. Esso indica la quota di varianza totale di *Y* dovuta alla relazione lineare fra *X* e *Y*, quindi indica quanto il modello di regressione lineare scelto è in grado di sintetizzare e rappresentare il fenomeno oggetto di studio.

$$r^2 = \frac{\dfrac{\sum_{i=1}^{n}\left(y_{1i} - \bar{y}\right)^2}{n}}{\dfrac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}{n}} = \frac{\sum_{i=1}^{n}\left(y_{1i} - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

la frazione rappresenta il rapporto fra la devianza di regressione e la devianza totale.

Facilmente gli studenti osservano che il rapporto è un numero puro compreso fra 0 ed 1, estremi inclusi, e che quanto più r^2 è prossimo 0 tanto minore è la

©INDIRE 2013

devianza totale spiegata dal modello lineare, mentre quanto più r^2 è prossimo a 1, tanto maggiore è la devianza totale spiegata dal modello.

Calcolando il valore dell'indice di determinazione lineare per il problema proposto nell'attività si ha $r^2 = \frac{183,821}{228,621} = 0,804$. L'insegnante aiuta alla interpretazione di r^2 affermando che esso, espresso in percentuale, indica che l'80,4% della variabilità di *Y* è dovuto al legame lineare con *X*. Nel caso analizzato l'80,4% della variabilità della lunghezza "gomito-punta mano", del collettivo, è spiegata dalla relazione lineare con l'altezza. Il 19,6% è la varianza non spiegata. L'insegnante può chiedere alla classe di esplicitare il legame tra il coefficiente di correlazione r calcolato nella fase 2 e l'indice di determinazione lineare r^2.

Gli studenti dovrebbero facilmente comprendere che l'indice di determinazione non è altro che il quadrato del coefficiente di correlazione r e quindi, in alternativa, noto quest'ultimo, per trovare il coefficiente di determinazione basta elevarlo al quadrato: ossia se r=0,897, r^2=0,804, che è esattamente quanto già ottenuto.

L'insegnante invita la classe a prestare attenzione alla diversa interpretazione dei due indici r è una misura della correlazione e varia fra -1e +1, l'indice di determinazione lineare dice quanta è la varianza totale spiegata della retta di regressione ed è compreso fra 0 ed 1.

Spunti per altre attività con gli studenti

L'insegnante può proporre come ulteriore attività una a scelta tra le seguenti:

A partire dai dati presenti in Tabella 2 costruire la retta dei minimi quadrati quando si sceglie come variabile indipendente la lunghezza "gomito-punta mano" e confrontare tutti i parametri della nuova analisi con quelli già ottenuti nell'attività proposta.

A partire dai dati presenti nella Tabella 1 l'insegnante potrà invitare gli studenti a verificare altre proprietà riferite al testo di Leonardo:

©INDIRE 2013

1. *Tanto ap(r)e l'omo nele b(r)accia, qua(n)to ella sua alteza.*
2. *La mag(g)iore larg(h)ez(z)a delle spalli chontiene insè la quarta parte dell'omo.*

Avendo già a disposizione le elaborazioni dei dati delle fasi 1-3 gli studenti potranno riadattare i fogli di lavoro con le nuove serie di dati e ottenere in tempi piuttosto rapidi i risultati.

Alternativamente si potranno dividere gli studenti in gruppi ed affidare a ciascuno dei gruppi una delle proposte in modo che poi, in sede di presentazione dei lavori, sia possibile poter confrontare le soluzioni di ciascuno dei gruppi che hanno sviluppato la stessa proposta.

Elementi per prove di verifica

Esercizio 1

Indica quali affermazioni sono vere e quali sono false e giustifica le risposte date.

Un grafico a dispersione:

a) può essere costruito per qualsiasi tipo di carattere; b) ci consente di valutare oggettivamente l'entità del legame tra i due caratteri rappresentati;

c) è la rappresentazione su un diagramma cartesiano delle modalità quantitative di due caratteri rilevati su ciascuna unità;

d) presenta tanti punti quante sono le unità del collettivo.

Esercizio 2

Indica quali affermazioni sono vere e quali sono false e giustifica le risposte date.

Il coefficiente di correlazione lineare (tra *X* e *Y*)

a) può essere calcolato tra una qualunque coppia di caratteri *X* e *Y;*

b) se è pari a zero, vuol dire che *X* e *Y* sono indipendenti;

c) risulta nullo se i due caratteri *X* e *Y* sono indipendenti;

d) è sempre compreso tra -1 e 1;

©INDIRE 2013

e) se è prossimo a -1 vuol dire che esiste un legame lineare tra *X* e *Y* molto debole;

f) se è prossimo a 1 vuol dire che esiste un forte legame lineare tra *X* e *Y;*

g) se è negativo vuol dire che al crescere di un carattere l'altra rimane costante.

Esercizio 3

Indica quali affermazioni sono vere e quali sono false e giustifica le risposte date.

In un collettivo di 10 famiglie è stato misurato il reddito (*X*) in Euro e la quota percentuale di spesa per consumi alimentari (*Y*).

Famiglia	Reddito (X)	Quota % spesa aliment. (Y)
1	1215	69.3
2	2510	44.6
3	1745	52.1
4	1875	48.0
5	3100	25.8
6	1350	55.3
7	2120	45.8
8	1970	37.7
9	2880	27.3
10	1505	60.1

a) tra i due caratteri esiste un forte legame lineare;

b) il coefficiente di correlazione lineare di Pearson `e pari a − 0.9254;

c) il coefficiente di correlazione segnala che la relazione lineare tra i due caratteri è crescente;

d) il coefficiente segnala la possibile presenza di legami (tra X e Y) diversi da quello lineare, se si rappresentassero le unità su un diagramma di dispersione, i punti si presenterebbero in ordine sparso;

f) la covarianza tra i due caratteri è negativa.

©INDIRE 2013

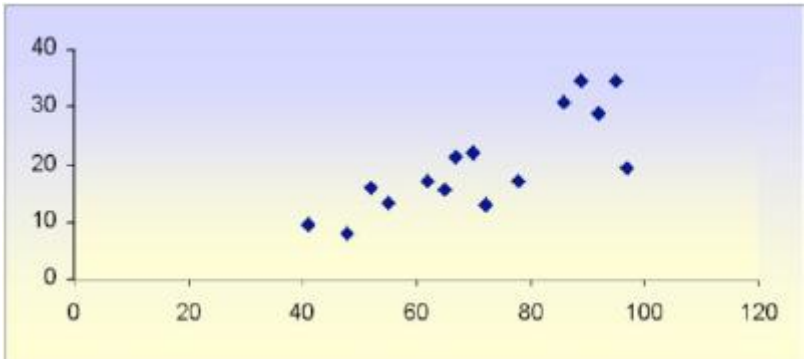
Esercizio 4

Indica quali affermazioni sono vere e quali sono false e giustifica le risposte date.

Si supponga di costruire la retta di regressione per i dati rappresentati nel diagramma a dispersione seguente.

Qual è la coppia di valori più ragionevole per l'intercetta con l'asse y e la pendenza?

(suggerimento: per individuare l'intercetta si provi a tracciare una retta passante fra i punti e prolungarla fino ad intersecare l'asse delle ordinate)



a) intercetta -20 e pendenza 2;

b) intercetta 0 e pendenza -2;

c) intercetta -20 e pendenza 0,5;

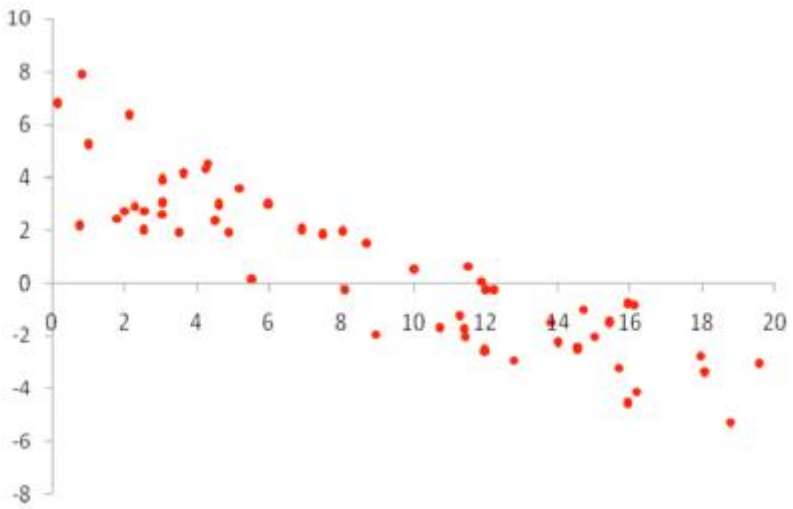
d) intercetta 20 e pendenza 2.

Esercizio 5

Indica quali affermazioni sono vere e quali sono false e giustifica le risposte date.

Se il diagramma a dispersione è come quello mostrato, qui di seguito, il coefficiente di correlazione è:

©INDIRE 2013



a) positivo;

b) negativo;

c) nullo;

d) prima positivo poi negativo.

Esercizio 6

A 32 studenti sono state chieste l'età della madre: X e età del padre: Y. Della variabile bidimensionale (X,Y) sono noti:

$\bar{x}$ = 49.3

$\bar{y}$ = 52.1

σ_x^2 =19.8

σ_y^2 =16.45

cov(X,Y)= 12

a) Commenta la situazione presentata.

b) È possibile con le informazioni date fornire la retta interpolante y_t = mx+q? Se sì, darne l'equazione.

c) Di uno studente si sa che la madre ha 37 anni, è possibile dare una

©INDIRE 2013

valutazione dell'età del padre? Come?

d) Calcola l'indice di determinazione lineare r^2.

Soluzioni

Esercizio 1

L'alternativa a) risulta falsa in quanto è necessario che i due caratteri siano entrambi quantitativi, essendo ciascun punto individuato da due valori numerici di cui il primo viene preso sull'asse delle ascisse, il secondo sull'asse delle ordinate.

L'alternativa b) risulta falsa l'osservazione del grafico ci consente solo di stabilire che tipo di legame è possibile ipotizzare tra le due variabili. Per valutare oggettivamente l'entità del legame lineare è necessario verificare il valore di r^2.

L'alternativa c) è vera per le motivazioni riportate nella giustificazione data nell'alternativa a).

L'alternativa d) è vera in generale, ogni punto del diagramma a dispersione rappresenta un'unità statistica su cui sono stati misurate le due variabili. Può succedere, tuttavia, che alcuni punti siano sovrapposti e questo si verifica se due o più unità statistiche presentano gli stessi valori per la variabile *x* e *y* rispettivamente, può perciò verificarsi che non vi siano tanti punti quante le unità del collettivo.

Esercizio 2

a) L'alternativa a) risulta falsa in quanto è necessario che le variabili siano entrambe quantitative.

b) L'alternativa b) è falsa in quanto è l'indipendenza tra due variabili che implica una correlazione nulla, mentre la correlazione nulla implica soltanto l'assenza di un legame lineare.

c) L'alternativa c) è vera per quanto detto prima.

d) L'alternativa d) è vera.

e) L'alternativa e) è falsa, tanto più vicino a -1 è il coefficiente di correlazione,

©INDIRE 2013

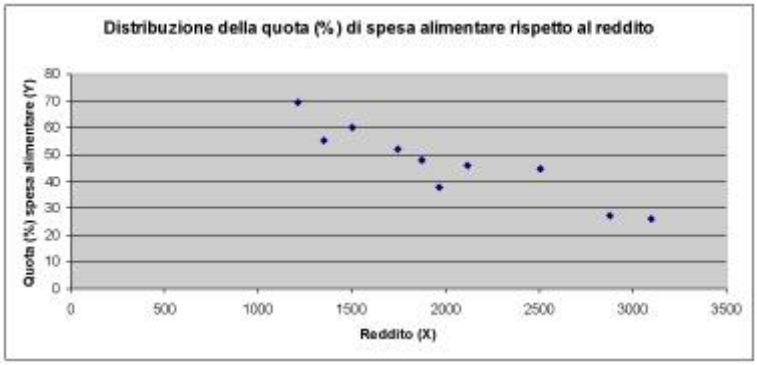
tanto più forte è il legame fra le due variabili, ma la relazione è decrescente.

f) L'alternativa f) è vera.

g) L'alternativa g) è falsa, quando il coefficiente è negativo significa che al crescere di una variabile l'altra decresce.

Esercizio 3

L'alternativa a) risulta vera. Dal diagramma a dispersione si può osservare che i punti sono addensati e ben interpolabili da una funzione lineare, la cui equazione risulta



essere pari a: $Y = -0{,}02x + 87{,}20$ che significa che all'aumentare del reddito di 1 la quota percentuale di spesa alimentare diminuisce di 0,02. L'indice r^2 è uguale a 0,8564 e significa che l'85,64% della variabilità totale della spesa alimentare è dovuta al legame lineare con il reddito.

Calcolando il coefficiente di correlazione si verifica che l'alternativa b) è vera. Dal segno del coefficiente si può verificare che l'alternativa c) è falsa in quanto la relazione è decrescente.

L'alternativa d) è falsa, il diagramma a dispersione evidenzia nettamente che il legame è interpolabile con una funzione lineare. L'alternativa f) è vera, la covarianza negativa sta ad indicare che le variabili variano in modo discorde: all'aumentare del reddito la quota percentuale di spesa alimentare tende mediamente a diminuire.

©INDIRE 2013

Esercizio 4

Osservando il grafico si comprende che l'alternativa b) è falsa in quanto la tendenza dei punti indica una relazione crescente e quindi il coefficiente angolare deve essere positivo.

Osservando la scala si comprende che la retta interpolatrice taglia l'asse delle y su un valore negativo resta da stabilire tra le alternative a) e c) quale sia quella corretta.

Osservando i punti si può dire che solo un coefficiente angolare pari a 0,5 è plausibile con l'andamento dei punti, infatti prendendo un valore della x pari a 40, in base ai parametri, otterremo un valore della y pari a -20+0,5*40 cioè pari a zero, si tratta cioè dell'intersezione con l'asse delle y. Ipotizzando un altro valore della x, ad esempio 80, otterremo, sempre nell'alternativa c) -20+0,5*80=20, valore plausibile con il grafico. Usando l'alternativa a) per x=80, si otterrebbe -20+2*80 uguale 140 ben al di fuori del range di valori previsto per la y.

Esercizio 5

Osservando il diagramma a dispersione si comprende che l'alternativa corretta è la b) essendo i dati disposti in modo decrescente.

L'alternativa a) è falsa perché in questo caso la nuvola dei punti avrebbe dovuto avere un andamento crescente.

L'alternativa c) deriva da due diverse tipologie di grafico. O un grafico con un cambiamento di andamento o un grafico in cui i punti sono disposti in maniera caotica e quindi che non mostrano andamento.

L'alternativa d) è falsa in quanto il coefficiente di correlazione è determinato in modo univoco e non può cambiare di segno nello stesso diagramma a dispersione.

©INDIRE 2013

Esercizio 6

a) Essendo la covarianza tra X e Y maggiore di 0, all'aumentare di una delle variabili aumenta mediamente anche l'altra. C'è concordanza o correlazione positiva fra età delle madri ed età dei padri.

b) Ricordando le espressioni dei parametri della retta dei minimi quadrati la risposta è affermativa, pertanto si ha:

$$m = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{cov}(x,y)}{\sigma_x^2} = \frac{12}{19{,}8} = 0{,}606 \text{ e}$$

$q = \bar{y} - m \cdot \bar{x} = 52{,}1 - 0{,}606*49{,}3 = 22{,}221$ anni;

la retta ha equazione: $y_t = 0{,}606*x + 22{,}221$

c) per uno studente la cui madre ha 37 anni si trova per il padre, mediamente, una età pari a :

$y_t = 0{,}606*37 + 22{,}221 = 44{,}65$ anni

d) l'indice richiesto può essere calcolato come quadrato dell'indice di correlazione lineare $r = \frac{\mathrm{cov}(X,Y)}{\sigma_x \cdot \sigma_y} = \frac{12}{\sqrt{19{,}8 \cdot 16{,}45}} = 0{,}6649$ da cui $r^2 = (0{,}6649)^2 = 0{,}4421$ che indica che il 44,21% della variabilità dell'età del padre è spiegata dalla relazione lineare con l'età della madre.

©INDIRE 2013

Risorse

Bibliografia

Leti, G., Cerbara, L. *Elementi di statistica descrittiva*. Il Mulino, Bologna 2009.

Ferrari, P., Nicolini, G., Tommasi, C. *Introduzione all'inferenza statistica*. G. Giappichelli Editore, Torino 2006.

Gnedenko, B. *Teoria della probabilità*. Editori Riuniti *Univ. Press,* Roma 2011.

Baldi, P. *Calcolo delle probabilità e statistica*. McGraw-Hill Companies, Milano 2003.

Cicchitelli, G. *Probabilità e Statistica*. Maggioli editore, Santarcangelo di R. (RN) 2004.

Boggio, A., Borello, G. *Statistica.* Vol.1 e 2. Petrini editore, 2008.

Matematica 2004. *La matematica del cittadino* (*Come quando fuori piove*).

Ottaviani, M.G. *Strumenti per l'analisi dei dati, II parte*.In: "Induzioni", 24, 2002, pp.49-80.

Sitografia

Retta di regressione

(Visitato nel giugno 2013)

Il Metodo dei Minimi Quadrati. Alcuni esempi svolti

(Visitato nel giugno 2013)

Antropometria. Universo del corpo

(Visitato nel giugno 2013)

©INDIRE 2013

Questo prodotto multimediale è stato realizzato nel 2013 da INDIRE con i fondi stanziati dal MIUR – Uff. VI nell'ambito del progetto m@t.abel – Apprendimenti di Base. *La grafica, i testi, le immagini, l'audio, i video e ogni altra informazione disponibile in qualunque formato sono utilizzabili a fini didattici e scientifici, purché non a scopo di lucro e sono protetti ai sensi della normativa in tema di opere dell'ingegno (legge 22 aprile 1941, n. 633).*

©INDIRE 2013