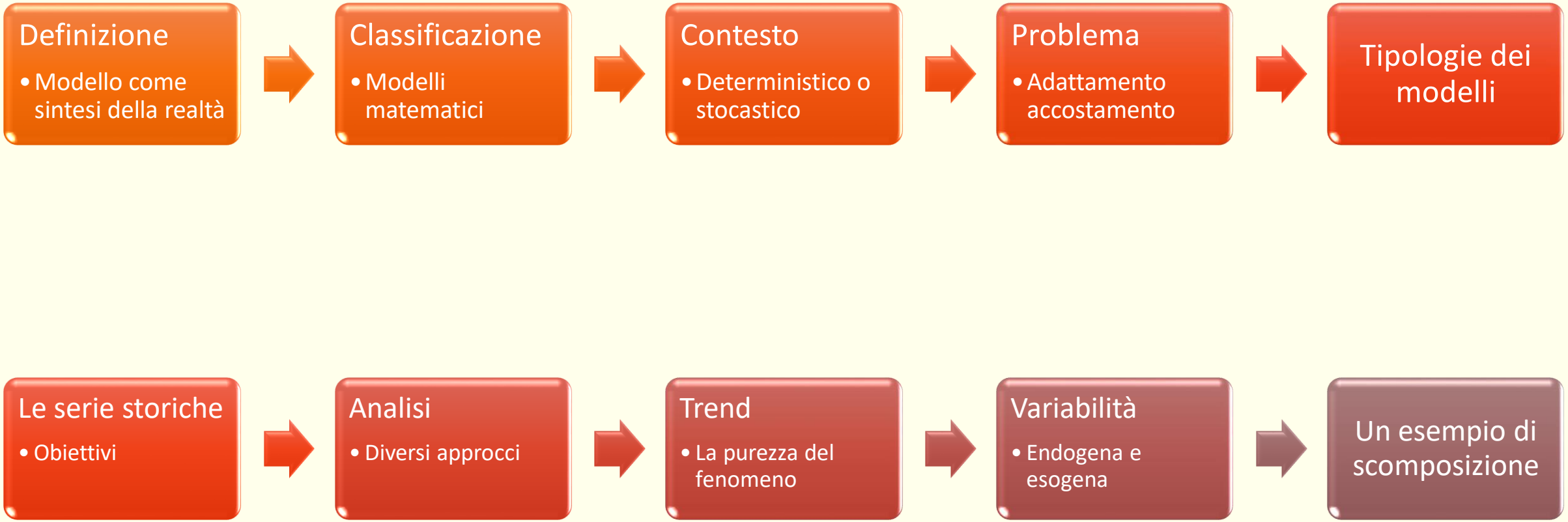
La modellizzazione matematica della realtà

un percorso di elaborazione applicativa dei contenuti didattici della matematica

INCONTRO III 24/06/2020


Definizione
• Modello come sintesi della realtà
Classificazione
• Modelli matematici
Contesto
• Deterministico o stocastico
Problema
• Adattamento accostamento
Tipologie dei modelli
Le serie storiche
• Obiettivi
Analisi
• Diversi approcci
Trend
• La purezza del fenomeno
Variabilità
• Endogena e esogena
Un esempio di scomposizione

Tipologie di modelli matematici stocastici

Nei modelli matematici rappresentati da una funzione occorre chiarire bene quali sono le variabili in gioco.

1

Analisi di una sola variabile

Questa analisi cerca di associare un modello matematico capace di descrivere le frequenze di un carattere nella popolazione. Particolarmente importante per lo studio delle popolazioni normali.

1+

$y = f(t)$

Analisi di una variabile nel tempo

Questa analisi considera il movimento di un carattere nel tempo. Particolarmente importante per lo studio delle popolazioni ad esempio. Se il carattere ha natura economica questo tipo di studio prende il nome di analisi delle serie storiche ed ha un suo sviluppo metodologico molto importante.

2

$y = f(x)$

Analisi di due variabili

Questo tipo di studio è legato al concetto di dipendenza di due variabili e di come si può spiegare una rispetto al valore dell'altra, in ambiente statistico prende il nome di regressione.

+

$y = f(x_i)$

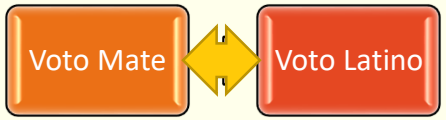
Analisi di più variabili

Questi modelli sono molto complicati da gestire e anche da rappresentare oltre al fatto che richiederebbero una introduzione allo studio di funzione in più variabili. Tra i più famosi, in ambito economico, ci sono i modelli che spiegano la produzione con le variabili capitale e lavoro (modelli di Cobb Douglas).

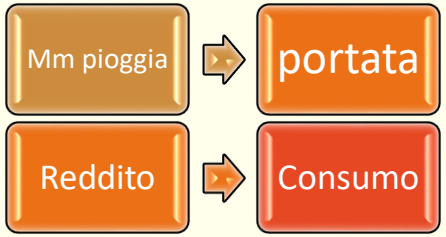
Il concetto di dipendenza

Gli approcci statistici all'analisi bivariata o multivariata vengono classificati in metodi simmetrici o asimmetrici

I primi si occupano solo di stabilire se si possono considerare le due variabili dipendenti o indipendenti tra loro, senza necessariamente entrare nel dettaglio della relazione di causa effetto.



Con i metodi asimmetrici viene supposta una relazione di causa effetto fra i caratteri osservati e si ipotizza che il valore di un carattere sia spiegato dal valore assunto da un altro carattere.



Prima della spiegazione y=f(x)

Prima di arrivare a definire un modello capace di spiegare la dipendenza di un carattere da un altro ci troviamo in una sorta di «ambiente primordiale» nel quale, attraverso l'osservazione e una prima organizzazione dei dati, può esserci suggerita l'idea di una possibile relazione fra i due caratteri.

Gli oggetti che ci suggeriscono questa relazione possono essere le tabelle o i grafici.

Come interpretare una tabella

Il suggerimento può arrivare attraverso la lettura critica di una tabella, prestando particolare attenzione alle distribuzioni marginali e a quelle condizionate.

L'idea da seguire è quella della evidente mancanza di proporzionalità tra i termini della distribuzione condizionata rispetto a quella marginale.

Infatti se la distribuzione condizionata rispettasse le stesse proporzioni della marginale allora il condizionamento della manifestazione del carattere sarebbe inesistente.

$X \setminus Y$	y_1	y_2	…	y_j	…	y_m	Tot	
x_1	$n_{1,1}$	$n_{1,2}$	…	$n_{1,j}$	…	$n_{1,m}$	$n_{1,-}$	
x_2	$n_{2,1}$	$n_{2,2}$	…	$n_{2,j}$	…	$n_{2,m}$	$n_{2,-}$	
…	…	…	…	…	…	…	…	
x_i	$n_{i,1}$	$n_{i,2}$	…	$n_{i,j}$	…	$n_{i,m}$	$n_{i,-}$	$\{y_j \mid x_i\}$
…	…	…	…	…	…	…	…	
x_k	$n_{k,1}$	$n_{k,2}$	…	$n_{k,j}$	…	$n_{k,m}$	$n_{k,-}$	
Tot	$n_{-,1}$	$n_{-,2}$	…	$n_{-,j}$	…	$n_{-,m}$	N	$\{y_j\}$
		$\{x_i \mid y_2\}$					$\{x_i\}$	

Il modello teorico di perfetta indipendenza

È un modello teorico che definisce quali sono le numerosità di ogni singola coppia di caratteri in modo tale che qualsiasi distribuzione condizionata risulti perfettamente proporzionale alla corrispondente distribuzione marginale.

$$n^*_{i,j} = \frac{n_{i,-} \cdot n_{-,j}}{N}$$

Analizzando poi gli scostamenti si possono notare e valutare le distanze. Più sono elevate maggiore è la distanza dal modello di perfetta indipendenza.

$X \setminus Y$	y_1	y_2	…	y_j	…	y_k	Tot
x_1	$n^*_{1,1}$	$n^*_{1,2}$	…	$n^*_{1,j}$	…	$n^*_{1,k}$	$n_{1,-}$
x_2	$n^*_{2,1}$	$n^*_{2,2}$	…	$n^*_{2,j}$	…	$n^*_{2,k}$	$n_{2,-}$
…	…	…	…	…	…	…	…
x_i	$n^*_{i,1}$	$n^*_{i,2}$	…	$n^*_{i,j}$	…	$n^*_{i,k}$	$n_{i,-}$
…	…	…	…	…	…	…	…
x_h	$n^*_{h,1}$	$n^*_{h,2}$	…	$n^*_{h,j}$	…	$n^*_{h,k}$	$n_{h,-}$
Tot	$n_{-,1}$	$n_{-,2}$	…	$n_{-,j}$	…	$n_{-,k}$	N

La tavola di contingenza

$$c_{i,j} = \frac{\left(n_{i,j} - n^*_{i,j}\right)^2}{n^*_{i,j}}$$

Seguendo il principio già visto nel primo incontro secondo il quale un modello teorico è un descrittore accettabile di una realtà se la distanza dei suoi valori da quelli reali è bassa, si può costruire una tabella detta di contingenza, che riporta per ogni cella una misura di distanza rapportata al valore teorico.

La sommatoria di quelle distanze è il chi-quadro, il suo valore più è piccolo maggiore sarà l'adesione della realtà al modello teorico di indipendenza.

$X \setminus Y$	y_1	y_2	…	y_j	…	y_k	Tot
x_1	$c_{1,1}$	$c_{1,2}$	…	$c_{1,j}$	…	$c_{1,k}$	$c_{1,-}$
x_2	$c_{2,1}$	$c_{2,2}$	…	$c_{2,j}$	…	$c_{2,k}$	$c_{2,-}$
…	…	…	…	…	…	…	…
X_i	$c_{i,1}$	$c_{i,2}$	…	$c_{i,j}$	…	$c_{i,k}$	$c_{i,-}$
…	…	…	…	…	…	…	…
X_h	$c_{h,1}$	$c_{h,2}$	…	$c_{h,j}$	…	$c_{h,k}$	$c_{h,-}$
Tot	$c_{-,1}$	$c_{-,2}$	…	$c_{-,j}$	…	$c_{-,k}$	χ^2

$$\chi^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} c_{i,j}$$

La normalizzazione del chi-quadro

Il valore ottenuto del chi quadro è di difficile interpretazione poiché risente della struttura della tabella e della dimensione della popolazione.

Un correttore che permette di evitare il condizionamento della dimensione della popolazione e di poter descrivere meglio la situazione consiste nel normalizzare il chi-quadro E renderlo un numero compreso tra 0 e 1 rapportandolo al massimo che può essere.

Un'ulteriore misura è l'indice V di Cramer (indice di connessione)

$$\chi^2_{norm} = \frac{\chi^2}{N \cdot (min\{h,k\} - 1)}$$

$$V = \sqrt{\frac{\chi^2}{N \cdot (min\{h,k\} - 1)}}$$

Un esempio

Osservazioni

	R	S	T	U	
a	7	8	7	2	24
b	4	3	4	2	13
c	5	9	8	8	30
d	2	4	1	9	16
	18	24	20	21	83

Tabella di indipendenza

	R	S	T	U	
a	5,14	6,86	5,71	6,00	24
b	2,79	3,71	3,10	3,25	13
c	6,43	8,57	7,14	7,50	30
d	3,43	4,57	3,81	4,00	16
	18	24	20	21	83

	R	S	T	U	
a	0,67	0,19	0,29	2,67	
b	0,53	0,14	0,26	0,48	
c	0,32	0,02	0,10	0,03	
d	0,60	0,07	2,07	6,25	
					14,69

$\chi^2 = 14.69$ $\chi^2_{norm} = 0{,}059$

$V = 0{,}243$

Osservazioni

	R	S	T	U	
a	17	7	0	0	24
b	1	11	1	0	13
c	0	6	19	5	30
d	0	0	1	15	16
	18	24	21	20	83

Tabella di indipendenza

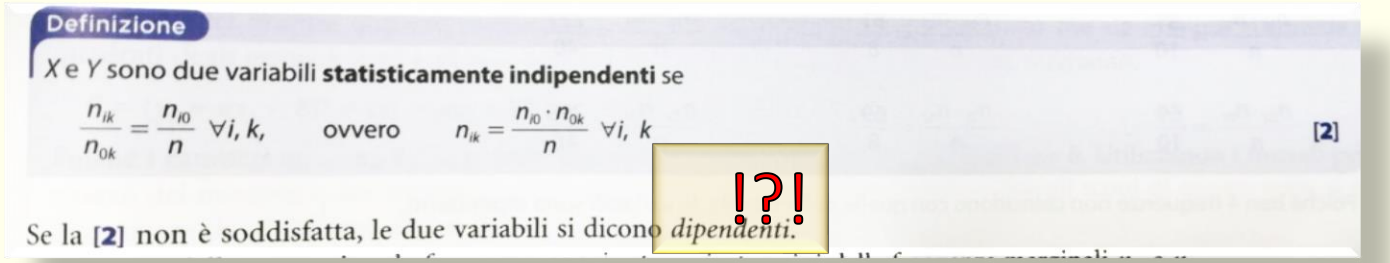
	R	S	T	U	
a	5,14	6,86	5,71	6,00	24
b	2,79	3,71	3,10	3,25	13
c	6,43	8,57	7,14	7,50	30
d	3,43	4,57	3,81	4,00	16
	18	24	20	21	83

	R	S	T	U	
a	27,34	0,00	5,71	6,00	
b	1,14	14,29	1,42	3,25	
c	6,43	0,77	19,68	0,83	
d	3,43	4,57	2,07	30,25	
					127,20

$\chi^2 = 127.20$ $\chi^2_{norm} = 0{,}511$

$V = 0{,}715$

Qualche misconcetto



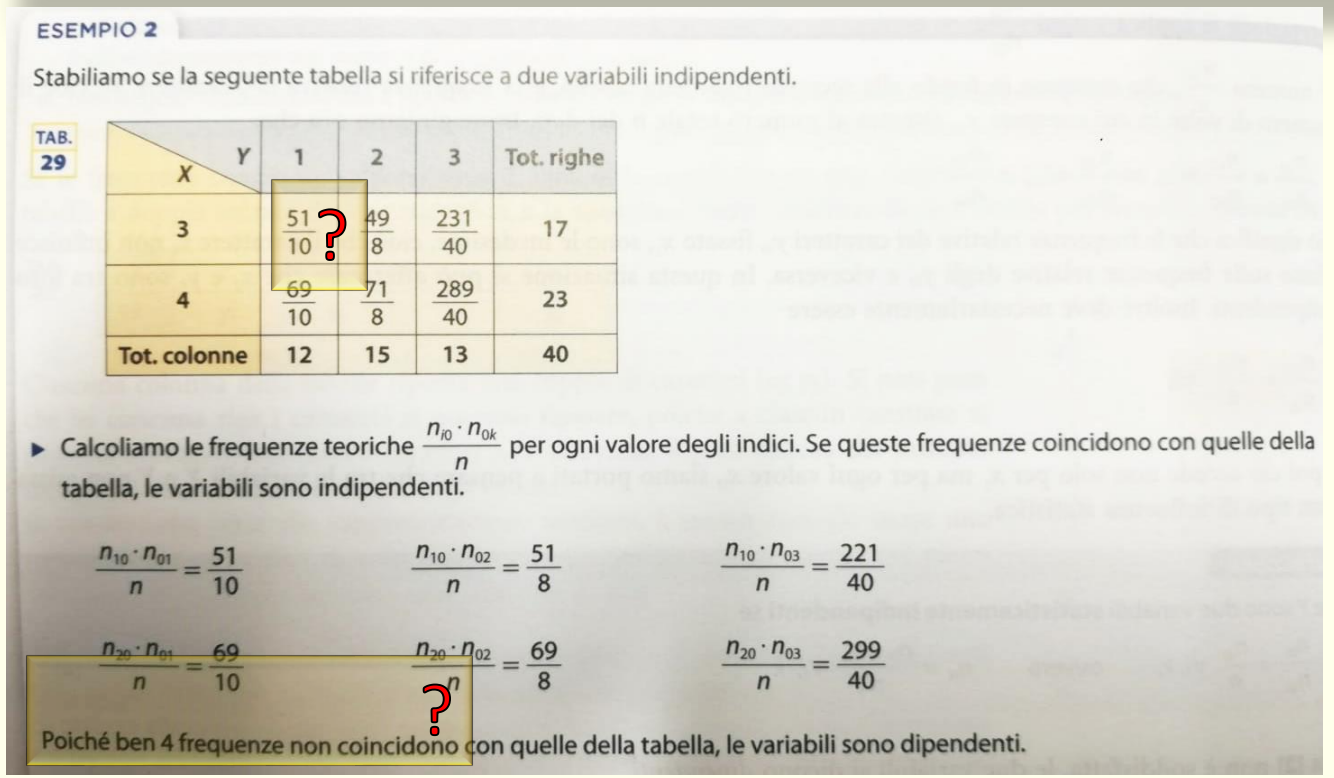
Definizione

X e Y sono due variabili **statisticamente indipendenti** se

$$\frac{n_{ik}}{n_{0k}} = \frac{n_{i0}}{n} \ \forall i, k, \quad \text{ovvero} \quad n_{ik} = \frac{n_{i0} \cdot n_{0k}}{n} \ \forall i, k \qquad [2]$$

!?!

Se la [2] non è soddisfatta, le due variabili si dicono *dipendenti*.



ESEMPIO 2

Stabiliamo se la seguente tabella si riferisce a due variabili indipendenti.

TAB. 29

X \ Y	1	2	3	Tot. righe
3	$\frac{51}{10}$?	$\frac{49}{8}$	$\frac{231}{40}$	17
4	$\frac{69}{10}$	$\frac{71}{8}$	$\frac{289}{40}$	23
Tot. colonne	12	15	13	40

▸ Calcoliamo le frequenze teoriche $\frac{n_{i0} \cdot n_{0k}}{n}$ per ogni valore degli indici. Se queste frequenze coincidono con quelle della tabella, le variabili sono indipendenti.

$$\frac{n_{10} \cdot n_{01}}{n} = \frac{51}{10} \qquad \frac{n_{10} \cdot n_{02}}{n} = \frac{51}{8} \qquad \frac{n_{10} \cdot n_{03}}{n} = \frac{221}{40}$$

$$\frac{n_{20} \cdot n_{01}}{n} = \frac{69}{10} \qquad \frac{n_{20} \cdot n_{02}}{n} = \frac{69}{8} \qquad \frac{n_{20} \cdot n_{03}}{n} = \frac{299}{40}$$

?

Poiché ben 4 frequenze non coincidono con quelle della tabella, le variabili sono dipendenti.

Il linguaggio è quello della matematica e del controesempio.

Questo errore è anche più grave perché snatura il contenuto delle celle (che è Naturale o un numero puro, ma in questo caso il totale deve essere 1).

Poiché ben 4??? Non è la quantità di celle che non coincidono ma quanto non coincidono.

Quali esercizi

Al di là di esercizi «tecnici» come quello proposto dallo stesso libro di testo che chiede di inserire i dati in modo tale da rendere massima l'indipendenza, forse accompagnerebbe più una riflessione la richiesta di inserire i dati per avere una dipendenza massima (anche dando un verso qualora i caratteri fossero ordinabili).

	R	S	T	U	
a					24
b					13
c					30
d					16
	18	24	21	20	83

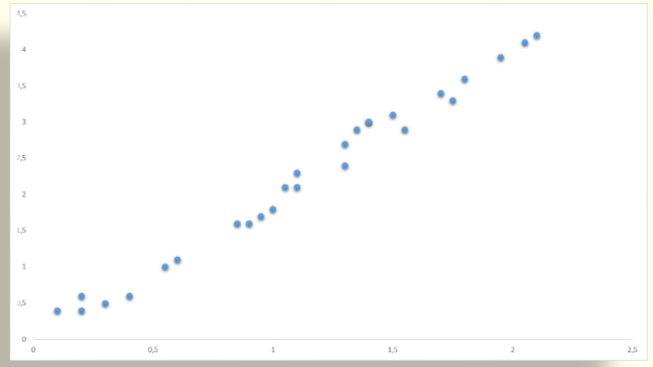
Come interpretare i grafici

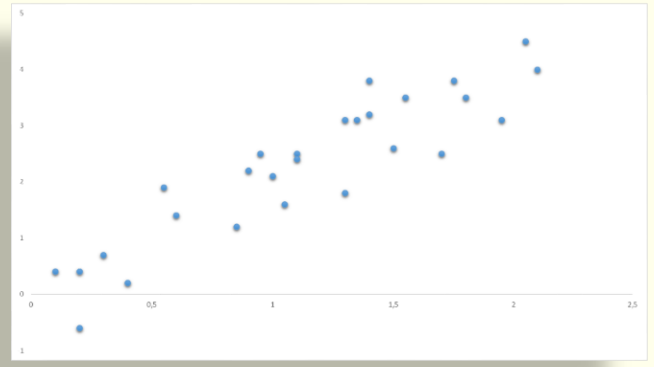
Il suggerimento in questo caso è più immediato, meno dettagliato e preciso, ma più diretto.

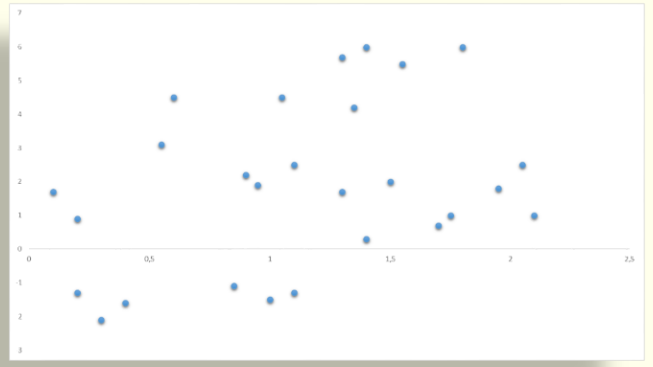
La forma della nuvola di punti è il primo aspetto che dobbiamo considerare.

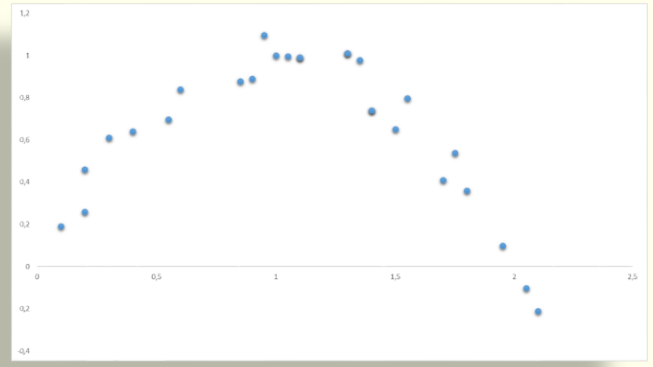
In particolare la forma distributiva ci suggerisce il modello matematico
La dispersione dei punti l'adattabilità del modello.

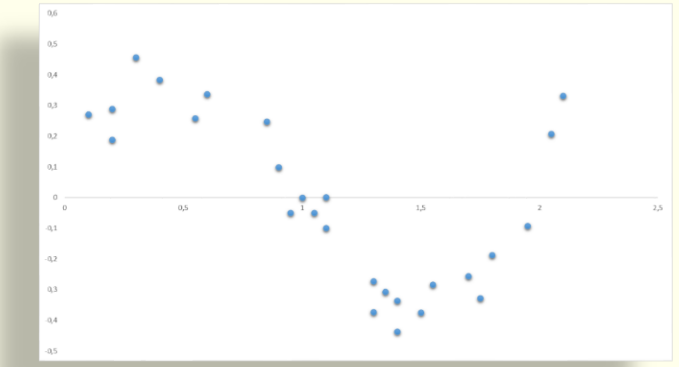
Alcune distribuzioni

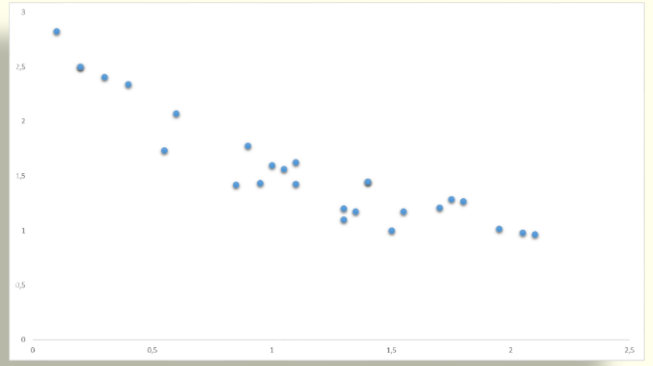












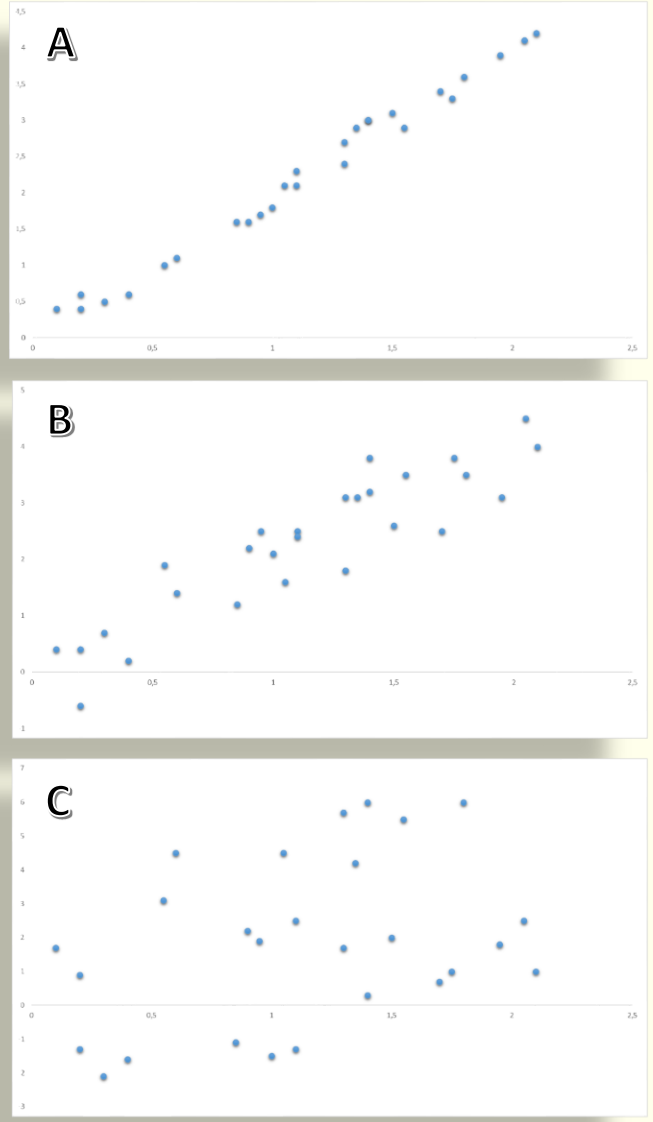
La diversa incidenza della dispersione

Consideriamo i tre grafici che hanno la stessa «ossatura», ma differiscono riguardo alla dispersione.

La domanda chiave che ci si deve porre è: fino a quanto è tollerabile la dispersione per poter disporre della relazione in modo corretto?

È la stessa domanda che dovremmo farci ogni volta che utilizziamo una media aritmetica: fino a che punto l'indicatore è potenzialmente descrittivo della centralità?

Come si può misurare la variabilità presente?



Il coefficiente di correlazione

È un esempio di una bellissima costruzione di un rapporto di composizione (un rapporto che lega una parte al tutto).

Parte dal calcolo della covarianza: essa rappresenta la media dei prodotti delle differenze di ogni coppia rilevata con la media aritmetica della distribuzione del singolo carattere.

In questo caso non possiamo chiamarle distanze perché potrebbe essere un valore negativo

Il valore ρ (o r in molti libri di testo) si chiama coefficiente di correlazione di Bravais Pearson ed è un indice variabile tra -1 e 1. I valori estremi individuano una perfetta dipendenza (diretta o inversa) mentre valori prossimi allo 0 indicano che c'è indipendenza fra i caratteri.

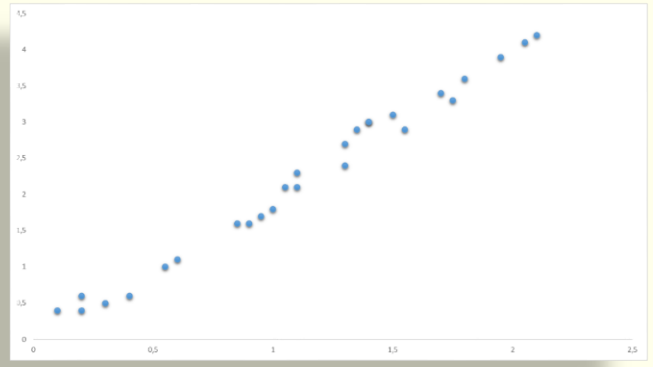
$$\sigma_{XY} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n}} \qquad \sigma_Y = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{n}}$$

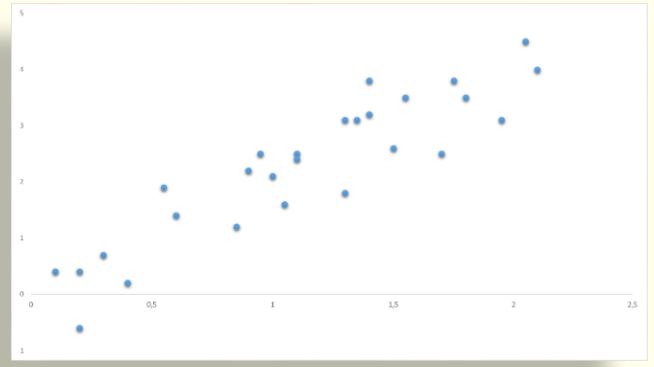
$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x}) \sum_{i=1}^{N}(y_i - \bar{y})}$$

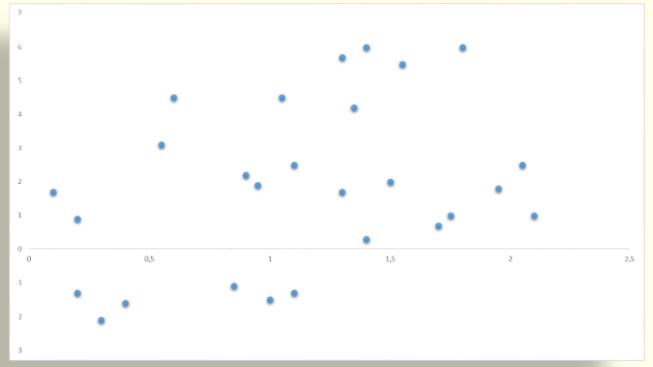
Il coefficiente di correlazione



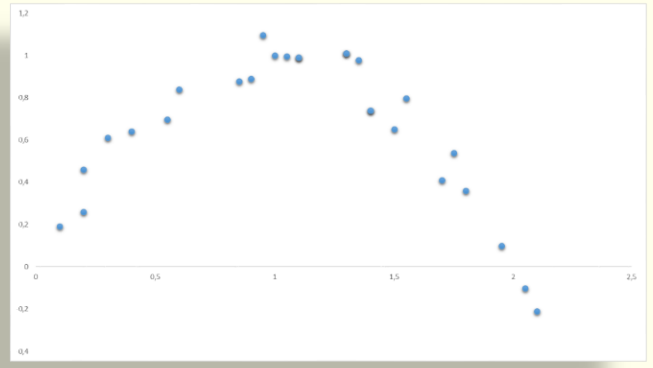
$\rho = 0{,}9925$



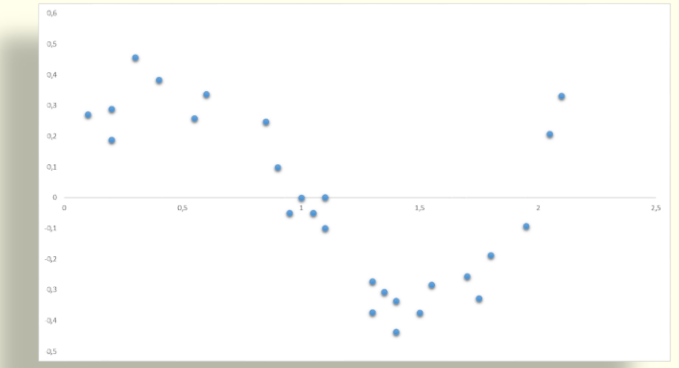
$\rho = 0{,}9097$



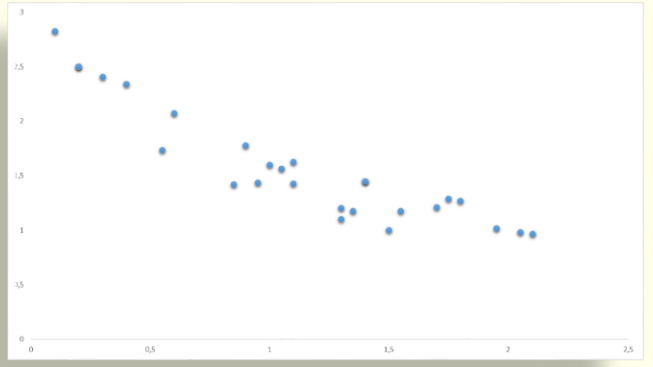
$\rho = 0{,}3563$



$\rho = -0{,}261$



$\rho = -0{,}581$



$\rho = -0{,}918$

Il modello lineare

Il coefficiente di correlazione riesce ad essere efficace solo nell'ambiente in cui si ipotizza una relazione lineare tra le variabili.

Con regressioni lineari semplici il quadrato del coefficiente di correlazione è il coefficiente di determinazione lineare il cui valore è strettamente legato alla bontà dell'adattamento della retta.

$$\rho^2 = R^2$$

In particolare possiamo sostenere che se il suo valore è superiore a 0,75 si può ritenere buono il modello lineare come spiegazione della dipendenza.

Per scelte di modelli non lineari esistono altri coefficienti di correlazione come quello di Spearman.

La regressione lineare

Nella seconda metà dell'ottocento Francis Galton osservò che la statura dei padri (X) e quella dei relativi figli (Y) era spiegabile da una dipendenza lineare.
Notò anche però che con altezze estreme dei padri (molto alti o molto bassi) le altezze dei figli tendevano ad assumere valori più vicini alla media o meglio «regredivano verso la media».
La figura di Galton è interessante: politico, cugino di Darwin, amico di Pearson, viene considerato il padre dell'eugenetica, della biometria, della psicometria.

Gli obiettivi sono gli stessi della modellizzazione: da un lato esplicativi di una relazione attraverso una causa effetto che lega due caratteri, dall'altro predittivi.

Storia

Obiettivi

- esplicativi
- predittivi

L’equazione del modello

$$y = \beta_0 + \beta_1 x + \varepsilon_i$$

Il parametro β_1 è il coefficiente di regressione, è il «traduttore» dell’unità di misura, indica di quanto varia la variabile dipendente in seguito ad una variazione unitaria della variabile indipendente.

Il parametro β_0 ha un significato statistico qualora il valore x=0 sia nel range delle osservazioni, in questo senso rappresenta il valor medio di y quando x=0, in altro casi non ha interpretazioni statistiche, è l’ordinata all’origine del modello.

Il valore di ε_i rappresenta l’errore (variabilità esogena, accidentale) dovuto agli elementi perturbatori del fenomeno che non risentono in nessun modo dei valori delle variabili, si assume che la media degli ε_i è nulla.

L'individuazione di β_1

Attraverso il metodo dei minimi quadrati si ottiene il valore di β_1

$$\beta_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Se la relazione fra le due variabili fosse simmetrica allora potremmo ricavare due coefficienti di regressione a seconda di quale variabile si consideri dipendente dall'altra.

$$\beta_{1,Y/X} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$\beta_{1,X/Y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

Mettendo in relazione il coefficiente di correlazione di Pearson con i due coefficienti di regressione si può notare che esso è la media geometrica dei due coefficienti

$$\rho = \sqrt{\beta_{1,Y|X} \cdot \beta_{1,X|Y}}$$

L’individuazione di β_0

Dopo aver individuato il coefficiente angolare e considerando il passaggio per il punto $M(\bar{x}, \bar{y})$ possiamo dedurre il valore di β_0

$$\beta_0 = \bar{y} - \beta_1 x$$

L’analisi degli errori

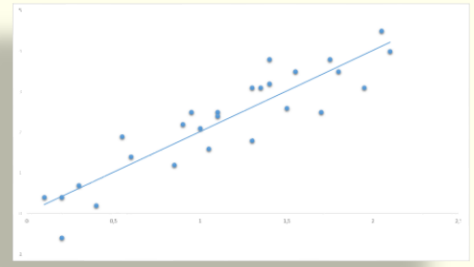
Sia il carattere X che il carattere Y hanno una loro variabilità insita nel carattere stesso.

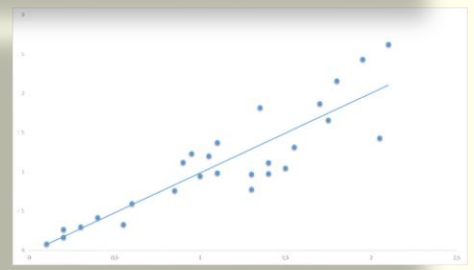
$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{Parte deterministica}} + \varepsilon_i$$

$$\hat{y} = \beta_0 + \beta_1 x$$

La variabilità che relativa all’errore rappresenta non è quella di X o di Y ma quella derivante dalla distanza tra la relazione causa effetto considerata in modo deterministico con le osservazioni realmente osservate

$$\varepsilon_i = y_i - \hat{y}_i x$$





Le ipotesi sulla distribuzione degli errori nel modello sono le seguenti:

- gli errori si distribuiscono normalmente intorno ai valori della retta;
- omoschedasticità cioè stessa incidenza della variabilità ad ogni valore del carattere dipendente;
- indipendenza degli errori dal valore del carattere X

La scomponibilità della varianza

$$\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$$

Se assumiamo la parte deterministica del modello di regressione possiamo costruire per ogni valore della X un corrispondente valore teorico $\hat{Y}$. La distribuzione dei valori teorici ha come media la stessa media della popolazione mentre la sua variabilità sarà minore perché non risente degli errori accidentali.

$$\sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$SQT = \sum_{i=1}^{N}(y_i - \bar{y}_i)^2$$

SQT rappresenta il valore della variabilità complessiva del carattere Y preso singolarmente

$$SQR = \sum_{i=1}^{N}(\hat{y}_i - \bar{y}_i)^2$$

SQR rappresenta il valore della variabilità complessiva del carattere $\hat{Y}$ è la variabilità del modello teorico.

$$SQE = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

SQE rappresenta il valore della variabilità che scaturisce dalla perturbazione accidentale del fenomeno.

La scomposizione come indicatore

$$SQR = \sum_{i=1}^{N} (\hat{y}_i - \bar{y}_i)^2$$

La quota di SQR sul totale (SQT) indica quanta variabilità totale è da ricondursi al modello e di conseguenza quanto il modello sia opportuno per la descrizione della relazione.

Si può dimostrare il rapporto tra questo indicatore e il suo totale è uguale al coefficiente di determinazione lineare

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

Come farne un uso didattico

Un docente subentra in una classe IV superiore e dopo la consegna della prima verifica dell'a.s. si scontra animatamente con gli studenti che si aspettavano una valutazione sensibilmente diversa.

È evidente un bassissimo livello di coscienziosità.

Il docente allora inizia a raccogliere ad ogni verifica le autovalutazioni degli studenti e le sue.

Calcola per ogni verifica il coefficiente di correlazione e alla fine ritiene che quando questo supera 0,9 allora si è raggiunto un buon livello di autovalutazione.