

Statistica Descrittiva III

Serie statistiche bi-variate

- Definizioni
- Rappresentazioni tabellari e grafiche
- Indici di posizione e di variabilità
- Dipendenza lineare: retta di regressione ed indice R.

Serie Bi-variate

- **Serie bi-variate**: serie statistica in cui da ogni unità statistica si rilevano due caratteri.

$$o_i = (x_i; y_i) \quad O = \{o_i\}$$

- Esempio 1:

– Popolazione: 15 studenti di matematica.

– Caratteri:

X: genere

Y: altezza

– Osservazioni:

(F; 155) (M; 173) (M; 191) (F; 163) (F; 159) (M; 178)
(M; 183) (M; 169) (F; 169) (M; 170) (M; 183) (M; 177)
(M; 175) (M; 176) (F; 170)

Serie Bi-variate: esempi - I

- Esempio 2:

- Popolazione: 16 cavie di laboratorio.

- Caratteri:

X: trattamento antibiotico

Y: stato dell'infezione

{Si ; No}

{Espansa, Stabile, Ridotta}

- Osservazioni:

$(S ; E)$ $(S ; R)$ $(S ; R)$ $(S ; R)$ $(S ; R)$ $(S ; S)$ $(S ; R)$ $(S ; S)$

$(S ; R)$ $(N ; S)$ $(N ; S)$ $(N ; E)$ $(N ; E)$ $(N ; S)$ $(N ; R)$ $(S ; E)$

- **Osservazione:** i caratteri della serie possono essere non omogenei.

Serie Bi-variate: esempi - II

- Esempio 3:

- Popolazione: 15 studenti di matematica.

- Caratteri:

X: peso

Y: altezza

- Osservazioni:

(55; 155) (78; 173) (100; 191) (60; 163) (50; 159) (78; 178)
(101; 183) (68; 169) (60; 169) (72; 170) (82; 183)
(82; 177) (75; 175) (72; 176) (65; 170)

- **Osservazione:** se i caratteri sono omogenei (es. quantitativi continui) possiamo estendere il concetto alla serie bivariata (serie quantitativa continua).

Scatter plot (diagrammi a dispersione)

- Rappresentazione grafica usata per serie quantitative, in cui ogni osservazione viene riportata in un piano cartesiano come fosse un punto.

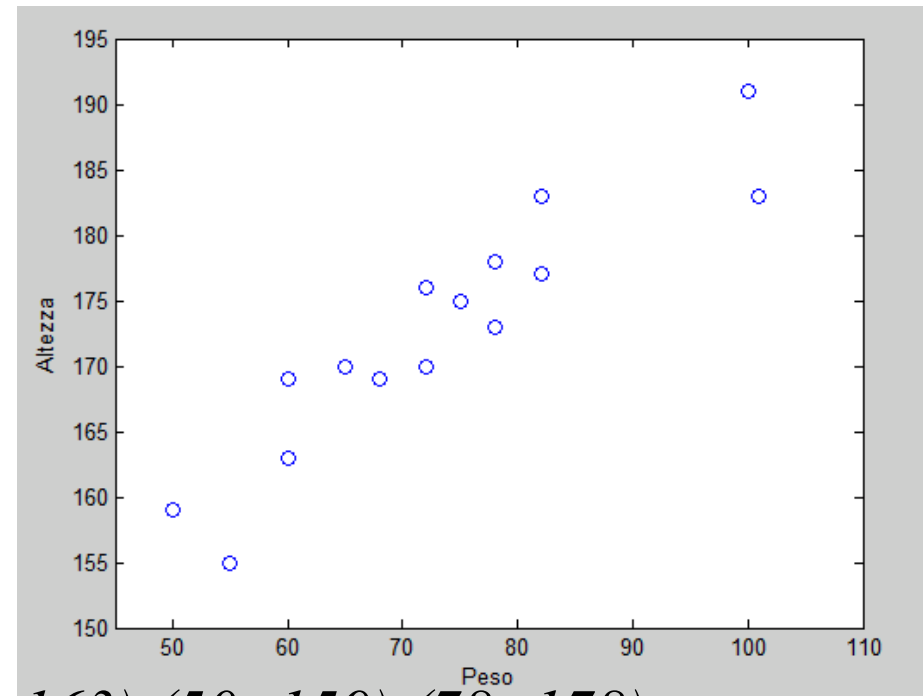
- Esempio 3:

- Caratteri:

- X: peso
 - Y: altezza

- Osservazioni:

(55; 155) (78; 173) (100; 191) (60; 163) (50; 159) (78; 178)
(101; 183) (68; 169) (60; 169) (72; 170) (82; 183) (82; 177)
(75; 175) (72; 176) (65; 170)



Modalità: definizione

- **Osservazione** il concetto di modalità è legato al carattere.
- **Problema**: è possibile estenderlo alla Serie?
 - in una monovariata serie e carattere coincidono.
 - una serie bi-variata è formata da due caratteri.
- **Conseguenza**: vi sono 3 tipi di modalità
 - M_x : modalità carattere X .
 - M_y : modalità carattere Y .
 - $M = M_x M_y$: modalità serie (# possibili coppie).

Modalità serie: esempio

- Esempio 2:

- Popolazione: 16 cavie di laboratorio.

- Modalità caratteri:

X: trattamento antibiotico

Y: stato dell'infezione

{Si; No}

{Espansa, Stabile, Ridotta}

$$M_x = 2$$

$$M_y = 3$$

- **Modalità serie:**

$$M = M_x M_y = 6$$

		Infezione		
		Espansa	Stabile	Ridotta
Trattamento	Si	(S; E)	(S; S)	(S; R)
	No	(N; E)	(N; S)	(N; R)

Modalità serie: classi

- Esempio 1:

- Popolazione: 15 studenti di matematica.

- Modalità caratteri:

X: Genere

Y: altezza

{M; F}

*{155;159;163;169;170;173;
175;176;177; 178;183;191}*

$$M_x = 2$$

$$M_y = 12$$

- **Modalità serie:**

$$M = M_x M_y = 24$$

- Osservazione: nel caso continuo spesso si introducono classi di modalità.

Modalità serie: classi

- Esempio 1:
 - Popolazione: 15 studenti di matematica.
 - Modalità caratteri:

X: Genere

Y: altezza

{M; F}

{155-165; 165-175; 175-185; 185-195}

$$M_x = 2$$

$$C_y = 4$$

- Modalità serie:

$$M = M_x C_y = 8$$

		Altezza			
		155 - 165	165 - 175	175 - 185	185-195
Gene re	M	$m_{1,1}$	$m_{1,2}$	$m_{1,3}$	$m_{1,4}$
	F	$m_{2,1}$	$m_{2,2}$	$m_{2,3}$	$m_{2,4}$

- Le diverse modalità si indicano con $m_{i,j}$.

Serie mono-variate: frequenza.

- **Frequenza**: un valore associato ad ogni modalità
 - **assoluta** (n_i): # osservazioni della modalità i .
 - **relativa** (f_i): frazione delle osservazioni della modalità i .
$$f_i = n_i / N$$
 - **cumulata** (F_i): frazione delle osservazioni delle modalità non più grandi di i . ($F_i = (n_1 + n_2 + \dots + n_i) / N$)
- Come estendere il concetto alle serie bi-variate?
- **Osservazione**: nelle bi-variate la singola modalità viene definita da due indici (i, j) .

Serie bi-variate: frequenza assoluta.

- **Frequenza:** un valore associato ad ogni modalità i, j
 - assoluta ($n_{i,j}$): # osservazioni della modalità i, j .
- **Esempio 2**
 - Osservazioni: (S;E) (S;R) (S;R) (S;R) (S;R) (S;S) (S;R)
(S;S) (S;R) (N;S) (N;S) (N;E) (N;E) (N;S) (N;R) (S;E)

		Infezione		
		Espansa	Stabile	Ridotta
Tratta mento	Si	$n_{1,1}$	$n_{1,2}$	$n_{1,3}$
	No	$n_{2,1}$	$n_{2,2}$	$n_{2,3}$

		Infezione		
		Espansa	Stabile	Ridotta
Tratta mento	Si	2	2	6
	No	2	3	1

Tabelle a doppia entrata.

- **Tabella a doppia entrata:** tabella delle frequenze assolute completata con i totali di riga e colonna.
- Osservazione: alcuni autori usano il termine tabella di contingenza.
- Esempio 2

– Osservazioni: (S;E) (S;R) (S;R) (S;R) (S;R) (S;S) (S;R)
 (S;S) (S;R) (N;S) (N;S) (N;E) (N;E) (N;S) (N;R) (S;E)

		Infezione			Totali
		Espansa	Stabile	Ridotta	
Tratta mento	Si	2	2	6	10
	No	2	3	1	6
	Tot	4	5	7	16

$n_{i,+}$: Frequenze assolute carattere trattamento

$n_{+,j}$: Frequenze assolute carattere infezione

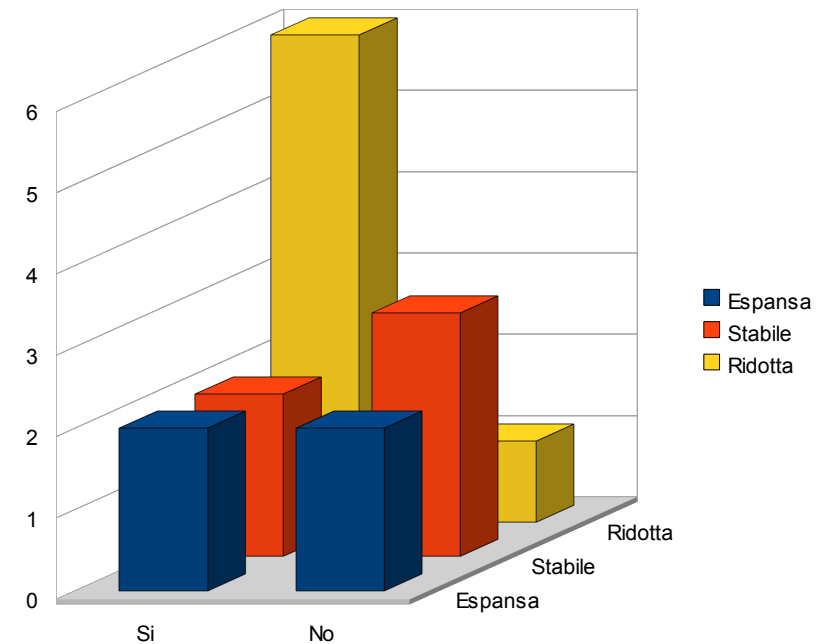
N : Totale osservazioni

Serie bi-variate: diagramma a barre

- **Diagramma a barre:** rappresentazione 3 D di frequenze assolute per serie bi-variate.
- **Osservazione:** usato serie a caratteri qualitativi e/o discreti.
- **Esempio 2**

– Osservazioni: (S;E) (S;R) (S;R)
 (S;R) (S;R) (S;S) (S;R) (S;S)
 (S;R) (N;S) (N;S) (N;E) (N;E)
 (N;S) (N;R) (S;E)

		Infezione		
		Espansa	Stabile	Ridotta
Trattamento	Si	2	2	6
	No	2	3	1



Serie bi-variate: frequenza assoluta.

- **Frequenza:** un valore associato ad ogni modalità i, j
 - **assoluta** ($n_{i,j}$): # osservazioni della modalità i,j .
 - **relativa** ($f_i = n_{i,j} / N$): frazione osservazioni modalità i,j .
- **Esempio 2**
 - Osservazioni: (S;E) (S;R) (S;R) (S;R) (S;R) (S;S) (S;R) (S;S)
(S;R) (N;S) (N;S) (N;E) (N;E) (N;S) (N;R) (S;E)

		Infezione			
		Espansa	Stabile	Ridotta	
Tratta mento	Si	$f_{1,1}$	$f_{1,2}$	$f_{1,3}$	$f_{1,+}$
	No	$f_{2,1}$	$f_{2,2}$	$f_{2,3}$	$f_{2,+}$
		$f_{+,1}$	$f_{+,2}$	$f_{+,3}$	1

		Infezione			
		Espansa	Stabile	Ridotta	
Tratta mento	Si	2/16	2/16	6/16	5/8
	No	2/16	3/16	1/16	3/8
		1/4	5/16	7/16	1

Serie mono-variate: indici di posizione.

- **Indici di posizione:** il valore “centrale” della serie S .
 - **moda** (mo): modalità con frequenza maggiore.
(tutti i dati – per dati continui si parla di classe/i modale)
 - **mediana** (me): osservazione che bipartisce la popolazione
(solo dati ordinabili)
 - **media** (\bar{S}): media delle osservazioni.
(solo dati quantitativi)

Come estendere il concetto alle serie bi-variate?

- **Osservazione:** in una bi-variata le osservazioni sono vettori.
- **Proposta:** cerco di estendere al caso multidimensionale il criterio di calcolo.
- **Osservazione:** debbo scegliere un indice che sia calcolabile per ambo i caratteri.

Serie bi-variate: moda.

- **Indici di posizione:** il valore “centrale” della serie S .
 - **moda** (mo): modalità con frequenza maggiore.
(tutti i dati – per dati continui si parla di classe/i modale)
- In una serie bi-variata ho definito sia modalità che frequenza
- Posso applicare direttamente la definizione

- Esempio 1

$mo = (Si ; Ridotta)$

		Infezione		
		Espansa	Stabile	Ridotta
Trattamento	Si	2	2	6
	No	2	3	1

- Esempio 2

$mo = (M ; 175-185)$

		Altezza			
		155 - 165	165 - 175	175 - 185	185-195
Genere	F	3	2	0	0
	M	0	3	6	1

Serie bi-variate: moda - osservazioni.

- Un una serie bivariata vi sono tre diverse mode.
- Esempio: data la serie a lato si hanno le seguenti

- Moda della Serie

$$mo = (C ; 10)$$

- Moda caratteri

- $mo_X = A$

- $mo_Y = 30$

		X			Totali
		A	B	C	
Y	10	1	2	6	9
	20	4	3	1	8
	30	5	3	2	10
Totali		10	8	9	27

- **Osservazione:** la moda di una serie bi-variate può non coincidere con la moda dei singoli caratteri.

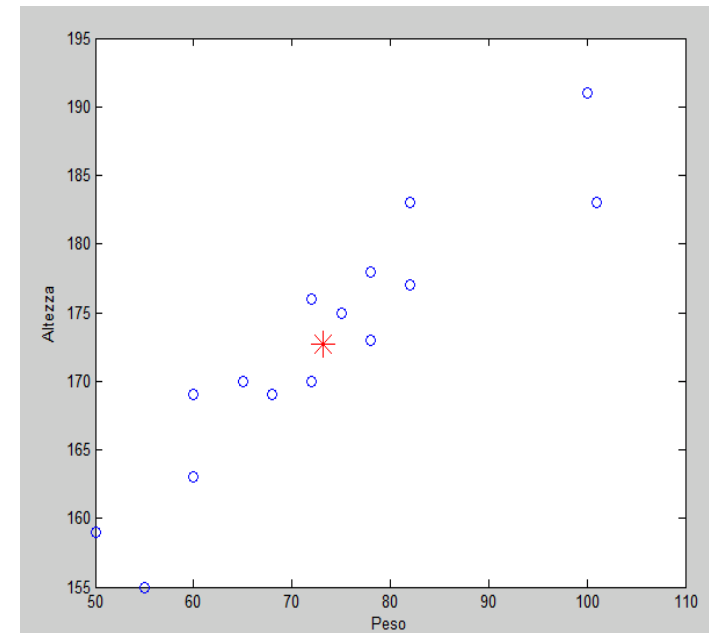
Serie bi-variate: mediana.

- **Indici di posizione:** il valore “centrale” della serie S .
 - **moda** (mo): modalità con frequenza maggiore.
(tutti i dati – per dati continui si parla di classe/i modale)
 - **mediana** (me): osservazione che bipartisce la popolazione
(solo dati ordinabili)
- **Osservazione:** le osservazioni sono vettori a 2 dimensioni.
- **Osservazione:** non è possibile ordinare in modo univoco vettori a due dimensioni.
- **Osservazione:** la mediana richiede l'ordinamento delle osservazioni.
- **Conseguenza:** la mediana non si applica a serie bi-variate.

Serie bi-variate: indici di posizione.

- **Indici di posizione:** il valore “centrale” della serie S .
 - **media** (\bar{S}): media delle osservazioni.
(solo dati quantitativi)
- Per applicare la definizione debbo
 - sommare più vettori
 - dividere un vettore per un numero.
- **Esempio 3**

$$\begin{aligned} & ((55; 155) + (78; 173) + \dots + (65; 170))/15 = \\ & = (55 + 78 + \dots + 65; 155 + 173 + \dots + 170)/15 = \\ & = ((55 + 78 + \dots + 65)/15; (155 + \dots + 170)/15) = \\ & = (73.20; 172.73) \end{aligned}$$



- **Osservazione:** la media della serie è data dalle medie dei caratteri

$$\bar{S} = (\bar{X}; \bar{Y})$$

Serie mono-variate: indici di variabilità.

- **Indici di variabilità:** quanto le osservazioni si discostano dal “centrale” della serie S .
 - **range** : osservazione maggiore meno osservazione minore
 - **distanza interquartile** (D): differenza terzo, primo quartile
 - **varianza** (σ^2): media degli scarti dalla media
- **Osservazione:** calcolabili solo per dati quantitativi.

Come estendere il concetto alle serie bi-variate?

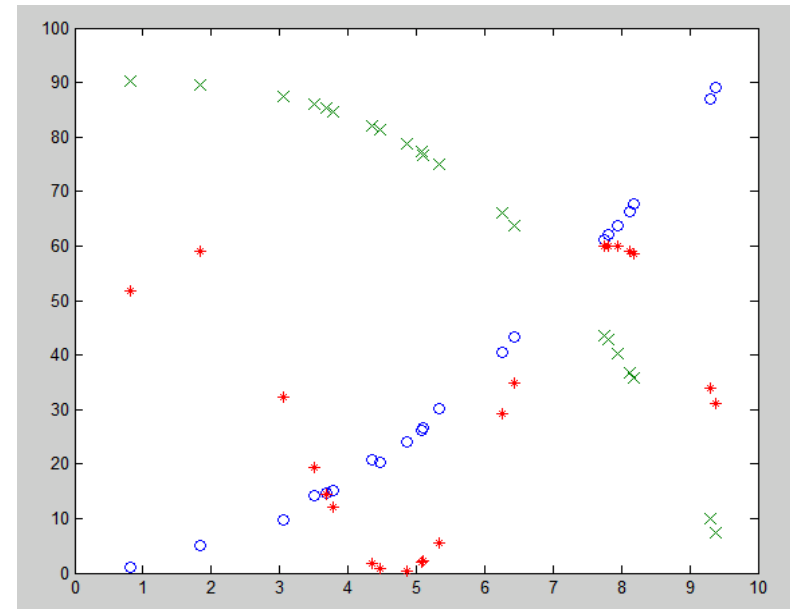
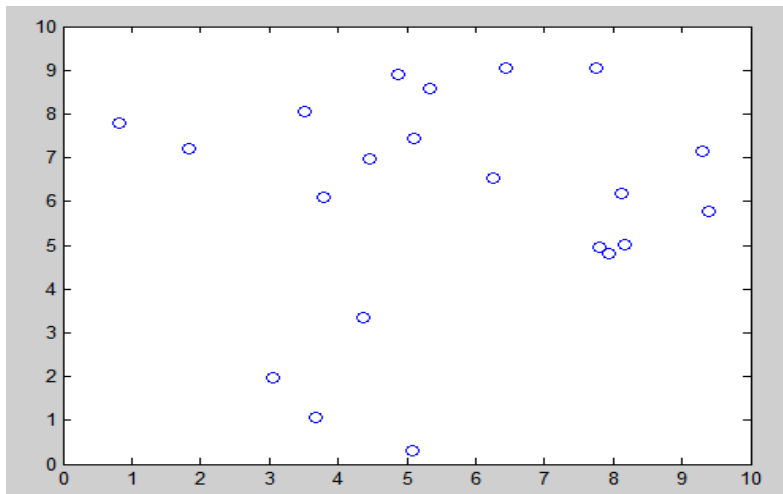
- **Osservazione:** non è possibile ordinare un gruppo di vettori a due dimensioni.
- **Proposta:** estendo solo la varianza.

Serie bi-variate: indice di variabilità-I.

- **Osservazione:** la variabilità in una serie bivariata può essere dovuta a tre fattori
 - Variabilità carattere X .
 - Variabilità carattere Y .
 - Mutua influenza dei due valori.
- **Esempio:**
 - Popolazione 4 persone
 - Caratteri
 - X = Altezza della persona in cm
 - Y = Altezza della persona in m
 - Osservazioni (150; 1.5) (170;1.7) (180;1.8) (160;1.6)
- Nota l'osservazione di X "ho" quella di Y , pertanto la variabilità di Y è legata a quella di X .

Mutua influenza: classificazione

- Due caratteri possono essere legati in diversi modi
 - Proporzionalità diretta:
al crescere di x
cresce y . (o)
 - Proporzionalità inversa:
al crescere di x
decrece y . (x)
 - Legame misto. (*)



- Due caratteri possono essere indipendenti

Covarianza: idea base.

- Come misuro il “grado di legame” fra i due caratteri ?
- **Proposta:** data un'osservazione $(x_i; y_i)$ valuto il prodotto dei due scarti (distanza dai rispettivi valori medi \bar{x} e \bar{y}).

$$(x_i - \bar{x})(y_i - \bar{y})$$

- **Osservazione:** il singolo prodotto è:
 - positivo: gli scarti sono concordi (hanno lo stesso segno)
proporzionalità diretta: ad una osservazione di x sopra media corrisponde una y sopra media e viceversa
 - negativo in caso contrario (proporzionalità inversa).
- Se i due caratteri sono legati, tutti i contributi sono concordi.
- Sperabilmente, se i caratteri sono indipendenti, si avranno sia contributi negativi che positivi.

Covarianza: definizione I.

- **Definizione:** data un serie bi-variata avente N osservazioni $(x_i; y_i)$ definisco covarianza

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

- Esempio
 - Osservazioni (150; 1.5) (170;1.7) (180;1.8) (160;1.6)
 - Media = (165,1.65)
 - Varianza $X = (150^2 + 170^2 + 180^2 + 160^2) / 4 - 165^2 = 125$
 - Varianza $Y = (1.5^2 + 1.7^2 + 1.8^2 + 1.6^2) / 4 - 1.65^2 = 0.0125$
 - Contributi = (-15*-0.15) (5*0.05) (15*0.15) (-5*-0.05)
 - Covarianza = (2.25 + 0.25 + 2.25 + 0.25)/4 = 1.25
- **Osservazione:** la definizione non usa le frequenze relative

Covarianza: definizione II.

- Definizione:** data un serie bi-variata $(x_i ; y_j)$ avente $M = M_x M_y$ modalità, con frequenza relativa $f_{i,j}$ definisco covarianza

$$\sigma_{XY} = \sum_{i=1}^{M_x} \sum_{j=1}^{M_y} (x_i - \bar{x})(y_j - \bar{y}) f_{i,j}$$

- Esempio 4

- S: numero di teste lanciando due monete
- N = 80

		Lancio 1 (X)		
		0	1	
Lancio 2 (Y)	0	22/80	20/80	42/80
	1	18/80	20/80	38/80
		40/80	40/80	1

$$\bar{x} = 0 \frac{40}{80} + 1 \frac{40}{80} = \frac{1}{2} \quad \bar{y} = 0 \frac{42}{80} + 1 \frac{38}{80} = \frac{19}{40} \quad \sigma_{XY} = \sum_{i=1}^2 \sum_{j=1}^2 (x_i - \frac{1}{2})(y_j - \frac{19}{40}) f_{i,j}$$

$$\sigma_{XY} = (0 - \frac{1}{2})(0 - \frac{19}{40}) \frac{22}{80} + (0 - \frac{1}{2})(1 - \frac{19}{40}) \frac{18}{80} + (1 - \frac{1}{2})(0 - \frac{19}{40}) \frac{20}{80} + (1 - \frac{1}{2})(1 - \frac{19}{40}) \frac{20}{80}$$

$$\sigma_{XY} = \frac{19}{40} \frac{22}{80} - \frac{1}{40} \frac{18}{80} - \frac{19}{40} \frac{20}{80} + \frac{1}{40} \frac{22}{80} = 0.013125$$

Serie bi-variate: indice di variabilità-II.

- **Osservazione:** la variabilità in una serie bivariata può essere dovuta a tre fattori
 - Variabilità carattere X .
 - Variabilità carattere Y .
 - Mutua influenza dei due valori.
- la variabilità viene definita mediante la seguente matrice detta matrice varianze/covarianza:

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}$$

- **Osservazione:** Σ è sempre simmetrica.
- **Osservazione:** si dimostra che Σ è sempre definita positiva.

Covarianza: osservazione.

Cosa succede se aumento la variabilità dei caratteri?

- **Desiderio:** σ_{XY} dovrebbe descrivere solo la mutua variabilità fra i caratteri quindi vorrei non variasse.
- **Verifica:** modifico la variabilità delle precedenti osservazioni.
 - Precedenti [150; 1.5] [170;1.7] [180;1.8] [160;1.6]
 - Osservazioni (140; 1.4) (170;1.7) (190;1.9) (160;1.6)
 - Media = (165,1.65) [(165,1.65)]
 - Varianza $X = 325$ [125]
 - Varianza $Y = 0.0325$ [0.0125]
 - Contributi = (-25*-0.25) (5*0.05) (25*0.25) (-5*-0.05)
 - Covarianza = (6.25 + 0.25 + 6.56 + 0.25)/4 = 3.25 [1.25]
- **Osservazione:** il valore della covarianza è legato alla variabilità.

Correlazione (di Pearson): definizione.

- **Definizione:** data un serie bi-variata $(x_i ; y_i)$ definisco correlazione (di Pearson) il valore

$$R = \frac{\sigma_{XY}}{\sigma_x \sigma_y}$$

- **Osservazione:** spesso di una serie dispongo della varianza invece che della deviazione standard.
- Alcuni autori preferiscono utilizzare la versione “quadrata” del coefficiente di correlazione per motivi pratici.

$$R^2 = \frac{\sigma_{XY}^2}{\sigma_x^2 \sigma_y^2}$$

Correlazione (di Pearson): esempi - I.

- Esempio:

- Osservazioni (150; 1.5) (170;1.7) (180;1.8) (160;1.6)

- Varianza $X = 125$

- Varianza $Y = 0.0125$

- Covarianza = 1.25

$$R^2 = \frac{\sigma_{XY}^2}{\sigma_x^2 \sigma_y^2} = \frac{1.25^2}{125 \cdot 0.00125} = 1$$

- Esempio modificato

- Osservazioni (140; 1.4) (170;1.7) (190;1.9) (160;1.6)

- Varianza $X = 325$

- Varianza $Y = 0.0325$

- Covarianza = 3.25

$$R^2 = \frac{\sigma_{XY}^2}{\sigma_x^2 \sigma_y^2} = \frac{3.25^2}{325 \cdot 0.00325} = 1$$

- **Osservazione:** R sembra meno influenzato dalla variabilità

Correlazione (di Pearson): esempi - II.

- **Osservazione:** la definizione di R si applica direttamente anche al caso avessi frequenze relative.

- Esempio 4

– S : numero di teste lanciando due monete

$$\bar{x} = \frac{1}{2} \quad \bar{y} = \frac{19}{20} \quad \sigma_{XY} = 0.013125$$

		Lancio 1 (X)		
		0	1	
Lancio 2 (Y)	0	22/80	20/80	42/80
	1	18/80	20/80	38/80
		40/80	40/80	1

$$\sigma_x^2 = 0^2 \frac{1}{2} + 1^2 \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4} \quad \sigma_y^2 = 0^2 \frac{21}{40} + 1^2 \frac{19}{40} - \left(\frac{19}{40}\right)^2 = \frac{399}{1600} \approx \frac{1}{4}$$

$$R = \frac{\sigma_{XY}}{\sqrt{\sigma_x^2 \sigma_y^2}} = 0.0525$$

- **Osservazione:** a bassi valori di R (o R^2) sembrano corrispondere legami deboli fra i caratteri.

Correlazione (di Pearson): proprietà.

- Si dimostrano le seguenti proprietà:
 - a) $-1 \leq R \leq 1$
 - b) Se per ogni osservazione $y_i = a x_i + b$ allora $|R| = 1$.
 - c) Se $(x; y)$ sono ottenute da realizzazioni di vv. cc. X ed Y indipendenti allora
 - R è una v.c.
 - $E[R] = 0$.
- **Osservazione:** la proprietà b) non implica che se $R = \pm 1$ i caratteri della serie sono affini (o in ogni caso legati).
- **Osservazione:** la proprietà c) non implica che se $R = 0$ i caratteri della serie sono indipendenti.
- **Deduzione:** R è un indicatore, non fornisce alcuna certezza.

Variabili dipendenti: esempio I - I.

- Esempio:

- Siano X ed Y due vv. cc. tali che $Y = 2X^2$

- Supponiamo che

- si abbiano 4 estrazioni di

- $x_1 = 1$; $x_2 = 2$; $x_3 = -2$; $x_4 = -1$

- cui corrispondono 4 osservazioni della bivariata (x_i, y_i)

- $(1;2)$ $(2;4)$ $(-2;4)$ $(-1;2)$

- Come sarà il coefficiente R^2 ?

- **Osservazione:** Le osservazioni sono dipendenti.

- **Osservazione:** Il legame fra le osservazioni è di tipo misto.

- **Deduzione:** ?

Variabili dipendenti: esempio I - II

- Osservazioni: $(1;2)$ $(2;4)$ $(-2;4)$ $(-1;2)$
- Calcolo momenti dei singoli caratteri

$$\bar{x} = \frac{1+2-2-1}{4} = 0 \quad \sigma_x^2 = \frac{1+4+4+1}{4} - 0^2 = \frac{5}{2}$$

$$\bar{y} = \frac{2+4+4+2}{4} = 3 \quad \sigma_y^2 = \frac{4+16+16+4}{4} - 3^2 = 1$$

- Calcolo covarianza

$$\sigma_{XY} = \frac{1}{4} [(1-0)(2-3) + (2-0)(4-3) + (-2-0)(4-3) + (-1-0)(2-3)]$$

$$\sigma_{XY} = \frac{-1+2-2+1}{4} = 0 \quad R^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} = \frac{0^2}{\frac{5}{2} \cdot 1} = 0$$

- **Osservazione:** nel caso di variabili con legame non lineare R può dare delle false certezze.

Variabili dipendenti: esempio II - I.

- Esempio 5:

- Siano X ed Y due vv. cc. tali che $Y = 2X + 0.5Z$
- Dove Z è la v.c. Normale standard
- Supponiamo che

- si abbiano 4 estrazioni di X e Z

$$(x_1 = 1; z_1 = 0.1) ; (x_2 = 2 ; z_2 = -0.1) (x_3 = -2 ; z_3 = -0.2) (x_4 = -1 ; z_4 = 0.2)$$

- cui corrispondono 4 osservazioni della bivariata (x_i, y_i)
 $(1; 2.05) (2; 3.95) (-2; -4.1) (-1; -1.9)$

- Come sarà il coefficiente R^2 ?
- **Osservazione:** Le osservazioni sono dipendenti
- **Osservazione:** L'influenza di X su Y è prevalente rispetto a quella di Z su Y . (Z può considerarsi un rumore o disturbo)
- **Osservazione:** Il legame fra le osservazioni è di tipo misto.

Variabili dipendenti: esempio II - II

• Osservazioni: $(1; 2.05)$ $(2; 3.95)$ $(-2; -4.1)$ $(-1; -1.9)$

• Calcolo momenti dei singoli caratteri

$$\bar{x} = \frac{1+2-2-1}{4} = 0 \qquad \bar{y} = \frac{2.05+3.95-4.1-1.9}{4} = 0$$

$$\sigma_x^2 = \frac{1+4+4+1}{4} - 0^2 = \frac{5}{2} \qquad \sigma_y^2 = \frac{4.2025+15.6025+16.81+4.41}{4} - 0^2 = 10.05625$$

• Calcolo covarianza

$$\sigma_{XY} = \frac{1}{4} [(1-0)(2.05-0) + (2-0)(3.95-0) + (-2-0)(-4.1-0) + (-1-0)(-1.9-0)]$$

$$\sigma_{XY} = \frac{2.05+7.9+8.2+1.9}{4} = \frac{20.05}{4} = 5.0125$$

• Calcolo correlazione

$$R^2 = \frac{\sigma_{XY}^2}{\sigma_x^2 \sigma_y^2} = \frac{5.0125^2}{2.5 \cdot 10.05625} = 1$$

Legami fra caratteri.

Come riconoscere in maniera quantitativa il legame fra i due caratteri di una bivariata?

- **Idea:** scelgo un tipo di legame utile e cerco quello che meglio approssima i dati.
- Per poter concretizzare l'idea debbo definire
 - Che cosa sia un legame
 - Che cosa voglia dire un legame utile
 - Come misuro la bontà dell'approssimazione.

Legami fra caratteri: modelli.

Come descrivere in maniera quantitativa un legame fra i due caratteri?

- **Osservazione:** ogni dato è una coppia di misure (x ed y).
- In “matematica” un legame fra due elementi di un insieme è definito funzione.
- **Soluzione:** considero legame una funzione $m: X \rightarrow Y$.
- Ad ogni misura x_i di X il legame “predice” una possibile misura $\hat{y}_i = m(x_i)$ chiamata **predizione**.
- Poiché $m(\cdot)$ “descrive” il legame viene chiamato **modello**.

Modelli: tassonomia.

Quanti modelli esistono?

- **Osservazione:** il numero di funzioni possibili è enorme
- Il problema di identificazione è troppo complesso!
- **Osservazione:** esistono diversi tipi di funzioni
 - Polinomiali
 - Esponenziali
 - Trigonometriche
 - ...
- Spezzo il problema in due:
 - Trovo il tipo di funzione (scelta umana)
 - es. polinomi di secondo grado
 - Trovo la funzione “migliore” fra le scelte. (data driven)
 - es. $\hat{y}_i = 14 x_i^2 + 12 x_i + 5$

Modelli: cosa vuol dire migliore? .

- In generale si predilige la semplicità.
- Modelli “semplici”
 - Affine $\hat{y}_i = a x_i + b$
 - Parabolico $\hat{y}_i = a x_i^2 + b x_i + c$
- **Osservazione:** ogni famiglia di modelli possiede un insieme di parametri
 - affine: $\{a, b\}$
 - parabolico: $\{a, b, c\}$.
- Osservazione: scelto il tipo di modello, la relazione viene definita una volta **identificati i parametri**.
- I parametri si identificano con un criterio di “bontà”.

Criterio di bontà di un modello - I.

Come si quantifica la bontà di un modello?

- **Osservazione:** poiché un modello realizza predizioni un criterio ragionevole è che le predizioni siano accurate.

$$\hat{y}_i = y_i$$



$$m(x_i) = y_i$$

- **Esempio 5: (errori di misura)**

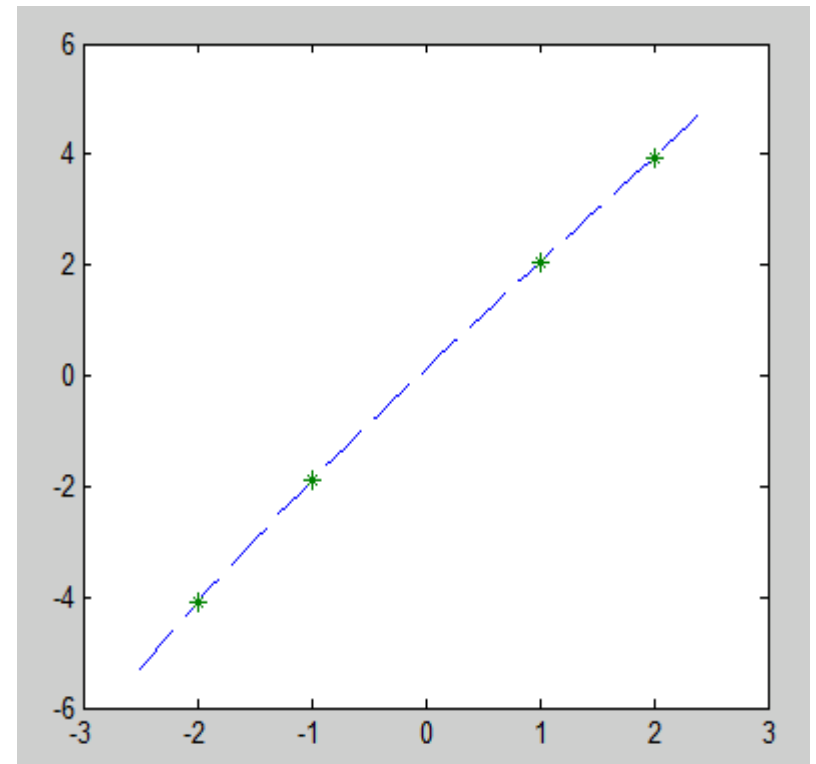
$$(x_1 = 1; y_1 = 2.05) ; (x_2 = 2 ; y_2 = 3.95)$$

$$(x_3 = -2 ; y_3 = -4.1); (x_4 = -1 ; y_4 = -1.9)$$

- **Famiglia: polinomi**

– Polinomio interpolante

$$\hat{y}_i = 0.0125x_i^3 - 0.05x_i^2 + 1.9625x_i + 0.1250$$



Funzione interpolante: osservazioni

- Si dimostra che dati N punti la funzione interpolante ha (nel caso generale) grado $N-1$.
- **Osservazione:** per avere risultati attendibili si usano tanti dati
- **Conseguenza:** modello interpolante complesso.
- **Osservazione:** nel caso di errori di misura la funzione interpolante descrive anche il rumore (cosa che non vorrei)
- **Conseguenza:** si preferiscono modelli di grado basso
- Si dice che le variabili “regrediscono” a
 - Rette (retta di regressione)
 - Parabole (parabola di regressione)
- **Considerazione:** serve un nuovo criterio per individuare il modello migliore nella famiglia di modelli.

Criterio di bontà di un modello - II.

- Dato un modello $m(x_i)$, definisco residuo

$$r_i = \hat{y}_i - y_i$$

- **Osservazione:** Un buon modello avrà dei residui non nulli (mi rappresentano l'errore di misura).
- **Osservazione:** Un buon modello avrà residui piccoli. (deve descrivere i dati)
- **Proposta:** i due risultati potrebbero essere ottenuti minimizzando la somma dei quadrati dei residui.

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

- SSR: sum squared residual

SSR: esempio

- Esempio: Data la bi-variata in tabella

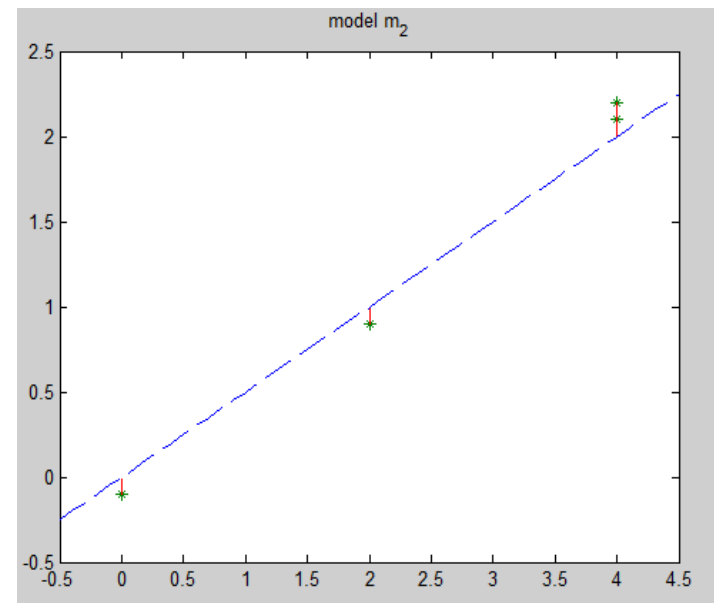
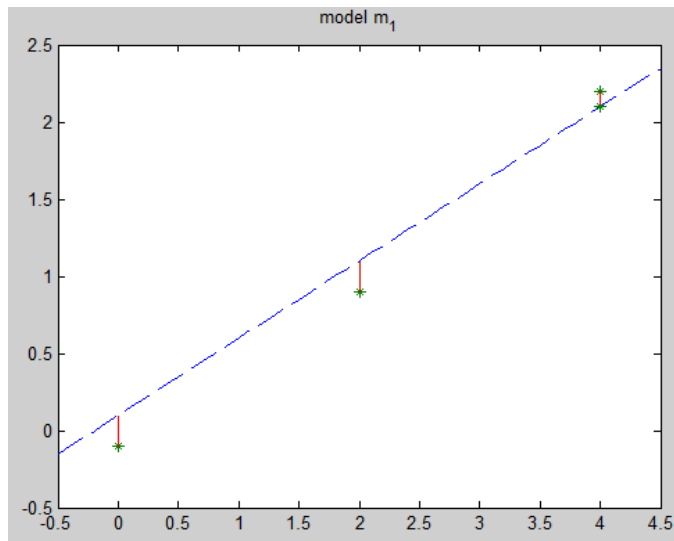
- Considerare i modelli

- $m_1: \hat{y}_i = 0.5 x_i + 0.1$

- $m_2: \hat{y}_i = 0.5 x_i$

- graficamente è r_i il segmento rosso

		m_1			m_2		
x_i	y_i	$m_1(x_i)$	r_i	r_i^2	$m_2(x_i)$	r_i	r_i^2
0	-0.1	0.1	0.2	0.04	0	0.1	0.01
4	2.1	2.1	0	0	2	0.1	0.01
4	2.2	2.1	-0.1	0.01	2	0.2	0.04
2	0.9	1.1	0.2	0.04	1	-0.1	0.01
Totali			0.3	0.09		0.3	0.07



- Il modello m_2 risulta migliore

Retta di regressione.

- Tecnica che prevede
 - Modello di tipo affine $m_1: \hat{y}_i = a x_i + b$
 - Parametri identificati rendendo minimo

$$SSR = \sum_{i=1}^N (\hat{y}_i - y_i)^2 = \sum_{i=1}^N (a x_i + b - y_i)^2$$

- In simboli $\arg \min_{a,b} \sum_{i=1}^N (a x_i + b - y_i)^2$

- Si dimostra che $a = \frac{\sigma_{xy}}{\sigma_x^2}$ $b = \bar{y} - a \bar{x}$

- La minimizzazione della SSR fu proposta da Gauss con il nome di **metodo dei minimi quadrati**.

Retta di regressione: esempio

- Esempio 5: (errori di misura)

$$(x_1 = 1; y_1 = 2.05) ; (x_2 = 2 ; y_2 = 3.95)$$

$$(x_3 = -2 ; y_3 = -4.1); (x_4 = -1 ; y_4 = -1.9)$$

– Da esercizio precedente

$$\bar{x} = 0 \quad \bar{y} = 0 \quad \sigma_{xy} = 5.0125$$

$$\sigma_x^2 = 2.5 \quad \sigma_y^2 = 10.05625$$

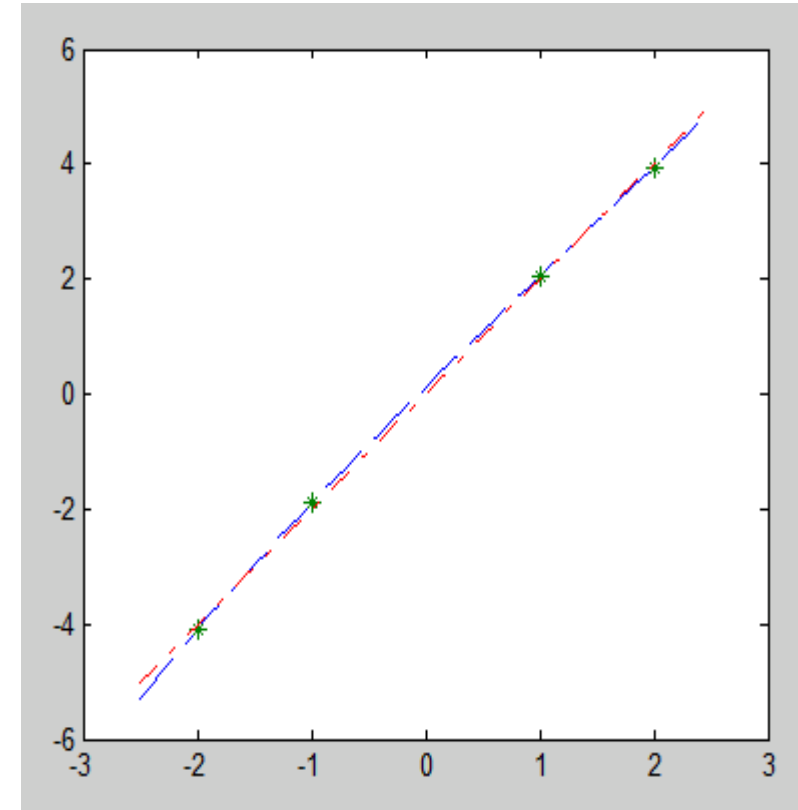
$$a = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{5.0125}{2.5} = 2.005$$

$$b = \bar{y} - a \bar{x} = 0$$

– retta di regressione $\hat{y}_i = 2.005 x_i$

– funzione interpolante $\hat{y}_i = 0.0125x_i^3 - 0.05x_i^2 + 1.9625x_i + 0.1250$

– legame vero $y_i = 2x_i + 0.5z_i$



Retta di regressione: proprietà

- Sia $m(x)$ una retta di regressione, si dimostra che
 - m passa sempre per la media della serie

$$\bar{y} = m(\bar{x})$$

- La somma degli scarti è nulla

$$\sum_{i=1}^N \hat{y}_i - y_i = \sum_{i=1}^N (a x_i + b - y_i) = 0$$

- Se $y_i = a x_i + b$ (osservazioni disposte su una retta) allora
 - $m(x)$ è la retta interpolante
 - $SSR = 0$.

Retta di regressione: validità

- **Osservazione:** date le osservazioni di una bivariata si può sempre calcolare la retta di regressione.

Come verificare la bontà del modello ?

- **Criterio grafico:** disegnare la retta insieme al diagramma a dispersione
- **Criterio numerico:** calcolare il coefficiente di Pearson
 - $|R| < 0.3$ scarsa probabilità di legame
 - $0.3 < |R| < 0.7$ moderata probabilità di legame lineare
 - $0.7 < |R|$ ottima probabilità di legame lineare

Ricapitolando - I

- Indici di posizione
 - Moda: modalità frequenza della serie maggiore
 - Media: unione delle medie delle modalità $\bar{S}=(\bar{X}; \bar{Y})$
- Indici di variabilità
 - Covarianza (varianza congiunta)
 - $\sigma_{XY} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$
 - $\sigma_{XY} = \sum_{i=1}^{M_x} \sum_{j=1}^{M_y} (x_i - \bar{x})(y_j - \bar{y}) f_{i,j}$
 - Matrice varianza/covarianza

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}$$

Ricapitolando - II

- Retta di regressione

- Modello di tipo affine $m: \hat{y}_i = a x_i + b$
- Parametri identificati rendendo minimo

$$SSR = \sum_{i=1}^N (\hat{y}_i - y_i)^2 = \sum_{i=1}^N (a x_i + b - y_i)^2$$

- Parametri $a = \frac{\sigma_{xy}}{\sigma_x^2}$ $b = \bar{y} - a \bar{x}$

- Indice di correlazione di Pearson

$$R = \frac{\sigma_{XY}}{\sigma_x \sigma_y} \qquad R^2 = \frac{\sigma_{XY}^2}{\sigma_x^2 \sigma_y^2}$$

- Usato per verificare la presenza di un legame
 - $|R| < 0.3$ scarsa probabilità di legame
 - $0.3 < |R| < 0.7$ moderata probabilità di legame
 - $0.7 < |R|$ ottima probabilità di legame lineare