ANALISI DI CORRELAZIONE

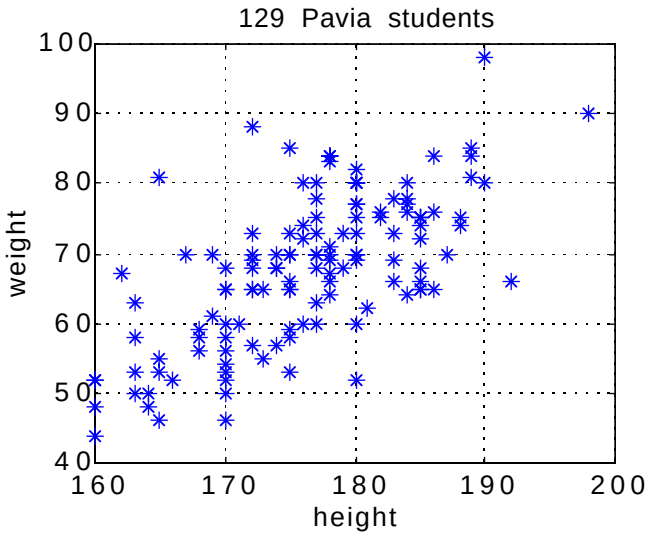
Esempio: Dati raccolti da $n = 129$ studenti di Pavia (A.A. 2001/02)

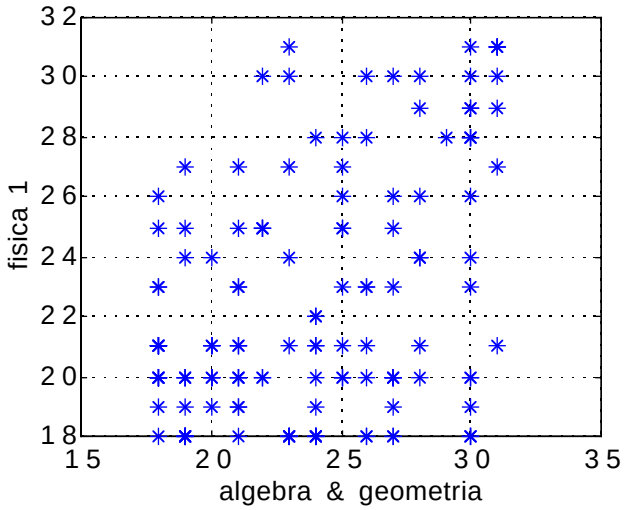
- Altezza (cm)
- Peso (Kg)
- Voto Algebra e Geometria
- Voto Fisica I

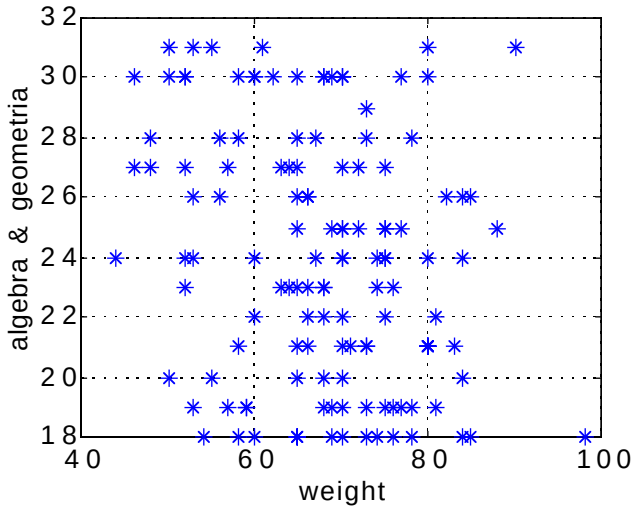
Valutare la correlazione delle seguenti coppie:

- Peso - Altezza
- Algebra e Geometria - Fisica I
- Peso - Algebra e Geometria

Innanzi tutto, esaminiamo gli scatter plot:







A prima vista si nota che:

- Chiara correlazione positiva tra peso e altezza
- Leggera correlazione positiva tra Algebra e Fisica
- Dubbia correlazione negativa (incorrelazione?) tra peso e Geometria.

Indice quantitativo di correlazione: Coefficiente di correlazione

$$r_{XY} := \frac{Cov[X,Y]}{\sqrt{Var[X]\,Var[Y]}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Per stimarlo, facciamo ricorso alla

Correlazione campionaria:

$$R_{XY} := \frac{S_{xy}}{S_x\, S_y}$$

$$S_{xy} := \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{n - 1}$$

Nota: Se $V = aX+b$, $W = cY+d$:

- $r_{VW} = r_{XY}$
- $R_{VW} = R_{XY}$

Tabella: Medie, varianze, SD

variabile	*media*	*varianza*	*SD*
altezza	176.3	60.0	7.8
peso	67.7	111.7	10.6
algebra	24.0	17.2	4.1
fisica I	22.6	16.0	4.0

Tabella: Correlazioni stimate

variabili	*correlazione campionaria*
altezza-peso	0.67
algebra-fisica	0.38
peso-algebra	- 0.25

Commenti:

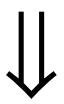
- Come previsto, c'è una buona correlazione tra altezza e peso.

- La correlazione non molto alta tra algebra e fisica significa che gli esami sono una "lotteria"?

- La correlazione negativa tra peso e algebra significa che perdere peso aiuta ad alzare la media?

Problema: come decidere se la correlazione è "significativamente diversa da zero"?

Approccio: Prendere le parti dell'avvocato del diavolo, fare l'ipotesi nulla che la coppia di variabili sia incorrelata ($r_{XY} = 0$) e verificare se è sufficientemente screditata dai dati.



Per verificare se l'ipotesi nulla è screditata, ho bisogno di una "distribuzione di riferimento".

Esperimento: (che mostra in che modo la distribuzione di riferimento di R_{xy} dipende dalla numerosità n del campione)

1) Usando un generatore di numeri casuali estraggo n coppie di variabili casuali *indipendenti* ($\Rightarrow r_{XY} = 0$), distribuite entrambe come normali standard

2) Calcolo la correlazione campionaria R_{xy}.

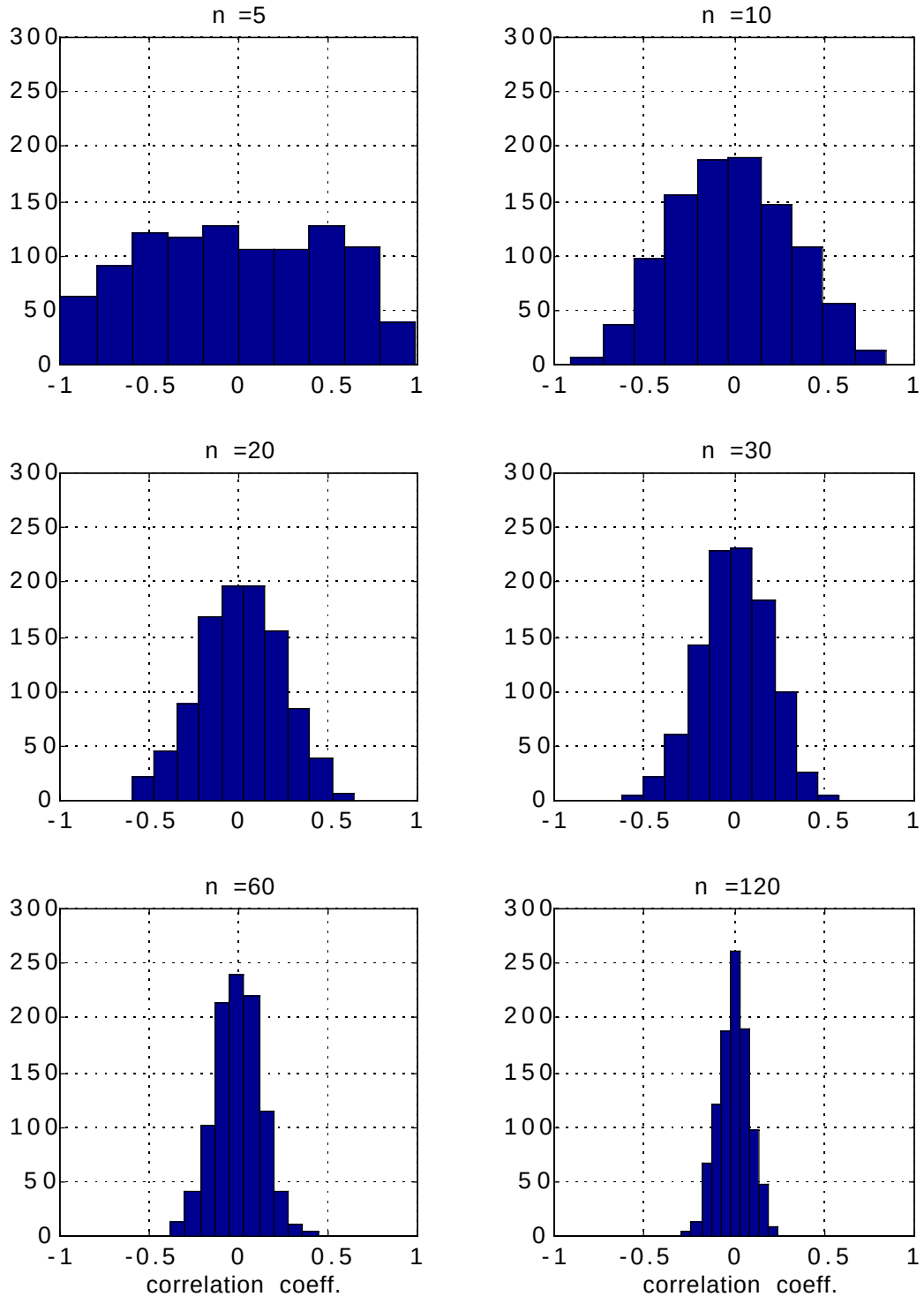
3) Ripeto l'estrazione ed il calcolo di R_{xy} altre 999 volte.

4) Costruisco l'istogramma delle 1000 correlazioni campionarie R_{xy} così ottenute

$\Downarrow$

L'istogramma approssima la distribuzione di riferimento per R_{xy} sotto l'ipotesi nulla che le variabili siano incorrelate

Distribuzioni di riferimento approssimate per diverse numerosità n del campione:



Osservazione: Se n non è abbastanza grande risulta relativamente facile ottenere valori di R_{xy} molto maggiori di zero anche se le variabili sono incorrelate.

Esempio: Se $n \leq 20$, un valore $R_{xy} = 0.4$ non è una buona garanzia che esista correlazione (a differenza di quanto accade per $n \geq 60$).

Diamo ora una valutazione quantitativa della significatività di R_{xy}.

Proprietà: Si considerino n coppie (x,y) dove x ed y sono congiuntamente gaussiane e indipendenti. Allora:

$$t = \frac{R_{xy}}{\sqrt{1 - R_{xy}^2}} \sqrt{n - 2} = t_{n-2}$$

(t_{n-2}: t di Student a n-2 gradi di libertà)

Test di correlazione (significatività α): Avendo calcolato t

- $|t| > t_{\alpha/2} \Rightarrow$ respingo l'ipotesi nulla $\Rightarrow R_{xy}$ è significativamente $\neq 0$
- $|t| \leq t_{\alpha/2} \Rightarrow$ non respingo l'ipotesi nulla $\Rightarrow R_{xy}$ non è significativamente $\neq 0$

Tabella: Valori di $t_{\alpha/2}$

$$P(t_{n-1} \geq t_{\alpha/2}) = \alpha/2$$
$$\text{ovvero } P(|t_{n-1}| \leq t_{\alpha/2}) = \alpha$$

n	$\alpha = 0.20$	$\alpha = 0.10$	$\alpha = 0.05$
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182
4	1.533	2.132	2.776
5	1.476	2.015	2.571
6	1.440	1.943	2.447
7	1.415	1.895	2.365
8	1.397	1.860	2.306
9	1.383	1.833	2.262
10	1.372	1.812	2.228
11	1.363	1.796	2.201
12	1.356	1.782	2.179
13	1.350	1.771	2.160
14	1.345	1.761	2.145
15	1.341	1.753	2.131
16	1.337	1.746	2.120
17	1.333	1.740	2.110
18	1.330	1.734	2.101
19	1.328	1.729	2.093
20	1.325	1.725	2.086
21	1.323	1.721	2.080
22	1.321	1.717	2.074
23	1.319	1.714	2.069
24	1.318	1.711	2.064
25	1.316	1.708	2.060
26	1.315	1.706	20.56
27	1.314	1.703	2.052
28	1.313	1.701	2.048
29	1.311	1.699	2.045
30	1.310	1.697	2.042
40	1.303	1.684	2.021
60	1.296	1.671	2.000
120	1.289	1.658	1.980
∞	1.282	1.645	1.960

Metodo più comodo: Se fisso $\alpha = 5\%$ posso calcolare i valori critici di R_{xy} per diversi valori di n.

Tabella: Valori di $r_{2.5\%}$

R_{xy} *è significativamente* $\neq 0$
se $|R_{xy}| > r_{2.5\%}$

n	$r_{2.5\%}$
3	0.997
4	0.95
5	0.88
6	0.81
7	0.75
8	0.71
9	0.67
10	0.63
11	0.60
12	0.58
13	0.55
14	0.53
15	0.51
16	0.50
17	0.48
18	0.47
19	0.46
20	0.44
21	0.43
22	0.42
23	0.41
24	0.40
25	0.40
26	0.39
27	0.38
28	0.37
29	0.37
30	0.36
60	0.25
120	0.18

Torniamo all'esempio:

Tabella: Correlazioni stimate (n = 129)

variabili	*correlaz. camp.*	
altezza-peso	0.67	*significativamente* $\neq 0$
algebra-fisica	0.38	*significativamente* $\neq 0$
peso-algebra	- 0.25	*significativamente* $\neq 0$*!*

Domanda: Cosa c'entra il peso con il voto di Algebra e Geometria??

"BEWARE THE LURKING VARIABLE!"

Attenzione alla variabile nascosta ("in agguato")!

- Il fatto che due variabili siano correlate non implica l'esistenza di una relazione causale:
  Esempio #1: numero di scarpe - capacità di lettura.
  Esempio #2: titolo di studio - durata disoccupazione negli USA durante la Grande Depressione

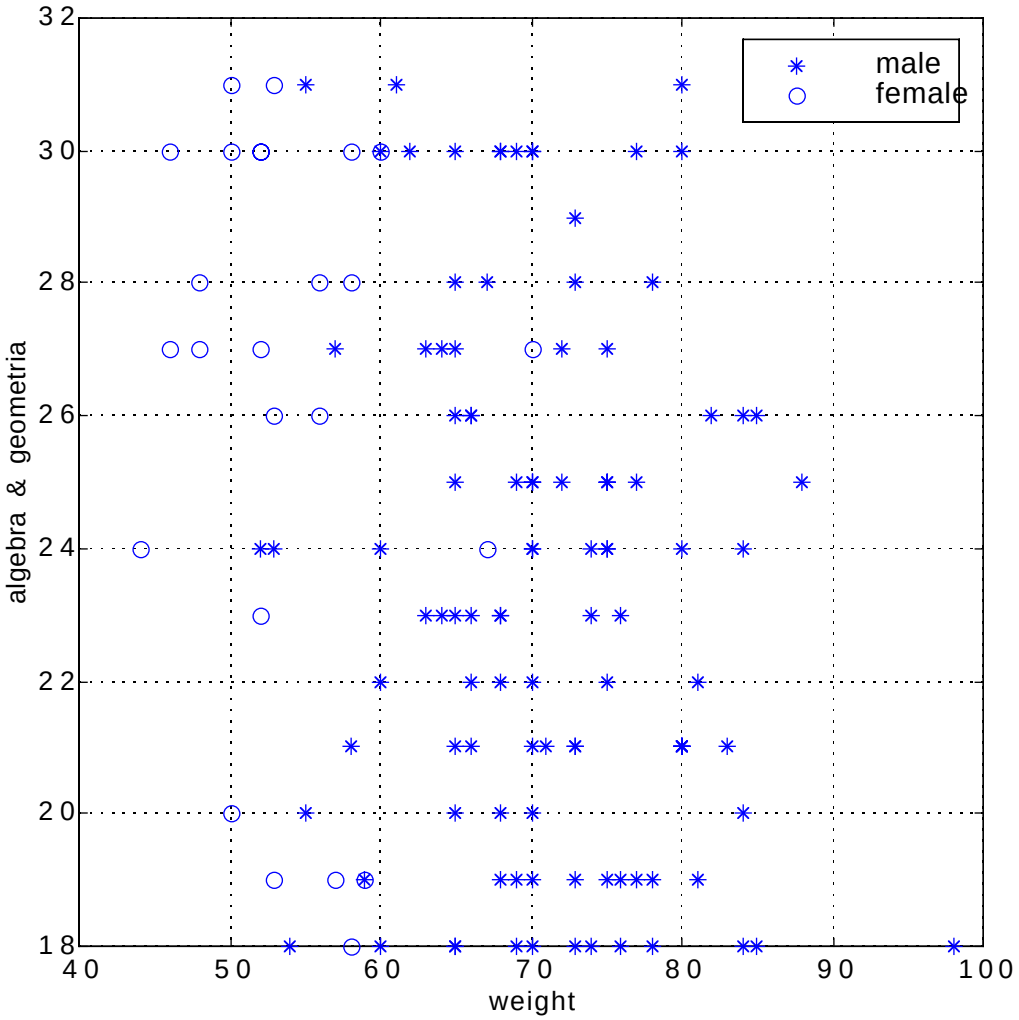
- In molti casi la correlazione può essere spiegata perché entrambe la variabili sono correlate con una terza variable (*the lurking variable*).
  Esempio #1: l'età.
  Esempio #2: ancora l'età perché i giovani erano più istruiti.

- Bisogna stare molto attenti, specialmente quando la popolazione è composta di sottopopolazioni.

Nel nostro caso gli studenti si suddividono
in maschi e femmine …

Rivediamo lo scatter plot distinguendo maschi e femmine:

(n_M = 104, n_F = 25)



Congettura: C'è correlazione tra peso e algebra perché le femmine hanno (mediamente) voti più alti e (mediamente) pesano di meno

Istogrammi

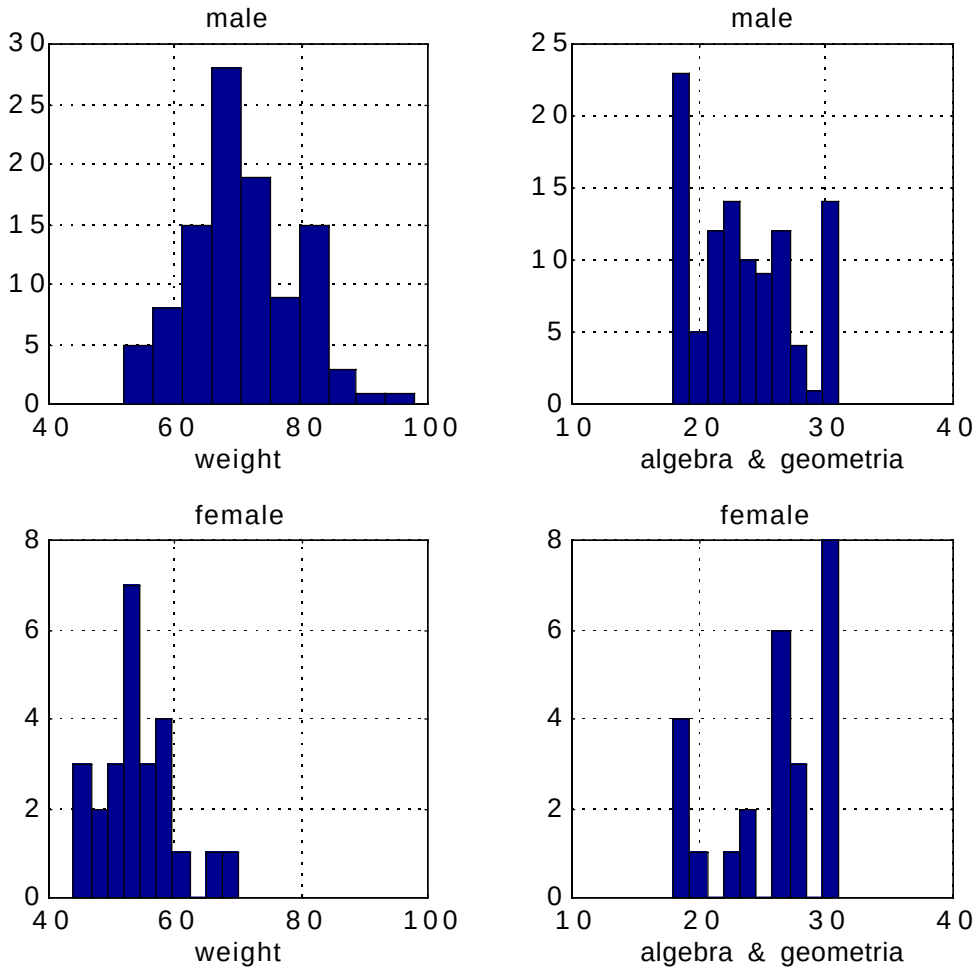

male
30
25
20
15
10
5
0
40
60
80
100
weight
male
25
20
15
10
5
0
10
20
30
40
algebra & geometria
female
8
6
4
2
0
40
60
80
100
weight
female
8
6
4
2
0
10
20
30
40
algebra & geometria

Tabella: Medie, varianze, SD

variabile	*media*	*varianza*	*SD*
peso *(femmine)*	53.9	38.2	6.2
peso *(maschi)*	71.1	72.3	8.5
algebra *(femmine)*	26.1	17.7	4.2
algebra *(maschi)*	23.5	15.9	4.0

Tabella: Correlazioni stimate

variabili	*correlazione campionaria*
peso-algebra *(femmine)*	- 0.17
peso-algebra *(maschi)*	- 0.11

Commenti:

- Analizzando separatamente le due sottopopolazioni (maschi e femmine) la correlazione peso-algebra torna in entrambi i casi sotto il limite di significatività.

- E' ragionevole pensare che il sesso funga da variabile nascosta.

- Per qualche ragione, le femmine sembrano ottenere (mediamente) voti più alti in algebra rispetto ai maschi.

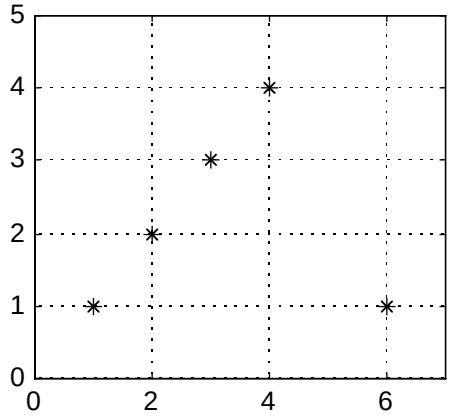
Come valutare la significatività della superiorità femminile?

1) Tecniche per testare le differenze tra due gruppi.

2) Vedi più avanti ("Valutare l'influenza di una variabile").

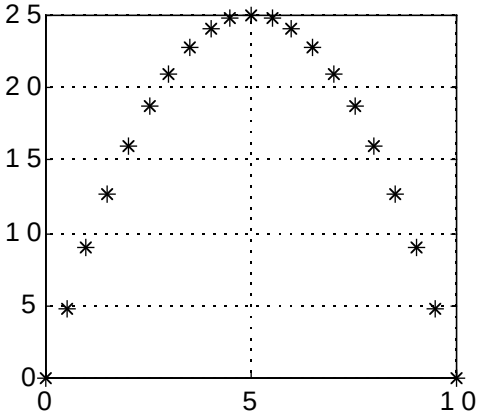
Considerazioni finali sull'analisi di correlazione

Attenzione! Gli outlier possono falsare le correlazioni.



$R_{xy} = 0.08$

Attenzione! Ci sono legami tra le variabili che non sono rivelati dal coefficiente di correlazione.



$R_{xy} = 3.0774\text{e-}17$

Morale:

- Il coefficiente di correlazione misura *l'associazione lineare* e non l'associazione in generale.

- In generale il suo uso è appropriato quando lo *scatter plot* è circa *ovale*.

- I migliori risultati si hanno per variabili congiuntamente gaussiane (per le quali incorrelazione $\Leftrightarrow$ indipendenza).