

# **La matrice dei dati**

## **Analisi Esplorativa**

Aldo Solari



- ① Tipologia di variabili
- ② Valori mancanti
- ③ Valori anomali
- ④ Matrice dei dati
- ⑤ Diagramma di dispersione



# I dati

I dati possono essere rappresentati con una tabella  $n \times p$

- $n$  osservazioni o unità statistiche: individui, aziende, etc.
- $p$  variabili o misurazioni o caratteristiche: altezza, sesso, etc.

	Variabile 1	...	Variabile $j$	...	Variabile $p$
Unità statistica 1	$x_{11}$	...	$x_{1j}$	...	$x_{1p}$
Unità statistica 2	$x_{21}$	...	$x_{2j}$	...	$x_{2p}$
...	...	...	...	...	...
Unità statistica $i$	$x_{i1}$	...	$x_{ij}$	...	$x_{ip}$
...	...	...	...	...	...
Unità statistica $n$	$x_{n1}$	...	$x_{nj}$	...	$x_{np}$

- $n$  = numerosità dei dati
- $p$  = dimensionalità dei dati



# Esempio

$n = 10$  individui e  $p = 5$  variabili:

	sexo	figli	occhi	salute	peso
1	Maschio	0	Azzurri	Molto Buona	68.04
2	Maschio	1	Neri	Molto Buona	72.57
3	Maschio	0	Marroni	Media	61.23
4	Maschio	0	Neri	Cattiva	63.50
5	Maschio	1	Azzurri	Buona	49.90
6	Femmina	0	Marroni	Buona	49.90
7	Femmina	2	Azzurri	Molto Buona	54.43
8	Femmina	0	Marroni	Media	54.43
9	Femmina	0	Neri	Media	47.63
10	Femmina	1	Neri	Buona	45.36



# Outline

- ➊ Tipologia di variabili
- ➋ Valori mancanti
- ➌ Valori anomali
- ➍ Matrice dei dati
- ➎ Diagramma di dispersione



# Tipologia di variabili

Le variabili si suddividono in due tipologie:

## Qualitative

- nominali (in R: `Factor`), se non esiste nessun ordinamento naturale tra le modalità ;
- ordinali (in R: `Ord.factor`), se esiste un ordinamento naturale tra le modalità .

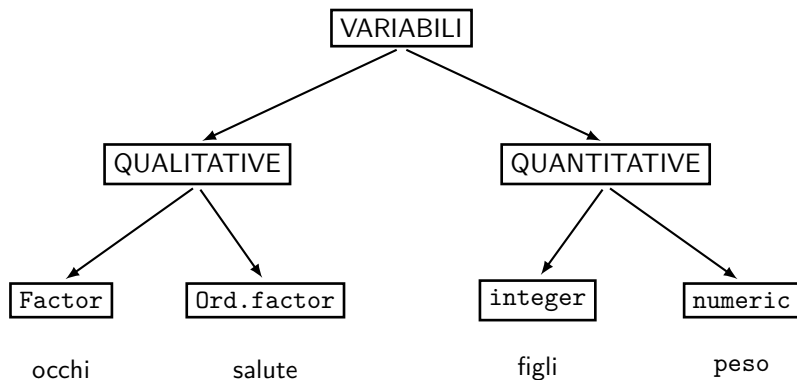
## Quantitative

- discrete (in R: `integer`), quando sono esprimibili da numeri interi
- continue (in R: `numeric`), quando sono esprimibili da numeri reali

Variabili **Dicotomiche**: quando le modalità sono solamente due



# Tipologia di variabili



# Outline

- 1 Tipologia di variabili
- 2 Valori mancanti**
- 3 Valori anomali
- 4 Matrice dei dati
- 5 Diagramma di dispersione





## Valori mancanti (*missing values*)

	sexo	figli	IQ	occhi	salute	peso
1	Maschio	0	120	Azzurri	Molto Buona	68.04
2	Maschio	1		Neri	Molto Buona	72.57
3	Maschio	0		Marroni	Media	61.23
4	Maschio	0	150	Neri	Cattiva	63.50
5	Maschio	1	92	Azzurri	Buona	49.90
6	Femmina	0	130	Marroni	Buona	49.90
7	Femmina			Azzurri	Molto Buona	54.43
8	Femmina	0		Marroni	Media	54.43
9	Femmina	0	84	Neri	Media	47.63
10	Femmina	1	70	Neri	Buona	45.36



# NA

In R, i valori mancanti vengono codificati con NA (*Not Available*)

	sexo	figli	IQ	occhi	salute	peso
1	Maschio	0	120	Azzurri	Molto Buona	68.04
2	Maschio	1	NA	Neri	Molto Buona	72.57
3	Maschio	0	NA	Marroni	Media	61.23
4	Maschio	0	150	Neri	Cattiva	63.50
5	Maschio	1	92	Azzurri	Buona	49.90
6	Femmina	0	130	Marroni	Buona	49.90
7	Femmina	NA	NA	Azzurri	Molto Buona	54.43
8	Femmina	0	NA	Marroni	Media	54.43
9	Femmina	0	84	Neri	Media	47.63
10	Femmina	1	70	Neri	Buona	45.36

Problema: le tecniche di analisi multivariata che andremo a considerare prevedono osservazioni con tutti i valori presenti.



## Esclusione di variabili incomplete

	sexo	figli	<del>IQ</del>	occhi	salute	peso
1	Maschio	<del>0</del>	<del>120</del>	Azzurri	Molto Buona	68.04
2	Maschio	<del>1</del>	<del>NA</del>	Neri	Molto Buona	72.57
3	Maschio	<del>0</del>	<del>NA</del>	Marroni	Media	61.23
4	Maschio	<del>0</del>	<del>150</del>	Neri	Cattiva	63.50
5	Maschio	<del>1</del>	<del>92</del>	Azzurri	Buona	49.90
6	Femmina	<del>0</del>	<del>130</del>	Marroni	Buona	49.90
7	Femmina	<del>NA</del>	<del>NA</del>	Azzurri	Molto Buona	54.43
8	Femmina	<del>0</del>	<del>NA</del>	Marroni	Media	54.43
9	Femmina	<del>0</del>	<del>84</del>	Neri	Media	47.63
10	Femmina	<del>1</del>	<del>70</del>	Neri	Buona	45.36

Diminuisce la dimensionalità  $p$  dei nostri dati. Però le variabili escluse potrebbero essere proprio quelle di interesse per l'analisi



# Esclusione di osservazioni incomplete

	sexo	figli	IQ	occhi	salute	peso
1	Maschio	0	120	Azzurri	Molto Buona	68.04
<del>2</del>	<del>Maschio</del>	<del>1</del>	<del>NA</del>	<del>Neri</del>	<del>Molto Buona</del>	<del>72.57</del>
<del>3</del>	<del>Maschio</del>	<del>0</del>	<del>NA</del>	<del>Marroni</del>	<del>Media</del>	<del>61.23</del>
4	Maschio	0	150	Neri	Cattiva	63.50
5	Maschio	1	92	Azzurri	Buona	49.90
6	Femmina	0	130	Marroni	Buona	49.90
<del>7</del>	<del>Femmina</del>	<del>NA</del>	<del>NA</del>	<del>Azzurri</del>	<del>Molto Buona</del>	<del>54.43</del>
<del>8</del>	<del>Femmina</del>	<del>0</del>	<del>NA</del>	<del>Marroni</del>	<del>Media</del>	<del>54.43</del>
9	Femmina	0	84	Neri	Media	47.63
10	Femmina	1	70	Neri	Buona	45.36

Diminuisce la numerosità  $n$  dei nostri dati. Vi vengono in mente altri potenziali problemi?

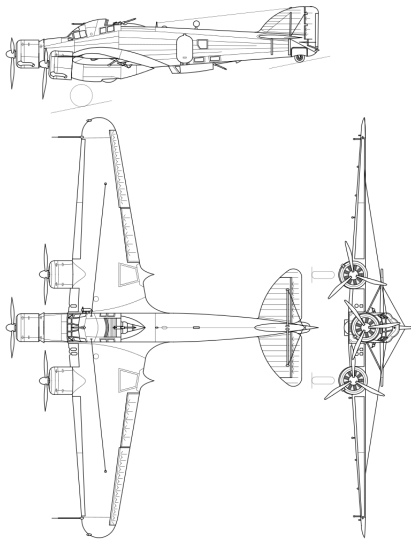


# WWII

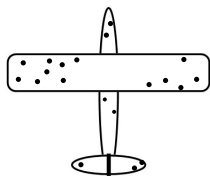
- Quanto segue è realmente accaduto durante la seconda guerra mondiale
- Obiettivo: proteggere gli aerei da caccia degli alleati negli scontri con i caccia della Luftwaffe
- Un caccia (Savoia-Marchetti S.M.79) è un velivolo leggero e agile
- Per evitare l'abbattimento, questi aerei venivano corazzati con robuste lastre di ferro
- Problema: quante corazze e dove le mettiamo? Se un aereo non è corazzato, è facile da abbattere; se è troppo corazzato, è difficile da manovrare
- Per un aereo abbiamo 4 settori: (A) ali (B) alimentazione (C) fusoliera (D) motore. Possiamo mettere la corazza in un solo settore. Dove la mettiamo?
- Guardiamo i dati degli aerei



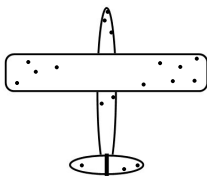
# Savoia-Marchetti S.M.79



# I dati



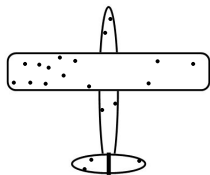
(a)



(b)



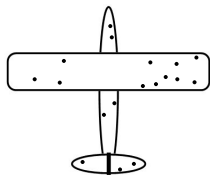
(c)



(d)



(e)



(f)



# Tabella dei dati

Zona dell'aereo	Numero di colpi/dm <sup>2</sup>
Ali	0.167
Alimentazione	0.143
Fusoliera	0.161
Motore	0.103

Nota: la media delle densità di colpi (numero di colpi per decimetro quadrato) è calcolata escludendo i valori (aerei) mancanti

Grazie a questa tabella, lo statistico Abraham Wald fu in grado di posizionare la protezione nel punto più rischioso

Fonte: D. Hand (2019) Il tradimento dei numeri. I dark data e l'arte di nascondere la verità. Rizzoli





# L'opinione di uno statistico

*The armor doesn't go where the bullet holes are.  
It goes where the bullet holes aren't.*

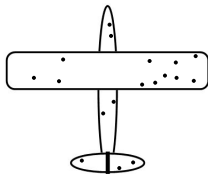
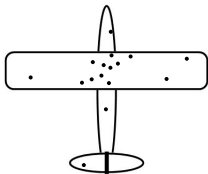
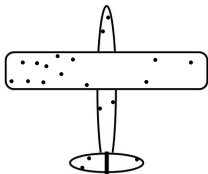
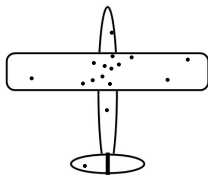
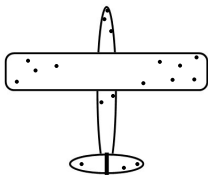
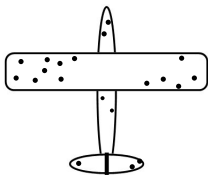
*The observed holes showed where the planes were strongest;  
that's where the planes could be shot and still survive the flight home.  
The missing holes showed where the planes were weaker;  
that's where the planes that didn't make it back were hit.*

Abraham Wald

Pensate di far visita ad un ospedale militare durante una guerra:  
vi aspettate di osservare più feriti alle gambe o alla testa?



# Gli aerei mancanti (non a caso)



# Valori mancanti (completamente) a caso

Si parla di valori mancanti (completamente) a caso se i valori mancanti sono un campione casuale dei  $n \times p$  valori possibili.

In tale situazione non ci sono problemi se escludiamo le osservazioni che presentano almeno un valore mancante (tranne il fatto che diminuisce la numerosità  $n$ )



# Dati Titanic



pclass	name	sex	age	sibsp	parch	ticket	fare	embarked
3	Storey	male	60.5	0	0	3701		S
1	Natsch	male	37.0	0	1	PC 17596	29.7	C
3	Johansson	male	31.0	0	0	347063	7.8	S
2	Clarke	female	28.0	1	0	2003	26.0	S
3	Danbom	female	28.0	1	1	347080	14.4	S



# Imputazione di dati mancanti

- Il passeggero Mr. Thomas Storey presenta un valore mancante sul prezzo del biglietto (variabile fare)
- Tuttavia sappiamo che si è imbarcato a Southampton (variabile embarked) e viaggiava in terza classe (variabile pclass).
- Potrebbe essere sensato sostituire il valore mancante con il prezzo mediano pari a 8.1

pclass	embarked	median fare
1	C	76.7
2	C	15.3
3	C	7.9
1	Q	90.0
2	Q	12.3
3	Q	7.8
1	S	52.0
2	S	15.4
3	S	8.1



# Outline

- 1 Tipologia di variabili
- 2 Valori mancanti
- 3 Valori anomali**
- 4 Matrice dei dati
- 5 Diagramma di dispersione



# Valori anomali (*outliers*)

Ogni insieme di valori ha un massimo e un minimo, però può capitare di osservare uno o più valori veramente anomali (*outliers*)

## Valore anomalo (*outlier*)

E' un valore che si discosta dal baricentro della distribuzione più di quanto possa essere giustificato dalla variabilità dei dati.



# Perchè sono valori anomali?

Ci possono essere diverse spiegazioni, ad esempio:

## **Errore di rilevazione**

e.g. per la variabile altezza, ho imputato 18.4 m invece di 1.84 m

## **Elevata variabilità intrinseca del fenomeno (code pesanti)**

e.g. pensate alla variabile reddito

## **Valori provenienti da una distribuzione diversa (contaminazione)**

e.g. pensate al peso per animali viventi e animali estinti (dinosauri)





# Come si individuano i valori anomali?

Metodi basati sull'esplorazione grafica:

Per una singola variabile

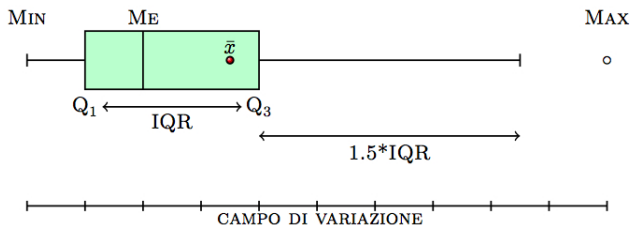
- Diagramma a scatola con baffi (*boxplot*)

Per due variabili

- Diagramma di dispersione
- *Bagplot*



# Diagramma a scatola con baffi (*boxplot*)



- $Me$ ,  $Q_1$  e  $Q_3$  sono la mediana, il primo e il terzo quartile
- $IQR = Q_3 - Q_1$  è il *range* interquartile
- Il baffo a sinistra è il valore massimo tra Min e  $Q_1 - 1.5 \cdot IQR$
- Il baffo a destra è il valore minimo tra Max e  $Q_3 + 1.5 \cdot IQR$



# Boxplot e valori anomali

Il diagramma a scatola e baffi (*boxplot*) identifica un valore anomalo (indicandolo con  $\circ$ ) con la seguente regola:

Un valore  $x_i$ ,  $i = 1, \dots, n$  è anomalo se:

- $x_i < Q_1 - 1.5 \cdot \text{IQR}$  oppure se
- $x_i > Q_3 + 1.5 \cdot \text{IQR}$



## Comando `boxplot()` con R

- Se la numerosità campionaria  $n$  è un numero dispari, la descrizione coincide con quella delle slides precedenti;
- Se invece la numerosità campionaria  $n$  è un numero pari, i valori di  $Q_1$  e  $Q_3$  che calcola il comando `boxplot()` potrebbero essere leggermente diversi dal primo e il terzo quartile
- Potete utilizzare il comando `boxplot.stats()` per ottenere i 5 valori che compongono il *boxplot* (Min, baffo sx, Me, baffo dx, Max)



# Dati Animals

- `Animals` è un *dataset* presente nella libreria MASS
- Per una descrizione del *dataset*, digitare `?Animals`
- *Average brain and body weights for 28 species of land animals*
- `body` : body weight in kg
- `brain` : brain weight in g
- $n = 28$  osservazioni misurate su  $p = 2$  variabili

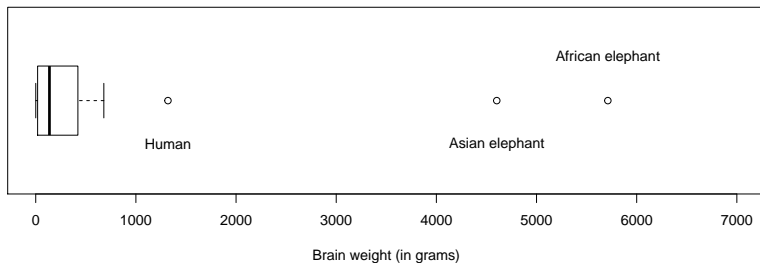


# Dati Animals

	body	brain
Mountain beaver	1.35	8.10
Cow	465.00	423.00
Grey wolf	36.33	119.50
Goat	27.66	115.00
Guinea pig	1.04	5.50
Dipliodocus	11700.00	50.00
Asian elephant	2547.00	4603.00
Donkey	187.10	419.00
Horse	521.00	655.00
Potar monkey	10.00	115.00
Cat	3.30	25.60
Giraffe	529.00	680.00
Gorilla	207.00	406.00
Human	62.00	1320.00
African elephant	6654.00	5712.00
Triceratops	9400.00	70.00
Rhesus monkey	6.80	179.00
Kangaroo	35.00	56.00
Golden hamster	0.12	1.00
Mouse	0.02	0.40
Rabbit	2.50	12.10
Sheep	55.50	175.00
Jaguar	100.00	157.00
Chimpanzee	52.16	440.00
Rat	0.28	1.90
Brachiosaurus	87000.00	154.50
Mole	0.12	3.00
Pig	192.00	180.00



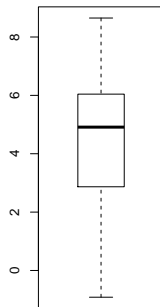
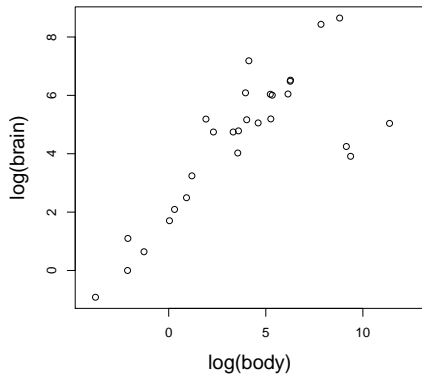
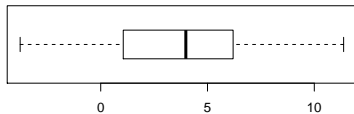
# Boxplot brain



```
library("MASS")  
boxplot(Animals$brain)  
boxplot.stats(Animals$brain)  
$stats  
[1] 0.40 18.85 137.00 421.00 680.00  
$out  
[1] 4603 1320 5712
```

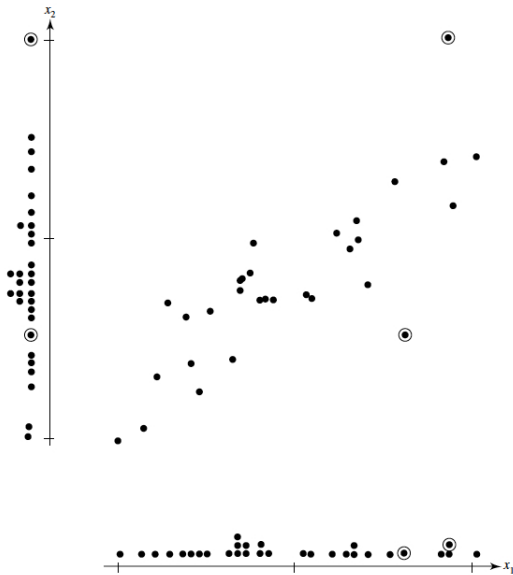


# Boxplot brain e body

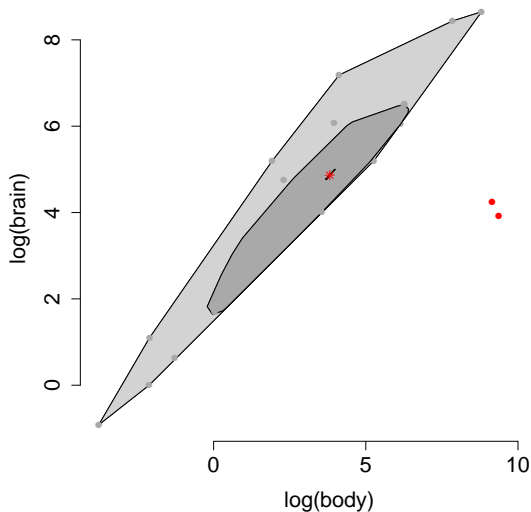




# Outlier bivariato



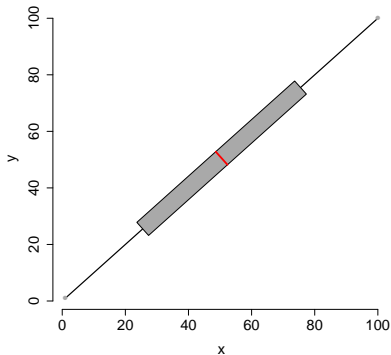
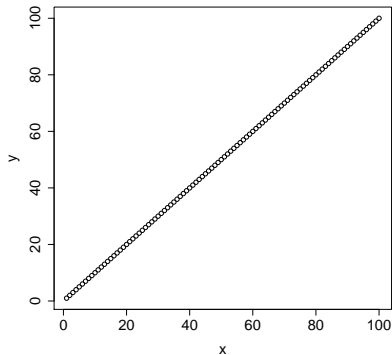
# Bagplot = boxplot bivariato



50% delle osservazioni all'interno del *bag* (area grigio scuro)



# Bagplot per dati unidimensionali



# Outline

- 1 Tipologia di variabili
- 2 Valori mancanti
- 3 Valori anomali
- 4 Matrice dei dati**
- 5 Diagramma di dispersione



# Matrice $X$

$$X_{n \times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$



# Medie e varianze

- Media per la  $j$ -sima variabile

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, \quad j = 1, \dots, p$$

- Varianza per la  $j$ -sima variabile

$$s_j^2 = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2, \quad j = 1, \dots, p$$



# Covarianze e correlazioni

- Covarianza tra la  $j$ -sima e la  $k$ -sima variabile

$$s_{jk} = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \quad j = 1, \dots, p, \quad k = 1, \dots, p$$

Si noti che  $s_{jk} = s_{kj}$  e che  $s_{jj} = s_j^2$

- Correlazione tra la  $j$ -sima e la  $k$ -sima variabile

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}, \quad j = 1, \dots, p, \quad k = 1, \dots, p$$

Si noti che  $-1 \leq r_{jk} \leq 1$



# Vettore delle medie

$$\bar{x}_{p \times 1} = \begin{bmatrix} \bar{x}_1 \\ \dots \\ \bar{x}_j \\ \dots \\ \bar{x}_p \end{bmatrix}$$





# Matrice di varianze/covarianze

$$S_{p \times p} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1j} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2j} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ s_{j1} & s_{j2} & \cdots & s_{jj} & \cdots & s_{jp} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \cdots \\ s_{p1} & s_{p2} & \cdots & s_{pj} & \cdots & s_{pp} \end{bmatrix}$$



# Matrice di correlazione

$$R_{p \times p} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1j} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2j} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ r_{j1} & r_{j2} & \cdots & 1 & \cdots & r_{jp} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pj} & \cdots & 1 \end{bmatrix}$$



# Esempio

Variabile 1 (prezzo in Dollari per libro):	42	52	48	58
Variabile 2 (numero di libri venduti):	4	5	4	3

Calcolare il vettore delle medie, la matrice di varianze/covarianze e di correlazione (svolgimento alla lavagna)



# Outline

- 1 Tipologia di variabili
- 2 Valori mancanti
- 3 Valori anomali
- 4 Matrice dei dati
- 5 Diagramma di dispersione**



# Dati

	$u_1$	$u_2$	$u_3$	$u_4$	$u_5$	$u_6$	$u_7$
$x_1$	3	4	2	6	8	2	5
$x_2$	5	5.5	4	7	10	5	7.5

Medie:  $\bar{x}_1 = 4.2$ ,  $\bar{x}_2 = 6.2$

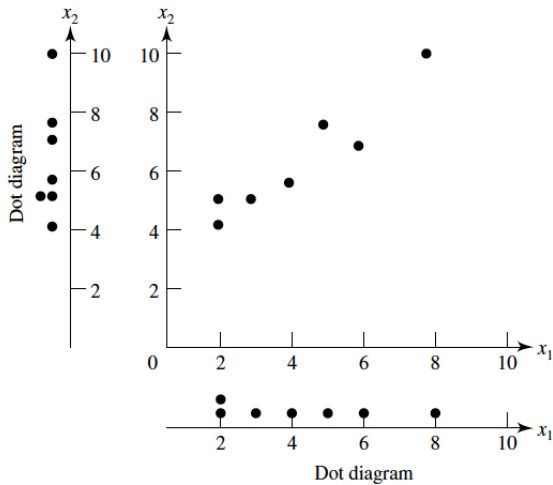
Varianze:  $s_{11} = 4.2$ ,  $s_{22} = 0.56$

Covarianza:  $s_{12} = 3.70$

Correlazione:  $r_{12} = 0.95$



# Diagramma di dispersione



# Dati

	$u_1$	$u_2$	$u_3$	$u_4$	$u_5$	$u_6$	$u_7$
$x_1$	5	4	6	2	2	8	3
$x_2$	5	5.5	4	7	10	5	7.5

Medie:  $\bar{x}_1 = 4.2$ ,  $\bar{x}_2 = 6.2$

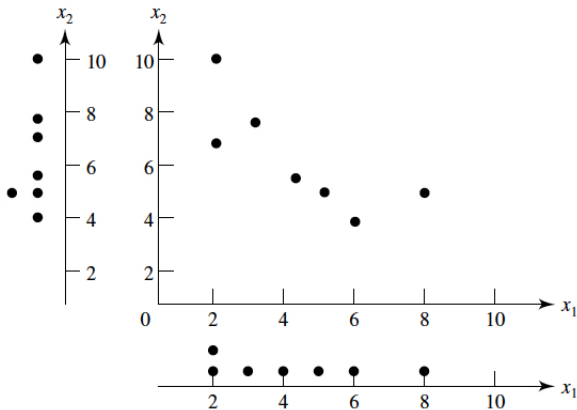
Varianze:  $s_{11} = 4.20$ ,  $s_{22} = 0.56$

Covarianza  $s_{12} = -3.01$

Correlazione  $r_{12} = -0.78$



# Diagramma di dispersione





# Indovina la correlazione

Guess the correlation



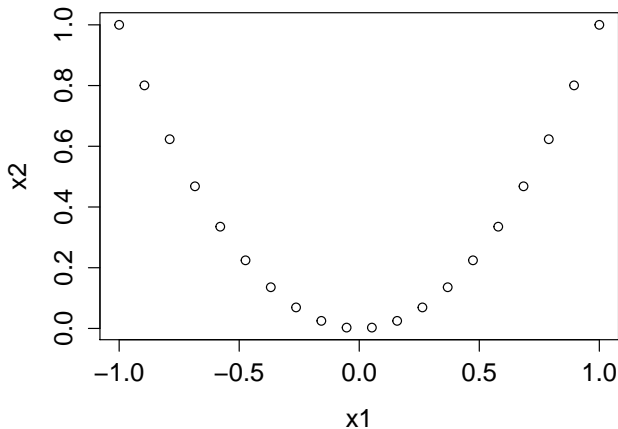
# Relazione quadratica

$$x_{1i} = -1 + 2 \frac{(i-1)}{(n-1)}$$
$$x_{2i} = x_{1i}^2, \quad i = 1, \dots, n$$



# Relazione quadratica

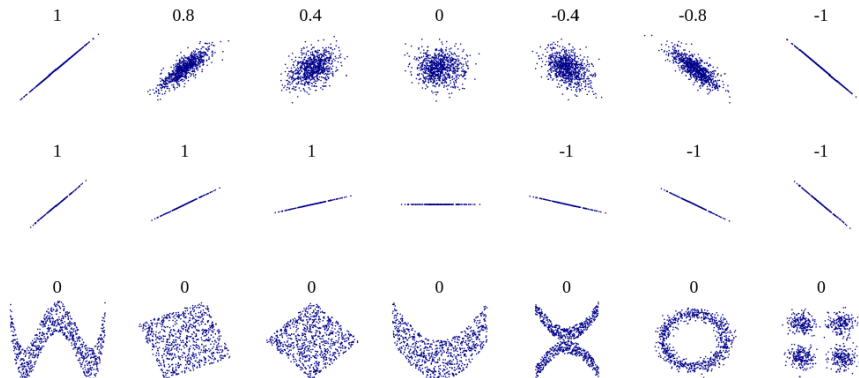
Per  $n = 20$ :



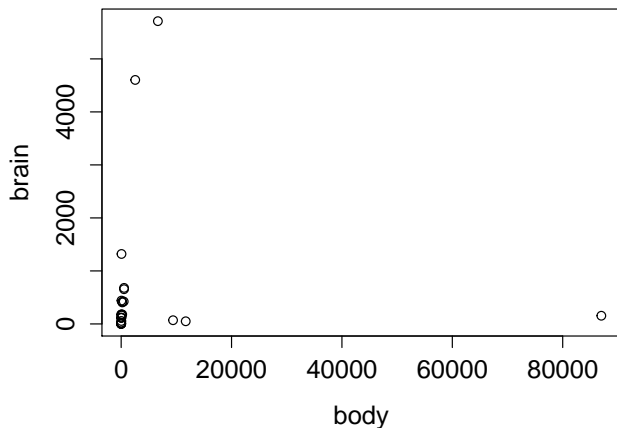
$$r_{12} \approx 0$$



# Correlazione = relazione LINEARE



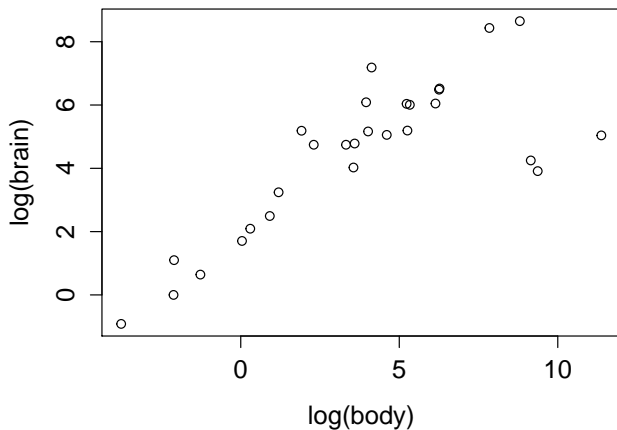
## Animals: diagramma di dispersione



$$r_{12} = -0.0053$$



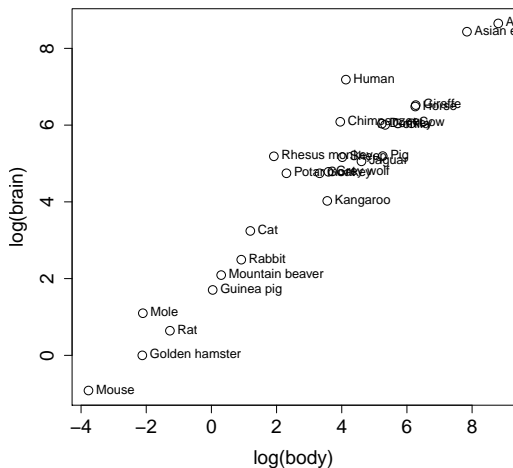
# Animals: trasformazione logaritmica



$$r_{12} = 0.779$$



# Animals: escludendo 3 osservazioni anomale



$$r_{12} = 0.932$$

