Relazioni Statistiche

G. M. Marchetti

2016-10-04

Grammatica Statistica

- La sensibilità statistica proviene dall'*analisi dei dati*
- La statistica *non è intuitiva*
- La statistica *non è matematica*
- Ma dietro ogni concetto statistico c'è una *definizione matematica precisa* basata sul calcolo delle probabilità

Regressione

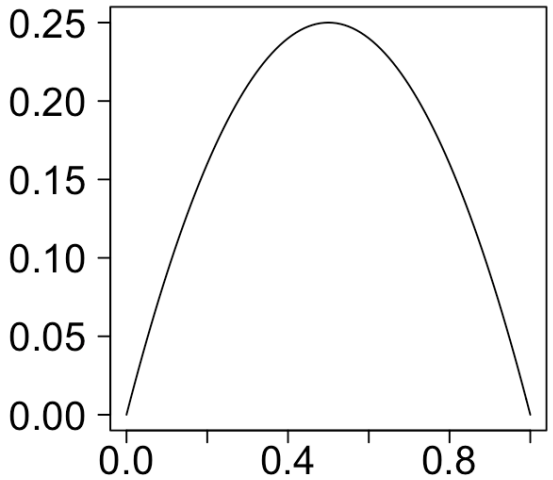
Distinzioni

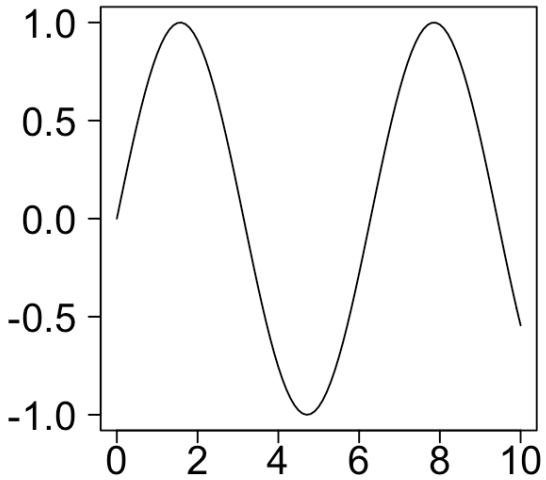
- Regressione: relazioni tra due variabili in cui una Y è dipendente e l'altra X è un potenziale predittore
- Correlazione: relazioni lineari tra due variabili X e Y considerate sullo stesso piano.

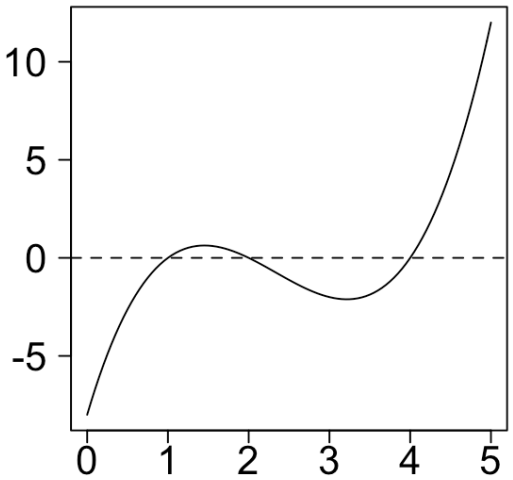
Relazioni matematiche

Il concetto base è quello di **funzione** $y = f(x)$

- Parabola
- Sinusoide
- Un polinomio





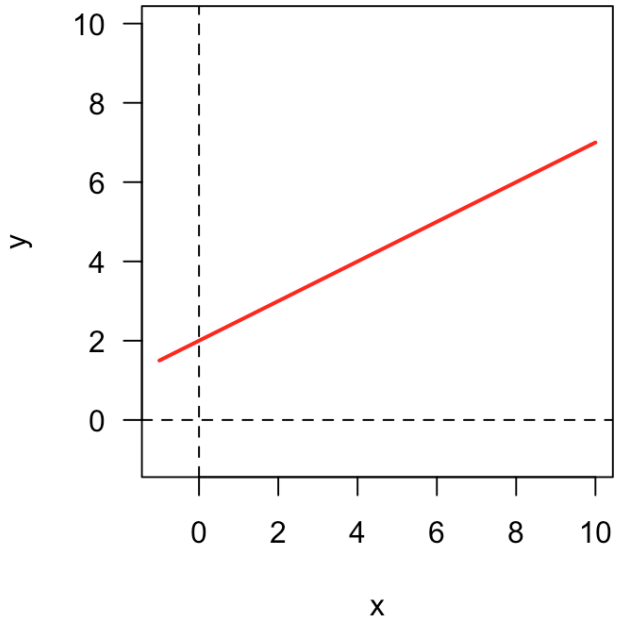


La retta

La relazione più semplice è quella *lineare*

$$y = a + bx$$

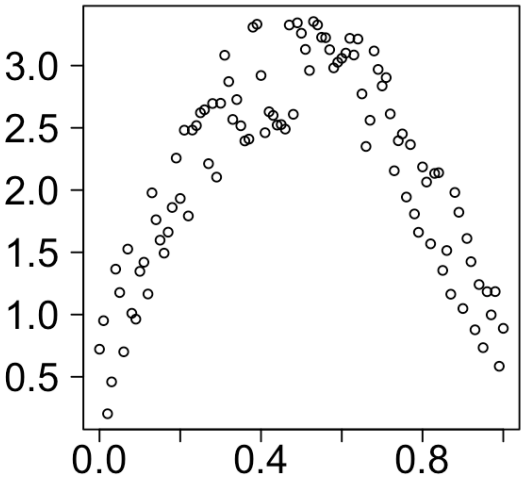
dove a è l'intercetta e b è la pendenza. (Trovale sulla figura seguente)

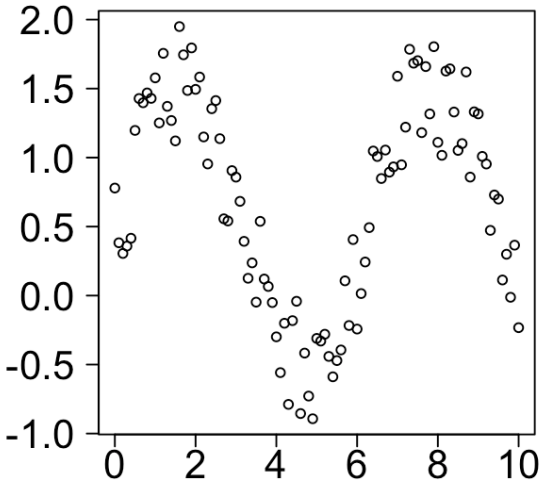


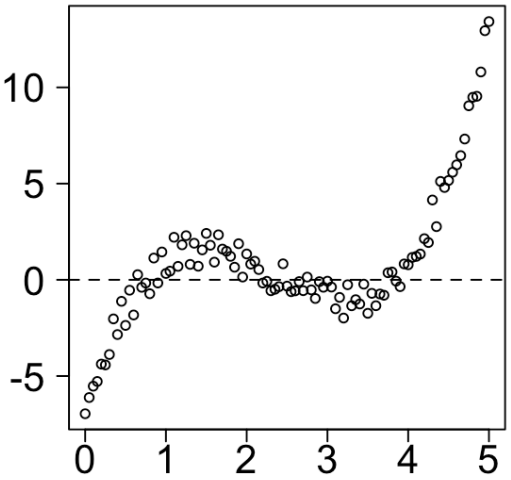
Relazioni statistiche

Segnale più *rumore*

Y = f(x) + residuo







Variabile dipendente e variabile esplicativa

- In **matematica**: y variabile dipendente, x variabile **indipendente**
- In **statistica**: y variabile dipendente, x variabile **esplicativa**.
- Le due variabili statistiche sono su due piani diversi

Esempio 1

Nel 1660, il fisico Robert Hooke ha enunciato la relazione esistente tra la forza F applicata a una molla e la sua deformazione d

La forza è proporzionale alla deformazione: $F = kd$

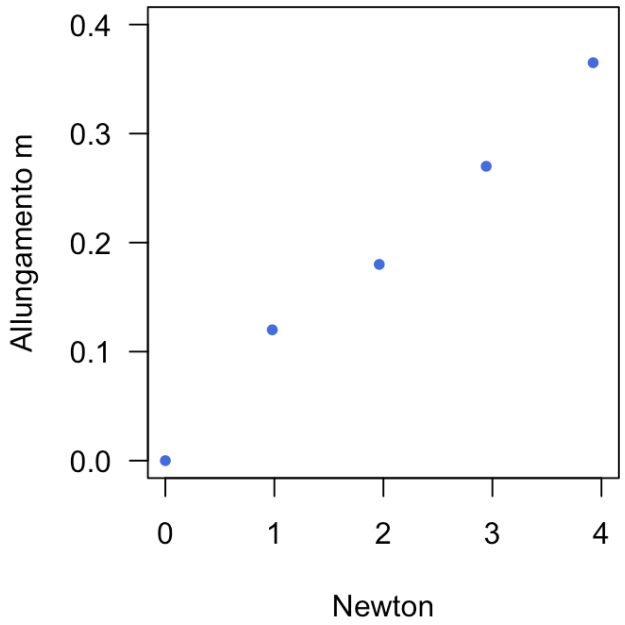
Relazione lineare esatta?

Esperimento

Applicare pesi x a una molla registrando l'allungamento y con uno strumento

Risultati dell'esperimento

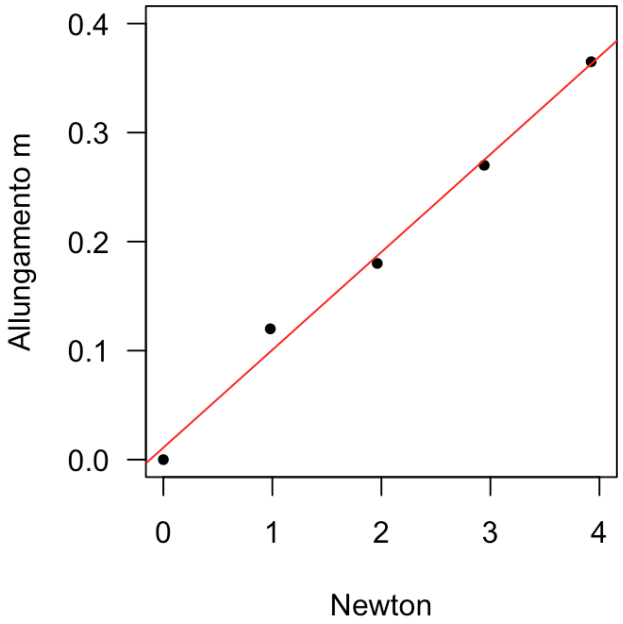
	Kg	Newton	Allungamento
1	0.0	0.000	0.000
2	0.1	0.981	0.120
3	0.2	1.962	0.180
4	0.3	2.943	0.270
5	0.4	3.924	0.365



- Quindi si rappresentano con un

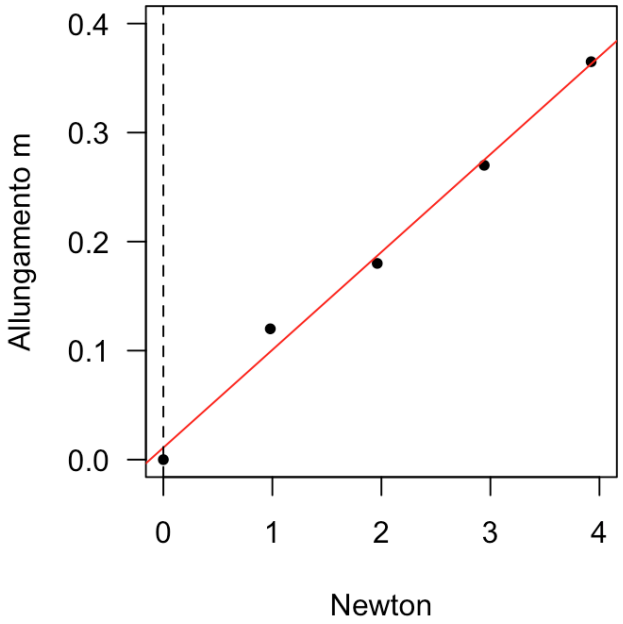
La risposta è che nella realtà ci sono imprecisioni dovute ad errori di misura *che vanno tenute presente*

Stima del coefficiente di elasticità



- Si adatta la retta dei minimi quadrati, la retta che passa fra i punti ed è ad essi più vicina
- Si stimano i due coefficienti a e b della retta dai dati:
  $\text{Lunghezza} = 0.011 + 0.089\ \text{Forza}$

Interpretazione dei risultati

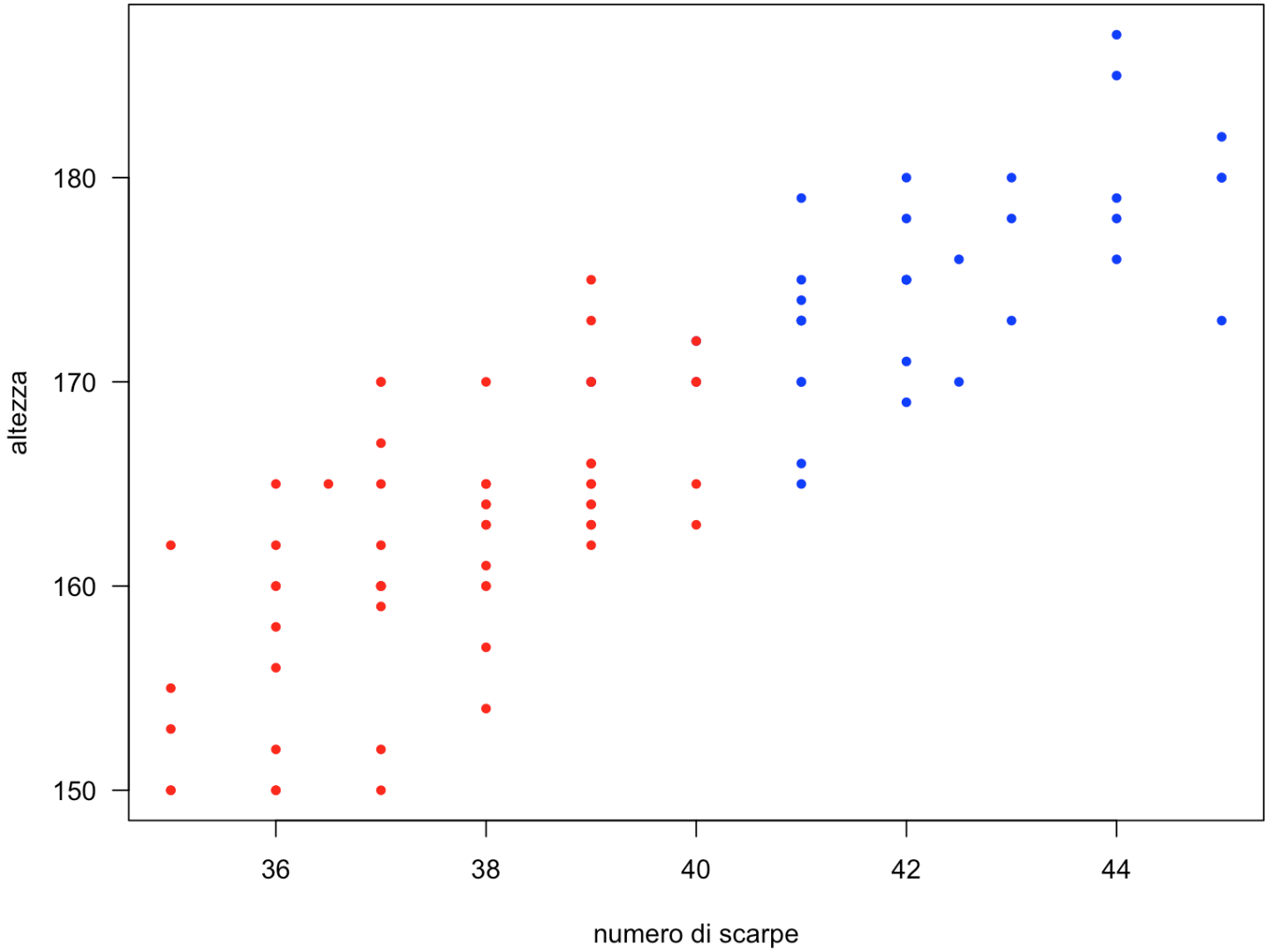


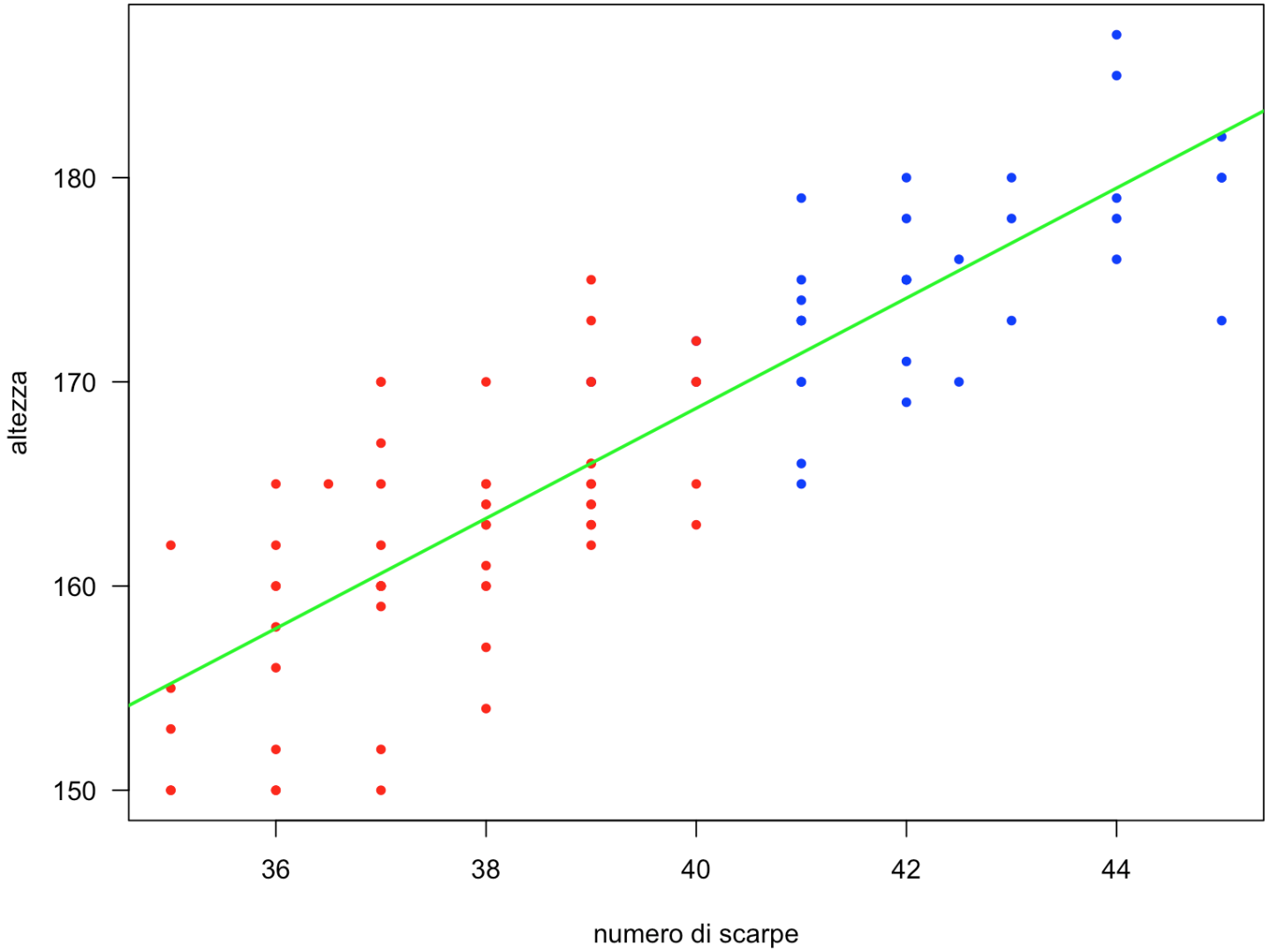
- $a = 0.011$ cm è l'allungamento della molla quando non viene applicata nessuna forza
- $b = 8.9$ cm/N è di quanti cm si deforma la molla per ogni incremento della forza di 1 N.
- La costante elastica è $k = 1/b = 1/(0.089) = 11.2$ N/m

Esempio 2 - C'è una relazione tra altezza e numero di scarpe?

- Si possono raccogliere i dati facilmente in classe
- La relazione contiene più eterogeneità di prima
- C'è evidenza di una relazione crescente tra altezza e numero di scarpa

```
  sesso scarpe altezza
1     1     39     170
2     2     40     170
3     2     37     162
4     2     38     160
5     2     38     157
6     1     42     169
```


altezza
180
170
160
150
36
38
40
42
44
numero di scarpe


altezza
180
170
160
150
36
38
40
42
44
numero di scarpe

Relazione tra altezza e n. di scarpa

Interpretazione

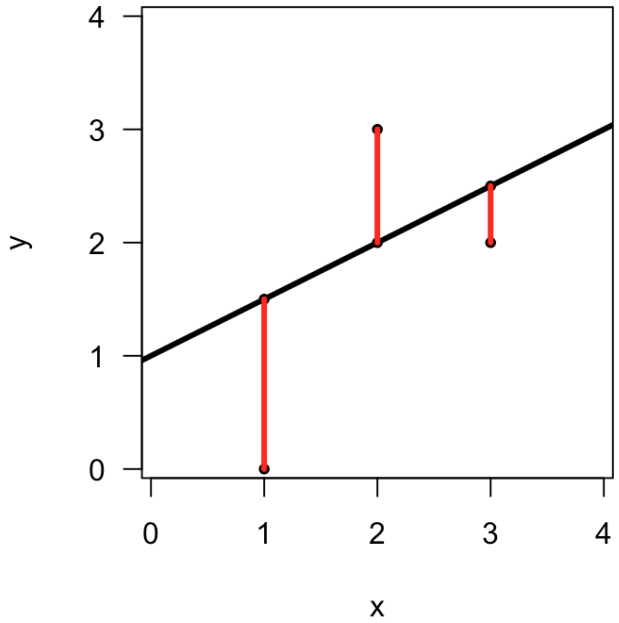
$$\text{altezza} = 60.9 + 2.7 \cdot \text{n. scarpa}$$

$b = 2.7$ cm in più per ogni numero di scarpa in più.

Minimi quadrati

Minimizzare $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ cioè

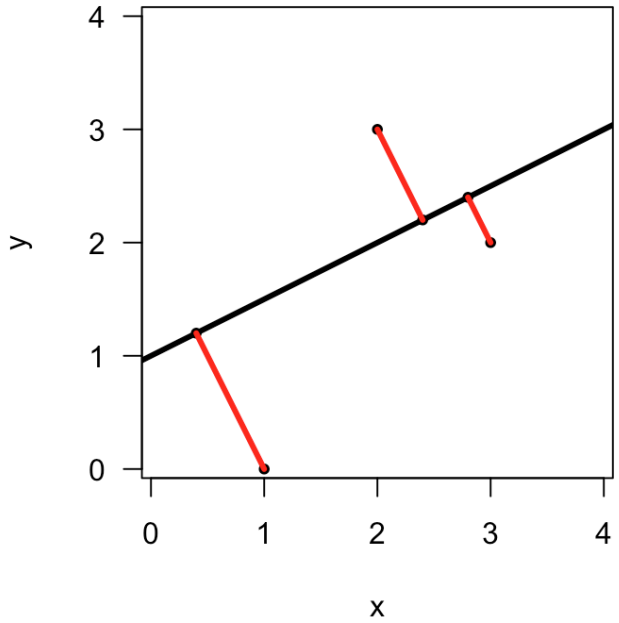
la somma dei quadrati delle distanze **verticali** tra i punti e la retta



Minimi quadrati

Minimizzare $\sum_{i=1}^{n}(y_i - a - bx_i)^2$ NON

la somma dei quadrati delle distanze tra i punti e la retta



Minimi quadrati

La retta dei minimi quadrati ha **due proprietà**

- Passa sicuramente per il punto $(\bar{x}, \bar{y})$ (media di X, media di Y). Quindi

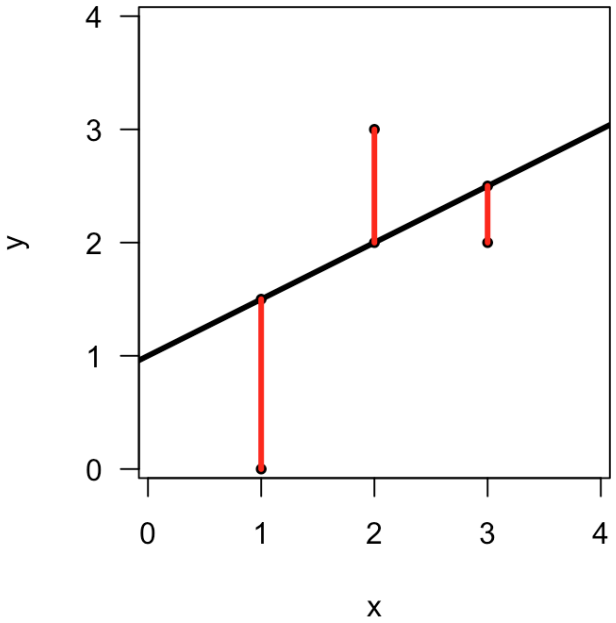
$$y = \bar{y} + b(x - \bar{x})$$

- ha pendenza

$$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}$$

Devianza residua

- La somma dei quadrati minimizzata si chiama **devianza residua** e misura la distanza verticale complessiva della retta dai punti



Esempio di calcolo

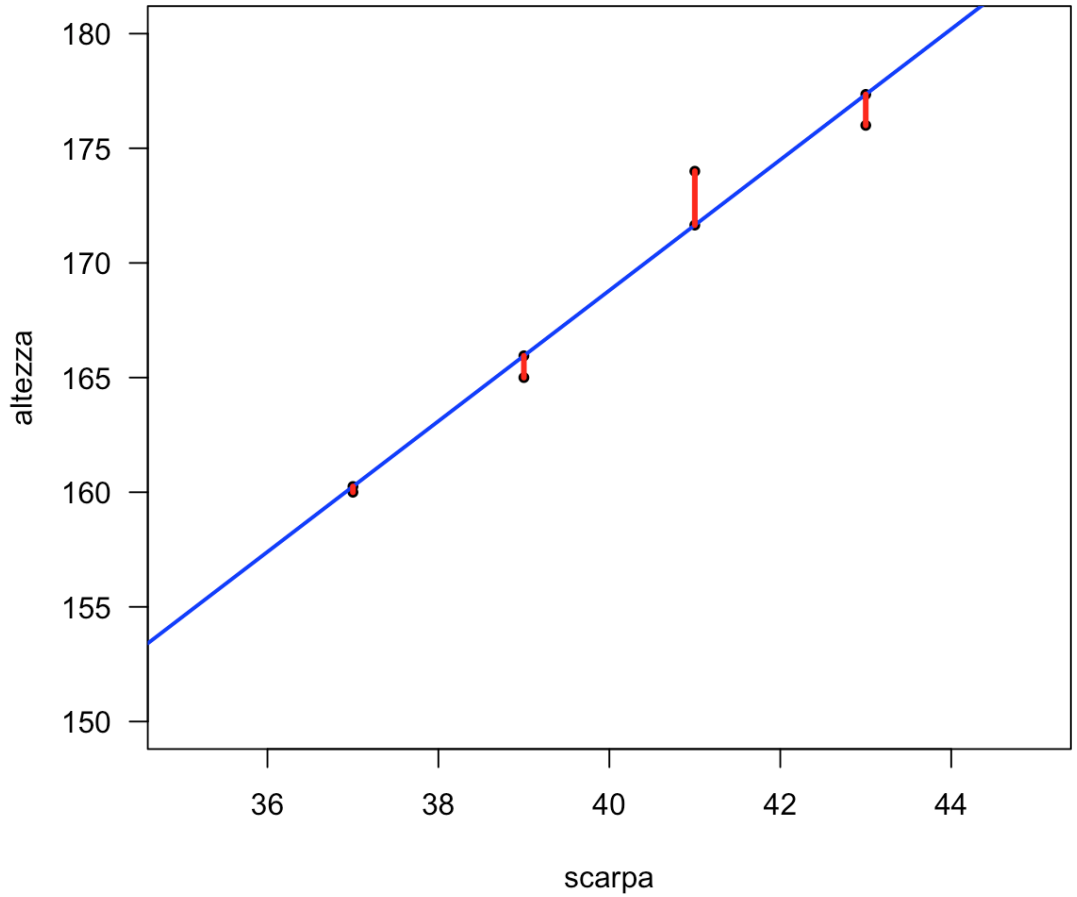
scarpa	altezza	scarti	scarti^2	altezza*scarti
37	160	-3	9	-480
39	165	-1	1	-165
41	174	1	1	174
43	176	3	9	528
	168.8		*20*	*57*

- Dunque $b = 57/20 = 2.85$ e quindi la relazione è

$$\text{altezza} = 168.8 + 2.85(\text{scarpa} - 40)$$

Grafico

- Con più dati il calcoli conviene farli con un computer

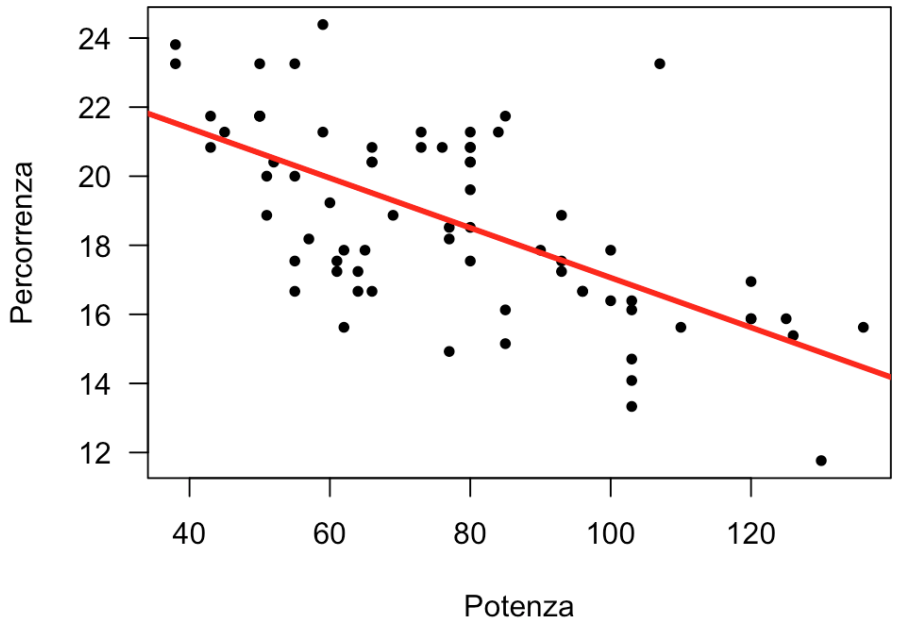


Esempio 3

Abbiamo visto situazioni in cui a valori sopra la media di X corrispondono valori sopra la media di Y

Esistono anche situazioni in cui le variabili hanno un andamento discordante

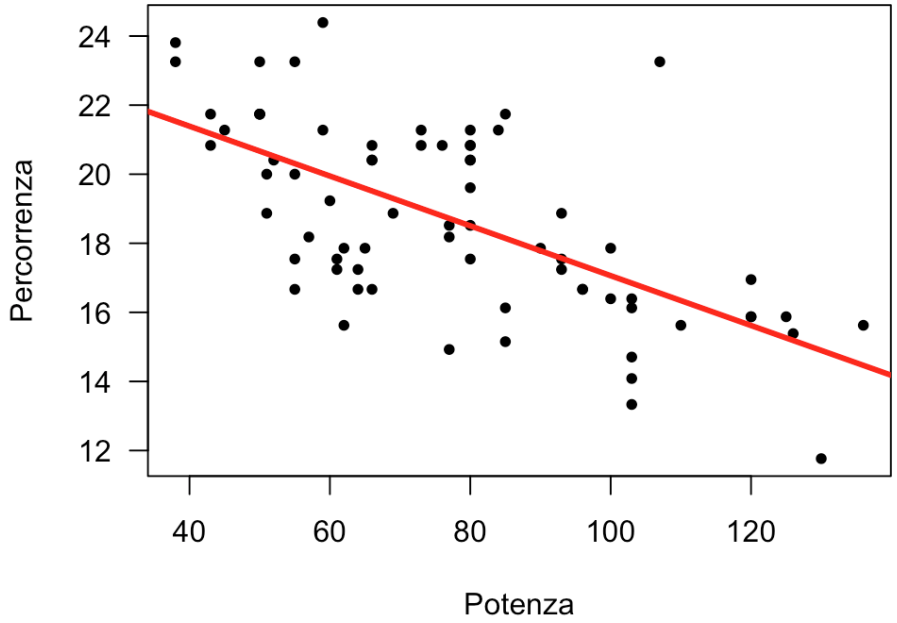
Relazione tra percorrenza di un auto (km/l) e potenza (kW)



$$\text{Percorrenza (km/l)} = 18.7 - 0.072\ (\text{Potenza (kW)} - 77.6)$$

Interpretazione

Relazione tra percorrenza di un auto (km/l) e potenza (kW)



$$\text{Percorrenza (km/l)} = 18.7 - 0.072\ (\text{Potenza (kW)} - 77.6)$$

per ogni 10 kW in più si percorrono in media 720 m in meno.

Variabilità residua

- La retta dei minimi quadrati si chiama anche **retta di regressione** (Galton)
- La sua pendenza si chiama **coefficiente di regressione** di Y da X
- Il metodo postula l'esistenza di una **variabilità attorno alla retta** che viene misurata con la somma dei quadrati dei residui:

$$\sum_i (y_i - [a + bx_i])^2$$

- la **forza dell'associazione lineare** si misura con un indice collegato che si chiama **coefficiente di correlazione lineare**.

Correlazione

Standardizzazione

Ricordiamo che una variabile X si **standardizza** se si centra rispetto alla media e si riscala con la deviazione standard

$$z_x = \frac{x - \text{media}(x)}{\sqrt{\text{var}(x)}}$$

- La media di z_x è zero e la varianza è 1

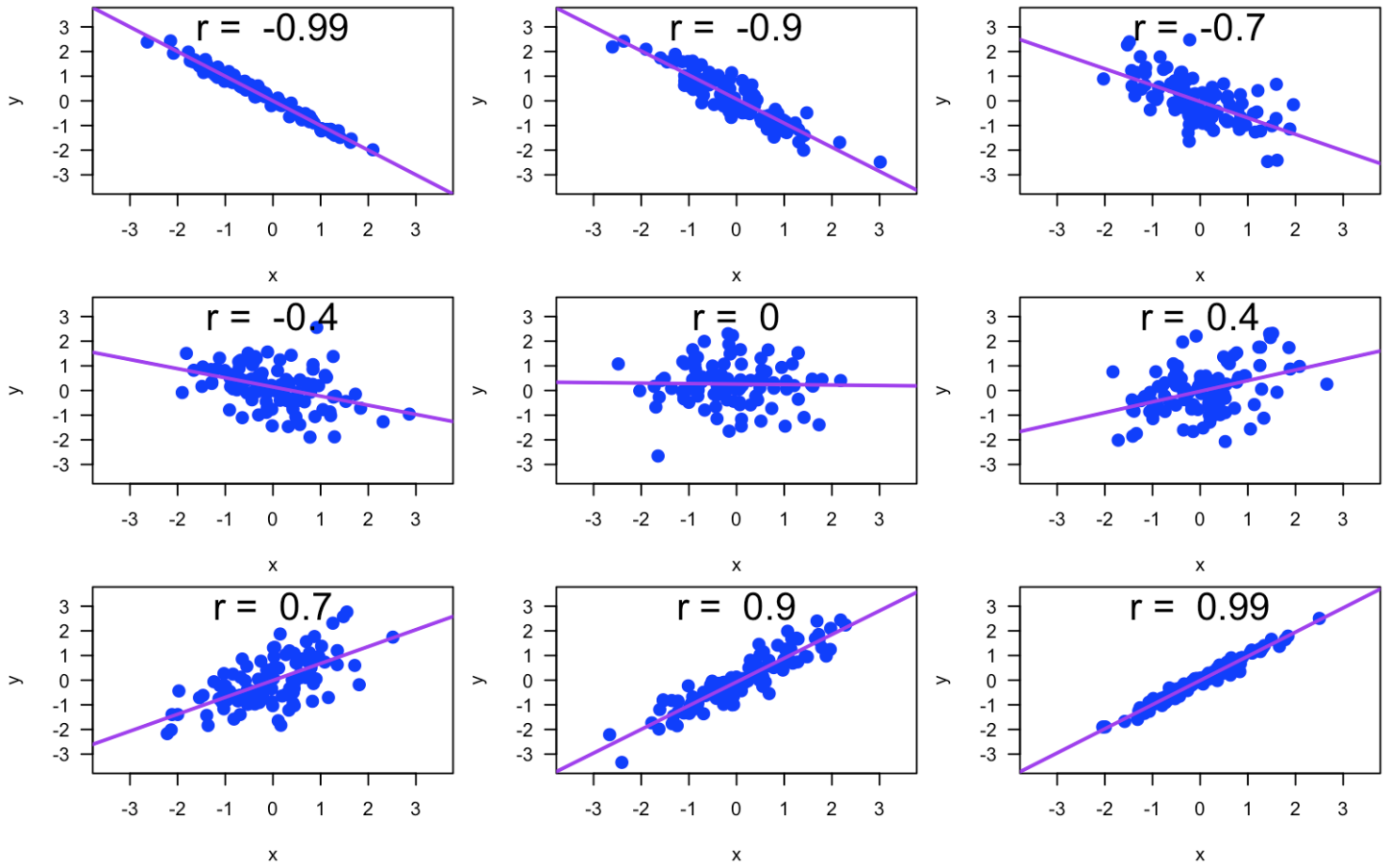
Coefficiente di correlazione

- Il **coefficiente di correlazione è la pendenza della retta di regressione tra Y e X dopo averle standardizzate entrambe**
- È uguale a

$$r = b\sqrt{\frac{\mathrm{var}(X)}{\mathrm{var}(Y)}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

-È normalizzato: $-1 \leq r \leq 1$ ed è un numero puro

Esempi di correlazione lineare



Confronto tra indici

- Il coefficiente di regressione è un indice **asimmetrico** cioè cambia se si scambiano variabile dipendente ed esplicativa:

$$b(y,x) \neq b(x,y)$$

  Per esempio il coefficienti di regressione

- tra altezza e n. di scarpe = 2.7 cm/numero
- tra n. di scarpe e altezza = 0.27 numero/cm
- NOTA non sono il reciproco l'uno dell'altro!

Confronto tra indici

- Il coefficiente di correlazione è un indice **simmetrico**

$$r(y, x) = r(x, y)$$

Per esempio il coefficiente di correlazione

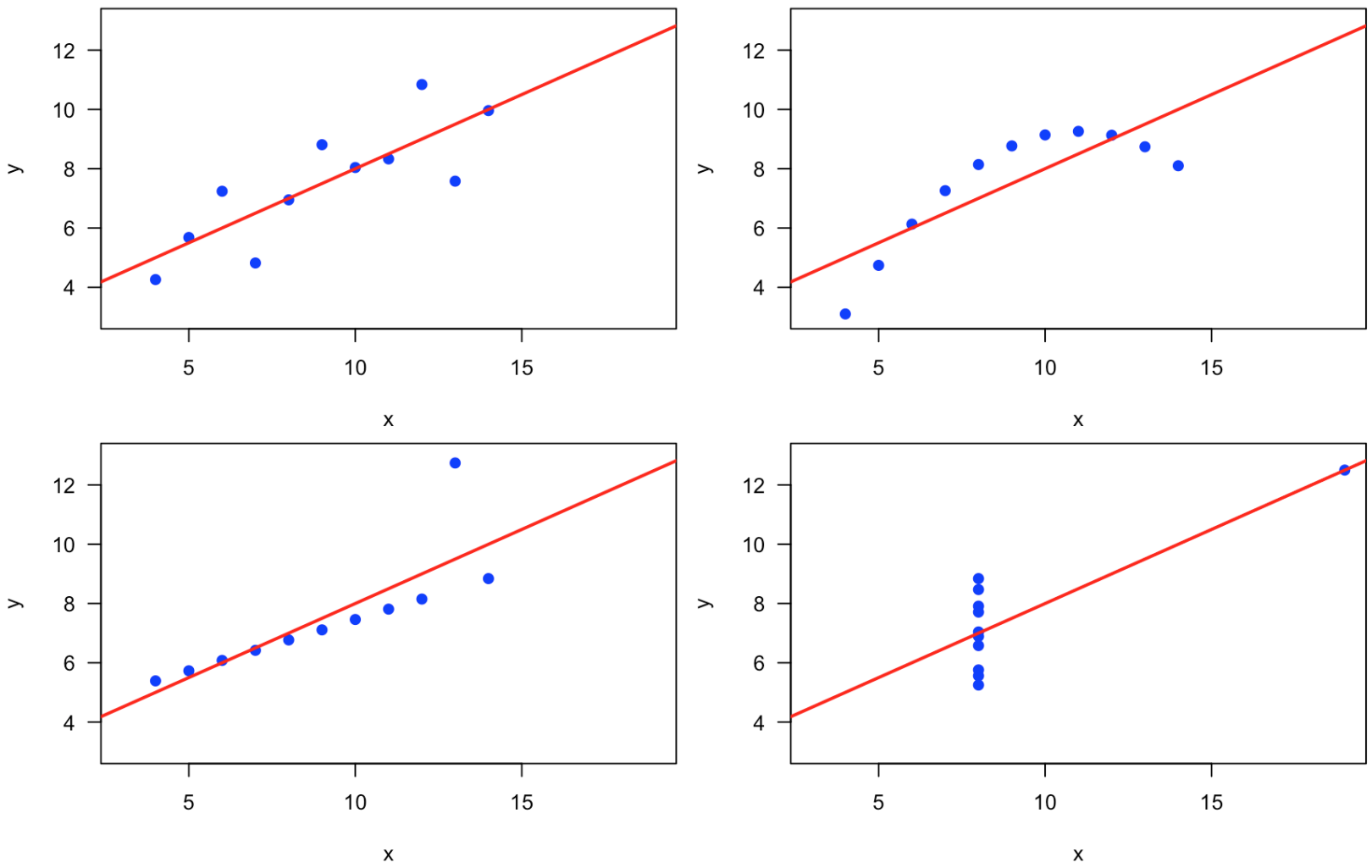
- tra altezza e n. di scarpe = 0.85 (numero puro)
- tra n. di scarpe e altezza = 0.85 (numero puro)

Esempi da ricordare!

Dati di Anscombe

- Stessa retta di regressione $y = 3 + 0.5x$
- Stesso coefficiente di correlazione $r = 0.82$

	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.10	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.10	4	5.39	19	12.50
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89


12
10
8
6
4
y
5
10
15
x
12
10
8
6
4
y
5
10
15
x
12
10
8
6
4
y
5
10
15
x
12
10
8
6
4
y
5
10
15
x

Alcuni strumenti utili sul web

- Inventori della regressione e della correlazione
- Francis Galton (1822-1911, cugino di Darwin)
- Karl Pearson (1857-1936)
- Dimostrazioni
- Mathematica Demonstration
- Applet

Sommario

- Date due variabili Y e X vogliamo studiare la dipendenza di Y da X.
- Spesso la relazione è lineare
- Utile per prevedere Y conoscendo X
- La retta si può stimare con il metodo dei minimi quadrati
- La retta si chiama **retta di regressione** di Y da X
- La pendenza misura la variazione della media Y per un incremento unitario di X.
- La retta passa fra i punti e permette l'esistenza di una **variabilità attorno alla retta** che va misurata.

Dipendenza e indipendenza

Tabelle a doppia entrata

- Orientamento politico e genere in USA, 2004

Orientamento	Democratici	Indipendenti	Repubblicani	Sum
Sesso				
F	573	516	422	1511
M	386	475	399	1260
Sum	959	991	821	2771

- Abitudini al fumo e all'alcool in una scuola

Alcool	Sì	No	Sum
Sigarette			
Sì	1449	500	1949
No	46	281	327
Sum	1495	781	2276

Costruzione

- Imparare a costruirle su esempi facili
- Costruirle raccogliendo dati
  - in classe
  - Dati GSS

```
   Sesso Fumo
1      F   no
2      M   no
3      F   sì
4      M   sì
5      F   sì
6      M   sì
7      M   sì
8      M   sì
9      M   no
10     F   sì
11     M   sì
12     M   sì
```

```
      Fumo no sì Sum
Sesso
F          1  3   4
M          2  6   8
Sum        3  9  12
```

Che domande ci possiamo fare?

- Come cambia l'orientamento politico tra maschi e femmine?
- La proporzione di fumatori nella classe è la stessa tra maschi e femmine?
- C'è una associazione tra consumo di alcool e fumo?
- Secondo voi c'è associazione tra sesso e proporzione di disoccupati?
- La proporzione di condannati alla pena capitale è la stessa per bianchi e neri?

Distribuzioni condizionate

- Come cambia l'orientamento politico tra maschi e femmine?

Sesso	Orientamento Democratici	Indipendenti	Repubblicani	Sum
F	573	516	422	1511
M	386	475	399	1260

Sesso	Orientamento Democratici	Indipendenti	Repubblicani	Sum
F	37.9	34.1	27.9	100.0
M	30.6	37.7	31.7	100.0

Si calcolano separatamente per i maschi e le femmine le proporzioni di democratici, indipendenti e repubblicani

Distribuzioni condizionate

- Come cambia la proporzione di femmine a seconda dell'orientamento politico?

Orientamento	Democratici	Indipendenti	Repubblicani
Sesso			
F	573	516	422
M	386	475	399
Sum	959	991	821

	Orientamento		
Sesso	Democratici	Indipendenti	Repubblicani
F	59.7	52.1	51.4
M	40.3	47.9	48.6
Sum	100.0	100.0	100.0

Si calcolano separatamente per orientamento le proporzioni di maschi e femmine

Indipendenza

- Se le proporzioni di maschi e femmine sono le stesse per ogni orientamento politico, l'orientamento è **indipendente** dal genere.
- In caso contrario c'è **associazione**.

	Orientamento		
Sesso	Democratici	Indipendenti	Repubblicani
F	59.7	52.1	51.4
M	40.3	47.9	48.6
Sum	100.0	100.0	100.0

"Evidentemente" c'è associazione tra genere e orientamento politico.

Esempio 1

Sesso	Fumo no	sì	Sum
F	1	3	4
M	2	6	8
Sum	3	9	12

Sesso	Fumo no	sì	Sum
F	25	75	100
M	25	75	100
Sum	25	75	100

Sesso	Fumo no	sì	Sum
F	33.3	33.3	33.3
M	66.7	66.7	66.7
Sum	100.0	100.0	100.0

Esempio 2

- Associazione tra pena di morte e colore della pelle (dati fittizi ma realistici)

Morte	sì	no	Sum
Razza			
bianca	7	63	70
nera	3	27	30
Sum	10	90	100

	Morte		
Razza	sì	no	Sum
bianca	10	90	100
nera	10	90	100

Indice tipico: rischio relativo

- È dato da

$$RR = \frac{\Pr(Insuccesso \mid Condizione1)}{\Pr(Insuccesso \mid Condizione2)}$$

- Nel caso della pena di morte

$$\text{RR} = \frac{10/100}{10/100} = 1$$

  La proporzione di condannati è uguale per i bianchi e per i neri.

- Indipendenza tra razza e pena capitale

Esempio 3

Incidenti (gravi) in un anno calssificati a seconda dell'esito per il conducente e l'uso delle cinture.

Esito	sopravvissuto	morto	Sum
Cinture			
sì	412368	510	412878
no	162527	1601	164128
Sum	574895	2111	577006

Esito			
Cinture	sopravvissuto	morto	Sum
sì	99.9	0.1	100.0
no	99.0	1.0	100.0

Interpretazione

$$\text{RR} = \frac{1.0}{0.1} = 10$$

La proporzione di deceduti è 10 volte più alta per quelli che non indossano le cinture di sicurezza

- Notare che RR $= 1$ rappresenta l'indipendenza
- L'allontanamento da 1 rappresenta associazione
- RR $= 10$ e RR $= 0.1$ rappresentano lo stesso grado di associazione.

Esempio di uso del rischio relativo

Secondo i dati dal 1999 al 2006 in USA la probabilità di morire per i maschi di 19 anni è 0.00135 e la stessa probabilità per le femmine è 0.00046.

Qual è il rischio relativo?

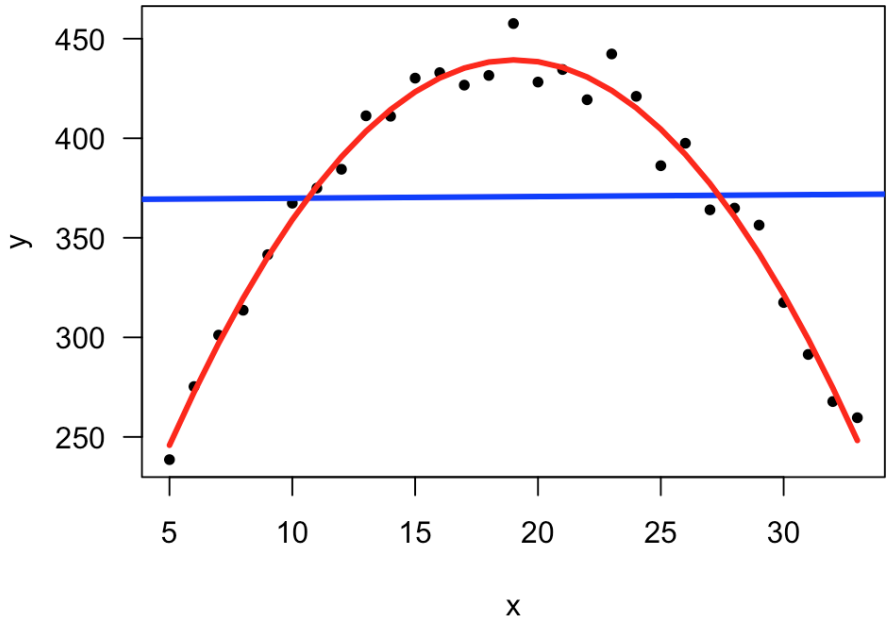
$$\text{RR} = \frac{0.00135}{0.00046} = 2.9$$

Il rischio è circa 3 volte più alto per i maschi

Relazioni spurie e distorte

Alcune convinzioni errate

- Se due variabili sono incorrelate allora non può esserci associazione fra esse
-



- Coefficiente di correlazione 0.01154

Alcune convinzioni errate

- Se due variabili sono molto associate allora ci deve essere una relazione di causa-effetto fra loro
- Le dita gialle sono molto associata al tumore ai polmoni, ma non sono la causa

Fumo	sì			no		
Tumore	sì	no	Sum	sì	no	Sum
Dita						
gialle	50	450	500	2	98	100
no	10	90	100	6	294	300
Sum	60	540	600	8	392	400

Tumore	sì	no	Sum
Dita			
gialle	52	548	600
no	16	384	400
Sum	68	932	1000

Tumore	sì	no	Sum
Dita			
gialle	8.7	91.3	100.0
no	4.0	96.0	100.0
Sum	6.8	93.2	100.0

Tumore	sì	no	Sum
Dita			
gialle	8.7	91.3	100.0
no	4.0	96.0	100.0
Sum	6.8	93.2	100.0

- La proporzione di tumore ai polmoni è più del doppio per quelli che hanno le dita gialle $\mathrm{RR} = \frac{8.7}{4.0} = 2.2$

Fumo	sì			no		
Tumore	sì	no	Sum	sì	no	Sum
Dita						
gialle	50	450	500	2	98	100
no	10	90	100	6	294	300
Sum	60	540	600	8	392	400

- La proporzione di tumori è il 2% per chi non fuma e del 10% per chi fuma indipendentemente dalle dita

Alcune convinzioni errate

Se due variabili sono indipendenti sicuramente non ci può essere una relazione fra di esse

La pena capitale appare indipendente dalla razza ma non in realtà lo è

Morte	sì	no	Sum
Accusato			
bianco	7	63	70
nero	3	27	30
Sum	10	90	100

	Morte		
Accusato	sì	no	Sum
bianco	10	90	100
nero	10	90	100

Se si scompongono i casi a seconda della razza della vittima la pena capitale dipende fortemente dalla razza dell'accusato

Vittima	Accusato	Morte sì	no	Sum
bianco	bianco	5	45	50
	nero	2	8	10
nero	bianco	2	18	20
	nero	1	19	20

Vittima	Accusato	Morte sì	no	Sum
bianco	bianco	10	90	100
	nero	20	80	100
nero	bianco	10	90	100
	nero	5	95	100

- Se l'accusato è bianco il rischio di essere condannato è il 10%
- Se l'accusato è nero e la vittima è un bianco il rischio è il doppio.

La dipendenza in senso statistico va saputa intendere

- In una classe si fa un dettato in francese e si dà il voto y (in base agli errori commessi). Poi si calcola la correlazione tra peso x dello studente e y
- La ricerca non ha senso, ma se la classe è composta di bambini delle elementari, e ragazzi delle medie e delle superiori si troverà probabilmente una correlazione positiva tra x e y
- Anche se prendiamo ragazzi/e tutti della stessa classe delle superiori c'è un problema: le ragazze pesano in media meno dei ragazzi, ma hanno voti in media più alti.
- Stavolta troviamo una correlazione inversa!
- In entrambi i casi la correlazione non ha un'interpretazione causale

Paradosso di Simpson

Sesso	Tratt	OK	sì	no	Sum
m	A		1500	2250	3750
	B		375	875	1250
f	A		1000	250	1250
	B		2625	1125	3750

Sesso	Tratt	OK	sì	no	Sum
m	A		40	60	100
	B		30	70	100
f	A		80	20	100
	B		70	30	100

- Il trattamento A è meglio sia per i maschi che per le femmine

Il trattamento fa bene o male?

Mettiamo insieme maschi e femmine

```
      OK    sì    no   Sum
Tratt
A         2500  2500  5000
B         3000  2000  5000
Sum       5500  4500 10000
```

```
     OK
Tratt  sì  no Sum
    A  50  50 100
    B  60  40 100
```

- Stavolta il trattamento B è meglio in complesso!