Relazioni statistiche tra 2 variabili quantitative

M.M.I.B. 2019/20
corso di laurea in Scienze Biologiche
Sapienza Università di Roma

Esempio 1

Un'azienda vuole indagare il rapporto tra temperatura ambientale e numero di parti difettose prodotte.

Esempio 1

Un'azienda vuole indagare il rapporto tra temperatura ambientale e numero di parti difettose prodotte.

Per 22 giorni si registrano le temperature massime e il numero di difetti riscontrati, e si definiscono le osservabili x_i e y_i rispettivamente temperatura e numero di difetti riscontrati.

Esempio 1

Un'azienda vuole indagare il rapporto tra temperatura ambientale e numero di parti difettose prodotte.

Per 22 giorni si registrano le temperature massime e il numero di difetti riscontrati, e si definiscono le osservabili x_i e y_i rispettivamente temperatura e numero di difetti riscontrati.

L'insieme dei dati (x_i, y_i) costituisce un campione **bivariato**.
I dati del campione bivariato sono rappresentati attraverso un **diagramma di dispersione**.

Scatter plot

Il diagramma di dispersione, detto **scatter plot** nella letteratura scientifica, permette di valutare qualitativamente se esiste una corispondenza, un'associazione, tra i dati.

Scatter plot

Il diagramma di dispersione, detto **scatter plot** nella letteratura scientifica, permette di valutare qualitativamente se esiste una corispondenza, un'associazione, tra i dati.

Si riportano in ascissa le misure osservate $x_1, x_2, \ldots, x_n$ della variabile x, in ordinata le corrispondenti misure osservate $y_1, y_2, \ldots, y_n$ della variabile y.

Scatter plot

Il diagramma di dispersione, detto **scatter plot** nella letteratura scientifica, permette di valutare qualitativamente se esiste una corispondenza, un'associazione, tra i dati.

Si riportano in ascissa le misure osservate $x_1, x_2, \ldots, x_n$ della variabile x, in ordinata le corrispondenti misure osservate $y_1, y_2, \ldots, y_n$ della variabile y.

Le singole osservazioni (x_i, y_i), cioè i dati del campione bivariato, vengono così rappresentate tramite dei punti su un piano cartesiano.

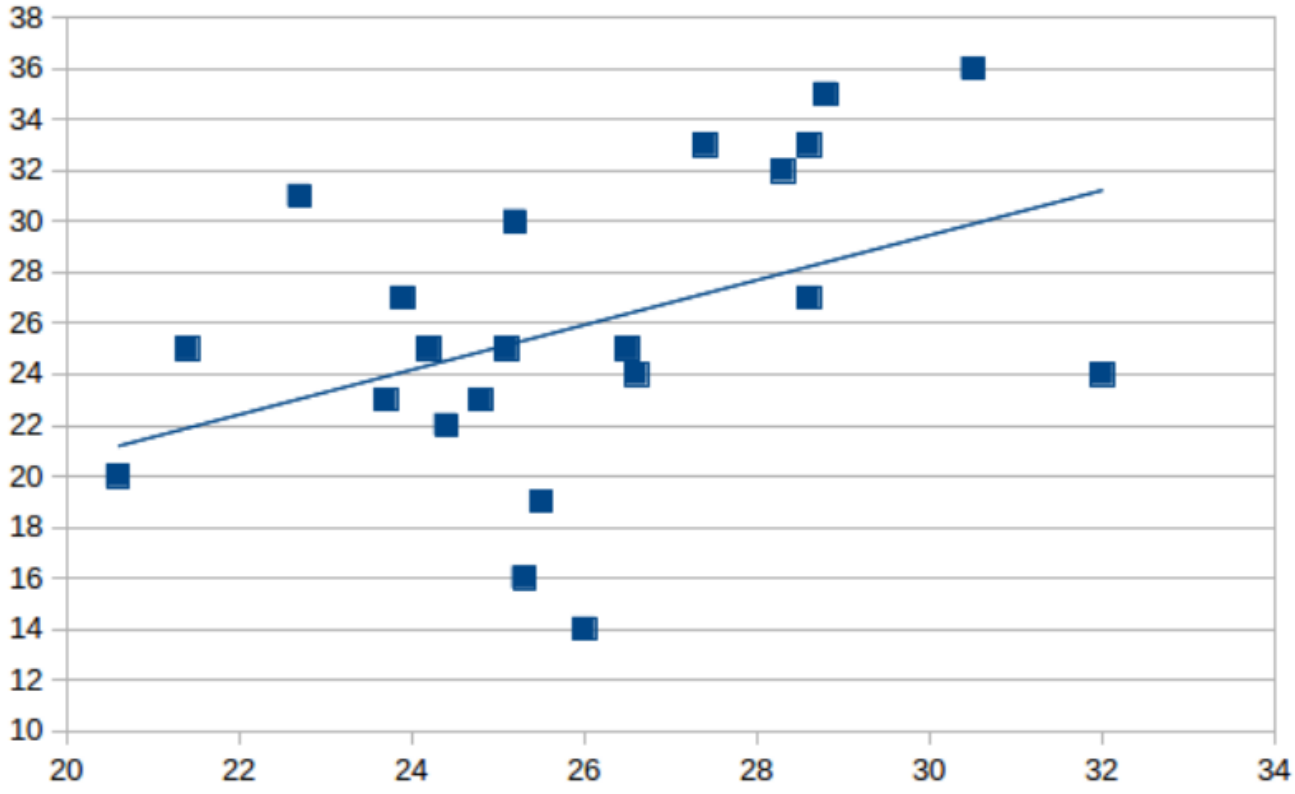
Esempio 1

Temperatura (Centigradi)	N parti difettose	Temper	N parti difettose
24.2	25	24.8	23
22.7	31	20.6	20
30.5	36	25.1	25
28.6	33	21.4	25
25.5	19	23.7	23
32.0	24	23.9	27
28.6	27	25.2	30
26.5	25	27.4	33
25.3	16	28.3	32
26.0	14	28.8	35
24.4	22	26.6	24

Esempio 1

Temperature	N parti difettose	Temperature	N parti difettose
20.6	20	25.5	19
21.4	25	26	14
22.7	31	26.5	25
23.7	23	26.6	24
23.9	27	27.4	33
24.2	25	28.3	32
24.4	22	28.6	33
24.8	23	28.6	27
25.1	25	28.8	35
25.2	30	30.5	36
25.3	16	32	24

Esempio 1



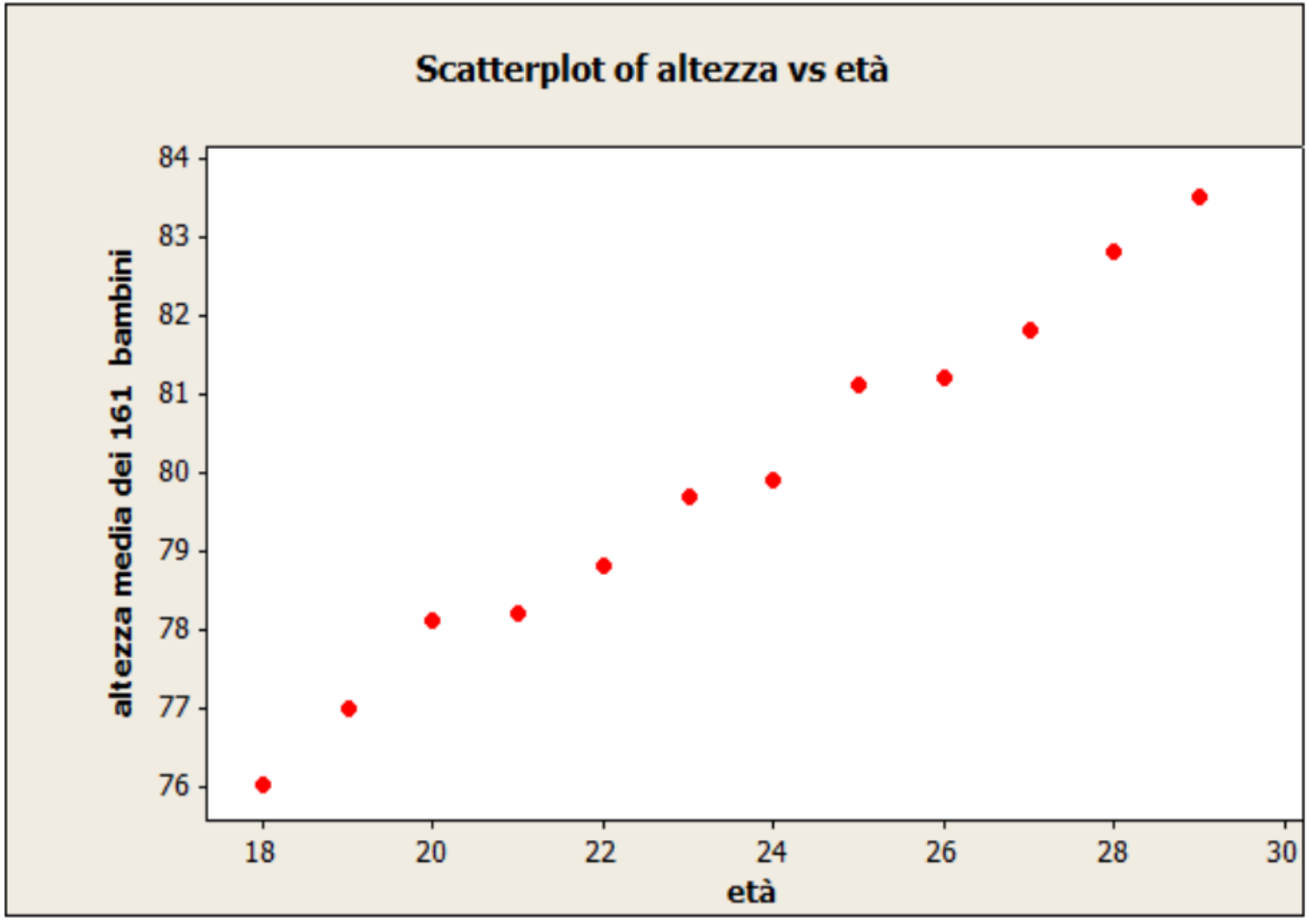
Esempio 2

Riportiamo i dati ralativi ad un gruppo di 161 bambini del villaggio di Kalama (Egitto), tali dati consistono nell'altezza media (in centimetri) del gruppo di bambini e nell'età (in mesi).

Esempio 2

età	altezza
18	76.01
19	77.00
20	78.10
21	78.20
22	78.80
23	79.70
24	79.90
25	81.10
26	81.20
27	81.80
28	82.80
29	83.50

Esempio 2


Scatterplot of altezza vs età
altezza media dei 161 bambini
84
83
82
81
80
79
78
77
76
18
20
22
24
26
28
30
età

Interpretazione del grafico

Si cerca di capire l'andamento generale (**trend**): in questo caso i punti sono disposti attorno a una retta (ci aspettiamo una relazione lineare)

Interpretazione del grafico

Si cerca di capire l'andamento generale (**trend**): in questo caso i punti sono disposti attorno a una retta (ci aspettiamo una relazione lineare)

Ad alti valori dell'età corrispondono alti valori dell'altezza (**correlazione positiva**)

Interpretazione del grafico

Si cerca di capire l'andamento generale (**trend**): in questo caso i punti sono disposti attorno a una retta (ci aspettiamo una relazione lineare)

Ad alti valori dell'età corrispondono alti valori dell'altezza (**correlazione positiva**)

L'intensità della correlazione è misurata da un indice numerico chiamato **coefficiente di correlazione**

Coefficiente di correlazione (campionaria)

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{s_x} \right] \left[\frac{y_i - \bar{y}}{s_y} \right]$$

Coefficiente di correlazione (campionaria)

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{s_x} \right] \left[\frac{y_i - \bar{y}}{s_y} \right]$$

Osserviamo che r è adimensionale, quindi, in particolare, non dipende dalle unità di misura delle variabili.

Coefficiente di correlazione (campionaria)

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{s_x} \right] \left[\frac{y_i - \bar{y}}{s_y} \right]$$

Osserviamo che r è adimensionale, quindi, in particolare, non dipende dalle unità di misura delle variabili.

Quando $r > 0$, i dati si dicono correlati positivamente

Quando $r < 0$, i dati si dicono correlati negativamente.

Osservazioni

Vale sempre che

$$-1 \leq r \leq +1$$

Osservazioni

Vale sempre che

$$-1 \leq r \leq +1$$

Se per tutte le coppie di dati (x_i, y_i) vale la relazione $y_i = ax_i + b$ con $a > 0$, allora $r = +1$,

Osservazioni

Vale sempre che

$$-1 \leq r \leq +1$$

Se per tutte le coppie di dati (x_i, y_i) vale la relazione $y_i = ax_i + b$ con $a > 0$, allora $r = +1$,

Se per tutte le coppie di dati (x_i, y_i) vale la relazione $y_i = ax_i + b$ con $a < 0$, allora $r = -1$,

Osservazioni

Vale sempre che

$$-1 \leq r \leq +1$$

Se per tutte le coppie di dati (x_i, y_i) vale la relazione $y_i = ax_i + b$ con $a > 0$, allora $r = +1$,

Se per tutte le coppie di dati (x_i, y_i) vale la relazione $y_i = ax_i + b$ con $a < 0$, allora $r = -1$,

I valori estremi di r sono raggiunti solo quando tra x e y c'è una relazione lineare.

Coefficiente di correlazione (campionaria)

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{s_x} \right] \left[\frac{y_i - \bar{y}}{s_y} \right]$$

Coefficiente di correlazione (campionaria)

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{s_x} \right] \left[\frac{y_i - \bar{y}}{s_y} \right]$$

Osserviamo che r è adimensionale, quindi, in particolare, non dipende dalle unità di misura delle variabili.

Coefficiente di correlazione (campionaria)

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{s_x} \right] \left[\frac{y_i - \bar{y}}{s_y} \right]$$

Osserviamo che r è adimensionale, quindi, in particolare, non dipende dalle unità di misura delle variabili.

Quando $r > 0$, i dati si dicono correlati positivamente

Quando $r < 0$, i dati si dicono correlati negativamente.

Esempio 3

Suppponiamo di sospettare che un certo farmaco abbia l'effetto indesiderato di innalzare la pressione arteriosa e cerchiamo di verificare questa affermazione tramite un esperimento.

Esempio 3

Suppponiamo di sospettare che un certo farmaco abbia l'effetto indesiderato di innalzare la pressione arteriosa e cerchiamo di verificare questa affermazione tramite un esperimento.

Formiamo 8 gruppi di 2 ratti a cui sommistriamo differenti dosi del farmaco: il primo gruppo non riceve alcuna sommistrazione e funge da gruppo di controllo, il secondo gruppo riceve una dose da $1mg/kg$, il terzo gruppo riceve una dose da $2mg/kg$, il quarto gruppo da $2mg/kg$ e via dicendo. I risultati sono riassunti nella seguente tabella

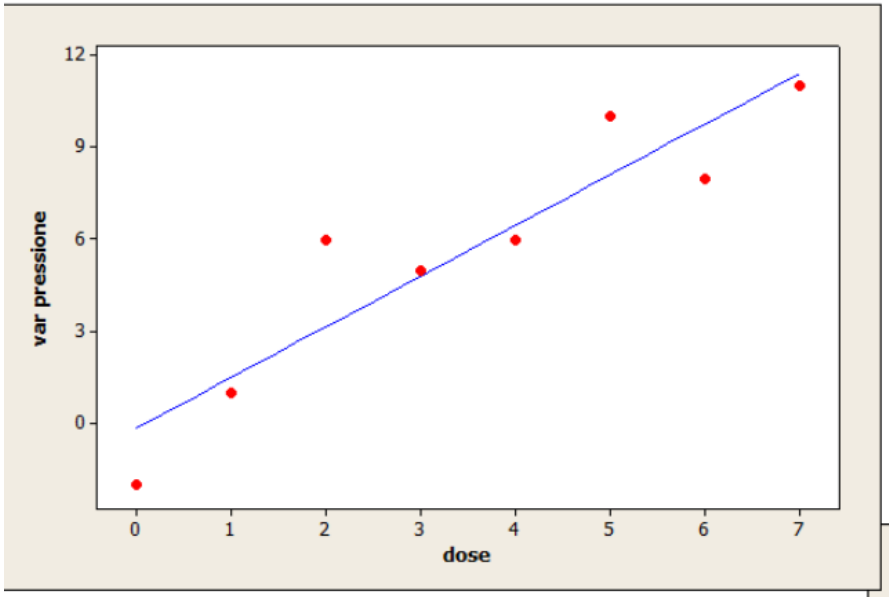
Esempio 3

Suppponiamo di sospettare che un certo farmaco abbia l'effetto indesiderato di innalzare la pressione arteriosa e cerchiamo di verificare questa affermazione tramite un esperimento.

Formiamo 8 gruppi di 2 ratti a cui sommistriamo differenti dosi del farmaco: il primo gruppo non riceve alcuna sommistrazione e funge da gruppo di controllo, il secondo gruppo riceve una dose da $1mg/kg$, il terzo gruppo riceve una dose da $2mg/kg$, il quarto gruppo da $2mg/kg$ e via dicendo. I risultati sono riassunti nella seguente tabella

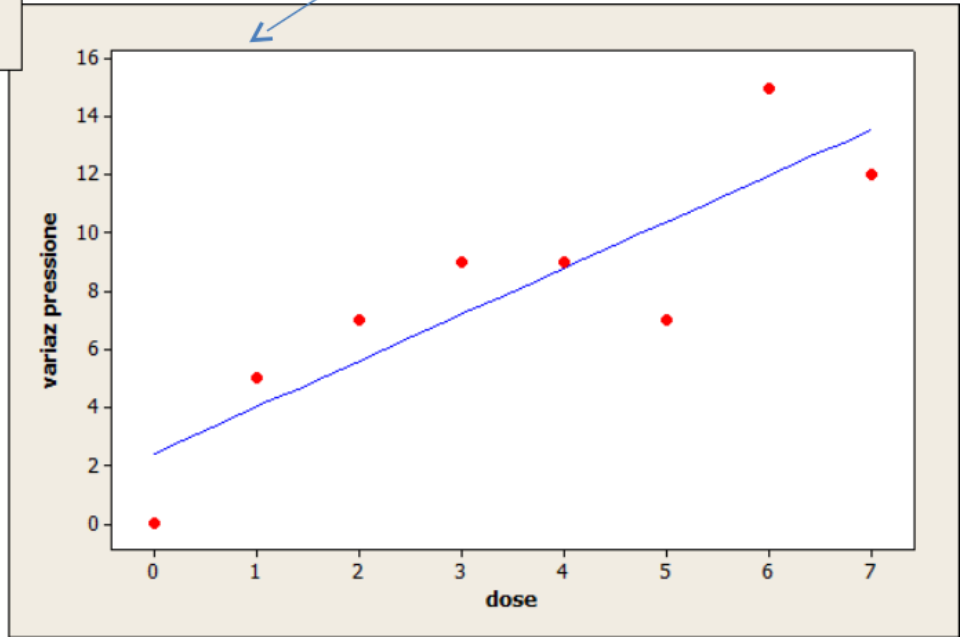
DOSE (mg/kg)	0	1	2	3	4	5	6	7
ratto 1	-2	1	6	5	6	10	8	11
ratto 2	0	5	7	9	9	7	15	12

Esempio 3



Ratto 1 r= 0.927

Ratto 2 r= 0.868



Esempio 3

Al crescere della dose la variazione di pressione aumenta, l'associazione lineare è positiva e forte in entrambi i casi, visto che abbiamo $r = 0.927$ e $r = 0.868$.

Esempio 3

Al crescere della dose la variazione di pressione aumenta, l'associazione lineare è positiva e forte in entrambi i casi, visto che abbiamo $r = 0.927$ e $r = 0.868$.

L'andamento generale è piuttosto regolare, si può pensare ad un modello (matematico, statistico) per descrivere tale andamento.

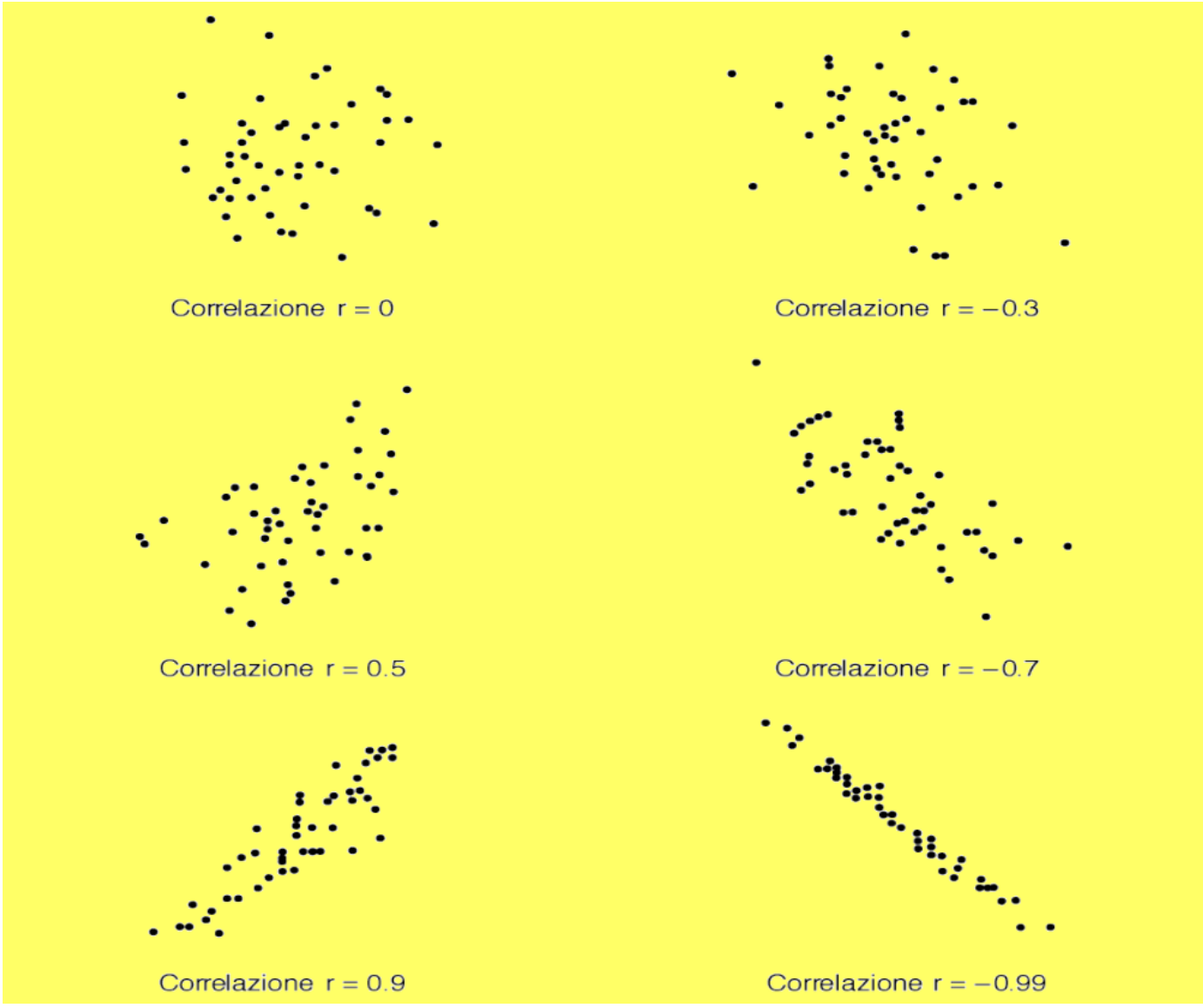
Esempio 3

Al crescere della dose la variazione di pressione aumenta, l'associazione lineare è positiva e forte in entrambi i casi, visto che abbiamo $r = 0.927$ e $r = 0.868$.

L'andamento generale è piuttosto regolare, si può pensare ad un modello (matematico, statistico) per descrivere tale andamento.

Potremmo usare il modello lineare suggerito dal grafico per predire la variazione di pressione al variare della dose del farmaco.

Osservazioni



Osservazioni

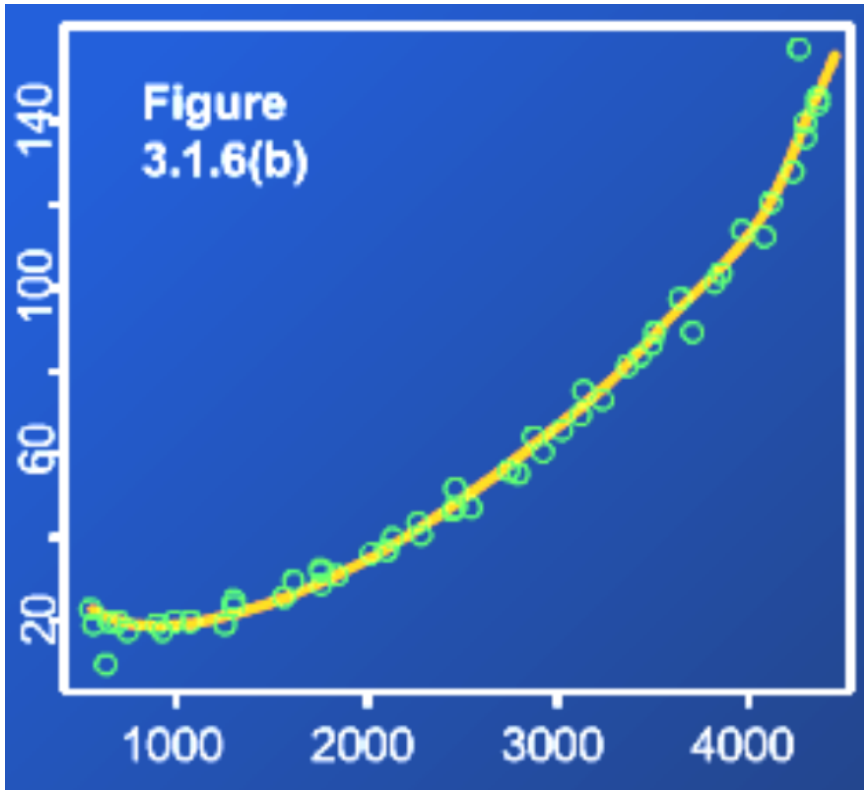
Una correlazione positiva non indica che una variabile influenza l'altra, ma solo che le 2 variabili evolvono nello stesso senso, la causa del comune andamento può essere un fatto non ancora preso in considerazione.

Osservazioni

Una correlazione positiva non indica che una variabile influenza l'altra, ma solo che le 2 variabili evolvono nello stesso senso, la causa del comune andamento può essere un fatto non ancora preso in considerazione.

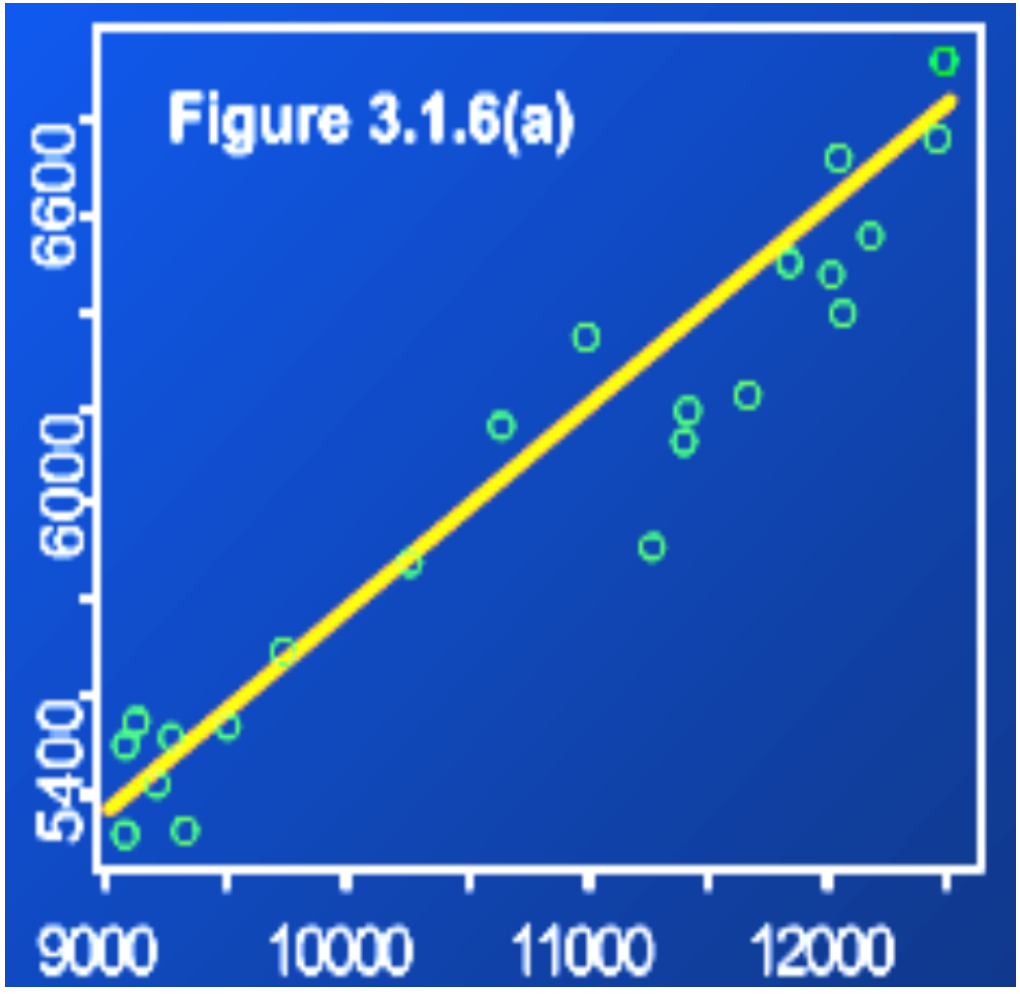
Per esempio se in un'indagine statistica si trova che il numero di figli per famiglia e il consumo di alcool pro capite per famiglia sono correlati positivamente, questo non vuol dire che una variabile influenza direttamente l'altra, la correlazione deve essere chiarita da ulteriori indagini e si può ipotizzare che la causa comune siano le condizioni economiche e culturali delle famiglie.

Cosa dice uno scatter plot?



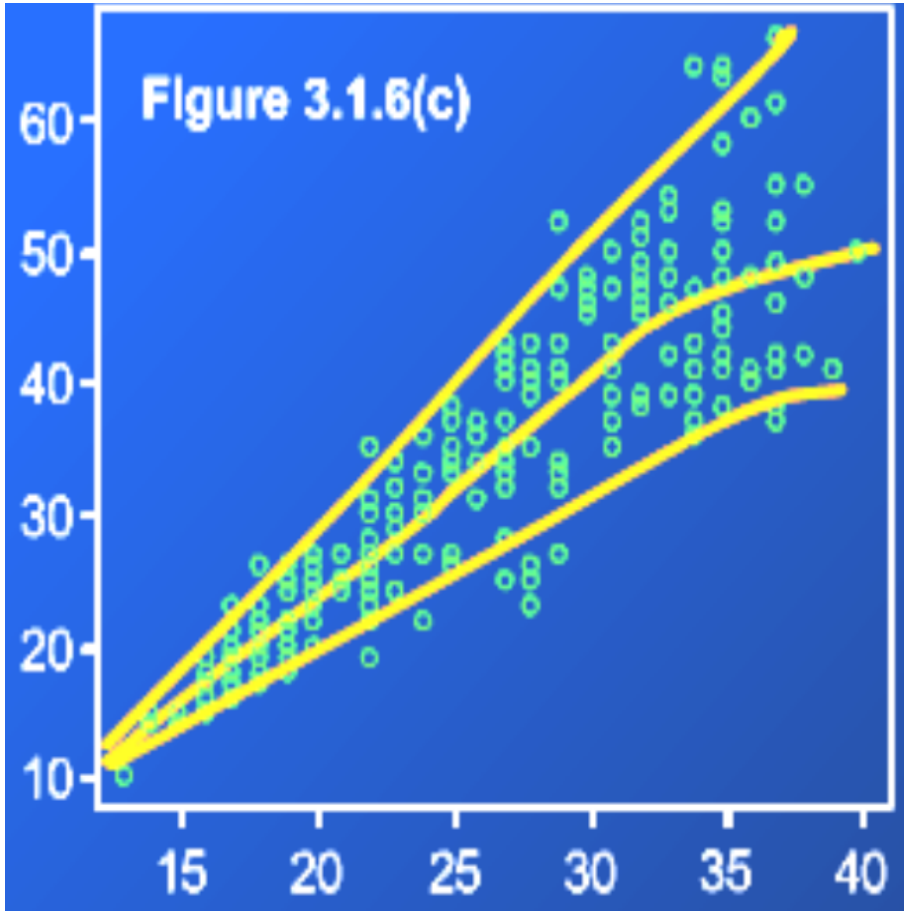
Trend non lineare con poca dispersione dei dati intorno alla curva, relazione forte

Cosa dice uno scatter plot?



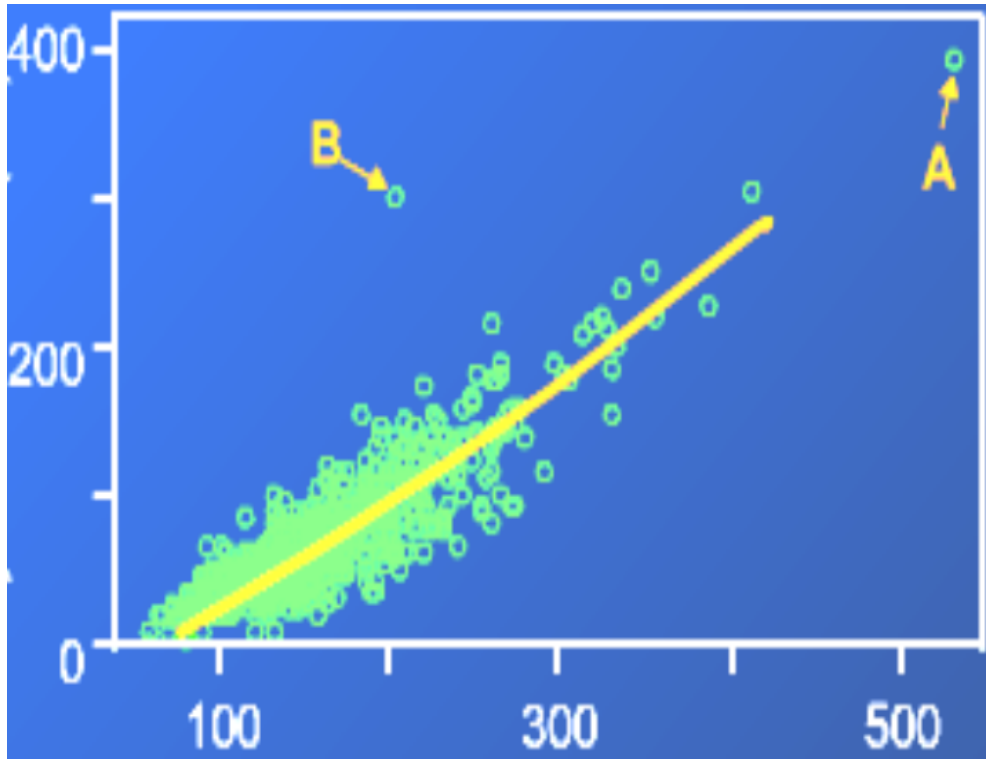
Trend lineare con una dispersione moderata e costante lungo la linea di tendenza

Cosa dice uno scatter plot?



Trend non lineare con dispersione non costante intorno alla curva, relazione debole

Cosa dice uno scatter plot?



A e B sono outlier, dopo un controllo B si è rivelato un errore, mentre A è sembrato un valore possibile

Per riassumere

Se $r = 0$ si dice che c'è una correlazione lineare non significativa, però è possibile che esista una relazione non lineare tra le variabili.

Per riassumere

Se $r = 0$ si dice che c'è una correlazione lineare non significativa, però è possibile che esista una relazione non lineare tra le variabili.

Non si può concludere che, poiché 2 variabili sono correlate in modo significativo, una sia necessariamente la causa dell'altra: un fattore nascosto può essere la causa della relazione tra le 2 variabili.

Per riassumere

Se $r = 0$ si dice che c'è una correlazione lineare non significativa, però è possibile che esista una relazione non lineare tra le variabili.

Non si può concludere che, poiché 2 variabili sono correlate in modo significativo, una sia necessariamente la causa dell'altra: un fattore nascosto può essere la causa della relazione tra le 2 variabili.
Karl Pearson (1857 London, 1936 Surrey)