Dipendenza tra variabili quantitative
Analisi della correlazione

SAPIENZA
UNIVERSITÀ DI ROMA

Prof. Annarita Vestri
Annarita.vestri@uniroma1.it

Correlazione

- studio la relazione tra indice di affollamento delle città e la mortalità infantile.
- Studio la relazione peso e altezza di individui di una popolazione;
- Studio le variabili di un censimento di una popolazione.

Il tasso di mortalità infantile (morti nel primo anno di vita) su 10.000 nati vivi «decresce quando il numero di abitanti/stanza (indice di affollamento) decresce» In altri termini «la mortalità infantile cresce con l'indice di affollamento».

Come concludo ? Esiste …

a) relazione di causa effetto (criteri di Bradford Hill).
b) associazione (generica).
c) correlazione (mutua influenza).

Correlazione

La regressione lineare è finalizzata all'analisi della dipendenza tra due variabili, delle quali
- una (**Y**) è a priori definita come **dipendente o effetto**,
- l'altra (**X**) è individuata come **indipendente o causa**.

L'interesse della ricerca è rivolta essenzialmente all'**analisi delle cause** o allo **studio predittivo delle quantità medie di Y**, che si ottengono come risposta al variare di X.
Spesso anche nella ricerca ambientale, biologica e medica, la relazione di causa-effetto non ha una direzione logica o precisa: potrebbe essere ugualmente applicata nei due sensi, da una variabile all'altra.

Le coppie di fidanzati o sposi di solito hanno altezza simile: la relazione di causa effetto può essere applicata sia dall'uomo alla donna che viceversa; coppie di gemelli hanno strutture fisiche simili e quella di uno può essere stimata sulla base dell'altro.

Altre volte, la causa può essere individuata in un terzo fattore, che agisce simultaneamente sui primi due, in modo diretto oppure indiretto, determinando i valori di entrambi e le loro variazioni, come la quantità di polveri sospese nell'aria e la concentrazione di benzene, entrambi dipendenti dall'intensità del traffico.
In altre ancora, l'interesse può essere limitato a **misurare come due serie di dati variano congiuntamente**, per poi andare alla ricerca delle eventuali cause, se la risposta fosse statisticamente significativa.
In tutti questi casi, è corretto utilizzare la **correlazione**

Differenze logiche nell'uso della regressione e della correlazione.

Esempio evidenziato da un ricercatore dei paesi nordici.

In un'ampia area rurale, per ogni comune durante il periodo invernale è stato contato il numero di cicogne e quello dei bambini nati. E' dimostrato che all'aumentare del primo cresce anche il secondo.

Ricorrere all'analisi della regressione su queste due variabili, indicando per ogni comune con X il numero di cicogne e con Y il numero di nati, implica una relazione di causa-effetto tra presenza di cicogne (X) e nascite di bambini (Y). Anche involontariamente si afferma che i bambini sono portati dalle cicogne; addirittura, stimando b, si arriva ad indicare quanti bambini sono portati mediamente da ogni cicogna.

In realtà durante i mesi invernali, nelle case in cui è presente un neonato, la temperatura viene mantenuta più alta della norma, passando indicativamente dai 16 ai 20 gradi centigradi. Soprattutto nei periodi più rigidi, le cicogne sono attratte dal maggior calore emesso dai camini e nidificano più facilmente su di essi o vi si soffermano più a lungo.

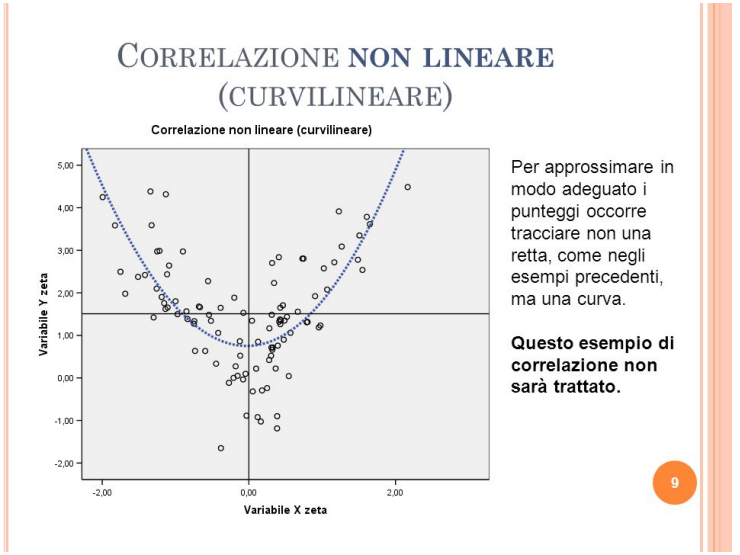
Con la correlazione si afferma solamente che le due variabili cambiano in modo congiunto.

L'analisi della correlazione misura solo il grado di associazione spaziale o temporale dei due fenomeni; ma lascia liberi nella scelta della motivazione logica, nel rapporto logico tra i due fenomeni.

Il coefficiente r è una misura dell'intensità dell'associazione tra le due variabili.

CORRELAZIONE

- Consente di misurare come due serie di dati (due variabili quantitative) variano congiuntamente.
- Prima cosa da fare è l'esplorazione grafica, per valutare visivamente se si può evidenziare l'eventuale correlazione tra le due variabili.



Nella correlazione, le due variabili possono essere indicate con X1 e X2, non più con X (causa) e Y (effetto), per rendere evidente l'assenza del concetto di dipendenza funzionale.

L'indice statistico (+r oppure –r) misura:
- il tipo (con il segno + o -)
- e il grado (con il valore assoluto) di interdipendenza tra due variabili.

Il segno indica il tipo di associazione:
- positivo, quando le due variabili aumentano o diminuiscono insieme,
- negativo, quando all'aumento dell'una corrisponde una diminuzione dell'altra o viceversa

Il valore assoluto varia da 0 a 1:
- - è massimo (uguale a 1) quando c'è una perfetta corrispondenza lineare tra X1 e X2;
- - tende a ridursi al diminuire della corrispondenza ed è zero quando essa è nulla.

L’indicatore della correlazione r è fondato sulla Codevianza e la Covarianza delle due variabili.

La Codevianza e la Covarianza tra X1 e X2 (CodX1/X2 e CovX1/X2) hanno la proprietà vantaggiosa di contenere queste due informazioni sul tipo (segno) ed sul grado (valore) di associazione; ma presentano anche lo svantaggio della regressione, poiché il loro valore risente in modo determinante della scala con la quale le due variabili X1 e X2 sono misurate.

Varianza

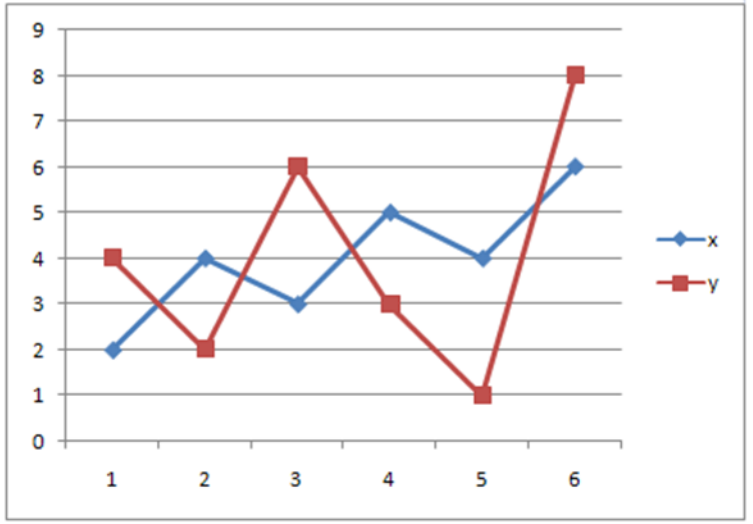
La varianza, detta anche media degli scarti al quadrato, è un indice di dispersione che è nullo solo nei casi in cui tutti i valori sono uguali alla loro media e cresce con il crescere delle differenze reciproche dei valori.

Varianza di *x* e di *y*:

$$s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}$$

	x	y			
1	2	4		media x=	4
2	4	2		media Y=	4
3	3	6			
4	5	3		varianza x=	1.67
5	4	1		varianza y=	5.67
6	6	8			



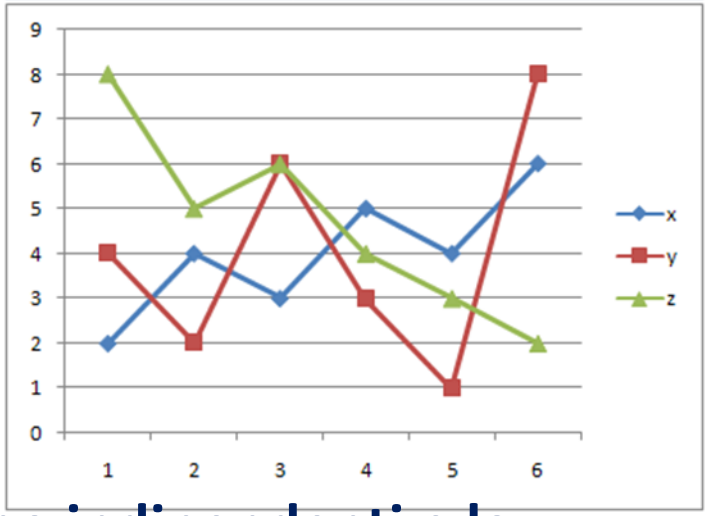
Covarianza

La covarianza analizza congiuntamente due caratteri e fornisce una misura della loro contemporanea variazione. **E' un indice simmetrico che misura la concordanza (o la discordanza) tra due caratteri quantitativi.** E' data dalla media del prodotto degli scarti tra le modalità dei due caratteri e le rispettive medie aritmetiche. Essa può assumere sia valori positivi che negativi. Nel primo caso indica che al crescere di una caratteristica statisticamente cresce anche l'altra, nel secondo caso accade il contrario.

Covarianza di *x* e *y*:

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}$$

	x	y	z		
1	2	4	8		
2	4	2	5		
3	3	6	6	covarianza (x,y)=	0.83
4	5	3	4	covarianza (x,z)=	-2.33
5	4	1	3	covarianza (y,z)=	-0.50
6	6	8	2		



Se non c'è nessuna relazione tra le due variabili si dice che x e y sono indipendenti e la loro covarianza è nulla

Esempio: una persona più alta della media, non è detto che abbia un QI elevato.

Viceversa, una persona più alta della media peserà probabilmente più della media.

La covarianza, in questo caso, è un *numero positivo* perché nella maggior parte dei casi $(x_i - \bar{x})$ e $(y_i - \bar{y})$ saranno ambedue positivi (per persone alte e pesanti) o ambedue negativi (per persone piccole e leggere).

Quantificando il peso in chilogrammi oppure in grammi e l'altezza in metri oppure in centimetri, si ottengono valori assoluti di Codevianza con dimensioni diverse, appunto perché fondati sugli scarti dalle medie

$$\sum_{i=1}^{n}(Xi-\overline{X})(Yi-\overline{Y})$$

E’ possibile pervenire a valori direttamente comparabili, qualunque sia la dimensione dei due fenomeni, cioè ottenere valori adimensionali, solo ricorrendo ad unità standard, quale appunto la variazione tra – 1 e +1 .
Si perviene ad essa, mediante il rapporto tra la codevianza e la media geometrica delle devianze di X1 e X2:

$$r = \frac{Codev\,(X,Y)}{\sqrt{Dev\,(X)\cdot Dev\,(Y)}}$$

$$= \frac{Covar\,(X,Y)}{\sqrt{V(X)\cdot V(Y)}}$$

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

r = +1	correlazione massima concorde
r= 0	correlazione assente
r = -1	correlazione massima discorde
r > 0	correlazione presente : all'aumentare di x aumenta y
r < 0	correlazione presente : all'aumentare di x diminuisce y

nota bene :

quando {y} è costante => r = indefinito

quando {x} è costante => r = indefinito

Coefficiente di correlazione di Pearson: r

CORRELAZIONE PARAMETRICA

Assunzioni:

- entrambe le variabili devono essere continue
- i dati devono essere secondo una scala a intervalli o rapporti
- entrambe le variabili devono seguire una distribuzione normale
- la relazione tra le variabili è lineare

E' importante ricordare che un valore assoluto basso o nullo di correlazione non deve essere interpretato come assenza di una qualsiasi forma di relazione tra le due variabili:

- è assente solo una relazione di tipo lineare,
- ma tra esse possono esistere relazioni di tipo non lineare, espresse da curve di ordine superiore, tra le quali la più semplice e frequente è quella di secondo grado.

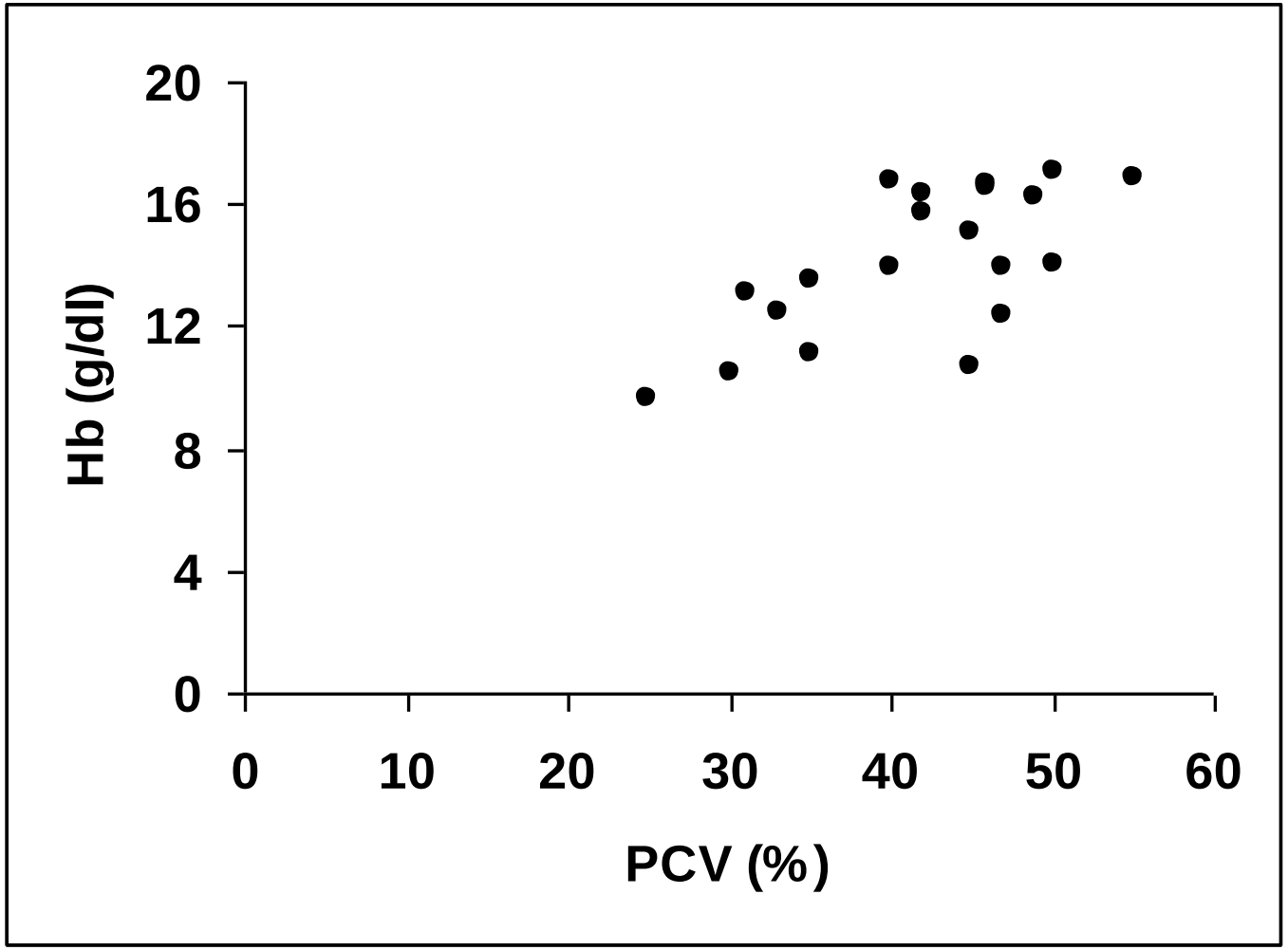
L'informazione contenuta in r riguarda solamente la quota espressa da una relazione lineare.

Esempio: campione di 20 donne

Gruppo di donne di una determinata area geografica invitate a sottoporsi a un prelievo di sangue per la determinazione del livello di emoglobina (Hb) e dell'ematocrito (PCV). Percentuale di adesione: circa il 90%.

Esempio: campione di 20 donne

ID	Hb (g/dl)	PCV (%)	Età (anni)
1	11,1	35	20
2	10,7	45	22
3	12,4	47	25
4	14,0	50	28
5	13,1	31	28
6	10,5	30	31
7	9,6	25	32
8	12,5	33	35
9	13,5	35	38
10	13,9	40	40
11	15,1	45	45
12	13,9	47	49
13	16,2	49	54
14	16,3	42	55
15	16,8	40	57
16	17,1	50	60
17	16,6	46	62
18	16,9	55	63
19	15,7	42	65
20	16,5	46	67



Si vuole analizzare la relazione fra Hb e PCV. Non ci chiediamo se Hb influenzi PCV o PCV influenzi Hb, o se un alto valore di PCV causi un alto valore di Hb, ma se le due variabili sono associate

Il coefficiente di correlazione del campione

r= coefficiente di correlazione di Pearson

ci permette di:

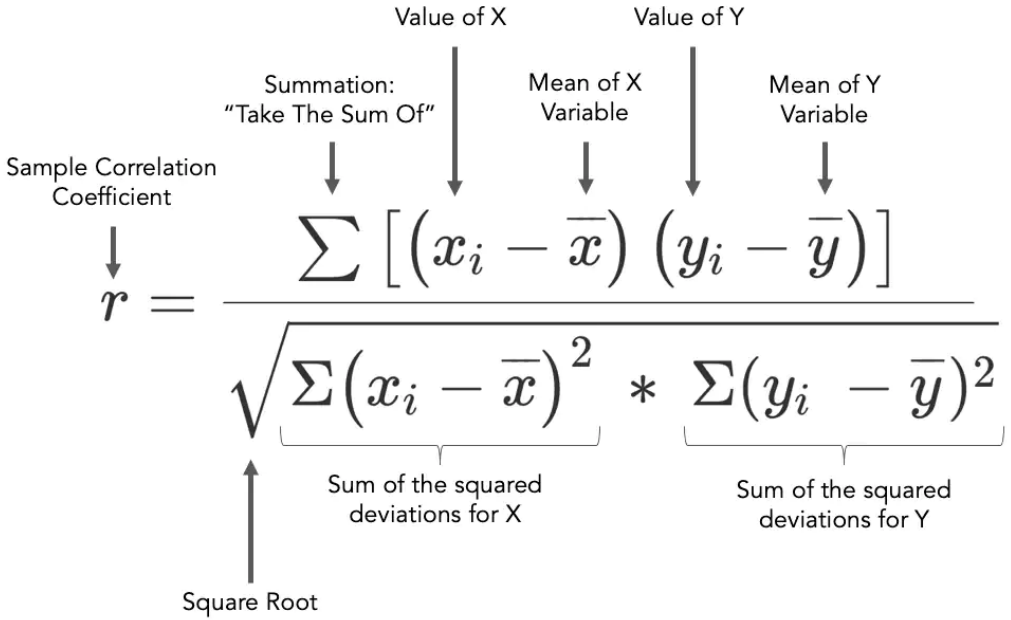
- verificare l'ipotesi che vi sia associazione fra le variabili o se l'apparente associazione possa essere dovuta al caso
- riassume la forza della relazione lineare fra le variabili

Calcolo di *r*

Dato un insieme di coppie di osservazioni

$(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$



Esempio: Hb = x, PCV = y $\bar{x} = 14.12$ $\bar{y} = 41.65$

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
11.1	35	-3.02	-6.65	20.083	9.12	44.2225
10.7	45	-3.42	3.35	-11.457	11.70	11.2225
12.4	47	-1.72	5.35	-9.202	2.96	28.6225
14.0	50	-0.12	8.35	-1.002	0.01	69.7225
13.1	31	-1.02	-10.65	10.863	1.04	113.4225
10.5	30	-3.62	-11.65	42.173	13.10	135.7225
9.6	25	-4.52	-16.65	75.258	20.43	277.2225
12.5	33	-1.62	-8.65	14.013	2.62	74.8225
13.5	35	-0.62	-6.65	4.123	0.38	44.2225
13.9	40	-0.22	-1.65	0.363	0.05	2.7225
15.1	45	0.98	3.35	3.283	0.96	11.2225
13.9	47	-0.22	5.35	-1.177	0.05	28.6225
16.2	49	2.08	7.35	15.288	4.33	54.0225
16.3	42	2.18	0.35	0.763	4.75	0.1225
16.8	40	2.68	-1.65	-4.422	7.18	2.7225
17.1	50	2.98	8.35	24.883	8.88	69.7225
16.6	46	2.48	4.35	10.788	6.15	18.9225
16.9	55	2.78	13.35	37.113	7.73	178.2225
15.7	42	1.58	0.35	0.553	2.50	0.1225
16.5	46	2.38	4.35	10.353	5.66	18.9225

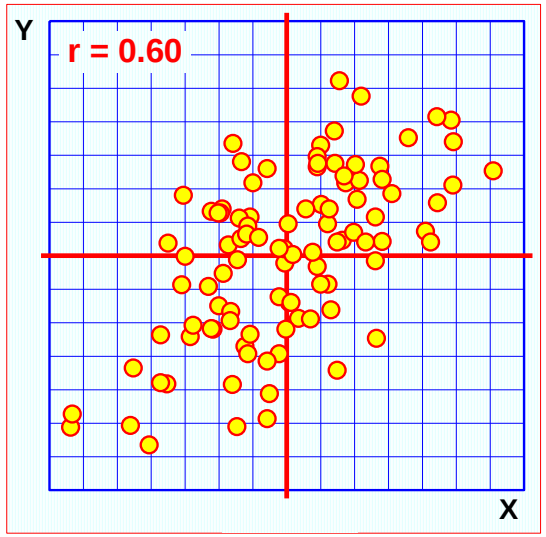
$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = 242.64$$

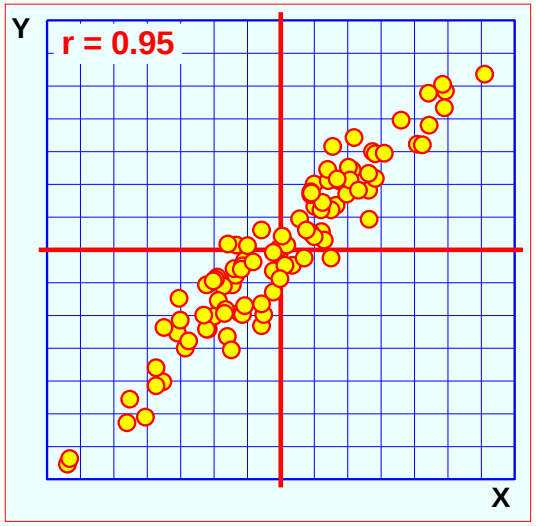
$$\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2} = 360.33$$

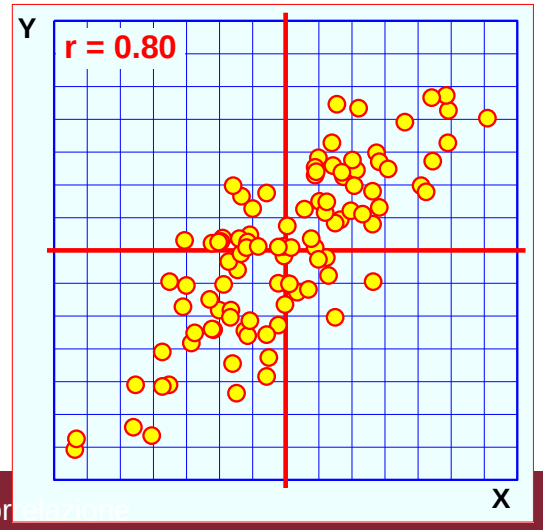
$$r = \frac{242.64}{360.33} = 0.67$$

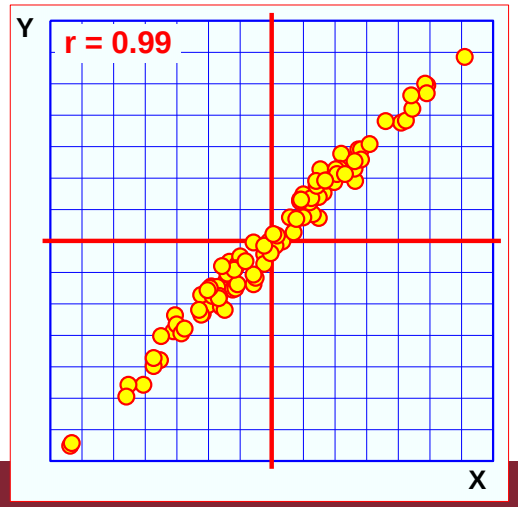
La FORZA e il TIPO dell'ASSOCIAZIONE

Grafici di dispersione per variabili a correlazione elevata o molto elevata



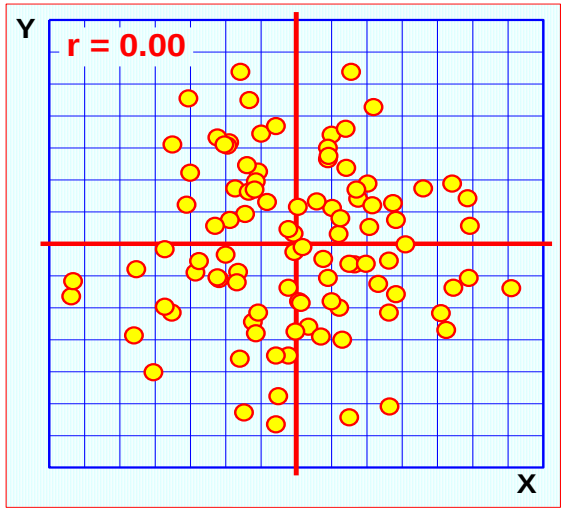


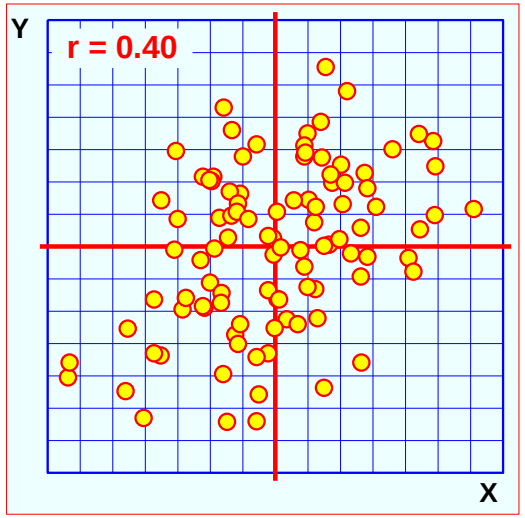


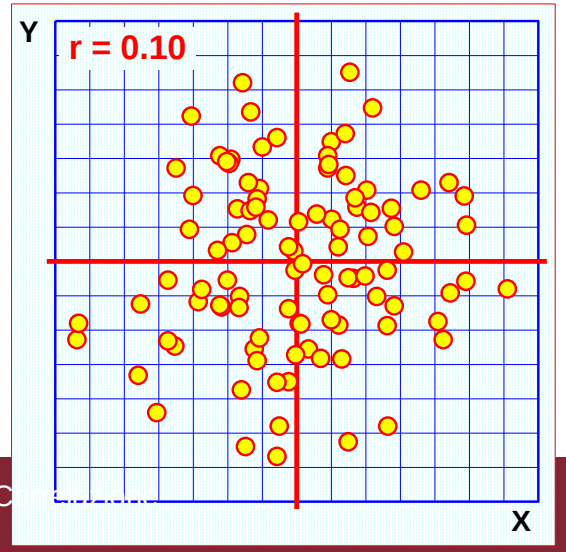


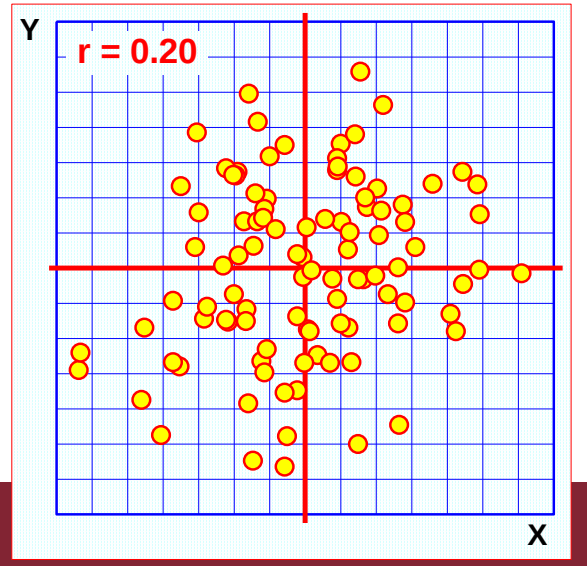
La FORZA e il TIPO dell'ASSOCIAZIONE

Grafici di dispersione per variabili a correlazione nulla o lieve.



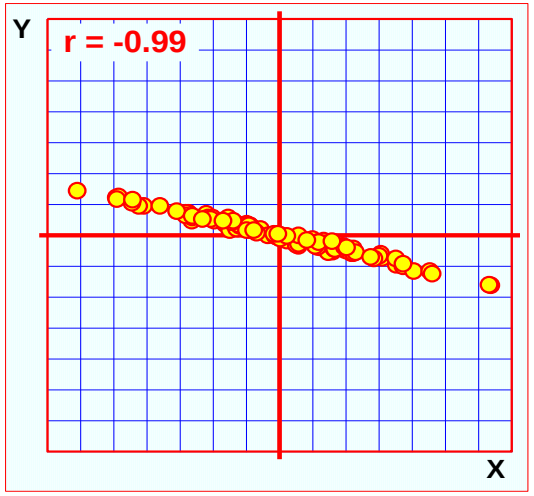


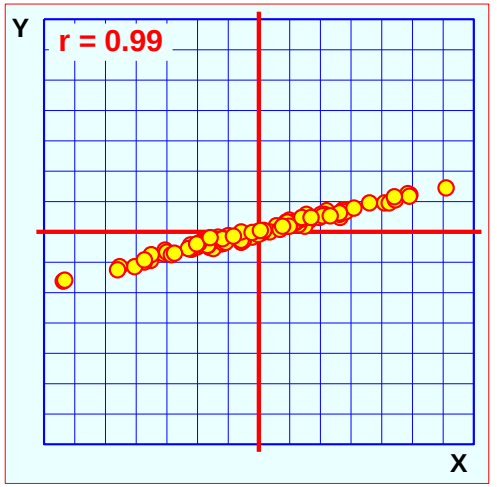


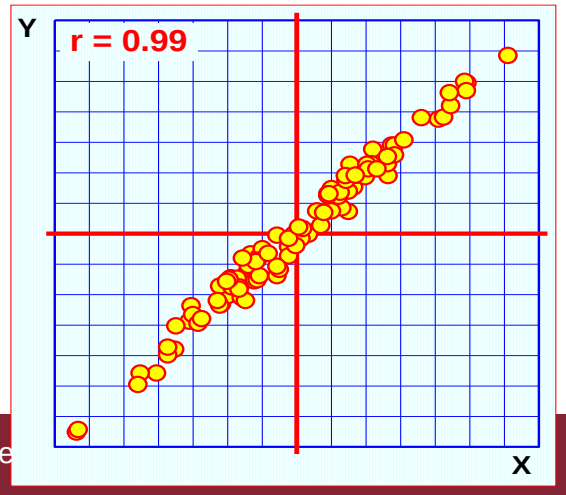


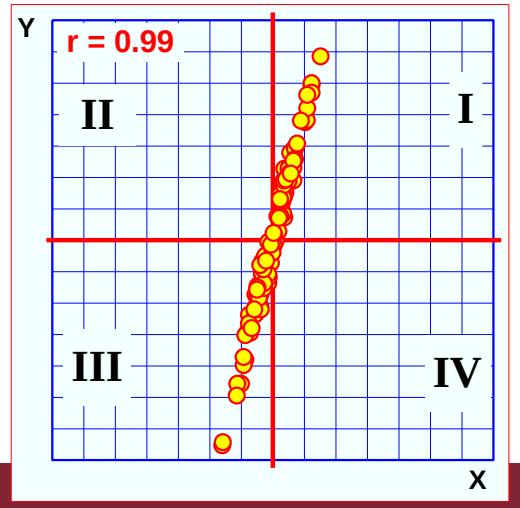
La FORZA e il TIPO dell'ASSOCIAZIONE

Il coefficiente di correlazione è **positivo se la retta giace nei quadranti I e III**, negativo in caso contrario. Se i punti si allineano **perfettamente su una retta parallela** ad uno dei due assi, il coefficiente di correlazione è **indeterminato**.



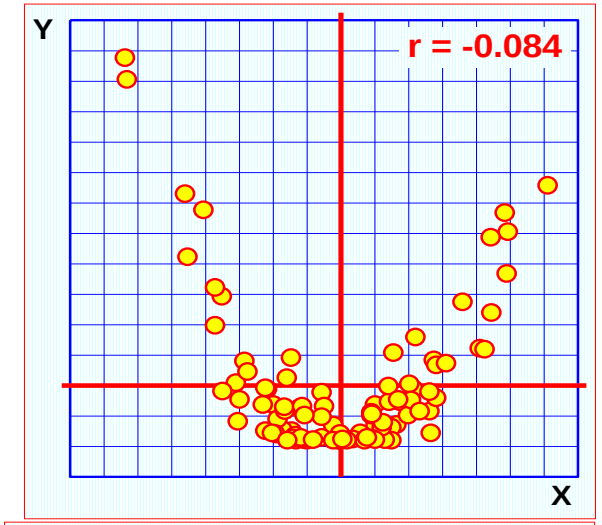


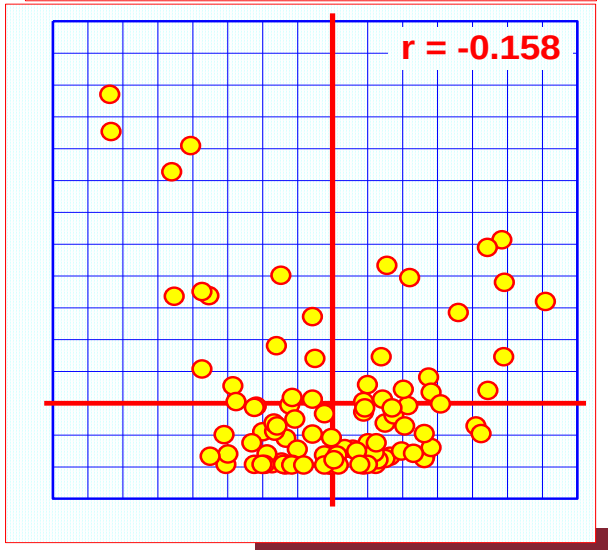


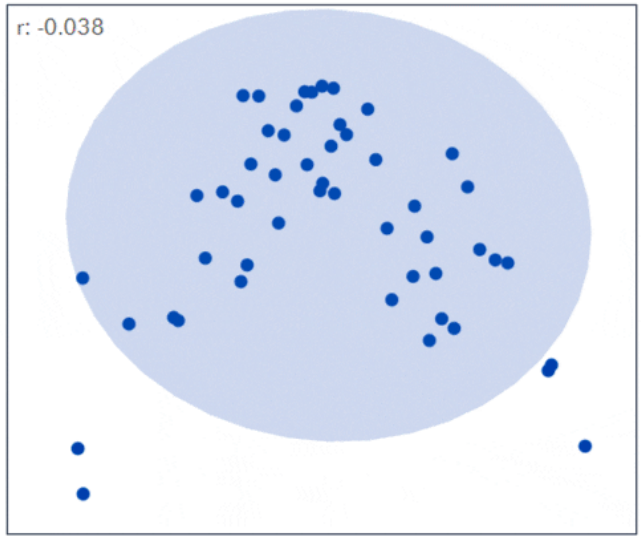


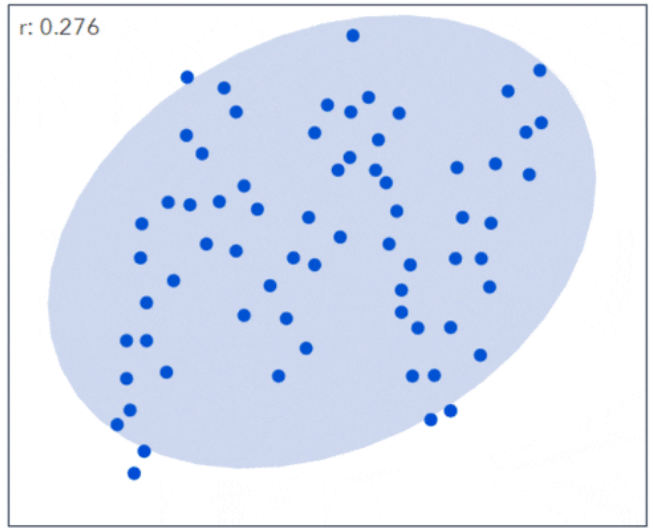
La FORZA e il TIPO dell'ASSOCIAZIONE

Il coefficiente di correlazione lineare **è indice di quanto i punti si allineano su di una retta**: vi possono essere associazioni anche forti, ma di tipo non lineare per le quali il coefficiente di correlazione è prossimo a 0.









IMPORTANTE:

Una elevata correlazione fra due variabili NON implica una relazione causa-effetto

- r quantifica solo una relazione lineare tra variabili
- r è sensibile a valori estremi
- la correlazione osservata non deve essere mai estrapolata oltre i range osservati delle variabili

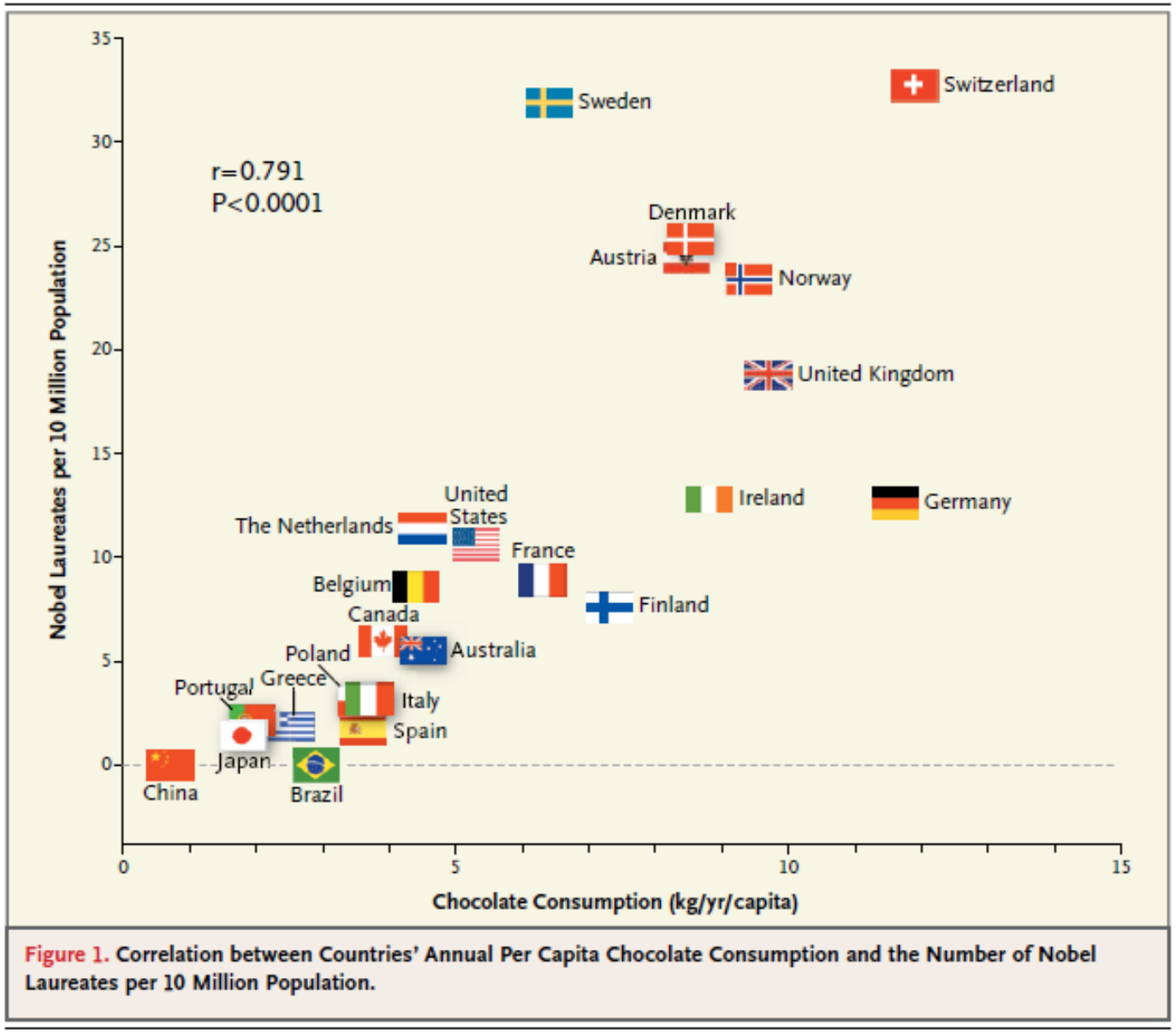


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.

The principal finding of this study is a surprisingly powerful correlation between chocolate intake per capita and the number of Nobel laureates in various countries. Of course, a correlation between X and Y does not prove causation but indicates that either X influences Y, Y influences X, or X and Y are influenced by a common underlying mechanism. However, since chocolate consumption has been documented to improve cognitive function, it seems most likely that in a dose-dependent way, chocolate intake provides the abundant fertile ground needed for the sprouting of Nobel laureates. Obviously, these findings are hypothesis-generating only and will have to be tested in a prospective, randomized trial.

N ENGL J MED 367;16 NEJM.ORG OCTOBER 18, 2012

esempio

Relazione tra percentuale di bambini che sono stati vaccinati per difterite, pertosse e tetano (DPT) in una Nazione e il corrispondente tasso di mortalità dei bambini al di sotto dei 5 anni di età

Tabella 17.1 Percentuale di bambini vaccinati contro DPT e tasso di mortalità al di sotto di 5 anni per 20 Paesi, 1992

Paese	Percentuale vaccinati	Tasso di mortalità per 1.000 nati vivi
Bolivia	77	118
Brasile	69	65
Cambogia	32	184
Canada	85	8
Cina	94	43
Egitto	89	55
Etiopia	13	208
Fed. Russa	73	32
Finlandia	95	7
Francia	95	9
Giappone	87	6
Grecia	54	9
India	89	124
Italia	95	10
Messico	91	33
Polonia	98	16
Regno Unito	90	9
Rep. Ceca	99	12
Senegal	47	145
Turchia	76	87

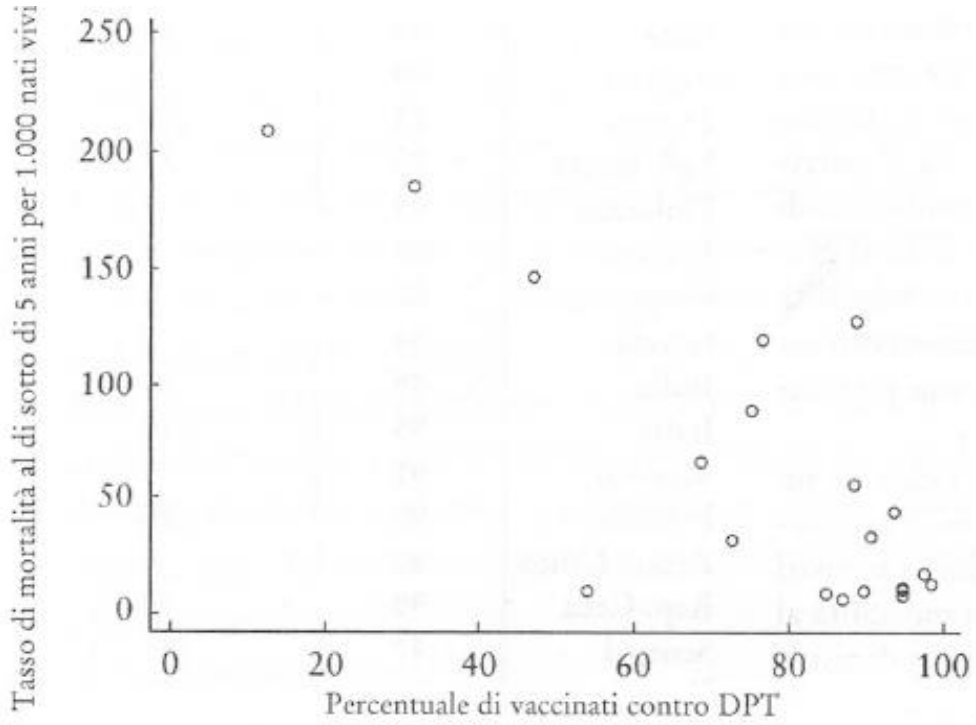


Figura 17.1 Tasso di mortalità al di sotto di 5 anni in funzione della percentuale di bambini vaccinati contro DPT per 20 Paesi, 1992

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= \frac{1}{20}\sum_{i=1}^{20} x_i$$

$$= 77{,}4\%,$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$= \frac{1}{20}\sum_{i=1}^{20} y_i$$

$$= 59{,}0 \text{ per } 1.000 \text{ nati vivi.}$$

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{20}(x_i-77{,}4)(y_i-59{,}0)$$

$$= -22.706,$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{20}(x_i-77{,}4)^2$$

$$= 10.630{,}8$$

Pertanto, il coefficiente di correlazione è:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]}}$$

$$= \frac{-22.706}{\sqrt{(10.630{,}8)(77.498)}}$$

$$= -0{,}79.$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{20}(y_i-59{,}0)^2$$

$$= 77.498.$$

Sembra esserci una forte relazione lineare tra la percentuale di bambini vaccinati contro DPT e il corrispondente tasso di mortalità al di sotto dei 5 anni, r è negativo quindi il tasso di mortalità diminuisce all'aumentare della percentuale di vaccinazioni.

Un efficace programma di vaccinazione potrebbe essere la causa principale della riduzione della mortalità o uno degli aspetti di un efficace sistema di assistenza sanitaria che a sua volta causa la riduzione

Observational Study

Medicine

OPEN

Relationships between serum omentin-1 concentration, body composition and physical activity levels in older women

Shuo Li, PhD[a], Jingjing Xue, PhD[b], Ping Hong, PhD[a,c,*]

Abstract

This study aimed to investigate the relationships between omentin-1, body composition and physical activity (PA) levels in older women.

Eighty-one older women (age = 64 ± 6 years; body mass index = 24.2 ± 3.2 kg/m^2; body fat percentage = 36.1 ± 5.7%) participated in this study. We divided the subjects into overweight/obesity and normal weight group. Body composition was measured by dual energy X-ray absorptiometry. Serum omentin-1 concentration was measured using enzyme linked immunosorbent assay. PA levels were obtained by using accelerometers. In addition, anthropometric and insulin resistance values were determined.

Omentin-1 level in overweight/obesity group was significantly lower than in the normal weight group ($P < .01$). Analysis of all subjects showed that serum omentin-1 was negatively correlated with body weight, BMI (body mass index), waist circumference (WC), WHR (waist-to-hip ratio), percentage of body fat, total body fat mass (FM), fat-free mass (FFM) ($r = -.571$, -0.569, -0.546, -0.382, -0.394, -0.484, -0.524, all $P < .01$), respectively. We also found a negative correlation between moderate-to-vigorous physical activity (MVPA) and total body FM ($r = -.233$, $P < .05$). However, no significant correlation was found between omentin-1 and sedentary behavior and MVPA (both $P > .05$). Moreover, the relationship between omentin-1, body composition and PA was analyzed by using multiple linear stepwise regressions. The results showed that serum omentin-1 concentration was inversely correlated with total body FM ($\beta = -0.334$, $P = .004$) in multiple linear stepwise regression analysis.

We found that total body FM was inversely related to serum omentin-1 concentration and PA levels, but there was no correlation between omentin-1 and PA levels. These results showed that PA may participate in the regulation of body composition, which may be also affected by serum omentin-1. However, the mechanism by which PA affects body composition may not be through omentin-1 and was more likely through other metabolic pathways.

Abbreviations: BMI = body mass index, DBP = diastolic blood pressure, FFM = fat-free mass, FM = fat mass, HDL-C = high-density lipoprotein cholesterol, HOMA-IR = Homoeostasis model of insulin resistance, LDL-C = low-density lipoprotein cholesterol, LPA = light-intensity physical activity, MPA = mderate-intensity physical activity, MVPA = moderate-to-vigorous physical activity, PA = physical activity, SB = Sedentary behavior, SBP = systolic blood pressure, VPA = Vigorous-intensity physical activity, WHR = Waist-to-hip ratio.

Observational Study

Medicine
OPEN

Relationships between serum omentin-1 concentration, body composition and physical activity levels in older women

Shuo Li, PhD[a], Jingjing Xue, PhD[b], Ping Hong, PhD[a,c,*]

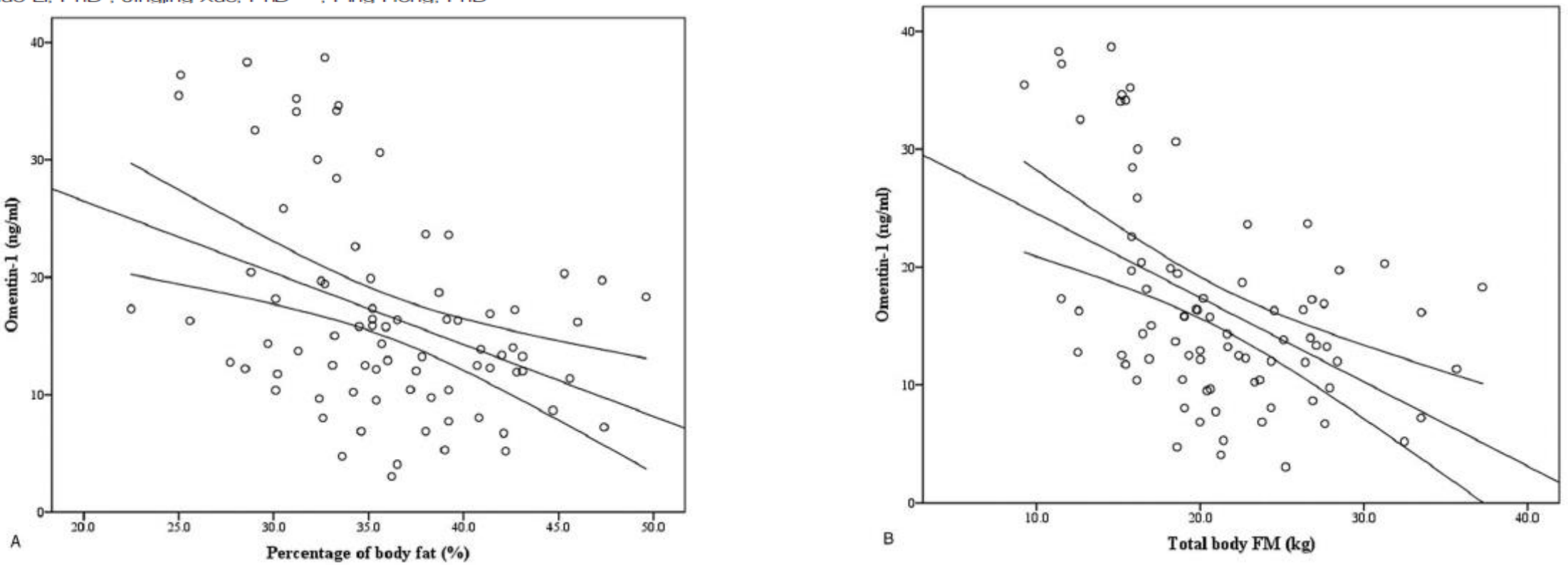


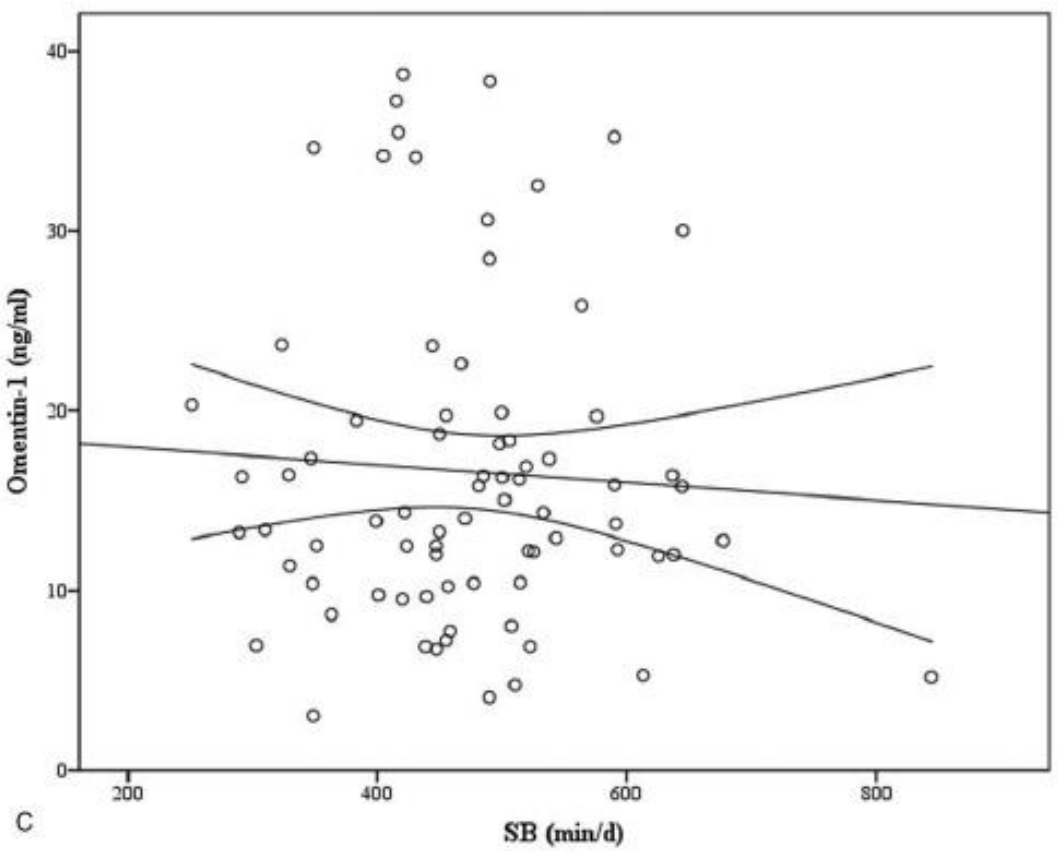
Figure 1. Correlation between circulating omentin-1 and a percent body fat ($r = -.394$, $P < .01$), B total body FM ($r = -.484$, $P < .01$) in all subjects. Correlation between circulating omentin-1 and C SB ($r = -.057$, $P > .05$), D MVPA ($r = .166$, $P > .05$) in all subjects. Correlation between E SB ($r = -.068$, $P > 0.05$), F MVPA ($r = -.233$, $P < .05$) and total body FM in all subjects.

Observational Study

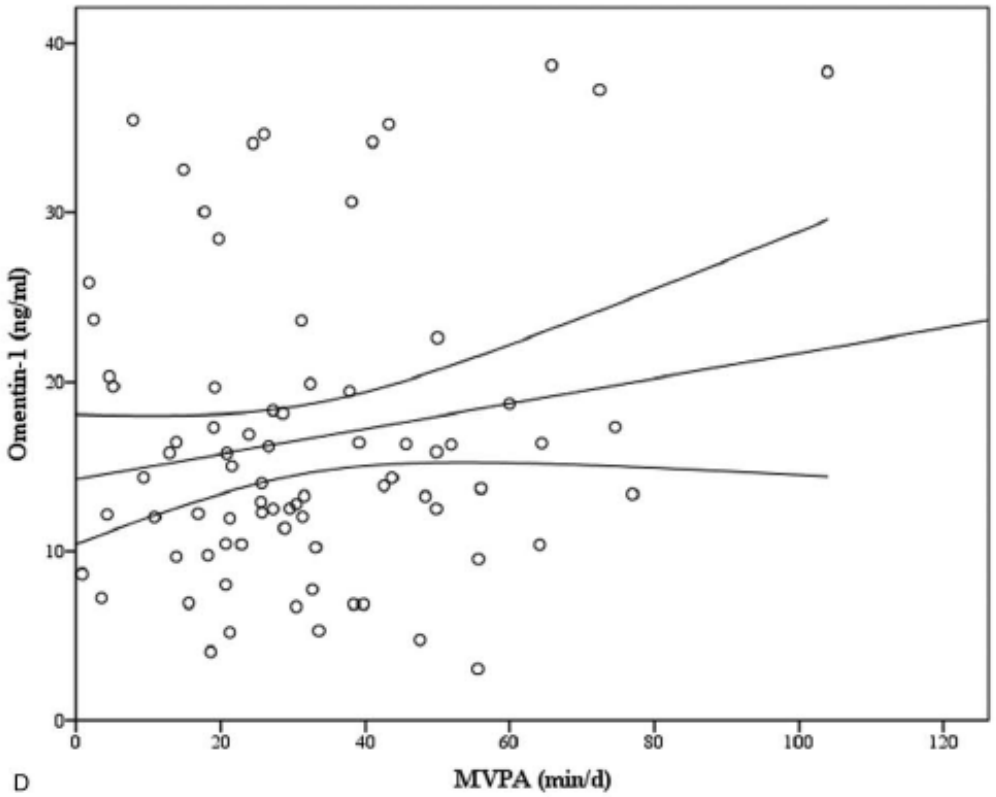
Medicine
OPEN

Relationships between serum omentin-1 concentration, body composition and physical activity levels in older women

Shuo Li, PhD[a], Jingjing Xue, PhD[b], Ping Hong, PhD[a,c,*]



C



D

Figure 1. Continued.