Analisi Dati con Excel

Giovanni Della Lunga

giovanni.dellalunga@gmail.com

La prima regola di ogni tecnologia è che l'automazione applicata ad un'operazione efficiente ne aumenterà l'efficienza. La seconda è che l'automazione applicata ad un'operazione inefficiente ne aumenterà l'inefficienza.

Bill Gates

Correlazione e Covarianza

Correlation IS NOT Causation!!!

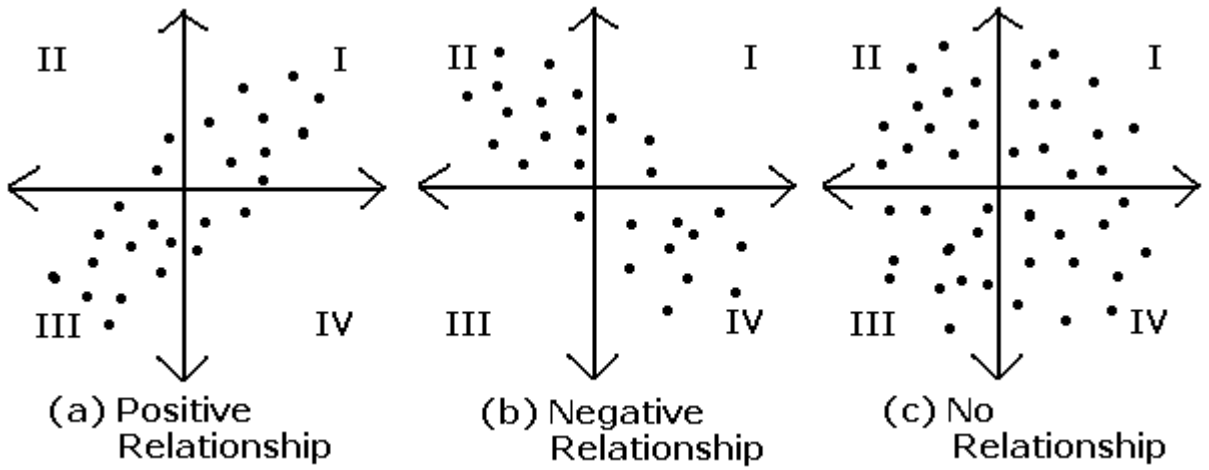
Covarianza

- Cerchiamo ora di capire cosa è la **covarianza**.
- Innanzitutto quando parliamo di covarianza non ci limitiamo più ad osservare una singola variabile, ma introduciamo un'altra variabile.
- La covarianza è un indice che mette in relazione due variabili; in particolare, a differenza di altri indici, **la covarianza definisce la relazione direzionale fra due variabili**.
- I valori ammessi della covarianza sono compresi fra -infinito e +infinito.

Covarianza

Per “relazione direzionale” intendiamo dire che:

- se la covarianza ha un **valore positivo**, le due variabili oggetto del nostro studio si muovono in maniera **concorde**;
- se la covarianza ha un **valore negativo**, le due variabili oggetto del nostro studio si muovono in maniera **discorde**;
- se la covarianza è pari a **zero**, le due variabili **non sono correlate fra loro**



Stima della Covarianza

Uno stimatore della covarianza è il seguente

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)$$

Calcolo della Covarianza in Excel

Passaggi per Calcolare la Covarianza in Excel

- Inserisci i Dati
- Supponiamo di avere due serie di dati nelle colonne A e B.
  - Colonna A: Valori della variabile X.
  - Colonna B: Valori della variabile Y.
- Utilizza la Funzione COVARIANZA.P o COVARIANZA.S
  - COVARIANZA.P: Calcola la covarianza per l'intera popolazione.
  - COVARIANZA.S: Calcola la covarianza per un campione.

Matrice di Covarianza

- Nel caso in cui abbiamo più variabili e vogliamo calcolare la covarianza fra di esse si parla di matrice di covarianza.
- Si tratta di una matrice quadrata in cui le intestazioni delle righe sono le nostre variabili e le colonne anche.
- Se ad esempio volessimo calcolare la matrice di covarianza delle variabili appena viste, avremmo questo:

$$C(x,y) = \begin{Bmatrix} \sigma_x^2 & cov_{x,y} \\ cov_{y,x} & \sigma_y^2 \end{Bmatrix}$$

Matrice di Covarianza

- È una matrice simmetrica (perchè *cov(x, y) = cov(y, x)*) e notiamo che sulla **diagonale principale** troviamo le **varianze** delle singole variabili X e Y.
- Gli altri elementi sono le covarianze delle variabili corrispondenti.

$$C(x,y) = \begin{Bmatrix} \sigma_x^2 & cov_{x,y} \\ cov_{y,x} & \sigma_y^2 \end{Bmatrix}$$

Correlazione

Coefficiente di Correlazione

- È un indice che presenta delle analogie e delle differenze con la già vista covarianza, e serve ad uno scopo ben preciso: fornire informazioni sulla presenza (e in caso affermativo dell'andamento) di una relazione fra due variabili casuali.

- La cosa insolita è che sotto il termine "correlazione" si possono nascondere formule e indici anche molto diversi fra loro a seconda della natura delle variabili sulle quali la correlazione si vuole calcolare.

Coefficiente di Correlazione Lineare

Per le variabili continue come quelle utilizzate nella maggior parte dei casi, il coefficiente di correlazione che si utilizza è di solito il coefficiente di Pearson, la cui formula è:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Coefficiente di Correlazione Lineare

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

- Come vediamo, la correlazione calcolata con il coefficiente di Pearson non è altro che una frazione in cui al numeratore c'è la covarianza fra le variabili X, Y considerate mentre al denominatore c'è il prodotto fra le deviazioni standard delle due variabili;
- Il coefficiente di Pearson serve a individuare ***se esiste una relazione lineare fra le due variabili casuali***. Se si vuole misurare la relazione non lineare si devono usare altri indici (come il coefficiente di Spearman).

Coefficiente di Correlazione Lineare

Il Coefficiente di Correlazione Lineare ha un intervallo ben definito. Esso è infatti **sempre compreso fra -1 e 1**.

In particolare, in maniera simile alla covarianza:

1. se la correlazione è compresa fra -1 e 0 vuol dire che le due variabili sono inversamente correlate: vale a dire che all'aumentare di una l'altra decresce
2. se la correlazione è uguale a 0 non esiste alcuna correlazione lineare fra le variabili
3. se la correlazione è compresa fra 0 e 1 vuol dire che le due variabili sono direttamente correlate: all'aumentare di una aumenta anche l'altra

Calcolo della Correlazione in Excel

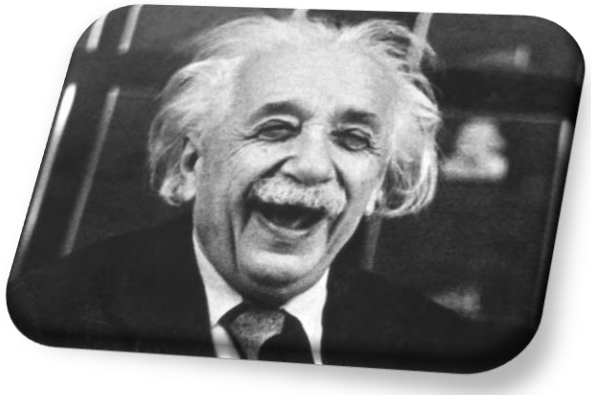
Passaggi per Calcolare la Correlazione in Excel

- Inserisci i Dati
- Supponiamo di avere due serie di dati nelle colonne A e B.
  - Colonna A: Valori della variabile X.
  - Colonna B: Valori della variabile Y.
- Utilizza la Funzione CORRELAZIONE
  - La funzione CORRELAZIONE calcola il coefficiente di correlazione di Pearson tra due serie di dati.

Correlation is NOT Causation!

- Occhio, **correlazione non significa causalità**.
- Ad esempio, quando la correlazione fra due variabili è vicina ad 1, **non significa che un cambio in una variabile generi automaticamente un cambio anche nella seconda**.
- Ad esempio, prendiamo due variabili come “Numero di gelati venduti giornalmente nell’arco di un anno” e “Numero di scottature giornaliere rilevate nell’arco di un anno”.
- Verosimilmente le due variabili avranno un’alta correlazione (all’aumentare di una aumenterà verosimilmente anche l’altra), ma un cambio in una delle due variabili sicuramente non si rifletterà nell’altra.
- In questo caso abbiamo Alta correlazione ma bassa causalità.

Correlazioni Spurie



spurious correlations

correlation is not causation

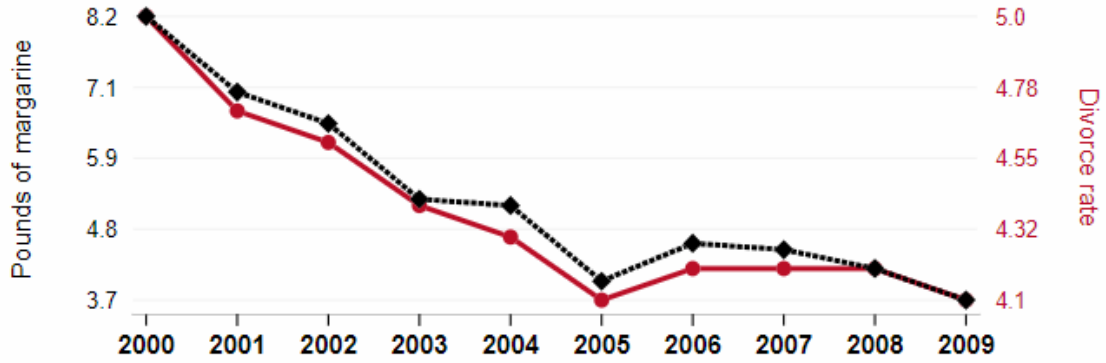
random · discover · *next page* →

don't miss spurious scholar,
where each of these is an academic paper

Per capita consumption of margarine

correlates with

The divorce rate in Maine



Per capita consumption of margarine in the United States · Source: US Department of Agriculture

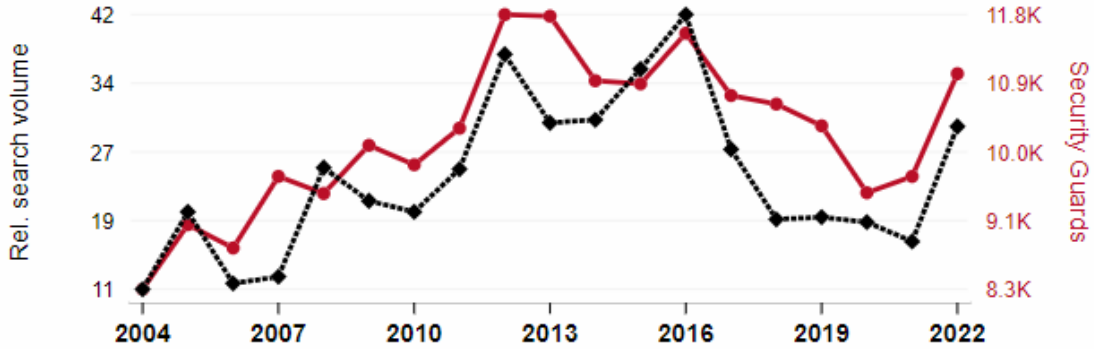
The divorce rate in Maine · Source: CDC National Vital Statistics

2000-2009, r=0.993, r²=0.985, p<0.01 · tylervigen.com/spurious/correlation/5920

Google searches for 'batman'

correlates with

The number of security guards in Oklahoma



Relative volume of Google searches for 'batman' (Worldwide, without quotes) · Source: Google Trends

BLS estimate of security guards in Oklahoma · Source: Bureau of Larbor Statistics

2004-2022, r=0.848, r²=0.719, p<0.01 · tylervigen.com/spurious/correlation/5227