Statistica I

Unità M: cenni alla correlazione parziale

Tommaso Rigon

Università Milano-Bicocca

Unità M

Argomenti affrontati

- La correlazione parziale
- Correlazione spuria
- Rapporti di causa ed effetto

Riferimenti al libro di testo

- §7.6

Una congettura

- Consideriamo nuovamente i dati analizzati nell'**unità J**, le province svizzere del 1888.
- Si supponga che un sociologo faccia le seguenti ipotesi sulle relazioni socio-economiche intercorrenti tra le variabili `agricoltura`, `istruzione` e `fertilità`.
- **Ipotesi 1**. Tra `agricoltura` ed `istruzione` esiste una sostanziale interdipendenza.
- Una possibile spiegazione è che la scuola venisse percepita come poco utile per il lavoro nei campi. Quindi, le province agricole rimanevano associate a bassi livelli di istruzione.
- Viceversa, un buon livello di `istruzione` potrebbe facilitare l'inserimento nelle attività secondarie e terziarie.
- È ragionevole quindi congetturare un **effetto diretto** tra `agricoltura` e `istruzione`. Scriveremo:

$$\texttt{agricoltura} \Longleftrightarrow \texttt{istruzione}.$$

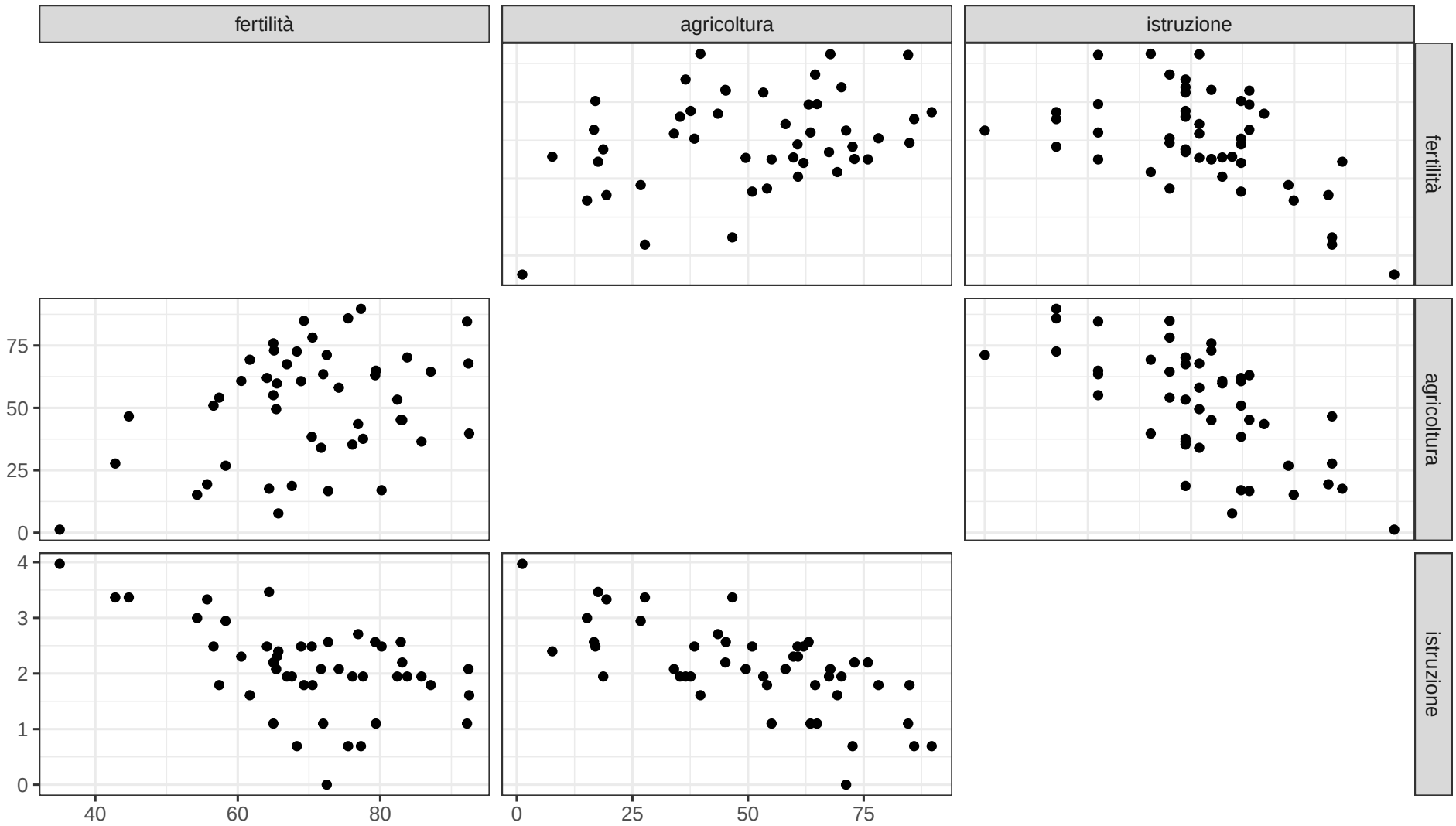
Una congettura

- **Ipotesi 2**. Anche tra `istruzione` e `fertilità` esiste una sostanziale interdipendenza.
- Una possibile spiegazione è che coppie con buona istruzione vogliono (e sono capaci di) controllare la natalità.
- Allo stesso tempo, famiglie con pochi figli hanno più disponibilità di reddito e quindi più inclini a istruirli.
- È ragionevole quindi congetturare un **effetto diretto** tra `fertilità` ed `istruzione`. Scriveremo:

$$\texttt{fertilità} \Longleftrightarrow \texttt{istruzione}.$$

- Cosa possiamo invece dire della relazione tra `agricoltura` e `fertilità`?

I diagrammi a dispersione



La correlazione spuria

- I grafici precedenti e i risultati dell'unità J evidenziano una **correlazione positiva** tra `fertilità` e `agricoltura`.
- **Ipotesi 3**. Per quanto riguarda la `fertilità`, congetturiamo che le province "molto agricole e colte" abbiano un comportamento simile a province "poco agricole ma colte".
- In altri termini, supponiamo non vi sia alcuna relazione diretta tra `agricoltura` e `fertilità`.
- La relazione osservata precedentemente è comunemente detta **correlazione spuria**.
- Infatti, l'associazione positiva tra queste due variabili è dovuto al fatto che l'agricoltura è legata all'`istruzione`, che a sua volta è legata alla `fertilità`.

Sommario della congettura

- Complessivamente, il sommario delle relazioni che il sociologo ipotizza tra le tre variabili è rappresentabile schematicamente come segue

$$\texttt{agricoltura} \Longleftrightarrow \texttt{istruzione} \Longleftrightarrow \texttt{fertilità}.$$

- Il punto cruciale della congettura è l'inesistenza di una freccia che metta in relazione diretta `agricoltura` e `fertilità`.

- È possibile trovare un qualche **indicatore statistico** a supporto di questa congettura?

- In altri termini, come possiamo valutare la "correlazione" tra `agricoltura` e `fertilità` al netto dell'effetto dell'`istruzione`?

Formalizzazione della congettura

- Una possibile modo di affrontare il problema è supporre la seguente decomposizione

  $$\texttt{agricoltura} = (\text{parte } \textbf{legata} \text{ ad } \texttt{istruzione}) + (\text{parte } \textbf{non legata} \text{ ad } \texttt{istruzione}),$$

  e parallelamente che

  $$\texttt{fertilità} = (\text{parte } \textbf{legata} \text{ ad } \texttt{istruzione}) + (\text{parte } \textbf{non legata} \text{ ad } \texttt{istruzione}).$$

- Se accettiamo questa formalizzazione, possiamo quindi tentare di identificare la "parte non legate ad `istruzione`" e studiarne le relazioni.

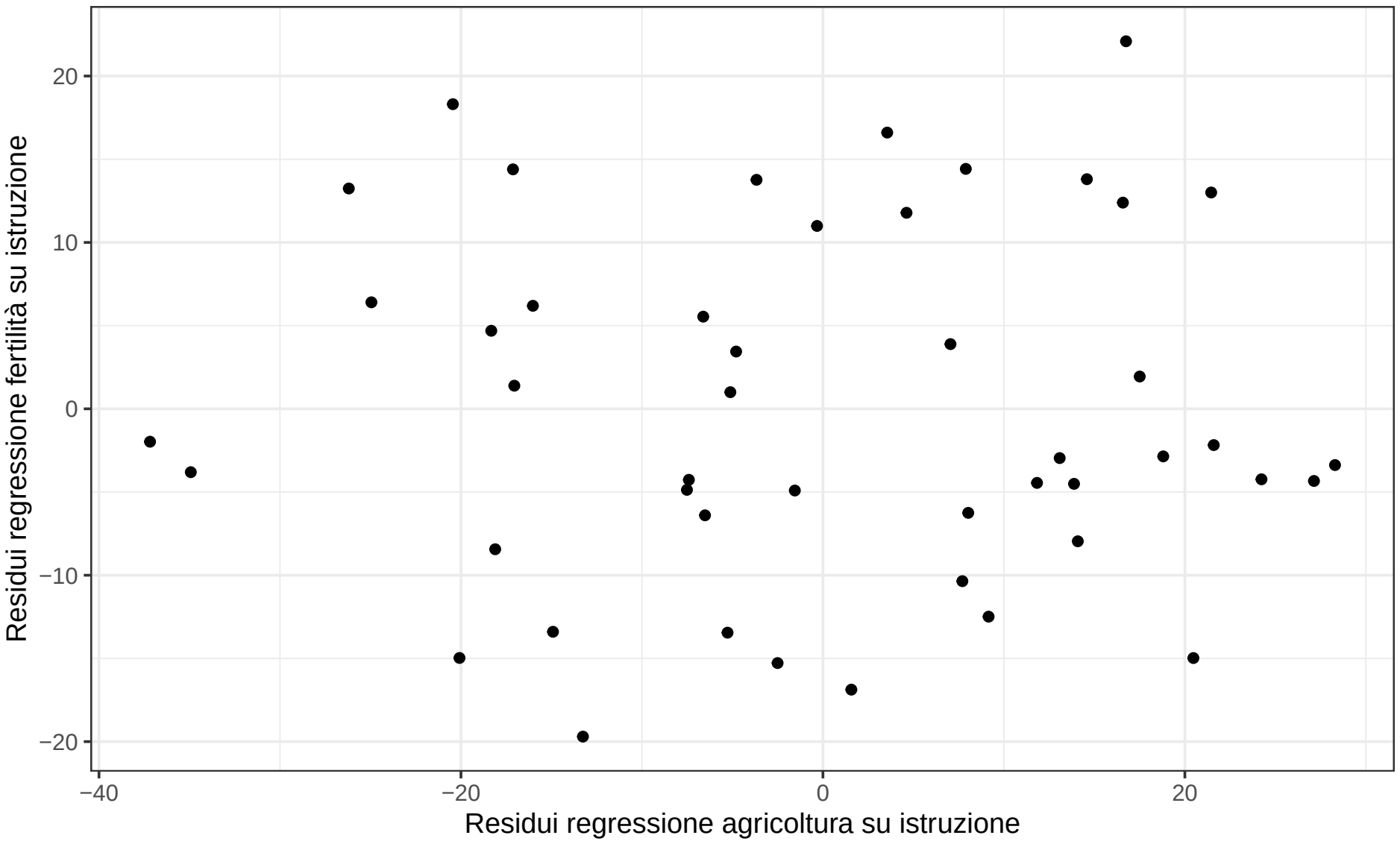
- Infatti, siamo interessati a studiare la relazione tra `agricoltura` e `fertilità` avendo eliminato l'effetto dell'`istruzione`.

- Questa decomposizione richiama al modello di **regressione lineare semplice**.

La correlazione parziale

- Supponendo che le relazioni intercorrenti tra le variabili siano lineari, un modo di procedere è il seguente.
- **Step 1**. Si costruisce un modello di regressione lineare usando `agricoltura` come variabile risposta e `istruzione` come variabile esplicativa.
- Si calcolano i **residui del modello**, che identificano la parte di `agricoltura` non legata all'`istruzione`.
- **Step 2**. In maniera analoga, si costruisce un modello di regressione lineare usando `fertilità` come variabile risposta e `istruzione` come variabile esplicativa.
- Si anche in questo caso i **residui del modello**, che identificano la parte di `fertilità` non legata all'`istruzione`.
- **Step 3**. Si calcola la **correlazione** tra i residui calcolati agli Step 1 e 2.

Diagramma a dispersione dei residui



La correlazione parziale

- Il coefficiente ottenuto allo Step 3 viene chiamato **coefficiente di correlazione parziale** tra `agricoltura` e `fertilità` data l'`istruzione`.
- Calcolato con i dati disponibili, tale coefficiente è pari a -0.0021.
- Il valore è molto vicino allo zero, che indica che tra i residui dei due modelli non esiste una relazione lineare importante, come confermato dal grafico precedente.
- I dati sembrano essere quindi in accordo con la congettura fatta: al netto dell'istruzione, pare non esserci alcun legame diretto tra `fertilità` e `istruzione`.

Esercizio: la correlazione parziale

- Siano x, y e z tre variabili numeriche rilevate sulle stesse unità statistiche. Si indichino $(x_1, \ldots, x_n)$ i dati su x e in maniera analoga per y ed z.

- **Esercizio - proprietà**. Si dimostri che

$$\text{cov}(\tilde{x}, \tilde{y}) = \text{cov}(x, y) - \frac{\text{cov}(x, z)\text{cov}(y, z)}{\text{var}(z)},$$

dove $\tilde{x}$ e $\tilde{y}$ indicano i residui della regressione rispettivamente di x su z e y su z, ovvero

$$\tilde{x}_i = x_i - \bar{x} - \frac{\text{cov}(x, z)}{\text{var}(z)}(z_i - \bar{z}), \qquad i = 1, \ldots, n,$$

e

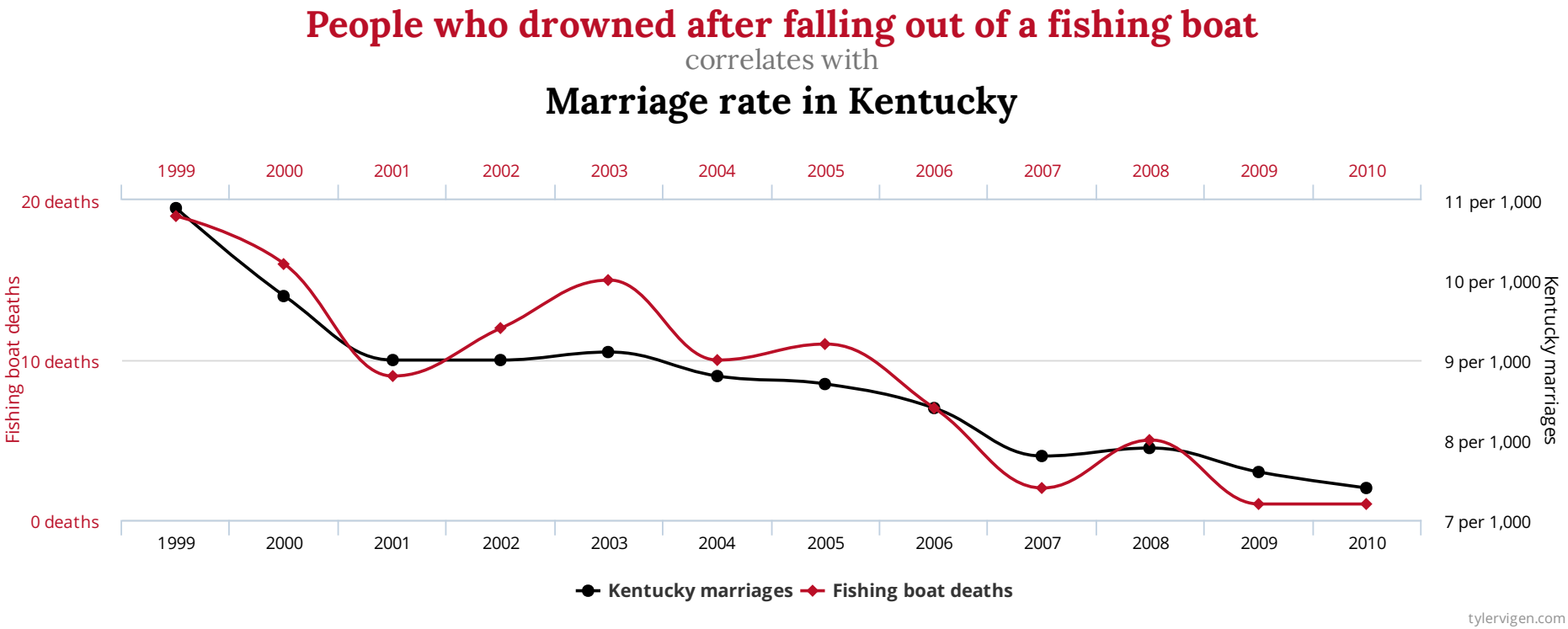
$$\tilde{y}_i = y_i - \bar{y} - \frac{\text{cov}(y, z)}{\text{var}(z)}(z_i - \bar{z}), \qquad i = 1, \ldots, n.$$

- **Nota**. Questa formula permette, insieme alle formule dell'unità K, di calcolare agevolmente il coefficiente di correlazione parziale tra x ed y dato z.

La correlazione spuria

- L'esempio delle province svizzere evidenzia l'esistenza delle **correlazioni spurie**.
- Ci sono sostanzialmente due ragioni per cui si osservano le correlazioni spurie.
- **Motivo I**. La correlazione (spuria) è dovuta all'effetto del caso e l'associazione è interamente dovuta ad oscillazioni casuali.
- Esistono degli strumenti per tenere sotto controllo questo fenomeno, che verranno illustrati nei corsi di inferenza statistica.
- **Motivo II**. La correlazione (spuria) è dovuta all'effetto di una causa comune.
- Questo è il caso delle province svizzere: la correlazione tra `agricoltura` e `fertilità` sembra fosse dovuta alla variabile `istruzione`.

Correlazione spuria: motivo I

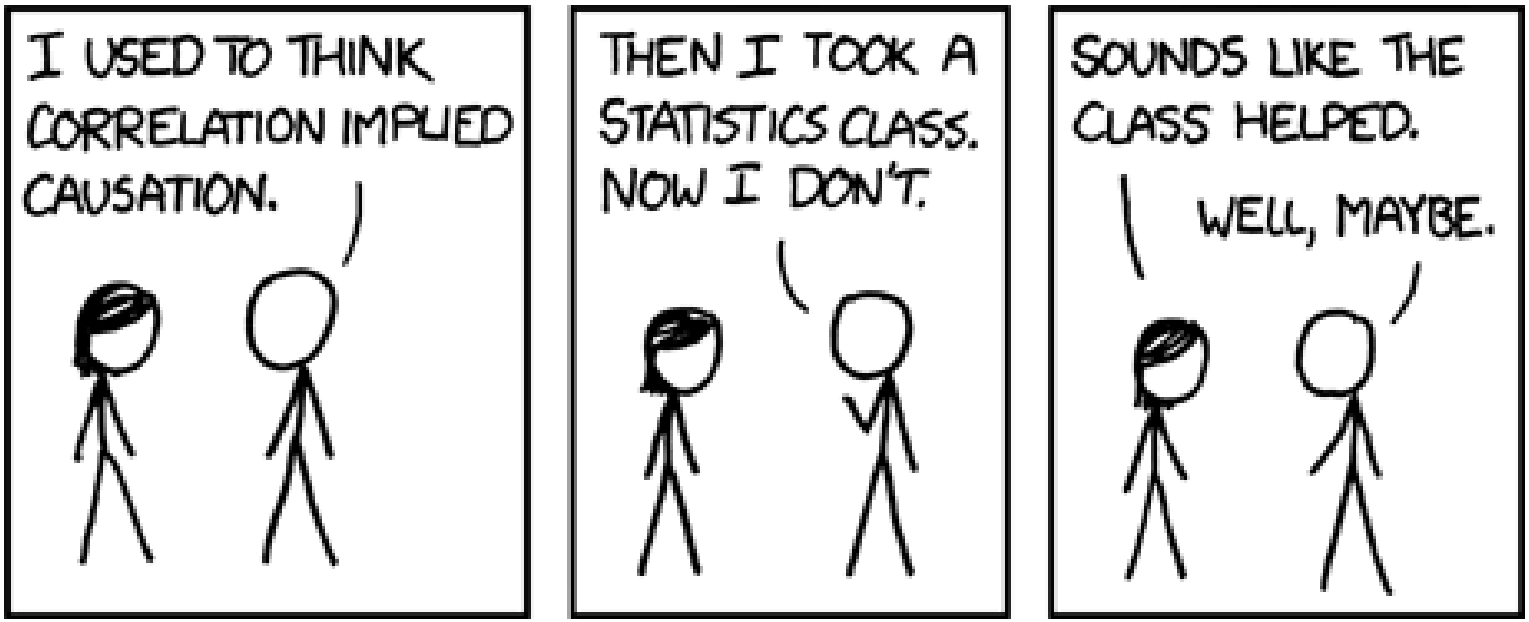


- La correlazione in questo caso è pari a $\rho = 0.95$.
- **Sito web**: `https://www.tylervigen.com/spurious-correlations`

Correlazione spuria: motivo II

- La presenza di correlazione tra due fenomeni **non** è sufficiente per stabilire un **rapporto di causa-effetto**.
- Come ulteriore esempio, pare ci sia una correlazione negativa tra inquinamento ed il numero di pirati nel mondo.
- La correlazione è spuria: i pirati non causano l'inquinamento né viceversa.
- Esiste tuttavia una causa comune, plausibilmente l'industrializzazione, che determina l'andamento di entrambi i fenomeni.
- La **correlazione parziale** aiuta a identificare correlazioni spurie di questo tipo, ma necessita di congetture fornite da esperti del settore.
- Identificare rapporti di causa-effetto è difficile e la statistica da sola non basta.

Correlazione non implica causalità



- **Sito web**: `https://xkcd.com/552/`