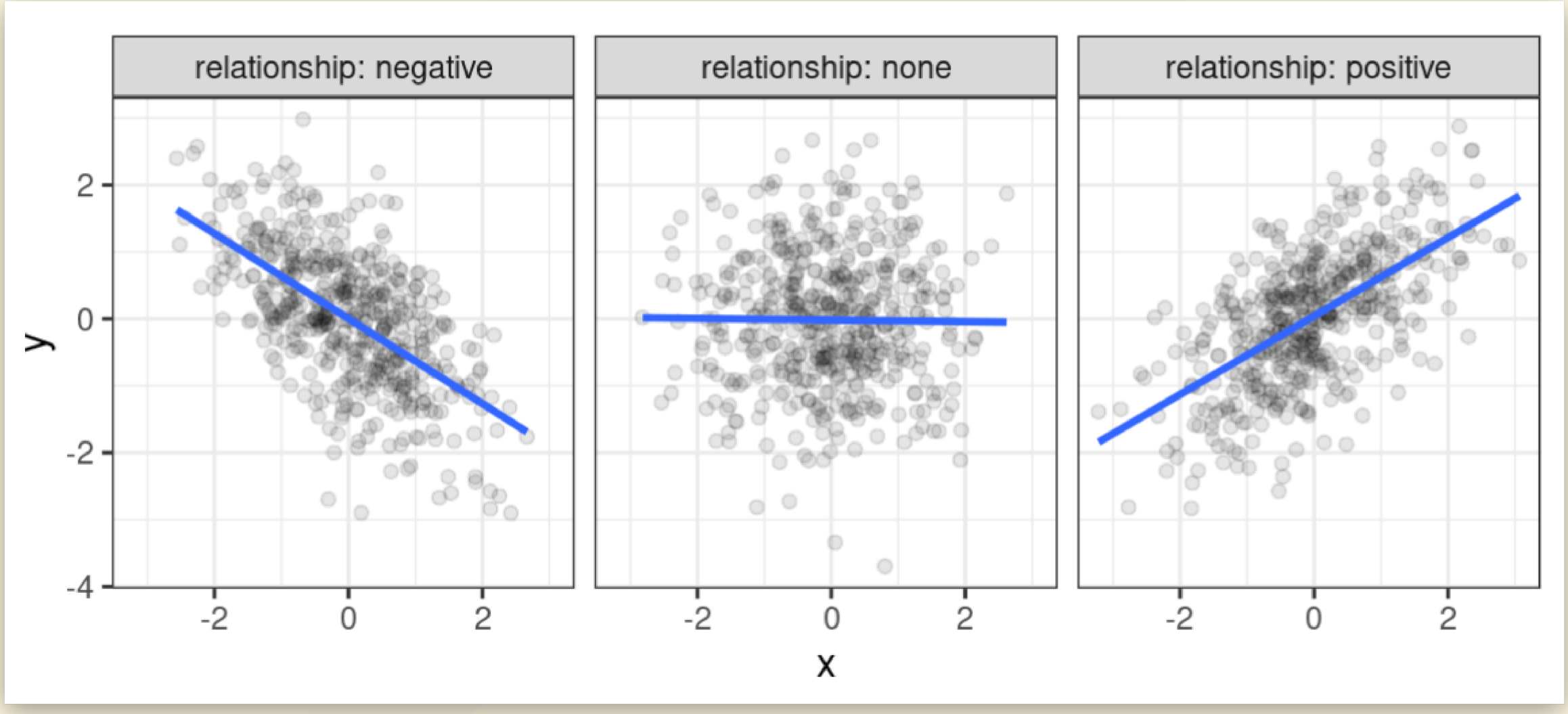
相关性和因果性

相关性

- 变量的变化是有关联的 (associate)
- 或，一个变量发生的变化与另一个变量发生的变化是有关联的。

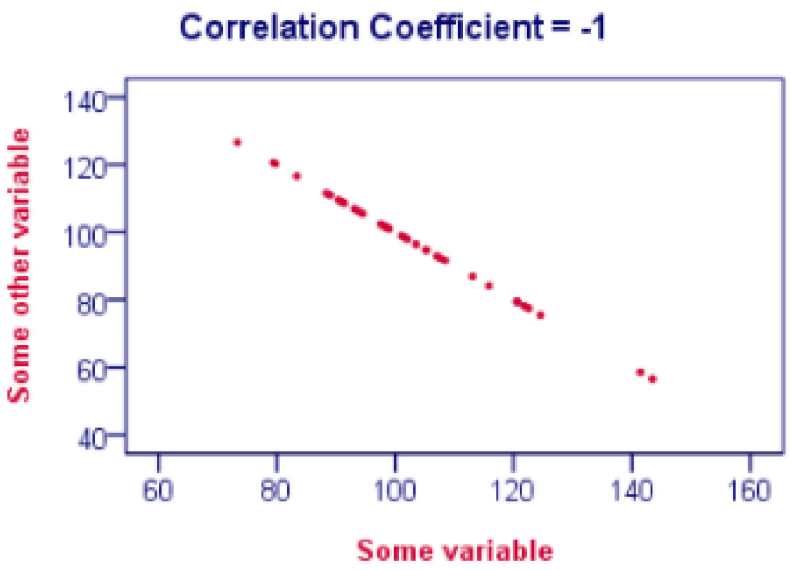
- i.e.
  - 变量：儿童的年龄 & 体重
  - 样本量：10位儿童的数据

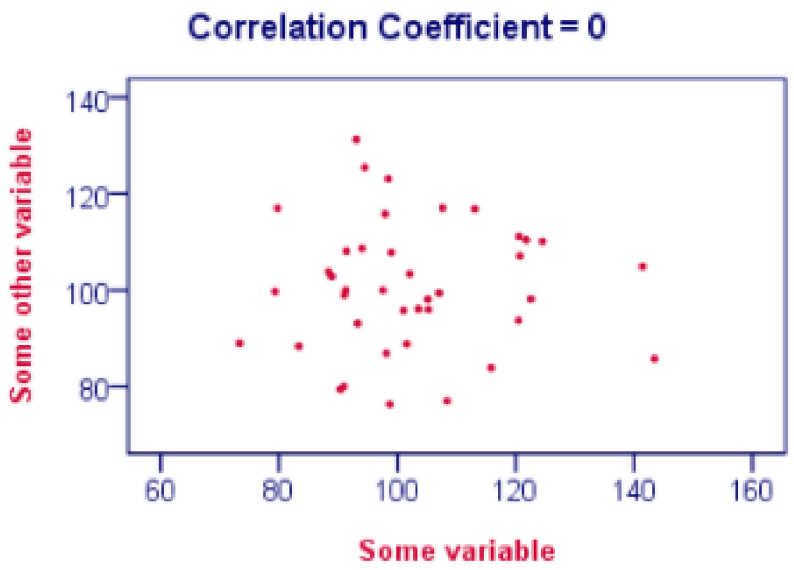

relationship: negative
relationship: none
relationship: positive
y
x
2
0
-2
-4
-2
0
2

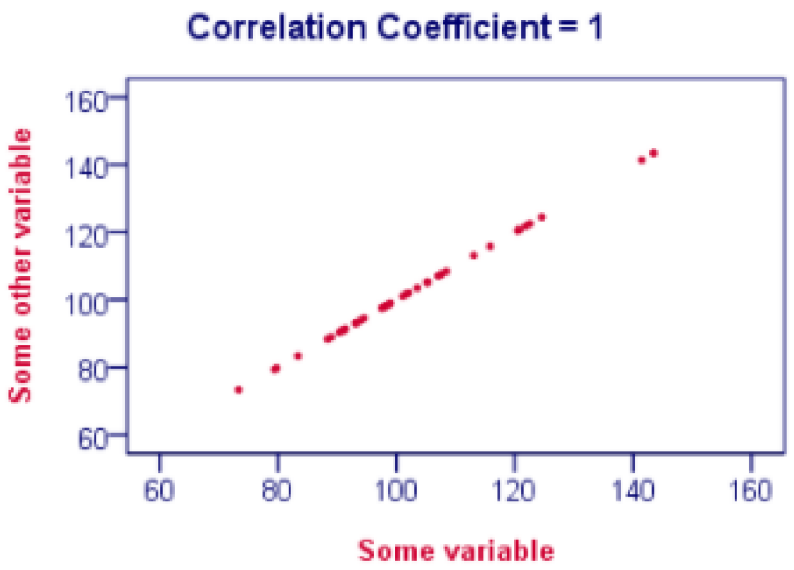
- **相关性的测量**
  - **皮尔逊相关系数** (Pearson correlation coefficient)
    - 衡量线性关系
    - 取值在[-1, 1]之间

$$r_{XY} = \frac{cov(X,Y)}{\sigma_X * \sigma_Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

- A Pearson correlation is a number between -1 and +1 that indicates to which extent 2 variables are **linearly related**.

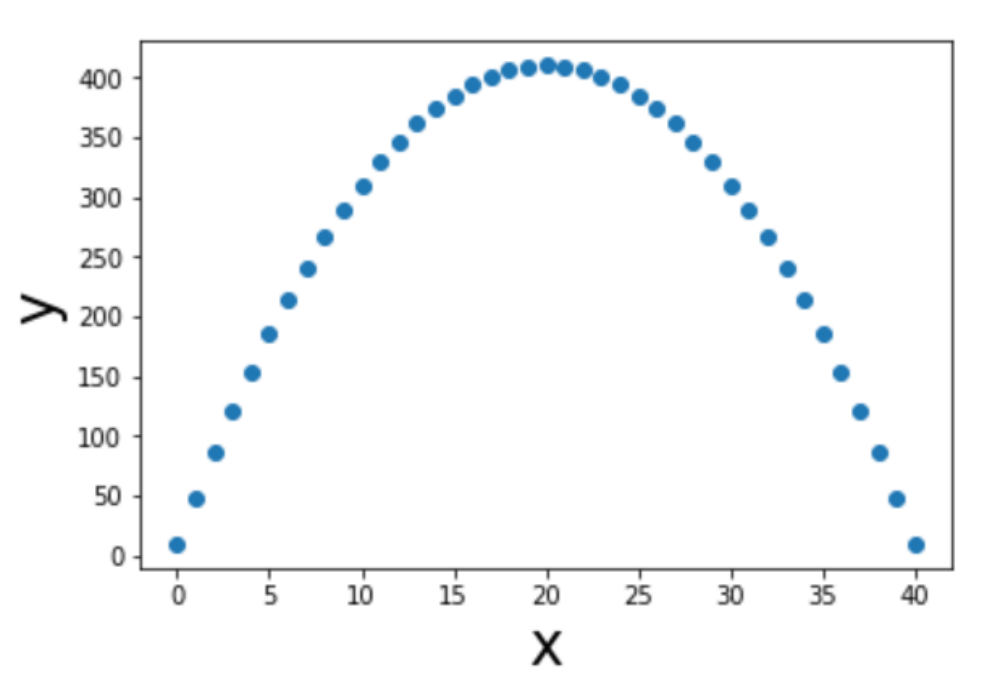

Correlation Coefficient = -1
Some other variable
140
120
100
80
60
40
60
80
100
120
140
160
Some variable


Correlation Coefficient = 0
Some other variable
140
120
100
80
60
80
100
120
140
160
Some variable


Correlation Coefficient = 1
Some other variable
160
140
120
100
80
60
60
80
100
120
140
160
Some variable

- NOTE: **相关性强不一定意味着相关系数高**
- i.e.
  - $y = 10 + 40x - x^2$
  - $x \in [0, 40]$
  - $r(x, y) = 0.00$

```
x = range(41)
y = [10 + 40*i - i**2 for i in x]
print('r(x,y): {}'.format(np.corrcoef(x,y)[0,1]))
```

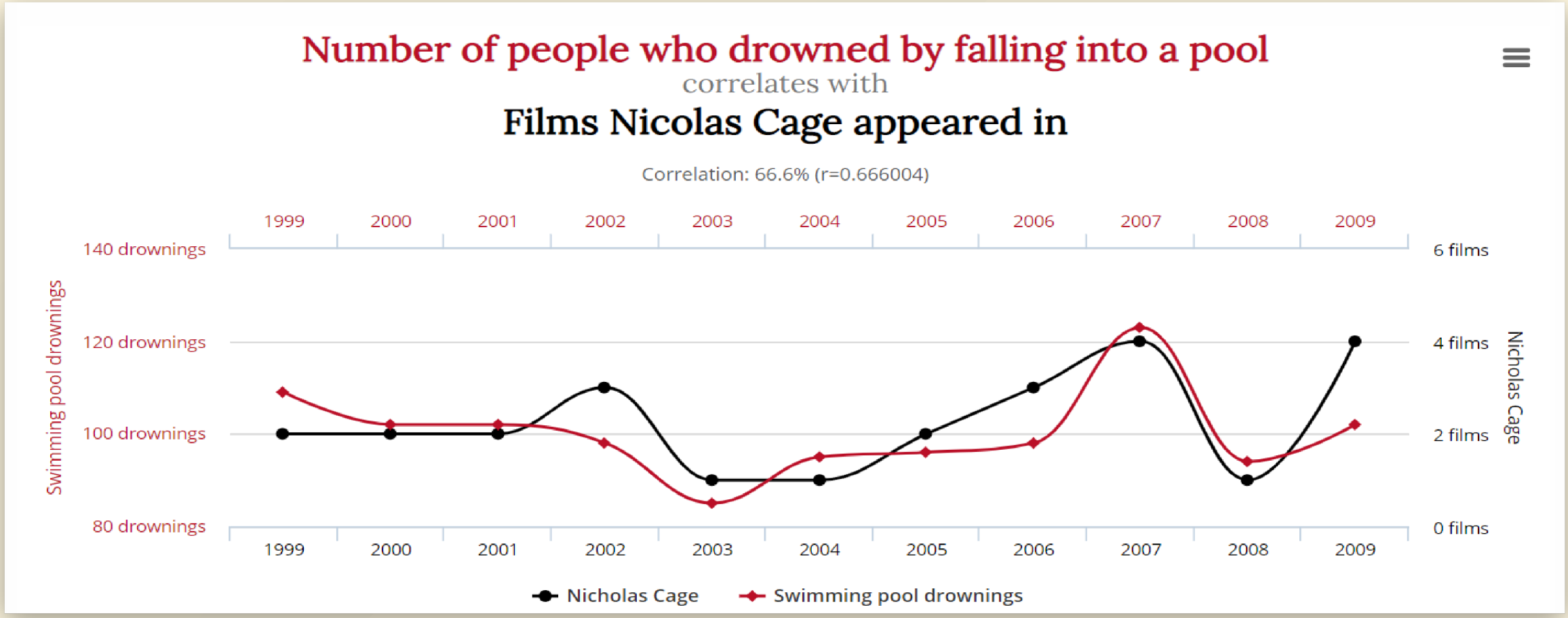


相关性 v.s. 因果性

- 相关性是【**对称的**】
  - r(x,y) = r(y,x)
- 因果关系【**可能不是对称的**】
  - 有方向的

- 如果变量A和B存在相关性，可能说明
  - A和B存在单向的因果关系
  - A和B存在双向的因果关系
  - 其它变量引起A和B的变化

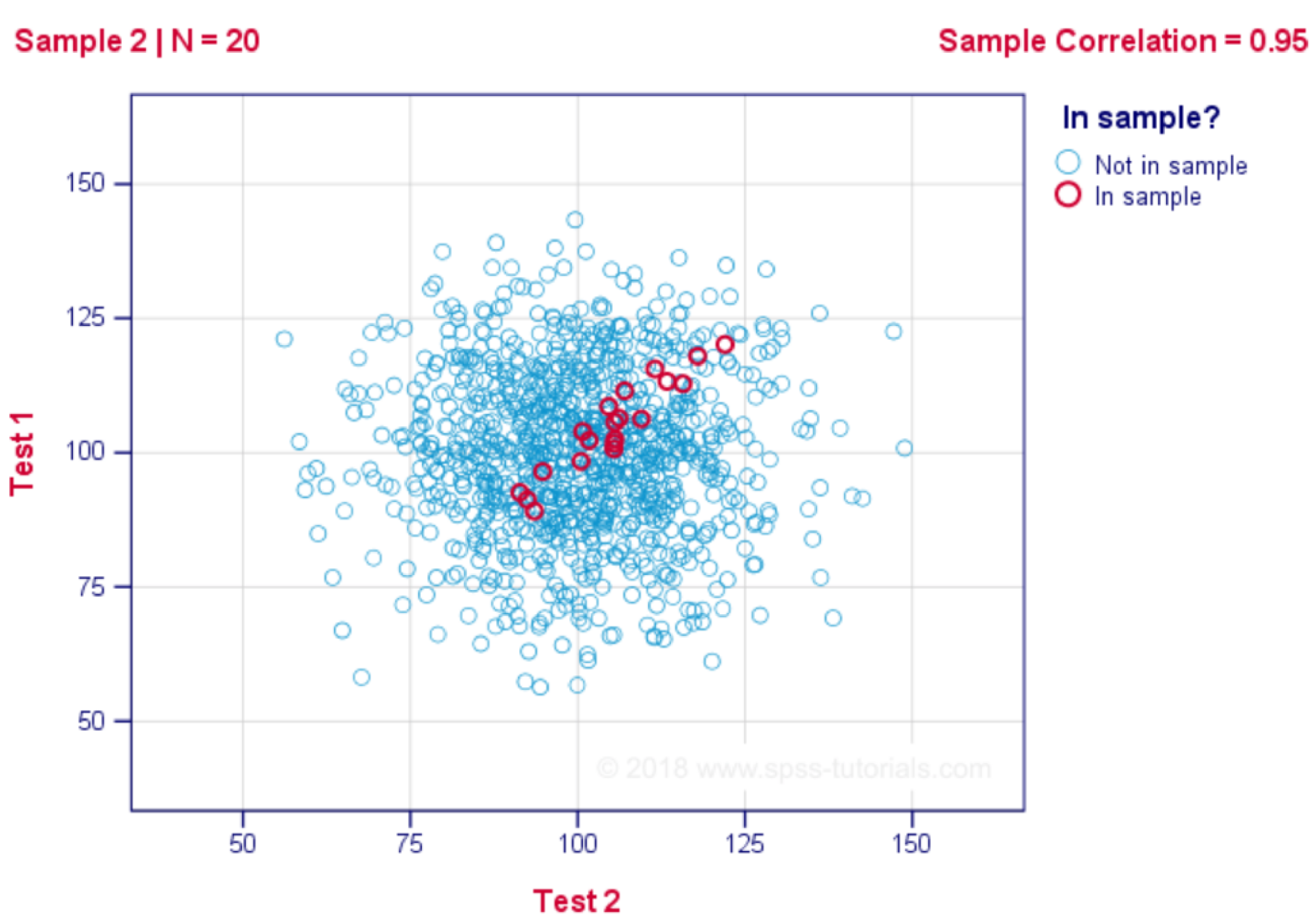
- 如果变量A和B存在相关性，可能说明
  - A和B没有任何关系



- i.e. 当两个随机变量样本量很小时，也有可能会出现相关性，但两者无任何关联。

```
import numpy as np
np.random.seed(10)
# get two random number and calculate their correlation coefficient; size: sample size
def get_r(size):
    ab = np.random.rand(2,size)
    return np.corrcoef(ab)[0,1]
sample_size = [10,100,1000,10000]
print('Pearson correlation coefficient:')
for i in sample_size:
    print('Sample_size = {:<5d}: {:.2f}'.format(i, get_r(i)))
```

- **Pearson correlation coefficient:**
  - **Sample_size = 10 : -0.64**
  - **Sample_size = 100 : 0.04**
  - **Sample_size = 1000 : 0.04**
  - **Sample_size = 10000: 0.00**


Sample 2 | N = 20
Sample Correlation = 0.95
In sample?
Not in sample
In sample
Test 1
Test 2
150
125
100
75
50
50
75
100
125
150
© 2018 www.spss-tutorials.com

相关性不代表因果性

因果关系也不一定含义着相关性

相关性不是因果关系的必要条件