相関

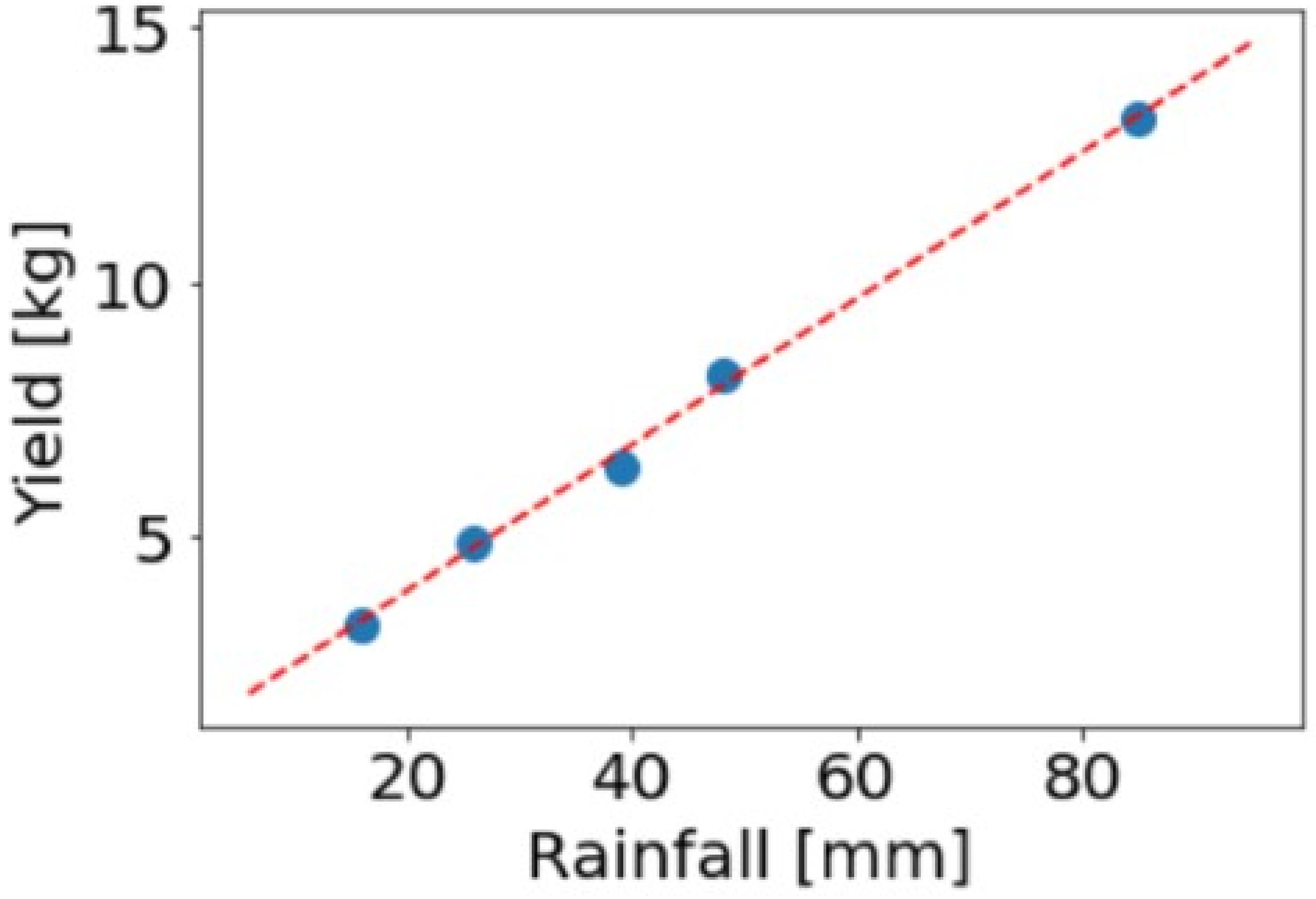
2データ間の関係を調べる

- あるデータが、別のデータと関連しているかどうかを調べたい、という場合を考えます。
- 例：月間降雨量とある作物の収穫量のデータがあり、月間降雨量が収穫量に影響しているかどうかを調べたい
  - 降雨量（mm）：　16,　85,　39,　26,　48,　…
  - 収穫量（kg）：　3.3,　13.2,　6.4,　4.9,　8.2,　…
- データをただ眺めているだけではわからないので…ともかく可視化してみましょう!

散布図で可視化してみると…

降水量（mm）：	16,　85,　39,　26,　48,　…
収穫量（kg）：	3.3,　13.2,　6.0,　5.3,　8.2,　…

- 降水量をx軸に、収穫量をy軸にして、散布図をプロットすると…
  - (16,3.3), (85,13.2), (39,6.0), (26,5.3), (48,8.2), …

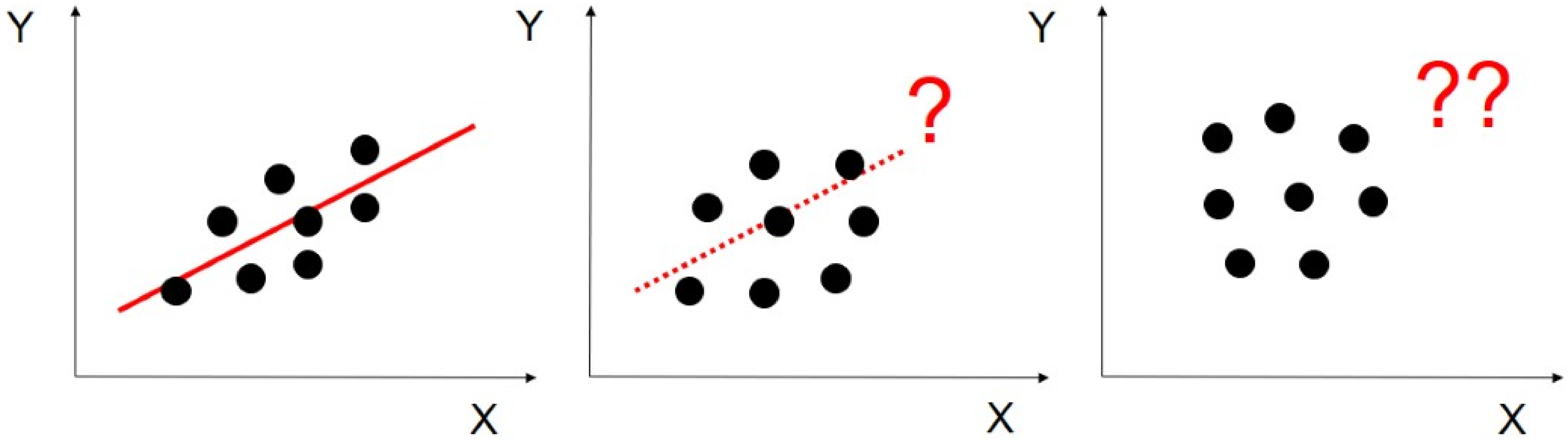


➡ 比例関係にあることが一目瞭然！

散布図を描くと2データ間の関係がよくわかります

関係の強さはいろいろ

- 2データ間の関係がいつも明確であるとは限りません



- このことから、「関係の強さ」を数値として示すことが必要であることがわかります

ピアソン相関係数

2データ間の関係の強さを表す典型的な指標に「ピアソン(Pearson)相関係数」（単に「相関係数」とも呼ばれる）があります

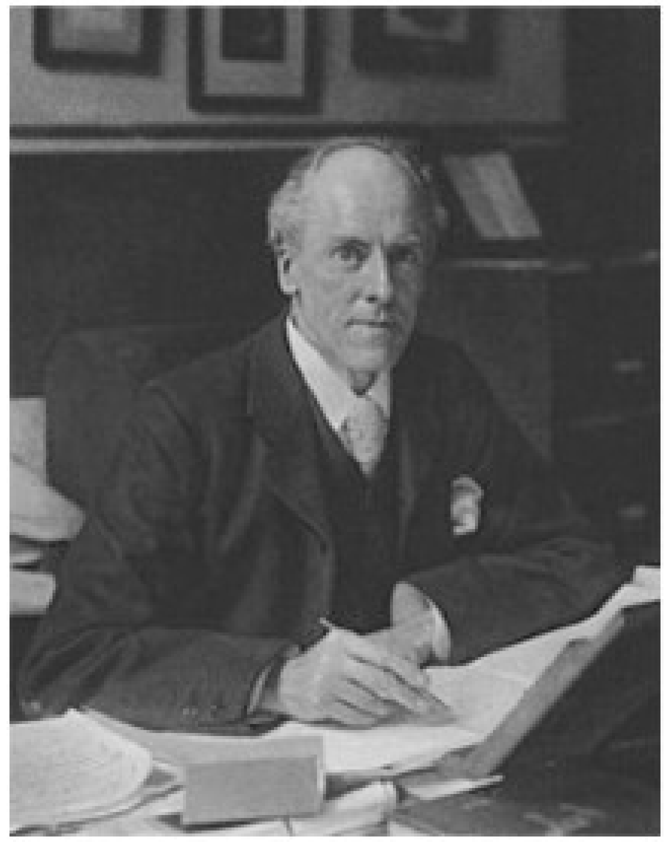
N組のデータ	データX：	$x_1, x_2, \ldots, x_N$
	データY：	$y_1, y_2, \ldots, y_N$

平均 $\bar{X} = \sum_{i=1}^{N} x_i \ / \ N$ $\quad \bar{Y} = \sum_{i=1}^{N} y_i \ / \ N$

変動 $D_x = \sum_{i=1}^{N} (x_i - \bar{X})^2 \quad D_y = \sum_{i=1}^{N} (y_i - \bar{Y})^2$

共変動 $D_{xy} = \sum_{i=1}^{N} (x_i - \bar{X})(y_i - \bar{Y})$

ピアソン相関係数 $r_{xy} = D_{xy} / (\sqrt{D_x} \sqrt{D_y})$

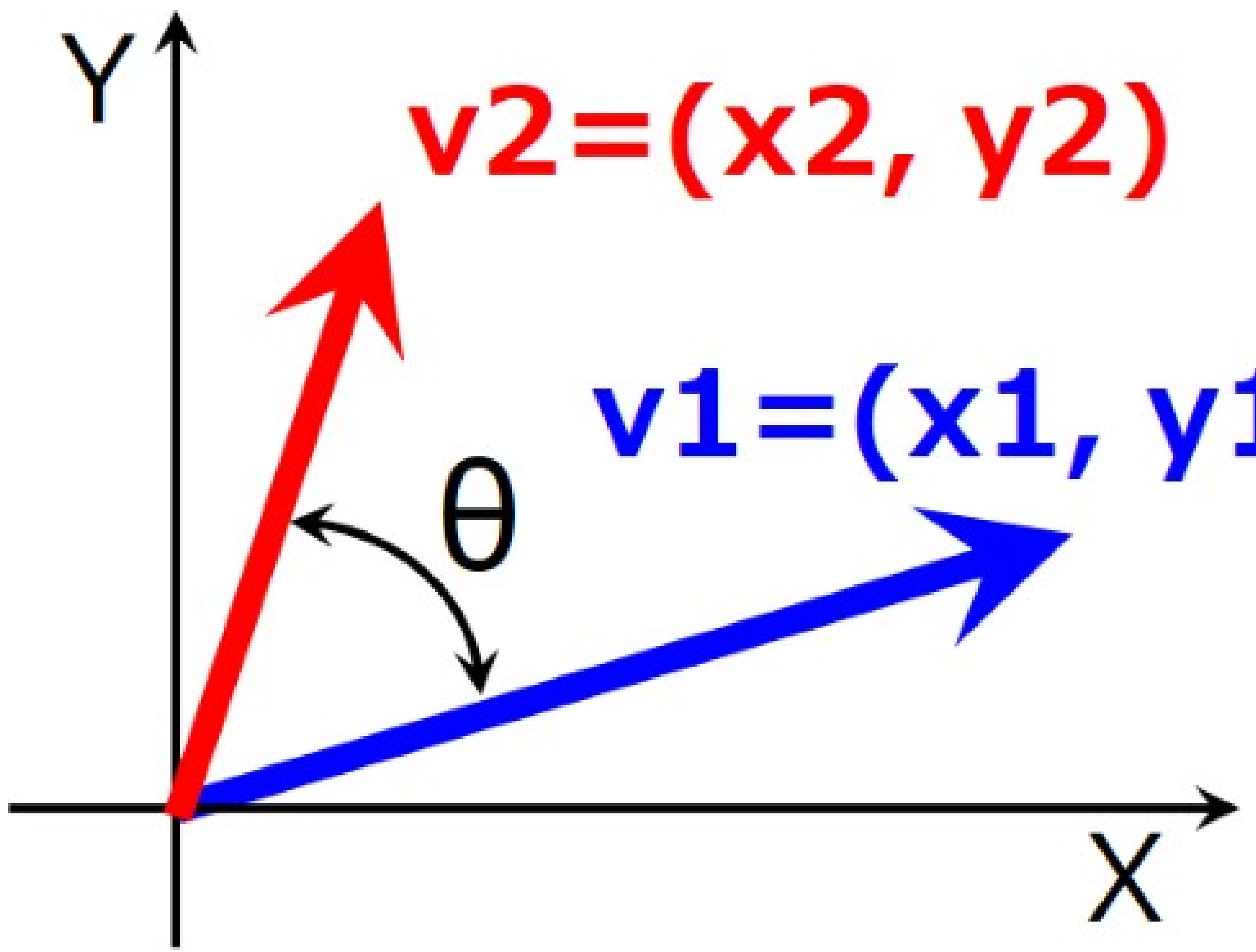


Karl Pearson
1857-1936 英
Wikipediaより

ピアソン相関係数

- -1から1までの値をとる
- 正の値のとき (正相関)
  - Xが増えると Yも増える傾向がある
  - 1 に近いほど、その傾向が強い
- 負の値のとき (負相関)
  - Xが増えると Yが減る傾向がある
  - -1 に近いほど、その傾向が強い
- 絶対値が小さいとき (0付近, 無相関)
  - Xの増減と Yの増減の (線形の) 関係は弱い

(復習) ベクトルの話



以下の手順でベクトル間の角度が簡単にわかる！

1.「内積」を求める

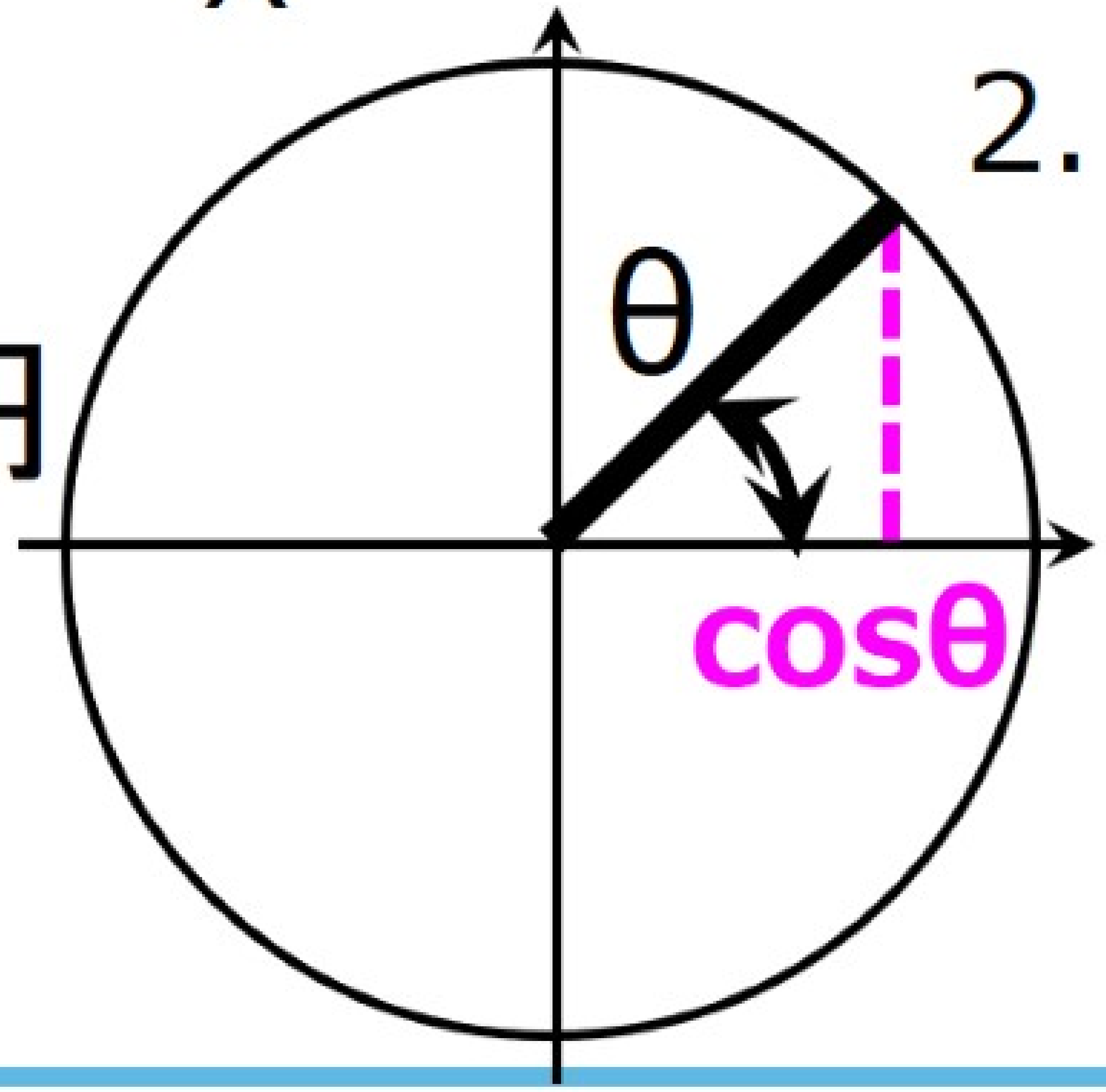
v1=(x1, y1)
*　*
v2=(x2, y2)

v1・v2=x1*x2+ y1*y2

2. 内積を、それぞれの長さで割ると cos

$$\cos\theta = \frac{v1 \cdot v2}{|v1| * |v2|}$$

半径1の円
（単位円）



3. x座標がちょうど cos θ になる
角度が θ

ピアソン相関係数の図形的理解

- X, Yそれぞれについて、平均からのずれ（偏差）ベクトルを考えると…

データXの偏差ベクトルΔX: $x_1 - \overline{X}, x_2 - \overline{X}, \ldots, x_N - \overline{X}$

データYの偏差ベクトルΔY: $y_1 - \overline{Y}, y_2 - \overline{Y}, \ldots, y_N - \overline{Y}$

相関係数 $r_{xy} = D_{xy} / (\sqrt{D_x} \sqrt{D_y})$

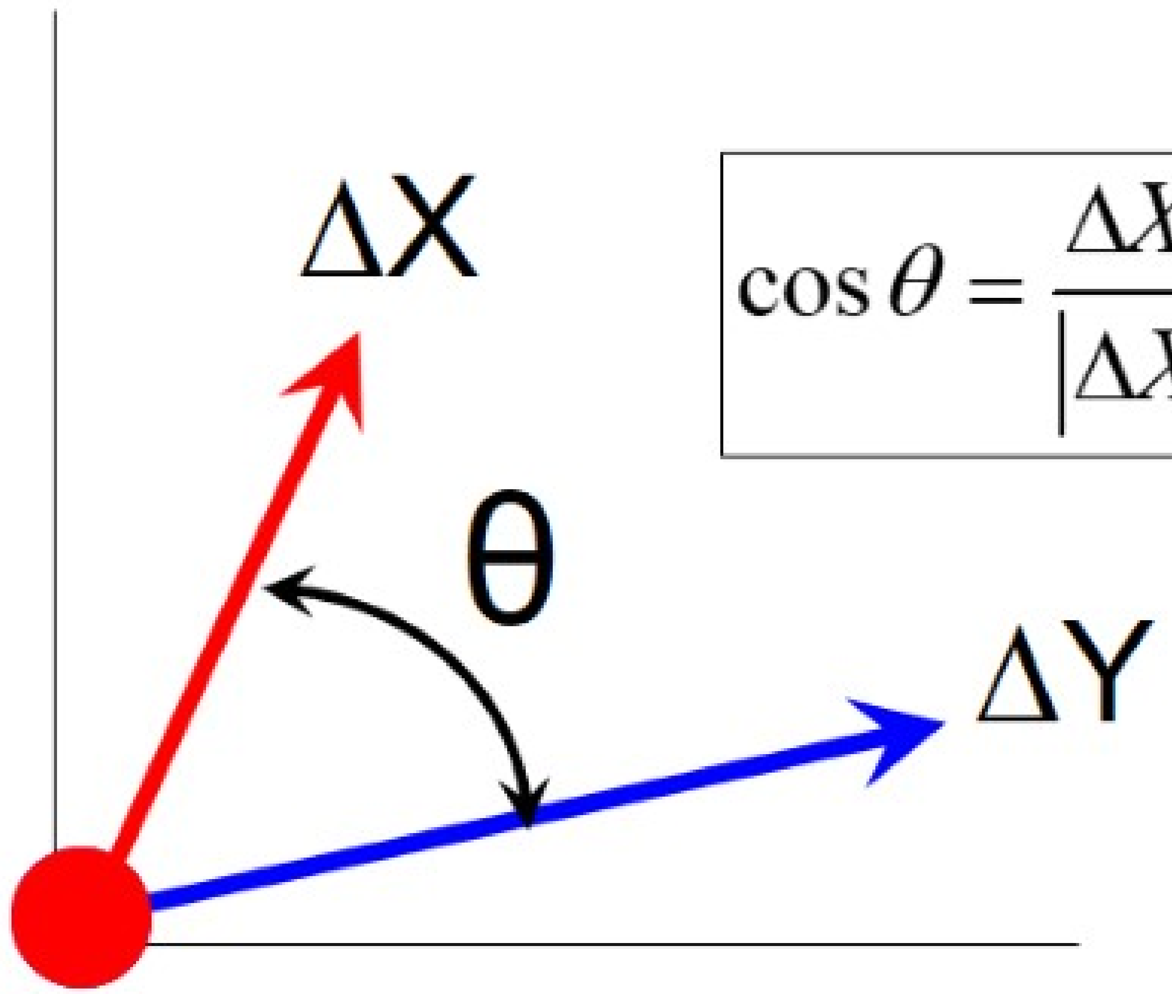
$$D_{xy} = \sum_{i=1}^{N} (x_i - \overline{X})(y_i - \overline{Y})$$

ΔX・ΔY（内積）

$$D_x = \sum_{i=1}^{N} (x_i - \overline{X})^2$$

ΔXの長さ

ΔYの長さ



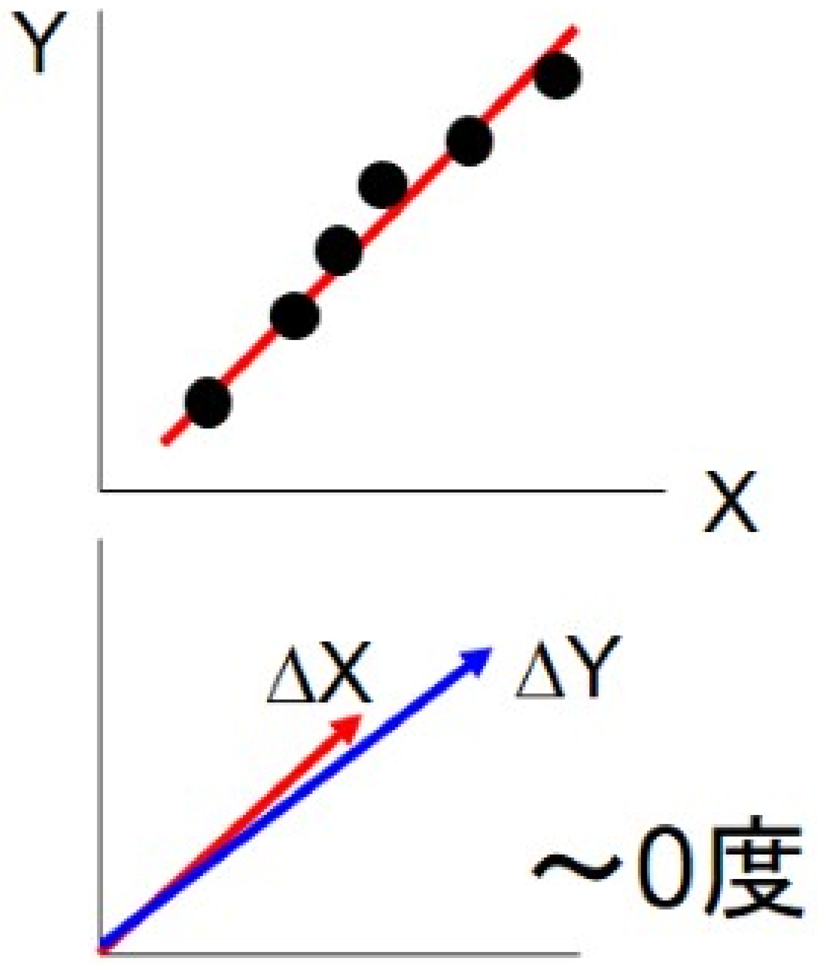
$$\cos\theta = \frac{\Delta X \cdot \Delta Y}{|\Delta X||\Delta Y|}$$

- 相関係数はX, Yの偏差ベクトルΔX・ΔY (平均からの差ベクトル)の内積を、それぞれのベクトルの長さで割ったもの
- すなわち偏差ベクトル間の角度（のコサイン）！

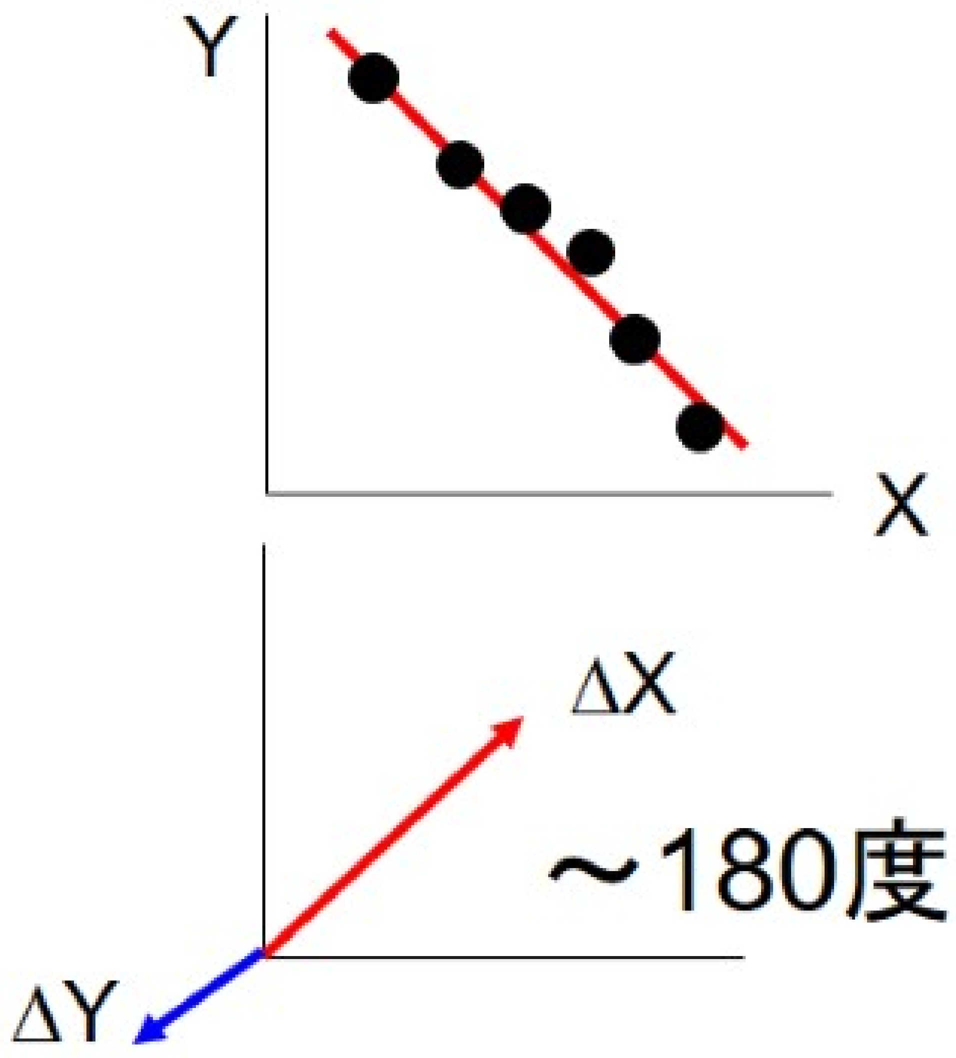
ピアソン相関係数の値の意味

- -1から1までの値をとる。正の値: Xが増えると Yも増える、負の値: Xが増えると Yは減る。
- 絶対値が大きいと、XとYは強く関係。絶対値が小さい (0付近)だとXとYは (線形の) 関係が弱い。

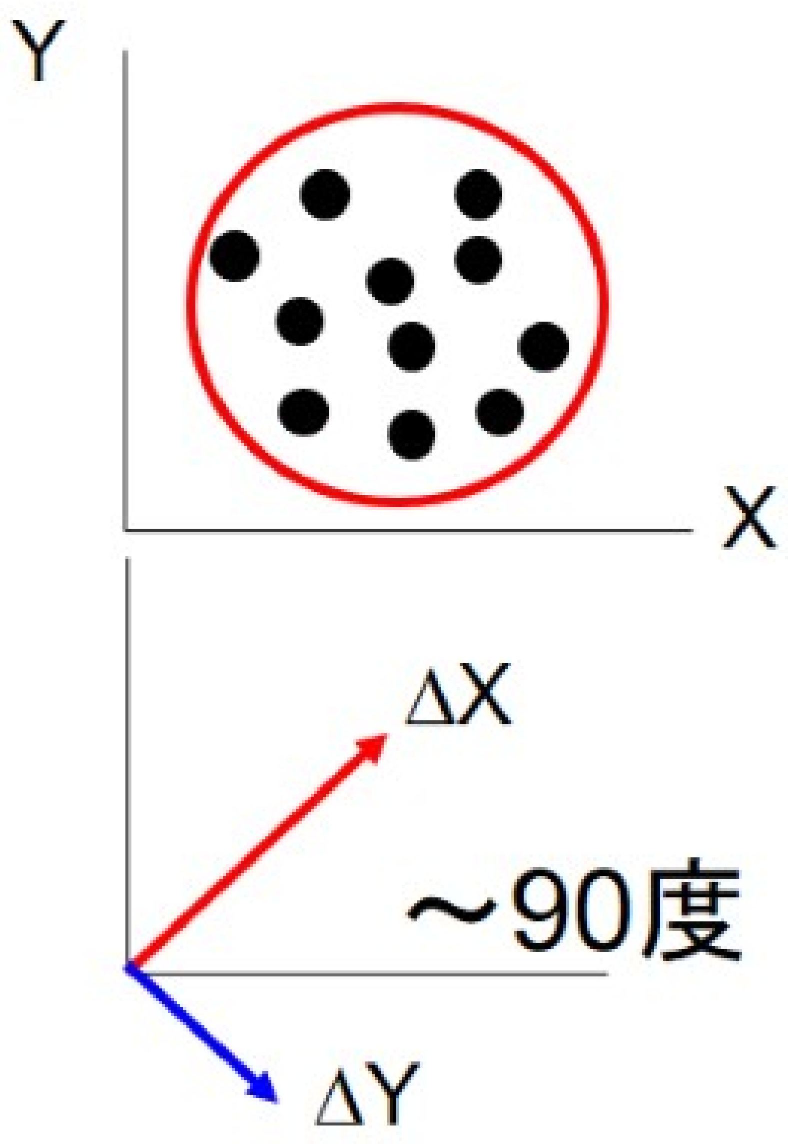
平均からのずれ方向が似ている → 両者は関係がある



相関係数 〜1
(強い正の相関)



相関係数 〜 −1
(強い負の相関)

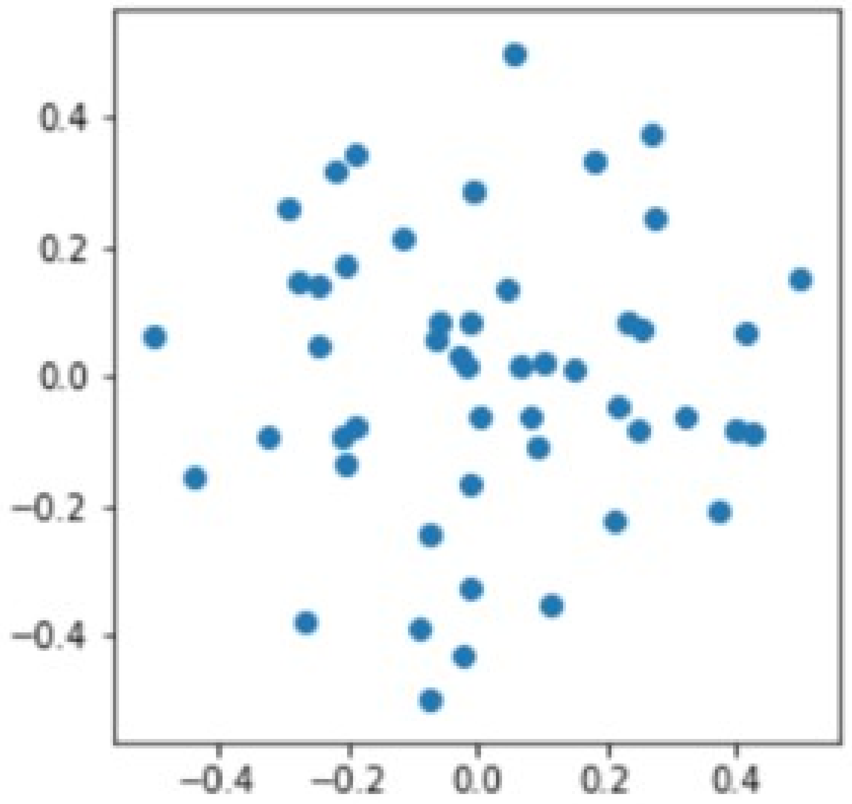


相関係数 〜 0
(線形の関係は弱い)

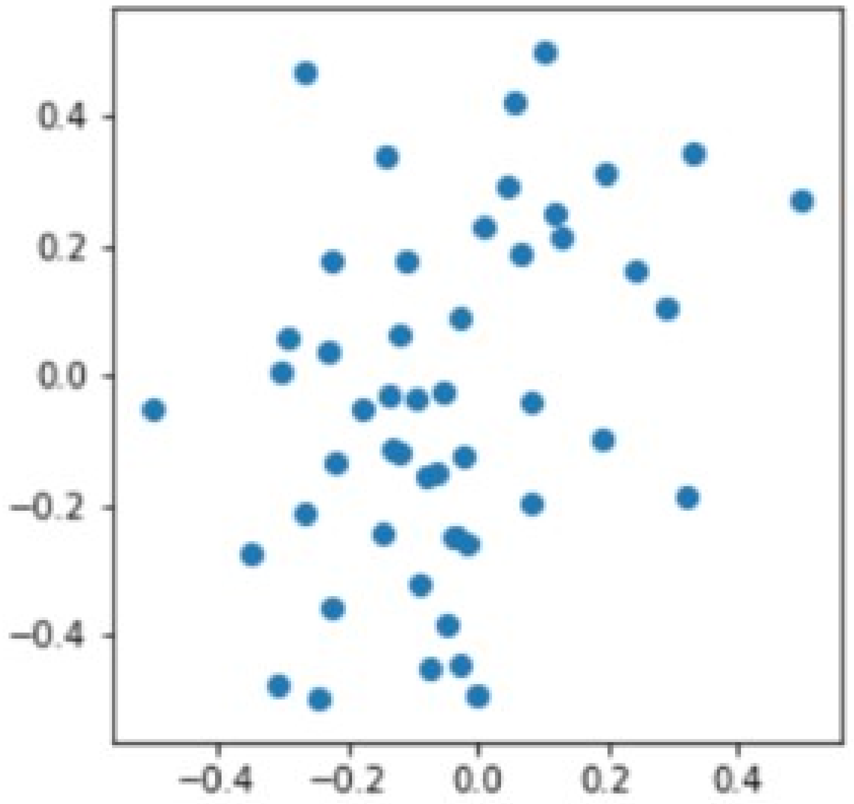
ピアソン相関係数の目安

- 一般に、ピアソン相関係数の絶対値が0.2以下だとほぼ相関なし、0.3～0.4付近だと弱い相関あり、0.5～0.7付近だと相関あり、0.7以上だと強い相関あり、というのが目安とされています（あくまで目安です）

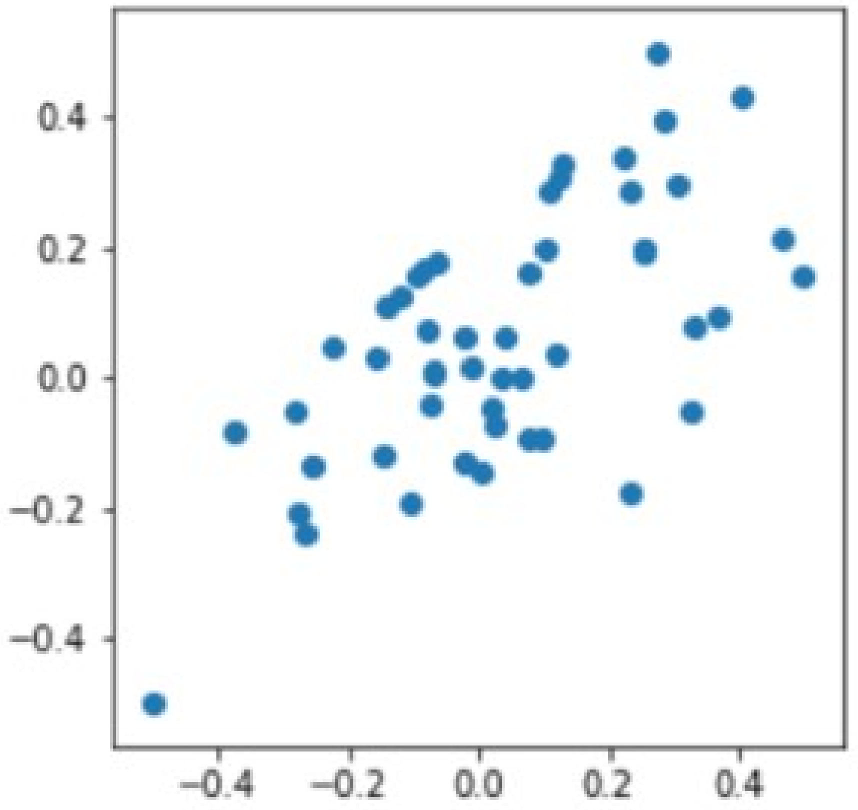
グラフ例



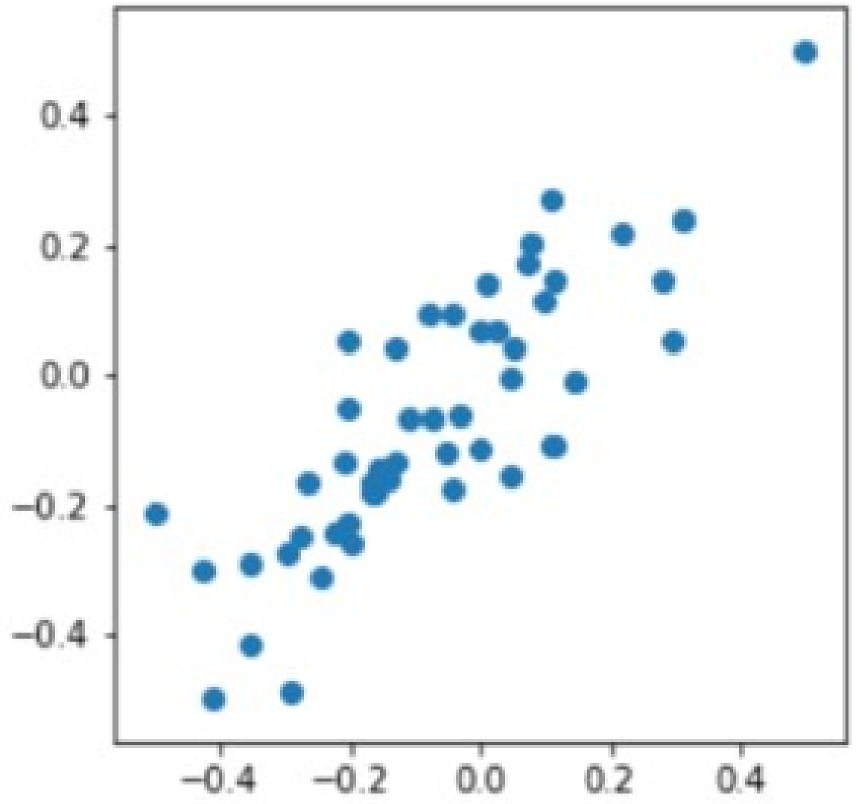
相関係数0.03



相関係数0.34



相関係数0.63



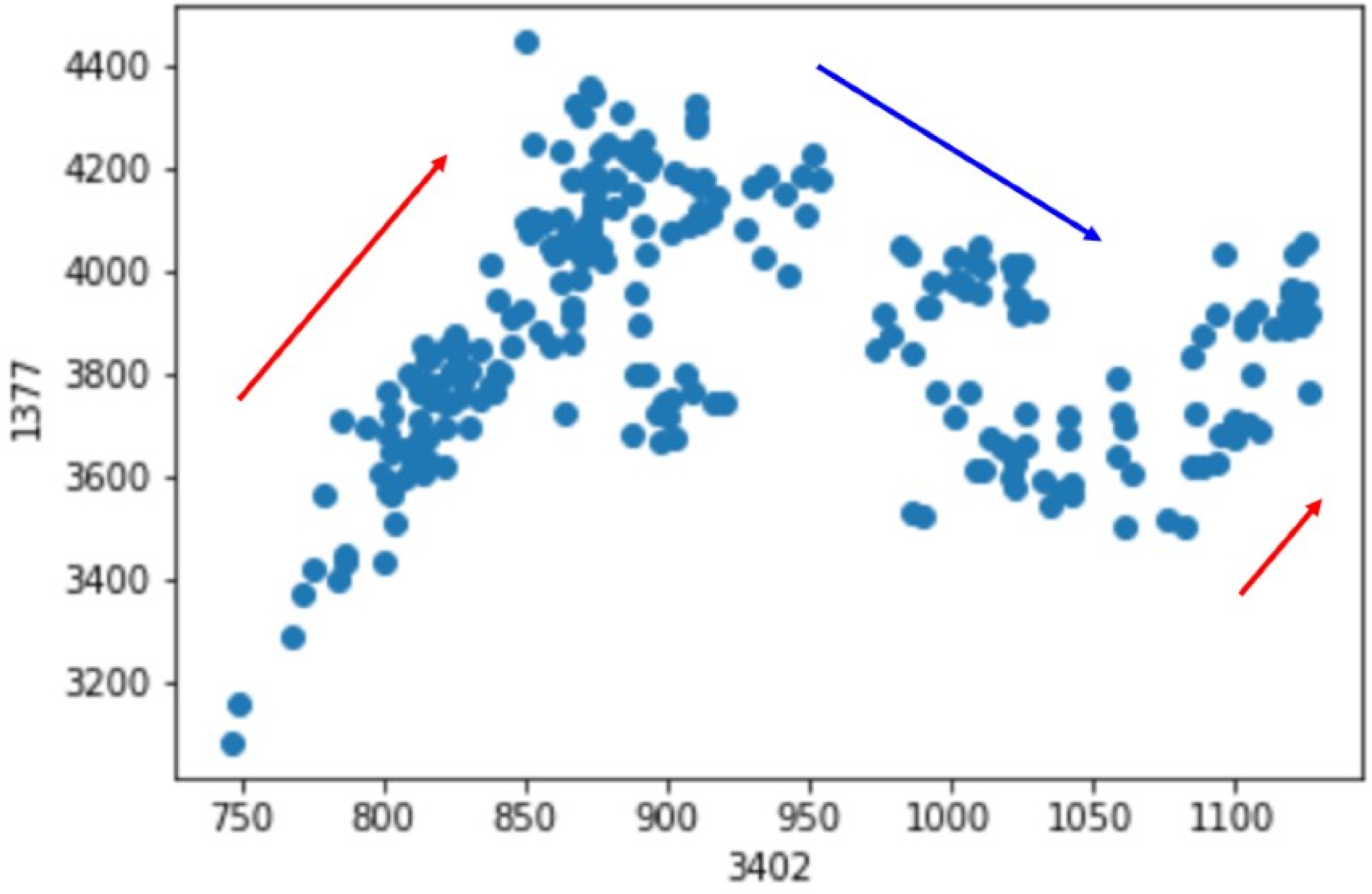
相関係数0.83

相関係数はあくまで「直線的な関係の強さ」

- 相関係数が0付近だからXとYの間に「関係が全くない」わけではない。あくまで「線形の関係」がないだけなので注意。
- 必ず散布図も描くこと。2変数が "ばらばら" で関係なし (独立) なら、相関係数は0付近。でも逆に相関係数が0付近だと2変数はばらばら、とは限らない。独立であれば無相関だが、無相関であれば独立、ではない。
- 非線形の関係はどのように扱う？ → 後に学ぶ機械学習法などを使う

例: 相関係数が0付近でも…

散布図では…



企業コード 1377 (サカタのタネ)
と
企業コード 3402 (東レ)
の2018年の株価

ピアソン相関係数: 0.04

しかし「独立」ではなさそう。

折れ線グラフでは…


負の相関っぽい
正の相関っぽい
正の相関っぽい
3402
1377
1.1
1.0
0.9
0.8
0.7
Ratio of closing price (adjusted)
2018-01
2018-03
2018-05
2018-07
2018-09
2018-11
2019-01
Date

(参考) その他の相関係数

- 数値同士ではなく、大小関係 (順位付け) がどれだけ一致しているかを調べる相関係数もあります。
  - スピアマンの順位相関係数
  - ケンドールの順位相関係数
  - …