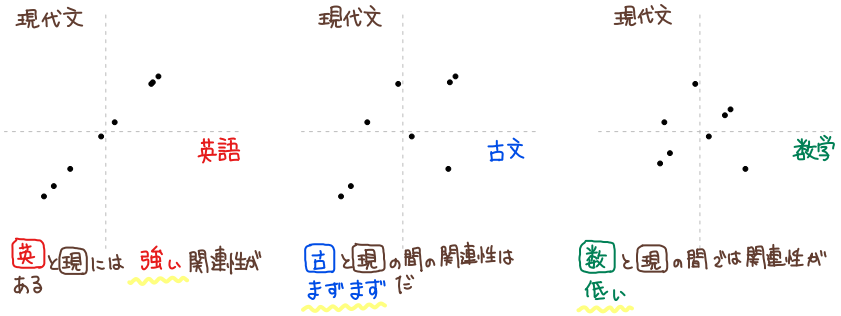
■　基本的な統計量 III（二つの変数の 相関）

【目的】x 軸と y 軸の関連性の方向と大きさを測る指標を作りたい。



（1）共分散

これは、平均と各点との間の 二次の距離（面積）の平均。

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

※共分散には最大値と最小値が存在する。

$$-s_x s_y \leq s_{xy} \leq s_x s_y$$

（2）相関係数（ピアソンの積率相関係数）

これは、共分散が最大値と比較したときに占める割合。

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

（s_{xy}：今回とった共分散、$s_x s_y$：関連度がMAXのときの共分散）

※相関係数には最大値と最小値が存在する。

$$-1 \leq r_{xy} \leq 1$$

（-1 ↑ −100%、1 ↑ 100%）

Ⓟ最大値で割ること

Aさん：90点（990点満点）
Bさん：90点（200点満点）

$\frac{90}{990} \fallingdotseq 0.09$（9%）

$\frac{90}{200} = 0.45$（45%）

最大値を分母

Ⓟ 共分散の弱点

単位に依存してしまう。

（ケース1）g, 1000, 1000×100 $= 10^5$, 100, cm

（ケース2）kg, 1, $|x| = 1$, m

Ⓟ 共分散の上限と下限

共分散 S_{xy} は，関連性が最大化するとき，次の値をとる

$S_{xy} = S_x S_y$（右肩あがり）

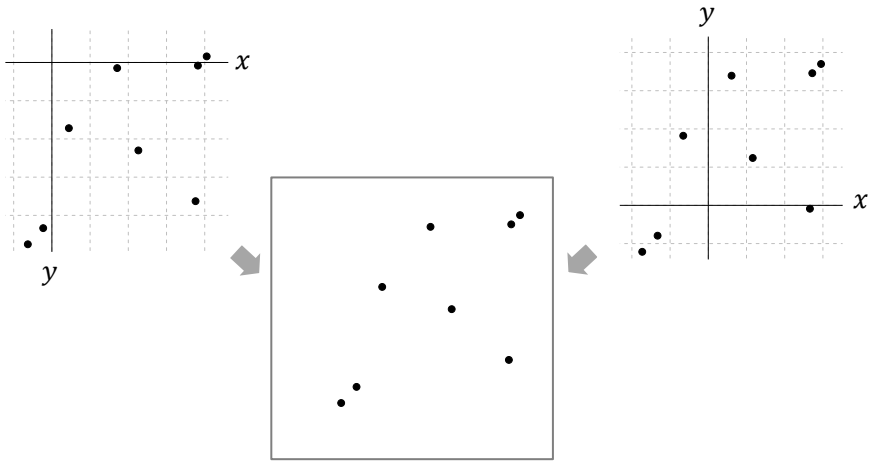
$S_{xy} = -S_x S_y$（右肩下がり）

Ⓟ 相関係数の良いところ

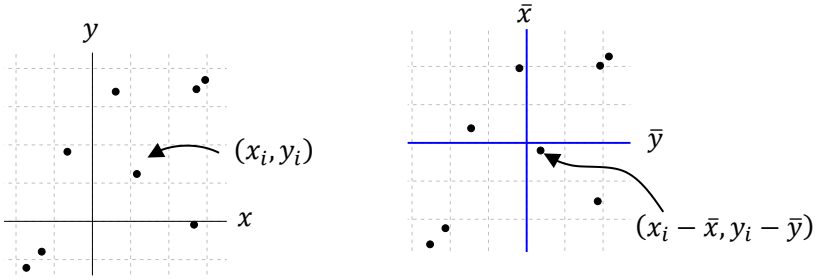
① 正負：関連の方向
② 値：関連度合の強さ
　1：相関が（正の方向に）MAX
　0：相関がない
　−1：相関が（負の方向に）MAX
③ 場合分け不用
④ サンプルサイズに依存しない
⑤ データの単位に依存しない

（準備）軸を取り払う

もともとの x 軸、y 軸の位置は関連度合いには無関係なので、取っ払ってしまいたい。

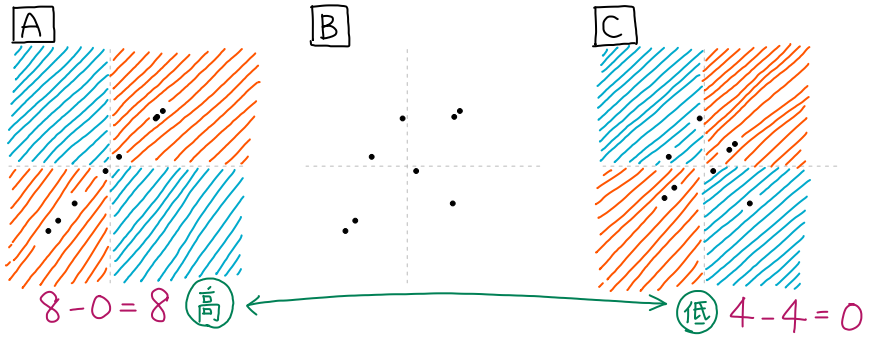


しかし、座標自体は維持したいので、新しく、横軸を$\bar{y}$、縦軸を$\bar{x}$の位置に据えた新しい座標を考える。これはもともとの座標を平行移動したことに相当する。



（第一案）球の個数の差を考える

（式１）　「第一／三象限の球の数」－「第二／四象限の球の数」



問題点：ケース B とケース C の違いを捉えられない。

Ⓟ 象限

第2象限　第1象限

第3象限　第4象限

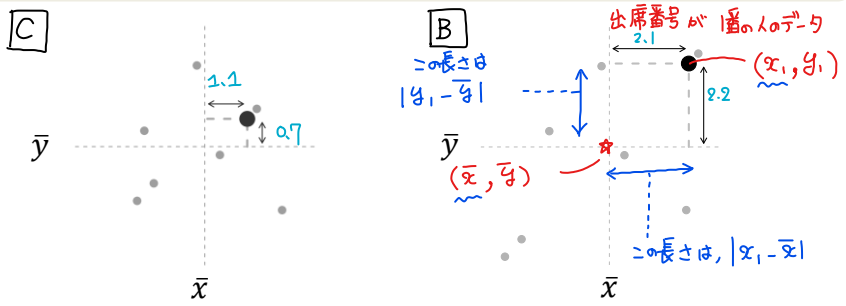
Ⓟ 第1案の良いところ

① 正負：関連の方向

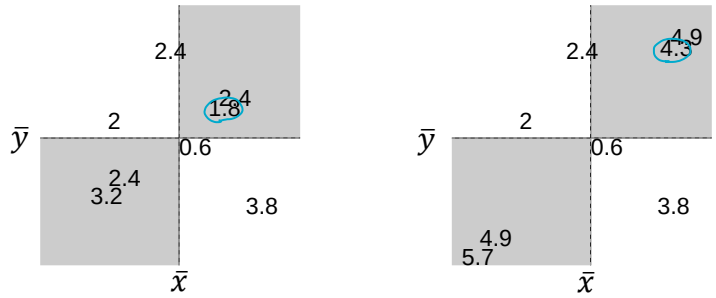
- 正：↗
- 0：↔
- 負：↘

② 値の大きさ：関連の強さ

（第二案）中心からの離れ具合を「足し算」で捉える

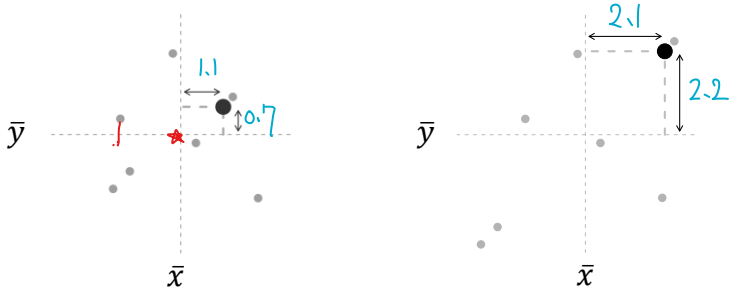


（式２）「灰色の領域の数字の和」－「白色の領域の数字の和」

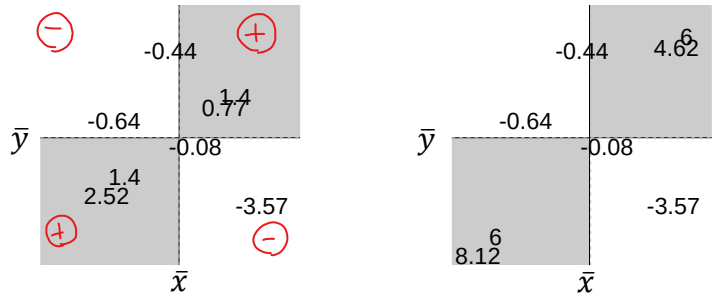


問題点：数式に場合分けが存在してしまう。

（第三案）中心からの離れ具合を「掛け算」で捉える



（式２）　「灰色の領域の数字の和」－「白色の領域の数字の和」



問題点：サンプルサイズに依存してしまう。

Ⓟ第2案の良いところ

① 正負：関連性の方向
② 値　：関連性の強さ
③ Bと Cを区別できる

Ⓟ数式で表現（第2案）

① 出席番号1の人の中心からの遠さ

$2.1 + 2.2$
$= |x_1-\bar{x}| + |y_1-\bar{y}|$

② 全ての人をまとめて，中心からの遠さ

$4.3 + 4.9 + 4.9 + 5.7 - 2.4 - 2 - 0.6 - 3.8 = 11.1$

$|x_1-\bar{x}| + |y_1-\bar{y}|$
$+ |x_2-\bar{x}| + |y_2-\bar{y}|$
$+ |x_5-\bar{x}| + |y_5-\bar{y}|$
$+ |x_8-\bar{x}| + |y_8-\bar{y}|$
$-(|x_3-\bar{x}| + |y_3-\bar{y}|)$
$-(|x_4-\bar{x}| + |y_4-\bar{y}|)$
$-(|x_6-\bar{x}| + |y_6-\bar{y}|)$
$-(|x_7-\bar{x}| + |y_7-\bar{y}|)$

Ⓟ数式で表現（第3案）

① 出席番号1の人

2.1×2.2
$= (x_1-\bar{x})(y_1-\bar{y})$

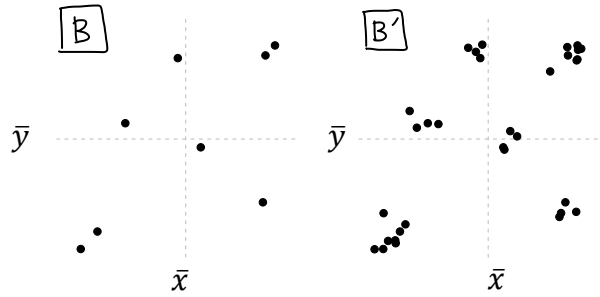
② 全ての人をまとめて

$6 + 4.62 + 6 + 8.12 + (-0.44) + (-0.64) + (-0.08) + (-3.57)$

$(x_1-\bar{x})(y_1-\bar{y})$
$+(x_2-\bar{x})(y_2-\bar{y})$
$+(x_5-\bar{x})(y_5-\bar{y})$
$+(x_8-\bar{x})(y_8-\bar{y})$
$+(x_3-\bar{x})(y_3-\bar{y})$
$+(x_4-\bar{x})(y_4-\bar{y})$
$+(x_7-\bar{x})(y_7-\bar{y})$
$+(x_8-\bar{x})(y_8-\bar{y})$

$= \sum_{i=1}^{8} (x_i-\bar{x})(y_i-\bar{y})$

（第四案）サンプルサイズの影響を無くす　採用！



（式３）　$\frac{1}{n}\times$（「奇数象限の数字の和」－「偶数象限の数字の和」）

共分散

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

準備：平行移動をする

第2案：中心からの遠さを考慮する

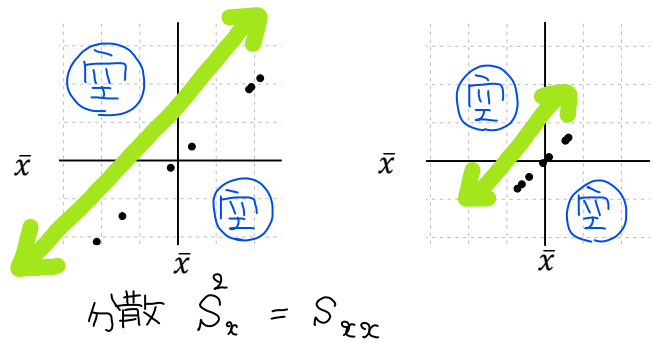
第3案：場合わけを回避

第4案：サンプルサイズの影響を取りのぞく

1つのデータではなくデータ全体を見よう

共分散の特殊なケースとしての分散

xとyという二変数の間ではなく、xとxという自分自身との間の共分散を考えてみたものが分散である。



分散　$S_x^2 = S_{xx}$

この値が大きいということは、その変数のばらつきが大きいということ。そこで、分散はばらつきの指標として使われる。