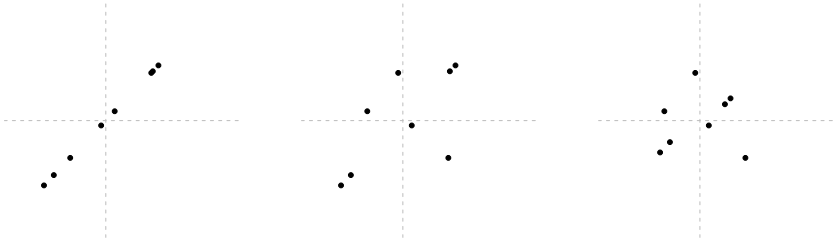
■　基本的な統計量 III（二つの変数の 相関）

【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。

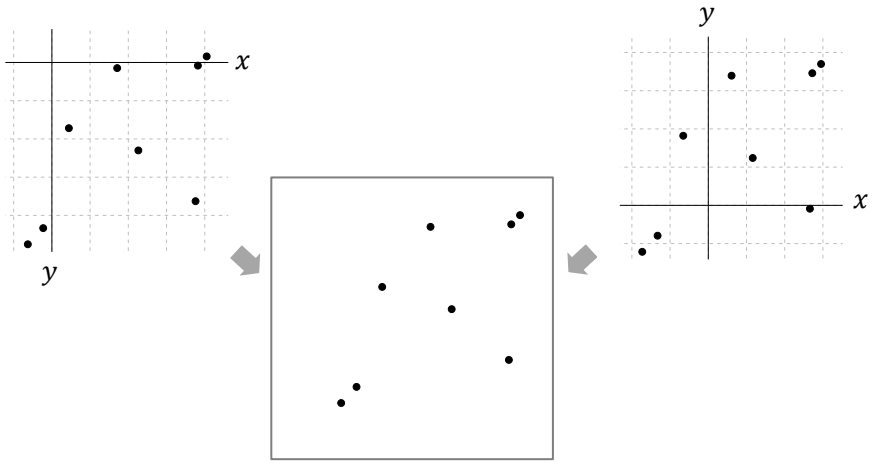
$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

※相関係数には最大値と最小値が存在する。

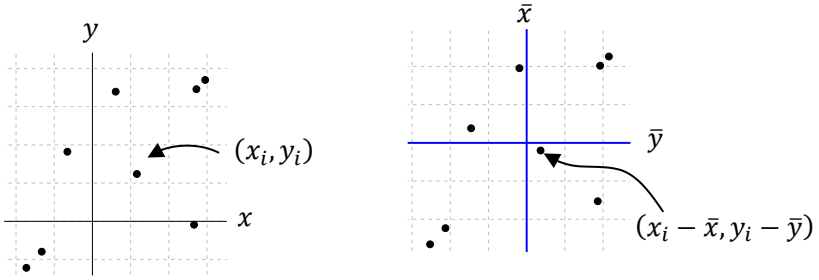
$$-1 \leq r_{xy} \leq 1$$

（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。

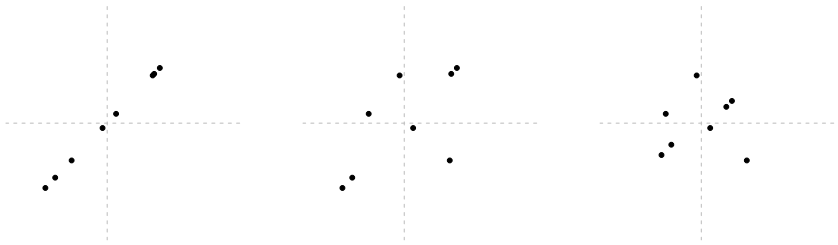


しかし、座標自体は維持したいので、新しく、横軸を$\bar{y}$、縦軸を$\bar{x}$の位置に据えた新しい座標を考える。これはもともとの座標を平行移動したことに相当する。



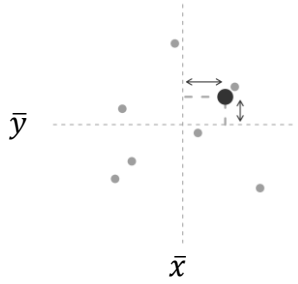
（第一案）球の個数の差を考える

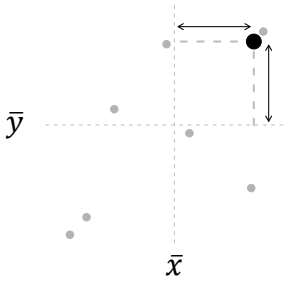
（式１）　「第一／三象限の球の数」－「第二／四象限の球の数」



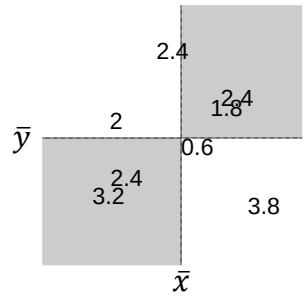
問題点：ケース B とケース C の違いを捉えられない。

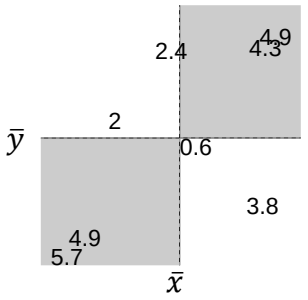
（第二案）中心からの離れ具合を「足し算」で捉える





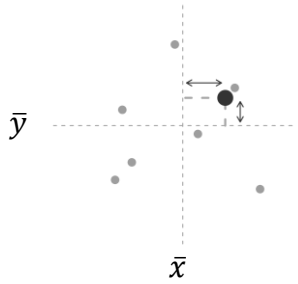
（式２）「灰色の領域の数字の和」－「白色の領域の数字の和」

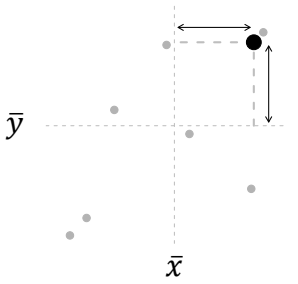




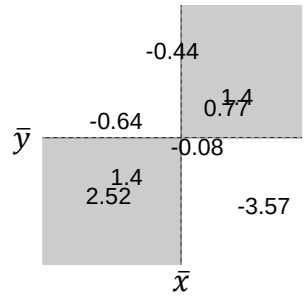
問題点：数式に場合分けが存在してしまう。

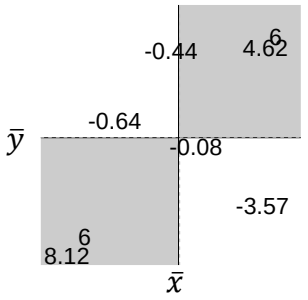
（第三案）中心からの離れ具合を「掛け算」で捉える





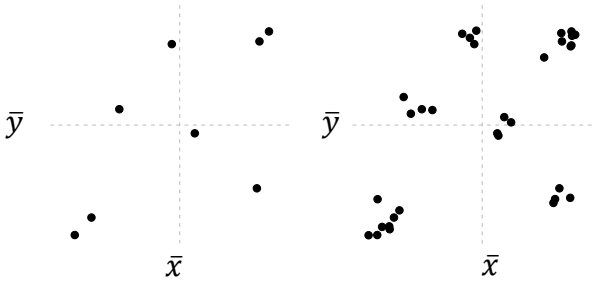
（式２）　「灰色の領域の数字の和」－「白色の領域の数字の和」





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！

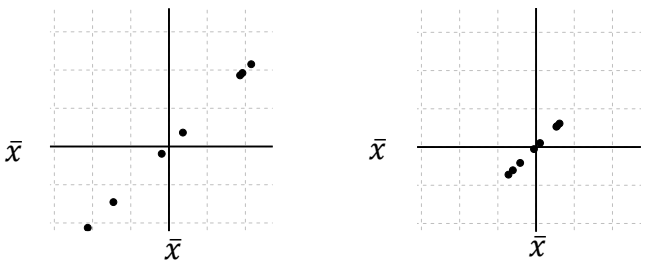


（式３）　$\frac{1}{n}\times$(「奇数象限の数字の和」－「偶数象限の数字の和」)

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

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。



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。