多変量解析

第5回　相関

萩原•篠田
情報理工学部

相関

- 3回生GPAと入試得点の関連性は？
- 3回生GPAと2回生GPAの関連性は？

keywords

相関係数、共分散・偏差積和（分散・偏差平方和）、内積

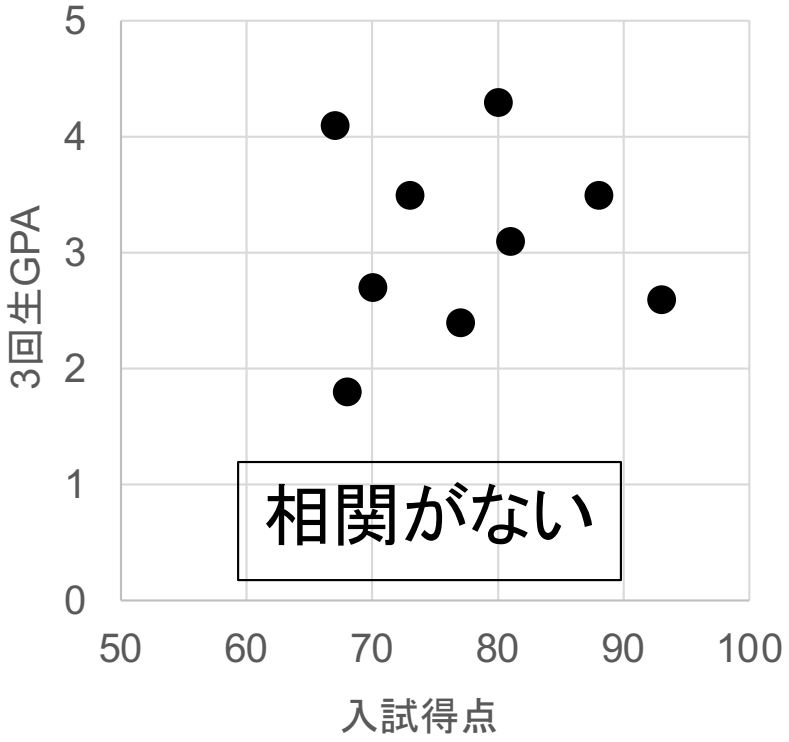
ID	3回生GPA y	入試得点 x_1	2回生GPA x_2	性別 x_3	出身高校 x_4
1	3.5	80	3.7	F	A高校
2	2.4	61	2.3	M	B高校
3	4.1	82	4.0	M	C高校
4	3.1	78	3.4	F	D高校
5	1.8	62	2.2	M	D高校
6	2.7	73	2.0	F	B高校
7	2.6	62	2.1	M	C高校
8	3.5	60	3.2	M	A高校
9	4.3	100	4.4	F	B高校

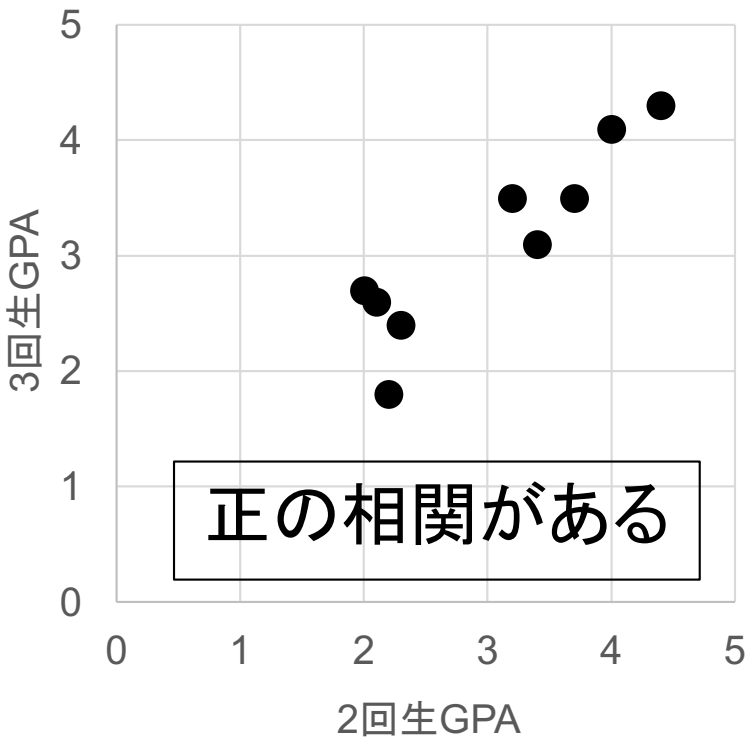
相関

- 3回生GPAと入試得点の関連性は？
- 3回生GPAと2回生GPAの関連性は？

keywords

相関係数、共分散・偏差積和（分散・偏差平方和）、内積





相関に入る前の復習

S_{xx} ； xの偏差平方和　$S_{xx} = \sum (x_i - \bar{x})^2$

S_{yy} ； yの偏差平方和　$S_{yy} = \sum (y_i - \bar{y})^2$

} $(n-1)$で割ると(標本)分散

$$s_{xx} = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$

分散の表記　s_{xx} , s_x^2 , V_x , $\mathrm{Var}(x)$

$$s_{yy} = \frac{1}{n-1}\sum (y_i - \bar{y})^2$$

分散の表記　s_{yy} , s_y^2 , V_y , $\mathrm{Var}(y)$

S_{xy} ； x, yの偏差積和　$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$

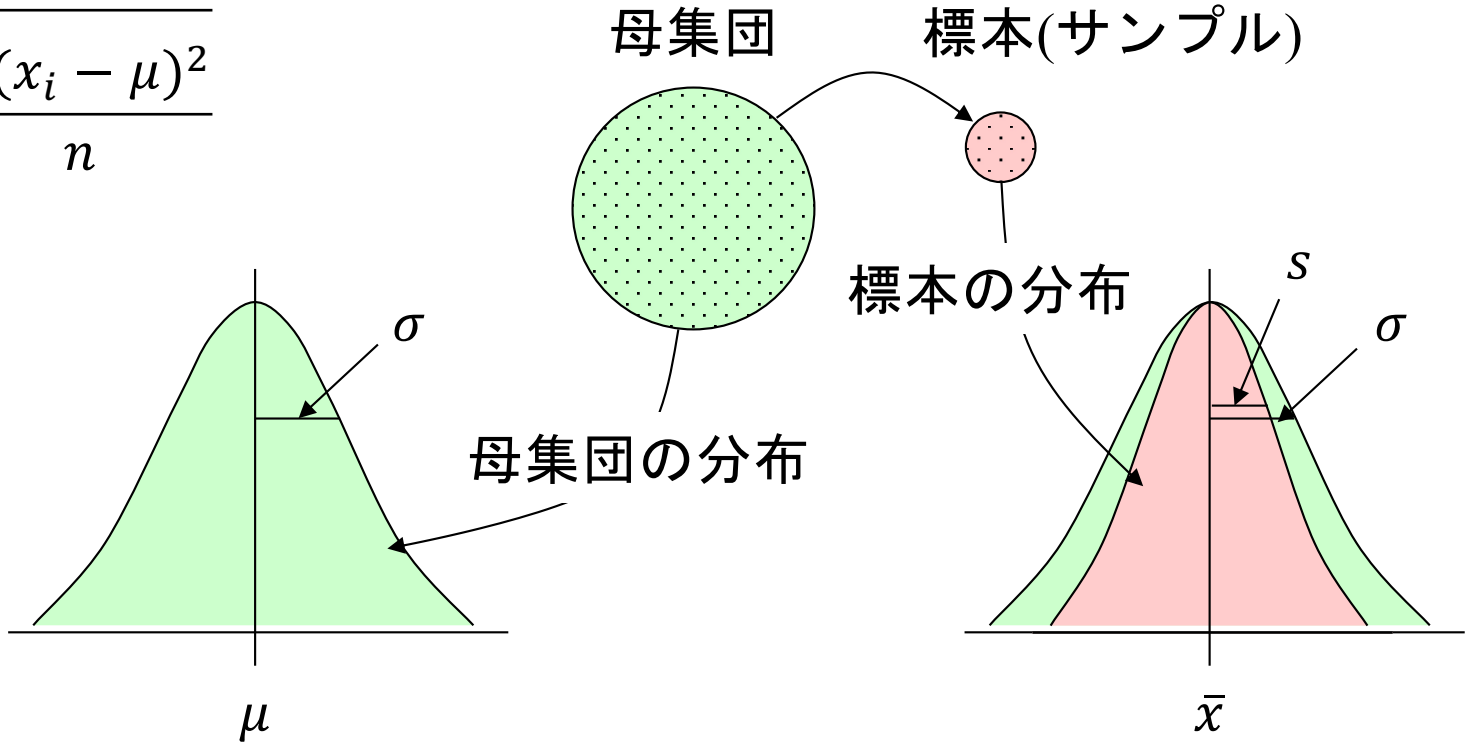
⟶ $(n-1)$で割ると(標本)共分散

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

共分散の表記　s_{xy} , C_{xy} , $\mathrm{Cov}(x,y)$

標準偏差；個々の点 x が母集団の平均値 μ から平均的にどの程度隔たっているかを示す

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$



$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

(標本)標準偏差 s は σ より小さく計算される。
σ の偏りのない推定値を得るには
偏差平方和 $\sum(x_i - \bar{x})^2$ を n ではなく $n-1$ で割ると良い
(数学的理論より証明されている)

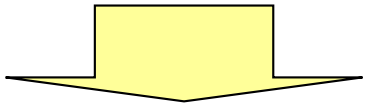
$$s^2 = s_{xx} = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

標本分散；s を2乗した値

相関分析

量的な（順序尺度を含む）2変数間の関係を分析する際に用いる。

2つの変量(x , y)の関係について、x , y ともに正規分布にしたがってばらつく量であるときには両者の関係を相関分析する。



相関分析では両変数間の関連の度合いを相関係数で評価することを主な目的とする。

相関係数

相関係数とは2変量 x , y の間の直線関係の強さを表す指標

x, yの偏差積和

$$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \cdot \sum(y_i-\bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{s_{xy}}{s_x s_y}$$

s_{xy}：共分散

s_x, s_y：標準偏差

xの偏差平方和　yの偏差平方和

r の範囲は、－1から1で、絶対値が1に近いほど点が直線的に配列していることになる。$r = 0$ で無相関となる。

相関係数の解釈

相関係数の絶対値	解釈
0.0～0.2	ほとんど相関関係がない
0.2～0.4	やや相関関係がある
0.4～0.7	かなり相関関係がある
0.7～1.0	強い相関関係がある

相関係数を理解するためには共分散を理解する必要がある

変量の関係を与える共分散

- 変量間の関連指標である共分散

  変量の関連を調べられる指標の代表的なものとして、共分散がある。

No	x	y	偏差積
1	x_1	y_1	$(x_1-\bar{x})(y_1-\bar{y})$
2	x_2	y_2	$(x_2-\bar{x})(y_2-\bar{y})$
⋮	⋮	⋮	⋮
n	x_n	y_n	$(x_n-\bar{x})(y_n-\bar{y})$

s_{xy} ； 共分散

x, yの偏差積和S_{xy}を
$n-1$で割る

$$s_{xy} = \frac{S_{xy}}{n-1}$$

$$= \frac{1}{n-1}\sum(x_i-\bar{x})(y_i-\bar{y})$$

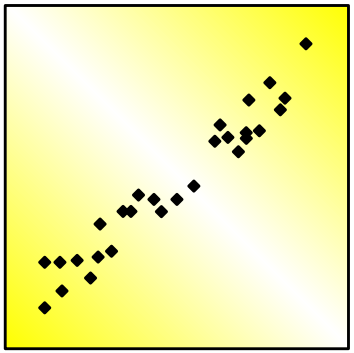
共分散と相関図

- 右上がりの分布
  $s_{xy} > 0$

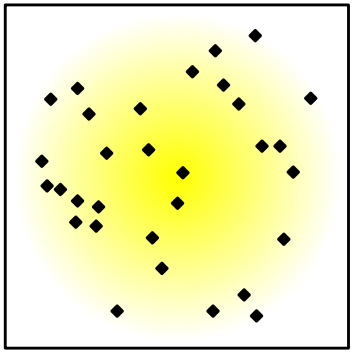
- 規則性がない分布
  $s_{xy} \fallingdotseq 0$

- 右下がりの分布
  $s_{xy} < 0$

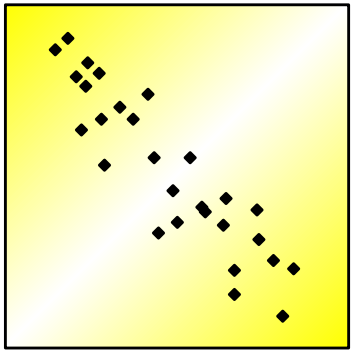
このような、分布を調べるグラフを
相関図いう。



正の相関



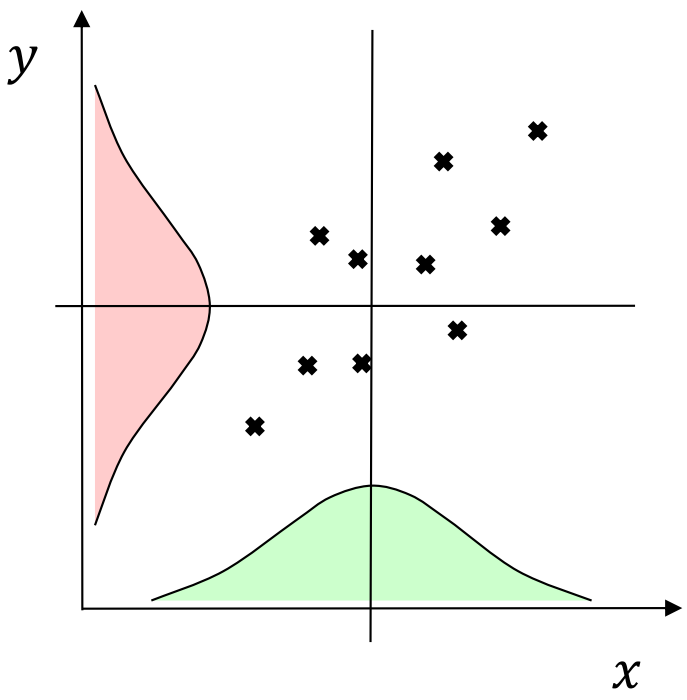
相関がない

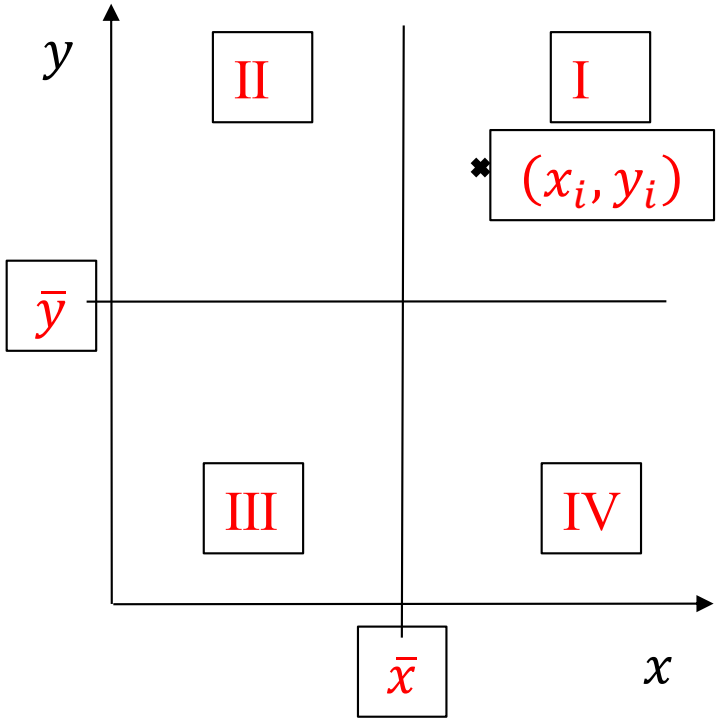


負の相関

共分散の性質

①偏差積の意味として、$(\bar{x}, \bar{y})$を中心にみて、点がその右上か左下の区画にあると、偏差積はプラスの値をとる。逆に点が左上か右下の区画にあると、偏差積はマイナスの値をとることがわかる。





偏差積

$(x_i - \bar{x})(y_i - \bar{y})$

I＞0

II＜0

III＞0

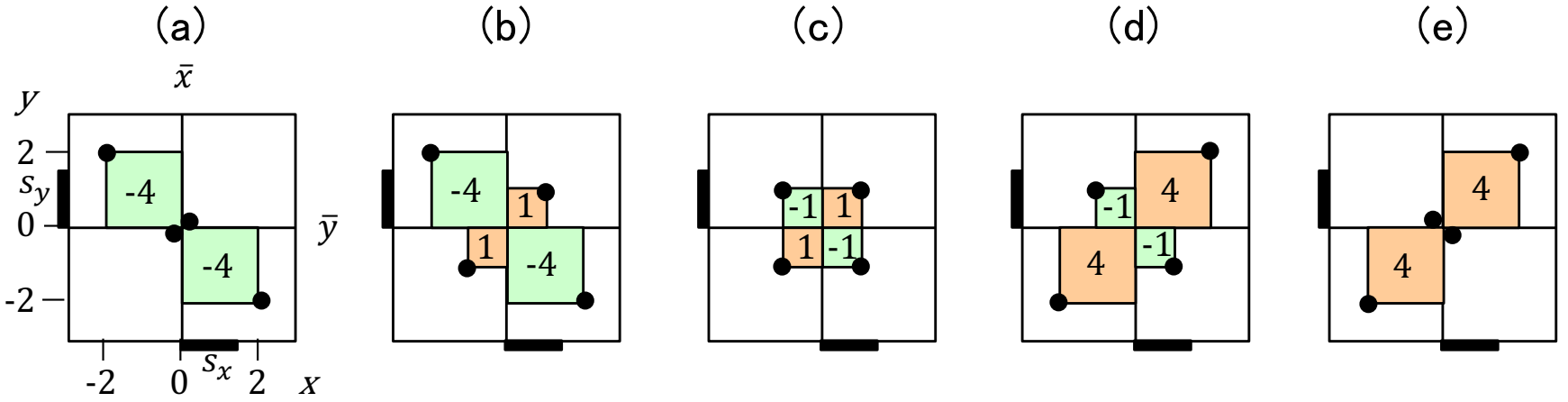
IV＜0

②偏差積和は、点が分布の中心$(\bar{x}, \bar{y})$の周りの4つの区画において

（a）、（b）：左上と右下に集まると負の値

（c）　　：均一に分布すると打ち消されて0

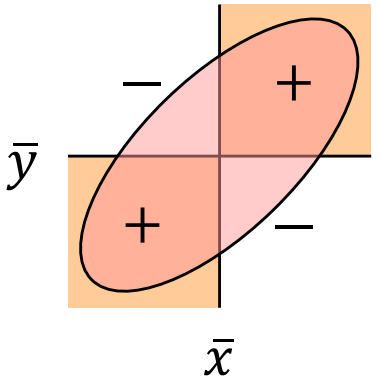
（d）、（e）：右上と左下に集まると正の値



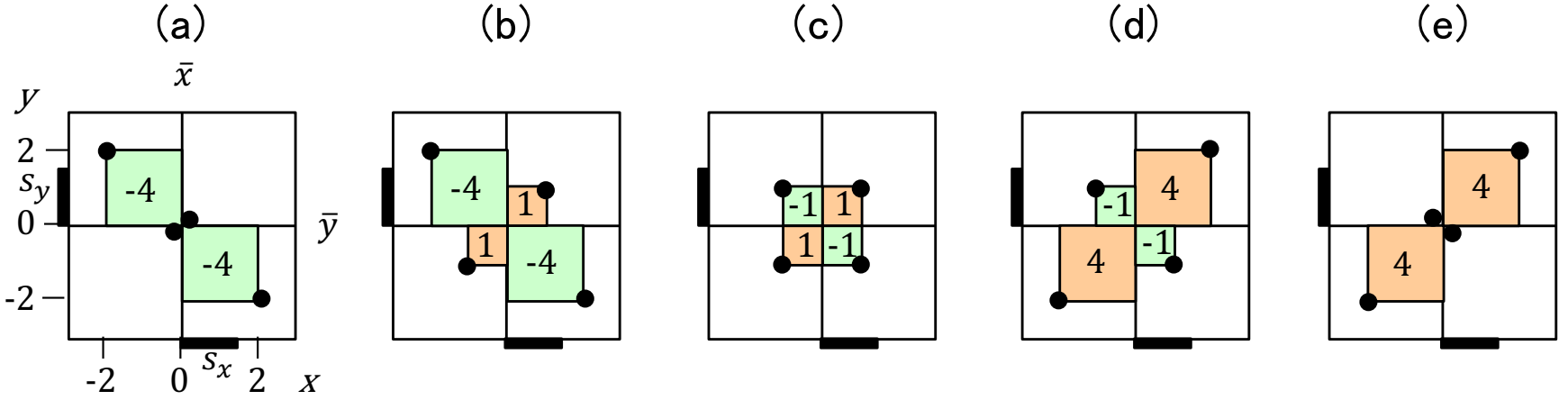
偏差積和　$S_{xy} = -8$　　$S_{xy} = -6$　　$S_{xy} = 0$　　$S_{xy}=6$　　$S_{xy} = 8$

偏差積和　$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$

$= -8$



③偏差積和や共分散は、分布の中心$(\bar{x}, \bar{y})$からみて、点が対角線に沿ってどの程度集中しているかを表しており、2変量の直線関係の強さを表すことがわかる。



	(a)	(b)	(c)	(d)	(e)
偏差積和	$S_{xy} = -8$	$S_{xy} = -6$	$S_{xy} = 0$	$S_{xy} = 6$	$S_{xy} = 8$
共分散	$s_{xy} = -8/3$	$s_{xy} = -6/3$	$s_{xy} = 0$	$s_{xy} = 6/3$	$s_{xy} = 8/3$
偏差平方和	$S_{xx} = S_{yy} = 8$	$S_{xx} = S_{yy} = 10$	$S_{xx} = S_{yy} = 4$	$S_{xx} = S_{yy} = 10$	$S_{xx} = S_{yy} = 8$
相関係数	$r = -1$	$r = -0.6$	$r = 0$	$r = 0.6$	$r = 1$

偏差積和

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = -8$$

共分散

$$s_{xy} = \frac{S_{xy}}{n-1} = -8/3$$

偏差平方和

$$S_{xx} = \sum (x_i - \bar{x})^2 = 8$$

相関係数

$$r = \frac{s_{xy}}{s_x s_y} = \frac{-8}{\sqrt{8 \cdot 8}} = -1$$

共分散と相関係数の関係

- 共分散の欠点； 値が変数の単位によって変わる。
- データを平均値が0、標準偏差が1となるよう標準化する。
- 標準化された共分散が相関係数。値は-1から+1の値をとる。

共分散の式を標準偏差で標準化すると

$$\frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{n-1}\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2 \cdot \frac{1}{n-1}\sum(y_i-\bar{y})^2}}$$

共分散を標準偏差で標準化

$$= \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \cdot \sum(y_i-\bar{y})^2}}$$

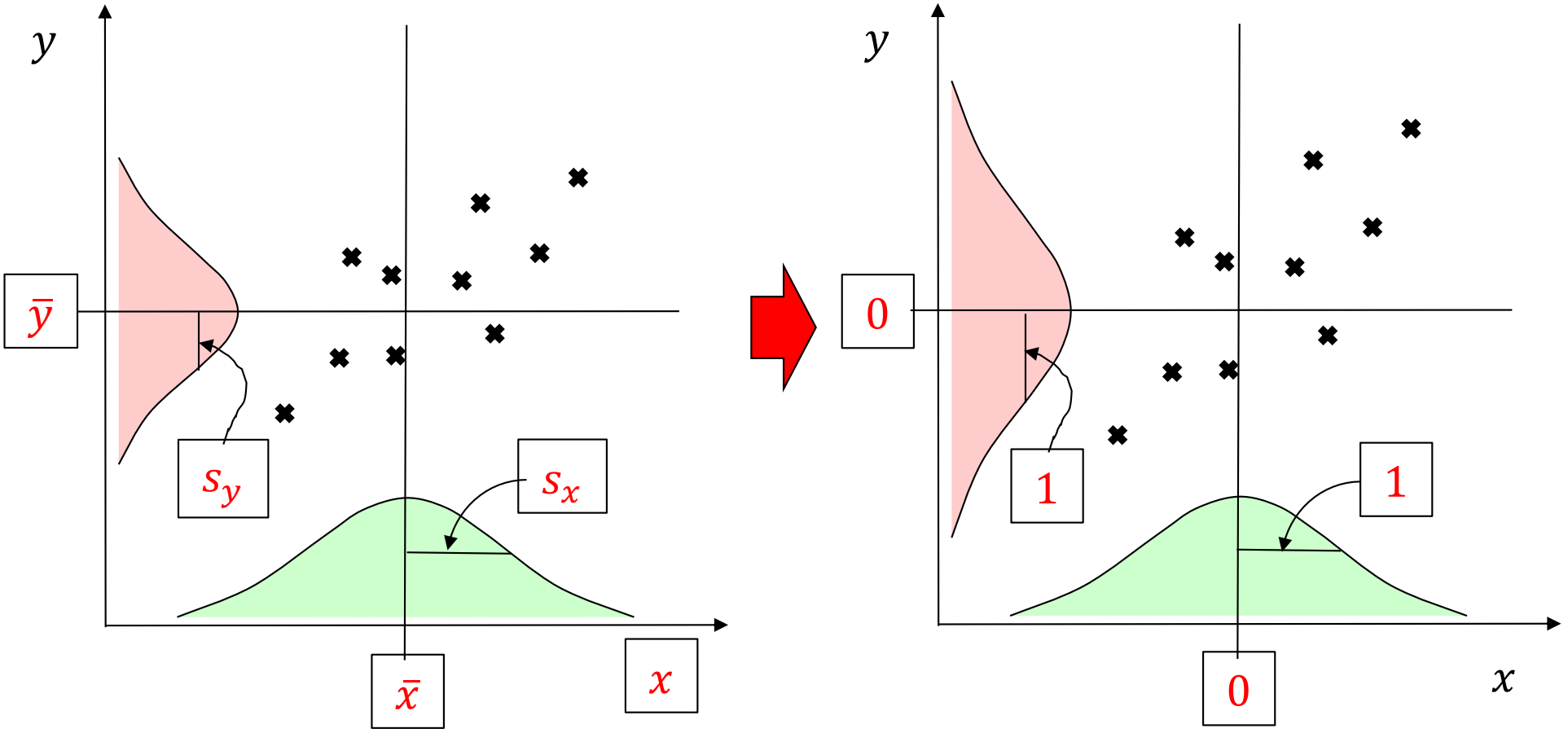
$$= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

$$= \mathrm{r}$$

前出の相関係数の式

相関係数の考え方

- データを平均値が0、標準偏差が1となるよう標準化。
- 標準化された共分散が相関係数。



まとめ

偏差積和

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

データ数が多くなると大きくなる傾向がある。
したがって、どの程度の直線化傾向があるか判断できない。

↓ (データ数)－1で割る

共分散

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

データの単位が小さいものは小さな値となり、データの単位が大きいものは大きい値となる。

↓ xの標準偏差とyの標準偏差の積で割る

相関係数

$$\frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n-1}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\Sigma(x_i - \bar{x})^2 \cdot \frac{1}{n-1}\Sigma(y_i - \bar{y})^2}}$$

$$= \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \cdot \Sigma(y_i - \bar{y})^2}}$$

$$= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

$$= \mathrm{r}$$

正の傾きの直線化傾向が強いほど＋1に近くなり、負の傾きの直線化傾向が強いほど－1に近づく。

相関係数のベクトル表現

ベクトルといえば列ベクトル（縦ベクトルを指す）
行ベクトル（横ベクトル）を表すときは、転置の記号(′)を用いる

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \boldsymbol{x}' = [x_1, x_2, \cdots, x_n]$$

2つのn×1ベクトル $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]'$ と $\boldsymbol{y} = [y_1, y_2, \cdots, y_n]'$ の内積を
次のように定義する

$$\boldsymbol{x}'\boldsymbol{y} = [x_1, x_2, \cdots, x_n] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = \sum_{i=1}^{n} x_i\, y_i$$

ベクトル $\boldsymbol{x}$ の長さは内積を用いて次のように表現できる

$$\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}'\boldsymbol{x}} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

ベクトル $\boldsymbol{x}$ と $\boldsymbol{y}$ を平均$\bar{x}$からの偏差ベクトルとして次のように定義する

$$\boldsymbol{x}=\begin{bmatrix} x_1-\bar{x} \\ x_2-\bar{x} \\ \vdots \\ x_n-\bar{x} \end{bmatrix} \qquad \boldsymbol{y}=\begin{bmatrix} y_1-\bar{y} \\ y_2-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{bmatrix}$$

このとき、「内積」とベクトルの「長さの2乗」を求めると次のようになる

$\boldsymbol{x}'\boldsymbol{y}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=S_{xy}$ （$\boldsymbol{x},\boldsymbol{y}$の偏差積和）

$\|\boldsymbol{x}\|^2=\sum_{i=1}^{n}(x_i-\bar{x})^2=S_{xx}$ （$\boldsymbol{x}$ の偏差平方和）

$\|\boldsymbol{y}\|^2=\sum_{i=1}^{n}(y_i-\bar{y})^2=S_{yy}$ （$\boldsymbol{y}$ の偏差平方和）

相関係数r_{xy}は

$$r_{xy}=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}=\frac{\boldsymbol{x}'\boldsymbol{y}}{\|\boldsymbol{x}\|\cdot\|\boldsymbol{y}\|}=\cos\theta$$ （θは$\boldsymbol{x}$と$\boldsymbol{y}$ のなす角）

と表現できる. n次元の高次元空間の偏差ベクトル間の角度と考えられる.

相関

①相関分析、相関係数とは何か
②共分散とは何か、またその性質
③共分散の欠点
④共分散と相関係数の関係
について説明せよ。