教養科目／B 自然の摂理の探求

2026 年度 統計学（集中１期）

第４回

前回の内容

・いろいろな代表値

a. 最頻値（モード）

b. 最小値，最大値

c. 範囲

d. 四分位数，四分位範囲

・はずれ値

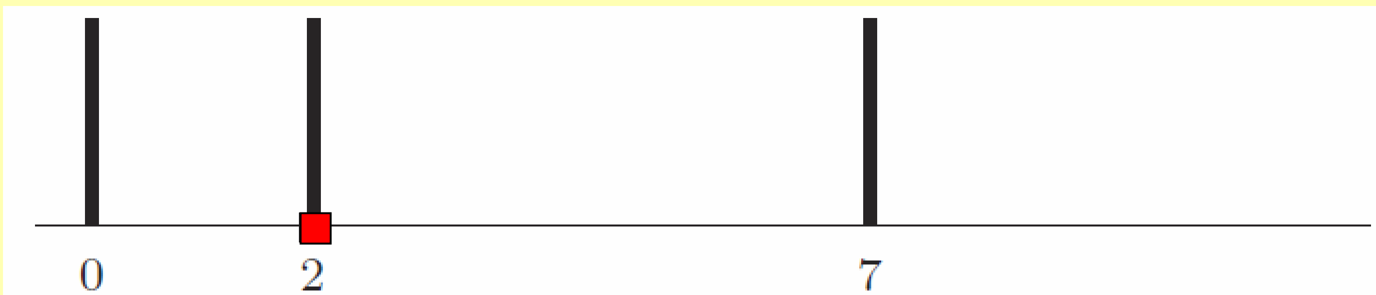
・箱ひげ図

中央値

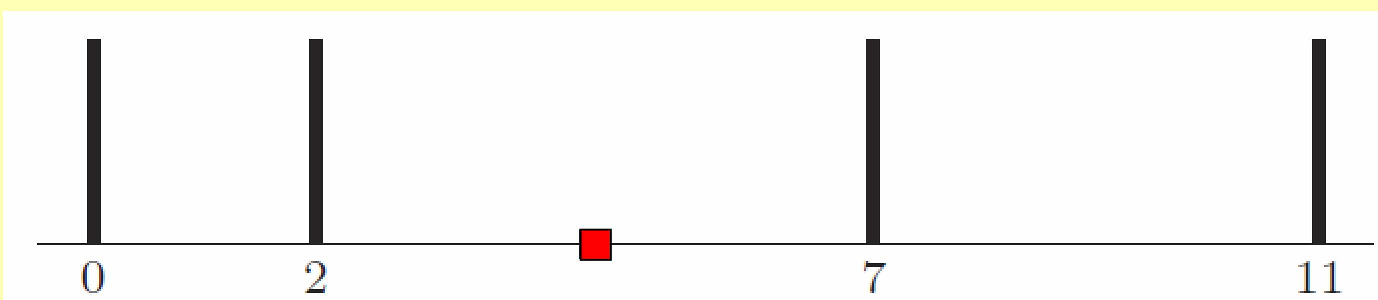
後者を**中央値またはメジアン**という．

■中央値

データの真ん中の位置にある値 (データの個数が奇数の場合)



$$\frac{\textbf{データの真ん中の値の２つの候補の和}}{2}$$ (データの個数が偶数の場合)



1．4．3．その他の代表値

$$\text{第１四分位数 } Q_1 = \begin{cases} \text{小さい方から 25\% に位置する値} \\ \qquad\qquad\qquad\qquad \text{(データの個数が 4 の倍数のとき)}, \\ \dfrac{\text{小さい方から 25\% に位置する値の２つの候補の和}}{2} \\ \qquad\qquad\qquad\qquad \text{(データの個数が 4 の倍数でないとき)}, \end{cases}$$

$$\text{第３四分位数 } Q_3 = \begin{cases} \text{大きい方から 25\% に位置する値} \\ \qquad\qquad\qquad\qquad \text{(データの個数が 4 の倍数のとき)}, \\ \dfrac{\text{大きい方から 25\% に位置する値の２つの候補の和}}{2} \\ \qquad\qquad\qquad\qquad \text{(データの個数が 4 の倍数でないとき)}, \end{cases}$$

$$\text{四分位範囲 } R_Q = \text{第３四分位数 } Q_3 - \text{第１四分位数 } Q_1.$$

1. 4. 5. 箱ひげ図

・箱ひげ図とは中央値，第１四分位数，第３四分位数，四分位範囲，はずれ値を図で表すことにより，データがもつ特徴をヒストグラムよりも簡潔に表す方法.

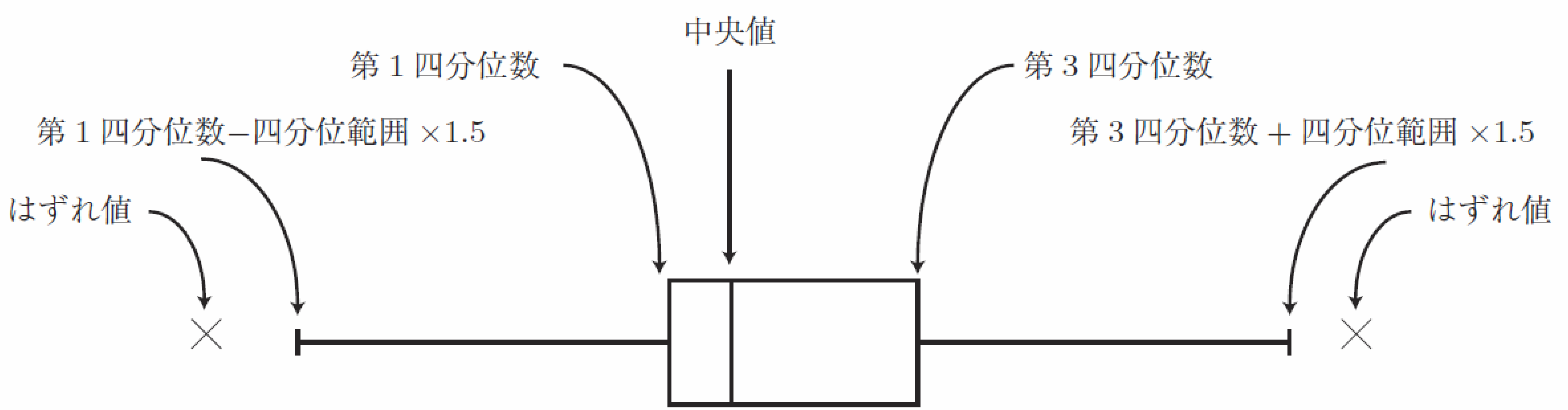


図 1.12　箱ひげ図のイメージ

１．５ ２次元データ

・２次元データ ： ２つのデータを対にしたもの.

例 1.10

表 1.11 はある地域の 7 日間の最高気温 x(℃) とその地域のあるアイスクリーム店の売り上げ額 y（千円）.

最高気温 (x)	26.1	20.0	23.2	25.1	25.3	27.2	28.1
売り上げ額 (y)	120	55	70	75	70	95	110

・「暑ければ暑いほど売り上げ額が多い.」とデータから読み取れるのか？

1．5　2次元データ

例 1.11

表　1.12 は国際教育到達度評価学会に参加している主な国・地域の中学２年生の数学と理科の成績である (2007 年国際数学・理科教育調査).

	台湾	韓国	シンガポール	香港	日本	ハンガリー	イングランド	露国
数学 (x)	598	597	593	572	570	517	513	512
理科 (y)	561	553	567	530	554	539	542	530

・数学が得意な人は理科も得意であり，また逆に理科が得意な人は数学も得意である．このことは本当か?

1．5　2次元データ

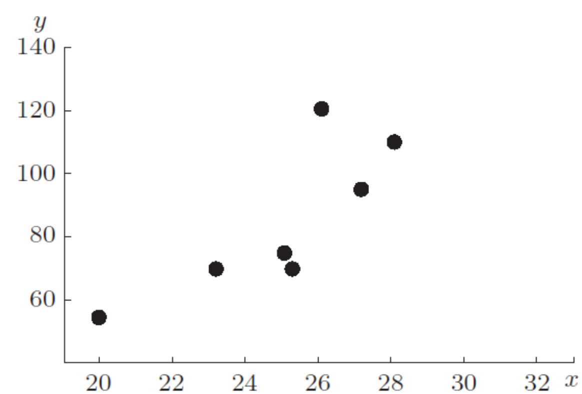
データそのものをみただけではデータがどのように集中していたり散らばっていたりしているかという特徴をつかむことは困難なので，データを整理することから始める．

最高気温 (x)	26.1	20.0	23.2	25.1	25.3	27.2	28.1
売り上げ額 (y)	120	55	70	75	70	95	110

最高気温 (x) と売り上げ額 (y) を対にした (x, y) は

$$(26.1, 120), \quad (20.0, 55), \quad \cdots, \quad (28.1, 110)$$

となる．これを xy 平面にプロットした図を与えると ⋯



右のような図を，**散布図**という．

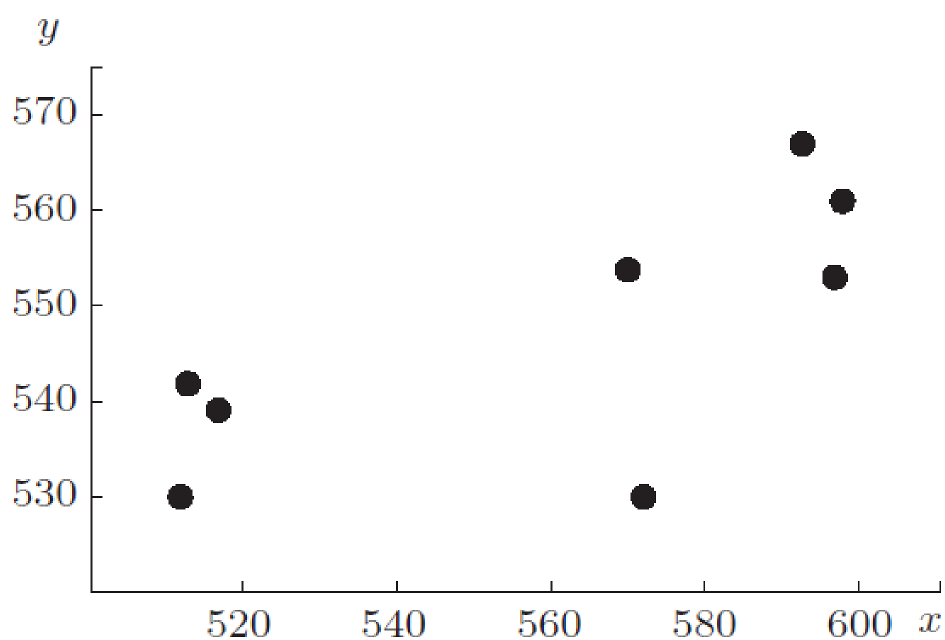
データの散らばりや集中を確認できる．

1．5　2次元データ

例 1.11

表　1.12 は国際教育到達度評価学会に参加している主な国・地域の中学２年生の数学と理科の成績である (2007 年国際数学・理科教育調査).

	台湾	韓国	シンガポール	香港	日本	ハンガリー	イングランド	露国
数学 (x)	598	597	593	572	570	517	513	512
理科 (y)	561	553	567	530	554	539	542	530



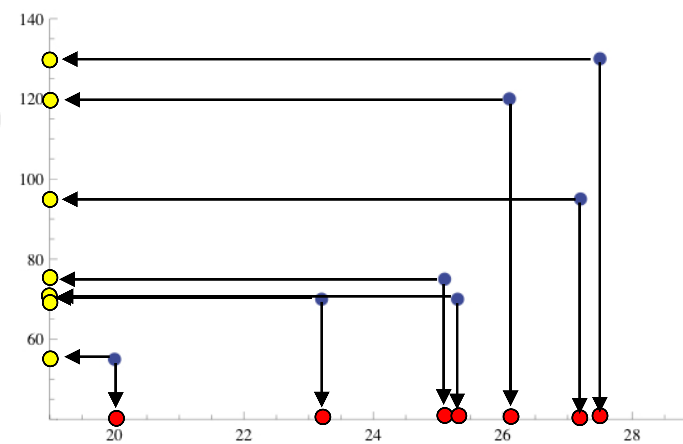
１．５　２次元データ

・２次元データの１つ目のデータだけ，または２つ目のデータだけを考えるとそれは１次元データである．つまり，それらの代表値が考えられる．

最高気温 (x)	26.1	20.0	23.2	25.1	25.3	27.2	27.5
売り上げ額 (y)	120	55	70	75	70	95	130

・最高気温の平均

$$\bar{x} = \frac{26.1+20.0+23.2+25.1+25.3+27.2+28.1}{7} = 25.0$$



・売り上げ額の平均

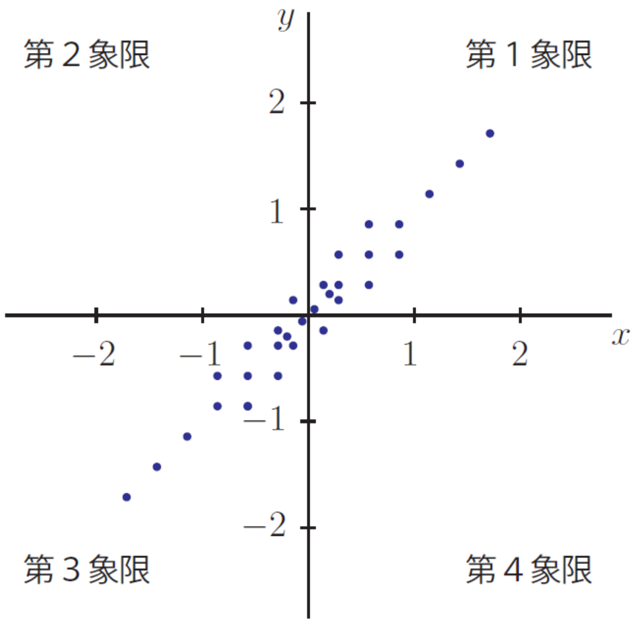
$$\bar{y} = \frac{120+55+70+75+70+95+110}{7} = 85$$

$$s_x^2 = \frac{1}{7} \times \{(26.1-25.0)^2 + \cdots + (28.1-25.0)^2\} = \frac{44}{7}$$

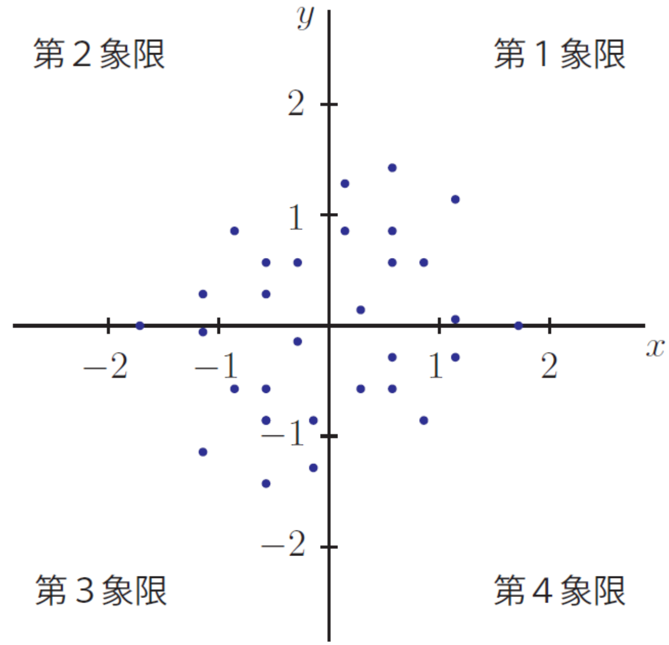
$$s_y^2 = \frac{1}{7} \times \left\{(120-85)^2 + \cdots + (110-85)^2\right\} = \frac{3400}{7}$$

1. 6 共分散

・x の平均も y の平均も 0 であるような 30 個のデータの散布図.



(i) 直線的な関係がある



(ii) 直線的な関係がない

・大雑把にいうと, (i) には直線的な関係があり, (ii) には直線的な関係がないといえる.

(直線的な関係とは, x の値が大きくなると y の値も大きくなり, その大きくなる様子が直線的に大きくなることを意味している.)

１．６　共分散

・共分散：直線的な関係の強さを表す代表値

$$\frac{x \times y \text{ の総和}}{2\text{ 次元データの個数}}$$

ここで,「$x \times y$ の総和」とは，２次元データの１つ目のデータと２つ目のデータの積をとり，それらの総和のことをいう.

・例

$$\text{データ}：(2,2), \quad (1,1), \quad (0,0), \quad (-1,-1), \quad (-2,-2)$$

⇩

$$x \times y \text{ の総和}：2 \times 2 + 1 \times 1 + 0 \times 0 + (-1) \times (-1) + (-2) \times (-2) = 10$$

⇩

$$\text{共分散}：\frac{10}{5} = 2$$

１．６　共分散

・共分散の定義

n 個のデータ：

$$(x_1, y_1),\quad (x_2, y_2),\quad \cdots,\quad (x_n, y_n)$$

に対して, x と y の共分散 c_{xy} は

$$c_{xy} = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

で定義される.

・**Note :** $x = y$ とすると標本分散となる.

1．7　相関係数

・共分散は単位に依存し，関係の度合いを比べるには適さない．

・例

・身長（cm）と体重（kg）の共分散は “cm×kg” の単位.

・例 1.10 のデータの共分散は “℃ × 千円” の単位.

⇓

2 組の 2 次元データの共分散を用いてどちらが関係の度合いが強いかを比較することはできない．

共分散を基準化し，単位に依存しないようにした値として相関係数を用いる．

1．7　相関係数の定義

相関係数は

$$x \text{ と } y \text{ の相関係数} = \frac{x \text{ と } y \text{ の共分散}}{(x \text{ の標本標準偏差}) \times (y \text{ の標本標準偏差})}$$

で与えられる.

式で書くと，x と y の相関係数 r_{xy} は

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

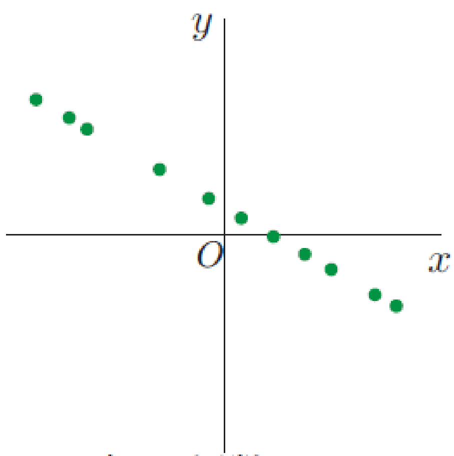
で与えられる.

１．７　相関係数の性質

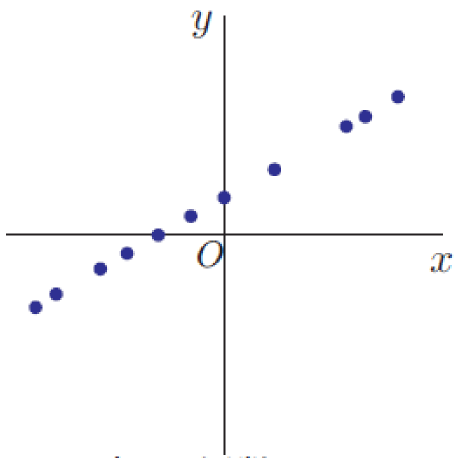
(a) 単位がない　（直線的な関係の強さを比較することができる）

(b) $-1 \leq$ 相関係数 ≤ 1

・２次元データ $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ が傾きが正の直線上にあるとき相関係数は 1 であり，傾きが負の直線上にあるときは相関係数は -1.



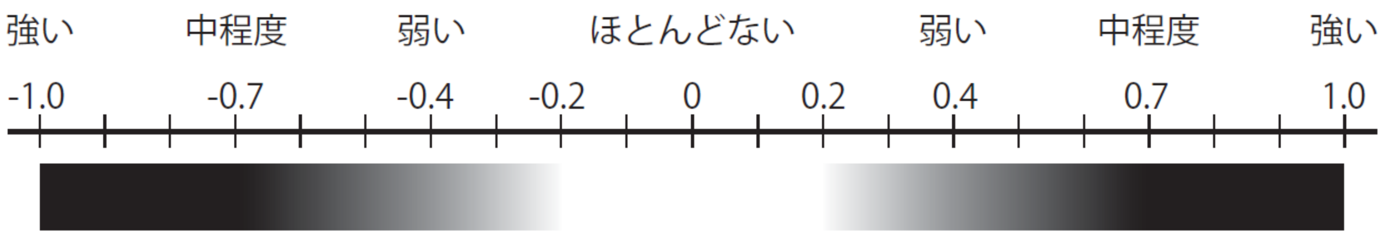
相関係数 =-1



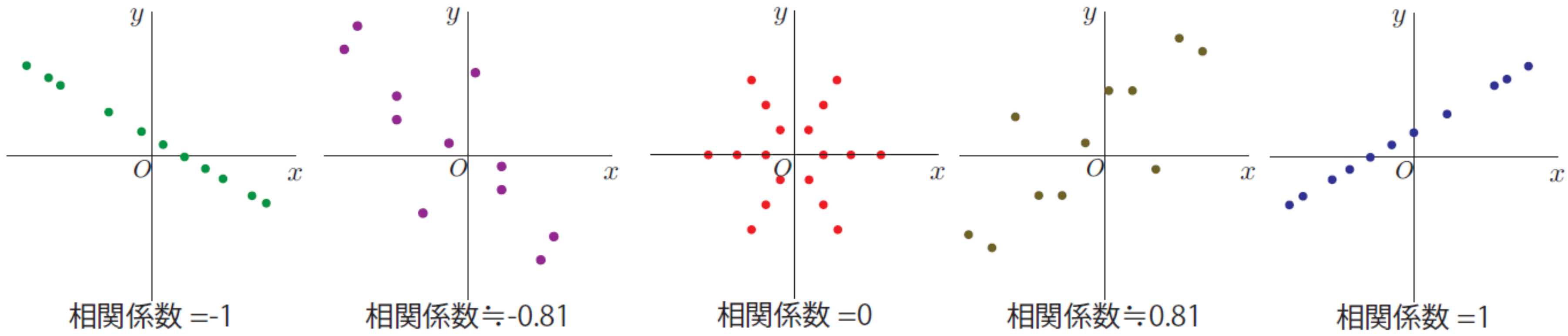
相関係数 =1

1．7　相関係数の性質

・相関係数の値と直線的な関係の強さを感覚的に表すと図 1.19 のようになる.



・図 1.20 は相関係数の値と散布図との関係のイメージである.



２次元データは相関係数が 1 に近い程傾きが正の直線の近くに位置し，相関係数が -1 に近い程傾きが負の直線の近くに位置する.

例 1. 11の相関係数

	台湾	韓国	シンガポール	香港	日本	ハンガリー	イングランド	露国
数学 (x)	598	597	593	572	570	517	513	512
理科 (y)	561	553	567	530	554	539	542	530

$$
\begin{aligned}
x \text{ の標本平均 } \bar{x} &= \frac{598 + 597 + \cdots + 512}{8} = 559, \\
y \text{ の標本平均 } \bar{y} &= \frac{561 + 553 + \cdots + 530}{8} = 547, \\
x \text{ の標本分散 } s_x^2 &= \frac{(598-559)^2 + (597-559)^2 + \cdots + (512-559)^2}{8} = 1312.5, \\
y \text{ の標本分散 } s_y^2 &= \frac{(561-547)^2 + (553-547)^2 + \cdots + (530-547)^2}{8} = 168.5, \\
x \text{ と } y \text{ の共分散 } c_{xy} &= \frac{1}{8}\{(598-559)\times(561-547) + (597-559)\times(553-547) \\
&\quad + \cdots + (512-559)\times(530-547)\} = 334.375 \fallingdotseq 334.4
\end{aligned}
$$

となるので,

$$
x \text{ と } y \text{ の相関係数 } r_{xy} = \frac{334.375}{\sqrt{1312.5}\times\sqrt{168.5}} \fallingdotseq 0.71.
$$

・直線的な関係の強さは「中程度」になる.

注意

曲線 $y = x^2$ 上にあるデータ：

$$(-2, 4),\ (-1, 1),\ (0, 0),\ (1, 1),\ (2, 4)$$

の相関係数を考える.

$$\bar{x} = 0,\ \bar{y} = 2,$$

$$s_x^2 = \frac{1}{5}\{(-2)^2 + (-1)^2 + 0 + 1^2 + 2^2\} = 2,$$

$$s_y^2 = \frac{1}{5}\{2 \times (4-2)^2 + 2 \times (1-2)^2 + (0-2)^2\} = 2.8,$$

$$\boxed{\begin{aligned} c_{xy} = \frac{1}{5}\{&(-2) \times (4-2) + (-1) \times (1-2) + 0 \times (0-2) \\ &+ 1 \times (1-2) + 2 \times (4-2)\} = 0 \end{aligned}}$$

よって, $r_{xy} = 0.$

- **相関係数が** 0 **であるからといって,** x **と** y **が無関係とはいえない.**

まとめ

・**2** 次元データにおける散布図，共分散，相関係数を扱った．

・相関係数は直線的な関係の強さを表している．

・相関係数の値が 0 に近い場合は，x と y の直線的な関係がほとんどないと考えるが，x と y は無関係という意味ではない．

付録

$$c_{xy} = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})}{n}$$

$$s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

$$s_y^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n}$$

$a_i = x_i - \bar{x}, \quad b_i = y_i - \bar{y}$ とすると

$$r_{xy} = \frac{a_1 b_1 + a_2 b_2 + \cdots + a_n b_n}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}\sqrt{b_1^2 + b_2^2 + \cdots + b_n^2}}$$