

Lecture 2

2 変量データの整理

2変量(2次元)データの例

新生児の身長, 体重のデータ

番号	身長	体重
1	46.0	2700
2	49.5	3220
3	50.0	3360
4	50.0	3500
5	49.0	3120
6	50.0	3160
7	53.0	4150
8	48.0	3310
9	49.0	2880
10	50.5	3090
11	49.5	3020
12	49.0	3360
13	50.0	3110
14	50.0	3560
15	47.5	2990

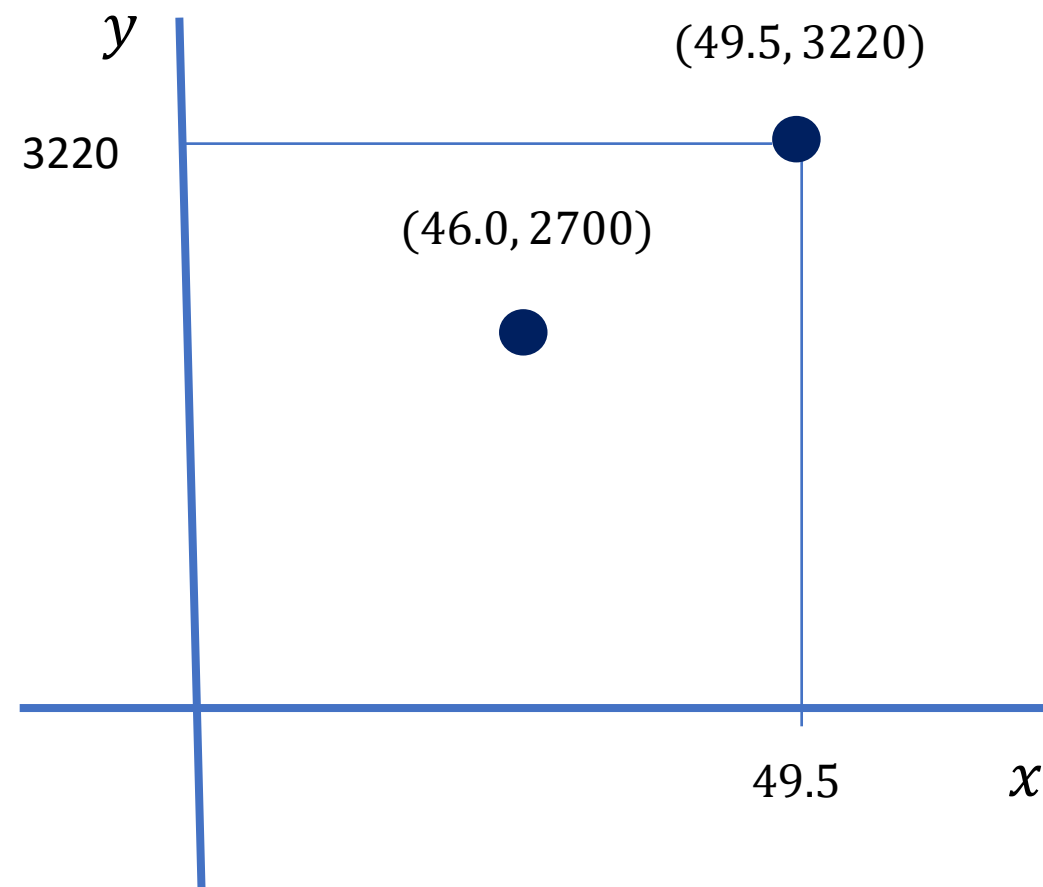
番号	身長	体重
16	50.5	3440
17	48.0	2920
18	49.0	3060
19	49.0	3360
20	50.0	3400
21	48.0	3200
22	50.5	2940
23	48.5	2850
24	50.5	3220
25	48.5	2750
26	49.0	3020
27	48.5	2570
28	48.5	3030
29	45.0	2410
30	51.0	3280

番号	身長	体重
31	50.5	3140
32	49.0	3040
33	52.0	3910
34	50.0	2770
35	46.5	2340
36	50.0	3140
37	50.5	3560
38	50.0	3390
39	50.0	3420
40	51.0	3450
41	49.5	3590
42	48.5	2830
43	48.0	3120
44	51.0	3190
45	50.0	3600

番号	身長	体重
46	47.0	2980
47	50.0	3090
48	51.0	3630
49	53.0	4060
50	50.0	3720
51	50.0	3400
52	50.5	3430
53	51.0	3250
54	48.0	2760
55	50.0	3320
56	49.0	2930
57	50.0	3320
58	48.0	2620
59	47.5	2860
60	48.0	2530

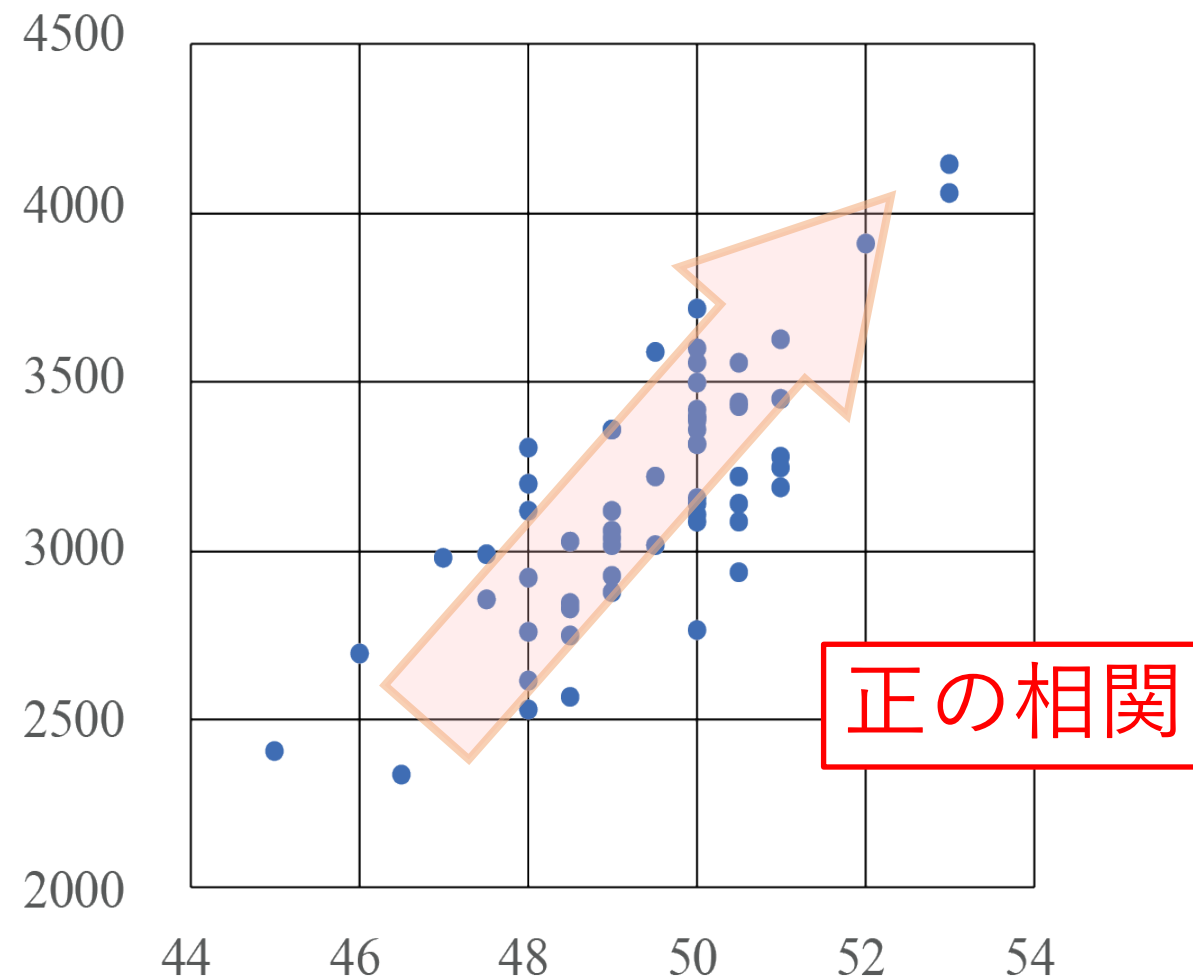
生データ $\Rightarrow$ 散布図(Scatter Plot)

番号	身長 x 座標	体重 y 座標
1	46.0	2700
2	49.5	3220
3	50.0	3360
$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i
$\vdots$	$\vdots$	$\vdots$
60	48.0	2530



生データ $\Rightarrow$ 散布図(Scatter Plot)

番号	身長 x 座標	体重 y 座標
1	46.0	2700
2	49.5	3220
3	50.0	3360
$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i
$\vdots$	$\vdots$	$\vdots$
60	48.0	2530



2変量の統計量

番号	身長 x	体重 y
1	46.0	2700
2	49.5	3220
3	50.0	3360
$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i
$\vdots$	$\vdots$	$\vdots$
60	48.0	2530

x の平均値と分散

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

n : データの大きさ

$$s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

y の平均値と分散

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$s_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

x と y の共分散

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

x と y の相関係数

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

共分散・相関係数の意味

共分散

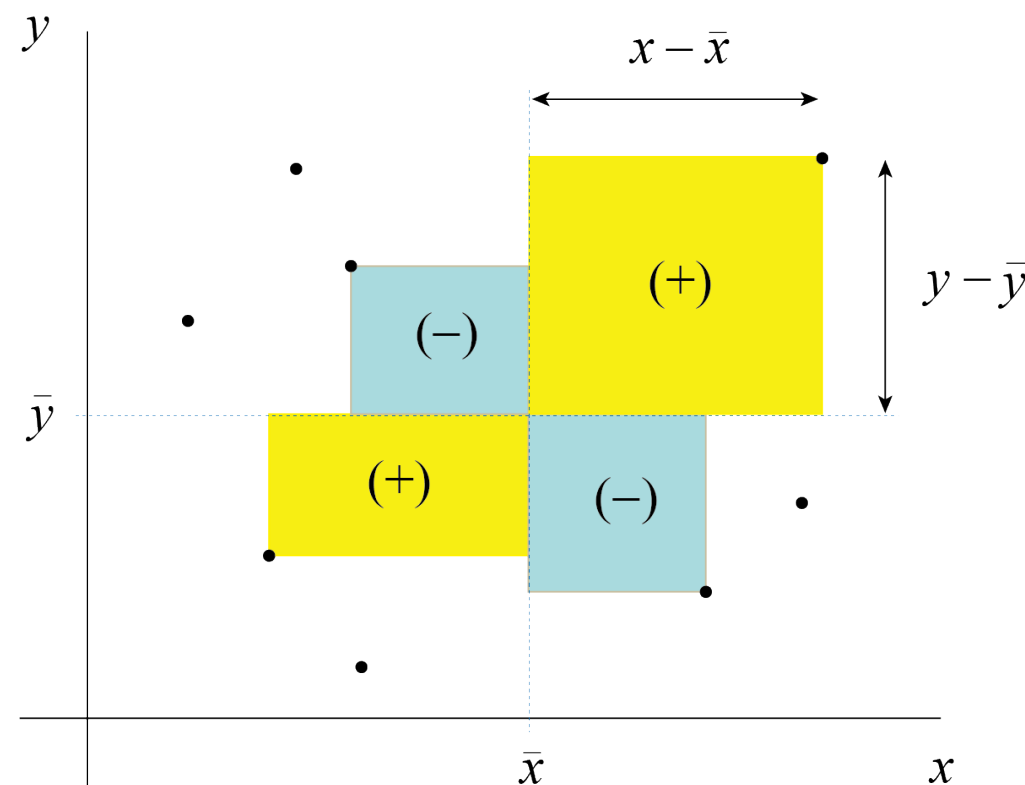
$$s_{xy} = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$s_{xy} > 0$: 概ね右上がりの傾向

$s_{xy} < 0$: 概ね右下がりの傾向

相関係数

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{1}{n} \sum \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$



重要

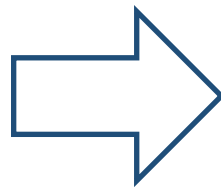
標準化された共分散といえる

変量の標準化（規準化）

生データ： $x_1, \dots, x_i, \dots, x_n$

平均値： $\bar{x}$

標準偏差： $s = s_x$



標準化または規準化（z 変換）

$$z_i = \frac{x_i - \bar{x}}{s}$$

定理 標準化された変数 $z_i = \frac{x_i - \bar{x}}{s}$ に対して

$$\text{平均値} = \bar{z} = 0$$

$$\text{標準偏差} = s_z = 1$$

証明 平均値の定義より

偏差の和はゼロ

$$\begin{aligned}\bar{z} &= \frac{1}{n} \sum z_i = \frac{1}{n} \sum \frac{x_i - \bar{x}}{s} = \frac{1}{s} \frac{1}{n} \sum (x_i - \bar{x}) = \frac{1}{s} \left(\frac{1}{n} \sum x_i - \frac{1}{n} \sum \bar{x} \right) \\ &= \frac{1}{s} (\bar{x} - \bar{x}) \\ &= 0\end{aligned}$$

定理 標準化された変数 $z_i = \frac{x_i - \bar{x}}{s}$ に対して

$$\text{平均値} = \bar{z} = 0$$

$$\text{標準偏差} = s_z = 1$$

証明 分散の定義より

$$\begin{aligned} s_z^2 &= \frac{1}{n} \sum (z_i - \bar{z})^2 = \frac{1}{n} \sum z_i^2 = \frac{1}{n} \sum \frac{(x_i - \bar{x})^2}{s^2} = \frac{1}{s^2} \times \frac{1}{n} \sum (x_i - \bar{x})^2 \\ &= \frac{1}{s^2} \times s^2 \\ &= 1 \end{aligned}$$

変量の標準化（規準化）

定理 標準化された変数

$$z_i = \frac{x_i - \bar{x}}{s}$$

に対して

$$\text{平均値} = \bar{z} = 0$$

$$\text{標準偏差} = s_z = 1$$

身長 x を cm で測定する

x_i の単位は cm

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{の単位は cm}$$

$$s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{の単位は cm}^2$$

s_x の単位は cm

標準化された変数 z

$$z_i = \frac{x_i^{\text{cm}} - \bar{x}^{\text{cm}}}{s^{\text{cm}}} \quad \text{は単位をもたない}$$

注目 標準化によって、測定単位が異なるデータの組を比較できる.

共分散・相関係数の意味（再）

共分散

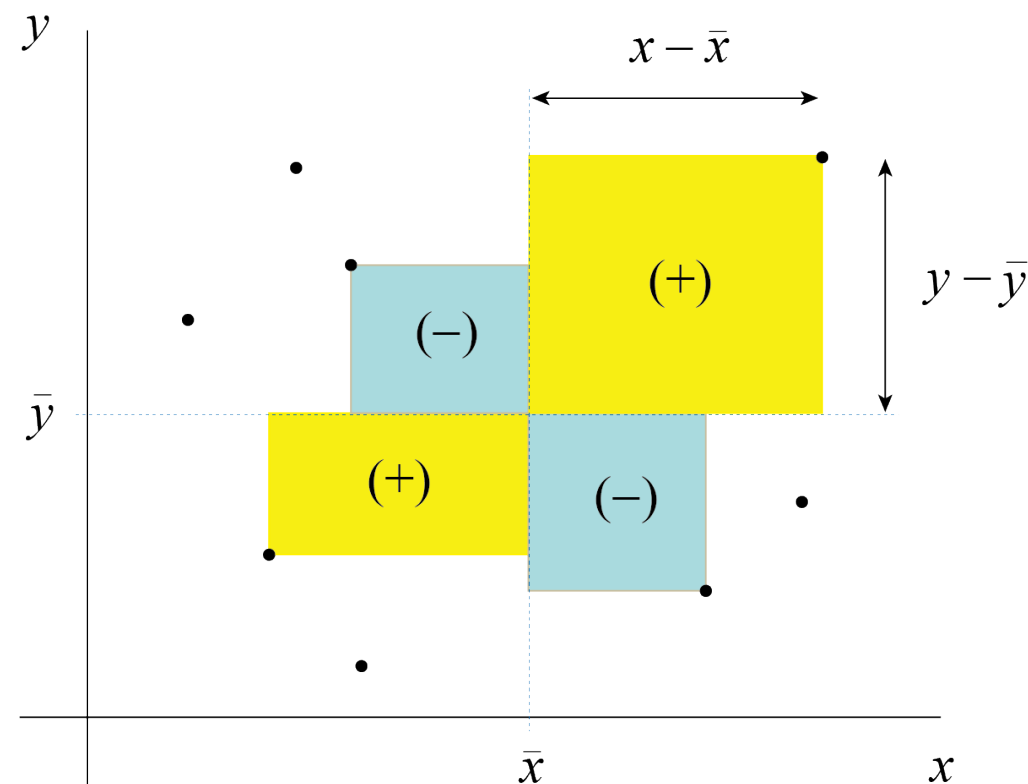
$$s_{xy} = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$s_{xy} > 0$: 概ね右上がりの傾向

$s_{xy} < 0$: 概ね右下がりの傾向

相関係数

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{1}{n} \sum \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$



納得!

標準化された共分散といえる

定理 (相関係数の基本性質)

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

証明 すべての実数 t に対して $\sum \{t(x_i - \bar{x}) + (y_i - \bar{y})\}^2 \geq 0$

$$t^2 \sum (x_i - \bar{x})^2 + 2t \sum (x_i - \bar{x})(y_i - \bar{y}) + \sum (y_i - \bar{y})^2 \geq 0$$

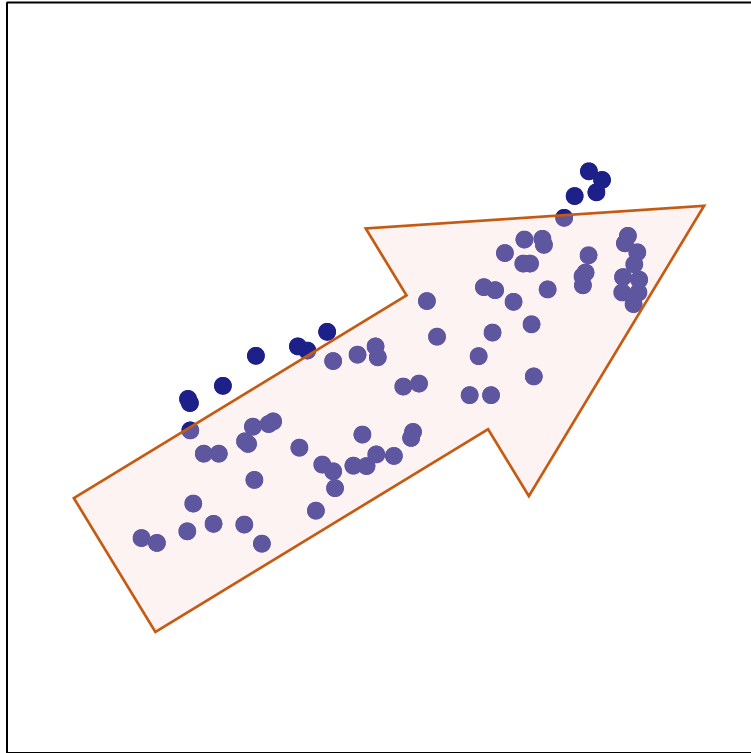
両辺を n で割って, $t^2 s_x^2 + 2t s_{xy} + s_y^2 \geq 0$

これがすべての実数 t に対して成り立つので, 判別式は $\frac{D}{4} = s_{xy}^2 - s_x^2 s_y^2 \leq 0$

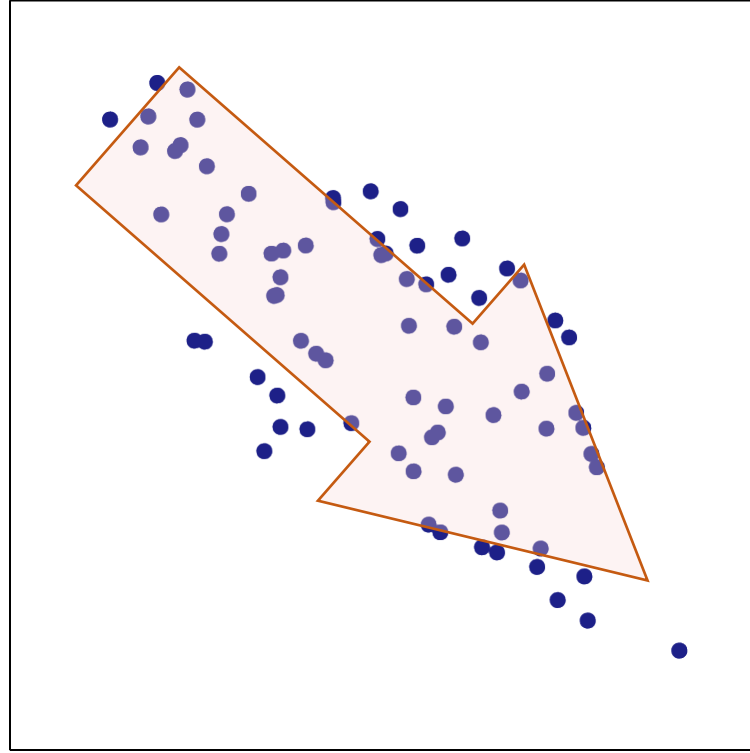
つまり, $s_{xy}^2 \leq s_x^2 s_y^2$

相関係数は $r_{xy} = \frac{s_{xy}}{s_x s_y}$ なので $r_{xy}^2 \leq 1$ が得られ, 証明が終わる.

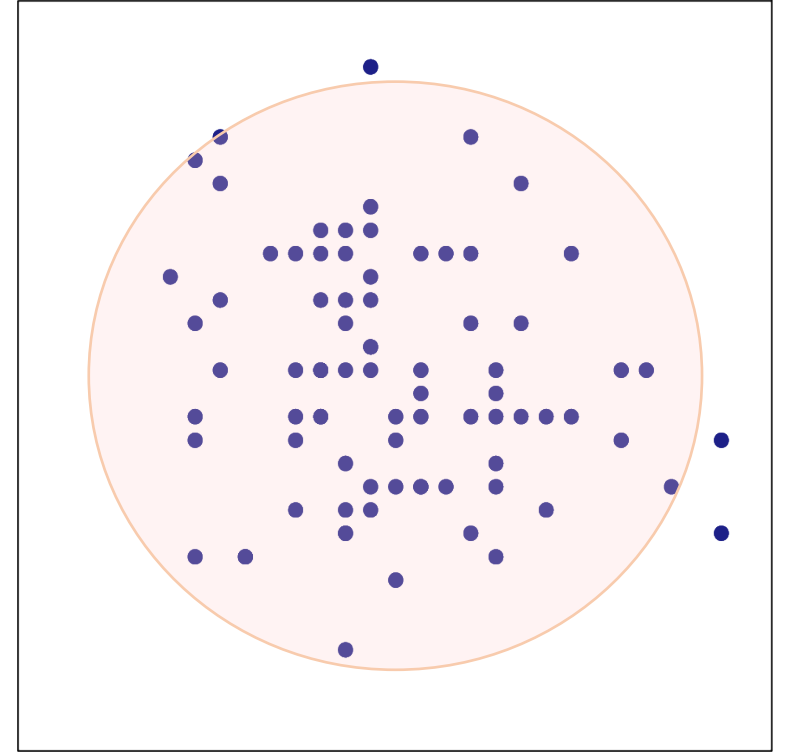
正の相関, 負の相関, 無相関



$r_{xy} > 0$: 正の相関

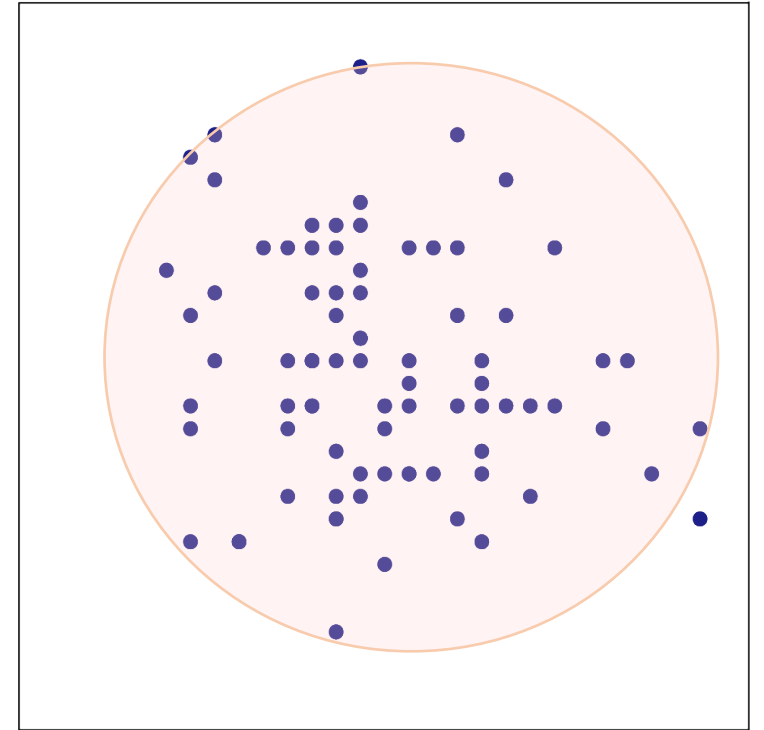
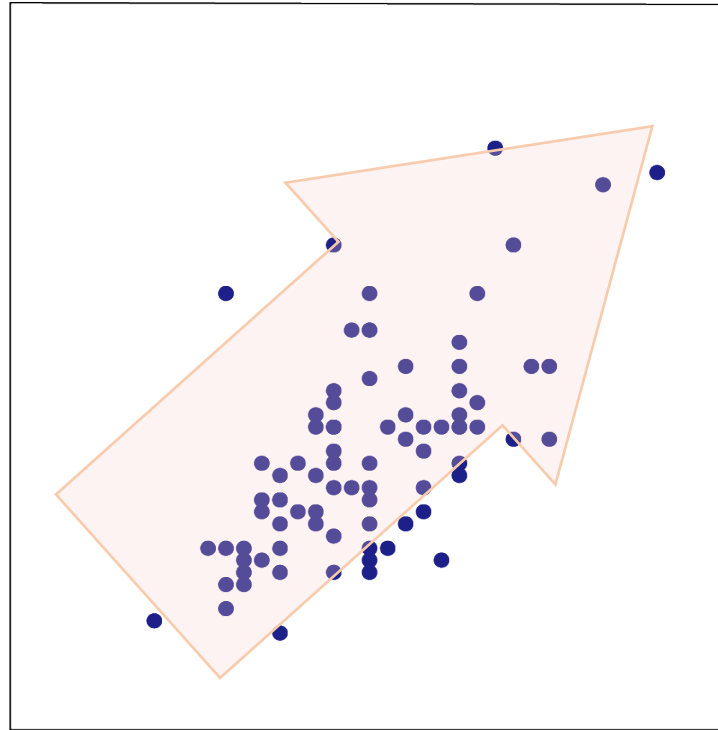
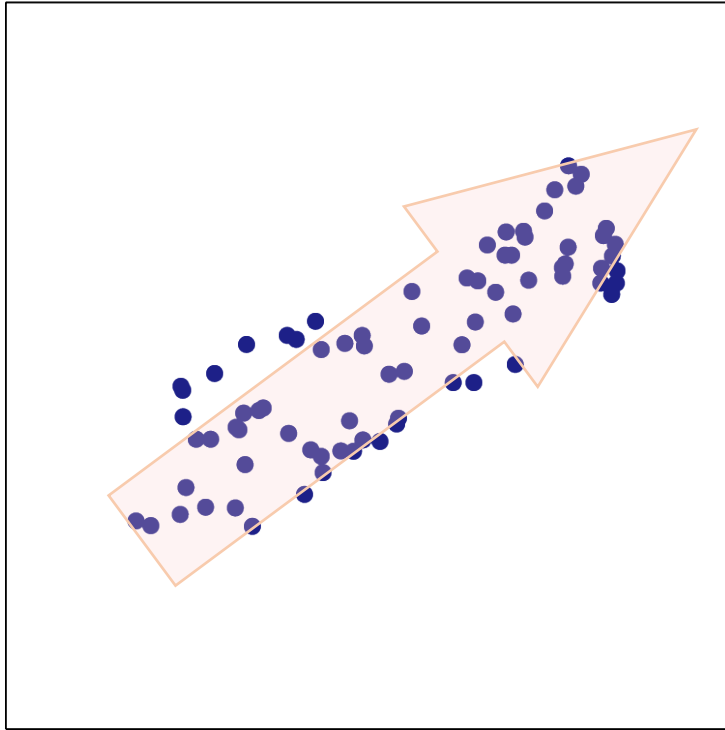


$r_{xy} < 0$: 負の相関



$|r_{xy}| \approx 0$: 無相関

強い相関, 弱い相関



$$|r_{xy}|$$

1

0

強い相関

弱い相関

無相関

定理（共分散の公式）

共分散は積の平均値から平均値の積を引く

$$s_{xy} = \overline{xy} - \bar{x} \bar{y}$$

分散公式との関係

$$s_{xx} = \overline{xx} - \bar{x} \bar{x} = \overline{x^2} - \bar{x}^2 = s_x^2$$

証明 共分散の定義より

$$\begin{aligned} s_{xy} &= \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x} \bar{y}) \\ &= \frac{1}{n} \sum x_i y_i - \bar{x} \frac{1}{n} \sum y_i - \bar{y} \frac{1}{n} \sum x_i + \frac{1}{n} \sum \bar{x} \bar{y} \\ &= \frac{1}{n} \sum x_i y_i - \bar{x} \bar{y} - \bar{y} \bar{x} + \bar{x} \bar{y} \\ &= \overline{xy} - \bar{x} \bar{y} \end{aligned}$$

例題 2.1

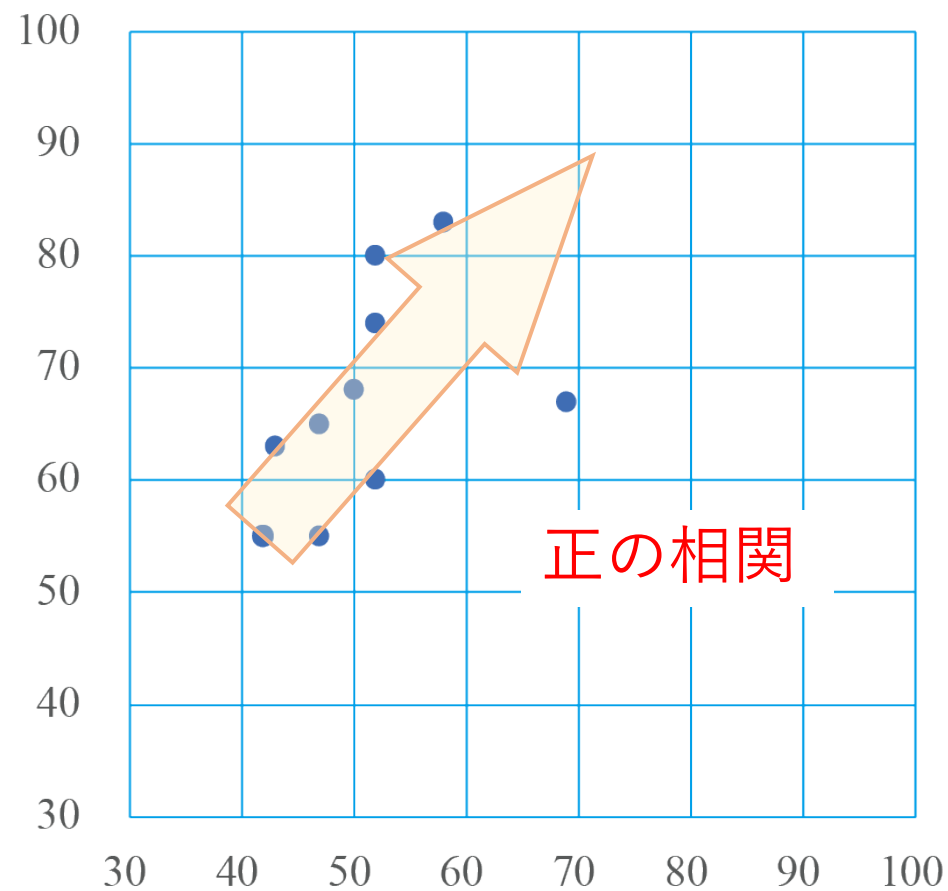
受講生10名の間接試験と期末試験の結果は次の通りであった。中間試験と期末試験の得点について相関はあるか？

中間試験 (x)	50	58	52	52	43	47	52	69	47	42
期末試験 (y)	68	83	74	80	63	55	60	67	65	55

StatData02_2.csv

中間試験 (x)	50	58	52	52	43	47	52	69	47	42
期末試験 (y)	68	83	74	80	63	55	60	67	65	55

散布図



統計量の計算

中間試験 (x)	50	58	52	52	43	47	52	69	47	42
期末試験 (y)	68	83	74	80	63	55	60	67	65	55

統計量の計算

統計量の計算											合計
中間試験 (x)	50	58	52	52	43	47	52	69	47	42	512
期末試験 (y)	68	83	74	80	63	55	60	67	65	55	670
x^2	2500	3364	2704	2704	1849	2209	2704	4761	2209	1764	26768
y^2	4624	6889	5476	6400	3969	3025	3600	4489	4225	3025	45722
xy	3400	4814	3848	4160	2709	2585	3120	4623	3055	2310	34624

$$\bar{x} = \frac{1}{n} \sum x_j = 51.2$$

$$\bar{y} = \frac{1}{n} \sum y_j = 67.0$$

$$\overline{xy} = \frac{1}{n} \sum x_j y_j = 3462.4$$

$$\overline{x^2} = \frac{1}{n} \sum x_j^2 = 2676.8$$

$$\overline{y^2} = \frac{1}{n} \sum y_j^2 = 4572.2$$

$$s_{xy} = \overline{xy} - \bar{x} \cdot \bar{y}$$

$$= 3462.4 - 51.2 \times 67.0 = 32.0$$

$$s_x^2 = \overline{x^2} - \bar{x}^2 = 55.36$$

$$s_y^2 = \overline{y^2} - \bar{y}^2 = 83.2$$

$$s_x = 7.44$$

$$s_y = 9.12$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = 0.47$$

Python を試してみる

StatData02_1.csv

番号	身長 (x)	体重 (y)
1	46.0	2700
2	49.5	3220
3	50.0	3360
$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i
$\vdots$	$\vdots$	$\vdots$
60	48.0	2530

[StatData02_1a - Jupyter Notebook.pdf](#)

StatData02_2.csv

中間試験 (x)	50	58	52	52	43	47	52	69	47	42
期末試験 (y)	68	83	74	80	63	55	60	67	65	55

[StatData02_2a - Jupyter Notebook.pdf](#)