

第3回 2変量データの整理 (3.1–3.3.5)

村澤 康友

2024 年 9 月 27 日

今日のポイント

1. 2 変量データの度数分布表を分割表という。量的な 2 変量データは散布図に表せる。
2. 2 つの変量の関係の強さは（積率）相関係数、2 つの順位の関係の強さは順位相関係数で表す。
3. 相関は必ずしも因果関係を意味しない。因果関係のない相関を見かけ上の相関という。

- 3.3.6 節の時系列データは扱わない（「数理統計学」と「時系列解析」の方法論は異なる）。
- 3.4 節の回帰分析は 13 章で扱うので今回は飛ばす。

目次

1	散布図 (p. 43)	
2	分割表 (p. 45)	
3	相関係数 (p. 47)	
3.1	共分散 (p. 49)	1
3.2	標準化 (p. 39)	2
3.3	（積率）相関係数 (p. 48)	2
3.4	順位相関係数 (p. 54)	3
3.5	相関と因果 (p. 50)	4
4	今日のキーワード	4
5	次回までの準備	4

1 散布図 (p. 43)

定義 1. 2 変量データを xy 平面上の座標で表した図を**散布図**という。

注 1. 量的変量に用いる。

注 2. 散布図から 2 変量の関係（相関関係）が読み取れる（図 1）。

例 1. 某大学 1 年生の英語と数学の入試成績（図 2）。

2 分割表 (p. 45)

定義 2. 2 変量データの度数分布表を**分割（クロス）表**という。

注 3. 相対度数は縦比・横比でみることもできる。

例 2. 東大（学部・院）の学生構成（表 1）。

定義 3. （該当数）／（非該当数）を**オッズ**という。

注 4. 該当の割合を p とすると $p/(1-p)$ 。

例 3. 検査の陽性／陰性のオッズ。

定義 4. 2 群のオッズの比を**オッズ比**という。

注 5. 第 1 群の該当の割合を p 、第 2 群の該当の割合を q とすると $[p/(1-p)]/[q/(1-q)]$ 。

例 4. 処置群と対照群の陽性／陰性のオッズ比。

3 相関係数 (p. 47)

3.1 共分散 (p. 49)

2 変量データを $((x_1, y_1), \dots, (x_n, y_n))$ とする。

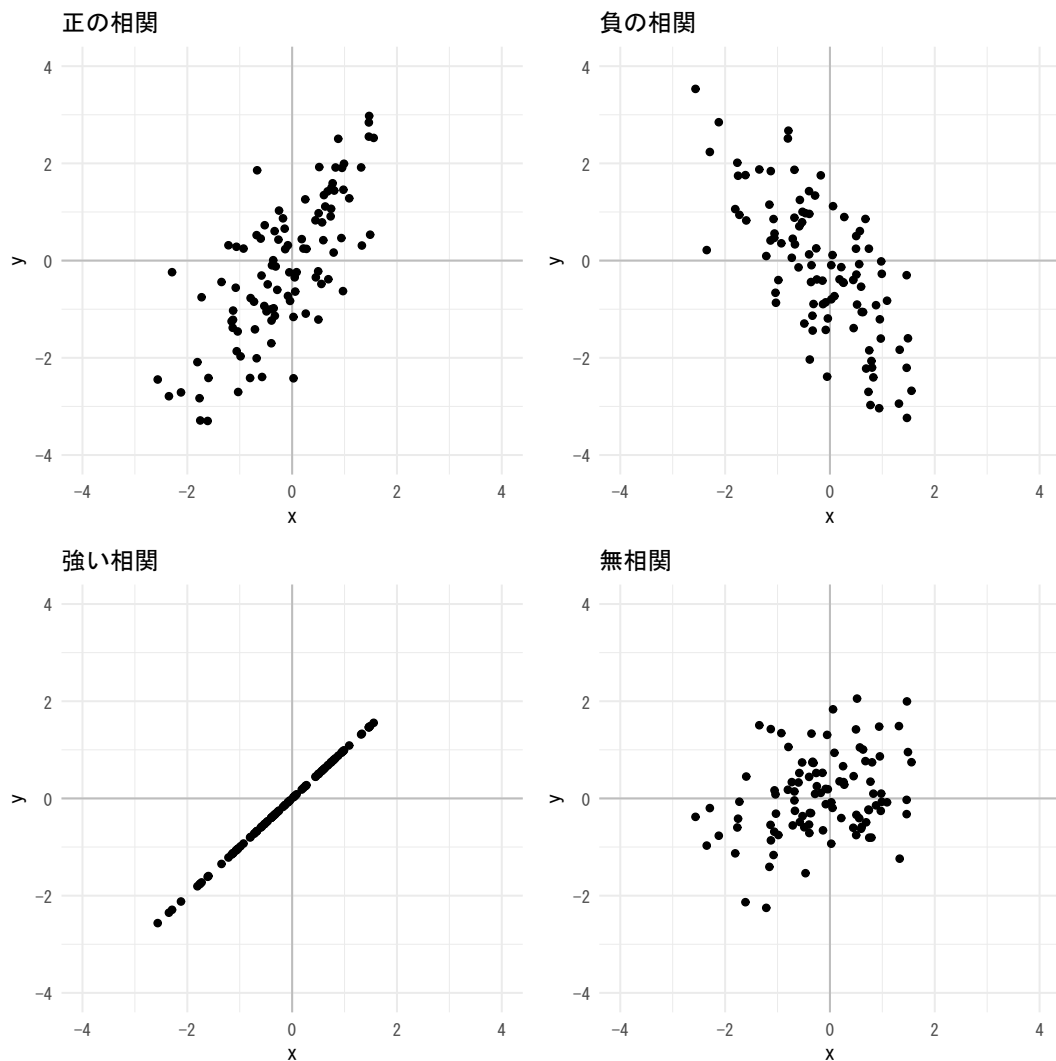


図 1: 散布図と相関関係

定義 5. 各変量の平均からの偏差の積の平均を**共分散**という.

注 6. 式で表すと

$$\sigma_{xy} := \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)$$

注 7. x_i が大きいと y_i も大きいなら共分散は正, x_i が大きいと y_i は小さいなら共分散は負, 「無関係」なら 0 となる.

3.2 標準化 (p. 39)

定義 6. 変量の値から平均を引き, 標準偏差で割る変換を**標準化**という.

注 8. 式で表すと

$$z_i := \frac{x_i - \mu_x}{\sigma_x}$$

注 9. 標準化した変量の平均は 0, 分散は 1 となる.

3.3 (積率) 相関係数 (p. 48)

定義 7. 標準化した 2 変量の共分散を (**ピアソンの積率**) 相関係数という.

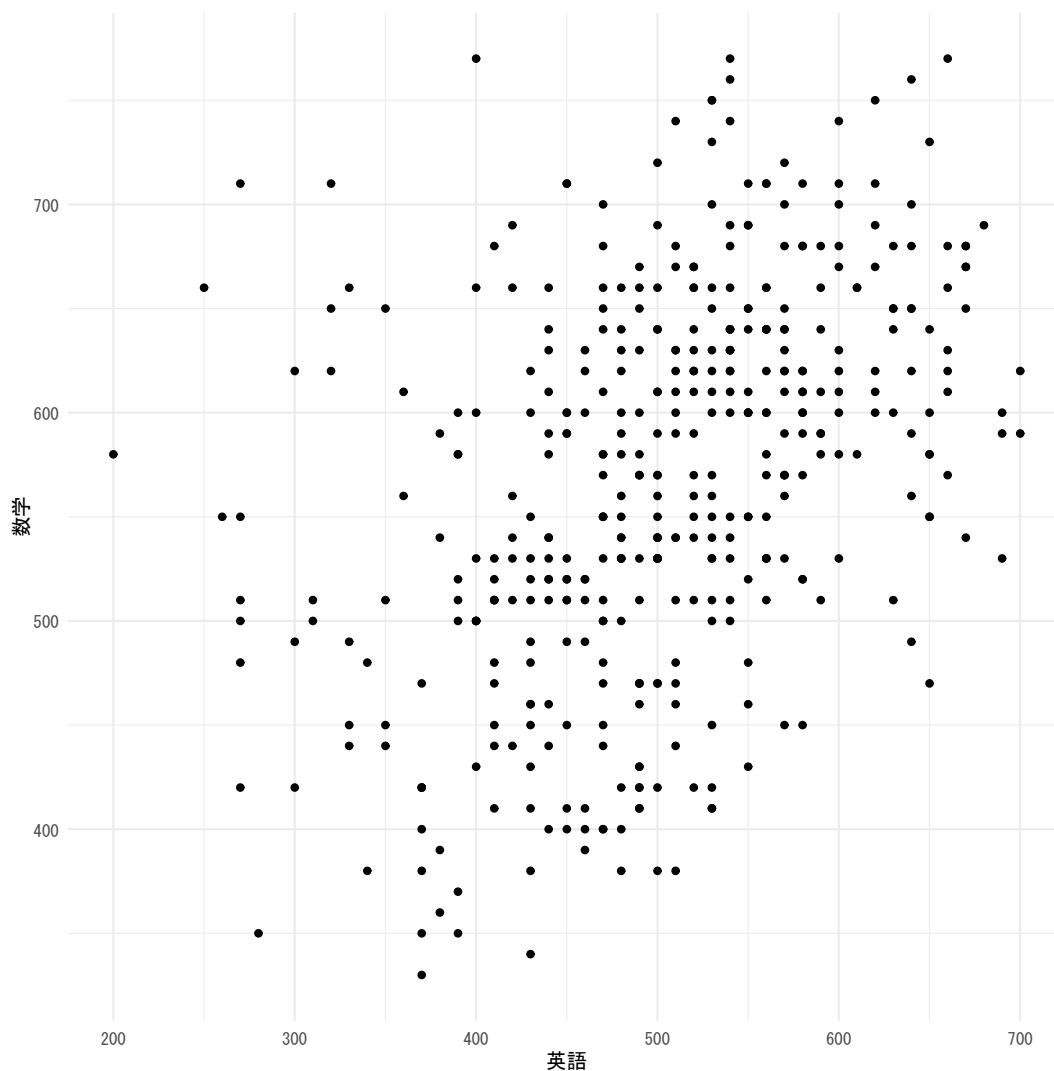


図 2: 某大学 1 年生の英語と数学の入試成績

注 10. 式で表すと

$$\begin{aligned}\rho_{xy} &= \frac{1}{n} \sum_{i=1}^n \frac{x_i - \mu_x}{\sigma_x} \frac{y_i - \mu_y}{\sigma_y} \\ &= \frac{(1/n) \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \sigma_y} \\ &= \frac{\sigma_{xy}}{\sigma_x \sigma_y}\end{aligned}$$

注 11. 「関係」が強いほど 1 か -1 に近くなる.

3.4 順位相関係数 (p. 54)

順位を表す 2 変量の相関を定義する (量的変量を順位に変換してもよい).

定義 8. 順位の (積率) 相関係数を **スピアマンの順位相関係数** という.

定理 1. 2 変量データ $((x_1, y_1), \dots, (x_n, y_n))$ が順位を表すなら

$$\rho_{xy} = 1 - \frac{6}{(n-1)n(n+1)} \sum_{i=1}^n (x_i - y_i)^2$$

証明. 省略.

□

注 12. $x_1 = y_1, \dots, x_n = y_n$ なら $\rho_{xy} = 1$.

表 1: 東大（学部・院）の学生構成

(a) 度数				(b) 相対度数			
	日本人	留学生	計		日本人	留学生	計
学部	14,871	96	14,967	学部	70.3	0.5	70.8
学部研究生	252	17	269	学部研究生	1.2	0.1	1.3
修士	2,415	274	2,689	修士	11.4	1.3	12.7
博士	2,002	620	2,622	博士	9.5	2.9	12.4
院研究生	143	454	597	院研究生	0.7	2.1	2.8
計	19,683	1,461	21,144	計	93.1	6.9	100.0

(c) 縦比				(d) 横比			
	日本人	留学生	計		日本人	留学生	計
学部	75.6	6.6	70.8	学部	99.4	0.6	100.0
学部研究生	1.3	1.2	1.3	学部研究生	93.7	6.3	100.0
修士	12.3	18.8	12.7	修士	89.8	10.2	100.0
博士	10.2	42.4	12.4	博士	76.4	23.6	100.0
院研究生	0.7	31.1	2.8	院研究生	24.0	76.0	100.0
計	100.0	100.0	100.0	計	93.1	6.9	100.0

定義 9. ケンドールの順位相関係数は

$$\tau_{xy} := \frac{\sum_{i=2}^n \sum_{j=1}^{i-1} \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j)}{nC_2}$$

注 13. $\text{sgn}(\cdot)$ は符号関数. すなわち

$$\text{sgn}(x) := \begin{cases} -1 & \text{for } x < 0 \\ 0 & \text{for } x = 0 \\ 1 & \text{for } x > 0 \end{cases}$$

注 14. 2つの観測値 (x_i, y_i) , (x_j, y_j) を取り出したとき, 「 $x_i > x_j$ だと $y_i > y_j$ 」なら順位相関係数は正, 「 $x_i > x_j$ だと $y_i < y_j$ 」なら順位相関係数は負となる.

3.5 相関と因果 (p. 50)

2変量が相関をもつ理由は2つ考えられる.

定義 10. 原因と結果の関係を因果関係という.

例 5. 身長→体重 (?), 年齢→血圧, 所得→消費, 人口→商店数.

定義 11. 因果関係のない相関を見かけ上の相関という.

注 15. 2変量の原因となる第3の変量が存在する場合に生じる.

例 6. 数学と理科の成績 (?), 飲食店数と金融機関店舗数.

4 今日のキーワード

散布図, 分割表, オッズ, オッズ比, 共分散, 標準化, (積率) 相関係数, 順位相関係数 (スピアマン, ケンドール), 因果関係, 見かけ上の相関

5 次回までの準備

提出 宿題 1

復習 教科書第3章 1~3.3.5 節, 復習テスト 3

予習 教科書第4章 1~4 節