総合科目　数理・自然 6 「統計学入門」講義レジュメ　(今野)

ふたつの変量 x, y の間の関係

1. 相関 – x, y の間に区別をもうけずに対等に観る方法．相互関係.
2. 回帰 – x から y を観ること． x から y が決定される様子.

散布図と分割表

身長	148	160	159	153	151	140	156	137	149	160	151	157	157	144
体重	41	49	45	43	42	29	49	31	47	47	42	39	48	36

1. 散布図 — 量的なデータの場合，各観測対象を平面上にプロットして，x と y の様子を視覚的に観察し，ふたつの変量間の関係を観る.
2. 分割表 — 変量のすくなくとも一方が質的データである場合に，項目別に分類されるデータの頻度をまとめたもの.

相関関係 — 一方の変量が増加すれば，他方も増加するという関係 (傾向) や一方が増加すれば，減少するという関係のこと．特に，直線関係に近い傾向が見られるときに,「相関関係がある」ということがおおい.

1. 正の相関関係 — 一方の変量が増加すれば，他方も増加するという関係
2. 負の相関関係 — 一方の変量が増加すれば，他方が減少するという関係

相関係数

相関係数には，(ピアソンの) 積率相関係数・(スピアマンの) 順位相関係数・(ケンドールの) 順位相関係数等がある.

(ピアソンの) 積率相関係数

観測値 (x_i, y_i), $i = 1, 2, \ldots, n$ を得たとする.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}$$

ただし，$\bar{x}_n = (1/n)\sum_{i=1}^{n} x_i$ と $\bar{y}_n = (1/n)\sum_{i=1}^{n} y_i$ である．これは

$$r_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}_n\bar{y}_n}{\sqrt{\sum_{i=1}^{n} x_i^2 - n(\bar{x}_n)^2}\sqrt{\sum_{i=1}^{n} y_i^2 - (\bar{y}_n)^2}}$$

とも表現できる.
(ピアソンの) 積率相関係数の性質

1. $-1 \leq r_{xy} \leq 1$
2. $r_{xy} = 1$ のとき，正の傾きの直線関係
3. $r_{xy} = 1$ のとき，負の傾きの直線関係
4. $|r_{xy}| \leq 0.5$ のとき，散布図上では相関関係があるように観えない
5. 注意：相関関係と因果関係は異なる – 見かけの相関

(スピアマンの) 順位相関係数

x の値	25	0	9	31	12	3	-9	11	-5	5
x の順位	9	3	6	10	8	4	1	7	2	5
y の値	22	0	29	39	12	9	-13	17	-8	13
y の順位	8	3	9	10	5	4	1	7	2	6

観測値 x_i の順位を $R_x(i)$，y_i の順位を $R_y(i)$ とおく．このとき，(スピアマンの) 順位相関係数を

$$r_S = \frac{\sum_{i=1}^n (R_x(i) - \bar{R}_x)(R_y(i) - \bar{R}_y)}{\sqrt{\sum_{i=1}^n (R_x(i) - \bar{R}_x)^2}\sqrt{\sum_{i=1}^n (R_y(i) - \bar{R}_y)^2}}$$

で定義する．ただし，

$$\bar{R}_x = \frac{1}{n}\sum_{i=1}^n R_x(i) = \frac{n(n+1)}{2}, \qquad \bar{R}_y = \frac{1}{n}\sum_{i=1}^n R_y(i) = \frac{n(n+1)}{2}$$

である．結果として

$$R_s = 1 - \frac{6}{n^3 - n}\sum_{i=1}^n (R_x(i) - R_y(i))^2$$

となる．

$R_x(i)$	9	3	6	10	8	4	1	7	2	5
$R_y(i)$	8	3	9	10	5	4	1	7	2	6
$(R_x(i) - R_y(i))^2$	1	0	9	0	9	0	0	0	0	1

よって

$$r_S = 1 - \frac{6}{10^3 - 10}(1 + 9 + 9 + 1) = 1 - \frac{12}{990} = 0.8787879$$

となる．

(ケンドールの) 順位相関係数

観測対象の対 (i, j) を考える．ただし，$1 \leq i < j \leq n$ である．各対 (i, j) に

1. $(R_x(i) > R_x(j)$ かつ $R_y(i) > R_u(j))$ または $(R_x(i) < R_x(j)$ かつ $R_y(i) < R_u(j))$ $\Longrightarrow$ $+1$ を対 (i, j) に与える．

2. $(R_x(i) > R_x(j)$ かつ $R_y(i) < R_u(j))$ または $(R_x(i) < R_x(j)$ かつ $R_y(i) > R_u(j))$ $\Longrightarrow$ -1 を対 $(i,\, j)$ に与える．

全対 $n(n-1)/2$ に与えられた $+1$ と -1 から

$$r_K = \frac{\{\ (+1)\ \text{を与えられた対の数}\ \}\ -\ \{\ (-1)\ \text{を与えられた対の数}\ \}}{n(n-1)/2}$$

x の値	-7	4	27	28	15
$R_x(i)$	1	2	4	5	3
y の値	5	10	11	-12	23
$R_y(i)$	2	3	4	1	5

$j \backslash i$	1	2	3	3
2	$+1$			
3	$+1$	$+1$		
4	-1	-1	-1	
5	$+1$	$+1$	-1	-1

したがって

$$r_K = \frac{4-5}{4 \times 5/2} = -0.1$$

となる．一方，$r_S = -0.1$ となる．

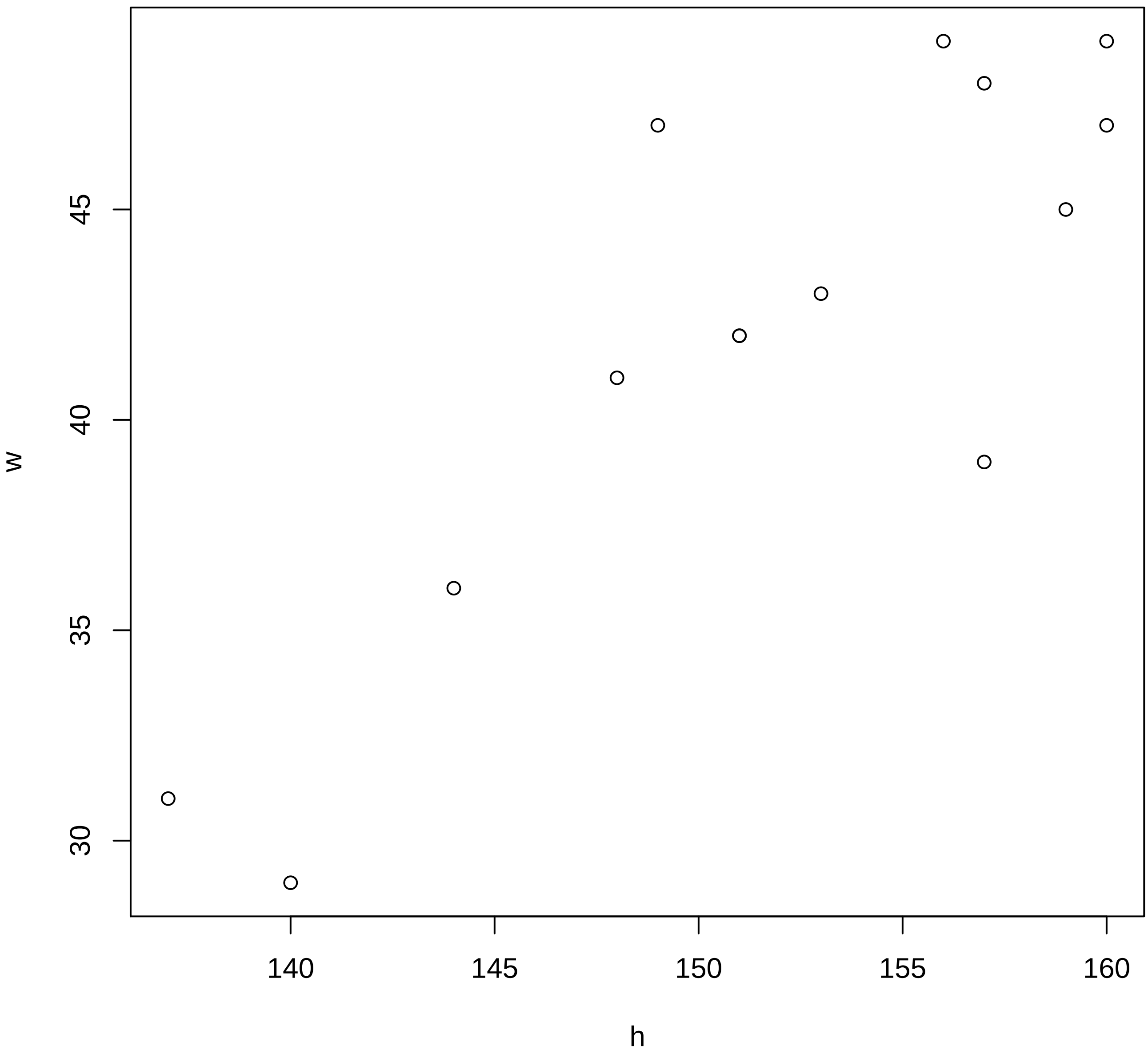
$\boxed{\text{問題}}$

散布図を作成し，どのような相関があるかを観察せよ．さらに，積率相関係数を求めよ．

x	0.27	1.41	2.19	2.83	2.19	1.81	0.85	3.05
y	2	3	3	6	4	2	1	5

以下のような表を作成すると便利である．

x	y	$x \cdot y$	x^2	y^2
0.27	2	0.54	0.0729	4
1.41	3	4.23	1.9881	4
2.19	3	6.57	4.7961	9
2.83	6	16.98	8.0089	36
2.19	4	8.76	4.7961	16
1.81	2	3.62	3.2761	4
0.85	1	0.85	0.7225	1
3.05	5	15.25	9.3025	5
14.60	26	56.80	32.9632	104


45
40
35
30
w
140
145
150
155
160
h