

## 数理統計学まとめ (その2)

### 3 相関関係

1. 散布図・相関表：2種類のデータの組  $\{(x_j, y_j); 1 \leq j \leq n\}$  (例えば身長と体重, 数学と英語のテストの点数など) が与えられたとき, 二つのデータの関連性を調べる方法.
  - (a) 散布図：そのまま縦軸と横軸に  $y_j$  と  $x_j$  をプロット.
  - (b) 相関表：階級ごとの度数分布表
2. 共分散  $S_{xy}$  :  $n$  個のデータ  $\{(x_j, y_j)\}_{j=1}^n$  が与えられたとき.

$$s_{xy} = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y}) \quad (\text{定義式})$$

$$s_{xy} = \frac{1}{n} \sum_{j=1}^n x_j y_j - \bar{x} \bar{y} \quad (\text{計算式})$$

度数分布表が与えられたとき： $x$  が  $J$  個の階級,  $y$  が  $K$  個の階級に別れ,  $x$  について  $j$  番目,  $y$  について  $k$  番目の階級の度数  $f_{jk}$ , 代表値  $x_j, y_k$  のとき

$$s_{xy} = \frac{1}{n} \sum_{j=1}^J \sum_{k=1}^K f_{jk} (x_j - \bar{x})(y_k - \bar{y}) \quad (\text{定義式})$$

$$s_{xy} = \frac{1}{n} \sum_{j=1}^J \sum_{k=1}^K f_{jk} x_j y_k - \bar{x} \bar{y} \quad (\text{計算式})$$

ただし, この場合

$$n = \sum_{j=1}^J \sum_{k=1}^K f_{jk}$$

である. ついでに言うと

$$f_{j\cdot} = \sum_{k=1}^K f_{jk}, \quad f_{\cdot k} = \sum_{j=1}^J f_{jk}$$

であり,

$$\bar{x} = \frac{1}{n} \sum_{j=1}^J f_j \cdot x_j$$
$$s_x^2 = \frac{1}{n} \sum_{j=1}^n f_j \cdot (x_j - \bar{x})^2 = \frac{1}{n} \sum_{j=1}^J f_j \cdot x_j^2 - \bar{x}^2$$

となり,  $\bar{y}$  や  $s_y^2$  を計算するときは  $f_j$  のかわりに  $f_k$  を使う.

3. Pearson の相関係数:  $s_{xy}$  を  $x, y$  それぞれの標準偏差  $s_x, s_y$  でわると標準化された  $x, y$  の  $z$ -スコアの共分散が得られる. これを相関係数という.

$$r_{x,y} = \frac{s_{xy}}{s_x s_y}$$

$|r_{xy}| \leq 1$  である. 等式は 任意の  $j$  に対して

$$\frac{x_j - \bar{x}}{s_x} = \frac{y_j - \bar{y}}{s_y},$$

つまり任意の  $j$  で  $y_j = cx_j + d$  となる  $c$  と  $d$  があるときに限る.  $r_{xy}$  が 1 に近いとき  $\{x_j\}$  と  $\{y_j\}$  は強い(正の)相関を持つという. ( $-1$  に近いときは強い負の相関があるという.)  $|r_{xy}|$  が小さいときは  $\{x_j\}$  と  $\{y_j\}$  は相関が弱いという.  $0$  に近いときは「無相関」に近いという. データ  $\{x_j\}$  と  $\{y_j\}$  が元々独立ならば  $r_{xy}$  は理論的には  $0$  になる.

### 3.1 相関係数の計算

$$u_j = \frac{x_j - a}{c}, \quad v_j = \frac{y_j - b}{d}$$

と変換する.  $c \neq 0, d \neq 0$  とする. このとき

$$s_u^2 = \frac{s_x^2}{c^2} \quad \therefore s_u = \frac{s_x}{c}$$

だったので,

$$\begin{aligned}r_{uv} &= \frac{1}{n} \sum_{j=1}^n \frac{(u_j - \bar{u})(v_j - \bar{v})}{s_u s_v} \\ &= \frac{1}{n} \sum_{j=1}^n \frac{\frac{x_j - \bar{x}}{c} \frac{y_j - \bar{y}}{d}}{\frac{s_x}{c} \frac{s_y}{d}} \\ &= \frac{1}{n} \sum_{j=1}^n \frac{(x_j - \bar{x})(y_j - \bar{y})}{s_x s_y} \\ &= r_{xy}\end{aligned}$$

となり,  $r_{uv}$  を計算すれば良い.

例 3.1 教科書 p.44 問題 2.12

y \ x	10	20	30	40	50
8	2	3	1		
12	2	4	1		
16		2	5	3	
20		1	2	3	1

サンプルの総数は 30 である .

$$u = \frac{x - 10}{10}, \quad v = \frac{y - 8}{4}$$

と置き , 相関表を作ると

v \ u	0	1	2	3	4
0	2	3	1		
1	2	4	1		
2		2	5	3	
3		1	2	3	1

計算してみると

$$\begin{aligned} \bar{u} &= \frac{10 + 18 + 18 + 4}{30} = \frac{5}{3} \\ \bar{v} &= \frac{7 + 20 + 21}{30} = \frac{8}{5} \\ s_u^2 &= \frac{10 + 36 + 54 + 16}{30} - \frac{25}{9} = \frac{98}{90} \\ s_v^2 &= \frac{7 + 40 + 63}{30} - \frac{64}{25} = \frac{83}{75} \\ s_{uv} &= \frac{4 + 4 + 3 + 2 + 20 + 12 + 18 + 27 + 12}{30} - \frac{8}{3} = \frac{22}{30} \\ r_{uv} &= \frac{22 \times 5\sqrt{3} \times 3\sqrt{10}}{30 \times \sqrt{83} \times \sqrt{98}} = 0.6680 \end{aligned}$$