第11回：データ処理と標本分布（1）

データ処理：実験や調査などによって得られたデータを整理して、意味のある情報を引き出す。

その代表例として、特に、相関係数を学ぶ。

母集団と標本：得られたデータを、「標本」と捉える。
⇒標本の平均や分散は、確率変数

データ処理・データ解析

- 実験や調査、計算結果のレポートや論文で、全く整理されていない「生」データを並べても、誰も見てくれない。
- レポート（他の人に読んで貰うべき報告書）では、目的と手段を明確にするとともに、得られた結論や主張も簡潔・明瞭でなければならない。
- 例えば、数値データは、その平均や分散などを示す。度数（頻度）分布や、そのグラフなど、視覚的にも有効な示し方をする。
- データから結論を導く際には、主観を入れず、客観的・統計的に処理する：データ処理・データ解析

標本平均と標本分散

得られたデータ＝標本(sample)
得られたデータの平均と分散
　＝標本平均・標本分散

いま、n個のデータ$x_1,x_2,\ldots,x_n$が得られたとき、

標本平均：$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$

標本分散：$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{S}{n-1}$　平方和：$S=\sum_{i=1}^{n}(x_i-\bar{x})^2$

標本標準偏差：$s=\sqrt{s^2}=\sqrt{\frac{S}{n-1}}$

テキストp.79 例題6.1

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

$$S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^{\,2}-2\bar{x}\sum_{i=1}^{n}x_i+n\cdot\bar{x}^2$$

$$=\sum_{i=1}^{n}x_i^{\,2}-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2$$

xについての平方和

$$s_x^{\,2}=\frac{S_{xx}}{n-1}$$

相関と相関係数

各データを構成する2つの変数(x, y)
　i番目のデータ(x_i, y_i)
xとyの間に「相関(correlation)」があるか？
　正の相関⇔xが増えるとyも増える
　負の相関⇔xが増えるとyは減る
散布図：各データをx-y平面上の1点として表した図。散布図によって、視覚的に（正負の）相関の有無がわかる。

相関係数：r

相関係数：直線的な相関の有無や程度を定量的に表す尺度

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{S_{xy}}{\sqrt{S_{xx}\cdot S_{yy}}}$$

ここで、積和　$S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})$

- xやyを全て定数倍しても、相関係数rは変わらない。
- rの正負は、相関の正負に対応している。

相関係数は $-1 \leqq r \leqq +1$

証明は、シュワルツの不等式から導かれる。

$$\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2 \geq \left|\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})\right|^2$$

$$\Leftrightarrow \quad |r| \equiv \left|\frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}\right| \leq 1$$

シュワルツの不等式の証明は、次ページで

$\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2 \geq \left|\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})\right|^2$ の証明

任意の実数λ,μに対して、以下が成り立つ:

$$\sum_{i=1}^{n}\{\lambda(x_i-\bar{x})-\mu(y_i-\bar{y})\}^2 \geq 0$$

$$\text{左辺} = \sum_{i=1}^{n}\{\lambda(x_i-\bar{x})-\mu(y_i-\bar{y})\}^2$$

$$= \lambda^2\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2 - 2\lambda\mu\cdot\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) + \mu^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2$$

これを、λ/μについての2次式とみなすと、常に≧0 が成り立つ

⇔ 判別式 $\leqq 0$

$$\Leftrightarrow \quad \left\{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right\}^2 - \sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2 \leq 0$$

証明おわり

証明より、また

等号が成り立つ($\Leftrightarrow r=\pm 1$)のは、以下の場合

ある実数λ,μに対して、以下の等号が全てのiで成り立つ:

$$\lambda(x_i-\bar{x})-\mu(y_i-\bar{y})=0$$

$\Leftrightarrow$ $\dfrac{y_i-\bar{y}}{x_i-\bar{x}}$ 比が、全ての i において一定

$\Leftrightarrow$ 全ての (x_i,y_i) が同じ直線上にある

- r の正負は、相関の正負に対応している。
- $|r|$ が1に近いほど(直線的な)相関が強く、0に近いほど相関は弱い。

データ処理の課題1

教科書p.79の例題6．1で、最初の3人に限定して、体重xと身長yについて、相関係数を求めなさい。

i	1	2	3	計
x	2250	3525	3005	8780
y	46.1	52.0	48.5	146.6
x^2	5,062,500	12,425,625	9,030,025	26,518,150
y^2	2125.21	2704.00	2352.25	7181.46
xy	103,725.0	183,300.0	145,742.5	432,767.5

$$\sum x = 8780,\ \sum y = 146.6,$$
$$\sum x^2 = 26518150,\ \sum y^2 = 7181.46,\ \sum xy = 432767.5$$

$$S_{xx} = \sum x^2 - \frac{1}{n}\cdot\left(\sum x\right)^2 = 26518150 - \frac{1}{3}\cdot 8780^2 = 822{,}016.7$$

$$S_{yy} = \sum y^2 - \frac{1}{n}\cdot\left(\sum y\right)^2 = 7181.46 - \frac{1}{3}\cdot 146.6^2 = 17.6$$

$$S_{xy} = \sum xy - \frac{1}{n}\cdot\sum x\cdot\sum y = 432767.5 - \frac{1}{3}\cdot 8780\cdot 146.6 = 3718.2$$

$$\therefore r = \frac{S_{xy}}{\sqrt{S_{xx}\cdot S_{yy}}} = \frac{3718.2}{\sqrt{822017\times 17.6}} = \frac{3718.2}{3803.6} = 0.978$$