データの分析

目次

本スライドの内容

このスライドは，次の書籍の第 4 章「データの分析」の内容に基づく．

- 『ガイダンス 確率統計：基礎から学び本質の理解へ』，発行：サイエンス社，ISBN：978-4-7819-1526-5.

書籍に関する最新の情報は，以下の URL から入手することができます．

https://www.saiensu.co.jp

この URL は，サイエンス社が運営しているホームページです．

概要

このスライドでは，得られたデータを処理することで，そのデータが持つ特徴を明らかにする統計分析手法について解説する．この手法は記述統計とよばれ，その考え方は全数調査などで活用されている．

データの代表値

気温や降水量，身長や体重などのように，ある集団を構成する人や物の特性を数量的に表すものを変量といい，調査や実験などで得られた変量の観測値や測定値の集まりをデータという．データを構成する観測値や測定値の個数を，そのデータの大きさという．データの分布の状態は，度数分布表やヒストグラムなどによって知ることができるが，データ全体の特徴を適当な１つの数値で表すこともあり，その数値をデータの代表値とよぶ．よく用いられる代表値として，平均値，分散，標準偏差，中央値，最頻値がある．

データの代表値：平均値，分散

変量 x についてのデータが n 個の値 $x_1, x_2, \cdots, x_n$ であるとき，このデータの平均値 $\overline{x}$ と分散 s_x^2 を

$$\overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n), \tag{4.1}$$

$$s_x^2 = \frac{1}{n}\left\{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2\right\} \tag{4.2}$$

と定義し，分散 s_x^2 の正の平方根 $\sqrt{s_x^2}$ をデータの標準偏差とよび s_x で表す．分散 s_x^2 は，データの散らばりの度合いを表す量であり，データの各値が平均値から離れるほど大きな値を取る．たとえば，変量 x の測定単位が cm であるとき，分散 s_x^2 の単位は cm^2 となるが，標準偏差 s_x の単位は変量 x の測定単位と同じ cm である．

データの代表値：中央値

次に，この変量 x のデータ $x_1, x_2, \cdots, x_n$ を小さい順に並び替えた $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ に対して，中央の位置に来る値は

$$\widetilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & (n \text{ が奇数}) \\ \frac{1}{2}\left\{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right\} & (n \text{ が偶数}) \end{cases}$$

で与えられ，この値 $\widetilde{x}$ はデータの中央値またはメジアンとよばれる．たとえば，ある学生の通学時間を，ある週の 5 日について調べた結果 (データ) が

42　38　40　44　96　　(単位は分)

であるとき，このデータの平均値は 52 分であるが，1 日だけ通学時間が極端に長かったために，この平均値は，他の 4 日の通学時間からは離れたものになっていて，このデータの代表値として適切とはいえない．一方で，このデータの中央値は 42 分であり，この値をデータの代表値とすることが考えられる．

データの代表値：最頻値

最も観測された個数の多いデータの値は，そのデータの最頻値またはモードとよばれる．服や靴の最も売れ行きのよいサイズなどを知りたい場合，代表値としては最頻値が適切である．たとえば，ある店での 1 週間の靴のサイズ別の販売数を調べたところ，次の表のようになったとすると，最頻値は 26 cm である．

サイズ (cm)	24	24.5	25	25.5	26	26.5	27	計
販売数	3	7	13	16	24	10	4	77

データの相関と回帰分析

2 つの変量 x, y が，n 個の x, y の観測値の組として

$$(x_1, y_1),\ (x_2, y_2),\ \cdots,\ (x_n, y_n) \qquad (4.3)$$

のように与えられているとする．データ $x_1, x_2, \cdots, x_n$ とデータ $y_1, y_2, \cdots, y_n$ の平均値をそれぞれ $\overline{x}$, $\overline{y}$，分散をそれぞれ s_x^2, s_y^2，標準偏差をそれぞれ s_x, s_y とする．ここで，x と y の共分散 s_{xy} を

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

と定義する．さらに，共分散 s_{xy} を，標準偏差 s_x と s_y の積 $s_x s_y$ で割った量 $r_{xy} = s_{xy}/(s_x s_y)$ は，x と y の相関係数とよばれる．相関係数 r_{xy} については，一般に不等式 $-1 \leq r_{xy} \leq 1$ が成り立つ．

データの相関と回帰分析

以下では，変量 x と変量 y の関係性を 1 次関数

$$y = f(x) = ax + b \quad (a, b \text{ は定数}) \qquad (4.5)$$

で捉えることを考える．このとき，x を説明変数とよび，y を被説明変数とよぶ．ただし，$f(x_i)$ と y_i が一致するとは限らないため，推定値 $f(x_i)$ と観測値 y_i の (回帰) 残差 ε_i を

$$y_i = f(x_i) + \varepsilon_i \quad (i = 1, 2, \cdots, n)$$

と定義する．残差平方和 $\mathrm{RSS}(a, b)$ を

$$\mathrm{RSS}(a, b) := \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \qquad (4.6)$$

と定義し，$\mathrm{RSS}(a, b)$ を最小にするように a, b を定めるのが最小二乗法の考え方である．

データの相関と回帰分析

RSS(a, b) を最小にするような a, b を a_0, b_0 とすると，直線 $y = a_0x + b_0$ で与えられる回帰モデルを線形回帰モデル，または「y の x への回帰直線」といい，定数 a_0 を回帰係数という．まず，(4.6) の右辺を展開することで，

$$\begin{aligned}&\mathrm{RSS}(a, b)\\ &= n\left\{s_y^2 + \overline{y}^2 + a^2(s_x^2 + \overline{x}^2) + b^2 - 2a(s_{xy} + \overline{x}\cdot\overline{y}) - 2b\overline{y} + 2ab\overline{x}\right\}\end{aligned}$$

が得られる．よって，RSS(a, b) の偏微分は

$$\begin{aligned}\frac{\partial}{\partial a}\mathrm{RSS}(a, b) &= 2n\left\{a(s_x^2 + \overline{x}^2) - (s_{xy} + \overline{x}\cdot\overline{y}) + b\overline{x}\right\},\\ \frac{\partial}{\partial b}\mathrm{RSS}(a, b) &= 2n\left\{b - \overline{y} + a\overline{x}\right\}\end{aligned}$$

と計算できる．

データの相関と回帰分析

RSS(a, b) を最小にするような a, b を求めるために，連立方程式

$$\frac{\partial}{\partial a}\mathrm{RSS}(a, b) = 0, \quad \frac{\partial}{\partial b}\mathrm{RSS}(a, b) = 0.$$

を解くと，次が得られる．

$$a_0 = \frac{s_{xy}}{s_x^2}, \quad b_0 = \overline{y} - a_0\overline{x}.$$

したがって，求めるべき線形回帰モデル $y = a_0x + b_0$ は

$$y - \overline{y} = a_0(x - \overline{x}) = \frac{s_{xy}}{s_x^2}(x - \overline{x}) = \frac{s_y}{s_x}r_{xy}(x - \overline{x})$$

で与えられる．