平均の再定義と拡張

――中央値，レンジと最小二乗法――

かめさん　@cogitoergosumkm

$x_1, x_2, \ldots, x_n$ をデータとする．

定義 0.1 (平均と分散)**.**
平均 μ，分散 σ は

$$\mu = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{0.1}$$

$$\sigma^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \mu)^2 \tag{0.2}$$

と定義される．

1　平均の再定義

定義 1.1 (平均と分散)**.**

$$g(\mu) = \sum_{k=1}^{n}(x_k - \mu)^2 \tag{1.1}$$

とおく．$g(\mu)$ を最小にする μ を $x_1, x_2, \ldots, x_n$ の平均，$\min_\mu\{g(\mu)/n\}$ を分散と定義する．

先の定義と同じことを確認する．

$$\begin{aligned} g(\mu) &= \sum_{k=1}^{n}(x_k - \mu)^2 && (1.2)\\ &= n\mu^2 - 2\mu\sum_{k=1}^{n} x_k + \sum_{k=1}^{n} x_k^2 && (1.3)\\ &= n\left(\mu - \frac{1}{n}\sum_{k=1}^{n} x_k\right)^2 + \sum_{k=1}^{n} x_k^2 - \frac{1}{n}\left(\sum_{k=1}^{n} x_k\right)^2 && (1.4) \end{aligned}$$

したがって，$\mu = \frac{1}{n}\sum_{k=1}^{n} x_k$ のときに $g(\mu)$ は最小値をとる．その値を計算すると以下のようになる．

$$\begin{aligned} \frac{1}{n}g\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right) &= \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)^2 && (1.5)\\ &= \frac{1}{n}\sum_{k=1}^{n}\left(x_k - \frac{1}{n}\sum_{k=1}^{n} x_k\right)^2 && (1.6) \end{aligned}$$

2　平均の拡張 ――中央値，レンジ――

この節においては，データを並べ替えて，$x_1 \leq x_2 \leq \ldots \leq x_n$ としておく．

定義 2.1 (p ノルム)**.**
$\boldsymbol{v} = (v_1, v_2, \dots, v_n)$ に対して，$1 \le p < \infty$ として，p ノルムを

$$\|\boldsymbol{v}\|_p = \left(\sum_{k=1}^{n} |v_k|^p\right)^{\frac{1}{p}} \tag{2.1}$$

と定義し，$p \to \infty$ を考えて，

$$\|\boldsymbol{v}\|_\infty = \max\{v_1, v_2, \dots, v_n\} \tag{2.2}$$

$\boldsymbol{x} = (x_1, x_2, \dots, x_n)$，$\boldsymbol{\eta} = (1, 1, \dots, 1)$ とする．$g_p(\mu) = \|\boldsymbol{x} - \mu\boldsymbol{\eta}\|_p$ とすると，前節の $g(\mu)$ は $g_2(\mu)$ と等しい．

定義 2.2.
$1 \le p \le \infty$ に対して，$g_p(\mu)$ を最小にする μ を μ_p と定義する．さらに，$\sigma_p = n^{-\frac{1}{p}} g_p(\mu_p)$ と定義する．

$p = 2$ のとき，μ_2 は平均となり，σ_2 は標準偏差，つまり σ_2^2 は分散となる．
ほかによく用いるものとして，$p = 1$ を考えると，

$$g_1(\mu) = \sum_{k=1}^{n} |x_k - \mu| \tag{2.3}$$

である．

$$\frac{d}{d\mu}|x_k - \mu| = \begin{cases} -1 & (\mu < x_k) \\ 1 & (\mu > x_k) \end{cases} \tag{2.4}$$

であるから，n が奇数なら，$\mu_1 = x_{\frac{n+1}{2}}$ となり，n が偶数なら，μ_1 は区間 $[x_{\frac{n}{2}}, x_{\frac{n}{2}+1}]$ 内の任意の数である．偶数の時は $\mu_1 = \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1})$ をとれば，μ_1 は中央値である．
さらに，$p = \infty$ のときは，

$$g_\infty(\mu) = \max_{1 \le i \le n}\{|x_i - \mu|\} \tag{2.5}$$

だから，$\mu_\infty = \frac{x_1 + x_n}{2}$ であり，$p \to \infty$ で $n^{-\frac{1}{p}} \to 0$ だから，$\sigma_\infty = g_\infty(\mu_\infty) = \frac{x_n - x_1}{2}$ である．これはレンジ $x_n - x_1$ の半分になっている．

3　平均の拡張 ――最小二乗法――

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ をデータとする．$X_1, X_2, \dots, X_n$ の平均を $\overline{X}$，分散を σ_X^2 と書くこととする．また，$Y_1, Y_2, \dots, Y_n$ もあるとき，共分散を

$$\begin{aligned} \mathrm{Cov}(X, Y) &\coloneqq \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})(Y_k - \overline{Y}) && (3.1) \\ &= \overline{(X - \overline{X})(Y - \overline{Y})} && (3.2) \\ &= \overline{X\,Y} - \overline{X}\,\overline{Y} && (3.3) \end{aligned}$$

と定義する．$g(\mu) = \sum_{k=1}^{n}(x_k - \mu)^2$ とおけば，g を最小化する μ として x の平均が得られ，$g(\mu) = \sum_{k=1}^{n}(y_k - \mu)^2$ とおけば，g を最小化する μ として y の平均が得られるが，これでは (x, y) とペアでデータを得たことが活かされない．そこで μ を関数だと考える．g は汎関数となる．

$$g[f] = \sum_{k=1}^{n}(y_k - f(x_k))^2 \tag{3.4}$$

とはいえ，任意の関数をとって良ければ，データに (x, y) と (x, y') のような，x は等しく，y は異なるデータが含まれない限り g の最小値は 0 となり，そのような f は無数に得られてしまう．$f_1(x), f_2(x), \ldots, f_m(x)$ をあらかじめ決めておき，

$$f(x; a_1, a_2, \ldots, a_m) = \sum_{k=1}^{m} a_k f_k(x) \tag{3.5}$$

と定義して，

$$g(a_1, a_2, \ldots, a_m) = \sum_{k=1}^{n} (y_k - f(x_k; a_1, a_2, \ldots, a_m)^2 \tag{3.6}$$

を最小化する $a_1, a_2, \ldots, a_m$ の組を求めて近似関数を得る．このような方法を最小二乗法という．ただし，$m < n$ である．もし，$m \geq n$ ならば，やはり g の最小値は 0 になってしまい，強引に形を合わせたに過ぎないこととなってしまう．
特に，直線で近似する場合を考える．

$$f(x; a, b) = ax + b \tag{3.7}$$

$$g(a, b) = \sum_{k=1}^{n} (y_k - ax_k - b)^2 \tag{3.8}$$

まず，$X = y - ax$ として，g を変形する．式 (1.4) より，

$$g(a, b) = \sum_{k=1}^{n} (X_k - b)^2 \tag{3.9}$$

$$= n\left(b - \frac{1}{n}\sum_{k=1}^{n} X_k\right)^2 + \sum_{k=1}^{n} X_k^2 - \frac{1}{n}\left(\sum_{k=1}^{n} X_k\right)^2 \tag{3.10}$$

$$\frac{1}{n}g(a, b) = \{b - (\overline{y - ax})\}^2 + \overline{X^2} - \overline{X}^2 \tag{3.11}$$

さらに，

$$\overline{X^2} - \overline{X}^2 = \overline{(y - ax)^2} - (\overline{y} - a\overline{x})^2 \tag{3.12}$$

$$= a^2(\overline{x^2} - \overline{x}^2) - 2a(\overline{xy} - \overline{x}\overline{y}) + (\overline{y^2} - \overline{y}^2) \tag{3.13}$$

$$= \sigma_x^2 a^2 - 2\mathrm{Cov}(x, y)a + \sigma_y^2 \tag{3.14}$$

$$= \sigma_x^2\left(a - \frac{\mathrm{Cov}(x, y)}{\sigma_x^2}\right)^2 - \frac{(\mathrm{Cov}(x, y))^2}{\sigma_x^2} + \sigma_y^2 \tag{3.15}$$

まとめれば，

$$\frac{1}{n}g(a, b) = (b - \overline{y} + a\overline{x})^2 + \sigma_x^2\left(a - \frac{\mathrm{Cov}(x, y)}{\sigma_x^2}\right)^2 - \frac{(\mathrm{Cov}(x, y))^2}{\sigma_x^2} + \sigma_y^2 \tag{3.16}$$

したがって，

$$a = \frac{\mathrm{Cov}(x, y)}{\sigma_x^2} \tag{3.17}$$

$$b = \overline{y} - \frac{\mathrm{Cov}(x, y)}{\sigma_x^2}\overline{x} \tag{3.18}$$

のとき，g は最小値をとり，$y = ax + b$ は直線での近似となる．