Publication: lifemater.com, Author: peekAboo, Creation: 2021/12/30

平均・分散・共分散

1. 平均

n 個のデータ $\{x_i\}$ について、その平均値 μ_x は以下の式で与えられる。

$$\mu_x = \frac{1}{n}\sum_i x_i \,. \tag{1-1}$$

ここで、a, b を定数として、

$$X_i = a(x_i - b), \tag{1-2}$$

とすると、その平均値 μ_X は、

$$\mu_X = \frac{1}{n}\sum_i X_i = \frac{1}{n}\sum_i a(x_i - b) = a\frac{1}{n}\sum_i (x_i - b) = a\frac{1}{n}\left(\sum_i x_i - \sum_i b\right)$$
$$= a\left(\frac{1}{n}\sum_i x_i - b\frac{1}{n}\sum_i 1\right) = a\left(\mu_x - b\frac{1}{n}n\right) = a(\mu_x - b)$$

となる。つまり、

$$\mu_X = a(\mu_x - b), \tag{1-3}$$

となる。

2. 分散

n 個のデータ $\{x_i\}$ について、その分散 s_{xx}^2 は以下の式で与えられる（ここでは、3 章の共分散と下付き添え字の対応を保つために x を二つ重ねている）。

$$s_{xx}^2 = \frac{1}{n}\sum_i (x_i - \mu_x)^2 \,. \tag{2-1}$$

この式を展開すると、Eq. (1-1)を用いれば、

$$s_{xx}^2 = \frac{1}{n}\sum_i \left(x_i - \mu_x\right)^2 = \frac{1}{n}\sum_i \left(x_i^2 - 2x_i\mu_x + \mu_x^2\right) = \frac{1}{n}\left(\sum_i x_i^2 - 2\sum_i x_i\mu_x + \sum_i \mu_x^2\right)$$
$$= \frac{1}{n}\sum_i x_i^2 - 2\mu_x \frac{1}{n}\sum_i x_i + \mu_x^2 \frac{1}{n}\sum_i 1 = \mu_{x^2} - 2\mu_x\mu_x + \mu_x^2 \frac{1}{n} n = \mu_{x^2} - 2\mu_x^2 + \mu_x^2 \quad ,$$
$$= \mu_{x^2} - \mu_x^2$$

となるので、

$$s_{xx}^2 = \mu_{x^2} - \mu_x^2 , \tag{2-2}$$

が成立する。分散s_{xx}^2の平方根をとった量は標準偏差σ_xであり、Eqs. (2-1, 2-2)より、

$$\sigma_x = \sqrt{\frac{1}{n}\sum_i \left(x_i - \mu_x\right)^2} = \sqrt{\mu_{x^2} - \mu_x^2} , \tag{2-3}$$

にて与えられる。

ここで、*a, b* を定数として、

$$X_i = a\left(x_i - b\right) , \tag{2-4}$$

とすると、その分散$s_{XX}{}^2$は、Eqs. (2-2, 1-3)より、

$$s_X{}^2 = \mu_{X^2} - \mu_X^2 = \mu_{X^2} - \left(a\left(\mu_x - b\right)\right)^2 = \mu_{X^2} - a^2\left(\mu_x^2 - 2b\mu_x + b^2\right),$$

となる。また、Eqs. (1-1, 2-4)より、

$$\mu_{X^2} = \frac{1}{n}\sum_i \left(X_i\right)^2 = \frac{1}{n}\sum_i \left(a\left(x_i - b\right)\right)^2 = a^2\frac{1}{n}\sum_i \left(x_i - b\right)^2 = a^2\frac{1}{n}\sum_i \left(x_i^2 - 2bx_i + b^2\right)$$
$$= a^2\frac{1}{n}\left(\sum_i x_i^2 - 2\sum_i bx_i + \sum_i b^2\right) = a^2\left(\frac{1}{n}\sum_i x_i^2 - 2b\frac{1}{n}\sum_i x_i + b^2\frac{1}{n}\sum_i 1\right) = a^2\left(\mu_{x^2} - 2b\mu_x + b^2\right) ,$$

となるので、Eq. (2-2)を用いて、

$$s_{XX}{}^2 = \mu_{X^2} - a^2\left(\mu_x^2 - 2b\mu_x + b^2\right) = a^2\left(\mu_{x^2} - 2b\mu_x + b^2\right) - a^2\left(\mu_x^2 - 2b\mu_x + b^2\right)$$
$$= a^2\left(\mu_{x^2} - \mu_x^2\right) = a^2 s_{xx}^2 ,$$

となる。つまり、

$$s_{XX}{}^2 = a^2 s_{xx}^2 , \tag{2-5}$$

となる。Eq. (2-5)は b によらないため、分散は b に依存しないことが分かる。

3. 共分散

n 個のデータ $\{(x_i, y_i)\}$ について、その共分散 s_{xy}^2 は以下の式で与えられる。

$$s_{xy}^2 = \frac{1}{n}\sum_i (x_i - \mu_x)(y_i - \mu_y). \tag{3-1}$$

この式を展開すると、Eq. (1-1)を用いれば、

$$\begin{aligned} s_{xy}^2 &= \frac{1}{n}\sum_i (x_i - \mu_x)(y_i - \mu_y) = \frac{1}{n}\sum_i (x_i y_i - x_i\mu_y - y_i\mu_x + \mu_x\mu_y) \\ &= \frac{1}{n}\left(\sum_i x_i y_i - \mu_y\sum_i x_i - \mu_x\sum_i y_i + \mu_x\mu_y\sum_i 1\right) = \frac{1}{n}(n\mu_{xy} - \mu_y n\mu_x - \mu_x n\mu_y + \mu_x\mu_y n), \\ &= \mu_{xy} - \mu_y\mu_x - \mu_x\mu_y + \mu_x\mu_y = \mu_{xy} - \mu_y\mu_x \end{aligned}$$

となるので、

$$s_{xy}^2 = \mu_{xy} - \mu_x\mu_y, \tag{3-2}$$

が成立する。

ここで、a, b, c, d を定数として、

$$X_i = a(x_i - b), \tag{3-3}$$

$$Y_i = c(y_i - d), \tag{3-4}$$

とすると、その共分散 $s_{XY}{}^2$ は、Eqs. (3-2, 1-3)より、

$$\begin{aligned} s_{XY}{}^2 &= \mu_{XY} - \mu_X\mu_Y = \mu_{XY} - a(\mu_x - b)c(\mu_y - d) \\ &= \mu_{XY} - ac(\mu_x\mu_y - d\mu_x - b\mu_y + bd) \end{aligned},$$

となる。また、Eqs. (1-1, 3-3, 3-4)より、

$$\mu_{\mathrm{XY}} = \frac{1}{n}\sum_i X_i Y_i = \frac{1}{n}\sum_i a(x_i - b)c(y_i - d) = ac\frac{1}{n}\sum_i (x_i y_i - x_i d - y_i b + bd)$$
$$= ac\frac{1}{n}\left(\sum_i x_i y_i - d\sum_i x_i - b\sum_i y_i + bd\sum_i 1\right) = ac\frac{1}{n}\left(n\mu_{\mathrm{xy}} - dn\mu_{\mathrm{x}} - bn\mu_{\mathrm{y}} + bdn\right),$$
$$= ac\left(\mu_{\mathrm{xy}} - d\mu_{\mathrm{x}} - b\mu_{\mathrm{y}} + bd\right)$$

となるので、Eq. (3-2)を用いて、

$$s_{\mathrm{XY}}{}^2 = \mu_{\mathrm{XY}} - ac\left(\mu_{\mathrm{x}}\mu_{\mathrm{y}} - d\mu_{\mathrm{x}} - b\mu_{\mathrm{y}} + bd\right)$$
$$= ac\left(\mu_{\mathrm{xy}} - d\mu_{\mathrm{x}} - b\mu_{\mathrm{y}} + bd\right) - ac\left(\mu_{\mathrm{x}}\mu_{\mathrm{y}} - d\mu_{\mathrm{x}} - b\mu_{\mathrm{y}} + bd\right)$$
$$= ac\left(\mu_{\mathrm{xy}} - \mu_{\mathrm{x}}\mu_{\mathrm{y}}\right) = acs_{\mathrm{xy}}^2$$

となる。つまり、

$$s_{\mathrm{XY}}{}^2 = acs_{\mathrm{xy}}^2, \tag{3-5}$$

となる。Eq. (3-5)は *b, d* によらないため、共分散は *b, d* に依存しない。

2 章の分散は共分散にて $y_i = x_i\ (\forall i)$ の場合に等しく、2 章の一連の説明は 3 章にて内包されている。

4. 相関係数

n 個のデータ $\{(x_i, y_i)\}$ について、その相関係数 ρ_{xy} は以下の式で与えられる。

$$\rho_{\mathrm{xy}} = \frac{s_{\mathrm{xy}}^2}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}}. \tag{4-1}$$

Eqs. (3-3, 3-4)と同様に X_i, Y_i を定義すると、その相関係数 ρ_{XY} は Eq. (4-1)より、以下のようになる。

$$\rho_{\mathrm{XY}} = \frac{s_{\mathrm{XY}}^2}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}},$$

Eqs. (2-5, 3-5, 4-1)より、

$$\rho_{\mathrm{XY}} = \frac{s_{\mathrm{XY}}^2}{\sigma_{\mathrm{X}}\sigma_{\mathrm{Y}}} = \frac{acs_{\mathrm{xy}}^2}{\sqrt{a^2 s_{\mathrm{xx}}^2}\sqrt{c^2 s_{\mathrm{yy}}^2}} = \frac{acs_{\mathrm{xy}}^2}{|a|\sqrt{s_{\mathrm{xx}}^2}|c|\sqrt{s_{\mathrm{yy}}^2}} = \frac{ac}{|a||c|}\frac{s_{\mathrm{xy}}^2}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} = \frac{ac}{|ac|}\rho_{\mathrm{xy}},$$

となる。つまり、

$$\rho_{\mathrm{XY}} = \frac{ac}{|ac|}\rho_{\mathrm{xy}}, \tag{4-2}$$

となる。つまり、以下の関係式が成立する。

$$\rho_{\mathrm{XY}} = \begin{cases} \rho_{\mathrm{xy}} & (ac > 0) \\ -\rho_{\mathrm{xy}} & (ac < 0) \end{cases}. \tag{4-3}$$

共分散は x, y の相関関係を反映しているが、Eq. (3-5)に示したように、a, c に比例する量であり、その絶対値はスケーリングに対して不変ではない。それに対して、相関係数は a, c の大きさには依存せず、スケーリングに対して不変である。

Cauchy–Schwarz の不等式より、n 要素の任意の実数列 $\{x_i\}, \{y_i\}$ について、以下の関係が成立する。

$$\left(\sum_i x_i y_i\right)^2 \le \left(\sum_i x_i^2\right)\left(\sum_i y_i^2\right), \tag{4-4}$$

したがって、Eq. (3-1)より、

$$\begin{aligned} \left(s_{\mathrm{xy}}^2\right)^2 &= \frac{1}{n^2}\left(\sum_i (x_i - \mu_{\mathrm{x}})(y_i - \mu_{\mathrm{y}})\right)^2 \\ &\le \frac{1}{n^2}\left\{\sum_i (x_i - \mu_{\mathrm{x}})^2\right\}\left\{\sum_i (y_i - \mu_{\mathrm{y}})^2\right\} = \frac{1}{n^2}\left\{ns_{\mathrm{xx}}^2\right\}\left\{ns_{\mathrm{yy}}^2\right\} = s_{\mathrm{xx}}^2 s_{\mathrm{yy}}^2 \end{aligned},$$

となる。よって、

$$\rho_{\mathrm{xy}}{}^2 = \frac{\left(s_{\mathrm{xy}}^2\right)^2}{\sigma_{\mathrm{x}}{}^2\sigma_{\mathrm{y}}{}^2} \le \frac{s_{\mathrm{xx}}^2 s_{\mathrm{yy}}^2}{s_{\mathrm{xx}}^2 s_{\mathrm{yy}}^2} = 1,$$

となる。つまり、

$$\left|\rho_{\mathrm{xy}}\right| \le 1. \tag{4-5}$$

となる。