＜ 2 次元正規分布に従うデータと回帰直線 2 ＞

前ページの性質 $\boxed{2}$ を示す。

各 (X_i, Y_i) は 2 次元正規分布に従い、その密度関数は

$$p(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-m_X}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-m_X}{\sigma_1}\right)\left(\frac{y-m_Y}{\sigma_2}\right) + \left(\frac{y-m_Y}{\sigma_2}\right)^2\right\}}$$

とする。これより P.16 定理 3 から

$$E[\,X_i\,] = m_X\ ,\quad E[\,Y_i\,] = m_Y\ ,\qquad V(X_i) = E\big[(X_i - m_X)^2\big] = {\sigma_1}^2$$

$$V(Y_i) = E\big[(Y_i - m_Y)^2\big] = {\sigma_2}^2\ ,\quad C_{ov}(X_i,\ Y_i) = E\big[(X_i - m_X)(Y_i - m_Y)\big] = \rho\sigma_1\sigma_2$$

となる。一方、2 次元データ (X_i, Y_i) $(1 \leqq i \leqq n)$ の統計量は

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\ ,\quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i\ ,\qquad S_{xx} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$S_{yy} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2\ ,\qquad S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

である。それらは確率変数であるから、その平均をとると

$$E[\,\overline{X}\,] = m_X\ ,\quad E[\,\overline{Y}\,] = m_Y\ ,\qquad E[\,S_{xx}\,] = \frac{n-1}{n}{\sigma_1}^2$$

$$E[\,S_{yy}\,] = \frac{n-1}{n}{\sigma_2}^2\ ,\qquad E[\,S_{xy}\,] = \frac{n-1}{n}\rho\sigma_1\sigma_2$$

となる。(証明は不偏分散の項でする。) 大数の法則より、$n \to \infty$ のとき平均に近づくので、n が十分大きければ

$$\overline{X} \doteqdot m_X\ ,\quad \overline{Y} \doteqdot m_Y\ ,\quad S_{xx} \doteqdot {\sigma_1}^2\ ,\quad S_{yy} \doteqdot {\sigma_2}^2\ ,\quad S_{xy} \doteqdot \rho\sigma_1\sigma_2$$

とみなせる。

一方、データ (X_i, Y_i) の回帰直線の方程式は

$$y = \frac{S_{xy}}{S_{xx}}(x - \overline{X}) + \overline{Y} \qquad \text{(データの回帰直線)}$$

であるが、傾きは $\dfrac{S_{xy}}{S_{xx}} \doteqdot \dfrac{\rho\sigma_1\sigma_2}{{\sigma_1}^2} = \rho\dfrac{\sigma_2}{\sigma_1}$ であり $\overline{X} \doteqdot m_X$, $\overline{Y} \doteqdot m_Y$ より

この直線は

$$y = \rho\frac{\sigma_2}{\sigma_1}(x - m_X) + m_Y \qquad \text{(前ページ直線②)}$$

で近似できるので、性質 $\boxed{2}$ が示された。