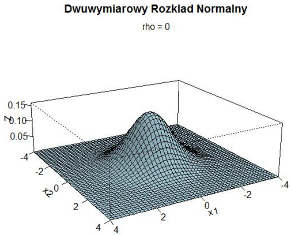
Rozkład normalny wielowymiarowy

Joanna Czarnowska[1]

[1]Uniwersytet Gdański
Instytut Matematyki

Rozkład normalny standardowy dwuwymiarowy



$(X, Y) \sim N(0, 0, 1, 1, 0)$

Gęstość: $f(x, y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}, \ (x, y) \in \mathbb{R}$

Wartość oczekiwana $\mu = (EX, EY) = (0, 0)$.

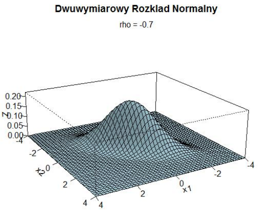
Macierz kowariancji

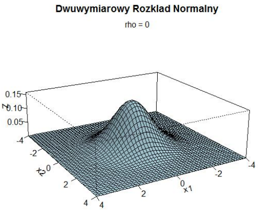
$$\Sigma = \begin{bmatrix} \mathrm{cov}(X, X) & \mathrm{cov}(X, Y) \\ \mathrm{cov}(Y, X) & \mathrm{cov}(Y, Y) \end{bmatrix}$$

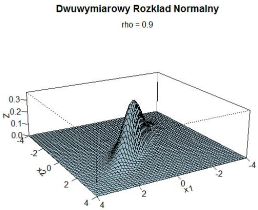
$$= \begin{bmatrix} \mathrm{Var}(X) & \mathrm{cov}(X, Y) \\ \mathrm{cov}(Y, X) & \mathrm{Var}(Y) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathrm{I}_2$$

Uwaga. Zobacz Przykład 1a) w skrypcie 07GaussWielowymiarowy-wyklad.R

Rozkład normalny dwuwymiarowy $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$







Gęstość

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right)$$

Wartość oczekiwana: $\mu = (EX, EY) = (\mu_1, \mu_2)$

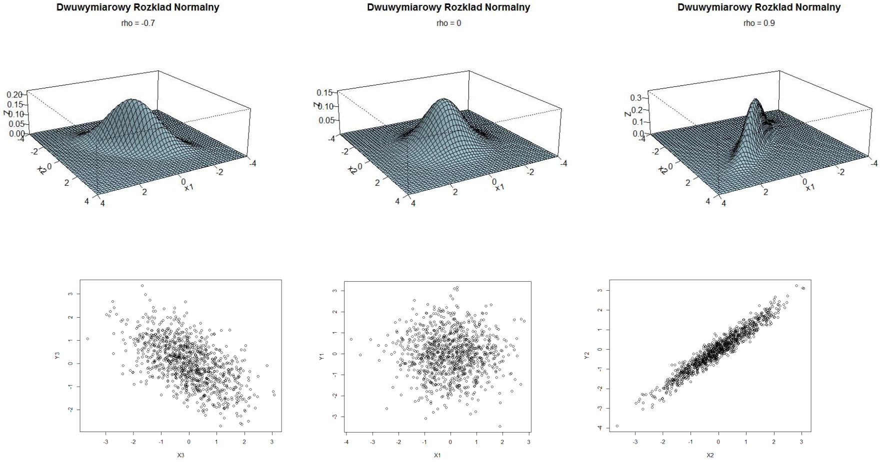
Macierz kowariancji

$$\Sigma = \begin{bmatrix} cov(X,X) & cov(X,Y) \\ cov(Y,X) & cov(Y,Y) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

gdzie σ_1, σ_2 to odchylenia standardowe i ρ współczynnik korelacji

Rozkład normalny dwuwymiarowy

Gęstości i próby wygenerowane z rozkładów: $N(0, 0, 1, 1, -0.7)$, $N(0, 0, 1, 1, 0)$ i $N(0, 0, 1, 1, 0.9)$.



Uwaga. Zobacz Przykład 1b) w skrypcie 07GaussWielowymiarowy-wyklad.R

Rozkłady brzegowe

Fakt. Jeśli $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$, to $X \sim N(\mu_1, \sigma_1)$ i $Y \sim N(\mu_2, \sigma_2)$

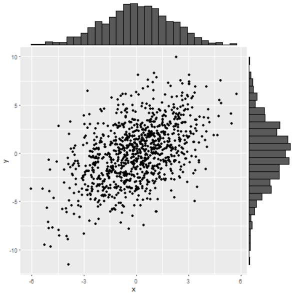
$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)dy = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2}}, \quad x \in \mathbb{R}$$

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)dx = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(y-\mu_2)^2}{2}}, \quad y \in \mathbb{R}$$

gdzie f_1, f_2, f są odpowiednio gęstościami zmiennych losowych X, Y oraz wektora (X, Y).

Na wykresie obok próba liczności 1000, wygenerowana z rozkładu $N(0, 0, 2, 3, 0.5)$ wraz z histogramami rozkładów brzegowych.

Uwaga. Zobacz Przykład 2 w skrypcie 07GaussWielowymiarowy-wyklad.R



Estymator kowariancji

Mamy próbę $(X_1, Y_1), \ldots, (X_n, Y_n)$ prostą z rozkładu o nieznanej kowariancji i nieznanym współczynniku korelacji ρ.

Estymator kowariancji

$$\tilde{S}_{xy} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X_n})(Y_i - \overline{Y_n}),$$

gdzie $\overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i$, $\overline{Y_n} = \frac{1}{n} \sum_{i=1}^{n} Y_i$.

Estymator $\tilde{S}_{xy}$ jest obciążonym estymatorem kowariancji, nieobciążonym jest

$$S_{xy} = \frac{n}{n-1} \tilde{S}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)(Y_i - \overline{Y}_n).$$

Estymator współczynnika korelacji

Estymator współczynnika korelacji

$$\hat{\rho} = R_{xy} = \frac{S_{xy}}{S_x \cdot S_y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X_n})(Y_i - \overline{Y_n})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X_n})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y_n})^2}},$$

gdzie S_x^2 i S_y^2 są nieobciążonymi estymatorami wariancji

$$S_x^2 = \hat{\sigma}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2, \qquad S_y^2 = \hat{\sigma}_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y_n})^2.$$

Estymatorami największej wiarygodności wektora μ i macierzy kowariancji Σ, dwuwymiarowego rozkładu normalnego są $\hat{\mu} = (\overline{X_n}, \overline{Y_n})$ oraz

$$\widehat{\Sigma} = \begin{bmatrix} \tilde{S}_x^2 & \tilde{S}_{xy} \\ \tilde{S}_{xy} & \tilde{S}_y^2 \end{bmatrix}, \tag{1}$$

gdzie $\tilde{S}_x^2$ i $\tilde{S}_y^2$ są obciążonymi estymatorami wariancji

$$\tilde{S}_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2, \qquad \tilde{S}_y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y_n})^2.$$

Jest to estymator obciążony ($E(\widehat{\Sigma}) = \frac{n-1}{n}\Sigma$), nieobciążonym estymatorem jest

$$\widehat{\Sigma} = \begin{bmatrix} S_x^2 & S_{xy} \\ S_{xy} & S_y^2 \end{bmatrix}. \tag{2}$$

Uwaga. Zobacz Przykład 3a) w skrypcie 07GaussWielowymiarowy-wyklad.R

Odległość Mahalanobisa

Dla danego wektora $\mathrm{x} = (x_1, x_2)$, wektora średnich $\mu = (\mu_1, \mu_2)$ oraz macierzy kowariancji Σ, definiujemy odległość Mahalanobisa wektora x od wektora μ, jako

$$d(\mathrm{x}, \mu) = \left(\begin{bmatrix} x_1 - \mu_1 & x_2 - \mu_2 \end{bmatrix} \Sigma^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right)^{1/2}$$
$$= \left((\mathrm{x} - \mu)^T \Sigma^{-1} (\mathrm{x} - \mu) \right)^{1/2},$$

gdzie Σ^{-1} jest macierzą odwrotną do macierzy Σ.

Jeśli macierz Σ jest macierzą jednostkową, to odległość Mahalanobisa jest zwykłą odległością euklidesową

$$d(\mathrm{x}, \mu) = \left((\mathrm{x} - \mu)^T (\mathrm{x} - \mu) \right)^{1/2} = \sqrt{(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2}.$$

Uwaga. Zobacz

- Przykład 3b) w skrypcie 07GaussWielowymiarowy-wyklad.R
- https://pl.wikipedia.org/wiki/Odległość_Mahalanobisa

Testowanie normalności rozkładu dwuwymiarowego

Mamy realizację $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ próby z rozkładu dwuwymiarowego o nieznanej dystrybuancie F, wartości oczekiwanej $\mu = (\mu_1, \mu_2)$ i macierzy kowariancji Σ. Chcemy zweryfikować hipotezę o normalności rozkładu F. Jedna z metod weryfikacji wykorzystuje odległość Mahalanobisa.

Fakt. Jeśli $\mathrm{X} = (X_1, X_2) \sim N(\mu, \Sigma)$, to kwadrat odległości Mahalanobisa wektoraz X od wektora μ ma rozkład chi-kwadrat o $n = 2$ stopniach swobody

$$d^2(\mathrm{X}, \mu) = (\mathrm{X} - \mu)^T \Sigma^{-1} (\mathrm{X} - \mu) \sim \chi^2(2). \tag{3}$$

Mając dane $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, obliczamy

- $\hat{\mu}$ oraz macierz kowariancji $\hat{\Sigma}$,
- $d_1, d_2, \ldots, d_n$ – kwadraty odległości Mahalanobisa punktów $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ od $\hat{\mu}$.

Odległości $d_1, d_2, \ldots, d_n$ wykorzystujemy do utworzenia wykresów diagnostycznych oraz weryfikacji hipotezy, że (3) ma rozkład $\chi^2(2)$. Odrzucenie tej hipotezy, daje podstawy do odrzucenia hipotezy o normalności F. Brak odrzucenia – nie daje takich podstaw.

Zobacz Przykład 3c)d) w skrypcie 07GaussWielowymiarowy-wyklad.R

Testowanie normalności rozkładu – test Mardii

Klasycznym testem normalności, dla rozkładu dwuwymiarowego, jest test Mardii oparty na teście skośności i kurtozy rozkładu. Statystyki testowe to

$$A = \frac{1}{6n} \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij}^3, \quad B = \sqrt{n} \left(\frac{1}{8n} \sum_{i=1}^{n} D_i^2 - 1 \right),$$

gdzie

$$D_{ij} = \begin{bmatrix} x_i - \mu_1 & y_i - \mu_2 \end{bmatrix} \widehat{\Sigma}^{-1} \begin{bmatrix} x_j - \mu_1 \\ y_j - \mu_2 \end{bmatrix},$$

$$D_i = \begin{bmatrix} x_i - \mu_1 & y_i - \mu_2 \end{bmatrix} \widehat{\Sigma}^{-1} \begin{bmatrix} x_i - \mu_1 \\ y_i - \mu_2 \end{bmatrix}.$$

Jeśli hipoteza zerowa o normalności rozkładu dwuwymiarowego jest prawdziwa, to asymptotyczne rozkłady statystyk wynoszą

$$A \sim \chi^2(4), \qquad B \sim N(0, 1).$$

Duże wartości statystyk świadczą przeciwko hipotezie zerowej.

Uwaga. Zobacz Przykład 4 w skrypcie 07GaussWielowym-wyklad.R