technische universität dortmund

Statistical Methods of Data Analysis

Multivariate Distributions

Prof. Dr. Dr. Wolfgang Rhode   Dr. Maximilian Linhoff

2023

Overview

Two-dimensional distributions

Multivariate distributions

Two-dimensional Gaussian distribution

Theorems and propositions

Overview

Two-dimensional distributions

- Given: Stochastic variables X and Y
- Wanted: $P((X < x) \wedge (Y < y))$
- Similar to the one-dimensional case, the CDF is given by

$$F(x, y) = P(X < x, Y < y)$$

Two-dimensional PDF and CDF

- If F is continuously differentiable, then it follows

$$f(x, y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F(x, y)$$

$$P(a \leq x < b, c \leq y < d) = \int_a^b \int_c^d f(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

Marginal distribution

- PDF of the marginal distributions

$$g(x) = \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}y$$
$$h(y) = \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}x$$

- It follows

$$P(a \leq x < b, -\infty \leq y < \infty) = \int_a^b \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}y\ \mathrm{d}x = \int_a^b g(x)\, \mathrm{d}x$$

Conditional probability

- Conditional probability for Y = conditional probability for Y with known X

$$P(y \leq Y \leq y + \mathrm{d}y \,|\, x \leq X \leq x + \mathrm{d}x)$$

- Corresponding PDF

$$f(y|x) = \frac{f(x,y)}{g(x)}$$

- Corresponding marginal distribution

$$h(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x = \int_{-\infty}^{\infty} f(y|x)\,g(x)\,\mathrm{d}x$$

Stochastic independence

- Stochastic variables X and Y are independent, if

$$f(x,y) = g(x) \cdot h(y)$$

- Thus, with independence

$$f(y|x) = \frac{f(x,y)}{g(x)} = \frac{g(x) \cdot h(y)}{g(x)} = h(y)$$

- Be aware: Independence $\neq$ No correlation

  Independence $\not\leftarrow$ No correlation

  Independence $\rightarrow$ No correlation

Expected value

- Analogous to the one-dimensional case

$$\mathrm{E}[H(x,y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} H(x,y)\,f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

- Example: $H(x,y) = ux + by$

$$\mathrm{E}[ux + by] = u\,\mathrm{E}[x] + b\,\mathrm{E}[y]$$

Variance

$$\sigma^2[H(x,y)] = \mathrm{E}\Big[H(x,y) - \mathrm{E}[H(x,y)]^2\Big]$$

- Example: $H(x,y) = ax + by$

$$\begin{aligned}
\sigma^2(ax+by) &= \mathrm{E}\Big[((ax+by) - \mathrm{E}[ax+by])^2\Big] \\
&= \mathrm{E}\Big[(a(x-\bar{x}) + b(y-\bar{y}))^2\Big] \\
&= \mathrm{E}\Big[a^2(x-\bar{x})^2 + b^2(y-\bar{y})^2 + 2ab(x-\bar{x})(y-\bar{y})\Big] \\
&= a^2\sigma^2(x) + b^2\sigma^2(y) + 2ab\cdot\mathrm{Cov}(x,y)
\end{aligned}$$

- Example: $H(x,y) = xy$ with x, y independent

$$\mathrm{E}[xy] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, g(x)h(y)\,\mathrm{d}x\,\mathrm{d}y = \int_{-\infty}^{\infty} xg(x)\,\mathrm{d}x \cdot \int_{-\infty}^{\infty} yh(y)\,\mathrm{d}y = \mathrm{E}[x]\cdot\mathrm{E}[y]$$

Overview

Two-dimensional distributions

Multivariate distributions

Two-dimensional Gaussian distribution

Theorems and propositions

Multivariate distributions

- CDF

$$F(x_1, x_2, ..., x_n) = P(X_1 < x_1, X_2 < x_2, ..., X_n < x_n)$$

- PDF

$$f(x_1, x_2, .., x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 ... \partial x_n} F(x_1, x_2, ..., x_n)$$

- Marginal distribution of one variable

$$g(x_r) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(x_1, x_2, ..., x_n)\, \mathrm{d}x_1\, \mathrm{d}x_2 \ldots \mathrm{d}x_{r-1}\, \mathrm{d}x_{r+1} \ldots \mathrm{d}x_n$$

Multivariate distributions

- Expected value

$$\mathrm{E}\left[H(x_1, \ldots, x_n)\right] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} H(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\,\mathrm{d}x_1 \ldots \mathrm{d}x_n$$

- Variance

$$\sigma^2[H(x_1, \ldots, x_n)] = \mathrm{E}\left[\left(H(x_1, \ldots, x_n) - \mathrm{E}\left[H(x_1, \ldots, x_n)\right]\right)^2\right]$$

Covariance

- Definition

$$\mathrm{Cov}(X_i, X_j) = \mathrm{E}\left[(X_i - \mathrm{E}[X_i]) \cdot (X_j - \mathrm{E}[X_j])\right]$$

- Alternative

$$\mathrm{Cov}(X_i, X_j) = \mathrm{E}\left[X_i \cdot X_j\right] - \mathrm{E}[X_i]\,\mathrm{E}[X_j]$$

Covariance

The covariance is

- positive, if $\begin{cases} X_i > \mathrm{E}[X_i] & X_j > \mathrm{E}[X_j] \\ X_i < \mathrm{E}[X_i] & X_j < \mathrm{E}[X_j] \end{cases}$
- negative, if $\begin{cases} X_i > \mathrm{E}[X_i] & X_j < \mathrm{E}[X_j] \\ X_i < \mathrm{E}[X_i] & X_j > \mathrm{E}[X_j] \end{cases}$
- zero, if X_i and X_j are independent

Pearson correlation coefficient

- Rough measure of the dependence of two stochastic variables

$$\rho(X_i, X_j) = \frac{\mathrm{Cov}(X_i, X_j)}{\sigma(X_i) \cdot \sigma(X_j)}$$

- Two stochastic variables are uncorrelated, if $\rho(X_i, X_j) = 0$

Covariance matrix

- General

$$\begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \dots & \mathrm{Cov}(X_1, X_n) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \dots & \mathrm{Cov}(X_n, X_n) \end{pmatrix}$$

- Non-correlation

$$\begin{pmatrix} \sigma^2(X_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2(X_n) \end{pmatrix}$$

Multivariate Gaussian distribution

- Vector with $\boldsymbol{n}$ variables $\boldsymbol{X} = (x_1, x_2, ..., x_n)$
- PDF

$$f(\boldsymbol{X}) = k_n \cdot \mathrm{e}^{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^\top \boldsymbol{B}(\boldsymbol{X}-\boldsymbol{\mu})} = k_n \cdot \mathrm{e}^{-\frac{1}{2} g(\boldsymbol{X})}$$

$$k_n = \left(\frac{\det B}{(2\pi)^n}\right)^{\frac{1}{2}}$$

- $\boldsymbol{\mu}$ refers to the vector of the expected values
- B is a symmetric and positive definite $\boldsymbol{n} \times \boldsymbol{n}$ matrix

Properties

- PDF symmetric with respect to $\boldsymbol{X} = \boldsymbol{\mu}$
- Expected values

$$\mathrm{E}[\boldsymbol{X} - \boldsymbol{\mu}] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(\boldsymbol{X})\, \mathrm{d}\boldsymbol{x} = 0$$

$$\mathrm{E}[\boldsymbol{X}] = \boldsymbol{\mu}$$

- Covariance matrix

$$\Sigma = \mathrm{E}\left[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{\top}\right] = B^{-1}$$

Standardized variable

- If stochastic variables are not independent, it is reasonable to use standardized variables

$$u_i = \frac{x_i - a_i}{\sigma_i}, i = 1, 2, \ldots$$

$$\Phi(u_1, u_2) = k_n \cdot \mathrm{e}^{-\frac{1}{2}\boldsymbol{u}^\top B \boldsymbol{u}} = k_n \cdot \mathrm{e}^{-\frac{1}{2} g(\boldsymbol{u})}$$

$$\rho = \frac{\mathrm{Cov}(x_1, x_2)}{\sigma_1 \sigma_2} = \mathrm{Cov}(u_1, u_2)$$

$$B = \frac{1}{1 - \rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$$

Overview

Definition and properties

$$\Sigma = B^{-1} = \begin{pmatrix} \sigma_1^2 & \mathrm{Cov}(x_1, x_2) \\ \mathrm{Cov}(x_1, x_2) & \sigma_2^2 \end{pmatrix}$$

$$B = \frac{1}{\sigma_1^2\sigma_2^2 - \mathrm{Cov}(x_1, x_2)^2} \begin{pmatrix} \sigma_2^2 & -\mathrm{Cov}(x_1, x_2) \\ -\mathrm{Cov}(x_1, x_2) & \sigma_1^2 \end{pmatrix}$$

- With independent variables

$$B = B_0 = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix}$$

Definition and properties

- PDF

$$\Phi(x) = k \cdot e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top} B_0 (\boldsymbol{x}-\boldsymbol{\mu})} = k \cdot e^{-\frac{1}{2}\frac{(x_1-\mu_1)^2}{\sigma_1^2}} e^{-\frac{1}{2}\frac{(x_2-\mu_2)^2}{\sigma_2^2}}$$

- With normalization

$$k = k_0 = \frac{1}{2\pi\sigma_1\sigma_2}$$

Definition and properties

- Lines of equal probability → contour lines

$$\Phi(u_1, u_2) = \text{const} \rightarrow -\frac{1}{2}g(\boldsymbol{u}) = \text{const}$$

$$\rightarrow -\frac{1}{2}\frac{1}{1-\rho^2}\left(u_1^2 + u_2^2 - 2u_1u_2\rho\right) = \text{const}$$

- Consider $\boldsymbol{g(u) = 1}$, then we obtain

$$\frac{(x_1-\mu_1)}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)}{\sigma_1}\cdot\frac{(x_2-\mu_2)}{\sigma_2} + \frac{(x_2-\mu_2)}{\sigma_2^2} = 1-\rho^2$$

- This is an elliptic equation!

Properties of the ellipse

- Mean value u_1, u_2
- Angle α between the ellipse major axis and the coordinate axis
- Length p_1, p_2 of the ellipse semi-major / semi-minor axes

Special ellipse

- Covariance ellipse

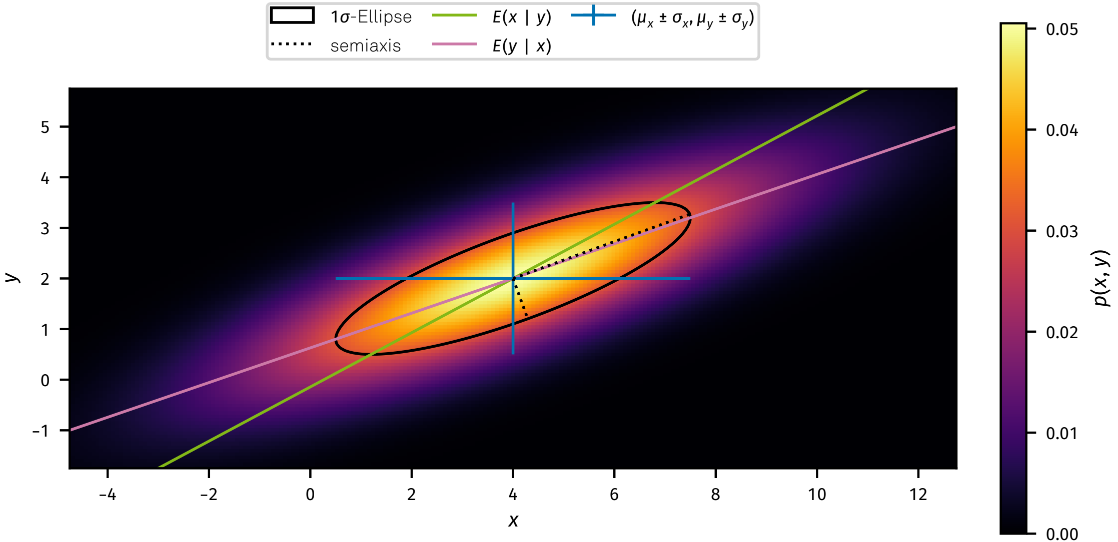
$$\alpha = \frac{1}{2}\arctan\left(\frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}\right)$$

$$p_1^2 = (1-\rho^2)\left(\frac{\cos^2(\alpha)}{\sigma_1^2} - \frac{2\rho\sin(\alpha)\cos(\alpha)}{\sigma_1\sigma_2} + \frac{\sin^2(\alpha)}{\sigma_2^2}\right)^{-1}$$

$$p_2^2 = (1-\rho^2)\left(\frac{\sin^2(\alpha)}{\sigma_1^2} + \frac{2\rho\sin(\alpha)\cos(\alpha)}{\sigma_1\sigma_2} + \frac{\cos^2(\alpha)}{\sigma_2^2}\right)^{-1}$$

- Within the **1 σ**-range, there is about 40 % probability content in the two-dim. case

technische universität
dortmund



Two-dimensional Gaussian distribution

Overview

Two-dimensional distributions

Multivariate distributions

Two-dimensional Gaussian distribution

Theorems and propositions

Theorems and propositions

- Chebyshev's inequality
- Law of large numbers
- Central limit theorem

Chebyshev’s inequality

- Upper bound for the probability that a stochastic variable deviates more than k standard deviations from the mean value
- The probability for a stochastic variable originating from $|x - \langle x\rangle| \geq k\sigma$ is given by

$$\int_{-\infty}^{\langle x\rangle - k\sigma} f(x)\,\mathrm{d}x + \int_{\langle x\rangle + k\sigma}^{\infty} f(x)\,\mathrm{d}x \leq \frac{1}{k^2}$$

- Applies under very general conditions (for all PDFs)
- However, is a very weak condition in return

Derivation

- The derivation is based on the definition of variance

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \langle x\rangle)^2 \cdot f(x)\,\mathrm{d}x$$
$$= \left(\int_{-\infty}^{\langle x\rangle - k\sigma} + \int_{\langle x\rangle - k\sigma}^{\langle x\rangle + k\sigma} + \int_{\langle x\rangle + k\sigma}^{\infty}\right)(x - \langle x\rangle)^2 \cdot f(x)\,\mathrm{d}x$$

- Omitting the middle term then leads to an inequality

$$\sigma^2 \geq \left(\int_{-\infty}^{\langle x\rangle - k\sigma} + \int_{\langle x\rangle + k\sigma}^{\infty}\right)(x - \langle x\rangle)^2 \cdot f(x)\,\mathrm{d}x$$

Derivation

- For the integrals is now valid due to the limits

$$x < \langle x \rangle - k\sigma \qquad\qquad x > \langle x \rangle + k\sigma$$

$$x - \langle x \rangle < -k\sigma \qquad\qquad x - \langle x \rangle > k\sigma$$

$$(x - \langle x \rangle) > k^2\sigma^2 \qquad\qquad (x - \langle x \rangle) > k^2\sigma^2$$

- Substituting then yields the inequality

$$\sigma^2 \geq k^2\sigma^2\left(\int_{-\infty}^{\langle x \rangle - k\sigma} f(x)\,\mathrm{d}x + \int_{\langle x \rangle + k\sigma}^{\infty} f(x)\,\mathrm{d}x\right)$$

Law of large numbers

- Let n independent experiments be given, in which the event j has occurred n_j times
- Let the n_j be binomially distributed and let $h_j = n_j / n$ be the corresponding stochastic variable
- From this follows for the expected value of h_j

$$E\left[h_j\right] = \frac{1}{n} E\left[n_j\right] = p_j$$

- How does this estimate the unknown probability p_j?

Law of large numbers

- Compute variance of h_j

$$V(h_j) = \sigma^2(h_j) = \sigma^2(n_j/n) = \frac{1}{n^2}\sigma^2(n_j) = \frac{1}{n^2}np_j(1-p_j)$$

- Since $p_j(1-p_j) \leq \frac{1}{4}$ stands, it applies for the variance $\sigma^2(h_j) \leq \frac{1}{4n}$
- Thus, for large numbers ($n \to \infty$), the error of the estimate h_j can be made as small as desired
- The error is bounded by $1/2\sqrt{n}$

Central limit theorem

- The probability density of the sum $\omega = \sum_{i=1}^{n} x_i$ of a sample of n independent stochastic variables x_i with an arbitrary probability density with the mean $\langle x \rangle$ and variance σ^2 approaches a Gaussian distribution with the mean $\langle \omega \rangle = n \cdot \langle x \rangle$ and a variance $V(\omega) = n\sigma^2$ in the limiting case $n \to \infty$
- Quantities based on sums of randomly distributed events are Gaussian distributed

Central limit theorem