Vorlesung
Statistische Methoden der Datenanalyse
Prof. Dr. Dr. Wolfgang Rhode

Wahrscheinlichkeitsverteilungen

Überblick

- Definitionen von Wahrscheinlichkeiten
- Kombination von Wahrscheinlichkeiten
- Eindimensionale Verteilungen
  - Verteilungsfunktion und Wahrscheinlichkeitsdichte
  - Momente
  - Regeln über Mittelwerte und Varianzen
  - Bestimmte eindimensionale Verteilungen
- Mehrdimensionale Verteilungen
  - Verteilungsfunktion und Wahrscheinlichkeitsdichte
  - Bedingte Wahrscheinlichkeit und Randverteilungen
  - Erwartungswert, Varianz, Kovarianz
  - Unabhängigkeit und Korrelation
  - Bestimmte mehrdimensionale Verteilungen

Und wie sieht es mit mehrdimensionalen Verteilungen aus?

- Zunächst der **zweidimensionale** Fall:
  - Gegeben: Zufallsvariablen X und Y
  - Gesucht: $P((X<x)\wedge(Y<y))$
- Analog zum eindimensionalen Fall, ist Verteilungsfunktion definiert durch:

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

Zweidim. Wahrscheinlichkeitsdichte und Verteilungsfunktion

- *F* sei stetig differenzierbar, dann

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

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

Randverteilungen

- Wahrscheinlichkeitsdichten der Randverteilungen:

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

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

- Daraus folgt beispielsweise:

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

Bedingte Wahrscheinlichkeit

- Bedingte Wahrscheinlichkeit für *Y*
  = Wahrscheinlichkeit für *Y* bei bekanntem *X*

$$P(y \leq Y \leq y + dy | x \leq X \leq x + dx)$$

- Entsprechende Wahrscheinlichkeitsdichte:

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

- Daraus folgt für die Randverteilung:

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

Stochastische Unabhängigkeit

- Zufallsvariablen *X* und *Y* sind **unabhängig**, wenn gilt:

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

- Bei Unabhängigkeit gilt also:

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

- Achtung:

$$\text{Unabhängigkeit} \neq \text{Unkorreliertheit}$$
$$\text{Unabhängigkeit} \not\leftarrow \text{Unkorreliertheit}$$
$$\text{Unabhängigkeit} \rightarrow \text{Unkorreliertheit}$$

Erwartungswert

- Analog zu 1-dim. Fall:

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

- Beispiel: *H(x,y)=ux+by*

$$E(ux + by) = uE(x) + bE(y)$$

Varianz

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

- Beispiel: *H(x,y)=ax+by*

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

- Beispiel: *H(x,y)=xy* mit *x,y* unabh.

$$E[xy] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, g(x)h(y)\, dxdy = \int_{-\infty}^{\infty} x\, g(x)\, dx \cdot \int_{-\infty}^{\infty} y\, h(y)\, dy = E[x]\cdot E[y]$$

Mehrdimensionale Verteilungen

- Noch einen Schritt weiter...
  Die *n*-dim Verteilungen!
- Verteilungsfunktion:

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

- Wahrscheinlichkeitsdichte:

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

Mehrdimensionale Verteilungen

- Randverteilung einer Variablen

$$g(x_r) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, x_2, ..., x_n) dx_1 dx_2 ... dx_{r-1} dx_{r+1} ... dx_n$$

Mehrdimensionale Verteilungen

- Erwartungswert:

$$E[H(x_1, \dots, x_n)] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} H(x_1, \dots, x_n) f(x_1, \dots, x_n) dx_1 \dots dx_n$$

Mehrdimensionale Verteilungen

- Varianz:

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

Kovarianz

- Es gilt:

$$cov(X_i, X_j) = E[(X_i - E[X_i]) \cdot (X_j - E[X_j])]$$

- Alternativ (über den Verschiebungssatz):

$$cov(X_i, X_j) = E[X_i \cdot X_j] - E[X_i]E[X_j]$$

Kovarianz

Die Kovarianz ist:

- positiv, wenn $X_i > (<)\ E[X_i]$ mit $X_j > (<)\ E[X_j]$;
- negativ, wenn $X_i > (<)\ E[X_i]$ mit $X_j < (>)\ E[y]$
- =0, wenn $X_i,\ X_j$ unabhängig sind.

Korrelationskoeffizient

- Grobes Maß für die Abhängigkeit zweier Zufallsvariablen

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

- Unkorreliertheit zweier Variablen, wenn der Korrelationskoeffizient Null ist

Kovarianzmatrix

- Allgemein:

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

- Bei Unkorreliertheit:

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

Mehrdimensionale Gaußverteilung

- Vektor mit *n* Variablen $\overrightarrow{X} = (x_1, x_2, ..., x_n)$
- Wahrscheinlichkeitsdichte:

$$f\left(\vec{X}\right) = k \cdot e^{-\frac{1}{2}\left(\vec{X}-\vec{a}\right)^\top B\left(\vec{X}-\vec{a}\right)} = k \cdot e^{-\frac{1}{2}g\left(\vec{X}\right)}$$

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

$$\overrightarrow{a}: \quad n - \text{Komponenten Vektor,}$$

$$B: \quad n \times n - \text{Matrix, symmetrisch und positiv definit}$$

Mehrdimensionale Gaußverteilung

- Eigenschaften
  - Wahrscheinlichkeitsdichte symmetrisch um $\vec{X} = \vec{a}$
  - Erwartungswerte:

$$E(\vec{X} - \vec{a}) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f(\vec{X}) d\vec{x} = 0$$

$$E(\vec{X}) = \vec{a}$$

  - Kovarianzmatrix *C*:

$$C = E[(\vec{X} - \vec{a})(\vec{X} - \vec{a})^\top] = B^{-1}$$

Mehrdimensionale Gaußverteilung

- Wenn Zufallsvariablen abhängig voneinander:
  Übergang zu standardisierten Variablen sinnvoll

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

$$\phi(u_1, u_2) = k \cdot e^{-\frac{1}{2}\vec{u}^\top B \vec{u}} = k \cdot e^{-\frac{1}{2}g(\vec{u})}$$

Mehrdimensionale Gaußverteilung

- Bei Verwendung standardisierter Variablen:

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

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

technische universität dortmund
Experimentelle Physik Vb
Astroteilchenphysik

Zweidimensionale Gaußverteilung

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

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

- Wenn Kovarianzen verschwinden: diagonale Matrizen

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

Prof. Dr. Dr. W. Rhode
Wahrscheinlichkeitsverteilungen
Statistische Methoden der Datenanalyse

technische universität dortmund
Experimentelle Physik Vb
Astroteilchenphysik

Zweidimensionale Gaußverteilung

- Daraus folgt die Wahrscheinlichkeitsdichte:

$$\phi(x) = k \cdot e^{-\frac{1}{2}(\vec{X}-\vec{a})^\top B_0 (\vec{X}-\vec{a})} = k \cdot e^{-\frac{1}{2}\frac{(x_1-a_1)^2}{\sigma_1^2}} e^{-\frac{1}{2}\frac{(x_2-a_2)^2}{\sigma_2^2}},$$

- Mit der Normierung:

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

Prof. Dr. Dr. W. Rhode
Wahrscheinlichkeitsverteilungen
Statistische Methoden der Datenanalyse

technische universität dortmund
Experimentelle Physik Vb
Astroteilchenphysik

Zweidimensionale Gaußverteilung

- Linien gleicher Wahrscheinlichkeit als Höhenlinien

$$\phi(u_1, u_2) = \mathrm{const} \;\Rightarrow\; -\frac{1}{2} g(\vec{u}) = \mathrm{const}$$

$$\Rightarrow\; -\frac{1}{2}\frac{1}{1-\rho^2}(u_1^2 + u_2^2 - 2u_1 u_2 \rho) = \mathrm{const}$$

Prof. Dr. Dr. W. Rhode
Wahrscheinlichkeitsverteilungen
Statistische Methoden der Datenanalyse

Zweidimensionale Gaußverteilung

- Betrachte $g(\vec{u}) = 1$, so ergibt sich:

$$\frac{(x_1 - a_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1 - a_1)}{\sigma_1} \cdot \frac{(x_2 - a_2)}{\sigma_2} + \frac{(x_2 - a_2)^2}{\sigma_2^2} = 1 - \rho^2$$

- Dies ist eine Ellipsengleichung!

technische universität dortmund
Experimentelle Physik Vb
Astroteilchenphysik

Zweidimensionale Gaußverteilung

- Eigenschaften der Ellipse:
  - Mittelwert $u_1 u_2$
  - Winkel α zwischen Ellipsen-Hauptachsen und Koordinatenachsen
  - Halbmesser $p_1,\ p_2$ der Ellipsen-Hauptachsen

Prof. Dr. Dr. W. Rhode
Wahrscheinlichkeitsverteilungen
Statistische Methoden der Datenanalyse

technische universität dortmund
Experimentelle Physik Vb
Astroteilchenphysik

Zweidimensionale Gaußverteilung

- Spezielle Ellipse: Kovarianzellipse

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

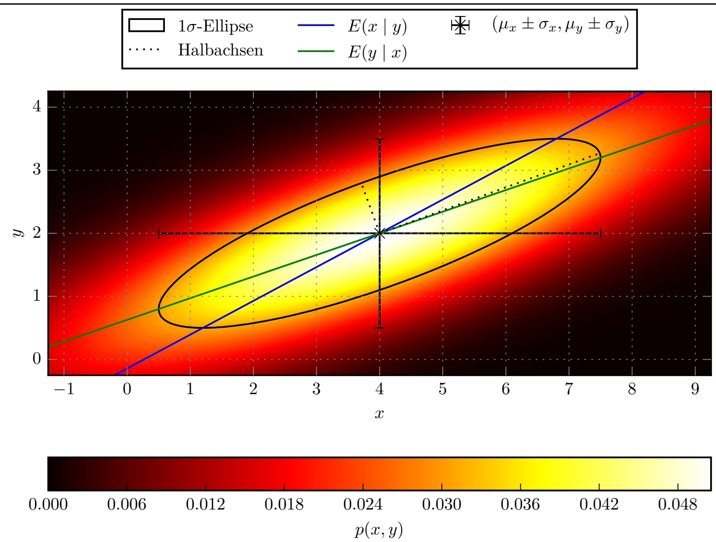
$$p_1^2 = (1-\rho^2)\left(\frac{\cos^2\alpha}{\sigma_1^2} - \frac{2\rho\sin\alpha\cos\alpha}{\sigma_1\sigma_2} + \frac{\sin^2\alpha}{\sigma_2^2}\right)^{-1}$$

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

- Innerhalb des 1σ-Bereichs liegen analog zum eindim. Fall 68,39%

Prof. Dr. Dr. W. Rhode
Wahrscheinlichkeitsverteilungen
Statistische Methoden der Datenanalyse

technische universität dortmund
Experimentelle Physik Vb
Astroteilchenphysik



Prof. Dr. Dr. W. Rhode
Wahrscheinlichkeitsverteilungen
Statistische Methoden der Datenanalyse

Theoreme und Sätze

- Tschebyscheff-Ungleichung
- Gesetz der großen Zahlen
- Zentraler Grenzwertsatz

Tschebyscheff-Ungleichung

- Obere Schranke für die Wahrscheinlichkeit, dass eine Zufallsvariable mehr als *k* Standardabweichungen vom Mittelwert abweicht
- Für die Wahrscheinlichkeit, dass eine Zufallsvariable einen Wert aus dem Bereich $|x - \langle x \rangle| \geq k\sigma$ stammt, ist gegeben durch:

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

- Gilt unter sehr allgemeinen Bedingungen (für alle Wahrscheinlichkeitsdichten)
- Ist jedoch im Gegenzug eine sehr schwache Bedingung

Tschebyscheff-Ungleichung – Herleitung

- Die Herleitung basiert auf der Definition der Varianz:

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

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

- Das Weglassen des mittleren Terms führt dann auf eine Ungleichung:

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

Tschebyscheff-Ungleichung – Herleitung

- Für die Integrale gilt nun aufgrund der Grenzen:

$$\begin{aligned} x &< \langle x \rangle - k\sigma \\ x - \langle x \rangle &< -k\sigma \\ (x - \langle x \rangle)^2 &> k^2\sigma^2 \end{aligned} \quad \text{und} \quad \begin{aligned} x &> \langle x \rangle + k\sigma \\ x - \langle x \rangle &> k\sigma \\ (x - \langle x \rangle)^2 &> k^2\sigma^2 \end{aligned}$$

- Einsetzen liefert dann die Ungleichung:

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

Gesetz der großen Zahlen

- Gegeben seien *n* unabhängige Experimente, in denen das Ereignis *j* n_j mal aufgetreten ist
- Die n_j seien binomialverteilt und $h_j = n_j / n$ sei die entsprechende Zufallsvariable
- Dann gilt für den Erwartungswert von h_j: $E(h_j) = \frac{1}{n} E(n_j) = p_j$
- Wie genau wird die unbekannte Wahrscheinlichkeit p_j damit geschätzt?
- Berechne die Varianz von h_j:
$$V(h_j) = \sigma^2(h_j) = \sigma^2(n_j/n) = \frac{1}{n^2} \cdot \sigma^2(n_j) = \frac{1}{n^2} n p_j (1 - p_j)$$
- Da $p_j(1-p_j) \leq \frac{1}{4}$ ist, gilt für die Varianz: $\sigma^2(h_j) \leq \frac{1}{4n}$
- Somit kann für große Zahlen ($n \to \infty$) der Fehler der Schätzung h_j so klein gemacht werden wie gewünscht. Der Fehler ist durch $1/2\sqrt{n}$ beschränkt

technische universität dortmund
Experimentelle Physik Vb
Astroteilchenphysik

Der zentrale Grenzwertsatz

- Die Wahrscheinlichkeitsdichte der Summe $\omega = \sum_{i=1}^{n} x_i$ einer Stichprobe aus n unabhängigen Zufallsvariablen x_i mit einer beliebigen Wahrscheinlichkeitsdichte mit dem Mittelwert $\langle x \rangle$ und der Varianz σ^2 geht im Grenzfall $n \to \infty$ gegen eine Gaußverteilung mit dem Mittelwert $\langle \omega \rangle = n \cdot \langle x \rangle$ und einer Varianz $V(\omega) = n\sigma^2$
- Größen, die auf Summen von zufallsverteilten Ereignissen basieren sind gaußverteilt