Vorlesung 12

Inhalt

- Unabhängigkeit
- Erwartungswert, Varianz
- Kovarianz und Korrelation

Lernziele

- Das Konzept der stochastischen Unabhängigkeit kenne
- Erwartungswert, Varianz, Kovarianz und Korrelation berechnen können

Benötigte Vorkenntnisse

- Grundlagen Wahrscheinlichkeitsrechnung

Stochastische Unabhängigkeit

Definition

Zwei Zufallsgrößen X und Y heißen (stochastisch) unabhängig, wenn für alle Ereignisse $\{X \in A\}$, $\{Y \in B\}$ gilt

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B)$$

Beispiel:

- Werfen zweier Würfel:
  $X =$ Augenzahl Würfel 1, $Y =$ Augenzahl Würfel 2.

$$\mathbb{P}(X = 2,\ Y = 5) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = \mathbb{P}(X = 2) \cdot \mathbb{P}(Y = 5)$$

Stochastische Unabhängigkeit

In der Praxis wendet man häufig Resultate an, die Unabhängigkeit einer Stichprobe voraussetzen.

Beispiele:

- Für eine Studie wird eine zufällige Person in München und eine zufällige Person in Hamburg befragt. Die Antworten dürfen als unabhängig voneinander angenommen werden.
- Befragt man zwei Schwestern oder nahe Verwandte (getrennt voneinander), so werden die Antworten nicht unabhängig voneinander sein.

Definition (Erwartungswert)

Sei X eine Zufallsvariable mit endlichem oder aufzählbarem Wertebereich $\mathcal{S} = \{a_1, a_2, a_3 \dots\} \subseteq \mathbb{R}$. Dann ist der *Erwartungswert* von X definiert durch

$$\mathbb{E}[X] = \sum_{a \in \mathcal{S}} a \cdot \Pr(X = a)$$

Manchmal schreibt man auch μ_X statt $\mathbb{E}[X]$.

Ersetzt man in der Definition die Wahrscheinlichkeit durch relative Häufigkeiten, so erhält man die bekannte Formel

$$\text{Erwartungswert} = \frac{\text{Summe der Werte}}{\text{Anzahl der Werte}} \; :$$

Erwartungswert

Definition (Erwartungswert)

Sei X eine Zufallsvariable mit endlichem oder aufzählbarem Wertebereich $S = \{a_1, a_2, a_3 \dots\} \subseteq \mathbb{R}$. Dann ist der *Erwartungswert* von X definiert durch

$$\mathbb{E}[X] = \sum_{a \in \mathcal{S}} a \cdot \mathbb{P}(X = a)$$

Beispiele:

- Sei W die Augenzahl bei einem Würfelwurf. Dann gilt

$$\begin{aligned}\mathbb{E}[W] &= 1 \cdot \mathbb{P}(W = 1) + 2 \cdot \mathbb{P}(W = 2) + \cdots + 6 \cdot \mathbb{P}(W = 6) \\ &= \frac{1 + \cdots + 6}{6} = \frac{21}{6} = 3.5\end{aligned}$$

Erwartungswert und Funktionen

Sei X eine Zufallsvariable mit endlichem Wertebereich $\mathcal{S} \subseteq \mathbb{R}$. Sei $f : \mathcal{S} \to \mathbb{R}$ eine Funktion. Dann ist der *Erwartungswert* von $f(X)$ gegeben durch

$$\mathbb{E}[f(X)] = \sum_{a \in \mathcal{S}} f(a) \cdot \mathbb{P}(X = a)$$

Beispiel:
Sei W die Augenzahl bei einem Würfelwurf. Dann gilt

$$\begin{aligned}\mathbb{E}[W^2] &= 1^2 \cdot \mathbb{P}(W = 1) + 2^2 \cdot \mathbb{P}(W = 2) + \cdots + 6^2 \cdot \Pr(W = 6) \\ &= \frac{1^2 + \cdots + 6^2}{6} = \frac{91}{6} = 15.1\overline{6}\end{aligned}$$

Rechnen mit Erwartungswerten

Satz[Linearität der Erwartung] Sind X und Y Zufallsvariablen mit Werten in $\mathbb{R}$ und ist $a \in \mathbb{R}$, so gilt:

- $\mathbb{E}\left[a \cdot X\right] = a \cdot \mathbb{E}[X]$
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

Satz[Nur für Unabhängige!] Sind X und Y **stochastisch unabhängige** Zufallsvariablen mit Werten in $\mathbb{R}$, so gilt

- $\mathbb{E}\left[X \cdot Y\right] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$.

Varianz, Kovarianz und Korrelation

Definition (Varianz, Kovarianz und Korrelation)

Die *Varianz* einer $\mathbb{R}$-wertigen Zufallsgröße X ist

$$\mathbb{V}[X] = \sigma_X^2 = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big)^2\Big].$$

$\sigma_X = \sqrt{\mathbb{V}[X]}$ ist die *Standardabweichung*.

Ist Y eine weitere reellwertige Zufallsvariable, so ist

$$\mathrm{Cov}[X, Y] = \mathbb{E}\Big[\big(X - \mathbb{E}[X]\big) \cdot \big(Y - \mathbb{E}[Y]\big)\Big]$$

die *Kovarianz* von X und Y.

Die *Korrelation* von X und Y ist

$$\mathrm{Cor}[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sigma_X \cdot \sigma_Y}.$$

Die Varianz

$$\mathbb{V}[X] = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]$$

ist die mittlere quadrierte Abweichung vom Mittelwert.

Die Korrelation

$$\mathrm{Cor}[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sigma_X \cdot \sigma_Y}$$

liegt immer im Intervall $[-1, 1]$. Die Variablen X und Y sind

- positiv korreliert, wenn X und Y tendenziell entweder beide überdurchschnittlich große Werte oder beide unterdurchschnittlich große Werte annehmen.
- negativ korreliert, wenn X und Y tendenziell auf verschiedenen Seiten ihrer Erwartungswerte liegen.

Sind X und Y unabhängig, so sind sie auch unkorreliert, d.h. $\mathrm{Cor}[X, Y] = 0$.

Daten und Schätzer

Fasst man Daten $x_1, ..., x_n$ als Stichprobe aus einer Population auf, oder als Realisierungen von Zufallsvariablen, so kann man die Parameter der Population bzw. der Verteilung schätzen.

Dabei ist das Stichprobenmittel $\bar{\mu}$ ein Schätzer für den Erwartungswert, und die (korrigierte) Stichprobenvarianz ein Schätzer für die Varianz.

Korrelationen zwischen Daten

Seien $x = (x_1, ..., x_n)$ und $y = (y_1, ..., y_n)$ gegeben. Die empirische Kovarianz ist definiert als

$$c_{xy} := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{\mu}_x)(y_i - \bar{\mu}_y),$$

wobei $\bar{\mu}_x$ das empirische Mittel der x_i und $\bar{\mu}_y$ das empirische Mittel der y_i ist.

Der empirische Korrelationskoeffizient ist definiert als

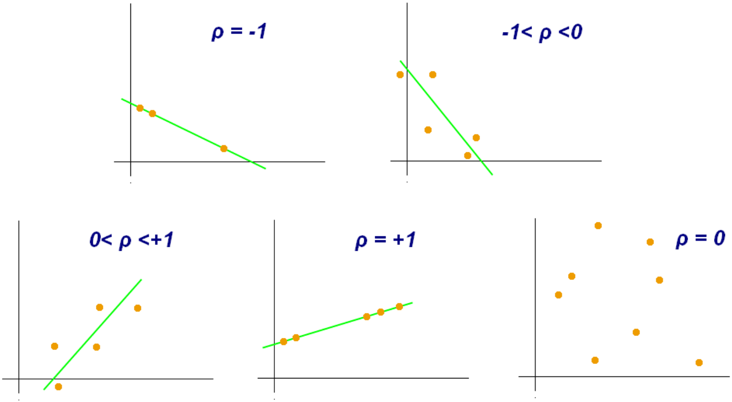
$$r_{xy} = \frac{c_{xy}}{\bar{s}_x \bar{s}_y},$$

wobei $\bar{s}_x = \sqrt{\bar{s}_x^2}$ die empirische (korrigierte) Standardabweichung von x ist (und analog $\bar{s}_y$ für y).

Linearer Zusammenhang ($\rho = r$)

Korrelation kann positiv oder negativ sein, $-1 \leq r \leq 1$

(Quelle: Wikipedia)



Stärke der Korrelation

Faustregel für die Stärke der Korrelation:

- $r > 0.75$ starke positive Korrelation
- $0.5 < r < 0.75$ mittlere positive Korrelation
- $0.25 < r < 0.5$ schwache positive Korrelation
- $r < 0.75$ starke negative Korrelation, etc.

Excel: =KORREL(Bereich1;Bereich2) gibt den Korrelationskoeffizienten für die Daten in Bereich 1 (x) und Bereich 2 (y) aus.

Lineare Regression

Gegeben: $x = (x_1, ..., x_n), y = (y_1, ..., y_n)$

Gesucht: Die Zahlen $a, b \in \mathbb{R}$, so dass die Gerade mit der Gleichung

$$y = ax + b$$

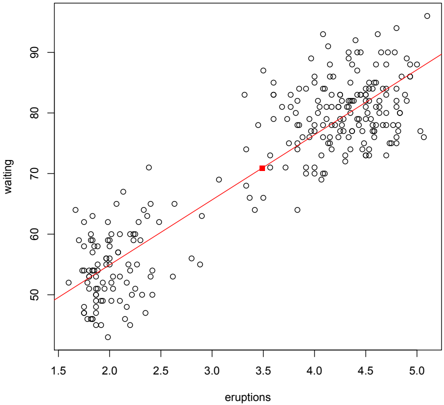
zwei Bedingungen erfüllt:

- $(\bar{\mu}_x, \bar{\mu}_y)$ liegt auf der Geraden, also $\bar{\mu}_y = a \cdot \bar{\mu}_x + b$,
- Die Abstände der Punkte von der Geraden sind minimal.

Die beiden Bedingungen legen a und b eindeutig fest, und zwar als

$$a = \frac{c_{xy}}{\bar{s}_x^2}, \quad b = \bar{\mu}_y - a \cdot \bar{\mu}_x.$$

Lineare Regression



In EXCEL: =RGP(Y-Werte;X-Werte:WAHR) gibt die Regressionsgerade (Werte für *a* und *b*) aus.