Mathematische Grundlagen der Computerlinguistik

Grundbegriffe der Wahrscheinlichkeitstheorie - Charakterisierung von Zufallsvariablen und Schätztheorie

Michael Staniek

ICL, Universität Heidelberg, SoSe 2019

12.06.2019

Commitment-Frist

- 08.07.2018

Inhalte der heutigen Vorlesung

- Schätztheorie und der Likelihood-Begriff
- Maximum-Likelihood-Schätzer
- Monty-Hall-Problem
- Bedingte Wahrscheinlichkeit
- Anwendung wahrscheinlichkeitstheoretischer Konzepte auf computerlinguistische Fragen
- Der zentrale Grenzwertsatz
- Signifikanztests - ein Beispiel

Schätztheorie

Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum. Das Tripel

Schätztheorie

Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum. Das Tripel

$$(\mathcal{X}, \mathcal{G}, (P_\lambda)\lambda \in \Lambda)$$

Schätztheorie

Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum. Das Tripel

$$(\mathcal{X}, \mathcal{G}, (P_\lambda)\lambda \in \Lambda)$$

mit der Familie von Wahrscheinlichkeitsverteilungen $(P_\lambda)\lambda \in \Lambda$ ist ein statistisches Modell.

Schätztheorie

Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum. Das Tripel

$$(\mathcal{X}, \mathcal{G}, (P_\lambda)\lambda \in \Lambda)$$

mit der Familie von Wahrscheinlichkeitsverteilungen $(P_\lambda)\lambda \in \Lambda$ ist ein statistisches Modell.

- Die Schätztheorie beschäftigt sich mit der Frage, wie dasjenige $(P_\lambda)\lambda \in \Lambda$ gefunden werden kann, das P "am besten modelliert"

Schätztheorie

Sei $(\Omega, \mathcal{F}, P)$ ein Wahrscheinlichkeitsraum. Das Tripel

$$(\mathcal{X}, \mathcal{G}, (P_\lambda)\lambda \in \Lambda)$$

mit der Familie von Wahrscheinlichkeitsverteilungen $(P_\lambda)\lambda \in \Lambda$ ist ein statistisches Modell.

- Die Schätztheorie beschäftigt sich mit der Frage, wie dasjenige $(P_\lambda)\lambda \in \Lambda$ gefunden werden kann, das P "am besten modelliert"
- Im Kontext der Computerlinguistik heißt dies oft, dasjenige $(P_\lambda)\lambda \in \Lambda$ zu finden, das für ungesehene Stichproben von Zufallsvariablen auf $(\Omega, \mathcal{F}, P)$ die besten Vorhersagen hinsichtlich deren Wahrscheinlichkeiten trifft

Der Likelihood-Begriff

- Wahrscheinlichkeitsmaß auf einem Wahrscheinlichkeitsraum weist bestimmten Beobachtungen eine Wahrscheinlichkeit zu

Der Likelihood-Begriff

- Wahrscheinlichkeitsmaß auf einem Wahrscheinlichkeitsraum weist bestimmten Beobachtungen eine Wahrscheinlichkeit zu
- Wahrscheinlichkeit (Engl. “probability”) weist Beobachtungen *unter Annahme bestimmter Modellparameter* eine Wahrscheinlichkeit zu

Der Likelihood-Begriff

- Wahrscheinlichkeitsmaß auf einem Wahrscheinlichkeitsraum weist bestimmten Beobachtungen eine Wahrscheinlichkeit zu
- Wahrscheinlichkeit (Engl. “probability”) weist Beobachtungen *unter Annahme bestimmter Modellparameter* eine Wahrscheinlichkeit zu
- Im Gegensatz dazu bezeichnet Likelihood die *Wahrscheinlichkeit von Modellparametern unter der Annahme bestimmter Beobachtungen*

Der Likelihood-Begriff

- Wahrscheinlichkeitsmaß auf einem Wahrscheinlichkeitsraum weist bestimmten Beobachtungen eine Wahrscheinlichkeit zu
- Wahrscheinlichkeit (Engl. "probability") weist Beobachtungen *unter Annahme bestimmter Modellparameter* eine Wahrscheinlichkeit zu
- Im Gegensatz dazu bezeichnet Likelihood die *Wahrscheinlichkeit von Modellparametern unter der Annahme bestimmter Beobachtungen*
- Für das statistische Modell $(\mathcal{X}, \mathcal{G}, (P_\lambda)\lambda \in \Lambda)$ wird

Der Likelihood-Begriff

- Wahrscheinlichkeitsmaß auf einem Wahrscheinlichkeitsraum weist bestimmten Beobachtungen eine Wahrscheinlichkeit zu
- Wahrscheinlichkeit (Engl. "probability") weist Beobachtungen *unter Annahme bestimmter Modellparameter* eine Wahrscheinlichkeit zu
- Im Gegensatz dazu bezeichnet Likelihood die *Wahrscheinlichkeit von Modellparametern unter der Annahme bestimmter Beobachtungen*
- Für das statistische Modell $(\mathcal{X}, \mathcal{G}, (P_\lambda)\lambda \in \Lambda)$ wird

$$\ell_x(\lambda) : \Lambda \to [0,1], x \in \mathcal{X}$$

$$\lambda \longmapsto P_\lambda(\{x\}), x \in \mathcal{X}$$

Der Likelihood-Begriff

- Wahrscheinlichkeitsmaß auf einem Wahrscheinlichkeitsraum weist bestimmten Beobachtungen eine Wahrscheinlichkeit zu
- Wahrscheinlichkeit (Engl. "probability") weist Beobachtungen *unter Annahme bestimmter Modellparameter* eine Wahrscheinlichkeit zu
- Im Gegensatz dazu bezeichnet Likelihood die *Wahrscheinlichkeit von Modellparametern unter der Annahme bestimmter Beobachtungen*
- Für das statistische Modell $(\mathcal{X}, \mathcal{G}, (P_\lambda)\lambda \in \Lambda)$ wird

$$\ell_x(\lambda) : \Lambda \to [0,1], x \in \mathcal{X}$$

$$\lambda \longmapsto P_\lambda(\{x\}), x \in \mathcal{X}$$

die Likelihood-Funktion zur Beobachtung x genannt.

Maximum-Likelihood-Schätzer

Für das statistische Modell $(\mathcal{X}, \mathcal{G}, (P_\lambda)\lambda \in \Lambda)$ wird ein konkretes $\hat{\lambda}$ als Schätzer bezeichnet. Wenn $\hat{\lambda}$ dasjenige Element in Λ ist, was die Wahrscheinlichkeit der Beobachtungsdaten maximiert, also wenn gilt:

Maximum-Likelihood-Schätzer

Für das statistische Modell $(\mathcal{X}, \mathcal{G}, (P_\lambda)\lambda \in \Lambda)$ wird ein konkretes $\hat{\lambda}$ als Schätzer bezeichnet. Wenn $\hat{\lambda}$ dasjenige Element in Λ ist, was die Wahrscheinlichkeit der Beobachtungsdaten maximiert, also wenn gilt:

$$P_{\hat{\lambda}}(\{x\}) = \underset{\lambda \in \Lambda}{\operatorname{argmax}} P_\lambda(\{x\})$$

Maximum-Likelihood-Schätzer

Für das statistische Modell $(\mathcal{X}, \mathcal{G}, (P_\lambda)\lambda \in \Lambda)$ wird ein konkretes $\hat{\lambda}$ als Schätzer bezeichnet. Wenn $\hat{\lambda}$ dasjenige Element in Λ ist, was die Wahrscheinlichkeit der Beobachtungsdaten maximiert, also wenn gilt:

$$P_{\hat{\lambda}}(\{x\}) = \underset{\lambda \in \Lambda}{\operatorname{argmax}} P_\lambda(\{x\})$$

ist $\hat{\lambda}$ ein Maximum-Likelihood-Schätzer zu den Beobachtungen x.

Maximum-Likelihood-Schätzer

Für das statistische Modell $(\mathcal{X}, \mathcal{G}, (P_\lambda)\lambda \in \Lambda)$ wird ein konkretes $\hat{\lambda}$ als Schätzer bezeichnet. Wenn $\hat{\lambda}$ dasjenige Element in Λ ist, was die Wahrscheinlichkeit der Beobachtungsdaten maximiert, also wenn gilt:

$$P_{\hat{\lambda}}(\{x\}) = \operatorname*{argmax}_{\lambda \in \Lambda} P_\lambda(\{x\})$$

ist $\hat{\lambda}$ ein Maximum-Likelihood-Schätzer zu den Beobachtungen x.

- Maximum-Likelihood-Schätzer sind oft "vernünftige" Modellparametrisierungen

Maximum-Likelihood-Schätzer

Für das statistische Modell $(\mathcal{X}, \mathcal{G}, (P_\lambda)\lambda \in \Lambda)$ wird ein konkretes $\hat{\lambda}$ als Schätzer bezeichnet. Wenn $\hat{\lambda}$ dasjenige Element in Λ ist, was die Wahrscheinlichkeit der Beobachtungsdaten maximiert, also wenn gilt:

$$P_{\hat{\lambda}}(\{x\}) = \underset{\lambda \in \Lambda}{\operatorname{argmax}} P_\lambda(\{x\})$$

ist $\hat{\lambda}$ ein Maximum-Likelihood-Schätzer zu den Beobachtungen x.

- Maximum-Likelihood-Schätzer sind oft "vernünftige" Modellparametrisierungen
- Allerdings haben sie den Nachteil, "aus Zufall" ungesehenen Daten tendenziell zu wenig Wahrscheinlichkeit zuzuweisen

Maximum-Likelihood-Schätzer

Für das statistische Modell $(\mathcal{X}, \mathcal{G}, (P_\lambda)\lambda \in \Lambda)$ wird ein konkretes $\hat{\lambda}$ als Schätzer bezeichnet. Wenn $\hat{\lambda}$ dasjenige Element in Λ ist, was die Wahrscheinlichkeit der Beobachtungsdaten maximiert, also wenn gilt:

$$P_{\hat{\lambda}}(\{x\}) = \operatorname*{argmax}_{\lambda \in \Lambda} P_\lambda(\{x\})$$

ist $\hat{\lambda}$ ein Maximum-Likelihood-Schätzer zu den Beobachtungen x.

- Maximum-Likelihood-Schätzer sind oft "vernünftige" Modellparametrisierungen
- Allerdings haben sie den Nachteil, "aus Zufall" ungesehenen Daten tendenziell zu wenig Wahrscheinlichkeit zuzuweisen
- Beispiel: Taxiproblem (Schätzung einer Laplace-Verteilung auf einer endlichen Menge unbekannter Kardinalität)

Maximum-Likelihood-Schätzer

Für das statistische Modell $(\mathcal{X}, \mathcal{G}, (P_\lambda)\lambda \in \Lambda)$ wird ein konkretes $\hat{\lambda}$ als Schätzer bezeichnet. Wenn $\hat{\lambda}$ dasjenige Element in Λ ist, was die Wahrscheinlichkeit der Beobachtungsdaten maximiert, also wenn gilt:

$$P_{\hat{\lambda}}(\{x\}) = \operatorname*{argmax}_{\lambda \in \Lambda} P_\lambda(\{x\})$$

ist $\hat{\lambda}$ ein Maximum-Likelihood-Schätzer zu den Beobachtungen x.

- Maximum-Likelihood-Schätzer sind oft "vernünftige" Modellparametrisierungen
- Allerdings haben sie den Nachteil, "aus Zufall" ungesehenen Daten tendenziell zu wenig Wahrscheinlichkeit zuzuweisen
- Beispiel: Taxiproblem (Schätzung einer Laplace-Verteilung auf einer endlichen Menge unbekannter Kardinalität)
- Es gibt viele weitere gute und vernünftige Schätzverfahren neben der Maximum-Likelihood-Schätzung

Monty Hall Problem

- Drei Türen, zwei mit Ziegen, eine mit Sportwagen

Monty Hall Problem

- Drei Türen, zwei mit Ziegen, eine mit Sportwagen
- Gast wählt eine Türe aus

Monty Hall Problem

- Drei Türen, zwei mit Ziegen, eine mit Sportwagen
- Gast wählt eine Türe aus
- Moderator öffnet eine der anderen Türen, hinter der immer eine Ziege steht

Monty Hall Problem

- Drei Türen, zwei mit Ziegen, eine mit Sportwagen
- Gast wählt eine Türe aus
- Moderator öffnet eine der anderen Türen, hinter der immer eine Ziege steht
- Moderator lässt Gast entscheiden, ob er bei seiner Wahl bleiben, oder die Türe wechseln möchte

Das Monty-Hall-Problem

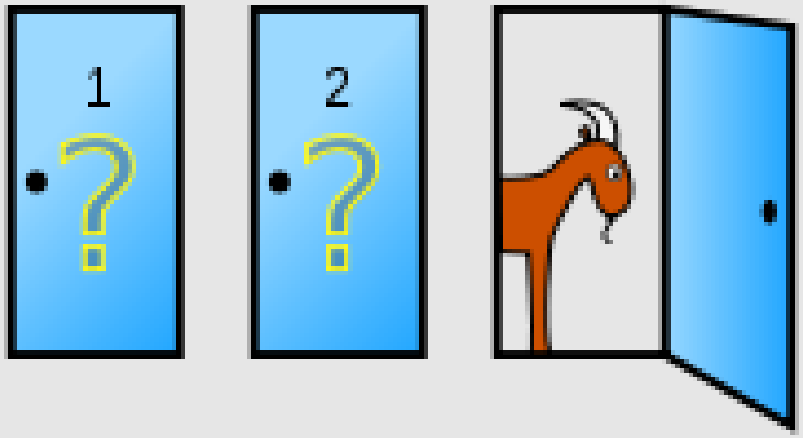

Abbildung: Das Monty-Hall-Problem. Quelle: Wikimedia Commons:
https://commons.wikimedia.org/wiki/File:Monty_open_door.svg

Das Monty-Hall-Problem

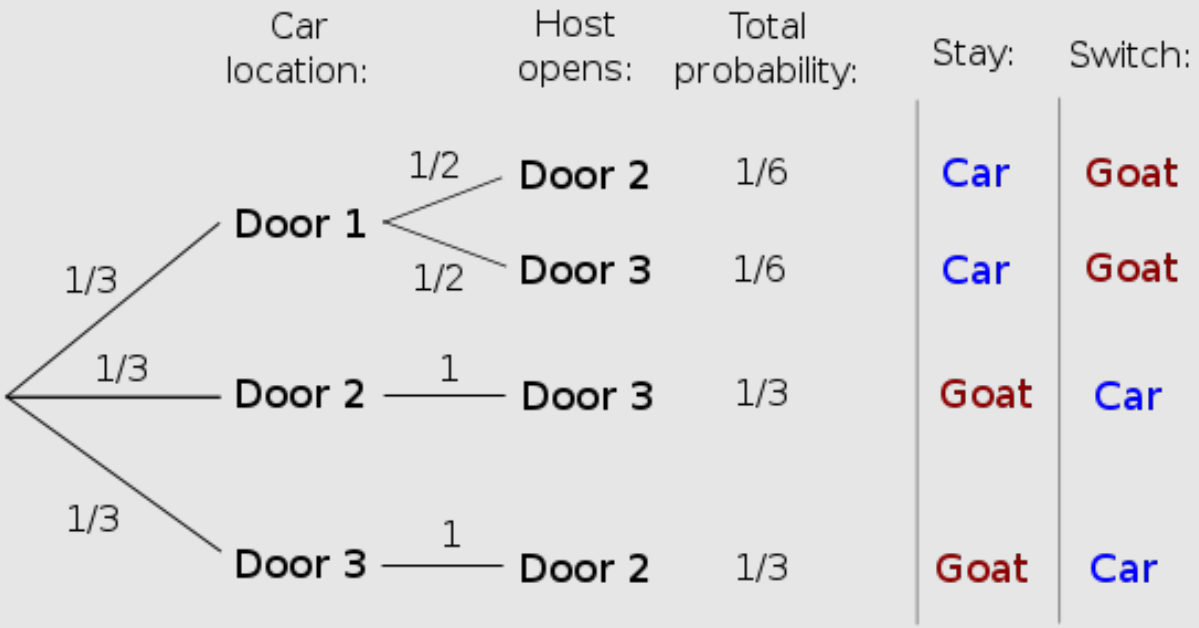

Abbildung: Das Monty-Hall-Problem. Quelle: Wikimedia Commons:
`https://commons.wikimedia.org/wiki/File:Monty_tree_door1.svg`

Bedingte Wahrscheinlichkeit

Sei $(\Omega, \mathcal{F})$ ein messbarer Raum und A, B Ereignismengen $\in \mathcal{F}$. Die bedingte Wahrscheinlichkeit $P(B|A)$ ist die Wahrscheinlichkeit, dass B eintritt, unter der Annahme, dass A eintritt und ist gegeben durch:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A, B)}{P(A)}$$

Das Monty-Hall-Problem: Analyse mit bedingter Wahrscheinlichkeit

- Sportwagen hinter Türe k: S_k
- Moderator wählt Türe k: M_k
- Gast wählt (zu Beginn) Türe k: G_k

Somit gilt:

Das Monty-Hall-Problem: Analyse mit bedingter Wahrscheinlichkeit

- Sportwagen hinter Türe k: S_k
- Moderator wählt Türe k: M_k
- Gast wählt (zu Beginn) Türe k: G_k

Somit gilt:

- $P(M_3|S_1, G_1) = \frac{1}{2}$

Das Monty-Hall-Problem: Analyse mit bedingter Wahrscheinlichkeit

- Sportwagen hinter Türe k: S_k
- Moderator wählt Türe k: M_k
- Gast wählt (zu Beginn) Türe k: G_k

Somit gilt:

- $P(M_3|S_1, G_1) = \frac{1}{2}$
- $P(M_3|S_2, G_1) = 1$

Das Monty-Hall-Problem: Analyse mit bedingter Wahrscheinlichkeit

- Sportwagen hinter Türe k: S_k
- Moderator wählt Türe k: M_k
- Gast wählt (zu Beginn) Türe k: G_k

Somit gilt:

- $P(M_3|S_1, G_1) = \frac{1}{2}$
- $P(M_3|S_2, G_1) = 1$
- $P(M_3|S_3, G_1) = 0$

Das Monty-Hall-Problem: Analyse mit bedingter Wahrscheinlichkeit

Somit gilt auch:

$$P(S_2|M_3, G_1) = \frac{P(M_3|S_2, G_1)P(S_2 \cap G_1)}{P(M_3 \cap G_1)}$$

Das Monty-Hall-Problem: Analyse mit bedingter Wahrscheinlichkeit

Somit gilt auch:

$$P(S_2|M_3, G_1) = \frac{P(M_3|S_2, G_1)P(S_2 \cap G_1)}{P(M_3 \cap G_1)}$$

$$= \frac{P(M_3|S_2, G_1)P(S_2 \cap G_1)}{P(M_3|S_1, G_1)P(S_1 \cap G_1) + P(M_3|S_2, G_1)P(S_2 \cap G_1) + P(M_3|S_3, G_1)P(S_3 \cap G_1)}$$

Das Monty-Hall-Problem: Analyse mit bedingter Wahrscheinlichkeit

Somit gilt auch:

$$P(S_2|M_3, G_1) = \frac{P(M_3|S_2, G_1)P(S_2 \cap G_1)}{P(M_3 \cap G_1)}$$

$$= \frac{P(M_3|S_2, G_1)P(S_2 \cap G_1)}{P(M_3|S_1, G_1)P(S_1 \cap G_1) + P(M_3|S_2, G_1)P(S_2 \cap G_1) + P(M_3|S_3, G_1)P(S_3 \cap G_1)}$$

$P(S_k \cap G_j) = P(S_k)P(G_j) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$ daher gilt:

Das Monty-Hall-Problem: Analyse mit bedingter Wahrscheinlichkeit

Somit gilt auch:

$$P(S_2|M_3, G_1) = \frac{P(M_3|S_2, G_1)P(S_2 \cap G_1)}{P(M_3 \cap G_1)}$$

$$= \frac{P(M_3|S_2, G_1)P(S_2 \cap G_1)}{P(M_3|S_1, G_1)P(S_1 \cap G_1) + P(M_3|S_2, G_1)P(S_2 \cap G_1) + P(M_3|S_3, G_1)P(S_3 \cap G_1)}$$

$P(S_k \cap G_j) = P(S_k)P(G_j) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$ daher gilt:

$$P(S_2|M_3, G_1) = \frac{P(M_3|S_2, G_1)}{P(M_3|S_1, G_1) + P(M_3|S_2, G_1) + P(M_3|S_3, G_1)}$$

Das Monty-Hall-Problem: Analyse mit bedingter Wahrscheinlichkeit

Somit gilt auch:

$$P(S_2|M_3, G_1) = \frac{P(M_3|S_2, G_1)P(S_2 \cap G_1)}{P(M_3 \cap G_1)}$$

$$= \frac{P(M_3|S_2, G_1)P(S_2 \cap G_1)}{P(M_3|S_1, G_1)P(S_1 \cap G_1) + P(M_3|S_2, G_1)P(S_2 \cap G_1) + P(M_3|S_3, G_1)P(S_3 \cap G_1)}$$

$P(S_k \cap G_j) = P(S_k)P(G_j) = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9}$ daher gilt:

$$P(S_2|M_3, G_1) = \frac{P(M_3|S_2, G_1)}{P(M_3|S_1, G_1) + P(M_3|S_2, G_1) + P(M_3|S_3, G_1)}$$

$$= \frac{1}{\frac{1}{2} + 1 + 0} = \frac{2}{3}$$

Wahrscheinlichkeitstheoretische Erweiterungen der Computerlinguistik

Wahrscheinlichkeitstheoretische Erweiterungen der Computerlinguistik

- Sprachmodelle

Wahrscheinlichkeitstheoretische Erweiterungen der Computerlinguistik

- Sprachmodelle
- Probabilistische Grammatiken

Wahrscheinlichkeitstheoretische Erweiterungen der Computerlinguistik

- Sprachmodelle
- Probabilistische Grammatiken
- Maschinelles Lernen

Wahrscheinlichkeitstheoretische Erweiterungen der Computerlinguistik

- Sprachmodelle
- Probabilistische Grammatiken
- Maschinelles Lernen
- und viele weitere

N-Gramm Modelle zur Sprachmodellierung

Grundidee 1 (Markovkette): Wahrscheinlichkeit einer Sequenz lässt sich in ein Produkt bedingter Wahrscheinlichkeiten zerlegen

$$p(w_0, ..., w_n) = p(w_0)p(w_1|w_0)p(w_2|w_0, w_1)...p(w_n|w_0, ..., w_{n-1})$$

N-Gramm Modelle zur Sprachmodellierung

Grundidee 1 (Markovkette): Wahrscheinlichkeit einer Sequenz lässt sich in ein Produkt bedingter Wahrscheinlichkeiten zerlegen

$$p(w_0, ..., w_n) = p(w_0)p(w_1|w_0)p(w_2|w_0, w_1)...p(w_n|w_0, ..., w_{n-1})$$

Grundidee 2: Bedingte Wahrscheinlichkeit einer Wortsequenz hängt im Wesentlichen von begrenzter Anzahl von m Wörtern ab:

$$p(w_n|w_0, ..., w_{n-1}) \approx p(w_n|w_{n-m-1}, ...w_{n-1})$$

N-Gramm Modelle zur Sprachmodellierung

Beispiel: $m = 3$ (Trigramm-Modell):

$$p(w_n|w_0, ..., w_{n-1}) = p(w_n|w_{n-2}, w_{n-1})$$

N-Gramm Modelle zur Sprachmodellierung

Beispiel: $m = 3$ (Trigramm-Modell):

$$p(w_n|w_0, ..., w_{n-1}) = p(w_n|w_{n-2}, w_{n-1})$$

- Annahmen:

N-Gramm Modelle zur Sprachmodellierung

Beispiel: $m = 3$ (Trigramm-Modell):

$$p(w_n|w_0, ..., w_{n-1}) = p(w_n|w_{n-2}, w_{n-1})$$

- Annahmen:
  - Markov-Annahme: Die Gesamtwahrscheinlichkeit der Sequenz lässt sich durch den auf m Wörter begrenzten Kontext gut modellieren.

N-Gramm Modelle zur Sprachmodellierung

Beispiel: $m = 3$ (Trigramm-Modell):

$$p(w_n|w_0, ..., w_{n-1}) = p(w_n|w_{n-2}, w_{n-1})$$

- Annahmen:
  - Markov-Annahme: Die Gesamtwahrscheinlichkeit der Sequenz lässt sich durch den auf m Wörter begrenzten Kontext gut modellieren.
  - Dass die Markov-Annahme nicht immer zutrifft, lässt sich gut am Extrembeispiel $m = 1$ (Unigramm-Modell) verdeutlichen.

N-Gramm Modelle zur Sprachmodellierung

Maximum-Likelihood-Schätzer für Trigramm-Modell:

$$p(w_n|w_{n-2}, w_{n-1}) = \frac{\#(w_{n-2}, w_{n-1}, w_w)}{\#(w_{n-2}, w_{n-1})}$$

N-Gramm Modelle zur Sprachmodellierung

Maximum-Likelihood-Schätzer für Trigramm-Modell:

$$p(w_n|w_{n-2}, w_{n-1}) = \frac{\#(w_{n-2}, w_{n-1}, w_w)}{\#(w_{n-2}, w_{n-1})}$$

- Warum ist dies ein Maximum-Likelihood-Schätzer?

N-Gramm Modelle zur Sprachmodellierung

Maximum-Likelihood-Schätzer für Trigramm-Modell:

$$p(w_n|w_{n-2}, w_{n-1}) = \frac{\#(w_{n-2}, w_{n-1}, w_w)}{\#(w_{n-2}, w_{n-1})}$$

- Warum ist dies ein Maximum-Likelihood-Schätzer?
  - Es gibt keine Möglichkeit, die Wahrscheinlichkeitsmasse auf alle Trigramme zu verteilen und dem Gesamtkorpus eine höhere Wahrscheinlichkeit zuzuweisen

N-Gramm Modelle zur Sprachmodellierung

Maximum-Likelihood-Schätzer für Trigramm-Modell:

$$p(w_n|w_{n-2}, w_{n-1}) = \frac{\#(w_{n-2}, w_{n-1}, w_w)}{\#(w_{n-2}, w_{n-1})}$$

- Warum ist dies ein Maximum-Likelihood-Schätzer?
  - Es gibt keine Möglichkeit, die Wahrscheinlichkeitsmasse auf alle Trigramme zu verteilen und dem Gesamtkorpus eine höhere Wahrscheinlichkeit zuzuweisen
- Ist ein Maximum-Likelihood-Schätzer hier angebracht?

N-Gramm Modelle zur Sprachmodellierung

Maximum-Likelihood-Schätzer für Trigramm-Modell:

$$p(w_n|w_{n-2}, w_{n-1}) = \frac{\#(w_{n-2}, w_{n-1}, w_w)}{\#(w_{n-2}, w_{n-1})}$$

- Warum ist dies ein Maximum-Likelihood-Schätzer?
  - Es gibt keine Möglichkeit, die Wahrscheinlichkeitsmasse auf alle Trigramme zu verteilen und dem Gesamtkorpus eine höhere Wahrscheinlichkeit zuzuweisen
- Ist ein Maximum-Likelihood-Schätzer hier angebracht?
  - Nein! Ungesehenen Trigrammen wird die Wahrscheinlichkeit 0 zugewiesen

N-Gramm Modelle zur Sprachmodellierung

Maximum-Likelihood-Schätzer für Trigramm-Modell:

$$p(w_n|w_{n-2}, w_{n-1}) = \frac{\#(w_{n-2}, w_{n-1}, w_w)}{\#(w_{n-2}, w_{n-1})}$$

- Warum ist dies ein Maximum-Likelihood-Schätzer?
  - Es gibt keine Möglichkeit, die Wahrscheinlichkeitsmasse auf alle Trigramme zu verteilen und dem Gesamtkorpus eine höhere Wahrscheinlichkeit zuzuweisen
- Ist ein Maximum-Likelihood-Schätzer hier angebracht?
  - Nein! Ungesehenen Trigrammen wird die Wahrscheinlichkeit 0 zugewiesen
  - Alternativen:

N-Gramm Modelle zur Sprachmodellierung

Maximum-Likelihood-Schätzer für Trigramm-Modell:

$$p(w_n|w_{n-2}, w_{n-1}) = \frac{\#(w_{n-2}, w_{n-1}, w_w)}{\#(w_{n-2}, w_{n-1})}$$

- Warum ist dies ein Maximum-Likelihood-Schätzer?
  - Es gibt keine Möglichkeit, die Wahrscheinlichkeitsmasse auf alle Trigramme zu verteilen und dem Gesamtkorpus eine höhere Wahrscheinlichkeit zuzuweisen
- Ist ein Maximum-Likelihood-Schätzer hier angebracht?
  - Nein! Ungesehenen Trigrammen wird die Wahrscheinlichkeit 0 zugewiesen
  - Alternativen: $n + 1$-Smoothing, $n + \alpha$-Smoothing, Kneser-Ney-Smoothing, etc.

Recap: Kontextfreie Grammatik

Kontextfreie Grammatik

Eine kontextfreie Grammatik G ist ein 4-Tupel (N, Σ, R, S) wobei

- Σ eine endliche Menge von Terminalsymbolen, das Alphabet von G ist

Recap: Kontextfreie Grammatik

Kontextfreie Grammatik

Eine kontextfreie Grammatik G ist ein 4-Tupel (N, Σ, R, S) wobei

- Σ eine endliche Menge von Terminalsymbolen, das Alphabet von G ist
- N eine endliche Menge von Nichtterminalsymbolen ist

Recap: Kontextfreie Grammatik

Kontextfreie Grammatik

Eine kontextfreie Grammatik G ist ein 4-Tupel (N, Σ, R, S) wobei

- Σ eine endliche Menge von Terminalsymbolen, das Alphabet von G ist
- N eine endliche Menge von Nichtterminalsymbolen ist
- $S \in N$ das Startsymbol von G ist

Recap: Kontextfreie Grammatik

Kontextfreie Grammatik

Eine kontextfreie Grammatik G ist ein 4-Tupel (N, Σ, R, S) wobei

- Σ eine endliche Menge von Terminalsymbolen, das Alphabet von G ist
- N eine endliche Menge von Nichtterminalsymbolen ist
- $S \in N$ das Startsymbol von G ist
- R eine endliche Menge von Produktionsregeln ist:
  - $R \subseteq N \times (\Sigma \cup N)^*$
- A^* bezeichnet die Kleenesche Hülle einer Sprache A:

$$A^* := \cup_{i \in N_0} A^i, A^0 := \epsilon, A^{n+1} := A^n \circ A$$

Recap: Kontextfreie Grammatik

Kontextfreie Grammatik

Eine kontextfreie Grammatik G ist ein 4-Tupel (N, Σ, R, S) wobei

- Σ eine endliche Menge von Terminalsymbolen, das Alphabet von G ist
- N eine endliche Menge von Nichtterminalsymbolen ist
- $S \in N$ das Startsymbol von G ist
- R eine endliche Menge von Produktionsregeln ist:
  - $R \subseteq N \times (\Sigma \cup N)^*$
- A^* bezeichnet die Kleenesche Hülle einer Sprache A:

$$A^* := \cup_{i \in N_0} A^i, A^0 := \epsilon, A^{n+1} := A^n \circ A$$

- Alle Zeichenketten, die durch rekursive Anwendungen von Regeln in R auf S erzeugt werden können, sind Teil der von G definierten kontextfreien Sprache

PCFG: Probabilistische Kontextfreie Grammatik

Probabilistische Kontextfreie Grammatik

Eine kontextfreie Grammatik G ist ein 5-Tupel (N, Σ, R, S, P) wobei

- Σ eine endliche Menge von Terminalsymbolen, das Alphabet von G ist

PCFG: Probabilistische Kontextfreie Grammatik

Probabilistische Kontextfreie Grammatik

Eine kontextfreie Grammatik G ist ein 5-Tupel (N, Σ, R, S, P) wobei

- Σ eine endliche Menge von Terminalsymbolen, das Alphabet von G ist
- N eine endliche Menge von Nichtterminalsymbolen ist

PCFG: Probabilistische Kontextfreie Grammatik

Probabilistische Kontextfreie Grammatik

Eine kontextfreie Grammatik G ist ein 5-Tupel (N, Σ, R, S, P) wobei

- Σ eine endliche Menge von Terminalsymbolen, das Alphabet von G ist
- N eine endliche Menge von Nichtterminalsymbolen ist
- $S \in N$ das Startsymbol von G ist

PCFG: Probabilistische Kontextfreie Grammatik

Probabilistische Kontextfreie Grammatik

Eine kontextfreie Grammatik G ist ein 5-Tupel (N, Σ, R, S, P) wobei

- Σ eine endliche Menge von Terminalsymbolen, das Alphabet von G ist
- N eine endliche Menge von Nichtterminalsymbolen ist
- $S \in N$ das Startsymbol von G ist
- R eine endliche Menge von Produktionsregeln ist:
  - $R \subseteq N \times (\Sigma \cup N)^*$

PCFG: Probabilistische Kontextfreie Grammatik

Probabilistische Kontextfreie Grammatik

Eine kontextfreie Grammatik G ist ein 5-Tupel (N, Σ, R, S, P) wobei

- Σ eine endliche Menge von Terminalsymbolen, das Alphabet von G ist
- N eine endliche Menge von Nichtterminalsymbolen ist
- $S \in N$ das Startsymbol von G ist
- R eine endliche Menge von Produktionsregeln ist:
  - $R \subseteq N \times (\Sigma \cup N)^*$
- P ein bedingtes Wahrscheinlichkeitsmaß auf den Produktionsregeln der folgenden Form ist:

$$p(\alpha \to \beta|\alpha), \alpha \in N, \beta \in (N \cup \Sigma)^*$$

PCFG: Probabilistische Kontextfreie Grammatik

- Wie könnte ein Maximum-Likelihood-Schätzer für PCFGs aussehen?

PCFG: Probabilistische Kontextfreie Grammatik

- Wie könnte ein Maximum-Likelihood-Schätzer für PCFGs aussehen?

Maximum-Likelihood-Schätzer für PCFGs

$$p(\alpha \to \beta|\alpha) = \frac{\#\alpha \to \beta}{\#\alpha}$$

PCFG: Probabilistische Kontextfreie Grammatik

- Wie könnte ein Maximum-Likelihood-Schätzer für PCFGs aussehen?

Maximum-Likelihood-Schätzer für PCFGs

$$p(\alpha \rightarrow \beta|\alpha) = \frac{\#\alpha \rightarrow \beta}{\#\alpha}$$

- Ist ein Maximum-Likelihood-Schätzer hier angebracht? Welche Probleme könnten sich ergeben?

PCFG: Probabilistische Kontextfreie Grammatik

- Wie könnte ein Maximum-Likelihood-Schätzer für PCFGs aussehen?

Maximum-Likelihood-Schätzer für PCFGs

$$p(\alpha \to \beta|\alpha) = \frac{\#\alpha \to \beta}{\#\alpha}$$

- Ist ein Maximum-Likelihood-Schätzer hier angebracht? Welche Probleme könnten sich ergeben?
- Welche Alternativen gäbe es zur Maximum-Likelihood-Schätzung?

Recap: Binomial-Verteilung

- Abkürzung: $B(N, p)$
- Parameter:
  - $N \in \mathbb{N}$
  - $p \in [0, 1]$

$$\Omega = \{0, 1, ...N\}$$

$$p_k = \binom{N}{k} p^k (1-p)^{N-k}$$

- $\binom{N}{k} = \Pi_{j=1}^{k} \frac{N+1-j}{j}$: Anzahl der verschiedenen Arten, auf die man k Objekte aus einer Menge von N Objekten ohne Zurücklegen auswählen kann (anzahl der k-elementigen Teilmengen von N).

Recap: Binomial-Verteilung

- Abkürzung: $B(N, p)$
- Parameter:
  - $N \in \mathbb{N}$
  - $p \in [0, 1]$

$$\Omega = \{0, 1, ...N\}$$

$$p_k = \binom{N}{k} p^k (1-p)^{N-k}$$

- $\binom{N}{k} = \Pi_{j=1}^{k} \frac{N+1-j}{j}$: Anzahl der verschiedenen Arten, auf die man k Objekte aus einer Menge von N Objekten ohne Zurücklegen auswählen kann (anzahl der k-elementigen Teilmengen von N).
- Anzahl der Erfolge bei der N-maligen Wiederholung eines Bernoulli-Experiments mit Erfolgswahrscheinlichkeit p

Recap: Binomial-Verteilung

Binomial-Verteilungen: Beispiele

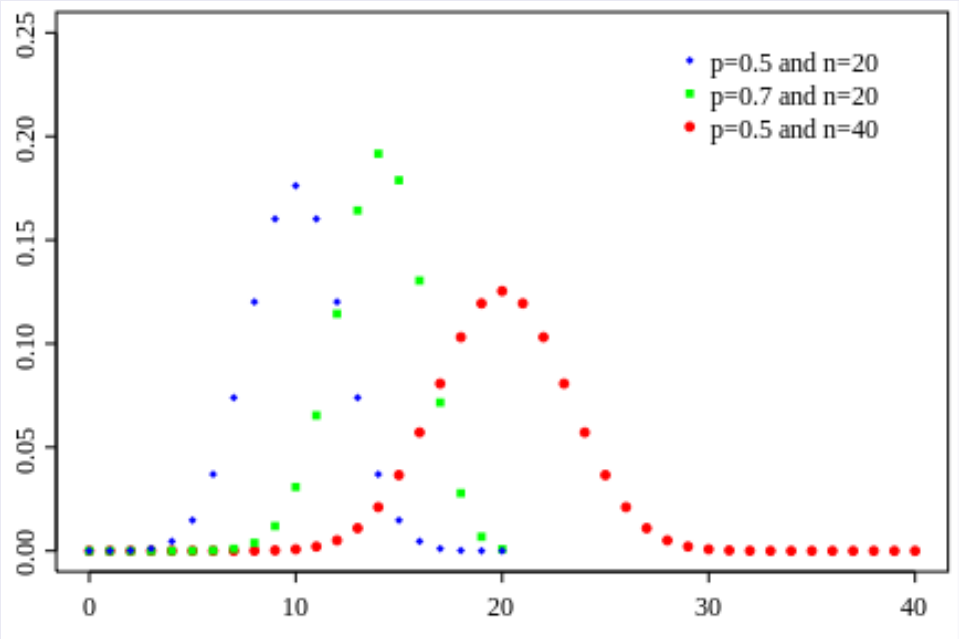

Abbildung: Binomial-Verteilungen mit verschiedenen Parameterkonfigurationen. Quelle: Wikimedia Commons: `https://commons.wikimedia.org/wiki/File:Binomial_distribution_pmf.svg`

- Welche Tendenzen ergeben sich, wenn $n \to \infty$?

Der zentrale Grenzwertsatz

- Mit zunehmendem n scheinen sich die Binomialverteilungen für unterschiedliche p immer weiter zu ähneln.

Der zentrale Grenzwertsatz

- Mit zunehmendem n scheinen sich die Binomialverteilungen für unterschiedliche p immer weiter zu ähneln.
- Zur Erinnerung: Eine binomial verteile Zufallsvaiable lässt sich als Summe von Bernoulli-verteilten Zufallsvariablen auffassen.

Der zentrale Grenzwertsatz

- Mit zunehmendem n scheinen sich die Binomialverteilungen für unterschiedliche p immer weiter zu ähneln.
- Zur Erinnerung: Eine binomial verteile Zufallsvaiable lässt sich als Summe von Bernoulli-verteilten Zufallsvariablen auffassen.
- Es lässt sich zeigen, dass diese Konvergenzeigenschaft auch allgemein für viele Summen von Zufallsvariablen gilt.

Der zentrale Grenzwertsatz

- Mit zunehmendem n scheinen sich die Binomialverteilungen für unterschiedliche p immer weiter zu ähneln.
- Zur Erinnerung: Eine binomial verteile Zufallsvaiable lässt sich als Summe von Bernoulli-verteilten Zufallsvariablen auffassen.
- Es lässt sich zeigen, dass diese Konvergenzeigenschaft auch allgemein für viele Summen von Zufallsvariablen gilt.
- Dies besagt der sogenannte *zentrale Grenzwertsatz*. Die aus der Summierung vieler solcher Zufallsvariablen resultierende Zufallsvariable hat eine sogenannte *Normalverteilung*

Normalverteilung

- Die Normalverteilung $N(\mu, \sigma^2)$ mit Parametern μ (Erwartungswert) und σ^2 (Varianz) ist eine kontinuierliche Wahrscheinlichkeitsverteilung auf $\mathbb{R}$ mit einer Dichte $f(x)$ bezüglich des Lebesque-Maßes:

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normalverteilung

- Die Normalverteilung $N(\mu, \sigma^2)$ mit Parametern μ (Erwartungswert) und σ^2 (Varianz) ist eine kontinuierliche Wahrscheinlichkeitsverteilung auf $\mathbb{R}$ mit einer Dichte $f(x)$ bezüglich des Lebesque-Maßes:

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- $N(0, 1)$ wird als *Standard-Normalverteilung* bezeichnet. Jede Normalverteilung kann durch entsprechende Skalierung zu einer Standard-Normalverteilung umgeformt werden.

Normalverteilung

Normalverteilungen: Beispiele

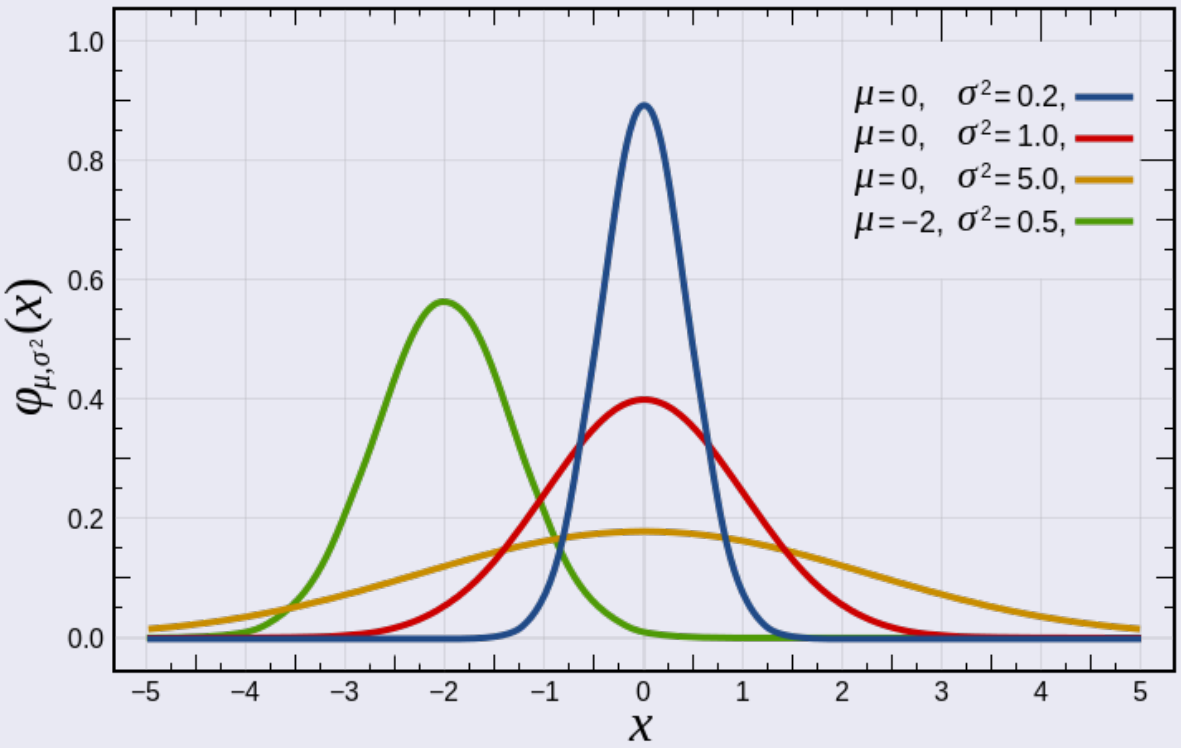

Abbildung: Normalverteilungen mit verschiedenen Parametern. Quelle: Wikimedia Commons: `https://commons.wikimedia.org/wiki/File:Normal_Distribution_PDF.svg`

Zentraler Grenzwertsatz

Zentraler Grenzwertsatz nach Lindeberg und Levy:[1]

Sei $X_1, X_2, ...$ eine Folge von unabhängigen, identisch verteilen Zufallsvariablen mit Erwartungswert $\mathbb{E}[X_i] = \mu$ und Varianz $Var(X_i) = \sigma^2 < \infty$. Dann gilt:

$$\lim_{n\to\infty} \frac{1}{\sqrt{n\sigma^2}}((\frac{1}{n}\sum_{i=1}^{n} X_i) - \mu) \stackrel{d}{=} X$$

(in Verteilung), wobei

$$P_X = N(0, \sigma^2)$$

[1] Dies ist ein Spezialfall des allgemeineren zentralen Grenzwertsatzes nach Lindeberg und Levy. Siehe auch:
`http://mitschriebwiki.nomeata.de/Stochastik2.pdf.6.pdf`

Statistische Tests von Hypothesen

- Wozu statistische Signifikanztests?

Statistische Tests von Hypothesen

- Wozu statistische Signifikanztests?
- Statistische Signifikanztests sind Verfahren, mit denen zufällige Fluktuation oder "Rauschen" in Daten und realen Effekten unterschieden werden sollen

Statistische Tests von Hypothesen

- Wozu statistische Signifikanztests?
- Statistische Signifikanztests sind Verfahren, mit denen zufällige Fluktuation oder "Rauschen" in Daten und realen Effekten unterschieden werden sollen
- Ohne statistische Tests ist es sehr leicht, zufällige Muster in Daten fälschlicherweise als reale Effekte zu interpretieren

Statistische Tests von Hypothesen

- Studie zur Wirksamkeit eines Antibiotikums:

Statistische Tests von Hypothesen

- Studie zur Wirksamkeit eines Antibiotikums:
  - 100 Patienten erkrankt, ohne Behandlung mit Antibiotikum:

Statistische Tests von Hypothesen

- Studie zur Wirksamkeit eines Antibiotikums:
  - 100 Patienten erkrankt, ohne Behandlung mit Antibiotikum:
    - Beobachteter Erwartungswert von 1.2 Krankheitserregern pro Milliliter Blut

Statistische Tests von Hypothesen

- Studie zur Wirksamkeit eines Antibiotikums:
  - 100 Patienten erkrankt, ohne Behandlung mit Antibiotikum:
    - Beobachteter Erwartungswert von 1.2 Krankheitserregern pro Milliliter Blut
  - 100 Patienten erkrankt, Behandlung mit Antibiotikum:

Statistische Tests von Hypothesen

- Studie zur Wirksamkeit eines Antibiotikums:
  - 100 Patienten erkrankt, ohne Behandlung mit Antibiotikum:
    - Beobachteter Erwartungswert von 1.2 Krankheitserregern pro Milliliter Blut
  - 100 Patienten erkrankt, Behandlung mit Antibiotikum:

Statistische Tests von Hypothesen

- Studie zur Wirksamkeit eines Antibiotikums:
  - 100 Patienten erkrankt, ohne Behandlung mit Antibiotikum:
    - Beobachteter Erwartungswert von 1.2 Krankheitserregern pro Milliliter Blut
  - 100 Patienten erkrankt, Behandlung mit Antibiotikum:
    - 1 Dosis, Wartezeit 1 Stunde

Statistische Tests von Hypothesen

- Studie zur Wirksamkeit eines Antibiotikums:
  - 100 Patienten erkrankt, ohne Behandlung mit Antibiotikum:
    - Beobachteter Erwartungswert von 1.2 Krankheitserregern pro Milliliter Blut
  - 100 Patienten erkrankt, Behandlung mit Antibiotikum:
    - 1 Dosis, Wartezeit 1 Stunde
    - Beobachteter Erwartungswert von 1.05 Krankheitserregern pro Milliliter Blut
  - Standardabweichungen beider Gruppen: 0.5 Krankheitserreger pro Milliliter Blut

Statistische Tests von Hypothesen

- Studie zur Wirksamkeit eines Antibiotikums:
  - 100 Patienten erkrankt, ohne Behandlung mit Antibiotikum:
    - Beobachteter Erwartungswert von 1.2 Krankheitserregern pro Milliliter Blut
  - 100 Patienten erkrankt, Behandlung mit Antibiotikum:
    - 1 Dosis, Wartezeit 1 Stunde
    - Beobachteter Erwartungswert von 1.05 Krankheitserregern pro Milliliter Blut
  - Standardabweichungen beider Gruppen: 0.5 Krankheitserreger pro Milliliter Blut
- H_0: (Nullhypothese): Das Antibiotikum hat keine positive Wirkung

Statistische Tests von Hypothesen

- Studie zur Wirksamkeit eines Antibiotikums:
  - 100 Patienten erkrankt, ohne Behandlung mit Antibiotikum:
    - Beobachteter Erwartungswert von 1.2 Krankheitserregern pro Milliliter Blut
  - 100 Patienten erkrankt, Behandlung mit Antibiotikum:
    - 1 Dosis, Wartezeit 1 Stunde
    - Beobachteter Erwartungswert von 1.05 Krankheitserregern pro Milliliter Blut
  - Standardabweichungen beider Gruppen: 0.5 Krankheitserreger pro Milliliter Blut
- H_0: (Nullhypothese): Das Antibiotikum hat keine positive Wirkung
- H_1: Das Antibiotikum verringert die Konzentration der Krankheitserreger

Statistische Tests: Der Z-Test für normalverteilte Daten

- $\bar{\mu}$: Empirischer Erwartungswert der Hintergrund-Population: 1.2

Statistische Tests: Der Z-Test für normalverteilte Daten

- $\bar{\mu}$: Empirischer Erwartungswert der Hintergrund-Population: 1.2
- $\hat{\mu}$: Empirischer Erwartungswert der Stichprobe: 1.05

Statistische Tests: Der Z-Test für normalverteilte Daten

- $\bar{\mu}$: Empirischer Erwartungswert der Hintergrund-Population: 1.2
- $\hat{\mu}$: Empirischer Erwartungswert der Stichprobe: 1.05
- Standardabweichung der Hintergrund-Population ($\bar{\sigma}$) und Stichprobe ($\hat{\sigma}$): 0.5

Statistische Tests: Der Z-Test für normalverteilte Daten

- $\bar{\mu}$: Empirischer Erwartungswert der Hintergrund-Population: 1.2
- $\hat{\mu}$: Empirischer Erwartungswert der Stichprobe: 1.05
- Standardabweichung der Hintergrund-Population ($\bar{\sigma}$) und Stichprobe ($\hat{\sigma}$): 0.5
- Wir nehmen an, dass die Population und Stichprobe normalverteilt sind (Komplexes Phänomen, kumulativer Effekt vieler Zufallsvariablen)

Statistische Tests: Der Z-Test für normalverteilte Daten

- $\bar{\mu}$: Empirischer Erwartungswert der Hintergrund-Population: 1.2
- $\hat{\mu}$: Empirischer Erwartungswert der Stichprobe: 1.05
- Standardabweichung der Hintergrund-Population ($\bar{\sigma}$) und Stichprobe ($\hat{\sigma}$): 0.5
- Wir nehmen an, dass die Population und Stichprobe normalverteilt sind (Komplexes Phänomen, kumulativer Effekt vieler Zufallsvariablen)
- Der Erwartungswert einer zufälligen Stichprobe der Größe N einer normalverteilten Zufallsvariablen mit Erwartungswert μ und Standardabweichung σ ist selbst normalverteilt mit Erwartungswert μ und Standardabweichung s_N:

$$s_N = \frac{\sigma}{\sqrt{N}}$$

Statistische Tests: Der Z-Test für normalverteilte Daten

- Daraus folgt:

Statistische Tests: Der Z-Test für normalverteilte Daten

- Daraus folgt:

$$\bar{s}_{N=100} = \frac{\bar{\sigma}}{\sqrt{100}} = \frac{0.5}{10} = 0.05$$

- Unter H_0 ist die Stichprobe eine zufällige Stichprobe aus $N(\bar{\mu}, \bar{\sigma}^2)$

Statistische Tests: Der Z-Test für normalverteilte Daten

- Daraus folgt:

$$\bar{s}_{N=100} = \frac{\bar{\sigma}}{\sqrt{100}} = \frac{0.5}{10} = 0.05$$

- Unter H_0 ist die Stichprobe eine zufällige Stichprobe aus $N(\bar{\mu}, \bar{\sigma}^2)$
- $\hat{\mu}$ weicht also um drei Standardabweichungen (3×0.05) von seinem Erwartungswert unter H_0 ab:

$$\bar{\mu} - \hat{\mu} = 1.2 - 1.05 = 0.15 = 3 \times 0.05 = 3 \times \bar{s}_{N=100}$$

Statistische Tests: Der Z-Test für normalverteilte Daten

- Daraus folgt:

$$\bar{s}_{N=100} = \frac{\bar{\sigma}}{\sqrt{100}} = \frac{0.5}{10} = 0.05$$

- Unter H_0 ist die Stichprobe eine zufällige Stichprobe aus $N(\bar{\mu}, \bar{\sigma}^2)$
- $\hat{\mu}$ weicht also um drei Standardabweichungen (3×0.05) von seinem Erwartungswert unter H_0 ab:

$$\bar{\mu} - \hat{\mu} = 1.2 - 1.05 = 0.15 = 3 \times 0.05 = 3 \times \bar{s}_{N=100}$$

- Mithilfe dieser sogenannten Z-Statistik lässt sich aus einer Tabelle ablesen, dass die Wahrscheinlichkeit für H_0 nur etwa 0.1% beträgt.

Statistische Tests: Der Z-Test für normalverteilte Daten

- Daraus folgt:

$$\bar{s}_{N=100} = \frac{\bar{\sigma}}{\sqrt{100}} = \frac{0.5}{10} = 0.05$$

- Unter H_0 ist die Stichprobe eine zufällige Stichprobe aus $N(\bar{\mu}, \bar{\sigma}^2)$
- $\hat{\mu}$ weicht also um drei Standardabweichungen (3×0.05) von seinem Erwartungswert unter H_0 ab:

$$\bar{\mu} - \hat{\mu} = 1.2 - 1.05 = 0.15 = 3 \times 0.05 = 3 \times \bar{s}_{N=100}$$

- Mithilfe dieser sogenannten Z-Statistik lässt sich aus einer Tabelle ablesen, dass die Wahrscheinlichkeit für H_0 nur etwa 0.1% beträgt.
- Zu unterscheiden sind einseitige von zweiseitigen Z-Tests: (*Wirkung* vs. *positive Wirkung*)

Normalverteilung

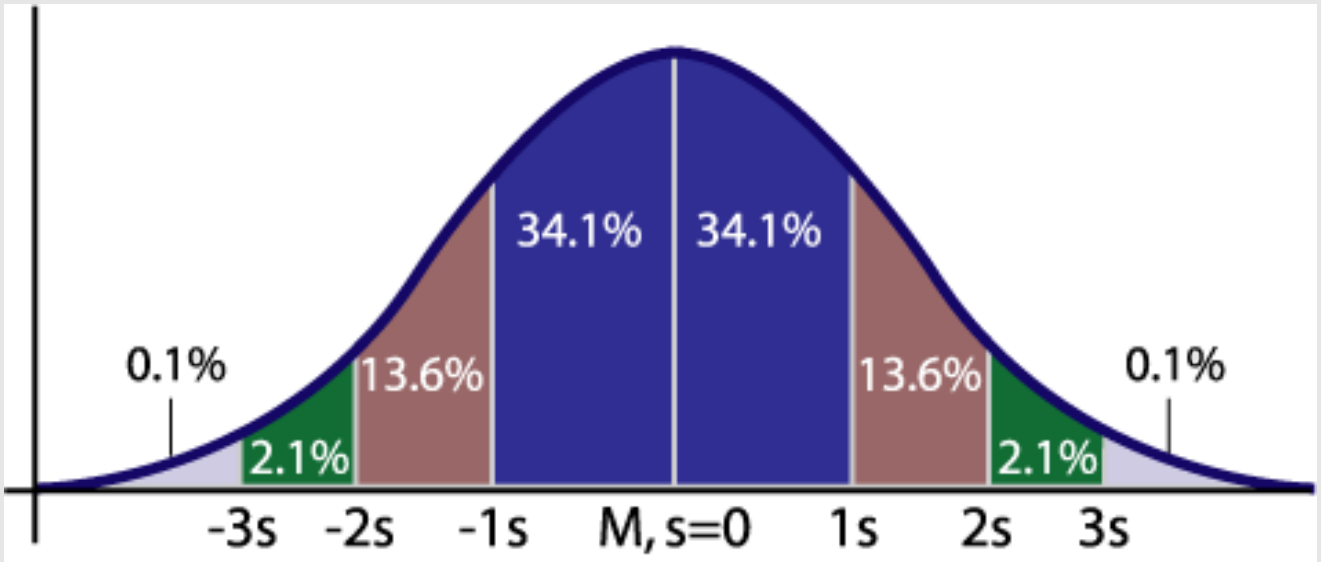

Abbildung: Normalverteilung mit Standardabweichungen und deren Wahrscheinlichkeitsmassen: `https://commons.wikimedia.org/wiki/File:Diagramma_standaardafwijking.png`

Statistische Tests: Anmerkungen

- Die errechnete Wahrscheinlichkeit von H_0 wird allgemein als p-Wert bezeichnet

Statistische Tests: Anmerkungen

- Die errechnete Wahrscheinlichkeit von H_0 wird allgemein als p-Wert bezeichnet
- p-Werte von unter 0.05 gelten in den meisten empirischen Wissenschaften als *statistisch signifikant* (Reine Konvention, kein Naturgesetz)

Statistische Tests: Anmerkungen

- Die errechnete Wahrscheinlichkeit von H_0 wird allgemein als p-Wert bezeichnet
- p-Werte von unter 0.05 gelten in den meisten empirischen Wissenschaften als *statistisch signifikant* (Reine Konvention, kein Naturgesetz)
- Weitere wichtige statistische Tests:

Statistische Tests: Anmerkungen

- Die errechnete Wahrscheinlichkeit von H_0 wird allgemein als p-Wert bezeichnet
- p-Werte von unter 0.05 gelten in den meisten empirischen Wissenschaften als *statistisch signifikant* (Reine Konvention, kein Naturgesetz)
- Weitere wichtige statistische Tests:
  - Student'scher t-Test für kleine, pseudo-normalverteilte Datensätze

Statistische Tests: Anmerkungen

- Die errechnete Wahrscheinlichkeit von H_0 wird allgemein als p-Wert bezeichnet
- p-Werte von unter 0.05 gelten in den meisten empirischen Wissenschaften als *statistisch signifikant* (Reine Konvention, kein Naturgesetz)
- Weitere wichtige statistische Tests:
  - Student'scher t-Test für kleine, pseudo-normalverteilte Datensätze
  - χ^2-Test (Engl.: "Chi-squared Test") für kategoriale Datensätze

Statistische Tests: Anmerkungen

- Computerlinguistische Daten folgen selten Normalverteilungen, in solche Fällen kann eine Technik Namens *Bootstrap-Resampling* verwendet werden, um normalverteilte Statistiken für Signifikanztests zu erzeugen.
  - Mehr dazu zum Beisiel hier:[2] `http://www.cl.uni-heidelberg.de/~riezler/publications/papers/ACL05WS.pdf`

[2] Stefan Riezler and John Maxwell: "On Some Pitfalls in Automatic Evaluation and Significance Testing for MT". In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Methods for MT and Summarization (MTSE) at the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Ann Arbor, Michigan., 2005

Noch Fragen?

Vielen Dank für die Aufmerksamkeit!

Weiterführende Literatur

Karl Oelschläger, *Einführung in die Wahrscheinlichkeitstheorie und die Statistik*, Vorlesungsskript, Universität Heidelberg, 2016. Kapitel 7. `http://www.math.uni-heidelberg.de/studinfo/oelschlaeger/Einf_WTheorie_Statistik_SS_18/Einf.WTheorie.Statistik.Skript.SS_16.pdf`