Stochastik

Kapitel 1: Einführung und Motivation

Prof. Dr. Reinhard Klein
Folien von Prof. Dr. Michael Clausen

Sommersemester 2016

Institut für Informatik II, Universität Bonn

Stochastik: Grobüberblick

- **Stochastik**: Lehre von Zufallsprozessen
- Teilgebiete: Wahrscheinlichkeitstheorie und Statistik

- **W-Theorie**: untersucht Zufallsprozesse mit bekannten Wahrscheinlichkeiten
  - **Beispiel**: zweimaliges Würfeln mit fairem Würfel
  - typische Frage: Wie groß ist die Wahrscheinlichkeit, eine Augensumme ≥ 10 zu würfeln?

- **Statistik**: untersucht Zufallsprozesse mit unbekannten Wahrscheinlichkeiten
  - **Beispiel**: zweimaliges Würfeln mit unfairem Würfel
  - mögliche Aufgabe: schätze anhand vieler Versuche Wahrscheinlichkeiten ... etwa zu Prognosezwecken

Stochastik in der Informatik?

- Randomisierte Algorithmen
- Modellierung
- Simulation
- Informationsmaße
- Datenkompression
- Stochastische Suche
- Mustererkennung
- Data Mining
- Information Retrieval
- Artifical Life . . .

Randomisierte Algorithmen

- Randomisierte Algorithmen arbeiten mit Zufallsbits
- Zufall wirkt oft effizienzsteigernd
- Zufall führt oft zu einfachen Lösungen
- zwei wichtige Typen randomisierter Algorithmen sind
  Monte Carlo und **Las Vegas**
- **Monte-Carlo-Algorithmen** geben immer eine Antwort,
  ... aber nicht jede Antwort ist richtig!
- **Las-Vegas-Algorithmen** geben stets richtige Antworten,
  ... aber die Antworten lassen manchmal lange auf sich warten!
- Das Verhalten randomisierter Algorithmen klingt zunächst absurd, wird aber durch folgende Beispiele ins rechte Licht gerückt.

Ein Monte-Carlo-Gleichheitstest

Problemstellung: Zu drei gegebenen $n \times n$ Matrizen A, B, C entscheide, ob $A \cdot B = C$ ist.

Naive Lösung: direkte algorithmische Umsetzung der Definition des Matrizenproduktes mit anschließenden Koeffizientenvergleichen:

$$c_{ik} \stackrel{?}{=} \sum_{j=1}^{n} a_{ij} \cdot b_{jk}.$$

Arithmetischer Aufwand: $\Theta(n^3)$

Aufwand an Vergleichen: $\Theta(n^2)$

Geht es schneller? **Ja, randomisiert!**

Freivalds Algorithmus (1)

1. Wähle Zufallsvektor $x \in \{0,1\}^n$.
2. Berechne $A \cdot (B \cdot x)$ und $C \cdot x$.
3. Ist $A \cdot (B \cdot x) \neq C \cdot x$, antworte mit $AB \neq C$.
4. Ist $A \cdot (B \cdot x) = C \cdot x$, antworte mit $AB = C$.

Bemerkungen:

- Arithmetischer Aufwand $\Theta(n^2)$
- Aufwand an Vergleichen $O(n)$
- $A \cdot (B \cdot x) = (A \cdot B) \cdot x$ (Assoziativität)
- Antwort $AB \neq C$ unproblematisch, da $ABx \neq Cx$.
- Antwort $AB = C$ problematisch, da von $ABx = Cx$ auf $ABy = Cy$ für alle Zufallsvektoren y geschlossen wurde.

Das Urnenmodell

... dient der Veranschaulichung des weiteren Vorgehens.

Eine Urne $U_{A,B,C}$ enthält zu jedem Zufallsvektor $x \in \{0,1\}^n$ eine gefärbte Kugel K_x:

- falls $ABx = Cx$, ist K_x blau,
- falls $ABx \neq Cx$, ist K_x rot.

Im Fall $AB = C$ sind **alle** Kugeln blau.

Satz von Freivalds

Im Fall $AB \neq C$ ist höchstens die Hälfte aller Kugeln blau.

Es folgt ein einfacher Beweis. Eine schärfere Aussage wird in den Übungen besprochen.

Satz von Freivalds: Beweis

- Sei $D := AB - C \neq 0$. Dann existiert eine Matrixposition (i, j) mit $d_{ij} \neq 0$.
- Sei $R_j := \{x \in \{0,1\}^n \mid x_j = 0\}$ und $e_j :=$ j-ter Einheitsvektor.
- Dann ist $|R_j| = 2^{n-1} = |\{0,1\}^n|/2$ und für $x \in R_j$ ist $x + e_j \in \{0,1\}^n \setminus R_j$.
- **Behauptung**: Für jedes $x \in R_j$ ist $Dx \neq 0$ oder $D(x + e_j) \neq 0$.
  - Im Fall $Dx \neq 0$ gilt die Behauptung.
  - Im Fall $Dx = 0$ ist

$$D(x + e_j) = Dx + De_j = 0 + (d_{1j}, \dots, d_{ij}, \dots, d_{nj})^\top \neq 0.$$

- Also ist zu **jedem** $x \in R_j$ mindestens einer der beiden Vektoren $x, x + e_j$ ein Zeuge für $AB \neq C$.
- Schließlich beachte man, dass $\{0,1\}^n = R_j \sqcup (R_j + e_j)$.

Das beweist den Satz von Freivalds. $\square$

Freivalds Algorithmus (2)

1. Wähle eine positive ganze Zahl k.
2. Greife k-mal unabhängig in die Urne $U_{A,B,C}$.
3. Sind alle gezogenen Kugeln blau, so antworte mit $AB = C$.
4. Ist eine gezogene Kugel rot, so antworte mit $AB \neq C$.

Bemerkungen:

- Arithmetischer Aufwand $\Theta(k \cdot n^2)$
- Aufwand an Vergleichen $O(k \cdot n)$
- Antwort $AB \neq C$ unproblematisch.
- Antwort $AB = C$ bei hinreichend großem k marginal problematisch, da im Fall $AB \neq C$ die Wahrscheinlichkeit, immer eine blaue Kugel zu greifen, höchstens 2^{-k} ist.

Ein Färbungsproblem: konkret

- Wir kommen nun zu einem Problem, das wir letztlich mit einem Las-Vegas Algorithmus auf erstaunlich einfache und schnelle Art lösen werden.
- Die Fachabteilung F_j hat entweder einen Experten oder eine Expertin in **alle** in Spalte j mit • gekennzeichneten Arbeitsgruppen A_i abzuordnen.

	F_1	F_2	F_3	F_4	F_5	F_6	F_7
A_1	•		•	•		•	
A_2	•	•		•			•
A_3		•	•		•	•	
A_4	•			•	•		•
A_5	•	•	•			•	
A_6			•	•	•		•
A_7	•		•		•		•
A_8		•		•	•	•	

Frage: Gibt es eine Gesamtabordnung, so dass in den Arbeitsgruppen keine reinen Herrenrunden und keine reinen Damenkränzchen entstehen?

Eine zulässige Färbung

	F_1	F_2	F_3	F_4	F_5	F_6	F_7
A_1	$\color{blue}\bullet$		$\color{red}\bullet$	$\color{blue}\bullet$		$\color{blue}\bullet$	
A_2	$\color{blue}\bullet$	$\color{red}\bullet$		$\color{blue}\bullet$			$\color{blue}\bullet$
A_3		$\color{red}\bullet$	$\color{red}\bullet$		$\color{red}\bullet$	$\color{blue}\bullet$	
A_4	$\color{blue}\bullet$			$\color{blue}\bullet$	$\color{red}\bullet$		$\color{blue}\bullet$
A_5	$\color{blue}\bullet$	$\color{red}\bullet$	$\color{red}\bullet$			$\color{blue}\bullet$	
A_6			$\color{red}\bullet$	$\color{blue}\bullet$	$\color{red}\bullet$		$\color{blue}\bullet$
A_7	$\color{blue}\bullet$		$\color{red}\bullet$		$\color{red}\bullet$		$\color{blue}\bullet$
A_8		$\color{red}\bullet$		$\color{blue}\bullet$	$\color{red}\bullet$	$\color{blue}\bullet$	

Obige Färbung ist zulässig, denn alle Spalten sind einfarbig, während alle Zeilen zweifarbig sind.

Ein Färbungsproblem: abstrakt

- Gegeben sei eine n-elementige Menge F.
- konkret: $F = [1:7]$ identifiziert die Fachabteilungen $F_1, \ldots, F_7$.
- Ferner seien $S_1, \ldots, S_m$ paarweise verschiedene r-elementige Teilmengen von F.
- konkret: S_i beschreibt die Zusammensetzung der i-ten Arbeitsgruppe. Im Beispiel sind etwa $S_1 = \{1,3,4,6\}$, $S_2 = \{1,2,4,7\}$ usw.
- **Aufgabe**: Färbe jedes Element von F jeweils mit einer von zwei Farben (rot, blau) so ein, dass keine der Teilmengen $S_1, S_2, \ldots, S_m$ einfarbig ist.
  [Eine solche Färbung heiße $(S_1, \ldots, S_m)$-zulässig.]
- konkret: S_i einfarbig rot: reines Damenkränzchen,
  S_i einfarbig blau: reine Herrenrunde.

Frage: Existiert überhaupt eine $(S_1, \ldots, S_m)$-zulässige Färbung und wenn ja, wie findet man schnell eine solche?

Ein Las-Vegas-Färbungsalgorithmus

1. **Eingabe**: n-elementige Menge F sowie r-elementige, paarweise verschiedene Teilmengen $S_1, \ldots, S_m$ von F.
2. Färbe die Elemente von F gemäß Gleichverteilung blau oder rot.
3. Teste, ob Gesamtfärbung $(S_1, \ldots, S_m)$-zulässig.
4. Falls zulässig, Färbung gefunden, sonst weiter mit 2.

Satz

Es sei $r \geq 2$ und $m \leq 2^{r-2}$. Dann gilt:

(1) Es gibt eine $(S_1, \ldots, S_m)$-zulässige Gesamtfärbung.
(2) Im Mittel ist der Las-Vegas-Algorithmus spätestens nach zwei (!) Färbungsversuchen erfolgreich.

Beachte: Effiziente Konstruktion durch Münzwurf!

Beweisskizze (1)

- Es sei Ω die Menge aller Gesamtfärbungen von F und R_i bzw. B_i sei die Teilmenge von Ω, bei denen die Elemente von S_i komplett rot bzw. blau gefärbt wurden. P sei die Gleichverteilung auf Ω. Dann gilt für die Wahrscheinlichkeit, dass alle Elemente von S_i komplett rot eingefärbt wurden:

$$P(R_i) = \frac{|R_i|}{|\Omega|} = 2^{-r}. \quad \text{Analog ist} \quad P(B_i) = \frac{|B_i|}{|\Omega|} = 2^{-r}.$$

- Die Wahrscheinlichkeit, dass ein fest vorgegebenes S_i komplett rot oder komplett blau gefärbt ist, beträgt $P(R_i \sqcup B_i) = P(R_i) + P(B_i) = 2 \cdot 2^{-r} = 2^{1-r}$.
- Für die Wahrscheinlichkeit q, dass mindestens eine der Mengen $S_1, \ldots, S_m$ einfarbig ist (S_i einfarbig: $E_i := R_i \sqcup B_i$), ergibt sich mit der Voraussetzung $m \leq 2^{r-2}$ die Abschätzung

$$q = P(E_1 \cup \ldots \cup E_m) \leq P(E_1) + \ldots + P(E_m) \leq m \cdot 2^{1-r} \leq 2^{r-2} \cdot 2^{1-r} = \tfrac{1}{2}.$$

Dies beweist bereits die Existenz einer $(S_1, \ldots, S_m)$-zulässigen Gesamtfärbung, denn die Wahrscheinlichkeit $p := 1 - q$ für eine erfolgreiche Färbung ist positiv: $p \geq \frac{1}{2}$.

Beweisskizze (2)

- Die Wahrscheinlichkeit, beim k-ten Gesamtfärbungsversuch die erste erfolgreiche Gesamtfärbung zu erzielen, ist $p \cdot q^{k-1}$.
  - Hintergrund: mehrstufige Experimente, stochastische Unabhängigkeit
- Als Erwartungswert ($=$ gewichteter Mittelwert) für die Anzahl der Versuche bis zur ersten erfolgreichen Gesamtfärbung erhält man also

$$\sum_{k \geq 1} kpq^{k-1} = p \sum_{k \geq 1} kq^{k-1}.$$

- Um diesen Erwartungswert auszurechnen, betrachten wir die für $|x| < 1$ gültige Beziehung der geometrischen Reihe

$$g(x) := \sum_{k \geq 0} x^k = \frac{1}{1-x}.$$

Beweisskizze (3)

- Für die Ableitung von $g(x) = (1-x)^{-1}$ ergibt sich mit der Kettenregel einerseits

$$g'(x) = \frac{\mathrm{d}}{\mathrm{d}x}(1-x)^{-1} = (1-x)^{-2},$$

- andererseits führt termweises Differenzieren von $g(x) = \sum_{k\geq 0} x^k$ zu

$$g'(x) = \sum_{k\geq 1} kx^{k-1}.$$

- Mit $p = 1-q$ erhalten wir als obere Schranke für die mittlere Anzahl von Färbungsversuchen bis zum ersten Erfolg:

$$\sum_{k\geq 1} kpq^{k-1} = p\cdot g'(q) = p\cdot(1-q)^{-2} = \tfrac{1}{p} \leq 2. \qquad \square$$

Nachbemerkungen

- Der Las-Vegas-Algorithmus löst auf sehr einfache Weise das Färbungsproblem.
- Die gerade bewiesene Ungleichung ist eine obere Schranke für den Mittelwert. In seltenen Fällen kann der Erfolg sehr lange auf sich warten lassen: Mit Wahrscheinlichkeit $pq^{k-1} > 0$ muss man k Schritte bis zum ersten Erfolg warten.
- Beachte: Wegen $p \leq 1$ und $q \leq \frac{1}{2}$ ist die Wahrscheinlichkeit $pq^{k-1} \leq 2^{1-k}$ für große k verschwindend klein!
- Beweisdetails werden nachgeliefert, sobald wir erste Grundkenntnisse in W-Theorie erworben haben.

Stochastik

Kapitel 2: Wahrscheinlichkeitsräume

Prof. Dr. Reinhard Klein
Folien von Prof. Dr. Michael Clausen

Sommersemester 2016

INSTITUT FÜR INFORMATIK II, UNIVERSITÄT BONN

Wahrscheinlichkeitsräume: Grobübersicht

Ein **W-Raum** modelliert ein Zufallsexperiment durch Angabe von drei Komponenten:

$$(\Omega, \mathcal{A}, P).$$

- Ω: **Ergebnisraum**
- $\mathcal{A}$: **Ereignis-Algebra**
- P: **W-Maß**

Im folgenden gehen wir auf die einzelnen Komponenten ein.

2.1 Ergebnisräume

Bei der **Modellierung von Zufallsexperimenten** hat man zunächst u.a. folgende Fragen zu beantworten:

- Wie soll der Ausgang eines Experiments beschrieben werden?
- Was ist vom jeweiligen Ausgang des Experiments von Interesse?
- Wie kann man die Gesamtheit aller relevanten Resultate mengentheoretisch beschreiben?

Dies illustrieren wir anhand von drei Beispielen.

Beispiel 1: Einmaliges Würfeln

- relevant: oben sichtbare Augenzahl
- irrelevant: Position des Würfels auf dem Tisch
- **Ergebnismenge**:

$$\Omega = [1 : 6].$$

Beispiel 2: n-maliges Würfeln

Variante 1:

- relevant: die Folge der gewürfelten Augenzahlen
- **Ergebnismenge**:

$$\Omega_n = [1:6]^n = \{(\omega_1, \ldots, \omega_n) \mid \forall i : \omega_i \in [1:6]\}.$$

Variante 2 (gröbere Sichtweise):

- relevant: Häufigkeit der gewürfelten Augenzahlen
- **Ergebnismenge**:

$$\Omega'_n = \{(h_1, \ldots, h_6) \in [0:n]^6 \mid h_1 + \ldots + h_6 = n\}.$$

Ω'_n ist dabei das Bild der Vergröberungsabbildung

$$\Omega_n \ni \omega \mapsto (h_j := |\{i : \omega_i = j\}|)_{j \in [1:6]},$$

die jeder Augenzahlfolge das zugehörige **Histogramm** zuordnet.

Beispiel 3: Münzwurffolgen

- unendlich oft wiederholtes Münzwurfexperiment
- **Idealisierung**: gut zum Studium von Gesetzmäßigkeiten für sehr lange Münzwurfreihen
- **Ergebnismenge**:

$$\Omega = \{0,1\}^{\mathbb{N}} = \{\omega = (\omega_i)_{i\in\mathbb{N}} \mid \forall i : \omega_i \in \{0,1\}\}.$$

Satz 1.1. $\Omega = \{0,1\}^{\mathbb{N}}$ ist überabzählbar unendlich.

Beweis. Wäre $\{0,1\}^{\mathbb{N}}$ abzählbar unendlich, so könnten wir die Elemente von $\{0,1\}^{\mathbb{N}}$ durch eine binärwertige $\mathbb{N} \times \mathbb{N}$-Matrix (ω_{ij}) beschreiben. Dabei gehört die i-te Zeile dieser Matrix zum i-ten Element von $\{0,1\}^{\mathbb{N}}$. Mit $\overline{1} := 0$ und $\overline{0} := 1$ kann die Folge

$$(\overline{\omega_{11}}, \overline{\omega_{22}}, \ldots)$$

der negierten Diagonalelemente in keiner Zeile der Matrix vorkommen. $\square$

Ergebnisräume: Fazit

Die relevanten Ergebnisse eines Zufallsexperiments werden zu einer Menge Ω zusammengefasst.

Diese Menge Ω heißt **Ergebnisraum** oder **Stichprobenraum**.

- Damit haben wir die erste Komponente $(\Omega, \mathcal{A}, P)$ eines W-Raums erläutert.
- Wir wenden uns nun der Erläuterung der zweiten Komponente $\mathcal{A}$ zu.

2.2 Ereignisse

- **Ereignisse** entsprechen Teilmengen A des Ergebnisraums Ω.
- Ist $\omega \in \Omega$ das Ergebnis eines Zufallsexperiments, so sagt man im Fall
  - $\omega \in A$: das Ereignis A ist **eingetreten**.
  - $\omega \notin A$: das Ereignis A ist **nicht eingetreten**.

Beispiele für Ereignisse:

- Beim einmaligen Würfeln beschreibt die Teilmenge $A := \{2, 4, 6\}$ des Ergebnisraums $\Omega = [1 : 6]$ das Ereignis

Es wurde eine gerade Augenzahl gewürfelt.

- Das Ereignis *Bei n Münzwürfen fällt mindestens k-mal Zahl* wird beschrieben durch die Teilmenge

$$A = \{\omega \in \{0, 1\}^n \mid \omega_1 + \ldots + \omega_n \geq k\}$$

des Ergebnisraums $\Omega = \{0, 1\}^n$. [$0 \cong$ Kopf, $1 \cong$ Zahl.]

Las-Vegas-Beispiel: Ergebnisraum & Ereignisse

- **Ergebnisraum**: alle rot-blau-Färbungen der n-elementigen Menge F:

$$\Omega = \{\varphi \mid \varphi : F \to \{\text{rot,blau}\}\} =: \{\text{rot,blau}\}^F.$$

- **Wichtige Ereignisse**, die im Verlauf des Beweises benutzt wurden:
(Memo: $S_1, \ldots, S_m \subset F$ sind r-elementig & paarweise verschieden)

$$\begin{aligned}
R_i &:= \{\varphi \in \Omega \mid \varphi \downarrow S_i \equiv \text{rot}\} \\
&\ \text{Menge aller Färbungen, bei denen } S_i \text{ komplett rot gefärbt ist,} \\
B_i &:= \{\varphi \in \Omega \mid \varphi \downarrow S_i \equiv \text{blau}\} \\
&\ \text{Menge aller Färbungen, bei denen } S_i \text{ komplett blau gefärbt ist,} \\
E_i &:= R_i \sqcup B_i \\
&\ \text{Menge aller Färbungen, bei denen } S_i \text{ einfarbig ist,} \\
E &:= E_1 \cup \ldots \cup E_m \\
&\ \text{Menge aller Färbungen, bei denen mindestens ein } S_i \text{ einfarbig ist.}
\end{aligned}$$

- Diese Ereignisse wurden später im Beweis wahrscheinlichkeitstheoretisch bewertet:
$$P(R_i), \quad P(B_i), \quad P(R_i \sqcup B_i), \quad P(E_1 \cup \ldots \cup E_m).$$

Ereignissysteme

- **Ziel**: Festlegung eines Systems $\mathcal{A}$ von Ereignissen, so dass man jedem Ereignis $A \in \mathcal{A}$ in konsistenter Weise eine Wahrscheinlichkeit $P(A) \in [0,1]$ für das Eintreten von A zuordnen kann.
- Ein Ereignissystem $\mathcal{A}$ ist zunächst einmal eine Teilmenge der **Potenzmenge** von Ω:

$$\mathcal{A} \subseteq 2^{\Omega} := \{X \mid X \subseteq \Omega\}.$$

  Warum es nicht sinnvoll ist, immer die Potenzmenge als Ereignissystem zu wählen, wird in den Übungen diskutiert.
- Gewisse mengentheoretische Operationen mit Ereignissen aus $\mathcal{A}$ sollten wieder in $\mathcal{A}$ liegen, damit man das Resultat der Operation wieder w-theoretisch bewerten kann.
- Es hat sich herausgestellt, dass sich solche Mengensysteme $\mathcal{A} \subseteq 2^{\Omega}$ eignen, die Ω enthalten und unter Komplement-, sowie endlicher und abzählbar unendlicher Vereinigungs- und Durchschnittsbildung abgeschlossen sind.
- Mengensysteme mit diesen Eigenschaften nennt man σ-**Algebren**. Damit beschäftigen wir uns als nächstes.

σ-Algebren als Ereignissysteme

Definition. Es sei Ω eine nichtleere Menge. Ein System $\mathcal{A}$ von Teilmengen von Ω heißt σ-**Algebra über** Ω, wenn gilt:

(a) Das sichere Ereignis gehört zu $\mathcal{A}$: $\Omega \in \mathcal{A}$.

(b) Das Gegenereignis eines Ereignisses aus $\mathcal{A}$ gehört wieder zu $\mathcal{A}$:

$$A \in \mathcal{A} \Rightarrow A^c := \Omega \setminus A \in \mathcal{A}.$$

(c) $\mathcal{A}$ ist unter abzählbar unendlicher Vereinigungsbildung abgeschlossen:

$$A_1, A_2, \ldots \in \mathcal{A} \Rightarrow \bigcup_{i \geq 1} A_i \in \mathcal{A}.$$

Das Paar $(\Omega, \mathcal{A})$ heißt dann auch **Ereignisraum** oder **Messraum**.

Ist $(\Omega, \mathcal{A})$ ein Ereignisraum, so heißen die Elemente $A \in \mathcal{A}$ auch die **messbaren Ereignisse**.

Erste Eigenschaften von σ-Algebren

Satz 1.2. Es sei $\mathcal{A}$ eine σ-Algebra über Ω. Dann gilt:

(a) $\emptyset \in \mathcal{A}$.
(b) $A, B \in \mathcal{A}$ impliziert $A \cup B \in \mathcal{A}$, $A \cap B \in \mathcal{A}$, $A \setminus B \in \mathcal{A}$.
(c) $\mathcal{A}$ ist unter abzählbar unendlicher Durchschnittsbildung abgeschlossen.

Beweis. (a) Da $\Omega \in \mathcal{A}$ und $\mathcal{A}$ unter Komplementbildung abgeschlossen ist, folgt $\emptyset = \Omega^c \in \mathcal{A}$.

(b) Die Aussagen folgen aus

$$\begin{aligned} A \cup B &= A \cup B \cup \emptyset \cup \emptyset \cup \ldots \in \mathcal{A}, \\ A \cap B &= (A^c \cup B^c)^c \in \mathcal{A}, \\ A \setminus B &= A \cap B^c \in \mathcal{A}. \end{aligned}$$

(c) Übung! □

Fazit: Eine σ-Algebra $\mathcal{A}$ über Ω ist ein Mengensystem $\mathcal{A} \subseteq 2^\Omega$, das Ω enthält und unter Komplementbildung sowie unter endlicher und abzählbar unendlicher Durchschnitts- und Vereinigungsbildung abgeschlossen ist.

Das System aller σ-Algebren über Ω (1)

Satz 1.3. Bezeichnet $\Sigma(\Omega) := \{\mathcal{A} \mid \mathcal{A} \text{ ist } \sigma\text{-Algebra über } \Omega\}$ das System aller σ-Algebren über Ω, so gilt:

(1) $\Sigma(\Omega)$ ist bezüglich Mengeninklusion halbgeordnet.
(2) $\{\emptyset, \Omega\}$ ist die kleinste σ-Algebra über Ω.
(3) 2^Ω ist die größte σ-Algebra über Ω.
(4) Der Durchschnitt beliebig vieler σ-Algebren über Ω ist wieder eine σ-Algebra über Ω.
(5) Zu jedem System $\mathcal{G}$ von Teilmengen von Ω gibt es eine kleinste, $\mathcal{G}$ enthaltende σ-Algebra über Ω. Diese von $\mathcal{G}$ erzeugte σ-Algebra $\sigma(\mathcal{G})$ ist der Durchschnitt aller $\mathcal{G}$ enthaltenden σ-Algebren über Ω:

$$\sigma(\mathcal{G}) = \bigcap_{\mathcal{G} \subseteq \mathcal{A} \in \Sigma(\Omega)} \mathcal{A}.$$

Das System aller σ-Algebren über Ω (2)

Beweis. **(1)** $\Sigma(\Omega)$ ist bezüglich Mengeninklusion halbgeordnet, da Mengeninklusion eine reflexive, transitive und antisymmetrische Relation ist.

(2) Jede σ-Algebra über Ω enthält Ω und die leere Menge. Andererseits ist $\{\emptyset, \Omega\}$ eine σ-Algebra über Ω.

(3) 2^Ω ist die größte σ-Algebra über Ω: klar.

(4) Ist $(\mathcal{A}_i)_{i\in I}$ eine Familie von σ-Algebren über Ω, so ist auch $\mathcal{A} := \cap_{i\in I}\mathcal{A}_i$ wieder eine σ-Algebra über Ω:

- Da Ω in allen $\mathcal{A}_i$ liegt, gilt auch $\Omega \in \mathcal{A}$.
- Liegt A in $\mathcal{A} = \cap_{i\in I}\mathcal{A}_i$, so liegt A auch in allen $\mathcal{A}_i$. Da jedes $\mathcal{A}_i$ unter Komplementbildung abgeschlossen ist, liegt A^c in allen $\mathcal{A}_i$, also auch in $\mathcal{A}$.
- Ist schließlich $A_1, A_2, \ldots$ eine abzählbar unendliche Folge von Elementen aus $\mathcal{A}$, so ist $A_1, A_2, \ldots$ auch eine Folge von Elementen in $\mathcal{A}_i$, für alle $i \in I$. Daher liegt $\cup_{j\geq 1}A_j$ in $\mathcal{A}_i$, für alle $i \in I$. Folglich liegt $\cup_{j\geq 1}A_j$ auch in $\mathcal{A}$.

(5) ist klar wegen (4). $\square$

Beispiele von σ-Algebren

- Im Fall $\emptyset \neq A \subset \Omega$ gilt für die von $\mathcal{G} = \{A\}$ erzeugte σ-Algebra:

$$\sigma(\mathcal{G}) = \{\emptyset, A, A^c, \Omega\}.$$

- Ist Ω höchstens abzählbar unendlich, so ist die von $\mathcal{G} := \{\{\omega\} \mid \omega \in \Omega\}$ erzeugte σ-Algebra die Potenzmenge von Ω:

$$\sigma(\mathcal{G}) = 2^{\Omega}.$$

- Sei $\Omega = \mathbb{R}^n$ und $\mathcal{G} := \{\prod_{i=1}^n [a_i, b_i] : a_i, b_i \in \mathbb{Q} \wedge a_i < b_i\}$ das System aller achsenparallelen kompakten Quader in $\mathbb{R}^n$ mit rationalen Endpunkten. Dann heißt

$$\mathcal{B}^n := \sigma(\mathcal{G})$$

die **Borelsche σ-Algebra** auf $\mathbb{R}^n$ und jedes $A \in \mathcal{B}^n$ heißt eine **Borel-Menge**.

Typische σ-Algebren

Zwei Typen von σ-Algebren spielen für uns eine zentrale Rolle:

- **Diskreter Fall**: Ω ist höchstens abzählbar unendlich. Hier setzt man $\mathcal{A} = 2^{\Omega}$.
- **Reeller Fall**: $\emptyset \neq \Omega \subseteq \mathbb{R}^n$. Dann wählt man

$$\mathcal{A} = \mathcal{B}^n_\Omega := \{A \cap \Omega \mid A \in \mathcal{B}^n\}.$$

Übung: $\mathcal{B}^n_\Omega$ ist eine σ-Algebra.

Damit haben wir die zweite Komponente eines W-Raums $(\Omega, \mathcal{A}, P)$ hinreichend diskutiert und kommen nun zur letzten und entscheidenden Komponente.

2.3 Wahrscheinlichkeitsmaße

W-Maße P bewerten den Grad der Wahrscheinlichkeit eines Ereignisses A durch eine Maßzahl $P(A) \in [0,1]$. Je höher diese Zahl, desto wahrscheinlicher ist das Eintreten dieses Ereignisses.

Definition. Es sei $(\Omega, \mathcal{A})$ ein Messraum. Eine Funktion $P : \mathcal{A} \to [0,1]$ heißt ein **W-Maß** oder **W-Verteilung** auf $(\Omega, \mathcal{A})$, wenn gilt:

(N) Normierung: $P(\Omega) = 1$.

(A) σ-Additivität: Für paarweise disjunkte Ereignisse $A_1, A_2, \ldots \in \mathcal{A}$ gilt

$$P\Big(\bigsqcup_{i\geq 1} A_i\Big) = \sum_{i\geq 1} P(A_i).$$

Das Tripel $(\Omega, \mathcal{A}, P)$ heißt dann **Wahrscheinlichkeitsraum** (kurz: **W-Raum**).

Eigenschaften von W-Maßen (1)

Satz 1.4. Für beliebige Ereignisse $A, B, A_1, A_2, \ldots$ eines W-Raums $(\Omega, \mathcal{A}, P)$ gilt:

(1) $P(\emptyset) = 0$.

(2) $P(A \cup B) + P(A \cap B) = P(A) + P(B)$; insbesondere ist

$$P(A) + P(A^c) = 1.$$

(3) **Monotonie**: $A \subseteq B \Rightarrow P(A) \leq P(B)$.

(4) σ-**Subadditivität**: $P(\cup_{i\geq 1} A_i) \leq \sum_{i\geq 1} P(A_i)$.

(5) σ-**Stetigkeit**: Wenn $A_1 \subseteq A_2 \subseteq \ldots$ und $A = \cup_{i\geq 1} A_i$ bzw. $A_1 \supseteq A_2 \supseteq \ldots$ und $A = \cap_{i\geq 1} A_i$, so gilt

$$\lim_{n\to\infty} P(A_n) = P(A).$$

Es folgen die Beweise von (1) bis (4). Der Beweis von (5) wird in den Übungen besprochen.

Eigenschaften von W-Maßen (2)

Beweis von (1):

$\emptyset, \emptyset, \ldots$ ist Folge paarweise disjunkter Ereignisse. (A) ergibt:

$$P(\emptyset) = P(\emptyset \cup \emptyset \cup \ldots) = \sum_{i \geq 1} P(\emptyset).$$

Daraus folgt $P(\emptyset) = 0$.

Beweis von (2):

Sind A und B disjunkt, so folgt wegen (A) zusammen mit (1):

$$P(A \sqcup B) = P(A \sqcup B \sqcup \emptyset \sqcup \emptyset \sqcup \ldots) = P(A) + P(B) + 0 + 0 \ldots.$$

Allgemein gilt wegen $A \cup B = (A \setminus B) \sqcup (B \setminus A) \sqcup (A \cap B)$:

$$\begin{aligned} P(A \cup B) + P(A \cap B) &= P(A \setminus B) + P(B \setminus A) + 2P(A \cap B) \\ &= P(A) + P(B). \end{aligned}$$

Schließlich ist $1 = P(\Omega) = P(A \sqcup A^c) = P(A) + P(A^c)$.

Eigenschaften von W-Maßen (3)

Beweis von (3):
Aus $A \subseteq B$ folgt wegen (2) und der Nichtnegativität von Wahrscheinlichkeiten:

$$P(B) = P(A \sqcup (B \setminus A)) = P(A) + P(B \setminus A) \geq P(A).$$

Beweis von (4):

$$\begin{aligned} P(\bigcup_{i\geq 1} A_i) &= P(\bigsqcup_{i\geq 1}(A_i \setminus \cup_{j<i} A_j)) \\ &= \sum_{i\geq 1} P(A_i \setminus \cup_{j<i} A_j) \leq \sum_{i\geq 1} P(A_i). \end{aligned}$$

Beweis von (5): Übung! $\square$

Konstruktion von W-Maßen: diskreter Fall 1

Satz 1.5. Es sei Ω eine höchstens abzählbare Menge. Dann gilt:

(1) Jedes W-Maß P auf $(\Omega, 2^\Omega)$ ist durch die Folge $(P(\{\omega\}))_{\omega\in\Omega}$ bereits eindeutig bestimmt, denn für $A \subseteq \Omega$ ist

$$P(A) = \sum_{\omega\in A} P(\{\omega\}).$$

(2) Umgekehrt liefert jede Funktion $p : \Omega \to [0, 1]$ mit $\sum_{\omega\in\Omega} p(\omega) = 1$ vermöge

$$P(A) := \sum_{\omega\in A} p(\omega)$$

ein W-Maß auf Ω.

Definition. Eine Funktion $p : \Omega \to [0, 1]$ mit $\sum_{\omega\in\Omega} p(\omega) = 1$ heißt **W-Funktion**, **W-Vektor** oder **Zähldichte**.

Konstruktion von W-Maßen: diskreter Fall 2

Beweis von (1). Folgt aus der σ-Additivität und der Voraussetzung, dass Ω höchstens abzählbar unendlich ist.

Beweis von (2). Nach Definition von P ergibt sich: $P(\Omega) = \sum_{\omega \in \Omega} p(\omega) = 1$. Weiter ist p absolut summierbar und nach dem Umordnungssatz gilt für eine Folge $A_1, A_2, \ldots$ von paarweise disjunkten Teilmengen von Ω mit $A := \sqcup_{i \geq 1} A_i$ sofort

$$P(A) = \sum_{\omega \in A} p(\omega) = \sum_{i \geq 1} \Big(\sum_{\omega \in A_i} p(\omega) \Big) = \sum_{i \geq 1} P(A_i).$$

$\square$

Geometrischer Hintergrund: Konvexität

Wir wollen uns die Geometrie aller n-komponentigen Zähldichten $(p_1, \ldots, p_n)$ klarmachen. Dazu benötigen wir einige Begriffe über Konvexität.

- Eine Menge $K \subseteq \mathbb{R}^m$ heißt **konvex**, wenn mit zwei Elementen auch jedes Element der Verbindungsstrecke zu K gehört. Formal bedeutet das:

$$x, y \in K \quad \text{und} \quad \lambda \in [0,1] \quad \text{impliziert} \quad \lambda x + (1-\lambda) y \in K.$$

- Da $\mathbb{R}^m$ konvex ist und der Durchschnitt von konvexen Mengen wieder konvex ist, gibt es zu jeder Menge $X \subseteq \mathbb{R}^m$ eine kleinste, X enthaltende konvexe Menge, die sog. **konvexe Hülle** $\mathrm{conv}(X)$ von X. Diese ergibt sich als Durchschnitt aller X enthaltenden konvexen Teilmengen K von $\mathbb{R}^m$:

$$\mathrm{conv}(X) = \bigcap_{X \subseteq K,\ K \text{ konvex}} K.$$

- Ist $X = \{x_0, \ldots, x_r\}$ endlich, so kann man die konvexe Hülle von X auch beschreiben als Gesamtheit aller **Konvexkombinationen** von $x_0, \ldots, x_r \in \mathbb{R}^m$, das sind Vektoren der Form $\lambda_0 x_0 + \ldots + \lambda_r x_r$, wobei $(\lambda_0, \ldots, \lambda_r)$ eine Zähldichte ist.

Wahrscheinlichkeitssimplexe (1)

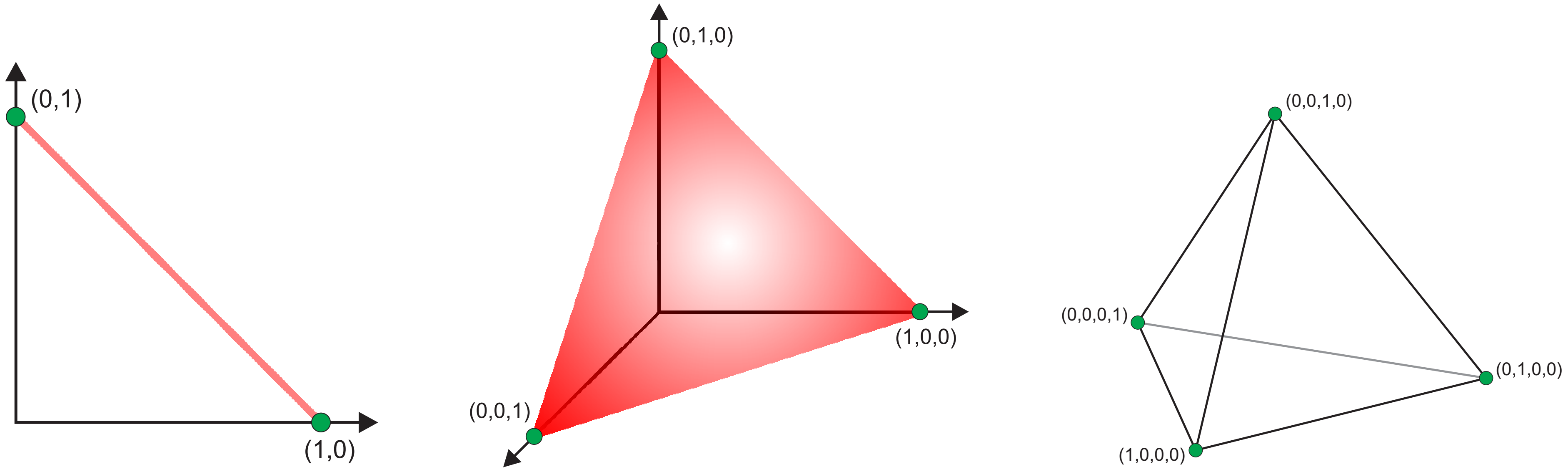
Es bezeichne $\Delta(\Omega)$ die Gesamtheit aller W-Funktionen auf der höchstens abzählbar unendlichen Menge Ω. Für $n \in \mathbb{N}$ gilt:

- $\Delta([1:n]) = \{(p_1, \ldots, p_n) \in [0,1]^n \mid p_1 + \ldots + p_n = 1\}$.
- $\Delta([1:n])$ ist die konvexe Hülle aller Einheitsvektoren $e_1, \ldots, e_n$ im $\mathbb{R}^n$.
- $\Delta([1:2])$ ist die Verbindungsstrecke zwischen e_1, e_2.
- $\Delta([1:3])$ ist das gleichseitige Dreieck mit den Ecken e_1, e_2, e_3.
- $\Delta([1:4])$ ist das reguläre Tetraeder mit den Ecken e_1, e_2, e_3, e_4.

Statt $\Delta([1:n])$ schreibt man oft auch kurz Δ_{n-1} und drückt damit aus, dass es sich dabei um das $(n-1)$-dimensionale **Einheitssimplex** handelt, das hier auch **Wahrscheinlichkeitssimplex** genannt wird.

Wahrscheinlichkeitssimplexe (2)

Das folgende Schaubild illustriert die Wahrscheinlichkeitssimplexe Δ_1, Δ_2 und Δ_3:

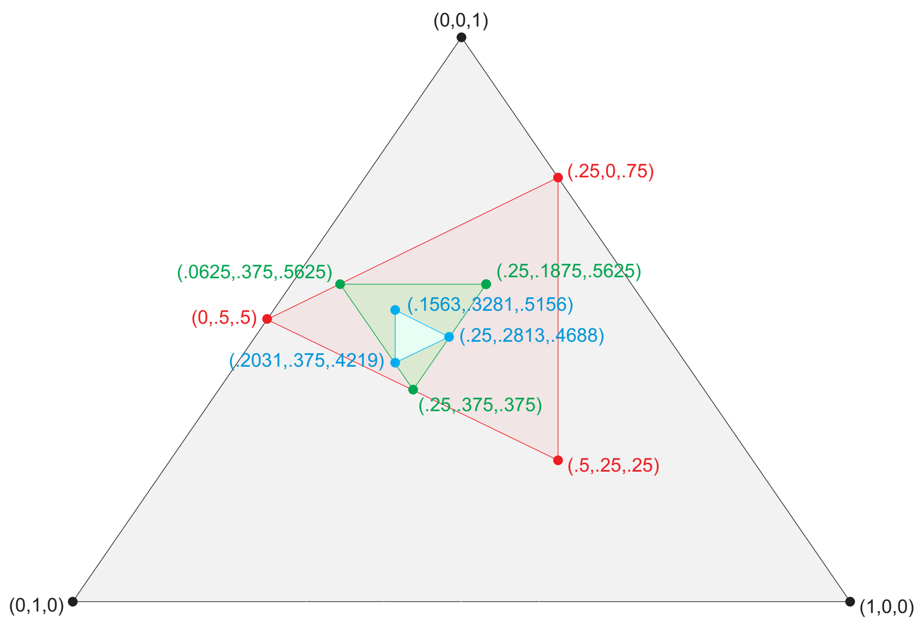


Beachte: man kann Δ_3 als Teilmenge des $\mathbb{R}^4$ komplett im $\mathbb{R}^3$ sehen!

Informatikbeispiele

In der Informatik kommen sehr oft Zähldichten vor. Hier sind zwei Beispiele:

- **Information Retrieval**: relative Termhäufigkeit in Textdokumenten
- **Google-Suche**: Web-Surfen als stochastischer Prozess. Jede Webseite steuert einen Zähldichte zur Beschreibung des Web-Surfens bei. Der sog. **Page-Rank** bewertet die Wichtigkeit einer Webseite und basiert auf einem Schrumpfungsprozess, der sich geometrisch-stochastisch sehr anschaulich beschreiben lässt, siehe nachfolgendes Schaubild. Näheres später!



Kontinuierliche Zähldichten: Dichtefunktionen

- Bisher haben wir nur im diskreten Fall W-Maße diskutiert.
- Dies führte auf Zähldichten: $p = (p_1, \ldots, p_n)$.
- Solche Zähldichten kann man sich als Säulengramme veranschaulichen: Für jedes $i \in \Omega := [1:n]$ zeichnet man über dem reellen Intervall $(i-1, i]$ eine Säule der Höhe p_i.
- Die Wahrscheinlichkeit, dass das Ereignis $A \subseteq \Omega$ eintritt, ist dann gleich dem Flächeninhalt aller Säulen zu $i \in A$.
- Säulendiagramme sind Treppenfunktionen.
- Im Folgenden geht es um kontinuierliche Varianten: statt Treppenfunktionen lassen wir allgemeinere Funktionen, sog. **Dichtefunktionen**, zu.
- Bei kontinuierlichen Funktionen ist das Integral das adäquate Werkzeug zur Messen von Flächeninhalten. Darauf gehen wir zunächst ein.

Intermezzo: Lebesgue-Integral

Für jede $(\mathcal{B}^n, \mathcal{B})$-messbare Funktion $f : \mathbb{R}^n \to [0, \infty]$, d.h.

$$\text{für alle } c > 0 \text{ gilt } \{x \in \mathbb{R}^n \mid f(x) \leq c\} \in \mathcal{B}^n,$$

ist das **Lebesgue-Integral** $\int f(x)dx \in [0, \infty]$ definiert.

Dieses Integral besitzt folgende Eigenschaften:

(ρ) Für jede Riemann-integrierbare Funktion f stimmt $\int f(x)dx$ mit dem Riemann-Integral von f überein.

(σ) Für jede Folge $f_1, f_2, \ldots$ von nichtnegativen messbaren Funktionen gilt

$$\int \sum_{n \geq 1} f_n(x)dx = \sum_{n \geq 1} \int f_n(x)dx.$$

- Das Riemann-Integral erfüllt (σ) nicht.

Intermezzo: Lebesgue-Maß

- Ist $A \in \mathcal{B}^n$ und bezeichnet 1_A die Indikatorfunktion zu A [d.h. $1_A(x) = 1$, falls $x \in A$ und $1_A(x) = 0$ sonst], so setzt man

$$\int_A f(x)dx := \int 1_A(x) f(x) dx.$$

- $\lambda^n(A) := \int 1_A(x)dx$ definiert das **Lebesgue-Maß auf** $(\mathbb{R}^n, \mathcal{B}^n)$.
- Insbesondere stimmt das Volumen eines Quaders Q in $\mathbb{R}^n$, also das Produkt der Kantenlängen von Q, mit $\lambda^n(Q)$ überein.
- Für $\Omega \in \mathcal{B}^n$ heißt die Einschränkung λ^n_Ω von λ^n auf $\mathcal{B}^n_\Omega$ das **Lebesgue-Maß auf** Ω.

Konstruktion von W-Maßen: stetiger Fall 1

Satz 1.6. Ist $\Omega \subseteq \mathbb{R}^n$ Borelsch, so bestimmt jede Funktion $\rho : \Omega \to [0, \infty)$ mit

(a) $\{x \in \Omega \mid \rho(x) \leq c\} \in \mathcal{B}^n_\Omega$ für alle $c > 0$
(b) $\int_\Omega \rho(x) dx = 1$

genau ein W-Maß P auf $(\Omega, \mathcal{B}^n_\Omega)$ vermöge

$$P(A) = \int_A \rho(x) dx \quad \text{für } A \in \mathcal{B}^n_\Omega.$$

Definition. Die Funktion ρ heißt dann **Dichtefunktion** von P oder auch **W-Dichte**.

Konstruktion von W-Maßen: stetiger Fall 2

Beweis des Satzes.

$P(\Omega) = 1$ folgt aus (b).

Ist $A_1, A_2 \ldots$ eine Familie paarweise disjunkter Borel-Mengen, so folgt mit $A := \sqcup_{i\geq 1} A_i$ und $1_A = \sum_{i\geq 1} 1_{A_i}$ mittels der Eigenschaft (σ) bei Lebesgue-Integralen:

$$\begin{aligned} P(\bigsqcup_{i\geq 1} A_i) &= \int_A \rho(x)dx = \int 1_A(x)\rho(x)dx \\ &= \int \sum_{i\geq 1} 1_{A_i}(x)\rho(x)dx = \sum_{i\geq 1} \int 1_{A_i}(x)\rho(x)dx \\ &= \sum_{i\geq 1} \int_{A_i} \rho(x)dx = \sum_{i\geq 1} P(A_i). \end{aligned}$$

$\square$

Gleichverteilungen

Endlicher Fall.
Ist $(\Omega, 2^\Omega)$ ein endlicher W-Raum, so heißt das W-Maß zur konstanten W-Funktion

$$\rho(\omega) := \frac{1}{|\Omega|}$$

die (diskrete) **Gleichverteilung** auf Ω.

Dieses Maß wird mit $\mathcal{U}_\Omega$ bezeichnet. [$\mathcal{U}$: **u**niform distribution.]

Für $A \subseteq \Omega$ ist dann $\mathcal{U}_\Omega(A) = |A|/|\Omega|$ (# günstigen Fälle / # möglichen Fälle).

Borel-Mengen. Ist $\Omega \subset \mathbb{R}^n$ eine Borel-Menge mit Volumen $0 < \lambda^n(\Omega) < \infty$, so heißt das W-Maß auf $(\Omega, \mathcal{B}^n_\Omega)$ mit der konstanten Dichtefunktion

$$\rho(x) := \frac{1}{\lambda^n(\Omega)}$$

die (stetige) **Gleichverteilung** auf Ω. Sie wird ebenfalls mit $\mathcal{U}_\Omega$ bezeichnet.

Für eine Borel-Untermenge A von Ω ist dann $\mathcal{U}_\Omega(A) = \lambda^n(A)/\lambda^n(\Omega)$ (günstiges Volumen / mögliches Volumen).

Stochastik

Kapitel 3: Zufallsvariablen

Prof. Dr. Reinhard Klein
Folien von Prof. Dr. Michael Clausen

Sommersemester 2017

INSTITUT FÜR INFORMATIK II, UNIVERSITÄT BONN

Motivierendes Beispiel

Beim n-maligen Münzwurf (mit fairer Münze) kann man das Zufallsgeschehen unterschiedlich genau protokollieren:

- **detailliert**: Münzwurffolge
  Ergebnisraum: $\Omega = \{0,1\}^n$ (*Zahl* entspricht 1),
  Ereignisalgebra: 2^Ω, W-Maß: Gleichverteilung.
- **gröber**: Anzahlprotokoll (Wie oft ist *Zahl* gefallen?)
  Ergebnisraum: $\Omega' = \{0,1,\ldots,n\}$,
  Ereignisalgebra: $2^{\Omega'}$, W-Maß: **?**.

Idee: Die Ereignisräume $(\Omega, 2^\Omega)$ und $(\Omega', 2^{\Omega'})$ hängen zusammen über die Zuordnung

$$X : \Omega \ni (\omega_1, \ldots, \omega_n) \mapsto \omega_1 + \ldots + \omega_n \in \Omega'$$

und die W-Funktion $p = (p_0, p_1, \ldots, p_n)$ zum Anzahlprotokoll leitet sich aus der Gleichverteilung auf $(\Omega, 2^\Omega)$ ab:

$$p_k = \frac{|\{\omega \in \Omega | X(\omega) = k\}|}{|\Omega|} = \binom{n}{k} \cdot 2^{-n}, \quad (k \in [0:n]).$$

Allgemeines Szenario

Gegeben:

- W-Raum $(\Omega, \mathcal{A}, P)$ [Detailsicht]
- Ereignisraum $(\Omega', \mathcal{A}')$ [Grobsicht: Modellausschnitt]
- Abbildung $X : \Omega \to \Omega'$ [Informationskompression]

Fragen:

- Wann ist X mit den Ereignisräumen $(\Omega, \mathcal{A})$ und $(\Omega', \mathcal{A}')$ verträglich?
  - $\to$ messbare Abbildungen, Zufallsvariablen
- Wie kann man auf $(\Omega', \mathcal{A}')$ in Abhängigkeit von X und $(\Omega, \mathcal{A}, P)$ sinnvoll ein W-Maß definieren?
  - $\to$ Bildmaße, Verteilungen

Urbilder

Gegeben (wie oben):

- W-Raum $(\Omega, \mathcal{A}, P)$, Ereignisraum $(\Omega', \mathcal{A}')$, Vergröberungsabbildung $X : \Omega \to \Omega'$, **Ereignis** $A' \in \mathcal{A}'$.

Frage:

- Wie wahrscheinlich ist es, dass das Ergebnis eines Zufallsexperiments in Ω nach Abbildung durch X in A' liegt?
- Ein $\omega \in \Omega$ landet nach Vergröberung durch X in A' gdw. $X(\omega) \in A'$ gilt.
- Daher bietet es sich an, die Menge

$$X^{-1}[A'] := \{\omega \in \Omega \mid X(\omega) \in A'\}$$

  aller X-**Urbilder** heranzuziehen und die Wahrscheinlichkeit, dass man in A' landet mit $P(X^{-1}[A'])$ zu definieren.
- Einziger **Haken**: P darf nur an Ereignissen aus $\mathcal{A}$ ausgewertet werden!
- Das führt zur **Forderung**: $X^{-1}[A']$ muss zu $\mathcal{A}$ gehören, für jedes $A' \in \mathcal{A}'$.

Messbare Abbildungen

Definition. Es seien $(\Omega, \mathcal{A})$ und $(\Omega', \mathcal{A}')$ Messräume.
$X : \Omega \to \Omega'$ heißt $(\mathcal{A}, \mathcal{A}')$-**messbar** oder eine **Zufallsvariable**, kurz: **ZV**, wenn das X-Urbild von jedem $A' \in \mathcal{A}'$ zu $\mathcal{A}$ gehört:

$$\forall A' \in \mathcal{A}' : \quad X^{-1}[A'] := \{\omega \in \Omega \mid X(\omega) \in A'\} \in \mathcal{A}.$$

Statt $X^{-1}[A']$ schreibt man oft suggestiver: $\{X \in A'\}$.

Bemerkungen:

- Im Fall $\mathcal{A} = 2^\Omega$ ist jede Abbildung eine Zufallsvariable.
- **Sparversion der Messbarkeitsbedingung**:
  Wird die σ-Algebra $\mathcal{A}'$ von $\mathcal{G}'$ erzeugt, $\sigma(\mathcal{G}') = \mathcal{A}'$, so ist $X : \Omega \to \Omega'$ bereits dann eine ZV, wenn gilt (Übung):

$$\forall A' \in \mathcal{G}' : \quad X^{-1}[A'] \in \mathcal{A}.$$

- Die $(\mathcal{A}, \mathcal{A}')$-Messbarkeit von $X : \Omega \to \Omega'$ schreibt man oft kurz als $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$.

Beispiele von Zufallsvariablen

- **Diskreter Fall**: Hier ist jede Abbildung eine Zufallsvariable.
- **Reellwertiger Fall**: Ist $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B})$, so ist $X : \Omega \to \mathbb{R}$ bereits ZV, wenn alle Mengen $\{X \leq c\}$ zu $\mathcal{A}$ gehören. (Begründung: Sparversion!)
- **Stetiger Fall**: Stetige Funktionen $X : \Omega \to \mathbb{R}$, $\Omega \subseteq \mathbb{R}^n$, sind $(\mathcal{B}^n_\Omega, \mathcal{B})$-messbar, also ZVs, denn für $c \in \mathbb{R}$ ist $\{X \leq c\}$ als Urbild der abgeschlossenen Menge $(-\infty, c]$ selbst abgeschlossen in Ω, gehört also zu $\mathcal{B}^n_\Omega$.
  - Hier haben wir folgende Charakterisierung der Stetigkeit benutzt:
    Eine Funktion ist genau dann stetig, wenn das Urbild jeder abgeschlossenen Menge wieder abgeschlossen ist.

Damit haben wir die erste Frage beantwortet.

Wir kommen nun zum zweiten Problem.

Bildmaße

Satz. Es sei $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ ZV und P W-Maß auf $(\Omega, \mathcal{A})$. Dann wird durch

$$P'(A') := P(X^{-1}[A']) = P(\{X \in A'\}) =: P(X \in A') \quad \text{für } A' \in \mathcal{A}'$$

ein W-Maß P' auf $(\Omega', \mathcal{A}')$ definiert, das **Bildmaß** zu P bzgl. X.

Beweis.

- $P'(A') := P(X^{-1}[A'])$ macht Sinn, denn für $A' \in \mathcal{A}'$ liegt $X^{-1}[A']$ in $\mathcal{A}$, weil X ZV ist.
- P' erfüllt (N): $P'(\Omega') = P(X \in \Omega') = P(\Omega) = 1$.
- P' erfüllt (A): Sind $A'_1, A'_2, \ldots \in \mathcal{A}'$ paarweise disjunkt, so sind auch die Urbilder $X^{-1}[A'_1], X^{-1}[A'_2], \ldots$ paarweise disjunkt. Folglich ist

$$\begin{aligned} P'(\sqcup_{i\geq 1} A'_i) &= P(X^{-1}[\sqcup_{i\geq 1} A'_i]) = P(\sqcup_{i\geq 1} X^{-1}[A'_i]) \\ &= \sum_{i\geq 1} P(X^{-1}[A'_i]) = \sum_{i\geq 1} P'(A'_i). \end{aligned}$$

Also ist P' W-Maß. □

Verteilungen & identisch verteilte Zufallsvariablen

Definition. Das Bildmaß zu P bzgl. X wird auch die **Verteilung** von X genannt und manchmal mit P_X oder mit $P \circ X^{-1}$ bezeichnet.

Es sei $((\Omega_i, \mathcal{A}_i, P_i))_{i\in I}$ eine Familie von W-Räumen sowie $(X_i)_{i\in I}$ eine Familie von ZVs, die alle in denselben Messraum abbilden

$$X_i : (\Omega_i, \mathcal{A}_i) \to (\Omega, \mathcal{A}).$$

Die Familie $(X_i)_{i\in I}$ heißt **identisch verteilt**, wenn alle Verteilungen $P_i \circ X_i^{-1}$ übereinstimmen.

Beispiel. n-maliger Münzwurf mit fairer Münze.

- Hier ist $I = [1 : n]$, $(\Omega_i = \{0,1\}^n, \mathcal{A}_i = 2^{\Omega_i}, P_i \equiv$ Gleichverteilung) für alle i. $(\Omega, \mathcal{A}) = (\{0,1\}, 2^{\{0,1\}})$.
- $X_i : \{0,1\}^n \to \{0,1\}$ sei die Projektion $(\omega_1, \ldots, \omega_n) \mapsto \omega_i$.
- Dann sind $X_1, \ldots, X_n$ identisch verteilte ZVs, denn die Verteilung von jedem X_i ist die Gleichverteilung auf $\Omega = \{0,1\}$.

Gemeinsame Verteilungen

- Hat man Ereignisse, die von mehreren ZVs abhängen, so reicht nicht die Kenntnis der einzelnen Verteilungen aus. (Ein Beispiel dazu findet sich auf der letzten Folie dieses Kapitels.) Man braucht vielmehr folgendes Konzept.
- Es sei $(\Omega, \mathcal{A}, P)$ ein W-Raum und $X_1, \ldots, X_n$ seien ZVs $X_i : (\Omega, \mathcal{A}) \to (\Omega_i, \mathcal{A}_i)$.
- Die **Produktabbildung** $X := X_1 \otimes \ldots \otimes X_n : \Omega \to \Omega_1 \times \ldots \times \Omega_n$, definiert durch

$$X(\omega) := (X_1(\omega), \ldots, X_n(\omega))$$

  ist eine Zufallsvariable $X \colon (\Omega, \mathcal{A}) \to (\Omega_1 \times \ldots \times \Omega_n, \mathcal{A}_1 \otimes \ldots \otimes \mathcal{A}_n)$, deren Verteilung P_X die **gemeinsame Verteilung** der $X_1, \ldots, X_n$ genannt wird.
- Hier ist

$$\mathcal{A}_1 \otimes \ldots \otimes \mathcal{A}_n := \sigma\Big(\bigcup_{j=1}^{n} \pi_j^{-1}[\mathcal{A}_j]\Big) \subseteq 2^{\Omega_1 \times \ldots \times \Omega_n}$$

  die **Produkt-$\boldsymbol{\sigma}$-Algebra**, bei der alle Projektionen $\pi_j : (\omega_1, \ldots, \omega_n) \mapsto \omega_j$ $(\mathcal{A}, \mathcal{A}_j)$-messbar sind.

Messbarkeit der Produktabbildung

Satz. Die obige Produktabbildung $X := X_1 \otimes \ldots \otimes X_n : \Omega \to \Omega_1 \times \ldots \times \Omega_n$ ist $(\mathcal{A}, \mathcal{A}_1 \otimes \ldots \otimes \mathcal{A}_n)$-messbar.

Beweis.

- $\mathcal{A}_1 \otimes \ldots \otimes \mathcal{A}_n$ wird erzeugt von allen $A_1 \times \ldots \times A_n$, wobei $A_i \in \mathcal{A}_i$, für alle $i \in [1 : n]$.
- $X^{-1}[A_1 \times \ldots \times A_n] = X_1^{-1}[A_1] \cap \ldots \cap X_n^{-1}[A_n]$.
- Wegen der $(\mathcal{A}, \mathcal{A}_i)$-Messbarkeit von $X_i = \pi_i \circ X$ liegt $X_i^{-1}[A_i]$ in $\mathcal{A}$.
- Da $\mathcal{A}$ unter endlicher Durchschnittsbildung abgeschlossen ist, liegt auch $X^{-1}[A_1 \times \ldots \times A_n]$ in $\mathcal{A}$.
- Nach der Sparversion der Messbarkeitsbedingung (siehe Folie 5) ist X also $(\mathcal{A}, \mathcal{A}_1 \otimes \ldots \otimes \mathcal{A}_n)$-messbar. $\square$

Gemeinsame Verteilung: Beispiel (1)

n-**maliger Münzwurf mit fairer Münze**. Betrachte zwei ZVs:

- **Anzahl der Erfolge** (Zahl $\equiv 1$): $X_1 : \{0,1\}^n \to [0:n]$,
$$X_1 : (\omega_1, \ldots, \omega_n) \mapsto \omega_1 + \ldots + \omega_n.$$
- **Wartezeit bis zum ersten Erfolg**: $X_2 : \{0,1\}^n \to [1 : n+1]$,
$$X_2 : (\omega_1, \ldots, \omega_n) \mapsto \min\{j \mid \omega_j = 1\} \quad \text{bzw.} \quad (0, \ldots, 0) \mapsto n+1.$$
- Die ZV $X := X_1 \otimes X_2$ ordnet $0 \neq \omega \in \{0,1\}^n$ das Paar
$$(\omega_1 + \ldots + \omega_n, \min\{j \mid \omega_j = 1\}) \in [1:n] \times [1:n]$$
zu; weiter ist $X((0, \ldots, 0)) = (0, n+1)$.
- Damit ist
$$X = X_1 \otimes X_2 : \{0,1\}^n \to [0:n] \times [1:n+1].$$

Gemeinsame Verteilung: Beispiel (2)

Satz. Für die zu P_X gehörige Zähldichte p_X auf $[0:n] \times [1:n+1]$ gilt für $(k,h) \in [1:n]^2$

$$p_X(k,h) = \binom{n-h}{k-1} \cdot 2^{-n}.$$

Weiter ist $p_X(0,h) = 0$, $p_X(k,n+1) = 0$. Schließlich ist $p_X(0,n+1) = 2^{-n}$.

Beweis. **Fall 1**: $(k,h) \in [1:n]^2$

- Das X-Urbild von (k,h) besteht dann aus denjenigen n-Tupeln, die mit $h-1$ Nullen beginnen, dann folgt eine Eins,
- von den restlichen $n-h$ Stellen sind $k-1$ mit Einsen, die anderen mit Nullen belegt. Das beweist $p_X(k,h) = \binom{n-h}{k-1} \cdot 2^{-n}$.

Fall 2: Für $h, k \in [1:n]$ ist $p_X(0,h) = 0$, $p_X(k,n+1) = 0$: klar!

Fall 3: $k = 0$ und $h = n+1$

- Hier ist $\omega = (0,\ldots,0)$ einziges ω mit $X(\omega) = (0,n+1)$. Also ist $p_X(0,n+1) = 2^{-n}$. $\square$

Randverteilungen

- Es sei $(\Omega, \mathcal{A}, P)$ W-Raum und $X_1, \dots, X_n$ seien ZVs, $X_i : (\Omega, \mathcal{A}) \to (\Omega_i, \mathcal{A}_i)$.
- $X := X_1 \otimes \ldots \otimes X_n$ bezeichne das Produkt der ZVs $X_1, \dots, X_n$.
- Sind $1 \leq i_1 < \ldots < i_k \leq n$, so heißt die gemeinsame Verteilung des Teilprodukts

$$X_{i_1} \otimes \ldots \otimes X_{i_k} : \Omega \to \Omega_{i_1} \times \ldots \times \Omega_{i_k}$$

eine k-**dimensionale Rand- oder Marginalverteilung** von $X_1, \dots, X_n$.

Satz. Im diskreten Fall ergeben sich die Marginalverteilungen, indem man durch komplettes Aufsummieren die irrelevanten Komponenten eliminiert. So ergibt sich etwa im Fall $(i_1, \dots, i_k) = (1, \dots, k)$ für die zugehörigen Zähldichte $p_{X_1 \otimes \ldots \otimes X_k}$:

$$p_{X_1 \otimes \ldots \otimes X_k}(\omega_1, \dots, \omega_k) = \sum p_{X_1 \otimes \ldots \otimes X_n}(\omega_1, \dots, \omega_k, \omega_{k+1}, \dots, \omega_n),$$

dabei wird – bei festem $(\omega_1, \dots, \omega_k)$ – über alle $(\omega_{k+1}, \dots, \omega_n) \in \Omega_{k+1} \times \ldots \times \Omega_n$ summiert. Diesen Vorgang nennt man **Marginalisierung**.

Marginalisierung

Beweis des letzten Satzes:

$$\begin{aligned} p_{X_1\otimes\ldots\otimes X_k}(\omega_1,\ldots,\omega_k) &= P(\{\omega\in\Omega \mid \forall i\in[1:k]\colon X_i(\omega)=\omega_i\}) \quad \text{(nach Definition)} \\ &= P(\{\omega\in\Omega \mid \forall i\in[1:k]\colon X_i(\omega)=\omega_i \wedge \forall i>k : X_i(\omega)\in\Omega_i\}) \\ &= P(\bigsqcup_{\omega_{k+1},\ldots,\omega_n} \{\omega\in\Omega \mid \forall i\in[1:n]\colon X_i(\omega)=\omega_i\}) \\ &= \sum_{\omega_{k+1},\ldots,\omega_n} P(\{\omega\in\Omega \mid \forall i\in[1:n]\colon X_i(\omega)=\omega_i\}) \\ &= \sum_{\omega_{k+1},\ldots,\omega_n} p_{X_1\otimes\ldots\otimes X_n}(\omega_1,\ldots,\omega_k,\omega_{k+1},\ldots,\omega_n), \end{aligned}$$

wobei die disjunkte Vereinigung und die beiden Summen jeweils über alle $(\omega_{k+1},\ldots,\omega_n)\in\Omega_{k+1}\times\ldots\times\Omega_n$ laufen.

Damit ist der Satz bewiesen. $\square$

Beispiel: Gemeinsame Verteilung und Randverteilungen

- Wir betrachten wieder den n-maligen Münzwurf und die beiden ZVs X_1 = Anzahl der Erfolge und X_2 = Wartezeit bis zum ersten Erfolg.
- Für die Zähldichte der gemeinsamen Verteilung von X_1 und X_2, also die Verteilung des Produkts $X := X_1 \otimes X_2 : [0 : n] \times [1 : n+1]$, hatten wir bereits die folgenden Formeln bewiesen ($k, h \in [1 : n]$):

$$p_X(k,h) = \binom{n-h}{k-1} \cdot 2^{-n}, \quad p_X(0,n+1) = 2^{-n}, \quad p_X(k,n+1) = 0.$$

- Im Fall $n = 3$ ergibt sich folgende Zähldichte für die gemeinsame Verteilung (alle farbigen Zahlen sind noch mit $1/8$ zu multiplizieren):

$$p_X = \begin{array}{c|cccc|c} & 1 & 2 & 3 & 4 & \\ \hline 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 3 \\ 2 & 2 & 1 & 0 & 0 & 3 \\ 3 & 1 & 0 & 0 & 0 & 1 \\ \hline & 4 & 2 & 1 & 1 & \end{array}$$

- Die beiden Zähldichten zu den 1-dimensionalen Marginalverteilungen P_{X_1} und P_{X_2} ergeben sich als die Zeilen- bzw. Spaltensummen.

Beispiel: Marginalverteilungen ergeben nicht gemeinsame Verteilung

- Wir zeigen anhand des letzten Beispiels, dass man i.a. nicht von den Marginalverteilungen auf die gemeinsame Verteilung schließen kann.

$$p_X = \begin{array}{c|cccc|c} & 1 & 2 & 3 & 4 & \\ \hline 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 3 \\ 2 & 2 & 1 & 0 & 0 & 3 \\ 3 & 1 & 0 & 0 & 0 & 1 \\ \hline & 4 & 2 & 1 & 1 & \end{array}$$

- Die folgende Verteilung P_Y hat dieselben Marginalverteilungen wie X:

$$p_Y = \begin{array}{c|cccc|c} & 1 & 2 & 3 & 4 & \\ \hline 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 2 & 1 & 0 & 0 & 3 \\ 2 & 1 & 1 & 1 & 0 & 3 \\ 3 & 0 & 0 & 0 & 1 & 1 \\ \hline & 4 & 2 & 1 & 1 & \end{array}$$

Memo: alle farbigen Zahlen sind noch mit $1/8$ zu multiplizieren!

Beispiel zum Nacharbeiten(1)

Eine faire Münze wird n-mal geworfen. "'Zahl"' werde als Erfolg (codiert durch 1) und "'Kopf"' als Misserfolg (codiert durch 0) angesehen. Wir betrachten zwei Zufallsvariablen:

- X_1: Anzahl der Erfolge.
- X_2: Wartezeit bis zum ersten Erfolg.

Aufgaben:

1. Gebe den zu X_1 und X_2 gehörigen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ und die Messräume $(\Omega_1, \mathcal{A}_1)$ und $(\Omega_2, \mathcal{A}_2)$ an.
2. Beschreibe X_1 und X_2 konkret als Abbildungen

$$X_1 : (\Omega, \mathcal{A}) \to (\Omega_1, \mathcal{A}_1), \quad X_2 : (\Omega, \mathcal{A}) \to (\Omega_2, \mathcal{A}_2).$$

3. Bestimme die Verteilungen von X_1 und X_2.

Beispiel zum Nacharbeiten(2)

1. Beiden Zufallsvariablen liegt derselbe W-Raum $(\Omega, \mathcal{A}, P)$ als Definitionsbereich zugrunde, mit $\Omega = \{0,1\}^n$, $\mathcal{A} = 2^\Omega$, $P =$ Gleichverteilung. Außerdem für

$$X_1 : (\Omega, \mathcal{A}) \to (\Omega_1, \mathcal{A}_1), \quad X_2 : (\Omega, \mathcal{A}) \to (\Omega_2, \mathcal{A}_2).$$

   gilt: $\Omega_1 = [0:n] = \{0, 1, ..., n\}$, $\mathcal{A}_1 = 2^{\Omega_1}$ und $\Omega_2 = [1 : n+1] = \{0, 1, ..., n\}$, $\mathcal{A}_2 = 2^{\Omega_2}$.
2. $X_1 : \Omega \ni (\omega_1, ..., \omega_n) \mapsto \sum_{i=1}^n \omega_i \in \Omega_1$.
   $X_2 : \Omega \ni (\omega_1, ..., \omega_n) \mapsto \min\{j \in [2:n] \mid \omega_j = 1\}$ und
   $\Omega \ni (0, ..., 0) \mapsto n+1$
3. Für die Verteilungen $P_1 := P_{X_1}$ und $P_2 := P_{X_2}$ ergibt sich:

$$\begin{aligned} P_1(k) &= P\left(\left\{\omega \in \Omega \mid \sum_{i=1}^n \omega_i = k\right\}\right) \\ &= \binom{n}{k}\left(\frac{1}{2}\right)^k \cdot \left(\frac{1}{2}\right)^{n-k} = \binom{n}{k} \cdot 2^{-n}, \quad k \in [0:n] \end{aligned}$$

Beispiel zum Nacharbeiten(3)

$$
\begin{aligned}
P_2(n+1) &= P\left(\left\{\omega \in \Omega \mid \sum_{i=1}^{n} \omega_i = 0\right\}\right) \\
&= P_1(0) = 2^{-n}.
\end{aligned}
$$

Für $j \in [1:n]$ und $\omega \in \Omega$ mit $X_2(\omega) = j$ ist

$$
\omega = (\underbrace{\omega_1, \ldots, \omega_{j-1}}_{j-1 \text{ Nullen}}, \underbrace{\omega_j = 1}_{\text{erste "`1"'}}, \underbrace{\omega_{j+1}, \ldots, \omega_n}_{\text{beliebiger Binärvektor}}),
$$

d.h.

$$
\begin{aligned}
P_2(j) &= P\left(\left\{\omega \in \Omega \mid \sum_{i=1}^{j-1} \omega_i = 0, w_j = 1\right\}\right) \\
&= \left(\frac{1}{2}\right)^{j-1} \cdot \left(\frac{1}{2}\right)^{1} \cdot 1 = \left(\frac{1}{2}\right)^{j}
\end{aligned}
$$

Stochastik

Kapitel 4: Stochastische Standardmodelle

Prof. Dr. Reinhard Klein
Folien von Prof. Dr. Michael Clausen

Sommersemester 2017

INSTITUT FÜR INFORMATIK II, UNIVERSITÄT BONN

Ziele

- Beschreibung klassischer stochastischer Modelle
- Herleitung zugehöriger W-Maße und Verteilungen
- Erläuterung von Zusammenhängen zwischen den Modellen
- Aufzeigen typischer Anwendungen

Laplace-Räume

Dies sind endliche W-Räume mit Gleichverteilung. Genauer:

- Ω sei eine endliche, nichtleere Menge.
- $(\Omega, 2^\Omega, \mathcal{U}_\Omega)$ wird zum W-Raum mit Gleichverteilung, auch **Laplace-Raum** genannt, indem man für $A \subseteq \Omega$ festlegt:

$$\mathcal{U}_\Omega(A) := \frac{|A|}{|\Omega|} = \frac{\text{Anzahl der günstigen Fälle}}{\text{Anzahl der möglichen Fälle}}.$$

- Dieses Modell wird gewählt, wenn aus Symmetriegründen (fairer Würfel, faire Münze, usw.) alle Ergebnisse gleichwahrscheinlich sind.
- Zugehörige W-Funktion ist die auf Ω konstante Funktion

$$\omega \mapsto \frac{1}{|\Omega|}.$$

Beispielanwendungen

- n-maliger Wurf einer fairen Münze: $\Omega = \{0, 1\}^n$, $|\Omega| = 2^n$
- n-maliger Wurf eines fairen Würfels: $\Omega = [1 : 6]^n$, $|\Omega| = 6^n$
- Zahlenlotto: $\Omega = \{Z \mid Z \subset [1 : 49] \wedge |Z| = 6\}$, $|\Omega| = \binom{49}{6} = \frac{49!}{43! \cdot 6!}$
- Reihenfolge eines gut gemischten Skatblatts: $\Omega = \{\pi \mid \pi \text{ ist Permutation von } [1 : 32]\}$, $|\Omega| = 32!$

Gleichverteilung: kontinuierlicher Fall

- $\Omega \subset \mathbb{R}^n$ sei Borel-Menge mit endlichem Volumen: $0 < \lambda^n(\Omega) < \infty$.
- $(\Omega, \mathcal{B}^n_\Omega, \mathcal{U}_\Omega)$ wird zum W-Raum mit kontinuierlicher Gleichverteilung, wenn man für jede Borel-Menge $A \subseteq \Omega$ festlegt:

$$\mathcal{U}_\Omega(A) := \frac{\lambda^n(A)}{\lambda^n(\Omega)} = \frac{\text{günstiges Volumen}}{\text{mögliches Volumen}}.$$

- Dieses Modell wird gewählt, wenn aus Symmetriegründen die Wahrscheinlichkeit nur von der Größe des Volumens abhängt.

Beispielanwendung

- Zufällige Wahl einer Richtung durch Drehung eines Roulette-Rades.
- In welche Richtung zeigt die Null?
- **Ergebnismenge**: $\Omega = [0, 2\pi)$ (Drehwinkel)
- Aus Symmetriegründen wählt man als W-Raum: $(\Omega, \mathcal{B}^1_\Omega, \mathcal{U}_\Omega)$
- Wegen $\lambda^1(\Omega) = 2\pi$ ist

$$\mathcal{U}_\Omega(A) = \frac{\lambda^1(A)}{2\pi}$$

für jede Borel-Teilmenge A von $[0, 2\pi)$.

Urnenmodelle: allgemein

- einfache stochastische Modellklasse mit endlichen Ergebnisräumen
- Zufallsereignis: Mehrmaliges Ziehen von gleichartigen, aber (teilweise) verschiedenfarbigen Kugeln aus einer Urne
- Vier Varianten:
  - Ziehen mit (Z) oder ohne (z) **Z**urücklegen sowie
  - Ziehen mit (R) oder ohne (r) Berücksichtigung der **R**eihenfolge
- Das liefert folgende Typen von Ergebnisräumen:

$$\Omega_{ZR}, \quad \Omega_{Zr}, \quad \Omega_{zR}, \quad \Omega_{zr}.$$

Urnenmodelle: Modellparameter

- N: Anzahl der Kugeln in der Urne
- F: Menge der verschiedenen Farben der Kugeln in der Urne
- N_f: Anzahl der f-farbigen Kugeln in der Urne ($f \in F$)
- Beachte: $N = \sum_{f \in F} N_f$.
- n: Anzahl der Ziehungen

Beispiel. Der Urneninhalt ●●●●●●● (4 rote, 1 blaue, 2 schwarze Kugeln) wird durch folgende Parameter beschrieben:

- $N = 7$: Anzahl der Kugeln
- $F = \{\bullet, \bullet, \bullet\} \equiv \{\text{rot, blau, schwarz}\}$
- $N_\bullet \equiv N_{\text{rot}} = 4, \quad N_\bullet \equiv N_{\text{blau}} = 1, \quad N_\bullet \equiv N_{\text{schwarz}} = 2;$
- $N = 7 = 4 + 1 + 2.$

Trick: fiktiver Laplace-Raum (Z)

Um die verschiedenen Urnenmodelle einheitlich behandeln zu können, gehen wir bei n Ziehungen mit **Z**urücklegen von einem fiktiven Laplace-Raum aus, indem wir uns die N Kugeln zusätzlich noch durchnummeriert denken.

- Ausgangspunkt bei den **Modellen mit Zurücklegen** ist die Menge $\Omega := [1 : N]^n$ mit Gleichverteilung: $(\Omega, 2^\Omega, \mathcal{U}_\Omega)$
- F: Menge der verschiedenen Farben der Kugeln in der Urne
- $\phi : [1 : N] \to F$ sei surjektive Färbungsfunktion.
- ϕ zerlegt $[1 : N]$ disjunkt in gleichfarbige Nummernbereiche ($f \in F$):

$$F_f := \{i \in [1 : N] \mid i\text{-te Kugel ist } f\text{-farbig}\} = \phi^{-1}[\{f\}].$$

Vorteil der Fiktion: W-Raum bekannt, **Nachteil**: Realität $\neq$ Fiktion

Idee: Verbinde Fiktion mit Realität durch geeignete **Zufallsvariablen !**

Memo: Bildmaße im diskreten Fall

- Es sei $X : \Omega \to \Omega'$ ZV zwischen zwei diskreten Messräumen.
- Ist P W-Maß auf $(\Omega, 2^{\Omega})$, p die zugehörige Zähldichte, $p(\omega) = P(\{\omega\})$, so ist die Zähldichte p_X des Bildmaßes P_X auf $(\Omega', 2^{\Omega'})$ eindeutig beschrieben durch

$$p_X(\omega') = P(X = \omega') = \sum_{\omega : X(\omega) = \omega'} p(\omega), \qquad \text{für } \omega' \in \Omega'.$$

- Für $A' \subseteq \Omega'$ ist dann

$$P_X(A') = \sum_{\omega' \in A'} p_X(\omega').$$

- Von diesen Formeln werden wir wiederholt Gebrauch machen.
- Desweiteren werden wir ab jetzt - etwas ungenau - meistens die Zähldichte mit demselben Buchstaben bezeichnen wie das zugehörige W-Maß:

$$P_X(A') = \sum_{\omega' \in A'} P_X(\omega').$$

Stichproben mit Zurücklegen

Ausgehend vom fiktiven Laplace-Raum $\Omega = [1:N]^n$ versehen mit fester surjektiver Färbungsfunktion $\phi : [1:N] \to F$ werden wir nun schrittweise mittels geeigneter ZVs die Information komprimieren:

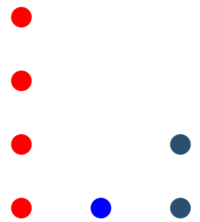
$$[1:N]^n \xrightarrow{X_{ZR}} \Omega_{ZR} \xrightarrow{X_{Zr}} \Omega_{Zr}.$$

Konkret wird einer Folge farbiger Kugelnummern mittels X_{ZR} die zugehörige **Farbenfolge** zugeordnet:

$$(\textcolor{red}{2}, \textcolor{blue}{4}, \textcolor{red}{5}, \textcolor{red}{2}, 1, \textcolor{red}{2}, 3) \xrightarrow{X_{ZR}} (\textcolor{red}{\bullet}, \textcolor{blue}{\bullet}, \textcolor{red}{\bullet}, \textcolor{red}{\bullet}, \bullet, \textcolor{red}{\bullet}, \bullet)$$

die ihrerseits mittels X_{Zr} zum zugehörigen **Farbenhistogramm** vergröbert wird:

$$(\textcolor{red}{\bullet}, \textcolor{blue}{\bullet}, \textcolor{red}{\bullet}, \textcolor{red}{\bullet}, \bullet, \textcolor{red}{\bullet}, \bullet) \xrightarrow{X_{Zr}}$$



Geordnete Stichproben mit Zurücklegen

- $\Omega_{ZR} = F^n$ ist der **Raum der Farbenfolgen**, der mittels

$$X_{ZR} : [1:N]^n \ni (\omega_1, \ldots, \omega_n) \mapsto (\phi(\omega_1), \ldots, \phi(\omega_n)) \in F^n$$

  mit dem Bildmaß, hier kurz mit P_{ZR} bezeichnet, zu versehen ist.
- Dabei ist zu beachten, dass auf $\Omega = [1:N]^n$ die Gleichverteilung $\mathcal{U}_\Omega$ vorgegeben ist.
- Für die Farbenfolge $\mathbf{f} = (f_1, \ldots, f_n) \in F^n$ ist

$$\{X_{ZR} = \mathbf{f}\} = \{\omega \in [1:N]^n \mid X_{ZR}(\omega) = \mathbf{f}\} = F_{f_1} \times \ldots \times F_{f_n}.$$

  Daher erhalten wir für das Bildmaß P_{ZR}:

$$P_{ZR}(\mathbf{f}) = \mathcal{U}_\Omega(X_{ZR} = \mathbf{f}) = \frac{|F_{f_1} \times \ldots \times F_{f_n}|}{|[1:N]^n|} = \prod_{i=1}^{n} \frac{N_{f_i}}{N}.$$

- Dies ist gerade das n-**fache Produktmaß** zur W-Funktion $f \mapsto N_f/N$ auf F. (Siehe übernächste Folie.)
- Zunächst folgt ein konkretes Beispiel.

universitätbonn

Geordnete Stichproben mit Zurücklegen: Beispiel

- Es sei $N = 7$ und die nummerierten Kugeln in $[1:7]$ seien mittels ϕ wie folgt gefärbt (r: rot, b: blau, s: schwarz):

$$\phi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ r & r & r & r & b & s & s \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ {\color{red}\bullet} & {\color{red}\bullet} & {\color{red}\bullet} & {\color{red}\bullet} & {\color{blue}\bullet} & \bullet & \bullet \end{pmatrix}.$$

- Also ist $F = \{{\color{red}\bullet}, {\color{blue}\bullet}, \bullet\} \equiv \{r, b, s\}$ und ϕ zerlegt $[1:7]$ in folgende gleichfarbige Nummernbereiche: $F_r = \{1, 2, 3, 4\}, \quad F_b = \{5\}, \quad F_s = \{6, 7\}$.
- Im Fall $n = 5$ (fünfmaliges Ziehen) ergibt sich als Raum der Farbenfolgen $\Omega_{ZR} = F^5$.
- Für die Farbenfolge $\mathbf{f} := (s, r, r, b, s) \in F^5$ ist

$$\{X_{ZR} = \mathbf{f}\} = F_s \times F_r \times F_r \times F_b \times F_s.$$

Daher erhalten wir für das Bildmaß P_{ZR} an der Stelle $\mathbf{f}$:

$$P_{ZR}(\mathbf{f}) = \frac{2}{7} \cdot \frac{4}{7} \cdot \frac{4}{7} \cdot \frac{1}{7} \cdot \frac{2}{7}.$$

Produktmaße

- **Satz/Definition**. Es sei ρ eine Zähldichte auf dem endlichen Ergebnisraum Ω. Dann wird durch

$$\rho^{\otimes n}(\omega) = \rho^{\otimes n}(\omega_1, \ldots, \omega_n) := \prod_{i=1}^{n} \rho(\omega_i)$$

eine Zähldichte auf Ω^n definiert, diese heißt n-**fache Produktdichte** von ρ und das zugehörige W-Maß auf der Potenzmenge von Ω^n heißt das n-**fache Produktmaß** zu ρ.

- **Beweis**. $\rho^{\otimes n}$ ist Zähldichte, denn $\rho^{\otimes n}(\omega) \geq 0$ für alle $\omega \in \Omega^n$. Ferner ist

$$\begin{aligned} \sum_{\omega \in \Omega^n} \rho^{\otimes n}(\omega) &= \sum_{\omega_1 \in \Omega} \cdots \sum_{\omega_n \in \Omega} \rho(\omega_1) \cdot \ldots \cdot \rho(\omega_n) \\ &= \underbrace{\Big(\sum_{\omega_1 \in \Omega} \rho(\omega_1) \Big)}_{=1} \cdot \ldots \cdot \underbrace{\Big(\sum_{\omega_n \in \Omega} \rho(\omega_n) \Big)}_{=1} = 1. \end{aligned}$$

Damit ist der Satz bewiesen. □

Bernoulli-Verteilung

Hierbei handelt es sich um folgende gewichtete Version geordneter Stichproben mit Zurücklegen, das auf einen Spezialfall von Produktmaßen hinausläuft:

- Es gibt nur zwei Farben: $F = \{0, 1\}$, wobei 1 als Erfolg und 0 als Misserfolg interpretiert wird.
- Die Erfolgswahrscheinlichkeit wird mit p, die Misserfolgswahrscheinlichkeit mit $q = 1 - p$ bezeichnet.
- Die Wahrscheinlichkeit für die Stichprobenfolge $\omega = (\omega_1, \ldots, \omega_n)$ ist

$$p^k \cdot q^{n-k},$$

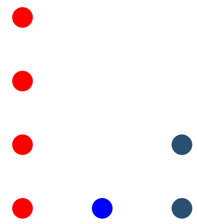
wenn in ω genau k-mal eine Eins vorkommt.

Diese Verteilung auf $\{0, 1\}^n$ heißt **Bernoulli-Verteilung** für n Alternativ-Versuche mit Erfolgswahrscheinlichkeit p.

Ungeordnete Stichproben mit Zurücklegen

Ein Farbenhistogramm H beschreibt, welche Farbe mit welcher Häufigkeit vorkommt:

$$H \quad = \quad \begin{pmatrix} \color{red}{\bullet} & \color{blue}{\bullet} & \bullet \\ 4 & 1 & 2 \end{pmatrix} \quad \equiv$$



Der **Raum der Farbenhistogramme** ist also spezifiziert durch

$$\Omega_{Zr} = \{H \in [0:n]^F \mid \sum_{f \in F} H(f) = n\}.$$

Die ZV X_{Zr} bildet wie folgt Farbenfolgen auf Histogramme ab:

$$X_{Zr} : F^n \ni (f_1, \ldots, f_n) \mapsto (F \ni f \mapsto |\{i \mid f_i = f\}|).$$

Mittels X_{Zr} versehen wir nun der Raum der Farbenhistogramme mit dem zugehörigen Bildmaß, das wir mit P_{Zr} bezeichnen.

Multinomialverteilung

Satz. Kommt in der Urne mit N Kugeln die Farbe $f \in F$ genau N_f-mal vor, so ergibt sich bei n-maligem Ziehen mit Zurücklegen das Farbenhistogramm H mit Wahrscheinlichkeit

$$P_{Zr}(H) = \binom{n}{H} \cdot \prod_{f \in F} \left(\frac{N_f}{N}\right)^{H(f)}.$$

Dabei ist der **Multinomialkoeffizient** definiert durch:

$$\binom{n}{H} := \frac{n!}{\prod_{f \in F} H(f)!}.$$

Definition. Dieses W-Maß auf dem Raum der Farbenhistogramme heißt die **Multinomialverteilung** für n Stichproben (mit Zurücklegen) zur W-Funktion $(F \ni f \mapsto N_f/N)$.

Allgemeiner kann man die Multinomialverteilung definieren, wenn man von beliebiger W-Funktion ρ auf F ausgeht.

Multinomialverteilung: Beispiel

- Eine Urne enthalte 4 rote, 1 blaue und 3 schwarze Kugeln.
- Hier ist $N = 8 = 4 + 1 + 3$ die Gesamtanzahl von Kugeln,
- $N_r = 4$, $N_b = 1$ und $N_s = 3$.
- Bei sechsmaligem Ziehen ($n = 6$) mit Zurücklegen werde 3-mal eine rote, einmal eine blaue und 2-mal eine schwarze Kugel gezogen. Hier ist

$$H = \begin{pmatrix} \textcolor{red}{\bullet} & \textcolor{blue}{\bullet} & \bullet \\ 3 & 1 & 2 \end{pmatrix} \quad \text{und} \quad \binom{n}{H} = \frac{6!}{3! \cdot 1! \cdot 2!}.$$

- Damit ergibt sich für die Wahrscheinlichkeit $P_{Zr}(H)$, bei sechsmaligem Ziehen mit Zurücklegen dieses Farbenhistogramm H zu erhalten:

$$P_{Zr}(H) = \frac{6!}{3! \cdot 1! \cdot 2!} \cdot \underbrace{\left(\frac{4}{8}\right)^3}_{\text{rot}} \cdot \underbrace{\left(\frac{1}{8}\right)^1}_{\text{blau}} \cdot \underbrace{\left(\frac{3}{8}\right)^2}_{\text{schwarz}}.$$

.

Multinomialkoeffizienten

Satz. Die Anzahl der Farbenfolgen der Länge n, in denen die c Farben $1, 2, \ldots, c$ mit Vielfachheiten $h_1, h_2, \ldots, h_c$ vorkommen, ist gleich dem Multinomialkoeffizienten

$$\binom{n}{h_1, \ldots, h_c} = \frac{n!}{h_1! \cdot \ldots \cdot h_c!}.$$

Beweis. h_1 der n Positionen werden mit Farbe 1 belegt. Das geht auf $\binom{n}{h_1}$ Weisen. Von den restlichen $n - h_1$ Positionen werden h_2 mit Farbe 2 belegt. Das geht auf $\binom{n-h_1}{h_2}$ Arten. Insgesamt erhält man durch Induktion:

$$\begin{aligned}
\binom{n}{h_1, \ldots, h_c} &= \binom{n}{h_1} \cdot \binom{n-h_1}{h_2} \cdot \binom{n-h_1-h_2}{h_3} \cdot \ldots \cdot \binom{h_c}{h_c} \\
&= \frac{n!}{h_1!(n-h_1)!} \cdot \frac{(n-h_1)!}{h_2!(n-h_1-h_2)!} \cdot \frac{(n-h_1-h_2)!}{h_3!(n-h_1-h_2-h_3)!} \cdot \ldots \\
&= \frac{n!}{h_1! \cdot \ldots \cdot h_c!}
\end{aligned}$$

Damit ist der Satz bewiesen. $\square$

Multinomialverteilung: Beweis

Für die Wahrscheinlichkeit, dass sich das Farbenhistogramm H ergibt, gilt:

$$
\begin{aligned}
P_{Zr}(H) &= P_{ZR}(X_{Zr} = H) \quad = \quad P_{ZR}(\{\mathbf{f} \in F^n \mid X_{Zr}(\mathbf{f}) = H\}) \\
&= \sum_{\mathbf{f} \in F^n : X_{Zr}(\mathbf{f}) = H} P_{ZR}(\mathbf{f}) \\
&= \sum_{\mathbf{f} \in F^n : X_{Zr}(\mathbf{f}) = H} \prod_{i=1}^{n} \frac{N_{f_i}}{N} \quad (P_{ZR} \text{ ist Produktmaß auf } F^n) \\
&= \sum_{\mathbf{f} \in F^n : X_{Zr}(\mathbf{f}) = H} \prod_{f \in F} \left(\frac{N_f}{N} \right)^{H(f)} \\
&= \binom{n}{H} \cdot \prod_{f \in F} \left(\frac{N_f}{N} \right)^{H(f)},
\end{aligned}
$$

denn die Summanden im vorletzten Glied der Gleichungskette sind für alle $\mathbf{f} \in F^n$ mit $X_{Zr}(\mathbf{f}) = H$ gleich. Ferner ist die Anzahl der Summanden der Multinomialkoeffizient $\binom{n}{H}$.
□

Binomialverteilung

Hierbei handelt es sich um folgenden Spezialfall ungeordneter Stichproben mit Zurücklegen:

- Es gibt nur zwei Farben: $F = \{0, 1\}$, wobei 1 als Erfolg und 0 als Misserfolg interpretiert wird.
- Die Erfolgswahrscheinlichkeit wird wieder mit p, die Misserfolgswahrscheinlichkeit mit $q = 1 - p$ bezeichnet.
- Die Farbenhistogramme bei n Stichproben bestehen jetzt aus zwei Balken, die zusammen die Höhe n haben. Also reicht es, nur die Anzahl der Erfolge zu protokollieren. Das führt zum modifizierten Raum $[0 : n]$.
- Die Wahrscheinlichkeit bei n Stichproben k-mal Erfolg zu haben ist

$$B_{n,p}(k) := \binom{n}{k} p^k (1-p)^{n-k}.$$

Diese Verteilung auf $[0 : n]$ heißt die **Binomialverteilung** zur Erfolgswahrscheinlichkeit p.

Stichproben ohne Zurücklegen 1

Ausgehend vom fiktiven Laplace-Raum der injektiven n-Tupel

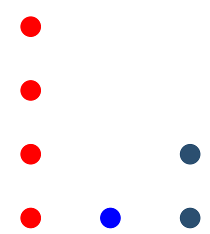
$$\Omega_{\neq} := \{\omega \in [1:N]^n | i < j \Rightarrow \omega_i \neq \omega_j\},$$

der $(N)_n := \prod_{k=0}^{n-1}(N-k)$-elementig ist, könnten wir nun schrittweise mittels geeigneter ZVs die Information komprimieren:

$$\Omega_{\neq} \xrightarrow{X_{zR}} \Omega_{zR} \xrightarrow{X_{zr}} \Omega_{zr}.$$

Konkret wird einer injektiven Folge farbiger Kugelnummern mittels X_{zR} die zugehörige Farbenfolge zugeordnet, die ihrerseits mittels X_{zr} zum zugehörigen Farbenhistogramm vergröbert wird:

$$(\color{red}{2}, \color{blue}{4}, \color{red}{5}, \color{red}{6}, 1, \color{red}{7}, 3) \xrightarrow{X_{zR}} (\color{red}{\bullet}, \color{blue}{\bullet}, \color{red}{\bullet}, \color{red}{\bullet}, \bullet, \color{red}{\bullet}, \bullet) \xrightarrow{X_{zr}}$$



Wir beschreiben kurz den ersten Übergang $\Omega_{\neq} \xrightarrow{X_{zR}} \Omega_{zR}$ inklusive Bildmaß, überlassen die Ausführung der zweiten Informationskompression $\Omega_{zR} \xrightarrow{X_{zr}} \Omega_{zr}$ als Übung, und werden schließlich einen einfacheren, alternativen Weg zur Beschreibung von (Ω_{zr}, P_{zr}) vorstellen.

Stichproben ohne Zurücklegen 2

Da nicht zurückgelegt wird, besteht Ω_{zR} aus allen n-komponentigen Farbenfolgen $\mathbf{f} = (f_1, \ldots, f_n)$, in denen die Farbe $f \in F$ höchstens N_f-mal vorkommt:
$\Omega_{zR} = \{\mathbf{f} \in F^n \mid \forall f \in F : |\{i \in [1:n] \mid f_i = f\}| \le N_f\}$.

Satz. Ist $\mathbf{f} \in \Omega_{zR}$ und kommt die Farbe f in $\mathbf{f}$ genau n_f-mal vor, so ist

$$\left|X_{zR}^{-1}(\mathbf{f})\right| = \prod_{f \in F} (N_f)_{n_f} \quad \text{sowie} \quad P_{zR}(\mathbf{f}) = \frac{\prod_{f \in F} (N_f)_{n_f}}{(N)_n} .$$

Beweis. Kommt die Farbe $f \in F$ an genau n_f Stellen in $\mathbf{f}$ vor, so stammt dieser einfarbige Bereich von einer n_f-elementigen Teilmenge der Kugeln mit Kugelnummer aus der N_f-elementigen Nummernmenge F_f. Diese n_f-elementige Teilmenge kann man auf $n_f!$ Weisen anordnen. Insgesamt ergibt sich daher

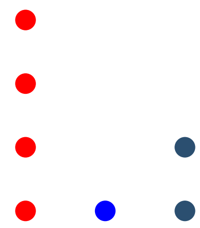
$$\left|X_{zR}^{-1}(\mathbf{f})\right| = \prod_{f \in F} \binom{N_f}{n_f} \cdot n_f! = \prod_{f \in F} (N_f)_{n_f} .$$

Dies beweist die erste Behauptung. Die zweite folgt mit $|\Omega_{\neq}| = (N)_n$ aus der Formel für Bildmaße. $\square$

Stichproben ohne Zurücklegen 3

Wir gehen wieder vom fiktiven Laplace-Raum $\Omega_{\neq}$ der injektiven n-Tupel aus und ordnen jetzt einer injektiven Folge farbiger Kugelnummern zunächst die zugehörige n-elementige Menge zu, bevor wir wieder zum Farbenhistogramm übergehen. Konkret:

$$(2, 4, 5, 6, 1, 7, 3) \xrightarrow{X} \{2, 4, 5, 6, 1, 7, 3\} \xrightarrow{Y}$$



Bezeichnet Ω' den Raum aller n-elementigen Teilmengen von $[1 : N]$ und $\Omega_{zr} = \{H \in \prod_{f \in F} [0 : N_f] \mid \sum_{f \in F} H(f) = n\}$ den Raum aller möglichen Histogramme im zr-Szenario, so sind beide ZVs genauer spezifiziert durch:

$$\begin{aligned} X &:= (\Omega_{\neq} \ni (\omega_1, \ldots, \omega_n) \mapsto \{\omega_1, \ldots, \omega_n\} \in \Omega') \\ Y &:= (\Omega' \ni A \mapsto (F \ni f \mapsto |A \cap F_f|) \in \Omega_{zr}), \end{aligned}$$

wobei $F_f := \{i \in [1 : N] \mid \phi(i) = f\}$.

Frage: Wie sehen die Bildmaße P_X und P_Y zu den ZVs X bzw. Y aus?

Hypergeometrische Verteilung (1)

Satz 1. P_X ist die Gleichverteilung auf Ω'.

Beweis.

- Jede n-elementige Teilmenge von $[1:N]$ hat $n!$ verschiedene X-Urbilder.
- Der Laplace-Raum $\Omega_{\neq}$ hat $(N)_n$ Elemente.
- Also ist $P_X(A) = n!/(N)_n = \binom{N}{n}^{-1}$, für alle $A \in \Omega'$. D.h. alle Ergebnisse in Ω' sind gleichwahrscheinlich. $\square$

Satz 2. Enthält eine Urne N gefärbte Kugeln und kommt die Farbe $f \in F$ mit Vielfachheit N_f vor, so ist die Wahrscheinlichkeit, dass bei n Ziehungen ohne Rücklegen h_f-mal Farbe f gezogen wird (d.h. es resultiert das Histogramm $H = (h_f)_{f\in F}$), gegeben durch

$$P_Y(H) = \frac{\prod_{f\in F} \binom{N_f}{h_f}}{\binom{N}{n}}.$$

Diese Verteilung heißt die **hypergeometrische Verteilung** zum Parametersatz $(N_f)_{f\in F}$.

Hypergeometrische Verteilung (2)

Beweis.

- $[1:N] = \sqcup_{f\in F} F_f$ ist die Zerlegung der Menge der Kugelnummern in gleichfarbige Bereiche.
- Zum Histogramm $H = (h_f)_{f\in F}$ erhält man alle Y-Urbilder A, indem man zu jeder Farbe f eine h_f-elementige Teilmenge A_f von F_f wählt und $A = \sqcup_{f\in F} A_f$ bildet.
- Für A_f gibt es $\binom{N_f}{h_f}$ viele Möglichkeiten.
- Der Nenner ist die Kardinalität von Ω'.
- Die Behauptung ergibt sich nun aus der Formel für Bildmaße und Satz 1. $\square$

Hypergeometrische Verteilung: Beispiel

- Beim Skatspiel enthält jeder der drei Spieler zehn Karten aus einem Pack mit 32 Karten. Zwei Karten (der Skat) werden zunächst beiseite gelegt. Es gibt 4 Asse.
- Wie groß ist die Wahrscheinlichkeit, dass Spieler A genau drei Asse erhält?
- **Urnenmodell**: Ziehen ohne Rücklegen und ohne Berücksichtigung der Reihenfolge.
- Die Urne enthält 32 Kugeln, von denen 4 blau (Ass) und 28 rot (kein Ass) sind. Also ist

$$N = 32, \quad N_{\text{blau}} = 4, \quad N_{\text{rot}} = 28.$$

- Bezogen auf Spieler A wird 10 mal gezogen. Gesucht ist die Wahrscheinlichkeit, dass 3 blaue und 7 rote Kugeln gezogen werden. Also ist

$$n = 10, \quad h_{\text{blau}} = 3, \quad h_{\text{rot}} = 7.$$

- Die hypergeometrische Verteilung liefert für die gesuchte Wahrscheinlichkeit den Wert

$$\frac{\binom{4}{3} \cdot \binom{28}{7}}{\binom{32}{10}} = \frac{66}{899} \approx 0,0734149.$$

Geometrische Verteilung

Ab jetzt geht es um Verteilungen auf unendlichen Ergebnisräumen.

- Wir betrachten ein Bernoulli-Experiment mit Erfolgswahrscheinlichkeit $p \in (0, 1)$.
- **Frage**: Wie groß ist die Wahrscheinlichkeit $P(k)$, erstmals im k-ten Schritt Erfolg zu haben?
- **Lösung**: Messraum $(\mathbb{N}, 2^{\mathbb{N}})$, W-Maß (genauer: Zähldichte): $P(k) = p \cdot (1 - p)^{k-1}$. Dieses W-Maß heißt die **geometrische Verteilung** auf $\mathbb{N}$.

Satz. Die geometrische Verteilung ist eine Verteilung.

Beweis. Wegen $P(k) \geq 0$ bleibt zu zeigen, dass $P(\mathbb{N}) = 1$. Dies folgt aber aus $0 < p < 1$ und den Eigenschaften der geometrischen Reihe:

$$\begin{aligned} P(\mathbb{N}) &= \sum_{k=1}^{\infty} P(k) = \sum_{k=1}^{\infty} p(1-p)^{k-1} = p \sum_{k=0}^{\infty} (1-p)^k \\ &= p \cdot \frac{1}{1 - (1-p)} = 1. \end{aligned}$$

$\square$

Beispiel zum Nacharbeiten(1)

Man kann die Zahlen 1,. . . ,49 eines Lottoscheins in die drei *disjunkten* Mengen

- $S_0 :=$ "Zahlen, die weder durch 3 noch durch 7 teilbar sind"
- $S_3 :=$ "Zahlen, die durch 3, nicht aber durch 7 geteilt werden können"
- $S_7 :=$ "Zahlen, die durch 7 geteilt werden können"

einteilen mit den Mächtigkeiten $|S_0| = 26$, $|S_3| = 16$ und $|S_7| = 7$. Nun werden gleichverteilt 6 unterschiedliche Felder markiert. Modelliere die Wahrscheinlichkeit, dass genau H_0, H_3 und H_7 Felder in den jeweiligen Mengen markiert wurden.

1. Nenne das passende Standardmodell beim Namen.
2. Gebe einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ an, der dieses Zufallsexperiment modelliert.
3. Wie hoch ist die Wahrscheinlichkeit für das Ereignis $H_0 = 3$, $H_3 = 2$, $H_7 = 1$?

Beispiel zum Nacharbeiten(2)

1. Da wir ohne Zurücklegen ziehen und die Reihenfolge nicht berücksichtigen handelt es sich um eine hypergeometrische Verteilung.
2. Der Ergebnisraum enthält Histogramme, die σ-Algebra ist seine Potenzmenge und das Wahrscheinlichkeitsmaß $P : \mathcal{A} \to [0,1]$ definieren wir über eine Zähldichte $p : \Omega \to [0,1]$.

$$\Omega := [0:26] \times [0:16] \times [0:7]$$

$$\mathcal{A} := 2^{\Omega}$$

$$p(H_0, H_3, H_7) := \frac{\binom{N_0}{H_0}\binom{N_3}{H_3} \cdot \binom{N_7}{H_7}}{\binom{N}{n}} \qquad \forall (H_0, H_3, H_7) \in \Omega$$

$$P(A) := \sum_{\omega \in A} p(\omega) \qquad \forall A \in \mathcal{A}$$

3. Es ist nach der Wahrscheinlichkeit des Histograms mit $H_0 = 3$, $H_3 = 2$ und $H_7 = 1$ gefragt. Diese ist gegeben durch:

$$p((3,2,1)) = \frac{\binom{26}{3} \cdot \binom{16}{2} \cdot \binom{7}{1}}{\binom{49}{6}} = \frac{26!}{23!\,3!}\,\frac{16!}{14!\,2!}\,\frac{7!}{6!\,1!}\,\frac{43!\,6!}{49!} = \frac{13\,5^3\,8}{47\,23\,11\,7} \approx 15.6\%$$

Poisson-Verteilung: Motivation

Wir beginnen zur Motivation mit einem Informatik-Beispiel.

- Wieviele E-Mails werden in einem festen Zeitintervall $(0,t]$, $t>0$, über einen Mail-Server geleitet?
- Ergebnisraum: $\Omega := \mathbb{Z}_{\geq 0}$ (Anzahl der weiterzuleitenden E-Mails).
- Ereignisalgebra: 2^{Ω}.
- W-Maß: **?**

Heuristische Vorbetrachtungen:

- Zerlege das Zeitintervall $(0,t]$ in n Teilintervalle der Länge t/n.
- **Annahme** (bei großem n): pro Teilintervall ist höchstens eine E-Mail weiterzuleiten.
- **Ansatz**: Die Wahrscheinlichkeit, eine E-Mail in einem bestimmten Teilintervall der Länge t/n weiterleiten zu müssen ist proportional zur Teilintervalllänge: $\alpha \cdot t/n$, für eine Proportionalitätskonstante $\alpha > 0$.

Poisson-Verteilung: Modellierung

- Die Gesamtanzahl der weiterzuleitenden E-Mails im Zeitintervall $(0, t]$ ist daher grob modellierbar als das Ziehen von n Kugeln (mit Zurücklegen) aus einer Urne, die zwei Sorten von Kugeln enthält:
- Sorte 1: eine E-Mail ist im aktuellen Teilintervall weiterzuleiten;
  Sorte 0: keine E-Mail ist im aktuellen Teilintervall weiterzuleiten.
- Kugelanteil der Sorte 1 sollte $\alpha \cdot t/n$ betragen.
- Also handelt es sich hier grob um die Binomialverteilung B_{n,p_n} zum Erfolgsparameter $p_n = \alpha \cdot t/n$.

Satz. Es sei $\lambda := \alpha \cdot t$ und $p_n := \lambda/n$. Dann ist für jedes $k \in \mathbb{Z}_{\geq 0}$

$$\lim_{n\to\infty} B_{n,p_n}(k) = e^{-\lambda}\frac{\lambda^k}{k!} =: P_\lambda(k).$$

P_λ definiert ein W-Maß auf $\mathbb{Z}_{\geq 0}$, die sog. **Poisson-Verteilung** zum Parameter λ.

Poisson-Verteilung: Beweisskizze

Für jedes feste $k \in \mathbb{Z}_{\geq 0}$ erhält man mit $n \cdot p_n = \lambda$ für $n \to \infty$:

$$\begin{aligned} B_{n,p_n}(k) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \approx \frac{n^k}{k!} p_n^k (1-p_n)^{n-k} \approx \frac{\lambda^k}{k!}(1-p_n)^n \\ &= \frac{\lambda^k}{k!}(1-\frac{np_n}{n})^n = \frac{\lambda^k}{k!}(1-\frac{\lambda}{n})^n \to \frac{\lambda^k}{k!}e^{-\lambda}. \end{aligned}$$

Also existiert für alle k der Grenzwert und dieser stimmt mit $P_\lambda(k)$ überein.

Es bleibt zu zeigen, dass P_λ ein W-Maß auf $\Omega = \mathbb{Z}_{\geq 0}$ definiert. Die σ-Additivität ist für alle diskreten Fälle klar (Stichwort: großer Umordnungssatz!). Zeigen noch $P_\lambda(\Omega) = 1$:

$$P_\lambda(\Omega) = \sum_{k \geq 0} P_\lambda(k) = \sum_{k \geq 0} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k \geq 0}\frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1. \qquad \square$$

- Allgemein ist die Poisson-Verteilung ein natürliches Modell für die Anzahl von rein zufälligen Zeitpunkten in einem Zeitintervall.
- Für kleine Erfolgswahrscheinlichkeiten p_n (hier: $p_n = \lambda/n$) ist die Poisson-Verteilung eine gute Approximation der Binomialverteilung.

Poisson-Verteilung: seltene Ereignisse

In der Praxis verwendet man die Poisson-Verteilung als Modell überall dort, wo gezählt wird, wie viele von vielen möglichen, aber einzeln relativ unwahrscheinlichen unabhängigen Ereignissen eintreten. Es folgen einige typische Szenarien, in denen sinnvollerweise mit der Poisson-Verteilung modelliert wird.

- Die Wahrscheinlichkeit, dass ein bestimmter Einwohner einer Großstadt an einem bestimmten Abend in ein bestimmtes Kino der Stadt geht, ist sehr gering. Da aber in der Stadt sehr viele Menschen leben, liegt die Zahl der Leute, die an diesem Abend in dieses Kino gehen, typischerweise in einer uns vertrauten Größenordnung.
- Über eine bestimmte Straßenkreuzung fahren jedes Jahr sehr viele Fahrzeuge und pro Jahr passieren dort durchschnittlich nur 2 Autounfälle. Frage: Wie groß ist die Wahrscheinlichkeit, dass es nächstes Jahr

  (1) zu 0, **(2)** zu 4, **(3)** zu weniger als 3 Unfällen kommt?

- In einem Land gibt es pro Jahr 30 Selbstmorde pro 100.000 Einwohner. Wie hoch ist die Wahrscheinlichkeit, dass in einer typischen Stadt mit 120.000 Einwohnern im nächsten Jahr 40 Selbstmorde passieren?

Poisson-Verteilung: Beispiel

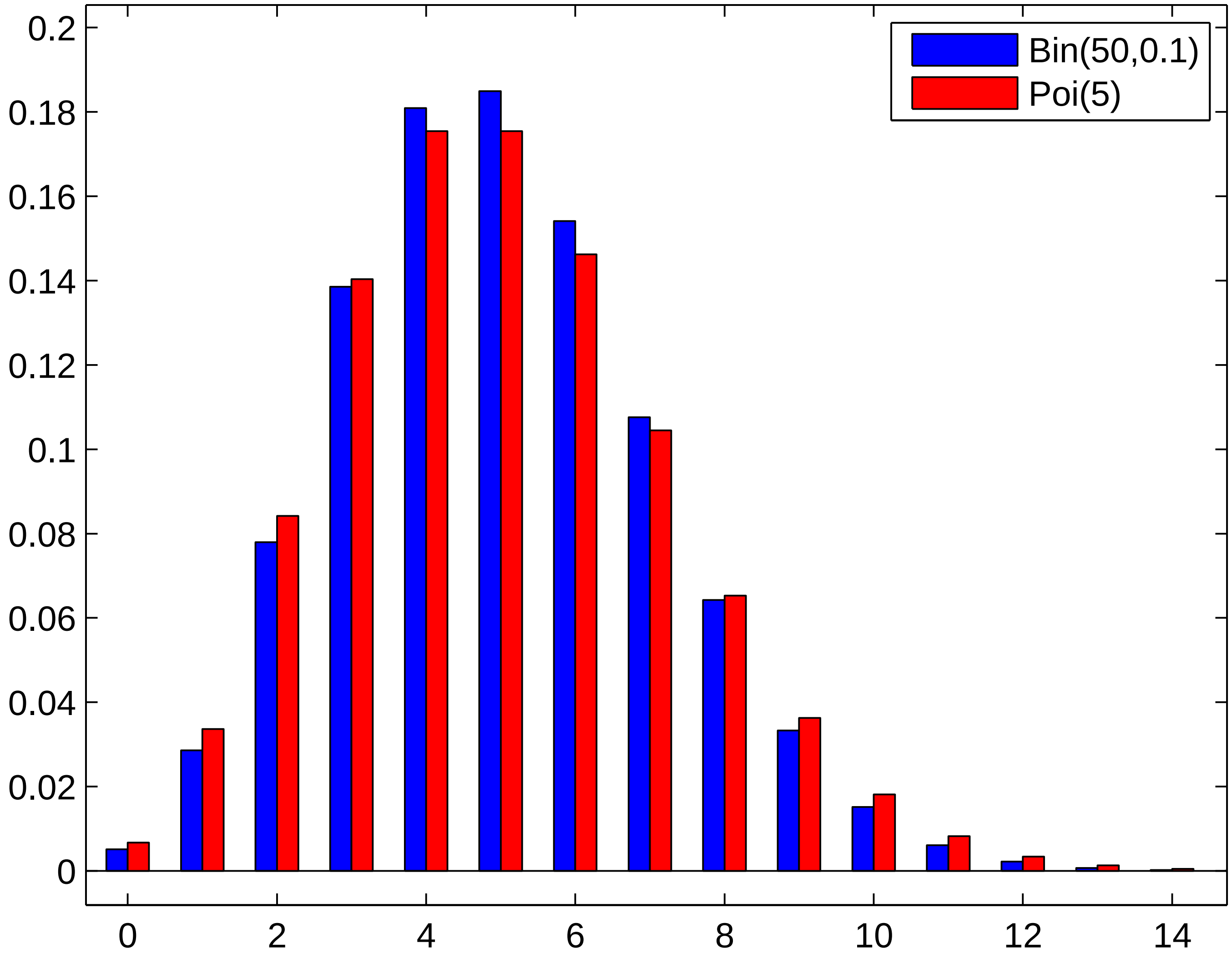
- In einer Stochastik-Vorlesung sitzen $n = 91$ Studierende. Die Wahrscheinlichkeit p, heute Geburtstag zu haben, ist (idealisiert) $p = 1/365$.
- Gesucht ist die Wahrscheinlichkeit $P(k)$, dass genau k Studierende heute Geburtstag haben.
- **Exakte Lösung**: Bernoulli-Experiment mit $n = 91$ Stichproben und Erfolgswahrscheinlichkeit $p = 1/365$. Gesucht ist für $k \in [0:91]$ der Wert der Binomialverteilung:

$$B_{91,1/365}(k) = \binom{91}{k} \cdot 365^{-k} \cdot \left(\tfrac{364}{365}\right)^{91-k}.$$

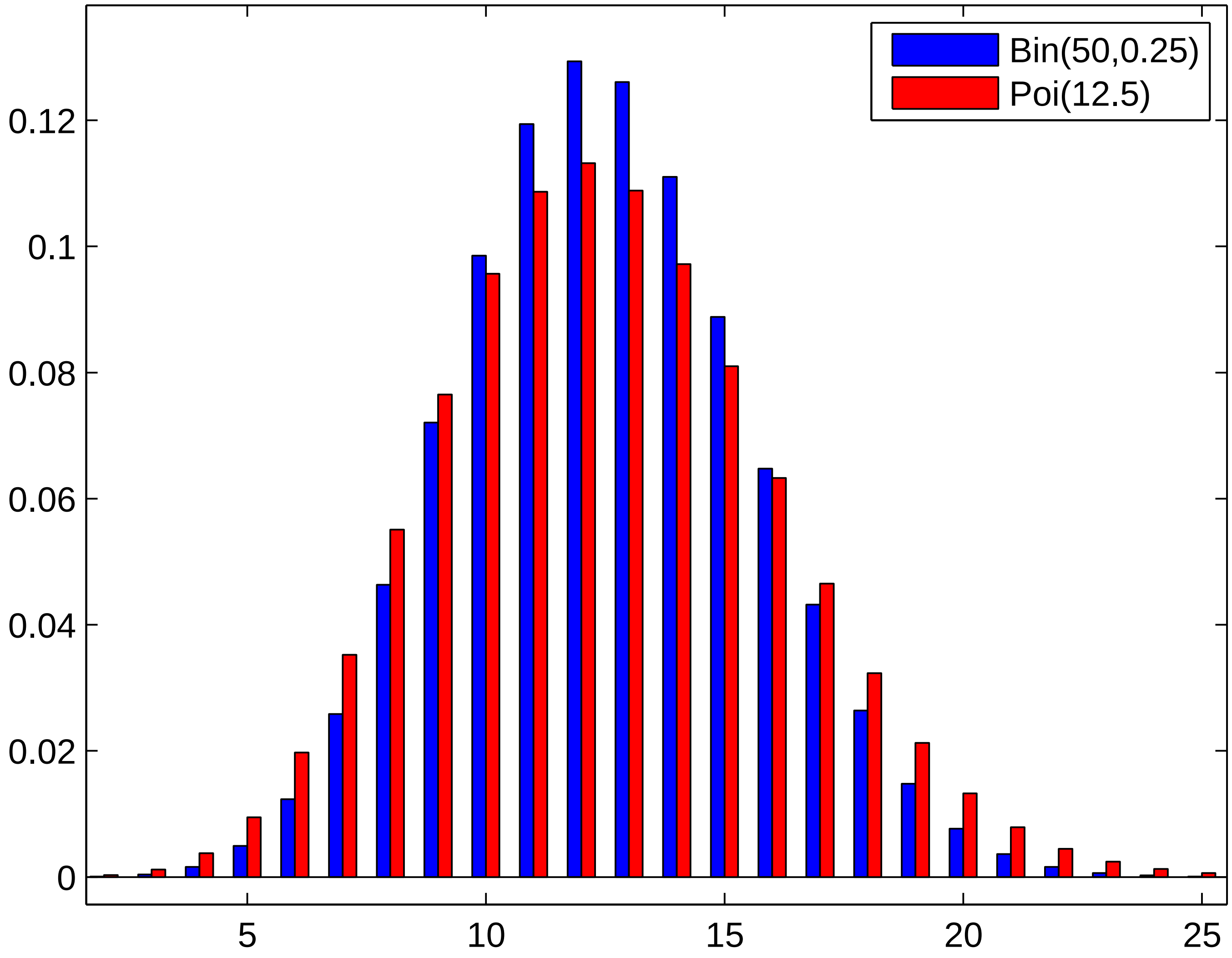
- **Approximative Lösung**: Wegen $p \ll 1$ und $\lambda = n \cdot p = 91/365 \approx 0.25$ wird die Binomialverteilung hier gut durch die Poisson-Verteilung $P_{0.25}$ approximiert:

$$B_{91,1/365}(k) \approx P_{0.25}(k) \approx e^{-0.25} \cdot \frac{0.25^k}{k!}.$$

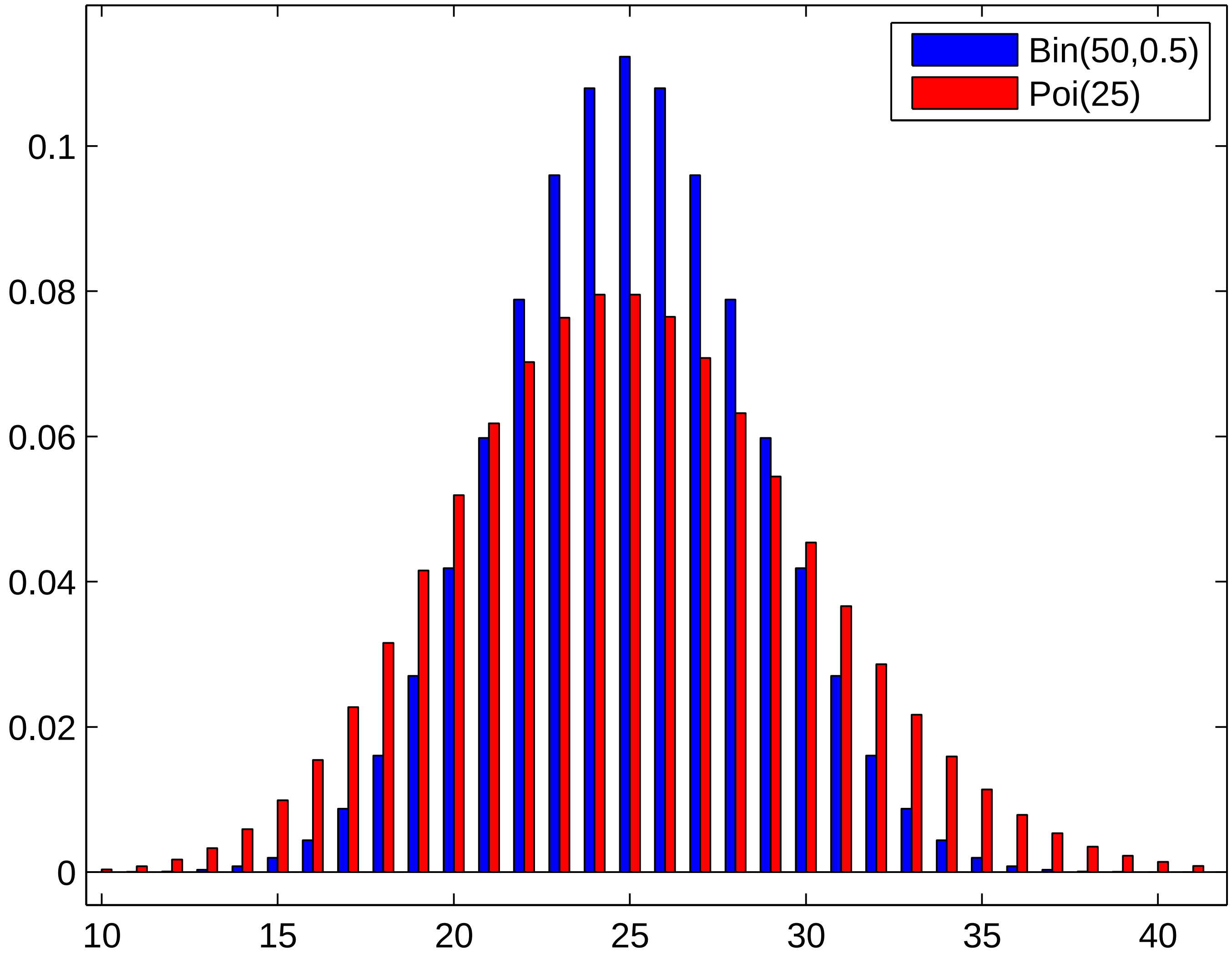
Binomial- & Poissonverteilung im Vergleich (1)



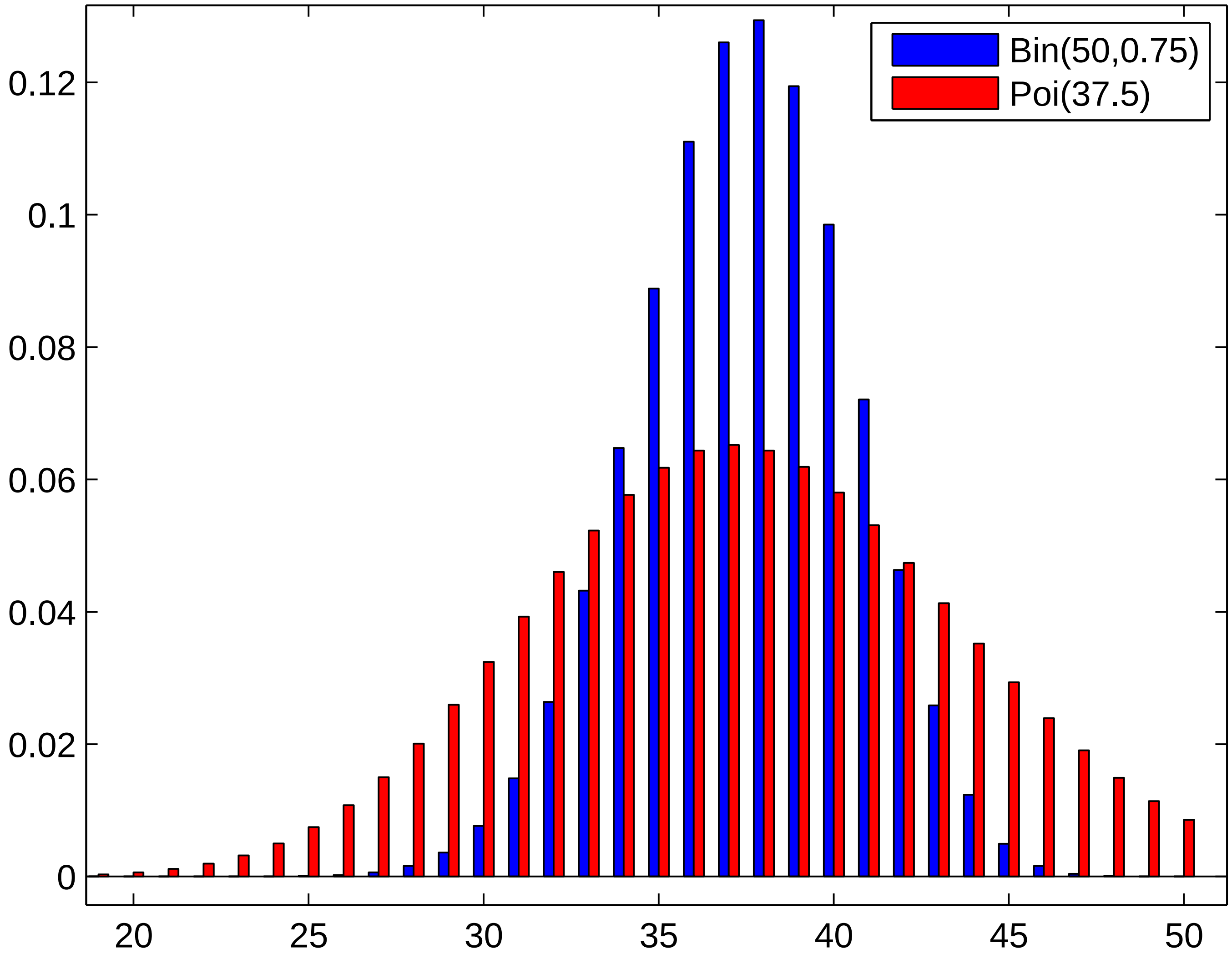
Binomial- & Poissonverteilung im Vergleich (2)



Binomial- & Poissonverteilung im Vergleich (3)



Binomial- & Poissonverteilung im Vergleich (4)



Gauß-Verteilungen

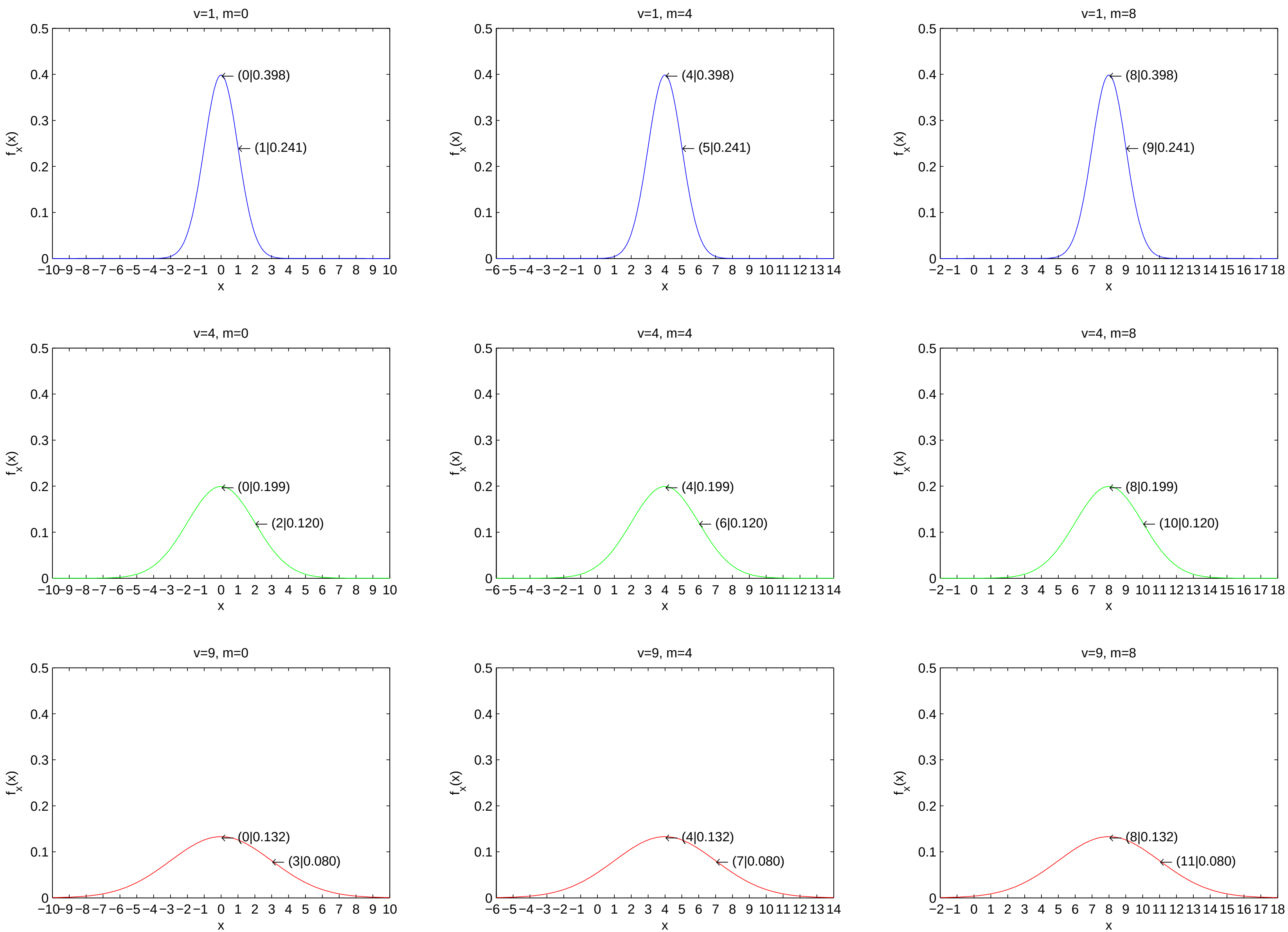
Von fundamentaler Bedeutung ist folgende(r)

- **Satz/Definition**. Es sei $m \in \mathbb{R}$ und $v > 0$ ein reeller Parameter. Dann wird durch

$$\phi_{m,v}(x) := \frac{1}{\sqrt{2\pi v}} e^{-(x-m)^2/(2v)}$$

  auf $(\mathbb{R}, \mathcal{B})$ eine Dichtefunktion definiert.
- Das zugehörige W-Maß $\mathcal{N}_{m,v}$ heißt die sog. **Gauß-Verteilung** oder **Normalverteilung** mit **Erwartungswert** m und **Varianz** v.
- Im Spezialfall $\mathcal{N}_{0,1}$ spricht man auch von der **Standard-Normalverteilung**.
- **Beweis** sowie Rechtfertigung der Begriffe Erwartungswert und Varianz: später.
- Auf der folgenden Folie zeigen wir einige Normalverteilungen.

Gauß-Verteilungen: Beispiele



Stochastik

Kapitel 5: Bedingte Wahrscheinlichkeit und Unabhängigkeit

Prof. Dr. Reinhard Klein
Folien von Prof. Dr. Michael Clausen

Sommersemester 2017

INSTITUT FÜR INFORMATIK II, UNIVERSITÄT BONN

Neubewertung von Wahrscheinlichkeiten 1

Die Information, dass ein Ereignis B eingetreten ist, sollte zu einer Neubewertung der Wahrscheinlichkeiten aller Ereignisse A führen.

Beispiel. Zweimaliges Ziehen ohne Zurücklegen aus einer Urne mit $N = b + r$ Kugeln: b blaue und r rote.

Wir denken uns die Kugeln durchnummeriert:

$$[1 : b] \quad \text{bzw.} \quad [b + 1 : N]$$

seien die Mengen der blauen bzw. roten Kugelnummern.

Das führt zum **Ergebnisraum**:

$$\Omega = \{(i, j) \in [1 : N]^2 \mid i \neq j\} \quad \text{mit Gleichverteilung.}$$

Wir betrachten folgende Ereignisse:

$$\begin{aligned} A &= \{\text{die erste Kugel ist blau}\} = \{(i, j) \in \Omega \mid i \leq b\} \\ B &= \{\text{die zweite Kugel ist blau}\} = \{(i, j) \in \Omega \mid j \leq b\} \end{aligned}$$

Neubewertung von Wahrscheinlichkeiten 2

- Vor Beginn des Experiments rechnet man mit Wahrscheinlichkeit

$$P(B) = \frac{(N-1)b}{N(N-1)} = \frac{b}{N}$$

  mit dem Eintreten von $B = \bigsqcup_{j=1}^{b}([1:N] \setminus \{j\}) \times \{j\}$.
- Wenn man beim ersten Zug eine blaue Kugel gezogen hat, d.h. A ist bereits eingetreten, wird man die Wahrscheinlichkeit des Eintretens von B revidieren, denn jetzt sind unter den $N-1$ Kugeln nur noch $b-1$ blaue Kugeln in der Urne.
- Die Neubewertung sollte also zur Wahrscheinlichkeit

$$\frac{b-1}{N-1}$$

  für das Eintreten des Ereignisses B führen.

Anforderungen an revidiertes W-Maß

Ausgehend vom W-Raum $(\Omega, \mathcal{A}, P)$ sollte das noch zu definierende revidierte W-Maß P_A in Abhängigkeit von der Information, dass A bereits eingetreten ist, folgende Eigenschaften haben:

(a) $P_A(A) = 1$, denn A ist jetzt ein sicheres Ereignis!

(b) Die Neubewertung der Teilereignisse B von A ist proportional zur ursprünglichen Bewertung:

$$\exists c_A > 0 \, \forall B \in \mathcal{A}, B \subseteq A : \; P_A(B) = c_A P(B).$$

Satz. (Neubewertung von Ereignissen) Es sei $(\Omega, \mathcal{A}, P)$ ein W-Raum und $A \in \mathcal{A}$ mit $P(A) > 0$. Dann gibt es genau ein W-Maß P_A auf $(\Omega, \mathcal{A})$ mit den Eigenschaften (a) und (b), nämlich

$$P_A(B) := \frac{P(A \cap B)}{P(A)} \quad \text{für } B \in \mathcal{A}.$$

Beweis. Eindeutigkeit: Wegen $1 = P_A(A) = c_A P(A)$ ist $c_A = 1/P(A)$. Also gibt es höchstens ein solches W-Maß. **Existenz**: siehe nächste Folie. $\square$

Bedingte Wahrscheinlichkeit

- **Definition**. Es sei $(\Omega, \mathcal{A}, P)$ ein W-Raum und $A \in \mathcal{A}$ mit $P(A) > 0$. Dann heißt

$$P(B|A) := \frac{P(A \cap B)}{P(A)} \quad \text{für } B \in \mathcal{A}$$

die **bedingte Wahrscheinlichkeit** von B unter der Bedingung A bzgl. P.
- **Satz**. Durch $B \mapsto P(B|A)$ wird ein W-Maß P_A auf $(\Omega, \mathcal{A})$ definiert.
- **Beweis**. $P_A(\Omega) = P(A \cap \Omega)/P(A) = 1$. Sei $B = \sqcup_{i\geq 1} B_i$ mit $B_i \in \mathcal{A}$. Dann gilt mit der σ-Additivität von P:

$$\begin{aligned} P_A(B) &= P_A(\sqcup_{i\geq 1} B_i) = \frac{P(A \cap \sqcup_{i\geq 1} B_i)}{P(A)} \\ &= \frac{P(\sqcup_{i\geq 1} A \cap B_i)}{P(A)} = \frac{\sum_{i\geq 1} P(A \cap B_i)}{P(A)} \\ &= \sum_{i\geq 1} P_A(B_i). \end{aligned}$$

Damit ist der Satz bewiesen. $\square$

Interpretation bedingter Wahrscheinlichkeiten

- **frequentistisch**:

  Bei oftmaliger Wiederholung des Zufallsexperiments ist $P(B|A)$ ungefähr

$$\frac{\text{Anzahl der Fälle, in denen } A \textbf{ und } B \text{ eingetreten sind}}{\text{Anzahl der Fälle, in denen } A \text{ eingetreten ist}}$$

- **subjektiv**:

  Ist P meine Einschätzung der Lage vor Beginn des Experiments, so ist $P(\cdot|A)$ meine Einschätzung, nachdem ich über das Eintreten von A **informiert** bin.

  - **Warnung**: ich selbst kenne den konkreten Ausgang ω des Zufallsexperiments nicht, denn sonst würde ich **wissen**, dass $\omega \in A$ das Ergebnis ist, womit **alles** zur **Gewissheit** würde.

Formel von Bayes

Elementar aber wichtig sind folgende Formeln.

Satz. Es sei $(\Omega, \mathcal{A}, P)$ ein W-Raum und $\Omega = \sqcup_{i\in I} B_i$ eine disjunkte Zerlegung von Ω in höchstens abzählbar unendlich viele Ereignisse $B_i \in \mathcal{A}$ mit $P(B_i) > 0$. Dann gilt:

(1) (**Fallunterscheidungsformel**) Für alle $A \in \mathcal{A}$ gilt

$$P(A) = \sum_{i\in I} P(B_i) P(A|B_i).$$

(2) (**Formel von Bayes**) Für alle $A \in \mathcal{A}$ mit $P(A) > 0$ und alle $k \in I$ gilt:

$$P(B_k|A) = \frac{P(B_k)P(A|B_k)}{\sum_{i\in I} P(B_i)P(A|B_i)}.$$

Formel von Bayes: Beweis

Beweis. (1) folgt aus der Definition der bedingten Wahrscheinlichkeit und der σ-Additivität von P:

$$\sum_{i\in I} P(B_i)P(A|B_i) = \sum_{i\in I} P(B_i)\cdot\frac{P(A\cap B_i)}{P(B_i)} = \sum_{i\in I} P(A\cap B_i) = P(\bigsqcup_{i\in I} A\cap B_i) = P(A).$$

(2) folgt aus (1) und der Definition der bedingten Wahrscheinlichkeit:

$$\begin{aligned} P(B_k|A) &= \frac{P(B_k\cap A)}{P(A)} \\ &\overset{(1)}{=} \frac{P(B_k\cap A)}{\sum_{i\in I} P(B_i)P(A|B_i)} \\ &\overset{\text{Def.}}{=} \frac{P(B_k)P(A|B_k)}{\sum_{i\in I} P(B_i)P(A|B_i)}. \qquad \square \end{aligned}$$

Beide Sätze wollen wir uns nun im Wegemodell veranschaulichen.

Das Wegemodell 1

Ausgehend von einem Startknoten • wird für $k \in I$ mit Wahrscheinlichkeit $P(B_k)$ zum Ereignis B_k verzweigt, von B_k geht es mit Wahrscheinlichkeit $P(A|B_k)$ zum Ereignis A:

$$\begin{tikzcd}
 & B_1 \arrow[dr, "P(A|B_1)"] & \\
\bullet \arrow[ur, "P(B_1)"] \arrow[r, "P(B_2)"] \arrow[dr, "P(B_3)"'] & B_2 \arrow[r, "P(A|B_2)"] & A \\
 & B_3 \arrow[ur, "P(A|B_3)"'] &
\end{tikzcd}$$

Insgesamt ist die Wahrscheinlichkeit, vom Startknoten über B_k nach A zu gelangen das Produkt der Wahrscheinlichkeiten entlang des Weges,

$$P(B_k) \cdot P(A|B_k) = P(B_k) \cdot \frac{P(A \cap B_k)}{P(B_k)} = P(A \cap B_k).$$

Die Fallunterscheidungsformel spaltet $P(A)$ also auf in die Wahrscheinlichkeiten verschiedener Wege nach A.

Das Wegemodell 2

Der Satz von Bayes besagt, dass für alle $A \in \mathcal{A}$ mit $P(A) > 0$ und alle $k \in I$ gilt:

$$P(B_k|A) = \frac{P(B_k)P(A|B_k)}{\sum_{i\in I} P(B_i)P(A|B_i)}.$$

Im gerade beschriebenen Wegemodell ist $P(B_k|A)$ also die Wahrscheinlichkeit des Weges über B_k im Verhältnis zur Gesamtwahrscheinlichkeit aller Wege:

$$P(B_k|A) = \frac{\text{Wahrscheinlichkeit des Wegs über } B_k}{\text{Gesamtwahrscheinlichkeit aller Wege}}.$$

Mehrstufige Experimente: Motivation

Bedingte Wahrscheinlichkeiten ermöglichen es uns, mehrstufige Zufallsexperimente systematisch in Angriff zu nehmen. Wir beginnen mit einem vieldiskutierten Beispiel, dem **Ziegenproblem**.

Bei einer Spielshow gibt es drei verschlossene Türen. Hinter zwei Türen befindet sich jeweils eine Ziege (Niete), hinter einer dritten der Hauptgewinn: ein Auto. Die Kandidatin hat die Möglichkeit, durch die richtige Türwahl das Auto zu gewinnen. Die Türwahl verläuft in Etappen:

- Zunächst wählt die Kandidatin eine der drei Türen. Diese Tür bleibt zunächst weiter verschlossen.
- Der Moderator öffnet eine andere Tür, hinter der sich eine Ziege befindet.
- Jetzt wird die Kandidatin aufgefordert, ihre endgültige Wahl zwischen den beiden noch verschlossenen Türen zu treffen.

Frage: Wie soll sich die Kandidatin verhalten? Soll sie bei ihrer ersten Wahl bleiben oder lohnt es, dass sie sich umentscheidet?

Charakteristika mehrstufiger Experimente

Bevor wir auf die Lösung des Ziegenproblems eingehen, wollen wir Charakteristika mehrstufiger Experimente hervorheben:

- Das Gesamtexperiment besteht aus n nacheinander ausgeführten Teilexperimenten (Ziegenproblem: $n = 3$).
- Auf dem noch zu spezifizierenden W-Raum $(\Omega, \mathcal{A}, P)$ für das Gesamtexperiment hat man ZVs $X_1, \ldots, X_n$, die die Ergebnisse der Teilexperimente beschreiben.
- Während das erste Teilexperiment X_1 vorab eindeutig beschrieben ist (Ziegenproblem: Wahl einer Tür gemäß Gleichverteilung), kann die Verteilung des Teilexperiments X_k vom Ausgang der bisher durchgeführten Teilexperimente $X_1, \ldots, X_{k-1}$ abhängen.

Wir veranschaulichen das am Beispiel des Ziegenproblems.

Das Ziegenproblem

Bezeichnet man die drei Türen mit

$$a \text{ (Auto)}, \; n \text{ (1. Niete) sowie } z \text{ (2. Niete = Ziege)},$$

so lässt sich das Ergebnis eines Gesamtexperiments schreiben als Tripel

$$(\omega_1, \omega_2, \omega_3) \in \{a, n, z\}^3.$$

- Wählt die Kandidatin gemäß Gleichverteilung eine der drei Türen aus, so ist $X_1 : (\omega_1, \omega_2, \omega_3) \mapsto \omega_1$ gleichverteilt auf $\Omega_1 := \{a, n, z\}$.
- Fall $\omega_1 \in \{n, z\}$: Hier **muss** der Moderator in der zweiten Etappe die Tür mit der anderen Niete aufdecken.
- Fall $\omega_1 = a$: Hier hat der Moderator in der zweiten Etappe die Wahl zwischen der n- und der z-Tür.
- Die ZV $X_2 : (\omega_1, \omega_2, \omega_3) \mapsto \omega_2$ führt in Abhängigkeit von ω_1 zu unterschiedlichen Verteilungen.

Multiplikationsformel

Vorbereitend benötigen wir noch folgenden

Satz. (Multiplikationsformel) Ist $(\Omega, \mathcal{A}, P)$ W-Raum, so gilt für $A_1, \ldots, A_n \in \mathcal{A}$ und $A := A_1 \cap \ldots \cap A_n$ mit $P(A) > 0$:

$$P(A) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot \ldots \cdot P(A_n|A_1 \cap \ldots \cap A_{n-1}).$$

Beweis.

- Nach Voraussetzung sind alle bedingten Wahrscheinlichkeiten rechts definiert und von Null verschieden.
- Wendet man die Definition der bedingten Wahrscheinlichkeit an, so sieht man, dass sich rechts alles weghebt bis auf den Faktor $P(A_1 \cap \ldots \cap A_n) = P(A)$, der auf der linken Seite steht:

$$P(A_1) \cdot \frac{P(A_1 \cap A_2)}{P(A_1)} \cdot \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)} \cdot \ldots \cdot \frac{P(A_1 \cap \ldots \cap A_n)}{P(A_1 \cap \ldots \cap A_{n-1})}.$$

Damit ist die Multiplikationsformel bewiesen. $\square$

Konstruktion mehrstufiger Modelle

Satz. Gegeben seien abzählbare Ereignisräume $\Omega_1, \ldots, \Omega_n$, $n \geq 2$. Weiter sei p_1 eine W-Funktion auf Ω_1 und für $k \in [2:n]$ und $\omega_i \in \Omega_i$ für $i < k$ sei $p_{k|\omega_1,\ldots,\omega_{k-1}}$ eine W-Funktion auf Ω_k. Sei ferner $\Omega = \prod_{i=1}^n \Omega_i$ der Produktraum und $X_i : \Omega \to \Omega_i$ die i-te Projektion. Dann existiert genau ein W-Maß P auf $(\Omega, 2^\Omega)$ mit

(a) Für alle $\omega_1 \in \Omega_1$ gilt $P(X_1 = \omega_1) = p_1(\omega_1)$.

(b) Für alle $k \in [2:n]$ und $\omega_i \in \Omega_i$ gilt

$$P(X_k = \omega_k | X_1 = \omega_1, \ldots, X_{k-1} = \omega_{k-1}) = p_{k|\omega_1,\ldots,\omega_{k-1}}(\omega_k),$$

sofern $P(X_1 = \omega_1, \ldots, X_{k-1} = \omega_{k-1}) > 0$.

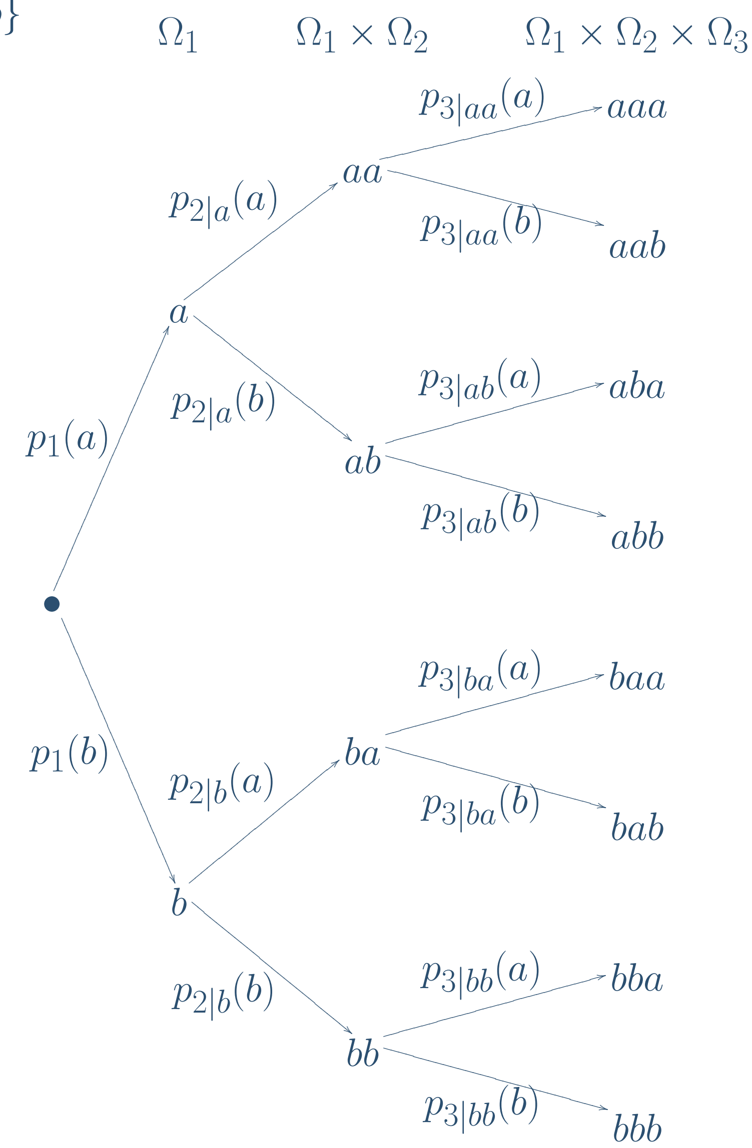
Das W-Maß P ist gegeben durch

$$P(\{\omega\}) = p_1(\omega_1) \cdot p_{2|\omega_1}(\omega_2) \cdot p_{3|\omega_1,\omega_2}(\omega_3) \cdot \ldots \cdot p_{n|\omega_1,\ldots,\omega_{n-1}}(\omega_n)$$

für $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$.

Mehrstufige Modelle als Baumdiagramm

$$\Omega_1 = \Omega_2 = \Omega_3 = \{a, b\}$$



Beweis des Satzes (1)

Eindeutigkeit von P: Erfüllt das W-Maß P die Eigenschaften (a) und (b), so folgt die behauptete Formel für $P(\{\omega\})$ aus $\{\omega\} = \cap_{i=1}^n \{X_i = \omega_i\}$ und der Multiplikationsformel für die Ereignisse $A_i := \{X_i = \omega_i\}$.

Existenz von P: Sei P für $\omega = (\omega_1, \ldots, \omega_n) \in \Omega := \prod_{i=1}^n \Omega_i$ definiert durch

$$P(\{\omega\}) := p_1(\omega_1) \cdot p_{2|\omega_1}(\omega_2) \cdot p_{3|\omega_1,\omega_2}(\omega_3) \cdot \ldots \cdot p_{n|\omega_1,\ldots,\omega_{n-1}}(\omega_n).$$

Dann folgt für alle $k \in [1:n]$ und alle $(\omega_1, \ldots, \omega_k) \in \Omega_1 \times \ldots \times \Omega_k$:

$$\begin{aligned} P(X_1 = \omega_1, \ldots, X_k = \omega_k) &= \sum_{\omega_{k+1}\in\Omega_{k+1},\ldots,\omega_n\in\Omega_n} P(\{(\omega_1, \ldots, \omega_n)\}) \\ &= p_1(\omega_1) \cdots p_{k|\omega_1,\ldots,\omega_{k-1}}(\omega_k) \cdot \\ &\quad \underbrace{\sum_{\omega_{k+1}\in\Omega_{k+1}} p_{k+1|\omega_1,\ldots,\omega_k}(\omega_{k+1})}_{=1 \text{ da } p_{k+1|\ldots} \text{ W-Fkt.}} \cdots \underbrace{\sum_{\omega_n\in\Omega_n} p_{n|\omega_1,\ldots,\omega_{n-1}}(\omega_n)}_{=1, \text{ da } p_{n|\ldots,} \text{ W-Fkt.}} . \end{aligned} \quad (1)$$

Beweis des Satzes (2)

Im Fall $k = 1$ ergibt sich $P(X_1 = \omega_1) = p_1(\omega_1)$, also (a). Indem man über alle ω_1 summiert, erhält man, dass P ein W-Maß auf Ω definiert:

$$\begin{aligned} P(\Omega) &= P\Big(\bigsqcup_{\omega_1 \in \Omega_1} \{\omega_1\} \times \Omega_2 \times \ldots \times \Omega_n\Big) \\ &= \sum_{\omega_1 \in \Omega_1} P(X_1 = \omega_1) = \sum_{\omega_1 \in \Omega_1} p_1(\omega_1) = 1. \end{aligned}$$

Für $k > 1$ ist wegen der Gleichung (1) auf der letzten Folie

$$P(X_1 = \omega_1, \ldots, X_k = \omega_k) = P(X_1 = \omega_1, \ldots, X_{k-1} = \omega_{k-1}) p_{k|\omega_1, \ldots, \omega_{k-1}}(\omega_k),$$

woraus (b) folgt. $\square$

Anschaulich zeigt das Baumdiagramm:

Die Wahrscheinlichkeit von $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$ im mehrstufigen Modell ist das Produkt der Kantengewichte (Übergangswahrscheinlichkeiten) entlang des Weges vom Startpunkt $\bullet$ zum n-Tupel ω.

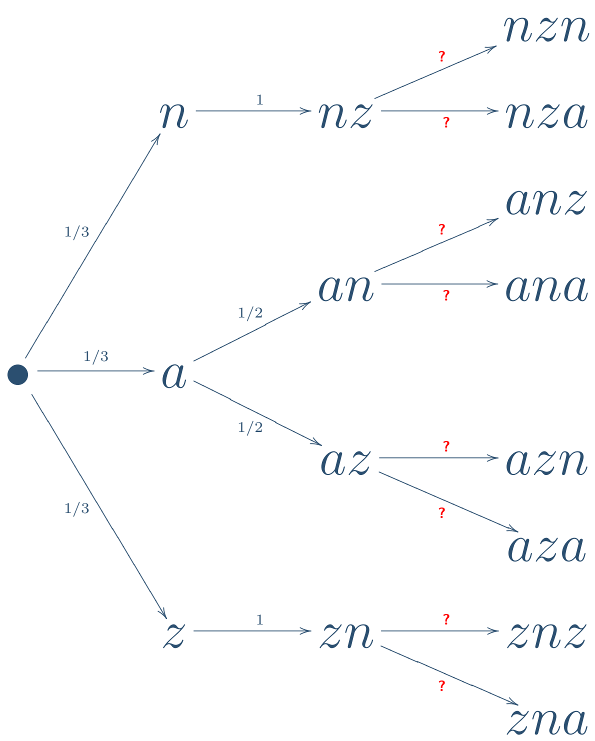
Variantes des Ziegenproblems

Im folgenden werden wir für drei Varianten des Ziegenproblems die Erfolgswahrscheinlichkeiten berechnen:

- Stufe 1: Die Kandidatin wählt eine Tür gemäß Gleichverteilung.
  (Gilt für alle drei Varianten.)
- Stufe 2: Der Moderator wählt eine der verbleibenden Nieten-Türen gemäß Gleichverteilung.
  (Gilt für alle drei Varianten.)
- Stufe 3:
  - Variante 1: Kandidatin bleibt **immer** bei ihrer ersten Wahl.
  - Variante 2: Kandidatin entscheidet per Münzwurf.
  - Variante 3: Kandidatin bleibt **nie** bei ihrer ersten Wahl.

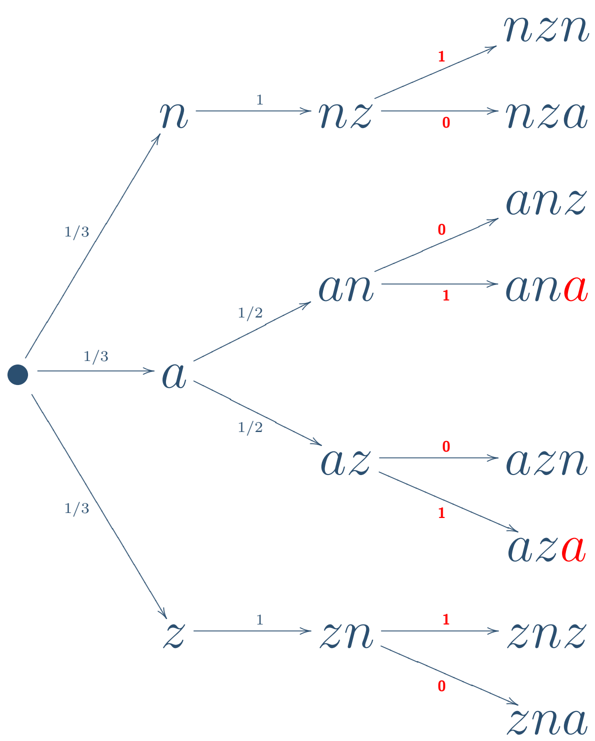
Wir zeigen zunächst das für alle Varianten gleiche Grundgerüst der Baumdiagramme.

Baumdiagrammschablone zum Ziegenproblem



Die Gewichte der mit **?** gekennzeichneten Pfeile sind variantenabhängig.

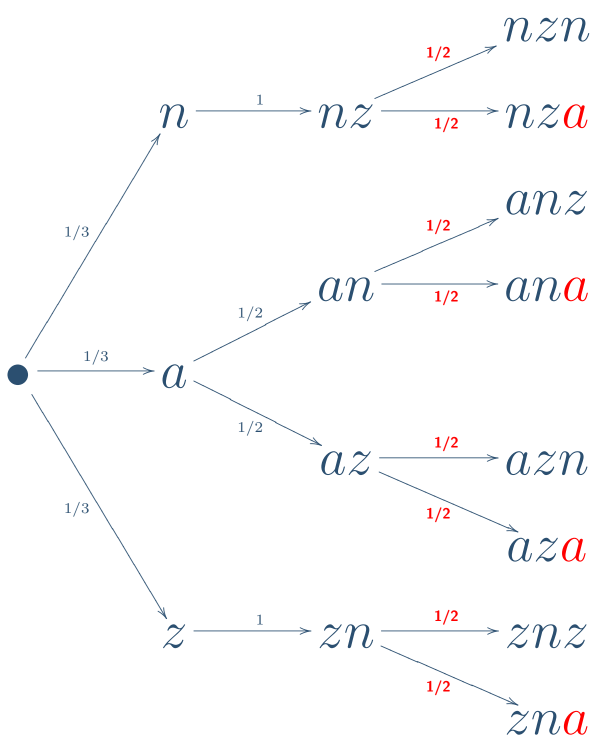
Ziegenproblem: Variante 1



Bei Variante 1 ergibt sich damit folgende Erfolgswahrscheinlichkeit:

$$\underbrace{\frac{1}{3}\cdot\frac{1}{2}\cdot 1}_{ana}+\underbrace{\frac{1}{3}\cdot\frac{1}{2}\cdot 1}_{aza}=\frac{1}{3}.$$

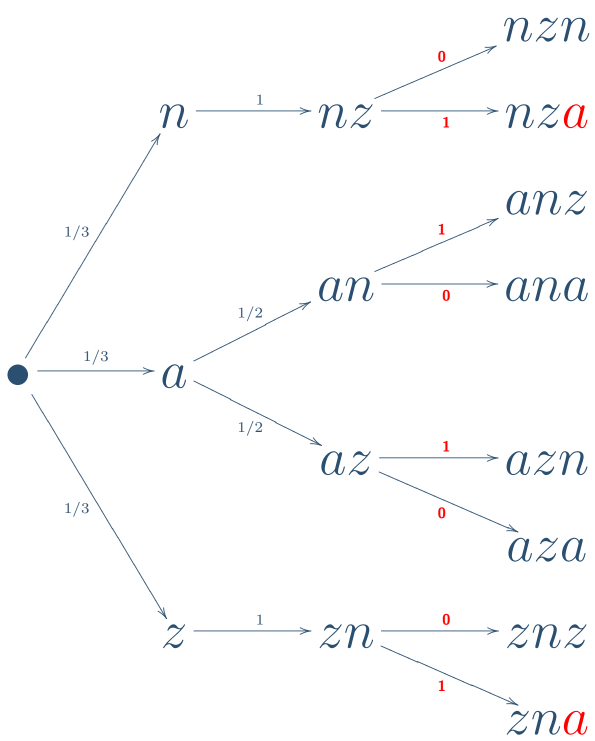
Ziegenproblem: Variante 2



Bei Variante 2 ergibt sich damit folgende Erfolgswahrscheinlichkeit:

$$\underbrace{\frac{1}{3}\cdot 1\cdot\frac{1}{2}}_{nz\color{red}{a}}+\underbrace{\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{1}{2}}_{an\color{red}{a}}+\underbrace{\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{1}{2}}_{az\color{red}{a}}+\underbrace{\frac{1}{3}\cdot 1\cdot\frac{1}{2}}_{zn\color{red}{a}}=\frac{1}{2}.$$

Ziegenproblem: Variante 3



Bei Variante 3 ergibt sich damit folgende Erfolgswahrscheinlichkeit:

$$\underbrace{\frac{1}{3}\cdot 1\cdot 1}_{nza} + \underbrace{\frac{1}{3}\cdot 1\cdot 1}_{zna} = \frac{2}{3}.$$

Ziegenproblem: Kurzversion der Lösungen

Variante 1: Da die Kandidatin bei ihrer ersten Wahl bleibt, ist das Verhalten des Moderators irrelevant. Wegen der Gleichverteilung ist $P(a) = 1/3$, $P(n) = 1/3$ und $P(z) = 1/3$.

Erfolgswahrscheinlichkeit ist daher $1/3$.

Variante 2: Da die Kandidatin im letzten Schritt gemäß Gleichverteilung auf $\{a, n\}$ oder $\{a, z\}$ entscheidet,

ergibt sich hier als Erfolgswahrscheinlichkeit $1/2$.

Variante 3: Da sich die Kandidatin im letzten Schritt immer umentscheidet, wird aus der Anfangswahl a eine Niete, aber aus n ein a sowie aus z ein a.

Die Erfolgswahrscheinlichkeit beträgt daher $2/3$.

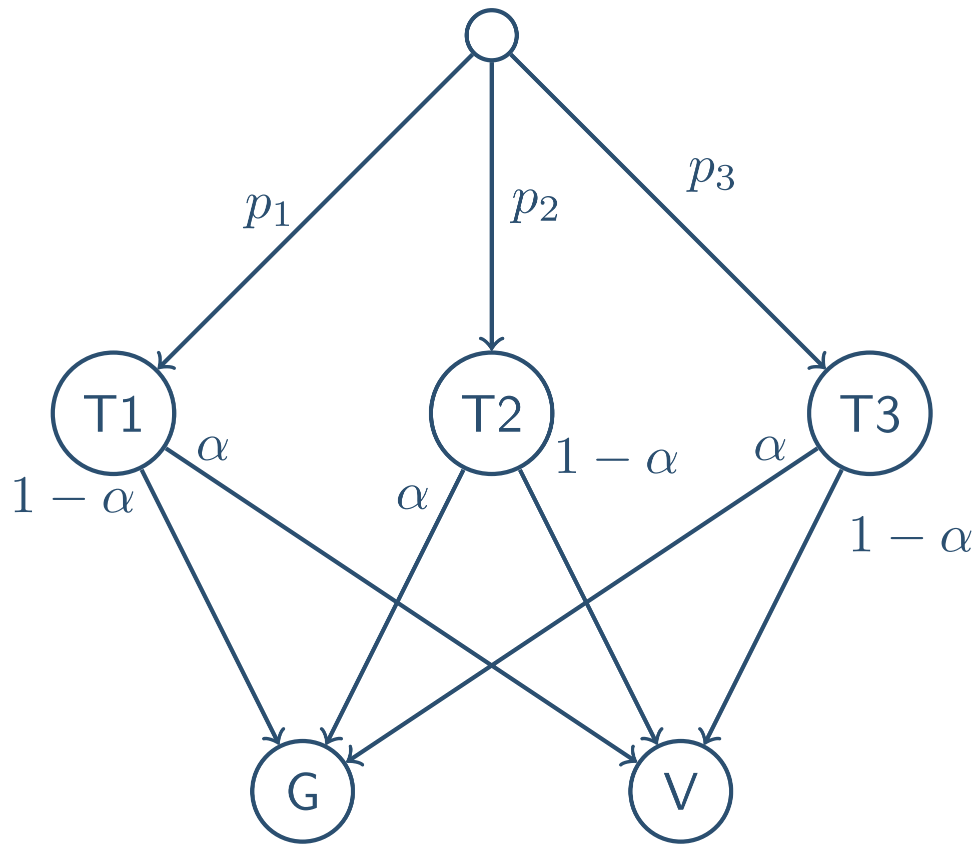
Beispiel zum Nacharbeiten(1)

Einem Kandidaten einer Spielshow werden drei Türen präsentiert, hinter denen genau ein Gewinn versteckt ist. Im Unterschied zum Ziegenproblem sind die Wahrscheinlichkeiten den Gewinn hinter den jeweiligen Türen zu finden nicht notwendig gleich und durch $p_1 \geq p_2 \geq p_3 \geq 0$ mit $p_1 + p_2 + p_3 = 1$ gegeben. Im Wissen um die Wahrscheinlichkeiten entscheidet sich der Kandidat für die erste Tür. Daraufhin öffnet der Moderator eine der verbleibenden Türen, hinter denen der Gewinn nicht versteckt ist. Nun hat der Kandidat die Möglichkeit seine Wahl auf die noch verschlossene Tür zu ändern, welche er mit Wahrscheinlichkeit α wahrnimmt. Uns interessiert, ob er schließlich gewinnt.

1. Modelliere dieses Experiment im Wegemodell. Unterscheide hierbei in der ersten Stufe die Fälle $\{T1, T2, T3\}$ die angeben hinter welcher Tür der Gewinn war und in der zweiten Stufe die Fälle $\{G, V\}$, die für Gewonnen und Verloren stehen. Beachte, dass die Wahrscheinlichkeit zu gewinnen von der Wahrscheinlichkeit α abhängt, die angibt ob der Kandidat wechselt.
2. Gib die Wahrscheinlichkeit, dass der Kandidat gewinnt, in Abhängigkeit von p_1, p_2, p_3 und α an.
3. Ist der Kandidat für $p_1 = \frac{1}{2}$ besser beraten zu wechseln, zu verbleiben oder ist es irrelevant?
4. Sei $p_1 = 0.4$, $p_2 = p_3 = 0.3$ und $\alpha = 0.5$. Falls der Kandidat gewonnen hat, wie groß ist die Wahrscheinlichkeit, dass der Gewinn hinter der ersten Tür war?

Beispiel zum Nacharbeiten(2)

1.



2. Für die Wahrscheinlichkeit zu gewinnen summieren wir die möglichen Pfade:

$$P(G) = p_1(1-\alpha) + p_2\alpha + p_3\alpha = p_1 + \alpha(p_2 + p_3 - p_1) = p_1 + \alpha(1 - 2p_1)$$

Beispiel zum Nacharbeiten(3)

3. Für $p_1 = 0.5$ ist die Wahrscheinlichkeit zu gewinnen unabhängig von α. Deshalb ist es irrelevant ob er wechselt.
4. $$P(T1|G) = \frac{P(G|T1)\,P(T1)}{P(G)} = \frac{p_1(1-\alpha)}{p_1 + \alpha(1-2p_1)} = \frac{2}{5}$$

Stochastische Unabhängigkeit

Dies ist ein weiterer fundamentaler Begriff der Stochastik.

- **Intuitiv**: Zwei Ereignisse A und B sind stochastisch unabhängig, wenn die Wahrscheinlichkeit des Eintretens von A nicht beeinflusst wird durch die Information, dass B eingetreten ist, und umgekehrt.
- Das bedeutet:

$$P(A|B) = P(A) \quad \text{und} \quad P(B|A) = P(B), \quad \text{falls } P(A), P(B) > 0.$$

- Mit $P(A|B) = P(A \cap B)/P(B)$ und $P(B|A) = P(A \cap B)/P(A)$ ergibt das sofort $P(A \cap B) = P(A)P(B)$.
- Diese letzte Formel kommt vorteilhafterweise ohne die Positivitätsbedingung $P(A), P(B) > 0$ aus. Daher vereinbart man:

Definition. Es sei $(\Omega, \mathcal{A}, P)$ ein W-Raum. Ereignisse $A, B \in \mathcal{A}$ heißen (stochastisch) **unabhängig** bezüglich P, wenn

$$P(A \cap B) = P(A)P(B).$$

Beispiele zur Unabhängigkeit

Zufallsexperiment: Ziehen von zwei Kugeln aus einer Urne mit $N = b + r$ Kugeln, b blaue und r rote. Es sei $0 < b < N$.
Wir betrachten folgende Ereignisse:

$$\begin{aligned} A &= \{\text{die erste Kugel ist blau}\} \\ B &= \{\text{die zweite Kugel ist blau}\} \end{aligned}$$

Beispiel 1. Ziehen **mit** Zurücklegen und mit Berücksichtigung der Reihenfolge. W-Raum: $[1:N]^2$ mit Gleichverteilung P. Hier ist

$$P(A \cap B) = \frac{b^2}{N^2} = P(A)P(B),$$

also sind A und B unabhängig bzgl. P.

Beispiel 2. Ziehen **ohne** Zurücklegen aber mit Berücksichtigung der Reihenfolge. W-Raum: $\{(i,j) \in [1:N]^2 | i \neq j\}$ mit Gleichverteilung P'. A und B sind nicht unabhängig bzgl. P', denn wegen $b < N$ ist

$$P'(A \cap B) = \frac{b(b-1)}{N(N-1)} < \frac{b^2}{N^2} = P'(A)P'(B).$$

Unabhängigkeit: Verallgemeinerungen

Definition. (Unabhängigkeit von Ereignisfamilien)
Es sei $(\Omega, \mathcal{A}, P)$ ein W-Raum. Eine Familie $(A_i)_{i\in I}$ von Ereignissen in $\mathcal{A}$ heißt (stochastisch) **unabhängig** bezüglich P, wenn für jede endliche Teilmenge $\emptyset \neq J \subseteq I$ gilt:

$$P\left(\bigcap_{j\in J} A_j\right) = \prod_{j\in J} P(A_j).$$

Bei der nächsten Verallgemeinerung geht es um die Unabhängigkeit von Teilexperimenten.

Definition. (Unabhängigkeit von ZV-Familien)
Es sei $(\Omega, \mathcal{A}, P)$ ein W-Raum. Eine Familie $(Y_i)_{i\in I}$ von ZVs $Y_i : (\Omega, \mathcal{A}) \to (\Omega_i, \mathcal{A}_i)$ heißt **unabhängig** bezüglich P, wenn für jede endliche Teilmenge $\emptyset \neq J \subseteq I$ und alle $A_j \in \mathcal{A}_j$ mit $j \in J$ gilt:

$$P\left(\bigcap_{j\in J} \{Y_j \in A_j\}\right) = \prod_{j\in J} P(Y_j \in A_j).$$

Unabhängigkeitskriterien

Satz. Es sei $(\Omega, \mathcal{A}, P)$ ein W-Raum und $Y_1, \ldots, Y_n$ eine endliche Familie von ZVs $Y_i : (\Omega, \mathcal{A}) \to (\Omega_i, \mathcal{A}_i)$. Dann gilt:

(a) **Diskreter Fall**: Hat jedes Y_i einen endlichen Wertebereich Ω_i und ist $\mathcal{A}_i = 2^{\Omega_i}$, so ist $(Y_i)_{i \in [1:n]}$ genau dann unabhängig, wenn für beliebige $\omega_i \in \Omega_i$ gilt:

$$P(Y_1 = \omega_1, \ldots, Y_n = \omega_n) = \prod_{i=1}^{n} P(Y_i = \omega_i).$$

(b) **Reeller Fall**: Ist jedes Y_i reellwertig und ist $\mathcal{A}_i = \mathcal{B}^1_{\Omega_i}$, so ist $(Y_i)_{i \in [1:n]}$ genau dann unabhängig, wenn für beliebige $c_i \in \mathbb{R}$ gilt:

$$P(Y_1 \leq c_1, \ldots, Y_n \leq c_n) = \prod_{i=1}^{n} P(Y_i \leq c_i).$$

Beweis: Übung! $\square$

Produktmaße und Unabhängigkeit

Hier geht es um die Unabhängigkeit von ZVs im diskreten Fall.

Es sei p eine W-Funktion auf der endlichen Menge E. Für festes $n \geq 2$ bezeichne $P = p^{\otimes n}$ das n-fache Produktmaß auf $\Omega = E^n$. Es bezeichne $Y_i : \Omega \to E$ die i-te Projektion. Für $\omega \in \Omega$ ist

$$\{(\omega_1, \dots, \omega_n)\} = A_1 \cap \dots \cap A_n \quad \text{mit} \quad A_i = E^{i-1} \times \{\omega_i\} \times E^{n-i}.$$

Zusammen mit der Definition von Produktmaßen folgt damit

$$\begin{aligned}
\prod_{i=1}^{n} P(Y_i = \omega_i) &= \prod_{i=1}^{n} P(A_i) = \prod_{i=1}^{n} p(\omega_i) \sum_{\omega' \in A_i} \prod_{j \neq i} p(\omega'_j) \\
&= \prod_{i=1}^{n} p(\omega_i) \prod_{j \neq i} \underbrace{\Big(\sum_{\omega'_j \in E} p(\omega'_j) \Big)}_{=1} = P(\{(\omega_1, \dots, \omega_n)\}) \\
&= P(Y_1 = \omega_1, \dots, Y_n = \omega_n).
\end{aligned}$$

Also sind die Projektionen $Y_1, \dots, Y_n$ (wie erwartet) unabhängig.

Unabhängigkeit im Reellen

Hier geht es um die Polarkoordinaten eines zufälligen Punktes der Kreisscheibe. Es sei $(\Omega, \mathcal{A}, P)$ ein W-Raum und

$$K := \{x \in \mathbb{R}^2 \mid x_1^2 + x_2^2 \leq 1\}$$

die Einheitskreisscheibe. Es sei $Z = (Z_1, Z_2) : (\Omega, \mathcal{A}) \to (K, \mathcal{B}_K^2)$ eine ZV mit Gleichverteilung $\mathcal{U}_K$ auf K, Z_i sei i-te Projektion. Wir definieren neue ZVs $R : (\Omega, \mathcal{A}) \to ([0,1], \mathcal{B}_{[0,1]})$ und $\Psi : (\Omega, \mathcal{A}) \to ([0, 2\pi), \mathcal{B}_{[0,2\pi)})$ durch Übergang zu Polarkoordinaten:

$$R := |Z| = \sqrt{Z_1^2 + Z_2^2} \quad \text{und} \quad \Psi := \arg(Z_1 + iZ_2).$$

Dann ist für $r \in [0,1]$ und $\psi \in [0, 2\pi)$ (günstige Fläche/mögliche Fläche):

$$P(R \leq r, \Psi \leq \psi) = \frac{\pi r^2 \cdot \frac{\psi}{2\pi}}{\pi} = r^2 \cdot \frac{\psi}{2\pi} = P(R \leq r) \cdot P(\Psi \leq \psi).$$

Nach dem Unabhängigkeitskriterium sind R und Ψ unabhängig und zudem gleichverteilt auf den jeweiligen Bildmessräumen.

Stochastik

Kapitel 6: Erwartungswert und Varianz

Prof. Dr. Reinhard Klein
Folien von Prof. Dr. Michael Clausen

Sommersemester 2017

INSTITUT FÜR INFORMATIK II, UNIVERSITÄT BONN

Kenngrößen reellwertiger Zufallsvariablen

Eine reellwertige ZV X wird durch zwei fundamentale Größen grob gekennzeichnet:

- der **Erwartungswert** $\mathbf{E}_P(X)$ gibt den (P-gewichteten) mittleren Wert von X an, und
- die **Varianz** $\mathbf{V}_P(X)$ misst, wie stark die Werte von X typischerweise vom Erwartungswert abweichen.

- **Warnung**:

Nicht jede reellwertige ZV X besitzt einen Erwartungswert und eine Varianz!

Erwartungswert: diskreter Fall

Im folgenden sei $(\Omega, \mathcal{A}, P)$ W-Raum, $X : \Omega \to \mathbb{R}$ eine **diskrete** ZV, d.h. das Bild $X[\Omega] := \{X(\omega) \mid \omega \in \Omega\}$ von X ist höchstens abzählbar unendlich.

Definition. X **besitzt** einen Erwartungswert, wenn

$$\sum_{x \in X[\Omega]} |x| P(X = x) < \infty.$$

In diesem Fall ist nach dem Umordnungssatz die Summe

$$\mathbf{E}(X) = \mathbf{E}_P(X) := \sum_{x \in X[\Omega]} x P(X = x)$$

wohldefiniert und heißt der **Erwartungswert** von X bzgl. P.

$$\mathscr{L}^1(P) := \{X : \Omega \to \mathbb{R} \mid \mathbf{E}_P(|X|) < \infty\} = \mathscr{L}^1$$

bezeichnet die Menge aller ZVs X, für die der Erwartungswert bzgl. P existiert.

Erwartungswert als Schwerpunkt

- Der Erwartunsgwert besitzt im Fall einer diskreten ZV X folgende physikalische Interpretation:
- Wenn man $P \circ X^{-1}$ als diskrete Massenverteilung (mit Gesamtmasse 1) auf der (gewichtslosen) reellen Achse auffasst, so ist $\mathbf{E}_P(X)$ gerade der Schwerpunkt der Massenverteilung.
- Illustration in der Vorlesung.

Beispiele

Indikatorfunktion: Es sei $(\Omega, \mathcal{A}, P)$ W-Raum. Die Indikatorfunktion 1_A zu $A \in \mathcal{A}$ ist konstant 1 auf A und konstant 0 auf $\Omega \setminus A$. Es gilt:

$$\mathbf{E}_P(1_A) = 0 \cdot P(1_A = 0) + 1 \cdot P(1_A = 1) = P(A).$$

Dies verbindet die Begriffe Erwartungswert und Wahrscheinlichkeit.

Beispiel. Ist $c \in \mathbb{R}$, so hat die konstante Funktion $c1_\Omega := (\Omega \ni \omega \mapsto c)$ den Erwartungswert c: denn $\mathbf{E}_P(c1_\Omega) = cP(\Omega) = c \cdot 1 = c$.

Abzählbarer Definitionsbereich Ω: Dann ist jede ZV X diskret und es gilt:

$$\sum_{\omega \in \Omega} |X(\omega)| P(\omega) = \sum_{x \in X[\Omega]} |x| \sum_{\omega \in \{X = x\}} P(\omega) = \sum_{x \in X[\Omega]} |x| P(X = x).$$

Also ist $X \in \mathscr{L}^1(P)$, gdw. $\sum_{\omega \in \Omega} |X(\omega)| P(\omega) < \infty$. In dem Fall erhält man:

$$\mathbf{E}_P(X) = \sum_{\omega \in \Omega} X(\omega) P(\omega).$$

Rechenregeln für Erwartungswerte

Satz. Für diskrete ZVs $X, Y, X_n, Y_n : \Omega \to \mathbb{R}$ in $\mathscr{L}^1(P)$ gilt:

(1) **Monotonie**: Aus $X \leq Y$, d.h. $X(\omega) \leq Y(\omega)$ für alle $\omega \in \Omega$, folgt $\mathbf{E}_P(X) \leq \mathbf{E}_P(Y)$.

(2) **Linearität**: $\mathscr{L}^1(P)$ ist ein reeller Vektorraum und $\mathbf{E}_P : \mathscr{L}^1(P) \to \mathbb{R}$ ist linear:
$\mathbf{E}_P(c_1 X_1 + c_2 X_2) = c_1 \mathbf{E}_P(X_1) + c_2 \mathbf{E}_P(X_2)$ für $c_1, c_2 \in \mathbb{R}$.

(3) σ-**Additivität**: Sind alle $X_n \geq 0$ und ist $X = \sum_{n \geq 1} X_n$, so gilt
$\mathbf{E}_P(X) = \sum_{n \geq 1} \mathbf{E}_P(X_n)$.

(4) **Monotone Konvergenz**: Wenn $Y_n \uparrow Y$ für $n \uparrow \infty$, so folgt $\mathbf{E}_P(Y) = \lim_{n \to \infty} \mathbf{E}_P(Y_n)$.

(5) **Produktregel**: Sind X und Y unabhängig, so ist $X \cdot Y \in \mathscr{L}^1(P)$ und es gilt:
$\mathbf{E}_P(X \cdot Y) = \mathbf{E}_P(X) \mathbf{E}_P(Y)$.

Beweis der Monotonieregel

Aus der Definition des Erwartungswerts und der σ-Additivität von P ergibt sich (Begründung folgt unten):

$$\begin{aligned}
\mathbf{E}_P(X) &= \sum_{x\in X[\Omega]} xP(X=x) \\
&= \sum_{x\in X[\Omega],y\in Y[\Omega]} xP(X=x,Y=y) \\
&\leq \sum_{x\in X[\Omega],y\in Y[\Omega]} yP(X=x,Y=y) = \mathbf{E}_P(Y).
\end{aligned}$$

Man beachte, dass die Summationsreihenfolge wegen der absoluten Konvergenz der Reihe irrelevant ist. Da nach Voraussetzung $P(X=x,Y=y)=0$ außer wenn $x\leq y$, gilt die Ungleichung, womit die Monotonieregel bewiesen ist.

Beweis der Linearität (1)

- $X \in \mathscr{L}^1(P)$, $c \in \mathbb{R}$ impliziert $cX \in \mathscr{L}^1(P)$: klar.
- O.E. sei $c \neq 0$. Dann gilt:

$$\mathbf{E}_P(cX) = \sum_{x \in X[\Omega]} cxP(cX = cx) = c \sum_{x \in X[\Omega]} xP(X = x) = c\mathbf{E}_P(X).$$

- $X, Y \in \mathscr{L}^1(P)$ impliziert $X + Y \in \mathscr{L}^1(P)$:
  - $X + Y$ diskret, denn mit $X[\Omega]$ und $Y[\Omega]$ ist nach dem Cantorschen Diagonalverfahren auch $X[\Omega] \times Y[\Omega]$ abzählbar.
  - Zeigen als nächstes $\mathbf{E}_P(|X + Y|) < \infty$. Dabei benutzen wir die disjunkten Zerlegungen

$$\Omega = \bigsqcup_{x \in X[\Omega]} \{X = x\} \quad \text{sowie} \quad \Omega = \bigsqcup_{y \in Y[\Omega]} \{Y = y\}. \qquad (*)$$

Beweis der Linearität (2)

$$
\begin{aligned}
\mathbf{E}_P(|X+Y|) &= \sum_{z\in(X+Y)[\Omega]} |z|P(X+Y=z) \\
&\leq \sum_{z\in(X+Y)[\Omega]} \sum_{x\in X[\Omega]} (|x|+|z-x|)P(X=x, Y=z-x) \\
&\leq \sum_{x\in X[\Omega]} \sum_{y\in Y[\Omega]} (|x|+|y|)P(X=x, Y=y) \\
&= \sum_{x\in X[\Omega]} |x| \sum_{y\in Y[\Omega]} P(X=x, Y=y) + \sum_{y\in Y[\Omega]} |y| \sum_{x\in X[\Omega]} P(X=x, Y=y) \\
&\overset{(*)}{=} \underbrace{\sum_{x\in X[\Omega]} |x|P(X=x)}_{=\mathbf{E}_P(|X|)<\infty} + \underbrace{\sum_{y\in Y[\Omega]} |y|P(Y=y)}_{=\mathbf{E}_P(|Y|)<\infty} < \infty.
\end{aligned}
$$

Also ist auch $X+Y \in \mathscr{L}^1(P)$. Eine ähnliche Rechnung ohne Betragsstriche ergibt die Additivität des Erwartungswertoperators.

Beweis der Produktregel

- XY ist diskrete ZV:

denn mit $X[\Omega]$ und $Y[\Omega]$ ist auch $X[\Omega] \times Y[\Omega]$ abzählbar also auch das Bild von $\omega \mapsto X(\omega)Y(\omega)$.

- XY liegt in $\mathscr{L}^1(P)$:

denn

$$
\begin{aligned}
\sum_z \quad |z| \quad P(XY = z) &= \sum_{z \neq 0} |z| \sum_{x \neq 0} P(X = x, Y = z/x) \\
&= \sum_{x \neq 0, y \neq 0} |x||y| P(X = x) P(Y = y) \quad \text{da } X \text{ u. } Y \text{ unabhängig} \\
&= \sum_{x \neq 0} |x| P(X = x) \sum_{y \neq 0} |y| P(Y = y) < \infty,
\end{aligned}
$$

da $X, Y \in \mathscr{L}^1(P)$. Also liegt auch XY in $\mathscr{L}^1(P)$. Die entsprechende Rechnung ohne Betragsstriche liefert die Behauptung.

Die Beweise der restlichen Behauptungen werden als Übung empfohlen. $\square$

Beispiel zum Nacharbeiten

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und seien $X, Y : \Omega \to \mathbb{Z}$ zwei Zufallsvariablen mit $X, Y \in \mathcal{L}^1(P)$. Beweise, dass gilt: $\mathbf{E}_P(X+Y) = \mathbf{E}_P(X) + \mathbf{E}_P(Y)$.

Lösung:

$$\begin{aligned}
\mathbf{E}_P(X+Y) &= \sum_{x,y\in\mathbb{Z}} (x+y)P(X=x, Y=y) \\
&= \sum_{x,y\in\mathbb{Z}} xP(X=x, Y=y) + \sum_{x,y\in\mathbb{Z}} yP(X=x, Y=y) \\
&= \sum_{x\in\mathbb{Z}} xP(X=x) + \sum_{y\in\mathbb{Z}} yP(Y=y) \\
&= \mathbf{E}_P(X) + \mathbf{E}_P(Y)
\end{aligned}$$

Erwartete Erfolgsanzahl bei Bernoulli-Experimenten

- Sei $X_1, \ldots, X_n$ eine endliche Bernoulli-Folge zur Erfolgswahrscheinlichkeit p.
- Dann beschreibt die ZV

$$S := \sum_{i=1}^{n} X_i$$

  die Anzahl der Erfolge.
- Mit Hilfe der Linearität des Erwartungswertoperators ergibt sich für die zu erwartende Erfolgsanzahl:

$$\mathbf{E}_P(S) = \sum_{i=1}^{n} \mathbf{E}_P(X_i) = \sum_{i=1}^{n} P(X_i = 1) = np.$$

Erwartungswert: allgemein

Wir gehen jetzt auf den Fall einer beliebigen, nicht notwendig diskreten, reellwertigen ZV X ein. In diesem Fall kann der Erwartungswert von X i.a. nicht mehr über eine Summe definiert werden, sondern ist nur noch Limes von Erwartungswerten approximativer Varianten $X_{(n)}$ von X. Diese Varianten definieren wir zunächst.

Definition. Zur ZV $X : \Omega \to \mathbb{R}$ und $n \in \mathbb{N}$ wird die $1/n$-**Diskretisierung** von X definiert durch

$$X_{(n)}(\omega) := \frac{k}{n}, \quad \text{wobei } k \in \mathbb{Z} \text{ und } \quad \frac{k}{n} \leq X(\omega) < \frac{k+1}{n}.$$

Anschaulich: Die reelle Achse wird partitioniert:

$$\mathbb{R} = \bigsqcup_{k \in \mathbb{Z}} [\frac{k}{n}, \frac{k+1}{n})$$

und die X-Werte im Intervall $[\frac{k}{n}, \frac{k+1}{n})$ werden abgerundet zum Wert $\frac{k}{n}$.

Diskrete Approximation des Erwartungswertes

Satz. Für die $1/n$-Diskretisierungen $X_{(n)}$ von X gilt:

(1) Für alle $n \in \mathbb{N}$ ist $X_{(n)} \leq X < X_{(n)} + \frac{1}{n}$.
(2) Ist $X_{(m)} \in \mathscr{L}^1(P)$ für ein m, so ist $X_{(n)} \in \mathscr{L}^1(P)$ für alle n und in diesem Fall ist $(\mathbf{E}_P(X_{(n)}))_{n \geq 1}$ eine Cauchy-Folge.

Beweis. (1) ist klar. Hieraus folgt (Übung!)

$$|X_{(n)}| < |X_{(m)}| + \max(\frac{1}{m}, \frac{1}{n}).$$

Wenn also $X_{(m)} \in \mathscr{L}^1(P)$ für ein m, dann ist wegen der Monotonieregel sogar $X_{(n)} \in \mathscr{L}^1(P)$ für alle n. Weiter ergibt sich $\mathbf{E}_P(X_{(m)}) \leq \mathbf{E}_P(X_{(n)}) + \frac{1}{n}$. Vertauscht man in letzter Formel n und m, ergibt sich insgesamt:

$$|\mathbf{E}_P(X_{(m)}) - \mathbf{E}_P(X_{(n)})| \leq \max(\frac{1}{m}, \frac{1}{n}).$$

Das zeigt (2). □

Erwartungswert: allgemeine Definition

Definition. Die ZV $X : \Omega \to \mathbb{R}$ **besitzt** einen Erwartungswert, kurz: $X \in \mathscr{L}^1(P)$, wenn $X_{(n)} \in \mathscr{L}^1(P)$ für ein n (und damit für alle n) gilt. In dem Fall heißt

$$\mathbf{E}_P(X) := \lim_{n\to\infty} \mathbf{E}_P(X_{(n)})$$

der **Erwartungswert** von X bzgl. P.

Aus Sicht der Integrationstheorie ist $\mathbf{E}_P(X)$ das Integral von X bzgl. P:

$$\mathbf{E}_P(X) = \int X\, dP.$$

Satz. Auch für den allgemeinen Fall gelten die obigen Rechenregeln.

Beweis. Übung! $\square$

Erwartungswert bei W-Maßen mit Dichtefunktion

Satz.

- Es sei $\Omega \subset \mathbb{R}^d$ Borelsch und P das W-Maß auf $(\Omega, \mathcal{B}_\Omega^d)$ zur Dichtefunktion ρ, d.h. $P(A) = \int_A \rho(\omega)d\omega$ für alle $A \in \mathcal{B}_\Omega^d$.
- Weiter sei X eine reelle ZV auf Ω. Dann gilt $X \in \mathscr{L}^1(P)$ gdw. $\int_\Omega |X(\omega)|\rho(\omega)d\omega < \infty$.
- Im Fall $X \in \mathscr{L}^1(P)$ ist
$$\mathbf{E}_P(X) = \int_\Omega X(\omega)\rho(\omega)d\omega.$$
- Im Spezialfall $\Omega = \mathbb{R}$ und $X = \mathrm{id}_\mathbb{R}$ ist
$$\mathbf{E}_P(\mathrm{id}_\mathbb{R}) = \int_\mathbb{R} x\rho(x)dx.$$

Dieser Wert, der nur von P abhängt, wird auch **Erwartungswert des W-Maßes** P genannt und mit $\mathbf{E}(P)$ abgekürzt.
[Analog definiert man später die **Varianz** $\mathbf{V}(P)$ von P.]

Höhere Momente von Zufallsvariablen (1)

- Es sei $(\Omega, \mathcal{A}, P)$ W-Raum und $X : \Omega \to \mathbb{R}$ reelle ZV und $r \in \mathbb{N}$.
- Ist $X^r \in \mathscr{L}^1(P)$, wobei $X^r(\omega) := X(\omega)^r$, so nennt man $\mathbf{E}_P(X^r)$ das **r-te Moment** von X.
- Die Gesamtheit aller ZVs, für die das r-te Moment existiert, fasst man zu einer Menge zusammen:

$$\mathscr{L}^r(P) := \{X : \Omega \to \mathbb{R} \mid X^r \in \mathscr{L}^1(P)\}.$$

Satz. Es sei $(\Omega, \mathcal{A}, P)$ W-Raum. Dann gilt:

(1) Für $r \in \mathbb{N}$ ist $\mathscr{L}^r(P)$ ein reeller Untervektorraum von $\mathscr{L}^1(P)$.

(2) Die Räume $\mathscr{L}^r(P)$ bilden eine Kette:

$$\mathscr{L}^1(P) \supseteq \mathscr{L}^2(P) \supseteq \mathscr{L}^3(P) \supseteq \ldots$$

Bemerkung: Uns interessiert besonders der Fall $r = 2$.

Höhere Momente von Zufallsvariablen (2)

Beweis. **(1)** Seien $X, Y \in \mathscr{L}^r(P)$ und $a \in \mathbb{R}$. Dann liegt aX in $\mathscr{L}^r(P)$, denn $a^r X^r \in \mathscr{L}^1(P)$. Weiter ist

$$\begin{aligned} |X+Y|^r &\leq (|X|+|Y|)^r \leq (2\max\{|X|,|Y|\})^r \leq 2^r \max\{|X|^r,|Y|^r\} \\ &\leq 2^r(|X|^r+|Y|^r). \end{aligned}$$

Daraus ergibt sich $X+Y \in \mathscr{L}^r(P)$, denn

$$\int |X+Y|^r dP \leq 2^r(\underbrace{\int |X|^r dP}_{<\infty} + \underbrace{\int |Y|^r dP}_{<\infty}) < \infty.$$

Also ist $\mathscr{L}^r(P)$ ein $\mathbb{R}$-Vektorraum.

(2) Für $r < s$ ist $|X|^r \leq 1 + |X|^s$: Für $|X| \geq 1$ ist dies trivial. Im Fall $|X| < 1$ folgt die Ungleichung aus $|X|^r < 1$ und $|X|^s < 1$. Ist $X \in \mathscr{L}^s(P)$ und $r < s$, so ist $X \in \mathscr{L}^r(P)$:

$$\int |X|^r dP \leq \int (1+|X|^s) dP = \underbrace{\int 1 dP}_{=1} + \underbrace{\int |X|^s dP}_{<\infty} < \infty. \quad \square$$

Varianz und Kovarianz

Definition. Für $X, Y \in \mathscr{L}^2(P)$ heißt

(1) $\mathbf{V}(X) = \mathbf{V}_P(X) := \mathbf{E}_P([X - \mathbf{E}_P(X)]^2) = \mathbf{E}_P(X^2) - \mathbf{E}_P(X)^2$ die **Varianz** von X bzgl. P,

(2) $\sqrt{\mathbf{V}(X)}$ die **Standardabweichung** von X bzgl. P, sowie

(3) $\mathrm{Cov}_P(X, Y) := \mathbf{E}_P([X - \mathbf{E}_P(X)][Y - \mathbf{E}_P(Y)]) = \mathbf{E}_P(XY) - \mathbf{E}_P(X)\mathbf{E}_P(Y)$ die **Kovarianz** von X und Y bzgl. P.

(4) Ist $\mathrm{Cov}_P(X, Y) = 0$, so heißen X und Y **unkorreliert**.

Bemerkung. Wegen $|XY| \leq X^2 + Y^2$ gilt

$$\int |XY| dP \leq \int (X^2 + Y^2) dP = \underbrace{\int X^2 dP}_{<\infty} + \underbrace{\int Y^2 dP}_{<\infty} < \infty.$$

Aus $X, Y \in \mathscr{L}^2(P)$ folgt also $X \cdot Y \in \mathscr{L}^1(P)$. Daher ist die Kovarianz wohldefiniert.

Beispiel zum Nacharbeiten

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum und seien $X, Y : \Omega \to \mathbb{Z}$ zwei Zufallsvariablen mit $X, Y \in \mathcal{L}^1(P)$. Beweise, dass gilt $\mathbf{E}_P([X - \mathbf{E}_P(X)]^2) = \mathbf{E}_P(X^2) - \mathbf{E}_P(X)^2$.

Lösung: Die Äquivalenz folgt direkt aus der Linearität des Erwartungswerts:

$$
\begin{aligned}
\mathbf{E}_P([X - \mathbf{E}_P(X)]^2) &= \mathbf{E}_P(X^2 - 2\mathbf{E}_P(X)X + \mathbf{E}_P(X)^2) \\
&= \mathbf{E}_P(X^2) - 2\mathbf{E}_P(X)\mathbf{E}_P(X) + \mathbf{E}_P(X)^2 \\
&= \mathbf{E}_P(X^2) - \mathbf{E}_P(X)^2
\end{aligned}
$$

Varianz und Kovarianz: Bemerkungen

Die Varianz misst, wie weit die Werte von X im Schnitt auseinanderliegen:

Beispiel. Ist X gleichverteilt auf $\{x_1, \dots, x_n\} \subset \mathbb{R}$, so ist

$$\mathbf{E}(X) = \overline{x} := \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{und} \quad \mathbf{V}(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2,$$

d.h. die Varianz ist hier gerade die **mittlere quadratische Abweichung vom Mittelwert**.

Ist zudem Y gleichverteilt auf $\{y_1, \dots, y_n\} \subset \mathbb{R}$, so ist $n\mathrm{Cov}(X, Y)$ gerade das euklidische Skalarprodukt der zentrierten Vektoren $(x_i - \overline{x})_{i \in [1:n]}$ und $(y_i - \overline{y})_{i \in [1:n]}$.

Der sog. **Korrelationskoeffizient**

$$\rho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathbf{V}(X)\mathbf{V}(Y)}}$$

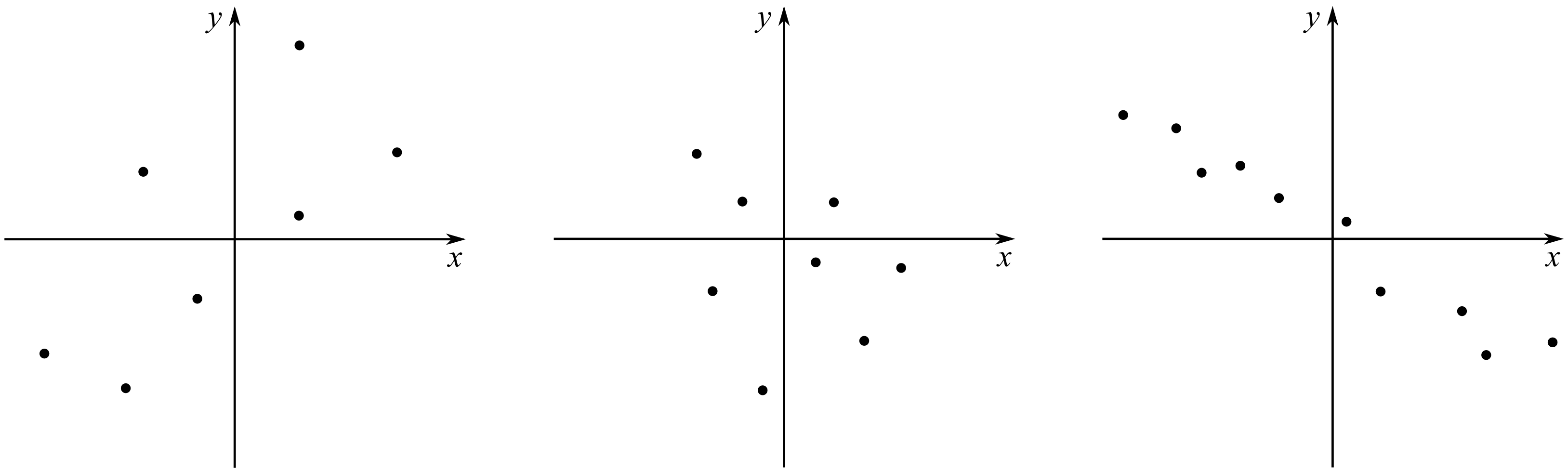
entspricht dem Cosinus des Winkels zwischen diesen Vektoren.

Die Unkorreliertheit ist gerade die Orthogonalität.

Korrelationskoeffizienten: Interpretation

Für ZVs $X : [1 : n] \to \mathbb{R}$ und $Y : [1 : n] \to \mathbb{R}$ bedeutet

- große positive Korrelation ($\rho(X, Y) \approx 1$) (bzw. große negative Korrelation ($\rho(X, Y) \approx -1$)), dass die Punktepaare $(X(k), Y(k))$ ziemlich gut auf einer Geraden mit positiver (bzw. negativer) Steigung liegen.
- Ein Wert $\rho(X, Y) \approx 0$ besagt, dass kein geradenartiger Trend vorliegt.
- Das folgende Schaubild zeigt links den Fall $\rho(X, Y) > 0$, rechts den Fall $\rho(X, Y) \approx -1$ und mittig den Fall $\rho(X, Y) \approx 0$.



Beispiel zum Nacharbeiten

Es seien X und Y Zufallsvariablen mit $Y = -aX + b$ für eine positive reelle Zahl $a > 0$, eine reelle Zahl b und $\mathbf{V}(X) \neq 0$. Zeige unter Verwendung der Rechenregeln für **E**, **V** und Cov, dass für den Korrelationskoeffizienten ρ_{XY} gilt: $\rho_{XY} = -1$.
Wie kann diese Aussage geometrisch gedeutet werden?
Lösung:

$$\begin{aligned}
\mathbf{V}(Y) &= \mathbf{V}(-aX+b) = (-a)^2\mathbf{V}(X) = a^2\mathbf{V}(X) \\
\mathrm{Cov}(X,Y) &= \mathrm{Cov}(X,-aX+b) = -a\mathrm{Cov}(X,X) = -a\mathbf{V}(X) \\
\rho_{XY} &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathbf{V}(X)}\sqrt{\mathbf{V}(Y)}} = \frac{-a\mathbf{V}(X)}{a\mathbf{V}(X)} = -1
\end{aligned}$$

Geometrisch bedeutet diese Aussage: Für $\omega \in \Omega$ liegen die Punkte $(X(\omega), Y(\omega))$ alle auf einer Geraden mit negativer Steigung.

Varianz und Kovarianz: Rechenregeln

Satz. Seien $X, Y, X_i \in \mathscr{L}^2(P)$ und $a, b, c, d \in \mathbb{R}$. Dann gilt:

(1) $aX + b$, $cY + d$ liegen in $\mathscr{L}^2(P)$ und $\mathrm{Cov}_P(aX + b, cY + d) = ac\mathrm{Cov}_P(X, Y)$.
Insbesondere gilt:

$$\mathbf{V}_P(aX + b) = a^2\mathbf{V}_P(X).$$

(2) $\mathrm{Cov}_P(X, Y)^2 \leq \mathbf{V}_P(X)\mathbf{V}_P(Y)$ (**Cauchy-Schwarz-Ungleichung**).

(3) $\sum_{i=1}^n X_i \in \mathscr{L}^2(P)$ und

$$\mathbf{V}_P(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathbf{V}_P(X_i) + \sum_{j \neq i} \mathrm{Cov}_P(X_i, X_j).$$

Sind $X_1, \ldots, X_n$ paarweise unkorreliert, so gilt der **Satz des Pythagoras**:

$$\mathbf{V}_P(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathbf{V}_P(X_i).$$

(4) Sind X und Y unabhängig, so sind X und Y auch unkorreliert.
(Die Umkehrung gilt i.a. nicht!)

Kovarianzregeln: Beweis von (1)

Mit der Linearität des Erwartungswertoperators folgt:

$$
\begin{aligned}
& \mathrm{Cov}_P(aX+b, cY+d) \\
= \; & \mathbf{E}_P((aX+b)\cdot(cY+d)) - \mathbf{E}_P(aX+b)\mathbf{E}_P(cY+d) \\
= \; & ac\mathbf{E}_P(XY) + ad\mathbf{E}_P(X) + bc\mathbf{E}_P(Y) + bd \\
& -(a\mathbf{E}_P(X)+b)(c\mathbf{E}_P(Y)+d) \\
= \; & ac(\mathbf{E}_P(XY) - \mathbf{E}_P(X)\mathbf{E}_P(Y)) \\
= \; & ac\mathrm{Cov}_P(X,Y).
\end{aligned}
$$

Dies beweist (1). $\square$

Kovarianzregeln: Beweis von (2)

Cauchy-Schwarzsche Ungleichung: Für $X, Y \in \mathscr{L}^2(P)$ gilt:

$$\mathbf{E}_P(XY)^2 \leq \mathbf{E}_P(X^2) \cdot \mathbf{E}_P(Y^2).$$

Beweisskizze der CSU:

Sei $a := \mathbf{E}_P(Y^2)$ und $b := -\mathbf{E}_P(XY)$.
O.B.d.A. können wir $a > 0$ voraussetzen, da sonst beide Seiten der Ungleichung Null sind. (Dies folgt aus der Tatsache (Übung!), dass $\mathbf{E}_P(Y^2) = 0$ stets $\mathbf{E}_P(XY) = 0$ impliziert.)
Rechnet man mittels der Regeln für Erwartungswerte die rechte Seite von

$$0 \leq \mathbf{E}_P((aX + bY)^2)$$

weiter aus (Übung!), so folgt die CSU. □

Wendet man die CSU auf die **zentrierten ZVs** $X - \mathbf{E}_P(X)$ und $Y - \mathbf{E}_P(Y)$ an, so folgt (2). □

Die Beweise der restlichen Regeln werden als Übungsaufgaben empfohlen!

Standardisierung und Normalverteilung

Ist $X \in \mathscr{L}^2(P)$ mit $\mathbf{V}_P(X) > 0$, so folgt aus der Regel (1), dass die ZV

$$X^* := \frac{X - \mathbf{E}_P(X)}{\sqrt{\mathbf{V}_P(X)}}$$

standardisiert ist, d.h. es gilt $\mathbf{E}_P(X^*) = 0$ und $\mathbf{V}_P(X^*) = 1$.

Satz. Es seien $m, v \in \mathbb{R}$ und $v > 0$. Dann hat die Normalverteilung $\mathcal{N}_{m,v}$ zur Dichtefunktion $\phi_{m,v}(x) := \frac{1}{\sqrt{2\pi v}} e^{-(x-m)^2/(2v)}$ den Erwartungswert m und die Varianz v:

$$\begin{aligned} \mathbf{E}(\mathcal{N}_{m,v}) &= \int_{\mathbb{R}} x\phi_{m,v}(x)dx = m \\ \mathbf{V}(\mathcal{N}_{m,v}) &= \int_{\mathbb{R}} (x-m)^2\phi_{m,v}(x)dx = v. \end{aligned}$$

Insbesondere ist $\mathcal{N}_{0,1}$ die standardisierte Normalverteilung.

Zum Beweis benötigen wir folgendes fundamentale Resultat:

Gauß-Integral: $\int_{\mathbb{R}} e^{-x^2/(2v)} dx = \sqrt{2\pi v}$

Gauß-Integral: Beweisskizze

Durch Übergang von kartesischen zu Polarkoordinaten, $x(r,\phi) = r\cos(\phi)$ und $y(r,\phi) = r\sin(\phi)$, und mit dem Transformationssatz erhalten wir im Fall $v = 1$ ($v \neq 1$ analog) mit $\cos(\phi)^2 + \sin(\phi)^2 = 1$:

$$\begin{aligned}
\left(\int_{\mathbb{R}} e^{-x^2/2}dx\right)^2 &= \left(\int_{\mathbb{R}} e^{-x^2/2}dx\right)\left(\int_{\mathbb{R}} e^{-y^2/2}dy\right) = \int_{\mathbb{R}}\int_{\mathbb{R}} e^{-(x^2+y^2)/2}dxdy \\
&= \int_{r=0}^{\infty}\int_{\phi=0}^{2\pi} e^{-r^2/2}\cdot\left|\det\begin{pmatrix} \frac{\partial}{\partial r}x(r,\phi) & \frac{\partial}{\partial \phi}x(r,\phi) \\ \frac{\partial}{\partial r}y(r,\phi) & \frac{\partial}{\partial \phi}y(r,\phi)\end{pmatrix}\right| drd\phi \\
&= \int_{r=0}^{\infty}\int_{\phi=0}^{2\pi} e^{-r^2/2}\cdot\underbrace{\left|\det\begin{pmatrix} \cos(\phi) & -r\sin(\phi) \\ \sin(\phi) & r\cos(\phi)\end{pmatrix}\right|}_{=r} drd\phi \\
&= 2\pi\int_{r=0}^{\infty} r\cdot e^{-r^2/2}dr = 2\pi\left[-e^{-r^2/2}\right]_0^{\infty} = 2\pi.
\end{aligned}$$

Erwartung und Varianz von Normalverteilungen

Beweisskizze von $\mathbf{E}(\mathcal{N}_{m,v}) = m$.
Substitution $y = (x-m)/\sqrt{v}$ und Ungeradheit von $f(y) := y \cdot e^{-y^2/2}$ ergibt $\int_{\mathbb{R}} f(y)dy = 0$, und somit:

$$\begin{aligned}
\mathbf{E}(\mathcal{N}_{m,v}) &= \frac{1}{\sqrt{2\pi v}} \int x e^{-(x-m)^2/(2v)} dx \\
&= \frac{1}{\sqrt{2\pi}} \int (m + \sqrt{v}y) e^{-y^2/2} dy = m.
\end{aligned}$$

Beweisskizze von $\mathbf{V}(\mathcal{N}_{m,v}) = v$.
Die Behauptung über die Varianz folgt mit obiger Substitution und mittels partieller Integration ($\int_{\mathbb{R}} p(y) \cdot q'(y)dy = p(y) \cdot q(y)\big|_{-\infty}^{\infty} - \int_{\mathbb{R}} p'(y) \cdot q(y)dy$ mit $p(y) = y$ und $q(y) = -e^{-y^2/2}$) unter Benutzung des Wertes des Gauß-Integrals:

$$\begin{aligned}
\mathbf{V}(\mathcal{N}_{m,v}) &= \frac{1}{\sqrt{2\pi v}} \int (x-m)^2 e^{-(x-m)^2/(2v)} dx = \frac{v}{\sqrt{2\pi}} \int y^2 e^{-y^2/2} dy \\
&= \frac{v}{\sqrt{2\pi}} \left(\left[-y e^{-y^2/2} \right]_{-\infty}^{\infty} + \int e^{-y^2/2} dy \right) = v.
\end{aligned}$$

Beispiel zum Nacharbeiten(1)

Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, $\lambda > 0$ und $X : \Omega \to \mathbb{Z}_{\geq 0}$ eine Poisson-verteilte Zufallsvariable, d.h. für alle $k \in \mathbb{Z}_{\geq 0}$ gilt

$$P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}.$$

Zeige, dass die Poissonverteilung Erwartungswert λ hat, d.h. $\mathbf{E}_P(X) = \lambda$.

Lösung:

Wir nutzen die Reihendarstellung der Exponentialfunktion:

$$\mathbf{E}_P(X) = \sum_{k=0}^{\infty} k \cdot P_\lambda(k) = \sum_{k=1}^{\infty} e^{-\lambda} \cdot \frac{\lambda^k}{(k-1)!}$$

$$= e^{-\lambda} \cdot \lambda \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda} \cdot \lambda \cdot e^{\lambda} = \lambda$$

Stochastik

Kapitel 7: Gesetze der großen Zahl

Prof. Dr. Reinhard Klein
Folien von Prof. Dr. Michael Clausen

Sommersemester 2017

Institut für Informatik II, Universität Bonn

Ziele

In diesem Abschnitt wollen wir (unter anderem als wichtige Vorbereitung für Statistik-Anwendungen)

- den Zusammenhang zwischen **Wahrscheinlichkeit** und **relativer Häufigkeit** genauer beleuchten,
- die **Langzeit-Mittelwerte** von unabhängigen, identisch verteilten reellen ZVs genauer studieren und
- den **zentralen Grenzwertsatz** diskutieren, der die universelle Bedeutung der **Normalverteilung** unterstreicht.

Dabei werden wir auf verschiedene **Konvergenzbegriffe** der Stochastik stoßen.

Relative Häufigkeit bei Bernoulli-Experimenten

Sei $(X_i)_{i\geq 1}$ eine Bernoulli-Folge zur Erfolgswahrscheinlichkeit $p = \frac{1}{2}$.

Die Erfahrung zeigt, dass bei n Bernoulli-Versuchen **ungefähr** in der Hälfte der Fälle ein Erfolg eintritt.

Für die Wahrscheinlichkeit, dass Erfolg in **genau** der Hälfte der Fälle eintritt, ergibt sich mit der **Stirlingschen Formel**:

$$n! = \sqrt{2\pi n} \cdot n^n \cdot e^{-n+\eta(n)} \quad \text{mit } \tfrac{1}{12n+1} < \eta(n) < \tfrac{1}{12n}$$

bei gerader Versuchsanzahl $2n$ nach leichter Rechnung

$$P\left(\frac{1}{2n}\sum_{i=1}^{2n} X_i = \frac{1}{2}\right) = \binom{2n}{n} 2^{-2n} \sim \frac{1}{\sqrt{\pi n}} \xrightarrow{n\to\infty} 0,$$

d.h. die relative Häufigkeit liegt bei großem n nur mit sehr geringer Wahrscheinlichkeit **genau** bei $\frac{1}{2}$.

Stochastische Konvergenz

Die Erfahrung, dass die relative Erfolgshäufigkeit im obigen Beispiel **ungefähr** bei $1/2$ liegt, wird präzisiert durch

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}_P(X_1)\right| < \epsilon\right) \xrightarrow{n\to\infty} 1,$$

für alle $\epsilon > 0$.

Diese Formel würde besagen:
Für großes n liegt der Mittelwert mit großer Wahrscheinlichkeit nahe beim Erwartungswert.

Dies motiviert mit $Y_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ und $Y := \mathbf{E}_P(X_1)$ folgende allgemeinere

Definition. Es sei $(\Omega, \mathcal{A}, P)$ W-Raum, $Y, Y_1, Y_2, \ldots$ ZVs $\Omega \to \mathbb{R}$.
Die Folge $(Y_n)_n$ heißt **stochastisch konvergent** (oder **konvergent in Wahrscheinlichkeit**, genauer ***P*-konvergent**) gegen Y, kurz: $Y_n \xrightarrow{P} Y$, wenn für alle $\epsilon > 0$ gilt:

$$\boxed{\lim_{n\to\infty} P(|Y_n - Y| < \epsilon) = 1}.$$

Schwaches Gesetz der großen Zahl

Die oben erwähnte Erfahrung wird bestätigt und präzisiert durch folgenden

Satz. Es sei $(X_i)_{i\geq 1}$ eine Folge paarweise unkorrelierter (z.B. unabhängiger) ZVs in $\mathscr{L}^2(P)$ mit beschränkten Varianzen: $\mathbf{V}_P(X_i) \leq v < \infty$, für alle i. Dann ist die Wahrscheinlichkeit, dass das arithmetische Mittel der ersten n zentralisierten ZVs mindestens um ϵ von Null abweicht, nach oben beschränkt durch $v/(n\epsilon^2)$:

$$\boxed{P\left(\left|\tfrac{1}{n}\textstyle\sum_{i=1}^n(X_i - \mathbf{E}_P(X_i))\right| \geq \epsilon\right) \leq \tfrac{v}{n\epsilon^2} \xrightarrow{n\to\infty} 0}.$$

Grob gesprochen heißt das:

$$\frac{1}{n}\sum_{i=1}^n (X_i - \mathbf{E}_P(X_i)) \xrightarrow{P} 0.$$

Sind alle Erwartungswerte $\mathbf{E}_P(X_i)$ gleich, so folgt insbesondere

$$\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{P} \mathbf{E}_P(X_1).$$

Relative Häufigkeit bei Bernoulli-Versuchen

- Wir wenden das schwache Gesetz der großen Zahl an auf n Bernoulli-Versuche $X_1, \dots, X_n$ zur Erfolgswahrscheinlichkeit p. (Es ist also $P(X_i = 1) = p$ und $P(X_i = 0) = 1 - p$.)
- Dann misst die ZV $Y_n := (X_1 + \ldots + X_n)/n$ die relative Häufigkeit der Erfolge.
- Wegen $\mathbf{E}_P(X_i) = p$ und $\mathbf{V}_P(X_i) = p - p^2$ (wieso?) gilt nach dem schwachen Gesetz der großen Zahl für beliebiges $\epsilon > 0$:

$$P(|Y_n - p| \geq \epsilon) \leq \frac{\mathbf{V}_P(X_1)}{n\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2},$$

  denn $p(1-p) \leq \frac{1}{4}$.
- **Für große n ist also die Wahrscheinlichkeit, dass sich die relative Erfolgshäufigkeit um mehr als ϵ von der Erfolgswahrscheinlichkeit p unterscheidet, sehr klein.**

Tschebyscheff-Ungleichung

Zum Beweis des schwachen Gesetzes benötigen wir folgenden

Satz. Es sei $(\Omega, \mathcal{A}, P)$ W-Raum, $X \in \mathscr{L}^2(P)$, $\epsilon > 0$. Dann gilt:

$$\boxed{P(|X - \mathbf{E}_P(X)| \geq \epsilon) \leq \frac{\mathbf{V}_P(X)}{\epsilon^2}}.$$

Beweis. Wir beschränken uns auf den Fall einer diskreten ZV X.
In dem Fall ist auch $Y := (X - \mathbf{E}_P(X))^2$ diskret, liegt in $\mathscr{L}^1(P)$ (wieso?) und es gilt:

$$\begin{aligned}
\mathbf{V}_P(X) &= \mathbf{E}_P(Y) = \sum_{y \in Y[\Omega]} y P(Y = y) \\
&\geq \sum_{y \in Y[\Omega]: y \geq \epsilon^2} \epsilon^2 \cdot P(Y = y) + \sum_{y \in Y[\Omega]: y < \epsilon^2} 0 \cdot P(Y = y) \\
&= \epsilon^2 \sum_{y \in Y[\Omega]: y \geq \epsilon^2} P(Y = y) = \epsilon^2 P(Y \geq \epsilon^2) \\
&= \epsilon^2 P((X - \mathbf{E}_P(X))^2 \geq \epsilon^2) = \epsilon^2 P(|X - \mathbf{E}_P(X)| \geq \epsilon). \qquad \square
\end{aligned}$$

Schwaches Gesetz der großen Zahl: Beweis

Setze $Y_n := \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbf{E}_P(X_i))$.

Dann ist $Y_n \in \mathscr{L}^2(P)$ und mit der Linearität der Erwartungswertbildung folgt $\mathbf{E}_P(Y_n) = 0$.

Aus $\mathbf{V}_P(aX+b) = a^2\mathbf{V}_P(X)$ und $\mathbf{V}_P(\sum_i X_i) = \sum_i \mathbf{V}_P(X_i)$ für unkorrelierte ZVs folgt:

$$\mathbf{V}_P(Y_n) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{V}_P(X_i) \le \frac{v}{n}.$$

Mit der Tschebyscheff-Ungleichung ergibt sich:

$$P\left(\left|Y_n - \underbrace{\mathbf{E}_P(Y_n)}_{=0}\right| \ge \epsilon\right) \le \frac{\mathbf{V}_P(Y_n)}{\epsilon^2} \le \frac{v}{n\epsilon^2}.$$

Damit ist das schwache Gesetz der großen Zahl bewiesen. $\square$

Monte-Carlo-Integration 1

- Als erste Anwendung des schwachen Gesetzes der großen Zahl diskutieren wir die Monte-Carlo-Integration.
- Es sei $f : [0,1] \to [0,c]$ messbar.
- **Ziel**: numerische Berechnung des Integrals $\int_0^1 f(x)dx$.
- Zunächst interpretieren wir dieses **Integral als Erwartungswert**:
- Ist $\mathcal{U}$ die Gleichverteilung auf $[0,1]$, so hat diese die Indikatorfunktion von $[0,1]$ als Dichtefunktion ρ: $\rho(x) = 1$.
- Daher ist (siehe Kapitel 6: Erwartungswert bei W-Maßen mit Dichtefunktion)

$$\mathbf{E}_{\mathcal{U}}(f) = \int_0^1 f(x)\,\underbrace{\rho(x)}_{=1}\,dx = \int_0^1 f(x)dx.$$

Monte-Carlo-Integration 2

- Sind nun $X_1, \ldots, X_n$ unabhängige, auf $[0,1]$ gleichverteilte ZVs, d.h.

$$X_i := ([0,1]^n \ni \xi \mapsto \xi_i \in [0,1])$$

und $P_{X_i} = \mathcal{U}_n \circ X_i^{-1} = \mathcal{U}$, so folgt aus dem schwachen Gesetz der großen Zahl:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \int_0^1 f(x)dx\right| \geq \epsilon\right) \leq \frac{\mathbf{V}_{\mathcal{U}}(f \circ X_1)}{n\epsilon^2} \leq \frac{c^2}{n\epsilon^2}.$$

- Letzte Ungleichung: beachte, dass $f : [0,1] \to [0,c]$ und

$$\begin{aligned} \mathbf{V}_{\mathcal{U}}(f \circ X_1) &= \mathbf{E}_{\mathcal{U}}[(f \circ X_1)^2] - \mathbf{E}_{\mathcal{U}}[f \circ X_1]^2 \\ &\leq \mathbf{E}_{\mathcal{U}}[(f \circ X_1)^2] = \int_0^1 f(x)^2 dx \leq c^2. \end{aligned}$$

- Man kann also durch zufällige Wahl von vielen Punkten $x_1, \ldots, x_n \in [0,1]$ das Integral $\int_0^1 f(x)dx$ durch das arithmetische Mittel von $f(x_1), \ldots, f(x_n)$ mit hoher Wahrscheinlichkeit sehr gut approximieren!

Fast sichere Konvergenz

Der folgende (stärkere) Konvergenzbegriff liegt dem starken Gesetz der großen Zahl zugrunde:

Definition. Seien $Y, Y_1, Y_2, \ldots$ reelle ZVs auf $(\Omega, \mathcal{A}, P)$.
Die Folge $(Y_n)_{n\geq 1}$ **konvergiert** P-**fast sicher** gegen Y, wenn für die Menge

$$A := \{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}$$

aller Stellen $\omega \in \Omega$, an denen punktweise Konvergenz herrscht, gilt

$$P(A) = 1.$$

Bemerkungen:

- Allgemein sagt man, eine Aussage gelte **fast sicher**, wenn sie mit Wahrscheinlichkeit 1 zutrifft.
- Das Ereignis A liegt in $\mathcal{A}$ (Begründung: nächste Folie), also macht die Definition Sinn.

Fast sichere Konvergenz impliziert stochastische Konvergenz

Beweis. $(Y_n)_{n\geq 1}$ konvergiere P-fast sicher gegen Y. Das bedeutet $P(A) = 1$, wobei $A := \{\omega \in \Omega \mid \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}$. Für $\epsilon > 0$ ist

$$A_\epsilon := \{\omega \in \Omega \mid \exists N\, \forall n \geq N : |Y_n(\omega) - Y(\omega)| < \epsilon\} = \bigcup_{N\geq 1} \bigcap_{n\geq N} (Y_n - Y)^{-1}(-\epsilon, +\epsilon) \in \mathcal{A}.$$

Dann ist $A = \bigcap_{\epsilon>0} A_\epsilon \in \mathcal{A}$. Wegen $A_\epsilon \supseteq A$ ist auch $P(A_\epsilon) = 1$. Mit Hilfe von $B_N^\epsilon := \bigcap_{n\geq N}\{|Y_n - Y| < \epsilon\}$ zerlegen wir A_ϵ weiter und erhalten

$$1 = P(A_\epsilon) = P\Big(\bigcup_{N\geq 1} B_N^\epsilon\Big).$$

Da $(B_N^\epsilon)_{N\geq 1}$ monoton steigend ist und $B_N^\epsilon \uparrow A_\epsilon = \bigcup_{N\geq 1} B_N^\epsilon$ gilt, folgt aus der σ-Stetigkeit von P (siehe nächste Folie):

$$\begin{aligned} 1 &= P\Big(\bigcup_{N\geq 1} B_N^\epsilon\Big) = \lim_{N\to\infty} P\Big(B_N^\epsilon\Big) \leq \lim_{N\to\infty} \inf_{n\geq N} P(|Y_n - Y| < \epsilon) \\ &= \liminf_{n\to\infty} P(|Y_n - Y| < \epsilon) \leq 1. \end{aligned}$$

Folglich ist $\lim_{n\to\infty} P(|Y_n - Y| < \epsilon) = 1$, für alle $\epsilon > 0$. D.h. $(Y_n)_{n\geq 1}$ konvergiert stochastisch gegen Y. $\square$

Memo: Sigma-Stetigkeit von W-Maßen

Satz. Sei $(\Omega, \mathcal{A}, P)$ ein W-Raum und $(A_n)_{n\geq 1}$ eine aufsteigende Kette von Ereignissen $A_n \in \mathcal{A}$. Ist $A := \bigcup_{n\geq 1} A_n$, so gilt:

$$P(A) = \lim_{n\to\infty} P(A_n).$$

Beweis. Mit $A_0 := \emptyset$ gilt unter Verwendung der σ-Additivität von P:

$$\begin{aligned} P(A) &= P\Big(\bigsqcup_{i\geq 1}(A_i \setminus A_{i-1})\Big) \\ &= \sum_{i\geq 1} P(A_i \setminus A_{i-1}) = \lim_{n\to\infty} \sum_{i=1}^{n} P(A_i \setminus A_{i-1}) \\ &= \lim_{n\to\infty} P\Big(\bigsqcup_{i=1}^{n}(A_i \setminus A_{i-1})\Big) \\ &= \lim_{n\to\infty} P(A_n). \end{aligned}$$

Damit ist die σ-Stetigkeit von W-Maßen bewiesen. $\square$

P-Konvergenz impliziert nicht fast sichere Konvergenz

Das Einheitsintervall $(0,1]$ sei mit der Gleichverteilung $\mathcal{U}$ versehen, und für $k = 2^n + m$ mit $m \in [0 : 2^n - 1]$ sei Y_k die Indikatorfunktion von $(m2^{-n}, (m+1)2^{-n}] \subset (0,1]$.

Satz. $(Y_k)_{k\in\mathbb{N}}$ konvergiert stochastisch aber nicht fast sicher gegen $Y = 0$.

Beweis. Beachte, dass für alle k zwischen 2^n und $2^{n+1} - 1$ die ZV Y_k auf einem 2^n-tel des Einheitsintervalls Eins und auf dem Rest Null ist. Daher konvergiert die Folge der Y_k stochastisch gegen $Y = 0$. Genauer ergibt sich für $k \in [2^n : 2^{n+1} - 1]$ wegen $k/2 < 2^n$:

$$\mathcal{U}(|Y_k - Y| \leq \epsilon) \geq 1 - 2^{-n} > 1 - \tfrac{2}{k}.$$

Die Folge $(Y_k)_{k\geq 1}$ ist aber nicht fast sicher konvergent, da für beliebiges $\omega \in (0,1]$ beide Werte 0 und 1 unendlich oft in der Folge $(Y_k(\omega))_k$ vorkommen. Hier ist also

$$A := \{\omega \in \Omega : \lim_{k\to\infty} Y_k(\omega) = Y(\omega)\} = \emptyset.$$

Für die fast sichere Konvergenz müsste $P(A) = 1$ gelten. Stattdessen haben wir $P(A) = P(\emptyset) = 0$. (!) $\square$

Starkes Gesetz der großen Zahl

Vor dem Hintergrund des letzten Beispiels ist das folgende Resultat um so erstaunlicher:

Satz. Sei $(X_i)_{i\geq 1}$ eine Folge paarweise unkorrelierter ZVs in $\mathscr{L}^2(P)$ mit beschränkten Varianzen: $\sup_{i\geq 1} \mathbf{V}_P(X_i) \leq v < \infty$. Dann gilt:

$$\boxed{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbf{E}_P(X_i)) \to 0 \quad P\text{-fast sicher.}}$$

- Der **Beweis** verwendet neben dem schwachen Gesetz und der Tschebyscheff-Ungleichung das Borel-Cantelli Lemma.
- Wir beginnen daher mit diesem Lemma, das zu den sog. **0-1-Gesetzen** der W-Theorie zählt.

Das Borel-Cantelli Lemma

Lemma (Borel-Cantelli). Es sei $(\Omega, \mathcal{A}, P)$ ein W-Raum und $A_k \in \mathcal{A}$, für alle $k \geq 1$. Setze

$$A \equiv \limsup_{k\to\infty} A_k := \{\omega \in \Omega \mid \omega \in A_k \text{ für unendlich viele } k\}.$$

(0) Ist $\sum_{k\geq 1} P(A_k) < \infty$, so ist $P(A) = 0$.
(1) Ist $\sum_{k\geq 1} P(A_k) = \infty$ und sind die Mitglieder der Ereignisfamilie $(A_k)_{k\geq 1}$ paarweise stochastisch unabhängig, so gilt $P(A) = 1$.

Beweis von (0). Für alle $m \in \mathbb{N}$ gilt $A \subseteq \bigcup_{k\geq m} A_k$ und daher $P(A) \leq \sum_{k\geq m} P(A_k)$. Aus $\sum_{k\geq 1} P(A_k) < \infty$ folgt unmittelbar

$$0 \leq P(A) \leq \lim_{m\to\infty} \sum_{k\geq m} P(A_k) = 0.$$

Damit ist (0) bewiesen.

Da wir Aussage (1) später nicht benötigen, verzichten wir auf den Beweis dieser Aussage und verweisen stattdessen auf die Literatur. $\square$

Beweis des starken Gesetzes: Vorbereitungen

- Zunächst kann man ohne Einschränkung annehmen, dass die X_i zentriert sind: $\mathbf{E}_P(X_i) = 0$. Sonst argumentiere man mit $X_i' := X_i - \mathbf{E}_P(X_i)$.
- Beachte: auch die X_i' sind unkorreliert mit beschränkten Varianzen.
- Setze $Y_n := \frac{1}{n} \sum_{i=1}^n X_i$.
- Der Beweis geschieht nun in zwei Schritten:
- **Schritt 1**: Für die Teilfolge $(Y_{n^2})_{n \geq 1}$ wird gezeigt, dass P-fast sicher $Y_{n^2} \to 0$ gilt.
- **Schritt 2**: Mit Schritt 1 wird dann interpoliert und gezeigt, dass P-fast sicher $Y_n \to 0$ gilt.

Beweis des starken Gesetzes: Schritt 1

- Nach dem schwachen Gesetz gilt für alle $\epsilon > 0$

$$P(|Y_{n^2}| \geq \epsilon) \leq \frac{v}{n^2\epsilon^2}\,.$$

- Wegen

$$\sum_{n\geq 1} P(|Y_{n^2}| \geq \epsilon) \leq \frac{v}{\epsilon^2}\sum_{n\geq 1} n^{-2} \leq \frac{v}{\epsilon^2}\left(1 + \lim_{K\to\infty}\int_1^K x^{-2}dx\right) \leq \frac{2v}{\epsilon^2} < \infty$$

folgt aus Aussage (0) des Borel-Cantelli Lemmas $P(A_\epsilon) = 0$, wobei

$$A_\epsilon := \{|Y_{n^2}| \geq \epsilon \text{ für } \infty \text{ viele } n\} = \{\omega \in \Omega \mid \forall N \exists n \geq N : |Y_{n^2}(\omega)| \geq \epsilon\}.$$

Also hat für alle $\epsilon > 0$ das Komplement

$$\Omega \setminus A_\epsilon = \{\omega \in \Omega \mid \exists N \forall n \geq N : |Y_{n^2}(\omega)| < \epsilon\}$$

Wahrscheinlichkeit 1. D.h. $Y_{n^2} \to\ 0$ P-fast sicher.

Beweis des starken Gesetzes: Schritt 2.1

- Für $m \in \mathbb{N}$ sei $n = n(m)$ so gewählt, dass $n^2 \le m < (n+1)^2$.
- Wir vergleichen Y_m mit Y_{n^2} und setzen $S_k := k \cdot Y_k = \sum_{i=1}^k X_i$.
- Aus der Tschebyscheff-Ungleichung folgt mit der Unkorreliertheit der X_i dann für $S_m - S_{n^2} = X_{n^2+1} + \ldots + X_m$:

$$P(|S_m - S_{n^2}| \ge \epsilon \cdot n^2) \le \frac{1}{\epsilon^2 n^4} \mathbf{V}_P \Big(\sum_{n^2 < i \le m} X_i \Big) \le \frac{v(m-n^2)}{\epsilon^2 n^4}.$$

- Daraus ergibt sich weiter (beachte: $(n+1)^2 - 1 - n^2 = 2n$):

$$\begin{aligned}
\sum_{m \ge 1} P(|S_m - S_{n(m)^2}| \ge \epsilon \cdot n(m)^2) \quad &\le \quad \frac{v}{\epsilon^2} \sum_{n \ge 1} \sum_{m=n^2+1}^{(n+1)^2-1} \frac{m-n^2}{n^4} \\
&= \quad \frac{v}{\epsilon^2} \sum_{n \ge 1} \sum_{k=1}^{2n} \frac{k}{n^4} = \frac{v}{\epsilon^2} \sum_{n \ge 1} \frac{(2n)(2n+1)}{2n^4} < \infty \,.
\end{aligned}$$

Beweis des starken Gesetzes: Schritt 2.2

- Wegen $S_k/k = Y_k$ folgt mit Borel-Cantelli wie in Schritt 1

$$P\Big(\Big|\frac{S_m}{n(m)^2} - Y_{n(m)^2}\Big| \xrightarrow{m\to\infty} 0\Big) = 1.$$

- **Hilfssatz**. Für $A, B \in \mathcal{A}$ gilt: $P(A) = P(B) = 1$ impliziert $P(A \cap B) = 1$.

$$\textbf{Beweis}: \quad \begin{aligned} 1 &= P(A \cup B) = P(A \setminus B) + P(B \setminus A) + P(A \cap B) \\ &\leq P(\Omega \setminus B) + P(\Omega \setminus A) + P(A \cap B) \\ &= 1 - P(B) + 1 - P(A) + P(A \cap B) = P(A \cap B) \leq 1. \quad \square \end{aligned}$$

- Auf $A := \{Y_{n^2} \xrightarrow{n\to\infty} 0\}$ (wegen Schritt 1) und $B := \{|\frac{S_m}{n(m)^2} - Y_{n(m)^2}| \xrightarrow{m\to\infty} 0\}$ ist der Hilfssatz anwendbar und ergibt

$$P\Big(\frac{S_m}{n(m)^2} \xrightarrow{m\to\infty} 0\Big) = 1.$$

- Wegen $|Y_m| = |S_m|/m \leq |S_m|/n(m)^2$ folgt $P(Y_m \xrightarrow{m\to\infty} 0) = 1.$

Damit ist das starke Gesetz der großen Zahl bewiesen. $\square$

Die Verteilungsfunktion der Gauß-Glocke

Definition. Ist P ein W-Maß auf $(\mathbb{R}, \mathcal{B})$, so heißt

$$F_P := (\mathbb{R} \ni c \mapsto P((-\infty, c]) \in [0,1]$$

die **Verteilungsfunktion** zu P.

Die Funktion

$$\phi(x) := \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

ist die **Dichtefunktion zur Standardnormalverteilung** $\mathcal{N}_{0,1}$.
Für $c \in \mathbb{R}$ setze

$$\Phi(c) := \mathcal{N}_{0,1}((-\infty, c]) = \int_{-\infty}^{c} \phi(x)dx.$$

Dann ist Φ die **Verteilungsfunktion der Standardnormalverteilung**.

Zentraler Grenzwertsatz

Der **Supremumsabstand** zweier Verteilungsfunktionen F und G ist definiert durch

$$d(F,G) := \sup_{c\in\mathbb{R}} |F(c) - G(c)|.$$

Satz. Es sei $(X_i)_{i\geq 1}$ eine Folge unabhängiger, identisch verteilter reellwertiger ZVs in $\mathscr{L}^2(P)$ mit $\mathbf{E}_P(X_i) = m$ und $\mathbf{V}_P(X_i) = v > 0$. Sei $S_n := X_1 + \ldots + X_n$ die n-te Partialsumme und

$$S_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - m}{\sqrt{v}}$$

die Standardisierung von S_n. Bezeichnet F_n die Verteilungsfunktion von S_n^*, so konvergiert F_n im Supremumsabstand gegen die Verteilungsfunktion Φ der Standardnormalverteilung:

$$\lim_{n\to\infty} d(F_n, \Phi) = 0.$$

Für den etwas längeren **Beweis** verweisen wir auf die Literatur.

Stochastik

Kapitel 8: Markov-Ketten

Prof. Dr. Reinhard Klein
Folien von Prof. Dr. Michael Clausen

Sommersemester 2017

INSTITUT FÜR INFORMATIK II, UNIVERSITÄT BONN

Einführung

- Bei den Produktmodellen waren die Projektionen $X_1, X_2, \ldots$ Folgen **unabhängiger** ZVs.
- In diesem Abschnitt geht es um ZV-Folgen mit einer besonders einfachen Art der stochastischen **Abhängigkeit**:
  Eine **Markov-Kette** ist - grob gesprochen - eine ZV-Folge mit **kurzem Gedächtnis**: das Verhalten zum Zeitpunkt $n+1$ hängt nur vom Zustand zum Zeitpunkt n ab.
- Von Interesse ist besonders das Langzeitverhalten solcher Folgen: wir diskutieren hier nur die **Konvergenz ins Gleichgewicht**.
- Der Aspekt der **Absorption in einer Falle** kann aus Zeitgründen nur angedeutet werden.

Stochastische Matrizen

Es sei $V \neq \emptyset$ abzählbar und $\Pi = (\Pi(x,y))_{x,y \in V}$ eine reellwertige Matrix.

Definition. Π heißt **zeilenstochastisch**, wenn in jeder Zeile der Matrix Π eine W-Funktion auf V steht. Das heißt:

- alle Einträge von Π liegen im Intervall $[0,1]$: $\Pi \in [0,1]^{V \times V}$.
- für alle $x \in V$ ist $\sum_{y \in V} \Pi(x,y) = 1$.

Wir betrachten jetzt den Zufallsprozess in V, der bei jedem Schritt mit Wahrscheinlichkeit $\Pi(x,y)$ vom Zustand x zum Zustand y springt. Symbolisch:

$$x \xrightarrow{\Pi(x,y)} y.$$

Markov-Ketten

Definition. Eine Folge $X_0, X_1, \ldots$ von ZVs auf dem W-Raum $(\Omega, \mathcal{A}, P)$ mit Werten in V, d.h. $X_i : (\Omega, \mathcal{A}) \to (V, 2^V)$, heißt **Markov-Kette** mit **Zustandsraum** V und **Übergangsmatrix** Π,
wenn für alle $n \geq 0$ und für alle $x_0, \ldots, x_{n+1} \in V$ folgende **Markov-Eigenschaft** erfüllt ist:

$$\begin{aligned} & P(X_{n+1} = x_{n+1} | X_0 = x_0, \ldots, X_n = x_n) \\ = \quad & P(X_{n+1} = x_{n+1} | X_n = x_n), \end{aligned}$$

sofern $P(X_0 = x_0, \ldots, X_n = x_n) > 0$.

Die Verteilung $\alpha := P \circ X_0^{-1}$ von X_0 heißt die **Startverteilung** der Markov-Kette.

Markov-Eigenschaft

- Die bedingte Verteilung von X_{n+1} bei bekannter Vorgeschichte $x_0, \ldots, x_n$ hängt nur von der **Gegenwart** x_n aber nicht von der **Vergangenheit** $x_0, \ldots, x_{n-1}$ ab.
- Die bedingten Verteilungen hängen nicht vom Zeitpunkt n ab. Diese **Zeitinvarianz** von Π bezeichnet man als Fall der **stationären Übergangswahrscheinlichkeiten**.
- **Fazit**: Eine Markov-Kette $(X_n)_{n\geq 0}$ ist ein stochastischer Prozess mit kurzem Gedächtnis von genau einer Zeiteinheit und ohne innere Uhr.

Mehrstufige Modelle

Die Konstruktion mehrstufiger Modelle umfasst Markov-Ketten als Spezialfall. Setzt man

$$p_{k|\omega_0,\ldots,\omega_{k-1}}(\omega_k) := \Pi(\omega_{k-1}, \omega_k),$$

so sieht man, dass die Markov-Eigenschaft ein Spezialfall der Bedingung (b) im Satz über die Konstruktion mehrstufiger Modelle ist.

Nach diesem Satz existiert zu jeder Startverteilung α auf V genau ein W-Maß P^α auf

$$(\Omega, \mathcal{A}) := \Big(\prod_{k\geq 0} V, \bigotimes_{k\geq 0} 2^V\Big)$$

derart, dass die Projektionen $X_n : \Omega \to V$,

$$X_n : (\omega_k)_{k\geq 0} \mapsto \omega_n,$$

eine Markov-Kette zu Π und α bilden.

Ist $\alpha = \delta_x$ für ein $x \in V$ (*sicherer Start* in x), so schreibt man kurz P^x statt P^α.

Matrixpotenzen

Satz. Die n-te Potenz Π^n der zeilenstochastischen Matrix Π enthält an der Position (x, y) die Wahrscheinlichkeit, in genau n Schritten vom Zustand x in den Zustand y zu gelangen:

$$P^x(X_n = y) = \Pi^n(x, y).$$

Beweis. Für beliebiges $n \geq 1$ und $x_1, \ldots, x_n \in V$ gilt nach der Produktformel

$$\begin{aligned}
& P^x(X_1 = x_1, \ldots, X_n = x_n) \\
= \quad & P^x(X_1 = x_1 | X_0 = x) \cdot P^x(X_2 = x_2 | X_0 = x, X_1 = x_1) \cdot \ldots \\
& \ldots \cdot P^x(X_n = x_n | X_0 = x, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) \\
= \quad & \Pi(x, x_1) \cdot \Pi(x_1, x_2) \cdot \ldots \cdot \Pi(x_{n-1}, x_n).
\end{aligned}$$

Durch Summation über $x_1, \ldots, x_{n-1} \in V$ erhält man für alle $x, y \in V$:
$P^x(X_n = y) = \Pi^n(x, y)$. $\square$

Übergangsgraphen

Im endlichen Fall ($|V| < \infty$) veranschaulicht man gerne die zu $\Pi \in [0,1]^{V \times V}$ gehörige Markov-Kette durch einen gerichteten, kantengewichteten Graphen $G = (V, E, \Pi)$, den sog. **Übergangsgraphen** zu Π. Genauer:

- V ist die endliche Knotenmenge (engl.: **V**ertices).
- $E := \{(x,y) \in V \times V | \Pi(x,y) > 0\}$ ist die Kantenmenge (engl.: **E**dges).
- $\Pi(x,y)$ ist das Kantengewicht von $(x,y) \in E$.

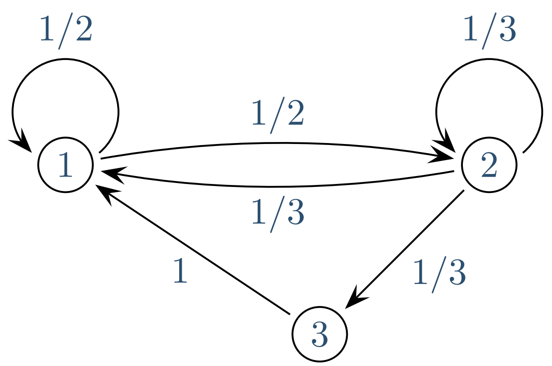
Es folgen drei Beispiele.

Übergangsgraphen: Beispiel 1

Es sei $V := \{1, 2, 3\}$ und

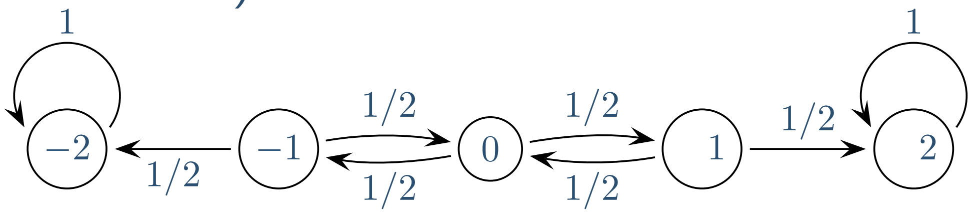
$$\Pi = \frac{1}{6}\begin{pmatrix} 3 & 3 & 0 \\ 2 & 2 & 2 \\ 6 & 0 & 0 \end{pmatrix}.$$

Der zugehörige Übergangsgraph sieht so aus:



Übergangsgraphen: Beispiel 2

Hier geht es um folgendes Münzwurfspiel:
Spieler A und B haben a bzw. b Euro. Nach jedem Münzwurf zahlt - je nach Ergebnis - der Verlierer dem Gewinner 1 Euro. Das Spiel ist beendet, wenn ein Spieler sein Kapital verspielt hat. Es sei X_n der Gewinn von Spieler A nach n Spielen. Die X_n haben Werte in $V = [-a : b]$ und bilden eine Markov-Kette mit Übergangsgraph (hier gezeigt im Spezialfall $a = b = 2$):



Der Übergangsgraph zeigt, dass die Zustände $x = -a$ und $x = b$ sog. **Fallen** sind: **einmal drin, immer drin**!

In diesem Beispiel interessiert die sog. **Ruinwahrscheinlichkeit**

$$P(\text{Spiel endet mit Ruin von Spieler } A).$$

Das WWW als gerichteter Graph

Das Wold Wide Web kann man als gerichteten Graphen $G = (V, E)$ modellieren:

- Knotenmenge V ist die Menge der Webseiten, o.E. $V = [1 : N]$.
- Kantenmenge E ist die Menge der Links.

Für $x \in V$ bezeichne

$$G[x] := \{y \in V | (x, y) \in E\} \cup \{x\}$$

die Menge der von x aus verlinkten Seiten inklusive der Ausgangsseite x.

Websurfen als Markov-Prozess

Das Websurfen kann man idealisiert folgendermaßen modellieren:
Ein Surfer, der sich auf Seite x befindet

- gibt mit Wahrscheinlichkeit $p \in (0,1)$ gemäß Gleichverteilung eine neue URL ein, oder
- er folgt mit Wahrscheinlichkeit $1-p$ gemäß Gleichverteilung einem Link von x auf eine andere Webseite oder bleibt auf x.

Dies ist eine Markov-Kette mit Übergangswahrscheinlichkeiten

$$\Pi(x,y) := \begin{cases} \frac{p}{N} & \text{falls } (x,y) \notin E \\ \frac{p}{N} + \frac{1-p}{|G[x]|} & \text{falls } (x,y) \in E. \end{cases}$$

Bemerkung. Π ist zeilenstochastisch, denn für $x \in V$ ist

$$\sum_{y \in V} \Pi(x,y) = \underbrace{\sum_{y=1}^{N} \frac{p}{N}}_{=p} + \underbrace{\sum_{y \in G[x]} \frac{1-p}{|G[x]|}}_{=1-p} = 1.$$

Der Ergodensatz

Ergodensatz. Es sei $\Pi = (\pi_{ij}) \in [0,1]^{N \times N}$ zeilenstochastisch. Ferner gebe es ein $L \geq 1$, so dass alle Einträge in Π^L positiv sind. Dann gibt es eine W-Funktion $\rho = (\rho_1, \ldots, \rho_N)$, die sog. **Grenzverteilung**, mit folgenden Eigenschaften:

- $\lim_{m\to\infty} \Pi^m =: \Pi^\infty$ existiert und in jeder Zeile von Π^∞ steht die Grenzverteilung:

$$\Pi^\infty = \begin{pmatrix} \rho_1 & \rho_2 & \ldots & \rho_N \\ \rho_1 & \rho_2 & \ldots & \rho_N \\ \vdots & \vdots & \ldots & \vdots \\ \rho_1 & \rho_2 & \ldots & \rho_N \end{pmatrix}.$$

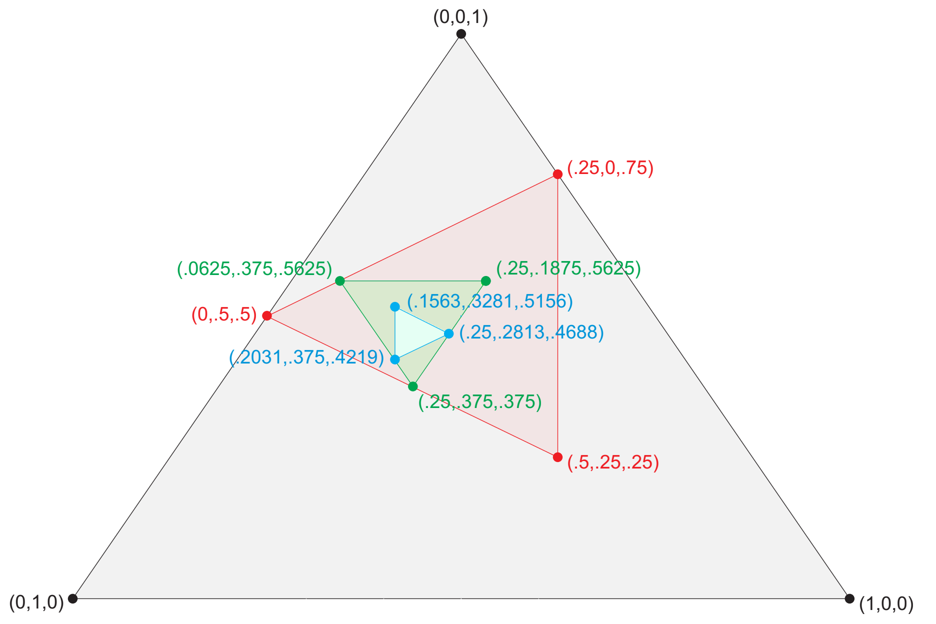
- Die Matrizenfolge $(\Pi^m)_{m\geq 1}$ konvergiert exponentiell schnell gegen Π^∞.
- Die Grenzverteilung ρ ist die eindeutig bestimmte W-Funktion mit $\rho\Pi = \rho$.

Geometrischer Hintergrund des Ergodensatzes

Ist $\Pi = (\pi_{ij}) \in [0,1]^{n\times n}$ zeilenstochastisch mit nur positiven Einträgen, und bezeichnen wir mit $\Pi_1, \ldots, \Pi_n$ die Zeilen der Matrix Π, so sind die Zeilen Q_i der Matrix $Q := \Pi^2$ Konvexkombinationen der Zeilen von Π, liegen also in der konvexen Hülle von $\Pi_1, \ldots, \Pi_n$:

$$Q_i := \sum_{j=1}^{n} \pi_{ij} \Pi_j .$$

Es zeigt sich, dass das wiederholte Bilden von Konvexkombinationen geometrisch einen Schrumpfungsprozess auslöst. Dies wird durch folgendes Schaubild illustriert:



Beweis des Ergodensatzes (1)

Für $N = 1$ ist die Behauptung trivialerweise richtig. Sei also $N \geq 2$.

Wir betrachten für festes $j \in [1 : N]$ die Folge der Minima und Maxima in den j-ten Spalten der Matrizen $\Pi, \Pi^2, \Pi^3, \ldots$. Dazu sei $\Pi^n =: (\pi_{ij}^{(n)})$ und

$$m_j^{(n)} := \min_i \pi_{ij}^{(n)} \quad \text{sowie} \quad M_j^{(n)} := \max_i \pi_{ij}^{(n)}.$$

Schritt 1: Bei festem j ist die Folge $(m_j^{(n)})_{n \geq 1}$ monoton steigend und die Folge $(M_j^{(n)})_{n \geq 1}$ monoton fallend.

Zunächst gilt für $(i, j) \in [1 : N]^2$ und alle $a, b \in \mathbb{N}$ die Formel

$$\pi_{ij}^{(a+b)} = \sum_{\ell=1}^{N} \pi_{i\ell}^{(a)} \pi_{\ell j}^{(b)}. \tag{1}$$

[Benutze $\Pi^{a+b} = \Pi^a \cdot \Pi^b$ und die Definition $\Pi^n =: (\pi_{ij}^{(n)})$.]

Beweis des Ergodensatzes (2)

Damit ergeben sich aus der Zeilenstochastizität von Π folgende Monotoniebeziehungen:

$$m_j^{(n+1)} = \min_i \sum_{\ell=1}^{N} \pi_{i\ell}\pi_{\ell j}^{(n)} \geq \min_i \sum_{\ell=1}^{N} \pi_{i\ell} m_j^{(n)} = m_j^{(n)}$$

und analog

$$M_j^{(n+1)} = \max_i \sum_{\ell=1}^{N} \pi_{i\ell}\pi_{\ell j}^{(n)} \leq \max_i \sum_{\ell} \pi_{i\ell} M_j^{(n)} = M_j^{(n)}.$$

Also ist die Minima-Folge $(m_j^{(n)})_{n\geq 1}$ monoton steigend, während umgekehrt die Maxima-Folge $(M_j^{(n)})_{n\geq 1}$ monoton fallend ist.

Im **nächsten Schritt** betrachten wir die Teilfolge $(\Pi^{\lambda L})_{\lambda \geq 1}$ und zeigen, dass für jeden Spaltenindex j die Minima und Maxima in den j-ten Spalten gegen denselben Wert ρ_j konvergieren:

$$\lim_{\lambda\to\infty} m_j^{(\lambda L)} = \lim_{\lambda\to\infty} M_j^{(\lambda L)}.$$

Beweis des Ergodensatzes (3)

Schritt 2: Es gibt ein $\delta \in (0,1)$ mit $M_j^{(\lambda L)} - m_j^{(\lambda L)} \leq (1-\delta)^\lambda$, für alle $\lambda \in \mathbb{Z}_{\geq 1}$.
[Daraus folgt dann $\lim_{n\to\infty} m_j^{(n)} = \lim_{n\to\infty} M_j^{(n)} =: \rho_j$, womit (1) bewiesen wäre.]

Es sei $\delta := \min_{(i,j)} \pi_{ij}^{(L)}$. Da alle $\pi_{ij}^{(L)}$ positiv sind und Π^L wieder zeilenstochastisch ist (Übung!), gilt zusammen mit $N \geq 2$:

$$\forall (i,j): \quad 0 < \delta \leq \pi_{ij}^{(L)} < 1. \tag{2}$$

Insbesondere gilt für alle j stets $m_j^{(L)} \geq \delta$. Für Zeilenindizes $h, i \in [1:N]$ seien

$$I^+(h,i) := \{k \mid \pi_{hk}^{(L)} \geq \pi_{ik}^{(L)}\} \quad \text{und} \quad I^-(h,i) := [1:N] \setminus I^+(h,i).$$

Damit gilt (nach getrennter Zusammenfassung der positiven bzw. negativen Summanden und wegen $[1:N] = I^+(h,i) \sqcup I^-(h,i)$):

$$\underbrace{\sum_{k\in I^+(h,i)} (\pi_{hk}^{(L)} - \pi_{ik}^{(L)})}_{=:A} + \underbrace{\sum_{k\in I^-(h,i)} (\pi_{hk}^{(L)} - \pi_{ik}^{(L)})}_{=:B} = 1 - 1 = 0. \tag{3}$$

Beweis des Ergodensatzes (4)

Ist nun für festes n und j das Indexpaar (h,i) so gewählt, dass

$$M_j^{(n+L)} = \pi_{hj}^{(n+L)} \quad \text{und} \quad m_j^{(n+L)} = \pi_{ij}^{(n+L)},$$

so ist

$$\begin{aligned}
M_j^{(n+L)} - m_j^{(n+L)} &= \pi_{hj}^{(n+L)} - \pi_{ij}^{(n+L)} = \sum_k (\pi_{hk}^{(L)} - \pi_{ik}^{(L)})\pi_{kj}^{(n)} \quad \text{wg. (1)} \\
&= \sum_{k \in I^+(h,i)} \underbrace{(\pi_{hk}^{(L)} - \pi_{ik}^{(L)})}_{\geq 0}\pi_{kj}^{(n)} + \sum_{k \in I^-(h,i)} \underbrace{(\pi_{hk}^{(L)} - \pi_{ik}^{(L)})}_{<0}\pi_{kj}^{(n)} \\
&\leq \sum_{k \in I^+(h,i)} (\pi_{hk}^{(L)} - \pi_{ik}^{(L)}) M_j^{(n)} + \sum_{k \in I^-(h,i)} (\pi_{hk}^{(L)} - \pi_{ik}^{(L)}) m_j^{(n)} \\
&= \sum_{k \in I^+(h,i)} (\pi_{hk}^{(L)} - \pi_{ik}^{(L)})(M_j^{(n)} - m_j^{(n)}) \quad \text{wg. } B = -A \text{ in (3)} \\
&\leq (1-\delta)(M_j^{(n)} - m_j^{(n)}). \quad \text{(Begründung folgt!)}
\end{aligned}$$

Beweis des Ergodensatzes (5)

Begründung:

- Da Π^L zeilenstochastisch ist, folgt $\sum_{k\in I^+(h,i)} \pi_{hk}^{(L)} \leq 1$.
- Weiter ist $I^+(h,i) \neq \emptyset$:
  - denn aus $I^+(h,i) = \emptyset$ würde $I^-(h,i) = [1:N]$ folgen, d.h. für alle $k \in [1:N]$ wäre $\pi_{hk}^{(L)} < \pi_{ik}^{(L)}$, was wegen der Zeilenstochastizität von Π^L nicht geht!
- Also ist $\sum_{k\in I^+(h,i)} \pi_{ik}^{(L)} \geq |I^+(h,i)| \cdot \delta \geq \delta$, woraus insgesamt $\sum_{k\in I^+(h,i)}(\pi_{hk}^{(L)} - \pi_{ik}^{(L)}) \leq 1-\delta$ folgt.

Induktiv folgt für alle $\lambda \geq 1$ zusammen mit $m_j^{(L)} \geq \delta$:

$$M_j^{(\lambda L)} - m_j^{(\lambda L)} \leq (1-\delta)^{\lambda-1}(M_j^{(L)} - m_j^{(L)}) \leq (1-\delta)^{\lambda}.$$

Wegen der Monotonie der Folgen $(m_j^{(n)})_{n\geq 1}$ und $(M_j^{(n)})_{n\geq 1}$ folgt damit die Konvergenz beider Folgen gegen ein und denselben Grenzwert, den wir mit ρ_j bezeichnen.

Beweis des Ergodensatzes (6)

Schritt 3: Exponentiell schnelle Konvergenz: $\exists \eta \in (0,1)\ \exists c > 0\ \forall j$:

$$M_j^{(n)} - m_j^{(n)} \leq c \cdot \eta^n.$$

Es sei $n = \lambda L + \nu$ mit $0 \leq \nu < L$. Damit gilt unter Verwendung von $M_j^{(n+L)} - m_j^{(n+L)} \leq (1-\delta)(M_j^{(n)} - m_j^{(n)})$:

$$\begin{aligned} M_j^{(n)} - m_j^{(n)} &= M_j^{(\lambda L+\nu)} - m_j^{(\lambda L+\nu)} \\ &\leq (1-\delta)^{\lambda}(M_j^{(\nu)} - m_j^{(\nu)}) \\ &= (1-\delta)^{n/L} \cdot (1-\delta)^{-\nu/L}(M_j^{(\nu)} - m_j^{(\nu)}) \\ &\leq \eta^n (1-\delta)^{-1}, \end{aligned}$$

wobei wir $\eta := (1-\delta)^{1/L}$ gesetzt und $(1-\delta)^{-\nu/L} \leq (1-\delta)^{-1}$ sowie $M_j^{(\nu)} - m_j^{(\nu)} \leq 1$ ausgenutzt haben. Da $\eta < 1$ ist, folgt die exponentiell schnelle Konvergenz sowohl der Maxima- und Minima- als auch damit aller Koeffizientenfolgen in den Potenzen von Π gegen die jeweiligen Grenzwerte.

Durch Grenzübergang $n \to \infty$ folgt aus $\pi_{ik}^{(n+1)} = \sum_j \pi_{ij}^{(n)} \pi_{jk}$ die Gleichung $\rho = \rho \Pi$.

Beweis des Ergodensatzes (7)

- Als Grenzwert von W-Verteilungen im $\mathbb{R}^N$ ist auch ρ eine W-Verteilung.
- Diese W-Verteilung ρ ist eindeutig bestimmt: für jede W-Verteilung ρ' mit $\rho' = \rho'\Pi$ folgt $\rho' = \rho'\Pi^n$. Durch Grenzübergang folgt $\rho' = \rho'\Pi^\infty$, woraus sich wegen

$$\rho'_k = \sum_j \rho'_j \pi_{jk}^{(\infty)} = \sum_j \rho'_j \rho_k = \rho_k$$

  sofort $\rho = \rho'$ ergibt.

Damit ist der Ergodensatz vollständig bewiesen. $\square$

Bemerkung. Indem man im obigen Satz sämtliche Matrizen und Vektoren transponiert und die für das Transponieren von Matrizenprodukten gültige Formel $(A \cdot B)^\top = B^\top \cdot A^\top$ anwendet, erhält man ein entsprechendes Resultat für spaltenstochastische Matrizen.

Google: PageRank

Google benutzt eine Modifikation der obigen Markov-Kette bezogen auf den **Google-Graphen**, das ist ein Teilgraph des Webgraphen. Die zugehörige Grenzverteilung nutzt Google zur Bewertung der Wichtigkeit einer Webseite. Diese Bewertung ist ein (wichtiger) Bestandteil des Gesamtrankings von Treffern. Genaueres zu diesem **PageRank** genannten Verfahren findet man in:

Lawrence Page, Sergey Brin, Rajeew Motwani, Terry Winograd:
The pagerank citation ranking: Bringing order to the web.
Technical Report, Stanford Digital Library Technologies Project, 1998

pdf: `http://dbpubs.stanford.edu:8090/pub/1999-66`

Stochastik

Kapitel 9: Steilkurs Statistik

Prof. Dr. Reinhard Klein
Folien von Prof. Dr. Michael Clausen

Sommersemester 2017

INSTITUT FÜR INFORMATIK II, UNIVERSITÄT BONN

W-Theorie vs. Statistik

In der **W-Theorie** geht man von **bekannten** W-Räumen, Experimenten und Verteilungen aus:

- Laplace-Räume, Bernoulli-Experimente, ...
- Verteilungen: multinomial, (hyper-)geometrisch, Poisson, Gauß, ...

und analysiert diese weiter:

- Unabhängigkeit, Erwartungswert, Varianz, Unkorreliertheit, ...
- Gesetze der großen Zahl, zentraler Grenzwertsatz, ...

In der **Statistik** sind diese W-Räume, insbesondere die zugrunde liegenden W-Maße, **nicht bekannt**, sondern müssen erst aus zufälligen Beobachtungen ermittelt oder geschätzt werden. Allerdings kennt man oft die Natur des zugrundeliegenden W-Maßes.

Es folgen zunächst einige Beispiele zur Modellierung der Schätzungsproblematik.

Typische Statistikprobleme 1

Beispiel 1: Qualitätskontrolle. Eine große Lieferung aus N Orangen, enthalte f faule und $g = N - f$ gute. N ist bekannt, aber nicht f und g.
Frage: Wie soll der Empfänger durch n Stichproben (ohne Zurücklegen) f (und damit auch g) schätzen?

Modellierung: **Hypergeometrische Verteilungen** zu den bekannten Parametern (N, n) und den unbekannten Parametern (f, g), wobei $f + g = N$.

Beispiel 2: Werfen einer Reißzwecke. Entweder landet sie mit der Spitze nach oben (o) oder mit der Spitze schräg nach unten (u).
Frage: Wie kann man durch n-maliges Werfen die Wahrscheinlichkeit p von o schätzen?

Modellierung: **Bernoulli-Verteilungen** auf $\{o, u\}^n$ mit unbekannter Erfolgswahrscheinlichkeit p.

In beiden Problemen geht es um die Schätzung der unbekannten Parameter f bzw. p, durch die dann ein sinnvolles W-Maß aus einer Familie möglicher W-Maße festgelegt ist.

Typische Statistikprobleme 2

Beispiel 3: Qualitätskontrolle. Wieder geht es um die Orangenlieferung.
Frage: Wenn vertraglich vereinbart wurde, dass der Empfänger nur dann den vollen Preis zu bezahlen hat, wenn höchstens fünf Prozent der Orangen faul sind, wie soll sich der Empfänger entscheiden, wenn er unter den n Stichproben a faule findet?

Gut wäre eine **Entscheidungsregel** der Form:

$$\begin{aligned} \text{höchstens } c \text{ Orangen faul} \quad &\Rightarrow \quad \text{Lieferung akzeptieren} \\ \text{mehr als } c \text{ Orangen faul} \quad &\Rightarrow \quad \text{Preisnachlass fordern} \end{aligned}$$

Zentrale Statistikaufgaben

Diese Beispiele führen auf folgende zentrale Aufgaben:

- **Parameterschätzung**: Aufgabe ist die Entwicklung von Methoden, mit denen man aus zufallsgesteuerten Beobachtungen auf die zugrunde liegenden, oft parameterabhängigen W-Maße schließen kann.
- **Konfidenzbereiche**: Man schätzt nicht einen Parameter sondern ein Intervall, in dem der Parameter mit sehr hoher Wahrscheinlichkeit liegt.
- **Testen von Hypothesen**: Hier geht es um statistische Entscheidungsverfahren. (z.B.: Soll die Orangenlieferung akzeptiert werden oder nicht?)

Statistische Modelle

Definition. Ein **statistisches Modell** ist ein Tripel

$$\mathcal{M} = (\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \Theta})$$

bestehend aus

- einem **Stichprobenraum** $\mathcal{X}$,
- einer σ-**Algebra** $\mathcal{A}$ auf $\mathcal{X}$ und
- einer **Familie** $(P_\theta)_{\theta \in \Theta}$ **von W-Maßen** auf $(\mathcal{X}, \mathcal{A})$, $|\Theta| \geq 2$.

- **Notationswechsel** $\mathcal{X}$ statt Ω beruht auf Vorstellung, dass die Beobachtung durch ZV $X : \Omega \to \mathcal{X}$ gegeben ist, wobei Ω eine detaillierte Beschreibung und $\mathcal{X}$ die tatsächlich beobachtbaren Ergebnisse enthält.

Konvention: Wir schreiben kurz $\mathbf{E}_\theta$ statt $\mathbf{E}_{P_\theta}$ und $\mathbf{V}_\theta$ statt $\mathbf{V}_{P_\theta}$.

Modellklassen

- Ein statistisches Modell $\mathcal{M} = (\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ heißt **parametrisches Modell**, wenn $\Theta \subset \mathbb{R}^d$ für ein $d \in \mathbb{N}$.
  Ist $d = 1$, so heißt $\mathcal{M}$ ein **einparametrisches Modell**.
- $\mathcal{M}$ heißt ein **diskretes Modell**, wenn $\mathcal{X}$ diskret, d.h. $|\mathcal{X}| \leq |\mathbb{N}|$ und $\mathcal{A} = 2^{\mathcal{X}}$ ist. Dann ist jedes P_θ durch die W-Funktion $p_\theta : x \mapsto p_\theta(x) := P_\theta(\{x\})$ bestimmt.
- $\mathcal{M}$ heißt ein **stetiges Modell**, wenn $\mathcal{X}$ eine Borel-Teilmenge eines $\mathbb{R}^n$ ist, $\mathcal{A} = \mathcal{B}^n_{\mathcal{X}}$ die auf $\mathcal{X}$ eingeschränkte Borel-σ-Algebra von $\mathbb{R}^n$, und jedes P_θ eine Dichtefunktion p_θ besitzt.
- Ist $\mathcal{M}$ diskret oder stetig, so sprechen wir von $\mathcal{M}$ als ein **Standardmodell**.

Produktmodelle

Oft werden statistische Modelle betrachtet, die die unabhängige Wiederholung von identischen Einzelexperimenten beschreiben.
Das führt zu folgender

Definition. Ist $(E, \mathcal{E}, (Q_\theta)_{\theta\in\Theta})$ ein statistisches Modell und $n \geq 2$, so heißt

$$(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta}) := (E^n, \mathcal{E}^{\otimes n}, (Q_\theta^{\otimes n})_{\theta\in\Theta})$$

das zugehörige n-**fache Produktmodell**.

In dem Fall bezeichne $X_i : \mathcal{X} \to E$ die Projektion auf die i-te Koordinate. Diese Projektion beschreibt den Ausgang des i-ten Teilexperiments. Die $X_1, \ldots, X_n$ sind dann bzgl. jedes $P_\theta = Q_\theta^{\otimes n}$ unabhängig und identisch verteilt mit Verteilung Q_θ.

Bemerkung. Das Produktmodell eines parametrischen Modells ist wieder parametrisch und das eines Standardmodells wieder ein Standardmodell.

Produktmodell: Beispiel

- In einer Fernsehshow führt der Moderator einen Apparat vor, der Zufallszahlen im Intervall $[0, \theta]$ gemäß Gleichverteilung ausspuckt, wenn er vom Moderator auf den Wert $\theta > 0$ eingestellt wurde.
- Zwei Spieler dürfen den Apparat $n = 10$ mal bedienen und sollen dann θ möglichst gut schätzen.
- Wer besser rät, hat gewonnen.

Modellierung:

- **Stichprobenraum**: $\mathcal{X} = [0, \infty)^n$,
- σ-**Algebra**: $\mathcal{A} = \mathcal{B}_{[0,\infty)}^{\otimes n}$,
- **W-Maß-Familie**: $(P_\theta)_{\theta \in \Theta} = (\mathcal{U}_{[0,\theta]}^{\otimes n})_{\theta > 0}$.

Zufallszahlen: Erwartungswert und Varianz

- Sei $\Omega = [0,\theta]^n$ und P_θ die Gleichverteilung auf Ω.
- Dann ist $\rho_\theta(\omega) := \theta^{-n}$ $(\omega \in \Omega)$ die Dichtefunktion zu P_θ.
- $X := \mathrm{id}_\Omega$ modelliert das Gesamtprotokoll der n Stichproben, während $X_i := p_i \circ X := (\Omega \ni \omega \mapsto \omega_i)$ das Ergebnis der i-ten Stichprobe beschreibt.

Satz. $\mathbf{E}_\theta(X_i) = \theta/2$ und $\mathbf{V}_\theta(X_i) = \theta^2/12$.

Beweis. Mit $\Omega_j := [0,\theta]$ ergibt sich, da ρ_θ die Dichtefunktion zu P_θ ist (siehe Kapitel 6, Folie 15):

$$\begin{aligned}
\mathbf{E}_\theta(X_i) &= \mathbf{E}_\theta(p_i \circ X) = \int_\Omega p_i(\omega)\rho_\theta(\omega)d\omega = \int_{\Omega_1} \dots \int_{\Omega_n} \omega_i \theta^{-n} d\omega_1 \dots d\omega_n \\
&= \prod_{j\neq i} \left(\int_0^\theta \frac{1}{\theta} d\omega_j \right) \cdot \int_0^\theta \frac{1}{\theta}\omega_i d\omega_i = \frac{1}{\theta}\left[\frac{\omega_i^2}{2}\right]_0^\theta = \frac{\theta}{2}. \\
\mathbf{V}_\theta(X_i) &= \mathbf{E}_\theta(X_i^2) - \mathbf{E}_\theta(X_i)^2 = \int_0^\theta \frac{1}{\theta}\omega_i^2 d\omega_i - \left(\frac{\theta}{2}\right)^2 = \frac{\theta^2}{12}. \qquad \square
\end{aligned}$$

Ratestrategien

- **Spieler A** erinnert sich an das schwache Gesetz der großen Zahl. Wegen $\mathbf{E}_\theta(\mathrm{id}_{[0,\theta]}) = \theta/2$ gilt für den doppelten Mittelwert $T_n := 2M = \frac{2}{n}\sum_{k=1}^n X_k$

$$\begin{aligned} P_\theta(|T_n - \theta| \geq \epsilon) &= P_\theta(|\frac{1}{n}\sum_{k=1}^n (X_k - \mathbf{E}_\theta(X_k))| \geq \epsilon/2) \\ &\leq \frac{\theta^2/12}{n(\epsilon/2)^2} = \frac{\theta^2}{3\epsilon^2} \cdot \frac{1}{n} \xrightarrow{n\to\infty} 0 \quad \text{für alle } \epsilon > 0. \end{aligned}$$

  Spieler A wählt T_n als Schätzer und hofft, dass dies bereits für $n = 10$ vernünftig ist.
- **Spieler B** nimmt das Beobachtungsmaximum $\tilde{T}_n := \max(X_1, \ldots, X_n)$ als Schätzer mit der Begründung, dass wegen der Unabhängigkeit von $X_1, \ldots, X_n$ und wegen $P_\theta(X_i \leq \theta - \epsilon) = (\theta - \epsilon)/\theta$ für alle $\theta > 0$ und alle $\epsilon \in (0, \theta]$ gilt:

$$P_\theta(\tilde{T}_n \leq \theta - \epsilon) = P_\theta(X_1 \leq \theta - \epsilon, \ldots, X_n \leq \theta - \epsilon) = \left(\frac{\theta - \epsilon}{\theta}\right)^n \xrightarrow{n\to\infty} 0.$$

Analyse der Ratestrategien 1

- Beide Schätzer sind **konsistent** in dem Sinne, dass für $n \to \infty$

$$T_n \xrightarrow{P_\theta} \theta \quad \text{und} \quad \tilde{T}_n \xrightarrow{P_\theta} \theta.$$

- T_n ist **erwartungstreu** (engl. unbiased) im Sinne von

$$\mathbf{E}_\theta(T_n) = \frac{2}{n} \sum_{k=1}^{n} \mathbf{E}_\theta(X_k) = \frac{2}{n} \sum_{k=1}^{n} \frac{\theta}{2} = \theta \quad \text{für alle } \theta \in \Theta.$$

- $\tilde{T}_n$ ist zwar **verzerrt** (engl. biased), aber fast erwartungstreu, denn es gilt (Beweis nächste Folie):

$$\mathbf{E}_\theta(\tilde{T}_n) = \frac{n}{n+1}\theta.$$

Analyse der Ratestrategien 2

- Wir berechnen Erwartungswert und Varianz des zweiten Schätzers.
- Wegen $P_\theta(\tilde{T}_n \leq x) = (x/\theta)^n$ $(\forall x \in [0,\theta])$ hat P_θ die Verteilungsdichte

$$\delta_\theta(x) = \frac{d}{dx}(x/\theta)^n = n \cdot \theta^{-n} x^{n-1}.$$

- Damit ist

$$\begin{aligned}
\mathbf{E}_\theta(\tilde{T}_n) &= \int_0^\theta x\delta_\theta(x)dx = n\theta^{-n}\int_0^\theta x^n dx = \frac{n}{n+1}\cdot\theta. \\
\mathbf{V}_\theta(\tilde{T}_n) &= \mathbf{E}_\theta(\tilde{T}_n^2) - \mathbf{E}_\theta(\tilde{T}_n)^2 = \int_0^\theta x^2\delta_\theta(x)dx - \left(\frac{n\theta}{n+1}\right)^2 \\
&= \frac{n}{(n+1)^2(n+2)}\cdot\theta^2. \qquad \square
\end{aligned}$$

Ein neuer Schätzer

- Dies motiviert den neuen erwartungstreuen und konsistenten Schätzer:

$$T_n^* := \frac{n+1}{n}\tilde{T}_n.$$

- Die folgende Tabelle vergleicht die Erwartungswerte, Varianzen und mittleren quadratischen Fehler der drei Schätzer:

	$\mathbf{E}_\theta(T)$	$\mathbf{V}_\theta(T)$	$\mathbf{E}_\theta((T-\theta)^2)$
T_n	θ	$\frac{\theta^2}{3n}$	$=\mathbf{V}_\theta(T_n)$
$\tilde{T}_n$	$\frac{n}{n+1}\cdot\theta$	$\frac{n\theta^2}{(n+1)^2(n+2)}$	$\frac{2\theta^2}{(n+1)(n+2)}$
T_n^*	θ	$\frac{\theta^2}{n(n+2)}$	$=\mathbf{V}_\theta(T_n^*)$

Beachte, dass

$$\mathbf{V}_\theta(T_n^*) = \mathbf{V}_\theta\left(\tfrac{n+1}{n}\tilde{T}_n\right) = \left(\tfrac{n+1}{n}\right)^2\mathbf{V}_\theta(\tilde{T}_n).$$

Vergleich der drei Schätzer

- Für die mittleren quadratischen Fehler der drei Schätzer ergibt sich im Fall $n = 10$:

$$\mathbf{E}_\theta((T_n - \theta)^2) = \frac{\theta^2}{30},$$

$$\mathbf{E}_\theta((\tilde{T}_n - \theta)^2) = \frac{\theta^2}{66},$$

$$\mathbf{E}_\theta((T_n^* - \theta)^2) = \frac{\theta^2}{120}.$$

Fazit: Strategie T_n^* ist beiden anderen Schätzern vorzuziehen.

Statistiken und Schätzer

Definition. Es sei $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta \in \Theta})$ ein statistisches Modell und $(\Sigma, \mathcal{S})$ ein Messraum.

- Eine beliebige ZV $S : (\mathcal{X}, \mathcal{A}) \to (\Sigma, \mathcal{S})$ heißt eine **Statistik**.
- Sei $\tau : \Theta \to \Sigma$ eine Abbildung, die jedem $\theta \in \Theta$ eine Kenngröße $\tau(\theta) \in \Sigma$ zuordnet. Eine Statistik $T : \mathcal{X} \to \Sigma$ heißt dann ein **Schätzer** für τ. (Oft ist $\tau = \mathrm{id}_\Theta$; T heißt dann auch Schätzer für θ.)

Bemerkung: Neue Namensgebungen (Statistik statt ZV, Schätzer statt Statistik) wegen neuer Interpretationen:

- **ZV**: beschreibt unvorhersehbare Ergebnisse;
- **Statistik**: ist eine vom Statistiker wohlkonstruierte Abbildung, die aus den Beobachtungsdaten Essentielles extrahiert.
- Statistiken gibt es viele, ein **Schätzer** ist zugeschnitten auf das Schätzen von τ.

Maximum-Likelihood-Schätzer 1

Es sei $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$ ein statistisches Standardmodell; dann ist jedes P_θ durch eine W-Funktion oder Dichte p_θ gekennzeichnet.

Idee: Wird $x \in \mathcal{X}$ beobachtet, so bestimme den Schätzwert $T(x) \in \Theta$ so, dass

$$p_{T(x)}(x) = \max_{\theta\in\Theta} p_\theta(x).$$

Bemerkung: Da im stetigen Modell $P_\theta(\{x\})$ typischerweise gleich Null ist, sind wir zu Dichten übergegangen.

Definition.

- Die Funktion $p : \mathcal{X} \times \Theta \to [0, \infty)$ mit

  $$p(x, \theta) := p_\theta(x)$$

  heißt die zugehörige **Likelihood-Funktion**.
- $p(x, \cdot) : \Theta \to [0, \infty)$ heißt die **Likelihood-Funktion zum Beobachtungswert** $x \in \mathcal{X}$.

Maximum-Likelihood-Schätzer 2

Definition.

Ein Schätzer $T : \mathcal{X} \to \Theta$ für θ heißt **Maximum-Likelihood-Schätzer**, (engl.: Maximum Likelihood Estimator), kurz: MLE, wenn für jedes $x \in \mathcal{X}$ stets $T(x)$ eine Maximalstelle von $p(x, \cdot)$ ist, d.h.:

$$p(x, T(x)) = \max_{\theta \in \Theta} p(x, \theta).$$

Bemerkung. Zur MLE-Bestimmung ist es oft bequem, mit der sogenannten **Log-Likelihood-Funktion** $\log p(x, \cdot)$ zu rechnen, die wegen der Monotonie der Logarithmus-Funktion dieselben Maximalstellen wie $p(x, \cdot)$ hat.

Dazu folgt ein Beispiel.

Schätzung der Erfolgswahrscheinlichkeit

Eine Reißzwecke falle mit unbekannter Wahrscheinlichkeit θ auf die Spitze. Gesucht ist ein Schätzer für θ bei Beobachtung von n Würfen. Als statistisches Modell wählen wir das **Binomialmodell**

$$([0:n], 2^{[0:n]}, (B_{n,\theta})_{\theta\in[0,1]})$$

mit der Likelihood-Funktion

$$p(k,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

Zur MLE-Bestimmung bei k Erfolgen betrachten wir die Nullstellen der Ableitung der Log-Likelihood-Funktion (beachte: k ist eine Konstante, θ ist die Variable!): $0 = \frac{d}{d\theta}\log p(k,\theta)$ und erhalten mit der Kettenregel und $\frac{d}{dx}\log x = \frac{1}{x}$:

$$0 = \frac{d}{d\theta}(k\log\theta + (n-k)\log(1-\theta)) = \frac{k}{\theta} - \frac{n-k}{1-\theta}.$$

Einzige Nullstelle ist $\theta = k/n$. Da die zweite Ableitung an θ kleiner Null ist, liegt dort ein Maximum vor. Also ist $T(k) := k/n$ einziger MLE für θ.

Schätzen des Zufallszahlenbereichs

Wir betrachten wieder das Beispiel vom Zufallszahlenautomaten in der Fernsehshow. Als statistisches Modell haben wird das Produktmodell

$$([0,\infty)^n, \mathcal{B}_{[0,\infty)}^{\otimes n}, (\mathcal{U}_{[0,\theta]}^{\otimes n})_{\theta>0})$$

der skalierten Gleichverteilungen gewählt. Die Likelihood-Funktion ist somit

$$p((x_1,\ldots,x_n),\theta) = \begin{cases} \theta^{-n} & \text{falls } x_1,\ldots,x_n \le \theta, \\ 0 & \text{sonst.} \end{cases}$$

Bei festem $x = (x_1,\ldots,x_n)$ ist $p(x,\theta)$ genau für $\theta = \max(x_1,\ldots,x_n)$ maximal, denn $(0,\infty) \ni \theta \mapsto \theta^{-n}$ ist streng monoton fallend. Wegen $x_1,\ldots,x_n \le \theta$ muss $\theta \ge \max(x_1,\ldots,x_n)$ sein. Also ist

$$\tilde{T}_n(x) = \max(x_1,\ldots,x_n)$$

der MLE.

Die obige Diskussion hat gezeigt, dass es hier einen besseren Schätzer gibt, nämlich T_n^*.

Alternativen zu MLE

Es gibt zahlreiche Methoden zur Parameterschätzung. Einen schnellen Überblick kann man sich bei Wikipedia verschaffen unter dem Stichwort:

Estimation Theory.

Konfidenzintervalle & Hypothesentests

- Aus Zeitgründen konnten die Themen Konfidenzintervalle & Hypothesentests leider nicht mehr diskutiert werden.
- Hier verweisen wir auf das sehr empfehlenswerte Buch:

Hans-Otto Georgii: Stochastik, 3. Auflage, de Gryter.