SS 2021

Diskrete Wahrscheinlichkeitstheorie

Susanne Albers

Fakultät für Informatik
TU München

http://wwwalbers.in.tum.de/lehre/2021SS/dwt/index.html.de

Sommersemester 2021

Kapitel 0 Organisatorisches

Es findet kein Präsenzbetrieb statt.

- Vorlesungen:
  - Die Vorlesung ist als Live-Stream verfügbar und wird gleichzeitig aufgezeichnet.
    Fr 12:00–14:00 und Fr 14:00–15:00
    Pflichtvorlesung Bachelor IN, Bioinformatik
    Modulnr.: IN0018
- Übung:
  - 2SWS Tutorübung: siehe Webseite zur Übung
  - Übungsleitung: Jens Quedenfeld und Sebastian Schubert
- Umfang:
  - 3V+2TÜ, 6 ECTS-Punkte
- Sprechstunde:
  - nach Vereinbarung

- Übungsaufgaben:
  - Ausgabe jeweils am Freitag auf der Webseite der Vorlesung, ab 18:00 Uhr
  - Abgabe eine Woche später, jeweils Montag bis 10:00 Uhr.
  - Vorbereitung in den Übung
  - vorauss. 12 Übungsblätter, das letzte am 02. Juli 2021, jedes 20 Punkte
  - Bearbeitung in Teams von zwei Studierenden
  - Bonusregelung: Werden bei den ersten sechs und zweiten sechs Übungsblättern jeweils mindestens 50% der insgesamt erreichbaren Punkte erzielt, so verbessert sich die Note einer *bestandenen* Erstklausur (03. August 2021) um 1/3 Notenstufe.
- Klausur:
  - Klausur am 03. August 2021, 14:15–16:15 Uhr
  - Wiederholungsklausur: wird noch bekannt gegeben
  - bei den Klausuren sind *keine* Hilfsmittel außer einem handbeschriebenen DIN-A4-Blatt zugelassen

- Vorkenntnisse:
  - Einführung in die Informatik I/II
  - Diskrete Strukturen
- Weiterführende Vorlesungen:
  - Effiziente Algorithmen und Datenstrukturen
  - Randomisierte Algorithmen
  - Online- und Approximationsalgorithmen
  - Komplexitätstheorie
  - ...
- Webseite:

  http://wwwalbers.in.tum.de/lehre/2021SS/dwt/index.html.de

1. Vorlesungsinhalt

- Diskrete Wahrscheinlichkeitsräume
  - Wahrscheinlichkeitsraum, Ereignis, Zufallsvariable
  - spezielle Verteilungen
  - Ungleichungen von Markov und Chebyshev
- Kontinuierliche Wahrscheinlichkeitsräume
  - Normalverteilung, Exponentialverteilung
  - Zentraler Grenzwertsatz
- Statistik
  - Schätzvariablen
  - Konfidenzintervalle
  - Testen von Hypothesen
- Stochastische Prozesse
  - Markovketten
  - Warteschlangen

2. Literatur

- T. Schickinger, A. Steger:
  Diskrete Strukturen - Band 2,
  Springer Verlag, 2001

- M. Greiner, G. Tinhofer:
  Stochastik für Informatiker,
  Carl Hanser Verlag, 1996

- H. Gordon:
  Discrete Probability,
  Springer-Verlag, 1997

- M. Mitzenmacher, E. Upfal:
  Probability and Computing: Randomized Algorithms and Probabilistic Analysis,
  Cambridge University Press, 2005

R. Motwani, P. Raghavan:
Randomized Algorithms,
Cambridge University Press, 1995

M. Hofri:
Probabilistic Analysis of Algorithms,
Springer Verlag, 1987

L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz:
Statistik - Der Weg zur Datenanalyse,
Springer-Verlag, 1997

3. Einleitung

Was bedeutet Zufall?

- Unkenntnis über den Ausgang eines durchgeführten Experiments
- Ein Experiment wird vielfach mit eventuell sich änderndem Ergebnis ausgeführt
- Ereignisse stehen in keinem kausalen Zusammenhang
- physikalischer Zufall (Rauschen, Kernzerfall)

Zufall in der *diskreten* Informatik

- Die Eingabe für einen bestimmten Algorithmus wird aus einer großen Menge möglicher Eingaben zufällig gewählt:

  average case

- Kombination von Worst-Case- und Average-Case-Analyse, in der Eingaben gemäß einer Verteilung leicht pertubiert werden:

  smoothed analysis

- Der Algorithmus verwendet Zufallsbits, um mit großer Wahrscheinlichkeit gewisse Problemsituationen zu vermeiden:

  Randomisierung

Kapitel I Diskrete Wahrscheinlichkeitsräume

1. Grundlagen

Definition 1

1. Ein diskreter Wahrscheinlichkeitsraum ist durch eine Ergebnismenge $\Omega = \{\omega_1, \omega_2, \ldots\}$ von Elementarereignissen gegeben.
2. Jedem Elementarereignis ω_i ist eine (Elementar-)Wahrscheinlichkeit $\Pr[\omega_i]$ zugeordnet, wobei wir fordern, dass $0 \leq \Pr[\omega_i] \leq 1$ und

$$\sum_{\omega \in \Omega} \Pr[\omega] = 1.$$

3. Eine Menge $E \subseteq \Omega$ heißt Ereignis. Die Wahrscheinlichkeit $\Pr[E]$ eines Ereignisses ist durch

$$\Pr[E] := \sum_{\omega \in E} \Pr[\omega]$$

definiert.

Beispiel 2

Zwei faire Würfel (einer weiß, einer schwarz) werden geworfen. Wir sind an der Gesamtzahl der angezeigten Augen interessiert:

$$
\begin{aligned}
\Omega \quad = \{ \quad & (1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\
& (2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\
& (3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\
& (4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\
& (5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\
& (6,1),(6,2),(6,3),(6,4),(6,5),(6,6) \quad \}
\end{aligned}
$$

1. Die Wahrscheinlichkeit $\Pr((i,j))$ eines jeden Elementarereignisses (i,j) ist $\frac{1}{36}$.

2. Die Wahrscheinlichkeit $\Pr(E)$ des Ereignisses

$$E = \{\text{Die Gesamtzahl der Augen ist } 10\}$$

   ist $\frac{1}{12}$.

Wir hätten aber auch sagen können:

$$\Omega = \{2, 3, 4, \ldots, 10, 11, 12\}$$

Die Wahrscheinlichkeiten der Elementarereignisse sind dann aber nicht mehr gleich. Es ist z.B.

1. $\Pr(2) = \frac{1}{36}$;
2. $\Pr(4) = \frac{1}{12}$;
3. $\Pr(7) = \frac{1}{6}$.

Beispiel 3

Eine faire Münze wird so lange geworfen, bis die gleiche Seite zweimal hintereinander fällt. Dann ist

$$\Omega = \{\text{hh, tt, htt, thh, thtt, hthh, hthtt, ththh}, \ldots\}$$

Frage: Was sind die Wahrscheinlichkeiten der einzelnen Elementarereignisse?

$\bar{E}$ heißt komplementäres Ereignis zu E.
Allgemein verwenden wir bei der Definition von Ereignissen alle bekannten Operatoren aus der Mengenlehre. Wenn also A und B Ereignisse sind, dann sind auch $A \cup B$, $A \cap B$, $A \setminus B$ etc. Ereignisse.
Zwei Ereignisse A und B heißen disjunkt oder auch unvereinbar, wenn $A \cap B = \emptyset$ gilt.

Definition 4

$$\text{relative Häufigkeit von } E := \frac{\text{absolute Häufigkeit von } E}{\text{Anzahl aller Beobachtungen}} = \frac{\text{Anzahl Eintreten von } E}{\text{Anzahl aller Beobachtungen}}.$$

Definition 5

Ein Wahrscheinlichkeitsraum mit $\Omega = \{\omega_1, \ldots, \omega_n\}$ heißt endlicher Wahrscheinlichkeitsraum.

Bei unendlichen Wahrscheinlichkeitsräumen werden wir gewöhnlich nur den Fall $\Omega = \mathbb{N}_0$ betrachten. Dies stellt keine große Einschränkung dar, da wir statt einer Ergebnismenge $\Omega = \{\omega_1, \omega_2, \ldots\}$ auch $\mathbb{N}_0$ als Ergebnismenge verwenden können, indem wir ω_i mit $i - 1$ identifizieren. Wir sagen, dass durch die Angabe der Elementarwahrscheinlichkeiten ein Wahrscheinlichkeitsraum auf Ω definiert ist.

Beispiel 6

Wir beobachten die an einer Straße in Bayern vorbeifahrenden Autos. Dabei gelte:

1. Es fahren doppelt so viele Autos von links nach rechts wie von rechts nach links.
2. Von zehn Autos haben zwei die Farbe hellelfenbein, die übrigen eine andere Lackierung.

- Das Ereignis "*Wir beobachten ein von links nach rechts fahrendes Auto*" hat die Wahrscheinlichkeit $\frac{2}{3}$.
- Das Ereignis "*Das nächste Auto ist ein Taxi von rechts*" passiert mit Wahrscheinlichkeit

$$\frac{1}{3} \cdot \frac{1}{5}\,.$$

Beispiel 7 (Unendlicher Wahrscheinlichkeitsraum)

Wir betrachten eine Münze, die mit Wahrscheinlichkeit p Kopf zeigt und mit Wahrscheinlichkeit $q := 1 - p$ Zahl.

Wir führen Versuche aus, indem wir die Münze wiederholt solange werfen, bis *Zahl* fällt. Das *Ergebnis* eines solchen Versuchs ist die Anzahl der durchgeführten Münzwürfe.

Damit ergibt sich hier als Ergebnismenge

$$\Omega = \mathbb{N} = \{1, 2, 3, \ldots\} \,.$$

Beispiel 7 (Forts.)

Sei, für $i \in \mathbb{N}$, ω_i das Elementarereignis

$$\omega_i \;\widehat{=}\; \text{Die Münze wird } i\text{-mal geworfen} \; .$$

Dann gilt:

$$\Pr[\omega_i] = p^{i-1}q \; ,$$

und

$$\sum_{\omega \in \Omega} \Pr[\omega] = \sum_{i=1}^{\infty} p^{i-1}q = q \cdot \sum_{i=0}^{\infty} p^i = \frac{q}{1-p} = 1 \; .$$

(wie es sein soll!)

Lemma 8

Für Ereignisse $A, B, A_1, A_2, \ldots$ *gilt:*

1. $\Pr[\emptyset] = 0$, $\Pr[\Omega] = 1$.
2. $0 \le \Pr[A] \le 1$.
3. $\Pr[\bar{A}] = 1 - \Pr[A]$.
4. *Wenn* $A \subseteq B$, *so folgt* $\Pr[A] \le \Pr[B]$.

Lemma 8 (Forts.)

5. *(Additionssatz) Wenn die Ereignisse $A_1, \ldots, A_n$ paarweise disjunkt sind (also wenn für alle Paare $i \neq j$ gilt, dass $A_i \cap A_j = \emptyset$), so folgt*

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} \Pr[A_i].$$

Für disjunkte Ereignisse A, B erhalten wir insbesondere

$$\Pr[A \cup B] = \Pr[A] + \Pr[B]\ .$$

Für eine unendliche Menge von disjunkten Ereignissen $A_1, A_2, \ldots$ gilt analog

$$\Pr\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \Pr[A_i]\ .$$

Beweis:

Die Aussagen folgen unmittelbar aus Definition 1, den Eigenschaften der Addition und der Definition der Summe. $\square$

Eigenschaft 5 in Lemma 8 gilt nur für disjunkte Ereignisse. Für den allgemeinen Fall erhalten wir folgenden

Satz 9 (Siebformel, Prinzip der Inklusion/Exklusion)

Für Ereignisse $A_1, \ldots, A_n$ $(n \geq 2)$ gilt:

$$
\begin{aligned}
\Pr\left[\bigcup_{i=1}^{n} A_i\right] \quad &= \quad \sum_{i=1}^{n} \Pr[A_i] - \sum_{1 \leq i_1 < i_2 \leq n} \Pr[A_{i_1} \cap A_{i_2}] + - \ldots \\
&+ \quad (-1)^{l-1} \sum_{1 \leq i_1 < \ldots < i_l \leq n} \Pr[A_{i_1} \cap \ldots \cap A_{i_l}] + - \ldots \\
&+ \quad (-1)^{n-1} \cdot \Pr[A_1 \cap \ldots \cap A_n] \,.
\end{aligned}
$$

Satz 9 (Forts.)

Insbesondere gilt für zwei Ereignisse A und B

$$\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]\ .$$

Für drei Ereignisse A_1, A_2 und A_3 erhalten wir

$$\begin{aligned}\Pr[A_1 \cup A_2 \cup A_3] \quad = \quad & \Pr[A_1] + \Pr[A_2] + \Pr[A_3] \\ & -\Pr[A_1 \cap A_2] - \Pr[A_1 \cap A_3] \\ & -\Pr[A_2 \cap A_3] \\ & +\Pr[A_1 \cap A_2 \cap A_3]\ .\end{aligned}$$

Beweis:

Wir betrachten zunächst den Fall $n = 2$. Dazu setzen wir $C := A \setminus B = A \setminus (A \cap B)$. Gemäß dieser Definition gilt, dass C und $A \cap B$ sowie C und B disjunkt sind. Deshalb können wir Eigenschaft 5 von Lemma 8 anwenden:

$$\Pr[A] = \Pr[C \cup (A \cap B)] = \Pr[C] + \Pr[A \cap B]\,.$$

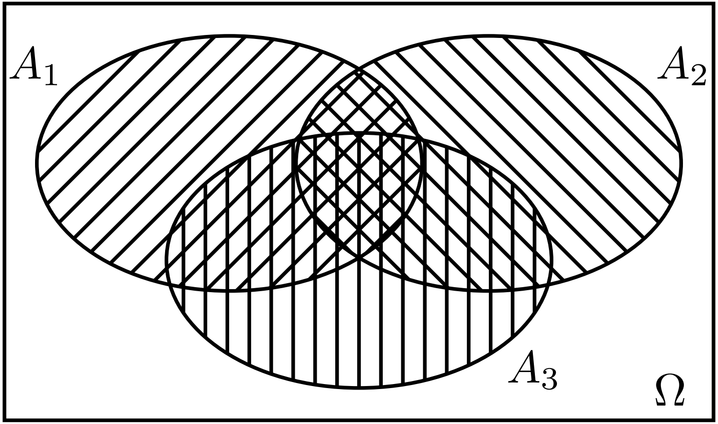
Wegen $A \cup B = C \cup B$ folgt daraus

$$\begin{aligned}\Pr[A \cup B] = \Pr[C \cup B] = \Pr[C] + \Pr[B] = \\ \Pr[A] - \Pr[A \cap B] + \Pr[B]\end{aligned}$$

und wir haben die Behauptung für $n = 2$ gezeigt.

Beweis (Forts.):

Der Fall $n = 3$:



Man beachte, dass durch die im Satz angegebene Summe jedes Flächenstück insgesamt genau einmal gezählt wird.

Beweis (Forts.):

Der allgemeine Fall kann nun durch Induktion über n gezeigt werden (was wir aber hier nicht ausführen!).

Satz 9 findet man manchmal auch unter der Bezeichung *Satz von Poincaré-Sylvester*, nach dem Franzosen

Jules Henri Poincaré (1854–1912)

und dem Engländer

James Joseph Sylvester (1814–1897)

benannt.

Boolesche Ungleichung:

Die folgende Abschätzung ist nach George Boole (1815–1864) benannt:

Korollar 10

Für Ereignisse $A_1, \ldots, A_n$ gilt

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} \Pr[A_i]\,.$$

Analog gilt für eine unendliche Folge von Ereignissen $A_1, A_2, \ldots$, dass

$$\Pr\left[\bigcup_{i=1}^{\infty} A_i\right] \leq \sum_{i=1}^{\infty} \Pr[A_i]\,.$$

Beweis:

Zunächst betrachten wir die linke Seite der Ungleichung für den endlichen Fall und erhalten

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{\omega \in \bigcup_{i=1}^{n} A_i} \Pr[\omega]\,.$$

Für die rechte Seite gilt

$$\sum_{i=1}^{n} \Pr[A_i] = \sum_{i=1}^{n} \sum_{\omega \in A_i} \Pr[\omega]\,.$$

Jedes Elementarereignis kommt links also genau einmal und rechts mindestens einmal vor. $\square$

1.1 Wahl der Wahrscheinlichkeiten

Frage: Wie können Wahrscheinlichkeiten sinnvoll festgelegt werden?

Prinzip von Laplace (Pierre Simon Laplace (1749–1827)): *Wenn nichts dagegen spricht, gehen wir davon aus, dass alle Elementarereignisse gleich wahrscheinlich sind.*

Also:

$$\Pr[E] = \frac{|E|}{|\Omega|}$$

1.2 Historische Anfänge der Wahrscheinlichkeitstheorie

Die ersten Hinweise auf mathematische Untersuchungen zu Problemen der Wahrscheinlichkeitstheorie finden sich in einem Briefwechsel zwischen den französischen Mathematikern

Pierre Fermat (1601–1665)

und

Blaise Pascal (1623–1662).

Pascal beschäftigte sich neben der Mathematik auch mit Fragestellungen aus dem Bereich der Physik und auch aus der Informatik! Sein Vater hatte als Steuerinspektor in Rouen umfangreiche Rechnungen durchzuführen und so wurde Pascal zum Bau einer mechanischen Rechenmaschine, der so genannten *Pascaline*, motiviert.

In dem Briefwechsel taucht bereits der Ansatz $\Pr[E] = |E|/|\Omega|$ zur Berechnung der Wahrscheinlichkeit von E auf. Auch den Begriff des Erwartungswerts kann man dort schon finden. Weder Fermat noch Pascal publizierten ihre Überlegungen zur Wahrscheinlichkeitstheorie. Der Niederländer

Christiaan Huygens (1629–1695)

entwickelte ebenfalls Methoden zum Arbeiten mit Wahrscheinlichkeiten aus. Er publizierte im Jahre 1657 auch eine kleine Arbeit mit dem Titel „De ratiociniis in ludo aleae“ (Über die Gesetzmäßigkeiten beim Würfelspiel).

2. Bedingte Wahrscheinlichkeiten

Beispiel 11

A und B spielen Poker (52 Karten, 5 Karten pro Spieler, keine getauschten Karten). A hält vier Asse und eine Herz Zwei in der Hand. B kann dieses Blatt nur überbieten, wenn er einen Straight Flush (fünf Karten *einer* Farbe in aufsteigender Reihenfolge) hat. Die Wahrscheinlichkeit für das Ereignis $F :=$ „B hat einen Straight Flush" beträgt

$$\Pr[F] = \frac{|F|}{|\Omega|} = \frac{3 \cdot 8 + 7}{\binom{52-5}{5}} = \frac{31}{1533939} = 2{,}02.. \cdot 10^{-5}.$$

Beispiel 11 (Forts.)

A hat die Karten allerdings gezinkt und weiß, dass B nur Kreuz in der Hand hält. Bezeichne nun Ω' den Wahrscheinlichkeitsraum aller Möglichkeiten für B und F' das Ereignis, dass B einen Straight Flush der Farbe Kreuz hat:

$$\Pr[F'] = \frac{|F'|}{|\Omega'|} = \frac{8}{\binom{12}{5}} = \frac{8}{792} \approx 0{,}01 \;!!$$

Für $\Pr[A|B]$ erforderliche Eigenschaften:

1. $\Pr[B|B] = 1$;
2. $\Pr[A|\Omega] = \Pr[A]$;
3. für festes B ist $\Pr[A|B]$ proportional zu $\Pr[A \cap B]$.

Definition 12

A und B seien Ereignisse mit $\Pr[B] > 0$. Die bedingte Wahrscheinlichkeit $\Pr[A|B]$ von A gegeben B ist definiert als

$$\Pr[A|B] := \frac{\Pr[A \cap B]}{\Pr[B]} \,.$$

Die bedingten Wahrscheinlichkeiten $\Pr[\cdot|B]$ bilden für ein beliebiges Ereignis $B \subseteq \Omega$ mit $\Pr[B] > 0$ einen neuen Wahrscheinlichkeitsraum über Ω.
Es ist leicht nachzurechnen, dass dadurch die Definition eines *diskreten Wahrscheinlichkeitsraums* erfüllt ist:

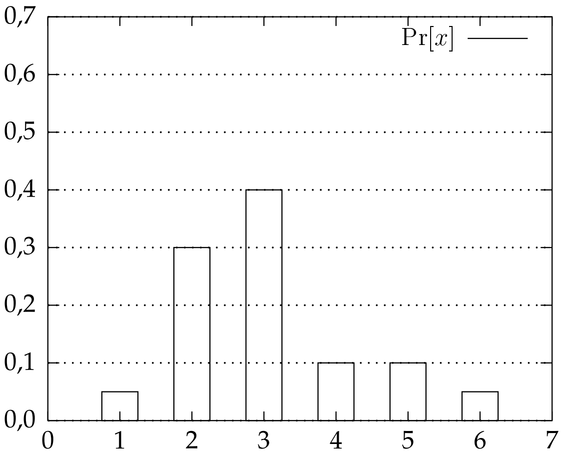
$$\sum_{\omega\in\Omega} \Pr[\omega|B] = \sum_{\omega\in\Omega} \frac{\Pr[\omega \cap B]}{\Pr[B]} = \sum_{\omega\in B} \frac{\Pr[\omega]}{\Pr[B]} = \frac{\Pr[B]}{\Pr[B]} = 1.$$

Damit gelten alle Rechenregeln für Wahrscheinlichkeiten auch für bedingte Wahrscheinlichkeiten. Beispielsweise:

$$\Pr[\emptyset|B] = 0 \text{ sowie } \Pr[\bar{A}|B] = 1 - \Pr[A|B]\ .$$

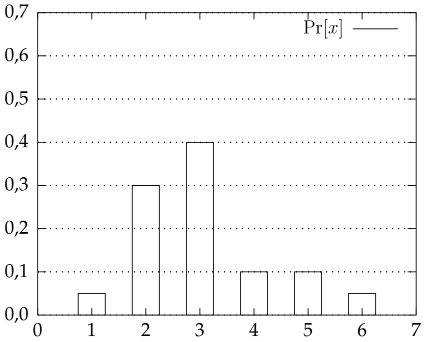
Beispiel 13 (Reskalierung bei bedingten Wahrscheinlichkeiten)

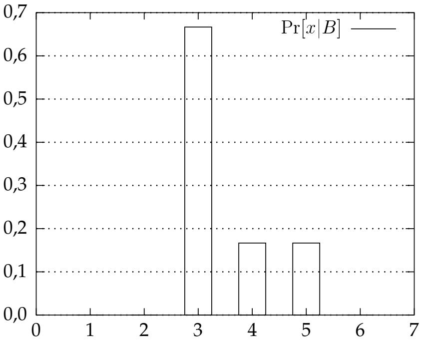
Betrachte folgenden gezinkten Würfel:



Beispiel 13 (Forts.)

Wir betrachten nun den durch $B := \{3, 4, 5\}$ gegebenen bedingten Wahrscheinlichkeitsraum:





Was genau war die Bedingung?

Beispiel 14 (Zweikinderproblem)

Wir nehmen an, dass bei der Geburt eines Kindes beide Geschlechter gleich wahrscheinlich sind. Wir wissen, dass eine bestimmte Familie zwei Kinder hat und eines davon ein Mädchen ist. Wie groß ist die Wahrscheinlichkeit, dass beide Kinder der Familie Mädchen sind?

Natürlich $\frac{1}{2}$.

Wirklich?

Beispiel 14 (Forts.)

Eigentlich gilt:

$$\Omega := \{mm, mj, jm, jj\}$$

und

$$M := \{mm, mj, jm\}\ .$$

Wir bedingen auf M, und damit gilt für $A := \{mm\}$:

$$\Pr[A|M] = \frac{\Pr[A \cap M]}{\Pr[M]} = \frac{1/4}{3/4} = \frac{1}{3}.$$

Beispiel 15 (Ziegenproblem)

Sie nehmen an einer Spielshow im Fernsehen teil, bei der Sie eine von drei verschlossenen Türen auswählen sollen. Hinter einer Tür wartet der Preis, ein Auto, hinter den beiden anderen stehen Ziegen. Sie zeigen auf eine Tür, sagen wir Nummer eins. Sie bleibt vorerst geschlossen. Der Moderator weiß, hinter welcher Tür sich das Auto befindet; mit den Worten "Ich gebe Ihnen mal einen kleinen Hinweis" öffnet er eine andere Tür, zum Beispiel Nummer drei, und eine Ziege schaut heraus und meckert. Er fragt: "Bleiben Sie bei Nummer eins, oder wählen sie Nummer zwei? "

Frage: Welche Strategie ist günstiger:

- S1 Der Spieler bleibt immer bei seiner ursprünglichen Wahl.
- S2 Der Spieler wechselt stets die ausgewählte Tür.

Beispiel (Forts.)

Wir betrachten hier eine Diskussion des Ziegenproblems mit Hilfe von bedingten Wahrscheinlichkeiten. Wir betrachten bei jeder Variante den Fall, dass der Spieler

a) *die "richtige",*

b) *eine falsche Tür gewählt hat.*

Ersteres geschieht mit Wahrscheinlichkeit $\frac{1}{3}$, *Letzteres mit Wahrscheinlichkeit* $\frac{2}{3}$.
Mit der vom Moderator gegebenen Information ergeben sich für die beiden Strategien die folgenden Gewinnwahrscheinlichkeiten:

	S1	**S2**
a)	*?*	*?*
b)	*?*	*?*

Häufig verwendet man die Definition der bedingten Wahrscheinlichkeit in der Form

$$\Pr[A \cap B] = \Pr[B|A] \cdot \Pr[A] = \Pr[A|B] \cdot \Pr[B] \,. \tag{1}$$

Damit:

Satz 16 (Multiplikationssatz)

Seien die Ereignisse $A_1, \ldots, A_n$ *gegeben. Falls* $\Pr[A_1 \cap \ldots \cap A_n] > 0$ *ist, gilt*

$$\Pr[A_1 \cap \ldots \cap A_n] = \Pr[A_1] \cdot \Pr[A_2|A_1] \cdot \Pr[A_3|A_1 \cap A_2] \cdot \ldots \cdot \Pr[A_n|A_1 \cap \ldots \cap A_{n-1}] \,.$$

Beweis:

Zunächst halten wir fest, dass alle bedingten Wahrscheinlichkeiten wohldefiniert sind, da $\Pr[A_1] \geq \Pr[A_1 \cap A_2] \geq \ldots \geq \Pr[A_1 \cap \ldots \cap A_n] > 0$.

Die rechte Seite der Aussage im Satz können wir umschreiben zu

$$\frac{\Pr[A_1]}{1} \cdot \frac{\Pr[A_1 \cap A_2]}{\Pr[A_1]} \cdot \frac{\Pr[A_1 \cap A_2 \cap A_3]}{\Pr[A_1 \cap A_2]} \cdot \ldots \cdot \frac{\Pr[A_1 \cap \ldots \cap A_n]}{\Pr[A_1 \cap \ldots \cap A_{n-1}]}.$$

Offensichtlich kürzen sich alle Terme bis auf $\Pr[A_1 \cap \ldots \cap A_n]$. $\square$

Beispiel 17 (Geburtstagsproblem)

Wie groß ist die Wahrscheinlichkeit, dass in einer m-köpfigen Gruppe zwei Personen am selben Tag Geburtstag haben?

Umformulierung:
Man werfe m Bälle zufällig und gleich wahrscheinlich in n Körbe. Wie groß ist die Wahrscheinlichkeit, dass nach dem Experiment jeder Ball allein in seinem Korb liegt?

Für das Geburtstagsproblem: $n = 365$

Offensichtlich muss $m \leq n$ sein, damit überhaupt jeder Ball allein in einem Korb liegen kann.
Wir nehmen an, dass die Bälle nacheinander geworfen werden. A_i bezeichne das Ereignis „Ball i landet in einem noch leeren Korb“. Das gesuchte Ereignis „Alle Bälle liegen allein in einem Korb“ bezeichnen wir mit A. Nach Satz 16 können wir $\Pr[A]$ berechnen durch

$$\begin{aligned}\Pr[A] &= \Pr\left[\cap_{i=1}^m A_i\right] \\ &= \Pr[A_1] \cdot \Pr[A_2|A_1] \cdot \ldots \cdot \Pr[A_m| \cap_{i=1}^{m-1} A_i].\end{aligned}$$

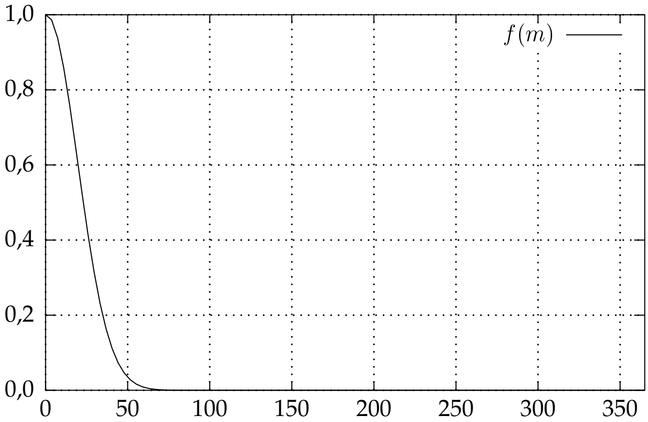
Unter der Bedingung, dass die ersten $j-1$ Bälle jeweils in einem leeren Korb gelandet sind, bedeutet A_j, dass der j-te Ball in eine der $n-(j-1)$ leeren Körbe fallen muss, die aus Symmetriegründen jeweils mit derselben Wahrscheinlichkeit gewählt werden.

Daraus folgt

$$\Pr[A_j | \cap_{i=1}^{j-1} A_i] = \frac{n-(j-1)}{n} = 1 - \frac{j-1}{n}.$$

Mit der Abschätzung $1 - x \leq e^{-x}$ und wegen $\Pr[A_1] = 1$ erhalten wir

$$\begin{aligned}
\Pr[A] &= \prod_{j=1}^{m} \left(1 - \frac{j-1}{n}\right) \\
&\leq \prod_{j=2}^{m} e^{-(j-1)/n} = e^{-(1/n)\cdot\sum_{j=1}^{m-1} j} \\
&= e^{-m(m-1)/(2n)} =: f(m) \;.
\end{aligned}$$



Verlauf von $f(m)$ für $n = 365$

Ausgehend von der Darstellung der bedingten Wahrscheinlichkeit in Gleichung 1 zeigen wir:

Satz 18 (Satz von der totalen Wahrscheinlichkeit)

Die Ereignisse $A_1, \ldots, A_n$ *seien paarweise disjunkt und es gelte* $B \subseteq A_1 \cup \ldots \cup A_n$. *Dann folgt*

$$\Pr[B] = \sum_{i=1}^{n} \Pr[B|A_i] \cdot \Pr[A_i]\ .$$

Analog gilt für paarweise disjunkte Ereignisse $A_1, A_2, \ldots$ *mit* $B \subseteq \bigcup_{i=1}^{\infty} A_i$*, dass*

$$\Pr[B] = \sum_{i=1}^{\infty} \Pr[B|A_i] \cdot \Pr[A_i]\ .$$

Beweis:

Wir zeigen zunächst den endlichen Fall. Wir halten fest, dass

$$B = (B \cap A_1) \cup \ldots \cup (B \cap A_n)\ .$$

Da für beliebige i, j mit $i \neq j$ gilt, dass $A_i \cap A_j = \emptyset$, sind auch die Ereignisse $B \cap A_i$ und $B \cap A_j$ disjunkt. Wegen (1) folgt $\Pr[B \cap A_i] = \Pr[B|A_i] \cdot \Pr[A_i]$ (auch für den Fall, dass $\Pr[A_i] = 0$!). Wir wenden nun den Additionssatz (Lemma 8, Teil 5) an

$$\Pr[B] = \Pr[B \cap A_1] + \ldots + \Pr[B \cap A_n] =$$
$$\Pr[B|A_1] \cdot \Pr[A_1] + \ldots + \Pr[B|A_n] \cdot \Pr[A_n]$$

und haben damit die Behauptung gezeigt. Da der Additionssatz auch für unendlich viele Ereignisse $A_1, A_2, \ldots$ gilt, kann dieser Beweis direkt auf den unendlichen Fall übertragen werden. □

Mit Hilfe von Satz 18 erhalten wir leicht einen weiteren nützlichen Satz:

Satz 19 (Satz von Bayes)

Die Ereignisse $A_1, \ldots, A_n$ *seien paarweise disjunkt, mit* $\Pr[A_j] > 0$ *für alle* j*. Ferner sei* $B \subseteq A_1 \cup \ldots \cup A_n$ *ein Ereignis mit* $\Pr[B] > 0$*. Dann gilt für ein beliebiges* $i = 1, \ldots, n$

$$\Pr[A_i|B] = \frac{\Pr[A_i \cap B]}{\Pr[B]} = \frac{\Pr[B|A_i] \cdot \Pr[A_i]}{\sum_{j=1}^{n} \Pr[B|A_j] \cdot \Pr[A_j]} \,.$$

Analog gilt für paarweise disjunkte Ereignisse $A_1, A_2, \ldots$ *mit* $B \subseteq \bigcup_{i=1}^{\infty} A_i$*, dass*

$$\Pr[A_i|B] = \frac{\Pr[A_i \cap B]}{\Pr[B]} = \frac{\Pr[B|A_i] \cdot \Pr[A_i]}{\sum_{j=1}^{\infty} \Pr[B|A_j] \cdot \Pr[A_j]} \,.$$

Mit dem Satz von Bayes dreht man gewissermaßen die Reihenfolge der Bedingung um. Gegeben die Wahrscheinlichkeit von B unter den Bedingungen A_i (sowie die Wahrscheinlichkeiten der A_i selbst), berechnet man die Wahrscheinlichkeit von A_i bedingt auf das Ereignis B.

Thomas Bayes (1702–1761) war ein bekannter Theologe und Mitglied der Royal Society. Als sein bedeutendstes Werk gilt sein Beitrag zur Wahrscheinlichkeitstheorie „Essay Towards Solving a Problem in the Doctrine of Chances". Diese Arbeit wurde erst 1763 publiziert.

3. Unabhängigkeit

Bei einer bedingten Wahrscheinlichkeit $\Pr[A|B]$ kann der Fall auftreten, dass die Bedingung auf B, also das Vorwissen, dass B eintritt, keinen Einfluss auf die Wahrscheinlichkeit hat, mit der wir das Eintreten von A erwarten. Es gilt also $\Pr[A|B] = \Pr[A]$, und wir nennen dann die Ereignisse A und B unabhängig.

Beispiel 20 (Zweimaliges Würfeln)

$$\Omega := \{(i,j) \mid 1 \le i,j \le 6\} \;.$$

Alle Elementarereignisse erhalten nach dem Prinzip von Laplace die Wahrscheinlichkeit $\frac{1}{36}$.
Wir definieren die Ereignisse

$$\begin{aligned} A &:= \text{Augenzahl im ersten Wurf ist gerade,} \\ B &:= \text{Augenzahl im zweiten Wurf ist gerade,} \\ C &:= \text{Summe der Augenzahlen beider Würfe beträgt 7.} \end{aligned}$$

Es gilt $\Pr[A] = \Pr[B] = \frac{1}{2}$ und $\Pr[C] = \frac{1}{6}$. Wie groß ist $\Pr[B|A]$?

Beispiel 20 (Forts.)

Nach unserer Intuition beeinflusst der Ausgang des ersten Wurfs den zweiten Wurf nicht. Daher gewinnen wir durch das Eintreten von A keine Information in Bezug auf das Ereignis B hinzu:

$$B \cap A = \{(2,2), (2,4), (2,6), (4,2), (4,4), (4,6), (6,2), (6,4), (6,6)\}.$$

Daraus folgt

$$\Pr[B|A] = \frac{\Pr[B \cap A]}{\Pr[A]} = \frac{\frac{9}{36}}{\frac{1}{2}} = \frac{1}{2} = \Pr[B]\,.$$

Das Eintreffen des Ereignisses B hat mit dem Ereignis A „nichts zu tun".

Definition 21

Die Ereignisse A und B heißen unabhängig, wenn gilt

$$\Pr[A \cap B] = \Pr[A] \cdot \Pr[B] \ .$$

Falls $\Pr[B] \neq 0$, so können wir diese Definition zu

$$\Pr[A] = \frac{\Pr[A \cap B]}{\Pr[B]} = \Pr[A|B]$$

umschreiben.

Beispiel 20 (Zweimaliges Würfeln, Forts.)

Zur Erinnerung:

$$\begin{aligned} A &:= \text{Augenzahl im ersten Wurf ist gerade,}\\ B &:= \text{Augenzahl im zweiten Wurf ist gerade,}\\ C &:= \text{Summe der Augenzahlen beider Würfe beträgt 7.} \end{aligned}$$

Bei den Ereignissen A und B ist die Unabhängigkeit klar, da offensichtlich kein kausaler Zusammenhang zwischen den Ereignissen besteht. Wie steht es mit A und C?

$$A \cap C = \{(2,5),(4,3),(6,1)\}$$

und damit

$$\Pr[A \cap C] = \frac{3}{36} = \frac{1}{2} \cdot \frac{1}{6} = \Pr[A] \cdot \Pr[C] \text{ bzw. } \Pr[C|A] = \Pr[C]\,.$$

Beispiel 20 (Forts.)

Also sind auch A und C (und analog B und C) unabhängig.

Bemerkung: Im Beispiel ist $A \cap C \neq \emptyset$.
Es gilt sogar allgemein für zwei unabhängige Ereignisse A und B mit $\Pr[A], \Pr[B] > 0$, dass sie gar nicht disjunkt sein können, da ansonsten

$$0 = \Pr[\emptyset] = \Pr[A \cap B] \neq \Pr[A] \cdot \Pr[B] \; .$$

Beispiel 20 (Zweimaliges Würfeln (Forts.))

Zur Erinnerung:

$$A := \text{Augenzahl im ersten Wurf ist gerade,}$$
$$B := \text{Augenzahl im zweiten Wurf ist gerade,}$$
$$C := \text{Summe der Augenzahlen beider Würfe beträgt 7.}$$

Wir betrachten das Ereignis $A \cap B \cap C$. Wenn $A \cap B$ eintritt, so sind beide gewürfelten Augenzahlen gerade und somit ergibt auch die Summe davon eine gerade Zahl. Daraus folgt $\Pr[A \cap B \cap C] = 0$ bzw. $\Pr[C|A \cap B] = 0 \neq \Pr[C]$. Das Ereignis $A \cap B$ liefert uns also Information über das Ereignis C.

Definition 22

Die paarweise verschiedenen Ereignisse $A_1, \ldots, A_n$ heißen unabhängig, wenn für alle Teilmengen $I = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$ mit $i_1 < i_2 < \ldots < i_k$ gilt, dass

$$\Pr[A_{i_1} \cap \ldots \cap A_{i_k}] = \Pr[A_{i_1}] \cdot \ldots \cdot \Pr[A_{i_k}]. \tag{2}$$

Eine unendliche Familie von paarweise verschiedenen Ereignissen A_i mit $i \in \mathbb{N}$ heißt unabhängig, wenn (2) für jede endliche Teilmenge $I \subseteq \mathbb{N}$ erfüllt ist.

Lemma 23

Die (paarweise verschiedenen) Ereignisse $A_1, \ldots, A_n$ *sind genau dann unabhängig, wenn für alle* $(s_1, \ldots, s_n) \in \{0,1\}^n$ *gilt, dass*

$$\Pr[A_1^{s_1} \cap \ldots \cap A_n^{s_n}] = \Pr[A_1^{s_1}] \cdot \ldots \cdot \Pr[A_n^{s_n}], \tag{3}$$

wobei $A_i^0 = \bar{A}_i$ *und* $A_i^1 = A_i$.

Beweis:

Zunächst zeigen wir, dass aus (2) die Bedingung (3) folgt. Wir beweisen dies durch Induktion über die Anzahl der Nullen in $s_1, \ldots, s_n$. Wenn $s_1 = \ldots = s_n = 1$ gilt, so ist nichts zu zeigen. Andernfalls gelte ohne Einschränkung $s_1 = 0$. Aus dem Additionssatz folgt dann

$$\begin{aligned}\Pr[\bar{A}_1 \cap A_2^{s_2} \cap \ldots \cap A_n^{s_n}] &= \Pr[A_2^{s_2} \cap \ldots \cap A_n^{s_n}] \\ &\quad -\Pr[A_1 \cap A_2^{s_2} \cap \ldots \cap A_n^{s_n}].\end{aligned}$$

Darauf können wir die Induktionsannahme anwenden und erhalten

$$\begin{aligned}&\Pr[\bar{A}_1 \cap A_2^{s_2} \cap \ldots \cap A_n^{s_n}] \\ &= \Pr[A_2^{s_2}] \cdot \ldots \cdot \Pr[A_n^{s_n}] - \Pr[A_1] \cdot \Pr[A_2^{s_2}] \cdot \ldots \cdot \Pr[A_n^{s_n}] \\ &= (1 - \Pr[A_1]) \cdot \Pr[A_2^{s_2}] \cdot \ldots \cdot \Pr[A_n^{s_n}],\end{aligned}$$

woraus die Behauptung wegen $1 - \Pr[A_1] = \Pr[\bar{A}_1]$ folgt.

Beweis (Forts.):

Für die Gegenrichtung zeigen wir nur, dass aus (3) $\Pr[A_1 \cap A_2] = \Pr[A_1] \cdot \Pr[A_2]$ folgt. Es gilt wegen des Satzes von der totalen Wahrscheinlichkeit, dass

$$\begin{aligned}
\Pr[A_1 \cap A_2] &= \sum_{s_3,\ldots,s_n \in \{0,1\}} \Pr[A_1 \cap A_2 \cap A_3^{s_3} \cap \ldots \cap A_n^{s_n}] \\
&= \sum_{s_3,\ldots,s_n \in \{0,1\}} \Pr[A_1] \cdot \Pr[A_2] \cdot \Pr[A_3^{s_3}] \cdot \ldots \cdot \Pr[A_n^{s_n}] \\
&= \Pr[A_1] \cdot \Pr[A_2] \cdot \sum_{s_3=0,1} \Pr[A_3^{s_3}] \cdot \ldots \cdot \sum_{s_n=0,1} \Pr[A_n^{s_n}] \\
&= \Pr[A_1] \cdot \Pr[A_2],
\end{aligned}$$

und es folgt die Behauptung. $\square$

Aus der Darstellung in Lemma 23 folgt die wichtige Beobachtung, dass für zwei unabhängige Ereignisse A und B auch die Ereignisse $\bar{A}$ und B (und analog auch A und $\bar{B}$ bzw. $\bar{A}$ und $\bar{B}$) unabhängig sind!

Ebenso folgt:

Lemma 24

Seien A, B und C unabhängige Ereignisse. Dann sind auch $A \cap B$ und C bzw. $A \cup B$ und C unabhängig.

Beweis:

Die Unabhängigkeit von $A \cap B$ und C folgt unmittelbar aus Definition 22.
Aus

$$\begin{aligned}
\Pr[(A \cup B) \cap C] &= \Pr[(A \cap C) \cup (B \cap C)] \\
&= \Pr[A \cap C] + \Pr[B \cap C] - \Pr[A \cap B \cap C] \\
&= \Pr[C] \cdot (\Pr[A] + \Pr[B] - \Pr[A \cap B]) \\
&= \Pr[A \cup B] \cdot \Pr[C]
\end{aligned}$$

folgt die Unabhängigkeit von $A \cup B$ und C. $\square$

4. Zufallsvariablen

4.1 Grundlagen

Anstatt der Ereignisse selbst sind wir oft an „Auswirkungen“ oder „Merkmalen“ der (Elementar)Ereignisse interessiert.

Definition 25

Sei ein Wahrscheinlichkeitsraum auf der Ergebnismenge Ω gegeben. Eine Abbildung

$$X : \Omega \to \mathbb{R}$$

heißt (numerische) Zufallsvariable.
Eine Zufallsvariable X über einer endlichen oder abzählbar unendlichen Ergebnismenge Ω heißt diskret.

Bei diskreten Zufallsvariablen ist der Wertebereich

$$W_X := X(\Omega) = \{x \in \mathbb{R};\ \exists\, \omega \in \Omega \text{ mit } X(\omega) = x\}$$

ebenfalls wieder endlich (bzw. abzählbar unendlich).

Beispiel 26

Wir werfen eine ideale Münze drei Mal. Als Ergebnismenge erhalten wir $\Omega := \{H, T\}^3$. Die Zufallsvariable Y bezeichne die Gesamtanzahl der Würfe mit Ergebnis „Head".

Beispielsweise gilt also $Y(HTH) = 2$ und $Y(HHH) = 3$. Y hat den Wertebereich $W_Y = \{0, 1, 2, 3\}$.

Für $W_X = \{x_1, \ldots, x_n\}$ bzw. $W_X = \{x_1, x_2, \ldots\}$ betrachten wir (für ein beliebiges $1 \leq i \leq n$ bzw. $x_i \in \mathbb{R}$) das Ereignis

$$A_i := \{\omega \in \Omega;\ X(\omega) = x_i\} = X^{-1}(x_i).$$

Bemerkung: Anstelle von $\Pr[X^{-1}(x_i)]$ verwendet man häufig auch die Schreibweise $\Pr[„X = x_i“]$. Analog setzt man

$$\begin{aligned}\Pr[„X \leq x_i“] &= \sum_{x \in W_X : x \leq x_i} \Pr[„X = x“] \\ &= \Pr[\{\omega \in \Omega;\ X(\omega) \leq x_i\}]\,.\end{aligned}$$

Oft lässt man auch die Anführungszeichen weg.

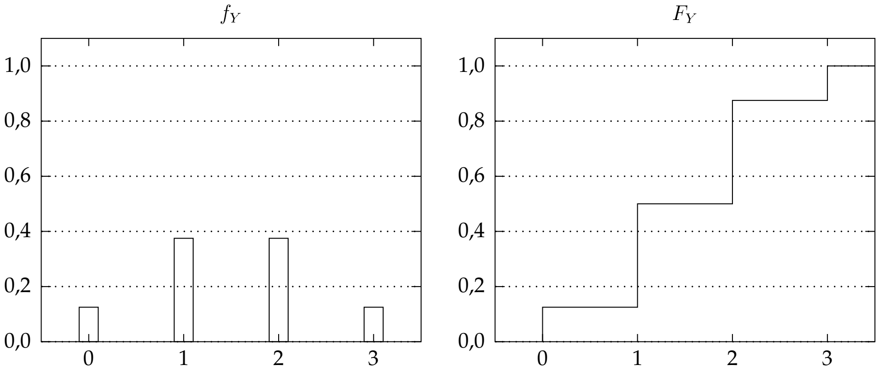
Definition 27

- Die Funktion
$$f_X : \mathbb{R} \ni x \mapsto \Pr[X = x] \in [0,1] \tag{4}$$
nennt man (diskrete) Dichte(funktion) der Zufallsvariablen X.
- Die Funktion
$$F_X : \mathbb{R} \ni x \mapsto \Pr[X \le x] = \sum_{x' \in W_X \,:\, x' \le x} \Pr[X = x'] \in [0,1] \tag{5}$$
heißt Verteilung(sfunktion) der Zufallsvariablen X.

Beispiel 28

Für die Zufallsvariable Y erhalten wir

$$\begin{aligned}
\Pr[Y=0] &= \Pr[TTT] = \frac{1}{8},\\
\Pr[Y=1] &= \Pr[HTT] + \Pr[THT] + \Pr[TTH] = \frac{3}{8},\\
\Pr[Y=2] &= \Pr[HHT] + \Pr[HTH] + \Pr[THH] = \frac{3}{8},\\
\Pr[Y=3] &= \Pr[HHH] = \frac{1}{8}.
\end{aligned}$$



Dichte und Verteilung von Y

Bemerkung: Man kann statt Ω auch den zugrunde liegenden Wahrscheinlichkeitsraum über W_X betrachten.

4.2 Erwartungswert und Varianz

Definition 29

Zu einer Zufallsvariablen X definieren wir den Erwartungswert $\mathbb{E}[X]$ durch

$$\mathbb{E}[X] := \sum_{x \in W_X} x \cdot \Pr[X = x] = \sum_{x \in W_X} x \cdot f_X(x) \,,$$

sofern $\sum_{x \in W_X} |x| \cdot \Pr[X = x]$ konvergiert.

Beispiel 30

$$\begin{aligned}
\mathbb{E}[Y] &= \sum_{i=0}^{3} i \cdot \Pr[Y = i] \\
&= 1 \cdot \Pr[Y = 1] \; + \; 2 \cdot \Pr[Y = 2] \; + \; 3 \cdot \Pr[Y = 3] \\
&= 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} \; = \; \frac{3}{2} \,.
\end{aligned}$$

Beispiel 31

Eine Münze wird so lange geworfen, bis sie zum ersten Mal „Head“ zeigt. Sei k die Anzahl der durchgeführten Würfe. Wenn k ungerade ist, zahlt der Spieler an die Bank k Euro. Andernfalls (k gerade) zahlt die Bank k Euro an den Spieler.

$$G := \begin{cases} k & \text{falls } k \text{ ungerade,} \\ -k & \text{falls } k \text{ gerade.} \end{cases}$$

Wie schon gesehen, gilt dann

$$\Pr[\text{„Anzahl Würfe} = k\text{“}] = (1/2)^k \,.$$

Damit erhalten wir

$$\mathbb{E}[G] = \sum_{k=1}^{\infty} (-1)^{k-1} \cdot k \cdot \left(\frac{1}{2}\right)^k \,.$$

Da

$$\sum_{k=1}^{\infty} |(-1)^{k-1} \cdot k| \cdot \left(\frac{1}{2}\right)^k \le \sum_{k=1}^{\infty} k \cdot \left(\frac{1}{2}\right)^k ,$$

existiert der Erwartungswert $\mathbb{E}[G]$.
Es gilt

$$\begin{aligned}
\mathbb{E}[G] &= \sum_{j=1}^{\infty} \left[(2j-1) \cdot \left(\frac{1}{2}\right)^{2j-1} - 2j \cdot \left(\frac{1}{2}\right)^{2j} \right] \\
&= \sum_{j=1}^{\infty} \left(\frac{1}{2}\right)^{2j-1} \cdot [(2j-1) - j] \\
&= \frac{1}{2} \cdot \sum_{j=1}^{\infty} (j-1) \cdot \left(\frac{1}{4}\right)^{j-1} = \frac{1}{2} \cdot \frac{\frac{1}{4}}{\left(1-\frac{1}{4}\right)^2} = \frac{2}{9} .
\end{aligned}$$

Wird jedoch, um das Risiko zu steigern, der zu zahlende Betrag von k Euro jeweils auf 2^k Euro erhöht, also

$$G' := \begin{cases} 2^k & \text{falls } k \text{ ungerade,} \\ -2^k & \text{falls } k \text{ gerade ,} \end{cases}$$

dann existiert $\mathbb{E}[G']$ nicht, da

$$\begin{aligned} \mathbb{E}[G'] &= \sum_{k=1}^{\infty} (-1)^{k-1} \cdot 2^k \cdot \left(\frac{1}{2}\right)^k \\ &= \sum_{k=1}^{\infty} (-1)^{k-1} = +1 - 1 + 1 - 1 + - \ldots . \end{aligned}$$

Berechnung des Erwartungswerts:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x \in W_X} x \cdot \Pr[X = x] = \sum_{x \in W_X} x \cdot f_X(x)\\ &= \sum_{x \in W_X} x \sum_{\omega \in \Omega : X(\omega) = x} \Pr[\omega]\\ &= \sum_{\omega \in \Omega} X(\omega) \cdot \Pr[\omega]\,.\end{aligned}$$

Bei unendlichen Wahrscheinlichkeitsräumen ist dabei analog zur Definition des Erwartungswerts erforderlich, dass $\sum_{\omega \in \Omega} |X(\omega)| \cdot \Pr[\omega]$ konvergiert (absolute Konvergenz).

Satz 32 (Monotonie des Erwartungswerts)

Seien X und Y Zufallsvariablen über dem Wahrscheinlichkeitsraum Ω mit $X(\omega) \leq Y(\omega)$ für alle $\omega \in \Omega$. Dann gilt $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

Beweis:

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \cdot \Pr[\omega] \leq \sum_{\omega \in \Omega} Y(\omega) \cdot \Pr[\omega] = \mathbb{E}[Y]\,.$$

$\square$

Aus Satz 32 folgt insbesondere, dass $a \leq \mathbb{E}[X] \leq b$ gilt, wenn für die Zufallsvariable X die Eigenschaft $a \leq X(\omega) \leq b$ für alle $\omega \in \Omega$ erfüllt ist.

4.2.1 Rechenregeln für den Erwartungswert

Oft betrachtet man eine Zufallsvariable X nicht direkt, sondern wendet noch eine Funktion darauf an:

$$Y := f(X) = f \circ X\,,$$

wobei $f : \mathcal{D} \to \mathbb{R}$ eine beliebige Funktion sei mit $W_X \subseteq \mathcal{D} \subseteq \mathbb{R}$.

Beobachtung: $f(X)$ ist wieder eine Zufallsvariable.

Aus

$$\Pr[Y = y] = \Pr[\{\omega \mid f(X(\omega)) = y\}] \;=\; \sum_{x\,:\,f(x)=y} \Pr[X = x]$$

folgt

$$\begin{aligned}
\mathbb{E}[f(X)] = \mathbb{E}[Y] &= \sum_{y \in W_Y} y \cdot \Pr[Y = y] \\
&= \sum_{y \in W_Y} y \cdot \sum_{x\,:\,f(x)=y} \Pr[X = x] = \sum_{x \in W_X} f(x) \cdot \Pr[X = x] \\
&= \sum_{\omega \in \Omega} f(X(\omega)) \cdot \Pr[\omega]\,.
\end{aligned}$$

Satz 33 (Linearität des Erwartungswerts, einfache Version)

Für eine beliebige Zufallsvariable X und $a, b \in \mathbb{R}$ gilt

$$\mathbb{E}[a \cdot X + b] = a \cdot \mathbb{E}[X] + b\,.$$

Beweis:

$$\begin{aligned}
\mathbb{E}[a \cdot X + b] &= \sum_{x \in W_X} (a \cdot x + b) \cdot \Pr[X = x] \\
&= a \cdot \sum_{x \in W_X} x \cdot \Pr[X = x] + b \cdot \sum_{x \in W_X} \Pr[X = x] \\
&= a \cdot \mathbb{E}[X] + b\,.
\end{aligned}$$

$\square$

Satz 34

Sei X eine Zufallsvariable mit $W_X \subseteq \mathbb{N}_0$. Dann gilt

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \Pr[X \geq i]\,.$$

Beweis:

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} i \cdot \Pr[X = i] = \sum_{i=0}^{\infty} \sum_{j=1}^{i} \Pr[X = i]$$
$$= \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} \Pr[X = i] = \sum_{j=1}^{\infty} \Pr[X \geq j]\,.$$

$\square$

Definition 35

Sei X eine Zufallsvariable und A ein Ereignis mit $\Pr[A] > 0$. Die bedingte Zufallsvariable $X|A$ besitzt die Dichte

$$f_{X|A}(x) := \Pr[X = x \mid A] = \frac{\Pr[„X = x“ \cap A]}{\Pr[A]}\,.$$

Die Definition von $f_{X|A}$ ist zulässig, da

$$\sum_{x \in W_X} f_{X|A}(x) = \sum_{x \in W_X} \frac{\Pr[„X = x“ \cap A]}{\Pr[A]} = \frac{\Pr[A]}{\Pr[A]} = 1\,.$$

Der Erwartungswert $\mathbb{E}[X|A]$ der Zufallsvariablen $X|A$ berechnet sich entsprechend:

$$\mathbb{E}[X|A] = \sum_{x \in W_X} x \cdot f_{X|A}(x)\,.$$

Satz 36

Sei X *eine Zufallsvariable. Für paarweise disjunkte Ereignisse* $A_1, \ldots, A_n$ *mit* $A_1 \cup \ldots \cup A_n = \Omega$ *und* $\Pr[A_1], \ldots, \Pr[A_n] > 0$ *gilt*

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X|A_i] \cdot \Pr[A_i]\,.$$

Für paarweise disjunkte Ereignisse $A_1, A_2, \ldots$ *mit* $\bigcup_{i=1}^{\infty} A_k = \Omega$ *und* $\Pr[A_1]$*,* $\Pr[A_2], \ldots > 0$ *gilt analog*

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{E}[X|A_i] \cdot \Pr[A_i],$$

sofern die Erwartungswerte auf der rechten Seite alle existieren und die Summe $\sum_{i=1}^{\infty} |\mathbb{E}[X|A_i]| \cdot \Pr[A_i]$ *konvergiert.*

Beweis:

$$\mathbb{E}[X] = \sum_{x \in W_X} x \cdot \Pr[X = x] = \sum_{x \in W_X} x \cdot \sum_{i=1}^{n} \Pr[X = x | A_i] \cdot \Pr[A_i]$$
$$= \sum_{i=1}^{n} \Pr[A_i] \sum_{x \in W_X} x \cdot \Pr[X = x | A_i] = \sum_{i=1}^{n} \Pr[A_i] \cdot \mathbb{E}[X | A_i].$$

Der Beweis für den unendlichen Fall verläuft analog. $\square$

Beispiel 37

Wir werfen eine Münze so lange, bis zum ersten Mal „Kopf“ erscheint. Dies geschehe in jedem Wurf unabhängig mit Wahrscheinlichkeit p. Wir definieren dazu die Zufallsvariable $X :=$ „Anzahl der Würfe“. Wir haben bereits gesehen, dass

$$\Pr[X = k] = p(1-p)^{k-1}$$

und damit

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k \cdot p(1-p)^{k-1} = p \cdot \frac{1}{(1-(1-p))^2} = \frac{1}{p}\,.$$

Beispiel 37

Andere Berechnungsmethode: (gestützt auf Satz 36)
Definiere das Ereignis

$$K_1 := \text{„Im ersten Wurf fällt Kopf"}\,.$$

Offensichtlich gilt $\mathbb{E}[X|K_1] = 1$.
Nehmen wir nun an, dass im ersten Wurf *nicht* „Kopf" gefallen ist. Wir starten das Experiment neu.

Beispiel 37

Sei X' die Anzahl der Würfe bis zum ersten Auftreten von „Kopf“ im neu gestarteten Experiment. Wegen der Gleichheit der Experimente gilt $\mathbb{E}[X'] = \mathbb{E}[X]$. Damit schließen wir

$$\mathbb{E}[X|\bar{K}_1] = 1 + \mathbb{E}[X'] = 1 + \mathbb{E}[X]$$

und erhalten mit Satz 36:

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[X|K_1] \cdot \Pr[K_1] + \mathbb{E}[X|\bar{K}_1] \cdot \Pr[\bar{K}_1] \\
&= 1 \cdot p + (1 + \mathbb{E}[X]) \cdot (1 - p)\,.
\end{aligned}$$

Daraus ergibt sich wiederum $\mathbb{E}[X] = 1/p$.

4.2.2 Varianz

Wir betrachten die beiden folgenden Zufallsexperimente:

1. Wir würfeln (mit einem fairen Würfel), bei gerader Augenzahl erhalten wir 1 Euro, bei ungerader Augenzahl müssen wir 1 Euro bezahlen.
2. Wir würfeln (mit einem fairen Würfel), bei 6 Augen erhalten wir 5 Euro, ansonsten müssen wir 1 Euro bezahlen.

Beobachtung:

In beiden Fällen ist der erwartete Gewinn $= 0$.

Dennoch sind die „Schwankungen" im ersten Fall geringer als im zweiten.

Eine nahe liegende Lösung wäre,

$$\mathbb{E}[|X - \mu|]$$

zu berechnen, wobei $\mu = \mathbb{E}[X]$ sei. Dies scheitert jedoch meist an der „unhandlichen" Betragsfunktion. Aus diesem Grund betrachtet man stattdessen $\mathbb{E}[(X - \mu)^2]$, also die quadratische Abweichung vom Erwartungswert.

Definition 38

Für eine Zufallsvariable X mit $\mu = \mathbb{E}[X]$ definieren wir die *Varianz* $\mathrm{Var}[X]$ durch

$$\mathrm{Var}[X] := \mathbb{E}[(X - \mu)^2] = \sum_{x \in W_X} (x - \mu)^2 \cdot \Pr[X = x]\,.$$

Die Größe $\sigma := \sqrt{\mathrm{Var}[X]}$ heißt *Standardabweichung* von X.

Satz 39

Für eine beliebige Zufallsvariable X gilt

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2\,.$$

Beweis:

Sei $\mu := \mathbb{E}[X]$. Nach Definition gilt

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2 - 2\mu \cdot X + \mu^2] \\
&= \sum_{x \in W_X} (x^2 - 2\mu \cdot x + \mu^2) \cdot \Pr[X = x] \\
&= \sum_{x \in W_X} x^2 \cdot \Pr[X = x] - \sum_{x \in W_X} 2\mu \cdot x \cdot \Pr[X = x] + \sum_{x \in W_X} \mu^2 \cdot \Pr[X = x] \\
&= \mathbb{E}[X^2] - 2\mu \cdot \mathbb{E}[X] + \mu^2 \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2\,.
\end{aligned}$$

Beispiel 40

1. Wir würfeln (mit einem fairen Würfel), bei gerader Augenzahl erhalten wir 1 Euro, bei ungerader Augenzahl müssen wir 1 Euro bezahlen. Es ist

$$\mu = 0 \text{ und } \mathrm{Var}[X] = \frac{1}{2} \cdot 1^2 + \frac{1}{2} \cdot (-1)^2 = 1\,.$$

2. Wir würfeln (mit einem fairen Würfel), bei 6 Augen erhalten wir 5 Euro, ansonsten müssen wir 1 Euro bezahlen.
   Es ist

$$\mu = 0 \text{ und } \mathrm{Var}[X] = \frac{1}{6} \cdot 5^2 + \frac{5}{6} \cdot (-1)^2 = 5\,.$$

Satz 41

Für eine beliebige Zufallsvariable X und $a, b \in \mathbb{R}$ gilt

$$\mathrm{Var}[a \cdot X + b] = a^2 \cdot \mathrm{Var}[X]\,.$$

Beweis:

Aus der in Satz 33 gezeigten Linearität des Erwartungswerts folgt $\mathbb{E}[Y+b] = \mathbb{E}[Y]+b$. Zusammen mit der Definition der Varianz ergibt sich damit sofort

$$\mathrm{Var}[Y+b] = \mathbb{E}[(Y+b-\mathbb{E}[Y+b])^2] = \mathbb{E}[(Y-\mathbb{E}[Y])^2] = \mathrm{Var}[Y]\,.$$

Weiter folgt mit Satz 39:

$$\mathrm{Var}[a\cdot X] = \mathbb{E}[(aX)^2] - \mathbb{E}[aX]^2 = a^2\mathbb{E}[X^2] - (a\mathbb{E}[X])^2 = a^2\cdot \mathrm{Var}[X]\,,$$

und daraus zusammen die Behauptung. $\square$

Der Erwartungswert und die Varianz gehören zu den so genannten Momenten einer Zufallsvariablen:

Definition 42

Für eine Zufallsvariable X nennen wir $\mathbb{E}[X^k]$ das k-te Moment und $\mathbb{E}[(X-\mathbb{E}[X])^k]$ das k-te zentrale Moment.

Der Erwartungswert ist also identisch zum ersten Moment, während die Varianz dem zweiten zentralen Moment entspricht.

4.3 Mehrere Zufallsvariablen

Beispiel 43

Aus einem Skatblatt mit 32 Karten ziehen wir zufällig eine Hand von zehn Karten sowie einen Skat von zwei Karten. Unter den Karten gibt es vier Buben. Die Zufallsvariable X zählt die Anzahl der Buben in der Hand, während Y die Anzahl der Buben im Skat angibt. Die Werte von X und Y hängen offensichtlich stark voneinander ab. Beispielsweise muss $Y = 0$ sein, wenn $X = 4$ gilt.

Wie kann man mit mehreren Zufallsvariablen über demselben Wahrscheinlichkeitsraum rechnen, auch wenn sie, wie im obigen Beispiel, sehr voneinander abhängig sind?
Wir untersuchen Wahrscheinlichkeiten der Art

$$\Pr[X = x, Y = y] = \Pr[\{\omega;\ X(\omega) = x, Y(\omega) = y\}]\,.$$

Beispiel 44

Wenn wir nur die Zufallsvariable X betrachten, so gilt für $0 \leq x \leq 4$

$$\Pr[X = x] = \frac{\binom{4}{x}\binom{28}{10-x}}{\binom{32}{10}}.$$

Allgemein nennt man Zufallsvariablen mit der Dichte

$$\Pr[X = x] = \frac{\binom{b}{x}\binom{a}{r-x}}{\binom{a+b}{r}}$$

hypergeometrisch verteilt. Durch diese Dichte wird ein Experiment modelliert, bei dem r Elemente ohne Zurücklegen aus einer Grundmenge der Mächtigkeit $a + b$ mit b besonders ausgezeichneten Elementen gezogen werden.

Beispiel 44 (Forts.)

Die Zufallsvariable Y ist für sich gesehen ebenfalls hypergeometrisch verteilt mit $b = 4$, $a = 28$ und $r = 2$.
Für X und Y zusammen gilt jedoch z.B.

$$\Pr[X = 4, Y = 1] = 0,$$

und allgemein

$$\Pr[X = x, Y = y] = \frac{\binom{4}{x}\binom{28}{10-x}\binom{4-x}{y}\binom{28-(10-x)}{2-y}}{\binom{32}{10}\binom{22}{2}}\,.$$

Bemerkung: Die Schreibweise $\Pr[X = x, Y = y]$ stellt eine Abkürzung von $\Pr[„X = x \wedge Y = y“]$ dar. Ein anderes Beispiel ist

$$\Pr[X \le x, Y \le y_1, \sqrt{Y} = y_2]\,.$$

Die Funktion

$$f_{X,Y}(x,y) := \Pr[X = x, Y = y]$$

heißt gemeinsame Dichte der Zufallsvariablen X und Y.
Aus der gemeinsamen Dichte $f_{X,Y}$ kann man ableiten

$$f_X(x) = \sum_{y \in W_Y} f_{X,Y}(x,y) \quad \text{bzw.} \quad f_Y(y) = \sum_{x \in W_X} f_{X,Y}(x,y)\,.$$

Die Funktionen f_X und f_Y nennt man Randdichten.

Die Ereignisse „$Y = y$“ bilden eine Partitionierung des Wahrscheinlichkeitsraumes, und es gilt daher

$$\Pr[X = x] = \sum_{y \in W_Y} \Pr[X = x, Y = y] = f_X(x)\,.$$

Die Dichten der einzelnen Zufallsvariablen entsprechen also genau den Randdichten.

Für zwei Zufallsvariablen definiert man die gemeinsame Verteilung

$$\begin{aligned} F_{X,Y}(x, y) &= \Pr[X \le x, Y \le y] = \Pr[\{\omega;\ X(\omega) \le x, Y(\omega) \le y\}] \\ &= \sum_{x' \le x} \sum_{y' \le y} f_{X,Y}(x', y')\,. \end{aligned}$$

Die Randverteilung ergibt sich gemäß

$$F_X(x) = \sum_{x' \le x} f_X(x') = \sum_{x' \le x} \sum_{y \in W_Y} f_{X,Y}(x', y)$$

sowie

$$F_Y(y) = \sum_{y' \le y} f_Y(y') = \sum_{y' \le y} \sum_{x \in W_X} f_{X,Y}(x, y')\,.$$

4.3.1 Unabhängigkeit von Zufallsvariablen

Definition 45

Die Zufallsvariablen $X_1, \ldots, X_n$ heißen unabhängig, wenn für alle $(x_1, \ldots, x_n) \in W_{X_1} \times \ldots \times W_{X_n}$ gilt

$$\Pr[X_1 = x_1, \ldots, X_n = x_n] = \Pr[X_1 = x_1] \cdot \ldots \cdot \Pr[X_n = x_n]\,.$$

Alternativ:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdot \ldots \cdot f_{X_n}(x_n)\,.$$

Bei unabhängigen Zufallsvariablen ist also die gemeinsame Dichte gleich dem Produkt der Randdichten. Ebenso gilt

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdot \ldots \cdot F_{X_n}(x_n)\,.$$

Satz 46

Seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen und $S_1, \ldots, S_n$ beliebige Mengen mit $S_i \subseteq W_{X_i}$. Dann sind die Ereignisse „$X_1 \in S_1$", ..., „$X_n \in S_n$" unabhängig.

Beweis:

$$
\begin{aligned}
&\Pr[X_1 \in S_1, \ldots, X_n \in S_n] \\
&= \sum_{x_1 \in S_1} \ldots \sum_{x_n \in S_n} \Pr[X_1 = x_1, \ldots, X_n = x_n] \\
&\overset{\text{Unabh.}}{=} \sum_{x_1 \in S_1} \ldots \sum_{x_n \in S_n} \Pr[X_1 = x_1] \cdot \ldots \cdot \Pr[X_n = x_n] \\
&= \left(\sum_{x_1 \in S_1} \Pr[X_1 = x_1] \right) \cdot \ldots \cdot \left(\sum_{x_n \in S_n} \Pr[X_n = x_n] \right) \\
&= \Pr[X_1 \in S_1] \cdot \ldots \cdot \Pr[X_n \in S_n] \,.
\end{aligned}
$$

$\square$

Satz 47

Seien $f_1, \ldots, f_n$ reellwertige Funktionen ($f_i : \mathbb{R} \to \mathbb{R}$ für $i = 1, \ldots, n$). Wenn die Zufallsvariablen $X_1, \ldots, X_n$ unabhängig sind, dann gilt dies auch für $f_1(X_1), \ldots, f_n(X_n)$.

Beweis:

Sei $z_i \in W_{f_i(X_i)}$ für $i = 1, \ldots, n$ und $S_i = \{x;\ f_i(x) = z_i\}$.

$$
\begin{aligned}
&\Pr[f_1(X_1) = z_1, \ldots, f_n(X_n) = z_n] \\
&= \Pr[X_1 \in S_1, \ldots, X_n \in S_n] \\
&\overset{\text{Unabh.}}{=} \Pr[X_1 \in S_1] \cdot \ldots \cdot \Pr[X_n \in S_n] \\
&= \Pr[f_1(X_1) = z_1] \cdot \ldots \cdot \Pr[f_n(X_n) = z_n]\,.
\end{aligned}
$$

$\square$

4.3.2 Zusammengesetzte Zufallsvariablen

Beispiel 48

Ein Würfel werde zweimal geworfen. X bzw. Y bezeichne die Augenzahl im ersten bzw. zweiten Wurf. Sei $Z := X + Y$ die Summe der gewürfelten Augenzahlen.

Für Z gilt z.B.:
$\Pr[Z = 1] = \Pr[\emptyset] = 0$, $\Pr[Z = 4] = \Pr[\{(1,3),(2,2),(3,1)\}] = \frac{3}{36}$.

Für die Verteilung der Summe zweier unabhängiger Zufallsvariablen gilt der folgende Satz:

Satz 49

Für zwei unabhängige Zufallsvariablen X und Y sei $Z := X + Y$. Es gilt

$$f_Z(z) = \sum_{x \in W_X} f_X(x) \cdot f_Y(z - x)\,.$$

Beweis:

Mit Hilfe des Satzes von der totalen Wahrscheinlichkeit folgt, dass

$$\begin{aligned} f_Z(z) = \Pr[Z = z] &= \sum_{x \in W_X} \Pr[X + Y = z \mid X = x] \cdot \Pr[X = x] \\ &= \sum_{x \in W_X} \Pr[Y = z - x] \cdot \Pr[X = x] \\ &= \sum_{x \in W_X} f_X(x) \cdot f_Y(z - x)\,. \end{aligned}$$

□

Den Ausdruck $\sum_{x \in W_X} f_X(x) \cdot f_Y(z - x)$ aus Satz 49 nennt man in Analogie zu den entsprechenden Begriffen bei Potenzreihen auch Faltung oder Konvolution der Dichten f_X und f_Y.

Beispiel (Forts.)

Berechne die Dichte von $Z = X + Y$:

$$\Pr[Z = z] = \sum_{x \in W_X} \Pr[X = x] \cdot \Pr[Y = z - x]$$

$$= \sum_{x=1}^{6} \frac{1}{6} \cdot \Pr[Y = z - x] = \sum_{x=\max\{1,z-6\}}^{\min\{6,z-1\}} \frac{1}{36}.$$

Für $2 \le z \le 7$ erhalten wir

$$\Pr[Z = z] = \sum_{i=1}^{z-1} \frac{1}{36} = \frac{z-1}{36}.$$

Und für $7 < z \le 12$:

$$\Pr[Z = z] = \frac{13-z}{36}.$$

4.3.3 Momente zusammengesetzter Zufallsvariablen

Satz 50 (Linearität des Erwartungswerts)

Für Zufallsvariablen $X_1, \ldots, X_n$ *und* $X := a_1 X_1 + \cdots + a_n X_n$ *mit* $a_1, \ldots, a_n \in \mathbb{R}$ *gilt*

$$\mathbb{E}[X] = a_1 \mathbb{E}[X_1] + \cdots + a_n \mathbb{E}[X_n]\,.$$

Beweis:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{\omega \in \Omega} (a_1 \cdot X_1(\omega) + \ldots + a_n \cdot X_n(\omega)) \cdot \Pr[\omega] \\
&= a_1 \cdot \left(\sum_{\omega \in \Omega} X_1(\omega) \cdot \Pr[\omega] \right) + \cdots + a_n \cdot \left(\sum_{\omega \in \Omega} X_n(\omega) \cdot \Pr[\omega] \right) \\
&= a_1 \cdot \mathbb{E}[X_1] + \ldots + a_n \cdot \mathbb{E}[X_n]\,.
\end{aligned}$$

□

Beispiel 51

n betrunkene Seeleute torkeln nach dem Landgang in ihre Kojen. Sie haben völlig die Orientierung verloren, weshalb wir annehmen, dass jede Zuordnung der Seeleute zu den n Betten gleich wahrscheinlich ist (genau ein Seemann pro Bett). Wie viele Seeleute liegen im Mittel im richtigen Bett?

Die Anzahl der Seeleute im richtigen Bett zählen wir mit der Zufallsvariablen X, die als Summe der Zufallsvariablen $X_1, \ldots, X_n$ dargestellt wird, wobei

$$X_i := \begin{cases} 1 & \text{falls Seemann } i \text{ in seinem Bett liegt,} \\ 0 & \text{sonst.} \end{cases}$$

Offenbar gilt $X := X_1 + \cdots + X_n$.

Beispiel 51

Für die Variablen X_i erhalten wir $\Pr[X_i = 1] = \frac{1}{n}$, da jedes Bett von Seemann i mit gleicher Wahrscheinlichkeit aufgesucht wird.
Daraus folgt

$$\mathbb{E}[X_i] = 0 \cdot \Pr[X_i = 0] \;+\; 1 \cdot \Pr[X_i = 1] = \frac{1}{n},$$

und somit

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \frac{1}{n} = 1\,.$$

Im Mittel hat also nur ein Seemann sein eigenes Bett aufgesucht.

Satz 52 (Multiplikativität des Erwartungswerts)

Für unabhängige Zufallsvariablen $X_1, \ldots, X_n$ *gilt*

$$\mathbb{E}[X_1 \cdot \cdots \cdot X_n] = \mathbb{E}[X_1] \cdot \cdots \cdot \mathbb{E}[X_n]\,.$$

Beweis:

Wir beweisen den Fall $n = 2$. Der allgemeine Fall ist analog.

$$\begin{aligned}
\mathbb{E}[X \cdot Y] &= \sum_{x \in W_X} \sum_{y \in W_Y} xy \cdot \Pr[X = x, Y = y] \\
&\overset{\text{Unabh.}}{=} \sum_{x \in W_X} \sum_{y \in W_Y} xy \cdot \Pr[X = x] \cdot \Pr[Y = y] \\
&= \sum_{x \in W_X} x \cdot \Pr[X = x] \sum_{y \in W_Y} y \cdot \Pr[Y = y] \\
&= \mathbb{E}[X] \cdot \mathbb{E}[Y]\,.
\end{aligned}$$

□

Dass für die Gültigkeit von Satz 52 die Unabhängigkeit der Zufallsvariablen wirklich notwendig ist, sieht man beispielsweise am Fall $Y = -X$ für eine Zufallsvariable mit einer von Null verschiedenen Varianz. Dann gilt

$$\mathbb{E}[X \cdot Y] = -\mathbb{E}[X^2] \neq -(\mathbb{E}[X])^2 = \mathbb{E}[X] \cdot \mathbb{E}[Y]\,.$$

Definition 53

Zu einem Ereignis A heißt die Zufallsvariable

$$I_A := \begin{cases} 1 & \text{falls } A \text{ eintritt,} \\ 0 & \text{sonst} \end{cases}$$

Indikatorvariable des Ereignisses A.

Beobachtung:

Für die Indikatorvariable I_A gilt nach Definition

$$\mathbb{E}[I_A] = 1 \cdot \Pr[A] + 0 \cdot \Pr[\bar{A}] = \Pr[A]\,.$$

Ebenso gilt

$$\mathbb{E}[I_{A_1} \cdot \ldots \cdot I_{A_n}] = \Pr[A_1 \cap \ldots \cap A_n],$$

da das Produkt von Indikatorvariablen genau dann gleich 1 ist, wenn alle entsprechenden Ereignisse eintreten.

Beispiel (Forts.)

Wir betrachten wieder das Beispiel der total betrunkenen Matrosen.
Sei A_i das Ereignis, dass der i-te Seemann im richtigen Bett liegt. Mit der Notation der Indikatorvariablen sei $X_i = I_{A_i}$. Dann gilt für beliebige $i, j \in \{1, \ldots, n\}$, $i \neq j$:

$$\mathbb{E}[X_i X_j] = \mathbb{E}[I_{A_i} I_{A_j}] = \Pr[A_i \cap A_j] = \frac{1}{n(n-1)},$$

sowie

$$\mathbb{E}[X_i^2] = 0^2 \cdot \Pr[\bar{A}_i] + 1^2 \cdot \Pr[A_i] = \Pr[A_i] = 1/n.$$

Beispiel (Forts.)

Daraus folgt wegen der Linearität des Erwartungswerts für $X = X_1 + \cdots + X_n$*:*

$$\begin{aligned}\mathbb{E}[X^2] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} \sum_{j \neq i} X_i X_j\right] \\ &= n \cdot \frac{1}{n} + n(n-1) \cdot \frac{1}{n(n-1)} = 2\,.\end{aligned}$$

Für die Varianz erhalten wir somit den Wert

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 2 - 1 = 1.$$

Einfacher Beweis für Satz 9 mit Hilfe von Indikatorvariablen:
Zur Erinnerung:

Satz 9 (Siebformel, Prinzip der Inklusion/Exklusion)

Für Ereignisse $A_1, \ldots, A_n$ $(n \geq 2)$ gilt:

$$\begin{aligned}\Pr\left[\bigcup_{i=1}^{n} A_i\right] &= \sum_{i=1}^{n} \Pr[A_i] - \sum_{1 \leq i_1 < i_2 \leq n} \Pr[A_{i_1} \cap A_{i_2}] + - \ldots \\ &\quad + (-1)^{l-1} \sum_{1 \leq i_1 < \ldots < i_l \leq n} \Pr[A_{i_1} \cap \ldots \cap A_{i_l}] + - \ldots \\ &\quad + (-1)^{n-1} \cdot \Pr[A_1 \cap \ldots \cap A_n]\,.\end{aligned}$$

Beweis:

Zur Erinnerung: Zu Ereignissen $A_1, \ldots, A_n$ wollen wir die Wahrscheinlichkeit $\Pr[B]$ des Ereignisses $B := A_1 \cup \ldots \cup A_n$ ermitteln.

Wir betrachten die Indikatorvariablen $I_i := I_{A_i}$ der Ereignisse $A_1, \ldots, A_n$ und die Indikatorvariable $I_{\bar{B}}$ des Ereignisses $\bar{B}$.

Das Produkt $\prod_{i=1}^{n}(1 - I_i)$ ist genau dann gleich 1, wenn $I_1 = \ldots = I_n = 0$, d.h. wenn B nicht eintritt. Somit gilt $I_{\bar{B}} = \prod_{i=1}^{n}(1 - I_i)$ und wir erhalten:

$$I_{\bar{B}} = 1 - \sum_{1 \le i \le n} I_i + \sum_{1 \le i_1 < i_2 \le n} I_{i_1} I_{i_2} - + \ldots + (-1)^n I_1 \cdot \ldots \cdot I_n\,,$$

also

$$\begin{aligned} I_B &= 1 - I_{\bar{B}} \\ &= \sum_{1 \le i \le n} I_i - \sum_{1 \le i_1 < i_2 \le n} I_{i_1} I_{i_2} + - \ldots + (-1)^{n-1} I_1 \cdot \ldots \cdot I_n\,. \end{aligned}$$

Beweis:

Wegen der Eigenschaften von Indikatorvariablen gilt

$$\Pr[B] = 1 - \Pr[\bar{B}] = 1 - \mathbb{E}[I_{\bar{B}}].$$

Mit Hilfe von Satz 50 „verteilen" wir den Erwartungswert auf die einzelnen Produkte von Indikatorvariablen. Wenn wir nun $\mathbb{E}[I_i]$ durch $\Pr[A_i]$ und allgemein $\mathbb{E}[I_{i_1} \cdot \ldots \cdot I_{i_k}]$ durch $\Pr[A_{i_1} \cap \ldots \cap A_{i_k}]$ ersetzen, haben wir Satz 9 (dieses Mal vollständig) bewiesen.

□

Satz 54

Für unabhängige Zufallsvariablen $X_1, \ldots, X_n$ *und* $X := X_1 + \ldots + X_n$ *gilt*

$$\mathrm{Var}[X] = \mathrm{Var}[X_1] + \ldots + \mathrm{Var}[X_n]\,.$$

Beweis:

Wir betrachten nur den Fall $n = 2$ mit den Zufallsvariablen X und Y.

$$\mathbb{E}[(X+Y)^2] = \mathbb{E}[X^2 + 2XY + Y^2] = \mathbb{E}[X^2] + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y^2]$$
$$\mathbb{E}[X+Y]^2 = (\mathbb{E}[X] + \mathbb{E}[Y])^2 = \mathbb{E}[X]^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y]^2$$

Wir ziehen die zweite Gleichung von der ersten ab und erhalten

$$\mathbb{E}[(X+Y)^2] - \mathbb{E}[X+Y]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Y^2] - \mathbb{E}[Y]^2\,.$$

Mit Hilfe von Satz 39 folgt die Behauptung. $\square$

Für abhängige Zufallsvariablen $X_1, \ldots, X_n$ gilt Satz 54 im Allgemeinen nicht. Als Beispiel funktioniert wiederum der Fall $X = -Y$:

$$\mathrm{Var}[X+Y] = 0 \neq 2 \cdot \mathrm{Var}[X] = \mathrm{Var}[X] + \mathrm{Var}[Y]\,.$$

5. Wichtige diskrete Verteilungen

Wir diskutieren nun einige wichtige *diskrete* Verteilungen. Bei diesen Verteilungen handelt es sich um Funktionen, die von gewissen *Parametern* abhängen. Eigentlich betrachten wir also immer eine ganze Familie von ähnlichen Verteilungen.

5.1 Bernoulli-Verteilung

Eine Zufallsvariable X mit $W_X = \{0, 1\}$ und der Dichte

$$f_X(x) = \begin{cases} p & \text{für } x = 1, \\ 1 - p & \text{für } x = 0. \end{cases}$$

heißt Bernoulli-verteilt. Den Parameter p nennen wir Erfolgswahrscheinlichkeit.

Eine solche Verteilung erhält man z.B. bei einer einzelnen Indikatorvariablen. Es gilt mit $q := 1 - p$

$$\mathbb{E}[X] = p \text{ und } \mathrm{Var}[X] = pq,$$

wegen $\mathbb{E}[X^2] = p$ und $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2$.

Der Name der Bernoulli-Verteilung geht zurück auf den Schweizer Mathematiker Jakob Bernoulli (1654–1705). Wie viele andere Mathematiker seiner Zeit hätte auch Bernoulli nach dem Wunsch seines Vaters ursprünglich Theologe werden sollen. Sein Werk *ars conjectandi* stellt eine der ersten Arbeiten dar, die sich mit dem Teil der Mathematik beschäftigen, den wir heute als Wahrscheinlichkeitstheorie bezeichnen.

5.2 Binomialverteilung

Eine Bernoulli-verteilte Zufallsvariable entspricht der Verteilung *einer* Indikatorvariablen. Häufig betrachtet man jedoch Summen von Indikatorvariablen.

Definition 55

Sei $X := X_1 + \ldots + X_n$ als Summe von n unabhängigen, Bernoulli-verteilten Zufallsvariablen mit gleicher Erfolgswahrscheinlichkeit p definiert. Dann heißt X binomialverteilt mit den Parametern n und p. In Zeichen schreiben wir

$$X \sim \mathrm{Bin}(n, p)\,.$$

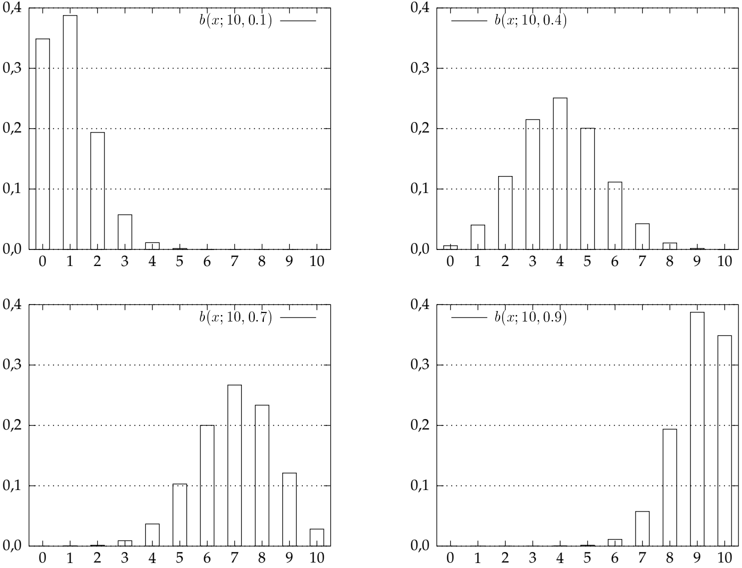
Es gilt $W_X = \{0, \ldots, n\}$. Die Binomialverteilung besitzt die Dichte

$$f_X(x) := b(x; n, p) = \binom{n}{x} p^x q^{n-x}$$

mit $q := 1 - p$. Da die Binomialverteilung eine sehr wichtige Rolle spielt, führen wir für die Dichtefunktion die Abkürzung $b(x; n, p)$ ein.

Mit den Sätzen über Erwartungswert und Varianz von Summen unabhängiger Zufallsvariablen erhalten wir sofort

$$\mathbb{E}[X] = np \quad \text{und} \quad \mathrm{Var}[X] = npq\,.$$



Dichte der Binomialverteilung

Satz 56

Wenn $X \sim \mathrm{Bin}(n_x, p)$ *und* $Y \sim \mathrm{Bin}(n_y, p)$ *unabhängig sind, dann gilt für* $Z := X + Y$*, dass* $Z \sim \mathrm{Bin}(n_x + n_y, p)$*.*

Beweis:

Die Aussage folgt sofort, wenn man gemäß der Definition der Binomialverteilung X und Y als Summen von Indikatorvariablen darstellt. Z ist dann offensichtlich wieder eine Summe von unabhängigen Indikatorvariablen. $\square$

5.3 Geometrische Verteilung

Man betrachte ein Experiment, das so lange wiederholt wird, bis Erfolg eintritt. Gelingt ein einzelner Versuch mit Wahrscheinlichkeit p, so ist die Anzahl der Versuche bis zum Erfolg geometrisch verteilt.

Definition 57

Eine geometrisch verteilte Zufallsvariable X mit Parameter (Erfolgswahrscheinlichkeit) $p \in (0,1]$ und $q := 1 - p$ hat die Dichte

$$f_X(i) = pq^{i-1} \quad \text{für } i \in \mathbb{N}\,.$$

Für Erwartungswert und Varianz geometrisch verteilter Zufallsvariablen gilt

$$\mathbb{E}[X] = \frac{1}{p} \quad \text{und} \quad \mathrm{Var}[X] = \frac{q}{p^2}.$$

Es gilt

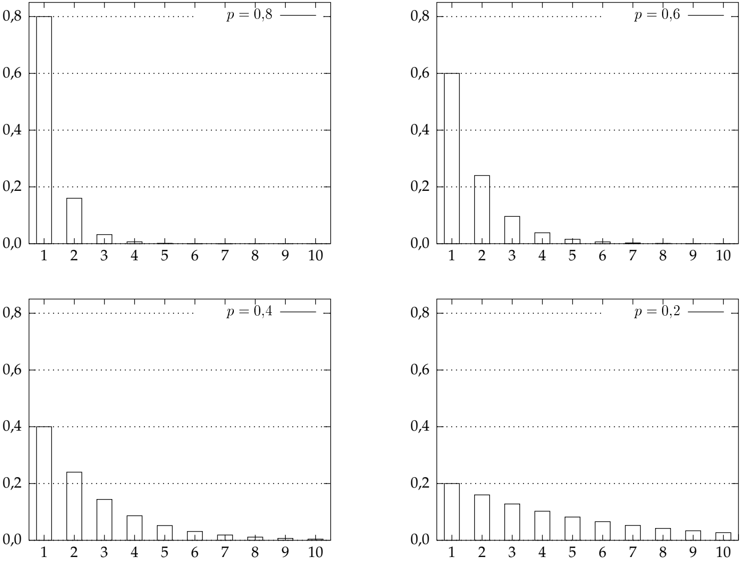
$$\mathbb{E}[X] = \sum_{i=1}^{\infty} i \cdot pq^{i-1} = p \cdot \sum_{i=1}^{\infty} i \cdot q^{i-1} = p \cdot \frac{1}{(1-q)^2} = \frac{1}{p}.$$

Ferner ist

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{i=1}^{\infty} i^2 \cdot pq^{i-1} \\ &= p \cdot \left(\sum_{i=1}^{\infty} i(i+1) \cdot q^{i-1} - \sum_{i=1}^{\infty} i \cdot q^{i-1} \right) \\ &= p \cdot \left(\frac{2}{p^3} - \frac{1}{p^2} \right) = \frac{2-p}{p^2},\end{aligned}$$

und damit

$$\mathrm{Var}[X] = \frac{q}{p^2}.$$



Dichte der geometrischen Verteilung

Sei X wieder geometrisch verteilt mit Erfolgswahrscheinlichkeit p. Dann ist $\Pr[X = k]$ die Wahrscheinlichkeit, dass wir bei einem binären Experiment mit Erfolgswahrscheinlichkeit p genau in der k-ten unabhängigen Wiederholung das erste Mal erfolgreich sind.

Wie groß ist die Wahrscheinlichkeit $\Pr[X > y + x \mid X > x]$?

Da bei den ersten x Versuchen kein Erfolg eintrat, stellen wir uns vor, dass das „eigentliche" Experiment erst ab dem $(x + 1)$-ten Versuch beginnt. Die Zeit bis zum ersten Erfolg bei diesem neuen Experiment nennen wir X'. Damit $X > y + x$ gilt, muss $X' > y$ gelten. Es ist intuitiv, dass X' wieder geometrisch verteilt ist mit Erfolgswahrscheinlichkeit p, dass also für $x, y \in \mathbb{N}$ gilt:

$$\Pr[X > y + x \mid X > x] = \Pr[X' > y]. \tag{6}$$

Formal gilt

$$\Pr[X > x] = \sum_{i=x+1}^{\infty} (1-p)^{i-1} p = (1-p)^x p \cdot \sum_{i=0}^{\infty} (1-p)^i$$
$$= (1-p)^x p \cdot \frac{1}{1-(1-p)} = (1-p)^x \,,$$

sowie

$$\begin{aligned}\Pr[X > y + x \mid X > x] &= \frac{\Pr[X > y + x, X > x]}{\Pr[X > x]} \\ &= \frac{\Pr[X > y + x]}{\Pr[X > x]} \\ &= (1-p)^{y+x} \cdot (1-p)^{-x} = (1-p)^y \\ &= \Pr[X > y] \,.\end{aligned}$$

Diese Eigenschaft nennt man *Gedächtnislosigkeit*, da eine geometrisch verteilte Zufallsvariable gewissermaßen vergisst, dass sie schon x Misserfolge hinter sich hat und sich deshalb zum Zeitpunkt $y + x$ genauso verhält wie ursprünglich zur Zeit y.

Warten auf den n-ten Erfolg.

Wir betrachten n unabhängige Zufallsvariablen $X_1, \ldots, X_n$, die jeweils geometrisch verteilt sind mit Parameter p, und bestimmen die Dichte der Zufallsvariablen $Z := X_1 + \cdots + X_n$. Damit bezeichnet Z also die Anzahl der Versuche bis zum n-ten erfolgreichen Experiment (einschließlich).

Falls $Z = z$ ist, so werden also genau n erfolgreiche und $z - n$ nicht erfolgreiche Experimente durchgeführt. Dafür gibt es genau $\binom{z-1}{n-1}$ Möglichkeiten, von denen jede mit Wahrscheinlichkeit $p^n(1-p)^{z-n}$ eintritt. Es gilt also

$$f_Z(z) = \binom{z-1}{n-1} \cdot p^n(1-p)^{z-n}\,.$$

Die Zufallsvariable Z nennt man negativ binomialverteilt mit Ordnung n.

Das Coupon-Collector-Problem

In manchen Branchen legen Firmen den Verpackungen ihrer Produkte oft kleine Bilder oder andere Gegenstände bei, um den Käufer zum Sammeln anzuregen. Wenn es insgesamt n verschiedene solche Beilagen gibt, wie viele Packungen muss man im Mittel erwerben, bis man eine vollständige Sammlung besitzt? Hierbei nehmen wir an, dass bei jedem Kauf jede Beilage mit gleicher Wahrscheinlichkeit auftritt.

Bezeichne

- X die Anzahl der zu tätigenden Käufe und
- Phase i die Schritte vom Erwerb der $(i-1)$-ten Beilage (ausschließlich) bis zum Erwerb der i-ten Beilage (einschließlich).

Sei etwa $n = 4$, und seien die Beilagen mit den Zahlen $1, 2, 3, 4$ identifiziert. Ein Experiment ist z.B.:

$$\underbrace{2}_{1}, \underbrace{2, 1}_{2}, \underbrace{2, 2, 3}_{3}, \underbrace{1, 3, 2, 3, 1, 4}_{4} .$$

Beobachtung:
Phase i endet genau dann, wenn wir eine der $n - i + 1$ Beilagen erhalten, die wir noch nicht besitzen.

Somit ist X_i geometrisch verteilt mit Parameter $p = \frac{n-i+1}{n}$ und es gilt $\mathbb{E}[X_i] = \frac{n}{n-i+1}$.

Damit folgt aber sofort

$$\begin{aligned}\mathbb{E}[X] &= \sum_{i=1}^{n} \mathbb{E}[X_i] \\ &= \sum_{i=1}^{n} \frac{n}{n-i+1} \\ &= n \cdot \sum_{i=1}^{n} \frac{1}{i} = n \cdot H_n,\end{aligned}$$

wobei $H_n := \sum_{i=1}^{n} \frac{1}{i}$ die n-te harmonische Zahl bezeichnet. Da $H_n = \ln n + O(1)$, folgt $\mathbb{E}[X] = n \ln n + O(n)$.

5.4 Poisson-Verteilung

Die Poisson-Verteilung kann verwendet werden, um die Anzahl von Ereignissen zu modellieren, welche mit konstanter Rate und unabhängig voneinander in einem Zeitintervall auftreten.

Eine Poisson-verteilte Zufallsvariable X mit dem Parameter $\lambda \geq 0$ hat den Wertebereich $W_X = \mathbb{N}_0$ und besitzt die Dichte

$$f_X(i) = \frac{e^{-\lambda}\lambda^i}{i!} \quad \text{für } i \in \mathbb{N}_0.$$

f_X ist eine zulässige Dichte, da

$$\begin{aligned}\sum_{i=0}^{\infty} f_X(i) &= \sum_{i=0}^{\infty} \frac{e^{-\lambda}\lambda^i}{i!}\\ &= e^{-\lambda} \cdot e^{\lambda} = 1\,.\end{aligned}$$

Für den Erwartungswert erhalten wir

$$
\begin{aligned}
\mathbb{E}[X] &= \sum_{i=0}^{\infty} i \cdot \frac{e^{-\lambda}\lambda^i}{i!} \\
&= \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} \\
&= \lambda e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^{i}}{i!} \\
&= \lambda e^{-\lambda} e^{\lambda} = \lambda\,.
\end{aligned}
$$

Da

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{i=0}^{\infty} i(i-1)\cdot\frac{e^{-\lambda}\lambda^i}{i!}\\ &= \lambda^2 e^{-\lambda}\sum_{i=2}^{\infty}\frac{\lambda^{i-2}}{(i-2)!}\\ &= \lambda^2 e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^{i}}{i!}\\ &= \lambda^2 e^{-\lambda}e^{\lambda} = \lambda^2\end{aligned}$$

und

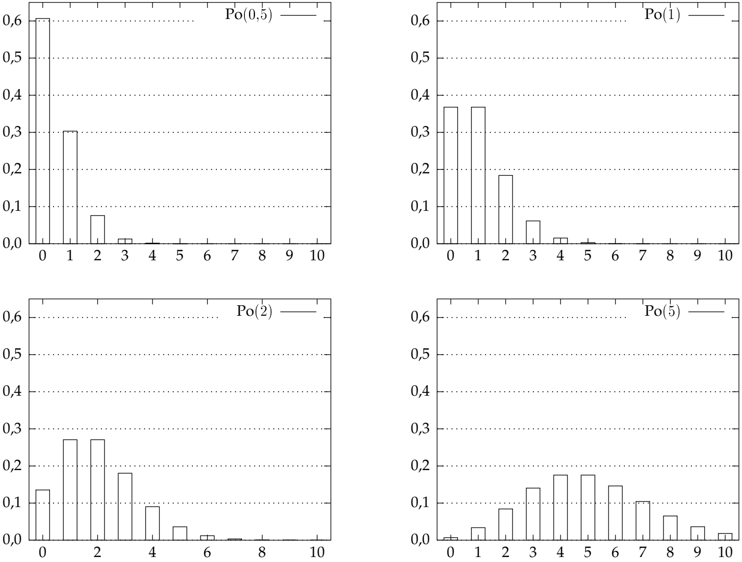
$$\begin{aligned}&\mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2\\ &= \mathbb{E}[X^2] - \mathbb{E}[X] + \mathbb{E}[X] - \mathbb{E}[X]^2 = \mathrm{Var}[X]\,,\end{aligned}$$

folgt

$$\operatorname{Var}[X] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda. \tag{7}$$

Dafür, dass eine Zufallsvariable X Poisson-verteilt mit Parameter λ ist, schreiben wir auch

$$X \sim \mathrm{Po}(\lambda).$$



Dichte der Poisson-Verteilung

5.4.1 Poisson-Verteilung als Grenzwert der Binomialverteilung

Wir betrachten eine Folge von binomialverteilten Zufallsvariablen X_n mit $X_n \sim \mathrm{Bin}(n, p_n)$, wobei $p_n = \lambda/n$. Für ein beliebiges k mit $0 \leq k \leq n$ ist die Wahrscheinlichkeit, dass X_n den Wert k annimmt, gleich

$$\begin{aligned} b(k; n, p_n) &= \binom{n}{k} \cdot p_n^k \cdot (1 - p_n)^{n-k} \\ &= \frac{(n \cdot p_n)^k}{k!} \cdot \frac{n^{\underline{k}}}{n^k} \cdot (1 - p_n)^{-k} \cdot (1 - p_n)^n \\ &= \frac{\lambda^k}{k!} \cdot \frac{n^{\underline{k}}}{n^k} \cdot \left(1 - \frac{\lambda}{n}\right)^{-k} \cdot \left(1 - \frac{\lambda}{n}\right)^n . \end{aligned}$$

Wir betrachten nun $n \to \infty$ und erinnern uns, dass

$$\lim_{n\to\infty} \frac{n^{\underline{k}}}{n^k} = 1,$$

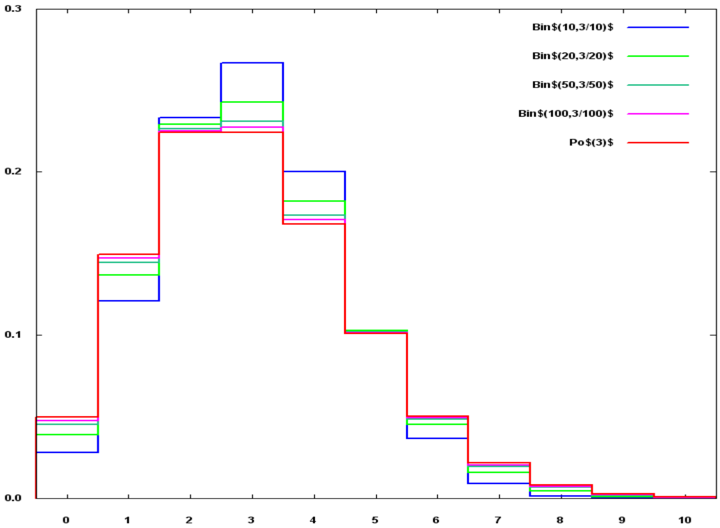
$$\lim_{n\to\infty} (1 - \frac{\lambda}{n})^{-k} = 1, \text{ und}$$

$$\lim_{n\to\infty} (1 - \frac{\lambda}{n})^{n} = e^{-\lambda}\,.$$

Damit folgt

$$\lim_{n\to\infty} b(k; n, p_n) = \lim_{n\to\infty} \binom{n}{k} \cdot p_n^k \cdot (1 - p_n)^{n-k} = e^{-\lambda} \cdot \frac{\lambda^k}{k!}.$$

Die Wahrscheinlichkeit $b(k; n, p_n)$ konvergiert also für $n \to \infty$ gegen die Wahrscheinlichkeit, dass eine Poisson-verteilte Zufallsvariable mit Parameter λ den Wert k annimmt. Insgesamt folgt somit, dass die Verteilung einer Zufallsvariablen $X \sim \mathrm{Bin}(n, \lambda/n)$ sich für $n \to \infty$ der Poisson-Verteilung $\mathrm{Po}(\lambda)$ annähert.



Vergleich von Binomial- und Poisson-Verteilung

Ist also n im Vergleich zu λ hinreichend groß, so kann man die Poisson-Verteilung als Approximation der Binomialverteilung verwenden.

Diese Tatsache wird manchmal auch als Gesetz seltener Ereignisse bezeichnet, da die Wahrscheinlichkeit eines einzelnen Treffers $p_n = \lambda/n$ relativ klein sein muss, wenn die Approximation gute Ergebnisse liefern soll.

Die folgenden Voraussetzungen müssen erfüllt sein, damit die Annahme der Poisson-Verteilung gerechtfertigt ist:

- Die Ereignisse treten nie zur gleichen Zeit auf.
- Die Wahrscheinlichkeit, dass ein Ereignis in einem (kleinen) Zeitintervall auftritt, ist proportional zur Länge des Intervalls.
- Die Anzahl der Ereignisse in einem festen Zeitintervall hängt nur von dessen Länge ab, nicht aber von der Lage auf der Zeitachse.
- Wenn man zwei disjunkte Zeitintervalle betrachtet, so sind die Anzahlen der Ereignisse in diesen Zeiträumen voneinander unabhängig.

Beispiel 58

Wir wollen wissen, wie oft eine bestimmte Gegend im Durchschnitt von einer Naturkatastrophe (z.B. Vulkanausbruch) getroffen wird. Aus Statistiken entnehmen wir, dass so ein Ereignis im Mittel 10^{-4}-mal pro Jahr auftritt. Wir interessieren uns nun für die Wahrscheinlichkeit, dass die Region in einem Jahr mehr als einmal von einem solchen Unglück heimgesucht wird.
Die Voraussetzungen scheinen erfüllt zu sein, die Anzahl X der Katastrophen durch eine Poisson-Verteilung mit Parameter $\lambda = 10^{-4}$ zu modellieren.
Damit gilt

$$\begin{aligned}\Pr[X \geq 2] &= 1 - \Pr[X = 0] - \Pr[X = 1] = 1 - e^{-\lambda} - \lambda e^{-\lambda}\\ &\approx 1 - 0{,}999900005 - 0{,}000099990 = 5 \cdot 10^{-9}.\end{aligned}$$

Summe von Poisson-verteilten Zufallsvariablen

Satz 59

Sind X und Y unabhängige Zufallsvariablen mit $X \sim \mathrm{Po}(\lambda)$ und $Y \sim \mathrm{Po}(\mu)$, dann gilt

$$Z := X + Y \sim \mathrm{Po}(\lambda + \mu)\,.$$

Beweis:

$$\begin{aligned}
f_Z(z) &= \sum_{x=0}^{\infty} f_X(x) \cdot f_Y(z-x) = \sum_{x=0}^{z} \frac{e^{-\lambda}\lambda^x}{x!} \cdot \frac{e^{-\mu}\mu^{z-x}}{(z-x)!} \\
&= e^{-(\lambda+\mu)} \cdot \frac{(\lambda+\mu)^z}{z!} \cdot \sum_{x=0}^{z} \frac{z!}{x!(z-x)!} \left(\frac{\lambda}{\lambda+\mu}\right)^x \left(\frac{\mu}{\lambda+\mu}\right)^{z-x} \\
&= e^{-(\lambda+\mu)} \cdot (\lambda+\mu)^z \frac{1}{z!} \cdot \sum_{x=0}^{z} \binom{z}{x} p^x (1-p)^{z-x},
\end{aligned}$$

wobei $p := \frac{\lambda}{\lambda+\mu}$.
Da die Summe gleich 1 ist, folgt

$$f_Z(z) = e^{-(\lambda+\mu)} \cdot (\lambda+\mu)^z \frac{1}{z!}\,.$$

□

Erläuterungen und Beispiele zur Poisson-Verteilung

- In der Wikipedia finden sich ein paar weitere Details und Beispiele.
- Eine Anwendung der Poisson-Verteilung auf die Fußball-Bundesliga wird in einem Artikel präsentiert, der im Spektrum der Wissenschaft, Heft Juni 2010, erschienen ist.

6. Abschätzen von Wahrscheinlichkeiten

6.1 Die Ungleichungen von Markov und Chebyshev

Satz 60 (Markov-Ungleichung)

Sei X eine Zufallsvariable, die nur nicht-negative Werte annimmt. Dann gilt für alle $t \in \mathbb{R}$ mit $t > 0$, dass

$$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

Äquivalent dazu:

$$\Pr[X \geq t \cdot \mathbb{E}[X]] \leq 1/t\,.$$

Beweis:

$$
\begin{aligned}
t \cdot \Pr[X \geq t] &= t \cdot \sum_{x \in W_X,\, x \geq t} \Pr[X = x] \\
&\leq \sum_{x \in W_X,\, x \geq t} x \cdot \Pr[X = x] \\
&\leq \sum_{x \in W_X} x \cdot \Pr[X = x] \\
&= \mathbb{E}[X]\,.
\end{aligned}
$$

□

Alternativer Beweis:

Es gilt

$$\mathbb{E}[X] = \mathbb{E}[X|X < t] \cdot \Pr[X < t] + \mathbb{E}[X|X \geq t] \cdot \Pr[X \geq t]\,.$$

Wegen $\mathbb{E}[X|X < t] \cdot \Pr[X < t] \geq 0$ und $\mathbb{E}[X|X \geq t] \geq t$ folgt sofort

$$\mathbb{E}[X] \geq t \cdot \Pr[X \geq t]\,.$$

Die Markov-Ungleichung ist nach Andrey Andreyevich Markov (1856–1922) benannt, der an der Universität von St. Petersburg bei Chebyshev studierte und später dort arbeitete. Neben seiner mathematischen Tätigkeit fiel Markov durch heftige Proteste gegen das Zaren-Regime auf, und nur sein Status als vermeintlich harmloser Akademiker schützte ihn vor Repressalien durch die Behörden. Im Jahr 1913 organisierte er parallel zum dreihundertjährigen Geburtstag der Zarenfamilie Romanov eine Feier zum zweihundertjährigen Geburtstag des Gesetzes der großen Zahlen (s.u.).

Die folgende Abschätzung ist nach Pavnuty Lvovich Chebyshev (1821–1894) benannt, der ebenfalls an der Staatl. Universität in St. Petersburg wirkte.

Satz 61 (Chebyshev-Ungleichung)

Sei X eine Zufallsvariable, und sei $t \in \mathbb{R}$ mit $t > 0$. Dann gilt

$$\Pr[|X - \mathbb{E}[X]| \geq t] \;\leq\; \frac{\mathrm{Var}[X]}{t^2}\,.$$

Äquivalent dazu:

$$\Pr[|X - \mathbb{E}[X]| \geq t\sqrt{\mathrm{Var}[X]}] \;\leq\; 1/t^2\,.$$

Beweis:

Wir stellen fest, dass

$$\Pr[|X - \mathbb{E}[X]| \geq t] \;=\; \Pr[(X - \mathbb{E}[X])^2 \geq t^2]\,.$$

Setze

$$Y := (X - \mathbb{E}[X])^2\,.$$

Dann gilt $\mathbb{E}[Y] = \mathrm{Var}[X]$, und damit mit der Markov-Ungleichung:

$$\Pr[|X - \mathbb{E}[X]| \geq t] \;=\; \Pr[Y \geq t^2] \;\leq\; \frac{\mathbb{E}[Y]}{t^2} \;=\; \frac{\mathrm{Var}[X]}{t^2}\,.$$

$\square$

Beispiel 62

Wir werfen 1000-mal eine faire Münze und ermitteln die Anzahl X der Würfe, in denen „Kopf" fällt.

X ist binomialverteilt mit $X \sim \mathrm{Bin}(1000, p = \frac{1}{2})$, also gilt

$$\mathbb{E}[X] = \frac{1}{2}n = 500 \text{ und } \mathrm{Var}[X] = \frac{1}{4}n = 250.$$

Wie groß ist die Wahrscheinlichkeit, dass mehr als 550-mal „Kopf" fällt?

Beispiel 62

Chebyshev-Ungleichung:

$$\Pr[X \geq 550] \leq \Pr[|X - 500| \geq 50] \leq \frac{250}{50^2} = 0{,}1\,.$$

Setze nun $n = 10000$ und betrachte wieder eine maximal 10%-ige Abweichung vom Erwartungswert:

$$\mathbb{E}[X] = 5000 \text{ und } \mathrm{Var}[X] = 2500, \text{ und damit}$$

$$\Pr[X \geq 5500] \leq \Pr[|X - 5000| \geq 500] \leq \frac{2500}{500^2} = 0{,}01\,.$$

6.2 Gesetz der großen Zahlen

Wir haben diskutiert, wie Wahrscheinlichkeiten als Grenzwerte von relativen Häufigkeiten aufgefasst werden können.

Satz 63 (Gesetz der großen Zahlen)

Gegeben sei eine Zufallsvariable X. Ferner seien $\varepsilon, \delta > 0$ beliebig aber fest. Dann gilt für alle $n \geq \frac{\mathrm{Var}[X]}{\varepsilon\delta^2}$:
Sind $X_1, \ldots, X_n$ unabhängige Zufallsvariablen mit derselben Verteilung wie X und setzt man

$$Z := \frac{X_1 + \ldots + X_n}{n},$$

so gilt

$$\Pr[|Z - \mathbb{E}[X]| \geq \delta] \;\leq\; \varepsilon.$$

Beweis:

Für Z gilt

$$\mathbb{E}[Z] = \frac{1}{n} \cdot (\mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n]) = \frac{1}{n} \cdot n \cdot \mathbb{E}[X] = \mathbb{E}[X],$$

sowie

$$\mathrm{Var}[Z] = \frac{1}{n^2} \cdot (\mathrm{Var}[X_1] + \ldots + \mathrm{Var}[X_n]) = \frac{1}{n^2} \cdot n \cdot \mathrm{Var}[X] = \frac{\mathrm{Var}[X]}{n}.$$

Mit der Chebyshev-Ungleichung erhalten wir

$$\Pr[|Z - \mathbb{E}[X]| \geq \delta] = \Pr[|Z - \mathbb{E}[Z]| \geq \delta] \leq \frac{\mathrm{Var}[Z]}{\delta^2} = \frac{\mathrm{Var}[X]}{n\delta^2} \leq \varepsilon,$$

nach Wahl von n. $\square$

Wahrscheinlichkeit und relative Häufigkeit.

Sei X eine Indikatorvariable für ein Ereignis A, $\Pr[A] = p$. Somit ist X Bernoulli-verteilt mit $\mathbb{E}[X] = p$.

$Z = \frac{1}{n}(X_1 + \ldots + X_n)$ gibt die relative Häufigkeit an, mit der A bei n Wiederholungen des Versuchs eintritt, denn

$$Z = \frac{\text{Anzahl der Versuche, bei denen } A \text{ eingetreten ist}}{\text{Anzahl aller Versuche}}.$$

Mit Hilfe des obigen Gesetzes der großen Zahlen folgt

$$\Pr[|Z - p| \geq \delta] \leq \varepsilon,$$

für genügend großes n. Also nähert sich die relative Häufigkeit von A bei hinreichend vielen Wiederholungen des Experiments mit beliebiger Sicherheit beliebig nahe an die „wahre" Wahrscheinlichkeit p an.

Die obige Variante eines Gesetzes der großen Zahlen geht auf Jakob Bernoulli zurück, der den Satz in seinem Werk ars conjectandi zeigte.

Es soll betont werden, dass das Gesetz der großen Zahlen die

$$\text{relative Abweichung } |\tfrac{1}{n}\sum_i X_i - p|$$

und nicht die

$$\text{absolute Abweichung } |\sum_i X_i - np|$$

abschätzt!

6.3 Chernoff-Schranken

6.3.1 Chernoff-Schranken für Summen von 0–1–Zufallsvariablen

Die hier betrachtete Art von Schranken ist nach Herman Chernoff (*1923) benannt. Sie finden in der komplexitätstheoretischen Analyse von Algorithmen eine sehr häufige Verwendung.

Satz 64

Seien $X_1, \ldots, X_n$ unabhängige Bernoulli-verteilte Zufallsvariablen mit $\Pr[X_i = 1] = p_i$ und $\Pr[X_i = 0] = 1 - p_i$. Dann gilt für $X := \sum_{i=1}^n X_i$ und $\mu := \mathbb{E}[X] = \sum_{i=1}^n p_i$, sowie jedes $\delta > 0$, dass

$$\Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}} \right)^\mu .$$

Beweis:

Für $t > 0$ gilt
$$\Pr[X \geq (1+\delta)\mu] = \Pr[e^{tX} \geq e^{t(1+\delta)\mu}]\,.$$

Mit der Markov-Ungleichung folgt

$$\Pr[X \geq (1+\delta)\mu] = \Pr[e^{tX} \geq e^{t(1+\delta)\mu}] \leq \frac{\mathbb{E}[e^{tX}]}{e^{t(1+\delta)\mu}}.$$

Wegen der Unabhängigkeit der Zufallsvariablen $X_1, \ldots, X_n$ gilt

$$\mathbb{E}[e^{tX}] = \mathbb{E}\left[\exp\left(\sum_{i=1}^{n} tX_i\right)\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{tX_i}\right] = \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}].$$

Weiter ist für $i \in \{1, \ldots, n\}$:
$$\mathbb{E}[e^{tX_i}] = e^{t\cdot 1}p_i + e^{t\cdot 0}(1-p_i) = e^t p_i + 1 - p_i = 1 + p_i(e^t - 1)\,,$$

Beweis (Forts.):

und damit

$$
\begin{aligned}
\Pr[X \geq (1+\delta)\mu] &\leq \frac{\prod_{i=1}^{n}(1+p_i(e^t-1))}{e^{t(1+\delta)\mu}} \\
&\leq \frac{\prod_{i=1}^{n}\exp(p_i(e^t-1))}{e^{t(1+\delta)\mu}} \\
&= \frac{\exp(\sum_{i=1}^{n} p_i(e^t-1))}{e^{t(1+\delta)\mu}} = \frac{e^{(e^t-1)\mu}}{e^{t(1+\delta)\mu}} =: f(t)\,.
\end{aligned}
$$

Wir wählen nun t so, dass $f(t)$ minimiert wird, nämlich

$$t = \ln(1+\delta)\,.$$

Damit wird

$$f(t) = \frac{e^{(e^t-1)\mu}}{e^{t(1+\delta)\mu}} = \frac{e^{\delta\mu}}{(1+\delta)^{(1+\delta)\mu}}\,.$$

□

Beispiel 65

Wir betrachten wieder das Beispiel, dass wir eine faire Münze n-mal werfen und abschätzen wollen, mit welcher Wahrscheinlichkeit „Kopf"

$$\frac{n}{2}(1 + 10\%)$$

oder öfter fällt.

n	Chebyshev	Chernoff
1000	$0{,}1$	$0{,}0889$
10000	$0{,}01$	$0{,}308 \cdot 10^{-10}$
n	$\frac{\frac{1}{4}n}{(0{,}1 \cdot \frac{1}{2}n)^2}$	$\left(\frac{e^{0{,}1}}{(1+0{,}1)^{1+0{,}1}}\right)^{\frac{1}{2}n}$

Satz 66

Seien $X_1, \ldots, X_n$ *unabhängige Bernoulli-verteilte Zufallsvariablen mit* $\Pr[X_i = 1] = p_i$ *und* $\Pr[X_i = 0] = 1 - p_i$. *Dann gilt für* $X := \sum_{i=1}^n X_i$ *und* $\mu := \mathbb{E}[X] = \sum_{i=1}^n p_i$, *sowie jedes* $0 < \delta < 1$, *dass*

$$\Pr[X \leq (1-\delta)\mu] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}.$$

Beweis:

Analog zum Beweis von Satz 64. □

Bemerkung: Abschätzungen, wie sie in Satz 64 und Satz 66 angegeben sind, nennt man auch tail bounds, da sie Schranken für die tails, also die vom Erwartungswert weit entfernten Bereiche angeben. Man spricht hierbei vom upper tail (vergleiche Satz 64) und vom lower tail (vergleiche Satz 66).

Die Chernoff-Schranken hängen exponentiell von μ ab!

Lemma 67

Für $0 \le \delta < 1$ *gilt*

$$(1-\delta)^{1-\delta} \ge e^{-\delta+\delta^2/2} \quad \textit{und} \quad (1+\delta)^{1+\delta} \ge e^{\delta+\delta^2/3}.$$

Beweis:

Wir betrachten

$$f(x) = (1-x)\ln(1-x) \text{ und } g(x) = -x + \frac{1}{2}x^2 .$$

Es gilt für $0 \le x < 1$:

$$g'(x) = x - 1 \le -\ln(1-x) - 1 = f'(x)$$

sowie

$$f(0) = 0 = g(0) ,$$

also im angegebenen Intervall $f(x) \ge g(x)$.
Die Herleitung der zweiten Ungleichung erfolgt analog. □

Korollar 68

Seien $X_1, \ldots, X_n$ *unabhängige Bernoulli-verteilte Zufallsvariablen mit* $\Pr[X_i = 1] = p_i$ *und* $\Pr[X_i = 0] = 1 - p_i$. *Dann gelten folgende Ungleichungen für* $X := \sum_{i=1}^n X_i$ *und* $\mu := \mathbb{E}[X] = \sum_{i=1}^n p_i$*:*

1. $\Pr[X \geq (1+\delta)\mu] \leq e^{-\mu\delta^2/3}$ *für alle* $0 < \delta \leq 1$,
2. $\Pr[X \leq (1-\delta)\mu] \leq e^{-\mu\delta^2/2}$ *für alle* $0 < \delta \leq 1$,
3. $\Pr[|X - \mu| \geq \delta\mu] \leq 2e^{-\mu\delta^2/3}$ *für alle* $0 < \delta \leq 1$,
4. $\Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e}{1+\delta}\right)^{(1+\delta)\mu}$ *und*
5. $\Pr[X \geq t] \leq 2^{-t}$ *für* $t \geq 2e\mu$.

Beweis:

1 und 2 folgen direkt aus Satz 64 bzw. 66 und Lemma 67.

Aus 1 und 2 zusammen folgt 3.

Die Abschätzung 4 erhalten wir direkt aus Satz 64, da für den Zähler gilt

$$e^{\delta} \leq e^{(1+\delta)}.$$

5 folgt aus 4, indem man $t = (1+\delta)\mu$ setzt, $t \geq 2e\mu$:

$$\left(\frac{e}{1+\delta}\right)^{(1+\delta)\mu} \leq \left(\frac{e}{t/\mu}\right)^{t} \leq \left(\frac{1}{2}\right)^{t}.$$

$\square$

Beispiel 69

Wir betrachten wieder balls into bins und werfen n Bälle unabhängig und gleichverteilt in n Körbe. Sei

$$X_i := \text{Anzahl der Bälle im } i\text{-ten Korb}$$

für $i = 1, \ldots, n$, sowie $X := \max_{1 \le i \le n} X_i$.
Für die Analyse von X_i ($i \in \{1, \ldots, n\}$ beliebig) verwenden wir Aussage 5 von Korollar 68, mit $p_1 = \ldots = p_n = \frac{1}{n}$, $\mu = 1$ und $t = 2\log n$. Es folgt

$$\Pr[X_i \ge 2\log n] \le 1/n^2.$$

Daraus ergibt sich

$$\Pr[X \ge 2\log n] = \Pr[X_1 \ge 2\log n \vee \ldots \vee X_n \ge 2\log n] \le n \cdot \frac{1}{n^2} = \frac{1}{n}.$$

Es gilt also mit Wahrscheinlichkeit $\ge 1 - 1/n$, dass $X < 2\log n$ ist.

Literatur:

Torben Hagerup, Christine Rüb:
A guided tour of Chernoff bounds
Inf. Process. Lett. **33**, pp. 305–308 (1990)

7. Erzeugende Funktionen

7.1 Einführung

Definition 70

Für eine Zufallsvariable X mit $W_X \subseteq \mathbb{N}_0$ ist die (wahrscheinlichkeits-)erzeugende Funktion definiert durch

$$G_X(s) := \sum_{k=0}^{\infty} \Pr[X = k] \cdot s^k = \mathbb{E}[s^X]\,.$$

Die obige Definition gilt für allgemeine $s \in \mathbb{R}$, wir werden uns aber auf $s \in [-1, 1]$ konzentrieren.

Eine wahrscheinlichkeitserzeugende Funktion ist also die (gewöhnliche) erzeugende Funktion der Folge $(f_i)_{i \in \mathbb{N}_0}$ mit $f_i := \Pr[X = i]$.

Bei wahrscheinlichkeitserzeugenden Funktionen haben wir kein Problem mit der Konvergenz, da für $|s| < 1$ gilt

$$\begin{aligned}|G_X(s)| &= \left|\sum_{k=0}^{\infty} \Pr[X = k] \cdot s^k\right| \\ &\le \sum_{k=0}^{\infty} \Pr[X = k] \cdot |s^k| \le \sum_{k=0}^{\infty} \Pr[X = k] = 1\,.\end{aligned}$$

Beobachtung:

Sei $Y := X + t$ mit $t \in \mathbb{N}_0$. Dann gilt

$$G_Y(s) = \mathbb{E}[s^Y] = \mathbb{E}[s^{X+t}] = \mathbb{E}[s^t \cdot s^X] = s^t \cdot \mathbb{E}[s^X] = s^t \cdot G_X(s)\,.$$

Ebenso lässt sich leicht nachrechnen, dass

$$\begin{aligned}
G'_X(s) &= \sum_{k=1}^{\infty} k \cdot \Pr[X = k] \cdot s^{k-1}, \text{ also}\\
G'_X(0) &= \Pr[X = 1], \text{ sowie}\\
G_X^{(i)}(0) &= \Pr[X = i] \cdot i!, \text{ also}\\
G_X^{(i)}(0)/i! &= \Pr[X = i]\,.
\end{aligned}$$

Satz 71 (Eindeutigkeit der w.e. Funktion)

Die Dichte und die Verteilung einer Zufallsvariablen X mit $W_X \subseteq \mathbb{N}$ sind durch ihre wahrscheinlichkeitserzeugende Funktion eindeutig bestimmt.

Beweis:

Folgt aus der Eindeutigkeit der Potenzreihendarstellung. $\square$

Bernoulli-Verteilung

Sei X eine Bernoulli-verteilte Zufallsvariable mit $\Pr[X = 0] = 1 - p$ und $\Pr[X = 1] = p$. Dann gilt

$$G_X(s) = \mathbb{E}[s^X] = (1-p) \cdot s^0 + p \cdot s^1 = 1 - p + ps\,.$$

Gleichverteilung auf $\{0, \ldots, n\}$

Sei X auf $\{0, \ldots, n\}$ gleichverteilt, d.h. für $0 \le k \le n$ ist $\Pr[X = k] = 1/(n+1)$. Dann gilt

$$G_X(s) = \mathbb{E}[s^X] = \sum_{k=0}^{n} \frac{1}{n+1} \cdot s^k = \frac{s^{n+1} - 1}{(n+1)(s-1)}\,.$$

Binomialverteilung

Für $X \sim \mathrm{Bin}(n, p)$ gilt nach der binomischen Formel

$$G_X(s) = \mathbb{E}[s^X] = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \cdot s^k = (1 - p + ps)^n .$$

Geometrische Verteilung

Sei X eine geometrisch verteilte Zufallsvariable mit Erfolgswahrscheinlichkeit p. Dann gilt

$$\begin{aligned} G_X(s) = \mathbb{E}[s^X] &= \sum_{k=1}^{\infty} p(1-p)^{k-1} \cdot s^k \\ &= ps \cdot \sum_{k=1}^{\infty} ((1-p)s)^{k-1} = \frac{ps}{1-(1-p)s} . \end{aligned}$$

Poisson-Verteilung

Für $X \sim \mathrm{Po}(\lambda)$ gilt

$$G_X(s) = \mathbb{E}[s^X] = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \cdot s^k = e^{-\lambda+\lambda s} = e^{\lambda(s-1)}.$$

Beispiel 72

Sei X binomialverteilt mit $X \sim \mathrm{Bin}(n, \lambda/n)$, Für $n \to \infty$ folgt

$$G_X(s) = \left(1 - \frac{\lambda}{n} + \frac{\lambda s}{n}\right)^n = \left(1 + \frac{\lambda(s-1)}{n}\right)^n \to e^{\lambda(s-1)}\,.$$

Man kann beweisen, dass aus der Konvergenz der wahrscheinlichkeitserzeugenden Funktion die Konvergenz der Verteilung folgt.

7.1.1 Zusammenhang zwischen der w.e. Funktion und den Momenten

Da

$$G_X(s) := \sum_{k=0}^{\infty} \Pr[X = k] \cdot s^k = \mathbb{E}[s^X]\,,$$

gilt

$$G'_X(1) = \sum_{k=1}^{\infty} k \cdot \Pr[X = k] = \mathbb{E}[X]\,.$$

Beispiel 73

Sei X binomialverteilt mit $X \sim \mathrm{Bin}(n,p)$, also

$$G_X(s) = (1 - p + ps)^n \,.$$

Dann gilt

$$G'_X(s) = n \cdot (1 - p + ps)^{n-1} \cdot p$$

und somit

$$\mathbb{E}[X] = G'_X(1) = np \,.$$

Beispiel 73

Ebenso ergibt sich

$$\mathbb{E}[X(X-1)\dots(X-i+1)] = G_X^{(i)}(1)\,,$$

also etwa

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2\\ &= G''_X(1) + G'_X(1) - (G'_X(1))^2\,.\end{aligned}$$

Andere Momente von X kann man auf ähnliche Art und Weise berechnen.

Momenterzeugende Funktionen

Definition 74

Zu einer Zufallsvariablen X ist die momenterzeugende Funktion gemäß

$$M_X(s) := \mathbb{E}[e^{Xs}]$$

definiert.

Es gilt

$$M_X(s) = \mathbb{E}[e^{Xs}] = \mathbb{E}\left[\sum_{i=0}^{\infty} \frac{(Xs)^i}{i!}\right] = \sum_{i=0}^{\infty} \frac{\mathbb{E}[X^i]}{i!} \cdot s^i$$

und für Zufallsvariablen X mit $W_X \subseteq \mathbb{N}_0$

$$M_X(s) = \mathbb{E}[e^{Xs}] = \mathbb{E}[(e^s)^X] = G_X(e^s)\,.$$

Satz 75 (Erzeugende Funktion einer Summe)

Für unabhängige Zufallsvariablen $X_1, \ldots, X_n$ *und die Zufallsvariable* $Z := X_1 + \ldots + X_n$ *gilt*

$$G_Z(s) = G_{X_1}(s) \cdot \ldots \cdot G_{X_n}(s)\,.$$

Ebenso gilt

$$M_Z(s) = M_{X_1}(s) \cdot \ldots \cdot M_{X_n}(s)\,.$$

Beweis:

Wegen der Unabhängigkeit von $X_1, \ldots, X_n$ gilt

$$G_Z(s) = \mathbb{E}[s^{X_1+\ldots+X_n}] = \mathbb{E}[s^{X_1}] \cdot \ldots \cdot \mathbb{E}[s^{X_n}] = G_{X_1}(s) \cdot \ldots \cdot G_{X_n}(s).$$

$\square$

Beispiel 76

Seien $X_1, \ldots X_k$ mit $X_i \sim \mathrm{Bin}(n_i, p)$ unabhängige Zufallsvariable und $Z := X_1 + \ldots + X_k$. Dann gilt

$$G_Z(s) = \prod_{i=1}^{k} (1 - p + ps)^{n_i} = (1 - p + ps)^{\sum_{i=1}^{k} n_i}$$

und somit

$$Z \sim \mathrm{Bin}(\sum_{i=1}^{k} n_i, p)$$

(vgl. Satz 56).

Seien $X_1, \ldots, X_k \sim \mathrm{Po}(\lambda)$ unabhängige Zufallsvariablen. Dann folgt für $Z := X_1 + \ldots + X_k$

$$G_Z(s) = \prod_{i=1}^{k} e^{\lambda(s-1)} = e^{k\lambda(s-1)}$$

und somit $Z \sim \mathrm{Po}(k\lambda)$ (vgl. Satz 59).

7.2.1 Zufällige Summen

Wir betrachten die Situation, dass $Z := X_1 + \ldots + X_N$, wobei N ebenfalls eine Zufallsvariable ist.

Satz 77

Seien $X_1, X_2, \ldots$ unabhängige und identisch verteilte Zufallsvariablen mit der wahrscheinlichkeitserzeugenden Funktion $G_X(s)$. N sei ebenfalls eine unabhängige Zufallsvariable mit der wahrscheinlichkeitserzeugenden Funktion $G_N(s)$. Dann besitzt die Zufallsvariable $Z := X_1 + \ldots + X_N$ die wahrscheinlichkeitserzeugende Funktion $G_Z(s) = G_N(G_X(s))$.

Beweis:

Nach Voraussetzung ist $W_N \subseteq \mathbb{N}_0$. Deshalb folgt mit Satz 36

$$\begin{aligned} G_Z(s) &= \sum_{n=0}^{\infty} \mathbb{E}[s^Z \mid N = n] \cdot \Pr[N = n] \\ &= \sum_{n=0}^{\infty} \mathbb{E}[s^{X_1+\ldots+X_n}] \cdot \Pr[N = n] \\ &= \sum_{n=0}^{\infty} \mathbb{E}[s^{X_1}] \cdot \ldots \cdot \mathbb{E}[s^{X_n}] \cdot \Pr[N = n] \\ &= \sum_{n=0}^{\infty} (G_X(s))^n \cdot \Pr[N = n] \\ &= \mathbb{E}[(G_X(s))^N] \\ &= G_N(G_X(s)) \, . \end{aligned}$$

$\square$

8. Formelsammlung

8.1 Gesetze zum Rechnen mit Ereignissen

Im Folgenden seien A und B, sowie $A_1, \ldots, A_n$ Ereignisse. Die Notation $A \uplus B$ steht für $A \cup B$ und zugleich $A \cap B = \emptyset$ (disjunkte Vereinigung). $A_1 \uplus \ldots \uplus A_n = \Omega$ bedeutet also, dass die Ereignisse $A_1, \ldots, A_n$ eine Partition der Ergebnismenge Ω bilden.

$\Pr[\emptyset] = 0$
$0 \leq \Pr[A] \leq 1$
$\Pr[\bar{A}] = 1 - \Pr[A]$
$A \subseteq B \implies \Pr[A] \leq \Pr[B]$

$\forall i \neq j : A_i \cap A_j = \emptyset \implies$ $\Pr\left[\bigcup_{i=1}^n A_i\right] = \sum_{i=1}^n \Pr[A_i]$	Additionssatz
$\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$ allgemeine Form: siehe Satz 9	Inklusion/Exklusion, Siebformel
$\Pr\left[\bigcup_{i=1}^n A_i\right] \leq \sum_{i=1}^n \Pr[A_i]$	Boolesche Ungleichung
$\Pr[A\|B] = \frac{\Pr[A \cap B]}{\Pr[B]}$ für $\Pr[B] > 0$	Def. bedingte Ws.

$B \subseteq A_1 \uplus \ldots \uplus A_n \implies$ $\Pr[B] = \sum_{i=1}^{n} \Pr[B\|A_i] \cdot \Pr[A_i]$	Satz von der totalen Wahrscheinlichkeit
$\Pr[B] > 0,\ B \subseteq A_1 \uplus \ldots \uplus A_n \implies$ $\Pr[A_i\|B] = \frac{\Pr[B\|A_i]\cdot\Pr[A_i]}{\sum_{i=1}^{n} \Pr[B\|A_i]\cdot\Pr[A_i]}$	Satz von Bayes
$\Pr[A_1 \cap \ldots \cap A_n] = \Pr[A_1] \cdot \Pr[A_2\|A_1] \cdot$ $\ldots \cdot \Pr[A_n\|A_1 \cap \ldots \cap A_{n-1}]$	Multiplikationssatz
A und B unabhängig $\iff$ $\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$	Definition Unabhängigkeit

8.2 Erwartungswert und Varianz diskreter Zufallsvariablen

Sei X eine diskrete Zufallsvariable. Für Erwartungswert und Varianz gelten die folgenden Formeln (sofern $\mathbb{E}[X]$ und $\mathrm{Var}[X]$ existieren).

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x \in W_X} x \cdot \Pr[X = x] \\ &= \sum_{\omega \in \Omega} X(\omega) \cdot \Pr[\omega] \\ \Big(&= \sum_{i=1}^{\infty} \Pr[X \geq i], \quad \text{falls } W_X \subseteq \mathbb{N}_0 \Big)\end{aligned}$$

Erwartungswert

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \sum_{x \in W_X} \Pr[X = x] \cdot (x - \mathbb{E}[X])^2\end{aligned}$$

Varianz

8.3 Gesetze zum Rechnen mit Zufallsvariablen

Seien a, b, $a_1, \ldots, a_n \in \mathbb{R}$, $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$.

$X_1, \ldots, X_n$ unabhängig $\iff$	für alle $(a_1, \ldots, a_n)$: $\Pr[X_1 = a_1, \ldots, X_n = a_n] = \Pr[X_1 = a_1] \cdot \ldots \cdot \Pr[X_n = a_n]$
$X_1, \ldots, X_n$ unabhängig $\implies$	$f_1(X_1), \ldots, f_n(X_n)$ unabhängig
$\mathbb{E}[a \cdot X + b] = a \cdot \mathbb{E}[X] + b$	

$X(\omega) \leq Y(\omega)$ für alle $\omega \in \Omega \implies \mathbb{E}[X] \leq \mathbb{E}[Y]$	Monotonie des Erwartungswerts
$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X\|A_i] \cdot \Pr[A_i]$	
$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$	
$\mathrm{Var}[a \cdot X + b] = a^2 \cdot \mathrm{Var}[X]$	

$\mathbb{E}[a_1 X_1 + \ldots + a_n X_n]$ $= a_1 \mathbb{E}[X_1] + \ldots + a_n \mathbb{E}[X_n]$	Linearität des Erwartungswerts
$X_1, \ldots, X_n$ unabhängig $\Longrightarrow$ $\mathbb{E}[X_1 \cdot \ldots \cdot X_n] = \mathbb{E}[X_1] \cdot \ldots \cdot \mathbb{E}[X_n]$	Multiplikativität des Erwartungswerts
$X_1, \ldots, X_n$ unabhängig $\Longrightarrow$ $\mathrm{Var}[X_1 + \ldots + X_n] = \mathrm{Var}[X_1] + \ldots + \mathrm{Var}[X_n]$	Varianz einer Summe

$X \geq 0 \implies$ $\Pr[X \geq t] \leq \mathbb{E}[X]/t$ für $t > 0$	Markov
$\Pr[\lvert X - \mathbb{E}[X]\rvert \geq t]$ $\leq \mathrm{Var}[X]/t^2$ für $t > 0$	Chebyshev
siehe Satz 63	Gesetz der großen Zahlen

Kapitel II Kontinuierliche Wahrscheinlichkeitsräume

1. Einführung

1.1 Motivation

Interpretation der Poisson-Verteilung als Grenzwert der Binomialverteilung.

Beispiel 78

Wir betrachten das Szenario: Bei einem Druckerserver kommen Aufträge in einer Warteschlange an, die alle $1/n$ Zeiteinheiten vom Server abgefragt wird. Der Server nimmt also zu den diskreten Zeitpunkte $1/n, 2/n, 3/n, \ldots$ neue Aufträge entgegen. Durch den Grenzwert $n \to \infty$ „verschmelzen“ diese diskreten Zeitpunkte zu einer kontinuierlichen Zeitachse, und für die Zufallsvariable T, welche die Zeitspanne bis zum Eintreffen des nächsten Auftrags misst, reicht eine diskrete Wertemenge W_T nicht mehr aus.

Definition 79

Eine kontinuierliche oder auch stetige Zufallsvariable X und ihr zugrunde liegender kontinuierlicher (reeller) Wahrscheinlichkeitsraum sind definiert durch eine integrierbare Dichte(-funktion) $f_X : \mathbb{R} \to \mathbb{R}_0^+$ mit der Eigenschaft

$$\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}\,x = 1.$$

Eine Menge $A \subseteq \mathbb{R}$, die durch Vereinigung $A = \bigcup_k I_k$ abzählbar vieler paarweise disjunkter Intervalle beliebiger Art (offen, geschlossen, halboffen, einseitig unendlich) gebildet werden kann, heißt Ereignis. Ein Ereignis A tritt ein, wenn X einen Wert aus A annimmt. Die Wahrscheinlichkeit von A ist bestimmt durch

$$\Pr[A] = \int_A f_X(x)\,\mathrm{d}\,x = \sum_k \int_{I_k} f_X(x)\,\mathrm{d}\,x.$$

Beispiel 80 (Gleichverteilung)

Eine besonders einfache kontinuierliche Dichte stellt die Gleichverteilung auf dem Intervall $[a, b]$ dar. Sie ist definiert durch

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } x \in [a, b], \\ 0 & \text{sonst.} \end{cases}$$

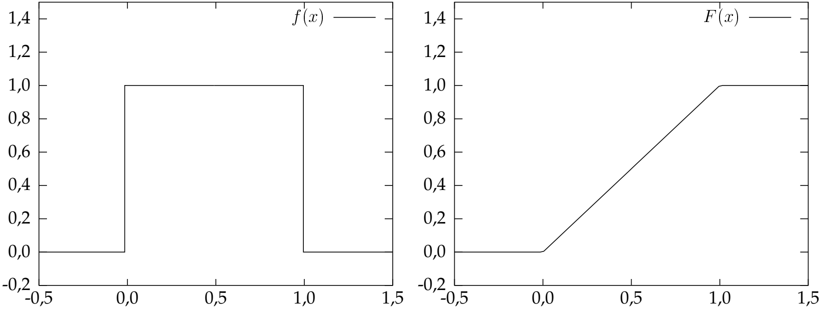
Analog zum diskreten Fall ordnen wir jeder Dichte f_X eine Verteilung oder Verteilungsfunktion F_X zu:

$$F_X(x) := \Pr[X \leq x] = \Pr[\{t \in \mathbb{R} \mid t \leq x\}] = \int_{-\infty}^{x} f_X(t) \, \mathrm{d}\, t.$$

Beispiel 81

Die Verteilungsfunktion der Gleichverteilung:

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}\,t = \begin{cases} 0 & \text{für } x < a, \\ \frac{x-a}{b-a} & \text{für } a \le x \le b, \\ 1 & \text{für } x > b. \end{cases}$$



Gleichverteilung über dem Intervall $[0, 1]$

Beobachtungen: *(Eigenschaften der Verteilungsfunktion)*

- F_X ist monoton steigend.
- F_X ist stetig. Man spricht daher auch von einer „stetigen Zufallsvariablen".
- Es gilt: $\lim_{x \to -\infty} F_X(x) = 0$ und $\lim_{x \to \infty} F_X(x) = 1$.
- Jeder (außer an endlich vielen Punkten) differenzierbaren Funktion F, welche die zuvor genannten Eigenschaften erfüllt, können wir eine Dichte f durch $f(x) = F'(x)$ zuordnen.

Es gilt

$$\Pr[a < X \leq b] = F_X(b) - F_X(a)\,.$$

Bei den von uns betrachteten Dichten besteht zwischen den Ereignissen „$a < X \leq b$“, „$a \leq X \leq b$“, „$a \leq X < b$“ und „$a < X < b$“ kein wesentlicher Unterschied, da

$$\int_{[a,b]} f(t)\,\mathrm{d}\,t = \int_{]a,b]} f(t)\,\mathrm{d}\,t = \int_{[a,b[} f(t)\,\mathrm{d}\,t = \int_{]a,b[} f(t)\,\mathrm{d}\,t.$$

1.3 Kolmogorov-Axiome und σ-Algebren

1.3.1 σ-Algebren

Definition 82

Sei Ω eine Menge. Eine Menge $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ heißt σ-Algebra über Ω, wenn folgende Eigenschaften erfüllt sind:

(E1) $\Omega \in \mathcal{A}$.

(E2) Wenn $A \in \mathcal{A}$, dann folgt $\bar{A} \in \mathcal{A}$.

(E3) Für $n \in \mathbb{N}$ sei $A_n \in \mathcal{A}$. Dann gilt auch $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$.

Für jede (endliche) Menge Ω stellt die Menge $\mathcal{P}(\Omega)$ eine σ-Algebra dar.

Für $\Omega = \mathbb{R}$ ist die Klasse der *Borel'schen Mengen*, die aus allen Mengen $A \subseteq \mathbb{R}$ besteht, welche sich durch abzählbare Vereinigungen und Schnitte von Intervallen (offen, halboffen oder geschlossen) darstellen lassen, eine σ-Algebra.

1.3.2 Kolmogorov-Axiome

Definition 83 (Wahrscheinlichkeitsraum, Kolmogorov-Axiome)

Sei Ω eine beliebige Menge und $\mathcal{A}$ eine σ-Algebra über Ω. Eine Abbildung

$$\Pr[.] \ : \ \mathcal{A} \to [0,1]$$

heißt Wahrscheinlichkeitsmaß auf $\mathcal{A}$, wenn sie folgende Eigenschaften besitzt:

1. (W1) $\Pr[\Omega] = 1$.
2. (W2) $A_1, A_2, \ldots$ seien paarweise disjunkte Ereignisse. Dann gilt
$$\Pr\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \Pr[A_i].$$

Für ein Ereignis $A \in \mathcal{A}$ heißt $\Pr[A]$ Wahrscheinlichkeit von A. Ein Wahrscheinlichkeitsraum ist definiert durch das Tupel $(\Omega, \mathcal{A}, \Pr)$.

Die in obiger Definition aufgelisteten Eigenschaften eines Wahrscheinlichkeitsmaßes wurden von dem russischen Mathematiker Andrei Nikolaevich Kolmogorov (1903–1987) formuliert. Kolmogorov gilt als einer der Pioniere der modernen Wahrscheinlichkeitstheorie, leistete jedoch auch bedeutende Beiträge zu zahlreichen anderen Teilgebieten der Mathematik. Informatikern begegnet sein Name auch im Zusammenhang mit der so genannten Kolmogorov-Komplexität, einem relativ jungen Zweig der Komplexitätstheorie.

Die Eigenschaften in obiger Definition nennt man auch Kolmogorov-Axiome.

Lemma 84

Sei $(\Omega, \mathcal{A}, \Pr)$ *ein Wahrscheinlichkeitsraum. Für Ereignisse* A, B, A_1, A_2, $\ldots$ *gilt*

1. $\Pr[\emptyset] = 0$, $\Pr[\Omega] = 1$.
2. $0 \leq \Pr[A] \leq 1$.
3. $\Pr[\bar{A}] = 1 - \Pr[A]$.
4. *Wenn* $A \subseteq B$, *so folgt* $\Pr[A] \leq \Pr[B]$.

Lemma 84

5. *(Additionssatz) Wenn die Ereignisse $A_1, \ldots, A_n$ paarweise disjunkt sind, so folgt*

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} \Pr[A_i].$$

Für disjunkte Ereignisse A, B erhalten wir insbesondere

$$\Pr[A \cup B] = \Pr[A] + \Pr[B].$$

Für eine unendliche Menge von paarweise disjunkten Ereignissen $A_1, A_2, \ldots$ gilt analog $\Pr\left[\bigcup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \Pr[A_i]$.

Beweis:

Wenn wir in Eigenschaft (W2) $A_1 = \Omega$ und $A_2, A_3, \ldots = \emptyset$ setzen, so ergibt die Eigenschaft, dass $\Pr[\Omega] + \sum_{i=2}^{\infty} \Pr[\emptyset] = \Pr[\Omega]$. Daraus folgt $\Pr[\emptyset] = 0$.

Regel 2 und Regel 5 gelten direkt nach Definition der Kolmogorov-Axiome und Regel 1.

Regel 3 erhalten wir mit Regel 5 wegen $1 = \Pr[\Omega] = \Pr[A] + \Pr[\bar{A}]$.

Für Regel 4 betrachten wir die disjunkten Ereignisse A und $C := B \setminus A$, für die gilt, dass $A \cup B = A \cup C$. Mit Regel 5 folgt die Behauptung. $\square$

1.3.3 Lebesgue-Integrale

Eine Funktion $f : \mathbb{R} \to \mathbb{R}$ heißt messbar, falls das Urbild jeder Borel'schen Menge ebenfalls eine Borel'sche Menge ist.
Z.B. ist für jede Borel'sche Menge A die Indikatorfunktion

$$I_A : x \mapsto \begin{cases} 1 & \text{falls } x \in A, \\ 0 & \text{sonst} \end{cases}$$

messbar. Jede stetige Funktion ist messbar. Auch Summen und Produkte von messbaren Funktionen sind wiederum messbar.
Jeder messbaren Funktion kann man ein Integral, das so genannte Lebesgue-Integral, geschrieben $\int f \,\mathrm{d}\,\lambda$, zuordnen.

Ist $f : \mathbb{R} \to \mathbb{R}_0^+$ eine messbare Funktion, so definiert

$$\Pr : A \mapsto \int f \cdot I_A \, \mathrm{d}\,\lambda$$

eine Abbildung auf den Borel'schen Mengen, die die Eigenschaft (W2) der Kolmogorov-Axiome erfüllt. Gilt daher zusätzlich noch $\Pr[\mathbb{R}] = 1$, so definiert f auf natürliche Weise einen Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \Pr)$, wobei $\Omega = \mathbb{R}$ und $\mathcal{A}$ die Menge der Borel'schen Mengen ist.

1.4 Rechnen mit kontinuierlichen Zufallsvariablen

1.4.1 Funktionen kontinuierlicher Zufallsvariablen

Sei $Y := g(X)$ mit einer Funktion $g : \mathbb{R} \to \mathbb{R}$.
Die Verteilung von Y erhalten wir durch

$$F_Y(y) = \Pr[Y \le y] = \Pr[g(X) \le y] = \int_C f_X(t)\,\mathrm{d}\,t.$$

Hierbei bezeichnet $C := \{t \in \mathbb{R} \mid g(t) \le y\}$ alle reellen Zahlen $t \in \mathbb{R}$, für welche die Bedingung „$Y \le y$" zutrifft. Das Integral über C ist nur dann sinnvoll definiert, wenn C ein zulässiges Ereignis darstellt. Aus der Verteilung F_Y können wir durch Differenzieren die Dichte f_Y ermitteln.

Beispiel 85

Sei X gleichverteilt auf dem Intervall $]0,1[$. Für eine Konstante $\lambda > 0$ definieren wir die Zufallsvariable $Y := -(1/\lambda)\ln X$.

$$\begin{aligned}
F_Y(y) &= \Pr[-(1/\lambda)\ln X \le y] = \Pr[\ln X \ge -\lambda y] \\
&= \Pr[X \ge e^{-\lambda y}] \\
&= 1 - F_X(e^{-\lambda y}) \\
&= \begin{cases} 1 - e^{-\lambda y} & \text{für } y \ge 0, \\ 0 & \text{sonst.} \end{cases}
\end{aligned}$$

Beispiel (Forts.)

Damit folgt mit $f_Y(y) = F'_Y(y)$ *sofort*

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda y} & \text{für } y \geq 0, \\ 0 & \text{sonst.} \end{cases}$$

Eine Zufallsvariable mit einer solchen Dichte f_Y *nennt man exponentialverteilt.*

Beispiel 86

Sei X eine beliebige Zufallsvariable. Für $a, b \in \mathbb{R}$ mit $a > 0$ definieren wir die Zufallsvariable $Y := a \cdot X + b$.
Es gilt

$$F_Y(y) = \Pr[aX + b \leq y] = \Pr\left[X \leq \frac{y-b}{a}\right] = F_X\left(\frac{y-b}{a}\right),$$

und somit

$$f_Y(y) = \frac{\mathrm{d}\, F_Y(y)}{\mathrm{d}\, y} = \frac{\mathrm{d}\, F_X((y-b)/a)}{\mathrm{d}\, y} = f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{a}\,.$$

Simulation von Zufallsvariablen

Unter der Simulation einer Zufallsvariablen X mit Dichte f_X versteht man die algorithmische Erzeugung von Zufallswerten, deren Verteilung der Verteilung von X entspricht.

Dazu nehmen wir an, dass die zu simulierende Zufallsvariable X eine stetige, im Bildbereich $]0,1[$ streng monoton wachsende Verteilungsfunktion F_X besitzt. Weiter nehmen wir an, dass U eine auf $]0,1[$ gleichverteilte Zufallsvariable ist, die wir simulieren können.

Aus unserer Annahme über F_X folgt, dass es zu F_X eine (eindeutige) inverse Funktion F_X^{-1} gibt mit $F_X(F_X^{-1}(x)) = x$ für alle $x \in]0,1[$.

Sei nun

$$\tilde{X} := F_X^{-1}(U)\,,$$

dann gilt

$$\begin{aligned}\Pr[\tilde{X} \le t] &= \Pr[F_X^{-1}(U) \le t]\\ &= \Pr[U \le F_X(t)]\\ &= F_U(F_X(t))\\ &= F_X(t)\,.\end{aligned}$$

Beispiel 87

Im obigen Beispiel der Exponentialverteilung gilt $F_X(t) = 1 - e^{-t}$ für $t \geq 0$, und wir erhalten auf $]0,1[$ die Umkehrfunktion $F_X^{-1}(t) = -\ln(1-t)$. Also gilt $\tilde{X} = F_X^{-1}(U) = -\ln(1-U)$.

Statt $\tilde{X}$ haben wir im Beispiel die Zufallsvariable $-\ln U$ betrachtet, die aber offensichtlich dieselbe Verteilung besitzt.

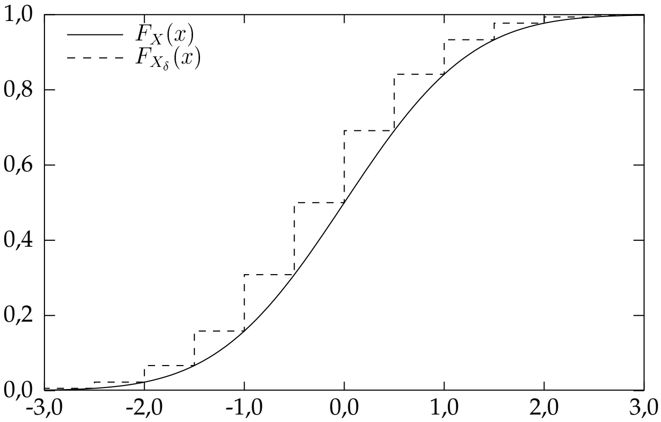
1.4.2 Kontinuierliche Zufallsvariablen als Grenzwerte diskreter Zufallsvariablen

Sei X eine kontinuierliche Zufallsvariable. Wir können aus X leicht eine diskrete Zufallsvariable konstruieren, indem wir für ein festes $\delta > 0$ definieren

$$X_\delta = n\delta \iff X \in [n\delta, (n+1)\delta[\text{ für } n \in \mathbb{Z}.$$

Für X_δ gilt

$$\Pr[X_\delta = n\delta] = F_X((n+1)\delta) - F_X(n\delta)\,.$$



Für $\delta \to 0$ nähert sich die Verteilung von X_δ der Verteilung von X immer mehr an.

1.4.3 Erwartungswert und Varianz

Definition 88

Für eine kontinuierliche Zufallsvariable X ist der Erwartungswert definiert durch

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} t \cdot f_X(t) \,\mathrm{d}\, t,$$

sofern das Integral $\int_{-\infty}^{\infty} |t| \cdot f_X(t) \,\mathrm{d}\, t$ endlich ist.

Für die Varianz gilt entsprechend

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_{-\infty}^{\infty} (t - \mathbb{E}[X])^2 \cdot f_X(t) \,\mathrm{d}\, t,$$

wenn $\mathbb{E}[(X - \mathbb{E}[X])^2]$ existiert.

Lemma 89

Sei X eine kontinuierliche Zufallsvariable, und sei

$$Y := g(X)\,.$$

Dann gilt

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} g(t) \cdot f_X(t)\,\mathrm{d}\,t\,.$$

Beweis:

Wir zeigen die Behauptung nur für den einfachen Fall, dass g eine lineare Funktion ist, also $Y := a \cdot X + b$ für $a, b \in \mathbb{R}$ und $a > 0$.

Es gilt (siehe obiges Beispiel)

$$\mathbb{E}[a \cdot X + b] = \int_{-\infty}^{\infty} t \cdot f_Y(t) \, \mathrm{d}\, t = \int_{-\infty}^{\infty} t \cdot f_X\left(\frac{t-b}{a}\right) \cdot \frac{1}{a} \, \mathrm{d}\, t.$$

Durch die Substitution $u := (t-b)/a$ mit $\mathrm{d}\, u = (1/a) \, \mathrm{d}\, t$ erhalten wir

$$\mathbb{E}[a \cdot X + b] = \int_{-\infty}^{\infty} (au + b) f_X(u) \, \mathrm{d}\, u.$$

$\square$

Beispiel 90

Für Erwartungswert und Varianz der Gleichverteilung ergibt sich

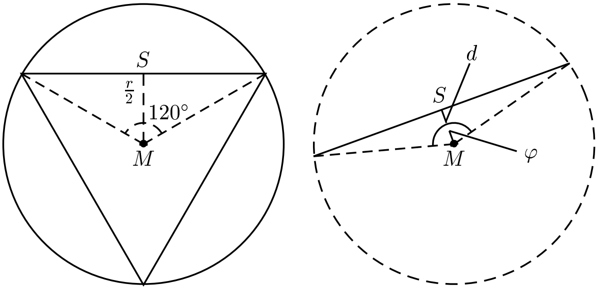
$$
\begin{aligned}
\mathbb{E}[X] &= \int_a^b t \cdot \frac{1}{b-a}\,\mathrm{d}\,t = \frac{1}{b-a} \cdot \int_a^b t \cdot \mathrm{d}\,t \\
&= \frac{1}{2(b-a)} \cdot [t^2]_a^b \\
&= \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}, \\
\mathbb{E}[X^2] &= \frac{1}{b-a} \cdot \int_a^b t^2 \cdot \mathrm{d}\,t = \frac{b^2+ba+a^2}{3}, \\
\mathrm{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \ldots = \frac{(a-b)^2}{12}\,.
\end{aligned}
$$

1.4.4 Laplace-Prinzip in kontinuierlichen Wahrscheinlichkeitsräumen

Das folgende Beispiel zeigt, dass im kontinuierlichen Fall die Bedeutung von „gleichwahrscheinlich“ nicht immer ganz klar sein muss.

Bertrand'sches Paradoxon

Wir betrachten einen Kreis mit einem eingeschriebenen gleichseitigen Dreieck. Was ist die Wahrscheinlichkeit, mit der die Länge einer zufällig gewählten Sehne die Seitenlänge dieses Dreiecks übersteigt (Ereignis A)?


S
$\frac{r}{2}$
120°
M
d
S
M
φ

Beobachtungen:

- Die Seiten des Dreiecks haben Abstand $\frac{r}{2}$ vom Mittelpunkt M.
- Die Lage jeder Sehne ist (bis auf Rotation um M) durch einen der folgenden Parameter festgelegt:
  - Abstand d zum Kreismittelpunkt,
  - Winkel φ mit dem Kreismittelpunkt.

Wir nehmen für jeden dieser Parameter Gleichverteilung an und ermitteln $\Pr[A]$.

1. Sei $d \in [0, r]$ gleichverteilt. A tritt ein, wenn $d < \frac{r}{2}$, und es folgt $\Pr[A] = \frac{1}{2}$.
2. Sei $\varphi \in [0°, 180°]$ gleichverteilt. Für A muss gelten $\varphi \in]120°, 180°]$, und es folgt somit $\Pr[A] = \frac{1}{3}$.

Siehe auch diese graphischen Darstellungen!

2. Wichtige stetige Verteilungen

2.1 Gleichverteilung

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für } x \in [a,b], \\ 0 & \text{sonst.} \end{cases}$$

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}\,t = \begin{cases} 0 & \text{für } x < a, \\ \frac{x-a}{b-a} & \text{für } a \le x \le b, \\ 1 & \text{für } x > b. \end{cases}$$

$$\mathbb{E}[X] = \frac{a+b}{2} \text{ und } \mathrm{Var}[X] = \frac{(a-b)^2}{12}\,.$$

2.2 Normalverteilung

Die Normalverteilung nimmt unter den stetigen Verteilungen eine besonders prominente Position ein.

Definition 91

Eine Zufallsvariable X mit Wertebereich $W_X = \mathbb{R}$ heißt normalverteilt mit den Parametern $\mu \in \mathbb{R}$ und $\sigma \in \mathbb{R}^+$, wenn sie die Dichte

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) =: \varphi(x; \mu, \sigma)$$

besitzt.

In Zeichen schreiben wir $X \sim \mathcal{N}(\mu, \sigma^2)$.
$\mathcal{N}(0, 1)$ heißt Standardnormalverteilung. Die zugehörige Dichte $\varphi(x; 0, 1)$ kürzen wir durch $\varphi(x)$ ab.

Die Verteilungsfunktion zu $\mathcal{N}(\mu, \sigma^2)$ ist

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \mathrm{d}\, t =: \Phi(x; \mu, \sigma)\,.$$

Diese Funktion heißt Gauß'sche Φ-Funktion (φ ist nicht geschlossen integrierbar).

Lemma 92

$$I := \int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}\,x = \sqrt{2\pi}.$$

Beweis:

Wir berechnen zunächst I^2:

$$\begin{aligned} I^2 &= \left(\int_{-\infty}^{\infty} e^{-x^2/2}\,\mathrm{d}\,x\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\,\mathrm{d}\,y\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,\mathrm{d}\,x\,\mathrm{d}\,y\,. \end{aligned}$$

Wir gehen nun zu Polarkoordinaten über und setzen $x := r\cos\phi$ und $y := r\sin\phi$. Dann ist

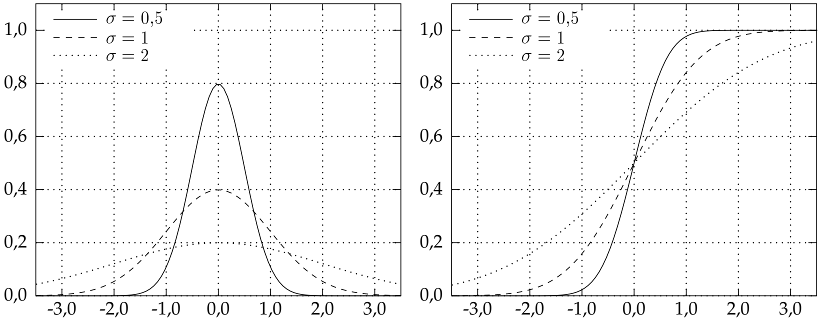
$$\begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} \\ \frac{\partial x}{\partial \phi} & \frac{\partial y}{\partial \phi} \end{vmatrix} = \begin{vmatrix} \cos\phi & \sin\phi \\ -r\sin\phi & r\cos\phi \end{vmatrix} = r(\cos^2\phi + \sin^2\phi) = r$$

Beweis (Forts.):

und wir erhalten

$$I^2 = \int_0^{2\pi}\int_0^{\infty} e^{-r^2/2} r \,\mathrm{d}\, r \,\mathrm{d}\,\phi \;=\; \int_0^{2\pi} \left[-e^{-r^2/2}\right]_0^{\infty} \mathrm{d}\,\phi$$
$$= \int_0^{2\pi} 1 \,\mathrm{d}\,\phi \;=\; 2\pi.$$

$\square$



Dichte und Verteilung von $\mathcal{N}(0, \sigma^2)$

Satz 93 (Lineare Transformation der Normalverteilung)

Sei X eine normalverteilte Zufallsvariable mit $X \sim \mathcal{N}(\mu, \sigma^2)$. Dann gilt für beliebiges $a \in \mathbb{R} \setminus \{0\}$ und $b \in \mathbb{R}$, dass $Y = aX + b$ normalverteilt ist mit $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

Beweis:

Wir betrachten zunächst den Fall „$a > 0$“:

$$\begin{aligned}\Pr[Y \leq y] = \Pr[aX + b \leq y] &= \Pr\left[X \leq \frac{y-b}{a}\right] \\ &= \frac{1}{\sqrt{2\pi}\sigma} \cdot \int_{-\infty}^{(y-b)/a} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) \mathrm{d}\, u.\end{aligned}$$

Nach der Substitution $u = (v - b)/a$ und d $u = (1/a) \cdot$ d v erhalten wir

Beweis (Forts.):

$$\Pr[Y \le y] = \frac{1}{\sqrt{2\pi}a\sigma} \cdot \int_{-\infty}^{y} \exp\left(-\frac{(v-a\mu-b)^2}{2a^2\sigma^2}\right) \mathrm{d}\, v\,.$$

Also $Y \sim \mathcal{N}(a\mu+b, a^2\sigma^2)$. Für $a<0$ verläuft der Beweis analog. $\square$

Sei also X eine beliebige $\mathcal{N}(\mu, \sigma^2)$-verteilte Zufallsvariable X und $Y := \frac{X-\mu}{\sigma}$.

Dann ist nach Satz 93 Y $\mathcal{N}(0,1)$-verteilt. Y heißt auch normiert.

Ferner gilt

$$\begin{aligned}
\Pr[a < X \leq b] &= \Pr\left[\frac{a-\mu}{\sigma} < Y \leq \frac{b-\mu}{\sigma}\right] \\
&= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) .
\end{aligned}$$

Satz 94

X sei $\mathcal{N}(0,1)$-verteilt. Dann gilt

$$\mathbb{E}[X] = 0 \textit{ und } \quad \mathrm{Var}[X] = 1.$$

Beweis:

$$\mathbb{E}[X] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \cdot \exp\left(-\frac{x^2}{2}\right) \mathrm{d}\, x.$$

Da der Integrand punktsymmetrisch zu $(0,0)$ ist, folgt $\mathbb{E}[X] = 0$.

Beweis (Forts.):

Mittels Lemma 92 und durch partielle Integration erhalten wir

$$\begin{aligned}\sqrt{2\pi} &= \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}\, x\\ &= \underbrace{x \exp\left(-\frac{x^2}{2}\right)\Big|_{-\infty}^{\infty}}_{=0} + \int_{-\infty}^{\infty} x^2 \cdot \exp\left(-\frac{x^2}{2}\right) \mathrm{d}\, x\end{aligned}$$

Daraus folgt, dass $\mathbb{E}[X^2] = 1$ ist und somit $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 1$. □

Satz 95

X sei $\mathcal{N}(\mu, \sigma^2)$-verteilt. Dann gilt

$$\mathbb{E}[X] = \mu \textit{ und } \mathrm{Var}[X] = \sigma^2 .$$

Beweis:

$Y := \frac{X-\mu}{\sigma}$ ist standardnormalverteilt. Ferner gilt gemäß der Rechenregeln für Erwartungswert und Varianz

$$\mathbb{E}[X] = \mathbb{E}[\sigma Y + \mu] = \sigma \cdot \mathbb{E}[Y] + \mu = \mu$$

und

$$\mathrm{Var}[X] = \mathrm{Var}[\sigma Y + \mu] = \sigma^2 \cdot \mathrm{Var}[Y] = \sigma^2.$$

$\square$

2.3 Exponentialverteilung

Die Exponentialverteilung ist in gewisser Weise das kontinuierliche Analogon zur geometrischen Verteilung. Wie die geometrische Verteilung ist sie „gedächtnislos". Sie spielt daher vor allem bei der Modellierung von Wartezeiten eine große Rolle.

Definition 96

Eine Zufallsvariable X heißt exponentialverteilt mit dem Parameter λ, $\lambda > 0$, wenn sie die Dichte

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda x} & \text{falls } x \geq 0, \\ 0 & \text{sonst} \end{cases}$$

besitzt.

Für die entsprechende Verteilungsfunktion gilt (für $x \geq 0$)

$$F(x) = \int_0^x \lambda \cdot e^{-\lambda t} \,\mathrm{d}\, t = \left[-e^{-\lambda t}\right]_0^x = 1 - e^{-\lambda x}.$$

Für $x < 0$ gilt selbstverständlich $F(x) = 0$.

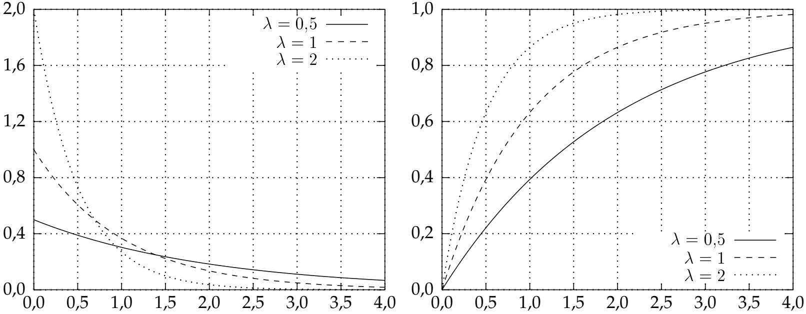
$$
\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty t \cdot \lambda \cdot e^{-\lambda t} \,\mathrm{d}\, t \\
&= \left[t \cdot (-e^{-\lambda t}) \right]_0^\infty + \int_0^\infty e^{-\lambda t} \,\mathrm{d}\, t \\
&= 0 + \left[-\frac{1}{\lambda} \cdot e^{-\lambda t} \right]_0^\infty = \frac{1}{\lambda}\,.
\end{aligned}
$$

Analog erhalten wir

$$\begin{aligned}\mathbb{E}[X^2] &= \int_0^\infty t^2 \cdot \lambda \cdot e^{-\lambda t}\,\mathrm{d}\,t \\ &= \left[t^2 \cdot (-e^{-\lambda t})\right]_0^\infty + \int_0^\infty 2t \cdot e^{-\lambda t}\,\mathrm{d}\,t \\ &= 0 + \frac{2}{\lambda} \cdot \mathbb{E}[X] = \frac{2}{\lambda^2}\end{aligned}$$

und somit

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{\lambda^2}\,.$$



Dichte und Verteilung der Exponentialverteilung

2.3.1 Eigenschaften der Exponentialverteilung

Satz 97 (Skalierung exponentialverteilter Variablen)

Sei X eine exponentialverteilte Zufallsvariable mit dem Parameter λ. Für $a > 0$ ist die Zufallsvariable $Y := aX$ wieder exponentialverteilt mit dem Parameter λ/a.

Beweis:

$$\begin{aligned} F_Y(x) &= \Pr[Y \leq x] = \Pr[aX \leq x] \\ &= \Pr\left[X \leq \frac{x}{a}\right] = F_X\left(\frac{x}{a}\right) \\ &= 1 - e^{-\frac{\lambda x}{a}}\,. \end{aligned}$$

$\square$

Gedächtnislosigkeit

Satz 98 (Gedächtnislosigkeit)

Eine (positive) kontinuierliche Zufallsvariable X mit Wertebereich $\mathbb{R}^+$ ist genau dann exponentialverteilt, wenn für alle $x, y > 0$ gilt, dass

$$\Pr[X > x + y \mid X > y] = \Pr[X > x]\,. \qquad (*)$$

Beweis:

Sei X exponentialverteilt mit Parameter λ. Dann gilt

$$\begin{aligned} \Pr[X > x + y \mid X > y] &= \frac{\Pr[X > x + y, X > y]}{\Pr[X > y]} \\ &= \frac{\Pr[X > x + y]}{\Pr[X > y]} \\ &= \frac{e^{-\lambda(x+y)}}{e^{-\lambda y}} = e^{-\lambda x} = \Pr[X > x]\,. \end{aligned}$$

Beweis (Forts.):

Sei umgekehrt X eine kontinuierliche Zufallsvariable, die die Gleichung $(*)$ erfüllt. Wir definieren $g(x) := \Pr[X > x]$. Für $x, y > 0$ gilt

$$\begin{aligned} g(x+y) &= \Pr[X > x+y] \\ &= \Pr[X > x+y \mid X > y] \cdot \Pr[X > y] \\ &= \Pr[X > x] \cdot \Pr[X > y] = g(x)g(y)\,. \end{aligned}$$

Daraus folgt durch wiederholte Anwendung

$$g(1) = g\Big(\underbrace{\frac{1}{n} + \cdots + \frac{1}{n}}_{n\text{-mal}}\Big) = \left(g\Big(\frac{1}{n}\Big)\right)^n \text{ für alle } n \in \mathbb{N}$$

und somit insbesondere auch $g(1/n) = (g(1))^{1/n}$.

Beweis (Forts.):

Da X nur positive Werte annimmt, muss es ein $n \in \mathbb{N}$ geben mit $g(1/n) > 0$. Wegen $0 < g(1) \leq 1$ muss es daher auch ein $\lambda \geq 0$ geben mit $g(1) = e^{-\lambda}$.

Nun gilt für beliebige $p, q \in \mathbb{N}$

$$g(p/q) = g(1/q)^p = g(1)^{p/q},$$

und somit $g(r) = e^{-\lambda r}$ für alle $r \in \mathbb{Q}^+$.

Aufgrund der Stetigkeit folgt daraus

$$g(x) = e^{-\lambda x} .$$

□

Beispiel 99

Über das Cäsium-Isotop $^{134}_{55}\mathrm{Cs}$ ist bekannt, dass es eine mittlere Lebensdauer von ungefähr 3,03 Jahren oder $1{,}55 \cdot 10^6$ Minuten besitzt. Die Zufallsvariable X messe die Lebenszeit eines bestimmten $^{134}_{55}\mathrm{Cs}$-Atoms. X ist exponentialverteilt mit dem Parameter

$$\lambda = \frac{1}{\mathbb{E}[X]} = \frac{1}{1{,}55 \cdot 10^6} \approx 0{,}645 \cdot 10^{-6} \left[\frac{1}{\mathsf{min}}\right]$$

Da λ den Kehrwert einer Zeit als Einheit besitzt, spricht man von der Zerfallsrate. Auch bei anderen Anwendungen ist es üblich, λ als Rate einzuführen.

2.3.2 Exponentialverteilung als Grenzwert der geometrischen Verteilung

Erinnerung: Die Poisson-Verteilung lässt sich als Grenzwert der Binomialverteilung darstellen.

Wir betrachten eine Folge geometrisch verteilter Zufallsvariablen X_n mit Parameter $p_n = \lambda/n$. Für ein beliebiges $k \in \mathbb{N}$ ist die Wahrscheinlichkeit, dass $X_n \le k \cdot n$, gleich

$$\begin{aligned}\Pr[X_n \le kn] &= \sum_{i=1}^{kn} (1-p_n)^{i-1} \cdot p_n \;=\; p_n \cdot \sum_{i=0}^{kn-1} (1-p_n)^i \\ &= p_n \cdot \frac{1-(1-p_n)^{kn}}{p_n} \;=\; 1 - \left(1 - \frac{\lambda}{n}\right)^{kn} .\end{aligned}$$

Wegen $\lim_{n\to\infty}(1-\frac{\lambda}{n})^n = e^{-\lambda}$ gilt daher für die Zufallsvariablen $Y_n := \frac{1}{n}X_n$, dass

$$\begin{aligned}\lim_{n\to\infty}\Pr[Y_n \le t] &= \lim_{n\to\infty}\Pr[X_n \le t\cdot n]\\ &= \lim_{n\to\infty}\left[1-\left(1-\frac{\lambda}{n}\right)^{tn}\right]\\ &= 1-e^{-\lambda t}\,.\end{aligned}$$

Die Folge Y_n der (skalierten) geometrisch verteilten Zufallsvariablen geht also für $n\to\infty$ in eine exponentialverteilte Zufallsvariable mit Parameter λ über.

3. Mehrere kontinuierliche Zufallsvariablen

3.1 Mehrdimensionale Dichten

Beobachtung

Zu zwei kontinuierlichen Zufallsvariablen X, Y wird der zugrunde liegende gemeinsame Wahrscheinlichkeitsraum über $\mathbb{R}^2$ durch eine integrierbare (gemeinsame) Dichtefunktion $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}_0^+$ mit

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) \,\mathrm{d}\, x \,\mathrm{d}\, y = 1$$

beschrieben. Für ein Ereignis $A \subseteq \mathbb{R}^2$ (das aus abzählbar vielen geschlossenen oder offenen Bereichen gebildet sein muss) gilt

$$\Pr[A] = \int_A f_{X,Y}(x,y) \,\mathrm{d}\, x \,\mathrm{d}\, y.$$

Unter einem Bereich B verstehen wir dabei Mengen der Art

$$B = \{(x, y) \in \mathbb{R}^2 \mid a \le x \le b, c \le y \le d\} \quad \text{mit } a, b, c, d \in \mathbb{R}.$$

Dabei können die einzelnen Intervallgrenzen auch „offen“ bzw. $\pm\infty$ sein.

Analog zum eindimensionalen Fall ordnen wir der Dichte $f_{X,Y}$ eine (gemeinsame) Verteilung $F_{X,Y} : \mathbb{R}^2 \to [0,1]$ zu:

$$F_{X,Y}(x,y) = \Pr[X \le x, Y \le y] = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u,v) \, \mathrm{d}\, u \, \mathrm{d}\, v.$$

3.2 Randverteilungen und Unabhängigkeit

Definition 100

Sei $f_{X,Y}$ die gemeinsame Dichte der Zufallsvariablen X und Y. Die Randverteilung der Variablen X ist gegeben durch

$$F_X(x) = \Pr[X \leq x] = \int_{-\infty}^{x} \left[\int_{-\infty}^{\infty} f_{X,Y}(u,v) \,\mathrm{d}\, v \right] \mathrm{d}\, u.$$

Analog nennen wir

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,v) \,\mathrm{d}\, v$$

die Randdichte von X. Entsprechende Definitionen gelten symmetrisch für Y.

Definition 101

Zwei kontinuierliche Zufallsvariablen X und Y heißen unabhängig, wenn

$$\Pr[X \leq x, Y \leq y] = \Pr[X \leq x] \cdot \Pr[Y \leq y]$$

für alle $x, y \in \mathbb{R}$ gilt.

Dies ist gleichbedeutend mit

$$F_{X,Y}(x, y) = F_X(x) \cdot F_Y(y)\,.$$

Differentiation ergibt

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)\,.$$

Für mehrere Zufallsvariablen $X_1, \ldots, X_n$ gilt analog: $X_1, \ldots, X_n$ sind genau dann unabhängig, wenn

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdot \ldots \cdot F_{X_n}(x_n)$$

bzw.

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdot \ldots \cdot f_{X_n}(x_n)$$

für alle $x_1, \ldots, x_n \in \mathbb{R}$.

Warten auf mehrere Ereignisse

Satz 102

Die Zufallsvariablen $X_1, \ldots, X_n$ *seien unabhängig und exponentialverteilt mit den Parametern* $\lambda_1, \ldots, \lambda_n$. *Dann ist auch* $X := \min\{X_1, \ldots, X_n\}$ *exponentialverteilt mit dem Parameter* $\lambda_1 + \ldots + \lambda_n$.

Beweis:

Der allgemeine Fall folgt mittels Induktion aus dem für $n = 2$. Für die Verteilungsfunktion F_X gilt:

$$\begin{aligned}
1 - F_X(t) &= \Pr[X > t] = \Pr[\min\{X_1, X_2\} > t] \\
&= \Pr[X_1 > t, X_2 > t] \\
&= \Pr[X_1 > t] \cdot \Pr[X_2 > t] \\
&= e^{-\lambda_1 t} \cdot e^{-\lambda_2 t} = e^{-(\lambda_1 + \lambda_2) t}.
\end{aligned}$$

□

Anschaulich besagt Satz 102, dass sich die Raten addieren, wenn man auf das erste Eintreten eines Ereignisses aus mehreren unabhängigen Ereignissen wartet. Wenn beispielsweise ein Atom die Zerfallsrate λ besitzt, so erhalten wir bei n Atomen die Zerfallsrate $n\lambda$ (wie uns auch die Intuition sagt).

Poisson-Prozess

Wir hatten bei der Diskussion der geometrischen und der Poisson-Verteilung festgestellt:

Wenn der zeitliche Abstand der Treffer geometrisch verteilt ist, so ist ihre Anzahl in einer festen Zeitspanne binomialverteilt.

Im Grenzwert $n \to \infty$, wobei wir die Trefferwahrscheinlichkeit mit $p_n = \lambda/n$ ansetzen, konvergiert die geometrische Verteilung gegen die Exponentialverteilung und die Binomialverteilung gegen die Poisson-Verteilung. Im Grenzwert $n \to \infty$ erwarten wir deshalb die folgende Aussage:

Wenn man Ereignisse zählt, deren zeitlicher Abstand exponentialverteilt ist, so ist die Anzahl dieser Ereignisse in einer festen Zeitspanne Poisson-verteilt.

Seien $T_1, T_2 \ldots$ unabhängige exponentialverteilte Zufallsvariablen mit Parameter λ. Die Zufallsvariable T_i modelliert die Zeit, die zwischen Treffer $i-1$ und i vergeht.

Für den Zeitpunkt $t > 0$ definieren wir

$$X(t) := \max\{n \in \mathbb{N} \mid T_1 + \ldots + T_n \leq t\}.$$

$X(t)$ gibt also an, wie viele Treffer sich bis zur Zeit t (von Zeit Null ab) ereignet haben. Es gilt:

Fakt 103

Seien $T_1, T_2, \ldots$ *unabhängige Zufallsvariablen und sei* $X(t)$ *für* $t > 0$ *wie oben definiert. Dann gilt:* $X(t)$ *ist genau dann Poisson-verteilt mit Parameter* $t\lambda$*, wenn es sich bei* $T_1, T_2, \ldots$ *um exponentialverteilte Zufallsvariablen mit Parameter* λ *handelt.*

Zum Zufallsexperiment, das durch $T_1, T_2, \ldots$ definiert ist, erhalten wir für jeden Wert $t > 0$ eine Zufallsvariable $X(t)$. Hierbei können wir t als Zeit interpretieren und $X(t)$ als Verhalten des Experiments zur Zeit t. Eine solche Familie $(X(t))_{t>0}$ von Zufallsvariablen nennt man allgemein einen stochastischen Prozess. Der hier betrachtete Prozess, bei dem $T_1, T_2, \ldots$ unabhängige, exponentialverteilte Zufallsvariablen sind, heißt Poisson-Prozess und stellt ein fundamentales und zugleich praktisch sehr bedeutsames Beispiel für einen stochastischen Prozess dar.

Beispiel 104

Wir betrachten eine Menge von Jobs, die auf einem Prozessor sequentiell abgearbeitet werden. Die Laufzeiten der Jobs seien unabhängig und exponentialverteilt mit Parameter $\lambda = 1/30[1/s]$. Jeder Job benötigt also im Mittel $30s$.

Gemäß Fakt 103 ist die Anzahl von Jobs, die in einer Minute vollständig ausgeführt werden, Poisson-verteilt mit Parameter $t\lambda = 60 \cdot (1/30) = 2$.

Die Wahrscheinlichkeit, dass in einer Minute höchstens ein Job abgearbeitet wird, beträgt in diesem Fall ($t\lambda = 2$)

$$e^{-t\lambda} + t\lambda e^{-t\lambda} \approx 0{,}406\ .$$

3.4 Summen von Zufallsvariablen

Satz 105

Seien X und Y unabhängige kontinuierliche Zufallsvariablen. Für die Dichte von $Z := X + Y$ gilt

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) \cdot f_Y(z - x) \, \mathrm{d}\, x \, .$$

Beweis:

Nach Definition der Verteilungsfunktion gilt

$$F_Z(t) = \Pr[Z \leq t] = \Pr[X + Y \leq t] = \int_{A(t)} f_{X,Y}(x, y) \, \mathrm{d}\, x \mathrm{d}\, y$$

wobei $A(t) = \{(x, y) \in \mathbb{R}^2 \mid x + y \leq t\}$.

Beweis (Forts.):

Aus der Unabhängigkeit von X und Y folgt

$$F_Z(t) = \int_{A(t)} f_X(x) \cdot f_Y(y) \, \mathrm{d}\, x \mathrm{d}\, y$$
$$= \int_{-\infty}^{\infty} f_X(x) \cdot \left(\int_{-\infty}^{t-x} f_Y(y) \, \mathrm{d}\, y \right) \mathrm{d}\, x.$$

Mittels der Substitution $z := x + y$, d z = d y ergibt sich

$$\int_{-\infty}^{t-x} f_Y(y) \, \mathrm{d}\, y = \int_{-\infty}^{t} f_Y(z - x) \, \mathrm{d}\, z$$

und somit

$$F_Z(t) = \int_{-\infty}^{t} \left(\int_{-\infty}^{\infty} f_X(x) f_Y(z - x) \, \mathrm{d}\, x \right) \mathrm{d}\, z \,.$$

□

Satz 106 (Additivität der Normalverteilung)

Die Zufallsvariablen $X_1, \ldots, X_n$ seien unabhängig und normalverteilt mit den Parametern μ_i, σ_i $(1 \le i \le n)$. Es gilt: Die Zufallsvariable

$$Z := a_1 X_1 + \ldots + a_n X_n$$

ist normalverteilt mit Erwartungswert $\mu = a_1\mu_1 + \ldots + a_n\mu_n$ und Varianz $\sigma^2 = a_1^2\sigma_1^2 + \ldots + a_n^2\sigma_n^2$.

Beweis:

Wir beweisen zunächst den Fall $n = 2$ und $a_1 = a_2 = 1$. Nach Satz 105 gilt für $Z := X_1 + X_2$, dass

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_{X_1}(z-y) \cdot f_{X_2}(y)\, \mathrm{d}\, y \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2} \underbrace{\left(\frac{(z-y-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)}_{=:v}\right) \mathrm{d}\, y. \end{aligned}$$

Beweis (Forts.):

Wir setzen

$$\begin{aligned}\mu &:= \mu_1 + \mu_2\\ \sigma^2 &:= \sigma_1^2 + \sigma_2^2\\ v_1 &:= (z-\mu)/\sigma\\ v_2^2 &:= v - v_1^2\end{aligned}$$

Damit ergibt sich unmittelbar

$$v_2^2 = \frac{(z-y-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - \frac{(z-\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2},$$

woraus wir

$$v_2 = \frac{y\sigma_1^2 - \mu_2\sigma_1^2 + y\sigma_2^2 - z\sigma_2^2 + \mu_1\sigma_2^2}{\sigma_1\sigma_2\sigma}$$

ermitteln.

Beweis (Forts.):

Damit folgt für die gesuchte Dichte

$$f_Z(z) = \frac{1}{2\pi \cdot \sigma_1 \cdot \sigma_2} \cdot \exp\left(-\frac{v_1^2}{2}\right) \cdot \int_{-\infty}^{\infty} \exp\left(-\frac{v_2^2}{2}\right) \mathrm{d}\, y.$$

Wir substituieren noch

$$t := v_2 \text{ und } \mathrm{d}\, t = \frac{\sigma}{\sigma_1 \sigma_2} \mathrm{d}\, y$$

und erhalten

$$f_Z(z) = \frac{1}{2\pi \cdot \sigma} \cdot \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right) \cdot \int_{-\infty}^{\infty} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}\, t.$$

Mit Lemma 92 folgt, dass $f_Z(z) = \varphi(z; \mu, \sigma)$ ist.

Beweis (Forts.):

Daraus erhalten wir die Behauptung für $n = 2$, denn den Fall $Z := a_1X_1 + a_2X_2$ für beliebige Werte $a_1, a_2 \in \mathbb{R}$ können wir leicht mit Hilfe von Satz 93 auf den soeben bewiesenen Fall reduzieren. Durch Induktion kann die Aussage auf beliebige Werte $n \in \mathbb{N}$ verallgemeinert werden. $\square$

3.5 Momenterzeugende Funktionen für kontinuierliche Zufallsvariablen

Für diskrete Zufallsvariablen X haben wir die momenterzeugende Funktion

$$M_X(s) = \mathbb{E}[e^{Xs}]$$

eingeführt. Diese Definition kann man unmittelbar auf kontinuierliche Zufallsvariablen übertragen. Die für $M_X(s)$ gezeigten Eigenschaften bleiben dabei erhalten.

Beispiel 107

Für eine auf $[a, b]$ gleichverteilte Zufallsvariable U gilt

$$M_U(t) = \mathbb{E}[e^{tX}] = \int_a^b e^{tx} \cdot \frac{1}{b-a} \,\mathrm{d}\, x$$

$$= \left[\frac{e^{tx}}{t(b-a)}\right]_a^b$$

$$= \frac{e^{tb} - e^{ta}}{t(b-a)}\,.$$

Beispiel (Forts.)

Für eine standardnormalverteilte Zufallsvariable $N \sim \mathcal{N}(0,1)$ *gilt*

$$\begin{aligned} M_N(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{t\xi} e^{-\xi^2/2} \,\mathrm{d}\,\xi \\ &= e^{t^2/2} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-(t-\xi)^2/2} \,\mathrm{d}\,\xi \\ &= e^{t^2/2} \,. \end{aligned}$$

Beispiel (Forts.)

Daraus ergibt sich für $Y \sim \mathcal{N}(\mu, \sigma^2)$ *wegen* $\frac{Y-\mu}{\sigma} \sim \mathcal{N}(0,1)$

$$\begin{aligned} M_Y(t) &= \mathbb{E}[e^{tY}] \\ &= e^{t\mu} \cdot \mathbb{E}[e^{(t\sigma)\cdot \frac{Y-\mu}{\sigma}}] \\ &= e^{t\mu} \cdot M_N(t\sigma) \\ &= e^{t\mu+(t\sigma)^2/2}\,. \end{aligned}$$

Weiterer Beweis von Satz 106:

Beweis:
Gemäß dem vorhergehenden Beispiel gilt

$$M_{X_i}(t) = e^{t\mu_i + (t\sigma_i)^2/2}\,.$$

Wegen der Unabhängigkeit der X_i folgt

$$\begin{aligned} M_Z(t) = \mathbb{E}[e^{t(a_1X_1+\cdots+a_nX_n)}] &= \prod_{i=1}^n \mathbb{E}[e^{(a_it)X_i}] \\ &= \prod_{i=1}^n M_{X_i}(a_it) \\ &= \prod_{i=1}^n e^{a_it\mu_i + (a_it\sigma_i)^2/2} \\ &= e^{t\mu + (t\sigma)^2/2}\,, \end{aligned}$$

mit $\mu = a_1\mu_1 + \cdots + a_n\mu_n$ und $\sigma^2 = a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2$. $\square$

4. Zentraler Grenzwertsatz

Satz 108 (Zentraler Grenzwertsatz)

Die Zufallsvariablen $X_1, \ldots, X_n$ besitzen jeweils dieselbe Verteilung und seien unabhängig. Erwartungswert und Varianz von X_i existieren für $i = 1, \ldots, n$ und seien mit μ bzw. σ^2 bezeichnet ($\sigma^2 > 0$).
Die Zufallsvariablen Y_n seien definiert durch $Y_n := X_1 + \ldots + X_n$ für $n \geq 1$. Dann folgt, dass die Zufallsvariablen

$$Z_n := \frac{Y_n - n\mu}{\sigma\sqrt{n}}$$

asymptotisch standardnormalverteilt sind, also $Z_n \sim \mathcal{N}(0, 1)$ für $n \to \infty$.

Etwas formaler ausgedrückt gilt: Die Folge der zu Z_n gehörenden Verteilungsfunktionen F_n hat die Eigenschaft

$$\lim_{n\to\infty} F_n(x) = \Phi(x) \text{ für alle } x \in \mathbb{R}.$$

Wir sagen dazu auch: Die Verteilung von Z_n konvergiert gegen die Standardnormalverteilung für $n \to \infty$.

Dieser Satz ist von großer Bedeutung für die Anwendung der Normalverteilung in der Statistik. Der Satz besagt, dass sich die Verteilung einer Summe beliebiger unabhängiger Zufallsvariablen (mit endlichem Erwartungswert und Varianz) der Normalverteilung umso mehr annähert, je mehr Zufallsvariablen an der Summe beteiligt sind.

Beweis:

Wir betrachten $X_i^* := (X_i - \mu)/\sigma$ für $i = 1, \ldots, n$ mit $\mathbb{E}[X_i^*] = 0$ und $\mathrm{Var}[X_i^*] = 1$. Damit gilt (gemäß vorhergehendem Beispiel)

$$\begin{aligned} M_Z(t) = \mathbb{E}[e^{tZ}] &= \mathbb{E}[e^{t(X_1^* + \ldots + X_n^*)/\sqrt{n}}] \\ &= M_{X_1^*}(t/\sqrt{n}) \cdot \ldots \cdot M_{X_n^*}(t/\sqrt{n})\,. \end{aligned}$$

Für beliebiges i betrachten wir die Taylorentwicklung von $M_{X_i^*}(t) =: h(t)$ an der Stelle $t = 0$

$$h(t) = h(0) + h'(0) \cdot t + \frac{h''(0)}{2} \cdot t^2 + \mathcal{O}(t^3).$$

Aus der Linearität des Erwartungswerts folgt

$$h'(t) = \mathbb{E}[e^{tX_i^*} \cdot X_i^*] \text{ und } h''(t) = \mathbb{E}[e^{tX_i^*} \cdot (X_i^*)^2].$$

Beweis (Forts.):

Damit gilt

$$h'(0) = \mathbb{E}[X_i^*] = 0 \text{ und } h''(0) = \mathbb{E}[(X_i^*)^2] = \mathrm{Var}[X] = 1.$$

Durch Einsetzen in die Taylorreihe folgt $h(t) = 1 + t^2/2 + \mathcal{O}(t^3)$, und wir können $M_Z(t)$ umschreiben zu

$$M_Z(t) = \left(1 + \frac{t^2}{2n} + \mathcal{O}\left(\frac{t^3}{n^{3/2}}\right)\right)^n \to e^{t^2/2} \text{ für } n \to \infty.$$

Aus der Konvergenz der momenterzeugenden Funktion folgt auch die Konvergenz der Verteilung. Damit ist Z asymptotisch normalverteilt.

Beweis (Forts.):

Die momenterzeugende Funktion existiert leider nicht bei allen Zufallsvariablen und unser Beweis ist deshalb unvollständig. Man umgeht dieses Problem, indem man statt der momenterzeugenden Funktion die so genannte charakteristische Funktion $\tilde{M}_X(t) = \mathbb{E}[e^{\mathrm{i}tX}]$ betrachtet. Für Details verweisen wir auf die einschlägige Literatur. $\square$

Der Zentrale Grenzwertsatz hat die folgende intuitive Konsequenz:

> *Wenn eine Zufallsgröße durch lineare Kombination vieler unabhängiger, identisch verteilter Zufallsgrößen entsteht, so erhält man näherungsweise eine Normalverteilung.*

Ein wichtiger Spezialfall das Zentralen Grenzwertsatzes besteht darin, dass die auftretenden Zufallsgrößen Bernoulli-verteilt sind.

Korollar 109 (Grenzwertsatz von de Moivre)

$X_1, \ldots, X_n$ seien unabhängige Bernoulli-verteilte Zufallsvariablen mit gleicher Erfolgswahrscheinlichkeit p. Dann gilt für die Zufallsvariable H_n mit

$$H_n := X_1 + \ldots + X_n$$

für $n \geq 1$, dass die Verteilung der Zufallsvariablen

$$H_n^* := \frac{H_n - np}{\sqrt{np(1-p)}}$$

für $n \to \infty$ gegen die Standardnormalverteilung konvergiert.

Beweis:

Die Behauptung folgt unmittelbar aus dem Zentralen Grenzwertsatz, da $\mu = \frac{1}{n}\mathbb{E}[H_n] = p$ und $\sigma^2 = \frac{1}{n}\mathrm{Var}[H_n] = p(1-p)$. $\square$

Bemerkung

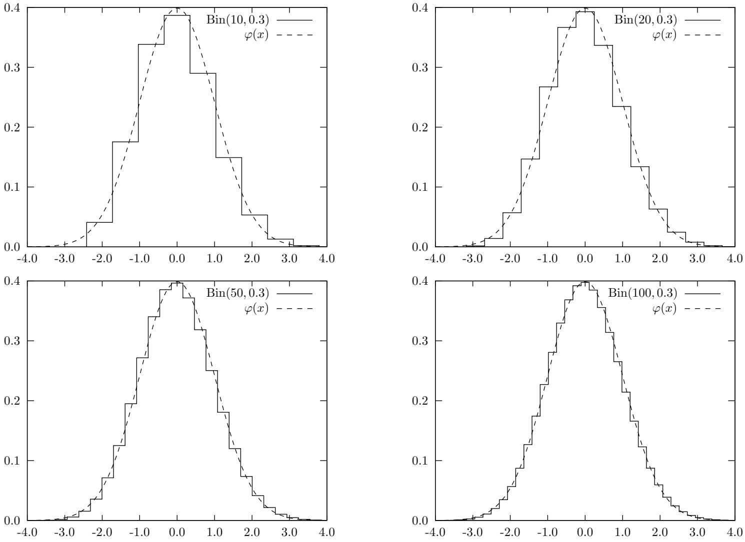
Wenn man $X_1, \ldots, X_n$ als Indikatorvariablen für das Eintreten eines Ereignisses A bei n unabhängigen Wiederholungen eines Experimentes interpretiert, dann gibt H_n die absolute Häufigkeit von A an.

4.1 Normalverteilung als Grenzwert der Binomialverteilung

Korollar 109 ermöglicht, die Normalverteilung als Grenzwert der Binomialverteilung aufzufassen. Die folgende Aussage ist eine Konsequenz von Korollar 109:

Korollar 110

Sei $H_n \sim \mathrm{Bin}(n, p)$ *eine binomialverteilte Zufallsvariable. Die Verteilung von* H_n/n *konvergiert gegen* $\mathcal{N}(p, p(1-p)/n)$ *für* $n \to \infty$.



Vergleich von Binomial- und Normalverteilung

$\mathrm{Bin}(n, 0.3)$ bei $0.3n$ zentriert, mit $\sqrt{0.3 \cdot 0.7n}$ horizontal gestaucht und vertikal gestreckt

Historisch gesehen entstand Korollar 109 vor Satz 108.

Für den Fall $p = 1/2$ wurde Korollar 109 bereits von Abraham de Moivre (1667–1754) bewiesen. De Moivre war gebürtiger Franzose, musste jedoch aufgrund seines protestantischen Glaubens nach England fliehen. Dort wurde er unter anderem Mitglied der Royal Society, erhielt jedoch niemals eine eigene Professur.

Die allgemeine Formulierung von Korollar 109 geht auf Pierre Simon Laplace (1749–1827) zurück. Allerdings vermutet man, dass die Lösung des allgemeinen Falls $p \neq 1/2$ bereits de Moivre bekannt war.

4.2 Elementarer Beweis des Grenzwertsatzes von de Moivre für $p = 1/2$

Wir betrachten die Wahrscheinlichkeit $\Pr[a \le H^*_{2n} \le b]$ für $p = 1/2$ und $a, b \in \mathbb{R}$ mit $a \le b$. Wenn die Verteilung von H^*_{2n}, wie in Korollar 109 angegeben, gegen $\mathcal{N}(0,1)$ konvergiert, so sollte $\Pr[a \le H^*_{2n} \le b] \approx \int_a^b \varphi(t)\,\mathrm{d}\,t$ für genügend große n gelten.

Wir schreiben $f(n) \sim_\infty g(n)$ für $\lim_{n\to\infty} f(n)/g(n) = 1$, wollen also zeigen:

$$\Pr[a \le H^*_{2n} \le b] \sim_\infty \int_a^b \varphi(t)\,\mathrm{d}\,t.$$

Da für $H_{2n} \sim \mathrm{Bin}(2n, 1/2)$ gilt, dass $\mathbb{E}[H_{2n}] = n$ und $\mathrm{Var}[H_{2n}] = n/2$ ist, erhalten wir

$$H^*_{2n} = \frac{H_{2n} - n}{\sqrt{n/2}},$$

und es folgt

$$\begin{aligned}\Pr[a \le H_{2n}^* \le b] &= \Pr[n + a\sqrt{n/2} \le H_{2n} \le n + b\sqrt{n/2}] \\ &= \sum_{i \in I_n} \Pr[H_{2n} = n + i]\end{aligned}$$

für $I_n := \{z \in \mathbb{Z} \mid a\sqrt{n/2} \le z \le b\sqrt{n/2}\}$. Damit ist

$$\Pr[a \le H_{2n}^* \le b] = \sum_{i \in I_n} \underbrace{\binom{2n}{n+i} \cdot \left(\frac{1}{2}\right)^{2n}}_{=:p_{n,i}}.$$

Es gilt

$$\max_i p_{n,i} \le p_n^* := \binom{2n}{n} \cdot \left(\frac{1}{2}\right)^{2n} = \frac{(2n)!}{(n!)^2} \cdot \left(\frac{1}{2}\right)^{2n},$$

und mit der Stirling'schen Approximation für $n!$

$$p_n^* \sim_\infty \frac{(2n)^{2n} \cdot e^{-2n} \cdot \sqrt{2\pi \cdot 2n}}{(n^n \cdot e^{-n} \cdot \sqrt{2\pi n})^2} \cdot \left(\frac{1}{2}\right)^{2n} = \frac{1}{\sqrt{\pi n}}.$$

Ersetzen wir nun die $p_{n,i}$ durch p_n^*, so entsteht dabei ein Fehler, den wir mit $q_{n,i} := \frac{p_{n,i}}{p_n^*}$ bezeichnen.

Für $i > 0$ gilt

$$\begin{aligned} q_{n,i} = \frac{\binom{2n}{n+i} \cdot \left(\frac{1}{2}\right)^{2n}}{\binom{2n}{n} \cdot \left(\frac{1}{2}\right)^{2n}} &= \frac{(2n)! \cdot n! \cdot n!}{(n+i)! \cdot (n-i)! \cdot (2n)!} \\ = \frac{\prod_{j=0}^{i-1}(n-j)}{\prod_{j=1}^{i}(n+j)} &= \prod_{j=1}^{i} \frac{n-j+1}{n+j} = \prod_{j=1}^{i} \left(1 - \frac{2j-1}{n+j}\right). \end{aligned}$$

Wegen der Symmetrie der Binomialkoeffizienten gilt $q_{n,-i} = q_{n,i}$, womit auch der Fall $i < 0$ abgehandelt ist.

Man macht sich leicht klar, dass $1 - 1/x \leq \ln x \leq x - 1$ für $x > 0$ gilt. Damit schließen wir, dass

$$
\begin{aligned}
\ln\left(\prod_{j=1}^{i}\left(1-\frac{2j-1}{n+j}\right)\right) &= \sum_{j=1}^{i}\ln\left(1-\frac{2j-1}{n+j}\right)\\
&\leq -\sum_{j=1}^{i}\frac{2j-1}{n+j} \leq -\sum_{j=1}^{i}\frac{2j-1}{n+i}\\
&= -\frac{i(i+1)-i}{n+i} = -\frac{i^2}{n}+\frac{i^3}{n(n+i)}\\
&= -\frac{i^2}{n}+\mathcal{O}\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}
$$

da $i = \mathcal{O}(\sqrt{n})$ für $i \in I_n$.

Ebenso erhalten wir

$$\begin{aligned}\ln\left(\prod_{j=1}^{i}\left(1-\frac{2j-1}{n+j}\right)\right) &\geq \sum_{j=1}^{i}\left(1-\left(1-\frac{2j-1}{n+j}\right)^{-1}\right)\\ &= \sum_{j=1}^{i}\frac{-2j+1}{n-j+1} \geq -\sum_{j=1}^{i}\frac{2j-1}{n-i}\\ &= -\frac{i^2}{n-i} = -\frac{i^2}{n}-\mathcal{O}\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Zusammen haben wir

$$e^{-\frac{i^2}{n-i}} = e^{-\frac{i^2}{n}-\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)} \leq q_{n,i} \leq e^{-\frac{i^2}{n}+\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)}$$

Wegen $e^{\pm\mathcal{O}(1/\sqrt{n})} = 1 \pm o(1)$ folgt daraus $q_{n,i} \sim_\infty e^{-i^2/n}$.

Damit schätzen wir nun $\Pr[a \le H_{2n}^* \le b]$ weiter ab:

$$\Pr[a \le H_{2n}^* \le b] = \sum_{i \in I_n} p_n^* \cdot q_{n,i} \ \sim_\infty \ \underbrace{\frac{1}{\sqrt{\pi n}} \cdot \sum_{i \in I_n} e^{-i^2/n}}_{=:S_n}.$$

Mit $\delta := \sqrt{2/n}$ können wir die Summe S_n umschreiben zu

$$S_n = \frac{1}{\sqrt{2\pi}} \cdot \sum_{i \in I_n} \delta e^{-(i\delta)^2 \cdot \frac{1}{2}}.$$

Diese Summe entspricht einer Näherung für $\int_a^b \varphi(t) \, \mathrm{d}\, t = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} \, \mathrm{d}\, t$ durch Aufteilung der integrierten Fläche in Balken der Breite δ. Für $n \to \infty$ konvergiert die Fläche der Balken gegen das Integral, d. h. $S_n \sim_\infty \int_a^b \varphi(t) \, \mathrm{d}\, t$.

q. e. d.

4.3 Verschiedene Approximationen der Binomialverteilung

Sei $H_n \sim \mathrm{Bin}(n, p)$ eine binomialverteilte Zufallsvariable mit der Verteilungsfunktion F_n. Für $n \to \infty$ gilt

$$\begin{aligned} F_n(t) &= \Pr[H_n/n \le t/n] \\ &\to \Phi\left(\frac{t/n - p}{\sqrt{p(1-p)/n}}\right) = \Phi\left(\frac{t - np}{\sqrt{p(1-p)n}}\right) . \end{aligned}$$

Wir können F_n somit für große n durch Φ approximieren. Diese Approximation ist in der Praxis deshalb von Bedeutung, da die Auswertung der Verteilungsfunktion der Binomialverteilung für große n sehr aufwändig ist, während für die Berechnung der Normalverteilung effiziente numerische Methoden vorliegen.

Beispiel 111

Wenn man die Wahrscheinlichkeit berechnen möchte, mit der bei 10^6 Würfen mit einem idealen Würfel mehr als 500500-mal eine gerade Augenzahl fällt, so muss man eigentlich folgenden Term auswerten:

$$T := \sum_{i=5{,}005\cdot 10^5}^{10^6} \binom{10^6}{i} \left(\frac{1}{2}\right)^{10^6}.$$

Dies ist numerisch kaum effizient möglich.

Die numerische Integration der Dichte φ der Normalverteilung ist hingegen relativ einfach. Auch andere Approximationen der Verteilung Φ, beispielsweise durch Polynome, sind bekannt. Entsprechende Funktionen werden in zahlreichen Softwarebibliotheken als „black box“ angeboten.

Beispiel

Mit der Approximation durch die Normalverteilung erhalten wir

$$
\begin{aligned}
T &\approx 1 - \Phi\left(\frac{5{,}005 \cdot 10^5 - 5 \cdot 10^5}{\sqrt{2{,}5 \cdot 10^5}}\right)\\
&= 1 - \Phi\left(\frac{5 \cdot 10^2}{5 \cdot 10^2}\right)\\
&= 1 - \Phi(1) \approx 0{,}1573\,.
\end{aligned}
$$

Bei der Approximation der Binomialverteilung mit Hilfe von Korollar 109 führt man oft noch eine so genannte Stetigkeitskorrektur durch. Zur Berechnung von $\Pr[X \le x]$ für $X \sim \mathrm{Bin}(n,p)$ setzt man

$$\Pr[X \le x] \approx \Phi\left(\frac{x + 0{,}5 - np}{\sqrt{np(1-p)}}\right)$$

statt

$$\Pr[X \le x] \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right)$$

an.

Der Korrekturterm läßt sich in der Histogramm-Darstellung der Binomialverteilung veranschaulichen. Die Binomialverteilung wird dort durch Balken angegeben, deren Fläche in etwa der Fläche unterhalb der Dichte φ von $\mathcal{N}(0,1)$ entspricht. Wenn man die Fläche der Balken mit „$X \leq x$“ durch das Integral von φ approximieren möchte, so sollte man bis zum Ende des Balkens für „$X = x$“ integrieren und nicht nur bis zur Mitte. Dafür sorgt der Korrekturterm $0{,}5$.

Approximationen für die Binomialverteilung

- Approximation durch die Poisson-Verteilung: $\mathrm{Bin}(n, p)$ wird approximiert durch $\mathrm{Po}(np)$. Diese Approximation funktioniert sehr gut für seltene Ereignisse, d. h. wenn np sehr klein gegenüber n ist. Als Faustregel fordert man $n \geq 30$ und $p \leq 0{,}05$.
- Approximation durch die Chernoff-Schranken: Bei der Berechnung der tails der Binomialverteilung liefern diese Ungleichungen meist sehr gute Ergebnisse. Ihre Stärke liegt darin, dass es sich bei den Schranken nicht um Approximationen, sondern um echte Abschätzungen handelt. Dies ist vor allem dann wichtig, wenn man nicht nur numerische Näherungen erhalten möchte, sondern allgemeine Aussagen über die Wahrscheinlichkeit von Ereignissen beweisen möchte.

- Approximation durch die Normalverteilung: Als Faustregel sagt man, dass die Verteilungsfunktion $F_n(t)$ von $\mathrm{Bin}(n, p)$ durch

$$F_n(t) \approx \Phi((t - np)/\sqrt{p(1-p)n})$$

approximiert werden kann, wenn $np \geq 5$ und $n(1-p) \geq 5$ gilt.

Kapitel III Induktive Statistik

1. Einführung

Das Ziel der induktiven Statistik besteht darin, aus gemessenen Zufallsgrößen auf die zugrunde liegenden Gesetzmäßigkeiten zu schließen. Im Gegensatz dazu spricht man von deskriptiver Statistik, wenn man sich damit beschäftigt, große Datenmengen verständlich aufzubereiten, beispielsweise durch Berechnung des Mittelwertes oder anderer abgeleiteter Größen.

2. Schätzvariablen

Wir betrachten die Anzahl X von Lesezugriffen auf eine Festplatte bis zum ersten Lesefehler und nehmen an, dass $\Pr[X = i] = (1-p)^{i-1}p$, setzen also für X eine geometrische Verteilung an. Dahinter verbirgt sich die Annahme, dass bei jedem Zugriff unabhängig und mit jeweils derselben Wahrscheinlichkeit p ein Lesefehler auftreten kann.

Unter diesen Annahmen ist die Verteilung der Zufallsvariablen X eindeutig festgelegt. Allerdings entzieht sich der numerische Wert des Parameters p noch unserer Kenntnis. Dieser soll daher nun empirisch geschätzt werden. Statt p können wir ebensogut $\mathbb{E}[X]$ bestimmen, da wir daraus nach den Eigenschaften der geometrischen Verteilung p mittels $p = \frac{1}{\mathbb{E}[X]}$ berechnen können.

Dazu betrachten wir n baugleiche Platten und die zugehörigen Zufallsvariablen X_i (für $1 \leq i \leq n$), d. h. wir zählen für jede Platte die Anzahl von Zugriffen bis zum ersten Lesefehler. Die Zufallsvariablen X_i sind dann unabhängig und besitzen jeweils dieselbe Verteilung wie X. Wir führen also viele Kopien eines bestimmten Zufallsexperiments aus, um Schlüsse auf die Gesetzmäßigkeiten des einzelnen Experiments ziehen zu können. Dies ist das Grundprinzip der induktiven Statistik. Die n Messungen heißen Stichproben, und die Variablen X_i nennt man Stichprobenvariablen.

Grundprinzip statistischer Verfahren

Wir erinnern an das Gesetz der großen Zahlen (Satz 63) bzw. den Zentralen Grenzwertsatz (Satz 108). Wenn man ein Experiment genügend oft wiederholt, so nähert sich der Durchschnitt der Versuchsergebnisse immer mehr dem Verhalten an, das man „im Mittel" erwarten würde. Je mehr Experimente wir also durchführen, umso genauere und zuverlässigere Aussagen können wir über den zugrunde liegenden Wahrscheinlichkeitsraum ableiten. Auf diesem Grundprinzip beruhen alle statistischen Verfahren.

Um $\mathbb{E}[X]$ empirisch zu ermitteln, bietet es sich an, aus den Zufallsvariablen X_i das arithmetische Mittel $\overline{X}$ zu bilden, das definiert ist durch

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Es gilt

$$\mathbb{E}[\overline{X}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X] = \mathbb{E}[X].$$

$\overline{X}$ liefert uns also im Mittel den gesuchten Wert $\mathbb{E}[X]$. Da wir $\overline{X}$ zur Bestimmung von $\mathbb{E}[X]$ verwenden, nennen wir $\overline{X}$ einen Schätzer für den Erwartungswert $\mathbb{E}[X]$. Wegen der obigen Eigenschaft ist $\overline{X}$ sogar ein so genannter erwartungstreuer Schätzer.

Definition 112

Gegeben sei eine Zufallsvariable X mit der Dichte $f(x;\theta)$. Eine Schätzvariable oder kurz Schätzer für den Parameter θ der Dichte von X ist eine Zufallsvariable, die aus mehreren (meist unabhängigen und identisch verteilten) Stichprobenvariablen zusammengesetzt ist. Ein Schätzer U heißt erwartungstreu, wenn gilt

$$\mathbb{E}[U] = \theta.$$

Bemerkung:

Die Größe $\mathbb{E}[U-\theta]$ nennt man Bias der Schätzvariablen U. Bei erwartungstreuen Schätzvariablen ist der Bias gleich Null.

Der Schätzer $\overline{X}$ ist also ein erwartungstreuer Schätzer für den Erwartungswert von X.
Ein wichtiges Maß für die Güte eines Schätzers ist die mittlere quadratische Abweichung, kurz MSE für mean squared error genannt. Diese berechnet sich durch $MSE := \mathbb{E}[(U - \theta)^2]$. Wenn U erwartungstreu ist, so folgt $MSE = \mathbb{E}[(U - \mathbb{E}[U])^2] = \mathrm{Var}[U]$.

Definition 113

Wenn die Schätzvariable A eine kleinere mittlere quadratische Abweichung besitzt als die Schätzvariable B, so sagt man, dass A effizienter ist als B.
Eine Schätzvariable heißt konsistent im quadratischen Mittel, wenn $\mathrm{MSE} \to 0$ für $n \to \infty$ gilt. Hierbei bezeichne n den Umfang der Stichprobe.

Für $\overline{X}$ erhalten wir wegen der Unabhängigkeit von $X_1, \ldots, X_n$

$$\begin{aligned} MSE = \mathrm{Var}[\overline{X}] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[X_i] = \frac{1}{n}\mathrm{Var}[X]. \end{aligned}$$

Bei jeder Verteilung mit endlicher Varianz folgt $MSE = \mathcal{O}(1/n)$ und somit $MSE \to 0$ für $n \to \infty$. Der Schätzer $\overline{X}$ ist also konsistent.

Aus der Konsistenz von $\overline{X}$ im quadratischen Mittel können wir mit Hilfe des Satzes von Chebyshev (siehe Satz 61) folgende Konsequenz ableiten. Sei $\varepsilon > 0$ beliebig, aber fest. Dann gilt

$$\Pr[|\overline{X} - \theta| \geq \varepsilon] = \Pr[|\overline{X} - \mathbb{E}[X]| \geq \varepsilon] \leq \frac{\mathrm{Var}[\overline{X}]}{\varepsilon^2} \to 0$$

für $n \to \infty$. Für genügend große n liegen also die Werte von $\overline{X}$ beliebig nahe am gesuchten Wert $\theta = \mathbb{E}[X]$. Diese Eigenschaft nennt man auch schwache Konsistenz, da sie aus der Konsistenz im quadratischen Mittel folgt.

Als nächstes betrachten wir eine weitere von $\overline{X}$ abgeleitete Schätzvariable:

$$S := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Wir zeigen, dass S^2 ein erwartungstreuer Schätzer für die Varianz von X ist. Sei $\mu := \mathbb{E}[X] = \mathbb{E}[X_i] = \mathbb{E}[\overline{X}]$.

$$\begin{aligned}
(X_i - \overline{X})^2 &= (X_i - \mu + \mu - \overline{X})^2 \\
&= (X_i - \mu)^2 + (\mu - \overline{X})^2 + 2(X_i - \mu)(\mu - \overline{X}) \\
&= (X_i - \mu)^2 + (\mu - \overline{X})^2 - \frac{2}{n}\sum_{j=1}^{n}(X_i - \mu)(X_j - \mu) \\
&= \frac{n-2}{n}(X_i - \mu)^2 + (\mu - \overline{X})^2 - \frac{2}{n}\sum_{j\neq i}(X_i - \mu)(X_j - \mu).
\end{aligned}$$

Für je zwei unabhängige Zufallsvariablen X_i, X_j mit $i \neq j$ gilt

$$\begin{aligned}
\mathbb{E}[(X_i - \mu)(X_j - \mu)] &= \mathbb{E}[X_i - \mu] \cdot \mathbb{E}[X_j - \mu] \\
&= (\mathbb{E}[X_i] - \mu) \cdot (\mathbb{E}[X_j] - \mu) \;=\; 0 \cdot 0 \;=\; 0.
\end{aligned}$$

Daraus folgt

$$\begin{aligned}
\mathbb{E}[(X_i - \overline{X})^2] &= \frac{n-2}{n} \cdot \mathbb{E}[(X_i - \mu)^2] + \mathbb{E}[(\mu - \overline{X})^2] \\
&= \frac{n-2}{n} \cdot \mathrm{Var}[X_i] + \mathrm{Var}[\overline{X}].
\end{aligned}$$

Wegen $\mathrm{Var}[X_i] = \mathrm{Var}[X]$ und $\mathrm{Var}[\overline{X}] = \frac{1}{n}\mathrm{Var}[X]$ folgt nun

$$\mathbb{E}[(X_i - \overline{X})^2] = \frac{n-1}{n} \cdot \mathrm{Var}[X],$$

und somit gilt für S^2

$$\begin{aligned}\mathbb{E}[S^2] &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}[(X_i - \overline{X})^2]\\ &= \frac{1}{n-1} \cdot n \cdot \frac{n-1}{n} \cdot \mathrm{Var}[X] = \mathrm{Var}[X].\end{aligned}$$

S^2 ist also eine erwartungstreue Schätzvariable für die Varianz von X.

Die vorangegangene Rechnung erklärt, warum man als Schätzer nicht

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \overset{!}{\neq} S^2$$

verwendet, wie man vielleicht intuitiv erwarten würde.

Definition 114

Die Zufallsvariablen

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \text{ und } S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

heißen Stichprobenmittel bzw. Stichprobenvarianz der Stichprobe $X_1, \ldots, X_n$. $\overline{X}$ und S^2 sind erwartungstreue Schätzer für den Erwartungswert bzw. die Varianz.

2.1 Maximum-Likelihood-Prinzip zur Konstruktion von Schätzvariablen

Wir betrachten nun ein Verfahren zur Konstruktion von Schätzvariablen für Parameter von Verteilungen. Sei

$$\vec{X} = (X_1, \ldots, X_n).$$

Bei $X_1, \ldots, X_n$ handelt es sich um unabhängige Kopien der Zufallsvariablen X mit der Dichte $f(x;\theta)$. Hierbei sei θ der gesuchte Parameter der Verteilung.
Sei X eine diskrete Zufallsvariable. Dann gilt

$$f(x;\theta) = \Pr[X = x],$$

wobei θ ein Parameter der Verteilung ist.

Wenn wir den Parameter explizit angeben wollen, so schreiben wir dafür auch $f(x;\theta) = \Pr_\theta[X = x]$. Eine Stichprobe liefert für jede Variable X_i einen Wert x_i. Diese Werte fassen wir ebenfalls zu einem Vektor $\vec{x} = (x_1, \ldots, x_n)$ zusammen.

Der Ausdruck

$$L(\vec{x};\theta) := \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \Pr_\theta[X_i = x_i]$$
$$\stackrel{\text{unabh.}}{=} \Pr_\theta[X_1 = x_1, \ldots, X_n = x_n]$$

entspricht der Wahrscheinlichkeit, dass wir die Stichprobe $\vec{x}$ erhalten, wenn wir den Parameter mit dem Wert θ belegen.
Für eine kontinuierliche Zufallsvariable X gilt analog

$$L(\vec{x};\theta) := \prod_{i=1}^{n} f(x_i;\theta).$$

Wir betrachten nun eine feste Stichprobe $\vec{x}$ und fassen $L(\vec{x};\theta)$ somit als Funktion von θ auf. In diesem Fall nennen wir L die Likelihood-Funktion der Stichprobe.

Es erscheint sinnvoll, zu einer gegebenen Stichprobe $\vec{x}$ den Parameter θ so zu wählen, dass $L(x;\theta)$ maximal wird.

Definition 115

Ein Schätzwert $\widehat{\theta}$ für den Parameter einer Verteilung $f(x;\theta)$ heißt Maximum-Likelihood-Schätzwert (ML-Schätzwert) für eine Stichprobe $\vec{x}$, wenn gilt

$$L(\vec{x};\theta) \leq L(\vec{x};\widehat{\theta}) \text{ für alle } \theta.$$

Beispiel 116

Wir konstruieren mit der ML-Methode einen Schätzer für den Parameter p der Bernoulli-Verteilung. Es gilt $\Pr_p[X_i = 1] = p$ und $\Pr_p[X_i = 0] = 1 - p$. Daraus schließen wir, dass $\Pr_p[X_i = x_i] = p^{x_i}(1-p)^{1-x_i}$, und stellen die Likelihood-Funktion

$$L(\vec{x}; p) = \prod_{i=1}^{n} p^{x_i} \cdot (1-p)^{1-x_i}$$

auf.

Wir suchen als Schätzer für p den Wert, an dem die Funktion L maximal wird. Wir erhalten

$$\begin{aligned}\ln L(\vec{x}; p) &= \sum_{i=1}^{n} (x_i \cdot \ln p + (1 - x_i) \cdot \ln(1-p)) \\ &= n\bar{x} \cdot \ln p + (n - n\bar{x}) \cdot \ln(1-p).\end{aligned}$$

Hierbei bezeichnet $\bar{x}$ das arithmetische Mittel $\frac{1}{n}\sum_{i=1}^{n} x_i$.

Beispiel (Forts.)

Wir finden das Maximum durch Nullsetzen der Ableitung:

$$\frac{d \ln L(\vec{x}; p)}{d p} = \frac{n\bar{x}}{p} - \frac{n - n\bar{x}}{1-p} = 0.$$

Diese Gleichung hat die Lösung $p = \bar{x}$.

Beispiel 117

Die Zufallsvariable X sei $\mathcal{N}(\mu, \sigma^2)$-verteilt, und wir suchen Schätzvariablen für die Parameter μ und σ. Nach Definition der Likelihood-Funktion gilt

$$L(\vec{x}; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \cdot \prod_{i=1}^{n} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right).$$

Durch Logarithmieren erhalten wir

$$\ln L(\vec{x}; \mu, \sigma^2) = -n(\ln\sqrt{2\pi} + \ln\sigma) + \sum_{i=1}^{n}\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right).$$

Beispiel 117

Für die Nullstellen der Ableitungen ergibt sich

$$\frac{\partial \ln L}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} \stackrel{!}{=} 0,$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^3} \stackrel{!}{=} 0,$$

also

$$\mu = \bar{x} \quad \text{und} \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Wir haben also durch die ML-Methode „fast“ das Stichprobenmittel und die Stichprobenvarianz erhalten. Allerdings besitzt der Schätzer für die Varianz hier den Vorfaktor $\frac{1}{n}$ statt $\frac{1}{n-1}$. Die ML-Schätzvariable für die Varianz ist somit nicht erwartungstreu.

3. Konfidenzintervalle

Bei der Verwendung von Schätzvariablen geht man davon aus, dass der erhaltene Schätzwert „nahe“ beim gesuchten Parameter θ liegt. Die Schätzungen werden „besser“, je größer die betrachtete Stichprobe ist. Diese Angaben sind aus quantitativer Sicht natürlich unbefriedigend, da nicht erkennbar ist, wie gut man sich auf den Schätzwert verlassen kann.

Die Lösung dieses Problems besteht darin, statt einer Schätzvariablen U zwei Schätzer U_1 und U_2 zu betrachten. U_1 und U_2 werden so gewählt, dass

$$\Pr[U_1 \leq \theta \leq U_2] \geq 1 - \alpha.$$

Die Wahrscheinlichkeit $1 - \alpha$ heißt Konfidenzniveau und kann dem „Sicherheitsbedürfnis“ angepasst werden.

Wenn wir für eine konkrete Stichprobe die Schätzer U_1 und U_2 berechnen und davon ausgehen, dass $\theta \in [U_1, U_2]$ ist, so ziehen wir höchstens mit Wahrscheinlichkeit α einen falschen Schluss. $[U_1, U_2]$ heißt Konfidenzintervall.

In vielen Fällen verwendet man nur eine Schätzvariable U und konstruiert mittels $U_1 := U - \delta$ und $U_2 := U + \delta$ ein symmetrisches Konfidenzintervall $[U - \delta, U + \delta]$.

Sei X eine $\mathcal{N}(\mu, \sigma^2)$-verteilte Zufallsvariable, und seien $X_1, \ldots, X_n$ n zugehörige Stichprobenvariablen. Gemäß der Additivität der Normalverteilung (siehe Satz 106) ist das Stichprobenmittel $\overline{X}$ ebenfalls normalverteilt mit $\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$. Wir suchen für $\overline{X}$ ein symmetrisches Konfidenzintervall.

Nach Satz 93 ist

$$Z := \sqrt{n} \cdot \frac{\overline{X} - \mu}{\sigma}$$

standardnormalverteilt.

Für Z betrachten wir das Konfidenzintervall $[-c, c]$ für ein geeignetes $c > 0$ und setzen

$$\Pr[-c \le Z \le c] \stackrel{!}{=} 1 - \alpha.$$

Auflösen nach μ ergibt

$$\Pr\left[\overline{X} - \frac{c\sigma}{\sqrt{n}} \le \mu \le \overline{X} + \frac{c\sigma}{\sqrt{n}}\right] \stackrel{!}{=} 1 - \alpha\,.$$

Das gesuchte Konfidenzintervall lautet also

$$K = [\overline{X} - \frac{c\sigma}{\sqrt{n}}, \overline{X} + \frac{c\sigma}{\sqrt{n}}]\,.$$

Den Parameter c wählen wir wie folgt:

$$\Pr[-c \leq Z \leq c] = \Phi(c) - \Phi(-c) \stackrel{!}{=} 1 - \alpha.$$

Wegen der Symmetrie von Φ gilt $\Phi(-x) = 1 - \Phi(x)$ und wir erhalten

$$\Phi(c) - \Phi(-c) = 2 \cdot \Phi(c) - 1 \stackrel{!}{=} 1 - \alpha \iff \Phi(c) = 1 - \frac{\alpha}{2},$$

also

$$c = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right).$$

Definition 118

X sei eine stetige Zufallsvariable mit Verteilung F_X. Eine Zahl x_γ mit

$$F_X(x_\gamma) = \gamma$$

heißt γ-Quantil von X bzw. der Verteilung F_X.

Definition 119

Für die Standardnormalverteilung bezeichnet z_γ das γ-Quantil.

Damit können wir das gesuchte Konfidenzintervall angeben durch

$$K = \left[\overline{X} - \frac{z_{(1-\frac{\alpha}{2})}\sigma}{\sqrt{n}}, \overline{X} + \frac{z_{(1-\frac{\alpha}{2})}\sigma}{\sqrt{n}}\right].$$

4. Testen von Hypothesen

4.1 Einführung

Bislang haben wir versucht, Parameter von Verteilungen zu schätzen. In der Praxis ist man jedoch oft an der eigentlichen Kenntnis dieser Parameter gar nicht interessiert, sondern man möchte gewisse, damit zusammenhängende Behauptungen überprüfen.

Im Folgenden stellen wir die Bestandteile eines statistischen Tests anhand eines abstrakten Beispiels vor. Wir betrachten dazu eine Zufallsvariable X mit $\Pr[X = 1] = p$ und $\Pr[X = 0] = 1 - p$. Durch einen Test soll überprüft werden, ob $p < 1/3$ oder $p \geq 1/3$ gilt.

Definition eines Tests

Wir betrachten eine Stichprobe von n unabhängigen Stichprobenvariablen $X_1, \ldots, X_n$, die dieselbe Verteilung wie die Zufallsvariable X besitzen. Zu einem zugehörigen Stichprobenvektor $\vec{x}$ müssen wir nun die Frage beantworten, ob wir für diesen Versuchsausgang die Hypothese „$p \geq 1/3$“ annehmen oder ablehnen.

Sei

$$K := \{\vec{x} \in \mathbb{R}^n;\ \vec{x} \text{ führt zur Ablehnung der Hypothese}\}.$$

K nennen wir den Ablehnungsbereich oder den kritischen Bereich des Tests.

Gewöhnlich wird K konstruiert, indem man die Zufallsvariablen $X_1, \ldots, X_n$ zu einer neuen Variablen T, der so genannten Testgröße, zusammenfasst. Dann unterteilt man den Wertebereich $\mathbb{R}$ von T in mehrere Bereiche, die entweder zur Ablehnung der Hypothese führen sollen oder nicht. Dabei betrachtet man meist ein einzelnes halboffenes oder abgeschlossenes Intervall und spricht dann von einem einseitigen bzw. von einem zweiseitigen Test.

Die Menge $\widetilde{K} \subseteq \mathbb{R}$ enthalte die Werte von T, die zur Ablehnung der Hypothese führen sollen. Da wir Tests immer über eine Testgröße definieren, werden wir der Einfachheit halber auch $\widetilde{K}$ als Ablehnungsbereich bezeichnen. $\widetilde{K} \subseteq \mathbb{R}$ entspricht direkt dem Ablehnungbereich $K = T^{-1}(\widetilde{K}) \subseteq \mathbb{R}^n$, wie wir ihn oben festgelegt haben.

Die zu überprüfende Hypothese bezeichnen wir mit H_0 und sprechen deshalb auch von der Nullhypothese. Bei manchen Tests formuliert man noch eine zweite Hypothese H_1, die so genannte Alternative. Im Beispiel können wir

$$H_0 : p \geq 1/3 \text{ und } H_1 : p < 1/3$$

setzen.

Manchmal verzichtet man darauf, H_1 anzugeben. Dann besteht die Alternative wie oben einfach darin, dass H_0 nicht gilt. In diesem Fall nennen wir H_1 triviale Alternative.

Ein echter, also nicht-trivialer Alternativtest läge beispielsweise vor, wenn wir ansetzen

$$H_0' : p \geq 1/3 \text{ und } H_1' : p \leq 1/6.$$

Beispiel 120

Wir untersuchen eine Festplatte, von der bekannt ist, dass sie zu einer von zwei Baureihen gehört. Die mittleren Zugriffszeiten dieser Baureihen betragen 9ms bzw. 12ms. Wir möchten nun herausfinden, zu welchem Typ die betrachtete Festplatte gehört, indem wir die Zugriffszeit bei n Zugriffen bestimmen. Hier würde man dann ansetzen: $H_0 : \mu \leq 9$ und $H_1 := \mu \geq 12$, wobei μ die mittlere Zugriffszeit bezeichnet.

Fehler bei statistischen Tests

Bei jedem statistischen Test können mit einer gewissen Wahrscheinlichkeit falsche Schlüsse gezogen werden. Dieser Fall tritt beispielsweise ein, wenn H_0 gilt, aber das Ergebnis $\vec{x}$ der Stichprobe im Ablehnungsbereich K liegt.
Dann spricht man von einem Fehler 1. Art.
Analog erhalten wir einen Fehler 2. Art, wenn H_1 gilt und $\vec{x}$ nicht im Ablehnungsbereich liegt.

Fehler 1. Art	:	H_0 gilt, wird aber abgelehnt.
Fehler 2. Art	:	H_0 wird nicht abgelehnt, obwohl H_1 gilt.

Für die Beurteilung eines Tests ist es wesentlich, mit welcher Wahrscheinlichkeit diese beiden Fehler eintreten können. Ziel ist es natürlich, diese Wahrscheinlichkeiten möglichst klein zu halten. Allerdings sind die Minimierung des Fehlers 1. Art und des Fehlers 2. Art gegenläufige Ziele, so dass ein vernünftiger Ausgleich zwischen beiden Fehlern gefunden werden muss. Wenn man beispielsweise $K = \emptyset$ setzt, so erhält man Wahrscheinlichkeit Null für den Fehler 1. Art, da H_0 immer angenommen wird. Allerdings tritt der Fehler 2. Art dann mit Wahrscheinlichkeit Eins ein, wenn H_1 gilt.

Die Wahrscheinlichkeit für den Fehler 1. Art wird mit α bezeichnet, und man spricht deshalb gelegentlich vom α-Fehler. α heißt auch Signifikanzniveau des Tests.

In der Praxis ist es üblich, sich ein Signifikanzniveau α vorzugeben (übliche Werte hierfür sind $0{,}05$, $0{,}01$ oder $0{,}001$) und dann den Test so auszulegen (also den Ablehnungsbereich K so zu bestimmen), dass die Wahrscheinlichkeit für den Fehler 1. Art den Wert α besitzt.

Konstruktion eines einfachen Tests

Wir konstruieren einen Test für den Parameter p einer Bernoulli-verteilten Zufallsvariablen X. Wir setzen

$$H_0 : p \geq p_0, \qquad H_1 : p < p_0.$$

Als Testgröße verwenden wir

$$T := X_1 + \ldots + X_n.$$

Für größere Wahrscheinlichkeiten p erwarten wir auch größere Werte für T. Deshalb ist es sinnvoll, einen Ablehnungsbereich der Art $K := [0, k]$ für T zu wählen, wobei $k \in \mathbb{R}$ geeignet festzulegen ist. Wir konstruieren hier also einen einseitigen Test, während für eine Nullhypothese $H_0 : p = p_0$ sowohl zu kleine als auch zu große Werte von T zur Ablehnung von H_0 führen sollten und somit ein zweiseitiger Test vorzuziehen wäre.

T ist binomialverteilt. Da wir von einem großen Stichprobenumfang n ausgehen, bietet es sich an, die Verteilung von T nach dem Grenzwertsatz von de Moivre (siehe Korollar 109) durch die Normalverteilung zu approximieren.

Sei

$$\tilde{T} := \frac{T - np}{\sqrt{np(1-p)}}.$$

$\tilde{T}$ ist annähernd standardnormalverteilt.

Wir berechnen für jeden Wert von k das zugehörige Signifikanzniveau α des Tests.

$$\begin{aligned}
\text{Fehlerwahrscheinlichkeit 1. Art} &= \max_{p \in H_0} \Pr_p[T \in K] \\
&= \max_{p \in H_0} \Pr_p[T \le k] \\
\text{Fehlerwahrscheinlichkeit 2. Art} &= \sup_{p \in H_1} \Pr_p[T \notin K] \\
&= \sup_{p \in H_1} \Pr_p[T > k]
\end{aligned}$$

Für den Fehler 1. Art α erhalten wir

$$\begin{aligned}
\alpha = \max_{p \geq p_0} \Pr_p[T \leq k] = \Pr_{p=p_0}[T \leq k] \\
= \Pr_{p=p_0}\left[\tilde{T} \leq \frac{k - np}{\sqrt{np(1-p)}}\right] \\
= \Pr\left[\tilde{T} \leq \frac{k - np_0}{\sqrt{np_0(1-p_0)}}\right] \approx \Phi\left(\frac{k - np_0}{\sqrt{np_0(1-p_0)}}\right).
\end{aligned}$$

Unter Verwendung der Quantile der Standardnormalverteilung ergibt sich damit:

- Ist k so gewählt, dass $(k - np_0)/\sqrt{np_0(1-p_0)} = z_\alpha$, so ist das Signifikanzniveau gleich α.
- Ist das gewünschte Signifikanzniveau α des Tests vorgegeben, so erhält man den Wert $k = k(n)$ in Abhängigkeit vom Umfang n der Stichprobe durch

$$k = z_\alpha \cdot \sqrt{np_0(1-p_0)} + np_0. \tag{8}$$

Kleinere Werte für k verkleinern zwar den Fehler 1. Art, vergrößern jedoch den Annahmebereich und damit die Wahrscheinlichkeit für einen Fehler 2. Art.

Verhalten der Testfehler

Wie verhalten sich die möglichen Testfehler des konstruierten Verfahrens? Was geschieht beispielsweise, wenn p nur geringfügig kleiner als p_0 ist?

In diesem Fall betrachten wir beim Fehler 2. Art die Wahrscheinlichkeit

$$\Pr_{p=p_0-\varepsilon}[T > k] \approx \Pr_{p=p_0}[T > k] \approx 1 - \alpha \;.$$

Wenn sich also die „wahren“ Verhältnisse nur minimal von unserer Nullhypothese unterscheiden, so werden wir diese „im Zweifelsfall“ annehmen.

Bei echten Alternativtests werden für hinreichend große Stichproben und einen geeignet eingestellten Ablehnungsbereich beide Testfehler klein.

Beispiel 121

Die Abbruchrate p der Transaktionen in einem Online-Datenbanksystem wurde bereits früher einmal ermittelt. Allerdings sind die entsprechenden Daten verloren gegangen und die Entwickler erinnern sich nur noch, dass das Ergebnis entweder $p = 1/3$ oder $p = 1/6$ lautete. Unter dieser Annahme würde man den Test wie folgt ansetzen:

$$H_0 : p \geq 1/3, \qquad H_1' : p \leq 1/6.$$

Beispiel (Forts.)

Für den Fehler 2. Art erhält man nun:

$$\textit{Fehlerwahrsch. 2. Art} = \max_{p \le 1/6} \Pr_p[T > k]$$
$$\approx 1 - \Phi\left(\frac{k - (1/6) \cdot n}{\sqrt{(1/6) \cdot (5/6) n}}\right).$$

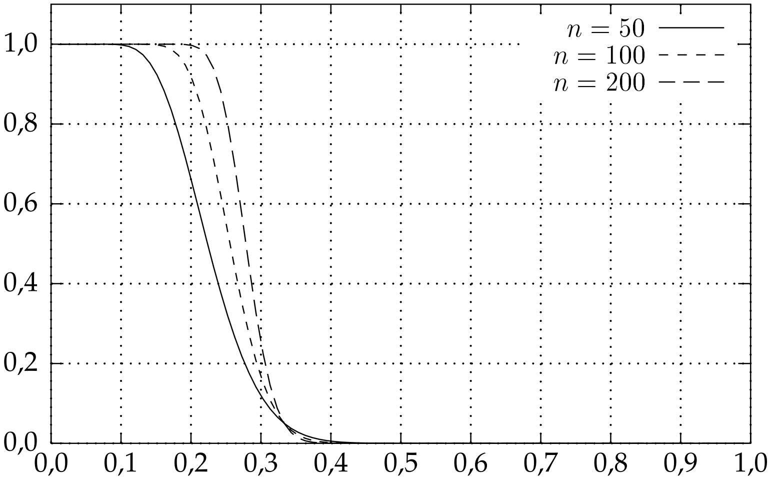
Mit den obigen Werten $k = 25$ *und* $n = 100$ *ergibt sich mit*

$$\Phi\left(\frac{150 - 100}{\sqrt{5} \cdot 10}\right) = \Phi(\sqrt{5}) \approx 0{,}9871$$

ein Fehler 2. Art der Größe $0{,}0129$, *während sich für die triviale Alternative* $H_1 : p < 1/3$ *ein Wert von etwa* $0{,}95$ *ergibt.*

Die so genannte Gütefunktion g gibt allgemein die Wahrscheinlichkeit an, mit der ein Test die Nullhypothese verwirft. Für unser hier entworfenes Testverfahren gilt

$$g(n,p) = \Pr_p[T \in K] = \Pr_p[T \leq k] \approx \Phi\left(\frac{k - np}{\sqrt{np(1-p)}}\right).$$



Gütefunktion $g(n, p)$ für verschiedene Werte von n

Man erkennt deutlich, dass für alle n der Wert von $k = k(n)$ genau so gewählt wurde, dass $g(n, 1/3) = 0{,}05$ gilt. Dies wird durch den in Gleichung 8 angegebenen Ausdruck erreicht.

Für Werte von p größer als $1/3$ wird $H_0 : p \geq 1/3$ mit hoher Wahrscheinlichkeit angenommen, während für Werte deutlich unter $1/3$ die Hypothese H_0 ziemlich sicher abgelehnt wird.

Ferner ist auffällig, dass g für größere Werte von n schneller von Eins auf Null fällt. Daran erkennt man, dass durch den Test die Fälle „H_0 gilt“ und „H_0 gilt nicht“ umso besser unterschieden werden können, je mehr Stichproben durchgeführt werden. Für Werte von p, bei denen $g(n, p)$ weder nahe bei Eins noch nahe bei Null liegt, kann der Test nicht sicher entscheiden, ob die Nullhypothese abzulehnen ist.

4.2 Praktische Anwendung statistischer Tests

Das im vorhergehenden Abschnitt konstruierte Testverfahren taucht in der Literatur unter dem Namen approximativer Binomialtest auf.

Die folgende Tabelle 1 gibt einen Überblick über die Eckdaten dieses Tests.

Tabelle: **Approximativer Binomialtest**

Annahmen:

$X_1, \ldots, X_n$ seien unabhängig und identisch verteilt mit $\Pr[X_i = 1] = p$ und $\Pr[X_i = 0] = 1-p$, wobei p unbekannt sei. n sei hinreichend groß, so dass die Approximation aus Korollar 109 brauchbare Ergebnisse liefert.

Hypothesen:

a) $H_0 : p = p_0$ *gegen* $H_1 : p \neq p_0$,
b) $H_0 : p \geq p_0$ *gegen* $H_1 : p < p_0$,
c) $H_0 : p \leq p_0$ *gegen* $H_1 : p > p_0$.

Testgröße:

$$Z := \frac{h - np_0}{\sqrt{np_0(1-p_0)}},$$

wobei $h := X_1 + \ldots + X_n$ die Häufigkeit bezeichnet, mit der die Ereignisse $X_i = 1$ aufgetreten sind.

Ablehnungskriterium für H_0 bei Signifikanzniveau α:

a) $|Z| > z_{1-\alpha/2}$,
b) $Z < z_\alpha$,
c) $Z > z_{1-\alpha}$.

4.3 Allgemeines Vorgehen bei statistischen Tests

1. Schritt: Formulierung von Annahmen. Ganz ohne Annahmen kommt man meist nicht aus. Übliche Annahmen betreffen meist die Verteilung der Stichprobenvariablen und deren Unabhängigkeit.

2. Schritt: Formulierung der Nullhypothese.

3. Schritt: Auswahl des Testverfahrens.

4. Schritt: Durchführung des Tests und Entscheidung.

4.4 Ausgewählte statistische Tests

4.4.1 Wie findet man das richtige Testverfahren?

Statistische Tests kann man nach mehreren Kriterien in Klassen einteilen.

- **Anzahl der beteiligten Zufallsgrößen**
  Sollen zwei Zufallsgrößen mit potentiell unterschiedlichen Verteilungen verglichen werden, für die jeweils eine Stichprobe erzeugt wird (Zwei-Stichproben-Test), oder wird nur eine einzelne Zufallsgröße untersucht (Ein-Stichproben-Test)?

Bei der Fragestellung

> *Beträgt die mittlere Zugriffszeit auf einen Datenbankserver im Mittel höchstens 10ms?*

hat man es mit einem Ein-Stichproben-Test zu tun, während die Untersuchung der Frage

> *Hat Datenbankserver A eine kürzere mittlere Zugriffszeit als Datenbankserver B?*

auf einen Zwei-Stichproben-Test führt.

Bei mehreren beteiligten Zufallsgrößen wird zusätzlich unterschieden, ob aus voneinander unabhängigen Grundmengen Stichproben erhoben werden oder nicht. Beim vorigen Beispiel werden unabhängige Messungen vorgenommen, sofern die Server A und B getrennt voneinander arbeiten. Wenn man jedoch die Frage

> *Läuft ein Datenbankserver auf einer Menge festgelegter Testanfragen mit Query-Optimierung schneller als ohne?*

untersucht, so spricht man von verbundenen Messungen.

Gelegentlich betrachtet man auch den Zusammenhang zwischen mehreren Zufallsgrößen. Beispielsweise könnte man sich für die Frage interessieren:

> *Wie stark wächst der Zeitbedarf für eine Datenbankanfrage im Mittel mit der (syntaktischen) Länge der Anfrage, d. h. führen kompliziertere Formulierungen zu proportional längeren Laufzeiten?*

Mit solchen Fragenstellungen, bei denen ein funktionaler Zusammenhang zwischen Zufallsgrößen ermittelt werden soll, beschäftigt sich die Regressionsanalyse. Wenn überhaupt erst zu klären ist, ob ein solcher Zusammenhang besteht oder ob die Zufallsgrößen vielmehr unabhängig voneinander sind, so spricht man von Zusammenhangsanalyse.

- **Formulierung der Nullhypothese**

  Welche Größe dient zur Definition der Nullhypothese? Hierbei werden in erster Linie Tests unterschieden, die Aussagen über verschiedene so genannte Lageparameter treffen, wie z.B. den Erwartungswert oder die Varianz der zugrunde liegenden Verteilungen.

  Im Zwei-Stichproben-Fall könnte man beispielsweise untersuchen, ob der Erwartungswert der Zufallsgröße A größer oder kleiner als bei Zufallsgröße B ist.

  Gelegentlich wird zur Formulierung der Nullhypothese auch der so genannte Median betrachtet: Der Median einer Verteilung entspricht dem (kleinsten) Wert x mit $F(x) = 1/2$.

  Neben solchen Tests auf Lageparameter gibt es z.B. auch Tests, die auf eine vorgegebene Verteilung oder auf ein Maß für die Abhängigkeit verschiedener Zufallsgrößen testen.

- **Annahmen über die Zufallsgrößen**
  Was ist über die Verteilung der untersuchten Größe(n) bekannt? Bei entsprechenden Annahmen könnte es sich z.B. um die Art der Verteilung, den Erwartungswert oder die Varianz handeln.

4.4.2 Ein-Stichproben-Tests für Lageparameter

Beim approximativen Binomialtest wird ausgenutzt, dass die Binomialverteilung für große n nach dem Grenzwertsatz von de Moivre (Korollar 109) gegen die Normalverteilung konvergiert. Aus diesem Grund kann man diesen Test auch als Spezialfall eines allgemeineren Testverfahrens ansehen, nämlich des Gaußtest, der nun dargestellt wird.

Tabelle: **Gaußtest**

Annahmen:

$X_1, \ldots, X_n$ *seien unabhängig und identisch verteilt mit* $X_i \sim \mathcal{N}(\mu, \sigma^2)$, *wobei* σ^2 *bekannt ist.* *Alternativ gelte* $\mathbb{E}[X_i] = \mu$ *und* $\mathrm{Var}[X_i] = \sigma^2$, *und* n *sei groß genug.*

Hypothesen:

a) $H_0 : \mu = \mu_0$ *gegen* $H_1 : \mu \neq \mu_0$,
b) $H_0 : \mu \geq \mu_0$ *gegen* $H_1 : \mu < \mu_0$,
c) $H_0 : \mu \leq \mu_0$ *gegen* $H_1 : \mu > \mu_0$.

Testgröße:

$$Z := \frac{\overline{X} - \mu_0}{\sigma} \sqrt{n} \,.$$

Ablehnungskriterium für H_0 bei Signifikanzniveau α:

a) $|Z| > z_{1-\alpha/2}$,
b) $Z < z_\alpha$,
c) $Z > z_{1-\alpha}$.

Der Gaußtest hat den Nachteil, dass man die Varianz σ^2 der beteiligten Zufallsgrößen kennen muss.

Wenn diese unbekannt ist, so liegt es nahe, die Varianz durch die Stichprobenvarianz S^2 (siehe Definition 114) anzunähern. Dies führt auf den so genannten t-Test, der in der folgenden Übersicht dargestellt ist.

Tabelle: *t*-Test

Annahmen:

$X_1, \ldots, X_n$ seien unabhängig und identisch verteilt mit $X_i \sim \mathcal{N}(\mu, \sigma^2)$.
Alternativ gelte $\mathbb{E}[X_i] = \mu$ und $\mathrm{Var}[X_i] = \sigma^2$, und n sei groß genug.

Hypothesen:

$$
\begin{array}{llll}
\text{a)} & H_0 : \mu = \mu_0 & \text{gegen} & H_1 : \mu \neq \mu_0, \\
\text{b)} & H_0 : \mu \geq \mu_0 & \text{gegen} & H_1 : \mu < \mu_0, \\
\text{c)} & H_0 : \mu \leq \mu_0 & \text{gegen} & H_1 : \mu > \mu_0.
\end{array}
$$

Testgröße:

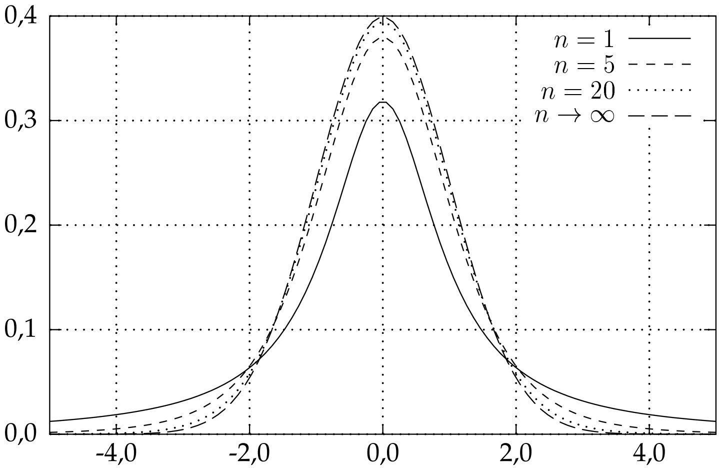
$$T := \frac{\overline{X} - \mu_0}{S} \sqrt{n}.$$

Ablehnungskriterium für H_0 bei Signifikanzniveau α:

$$
\begin{array}{ll}
\text{a)} & |T| > t_{n-1, 1-\alpha/2}, \\
\text{b)} & T < t_{n-1, \alpha}, \\
\text{c)} & T > t_{n-1, 1-\alpha}.
\end{array}
$$

Hierbei gibt $t_{n-1,1-\alpha}$ das $(1-\alpha)$-Quantil der t-Verteilung mit $n-1$ Freiheitsgraden an. Die t-Verteilung taucht manchmal auch unter dem Namen Student-Verteilung auf, da sie ursprünglich unter dem Pseudonym „Student“ publiziert wurde.

Wir gehen an dieser Stelle nicht darauf ein, wieso die Testgröße die t-Verteilung besitzt, sondern weisen nur darauf hin, dass die Dichte dieser Verteilung (eigentlich handelt es sich um eine ganze Familie von Verteilungen, da die Anzahl der Freiheitsgrade jeweils noch gewählt werden kann) der Dichte der Normalverteilung ähnelt. Für große n (Faustregel: $n \geq 30$) liegen die beiden Dichten so genau übereinander, dass man in der Praxis die t-Verteilung durch die Normalverteilung annähert.



Dichte der t-Verteilung mit n Freiheitsgraden

Als weitere Beispiele für gängige Ein-Stichproben-Tests zu Lageparametern seien der Wilcoxon-Test und der χ^2-Varianztest genannt. Ersterer dient zum Testen von Hypothesen zum Median, während der zweite Test Hypothesen zur Varianz beinhaltet.

4.4.3 Zwei-Stichproben-Tests für Lageparameter

Bei Zwei-Stichproben-Tests wollen wir das Verhältnis von Lageparametern untersuchen. Besonders wichtig sind hierbei Tests zum Erwartungswert. Für zwei Zufallsgrößen X und Y könnten wir beispielsweise die Frage untersuchen, ob für die Erwartungswerte μ_X und μ_Y gilt, dass $\mu_X = \mu_Y$ ist.

Tabelle: **Zwei-Stichproben-t-Test**

Annahmen:

$X_1, \ldots, X_m$ und $Y_1, \ldots, Y_n$ seien unabhängig und jeweils identisch verteilt, wobei $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$ und $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ gelte. Die Varianzen seien identisch, also $\sigma_X^2 = \sigma_Y^2$.

Hypothesen:

$$\begin{array}{lllll} a) & H_0 : \mu_X = \mu_Y & \textit{gegen} & H_1 : \mu_X \neq \mu_Y, \\ b) & H_0 : \mu_X \geq \mu_Y & \textit{gegen} & H_1 : \mu_X < \mu_Y, \\ c) & H_0 : \mu_X \leq \mu_Y & \textit{gegen} & H_1 : \mu_X > \mu_Y. \end{array}$$

Testgröße:

$$T := \sqrt{\frac{n+m-2}{\frac{1}{m}+\frac{1}{n}}} \cdot \frac{\overline{X} - \overline{Y}}{\sqrt{(m-1) \cdot S_X^2 + (n-1) \cdot S_Y^2}}.$$

Ablehnungskriterium für H_0 bei Signifikanzniveau α:

$$\begin{array}{ll} a) & |T| > t_{m+n-2, 1-\alpha/2}, \\ b) & T < t_{m+n-2, \alpha}, \\ c) & T > t_{m+n-2, 1-\alpha}. \end{array}$$

Vom Zwei-Stichproben-t-Test findet man in der Literatur noch zusätzliche Varianten, die auch dann einsetzbar sind, wenn die beteiligten Zufallsgrößen nicht dieselbe Varianz besitzen. Der beim Ein-Stichproben-Fall erwähnte Wilcoxon-Test kann ebenfalls auf den Zwei-Stichproben-Fall übertragen werden.

4.4.4 Nicht an Lageparametern orientierte Tests

Wir betrachten in diesem Abschnitt exemplarisch den χ^2-Anpassungstest. Bei einem Anpassungstest wird nicht nur der Lageparameter einer Verteilung getestet, sondern es wird die Verteilung als Ganzes untersucht.

Beim approximativen Binomialtest (siehe Tabelle 1) haben wir streng genommen bereits einen Anpassungstest durchgeführt. Bei der Nullhypothese $H_0 : p = p_0$ wird untersucht, ob es sich bei der betrachteten Zufallsgröße um eine Bernoulli-verteilte Zufallsvariable mit Parameter p_0 handelt. Beim χ^2-Test gehen wir nun einen Schritt weiter: Wir nehmen an, dass die Zufallsgröße X genau k verschiedene Werte annimmt. Ohne Beschränkung der Allgemeinheit sei $W_X = \{1, \ldots, k\}$. Die Nullhypothese lautet nun

$$H_0 : \Pr[X = i] = p_i \quad \text{für } i = 1, \ldots, k.$$

Tabelle: χ^2-**Anpassungstest**

Annahmen:

$X_1, \ldots, X_n$ seien unabhängig und identisch verteilt mit $W_{X_i} = \{1, \ldots, k\}$.

Hypothesen:

$$\begin{aligned} H_0 &: \Pr[X = i] = p_i \quad \text{für } i = 1, \ldots, k, \\ H_1 &: \Pr[X = i] \neq p_i \quad \text{für mindestens ein } i \in \{1, \ldots, k\}, \end{aligned}$$

Testgröße:

$$T = \sum_{i=1}^{k} \frac{(h_i - np_i)^2}{np_i},$$

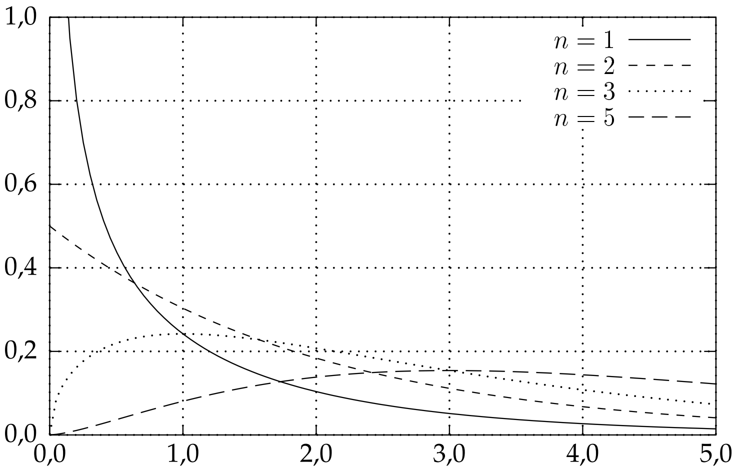
wobei h_i die Häufigkeit angibt, mit der $X_1, \ldots, X_n$ den Wert i angenommen haben.

Ablehnungskriterium für H_0 bei Signifikanzniveau α:

$$T > \chi^2_{k-1,1-\alpha};$$

dabei sollte gelten, dass $np_i \geq 1$ für alle i und $np_i \geq 5$ für mindestens 80% der Werte $i = 1, \ldots, k$.

Für die Testgröße T wird näherungsweise eine χ^2-Verteilung mit $k-1$ Freiheitsgraden angenommen. Die Werte dieser Verteilung finden sich in entsprechenden Tabellen in der Literatur. Damit diese Approximation gerechtfertigt ist, sollte gelten, dass $np_i \geq 1$ für alle i und $np_i \geq 5$ für mindestens 80% der Werte $i = 1, \ldots, k$. Das γ-Quantil einer χ^2-Verteilung mit k Freiheitsgraden bezeichnen wir mit $\chi^2_{k,\gamma}$.



Dichte der χ^2-Verteilung mit n Freiheitsgraden

Beispiel 122

Als Anwendung für den χ^2-Test wollen wir überprüfen, ob der Zufallszahlengenerator von Maple eine gute Approximation der Gleichverteilung liefert. Dazu lassen wir Maple $n = 100000$ Zufallszahlen aus der Menge $\{1, \ldots, 10\}$ generieren. Wir erwarten, dass jede dieser Zahlen mit gleicher Wahrscheinlichkeit $p_1 = \ldots = p_{10} = 1/10$ auftritt. Dies sei unsere Nullhypothese, die wir mit einem Signifikanzniveau von $\alpha = 0{,}05$ testen wollen.

Beispiel:

i	1	2	3	4	5	6	7	8	9	10
h_i	10102	10070	9972	9803	10002	10065	10133	9943	10009	9901

Für den Wert der Testgröße gilt $T = 8{,}9946$. Ferner erhalten wir $\chi^2_{9,0,95} \approx 16{,}919$. Der Test liefert also keinen Grund, die Nullhypothese abzulehnen.

Das Prinzip des χ^2-Anpassungstests kann in leicht abgewandelter Form auch noch zum Testen einiger anderer Hypothesen verwendet werden: Beim χ^2-Homogenitätstest wird überprüft, ob zwei oder mehrere Verteilungen identisch sind, während beim χ^2-Unabhängigkeitstest zwei Zufallsgrößen auf Unabhängigkeit untersucht werden. Beschreibungen dieser Tests findet man in der Literatur.

Kapitel IV Stochastische Prozesse

1. Einführung

Wir betrachten zeitliche Folgen von Zufallsexperimenten. Mathematisch beschreibt man diese durch einen so genannten stochastischen Prozess. Darunter versteht man eine Folge von Zufallsvariablen $(X_t)_{t \in T}$, die das Verhalten des Systems zu verschiedenen Zeitpunkten t angeben.

Wenn wir $T = \mathbb{N}_0$ annehmen, sprechen wir von einem stochastischen Prozess mit diskreter Zeit. Lässt man andererseits $T = \mathbb{R}_0^+$ zu, so spricht man von stochastischen Prozessen mit kontinuierlicher Zeit.

Eine besonders einfache Art von stochastischen Prozessen sind so genannte Markov-Ketten. Diese haben die Eigenschaft, dass der nächste Zustand des Prozesses zwar vom aktuellen Zustand abhängen darf, nicht aber von der Historie, d.h. davon, wie der aktuelle Zustand erreicht wurde.

2. Prozesse mit diskreter Zeit

2.1 Einführung

Definition 123

Eine (endliche) Markov-Kette (mit diskreter Zeit) über der Zustandsmenge $S = \{0, \ldots, n-1\}$ besteht aus einer unendlichen Folge von Zufallsvariablen $(X_t)_{t \in \mathbb{N}_0}$ mit Wertemenge S sowie einer Startverteilung q_0 mit $q_0^T \in \mathbb{R}^n$. Die Komponenten von q_0 sind hierbei ≥ 0 und addieren sich zu 1. Für jede Indexmenge $I \subseteq \{0, \ldots, t-1\}$ und beliebige Zustände i, j, s_k $(k \in I)$ gilt

$$\begin{aligned}\Pr[X_{t+1} = j \mid X_t = i, \ \forall k \in I: \ X_k = s_k] = \\ \Pr[X_{t+1} = j \mid X_t = i]\,.\end{aligned} \tag{9}$$

Sind die Werte

$$p_{ij} := \Pr[X_{t+1} = j \mid X_t = i]$$

von t unabhängig, so nennt man die Markov-Kette (zeit)homogen. In diesem Fall definiert man die Übergangsmatrix durch $P = (p_{ij})_{0 \le i,j < n}$. Wenn man $S = \mathbb{N}_0$ zulässt, so spricht man von einer unendlichen Markov-Kette.

Markov-Ketten sind nach Andrey Andreyevich Markov (1856–1922) benannt.

Bedingung (9) heißt Markov-Bedingung und besagt:
Wenn wir den Zustand i zum Zeitpunkt t kennen, so hängt die Übergangswahrscheinlichkeit zum Folgezustand j nur von i und j ab. Die Vergangenheit (Zustände zu Zeitpunkten $< t$) der Markov-Kette spielt keine Rolle. Das „Gedächtnis" der Markov-Kette besteht also nur aus ihrem aktuellen Zustand und sie „weiß" nicht, wie sie dorthin gekommen ist.

Bei einer zeithomogenen Markov-Kette hat die (absolute) Zeit t keinen Einfluss auf die Übergangswahrscheinlichkeiten p_{ij}, d.h. das Systemverhalten wird nur durch den aktuellen Zustand bestimmt und nicht durch eine absolute Uhr.

Wahrscheinlichkeitsraum einer Markov-Kette

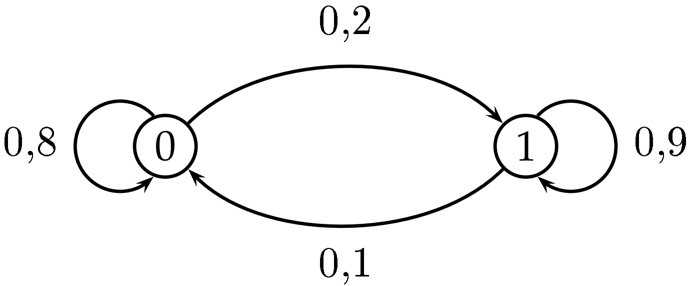
Nehmen wir an, dass wir die Kette von der Zeit 0 bis zur Zeit t_0 beobachten wollen. Wir bezeichnen die Folge von Zuständen, die von der Kette in dieser Zeit durchlaufen wurde, mit $\vec{x} = (x_0, x_1, \ldots, x_{t_0})$. $\Omega \subseteq S^{t_0+1}$ sei die Menge möglicher Zustandsfolgen. Einer beliebigen Folge $\omega := (x_0, x_1, \ldots, x_{t_0}) \in \Omega$ ordnen wir die Wahrscheinlichkeit

$$\Pr[\omega] = (q_0)_{x_0} \cdot \prod_{i=1}^{t_0} \Pr[X_i = x_i \mid X_{i-1} = x_{i-1}]$$

zu. Dadurch erhalten wir einen diskreten Wahrscheinlichkeitsraum im Sinne der Definition.

Beispiel 124

$$\Pr[X_{t+1} = 1 \mid X_t = 1] = 0{,}9,\ \Pr[X_{t+1} = 1 \mid X_t = 0] = 0{,}2$$
$$\Pr[X_{t+1} = 0 \mid X_t = 1] = 0{,}1,\ \Pr[X_{t+1} = 0 \mid X_t = 0] = 0{,}8$$



Einen bestimmten Ablauf des Systems kann man sich als so genannten Random Walk vorstellen.

Wenn wir uns beispielsweise zum Zeitpunkt $t = 0$ im Knoten 1 (also $X_0 = 1$) befinden, dann führen von dort zwei Kanten weiter, nämlich zu den Knoten 0 und 1. Diese Kanten sind mit Wahrscheinlichkeiten beschriftet, die sich zu Eins addieren. Gemäß dieser Wahrscheinlichkeiten entscheiden wir zufällig, wohin wir uns im nächsten Schritt begeben.

Wir können auch die Frage beantworten, mit welcher Wahrscheinlichkeit wir uns zum Zeitpunkt $t = 2$ im Knoten 1 befinden. Da wir vereinbarungsgemäß beim Knoten 1 starten, gibt es zwei mögliche Wege der Länge zwei durch den Graphen mit Endknoten 1, nämlich „111" und „101". Die Wahrscheinlichkeiten für diese Wege lauten $0{,}9 \cdot 0{,}9 = 0{,}9^2$ bzw. $0{,}1 \cdot 0{,}2$. Insgesamt erhalten wir also eine Wahrscheinlichkeit von $0{,}81 + 0{,}02 = 0{,}83$.

Auch eine Aussage über die erwartete Anzahl Schritte, die wir im Knoten 1 bis zum ersten Übergang zu Knoten 0 verbleiben, ist schnell getroffen. Die Wahrscheinlichkeit, dass man genau k Schritte verbleibt, ist $(0{,}9)^k \cdot 0{,}1$. Die Anzahl Schritte ist also geometrisch verteilt mit Erfolgswahrscheinlichkeit $0{,}1$. Der Erwartungswert ist daher $1/0{,}1 = 10$.

2.2 Berechnung von Übergangswahrscheinlichkeiten

Wir beschreiben die Situation zum Zeitpunkt t durch einen Zustandsvektor q_t (den wir als Zeilenvektor schreiben). Die i-te Komponente $(q_t)_i$ bezeichnet dabei die Wahrscheinlichkeit, mit der sich die Kette nach t Schritten im Zustand i aufhält.

Es gilt

$$\Pr[X_{t+1} = k] = \sum_{i=0}^{n-1} \Pr[X_{t+1} = k \mid X_t = i] \cdot \Pr[X_t = i],$$

also

$$(q_{t+1})_k = \sum_{i=0}^{n-1} p_{ik} \cdot (q_t)_i,$$

bzw. in Matrixschreibweise

$$q_{t+1} = q_t \cdot P.$$

Mit der Matrixschreibweise können wir q_t einfach durch die Startverteilung q_0 ausdrücken:

$$q_t = q_0 \cdot P^t \ .$$

Ebenso gilt wegen der Zeithomogenität allgemein für alle $t, k \in \mathbb{N}$:

$$q_{t+k} = q_t \cdot P^k.$$

Die Einträge von P^k geben an, mit welcher Wahrscheinlichkeit ein Übergang vom Zustand i zum Zustand j in genau k Schritten erfolgt.

$$p_{ij}^{(k)} := \Pr[X_{t+k} = j \mid X_t = i] = (P^k)_{ij}.$$

Exponentiation von Matrizen
Wenn P diagonalisierbar ist, so existiert eine Diagonalmatrix D und eine invertierbare Matrix B, so dass $P = B \cdot D \cdot B^{-1}$ gilt. Diese erhalten wir durch Berechnung der Eigenwerte und Eigenvektoren von P und durch Transformation von P in den Raum der Eigenvektoren.

Dann gilt

$$P^k = B \cdot D^k \cdot B^{-1} \; .$$

Beispiel 125

$$P = \begin{pmatrix} 0{,}8 & 0{,}2 \\ 0{,}1 & 0{,}9 \end{pmatrix}$$

Durch Bestimmung der Nullstellen des charakteristischen Polynoms der Matrix $(P - \lambda \cdot I)$ erhalten wir die Eigenwerte $0{,}7$ und 1, sowie die zugehörigen (rechten) Eigenvektoren

$$\nu_1 = \begin{pmatrix} -2 \\ 1 \end{pmatrix} \text{ und } \nu_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Beispiel 125

Damit

$$D = \begin{pmatrix} 0{,}7 & 0 \\ 0 & 1 \end{pmatrix} \text{ und } B = \begin{pmatrix} -2 & 1 \\ 1 & 1 \end{pmatrix}$$

und

$$B^{-1} = \begin{pmatrix} -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}.$$

Damit ergibt sich beispielsweise

$$P^3 = \begin{pmatrix} -2 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0{,}7^3 & 0 \\ 0 & 1^3 \end{pmatrix} \begin{pmatrix} -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix} \approx \begin{pmatrix} 0{,}562 & 0{,}438 \\ 0{,}219 & 0{,}781 \end{pmatrix}$$

2.3 Ankunftswahrscheinlichkeiten und Übergangszeiten

Bei der Analyse von Markov-Ketten treten oftmals Fragestellungen auf, die sich auf zwei bestimmte Zustände i und j beziehen:

- Wie wahrscheinlich ist es, von i irgendwann nach j zu kommen?
- Wie viele Schritte benötigt die Kette im Mittel, um von i nach j zu gelangen?

Definition 126

Die Zufallsvariable

$$T_{ij} := \min\{n \geq 0 \mid X_n = j, \text{ wenn } X_0 = i\}$$

zählt die Anzahl der Schritte, die von der Markov-Kette für den Weg von i nach j benötigt werden. T_{ij} nennen wir die Übergangszeit (engl. hitting time) vom Zustand i zum Zustand j. Wenn j nie erreicht wird, setzen wir $T_{ij} = \infty$.

Ferner definieren wir $h_{ij} := \mathbb{E}[T_{ij}]$.

Die Wahrscheinlichkeit, vom Zustand i nach beliebig vielen Schritten in den Zustand j zu gelangen, nennen wir Ankunftswahrscheinlichkeit f_{ij}. Formal definieren wir

$$f_{ij} := \Pr[T_{ij} < \infty].$$

Im Fall $i = j$ gilt $T_{ii} = 0$ und somit auch $h_{ii} = 0$, sowie $f_{ii} = 1$. Anschaulich ist dies klar: Wenn Anfangs- und Zielzustand identisch sind, so ist die Übergangszeit gleich Null. Für viele Zwecke ist es andererseits auch interessant zu messen, wie lange es dauert, bis Zustand i zu einem *späteren* Zeitpunkt wieder besucht wird. Wir ergänzen Definition 126 für diesen Fall.

Definition 127

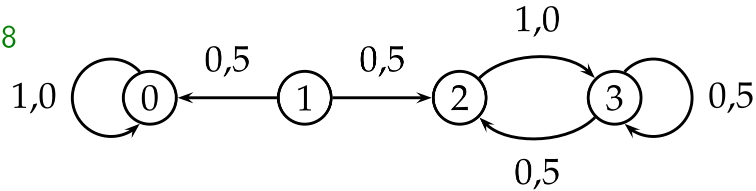
Die Zufallsvariable

$$T_i := \min\{n \geq 1 \mid X_n = i, \text{ wenn } X_0 = i\}$$

zählt die Anzahl Schritte, die von der Markov-Kette benötigt werden, um von i nach i zurückzukehren (Rückkehrzeit, engl. recurrence time). Der Erwartungswert sei $h_i := \mathbb{E}[T_i]$. Die Wahrscheinlichkeit, mit der T_i einen endlichen Wert annimmt, nennt man Rückkehrwahrscheinlichkeit:

$$f_i := \Pr[T_i < \infty].$$

Beispiel 128

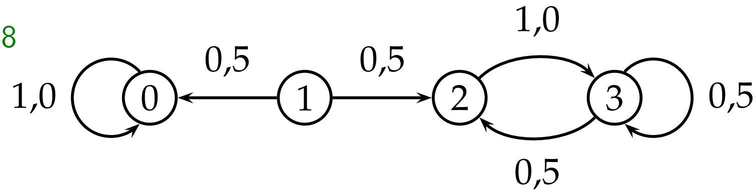


Beispiel zur Berechnung von f_{ij} und h_{ij}

Wir betrachten die obige Markov-Kette. Einige Besonderheiten fallen sofort auf:

- Beginnt man im Zustand 0, so kann man niemals einen der übrigen Zustände erreichen. Die Übergangszeiten T_{01}, T_{02} und T_{03} sind daher ∞.

Beispiel 128

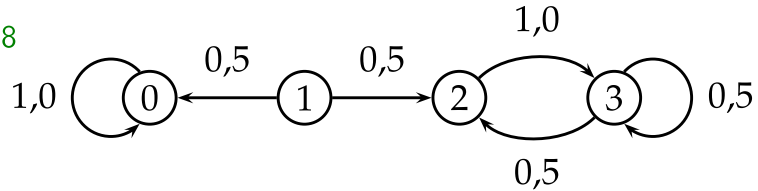


- Beginnt man im Zustand 1, so entscheidet sich im ersten Schritt, ob die Kette sich zukünftig im „linken Teil“ (Zustand 0) oder im „rechten Teil“ (Zustand 2 und 3) aufhält. Für die Übergangszeit T_{10} gilt daher

$$T_{10} = \begin{cases} 1 & \text{falls } X_1 = 0, \\ \infty & \text{falls } X_1 = 2. \end{cases}$$

Wegen $\Pr[X_1 = 0 \mid X_0 = 1] = 0{,}5$ folgt $f_{10} = 0{,}5$ und $\mathbb{E}[T_{10}]$ existiert nicht.

Beispiel 128



- Beginnt man im Zustand 2 oder 3, so wird die Kette auch weiterhin zwischen den Zuständen 2 und 3 „hin und her pendeln". Genauer:
  Die Anzahl der Schritte, in denen die Kette im Zustand 3 bleibt, ist geometrisch verteilt mit Parameter $0{,}5$. Der Zustand 3 wird daher im Mittel nach $1/0{,}5 = 2$ Schritten verlassen. Da Zustand 2 der einzige Nachbar von 3 ist, folgt $h_{32} = 2$ und somit insbesondere auch $f_{32} = 1$.

Lemma 129

Für die erwarteten Übergangs-/Rückkehrzeiten gilt

$$h_{ij} = 1 + \sum_{k \neq j} p_{ik} h_{kj} \text{ für alle } i, j \in S, i \neq j,$$

$$h_j = 1 + \sum_{k \neq j} p_{jk} h_{kj} \ ,$$

sofern die Erwartungswerte h_{ij} und h_{kj} existieren.

Für die Ankunfts-/Rückkehrwahrscheinlichkeiten gilt analog

$$f_{ij} = p_{ij} + \sum_{k \neq j} p_{ik} f_{kj} \text{ für alle } i, j \in S, i \neq j;$$

$$f_j = p_{jj} + \sum_{k \neq j} p_{jk} f_{kj} \ .$$

Beweis:

Sei $i \neq j$. Wir bedingen auf das Ergebnis des ersten Schritts der Markov-Kette und erhalten aufgrund der Gedächtnislosigkeit $\Pr[T_{ij} < \infty \mid X_1 = k] = \Pr[T_{kj} < \infty]$ für $k \neq j$ sowie $\Pr[T_{ij} < \infty \mid X_1 = j] = 1$.

$$\begin{aligned} f_{ij} = \Pr[T_{ij} < \infty] &= \sum_{k \in S} \Pr[T_{kj} < \infty \mid X_1 = k] \cdot p_{ik} \\ &= p_{ij} + \sum_{k \neq j} \Pr[T_{kj} < \infty] \cdot p_{ik} = p_{ij} + \sum_{k \neq j} p_{ik} f_{kj} \,. \end{aligned}$$

Die Ableitung für f_j (also $i = j$) ist analog.

Beweis:

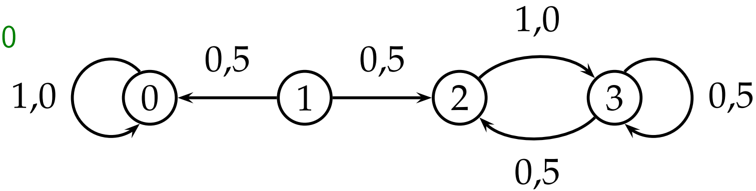
Sei wiederum $i \neq j$. Wegen der Gedächtnislosigkeit folgt $\mathbb{E}[T_{ij} \mid X_1 = k] = 1 + \mathbb{E}[T_{kj}]$ für $k \neq j$. Ferner gilt $\mathbb{E}[T_{ij} \mid X_1 = j] = 1$.

Bedingen wir wieder auf das Ergebnis des ersten Schritts, so folgt (siehe Satz 36):

$$\begin{aligned} h_{ij} = \mathbb{E}[T_{ij}] &= \sum_{k \in S} \mathbb{E}[T_{ij} \mid X_1 = k] \cdot p_{ik} \\ &= p_{ij} + \sum_{k \neq j} (1 + \mathbb{E}[T_{kj}]) \cdot p_{ik} = 1 + \sum_{k \neq j} h_{kj} \cdot p_{ik}. \end{aligned}$$

Wiederum ist die Herleitung für h_j analog. $\square$

Beispiel 130



Für die Berechnung der Übergangszeiten für die Zustände 2 und 3 erhalten wir die Gleichungen

$$h_2 = 1 + h_{32}, \qquad h_3 = 1 + \tfrac{1}{2} \cdot h_{23}$$

und

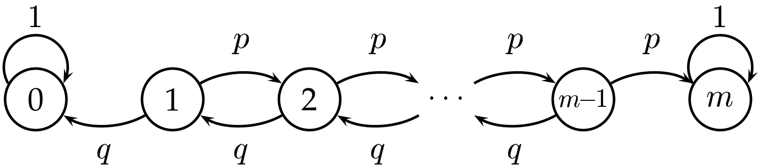
$$h_{23} = 1, \qquad h_{32} = 1 + \tfrac{1}{2} h_{32} = 2\,.$$

Durch Lösen dieses Gleichungssystems erhalten wir die Werte $h_2 = 3$, $h_3 = 1{,}5$, $h_{23} = 1$ und $h_{32} = 2$, die man leicht verifiziert. Die Ankunftswahrscheinlichkeiten lassen sich analog herleiten. Man erhält $f_2 = f_3 = f_{23} = f_{32} = 1$.

2.4 Das Gambler's Ruin Problem

Anna und Bodo spielen Poker, bis einer von ihnen bankrott ist. A verfügt über Kapital a, und B setzt eine Geldmenge in Höhe von $m - a$ aufs Spiel. Insgesamt sind also m Geldeinheiten am Spiel beteiligt. In jeder Pokerrunde setzen A und B jeweils eine Geldeinheit. A gewinnt jedes Spiel mit Wahrscheinlichkeit p. B trägt folglich mit Wahrscheinlichkeit $q := 1 - p$ den Sieg davon. Wir nehmen an, dass diese Wahrscheinlichkeiten vom bisherigen Spielverlauf und insbesondere vom Kapitalstand der Spieler unabhängig sind.

Wir modellieren das Spiel durch die Markov-Kette



A interessiert sich für die Wahrscheinlichkeit, mit der sie B in den Ruin treibt, also für die Wahrscheinlichkeit $f_{a,m}$ (wir schreiben hier der Deutlichkeit halber $f_{i,j}$ statt f_{ij}).

Wir erhalten:

$$\begin{aligned} f_{i,m} &= p \cdot f_{i+1,m} + q \cdot f_{i-1,m} \text{ für } 1 \leq i < m-1, \\ f_{m-1,m} &= p + q \cdot f_{m-2,m}, \\ f_{0,m} &= 0. \end{aligned} \tag{10}$$

Wir wollen nun $f_{i,m}$ allgemein als Funktion von m berechnen. Dazu beobachten wir zunächst, dass wir (10) wegen $f_{m,m} = 1$ umschreiben können zu

$$f_{i+1,m} = (1/p) \cdot f_{i,m} - (q/p) \cdot f_{i-1,m} \text{ für } 1 \leq i < m. \tag{11}$$

Wir ergänzen (11) um die Anfangswerte

$$f_{0,m} = 0 \text{ und } f_{1,m} = \xi.$$

(Für den Moment fassen wir ξ als Variable auf. Nach Lösung der Rekursion werden wir ξ so wählen, dass die Bedingung $f_{m,m} = 1$ erfüllt ist.)

Als Lösung dieser linearen homogenen Rekursionsgleichung 2. Ordnung (11) ergibt sich für $p \neq 1/2$:

$$f_{i,m} = \frac{p \cdot \xi}{2p-1} \cdot \left(1 - \left(\frac{1-p}{p}\right)^i\right).$$

Setzen wir nun $i = m$, so folgt aus $f_{m,m} = 1$, dass

$$\xi = \frac{2p-1}{p \cdot \left(1 - \left(\frac{1-p}{p}\right)^m\right)}$$

gelten muss.

Insgesamt erhalten wir somit das Ergebnis:

$$f_{j,m} = \frac{1 - \left(\frac{1-p}{p}\right)^j}{1 - \left(\frac{1-p}{p}\right)^m}.$$

Für $p = 1/2$ verläuft die Rechnung ähnlich.

Beispiel 131

Wir wollen berechnen, wie lange A und B im Mittel spielen können, bis einer von ihnen bankrott geht.

$h_{a,m}$ eignet sich dazu i.a. nicht (warum?).

Wir betrachten stattdessen:

$$T_i' := \text{„Anzahl der Schritte von Zustand } i \text{ nach Zustand 0 oder } m\text{“}$$

und setzen

$$d_i := \mathbb{E}[T_i'].$$

Offensichtlich gilt $d_0 = d_m = 0$ und für $1 \leq i < m$

$$d_i = qd_{i-1} + pd_{i+1} + 1\ .$$

Beispiel (Forts.)

Wir betrachten nun nur den Fall $p = q = 1/2$ und erhalten

$$d_i = i \cdot (m - i) \text{ für alle } i = 0, \ldots, m.$$

Wegen $d_i \leq mi \leq m^2$ folgt also, dass das Spiel unabhängig vom Startzustand im Mittel nach höchstens m^2 Schritten beendet ist.

2.5 Stationäre Verteilung

Reale dynamische Systeme laufen oft über eine lange Zeit. Für solche Systeme ist es sinnvoll, das Verhalten für $t \to \infty$ zu berechnen.

Wir betrachten wieder die Markov-Kette aus unserem Beispiel. Wir hatten gezeigt, dass für die Übergangsmatrix P gilt:

$$P = B \cdot D \cdot B^{-1} = \begin{pmatrix} -2 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \frac{7}{10} & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}.$$

Daraus folgt

$$P^t = B \cdot D^t \cdot B^{-1} = \begin{pmatrix} -2 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \left(\frac{7}{10}\right)^t & 0 \\ 0 & 1^t \end{pmatrix} \cdot \begin{pmatrix} -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix},$$

und für $t \to \infty$ erhalten wir

$$\lim_{t\to\infty} P^t = \begin{pmatrix} -2 & 1 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}.$$

Für eine beliebige Startverteilung $q_0 = (a, 1-a)$ folgt

$$\begin{aligned}\lim_{t\to\infty} q_t = \lim_{t\to\infty} q_0 \cdot P^t &= (a, 1-a) \cdot \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix} \\ &= \left(\frac{1}{3}a + \frac{1}{3}(1-a),\ \frac{2}{3}a + \frac{2}{3}(1-a)\right) = (\frac{1}{3},\ \frac{2}{3}).\end{aligned}$$

Das System konvergiert also unabhängig vom Startzustand in eine feste Verteilung. Der zugehörige Zustandsvektor $\pi = (\frac{1}{3}, \frac{2}{3})$ hat eine interessante Eigenschaft:

$$\pi \cdot P = (\frac{1}{3}, \frac{2}{3}) \cdot \begin{pmatrix} 0{,}8 & 0{,}2 \\ 0{,}1 & 0{,}9 \end{pmatrix} = (\frac{1}{3}, \frac{2}{3}) = \pi.$$

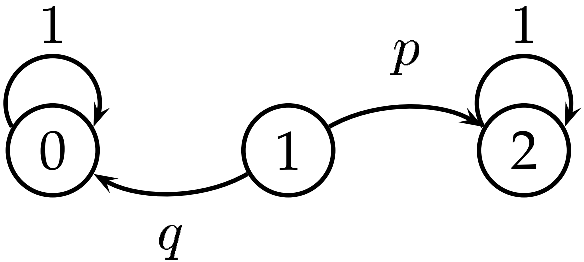
π ist also ein Eigenvektor der Matrix P zum Eigenwert 1 bezüglich Multiplikation von links. Dies bedeutet: Wenn die Kette einmal den Zustandsvektor π angenommen hat, so bleibt dieser bei allen weiteren Übergängen erhalten.

Definition 132

P sei die Übergangsmatrix einer Markov-Kette. Einen Zustandsvektor π mit $\pi = \pi \cdot P$ nennen wir stationäre Verteilung der Markov-Kette.

Besitzen alle Markov-Ketten die Eigenschaft, dass sie unabhängig vom Startzustand in eine bestimmte stationäre Verteilung konvergieren?

Nein!



Eine Markov-Kette mit absorbierenden Zuständen

Die Abbildung zeigt die Kette aus dem „gamblers ruin problem" für $m = 2$. Man sieht sofort, dass hier sowohl $\pi_1 = (1, 0, 0)$ als auch $\pi_2 = (0, 0, 1)$ stationäre Verteilungen sind. Die beiden Zustände 0 und 2 haben jeweils keine ausgehenden Kanten. Solche Zustände heißen absorbierend.

Definition 133

Wir bezeichnen einen Zustand i als absorbierend, wenn aus ihm keine Übergänge herausführen, d.h. $p_{ij} = 0$ für alle $j \neq i$ und folglich $p_{ii} = 1$.

Ein Zustand i heißt transient, wenn $f_i < 1$, d.h. mit positiver Wahrscheinlichkeit $1 - f_i > 0$ kehrt der Prozess nach einem Besuch von i nie mehr dorthin zurück.

Ein Zustand i mit $f_i = 1$ heißt rekurrent.

Definition 134

Eine Markov-Kette heißt irreduzibel, wenn es für alle Zustandspaare $i, j \in S$ eine Zahl $n \in \mathbb{N}$ gibt, so dass $p_{ij}^{(n)} > 0$.

Die Definition besagt anschaulich, dass jeder Zustand von jedem anderen Zustand aus mit positiver Wahrscheinlichkeit erreicht werden kann, wenn man nur genügend viele Schritte durchführt. Dies ist bei endlichen Markov-Ketten genau dann der Fall, wenn der gerichtete Graph des Übergangsdiagramms stark zusammenhängend ist.

Lemma 135

Für irreduzible endliche Markov-Ketten gilt: $f_{ij} = \Pr[T_{ij} < \infty] = 1$ *für alle Zustände* $i, j \in S$. *Zusätzlich gilt auch, dass die Erwartungswerte* $h_{ij} = \mathbb{E}[T_{ij}]$ *alle existieren.*

Beweis:

Wir betrachten zunächst den Beweis für die Existenz von h_{ij}.
Für jeden Zustand k gibt es nach Definition der Irreduzibilität ein n_k, so dass $p_{kj}^{(n_k)} > 0$. Wir halten n_k fest und setzen $n := \max_k n_k$ und $p := \min_k p_{kj}^{(n_k)}$.

Von einem beliebigen Zustand aus gelangen wir nach höchstens n Schritten mit Wahrscheinlichkeit mindestens p nach j. Wir unterteilen die Zeit in Phasen zu n Schritten und nennen eine Phase erfolgreich, wenn während dieser Phase ein Besuch bei j stattgefunden hat. Die Anzahl von Phasen bis zur ersten erfolgreichen Phase können wir durch eine geometrische Verteilung mit Parameter p abschätzen. Die erwartete Anzahl von Phasen ist somit höchstens $1/p$, und wir schließen $h_{ij} \leq (1/p)n$.
Daraus folgt sofort, dass auch $f_{ij} = \Pr[T_{ij} < \infty] = 1$. $\square$

Satz 136

Eine irreduzible endliche Markov-Kette besitzt eine eindeutige stationäre Verteilung π, und es gilt $\pi_j = 1/h_{jj}$ für alle $j \in S$.

Beweis:

Wir zeigen zunächst, dass es einen Vektor $\pi \neq 0$ mit $\pi = \pi P$ gibt. Sei $e := (1, \ldots, 1)^T$ der All-1-Vektor und I die Einheitsmatrix. Für jede Übergangsmatrix P gilt $P \cdot e = e$, da sich die Einträge der Zeilen von P zu Eins addieren. Daraus folgt $0 = Pe - e = (P - I)e$, und die Matrix $P - I$ ist somit singulär. Damit ist auch die transponierte Matrix $(P - I)^T = P^T - I$ singulär. Es gibt also einen (Spalten-)Vektor $\pi \neq 0$ mit $(P^T - I) \cdot \pi = 0$ bzw. $\pi^T P = \pi^T$. Wir betrachten zunächst den Fall, dass $\sum_i \pi_i \neq 0$. Dann können wir o.B.d.A. annehmen, dass π normiert ist, also dass $\sum_i \pi_i = 1$ gilt.

Beweis (Forts.):

Wegen Lemma 135 existieren die Erwartungswerte h_{ij}. Für jeden Zustand $j \in S$ gelten somit nach Lemma 129 die Gleichungen

$$\pi_i h_{ij} = \pi_i\Big(1 + \sum_{k \neq j} p_{ik} h_{kj}\Big) \quad \text{für } i \in S,\ i \neq j.$$

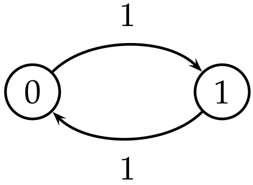
Wir addieren diese Gleichungen und erhalten wegen $\sum_i \pi_i = 1$

$$\begin{aligned} \pi_j h_j + \sum_{i \neq j} \pi_i h_{ij} &= 1 + \sum_{i \in S} \sum_{k \neq j} \pi_i p_{ik} h_{kj} \\ &= 1 + \sum_{k \neq j} h_{kj} \sum_{i \in S} \pi_i p_{ik} = 1 + \sum_{k \neq j} \pi_k h_{kj}. \end{aligned}$$

Wegen $h_j > 0$ ist auch $\pi_j = 1/h_j$ positiv, und π stellt somit einen zulässigen Zustandsvektor dar.

Für den Fall $\sum_i \pi_i = 0$ zeigt die entsprechende Rechnung wie zuvor, dass $\pi_j = 0$ für alle $j \in S$ gilt. Dies steht im Widerspruch zu $\pi \neq 0$. $\square$

Auch wenn eine Markov-Kette irreduzibel ist und somit eine eindeutige stationäre Verteilung besitzt, so muss sie nicht zwangsläufig in diese Verteilung konvergieren.



Eine Markov-Kette mit periodischen Zuständen

Als Startverteilung nehmen wir $q_0 = (1, 0)$ an. Es gilt:

$$q_t = \begin{cases} (1,0) & \text{falls } t \text{ gerade,} \\ (0,1) & \text{sonst.} \end{cases}$$

Die Kette pendelt also zwischen den beiden Zustandsvektoren $(1, 0)$ und $(0, 1)$ hin und her.

Definition 137

Die Periode eines Zustands j ist definiert als die größte Zahl $\xi \in \mathbb{N}$, so dass gilt:

$$\{n \in \mathbb{N}_0 \mid p_{jj}^{(n)} > 0\} \subseteq \{i \cdot \xi \mid i \in \mathbb{N}_0\}$$

Ein Zustand mit Periode $\xi = 1$ heißt aperiodisch. Wir nennen eine Markov-Kette aperiodisch, wenn alle Zustände aperiodisch sind.

Für ein $n \in \mathbb{N}$ gilt $p_{ii}^{(n)} > 0$ genau dann, wenn es im Übergangsdiagramm einen geschlossenen Weg von i nach i der Länge n gibt.
Damit folgt insbesondere:
Ein Zustand $i \in S$ einer endlichen Markov-Kette ist sicherlich dann aperiodisch, wenn er im Übergangsdiagramm

- eine Schleife besitzt (also $p_{ii} > 0$) oder
- auf mindestens zwei geschlossenen Wegen W_1 und W_2 liegt, deren Längen l_1 und l_2 teilerfremd sind (für die also $\mathrm{ggT}(l_1, l_2) = 1$ gilt).

Lemma 138

Ein Zustand $i \in S$ ist genau dann aperiodisch, falls gilt: Es gibt ein $n_0 \in \mathbb{N}$, so dass $p_{ii}^{(n)} > 0$ für alle $n \in \mathbb{N}, n \geq n_0$.

Beweis:

Da je zwei aufeinanderfolgende natürliche Zahlen teilerfremd sind, folgt aus der Existenz eines n_0 mit der im Lemma angegebenen Eigenschaft sofort die Aperiodizität des Zustands. Nehmen wir daher umgekehrt an, dass der Zustand i aperiodisch ist. Mit Hilfe des erweiterten euklidischen Algorithmus kann man die folgende Aussage zeigen. Für je zwei natürliche Zahlen $a, b \in \mathbb{N}$ gibt es ein $n_0 \in \mathbb{N}$, so dass gilt: Bezeichnet $d := \mathrm{ggT}(a, b)$ den größten gemeinsamen Teiler von a und b, so gibt es für alle $n \in \mathbb{N}, n \geq n_0$ nichtnegative Zahlen $x, y \in \mathbb{N}_0$ mit $nd = xa + yb$.

Beweis (Forts.):

Wegen $p_{ii}^{(xa+yb)} \geq (p_{ii}^{(a)})^x \cdot (p_{ii}^{(b)})^y$ folgt daraus unmittelbar: Gilt für $a, b \in \mathbb{N}$, dass sowohl $p_{ii}^{(a)}$ als auch $p_{ii}^{(b)}$ positiv sind, so gilt auch $p_{ii}^{(nd)} > 0$ für alle $n \in \mathbb{N}$, $n \geq n_0$.

Aus der Aperiodizität des Zustand i folgt andererseits, dass es Werte $a_0, \ldots, a_k$ geben muss mit $p_{ii}^{(a_i)} > 0$ und der Eigenschaft, dass für $d_1 = \mathrm{ggT}(a_0, a_1)$ und $d_i := \mathrm{ggT}(d_{i-1}, a_i)$ für $i = 2, \ldots, k$ gilt: $d_1 > d_2 > \cdots > d_k = 1$.
Aus beiden Beobachtungen zusammen folgt die Behauptung. $\square$

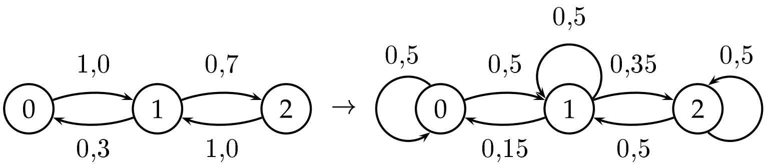
Korollar 139

Für irreduzible, aperiodische endliche Markov-Ketten gilt: Es gibt ein $t \in \mathbb{N}$, so dass unabhängig vom Startzustand $(q_t)_i > 0$ für alle $i \in S$.

Beweis:

Aus der Irreduzibilität folgt, dass die Markov-Kette jeden Zustand $i \in S$ irgendwann besuchen wird. Wegen Lemma 138 wissen wir ferner, dass die Kette hinreichend viele Schritte nach dem ersten Besuch in i in jedem folgenden Zeitschritt mit positiver Wahrscheinlichkeit zu i zurückkehren wird. Da die Kette endlich ist, gibt es daher ein n_0, so dass die Kette sich unabhängig vom Startzustand für alle $n \geq n_0$ in jedem Zustand $i \in S$ mit positiver Wahrscheinlichkeit aufhält. □

Die Aperiodizität einer irreduziblen Markov-Kette kann auf einfache Weise sichergestellt werden. Man fügt an alle Zustände so genannte Schleifen an. Diese versieht man mit der Übergangswahrscheinlichkeit $p = 1/2$ und halbiert die Wahrscheinlichkeiten an allen übrigen Kanten.



Einführung von Schleifen

Bei irreduziblen Ketten genügt es, eine einzige Schleife einzuführen, um die Aperiodizität der Kette sicherzustellen.

Definition 140

Irreduzible, aperiodische Markov-Ketten nennt man ergodisch.

Satz 141 (Fundamentalsatz für ergodische Markov-Ketten)

Für jede ergodische endliche Markov-Kette $(X_t)_{t\in\mathbb{N}_0}$ *gilt unabhängig vom Startzustand*

$$\lim_{n\to\infty} q_n = \pi,$$

wobei π *die eindeutige stationäre Verteilung der Kette bezeichnet.*

Beweis:

Gemäß Satz 136 existiert eine stationäre Verteilung π. Wir zeigen, dass für beliebige Zustände i und k gilt

$$p_{ik}^{(n)} \to \pi_k \quad \text{für } n \to \infty.$$

Daraus folgt die Behauptung, da

$$(q_n)_k = \sum_{i\in S} (q_0)_i \cdot p_{ik}^{(n)} \to \pi_k \cdot \sum_{i\in S} (q_0)_i = \pi_k.$$

Beweis (Forts.):

$(Y_t)_{t\in\mathbb{N}_0}$ sei eine unabhängige Kopie der Kette $(X_t)_{t\in\mathbb{N}_0}$. Für den Prozess $Z_t := (X_t, Y_t)$ $(t \in \mathbb{N}_0)$, bei dem die Ketten X_t und Y_t gewissermaßen „parallel" betrieben werden, gilt also

$$\begin{aligned}
&\Pr[(X_{t+1}, Y_{t+1}) = (j_x, j_y) \mid (X_t, Y_t) = (i_x, i_y)] \\
&\quad= \Pr[X_{t+1} = j_x \mid X_t = i_x] \cdot \Pr[Y_{t+1} = j_y \mid Y_t = i_y] \\
&\quad= p_{i_x j_x} \cdot p_{i_y j_y}.
\end{aligned}$$

$(Z_t)_{t\in\mathbb{N}_0}$ ist daher ebenfalls eine Markov-Kette. Für die Wahrscheinlichkeit, in n Schritten von (i_x, i_y) nach (j_x, j_y) zu gelangen, erhält man analog $p^{(n)}_{i_x j_x} p^{(n)}_{i_y j_y}$, was für genügend großes n gemäß Lemma 138 positiv ist. $(Z_t)_{t_0\in\mathbb{N}}$ ist daher ebenfalls ergodisch.

Beweis (Forts.):

Wir starten nun Z_t so, dass die Ketten X_t und Y_t in verschiedenen Zuständen i_x bzw. i_y beginnen, und interessieren uns für den Zeitpunkt H, bei dem sich X_t und Y_t zum ersten Mal im gleichen Zustand befinden.

Die Menge der Zustände von Z_t ist gegeben durch $S \times S$. Wir definieren die Menge

$$M := \{(x, y) \in S \times S \mid x = y\}.$$

von Zuständen der Kette Z_t, an denen sich X_t und Y_t „treffen". Definieren wir nun die Treffzeit H durch

$$H := \max\{T_{(i_x,i_y),(j_x,j_y)} \mid (i_x, i_y) \in S \times S, (j_x, j_y) \in M\},$$

so folgt aus Lemma 135 und der Endlichkeit der Markov-Kette sofort, dass $\Pr[H < \infty] = 1$ und $\mathbb{E}[H] < \infty$.

Beweis (Forts.):

Da die weitere Entwicklung der Ketten X_t und Y_t ab dem Zeitpunkt H nur vom Zustand $X_H = Y_H$ und der Übergangsmatrix abhängt, wird jeder Zustand $s \in S_Z$ zu den Zeiten $t \geq H$ von X_t und Y_t mit derselben Wahrscheinlichkeit angenommen. Es gilt also $\Pr[X_t = s \mid t \geq H] = \Pr[Y_t = s \mid t \geq H]$ und somit auch

$$\Pr[X_t = s, t \geq H] = \Pr[Y_t = s, t \geq H]. \tag{12}$$

Als Startzustand wählen wir für die Kette X_t den Zustand i, während Y_t in der stationären Verteilung π beginnt (und natürlich auch bleibt). Damit erhalten wir für einen beliebigen Zustand $k \in S$ und $n \geq 1$

$$\begin{aligned}
|p_{ik}^{(n)} - \pi_k| &= |\Pr[X_n = k] - \Pr[Y_n = k]| \\
&= |\Pr[X_n = k, n \geq H] + \Pr[X_n = k, n < H] \\
&\quad -\Pr[Y_n = k, n \geq H] - \Pr[Y_n = k, n < H]|.
\end{aligned}$$

Beweis (Forts.):

Nun können wir (12) anwenden und schließen, dass

$$|p_{ik}^{(n)} - \pi_k| = |\Pr[X_n = k, n < H] - \Pr[Y_n = k, n < H]|.$$

Zur Abschätzung dieses Ausdrucks benutzen wir die Abschätzung

$$|\Pr[A \cap B] - \Pr[A \cap C]| \leq \Pr[A].$$

für beliebige Ereignisse A, B und C (die offensichtlich ist).

Wir erhalten

$$|p_{ik}^{(n)} - \pi_k| \leq \Pr[n < H].$$

Da $\Pr[H < \infty] = 1$, gilt $\Pr[n < H] \to 0$ für $n \to \infty$, d.h. die Wahrscheinlichkeiten $p_{ik}^{(n)}$ konvergieren für $n \to \infty$ gegen π_k. $\square$

2.6 Doppeltstochastische Matrizen

Wie berechnet man die nach Satz 141 (eindeutig bestimmte) stationäre Verteilung, gegen die ergodische endliche Markov-Ketten für jede Startverteilung konvergieren?

Eine Möglichkeit besteht darin, das lineare Gleichungssystem $\pi \cdot P = \pi$ aufzustellen und zu lösen. Für größere Matrizen ist dieses Verfahren allerdings im Allgemeinen sehr aufwändig.
Wir stellen hier einen anderen Ansatz vor.

Definition 142

Eine $n \times n$ Matrix $P = (p_{ij})_{0 \le i,j < n}$ heißt stochastisch, falls alle Einträge p_{ij} nichtnegativ und alle Zeilensummen gleich Eins sind:

$$\sum_{j=0}^{n-1} p_{ij} = 1 \text{ für alle } i = 0, \ldots, n-1.$$

Sind zusätzlich auch alle Spaltensummen gleich 1, also

$$\sum_{i=0}^{n-1} p_{ij} = 1 \text{ für alle } j = 0, \ldots, n-1,$$

so nennt man P doppeltstochastisch.

Die Übergangsmatrix einer Markov-Kette ist immer stochastisch, und umgekehrt.

Lemma 143

Ist P eine doppeltstochastische $n \times n$ Matrix, so ist $\pi = (\frac{1}{n}, \ldots, \frac{1}{n})$ ein Eigenvektor zum Eigenwert 1 bezüglich Multiplikation von links:

$$\pi = \pi \cdot P.$$

Beweis:

Für alle $0 \leq k < n$ gilt:

$$(\pi \cdot P)_k = \sum_{i=0}^{n-1} \pi_i \cdot p_{ik} = \frac{1}{n} \underbrace{\sum_{i=0}^{n-1} p_{ik}}_{=1} = \frac{1}{n} = \pi_k.$$

$\square$

Zusammen mit Satz 141 erhalten wir damit sofort:

Satz 144

Für jede ergodische endliche Markov-Kette $(X_t)_{t\in\mathbb{N}_0}$ mit doppeltstochastischer Übergangsmatrix gilt unabhängig vom Startzustand

$$\lim_{t\to\infty} q_t = (\tfrac{1}{n}, \ldots, \tfrac{1}{n}),$$

wobei n die Kardinalität der Zustandsmenge bezeichne.

Beweis:

Klar! $\square$

Beispiel 145

Anna und Bodo verabreden sich wieder einmal zu einer Partie Poker. Misstrauisch geworden durch ihre Verluste beim letzten Rendezvous verdächtigt Anna mittlerweile ihren Spielpartner, beim Mischen zu mogeln. Um ganz sicher zu gehen, dass die Karten zukünftig auch wirklich gut gemischt werden, schlägt sie folgendes Verfahren vor: Der Stapel mit Karten wird verdeckt hingelegt; dann werden m-mal jeweils zwei Karten daraus zufällig ausgewählt und vertauscht. Soll Bodo dieser Prozedur zustimmen?

Beispiel 145

Wir modellieren den oben skizzierten Mischvorgang durch eine Markov-Kette. Als Zustandsmenge S wählen wir alle möglichen Anordnungen der Karten. Identifizieren wir die Karten mit den Zahlen $[n] = \{1, \ldots, n\}$, so besteht S aus der Menge aller Permutationen der Menge $[n]$.

Betrachten wir nun zwei verschiedene Permutationen $\sigma, \rho \in S$. Nach Definition der Markov-Kette ist die Übergangswahrscheinlichkeit $p_{\sigma,\rho}$ genau dann positiv, wenn es $i, j \in [n]$, $i \neq j$, gibt, so dass

$$\rho(k) = \begin{cases} \sigma(j) & \text{falls } k = i, \\ \sigma(i) & \text{falls } k = j, \\ \sigma(k) & \text{sonst.} \end{cases}$$

Beispiel 145

Da nach Voraussetzung i und j zufällig gewählt werden (und es genau $\binom{n}{2}$ solcher Paare i, j gibt), gilt in diesem Fall $p_{\sigma,\rho} = 1/\binom{n}{2}$.

Da man jede Vertauschung zweier Karten durch nochmaliges Vertauschen wieder rückgängig machen kann, sieht man auch sofort ein, dass $p_{\sigma,\rho} = p_{\rho,\sigma}$ gilt. Die Übergangsmatrix P ist also symmetrisch und damit insbesondere auch doppeltstochastisch. Aus Satz 144 folgt somit, dass die Markov-Kette unabhängig von der Startverteilung zur Gleichverteilung konvergiert.

Der von Anna vorgeschlagene Mischvorgang ist also in der Tat sinnvoll: Für $m \to \infty$ konvergiert die Wahrscheinlichkeitsverteilung für die sich ergebende Kartenreihenfolge gegen die Gleichverteilung, die Karten sind also bestens gemischt!

Beispiel 145

Anmerkung: Man kann zeigen, dass für n Karten bereits $m = O(n \log n)$ Vertauschungen genügen, um einen gut durchmischten Kartenstapel zu erhalten.