Wahrscheinlichkeitstheorie und stochastische Prozesse für Informatik
Skriptum

Karl Grill
Institut für Stochastik und Wirtschaftsmathematik
TU Wien

8. Februar 2022
Version 0.2022.1

Inhaltsverzeichnis

Kapitel 1

Einleitung

Als um das Jahr 2011 herum davon die Rede war, dass im Bachelorstudium "Technische Informatik" eine Einführungsvorlesung in Wahrscheinlichkeitstheorie und die Theorie der stochastischen Prozesse geschaffen werden soll, ist für mich ein lang gehegter Wunsch in Erfüllung gegangen — ich hatte schon lange davon geträumt, die Grundideen der Theorie der stochastischen Prozesse auf elementarem Niveau und mit einer Verbindung zur Anwendung zu unterrichten. Ich bedanke mich hier bei allen, die das möglich gemacht haben, ganz besonders bei dem Studienplankoordinator für die Technische Informatik, Ulrich Schmid, und bei meinem Kollegen Norbert Kusolitsch, die auch die ersten Überlegungen dazu angestellt haben, für ihre Unterstützung.

Ich selbst bin nun (Anfang 2018, im sechsten Jahr dieser Vorlesung) immer noch dabei, zu lernen und an der Gewichtung der Inhalte zu arbeiten. Ich danke allen meinen Studentinnen und Studenten für ihre Mitarbeit und für das hilfreiche Feedback, und bei meinen Kollegen, allen voran Zsolt Saffer, für die inhaltlichen und methodischen Diskussionen.

Dieses Skriptum und die Lehrveranstaltung dazu haben eine recht ambitionierte Inhaltsangabe: jedes der vier Kapitel Wahrscheinlichkeitstheorie, stochastische Prozesse, Informationstheorie und Statistik wäre für eine eigene Vorlesung und in einem Ausmaß, das über diese hinausgeht. Es ist also hier notwendig, eine sehr selektive Auswahl des Stoffes zu treffen, und an einigen Stellen habe ich mich gezwungen gesehen, auf eine mathematisch strenge Herleitung der Ergebnisse zu verzichten, um mich mehr mit der Anwendung der jeweiligen Resultate beschäftigen zu können, etwa bei den Gesetzen der großen Zahlen oder beim zentralen Grenzwertsatz. Zu meiner eigenen Beruhigung und für interessierte Leserinnen und Leser werden nach und nach im Anhang über mathematische Hintergründe zumindest skizzenhaft die Details ergänzt werden. Einstweilen (Jänner 2018) ist das noch Zukunftsmusik, momentan kämpfe ich noch damit, dieses Skriptum aus einer Sammlung von Überschriften und Schlagwörtern für diejenigen, die die Vorlesung besucht haben, in einen einigermaßen selbst stehenden Einführungstext zu verwandeln.

Zur Genderfrage: eigentlich habe ich das Glück, in der Mathematik zu arbeiten, die geradezu der Inbegriff von Körper- und damit Geschlechtslosigkeit ist (was manchmal zu recht interessanten Assoziationen führt, wie etwa in Wilhelm Reichs "Charakteranalyse" nachzulesen ist). Dennoch sind in manchen Beispielen, die Bezug auf die reale Welt nehmen, die handelnden Personen eben handelnde Personen, und gelegentlich möchte ich mich auch an Sie, liebe Leserinnen und Leser, wenden. Da bin ich nun hin-und hergerissen zwischen der Überzeugung, dass nur ein konsequenter Gebrauch von weiblicher und männlicher Form nebeneinander einen korrekten Umgang mit dem Problem darstellt (und nicht irgendwelche "I-" und "/"-Konstruktionen; der oft angepriesene Ausweg, nur geschlechtsneutrale Formulierungen zu verwenden, ist für mich nichts anderes als eine Flucht), und dem ästhetisch-pragmatischen Problem, dass Sätze mit vielen "sie und er"-Kombinationen holprig und schwer lesbar sind. Deswegen möchte ich es so halten, dass

grundsätzlich Formulierungen à la “Studentinnen und Studenten” oder “Studentin oder Student” verwendet werden, aber in Beispielen, die sonst zu mühsam zu lesen wären, die Rollen zur Hälfte weiblich und zur Hälfte männlich besetzt werden (wobei im Zweifelsfall die Frauen die besseren Rollen bekommen).

Kapitel 2

Grundlagen der Wahrscheinlichkeitstheorie

2.1 Die Axiome von Kolmogorov

Wir beginnen unsere Erkundung der Welt der Wahrscheinlichkeitstheorie mit einem einfachen Beispiel, dem Werfen von zwei Würfeln. Dabei gibt es insgesamt 36 mögliche Ausgänge — 6 Möglichkeiten für die Augenzahl des ersten Würfels, 6 für die des zweiten. Wenn wir diesen Versuch sehr oft wiederholen, etwa 36000 mal, werden wir feststellen, dass die Häufigkeiten der einzelnen Ausgänge sich nicht allzu sehr von 1000 unterscheiden (mit dem Wissen, das wir in dieser Vorlesung erwerben, würde es uns sehr verwundern, wenn die Abweichung größer als 100 wäre). Die relative Häufigkeit der einzelnen Ausgänge liegt also ungefähr bei 1/36. Die Feststellung, dass sich die relativen Häufigkeiten der einzelnen Versuchsausgänge in einer großen Anzahl von Versuchen bei gewissen Werten, den Wahrscheinlichkeiten, "einpendeln", ist als das "empirische Gesetz der großen Zahlen" bekannt. Dabei handelt es sich zwar nicht um einen mathematischen Satz, und deshalb hat es in einer strengen Behandlung des Themas eigentlich nichts verloren, wir werden es aber gelegentlich benutzen, um gewisse Begriffe zu motivieren oder zu veranschaulichen.

Was wir aus dem empirischen Gesetz der großen Zahlen mitnehmen können, ist, dass Wahrscheinlichkeiten etwas sind, mit dem man rechnen kann wie mit relativen Häufigkeiten. Dazu fassen wir zunächst die möglichen Versuchsausgänge zu einer Menge Ω zusammen, die wir die Grundmenge nennen. In unserem Beispiel ist

$$\Omega = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\} = \{(1,1),(1,2),\ldots,(6,6)\}.$$

Die Elemente von Ω heißen Elementarereignisse. Wir können nun beliebige Ereignisse definieren, etwa "die Summe der Augenzahlen ist 9" oder "der erste Würfel zeigt Augenzahl 5". Wir können solche Ereignisse festlegen, indem wir die Menge der Elementarereignisse angeben, bei denen sie eintreten. Aus diesem Grund werden für uns in Hinkunft Ereignisse einfach Teilmengen der Grundmenge Ω sein.

Die beiden Ereignisse, die wir genannt haben, sind dann

$$A = \{(3,6),(4,5),(5,4),(6,3)\}$$

und

$$B = \{(5,1),(5,2),(5,3),(5,4),(5,5),(5,6)\}.$$

Da wir unsere Ereignisse als Mengen definieren, können wir sie auch mit den Mengenoperationen verknüpfen. Wir können also von den Ereignissen A^C ("A tritt nicht ein"), $A \cap B$ ("A und B treten ein") und $A \cup B$ ("A oder B tritt ein") sprechen — wie es in der Mathematik üblich ist, verstehen wir das "oder" inklusiv. Ein "exklusives Oder" für Ereignisse gibt es allerdings auch, als Mengenoperation heißt es "symmetrische Differenz":

$$A \triangle B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B).$$

Die leere Menge heißt auch das unmögliche Ereignis, die Grundmenge Ω das sichere Ereignis.

Diesen Ereignissen wollen wir Zahlenwerte — Wahrscheinlichkeiten — zuordnen. Diese sollen sich verhalten wie relative Häufigkeiten, insbesondere zwischen 0 und 1 liegen und additiv sein — die Wahrscheinlichkeit der Vereinigung von disjunkten Ereignissen ist die Summe der einzelnen Wahrscheinlichkeiten. Wir verlangen zusätzlich, dass diese Additivität auch für abzählbar unendlich viele Ereignisse gilt:

Definition 2.1: Axiome von Kolmogorov

Ω sei eine beliebige (nichtleere) Menge. Eine Funktion $\mathbb{P}$, die Mengen $A \subseteq \Omega$ reelle Zahlen zuordnet, heißt Wahrscheinlichkeit (oder Wahrscheinlichkeitsmaß), wenn die folgenden Axiome gelten:

1. $0 \leq \mathbb{P}(A) \leq 1$.
2. $\mathbb{P}(\emptyset) = 0$.
3. $\mathbb{P}(\Omega) = 1$.
4. (Additivität) Wenn A und B disjunkte Ereignisse ($A \cap B = \emptyset$) sind, dann gilt
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B).$$
4a. (Abzählbare Additivität, Sigmaadditivität) Wenn $A_n, n \in \mathbb{N}$ disjunkte Ereignisse sind (d.h., $A_i \cap A_j = \emptyset, i \neq j$), dann gilt
$$\mathbb{P}(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n).$$

Das, was den Ansatz von Kolmogorov über frühere Zugänge zur Wahrscheinlichkeitstheorie hinaushebt, ist die abzählbare Additivität. Es sollte leicht zu sehen sein (wenn man alle A_n bis auf zwei gleich der leeren Menge setzt), dass aus ihr die Additivität (die zur Unterscheidung auch manchmal "endliche Additivität" genannt wird) folgt; in der umgekehrten Richtung lässt sich zwar aus der Additivität folgern, dass für jedes $N \in \mathbb{N}$ und disjunkte Ereignisse $A_1, \ldots, A_N$ die Gleichung
$$\mathbb{P}(\bigcup_{n=1}^{N} A_n) = \sum_{n=1}^{N} \mathbb{P}(A_n)$$
gilt, aber der Grenzübergang für $N \to \infty$ ist nicht so ohne weiteres erlaubt. Wir fassen noch einmal zusammen:

Definition 2.2: Wahrscheinlichkeitsraum

Ein Wahrscheinlichkeitsraum $(\Omega, \mathbb{P})$ besteht aus einer Menge Ω, der *Grundmenge* und einem Wahrscheinlichkeitsmaß $\mathbb{P}$.

Die Elemente $\omega \in \Omega$ heißen *Elementarereignisse*, Teilmengen $A \subseteq \Omega$ *Ereignisse*. Das Wahrscheinlichkeitsmaß $\mathbb{P}$ ist auf der Menge dieser Ereignisse definiert (und erfüllt die Axiome von Kolmogorov).

Anmerkungen:

1. In dieser Definition haben wir keine genauen Angaben über den Definitionsbereich von $\mathbb{P}$ gemacht, es ist mit Absicht offen gelassen worden, welche Teilmengen von Ω wir als Ereignisse ansehen wollen bzw. können. Wir hoffen natürlich, dass wir für alle Teilmengen von Ω eine Wahrscheinlichkeit definieren können. Wenn Ω unendlich (überabzählbar) ist, gibt es damit allerdings Probleme. Wir werden in dieser Vorlesung darauf vertrauen, dass alle Mengen, mit denen wir es zu tun haben, brav genug sind, dass sie eine Wahrscheinlichkeit verdienen, und ansonsten nicht weiter darüber nachdenken.

   In der Mathematik wird das Problem so gelöst, dass der Definitionsbereich des Wahrscheinlichkeitsmaßes auf eine Teilmenge der Potenzmenge von Ω eingeschränkt wird (siehe Anhang).

2. Ein simples Beispiel für einen Wahrscheinlichkeitsraum mit (abzählbar) unendlich vielen Elementen ist das Werfen einer Münze, bis zum ersten Mal "Kopf" erscheint. Die Anzahl der Würfe kann jede beliebige positive ganze Zahl sein, also ist $\Omega = \mathbb{N}$ (und $\mathbb{P}(\{n\}) = 2^{-n}$, wenn wir annehmen, dass alle möglichen Münzwurffolgen der Länge n gleiche Wahrscheinlichkeit haben).

3. Ein Beispiel für einen überabzählbaren Wahrscheinlichkeitsraum gibt das Intervall $[0, 1]$, wobei wir $\mathbb{P}([a, b]) = b - a$ setzen (die Wahrscheinlichkeit eines Intervalls ist also gleich seiner Länge). Im nächsten Beispiel werden wir einen Weg sehen, wie man dazu kommt.

4. Die Elementar-"Ereignisse" sind eigentlich keine Ereignisse, weil sie ja *Elemente* und keine *Teilmengen* von Ω sind. So ist etwa für $\Omega = \{1, 2, 3\}$ das Elementarereignis "1" kein Ereignis, um es zu einem Ereignis zu machen, müssen wir Mengenklammern darum setzen: $\{1\}$ ist das Ereignis mit dem einen Element 1. In der Zukunft wird allerdings wenig Anlass bestehen, über Elementarereignisse im eigentlichen Sinn zu sprechen, wir werden uns also die Freiheit nehmen, auch die Mengen der Form $\{\omega\}$ als "Elementarereignisse" zu bezeichnen. Diese kleine Schlamperei können wir uns erlauben, weil stets aus dem Kontext klar sein sollte, in welchem Sinn dieser Begriff zu verstehen ist (gewissermaßen als objektorientierte Überladung). Etwas schöner ist es, die *Elemente* $\omega \in \Omega$ weiterhin als "Elementarereignisse" zu bezeichnen, und für die einelementigen *Mengen* $\{\omega\}$ einen eigenen Namen einzuführen, etwa "Atome".

In vielen Fällen (etwa beim Würfeln) ist es aus Symmetriegründen plausibel, dass alle Elementarereignisse dieselbe Wahrscheinlichkeit haben müssen; wegen der Additivität muss dann die Wahrscheinlichkeit eines Ereignisses proportional zu seiner Mächtigkeit sein, also kommen wir zu der Definition

Definition 2.3: Laplacescher Wahrscheinlichkeitsraum

Wenn Ω endlich ist und

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

(d.h. alle Elementarereignisse haben dieselbe Wahrscheinlichkeit), dann heißt der Wahrscheinlichkeitsraum ein Laplacescher Wahrscheinlichkeitsraum.

Diese Annahme liegt den meisten "Spielzeugen" der Wahrscheinlichkeitstheoretiker zugrunde: wir werden meist annehmen, dass die Seiten eines Würfels, die Seiten einer Münze, die Kugeln in einer Urne... alle dieselbe Wahrscheinlichkeit haben, geworfen bzw. gezogen zu werden. Gelegentlich werden wir diese Eigenschaft hervorheben, indem wir von einem "fairen" Würfel oder einer "fairen" Münze sprechen. Im Beispiel am Anfang unseres Kapitels würden wir also dem Ereignis A ("Augensumme 9") die Wahrscheinlichkeit $4/36 = 1/9$ und dem Ereignis B ("erster Würfel 5") die Wahrscheinlichkeit $6/36 = 1/6$ zuordnen.

Die Definition eines Laplaceschen Wahrscheinlichkeitsraumes ist nur für endliche Mengen Ω sinnvoll. Die Wahrscheinlichkeiten in Anmerkung 3 kann man zwar auch als "gleichmäßig verteilt" ansehen, aber die Analogie zum endlichen Fall ist nicht perfekt: für endliche Laplacesche Räume ergibt sich durch eine umkehrbar eindeutige Abbildung etwa wieder ein Laplacescher Raum, im unendlichen Beispiel stimmt das (etwa mit der Abbildung $x \mapsto x^3$) nicht.

Beispiel 2.1

Bei abzählbar unendlichen Wahrscheinlichkeitsräumen ist das Laplacesche Konzept nicht direkt anwendbar. Hier — und auch in endlichen Räumen kann jede Menge als abzählbare Vereinigung von einpunktigen Mengen dargestellt werden:

$$A = \bigcup_{x \in A} \{x\}$$

und daher

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

Zur Festlegung eines Wahrscheinlichkeitsmaßes müssen wir in diesen "diskreten" Räumen also nur die "Punktwahrscheinlichkeiten" $\mathbb{P}(\{x\})$ angeben. In einem Laplaceschen Raum haben alle diese Wahrscheinlichkeiten denselben Wert p; Ist Ω (abzählbar) unendlich, dann ist das nicht möglich. Aber auch da kann das Konzept des Laplaceschen Raumes nützlich sein: in Anmerkung 2 ist es es plausibel, dass alle 2^n möglichen Folgen von n Münzwürfen gleich wahrscheinlich sind, eine davon — $n-1$ mal "Zahl", dann "Kopf". — ist günstig für das Ereignis "n Würfe bis zum ersten Kopf", also $\mathbb{P}(\{n\}) = 1/2^n$.

Damit können wir die Wahrscheinlichkeiten komplizierterer Ereignisse bestimmen, etwa, dass eine gerade Anzahl von Würfen nötig ist, also

$$A = \{2, 4, 6, , \ldots\} = \{2n, n \in \mathbb{N}\} :$$

$$\mathbb{P}(A) = \sum_{x \in A} 2^{-x} = \sum_{n=1}^{\infty} 2^{-2n} = \frac{1}{3}.$$

Auf der anderen Seite lässt sich das Wahrscheinlichkeitsmaß in Anmerkung 3 als Grenzfall von Laplaceschen Maßen (etwa auf $\Omega_n = \{k2^{-n}, 1 \leq k \leq 2^n\}$ ansehen. Man muss dazu nur das Maß zu einem Maß auf ganz $[0, 1]$ ergänzen, indem das Komplement von Ω_n (und auch jede seiner Teilmengen) Maß 0 erhält. Dann überzeugt man sich leicht, dass etwa die Wahrscheinlichkeit von $[0, x]$ tatsächlich gegen x konvergiert (für $0 \leq x \leq 1$).

Beispiel 2.2

Man kann sich etwa vorstellen, dass ein Mann zu einem zufälligen Zeitpunkt zwischen 8 und 9 in seinem bevorzugten Bar Tabacchi eintrifft. Wenn die Ankunftszeit in Minuten gemessen wird, sollte jede der 60 Minuten gleich wahrscheinlich sein, bei immer genauerer Messung sollte auch jede Sekunde, Zehntel-, Hundertstel-, Millionstel-, ..., Abstrusillionstel-Sekunde dieselbe Wahrscheinlichkeit wie alle anderen haben (die natürlich für kleinere Einheiten auch immer kleiner werden). Im Grenzfall (wenn wir diese Zeiteinheit gegen 0 gehen lassen) ergibt sich so genau das Beispiel aus Anmerkung 3. Lassen wir nun eine Kollegin ebenfalls zufällig zwischen 8 und 9 dort ankommen, und nehmen wir weiter an, dass beide jeweils 10 Minuten bei ihrem Kaffee verbringen. Wir wollen wissen, wie groß die Wahrscheinlichkeit ist, dass beide zusammentreffen. Dazu tragen wir die beiden Ankunftszeiten in einem rechtwinkeligen Koordinatensystem ein (Abb. 2.1).

Die gesuchte Wahrscheinlichkeit bestimmen wir als das Verhältnis der schraffierten Fläche ($|x - y| \leq 10$) zur Gesamtfläche des Quadrats (also 11/36).

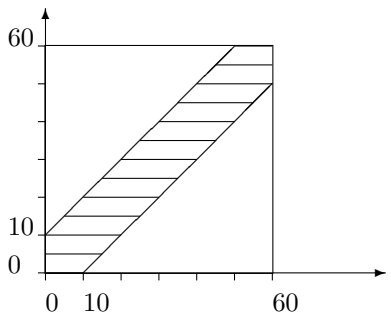

Abbildung 2.1: Geometrische Wahrscheinlichkeit

Solche "geometrischen Wahrscheinlichkeiten" sind in einfachen Fällen wie diesem eine brauch-

bare Veranschaulichung. Die Berechnung der Wahrscheinlichkeit über die Fläche entspricht der Annahme, dass die beiden Ankunftszeiten unabhängig sind — dieser Begriff wird uns im nächsten Abschnitt beschäftigen.

In komplizierteren Fällen ist nicht immer eindeutig klar, was unter einer "gleichmäßigen" Verteilung zu verstehen ist, und unterschiedliche Interpretationen geben unterschiedliche Resultate (Bertrandsches Paradox).

Aus den Axiomen von Kolmogorov ergeben sich einige elementare Folgerungen:

Satz 2.1

1. $\mathbb{P}(A^C) = 1 - \mathbb{P}(A)$.
2. Wenn $A \subseteq B$, dann gilt $\mathbb{P}(A) \leq \mathbb{P}(B)$.
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
4. Für Ereignisse A_n mit $A_n \subseteq A_{n+1}$ gilt
$$\mathbb{P}(\bigcup_n A_n) = \lim_n \mathbb{P}(A_n).$$
5. Für Ereignisse A_n mit $A_n \supseteq A_{n+1}$ gilt
$$\mathbb{P}(\bigcap_n A_n) = \lim_n \mathbb{P}(A_n).$$
6. Für beliebige Ereignisse $A_n, n \in \mathbb{N}$ gilt
$$\mathbb{P}(\bigcup_n A_n) \leq \sum_n \mathbb{P}(A_n).$$

Beweis:

1. folgt aus $\Omega = A \cup A^C$.
2. $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A)$.
3. folgt aus den Gleichungen
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A)$$
und
$$\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A).$$
Insbesondere ist
$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$
4. Mit $B_1 = A_1, B_n = A_n \setminus A_{n-1} (n \geq 2)$ ergibt sich
$$A_n = \bigcup_{i \leq n} B_i,$$
$$\bigcup_n A_n = \bigcup_i B_i,$$
und wegen 2
$$\mathbb{P}(B_n) \leq \mathbb{P}(A_n),$$
also
$$\mathbb{P}(\bigcup_n A_n) = \mathbb{P}(\bigcup_n B_n) = \sum_n \mathbb{P}(B_n) = \lim_{N\to\infty} \sum_{n \leq N} \mathbb{P}(B_n) = \lim_{N\to\infty} \mathbb{P}(\bigcup_{n \leq N} B_n) = \lim_{N\to\infty} \mathbb{P}(A_n).$$

5. folgt aus 1 und 4.

6. Wiederholte Anwendung der Folgerung zu Punkt 3 liefert

$$\mathbb{P}(\bigcup_{n\leq N} A_n) \leq \sum \mathbb{P}(A_n),$$

und wegen 4 kann man hier N gegen ∞ gehen lassen.

Punkt 3 des letzten Satzes lässt sich verallgemeinern, etwa ergibt sich für die Vereinigung von 3 Ereignissen

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C).$$

Allgemein gilt

Satz 2.2: Additionstheorem

$A_1, \ldots A_n$ seien beliebige Ereignisse. Dann gilt

$$\mathbb{P}(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} (-1)^{i-1} S_i$$

mit

$$S_i = \sum_{1\leq j_1<j_2<\ldots<j_i\leq n} \mathbb{P}(A_{j_1} \cap \ldots \cap A_{j_i}).$$

Es werden also alle Wahrscheinlichkeiten summiert, davon wird die Summe aller Wahrscheinlichkeiten von Durchschnitten von zwei Mengen abgezogen, dann kommen die dreifachen Durchschnitte dazu, die vierfachen weg, bis schließlich der Durchschnitt aller n Mengen dazugezählt (n ungerade) bzw. abgezogen (n gerade) wird.

Beispiel 2.3

Als Anwendung dieses Satzes berechnen wir die Wahrscheinlichkeit, dass eine zufällig gewählte Permutation von n Elementen keinen Fixpunkt hat. Diese Frage wird gern in der Form präsentiert, dass eine Anzahl (10) von Ehepaaren sich zu einer Tanzveranstaltung treffen, und nach einiger Zeit, in der nur die Ehepartner miteinander tanzen, beschließen, die Tanzpartner durch das Los zu bestimmen. In dieser Einkleidung wird unsere Frage zu der nach der Wahrscheinlichkeit, dass kein Ehepaar miteinander tanzt.

Wir betrachten das Gegenereignis A, dass mindestens ein Fixpunkt existiert. Dieses können wir wieder als Vereinigung der Ereignisse A_i, dass i ein Fixpunkt ist, schreiben, also

$$\mathbb{P}(A) = \mathbb{P}(\bigcup_{i=1}^{n} A_i).$$

Diese Wahrscheinlichkeit bestimmen wir mit dem Additionstheorem. Dazu müssen wir die Summen S_k berechnen. Das Ereignis $A_{i_1} \cap \ldots \cap A_{i_k}$ tritt ein, wenn $i_1, \ldots i_k$ Fixpunkte sind, die anderen $n-k$ Elemente können beliebig vertauscht werden. Das gibt $(n-k)!$ günstige Möglichkeiten von insgesamt $n!$ und somit eine Wahrscheinlichkeit $(n-k)!/n!$. Es gibt $\binom{n}{k}$ solcher Summanden, also

$$S_k = \binom{n}{k} \frac{(n-k)!}{n!} = \frac{1}{k!}.$$

Insgesamt ist

$$\mathbb{P}(A) = \sum_{k=1}^{n} (-1)^{k-1} \frac{1}{k!}$$

und die Wahrscheinlichkeit, die wir suchen,

$$\mathbb{P}(A^C) = 1 - \mathbb{P}(A) = \sum_{k=0}^{n} (-1)^k \frac{1}{k!}.$$

Für großes n ist das näherungsweise $1/e$. Die Näherung ist so gut, dass sich die Anzahl der Permutationen ohne Fixpunkt (für $n \geq 1$) bestimmen lässt, indem man $n!/e$ auf die nächste ganze Zahl rundet.

2.2 Bedingte Wahrscheinlichkeiten

Definition 2.4

A und B seien zwei Ereignisse mit $\mathbb{P}(B) > 0$. Dann heißt

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

die bedingte Wahrscheinlichkeit von A unter (der Bedingung) B.

Zur Motivation dieser Definition geben wir vor, an das empirische Gesetz der großen Zahlen zu glauben. Unter N Versuchen sind dann etwa $N\mathbb{P}(B)$ Versuche, bei denen B eintritt. Die Information, dass B eingetreten ist, sagt uns jetzt, dass unser Versuch zu diesen $N\mathbb{P}(B)$ gehört. Von diesen sind wiederum etwa $N\mathbb{P}(A \cap B)$ solche, bei denen auch A eintritt. Die relative Häufigkeit von allen Experimenten, bei denen A beobachtet wird, unter denen, bei denen B eintritt ist also ungefähr

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

und so ergibt sich die Formel in unserer Definition.

Die Definition kann man ausmultiplizieren und erhält

$$\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A|B).$$

In dieser Formulierung ist es nicht mehr nötig, anzunehmen, dass $\mathbb{P}(B) > 0$ gilt, für $\mathbb{P}(B) = 0$ ist die Gleichung sogar immer richtig, egal, was wir für die bedingte Wahrscheinlichkeit einsetzen. Um etwas Sinnvolles damit anzufangen, muss die bedingte Wahrscheinlichkeit auf anderem Weg bestimmt werden, so, wie es etwa in den Urnenbeispielen weiter unten passiert, oder vorgegeben sein, wie es im Kapitel über Markovketten der Fall ist.

Mehrfache Anwendung dieser Formel liefert den

Satz 2.3: Multiplikationssatz

$A_1, \ldots A_n$ seien Ereignisse mit $\mathbb{P}(A_1 \cap \ldots \cap A_{n-1}) > 0$. Dann gilt

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

Beispiel 2.4: Ziehen ohne Zurücklegen

Ein typisches Beispiel ist das Ziehen ohne Zurücklegen: Es seien etwa in einer Urne zwei schwarze und drei weiße Kugeln. Es wird dreimal ohne Zurücklegen gezogen, und wir wollen die Wahrscheinlichkeit bestimmen, dass alle gezogenen Kugeln weiß sind. Wir setzen also A_i gleich dem Ereignis, dass die i-te gezogene Kugel weiß ist, und suchen $\mathbb{P}(A_1 \cap A_2 \cap A_3)$. Am Anfang sind in der Urne fünf Kugeln, drei davon sind weiß, also

$$\mathbb{P}(A_1) = \frac{3}{5}.$$

Nach der ersten Ziehung (mit Ergebnis A_1) sind noch vier Kugeln in der Urne, davon sind zwei weiß, also

$$\mathbb{P}(A_2|A_1) = \frac{2}{4}.$$

Schließlich sind nach den ersten beiden Ziehungen noch eine weiße und zwei schwarze Kugeln in der Urne, und die Wahrscheinlichkeit, nochmals weiß zu ziehen, ist

$$\mathbb{P}(A_3|A_1 \cap A_2) = \frac{1}{3}.$$

Insgesamt ergibt sich

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \frac{3 \cdot 2 \cdot 1}{5 \cdot 4 \cdot 3} = \frac{1}{10}.$$

Wenn wir nach der Wahrscheinlichkeit fragen, dass zwei weiße kugeln unter den drei gezogenen sind, dann stellen wir zuerst fest, dass dieses Ereignis auf drei Arten eintreten kann — die schwarze Kugel kann die erste, zweite oder dritte gezogene sein. Das ergibt

$$\mathbb{P}(\text{“2 weiße”}) = \mathbb{P}(A_1^C \cap A_2 \cap A_3) + \mathbb{P}(A_1 \cap A_2^C \cap A_3) + \mathbb{P}(A_1 \cap A_2 \cap A_3^C) =$$

$$\frac{2 \cdot 3 \cdot 2 + 3 \cdot 2 \cdot 2 + 3 \cdot 2 \cdot 2}{5 \cdot 4 \cdot 3} = \frac{3}{5}.$$

Es fällt auf, dass die drei Summanden den gleichen Wert haben — die Faktoren treten nur in unterschiedlicher Reihenfolge auf.

In einer allgemeinen Formulierung — in der Urne sind N Kugeln, davon sind A weiß, n werden gezogen, x gezogene sind weiß — ist die Situation genauso. Die weißen Kugeln können auf $\binom{n}{x}$ Arten auf die n Ziehungen verteilt werden, und in jeder dieser Wahrscheinlichkeiten treten dieselben Faktoren auf, es ergibt sich also

$$\mathbb{P}(x \text{ weiße}) = \binom{n}{x} \frac{A \cdot (A-1) \cdots (A-x+1)(N-A) \cdot (N-A-1) \cdots (N-A-n+x+1)}{N \cdot (N-1) \cdots (N-n+1)} =$$

$$\frac{\binom{A}{x}\binom{N-A}{n-x}}{\binom{N}{n}}. \tag{2.1}$$

Die letzte Gleichung lässt sich etwas einfacher so ableiten: aus Symmetriegründen sind alle $\binom{N}{n}$ Auswahlen von n gezogenen Kugeln aus den N vorhandenen gleich wahrscheinlich; günstig sind die, bei denen aus den A weißen Kugeln genau x Kugeln gezogen werden und die übrigen $n-x$ aus den $N-A$ schwarzen. Das geht auf $\binom{A}{x}\binom{N-A}{n-x}$ Arten. Eigentlich müssten wir hier noch Grenzen für x vorgeben (in unserem Beispiel mit $N = 5$, $A = 3$ und $n = 3$ muss $1 \leq x \leq 3$ gelten), aber wir können darauf verzichten, wenn wir

$$\binom{n}{k} = 0$$

setzen, wenn $k < 0$ oder $k > n$ ist.

Im Multiplikationssatz ist, wie wir schon festgestellt haben, es nicht zwingend notwendig, dass die bedingenden Ereignisse positive Wahrscheinlichkeiten haben. Es muss dann natürlich die bedingte Wahrscheinlichkeit auf einem anderen Weg gefunden werden, etwa so wie in diesem Beispiel:

Beispiel 2.5

Ein Würfel wird einmal geworfen. Anschließend werden eine schwarze Kugel und so viele weiße Kugeln, wie die Augenzahl des Würfels angibt, in eine Urne gelegt. Aus dieser Urne wird dann eine Kugel zufällig gezogen. Wir definieren die Ereignisse A_n:“es wird eine n gewürfelt” und B:“die gezogene Kugel ist schwarz”. Dann ist

$$\mathbb{P}(B|A_n) = \frac{1}{n+1}$$

(es sind ja n weiße und eine schwarze Kugel in der Urne), und

$$\mathbb{P}(A_n \cap B) = \mathbb{P}(A_n)\mathbb{P}(B|A_n) = \frac{1}{6(n+1)}$$

für $n = 1, \ldots, 6$. Wir können aber auch ein $n > 6$ einsetzen. Dann ist nämlich $\mathbb{P}(A_n) = 0$, und damit auch $\mathbb{P}(A_n \cap B) = 0$ und

$$\mathbb{P}(A_n)\mathbb{P}(B|A_n) = \mathbb{P}(A_n)\frac{1}{n+1} = 0.$$

Es gilt also auch im Fall $\mathbb{P}(A) = 0$

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A).$$

Für die Gültigkeit dieser Gleichung ist sogar unwichtig, welchen Wert wir der bedingten Wahrscheinlichkeit geben. Das ist hier nicht besonders wichtig, im Kapitel über Markovketten werden wir diese Tatsache (meist stillschweigend) verwenden. Insbesondere enthebt sie uns der Notwendigkeit, im Satz von der vollständigen Wahrscheinlichkeit darauf zu achten, dass alle Ereignisse B_i positive Wahrscheinlichkeit haben (immer vorausgesetzt, dass wir einen alternativen Weg haben, zu den bedingten Wahrscheinlichkeiten zu kommen).

Wenn die bedingte Wahrscheinlichkeit $\mathbb{P}(A|B)$ gleich der unbedingten $\mathbb{P}(A)$ ist, wenn also das Wissen um das Eintreten von B unsere Einschätzung der Wahrscheinlichkeit von A nicht ändert, werden wir sagen, dass A von B unabhängig ist. Die entsprechende Gleichung können wir ausmultiplizieren, dadurch können wir auf die Forderung $\mathbb{P}(B) > 0$ verzichten und sehen, dass die Rollen von A und B symmetrisch sind:

Definition 2.5: Unabhängigkeit

Zwei Ereignisse A und B heißen unabhängig, wenn

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Die Ereignisse $A_1, \ldots, A_n$ heißen unabhängig, wenn für alle $k \leq n$ und $1 \leq i_1 < i_2 \ldots < i_k \leq n$ gilt

$$\mathbb{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \ldots \mathbb{P}(A_{i_k}).$$

Die Ereignisse $A_1, \ldots, A_n$ heißen paarweise unabhängig, wenn für alle $1 \leq i < j \leq n$

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j).$$

Unendlich viele Ereignisse nennen wir unabhängig, wenn jede endliche Auswahl daraus unabhängig ist.

In unserem Urnenbeispiel sind die Ereignisse A_i ("die i-te Kugel ist weiß") nicht unabhängig. Um das für A_1 und A_2 zu verifizieren, bestimmen wir die Wahrscheinlichkeit von A_2. Das kann mit einem Symmetrieargument geschehen (auch beim zweiten, dritten,... Zug muss jede Kugel mit gleicher Wahrscheinlichkeit — $1/5$ — gezogen werden) geschehen, andererseits können wir die Wahrscheinlichkeit von A_2 auch so erhalten:

$$\mathbb{P}(A_2) = \mathbb{P}(\Omega \cap A_2) = \mathbb{P}((A_1 \cup A_1^C) \cap A_2) = \mathbb{P}(A_1 \cap A_2) + \mathbb{P}(A_1^C \cap A_2) =$$

$$\mathbb{P}(A_1)\mathbb{P}(A_2|A_1) + \mathbb{P}(A_1^C)\mathbb{P}(A_2|A_1^C) = \frac{3 \cdot 2 + 2 \cdot 3}{5 \cdot 4} = \frac{3}{5}.$$

Damit sehen wir

$$\mathbb{P}(A_1 \cap A_2) = 0.3 \neq \mathbb{P}(A_1)\mathbb{P}(A_2) = 0.36,$$

und A_1 und A_2 sind nicht unabhängig.

Das Vorgehen, das wir zur Berechnung von $\mathbb{P}(A_2)$ verwendet haben, lässt sich auch in allgemeiner Form anwenden: wir nehmen an, dass (endlich oder abzählbar viele) Ereignisse B_i gegeben

sind, von denen genau eines eintritt, und dass wir sowohl die Wahrscheinlichkeiten der Ereignisse B_i und die bedingten Wahrscheinlichkeiten eines weiteren Ereignisses A bezüglich jedes dieser Ereignisse kennen (oder leicht berechnen können). Dann gilt

Satz 2.4: Satz von der vollständigen Wahrscheinlichkeit

B_i seien disjunkte Ereignisse mit $\mathbb{P}(B_i) > 0$ und $\bigcup_i B_i = \Omega$ und A ein beliebiges Ereignis. Dann gilt

$$\mathbb{P}(A) = \sum_i \mathbb{P}(B_i)\mathbb{P}(A|B_i).$$

Diese Berechnungen lassen sich (wenn es nicht zu viele Möglichkeiten gibt) mit einem Baum graphisch darstellen (Abb. 2.2): An den Kanten stehen die bedingten Wahrscheinlichkeiten, an den Knoten die Wahrscheinlichkeiten der einzelnen Zweige, die sich als Produkt der Werte an den Kanten (von der Wurzel ausgehend) berechnen. Die Wahrscheinlichkeit des gesuchten Ereignisses A ergibt sich als die Summe der Wahrscheinlichkeiten aller Endknoten, die zu A gehören.

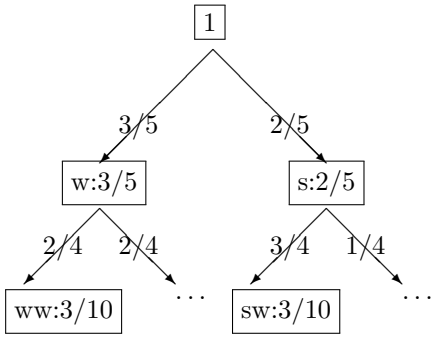

Abbildung 2.2: Vollständige Wahrscheinlichkeit

Wir können dann umgekehrt die bedingte Wahrscheinlichkeit ausrechnen, dass eines der Ereignisse B_i eingetreten ist, wenn A beobachtet wurde:

Satz 2.5: Satz von Bayes

Unter denselben Voraussetzungen wie im vorigen Satz gilt (wenn der Nenner positiv ist)

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(B_j)\mathbb{P}(A|B_j)}{\sum_i \mathbb{P}(B_i)\mathbb{P}(A|B_i)}.$$

Beispiel 2.6: Blutgruppen

Wir wenden diesen Satz auf eine Frage an, die mich einige Zeit beschäftigt hat: ich habe lange Zeit meine Blutgruppe nicht gekannt, aber die meiner Frau (A) und die meines Sohnes (0). Wir wollen also nun die bedingte Wahrscheinlichkeit für die möglichen Ausprägungen meiner Blutgruppe bestimmen. Dazu müssen wir zuerst einige Informationen zu den Blutgruppen einholen: wie jede genetisch bedingte Eigenschaft wird auch die Blutgruppe durch zwei Gene (eins vom Vater, eins von der Mutter) bestimmt. Diese können die Ausprägungen a, b und o besitzen. Bei einem zufällig gewählten Menschen nehmen die beiden Gene unabhängig voneinander die einzelnen Werte mit Wahrscheinlichkeiten p_a, p_b und p_0 an. Die Kombinationen aa, ao und oa resultieren in Blutgruppe A, bb, $b0$ und $0b$ ergeben B, ab und ba AB und schließlich oo 0. Pschyrembels medizinisches Wörterbuch liefert, dass in Europa 47% der Bevölkerung Blutgruppe A haben, 9% B, 4% AB und 40% 0. Daraus ergibt sich (ungefähr, die Gleichungen sind überbestimmt und nicht exakt zu lösen)

$$p_a = 0.300, p_b = 0.067, p_0 = 0.633.$$

Aus den Informationen über die Blutgruppen meiner Frau und meines Sohnes können wir folgern, dass mein Sohn von mir ein Gen o haben muss. Meine Ausstattung muss also entweder oa/ao (wir bezeichnen dieses Ereignis mit A_a), ob/bo (Ereignis A_b) oder oo (Ereignis A_0) sein. Die a-priori-Wahrscheinlichkeiten dafür sind $2p_ap_0$, $2p_bp_0$ und p_0^2. In den ersten beiden Fällen ist die bedingte Wahrscheinlichkeit für die Weitergabe einer o (Ereignis B) 1/2, im letzten 1.

Jetzt haben wir alles beisammen, was in den Satz von Bayes einzusetzen ist, und erhalten

$$\mathbb{P}(A_a|B) = \frac{\frac{1}{2}2p_ap_0}{\frac{1}{2}2p_ap_0 + \frac{1}{2}2p_bp_0 + 1p_0^2} = .300$$

und analog

$$\mathbb{P}(A_b|B) = .067$$

und

$$\mathbb{P}(A_0|B) = .633.$$

Wir würden also am ehesten darauf wetten, dass ich Blutgruppe 0 habe. Inzwischen habe ich mich natürlich pieksen lassen (und für die Bestimmung geblecht), und es wäre schön zu berichten, dass unsere Analyse uns zur korrekten Vermutung geführt hat, aber solche Bilderbuchenden gibt es nicht immer — ich darf mich über Blutgruppe A freuen.

2.3 Zufallsvariable

Eine Zufallsvariable ist im wesentlichen eine zufälliger Zahlenwert (mit dem man rechnen kann). Formal heißt das, mit einer beliebigen Grundmenge Ω und einem Wahrscheinlichkeitsmaß $\mathbb{P}$ darauf:

Definition 2.6

Eine Zufallsvariable X ist eine Abbildung von Ω nach $\mathbb{R}^d$.

Bemerkungen:

1. Wenn Ω überabzählbar ist, muss man von X eine zusätzliche Eigenschaft verlangen, die Messbarkeit (Anhang).
2. Meistens werden wir $d = 1$ haben, also reellwertige Zufallsvariable. Im Fall $d > 1$ haben wir $X = (X_1, \ldots, X_d)$, also einen Vektor von reellen Zufallsvariablen. Zur Unterscheidung gegenüber dem Fall $d = 1$ spricht man auch gerne von einem Zufallsvektor.

Ein zentraler Begriff ist die Verteilung einer Zufallsvariable:

Definition 2.7

Die Verteilung einer Zufallsvariable ist das Wahrscheinlichkeitsmaß

$$\mathbb{P}_X(A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega : X(\omega) \in A\})(A \subseteq \mathbb{R}^d).$$

Definition 2.8

Wenn der Wertebereich von X endlich oder höchstens abzählbar ist, dann nennen wir X diskret.

In diesem Fall kann die Verteilung von X durch die Wahrscheinlichkeitsfunktion

$$p_X(x) = \mathbb{P}(X = x) = \mathbb{P}_X(\{x\})$$

angegeben werden.

Wir können etwa das Urnenbeispiel aus dem vorigen Abschnitt unter diesem Gesichtspunkt betrachten, indem wir eine Zufallsvariable X definieren, die die Anzahl der weißen Kugeln unter den 3 gezogenen ist. Wir haben schon die Wahrscheinlichkeiten

$$p_X(3) = \frac{1}{10}, p_X(2) = \frac{3}{5},$$

die wir nach der allgemeinen Formel (2.1) durch

$$p_X(1) = \frac{3}{10}, p_X(0) = 0$$

ergänzen können. Diese Wahrscheinlichkeitsfunktion ist in Abbildung 2.3 graphisch dargestellt.

Die allgemeine Form (2.1) ist die erste Verteilung, für die wir einen Namen haben:

Definition 2.9

Die hypergeometrische Verteilung $H(N, A, n)$ $(n, A, N \in \mathbb{N},\ 0 \leq n, A \leq N)$ hat die Wahrscheinlichkeitsfunktion

$$p(x) = \frac{\binom{A}{x}\binom{N-A}{n-x}}{\binom{N}{n}}.$$

Diese Verteilung tritt auf, wenn aus einer Grundgesamtheit mit N Elementen, von denen A "günstig" sind, n ohne Zurücklegen gezogen werden, und X die Anzahl der "günstigen" Elemente unter den gezogenen ist.

Wird mit Zurücklegen gezogen, dann sind die einzelnen Ziehungen unabhängig voneinander mit Wahrscheinlichkeit $p = A/N$ "günstig" ("Erfolge"); wir können etwas allgemeiner den Fall betrachten, dass n unabhängige Versuche gemacht werden, die jeweils mit Wahrscheinlichkeit p (die nicht rational sein muss) einen Erfolg ergeben, und X die Anzahl der Erfolge in diesen n Versuchen ist. Das führt zu

Definition 2.10

Die Binomialverteilung $B(n, p)$:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Einen besonders einfachen Fall erhalten wir, wenn wir in der Binomialverteilung (oder in der hypergeometrischen Verteilung) $n = 1$ setzen:

Definition 2.11

X heißt alternativverteilt $(X \sim A(p))$ wenn

$$\mathbb{P}(X = 1) = p, \mathbb{P}(X = 0) = 1 - p.$$

Eine alternativverteilte Zufallsvariable lässt sich als *Indikatorvariable* darstellen:

Definition 2.12

Für ein Ereignis A heißt die Funktion

$$I_A : \Omega \to \{0, 1\}$$

mit

$$I_A(\omega) = \begin{cases} 1 & \text{wenn } \omega \in A, \\ 0 & \text{wenn } \omega \notin A \end{cases}$$

Indikator von A.

Für allgemeinere Fälle (nicht diskrete Zufallsvariable) ist die Wahrscheinlichkeitsfunktion nicht brauchbar. Immer funktioniert die Verteilungsfunktion:

Definition 2.13

Die Verteilungsfunktion von X ist gegeben durch

$$F_X(x) = \mathbb{P}_X((-\infty, x]) = \mathbb{P}(X \leq x)$$

(wenn X d-dimensional ist, ist auch x d-dimensional, und die Ungleichung ist komponentenweise zu verstehen, also $X \leq x$ wenn $X_i \leq x_i$ für alle $i = 1, \ldots, d$ und $(-\infty, x] = (-\infty, x_1] \times \ldots \times (-\infty, x_d]$). Damit ist

$$F_X(x) = F_{(X_1,\ldots,X_d)}(x_1, \ldots, x_d) = \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d).$$

Für $d = 1$ kann man die Verteilungsfunktionen einfach charakterisieren:

Satz 2.6

$F : \mathbb{R} \to \mathbb{R}$ ist genau dann eine Verteilungsfunktion, wenn

1. $0 \leq F(x) \leq 1$ für alle x,
2. F ist monoton nichtfallend,
3. F ist rechtsstetig,
4. $\lim_{x\to-\infty} F(x) = 0$,
5. $\lim_{x\to\infty} F(x) = 1$.

Zur Veranschaulichung ist in Abbildung 2.3 neben der Wahrscheinlichkeitsfunktion die Verteilungsfunktion der hypergeometrischen Verteilung graphisch dargestellt. Sie zeigt die typische Form der Verteilungsfunktion einer diskreten Zufallsvariable: die Funktion hat im Punkt x einen Sprung der Höhe $p(x)$ und ist zwischen den Sprungstellen konstant.

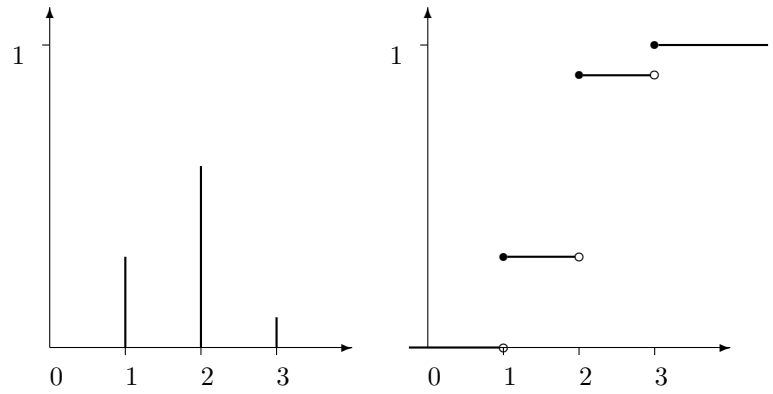

Abbildung 2.3: Wahrscheinlichkeits- und Verteilungsfunktion der hypergeometrischen Verteilung $H_{5,3,3}$

Mithilfe der Verteilungsfunktion kann man Wahrscheinlichkeiten berechnen:

$$\mathbb{P}(X \leq a) = F_X(a),$$

$$\mathbb{P}(X < a) = F_X(a - 0),$$

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a),$$
$$\mathbb{P}(a < X < b) = F_X(b-0) - F_X(a),$$
$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a-0),$$
$$\mathbb{P}(a \leq X < b) = F_X(b-0) - F_X(a-0).$$
$$\mathbb{P}(X = a) = F_X(a) - F_X(a-0).$$

Dabei ist $F(x-0) = \lim_{h \downarrow 0} F(x-h)$ der linksseitige Grenzwert von F in x.

Mehrdimensionale Verteilungsfunktionen haben zusätzliche Eigenschaften, wir betrachten hier nur den Fall $d = 2$:

Satz 2.7

$F : \mathbb{R}^2 \to \mathbb{R}$ ist genau dann eine Verteilungsfunktion, wenn

1. $0 \leq F(x_1, x_2) \leq 1$ für alle x_1, x_2,
2. F ist monoton nichtfallend in jeder Argumentvariable,
3. F ist rechtsstetig,
4. $\lim_{x_1 \to -\infty} F(x_1, x_2) = \lim_{x_2 \to -\infty} F(x_1, x_2) = 0$,
5. $\lim_{x_1, x_2 \to \infty} F(x_1, x_2) = 1$,
6. Für $a_1 < b_1, a_2 < b_2$ gilt

$$F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2) \geq 0.$$

Diese Eigenschaften sind analog zu denen von eindimensionalen Verteilungsfunktionen, nur die letzte ist neu. Der Ausdruck der dort steht, ist genau die Wahrscheinlichkeit

$$\mathbb{P}(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2).$$

Im diskreten Fall (und für $d = 1$) hat F_X Sprünge der Höhe $p_X(x)$ an den Punkten x, die mit positiver Wahrscheinlichkeit angenommen werden, und ist dazwischen konstant. Wenn F (stückweise stetig) differenzierbar ist, dann können wir F durch die Ableitung festlegen:

Definition 2.14

Wenn F_X in der Form

$$F_X(x) = \int_{-\infty}^{x} f_X(u) du$$

bzw.

$$F_X(x) = \int_{-\infty}^{x_d} \dots \int_{-\infty}^{x_1} f_X(u_1, \dots, u_d) du_1 \dots du_d$$

(falls $X = (X_1, \dots, X_d)$ mehrdimensional ist) darstellbar ist, dann heißt f_X die Dichte der Verteilung von X, und wir nennen X stetig (verteilt).

Wahrscheinlichkeits- und Dichtefunktionen sind leicht zu charakterisieren:

Satz 2.8

1. Die Funktion p ist genau dann eine Wahrscheinlichkeitsfunktion, wenn

$$p(x) \geq 0 \text{ für alle } x \in \mathbb{R}^d \text{ und } \sum_x p(x) = 1.$$

2. Die Funktion f ist genau dann eine Dichtefunktion, wenn

$$f(x) \geq 0 \text{ für alle } x \in \mathbb{R}^d \text{ und } \int f(x)dx = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1 \dots, x_d)dx_1 \dots dx_d = 1.$$

Beispiele für stetige Verteilungen:

- Die stetige Gleichverteilung $U(a, b)$:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{wenn } a \leq x \leq b, \\ 0 & \text{sonst.} \end{cases}$$

  Diese Verteilung (mit $a = 0, b = 1$) ist uns schon als Beispiel für einen Wahrscheinlichkeitsraum mit unendlich vielen Elementen begegnet. Ihre Dichte und Verteilungsfunktion sind in Abbildung 2.3 dargestellt.

- Die Exponentialverteilung $Ex(\lambda)$:

$$f(x) = \lambda e^{-\lambda x}[x \geq 0].$$

- Die Normalverteilung (eine der wichtigsten Verteilungen) $N(\mu, \sigma^2)$:

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

  Der Spezialfall $N(0, 1)$ (also $\mu = 0, \sigma^2 = 1$) wird als Standardnormalverteilung bezeichnet, für ihre Dichte bzw. Verteilungsfunktion haben sich die Bezeichnungen φ und Φ eingebürgert, also

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \Phi(x) = \int_{-\infty}^{x} \phi(u)du.$$

  Die Dichte der Normalverteilung kann nicht geschlossen integriert werden. Für die Berechnung der Verteilungsfunktion werden Näherungsformeln bzw. Tabellen herangezogen. Eine Tabelle für die Verteilungsfunktion $\Phi(x)$ der Standardnormalverteilung finden Sie (neben anderen) in Anhang A.

  Diese Tabelle ist so organisiert: das Argument von Φ wird auf zwei Stellen gerundet (wenn man nicht interpolieren möchte); die Zahl mit der ersten Nachkommastelle bestimmt die Zeile, die zweite Nachkommastelle die Spalte. $\Phi(1.54) = 0.938$ finden wir etwa in der Zeile "1.5" in Spalte "4".

Durch Integrieren der Dichte erhält man für eine Zufallsvariable mit einer Normalverteilung $N(\mu, \sigma^2)$ die Verteilungsfunktion

$$F_X(x) = \Phi(\frac{x-\mu}{\sigma}).$$

Wir berechnen als Beispiel die Wahrscheinlichkeit, dass eine Zufallsvariable X mit Verteilung $N(4, 25)$ Werte zwischen 2 und 7 annimmt:

$$\mathbb{P}(2 \leq x \leq 7) = \Phi(\frac{7-4}{5}) - \Phi(\frac{2-4}{5}) = \Phi(0.6) - \Phi(-0.4) = 0.726 - (1 - 0.655) = 0.381.$$

Es kann sein, dass eine Zufallsvariable sowohl diskrete als auch stetige Anteile hat:

Definition 2.15

Wenn F_X sowohl Sprünge als auch eine nichtverschwindende Ableitung hat, dann nennen wir X gemischt verteilt. In diesem Fall gibt es sowohl eine Wahrscheinlichkeitsfunktion als auch eine Dichte.

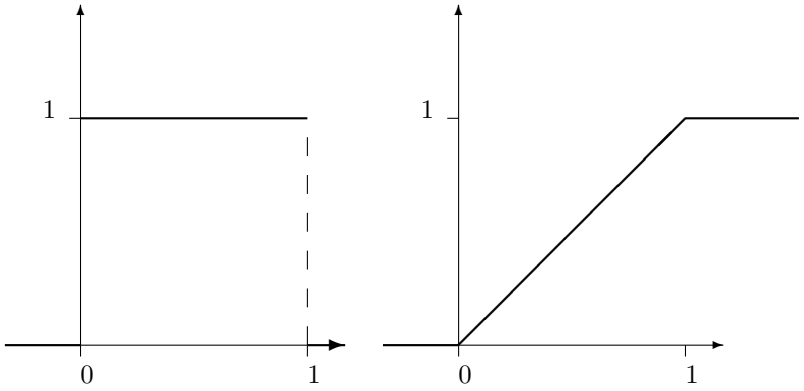

Abbildung 2.4: Dichte- und Verteilungsfunktion der stetigen Gleichverteilung $U(0,1)$

Anmerkung: Die Wahrscheinlichkeitsfunktion und Dichte einer gemischten Verteilung sind unvollständig: die Summe bzw. das Integral dieser Funktionen sind jeweils kleiner als 1, ihre Summe muss natürlich 1 ergeben. Manchmal werden diese unvollständigen Funktionen als modifizierte Dichte bzw. Wahrscheinlichkeitsfunktion bezeichnet.

Ein typisches Beispiel für eine gemischte Verteilung ist die Wartezeit bei einer Ampel (bei der wir zufällig eintreffen): mit positiver Wahrscheinlichkeit ist die Ampel grün und die Wartezeit 0; wenn gewartet werden muss, ist die Wartezeit stetig gleichverteilt (Abb. 2.5).

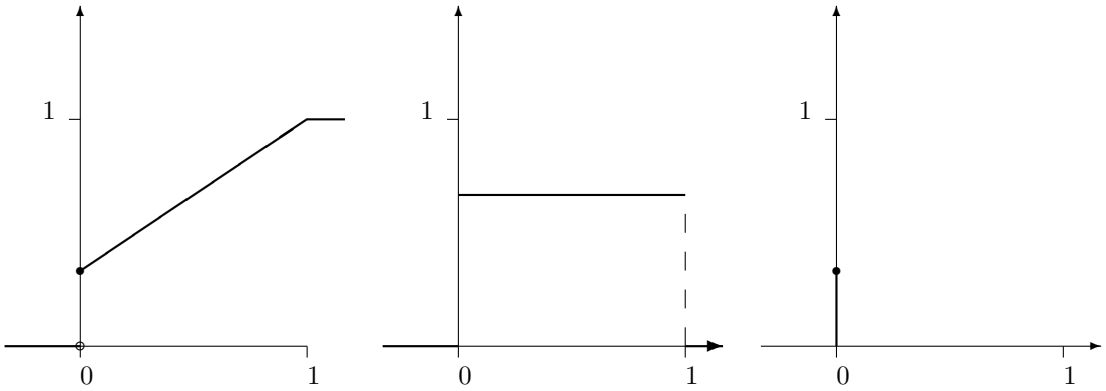

Abbildung 2.5: Verteilungs-, Dichte- und Wahrscheinlichkeitsfunktion einer gemischten Verteilung

Mithilfe der Wahrscheinlichkeits- bzw. Dichtefunktion kann die Wahrscheinlichkeit, dass eine Zufallsvariable X Werte in der Menge A annimmt, als Summe bzw. Integral dargestellt werden:

Für diskretes X:

$$\mathbb{P}(X \in A) = \sum_{x \in A} p_X(x),$$

für stetiges X:

$$\mathbb{P}(X \in A) = \int_A f_X(x)dx,$$

bzw. im mehrdimensionalen Fall

$$\mathbb{P}(X \in A) = \int_A f_X(x_1, \ldots, x_d)dx_1 \ldots dx_d,$$

und für gemischte Verteilungen

$$\mathbb{P}(X \in A) = \sum_{x \in A} p_X(x) + \int_A f_X(x)dx.$$

Wenn X und Y eine gemeinsame Verteilung mit der Dichte $f_{X,Y}(x,y)$ haben, dann können wir die Verteilungsfunktion von X wie folgt berechnen:

$$F_X(x) = \mathbb{P}(X \leq x, -\infty < Y < \infty) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy dx,$$

und durch Differenzieren ergibt sich die Dichte von X (und analog für Y) als

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy, f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx.$$

Für diskrete Verteilungen gilt

$$p_X(x) = \sum_y p_{X,Y}(x,y), p_Y(y) = \sum_x p_{X,Y}(x,y).$$

Wenn die Verteilung von X (oder Y) in dieser Weise aus der gemeinsamen Verteilung erhalten wird, nennt man sie auch die Randverteilung (im stetigen Fall heißt die Dichte dieser Verteilung Randdichte). Diese Bezeichnung kommt daher, dass man für diskrete Variable die gemeinsame Verteilung in Form einer Tabelle aufschreiben kann, die man durch eine zusätzliche Zeile und Spalte für die Summen ergänzt.

Als Beispiel würfeln wir dreimal, X soll die Anzahl der Einsen sein, Y die Anzahl der Sechsen (Abb.2.6).

	Y				
X	0	1	2	3	p_X
0	$\frac{64}{216}$	$\frac{48}{216}$	$\frac{12}{216}$	$\frac{1}{216}$	$\frac{125}{216}$
1	$\frac{48}{216}$	$\frac{24}{216}$	$\frac{3}{216}$	0	$\frac{75}{216}$
2	$\frac{12}{216}$	$\frac{3}{216}$	0	0	$\frac{15}{216}$
3	$\frac{1}{216}$	0	0	0	$\frac{1}{216}$
p_Y	$\frac{125}{216}$	$\frac{75}{216}$	$\frac{15}{216}$	$\frac{1}{216}$	1

Abbildung 2.6: Eine mehrdimensionale diskrete Verteilung, rechts und unten die beiden Randverteilungen

Diese Verteilung ist ein Spezialfall der einzigen mehrdimensionalen diskreten Verteilung, für die wir einen Namen haben:

Definition 2.16: Multinomialverteilung

Die Multinomialverteilung $M(n, p_1, \ldots, p_k)$ ($k \geq 2, p_i \geq 0, \sum_{i=1}^k p_i = 1$) ist die k-dimensionale Verteilung mit der Wahrscheinlichkeitsfunktion

$$p_{(X_1,\ldots,X_k)}(x_1, \ldots, x_k) =$$

$$\mathbb{P}(X_1 = x_1, \ldots, X_k = x_k) = \begin{cases} \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k} & \text{wenn } x_i \geq 0, \sum_{i=1}^k x_i = n, \\ 0 & \text{sonst.} \end{cases}$$

Diese Verteilung tritt auf, wenn n unabhängige Versuche angestellt werden, die jeweils mit Wahrscheinlichkeit p_i den Ausgang i ($i = 1, \ldots k$) liefern, und X_i die Anzahl der Ausgänge i in diesen n Versuchen ist. Die eindimensionalen Randverteilungen dieser Verteilung sind natürlich Binomialverteilungen.

Ein wichtiger Begriff ist die Unabhängigkeit von Zufallsvariablen:

Definition 2.17

Die Zufallsvariablen $(X_1, \ldots, X_n)$ heißen unabhängig, wenn für alle $x_1, \ldots x_n$

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i).$$

Die unendliche Folge $(X_n, n \in \mathbb{N})$ heißt unabhängig, wenn jede endliche Teilfolge unabhängig ist.

Wenn die gemeinsame Verteilung diskret bzw. stetig ist, kann man in dieser Definition die Verteilungsfunktion durch die Wahrscheinlichkeits- bzw. Dichtefunktion ersetzen.

Wenn wir von Unabhängigkeit sprechen, können wir auch über bedingte Wahrscheinlichkeiten nachdenken, die wir in diesem Fall als "bedingte Verteilung" bezeichnen. Im diskreten Fall kann man etwa die bedingte Wahrscheinlichkeit

$$\mathbb{P}(X \leq a | Y = y)$$

nach der üblichen Formel berechnen. Für stetige Verteilungen macht dieser Ausdruck vordergründig keinen Sinn, weil das bedingende Ereignis Wahrscheinlichkeit 0 hat. Wir können aber versuchen, diese Wahrscheinlichkeit als Grenzwert von

$$\mathbb{P}(X \leq a | y - \epsilon \leq Y \leq y + \epsilon)$$

für $\epsilon \to 0$ zu berechnen. Das funktioniert auch, und führt uns zu

Definition 2.18

X, Y seien stetig verteilt mit Dichte $f_{X,Y}$. Die bedingte Dichte von X unter $Y = y$ ist

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

Damit erhalten wir die bedingte Wahrscheinlichkeit als

$$\mathbb{P}(X \leq a | Y = y) = \int_{-\infty}^{a} f_X(x|Y = y) dx.$$

Ein Beispiel:

Beispiel 2.7: Zweidimensionale Dichte

X und Y haben eine gemeinsame Verteilung mit der Dichte

$$f_{X,Y}(x, y) = cxy[0 \leq x \leq y \leq 1].$$

Wir wollen c, die beiden Randdichten und die Wahrscheinlichkeiten

$$\mathbb{P}(X + Y < 1)$$

und

$$\mathbb{P}(0.2 < X < 0.4 | Y = 0.6)$$

bestimmen.

Für die Bestimmung von c verwenden wir, dass das Integral der Dichte 1 sein muss:

$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = \int_0^1 \int_0^y cxy\, dx\, dy = \int_0^1 \frac{cy^3}{2} dy = \frac{c}{8},$$

also $c = 8$.

Randdichte von Y:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_0^y 8xy\,dx = 4y^3$$

(das haben wir eigentlich schon in der Zeile darüber als inneres Integral berechnet).

Randdichte von X:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_x^1 8xy\,dy = 4x(1-x^2).$$

Die Wahrscheinlichkeit von $[X+Y<1]$ ergibt sich als

$$\mathbb{P}(X+Y<1) = \int\int_{[x+y<1]} f(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{1-y} f(x,y)dxdy.$$

$f(x,y)$ ist nur für $0 \le x \le y \le 1$ von 0 verschieden. Also

$$\mathbb{P}(X+Y<1) = \int_0^1 \int_0^{\min(y,1-y)} 8xy\,dx\,dy = \int_0^{1/2}\int_0^y 8xy\,dx\,dy + \int_{1/2}^1\int_0^{1-y} 8xy\,dx\,dy =$$

$$= \frac{1}{16} + \frac{5}{48} = \frac{1}{6}.$$

Die bedingte Dichte macht nur für $0 < y \le 1$ Sinn und ist dort

$$f_X(x|Y=y) = \frac{f(x,y)}{f_Y(y)} = \frac{8xy[0\le x\le y]}{4y^3} = \frac{2x[0\le x\le y]}{y^2}.$$

Damit ist

$$\mathbb{P}(0.2 \le X \le 0.4 | Y = 0.6) = \int_{0.2}^{0.4} f_X(x|Y=0.6)\,dx = \int_{0.2}^{0.4} \frac{2x}{0.36}\,dx = \frac{1}{3}.$$

2.4 Transformationen

Wir betrachten jetzt den Fall, dass die Verteilung einer Zufallsvariable X bekannt ist und wir die Verteilung der transformierten Zufallsvariable

$$Y = g(X)$$

bestimmen wollen, wobei g irgendeine Funktion (im einfachsten Fall $g : \mathbb{R} \to \mathbb{R}$ ist. Falls X diskret ist, dann hat auch Y eine diskrete Verteilung. In diesem Fall gilt klarerweise

Satz 2.9

Wenn X eine diskrete Verteilung mit Wahrscheinlichkeitsfunktion p_X hat und $Y = g(X)$ gilt, dann ist

$$p_Y(y) = \mathbb{P}(Y=y) = \mathbb{P}(g(X)=y) = \sum_{x:g(x)=y} p_X(x).$$

Beispiel 2.8

Zur Illustration verwenden wir ein Beispiel, das mit unseren späteren Diskussionen von Erwartungswert und Varianz im Zusammenhang steht:

X ist binomialverteilt mit Parametern $n = 6$ und $p = 1/3$. Wir fragen nach der Verteilung von $Y = (X-2)^2$. X hat die möglichen Werte $0, 1, \ldots, 6$, daher gibt es für Y die möglichen

Werte $0, 1, 4, 9, 16$, und wir erhalten die Wahrscheinlichkeitsfunktion p_Y als

$$p_Y(0) = \mathbb{P}(Y = 0) = \mathbb{P}(X = 2) = \binom{6}{2}(\frac{1}{3})^2(\frac{2}{3})^4 = 15\frac{16}{729} = \frac{240}{729},$$

$$p_Y(1) = \mathbb{P}(Y = 1) = \mathbb{P}((X-2)^2 = 1) = \mathbb{P}(X = 1) + \mathbb{P}(X = 3) =$$
$$\binom{6}{1}(\frac{1}{3})^1(\frac{2}{3})^5 + \binom{6}{3}(\frac{1}{3})^3(\frac{2}{3})^3 = \frac{352}{729},$$

$$p_Y(4) = \mathbb{P}(X = 0) + \mathbb{P}(X = 4) = \frac{124}{729},$$

$$p_Y(9) = \mathbb{P}(X = 5) = \frac{12}{729},$$

$$p_Y(16) = \mathbb{P}(X = 6) = \frac{1}{729}.$$

Wenn X stetig verteilt ist, dann kann $Y = g(X)$ immer noch diskret verteilt sein (wenn g nur endlich viele Werte annimmt, etwa falls g die Vorzeichenfunktion ist). Wir sind allerdings besonders an dem Fall interessiert, dass Y ebenfalls eine stetige Verteilung besitzt. In einfachen Fällen können wir diese Verteilung direkt erhalten:

Beispiel 2.9: Quadrat einer Normalverteilung

Wir nehmen an, dass X eine Standardnormalverteilung hat, und wollen die Verteilung von $Y = X^2$ bestimmen. Wir beginnen mit der Verteilungsfunktion von Y. Für $y \geq 0$ gilt

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = 2\Phi(\sqrt{y}) - 1.$$

Klarerweise ist $F_Y(y) = 0$ für $y < 0$.

Durch Ableiten der Verteilungsfunktion erhalten wir die Dichte

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi y}} e^{-y/2} & \text{für } y > 0. \\ 0 & \text{sonst.} \end{cases}$$

Diese Verteilung ist ein Spezialfall der Gammaverteilung $\Gamma(\alpha, \lambda)$ mit der Dichte

$$f(x) = \frac{x^{\alpha-1}\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x}[x > 0],$$

wobei Γ die Gammafunktion

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$$

ist. Diese erfüllt die Gleichung

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$$

und insbesondere

$$\Gamma(n) = (n-1)!(n \in \mathbb{N}).$$

Die Verteilung von Y ist eine Gammaverteilung mit $\alpha = \lambda = 1/2$.

In diesem Fall konnten wir unser Wissen über die spezielle Funktion $g(x) = x^2$ ausnützen. Für allgemeine Funktionen g kann die Bestimmung der Verteilung von $Y = g(X)$ kompliziert sein, weil die Menge aller x mit $g(x) \leq y$ bestimmt werden muss. Dies ist relativ einfach, wenn wir annehmen, dass g stetig differenzierbar und eindeutig umkehrbar, also streng monoton ist. Dann gibt es die beiden Möglichkeiten, dass g entweder steigend oder fallend sein kann. Wir führen die Rechnung für den etwas komplizierteren zweiten Fall durch.

Das Bild der reellen Zahlen unter g ist

$$g(\mathbb{R}) = \{g(x) : x \in \mathbb{R}\} = (a, b)$$

mit

$$a = \lim_{x \to \infty} g(x)$$

und

$$b = \lim_{x \to -\infty} g(x).$$

Für $a < y < b$ gilt

$$F_Y(y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)).$$

Ableiten gibt die Dichte:

$$f_Y(y) = \begin{cases} -(g^{-1})'(y) f_X(g^{-1}(y)) = |(g^{-1})'(y)| f_X(g^{-1}(y)) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|} & \text{wenn } x \in g(\mathbb{R}) \\ 0 & \text{sonst.} \end{cases}$$

Falls g steigend ist, dann ergibt sich dieselbe Darstellung.

Dieses Ergebnis gilt auch in mehreren Dimensionen:

Satz 2.10: Transformationssatz für Dichten

$X = (X_1, \ldots, X_n)$ sei stetig verteilt mit der Dichte f_X. $g : \mathbb{R}^n \to \mathbb{R}^n$ sei stetig differenzierbar und eindeutig umkehrbar. $Y = g(X)$ (d.h. $Y_i = g_i(X_1, \ldots, X_n)$) ist dann ebenfalls stetig verteilt mit der Dichte

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) |\frac{\partial g^{-1}}{\partial y}(y)| = f_X(g^{-1}(y)) \frac{1}{|\frac{\partial g}{\partial x}(g^{-1}(y))|} & \text{wenn } y \in g(\mathbb{R}^n), \\ 0 & \text{sonst.} \end{cases}$$

Dabei ist

$$\frac{\partial g}{\partial x} = \det((\frac{\partial g_i}{\partial x_j})_{n \times n})$$

die Funktionaldeterminante.

Damit können wir die Verteilung einer Summe von zwei unabhängigen Zufallsvariablen X und Y bestimmen: mit dem Transformationssatz kann die gemeinsame Dichte von X und $X+Y$ bestimmt werden, und die Verteilung von $X + Y$ als Randverteilung davon:

Wir setzen also

$$g(x, y) = (g_1(x, y), g_2(x, y) = (x, x + y).$$

Diese Transformation hat die Funktionaldeterminante

$$\begin{vmatrix} \frac{\partial g_1}{\partial x} & \frac{\partial g_1}{\partial y} \\ \frac{\partial g_2}{\partial x} & \frac{\partial g_2}{\partial y} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1$$

und die Umkehrfunktion

$$g^{-1}(u, v) = (u, v - u).$$

Weil X und Y unabhängig sind, ergibt sich ihre gemeinsame Dichte als Produkt:

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

also ergibt sich die gemeinsame Dichte von $U = X, Z = X + Y$ als

$$f_{Z,U}(z, u) = f_X(u) f_Y(z - u),$$

und durch Integrieren ergibt sich die Randdichte von Z:

Satz 2.11

X und Y seien unabhängig mit Dichte f_X und f_Y. Dann ist die Dichte von $X + Y$

$$f_{X+Y}(z) = f_X * f_Y(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx.$$

Definition 2.19

f und g seien zwei Dichten. Die Dichte $f * g$ mit

$$f * g(z) = \int_{-\infty}^{\infty} f(x)g(z-x)dx.$$

heißt die Faltung von f und g.

Die Dichte der Summe von zwei stetigen Zufallsvariablen ist also die Faltung der Dichten der beiden Summanden.

Beispiel 2.10: Summe von exponentialverteilten Zufallsvariablen

Als Beispiel wollen wir die Faltung zweier Exponentialdichten mit Parameter λ ($f(x) = \lambda e^{-\lambda x}[x \geq 0]$) bestimmen. Für $z < 0$ ist $f * f(z) = 0$, für $z \geq 0$ erhalten wir

$$f * f(z) = \int_{-\infty}^{\infty} f(z-x)f(x)dx = \int_0^z \lambda e^{-\lambda(z-x)}\lambda e^{-\lambda x}dx = \int_0^z \lambda^2 e^{-\lambda z}dx = z\lambda^2 e^{-\lambda z}.$$

Auch diese Verteilung ist eine Gammaverteilung (mit $\alpha = 2$); auch die Exponentialverteilung ist ein Spezialfall der Gammaverteilung für $\alpha = 1$. In ähnlicher Weise ergibt sich die Dichte von X/Y: wir setzen $u = y, z = x/y$. Die Umkehrfunktion $y = u, x = zu$ hat die Funktionaldeterminante $-u$, und wir erhalten

Satz 2.12

X und Y seien unabhängig mit Dichte f_X und f_Y. Dann hat $Z = X/Y$ die Dichte

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(y)f_X(zy)|y|dy.$$

Beispiel 2.11: Quotient von exponentialverteilten Zufallsvariablen

Setzen wir wieder für X und Y exponentialverteilte Zufallsvariable ein, dann ergibt sich (für $z \geq 0$)

$$f_{X/Y}(z) = \int_0^{\infty} \lambda^2 y e^{-y(1+z)}dy = \frac{1}{(1+z)^2}.$$

Beispiel 2.12: Gemeinsame Verteilung von Summe und Produkt

Wir können auch die gemeinsame Verteilung von $Q = X/Y$ und $S = X + Y$ bestimmen. Die Umkehrfunktion zu

$$q = x/y, s = x + y$$

ist

$$x = \frac{qs}{1+q}, y = \frac{s}{1+q}.$$

Die Funktionaldeterminante dazu ist

$$\frac{\partial(x,y)}{\partial(q,s)} = \begin{vmatrix} \frac{s}{(1+q)^2} & \frac{q}{1+q} \\ -\frac{s}{(1+q)^2} & \frac{1}{1+q} \end{vmatrix} = \frac{s}{(1+q)^2}.$$

Insgesamt ist

$$f_{Q,S}(q,s) = \frac{\lambda^2 s e^{-\lambda s}}{(1+q)^2}[s \geq 0][q \geq 0].$$

Diese gemeinsame Dichte ist das Produkt der beiden Dichten, die wir in den vorangegangenen Beispielen berechnet haben. S und Q sind also unabhängig. Für diese Erkenntnis hätten wir die Randdichten gar nicht berechnen müssen, es genügt die Erkenntnis, dass die gemeinsame Dichte sich in zwei Faktoren zerlegen lässt, von denen einer nur von s und der andere nur von q abhängt.

Es gilt allgemeiner: wenn $X \sim \Gamma(\alpha, \lambda)$ und $Y \sim \Gamma(\beta, \lambda)$ unabhängig sind, dann sind $S = X + Y$ und $Q = X/Y$ unabhängig, $S \sim \Gamma(\alpha + \beta, \lambda)$, und Q hat eine Betaverteilung 2. Art ($B_2(\alpha, \beta)$) mit der Dichte

$$f_{\alpha,\beta}(x) = \frac{x^{\alpha-1}}{B(\alpha,\beta)(1+x)^{\alpha+\beta}}[x > 0].$$

Wird statt durch Y durch die Summe $X + Y$ dividiert, dann ergibt sich für $X/(X + Y)$ eine Betaverteilung 1. Art ($B_1(\alpha, \beta)$) mit der Dichte

$$f_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}[0 < x < 1].$$

In den letzten beiden Gleichungen ist B die Betafunktion

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Eine ähnliche Formel wie die Faltung zweier Dichten gibt es auch für die Summe von zwei diskreten Zufallsvariablen:

Satz 2.13

X und Y seien unabhängig mit Wahrscheinlichkeitsfunktion p_X und p_Y. Dann ist die Wahrscheinlichkeitsfunktion von $X + Y$ die (diskrete) Faltung von p_X und p_Y:

$$p_{X+Y}(z) = p_X * p_Y(z) = \sum_x p_X(x)p_Y(z-x).$$

Wir nehmen jetzt an, dass die Verteilungsfunktion F von X stetig und streng monoton ist (und daher umkehrbar), und betrachten

$$Y = F(X).$$

Die Verteilungsfunktion von Y berechnet sich als

$$\mathbb{P}(Y \le y) = \mathbb{P}(F(X) \le y) = \mathbb{P}(X \le F^{-1}(y)) = F(F^{-1}(y)) = y,$$

Y ist also gleichverteilt auf $[0, 1]$. Umgekehrt ist $F^{-1}(Y)$ nach F verteilt, wenn Y auf $[0, 1]$ gleichverteilt ist. Dieses Ergebnis gilt auch für allgemeine Verteilungen, allerdings muss man dazu die Inverse neu definieren: es kann ja einerseits die Gleichung

$$F(x) = y$$

gar keine Lösung x haben, und andererseits kann es mehrere Lösungen x geben. Der erste Fall tritt auf, wenn F unstetig ist, also Sprünge hat, und wenn y innerhalb eines Sprunges liegt. Mit anderen Worten, es gibt einen Punkt x, für den der linksseitige Grenzwert $F(x - 0)$ kleiner als y ist und der Funktionswert $F(x)$ größer als y. In diesem Fall ist dieses x die natürliche Wahl für $F^{-1}(y)$. Wenn es mehrere x mit $F(x) = y$ gibt, dann bilden alle diese x ein Intervall. Ein möglicher Zugang besteht darin, alle diese x gleichberechtigt zu behandeln. Das führt zur Definition der *Quantile*.

Definition 2.20: Quantil

x heißt p-Quantil $(0 > p > 1)$ der Verteilung mit der Verteilungsfunktion F, wenn

$$F(x-0) \le p \le F(x)$$

(d.h., wenn $X \sim F$, dann ist $\mathbb{P}(X < x) \leq p \leq \mathbb{P}(X \leq x)$).

Für das p-Quantil wird gerne die Notation x_p verwendet. Dass dieses nicht eindeutig bestimmt ist, stört oft nicht. Muss man die Definition eindeutig machen, gibt es im wesentlichen drei Möglichkeiten, das p-Quantil zu definieren: der Mittelpunkt des Intervalls, in dem $F(x) = p$ gilt, oder einer der Endpunkte. Die erste Wahl hat etwas Salomonisches und wird gelegentlich in der beschreibenden Statistik verwendet. Mathematisch ergiebiger sind die Endpunkte, und zu unserer Definition der Verteilungsfunktion ($F_X(x) = \mathbb{P}(X \leq x)$, man könnte auch $\mathbb{P}(X < x)$ verwenden) passt am besten der linke Endpunkt. Formal gibt das

Definition 2.21

Die verallgemeinerte Inverse der Verteilungsfunktion F ist gegeben durch

$$F^{-1}(y) = \inf\{x : F(x) \geq y\}.$$

Satz 2.14

Wenn U auf $[0, 1]$ gleichverteilt ist, dann hat

$$X = F^{-1}(U)$$

die Verteilungsfunktion F.

Dieser Satz gibt uns die wichtigste Methode zur Erzeugung von Zufallszahlen mit beliebiger Verteilung, vorausgesetzt, dass wir einen Zufallsgenerator haben, der unabhängig gleichverteilte Zufallszahlen liefert. Wenn wir etwa exponentialverteilte Zufallszahlen erzeugen wollen, dann können wir die Verteilungsfunktion

$$F(x) = 1 - e^{-\lambda x}$$

invertieren

$$F^{-1}(y) = -\frac{1}{\lambda}\log(1 - y).$$

Wir haben also mit

$$X = -\frac{1}{\lambda}\log(1 - U)$$

die gesuchte exponentialverteilte Zufallszahl. Diese Formel lässt sich noch etwas vereinfachen, wenn man bedenkt, dass $1 - U$ so wie U auf $[0, 1]$ gleichverteilt ist, also funktioniert auch

$$X = -\frac{1}{\lambda}\log(U).$$

Die Subtraktion, die man sich dabei erspart, fällt zwar vom Aufwand her gegenüber dem Logarithmus und der Division nicht wirklich ins Gewicht, aber es gibt halt Prinzipien...

Etwas, das man bei der Anwendung der Inversenmethode beachten sollte, ist, dass zwar die theoretische Gleichverteilung einzelne Werte mit Wahrscheinlichkeit 0 annimmt, die (eigentlich diskreten, aber mit sehr vielen möglichen Werten) Zufallszahlen aus dem Generator aber nicht, und es kann sein, dass die Werte 0 und/oder 1 tatsächlich auftreten, was bei der Inversion zu unliebsamen Reaktionen der Laufzeitbibliothek führen kann. Weil so etwas nur etwa ein Mal in 10^9 Versuchen auftritt, ist das einer der Fehler, die oft erst beim "Gammatesten" erkannt werden.

2.5 Erwartungswert und Varianz

Auch hier wollen wir uns zur Motivation den frequentistischen Standpunkt zu eigen machen: stellen wir uns vor, dass wir 6000000 mal würfeln und den Mittelwert

$$\bar{X} = \frac{1}{6000000}\sum_{i=1}^{6000000} X_i$$

der Augenzahlen berechnen. In dieser "Stichprobe" wird jede Augenzahl etwa 1000000 mal vorkommen, also gilt

$$\bar{X} \approx \frac{1}{6000000}(1000000 \cdot 1 + \ldots + 1000000 \cdot 6) = \frac{1}{6} \cdot 1 + \ldots \frac{1}{6} \cdot 6 = 3.5.$$

Die rechte Seite ist unschwer als die Summe aus den Produkten der einzelnen Werte mit ihren Wahrscheinlichkeiten zu erkennen. Als Frequentisten glauben wir natürlich daran, dass diese Gleichung für $n \to \infty$ exakt wird. Wenn wir wieder zu Sinnen kommen, können wir das als Definition niederschreiben (und dezent verschweigen, wie wir dazu gekommen sind):

Definition 2.22

Der Erwartungswert einer Zufallsvariable X ist

$$\mathbb{E}(X) = \sum_x x p_X(x)$$

für diskrete Zufallsvariable, und

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

für stetige Zufallsvariable. Falls X gemischt verteilt ist, gilt

$$\mathbb{E}(X) = \sum_x x p_X(x) + \int_{-\infty}^{\infty} x f_X(x) dx.$$

Die Definition für den stetigen Fall lässt sich entweder ohne großes Nachdenken durch die formale Ableitung "ersetze Summen durch Integrale und die Wahrscheinlichkeitsfunktion durch die Dichtefunktion", oder, indem man die Zufallsvariable X auf eine gewisse Anzahl von Stellen rundet. Die gerundete Variable hat eine diskrete Verteilung, und ihr Erwartungswert konvergiert gegen das Integral $\int x f(x) dx$, wenn die Anzahl der Stellen gegen unendlich geht.

Die Summen und Integrale in dieser Definition sind so zu verstehen, dass sie zuerst in die Teile $x > 0$ und $x < 0$ zerlegt werden; jeder Teil kann einen endlichen oder unendlichen Wert ergeben, und das Endergebnis ist die Summe dieser beiden; wenn beide Summanden unendlich sind, also eine unbestimmte Form $\infty - \infty$ auftritt, dann ist das Ergebnis nicht definiert, und der Erwartungswert existiert nicht. In allen anderen Fällen hat der Erwartungswert einen bestimmten Wert, der auch $+\infty$ oder $-\infty$ sein kann.

Beispiel 2.13

Wir berechnen als Beispiel den Erwartungswert der Binomialverteilung:

$$\mathbb{E}(X) = \sum_{i=0}^{n} i \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=1}^{n} i \binom{n}{i} p^i (1-p)^{n-i} =$$

$$= \sum_{i=1}^{n} n \binom{n-1}{i-1} p^i (1-p)^{n-i}|_{i=j+1} = \sum_{j=0}^{n-1} n \binom{n-1}{j} p^{j+1} (1-p)^{n-1-j} =$$

$$np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-1-j} = np.$$

Beispiel 2.14

Wenn speziell $X \sim B(6, 1/3)$, dann ergibt sich $\mathbb{E}(X) = 2$.

Wir gehen einen Schritt weiter und berechnen den Erwartungswert von $Y = (X - \mathbb{E}(X))^2 =$

$(X-2)^2$, den wir später die Varianz von X nennen werden. Die Verteilung von Y haben wir schon in Beispiel 2.8 berechnet, und damit ergibt sich:

$$\mathbb{E}((X-\mathbb{E}(X))^2) = \mathbb{E}(Y) = 0\frac{240}{729} + 1\frac{352}{729} + 4\frac{124}{729} + 9\frac{12}{729} + 16\frac{1}{729} = \frac{4}{3}.$$

Das stimmt beruhigend mit dem Wert $6 \cdot \frac{1}{3}(1-\frac{1}{3})$ überein, den wir weiter unten auf einem leicht unterschiedlichen (aber natürlich äquivalenten) Weg erhalten werden.

Beispiel 2.15

Die stetige Gleichverteilung $U(a,b)$ hat Erwartungswert

$$\int_a^b \frac{x}{b-a} dx = \frac{a+b}{2}.$$

Beispiel 2.16

Für die Cauchyverteilung mit der Dichte

$$f(x) = \frac{1}{\pi(1+x^2)}$$

existiert der Erwartungswert nicht, weil

$$\int_0^\infty x f(x) dx = \infty$$

und

$$\int_{-\infty}^0 x f(x) dx = -\infty$$

gilt.

Satz 2.15: Eigenschaften des Erwartungswerts

1. Linearität: $\mathbb{E}(aX+b) = a\mathbb{E}(X) + b$,
2. Additivität: $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$,
3. Monotonie: wenn $X \leq Y$, dann ist auch $\mathbb{E}(X) \leq \mathbb{E}(Y)$,
4. wenn X und Y unabhängig sind, dann gilt $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Die Beweise für diese Behauptungen werden sehr viel einfacher, wenn wir zuerst die folgende Aussage für den Erwartungswert einer Funktion von X zeigen:

Satz 2.16: Satz vom unachtsamen Statistiker

X sei eine Zufallsvariable, $g : \mathbb{R} \to \mathbb{R}$ eine Funktion und$Y = g(X)$. Dann ist

1. Wenn X diskret verteilt ist mit Wahrscheinlichkeitsfunktion p_X,

$$\mathbb{E}(Y) = \sum_x g(x) p_X(x).$$

2. Wenn X stetig verteilt ist mit Dichte f_X,

$$\mathbb{E}(Y) = \int g(x) f_X(x) dx.$$

Diese Formel gilt sinngemäß auch für mehrdimensionales $X \in \mathbb{R}^k$ und $g : \mathbb{R}^k \to \mathbb{R}$:

$$\mathbb{E}(Y) = \int g(x) f_X(x) dx = \int \dots \int g(x_1, \dots, x_k) f_{X_1,\dots,X_k}(x_1, \dots, x_k) dx_1 \dots dx_k.$$

3. Wenn X gemischt verteilt ist mit Wahrscheinlichkeitsfunktion p_X und Dichte f_X,

$$\mathbb{E}(Y) = \sum_x g(x) p_X(x) + \int g(x) f_X(x) dx.$$

Wir müssen also nicht erst die Verteilung von Y bestimmen, um den Erwartungswert zu berechnen, sondern können in der Formel für $\mathbb{E}(X)$ einfach x durch $g(x)$ ersetzen.

Wir beweisen diesen Satz für den Fall, dass X diskret ist. Offensichtlich gilt

$$p_Y(y) = \mathbb{P}(Y = y) = \sum_{x : f(x) = y} p_X(x),$$

damit ist

$$\mathbb{E}(Y) = \sum_y y p_Y(y) = \sum_y \sum_{x : f(x) = y} y p_X(x) = \sum_y \sum_{x : f(x) = y} f(x) p_X(x) = \sum_x f(x) p_X(x).$$

Dieser Satz ist hilfreich beim Beweis der Eigenschaften des Erwartungswerts. Die Additivität ergibt sich etwa so (wieder für diskrete Verteilungen):

$$\mathbb{E}(X + Y) = \sum_{x,y} (x + y) \mathbb{1}_{X,Y}(x, y) = \sum_x x \sum_y p_{X,Y}(x, y) + \sum_y y \sum_x p_{X,Y}(x, y) =$$

$$= \sum_x x p_X(x) + \sum_y p_Y(y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Definition 2.23

Die Varianz von X ist

$$\mathbb{V}(X) = \mathbb{E}((X - \mathbb{E}(X))^2).$$

Satz 2.17: Steinerscher Verschiebungssatz

Für beliebiges reelles a gilt

$$\mathbb{E}((X - a)^2) = \mathbb{V}(X) + (\mathbb{E}(X) - a)^2.$$

Satz 2.18: Eigenschaften der Varianz

1. $\mathbb{V}(X) \geq 0$,
2. $\mathbb{V}(X) = 0$ genau dann, wenn $\mathbb{P}(X = \mathbb{E}(X)) = 1$,
3. $\mathbb{V}(aX + b) = a^2 \mathbb{V}(X)$,
4. wenn X und Y unabhängig sind, dann ist $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$.

Diese Behauptungen sind nicht schwer nachzurechnen und bleiben der Leserin oder dem Leser überlassen.

Wenn man im Steinerschen Verschiebungssatz $a = 0$ setzt, dann ergibt sich

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2,$$

eine Formel, die wir immer wieder gerne verwenden, um Varianzen zu berechnen.

Aus dem Steinerschen Verschiebungssatz ist unmittelbar zu erkennen, dass $\mathbb{E}((X-a)^2)$ minimal wird, wenn $a = \mathbb{E}(X)$ ist. Das kann man auch so betrachten: wir suchen eine Zahl a, die als "Vertreter" oder "typischer Wert" für die Zufallsvariable X fungieren soll. Diese eine Zahl a soll X "möglichst gut" wiedergeben. Damit man mit dieser Idee arbeiten kann, braucht man eine Definition dessen, was unter "möglichst gut" zu verstehen ist. Ein möglicher Weg besteht darin, den Abstand zwischen dem tatsächlichen Wert X und dem Vertreter a zu messen, etwa durch den Abstand $|X-a|$ (den "absoluten Fehler") oder durch sein Quadrat $(X-a)^2$ (den "quadratischen Fehler"). Das sind wieder Zufallsvariable, und wir nehmen davon den Erwartungswert, um einen Zahlenwert zu bekommen, der die Güte der Anpassung wiedergibt. Wir kommen so zum "mittleren absoluten Fehler" (mean absolute deviation, MAD) $\mathbb{E}(|X-a|)$ und zum "mittleren quadratischen Fehler" (mean square error, MSE) $\mathbb{E}((X-a)^2)$. Der mittlere absolute Fehler erscheint etwas natürlicher, dafür ist die mathematische Behandlung einfacher. Unsere Überlegungen können wir dann auch so interpretieren, dass der Erwartungswert der beste Vertreter für X ist, wenn der mittlere quadratische Fehler als Gütekriterium verwendet wird (wobei natürlich kleinere Werte besser sind).

Zahlenwerte wie der Erwartungswert, die angeben, wie groß die Werte, die X annimmt, typischerweise sind, heißen in der Wahrscheinlichkeitstheorie und speziell in der Statistik "Lageparameter" und stellen Variationen des Themas "Mittelwert" dar. Dem gegenüber stehen Konzepte wie die Varianz und die mittlere absolute Abweichung, die die typische Abweichung der einzelnen Werte voneinander angeben, diese werden als "Streuungsmaße" oder "Streuungsparameter" bezeichnet.

Die Eigenschaften von Erwartungswert und Varianz können verwendet werden, um manche Berechnungen einfacher zu machen. Wir beginnen als Beispiel mit der Berechnung der Varianz der Binomialverteilung.

Beispiel 2.17

Es sei also X wieder binomialverteilt, $X \sim B(n,p)$. Wir wissen bereits

$$\mathbb{E}(X) = np,$$

also fehlt uns nur noch $\mathbb{E}(X^2)$:

$$\mathbb{E}(X^2) = \sum_{i=0}^{n} i^2 \binom{n}{i} p^i (1-p)^{n-i} = np \sum_{i=1}^{n} i \binom{n-1}{i-1} p^{i-1} (1-p)^{n-1-(i-1)} =$$

$$np \sum_{j=0}^{n-1} (j+1) \binom{n-1}{j} p^j (1-p)^{n-1-j} = np((n-1)p+1) = n^2p^2 + np - np^2.$$

Also

$$\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = np - np^2 = np(1-p).$$

Beispiel 2.18: Methode der Indikatorvariablen

Erwartungswert und Varianz der Binomialverteilung lassen sich auch auf einem anderen Weg finden, dazu erinnern wir uns daran, wie wir zu dieser Verteilung gekommen sind: es gibt n unabhängige Ereignisse $A_1, \ldots, A_n$ und X zählt, wie viele dieser Ereignisse eintreten. Wir

führen jetzt die Indikatorvariablen

$$Y_i = I_{A_i} = \begin{cases} 1 & \text{wenn } A_i \text{ eintritt,} \\ 0 & \text{wenn } A_i \text{ nicht eintritt} \end{cases}$$

ein. $Y_1, \ldots, Y_n$ sind unabhängig mit $\mathbb{P}(Y_i = 1) = p$, $\mathbb{P}(Y_i = 0) = 1 - p$ und daher

$$\mathbb{E}(Y_i) = (Y_i^2) = p$$

und

$$\mathbb{V}(Y_i) = p - p^2 = p(1-p).$$

Wir können X durch die Y_i darstellen:

$$X = \sum_{i=1}^{n} Y_i.$$

Die Rechenregeln für Erwartungswert und Varianz liefern

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(Y_i) = np,$$

und wegen der Unabhängigkeit der Summanden auch

$$\mathbb{V}(X) = \sum_{i=1}^{n} \mathbb{V}(Y_i) = np(1-p).$$

Diese Methode kann erstaunliche Abkürzungen liefern, wo sie anwendbar ist, etwa auch für die hypergeometrische Verteilung: für den Erwartungswert ist das praktisch dieselbe Rechnung wie für die Binomialverteilung, für die Varianz ist die Rechnung ein wenig komplizierter, weil die Summanden jetzt nicht mehr unabhängig sind und daher auch die Kovarianzen berücksichtigt werden müssen.

Für Erwartungswert und Varianz gibt es Standardwerte:

Definition 2.24

Wenn $\mathbb{E}(X) = 0$, dann heißt X zentriert.

Wenn $\mathbb{E}(X^2) = 1$, dann heißt X normiert.

Wenn $\mathbb{E}(X) = 0$ und $\mathbb{V}(X) = 1$, dann heißt X standardisiert.

Wenn $\mu = \mathbb{E}(X)$ endlich ist, dann ist $X^o = X - \mu$ zentriert (die Zentrierung von X).

Wenn $\mathbb{E}(X^2)$ endlich und positiv ist, dann ist $X^* = \frac{X}{\sqrt{\mathbb{E}(X^2)}}$ normiert (die Normierung von X).

Wenn $\sigma^2 = \mathbb{V}(X)$ endlich und positiv ist, dann ist $X^{o*} = \frac{X-\mu}{\sigma}$ standardisiert (die Standardisierung von X).

Definition 2.25

X und Y seien Zufallsvariable mit endlicher Varianz. Dann heißt

$$\mathbf{Cov}(X,Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

die Kovarianz von X und Y.

Damit erhalten wir

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\mathbf{Cov}(X,Y).$$

Wenn die Kovarianz von zwei Zufallsvariablen gleich 0 ist, dann nennen wir sie unkorreliert. Aus der Unabhängigkeit folgt die Unkorreliertheit, die umgekehrte Aussage gilt nicht, wie das Beispiel $X \sim N(0,1), Y = X^2$ zeigt.

Aus der Kovarianz erhalten wir

Definition 2.26

X und Y seien zwei Zufallsvariable mit positiver endlicher Varianz. Dann heißt

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

der Korrelationskoeffizient von X und Y.

Satz 2.19

Es gilt

$$-1 \leq \rho(X,Y) \leq 1.$$

$\rho(X,Y) = 1$ genau dann, wenn es Konstante $a > 0$ und b gibt mit $Y = aX + b$.
$\rho(X,Y) = -1$ genau dann, wenn es Konstante $a < 0$ und b gibt mit $Y = aX + b$.
$\rho(X,Y) = 0$ genau dann, wenn X und Y unkorreliert sind.

Dieser Satz folgt direkt aus

Satz 2.20: Cauchy-Schwarz Ungleichung

Wenn $\mathbb{E}(X^2)$ und $\mathbb{E}(Y^2)$ endlich sind, dann gilt

$$\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$

Gleichheit gilt genau dann, wenn es ein $a > 0$ gibt, sodass $Y = aX$.

Wenn $E(X^2) = 0$ (oder $\mathbb{E}(Y^2) = 0$), dann ist mit Wahrscheinlichkeit 1 $X = 0$, also auch $XY = 0$ und daher auch $\mathbb{E}(XY) = 0$. In diesem Fall gilt die Ungleichung.

Im anderen Fall ist die Ungleichung äquivalent dazu, dass für die Normierungen X^* und Y^* die Ungleichung

$$\mathbb{E}(X^*Y^*) \leq 1$$

erfüllen. Wir können also o.B.d.A. annehmen, dass X und Y normiert sind. Quadrate sind nichtnegativ, also

$$(X-Y)^2 \geq 0$$

und damit auch

$$0 \leq \mathbb{E}((X-Y)^2) = \mathbb{E}(X^2) - 2\mathbb{E}(XY) + \mathbb{E}(Y^2) = 2 - 2\mathbb{E}(XY).$$

Also

$$\mathbb{E}(XY) \leq 1,$$

was wir zeigen wollten. Für die normierten Variablen gilt also Gleichheit genau dann, wenn X und Y mit Wahrscheinlichkeit 1 übereinstimmen. Für den allgemeinen Fall bedeutet dass, dass

$$Y = \frac{\sqrt{\mathbb{E}(Y^2)}}{\sqrt{\mathbb{E}(X^2)}}X,$$

also $Y = aX$ mit $a > 0$. Umgekehrt rechnet man leicht nach, dass in diesem Fall in der Ungleichung Gleichheit herrscht.

Beispiel 2.19

Wir berechnen die Varianz der Binomialverteilung $B(n,p)$. Wir wissen bereits, dass

$$\mathbb{E}(X) = np.$$

Wir berechnen also

$$\mathbb{E}(X^2) = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} =$$

$$np \sum_{k=1}^{n} k \binom{n-1}{k-1} p^{k-1} p^{n-k} = np \sum_{l=0}^{n-1} (l+1) \binom{n-1}{l} p^l p^{n-1-l} = np((n-1)p+1) = n(n-1)p^2 + np.$$

Damit ergibt sich

$$\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = n(n-1)p^2 + np - n^2p^2 = np(1-p).$$

Satz 2.21: Ungleichung von Markov

X sei eine nichtnegative Zufallsvariable, $\lambda > 0$. Dann ist

$$\mathbb{P}(X \geq \lambda) \leq \frac{1}{\lambda} \mathbb{E}(X).$$

Zum Beweis definieren wir die Zufallsvariable

$$Y = \begin{cases} \lambda & \text{wenn } X \geq \lambda, \\ 0 & \text{sonst.} \end{cases}$$

Es gilt $Y \leq X$ und daher

$$\mathbb{E}(X) \geq \mathbb{E}(Y) = \lambda \mathbb{P}(Y = \lambda) = \lambda \mathbb{P}(X \geq \lambda).$$

Satz 2.22: Ungleichung von Chebychev

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \lambda) \leq \frac{\mathbb{V}(X)}{\lambda^2}.$$

Wir wenden die Ungleichung von Markov auf die Zufallsvariable

$$Y = (X - \mathbb{E}(X))^2$$

an:

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \lambda) = \mathbb{P}(Y \geq \lambda^2) \leq \frac{\mathbb{E}(Y)}{\lambda^2} = \frac{\mathbb{V}(X)}{\lambda^2}.$$

Satz 2.23: Ungleichung von Kolmogorov

$X_1, \ldots, X_n$ seien unabhängig mit Erwartungswert 0, $S_0 = 0$, $S_n = X_1 + \ldots + X_n$. Dann ist

$$\mathbb{P}(\max_{k \leq n} |S_k| \geq \lambda) \leq \frac{\mathbb{V}(S_n)}{\lambda^2}.$$

Für $i = 1, \ldots, n$ setzen wir

$$Y_i = \begin{cases} 1 & \text{wenn } |S_i| \geq \lambda \text{ und } |S_j| < \lambda \text{ für } j < i, \\ 0 & \text{sonst,} \end{cases}$$

und

$$Y = \sum_{i=1}^{n} Y_i.$$

Y ist genau dann 1, wenn $\max_{0\le k\le n}|S_k| \ge \lambda$, und sonst 0.

Weil $\mathbb{E}(S_n) = 0$ ist, gilt

$$\mathbb{V}(S_n) = \mathbb{E}(S_n^2) \ge \mathbb{E}(YS_n^2) = \sum_{i=1}^{n} \mathbb{E}(Y_i S_n^2).$$

Für die Summanden ergibt sich

$$\mathbb{E}(Y_i S_n^2) = \mathbb{E}(Y_i(S_i + (S_n - S_i))^2) = \mathbb{E}(Y_i S_i^2) + 2\mathbb{E}(Y_i S_i(S_n - S_i)) + \mathbb{E}(Y_i(S_n - S_i)^2).$$

Im ersten Summanden ist $Y_i S_i^2 \ge Y_i \lambda^2$, im zweiten sind $Y_i S_i$ und $S_n - S_i$ unabhängig, und daher ist der Erwartungswert 0, und der letzte Summand kann einfach mit 0 nach unten abgeschätzt werden. Das ergibt

$$\mathbb{V}(S_n) \ge \sum_{i=1}^{n} \lambda^2 \mathbb{E}(Y_i) = \lambda^2 \mathbb{E}(Y) = \lambda^2 \mathbb{P}(\max_{0\le k\le n} |S_n| \ge \lambda).$$

2.6 Momente

Neben den Erwartungswerten von X und X^2 kann man auch die von höheren Potenzen betrachten:

Definition 2.27

$$M_n(X) = \mathbb{E}(X^n)$$

heißt das n-te Moment von X,

$$m_n(X) = \mathbb{E}((X - \mathbb{E}(X))^n)$$

das n-te zentrale Moment von X.

Wir müssen aber X nicht unbedingt als Basis für die Potenzen nehmen, es kann auch im Exponenten stehen. Das gibt die folgenden Definitionen:

Definition 2.28

X sei eine diskrete Zufallsvariable mit nichtnegativen ganzzahligen Werten. Dann heißt die Funktion

$$p_X^*(z) = \mathbb{E}(z^X) = \sum_{n=0}^{\infty} z^n p_X(n)$$

(für reelles oder auch komplexes z) die *wahrscheinlichkeitserzeugende Funktion* von X.

Generell heißt die Funktion

$$a^*(z) = \sum_{n=0}^{\infty} a_n z^n$$

die erzeugende Funktion der Folge a_n, In solchen Potenzreihen ist nicht immer garantiert, dass sie für irgendein $z \neq 0$ konvergieren. Im Fall der wahrscheinlichkeitserzeugenden Funktion sind aber die Koeffizienten nichtnegativ und haben Summe 1, deshalb konvergiert diese jedenfalls für alle z mit $|z| \le 1$. Sie legt die Verteilung von X fest — die Werte der Wahrscheinlichkeitsfunktion können ja als Taylorkoeffizienten erhalten werden.

Für alle Zufallsvariablen (diskret,stetig oder gemischt) kann die folgende Funktion definiert werden.

Definition 2.29

Die Funktion

$$M_X(t) = \mathbb{E}(e^{Xt})$$

heißt die *momentenerzeugende Funktion* von X.

Diese Funktion muss für kein $t \neq 0$ endlich sein (was etwa bei der Cauchyverteilung der Fall ist). Wenn Sie für ein $t \neq 0$ endlich ist, dann auch für alle Argumente zwischen 0 und t, und dann legt sie die Verteilung von X fest.

Ihr Name erklärt sich aus der Potenzreihenentwicklung der Exponentialfunktion: es gilt ja

$$e^{Xt} = \sum_{n=0}^{\infty} \frac{X^n t^n}{n!}$$

und daher

$$M_X(t) = \sum_{n=0}^{\infty} \frac{M_n(X) t^n}{n!}.$$

Ist $X \geq 0$, dann existiert M_X jedenfalls für $t \leq 0$. In diesem Fall wird $L(t) = M_X(-t)$ auch gelegentlich als Laplacetransformierte der Verteilung von X bezeichnet. Für stetig verteiltes X ist ja

$$L(t) = \mathbb{E}(e^{-Xt}) = \int_0^{\infty} e^{-xt} f_X(x) dx,$$

als die Laplacetransformierte von f_X. Ist X diskret mit nichtnegativen ganzzahligen Werten, dann gilt natürlich

$$M_X(t) = p_X^*(e^t).$$

Die wahrscheinlichkeitserzeugende und die momentenerzeugende Funktion haben eine nette Eigenschaft gemeinsam: wenn X und Y unabhängig sind, dann gilt für die Summe

$$p_{X+Y}^*(z) = p_X^*(z) p_Y^*(z)$$

und

$$M_{X+Y}(t) = M_X(t) M_Y(t),$$

vorausgesetzt, dass die rechte Seite Sinn macht. Es ergibt sich also die Momentenerzeugende bzw. Wahrscheinlichkeitserzeugende der Summe als das Produkt der entsprechenden Funktionen der einzelnen Summanden. Das macht diese Funktionen zu einem praktischen Hilfsmittel für die Untersuchung von Summen von unabhängigen Zufallsvariablen, weil die Multiplikation sehr viel leichter zu handhaben ist als die etwas umständliche Faltung. Das ist so schön, dass wir gerne für alle Verteilungen so etwas hätten, und ein kleiner Ausflug in die Welt der komplexen Zahlen beschert es uns:

Definition 2.30

Die charakteristische Funktion ϕ_X der Zufallsvariable X ist gegeben durch

$$\phi_X(t) = \mathbb{E}(e^{iXt}) = \mathbb{E}(\cos(Xt)) + i\mathbb{E}(\sin(Xt)), t \in \mathbb{R}.$$

Weil hier der Integrand beschränkt ist, gibt es keine Probleme mit der Existenz des Erwartungswerts. Wie vorher gilt, dass die charakteristische Funktion die Verteilung eindeutig festlegt und dass die charakteristische Funktion einer Summe von unabhängigen Zufallsvariablen gleich dem Produkt der charakteristischen Funktionen der Summanden ist. So wie die Momentenerzeugende eine Analogie zur Laplacetransformation aufweist, ist die charakteristische Funktion analog zur Fouriertransformation.

Bevor wir uns an einige Beispiele heranwagen, fassen wir einige Eigenschaften der Momentenerzeugenden und der charakteristischen Funktion zusammen:

Satz 2.24

F[r die momentenerzeugende Funktion und die charakteristische Funktion gilt:

1. Wenn X und Y unabhängig sind, dann gilt
$$M_{X+Y}(t) = M_X(t)M_Y(t),$$
$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t).$$
2. Für $Y = aX + b$, $a, b \in \mathbb{R}$:
$$M_Y(t) = e^{bt}M_X(at),$$
$$\phi_Y(t) = e^{ibt}\phi_X(at).$$
3. Wenn $m_k = \mathbb{E}(X^k)$ endlich ist, dann ist ϕ_X k-mal differenzierbar, und
$$\phi_X^{(k)}(0) = i^k m_k.$$
Wenn wir zusätzlich annehmen, dass $M_X(t)$ für ein $t \neq 0$ endlich ist, dann gilt auch
$$M_X^{(k)}(0) = m_k.$$
(unter Umständen, wenn die Momentenerzeugende nur auf einer Seite der Null endlich ist, ist hier die einseitige Ableitung zu verwenden. Wenn das nicht der Fall ist, wenn es also negative und positive Werte von t gibt, für die $M_X(t)$ endlich ist, dann sind auch alle Momente m_k endlich).

Beispiel 2.20: Die Laplaceverteilung und die Cauchyverteilung

Die Laplaceverteilung hat die Dichte
$$f(x) = \frac{1}{2}e^{-|x|}, x \in \mathbb{R}.$$
Sie wird auch manchmal "doppelte Exponentialverteilung" genannt, obwohl sie um diesen Namen mit der Gumbelverteilung mit der Verteilungsfunktion $e^{-e^{-x}}$ streiten muss.

Ihre Momente sind nicht allzu schwer zu berechnen. Weil die Verteilung symmetrisch ist, sind die ungeraden Momente 0, und für gerades $n = 2k$ gilt
$$\mathbb{E}(X^n) = \int_{-\infty}^{\infty} x^n \frac{1}{2}e^{-|x|}dx = 2\int_0^{\infty} x^n \frac{1}{2}e^{-x}dx = n!.$$
Die Momentenerzeugende ergibt sich (für $-1 < t < 1$, für andere Werte ist sie unendlich)
$$M_X(t) = \frac{1}{1-t^2}.$$
Daraus erhält man, wenn man einfach t durch it ersetzt, die charakteristische Funktion
$$\phi_X(t) = \frac{1}{1-(it)^2} = \frac{1}{1+t^2}.$$
Das ist, bis auf einen Faktor $1/\pi$, die Dichte der Cauchyverteilung. Wenn wir die Umkehrformel für die Fouriertransformation anwenden, ergibt sich
$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{1+t^2}e^{-ixt}dt,$$

und nach trivialen Umformungen

$$\int_{-\infty}^{\infty} \frac{1}{\pi(1+t^2)} e^{ixt} dt = e^{-|x|}.$$

Damit haben wir die charakteristische Funktion der Cauchyverteilung erhalten (mit x als Argument statt des üblichen t).

Beispiel 2.21

Dieses Kapitel wäre nicht komplett, wenn unsere Lieblingsverteilung, die Normalverteilung, nicht zur Sprache käme.

Wir beginnen mit der Standardnormalverteilung und bestimmen zuerst die momentenerzeugende Funktion:

$$M_X(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{xt} e^{-x^2/2} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{t^2-(x-t)^2}{2}} dx = e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-t)^2}{2}} dx = e^{t^2/2},$$

weil das letzte Integral als Integral der Dichte einer Normalverteilung $N(t,1)$ den Wert 1 hat.

Wenn wir t durch it ersetzen, erhalten wir die charakteristische Funktion

$$\phi_X(t) = e^{(it)^2/2} = e^{-t^2/2}.$$

Aus der Reihenentwicklung der momentenerzeugenden Funktion können wir die Momente ablesen:

$$\sum_{k=0}^{\infty} \frac{m_k t^k}{k!} = e^{t^2/2} = \sum_{n=0}^{\infty} \frac{t^{2n}}{2^n n!},$$

also

$$m_k = \begin{cases} 0 & \text{wenn } k \text{ ungerade,} \\ \frac{k!}{2^k (k/2)!} = k(k-2)\ldots 3 \cdot 1 & \text{wenn } k \text{ gerade.} \end{cases}$$

Falls Y eine allgemeine Normalverteilung $N(\mu, \sigma^2)$ hat, dann ist $X = (Y-\mu)/\sigma$ standardnormalverteilt, und wir erhalten

$$M_Y(t) = M_{\sigma X + \mu}(t) = e^{\mu t + \sigma^2 t^2/2}$$

und

$$\phi_Y(t) = e^{i\mu t - \sigma^2 t^2/2}.$$

2.7 Folgen von Zufallsvariablen

In diesem Abschnitt begegnen uns zwei der klassischen "großen" Sätze der Wahrscheinlichkeitstheorie. In diesen ist der Ausgangspunkt eine Folge $(X_n, n \in \mathbb{N})$ von unabhängigen Zufallsvariablen, die alle dieselbe Verteilung haben. Aus diesen werden die Partialsummen

$$S_n = \sum_{i=1}^{n} X_i$$

gebildet.

Die klassischen Sätze der Wahrscheinlichkeitstheorie beschreiben, wie sich diese Summen bzw. die *Stichprobenmittel*

$$\bar{X}_n = \frac{S_n}{n}$$

für großes n verhalten.

Satz 2.25: schwaches Gesetz der großen Zahlen

$(X_n, n \in \mathbb{N})$ sei eine Folge von unabhängigen, identisch verteilten Zufallsvariablen mit Erwartungswert μ und endlicher Varianz, $S_n = X_1 + \ldots + X_n$. Dann gilt

$$\mathbb{P}(|\frac{S_n}{n} - \mu| \geq \epsilon) \to 0$$

für jedes $\epsilon > 0$.

Der Beweis dieses Satzes ist nicht schwer, wir verwenden einfach die Ungleichung von Chebychev:

$$\mathbb{P}(|\frac{S_n}{n} - \mu| \geq \epsilon) = \mathbb{P}(|S_n - n\mu| \geq n\epsilon) = \mathbb{P}(|S_n - \mathbb{E}(S_n)| \geq n\epsilon) \leq \frac{\mathbb{V}(S_n)}{(n\epsilon)^2} = \frac{\mathbb{V}(X_1)}{n\epsilon^2},$$

und das geht natürlich gegen 0.

Die Beschreibung der Schlussfolgerung im letzten Satz gibt Anlass zu einer Definition und noch zwei weiteren:

Definition 2.31

Wir betrachten Zufallsvariable $X_n, n \in \mathbb{N}$ und X. Die Folge X_n konvergiert gegen X

1. in Wahrscheinlichkeit, wenn für alle $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

2. mit Wahrscheinlichkeit 1 (oder "fast sicher"), wenn

$$\mathbb{P}(X_n \text{ konvergiert gegen } X) = 1.$$

3. in Verteilung, wenn für alle Stetigkeitspunkte x von F_X

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

Mit diesen Definitionen können wir das schwache Gesetz der großen Zahlen so formulieren: die Folge S_n/n konvergiert in Wahrscheinlichkeit gegen den Erwartungswert μ.

Von diesen drei Konvergenzarten ist die Konvergenz mit Wahrscheinlichkeit 1 die stärkste, die Konvergenz in Verteilung die schwächste: aus der Konvergenz mit Wahrscheinlichkeit 1 folgt die in Wahrscheinlichkeit und aus dieser die in Verteilung. Die umgekehrten Implikationen gelten nicht. Wenn wir X so wählen, dass $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$ gilt, und wir für alle $n \in \mathbb{N}$ $X_n = -X$ setzen, dann haben alle X_n und X dieselbe Verteilung, und deshalb konvergiert X_n in Verteilung gegen X. Andererseits ist offensichtlich für alle n $|X_n - X| = 2$, und daher konvergiert X_n nicht in Wahrscheinlichkeit gegen X.

Eine Folge, die in Wahrscheinlichkeit, aber nicht fast sicher konvergiert, lässt sich so konstruieren: für $k \in \mathbb{N}$ wählen wir N_k gleichverteilt zwischen $2^k + 1$ und 2^{k+1}, also

$$\mathbb{P}(N_k = 2^k + i) = \frac{1}{2^k}, i = 1, \ldots, 2^k.$$

Wir setzen $X_n = 1$, wenn es ein k gibt mit $n = N_k$, sonst setzen wir $X_n = 0$. Für jedes $n \geq 3$ können wir ein k finden, für das $2^k + 1 \leq n \leq 2^{k+1}$ gilt. Daraus folgt

$$\mathbb{P}(X_n \neq 0) = \mathbb{P}(X_n = 1) = \frac{1}{2^k} \leq \frac{2}{n} \to 0.$$

Deshalb konvergiert die Folge (X_n) in Wahrscheinlichkeit gegen $X = 0$. Es gibt aber in jedem Fall für jedes k ein n zwischen $2^k + 1$ und 2^{k+1}, für das $X_n = 1$ gilt. Es ist also für unendlich viele n

$X_n = 1$. Daher konvergiert die Folge X_n nicht nur nicht mit Wahrscheinlichkeit 1, sondern sogar mit Wahrscheinlichkeit 1 nicht.

Die Formulierung der Konvergenz in Verteilung bedarf auch einer gewissen Erklärung: es wäre auf den ersten Blick natürlicher, die Konvergenz $F_{X_n}(x) \to F_X(x)$ für alle reellen x zu fordern. Diese Forderung ist aber schon für einfache Beispiele zu stark. Wenn wir etwa $X_n = 1/n$ setzen, dann konvergiert X_n mit Wahrscheinlichkeit 1 gegen $X = 0$. Dann ist

$$F_{X_n}(x) = \begin{cases} 1 & \text{wenn } x \geq 1/n, \\ 0 & \text{sonst,} \end{cases}$$

und daher

$$\lim_{n\to\infty} F_{X_n}(x) = \begin{cases} 1 & \text{wenn } x > 0, \\ 0 & \text{sonst.} \end{cases}$$

Das stimmt für alle $x \neq 0$ mit $F_X(x)$ überein, aber für $x = 0$ haben wir

$$\lim_{n\to\infty} F_{X_n}(0) = 0 \neq 1 = F_X(0).$$

Die Konvergenz von X_n gegen X in Verteilung ist äquivalent dazu, dass für jede beschränkte stetige Funktion g die Konvergenz

$$\lim_{n\to\infty} \mathbb{E}(g(X_n)) = \mathbb{E}(g(X))$$

gilt. In der Mathematik wird diese Eigenschaft als Definition der Konvergenz in Verteilung verwendet, und die Konvergenz der Verteilungsfunktionen daraus abgeleitet.

Satz 2.26: Starkes Gesetz der großen Zahlen

$(X_n, n \in \mathbb{N})$ sei eine Folge von unabhängigen, identisch verteilten Zufallsvariablen mit endlichem Erwartungswert $\mu = \mathbb{E}(X_n)$, $S_n = X_1 + \ldots + X_n$. Dann konvergiert $\frac{S_n}{n}$ mit Wahrscheinlichkeit 1 gegen μ.

Dieser Satz liefert einerseits eine mathematisch strenge Version des "empirischen Gesetzes der großen Zahlen", andererseits bildet er auch die Grundlage zur Berechnung von Erwartungswerten und Integralen durch Simulation. Wenn wir den Erwartungswert einer Verteilung bestimmen wollen, erzeugen wir einfach n unabhängige Zufallszahlen mit dieser Verteilung. Das Stichprobenmittel $\bar{X}_n$ sollte dann für hinreichend großes n nahe beim gesuchten Erwartungswert liegen. Integrale der Form

$$I = \int_0^1 \int f(x)dx$$

kann man als Erwartungswert

$$I = \mathbb{E}(f(X))$$

betrachten, wobei X auf $[0, 1]$ gleichverteilt ist. Diese Idee ist natürlich nur dann brauchbar, wenn wir auch die Genauigkeit dieser Näherung angeben können. Wie bei allen wahrscheinlichkeitstheoretischen Verfahren ist das nur mit einer gewissen Wahrscheinlichkeit möglich. Eine grobe Abschätzung dieser Genauigkeit liefert die Ungleichung von Chebychev. Etwas besser geht das mit dem nächsten Satz. Es wird sich herausstellen, dass sich mit n eine Genauigkeit in der Größenordnung $1/\sqrt{n}$ erreichen lässt. Das ist auf den ersten Blick nicht gar so berauschend, die Trapez- oder Simpsonregel garantieren bessere Raten. Allerdings macht die Simulation keine Voraussetzungen über die Funktion f, während die deterministischen Verfahren Annahmen über die Ableitungen von f benötigen. Theoretisch müsste für die Simulation f nicht einmal stetig sein, in der Praxis sind die Verteilungen der "gleichverteilten" Zufallszahlen, die wir für die Simulation verwenden, allerdings nur diskrete Approximationen der stetigen Gleichverteilung, also ist doch etwas Stetigkeit notwendig. Die wahre Stärke der Simulationsmethode zeigt sich aber bei mehrdimensionalen Integralen: um ein Integral wie

$$J = \int_0^1 \ldots \int_0^1 f(x_1, \ldots, x_{20})dx_1 \ldots dx_{20}$$

mit der Trapezmethode zu bestimmen, müssen wir ein 20-dimensionales Gitter verwenden. Bei auch nur zwei Stützwerten in jeder Dimension sind das schon $2^{20} = 1048576$ Punkte. Von der anderen Seite her betrachtet: wenn wir n Punkte verwenden, sind das $n^{1/d}$ Stützwerte pro Dimension, und etwa mit der Simpsonregel erhalten wir einen Fehler in der Größenordnung $n^{-4/d}$. Mit der Simulationsmethode können wir das Integral durch

$$\hat{J}_n = \frac{1}{n}\sum_{i=0}^{n-1} f(X_{id+1}, \ldots, X_{id+d})$$

annähern, und es gilt immer noch die Fehlerabschätzung in der Größenordnung $n^{-1/2}$.

Satz 2.27: Zentraler Grenzwertsatz

$(X_n, n \in \mathbb{N})$ sei eine Folge von unabhängigen, identisch verteilten Zufallsvariablen mit $\mathbb{E}(X) = \mu$, $\mathbb{V}(X) = \sigma^2$. Dann ist S_n näherungsweise normalverteilt mit Mittel $n\mu$ und Varianz $n\sigma^2$, d.h., es gilt

$$\lim_{n\to\infty} \mathbb{P}(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du.$$

Einen wichtigen Sonderfall bildet die Binomialverteilung: Wenn wir $X_n \sim A(p)$ setzen, dann ist $S_n \sim B(n,p)$, und der zentrale Grenzwertsatz sagt uns, dass die Binomialverteilung für großes n durch eine Normalverteilung $N(np, np(1-p))$ approximiert werden kann. In diesem Fall kann man aber mehr bekommen, die Approximation gilt nicht nur für die Verteilungsfunktion, sondern auch für die Wahrscheinlichkeitsfunktion:

Satz 2.28: deMoivre-Laplace

X_n sei binomialverteilt: $X_n \sim B(n,p)$. Dann gilt

$$\mathbb{P}(X_n = k) = \frac{1}{\sqrt{2n\pi p(1-p)}} e^{-\frac{(k-np)^2}{2np(1-p)}} + o(\frac{1}{\sqrt{n}})$$

gleichmäßig in k für $n \to \infty$.

Dieser Satz ist stärker als der zentrale Grenzwertsatz, weil er nicht nur die Konvergenz der Verteilungsfunktionen behauptet, sondern auch , dass sich die die Wahrscheinlichkeitsfunktion durch die Dichtefunktion der Normalverteilung annähern lässt. Solche Sätze werden als "lokale Grenzwertsätze" bezeichnet. Um den Satz von deMoivre-Laplace zu beweisen, kann etwa die Stirling-Formel

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

und die Taylor-Entwicklung verwendet werden. Die Details sind ein wenig umständlich, deshalb wird der Beweis hier nicht aufgeführt.

Für den zentralen Grenzwertsatz in der allgemeinen Form gibt es eine Vielzahl an Beweisen, darunter sind elementare, die allerdings etwas umständlich sind; mithilfe der charakteristischen Funktion lässt sich ein sehr kurzer Beweis führen (was natürlich geschummelt ist: der Beweis verwendet Eigenschaften der charakteristischen Funktion, die auch erst bewiesen werden müssen). Auch diesen Beweis werden wir nicht führen.

Für uns sind auch die Beweise dieser Aussagen nicht so wichtig wie ihre Anwendung: damit lassen sich Wahrscheinlichkeiten, in denen die Summen von unabhängigen Zufallsvariablen vorkommen, näherungsweise berechnen. Dazu muss die Anzahl der Summanden hinreichend groß sein — für die Binomialverteilung und die Poissonverteilung gibt es die Faustregel, dass die Varianz mindestens 9 sein soll; die Regel $n \geq 30$, die oft für den allgemeinen Fall kolportiert wird, ist eher mit Vorsicht zu genießen — generell kann es keine allgemeingültige "$n >$"-Regel geben: eine Million Summanden mit einer $P(0.000001)$-Verteilung ergeben eine Summe mit einer Poisson-Verteilung mit Parameter 1. Etwas besser funktionieren Abschätzungen des Fehlers mit Hilfe höherer Momente, aber diese tendieren dazu, etwas zu pessimistisch zu sein.

Beispiel 2.22

Als Beispiel berechnen wir die Wahrscheinlichkeit, dass in 1000 Würfen eines fairen Würfels mehr als 200 mal eine Sechs erscheint:

$$\mathbb{P}(X > 200) \approx 1 - \Phi(\frac{200 - 1000/6}{\sqrt{1000 * 5/36}}) = 1 - \Phi(2.83) = 0.002.$$

Beispiel 2.23

Nehmen wir an, wir wollen das Integral

$$I = \int_0^\pi \sin(x)dx$$

durch Simulation bestimmen. Wir möchten wissen, wie groß wir n wählen müssen, um mit 99% Wahrscheinlichkeit eine Genauigkeit von zwei Nachkommastellen (also ± 0.005) zu garantieren. Zuerst transformieren wir das Integral auf das Intervall $[0, 1]$:

$$I = \int_0^1 \pi \sin(\pi x)dx.$$

Unsere Näherung ist

$$\hat{I}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

mit

$$Y_i = f(X_i) = \pi \sin(\pi X_i),$$

und X_i sind die unabhängigen $U(0, 1)$-Zufallszahlen aus unserem Generator.

Für die Fehlerabschätzung haben wir zwei Möglichkeiten. Die erste führt über die Ungleichung von Chebychev: diese liefert pessimistischere Ergebnisse, hat aber den Vorteil, dass die geforderte Genauigkeit unter allen Umständen garantiert ist. Die andere Möglichkeit verwendet den zentralen Grenzwertsatz und liefert realistischere Ergebnisse, wenn Sie anwendbar ist, aber sie kann in extremen Fällen zu optimistisch sein.

Beide Methoden verwenden die Varianz von Y_i. Diese berechnet sich als

$$\sigma^2 = \mathbb{E}(Y^2) = \mathbb{E}(Y)^2 = \int_0^1 f(x)^2 dx - I^2.$$

Da wir das Integral I erst berechnen wollen, können wir im allgemeinen nicht annehmen, dass wir σ^2 kennen. Eine Möglichkeit wäre eine grobe Abschätzung: hier wissen wir etwa, dass $|f| \leq \pi$ gilt, was die Abschätzung $\sigma \leq \pi$ liefert, wenn man auch noch bemerkt, dass $f \geq 0$ gilt, dann lässt sich diese Abschätzung auf $\sigma \leq \pi/2$ verbessern. Wir werden hier schummeln und die exakten Werte einsetzen. In der Praxis geht man meist so vor, dass auch für die Varianz ein Näherungswert aus der Simulation berechnet und damit der Fehler abgeschätzt wird. Das lässt sich sehr schön iterativ bewerkstelligen.

Wir wollen (mit $\delta = 0.005$)

$$\mathbb{P}(I - \delta \leq \hat{I}_n \leq I + \delta) \geq 0.99$$

Die Ungleichung von Chebychev liefert die Abschätzung

$$\mathbb{P}(|\hat{I}_n - I| \geq \delta) \leq \frac{\mathbb{V}(I_n)}{\delta^2} = \frac{\sigma^2}{n\delta^2}.$$

Wir machen diesen letzten Ausdruck kleiner als 0.01, dafür muss

$$n \geq \frac{100\sigma^2}{\delta^2} = 3740000$$

sein.

Aus dem zentralen Grenzwertsatz erhalten wir, dass I_n näherungsweise normalverteilt (mit Mittel I und Varianz σ^2/n) ist, und daher

$$\mathbb{P}(-\delta \leq \hat{I}_n - I \leq \delta) = \mathbb{P}(-\frac{\delta}{\sigma/\sqrt{n}} \leq \frac{\hat{I}_n - I}{\sigma/\sqrt{n}} \leq \frac{\delta}{\sigma/\sqrt{n}}) \approx \Phi(\frac{\delta}{\sigma/\sqrt{n}}) - \Phi(-\frac{\delta}{\sigma/\sqrt{n}}) =$$

$$2\Phi(\frac{\delta}{\sigma/\sqrt{n}}) - 1.$$

Das soll 0.99 sein, also

$$\Phi(\frac{\delta}{\sigma/\sqrt{n}}) = 0.995,$$

und

$$\frac{\delta}{\sigma/\sqrt{n}} = \Phi^{-1}(0.995) = 2.576,$$

also

$$n = \frac{2.576^2\sigma^2}{\delta^2}.$$

Wir verwenden wieder unser Wissen, dass $\sigma^2 = \frac{\pi^2}{2} - 4 = 0.935$ gilt:

$$n = \frac{2.576^2 \cdot 0.935}{0.005^2} = 248178.$$

Eine schnelle Berechnung mit dem Statistikpaket R ergab als Ergebnis einer Simulation mit diesem n

$$\hat{I}_n = 2.002622,$$

drei weitere Versuche die Werte 2.000864, 1.999885 und 2.001498.

2.8 Spezielle Verteilungen

In diesem Abschnitt sammeln wir einige wichtige Beispiele für Verteilungen, die oft gebraucht werden. Bei den Verteilungen, die nicht schon früher in diesem Kapitel erwähnt worden sind, geben wir auch an, wie man dazu kommt.

2.8.1 Diskrete Verteilungen

Die Binomialverteilung und die Hypergeometrische Verteilung wurden schon besprochen.

Die geometrische Verteilung tritt auf, wenn ein Experiment mit Erfolgswahrscheinlichkeit p so lange wiederholt wird, bis der erste Erfolg eintritt. Wird dabei X als die Anzahl der Versuche definiert, dann hat X die Verteilung $G^*(p)$. Man kann aber auch die Anzahl der *Misserfolge* zählen, die man in Kauf nehmen muss, bis der erste Erfolg eintritt. Diese Anzahl ist dann nach $G(p)$ verteilt. Auf den ersten Blick erscheint die erste Betrachtungsweise natürlicher, aber die Verteilung, die bei 0 anfängt, tritt öfter auf als die, die bei 1 anfängt, deshalb ist für uns $G(p)$ die Standardversion der geometrischen Verteilung, und $G^*(p)$ die alternative Form (weil sie aus $G(p)$ durch eine Verschiebung um eins nach rechts entsteht, heißt sie auch die "verschobene geometrische Verteilung". Dasselbe gilt für die negative Binomialverteilung: NB^* zählt die Anzahl der Versuche bis zum n-ten Erfolg, NB die Anzahl der Misserfolge, die man in Kauf nehmen muss, bis der n-te Erfolg eintritt. In der negativen Binomialverteilung NB kann man auch nicht ganzzahliges n verwenden,

wenn man den Binomialkoeffizienten für beliebiges x als

$$\binom{x}{k} = \frac{x(x-1)\dots(x-k+1)}{k!}$$

definiert.

Die Poissonverteilung ergibt sich als Grenzfall der Binomialverteilung für $n \to infty$ und $np \to \lambda$. Das wird in den Übungen nachgewiesen.

Name	Symbol	$p(x)$	$\mathbb{E}(X)$	$\mathbb{V}(X)$
Binomial $n \in \mathbb{N}, 0 \leq p \leq 1$	$B(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}(0 \leq x \leq n)$	np	$np(1-p)$
Gleichverteilung (diskret) $a \leq b, a, b \in \mathbb{Z}$	$D(a,b)$	$\frac{1}{b-a+1}(a \leq x \leq b)$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$
Geometrisch $0 < p < 1$	$G(p)$	$p(1-p)^x, (x \geq 0)$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Geometrisch (alternativ) $0 < p < 1$	$G^*(p)$	$p(1-p)^{x-1}, (x \geq 1)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negativ Binomial $n > 0, 0 < p < 1$	$NB(n,p)$	$\binom{n+x-1}{n}p^n(1-p)^x, (x \geq 0)$	$\frac{n(1-p)}{p}$	$\frac{n(1-p)}{p^2}$
Negativ Binomial (alternativ) $n \in \mathbb{N}, 0 < p < 1$	$NB^*(n,p)$	$\binom{x-1}{n-1}p^{n-x}(1-p)^x, (x \geq n)$	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$
Poisson $\lambda > 0$	$P(\lambda)$	$\frac{\lambda^x e^{-\lambda}}{x!}$	λ	λ
Hypergeometrisch $N, A, n \in \mathbb{N}$, $n, A \leq N$	$H(N,A,n)$	$\frac{\binom{A}{x}\binom{N-A}{n-x}}{\binom{N}{n}}(0 \leq x \leq n)$	$\frac{nA}{N}$	$\frac{nA(N-A)(N-n)}{N^2(N-1)}$

2.8.2 Stetige Verteilungen

Die stetige Gleichverteilung ist ein Konzept, das uns schon in unseren ersten Überlegungen begegnet ist — eine Zahl wird zufällig aus dem Intervall $[a,b]$ gewählt, und alle diese Zahlen sind "gleichberechtigt". Eine alternative Betrachtungsweise ist, diese Verteilung als Grenzfall von diskreten Gleichverteilungen auf einem immer feineren Gitter anzusehen, so, wie wir es in Beispiel 2.2 getan haben.

Auch über die Normalverteilung haben wir schon gesprochen; der zentrale Grenzwertsatz gibt uns Anlass, viele Größen, die als Summe von vielen kleinen und unabhängigen Einflüssen zustande kommen, als zumindest näherungsweise normalverteilt anzusehen. Das und auch die Tatsache, dass sie in der Statistik sehr angenehm zu behandeln ist, machen sie zur wichtigsten Verteilung von allen.

Einige der anderen Verteilungen in der folgenden Tabelle sind von der Normalverteilung abgeleitet: Die Chiquadratverteilung ergibt sich als die Verteilung der Summe

$$X_1^2 + \dots + X_n^2,$$

wobei $(X_1, \dots, X_n)$ unabhängig standardnormalverteilt $(N(0,1))$ sind.

Die t-Verteilung ist die Verteilung von $X/\sqrt{Y/n}$ mit X, Y unabhängig $X \sim N(0,1)$, $Y \sim \chi^2_n$.

Die F-Verteilung ist die Verteilung von $\frac{X/n}{Y/m}$, X, Y unabhängig, $X \sim \chi^2_n$, $Y \sim \chi^2_m$.

Diese Verteilungen sind in der Statistik von großer Bedeutung. Ihre Parameter (n bei χ^2 und t, n und m bei F) werden gemeinhin als "Freiheitsgrade" bezeichnet.

Die Erlangverteilung $Er(n,\lambda)$ ist die Verteilung einer Summe von n unabhängigen exponentialverteilten Zufallsvariablen mit Parameter λ.

Für die Gammaverteilung gibt es ein Additionstheorem: sind $X \sim \Gamma(\alpha, \lambda)$ und $Y \sim \Gamma(\beta, \lambda)$ unabhängig, dann gilt $X + Y \sim \Gamma(\alpha + \beta, \lambda)$. Unter denselben Voraussetzungen sind die Betaverteilungen als die Verteilungen von Quotienten zu erhalten: $X/(X+Y) \sim B_1(\alpha, \beta)$, $X/Y \sim B_2(\alpha, \beta)$.

Die Exponentialverteilung kann als Grenzfall aus der geometrischen Verteilung erhalten werden und steht in enger Verbindung mit der geometrischen Verteilung; die Details dazu werden im nächsten Kapitel erörtert. Dort, speziell in der Theorie der Markovketten, hat die Exponentialverteilung eine zentrale Rolle.

Name	Symbol	$f(x)$	$\mathbb{E}(X)$	$\mathbb{V}(X)$
Gleichverteilung (stetig) $a < b$	$U(a,b)$	$\frac{1}{b-a}(a \le x \le b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential $\lambda > 0$	$E(\lambda)$	$\lambda e^{-\lambda x}(x \ge 0)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma $\alpha, \lambda > 0$	$\Gamma(\alpha, \lambda)$	$\frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x}(x \ge 0)$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Cauchy $a > 0$	$C(a)$	$\frac{a}{\pi(x^2+a^2)}$	N.A.	N.A.
Normal $\mu \in \mathbb{R},\ \sigma^2 > 0$	$N(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$	μ	σ^2
Beta 1. Art $\alpha, beta > 0$	$B_1(\alpha, \beta)$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}(0 \le x \le 1)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$
Beta 2. Art $\alpha, beta > 0$	$B_1(\alpha, \beta)$	$\frac{x^{\alpha-1}(1+x)^{-\alpha-\beta}}{B(\alpha,\beta)}(0 \le x)$	$\frac{\alpha}{\beta-1}\ (\beta > 1)$	$\frac{\alpha(\alpha+\beta-1)}{(\beta-2)(\beta-1)^2}\ (\beta > 2)$
Chiquadrat $n \in \mathbb{N}$	χ^2_n	$= \Gamma(n/2, 1/2)$		
Erlang $n \in \mathbb{N},\ \lambda > 0$	$Er(n, \lambda)$	$= \Gamma(n, \lambda)$		
t-Verteilung $n \in \mathbb{N}$	t_n	$\frac{2}{\sqrt{n}B(n/2,1/2)(1+x^2/n)^{(n+1)/2}}$	$0\ (n > 1)$	$\frac{n}{n-2}\ (n > 2)$
F-Verteilung $m, n \in \mathbb{N}$	$F_{n.m}$	$\frac{x^{n/2-1} n^{n/2} (1+nx/m)^{-(m+n)/2}}{m^{n/2} B(n/2, m/2)}(0 \le x)$	$\frac{m}{m-2}\ (m > 2)$	$\frac{m^2(m+n-2)}{n(m-4)(m-2)^2}\ (m > 4)$

2.9 Wiederholungsfragen

1. Was ist ein Wahrscheinlichkeitsraum?
2. Was ist ein Ereignis?
3. Was besagen die Axiome von Kolmogorov?
4. Welche Eigenschaften von Wahrscheinlichkeitsmaßen kennen Sie?
5. Wie kann man die Wahrscheinlichkeit einer Vereinigung berechnen?
6. Wie ist die bedingte Wahrscheinlichkeit definiert?
7. Wie kann man die Wahrscheinlichkeit eines Durchschnitts berechnen?
8. Wann heißen zwei Ereignisse unabhängig?
9. Wann heißen mehr als zwei Ereignisse unabhängig?
10. Wann heißen mehr als zwei Ereignisse paarweise unabhängig?
11. Was besagt der Satz von der vollständigen Wahrscheinlichkeit?
12. Was besagt der Satz von Bayes?

13. Was ist eine Zufallsvariable?
14. Welche Typen von Zufallsvariablen gibt es?
15. Wie kann man die Verteilung einer Zufallsvariable angeben?
16. Welche Eigenschaften hat eine Verteilungsfunktion?
17. Welche Eigenschaften hat eine Wahrscheinlichkeitsfunktion?
18. Welche Eigenschaften hat eine Dichtefunktion?
19. Wann heißen zwei/mehrere Zufallsvariable unabhängig?
20. Was versteht man unter einer "Randverteilung"?
21. Wie kann man die Dichte einer transformierten Zufallsvariable bestimmen?
22. Wie berechnet man die bedingte Dichte?
23. Wie ist der Erwartungswert einer Zufallsvariable definiert?
24. Wie ist die Varianz einer Zufallsvariable definiert?
25. Wie ist die Kovarianz von zwei Zufallsvariablen definiert?
26. Welche Eigenschaften des Erwartungswerts kennen Sie?
27. Welche Eigenschaften der Varianz kennen Sie?
28. Was besagt die Ungleichung von Markov?
29. Was besagt die Ungleichung von Chebychev?
30. Was besagt die Ungleichung von Kolmogorov?
31. Wie sind die Momente einer Zufallsvariable definiert?
32. Was ist die momentenerzeugende Funktion einer Zufallsvariable?
33. Was ist die wahrscheinlichkeitserzeugende Funktion einer diskreten Zufallsvariable?
34. Was ist die charakteristische Funktion einer Zufallsvariable?
35. Was besagt das schwache Gesetz der großen Zahlen?
36. Was besagt das starke Gesetz der großen Zahlen?
37. Was besagt der zentrale Grenzwertsatz?

Kapitel 3

Stochastische Prozesse

Beispiel 3.1: Erneuerungsprozess

Wir betrachten ein technisches Gerät, etwa einen Computer oder auch nur eine Glühbirne, das ununterbrochen in Betrieb ist. Es hat eine gewisse Lebensdauer, und nach deren Ablauf wird es kaputt. Wir nehmen an, dass dann unverzüglich ein neues Gerät vom selben Typ installiert wird und dass die einzelnen Lebensdauern unabhängig und identisch verteilt sind. Wenn wir die verbrauchten Geräte von 1 an durchnummerieren und die Lebensdauer des n-ten Geräts mit T_n bezeichnen, dann wird das Gerät mit Nummer n zum Zeitpunkt S_{n-1} in Betrieb genommen und fällt zur Zeit S_n aus, wobei wir wie üblich $S_n = T_1 + \ldots + T_n$ (und $S_0 = 0$) setzen. Wir können diese Situation auch anders betrachten und uns nach der Anzahl N_t von Geräten fragen, die bis zum Zeitpunkt t verbraucht worden sind. Klarerweise gilt

$$N_t = n \text{ wenn } S_n \leq t < S_{n+1}$$

(hier könnte man noch darüber diskutieren, wie der Wert an den Sprungstellen S_n gewählt werden soll, aber unsere Definition, die N_t als Funktion von t rechtsstetig macht, ist die übliche). Wir haben hier also für jedes reelle $t \geq 0$ eine Zufallsvariable N_t, also eine ganze Familie von Zufallsvariablen $(N_t, t \geq 0)$. Eine solche Familie von Zufallsvariablen nennen wir einen stochastischen Prozess:

Definition 3.1

Ein stochastischer Prozess ist eine Familie $(X_t, t \in T)$ von Zufallsvariablen. Die Indexmenge T wird Parameterraum genannt und soll eine Teilmenge der reellen Zahlen sein. Der Wertebereich Ω_X von X_t heißt Zustandsraum oder Phasenraum. Wenn T endlich oder abzählbar (etwa $\mathbb{N}$) ist, sprechen wir von einem Prozess in diskreter Zeit, wenn T ein ganzes (endliches oder unendliches) Intervall ist, von einem Prozess in stetiger Zeit,

Das spezielle Modell, das wir in unserem ersten Beispiel beschrieben haben, bekommt gleich einen Namen:

Definition 3.2

(T_n) sei eine Folge von unabhängigen identisch verteilten Zufallsvariablen mit $T_n > 0$. Wir setzen $S_0 = 0$, $S_n = T_1 + \ldots + T_n$, und für $n = 0, \ldots$

$$X(t) = n \text{ für } S_n \leq t < S_{n+1}.$$

Wir nennen X einen Erneuerungsprozess.

Stochastische Prozesse in diskreter Zeit sind einfach Folgen von Zufallsvariablen. Der Unterschied zu unseren früheren Überlegungen besteht darin, dass wir nicht mehr annehmen, dass die einzelnen

Zufallsvariablen voneinander unabhängig sind. Wir müssen also die Abhängigkeiten zwischen den einzelnen Zufallsvariablen festlegen, das heißt, wir müssen gewisse Annahmen über die gemeinsame Verteilung von $(X_{t_1}, \ldots, X_{t_n})$ mit $t_1 < \ldots < t_n$ treffen. In einem gewissen Sinn wird ein stochastischer Prozess durch diese "endlichdimensionalen Randverteilungen" festgelegt.

Mathematisch streng genommen ist die Zufallsvariable X_t selbst eine Funktion, die aus einem Wahrscheinlichkeitsraum Ω in die reellen Zahlen abbildet, also haben wir es mit einer Funktion $X_t(\omega)$ von zwei Variablen zu tun. Unsere Definition eines stochastischen Prozesses als Familie von Zufallsvariablen bedeutet nichts anderes als dass wir diese Funktion für fixes t als Funktion von ω ansehen. Wir können uns aber auch die umgekehrte Betrachtungsweise zu eigen machen und $X_t(\omega)$ für fixes ω als Funktion von t ansehen:

Definition 3.3

$X_t(\omega)$ sei ein stochastischer Prozess. Wir nennen die Abbildung $t \mapsto X_t(\omega)$ als (als Funktion von t für fixes ω) einen Pfad, eine Trajektorie bzw. eine Realisation des Prozesses.

Diese Betrachtungsweise ist vor allem in stetiger Zeit interessant: dort wünscht man sich natürlich, dass die Trajektorien eines Prozesses stetig oder zumindest rechtsstetig sind. In unseren Betrachtungen werden diese Fragen nicht allzuviel Raum einnehmen, weil wir es nur mit Prozessen zu tun bekommen, bei denen die Zeit oder der Zustandsraum diskret ist, und da sind die Trajektorien notgedrungen Stückweise konstant mit Sprüngen zwischen den einzelnen konstanten Abschnitten, aber wir können (und werden) sie rechtsstetig wählen.

Die Definition eines stochastischen Prozesses ist so allgemein, dass daraus noch nicht allzu viele Schlüsse gezogen werden können. Damit wir mehr damit anfangen können, definieren wir unterschiedliche Typen oder Klassen von stochastischen Prozessen, die yusätzliche Struktur mitbringen. Die meisten Eigenschaften, die unsere Klassen von stochastischen Prozessen definieren, sind von Eigenschaften einer Folge X_n von unabhängigen Zufallsvariablen oder der Folge ihrer Partialsummen $S_n = X_1 + \ldots + X_n$ abgeleitet. Das direkte Analog zu einer Folge von unabhängigen Zufallsvariablen wäre ein stochastischer Prozess, bei dem alle X_t unabhängig sind. In stetiger Zeit sind solche Prozesse ziemlich unangenehm, beispielsweise sind ihre Trajektorien in allen Punkten unstetig, ja nicht einmal einseitig stetig.

3.1 Stationäre Prozesse

Wir betrachten eine Folge $(X_1, X_2, \ldots)$ von unabhängigen identisch verteilten Zufallsvariablen. Wenn wir vom Anfang dieser Folge ein paar Glieder entfernen, dann ist etwa $(X_n, X_{n+1}, \ldots)$ ebenfalls eine Folge von unabhängig identisch verteilten Zufallsvariablen mit derselben gemeinsamen Verteilung. Durch diese Verschiebung ändert sich also die Verteilung der Folge nicht. Diese Eigenschaft können wir für sich betrachten:

Definition 3.4

Der Prozess $X_t, t \in T$ heißt stationär, wenn wenn für $t_1 < t_2 < \ldots < t_n$ und $h > 0$ die gemeinsame Verteilung von $(X(t_1), \ldots, X(t_n))$ mit der von $(X(t_1 + h), \ldots, X(t_n + h))$ übereinstimmt.

Diese simple Forderung hat weitreichende Konsequenzen: sie genügt, um einen Satz im Geiste des Gesetzes der großen Zahlen zu beweisen:

Satz 3.1: Ergodensatz von Birkhoff

Wenn die Folge (X_n) stationär ist und endlichen Erwartungswert $m = \mathbb{E}(X_n)$ hat, dann existiert

$$X_\infty = \lim_{n \to \infty} \bar{X}_n$$

mit Wahrscheinlichkeit 1 und

$$\mathbb{E}(X_\infty) = m.$$

Es gilt also eine ähnliche Aussage wie im Gesetz der großen Zahlen, allerdings ist der Grenzwert im allgemeinen eine Zufallsvariable. Wenn er deterministisch ist, muss er natürlich gleich $\mathbb{E}(X_1)$ sein. Stationäre Folgen, in denen dieser Grenzwert deterministisch ist (nicht nur für X_n selbst, sondern auch für alle beschränkten Funktionen $f(X_n, \dots, X_{n+k})$), heißen ergodisch.

Ein sehr einfaches Beispiel für einen nicht ergodischen Prozess ist eine konstante Folge, also $X_n = X_1$ mit einer beliebigen Zufallsvariable X_1. Dann ist natürliche auch $\bar{X}_n = X_1$ und auch $X_\infty = 1$. Auf der anderen Seite ist eine Folge von unabhängigen identisch verteilten Zufallsvariablen ergodisch.

Wir werden uns hier nicht weiter mit der Theorie der stationären Prozesse auseinandersetzen, gelegentlich werden sie in den folgenden Kapiteln (speziell bei der Informationstheorie) wieder auftauchen.

3.2 Markovprozesse

Die nächste Definition leitet sich nicht direkt von der Folge (X_n) von unabhängigen Zufallsvariablen ab, sondern von der Folge der Partialsummen $S_n = X_1 + \ldots + X_n$. Wenn wir für $m > n$ die Differenz $S_m - S_n$ betrachten, dann hängt diese nur von den Zufallsvariablen $X_{n+1}, \dots, X_m$ ab, und die sind unabhängig von den Variablen $X_1, \dots, X_n$, die in S_n stecken. S_n und $S_m - S_n$ sind also unabhängig. Genauso kann man sich überlegen, dass für $n_1 < n_2 < \ldots < n_k$ die Zufallsvariablen

$$S_{n_1}, S_{n_2} - S_{n_1}, \dots, S_{n_k} - S_{n_{k-1}}$$

unabhängig sind. Diese Eigenschaft kann man fast wörtlich übertragen:

Definition 3.5

Der Prozess $(X_t, t \in T)$ heißt Prozess mit unabhängigen Zuwächsen, wenn für $t_1 < t_2 < \ldots < t_n$ die Zufallsvariablen

$$X(t_1), X(t_2) - X(t_1), X(t_3) - X(t_2), \dots, X(t_n) - X(t_{n-1})$$

unabhängig sind.

Wie wir schon am Anfang erwähnt haben, verwenden wir die Notation $X(t)$ gleichbedeutend mit X_t, weil komplexe Ausdrücke im Index schlecht lesbar sind.

Etwas allgemeiner kann man sagen, dass die Verteilung von

$$S_m = S_n + S_m - S_n$$

von den Werten von S_i mit $i \leq n$ nur durch den Wert von S_n abhängt. Diese Eigenschaft heißt die Markoveigenschaft:

Definition 3.6

Der Prozess $X(t)$ heißt Markovprozess, wenn für $t_1 < t_2 < \ldots < t_n$

$$\mathbb{P}(X(t_n) \leq x_n | X(t_1) = x_1, \dots, X(t_{n-1}) = x_{n-1}) = \mathbb{P}(X(t_n) \leq x_n | X(t_{n-1}) = x_{n-1}).$$

Die Zukunft hängt also von der Vergangenheit nur über den letzten Wert $X(t_{n-1})$ ab. Ein Beispiel für Markovprozesse sind natürlich Prozesse mit unabhängigen Zuwächsen.

Wir sehen uns ein einfaches Beispiel für einen Prozess in stetiger Zeit an. So wie schon bei anderer Gelegenheit beginnen wir mit einem diskreten Modell und machen es immer feiner (wir erhöhen schrittweise die Auflösung): Wir teilen das unendliche Intervall $[0, \infty[$ in Teilintervalle der Länge 2^{-N}, also $[0, 2^{-N}[, [2^{-N}, 2 \cdot 2^{-N}[, \ldots$. Jedes Intervall markieren wir mit Wahrscheinlichkeit $\lambda 2^{-N}$

unabhängig von allen anderen (wir müssen natürlich N so groß wählen, dass diese Wahrscheinlichkeit kleiner als 1 ist). Für jedes N ist die mittlere Anzahl von Markierungen (ihr Erwartungswert) im Intervall $[0,1]$ gleich λ. Dasselbe gilt zumindest näherungswese für jedes Intervall der Länge 1; es stimmt für ein bestimmtes N genau, wenn die Endpunkte des Intervalls ganzzahlige Vielfache von 2^{-N} sind. Wir müssen uns aber nicht mit Erwartungswerten begnügen, wir wissen ja, dass die Anzahl der Markierungen im Intervall $[0,1[$ binomialverteilt ist mit Parametern 2^N und $\lambda 2^{-N}$. Für $N \to \infty$ konvergiert das gegen eine Poissonverteilung mit Parameter λ, wie wir in den Übungen gezeigt haben. Genauso kann man erhalten, dass für irgendein Itervall der Länge t die Anzahl der Markierungen darin für großes N näherungsweise Poissonverteilt ist mit Parameter $t\lambda$. Klarerweise sind die Anzahlen in disjunkten Intervallen unabhängig (zumindest wenn sie positiven Abstand voneinander haben; wenn nicht, dann gibt es nur eines der kleinen Intervalle, dessen Markierung bei beiden zählt, also ist die Abhängigkeit zwischen beiden Anzahlen nur schwach). Für $N \to \infty$ ergibt sich für $X(t)$, die Anzahl der Markierungen zwischen 0 und t, dass die Anzahl $X(t) - X(s)$ von Markierungen zwischen s und t, von allem unabhängig ist, was vor s passiert. Wir haben also einen Prozess mit unabhängigen Zuwächsen, bei dem $X(t) - X(s)$ mit der Poissonverteilung $P(\lambda(t-s))$ verteilt ist. Diesen Prozess nennen wir den Poissonprozess:

Definition 3.7

Der Poissonprozess mit Rate $\lambda > 0$ ist ein Prozess mit unahängigen Zuwächsen, mit $X(0) = 0$ und $X(t) - X(s) \sim P(\lambda(t-s))(s < t)$.

Wir können diesen Prozess auch anders ansehen: Wir zählen die Anzahl Y_1 von kleinen Intervallen, die wir durchlaufen müssen, bis wir die erste Markierung finden, danach zählen wir die Anzahl Y_2, die zwischen der ersten und der zweiten Markierung liegen, usw. Die Position der ersten Markierung ist dann $T_1 = 2^{-N}Y_1$, der Abstand zwischen erster und zweiter $T_2 = 2^{-N}Y_2$, etc. $Y_1, Y_2, \ldots$ sind unabhängig und geometrisch verteilt mit Erfolgswahrscheinlichkeit $\lambda 2^{-N}$, also

$$\mathbb{P}(Y_i > k) = (1 - \frac{\lambda}{2^N})^k$$

und daher

$$\mathbb{P}(T_i > x) = (1 - \frac{\lambda}{2^N})^{x2^N},$$

und das konvergiert für $N \to \infty$ gegen $e^{-\lambda x}$. Wir kommen so zu einer Folge $T_1, T_2, \ldots$ von unabhängigen exponentialverteilten Zufallsvariablen, und $X(t) = 0$ für $t < T_1$, $X(t) = 1$ für $T_1 \leq t < T_1 + T_2$, $X(t) = k$ für $T_1 + \ldots + T_k \leq t < T_1 + \ldots + T_{k+1}$. Wir können den Poissonprozess also auch als Erneuerungsprozess mit exponentialverteilten Lebensdauern definieren.

3.3 Markovketten in diskreter Zeit

3.3.1 Übergangswahrscheinlichkeiten

Markovprozesse mit diskretem Zustandsraum nennen wir Markovketten. Wir können noch zwischen Markovketten in diskreter und in stetiger Zeit unterscheiden. In beiden Fällen kann man die diskrete Verteilung der einzelnen Zufallsvariablen durch ihre Wahrscheinlichkeitsfunktion beschreiben, deshalb erhält die Markoveigenschaft die besonders einfache Form

$$\mathbb{P}(X_n = x_n | X_1 = x_1, \ldots, X_{n-1} = x_{n-1}) = \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}).$$

Definition 3.8

Die Wahrscheinlichkeiten

$$\mathbb{P}(X_{n+1} = j | X_n = i)$$

nennen wir die Übergangswahrscheinlichkeiten der Markovkette. Wenn diese nicht von n

abhängen, sprechen wir von einer homogenen Markovkette und setzen

$$p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i)$$

Die Wahrscheinlichkeiten

$$p_{ij}(t) = \mathbb{P}(X_{n+t} = j | X_n = i)$$

nennen wir die t-stufigen Übergangswahrscheinlichkeiten.

Wir können eine homogene Markovkette als *dynamisches System* betrachten:

Definition 3.9

Wir sagen, dass die Folge (X_n) ein (stochastisches) dynamisches System bildet, wenn es eine Folge (Y_n) von unabhängigen Zufallsvariablen und eine Funktion $f : \Omega_X \times \mathbb{R} \to \Omega_X$ gibt, sodass

$$X_{n+1} = f(X_n, Y_n)$$

gilt.

Es ist leicht einzusehen, dass ein solches dynamisches System eine Markovkette bildet, umgekehrt können wir jede Markovkette mit Übergangsmatrix P als dynamisches System erhalten, indem wir (Y_n) als Folge von unabhängigen $U(0,1)$-verteilten Zufallsvariablen ansetzen, und $f(i,y) = j$ setzen, wenn

$$\sum_{k<j} p_{ik} < y \leq \sum_{k \leq j} p_{ik}.$$

Aus dem Satz von der vollständigen Wahrscheinlichkeit erhalten wir

Satz 3.2: Chapman-Kolmogorov Gleichungen

$$p_{ij}(s+t) = \sum_{k \in \Omega_X} p_{ik}(s) p_{kj}(t).$$

in Matrixnotation mit den *t-stufigen Übergangsmatrizen*

$$P(t) = (p_{ij}(t))_{\Omega_X \times \Omega_X}$$

lauten die Chapman-Kolmogorov Gleichungen

$$P(t+s) = P(t)P(s).$$

Für die einstufige Übergangsmatrix schreiben wir kurz $P = P(1)$ und nennen sie einfach die "Übergangsmatrix". Die Chapman-Kolmogorov Gleichungen liefern

$$P(t) = P^t.$$

Wir setzen zusätzlich für die Verteilung von X_t $p_i(t) = \mathbb{P}(X_t = i)$ und $p(t) = (p_i(t), i \in \Omega_X)$ (als Zeilenvektor). Wieder mit dem Satz von der vollständigen Wahrscheinlichkeit erhalten wir

$$p(t) = p(0)P^t.$$

Durch $p(0)$ und P werden alle endlichdimensionalen Verteilungen festgelegt.

Beispiel 3.2

Der einfachste nichttriviale Fall ist eine Markovkette mit zwei Zuständen. Ihre Übergangsmatrix hat die allgemeine Form

$$P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}.$$

Die Potenzen dieser Matrix berechnen wir mithilfe der Eigenwertzerlegung: Wenn A diagonalisierbar ist, als so viele Eigenvektoren hat wie Zeilen/Spalten, dann gilt die Darstellung

$$A = X\Lambda X^{-1},$$

wobei X die Matrix aus den (Spalten-) Eigenvektoren und Λ die Diagonalmatrix der Eigenwerte ist. Die Matrix P hat einen Eigenwert 1, den es bei jeder Übergangsmatrix gibt, der zweite ist $1-(a+b)$. Die Eigenvektoren dazu sind $\binom{1}{1}$ und $\binom{a}{-b}$. Insgesamt ergibt sich

$$P^t = \frac{1}{a+b}\left(\begin{pmatrix} b & a \\ b & a \end{pmatrix} + (1-a-b)^t \begin{pmatrix} a & -a \\ -b & b \end{pmatrix}\right).$$

Diese Darstellung gilt natürlich nur, wenn $a+b>0$. Der verbleibende Fall $a=b=0$ ist der Leserin/dem Leser überlassen.

Ein einfaches Beispiel für eine Markovkette mit unendlichem Zustandsraum bietet der Random Walk:

Definition 3.10

Der einfache Random Walk (auch: einfache Irrfahrt) ist der Partialsummenprozess

$$X(t) = X(0) + \sum_{i=1}^{t} Y_i,$$

wobei $(Y_1, Y_2, \ldots)$ unabhängige Zufallsvariable mit

$$\mathbb{P}(Y_i = 1) = p, \mathbb{P}(Y = -1) = 1-p$$

sind. Der Spezialfall $p = 1/2$ heißt einfacher symmetrischer Random Walk.

Beispiel 3.3

Die einstufigen Übergangswahrscheinlichkeiten des einfachen Random Walk sind

$$p_{ij} = \begin{cases} p & \text{wenn } j = i+1, \\ 1-p & \text{wenn } j = i-1, \\ 0 & \text{sonst.} \end{cases}$$

Für unendlich Matrizen gibt es leider keine so praktische Methode zur Berechnung der Potenzen wie die Eigenwertzerlegung für endliche Matrizen. Es sind von Fall zu Fall individuelle Methoden notwendig. Für den Random Walk hilft uns unser Wissen über die Binomialverteilung: um in n Schritten von i zu j zu gelangen, muss die Differenz zwischen der Anzahl der Schritte $+1$ und der Anzahl der Schritte -1 Wert $j-i$ haben, und weil ihre Summe n ist, ergibt sich

$$p_{ij}(n) = \binom{n}{\frac{n+j-i}{2}} p^{\frac{n+j-i}{2}} (1-p)^{\frac{n+i-j}{2}},$$

wenn $n+j-i$ gerade ist.

Wird die spezielle Annahme an die Verteilung von Y_i fallengelassen (wenn also nur angenommen wird, dass (Y_n) eine Folge von unabhängigen, identisch verteilten Zufallsvariablen mit ganzzahligen Werten ist), dann ist $(X(t))$ ein allgemeiner Random Walk. Eine zweite Verallgemeinerung ist der *inhomogene Random Walk*, der auch als diskreter Geburts- und Todesprozess bezeichnet wird: hier werden die Übergänge zu den Nachbarzuständen beibehalten, aber die Wahrscheinlichkeiten hängen vom Zustand ab, also $p_{i,i+1} = p_i$, $p_{i,i-1} = q_i = 1-p_i$.

In Hinkunft verwenden wir für Wahrscheinlichkeiten und Erwartungswerte im Zusammenhang mit Markovketten, die ja von der konkreten Wahl des Anfangszustands abhängen, zur Abkürzung

die Notationen

$$\mathbb{P}_i(A) = \mathbb{P}(A|X_0 = i)$$

und

$$\mathbb{E}_i(Y) = \mathbb{E}(Y|X_0 = i).$$

3.3.2 Klasseneigenschaften

Wir definieren

Definition 3.11

Der Zustand j heißt Nachfolger von i ($i \to j$), wenn j von i aus mit positiver Wahrscheinlichkeit erreicht werden kann, also, wenn es ein $t \geq 0$ gibt, sodass $p_{ij}(t) > 0$.

Wenn sowohl $i \to j$ als auch $j \to i$ gilt, dann heißen i und j verbunden oder kommunizierend.

Das Kommunizieren ist eine Äquivalenzrelation, wir können daher den Phasenraum in die Äquivalenzklassen zerlegen, die wir Rekurrenzklassen oder kurz Klassen nennen. Gibt es nur eine Klasse (wenn also alle Zustände miteinander kommunizieren), heißt die Markovkette irreduzibel. Ein Zustand mit $p_{ii} = 1$ heißt absorbierender Zustand. Ein solcher Zustand ist offensichtlich eine Klasse für sich.

Definition 3.12

Eine Eigenschaft heißt Klasseneigenschaft, wenn sie entweder für alle Zustände einer Klasse oder für keinen gilt.

Ein einfaches Beispiel einer Klasseneigenschaft ist die Periode:

Definition 3.13

Die Periode eines Zustandes ist

$$d(i) = \text{ggT}\{t \geq 0 : p_{ii}(t) > 0\}.$$

Wenn $d(i) = 1$ gilt, dann heißt der Zustand i aperiodisch, sonst periodisch.

Beispiel 3.4

Der einfache Random Walk ist periodisch mit Periode 2. Wenn wir beispielsweise in 0 starten, dann ist $X(t)$ für gerades t gerade und für ungerades t ungerade, also ist $X(t) = 0$ nur für gerades t möglich.

Dass die Periode eine Klasseneigenschaft ist, ist einfach zu beweisen: wenn i und j verbunden sind, dann gibt es t_1 und t_2 mit $p_{ij}(t_1) > 0$ und $p_{ji}(t_2) > 0$. Weiters sei t eine Zahl mit $p_{jj}(t) > 0$. Es gilt

$$p_{ii}(t_1 + t_2) \geq p_{ij}(t_1)p_{ji}(t_2) > 0$$

und

$$p_{ii}(t_1 + t + t_2) \geq p_{ij}(t_1)p_{jj}(t)p_{ji}(t_2) > 0,$$

deshalb gilt

$$d(i) \,|\, t_1 + t_2$$

und

$$d(i) \,|\, t_1 + t + t_2,$$

deshalb auch $d(i) \,|\, t$. Weil das für jedes t mit $p_{jj}(t) > 0$ gilt, ergibt sich $d(i) \,|\, d(j)$. Weil auch umgekehrt $d(j) \,|\, d(i)$ ergibt sich $d(i) = d(j)$.

Etwas spannender ist die Rekurrenz: wir fragen uns nach dem Rückkehrverhalten unserer Markovkette. Ansatzweise ist das schon geschehen, bei der Nachfolger- und Verbundenheitsrelation geht es ja darum, ob man einen Zustand j von einem Zustand i aus erreichen kann (oder ob man wieder nach i zurückkehren kann). Jetzt fragen wir uns, ob der Prozess zurückkehren *muss*, oder ob es eine positive Wahrscheinlichkeit gibt, dass der Prozess nicht zurückkehrt.

Wir nennen einen Zustand i rekurrent, wenn der Prozess, ausgehend von i, mit Wahrscheinlichkeit 1 wieder nach i zurückkehren muss. Es ist intuitiv klar, dass in diesem Fall der Prozess den Zustand i unendlich oft besucht. Im nächsten Satz wird diese Vorstellung präzise gemacht. Dazu definieren wir zuerst

$$\tau_i = \inf\{t > 0 : X_t = i\},$$

die Übergangs- bzw. Rückkehrzeit (je nachdem, ob $X_0 \neq i$ ist oder nicht) nach i, und

$$\nu_i = \#\{t > 0 : X_t = i\},$$

die Anzahl der Besuche in i.

Satz 3.3

Die folgenden Bedingungen sind äquivalent:

1. $\mathbb{P}_i(\tau_i < \infty) = 1$,
2. $\mathbb{P}_i(\nu_i = \infty) = 1$,
3. $\mathbb{E}_i(\nu_i) = \infty$,
4. $\sum_t p_{ii}(t) = \infty$.

Wenn diese Bedingungen erfüllt sind, dann heißt i rekurrent, sonst transient. Rekurrenz und Transienz sind Klasseneigenschaften.

Beispiel 3.5

Für den Random Walk haben wir schon die n-stufigen Übergangswahrscheinlichkeiten berechnet. Speziell gilt

$$p_{ii}(2n) = \binom{2n}{n}(p(1-p))^n.$$

Mithilfe der Stirling-Formel ergibt sich

$$p_{ii}(2n) \approx \frac{(4p(1-p))^n}{\sqrt{\pi n}}.$$

Falls $p \neq 1/2$, dann ist $4p(1-p) < 1$, und

$$\sum_n p_{ii}(2n) < \infty,$$

also ist in diesem Fall der Random Walk transient. Für $p = 1/2$ gilt $p_{ii}(2n) \approx 1/\sqrt{\pi n}$, also

$$\sum_n p_{ii}(2n) = \infty,$$

und der Random Walk ist rekurrent.

Der transiente Fall lässt sich auch aus dem Gesetz der großen Zahlen ableiten. Dieses gibt uns nämlich mit Wahrscheinlichkeit 1

$$X_n/n \to 2p - 1.$$

Für $p > 1/2$ gilt also $X_n \to \infty$, für $p < 1/2$ $X_n \to -\infty$.

Beispiel 3.6

Jede irreduzible endliche Markovkette ist rekurrent. Es ist nämlich für einen fixen Zustand i und alle n

$$\sum_j p_{ij}(n) = 1$$

und daher

$$\sum_j \sum_n p_{ij}(n) = \sum_j \sum_n p_{ij}(n) = \infty.$$

Es muss also eine der endlich vielen Summen $\sum_n p_{ij}(n)$ unendlichen Wert haben. Wir können ein k mit $p_{ji}(k) > 0$ wählen und erhalten

$$\sum_n p_{jj}(n) \geq \sum_{n \geq k} p_{jj}(n) \geq \sum_{n \geq k} p_{ji}(k) p_{ij}(n-k) = \infty.$$

Wir haben also einen rekurrenten Zustand gefunden, und daher ist der ganze (irreduzible) Prozess rekurrent.

Im allgemeinen ist in endlichen Markovketten jede Klasse, aus der keine Verbindungen herausführen (die also keine Nachfolger hat) rekurrent, jede Klasse, aus der es Verbindungen nach außen gibt, transient. In unendlichen Ketten ist das nicht richtig, wie das Beispiel des asymmetrischen Random Walks zeigt.

Bei der Rekurrenz kann man weiter unterscheiden:

Definition 3.14

i sei ein rekurrenter Zustand. Wenn

$$\mathbb{E}_i(\tau_i) < \infty$$

gilt, dann heißt i positiv rekurrent, sonst nullrekurrent.

Definition 3.15

$(\pi_i, i \in \Omega_X)$ heißt stationäre Verteilung, wenn

$$\pi_i \geq 0,$$

$$\sum_i \pi_i = 1$$

und

$$\sum_i \pi_i p_{ij} = \pi_j.$$

Satz 3.4

Wenn (X_n) irreduzibel und aperiodisch ist, dann existieren die Grenzwerte

$$\lim_{n \to \infty} p_{ij}(t) = \pi_j = \frac{1}{\mathbb{E}_j(\tau_j)}.$$

Im periodischen Fall gilt

$$\pi_j = \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} p_{ij}(t).$$

Wenn (π_i) nicht identisch verschwindet (also wenn die Kette positiv rekurrent ist), dann ist es eine stationäre Verteilung. Umgekehrt folgt aus der Existenz einer stationären Verteilung die positive Rekurrenz. Die positive Rekurrenz bzw. Nullrekurrenz ist ebenfalls eine Klasseneigenschaft.

Beispiel 3.7: Random Walk

Wir wissen bereits, dass der einfache symmetrische Random Walk rekurrent ist. Natürlich fragen wir uns jetzt, ob er positiv rekurrent oder nullrekurrent ist. Wir suchen also nach einer stationären Verteilung. Diese muss die Gleichungen

$$\pi_i = \frac{1}{2}(\pi_{i+1} + \pi_{i-1}).$$

Das lässt sich auch so schreiben:

$$\pi_{i+1} - \pi_i = \pi_i - \pi_{i-1},$$

also ist die Folge der Differenzen $d_i = \pi_{i+1} - \pi_i$ konstant, und, wenn wir ihren gemeinsamen Wert mit d bezeichnen, ergibt sich

$$\pi_i = \pi_0 + id.$$

Damit π_i nichtnegativ und beschränkt bleibt, muss $d = 0$ gelten, also π_i konstant, und ihre Summe kann nur endlich sein, wenn $\pi_i = 0$ gilt. Wir haben hier also Nullrekurrenz.

Beispiel 3.8

Eine irreduzible endliche Markovkette ist immer positiv rekurrent. In der endlichen Summe

$$\sum_j p_{ij}(n) = 1$$

kann man ohne weiteres $n \to \infty$ laufen lassen, und es ergibt sich

$$\sum_i \pi_i = 1.$$

Das bedeutet natürlich auch, dass Nullrekurrenz nur auftreten kann, wenn die entsprechende Klasse unendlich viele Zustände besitzt.

3.3.3 Absorptionswahrscheinlichkeiten

Für einen absorbierenden Zustand i_0 definieren wir die Absorptionswahrscheinlichkeit a_i als die Wahrscheinlichkeit, bei Start in i den Zustand i_0 zu erreichen, also dort absobiert zu werden:

$$a_i = \mathbb{P}_i(\tau_{i_0} < \infty) = \mathbb{P}_i(X\text{wird in } i_0 \text{ absorbiert}).$$

Satz 3.5

Die Absorptionswahrscheinlichkeiten sind die kleinste nichtnegative Lösung des Gleichungssystems

$$a_{i_0} = 1,$$
$$a_i = \sum_j p_{ij} a_j, i \neq i_0.$$

Das gibt uns eine Möglichkeit, die Transienz oder Rekurrenz einer irreduziblen Markovkette zu entscheiden. Wir wählen einen Zustand und machen ihn zu einem absorbierenden Zustand und

bestimmen für die modifizierte Übergangsmatrix die Absorptionswahrscheinlichkeiten. Sind diese gleich 1, ist die Kette rekurrent.

Beispiel 3.9: Random Walk mit Reflexion

Wir ändern die Dynamik des Random Walks, indem wir aus dem Yustand 0 nur nach rechts gehen können. Den Zustandsraum verkleinern wir entsprechend zu $\Omega_X = \mathbb{N}_0$. Die übergangswahrscheinlichkeiten sind also

$$p_{ij} = \begin{cases} 1 & \text{wenn } i = 0,\ j = 1, \\ p & \text{wenn } i \geq 1,\ j = i+1, \\ q = 1-p & \text{wenn } i \geq 1,\ j = i-1, \\ 0 & \text{sonst.} \end{cases}$$

Wir machen 0 zum absorbierenden Zustand, und bestimmen die Absorptionswahrscheinlichkeiten. Dafür haben wir die Gleichungen

$$a_0 = 1,$$

$$a_i = pa_{i+1} + qa_{i-1}.$$

Wir lösen diese Gleichungen mit dem Ansatz

$$a_i = z^i.$$

Das gibt die Gleichung

$$pz^2 - z + q = 0$$

mit den Lösungen $z_1 = 1$ und $z_2 = q/p$. Wenn $p \neq 1/2$, dann sind $a_n = z_1^n = 1$ und $a_n = z_2^n$ zwei linear unabhängige Lösungen der Rekursiongleichung, und wir können jede ihrer Lösungen in der Form

$$a_n = C_1 z_1^n + C_2 z_2^n = C_1 + C_2 (\frac{q}{p})^n$$

erhalten. Für $p < 1/2$ ist $z_2 > 1$, und damit a_n beschränkt bleibt, muss $C_2 = 0$ gelten. Dann ist $a_n = a_0 = 1$, und wir haben Rekurrenz. Für $p > 1/2$ ist $\lim_{n\to\infty} a_n = C_1$, und deshalb muss $0 \leq C_1$ gelten, außerdem ist $a_0 = C_1 + C_2 = 1$. Das gibt

$$a_n = z_2^n + C_1(1 - z_2^n),$$

und die kleinste Lösung ergibt sich für $C_1 = 0$. In diesem Fall ist der Prozess also transient.

Für $p = 1/2$ hat die quadratische Gleichung die Doppellösung $z_{1,2} = 1$. In diesem Fall erhlaten wir eine zweite Lösung mit $a_n = n$. Genauso wie im Fall $p < 1/2$ ergibt sich, dass $a_n = 1$ gelten muss, also der Prozess auch rekurrent ist.

Man kann weiter feststellen, dass für $p < 1/2$ der Prozess positiv rekurrent ist und für $p = 1/2$ nullrekurrent.

Wir nehmen an, dass i_0 der einzige absorbierende Zustand ist und alle anderen Zustände kommunizieren und die Absorptionswahrscheinlichkeiten 1 sind. Dann erhalten wir für die mittlere Zeit bis zur Absorption

$$m_i = \mathbb{E}_i(\tau_{i_0})$$

eine ähnliche Gleichung:

$$m_{i_0} = 0,$$

$$m_i = 1 + \sum_j p_{ij} m_j, i \neq i_0.$$

Allgemeiner kann man eine Zufallsvariable $S = \sum_{t=1}^{\tau-1} x_{t,t+1}$ betrachten (d.h., wir "sammeln" auf unserem Weg zur Absorption Beträge x_{ij}, die von dem jeweiligen Übergang abhängen dürfen).

Wir setzen

$$m_i = \mathbb{E}_i(S)$$

und

$$v_i = \mathbb{V}_i(S).$$

Dann sind m_i und v_i die Lösungen der Gleichungen

$$m_{i_0} = 0, v_i(i_0) = 0,$$

$$m_i = \sum_j p_{ij}x_{ij} + \sum_j p_{ij}m_j, i \neq i_0.$$

$$v_i = \sum_j p_{ij}(x_{ij} + m_j - m_i)^2 + \sum_j p_{ij}v_j, i \neq i_0.$$

Wenn es mehrere absorbierende Zustände gibt, $A = \{i_1, \ldots, i_k\}$, dann kann man die Wahrscheinlichkeiten der Absorption $a_i(i_j)$ in dem absorbierenden Zustand i_j oder allgemeiner in einem der Zustände in der Menge $B \subseteq A$ (wenn wir auf die konkrete Menge B hinweisen wollen, schreiben wir für a_i $a_i(B)$). Das gibt die Gleichungen

$$a_i = \sum_j p_{ij}a_j, i \notin A$$

$$a_i = 1, i \in B$$

$$a_i = 0, i \in A \setminus B.$$

Manchmal ist es interessant (wieviel Geld werde ich haben, wenn ich nicht bankrott gehe?), den Erwartungswert von S unter der Bedingung, dass die Absorption in einem bestimmten Zustand oder in einer bestimmten Teilmenge der Menge der absorbierenden Zustände erfolgt. Da hilft es, dass $X(t)$ unter der Bedingung der Absorption in B wieder eine Markovkette bildet, mit den Übergangswahrscheinlichkeiten

$$p_{ij}^B = \frac{p_{ij}a_j(B)}{a_i(B)}.$$

Mit diesen modifizierten Übergangswahrscheinlichkeiten (wobei die absorbierenden Zustände $\notin B$ weggelassen werden) kann man die Gleichungen für m_i und v_i lösen und erhält so die bedingte Erwartung bzw. Varianz.

Beispiel 3.10

Ein Münzwurfspiel: Angelina und Brad werfen eine Münze. Zeigt diese "Kopf", dann erhält Angelina einen Euro von Brad, bei Kopf gibt sie ihm einen Euro. Am Anfang hat Angelina a Euro, Brad b Euro. Das Spiel wird so lange wiederholt, bis Angelina oder Brad kein Geld mehr hat.

Bei diesem Spiel bildet das Kapital X_n von Angelina nach der n-ten Runde eine Markovkette: es verändert sich ja aufgrund eines Münzwurfes, der von der Vorgeschichte unabhängig ist. Die Übergangswahrscheinlichkeiten sind leicht bestimmt: aus einem der Zustände $1, \ldots, a + b - 1$ wird mit Wahrscheinlichkeit $1/2$ in einen der Nachbarzustände gesprungen, die Zustände 0 und $a + b$ sind absorbierend, weil bei ihrem Erreichen das Spiel endet, also

$$p_{i,i+1} = p_{i,i-1} = 1/2 \ (1 \leq i \leq a + b - 1),$$

$$p_{00} = p_{a+b,a+b} = 1,$$

alle anderen Übergangswahrscheinlichkeiten sind 0.

Als erstes bestimmen wir die Wahrscheinlichkeit, dass Angelina gewinnt: das ist die Absorptionswahrscheinlichkeit in $a + b$. Also setzen wir

$$a_0 = 0, a_{a+b} = 1,$$

$$a_i = \frac{a_{i+1} + a_{i-1}}{2} (i = 1, \ldots, a + b - 1.$$

Wir lösen die Rekursionsgleichung wieder mit dem Ansatz $a_i = z^i$. Die charakteristische Gleichung hat die doppelte Lösung $z = 1$, also ist $a_i = ci + d$. Die Koeffizienten c und d ergeben sich aus den Randbedingungen $a_0 = 0$ und $a_{a+b} = 1$, also

$$a_i = \frac{i}{a + b}.$$

Am Anfang hat Angelina a Euro, also ist ihre Gewinnwahrscheinlichkeit

$$a_a = \frac{a}{a + b}.$$

Als nächstes interessiert uns der Erwartungswert der Spieldauer, also die mittlere Absorptionszeit. Wir haben die Gleichungen

$$m_0 = m_{a+b} = 0,$$

$$m_i = \frac{m_{i+1} + m_{i-1}}{2} + 1(1 \leq i \leq a + b - 1).$$

Wir kennen schon die allgemeine Lösung der $\tilde{m}_n = cn + d$ homogenen Gleichung (in der der Störterm $+1$ nicht vorhanden ist). Wenn wir eine Lösung der inhomogenen Gleichung kennen, dann kann man alle Lösungen erhalten, indem man zu dieser einen Lösung eine beliebige Lösung der homogenen Gleichung addiert. Im allgemeinen lässt sich eine Lösung des inhomogenen Problems für Störterme der Form $P(n)z^n$ mit einem Polynom P in der Form $Q(n)z^n$ finden, wobei Q vom gleichen Grad wie P. Wenn z aber eine Nullstelle des charakteristischen Polynoms ist, muss der Grad von Q um die Vielfachheit dieser Nullstelle erhöht werden. In unserem Fall ist $z = 1$ (der Störterm 1 ist ja $1 \cdot 1^n$) zweifache Nullstelle des charakteristischen Polynoms, also Versuchen wir eine Lösung der Form $\bar{m}_n = kn^2$. Einsetzen ergibt $k = -1$. Diese Lösung erfüllt noch nicht alle Randbedingungen, aber weil alle Lösungen von der Form

$$m_n = \bar{m}_n + \tilde{m}_n = -n^2 + c_n + d$$

sein müssen, bleiben nur noch c und d zu bestimmen, und es ergibt sich

$$m_i = i(a + b - i),$$

und die gesuchte mittlere Spieldauer ist

$$m_a = ab.$$

Für die Varianz der Spieldauer erhalten wir (nach einigen Rechnungen)

$$v_i = \frac{v_{i-1} + v_{i+1}}{2} + (a + b - 2i)^2, 1 \leq i \leq a + b - 1,$$

$$v_0 = v_{a+b} = 0.$$

Diesmal ist der Störterm quadratisch, also müssen wir die Lösung als Polynom vierten Grades ansetzen. Nach einigem Rechnen ist

$$v_i = \frac{1}{3} i(a + b - i)(4i^2 + (a + b)^2 - 2).$$

Die letzte Frage lautet: wie lange dauert das Spiel im Durchschnitt, wenn Angelina gewinnt? Wir fragen also nach dem bedingten Erwartungswert der Absorptionszeit unter der Bedingung

der Absorption in $a+b$. Dazu berechnen wir die Übergangswahrscheinlichtkeiten für die "bedingte Markovkette"

$$p_{ij}^{\{a+b\}} = \frac{p_{ij}a_j}{a_i},\ i,j = 1,\dots,a+b$$

(den "falschen" absorbierenden Zustand 0 haben wir entfernen müssen), und wir lösen damit die Gleichungen

$$m_{a+b} = 0,$$

$$m_i = 1 + \sum_{i=1}^{a+b} p_{ij}m_j,\ i = 1,\dots,a+b-1.$$

In vielen Fällen (und auch hier) ist es günstig, statt m_i die Variablen $y_i = a_i x_i$ zu verwenden. Dann kann man nämlich die ursprünglichen p_{ij} verwenden, und wir müssen auch die "schlechten" absorbierenden Zustände nicht entfernen (weil sie automatisch $y_i = 0$ bekommen). Das gibt im allgemeinen Fall die Gleichungen

$$y_i = \begin{cases} 0 & \text{wenn } i \text{ absorbierend ist,} \\ \sum_j p_{ij}y_j + a_i x_{ij} & \text{sonst.} \end{cases}$$

In unserem Fall ist das

$$y_0 = y_{a+b} = 0,$$

$$y_i = \frac{y_{i+1}+y_{i-1}}{2} + a_i = \frac{y_{i+1}+y_{i-1}}{2} + \frac{i}{a+b},\ i = 1,\dots,a+b.$$

Diemal müssen wir y als Polynom dritten Grades ansetzen und erhalten

$$y_i = \frac{i((a+b)^2 - i^2)}{3(a+b)}$$

und

$$m_i^{\{a+b\}} = \frac{y_i}{a_i} = \frac{(a+b)^2 - i^2}{3},$$

und wieder speziell für $i = a$

$$m_a^{\{a+b\}} = \frac{b(2a+b}{3}.$$

3.3.4 Markov Chain Monte Carlo

3.4 Markovketten in stetiger Zeit

Wir betrachten jetzt Markovketten, bei denen die Zeit nicht mehr nur ganzzahlige, sondern beliebige reelle (nichtnegative) Werte annehmen darf. Wir bleiben bei unserer Annahme, dass es sich um einen homogenen Prozess handelt, also dass die Übergangswahrscheinlichkeiten

$$\mathbb{P}(X_t = j | X_s = i)$$

nur von der Differenz $t-s$ abhängen, und setzen wieder

$$p_{ij}(t) = \mathbb{P}(X_{s+t} = j | X(s) = i)$$

und fassen die Übergangswahrscheinlichkeiten zu den Übergangsmatrizen

$$P(t) = (p_{ij}(t))_{\Omega_X \times \Omega_X}$$

zusammen.

Wie im diskreten Fall gelten die Chapman-Kolmogorov Gleichungen

$$P(s+t) = P(s)P(t).$$

Leider gibt es im stetigen Fall nicht so wie in diskreter Zeit die Möglichkeit, alle Übergangsmatrizen als Potenzen einer einzigen Matrix zu erhalten. Wir nehmen daher Zuflucht zur Differentialrechnung. Zuerst verlangen wir die Stetigkeit der Übergangswahrscheinlichkeiten bei 0, also dass

$$\lim_{t\to 0} p_{ij}(t) = p_{ij}(0) = \begin{cases} 1 & \text{wenn } i = j, \\ 0 & \text{sonst.} \end{cases}$$

Diese Annahme genügt schon für die Existenz der Ableitungen:

Satz 3.6

Wenn die Übergangswahrscheinlichkeiten bei 0 stetig sind, dann existieren die Grenzwerte

$$q_{ij} = \lim_{t\to 0} \frac{p_{ij}(t) - p_{ij}(0)}{t}.$$

Die Zahlen q_{ij} heißen die infinitesimalen Parameter, die Matrix

$$Q = (q_{ij})_{\Omega_X \times \Omega_X}$$

die infinitesimale Matrix oder der infinitesimale Erzeuger der Markovkette.

Offensichtlich gilt $q_{ii} \leq 0$ und $q_{ij} \geq 0$ für $i \neq j$. Wenn der Wertebereich Ω_X endlich ist, dann können wir in den Gleichungen

$$\sum_j \frac{p_{ij}(t) - p_{ij}(0)}{t} = 0$$

und

$$\frac{P(t+s) - P(t)}{s} = P(t)\frac{P(s) - I}{s} = \frac{P(s) - I}{s}P(t).$$

und erhalten so

$$\sum_j q_{ij} = 0$$

und

Definition 3.16: Kolmogorovsche Differentialgleichungen

Die Gleichung

$$P'(t) = QP(t)$$

heißt die Rückwärtsgleichung,

$$P'(t) = P(t)Q$$

die Vorwärtsgleichung von Kolmogorov.

Wenn Ω_X endlich ist, ist Q eine gewöhnliche Matrix, und die Lösung der Kolmogorov-Gleichungen ist

$$P(t) = e^{Qt} = \sum_{n=0}^{\infty} \frac{Q^n t^n}{n!}.$$

Wenn Q diagonalisierbar ist, also

$$Q = E\Lambda E^{-1}$$

mit der Diagonalmatrix Λ der Eigenwerte und der Matrix E, der Spalten die zugehörigen Eigenvektoren sind, dann ist

$$P(t) = e^{Qt} = E\Lambda^t E^{-1} = E \begin{pmatrix} e^{t\lambda_1} & & \\ & \ddots & \\ & & e^{t\lambda_d} \end{pmatrix} E^{-1}.$$

Für $d = 2$, also eine Kette mit zwei Zuständen, hat Q die allgemeine Form

$$Q = \begin{pmatrix} -a & a \\ b & -b \end{pmatrix} (a, b \geq 0).$$

Diese hat die Eigenwerte 0 und $-(a + b)$. Für $a = b = 0$ ergibt sich $P(t) = I$, sonst

$$P(t) = \frac{1}{a+b} \begin{pmatrix} b + ae^{-(a+b)t} & a(1 - e^{-(a+b)t}) \\ b(1 - e^{-(a+b)t}) & a + be^{-(a+b)t} \end{pmatrix}.$$

Wenn Ω_X unendlich ist, gibt es in diesen Gleichungen einen zweifachen Grenzübergang (durch die Ableitung und eine Summation), deren Vertauschbarkeit nicht garantiert ist. Man kann jedenfalls zeigen, dass für $i \neq j$ $q_{ij} < \infty$ sein muss. Für die Diagonalelemente können unendliche Werte auftreten, also $q_{ii} = -\infty$. Auch wenn q_{ii} endlich Wert, kann es sein, dass die Zeilensumme echt kleiner als 0 ist. Diese Komplikationen wollen wir vermeiden und definieren

Definition 3.17

Die Markovkette X_t bzw. ihr Erzeuger Q heißt konservativ, wenn für alle i

$$q_{ii} > -\infty$$

und

$$\sum_j q_{ij} = 0$$

gilt.

Auch die Kolmogorov-Gleichungen sind nicht immer gültig. Wir haben:

Satz 3.7

Wenn die Markovkette X_t konservativ ist, dann erfüllen die Übergangswahrscheinlichkeiten $P(t)$ die Rückwärtsgleichung.

Wenn die infinitesimalen Parameter beschränkt sind, dann erfüllen die Übergangswahrscheinlichkeiten $P(t)$ die Vorwärtsgleichung.

Im letzten Fall gilt auch bei unendlichem Zustandsraum die Gleichung

$$P(t) = e^{Qt} = \sum_{n=0}^{\infty} \frac{Q^n t^n}{n!}.$$

Wenn $X(t)$ eine Markovkette in stetiger Zeit ist, dann ist

$$Y(n) = X(an)$$

für jedes $a > 0$ eine Markovkette in diskreter Zeit. Wir können die Eigenschaften, die wir von den Markovketten in diskreter Zeit kennen, auf den stetigen Fall übertragen, indem wir etwa den Zustand i rekurrent nennen, wenn er für Y rekurrent ist. Wegen

$$p_{ii}(t) \geq p_{ii}(\frac{t}{n})^n$$

ist für alle t $p_{ii}(t) > 0$, also ist Y immer aperiodisch. Bei den anderen Eigenschaften ist eigentlich noch zu zeigen, dass ihre Definition nicht von der konkreten Diskretisierung abhängt, also dass sich für jedes $a > 0$ dasselbe ergibt. Wir werden versuchen, diese Eigenschaften mit Hilfe von Q auszudrücken. Zuerst sehen wir uns an, wie wir den Prozess X_t simulieren können, also einen typischen Pfad erzeugen. Da wir die Übergangswahrscheinlichkeiten stetig angesetzt haben, ist es plausibel, anzunehmen, dass der Prozess eine gewisse Zeit in dem Zustand verbringt und dann

in einen anderen Zustand springt. Wir beginnen mit der Wahrscheinlichkeit, dass für alle $s \leq t$ $X(s) = i$ gilt. Dazu bestimmen wir zuerst

$$\mathbb{P}_i(X(\frac{kt}{2^n}) = i, k = 1, \ldots, 2^n) = p_{ii}(\frac{t}{2^n})^{2^n}.$$

Für $n \to \infty$ konvergiert dass wegen $p_{ii}(t/2^n) \approx 1 + q_{ii}t/2^n$ gegen $e^{q_{ii}t}$. Die Zeit, die der Prozess in i verbringt, ist also exponentialverteilt mit Parameter $-q_{ii}$, wenn q_{ii} negativ ist. Wenn $q_{ii} = 0$ gilt, dann gilt $p_{ii}(t) = 1$, also ist i absorbierend. Im anderen Fall können wir die Wahrscheinlichkeit

$$\mathbb{P}_i(X(t) = j | X(s) = i, s < t, X(t) \neq i)$$

als

$$\lim_{n\to\infty} \frac{\mathbb{P}_i(X(kt/2^n) = i, k = 1, \ldots, 2^n - 1, X(t) = j)}{\mathbb{P}_i(X(kt/2^n) = i, k = 1, \ldots, 2^n - 1, X(t) \neq i)} = \frac{q_{ij}}{-q_{ii}}$$

bestimmen. Wir definieren eine Übergangsmatrix R durch

$$r_{ij} = \begin{cases} \frac{q_{ij}}{-q_{ii}} & \text{wenn } q_{ii} < 0,\ j \neq i, \\ 1 & \text{wenn } q_{ii} = 0,\ j = i, \\ 0 & \text{sonst.} \end{cases}$$

Insgesamt sieht unsere Vorschrift für die Simulation einer Markovkette so aus: in einem nicht absorbierenden Zustand i bleibt der Prozess eine Zeit, die exponentialverteilt ist mit Parameter $-q_{ii}$, dann wird mit Wahrscheinlichkeit r_{ij} in den Zustand j gesprungen.

Die diskrete Markovkette mit R, die die Sprünge von X beschreibt, heißt die *eingebettete Markovkette*.

Diese Beschreibung funktioniert natürlich nur für konservative Markovketten. Wir können hier kurz darüber nachdenken, was geschieht, wenn die Konservativität nicht gilt: wenn $q_i i = -\infty$, dann ist für jedes t die Wahrscheinlichkeit, dass $X(s) = i$ für alle $s < t$ gilt, gleich 0. Der Prozess verlässt also den Zustand i sofort nach dem Betreten. Wir nennen einen solchen Zustand instabil. Ist q_{ii} endlich aber die Zeilensumme nicht 0, dann ist die Zeit, die X in i verbringt, exponentialverteilt, aber die Summe der Übergangswahrscheinlichkeiten r_{ij} ist kleiner als 1, also springt der Prozess mit positiver Wahrscheinlichkeit in keinen der anderen Zustände; er verschwindet also im Nirgendwo, von wo er allerdings sofort wieder zurückkommen muss.

Besonders schön ist der Fall, dass die eingebettete Markovkette sehr einfache Form hat, etwa, dass die Übergänge nur in Nachbarzustände erfolgen können. Ein solcher Prozess heißt Geburts- und Todesprozess.

Die allgemeine Definition ist

Definition 3.18

Ein Geburts- und Todesprozess ist eine Markovkette mit Zustandsraum $\{0, 1, \ldots\}$ und infinitesimalen Parametern

$$q_{00} = -\lambda_0, q_{01} = \lambda_0,$$

$$q_{i,i-1} = \mu_i, q_{ii} = -(\lambda_i + \mu_i), q_{i,i+1} = \lambda_i.$$

Die Zahlen λ_i heißen die Geburtsraten, μ_i die Todesraten. Sind alle $\mu_i = 0$, dann sprechen wir von einem reinen Geburtsprozess, sind alle $\lambda_i = 0$, von einem reinen Todesprozess.

Reine Geburtsprozesse sind besonders einfach zu analysieren, weil jeweils nach einer exponentialverteilten Zeit τ_i mit Parameter λ_i in den nächsthöheren Zustand $i + 1$ gesprungen wird. Bis zum erreichen des Zustands n (wenn in 0 gestartet wird), vergeht die Zeit

$$\tau_0 + \ldots + \tau_{n-1}$$

mit Erwartungswert

$$\frac{1}{\lambda_0} + \ldots + \frac{1}{\lambda_{n-1}}.$$

Wenn die Summe

$$\sum_{i=0}^{\infty} \frac{1}{\lambda_i}$$

endlich ist, dann hat die Zeit

$$\sum_{n=0}^{\infty} \tau_n$$

die benötigt wird, bis alle Zustände durchlaufen sind, endlichen Erwartungswert und ist daher endlich. Die Markovkette macht also in endlicher Zeit unendlich viele Sprünge, eine Erscheinung, die als "Explosion" bezeichnet wird. Nach diesen unendlich vielen Sprüngen ist unser Geburtsprozess erst einmal in der Unendlichkeit verschwunden. Von dort muss er natürlich sofort ins Endliche zurückkehren, das allerdings kann auf unterschiedliche Weise geschehen: wir können uns einen bestimmten Zustand aussuchen, in den gesprungen wird, oder aus mehreren nach einer Verteilung auswählen. Kurz gesagt: solange der Prozess nur endliche viele Sprünge absolviert hat, ist alles eindeutig, gibt es unendlich viele Sprünge, dann nicht mehr. Das führt uns zu der Definition

Definition 3.19

Eine Markovkette in stetiger Zeit heißt regulär, wenn sie in endlicher Zeit nur endlich viele Sprünge ausführt, im anderen Fall singulär..

Unser Beispiel eines singulären Geburtsprozesses (bei dem wir aus der Unendlichkeit nach 0 zurückspringen) verläuft also so, dass er immer wieder von 0 nach oben durch alle Zustände läuft. Wenn wir diesen Prozess rückwärts laufen lassen, dann haben wir ein Beispiel für einen nicht konservativen Prozess: jetzt laufen die Zyklen aus dem Unendlichen‘ durch alle Zustände von oben nach unten und springen dann aus dem Zustand 0 nach unendlich und so weiter. Es wird immer in 0 eine exponentialverteilte Zeit verbracht, was $q_{00} = -\lambda_0$ gibt, aber um von 0 nach j zu kommen, müssen die Zeiten τ_j und τ_{j+1} vergehen. Die Wahrscheinlichkeit, dass sich das in Zeit $\leq t$ ausgeht, ist durch $\lambda_j \lambda_{j+1} t^2$ beschränkt, und daher gilt $q_{0j} = 0$ (und wir sehen unsere frühere Beschreibung dessen, was in einem nicht konservativen Markovkette passiert: "springt nach unendlich und läuft von dort zurück", bestätigt).

In der Zukunft werden wir uns auf reguläre Ketten beschränken. Diese sind recht angenehm:

Satz 3.8

Ist X regulär, dann gelten die Vorwärts- und die Rückwärtsgleichung, und sie sind eindeutig lösbar.

Für reguläre Markovketten lassen sich alle Eigenschaften der Kette aus denen der eingebetteten Markovkette ablesen. Die Markovkette X ist etwa genau dann rekurrent, wenn die eingebettete Markovkette rekurrent ist (eine Richtung funktioniert immer: wenn die eingebettete Kette rekurrent ist, dann auch X, und X ist regulär: bis zur Explosion muss ja ein Zustand i unendlich oft besucht werden, und das braucht jedesmal im Mittel $-1/q_i i$ Zeiteinheiten, also insgesamt unendlich lange). Für die positive Rekurrenz gilt das nicht — es ist nicht schwer, Beispiele zu konstruieren, in denen jeweils eine der beiden Ketten positiv rekurrent ist, die andere nullrekurrent. Absorptionswahrscheinlichkeiten stimmen bei beiden Ketten überein, die mittleren Absorptionszeiten haben unterschiedliche Werte.

Jedenfalls ist in regulären Ketten der Zustand der Zustand j genau dann Nachfolger von i, wenn er es in der eingebetteten Kette ist. Das ist für $j \neq i$, dass es Zustände $i_0 = i, \ldots, i_n = j$ gibt mit $q_{i_{l-1} i_l} \neq 0, i = 1, \ldots, n$.

Entweder über die eingebettete Markovkette, oder aus der Gleichung für die diskretisierte Kette, deren Schrittweite wir gegen 0 gehen lassen, erhalten wir die Formeln

Für die stationäre Verteilung:

$$\pi Q = 0.$$

Für die Absorptionswahrscheinlichkeiten: es wird wieder $a_i = 1$ gesetzt, wenn i ein "guter" absorbierender Zustand ist, $a_i = 0$ für "schlechte", und wenn i nicht absorbierend ist, dann

$$\sum_j q_{ij} a_j = 0.$$

Wie im diskreten Fall kann es sein, dass diese Gleichung nicht eindeutig lösbar ist. Dann ist wieder die kleinste mit $a_i \geq 0$ die richtige.

Mittlere Absorptionszeiten: m_i ist wieder 0, wenn i absorbierend ist, sonst gilt

$$\sum_j q_{ij} m_j = 0.$$

3.5 Wiederholungsfragen

1. Was ist ein stochastischer Prozess?
2. Was ist ein Erneuerungsprozess?
3. Wie ist ein stationärer Prozess definiert?
4. Wie ist ein Prozess mit unabhängigen Zuwächsen definiert?
5. Wie ist die Markoveigenschaft definiert?
6. Was ist eine Markovkette?
7. Wie lauten die Chapman-Kolmogorov-Gleichungen?
8. Wie ist der einfache Random Walk definiert, und wie lauten seine Übergangswahrscheinlichkeiten?
9. Wann heißt ein Zustand Nachfolger eines anderen?
10. Wann heißen zwei Zustände kommunizierend?
11. Was ist eine Klasseneigenschaft?
12. Wann heißt eine Markovkette irreduzibel?
13. Wie ist die Periode eines Zustands definiert?
14. Wann heißt ein Zustand rekurrent?
15. Wie ist positive Rekurrenz definiert?
16. Wie verhalten sich die Übergangswahrscheinlichkeiten $p_{ij}(t)$ in einer irreduziblen aperiodischen Markovkette für $t \to \infty$?
17. Was ist eine stationäre Verteilung?
18. Wie kann man die positive Rekurrenz anders charakterisieren?
19. Was kann man über die Rekurrenz des Random Walk aussagen?
20. Was kann man über dir Rekurrenzeigenschaften von endlichen Markovketten haben.
21. Wie sind Absorptionswahrscheinlichkeiten definiert?
22. Wie kann man die Absorptionswahrscheinlichkeiten berechnen?
23. Welche speziellen Annahmen machen wir für Markovketten in stetiger Zeit?
24. Wie kann man Rekurrenz, Transienz, positive Rekurrenz und Nullrekurrenz für Markovketten in stetiger Zeit definieren?

25. Was sind die infinitesimalen Parameter einer Markovkette?
26. Wie lauten die Kolmogorovschen Differentialgleichungen?
27. Was kann man über die Gültigkeit der Kolmogorovschen Differentialgleichungen sagen?
28. Wie kann man für endliche Markovketten die Übergangsmatrizen aus dem infinitesimalen Erzeuger erhalten?
29. Wie kann man die infinitesimalen Parameter anschaulich (?) interpretieren?
30. Wie kann mithilfe des infinitesimalen Erzeugers die Absorptionswahrscheinlichkeiten berechnen?
31. Wie kann mithilfe des infinitesimalen Erzeugers die stationäre Verteilung berechnen?

Kapitel 4

Informationstheorie

Was sind das für Zeiten, wo
Ein Gespräch über Bäume fast ein
Verbrechen ist
Weil es ein Schweigen über so viele
Untaten einschließt!

B. Brecht, "An die Nachgeborenen"

4.1 Entropie und Information

Das Bar Kochba Spiel: Spieler A wählt einen aus einer Menge M von m Gegenständen. Spieler B muss herausfinden, welcher es ist, und darf dazu nur Fragen stellen, die A Mit ja oder nein beantworten kann. Wenn A betrügen kann und sich nur nicht widersprechen darf, dann sieht man leicht, dass mit optimaler Strategie ein Spiel genau

$$H^*(m) = \lceil \log_2(m) \rceil$$

Runden dauert: einerseits gibt es bei k Fragen 2^k mögliche Kombinationen von "ja"- und "nein"-Antworten. Für jedes der m zu wählenden Elemente muss es eine solche Kombination geben, die es definiert, also ergibt sich $2^k \geq m$. Auf der anderen Seite können wir mit jeder Frage die Anzahl der noch in Frage kommenden Gegenstände (ungefähr) halbieren: wenn $2^{k-1} < m \leq 2^k$ ist, dann wäre die erste Frage beispielsweise "ist die Nummer (wir müssen die Gegenstände zuerst von 0 bis $m-1$ durchnumerieren) $< 2^{k-1}$?", je nach Ausgang dieser Frage ist die nächste Frage "ist die Nummer $< 2^{k-2}$?" oder "ist die Nummer $< 3 \cdot 2^{k-2}$?" und so weiter — das ist äquivalent dazu, dass man die k Binärziffern der Nummer der Reihe nach abfragt.

Durch einen Trick können wir das Aufrunden loswerden: statt nur nach einem Element der Menge M zu fragen, fragen wir nach n Elementen gleichzeitig, als nach einer Folge $(x_1, \ldots, x_n) \in M^n$. Es gibt m^n solche Folgen. Um eine davon zu identifizieren, benötigen wir $H^*(m^n)$ Fragen. Das liegt zwischen $\log_2(m^n) = n \log_2(m)$ und $n \log_2(m) + 1$. Pro einzelnem Element sind das $H^*(m^n)/n$ Fragen, und das konvergiert für $n \to \infty$ gegen $\log_2(m)$. Das führt uns zu der Definition

Definition 4.1

$$H^*(m) = \lceil \log_2(m) \rceil$$

heißt die maximale Unbestimmtheit,

$$H(m) = \log_2(m)$$

die Entropie von m.

Diese Situation wird uns in der Informationstheorie öfter begegnen: es gibt einen optimalen Wert, der nicht in allen Fällen direkt erreicht werden kann, an den man aber mit großen Blöcken beliebig nahe herankommen kann (zumindest mit einer Wahrscheinlichkeit, die beliebig nahe an 1 gewählt werden kann).

Von nun an werden wir die Elemente von M durchnumerieren und durch ihre Nummern ersetzen, wir setzen also $M = \{1, \ldots, m\}$. Die Fragen, die wir stellen, können wir durch die Teilmenge von M darstellen, für die die Frage mit "ja" beantwortet wird. Wenn die Sprache, die wir verwenden, um unsere Fragen zu formulieren, hinreichend aussagekräftig ist, können wir auch zu jeder Menge eine Frage finden, die genau dieser Menge entspricht (die langweiligste Idee ist wohl: "ist x in A enthalten?").

Fragestrategien können wir durch Binärbäume darstellen:

Definition 4.2

Ein Binärbaum ist ein gerichteter zyklusfreier Graph, bei dem aus jedem Knoten höchstens zwei Kanten entspringen. Es gibt einen ausgezeichneten Knoten, die Wurzel, in der keine Kante endet. In jedem anderen Knoten endet genau eine Kante. Ein Knoten, aus dem keine Kante entspringt, heißt Endknoten oder Blatt, alle anderen heißen innere Knoten. Wenn aus jedem inneren Knoten genau zwei Kanten entspringen, dann nennen wir den Baum vollständig. Der Abstand eines Blattes von der Wurzel heißt die Blattlänge dieses Blattes, das Maximum aller Blattlängen ist die Höhe des Baums.

Unsere Fragebäume werden nun so konstruiert: an die inneren Knoten, ausgehend Wurzel des Baumes setzen wir die jeweilige Frage, ist die Antwort "ja", dann gehen wir zum linken Nachfolger, sonst zum rechten. An die Endknoten setzen wir die Elemente, die wir dort eindeutig bestimmt haben. In Abbildung 4.1 ist der Fragebaum für eine optimale Strategie für $m = 5$ dargestellt.

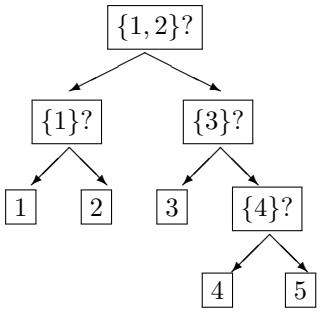

Abbildung 4.1: Ein Fragebaum

Wir ändern die Regeln des Spiels jetzt so ab, dass A nicht mehr betrügen darf und seine Auswahl zufällig mit Verteilung $P = (p_1, \ldots, p_m)$ wählt. Wir wollen die Strategie finden, bei der der Erwartungswert der Anzahl der Fragen minimal wird. Diesen Minimalwert nennen wir die mittlere Unbestimmtheit $H^*(P)$. Um ihn zu bestimmen, müssen wir die optimale Strategie finden. Das geht mit dem Huffman-Algorithmus: zuerst stellen wir fest, dass wir jede Fragestrategie durch einen Binärbaum repräsentieren können, und dass umgekehrt jedem solchen Baum eine Fragestrategie entspricht. Ist nämlich ein Baum gegeben, dessen Blätter von 1 bis m numeriert sind, können wir in zu einem Fragebaum machen, indem wir an jeden inneren Knoten als Frage die Menge aller Blätter setzen, die zum linken Teilbaum unter diesem Knoten gehören.

Die Blattlänge l_i des Blattes, das dem Element i entspricht, ist genau die Anzahl der Fragen, die benötigt werden, um dieses Element zu identifizieren. Die mittlere Anzahl von Fragen ist also gleich der mittleren Blattlänge

$$\sum_{i=1}^{m} p_i l_i.$$

Wir suchen also einen Binärbaum mit m Blättern, für den die mittlere Blattlänge minimal ist. Wir könnten versuchen, die Idee zu kopieren, die bei der maximalen Unbestimmtheit erfolgreich war, nur wollen wir jetzt statt der Anzahlen die Wahrscheinlichkeiten aufteilen. Versuchen wir das an der Verteilung $P = (0.1, 0.2, 0.3, 0.4)$: wir können hier sogar perfekt halbieren, indem wir 1 und 4 bzw. 2 und 3 zusammenfassen. Der komplette Baum sieht dann so aus wie in Abbildung 4.2.

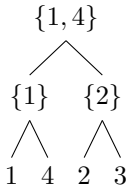

Abbildung 4.2: Halbieren: leider nicht optimal

Dieser Baum ist offensichtlich nicht optimal: die mittlere Blattlänge ist 2 und damit größer als für die Strategie in Abbildung 4.3, in der wir einfach der Reihe nach fragen ob es sich um 4, 3 oder 2 handelt, was uns eine mittlere Fragenanzahl von $0.4 \cdot 1 + 0.3 \cdot 2 + (0.2 + 0.1) \cdot 3 = 1.9$ liefert.

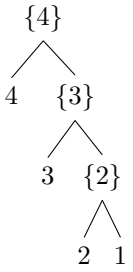

Abbildung 4.3: Eine bessere Strategie

Wir denken also darüber nach, wie ein optimaler Baum aussehen muss. Zuerst kann man ohne weitere Veränderungen am Baum versuchen, die Zuordnung der Zahlen 1 bis m zu den Endknoten zu optimieren: es ist recht offensichtlich, dass die kleinste mittlere Blattlänge dann erreicht wird, wenn die Wahrscheinlichkeiten in umgekehrter Reihenfolge geordnet werden wie die Blattlängen. Wenn die Numerierung so vorgenommen wird, dass $p_1 \geq p_2 \geq \ldots \geq p_m$ gilt, dann muss die Zuordnung zu den Blättern so vorgenommen werden, dass $l_1 \geq l_2 \geq \ldots \geq l_m$ gelten. Als nächstes geht es an die Struktur des Baumes. Ein optimaler Baum muss vollständig sein: ein innerer Knoten, von dem nur eine Kante ausgeht, entspricht einer Frage, deren Antwort man schon kennt, und kann daher weggelassen werden. Dann betrachten wir die Blätter m und $m-1$. Diese haben die kleinsten Wahrscheinlichkeiten p_m und p_{m-1} und damit die beiden größten Blattlängen l_m und l_{m-1}. Weil der Baum vollständig ist, muss $l_m = l_{m-1}$ gelten. Wir können ohne Beschränkung der Allgemeinheit annehmen, dass die Knoten m und $m-1$ am selben (direkten) Vorgänger hängen. Wenn wir jetzt diese beiden Knoten entfernen, erhalten wir eine Baum mit $m-1$ Blättern, in dem der gemeinsame Vorgänger von m und $m-1$ zu einem neuen Blatt (mit Gewicht $p_m + b_{m-1}$ geworden ist. Man kann also den (einen) optimalen Baum rekursiv mit dem Huffman-Algorithmus konstruieren:

Satz 4.1: Huffman

Einen optimalen Baum für die Verteilung $P = (p_1, \ldots, p_m)$ kann man folgendermaßen konstruieren:

1. Wenn $m = 1$, dann besteht der Baum nur aus der Wurzel, Ende.
2. Ordne die Wahrscheinlichkeiten: $p_1 \geq \ldots \geq p_m$.

3. Fasse die kleinsten Wahrscheinlichkeiten zusammen: $p^*_{m-1} = p_{m-1} + p_m$,
4. Konstruiere den optimalen Baum für $P^* = (p_1, \ldots, p_{m-2}, p^*_{m-1})$
5. Ersetze Blatt $m-1$ durch einen inneren Knoten mit den Blättern $m-1$ und m.

Wenn wir dieses Verfahren auf unser Beispiel $P = (0.1, 0.2, 0.3, 0.4)$ anwenden, dann müssen wir zuerst die Knoten 1 und 2 zusammenfassen. Das ergibt jetzt drei Knoten mit den Wahrscheinlichkeiten $(0.3, 0.3, 0.4)$. Im nächsten Schritt werden wieder die beiden ersten Knoten zusammengefasst. Im Endeffekt ergibt sich genau der Baum aus Abbildung 4.3, der sich also jetzt als optimal herausstellt.

Die Bestimmung des Huffman-Baums erfordert zumindest eine Sortierung, die relativ kostspielig ist, was den Zeitaufwand betrifft (das Sortieren in jedem Schritt, das in unserer Beschreibung steht, lässt sich leicht vermeiden), also wäre es schön, wenn wir zumindest eine gute Abschätzung hätten, die sich mit weniger Aufwand berechnen lässt. Wir kehren also wieder zurück zu dem Problem, dass wir unter allen Binärbäumen denjenigen suchen, für den

$$\sum p_i l_i$$

minimal wird, wobei l_i die Blattlänge des Blatts mit dem Index i ist. Für diese Aufgabe ist es nützlich zu wissen, wann ein Binärbaum mit den Blattlängen $l_1, \ldots l_m$ existiert:

Satz 4.2: Ungleichung von Kraft

Ein Binärbaum mit den Blattlängen $l_1, \ldots, l_m$ existiert genau dann, wenn

$$\sum_{i=1}^{m} 2^{-l_i} \leq 1.$$

Gleichheit gilt genau dann, wenn der Baum vollständig ist.

In jeder Richtung erfolgt der Beweis durch Induktion nach der Höhe des Baumes. Die eine Richtung (jeder Baum erfüllt die Ungleichung mit Gleichheit, wenn der Baum vollständig ist) ist für Höhe 0 offensichtlich: dann gibt es einen Knoten, die Wurzel, die gleichzeitig das einzige Blatt ist und Höhe 0 hat. Dieser Baum ist auch vollständig, und die Kraftsche Ungleichung gilt mit Gleichheit. Ist die Höhe positiv, dann ist die Wurzel ein innerer Knoten, an dem ein oder zwei Teilbäume hängen. Der linke davon soll m' Blätter mit den Längen $l'_1, \ldots, l'_{m'}$, der rechte m'' Blätter mit den Längen $l''_1, \ldots, l''_{m''}$. Nach der Induktionsvoraussetzung gilt in jedem dieser Teilbäume die Kraftsche Ungleichung; für den Ursprünglichen Baum gilt $m = m' + m''$ und $l_i = 1 + l'_i$ für $i \leq m'$ und $l_i = 1 + l''_{i-m'}$ für $m' < i \leq m' + m''$. Das ergibt

$$\sum_{i=1}^{m} 2^{-l_i} = \frac{1}{2}(\sum_{i=1}^{m'} 2^{-l'_i} + \sum_{i=1}^{m''} 2^{-l''_i}) \leq 1.$$

Ist der Baum vollständig, dann sind das auch die Teilbäume, und es gilt Gleichheit.

Für die andere Richtung beginnen wir wieder mit Höhe 0; wenn die Kraftsche Ungleichung gilt, muss $m = 1$ und $l_1 = 0$ gelten, und wir sind wieder bei unserem einen Knoten. Wir nehmen an, dass $l_1 \geq \ldots \geq l_m$ und

$$S = \sum_{i=1}^{m} 2^{-l_i} \leq 1$$

gilt. l_m ist die Höhe des Baums. Es gilt

$$S = \sum_{i=1}^{m} 2^{l_m - l_i} \leq 2^{l_m}.$$

Auf beiden Seiten stehen ganze Zahlen, die rechte Seite ist gerade, und auf der linken Seite steht für jedes l_i, das mit l_m übereinstimmt, eine 1 und für jedes andere eine gerade Zahl. Ist die linke Seite

ungerade, dann können wir noch eine Zahl $l_{m+1} = l_m$ hinzufügen, und die Kraftsche Ungleichung gilt mit diesem zusätzlichen Summanden immer noch. Wir dürfen also annehmen, dass es eine gerade Anzahl von Indizes i mit $l_i = l_m$ gibt. Wir ersetzen jeweils zwei Zahlen l_m durch eine mit dem Wert $l_m - 1$. Dadurch behält die Summe S ihren Wert, und weil wir die Höhe um 1 kleiner gemacht haben, gibt es nach Induktionsvoraussetzung einen Baum mit diesen kürzeren Blattlängen. Wenn wir jetzt an die Blätter mit Blattlänge $l_m - 1$ die für zwei Werte l_m stehen, zwei Nachfolger anhängen, erhalten wir einen Baum mit den ursprünglichen $l_1, \ldots, l_m$ als Blattlängen. Es folgt also aus der Gültigkeit der Kraftschen Ungleichung die Existenz eines entsprechenden Baumes. Ist $S = 1$, dann ist auch für den verkürzten Baum in der Kraftschen Ungleichung Gleichheit vorhanden, also ist dieser verkürzte Baum vollständig, und er bleibt es auch, wenn wir aus einigen seiner Blätter innere Knoten mit zwei Blättern als Nachfolger machen.

Die Suche nach der optimalen Strategie besteht also darin,

$$\sum p_i l_i$$

unter der Nebenbedingung

$$\sum 2^{-l_i} \leq 1$$

zu minimieren. Es ist leicht einzusehen, dass im optimalen Fall in der letzten Ungleichung Gleichheit gelten muss (sonst könnten wir einfach dass größte l_i verkleinern). Wenn wir die Forderung, dass l_i ganzzahlig sein muss, beiseite lassen, lässt sich das Minimum mit der Lagrange-Methode bestimmen und hat den Wert

$$\sum p_i \log_2(1/p_i).$$

Deswegen definieren wir

Definition 4.3

Die Entropie der Verteilung P ist

$$H(P) = \sum_{i=1}^{m} p_i \log_2(1/p_i) = -\sum_{i=1}^{m} p_i \log_2(p_i).$$

Für eine Zufallsvariable X mit Verteilung P_X ist

$$H(X) = H(P_X).$$

Satz 4.3

$$H(P) \leq H^*(P) \leq H(P) + 1.$$

Die untere Abschätzung haben wir oben gezeigt, die obere folgt aus der Tatsache, dass $l_i = \lceil \log_2(1/pi) \rceil$ die Kraftsche Ungleichung erfüllt.

Der Summand 1 in der oberen Abschätzung stört ein wenig; man kann diese Differenz verringern, indem man statt eines einzelnen Elements mehrere (unabhängige und identisch verteilte) errät, sagen wir n. Die Entropie der gemeinsamen Verteilung ist (wie weiter unten gezeigt wird) $nH(P)$, damit lässt sich die mittlere Anzahl der Fragen, um den gesamten Block zu erraten, mit $nH(P)+1$ abschätzen, pro Element ergibt das $H(P) + 1/n$, was beliebig nahe an die Entropie herankommt.

Die Formel für die Entropie legt nahe, dass in der optimalen Strategie für das erraten von i etwa $\log_2(\frac{1}{p_i})$ Fragen nötig sind. Das ist zwar nicht ganz richtig (einerseits ist für $m = 2$ immer $l_i = 1$, egal wie groß oder klein die Wahrscheinlichkeiten sind; in der anderen Richtung wird in den Übungen gezeigt, dass

$$l_i \leq \log_\tau(\frac{1}{p_i}) + 1$$

gilt. Dabei ist $\tau = \frac{1+\sqrt{5}}{2} \approx 1.618$ der goldene Schnitt. Dieser Logarithmus ist etwa um den Faktor 1.44 größer als der Zweierlogarithmus), ist aber gelegentlich hilfreich, um Ergebnisse zu motivieren.

Wir betrachten etwa den Fall, dass wir eine Zufallsvariable X mit Verteilung P erraten müssen, aber die optimale Strategie für eine andere Verteilung Q verwenden. Wenn wir annehmen, dass die Anzahl der Fragen für Ausgang i gleich $\log_2(1/q_i)$ ist, dann brauchen wir im Mittel

$$\sum p_i \log_2(1/q_i)$$

Fragen statt

$$\sum p_i \log_2(1/p_i),$$

also um

$$\sum p_i \log_2(p_i/q_i)$$

Fragen zu viel.

Das führt uns zu der Definition

Definition 4.4

$$D(P,Q) = \sum_i p_i \log_2(p_i/q_i).$$

heißt die Informationsdivergenz (I-Divergenz, Kullback-Leibler Information, relative Entropie, Strafe des Irrtums) zwischen P und Q.

Die plausible Folgerung aus unseren vorigen Überlegungen ist

Satz 4.4

Die I-Divergenz ist nichtnegativ

$$D(P,Q) \geq 0,$$

mit Gleichheit für $P = Q$.

Für den Beweis verwenden wir die Ungleichung

$$\log(x) \leq x - 1$$

(mit Gleichheit für $x = 0$), die aus der bekannten Ungleichung

$$e^x \geq 1 + x$$

folgt. Damit ist

$$-D(P,Q) = \frac{1}{\log 2} \sum_i p_i \log(\frac{q_i}{p_i}) \leq \frac{1}{\log 2} \sum_i p_i (\frac{q_i}{p_i} - 1) = \frac{1}{\log 2} \sum_i (p_i - q_i) = 1 - 1 - 0.$$

Damit Gleichheit gilt, müssen alle Ungleichungen für die Logarithmen in der Summe Gleichungen sein, also $q_i/p_i = 1$ bzw. $p_i = q_i$.

Wenn X und Y zwei Zufallsvariable sind, können wir (X,Y) als eine Zufallsvariable mit endlich vielen Werten ansehen und die gemeinsame Entropie $H(X,Y) = H((X,Y))$ betrachten. Wir setzen

$$p(x|y) = \mathbb{P}(X = x | Y = y)$$

und definieren

Definition 4.5

$$H(X|Y=y) = -\sum_x p(x|y) \log_2(p(x|y))$$

und nennen

$$H(X|Y) = \sum_x p_Y(y) H(X|Y=y)$$

die bedingte Entropie von X unter Y.

Es gilt

Satz 4.5

$$H(X,Y) = H(Y) + H(X|Y),$$

$$\max(H(X), H(Y)) \leq H(X,Y) \leq H(X) + H(Y), \text{‘}$$

$$H(X|Y) \leq H(X),$$

$$H(X|Y,Z) \leq H(X|Y).$$

$$H(X,Y) = H(X) + H(Y)$$

gilt genau dann, wenn X und Y unabhängig sind.

$$H(X,Y) = H(X)$$

gilt genau dann, wenn es eine Funktion g gibt, sodass $Y = g(X)$.

Definition 4.6

$$I(X,Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

heißt die Information zwischen X und Y.

Der letzte Satz impliziert

Satz 4.6

$$0 \leq I(X,Y) \leq \min(H(X), H(Y)).$$

Die Information ist genau dann 0, wenn X und Y unabhängig sind. $I(X,Y) = H(X)$ gilt genau dann, wenn $X = g(Y)$ gilt, also wenn X eindeutig aus Y bestimmt werden kann.

Wenn X aus Y zwar nicht mit absoluter Sicherheit, aber mit großer Wahrscheinlichkeit bestimmt werden kann, dann unterscheidet sich die Information nur wenig von der Entropie von X:

Satz 4.7

Wenn $\mathbb{P}(X \neq Y) \leq \epsilon$ ist, dann gilt

$$I(X,Y) \geq H(X) - H(\epsilon, 1-\epsilon) - \epsilon \log_2(m).$$

4.2 Codes

Eine Fragestrategie kann auch unter einem anderen Gesichtspunkt gesehen werden, nämlich indem für jede Frage die Antwort "nein" mit einer 0, die Antwort "ja" mit einer 1 codiert wird. Codes, die auf diese Weise gewonnen werden, haben eine wichtige Eigenschaft — sie sind präfixfrei, d.h. kein Codewort ist Präfix (Anfangsstück) eines anderen. Diese Eigenschaft wird auch als fortlaufende Entzifferbarkeit bezeichnet, weil an jeder Stelle der codierten Nachricht festgestellt werden kann, ob dort ein Codewort endet, ohne dass die nachfolgenden Zeichen bekannt sind (man erkennt leicht, dass dies genau bei den präfixfreien Codes der Fall ist. Der Huffmancode ist damit der

optimale präfixfreie Code, also der mit der kleinsten mittleren Codewortlänge. Allerdings gibt es nicht "den" Huffmancode, weil sich durch Spiegeln des Baumes oder von Teilen davon andere Codes ergeben. Außerdem kann es mehrere äquivalente Möglichkeiten geben, Knoten zusammenzufassen, etwa für die Verteilung $P = (0.2, 0.2, 0.2, 0.4)$, für die $(00, 01, 10, 11)$ und $(000, 001, 01, 1)$ beides Huffmancodes sind. Das zweite Problem ist nicht so groß, für das erste gibt es den kanonischen Huffmancode.

Definition 4.7

Für einen Präfixfreien Code mit den Codewortlängen $l_1, \ldots, l_m$ wird der kanonische Code folgendermaßen konstruiert:

1. Die Wortlängen werden aufsteigend geordnet: $l_1 \leq \ldots \leq l_m$.
2. Bei gleichen Wortlängen werden die entsprechenden Quellbuchstaben in ihrer natürlichen Reihenfolge geordnet.
3. c_1 besteht aus l_1 Nullen.
4. c_{i+1} wird aus c_i erhalten, indem 1 addiert wird (c_i wird als Binärzahl interpretiert), und dann werden $l_{i+1} - l_i$ Nullen angehängt.

Das ist gleichbedeutend damit, dass c_i aus den ersten l_i Nachkommastellen von

$$\sum_{j=1}^{i-1} 2^{-l_j}$$

besteht. Der kanonische Huffmancode wird erhalten, wenn die Wortlängen mit dem Huffmanalgorithmus bestimmt werden.

Kanonische Codes haben den Vorteil, dass statt der kompletten Codewörter für jeden Quellbuchstaben nur ihre Längen übertragen werden müssen.

Beispiel 4.1

Es muss wieder $P = (0.1, 0.2, 0.3, 0.4)$ herhalten:

i	x_i	l_i	c_i
1	4	1	0
2	3	2	10
3	2	3	110
4	1	3	111

Einen anderen Code mit fast optimaler Codewortlänge erhalten wir, wenn wenn wir von unserer oberen Abschätzung für die Entropie ausgehen. Wir wollen also einen Code mit Codewortlängen $l_i = \lceil \log_2(1/p_i) \rceil$ explizit angeben. Dazu ordnen wir die Wahrscheinlichkeiten absteigend ($p_1 \geq \ldots \geq p_m$) und setzen $f_i = \sum_{j=1}^{i} p_i$. Das Codewort c_i erhalten wir, indem wir f_{i-1} als Binärzahl darstellen und die ersten l_i Nachkommastellen als Code verwenden. Es ist nicht schwer einzusehen, dass dadurch ein präfixfreier Code definiert wird, der Shannon-Code. Der einzige Schönheitsfehler dabei ist, dass die Wahrscheinlichkeiten geordnet werden müssen. Diesen Schönheitsfehler behebt der Fano-Code: mit denselben Notationen wie vorhin (abgesehen davon, dass die Wahrscheinlichkeiten nicht geordnet werden müssen) codiert man $(f_{i-1} + f_i)/2$ mit $\lceil \log_2(1/p_i) \rceil + 1$ Bits.

Beispiel 4.2

Diese Codes werden wir jetzt für unser Standardbeispiel $P = (0.1, 0.2, 0.3, 0.4)$ konstruieren. Zuerst den Shannon-Code. Für diesen müssen wir die Wahrscheinlichkeiten absteigend ordnen:

i	x_i	p_i	l_i	f_{i-1}		Codewort
				dezimal	binär	
1	4	0.4	2	0	0.00000000...	00
2	3	0.3	2	0.4	0.01100110...	01
3	2	0.2	3	0.7	0.10110011...	101
4	1	0.1	4	0.9	0.11100110...	1110

Für den Fanocode brauchen wir nicht zu ordnen, aber dafür sind die Codewörter einen Buchstaben länger.

i	p_i	l_i	$(f_i + f_{i-1})/2$		Codewort
			dezimal	binär	
1	0.1	5	0.05	0.00001101...	00001
2	0.2	4	0.2	0.00110011...	0011
3	0.3	3	0.45	0.01110011...	011
4	0.4	3	0.8	0.11001100...	110

In diesen Tabellen haben wir die Darstellung als Binärbruch aus Anschaulichkeitsgründen auf 8 Stellen ausgeschrieben. Etwas einfacher ist es, die Wahrscheinlichkeitswerte mit 2^{l_i} zu multiplizieren und den ganzzahligen Anteil des Produkts als Binärzahl, eventuell mit führenden Nullen, darzustellen. Für die Codierung von "2" im Shannon-Code sieht das so aus: Der f-Wert 0.7 wir mit $2^3 = 8$ multipliziert, und das Codewort wird erhalten, indem der ganzzahlige Anteil (5) mit drei Binärstellen geschrieben wird.

Diese Idee kann man auf das Kodieren von ganzen Blöcken anwenden, im Extremfall wird die ganze Nachricht als einzelner Block kodiert. Im Vergleich zum Huffman-Code ist das hier möglich, weil nicht der ganze Code generiert werden muss, sondern nur das eine Wort, das die Nachricht kodiert. Verfahren, die auf dieser Idee beruhen, werden als arithmetische Codes bezeichnet.

Außer den präfixfreien Codes gibt es auch noch andere, die eindeutig entziffert werden können. Wir definieren

Definition 4.8

1. Ein Code heißt endlich eindeutig entzifferbar, wenn jede endliche Aneinanderreihung von Codewörtern eindeutig in Codewörter zerlegt werden kann.

2. Ein Code heißt eindeutig entzifferbar (manchmal zur Unterscheidung von 1. unendlich eindeutig entzifferbar), wenn jede endliche oder unendliche Aneinanderreihung von Codewörtern eindeutig zerlegt werden kann.

Der Code $\{0, 01\}$ ist offensichtlich nicht präfixfrei, aber trotzdem eindeutig entzifferbar, für die korrekte Zerlegung muss man allerdings das nachfolgende Zeichen kennen. Der Code $\{0, 01, 11\}$ ist endlich eindeutig entzifferbar, aber nicht eindeutig entzifferbar.

Wir können nun die Frage stellen, ob durch den Verzicht auf die fortlaufende Entzifferbarkeit etwas gewonnen werden kann, also ob es einen endlich eindeutig entzifferbaren Code gibt, der kleiner mittlere Codewortlänge hat als der Huffman-Code. Diese Hoffnung ist allerdings vergebens, denn es gilt

Satz 4.8

Die Codewortlängen in einem endlich eindeutig entzifferbaren Code erfüllen die Kraftsche Ungleichung.

Aus diesem Satz folgt, dass es zu jedem endlich eindeutig entzifferbaren Code einen präfixfreien Code mit denselben Codewortlängen (und damit mit derselben mittleren Codewortlänge) gibt.

(universelle Codes)

4.3 Informationsquellen

Definition 4.9

Eine Informationsquelle ist eine Folge $\mathcal{X} = (X_1, \ldots)$ von Zufallsvariablen.

Die verschiedenen Möglichkeiten für die Abhängigkeitsstruktur dieser Folge ergeben die Definitionen

Definition 4.10

Wenn die die Zufallsvariablen X_n unabhängig und identisch verteilt sind, dann heißt $\mathcal{X}$ gedächtnislos.

Wenn (X_n) eine Markovkette bilden oder stationär sind, dann heißt die Quelle Markovquelle bzw. stationäre Quelle. Eine irreduzible Markovquelle nennen wir ergodisch.

Eine wichtige Größe ist die Entropie einer Quelle. Im Sinne der der Idee, die wir bei der Einführung der Entropie verwendet haben, definieren wir sie über die mittlere Codewortlänge beim optimalen Kodieren von langen Blöcken:

Definition 4.11

Die Entropie der Quelle $\mathcal{X}$ ist

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n),$$

wenn dieser Grenzwert existiert (wenn nicht, dann hat $\mathcal{X}$ keine Entropie).

Für eine gedächtnislose Quelle gilt

$$H(\mathcal{X}) = H(X_1).$$

Für eine (irreduzible) Markovquelle mit Übergangsmatrix P ergibt sich

$$H(X) = \sum \pi_i H(P_i),$$

wobei P_i die i-te Zeile von P und π die stationäre Verteilung ist.

Für eine stationäre Quelle existiert die Entropie.

Satz 4.9: Shannon-MacMillan

Für eine stationäre Quelle gilt

$$\lim_{n\to\infty} \frac{1}{n} \log_2 p(X_1, \ldots, X_n) = -nH(\mathcal{X})$$

in Wahrscheinlichkeit.

Das kann man so sehen, dass mit hoher Wahrscheinlichkeit gilt, dass die Wahrscheinlichkeit, genau die Folge $(X_1, \ldots, X_n)$ zu ziehen $\approx 2^{-nH(\mathcal{X})}$ ist.

4.4 Blockcodes

Little boxes on the hillside
Little boxes made of ticky-tacky
Little boxes on the hillside
Little boxes all the same

Pete Seeger

Unsere optimalen Codes haben schöne statistische Eigenschaften, aber ihre variable Codewortlänge sieht etwas unordentlich aus. Codes mit konstanter Länge, also Blockcodes, sind da sehr

viel ansehnlicher. Wir wollen uns ansehen, was sie an Effizienz zu bieten haben, wobei wir wieder annehmen, dass die Buchstaben der Quellnachricht (mit Länge n) unabhängig mit Verteilung P gewählt werden. Unsere erste Erkenntnis ist ernüchternd: wenn wir auf eindeutiger Entzifferbarkeit bestehen, dann brauchen wir mindestens $\log_2(m) = H(m)$ Bits pro Quellzeichen, weil ja für jede mögliche Quellnachricht ein eigenes Codewort zur Verfügung gestellt werden muss. Es muss also $2^l \geq m^n$ und daher $l \geq n \log_2(m)$ gelten. Im Vergleich dazu geben uns die optimalen Codes mit variabler Länge Raten nahe an der Entropie $H(P)$, die im allgemeinen kleiner als $H(m)$ ist.

Wir müssen also bescheidener werden, und das sieht so aus, dass wir statt die unbedingte Entzifferbarkeit zu fordern nur mit einer Wahrscheinlichkeit nahe bei 1 die ursprüngliche Nachricht rekonstruieren können wollen. Dies ist ein erster Schritt ins Reich dessen, was man "verlustbehaftete Kompression" nennt — um bessere Kompressionsraten zu erreichen, begnügt man sich damit, nach der Decodierung etwas zu erhalten, was nur "so ähnlich" ist wie das Original. Für uns ist einstweilen "so ähnlich" gleichbedeutend mit "fast immer dasselbe". Wenn man entsprechendes Wissen aus der Wahrnehmungspsychologie hineinsteckt, lassen sich etwa bei Bildern oder Sounddateien erstaunlich große Teile der ursprünglichen Information weglassen, ohne dass ein merklicher Qualitätsverlust eintritt.

Das Lemma von Fano sagt uns, dass wir keine Rate erreichen können, die wesentlich besser als die Entropie ist. Wir werden sehen, dass wir auch mit Blockcodes beliebig nahe an die Entropie kommen können.

Zuerst wollen wir annehmen, dass der Code c komplett festgelegt ist: es sind also die Parameter $k, n, r(=2), l$ und auch die Abbildung $c : M^n \to \{0,1\}^l$, die dem Quellwort das Codewort zuordnet, fixiert. Wir suchen die optimale Decodierungsfunktion $d : \{0,1\}^l \to M^n$. Wir wollen die Wahrscheinlichkeit, dass korrekt decodiert wird, maximieren, also

$$\mathbb{P}(d(c(X)) = X) = \sum_{x:d(c(x))=x} p(x) = \sum_{y\in\{0,1\}^l} \sum_{x:c(x)=y} [x = d(y)]p(x).$$

Offensichtlich kann $d(y)$ beliebig gewählt werden, wenn y nicht als Codewort auftritt, also wenn es kein x mit $y = c(x)$ gibt. Andernfalls wird die Wahrscheinlichkeit maximal, wenn wir $d(y)$ so wählen, dass $p(d(y))$ das Maximum der Wahrscheinlichkeiten $p(x)$ unter allen x mit $c(x) = y$ ist. Aus nachvollziehbaren Gründen sprechen wir von "Maximum Likelihood Decodierung".

Noch einmal kommt die Maximum Likelihood Idee zu Ehren: zur Festlegung des Codes können wir die 2^l Wörter, die wir korrekt decodieren wollen, frei wählen. Die optimale Wahl besteht offensichtlich darin, die Wörter mit den größten Wahrscheinlichkeiten zu wählen. Der Satz von Shannon-MacMillan sagt uns, dass die (höchstens $2^{n(H(P)+\epsilon}$ Blöcke mit Wahrscheinlichkeit $\geq 2^{-n(H(P)+\epsilon)}$ eine Wahrscheinlichkeit haben, die für $n \to \infty$ gegen 1 geht. Wir können also (für hinreichend großes n) gleichzeitig eine Rate beliebig nahe an der Entropie und eine Fehlerwahrscheinlichkeit beliebig nahe bei 0 erreichen.

Diese Überlegungen führen uns zu der Erkenntnis:

Satz 4.10

Für jedes $\epsilon > 0$ und $\delta > 0$ gibt es für hinreichend großes n einen $(n, m, l, 2)$-Blockcode mit Wortlänge $l \leq n(H(X) + \delta)$ und Fehlerwahrscheinlichkeit kleiner als ϵ.

4.5 Kanalcodierung

4.6 Natürliche Sprachen als Informationsquellen

Kapitel 5

Statistik

5.1 Motivation: Wahlumfragen

Hier in Österreich finden Wahlen oft relativ spät im Jahr statt. Deshalb es ist nicht ganz unwahrscheinlich, dass zu der Zeit, zu der dieser Stoff durchgenommen wird, irgendeine Wahl kurz bevorsteht oder gerade vorbei ist. Zwei Dinge, die dabei von einer gewissen Mystik umgeben sind, sind die beiden Wege, wie Voraussagen über den Ausgang von solchen Wahlen getroffen werden: da sind einerseits die Umfragen, die schon vor der Wahl (und zum Teil auch als "Exit-Polls" am Wahltag selbst) kursieren, und andererseits die Hochrechnungen, die erst am Wahlabend auf der Grundlage der schon bekannten Stimmen während der Auszählung berechnet werden. Bei diesen besteht ein gewisses Missverhältnis: Umfragen sind notorisch inexakt, wogegen die Hochrechnungen erstaunlich präzise sind. Trotzdem wollen wir uns hier nur mit den Umfragen näher beschäftigen, das Konzept, das hinter den Hochrechnungen steht, ist komplizierter, weil dort Erfahrungen aus früheren Wahlgängen mit verwendet werden.

Wir werden uns die Sache auch gleich noch einfacher machen, indem wir uns nur um eine einzige Partei kümmern, über deren WählerInnenanteil wir etwas erfahren wollen: wir haben also eine große Anzahl N von Wahlberechtigten, und ein Prozentsatz p davon wählt die gefragte Partei. Es ist zumindest konzeptuell nicht allzu schwer, p herauszufinden: man muss einfach nur alle N Personen befragen, ob Sie dieser Partei ihre Stimme geben werden. Das ist allerdings langwierig und teuer, etwas, das man alle vier Jahre bei der Wahl macht und nicht jede Woche, um das Ergebnis der bevorstehenden Wahl vorherzusagen. Viel schöner wäre es doch, wenn man mit deutlich weniger als eine solchen "Vollerhebung" auskäme. Man könnte etwa versuchen, die Leser der eigenen Zeitung aufzufordern, ihre Meinung per Post oder am Telefon der Redaktion mitzuteilen. Das ist durchaus schon gemacht worden und wird auch immer wieder noch gemacht, hat aber gleich zwei Nachteile: einerseits könnten manche Gruppen der Bevölkerung ein stärkeres Bedürfnis haben, ihre Meinung kundzutun, als andere, und wenn dort der Anteil an Wählern unserer Partei höher oder niedriger ist als bei anderen, ergibt sich bei unserer Umfrage ein verzerrtes Bild. Andererseits kann es auch der Fall sein, dass die Personen, die schreiben oder anrufen, nicht die Wahrheit über ihre Präferenzen sagen. Das zweite Problem ist eines, das jede Art von Vorhersage hat (und eines, das die Hochrechnung nicht hat, die verwendet ja nur die tatsächlich abgegebenen Stimmen, wenn auch nicht alle), aber am ersten kann man arbeiten: wenn man etwa weiß, dass ein Drittel der Bevölkerung in großen Städten wohnt und zwei Drittel auf dem Land, dann wird man versuchen, diese Anteile auch in der Stichprobe wiederzugeben. Hier ist allerdings nicht immer klar, welche der vielen Eigenschaften, die es gibt, mit der Wahlentscheidung im Zusammenhang stehen. Wenn man hier zu viele Merkmale zu berücksichtigen versucht, wird die Sache kompliziert.

In dieser Situation kommt uns der Zufall zu Hilfe: wenn wir eine Person zufällig (gleichverteilt) auswählen, also jede der N Personen mit Wahrscheinlichkeit $1/N$, dann erwischen wir mit Wahrscheinlichkeit p eine Person, die unsere Partei wählt. Wenn wir das n mal unabhängig wiederholen, dann ist die Anzahl X der Wählerinnen und Wähler der fraglichen Partei in unserer Stichprobe binomialverteilt; der relative Anteil X/n hat Erwartungswert p und Varianz $\frac{p(1-p)}{n}$. Im Lichte des zentralen Grenzwertsatzes bzw. der Ungleichung von Chebychev dürfen wir also erwarten, dass der Fehler in der Größenordnung $1/\sqrt{n}$ liegt. Wenn wir n hinreichend groß wählen, dann können wir

diesen Fehler beliebig klein machen, zumindest mit großer Wahrscheinlichkeit. Ganz grob kann man etwa so rechnen: wir wollen, dass wir mit 95-prozentiger Wahrscheinlichkeit aus unserer Stichprobe einen Schätzwert für p erhalten, der sich um höchsten 5 Prozentpunkte von p unterscheidet. Wir nähern die Verteilung von X bzw. von X/n durch eine Normalverteilung an. Von dieser wissen Wir, dass sich zwischen $\mu - 2\sigma$ und $\mu + 2\sigma$ 95% der Verteilung befinden(eigentlich sollte hier statt 2 1.96 stehen, aber so tun wir uns mit dem Rechnen leichter). Für die Berechnung von

$$\sigma = \frac{p(1-p)}{n}$$

brauchen wir eigentlich p, das wir nicht kennen. Aber das Produkt $p(1-p)$ ist nicht schwer abzuschätzen: bekanntlich nimmt das für $p = 1/2$ seinen Maximalwert $1/4$ an. Die 2σ, die wir zur Abschätzung des Fehlers verwenden, sind also nicht mehr als $1/\sqrt{n}$. Das soll nicht mehr als 0.05 (5 Prozentpunkte) ausmachen, also sollten wir $n \geq 1/0.05^2 = 400$ wählen. Wären wir weniger bescheiden gewesen, und hätten etwa eine Genauigkeit von einem Prozentpunkt verlangt, hätten wir $n \geq 10000$ verwenden müssen. Dabei nehmen wir immer noch in Kauf, dass mit 5% Wahrscheinlichkeit der Fahler größer als diese Schranke sein kann. Wenn wir diese Wahrscheinlichkeit kleiner machen wollen, müssen wir n natürlich auch vergrößern, aber diese Entscheidung hat einen weniger dramatischen Einfluss. Die beiden Zahlen, die wir bekommen haben, geben ungefähr die Grenzen dessen wieder, was in der Praxis geschieht: die meisten Umfragen arbeiten mit einer Anzahl von Befragten, die zwischen 200 und 2000 liegt. Was darunter liegt, ist zu ungenau, und mehr ist meistens zu teuer.

5.2 Grundlagen der Statistik

Das Beispiel der Wahlumfragen gibt uns einige Anhaltspunkte für die Disziplin, zu der solche Umfragen gehören:

Die Statistik (genauer: die schließende Statistik, mit der wir uns hier beschäftigen; es gibt auch die beschreibende Statistik mit der Aufgabe, große Datenmengen überschaubar zusammenzufassen) hat die Aufgabe, aufgrund einer Stichprobe Aussagen über die Grundgesamtheit zu treffen. Wenn aus einer endlichen Menge mit Zurücklegen gezogen wird, dann sind die einzelnen Ziehungsergebnisse unabhängig und haben als Verteilung die Häufigkeitsverteilung aus der Grundgesamtheit. Unsere Annahmen über diese zugrundeliegende Verteilung fassen wir in einem statistischen Modell zusammen:

Definition 5.1

Ein statistisches Modell ist eine Menge $\mathcal{P}$ von Verteilungen. Wenn diese Verteilungen durch endlich viele reelle Zahlen (die Parameter) beschrieben werden können, also

$$\mathcal{P} = \{P_\theta, \theta \in \Theta\}$$

mit $\Theta \subseteq \mathbb{R}^d$, dann sprechen wir von einem parametrischen Modell, sonst von einem nichtparametrischen Modell.

Eine Stichprobe ist eine Folge $(X_1, \ldots, X_n)$ von unabhängigen Zufallsvariablen mit einer (unbekannten) Verteilung aus $\mathcal{P}$.

Wenn wir aus einer endlichen Grundgesamtheit ziehen, dann kann das Ergebnis nur einen von endlich vielen möglichen Werten haben, es gibt in diesem Zusammenhang also eigentlich nur diskrete Verteilungen. Wir lassen aber für unsere statistischen Modelle auch stetige Verteilungen zu — das kann man entweder als Annäherung an die tatsächlichen Verhältnisse sehen, oder man kann sich vorstellen, dass die Elemente der Grundgesamtheit selbst durch einen Zufallsmechanismus zustandekommen: wenn wir etwa die Körperlänge oder das Gewicht eines Babyelefanten betrachten, dann können wir uns vorstellen (der Elefant wächst jeden Tag ein bisserl), dass diese Größe als eine Summe von vielen kleinen Zuwächsen zustande kommt, die mehr oder weniger unabhängig voneinander sind. Der zentrale Grenzwertsatz macht es dann plausibel, dass die Normalverteilung für solche Messwerte zumindest eine brauchbare Näherung sein sollte. Diese Vorstellung ist einer der

Gründe dafür, dass die Normalverteilung in der klassischen Statistik eine zentrale Rolle spielt (ein zweiter Grund liegt in der Tatsache, dass die Normalverteilung rechnerisch sehr gut zu behandeln ist).

Damit stehen wir vor zwei klassischen Standardmodellen:

Beispiel 5.1: Die Alternativverteilung

Unsere Überlegungen zu den Wahlumfragen und viele andere Fragen, in denen wir es mit einer Wahl zwischen zwei Möglichkeiten zu tun haben (ja/nein, gut/schlecht, etc.), können wir eine der beiden Möglichkeiten mit "0" codieren, die andere mit "1". So erhalten wir für die Beobachtungen eine Alternativverteilung $A(p)$, die offenbar durch die eine Zahl p zwischen 0 und 1 beschrieben werden kann. Wir haben es also hier mit einem parametrischen Modell zu tun, und der Parameterraum ist $\Theta = [0, 1]$. Aus Tradition schreiben wir hier für den Parameter nicht θ, sondern p. Ähnliches gilt für die anderen klassischen parametrischen Familien, wie die Exponential-, Gamma- oder Poissonverteilung. Auch dort werden wir statt θ die üblichen Namen der Parameter verwenden.

Wenn wir keine parametrischen Modelle verwenden wollen oder können, bleibt uns eigentlich nichts anderes übrig als jede mögliche Verteilung zuzulassen. Diese können wir durch ihre Verteilungsfunktion F repräsentieren. Diese Funktion hat für jedes x einen Wert $F(x)$, den man zwar nicht ganz frei wählen kann (schließlich ist F nichtfallend und rechtsstetig), aber offensichtlich sind endlich die Werte von F an endlich vielen Punkten $x_1, \ldots, x_n$ nicht genug, um F festzulegen. Im Gegensatz dazu sind die Verteilungen in einem parametrischen Modell durch endlich viele reelle Zahlen festgelegt. In diesem Sinn sind parametrische Modelle von endlicher Dimension, nichtparametrische unendlichdimensional.

Beispiel 5.2: Die Normalverteilung

Wir haben schon besprochen, dass es mehr oder weniger gute Gründe gibt, warum Ergebnisse von Messungen zumindest näherungsweise normalverteilt sein sollten. Das macht die Normalverteilung zu einem der wichtigsten und auch am meisten verwendeten statistischen Modelle. Dieses Modell hat zwei Parameter: μ und σ^2, wobei μ eine beliebige reelle Zahl sein darf und σ^2 positiv. Wie in anderen Fällen bezeichnen wir die Parameter mit (μ, σ^2) und nicht mit (θ_1, θ_2). In jedem Fall ist der Parameterraum $\Theta = \mathbb{R} \times \mathbb{R}_+$.

In der mathematischen Statistik dreht sich alles um eine Stichprobe. Insbesondere werden wir mit den Zahlen in der Stichprobe Berechnungen vornehmen. Das ist so häufig der Fall, dass es uns eine eigene Definition wert ist:

Definition 5.2

Eine Statistik T ist eine Zufallsvariable, die aus der Stichprobe berechnet werden kann:

$$T = T(X_1, \ldots, X_n)$$

Insbesondere dürfen in dieser Funktion die unbekannten Parameter nicht vorkommen.

Beispiel 5.3

Wenn wir die Normalverteilung als Modell zugrundelegen, dann sind die folgenden Zufallsvariablen Statistiken:

$$T_1 = X_1 - X_2, T_2 = \bar{X}_n = \frac{X_1 + \ldots + X_n}{n}, T_3 = S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2,$$

$$T_4 = \max(X_1, \ldots, X_n).$$

Andererseits sind

$$T_5 = \frac{X_1 - \mu}{\sigma}, T_6 = X_1 - \mu, T_7 = \frac{X_1 - X_2}{\sigma}$$

keine Statistiken, weil in Ihrer Definition die unbekannten Parameter μ und σ^2 auftreten. Wenn wir aber, wie wir es gelegentlich (und nicht sehr realitätsnah, aber rechnerisch angenehm) tun werden, annehmen, dass σ^2 bekannt ist, dann wird T_7 doch zu einer Statistik, weil ja σ nicht mehr unbekannt ist.

Wenn wir nun ein bestimmtes statistisches Modell $\mathcal{P}$ zugrundelegen, dann hängt die Verteilung von $(X_1, \ldots, X_n)$ und damit auch die einer Statistik T von der (unbekannten) Verteilung $P \in \mathcal{P}$ ab, nach der die Beobachtungen $X_1, \ldots, X_n$ verteilt sind, im Falle eines parametrischen Modells wird diese Verteilung durch einen Parameter $\theta \in \Theta$ festgelegt. Wenn wir nun ein Ereignis A oder eine Zufallsvariable Y durch $X_1, \ldots, X_n$ definieren können, dann hängt die Wahrscheinlichkeit von A bzw. der Erwartungswert von Y von P bzw. θ ab. Um diese Abhängigkeit deutlich zu machen, verwenden wir die Notationen $\mathbb{P}_P, \mathbb{E}_P, \mathbb{P}_\theta, \mathbb{E}_\theta$ und auch $\mathbb{V}_P$ und $\mathbb{V}_\theta$ für die Varianz.

5.3 Schätztheorie

5.3.1 Punktschätzung

Die erste Aufgabe, mit der wir uns beschäftigen, besteht darin, aus einer Stichprobe Schätzwerte für den unbekannten Parameter zu bestimmen. Wir definieren

Definition 5.3

Ein Schätzer ist eine Folge $(\hat{\theta}_n)$ von Statistiken $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$.

Anmerkungen

1. Diese Definition erscheint auf den ersten Blick ein wenig kompliziert. Der Sinn ist, dass wir für jeden möglichen Stichprobenumfang eine Schätzstatistik $\hat{\theta}_n$ haben, und speziell die Möglichkeit haben, n gegen unendlich gehen zu lassen.
2. Neben dem Schätzen eines Parameters kann man die etwas allgemeinere Frage betrachten, dass nicht der Parameter selbst, sondern eine Funktion $\tau = \tau(\theta)$ geschätzt werden soll.
3. Es kann sein, dass für einige (kleine) Werte des Stichprobenumfangs n der Schätzer nicht definiert ist. Es macht etwa die Formel für die Stichprobenvarianz mit dem Nenner $n-1$ für $n = 1$ keinen Sinn. Da wir hauptsächlich an großen Werten von n interessiert sind, macht uns das in den meisten Fällen nichts aus. Wenn man ganz genau sein will, kann man für die fehlenden Werte per Fallunterscheidung einen Ersatz angeben.
4. Die Definition lässt recht dumme Schätzer zu (z.B. eine Konstante, die gar keine Rücksicht auf die Stichprobe nimmet, etwa 42, vgl. Douglas Adams), deshalb definieren wir einige Eigenschaften, die wir von Schätzern verlangen können:

Definition 5.4

Ein Schätzer $\hat{\theta}_n$ heißt

- schwach konsistent, wenn $\hat{\theta}_n \to \theta$ in Wahrscheinlichkeit konvergiert (also $\mathbb{P}_\theta(|\hat{\theta}_n - \theta| > \epsilon) \to 0$ für alle $\epsilon > 0$).
- stark konsistent, wenn $\hat{\theta}_n \to \theta$ mit Wahrscheinlichkeit 1 konvergiert (also $\mathbb{P}_\theta(\hat{\theta}_n \to \theta) = 1$.
- erwartungstreu, wenn für alle $\theta \in \Theta$

$$\mathbb{E}_\theta(\hat{\theta}_n) = \theta.$$

- effizient, wenn er erwartungstreu ist und die kleinste Varianz unter allen erwartungstreuen Schätzern hat, also wenn für jeden weiteren erwartungstreuen Schätzer $\tilde{\theta}_n$

$$\mathbb{V}_\theta(\hat{\theta}_n) \leq \mathbb{V}_\theta(\tilde{\theta}_n) \text{ für alle } \theta. \tag{5.1}$$

Für viele klassische Verteilungen existieren erwartungstreue und sogar effiziente Schätzer. Das ist nicht selbstverständlich; die Effizienz verlangt ja, dass $\mathbb{V}_\theta(\hat{\theta}_n)$ für alle θ minimal ist. Es kann aber durchaus der Fall eintreten, dass für einen Wert von θ ein Schätzer minimale Varianz hat, für einen anderen Wert von θ ein anderer, und keiner kann beides. Es kommt aber noch schlimmer: auch erwartungstreue Schätzer müssen nicht existieren. Beispiele dazu finden Sie am Ende des Kapitels.

Nach so viel Negativem und Kompliziertem sehen wir uns einige einfache Beispiele an:

Beispiel 5.4: Das Stichprobenmittel

Wir betrachten eine Stichprobe aus einer Verteilung mit Erwartungswert μ (dieser ist im allgemeinen eine Funktion des Parameters θ). Das Stichprobenmittel

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

hat Erwartungswert μ, und das Gesetz der großen Zahlen sagt uns, das $\bar{X}_n$ für $n \to \infty$ mit Wahrscheinlichkeit eins gegen μ konvergiert. Das Stichprobenmittel ist also ein erwartungstreuer und konsistenter Schätzer für den Erwartungswert. Ob er effizient ist, lässt sich nicht in allgemeiner Form beantworten, die Antwort auf diese Frage hängt vom zugrundeliegenden Modell ab.

Beispiel 5.5: Die Stichprobenvarianz

Wir nehmen jetzt auch noch an, dass $\sigma^2 = \mathbb{V}(X)$ endlich ist.

Wie das Stichprobenmittel für den Erwartungswert ist

$$T_1 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

ein erwartungstreuer und konsistenter Schätzer für $\mathbb{E}(X^2)$. Daraus folgt, dass

$$T_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = T_1 - \bar{X}_n^2$$

ein konsistenter Schätzer für die Varianz ist. Er ist allerdings nicht erwartungstreu. Sein Erwartungswert ist

$$\mathbb{E}(T_2) = \mathbb{E}(X^2) - \mathbb{E}(\bar{X}_n^2) = \mu^2 + \sigma^2 - (\mu^2 + \frac{\sigma^2}{n}) = \sigma^2\frac{n-1}{n}.$$

Für Stichprobenumfänge $n > 1$ kann man daraus einen erwartungstreuen Schätzer machen, indem man den störenden Faktor wegmultipliziert:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Diese Statistik heißt Stichprobenvarianz und ist ein erwartungstreuer Schätzer für die Varianz.

Die Überlegungen zur Konsistenz des Stichprobenmittels liefern uns eine Methode, um Schätzer zu konstruieren, die Momentenmethode. Sie nützt aus, dass das Stichprobenmittel konsistent für den Erwartungswert ist.

Definition 5.5: Momentenmethode

θ sei ein einzelner reeller Parameter (also der Parameterraum Θ eindimensional). Dann können wir den Erwartungswert von X als Funktion von θ schreiben:

$$\mathbb{E}_\theta(X) = m(\theta).$$

Wenn die Funktion m stetig umkehrbar ist, dann ist

$$\hat{\theta}_n = m^{-1}(\bar{X}_n)$$

ein (stark) konsistenter Schätzer, der Momentenschätzer.

Wenn es $d > 1$ Parameter gibt, dann verwendet man zusätzlich höhere Momente:

$$m_i(\theta) = m_i(\theta_1, \ldots, \theta_d) = \mathbb{E}_\theta(X^i).$$

Wir ersetzen die theoretischen Momente durch ihre Schätzer, das ergibt die d Gleichungen

$$m_i(\theta_1, \ldots, \theta_d) = \frac{1}{n}\sum_{j=1}^{n} X_j^i,\ i = 1, \ldots, d$$

in den d Variablen $\theta_1, \ldots, \theta_d$. Auflösen dieses Gleichungssytems nach θ ergibt den Momentenschätzer für $\theta = (\theta_1, \ldots, \theta_d)$. Wenn das Ergebnis stetig von den rechten Seiten des Gleichungssystems abhängt, erhalten wir dadurch wieder einen konsistenten Schätzer.

Beispiel 5.6

Wir betrachten die Normalverteilung. Diese hat die zwei Parameter μ und σ^2. Wir brauchen also auch zwei Gleichungen, und dafür verwenden wir die ersten beiden Momente:

$$\mathbb{E}_{\mu,\sigma^2}(X) = \mu$$

und

$$\mathbb{E}_{\mu,\sigma^2}(X^2) = \mu^2 + \sigma^2.$$

Das führt uns zu den Gleichungen

$$\hat{\mu} = \bar{X}_n$$

und

$$\hat{\mu}^2 + \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Das ist leicht aufzulösen und gibt

$$\hat{\mu} = \bar{X}_n$$

und

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2.$$

Das sind die Schätzer für Erwartungswert und Varianz, die wir schon früher besprochen haben.

Für die zweite Methode zur Konstruktion von Schätzern betrachten wir ein diskretes Modell mit der Wahrscheinlichkeitsfunktion $p_\theta(x) = \mathbb{P}_\theta(X_i = x)$. Die gemeinsame Verteilung der Werte in der Stichprobe ist durch die gemeinsame Wahrscheinlichkeitsfunktion gegeben. In diesem Zusammenhang nennen wir sie die Likelihoodfunktion:

$$L(x_1, \ldots, x_n; \theta) = \mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n) = p_\theta(x_1)p_\theta(x_2)\ldots p_\theta(x_n).$$

Dass hier für etwas, das es schon gibt, ein neuer Name vergeben wird, hat zwei Gründe: einerseits wollen wir einen einheitlichen Namen für den stetigen und den diskreten Fall haben, und

andererseits wollen wir die Stichprobe selbst in diese Funktion einsetzen. Im diskreten Fall, den wir hier betrachten, ist das die Wahrscheinlichkeit, *genau diese* Stichprobe zu ziehen. Wenn diese Wahrscheinlichkeit für einen Wert von θ gleich 0 ist, dann ist es unmöglich (so gut wie), diese Stichprobe zu erhalten, also kann dieser Wert von θ nicht der wahre Wert sein. Dieses extreme Beispiel macht es plausibel, dass kleine Werte der Likelihoodfunktion gegen einen bestimmten Parameter sprechen, und also umgekehrt große Werte dafür. Der Wert des Parameters, für den am meisten spricht, soll dann unser Schätzer sein. Wir definieren also

Definition 5.6: Likelihoodfunktion

Die Likelihoodfunktion ist

$$L(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} p_\theta(X_i),$$

wenn P_θ diskret mit Wahrscheinlichkeitsfunktion p ist, und

$$L(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} f_\theta(X_i),$$

wenn P_θ stetig mit Dichte f ist.

Definition 5.7: Maximum-Likelihood-Methode

Der Maximum-Likelihood (ML-) Schätzer ist der Wert von θ, der die Likelihoodfunktion maximiert.

Wir betrachten wieder unsere Lieblingsverteilung:

Beispiel 5.7

Die Normalverteilung $N(\mu, \sigma^2)$ hat die Dichte

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

Daraus ergibt sich die Likelihoodfunktion

$$L(X_1, \ldots, X_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(X_i-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \exp\left(-\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}.\right)$$

Diese Funktion soll maximiert werden. Oft, und auch hier, ist das klassische Verfahren "Ableiten und Nullsetzen" zielführend, und dann ist es hilfreich, zuerst zu logarithmieren.

$$\log(L(X_1, \ldots, X_n; \mu, \sigma^2)) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}.$$

Wir leiten nach μ und σ ab und setzen die Ableitungen gleich 0:

$$\frac{\partial \log(L(X_1, \ldots, X_n; \mu, \sigma^2))}{\partial \mu} = \sum_{i=1}^{n} \frac{\mu - X_i}{\sigma^2} = 0$$

und

$$\frac{\partial \log(L(X_1, \ldots, X_n; \mu, \sigma^2))}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^3} = 0.$$

Aus der ersten Gleichung ergibt sich

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n,$$

und aus der zweiten,

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2,$$

und Einsetzen liefert

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Also wieder etwas, das wir schon kennen.

Beim nächsten Beispiel funktioniert die Standardmethode "Ableiten und Nullsetzen" nicht, aber wir können immer noch die Likelihoodfunktion maximieren:

Beispiel 5.8

Wir betrachten die Gleichverteilung auf $[0, \theta]$ für $\theta > 0$. Hier ist die Dichte

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} & \text{wenn } 0 \le x \le \theta, \\ 0 & \text{sonst.} \end{cases}$$

Daraus ergibt sich die Likelihoodfunktion

$$L(X_1, \ldots, X_n; \theta) = \begin{cases} \frac{1}{\theta^n} & \text{wenn } 0 \le X_i \le \theta \text{ für alle } i = 1, \ldots, n, \\ 0 & \text{sonst.} \end{cases}$$

$\frac{1}{\theta^n}$ ist eine fallende Funktion von θ. Um die Likelihoodfunktion zu maximieren, müssen wir also θ möglichst klein wählen. Dabei müssen wir darauf achten, dass für alle i die Ungleichung $X_i \le \theta$ gelten muss, weil sonst die Likelihoodfunktion den Wert 0 hat. Diese Bedingung ist äquivalent zu $\max(X_1, \ldots, X_n) \le \theta$, also erhalten wir als kleinstes θ, für das die Likelihoodfunktion positiv ist,

$$\hat{\theta}_n = \max(X_1, \ldots, X_n).$$

Die Schätzer, die wir in den letzten Beispielen erhalten haben, sind konsistent. Das ist nicht selbstverständlich: der Maximum-Likelihood-Schätzer muss nicht konsistent sein. Auch dafür findet sich am Ende des Kapitels ein Beispiel.

Wenn man nach effizienten Schätzern sucht, ist es gut zu wissen, wie klein die Varianz eines erwartungstreuen Schätzers sein kann. Mithilfe der Likelihoodfunktion lässt sich in vielen Fällen eine untere Abschätzung finden, das leistet der folgende Satz:

Satz 5.1: Cramér-Rao

Wenn p_θ bzw. f_θ zweimal nach θ differenzierbar ist und zusätzliche Regularitätsvoraussetzungen erfüllt, dann gilt für jeden Erwartungstreuen Schätzer $\hat{\theta}_n$

$$\mathbb{V}(\hat{\theta}_n) \ge \frac{1}{I_n(\theta)} = \frac{1}{nI(\theta)}.$$

Dabei ist

$$I_n(\theta) = \mathbb{E}_\theta\left(\left(\frac{\partial \log(L(X_1, \ldots, X_n; \theta))}{\partial \theta}\right)^2\right) = -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial \theta^2} \log(L(X_1, \ldots, X_n; \theta))\right)$$

und $I(\theta) = I_1(\theta)$, also

$$I(\theta) = -\mathbb{E}_\theta((\frac{\partial \log(f_\theta(X))}{\partial \theta})^2) = -\mathbb{E}_\theta(\frac{\partial^2}{\partial \theta^2} \log(f_\theta(X)))$$

bzw.

$$I(\theta) = -\mathbb{E}_\theta((\frac{\partial \log(p_\theta(X))}{\partial \theta})^2) = -\mathbb{E}_\theta(\frac{\partial^2}{\partial \theta^2} \log(p_\theta(X))).$$

$I(\theta)$ bzw. $I_n(\theta)$ wird als die Fisherinformation bezeichnet.

Anmerkung: Es ist nicht ganz einfach, die "Regularitätsvoraussetzungen" auszuschreiben; ein wichtiger Teil davon ist, dass die Menge der x, für die $p_\theta(x)$ bzw. $f_\theta(x)$ den Wert 0 hat, für alle θ dieselbe ist. In diesem Fall kann man nämlich guten Gewissens die Punkte, an denen die Likelihoodfunktion den Wert 0 hat, ausschließen und hat so keine Probleme mit dem Logarithmieren. Ein klassisches Beispiel, in dem das nicht der Fall ist, ist die Gleichverteilung auf $[0, \theta]$. Dort gibt es dann auch einen erwartungstreuen Schätzer mit einer "zu kleinen" Varianz: wenn der Maximum-Likelihood-Schätzer so korrigiert wird, dass er erwartungstreu ist, hat er eine Varianz, die wie $1/n^2$ gegen 0 geht; wenn der Satz von Cramér-Rao anwendbar ist, kann diese aber nicht schneller als $1/n$ fallen.

Für den Beweis nehmen wir an, dass X_n diskret ist mit endlich vielen möglichen Werten. Wir schreiben kurz x für $(x_1, \ldots, x_n)$ und $L(x; \theta)$ für $L(x_1, \ldots, x_n, \theta)$, etc. Wenn nun

$$\hat{\theta}_n = T(X_1, \ldots, X_n)$$

ein erwartungstreuer Schätzer ist, dann ist

$$\mathbb{E}_\theta(\hat{\theta}_n) = \sum_x T(x) L(x; \theta) = \theta.$$

Aus den Eigenschaften der Wahrscheinlichkeitsfunktion wissen wir

$$\sum_x L(x, \theta) = 1.$$

Weil die Summen endlich sind, haben wir hier kein Problem damit, dass wir nach θ differenzieren (und dies ist der Ort, an dem im allgemeinen Fall die "Regularitätsbedingungen" ins Spiel kommen, nämlich um das Differenzieren in der unendlichen Summe bzw. dem Integral zu rechtfertigen):

$$\sum_x T(x) \frac{\partial L(x, \theta)}{\partial \theta} = 1$$

und

$$\sum_x \frac{\partial L(x, \theta)}{\partial \theta} = 0.$$

Wir multiplizieren die zweite Gleichung mit θ und subtrahieren sie von der ersten:

$$1 = \sum_x (T(x) - \theta) \frac{\partial L(x, \theta)}{\partial \theta} = \sum_x (T(x) - \theta) \frac{\partial \log L(x, \theta)}{\partial \theta} L(x, \theta) = \mathbb{E}_\theta((\hat{\theta}_n - \theta) \frac{\partial \log L(X_1, \ldots, X_n; \theta)}{\partial \theta}).$$

Die Cauchy-Schwarz Ungleichung ergibt

$$1 \leq \mathbb{E}_\theta((\hat{\theta}_n - \theta)^2) \mathbb{E}\left(\left(\frac{\partial \log L(X_1, \ldots, X_n; \theta)}{\partial \theta}\right)^2\right) = \mathbb{V}_\theta(\hat{\theta}_n) I_n(\theta).$$

Wenn wir einen erwartungstreuen Schätzer finden können, dessen Varianz mit der Cramér-Rao Schranke übereinstimmt, dann können wir sicher sein, dass er effizient ist. Allerdings gibt es einen solchen Schätzer nicht immer, die Dichte bzw. Wahrscheinlichkeitsfunktion muss dazu von einer bestimmten Form sein.

Und hier können wir zur Rehabilitation des ML-Schätzers antreten: Wenn es einen erwartungstreuen Schätzer gibt, der die Cramér-Rao Schranke erreicht, dann stimmt er mit dem Maximum-Likelihood Schätzer überein. Außerdem lässt sich zeigen, dass der ML-Schätzer — wieder unter geeigneten Regularitätsvoraussetzungen — asymptotisch normalverteilt ist mit Mittel θ und Varianz $1/I_n(\theta)$, also gewissermaßen "asymptotisch effizient".

Für uns ist der Satz von Cramér-Rao das Mittel der Wahl, um Effizienz nachzuweisen.

Beispiel 5.9

Wir betrachten als Beispiel wieder die Normalverteilung, speziell die Schätzung des Mittelwerts μ: der Logarithmus der Likelihoodfunktion ist uns schon bekannt:

$$\log(L(X_1,\ldots,X_n;\mu,\sigma^2)) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{\sum_{i=1}^n (X_i-\mu)^2}{2\sigma^2},$$

und auch seine Ableitung nach μ:

$$\frac{\partial \log(L(X_1,\ldots,X_n;\mu,\sigma^2))}{\partial \mu} = \sum_{i=1}^n \frac{X_i-\mu}{\sigma^2}.$$

Wir leiten noch einmal nach μ ab:

$$\frac{\partial^2 \log(L(X_1,\ldots,X_n;\mu,\sigma^2))}{\partial \mu^2} = -\frac{n}{\sigma^2}.$$

Zur Berechnung der Cramér-Rao Schranke können wir die Erwartung des Quadrats der ersten Ableitung bilden — das führt auf die Varianz der Summe $\sum_{i=1}^n X_i$, die durch σ^4 geteilt wird — oder die der zweiten Ableitung — diese ist praktischerweise konstant und damit ihr eigener Erwartungswert. In jedem Fall ist

$$I_n(\mu) = \frac{n}{\sigma^2}$$

und die Cramér-Rao Schranke

$$\frac{1}{I_n} = \frac{\sigma^2}{n}.$$

Wir wissen bereits, dass das Stichprobenmittel $\bar{X}_n$, das wir sowohl als Momenten- als auch als Maximum-Likelihood-Schätzer erhalten haben, erwartungstreu ist. Es ist auch bekannt, dass seine Varianz

$$\mathbb{V}(\bar{X}_n) = \frac{\mathbb{V}(X_1)}{n} = \frac{\sigma^2}{n}$$

ist. Das stimmt mit der Cramér-Rao Schranke überein, daher ist $\bar{X}_n$ in diesem Fall effizient.

Beispiel 5.10

Wir betrachten nun auch ein diskretes Beispiel, die Alternativverteilung $A(p)$. Ihre Wahrscheinlichkeitsfunktion ist

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

Wenn wir bedenken, dass x nur die Werte 0 und 1 annimmt, können wir das auch als

$$p_X(x) = p^x(1-p)^{1-x}$$

schreiben. Damit wird die Likelihoodfunktion zu

$$L(X_1,\ldots,X_n,p) = \prod_{i=1}^n p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}.$$

Wir logarithmieren und setzen die Ableitung gleich 0:

$$0 = \frac{\partial}{\partial p}\log(L) = \frac{\partial}{\partial p}(\sum_{i=1}^{n} X_i \log(p) + (n - \sum_{i=1}^{n} X_i)\log(1-p)) = \frac{\sum_{i=1}^n X_i}{p} - \frac{n - \sum_{i=1}^n X_i}{1-p} =$$

$$\frac{\sum_{i=1}^n X_i - np}{p(1-p)},$$

also

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}_n.$$

Für die Bestimmung der Cramér-Rao Schranke bestimmen wir die zweite Ableitung der logarithmierten Likelihoodfunktion:

$$\frac{\partial^2}{\partial p^2}\log(L) = -\frac{\sum_{i=1}^n X_i}{p^2} - \frac{n - \sum_{i=1}^n X_i}{(1-p)^2},$$

diese hat Erwartungswert

$$-\frac{np}{p^2} - \frac{n-np}{(1-p)^2} = -\frac{n}{p} - \frac{n}{1-p} = -\frac{n}{p(1-p)},$$

daher ist die Cramér-Rao Schranke

$$\frac{1}{I_n} = \frac{p(1-p)}{n},$$

und das ist genau die Varianz von $\hat{p} = \bar{X}_n$. Auch hier ist also $\hat{p}$ effizient.

5.3.2 Suffizienz

Ein besonders interessanter Fall liegt dann vor, wenn eine eine einzelne Statistik schon alle Information enthält, die die Stichprobe über den Parameter liefern kann. Formal bedeutet das

Definition 5.8

Die Statistik $T = T(X_1, \ldots, X_n)$ ist *suffizient* für den Parameter θ, wenn die bedingte Verteilung von $(X_1, \ldots, X_n)$ unter T nicht von θ abhängt.

Satz 5.2: Faktorisierungssatz von Neyman

T ist genau dann suffizient für θ, wenn die Likelihoodfunktion in der Form

$$L(X_1, \ldots, X_n; \theta) = g(T, \theta)h(X_1, \ldots, X_n)$$

dargestellt werden kann.

Beispiel 5.11

Die Exponentialverteilung hat die Likelihoodfunktion

$$L(X_1, \ldots, X_n; \lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n X_i}[X_i \geq 0, i = 1, \ldots, n].$$

Das passt in den Faktorisierungssatz mit

$$g(t, \lambda) = \lambda^n e^{-\lambda t},$$

$$h(x_1,\dots,x_n) = \begin{cases} 1 & x_i \geq 0, i = 1,\dots,n \\ 0 & \text{sonst} \end{cases}$$

und $T = \sum_{i=1}^{n} X_i$. T ist also suffizient.

Beispiel 5.12

Für die Gleichverteilung $U(0,\theta)$ kann die Likelihoodfunktion als

$$L(X_1,\dots,X_n,\theta) = \begin{cases} \frac{1}{\theta^n} & X_i \geq 0, i = 1,\dots,n, \max(X_1,\dots,X_n) \leq \theta, \\ 0 & \text{sonst} \end{cases}$$

geschrieben werden. Hier ist $T = \max(X_1,\dots,X_n)$ suffizient.

Beispiel 5.13

Die Likelihoodfunktion der Normalverteilung ist

$$L(X_1,\dots,X_n;\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\frac{(n-1)S_n^2 + n(\bar{X}_n - \mu)^2}{2\sigma^2}).$$

Es ist also das Paar $(\bar{X}_n, S_n^2)$ suffizient für den zweidimensionalen Parameter $\theta = (\mu,\sigma^2)$.

Wenn es eine suffiziente Statistik gibt, dann lassen sich viele wichtige Größen als Funktion der suffizienten Statistik schreiben. Man sieht etwa unmittelbar, dass der Maximum-Likelihood-Schätzer eine Funktion der suffizienten Statistik ist. Auch bei der Suche nach einem effizienten Schätzer kann eine suffiziente Statistik helfen, es gilt nämlich

Satz 5.3

Zu jedem Schätzer $\hat{\theta}$ gibt es einen Schätzer $\tilde{\theta}$, der eine Funktion der suffizienten Statistik ist und

$$\mathbb{E}_\theta(\tilde{\theta}) = \mathbb{E}_\theta(\hat{\theta}), \mathbb{V}_\theta(\tilde{\theta}) \leq \mathbb{V}_\theta(\hat{\theta}) \forall \theta$$

erfüllt. Insbesondere ist ein effizienter Schätzer eine Funktion der suffizienten Statistik.

Beispiel 5.14

Wenn die Exponentialverteilung mit $\theta = 1/\lambda$ parametrisiert wird, dann ergibt sich aus dem Satz von Cramér-Rao, dass der Maximum-Likelihood-Schätzer $\hat{\theta} = \bar{X}_n$ effizient ist (das wir in den Übungen gezeigt). Mit der ursprünglichen Parametrisierung ist der Maximum-Likelihood-Schätzer $\hat{\lambda} = 1/\bar{X}_n$ nicht erwartungstreu, er hat Erwartungswert $\lambda n/(n-1)$. Für $n > 1$ erhalten wir daraus den erwartungstreuen Schätzer

$$\hat{\lambda}_n = \frac{n-1}{\sum_{i=1}^{n} X_i}.$$

Dieser ist eine Funktion der suffizienten Statistik und erwartungstreu. Wir fragen uns jetzt, welche Funktionen $g(T)$ der suffizienten Statistik erwartungstreue Schätzer sind. Wir wissen, dass T eine Gammaverteilung $\Gamma(n,\lambda)$, also erhalten wir für die Erwartungstreue die Gleichung

$$\lambda = \mathbb{E}_\lambda(g(T)) = \int_0^\infty \frac{\lambda^n t^{n-1} g(t)}{(n-1)!} e^{-\lambda t} dt = \lambda,$$

oder

$$\int_0^\infty t^{n-1} g(t) e^{-\lambda t} dt = \frac{(n-1)!}{\lambda^{n-1}}$$

für alle $\lambda > 0$. Die linke Seite ist die Laplace-Transformierte der Funktion $f(t) = t^{n-1}g(t)$. Wegen des Eindeutigkeitssatzes für Laplace-Transformationen ist f und damit g eindeutig bestimmt. Das bedeutet, dass unser Schätzer der einzige erwartungstreue Schätzer, der eine Funktion der suffizienten Statistik ist, und daher der mit der kleinsten Varianz, also effizient.

5.3.3 Intervallschätzung

Die Theorie aus dem vorigen Abschnitt gibt uns einen einzelnen Schätzwert. Manchmal möchte man auch Angaben über die Genauigkeit eines Schätzwertes machen können, also ein Intervall bestimmen, in dem der gesuchte Parameter liegt. Leider kann so etwas nicht mit absoluter Sicherheit geschehen, weil in den meisten Fällen für jeden Wert des Parameters jede Stichprobe auftreten kann. In unserem einfachsten Modell, der Alternativverteilung $A(p)$, hat für jedes p (echt) zwischen 0 und 1, jede der 2^n möglichen 0-1-Folgen positive Wahrscheinlichkeit, als Stichprobe aufzutreten. Für keine dieser Stichproben können wir irgendein p außer den Extremen $p = 0$ oder $p = 1$ ausschließen.

Wir werden uns also damit begnügen müssen, ein Intervall zu bestimmen, das den gesuchten Parameter mit einer gewissen Wahrscheinlichkeit enthält.

Definition 5.9

$a = a(X_1, \ldots, X_n) \leq b = b(X_1, \ldots, X_n)$ seien zwei Statistiken. Das Intervall $[a, b]$ heißt Konfidenzintervall für θ mit Überdeckungswahrscheinlichkeit γ, wenn

$$\mathbb{P}_\theta(a \leq \theta \leq b) \geq \gamma.$$

Wenn in dieser Ungleichung Gleichheit gilt, sprechen wir von einem exakten Konfidenzintervall.

Wir beginnen diesmal mit der Normalverteilung. Diese hat zwei Parameter, μ und σ^2, für die wir versuchen können, Konfidenzintervalle zu finden. Wir beginnen mit μ. Ausgangspunkt ist der Schätzer für μ, zu dem wir auf mehreren Wegen gelangt sind, das Stichprobenmittel $\bar{X}_n$. Zuerst nehmen wir der Einfachheit halber an, dass wir die Varianz σ^2 kennen. Wir wollen unser Konfidenzintervall symmetrisch um den Schätzwert herumlegen, also setzen wir die Grenzen als $a = \bar{X}_n - c$ und $b = \bar{X}_n + c$ an. Wir wissen, dass $\bar{X}_n$ eine Normalverteilung mit Mittel μ und Varianz σ^2/n hat. Damit ergibt sich

$$\gamma = \mathbb{P}(a \leq \mu \leq b) = \mathbb{P}(\bar{X}_n - c \leq \mu \leq \bar{X}_n + c) = \mathbb{P}(-c \leq \bar{X}_n - \mu \leq c) =$$

$$\mathbb{P}(-\frac{c}{\sqrt{\sigma^2/n}} \leq \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \leq \frac{c}{\sqrt{\sigma^2/n}}) = \Phi(\frac{c}{\sqrt{\sigma^2/n}}) - \Phi(-\frac{c}{\sqrt{\sigma^2/n}}) = 2\Phi(\frac{c}{\sqrt{\sigma^2/n}}) - 1.$$

Wir haben also

$$\Phi(\frac{c}{\sqrt{\sigma^2/n}}) = \frac{1+\gamma}{2}$$

also

$$\frac{c}{\sqrt{\sigma^2/n}} = z_{\frac{1+\gamma}{2}},$$

$$c = z_{\frac{1+\gamma}{2}} \sqrt{\frac{\sigma^2}{n}},$$

und schließlich ist das Konfidenzintervall

$$[\bar{X}_n - z_{\frac{1+\gamma}{2}} \sqrt{\frac{\sigma^2}{n}}, , \bar{X}_n + z_{\frac{1+\gamma}{2}} \sqrt{\frac{\sigma^2}{n}}].$$

Diese Rechnung war nicht so schwer, aber die Annahme, dass σ^2 bekannt ist, ist natürlich recht unrealistisch. In realistischen Situationen ist σ^2 unbekannt und muss aus der Stichprobe geschätzt werden. Von dieser Erkenntnis aus gibt es zwei mögliche Wege: einerseits können wir einfach in

der Formel für das Konfidenzintervall die exakte Varianz σ^2 durch ihren Schätzer, die Stichprobenvarianz S_n^2 ersetzen und das Ergebnis als approximatives Konfidenzintervall behandeln (für großes n). Dieser Zugang funktioniert in vielen Situationen, wir werden ihn auch weiter unten für die Alternativverteilung verwenden. Die Normalverteilung schenkt uns den Luxus, dass wir auch exakt rechnen können. Wir beginnen mit dem Ansatz

$$a = \bar{X}_n - c\sqrt{\frac{S_n^2}{n}}, b = \bar{X}_n + c\sqrt{\frac{S_n^2}{n}}.$$

Wir formen die Gleichung für die Überdeckungswahrscheinlichkeit wieder um und erhalten

$$[\mathbb{P}(-c \leq \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}} \leq c) = \gamma. \tag{5.2}$$

An dieser Stelle hilft uns der folgende Satz, der sich mit dem Transformationssatz für Dichten beweisen lässt:

Satz 5.4

$X - 1, \ldots, X_n$ seien unabhängig $N(\mu, \sigma^2)$-verteilt. Dann sind $\bar{X}_n$ und S_n^2 unabhängig, $\bar{X}_n \sim N(\mu, \sigma^2/n)$ und $(n-1)S_n^2/\sigma^2 \sim \chi^2_{n-1}$ (Chi-Quadrat mit $n-1$ Freiheitsgraden).

$$T = \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}}$$

hat dann eine t-Verteilung mit $n-1$ Freiheitsgraden.

T ist genau der mittlere Term in (5.2). Mit der Symmetrie der t-Verteilung ergibt sich

$$c = t_{n-1;\frac{1+\gamma}{2}}$$

und das Konfidenzintervall

$$[\bar{X}_n - t_{n-1;\frac{1+\gamma}{2}}\sqrt{\frac{S_n^2}{n}},, \bar{X}_n + t_{n-1;\frac{1+\gamma}{2}}\sqrt{\frac{S_n^2}{n}}].$$

Für die Varianz gehen wir von der Stichprobenvarianz aus und setzen das Konfidenzintervall in der Form $[aS_n^2, bS_n^2]$ an. Die Gleichung für die Überdeckungswahrscheinlichkeit liefert

$$\gamma = \mathbb{P}(aS_n^2 \leq \sigma^2 \leq bS_n^2) = \mathbb{P}(\frac{n-1}{b} \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \frac{n-1}{a}).$$

Wir können jetzt unser Wissen über die Verteilung von S_n^2 ausnützen. Vorher müssen wir noch mit dem Problem fertig werden, dass wir nur eine Gleichung für die zwei Variablen a und b haben. Beim Mittelwert haben wir dieses Problem vermieden, indem wir das Intervall symmetrisch um den Schätzwert gelegt haben. Hier ist die Verteilung von S_n^2 leider nicht so symmetrisch wie die Normalverteilung des Stichprobenmittels, deshalb ist dieser Weg nicht so attraktiv. Eine bessere Idee ist es, die Fehlerwahrscheinlichkeit $1-\gamma$ symmetrisch aufzuteilen. Wir wollen also mit Wahrscheinlichkeit $\frac{1-\gamma}{2}$ mit σ^2 unter der linken Schranke zu liegen kommen, und mit der gleichen Wahrscheinlichkeit über der oberen, also

$$\mathbb{P}(\frac{(n-1)S_n^2}{\sigma^2} < \frac{n-1}{b}) = \frac{1-\gamma}{2}$$

und

$$\mathbb{P}(\frac{(n-1)S_n^2}{\sigma^2} > \frac{n-1}{a}) = \frac{1-\gamma}{2},$$

also

$$\mathbb{P}(\frac{(n-1)S_n^2}{\sigma^2} \leq \frac{n-1}{a}) = \frac{1+\gamma}{2}.$$

Wir wissen, dass

$$\frac{(n-1)S_n^2}{\sigma^2}$$

eine Chiquadratverteilung mit $n-1$ Freiheitsgraden hat, also können wir a und b durch deren Quantile ausdrücken:

$$\frac{n-1}{a} = \chi^2_{n-1;\frac{1+\gamma}{2}}$$

und

$$\frac{n-1}{b} = \chi^2_{n-1;\frac{1-\gamma}{2}}.$$

Damit wird unser Konfidenzintervall für σ^2:

$$[\frac{(n-1)S_n^2}{\chi^2_{n-1;\frac{1+\gamma}{2}}}, \frac{(n-1)S_n^2}{\chi^2_{n-1;\frac{1-\gamma}{2}}}].$$

Zusammengefasst erhalten wir

Satz 5.5

Konfidenzintervalle für die Normalverteilung $N(\mu, \sigma^2)$:
Für μ, wenn σ^2 bekannt ist:

$$[\bar{X}_n - z_{\frac{1+\gamma}{2}}\sqrt{\frac{\sigma^2}{n}},, \bar{X}_n + z_{\frac{1+\gamma}{2}}\sqrt{\frac{\sigma^2}{n}}],$$

Für μ, wenn σ^2 unbekannt ist:

$$[\bar{X}_n - t_{n-1;\frac{1+\gamma}{2}}\sqrt{\frac{S_n^2}{n}},, \bar{X}_n + t_{n-1;\frac{1+\gamma}{2}}\sqrt{\frac{S_n^2}{n}}],$$

für σ^2:

$$[\frac{(n-1)S_n^2}{\chi^2_{n-1;\frac{1+\gamma}{2}}}, \frac{(n-1)S_n^2}{\chi^2_{n-1;\frac{1-\gamma}{2}}}].$$

Das Beispiel der Normalverteilung gibt uns einen groben Rahmen zur Konstruktion von Konfidenzintervallen: ein möglicher Ausgangspunkt ist ein Schätzer für θ. Unter sehr günstigen Bedingungen (wie oben für die Normalverteilung) kann die Verteilung dieses Schätzers exakt bestimmt werden, in anderen Fällen (immer noch günstig) ist er zumindest asymptotisch normalverteilt. In diesem Fall können wir ein approximatives Konfidenzintervall in der Form

$$[\hat{\theta}_n - z_{\frac{1+\gamma}{2}}\sigma_n,, \hat{\theta}_n + z_{\frac{1+\gamma}{2}}\sigma_n],$$

wobei σ_n^2 die Varianz von $\hat{\theta}_n$ ist. Diese hängt in den meisten Fällen vom unbekannten θ ab, das wir durch $\hat{\theta}_n$ ersetzen.

Für Anteilswerte erhalten wir so das approximative Konfidenzintervall:

Satz 5.6

$(X_1, \ldots, X_n)$ sei eine Stichprobe einer Alternativverteilung $A(p), 0 < p < 1$.

$$\hat{p} = \bar{X}_n$$

sei der (erwartungstreue und effiziente) Schätzer für p. Dann ist ein approximatives Konfidenzintervall gegeben durch

$$[\hat{p} - z_{\frac{1+\gamma}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},, \hat{p} + z_{\frac{1+\gamma}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}].$$

Dieser Ansatz hat zwei Nachteile: einerseits kann der Fall eintreten, dass der Schätzwert $\hat{p}$ genau 0 oder 1 wird, dann hat das Konfidenzintervall Länge 0. Wenn p nicht genau 0 ist, aber trotzdem sehr klein, dann wird die Untergrenze des Konfidenzintervalls negativ, was auch nicht sehr befriedigend ist (und die analoge Situation gibt es in der Nähe von 1).

Am Ende des nächsten Kapitels werden wir einen Zusammenhang zwischen Tests und Konfidenzintervallen kennenlernen, der uns die folgende etwas esoterisch anmutende Regel liefert: wir lesen

$$p = \hat{p} \pm z_{\frac{1+\gamma}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

als Gleichung für p. Diese wird durch Routineumformungen zu einer quadratischen, und deren zwei Lösungen sind die Grenzen des Konfidenzintervalls.

Für und genügt aber die simple Version des Satzes.

Dazu ein schnelles Beispiel mit einer nicht ganz plausiblen Stichprobe:

Beispiel 5.15

1 2 3 4 5

sei eine Stichprobe einer Normalverteilung. Wir wollen Konfidenzintervalle für μ und σ^2 bestimmen. Wie immer, wenn die Überdeckungswahrscheinlichkeit nicht explizit genannt wird, verwenden wir den Standardwert $\gamma = 0.95$.

Das geht nach Kochrezept:

Wir sammeln zuerst die Information aus der Stichprobe, den Stichprobenumfang

$$n = 5,$$

das Stichprobenmittel

$$\bar{X} = 3$$

und die Stichprobenvarianz

$$S^2 = \frac{1}{5-1}(4+1+0+1+4) = 2.5;$$

außerdem brauchen wir die Quantile der t- und χ^2-Verteilungen mit 4 Freiheitsgraden aus den Tabellen in Anhang A:

$$t_{n-1;\frac{1+\gamma}{2}} = t_{4;0.975} = 2.776, \chi^2_{n-1;\frac{1+\gamma}{2}} = \chi^2_{4;0.975} = 11.143, \chi^2_{n-1;\frac{1-\gamma}{2}} = \chi^2_{4;0.025} = 0.484.$$

Das gibt als Konfidenzintervall für μ

$$[3 - 2.776\sqrt{\frac{2.5}{5}}, 3 + 2.776\sqrt{\frac{2.5}{5}}] = [1.04, 4.96]$$

und für σ^2

$$\frac{4 \cdot 2.5}{11.143}, \frac{4 \cdot 2.5}{0.484}] = [0.897, 20.66].$$

Auch zu den Anteilswerten:

Beispiel 5.16

In einer Packung mit 100 Gummibärchen waren 16 rote. Wir wollen ein 95%-Konfidenzintervall für den Anteil p der roten in der Produktion bestimmen.

Hier brauchen wir für unser Kochrezept nur $\hat{p} = 16/100 = 0.16$ und $z_{0.975} = 1.96$. Also ist

das Konfidenzintervall

$$0.16 - 1.96\sqrt{\frac{0.16 \cdot 0.84}{100}}, 0.16 + 1.96\sqrt{\frac{0.16 \cdot 0.84}{100}}] = [0.123, 0.197].$$

5.4 Tests

5.4.1 Grundlagen

Bei dieser Art von Problemen geht es nicht mehr darum, einen Näherungswert für einen unbekannten Parameter zu bestimmen, sondern es soll eine Aussage über den Parameter überprüft werden, etwa "der Ausschussanteil ist kleiner als 1%" oder "das mittlere Gewicht ist 1kg".

Beispiel 5.17

Als (nicht ganz ernstes) Beispiel wollen wir uns vorstellen, dass wir am Lieblingsspiel der Mathematiker, dem Münzwurfspiel, teilnehmen wollen. Um Sie ordentlich auszunehmen, lassen wir von einem befreundeten Münzschläger eine Münze prägen, die mit Wahrscheinlichkeit 3/4 "Kopf" zeigt und mit Wahrscheinlichkeit 1/4 "Zahl". Nun ist aber die Münzschlägerei eine Arbeit, die durch Anstrengung und Hitze zum Trinken verleitet, und auch unser Freund ist dagegen nicht gefeit, daher ist durch seinen getrübten Bewusstseinszustand nicht zu sagen, ob er den Rohling korrekt in in den Unterstempel eingelegt hat oder nicht, wir wissen also nicht, ob jetzt Kopf oder Zahl die bevorzugte Seite ist.

Als gestandene Statistiker*en wissen wir natürlich einen eleganten Ausweg: wir führen eine gewisse Anzahl von Münzwürfen aus, und entscheiden uns für die Seite, die öfter vorkommt. Damit wir bei fifty-fifty nicht in Entscheidungsnot kommen, wählen wir eine ungerade Anzahl von Würfen, sagen wir 51. Wie bei allen statistischen Verfahren gibt es auch hier eine gewisse Chance, einen Fehler zu begehen, und zwar machen wir genau dann einen Fehler, wenn die Seite, die mit Wahrscheinlichkeit 3/4 geworfen wird, höchstens 25 mal auftritt. Der Einfachheit halber verwenden wir die Approximation der Binomialverteilung durch die Normalverteilung (die allerdings hier nicht wirklich anwendbar ist) und erhalten

$$\mathbb{P}(X \leq 25) \approx \Phi\left(\frac{25.5 - 51 \cdot \frac{3}{4}}{\sqrt{51 \cdot \frac{3}{4} \cdot \frac{1}{4}}}\right) = 1.8 \cdot 10^{-5}.$$

(mit der exakten Binomialverteilung ergibt sich der mehr als dreimal so große Wert $5.9 \cdot 10^{-5}$, Das liegt daran, dass die Grenze $n/2$ zu weit vom Erwartungswert $3n/4$ entfernt ist, als dass der zentrale Grenzwertsatz noch gelten könnte; für solche extremen Fälle ist die "Theorie der großen Abweichungen" zuständig, die genauere Näherungen ermöglicht).

Etwas näher an der Realität ist das folgende Beispiel

Beispiel 5.18

Ein Flasche Bier soll 500 ml enthalten. Wir wollen das überprüfen und messen den Inhalt von 25 Flaschen mit folgendem Ergebnis:

509 522 507 476 508 492 480 523 474 506
513 510 517 476 479 516 496 531 464 497
482 495 492 495 483

Wir müssen hier einige Entscheidungen treffen: die erste ist die, was wir genau überprüfen wollen. Von unserem Standpunkt als Konsumenten her stört es uns eigentlich nicht (im Gegenteil), wenn wir mehr als das Soll bekommen, Abweichungen nach unten sind das, was uns wirklich stört.

Mindestens genau so wichtig ist die Frage nach dem Modell, das wir unseren Untersuchungen zugrunde legen wollen: wie immer, wenn es um Messwerte geht, denken wir zuerst

einmal an die Normalverteilung. Den typischen Wert 500 setzen wir als Mittelwert μ der Normalverteilung ein.

Unsere Frage formulieren wir also so: wir nehmen an, dass wir es hier mit einer Normalverteilung zu tun haben und wollen überprüfen, ob $\mu = 500$ gilt. Für diesen Mittelwert haben wir mit dem Stichprobenmittel einen guten Schätzer. Wir werden eine Grenze vorgeben und uns für "OK" entscheiden, wenn der Schätzer über dieser Grenze liegt, und für "schlecht", wenn er darunter ist. Wie diese Grenze gesetzt werden soll, müssen wir uns noch überlegen.

Wir definieren zuerst einige allgemeine Begriffe im Zusammenhang mit statistischen Tests:

Definition 5.10

Eine *Hypothese* ist eine Teilmenge des Parameterraums Θ.

Wir schreiben Hypothesen meistens nicht in Mengennotation, sondern als Aussage (meistens eine Gleichung oder Ungleichung) für den Parameter. Wir unterscheiden einseitige Hypothesen (von der Form $\theta \leq c$ oder $\theta > c$ etc.) und zweiseitige Hypothesen ($\theta \neq c$). Enthält die Hypothese nur einen Parameterwert ($\theta = c$), nennen wir sie einfach.

Ein Test wird als Entscheidung zwischen zwei Hypothesen formuliert, der Nullhypothese H_0 und der Gegenhypothese oder Alternative H_1. Die Rollen der beiden Hypothesen sind nicht symmetrisch — der übliche Sprachgebrauch ist "die Nullhypothese wird angenommen" oder "die Nullhypothese wird verworfen".

Beispiel 5.19

Wir formulieren die Hypothesen für die Fragestellungen der beiden letzten Beispiele.

Beim Münzwurfbeispiel herrscht so viel Symmetrie, dass wir es uns eigentlich aussuchen könnten, aber es ist üblich, dass das, was "unverändert" oder "in Ordnung" ist, als Nullhypothese formuliert wird, und die "Veränderung" oder das "Fehlerhafte" (generell das, was bewiesen werden soll) als Alternative. Mit dieser Philosophie würden wir also, wenn wir mit p die Wahrscheinlichkeit für "Kopf" bezeichnen, die einfachen Hypothesen $H_0 : p = 3/4$ gegen $H_1 : p = 1/4$ testen.

Im Normalverteilungsbeispiel wollen wir beweisen, dass wir betrogen worden sind, also dass wir zu wenig bekommen haben, und das formulieren wir als Alternative: $H_1 < 500$, und die Nullhypothese ist dann natürlich "500 oder besser": $H_0 : \mu \geq 500$.

Ein Test kann durch die Menge der möglichen Stichprobenwerte angegeben werden, bei denen die Nullhypothese angenommen wird, den Annahmebereich, bzw. durch sein Komplement, den Verwerfungsbereich. Oft ist es einfacher, eine Teststatistik anzugeben, und die Nullhypothese zu verwerfen, wenn diese Statistik einen kritischen Wert überschreitet (oder unterschreitet).

Beim Testen kann man zwei Arten von Fehlern begehen: Fehler erster Art — die Nullhypothese wird verworfen, obwohl sie zutrifft, und Fehler zweiter Art — die Nullhypothese wird angenommen, obwohl sie nicht zutrifft. Man möchte natürlich die Wahrscheinlichkeit für beide Fehler möglichst klein halten. Leider geht das nicht gleichzeitig — die Wahrscheinlichkeit für einen Fehler erster Art kann (zumindest ab einem gewissen Punkt) nur verkleinert werden, indem der Annahmebereich vergrößert wird, und dadurch wächst die Wahrscheinlichkeit für einen Fehler zweiter Art. In der Statistik wird dieses Dilemma gelöst, indem man eine Schranke für die Wahrscheinlichkeit eines Fehlers erster Art angibt:

Definition 5.11

Ein Test heißt vom Niveau α, wenn die Wahrscheinlichkeit für einen Fehler erster Art (die bei zusammengesetzten Hypothesen eine Funktion davon θ ist) nicht größer als α ist.

Wir können eine generelle Vorgangsweise aufstellen, um Tests durchzuführen:

1. Hypothesen formulieren: nach Möglichkeit wird das, was zu beweisen ist, als Alternative formuliert.

2. Auswahl der Teststatistik: das sollte eine Statistik sein, die unter der Nullhypothese deutlich andere Werte als unter der Alternative annimmt.
3. Festlegen des Signifikanzniveaus
4. Bestimmung der Verteilung der Teststatistik unter der Nullhypothese und damit des kritischen Wertes
5. Ziehen einer Stichprobe
6. Berechnung der Teststatistik aus der Stichprobe
7. Vergleich des Werts aus der Stichprobe mit dem kritischen Wert und Annehmen oder Verwerfen der Nullhypothese
8. Interpretation des Ergebnisses.

Es kann natürlich kleine Abweichungen von dieser Liste geben. Eine der deutlichsten ergibt sich bei der Verwendung von Statistiksoftware: dort gibt es keine Möglichkeit, das Niveau anzugeben. Stattdessen wird der sogenannte p-Wert berechnet.

Definition 5.12: p-Wert

T sei eine Teststatistik, T_{stp} sei der Wert von T, der aus einer konkreten Stichprobe berechnet wurde. Der p-Wert ist die Wahrscheinlichkeit, einen Wert für T zu erhalten, der mindestens so stark gegen die Nullhypothese spricht wie der aus der Stichprobe. Was das konkret bedeutet, hängt von der Semantik von T ab:

- Wenn T einseitig ist, und H_0 verworfen wird, wenn $T > t_c$ ist, dann ist der p-Wert $\mathbb{P}(T \geq T_{stp})$,
- Wenn T in der anderen Richtung einseitig ist, also H_0 verworfen wird, wenn $T < t_c$ ist, dann ist der p-Wert $\mathbb{P}(T \leq T_{stp})$,
- Wenn T zweiseitig ist, dann ist der p-Wert $2\min(\mathbb{P}(T \geq T_{stp}), \mathbb{P}(T \leq T_{stp}))$.

Die Nullhypothese wird auf Niveau α verworfen, wenn der p-Wert kleiner als α ist.

Die Auswahl der Teststatistik gibt uns einiges an Raum zur Kreativität. Wie schon gesagt wurde, ist die einzige Anforderung, dass ihre Verteilung durch den Parameter, der getestet wird, möglichst deutlich beeinflusst wird. Ein Schätzer für den Parameter ist im allgemeinen ein guter Ausgangspunkt.

Beispiel 5.20

Wir betrachten die Normalverteilung, konkret das Bier-Beispiel (Beispiel 5.18). Wir betrachten die allgemeine Frage, die Nullhypothese $H_0 : \mu \geq \mu_0$ gegen die Alternative $H_1 : \mu < \mu_0$ zu testen (im Beispiel ist $\mu_0 = 500$). Wir beginnen mit dem Schätzer für μ, dem Stichprobenmittel $\bar{X}_n$. Von diesem wissen wir, dass es eine Normalverteilung $N(\mu, \sigma^2/n)$ hat. Wir suchen dafür einen kritischen Wert c, und logischerweise werden wir H_0 verwerfen, wenn $\bar{X}_n < c$ gilt. Zur Bestimmung von c nehmen wir an, dass wir σ^2 kennen, und es ist auch plausibel, dass wir die größte Wahrscheinlichkeit für einen Fehler erster Art für $\mu = \mu_0$, den kleinsten Wert in der Nullhypothese erhalten. Wir rechnen also

$$\alpha = \mathbb{P}_{\mu_0}(\bar{X}_n \leq c) = \Phi(\frac{c - \mu_0}{\sqrt{\sigma^2/n}}),$$

daher

$$\frac{c - \mu_0}{\sqrt{\sigma^2/n}} = z_\alpha = -z_{1-\alpha},$$

und

$$c = \mu_0 - z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}}.$$

Für die weitere Diskussion ist es günstig, die letzte Umformung wieder rückgängig zu machen, und die neue Teststatistik

$$T = \frac{\bar{X}_n - \mu_0}{\sqrt{\sigma^2/n}}$$

zu verwenden. Diese hat eine Standardnormalverteilung, und der kritische Wert ist das α-Quantil $z_\alpha = -z_{1-\alpha}$.

Der Grund, warum wir das tun, liegt daran, dass wir auf den realistischeren Fall kommen wollen, dass wir die Varianz nicht kennen. Wie üblich schätzen wir diese unbekannte Varianz durch die Stichprobenvarianz, und dadurch kommen wir zu der Teststatistik

$$T = \frac{\bar{X}_n - \mu_0}{\sqrt{S_n^2/n}}.$$

Von dieser wissen wir nicht nur, dass Sie (unter sehr allgemeinen Bedingungen) für großes n näherungsweise standardnormalverteilt ist, sondern wir kennen sogar ihre exakte Verteilung, eine t-Verteilung mit $n-1$ Freiheitsgraden. Wir verwerfen also unsere Nullhypothese, wenn $T < t_{n-1;\alpha} = -t_{n-1,1-\alpha}$ gilt. Ähnlich lassen sich die umgekehrte einseitige und die zweiseitige Frage behandeln, die Ergebnisse werden weiter unten als "Kochrezepte" zusammengefasst. Alle diese Varianten sind als t-Test bekannt.

Zum Abschluss setzen wir die Zahlen aus unserem Beispiel ein: wir berechnen aus der Stichprobe

$$\bar{X}_n = 497.72, S_n^2 = 319.54,$$

und damit

$$T = \frac{497.72 - 500}{\sqrt{319.54/25}} = -0.6377.$$

Der kritische Wert ist $t_{24,0.05} = -1.711$. T ist nicht kleiner als dieser kritische Wert, also wird die Nullhypothese angenommen.

Zur Illustration soll hier auch noch gezeigt werden, wie dieser Test in einem Statistikpaket behandelt wird. Ich verwende gerne R (http://www.r-project.org). Da funktioniert das etwa so:

In der Variablen `bier` sind die 25 Stichprobenwerte als Vektor gespeichert. Für den t-Test ist — Überraschung — die Funktion `t.test()` zuständig; beim Aufruf muss man noch den hypothetischen Mittelwert und die Alternative angeben, etwa so

```
t.test(bier, alternative="less", mu=500)
```

mit der Ausgabe

```
        One Sample t-test

data:  bier
t = -0.63773, df = 24, p-value = 0.2648
alternative hypothesis: true mean is less than 500
95 percent confidence interval:
     -Inf 503.8367
sample estimates:
mean of x
   497.72

>
```

Der p-Wert 0.2648 sagt uns, dass wir die Nullhypothese auf keinem Signifikanzniveau verwerfen können, das darunter liegt, also auf keinem vernünftigen.

Eine Möglichkeit, einen Test zu konstruieren, liefert uns die Likelihood-Methode: die Grundidee besteht darin, sich für die Hypothese zu entscheiden, für die die aktuelle Stichprobe die größere Wahrscheinlichkeit hat. Damit man das Niveau α einstellen kann, wird noch ein zusätzlicher Faktor eingeführt:

Definition 5.13

Die Likelihoodquotientenstatistik ist

$$\ell = \frac{\sup_{\theta \in H_0} L(X_1, \ldots, X_n, \theta)}{\sup_{\theta \in H_1} L(X_1, \ldots, X_n, \theta)}$$

bzw.

$$\ell = \frac{\sup_{\theta \in H_0} L(X_1, \ldots, X_n, \theta)}{\sup_{\theta \in \Theta} L(X_1, \ldots, X_n, \theta)}$$

(die zweite Form ist oft einfacher zu berechnen, und für große Stichprobenumfänge sind die Tests identisch).

Der Likelihoodquotiententest verwirft die Nullhypothese, wenn ℓ kleiner ist als ein kritischer Wert.

In einem speziellen Fall ist der Likelihoodquotiententest optimal:

Satz 5.7: Neyman-Pearson

Falls sowohl H_0 als auch H_1 einfach ist,dann ist der Likelihoodquotiententest optimal, d.h., er hat unter allen Tests mit demselben Niveau die minimale Wahrscheinlichkeit für einen Fehler zweiter Art (in diesem Fall wird der Likelihoodquotient einfach als

$$L(X_1, \ldots X_n, \theta_0)/L(X_1, \ldots X_n, \theta_1)$$

berechnet)

Beispiel 5.21

Wieder einmal soll die Normalverteilung herhalten, und zwar testen wir die Nullhypothese $H_0 : \mu = \mu_0$ gegen die zweiseitige Alternative $H_1 : \mu \neq \mu_0$. Hier ist die Anmerkung fällig, dass es sich hier trotz der etwas irreführenden Form um eine zusammengesetzte Nullhypothese handelt, weil zwar μ festgelegt wird, aber σ^2 immer noch beliebige Werte annehmen kann.

Die Likelihoodfunktion für die Normalverteilung wurde schon bei der Besprechung des Maximum-Likelihood-Schätzers in Beispiel 5.7 berechnet. Die Maximierung, die wir dort durchgeführt haben, ist genau das, was wir für den Nenner der Likelihoodquotientenstatistik benötigen. Dieser wird

$$\left(\frac{n}{2\pi \sum_{i=1}^n (X_i - \bar{X}_n)^2} \right)^{n/2} e^{-n/2}.$$

Im Zähler ist $\mu = \mu_0$ fixiert, und wir müssen nur bezüglich σ maximieren, und das ergibt

$$\left(\frac{n}{2\pi \sum_{i=1}^n (X_i - \mu_0)^2} \right)^{n/2} e^{-n/2}.$$

Der Steinersche Verschiebungssatz liefert uns

$$\sum_{i=1}^{n}(X_i-\mu_0)^2=\sum_{i=1}^{n}(X_i-\bar{X}_n)^2+n(\bar{X}_n-\mu_0)^2.$$

Alles eingesetzt ergibt für die Likelihoodquotientenstatistik

$$\ell=\left(\frac{\sum_{i=1}^{n}(X_i-\bar{X}_n)^2}{\sum_{i=1}^{n}(X_i-\bar{X}_n)^2+n(\bar{X}_n-\mu_0)^2}\right)^{n/2}=\left(1+\frac{n-1}{n}T^2\right)^{-n/2},$$

wobei T die t-Teststatistik aus dem letzten Beispiel ist. Die Nullhypothese wird verworfen, wenn ℓ kleiner ist als ein kritischer Wert, und das ist genau dann der Fall, wenn T dem Betrag nach größer ist als ein passend gewählter Wert, und damit haben wir hier den t-Test erhalten, zu dem wir früher schon auf etwas heuristischeren Wegen gekommen sind.

5.4.2 Spezielle Tests

Für die Standardmodelle Alternativverteilung und Normalverteilung haben wir auch Standardtests, die wir durch die Likelihoodmethode und auch auf anderen Wegen finden können. Für die Alternativverteilung haben wir es uns einfach gemacht und die Näherung durch die Normalverteilung verwendet, was natürlich nur für hinreichend großen Stichprobenumfang erlaubt ist. Mit nicht allzugroßer Mühe, und speziell, wenn wir mit p-Werten arbeiten, könnten wir hier auch die exakte Binomialverteilung verwenden.

Für den Mittelwert einer Normalverteilung, wenn σ^2 unbekannt ist:

$$T=\frac{\bar{X}_n-\mu_0}{\sqrt{S_n^2/n}}$$

$H_0:\mu\le\mu_0$ gegen $H_1:\mu>\mu_0$: verwerfen, wenn $T>t_{n-1;1-\alpha}$.
$H_0:\mu\ge\mu_0$ gegen $H_1:\mu<\mu_0$: verwerfen, wenn $T<-t_{n-1;1-\alpha}$.
$H_0:\mu=\mu_0$ gegen $H_1:\mu\ne\mu_0$: verwerfen, wenn $|T|>t_{n-1;1-\alpha/2}$.
Für die Varianz einer Normalverteilung:

$$T=\frac{(n-1)S_n^2}{\sigma_0^2}$$

$H_0:\sigma^2\ge\sigma_0^2$ gegen $H_1:\sigma^2<\sigma_0^2$: verwerfen, wenn $T<\chi^2_{n-1;\alpha}$.
$H_0:\sigma^2=\sigma_0^2$ gegen $H_1:\sigma^2\ne\sigma_0^2$: verwerfen, wenn $T>\chi^2_{n-1;1-\alpha/2}$ oder $T<\chi^2_{n-1;\alpha/2}$.
$H_0:\sigma^2\le\sigma_0^2$ gegen $H_1:\sigma^2>\sigma_0^2$: verwerfen, wenn $T>\chi^2_{n-1;1-\alpha}$.
Für Anteilswerte:

$$T=\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}.$$

$H_0:p\le p_0$ gegen $H_1:p>p_0$: verwerfen, wenn $T>z_{1-\alpha}$.
$H_0:p\ge p_0$ gegen $H_1:p<p_0$: verwerfen, wenn $T<-z_{1-\alpha}$.
$H_0:p=p_0$ gegen $H_1:p\ne p_0$: verwerfen, wenn $|T|>z_{1-\alpha/2}$.

5.4.3 Der Chi-Quadrat-Anpassungstest

Wir machen jetzt einen kleinen Abstecher in die Welt der nichtparametrischen Statistik. Die Frage, die wir untersuchen, ist, ob eine Stichprobe aus einer gegebenen Verteilung stammt. Tests, die diese Hypothese überprüfen, werden *Anpassungstests* genannt.

Das geht am einfachsten, wenn es sich bei der hypothetischen Verteilung um eine diskrete Verteilung mit endlich vielen Werten $1,\dots,k$ handelt. Wir testen also

$$H_0: X\sim P=(p_1,\dots,p_k)$$

gegen die Alternative $X\not\sim P$. Unser ursprünglich nichtparametrisches Problem ist damit zu der parametrischen Frage nach den endlich vielen Wahrscheinlichkeiten $p_1,\dots,p_k$ geworden. Diese kann

man mit der Likelihoodquotientenmethode behandeln; der Test, den wir verwenden, kann man als Approximation des Likelihoodquotiententests erhalten.

Wir führen die Häufigkeiten

$$Y_i = \#\{j \leq n : X_j = i\}$$

ein. Für großes n ist $Y_i \approx np_i$ (und approximativ normalverteilt), wenn H_0 zutrifft. Wir wollen alle Differenzen zwischen Y_i und np_i gleichzeitig überprüfen. Dazu bilden wir eine gewichtete Quadratsumme:

$$T = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i}.$$

Diese Statistik ist asymptotisch χ^2-verteilt mit $k-1$ Freiheitsgraden. Die Nullhypothese wird abgelehnt, wenn

$$T > \chi^2_{k-1;1-\alpha}.$$

Da diese Aussagen nur asymptotisch gelten, muss n hinreichend groß sein. Die übliche Faustregel ist, dass np_i mindestens 5 sein soll.

Wenn diese Bedingung nicht erfüllt ist, oder wenn die Verteilung, die wir testen wollen, stetig ist, werden die möglichen Werte in Klassen eingeteilt. Bei stetigen Verteilungen kann man die Klassengrenzen so setzen, dass alle Klassen gleiche Wahrscheinlichkeit haben, was die Rechnung vereinfachen kann.

Wenn die Verteilung, auf die wir testen, nicht vollständig spezifiziert ist, etwa, wenn man testen will, ob eine Normalverteilung vorliegt, von der wir Mittel und Varianz nicht kennen, dann müssen die Parameter nach der Maximum-Likelihood Methode geschätzt werden. Mit diesen geschätzten Parametern können dann die Wahrscheinlichkeiten berechnet werden. Die Anzahl der Freiheitsgrade muss dann korrigiert werden, indem die Anzahl der geschätzten Parameter abgezogen wird, es sind also statt $k-1$ $k-1-d$ Freiheitsgrade, wobei d die Anzahl der geschätzten Parameter ist (im Normalverteilungsbeispiel ist $d = 2$).

Beispiel 5.22

Der Klassiker: ist der Würfel fair? Ein Würfel wird 600 mal geworfen, mit den folgenden Häufigkeiten für die einzelnen Augenzahlen:

1	2	3	4	5	6
90	85	80	120	115	110

Das ergibt die Chi-Quadrat-Statistik

$$\chi^2 = 1 + 2.25 + 4 + 4 + 2.25 + 1 = 15.$$

Das ist größer als $\chi^2_{5,0.95} = 11.0705$, und die Nullhypothese wird verworfen.

Beispiel 5.23

Auch klassisch: ist die Stichprobe normalverteilt?

Wir verwenden die Stichprobe aus dem Bierbeispiel (Beispiel 5.18). Diese hat Umfang $n = 25$. Weil jede Klasse eine erwartete Häufigkeit np_i von 5 haben soll, können wir höchstens 5 Klassen bilden, und wenn wir es tun, müssen wir die Wahrscheinlichkeiten für die einzelnen Klassen auf 1/5 einstellen. Wir schätzen μ und σ^2 durch Mittelwert und Stichprobenvarianz (wenn man der Theorie ganz genau folgt, sollte eigentlich der unkorrigierte ML-Schätzer für die Varianz, der mit Nenner n, verwendet werden) zur Berechnung dieser Wahrscheinlichkeiten. Damit wir für alle Klassen die gleichen Wahrscheinlichkeiten 1/5 erhalten, legen wird die Klassengrenzen an die Quantile x_p der Normalverteilung $N(\bar{X}_n, S_n^2) = N(497.72, 319.54)$ zu den Wahrscheinlichkeitswerten $0.2, 0.4, 0.6, 0.8$, also 482.68, 493.19, 502.25, 512.76. Die Häufigkeiten der einzelnen Klassen betragen 7, 3, 4, 5 und 6. Damit berechnen wir

$$\chi^2 = 0.8 + 0.8 + 0.1 + 0 + 0.1 = 1.8.$$

Die Anzahle der Freiheitsgrade ist $k-d-1 = 5-2-1 = 2$. Der kritische Wert zum üblichen Niveau $\alpha = 0.05$ ist $\chi^2_{2,0.95} = 5.99$. Der Chi-Quadrat-Wert aus der Stichprobe ist kleiner, wir nehmen also die Nullhypothese an.

5.4.4 Tests und Konfidenzintervalle

Es gibt einen interessanten Zusammenhang zwischen Tests und Konfidenzintervallen. Wenn wir etwa den t-Test für die zweiseitige Alternative und das Konfidenzintervall für μ vergleichen, dann wird klar, dass wir $H_0 : \mu = \mu_0$ genau dann auf Niveau α verwerfen, wenn μ im Konfidenzintervall mit Überdeckungswahrscheinlichkeit $\gamma = 1 - \alpha$ liegt. Das ist im allgemeinen so:

Satz 5.8

Wenn $I = [a(X_1, \ldots, X_n), b(X_1, \ldots, X_n)]$ ein Konfidenzintervall mit Überdeckungswahrscheinlichkeit $\gamma = 1 - \alpha$ ist, dann ist für jedes $\theta_0 \in \Theta$ durch die Regel "verwerfe, wenn $\theta_0 \notin I$" ein Test mit Niveau α für $H_0 : \theta = \theta_0$ gegeben.

Ist umgekehrt für jedes θ_0 ein Test mit Niveau α für die Nullhypothese $H_0 : \theta = \theta_0$ gegeben, dann ist die Menge I aller θ_0, für die $H_0 : \theta = \theta_0$ nicht verworfen wird, ein Konfidenzbereich mit Überdeckungswahrscheinlichkeit $\gamma = 1 - \alpha$.

Der Beweis ist nur ein Einsetzen in die Definitionen. Wenn I ein Konfidenzintervall mit Überdeckungswahrscheinlichkeit $1 - \alpha$ ist, dann ist für alle θ

$$\mathbb{P}_\theta(\theta \in I) \geq 1 - \alpha,$$

und daher ist tatsächlich für jedes θ_0 durch die Vorschrift "verwerfen, wenn $\theta_0 \notin I$" tatsächlich ein Test mit Niveau α gegeben.

Umgekehrt sei ϕ_{θ_0} der Test für die Nullhypothese $\theta = \theta_0$. Dann ist

$$I = \{\theta : \phi_\theta \text{ akzeptiert}\}$$

und

$$\mathbb{P}_\theta(\theta \in I) = \mathbb{P}_\theta(\phi_\theta \text{ akzeptiert}) \geq 1 - \alpha,$$

und I ein Konfidenzbereich mit Überdeckungswahrscheinlichkeit $1 - \alpha$.

Es ist allerdings nicht garantiert, dass der Konfidenzbereich, den wir hier erhalten, tatsächlich ein Intervall ist, man kann hier mit ausreichend viel bösem Willen allerhand an Grauslichkeiten herzaubern. Ein Beispiel, bei dem es funktioniert, geben uns die Anteilswerte. Wir haben den (näherungsweisen) Test für $H_0 : p = p_0$, der akzeptiert, wenn

$$p_0 - z_{1-\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}} \leq \hat{p} \leq p_0 + z_{1-\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}.$$

Die Menge aller p_0, für die diese Ungleichung gilt, wird begrenzt von den Werten, für die auf einer der beiden Seiten Gleichheit herrscht. Das führt genau auf die Gleichung, die nach Satz 5.6 erwähnt wurde.

5.5 Ergänzungen

5.5.1 Stichprobenvarianz

Dass es für die Varianz zwei Formeln gibt, eine mit Nenner n und eine mit Nenner $n-1$, also

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \text{ vs. } \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

kann etwas verwirrend erscheinen. Der Nenner $n-1$ ist für Stichproben richtig: da ist der Mittelwert, der abgezogen wird, nur ein Schätzwert, und er liegt näher an der Stichprobe (die Quadratsumme

kann als Abstand interpretiert werden) als der Erwartungswert. Der Nenner n ist korrekt, wenn der Mittelwert, der abgezogen wird, mit dem exakten Erwartungswert übereinstimmt. Das ist der Fall, wenn wir in der Wahrscheinlichkeitstheorie jeden der n Werte $x_1, \ldots, x_n$ mit gleicher Wahrscheinlichkeit ziehen, oder wenn wir in der Statistik eine Vollerhebung machen (also in einer Grundgesamtheit vom Umfang n alle Daten erheben und nicht nur eine Stichprobe). Die letzte Ansicht gibt uns auch einen Weg, wie wir die beiden Formel zusammenbringen können: wenn wir nämlich eine Stichprobe vom Umfang n aus einer Grundgesamtheit vom Umfang N *ohne* Zurücklegen ziehen, dann ist der Erwartungstreue Schätzer für die Varianz

$$\frac{1}{n-1}\frac{N-1}{N}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Das liegt immer zwischen den beiden Extremen, und für $N \to \infty$ nähert sich das Ziehen ohne Zurücklegen an das mit Zurücklegen an, und am anderen Ende $N = n$ steht die Vollerhebung.

5.5.2 Gegenbeispiele für Schätzer

Wir beginnen mit einem Beispiel, in dem es keinen effizienten Schätzer gibt, weil die kleinste Varianz für unterschiedliche Werte des Parameters von unterschiedlichen Schätzern erreicht wird und kein Schätzer für alle θ die kleinste Varianz liefern kann:

Beispiel 5.24

$(X_1, \ldots, X_n)$ sei eine Stichprobe einer Gleichverteilung aus $[\theta, \theta + 1]$. Wir setzen

$$\hat{\theta}_n = \lfloor X_1 \rfloor.$$

Dieser Schätzer ist erwartungstreu: wir setzen $\theta = k + a$ mit $k \in \mathbb{Z}$ und $0 \leq a < 1$. X_1 nimmt Werte zwischen θ und $\theta+1$; wenn X_1 zwischen $k+a$ und $k+1$ liegt, dann hat $\hat{\theta}_n = k$, wenn X_1 zwischen $k+1$ und $k+a+1$ liegt, gilt $\hat{\theta}_n = k+1$. Der erste Fall tritt mit Wahrscheinlichkeit $1 - a$ ein, der zweite mit Wahrscheinlichkeit a. Das ergibt

$$\mathbb{E}_\theta(\hat{\theta}_n) = k(1-a) + (k+1)a = k + a = \theta,$$

und $\hat{\theta}_n$ ist tatsächlich erwartungstreu. Wenn θ ganzzahlig ist, dann liegt X_1 zwischen θ und $\theta + 1$. Daher ist für ganzzahliges θ mit Wahrscheinlichkeit 1 $\hat{\theta}_n = \theta$ und daher $\mathbb{V}_\theta(\hat{\theta}_n) = 0$.

Der etwas kompliziertere Schätzer

$$\hat{\theta}_n^{(a)} = \lfloor X_1 - a \rfloor + a$$

mit $0 < a < 1$ ist, wie man ganz analog nachrechnet, ebenfalls erwartungstreu, und dieser Schätzer hat Varianz 0 wenn $\theta = k + a$ mit ganzzahligem k. Wenn es jetzt einen effizienten Schätzer gäbe, müsste dieser für jedes θ Varianz 0 haben, also mit Wahrscheinlichkeit 1 den Wert θ liefert. Das ist aber offensichtlich unmöglich, also existiert hier kein effizienter Schätzer.

Im nächsten Beispiel existiert kein erwartungstreuer Schätzer. Es macht zwar einen recht konstruierten Eindruck, ist aber nicht ganz so abwegig, wie es auf den ersten Blick scheinen mag (es ist eng mit einem statistischen Vorhersageverfahren namens “logistische Rekursion” verbunden).

Beispiel 5.25

Wir führen in der Alternativverteilung $A(p)$ den neuen Parameter

$$\theta = \log(\frac{p}{1-p})$$

ein. Das lässt sich umkehren:

$$p = p(\theta) = \frac{e^\theta}{e^\theta + 1}.$$

Das ist ein Weg, wie man den ursprünglichen Parameterraum $0 < p < 1$ durch ganz $\mathbb{R}$ ersetzen kann. Im allgemeinen kann man dazu $p = F(\theta)$ mit einer stetigen und streng monotonen Verteilungsfunktion F setzen. Die spezielle Form hier hat die schöne Eigenschaft, dass die Likelihoodfunktion die Form

$$L = (1 + e^{\theta})^{-n} e^{\theta \sum_i X_i}$$

hat, bei der im Exponenten θ als Faktor bei der suffizienten Statistik $T = \sum_i X_i$ steht. In einem solchen Fall nennt man θ den "natürlichen Parameter" des Modells.

Die Elemente $X_1, \ldots, X_n$ können jeweils nur die Werte 0 und 1 annehmen. Das gibt 2^n unterschiedliche Stichproben

$$x = (x_1, \ldots, x_n) \in \{0, 1\}^n.$$

Für jedes solche x können wir den Wert $\hat{\theta}(x)$ angeben, den der Schätzer $\hat{\theta}_n$ annehmen soll, wenn $(X_1, \ldots, X_n) = x = (x_1, \ldots, x_n)$ gilt. Das passiert mit der Wahrscheinlichkeit

$$\mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n) = p(\theta)^{\sum_{i=1}^n x_i} (1 - p(\theta))^{n - \sum_{i=1}^n x_i}.$$

Daraus ergibt sich

$$\mathbb{E}_\theta(\hat{\theta}_n) = \sum_{x \in \{0,1\}^n} \hat{\theta}(x) p(\theta)^{\sum_{i=1}^n x_i} (1 - p(\theta))^{n - \sum_{i=1}^n x_i}.$$

Das ist ein Polynom in p; damit $\hat{\theta}_n$ erwartungstreu ist, sollte es aber für alle p mit der Funktion $\log(\frac{p}{1-p})$ übereinstimmen, die aber nicht als Polynom darstellbar ist. Es kann also hier keinen erwartungstreuen Schätzer geben.

Beispiel 5.26

Wir betrachten ein Modell, in dem zwei Normalverteilungen gemischt werden: eine mit Mittel θ und Varianz 1, und die andere mit Mittel θ und Varianz $\sigma_\theta = e^{-1/\theta^2}$. Wir haben die Dichten

$$f_\theta(x) = \frac{\theta^2}{1 + \theta^2} \frac{1}{\sqrt{2\pi\sigma_\theta^2}} e^{-\frac{(x-\theta)^2}{2\sigma_\theta^2}} + \frac{1}{1 + \theta^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Für $\theta = 0$ soll der erste Summand 0 sein.

Nehmen wir an, dass wir eine Stichprobe für $\theta = 1$ vor uns haben. In der Umgebung von 0 gilt, sehr grob abgeschätzt, $f_1(x) \geq 0.1$. Für hinreichend großes n ist die Wahrscheinlichkeit, dass X_i zwischen $1/n$ und $2/n$ liegt, größer als $0.1/n$. Die Wahrscheinlichkeit, dass keines der X_i mit $i = 1, \ldots, n$ zwischen $1/n$ und $2/n$ liegt ist kleiner als $(1 - 0.1/n)^n \leq e^{-0.1} < 1$. Mit positiver Wahrscheinlichkeit gibt es also einen Wert in der Stichprobe, der zwischen $1/n$ und $2/n$ liegt. Wenn wir dieses Element der Stichprobe für θ einsetzen, gibt der entsprechende Faktor in der Likelihoodfunktion eine Größenordnung $e^{n^2/8}$, und die übrigen Faktoren können das schlimmstenfalls um e^{-cn} verringern. In der Umgebung von 1 ist die Likelihoodfunktion jedenfalls kleiner als 1, und kann mit dem quadratischen Wachstum im Exponenten nicht mithalten. Mit positiver Wahrscheinlichkeit hat daher der Maximum-Likelihood-Schätzer einen nach unten beschränkten Abstand vom tatsächlichen Parameterwert, und ist daher nicht einmal schwach konsistent.

5.6 Wiederholungsfragen

1. Woraus besteht ein statistisches Modell?
2. Was ist der Unterschied zwischen einem parametrischen und einem nichtparametrischen Modell?
3. Geben Sie Beispiele für statistische Modelle an.

4. Was ist eine Statistik?
5. Was ist ein Schätzer?
6. Welche Eigenschaften können wir von einem Schätzer verlangen?
7. Wie sind Stichprobenmittel und Stichprobenvarianz definiert? Warum wird bei der Stichprobenvarianz durch $n-1$ dividiert und nicht durch n?
8. Wie ist die Likelihoodfunktion definiert?
9. Welche Methoden zur Konstruktion von Schätzern kennen Sie?
10. Wie klein kann die Varianz eines erwartungstreuen Schätzers werden?
11. Wie ist ein Konfidenzintervall definiert?
12. Geben Sie ein Konfidenzintervall für den Mittelwert μ einer Normalverteilung an.
13. Geben Sie ein Konfidenzintervall für die Varianz σ^2 einer Normalverteilung an.
14. Geben Sie ein approximatives Konfidenzintervall für die Erfolgswahrscheinlichkeit p einer Alternativverteilung an.
15. Was ist ein statistischer Test?
16. Was ist eine Hypothese?
17. Welche Typen von Hypothesen gibt es?
18. Welche Fehler können beim Testen auftreten?
19. Was ist das Niveau (Signifikanzniveau) eines Tests?
20. Was ist der p-Wert und wie wird er verwendet?
21. Was versteht man unter einem Anpassungstest?
22. Wie wird der Chi-Quadrat-Anpassungstest durchgeführt?
23. Wie hängen Tests und Konfidenzintervalle zusammen?

Anhang A

Tabellen

Die Verteilungsfunktion der Standardnormalverteilung:

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

	0	1	2	3	4	5	6	7	8	9
0.0	.500	.504	.508	.512	.516	.520	.524	.528	.532	.536
0.1	.540	.544	.548	.552	.556	.560	.564	.567	.571	.575
0.2	.579	.583	.587	.591	.595	.599	.603	.606	.610	.614
0.3	.618	.622	.626	.629	.633	.637	.641	.644	.648	.652
0.4	.655	.659	.663	.666	.670	.674	.677	.681	.684	.688
0.5	.691	.695	.698	.702	.705	.709	.712	.716	.719	.722
0.6	.726	.729	.732	.736	.739	.742	.745	.749	.752	.755
0.7	.758	.761	.764	.767	.770	.773	.776	.779	.782	.785
0.8	.788	.791	.794	.797	.800	.802	.805	.808	.811	.813
0.9	.816	.819	.821	.824	.826	.829	.831	.834	.836	.839
1.0	.841	.844	.846	.848	.851	.853	.855	.858	.860	.862
1.1	.864	.867	.869	.871	.873	.875	.877	.879	.881	.883
1.2	.885	.887	.889	.891	.893	.894	.896	.898	.900	.901
1.3	.903	.905	.907	.908	.910	.911	.913	.915	.916	.918
1.4	.919	.921	.922	.924	.925	.926	.928	.929	.931	.932
1.5	.933	.934	.936	.937	.938	.939	.941	.942	.943	.944
1.6	.945	.946	.947	.948	.949	.951	.952	.953	.954	.954
1.7	.955	.956	.957	.958	.959	.960	.961	.962	.962	.963
1.8	.964	.965	.966	.966	.967	.968	.969	.969	.970	.971
1.9	.971	.972	.973	.973	.974	.974	.975	.976	.976	.977
2.0	.977	.978	.978	.979	.979	.980	.980	.981	.981	.982
2.1	.982	.983	.983	.983	.984	.984	.985	.985	.985	.986
2.2	.986	.986	.987	.987	.987	.988	.988	.988	.989	.989
2.3	.989	.990	.990	.990	.990	.991	.991	.991	.991	.992
2.4	.992	.992	.992	.992	.993	.993	.993	.993	.993	.994
2.5	.994	.994	.994	.994	.994	.995	.995	.995	.995	.995
2.6	.995	.995	.996	.996	.996	.996	.996	.996	.996	.996
2.7	.997	.997	.997	.997	.997	.997	.997	.997	.997	.997
2.8	.997	.998	.998	.998	.998	.998	.998	.998	.998	.998
2.9	.998	.998	.998	.998	.998	.998	.998	.999	.999	.999

Quantile z_p der Standardnormalverteilung:

p	z_p	p	z_p	p	z_p
.51	.025	.71	.553	.91	1.341
.52	.050	.72	.583	.92	1.405
.53	.075	.73	.613	.93	1.476
.54	.100	.74	.643	.94	1.555
.55	.126	.75	.674	.95	1.645
.56	.151	.76	.706	.96	1.751
.57	.176	.77	.739	.97	1.881
.58	.202	.78	.772	.975	1.960
.59	.228	.79	.806	.98	2.054
.60	.253	.80	.842	.99	2.326
.61	.279	.81	.878	.991	2.366
.62	.305	.82	.915	.992	2.409
.63	.332	.83	.954	.993	2.457
.64	.358	.84	.994	.994	2.512
.65	.385	.85	1.036	.995	2.576
.66	.412	.86	1.080	.996	2.652
.67	.440	.87	1.126	.997	2.748
.68	.468	.88	1.175	.998	2.878
.69	.496	.89	1.227	.999	3.090
.70	.524	.90	1.282	.9999	3.719

Quantile $t_{n;p}$ der t-Verteilung mit n Freiheitsgraden:

n	.9	.95	.975	.99	.995	n	.9	.95	.975	.99	.995
1	3.078	6.314	12.706	31.821	63.675	26	1.316	1.706	2.056	2.479	2.779
2	1.886	2.920	4.303	6.965	9.725	27	1.314	1.703	2.052	2.473	2.467
3	1.638	2.353	3.183	4.541	5.841	28	1.313	1.701	2.048	2.467	2.763
4	1.533	2.132	2.776	3.747	4.604	29	1.311	1.699	2.045	2.462	2.756
5	1.476	2.015	2.571	3.365	4.032	30	1.310	1.697	2.042	2.457	2.750
6	1.440	1.943	2.447	3.143	3.707	31	1.309	1.696	2.040	2.453	2.744
7	1.415	1.895	2.365	2.998	3.499	32	1.309	1.694	2.037	2.449	2.738
8	1.397	1.860	2.306	2.896	3.355	33	1.308	1.692	2.035	2.445	2.733
9	1.383	1.833	2.262	2.821	3.250	34	1.307	1.691	2.032	2.441	2.728
10	1.372	1.812	2.228	2.764	3.169	35	1.306	1.690	2.030	2.438	2.724
11	1.363	1.796	2.201	2.718	3.106	40	1.303	1.684	2.021	2.423	2.704
12	1.356	1.782	2.179	2.681	3.055	45	1.301	1.679	2.014	2.412	2.690
13	1.350	1.771	2.160	2.650	3.012	50	1.299	1.676	2.009	2.403	2.678
14	1.345	1.761	2.145	2.624	2.977	55	1.297	1.673	2.004	2.396	2.668
15	1.341	1.753	2.131	2.602	2.947	60	1.296	1.671	2.000	2.390	2.660
16	1.337	1.746	2.120	2.583	2.921	65	1.295	1.669	1.997	2.385	2.654
17	1.333	1.740	2.110	2.567	2.898	70	1.294	1.667	1.994	2.381	2.648
18	1.330	1.734	2.101	2.552	2.878	75	1.293	1.665	1.992	2.377	2.643
19	1.328	1.729	2.093	2.539	2.861	80	1.292	1.664	1.990	2.374	2.639
20	1.325	1.725	2.086	2.528	2.845	85	1.292	1.663	1.988	2.371	2.635
21	1.323	1.721	2.080	2.518	2.831	90	1.291	1.662	1.987	2.368	2.632
22	1.321	1.717	2.074	2.508	2.819	95	1.291	1.661	1.985	2.366	2.629
23	1.319	1.714	2.069	2.500	2.807	100	1.290	1.660	1.984	2.364	2.626
24	1.318	1.711	2.064	2.492	2.797	105	1.290	1.659	1.983	2.362	2.623
25	1.316	1.708	2.060	2.485	2.787	∞	1.282	1.645	1.960	2.326	2.576

Quantile $\chi^2_{n;p}$ der χ^2-Verteilung mit n Freiheitsgraden:

n	.005	.01	.02	.025	.05	.1	.5	.9	.95	.975	.98	.99	.995
1	.000	.000	.001	.001	.004	.016	.455	2.706	3.841	5.024	5.412	6.635	7.879
2	.010	.020	.040	.051	.103	.211	1.386	4.605	5.991	7.378	7.824	9.210	10.597
3	.072	.115	.185	.216	.352	.584	2.366	6.251	7.815	9.348	9.837	11.345	12.838
4	.207	.297	.429	.484	.711	1.064	3.357	7.779	9.488	11.143	11.668	13.277	14.860
5	.412	.554	.752	.831	1.145	1.610	4.351	9.236	11.070	12.832	13.308	15.086	16.750
6	.676	.872	1.134	1.237	1.635	2.204	5.348	10.645	12.592	14.449	15.033	16.812	18.548
7	.989	1.239	1.564	1.690	2.167	2.833	6.346	12.017	14.067	16.013	16.622	18.475	20.278
8	1.344	1.646	2.032	2.180	2.733	3.490	7.344	13.362	15.507	17.535	18.168	20.090	21.955
9	1.735	2.088	2.532	2.700	3.325	4.168	8.343	14.684	16.919	19.023	19.679	21.666	23.589
10	2.156	2.558	3.059	3.247	3.940	4.865	9.342	15.987	18.307	20.483	21.161	23.209	25.188
11	2.603	3.053	3.609	3.816	4.575	5.578	10.341	17.275	19.675	21.920	22.618	24.725	26.757
12	3.074	3.571	4.178	4.404	5.226	6.304	11.340	18.549	21.026	23.336	24.054	26.217	28.300
13	3.565	4.107	4.765	5.009	5.892	7.042	12.340	19.812	22.362	24.736	25.472	27.688	29.819
14	4.075	4.660	5.368	5.629	6.571	7.790	13.339	21.064	23.685	26.119	26.873	29.141	31.319
15	4.601	5.229	5.985	6.262	7.261	8.547	14.339	22.307	24.996	27.488	28.259	30.578	32.801
16	5.142	5.812	6.614	6.908	7.962	9.312	15.338	23.542	26.269	28.845	29.633	32.000	34.267
17	5.697	6.408	7.255	7.564	8.672	10.085	16.338	24.769	27.587	30.191	30.995	33.409	35.718
18	6.265	7.015	7.906	8.231	9,390	10.835	17.338	25.909	28.869	31.526	32.346	34.805	37.156
19	6.844	7.633	8.567	8.907	10.117	11.651	18.338	27.204	30.144	32.852	33.687	36.191	38.582
20	7.434	8.260	9.237	9.591	10.851	12.443	19.337	28.412	31.410	34.170	35.020	37.566	39.997
21	8.034	8.897	9.915	10.283	11.591	13.240	20.337	29.615	32.671	35.479	36.343	38.932	41.401
22	8.643	9.542	10.600	10.982	12.338	14.041	21.337	30.813	33.924	36.781	37.659	40.289	42.796
23	9.260	10.196	11.293	11.689	13.091	14.848	22.337	32.007	35.172	38.076	38.968	41.638	44.181
24	9.886	10.856	11.992	12.401	13.848	15.659	23.337	33.196	36.415	39.364	40.270	42.980	45.559
25	10.520	11.524	12.697	13.120	14.611	16.473	24.337	34.382	37.652	40.646	41.566	44.324	46.928
26	11.160	12.198	13.409	13.844	15.379	17.292	25.336	35.563	38.885	41.923	42.856	45.642	48.290
27	11.808	12.879	14.125	14.573	16.151	18.114	26.336	36.741	40.113	43.194	44.140	46.963	49.645
28	12.461	13.565	14.847	15.308	16.928	18.939	27.336	37.916	41.337	44.461	45.419	48.278	50.993
29	13.121	14.256	15.574	16.047	17.708	19.768	28.336	39.087	42.557	45.722	46.693	49.588	52.336
30	13.787	14.953	16.306	16.791	18.493	20.599	29.336	40.256	43.773	46.979	47.962	50.892	53.672
31	14.458	15.655	17.042	17.539	19.281	21.434	30.336	41.422	44.985	48.232	49.226	52.191	55.003
32	15.134	16.362	17.783	18.291	20.072	22.271	31.336	42.585	46.194	49.480	50.487	53.486	56.328
33	15.815	17.074	18.527	19.047	20.867	23.110	32.336	43.745	47.400	50.725	51.743	54.776	57.648
34	16.501	17.789	19.275	19.806	21.664	23.952	33.336	44.903	48.602	51.966	52.995	56.061	58.964
35	17.192	18.509	20.027	20.569	22.465	24.797	34.336	46.059	49.802	53.203	54.244	57.342	60.275
40	20.707	22.164	23.838	24.433	26.509	29.051	39.335	51.805	55.758	59.342	60.436	63.691	66.766
45	24.311	25.901	27.720	28.366	30.612	33.350	44.335	57.505	61.656	65.410	66.555	69.957	73.166
50	27.991	29.707	31.664	32.357	34.764	37.689	49.335	63.167	67.505	71.420	72.613	76.154	79.490
55	31.735	33.570	35.659	36.398	38.958	42.060	54.335	68.796	73.311	77.380	78.619	82.292	85.749
60	35.534	37.485	39.699	40.482	43.188	46.459	59.335	74.397	79.082	83.298	84.580	88.397	91.952
65	39.383	41.444	43.779	44.603	47.450	50.883	64.335	79.973	84.821	89.177	90.501	94.422	98.105
70	43.275	45.442	47.893	48.758	51.739	55.329	69.334	85.527	90.531	95.023	96.388	100.425	104.215
75	47.206	49.475	52.039	52.942	56.054	59.795	74.334	91.061	96.217	100.839	102.243	106.393	110.286
80	51.172	53.540	56.213	57.153	60.391	64.278	79.334	96.578	101.879	106.629	108.069	112.329	116.321
85	55.170	57.634	60.412	61.389	64.749	68.777	84.334	102.079	107.522	112.393	113.871	118.236	122.325
90	59.196	61.754	64.635	65.647	69.126	73.291	89.334	107.565	113.145	118.136	119.649	124.116	128.299
95	63.250	65.898	68.879	69.925	73.520	77.818	94.334	113.038	118.752	123.858	125.405	129.973	134.247
100	67.328	70.065	73.142	74.222	77.929	82.358	99.334	118.498	124.342	129.561	131.142	135.806	140.169

Anhang B

Mathematische Hintergründe

B.1 Wahrscheinlichkeitsräume

Wir haben uns bei der Definitions eines Wahrscheinlichkeitsraums nicht im Detail mit dem Definitionsbereich unserer Wahrscheinlichkeitsmaße beschäftigt bzw. stillschweigend angenommen, dass wir jeder Teilmenge von Ω eine Wahrscheinlichkeit zuordnen können. Darauf kann man sich allerdings nicht verlassen: es gibt etwa den folgenden

Satz B.1: Ulam

Es gibt kein Wahrscheinlichkeitsmaß $\mathbb{P}$, dass für *alle* Teilmengen von $[0,1]$ und die Bedingungen

$$\mathbb{P}([0,1]) = 1$$

und

$$\mathbb{P}(\{x\}) = 0$$

für alle $x \in [0,1]$ erfüllt.

Dieser Satz braucht allerdings eine zusätzliche Voraussetzung aus der Mengentheorie, die Kontinuumshypothese (oder eine Abschwächung davon).

In der Mathematik wird diese Frage so gelöst, dass eben nicht alle Teilmengen von Ω als "Ereignisse" zugelassen werden, sondern dass $\mathbb{P}$ auf einer Teilmenge $\mathfrak{S}$ der Potenzmenge 2^Ω definiert wird. Es ist angenehm, wenn diese Teilmenge ein wenig Struktur mitbringt: idealerweise sollte $\mathfrak{S}$ bezüglich der Bildung von Komplementen und abzählbaren Vereinigungen abgeschlossen sein. Ein solches System heißt Sigmaalgebra und ist auch gegenüber der Bildung von Differenzen zweier Mengen und von (endlichen oder abzählbaren) Durchschnitten abgeschlossen. Es ist allerdings nicht ganz wünschenswert, den Wert des Wahrscheinlichkeitsmaßes für alle die vielen und komplizierten Mengen aus $\mathfrak{S}$ angeben zu müssen, besser wäre es, wenn diese Angabe nur für wenige einfache Mengen notwendig wäre. Im Spezialfall $\Omega = \mathbb{R}$ wären etwa die Intervalle der Form $(a,b]$ einfach genug, und ihre Wahrscheinlichkeit lässt sich mit einer Verteilungsfunktion F als

$$\mathbb{P}((a,b]) = F(b) - F(a)$$

festlegen. Es stellt sich heraus, dass damit für jede Menge in der kleinsten Sigmaalgebra $\mathfrak{B}$, die alle diese Intervalle enthält, die Wahrscheinlichkeit eindeutig festgelegt wird. $\mathfrak{B}$ heißt die Borelsche Sigmaalgebra bzw. die Sigmaalgebra der Borelmengen. Sie ist sehr reichhaltig, enthält etwa alle offenen und abgeschlossenen Mengen und noch sehr viel mehr.

Index