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 allzusehr 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 bei einer großen Anzahl von Versuchen bei gewissen Werten, den Wahrscheinlichkeiten, "einpendeln", ist als das "empirische Gesetzder großen Zahlen" bekannt. Dabei handelt es sich zwar nicht um einen mathematischen Satz, 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 M zusammen, die wir die Grundmenge nennen. In unserem Beispiel ist

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

Die Elemente von M 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. 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 M 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) *M sei eine beliebige (nichtleere) Menge. Eine Funktion $\mathbb{P}$, die $A \subseteq M$ 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. $\mathbb{P}(M) = 1$,
4. *wenn $A_n, n \in \mathbb{N}$ disjunkte Ereignisse sind (d.h., $A_i \cap A_j = \emptyset, i \neq j$), dann gilt*

$$\mathbb{P}(\bigcup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n).$$

Anmerkungen:

1. In dieser Definition haben wir keine genauen Angaben über den Definitionsbereich von $\mathbb{P}$ gemacht. Wir hoffen natürlich, dass wir für alle Teilmengen von M eine Wahrscheinlichkeit definieren können. Wenn M unendlich (überabzählbar) ist, gibt es damit allerdings Probleme. Wir werden hier darauf vertrauen, dass alle Mengen, mit denen wir es zu tun haben, brav genug sind, dass sie eine Wahrscheinlichkeit verdienen. In der Mathematik wird das Problem so gelöst, dass der Definitionsbereich des Wahrscheinlichkeitsmaßes auf eine Teilmenge der Potenzmenge von M eingeschränkt wird (siehe Anhang).
2. Ein simples Beispiel für einen Wahrscheinlichkeitsraum mit 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 $M = \mathbb{N}$ (und $\mathbb{P}(\{n\}) = 2^{-n}$).
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).

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 2.2 (Laplacescher Wahrscheinlichkeitsraum) *Wenn M endlich ist und*

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

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

Diese Definition ist nur für endliche Mengen M 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 Laplace'sche 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.

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

Man kann sich etwa vorstellen, dass ein Mann zu einem zufälligen Zeitpunkt zwischen 8 und 9 in seiner 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). 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. Wir können die beiden Ankunftszeiten in einem rechtwinkeligen Koordinatensystem eintragen:

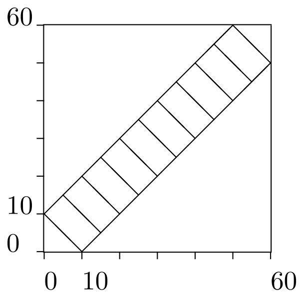

Abbildung 2.1: Geometrische Wahrscheinlichkeit

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

Solche "geometrischen Wahrscheinlichkeiten" sind in einfachen Fällen wie diesem eine brauchbare Veranschaulichung — die Berechnung der Wahrscheinlichkeit über die Fläche entspricht der Annahme, dass die beiden Koordinaten unabhängig sind. In komplizierteren Fällen ist nicht immer eindeutig klar, was unter einer "gleichmäßigen" Verteilung zu verstehen ist, und unterschiedliche Interpretationen geben unterschiedliche Resultate (Bertrand'sches 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 $M = A \cup A^C$.
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
$$S_i = \sum_{1 \leq j_1 < j_2 < \ldots < j_i \leq n} \mathbb{P}(A_{j_1} \cap \ldots \cap A_{j_i}).$$

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,

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

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 2.3 *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. Nach der Formel "günstige durch mögliche Fälle" ergibt sich unsere Formel.

Die Definition kann man ausmultiplizieren und erhält

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

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*

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

Ein schönes 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

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

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

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 2.4 (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 $1 \leq i < j \leq n$

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

In unserem Urnenbeispiel sind die Ereignisse A_i (“die i-te Kugel ist weiß”) nicht unabhängig. Um das zu verifizieren, müssen wir die unbedingten Wahrscheinlichkeiten finden. Das kann einerseits mit einem Symmetrieargument (auch beim zweiten, dritten,... Zug muss jede Kugel mit gleicher Wahrscheinlichkeit $1/5$ gezogen werden) geschehen, andererseits können wir etwa die Wahrscheinlichkeit von A_2 so erhalten:

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

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

Dieses Vorgehen 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 $\bigcup_i B_i = M$ und A ein beliebiges Ereignis. Dann gilt*

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

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)*

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

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 voneinader 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, mein Sohn von mir ein Gen o haben muss. Meine Ausstattung muss also entweder oa (wir bezeichnen dieses Ereignis mit A_a), ob (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)von $1/2$, 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 M und einem Wahrscheinlichkeitsmaß $\mathbb{P}$ darauf:

Definition 2.5 *Eine Zufallsvariable X ist eine Abbildung von M nach $\mathbb{R}^d$.*

Bemerkungen:

1. Wenn M 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.

Ein zentraler Begriff ist die Verteilung einer Zufallsvariable:

Definition 2.6 *Die Verteilung einer Zufallsvariable ist das Wahrscheinlichkeitsmaß*

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

Definition 2.7 *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. Die allgemeine Form (2.1) ist die erste Verteilung, für die wir einen Namen haben:

Definition 2.8 *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 2.9 *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 2.10 *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 2.11 *Für ein Ereignis A heißt die Funktion*

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

mit

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

Indikator von A.

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

Definition 2.12 *Die Verteilungsfunktion von X ist gegeben durch*

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

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

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

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. *$0 \le F(x_1, x_2) \le 1$ 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*

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

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

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

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 2.13 *Wenn F_X in der Form*

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

bzw.

$$F_X(x) = \int_{-\infty}^{x_d} \dots \int_{-\infty}^{x_1} f_X(u_1, \dots, x_d)dx_1 \dots dx_d$$

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

Beispiele für stetige Verteilungen:

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

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

  Diese Verteilung (mit $a = 0, b = 1$) ist uns schon als Beispiel für einen Wahrscheinlichkeitsraum mit unendlich vielen Elementen begegnet.

- Die Exponentialverteilung $Ex(\lambda)$:

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

- 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.

Die Dichte der Normalverteilung kann nicht geschlossen integriert werden. Es kann sein, dass eine Zufallsvariable sowohl diskrete als auch stetige Anteile hat:

Definition 2.14 *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.*

Anmerkung: Die Wahrscheinlichkeitsfunktion und Dichte einer gemischten Verteilung sind unvollständig: die Summe bzw. das Integral dieser Funktionen sind kleiner als 1, ihre Summe muss natürlich 1 ergeben.

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.

Mithilfe der Verteilungsfunktion kann man Wahrscheinlichkeiten berechnen:

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

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

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

Mithilfe der Wahrscheinlichkeits- bzw. Dichtefunktion können die Wahrscheinlichkeiten 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 ergibt sich die Dichte von X bzw. 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 Verteilug 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.

Definition 2.15 *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 2.16 *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

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

Satz 2.8 (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 Funktionaldeterminate.

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

Satz 2.9 *X und Y seien unabhängig mit Dichte f_X und f_Y. Dann ist die Dichte von $X + Y$ die Faltung von f_X und f_Y:*

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

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

$$f * f(z) = \int_{-\infty}^{\infty} f(z - x)f(x)dx = \int_{0}^{z} \lambda e^{-\lambda(z-x)}\lambda e^{-\lambda x}dx = \int_{0}^{z} \lambda^2 e^{-\lambda z}dx = z\lambda^2 e^{-\lambda z}.$$

Eine ähnliche Formel gibt es auch für diskrete Zufallsvariable:

Satz 2.10 *X und Y seien unabhängig mit Wahrscheinlichkeitsfuktion p_X und p_Y. Dann ist die Wahrscheinlichkeitsfunktion von $X + Y$ die (diskrete) Faltung von p_X und p_Y:*

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

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

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

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:

Definition 2.17 *Die verallgemeinerte Inverse der Verteilungsfunktion F ist gegeben durch*

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

Satz 2.11 *Wenn U auf $[0, 1]$ gleichverteilt ist, dann hat*

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

die Verteilungsfunktion F.

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

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

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 2.18 *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

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

Satz 2.12 (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 $X \le Y$, dann ist auch $\mathbb{E}(X) \le \mathbb{E}(Y)$,

4. wenn X und Y unabhängig sind, dann gilt $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Satz 2.13 (Satz vom unachtsamen Statistiker) *X sei stetig verteilt mit der Dichte f_X, $Y = g(X)$. Dann ist*

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

Wir müssen also nicht erst die Verteilung von Y bestimmen, um den Erwartungswert zu berechnen.

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 Addiditivität ergibt sich etwa so (wieder für diskrete Verteilungen):

$$mathbbE(X + Y) = \sum_{x,y} (x + y) {}_{\mathsf{I}X,Y}(x, y) = \sum_x x \sum_y p_{X,Y}(x, y) + \sum_y y \sum_x p_{X,Y}(x, y) =$$

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

Wir berechnen als Beispiel den Erwartungswert der Binomialverteilung:

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

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

Definition 2.19 *Die Varianz von X ist*

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

Satz 2.14 (Steinerscher Verschiebungssatz) *Für beliebiges reelles x gilt*

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

Satz 2.15 (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 $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$.

Definition 2.20 *X und Y seien Zufallsvariable mit endlicher Varianz. Dann heißt*

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

ie 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.

Satz 2.16 (Ungleichung von Markov) *X sei eine nichtnegative Zufallsvariable, $\lambda > 0$. Dann ist*

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

Satz 2.17 (Ungleichung von Chebychev)

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

Satz 2.18 (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*

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

2.5 Folgen von Zufallsvariablen

Satz 2.19 (schwaches Gesetz der großen Zahlen) *$(X_n, n \in \mathbb{N})$ sei eine Folge von unabhängigen, identisch verteilten Zufallsvariablen mit endlicher Varianz, $S_n = X_1 + \ldots + X_n$. Dann gilt*

$$\mathbb{P}(|\frac{S_n}{n} - \mathbb{E}(X_1)| \geq \epsilon) \to 0$$

für jedes $\epsilon > 0$.

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

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

und das geht natürlich gegen 0.

Satz 2.20 (Starkes Gesetz der großen Zahlen) $(X_n, n \in \mathbb{N})$ *sei eine Folge von unabhängigen, identisch verteilten Zufallsvariablen mit endlichem Erwartungswert,* $S_n = X_1 + \ldots + X_n$*. Dann konvergiert* $\frac{S_n}{n}$ *mit Wahrscheinlichkeit 1 gegen* $\mathbb{E}(X_1)$*.*

Satz 2.21 (Zentraler Grenzwertsatz) $(X_n, n \in \mathbb{N})$ *sei eine Folge von unabhängigen, identisch verteilten Zeufallsvariablen mit* $\mathbb{E}(X) = \mu$*,* $\mathbb{V}(X) = \sigma^2$*. Dann gilt*

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

Eine Summe von unabhängigen identisch verteilten Zufallsvariablen ist also näherungsweise normalverteilt.

2.6 Spezielle Verteilungen

2.6.1 Diskrete Verteilungen

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

2.6.2 Stetige Verteilungen

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

Die Chiquadratverteilung ergibt sich als die Verteilung der Summe

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

wobei $(X_1, \ldots, X_n)$ 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 werden gemeinhin als “Freiheitsgrade” bezeichnet.

Kapitel 3

Markovketten

3.1 Stochastische Prozesse

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* M_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,*

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 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 (Ein berühmter Satz von Kolmogorov besagt, dass durch die Angabe dieser "endlichdimensionalen Randverteilungen" ein stochastischer Prozess festgelegt wird). Einige Möglichkeiten zählen wir jetzt auf:

Definition 3.2 *Der Prozess* $(X_t, t \in T)$ *heißt Prozess mit unabhängigen Zuwächsen, wenn für* $t_1 < t_2 < \ldots > t_n$ *die Zufallsvariablen*

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

unabhängig sind.

Hier verwenden wir die Notation $X(t)$ für X_t, damit die Indizes nicht zu sehr überladen werden. Prozesse mit unabhängigen Zuwächsen in diskreter Zeit sind einfach Summen von unabhängigen Zufallsvariablen, wie wir sie im letzten Kapitel untersucht haben.

Definition 3.3 *Der Prozess* $X(t)$ *heißt Markovprozess, wenn für wenn für* $t_1 < t_2 < \ldots > t_n$

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

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

3.1.1 Stationäre Prozesse

Eine Eigenschaft einer Folge $(X_1, X_2, \ldots)$ von unabhängigen identisch verteilten Zufallsvariablen ist, dass $(X_n, X_{n+1}, \ldots)$ ebenfalls eine folge von unabhängig identisch verteilten Zufallsvariablen ist. 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 Vorstellung hat weitreichende Konsequenzen:

Satz 3.1 (Ergodensatz von Birkhoff) *Wenn die Folge (X_n) stationär ist und endlichen Erwartungswert hat, dann existiert*

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

mit Wahrscheinlichkeit 1 und

$$\mathbb{E}(X_\infty) = \mathbb{E}(X_1).$$

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 $f(X_n, \ldots, X_{n+k})$), heißen ergodisch.

3.2 Markovketten in diskreter Zeit

3.2.1 Übergangswahrscheinlichkeiten

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

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

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. Aus dem Satz von der vollständigen Wahrscheinlichkeit erhalten wir die Chapman-Kolmogorov Gleichungen

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

in Matrixnotation mit

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

lauten die Chapman-Kolmogorov Gleichungen

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

und mit $P = P(1)$

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

Wir nennen P die Übergangsmatrix und $P(t)$ die t-stufige Übergangsmatrix.

Wir setzen zusätzlich $p_i(t) = \mathbb{P}(X_t = i)$ und $p(t) = (p_i(t), i \in M_X)$ (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.

In Hinkunft verwenden wir zur Abkürzung die Notationen

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

und

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

3.2.2 Klasseneigenschaften

Wir definieren

Definition 3.5 *Der Zustand j heißt Nachfolger von i ($i \to j$), 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 (no pun intended).

Definition 3.6 *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 3.7 *Die Periode eines Zustandes ist*

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

Etwas spannender ist die Rekurrenz: 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 3.2 *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.

Bei der Rekurrenz kann man weiter unterscheiden:

Definition 3.8 *i sei ein rekurrenter Zustand. Wenn*

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

gilt, dann heißt i positiv rekurrent, sonst nullrekurrent.

Definition 3.9 *$(\pi_i, i \in M_X)$ heißt stationäre Verteilung, wenn*

$$\pi_i \geq 0,$$

$$\sum_i \pi_i = 1$$

und

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

Satz 3.3 *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.

Für einen absorbierenden Zustand i_0 definieren wir die Absorptionswahrscheinlichkeit a_i als

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

Satz 3.4 *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.

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

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

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

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

Wenn es mehrere absorbierende Zustände 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 $B \subseteq A$ $(a_i = a_i(B))$. 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 v_i lösen und erhält so die bedingte Erwartung bzw. Varianz.

3.2.3 Markov Chain Monte Carlo

Kapitel 4

Statistik

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 4.1 *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 $\mathcal{P}$.

Definition 4.2 *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.

4.1 Schätztheorie

4.1.1 Punktschätzung

Die erste Aufgabe, mit der wir uns beschäftigen, besteht darin, aus einer Stichprobe Schätzwerte für den unbekannten Parameter zu bestimmen. Wir definieren

Definition 4.3 *Ein Schätzer ist eine Folge $(\hat{\theta}_n)$ von Statistiken.*

Diese Definition lässt recht dumme Schätzer zu (z.B. 42, vgl. Douglas Adams), deshalb definieren wir einige Eigenschaften, die wir von Schätzern verlangen können:

Definition 4.4 *Ein Schätzer 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*

$$\mathbb{E}_\theta(\hat{\theta}_n) = \theta.$$

- *effizient, wenn er erwartungstreu ist und die kleinste Varianz unter allen erwartungstreuen Schätzern hat.*

Eine Methode, um Schätzer zu konstruieren, ist die Momentenmethode. Ihr liegt das Gesetz der großen Zahlen zugrunde. Am einfachsten geht das, wenn es nur einen Parameter gibt. 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.

Wenn es $d > 1$ Parameter gibt, dann verwendet man zusätzlich höhere Momente:

$$\mathbb{E}_\theta(X^i) = \frac{1}{n}\sum_{j=1}^{n} X_j^i,\ i = 1, \ldots, d$$

und löst nach θ auf.

Die andere Methode, die wir hier betrachten, ist die Maximum Likelihood Methode: Wenn die Wahrscheinlichkeit, die aktuelle Stichprobe zu erhalten, für einen Parameter sehr klein ist, spricht das gegen diesen Parameter. Umgekehrt heißt das, dass ein Parameterwert umso plausibler ist, je größer die Wahrscheinlichkeit der Stichprobe unter diesem Parameter ist.

Definition 4.5 *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 4.6 *Der Maximum-Likelihood (ML-) Schätzer ist der Wert von θ, der die Likelihoodfunktion maximiert.*

Damit der ML-Schätzer konsistent ist, müssen die Dichten gewisse Regularitätsvoraussetzungen erfüllen. Bei der Suche nach einem effizienten Schätzer hilft der

Satz 4.1 (Cramér-Rao) *Wenn p_θ bzw. f_θ zweimal nach θ differenzierbar ist und zusätzliche Regularitätsvoraussetzungen erfüllt, dann gilt für jeden Erwartungstreuen Schätzer $\hat{\theta}_n$*

$$\mathbb{V}(\hat{\theta}_n) \geq \frac{1}{I_n(\theta)} = \frac{1}{nI(\theta)}.$$

Dabei ist

$$I_n(\theta) = -\mathbb{E}_\theta(\frac{\partial^2}{\partial\theta^2}\log(L(X_1, \ldots, X_n; \theta)))$$

und

$$I(\theta) = -\mathbb{E}_\theta(\frac{\partial^2}{\partial\theta^2}\log(f_\theta(X)))$$

bzw.

$$I(\theta) = -\mathbb{E}_\theta(\frac{\partial^2}{\partial\theta^2}\log(p_\theta(X)))$$

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. Wenn es einen solchen Schätzer gibt, dann stimmt er mit dem Maximum-Likelihood Schätzer überein (überhaupt ist der ML-Schätzer unter gewissen Regularitätsvorraussetzungen asymptotisch normalverteilt mit Mittel θ und Varianz $1/I_n(\theta)$, also gewissermaßen "asymptotisch effizient").

4.1.2 Intervallschätzung

Die Theorie aus dem vorigen Abschnitt gibt ins 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 positive (wenn auch sehr kleine) Wahrscheinlichkeit hat, gezogen zu werden.

Wir werden uns also damit begnügen müssen, ein Intervall zu bestimmen, das den gesuchten Parameter mit einer gewissen Wahrscheinlichkeit enthält.

Definition 4.7 *$a = a(X_1, \ldots, X_n) \leq b = b(X_1, \ldots, X_n)$ 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.

Ein möglicher Ausgangspunkt für die Konstruktion von Konfidenzintervallen ist ein Schätzer für θ. Unter sehr günstigen Bedingungen 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 den Spezialfall einer Normalverteilung können wir exakte Konfidenzintervalle angeben:

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:

$$[\frac{(n-1)S_n^2}{\chi^2_{n-1;\frac{1+\gamma}{2}}}, \frac{(n-1)S_n^2}{\chi^2_{n-1;\frac{1-\gamma}{2}}}].$$

Für Anteilswerte verwenden wir das approximative Konfidenzintervall:

$$[\hat{p} - z_{\frac{1+\gamma}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, , \hat{p} - z_{\frac{1+\gamma}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}].$$

Ein anderer Weg, um Konfidenzintervalle zu erhalten, führt über die Theorie der statistischen Tests.

4.2 Tests

4.2.1 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".

Wir definieren zuerst

Definition 4.8 *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".

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 4.9 *Ein Test heißt vom Niveau α, wenn die Wahrscheinlichkeit für einen Fehler erster Art (die bei zusammengesetzten Hypothesen eine Funktion dvon θ ist) nicht größer als α ist.*

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 4.10 *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 4.2 (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 $L(X_1, dotsX_n, \theta_0)/L(X_1, \ldots X_n, \theta_1)$ berechnet)*

4.2.2 Spezielle Tests

Für den Mittelwert einer Normalverteilung, wenn σ^2 unbekannt ist:

$$T = \frac{\bar{X}_n - \mu_0}{\sqrt{S_n^2/n}}$$

$H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu > \mu_0$: verwerfen, wenn $T > t_{n-1;1-\alpha}$.

$H_0 : \mu \geq \mu_0$ gegen $H_1 : \mu < \mu_0$: verwerfen, wenn $T < -t_{n-1;1-\alpha}$.

$H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$: verwerfen, wenn $|T| > t_{n-1;1-\alpha/2}$.
Für die Varianz einer Normalverteilung:

$$T = \frac{(n-1)S_n^2}{\sigma_0^2}$$

$H_0 : \sigma^2 \geq \sigma_0^2$ gegen $H_1 : \sigma^2 < \sigma_0^2$: verwerfen, wenn $T < \chi^2_{n-1;\alpha}$.
$H_0 : \sigma^2 = \sigma_0^2$ gegen $H_1 : \sigma^2 \neq \sigma_0^2$: verwerfen, wenn $T > \chi^2_{n-1;1-\alpha/2}$ oder $T < \chi^2_{n-1;\alpha/2}$.
$H_0 : \sigma^2 \leq \sigma_0^2$ gegen $H_1 : \sigma^2 > \sigma_0^2$: verwerfen, wenn $T > \chi^2_{n-1;1-\alpha}$.
Für Anteilswerte:

$$T = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

$H_0 : p \leq p_0$ gegen $H_1 : p > p_0$: verwerfen, wenn $T > z_{1-\alpha}$.
$H_0 : p \geq p_0$ gegen $H_1 : p < p_0$: verwerfen, wenn $T < -z_{1-\alpha}$.
$H_0 : p = p_0$ gegen $H_1 : p \neq p_0$: verwerfen, wenn $|T| > z_{1-\alpha/2}$.

4.2.3 Der Chi-Quadrat-Anpassungstest

Die Frage, die wir hier untersuchen, ist, ob eine Stichprobe aus einer gegebenen Verteilung stammt. Das geht am einfachsten, wenn es sich bei der hypothetischen Verteilung um eine diskrete Verteilung mit endlich vielen Werten $1, \ldots, k$ handelt. Wir testen also

$$H_0 : X \sim P = (p_1, \ldots, p_k)$$

gegen die Alternative $X \not\sim P$. Diese Frage 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 bagezogen wird, es sind also statt $k-1$ $k-1-d$ Freiheitsgrade, wobei d die Anzahl der geschätzten Parameter ist (im Normalverteilunsgbeispiel ist $d = 2$).

4.2.4 Tests und Konfidenzintervalle

Es gibt einen interessanten Zusammenhang zwischen Tests und Konfidenzintervallen:

Satz 4.3 *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 aller* θ_0*, für die* $H_0 : \theta = \theta_0$ *nicht verworfen wird, ein Konfidenzintervall mit Überdeckungswahrscheinlichkeit* $\gamma = 1 - \alpha$*.*

Kapitel 5

Informationstheorie

5.1 Entropie und Information

Das Bar Kochba Spiel: Spieler A wählt einen 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.

Interessanter wird es, wenn 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 Fragestrategi entspricht. Wir suchen also einen Binärbaum, für den die mittlere Blattlänge minimal ist. Diesen kann man rekursiv mit dem Huffman-Algorithmus 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 $P^* = (p_2, \ldots, p_{m-1}, p^*_{m-2})$
5. Ersetze Blatt $m - 1$ durch einen inneren Knoten mit den Blättern $m - 1$ und m.

Wir können unsere Fragenstrategien durch Binärbäume darstellen, dabei entspricht jeder innere Knoten einer Frage, jeder Endknoten ("Blatt") einem Wert von X. Die Suche nach der optimalen Strategie lässt sich so formulieren, 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, also seine Entfernung von der Wurzel. Für diese Aufgabe ist es nützlich zu wissen, wann ein Binärbaum mit den Blattlängen $l_1, \ldots l_m$ existiert.

Satz 5.1 (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.

Die Suche nach der optimalen Strategie besteht also darin,

$$\sum p_i l_i$$

under der Nebenbedingung

$$\sum 2^{-l_i} \leq 1$$

zu minimieren. Es ist leicht einzusehen, dass im optimalen Fall in der letzten Ungleichun 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 5.1 *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 5.2

$$H(P) \leq H^*(P) \leq H(P) + 1.$$

Die untere Abschätzung haben wir oben gezeigt, die obere folgt aus der Tatsache, dass $l_i = \lceil \log_2(1/pi) \rceil$ die Kraft'sche 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) Entropie $nH(P)$, 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.

Wenn X und Y zwei Zufallsvariable sind, können wir (X, Y) als eine Zufallsvariable mit endlich vielen Werten betrachten 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 5.2

$$H(X|Y = y) = -\sum_x p(x|y) \log_2(p(x|y))$$

und nennen

$$H(X|Y) = \sum_x p_Y(y) H(X|Y = y)$$

die bedingte Entropie von X unter Y.

Es gilt

Satz 5.3

$$H(X, Y) = H(Y) + H(X|Y),$$

$$\max(H(X), H(Y)) \leq H(X, Y) \leq H(X) + H(Y), `$$

$$H(X|Y) \leq H(X),$$

$$H(X|Y,Z) \leq H(X|Y).$$

$$H(X,Y) = H(X) + H(Y)$$

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 5.3

$$I(X,Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

heißt die Information zwischen X und Y.

Der letzte Satz impliziert

Satz 5.4

$$0 \leq I(X,Y) \leq \min(H(X), H(Y)).$$

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 5.5 *Wenn $\mathbb{P}(X \neq Y) \leq \epsilon$ ist, dann gilt*

$$I(X,Y) \geq H(X) - H(\epsilon, 1-\epsilon) - \epsilon \log_2(m).$$

Wir betrachten nun 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 (das stimmt nicht genau, aber wir können etwa eine fast optimale Strategie finden, in der sich die Anzahl der Fragen um weniger als 1 davon unterscheidet), 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 zuviel.

Das führt uns zu der Definition

Definition 5.4

$$D(P,Q) = \sum_i p_i \log_2(p_i/q_i).$$

heißt die Informationsdivergenz (I-divergenz, Kullback-Leibler Distanz, relative Entropie, Strafe des Irrtums) zwischen P und Q.

5.2 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 Pr—afix (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.

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 Binaärzahl 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 werdn müssen) codiert man $(f_{i-1} + f_i)/2$ mit $\lceil \log_2(1/p_i) \rceil + 1$ Bits.

Diese Idee kann man auf das Kodoeren von ganzen Blöcken anwenden, im Estremfall 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, das die Nachricht kodiert. Verahren, 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 5.5
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 5.6 *Die Codewortlängen in einem endlich eindeutig entzifferbaren Code erfüllen die Kraft'sche Ungleichung.*

Aus diesem Satz folgt, dass es zu jedem endlich eindeutige entzifferbaren Code einen präfixfreien Code mit denselben Codewortlängen (und damit mit derselben mittleren Codewortlänge) gibt.

(universelle Codes)

5.3 Informationsquellen

Definition 5.6 *Eine Informationsquelle ist eine Folge $\mathcal{X} = (X_1, \ldots)$ von Zufallsvariablen.*

Die verschidenen Möglichkeiten für die Abhängigkeitsstruktur dieser Folge ergeben die Definitionen

Definition 5.7 *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 5.8 *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 § 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 5.7 (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.

5.4 Blockcodes

5.5 Kanalcodierung

5.6 Natürliche Sprachen als Informationsquellen

Die Verteilungsfunktion der Standardnormalverteilung:

$$\Phi(x) = \int_{-\infty}^{x} e^{-\frac{x^2}{2}}$$

	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