

Algorithmische Bioinformatik

DRAFT - Do not distribute!

Sven Rahmann und Martin Vingron

Skript zur Vorlesung
im Wintersemester 2002/2003

Fakultät für Mathematik und Informatik
Freie Universität Berlin

Berlin, im Februar 2003

Inhaltsverzeichnis

1 Grundlagen der Wahrscheinlichkeitstheorie	3
1.1 Wahrscheinlichkeiten und Ereignisse	3
1.2 Bedingte Wahrscheinlichkeiten	5
1.2.1 Definition und Eigenschaften	5
1.2.2 Bayes'sche Formel	6
1.2.3 Unabhängigkeit	7
1.3 Zufallsvariablen	8
1.3.1 Wichtige diskrete Zufallsvariablen	9
1.3.2 Stetige Zufallsvariablen	10
1.3.3 Momente	10
1.3.4 Wichtige Beispiele	10
1.4 Markovketten	10
1.5 Markovprozesse	10
2 Grundlagen der Statistik	11
A IUPAC-Symbole	12
A.1 IUPAC Nukleotid-Symbole	12
A.2 IUPAC Aminosäure-Symbole	13
B Glossar	14

Kapitel 1

Grundlagen der Wahrscheinlichkeitstheorie

Dieses Kapitel fasst kurz die wichtigsten Definitionen und Sätze aus der Wahrscheinlichkeitsrechnung zusammen, die in bioinformatischen Anwendungen häufig benötigt werden. Dem mit der Materie nicht vertrauten Leser seien die einführenden Lehrbücher [?] und [?] empfohlen.

1.1 Wahrscheinlichkeiten und Ereignisse

Wenn im folgenden von Zufallsexperimenten die Rede ist, kann man stets an eines der folgenden Beispiele denken.

1. Ergebnis eines Würfelwurfs; eine Zahl zwischen 1 und 6.
2. Anzahl der Münzwürfe (Kopf oder Zahl), bis zum ersten Mal Kopf erscheint; eine ganze nichtnegative Zahl.
3. Das Gewicht eines neugeborenen Babys in Kilogramm; potenziell eine beliebige nichtnegative Zahl. Alternatives Beispiel: Eine zufällige reelle Zahl im Intervall $[0, 1]$.

Die möglichen Ausgänge eines Zufallsexperiments bilden den *Stichprobenraum* (*sample space*) Ω .

Im ersten Beispiel ist der Stichprobenraum endlich ($\Omega = \{1, 2, 3, 4, 5, 6\}$), im zweiten abzählbar¹ ($\Omega = \{1, 2, 3, \dots\}$), und im dritten überabzählbar² ($\Omega = [0, \infty)$).

Eine Teilmenge $E \subset \Omega$ nennt man *Ereignis* (*event*). Im Fall eines überabzählbaren Stichprobenraumes kann es manchmal zu Paradoxa führen, wenn man alle Teilmengen als Ereignisse zulässt. In der Praxis interessieren aber auch nicht alle Teilmengen (zum Beispiel beschränkt man sich bei $\Omega = \mathbb{R}$ auf Intervalle und daraus abgeleitete Mengen). Ein sinnvolles Ereignissystem (man nennt es σ -*Algebra*) ist eine Menge \mathcal{A} von Teilmengen von Ω mit den folgenden Mindesteigenschaften: (1) $\Omega \in \mathcal{A}$, (2) mit einem Ereignis A ist auch sein Komplement $A^c := \Omega \setminus A$ in \mathcal{A} , und (3) ist $(A_i)_{i=1,2,\dots}$ eine abzählbare Familie von Ereignissen aus \mathcal{A} , so ist auch ihre Vereinigung und ihr Durchschnitt in \mathcal{A} . Weitere Eigenschaften werden in der Maßtheorie untersucht; wir gehen nicht näher darauf ein.

Im ersten Beispiel bezeichnet $E := \{2, 4, 6\}$ das Ereignis “Die gewürfelte Augenzahl ist gerade”. Die leere Menge $E = \emptyset$ heißt *unmögliches Ereignis*; der Stichprobenraum Ω selbst heißt *sicheres*

¹Eine Menge A heißt abzählbar, wenn es eine injektive Abbildung von A auf die natürlichen Zahlen gibt.

²Jedes Intervall reeller Zahlen bildet eine überabzählbare Menge.

Ereignis. Zwei Ereignisse A, B heißen *unvereinbar* oder *disjunkt*, wenn ihr Durchschnitt $A \cap B$ das unmögliche Ereignis \emptyset ist.

Die σ -Algebra aller zugelassenen Ereignisse auf Ω bezeichnen wir mit \mathcal{A} . Im Falle eines endlichen oder abzählbaren Ω sei stets $\mathcal{A} = 2^\Omega$, d.h., wir betrachten alle Teilmengen von Ω als potenzielle Ereignisse³. Im Falle eines überabzählbaren Ω gilt oft $\mathcal{A} \subset 2^\Omega$; insbesondere im Falle $\Omega = \mathbb{R}$ beschränken wir uns auf “sinnvolle” Ereignisse (genauer: die kleinste σ -Algebra, die alle Intervalle enthält); die sogenannte *Borel’sche σ -Algebra*. Es ist gar nicht so einfach, eine Teilmenge von \mathbb{R} zu konstruieren, die nicht zur Borel’schen σ -Algebra gehört; daher ist diese für unsere Zwecke vollkommen ausreichend.

Definition 1.1 (Verteilung). Ein *Wahrscheinlichkeitsmaß* oder eine (*Wahrscheinlichkeits*)-*Verteilung* (*probability measure* or (*probabilty*) *distribution*) auf Ω ist eine Funktion $P : 2^\Omega \rightarrow [0, 1]$, die jedem Ereignis $E \subset \Omega$ eine Wahrscheinlichkeit $P(E)$ mit den folgenden Eigenschaften zuordnet:

$$P(\Omega) = 1, \tag{1.1}$$

$$0 \leq P(A) \leq 1 \quad \text{für alle } A \in \mathcal{A}, \tag{1.2}$$

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i) \quad \begin{array}{l} \text{für alle abzählbaren Familien } (A_i) \\ \text{von paarweise disjunkten Ereignissen (“ σ -Additivität”).} \end{array} \tag{1.3}$$

Aus der σ -Additivität folgt insbesondere auch die endliche Additivität: $P(A \cup B) = P(A) + P(B)$, wenn A und B unvereinbar sind.

Das Tripel (Ω, \mathcal{A}, P) heißt *Wahrscheinlichkeitsraum* (*Probability space*). Ist Ω endlich oder abzählbar, so spricht man von einem *diskreten* (*discrete*) Wahrscheinlichkeitsraum.

Beispiel 1.1 (Würfeln). Im Falle eines endlichen Ω , wie z.B., beim Würfeln, gibt man P oft in Form eines Vektors an, der jedem *Elementarereignis* $\omega \in \Omega$ eine Wahrscheinlichkeit zuordnet. Beim Würfeln ist $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Wir betrachten zunächst einen fairen Würfel, bei dem die Wahrscheinlichkeit jedes Elementarereignisses $P(\{i\}) = 1/6$ für alle $i \in \Omega$ ist. Wegen der Additivität folgt für beliebige Ereignisse $P(A) = |A|/6$, wobei $|A|$ die Anzahl der Elemente in A ist.

Betrachten wir nun einen unfairen Würfel, der höhere Augenzahlen bevorzugt. Genauer gelte $P(\{i\}) = p_i$ mit $p = (\frac{10}{147}, \frac{12}{147}, \frac{15}{147}, \frac{20}{147}, \frac{30}{147}, \frac{60}{147})$ für Elementarereignisse $i \in \{1, 2, 3, 4, 5, 6\}$. Das Ereignis, eine gerade Zahl zu würfeln, hat somit die Wahrscheinlichkeit $P(\{2, 4, 6\}) = 92/147$.

Übung 1.1. Bestätigen Sie, dass der Vektor p im vorigen Beispiel ein Wahrscheinlichkeitsmaß definiert. Wie groß ist die Wahrscheinlichkeit, mindestens eine Vier zu würfeln?

Beispiel 1.2 (Zufallszahlen). Jetzt sei Ω das überabzählbare Intervall $[0, 1]$. Jedes “Elementarereignis” $\{x\}$ ($0 \leq x \leq 1$) hat nun Wahrscheinlichkeit null, ebenso alle endlichen oder abzählbaren Vereinigungen von Elementarereignissen. Die “interessanten” Ereignisse (die also auch in \mathcal{A} enthalten sein müssen) sind jetzt zum Beispiel Teilintervalle $[a, b]$ mit $0 \leq a < b \leq 1$. Die Gleichverteilung auf $[0, 1]$ (“zufälliges Ziehen einer Zahl”) ist so definiert, dass $P([a, b]) = b - a$ gilt.

Das Beispiel zeigt, dass auch Ereignisse $A \neq \emptyset$ die Wahrscheinlichkeit null, und auch Ereignisse $B \neq \Omega$ die Wahrscheinlichkeit eins haben können. Ein Ereignis B mit $P(B) = 1$ heißt *fast sicher* (*almost sure*).

Übung 1.2. Wie groß ist die Wahrscheinlichkeit, dass die im vorigen Beispiel gezogene Zufallszahl rational ist, d.h., sich als Bruch p/q mit natürlichen Zahlen p, q darstellen lässt? Hinweis: σ -Additivität.

³Für eine beliebige Menge A bezeichnet 2^A die Menge aller Teilmengen von A . Ist A endlich und enthält n Elemente, so enthält 2^A genau 2^n Elemente.

Gleichung (1.3) zeigt, wie man die Wahrscheinlichkeit von Ereignisvereinigungen ausrechnet, wenn die Ereignisse disjunkt sind. Trifft dies nicht zu, muss man die Ereignisse passend zerlegen. Im folgenden soll das Symbol \uplus andeuten, dass die vereinigten Mengen disjunkt sind. Es ist $A \cup B = (A \setminus B) \uplus (A \cap B) \uplus (B \setminus A)$, daher $P(A \cup B) = P(A \setminus B) + P(A \cap B) + P(B \setminus A) = P(A \setminus B) + P(A \cap B) + P(B \setminus A) + P(A \cap B) - P(A \cap B) = P(A) + P(B) - P(A \cap B)$.

Beispiel 1.3 (Skat-Karten). Ein Skatspiel hat 32 Karten; jede Karte hat eine Farbe (Kreuz, Pik, Herz, Karo) und einen Typ (7, 8, 9, 10, Bube, Dame, König, Ass). Wie groß ist die Wahrscheinlichkeit, dass eine zufällig gezogene Karte entweder ein König (Ereignis A) oder eine Pik-Karte (Ereignis B) ist? Es ist $P(A) = 4/32$ und $P(B) = 8/32$; die Ereignisse sind aber nicht disjunkt. Addiert man einfach die Wahrscheinlichkeiten, zählt man den Pik König ($A \cap B$) doppelt; daher muss seine Wahrscheinlichkeit einmal von der Summe subtrahiert werden: $P(A \cup B) = \frac{4+8-1}{32} = \frac{11}{32}$.

Der obige Ansatz lässt sich auf Vereinigungen von mehr als zwei Ereignissen verallgemeinern.

Lemma 1.1 (Vereinigungen von Ereignissen). Seien A_1, \dots, A_n und A, B, C Ereignisse.

$$P(A_1 \cup \dots \cup A_n) = \sum_{i=1}^n (-1)^{i-1} p_i, \quad \text{wobei} \quad (1.4)$$

$$p_i := \sum_{j_1 < j_2 < \dots < j_i} P(A_{j_1} \cap \dots \cap A_{j_i}).$$

$$\text{Insbesondere} \quad P(A \cup B) = P(A) + P(B) - P(A \cap B), \quad (1.5)$$

$$\text{und} \quad P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(B \cap C) - P(A \cap C) + P(A \cap B \cap C). \quad (1.6)$$

Übung 1.3 (Würfeln mit 2 Würfeln). Man würfelt mit 2 Würfeln. Der Stichprobenraum Ω besteht aus allen 36 möglichen Paaren von Augenzahlen, P sei die Gleichverteilung auf Ω . Sei A das Ereignis, einen Pasch zu würfeln, B das Ereignis, die Augensumme 8 zu erzielen, und C das Ereignis, dass der zweite Würfel die Augenzahl 5 zeigt. Berechnen Sie $P(A \cup B \cup C)$ (a) durch explizites Aufzählen aller Elementarereignisse, auf die A , B , oder C zutrifft, und (b) mit Hilfe von Lemma 1.1.

1.2 Bedingte Wahrscheinlichkeiten

Oft stehen vor der Durchführung (oder Beobachtung) eines Zufallsexperiments schon gewisse Vorabinformationen zur Verfügung. Man interessiert sich dann für die Verteilung unter Berücksichtigung dieser Informationen. Dies wird formalisiert durch die *bedingte Wahrscheinlichkeit (conditional probability)*.

1.2.1 Definition und Eigenschaften

Sei (Ω, \mathcal{A}, P) ein Wahrscheinlichkeitsraum und B ein Ereignis mit $P(B) > 0$ (die ‘‘Vorabinformation’’). Uns interessiert die Verteilung P unter der Annahme (Bedingung), dass B eingetreten ist; wir bezeichnen die resultierende Verteilung mit P_B . Konkret bedeutet das, dass wir den Stichprobenraum auf B einschränken. Von jedem Ereignis A betrachten wir somit nur noch den Teil, der auch in B liegt. Das Ereignis B selbst wird zum sicheren Ereignis.

Definition 1.2 (Bedingte Wahrscheinlichkeit). Sei (Ω, \mathcal{A}, P) ein Wahrscheinlichkeitsraum und B ein Ereignis mit $P(B) > 0$. Dann heißt

$$P(A | B) := \frac{P(A \cap B)}{P(B)} \quad (1.7)$$

die bedingte Wahrscheinlichkeit von A unter (gegeben) B .

Beispiel 1.4 (Würfelwurf). Es sei bekannt, dass die gewürfelte Augenzahl mindestens vier beträgt (B). Wie groß ist die bedingte Wahrscheinlichkeit, eine gerade Zahl (A) zu würfeln? Wir haben $\Omega = \{1, 2, 3, 4, 5, 6\}$, $B = \{4, 5, 6\}$ und $A = \{2, 4, 6\}$. Es folgt $P(A | B) = P(A \cap B) / P(B) = P(\{4, 6\}) / P(\{2, 4, 6\}) = \frac{2/6}{3/6} = 2/3$. Auf dem eingeschränkten Wahrscheinlichkeitsraum B sind 2 von 3 möglichen Augenzahlen gerade.

Bemerkung. Zur Notation: Statt $A \cap B$ schreiben wir in Zukunft oft einfach A, B ; gemeint ist nach wie vor das gemeinsame Auftreten der Ereignisse A und B .

Oft kann man bedingte Wahrscheinlichkeiten dazu benutzen, um die Wahrscheinlichkeit von Durchschnitten von Ereignissen zu berechnen. Wir können (1.7) wie folgt umschreiben:

$$P(A, B) = P(B) \cdot P(A | B). \quad (1.8)$$

Beispiel 1.5 (Ziehen ohne Zurücklegen). Eine Urne enthält zwei rote und drei grüne Kugeln. Nacheinander werden zwei Kugeln ohne Zurücklegen gezogen. Uns interessiert die Wahrscheinlichkeit des Ereignisses, dass die erste Kugel rot (B) und die zweite Kugel grün (A) ist. Es ist $P(B) = 2/5$ und $P(A | B) = 3/4$ (drei von den vier verbleibenden Kugeln sind grün), also $P(A, B) = 3/10$. Es wäre fatal gewesen, hier auf A statt auf B zu bedingen, denn $P(A)$ muss man erst mühsam ausrechnen, während man $P(B)$ sofort aus der Problemstellung ablesen kann.

Tipp: Gestaltet sich die Berechnung der Wahrscheinlichkeit eines Ereignisses schwierig, hilft oft geschicktes Zerlegen und Bedingen! Man kann stets ein Ereignis A wie folgt in Ω disjunkt zerlegen:

$$A = (A, B) \uplus (A, B^c), \quad \text{wobei } B^c := \Omega \setminus B.$$

Mit (1.3) und (1.8) folgt der

Satz 1.1 (Satz von der totalen Wahrscheinlichkeit). Sei B ein Ereignisse mit $P(B) > 0$ und $P(B^c) > 0$. Allgemeiner sei (B_i) eine disjunkte Zerlegung von Ω mit $P(B_i) > 0$ für alle i . Dann gilt

$$P(A) = P(B) \cdot P(A | B) + P(B^c) \cdot P(A | B^c), \quad \text{bzw.} \quad (1.9)$$

$$P(A) = \sum_i P(B_i) \cdot P(A | B_i). \quad (1.10)$$

Beispiel 1.6 (Ziehen ohne Zurücklegen; Fortsetzung). Eine Urne enthält zwei rote und drei grüne Kugeln. Nacheinander werden zwei Kugeln ohne Zurücklegen gezogen. Wie groß ist die Wahrscheinlichkeit, dass die zweite Kugel grün ist (A)? Wir bedingen auf die Farbe der ersten Kugel. Sei B das Ereignis, dass die erste Kugel rot ist. Mit dem Satz von der totalen Wahrscheinlichkeit folgt

$$P(A) = P(B) \cdot P(A | B) + P(B^c) \cdot P(A | B^c) = \frac{2}{5} \cdot \frac{3}{4} + \frac{3}{5} \cdot \frac{2}{4} = \frac{3}{5}.$$

1.2.2 Bayes'sche Formel

Da $P(A, B)$ symmetrisch in A und B ist, spielt es in (1.8) keine Rolle, ob man auf A oder B bedingt (sofern $P(A) > 0$ und $P(B) > 0$). Man erhält also $P(A, B) = P(B) \cdot P(A | B) = P(A) \cdot P(B | A)$ und daraus den folgenden wichtigen Satz.

Satz 1.2 (Bayes'sche Formel). Es gelte $P(A) > 0$, $P(B) > 0$. Dann gilt

$$P(B | A) = P(A | B) \cdot \frac{P(B)}{P(A)}, \quad \text{und mit (1.9),} \quad (1.11)$$

$$= \frac{P(A | B) \cdot P(B)}{P(A | B) \cdot P(B) + P(A | B^c) \cdot P(B^c)}. \quad (1.12)$$

Die Bayes'sche Formel ist deswegen so wichtig, weil sie es erlaubt, Ereignis und Bedingung zu vertauschen. Das folgende Beispiel sollte man nicht so schnell vergessen.

Beispiel 1.7 (Erkennung von Transkriptionsfaktor-Bindestellen). Angenommen, wir haben einen Algorithmus der uns nach der Eingabe einer kurzen DNA-Sequenz ("Sequenz-Fenster") eine Entscheidung liefert, ob dieses Fenster eine spezielle Transkriptionsfaktor-Bindestelle (TFBS) darstellt oder nicht. Es ist bekannt, dass diese TFBS im Schnitt nur alle 1000 Fenster auftaucht. Wenn eine TFBS vorliegt, erkennt der Algorithmus dies mit 99%iger Wahrscheinlichkeit. Leider "erkennt" der Algorithmus auch dann, wenn die TFBS nicht vorliegt, bisweilen eine solche, und zwar mit 1% Wahrscheinlichkeit. Uns interessiert die Wahrscheinlichkeit, dass tatsächlich eine TFBS vorliegt, wenn der Algorithmus eine meldet. Sei A das Ereignis "Algorithmus meldet TFBS" und B das Ereignis "TFBS liegt tatsächlich vor". Mit der Bayes'schen Formel folgt

$$\begin{aligned} P(B | A) &= \frac{P(A | B) \cdot P(B)}{P(A | B) \cdot P(B) + P(A | B^c) \cdot P(B^c)} \\ &= \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.01 \cdot 0.999} \approx 0.09 \end{aligned}$$

Obwohl der Algorithmus recht sensitiv und selektiv zu sein scheint, sind nur ca. 9% aller "erkannten" TFBSn wirkliche TFBSn.

Beispiel 1.8 (Test auf seltene Krankheit.). Wir nehmen zahlenmässig exakt dasselbe Beispiel wie eben, aber ersetzen B durch "Der Patient hat Krötzeritis" (eine fiktive Krankheit, mit der jeder Tausendste infiziert ist) und A durch "Der Krötzeritis-Test testet positiv". Obwohl der Test mit den oben angegebenen Werten relativ zuverlässig zu sein scheint, sind nur ca. 9% der positiv getesteten Patienten tatsächlich infiziert.

Übung 1.4. Wie groß ist die Wahrscheinlichkeit, an Krötzeritis erkrankt zu sein, wenn der Test negativ ausfällt?

1.2.3 Unabhängigkeit

Zwei Ereignisse A, B sind intuitiv dann unabhängig, wenn das Eintreten des einen Ereignisses die Wahrscheinlichkeit des Eintretens des anderen Ereignisses nicht beeinflusst, also wenn $P(A | B) = P(A)$ und $P(B | A) = P(B)$. Um Probleme mit Ereignissen der Wahrscheinlichkeit null zu vermeiden, definiert man Unabhängigkeit wie folgt.

Definition 1.3 (Unabhängige Ereignisse). Zwei Ereignisse A, B heissen *unabhängig (independent)*, wenn $P(A, B) = P(A) \cdot P(B)$.

Eine Familie $(A_i)_{i \in I}$ (I eine beliebige Indexmenge) von Ereignissen heißt unabhängig, wenn für jede endliche Teilmenge $J \subset I$ die Produktformel $P(\cap_{j \in J} A_j) = \prod_{j \in J} P(A_j)$ gilt.

Beispiel 1.9. Sei A ein beliebiges Ereignis. Dann sind A und Ω unabhängig. Ebenso sind A und \emptyset unabhängig. Sei weiter B ein zu A disjunktes Ereignis. Dann sind im allgemeinen A und B abhängig, denn $P(A, B) = 0$ und im allgemeinen $P(A) \cdot P(B) \neq 0$.

Bei mehr als zwei Ereignissen genügt es nicht, dass jedes Paar für sich unabhängig ist (in dem Fall spricht man von *paarweiser Unabhängigkeit (pairwise independence)*); dies ist eine schwächere Eigenschaft als Unabhängigkeit.

Beispiel 1.10. Eine Urne enthält vier Bälle, die von 1 bis 4 durchnummeriert sind. Ein Ball wird zufällig gezogen. Wir betrachten die Ereignisse $A = \{1, 2\}$, $B = \{1, 3\}$, und $C = \{1, 4\}$. Man kann leicht nachrechnen, dass jedes Paar unabhängig ist. Das Tripel (A, B, C) als ganzes ist aber nicht unabhängig, da $1/4 = P(A, B, C) \neq P(A)P(B)P(C) = (1/2)^3 = 1/8$.

Anhängigkeit hat nichts mit Kausalität zu tun. Beim Ziehen ohne Zurücklegen (Beispiel 1.5) ist es genauso richtig zu sagen, “Die Verteilung der Farbe der ersten Kugel hängt von der Farbe der zweiten Kugel ab” wie “Die Verteilung der Farbe der zweiten Kugel hängt von der Farbe der ersten Kugel ab”.

Oft wird Unabhängigkeit als Modellannahme benutzt. Zum Beispiel modellieren wir aufeinanderfolgende Würfelwürfe meistens als unabhängig.

1.3 Zufallsvariablen

Oft ist für uns nicht das genaue Ergebnis eines Wahrscheinlichkeitsexperiments interessant, sondern vielmehr eine Funktion des Ergebnisses, wie zum Beispiel die Summe der Augenzahlen beim Würfeln mit zwei Würfeln.

Definition 1.4 (Zufallsvariable). Sei (Ω, \mathcal{A}, P) ein Wahrscheinlichkeitsraum.

Ist Ω diskret und \mathcal{X} eine beliebige Menge, so heißt die Abbildung $X : \Omega \rightarrow \mathcal{X}$ eine *diskrete Zufallsvariable* (*discrete random variable, r.v.*) mit Werten in \mathcal{X} . Ist $\mathcal{X} = \mathbb{R}$, spricht man von einer (reellen) Zufallsvariablen; ist $\mathcal{X} = \mathbb{R}^n$, von einem *Zufallsvektor* (*random vector*).

Durch eine Zufallsvariable X wird auch \mathcal{X} zu einem Wahrscheinlichkeitsraum, denn X definiert eine neue Verteilung $Q := X(P)$ auf Ereignissen $B \subset \mathcal{X}$ durch $Q(B) := P(A)$, wobei $A := X^{-1}(B)$ das Urbild von B unter X ist.

Ist Ω überabzählbar, muss man noch sogenannte Messbarkeitsanforderungen an X stellen; nur bestimmte Mengen $B \in \mathcal{B} \subset 2^{\mathcal{X}}$ werden messbar sein. Man muss fordern, dass für alle $B \in \mathcal{B}$ das Urbild $X^{-1}(B)$ in Ω messbar ist, also in \mathcal{A} liegt. In unseren Beispielen wird dies stets erfüllt sein, und wir müssen uns darüber keine Gedanken machen.

Beispiel 1.11 (2 Würfel). Sei Ω die Menge aller Ergebnisse beim Würfeln mit 2 Würfeln ($\Omega = \{(a, b) : a = 1, \dots, 6; b = 1, \dots, 6\}$) und P die Gleichverteilung auf Ω . Wir definieren X als die Summe der Augenzahlen: $X(\omega) := a + b$ für $\omega = (a, b) \in \Omega$.

Damit ist der Wertebereich von X gegeben durch $\mathcal{X} = \{2, 3, \dots, 12\}$. Die Verteilung Q ist auf den Elementarereignissen $x \in \mathcal{X}$ gegeben durch $Q(x) = |A|/36$, wobei A die Menge der Paare (a, b) mit $a + b = x$ ist. Konkret ergibt sich:

x	2	3	4	5	6	7	8	9	10	11	12
$Q(\{x\})$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

Beispiel 1.12 (Wartezeit bis zum ersten Erfolg). Wir betrachten potenziell unendliche Folgen von unabhängigen Münzwürfen. Der Wahrscheinlichkeitsraum ist überabzählbar. Als Ereignisse betrachten wir nur solche, die sich aus endlichen Teilfolgen konstruieren lassen. Die Wahrscheinlichkeit, dass bei einem bestimmten Wurf Kopf erscheint, sei p (damit erhält Zahl die Wahrscheinlichkeit $q := 1 - p$). Die Wahrscheinlichkeit, dass die Wurffolge mit (Zahl, Zahl, Kopf) beginnt, ist wegen der Unabhängigkeit q^2p .

Uns interessiert Wartezeit X bis zum ersten Auftreten von Kopf (wir werfen so lange, bis zum ersten Mal Kopf erscheint, und brechen dann ab). Das Ereignis $\{X = n\}$ bedeutet, wir müssen $n - 1$ mal Zahl und dann Kopf beobachten, entspricht also der Menge aller unendlichen Folgen, die mit $n - 1$ mal Zahl und einmal Kopf beginnen und sich dann beliebig fortsetzen. Die Wahrscheinlichkeit ergibt sich zu $Q(n) = q^{n-1}p$.

Obwohl der zugrundeliegende Wahrscheinlichkeitsraum überabzählbar ist, ist der durch X gegebene Raum diskret.

Übung 1.5. Man rechne nach, dass Q tatsächlich eine Verteilung auf $\mathcal{X} = \{1, 2, \dots\}$ definiert, also $\sum_{n=1}^{\infty} Q(n) = 1$ gilt (geometrische Reihe).

Bemerkung. Wenn einmal die Verteilung von X bekannt ist, interessiert man sich oft gar nicht mehr für den zugrundeliegenden Wahrscheinlichkeitsraum und denkt nur noch an die Verteilung Q in \mathcal{X} . Trotzdem hat sich die Notation $P(X = n)$ als leicht lesbare Form für $Q(\{n\})$ oder $P(X^{-1}(\{n\}))$ eingebürgert.

Aus der Definition der Unabhängigkeit von Ereignissen leitet sich die Definition der Unabhängigkeit von Zufallsvariablen ab.

Definition 1.5 (Unabhängigkeit von Zufallsvariablen). Eine Familie $(X_i)_{i \in I}$ von Zufallsvariablen (X_i habe Werte in \mathcal{X}_i) heißt unabhängig, wenn für jede Wahl von Ereignissen $B_i \subset \mathcal{X}_i$ die Ereignisfamilie $(\{X_i \in B_i\})_{i \in I}$ unabhängig ist.

Beispiel 1.13. Sind X und Y unabhängig, so gilt $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ für alle Wahlen von x und y .

1.3.1 Wichtige diskrete Zufallsvariablen

Wir stellen jetzt einige wichtige diskrete Zufallsvariablen und ihre Eigenschaften vor.

Definition 1.6 (Gleichverteilung, Uniform-Verteilung). Sei \mathcal{X} eine endliche Menge mit n Elementen, z.B. $\mathcal{X} = \{1, \dots, n\}$. Die Zufallsvariable X ist *gleichverteilt* oder *uniform verteilt* (*uniformly distributed*) auf \mathcal{X} , wenn $P(X = x) = 1/n$ für alle $x \in \mathcal{X}$ gilt.⁴

Definition 1.7 (Binomialverteilung). Sei $\text{Cal}X$ eine Menge mit n Elementen; sei $p \in [0, 1]$. Man sagt, X hat die *Binomialverteilung* (*binomial distribution*) mit Parametern n und p (auch geschrieben als $X \sim b_{n,p}$), wenn $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ gilt.

Beispiel 1.14. Die Binomialverteilung ergibt sich, wenn X die Anzahl der Erfolge in n Versuchen zählt, wenn die Versuche unabhängig sind und jeder Versuch mit Wahrscheinlichkeit p zum Erfolg führt. Das sieht man so: Jedes Ereignis mit genau k Erfolgen und $n - k$ Misserfolgen hat die Wahrscheinlichkeit $p^k (1-p)^{n-k}$, und es gibt $\binom{n}{k}$ solche Ereignisse.

Beispiel 1.15 (Bernoulli-Verteilung). Im Fall $n = 1$ spricht man auch von der *Bernoulli-Verteilung* mit Parameter p . Es gilt dann $P(X = 1) = p$ und $P(X = 0) = 1 - p =: q$. Seien nun X_1, \dots, X_n unabhängig Bernoulli-verteilt mit Parameter p . Dann zählt $X := X_1 + \dots + X_n$ die Gesamtzahl der Erfolge und ist damit $b_{n,p}$ -verteilt. Merke: Eine Binomial-verteilte Zufallsvariable lässt sich als Summe von unabhängigen Bernoulli-verteilten Zufallsvariablen mit derselben Erfolgswahrscheinlichkeit schreiben.

Definition 1.8 (Geometrische Verteilung). Sei $\mathcal{X} = \{1, 2, \dots\}$. Dann besitzt X die *geometrische Verteilung* (*geometric distribution*) mit Parameter p ($0 < p < 1$), wenn $P(X = n) = p \cdot (1-p)^{n-1}$ gilt⁵.

Lemma 1.2 (Gedächtnislosigkeit). *Ist X geometrisch verteilt mit Parameter p und ist $q := 1 - p$, so gilt*

$$P(X \geq n) = \sum_{k=n}^{\infty} P(X = k) = q^{n-1},$$

$$P(X \geq n+k \mid X \geq n) = q^k \quad \text{unabhängig von } n. \quad (1.13)$$

Eigenschaft (1.13) wird auch Gedächtnislosigkeit (memoryless property) genannt; sie charakterisiert die geometrische Verteilung mit Parameter $p = 1 - q$.

⁴Man beachte den Unterschied zwischen gleichverteilt und gleich verteilt (identically distributed). Eine Zufallsvariable X kann auf $\text{Cal}X$ gleichverteilt sein; zwei Zufallsvariablen X und Y können gleich verteilt sein (sie haben sie gleiche Verteilung).

⁵Bisweilen wird auch X auf $\{0, 1, 2, \dots\}$ mit $P(X = n) = p \cdot (1-p)^n$ geometrisch verteilt genannt.

Ein wichtiges Beispiel für die geometrische Verteilung haben wir bereits in Beispiel 1.12 kennengelernt: Die Wartezeit auf den ersten Erfolg. Die Gedächtnislosigkeit sagt in dem Fall: Auch wenn man schon vergeblich n -mal auf Erfolg gewartet hat, so ist die Wahrscheinlichkeit, dass sich in den nächsten k Schritten Erfolg einstellt, genauso groß wie zu Beginn; die n Misserfolge sind “vergessen”.

Bild.

Definition 1.9 (Poisson-Verteilung). Sei $\mathcal{X} = \{0, 1, 2, \dots\}$. Dann besitzt X die *Poisson-Verteilung (Poisson distribution)* mit Parameter $\lambda > 0$, wenn $P(X = k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$ gilt. Die Poisson-Verteilung heißt auch *Verteilung der seltenen Ereignisse (law of rare events)*, weil die Anzahl des Eintretens seltener Ereignisse sich oft sehr gut durch eine Poisson-Verteilung approximieren lässt.

Beispiel 1.16 (Poisson-Approximation der Binomialverteilung). Es sei $X \sim b_{n,p}$ mit n groß und p klein (formal fixieren wir ein $\lambda > 0$, setzen $P := \lambda/n$, und betrachten den Grenzübergang $n \rightarrow \infty$. Dann gilt im Limes $X \sim \mathcal{P}_\lambda$; siehe Abbildung ??).

1.3.2 Stetige Zufallsvariablen

Uniform; Exponential; Gamma; Normal.

1.3.3 Momente

Erwartungswert; Varianz

1.3.4 Wichtige Beispiele

1.4 Markovketten

Chapman-Kolmogorov. Zeitreversibilität. MCMC.

1.5 Markovprozesse

Kapitel 2

Grundlagen der Statistik

Anhang A

IUPAC-Symbole

A.1 IUPAC Nukleotid-Symbole

Die *International Union of Pure and Applied Chemistry (IUPAC)*, www.iupac.org, und das *Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB)* haben folgende Symbole für Nukleotidsequenzen vorgeschlagen [1], die inzwischen auch zum Quasi-Standard avanciert sind.

Symbol	Nukleotide	Beschreibung	Komplement
A	A	Adenin	T
C	C	Cytosin	G
G	G	Guanin	C
T/U	T/U	Thymin / Uracil	A
R	A, G	Purin	Y
Y	C, T/U	Pyrimidin	R
M	A, C	Amino	K
K	G, T/U	Keto	M
S	C, G	Strong, 3 H-Bindungen	S
W	A, T	Weak, 2 H-Bindungen	W
B	C, G, T/U	nicht A	V
D	A, G, T/U	nicht C	H
H	A, C, T/U	nicht G	D
V	A, C, G	nicht T/U	B
N	A, C, G, T/U	any	N

Mit diesen 15 Symbolen lässt sich jede nichtleere Teilmenge des DNA-Alphabets {A,C,G,T} beschreiben. Mit Komplement ist nicht das mengentheoretische Komplement, sondern das Watson-Crick-Komplement gemeint. In der Sequenzanalyse findet man auch das Symbol X, das manchmal dieselbe Bedeutung wie N hat, oft aber für eine maskierte Sequenz steht, also für die leere Nukleotidmenge. Bei Mustererkennungsverfahren soll X auf keines der Nukleotide passen.

Die Einführung der mehrdeutigen Symbole R, Y, ..., N erweist sich bei vielen Anwendungen als nützlich, z.B. zur Beschreibung der Erkennungssequenzen für Restriktionsenzyme, zur kompakten Beschreibung genetischer Codes, für Konsensussequenzen, oder einfach, um Unsicherheit bei Sequenzdaten auszudrücken.

Beispiel A.1. Das Enzym *AvaI* erkennt die vier Sequenzen 5'-CCCGGG-3', 5'-CCCGAG-3', 5'-CTCGGG-3', und 5'-CTCGAG-3'. Diese lassen sich als 5'-CYCGRC-3' zusammenfassen.

Die hier vorgestellten Symbole erlauben es, Sequenzmengen kompakt darzustellen. Ihr Zweck ist *nicht*, zwischen DNA und RNA zu trennen (auch RNA-Sequenzen werden formal bisweilen mit T statt U geschrieben), zwischen Basen, Nukleotiden und Nukleosiden zu unterscheiden, oder Modifikationen wie Methylierung anzuzeigen.

A.2 IUPAC Aminosäure-Symbole

Analog zu den vier Nukleotidsymbolen gibt es 20 Symbole für die Aminosäuren, um Proteinsequenzen zu beschreiben. Hier ist es nicht sinnvoll, einzelne Symbole für alle 2^{20} (ca. 1 Million) Teilmengen einzuführen. Nur drei solcher Symbole (B,Z, und X) sind definiert.

Aminosäure	Symbol	3er-Code	Codon(s) (5'→3')
Glycin	G	Gly	GGN
Alanin	A	Ala	GCN
Valin	V	Val	GTN
Leucin	L	Leu	CTN und TTR
Isoleucin	I	Iso	ATH
Prolin	P	Pro	CCN
Phenylalanin	F	Phe	TTY
Tyrosin	Y	Tyr	TAY
Cystein	C	Cys	TGY
Methionin	M	Met	ATG
Histidin	H	His	CAY
Lysin	K	Lys	AAR
Arginin	R	Arg	CGN und AGR
Tryptophan	W	Trp	TGG
Serin	S	Ser	TCN und AGY
Threonin	T	Thr	ACN
Asparaginsäure	D	Asp	GAY
Glutaminsäure	E	Glu	GAR
Asparagin	N	Asn	AAY
Glutamin	Q	Gln	CAR
STOP	.		TAR und TGA
Asparagin(säure)	B		RAY
Glutamin(säure)	Z		SAR
unbekannt / alle	X		NNN

Die Codon-Angaben beziehen sich auf den genetischen Standard-Code (Wirbeltiere).
Bei anderen Organismen kann es Abweichungen geben.

Übung A.1. Warum lässt sich das Codon für Leucin (CTN und TTR) nicht als YTN schreiben (analog: Arginin (CGN und AGR) als MGN, Serin (TCN und AGY) als WSN, das STOP-Codon (TAR und TGA) als TRR)? Wo gibt es jeweils Konflikte?

Anhang B

Glossar

Base (*base*).

Nukleotid (*nucleotide*).

Nukleosid (*nucleoside*).

Literaturverzeichnis

- [1] International Union of Pure and Applied Chemistry and International Union of Biochemistry and Molecular Biology (IUPAC-IUBMB) Joint Commission on Biochemical Nomenclature and Nomenclature Commission of IUBMB. *Biochemical Nomenclature and Related Documents*, pages 122–126. Portland Press, 2nd edition, 1992.