

# Stochastik für Studierende der Informatik

VON PETER PFAFFELHUBER

Version: 12. Juli 2016

## Inhaltsverzeichnis

<b>1</b>	<b>Grundlegendes</b>	<b>3</b>
1.1	Häufige Modelle: Münzwurf, Würfeln, Urnen . . . . .	3
1.2	Kombinatorische Formeln . . . . .	5
1.3	Rekursionen . . . . .	7
1.4	Wahrscheinlichkeitsräume und Zufallsvariablen . . . . .	9
1.5	Gemeinsame Verteilung, Bildverteilung, Unabhängigkeit . . . . .	12
<b>2</b>	<b>Verteilungen und deren Eigenschaften</b>	<b>16</b>
2.1	Grundlegendes . . . . .	16
2.2	Uniforme Verteilungen, Laplace-Experimente . . . . .	17
2.3	Die Binomial-Verteilung . . . . .	18
2.4	Die Normal-Verteilung . . . . .	19
<b>3</b>	<b>Kenngößen von Zufallsvariablen</b>	<b>19</b>
3.1	Der Erwartungswert . . . . .	19
3.2	Die Varianz und Kovarianz . . . . .	23
3.3	Die Normal-Verteilung . . . . .	28
3.4	Erzeugung von Zufallszahlen . . . . .	30
<b>4</b>	<b>Approximationssätze</b>	<b>31</b>
4.1	Das Gesetz der großen Zahlen . . . . .	32
4.2	Der zentrale Grenzwertsatz und die Normal-Verteilung . . . . .	33
<b>5</b>	<b>Bedingte Wahrscheinlichkeiten</b>	<b>35</b>
5.1	Grundlegendes . . . . .	35
5.2	Ein Ausflug in die Bayes'sche Statistik . . . . .	40
5.3	Markov-Ketten . . . . .	40
5.4	Der Kalman-Filter . . . . .	45
<b>6</b>	<b>Statistik</b>	<b>48</b>
6.1	Grundlagen . . . . .	48
6.2	Schätzprobleme . . . . .	50
6.3	Testprobleme . . . . .	54



## 1 Grundlegendes

Die Stochastik beschäftigt sich mit zufälligen Ereignissen. Zunächst mag es erstaunen, dass man Regeln im Zufall finden kann. Auch wenn man nicht genau weiß, *welches* zufällige Ereignis passiert, kann man häufig angeben, *mit welcher Wahrscheinlichkeit* es passiert. Zentral ist hier der Begriff der *Wahrscheinlichkeitsverteilung*, die angibt, mit welchen (deterministischen) Wahrscheinlichkeiten zufällige Ereignisse passieren. Wir beschäftigen uns nicht damit, wie man diese Wahrscheinlichkeiten festlegt, sondern werden vor allem mit ihnen rechnen. Ein weiterer zentraler Begriff ist die *Zufallsvariable*, die ein zufälliges Element einer Menge beschreibt. In diesem ersten Kapitel werden wir neben der Einführung dieser Begriffe ein paar klassische Modelle kennen lernen, etwa das Werfen von Münzen. Dies wird bei der Behandlung von konkreten (Wahrscheinlichkeits-)Verteilungen sehr hilfreich sein.

In der Informatik werden stochastische Methoden in vielerlei Hinsicht benötigt. Wir besprechen etwa das grundlegende Gebiet der Kombinatorik, das auch für Datenstrukturen und Komplexitätsanalysen wichtig ist. Weiter gibt es etwa die average-case-Analyse von Algorithmen, bei der immer von einem zufälligen Input ausgegangen wird. Eine verwandte Frage ist die Analyse zufälliger Algorithmen, also Methoden, bei denen zufällige Entscheidungen getroffen werden (obwohl der Input deterministisch ist). Letztlich spielen statistische Fragestellungen, gerade wenn sie rechenintensiv sind, eine größer werdende Rolle in der Informatik. Hier verweisen wir auf das sich stetig fortentwickelnde Gebiet des Maschinellen Lernens.

### 1.1 Häufige Modelle: Münzwurf, Würfeln, Urnen

Bevor wir formal einführen, was Wahrscheinlichkeiten und Zufallsvariablen sind, behandeln wir häufige Beispiele.

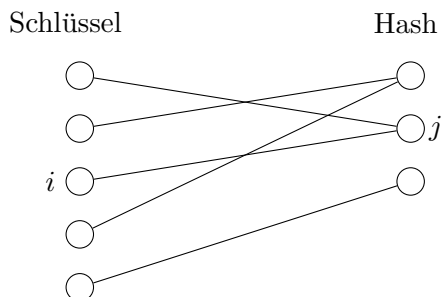
**Beispiel 1.1 (Münzwurf).** Wenn man eine Münze wirft, zeigt sie bekanntlich *Kopf* oder *Zahl*. Bei einem  $p$ -Münzwurf ist die Wahrscheinlichkeit für *Kopf* gerade  $p$ . Damit ist die Wahrscheinlichkeit für *Zahl* gerade  $q := 1 - p$ . Bei üblichen Münzen denkt man oft an den Fall  $p = 1/2$ , weil Kopf und Zahl mit derselben Wahrscheinlichkeit oben zu liegen kommen. Denkt man jedoch, dass man eine Reißzwecke wirft, die entweder mit der spitzen Seite oben oder unten zu liegen kommt, ist klar, dass auch  $p \neq 1/2$  sein kann. Wirft man zweimal mit derselben Münze, kann man nach der Wahrscheinlichkeit fragen, zweimal *Kopf* zu werfen. Dies ist  $p^2$ .

Etwas informatischer betrachten wir eine Folge zufälliger Bits. Will man etwa eine zufällige (natürliche) Zahl zwischen 0 und 31 erhalten, ist diese durch ihre Binärdarstellung durch einen fünf-fachen Münzwurf mit  $p = 1/2$  gegeben.

**Beispiel 1.2 (Würfeln).** Eine bekannte Herausforderung ist es (etwa bei *Mensch ärgere Dich nicht*), beim Würfeln mit einem Würfel möglichst schnell eine 6 zu werfen. Dies ist ganz ähnlich wie beim  $(1/6)$ -Münzwurf, schließlich gibt es nur zwei Möglichkeiten *6* oder *nicht 6*. Damit ist die Wahrscheinlichkeit,  $k$  mal keine 6 zu werfen, gerade  $(5/6)^k$ . Wir berechnen noch die Wahrscheinlichkeit, mit zwei Würfeln ein *Mädchen* zu werfen, also gerade die Kombination 1,2. Diese ist

$$2 \cdot \frac{1}{6} \cdot \frac{1}{6}.$$

Denn: Entweder der erste Würfel zeigt eine 1 (Wahrscheinlichkeit  $1/6$ ) und der zweite Würfel zeigt eine 2 (Wahrscheinlichkeit auch  $1/6$ ), oder andersherum (deswegen der Faktor 2).



**Abbildung 1.1:** Eine Hash-Funktion weist einer Menge von Schlüsseln (meist nicht injektiv) die Hashes zu.

Um eine Anwendung eines solchen (verallgemeinerten) Würfel-Experimentes zu geben, betrachten wir eine Hash-Funktion; siehe auch Abbildung 1.1. Diese weist bekanntlich einer Eingabemenge (oder Schlüsseln) die Hashwerte zu, und zwar meist nicht injektiv (so dass mehrere Schlüssel denselben Hash bekommen können). Die Eingabe sei so, dass jeder Schlüssel mit Wahrscheinlichkeit  $p_i$  den Hash  $i$  zugeordnet bekommt,  $i \in I$ . (Hier ist  $I$  die Menge der Hashes.) Die Wahrscheinlichkeit, dass zwei Schlüssel demselben Hash  $i$  zugeordnet werden, ist  $p_i^2$ .

**Beispiel 1.3 (Urne).** In einer Urne befinden sich  $N$  Kugeln, und zwar  $K_1$  Kugeln der Farbe 1, ...,  $K_n$  Kugeln der Farbe  $n$ . (Also ist  $K_1 + \dots + K_n = N$ .) Zieht man (mit verschlossenen Augen) aus der Urne, zieht man eine Kugel der Farbe  $i$  mit Wahrscheinlichkeit  $K_i/N$ . Zieht man anschließend nochmal aus der Urne (wobei die erste Kugel nicht zurückgelegt wurde), ist die Wahrscheinlichkeit nochmal eine Kugel der Farbe  $i$  zu ziehen  $(K_i - 1)/(N - 1)$ . Damit ist die Wahrscheinlichkeit, dass man zwei Kugeln derselben Farbe zu ziehen, gerade

$$\sum_{i=1}^n \frac{K_i}{N} \frac{K_i - 1}{N - 1}.$$

Hat man etwa  $N$  Aufgaben, die von einem Prozessor (in zufälliger Reihenfolge) bearbeitet werden sollen, und zwar  $K_1$  Aufgaben vom Typ 1 (etwa Sortieren), ...,  $K_n$  Aufgaben vom Typ  $n$ , so ist dies auch die Wahrscheinlichkeit, dass zunächst zwei Aufgaben vom selben Typ bearbeitet werden.

**Beispiel 1.4 (Geburtstagsproblem).** Es werden  $k$  Schlüssel in eine Liste mit  $n$  Hashes geschrieben. Hierbei soll die Wahrscheinlichkeit, dass Schlüssel  $i$  dem Hash  $j$  zugeordnet wird, gerade  $1/n$  sein. Wie groß ist die Wahrscheinlichkeit, dass es zu keiner Kollision kommt, d.h. dass alle Schlüssel verschiedenen Hashes zugeordnet werden?<sup>1</sup>

Wir gehen der Reihe nach die Schlüssel durch. Die Wahrscheinlichkeit, dass der zweite Schlüssel einen anderen Hash als der erste Schlüssel bekommt, ist  $(n - 1)/n$ . Die Wahrscheinlichkeit, dass der dritte Schlüssel einen Hash bekommt, der verschieden ist vom Hash der ersten beiden Schlüssel, ist  $(n - 2)/n$  etc. Insgesamt ergibt sich also für die gesuchte Wahrscheinlichkeit

$$\frac{n - 1}{n} \cdot \frac{n - 2}{n} \cdots \frac{n - k + 1}{n}.$$

<sup>1</sup>In der Stochastik ist dies das Geburtstagsproblem: Wie groß ist die Wahrscheinlichkeit, dass  $k$  Personen alle an unterschiedlichen Tagen Geburtstag haben? (Hierbei ist dann  $n = 365$ .)

Als Beispiel nehmen wir  $k = 10$  Schlüssel, die auf  $n = 100$  Hashes verteilt werden. (Dies ist natürlich untypisch, da normalerweise  $n < k$  Hashes verwendet werden.) Dann ergibt sich für die Wahrscheinlichkeit keiner Kollision

$$\frac{99 \cdot \dots \cdot 91}{100^9} \approx 0.629.$$

## 1.2 Kombinatorische Formeln

In allen obigen Beispielen wird klar, dass man oftmals etwas abzählen muss, um Wahrscheinlichkeiten zu berechnen. Diese Kunst wird auch als Kombinatorik bezeichnet. Wir behandeln nun ein paar ausgewählte Fälle, eine Zusammenfassung findet man in Tabelle 6.1. Anschließend geben wir Beispiele.

**Lemma 1.5 (Kombinatorik ohne Wiederholungen).** Sei  $\underline{x} = (x_1, \dots, x_n)$  ein Vektor der Länge  $n$ , deren Einträge alle verschieden sind.

1. Es gibt  $n! := n \cdot (n-1) \cdot \dots \cdot 1$  verschiedene Vektoren der Länge  $n$ , die dieselben Einträge wie  $\underline{x}$  haben.
2. Es gibt  $n \cdot (n-1) \cdot \dots \cdot (n-k+1)$  verschiedene Vektoren der Länge  $k$ , die aus den Einträgen von  $\underline{x}$  gewonnen werden können.
3. Es gibt  $\binom{n}{k} := \frac{n!}{k!(n-k)!}$  verschiedene,  $k$ -elementige Teilmengen von  $\{x_1, \dots, x_n\}$ .

*Beweis.* 1. Stellt man einen Vektor zusammen, so hat der erste Eintrag genau  $n$  Möglichkeiten. Für jede dieser Möglichkeiten hat der zweite Eintrag dann  $n-1$  Möglichkeiten usw. Daraus ergibt sich die Formel.

2. folgt genau wie 1. nur dass man nach dem  $k$ -ten Eintrag abbrechen muss.

3. Geht man von 2. aus, so gibt es  $\frac{n!}{(n-k)!}$  verschiedene Möglichkeiten, Vektoren der Länge  $k$  aus den Einträgen von  $\underline{x}$  zu bekommen. Von diesen Vektoren enthalten jeweils  $k!$  dieselben Elemente. Will man also die Menge der  $k$ -elementigen Teilmengen zählen, so liefert jede dieser Teilmengen genau  $k!$  Vektoren, also folgt das Ergebnis.  $\square$

**Lemma 1.6 (Kombinatorik mit Wiederholungen).** Gegeben seien  $n$  verschiedene Objekte,  $x_1, \dots, x_n$ .

1. Sind von den  $n$  Objekten genau  $r$  verschieden, und kommt das  $i$ -te Objekt genau  $k_i$  mal vor,  $i = 1, \dots, r$ , ist  $k_1 + \dots + k_r = n$  und es gibt genau  $\frac{n!}{k_1! \dots k_r!}$  verschiedene Möglichkeiten, die Objekte in eine Reihenfolge zu bringen.
2. Zieht man aus den  $n$  Objekten genau  $k$ -mal mit Zurücklegen, so gibt es genau  $n^k$  mögliche Ausgänge, wenn man die Reihenfolge der gezogenen Objekte beachtet.
3. Zieht man aus den  $n$  Objekten genau  $k$ -mal mit Zurücklegen, so gibt es genau  $\binom{n+k-1}{k}$  mögliche Ausgänge, wenn man die Reihenfolge der gezogenen Objekte nicht beachtet.

*Beweis.* 1. Könnte man alle Objekte unterscheiden, so gäbe es  $n!$  Möglichkeiten. Da man jedoch von dieser Anzahl jeweils  $k_1!, \dots, k_r!$  für die möglichen Reihenfolgen der identischen Objekte zusammenfassen kann, folgt das Ergebnis.

2. ist klar.

	Permutation	Variation	Kombination
ohne Wiederholung  Beispiel	$n!$  Anordnung von $n$ Zahlen, wobei jede Zahl nur einmal vorkommen darf	$\frac{n!}{(n-k)!}$  Ziehung von $k$ Zahlen aus $n$ Möglichen, wobei jede Zahl <i>nur einmal</i> vorkommen darf, <i>mit</i> Beachtung der Reihenfolge der Ziehung	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$  Ziehung von $k$ Zahlen aus $n$ Möglichen, wobei jede Zahl <i>nur einmal</i> vorkommen darf, <i>ohne</i> Beachtung der Reihenfolge der Ziehung
mit Wiederholung  Beispiel	$\frac{n!}{k_1! \cdot \dots \cdot k_r!}$  Anordnung von $n$ Zahlen, die nicht alle verschieden sind	$n^k$  Ziehung von $k$ Zahlen aus $n$ Möglichen, wobei jede Zahl <i>beliebig oft</i> vorkommen darf, <i>mit</i> Beachtung der Reihenfolge der Ziehung	$\binom{n+k-1}{k}$  Ziehung von $k$ Zahlen aus $n$ Möglichen, wobei jede Zahl <i>beliebig oft</i> vorkommen darf, <i>ohne</i> Beachtung der Reihenfolge der Ziehung

**Tabelle 1.1:** Kombinatorik beschäftigt sich mit dem Abzählen von Möglichkeiten. Generell unterscheidet man solche mit und ohne Wiederholungen, und mit (Permutationen und Variationen) und ohne (Kombinationen) Beachtung der Reihenfolge.

3. Hier muss man etwas genauer nachdenken, wobei folgende Überlegung entscheidend ist: ist etwa  $n = 5, k = 3$ , so lässt sich jede gesuchte Möglichkeit durch einen Code der Art  $\bullet\bullet**\bullet**$  darstellen. Dies bedeutet, dass vom ersten Objekt zwei Kopien gewählt wurden (deshalb  $\bullet\bullet$ ), vom zweiten Objekt allerdings gar keines (deshalb ist zwischen den  $*$ 's kein  $\bullet$ ), vom dritten eines, und vom vierten und fünften wiederum keines. Dieser Code ist also eine eindeutige Zuordnung der gesuchten Anordnungen auf die Menge der Vektoren der Länge  $n + k - 1$  mit  $k$  Einträgen  $\bullet$  und  $n - 1$  Einträgen  $*$ . (Man beachte, dass  $*$  ja hier ein Trennzeichen ist, und man  $n - 1$  Trennzeichen benötigt um  $n$  Felder zu trennen.) Die Anzahl dieser Vektoren ist gerade die gesuchte Anzahl, weil man nun die  $k$ -elementigen Teilmengen aller  $n + k - 1$  Einträge sucht. □

**Bemerkung 1.7 (Permutation).** Wir bezeichnen jede bijektive Abbildung einer Menge  $\mathcal{S}$  als Permutation von  $\mathcal{S}$ . Lemma 1.5.1 besagt, dass die Menge der Bijektionen von  $\{1, \dots, n\}$  gerade  $n!$  ist.

**Bemerkung 1.8 (Wie groß ist eigentlich  $n!$ ).** Um etwa die Anzahl der möglichen Listen anzugeben, die ein Sortieralgorithmus vorfinden kann (d.h. die Anzahl der möglichen Permutationen von  $n$  Objekten) oder die Komplexität von Algorithmen abschätzen zu können,

benötigt man eine Vorstellung, wie groß  $n!$  ist. Hierzu ist die Stirling-Formel hilfreich. Diese besagt, dass

$$\lim_{n \rightarrow \infty} \frac{n!}{\left(\frac{n}{e}\right)^n \sqrt{2\pi n}} = 1.$$

Wir werden dies nicht beweisen, aber immerhin einsehen (was für Komplexitätsaussagen meist genügt), dass  $n! \approx (n/e)^n$ . Wir schreiben

$$\begin{aligned} \log(n!) &= \sum_{i=1}^n \log(i) = \int_1^n \log(x) dx + O(n) \\ &= x \log x - x \Big|_{x=1}^n + O(\log(n)) = n \log(n) - n + O(\log(n)). \end{aligned}$$

Daraus erhalten wir

$$n! = e^{\log(n!)} = e^{n \log(n) - n} O(n) = O(n) \left(\frac{n}{e}\right)^n.$$

**Beispiel 1.9 (Passwörter).** Es soll ein Passwort mit acht Zeichen aus den Zeichen [a-z, A-Z, 0-9] generiert werden. Dafür hat man für jede Stelle insgesamt  $26 + 26 + 10 = 62$  Möglichkeiten, insgesamt also  $62^8 \approx 2.18 \cdot 10^{14}$  Möglichkeiten.

**Beispiel 1.10 (Schnitt zweier Listen).** Sei  $A$  eine Liste von  $k$  und  $B$  eine Liste von  $n$  Elementen. Wieviele Möglichkeiten gibt es, die Elemente der Liste  $B$  in die Liste  $A$  einzuordnen. Die Elemente beider Listen seien dabei nicht unterscheidbar.

Jede dieser Möglichkeiten lässt sich schreiben als Vektor, etwa  $AAABBABAAABBB \dots$ , der Länge  $n + k$ , wobei  $A$  bedeutet, dass ein Element der Liste  $A$  und  $B$ , dass ein Element der Liste  $B$  vorkommt. Da es insgesamt  $\binom{n+k}{k}$  solcher Vektoren gibt, ist dies auch die gesuchte Anzahl.

### 1.3 Rekursionen

Oftmals lassen sich Abzählaufgaben durch eine Rekursion lösen. Wir geben hier nur ein Beispiel an.

**Beispiel 1.11 (Anzahl binärer Suchbäume).** Ein binärer Suchbaum (Binary Search Tree, BST) ist eine bekannte Datenstruktur von geordneten Daten, in der man vorhandene Elemente leicht wiederfindet und auch neue einsortieren kann. Es handelt sich um einen binären Baum, bei dem die Wurzel Grad 2 hat, und alle anderen Knoten entweder Grad 3 (interne Knoten) oder Grad 1 (externe Knoten oder Blätter) haben. Jeder interne Knoten hat ein linkes und ein rechtes Kind. Die Einträge sind so geordnet, dass das linke Kind ein kleineres und das rechte ein größeres Label bekommt. Die Blätter sind Platzhalter für neue Elemente.

Wir wollen nun die Anzahl binärer Suchbäume mit einer vorgegebenen Anzahl von Blättern zählen, wobei wir die Einträge nicht betrachten wollen. Dabei gehen wir ähnlich wie bei Divide-and-Conquer-Algorithmen rekursiv vor. Das bedeutet: Jeder BST mit  $n$  Blättern (und damit  $n - 1$  internen Knoten) enthält einen linken und einen rechten Teilbaum. Hat der linke Teilbaum  $i$  Blätter, so hat der rechte Teilbaum  $n - i$  Blätter. Ist  $x_n$  die Anzahl der BSTe mit  $n$  Blättern, so gilt also

$$x_1 = x_2 = 1 \text{ und } x_n = \sum_{i=1}^{n-1} x_i x_{n-i}. \quad (1.1)$$

Wir zeigen nun, dass

$$x_{n+1} = \frac{1}{n+1} \binom{2n}{n} = \frac{(2n)!}{n!(n+1)!}. \quad (1.2)$$

Mit anderen Worten handelt es sich bei der gesuchten Anzahl von BSTs um die Catalan-Zahlen. Wir betrachten die Funktion

$$f(z) := \sum_{n=0}^{\infty} x_{n+1} z^n = \sum_{n=1}^{\infty} x_n z^{n-1},$$

die auch Erzeugendenfunktion der Zahlenfolge  $x_1, x_2, \dots$  genannt wird. Es gilt nämlich wegen der Rekursion (1.1)

$$\begin{aligned} f(z) &= 1 + \sum_{n=1}^{\infty} \sum_{i=1}^n x_i x_{n+1-i} z^n = 1 + z \sum_{i=1}^{\infty} \sum_{n=i}^{\infty} x_i z^{i-1} x_{n+1-i} z^{n-i} \\ &= 1 + z \sum_{i=1}^{\infty} x_i z^{i-1} \sum_{n=0}^{\infty} x_{n+1} z^n = 1 + z f(z)^2, \end{aligned}$$

also  $z f(z)^2 - f(z) + 1 = 0$  und damit (wegen  $f(z) \xrightarrow{z \rightarrow 0} 1$ )

$$f(z) = \frac{1 - \sqrt{1 - 4z}}{2z}.$$

Mit der Reihendarstellung<sup>2</sup> der Funktion  $y \mapsto \sqrt{1+y}$  ergibt sich

$$f(z) = \frac{1}{2z} 2 \sum_{n=0}^{\infty} \frac{(2n)!}{n!(n+1)!} z^{n+1} = \sum_{n=0}^{\infty} \frac{(2n)!}{n!(n+1)!} z^n.$$

Vergleicht man nun die letzte Zeile mit der Definition von  $f$  ergibt sich (1.2).

---

<sup>2</sup>Hier erinnern wir an die Mathematik-Vorlesung. Dort wurde die Taylor-Reihe einer beliebig oft differenzierbaren Funktion  $y \mapsto f(y)$  besprochen. Diese ist im Entwicklungspunkt  $y_0 = 0$  gegeben durch

$$T_{\infty} f(y) := \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} y^n,$$

wobei  $f^{(n)}(0)$  die  $n$ -te Ableitung von  $f$  in 0 bedeutet (und  $f^{(0)} := f$ ). Die Funktion  $g : y \mapsto \sqrt{1+y}$  etwa lässt sich durch  $T_{\infty} g$  approximieren und es ergibt sich

$$\begin{aligned} \sqrt{1+y} &= 1 + \frac{1}{2}y - \frac{1}{2^2 \cdot 2!}y^2 + \frac{3 \cdot 1}{2^3 \cdot 3!}y^3 - \frac{5 \cdot 3 \cdot 1}{2^4 \cdot 4!}y^4 \dots \\ &= 1 - \binom{-\frac{y}{2}}{1} - \frac{1}{2!} \binom{-\frac{y}{2}}{2} - \frac{3 \cdot 1}{3!} \binom{-\frac{y}{2}}{3} - \frac{5 \cdot 3 \cdot 1}{4!} \binom{-\frac{y}{2}}{4} \dots \\ &= 1 - \frac{0!}{2^0 0! 1!} \binom{-\frac{y}{2}}{1} - \frac{2!}{2^1 1! 2!} \binom{-\frac{y}{2}}{2} - \frac{4!}{2^2 2! 3!} \binom{-\frac{y}{2}}{3} - \frac{6!}{2^3 3! 4!} \binom{-\frac{y}{2}}{4} \dots \\ &= 1 - 2 \sum_{n=0}^{\infty} \frac{(2n)!}{n!(n+1)!} \left(-\frac{y}{4}\right)^{n+1}. \end{aligned}$$

Ist



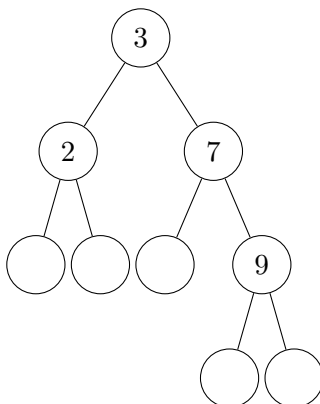


Abbildung 1.2: Ein binärer Suchbaum mit fünf Blättern.

## 1.4 Wahrscheinlichkeitsräume und Zufallsvariablen

**Definition 1.12 (Zufallsvariable).** Sei  $E$  eine Menge. Ein zufälliges Element von  $E$  heißt ( $E$ -wertige) Zufallsvariable.

**Bemerkung 1.13 (Zufälligkeit).** 1. Eine Illustration einer Zufallsvariablen findet sich in Abbildung 1.3.

2. In obiger Definition ist zunächst unklar, was genau *zufällig* bedeuten soll. Dies wird nun durch den Begriff der *Verteilung* der Zufallsvariable behoben. Wir unterscheiden hierbei zwei Fälle: Entweder ist  $E$  höchstens abzählbar, oder Teilmenge der reellen Zahlen.

**Definition 1.14 (Verteilung einer Zufallsvariablen).** Sei  $X$  eine  $E$ -wertige Zufallsvariable.

1. Angenommen,  $E$  ist höchstens abzählbar. Dann heißt  $X$  diskrete Zufallsvariable und die Funktion  $x \mapsto \mathbf{P}[X = x]$  die Verteilung von  $X$ . Hier ist  $\mathbf{P}[X = x]$  die Wahrscheinlichkeit, dass  $X$  den Wert  $x$  annimmt. Diese Abbildung heißt auch Zähldichte.
2. Angenommen,  $E \subseteq \mathbb{R}$ . Gilt für jedes Intervall  $A \subseteq \mathbb{R}$ , dass

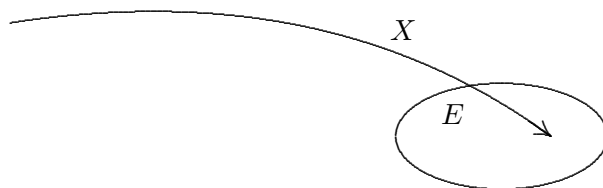
$$\mathbf{P}[X \in A] = \int_A f(x) dx,$$

so heißt  $f$  Dichte der Verteilung von  $X$  und  $X$  eine stetige Zufallsvariable. Wir schreiben dann auch

$$\mathbf{P}[X \in dx] = f(x) dx.$$

3. Wir bezeichnen mit  $\mathcal{P}(E)$  die Potenzmenge von  $E$ . Sei  $\mathbf{P} : \mathcal{A} \subseteq \mathcal{P}(E) \rightarrow [0, 1]$ . Gilt  $\mathbf{P}[E] = 1$ ,  $\mathbf{P}[A^c] = 1 - \mathbf{P}[A]$  und  $\mathbf{P}[\bigcup_{n \in \mathbb{N}} A_n] = \sum_{n=1}^{\infty} \mathbf{P}[A_n]$  für paarweise disjunkte  $A_1, A_2, \dots$  (falls alle Werte definiert sind), so heißt  $\mathbf{P}$  eine Wahrscheinlichkeits-Verteilung auf  $E$ . Sind  $\mathbf{P}$  und  $\mathbf{Q}$  Wahrscheinlichkeits-Verteilungen und ist  $X$  eine Zufallsvariable mit  $\mathbf{P}[X \in A] = \mathbf{Q}[A]$ , so schreiben wir auch  $X \sim \mathbf{Q}$  oder  $X \sim_{\mathbf{P}} \mathbf{Q}$ .

**Bemerkung 1.15 (Zusammenhang (Zähl-)Dichten und Wahrscheinlichkeitsverteilungen).** Offenbar definieren sowohl Zähldichten von diskreten Zufallsvariablen als auch Dichten von stetigen Zufallsvariablen Wahrscheinlichkeitsverteilungen.



**Abbildung 1.3:** Eine Zufallsvariable ist ein zufälliges Element einer Menge.

**Beispiel 1.16 (Münzwurf).** Bei einem zweimaligen  $p$ -Münzwurf sei  $X = (X_1, X_2)$  der zufällige, {Kopf, Zahl}-wertige Vektor, der den Ausgang des Wurfes beschreibt. Dann gilt etwa

$$\mathbf{P}[X = (\text{Kopf}, \text{Kopf})] = \mathbf{P}[X_1 = \text{Kopf}, X_2 = \text{Kopf}] = p^2.$$

**Beispiel 1.17 (Normal-Verteilung).** Vor allem in statistischen Anwendungen spielt die Normal-Verteilung eine große Rolle. Eine Zufallsvariable  $X$  heißt normalverteilt mit Erwartungswert  $\mu \in \mathbb{R}$  und Varianz  $\sigma^2 \in \mathbb{R}_+$ , falls die Verteilung von  $X$  die Dichte

$$f(x) := f_{\mu, \sigma^2}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2}\right) \quad (1.3)$$

besitzt.<sup>3</sup> Wir schreiben dann  $X \sim N(\mu, \sigma^2)$ . Ist  $\mu = 0$  und  $\sigma^2 = 1$ , so heißt die Verteilung auch Standard-Normal-Verteilung. Es gibt auch mehrdimensionale Normal-Verteilungen, die wir später behandeln werden.

**Bemerkung 1.18 (Mathematische Definition von Verteilungen).** In der Mathematik definiert man Zufallsvariablen und Verteilungen etwas anders. Während oben der Begriff der Zufallsvariable im Vordergrund steht, führt man in mathematischen Vorlesungen zur Stochastik zunächst Wahrscheinlichkeitsverteilungen ein. Für eine Menge  $\Omega$  sind dies Abbildungen  $\mathbf{P} : \mathcal{A} \rightarrow [0, 1]$ , wobei  $\mathcal{A} \subseteq \mathcal{P}[\Omega]$  eine  $\sigma$ -Algebra, also ein Mengensystem mit  $\emptyset \in \mathcal{A}, A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}, A_1, A_2, \dots \in \mathcal{A} \Rightarrow \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$  ist. Für die Abbildung  $\mathbf{P}$  fordert man nun  $\mathbf{P}[\emptyset] = 0, \mathbf{P}[A^c] = 1 - \mathbf{P}[A]$  und  $\mathbf{P}[\bigcup_{n \in \mathbb{N}} A_n] = \sum_{n=1}^{\infty} \mathbf{P}[A_n]$  für paarweise disjunkte  $A_1, A_2, \dots$ . Hier sind nun automatisch alle Werte definiert, da  $\mathcal{A}$  eine  $\sigma$ -Algebra ist.

In dieser Form der Modellierung ist eine  $E$ -wertige Zufallsvariable eine Abbildung  $X : \Omega \rightarrow E$ ; siehe auch Abbildung 1.4. Die Verteilung von  $X$  ist dann gegeben durch

$$\mathbf{P}[X \in A] = \mathbf{P}[X^{-1}(A)],$$

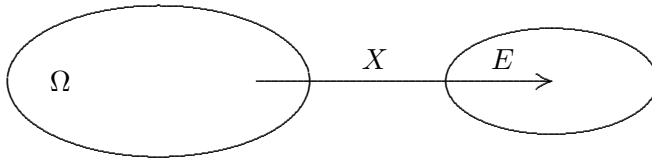
wobei im letzten Ausdruck  $X^{-1}(A) \in \mathcal{A}$  gelten muss.

Wir werden jedoch (wie oben) den intuitiveren Begriff der Zufallsvariable in den Vordergrund stellen. Ganz egal wie man modelliert, wichtig ist, dass auf jeden Fall die Rechenregeln aus Definition 1.14.3 gelten.

<sup>3</sup>Aus der Analysis bekannt ist

$$\int \exp\left(-\frac{x^2}{2}\right) dx = \sqrt{2\pi}.$$

Damit zeigt man durch Substitution  $z = (x - \mu)/\sqrt{\sigma^2}$  leicht, dass es sich bei (1.3) um eine Dichte handelt.



**Abbildung 1.4:** In der klassischen Modellierung ist eine Zufallsvariable eine Abbildung  $X : \Omega \mapsto E$ .

Wir lernen nun noch eine wichtige Formel zum Rechnen mit Wahrscheinlichkeiten kennen.

**Proposition 1.19 (Einschluss- Ausschluss-Formel).** Sei  $X$  eine  $E$ -wertige Zufallsvariable und  $A_1, \dots, A_n \subseteq E$ . Dann gilt

$$\begin{aligned} & \mathbf{P}[X \in A_1 \cup \dots \cup A_n] \\ &= \sum_{i=1}^n \mathbf{P}[X \in A_i] - \sum_{1 \leq i < j \leq n} \mathbf{P}[X \in A_i \cap A_j] + \dots \pm \mathbf{P}[X \in A_1 \cap \dots \cap A_n]. \end{aligned}$$

*Beweis.* Zunächst sei  $n = 2$ . Klar ist, dass  $\mathbf{P}[X \in A_2 \setminus (A_1 \cap A_2)] = \mathbf{P}[X \in A_2] - \mathbf{P}[X \in A_1 \cap A_2]$ , da  $A_1 \cap A_2$  und  $A_2 \setminus (A_1 \cap A_2)$  disjunkt sind. Daraus folgt nun

$$\begin{aligned} \mathbf{P}[X \in A_1 \cup A_2] &= \mathbf{P}[X \in A_1] + \mathbf{P}[X \in A_2 \setminus (A_1 \cap A_2)] \\ &= \mathbf{P}[X \in A_1] + \mathbf{P}[X \in A_2] - \mathbf{P}[X \in A_1 \cap A_2]. \end{aligned}$$

Wir beweisen die Aussage mit Hilfe von vollständiger Induktion. Es gilt wie im Fall  $n = 2$

$$\begin{aligned} \mathbf{P}[X \in A_1 \cup \dots \cup A_{n+1}] &= \mathbf{P}[X \in A_1 \cup \dots \cup A_n] + \mathbf{P}[X \in A_{n+1}] \\ &\quad - \mathbf{P}[X \in (A_1 \cap A_{n+1}) \cup \dots \cup (A_n \cap A_{n+1})] \\ &= \sum_{i=1}^{n+1} \mathbf{P}[X \in A_i] - \sum_{1 \leq i < j \leq n} \mathbf{P}[X \in A_i \cap A_j] - \sum_{i=1}^n \mathbf{P}[X \in A_i \cap A_{n+1}] \\ &\quad \dots \pm \mathbf{P}[X \in A_1 \cap \dots \cap A_n] \pm \sum_{i=1}^n \mathbf{P}[X \in A_1 \cap \dots \cap A_{i-1} \cap A_{i+1} \cap \dots \cap A_{n+1}] \\ &\quad \mp \mathbf{P}[X \in A_1 \cap \dots \cap A_{n+1}]. \end{aligned}$$

Damit ist die Aussage gezeigt. □

**Beispiel 1.20 (Runs).** Ein Algorithmus wird wiederholt (mit zufälligem Input) ausgeführt. Dabei bricht er mit Wahrscheinlichkeit  $p$  ab, mit Wahrscheinlichkeit  $q := 1 - p$  liefert er das richtige Ergebnis. Wie groß ist die Wahrscheinlichkeit, dass er bei  $n$ -maliger Ausführung mindestens  $k$ -mal in Folge abbricht?

Etwas mathematischer können wir das auch so formulieren: Wir werden eine Münze mit

Erfolgswahrscheinlichkeit  $p$  genau  $n$ -mal. Wie groß ist die Wahrscheinlichkeit, dass es zu einem *Run* von mindestens  $k$  Erfolgen kommt?

Um diese Wahrscheinlichkeit zu berechnen, ist die Einschluss- Ausschlussformel hilfreich. Sei  $\underline{X}$  ein zufälliger Vektor der Länge  $n$  mit Einträgen in  $\{0, 1\}$  (wobei 1 einen Erfolg und 0 einen Misserfolg bezeichnet), mit

$$\mathbf{P}[\underline{X} = (x_1, \dots, x_n)] = p^{\sum_{i=1}^n x_i} q^{n - \sum_{i=1}^n x_i}.$$

Weiter sei (mit  $x_0 := 0$  und festem  $k$ )

$$A_i = \{(x_1, \dots, x_n) : x_{i-1} = 0, x_i = \dots = x_{i+k-1} = 1\},$$

also ist  $\underline{X} \in A_i$  das Ereignis, dass im Vektor  $\underline{X}$  ab Position  $i$  einen *Run* der Länge mindestens  $k$  beginnt. Wir wollen also

$$\mathbf{P}[\underline{X} \in A] \text{ für } A := A_1 \cup \dots \cup A_{n-k+1}$$

berechnen. Es ist nun (da das Ereignis  $A_i$  impliziert, dass der Run in  $i$  startet) für  $i < j$

$$\{\underline{X} \in A_i \cap A_j\} = \emptyset \text{ falls } j - i \leq k$$

und damit (da es sich für  $|j - i| > k$  bei den Münzwürfen  $i, \dots, i + k - 1$  und  $j, \dots, j + k - 1$  um unterschiedliche Würfel handelt) für  $i < j$

$$\mathbf{P}[\underline{X} \in A_i \cap A_j] = \begin{cases} 0, & \text{falls } j - i \leq k, \\ q^2 p^{2k}, & \text{falls } j - i > k, i > 1, \\ qp^{2k}, & \text{falls } j - i > k, i = 1. \end{cases}$$

Nun kommt die Einschluss-Ausschlussformel zum Einsatz und wir schreiben

$$\begin{aligned} \mathbf{P}[\underline{X} \in A] &= p^k + (n - k)qp^k - (n - 2k)qp^{2k} - \binom{n - 2k - 1}{2} q^2 p^{2k} \\ &\quad + \binom{n - 2k}{2} q^2 p^{3k} + \binom{n - 3k - 1}{3} q^3 p^{3k} \dots \end{aligned}$$

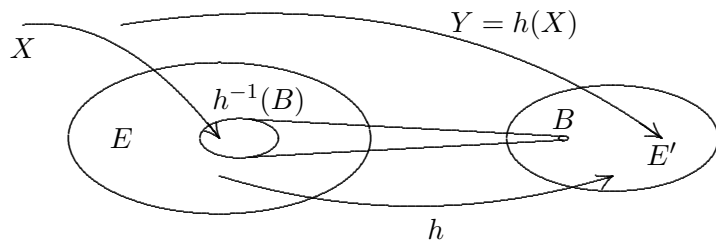
## 1.5 Gemeinsame Verteilung, Bildverteilung, Unabhängigkeit

Bereits in den Beispielen 1.1, 1.2, 1.4, 1.20 haben wir Wahrscheinlichkeiten miteinander multipliziert. Etwa hatten wir gesagt, dass die Wahrscheinlichkeit beim  $p$ -Münzwurf für das zweimalige Auftreten von *Kopf* gerade  $p^2$  ist. Begründet haben wir dies mit der Unabhängigkeit der Münzwürfe. Um dieses Konzept einzuführen, benötigen wir zunächst die gemeinsame Verteilung von Zufallsvariablen.

**Definition 1.21 (Gemeinsame Verteilung).** Ist  $\underline{X} = (X_1, \dots, X_n)$  mit Zielbereich  $E_1 \times \dots \times E_n$ , so heißt die Verteilung von  $\underline{X}$  auch gemeinsame Verteilung von  $X_1, \dots, X_n$ . Die Abbildung

$$A_i \mapsto \mathbf{P}[X_i \in A_i]$$

ist die  $i$ -te Rand- oder Marginalverteilung der gemeinsamen Verteilung von  $X_1, \dots, X_n$ .



**Abbildung 1.5:** Das Bild einer Zufallsvariablen  $X$  unter einer Abbildung  $h$  ist wieder eine Zufallsvariable, nämlich  $h(X)$ .

Gilt für alle Intervalle  $A_1, \dots, A_n \subseteq \mathbb{R}$ , dass

$$\mathbf{P}[\underline{X} \in A_1 \times \dots \times A_n] = \int_A f(x_1, \dots, x_n) dx_1 \cdots dx_n,$$

so heißt  $f$  Dichte der gemeinsamen Verteilung von  $X_1, \dots, X_n$ . Wir schreiben dann auch

$$\mathbf{P}[\underline{X} \in dx_1 \cdots dx_n] = f(x_1, \dots, x_n) dx_1 \cdots dx_n.$$

**Definition 1.22 (Bildverteilung).** Ist  $X$  eine  $E$ -wertige Zufallsvariable und

$$h : E \rightarrow E',$$

so ist  $Y := h(X)$  eine  $E'$ -wertige Zufallsvariable und

$$\{Y = y\} = \{X \in h^{-1}(y)\},$$

also ist die Verteilung von  $Y$  gegeben durch

$$\mathbf{P}[Y \in B] = \mathbf{P}[X \in h^{-1}(B)].$$

Eine Illustration von Bildern von Zufallsvariablen findet sich in Abbildung 6.2.

**Bemerkung 1.23 (Interpretation).** Auch wenn die Definition abstrakt erscheint, sind Bilder von Zufallsvariablen häufig anzutreffen. Sei etwa  $(X_1, \dots, X_n)$  das Ergebnis eines  $n$ -fachen  $p$ -Würfelwurfs und setzen  $q := 1 - p$ . Also ist  $X = (X_1, \dots, X_n)$  eine  $\{0, 1\}^n$ -wertige Zufallsvariable mit

$$\mathbf{P}[X = (x_1, \dots, x_n)] = p^{\sum_{i=1}^n x_i} (1 - p)^{n - \sum_{i=1}^n x_i}.$$

Sei  $h(x_1, \dots, x_n) = x_1 + \dots + x_n$ . Dann gilt  $|h^{-1}(k)| = \binom{n}{k}$ , also

$$\begin{aligned} \mathbf{P}(h(X) = k) &= \mathbf{P}(X \in h^{-1}(k)) \\ &= \sum_{(x_1, \dots, x_n) \in h^{-1}(k)} p^k q^{n-k} = \binom{n}{k} p^k q^{n-k}. \end{aligned}$$

Dies gibt also die Verteilung von  $h(X)$  an.

**Beispiel 1.24 (QUICKSORT).** Wir geben noch ein informatives Beispiel, nämlich die Verteilung der Vergleiche bei QUICKSORT: Der Input ist ein Vektor verschiedener Zahlen  $x_1, \dots, x_n$ . Ziel ist es, diese zu ordnen und wird durch den Pseudocode in Algorithmus 1 beschrieben. Hierbei nennt man jeweils  $a$  das Pivot-Element. Wir geben ein Beispiel mit

$$(x_1, \dots, x_7) = (17, 4, 32, 19, 8, 65, 44).$$

Zunächst wählen wir das Pivot-Element  $a = 17$ , vergleichen alle anderen Elemente des Vektors mit dem Pivot-Element (das ergibt 6 Vergleiche) und vertauschen Paare von Zahlen, die größer und kleiner als das Pivot-Element sind, so dass der Vektor

$$(17, 4, 8, 19, 32, 65, 44)$$

entsteht. Hier sind Elemente, die kleiner als das Pivot-Element sind, links von Elementen, die größer sind. Nun setzen wir das Pivot-Element zwischen diese beiden Mengen, so dass sich

$$(4, 8, 17, 19, 32, 65, 44)$$

ergibt. Anschließend iterieren wir das Ganze mit den Vektoren links und rechts des Pivot-Elements. Die folgenden Schritte sind

(4, 8, 17, 19, 32, 65, 44) mit 1 Vergleich links von 17 und 3 Vergleichen rechts von 17 ,

(4, 8, 17, 19, 32, 65, 44) mit 2 Vergleichen rechts von 19,

(4, 8, 17, 19, 32, 44, 65) mit 1 Vergleich rechts von 32.

Insgesamt ergeben sich also  $6 + 1 + 3 + 2 + 1 = 13$  Vergleiche.

---

**Algorithm 1** QUICKSORT

---

**INPUT:** Unsortierter Vektor von  $r - l + 1$  Zahlen  $x_l, \dots, x_r$

**OUTPUT:** Sortierter Vektor

```

1: if  $r - l > 0$  then
2:    $a \leftarrow x_l$ 
3:    $(x_l, \dots, x_r) \leftarrow$  Vektor, in dem alle Elemente, die kleiner (größer) als  $a$  sind, links (rechts)
      von  $a$  stehen.
4:    $q \leftarrow$  Position von  $a$  in  $(x_l, \dots, x_r)$ 
5:   QUICKSORT( $x_l, \dots, x_{q-1}$ )
6:   QUICKSORT( $x_{q+1}, \dots, x_r$ )
7: end if

```

---

Wir betrachten nun QUICKSORT mit randomisiertem Input  $\underline{X}$ . Das bedeutet, dass für gegebene Zahlen  $z_1 < \dots < z_n$  der Input eine zufällige Permutation  $\underline{\Sigma} = (\sigma(1), \dots, \sigma(n))$ , also  $\underline{z} \circ \underline{\Sigma} := (z_{\sigma(1)}, \dots, z_{\sigma(n)})$  dieser Zahlen ist. Mit anderen Worten kommt jede Umordnung der Zahlen  $(z_1, \dots, z_n)$  mit Wahrscheinlichkeit  $1/n!$  vor, also (mit

$$\mathbf{P}(\underline{X} = \underline{z} \circ \sigma) = \mathbf{P}(\underline{\Sigma} = \sigma) = \frac{1}{n!}, \quad \sigma \in \mathcal{S}_n.$$

Nun ist  $\sigma$  eine  $\mathcal{S}_n$ -wertige Zufallsvariable ist und

$$h : \begin{cases} \mathcal{S}_n & \rightarrow \mathbb{N} \\ \sigma & \mapsto \# \text{Vergleiche, die QUICKSORT mit Input } \underline{z} \circ \sigma \text{ benötigt.} \end{cases}$$

Als Beispiel für  $z = (4, 8, 17, 19, 32, 44, 65)$  und  $\sigma = (2, 5, 1, 4, 3, 7, 6)$  ist etwa  $z \circ \sigma = (17, 4, 32, 19, 8, 65, 44)$  und  $h(\sigma) = 13$ . Da die Verteilung von  $\Sigma$  unbekannt ist (nämlich die Gleichverteilung auf  $\mathcal{S}_n$ ), können wir nun die Bildverteilung von  $h(\Sigma)$  betrachten. Es ergibt sich etwa

$$\mathbb{P}(h(\Sigma) = 21) = \frac{2}{7!},$$

da (wie aus der Algorithmik-Vorlesung bekannt) QUICKSORT genau dann die maximale Anzahl von  $6 + 5 + 4 + 3 + 2 + 1 = 21$  Vergleiche benötigt, wenn der Input geordnet war, also  $h^{-1}(21) = \{(1, 2, 3, 4, 5, 6, 7), (7, 6, 5, 4, 3, 2, 1)\}$ .

**Definition 1.25 (Unabhängigkeit).** Seien  $X_1, \dots, X_n$  Zufallsvariablen. Gilt

$$\mathbf{P}[X_1 \in A_1, \dots, X_n \in A_n] = \mathbf{P}[X_1 \in A_1] \cdots \mathbf{P}[X_n \in A_n] \quad (1.4)$$

für beliebige  $A_1, \dots, A_n$ , so heißt  $(X_1, \dots, X_n)$  unabhängig.

**Bemerkung 1.26 (Unabhängigkeit von diskreten und stetigen Zufallsvariablen).**

Man überlegt sich einfach:

Sind  $X_1, \dots, X_n$  diskrete Zufallsvariable. Dann ist  $(X_1, \dots, X_n)$  genau dann unabhängig, wenn

$$\mathbf{P}[X_1 = x_1, \dots, X_n = x_n] = \mathbf{P}[X_1 = x_1] \cdots \mathbf{P}[X_n = x_n]$$

für beliebige Wahl von  $x_1, \dots, x_n$ .

Weiter gilt:

Sind  $X_1, \dots, X_n$  stetige Zufallsvariablen, so dass die gemeinsame Verteilung eine Dichte  $f$  besitzt. Dann ist  $(X_1, \dots, X_n)$  genau dann unabhängig, wenn es Funktionen  $f_1, \dots, f_n$  gibt mit  $f(x_1, \dots, x_n) = f_1(x_1) \cdots f_n(x_n)$ . Dabei gilt

$$f_i(x_i) = \int f(x_1, \dots, x_n) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n.$$

**Beispiel 1.27.** 1. Beispiele für unabhängige Zufallsvariable haben wir bereits kennen gelernt. Beispielsweise ist ein  $p$ -Münzwurf  $(X_1, \dots, X_n)$  unabhängig.

2. Seien  $X, Y$  gemäß

$$\mathbf{P}[X \in dx, Y \in dy] = f(x) \cdot f(y) dx dy$$

verteilt, wobei  $f := f_{0,1}$  die Dichte der Standardnormalverteilung ist. Dann sind  $X, Y$  unabhängig. Allerdings sind die Zufallsvariablen  $\min(X, Y), \max(X, Y)$  nicht unabhängig. Es gilt etwa für  $a \leq b$

$$\begin{aligned} \mathbf{P}[\min(X, Y) \leq a, \max(X, Y) \leq b] &= \mathbf{P}[X \leq Y, X \leq a, Y \leq b] + \mathbf{P}[Y \leq X, Y \leq a, X \leq b] \\ &= \int_{-\infty}^a \int_{-\infty}^b 1_{x \leq y} f(x) f(y) dy dx + \int_{-\infty}^b \int_{-\infty}^a 1_{y \leq x} f(y) f(x) dy dx \\ &= 2 \int_{-\infty}^a \int_{-\infty}^b 1_{x \leq y} f(x) f(y) dy dx. \end{aligned}$$

Damit hat  $(\min(X, Y), \max(X, Y))$  die gemeinsame Dichte  $2 \cdot 1_{x \leq y} f(x) f(y)$ . Dies lässt sich jedoch nicht als Produkt zweier Funktionen in  $x$  bzw.  $y$  schreiben.

## 2 Verteilungen und deren Eigenschaften

### 2.1 Grundlegendes

Einen besonderen Stellenwert haben reellwertige Zufallsvariablen und deren Verteilungen. Für diese lässt sich die Verteilungsfunktion definieren.

**Definition 2.1 (Verteilungsfunktion und Quantile).** Sei  $E \subseteq \mathbb{R}$  und  $X$  eine  $E$ -wertige Zufallsvariable. Dann heißt die Funktion

$$x \mapsto F_X(x) := \mathbf{P}[X \leq x]$$

Verteilungsfunktion von  $X$ . Sei weiter  $\alpha \in (0, 1)$ . Jedes  $q_\alpha \in \mathbb{R}$  mit

$$\mathbf{P}[X < q_\alpha] \leq \alpha \leq \mathbf{P}[X \leq q_\alpha]$$

heißt  $\alpha$ -Quantil von (der Verteilung von)  $X$ . Jedes 0.5-Quantil heißt Median von (der Verteilung von)  $X$ .

**Lemma 2.2 (Eigenschaften von Verteilungsfunktionen).** Sei  $X$  eine (diskrete oder stetige) reellwertige Zufallsvariable. Dann gilt

$$\begin{aligned} \lim_{x \rightarrow \infty} F_X(x) &= 1, \\ \lim_{x \rightarrow -\infty} F_X(x) &= 0. \end{aligned}$$

Ist  $X$  eine stetige Zufallsvariable mit Dichte  $f(x)dx$ , so gilt

$$F_X(x) = \int_{-\infty}^x f(z)dz.$$

Insbesondere ist also

$$\int_{-\infty}^{\infty} f(z)dz = 1.$$

Weiter ist  $F' = f$  und die Verteilungsfunktion damit differenzierbar.

*Beweis.* Wir zeigen die ersten beiden Aussagen für diskrete Zufallsvariable. Sei  $E \subseteq \mathbb{R}$  der diskrete Wertebereich von  $X$ . Dann gilt schließlich  $\sum_{z \in E} \mathbf{P}[X = z] = 1$  und damit  $\mathbf{P}[X \leq x] = \sum_{z \in E, z \leq x} \mathbf{P}[X = z] \xrightarrow{x \rightarrow \infty} 1$ , da andernfalls die Gesamtsumme von  $\mathbf{P}[X = z]$  nicht konvergiert. Die zweite Gleichung folgt analog.

Die Aussage für stetige Zufallsvariable ist klar, da  $(-\infty, x]$  ein Intervall ist und damit

$$\mathbf{P}[X \leq x] = \mathbf{P}[X \in (-\infty, x]] = \int_{-\infty}^x f(z)dz$$

gilt. □



## 2.2 Uniforme Verteilungen, Laplace-Experimente

Man stelle sich einen Würfelwurf mit einem fairen Würfel vor. Hier sind alle möglichen Ausgänge gleichwahrscheinlich. In einem solchen Fall spricht man auch von einem Laplace-Experiment. Solche Experimente führen zu uniformen Verteilungen, die wir nun kennen lernen.

**Definition 2.3 (Diskrete uniforme Verteilung).** Sei  $E$  eine endliche Menge. Ist  $x \mapsto \mathbf{P}[X = x]$  konstant, so heißt die Verteilung von  $X$  diskrete uniforme Verteilung. In diesem Fall gilt

$$\mathbf{P}[X \in A] = \frac{|A|}{|E|}.$$

**Beispiel 2.4 (Zufällige Permutationen).** Etwa bei der Analyse von QUICKSORT mit zufälligem Input haben wir eine zufällige Permutation  $\Sigma$  betrachtet. Diese hatte eine diskrete uniforme Verteilung auf  $\mathcal{S}_n$ , der Menge der Permutationen von  $\{1, \dots, n\}$ . Ist also etwa

$$A_j = \{\sigma : \sigma(1) = \max(\sigma(1), \dots, \sigma(j))\},$$

so folgt  $|A_j| = \binom{n}{j}(n-j)!(j-1)! = \frac{n!}{j}$  (da die erste Stelle in  $\sigma(1), \dots, \sigma(j)$  bereits festgelegt ist. Deshalb ist

$$\mathbf{P}[\Sigma \in A_j] = \frac{n!}{n!j} = \frac{1}{j}.$$

**Definition 2.5 (Stetige uniforme Verteilung).** Sei  $E = [a, b]$  ein Intervall und  $X$  eine  $E$ -wertige Zufallsvariable. Gilt für  $a \leq c \leq d \leq b$

$$\mathbf{P}[X \in (c, d)] = \frac{d - c}{b - a},$$

so heißt die Verteilung von  $X$  stetige uniforme Verteilung. In diesem Fall hat  $X$  die Dichte

$$\frac{1}{b - a} dx \text{ für } x \in [a, b].$$

und wir schreiben auch  $X \sim U([a, b])$ . Weiter können wir auch die Verteilungsfunktion von  $X$  berechnen. Es gilt

$$\mathbf{P}[X \leq x] = \begin{cases} 0, & x < a, \\ \frac{1}{b-a} \int_a^x dz = \frac{x-a}{b-a}, & a \leq x < b, \\ 1, & x \geq b. \end{cases}$$

**Definition 2.6 (Beta-Verteilung).** Seien  $X_1, \dots, X_n \sim U([0, 1])$  unabhängig. Wir bezeichnen mit  $X_{(i)}$  die  $i$ -te Ordnungsstatistik der Zahlen  $X_1, \dots, X_n$ , d.h.  $X_{(i)}$  ist die  $i$ -t-kleinste Zahl aus  $\{X_1, \dots, X_n\}$ . Dann heißt die Verteilung von  $X_{(i)}$  auch Beta-Verteilung zu den Parametern  $i$  und  $n - i + 1$ . Wir schreiben auch  $X_{(i)} \sim \text{Beta}(i, n - i + 1)$ .

**Lemma 2.7 (Dichte der Beta-Verteilung).** Ist  $X \sim \text{Beta}(\alpha, \beta)$  für  $\alpha, \beta \in \mathbb{N}$ , so besitzt  $X$  die Dichte für  $x \in [0, 1]$

$$f(x) = \binom{\alpha + \beta - 1}{\alpha} \alpha x^{\alpha-1} (1-x)^{\beta-1}$$

**Bemerkung 2.8.** Für eine Dichte  $f$  ist bekanntlich immer  $\int f(x)dx = 1$ . Da wir oben die Dichte der Beta-Verteilung angegeben haben, folgt für  $i, j \in \mathbb{Z}_+$

$$\int_0^1 x^i(1-x)^j dx = \frac{1}{\binom{i+j+1}{i+1}(i+1)}.$$

*Beweis.* Wir müssen für unabhängige  $X_1, \dots, X_n \sim U([0, 1])$  zeigen, dass für  $a \in [0, 1]$  (und für die  $i$ -te Ordnungsstatistik  $X_{(i)}$ )

$$\mathbb{P}[X_{(i)} \leq a] = \int_0^a \binom{n}{i} i x^{i-1} (1-x)^{n-i} dx.$$

Wir schreiben

$$\begin{aligned} \mathbf{P}[X_{(i)} \geq a] &= \mathbf{P}[\text{höchstens } i-1 \text{ der Zahlen } X_1, \dots, X_n \text{ sind kleiner als } a] \\ &= \sum_{k=0}^{i-1} \binom{n}{k} a^k (1-a)^{n-k}. \end{aligned}$$

Da die Dichte gerade die Ableitung der Verteilungsfunktion ist, berechnen wir zunächst

$$\frac{d}{da} \mathbf{P}[X_{(i)} \geq a] = \sum_{k=0}^{i-1} \binom{n}{k} a^{k-1} (1-a)^{n-k-1} \underbrace{(k(1-a) - (n-k)a)}_{=k-na}.$$

Den Rest des Beweises führen Sie bitte als Übungsaufgabe. □

## 2.3 Die Binomial-Verteilung

Denkt man an ein  $n$ -mal unabhängig wiederholt durchgeführtes Spiel, so ist es möglich, zwischen 0 und  $n$ -mal zu gewinnen. Die Anzahl der Gewinne ist eine Zufallsvariable, deren Verteilung wir nun als Binomialverteilung kennen lernen. Diese haben wir schon in Beispiel 1.1 kennen gelernt. Die Verteilungsgewichte und die Verteilungsfunktion der Binomialverteilung sind in Abbildung 2.1 dargestellt.

**Definition 2.9 (Binomialverteilung).** Sei  $E = \{0, \dots, n\}$  und  $0 \leq p \leq 1$ . Ist  $X$  eine  $E$ -wertige Zufallsvariable mit

$$\mathbf{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}, \quad (2.1)$$

so heißt die Verteilung von  $X$  Binomialverteilung mit Parametern  $n$  und  $p$  und wir schreiben  $X \sim B(n, p)$ .

**Bemerkung 2.10.** Die  $B(n, p)$  Verteilung gibt die Anzahl der Erfolge in einem  $n$ -mal unabhängig wiederholt ausgeführten Spiel dar, wobei in jedem Spiel die Gewinnwahrscheinlichkeit gerade  $p$  beträgt.

Denn: Jeder Spielausgang des wiederholten Spiels kann man mit einem  $\{0, 1\}$ -wertigem Vektor der Länge  $n$  beschreiben. Ist etwa  $n = 5$ , so ist  $(1, 1, 0, 0, 1)$  ein Gewinn in den ersten beiden und im letzten Spiel. Ein solcher Spielausgang hat gerade die Wahrscheinlichkeit  $p \cdot p \cdot (1-p) \cdot (1-p) \cdot p = p^3(1-p)^2$ . Da es genau  $\binom{n}{k}$  Vektoren gibt, die aus genau  $k$  Ziffern 1 und  $n-k$  Ziffern 0 bestehen, folgt (2.1).

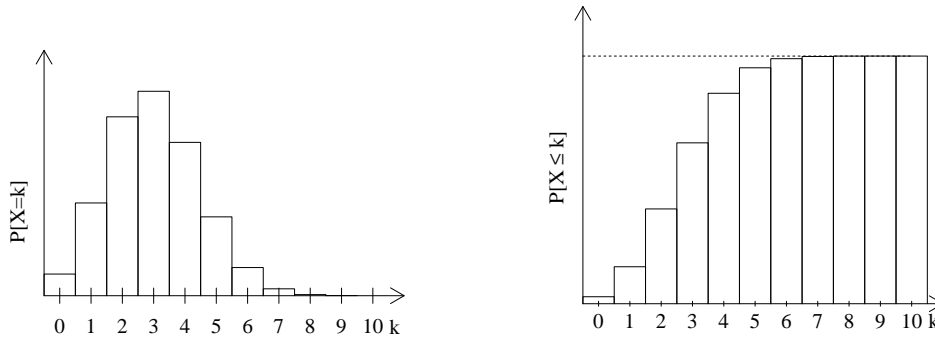


Abbildung 2.1: Verteilungsgewichte und Verteilungsfunktion von  $B(10, 0.2)$ .

## 2.4 Die Normal-Verteilung

Bereits in Beispiel 1.17 haben wir die ein-dimensionale Normal-Verteilung kennen gelernt. Wir erweitern diese nun auf mehrere Dimensionen. Diese Verteilung ist vor allem in der Statistik von höchstem Interesse.

**Definition 2.11 (Mehrdimensionale Normal-Verteilung).** Sei  $\underline{\mu} \in \mathbb{R}^n$  und  $\underline{\Sigma} \in \mathbb{R}^{n \times n}$  symmetrisch und positiv definit.<sup>4</sup> Gilt dann für eine  $\mathbb{R}^n$ -wertige Zufallsvariable  $\underline{X}$ , dass

$$\mathbf{P}[\underline{X} \in dx_1, \dots, dx_n] = \frac{1}{\sqrt{(2\pi)^k \det(\underline{\Sigma})}} \exp\left(-\frac{1}{2}(\underline{x} - \underline{\mu})^\top \underline{\Sigma}^{-1}(\underline{x} - \underline{\mu})\right) dx_1 \cdots dx_n,$$

so heißt  $\underline{X}$  multi-variant normalverteilt mit Erwartungswertvektor  $\underline{\mu}$  und Kovarianz-Matrix  $\underline{\Sigma}$ . Wir schreiben dann  $\underline{X} \sim N(\underline{\mu}, \underline{\Sigma})$ . Ist  $\underline{\mu} = \underline{0}$  und  $\underline{\Sigma} = \underline{E}_n$  (die  $n$ -dimensionale Einheitsmatrix), so heißt  $\underline{X}$  auch Standard-normalverteilt.

## 3 Kenngrößen von Zufallsvariablen

Die Verteilung einer Zufallsvariable ist oftmals kompliziert; man denke etwa an die (zufällige) Anzahl der Vergleiche, die QUICKSORT benötigt um einen Vektor zu sortieren. Deshalb ist es oftmals hilfreich, die Verteilung in Kenngrößen zusammenzufassen. Die wichtigsten sind sicherlich der Erwartungswert und die Varianz der Verteilung. Bei mehr-dimensionalen Verteilungen kommt weiterhin noch die Kovarianz hinzu.

### 3.1 Der Erwartungswert

**Definition 3.1 (Erwartungswert).** Sei  $X$  eine  $E$ -wertige Zufallsvariable und  $h : E \rightarrow \mathbb{R}$ .

<sup>4</sup>Eine Matrix  $\underline{A} = (A_{ij})_{i,j=1,\dots,n} \in \mathbb{R}^{n \times n}$  heißt symmetrisch, falls  $a_{ij} = a_{ji}, i, j = 1, \dots, n$ . Sie heißt positiv definit, falls  $\underline{x}^\top \underline{A} \underline{x} > 0$  für alle  $\underline{x} \in \mathbb{R}^n$ .

1. Ist  $E$  diskret, so heißt

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

Erwartungswert von  $h(X)$ , falls die Summe konvergiert.

2. Ist  $X$  stetig mit Dichte  $f$ , so ist

$$\mathbf{E}[h(X)] := \int h(x) f(x) dx$$

Erwartungswert von  $h(X)$ , falls das Integral existiert.

Ist speziell  $E \subseteq \mathbb{R}$  und  $h = id$ , so heißt

$$\mathbf{E}[X] := \sum_{x \in E} x \mathbf{P}[X = x],$$

$$\mathbf{E}[X] := \int x f(x) dx$$

Erwartungswert von  $X$ , falls Summe und Integral existieren.

**Beispiel 3.2 (Uniforme und Binomialverteilung).** 1. Sei  $X \sim U([a, b])$ . Dann ist

$$\mathbf{E}[X] = \frac{1}{b-a} \int_a^b x dx = \frac{1}{2} \frac{b^2 - a^2}{b-a} = \frac{1}{2}(a+b).$$

2. Sei  $X \sim B(n, p)$ . Dann ist mit  $q := 1 - p$

$$\begin{aligned} \mathbf{E}[X] &= \sum_{k=0}^n k \binom{n}{k} p^k q^{n-k} = np \sum_{k=1}^n \binom{n-1}{k-1} p^{k-1} q^{n-k} \\ &= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k q^{n-1-k} = np. \end{aligned} \tag{3.1}$$

**Bemerkung 3.3 (Transformationssatz).** Sei  $X$  eine  $E$ -wertige diskrete Zufallsvariable und  $h : E \rightarrow \mathbb{R}$ . Dann gilt. Dann gilt

$$\mathbf{E}[h(X)] = \sum_{y \in h(E)} y \mathbf{P}[h(X) = y]$$

falls die Summe existiert.

Denn: Es gilt nach Definition des Erwartungswertes

$$\begin{aligned} \mathbf{E}[h(X)] &= \sum_{x \in E} h(x) \mathbf{P}[X = x] = \sum_{y \in h(E)} \sum_{x \in h^{-1}(y)} h(x) \mathbf{P}[X = x] \\ &= \sum_{y \in h(E)} y \sum_{x \in h^{-1}(y)} \mathbf{P}[X = x] = \sum_{y \in h(E)} y \mathbf{P}[h(X) = y]. \end{aligned}$$

Als einfaches Beispiel des Transformationssatzes berechnen wir den Erwartungswert einer Indikatorfunktion.

**Beispiel 3.4 (Indikatorfunktion).** Sei  $X$  eine  $E$ -wertige Zufallsvariable und  $1_A : E \rightarrow \{0, 1\}$  gegeben durch

$$1_A(x) := \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

Dann ist

$$\mathbf{E}[1_A(X)] = \sum_{y=0,1} y \mathbf{P}[1_A(X) = y] = \mathbf{P}[1_A(X) = 1] = \mathbf{P}[X \in A].$$

**Lemma 3.5.** Sei  $X$  eine  $\mathbb{Z}_+$ -wertige Zufallsvariable. Dann gilt

$$\begin{aligned} \mathbf{E}[X] &= \sum_{x=0}^{\infty} \mathbf{P}[X > x], \\ \mathbf{E}[X(X-1)] &= 2 \cdot \sum_{x=0}^{\infty} x \cdot \mathbf{P}[X > x] \end{aligned}$$

falls jeweils die Summe existiert.

*Beweis.* Für die erste Behauptung schreiben wir

$$\begin{aligned} \mathbf{E}[X] &= \sum_{x=0}^{\infty} x \mathbf{P}[X = x] = \sum_{x=1}^{\infty} \sum_{y=1}^x \mathbf{P}[X = x] = \sum_{y=1}^{\infty} \sum_{x=y}^{\infty} \mathbf{P}[X = x] \\ &= \sum_{y=1}^{\infty} \mathbf{P}[X \geq y] = \sum_{y=0}^{\infty} \mathbf{P}[X > y] \end{aligned}$$

wobei die Umordnung der Doppelsumme wegen der absoluten Konvergenz der Reihe möglich ist. Für die zweite Behauptung ist analog

$$\begin{aligned} \mathbf{E}[X(X-1)] &= 2 \sum_{x=1}^{\infty} \frac{x(x-1)}{2} \cdot \mathbf{P}[X = x] = 2 \sum_{x=1}^{\infty} \sum_{y=0}^{x-1} y \cdot \mathbf{P}[X = x] \\ &= 2 \sum_{y=0}^{\infty} \sum_{x=y+1}^{\infty} y \cdot \mathbf{P}[X = x] = 2 \sum_{y=0}^{\infty} y \cdot \mathbf{P}[X > y]. \end{aligned}$$

□

**Beispiel 3.6 (Diskrete uniforme Verteilung).** Sei  $X$  uniform auf  $\{k, \dots, \ell\}$  verteilt. Dann ist

$$\begin{aligned} \mathbf{E}[X] &= \sum_{x=k}^{\ell} x \frac{1}{\ell - k + 1} = \frac{1}{\ell - k + 1} \left( \sum_{x=0}^{\ell} x - \sum_{x=0}^{k-1} x \right) = \frac{\binom{\ell+1}{2} - \binom{k}{2}}{\ell - k + 1} \\ &= \frac{(\ell - k + 1)(\ell + k)}{2(\ell - k + 1)} = \frac{\ell + k}{2}. \end{aligned}$$

Genauso können wir schreiben

$$\begin{aligned} \mathbf{E}[X] &= \sum_{x=0}^{\ell} \mathbf{P}[X > x] = \sum_{x=0}^{k-1} 1 + \sum_{x=k}^{\ell} \frac{\ell - x}{\ell - k + 1} = k + \frac{1}{\ell - k + 1} \sum_{x=0}^{\ell-k} x \\ &= k + \frac{\binom{\ell-k+1}{2}}{\ell - k + 1} = k + \frac{\ell - k}{2} = \frac{\ell + k}{2}. \end{aligned}$$

**Proposition 3.7 (Eigenschaften des Erwartungswertes).** *Seien  $X, Y$  (diskrete oder stetige) Zufallsvariablen.*

1. Sei  $a, b \in \mathbb{R}$ . Dann gilt

$$\mathbf{E}[aX + bY] = a\mathbf{E}[X] + b\mathbf{E}[Y],$$

falls alle Erwartungswerte existieren.

2. Gilt  $X \leq Y$ , so ist  $\mathbf{E}[X] \leq \mathbf{E}[Y]$ .

*Beweis.* Wir zeigen beide Aussagen nur im Fall von diskreten Zufallsvariablen. Es gilt

$$\begin{aligned} \mathbf{E}[aX + bY] &= \sum_{x,y \in E} (ax + by)\mathbf{P}[X = x, Y = y] \\ &= a \sum_{x,y \in E} x\mathbf{P}[X = x, Y = y] + b \sum_{x,y \in E} y\mathbf{P}[X = x, Y = y] \\ &= a \sum_{x \in E} \mathbf{P}[X = x] + b \sum_{y \in E} y\mathbf{P}[Y = y] = a\mathbf{E}[X] + b\mathbf{E}[Y]. \end{aligned}$$

Für 2. genügt es zu zeigen, dass  $\mathbf{E}[X] \geq 0$  für  $X \geq 0$ . (Dann ist nämlich  $\mathbf{E}[Y] - \mathbf{E}[X] = \mathbf{E}[Y - X] \geq 0$ .) Wir schreiben, falls  $X \geq 0$

$$\mathbf{E}[X] = \sum_{x \in E} x\mathbf{P}[X = x] = \sum_{x \in E, x \geq 0} x\mathbf{P}[X = x] \geq \sum_{x \in E, x \geq 0} 0 \cdot \mathbf{P}[X = x] = 0.$$

□

**Beispiel 3.8 ( $p$ -Münzwurf).** Ist  $X = (X_1, \dots, X_n)$  ein  $p$ -Münzwurf, so wissen wir bereits, dass  $Y := X_1 + \dots + X_n \sim B(n, p)$ . Also muss auch

$$\mathbf{E}[Y] = \mathbf{E}[X_1 + \dots + X_n] = \mathbf{E}[X_1] + \dots + \mathbf{E}[X_n] = n\mathbf{E}[X_1] = n\mathbf{P}[X_1 = 1] = np$$

in Übereinstimmung mit (3.1).

**Beispiel 3.9 (Average-Case-Analyse von QUICKSORT).** In Beispiel 1.24 haben wir bereits QUICKSORT für einen zufälligen Input  $\underline{X}$  (als eine zufällige Umordnung  $\Sigma$  der Zahlen  $z_1 < \dots < z_n$ ) kennen gelernt. OBdA nehmen wir nun an, dass  $z_1 = 1, \dots, z_n = n$ . (Das bedeutet, dass QUICKSORT den Vektor  $\sigma(1), \dots, \sigma(n)$  in den Vektor  $1, \dots, n$  umordnet.) Weiter haben wir die Verteilung der Anzahl der Vergleiche  $h(\Sigma)$  bei Vorliegen der Umordnung  $\Sigma$  definiert. Wir kommen nun zu einer Möglichkeit,  $\mathbf{E}[h(\Sigma)]$  zu berechnen. Hier spricht man auch von der Average-Case-Analyse, da es ja gerade um das mittlere Verhalten des Algorithmus geht.

Um  $\mathbf{E}[h(\Sigma)]$  zu berechnen, formalisieren wir die Funktion  $h(\sigma)$ . Wir erinnern an die Definition der Indikatorfunktion  $1_A$  aus Beispiel 3.4 und setzen für  $i < j$

$$\begin{aligned} A_{ij} &:= \{\sigma : \text{QUICKSORT vergleicht } i \text{ und } j\}, \\ \gamma_{ij}(\sigma) &:= 1_{A_{ij}}(\sigma). \end{aligned}$$

Da  $i$  und  $j$  höchstens einmal miteinander verglichen werden (nämlich genau dann, wenn entweder  $i$  oder  $j$  Pivot-Element ist), können wir

$$h(\sigma) = \sum_{i < j} \gamma_{ij}(\sigma)$$

schreiben. Betrachten wir nun das Ereignis  $A_{ij}$ . Wie in Beispiel 1.24 sei

$$\sigma = (3, 1, 5, 4, 2, 7, 6).$$

Da zunächst  $\sigma(1) = 3$  als Pivot-Element gewählt wird, mit dem alle anderen Zahlen verglichen werden, ist  $\gamma_{\sigma(1)j}(\sigma) = 1$  für  $j = \sigma(2), \dots, \sigma(7)$ . Bei der nächsten Iteration werden die Elemente kleiner und größer als 3 weiterbehandelt. Bereits hier sehen wir, dass  $\sigma(4) = 4$  nicht mehr mit  $\sigma(5) = 2$  verglichen wird, da nach einer Iteration  $\sigma(4)$  rechts und  $\sigma(5)$  links des ersten Pivot-Elementes  $\sigma(1)$  steht. So sieht man ein, dass  $i$  und  $j$  genau dann miteinander verglichen werden (und damit  $A_{ij}$  eintritt), wenn entweder  $i$  oder  $j$  im (durch  $\sigma$  gegebenen) Teilvektor, der die Zahlen  $i, \dots, j$  enthält (im Beispiel wäre das für  $i = 4, j = 2$  also  $(3, 4, 2)$ ), an erster Stelle stehen. (Denn: dann wird QUICKSORT bei einer Iteration  $i$  und  $j$  auf eine Seite eines Pivot-Elements legen und entweder  $i$  oder  $j$  als neues Pivot-Element bestimmen.) Daraus ergibt sich

$$\gamma_{ij}(\sigma) = 1 \text{ (Im Teilvektor von } \sigma, \text{ der } i, \dots, j \text{ enthält, kommt } i \text{ oder } j \text{ an erster Stelle)} \quad (\square)$$

und, da die Wahrscheinlichkeit, dass  $i$  oder  $j$  in einer Permutation von  $i, \dots, j$  an erster Stelle steht, gesucht ist,

$$\mathbf{P}[\Sigma \in A_{ij}] = \frac{2}{j - i + 1}.$$

Damit schreiben wir (siehe Beispiel 3.4 und Proposition 3.7)

$$\begin{aligned} \mathbf{E}[h(\Sigma)] &= \sum_{i < j} \mathbf{E}[\gamma_{ij}(\Sigma)] = \sum_{i < j} \mathbf{P}[\Sigma \in A_{ij}] = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{2}{j - i + 1} \\ &= \Omega\left(\int_1^n \int_x^n \frac{2}{y - x + 1} dy dx\right) = \Omega\left(\int_1^n 2 \log(n - x + 1) dx\right) \\ &= \Omega\left(\int_1^n 2 \log(x) dx\right) = \Omega(2n \log(n)). \end{aligned}$$

Also hat QUICKSORT eine durchschnittliche Laufzeit (die wohl proportional zur Anzahl der durchgeführten Vergleiche ist) von  $\Omega(2n \log n)$ .

### 3.2 Die Varianz und Kovarianz

**Definition 3.10 (Varianz).** Sei  $X$  entweder eine diskrete,  $E \subseteq \mathbb{R}$ -wertige Zufallsvariable, oder eine stetige Zufallsvariable und es existiere  $\mu := \mathbf{E}[X]$ . Dann ist

$$\sigma^2 := \mathbf{Var}[X] = \mathbf{E}[(X - \mu)^2]$$

die Varianz von  $X$ . In beiden Fällen heißt

$$\sigma := \sqrt{\mathbf{Var}[X]}$$

die Standardabweichung von  $X$ .

**Bemerkung 3.11.** Für diskrete Zufallsvariablen ist also

$$\mathbf{Var}[X] = \sum_{x \in E} (x - \mu)^2 \mathbf{P}[X = x],$$

für stetige mit Dichte  $f$  gerade

$$\mathbf{Var}[X] = \int (x - \mu)^2 f(x) dx$$

falls Summe und Integral existieren.

**Proposition 3.12 (Rechenregeln für die Varianz).** *Sei  $X$  entweder eine diskrete,  $E \subseteq \mathbb{R}$ -wertige Zufallsvariable oder eine stetige Zufallsvariable. Dann gilt*

$$\mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = \mathbf{E}[X(X-1)] + \mathbf{E}[X] - \mathbf{E}[X]^2$$

*falls alle Erwartungswerte existieren.*

*Beweis.* Die zweite Gleichheit ist klar wegen der Linearität des Erwartungswertes. Sei  $\mu := \mathbf{E}[X]$ . Dann schreiben wir

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mu)^2] = \mathbf{E}[X^2 - 2\mu X + \mu^2] = \mathbf{E}[X^2] - 2\mu\mathbf{E}[X] + \mu^2 = \mathbf{E}[X^2] - \mu^2.$$

Damit ist alles gezeigt. □

**Beispiel 3.13 (Uniforme und Binomialverteilung).** 1. Sei  $X \sim U([a, b])$ . Dann ist

$$\begin{aligned} \mathbf{Var}[X] &= \frac{1}{b-a} \int_a^b x^2 dx - \frac{(a+b)^2}{4} = \frac{1}{3} \frac{b^3 - a^3}{b-a} - \frac{(a+b)^2}{4} \\ &= \frac{4b^2 + 4ab + 4a^2 - 3a^2 - 6ab - 3b^2}{12} = \frac{(b-a)^2}{12}. \end{aligned}$$

2. Sei  $X \sim B(n, p)$ . Dann ist mit  $q := 1 - p$

$$\begin{aligned} \mathbf{E}[X(X-1)] &= \sum_{k=0}^n k(k-1) \binom{n}{k} p^k q^{n-k} \\ &= n(n-1)p^2 \sum_{k=2}^n \binom{n-2}{k-2} p^{k-2} q^{n-k} = n(n-1)p^2, \end{aligned}$$

also

$$\mathbf{Var}[X] = n(n-1)p^2 + np - (np)^2 = npq.$$

Wir verallgemeinern das letzte Resultat von der uniformen Verteilungen auf die Beta-Verteilung. (Die  $U([0, 1])$ -Verteilung ist ja ebenso die  $Beta(1, 1)$ -Verteilung).

**Lemma 3.14 (Erwartungswert und Varianz der Beta-Verteilung).** *Sei  $X \sim Beta(\alpha, \beta)$  Dann ist*

$$\mathbf{E}[X] = \frac{\alpha}{\alpha + \beta}, \quad \mathbf{V}[X] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

*Beweis.* Übung! □



**Definition 3.15 (Kovarianz und Korrelationskoeffizient).** Seien  $X, Y$  diskrete oder stetige Zufallsvariablen, deren Erwartungswerte  $\mu_X := \mathbf{E}[X], \mu_Y := \mathbf{E}[Y]$  sowie Varianzen  $\sigma_X^2 := \mathbf{Var}[X], \sigma_Y^2 := \mathbf{Var}[Y]$  existieren. Dann existiert auch

$$\mathbf{Cov}[X, Y] := \mathbf{E}[(X - \mu_X)(Y - \mu_Y)]$$

und heißt Kovarianz von  $X$  und  $Y$ . Gilt  $\mathbf{Cov}[X, Y] = 0$ , so heißen  $X, Y$  unkorreliert.

Außerdem nennen wir

$$\mathbf{Kor}[X, Y] := \frac{\mathbf{Cov}[X, Y]}{\sigma_X \cdot \sigma_Y}$$

Korrelationskoeffizient von  $X$  und  $Y$ .

**Lemma 3.16 (Rechenregeln für Kovarianzen).** Seien  $X, Y, Z$  (diskrete oder stetige) reellwertige Zufallsvariablen mit endlicher Varianz. Dann gilt

$$\begin{aligned} \mathbf{Cov}[X, Y] &= \mathbf{Cov}[Y, X], \\ \mathbf{Var}[X] &= \mathbf{Cov}[X, X], \\ \mathbf{Cov}[X, Y] &= \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y], \\ \mathbf{Cov}[X, aY + bZ] &= a\mathbf{Cov}[X, Y] + b\mathbf{Cov}[X, Z]. \end{aligned}$$

*Beweis.* Wir setzen  $\mu_X := \mathbf{E}[X], \mu_Y := \mathbf{E}[Y]$ . Die ersten beiden Gleichungen sind klar. Für die dritte Gleichung schreiben wir

$$\mathbf{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbf{E}[XY] - \mu_X \mathbf{E}[Y] - \mu_Y \mathbf{E}[X] + \mu_X \mu_Y = \mathbf{E}[XY] - \mu_X \mu_Y.$$

Die vierte folgt aus der Definition der Kovarianz und der Linearität des Erwartungswertes.  $\square$

**Lemma 3.17 (Cauchy–Schwartz Ungleichung).** Es gilt

$$\mathbf{Cov}[X, Y]^2 \leq \mathbf{E}[|(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])|]^2 \leq \mathbf{Var}[X]\mathbf{Var}[Y].$$

Insbesondere ist

$$-1 \leq \mathbf{Kor}[X, Y] \leq 1.$$

*Beweis.* Wir setzen oBdA  $\mathbf{E}[X] = \mathbf{E}[Y] = 0$ . Die erste Ungleichung ist klar, da  $XY \leq |XY|$ . Für die zweite genügt es den Fall  $\mathbf{E}[X^2] > 0$  zu betrachten. Andernfalls muss nämlich  $\mathbf{P}[X = 0] = 1$  gelten und dann ist nichts zu zeigen. Für jedes  $c \in \mathbb{R}$  gilt  $0 \leq (-c|X| + |Y|)^2 = c^2 X^2 - 2c|XY| + Y^2$ . Insbesondere gilt für  $c := \mathbf{E}[|XY|]/\mathbf{E}[X^2]$

$$\begin{aligned} 0 &\leq c^2 \mathbf{E}[X^2] - 2c \mathbf{E}[|XY|] + \mathbf{E}[Y^2] = \mathbf{E}[|XY|^2]/\mathbf{E}[X^2] - 2\mathbf{E}[|XY|]^2/\mathbf{E}[X^2] + \mathbf{E}[Y^2] \\ &= \mathbf{E}[Y^2] - \mathbf{E}[|XY|]^2/\mathbf{E}[X^2] \end{aligned}$$

woraus die erste Behauptung direkt folgt. Die zweite folgt dann aus der Definition des Korrelationskoeffizienten.  $\square$

**Proposition 3.18 (Unabhängigkeit und Unkorreliertheit).** Seien  $X, Y$  (diskrete oder stetige) reellwertige, unabhängige Zufallsvariablen, deren Varianz existiert. Dann sind  $X, Y$  auch unkorreliert, d.h.  $\mathbf{Cov}[X, Y] = 0$ .

*Beweis.* Wir zeigen die Behauptung nur für diskrete Zufallsvariablen. Es gilt

$$\begin{aligned}\mathbf{E}[XY] &= \sum_{x,y} xy\mathbf{P}[X = x, Y = y] = \sum_x x\mathbf{P}[X = x] \sum_y y\mathbf{P}[Y = y] \\ &= \mathbf{E}[X]\mathbf{E}[Y].\end{aligned}$$

□

**Proposition 3.19 (Varianz einer Summe).** *Seien  $X_1, \dots, X_n$  (diskrete oder stetige) reellwertige Zufallsvariablen mit endlichen Varianzen. Dann gilt*

$$\mathbf{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbf{Var}[X_i] + 2 \sum_{1 \leq i < j \leq n} \mathbf{Cov}[X_i, X_j].$$

*Ist insbesondere  $(X_1, \dots, X_n)$  paarweise unkorreliert, so gilt*

$$\mathbf{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbf{Var}[X_i].$$

*Beweis.* Wir verwenden Lemma 3.16 und schreiben

$$\begin{aligned}\mathbf{Var}\left[\sum_{i=1}^n X_i\right] &= \mathbf{Cov}\left[\sum_{i=1}^n X_i, \sum_{i=1}^n X_i\right] = \sum_{i=1}^n \sum_{j=1}^n \mathbf{Cov}[X_i, X_j] \\ &= \sum_{i=1}^n \mathbf{Cov}[X_i, X_i] + \sum_{i=1}^n \sum_{j=1}^{i-1} \mathbf{Cov}[X_i, X_j] + \sum_{i=1}^n \sum_{j=i+1}^n \mathbf{Cov}[X_i, X_j] \\ &= \sum_{i=1}^n \mathbf{Var}[X_i] + 2 \sum_{1 \leq i < j \leq n} \mathbf{Cov}[X_i, X_j].\end{aligned}$$

□

**Beispiel 3.20 (Varianz der Vergleiche bei QUICKSORT).** Bei QUICKSORT (mit  $\underline{z} = (1, \dots, n)$ ) hatten wir die Anzahl der Vergleiche  $h(\sigma)$  als (siehe (□))

$$h(\sigma) = \sum_{i < j} \gamma_{ij}(\sigma),$$

$$\gamma_{ij}(\sigma) = 1(i \text{ und } j \text{ werden beim Input } \sigma \text{ miteinander verglichen})$$

$$= 1(\text{Im Teilvektor von } \sigma, \text{ der } i, \dots, j \text{ enthält, kommt } i \text{ oder } j \text{ an erster Stelle}) \quad (\square)$$

geschrieben. Wir werden nun zeigen, dass (für zufälligen Input  $\Sigma$ )

$$\mathbf{V}[h(\Sigma)] \leq 3n(n-1).$$

Hierzu verwenden wir Proposition 3.19 verwenden. Zunächst ist mit Proposition 3.12 und, da  $1_A^2 = 1_A$  für eine Indikatorfunktion  $1_A$  immer gilt,

$$\mathbf{V}[\gamma_{ij}(\Sigma)] = \frac{2}{j-i+1} - \left(\frac{2}{j-i+1}\right)^2 \leq \frac{2}{j-i+1}.$$

Für die Kovarianz-Terme definieren wir

$$T_{ij} := T_{ij}(\sigma) := \text{erster Eintrag im Teilvektor } i, \dots, j \text{ in } \sigma$$

(womit  $\gamma_{ij} = 1$  genau dann gilt, wenn  $T_{ij} \in \{i, j\}$ ) und schreiben

$$\mathbf{Cov}[\gamma_{ij}(\Sigma), \gamma_{kl}(\Sigma)] = \mathbf{P}[T_{ij} \in \{i, j\}, T_{kl} \in \{k, l\}] - \frac{2}{j-i+1} \frac{2}{l-k+1}.$$

Sei  $i < j \neq k < l$ . Man überlegt sich, dass dann  $T_{ij}$  und  $T_{kl}$  unabhängig sind, da sie verschiedene Einträge in  $\sigma$  betreffen. Entsprechend ist in diesem Fall die Kovarianz 0. Wir untersuchen nun noch die Fälle  $i < j = k < l$ ,  $i = k < j < l$  und  $i < k < j = l$ . Im Fall  $j = k$  schreiben wir

$$\begin{aligned} \mathbf{E}[\gamma_{ij}\gamma_{jl}] &= \mathbf{P}[T_{il} = j] + \mathbf{P}[T_{il} = i, T_{jl} \in \{j, l\}] + \mathbf{P}[T_{il} = l, T_{ij} \in \{i, j\}] \\ &= \frac{1}{l-i+1} + \frac{1}{l-i+1} \frac{2}{l-j+1} + \frac{1}{l-i+1} \frac{2}{j-i+1} \\ &\leq \frac{1}{l-i+1} + \mathbf{E}[\gamma_{ij}]\mathbf{E}[\gamma_{jl}]. \end{aligned}$$

Im Fall  $i = k$  ist (für  $j < l$ )

$$\begin{aligned} \mathbf{E}[\gamma_{ij}\gamma_{il}] &= \mathbf{P}[T_{il} = i] + \mathbf{P}[T_{il} = l, T_{ij} \in \{i, j\}] \\ &= \frac{1}{l-i+1} + \frac{1}{l-i+1} \frac{2}{j-i+1} \\ &\leq \frac{1}{l-i+1} + \mathbf{E}[\gamma_{ij}]\mathbf{E}[\gamma_{il}]. \end{aligned}$$

Nun ist für  $j = l$  (und  $i < k$ )

$$\begin{aligned} \mathbf{E}[\gamma_{il}\gamma_{kl}] &= \mathbf{P}[T_{il} = l] + \mathbf{P}[T_{il} = i, T_{kl} \in \{k, l\}] \\ &= \frac{1}{l-i+1} + \frac{1}{l-i+1} \frac{2}{l-k+1} \\ &\leq \frac{1}{l-i+1} + \mathbf{E}[\gamma_{il}]\mathbf{E}[\gamma_{kl}]. \end{aligned}$$

Nun ist mit Proposition 3.19 (der Faktor 6 im zweiten Term erklärt sich mit einem Faktor 2 aus Proposition 3.19 und den drei verschiedenen Fällen  $j = k$ ,  $i = k$  und  $j = l$  von oben)

$$\begin{aligned} \mathbf{Var}[h(\Sigma)] &\leq \sum_{1 \leq i < j \leq n} \frac{2}{j-i+1} + \sum_{1 \leq i < j < l \leq n} \frac{6}{l-i+1} \\ &= \sum_{i < j} \frac{2}{j-i+1} + \sum_{i < l} \frac{6(l-i-1)}{l-i+1} \\ &\leq \sum_{i < j} 6 = 6 \binom{n}{2} \end{aligned}$$

### 3.3 Die Normal-Verteilung

Die Normal-Verteilung – besonders bekannt durch die Gauss'sche Glockenkurve – spielt in statistischen Anwendungen eine große Rolle. Diese Verteilung haben wir bereits in Beispiel 1.17 kennen gelernt. Zunächst wollen wir zeigen, dass es sich bei den Parametern  $\mu$  und  $\sigma^2$  genau um Erwartungswert und Varianz handelt.

**Proposition 3.21 (Erwartungswert und Varianz der Normal-Verteilung).** *Sei  $X \sim N(\mu, \sigma^2)$ . Dann ist*

$$\frac{X - \mu}{\sqrt{\sigma^2}} \sim N(0, 1)$$

und

$$\mathbf{E}[X] = \mu, \quad \mathbf{Var}[X] = \sigma^2.$$

*Beweis.* Für die erste Aussage schreiben wir mit  $Z \sim N(0, 1)$

$$\begin{aligned} \mathbf{P}\left(\frac{X - \mu}{\sqrt{\sigma^2}} \leq x\right) &= \mathbf{P}\left(X \leq x\sqrt{\sigma^2} + \mu\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^{x\sqrt{\sigma^2} + \mu} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right) dy \\ &= \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz = \mathbf{P}(Z \leq x). \end{aligned}$$

Zunächst ist mit partieller Integration

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

also  $\mathbf{E}[Z] = 0$ ,  $\mathbf{Var}[Z] = 1$ . Daraus und aus der ersten Aussage folgt

$$\begin{aligned} \mathbf{E}[X] &= \sqrt{\sigma^2} \mathbf{E}\left[\frac{X - \mu}{\sqrt{\sigma^2}}\right] + \mu = \sqrt{\sigma^2} \mathbf{E}[Z] + \mu = \mu, \\ \mathbf{Var}[X] &= \sigma^2 \mathbf{Var}\left[\frac{X - \mu}{\sqrt{\sigma^2}}\right] = \sigma^2 \mathbf{Var}[Z] = \sigma^2. \end{aligned}$$

□

Um einzusehen, dass für eine mehrdimensionale normalverteilte Zufallsvariable  $\underline{X} \sim N(\underline{\mu}, \underline{\Sigma})$  ähnliche Aussagen wie in der letzten Proposition gelten, würden wir die Theorie der mehrdimensionalen Integration benötigen. Da diese nicht zur Verfügung steht, bemühen wir ein praktisches Hilfsmittel, das uns ebenfalls zum Ziel führen wird.

**Definition 3.22 (Charakteristische Funktion).** *Für eine Zufallsvariable  $\underline{X}$  mit Werten in  $\mathbb{R}^n$  definieren wir ihre charakteristische Funktion  $\psi_{\underline{X}}$  durch*

$$t \mapsto \psi_{\underline{X}}(t) := \mathbf{E}[e^{it^\top \underline{X}}].$$

Hierbei ist  $e^{ix} := \cos(x) + i \sin(x)$  und  $i$  ist die imaginäre Einheit.<sup>5</sup>

Ohne Beweis werden wir folgendes wichtige Resultat verwenden. Es besagt, dass eine charakteristische Funktion die Verteilung einer Zufallsvariablen eindeutig bestimmt.

**Theorem 3.23 (Eindeutigkeit der charakteristischen Funktion).** *Seien  $\underline{X}, \underline{Y}$  Zufallsvariable mit Werten in  $\mathbb{R}^n$ . Gilt  $\psi_{\underline{X}}(\underline{t}) = \psi_{\underline{Y}}(\underline{t})$  für alle  $\underline{t} \in \mathbb{R}^n$ , so haben  $\underline{X}$  und  $\underline{Y}$  dieselbe Verteilung.*

Da die charakteristische Funktion eine Verteilung eindeutig bestimmt, kann man sie auch verwenden, um eine alternative Charakterisierung der Normal-Verteilung anzugeben.

**Theorem 3.24 (Alternative Definition der mehrdimensionalen Normal-Verteilung).** *Sei  $\underline{X}$  eine Zufallsvariable mit Werten in  $\mathbb{R}^n$ . Dann ist  $\underline{X} \sim N(\underline{\mu}, \underline{\Sigma})$  genau dann, wenn*

$$\psi_{\underline{X}}(\underline{t}) = e^{i\underline{t}^\top \underline{\mu} - \frac{1}{2} \underline{t}^\top \underline{\Sigma} \underline{t}}.$$

Nun können wir Proposition 3.21 ins Mehrdimensionale übertragen.

**Proposition 3.25 (Lineare Transformation normalverteilter Zufallsvariablen).** *Sei  $\underline{X} \sim N(\underline{\mu}, \underline{\Sigma})$  für  $\underline{\mu} \in \mathbb{R}^n$  und  $\underline{\Sigma} \in \mathbb{R}^{n \times n}$  symmetrisch und positiv definit. Dann ist*

$$\mathbf{E}[X_i] = \mu_i, \quad \mathbf{Cov}[X_i, X_j] = \Sigma_{ij}.$$

Weiter sei  $\underline{\nu} \in \mathbb{R}^k$  und  $\underline{B} \in \mathbb{R}^{k \times n}$  (mit  $k \leq n$ ) so, dass  $\underline{B} \underline{\Sigma} \underline{B}^\top$  invertierbar ist. Dann ist

$$\underline{Y} := \underline{\nu} + \underline{B} \underline{X} \sim N(\underline{\nu} + \underline{B} \underline{\mu}, \underline{B} \underline{\Sigma} \underline{B}^\top).$$

Ist insbesondere  $\underline{B}$  so, dass  $\underline{B} \underline{\Sigma} \underline{B}^\top = \underline{D}$  eine Diagonalmatrix ist<sup>6</sup>, dann sind  $Y_1, \dots, Y_n$  unabhängig.

*Beweis.* Wir beginnen mit der zweiten Aussage. Da  $\underline{X} \sim N(\underline{\mu}, \underline{\Sigma})$ , können wir die charakteristische Funktion von  $\underline{X}$  verwenden, um

$$\begin{aligned} \mathbf{E}[e^{i\underline{t}^\top \underline{Y}}] &= \mathbf{E}[e^{i\underline{t}^\top \underline{\nu}} e^{i(\underline{B}^\top \underline{t})^\top \underline{X}}] = e^{i\underline{t}^\top \underline{\nu}} e^{i(\underline{B}^\top \underline{t})^\top \underline{\mu} - \frac{1}{2} (\underline{B}^\top \underline{t})^\top \underline{\Sigma} (\underline{B}^\top \underline{t})} \\ &= e^{i\underline{t}^\top (\underline{\nu} + \underline{B} \underline{\mu}) - \frac{1}{2} \underline{t}^\top \underline{B} \underline{\Sigma} \underline{B}^\top \underline{t}} \end{aligned}$$

zu schreiben. Dies ist aber gerade die charakteristische Funktion von  $N(\underline{\nu} + \underline{B} \underline{\mu}, \underline{B} \underline{\Sigma} \underline{B}^\top)$ , d.h.  $\underline{Y}$  hat gerade diese Verteilung. Um die erste Aussage zu zeigen, wählen wir  $\underline{Z} \sim N(\underline{0}, \underline{E}_n)$ , wobei  $\underline{E}_n$  die  $n \times n$ -Einheitsmatrix ist. Hier ist klar, dass  $Z_1, \dots, Z_n$  unabhängig sind, da die

<sup>5</sup>Den Erwartungswert der  $\mathbb{C}^n$ -wertigen Zufallsvariable  $e^{i\underline{t}^\top \underline{X}}$  berechnen wir dann mittels

$$\mathbf{E}[e^{i\underline{t}^\top \underline{X}}] := \mathbf{E}[\cos(\underline{t}^\top \underline{X})] + i \mathbf{E}[\sin(\underline{t}^\top \underline{X})].$$

<sup>6</sup>Nach einem Satz aus der linearen Algebra (Hauptachsentransformation) gibt es für symmetrische Matrizen  $\underline{\Sigma}$  immer eine orthogonale Matrix  $\underline{B} \in \mathbb{R}^{n \times n}$ , so dass  $\underline{B} \underline{\Sigma} \underline{B}^\top = \underline{D}$  diagonal ist.

Dichte in Faktoren zerfällt. Wählen wir  $\underline{B}$  so, dass  $\underline{B}\underline{B}^\top = \underline{\Sigma}$ , so folgt mit dem obigen Argument, dass  $\underline{X} := \underline{\mu} + \underline{B}\underline{Z} \sim N(\underline{\mu}, \underline{\Sigma})$ . Nun schreiben wir

$$\begin{aligned} \mathbf{E}[X_i] &= \mathbf{E}\left[\mu_i + \sum_{k=1}^n B_{ik}Z_k\right] = \mu_i, \\ \mathbf{Cov}[X_i, X_j] &= \mathbf{Cov}\left[\sum_{k=1}^n B_{ik}Z_k, \sum_{l=1}^n B_{jl}Z_l\right] \\ &= \sum_{k=1}^n \sum_{l=1}^n B_{ik}B_{jl}\mathbf{1}_{k=l} = \sum_{k=1}^n B_{ik}B_{jk} = (\underline{B}\underline{B}^\top)_{ij} = \Sigma_{ij}. \end{aligned}$$

□

Wir formulieren eine besonders wichtige Folgerung aus der letzten Proposition.

**Korollar 3.26 (Unabhängigkeit, Unkorreliertheit normalverteilter Zufallsvariablen).** *Sei  $\underline{X}$  ein Vektor  $n$ -dimensional normalverteilter Zufallsvariablen. Dann sind  $X_i, X_j$  genau dann unabhängig, wenn sie unkorreliert sind.*

*Beweis.* Nach Proposition 3.18 folgt aus der Unabhängigkeit auch die Unkorreliertheit. Für die Umkehrung setzen wir zunächst  $\underline{B} = (\underline{e}_i, \underline{e}_j)^\top$ , d.h.  $\underline{B}$  besteht aus den  $i$ -ten und dem  $j$ -ten Einheitsvektor. Nun ist  $\underline{B}\underline{\Sigma}\underline{B}^\top = \begin{pmatrix} \Sigma_{ii} & \Sigma_{ij} \\ \Sigma_{ij} & \Sigma_{jj} \end{pmatrix}$ . Sind  $X_i, X_j$  unkorreliert, so ist diese Matrix eine Diagonalmatrix. Insbesondere zerfällt die Dichte von  $X_i - \mu_i, X_j - \mu_j$  in Faktoren und die Unabhängigkeit folgt. □

### 3.4 Erzeugung von Zufallszahlen

Will man ein stochastischen System numerisch untersuchen bieten sich mehrere Ansätze an. Als Beispiel stelle man sich die (zufällige) Tiefe eines Baumes im Quicksort Algorithmus vor. Die Verteilung dieser Tiefe ist schwer explizit zu bestimmen, jedoch kann man den Quicksort-Algorithmus (mit zufälliger Eingabesequenz) einfach simulieren, und damit auch die Verteilung der Tiefe des Baumes.

Um stochastische Systeme simulieren zu können, ist es notwendig Zufallsvariablen mit bestimmten Verteilungen mittels Zufallsgeneratoren zu ziehen. Während wir hier nicht über die Güte von Zufallsgeneratoren diskutieren wollen, stellen wir fest, dass diese meist uniform auf  $[0, 1]$  verteilte Zufallszahlen liefern. Ausgehen von diesen kann man mittels des nächsten Satzes nach beliebigen Verteilungen verteilte Zufallszahlen generieren. Eine Illustration hierzu findet sich in Abbildung 3.1.

**Theorem 3.27 (Simulationslemma).** *Sei  $X$  eine (diskrete oder stetige) reellwertige Zufallsvariable mit Verteilungsfunktion  $F$ . Wir definieren die Pseudoinverse von  $F$  für  $x \in [0, 1]$  durch*

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

*Dann ist  $F^{-1}(U)$  genauso verteilt wie  $X$ .*

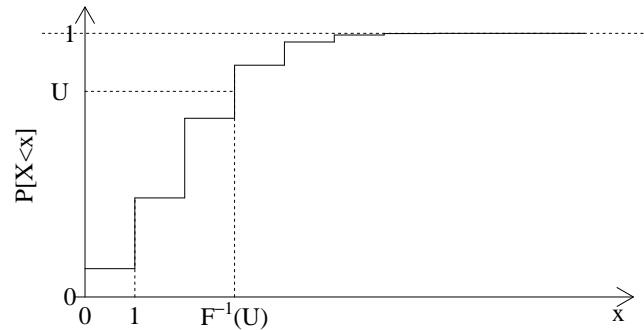


Abbildung 3.1: Illustration zum Simulationslemma, Theorem 3.27.

*Beweis.* Wir verwenden, dass die Verteilungsfunktion  $F$  die Verteilung der Zufallsvariable  $X$  eindeutig bestimmt. Da  $F$  nicht notwendigerweise injektiv ist, muss  $F^{-1}$  nicht die Umkehrfunktion von  $F$  sein. Es gilt jedoch wegen der Konstruktion  $F^{-1}(q) \leq x$  genau dann, wenn  $q \leq F(x)$ . Daraus folgt

$$\mathbf{P}[F^{-1}(U) \leq q] = \mathbf{P}[U \leq F(q)] = F(q).$$

Das bedeutet, dass  $F^{-1}(U)$  die Verteilungsfunktion  $F$  hat, woraus die Aussage folgt.  $\square$

**Bemerkung 3.28 (Anwendung).** Angenommen, von einer Verteilung ist die Verteilungsfunktion  $F$  (und damit die Pseudoinverse  $F^{-1}$ ) bekannt. Wir wollen unabhängige Zufallsvariablen  $X_1, X_2, \dots$  nach dieser Verteilung erzeugen. Verwenden dürfen wir hierzu jedoch nur die vom Computer bereit gestellten uniform verteilten Zufallsvariablen  $U_1, U_2, \dots$ . Das Simulationslemma besagt, dass in diesem Fall  $F^{-1}(U_1), F^{-1}(U_2), \dots$  genau die gewünschte Verteilung besitzen.

**Bemerkung 3.29 (Alternativen).** Während das Verfahren aus Theorem 3.27 allgemeingültig ist, gibt es für spezielle Verteilungen schnellere Verfahren, Zufallszahlen zu erzeugen. Wir erwähnen hier (ohne Beweis) nur eine, nämlich die Erzeugung normalverteilter Zufallszahlen. Hierzu seien  $U_1, U_2$  zwei unabhängige Zufallszahlen. Setzt man nun

$$X := \cos(2\pi U_1) \sqrt{-2 \log(U_2)},$$

so ist  $X \sim N(0, 1)$ .

## 4 Approximationssätze

Wir beschäftigen uns nun mit der Verteilung der Summe unabhängiger und identisch verteilter Zufallsvariablen. Dies ist eine Situation, die in der Praxis häufig vorkommt, etwa wenn man die Anzahl der Gewinne in einem Spiel zählt. Für unabhängige, identisch verteilte Zufallsvariable  $X_1, X_2, \dots$  sei nun  $Y_n := X_1 + \dots + X_n$ . Selbst bei bekannter Verteilung der  $X_i$  ist meistens die Verteilung von  $Y$  nicht bekannt. Allerdings gilt (falls Erwartungswert bzw. Varianz von

$X_1$  existieren)

$$\begin{aligned}\mathbf{E}[Y_n] &= \mathbf{E}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbf{E}[X_i] = n\mathbf{E}[X_1], \\ \mathbf{Var}[Y_n] &= \mathbf{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbf{Var}[X_i] = n\mathbf{Var}[X_1].\end{aligned}\tag{4.1}$$

Geht man also davon aus, dass die typische Schwankung der Zufallsvariable  $Y_n$  in etwa ihrer Standardabweichung entspricht, sieht man das  $\sqrt{n}$ -Gesetz:  $Y_n$  streut ihre Werte typischerweise in einem Bereich der Größenordnung  $\sqrt{n}$ .

#### 4.1 Das Gesetz der großen Zahlen

Das schwache Gesetz der großen Zahlen ist eine Konvergenzaussage für den Mittelwert von unabhängigen, identisch verteilten Zufallsgrößen. Als Beispiel betrachte man die relative Häufigkeit  $\frac{1}{n} \sum_{i=1}^n X_i$  der Köpfe in einem  $p$ -Münzwurf  $X_1, X_2, \dots$ . Intuitiv klar ist, dass

$$\frac{1}{n} \sum_{i=1}^n X_i \approx p$$

gelten sollte. (Die relative Häufigkeit der Köpfe entspricht also in etwa der Wahrscheinlichkeit Kopf zu werfen.) In welchem Sinne diese Konvergenz zu verstehen ist, werden wir hier zwar nicht beleuchten können, das schwache Gesetz der großen Zahlen formalisiert diese Aussage jedoch. Eine Illustration findet sich in Abbildung 4.1. Zunächst benötigen wir zwei wichtige Ungleichungen.

**Proposition 4.1 (Markov- und Tschebychev-Ungleichung).** *Sei  $X$  eine  $\mathbb{R}_+$ -wertige Zufallsvariable. Dann gilt für alle  $\varepsilon > 0$  die Markov-Ungleichung*

$$\mathbf{P}[X \geq \varepsilon] \leq \frac{1}{\varepsilon} \mathbf{E}[X].$$

*Existiert  $\mu := \mathbf{E}[X]$ , so gilt außerdem für  $\varepsilon > 0$  die Tschebychev-Ungleichung*

$$\mathbf{P}[|X - \mu| \geq \varepsilon] \leq \frac{\mathbf{Var}[X]}{\varepsilon^2}.$$

*Beweis.* Zum Beweis der Markov-Ungleichung bemerken wir  $X \geq \varepsilon \cdot 1_{X \geq \varepsilon}$ . Damit ist wegen der Monotonie des Erwartungswertes

$$\varepsilon \cdot \mathbf{P}[X \geq \varepsilon] = \mathbf{E}[\varepsilon \cdot 1_{X \geq \varepsilon}] \leq \mathbf{E}[X].$$

Die Tschebychev-Ungleichung folgt nun aus der Markov-Ungleichung, wenn man die Zufallsvariable  $(X - \mu)^2$  betrachtet.  $\square$

**Theorem 4.2 (Schwaches Gesetz großer Zahlen).** *Seien  $X_1, X_2, \dots$  reellwertige, unabhängige, identisch verteilte Zufallsvariable mit endlichem Erwartungswert  $\mathbf{E}[X_1] = \mu$  und endlicher Varianz. Dann gilt für alle  $\varepsilon > 0$*

$$\lim_{n \rightarrow \infty} \mathbf{P}\left[\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| \geq \varepsilon\right] = 0.$$



*Beweis.* Der Beweis erfolgt mittels der Chebyshev-Ungleichung und (4.1), denn

$$\mathbf{P}\left[\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right] \leq \frac{\mathbf{Var}\left[\frac{1}{n}(X_1 + \cdots + X_n)\right]}{\varepsilon^2} = \frac{\mathbf{Var}[X_1]}{n\varepsilon} \xrightarrow{n \rightarrow \infty} 0.$$

□

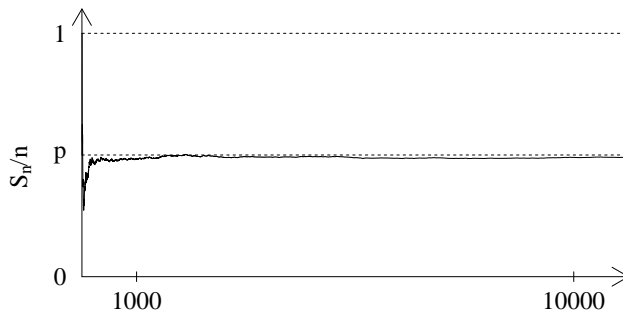


Abbildung 4.1: Illustration zum Gesetz der großen Zahlen, Theorem 4.2.

**Beispiel 4.3 (Anzahl der Vergleiche bei QUICKSORT).** Eine Art Gesetz großer Zahlen kann auch bei abhängigen Zufallsvariablen auftreten. Bei QUICKSORT (mit zufälligem Input  $\Sigma \in \mathcal{S}_n$ ) haben wir die Anzahl der Vergleiche (was wir proportional zur Laufzeit sehen) als

$$h(\Sigma) = \sum_{i < j} \gamma_{ij}(\Sigma)$$

geschrieben. Die Zufallsvariablen  $\gamma_{ij}(\Sigma)$   $i < j$  haben nun weder die gleiche Verteilung (siehe etwa deren Erwartungswert aus Beispiel 3.9, der von  $j - i$  abhängt), noch sind sie unabhängig (sonst wären ja alle Kovarianzen 0, was laut den Berechnungen aus Beispiel 3.20 nicht möglich ist). Trotzdem können wir mit der Tschebychev-Ungleichung schreiben

$$\begin{aligned} \mathbf{P}\left[\left|\frac{h(\Sigma)}{\mathbf{E}[h(\Sigma)]} - 1\right| > \varepsilon\right] &\leq \frac{1}{\varepsilon^2} \frac{\mathbf{V}[h(\Sigma)]}{\mathbf{E}[h(\Sigma)]^2} \leq \frac{1}{\varepsilon^2} \frac{3n(n-1)}{\Omega(4n^2(\log n)^2)} \\ &= \frac{1}{\varepsilon^2} \Omega\left(\frac{3}{4(\log n)^2}\right) \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

## 4.2 Der zentrale Grenzwertsatz und die Normal-Verteilung

Während das Gesetz der großen Zahlen eine Aussage über die Konvergenz des Mittelwertes von unabhängigen, identisch verteilten Zufallsvariablen  $X_1, X_2, \dots$  gegen deren Erwartungswert trifft, beschäftigt sich der zentrale Grenzwertsatz mit den Schwankungen in dieser Konvergenz. Da  $\mathbf{Var}\left[\frac{1}{n}(X_1 + \cdots + X_n)\right] = \mathbf{Var}X_1/n$ , eine typische Schwankung also von der Größenordnung  $1/\sqrt{n}$  ist, werden wir hier

$$\frac{\frac{1}{n} \sum_{i=1}^n X_i - \mathbf{E}[X_i]}{\sqrt{n \mathbf{Var}[X_1]}}$$

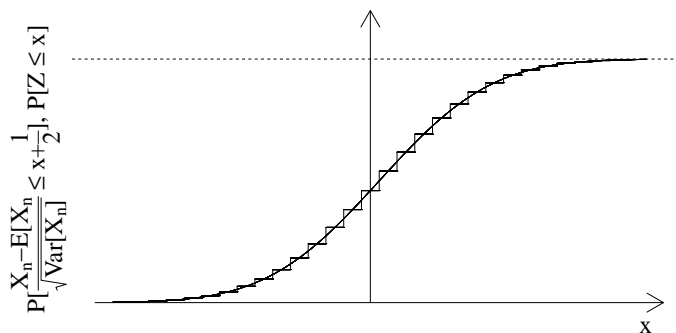
betrachten. Obwohl es möglich ist, ein allgemeines Resultat zu beweisen, beschränken wir uns auf den Fall von  $B(1, p)$ -verteilten Zufallsgrößen  $X_1, X_2, \dots$  und bemerken, dass  $X_1 + \dots + X_n \sim B(n, p)$ . Ohne Beweis werden wir außerdem die Stirling Formel zur Approximation von Fakultäten,

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

verwenden. Hierbei bedeutet  $\approx$ , dass

$$\frac{n!}{\left(\frac{n}{e}\right)^n \sqrt{2\pi n}} \xrightarrow{n \rightarrow \infty} 1.$$

Eine Illustration des Satzes findet sich in Abbildung 4.2.



**Abbildung 4.2:** Illustration zum Satz von deMoivre-Laplace, Proposition 4.4. Gezeigt sind jeweils die Verteilungsfunktionen von  $N(0, 1)$  (die glatte Kurve) und eine Transformation von  $B(100, 0.5)$ .

**Proposition 4.4 (Satz von deMoivre-Laplace).** Sei  $Z$  eine  $N(0, 1)$  verteilte Zufallsvariable und  $X_1, X_2, \dots$  eine Folge binomialverteilter Zufallsvariable mit  $\mathbf{Var}[X_n] \xrightarrow{n \rightarrow \infty} \infty$ . Dann ist für  $-\infty \leq c < d \leq \infty$

$$\mathbf{P}\left[c \leq \frac{X_n - \mathbf{E}[X_n]}{\sqrt{\mathbf{Var}[X_n]}} \leq d\right] \xrightarrow{n \rightarrow \infty} \mathbf{P}[c \leq Z \leq d].$$

*Beweisskizze.* Mit der Stirling-Formel und  $\eta(t) = t \ln \frac{t}{p} + (1-t) \ln \frac{1-t}{q}$  ist

$$\begin{aligned} \mathbf{P}[X = k] &= \binom{n}{k} p^k q^{n-k} \approx \sqrt{\frac{n}{2\pi k(n-k)}} \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k} \\ &= \frac{1}{\sqrt{2\pi n \frac{k}{n} \frac{n-k}{n}}} \exp\left(-n\eta\left(\frac{k}{n}\right)\right) \\ &\approx \frac{1}{\sqrt{2\pi n \frac{k}{n} \frac{n-k}{n}}} \exp\left(-\frac{1}{2} \left(\frac{k-np}{\sqrt{npq}}\right)^2\right), \end{aligned}$$

da  $\eta(p) = \eta'(p) = 0, \eta''(p) = \frac{1}{pq}$  mit einer Taylor-Entwicklung von  $\eta$  um  $p$ . Damit gilt für  $\mu_n = np_n, \sigma_n^2 = np_nq_n$  und  $z_{x,n} = \frac{x-\mu_n}{\sqrt{\sigma_n^2}}$

$$\begin{aligned} \mathbf{P}\left(c \leq \frac{X_n - \mathbf{E}[X_n]}{\sqrt{\mathbf{Var}[X_n]}} \leq d\right) &= \sum_{k:c \leq z_{k,n} \leq d} \mathbf{P}[X_n = k] \\ &\approx \sum_{k:c \leq z_{k,n} \leq d} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{z_{k,n}^2}{2}\right) \approx \frac{1}{\sqrt{2\pi}} \int_c^d e^{-\frac{z^2}{2}} dz, \end{aligned}$$

wobei wir die letzte Summe als Riemannsumme interpretiert haben.  $\square$

**Bemerkung 4.5 (Anwendung).** Sei  $X$  eine  $B(n, p)$  verteilte Zufallsgröße mit  $n$  groß,  $npq$  groß, sowie  $c, d \in \mathbb{Z}$ . Dann gilt

$$\begin{aligned} \mathbf{P}(c \leq X \leq d) &\approx \sum_{k=c}^d \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_{k,n}^2}{2}\right) (z_{k+\frac{1}{2},n} - z_{k-\frac{1}{2},n}) \\ &\approx \frac{1}{\sqrt{2\pi}} \int_{z_{c-1/2,n}}^{z_{d+1/2,n}} \exp\left(-\frac{z^2}{2}\right) dz = \Phi\left(\frac{d+\frac{1}{2}-np}{\sqrt{npq}}\right) - \Phi\left(\frac{c-\frac{1}{2}-np}{\sqrt{npq}}\right). \end{aligned}$$

Man beachte hierbei jeweils die Verschiebung um  $\frac{1}{2}$  in der Binomialverteilung. Mit ihr erreicht man eine bessere Approximationsgüte. Diese Korrektur wurde etwa auch in Abbildung 4.2 angewandt.

**Bemerkung 4.6 (Zentraler Grenzwertsatz).** Allgemeiner als der Satz von deMoivre-Laplace gilt auch der Zentraler Grenzwertsatz:

Seien  $X_1, X_2, \dots$  reellwertige, unabhängige, identisch verteilte Zufallsvariablen mit endlichem Erwartungswert  $\mathbf{E}[X_1] = \mu$  und Varianz  $\mathbf{Var}[X_1] = \sigma^2$ . Dann gilt für

$$Y_n^* := \frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}}$$

und alle  $-\infty \leq c < d \leq \infty$

$$\lim_{n \rightarrow \infty} \mathbf{P}[c \leq Y_n^* \leq d] = \mathbf{P}[c \leq Z \leq d],$$

wobei  $Z$  eine nach  $N(0, 1)$  verteilte Zufallsvariable ist.

## 5 Bedingte Wahrscheinlichkeiten

### 5.1 Grundlegendes

**Definition 5.1 (Bedingte Wahrscheinlichkeit).** Seien  $X$  und  $Y$  Zufallsvariablen.

1. Die bedingte Wahrscheinlichkeit von  $\{Y \in B\}$ , gegeben  $\{X \in A\}$ , ist durch

$$\mathbf{P}[Y \in B | X \in A] := \frac{\mathbf{P}[X \in A, Y \in B]}{\mathbf{P}[X \in A]}$$

gegeben. Ist  $\mathbf{P}[X \in A] = 0$ , so definieren wir die rechte Seite als 0.

2. Ist  $X$  diskret, so ist die bedingte Verteilung von  $Y$  gegeben  $X$  die Abbildung

$$\mathbf{P}[Y \in B|X] : x \mapsto \mathbf{P}[Y \in B|X = x].$$

Hat  $(X, Y)$  eine gemeinsame Dichte  $f(x, y)dxdy$ , so ist die bedingte Verteilung von  $Y$  gegeben  $X$  die Verteilung mit Dichte  $\frac{f(x, y)dy}{\int f(x, z)dz}$ , d.h.

$$\mathbf{P}[Y \in B|X] : x \mapsto \frac{\int_B f(x, y)dy}{\int f(x, z)dz}.$$

**Beispiel 5.2 ( $p$ -Münzwurf bedingt auf Anzahl der Erfolge).** Sei  $\underline{X} = (X_1, \dots, X_n)$  ein  $p$ -Münzwurf und  $S = X_1 + \dots + X_n$ . Die bedingte Verteilung von  $\underline{X}$  gegeben  $S$  ist für  $x_1 + \dots + x_n = k$

$$\mathbf{P}[\underline{X} = \underline{x}|S = k] = \frac{p^k(1-p)^{n-k}}{\binom{n}{k}p^k(1-p)^{n-k}} = \frac{1}{\binom{n}{k}}.$$

Damit ist  $\underline{X}$  gegeben  $S$  uniform auf allen  $k$ -elementigen Teilmengen von  $\{1, \dots, n\}$  verteilt.

**Beispiel 5.3 (Minimum und Maximum zweier uniform-verteilter Zufallsvariablen).** Seien  $U, V$  unabhängig und  $U([0, 1])$ -verteilt. Die Zufallsvariablen  $X = U \wedge V$  und  $Y = U \vee V$  haben die gemeinsame Dichte  $2 \cdot 1_{0 \leq x \leq y \leq 1}$ . Damit ist die Dichte der bedingten Verteilung von  $X$  gegeben  $Y = y$  durch

$$\frac{2 \cdot 1_{0 \leq x \leq y \leq 1}}{2 \int 1_{0 \leq x \leq y \leq 1} dx} = \frac{1}{y} 1_{0 \leq x \leq y}$$

gegeben. Dies ist aber gerade die Dichte eine  $U([0, y])$ -Verteilung.

**Lemma 5.4 (Einfache Eigenschaften bedingter Wahrscheinlichkeiten).** Seien  $X$  und  $Y$  Zufallsvariablen und  $A, B$  Mengen.

1. Ist  $Y$  diskret, so gilt

$$\mathbf{P}[Y \in B|X \in A] = \sum_{y \in B} \mathbf{P}[Y = y|X \in A].$$

2. Die Zufallsvariablen  $X$  und  $Y$  sind genau dann unabhängig, wenn

$$\mathbf{P}[Y \in B|X \in A] = \mathbf{P}[Y \in B]$$

für alle  $A, B$  gilt. Dies ist genau dann der Fall, wenn

$$\mathbf{P}[Y \in B|X] = \mathbf{P}[Y \in B]$$

für alle  $B$  gilt.

*Beweis.* Beide Eigenschaften folgen aus der Definition der bedingten Wahrscheinlichkeit.  $\square$

**Theorem 5.5 (Formel für die totalen Wahrscheinlichkeit und Bayes'sche Formel).** Seien  $X, Y$  diskrete Zufallsvariablen. Dann gilt die Formel von der totalen Wahrscheinlichkeit

$$\mathbf{P}[Y \in B] = \sum_x \mathbf{P}[Y \in B|X = x] \cdot \mathbf{P}[X = x]$$

für jede Menge  $B$ . Weiter gilt die Bayes'sche Formel für jede Menge  $A$  mit  $\mathbf{P}[X \in A] > 0$

$$\mathbf{P}[X \in A|Y \in B] = \frac{\mathbf{P}[Y \in B|X \in A] \cdot \mathbf{P}[X \in A]}{\sum_x \mathbf{P}[Y \in B|X = x] \cdot \mathbf{P}[X = x]}.$$

*Beweis.* Die Formel von der totalen Wahrscheinlichkeit ergibt sich durch Einsetzen der Definition von  $\mathbf{P}[Y \in B|X = x]$  in

$$\mathbf{P}[Y \in B] = \sum_x \mathbf{P}[Y \in B, X = x].$$

In der Bayes'schen Formel ist der Zähler der rechten Seite gleich  $\mathbf{P}[X \in A, Y \in B]$  nach der Definition der bedingten Wahrscheinlichkeit, und der Nenner ist

$$\sum_x \mathbf{P}[Y \in B|X = x] \cdot \mathbf{P}[X = x] = \sum_x \mathbf{P}[Y \in B, X = x] = \mathbf{P}[Y \in B].$$

□

**Beispiel 5.6 (Reihenuntersuchungen).** In einer Reihenuntersuchung werden Personen auf eine bestimmte Krankheit getestet. Dabei kann es fälschlicherweise vorkommen, dass gesunde Personen durch den Test als krank eingestuft werden, oder dass kranke Personen nicht als solche erkannt werden.

In einer Population sind insgesamt 0.8% der Personen erkrankt. Eine kranke Person wird in 90% der Fälle positiv getestet (der Test fällt also positiv aus), eine gesunde Person mit 7%.

Wie groß ist die Wahrscheinlichkeit, dass eine positiv getestete Person wirklich krank ist?

Um dies zu beantworten, setzen wir  $X = 1$  wenn die Person erkrankt ist (sonst  $X = 0$ ) und  $Y = 1$  wenn die Person positiv getestet wird (sonst  $Y = 0$ ). Die genannten Angaben übersetzen wir zu

$$\mathbf{P}[X = 1] = 0.008, \quad \mathbf{P}[Y = 1|X = 1] = 0.9, \quad \mathbf{P}[Y = 1|X = 0] = 0.07.$$

Mit der Formel von Bayes ergibt sich für die Wahrscheinlichkeit, dass eine positiv getestete Person wirklich erkrankt ist

$$\begin{aligned} \mathbf{P}[X = 1|Y = 1] &= \frac{\mathbf{P}[Y = 1|X = 1] \cdot \mathbf{P}[X = 1]}{\mathbf{P}[Y = 1|X = 1] \cdot \mathbf{P}[X = 1] + \mathbf{P}[Y = 1|X = 0] \cdot \mathbf{P}[X = 0]} \\ &= \frac{0.9 \cdot 0.008}{0.9 \cdot 0.008 + 0.07 \cdot 0.992} \approx 0.0939 \end{aligned}$$

Ein positiver Test ist also nicht unbedingt ein sicheres Zeichen dafür, ob eine Person erkrankt ist.

Situationen dieser Art lassen sich oft auch durch einen Wahrscheinlichkeitsbaum darstellen. Siehe Abbildung 5.1.

**Definition 5.7 (Bedingte Erwartung).** Seien  $X$  und  $Y$  Zufallsvariablen und  $B \mapsto \mathbf{P}[Y \in B|X]$  die bedingte Verteilung von  $Y$  gegeben  $X$ . Für eine Funktion  $h$  heißt dann der Erwartungswert von  $h(Y)$  bezüglich dieser bedingten Verteilung auch die bedingte Verteilung von  $h(Y)$  gegeben  $X$ . Wir schreiben hierfür im Falle diskreter Zufallsvariablen

$$\mathbf{E}[h(Y)|X] = \sum_y h(y)\mathbf{P}[Y = y|X],$$

beziehungsweise im Falle von stetigen Zufallsvariablen

$$\mathbf{E}[h(Y)|X] = \frac{\int h(y)f(X, y)dy}{\int f(X, y)dy},$$

falls die Erwartungswerte existieren.

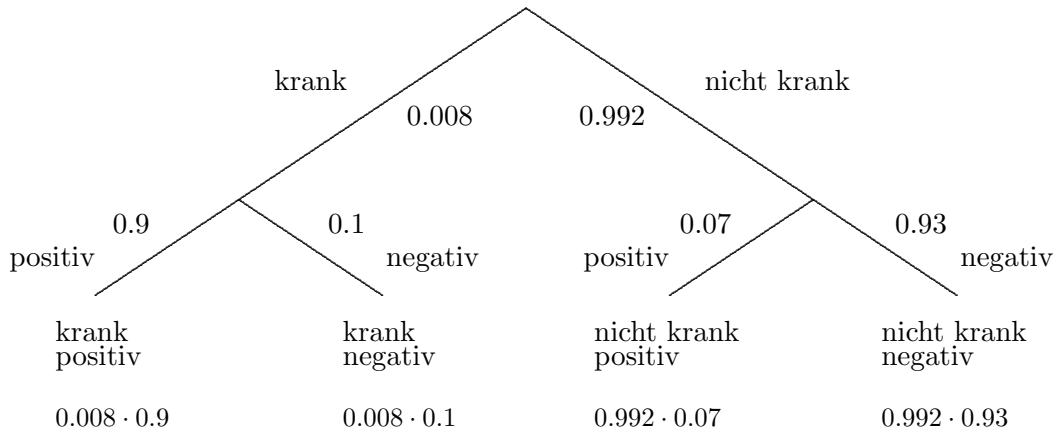


Abbildung 5.1: Ein Wahrscheinlichkeitsbaum für das Beispiel 5.6.

**Proposition 5.8 (Turmeigenschaft).** Seien  $X$  und  $Y$  Zufallsvariable und  $h$  eine Funktion. Dann gilt

$$\mathbf{E}[\mathbf{E}[h(Y)|X]] = \mathbf{E}[h(Y)].$$

*Beweis.* Wir führen den Beweis im Falle von diskreten Zufallsvariablen. Es gilt

$$\begin{aligned} \mathbf{E}[\mathbf{E}[h(Y)|X]] &= \sum_x \mathbf{E}[h(Y)|X = x] \cdot \mathbf{P}[X = x] = \sum_{x,y} h(y) \cdot \mathbf{P}[Y = y|X = x] \cdot \mathbf{P}[X = x] \\ &= \sum_{x,y} h(y) \cdot \mathbf{P}[Y = y, X = x] = \sum_y h(y) \cdot \mathbf{P}[Y = y] = \mathbf{E}[h(Y)]. \end{aligned}$$

□

**Beispiel 5.9 (Hashing).** Es werden  $n$  Objekte (Schlüssel) mittels einer Hash-Funktion  $r$  Hashes zugeordnet. Dabei ist der Schlüssel mit Wahrscheinlichkeit  $p_r$  so, dass er Hash  $r$  zugeordnet wird. Wir wählen nun rein zufällig eines der  $n$  Objekte aus. Mit wievielen Vergleichen  $X$  finden wir es mittels der Hash-Funktion? Zunächst bestimmen wir den Hash  $r$  und müssen dann alle Objekte mit diesem Hash vergleichen, bis wir das gesuchte Objekt finden. Wir werden zeigen, dass

$$\mathbf{E}[X] = \frac{n-1}{2} \sum_{j=1}^r p_j^2.$$

Seien  $\underline{Z} := (Z_1, \dots, Z_r)$  (mit  $Z_1 + \dots + Z_r = n$ ) die Anzahl der Objekte mit Hash  $1, \dots, r$ . Weiter sei  $X$  die gesuchte Anzahl der Vergleiche. Wir berechnen

$$\mathbf{P}[X = x|\underline{Z}] = \sum_{j=1}^r \frac{Z_j}{n} 1_{Z_j > x} \frac{1}{Z_j} = \frac{1}{n} \sum_{j=1}^r 1_{Z_j > x}.$$

Damit ist

$$\mathbf{E}[X|\underline{Z}] = \frac{1}{n} \sum_{j=1}^r \sum_x x 1_{Z_j > x} = \frac{1}{n} \sum_{j=1}^r \frac{Z_j(Z_j - 1)}{2}$$

Daraus können wir nun mit Hilfe der Turmeigenschaft den Mittelwert von  $X$  berechnen. Wir schreiben mit Proposition 3.12, da  $Z_j \sim B(n, p_j)$ ,

$$\begin{aligned}\mathbf{E}[X] &= \mathbf{E}[\mathbf{E}[X|\underline{Z}]] = \frac{1}{2n} \sum_{j=1}^r \mathbf{E}[Z_j(Z_j - 1)] = \frac{1}{2n} \sum_{j=1}^r np_j(1 - p_j) - np_j(1 - np_j) \\ &= \frac{1}{2n} \sum_{j=1}^r n(n-1)p_j^2 = \frac{n-1}{2} \sum_{j=1}^r p_j^2.\end{aligned}$$

**Beispiel 5.10 (Finden eines Elements in einem Binary Search Tree).** Eine zufällige Permutation der Zahlen  $1, \dots, n$  wird in einen Binary Search Tree (BST) einsortiert. Nun wählen wir eine der Zahlen  $1, \dots, n$  rein zufällig aus und suchen diese im BST. Wie viele Vergleiche müssen wir hier im Mittel machen, bis wir die Zahl gefunden haben?

Sei  $\Sigma$  die zufällige Permutation der Zahlen  $1, \dots, n$  und  $X_n$  die zufällige Anzahl der Vergleiche, die man zum Finden des zufälligen Elements benötigt. Zunächst wird  $\Sigma$  in den BST einsortiert, also wird  $\Sigma(1)$  zur Wurzel des BST gemacht. Alle Zahlen  $1, \dots, \Sigma(1) - 1$  werden unter das linke Kind der Wurzel, alle Zahlen  $\Sigma(1) + 1, \dots, n$  werden unter das rechte Kind der Wurzel eingehängt. Bei der Wahl der neuen Zahl ist es nun so, dass diese mit Wahrscheinlichkeit  $(\Sigma(1) - 1)/n$  im linken und mit Wahrscheinlichkeit  $(n - \Sigma(1))/n$  im rechten Teilbaum liegt. Im ersten Fall ist die Anzahl der Vergleiche  $X_n$  dieselbe wie die bei einem BST der Zahlen  $1, \dots, \Sigma(1) - 1$ , also so verteilt wie  $X_{\Sigma(1)-1}$ . Bedingen wir nun auf  $\Sigma(1)$ , so ergibt sich für die  $X_n$

$$\begin{aligned}\mathbf{E}[X_n] &= \mathbf{E}[\mathbf{E}[X_n|\Sigma(1)]] = 1 + \mathbf{E}\left[\frac{\Sigma(1) - 1}{n} X_{\Sigma(1)-1} + \frac{n - \Sigma(1)}{n} X_{n-\Sigma(1)}\right] \\ &= 1 + \frac{1}{n} \sum_{i=1}^n \frac{i-1}{n} \mathbf{E}[X_{i-1}] + \frac{n-i}{n} \mathbf{E}[X_{n-i}] \\ &= 1 + \frac{2}{n^2} \sum_{i=1}^{n-1} i \mathbf{E}[X_i]\end{aligned}$$

Hieraus ergibt sich

$$\begin{aligned}n^2 \mathbf{E}[X_n] - (n-1)^2 \mathbf{E}[X_{n-1}] &= 2n - 1 + 2(n-1) \mathbf{E}[X_{n-1}], \\ n^2 \mathbf{E}[X_n] &= 2n - 1 + (n-1)(n+1) \mathbf{E}[X_{n-1}], \\ \frac{n}{n+1} \mathbf{E}[X_n] - \frac{n-1}{n} \mathbf{E}[X_{n-1}] &= \frac{2n-1}{n(n+1)}.\end{aligned}$$

Summieren wir die letzte Gleichung bis  $n$ , so ergibt sich

$$\frac{n}{n+1} \mathbf{E}[X_n] = \sum_{k=1}^n \frac{2k-1}{k(k+1)}.$$

Für große  $n$  ist damit

$$\mathbf{E}[X_n] = \Omega(2 \log n).$$

## 5.2 Ein Ausflug in die Bayes'sche Statistik

Nehmen wir an, wir werfen eine Münze, deren Erfolgswahrscheinlichkeit  $p$  wir noch nicht kennen. Da wir uns völlig unklar darüber sind, wie groß  $p$  ist, sagen wir einfach, dies sei eine Zufallsvariable  $P$ , die uniform aus  $[0, 1]$  verteilt ist. Werfen wir nun mit der Münze und zählen die Anzahl  $X$  der Erfolge, lernen wir etwas über die Erfolgswahrscheinlichkeit. Wenn wir nun in  $k$  aus  $n$  Würfen einen Erfolg erzielt haben, wie beeinflusst das dann unseren ursprünglichen Glauben  $P \sim U([0, 1])$  an die Erfolgswahrscheinlichkeit?

**Proposition 5.11 (Konjugierte Verteilung).** *Sei  $P \sim \text{Beta}(\alpha, \beta)$  und (gegeben  $P$ ) sei  $X \sim B(n, P)$ . Dann ist die bedingte Verteilung von  $P$  gegeben  $X$  eine  $\text{Beta}(\alpha + k, \beta + n - k)$ -Verteilung, also*

$$\mathbf{P}(P \in dp | X = k) = \binom{\alpha + \beta + n + 1}{\alpha + k} (\alpha + k) p^{\alpha + k - 1} (1 - p)^{\beta + n - k - 1} dp.$$

Insbesondere ist

$$\mathbf{E}[P | X = k] = \frac{\alpha + k}{\alpha + \beta + n}, \quad \mathbf{V}[P | X = k] = \frac{(\alpha + k)(\beta + n - k)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}.$$

*Beweis.* Wir müssen den Satz von Bayes anwenden und schreiben (mit  $\sim$  als Proportionalitätszeichen)

$$\mathbf{P}(P \in dp | X = k) \sim \binom{n}{k} p^{\alpha - 1 + k} (1 - p)^{\beta - 1 + n - k} dp.$$

Daraus folgt bereits die erste Aussage. Erwartungswert und Varianz haben wir schon in Lemma 3.14 berechnet.  $\square$

**Bemerkung 5.12 (Interpretation).** Für große  $n$  berechnen wie mit Hilfe der Tchebychev-Ungleichung

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbf{P}[|P - X/n| > \varepsilon | X] &\leq \lim_{n \rightarrow \infty} \mathbf{P}\left[|P - \frac{\alpha + X}{\alpha + \beta + n}| > \varepsilon/2 | X\right] \leq \frac{4\mathbf{V}[P | X]}{\varepsilon^2} \\ &= \lim_{n \rightarrow \infty} \frac{4(\alpha + X)(\beta + n - X)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} = 0. \end{aligned}$$

Diese Konvergenz ist auch nicht erstaunlich, weil ja die Erfolgswahrscheinlichkeit  $p$  beim Münzwurf immer um  $X/n$  konzentriert ist.

## 5.3 Markov-Ketten

Wir betreten nun den Bereich der zeitdiskreten stochastischen Prozesse. Dies ist eine Familie  $X_0, X_1, X_2, \dots$  von Zufallsvariablen. Markov-Ketten zeichnen sich durch eine bestimmte Abhängigkeitsstruktur zwischen den Zufallsvariablen aus. Hier werden der Reihe nach Zufallsvariablen  $X_0, X_1, \dots$  realisiert, und zwar so, dass  $X_{t+1}$  nur von  $X_t$  abhängt. Man sagt auch,  $X_{t+1}$  ist unabhängig von  $X_1, \dots, X_{t-1}$ , wenn  $X_t$  bekannt ist. Diese Form der Abhängigkeit wird häufig zur stochastischen Modellierung verwendet. Beispielsweise könnte  $X_0, X_1, \dots$  der Preis einer Aktie an Tagen  $0, 1, \dots$  sein. Die Markov-Eigenschaft besagt in diesem Fall, dass



die Verteilung der Kursänderungen am Tag  $t + 1$  nur davon abhängt, wie der Kurs  $X_t$  am Tag  $t$  war.

Wir beginnen mit der Definition von Markov-Ketten. Wir werden vor allem den Fall von zeitlich homogenen Markov-Ketten behandeln. Bei solchen gibt es eine sich zeitlich nicht ändernde stochastische Übergangsvorschrift, wie die Verteilung des Zustandes  $X_{t+1}$  ist, wenn der Zustand  $X_t$  der Kette zur Zeit  $t$  bekannt ist. Diese Vorschrift wird mit Hilfe einer Matrix zusammengefasst, der Übergangsmatrix.

**Definition 5.13 (Stochastischer Prozess und Markov-Ketten).** 1. Seien  $I$  und  $E$  Mengen. Ein ( $E$ -wertiger) stochastischer Prozess (mit Indexmenge  $I$ ) ist eine Familie von Zufallsvariablen  $\mathcal{X} = (X_t)_{t \in I}$  mit Wertebereich  $E$ . Die Menge  $E$  heißt auch Zustandsraum von  $\mathcal{X}$  und  $I$  seine Indexmenge.

2. Sei  $I = \{0, 1, 2, \dots\}$ ,  $E$  abzählbar und  $\mathcal{X} = (X_t)_{t \in I}$  ein  $E$ -wertiger stochastischer Prozess. Falls

$$\mathbf{P}[X_{t+1} = i \mid X_0, \dots, X_t] = \mathbf{P}[X_{t+1} = i \mid X_t] \quad (5.1)$$

für alle  $i \in E$ , so heißt  $\mathcal{X}$  eine Markov-Kette. Sie heißt endlich, falls  $E$  endlich ist.

3. Sei  $\mathcal{X} = (X_t)_{t \in I}$  eine  $E$ -wertige Markov-Kette. Existiert eine Matrix  $P = (P_{ij})_{i,j \in E}$  mit

$$P_{ij} := \mathbf{P}[X_{t+1} = j \mid X_t = i]$$

(unabhängig von  $t$ ), so heißt  $\mathcal{X}$  zeitlich homogen und  $P$  heißt Übergangsmatrix von  $\mathcal{X}$ .

**Bemerkung 5.14 (Interpretation).** 1. Die Eigenschaft (5.1) bedeutet in Worten: die zukünftige Entwicklung von  $\mathcal{X}$  nach  $t$  hängt von  $X_1, \dots, X_t$  nur durch den aktuellen Zustand  $X_t$  ab.

2. Sei  $P$  die Übergangsmatrix einer homogenen Markov-Kette  $\mathcal{X}$  mit Zustandsraum  $E$ . Dann gilt

$$\begin{aligned} 0 \leq P_{ij} \leq 1, & \quad i, j \in E, \\ \sum_{j \in E} P_{ij} = 1, & \quad i \in E. \end{aligned} \quad (5.2)$$

Die erste Eigenschaft ist klar, da die Einträge in  $P$  Wahrscheinlichkeiten sind. Außerdem ist

$$1 = \mathbf{P}[X_{t+1} \in E \mid X_t = i] = \sum_{j \in E} \mathbf{P}[X_{t+1} = j \mid X_t = i] = \sum_{j \in E} P_{ij}.$$

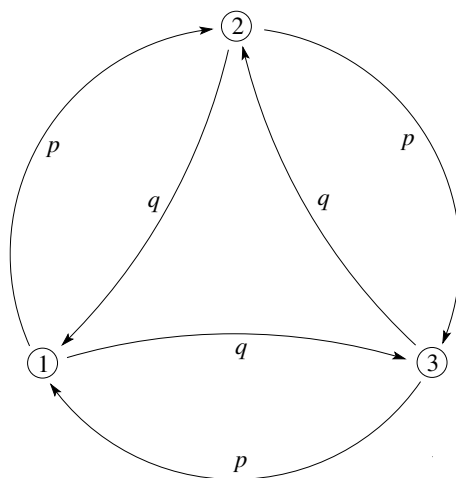
Matrizen  $P$  mit den Eigenschaften (5.2) heißen *stochastische Matrizen*.

3. Sei  $\mathcal{X}$  eine homogene Markov-Kette mit Übergangsmatrix  $P$ . Definiere einen gewichteten, gerichteten Graphen  $(E, K, W)$  wie folgt: die Menge der Knoten ist  $E$ , die Menge der (gerichteten) Kanten ist  $K := \{(i, j) : P_{ij} > 0\}$ . Das Gewicht der Kante  $(ij)$  ist  $w_{(ij)} := P_{ij}$  und  $W = (w_{(ij)})_{(ij) \in K}$ . Der Graph  $(E, K, W)$  heißt *Übergangsgraph* von  $\mathcal{X}$ .

**Beispiel 5.15 (Irrfahrt im Dreieck).** Betrachte eine homogene Markov-Kette  $\mathcal{X}$  mit Zustandsraum  $\{1, 2, 3\}$  und Übergangsmatrix

$$P = \begin{pmatrix} 0 & p & q \\ q & 0 & p \\ p & q & 0 \end{pmatrix} \quad (5.3)$$

für  $p \in (0, 1)$  und  $q := 1 - p$ . Die Kette  $\mathcal{X}$  veranschaulicht man sich am besten anhand des Übergangsgraphen; siehe Abbildung 5.2. In jedem Zustand 1, 2, 3 ist die Wahrscheinlichkeit, im Uhrzeigersinn zu wandern  $p$ , und die Wahrscheinlichkeit gegen den Uhrzeigersinn zu gehen ist  $q$ .  $\square$



**Abbildung 5.2:** Übergangsgraph der Markov-Kette aus Beispiel 5.15.

**Beispiel 5.16 (Ruinproblem).** Betrachte folgendes Ruinproblem: zwei Spieler spielen gegeneinander. Spieler 1 startet mit  $n$  Euro, Spieler 2 mit  $N - n$  Euro. Bei jedem Spiel gewinnt Spieler 1 mit Wahrscheinlichkeit  $p$  einen Euro, mit Wahrscheinlichkeit  $q = 1 - p$  verliert er einen Euro. Das Spiel endet, wenn einer der beiden pleite ist.

Sei  $X_t$  das Vermögen von Spieler 1 nach dem  $t$ -ten Spiel. In dieser Situation ist  $\mathcal{X} = (X_t)_{t=0,1,2,\dots}$  eine endliche, homogene Markov-Kette mit Zustandsraum  $\{0, \dots, N\}$  und Übergangsmatrix

$$P = \begin{pmatrix} 1 & 0 & & & & & & & & \\ q & 0 & p & & & & & & & \\ & q & 0 & p & & & & & & \\ & & \ddots & \ddots & \ddots & & & & & \\ & & & q & 0 & p & & & & \\ & & & & 0 & 1 & & & & \end{pmatrix}.$$

Der Übergangsgraph ist in Abbildung 5.3 dargestellt. Klar ist, dass nach langer Zeit entweder Spieler 1 oder Spieler 2 gewonnen hat, dass also  $X_t \xrightarrow{t \rightarrow \infty} 0$  oder  $X_t \xrightarrow{t \rightarrow \infty} N$ . Die Frage ist nur:

Mit welcher Wahrscheinlichkeit gewinnt Spieler 1?

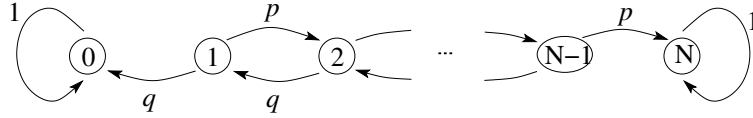


Abbildung 5.3: Übergangsgraph der Markov-Kette aus Beispiel 5.16.

Wir bezeichnen mit

$$p_n := \mathbf{P}[X_t \xrightarrow{t \rightarrow \infty} N \mid X_0 = n]$$

die Wahrscheinlichkeit, dass Spieler 1 gewinnt. Wir werden zeigen, dass

$$p_n = \begin{cases} \frac{n}{N}, & p = q = \frac{1}{2}, \\ \frac{1 - (q/p)^n}{1 - (q/p)^N}, & \text{sonst.} \end{cases} \quad (5.4)$$

Zunächst gilt für  $n = 1, \dots, N-1$

$$\begin{aligned} p_n &= q \cdot \mathbf{P}[X_t \xrightarrow{t \rightarrow \infty} N \mid X_0 = n, X_1 = n-1] + p \cdot \mathbf{P}[X_t \xrightarrow{t \rightarrow \infty} N \mid X_0 = n, X_1 = n+1] \\ &= q \cdot \mathbf{P}[X_t \xrightarrow{t \rightarrow \infty} N \mid X_0 = n-1] + p \cdot \mathbf{P}[X_t \xrightarrow{t \rightarrow \infty} N \mid X_0 = n+1] \\ &= q \cdot p_{n-1} + p \cdot p_{n+1}, \end{aligned}$$

also mit  $\Delta p_n := p_n - p_{n-1}$

$$q \Delta p_n = p \Delta p_{n+1}.$$

Im Fall  $p = q = \frac{1}{2}$  folgt mit  $\sum_{m=1}^N \Delta p_m = p_N - p_0 = 1$  daraus bereits

$$p_n = \frac{\sum_{m=1}^n \Delta p_m}{\sum_{m=1}^N \Delta p_m} = \frac{n \Delta p_1}{N \Delta p_1} = \frac{n}{N}.$$

Im Fall  $p \neq q$  setzen wir  $u := \frac{q}{p}$  und berechnen iterativ

$$\Delta p_n = u \Delta p_{n-1} = u^2 \Delta p_{n-2} \cdots = u^{n-1} \Delta p_1 = u^{n-1} p_1.$$

Weiter ist

$$1 = \sum_{m=1}^N \Delta p_m = p_1 \sum_{m=0}^{N-1} u^m = p_1 \frac{1 - u^N}{1 - u}.$$

Also

$$p_n = \sum_{m=1}^n \Delta p_m = p_1 \sum_{m=1}^n u^{m-1} = \frac{1-u}{1-u^N} \frac{1-u^n}{1-u} = \frac{1-u^n}{1-u^N}$$

und die Behauptung ist gezeigt.  $\square$

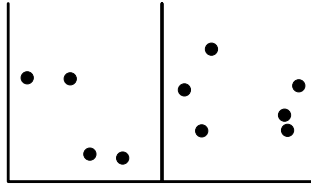
**Beispiel 5.17 (Ehrenfest'sche Urne).** Betrachte folgendes Urnenmodell: In einer Urne gibt es zwei durch eine Trennwand getrennte Kammern. Insgesamt liegen  $n$  Kugeln in den

beiden Kammern. Wir ziehen eine Kugel rein zufällig aus der Urne und legen sie anschließend in die andere Kammer zurück. In Abbildung 5.4 findet sich eine Illustration.

Wir betrachten den  $\{0, \dots, n\}$ -wertigen stochastischen Prozess  $\mathcal{X} = (X_t)_{t=0,1,2,\dots}$ , wobei  $X_t$  die Anzahl der Kugeln in der linken Kammer nach dem  $t$ -Schritt darstellt. Dann ist  $\mathcal{X}$  eine homogene Markov-Kette, denn der Ausgang des Schrittes  $t+1$  hängt nur von  $X_t$  ab. Die Übergangsmatrix von  $\mathcal{X}$  ist gegeben durch

$$P_{ij} = \begin{cases} \frac{i}{n}, & j = i - 1, \\ \frac{n-i}{n}, & j = i + 1, \\ 0, & \text{sonst.} \end{cases} \quad (5.5)$$

□



**Abbildung 5.4:** Die Ehrenfest'sche Urne mit  $n = 10$  Kugeln aus Beispiel 5.17. Im nächsten Schritt wird mit Wahrscheinlichkeit  $\frac{4}{10}$  eine Kugel von der linken Kammer in die rechte und mit Wahrscheinlichkeit  $\frac{6}{10}$  von der rechten in die linke Kammer gelegt.

Wir kommen nun zu einem wichtigen Zusammenhang zwischen der Verteilung einer Markov-Ketten zum Zeitpunkt  $t$  und Potenzen der Übergangsmatrix.

**Theorem 5.18.** Sei  $\mathcal{X} = (X_t)_{t=0,1,2,\dots}$  eine homogene Markov-Kette mit Übergangsmatrix  $P$  und  $\mu^{(t)} = (\mu^{(t)}(i))_{i \in E}$ ,

$$\mu^{(t)}(i) = \mathbf{P}[X_t = i]$$

für  $t = 0, 1, 2, \dots$ . Dann gilt

$$\mu^{(t)} = \mu^{(0)} P^t, \quad (5.6)$$

für  $t = 0, 1, 2, \dots$ , wobei die rechte Seite als Multiplikation des Zeilenvektors  $\mu^{(0)}$  und der Matrix  $P^t = \underbrace{P \cdots P}_{t \text{ mal}}$  zu verstehen ist.

*Beweis.* Der Beweis geht mittels Induktion über  $t$ . Für  $t = 0$  ist die Aussage klar. Ist (5.6) für  $t$  gezeigt, so gilt

$$\begin{aligned} \mu^{(t+1)}(j) &= \mathbf{P}[X_{t+1} = j] = \sum_{i \in E} \mathbf{P}[X_{t+1} = j \mid X_t = i] \cdot \mathbf{P}[X_t = i] \\ &= \sum_{i \in E} \mu^{(t)}(i) \cdot P_{ij} = \sum_{i \in E} (\mu^{(0)} P^t)_i \cdot P_{ij} = (\mu^{(0)} P^{t+1})_j. \end{aligned} \quad \square$$

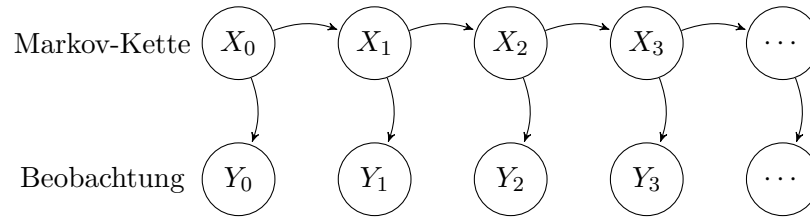


Abbildung 5.5: Illustration eines Hidden Markov Modells

**Beispiel 5.19 (Irrfahrt im Dreieck).** Betrachte die Markov-Kette  $\mathcal{X}$  aus Beispiel 5.15 und Übergangsmatrix  $P$  aus (5.3). Sei  $X_0 = 1$ , die Markov-Kette startet also in 1, d.h.  $\mu^{(0)} = (1, 0, 0)$ . Weiter berechnen wir

$$P = \begin{pmatrix} 0 & p & q \\ q & 0 & p \\ p & q & 0 \end{pmatrix}, \quad P^2 = \begin{pmatrix} 2pq & q^2 & p^2 \\ p^2 & 2pq & q^2 \\ q^2 & p^2 & 2pq \end{pmatrix}$$

und damit

$$\mu^{(2)} = (2pq, q^2, p^2).$$

Also ist etwa  $\mathbf{P}[X_2 = 1] = 2pq$ . Dies ist klar, weil  $X_2 = 1$  genau dann, wenn die ersten beiden Schritte im Übergangsgraphen aus Abbildung 5.2 einmal mit und einmal entgegen dem Uhrzeigersinn gingen. Es ist  $X_2 = 3$  genau dann, wenn zwei Schritte im Uhrzeigersinn realisiert wurden, was Wahrscheinlichkeit  $p^2$  hat.  $\square$

**Bemerkung 5.20 (Konvergenz von Markov-Ketten).** Mit Theorem 6.9 können wir bestimmen, welche Verteilung eine Markov-Kette nach  $t$  Zeiteinheiten hat. Von einer konvergenten Markov-Kette spricht man dann, wenn sich für große  $t$  diese Verteilung stabilisiert. Etwa sollte sich bei der Irrfahrt im Dreieck nach langer Zeit ergeben, dass die Markov-Kette mit etwa gleichen Wahrscheinlichkeiten in allen drei Zuständen befindet. Diese Konvergenz von Markov-Ketten wollen wir jedoch hier nicht weiter beleuchten.

## 5.4 Der Kalman-Filter

Als Anwendung bedingter Wahrscheinlichkeiten stellen wir nun den Kalman-Filter vor. Dieser wird eingesetzt, um verrauschte Messungen zu glätten. Dabei gehen wir von einem zugrunde liegenden Markov-Modell aus. Das bedeutet, dass wir daran interessiert sind, den Pfad einer Markov-Kette  $X_0, X_1, \dots$  aufzuzeichnen. Allerdings können wir die Kette nicht direkt, sondern nur verrauscht beobachten, d.h. wir beobachten  $Y_0, Y_1, \dots$ , wobei  $Y_k$  aus  $X_k$  mittels eines Übergangskernes entsteht. Die Messung  $Y_k$  hat also gegeben  $X_k$  eine Verteilung  $\mathbf{P}[Y_k \in A | X_k]$ . Da wir nur die Messungen  $Y_0, Y_1, \dots$  zur Verfügung haben, aber an der Markov-Kette  $X_0, X_1, \dots$  interessiert sind, spricht man hier auch von versteckten Markov-Modellen oder Hidden Markov Models (HMMs); siehe auch Abbildung 5.5.

Wir betrachten ein HMM  $(\underline{X}_n, \underline{Y}_n)_{n=0,1,2,\dots}$ . Wir wollen jeweils  $\underline{X}_n$  aus den Beobachtungen  $\underline{Y}_0, \dots, \underline{Y}_n$  schätzen. Wir bezeichnen mit

$$\pi_k(\cdot) := \mathbf{P}(\underline{X}_k \in \cdot | \underline{Y}_1, \dots, \underline{Y}_k), \quad \pi_{(k+1)|k}(\cdot) := \mathbf{P}(\underline{X}_{k+1} \in \cdot | \underline{Y}_1, \dots, \underline{Y}_k)$$

die bedingten Verteilungen von  $\underline{X}_k$  und  $\underline{X}_{k+1}$  gegeben  $\underline{Y}_1, \dots, \underline{Y}_k$ . Wir wollen nun eine Rekursion für  $\pi_k$  herleiten. Dies geht insbesondere dann, wenn alle  $\underline{X}_k, \underline{Y}_k$  normalverteilt sind. Wir erreichen dies über die (speziell beim Kalman-Filter getroffene) Annahme

$$\underline{X}_{k+1} = \underline{A}\underline{X}_k + \underline{R}\zeta_k, \quad \underline{Y}_k = \underline{B}\underline{X}_k + \underline{S}\eta_k$$

mit (bekannten) Matrizen  $\underline{A}, \underline{B}, \underline{R}, \underline{S}$ . Hierbei seien  $\zeta_k, \eta_k$  unabhängig und ebenfalls normalverteilt mit Einheitsmatrizen als Covarianzmatrizen.

**Theorem 5.21 (Kalman-Filter, Rekursionen).** *Angenommen,  $\pi_k = N(\widehat{\underline{X}}_k, \widehat{\underline{\Sigma}}_k)$ , so ist  $\pi_{(k+1)|k} = N(\widehat{\underline{X}}_{(k+1)|k}, \widehat{\underline{\Sigma}}_{(k+1)|k})$  und  $\pi_{k+1} = N(\widehat{\underline{X}}_{k+1}, \widehat{\underline{\Sigma}}_{k+1})$ , wobei sich  $\widehat{\underline{X}}_{(k+1)|k}, \widehat{\underline{\Sigma}}_{(k+1)|k}, \widehat{\underline{X}}_{k+1}, \widehat{\underline{\Sigma}}_{k+1}$  rekursiv wie folgt berechnen:*

$$\begin{aligned} \widehat{\underline{X}}_{(k+1)|k} &= \underline{A}\widehat{\underline{X}}_k, \\ \widehat{\underline{\Sigma}}_{(k+1)|k} &= \underline{A}\widehat{\underline{\Sigma}}_k\underline{A}^\top + \underline{R}\underline{R}^\top, \\ \widehat{\underline{X}}_{k+1} &= \widehat{\underline{X}}_{(k+1)|k} + \widehat{\underline{\Sigma}}_{(k+1)|k}\underline{B}^\top(\underline{B}\widehat{\underline{\Sigma}}_{(k+1)|k}\underline{B}^\top + \underline{S}\underline{S}^\top)^{-1}(\underline{Y}_{k+1} - \underline{B}\widehat{\underline{X}}_{(k+1)|k}), \\ \widehat{\underline{\Sigma}}_{k+1} &= \widehat{\underline{\Sigma}}_{(k+1)|k} - \widehat{\underline{\Sigma}}_{(k+1)|k}\underline{B}^\top(\underline{B}\widehat{\underline{\Sigma}}_{(k+1)|k}\underline{B}^\top + \underline{S}\underline{S}^\top)^{-1}\underline{B}\widehat{\underline{\Sigma}}_{(k+1)|k}. \end{aligned}$$

Um dieses Resultat zu erhalten, benötigen wir zunächst eine weitere wichtige Aussage über normalverteilte Zufallsvariable.

**Proposition 5.22 (Bedingte Normalverteilung ist Normalverteilung).** *Die  $\mathbb{R}^{m+n}$ -wertige Zufallsvariable  $(\underline{X}, \underline{Y})$  sei  $N((\underline{\mu}, \underline{\Sigma})$ -verteilt mit  $\underline{\mu} = (\underline{\mu}_X, \underline{\mu}_Y)$  und*

$$\underline{\Sigma} = \begin{pmatrix} \underline{\Sigma}_{XX} & \underline{\Sigma}_{XY} \\ \underline{\Sigma}_{YX} & \underline{\Sigma}_{YY} \end{pmatrix}.$$

*Dann ist die bedingte Verteilung von  $\underline{X}$  gegeben  $\underline{Y}$  eine*

$$N(\underline{\mu}_X + \underline{\Sigma}_{XY}\underline{\Sigma}_{YY}^{-1}(\underline{Y} - \underline{\mu}_Y), \underline{\Sigma}_{XX} - \underline{\Sigma}_{XY}\underline{\Sigma}_{YY}^{-1}\underline{\Sigma}_{YX}) - \text{Verteilung.}$$

*Beweis.* Als Linearkombination von normalverteilten Zufallsvariablen ist  $\underline{Z} := \underline{X} - \underline{\Sigma}_{XY}\underline{\Sigma}_{YY}^{-1}\underline{Y}$  normalverteilt mit Erwartungswert  $\underline{\mu}_X - \underline{\Sigma}_{XY}\underline{\Sigma}_{YY}^{-1}\underline{\mu}_Y$  und Covarianzmatrix

$$\underline{\Sigma}_{XX} - 2\underline{\Sigma}_{XY}\underline{\Sigma}_{YY}^{-1}\underline{\Sigma}_{YX} + \underline{\Sigma}_{XY}\underline{\Sigma}_{YY}^{-1}\underline{\Sigma}_{YX} = \underline{\Sigma}_{XX} - \underline{\Sigma}_{XY}\underline{\Sigma}_{YY}^{-1}\underline{\Sigma}_{YX}.$$

Es gilt

$$\mathbf{Cov}[\underline{Z}, \underline{Y}] = \underline{\Sigma}_{XY} - \underline{\Sigma}_{XY}\underline{\Sigma}_{YY}^{-1}\underline{\Sigma}_{YY} = \underline{0},$$

und damit sind  $\underline{Y}$  und  $\underline{Z}$  nach Korollar 3.26 unabhängig. Nun schreiben wir

$$\begin{aligned} \mathbf{E}[e^{it^\top \underline{X}} | \underline{Y}] &= \mathbf{E}[e^{it^\top \underline{Z}} e^{it^\top (\underline{\Sigma}_{XY}\underline{\Sigma}_{YY}^{-1}\underline{Y})} | \underline{Y}] = \mathbf{E}[e^{it^\top \underline{Z}}] e^{it^\top (\underline{\Sigma}_{XY}\underline{\Sigma}_{YY}^{-1}\underline{Y})} \\ &= e^{it^\top (\underline{\mu}_X + \underline{\Sigma}_{XY}\underline{\Sigma}_{YY}^{-1}(\underline{Y} - \underline{\mu}_Y)) - \frac{1}{2}t^\top (\underline{\Sigma}_{XX} - \underline{\Sigma}_{XY}\underline{\Sigma}_{YY}^{-1}\underline{\Sigma}_{YX})t}. \end{aligned}$$

Daraus folgt die Behauptung. □

*Beweis von Theorem 5.21.* Wir lassen im Folgenden die Unterstriche für die Vektor- und Matrix-Notation weg. Gegeben  $Y_1, \dots, Y_k$  ist  $X_k$  nach  $\pi_k$  verteilt, hat also eine Normalverteilung mit  $\widehat{X}_k$  und  $\widehat{\Sigma}_k$ . Nun ist nach Annahme  $X_{k+1}$  bedingt unter  $Y_1, \dots, Y_k$  eine Linearkombination der  $X_k$ . Nach Proposition 3.25 hat also  $X_{k+1}$  unter  $\pi_{(k+1)|k}$  eine Normalverteilung mit Erwartungswert und Covarianz

$$\begin{aligned} \mathbf{E}[X_{k+1}|\widehat{X}_k] &= A\widehat{X}_k, \\ \mathbf{Cov}[X_{k+1}, X_{k+1}|\widehat{X}_k] &= \mathbf{Cov}[A\widehat{X}_k + R\zeta_k, A\widehat{X}_k + R\zeta_k] = A\widehat{\Sigma}_k A^\top + RR^\top \end{aligned}$$

Im nächsten Schritt müssen wir nun die Verteilung von  $X_{k+1}$  bedingt unter  $Y_1, \dots, Y_{k+1}$  berechnen. Zunächst ist  $(X_{k+1}, Y_{k+1})$  bedingt unter  $Y_1, \dots, Y_k$  normalverteilt mit Erwartungswert  $(\widehat{X}_{(k+1)|k}, B\widehat{X}_{(k+1)|k})$  und Covarianzmatrix

$$\begin{pmatrix} \widehat{\Sigma}_{(k+1)|k} & (B\widehat{\Sigma}_{(k+1)|k})^\top \\ B\widehat{\Sigma}_{(k+1)|k} & B\widehat{\Sigma}_{(k+1)|k}B^\top + SS^\top \end{pmatrix}.$$

Nun müssen wir auf  $Y_{k+1}$  bedingen. Nach Proposition 5.22 ergibt sich dann eine Normalverteilung mit Erwartungswert

$$\widehat{X}_{(k+1)|k} + (B\widehat{\Sigma}_{(k+1)|k})^\top (B\widehat{\Sigma}_{(k+1)|k}B^\top + SS^\top)^{-1} (Y_{k+1} - B\widehat{X}_{(k+1)|k})$$

und Covarianzmatrix

$$\widehat{\Sigma}_{(k+1)|k} - (B\widehat{\Sigma}_{(k+1)|k})^\top (B\widehat{\Sigma}_{(k+1)|k}B^\top + SS^\top)^{-1} B\widehat{\Sigma}_{(k+1)|k}.$$

Dies war aber gerade die Behauptung.  $\square$

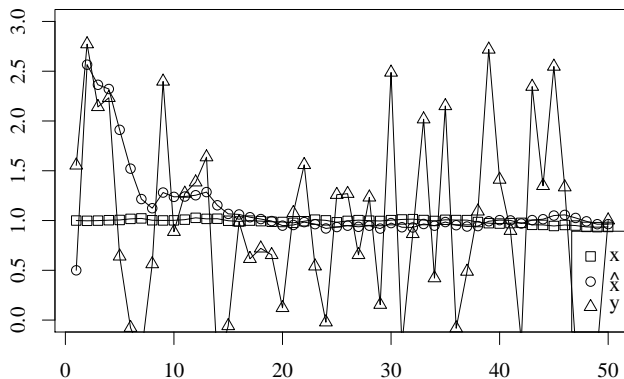
**Beispiel 5.23** ( $X_0, X_1, \dots$  **konstant**). Wir geben ein numerisches Beispiel in einem der einfachsten Fälle. Geht man davon aus, dass  $X_0, X_1, \dots$  reellwertig und nahezu konstant ist, so setzen wir  $AX = X$  und  $R = 10^{-2}, S = 0.3$ , und damit

$$X_{k+1} = X_k.$$

Die Rekursionen sind nun

$$\begin{aligned} \widehat{X}_{(k+1)|k} &= \widehat{X}_k, \\ \widehat{\Sigma}_{(k+1)|k} &= \widehat{\Sigma}_k + R^2, \\ \widehat{X}_{k+1} &= \widehat{X}_{(k+1)|k} + \widehat{\Sigma}_{(k+1)|k} (\widehat{\Sigma}_{(k+1)|k} + S^2)^{-1} (Y_{k+1} - \widehat{X}_{(k+1)|k}) \\ &= \widehat{X}_k + \frac{\widehat{\Sigma}_k + R}{\widehat{\Sigma}_k + R^2 + S^2} (Y_{k+1} - \widehat{X}_k) \\ \widehat{\Sigma}_{k+1} &= \widehat{\Sigma}_{(k+1)|k} - \widehat{\Sigma}_{(k+1)|k} (\widehat{\Sigma}_{(k+1)|k} + S^2)^{-1} \widehat{\Sigma}_{(k+1)|k} \\ &= (\widehat{\Sigma}_k + R) \left( 1 - \frac{\widehat{\Sigma}_k + R^2}{\widehat{\Sigma}_k + R^2 + S^2} \right). \end{aligned}$$

Eine numerische Anwendung dieser Rekursionen findet sich in Abbildung 5.6.



**Abbildung 5.6:** Der Kalman-Filter ermittelt aus verrauschten Daten  $Y$  eine Schätzung der wahren Wert  $X$ .

## 6 Statistik

Es ist nicht übertrieben zu behaupten, dass in der heutigen Welt immer mehr *Daten* jeglicher Art erhoben werden. Diese zu ordnen und aus Daten Schlussfolgerungen zu ziehen ist Aufgabe der Statistik.

Man teilt dabei diese Aufgaben in zwei Gebiete auf. Die *deskriptive Statistik* dient rein der Beschreibung der Daten, etwa durch geeignete Wahl von Statistiken, die die Daten zusammenfassen. Anders ist dies bei der hier behandelten *schließenden* oder *induktiven Statistik*. Die Aufgabe ist hier, mit Hilfe von stochastischen Modellen Aussagen darüber zu treffen, welchen Annahmen den Daten zugrunde liegen könnten.

### 6.1 Grundlagen

Wir beginnen mit einem Beispiel.

#### **Beispiel 6.1 (Erfolgswahrscheinlichkeit beim Münzwurf).**

Bevor wir statistische Konzepte einführen, betrachten wir folgendes Beispiel: eine Münze wird 100 mal geworfen. Dabei ist die Wahrscheinlichkeit für *Kopf* (was wir im Folgenden als Erfolg werten wollen) noch unbekannt. Von den 100 Würfeln sind 59 ein Erfolg.

Unsere statistischen Überlegungen gehen nun von der Vorstellung aus, dass die 100 Münzwürfe die Realisierung einer Zufallsvariable  $X = (X_1, \dots, X_{100})$  sind, wobei  $X_1, \dots, X_{100}$  unabhängig und identisch verteilt sind mit

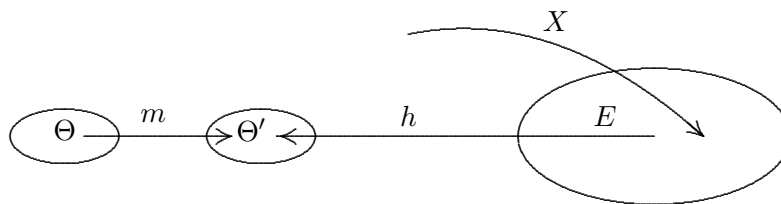
$$X_i = \begin{cases} 1, & \text{falls der } i\text{-te Wurf } \textit{Kopf} \text{ zeigt,} \\ 0, & \text{sonst.} \end{cases}$$

und es gilt

$$\mathbf{P}[X_i = 1] = p.$$

Jetzt ist  $X_1 + \dots + X_n$  die Gesamtzahl der Erfolge. Als Summe von  $n$  unabhängigen Bernoulli-verteilten Zufallsvariablen ist diese Summe  $B(n = 100, p)$ -verteilt. Wichtig ist, dass zwar





**Abbildung 6.1:** Veranschaulichung von statistischen Modellen

$n = 100$  bereits fest steht (schließlich wissen wir ja, dass wir 100 mal die Münze geworfen haben), nicht jedoch  $p$ . In dieser Situation gibt es zwei *statistische Probleme*.

- *Schätzproblem:* Wir können versuchen, den Erfolgsparameter  $p$  zu schätzen. Wir werden hierzu aus den Daten (59 Erfolge aus 100 Versuchen) einen Wert  $\hat{p}$  ableiten (*Punktschätzer*). Alternativ kann man auch ein Intervall  $[a, b]$  angeben, in dem der wahre Parameter  $p$  mit hoher Wahrscheinlichkeit liegt (*Intervallschätzer*). Hierauf werden wir allerdings nicht näher eingehen.
- *Testproblem:* Stellen wir uns vor, der Werfer der Münze behauptet, dass die Münze fair ist, also  $p = \frac{1}{2}$  gilt. Dieser Meinung können wir skeptisch gegenüber stehen, da ja sogar 59 aus 100 Würfeln ein Erfolg waren. Wir können versuchen, die Hypothese  $p = \frac{1}{2}$  zu testen. Das bedeutet, dass wir untersuchen, wie gut die Hypothese mit den Daten in Einklang steht.

Wie im Beispiel gesehen besteht ein statistisches Experiment aus einer (oder mehreren) Zufallsvariablen (etwa die Anzahl der Erfolge) und verschiedenen (möglichen) Verteilungen der Zufallsvariable; hier  $B(n = 100, p)$  für variierendes  $p$ . Dies führt zur Definition des statistischen Modells. Siehe auch Abbildung 6.1.

**Definition 6.2 (Statistisches Modell).** Seien  $E, \Theta, \Theta'$  Mengen. Ein statistisches Modell ist ein Paar  $(X, (\mathbf{P}_\vartheta)_{\vartheta \in \Theta})$ , wobei  $X$  eine Zufallsvariable mit Zielbereich  $E$  ist, bei deren Verteilung noch ein Parameter  $\vartheta \in \Theta$  frei, also unbestimmt, ist. Das bedeutet, dass es eine Funktion  $\vartheta \mapsto \rho_\vartheta$  gibt mit<sup>7</sup>

$$\mathbf{P}_\vartheta[X \in da] = \rho_\vartheta(a)da.$$

Die Menge  $\Theta$  heißt Parameterraum, die Menge  $E$  Beobachtungsraum. Jede Zufallsvariable  $h(X)$  mit  $h : E \rightarrow \Theta'$  heißt Statistik.

**Beispiel 6.3 (Erfolgswahrscheinlichkeit beim Münzwurf).** In Beispiel 6.1 ist einfach  $X$  die Anzahl der Erfolge und  $\mathbf{P}_p := B(n = 100, p)$  (wobei wir  $\vartheta$  durch  $p$  ersetzt haben).

<sup>7</sup>Wir wollen im Folgenden die dauernde Unterscheidung zwischen diskreten Zufallsvariablen und Zufallsvariablen mit Dichten durch die Notation  $\mathbf{P}[X \in da]$  vermeiden. Ist der Wertebereich  $E$  von  $X$  diskret und  $a \in E$ , ist damit

$$\mathbf{P}[X \in da] := \mathbf{P}[X = a]$$

gemeint. Hat  $X$  die Dichte  $f(a)da$ , ist

$$\mathbf{P}[X \in da] := f(a)da.$$

## 6.2 Schätzprobleme

Ein Schätzproblem ist das Folgende: Gegeben sei eine Funktion  $m : \Theta \rightarrow \Theta'$  (was meist die Identität und  $\Theta = \Theta'$  ist). Bestimme eine Funktion  $h : E \rightarrow \Theta'$ , so dass  $h(X)$  den Wert von  $m(\vartheta)$  (möglichst gut) schätzt. Interessant hierbei ist, was es denn überhaupt bedeutet, guter Schätzer zu sein.

**Beispiel 6.4 (Erfolgswahrscheinlichkeit beim Münzwurf).** In der Situation aus den Beispielen 6.1 und 6.3 heißt der freie Parameter  $p$  anstatt  $\vartheta$ . Es ist  $E = \{0, \dots, 100\}$  und  $X$  die Anzahl der Erfolge in  $n = 100$  Versuchen. Damit ist  $\mathbf{P}_p = B(n = 100, p)$  mit dem freien Parameter  $p \in \Theta = [0, 1]$ . Es ist mit  $\Theta' = \Theta$

$$h : \begin{cases} E & \rightarrow \Theta, \\ x & \mapsto \frac{1}{100}x. \end{cases}$$

und  $\hat{p} = h(X)$  ist ein Schätzer für  $p$ . (Im Folgenden werden wir einen Schätzer für einen Parameter  $\bullet$  meistens mit  $\hat{\bullet}$  bezeichnen.) Für unser Datenbeispiel ist also

$$\hat{p} = \frac{1}{n}(X_1 + \dots + X_n) = \frac{59}{100} = 0.59.$$

Da  $\hat{p}$  von den Daten abhängt, die wir uns als Realisierung von einer Zufallsvariable gedacht haben, ist  $\hat{p}$  also auch eine Zufallsvariable. Warum ist der Schätzer  $\hat{p}$  gut? Nehmen wir an, wir wüssten den wahren Parameter  $p$ . Dann leistet  $\hat{p}$  zumindest im Mittel das gewünschte: (Wir schreiben hier und im Folgenden  $\mathbf{P}_p[\cdot]$  und  $\mathbf{E}_p[\cdot]$ , wenn wir Wahrscheinlichkeiten und Erwartungswerte unter der Hypothese ausrechnen wollen, dass  $p$  der wahre Parameter ist.)

$$\mathbf{E}_p[\hat{p}] = \frac{1}{n}\mathbf{E}_p[X_1 + \dots + X_n] = p.$$

Wir sagen auch, der Schätzer  $\hat{p}$  ist erwartungstreu (oder unverzerrt oder unbiased).

Eine weitere wünschenswerte Eigenschaft eines Schätzers ist, dass er immer besser wird, je größer die zu Grunde liegende Datenmenge ist. Eine große Datengrundlage bedeutet in unserem Fall, dass die Münze oft geworfen wurde, also  $n$  groß ist. Aus dem schwachen Gesetz großer Zahlen wissen wir, dass

$$\mathbf{P}_p[|\hat{p} - p| \geq \varepsilon] = \mathbf{P}_p\left[\left|\frac{X_1 + \dots + X_n}{n} - \mathbf{E}_p[X_1]\right| \geq \varepsilon\right] \xrightarrow{n \rightarrow \infty} 0$$

für alle  $\varepsilon > 0$ . Die Eigenschaft, dass  $\hat{p}$  mit hoher Wahrscheinlichkeit immer näher am wahren Wert  $p$  liegt, wenn mehr Daten zur Verfügung stehen, nennen wir Konsistenz.

**Definition 6.5 (Punktschätzer, unverzerrte und konsistente Schätzer).**

1. Sei  $(X, (\mathbf{P}_\vartheta)_{\vartheta \in \Theta})$  ein statistisches Modell und  $m : \Theta \rightarrow \Theta'$ . Jede Statistik  $\hat{m} := \hat{m}(X)$  mit  $\hat{m} : E \rightarrow \Theta'$  heißt (Punkt-)Schätzer für  $m(\vartheta)$ .

Der Schätzer  $\hat{m}$  heißt unverzerrt (erwartungstreu, unbiased), falls

$$\mathbf{E}_\vartheta[\hat{m}] = m(\vartheta)$$

für alle  $\vartheta \in \Theta$ .

2. Sei  $(X^n, (\mathbf{P}_\vartheta^n)_{\vartheta \in \Theta})_{n=1,2,\dots}$  eine Folge statistischer Modelle mit derselben Parametermenge  $\Theta$  und  $\hat{m}_1(X^1), \hat{m}_2(X^2), \dots$  eine Folge von Schätzern für  $m(\vartheta)$ . Die Folge  $\hat{m}_1(X^1), \hat{m}_2(X^2), \dots$  heißt konsistent, falls

$$\mathbf{P}_\vartheta^n[|\hat{m}_n(X^n) - m(\vartheta)| \geq \varepsilon] \xrightarrow{n \rightarrow \infty} 0$$

für alle  $\varepsilon > 0, \vartheta \in \Theta$ .

3. Sei  $(X, (\mathbf{P}_\vartheta)_{\vartheta \in \Theta})$  ein statistisches Modell. Die Abbildung

$$L : \begin{cases} E \times \Theta & \rightarrow [0, 1] \\ (a, \vartheta) & \mapsto \mathbf{P}_\vartheta[X \in da] \end{cases}$$

heißt Likelihood-Funktion. Für eine Abbildung  $h : E \mapsto \Theta$  mit

$$L(a, h(a)) = \max_{\vartheta \in \Theta} L(a, \vartheta)$$

heißt  $\hat{\vartheta}_{ML} = h(X)$  Maximum-Likelihood-Schätzer von  $\vartheta$ .

**Bemerkung 6.6 (Interpretation von Maximum-Likelihood-Schätzern).** Sei  $X$  diskret und  $X = a$ . Da die Vorstellung die ist, dass die erhobenen Daten Ergebnis eines Zufallsexperiments (d.h. die Realisierung einer Zufallsvariable  $X$ ) sind, sagt man auch, dass die Daten  $X = a$  sind. Ein Maximum-Likelihood-Schätzer ist also ein Parameter  $\vartheta$ , unter dem die Wahrscheinlichkeit, die Daten  $X = a$  zu beobachten – das ist  $\mathbf{P}_\vartheta[X = a]$  – maximal ist.

**Beispiel 6.7 (Erfolgswahrscheinlichkeit beim Münzwurf).** Betrachten wir also wieder das Beispiel der Schätzung der Erfolgswahrscheinlichkeit beim Münzwurf. Wir hatten  $X = 59$  Erfolge bei  $n = 100$  Erfolgen verzeichnet. Unter  $\mathbf{P}_p$  ist  $X$  nach  $B(n = 100, p)$  verteilt, also ist

$$L(X, p) = \binom{n}{X} p^X (1-p)^{n-X}.$$

die Likelihood-Funktion; siehe auch Abbildung 6.2. Um diese für gegebenes  $X$  zu maximieren, berechnen wir den Wert  $p$ , für den  $\log L(X, p)$  maximal ist. Die Bestimmung des Maximums der log-Likelihood-Funktion  $\log L(a, p)$  genügt, da  $\log$  eine streng monotone Funktion ist. Wir berechnen

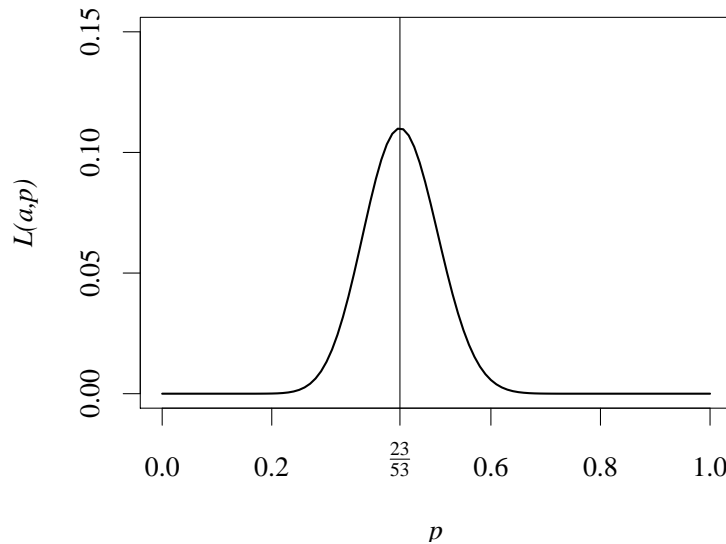
$$\frac{\partial \log L(X, p)}{\partial p} = \frac{\partial}{\partial p} \left( \log \binom{n}{X} + X \log p + (n - X) \log(1 - p) \right) = \frac{X}{p} - \frac{n - X}{1 - p}.$$

Am Maximum  $\hat{p}_{ML}$  muss  $\frac{\partial \log L(X, p)}{\partial p} = 0$  sein, also ist

$$(1 - \hat{p}_{ML})X = \hat{p}_{ML}(n - X), \quad \hat{p}_{ML} = \frac{X}{n}$$

ein Maximum-Likelihood-Schätzer für  $p$ . Da es nur ein einziges Maximum der Likelihood gibt, ist dies auch der einzige Maximum-Likelihood-Schätzer.

Soeben haben wir gesehen, dass der Maximum-Likelihood-Schätzer für eine Summe von Zufallsvariablen (was der Anzahl an Erfolgen im Münzwurf entspricht) gerade durch den Mittelwert gegeben ist. Die Verwendung des Mittelwertes für eine solche Schätzung ist generell eine gute Idee, wie wir nun zeigen werden.



**Abbildung 6.2:** Die Likelihood-Funktion beim Münzwurf für  $X = 23$  aus Beispiel 6.7.

**Definition 6.8 (Mittelwert und empirische Varianz).** Sei  $\underline{X} = (X_1, \dots, X_n)$  ein Vektor von Zufallsvariablen. Dann heißt

$$\bar{X} := \frac{1}{n}(X_1 + \dots + X_n)$$

Mittelwert von  $\underline{X}$  und

$$s^2(\underline{X}) := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

empirische Varianz von  $\underline{X}$ .

**Theorem 6.9 (Mittelwert und empirische Varianz als Schätzer).**

Sei  $(\underline{X} = (X_1, \dots, X_n), (\mathbf{P}_\vartheta)_{\vartheta \in \Theta})$  ein statistisches Modell, so dass  $X_1, \dots, X_n$  unter allen  $\mathbf{P}_\vartheta$  identisch verteilt sind und  $\mu_\vartheta := \mathbf{E}_\vartheta[X_1]$  existiert.

1. Es ist  $\bar{X}$  ein unverzerrter Schätzer für  $\mu_\vartheta$ . Sind außerdem  $X_1, \dots, X_n$  unter allen  $\mathbf{P}_\vartheta$  unabhängig mit endlichem  $\mathbf{V}_\vartheta[X_1]$ , so ist  $\bar{X}$  auch konsistent.
2. Sei  $n \geq 2$  und  $\underline{X} = (X_1, \dots, X_n)$  so, dass  $X_1, \dots, X_n$  unter allen  $\mathbf{P}_\vartheta$  paarweise unkorreliert und identisch verteilt sind mit  $\sigma_\vartheta^2 := \mathbf{V}_\vartheta[X_1] < \infty$ . Dann ist die empirische Varianz  $s^2(\underline{X})$  ein unverzerrter Schätzer für  $\sigma_\vartheta^2$ .

*Beweis.* 1. Zunächst ist

$$\mathbf{E}_\vartheta[\bar{X}] = \frac{1}{n}(\mathbf{E}_\vartheta[X_1] + \dots + \mathbf{E}_\vartheta[X_n]) = \mathbf{E}_\vartheta[X_1] = \mu_\vartheta,$$

was bereits die Unverzerrtheit von  $\bar{X}$  als Schätzer von  $\mu_\vartheta$  zeigt. Für die Konsistenz berechnen wir für  $\varepsilon > 0$

$$\mathbf{P}_\vartheta[|\bar{X} - \mu_\vartheta| \geq \varepsilon] = \mathbf{P}_\vartheta\left(\left|\frac{X_1 + \dots + X_n}{n} - \mathbf{E}_\vartheta[X_1]\right| \geq \varepsilon\right) \xrightarrow{n \rightarrow \infty} 0$$

nach dem schwachen Gesetz der großen Zahlen.

2. Wir schreiben zunächst

$$\mathbf{E}_\vartheta[s^2(\underline{X})] = \frac{1}{n-1} \sum_{i=1}^n \mathbf{E}_\vartheta[X_i^2 - 2X_i\bar{X} + \bar{X}^2] = \frac{n}{n-1} \mathbf{E}_\vartheta[X_1^2 - 2X_1\bar{X} + \bar{X}^2].$$

Nun ist

$$\begin{aligned} \mathbf{E}_\vartheta[X_1^2] &= \mu_\vartheta^2 + \sigma_\vartheta^2, \\ \mathbf{E}_\vartheta[X_1\bar{X}] &= \mu_\vartheta^2 + \frac{1}{n}\sigma_\vartheta^2, \\ \mathbf{E}_\vartheta[\bar{X}^2] &= \mathbf{E}_\vartheta[X_1\bar{X}], \end{aligned}$$

also

$$\mathbf{E}_\vartheta[s^2(\underline{X})] = \frac{n}{n-1} \mathbf{E}_\vartheta[X_1^2 - X_1\bar{X}] = \sigma_\vartheta^2,$$

was die Unverzerrtheit bereits zeigt.  $\square$

**Beispiel 6.10 (Erfolgswahrscheinlichkeit beim Münzwurf).** Betrachten wir noch einmal das einführende Beispiel der Schätzung des Erfolgsparameters in einem Münzwurf in  $n$  Versuchen. Wir wählen dazu die Notation aus den Beispielen 6.1 und 6.3. Der Schätzer  $\hat{p} = h(X)$  ist unverzerrt und konsistent, wie wir soeben gezeigt haben.

Ein weiterer Schätzer für  $p$  wäre  $\hat{p}' = X_1$ . Das bedeutet, dass  $\hat{p}'$  nur die beiden Werte, 0 und 1 annehmen kann, und genau dann 1 ist, wenn der erste Wurf einen Erfolg zeigt. Der Schätzer  $\hat{p}'$  ist ebenfalls erwartungstreu, denn

$$\mathbf{E}_p[\hat{p}'] = \mathbf{E}_p[X_1] = p.$$

Allerdings ist  $\hat{p}'$  nicht konsistent, da

$$\mathbf{P}_p(|\hat{p}' - p| > \varepsilon) = 1,$$

falls  $\varepsilon < \min(p, 1 - p)$ , unabhängig von  $n$ .

**Beispiel 6.11 (Maximum-Likelihood-Schätzer bei Normal-Verteilungen).**

Wir betrachten den Fall einer unabhängigen, normalverteilten Stichprobe. Sei also  $(\underline{X}, (\mathbf{P}_{(\mu, \sigma^2)})_{\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+})$  so, dass  $\underline{X} = (X_1, \dots, X_n)$  und  $X_1, \dots, X_n$  unter  $\mathbf{P}_{(\mu, \sigma^2)}$  unabhängig und identisch nach  $N(\mu, \sigma^2)$  verteilt sind.

Wir berechnen nun die Maximum-Likelihood-Schätzer für  $\mu$  und  $\sigma^2$ . Genau wie im letzten Beispiel berechnen wir zunächst die log-Likelihood-Funktion

$$\begin{aligned} \log L((X_1, \dots, X_n), (\mu, \sigma^2)) &= \log \left( \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left( - \sum_{i=1}^n \frac{(X_i - \mu)^2}{2\sigma^2} \right) \right) \\ &= -n \log \sigma - \sum_{i=1}^n \frac{(X_i - \mu)^2}{2\sigma^2} + C, \end{aligned}$$

wobei  $C$  weder von  $\mu$  noch von  $\sigma$  abhängt. Ableiten nach  $\mu$  und  $\sigma$  ergibt

$$\begin{aligned} \frac{\partial \log L((X_1, \dots, X_n), (\mu, \sigma^2))}{\partial \mu} &= \sum_{i=1}^n \frac{X_i - \mu}{\sigma^2}, \\ \frac{\partial \log L((X_1, \dots, X_n), (\mu, \sigma^2))}{\partial \sigma} &= -\frac{n}{\sigma} + \sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^3}. \end{aligned}$$

Für die Maximum-Likelihood-Schätzer  $\hat{\mu}_{ML}$  und  $\hat{\sigma}_{ML}^2$  gilt notwendigerweise

$$\sum_{i=1}^n (X_i - \hat{\mu}_{ML}) = 0,$$

$$\frac{n}{\hat{\sigma}_{ML}} - \sum_{i=1}^n \frac{(X_i - \hat{\mu}_{ML})^2}{\hat{\sigma}_{ML}^3} = 0.$$

Die Maximum-Likelihood-Schätzer sind also gegeben durch

$$\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X},$$

$$\hat{\sigma}_{ML}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} s^2(\underline{X}).$$

Insbesondere sehen wir, dass  $\bar{X}$  nicht nur erwartungstreu und konsistent (siehe Theorem 6.9) ist, sondern auch ein Maximum-Likelihood-Schätzer für  $\mu$ . Allerdings ist der Maximum-Likelihood-Schätzer für  $\sigma^2$  nicht erwartungstreu, wie man aus Theorem 6.9 abliest. Immerhin ist  $\hat{\sigma}_{ML}^2$  für große  $n$  annähernd erwartungstreu, da  $\hat{\sigma}_{ML}^2 - s^2(\underline{X}) \xrightarrow{n \rightarrow \infty} 0$ .

### 6.3 Testprobleme

Ein Testproblem ist das Folgende: Sei  $\Theta_0 \subset \Theta$ . Kann die Hypothese  $\vartheta \in \Theta_0$  auf Grundlage der Daten verworfen werden?

**Beispiel 6.12 (Erfolgswahrscheinlichkeit beim Münzwurf).** Nehmen wir an, der Werfer der Münze behauptet, sie sei fair, also  $p = \frac{1}{2}$ . Es ist also  $\Theta_0 = \{\frac{1}{2}\}$ . Können wir diese Hypothese aufgrund der Daten verwerfen? Zunächst stellen wir fest, dass wir prinzipiell zwei Arten von Fehlern mit unserer Entscheidung machen können. Wenn wir die Hypothese verwerfen, könnte sie doch wahr sein, und wenn wir die Hypothese nicht verwerfen, könnte sie doch falsch sein.

Da wir nicht grundlos dem Werfer der Münze widersprechen wollen, wollen wir die Wahrscheinlichkeit, dass wir die Hypothese ablehnen (wir dem Werfer der Münze widersprechen), obwohl sie wahr ist (die Hypothese des Werfers richtig ist), kontrollieren. Das bedeutet, dass

$$\mathbf{P}_{p=1/2}(\text{Hypothese verwerfen}) \leq \alpha$$

für ein anfangs gewähltes  $\alpha \in (0, 1)$  sein soll. Klar ist, dass damit die Hypothese umso seltener abgelehnt werden kann, je kleiner  $\alpha$  ist. Nun kommen wir zu der Regel, mit der wir die Hypothese ablehnen wollen. In unserem Beispiel haben wir für die Hypothese  $p = \frac{1}{2}$  zu viele (59 von 100) Erfolge. Wir würden die Hypothese ablehnen wollen, wenn

$$\mathbf{P}_{p=1/2}(\text{Daten mindestens so extrem wie tatsächliche Daten}) \leq \alpha. \quad (6.1)$$

Wir wählen  $\alpha = 5\%$ . Für  $X \sim B(n = 100, p = \frac{1}{2})$  verteilt, erwarten wir 50 Erfolge. Um die Wahrscheinlichkeit einer Abweichung, die größer ist als die der Daten zu berechnen, betrachten

wir eine nach  $N(0, 1)$  verteilte Zufallsvariable  $Z$  und berechnen

$$\begin{aligned} & \mathbf{P}_{p=1/2}(|X_1 + \cdots + X_n - 50| \geq 9) \\ &= 1 - \mathbf{P}_{p=1/2}\left(-\frac{9}{\sqrt{np(1-p)}} < \frac{X - np}{\sqrt{np(1-p)}} < \frac{9}{\sqrt{np(1-p)}}\right) \\ &\approx 1 - \mathbf{P}_{p=1/2}(-1.8 \leq Z \leq 1.8) \approx 7.19\% \end{aligned}$$

Da dieser Wert größer als  $\alpha = 5\%$  ist, kann die Hypothese nicht verworfen werden, siehe (6.1).

Wir beginnen mit der Einführung wichtiger Begriffe wie Teststatistik, Nullhypothese, Alternative, Ablehnungsbereich, Signifikanzniveau und  $p$ -Wert.

**Definition 6.13 (Statistischer Test).** Sei  $(X, (\mathbf{P}_\vartheta)_{\vartheta \in \Theta})$  ein statistisches Modell,  $E$  der Zielbereich von  $X$ , und  $\Theta_0, \Theta_A \subseteq \Theta$  disjunkt mit  $\Theta_0 \cup \Theta_A = \Theta$ .

1. Ein Paar  $(Y, C)$  mit  $Y = t(X)$  für  $t : E \rightarrow E'$  und  $C \subseteq E'$  heißt statistischer Test von

$$H_0 : \vartheta \in \Theta_0 \text{ gegen } H_A : \vartheta \in \Theta_A.$$

Hier heißt  $Y$  Teststatistik,  $C$  kritischer oder Ablehnungsbereich des Tests,  $H_0$  heißt Nullhypothese und  $H_A$  heißt Alternativhypothese. Man sagt, der Test  $(Y, C)$  hat (Signifikanz-)Niveau  $\alpha \in [0, 1]$ , falls

$$\sup_{\vartheta \in \Theta_0} \mathbf{P}_\vartheta(Y \in C) \leq \alpha.$$

Falls  $Y \in C$ , sagt man, dass  $H_0$  abgelehnt (und damit  $H_A$  angenommen) ist. Falls  $Y \notin C$ , sagt man, dass  $H_0$  nicht abgelehnt ist (und  $H_A$  abgelehnt ist).

Sei nun  $(Y, C)$  ein Test der Nullhypothese  $H_0$  gegen  $H_A$ .

2. Die Hypothese  $H : \vartheta \in \Theta_H$  (die entweder  $H_0$  oder  $H_A$  sein kann) heißt einfach wenn  $\Theta_H = \{\vartheta^*\}$  für ein  $\vartheta^* \in \Theta$ . Andernfalls heißt  $H$  zusammengesetzt.
3. Ist  $\Theta = (\underline{\vartheta}, \bar{\vartheta})$  ein Intervall (wobei  $\underline{\vartheta} = -\infty$  und  $\bar{\vartheta} = \infty$  zugelassen sind und die Intervalle auch abgeschlossen sein können), so heißt der Test  $(Y, C)$  einseitig, falls  $\Theta_H = (\underline{\vartheta}, \vartheta^*)$  oder  $\Theta_H = (\vartheta^*, \bar{\vartheta})$ . Falls  $\Theta_H = (\vartheta^+, \vartheta^*)$  mit  $\underline{\vartheta} < \vartheta^+ \leq \vartheta^* < \bar{\vartheta}$ , so heißt der Test  $(Y, C)$  zweiseitig.
4. Der Test  $(Y, C)$  heißt unverfälscht, falls

$$\mathbf{P}_{\vartheta_0}(Y \in C) \leq \mathbf{P}_{\vartheta_A}(Y \in C)$$

für alle  $\vartheta_0 \in \Theta_0, \vartheta_A \in \Theta_A$  gilt.

**Bemerkung 6.14 (Interpretation und Fehler eines Tests).**

1. Einen statistischen Test hat man sich am besten so vorzustellen (siehe auch das nächste Beispiel): die Daten sind gegeben durch die Zufallsvariable  $X$ . Diese Daten fasst man durch die meist reellwertige Funktion  $t$  zusammen zur Teststatistik  $Y = t(X)$ . Die Daten können entweder nach  $\mathbf{P}_\vartheta$  mit  $\vartheta \in \Theta_0$  (d.h. die Nullhypothese ist richtig) oder mit

	$H_0$ abgelehnt	$H_0$ nicht abgelehnt
$H_0$ richtig	Fehler erster Art	richtige Entscheidung
$H_0$ falsch	richtige Entscheidung	Fehler zweiter Art

**Tabelle 6.1:** Die möglichen Fehler eines statistischen Tests.

$\vartheta \in \Theta_A$  (d.h. die Alternativhypothese ist richtig) verteilt sein. Ziel ist es, die Nullhypothese genau dann (anhand der Daten  $X$ ) abzulehnen, wenn  $H_A$  richtig ist. Der Ablehnungsbereich  $C$  ist so gewählt, dass  $H_0$  genau dann abgelehnt wird, wenn  $Y \in C$ . Dabei können zwei verschiedene Arten von Fehler auftreten; siehe auch Tabelle 6.1.

Gehen wir zunächst davon aus, dass  $\vartheta \in \Theta_0$ . Hat der Test ein Niveau  $\alpha$ , so wissen wir, dass  $\mathbf{P}_\vartheta(Y \in C) \leq \alpha$ . Da  $H_0$  genau dann abgelehnt wird, wenn  $Y \in C$ , wissen wir also, dass die Nullhypothese höchstens mit Wahrscheinlichkeit  $\alpha$  abgelehnt wird, wenn sie zutrifft. Damit hat man also die Wahrscheinlichkeit, die Nullhypothese abzulehnen, falls sie zutrifft, durch  $\alpha$  beschränkt. Falls  $Y \in C$ , aber  $\vartheta \in \Theta_0$ , die Nullhypothese also irrtümlicherweise verworfen wird, sprechen wir über einen *Fehler erster Art* (dessen Wahrscheinlichkeit durch  $\alpha$  kontrolliert wird).

Geht man davon aus, dass  $\vartheta \in \Theta_A$ , liegt eine Fehlentscheidung genau dann vor, wenn  $Y \notin C$ , die Nullhypothese also nicht abgelehnt wird. In diesem Fall sprechen wir von einem *Fehler zweiter Art*. Das Niveau des Tests liefert keinen Anhaltspunkt dafür, mit welcher Wahrscheinlichkeit ein solcher Fehler auftritt.

2. Auf den ersten Blick besteht eine scheinbare Symmetrie zwischen  $H_0$  und  $H_A$ . Schließlich lehnen wir  $H_0$  genau dann ab (und nehmen  $H_A$  an), wenn  $Y \in C$  und wir lehnen  $H_0$  nicht ab (und lehnen damit  $H_A$  ab) wenn  $Y \notin C$ . Allerdings wird diese Symmetrie durch das Niveau des Tests gebrochen. Weiß man, dass  $(Y, C)$  ein Test zum Niveau  $\alpha$  ist, bedeutet das, dass die Wahrscheinlichkeit, die Nullhypothese  $H_0$  abzulehnen, obwohl sie wahr ist, höchstens  $\alpha$  ist. Mit anderen Worten ist die Wahrscheinlichkeit für einen Fehler erster Art höchstens  $\alpha$ . Allerdings hat man keine Kontrolle über den Fehler zweiter Art.

Wegen dieser Asymmetrie ist in der Praxis die Nullhypothese genau so zu wählen, dass eine Ablehnung der Nullhypothese möglichst sicher auf die Richtigkeit der Alternativhypothese zurückzuführen ist. Wir betrachten das Beispiel der Münzwürfe aus Beispiel 6.1. Bevor wir die Daten des Experimentes erhoben haben, haben wir die Vorstellung, dass der Würfel gezinkt sein könnte. Außerdem legen wir ein Signifikanzniveau  $\alpha$  fest (was in der Praxis oft  $\alpha = 5\%$  ist). Um unsere Vorstellung über die Münze zu überprüfen, testen wir

$$\begin{aligned}
 H_0 &: \text{die Münze ist nicht gezinkt, } p = \frac{1}{2} \\
 &\text{gegen} \\
 H_A &: \text{die Münze ist gezinkt, } p \neq \frac{1}{2}.
 \end{aligned}$$

Kommt es nämlich jetzt zu einer Ablehnung von  $H_0$ , so wissen wir, dass dies mit Wahrscheinlichkeit höchstens  $\alpha$  dann passiert, wenn  $H_0$  wahr ist, die Münze also nicht gezinkt ist. Damit können wir uns relativ sicher sein, dass die Ablehnung der Nullhypothese dar-



auf zurückzuführen ist, dass  $H_A$  zutrifft. Damit ist unsere Vorstellung, dass die Münze gezinkt ist, bei Ablehnung der Nullhypothese höchstwahrscheinlich bestätigt.

- Die Forderung von unverfälschten Tests ist klar zu verstehen: Da wir  $H_0$  dann ablehnen, wenn  $Y \in C$ , soll zumindest die Wahrscheinlichkeit, dass  $H_0$  abgelehnt wird, unter  $\mathbf{P}_{\vartheta_A}$ ,  $\vartheta_A \in \Theta_A$  größer sein als für  $\mathbf{P}_{\vartheta_0}$ ,  $\vartheta_0 \in \Theta_0$ .

**Bemerkung 6.15 ( $p$ -Werte und alternative Definition eines Tests).**

- Sei  $Y = y$ , d.h. dass die Teststatistik  $Y$ , angewendet auf die echten Daten, ergibt  $y$ . Dann heißt der Wert

$$p_y := \sup_{\vartheta \in \Theta_0} \mathbf{P}_{\vartheta}(Y \text{ mindestens so extrem wie } y)$$

$p$ -Wert des Tests für  $Y = y$ . Dabei hängt die Bedeutung davon, was 'extrem' heißt davon ab, was genau die Alternative ist. (Dies ist oftmals in konkreten Beispielen einfach zu verstehen, siehe etwa den Binomialtest, Abbildung 6.3.) Immer gilt jedoch  $p_y \leq p_{y'}$ , falls  $y$  mindestens so extrem wie  $y'$  ist. Es ist wichtig zu beachten, dass es dadurch einen engen Zusammenhang zwischen dem Niveau  $\alpha$  des Tests und dem  $p$ -Wert gibt. Ist nämlich  $(Y, C)$  ein Test zum Niveau  $\alpha$  und

$$C = \{y : y \text{ mindestens so extrem wie } y_0\}$$

für ein  $y_0$ , so wird  $H_0$  genau dann abgelehnt, wenn

$$\alpha \geq \sup_{\vartheta \in \Theta_0} \mathbf{P}_{\vartheta}(Y \text{ mindestens so extrem wie } y_0) = p_{y_0}.$$

Ist  $Y = y$  und gilt  $p_y \leq p_{y_0}$ , so wird  $H_0$  also abgelehnt. Es genügt also, für einen Test zum Niveau  $\alpha$  und  $Y = y$  den Wert  $p_y$  zu bestimmen. Ist  $p_y \leq \alpha$ , so wird  $H_0$  abgelehnt. Dieses Vorgehen wird bei vielen Statistik-Programmen angewendet, bei denen ausschließlich  $p$ -Werte ausgegeben werden. Dabei muss man meist angeben, was genau die Alternative ist (einseitig oder zweiseitig), damit das Programm weiß, in welche Richtungen Abweichungen als extrem zu betrachten sind.

Wir formalisieren nun noch das Eingangsbeispiel 6.1, was uns zum Binomialtest führt.

**Proposition 6.16 (Binomialtest).** Sei  $\alpha \in [0, 1]$ ,  $n \in \mathbb{N}$  und  $(X, (\mathbf{P}_p)_{p \in [0,1]})$  ein statistisches Modell, so dass  $X$  unter  $\mathbf{P}_p$  nach  $B(n, p)$  verteilt ist.

- Ist  $\Theta_0 = p^*$ ,  $\Theta_A = \Theta \setminus \Theta_0$ , so ist  $(X, \{0, \dots, k\} \cup \{l, \dots, n\})$  ein unverfälschter Test zum Niveau  $\alpha$ , falls

$$\mathbf{P}_{p^*}(X \leq k) \leq \alpha/2, \quad \mathbf{P}_{p^*}(X \geq l) \leq \alpha/2.$$

- Ist  $\Theta_0 = [0, p^*]$ ,  $\Theta_A = \Theta \setminus \Theta_0$ , so ist  $(X, \{k, \dots, n\})$  ein unverfälschter Test zum Niveau  $\alpha$ , falls

$$\mathbf{P}_{p^*}(X \geq k) \leq \alpha.$$

- Ist  $\Theta_0 = [p^*, 1]$ ,  $\Theta_A = \Theta \setminus \Theta_0$ , so ist  $(X, \{0, \dots, k\})$  ein unverfälschter Test zum Niveau  $\alpha$ , falls

$$\mathbf{P}_{p^*}(X \leq k) \leq \alpha.$$

**Der Binomialtest**

überprüft, ob bestimmte Erfolgswahrscheinlichkeiten einer Binomialverteilung angenommen werden.

Statistisches Modell  $X$  unter  $\mathbf{P}_p$  nach  $B(n, p)$  verteilt

Hypothesen (a)  $H_0 : p \in \{p^*\}$  gegen  $H_A : p \notin \{p^*\}$   
 (b)  $H_0 : p \in [0, p^*]$  gegen  $H_A : p \in (p^*, 1]$   
 (c)  $H_0 : p \in [p^*, 1]$  gegen  $H_A : p \in [0, p^*)$

Teststatistik  $X$  unter  $\mathbf{P}_p$  verteilt nach  $B(n, p)$

Ablehnungsbereich (a)  $\{0, \dots, k, l, \dots, n\}$  mit  $\mathbf{P}_{p^*}(X \leq k), \mathbf{P}_{p^*}(X \geq l) \leq \alpha/2$ ,  
 (b)  $\{l, \dots, n\}$  mit  $\mathbf{P}_{p^*}(X \geq l) \leq \alpha$   
 (c)  $\{0, \dots, k\}$  mit  $\mathbf{P}_{p^*}(X \leq k) \leq \alpha$ ,

$p$ -Wert, falls  $X = x$  (a)  $\mathbf{P}_{p^*}(X \leq x \wedge (2p^*n - x)) + \mathbf{P}_{p^*}(X \geq x \vee (2p^*n - x))$   
 (b)  $\mathbf{P}_{p^*}(X \geq x)$   
 (c)  $\mathbf{P}_{p^*}(X \leq x)$

**Abbildung 6.3:** Schematische Darstellung des Binomialtests aus Proposition 6.16

*Beweis.* Wir beweisen nur (c), die anderen beiden Aussagen folgen analog. Klar ist, dass der Test unverfälscht ist. Es ist außerdem

$$\sup_{p \in \Theta_0} \mathbf{P}_p(X \in \{0, \dots, k\}) = \mathbf{P}_{p^*}(X \leq k) \leq \alpha$$

nach Voraussetzung. Also folgt bereits die Aussage.  $\square$

**Beispiel 6.17 (Binomialtest).** Sei  $\alpha = 5\%$ ,  $n = 100$  und  $(X, (\mathbf{P}_p)_{p \in [0,1]})$  wie in der Proposition. Wir wollen nun

$$H_0 : p = 1/2 \text{ gegen } H_A : p \neq 1/2$$

testen, wenn wir in 100 Versuchen 59 Erfolge erzielt haben. Nach Proposition 6.16 ist der kritische Bereich von der Form  $\{0, \dots, k\} \cup \{l, \dots, 100\}$ . Es ist

$$\mathbf{P}_{p=1/2}(X \leq 40) + \mathbf{P}_{p=1/2}(X \geq 60) \approx 5.69\%.$$

Da  $40 < 59 < 60$ , liegt 59 nicht im Ablehnungsbereich von  $H_0$ . Damit kann die Nullhypothese aufgrund der Daten ( $X = 59$ ) nicht abgelehnt werden. Auf dasselbe Ergebnis kommt man mit Hilfe des  $p$ -Wertes. Es ist

$$\mathbf{P}_{p=1/2}(X \text{ mindestens so extrem wie } 59) = \mathbf{P}_{p=1/2}(X \leq 41) + \mathbf{P}_{p=1/2}(X \geq 59) \approx 8.86\%.$$

Da dieser Wert größer als  $\alpha = 5\%$  ist, kann man die Nullhypothese nicht ablehnen.

Wir behandeln nun noch einen Test der auf normalverteilten Daten beruht. Um die Verteilung der Teststatistik anzugeben, benötigen wir hierfür eine Definition.

**Definition 6.18 (Student- $t$ -Verteilung).** Sei  $n > 1$  und  $\underline{X} = (X_1, \dots, X_n)$  unabhängig und nach  $N(0, 1)$  verteilt, sowie  $\bar{X}$  der Mittelwert und  $s^2(\underline{X})$  die empirische Varianz. Dann heißt die Verteilung von

$$T := \frac{\bar{X}}{\sqrt{s^2(\underline{X})/n}}$$

(Student-) $t$ -Verteilung mit  $n - 1$  Freiheitsgraden. Diese bezeichnen wir auch mit  $t(n)$ .

**Bemerkung 6.19 (Transformation von  $N(\mu, \sigma^2)$ -verteilten Zufallsvariablen).** Sei  $n > 1$  und  $\underline{X} = (X_1, \dots, X_n)$  unabhängig und nach  $N(\mu, \sigma^2)$  verteilt. Man überlegt sich leicht, dass dann  $Y_1, \dots, Y_n$  mit  $Y_i := \frac{X_i - \mu}{\sqrt{\sigma^2}} \sim N(0, 1)$ . Deswegen gilt in dieser Situation, dass

$$\frac{\bar{Y}}{\sqrt{s^2(\underline{Y})}} = \frac{\bar{X} - \mu}{\sqrt{s^2(\underline{X})}} \sim t(n - 1).$$

**Proposition 6.20 (Einfacher  $t$ -Test).**

Sei  $\alpha \in [0, 1]$ ,  $\mu^* \in \mathbb{R}$  und  $(X = (X_1, \dots, X_n), (\mathbf{P}_{\mu, \sigma^2})_{\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+})$  ein statistisches Modell, so dass  $X_1, \dots, X_n$  unter  $\mathbf{P}_{\mu, \sigma^2}$  unabhängig und nach  $N(\mu, \sigma^2)$  verteilt sind. Weiter sei

$$T := \frac{\bar{X} - \mu^*}{\sqrt{s^2(\underline{X})/n}}$$

und  $t_{n,p}$  für  $p \in [0, 1]$  das  $p$ -Quantil von  $t(n)$ .

- (a) Ist  $\Theta_0 = \{\mu^*\} \times \mathbb{R}_+$ ,  $\Theta_A = \Theta \setminus \Theta_0$ , so ist  $(T, (-\infty, t_{n-1, \alpha/2}) \cup (t_{n-1, 1-\alpha/2}, \infty))$  ein unverfälschter Test zum Niveau  $\alpha$ .
- (b) Ist  $\Theta_0 = (-\infty, \mu^*] \times \mathbb{R}_+$ ,  $\Theta_A = \Theta \setminus \Theta_0$ , so ist  $(T, [t_{n-1, 1-\alpha}, \infty))$  ein unverfälschter Test zum Niveau  $\alpha$ .
- (c) Ist  $\Theta_0 = [\mu^*, \infty) \times \mathbb{R}_+$ ,  $\Theta_A = \Theta \setminus \Theta_0$ , so ist  $(T, (-\infty, t_{n-1, \alpha}])$  ein unverfälschter Test zum Niveau  $\alpha$ .

*Beweis.* Wieder beweisen wir nur (c), da die anderen beiden Aussagen analog folgen. Klar ist, dass der Test unverfälscht ist. Nach Definition der  $t$ -Verteilung folgt, dass  $T$  unter  $\mathbf{P}_{\mu^*, \sigma^2}$  nach  $t(n - 1)$  verteilt ist. Damit gilt

$$\sup_{\mu \geq \mu^*} \mathbf{P}_{\mu, \sigma^2}(T \leq t_{n-1, \alpha}) = \mathbf{P}_{\mu^*, \sigma^2}(T \leq t_{n-1, \alpha}) = \alpha,$$

woraus die Behauptung sofort folgt. □

**Beispiel 6.21 (Normalverteilte Schlafdauern).** Wir untersuchen ein Beispiel aus der Medizin. Ein Medikament wird daraufhin untersucht, ob es den Schlaf von Probanden verlängert. Dazu wird jeweils die Schlafdauerdifferenz bei zehn Patienten notiert. Man erhält

1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4.

**Der einfache  $t$ -Test**

überprüft, ob der Erwartungswert einer Normal-Verteilung (oder irgendeiner Verteilung bei approximativ unendlich großer Samplegröße) gleich einer vorgegebenen Größe  $\mu^*$  ist, wenn die Varianz unbekannt ist.

Statistisches Modell  $X = (X_1, \dots, X_n)$  und  $X_1, \dots, X_n$  unter  $\mathbf{P}_{\mu, \sigma^2}$  unabhängig und nach  $N(\mu, \sigma^2)$  verteilt

Hypothesen (a)  $H_0 : \mu \in \{\mu^*\}$  gegen  $H_A : \mu \notin \{\mu^*\}$   
 (b)  $H_0 : \mu \in (-\infty, \mu^*]$  gegen  $H_A : \mu \in (\mu^*, \infty)$   
 (c)  $H_0 : \mu \in [\mu^*, \infty)$  gegen  $H_A : \mu \in (-\infty, \mu^*)$

Teststatistik  $T = \frac{\bar{X} - \mu^*}{\sqrt{s^2(X)/n}}$  unter  $\mathbf{P}_{\mu^*, \sigma^2}$  verteilt nach  $t(n-1)$

Ablehnungsbereich (a)  $(-\infty, t_{n-1, \alpha/2}) \cup (t_{n-1, 1-\alpha/2}, \infty)$ ,  
 (b)  $(t_{n-1, 1-\alpha}, \infty)$   
 (c)  $(-\infty, t_{n-1, \alpha})$

$p$ -Wert, falls  $T = t$  (a)  $2(1 - \mathbf{P}[T \leq |t|])$ , wenn  $T$  nach  $t(n-1)$  verteilt ist  
 (b)  $\mathbf{P}[T \geq t]$   
 (c)  $\mathbf{P}[T \leq t]$

**Abbildung 6.4:** Schematische Darstellung des einfachen  $t$ -Tests aus Proposition 6.20

Diese Beobachtungen betrachten wir als Realisierungen von Zufallsvariablen  $X_1, \dots, X_{10}$ , die nach  $N(\mu, \sigma^2)$  verteilt sind, wobei weder  $\mu$  noch  $\sigma^2$  bekannt sind. Wir berechnen

$$\bar{X} = 2.33, \quad s^2(X) = 4.01.$$

Damit ist

$$T = 2.33/\sqrt{4.01/10} \approx 3.68.$$

Testet man also

$$H_0 : \mathbf{E}_{\mu, \sigma^2}[\bar{Y} - \bar{X}] = 0 \text{ gegen } \mathbf{E}_{\mu, \sigma^2}[\bar{Y} - \bar{X}] \neq 0,$$

auf dem Niveau 5%, so ist dessen Ablehnungsbereich  $C = (-\infty, -2.262) \cup (2.262, \infty)$ . Da  $3.68 \in C$ , kann man  $H_0$  auf dem Niveau 5% verwerfen. Das bedeutet, dass vermutlich eine Veränderung der Schlafdauer durch Einnahme des Medikamentes stattfand.

**Bemerkung 6.22 (Kontingenztafeln und Unabhängigkeitstests).** Wir kommen nun zum letzten hier behandelten Test, Fisher's exaktem Test. Gegeben seien Items mit zwei Merkmalen. (Beispielsweise Pflanzenblätter, die sowohl eine Farbe als auch eine bestimmte Form haben.) Solche Daten lassen sich in einer Kontingenztafel zusammenfassen; siehe unten. Falls beide Merkmale genau zwei verschiedene Möglichkeiten haben, lässt sich mit Hilfe des exakten Tests von Fisher bestimmen, ob beide Ausprägungen unabhängig sind.

	Merkmal 2, Möglichkeit 1	Merkmal 2 Möglichkeit 2	$\Sigma$
Merkmal 1, Möglichkeit 1	$S_{11}$	$S_{12}$	$S_{1\bullet}$
Merkmal 1, Möglichkeit 2	$S_{21}$	$S_{22}$	$S_{2\bullet}$
$\Sigma$	$S_{\bullet 1}$	$S_{\bullet 2}$	$S$

**Proposition 6.23 (Fisher's exakter Test).** Sei  $\alpha \in [0, 1]$ ,  $\mathcal{I}, \mathcal{J}$  endliche Mengen mit  $|\mathcal{I}| = |\mathcal{J}| = 2$  und

$$\Theta = \{\underline{p} = (p_{ij})_{i \in \mathcal{I}, j \in \mathcal{J}} \text{ Verteilungsgewichte einer Verteilung auf } \mathcal{I} \times \mathcal{J}\}.$$

Weiter sei  $((X, Y) = ((X_1, Y_1), \dots, (X_n, Y_n)), (\mathbf{P}_{\underline{p}})_{\underline{p} \in \Theta})$  ein statistisches Modell, so dass  $(X_1, Y_1), \dots, (X_n, Y_n)$  unter  $\mathbf{P}_{\underline{p}}$  unabhängig und nach  $\underline{p}$  verteilt sind (d.h.  $\mathbf{P}(X_k = i, Y_k = j) = p_{ij}$ ). Setze für  $i \in \mathcal{I}, j \in \mathcal{J}$

$$\begin{aligned} S_{ij} &:= |\{k : X_k = i, Y_k = j\}|, \\ S_{i\bullet} &:= |\{k : X_k = i\}| = \sum_{j \in \mathcal{J}} S_{ij}, \\ S_{\bullet j} &:= |\{k : Y_k = j\}| = \sum_{i \in \mathcal{I}} S_{ij} \end{aligned}$$

und  $p_{i\bullet} := \sum_{j \in \mathcal{J}} p_{ij}, p_{\bullet j} := \sum_{i \in \mathcal{I}} p_{ij}$  für  $\underline{p} \in \Theta$ . Sei

$$\Theta_0 = \{\underline{p} \in \Theta : p_{ij} = p_{i\bullet} \cdot p_{\bullet j} \text{ für } i \in \mathcal{I}, j \in \mathcal{J}\},$$

$\Theta_A = \Theta \setminus \Theta_0$ . Dann ist  $(S_{11}, C)$  ein unverfälschter Test zum Niveau  $\alpha$ , falls

$$H(S_{1\bullet}, S, S_{\bullet 1})(C) \leq \alpha. \tag{6.2}$$

*Beweis.* Gegeben sind Daten von  $S$  Objekten. Wir gehen davon aus, dass  $S_{1\bullet}, S_{2\bullet}, S_{\bullet 1}$  und  $S_{\bullet 2}$  bekannt sind. In diesem Fall gilt es, die  $S$  Objekte so auf die Kontingenztafel zu verteilen, dass die Randeinträge stimmen. Wir stellen uns vor, die  $S_{\bullet 1}$  Elemente mit Merkmal 2, Möglichkeit 1 seien weiß, die anderen  $S_{\bullet 2}$  schwarz. Aus diesen wählen wir  $S_{1\bullet}$  aus. Sind die Merkmale wirklich unabhängig, so erhalten wir eine  $H(S_{1\bullet}, S, S_{\bullet 1})$  verteilte Zufallsgröße als Eintrag von  $S_{11}$ . Ist diese entweder zu groß oder zu klein, können wir die Unabhängigkeit verwerfen.  $\square$

**Bemerkung 6.24.** In (6.2) besteht eine vermeintliche Unsymmetrie zwischen dem ersten und zweiten Merkmal, da die Parameter der hypergeometrischen Verteilung  $S_{1\bullet}$  die Anzahl der gezogenen und  $S_{\bullet 1}$  die Anzahl der weißen Kugeln bedeutet. Allerdings gilt

$$\begin{aligned} H(S_{1\bullet}, S, S_{\bullet 1})(k) &= \frac{\binom{S_{\bullet 1}}{k} \binom{S_{\bullet 2}}{S_{1\bullet} - k}}{\binom{S}{S_{1\bullet}}} = \frac{S_{\bullet 1}! S_{\bullet 2}! S_{1\bullet}! S_{2\bullet}!}{S! k! (S_{\bullet 1} - k)! (S_{1\bullet} - k)! (S - S_{\bullet 1} - S_{1\bullet} + k)!} \\ &= H(S_{\bullet 1}, S, S_{1\bullet})(k), \end{aligned}$$

was diese vermeintliche Unsymmetrie erklärt.

#### Fisher's exakter Test

überprüft, ob zwei Merkmale, die in jeweils zwei möglichen Ausprägungen vorliegen, stochastisch unabhängig sind.

Statistisches Modell  $(X_1, Y_1), \dots, (X_n, Y_n)$  unabhängig, identisch verteilt,

$$S_{ij} := |\{k : X_k = i, Y_k = j\}|,$$

$$S_{i\bullet} := |\{k : X_k = i\}| = \sum_{j \in \mathcal{J}} S_{ij},$$

$$S_{\bullet j} := |\{k : Y_k = j\}| = \sum_{i \in \mathcal{I}} S_{ij}.$$

Hypothesen  $H_0 : \mathbf{P}(X_k = i, Y_k = j) = \mathbf{P}(X_k = i) \cdot \mathbf{P}(Y_k = j)$  für alle  $i, j$

$H_A : \mathbf{P}(X_k = i, Y_k = j) \neq \mathbf{P}(X_k = i) \mathbf{P}(Y_k = j)$  für mindestens ein Paar  $i, j$

Teststatistik  $S_{11}$  ist – gegeben  $S_{1\bullet}, S_{2\bullet}, S_{\bullet 1}$  und  $S_{\bullet 2}$  – nach  $H(S_{1\bullet}, S, S_{\bullet 1})$  verteilt.

Ablehnungsbereich  $S_{11} \in C := \{0, \dots, k, \ell, \dots, S_{1\bullet} \wedge S_{\bullet 1}\}$  falls  $H(S_{1\bullet}, S, S_{\bullet 1})(C) \leq \alpha$ .

**Abbildung 6.5:** Schematische Darstellung von Fisher's exaktem Test

**Beispiel 6.25 (Mendel's Experimente).** Wir betrachten nochmal die beiden Merkmale der Blütenfarbe und Blattmuster einer Pflanze. Beispielsweise könnte ein Züchter von Pflanzen zählen, wieviele seiner Pflanzen rot und weiß blühen, und wieviele glatte und gesprenkelte Blätter haben. Wir gehen davon aus, dass  $(X_i, Y_i)$  die Farbe und die Blättermusterung der

$i$ -ten Pflanze ist. Ziel ist es herauszufinden, ob  $X_i$  und  $Y_i$  stochastisch unabhängig sind, d.h. ob etwa

$$\mathbf{P}(X_i = \text{rot}, Y_i = \text{glatt}) = \mathbf{P}(X_i = \text{rot}) \cdot \mathbf{P}(Y_i = \text{glatt})$$

(und ähnliche Gleichungen für die anderen Farben/Blattmusterungen) gilt. Bei einem Versuch, in dem man 400 Blumen als Stichprobe untersucht hat, erhält man folgende Daten, die wir in einer Kontingenztabelle zusammen fassen:

	glatt	gesprenkelt	$\Sigma$
weiß	41	74	115
rot	96	189	285
$\Sigma$	137	263	400

Wir bezeichnen diese Anzahlen mit  $S_{\text{weiß, glatt}}, \dots, S_{\text{rot, gesprenkelt}}$ . Die Marginalien bezeichnen wir mit  $S_{\text{weiß}\bullet}, \dots, S_{\text{rot}\bullet}$  und  $S_{\bullet\text{glatt}}, S_{\bullet\text{gesprenkelt}}$ . Kann aufgrund dieser Daten die Hypothese der Unabhängigkeit auf einem Signifikanzniveau von 5% verworfen werden? Wären die beiden Merkmale unabhängig, so wäre  $\mathbf{E}[S_{11}] = 400 \cdot \frac{137}{400} \cdot \frac{115}{400} \approx 39.39$ . Beobachtet wurden aber  $S_{11} = 41$ .

Zunächst stellen wir fest, dass

$$H(115, 400, 137)\{0, \dots, 30\} \approx 0.018 < 0.032 = H(115, 400, 137)\{0, \dots, 31\}$$

und

$$H(115, 400, 137)\{49, \dots, 115\} \approx 0.018 < 0.030 = H(115, 400, 137)\{48, \dots, 115\}.$$

Damit ist

$$C = \{0, \dots, 30, 49, \dots, 115\}$$

der Ablehnungsbereich für Fisher's exakten Test. Damit kann die Nullhypothese einer unabhängigen Ausprägung der Blütenfarbe und Blattmusterung nicht abgelehnt werden.