Anton Bovier

Stochastik für das Lehramt – ALMA2, Teil 1

Vorlesung Sommer 2015, Bonn

16. Juli 2015

universitätbonn iam

Inhaltsverzeichnis

Kapitel 1
Wahrscheinlichkeit

Il est remarquable qu'une science, qui a commencé par la considération des jeux, ce soit élévée aux plus importants objects des connaissances humaines[a].
Pierre Simon de Laplace, Théorie Analytique des Probabilités

[a] Es ist bemerkenswert, dass eine Wissenschaft, die mit der Betrachtung von Glücksspielen begonnen hat, sich zu einem der wichtigsten Gegenstände der menschlichen Erkenntnis erhoben hat.

In dieser Vorlesung werden wir ein Gebiet der Mathematik behandeln, dass sich von anderen dadurch hervorhebt, dass viele seiner Begriffe weitgehend Eingang in die Umgangssprache gefunden haben, ja, dass Fragen behandelt werden, die viele Menschen im täglichen Leben betreffen und von denen fast jedermann gewisse, ob falsche oder richtige, Vorstellungen hat.

Der zentrale Begriff, der uns hier beschäftigt, ist der des *Zufalls*. Was Zufall ist, oder ob es so etwas überhaupt gibt, ist eine tiefe philosphische Frage, der wir uns hier nur in wenigen Punkten annähern können; sie ist auch nicht der zentrale Gegenstand der Vorlesung. Grob gesprochen reden wir von "Zufall", wenn es sich um den Eintritt von *Ereignissen* handelt, die wir nicht oder nicht im Detail vorhersehen können. Typischerweise sind für ein solches Ereignis mehrere Varianten möglich, und wir reden von der Wahrscheinlichkeit des einen oder anderen Ausgangs. Ein beliebtes Beispiel ist etwa die Frage, ob es morgen regnet. In vielen Fällen ist dies möglich, aber nicht sicher. Der Wetterbericht macht darüber zwar Vorhersagen, aber auch diese treffen nur "mit einer gewissen Wahrscheinlichkeit ein". Wir können die Frage auch noch weiter spezifizieren, etwa danach wieviel Regen morgen fallen wird, und werden noch weniger sichere Vorhersagen bekommen. Gleiches gilt für sehr viele Vorkommnisse des täglichen Lebens. Der Begriff des Zufalls und der Wahrscheinlichkeit wird gebraucht, um solche Unsicherheiten qualitativ und quantitativ genauer zu beschreiben.

Unsicherheit tritt in vielen Situationen auf und wird sehr unterschiedlich wahrgenommen. Vielfach betrachten wir sie als Ärgernis und suchen eigentlich nach einer deterministischen Gesetzmässigkeit, die genauere Vorhersagen erlaubt. Dies betrifft insbesondere viele Bereiche von Naturwissenschaft und Technik, wo uns der Zufall vielfach nur in der Form von "Fehlern" und Ungenauigkeiten begegnet, und wir bestrebt sind seine Effekte möglichst zu eleminiern oder doch zu minimieren.

In anderen Fällen ist der Zufall wesentlicher Motor des Geschehens und seine Existenz ist sogar gewollt und wird gezielt ausgenutzt. Am ausgeprägtesten ist dies

sicher im *Glückspiel*, und in vieler Hinsicht ist hier die Wahrscheinlichkeitstheorie genuin zuhause and kann in ihrer reinsten Form beobachtet werden. Wie das Zitat von Laplace am Anfang dieses Kapitels belegt, sind die grundlegenden Prinzipien der Wahrscheinlichkeitstheorie zunächst in diesem Kontext entwickelt worden. In diesem Zusammenhang steht auch der Erfolg der Wahrscheinlichkeitstheorie unter dem Namen Finanzmathematik. Interessanterweise sind viele der mathematischen Prinzipien die hier entwickelt wurden, von der genauen Interpretation von Zufall gar nicht abhängig.

In dieser Vorlesung werden wir uns weitgehend in einem mathematisch recht einfachen zu bewältigenden Rahmen bewegen. Eine tiefere und wesentlich allgemeinere Darstellung wird in der Vorlesung Ëinführung in die Wahrscheinlichkeitstheorie gegeben werden".

Literaturhinweise: Es gibt eine grosse Zahl von Lehrbüchern zur Wahrscheinlichkeitstheorie. Eine elementare schöne Einführung ist ein neues Buch von Kersting und Wakolbinger [3]. Eine ebenfalls elementare Darstellung ist das Buch von Henze [2]. Wer mehr wissen möchte, kann das schöne Buch von Georgii [1]zu Rate ziehen.

1.1 Zufallsexperimente und Glückspiele

Die meisten klassischen Glückspiele beruhen auf einer Vorrichtung, die es erlaubt in unvorhersahbarer Weise wiederholbar eines aus einer Reihe möglicher Ausgänge eines Experiments zu produzieren. Typische Beispiele sind:

- **Münzwurf.** Eine Münze mit zwei unterschiedlich bedruckten Seiten ("Kopf" und "Zahl") wird in die Luft geworfen. Sie kommt schließlich auf dem Boden zu liegen und zeigt nun mit einer ihrer Seiten nach oben. Diese zwei möglichen Ausgänge stellen die zwei Ereignisse "Kopf" oder "Zahl" dar. Wir gehen davon aus, dass es uns nicht möglich ist den Ausgang vorherzusehen, wir betrachten diesen als völlig zufällig [dies mag eine Idealisierung sein, da ein sehr geschickter Münzwerfer den Ausgang des Experiments beeinflussen kann. Wir wollen hiervon aber absehen]. Wichtig ist hier, dass wir einen solchen Wurf beliebig oft wiederholen können, ohne irgendeine zusätzliche Information über den Ausgang des nächsten Wurfes zu bekommen.
- **Roulette.** Hier wird eine Kugel auf eine sich drehende Scheibe geworfen, die 37 numerierte identische Vertiefungen enthält, in einer von denen die Kugel am Ende des Experiments liegenbleibt. Auch hier wird eines der 37 möglichen Ereignisse in unvorhersehbarer Weise realisiert.
- **Würfeln.** Ähnlich wie der Münzwurf, es sind hier aber 6 Ereignisse möglich.
- **Lotto.** Aus einem Behälter, der 49 numerierte Kugeln enthält, werden 6 davon mit einem komplizierten Mechanismus herausgefischt. Aufgrund der Durchmischung am Anfang ist das Ergebnis nicht vorhersehbar. Die möglichen Ereignisse sind "sechs Zahlen aus den 49 ersten natürlichen Zahlen", zum Beispiel

$3,8,19,23,25,45$. Die Zahl der möglichen Ausgänge ist recht gross, nämlich $49!/43!/6! = \binom{49}{6} = 1\,398\,316$.

- **Zufallszahlengeneratoren.** Zufallszahlengeneratoren sind numerische Algorithmen, mit denen ein Computer Zahlenreihen (etwa aus $\{0,1\}$) produziert, die möglichst zufällig sein sollen. In Wirklichkeit sind diese Reihen allerdings völlig deterministisch, können aber sehr irregulär von einem Anfangswert ("seed") abhängen. Die Erzeugung von Zufallszahlen ist ein wichtiges Problem, dem wir uns aber zunächst nicht weiter widmen wollen.

Wir wollen die Durchführung eines solchen "Experiments" in Zukunft als *Zufallsexperiment* bezeichnen. Jedem Zufallsexperiment kommt eine Menge möglicher Ausgänge zu. Diese Menge bezeichnen wir meißt mit Ω.

Im Kontext von Glückspielen stellen wir uns vor, dass ein gegebenes Spiel beliebig oft in identischer Weise wiederholt werden kann, wobei der Ausgänge der einzelnen Spiele nicht voneinander abhängen und jeweils die gleichen "Wahrscheinlichkeiten"haben. Dabei ist eine Interpretation von Wahrscheinlichkeit einfach die "Häufigkeit"bzw. Frequenzmit der ein bestimmter Ausgang stattfindet. Nehmen wir als Beispiel den Wurf einer Münze mit den Ausgängen "Kopföder SZahl". Im Fall einer fairen Münze würden wir erwarten, dass

$$\frac{\#\{\text{Kopf in } n\text{Würfen}\}}{n} \to \frac{1}{2}, \tag{1.1}$$

d.h. auf Dauer kommen in der Hälfte der Würfe Kopf auf, was wir als "Kopf hat Wahrscheinlichkeit $1/2$" interpretieren können. Falls die Münze nicht fair ist, könnte die linke Seite in Gl. (1.1) gegen einen anderen Wert konvergieren, $p(\text{Kopf})$, und wir würden sagen, "Kopf hat Wahrscheinlichkeit $p(\text{Kopf})$".

Wir können aber auch eine andere Interpretation fuer diese Wahrscheinlichkeit finden. Dazu betrachten wir eine Einheitswetteäuf das Ereignis "Kopf": Gegen die Bezahlung eines bestimmten Preises p erhalten wir im Falle dass der Münzwurf mit Kopf ausgeht, einen Euro. Dieses Spiel ist für uns insbesondere dann interessant, wenn wir aus irgendeinem Grunde einen Euro verlieren, wenn das Ereignis Kopf eintritt. Gehen wir nämlich diese Einheitswette ein, so können wir den Euro aus unserem Gewinn bezahlen, unsere Kosten sind also unabhängig vom Ausgang der Spiels immer gerade der Preis, p, der Wette. Die Wette hat also in diesem Fall den Charakter einer Versicherung: gegen eine Prämie p versichern wir und gegen das Risiko, gegebenenfalls einen Euro bezahlen zu müssen. Wieviel sollten wir zu zahlen bereit sein? Im Kontext unseres wiederholbaren Glücksspiels kann man sich dies leicht überlegen. Dazu müssen wir überlegen, was passiert, wenn wir in n solchen Spielen jeweils eine solche Wette eingehen. Unser Kapital nach n Spielen ist dann nämlich

$$-np + \#\{\text{Kopf in } n\text{Würfen}\} = n\left(\frac{\#\{\text{Kopf in } n\text{Würfen}\}}{n} - p\right). \tag{1.2}$$

Falls (1.1) gilt, so wird dies jedenfalls nach $+\infty$ tendieren, falls $p < 1/2$, und nach $-\infty$, falls $p > 1/2$. Wir würden also jeden Preis der kleiner ist als $1/2$ gerne bezahlen

aber sicher keinen Preis der höher als $1/2$ ist akzeptieren. Umgekehrt würde und wohl niemand diese Wette für weniger als $1/2$ anbieten. Damit bleibt nur $p = 1/2$ als *fairer Preis* übrig. Falls die linke Seite von (1.1) gegen einen anderen Wert, $p(\text{Kopf})$ konvergiert, so wäre entsprechend dies der faire Preis für unsere Wette.

Wir können also auch postulieren, dass die Wahrscheinlichkeit einen Ereignisses der *faire Preis* einer Einheitswette auf das Eintreten dieses Ereignisses ist. Diese Definition hat den Vorteil, dass sie auch für den Fall nicht wiederholbarer Zufallsexperimente anwendbar ist. Die Frage, wie dieser faire Preis zu ermitteln ist, ist dann natürlich offen.

Das elementarste Zufallsmodell.

Wir wollen unser Münzwurfmodell nun etas mathematischer fassen. Wir benötigen die folgenden Ingredienzien.

- Mögliche Ausgänge des Spiels: $\Omega \equiv \{0,1\}$; (die unpraktischen "Kopf" und "Zahl" stellen wir lieber als 0 und 1 dar.
- Die möglichen Wetten: sinnvoll können wir nur auf die elementaren Ausgänge 0 und 1 setzen; daneben gäbe es aber noch die triviale Wette "der Ausgang ist entweder 0 oder 1, d.h. wir setzen auf Ω, und die idiotische Wette ës kommt weder 0 noch 1", wir sagen dann, dass wir auf die leere Menge $\emptyset$ setzen. Wir werden die Mengen der möglichen Wetten mit $\mathfrak{F}$ bezeichnen. In unserem Fall können wir diese mit der Potenzmenge von Ω,

$$\mathscr{P}(\{0,1\}) = \{\emptyset, \{0\}, \{1\}, \{0,1\}\}, \tag{1.3}$$

  identifizieren.
- Jede Wette braucht nun einen fairen Preis, bzw. eine Häufigkeit. Die Zuordnung eines fairen Preises zu den Elementen von $\mathfrak{F}$ werden wir *Wahrscheinlichkeitmaß* nennen und mit $\mathbb{P}$ bezeichnen. Klarerweise ist sowohl der Wert der Wette auf $\emptyset$ gleich Null als auch die Häufigkeit von $\emptyset$ gleich null. Genauso trivial ist, dass der Wert der Wette auf Ω gleich eins sein muss, da ja in jedem Spiel sicher der Betrag von einem Euro ausgezahlt wird. Wir müssen also nur noch die Wahrscheinlichkeiten von $p(0)$ und $p(1)$ festlegen.

Für die Wahl dieser Werte gibt es nun einige Einschränkungen. Klarerweise liegen $p(0)$ und $p(1)$ im Intervall $[0,1]$, und es muss gelten, dass

$$p(0) + p(1) = 1. \tag{1.4}$$

Letzteres ist offensichtlich, wenn die p's als Frequenz interpretiert werden. Aber auch als fairer Preis ist dies offenbar notwendig: Ist nämlich $p(0) + p(1) < 1$, so würden wir einfach beide Wetten eingehen, weniger als 1 Euro Einsatz zahlen, und sicher einen ganzen Euro Gewinn einziehen. Damit hätten wir einen sicheren Gewinn, was für den Verkäufer der Wetten einen sicheren Verlust bedeutete. Sollten wir dem Wettanbieter nicht trauen, wenn $p(0) + p(1) > 1$. Dann nämlich macht die

Bank einen größeren Gewinn als die Spieler: Das Spiel auf die 0, was ja für die Bank ein Spiel auf die 1 ist, gibt dieser

$$p(0) - \mathbb{1}_{\{0\}} = p(0) - 1 + \mathbb{1}_1 > -p(1) + \mathbb{1}_{\{1\}}, \tag{1.5}$$

wobei der letzte Ausdruck ja der Gewinn eines Spielers ist, der auf die 1 setzt. (In der Tat ist dies natürlich in Spielkasinos stets der Fall). Wir haben nun ein mathematisches Modell für das Zufallsexperiment Münzwurf"konstruiert. Es besteht $\Omega = \{0,1\}$, $\mathfrak{F} = \mathscr{P}(\{0,1\})$ und dem Wahrscheinlichkeitsmass $\mathbb{P}$ mit der Eigenschaft:

$$\mathbb{P}(\emptyset) = 0 \tag{1.6}$$
$$\mathbb{P}(\{0\}) = p(0) \in [0,1] \tag{1.7}$$
$$\mathbb{P}(\{1\}) = 1 - p(0) \tag{1.8}$$
$$\mathbb{P}(\{0,1\}) = 1. \tag{1.9}$$

Es enthält genau einen freien Parameter, $p(0) \in [0,1]$. Welchen Wert wir diesem Parameter für eine konkrete Münze zuweisen, könnten wir z.B. durch die Beobachtung einer Folge von Würfen dieser Münze entscheiden. Diese Anpassung verbleibender Parameter ist die Aufgabe der *Statisitik*. Damit werden wir uns in der zweiten Hälfte der Vorlesung beschäftigen.

Wahrscheinlickeiten auf endlichen Mengen.

Wir wollen diese Konstruktion nun auf den Fall wo die Menge Ω der Spielausgänge eine beliebige endliche Menge ist.

Zunächst wird der Begriff der *möglichen Wetten* zum Begriff der σ-Algebra erweitert.

Definition 1.1. Sei Ω eine Menge und sei $\mathfrak{F}$ eine Menge von Teilmengen ("Mengensystem"). von Ω. Man stattet $\mathfrak{F}$ mit den Operationen $\cup$ ("Vereinigung") und definiert als *Komplement*, A^c, die kleinste Teilmenge von Ω, so dass $A \cup A^c = \Omega$. Falls $\mathfrak{F}$ die leere Menge $\emptyset$ enthält, und mit $A, B \in \mathfrak{F}$ auch $A \cup B \in \mathfrak{F}$ und $A^c \in \mathfrak{F}$, so heisst $\mathfrak{F}$ eine (Mengen)-*Algebra*.

Aus Vereinigung und Komplementbildung kann man auch den Durchschnitt von Mengen konstruieren als $A \cap B = (A^c \cup B^c)^c$. Somit ist eine Mengenalgebra auch unter dem Durchschnitt abgeschlossen. Klarerweise entspricht $\cup$ der Addition und $\cap$ der Multiplikation. Dabei gilt das Assoziativgesetz für die Vereinigung: $(A \cup B) \cup C = A \cup (B \cup C)$ und für die Druchschnittsbildung, $A \cap (B \cap C) = (A \cap B) \cap C$. Weiterhin gilt das Distributivgesetz, $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$. Die Menge $\emptyset$ ist das neutrale Element der Addition und Ω das neutrale Element der Multiplikation. Allerdings gibt es kein inverses Element bezüglich der Vereinigung, weswegen einen Mengenalgebra bezüglich der Addition nur eine Halbgruppe und

keine Gruppe ist. Darüber hinaus kommutieren alle Operationen, und es gilt, dass $\emptyset \cap A = \emptyset$ für alle A. Damit is eine Mengenalgebra auch ein kommutativer Halbring.

Anmerkung. Im Sinne der Aussagenlogik entsprechen die Mengenoperationen der **Negation**, dem logischen **oder** und dem logischen **und**. Oft werden in der Wahrscheinlichkeitstheorie die Mengen A mit der Aussage "ein Zufallsexperiment hat einen Ausgang in der Menge A" identifiziert, und die Mengenoperationen daher mit den logischen Operationen bezeichnet.

In der Folge bezeichnen wir die Elemente von Ω gerne mit ω. Wir verstehen unter ω den Ausgang einen Zufallselement und schreiben $\{\omega \in A\}$ für "der Ausgang des Zufallsexperiments liegt in A. Im folgenden Benutzen wir auch gerne die Indikatorfunktion eines Ereignisses, $\mathbb{1}_A$. Wir schreiben ohne Unterscheidung

$$\mathbb{1}_{\omega \in A} = \mathbb{1}_A(\omega) = \begin{cases} 1, \text{wenn } \omega \in A, \\ 0, \text{ wenn } o \notin A. \end{cases} \tag{1.10}$$

$\mathbb{1}_A(\omega)$ ist dann gerade die Auszahlung der Einheitswette auf das Ereignis $\{\omega \in A\}$.

Mengenalgebren scheinen der richtige Spielplatz für die Wahrscheinlichkeitstheorie. Für den Fall endlicher Mengen Ω ist das auch so.

Definition 1.2. [Endliche Messräume] Sei Ω eine endliche Menge, und sei $\mathfrak{F}$ eine Algebra von Teilmengen (ein *"Mengensystem"*) von Ω. Dann heißt das Paar $(\Omega, \mathfrak{F})$ ein (endlicher) *Messraum.*

Die im Kontext endlicher Messräume häufig gewählte Algebra ist die Potenzmenge $\mathscr{P}(\Omega)$. Grundsätzlich kann man sich aber auf kleinere Algebren beschränken. Dies wird im Fall von überabzählbaren Mengen Ω sogar notwendig.

Definition 1.3 (Wahrscheinlichkeitsmaß). Sei $(\Omega, \mathfrak{F})$ ein endlicher Messraum, und sei $\mathbb{P} : \mathfrak{F} \to \mathbb{R}_+$ eine Abbildung von $\mathfrak{F}$ in die positiven reellen Zahlen, mit folgenden Eigenschaften:

(i) $\mathbb{P}(\Omega) = 1$.
(ii) Falls die Mengen $A, B \in \mathfrak{F}$ disjunkt sind, dann gilt

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B). \tag{1.11}$$

Dann heißt $\mathbb{P}$ ein *Wahrscheinlichkeitsmaß* auf dem Messraum $(\Omega, \mathfrak{F})$, und das Tripel $(\Omega, \mathfrak{F}, \mathbb{P})$ wird ein *Wahrscheinlichkeitsraum* genannt.

Im folgenden Lemma sammeln wir einige einfache Folgerungen aus der Definition.

Lemma 1.4. *Für eine Wahrscheinlichkeitsmaß auf einem endlichen Messraum gilt:*

(o)Für jedes $A \in \mathfrak{F}$ gilt $\mathbb{P}(A) \leq 1$.
(i) $\mathbb{P}(\emptyset) = 0$.

(ii)Für jedes $A \in \mathfrak{F}$ gilt $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
(iii)Für jede $A, B \in \mathfrak{F}$ gilt $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Beweis. Wegen der Eigenschaften (i) und (ii) aus der Definition 1.3 muss für jedes $A \in \mathfrak{F}$ gelten, dass

$$\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(A \cup A^c) = \mathbb{P}(\Omega) = 1. \tag{1.12}$$

Daraus folgt sofort (o) und (ii). Weiter gilt $\mathbb{P}(A) = \mathbb{P}(A \cup \emptyset) = \mathbb{P}(A) + \mathbb{P}(\emptyset)$, was (i) zur Folge hat. Schliesslich ist $A \cup B = A \cup (B \backslash (A \cap B))$, wobei wenn $C \subset B$ die Menge $B \backslash C = C^c \cap B$ gerade die kleinste Teilmenge von B ist so dass ihre Vereinigung mit C wieder B ergibt. Da A und $B \backslash (A \cap B)$ disjunkt sind, gilt

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \backslash (A \cap B)). \tag{1.13}$$

Andererseits sind $B \backslash (A \cap B)$ und $A \cap B$ disjunkt, und es gilt, dass $B = (B \backslash (A \cap B)) \cup (A \cap B)$. Daher gilt

$$\mathbb{P}(B) = \mathbb{P}(B \backslash (A \cap B)) + \mathbb{P}(A \cap B). \tag{1.14}$$

Wenn wir diese Gleichung in (1.13) einsetzen, so erhalten wir (iii). □

Als nächstes stellen wir fest, dass das so definierte Wahrscheinlichkeitsmass tatsächlich eine faire Bewertung von Einheitswetten auf alle Ereignisse in $\mathfrak{F}$ ist und dass die Bedingungen (i) und (ii) auch notwendig sind.

Wir definieren zunächst formal was wir unter fairen Preisen verstehen.

Definition 1.5. Sei $(\Omega, \mathfrak{F})$ ein Messraum. Wir sagen, dass eine Abbildung $P : \mathfrak{F} \to \mathbb{R}_+$ faire Preise für Einheitswetten sind, wenn folgendes gilt:

(i) Es gibt für den Spieler keine Möglichkeit durch kombinierte Wetten einen sicheren, vom Ausgang des Spiels unabhängigen Gewinn zu erzielen.
(ii)Der Gewinn (bzw. Verlust) eines Spieler, der auf ein Ereignis A wettet, ist unabhängig vom Ausgang des Spiels identisch zum Gewinn (Verlust) der Bank, wenn der Spieler auf A^c wettet.

Satz 1.6. *Sein $\mathbb{P}$ ein Wahrscheinlichkeitsmass auf dem endlichen Messraum $(\Omega, \mathfrak{F})$. Dann und nur dann sind die $\mathbb{P}(A)$, $A \in \mathfrak{F}$, faire Preise für Einheitswetten auf A.*

Beweis. Zunächst ist klar, dass (i) gelten muss. Wenn $\mathbb{P}(\Omega) < 1$, so liefert die Wette auf Ω einen sichern Gewinn. Ein Preis $\mathbb{P}(A) > 1$ für irgendein $A \in \mathfrak{F}$ ist offenbar sinnlos, da dies sicher zu einem Verlust führt. Also ist auch $\mathbb{P}(\Omega) \leq 1$, mithin gleich eins. Als nächstes sehen wir, das falls $\mathbb{P}(A) + \mathbb{P}(A^c) < 1$ eine kombinierte Wette auf A und auf A^c zu einem sicheren Gewinn führt. Umgekehrt ist im Fall $\mathbb{P}(A) + \mathbb{P}(A^c) > 1$ das Spiel unfair ist, mit demselben Argument wie im Münzwurf. Also muss gelten $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$.

Falls für disjunkte A, B, $\mathbb{P}(A) + \mathbb{P}(B) < \mathbb{P}(A \cup B)$ so gibt es nun die Strategie auf A, B und auch $(A \cup B)^c$ zu wetten. Da nach der vorherigen Überlegung $\mathbb{P}((A \cup B)^c) = 1 - \mathbb{P}(A \cup B)$, so erhalten wir in dieser Spielkombination folgenden, vom Spielausgang unabhängigen Kapitalfluss:

$$
\begin{aligned}
&-\mathbb{P}(A)-\mathbb{P}(B)-1+\mathbb{P}(A\cup B)+\mathbb{1}_A(\omega)+\mathbb{1}_B(\omega)+\mathbb{1}_{(A\cup B)^c}(\omega)\\
&=-\mathbb{P}(A)-\mathbb{P}(B)-1+\mathbb{P}(A\cup B)+1>0,
\end{aligned}
\tag{1.15}
$$

d.h. einen sicheren Gewinn. Im umgekehrten Fall, also wenn $\mathbb{P}(A)+\mathbb{P}(B)>\mathbb{P}(A\cup B)$, kann der Spieler auf $A\cup B$ sowie auf B^c und A^c wetten. Der Gewinn nach Abzug der Kosten ist dann

$$
\begin{aligned}
&-\mathbb{P}(A^c)-\mathbb{P}(B^c)-\mathbb{P}(A\cup B)+\mathbb{1}_{A^c}(\omega)+\mathbb{1}_{B^c}(\omega)+\mathbb{1}_{(A\cup B)}(\omega)\\
&=-2+\mathbb{P}(A)+\mathbb{P}(B)-\mathbb{P}(A\cup B)+\mathbb{1}_{A^c}(\omega)+\mathbb{1}_{B^c}(\omega)+\mathbb{1}_{(A\cup B)}(\omega)\\
&>\mathbb{P}(A\cup B)+\mathbb{1}_{A^c}(\omega)+\mathbb{1}_{B^c}(\omega)+\mathbb{1}_{(A\cup B)}(\omega)>0,
\end{aligned}
\tag{1.16}
$$

unabhängig vom Spielausgang. Von den drei Wetten auf A^c, B^c und $A\cup B$ werden nämlich stets zwei gewonnen: Fällt der Ausgang in A, so ist er nicht in B aber in $A\cup B$, so dass zwei Spiele gewonnen sind, ebenso falls der Ausgang in B liegt. Ist er weder in A noch in B, so gewinnen die Spiele auf A^c und B^c. Damit ist also auch hier ein sicherer Gewinn möglich und es muss gelten, dass $\mathbb{P}(A)+\mathbb{P}(B)=\mathbb{P}(A\cup B)$. Beide Bedingungen (i) und (ii) sind also notwendig.

Die Fairness der Preise ist schon wegen $\mathbb{P}(A)=1-\mathbb{P}(A^c)$ gezeigt. Um zu sehen, dass nun auch tatsachlich keine sicheren Gewinne moeglich sind, beobachten wir, das wir eine sichere Auszahlung von eine Euro dadurch erreichen können, dass wir eine Partition von Ω durch disjunkte Mengen $A_1,\ldots,A_k$ aus $\mathfrak{F}$ wählen und auf alle diese Ereignisse wetten. Dies kostet aber gerade $\sum_{i=1}^k \mathbb{P}(A_i)=1$, so dass der sichere Gewinn gerade Null ist. □

Terminologie. Man verwendet gemeinhin die Bezeichnungen *Wahrscheinlichkeitsmaß, Wahrscheinlichkeitsverteilung* oder auch einfach *Verteilung* synonym. Die ebenfalls synonyme Bezeichnung *Wahrscheinlichkeitsgesetz* ist im Deutschen eher veraltet, wird aber sowohl im Englischen *"probability law", "law"*, wie auch im Französischen *"loi de probabilités", "loi"*, noch gängig gebraucht.

Die Beschreibung eines Wahrscheinlichkeitsmaßes als Abbildung von $\mathfrak{F}$ nach $[0,1]$ mag sehr kompliziert sein, insbesondere da $\mathfrak{F}$ sehr gross sein kann. Oft wird $\mathfrak{F}$ die Potenzmenge von Ω sein, und diese hat die Kardinalität $2^{|\Omega|}$. Der folgende Satz sagt uns aber, dass viel weniger Parameter ausreichen um $\mathbb{P}$ vollständig zu charakterisieren.

Satz 1.7. *Sei Ω endlich und $\mathfrak{F}=\mathscr{P}(\Omega)$. Sei $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf $(\Omega,\mathfrak{F})$. Dann ist $\mathbb{P}$ vollständig durch die Angabe der Werte $\mathbb{P}(\{\omega\})=p(\omega),\omega\in\Omega$ festgelegt und es gilt $\sum_{\omega\in\Omega}p(\omega)=1$.*

Sei umgekehrt eine Menge von Parametern $p(\omega),\omega\in\Omega$ gegeben mit der Eigenschaft, dass für alle $\omega\in\Omega$, $p(\omega)\in[0,1]$ und $\sum_{\omega\in\Omega}p(\omega)=1$, dann existiert genau ein Wahrscheinlichkeitsmaß, $\mathbb{P}$ mit der Eigenschaft, dass für alle $\omega\in\Omega$, $\mathbb{P}(\{\omega\})=p(\omega)$. Es gilt dann, dass für jedes $A\in\mathfrak{F}$,

$$
\mathbb{P}(A)=\sum_{\omega\in A}p(\omega).
\tag{1.17}
$$

Beweis. Zunächst sind alle Elemente von Ω auch Elemente von $\mathfrak{F}$, so dass $\mathbb{P}(\{\omega\})$ definiert ist. Wegen der Additivitätseigenschaft (ii) in der Definition 1.3 muss dann für jedes $A \in \mathfrak{F}$ gelten, dass

$$\mathbb{P}(A) = \mathbb{P}\left(\cup_{\omega\in A}\{\omega\}\right) = \sum_{\omega\in A} \mathbb{P}(\{\omega\}) = \sum_{\omega\in A} p(\omega). \tag{1.18}$$

Weiter muss wegen (i) gelten, dass

$$1 = \mathbb{P}(\Omega) = \sum_{\omega\in\Omega} p(\omega). \tag{1.19}$$

Damit ist der erste Teil des Satzes gezeigt. Umgekehrt können wir für eine gegebene Kollektion $p(\omega), \omega \in \Omega$ mit den beschriebenen Eigenschaften definieren

$$\mathbb{P}(A) \equiv \sum_{\omega\in A} p(\omega). \tag{1.20}$$

Man sieht sofort, dass dies ein Wahrscheinlichkeitsmaß wie beschrieben liefert und wegen des ersten Teils des Satzes, ist dies auch eindeutig. □

Anmerkung. Wir sehen, dass Wahrscheinlichkeitsmaße auf endlichen Messräumen in einer eins-zu-eins Beziehung zu Funktionen $p : \Omega \to [0,1]$ mit der Eigenschaft $\sum_{\omega\in\Omega} p(\omega) = 1$ stehen. Wir nennen eine solche Funktion auf einer endlichen Menge manchmal auch einen *Wahrscheinlichkeitvektor*. Wir hätten nun Wahrscheinlichkeitsmaße über solche Funktionen *definieren* können, und vielfach wird dies in elementaren Darstellungen auch getan. Dies ist aber *konzeptuell* falsch und lässt sich nicht auf allgemeinere Grundräume übertragen. Man sollte sich daher frühzeitig daran gewöhnen, Wahrscheinlichkeitsmaße als Abbildungen von der Mengenalgebra nach $[0,1]$ zu verstehen.

Wir haben bisher das Konzept eines Wahrscheinlichkeitsmaßes mit einem Wettangebot identifiziert. Im Prinzip besteht damit noch überhaupt kein Zusammenhang zwischen einem solchen Maß und dem betrachteten Zufallsexperiment. Vielmehr ist es als eine subjektive Bewertung der Ereignisse durch die Spielbank zu betrachten. In den vorhergehenden Abschnitten haben wir nur gesehen, welche Restriktionen solche Bewertungen erfüllen müssen um überhaupt akzeptabel zu sein, ganz unabhängig vom den Eigenschaften des Zufallsexperiments.

Im Folgenden wollen wir einige spezielle Wahrscheinlichkeitsmaße betrachten.

1.2 Die Gleichverteilung.

Für eine endliche Menge Ω gibt es eine privilegierte Wahrscheinlichkeitsverteilung, die Gleichverteilung, wo jedes Element, i, von Ω dieselbe Wahrscheinlichkeit, $\mathbb{P}(i) = 1/|\Omega|$, zugeordnet bekommt. Im Roulette oder beim Würfeln entspricht es der anscheinenden Symmetrie des physikalischen Experiments, dass dem Spiel zu-

grunde liegt, dass jeder elementare Spielausgang gleich wahrscheinlich erscheint, und es a priori keinen Grund gibt, etwa die Zahl 2 anders zu bewerten als die 36. Im allgemeinen Sprachgebrauch werden die Begriffe "zufällig" und "gleichverteilt" oft synonym gebraucht.

Tatsächlich ist die Gleichverteilung die priviligierte Verteilung, die vom sogenannten "Baysianischen" Standpunkt zu verwenden ist, wenn wir keinerlei Information über den Ausgang eines Zufallsexperiments vorliegen haben. Im Fall des Roulettespiels gehen wir ja auch davon aus, dass das Gerät so konstruiert ist, dass die faire Bewertung gerade der Gleichverteilung auf $\{0,\dots,36\}$ entspricht.

In der *kombinatorischen Wahrscheinlichkeitstheorie* geht es dann darum, auf der Basis einer solchen angenommenen Gleichverteilung, Wahrscheinlichkeiten komplizierterer Mengen auszurechnen; also etwa die Wahrscheinlichkeit zu berechnen, dass, wenn k Münzen mit gleichverteiltem Ausgang 0 oder 1 geworfen werden, die Summe der Ergebnisse gerade m ist. Klarerweise ist ja in diesem Fall für jede Menge A, $\mathbb{P}(A) = |A|/|\Omega|$, und alles was wir tun müssen ist die Grösse uns interessierender Mengen zu berechnen. Dies kann allerdings schwierig genug sein.

Beispiel 1. Betrachte $n \in \mathbb{N}$ faire Münzen die geworfen werden. Die Menge Ω ist hier $\{0,1\}^n$. $\mathfrak{F}$ ist die Potenzmenge. Die Elemente $\omega \in \Omega$ schreiben wir als Vektoren $\omega = (\omega_1,\dots,\omega_n)$, mit $\omega_i \in \{0,1\}$. Jedes Ergebniss habe die gleiche Wahrscheinlichkeit 2^{-n}. Betrachte das Ereignis $\{k$Münzen zeigen Kopf$\}$. Dies entspricht eine Teilmenge von Ω die beschrieben ist durch $\{\omega \in \Omega : \sum_{i=1}^n \omega_i = k\}$. Dann ist

$$\mathbb{P}\left(\{\omega \in \Omega : \sum_{i=1}^n \omega_i = k\}\right) = \sum_{\omega\in\Omega:\sum_{i=1}^n \omega_i = k} 2^{-n} = 2^{-k}\binom{n}{k}. \tag{1.21}$$

Dabei habe wir das fundamentale kombinatorische Resultat benutzt, dass es gerade $\binom{n}{k}$ Möglichkeiten gibt aus einer Menge von n Objekten, $(\omega_1,\dots,\omega_n)$, k auszuwählen (und diesen eine 1 und den anderen die 0 zuzuordnen).

Beispiel 2. Hier ist eine klassisches Problem das Samuel Pepys im Jahr 1693 Isaac Newton gestellt hat:

Welches der folgenden drei Ereignisse hat die grösste Wahrscheinlichkeit?

(a) Beim Werfen von 6 Würfeln mindestens eine 6 zu erhalten;
(b) Beim Werfen von 12 Würfeln mindestens zwei 6en zu erhalten;
(c) Beim Werfen von 18 Würfeln mindestens drei 6en zu erhalten;

Man versteht hier, dass derjenige, der solche (komplizierten) kombinatorischen Probleme lösen kann, einen naiven Spieler übervorteilen kann.

1.3 Empirische Verteilung

Falls es keinen Grund gibt zu glauben, dass die Gleichverteilung oder ein anderes Wahrscheinlichkeitsmaß das richtige sind, können wir gegebenenfalls auf frühere Beobachtungen desselben Zufallsexperiments zurückgreifen. Dies ist sicher im Fall

von wiederholbaren Glückspielen möglich, in vielen anderen Situationen schwierig oder fraglich.

Wir betrachten also einen endlichen Messraum $(\Omega, \mathfrak{F})$. Es sei $\mathfrak{F} = \mathscr{P}(\Omega)$. Wir nehmen an, dass wir $n \in \mathbb{N}$ Realisierungen des betrachteten Zufallsexperiments beobachtet haben. Das bedeutet unserer Sprache, dass wir n Elemente, $\omega_1, \ldots, \omega_n$ aus Ω vorliegen haben.

Wir definieren nun das sogenannte Dirac-Maß, δ_ω mit der Eigenschaft, dass für alle $A \in \mathfrak{F}$,

$$\delta_\omega(A) = \begin{cases} 1, & \text{falls } \omega \in A, \\ 0, & \text{falls } \omega \notin A. \end{cases} \tag{1.22}$$

Es ist sehr einfach nachzuprüfen, dass δ_ω ein Wahrscheinlichkeitsmaß ist.

Mit dieser Definition können wir nun aus der Folge der Beobachtungen $\omega_1, \ldots, \omega_n$ das sogenannte *empirische Maß*

$$P_{\omega_1,\ldots,\omega_n} \equiv \frac{1}{n} \sum_{k=1}^{n} \delta_{\omega_k} \tag{1.23}$$

gewinnen.

Lemma 1.8. *Für jede Folge $\omega_1, \ldots, \omega_n$ mit $\omega_k \in \Omega$ ist $P_{\omega_1,\ldots,\omega_n}$ ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathscr{P}(\Omega))$.*

Der Beweis ist sehr einfach und wird als Übung gestellt. Wir benutzen die Gelegenheit um ein etwas allgemeineres Resultat zu erwähnen, dass ebenfalls sehr leicht zu beweisen ist.

Satz 1.9. *Sei $(\Omega, \mathscr{P}(\Omega))$ ein endlicher Messraum. Falls $\mathbb{P}_0, \mathbb{P}_1$ Wahrscheinlichkeitsmaße auf $(\Omega, \mathscr{P}(\Omega))$ sind, dann sind alle konvexen Linearkombinationen*

$$\mathbb{P}_\alpha \equiv \alpha \mathbb{P}_0 + (1-\alpha)\mathbb{P}_1, \quad \alpha \in [0,1], \tag{1.24}$$

ebenfalls Wahrscheinlichkeitsmaße auf $(\Omega, \mathscr{P}(\Omega))$. Die Menge aller Wahrscheinlichkeitsmaße auf $(\Omega, \mathscr{P}(\Omega))$ is ein $|\Omega|$-dimensionaler Simplex und die Dirac-Maße δ_ω, $\omega \in \Omega$ sind die Extremalpunkte dieses Simplex.

Dieses Satz ist nur eine Umformulierung von Satz 1.7.

Für $A \in \mathfrak{F}$ ist $P_{\omega_1,\ldots,\omega_n}(A) = \frac{1}{n}\sum_{k=1}^{n} \#\{1 \le k \le n : \omega_k \in A\}$, also die Häufigkeit mit der die beobachteten Spielausgänge in A lagen. Wenn diese Häufigkeiten mit zunehmender Zahl der Beobachtungen konvergieren, liefert der Grenzwert das aus frequentistischer Sicht richtige"Wahrscheinlichkeitsmaß.

Die Idee ist hier natürlich, dass man eine grosse Anzahl, sagen wir n, Experimente durchführt und sich so eine gute Approximation, genannt einen *Schätzer* für die tatsächlichen Wahrscheinlichkeiten verschafft.

1.4 Erzeugte Algebren

Üblicherweise geht man bei der Beschreibung einer Mengenalgebra so vor, dass man eine gewisse Menge von Teilmengen, die man in der Algebra haben möchte vorgibt, und diese dann zu einer Algebra ergänzt, indem man alle gemäß der Definition nötigen Mengen dazufügt.

Definition 1.10. Sei $\mathscr{E}$ eine Menge von Teilmengen von Ω. Die kleinste Algebra, die $\mathscr{E}$ enthält, heisst die von $\mathscr{E}$ *erzeugte* Algebra. Wir bezeichnen diese oft mit $\sigma(\mathscr{E})$. Für eine gegebene Algebra, $\mathfrak{F}$, heisst eine Menge von Mengen, $\mathscr{E}$, *Erzeuger* (oder *Generator*) von $\mathfrak{F}$, wenn $\sigma(\mathscr{E}) = \mathfrak{F}$.

Die folgenden Beobachtungen verallgemeinern den Satz 1.7 für den Fall, dass $\mathfrak{F}$ nicht die Potenzmange von Ω ist.

Lemma 1.11. *Sei $(\Omega, \mathfrak{F})$ ein endlicher Messraum. Dann enthält $\mathfrak{F}$ eine eindeutige minimale Partition, $\Pi = (\pi_1, \ldots, \pi_n)$, von Ω mit folgenden Eigenschaften:*

(i) $\bigcup_{i=1}^{n} \pi_i = \Omega$*;*
(ii) Für alle $B \in \mathfrak{F}$ *und alle* $k = 1, \ldots, n$*, gilt* $B \cap \pi_k \in \{\emptyset, \pi_k\}$*. Insbesondere gilt für alle* $i \neq j$*, dass* $\pi_i \cap \pi_j = \emptyset$*.*

Beweis. (Erst mal als Übung!) □

Proposition 1.12. *Sei Ω eine endliche Menge und $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Sei $\Pi = (\pi_1, \ldots, \pi_n)$ die Partition aus Lemma 1.11. Dann ist das Maß $\mathbb{P}$ eindeutig durch die Werte $p(i) = \mathbb{P}(\pi_i)$, $i = 1, \ldots, n$, festgelegt. Umgekehrt gibt es für jede Sammlung von Werten $p(i) \geq 0$, $i = 1, \ldots, n$, mit $\sum_{i=1}^{n} p(i) = 1$ genau ein Wahrscheinlichkeitsmaß auf $(\Omega, \mathfrak{F})$, so dass $P(\pi_i) = p(i)$.*

Beweis. Übung! □

Kapitel 2
Zufallsvariablen und Erwartungswerte

> *La probabilité des événements sert à déterminer l'espérance ou la crainte des personnes intéresées à leur existence. Le mot* espérance *a diverses acceptions: il exprime généralement l'avantage de celui qui attend un bien quelquonque, dans des suppositions qui ne sont que probables. Cet avantage, dans la théorie des hasards, est le produit de la somme espérée par la probabilité de l'obtenir : c'est la somme partielle qui doit revenir lorsqu'on ne veut pas courir les risques de l'événement, en supposant que la repartition se fasse proportionellement aux probabilités. Cette repartition est la seule équitable, lorsqu'on fait abstraction de toutes circonstances étrangères, parce qu'un égal degré de probabilité donne un droit égal sur la somme espérée. Nous nommerons cet avantage* espérance mathématique[a] Pierre Simon de Laplace, Théorie Analytique des Probabilités

[a] Die Wahrscheinlichkeit von Ereignissen dient zur Bestimmung der Erwartung oder der Furcht von Personen, die an ihrer Existenz interessiert sind. Das Wort. *Erwartung* hat verschiedene Bedeutungen: es drückt im allgemeinen den Vorteil desjenigen aus, der irgendeinen Vorteil erwartet, und zwar unter Annahmen, die nur wahrscheinlich sind. Dieser Vorteil ist in der Theorie der Zufälle das Produkt der erwarteten Summe und der Wahrscheinlichkeit sie zu erhalten: es ist die Teilsumme die man erhalten muss, wenn man das Risiko des Ereignisses nicht eingehen will, unter der Annahme, dass die Verteilung proportional zu den Wahrscheinlichkeiten erfolgt. Diese Verteilung ist die einzig gerechte, sofern man von allen fremden Umständen abstrahiert, da ein gleicher Grad von Wahrscheinlichkeit einen gleichen Anspruch an die erwartete Summe gibt. Wir nennen dieses Vorteil die *mathematische Erwartung*.

Wir haben im ersten Kapitel gesehen, dass unter einer vernünftig erscheinenden Definition des Wahrscheinlichkeitsbegriffes, in natürlicher Weise der Begriff eines Wahrscheinlichkeitsmaßes in der Form der Definition 1.3 auftaucht. Diese nunmehr *axiomatisch* definierten Objekte können nun mathematisch untersucht werden.

2.1 Messbare Funktionen

Wir wollen uns nun für Funktionen vom Ω in die reellen Zahlen interessieren. Schon im vorigen Abschnitt haben wir solche Funktionen betrachtet, nämlich die Indikatorfunktionen von Ereignissen $A \in \mathfrak{F}$, $\mathbb{1}_A$. Wir hatten gesagt dass diese gerade die Auszahlung einer Einheitswette auf das Ereignis $\omega \in A$ darstellt. Andererseits hat-

ten wir gesagt, dass $\mathbb{P}(A)$ gerade der Wert dieser Wette ist. Wir bezeichnen den Wert dieser Wette nun als *Erwartungswert* der Funktion $\mathbb{1}_A$ und schreiben

$$\mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A). \tag{2.1}$$

Man beachte, dass diese Definition nur dann Sinn macht, wenn $A \in \mathfrak{F}$. Wir können nun den Begriff des Erwartungswerts in natürlicher Weise auf Funktionen übertragen, die Linearkombinationen von Indikatorfunktionen von Mengen in $\mathfrak{F}$ sind.

Definition 2.1. Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein endlicher Wahrscheinlichkeitsraum. Sein $X : \Omega \to \mathbb{R}$ gegeben durch $X = \sum_{k=1}^{n} a_k \mathbb{1}_{A_k}$, wo alle $A_k \in \mathfrak{F}$. Dann definieren wir

$$\mathbb{E}[X] \equiv \sum_{k=1}^{n} a_k \mathbb{P}(A_k). \tag{2.2}$$

Wir definieren nun den Begriff der *messbaren Funktion*.

Definition 2.2. Sei $(\Omega, \mathfrak{F})$ ein endlicher Messraum und $X : \Omega \to \mathbb{R}$. Dann heisst X *messbar* (bezüglich $\mathfrak{F}$) falls fuer jedes $a \in \mathbb{R}$, die Mengen

$$X^{-1}(a) \equiv \{\omega \in \Omega : X(\omega) = a\} \in \mathfrak{F}. \tag{2.3}$$

Definition 2.3. Eine messbare Funktion heisst *Zufallsvariable*.

Beachte, dass $X^{-1}(a)$ die leere Menge ist, falls X den Wert a nicht annimmt. Da Ω endlich ist, nimmt X nur endlich viele Werte an.

Die folgende Beobachtung ist trivial.

Lemma 2.4. *Sei $(\Omega, \mathfrak{F})$ ein endlicher Messraum und $X : \Omega \to \mathbb{R}$ messbar. Dann ist X von der Form wie in Definition 2.1 wo $a_k, k = 1, \ldots, n$ die Werte sind, die X annimmt, and $A_k = \{\omega \in \Omega : X(\omega) = a_k\}$. Daher ist die Erwartung von X definiert und es gilt*

$$\mathbb{E}[X] \equiv \sum_{k=1}^{n} a_k \mathbb{P}(\{X(\omega) = a_k\}). \tag{2.4}$$

Wir wollen (2.4) vorläufig als Definition des Erwartungswerts einer Zufallsvariablen ansehen. Aus dieser Definition folgt sofort:

Korollar 2.5. *Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, und seien X_1, X_2 Zufallsvariablen. Dann sind für alle $\alpha, \beta \in \mathbb{R}$ die Linearkombinationen $\alpha X_1 + \beta X_2$ Zufallsvariablen, und es gilt, dass*

$$\mathbb{E}[\alpha X_a + \beta X_2] = \alpha \mathbb{E}[X_1] + \beta \mathbb{E}[X_2]. \tag{2.5}$$

Beweis. Der Beweis ist sehr einfach und wird als Übung gestellt. □

Die Definition der Erwartung mittels der Formel (2.4) hat den formalen Nachteil, dass sie die Kenntnis der Werte, die X annimmt, voraussetzt. Dies wird bei

der Verallgemeinerung auf allgemeine Messräume hinderlich sein. Wir können aber leicht eine Formel angeben, die mit (2.4) übereinstimmt, formal aber keine implizite Information über X voraussetzt.

Lemma 2.6. *Sei $(\Omega,\mathfrak{F},\mathbb{P})$ ein Wahrscheinlichkeitsraum, und sei $X:\Omega\to\mathbb{R}$ eine messbare Funktion bezüglich $\mathfrak{F}$, die nur endlich viele Werte annimmt. Dann gilt*

$$\mathbb{E}[X]=\lim_{\varepsilon\downarrow 0}\sum_{k=-\lceil\varepsilon^{-2}\rceil}^{+\lceil\varepsilon^{-2}\rceil} k\varepsilon\,\mathbb{P}(\{\omega\in\Omega: k\varepsilon\le X(\omega)<(k+1)\varepsilon\})\,. \tag{2.6}$$

Beweis. Der Beweis ist recht einfach. Wir nehmen an, dass X gerade die n Werte $a_1,\dots,a_n$ annimmt. Dann ist $\delta=\min_{i\neq j}|a_i-a_j|>0$ und $\max_{\omega\in\Omega}|X(\omega)|\equiv M<\infty$. Dann gilt zunächst, dass, für alle $0<\varepsilon<\delta/2$, jedes Intervall $(k\varepsilon,(k+1)\varepsilon]$ höchstens einen der Werte a_i enthalten kann. Ausserdem ist für $\varepsilon<M^{-1}/2$, in keinem Intervall $[k\varepsilon,(k+1)\varepsilon)$ mit $k>\varepsilon^{-2}$ ein Punkt $X(\omega)$. Für solche ε gibt es für jedes $l=1,\dots,n$ genau ein k_l, so dass $a_l\in(k_l\varepsilon,(k_l+1)\varepsilon]$. Dann ist aber

$$\begin{aligned}\sum_{l=1}^n a_l\mathbb{P}(\{\omega\in\Omega:X(\omega)=a_l\}) &= \sum_{l=1}^k a_l\mathbb{P}(\{\omega\in\Omega:X(\omega)\in(k_l\varepsilon,(k_l+1)\varepsilon]\})\\ &\ge \sum_{l=1}^n \varepsilon k_l\mathbb{P}(\{\omega\in\Omega:X(\omega)\in(k_l\varepsilon,(k_l+1)\varepsilon]\})\\ &= \sum_{k=-\lceil\varepsilon^{-2}\rceil}^{\lceil\varepsilon^{-2}\rceil}\varepsilon k\mathbb{P}(\{\omega\in\Omega:X(\omega)\in(k\varepsilon,(k+1)\varepsilon]\})\end{aligned}$$

sowie auch

$$\begin{aligned}\sum_{l=1}^n a_l\mathbb{P}(\{\omega\in\Omega:X(\omega)=a_l\}) &\le \sum_{l=1}^n \varepsilon(k_l+1)\mathbb{P}(\{\omega\in\Omega:X(\omega)\in(k_l\varepsilon,(k_l+1)\varepsilon]\})\\ &= \sum_{k=-\lceil\varepsilon^{-2}\rceil}^{\lceil\varepsilon^{-2}\rceil}\varepsilon k\mathbb{P}(\{\omega\in\Omega:X(\omega)\in(k\varepsilon,(k+1)\varepsilon]\})\\ &\quad+\varepsilon\sum_{k=-\lceil\varepsilon^{-2}\rceil}^{\lceil\varepsilon^{-2}\rceil}\mathbb{P}(\{\omega\in\Omega:X(\omega)\in(k\varepsilon,(k+1)\varepsilon])\\ &= \sum_{k=-\lceil\varepsilon^{-2}\rceil}^{\lceil\varepsilon^{-2}\rceil}\varepsilon k\mathbb{P}(\{\omega\in\Omega:X(\omega)\in[k\varepsilon,(k+1)\varepsilon)\})+\varepsilon\end{aligned}$$

da die letzte Summe gerade dass Maß von Ω, also 1 ist. Da diese Ungleichungen für jedes $\varepsilon<\min(\delta/2,1/(2M))$ gelten, folgt, dass

$$\begin{aligned}
&\limsup_{\varepsilon\downarrow 0} \sum_{k=-\lceil\varepsilon^{-2}\rceil}^{\lceil\varepsilon^{-2}\rceil} \varepsilon k \mathbb{P}(\{\omega\in\Omega : X(\omega)\in(k\varepsilon,(k+1)\varepsilon]\}) \\
&\leq \sum_{l=1}^{k} a_l \mathbb{P}(\{\omega\in\Omega : X(\omega)=a_l\}) \qquad (2.7)\\
&\leq \liminf_{\varepsilon\downarrow 0} \sum_{k=-\lceil\varepsilon^{-2}\rceil}^{\lceil\varepsilon^{-2}\rceil} \varepsilon k \mathbb{P}(\{\omega\in\Omega : X(\omega)\in(k\varepsilon,(k+1)\varepsilon]\}).
\end{aligned}$$

Dies beweist das Lemma und die Existenz des Limes in (2.6). □

Anmerkung. Die Formel (2.6) erscheint unsinnig kompliziert. Sie hat aber die schöne Eigenschaft, dass sie sich sehr gut auf kompliziertere Situationen, wenn Ω nicht endlich ist, übertragen lässt.

Manchmal spricht man auch vom *mathematischen Erwartung* oder dem *mathematischen Mittel* von X. Dies wird getan um den Unterschied zum sogenannten *empirischen Mittel* zu betonen, der das arithmetische Mittel der Funktion X über n Wiederholungen eines Experiments darstellt,

$$E^{\mathrm{emp}}_{\omega_1,\dots,\omega_n} X \equiv n^{-1}\sum_{k=1}^{n} f(X(\omega_k)). \qquad (2.8)$$

Der Zusammenhang zwischen mathematischem und empirischen Mittel ist eine der grundlegenden Fragen der Wahrscheinlichkeitstheorie.

Der Fall $\mathfrak{F} = \mathbb{P}(\Omega)$.

Wenn wir auf einem Wahrscheinlichkeitsraum arbeiten, in dem $\mathfrak{F}$ die Potenzmenge ist, gibt es einige Vereinfachungen.

Lemma 2.7. *Sei $(\Omega,\mathfrak{F},\mathbb{P})$ ein endlicher Wahrscheinlichkeitsraum und sei $\mathfrak{F}=\mathbb{P}(\Omega)$. Dann gilt:*

(i) Jede Funktion $X:\Omega\to\mathbb{R}$ ist messbar.
(ii) Der Erwartungswert von X kann geschrieben werden als

$$\mathbb{E}[X] = \sum_{\omega\in\Omega} X(\omega)p(\omega), \qquad (2.9)$$

wo p der zu $\mathbb{P}$ gehörige Wahrscheinlichkeitsvektor ist, also $p(\omega)=\mathbb{P}(\{\omega\})$.

Beweis. Übungsaufgabe! □

2.2 Zufallsvariablen als Abbildungen von Wahrscheinlichkeitsmaßen

Sei $\Omega, \mathfrak{F})$ ein Messraum und ein X eine Zufallsvariable. Sei S das Bild von Ω unter X, d.h. die Menge der Werte, die X annimmt. Per Definition gilt für jeden Punkt $x \in S$, dass $X^{-1}(x) \in \mathfrak{F}$, mithin gilt auch, dass für jede Teilmenge $B \in \mathscr{P}(S)$, $X^{-1}(B) \in \mathfrak{F}$. Wir können nun S zu einem Messraum $(S, \mathscr{P}(S))$ erweitern und X als Abbildung von $(\Omega, \mathfrak{F})$ nach $(S, \mathscr{P}(S))$ auffassen. Aufgrund der Messbarkeit transportiert X nun auch ein Wahrscheinlichkeitsmaß, $\mathbb{P}$, auf $(\Omega, \mathfrak{F})$ in ein Wahrscheinlichkeitsmaß, P_X, auf $(S, \mathscr{P}(S))$.

Definition 2.8. Sei $(\Omega, \mathfrak{F})$ ein endlicher Messraum und X eine Zufallsvariable mit Wertebereich S. Dann ist $P_X : \mathscr{P}(S) \to [0,1]$ definiert durch

$$P_X(B) \equiv \mathbb{P}\left(X^{-1}(B)\right) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in B\}\right), \tag{2.10}$$

für alle $B \in \mathscr{P}(S)$ wohldefiniert.

Satz 2.9. *Die in Definition 2.8 definierte Abbildung P_X ist ein Wahrscheinlichkeitsmaß auf $(S, \mathscr{P}(S))$.*

Beweis. Zunächst ist $P_X(S) = \mathbb{P}\left(X^{-1}(S)\right) = \mathbb{P}(\Omega) = 1$. Weiter ist für disjunkte Mengen B_1, B_2,

$$P_X(B_1 \cup B_2) = \mathbb{P}\left(X^{-1}(B_1 \cup B_2)\right). \tag{2.11}$$

Da X eine Funktion ist, gilt $X^{-1}(B_1 \cup B_2) = X^{-1}(B_1) \cup X^{-1}(B_2)$, und die Mengen $X^{-1}(B_i)$, $i = 1,2$, sind disjunkt. Daher ist die rechte Seite von Gl. (2.11) gleich $P_X(B_1) + P_X(B_2)$ und P_X ist ein Wahrscheinlichkeitsmaß. □

Das Wahrscheinlichkeitsmaß P_X nennt man auch die *Verteilung der Zufallsvariablen X* unter $\mathbb{P}$.

Beispiel. Wir betrachten wieder den Wahrscheinlichkeitsraum $(\{0,1\}^n, \mathbb{P}(\{0,1\}^n), \mathbb{P}_p)$, wobei diesmal $\mathbb{P}_p$ jedem $\omega = (\omega_1, \ldots \omega_n) \in \{0,1\}^n$ die Wahrscheinlichkeit (für $p \in [0,1]$)

$$\mathbb{P}_p(\omega) = p^{\sum_{i=1}^n \omega_i} (1-p)^{n - \sum_{i=1}^n \omega_i} \tag{2.12}$$

zuordnet. Wir definieren nun auf $\{0,1\}^n$ die Zufallsvariable X_n durch

$$X_n(\omega) \equiv \sum_{i=1}^n \omega_i. \tag{2.13}$$

Offenbar ist der Wertebereich von X_n die Menge $S_n \equiv \{0,1,2,\ldots,n\}$. Um den Erwartungswert von X_n zu berechnen, berechnen wir zunächst $P_{X_n}(\{k\})$, für $k \in S_n$. Dies ist einfach:

$$\begin{aligned} P_{X_n}(\{k\}) &= \mathbb{P}_p\left(\left\{\omega \in \Omega : \sum_{i=1}^{n} \omega_i = k\right\}\right) \\ &= \sum_{o\in\Omega:\sum_{i=1}^{n}=k} \mathbb{P}_p(\{\omega\}) = p^k(1-p)^{n-k} \sum_{o\in\Omega:\sum_{i=1}^{n}=k} 1 \\ &= p^k(1-p)^{n-k}\binom{n}{k}. \end{aligned} \tag{2.14}$$

Dies ist die allgemeine Form der Binomialverteilung, die wir im Fall $p = 1/2$ schon kennengelernt haben. Wir können nun auch die Erwartung von X_n ausrechnen.

$$\mathbb{E}[X_n] = \sum_{k=0}^{n} kP_{X_n}(\{k\}) = \sum_{k=0}^{n} kp^k(1-p)^{n-k}\binom{n}{k} = pn. \tag{2.15}$$

Anmerkung. Die Berechnung des Erwartungswerts von X_n ist zwar korrekt, aber unnötig kompliziert (die letzte Gleichung in (2.15) ist eben nicht so ganz trivial). Es geht nämlich viel einfacher. Dazu überlegen wir und, dass ja die ω_i, die in der Definition von X_n auftauchen, selbst auch Zufallsvariablen sind und dass gilt,

$$\mathbb{P}(\{o \in \Omega : \omega_i = 1\}) = p = 1 - \mathbb{P}(\{o \in \Omega : \omega_i = 0\}). \tag{2.16}$$

Insbesondere ist für $i = 1, \ldots, n$,

$$\mathbb{E}[\omega_i] = p. \tag{2.17}$$

Jetzt können wir die Linearität des Erwartungswerts ausnutzen und erhalten

$$\mathbb{E}[X_n] = \sum_{i=1}^{n} \mathbb{E}[\omega_i] = pn. \tag{2.18}$$

Das war wesentlich einfacher! Insbesondere mussten wir nicht einmal die Verteilung der Zufallsvariablen X_n ausrechnen, um den Erwartungswert zu erhalten, was immerhin auch die Lösung einer kombinatorischen Aufgabe erforderte. Diese Beobachtung wollen wir uns merken: Es ist oft einfacher, Erwartungswerte von Zufallsvariablen auszurechnen als ihre Verteilungen!!

Mittels Zufallsvariablen können wir also neue, interessante Verteilungen gewinnen. Die Berechnung der Verteilung von komplizierten Zufallsvariablen ist dann auch eine wesentliche Aufgabe der Wahrscheinlichkeitstheorie.

2.3 Verteilungsfunktionen

Wir erinnern uns dass wir Zufallsvariablen ursprünglich als Abbildungen von Ω nach $\mathbb{R}$ eingeführt haben. Dies erlaubt es, ein neues interessantes Objekt, die sogenannte *Verteilungsfunktion*, $F : \mathbb{R} \to [0,1]$, zu definieren.

Definition 2.10. Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und sei X eine Zufallsvariable. Dann ist die Funktion $F : \Omega \to [0,1]$ gegeben durch

$$F_X(x) \equiv \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}), \quad x \in \mathbb{R}, \tag{2.19}$$

wohldefiniert und heisst Verteilungsfunktion der Zufallsvariablen X.

Die Wahrscheinlichkeiten $\mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\})$ sind alle bekannt, da $\{\omega \in \Omega : X(\omega) \leq x\} = \{\omega \in \Omega : X(\omega) \in S \cap (-\infty, x]\}$, wo S das Bild von X ist, da $X(\omega) \in S \cap (-\infty, x] \subset S$.

In unserem Fall eines endlichen Zustandsraumes ist die Verteilungsfunktion jeder Zufallsvariablen eine Stufenfunktion mit endlich vielen Sprüngen. Die Sprungstellen sind genau die Werte, a_i, die X annimmt. Die Funktion F_X springt and der Stelle a_i um den Betrag $P_X(\{a_i\})$, d.h.

$$F_X(a_i) - \lim_{x \uparrow a_i} F_X(x) = P_X(a_i). \tag{2.20}$$

insbesondere ist *F nicht-fallend* und *rechtsstetig*.

2.4 Ungleichungen

Wir hatten oben bemerkt, dass Erwartungswerte oft einfach zu berechnen sind. Wir wollen nun zeigen, dass wir aus Erwartungswerten Informationen über Verteilungsfunktionen gewinnen können.

Lemma 2.11. *Sei $(\Omega, \mathfrak{F})$ ein Messraum und Sei $X : \Omega \to \mathbb{R}_+$ eine Zufallsvariable, die nur nicht-negative Werte annimmt. Sei $F_X : \mathbb{R} \to [0,1]$ die Verteilungsfunktion von X. Dann gilt für $x > 0$,*

$$1 - F_X(x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) > x\}) \leq \frac{\mathbb{E}[X]}{x}. \tag{2.21}$$

Beweis. Die erste Gleichung in (2.21) ist offensichtlich. Als nächstes schreiben wir

$$\mathbb{P}(\{\omega \in \Omega : X(\omega) > x\}) = \mathbb{E}\left[\mathbb{1}_{\{o \in \Omega : X(\omega) > x\}}\right]. \tag{2.22}$$

Nun ist aber für alle ω für die $\mathbb{1}_{\{o \in \Omega : X(\omega) > x\}} \neq 0$, $X(\omega)/x \geq 1$. Damit können wir schreiben

$$\mathbb{E}\left[\mathbb{1}_{\{\omega \in \Omega : X(\omega) > x\}}\right] \leq \mathbb{E}\left[\mathbb{1}_{\{\omega \in \Omega : X(\omega) > x\}} \frac{X(\omega)}{x}\right] \leq \mathbb{E}\left[\frac{X(\omega)}{x}\right] = \frac{\mathbb{E}[X(\omega)]}{x}. \tag{2.23}$$

Setzen wir dies in (2.22) ein, so erhalten wir (2.21). □

Alternativ können wir den Beweis auch so führen:

Beweis (Variante). Seien $S = \{a_1, \ldots, a_n\}$ der Wertebereich von X. Dann ist

$$\begin{aligned}
\mathbb{P}(\{\omega \in \Omega : X(\omega) > x\}) &= P(\cup_{a \in S: a > x} \{\omega \in \Omega : X(\omega) = a\}) \qquad (2.24)\\
&= \sum_{a \in S: a > x} \mathbb{P}(\{\omega \in \Omega : X(\omega) = a\})\\
&\leq \sum_{a \in S: a > x} \frac{a}{x} \mathbb{P}(\{\omega \in \Omega : X(\omega) = a\})\\
&\leq \sum_{a \in S} \frac{a}{x} \mathbb{P}(\{\omega \in \Omega : X(\omega) = a\})\\
&= \frac{1}{x} \sum_{a \in S} a \mathbb{P}(\{\omega \in \Omega : X(\omega) = a\})\\
&= \frac{1}{x} \mathbb{E}[X].
\end{aligned}$$

□

Die Ungleichung (2.21) heisst Chebychev-Ungleichung und ist trotz ihrer Einfachheit ein ganz wichtiges Werkzeug der Wahrscheinlichkeitstheorie. Wir können aus ihr sofort zwei weitere Ungleichungen gewinnen:

Korollar 2.12. *Sei $(\Omega, \mathfrak{F})$ ein endlicher Messraum und Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable. Dann gilt*

$$1 - F_X(x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) > x\}) \leq \frac{\mathbb{E}[X^2]}{x^2}. \qquad (2.25)$$

Beweis. Wenn X eine Zufallsvariable ist, so ist X^2 ebenfalls eine Zufallsvariable (nachprüfen!!) und X^2 ist positiv. Ausserdem gilt, dass

$$\{\omega \in \Omega : X(\omega) > x\} \subset \{\omega \in \Omega : X^2(\omega) > x^2\}. \qquad (2.26)$$

Daraus folgt

$$\mathbb{P}(\{\omega \in \Omega : X(\omega) > x\}) \leq \mathbb{P}\left(\{\omega \in \Omega : X^2(\omega) > x^2\}\right), \qquad (2.27)$$

und indem wir nun Lemma 2.11 anwenden, erhalten wir (2.28). □

Diese Ungleichung wird gelegentlich als Markov-Ungleichung bezeichnet.
Schliesslich erhalten wir als weiteres Korollar:

Korollar 2.13. *Sei $(\Omega, \mathfrak{F})$ ein endlicher Messraum und Sei $X : \Omega \to \mathbb{R}$ eine Zufallsvariable. Dann gilt*

$$\mathbb{P}(\{\omega \in \Omega : |X(\omega) - \mathbb{E}[X]| > x\}) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{x^2}. \qquad (2.28)$$

Beweis. Völlig analog zum Beweis von Korollar 2.12. □

Die Ungleichung sagt etwas darüber aus, wie stark eine Zufallsvariable von ihrem Erwartungswert abweicht. Die Größe

$$\mathbb{E}\left[(X-\mathbb{E}[X])^2\right] = \mathbb{E}[X^2] - (\mathbb{E}(X))^2 \equiv \mathrm{var}(X) \tag{2.29}$$

nennt man die *Varianz* der Zufallsvariablen X. Je kleiner die Varianz ist, um so unwahrscheinlicher sind starke Abweichungen der Zufallsvariablen von ihrem Erwartungswert. Die Varianz ist neben dem Erwartungswert eine wesentliche, oft leicht berechenbare Kenngröße einer Zufallsvariablen.
Notation: Zur Vereinfachung der Notation wollen wir in Zukunft nicht immer "$\mathbb{P}[\{o \in \Omega : X(\omega) \in B\}]$ßchreiben. Wir benutzen daher ohne Unterscheidung folgende Schreibweisen:

$$\mathbb{P}[\{\omega \in \Omega : X(\omega) \in B\}] = \mathbb{P}[\{X(\omega) \in B\}] = \mathbb{P}(X(\omega) \in B] = \mathbb{P}[X \in B]. \tag{2.30}$$

Beispiel. Als Anwendung dieser Ungleichungen betrachten wir wieder die Binomialverteilung, also wir betrachten die Verteilung der Zufallsvariablen X_n aus Gleichung (2.13). Wir habe schon gesehen, dass wir hier den Erwartungswert sehr einfach ausrechnen können. Es folgt dann aus Lemma 2.21, dass für alle

$$\mathbb{P}(X_n > x) = 1 - F_{X_n}(x) \leq \frac{pn}{x}. \tag{2.31}$$

Das ist noch nicht besonders aufschlussreich. Nun koennen wir aber auch den Erwartungswert von X_n^2 ausrechnen:

$$\mathbb{E}\left[X_n^2\right] = \mathbb{E}\left[\left(\sum_{k=1}^{n} \omega_k\right)^2\right] = \sum_{k=1}^{n}\sum_{\ell=1}^{n} \mathbb{E}\left[\omega_k \omega_\ell\right]. \tag{2.32}$$

Die Erwartungen in der Summe sind wieder leicht zu berechnen: Wenn $k = \ell$ dann ist $\omega_k \omega_\ell = \omega_k^2 = \omega_k$, also $\mathbb{E}[\omega_k \omega_\ell] = p$. Wenn $k \neq \ell$, dann ist $\omega_k \omega_\ell$ nur dann nicht gleich null, wenn $\omega_k = \omega_\ell = 1$, und dies hat die Wahrscheinlichkeit p^2. Damit ergibt sich aus (2.32)

$$\mathbb{E}\left[X_n^2\right] = \sum_{k=1}^{n} p + \sum_{k\neq\ell=1}^{n} p^2 = n(n-1)p^2 + np = n^2p^2 + np(1-p). \tag{2.33}$$

Also ist die Varianz

$$\mathrm{var}(X_n) = \mathbb{E}[X_n^2] - (\mathbb{E}[X_n])^2 = n^2p^2 + np(1-p) - n^2p^2 = n(p(1-p). \tag{2.34}$$

Wenn wir nun Korollar 2.13 anwenden, so erhalten wir die Ungleichung

$$\mathbb{P}\left(|X_n - pn| > x\right) \leq \frac{np(1-p)}{x^2}. \tag{2.35}$$

Die rechte Seite dieser Ungleichung wird klein, sobald x viel grösser ist als $\sqrt{n}$, d.h. die Wahrscheinlichkeit, dass sich die Zufallsvariable in einer Umgebung der Ordnung $\sqrt{n}$ um Ihren Mittelwert pn befindet, ist fast eins! Wir können dies genauer so formulieren:

$$\lim_{K\uparrow\infty}\lim_{n\uparrow\infty}\mathbb{P}\left(|X_n - pn| > K\sqrt{n}\right) = 0. \tag{2.36}$$

Vielleicht noch prägnanter wird diese Aussage, wenn wir eine Zufallsvariable

$$Y_n \equiv \frac{1}{n}X_n \tag{2.37}$$

definieren. Diese hat dann Erwartungswert p, und es gilt

$$\lim_{K\uparrow\infty}\lim_{n\uparrow\infty}\mathbb{P}\left(|Y_n - p| > K/\sqrt{n}\right) = 0. \tag{2.38}$$

Mit wachsendem n konzentriert sich Y_n also immer mehr um seinen Mittelwert p.

2.5 Wahrscheinlichkeiten auf $\mathbb{R}$

Nachdem wir gesehen haben, dass eine Zufallsvariable Wahrscheinlichkeitsmaße von $(\Omega,\mathfrak{F})$ nach $(S,\mathscr{P}(S))$ transportiert, liegt es nahe sich vorzustellen, dass wir damit auch ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$ konstruieren können. Tatsachlich können wir ja jedem Intervall $I \subset \mathbb{R}$ einen Wert $\mathbb{P}(\{\omega \in \Omega : X(\omega) \in I\}) \equiv P_X(I)$ zuordnen und als Wahrscheinlichkeit des Intervalls unter P_X interpretieren. Klarerweise erfüllt dies die Anforderungen $P_X(\mathbb{R}) = 1$ und $P_X(I \cup J) = P_X(I) + P_X(J)$, die wir an ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$ sicher stellen sollten. Allerdings bleiben noch einige Fragen offen. Zum einen müssten wir klären, auf welcher Mengenalgebra P_X eigentlich definiert sein sollte. Da $\mathbb{R}$ unendlich und sogar überabzählbar ist, führt dies auf recht komplizierte Fragen, die erst in der *Maßtheorie* behandelt werden. Wir können diese Probleme aber zunächst einmal ausklammern und pragmatisch ein Wahrscheinlichkeitsmaß auf $\mathbb{R}$ durch seine Werte auf Intervallen (und allem was wir daraus basteln können) beschrieben sehen. Ein solches Wahrscheinlichkeitsmass ist dann eindeutig durch die Angabe einer Verteilungsfunktion beschrieben.

Wir wollen an dieser Stelle die Definition von Mess- und Wahrscheinlichkeitsräumen auf den Fall einer beliebigen Menge Ω verallgemeinern.

Definition 2.14. Sei Ω eine Menge.

(a) Eine Menge von Teilmengen von Ω, $\mathfrak{F}$, heisst eine Sigma-Algebra, genau dann wenn gilt:

(i) $\emptyset \in \mathfrak{F}$;
(ii) Falls $A \in \mathfrak{F}$, dann ist auch $A^c \in \mathfrak{F}$;
(iii) Sei $(A_n)_{n\in\mathbb{N}}$ eine Folge von Elementen $A_n \in \mathfrak{F}$. Dann gilt dass $\cup_{n\in\mathbb{N}} A_n \in \mathfrak{F}$.

Ein Paar $(\Omega, \mathfrak{F})$ heisst Messraum.

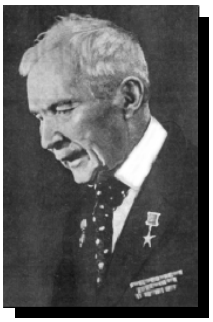
(b) Sei $(\Omega, \mathfrak{F})$ ein Messraum. Eine Abbildung $\mathbb{P} : \mathfrak{F} \to \mathbb{R}_+$ heisst ein Wahrscheinlichkeitsmaß, genau dann wenn folgendes gilt:

(i) $\mathbb{P}(\Omega) = 1$:

(ii) Falls $(A_n)_{n\in\mathbb{N}}$ eine Folge von paarweise disjunkten Elementen $A_n \in \mathfrak{F}$ (d.h., für alle $i, j \in \mathbb{N}$, gilt $A_i \cap A_j = \emptyset$). Dann gilt

$$\mathbb{P}(\cup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n). \tag{2.39}$$

Ein Tripel $(\Omega, \mathfrak{F}, \mathbb{P})$ heisst ein Wahrscheinlichkeitsraum.



Wir sehen, dass wir im Unterschied zum Fall wo Ω endlich ist zusätzliche Forderungen an abzählbar unendliche Vereinigungen von Mengen in $\mathfrak{F}$ stellen. Dies hat hat im wesentlichen mathematische Gründe und erlaubt gewissen Grenzwertbetrachtungen, zu denen wir aber in dieser Vorlesung nicht kommen werden. Im Fall nicht-abzählbarer Mengen Ω führt dies in der Regel zu recht komplizierten Sigma-Algebren, die man nicht ëxplizitängeben kann. Die damit verbundenen Probleme werden wir in dieser Vorlesung aber ignorieren. Die In der obigen Definition gestellten Forderungen an einen Wahrscheinlichkeitsraum gehen auf de russischen Mathematiker Andrey Nikolaevich Kolmogorov zurück und werden *Kolmogorov'sche Axiome* genannt.

Für den Fall, $\Omega = \mathbb{R}$ definieren wir die *Borel'sche* Sigma-Algebra, $\mathfrak{B}(\mathbb{R})$, als die kleinste Sigma-Algebra, die alle Intervalle in $\mathbb{R}$ enthält. Es gilt dann folgender Satz, den wir nicht beweisen werden:

Satz 2.15. *Sei $\mathbb{P}$ ein Wahrscheinlichkeitsmaß auf dem Messraum $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. Dann existiert genau eine monoton wachsende, rechts-stetige Funktion $F : \mathbb{R} \to [0,1]$ mit $\lim_{x\downarrow-\infty} F(x) = 0$ und $\lim_{x\uparrow+\infty} F(x) = 1$, so dass für jedes $a < b \in \mathbb{R}$,*

$$\mathbb{P}((a,b]) = F(b) - F(a). \tag{2.40}$$

umgekehrt gibt es für jede solche Funktion F ein einziges Wahrscheinlichkeitsmaß, so dass (2.40) gilt.

F heisst die Verteilungsfunktion von $\mathbb{P}$. Wir können uns Natürlich vorstellen, dass $\mathbb{P}$ die Verteilung einer Zufallsvariablen X ist. Es gibt drei wichtige Fälle.

- Falls F eine Stufenfunktion mit endlich vielen Sprüngen ist, dann ist $\mathbb{P}$ die Verteilung einer Zufallsvariablem die nur endlich viele Werte annimmt (siehe oben).
- Falls Falls F eine Stufenfunktion mit abzählbar vielen Sprüngen ist, dann ist $\mathbb{P}$ die Verteilung einer Zufallsvariablem die nur abzählbar viele Werte annimmt. Wir sagen dann $\mathbb{P}$ sein eine diskrete Verteilung.

- Es existiert eine nicht-negative Funktion $f : \mathbb{R} \to \mathbb{R}_+$, so dass für alle $a < b \in \mathbb{R}$

$$F(a) - F(b) = \int_a^b f(x)dx. \tag{2.41}$$

  Dann heisst *F absolut stetig* und wir sagen, $\mathbb{P}$ sei eine absolut stetige Verteilung. Die Funktion f heisst dann die *Dichte* der Verteilung, oder die Wahrscheinlichkeitsdichte von $\mathbb{P}$.

Wir wollen schliesslich noch kurz darauf eingehen, in wieweit sich auch das Konzept des Erwartungswertes einer Zufallsvariablen verallgemeinern lässt. Dazu bietet es sich an, die Formel (2.6) als *Definition* des Erwartungswertes zu verwenden. Dabei muss man sich natürlich nun die Frage stellen, ob und wann der Grenzwert in (2.6) auch existiert. Wir können jedenfalls erst einmal als vorläufige Definition festhalten.

Definition 2.16. Sei $(\Omega, \mathfrak{F}, \mathscr{P})$ ein Wahrscheinlichkeitsraum und X eine Zufallsvariable. Falls der Grenzwert in der rechten Seite der Gleichung (2.6) existiert, so nennen wir diesen den Erwartungswert von X, $\mathbb{E}[X]$.

Die folgende Beobachtung sagt uns, dass das keine so schlechte Idee ist.

Lemma 2.17. *Sei $(\Omega, \mathfrak{F}, \mathscr{P})$ ein Wahrscheinlichkeitsraum und X eine Zufallsvariable die nur nicht-negative Werte annimmt. Dann konvergiert die rechten Seite der Gleichung (2.6) zu einem Grenzwert in $[0, +\infty]$.*

Beweis. Da X nie negativ ist, so ist

$$\sum_{k=-\lceil \varepsilon^{-2} \rceil}^{+\lceil \varepsilon^{-2} \rceil} k\varepsilon \, \mathbb{P}(\{k\varepsilon \le X < (k+1)\varepsilon\}) = \sum_{k=0}^{+\lceil \varepsilon^{-2} \rceil} k\varepsilon \, \mathbb{P}(\{k\varepsilon \le X < (k+1)\varepsilon\}). \tag{2.42}$$

Die rechte Seite dieser Gleichung ist aber monoton wachsend in $1/\varepsilon$ (Warum?). Daher existiert der Grenzwert (möglicherweise als $+\infty$) wegen dem Satz von der monotonen Konvergenz. □

Anmerkung. Falls X eine Zufallsvariable ist, deren Wertebereich, S, abzählbar ist, so ist

$$\mathbb{E}[X] = \sum_{a \in S} a\mathbb{P}(X = a), \tag{2.43}$$

sofern die Summe auf der rechten Seite konvergiert. Wir sehen wieder, dass im Fall von positiven Zufallsvariablen, die Summe stets zu einem Wert in $[0, +\infty]$ konvergiert.

Als weiteres nützliches Ergebnis erhalten wir eine schöne Darstellung im Fall, wo die Verteilungsfunktion von X absolut stetig ist.

Lemma 2.18. *Sei $(\Omega, \mathfrak{F}, \mathscr{P})$ ein Wahrscheinlichkeitsraum und X eine Zufallsvariable mit absolut stetiger Verteilung F_X. Sei f die Dichte von F_X und es gelte*

$$\lim_{n\uparrow\infty} \int_{-n}^{n} x f(x) dx = E \tag{2.44}$$

existiert. Dann ist $E = \mathbb{E}[X]$.

Beweis. Wir haben

$$\begin{aligned}
\sum_{k=-\lceil\varepsilon^{-2}\rceil}^{+\lceil\varepsilon^{-2}\rceil} k\varepsilon\, \mathbb{P}(\{k\varepsilon \le X < (k+1)\varepsilon\}) &= \sum_{k=-\lceil\varepsilon^{-2}\rceil}^{+\lceil\varepsilon^{-2}\rceil} k\varepsilon \int_{k\varepsilon}^{(k+1)\varepsilon} f(x)dx \qquad (2.45)\\
&= \sum_{k=-\lceil\varepsilon^{-2}\rceil}^{+\lceil\varepsilon^{-2}\rceil} \int_{k\varepsilon}^{(k+1)\varepsilon} (k\varepsilon - x + x) f(x)dx\\
&= \sum_{k=-\lceil\varepsilon^{-2}\rceil}^{+\lceil\varepsilon^{-2}\rceil} \int_{k\varepsilon}^{(k+1)\varepsilon} x f(x)dx\\
&+ \sum_{k=-\lceil\varepsilon^{-2}\rceil}^{+\lceil\varepsilon^{-2}\rceil} \int_{k\varepsilon}^{(k+1)\varepsilon} (k\varepsilon - x) f(x)dx.
\end{aligned}$$

(Fuer die erste Summe haben wir

$$\sum_{k=-\lceil\varepsilon^{-2}\rceil}^{+\lceil\varepsilon^{-2}\rceil} \int_{k\varepsilon}^{(k+1)\varepsilon} x f(x)dx = \int_{-\lceil\varepsilon^{-2}\rceil}^{+\lceil\varepsilon^{-2}\rceil} x f(x)dx, \tag{2.46}$$

was nach Voraussetzung gegen E konvergiert. Für die zweite Summe gilt:

$$\begin{aligned}
&\left| \sum_{k=-\lceil\varepsilon^{-2}\rceil}^{+\lceil\varepsilon^{-2}\rceil} \int_{k\varepsilon}^{(k+1)\varepsilon} (k\varepsilon - x) f(x)dx \right| \qquad (2.47)\\
&\le \varepsilon \int_{-\lceil\varepsilon^{-2}\rceil}^{+\lceil\varepsilon^{-2}\rceil} f(x)dx = \varepsilon\,(F(+\infty) - F(-\infty)) \le \varepsilon,
\end{aligned}$$

was offenbar nach Null konvergiert (Wir haben hier $F(\pm\infty) \equiv \lim_{x\to\pm\infty} F(x)$ gesetzt). Daraus folgt aber sofort die Behauptung. □

Für Zufallsvariablem mit absolut stetiger Dichte ist also die Berechnung der Erwartungswerts auf die Berechnung eines Riemann-Integrals zurückgeführt. Dies ist natürlich ungemein praktisch, da wir nun die Techniken der Analysis einsetzen können.

2.6 Beispiele von Wahrscheinlichkeitsmaßen.

Das einfachste Wahrscheinlichkeitsmaß aus $\mathbb{R}$ ist wieder das *Dirac-Maß* an einem Punkt $t \in \mathbb{R}$, geschrieben δ_t. Es ist definiert durch

$$\delta_t(A) \equiv \mathbb{1}_A(t), \tag{2.48}$$

für jedes $A \in \mathfrak{B}(RR)$. Die zugehörige Verteilungsfunktion ist

$$F_t(x) \equiv \begin{cases} 0, & x < t, \\ 1, & x \geq t. \end{cases} \tag{2.49}$$

Einige besonders wichtige diskrete Verteilungen sind:

Bernoulli Verteilung $\mathrm{Ber}(p)$.

$$\mathbb{P}_p = p\,\delta_1 + (1-p)\,\delta_0. \tag{2.50}$$

Diese Verteilung kommt von einem Münzwurf, in dem mit Wahrscheinlichkeit p Kopf (und mit Wahrscheinlichkeit $(1-p)$ Zahl erscheint). Die Zufallsvariable X, definiert durch $X(\text{Kopf}) = 1, X(\text{Zahl}) = 0$ hat dann die Verteilung $\mathbb{P}$. Die Verteilungsfunktion ist gegeben durch

$$F_p(x) = \begin{cases} 0, & x < 0, \\ 1-p, & 0 \leq x < 1, \\ 1, & x \geq 1. \end{cases} \tag{2.51}$$

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

Die wichtige Binomialverteilung haben wir schon kennengelernt. Daraus sehen wir, dass die Verteilung von f gegeben ist durch

$$\mathbb{P}_{n,p} = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \delta_k. \tag{2.52}$$

Poissonverteilung $\mathrm{Poi}(\rho)$.

Eine weitere wichtige Verteilung is die Poissonverteilung, eingeführt von Simón-Denis Poisson (1781–1840). Die Poissonverteilung ist die Verteilung einer Zufallsvariablen, die Werte in den nicht-negativen ganzen Zahlen annimmt. Sie ist gegeben durch

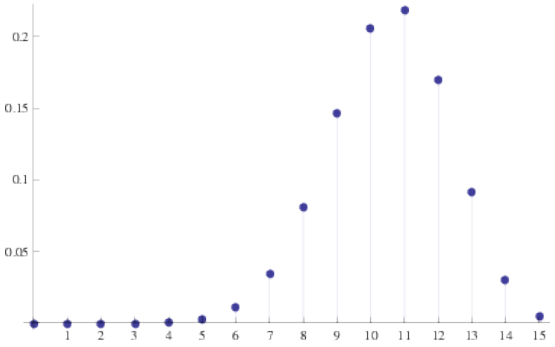


Abb. 2.1 Wahrscheinlichkeiten für $\mathrm{Bin}(n=15, p=0.7)$.

$$\mathbb{P}_\rho = \sum_{k=0}^{\infty} \frac{\rho^k}{k!} \mathrm{e}^{-\rho} \delta_k, \tag{2.53}$$

oder

$$\mathbb{P}_\rho(k) = \frac{\rho^k}{k!} \mathrm{e}^{-\rho}, \tag{2.54}$$

für $k \in \mathbb{N}_0$, wobei $\rho > 0$ ein Parameter ist. Die Poissonverteilung hängt mit der Binomialverteilung durch einen Grenzübergang zusammen. So können wir leicht sehen dass, wenn $p = \rho/n$ gewählt wird, die Koeffizienten $P_{n,\rho/n}(k)$ der Binomialverteilung gegen $P_\rho(k)$ (für festes k) konvergieren:

$$\lim_{n\uparrow\infty} \mathbb{P}_{n,\rho/n}(k) = \lim_{n\uparrow\infty} \frac{n!}{k!(n-k)!} \frac{\rho^k}{n^k} (1-\rho/n)^{n-k} = \frac{\rho^k}{k!} \mathrm{e}^{-\rho}. \tag{2.55}$$

Hier haben wir benutzt, dass

$$\frac{n!}{n^k(n-k)!} \to 1, \text{ wenn } n \uparrow \infty, \tag{2.56}$$

und

$$(1-\rho/n)^n \to \mathrm{e}^{-\rho}, \text{ wenn } n \uparrow \infty, \tag{2.57}$$

und schliesslich $(1-\rho/n)^{-k} \to 1$.

Die Poissonverteilung ist die kanonische Verteilung für die Häufigkeit des Auftretens sehr seltener Ereignisse.

Geometrische Verteilung $\mathrm{Geo}(q)$.

Dies ist wieder eine Verteilung auf den natürlichen Zahlen mit

$$\mathbb{P}_q(k) = q^{k-1}(1-q), \quad k \in \mathbb{N}. \tag{2.58}$$

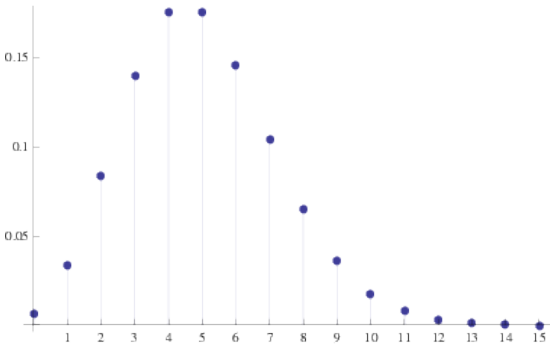


Abb. 2.2 Wahrscheinlichkeiten für Poi($\rho = 5$).

Sie hat eine wichtige Interpretation im Kontext des unendlich oft wiederholten Münzwurfs mit Parameter p: Wenn N die Nummer des Münzwurfs bezeichnet, bei dem erstmalig "Zahl" ($= 0$) erscheint, dann ist mit

$$\mathbb{P}(\{N = k\}) = p^{k-1}(1-p) = P_p(k).$$

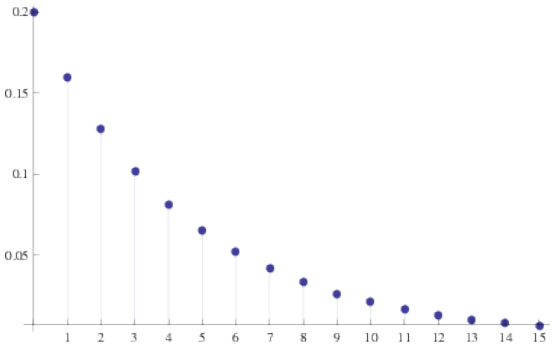


Abb. 2.3 Wahrscheinlichkeiten für Geo($q = 0.2$).

Eine Vielzahl in der Praxis verwendeter Wahrscheinlichkeitsmaße ist absolut stetig. Dies liegt, wenigstens zum Teil, daran, dass diese einfacher zu handhaben sind wenn es um konkrete Berechnungen geht. Wichtige Beispiele sind etwa:

Gleichverteilung $\mathscr{U}_I$.

Für ein Intervall $I = [a,b] \subset \mathbb{R}$ ist die Gleichverteilung auf I definiert als das W-Maß auf $\mathbb{R}$ mit Verteilungsfunktion

$$F(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \le x < b, \\ 1, & x \ge b. \end{cases} \tag{2.59}$$

Wie wir sehen, ist diese Verteilung absolut stetig mit der Wahrscheinlichkeitsdichte

$$f(x) = \frac{1}{|b-a|} \mathbb{1}_{[a,b]}(x). \tag{2.60}$$

Gauß-Verteilung $\mathcal{N}(m, \sigma^2)$.

Die mit Abstand wichtigste Verteilung ist ebenfalls absolut stetig mit Wahrscheinlichkeitsdichte

$$\phi_{m,\sigma^2}(x) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \tag{2.61}$$

wobei $m \in \mathbb{R}$ Mittelwert, $\sigma > 0$ Standardabweichung und σ^2 Varianz heisst. Man kann leicht nachprüfen, dass die Funktion

$$\Phi_{m,\sigma^2}(x) \equiv \int_{-\infty}^{x} \phi_{m,\sigma^2}(x)dx \tag{2.62}$$

in der Tat eine Verteilungsfunktion ist (man muss nur nachprüfen, dass $\lim_{x\uparrow\infty} \Phi_{m,\sigma^2}(x) = 1$). Eine Zufallsvariable X mit dieser Verteilung hat den Erwartungswert m und die Varianz σ^2, was man durch Berechnung der entsprechenden Integrale nachprüft:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\phi_{m,\sigma^2}(x)dx, \tag{2.63}$$

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \int_{-\infty}^{\infty} (x-m)^2 \phi_{m,\sigma^2}(x)dx. \tag{2.64}$$

Wenn $m = 0$ und $\sigma^2 = 1$, so nennt man die Verteilung die *Standardnormalverteilung*. Beachte, dass man alle Berechnungen durch die Substitution $(x-m)/\sigma = y$ auf diesen Fall reduzieren kann.

Aus vielen guten Gründen ist die Gauß-Verteilung die erste Wahl, wenn es um die Verteilung von Abweichungen um ein typisches Verhalten geht. Der Grund hierfür wird sich bei der Diskussion des zentralen Grenzwertsatzes offenbaren.

Interessanterweise wurde die Gauß-Verteilung von dem in England lebenden Franzosen Abraham de Moivre (26.05.1667–27.11.1754) 1733 als Approximation der Binomialverteilung eingeführt. Gauß benutzte sie erst 1794 (publiziert 1809) in der Fehlerrechnung (Methode der kleinsten Quadrate).

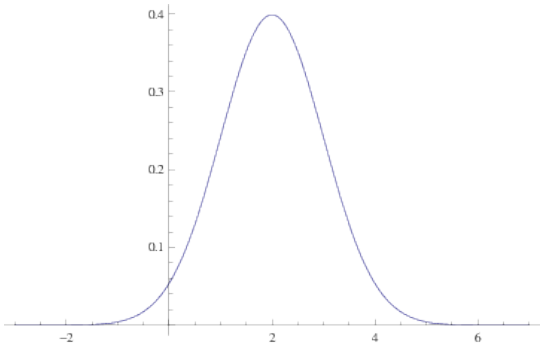


Abb. 2.4 Dichte der Gauß-Verteilung für $m = 2$ und $\sigma = 1$.

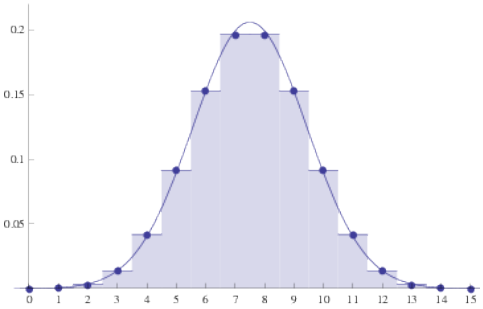


Abb. 2.5 Gauß-Verteilung und approximierende Binomialverteilung für $n = 15$ und $p = 0.5$.

Exponentialverteilung $\mathrm{Exp}(a)$.

Dies ist wieder eine absolut stetige Verteilung mit ist die Wahrscheinlichkeitsdichte

$$f_a(x) = a\mathrm{e}^{-ax}\mathbb{1}_{[0,\infty)}(x). \tag{2.65}$$

Im Gegensatz zur Gauß-Verteilung können wir hier die Verteilungsfunktion explizit ausrechnen:

$$F_a(x) = \begin{cases} 0, & x < 0, \\ 1 - \mathrm{e}^{-ax}, & x \geq 0. \end{cases} \tag{2.66}$$

Die Exponentialverteilung tritt insbesondere als Verteilung von Wartezeiten gerne auf. Ihr Charakteristikum ist die "Gedächtnislosigkeit". $a > 0$ is ein Parameter. Wie man nachrechnet ist der Erwartungswert einen Zufallsvariablen, deren Verteilung die Exponentialverteilung mit Parameter a is gegeben durch

$$\mathbb{E}[X] = \int_0^\infty xa\mathrm{e}^{-ax}dx = a^{-1}. \tag{2.67}$$

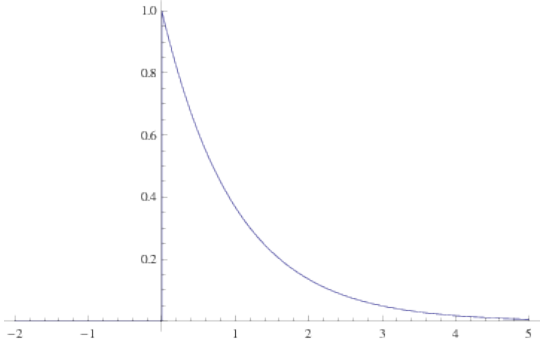


Abb. 2.6 Dichte der Exponentialverteilung mit $a = 1$.

Cauchy-Verteilung Cauchy(a).

Diese hat die Dichte

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

Diese Verteilung zeichnet sich dadurch aus, dass die Funktion x nicht gegen sie integrierbar ist, d.h. dass kein Mittelwert existiert.

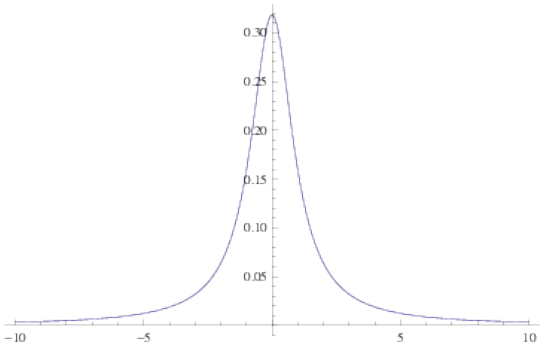
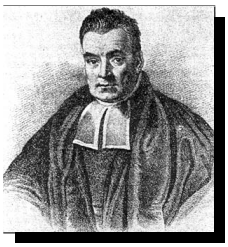


Abb. 2.7 Dichte der Cauchyverteilung mit $a = 1$.

Kapitel 3
Bedingte Wahrscheinlichkeiten, Unabhängigkeit, Produktmaße

In diesem Kapitel führen wir einige der wichtigsten Konzepte der Wahrscheinlichkeitstheorie ein. Unabhängige Zufallsvariablen sind die wichtigsten Bausteine um kompliziertere Modelle von zufälligen Ereignissen zu konstruieren.

3.1 Bedingte Wahrscheinlichkeiten



Wir betrachten nunmehr einen Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$. Es seien $A, B \in \mathfrak{F}$ zwei Ereignisse. Die Wahrscheinlichkeit von $A \cap B$, d.h. das gleichzeitige Eintreten beider Ereignisse ist $\mathbb{P}(A \cap B) \leq \min(\mathbb{P}(A), \mathbb{P}(B))$. Was uns nun interessiert ist, wie Information über das Ereignis B unsere Annahmen über das Ereignis A beeinflussen. Dazu definieren wir die *bedingte Wahrscheinlichkeit:*

Definition 3.1. Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und seien $A, B \in \mathfrak{F}$. Sei $\mathbb{P}(B) > 0$. Dann heisst

$$\mathbb{P}(A|B) \equiv \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{3.1}$$

die *bedingte Wahrscheinlichkeit* von A gegeben B.

Diese Definition der bedingten Wahrscheinlichkeit ist einleuchtend und kompatibel mit der frequentistischen Interpretation von Wahrscheinlichkeiten: Wenn $\mathbb{P}$ eine empirische Verteilung ist, dann stellt $\mathbb{P}(A|B)$ offenbar die Frequenz des Eintretens von A unter all den Experimenten mit Ausgang in B dar.

Die bedingte Wahrscheinlichkeit hat zwei wichtige Eigenschaften:

Satz 3.2. *Sei $B \in \mathfrak{F}$ mit $\mathbb{P}(B) > 0$.*

(i) Die bedingte Wahrscheinlichkeit, $\mathbb{P}(\cdot|B)$ definiert ein Wahrscheinlichkeitsmaß auf dem Raum $(B, \mathfrak{F}\cap B)$, wo

$$\mathfrak{F}\cap B \equiv \{A\cap B, A\in\mathfrak{F}\} \tag{3.2}$$

(ii) Sei $B_n \in \mathfrak{F}$, $n\in\mathbb{N}$, eine paarweise disjunkte Folge von Mengen, so dass (a) $\cup_{n\in\mathbb{N}} B_n = \Omega$, (b) $\mathbb{P}(B_n) > 0$, für alle n. Dann gilt, dass, für alle $A\in\mathfrak{F}$,

$$\sum_{n\in\mathbb{N}} \mathbb{P}(A|B_n)\mathbb{P}(B_n) = \mathbb{P}(A) \tag{3.3}$$

Beweis. Wir beweisen den Satz nur für den Fall, dass Ω endlich ist. Zunächst ist $\mathfrak{F}\cap B$ eine Algebra. Erstens ist $\emptyset = \emptyset\cap B \in \mathfrak{F}\cap B$. Sein $A\in\mathfrak{F}$. Dann ist das Komplement von $A\cap B$ in B gerade $B\backslash(A\cap B) = A^c\cap B \in \mathfrak{F}\cap B$. Schliesslich ist mit $A_1, A_2 \in \mathfrak{F}$ auch $(A_1\cap B)\cup(A_2\cap B) = (A_1\cup A_2)\cap B \in \mathfrak{F}\cap B$.

Als nächstes prüfen wir, ob $\mathbb{P}(\cdot|B)$ ein Wahrscheinlichkeitsmaß ist. Offenbar gilt $\mathbb{P}(B|B) = 1$ und $\mathbb{P}(\emptyset|B) = 0$. Weiterhin gilt, dass, für $A\subset$B,

$$\begin{aligned}\mathbb{P}(B\backslash A|B) &= \frac{\mathbb{P}(B\backslash A)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B)-\mathbb{P}(A)}{\mathbb{P}(B)} = 1-\mathbb{P}(A|B).\end{aligned} \tag{3.4}$$

Für A_1, A_2 disjunkte Teilmengen von B gilt

$$\mathbb{P}\Big(A_1\cup A_2\big|B\Big) = \frac{\mathbb{P}(A_1\cup A_2)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A_1)+\mathbb{P}(A_2)}{\mathbb{P}(B)} = \mathbb{P}(A_1|B)+P(A_2|B). \tag{3.5}$$

und somit gilt (i).

Wegen (ii) schreiben wir

$$\begin{aligned}\sum_{n\in\mathbb{N}} \mathbb{P}(A|B_n)\mathbb{P}(B_n) &= \sum_{n\in\mathbb{N}} \mathbb{P}(A\cap B_n) \\ &= \mathbb{P}(A\cap\cup_n B_n) = \mathbb{P}(A\cap\Omega) = \mathbb{P}(A).\end{aligned} \tag{3.6}$$

□

Definition 3.3. Zwei Ereignisse $A, B\in\mathfrak{F}$, mit $\mathbb{P}(B) > 0$ und $\mathbb{P}(A) > 0$, heissen *unabhängig*, genau dann wenn

$$\mathbb{P}(A|B) = \mathbb{P}(A), \tag{3.7}$$

beziehungsweise (was das gleiche ist), wenn

$$\mathbb{P}(A\cap B) = \mathbb{P}(A)\mathbb{P}(B). \tag{3.8}$$

Allgemeiner heissen n Ereignisse, $A_1, \ldots, A_n$ unabhängig, genau dann, wenn für alle $m\le n$, und $1\le i_1 < i_2 < \ldots < i_m \le n$ gilt

$$\mathbb{P}\left(\bigcap_{k=1}^{m} A_{i_k}\right) = \prod_{k=1}^{m} \mathbb{P}(A_{i_k}) \tag{3.9}$$

Es ist bequem den Begriff der Unabhängigkeit auch auf σ-Algebren zu übertragen,

Definition 3.4. Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum und seien $\mathfrak{F} \subset \mathfrak{F}$ und $\mathfrak{B} \subset \mathfrak{F}$ ebenfalls σ-Algebren. Dann heissen $\mathfrak{F}$ und $\mathfrak{B}$ *unabhängig*, genau dann wenn für alle Ereignisse $A \in \mathfrak{F}$ und $B \in \mathfrak{B}$, A und B unabhängig sind.

Ein triviales Korollar aus der Definition der bedingten Wahrscheinlichkeit ist die berühmte *Bayes'sche Formel:*

Satz 3.5. *Seien $A, B \in \mathfrak{F}$ und $\mathbb{P}(A) > 0$, $\mathbb{P}(B) > 0$. Dann gilt*

$$\mathbb{P}(B|A) = \mathbb{P}(A|B)\frac{\mathbb{P}(B)}{\mathbb{P}(A)} \tag{3.10}$$

Beweis. Der Beweis folgt durch Einsetzen der Definition der bedingten Wahrscheinlichkeit in beiden Seiten. □

Die Formel ist in der Statistik von grosser Bedeutung. Thomas Bayes (1702 - 1761) hat diesen Satz in seinem Werk *"Essay towards solving a problem in the doctrine of chances"* in einem speziellen Fall hergeleitet. Da Bayes von Beruf Priester war, ist sein Interesse an Wahrscheinlichkeiten wohl rein akademischer Natur gewesen. Ein Beispiel soll zeigen, dass man aus ihr durchaus nicht völlig intuitive Ergebnisse gewinnen kann.

Beispiel. Ein Test auf Vogelgrippe liefert mit Wahrscheinlichkeit von 99% ein korrektes Ergebnis. Ein bekanntes Pharmaunternehmen empfiehlt, sich sofort testen zu lassen, und bei positivem Resultat sofort Oseltamivirphosphat prophylaktisch einzunehmen. Für wen ist das sinnvoll?

Wir nehmen dazu an, dass der tatsächliche Durchseuchungsgrad x beträgt. Wir bezeichnen das Ereignis "krank" mit A und das Ereignis "Test richtig" mit B. Dann ist das Ereignis $C =$"positiv auf Vogelgrippe getestet" gegeben durch

$$C = (A \cap B) \cup (A^c \cap B^c)$$

Offenbar gilt

$$\mathbb{P}(A \cap B) = x \times 0.99$$

und

$$\mathbb{P}(A^c \cap B^c) = (1-x) \times 0.01$$

Insbesondere ist

$$\mathbb{P}(C) = 0.01 + x \times 0.98 \geq 1\%$$

, unabhängig vom tatsächlichen Wert von x.

Angenommen nun, eine Versuchsperson sei positiv getestet worden. Wie wahrscheinlich ist es, dass sie auch krank ist? Dazu müssen wir $\mathbb{P}(A|C)$ berechnen. Nach der Formel von Bayes ist dann

$$\mathbb{P}(A|C) = \mathbb{P}(C|A)\frac{\mathbb{P}(A)}{\mathbb{P}(C)} = \frac{\mathbb{P}(C\cap A)}{\mathbb{P}(C)} = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(C)} \tag{3.11}$$
$$= \frac{x\times 0.99}{x\times 0.99+(1-x)\times 0.01}.$$

Wenn $x \ll 1$ ist, dann ist im wesentlichen $\mathbb{P}(A|C) = 100\mathbb{P}(A) \ll 1$, d.h. der Test hat eigentlich keine neue Information gebracht, bzw. fast alle positiv getesteten erweisen sich im Nachhinein als gesund....

3.2 Unabhängige Zufallsvariablen

Wir kommen nun zu dem sowohl wichtigen als auch nützlichen Konzept der unabhängigen Zufallsvariablen.

Definition 3.6. Sei $(\Omega,\mathfrak{F},\mathbb{P})$ ein Wahrscheinlichkeitsraum und seien X_1,X_2 Zufallsvariablen. X_1 und X_2 heissen *unabhängig*, genau dann unabhängig sind, wenn für alle Mengen $B_1,B_2\in\mathfrak{B}$,

$$\mathbb{P}(\{X_1\in B_1\}\cap\{X_2\in B_2\}) = \mathbb{P}(\{X_1\in B_1\})\mathbb{P}(\{X_2\in B_2\}). \tag{3.12}$$

Anmerkung. Wenn die Zufallsvariablen nur abzählbar viele Werte annehmen, so genügt es dass für alle $a\in X_1(\Omega)$, $b\in X_2(\Omega)$ (also für a,b in den jeweiligen Wertebereichen der Zufallsvariablen),

$$\mathbb{P}(\{X_1=a\}\cap\{X_2=b\}) = \mathbb{P}(\{X_1=a\})\mathbb{P}(\{X_2=b\}). \tag{3.13}$$

Die Verallgemeinerung auf eine beliebige Anzahl von Zufallsvariablen ergibt sich analog:

Definition 3.7. Sei $(\Omega,\mathfrak{F},\mathbb{P})$ ein Wahrscheinlichkeitsraum und seien $X_1,X_2,\dots X_n$ eine Familie von Zufallsvariablen. Dann heisst diese Familie eine Familie von unabhängigen Zufallsvariablen, genau dann wenn für alle Mengen $B_1,\dots,B_n\in\mathfrak{B}$ die Ereignisse $X_1(B_1),\dots,X^{-1}(B_n)$ unabhängig sind.

Die Bemerkung oben überträgt sich analog.
Das folgende Lemma gibt eine alternative Charakterisierung der Unabhängigkeit.

Lemma 3.8. *Sei $(\Omega,\mathfrak{F},\mathbb{P})$ ein Wahrscheinlichkeitsraum, und seien X_1,X_2 unabhängige Zufallsvariablen. Seien g_1,g_2 messbare Funktionen von $(\mathbb{R},\mathfrak{B})$ nach $(\mathbb{R},\mathfrak{B})$. Es seien ferner $\mathbb{E}\,|g_i(X_i)|<\infty$. Dann gilt*

$$\mathbb{E}\,[g_1(X_1)g_2(X_2)] = \mathbb{E}\,[g_1(X_1)]\,\mathbb{E}\,[g_2(X_2)]\,. \tag{3.14}$$

Beweis. Wir beweisen diesen Satz nur für den Fall, dass die Zufallsvariablen X_i nur endlich viele Werte annehmen. Dann interessiert uns nur die Restriktion der

Funktionen g_i auf die jeweiligen Wertebereiche $S_i = X_i(\Omega)$. Daraus folgt aber, dass für jedes $o \in \Omega$,

$$g_i(X_i(\omega)) = \sum_{a \in S_i} g_i(a) \mathbb{1}_a(X_i(\omega)). \tag{3.15}$$

Damit gilt wegen der Linearität des Erwartungswerts

$$\begin{aligned}
\mathbb{E}[g_1(X_1)g_2(X_2)] &= \sum_{a_1 \in S_1} \sum_{a_2 \in S_2} g_1(a_1)g_2(a_2)\mathbb{E}[\mathbb{1}_{a_1}(X_1)\mathbb{1}_{a_2}(X_2)] \\
&= \sum_{a_1 \in S_1} \sum_{a_2 \in S_2} g_1(a_1)g_2(a_2)\mathbb{P}(\{X_1 = a_1\} \cap \{X_2 = a_2\}) \\
&= \sum_{a_1 \in S_1} \sum_{a_2 \in S_2} g_1(a_1)g_2(a_2)\mathbb{P}(X_1 \in A_1)\mathbb{P}(X_2 \in A_2) \\
&= \mathbb{E}[g_1(X_1)]\mathbb{E}[g_2(X_2)],
\end{aligned} \tag{3.16}$$

wo wir (3.13) benutzt haben. Dies liefert sofort (3.14). □

Anmerkung. Die Übertragung auf den Fall, dass X_i abzählbar viele Werte annimmt ist unproblematisch.

Übung. Beweisen Sie den Umkehrschluss zu Lemma 3.8, d.h., wenn (3.14) gilt für alle Wahl von g_1, g_2, dann sind X_1 und X_2 unabhängig.

Eine Eigenschaft, die der aus dem Lemma ähnlich sieht, aber deutlich schwächer ist, ist die sogenannte *Unkorreliertheit* von Zufallsvariablen.

Definition 3.9. Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum, und seien X_1, X_2 Zufallsvariablen. X_1 und X_2 heissen *unkorreliert*, genau dann wenn gilt

$$\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]. \tag{3.17}$$

Offensichtlich ist die Unkorreliertheit viel leichter nachzuprüfen als die Unabhängigkeit. Häufig wird erstere darum auch als erstes Indiz für die Unabhängigkeit benutzt. Allerdings muss man sich klarmachen, dass dieses Indiz keinesfalls schlüssig ist. So seien X, Y zwei unabhängige, gleichverteilte Zufallsvariablen, und $Z_+ \equiv X + Y$, $Z_- \equiv X - Y$. Dann sind Z_+, Z_- unkorreliert. Im allgemeinen sind sie aber nicht unabhängig. Dazu betrachten wir den Fall der Bernoulli Verteilung mit Parameter $p = 1/2$. Dann ist

$$\mathbb{P}(Z_- = 0 | Z_+ = 2) = 1 \quad \text{aber} \quad \mathbb{P}(Z_- = 0 | Z_+ = 1) = 0,$$

was sofort die Unabhängigkeit falsifiziert.

3.3 Produkträume

Unabhängige Zufallsvariablen können wir explizit konstruieren. Dazu betrachten wir zwei Wahrscheinlichkeitsräume, $(\Omega_1, \mathfrak{F}_1, \mathbb{P}_1)$ und $(\Omega_2, \mathfrak{F}_2, \mathbb{P}_2)$ und messbare

Funktionen $X_1 : \Omega_1 \to \mathbb{R}$, $X_2 : \Omega_2 \to \mathbb{R}$. Die Idee ist, einen Wahrscheinlichkeitsraum über dem Produktraum $\Omega_1 \times \Omega_2$ zu konstruieren, bezüglich dessen X_1 und X_2 unabhängige Zufallsvariablen sind. Dazu führen wir zunächst die entsprechende σ-Algebra ein.

Definition 3.10. Die *Produkt-σ-Algebra*, $\mathfrak{F}_1 \otimes \mathfrak{F}_2$, ist die kleinste σ-Algebra, die alle Mengen der Form $C = A \times B$ mit $A \in \mathfrak{F}_1, B \in \mathfrak{F}_2$ enthält.

Wir nennen Mengen der Form $A \times B$ gelegentlich Rechtecke, obwohl das etwas irreführend ist.

Der nächste Schritt ist die Konstruktion eines W-Maßes auf $(\Omega_1 \times \Omega_2, \mathfrak{F}_1 \otimes \mathfrak{F}_2)$ für das die Unter-σ-Algebren $\mathfrak{F}_1 \times \Omega_2$ und $\Omega_1 \times \mathfrak{F}_2$ unabhängig sind.

Sei $C \in \mathfrak{F}_1 \otimes \mathfrak{F}_2$. Für jedes $x \in \Omega_1$ und jedes $y \in \Omega_2$ führen wir die Mengen

$$C_x \equiv \{y \in \Omega_2 : (x,y) \in C\} \tag{3.18}$$

und

$$C^y \equiv \{x \in \Omega_1 : (x,y) \in C\} \tag{3.19}$$

ein. Diese Mengen sind gerade die "Scheiben"die man aus den in der Produkt-σ-Algebra enthaltenen Mengen herausschneiden kann. Entsprechend definieren wir auch für jede messbare Funktion X auf $\Omega_1 \times \Omega_2$ für jedes $x \in \Omega_1$ die Funktion $X_x(y) \equiv X(x,y)$ und für jedes $y \in \Omega_2$ die Funktion $X^y(x) \equiv X(x,y)$. Das nächste Lemma zeigt, dass die Menge C_x und C^y jeweils in $\mathfrak{F}_2$ bzw. $\mathfrak{F}_2$ liegen, sowie dass die Funktionen X_x und X^y messbar bez. der σ-Algebren $\mathfrak{F}_2$ bzw. $\mathfrak{F}_1$ sind.

Lemma 3.11. *Mit den Definitionen von oben gilt:*

(i) Für jedes $C \in \mathfrak{F}_1 \otimes \mathfrak{F}_2$ und $x \in \Omega_1, y \in \Omega_2$ ist $C_x \in \mathfrak{F}_2$ und $C^y \in \mathfrak{F}_1$.
(ii) Für jede messbare Funktion, $X : \Omega_1 \times \Omega_2 \to \mathbb{R}$, und $x \in \Omega_1, y \in \Omega_2$ ist X_x messbar bezüglich $\mathfrak{F}_2$ und X^y messbar bezüglich $\mathfrak{F}_1$.

Beweis. Wir setzen für $x \in \Omega_1$ (für $y \in \Omega_2$ ist das Beweis analog),

$$\mathfrak{C}_x \equiv \{C \in \mathfrak{F}_1 \otimes \mathfrak{F}_2 : C_x \in \mathfrak{F}_2\}.$$

Dann enthält $\mathfrak{C}_x$ sicher die einfachen Mengen $C = A \times B$ mit $A \in \mathfrak{F}_1$ und $B \in \mathfrak{F}_2$. Denn entweder ist dann $x \in A$ und $C_x = B$, oder $x \notin A$ und $C_x = \emptyset$. Beidemal ist $C_x \in \mathfrak{F}_2$. Nun kann man andererseits leicht nachweisen, dass $\mathfrak{C}_x$ eine σ-Algebra ist. Da diese aber den Erzeuger von $\mathfrak{F}_1 \otimes \mathfrak{F}_2$ enthält, andererseits per Konstruktion nicht grösser als $\mathfrak{F}_1 \otimes \mathfrak{F}_2$ ist, muss $\mathfrak{C}_x = \mathfrak{F}_1 \otimes \mathfrak{F}_2$ gelten.

Weiter ist für jede messbare Menge $D \subset \mathbb{R}$,

$$\begin{aligned} X_x^{-1}(D) &= \{y \in \Omega_2 : X_x(y) \in D\} = \{y \in \Omega_2 : X(x,y) \in D\} \\ &= \{y \in \Omega_2 : (x,y) \in X^{-1}(D)\} = (X^{-1}(D))_x, \end{aligned} \tag{3.20}$$

die aber nach (i) in $\mathfrak{F}_2$ liegt. Damit ist das Lemma bewiesen. □

Satz 3.12. *Seien $\mathbb{P}_1, \mathbb{P}_2$ Wahrscheinlichkeitsmaße auf $(\Omega_1, \mathfrak{F}_1)$, bzw. $(\Omega_2, \mathfrak{F}_2)$. Dann existiert ein einziges Wahrscheinlichkeitsmaß, $\mathbb{P} \equiv \mathbb{P}_1 \otimes \mathbb{P}_2$, genannt das* Produktmaß*, auf der Produkt-σ-Algebra, $\mathfrak{F}_1 \otimes \mathfrak{F}_2$, mit der Eigenschaft, dass für alle $A \in \mathfrak{F}_1$ und $B \in \mathfrak{F}_2$*

$$\mathbb{P}_1 \otimes \mathbb{P}_2(A \times B) = \mathbb{P}_1(A)\mathbb{P}_2(B). \tag{3.21}$$

Beweis. Wir beweisen die Aussage wieder nur für den Fall endlicher Mengen. Es folgt dann nämlich aus dem vorherigen Lemma, dass jede Menge in $\mathfrak{F}_1 \otimes \mathfrak{F}_2$ eine disjunkte Vereinigung von endlich vielen Rechtecken ist. Damit legt die Formel zusammen mit der Additivitätsregel die Masse jedes Elements der Produkt-σ-Algebra fest. Die Gültigkeit der Additivität für Rechtecke ist ebenfalls leicht nachzuprüfen, wobei man gut Lemma 1.11 verwenden kann. □

Der Punkt ist nun, dass, wenn X_i Zufallsvariablen auf $(\Omega_i, \mathfrak{F}_i)$, $i = 1, 2$, sind, dann sind X_1 und X_2 unabhängige Zufallsvariablen auf dem Wahrscheinlichkeitsraum $(\Omega_1 \times \Omega_2, \mathfrak{F}_1 \otimes \mathfrak{F}_2, \mathbb{P}_1 \otimes \mathbb{P}_2)$ sind. Dies ist die kanonische Konstruktion von unabhängigen Zufallsvariablen.

Es ist offensichtlich, dass durch Iteration die obige Konstruktion auf beliebige endliche Produkte von Wahrscheinlichkeitsmaßen ausgedehnt werden kann.

Im Fall endlicher Mengen und wenn die σ-Algebren die Potenzmengen sind, wird alles noch etwas einfacher:

Lemma 3.13. *Seien $(\Omega_1, \mathscr{P}(\Omega_1), \mathbb{P}_1), \ldots, (\Omega_n, \mathscr{P}(\Omega_n), \mathbb{P}_n)$ endliche Wahrscheinlichkeitsräume und seien $p_i, \ldots, p_n$ die jeweiligen Wahrscheinlichkeitsvektoren. Dann gilt:*

(i) $\mathscr{P}(\Omega_1) \otimes \cdots \otimes \mathscr{P}(\Omega_n) = \mathscr{P}(\Omega_1 \times \cdots \times \Omega_n)$*;*
(ii) Das Produktmass $\mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n$ hat den Wahrscheinlickeitsvektor

$$p(\omega) = p(\omega_1, \ldots, \omega_n) = p_1(\omega_1) \ldots p_n(\omega_n). \tag{3.22}$$

Beweis. Der Beweis ist sehr einfach und soll als Übung ausgeführt werden. □

Korollar 3.14. *Es seien $X_1, \ldots, X_n$ unabhängige Zufallsvariablen die nur endlich viele Werte annehmen. Für $x_k \in S_k \equiv X_k(\Omega)$ sei $p_k(x_k) = \mathbb{P}(X_k = x_k)$. Sein ferner $F : \mathbb{R}_n \to \mathbb{R}$ eine Funktion. Dann gilt*

$$\mathbb{E}[F(X_1, \ldots, X_n)] = \sum_{x_1 \in S_k} \ldots \sum_{x_n \in S_n} F(x_1, \ldots, x_n) p_1(x_1) \ldots p_n(x_n). \tag{3.23}$$

Beweis. Übung! □

Beispiel. Wir betrachten das Werfen von n Münzen. Der Zustandsraum jeder Münze ist $\Omega_i = \{0, 1\}$. Dann ist der Zustandsraum der n Würfe $\Omega_1 \times \cdots \times \Omega_n = \{0, 1\}^n$. Jede einzelne Münze hat eine Bernoulliverteillung mit Parameter p. Die Zufallsvariablen $X_1, \ldots, X_n$, wo $X_i(\omega_1, \ldots, \omega_n) = \omega_i$ sind dann unter dem n-fachen Produktmaß unabhängig und gleichverteilt.

Unabhängige Zufallsvariablen sind ein wesentlicher Baustein der Wahrscheinlichkeitstheorie. Vielfach wird im alltäglichen Sprachgebrauch der Begriff Unabhängigkeit mit dem der Zufälligkeit gleichgesetzt. So geht man stillschweigend davon aus, dass die sukzessiven Ausgänge eines Roulettspiels unabhängig sind, und wird dies als den zufälligen Charakter des Spiels betrachten.

Beispiel. (Gewinnen mit bedingter Wahrscheinlichkeit). Ein schönes Beispiel, das zeigt wie man Nutzen aus der Kenntnis des Konzepts der bedingten Wahrscheinlichkeit und Produktmaß ziehen kann, ist folgendes Spiel. Alice schreibt zwei Zahlen, auf je einen Zettel. Dann wirft sie eine faire Münze und zeigt Bob je nach Ausgang des Wurfs entweder den einen oder den anderen Zettel. Nennen wir die gezeigte Zahl im folgenden y und die versteckte Zahl x. Die Aufgabe von Bob besteht darin, zu erraten, ob $x > y$ oder ob $x < y$. Alice bietet Bob eine Wette mit Quote $1:2$ an. Soll Bob die Wette annehmen?

Die Antwort auf die Frage ist ja, und zwar weil Bob in der Lage ist, die richtige Antwort mit einer Wahrscheinlichkeit vom mehr als $1/2$ zu geben. Dazu muss er sich nur eine geschickte Strategie ausdenken!

Eine solche Strategie sieht so aus: Bob zieht gemäß einer Gaußverteilung $\mathcal{N}(0,100)$ eine Zufallszahl, Z. Nun vergleicht er x mit Z: Wenn $Z \geq y$, so rät er $y < x$, wenn $Z < y$ rät er $x < y$.

Um zu sehen, warum das funktioniert, wollen wir das ganze etwas formalisieren. Gegeben sind zwei Zahlen, $x_0 < x_1$. Ferner gibt es eine Bernoulli Zufallsvariable, B, mit Parameter $1/2$, definiert auf einem W-Raum $(\Omega_1, \mathfrak{F}_1, \mathbb{P}_1)$. Die Bob zugängliche Information ist nur die Zufallsvariable $Y = x_B$. Ziel des Spiels ist es, B zu schätzen, denn wenn Bob B kennt, kann es sagen, ob Y gleich x_0 oder x_1 ist, mithin ob es die grössere oder die kleinere Zahl war. Das bedeutet, dass Bob eine neue Zufallsvariable konstruieren will, die von Y abhängt und B voraussagen lässt. Dazu führt der Spieler einen neuen Wahrscheinlichkeitsraum $(\Omega_2, \mathfrak{F}_2, \mathbb{P}_2)$ ein, auf dem er eine Gauß'sche Zufallsvariable, Z konstruiert. Nun betrachten wir den Produktraum, $(\Omega_1 \times \Omega_2, \mathfrak{F}_1 \otimes \mathfrak{F}_2, \mathbb{P} \equiv \mathbb{P}_1 \otimes \mathbb{P}_2)$. Auf diesem sind die Zufallsvariablen B und Z unabhängig. Bob's Strategie ist es, auf diesem Produktraum eine neue Zufallsvariable, A, zu konstruieren, deren Wert nur von (den dem Spieler bekannten Werten von) Z und Y abhängt ist, die aber mit B positiv korreliert in dem Sinne, dass

$$\mathbb{P}(A = B) > 1/2.$$

Die Wahl von A ist

$$A \equiv \mathbb{1}_{Z<Y}$$

Wir sehen, dass, da Y ja von B abhängt, A und B nicht unabhängig sind. In der Tat ist

Nun können wir benutzen, dass, wenn $B = 1$, $Y = x_1$, und wenn $B = 0$, $Y = x_0$. Also folgt

$$\begin{aligned}\mathbb{P}(A=B) &= \mathbb{P}(\{Z<Y\}\cap\{B=1\})+\mathbb{P}(\{Z\geq Y\}\cap\{B=0\})\\ &= \frac{1}{2}\mathbb{P}(\{Z<x_B\}|\{B=1\})+\frac{1}{2}\mathbb{P}(\{Z\geq x_B\}|\{B=0\}\\ &= \frac{1}{2}\left(\mathbb{P}_2(Z<x_1)+\mathbb{P}_2(Z\geq x_0)\right)\\ &= \frac{1}{2}+\frac{1}{2}\mathbb{P}_2(x_0\leq Z<x_1)>\frac{1}{2}.\end{aligned}$$

Das wollten wir aber nur zeigen.

3.4 Unendliche Produkte

Natürlich würden wir letztlich gerne von der Verteilung von "beliebig", also "unendlich" vielen Zufallsexperimenten, etwa Münzwürfen, sprechen. Ist das wirklich so schwierig? Wir könnten zunächst geneigt sein, diese Frage zu verneinen. Nehmen wir dazu als einfache Räume Ω_i endliche Mengen (etwa $\Omega_i=\{0,1\}$). Die Frage ist dann, was die geeignete σ-Algebra für den unendlichen Produktraum $\otimes_{i=1}^{\infty}\Omega_i$ sein soll. Ähnlich wie im Fall der reellen Zahlen kann man hier nicht einfach die Potenzmenge nehmen, sonder man betrachtet eine kleinere σ-Algebra, die *Produkt-σ-Algebra* Wir wollen uns bei unserem Vorgehen von praktischen Erwägungen leiten lassen. Nun ist es ja so, dass wir auch wenn wir unendlich viele Münzwürfe durchführen wollen, uns stets zunächst für den Ausgang der ersten n davon interessieren, d.h. wie betrachten zunächst jeweils nur endlich viele auf einmal. Das heisst, dass unsere σ-Algebra sicher alle endlichen Produkte von Elementen der σ-Algebren der einfachen Mengen Ω_i enthalten soll. Wir können uns ohne weiteres auf den Standpunkt stellen, dass ausser diesen nur das Unvermeidliche noch dazu genommen werden soll, also dass die σ-Algebra $\mathfrak{B}(\prod_i\Omega_i)$ gerade die von diesen Mengen erzeugte σ-Algebra sein soll.

Definition 3.15. Seien $(\Omega_i,\mathfrak{F}_i)$, $i\in\mathbb{N}$, Messräume, $\widehat{\Omega}\equiv\otimes_{i=1}^{\infty}\Omega_i$ der unendliche Produktraum. Dann definieren wir die Produkt-σ-Algebra, $\widehat{\mathfrak{F}}$, über $\widehat{\Omega}$ als die kleinste σ-Algebra, die alle Teilmengen von $\widehat{\Omega}$ der Form

$$A=\bigotimes_{i\in I}A_i\bigotimes_{j\notin I}\Omega_j \tag{3.24}$$

enthält, wo $A_i\in\mathfrak{F}_i$ und $I=(i_1,\ldots,i_k)\subset\mathbb{N}$ endlich ist. Die Mengen A der Form (3.24) heissen *Zylindermengen*.

Notation: Die Notation in (3.24) bedeutet

$$\bigotimes_{i\in I}A_i\bigotimes_{j\notin I}\Omega_j=B_1\times B_2\times B_3\times\cdots \tag{3.25}$$

wobei $B_i=A_i$ falls $i\in I$ und $B_i=\Omega_i$ falls $i\notin I$.

Definition 3.16. Seien $(\Omega_i, \mathfrak{F}_i, \mathbb{P}_i)$, $i \in \mathbb{N}$, Wahrscheinlichkeitsräume. Dann definieren wir das unendliche Produktmaß, $\widehat{\mathbb{P}} \equiv \bigotimes_i \mathbb{P}_i$, auf $(\widehat{\Omega}, \widehat{\mathfrak{F}})$ dadurch, dass für alle Zylindermengen A der Form (3.24)

$$\widehat{\mathbb{P}}(A) = \prod_{i \in I} \mathbb{P}_i(A_i). \tag{3.26}$$

Die Produkt-σ-Algebra enthält eine äusserst reiche Klasse von Mengen, jedoch ist sie wieder, und zwar selbst in dem Fall, dass Ω endlich ist, kleiner als die Potenzmenge. In der Tat ist sie ihrer Natur nach der Borel'schen σ-Algebra vergleichbar.

Wir können mittels der Konstruktion unendlicher Produkträume nun unendliche Folgen von Zufallsvariablen konstruieren.

Definition 3.17. Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Dann heisst eine Folge $X_n, n \in \mathbb{N}_0$ von Zufallsvariablen $X_n \to S \subset \mathbb{R}$ ein *stochastischer Prozess* (mit diskreter Zeit).

Beachte, dass wir für einen stochastischen Prozess fordern, dass alle X_n auf demselben Wahrscheinlichkeitsraum definiert sein sollen. Dazu muss die σ-Algebra $\mathfrak{F}$ hinreichend gross sein, so dass Mengen $X_n^{-1}(S)$ in $\mathfrak{F}$ enthalten sind. Um dies sicherzustellen, liegt es nahe, für $(\Omega, \mathfrak{F})$ gerade einen unendlichen Produktraum zu wählen, im Fall, dass S endlich (oder abzählbar) ist etwa $(\widehat{S}, \widehat{\mathscr{P}(s)})$.

Definition 3.18. Eine Folge von Zufallsvariablen $X_n, n \in \mathbb{N}_0$ heisst eine Folge unabhängiger Zufallsvariablen genau dann, wenn für jede endliche Teilmenge $I \subset \mathbb{N}_0$, die Familie $(X_i, i \in I)$ eine Familie von unabhängigen Zufallsvariablen ist.

Beachte, dass wir immer nur endlich viele Zufallsvariablen auch Unabhängigkeit testen müssen.

Unendliche Folgen unabhängiger Zufallsvariablen sind die wichtigsten Bausteine der Wahrscheinlichkeitstheorie. Mit ihrer Hilfe können wir insbesondere die Folge der Ergebnisse von (beliebig oft) wiederholten identischen Zufallsexperimenten modellieren, also etwa wiederholte Münzwürfe, Roulettespiele, etc.

Im allgemeinen gilt für stochastische Prozesse, dass das die Wahrscheinlichkeitsverteilung eines solchen Prozesses schon bestimmt ist, wenn wir die Verteilung aller Famillien $X_i, i \in I$, wo I endliche Teilmenge von $\mathbb{N}_0$ sind, kennen. Dies formal zu beweisen würde über den Rahmen dieser Vorlesung allerdings hinausgehen.

3.5 Summen von unabhängigen Zufallsvariablen

Ein weiter Teil der Wahrscheinlichkeitstheorie behandelt die Eigenschaften von Funktionen von *unabhängigen* Zufallsvariablen. Insbesondere deren Summen, aber auch anderer, wie etwa der Maxima. In der Vorlesung werden wir uns im weiteren ebenfalls weitgehend darauf konzentrieren.

3.5.1 Die Irrfahrt

Gerne betrachten wir eine leichte Abwandlung der Summe S_n: wir wählen statt der Bernoulli-Variablen X_i die (manchmal[1]) sogenannten *Rademacher Variablen*, Y_i, mit der Eigenschaft, dass

$$\mathbb{P}(Y_i = 1) = 1 - \mathbb{P}(Y_i = -1) = p, \tag{3.27}$$

wobei der Fall $p = 1/2$ von besonderem Interesse ist. In diesem Fall nennen wir die Folge von Zufallsvariablen

$$S_n = \sum_{i=1}^{n} Y_i \tag{3.28}$$

die *einfache (falls $p = 1/2$ symmetrische) Irrfahrt* auf $\mathbb{Z}$. Beachte dass die Folge S_n, $n \in \mathbb{N}$ selbst wieder eine Zufallsfolge ist, allerdings natürlich keine unabhängigen. S_n ist unser erster *stochastische Prozess* neben unabhängigen Zufallsvariablen.

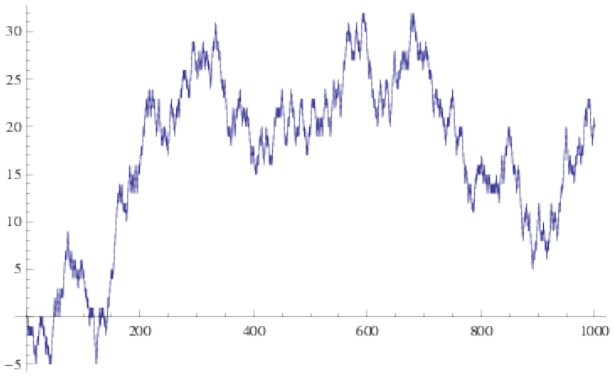


Abb. 3.1 Eine Realisierung der symmetrischen Irrfahrt: Abbildung von $\{(k, S_k), 0 \leq k \leq n = 1000\}$.

Das Interesse an S_n ist in natürlicher Weise dadurch begründet, dass es die Entwicklung des Gewinns (oder Verlustes) eines Spielers darstellt, der wiederholt auf den Ausgang von Münzwürfen wettet und dabei jeweils einen festen Betrag, 1, setzt, und er bei Gewinn 2 Euro erhält.

Unser Formalismus, d.h. die Modellierung von wiederholten Spielen durch unabhängige Zufallsvariablen, erlaubt es uns nun nicht nur einzelne Spiele, sondern ganze Folgen von Spielen zu analysieren. An dieser Stelle ist es vielleicht interessant, zwei Beispiele von Resultaten, die wir damit erhalten können zu betrachten.

Dir Irrfahrt kann auch als einfaches Modell für eine zufällige Bewegung gesehen werden, etwa als die Position einer Person, die zufällig Schritte der Länge 1 nach

[1] Oft werden auch die folgenden Rademacher Variablen als Bernoulli Variablen bezeichnet.

links oder rechts macht. Man kann dieses Modell leicht auch auf Bewegungen in höheren Dimensionen verallgemeinern.

3.6 Das Gesetz der großen Zahlen

In diesem Abschnitt werden wir den vielleicht wichtigsten Satz der Wahrscheinlichkeitstheorie beweisen, das sogenannte *Gesetz der großen Zahlen*. Das Gesetz der großen Zahlen macht für den Fall des Modells von unabhängigen Zufallsvariablen den Zusammenhang zwischen Wahrscheinlichkeit und Frequenz mathematisch rigoros.

Unser Ziel ist es den folgenden Satz zu beweisen.

Satz 3.19 (Schwaches Gesetz der großen Zahlen). *Sei $(\Omega, \mathfrak{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Seien X_i, $i \in \mathbb{N}$, unabhängige, identisch verteilte, Zufallsvariablen mit endlichem Erwartungswert $\mu = \mathbb{E}[X_i]$. Sei $S_n \equiv \sum_{i=1}^n X_i$ und endlicher Varianz $var(X_i) = \sigma^2 < \infty$. Dann konvergiert S_n/n* in Wahrscheinlichkeit *gegen μ, d.h. für jedes $\varepsilon > 0$ gilt,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|n^{-1}S_n - \mu\right| > \varepsilon\right) = 0. \tag{3.29}$$

Beweis. Die erste naheliegende Idee um ein Gesetz der großen Zahlen zu erhalten ist die Verwendung der Chebychev-Ungleichung, siehe Lemma (2.11). Wir können zunächst ohne Beschränkung der Allgemeinheit $\mu = 0$ annehmen. Nun sieht man schnell, dass man mit einer Abschätzung

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n X_i\right| > \varepsilon\right) \leq \frac{\mathbb{E}\left|\sum_{i=1}^n X_i\right|}{n\varepsilon} \leq \frac{\mathbb{E}|X_1|}{\varepsilon} \tag{3.30}$$

nicht weiterkommt, da die rechte Seite nicht von n abhängt.

Die nächste Idee wäre es mit der Markov-Ungleichung, Korollar 2.13 der Ordnung zwei zu versuchen, nämlich

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n X_i\right| > \varepsilon\right) \leq \frac{\mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^2\right]}{n^2\varepsilon^2}. \tag{3.31}$$

Wenn wir hier das Quadrat entwickeln, so sehen wir, das alle gemischten Terme $\mathbb{E}[X_iX_j]$, $i \neq j$ verschwinden, so dass wir die rechte Seite durch $\frac{\mathbb{E}X_1^2}{n\varepsilon^2} = \sigma^2/(n\varepsilon^2)$ abschätzen können. Dies geht für jedes $\varepsilon > 0$ gegen Null, wenn $n \uparrow \infty$, da wir angenommen hatten, dass $\sigma^2 < \infty$ ist. □

Definition 3.20. Wenn für eine Folge von Zufallsvariablen $Z_n, n \in \mathbb{N}$ und eine reelle Zahl a gilt, dass für jedes $\varepsilon > 0$

$$\lim_{n\uparrow\infty} \mathbb{P}(|Z_n - a| > \varepsilon) = 0, \tag{3.32}$$

sagen wir *"Z_n konvergiert in Wahrscheinlichkeit gegen a"*.

Wenn wir uns den Beweis anschauen, stellen wir fest, dass wir gar nicht benutzt haben, dass die X_i unabhängig sind, sondern nur, dass sie unkorreliert sind. Wir haben also sogar folgenden Satz bewiesen:

Satz 3.21. *Seien X_i, $i \in \mathbb{N}$, identische verteilte und paarweise unkorrelierte Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$ mit endlicher Varianz σ^2. Sei $S_n \equiv \sum_{i=1}^{n} X_i$. Dann konvergiert die Folge $n^{-1}S_n$ in Wahrscheinlichkeit gegen $\mathbb{E}[X_1]$.*

Kapitel 4
Markov Prozesse

Un des grands avantages du Calcul des Probabilités est d'apprendre à se défier des premiers aperçus. Comme on reconnaît qu'ils trompent souvent lorsqu'on peut les soumettre cu calcul, on doit en conclure que sur d'autres objets il ne faut s'y livrer qu'avec une circonspection extrème[a].
Pierre Simon de Laplace, Théorie Analytique des Probabilités

[a] Ein großen Nutzen der Wahrscheinlichkeitsrechnung ist es uns zu lehren den ersten Eindrücken zu misstrauen. Da man feststellt, dass diese da wo man sie mit mit Berechnungen konfrontieren kann, oft täuschen, so muss man schliessen, dass man sich ihnen in anderen Gegenständen nur mit der äusserster Umsicht ausliefern darf.

.

Bislang haben wir formal den Begriff eines stochastischen Prozesses als unendliche Folge von Zufallvariablen eingeführt. Als explizites Beispiel hatten wir unendliche Folgen von unabhängigen Zufallsvariablen konstruiert. Aus diesen konnten wir wiederum Summen bilden und die Irrfahrt als Model einer rein zufälligen Bewegung betrachten. In diesem Teil der Vorlesung wollen eine in vielen Anwendungen wichtige Klasse von *stochastischen Prozessen*, die sogenannten *Markov Prozesse* behandeln. Diese sind in vieler Hinsicht die wichtigsten stochastischen Prozesse überhaupt. Der Grund dafür ist, dass sie einerseits so vielseitig sind, dass sehr viele dynamischen Prozesse mit ihrer Hilfe modelliert werden können, andererseits aber mathematisch noch einigermaßen behandelbar sind. Wir werden in dieser Vorlesung natürlich nur einige wenige, einfache Beispiele dieser reichen Klasse betrachten. Markov Prozesse wurden von Andrey Andreyevich Markov (1856-1922) eingeführt.

4.1 Definitionen

Bausteine sind Familien von Zufallsvariable X_t, die für gegebenes t Werte in einem Raum $S \subset \mathbb{R}$, dem sogenannten *Zustandsraum*, annehmen. Wir betrachten hier nur den Fall, dass S eine endliche Menge ist. t nimmt Werte in einer sogenannten *Indexmenge*, I an. Die wichtigsten Beispiele sind $I = \mathbb{N}_0$ und $I = \mathbb{R}_+$, wobei wir uns hier auf den einfacheren Fall $I = \mathbb{N}_0$ einschränken wollen. Wir interpretieren den Index t als *Zeit*, und fassen X_t als Zustand eines Systems zur Zeit t auf. Der stochastische Prozess $\{X_t\}_{t\in I}$ ist als Familie von Zufallsvariablen definiert auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$ zu verstehen. Im Fall, dass $I = \mathbb{N}_0$ können wir natürlich $\Omega = S^{\mathbb{N}_0}$, und $\mathfrak{F} = \mathfrak{B}(S)^{\otimes \mathbb{N}_0}$, also den unendlichen Produktraum, wählen.

Eine wichtige Größe ist die Verteilung des Prozesses X, formal gegeben durch das Maß $P_X \equiv \mathbb{P} \circ X^{-1}$. P_X ist dann ein Wahrscheinlichkeitsmaß auf $(S^{\mathbb{N}_0}, \mathfrak{B}(S)^{\otimes \mathbb{N}_0})$.

Definition 4.1. Ein stochastischer Prozess mit diskreter Zeit und endlichem Zustandsraum S heißt eine *Markovkette*, genau dann, wenn, für alle $n \in \mathbb{N}_0$, und $x_1, \ldots, x_n \in S$, für die

$$\mathbb{P}[X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_0 = x_0] > 0, \tag{4.1}$$

ist, gilt,

$$\begin{aligned} &\mathbb{P}[X_n = x_n | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_0 = x_0] \\ &= \mathbb{P}[X_n = x_n | X_{n-1} = x_{n-1}] \equiv p_{n-1}(x_{n-1}, x_n). \end{aligned} \tag{4.2}$$

Anmerkung. Dieselbe Definition kann auch im Fall abzählbarer Zustandsräume verwandt werden.

Satz 4.2. *Die Wahrscheinlichkeitsverteilung einer Markovkette mit diskreter Zeit ist eindeutig bestimmt durch die Angabe der Anfangsverteilung, $\pi_0(x)$, $x \in S$ und der Übergangswahrscheinlichkeiten $p_n(x,y)$, $n \in \mathbb{N}, x, y \in S$. Umgekehrt gibt es für jedes Wahrscheinlichkeitsmaß π_0 auf $(S, \mathfrak{B}(S))$ und einer Sammlung von Zahlen $p_n(x,y)$ mit der Eigenschaft, dass, für alle $n \in \mathbb{N}$ und alle $x \in S$,*

$$\sum_{y \in S} p_n(x,y) = 1, \tag{4.3}$$

eine Markovkette mir Übergangswahrscheinlichkeiten $p_n(x,y)$ und Anfangsverteilung π_0.

Anmerkung. Mann bezeichnet p_n auch als Übergangsmatrix. Eine Matrix mit der Eigenschaft (4.3) nennt man auch *stochastische Matrix*.

Beweis. Wir zeigen, dass die endlich dimensionalen Verteilungen festgelegt sind. Da wir auf einem endlichen Raum S arbeiten, genügt es offenbar für alle $n \in \mathbb{N}$, und alle $x_i \in S$, $i \leq n$, alle Wahrscheinlichkeiten der Form

$$\mathbb{P}[X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0] \tag{4.4}$$

zu kennen. Nun ist aber wegen der Markoveigenschaft (4.2) und der Definition der bedingten Wahrscheinlichkeit

$$\begin{aligned} &\mathbb{P}[X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0] \\ &= \mathbb{P}[X_n = x_n | X_{n-1} = x_{n-1}] \mathbb{P}[X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0] \\ &= p_{n-1}(x_{n-1}, x_n) \mathbb{P}[X_{n-1} = x_{n-1}, \ldots, X_1 = x_1, X_0 = x_0] \\ &= p_{n-1}(x_{n-1}, x_n) p_{n-2}(x_{n-2}, x_{n-1}) \mathbb{P}[X_{n-2} = x_{n-2}, \ldots, X_1 = x_1, X_0 = x_0] \\ &= p_{n-1}(x_{n-1}, x_n) p_{n-2}(x_{n-2}, x_{n-1}) \ldots p_0(x_0, x_1) \mathbb{P}[X_0 = x_0] \\ &= p_{n-1}(x_{n-1}, x_n) p_{n-2}(x_{n-2}, x_{n-1}) \ldots p_0(x_0, x_1) \pi_0(x_0). \end{aligned} \tag{4.5}$$

□

Aus dieser Formel können wir nun eine Gleichung für die Verteilung von X_n gewinnen, indem wir über alle möglichen Zwischenwerte summieren:

Korollar 4.3. *Es gilt*

$$\begin{aligned}\mathbb{P}[X_n = x_n|] &= \sum_{x_0\in S} \pi_0(x_0) \sum_{x_1\in S} p_0(x_0,x_1) \sum_{x_2\in S} p_1(x_1,x_2)\dots \sum_{x_{n-1}\in S} p_{n-1}(x_{n-1},x_n) \\ &= (\pi_0 p_0 p_1 p_2 \dots p_{n-1})(x_n), \end{aligned} \tag{4.6}$$

wobei hier die Produkte als Matrixprodukte verstanden sind.

Die Matrix $p_0 p_1 \dots p_{n-1} \equiv P_{0,n}$ nennt man auch n-Schritt Übergangswahrscheinlichkeit. Für ihre Element gilt

$$P_{0,n}(x_0,x_n) = \mathbb{P}[X_n = x_n|X_0 = x_0]. \tag{4.7}$$

4.2 Markovketten mit stationären Übergangswahrscheinlichkeiten

Nach diesem allgemeinen Bemerkungen wollen wir uns zunächst nur mit dem einfachsten, aber bereits interessanten Spezialfall befassen, in dem

(i) der Zustandsraum, S, eine endlich Menge ist, also $S = \{1,\dots,d\}$, $d \in \mathbb{N}$, und
(ii) die Übergangswahrscheinlichkeiten $p_{n-1}(x,y)$ nicht von n abhängen.

Man nennst solche Markovketten *zeitlich homogene* Markovketten oder Markovketten mit *stationären Übergangswahrscheinlichkeiten.*

Beispiel. Ein sehr einfaches Beispiel für eine stationäre Markovkette ist folgendes (recht schlechtes) Klimamodell. Wir wollen dabei das Wetter auf die Grundfrage "Regen oder Sonnenschein" reduzieren. Das Wetter am Tag n soll also durch eine Zufallsvariable X_n die die Werte 0 (=Regen) und 1 (=Sonne) annimmt beschrieben werden. Versucht man diese durch unabhängige Zufallsvariablen zu beschreiben, stellt man fest, dass dies mit den Beobachtungen nicht kompatibel ist: längere Perioden mit konstantem Regen oder Sonnenschein treten in Wirklichkeit häufiger auf als das Modell vorhersagt. Man überlegt sich, dass es sinnvoll scheint, die Prognose des Wetters für morgen davon abhängig zu machen, wie das Wetter heute ist (aber nicht davon wie es gestern und vorgestern war). Dies führt auf die Beschreibung durch eine Markovkette mit den Zuständen 0 und 1, und Übergangswahrscheinlichkeiten

$$\begin{aligned} p(0,1) &= p_{0,1}, \quad p(0,0) = p_{0,0} = 1 - p_{0,1}, \\ p(1,0) &= p_{1,0}, \quad p(1,1) = p_{1,1} = 1 - p_{1,0}. \end{aligned} \tag{4.8}$$

Zusammen mit der Anfangsverteilung $\pi(0) = p_0, \pi(1) = p_1 = 1 - p_0$ legt dies eine Markovkette fest. Wie sehen, dass wir nun 3 freie Parameter zur Verfügung haben, mit denen wir im Zweifel das Wetter besser fitten können.

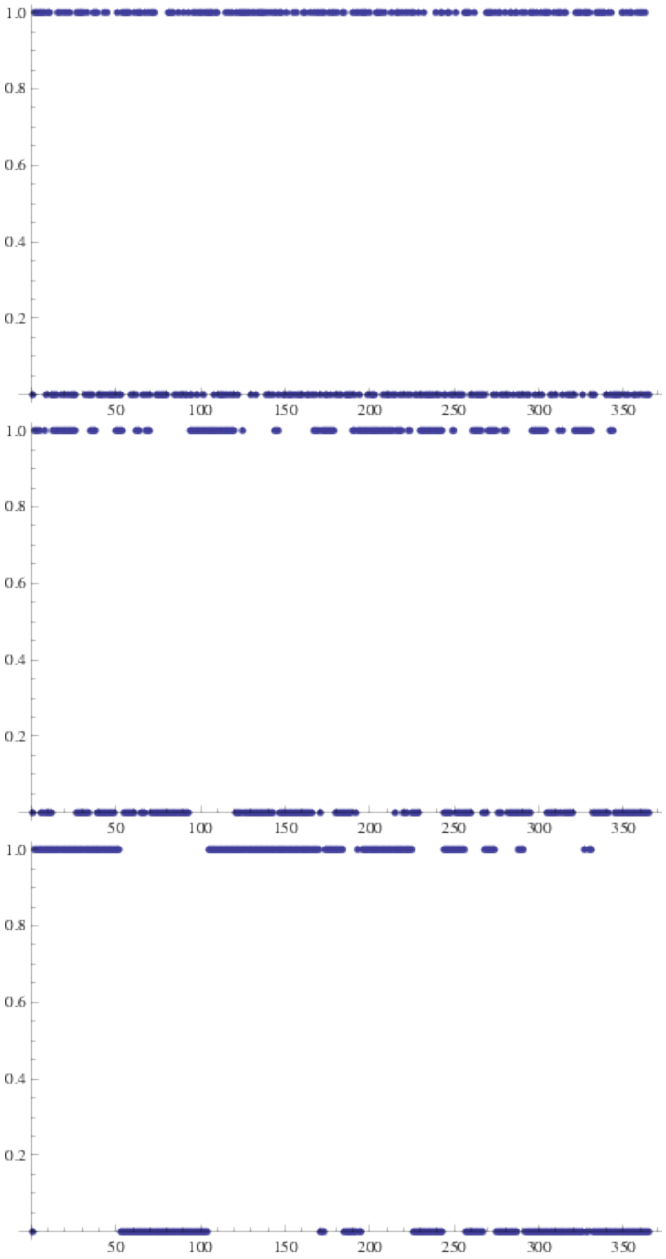


Abb. 4.1 Ein Jahresverlauf des "Wetters" in unserem Modell mit $p_{01} = p_{10} = 0.5$, 0.15, und 0.05.

Wir sehen, dass die Übergangswahrscheinlichkeiten einer stationären Markovkette eine $d \times d$ Matrix, P, bilden. Diese Matrix nennt man auch die Übergangsmatrix der Markovkette. Zusammen mit dem *Vektor* der Anfangsverteilung, π_0, legt diese die Wahrscheinlichkeitsverteilung einer Markovkette vollständig fest, d.h. Wahrscheinlichkeiten beliebiger Ereignisse lassen sich durch diese Objekte ausdrücken. Durch diese Beobachtung begründet sich ein enger Zusammenhang zwischen Markovketten und der linearen Algebra.

Übergangsmatrizen sind freilich keine beliebigen Matrizen, sondern sie haben eine Reihe von wichtigen Eigenschaften.

Lemma 4.4. *Sei P die Übergangsmatrix einer stationären Markovkette mit Zustandsraum $S = \{1, \dots, d\}$. Seien p_{ij} die Elemente von P. Dann gilt:*

(i) Für alle $i, j \in S$ gilt $1 \geq p_{ij} \geq 0$.
(ii) Für alle $i \in S$ gilt $\sum_{j \in S} p_{ij} = 1$.

Umgekehrt gilt: Jede Matrix die (i) und (ii) erfüllt, ist die Übergangsmatrix einer Markovkette.

Beweis. Die beiden ersten Eigenschaften sind offensichtlich, weil ja für jedes i, $p_{i,\cdot} = \mathbb{P}[X_{n+1} = \cdot | X_n = i]$ eine Wahrscheinlichkeitsverteilung auf S ist. Der Umkehrschluss folgt aus Satz 4.2. □

Matrizen die die Eigenschaften (i) und (ii) aus Lemma 4.4 erfüllen heissen *stochastische Matrizen*. Wir wollen uns die Übergangsmatrizen für einige Beispiele von Markovketten ansehen.

- **Unabhängige Zufallsvariablen.** Schon eine Folge unabhängiger, identisch verteilter Zufallsvariablen ist eine Markovkette. Hier ist

$$p_{ij} = \mathbb{P}[X_n = j | X_{n-1} = i] = \mathbb{P}[X_0 = j] = \pi_0(j),$$

  d.h. alle Zeilen der Matrix P sind identisch gleich dem Vektor der die Anfangsverteilung der Markovkette angibt.
- **Irrfahrt mit Rand.** Auch Summen unabhängiger Zufallsvariablen sind Markovketten. Wir betrachten den Fall, dass X_i unabhängige Rademachervariablen mit Parameter p sind, also eine Irrfahrt. In der Tat ist

$$\mathbb{P}[S_n = j | S_{n-i} = i] = \begin{cases} p, & \text{falls } j = i+1 \\ 1-p, & \text{falls } j = i-1 \\ 0, & \text{sonst} \end{cases} \tag{4.9}$$

  allerdings ist in diesem Fall der Zustandsraum abzählbar unendlich, nämlich $\mathbb{Z}$. Wir können eine Variante betrachten, in dem die Irrfahrt angehalten wird, wenn sie auf den Rand des endlichen Gebiets $[-L, L]$ trifft. Dazu modifizieren wir die Überangswahrscheinlichkeiten aus (4.9) für den Fall $i = \pm L$, so dass

$$\mathbb{P}[S_n = j | S_{n-i} = \pm L] = \begin{cases} 1, & \text{falls } i = \pm L \\ 0, & \text{sonst} \end{cases} \tag{4.10}$$

Die Übergangsmatrix hat dann folgende Gestalt:

$$P = \begin{pmatrix} 1 & 0 & 0 & \dots & \dots & \dots & 0 \\ 1-p & 0 & p & 0 & \dots & \dots & 0 \\ 0 & 1-p & 0 & p & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 1-p & 0 & p & 0 \\ 0 & \dots & \dots & 0 & 1-p & 0 & p \\ 0 & \dots & \dots & \dots & 0 & 0 & 1 \end{pmatrix}$$

- **Unser Wettermodell (4.8)**. Hier ist

$$P = \begin{pmatrix} 1-p_{0,1} & p_{0,1} \\ p_{1,0} & 1-p_{1,0} \end{pmatrix}$$

Das der Zusammenhang zwischen Markovketten und Matrizen nicht nur oberflächlich ist, zeigt sich daran, dass, wie wir schon in (4.6)gesehen haben, in der Berechnung verschiedener Wahrscheinlichkeiten tatsächlich Matrixoperationen auftauchen. So ist

$$\mathbb{P}[X_n = j | X_0 = i] \equiv P_n(i,j) = \sum_{i_1, i_2, \dots, i_{n-1}} p_{ii_1} p_{i_1 i_2} \dots p_{i_{n-2} i_{n-1}} p_{i_{n-1} j} = (P^n)_{ij}. \tag{4.11}$$

Es folgt, dass

$$\pi_n(j) \equiv \mathbb{P}[X_n = j] = \sum_{i \in S} \pi_0(i) P_n(ij) = (\pi_0 P^n)_j. \tag{4.12}$$

Wir sehen also, dass die Verteilung der Markovkette zur Zeit n durch die Wirkung der Matrix P^n von links auf die Anfangsverteilung gegeben ist.

4.3 Invariante Verteilungen

Grubdsätzlich ist eine zentrale Frage, wie sich die Verteilung π_n einer Markovkette verhält, wenn n nach unendlich geht. Eine der ersten Fragen, die man sich stellen wird, ist, ob Verteilungen, π_0, gibt, die unter der Wirkung der Markovkette *invariant* sind.

Definition 4.5. Sei X eine Markovkette mit diskreter Zeit, endlichem Zustandsraum S und stationären Übergangswahrscheinlichkeiten P. Dann heisst ein Wahrscheinlichkeitsmaß, π_0, *invariante Verteilung*, wenn für alle $n \in \mathbb{N}$ und alle $j \in S$,

$$\pi_n(j) = \pi_0(j), \tag{4.13}$$

gilt.

Offensichtlich ist wegen der Gleichung (4.12), die Frage nach invarianten Verteilungen äquivalent zur Frage nach links-Eigenwerten der Matrix P:

Lemma 4.6. *Sei P eine stochastische Matrix. Dann ist π_0 genau dann eine invariante Verteilung für eine stationäre Markovkette mit Übergangsmatrix P, wenn π_0 ein links-Eigenvektor von P zum Eigenwert 1 ist, mit $\pi_0(i) \geq 0$ und $\sum_{i\in S} \pi_0(i) = 1$.*

Beweis. Wir kombinieren (4.13) mit (4.12) und erhalten, dass π_0 invariant ist, wenn

$$\pi_0(i) = (\pi_0 P)_i . \tag{4.14}$$

Wenn andererseits ein Vektor mit positiven Komponenten deren Summe gleich eins ist die Gleichung (4.14) erfüllt, so liefert er eine invariante Anfangsverteilung. □

Satz 4.7. *Jede stationäre Markovkette mit endlichem Zustandsraum besitzt mindestens eine invariante Verteilung.*

Beweis. Wir müssen zeigen, dass jede stochastische Matrix einen Eigenwert eins besitzt, und dass der zugehörige links-Eigenvektor nur nicht-negative Einträge hat, nach Normierung also eine Wahrscheinlichkeitsverteilung liefert. Nun folgt aus der Definition einer stochastischen Matrix P sofort, dass der Vektor $(1,1,\ldots,1)$ rechts-Eigenvektor mit Eigenwert $+1$ ist. Dieser Eigenwert ist zugleich der größte Eigenwert von P. Setze $\|u\|_1 \equiv \sum_i |u_i|_1$. Wir zeigen, dass

$$\|uP\|_1 = \sum_i \left| \sum_j u_j p_{ji} \right| \leq \sum_i \sum_j |u_j| p_{ji} = \sum_j |u_j| = \|u\|_1 . \tag{4.15}$$

Falls u ein Eigenvektor mir Eigenwert λ ist, so folgt daraus

$$\lambda \|u\|_1 \leq \|u\|_1 , \tag{4.16}$$

also $|\lambda| \leq 1$. Wir müssen noch zeigen, dass alle Komponenten des rechts-Eigenvektors zum Eigenwert 1 das gleiche Vorzeichen haben. Dies folgt aber aus dem Satz von Perron-Frobenius, der besagt, dass für Matrizen mit nicht-negativen Einträgen stets ein Eigenvektor zum grössten Eigenwert existiert, der nur nicht-negative Einträge hat. □

4.3.1 Markovketten und Graphen. Klassifizierung der Zustände

Es erweist sich als instruktiv mit einer Übergangsmatrix einen gerichteten Graphen auf dem Zustandsraum S zu verbinden. Wir fassen die Menge S als Knotenmenge eines (gerichteten) Graphen, $(S,\mathscr{E})$ auf. Wir sagen, dass $\mathscr{E}$ die *Kante*, (i,j), $i \in S$, $j \in S$ enthält, $(i,j) \in \mathscr{E}$, wenn $p_{ij} > 0$. Graphisch stellen wir dies durch einen Pfeil dar.

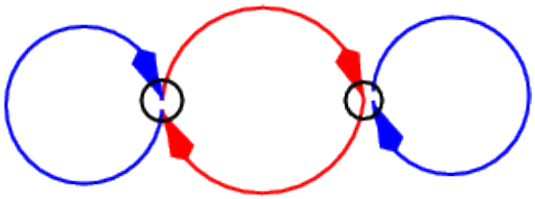


Abb. 4.2 Der Graph der Markovkette unseres Wettermodells

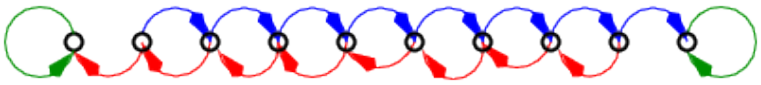


Abb. 4.3 Der Graph der am Rand gestoppten Irrfahrt

Definition 4.8. Ein *Pfad* γ in einem gerichteten Graphen $(S,\mathscr{E})$ ist eine Folge $\gamma = (e_1, e_2, \ldots, e_k)$ von Kanten $e_\ell \in \mathscr{E}$, so dass für jedes $\ell = 1, \ldots, k-1$ gilt, dass der Endpunkt von e_ℓ der Anfangspunkt von $e_{\ell+1}$ ist. γ verbindet i mit j falls der Anfangspunkt von e_1 i und der Endpunkt von e_k j ist.

Definition 4.9. Zwei Knoten, $i, j \in S$ einem gerichteten Graphen *kommunizieren*, wenn Pfade gibt, die i mit j verbinden und solche, die j mit i verbinden. Wir sagen auch, dass jeder Zustand mit sich selbst kommuniziert.

Man kann leicht nachprüfen, dass die Relation "kommunizieren" eine Äquivalenzrelation ist. Nun definiert eine Äquivalenzrelation eine Zerlegung der Menge S in Äquivalenzklassen. Wir bezeichnen die Äquivalenzklassen kommunizierender Zustände als *kommunizierende Klassen* oder einfach als *Klassen*.

Definition 4.10. Eine Markovkette heißt *irreduzibel* genau dann wenn der Zustandsraum aus einer einzigen Klasse besteht.

Anmerkung. Beachte, dass eine Markovkette deren Graph nicht zusammenhängend ist, auch nicht irreduzibel ist. Wenn der Graph einer Markovkette zusammenhängend ist, muss diese aber noch lange nicht irreduzibel sein.

Lemma 4.11. *Eine Markovkette ist genau dann irreduzibel, wenn es für jedes Paar,* $(i, j) \in S \times S$*, ein* $k \in \mathbb{N}_0$ *gibt, so dass* $\left(P^k\right)_{i,j} > 0$*.*

Beweis. Es gilt

$$\begin{aligned}\left(P^k\right)_{ij} &= \sum_{i_1, i_2, \ldots, i_{k-1}} p_{i i_1} p_{i_1 i_2} \cdots p_{i_{k-1} j} \\ &= \sum_{\substack{\gamma: i \to j \\ |\gamma| = k}} p_{e_1} p_{e_2} \cdots p_{e_k} \end{aligned} \tag{4.17}$$

Die rechte Seite ist offenbar genau dann positiv, wenn es einen solchen Weg gibt. Daraus folgt das Lemma direkt. □

4.3.2 Invariante Verteilungen für irreduzible Markovketten

Wir definieren die *ersten Eintrittszeiten* in Untermengen. Für $D \subset S$, so definieren wir

$$\tau_D \equiv \inf\{n > 0 | X_n \in D\}. \tag{4.18}$$

Beachte, dass

$$\{\tau_D = n\} = \{\forall_{k<n}, X_k \notin D\} \cap \{X_n \in D\}. \tag{4.19}$$

Das heisst, wir wissen, ob das Ereignis $\{\tau_D = n\}$ eintritt, wenn wir unseren Prozess bis zur Zeit n beobachtet haben. Zufallsvariablen mit dieser Eigenschaft nennt man auch *Stoppzeiten*.

Lemma 4.12. *Sei X eine irreduzible Markovkette mit endlichem Zustandsraum S. Sei $\tau_\ell \equiv \inf\{n > 0 | X_n = \ell\}$. Für $j, \ell \in S$, setze*

$$\mu(j) = \frac{\mathbb{E}_\ell \left[\sum_{t=1}^{\tau_\ell} \mathbb{1}_{X_t = j}\right]}{\mathbb{E}_\ell \tau_\ell} \tag{4.20}$$

Dann ist μ unahängig von ℓ und ist die eindeutige invariante Verteilung der Markovkette.

Beweis. Wir zeigen zunächst, dass $\mathbb{E}_\ell[\tau_\ell] < \infty$, und somit der Ausdruck auf der rechten Seite von (4.20) Sinn macht. Die Grundidee des Beweises ist folgende. Wir wissen, dass es für jedes $j \in S$ ein $k_j < \infty$ gibt, so dass $P_{j\ell}^{k_j} > 0$. Ausserdem ist $k^* \equiv \max_{j \in S} k_j$ endlich, und es gilt, dass für alle $j \in S$, $\exists_{k_j \leq k^*}$ so dass $P_{j\ell}^{k_j} > 0$. Das heisst, egal in welchem $j \in S$ die Markovkette startet ist die Wahrscheinlichkeit, dass sie den Zustand ℓ vor der Zeit k^* trifft größer als die Wahrscheinlichkeit, zur Zeit k_j in in ℓ zu sein, und diese ist strikt positiv. In Formeln

$$\max_{j \in S} \mathbb{P}_j(X_t \neq \ell, \forall_{t \leq k*}) \leq 1 - \min_{j \in S} \mathbb{P}_j\left(X_{k_j} = \ell\right) = 1 - \min_{j \in S} P_{j,\ell}^{k_j} < 1 - c, \tag{4.21}$$

für ein $c > 0$. Nun können wir dieses Argument iterieren: Zum Zeitpunkt $k*$ muss die Kette ja in irgendeinem Zustand sein. Wir betrachten nun das Ereignis, dass die Kette auch im Zeitintervall $[k^* + 1, 2k^*]$ nicht in ℓ ist. Dessen Wahrscheinlichkeit hängt von dem was bis $k*$ passiert ist aber nur über den Wert von X_{k^*} ab, und ist wieder beschränkt durch

$$\max_{j \in S} \mathbb{P}_j(X_t \neq \ell, \forall_{t \leq k*}) < 1 - c. \tag{4.22}$$

Es folgt daraus, dass

$$\mathbb{P}_\ell[\tau_\ell > 2k^*] = \mathbb{P}_\ell(X_t \neq \ell, \forall_{t \leq 2k*}) \leq (1-c)^2, \tag{4.23}$$

und durch weiteres iterieren

$$\begin{aligned}\mathbb{P}_\ell[\tau_\ell > t] &\le \mathbb{P}_\ell[X_t \neq \ell, \forall k \le t] \le \mathbb{P}_\ell(X_t \neq \ell, \forall t \le [t/k^*]k^*) \\ &\le \prod_{n=1}^{[t/k^*]} \max_{j\in S} \mathbb{P}_j(X_t \neq \ell, \forall_{t\le k*}) \le (1-c)^{[t/k^*]} \le \left[(1-c)^{1/k^*}\right]^{t-k^*} \quad (4.24)\end{aligned}$$

Damit ist dann natürlich $\mathbb{E}_\ell[\tau_\ell] = \sum_{t\ge 0} \mathbb{P}_\ell[\tau_\ell > t] \le (1-c)^{-1} \frac{1}{1-(1-c)^{1/k^*}} < \infty$.

Wir definieren $\nu_\ell(j) = \mathbb{E}_\ell\left[\sum_{t=1}^{\tau_\ell} \mathbb{1}_{X_t=j}\right]$. Wenn wir zeigen, dass $\nu_\ell(j)$ die Invarianzeigenschaft erfüllt, so tut dies auch μ, und nach Konstruktion ist μ eine Wahrscheinlichkeitsverteilung. Wir schreiben zunächst $1 = \sum_{m\in S} \mathbb{1}_{X_{t-1}=m}$, und

$$\begin{aligned}\nu_\ell(j) = \mathbb{E}_\ell\left[\sum_{t=1}^{\infty} \mathbb{1}_{X_t=j}\mathbb{1}_{t\le\tau_\ell}\right] &= \sum_{t=1}^{\infty} \mathbb{P}_\ell(X_t = j, t \le \tau_\ell) \\ &= \sum_{m\in S}\sum_{t=1}^{\infty} \mathbb{P}_\ell(X_{t-1} = m, X_t = j, t \le \tau_\ell).\end{aligned}$$

Nun ist das Ereignis $\{t \le \tau_\ell\} = \{\tau_\ell \le t-1\}^c$ nur davon abhängig, was bis zur Zeit $t-1$ passiert ist. Daher können wir die Markov-Eigenschaft zur Zeit $t-1$ anwenden und erhalten

$$\begin{aligned}\mathbb{P}_\ell(X_{t-1} = m, X_t = j, t \le \tau_\ell) &= \mathbb{P}_\ell(X_{t-1} = m, t \le \tau_\ell)\,\mathbb{P}_m(X_1 = j) \\ &= \mathbb{P}_\ell(X_{t-1} = m, t \le \tau_\ell)\, p_{m,j}. \quad (4.25)\end{aligned}$$

Damit ist aber

$$\nu_\ell(j) = \sum_{m\in S} \mathbb{E}_\ell\left[\sum_{t=1}^{\infty} \mathbb{1}_{X_{t-1}=m}\mathbb{1}_{t\le\tau_\ell}\right] p_{m,j} = \sum_{m\in S} \mathbb{E}_\ell\left[\sum_{t=1}^{\tau_\ell} \mathbb{1}_{X_{t-1}=m}\right] p_{m,j}. \quad (4.26)$$

Andererseits haben wir

$$\sum_{t=1}^{\tau_\ell} \mathbb{1}_{X_{t-1}=m} = \mathbb{1}_{X_0=m} + \sum_{t=1}^{\tau_\ell} \mathbb{1}_{X_t=m} - \mathbb{1}_{X_{\tau_\ell}=m} = \sum_{t=1}^{\tau_\ell} \mathbb{1}_{X_t=m} \quad (4.27)$$

weil $X_0 = X_{\tau_\ell} = \ell$. Somit ist aber

$$\nu_\ell(j) = \sum_{m\in S} \mathbb{E}_\ell\left[\sum_{t=1}^{\tau_\ell} \mathbb{1}_{X_t=m}\right] p_{m,j} = \sum_{m\in S} \nu_\ell(m) p_{m,j}. \quad (4.28)$$

Dies ist aber gerade die Gleichung für die invariante Verteilung. Daher ist $\nu_\ell(j)/\sum_{i\in S}\nu_\ell(i)$ eine invariante Wahrscheinlichkeitsverteilung. Nun ist aber

$$\sum_{i\in S} \nu_\ell(i) = \sum_{i\in S} \mathbb{E}_\ell\left[\sum_{t=1}^{\tau_\ell} \mathbb{1}_{X_t=i}\right] = \mathbb{E}_\ell\left[\sum_{t=1}^{\tau_\ell} \mathbb{1}_{X_t\in S}\right] = \mathbb{E}_\ell[\tau_\ell] \quad (4.29)$$

woraus folge, dass (4.20) eine invariante Wahrscheinlichkeit ist.

Es bleibt noch die Eindeutigkeit nachzuprüfen

Zunächst zeigen wir, dass wenn μ eine invariante Verteilung einer irreduziblen Markovkette ist, dann muss $\mu(j) > 0$ für alle $j \in S$. Klarerweise muss es ein $j \in S$ geben, so dass $\mu(j) > 0$. Dann aber gibt es für jedes $i \in S$ ein endliches t, so dass $P^t_{ji} > 0$ und $\mu(i) \geq \mu(j)P^t_{ji} > 0$.

Wir zeigen nun, dass ν_ℓ die einzige Lösung der Invarianzgleichung ist, für die $\nu_\ell(\ell) = 1$. Um dies zu tun, zeigen wir zunächst, dass für jede Lösung ν mit $\nu(\ell) = 1$ für alle j gilt, dass $\nu(j) \geq \nu_\ell(j)$. Dann ist auch $\rho_\ell \equiv \nu - \nu_\ell$ eine nicht-negative Lösung der Invarianzgleichung. Falls diese nicht strikt Null ist, liefert sie nach Normierung also eine invariante Verteilung, für die aber $\rho(\ell) = 0$ ist. Wir wissen aber, dass dies nicht möglich ist. Also ist $\nu = \nu_\ell$.

Nehmen wir nun ν mit $\nu(\ell) = 1$. Die Invarianzgleichung liefert

$$\nu(i) = \sum_{j \neq \ell} p(j,i)\nu(j) + p(\ell,i), \tag{4.30}$$

da ja $\nu(\ell) = 1$. Wir schreiben $p(\ell,i)$ als

$$p(\ell,i) = \mathbb{E}_\ell \left(\mathbb{1}_{\tau_\ell \geq 1} \mathbb{1}_{X_1 = i} \right).$$

Wir iterieren nun (4.30). Dies gibt

$$\begin{aligned} \nu(i) &= \sum_{j_1, j_2 \neq \ell} p(j_2, j_1)p(j_1, i)\nu(j_2) + \sum_{j_1 \neq \ell} p(\ell, j_1)p(j_1, i) + \mathbb{E}_\ell \left(\mathbb{1}_{\tau_\ell \geq 1} \mathbb{1}_{X_1 = i} \right) \\ &= \sum_{j_1, j_2 \neq \ell} p(j_2, j_1)p(j_1, i)\nu(j_2) + \mathbb{E}_\ell \left(\sum_{s=1}^{2 \wedge \tau_\ell} \mathbb{1}_{X_s = i} \right). \end{aligned} \tag{4.31}$$

Iterieren wir weiter erhalten wir schliesslich für jedes $n \in \mathbb{N}$

$$\begin{aligned} \nu(i) &= \sum_{j_1, j_2, \ldots j_n \neq \ell} p(j_n, j_{n-1}) \ldots p(j_2, j_1)p(j_1, i)\nu(j_n) + \mathbb{E}_\ell \left(\sum_{s=1}^{n \wedge \tau_\ell} \mathbb{1}_{X_s = i} \right) \\ &\geq \mathbb{E}_\ell \left(\sum_{s=1}^{n \wedge \tau_\ell} \mathbb{1}_{X_s = i} \right). \end{aligned} \tag{4.32}$$

Dies zeigt, dass $\nu(i) \geq \mathbb{E}_\ell \left(\sum_{s=1}^{\tau_\ell} \mathbb{1}_{X_s = i} \right) = \nu_\ell(i)$, also $\nu = \nu_\ell$. Der Beweis ist erbracht. □

Korollar 4.13. *Für eine irreduzible Markovkette mit endlichem Zustandsraum gilt*

$$\mu(j) = \frac{1}{\mathbb{E}_j[\tau_j]}. \tag{4.33}$$

Beweis. Formel (4.20) gilt für jede Wahl von ℓ. Indem wir $\ell = j$ wählen und benutzen, dass

$$\nu_j(j) = \mathbb{E}_j\left[\sum_{t=1}^{\tau_j} \mathbb{1}_{X_t=j}\right] = 1, \tag{4.34}$$

ist, weil aus der Definition von τ_j folgt $\mathbb{1}_{X_t=j} = \delta_{\tau_j,t}$ für $t = 1,\ldots,\tau_j$, erhalten wir (4.33). □

4.3.3 Der Ergodensatz

Wir sind nun in der Lage einen Ergodensatzes für irreduzible Markovketten zu formulieren, die in gewisser Weise das Analogon des Gesetzes der grossen Zahlen für Markovketten ist.

Satz 4.14 (Ergodensatz). *Sei X eine irreduzible Markovkette mit endlichem Zustandsraum S und invarianter Verteilung μ. Sei $f : S \to \mathbb{R}$ eine beschränkte messbare Funktion. Dann gilt*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \mathbb{E}\left[f(X)\right], \tag{4.35}$$

in Wahrscheinlichkeit, wo X eine Zufallsvariable mit Verteilung μ ist.

Anmerkung. Die Voraussetzungen an f sind angesichts der Endlichkeit des Zustandsraums natürlich trivial.

Beweis. Es genügt offenbar den Satz für Indikatorfunktionen $f = \mathbb{1}_i$, $i \in S$, zu beweisen. Sei nun T_ℓ eine Folge von Stoppzeiten definiert durch

$$\begin{aligned} T_0 &\equiv \inf\{k \geq 0 : X_k = i\}, \\ T_\ell &\equiv \inf\{k > T_{\ell-1} : X_k = i\}. \end{aligned} \tag{4.36}$$

Mit anderen Worten, die Zeiten T_ℓ sind genau die Zeiten, an denen X den Zustand i besucht. Offenbar ist dann

$$\sum_{k=1}^{n} f(X_k) = \sum_{k=1}^{n} \mathbb{1}_{X_k=i} = \max\{\ell : T_\ell \leq n\}. \tag{4.37}$$

Nun machen wir folgende wichtige Beobachtung: Setze $\sigma_\ell = T_\ell - T_{\ell-1}$. Dann sind für $\ell \geq 1$ die σ_ℓ unabhängige, identisch verteilte Zufallsvariablen. Das folgt aus der Markoveigenschaft, indem wir nachweisen, dass für beliebige integrierbare Funktionen, $g, h : \mathbb{N} \to \mathbb{R}$,

$$\mathbb{E}_\ell[g(\sigma_i)h(\sigma_j)] = \mathbb{E}_\ell[g(\sigma_i)]\,\mathbb{E}_\ell[h(\sigma_j)] \quad \text{für alle } i \neq j. \tag{4.38}$$

(Übung!). Es gilt $\mathbb{P}[\sigma_\ell \leq k] = \mathbb{P}[T_1 \leq k | X_0 = i] = \mathbb{P}_i[\tau_i \leq k]$. Wir wissen schon, dass $\mathbb{E}[\sigma_\ell] = \mathbb{E}_i[T_1] = \mathbb{E}_i[\tau_i] < \infty$. Daher gilt nach dem Gesetz der grossen Zahlen,

$$\lim_{n\to\infty} \frac{T_n}{n} = \mathbb{E}\left[T_1 | X_0 = i\right] = \mathbb{E}_i[\tau_i], \quad \text{in Wahrscheinlichkeit.} \tag{4.39}$$

Ausserdem ist für jedes ℓ,

$$\lim_{n\to\infty} \frac{\sigma_\ell}{n} = 0, \quad \text{in Wahrscheinlichkeit,} \tag{4.40}$$

Dann ist leicht einzusehen (Übung!), dass daraus folgt, dass

$$\lim_{n\to\infty} \frac{1}{n} \max\{\ell : T_\ell \le n\} = \frac{1}{\mathbb{E}_i[\tau_i]} = \mu(i), \quad \text{in Wahrscheinlichkeit,} \tag{4.41}$$

□

Anmerkung. Aus dem Ergodensatz folgt auch, indem wir auf der linken Seite die Erwartung nehmen, dass für irreduzible Markovketten gilt, dass

$$\lim_{n\uparrow\infty} \frac{1}{n} \sum_{k=1}^{n} \pi_0 P^k = \lim_{n\uparrow\infty} \frac{1}{n} \sum_{k=1}^{n} \pi_n = \mu, \tag{4.42}$$

das heisst, die Verteilung einer irreduziblem Markovkette konvergiert im Cesaro-Mittel stets gegen die invariante Verteilung konvergiert.

4.3.4 Wesentliche und unwesentliche Klassen.

Besitzt eine Markovkette mehrere Klassen, so kann man diese in zwei Gruppen einteilen: solche, aus denen man in eine andere Klasse austreten kann (aber nicht wieder zurück kann), und solche aus denen man nicht in eine andere Klasse eintreten kann (in die man aber ggf. aus anderen eintreten kann). Erstere heissen "unwesentlich", letztere "wesentlich".

Anmerkung. Im Fall endlichen Zustandsraums können wir wesentliche Klassen auch als rekurrent, unwesentliche als transient bezeichnen. Im Fall von Markovketten mit unendlichem Zustandsraum sind diese Begriffe aber zu unterscheiden.

Satz 4.15. *Sei X eine Markovkette mit Zustandsraum S. S zerfalle in die wesentlichen Klassen $C_1,\dots,C_\ell$ und die unwesentlichen Klassen $D_1,\dots,D_k$. Dann gibt es ℓ invariante Verteilungen $\mu_1,\dots,\mu_\ell$ mit Träger auf den wesentlichen Klassen $C_1,\dots,C_\ell$, und alle invarianten Verteilungen μ sind von der Form*

$$\mu = \sum_{i=1}^{\ell} \alpha_i \mu_i, \tag{4.43}$$

mit $\alpha_i \ge 0$ und $\sum_i \alpha_i = 1$.

Beweis. Es ist klar, dass es für jede wesentliche aperiodische Klasse genau eine invariante Verteilung gibt. Sei nämlich C eine wesentliche Klasse. Wenn die Anfangsverteilung π_0 so gewählt ist, dass für alle $i \notin C$, $\pi_0(i) = 0$, dann ist für alle Zeiten

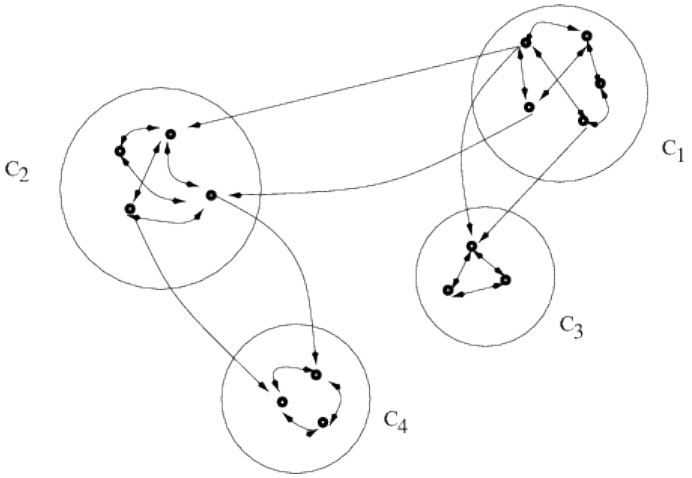


Abb. 4.4 Der Graph einer Markovkette mit vier Klassen C_1, C_2, C_3, C_4. Die Klassen C_1 und C_2 sind transient, C_3 und C_4 sind rekurrent.

für solche i, $\pi_t(i) = 0$. Die Matrix P eingeschränkt auf den von den Zuständen $j \in C$ aufgespannten Unterraum ist aber die Übergangsmatrix einer irreduziblen aperiodischen Markovkette mit Zustandsraum C. Also gibt es eine invariante Verteilung μ_C die C Maß eins gibt. Dies gilt für jede wesentliche Klasse separat.

Ebenso kann man sich leicht überzeugen, dass für jede invariante Verteilung μ und jede unwesentliche Klasse D gilt, dass $\mu(D) = \sum_{j\in D} \mu(j) = 0$. Sei nämlich $\mu(D) > 0$. Wir betrachten dazu zuächst solche unwesentliche Klassen, in die man aus keiner anderen Klasse eintreten kann (wegen der Endlichkeit des Zustandsraumes muss es mindestens eine solche geben). Sei D eine solche Klasse. Da μ invariant ist, muss $(\mu P)(D) = \mu(D)$ gelten. Nun ist aber

$$(\mu P)(D) = \sum_{j\in D}\sum_{i\in S} \mu(i)p_{i,j} = \sum_{j\in D}\sum_{i\in D} \mu(i)p_{i,j} + 0 \tag{4.44}$$

da ja für alle $j \in D$ und $i \notin D$, $p_{i,j} = 0$, gemäß unserer Annahme. Daher ist

$$(\mu P)(D) = \sum_{i\in D} \mu(i) \sum_{j\in D} p_{i,j} = \sum_{i\in D} \mu(i) - \sum_{i\in D} \mu(i) \sum_{j\notin D} p_{i,j} \leq \mu(D). \tag{4.45}$$

Dabei kann Gleichheit nur dann gelten, wenn für alle $i \in D$ für die es $j \in D^c$ gibt mit $p_{i,j} > 0$, $\mu(i) = 0$. Andererseits gilt für diese j dann

$$0 = \mu(i) = \sum_{j\in D} \mu(j)p_{j,i}, \tag{4.46}$$

weswegen $\mu(j) = 0$ auch für alle Zustände in D gilt die mit i verbunden sind; indem wir dieses Argument iterieren, und benutzen, dass D eine kommunizierende Klasse ist, folgt $\mu(j) = 0$ für alle $j \in D$.

Nachdem wir wissen, dass $\mu(D) = 0$ für alle unwesentlichen Klassen, in die man nicht eintritt, kann man nun diese D aus dem Zustandsraum aussondern, und die Restriktion der Markovkette auf den verbleibenden Zustandsraum $S\backslash D$ betrachten. Wenn dieser noch unwesentliche Klassen enthält, so gibt es mindestens eine, in die man nicht mehr eintreten kann, und man sieht, dass auf diesen die invariante Verteilung auch Null ist. Durch Iteration folgt, dass μ auf allen unwesentlichen Klassen verschwindet.

Nutzt man nun diese Information, so verbleibben als Gleichungssystem für die invarianten Verteilungen nur noch entkpoppelte Systeme für jede der verbleibenden wesentlichen irreduziblem Klassen. Daraus folgt die behauptete Struktur der invarianten Maße sofort. □

Beispiele. Wir schauen uns die Klassenzerlegung und invarianten Verteilungen für unsere drei Beispiele von vorher an.

- **Unabhängige Zufallsvariablen**. Hier ist die Markovkette irreduzibel und aperiodisch. Darüber hinaus ist die Übergangsmatrix bereits ein Projektor auf die einzige invariante Verteilung π_0.
- **Irrfahrt mit Rand**. Hier gibt es offenbar drei Klassen: $C_1 \equiv \{-L+1,\ldots,L-1\}$, $C_2 = \{-L\}$ und $C_3 = \{L\}$. Dabei ist C_1 unwesentlich und C_2 und C_3 sind wesentlich. Daher haben wir zwei invariante Verteilungen, μ_2 und μ_3, wobei

$$\mu_2(j) = \delta_{j,-L}, \qquad \mu_3(j) = \delta_{j,L}. \tag{4.47}$$

  Natürlich sind auch alle konvexen Linearkombinationen dieser zwei Verteilungen invariante Verteilungen. Da für jede invariante Verteilung $\mu(C_1) = 0$ gilt, erschöpfen diese offenbar die invarianten Verteilungen dieser Markovkette.

- **Wettermodell**. Seien zunächst $p_{0,1}, p_{1,0} \in (0,1)$. Dann ist die Markovkette wieder irreduzibel und aperiodisch, und die einzige invariante Verteilung ist

$$\mu = \frac{1}{(p_{0,1}+p_{1,0})}(p_{1,0}, p_{0,1}). \tag{4.48}$$

  dasselbe gilt wenn einer der beiden Parameter gleich eins ist, der andere aber in $(0,1)$ liegt.
  Wenn $p_{1,0}$ und $p_{0,1}$ gleich null sind, so gibt es zwei wesentliche Klassen mit den jeweils trivialen Verteilungen. Falls nur eine der beiden null ist, so gibt es eine wesentliche und eine unwesentliche Klasse.
  Wenn $p_{0,1} = p_{1,0} = 1$ ist, haben wir eine irreduzible, aber nicht aperiodische Klasse. Die Markovkette hat dann Periode zwei, wie schon oben beschrieben.

4.3.5 Markovketten Monte-Carlo Verfahren.

Eine in der Praxis wesentliche Anwendung des Ergodensatzes für Markovketten ist die Möglichkeit, mit seiner Hilfe Integrale bezüglich einer gewünschten Verteilung numerisch approximativ zu berechnen.

Bei der Berechnung von Erwartungswerten trifft man in der Praxis of auf zwei Probleme: (1) Der Zustandsrausm ist sehr gross (und hochdimensional) (etwa etwa in der statistischen Mechanik, Maße nur "bis auf die Normierung" explizit gegeben, eta in der Form

$$\rho(x) = \frac{1}{Z}\exp\left(-\beta H(x)\right), \tag{4.49}$$

wo $H(x)$ eine einfach zu berechende Funktion ist, die Konstante Z aber nur als $\sum_{x\in S}\exp\left(-\beta H(x)\right)$ gegeben ist, also etwa so schwer zu berechnen ist wie das Integral selbst.

Hier kommen nun die Markovketten und der Ergodensatz ins Spiel. Angenommen, wir fänden eine ergodische Markovkette mit Zustandraum S derart, das die invariante Verteilung der Kette gerade ρ ist. Da die Normierung für die Invarianzgleichung keine Rolle spielt, kann man eine solche konstruieren, ohne Z zu kennen. Dann wissen wir, dass

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} f(X_k) \equiv \mathbb{E}[f(X)]., \quad \text{in Wahrscheinlichkeit.} \tag{4.50}$$

Um eine systematiche Approximation unseres Integrals zu bekommen, benötigen wir also nur eine Realisierungen der Zufallsvariablen $X_1, X_2, \ldots$. Dabei gewinnen wir natürlich nur dann etwas, wenn die entsprechenden bedingten Verteilungen, also die Übergangwahrscheinlichkeiten der Markovkette, finden können. Dazu muss man natürlich in der Lage sein, diese Zufallsvariablen in einfacher Weise numerisch zu konstruieren. Dazu ist es nützlich, die Markovkette so zu konstruieren, dass man von einem gegebenen Zustand aus nur sehr wenige Zustände erreichen kann; im obigen Beispiel $S = \{-1,1\}^N$ wählt man die Markovkette etwa so, dass man in einem Schritt nur eine der Koordinaten des Vektors x ändern kann. Dann sind die Übergangswahrscheinlichkeiten effektiv Verteilungen auf nur N (statt 2^N) Zuständen, und somit viel leichter handhabbar. Im obigen Fall kann man z.B. die Übergangswahrscheinlichkeiten in der Form

$$p_{xy} = \frac{1}{N}\exp\left(-[H_N(y) - H_N(x)]_+\right), \text{wenn } |x-y| = 2, \tag{4.51}$$

$$p_{xx} = 1 - \sum_{y:|x-y|=2} p_{xy}, \tag{4.52}$$

und null sonst, wählen (Übung!).

Damit dieses Verfahren funktioniert, sollte natürlich die Konvergenz gegen die invariante Verteilung schnell genug erfolgen, so dass man tatsächlich rasch gute Approximationen erhält. Dies zu quantifizieren ist im Allgemeinen ein schwieriges

Problem. In vielen Fällen liefert dieses *Markovketten Monte-Carlo Verfahren* aber sehr gute Resultate. Monte-Carlo Verfahren sind ein wichtiges Hilfsmittel der stochastischen Numerik und werden in verschiedener Form sehr verbreitet eingesetzt.

Kapitel 5
Der zentrale Grenzwertsatz

On peut facilement, au moyen de ces formules, déterminer les bénéfices des loteries[a].
Pierre Simon de Laplace, Théorie Analytique des Probabilités

[a] Man kann mittels dieser Formeln leicht den Gewinn von Lotterien berechnen.

Wir kommen nun zu dem zweiten wichtigen Satz der Wahrscheinlichkeitstheorie, dem nicht ohne Grund so genannten *zentralen Grenzwertsatz*. Seine Bedeutung liegt zum einen wieder in den Implikationen für die Statistik, denn er rechtfertigt in vielen Fällen die Annahme einer Gauß'schen Verteilung (bzw. derer Derivate) für Zufallsgrössen die auf komplizierte Art und Weise zustande kommen. Zum anderen ist er ein weiteres Beispiel dafür, wie spezifische Gesetzmässigkeiten aus zufälligem Geschehen folgen.

5.1 Fehler im Gesetz der großen Zahlen

Der zentrale Grenzwertsatz kann als Verfeinerung des Gesetzes der großen Zahlen aufgefasst werden. Wir haben mit dem Gesetz der großen Zahlen gesehen, dass Summen, $S_n \equiv \sum_{i=1}^n X_i$, unabhängiger, identisch verteilter Zufallsvariablen, X_i, sich für große n annähernd wie $n\mathbb{E}[X_1]$ verhalten, in dem Sinn, dass $(S_n - n\mathbb{E}[X_1])/n \to 0$ konvetgiert. Es liegt nun nahe, die Frage nach der Konvergenzgeschwindigkeit zu stellen. Dies ist in vielen praktischen Anwendungen von grosser Bedeutung.

Beispiel. *Ein Hersteller von Computerchips muss eine Million funktionsfähige Chips liefern. Er weiss, dass im Durchschnitt* 20% *der produzierten Chips fehlerhaft sind. Wieviele Chips soll er produzieren, damit er mit* 99% *Wahrscheinlichkeit eine Million funktionsfähige Chips hat? Nehmen wir an, dass die Ereignisse* produzierter Chip Nummer k ist ok" *unabhängig sind. Dann ist die Zahl der guten Chips ei n produzierten Chips die Summe von n unabhängigen Bernoulli-Zufallsvariablen mit Parameter* $0,8$. *Das Gesetz der großen Zahlen sagt, dass bei n produzierten Chips die Zahl der brauchbaren ca.* $0,8n$ *ist, also ca.* 1.25 *Millionen Chips produziert werden müssen. Allerdings wissen wir nur, dass der Fehler klein gegen n ist, ohne dass wir*

genau wüssten, wie viel kleiner. *Nun macht es aber für die genaue Kostenkalkulation der Firma einen grossen Unterschied, ob nun* 1.3 *oder* 1.5 *Mio Chips produziert werden müssen!*

Wie können wir eine bessere Kontrolle über den Fehler bekommen? Ein erster Schritt ist die Berechnung der Varianz der Zufallsvariablen S_n. Wir nehmen hierzu an, dass die unabhängigen und gleichverteilten Zufallsvariablen X_i endlichen Erwartungswert μ und eine endliche Varianz σ^2 haben. Es gilt dann nämlich:

$$\mathrm{var}(S_n) \equiv \mathbb{E}\left[(S_n - \mathbb{E}[S_n])^2\right) = \sum_{k=1}^{n} \mathbb{E}\left[(X_k - \mathbb{E}[X_k])^2\right] = n\sigma^2. \tag{5.1}$$

Anders gesagt, die Zufallsvariablen $Z_n \equiv (S_n - n\mu)/\sqrt{n}$ haben alle die gleiche Varianz, nämlich σ^2. Dies besagt, dass die Abweichungen von S_n von seinem Mittelwert nur von der Ordnung $\sqrt{n}$ sind! Mit Hilfe der Markov-Ungleichung (Korollar 2.12) können wir dies sogar noch präziser machen:

Lemma 5.1. *Seien $(X_k, k \in \mathbb{N})$ unkorrelierte Zufallsvariablem mit Mittelwert $\mathbb{E}[X_k] = \mu$ und Varianz $\mathrm{var}(X_k) = \sigma^2 < \infty$. Dann gilt, für alle $\delta > 0$,*

$$\mathbb{P}\left(|S_n - n\mu| > \sqrt{n}\delta\right) \leq \frac{\sigma^2}{\delta^2}. \tag{5.2}$$

Beispiel (Fortsetzung). *Wir können nun der Lösung unseres Problems näher kommen, indem wir die Varianz der Bernoulli-Zufallsvariablen X_k ausrechnen. Dies ist*

$$\mathrm{var}(X_k) = \mathbb{E}[X_k^2] - 0,8^2 = 0,8 - 0,64 = 0,16. \tag{5.3}$$

Es folgt, dass

$$\mathbb{P}\left(|S_n - n0.8| \geq \delta\sqrt{n}\right) \leq 0.16/\delta^2. \tag{5.4}$$

Um eine Wahrscheinlichkeit von weniger als $0,01$ *zu erhalten, müssen wir also $\delta = 4$ wählen. Gl. (5.4) sagt uns dann, dass wir bei n produzierten Chips mit Wahrscheinlichkeit mindestens* 0.99*, zumindest $n0.8 - 4\sqrt{n}$ gute Chips haben. Lösen wir diese Gleichung nach n auf, so erhalten wir, dass nur* $1,25560$ *Mio Chips produziert werden müssen, gerade* 5600 *mehr als der Mittelwert. Dies ist deutlich besser, als das Gesetz der großen Zahlen nahegelegt hätte!*

Kann man noch mehr sagen? Die Tatsache, dass die Zufallsvariablen Z_n alle Mittelwert Null und Varianz σ^2 haben, lässt die Frage aufkommen, ob die Folge Z_n, $n \in \mathbb{N}$, vielleicht sogar gegen eine bestimmte Zufallsvariable Z (mit Mittelwert Null und Varianz σ^2) *konvergiert*, etwa in dem Sinne, dass, wenn $n \uparrow \infty$,

$$\mathbb{P}(Z_n \leq x) \to \mathbb{P}(Z \leq x)? \tag{5.5}$$

Definition 5.2. Eine Folge von Zufallsvariablen $(Z_n, n \in \mathbb{N})$ konvergiert *in Verteilung* gegen eine Zufallsvariable Z genau dann, wenn für alle $x \in \mathbb{R}$, für die $\mathbb{P}(Z = x) = 0$ gilt,

$$\lim_{n\uparrow\infty} \mathbb{P}(Z_n \leq x) = \mathbb{P}(Z \leq x). \tag{5.6}$$

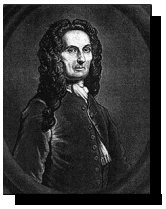
Mit dem Begriff der Verteilungsfunktion können wir alternativ formulieren: *Eine Folge von Zufallsvariablen* $(Z_n, n \in \mathbb{N})$ *konvergiert* in Verteilung *gegen eine Zufallsvariable Z genau dann, wenn für die Folgen der Verteilungsfunktionen* F_{Z_n} *gilt, dass*

$$\lim_{n\uparrow\infty} F_{Z_n}(x) = F_Z(x), \tag{5.7}$$

für alle $x \in \mathbb{R}$*, für die* F_Z *bei x stetig ist.*

Anmerkung. Es mag zunächst verwundern, dass wir nicht fordern, dass die Konvergenz an den Stellen gilt, an denen F_Z einen Sprung macht. Das dies aber sinnvoll ist, kann man sich an folgendem Beispiel klar machen. Es sei Z_n eine Folge von Zufallsvariablen, die die Wert $1/n$ und $1-1/n$ jeweils mit Wahrscheinlichkeit $1/2$ annehmen. Es ist bestimmt sinnvoll zu sagen, dass Z_n gegen eine Bernoulli-Zufallsvariable Z mit Parameter $1/2$ konvergiert. Nun ist aber $\mathbb{P}(Z_n \leq 0) = 0$, für alle $n \in \mathbb{N}$, und also $\lim_{n\uparrow\infty} \mathbb{P}(Z_n \leq 0) = 0$, während $\mathbb{P}(Z \leq 0) = \mathbb{P}(Z = 0) = 1/2$. Diese Tatsache soll uns aber nicht abhalten zu sagen, dass Z_n gegen Z konvergiert.

5.2 Der Satz von de Moivre-Laplace.



Die Antwort auf diese Frage wurde bereits im 17. Jahrhundert von de Moivre affirmativ beantwortet in dem Fall, dass die Zufallsvariablen X_k Bernoulli-Zufallsvariablen sind.

Satz 5.3 (Der Satz von de Moivre-Laplace). *Seien* X_i *eine Folge von unabhängigen Bernoulli Zufallsvariablen mit Parameter* $p \in (0,1)$*. Dann konvergiert die Folge* $Z_n \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - p)$ *in Verteilung gegen eine Gauß-verteilte Zufallsvariable, Z, mit Mittelwert Null und Varianz* $\sigma^2 = p(1-p)$*.*

Beweis. Wir wählen ein Intervall $I = [a,b]$, $a < b \in \mathbb{R}$. Wir wollen zeigen, dass

$$\lim_{n\uparrow\infty} \mathbb{P}(Z_n \in [a,b]) = \frac{1}{\sqrt{2\pi p(1-p)}} \int_a^b \mathrm{e}^{-\frac{x^2}{2p(1-p)}}\, dx. \tag{5.8}$$

Wir setzen $S_n \equiv \sum_{i=1}^{n} X_i$. Wir wissen schon, dass die Zufallsvariablen S_n binomial verteilt sind mit $Bin(n,p)$, also

$$\mathbb{P}(S_n = k) = p^k (1-p)^{n-k} \binom{n}{k}.$$

Insbesondere ist $\mathbb{E}[S_n] = pn$ und die Varianz von S_n, $\mathrm{var}(S_n) = np(1-p)$. Dann ist $Z_n = \frac{1}{\sqrt{n}}(S_n - pn)$ eine Folge von Zufallsvariablen mit Mittelwert $\mathbb{E}[Z_n] = 0$ und $\mathrm{var}(Z_n) = p(1-p)$. Offenbar ist

$$\mathbb{P}(Z_n \in [a,b]) = \mathbb{P}\left(S_n \in (\sqrt{n}a + pn, \sqrt{n}b + pn)\right) = \sum_{k:\, \frac{1}{\sqrt{n}}(k-pn)\in[a,b]} \mathbb{P}(S_n = k)\,. \tag{5.9}$$

Beachte, dass wir für unser Problem nur die Werte $k \in [np + \sqrt{n}a, np + \sqrt{n}b]$ benötigen. Für feste Zahlen a,b und $p \in (0,1)$ sind alle diese k von der Ordnung n, wenn n gross genug ist, und ebenso ist dann $n-k$ von der Ordnung n. Um zu verstehen, wie sich diese Wahrscheinlichkeiten für große n verhalten, müssen wir die Binomialkoeffizienten $\binom{n}{k} \equiv \frac{n!}{k!(n-k)!}$ untersuchen. Dabei sind scheinen die Fakultäten die Sache kompliziert zu machen. Glücklicherweise gibt es aber dafür sehr gute Annäherungen, nämlich die *Stirling'sche Approximation*.

Lemma 5.4 (Stirling Formel). *Es gilt für alle* $n \in \mathbb{N}$

$$\sqrt{2\pi}n^{n+1/2}\mathrm{e}^{-n}(1+1/(12n)) \le n! \le \sqrt{2\pi}n^{n+1/2}\mathrm{e}^{-n}(1+1/(12n-1)). \tag{5.10}$$

Wenden wir dies auf die Binomialkoeffizienten an, so erhalten wir

$$\begin{aligned}
\binom{n}{k} &= \frac{n!}{(n-k)!k!} = \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{(n-k)k}}\,\frac{n^n}{(n-k)^{n-k}k^k} \\
&\qquad \times (1+O(1/n)) \\
&= \frac{1}{\sqrt{2\pi n}}\sqrt{\frac{1}{(1-k/n)k/n}}\,\frac{1}{(1-k/n)^{n-k}(k/n)^k} \\
&\qquad \times (1+O(1/n)) \\
&= \frac{1}{\sqrt{2\pi n}}\sqrt{\frac{1}{(1-k/n)k/n}}\left(\frac{1}{(1-k/n)^{1-k/n}(k/n)^{k/n}}\right)^n \\
&\qquad \times (1+O(1/n))\,.
\end{aligned} \tag{5.11}$$

Hier bedeutet die Notation $A = B + O(1/n)$, dass es ein von n unabhängige Konstante c gibt, so dass $|A-B| \le c/n$.

Setzen wir nun $k/n = x$ und all dies in die Formel (5.9) für $\mathbb{P}(S_n = nx)$ ein, so ist

$$\begin{aligned}
\mathbb{P}(S_n = nx) &= \frac{1}{\sqrt{2\pi n}}\sqrt{\frac{1}{(1-x)x}}\left(\frac{p^x(1-p)^{1-x}}{(1-x)^{1-x}(x)^x}\right)^n (1+O(n^{-1})) \\
&= \frac{1}{\sqrt{2\pi n}}\sqrt{\frac{1}{(1-x)x}}\exp\left(-nI(p,x)\right)(1+O(n^{-1}))
\end{aligned} \tag{5.12}$$

wo

$$\begin{aligned}
I(p,x) &= \ln\left((x/p)^x[(1-x)/(1-p)]^{1-x}\right) \\
&= x\ln(x/p) + (1-x)\ln((1-x)/(1-p))
\end{aligned} \tag{5.13}$$

Folgende einfache Sachverhalte sind leicht nachzuprüfen (Übung!):

(i) $I(p,p) = 0$
(ii) $I(p,x)$ is konvex als Funktion von $x \in (0,1)$ und nimmt ihr einziges Minimum $x = p$ an.
(iii) $\frac{\partial^2 I(p,x)}{\partial x^2} = \frac{1}{x} + \frac{1}{1-x} = \frac{1}{x(1-x)} \geq 4$.
(iv) $I(p,x)$ ist unendlich oft differenzierbar im Intervall $(0,1)$.

Wir sehen an den obigen Rechnungen, dass $\mathbb{P}(S_n = nx)$ nur dann nicht exponentiell klein in n wird, wenn x sehr nahe bei p liegt.

Mittels der Taylorformel dritter Ordnung zeigt man nun, dass für alle Werte von k, die in der Summe (5.9) auftreten,

$$\left| I(p,k) - \frac{(k/n-p)^2}{2p(1-p)} \right| \leq Cn^{-3/2},$$

wo die Konstante C nur von p,a,b abhängt. Weiter ist für diese Werte

$$\left| \sqrt{\frac{1}{(1-k/n)k/n}} - \sqrt{\frac{1}{p(1-p)}} \right| \leq Cn^{-1/2}.$$

Damit erhalten wir

$$\begin{aligned} &\mathbb{P}(Z_n \in [a,b]) \qquad (5.14)\\ &= \sum_{k:\frac{1}{\sqrt{n}}(k-pn)\in[a,b]} \frac{1}{\sqrt{2\pi n}} \sqrt{\frac{1}{(1-p)p}} \exp\left(-n\frac{(k/n-p)^2}{2p(1-p)}(1+O(n^{-3/2})) \right) (1+O(n^{-1/2})) \end{aligned}$$

Wenn wir $k/n - p = y$ setzen, erkennen wir die Dichte der Gauß-Verteilung mit Varianz $\sigma^2 = (1-p)p$. Jetzt brauchen wir nur noch die Summe durch ein Integral zu ersetzen. Dazu bemerkt man, dass

$$\begin{aligned} &\frac{1}{n} \exp\left(-n\frac{(k/n-p)^2}{2p(1-p)}(1+O(n^{-3/2})) \right) (1+O(n^{-1/2})) \qquad (5.15)\\ &= \int_{k/n-p}^{(k+1)/n-p} \exp\left(-n\frac{y^2}{2p(1-p)}(1+O(n^{-3/2})) \right) (1+O(n^{-1/2}))\mathrm{d}y, \end{aligned}$$

da sich der Integrand zwischen den Integrationsgrenzen nur um einen Faktor höchstens der Form $1+O(1/n)$ unterscheidet. Somit haben wir

$$\begin{aligned}
&\sum_{k:\frac{1}{\sqrt{n}}(k-pn)\in[a,b]} \frac{1}{\sqrt{2\pi n}}\sqrt{\frac{1}{(1-p)p}}\exp\left(-n\frac{(k/n-p)^2}{2p(1-p)}(1+O(n^{-3/2}))\right)(1+O(n^{-1/2}))\\
&=\int_{a/\sqrt{n}}^{b/\sqrt{n}} \frac{\sqrt{n}}{\sqrt{2\pi p(1-p)}}\exp\left(-n\frac{y^2}{2p(1-p)}(1+O(n^{-3/2}))\right)(1+O(n^{-1/2}))\mathrm{d}y\\
&=\int_a^b \frac{1}{\sqrt{2\pi p(1-p)}}\exp\left(-\frac{x^2}{2p(1-p)}(1+O(n^{-1/2}))\right)(1+O(n^{-1/2}))\mathrm{d}x\\
&\to\int_a^b \frac{1}{\sqrt{2\pi p(1-p)}}\exp\left(-\frac{x^2}{2p(1-p)}\right)\mathrm{d}x \qquad (5.16)
\end{aligned}$$

Da dies für jedes Intervall $[a,b]$ gilt, folgt schliesslich auch die Konvergenz der Verteilungsfunktionen. Damit haben wir aber das behauptete Resultat bewiesen. □

Anmerkung. Wir haben im Beweis gesehen, dass die relativen Fehler in der Approximation durch die Gauß-Verteilung von der Ordnung $C/\sqrt{n}$ sind, wobei die Konstante im wesentlichen von p abhängt und umso größer wird, je näher p an Null oder 1 heranrückt. Die Aussage des Satzes gilt *nicht*, wenn p von n abhängt! So hatten wir bereits gesehen, dass für $p=\rho/n$ die Binomialverteilung gegen die Poisson-Verteilung (und also nicht die Gauß-Verteilung) konvergiert. Man kann sich das auch so erklären, dass in diesem Fall nur ein kleiner Bruchteil der Summanden, im Mittel nämlich gerada ρ, überhaupt von Null verschieden ist. Die tatsächliche Zahl ist dann Poisson verteilt.

Der Satz von de Moivre-Laplace kann nun zur Approximation von Wahrscheinlichkeiten von binomial verteilten Zufallsvariablen benutzt werden. Wir wollen etwa ausrechnen, was $\mathbb{P}(S_n > pn+x\sqrt{n})$ ist. Dies können wir umschreiben als

$$\begin{aligned}
\mathbb{P}(pn+a\sqrt{n}\le S_n\le pn+b\sqrt{n}) &= \mathbb{P}(a\le Z_n\le b)\\
&\sim \frac{1}{\sqrt{2\pi p(1-p)}}\int_a^b \mathrm{e}^{-\frac{y^2}{2p(1-p)}}dy\\
&= \frac{1}{\sqrt{2\pi}}\int_{a/\sqrt{p(1-p)}}^{b/\sqrt{p(1-p)}} \mathrm{e}^{-\frac{x^2}{2}}dx. \qquad (5.17)
\end{aligned}$$

Wenn wir noch $\mu_n = pn$ und $\sigma_n \equiv \sqrt{np(1-p)}$ definieren, erhalten wir die folgende praktische Formel

$$\mathbb{P}\left(\mu_n + r\sigma_n \le S_n \le \mu_n + t\sigma_n\right) \sim \frac{1}{\sqrt{2\pi}}\int_r^t \mathrm{e}^{-\frac{x^2}{2}}dx. \qquad (5.18)$$

Beispiel (Fortsetzung). *Der Satz von de Moivre-Laplace lässt sich auf unser Beispiel anwenden. Gewinnen wir dadurch etwas? Wir haben jetzt die Abschätzung*

$$\mathbb{P}\left(S_n \leq n0.8 - \delta\sqrt{n}\right) \sim \frac{1}{\sqrt{2\pi 0.16}} \int_{-\infty}^{-\delta} \mathrm{e}^{\frac{-y^2}{2\times 0.16}} dy$$
$$= \frac{1}{\sqrt{2\pi}} \int_{0.4\delta}^{\infty} \mathrm{e}^{-\frac{y^2}{2}} dy. \tag{5.19}$$

Die Funktion $\phi(x) = 1 - \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \mathrm{e}^{-\frac{y^2}{2}} dy$ (Φ heisst Fehlerfunktion", ϕ "komplementäre Fehlerfunktion") kann man zwar nicht explizit ausrechnen, aber gut numerisch berechnen. Es gibt auch schon seit langem Tabellen. Insbesondere ist $\phi(2.6) \sim 0.01$. Daher müssen wir $\delta = 0.4 \times 2.6 \sim 1.04$ wählen, um einen Fehler mit Wahrscheinlichkeit 0.99 genügend gute Chips zu erhalten. Das macht schon einen großen Unterschied gegenüber dem Wert $\delta = 4$, den wir aus der Markov-Ungleichung erhalten hatten.

Der Vorteil des zentralen Grenzwertsatzes wird deutlicher, wenn wir grössere Sicherheiten haben wollen. Angenommen, wir wollten genug Chips mit Wahrscheinlichkeit $1 - 10^{-6}$ haben. Dann würde die Markov Ungleichung ein $\delta = 400$ erfordern, was zu 1.947 Mio zu produzierender Chips führen würde. Mit dem zentralen Grenzwertsatz ergibt sich aber, dass $\delta = 2$ ausreicht, und somit nur ca. 1.2539Mio Chips produziert werden müssen, also weit über eine halbe Million weniger!

Anmerkung. Im Prinzip könnte man natürlich in unserem Fall exakte Formeln mit Hilfe der Binomialverteilung angeben. Allerdings sind die exakten Berechnungen der Binomialkoeffizienten bei so großen Zahlen praktisch unmöglich und nicht sinnvoll. Die Approximation mit der Gauß-Verteilung ist um vieles einfacher.

In der Tat ist der Satz von de Moivre-Laplace nur ein Spezialfall des viel allgemeineren *Zentralen Grenzwertsatzes*, den wir nur formulieren, aber nicht beweisen werden.

Satz 5.5. *Sei $(X_i, i \in \mathbb{N})$ eine Folge unabhängiger, identisch verteilter Zufallsvariablen und es gelte $\mathbb{E}[X_1] = \mu$, $\mathrm{var}(X_1) = \sigma^2 < \infty$. Dann konvergiert die Folge $Z_n \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu)$ in Verteilung gegen eine Gauß-verteilte Zufallsvariable mit Mittelwert Null und Varianz σ^2.*

Der zentrale Grenzwersatz ist eines der wichtigsten Ergebnisse der Wahrscheinlichkeitstheorie. Er macht die Gauß- Verteilung zu einer der wichtigsten Verteilungen und begründet, warum in sehr vielen Anwendungen, Zufallsgrößen mit Gauß'schen Zufallsvariablem modelliert werden.

Anmerkung. In dieser Allgemeinheit wurde der Zentrale Grenzwertsatz 1922 von Jarl Waldemar Lindeberg bewiesen, nachdem Lyapunov eine Version unter stärkeren Bedingungen schon 1901 gezeigt hatte.

Kapitel 6
Statistik

La probabilité de la plupart des événements simples est inconnue : en la considérant a priori*, elle nous paraît susceptible de toutes les valeurs comprises entre zéro et l'unité; mais, si l'on a observé un résultat composé de plusieurs de ces événements, la manière dont ils y entrent rend quelques-unes de ces valeurs plus probables que les autres. Ainsi, à mesure que le résultat observé se compose par le développement des événements simples, leur vraie possibilité se fait de plus en plus connaître, et il devient de plus en plus probable qu'elle tombe dans les limites qui, se resserant sans cesse, finiraient par coincider, si le nombre des événements simples devenait infini*[a].
Pierre Simon de Laplace, Théorie Analytique des Probabilités

[a] Die Wahrscheinlichkeit des meissten einfachen Ereignisse ist unbekannt: indem wir sie *a priori* betrachten, erscheinen alle Werte zwischen null und eins möglich; wenn man aber ein Ergebnis beobachtet, dass aus mehreren dieser Ereignisse zusammengesetzt ist, so macht die Art, wie diese eintreten, einige dieser Werte wahrscheinlicher als andere. So lässt sich, sofern das beobachtete Resultat sich aus der Entwicklung der einfachen Ereignisse zusammensetzt, ihre wirkliche Möglichkeit mehr und mehr erkennen, und es wird immer wahrscheinlicher, dass sie zwischen Schranken fällt, die, indem sie sich immer mehr zusammenziehen schlussendlich zusammenfielen, wenn die Zahl der einfachen Ereignisse unendlich würde.

Die Statistik verbindet Wahrscheinlichkeitstheoretische Modelle mit realen Daten. Sie liefert Methoden, um aus partiellen Beobachtungen realer Systeme Erkenntnisse über das System zu gewinnen. Alle Bereiche der empirischen Wissenschaften sind auf statistische Methoden angewiesen, aber auch im täglichen Leben sind wir ständig von statistischen Anwendungen umgeben. Elementare Kenntnisse statistischer Methoden und derer Probleme sind daher notwendige Grundkenntnisse für jedermann um auf Statistik beruhende Aussagen einschätzen zu können. Es ist daher extrem wichtig, diese Kompetenzen in der Schule zu vermitteln. in den folgenden Kapiteln benutzen wir neben Material aus [3] auch einiges aus dem zweiten Teil des Lehrbuchs von Georgii [1].

6.1 Statistische Modelle und Schätzer

Die Aufgabe der Statistik ist die Beschreibung von Beobachtungen von "Zufallsexperimenten" durch ein auf ein auf Zufallsvariablen basiertem *Modell*. Ganz allgemein gesprochen sieht das so aus. Gegeben sind eine Folge von *Beobachtun-*

gen (= Ausgänge von Zufallexperimenten), $Z_1, \ldots, Z_n$. Der Statistiker möchte diese als Realisierungen von n Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$ interpretieren. Er interessiert sich für die gemeinsame Verteilung der entsprechenden n Zufallsvariablen, die er a priori nicht kennt, sondern aus den Beobachtungen Z_i (interpretiert als einer Realisierung $\omega \in \Omega$), bestimmen, bzw. im statistischen Sprachgebrauch, *schätzen* muß.

Ohne weiteres ist dies praktisch nicht möglich, und man wird aufgrund von zusätzlichen "a priori" Informationen weitere Annahmen (Hypothesen) an die Zufallsvariablen machen. Im allgemeinen besteht ein statistisches Modell somit aus Modellannahmen und Modellparametern, wobei die Annahmen als wahr angesehen werden, und die Parameter zunächst unbekannt sind. Um die unbekannten Parameter zu bestimmen konstruiert der Statistiker nun sogenannte Schätzer, d.h. Funktionen der beobachteten Größen X_i, die die Werte der "wahren" Parameter annähern sollen. Die Schätzer, a_n, hängen dabei von n und von den Beobachtungen X_i, $i \leq n$ ab.

Eine wichtige Eigenschaft, die man von Schätzern fordert, ist die *Konsistenz*

Definition 6.1. Sei $X_n, i \in \mathbb{N}$ eine Families von Zufallsvariablen mit gemeinsamer Verteilung, die durch Parameter $a \in \mathbb{R}^k$ parametrisiert ist. Dann heisst eine Funktion $a_n : \mathbb{R}^n \to \mathbb{R}$ ein *konsistenter Schätzer* für die Parameter a, falls die Zufallsvariablen

$$a_n(X_1(\omega), \ldots, X_n(\omega)) \to a, \tag{6.1}$$

(in Wahrscheinlickeit), wenn $n \to \infty$.

Wir betrachten jetzt einige wichtige Beispiele.

6.1.1 Frequenzen

Seien unsere Beobachtungen X_i die Ausgänge von stets gleichen und sich nicht beeinflussenden Zufallsexperimenten, etwa eine Folge von Glücksspielen. Dann ist es eine plausible Annahme, dass die X_i durch unabhängige, gleichverteilte Zufallsvariablen mit gemeinsamer Verteilung ν zu modellieren sind. Hier ist also die Unabhängigkeit eine Modellannahmen, während die Verteilung, ν, zunächst ein unbekannter "Parameter" ist. Wie können wir aus den Beobachtungen ν schätzen?

Das Gesetz der großen Zahlen erlaubt es uns auf die Frage nach der Konvergenz der Frequenzen, die schon im ersten Abschnitt angesprochen war genauer einzugehen. Wir erinnern uns, dass wir in einer Reihe von n "identischen" Spiele (Zufallsexperimente) die Frequenzen der Ausgänge $X_i \in A$ definiert hatten als

$$\nu_n(A) \equiv \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_A(X_i). \tag{6.2}$$

Folgen unabhängiger, identisch verteilter Zufallsvariablen sind nun genau das statistische Modell für eine solche Folge identischer, sich nicht beeinflussender Zufalls-

experimente. Das Gesetz der großen Zahlen sagt uns dann, dass die Annahme der Konvergenz in der Tat korrekt war. Es gilt nämlich:

Lemma 6.2. *Seien $X_i, i \in \mathbb{N}$, eine Folge reellwertiger, unabhängiger, identisch verteilter Zufallsvariablen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$ mit Verteilung ν. Dann gilt, mit ν_n definiert durch (6.2),*

(i) Für jedes $A \in \mathfrak{B}(\mathbb{R})$ gilt

$$\nu_n(A) \to \nu(A), \tag{6.3}$$

und

(ii) ν ist die Wahrscheinlichkeitsverteilung von X_1, i.e. für alle $A \in \mathfrak{F}$ gilt

$$\nu(A) = \mathbb{P}[X_1 \in A].$$

Beweis. Der Beweis ist denkbar einfach: Die Funktionen $\mathbb{1}_A(X_i)$ sind selbst Zufallsvariablen, und zwar, wie man leicht nachprüft, unabhängige. Ihre Erwartung ist gerade

$$\mathbb{E}[\mathbb{1}_A(X_i)] = \mathbb{P}[X_i \in A] = \mathbb{P}[X_1 \in A].$$

Da diese endlich sind, folgen beide Aussagen des Lemmas aus dem Gesetz der großen Zahlen. □

Die Sammlung der $\nu_n(A)$ stellt für jede Realisierung der Zufallsvariablen X_i ein Wahrscheinlichkeitsmaß auf den reellen Zahlen dar. Also, im Rahmen des statistischen Modells, in dem die Ausgänge eines Zufallsexperiments unabhängige, gleichverteilte Zufallsvariablen sind, sind die empirischen Verteilungen, d.h. die Frequenzen, tatsächlich *Schätzer* für die gemeinsame Verteilung dieser Zufallsvariablen, und dieser Schätzer ist darüber hinaus konsistent.

Mit der Chebychev'schen Ungleichung erhalten wir sogar eine *Qualitätsabschätzung*.

Lemma 6.3. *Seien $X_i, i \in \mathbb{N}$, eine Folge rellwertiger, unabhängiger, identisch verteilter Zufallsvariablen mit Verteilungsfunktion F auf einem Wahrscheinlichkeitsraum $(\Omega, \mathfrak{F}, \mathbb{P})$. Dann gilt, für jede Borelmenge A, dass*

$$\mathbb{P}\left[|\nu_n(A) - \nu(A)| > c\nu(A)\right] \leq \frac{1}{nc^2\nu(A)}. \tag{6.4}$$

Beweis. Übung! □

Wie man an der Abschätzung sieht, sind die Schätzungen für Mengen kleiner Masse fehlerhafter als die von großer Masse. Dies ist nur natürlich: Ist $\nu(A)$ klein, so bedarf er vieler Experimente, bis überhaupt einmal ein Ergebnis in A fällt! Die Qualität des Schätzers hängt also von der erwarteten Zahl der Ereignisse, die in A fallen, eben $n\nu(A)$, direkt ab.

In der Statistik benutzt man gerne den Begriff des *Konfidenzintervalls*. Die ist der Bereich, in dem ein zu schätzender Parameter auf Grund der Schätzung mit einer vorgegebenen Wahrscheinlichkeit, oft 0.95%, liegen wird. Wir können für unseren

Fall Lemma 6.3 benutzen, um eine .95% Konfidenzintervall für $\nu(A)$ zu finden. Dazu schreiben wir die Ungleichung (6.4) um in die Form

$$\mathbb{P}\left[|\nu_n(A)-\nu(A)|>c\sqrt{\nu(A)}\right]\leq\frac{1}{nc^2}. \tag{6.5}$$

Dann bestimmen c, so dass die rechte Seite gleich 0.05 wird, also $c=1/\sqrt{0.05n}$. Das Komplement des Ereignisses in (6.5) is

$$\nu(A)-\nu_n(A)\leq c\sqrt{\nu(A)},\quad \text{und}\nu(A)-\nu_n(A)\geq -c\sqrt{\nu(A)}, \tag{6.6}$$

das heisst

$$\nu(A)\in\left[\nu_N(A)+c\sqrt{\nu_n+c^2/4}+c^2/2,\nu_N(A)-c\sqrt{\nu_n+c^2/4}+c^2/2\right], \tag{6.7}$$

mit Wahrscheinlichkeit mindestens 0.95.

6.1.2 Schätzen von Erwartungswert und Varianz

Wir haben gesehen, dass Erwartungswert und Varianz einer Zufallsvariable bereits wichtige Informationen über deren Verteilung enthalten. Es liegt also für einen Statistiker nahe, zunächst mal diese Kenngrößen zu schätzen, als gleich die ganze Verteilung. Das Gesetz der großen Zahlen liefert uns wieder Kandidaten für solche Schätzer sowie eine Rechtfertigung. Betrachten wir zunächst den Mittelwert einer Verteilung. Nach dem Gesetz der großen Zahlen konvergiert ja das *empirische Mittel*,

$$m_n\equiv n^{-1}\sum_{i=1}^{n}X_i \tag{6.8}$$

gegen $\mu\equiv\mathbb{E}X_1$, falls die X_i unabhängige, identisch Verteilte Zufallsvariablen sind. Damit ist die Zufallsvariable m_n, gut geeignet, um als *Schätzer* für den Mittelwert zu dienen. Betrachten wir nun wieder die Zuverlässigkeit des Schätzers. Wir begnügen uns mit dem Fall, dass die X_1 endliche zweite Momente haben. Dann liefert die Chebychev Ungleichung sofort:

Lemma 6.4. *Seien $X_i, i\in\mathbb{N}$, unabhängige, gleichverteilte Zufallsvariablen mit Mittelwert μ und mit endlicher Varianz σ^2. Dann ist m_n ein Schätzer für μ und es gilt*

$$\mathbb{P}[|m_n-\mu|>c\mu]\leq\frac{\sigma^2}{n\mu^2c^2}. \tag{6.9}$$

Wir sehen, dass die Qualität des Schätzers erheblich von Verhältnis σ^2/μ^2 abhängt. In der Praxis will man sich ja eine gewisse Genauigkeit der Schätzung vorgeben, und dann n so wählen, dass diese erzielt wird. Dabei soll natürlich n so

klein wie möglich sein, da in der Regel die Durchführung eines Zufallsexperimentes Kosten verursacht.

Nun kennen wir natürlich μ und σ^2 nicht, wir wollen μ ja gerade bestimmen. Was μ angeht, ist das nicht so tragisch, da wir ja zumindest den Schätzer m_n haben. Allerdings reicht das noch nicht aus, um eine "Stoppregel" für das benötigte n zu entwickeln, da wir dazu auch σ^2 brauchen. Also sollten wir besser auch gleich versuchen, einen Schätzer für die Varianz zu finden und gleich mitzuberechnen. Naheliegend ist wieder die *empirische Varianz*, d.h. die Varianz der empirischen Verteilung ν_n:

$$V_n \equiv \nu_n(X - \nu_n(X))^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m_n)^2\,, \tag{6.10}$$

wobei $X = (X_1, \ldots, X_n)$. Wir zeigen zunächst, dass dieser Schätzer gegen die Varianz konvergiert, falls σ^2 endlich ist.

Lemma 6.5. *Seien $X_i, i \in \mathbb{N}$, wie in Lemma 6.4 und sei* $\mathrm{var}(X_i) = \sigma^2$. *Dann konvergiert die Zufallsvariable V_n gegen σ^2.*

Beweis. Zum Beweis schreiben wir V_n leicht um:

$$V_n = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - m_n^2.$$

Nach Voraussetzung sind die X_i^2 unabhängige, gleichverteilte Zufallsvariablen mit endlicher Erwartung. Daher konvergiert die erste Summe, wegen dem Gesetz der großen Zahlen,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mathbb{E}X_1^2 \quad .$$

Andererseits wissen wir, dass $m_n \to \mu$ und somit auch $m_n^2 \to \mu^2$. Daraus folgt, dass

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - m_n^2 \to \mathbb{E}X_1^2 - (\mathbb{E}X_1)^2 = \sigma^2,$$

was wir behauptet haben. □

Immerhin sehen wir, dass wir mit Hilfe unserer Schätzer m_n und V_n^* bereits ein praktisches Verfahren zur qualitätskontrollierten Schätzung des Mittelwertes haben. Dazu ersetzen wir in der Abschätzung (6.9) für die Wahrscheinlichkeit einer Abweichung des Schätzers m_n vom wahren Wert μ, die Größen μ und σ^2 durch ihre Schätzer. Dies liefert uns einen Schätzer für den wahren Fehler, der zumindest die gute Eigenschaft hat, gegen eine obere Schranke zu konvergieren. Damit liegt folgende Strategie nahe: Wir suchen einen Schätzer für μ, der mit höchstens Wahrscheinlichkeit ε um mehr als $c\mu$ falsch liegt. Dann berechnen wir sukzessive m_n, V_n bis zu einem Wert n^* wo erstmals

$$\frac{V_{n^*}^2}{n^* m_{n^*}^2 c^2} < \varepsilon.$$

6.2 Parameterschätzung

Wir hatten im vorigen Kapitel gesehen, wie das Gesetz der großen Zahlen verwendet werden kann um Schätzer sowohl für Wahrscheinlichkeitsverteilungen als auch Erwartungswert und Varianz zu konstruieren. Allerdings hatten wir auch gesehen, dass es schwierig und aufwendig ist, Wahrscheinlichkeitsverteilungen zu schätzen. Es wäre für praktische Zwecke wesentlich einfacher, wenn wir bereits a priori etwas über die Wahrscheinlichkeitsverteilung der zugrundeliegenden Zufallsvariablen wüssten, und nur noch einige wenige *Parameter* identifizieren müssten. Der zentrale Grenzwertsatz ist *ein* wesentliches Resultat, dass in gewissen Situationen solche von wenigen Parametern indizierten Klassen von Verteilungen suggeriert, hier nämlich gerade die *Gauß-Verteilung*. Nehmen wir etwa als Model an, dass X_i eine Familie von unabhängigen und identisch Gauß-verteilten Zufallvariablen sein, so bleiben als Parameter nur noch Mittelwert und Varianz zu schätzen, was wir bereit können.

Ein interessanteres Beispiel ist die sogenannte lineare Regression. Wir betrachten etwa einen zeitabhängigen Vorgang, $f(t) \in \mathbb{R}$, $t \in \mathbb{R}_+$, zu gewissen Zeiten $t_1 < t_2 < \cdots < t_n$. Jede Beobachtung liefert einen Messwert z_i. Idealerweise wäre $z_i = f(t_i)$, aber durch Fehler ist diese Gleichung verfälscht und wir sollen annehmen, dass die Differenz eine Zufallsvariable ist. Unsere Aufgabe ist, aus den Beobachtungen einen Schätzer für f zu gewinnen, und gleichzeitig eine Qualitätsabschätzung für den Schätzer, sowie einen Schätzer für die Verteilung der Fehler, finden.

Ohne weitere Vorabinformation ist dieses Problem praktisch unlösbar, da es unendlich viele Parameter involviert. Wir müssen also vereinfachende Annahmen machen. Zunächst betrachten wir den Fall, in dem wir annehmen, dass $f(t) = a + bt$ eine lineare Funktion ist, wobei a und b unbekannte, zu bestimmende Parameter sind. Weiter nehmen wir an, dass die Messfehler unabhängige, identisch verteilte Zufallsvariablen, X_i sind. Dann sind unsere Beobachtungen (im Rahmen des Modells) beschrieben als Zufallsvariablen

$$Z_i = a + bt_i + X_i. \tag{6.11}$$

Eine weitere Vereinfachung träte ein, wenn wie einschränkende Annahmen an die Verteilung der X_i machen könnten. Hier greift nun der zentrale Grenzwertsatz: wenn wir der Überzeugung sind, dass die Fehler X_i sich als Summen vieler kleiner "Elementarfehler", die unseren Messapparat beeinflussen, ergeben, dann liegt es nahe anzunehmen, dass die X_i Gauß-verteilt sind, mit unbekanntem Mittelwert, μ, und Varianz, σ^2. Wir haben also ein vierparametriges *Modell* für unsere Beobachtungen, mit Parametern a, b, μ, σ^2 (wobei wir leicht sehen, dass wir in unserem Fall zwischen a und μ nicht unterscheiden können, und daher nur hoffen können, dass $\mu = 0$, d.h. dass unsere Messungen keinen systematischen Fehler aufweisen). Die Aufgabe der Statistik ist es nun, Schätzer für diese Parameter zu finden (also Familien von Zufallsvariablen, die, wenn die Z_i durch dieses Modell beschrieben werden), gegen diese Parameter konvergieren. Eine solche Familie von Schätzern nennt man *konsistent*. Letzlich ist dies eigentlich noch nicht genug: wir würden auch gerne wissen, ob unsere Modellannahmen plausibel waren!

6.2.1 Die Methode der kleinsten quadratischen Abweichung

Ein bereits von Gauß vorgeschlagenes Vorgehen zur Lösung unseres Regresssionsproblems ist das *Prinzip der kleinsten quadratischen Abweichungen*. Bei n Beobachtungen definieren wir Σ_n durch

$$\Sigma_n^2(a,b) = \sum_{i=1}^{n} (a + bt_i - Z_i)^2 . \tag{6.12}$$

Dann sind die geschätzten Werte für a und b, $a^*, b*$, gerade die, die den Ausdruck (6.12) minimieren:

$$\Sigma_n^2(a^*,b^*) = \inf_{a,b\in\mathbb{R}} \Sigma_n^2(a,b). \tag{6.13}$$

Das Prinzip bedeutet im Wesentlichen, dass wir die die Regressionsparameter so wählen, dass die Varianz der Summe der Variablen X_i minimal wird. In der Tat ist dann der Schätzer für die Varianz der X_i nun gerade

$$\bar{\sigma}^2 = \frac{1}{n}\Sigma(a^*,b^*). \tag{6.14}$$

Interessanterweise kann das Prinzip der kleinsten quadratischen Abweichungen aus einem allgemeinen Wahrscheinlichleitstheoretisch motivierten Prinzip hergeleitet werden, wie wir im Folgenden sehen werden.

6.2.2 Das Maximum-Likelihood Prinzip

Eine einleuchtende Idee, wie wir aus einer Familie von statistischen Modellen das "besteäuswählen könnten ist die folgende. Wir berechnen für alle unsere Modelle, wie groß die Wahrscheinlichkeit der beobachteten Werte in diesem Modell ist. Dann wählen wir dasjenige, dass diese Wahrscheinlichkeit maximiert. Die Wahrscheinlichkeit der beobachteten Daten in einem gegebenen Modell bezeichnet man als *likelihood"*.

Betrachten wir dazu zunächst ein sehr einfaches Beispiel: Wir beobachten eine Folge von Münzwürfen, $z_1,\ldots,z_n \in \{0,1\}$. Wir wollen diese modellieren als Realisierung von unabhängigen, identisch verteilten Bernoulli Zufallsvariablen, X_i, mit Parameter p. Aus den Beobachtungen wollen wir nun den Wert von p schätzen. Das Maximum-likelihood Prinzip sagt, man schätze $p = p(z_1,\ldots,z_n)$, so dass die Wahrscheinlichkeit der Beobachtungen maximal wird, also, dass

$$\begin{aligned} \rho_n(p;z_1,\ldots,z_n) &\equiv \mathbb{P}[X_1 = z_1 \wedge X_2 = z_2 \wedge \cdots \wedge X_n = z_n] \\ &= \prod_{i=1}^{n} p^{z_i}(1-p)^{1-z_i} \end{aligned} \tag{6.15}$$

maximal wird. Wir nennen $\rho_n(p;z_1,\dots,z_n)$ die *likelihood Funktion* für unser Modell.

Um dasjenige p zu bestimmen, dass $\rho_n(p;z_1,\dots,z_n)$ maximiert, suchen wir zunächst einen kritischen Punkt dieser Funktion, d.h. wir lösen die Gleichung

$$\begin{aligned} 0 = \frac{\mathrm{d}}{dp}\rho_n(p;z_1,\dots,z_n) &= \sum_{i=1}^n \left(\frac{z_i}{p} - \frac{1-z_i}{1-p}\right) \prod_{i=1}^n p^{z_i}(1-p)^{1-z_i} \\ &= \rho_n(p;z_1,\dots,z_n) \sum_{i=1}^n \left(\frac{z_i}{p(1-p)} - \frac{1}{1-p}\right). \end{aligned}$$

Diese Gleichung hat als einzige Lösung

$$p = p_n^* = p_n^*(z_1,\dots,z_n) = \frac{1}{n}\sum_{i=1}^n z_i.$$

Da $z_i \in \{0,1\}$ liegen, ist $z_i = \mathbb{1}_{z_i=1}$, so dass der Maximum-Likelihood Schätzer für die Wahrscheinlichkeit von $\{X_i = 1\}$ gerade gleich der Frequenz des Auftretens von 1 ist, der uns ja schon als konsistenter Schätzer bekannt ist. In diesem Fall liefert das Maximum-likelihood Prinzip also nichts neues, gibt aber eine interessante alternative Interpretation des Schätzers.

Als nächstes betrachten wir das interessantere Beispiel der Regression in dem oben beschriebenen Gauß'schen Modell. Hier ist es allerdings so, dass wegen der Stetigkeit der Gauß-Verteilung die Wahrscheinlichkeit jeder Beobachtung gleich null ist. Anstatt der Wahrscheinlichkeit der Beobachtungen wählt man daher als "likelihood Funktion" die Wahrscheinlichkeitsdichte an den beobachteten Daten, also in unserem Fall

$$\begin{aligned} \rho_n(a,b,\sigma^2;z_1,\dots,z_n) &\equiv \prod_{i=1}^n \rho_{0,\sigma^2}(z_i - a - bt_i) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z_i-a-bt_i)^2}{2\sigma^2}\right). \end{aligned} \tag{6.16}$$

Das maximum-likelihood Prinzip sagt nun, dass der maximum-likelihood Schätzer für a,b,σ^2, $a_n^*, b_n^*, (\sigma^2)_n^*$, dadurch gegeben ist, dass

$$\rho_n(a_n^*, b_n^*, (\sigma^2)_n^*; z_1,\dots,z_n) \equiv \max_{a,b\in\mathbb{R},\sigma^2\in\mathbb{R}_+} \rho_n(a,b,\sigma^2;z_1,\dots,z_n) \tag{6.17}$$

In unserem Fall ist die Lösung des Maximierungsproblems recht einfach. Es empfiehlt sich, anstatt direkt ρ_n zu maximieren, dessen Logarithmus,

$$\ln\rho_n(a,b,\sigma^2;z_1,\dots,z_n) = -\frac{n}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^n \frac{(z_i-a-bt_i)^2}{2\sigma^2},$$

zu maximieren. Dies führt auf die drei Gleichungen

$$\frac{\partial \ln \rho_n}{\partial a} = 0 \leftrightarrow \sum_{i=1}^{n} (z_i - a - bt_i)/\sigma^2 = 0,$$
$$\frac{\partial \ln \rho_n}{\partial b} = 0 \leftrightarrow \sum_{i=1}^{n} t_i(z_i - a - bt_i)/\sigma^2 = 0,$$
$$\frac{\partial \ln \rho_n}{\partial \sigma^2} = 0 \leftrightarrow \sum_{i=1}^{n} (z_i - a - bt_i)^2/2\sigma^4 - \frac{n}{2\sigma^2} = 0.$$

Es folgt

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (z_i - a - bt_i)^2 \tag{6.18}$$

$$a = \frac{1}{n} \sum_{i=1}^{n} (z_i - bt_i) \tag{6.19}$$

$$b = \frac{\sum_{i=1}^{n} t_i(z_i - a)}{\sum_{i=1}^{n} t_i^2} \tag{6.20}$$

und weiter, mit $T_n = \sum_{i=1}^{n} t_i$,

$$b_n^* = \frac{\sum_{i=1}^{n} t_i z_i - \frac{T_n}{n} \sum_{i=1}^{n} z_i}{\sum_{i=1}^{n} t_i^2 - \frac{T_n^2}{n}}. \tag{6.21}$$

Nachdem b explizit bekannt ist kann nun a und σ^2 ebenfalls explizit durch Einsetzen ausgerechnet werden:

$$a_n^* = \frac{1}{n} \sum_{i=1}^{n} (z_i - b_n^* t_i), \tag{6.22}$$

$$(\sigma^2)_n^* = \frac{1}{n} \sum_{i=1}^{n} (z_i - a_n^* - b_n^* t_i)^2. \tag{6.23}$$

Wesentlich zu bemerken ist aber, dass die Gleichungen (6.19) und (6.20) besagen, dass a und b so gewählt werden müssen, dass der durch (6.18) gegebene Ausdruck für σ^2 als Funktion von a und b minimiert wird. Letzterer ist aber gerade die Summe der Quadrate der Abweichung des Beobachtung vom theoretischen Wert. Mit anderen Worten, die maximum-likelihood Methode liefert im Fall der Gauß-Verteilung gerade die Methode der *kleinsten quadratischen Abweichungen* für die Schätzung der Parameter a und b.

6.3 Hypothesentests

Mit dem Begriff der Likelihood"haben wir nun auch bereits ein Hilfsmittel gefunden um eine weitere wichtige Aufgabe der Statistik anzugehen, das Testen von Hypothesen. Ganz allgemein gesprochen geht es dabei darum eine Modellannahme über zufällige Ereignisse anhand von Beobachtungen entweder zu akzeptieren oder zu verwerfen.

Beispiel 1.

In einer Urne befinden sich N rote und N grüne Kugeln. Eine Person, die behauptet rot-grün blind zu sein, zieht M Kugeln. Von diesen Kugeln sind K rot. Sollen wir glauben, dass die Person die Wahrheit gesagt hat? Dazu stellen wir folgende Überlegung an. Wir gehen davon aus, dass die Kugeln in der Urne unabhängig von Ihrer Farbe rein zufällig angeordnet sind. Eine farbenblinde Person würde daher eine zufällige Teilmenge der Kardinalität M der $2N$ Kugeln auswählen. Wir müssen berechnen, was die Wahrscheinlichkeit ist, dass darin gerade K rote Kugeln liegen. Die Zahl der roten Kugeln, Z, ist aber gerade *hypergeometrisch verteilt*:

$$\mathbb{P}(Z=K)=\frac{\binom{N}{K}\binom{N}{M-K}}{\binom{2N}{M}}. \tag{6.24}$$

Dies kann man wie folgt verstehen. Die Anzahl der Teilmengen von M Kugeln ist $\binom{2N}{M}$ Wenn sich in dieser Teilmenge gerade K rote Kugeln befinden, so gab es $\binom{N}{K}$ Möglichkeiten, gerade K der roten Teilmenge und $\binom{N}{M-K}$ Möglichkeiten, $M-K$ der grünen Teilmenge auszuwählen.

Betrachten wir nun den Fall $N=10$, $M=10$ und $K=9$. Dann ist

$$\mathbb{P}(Z=9)=\frac{\binom{10}{9}\binom{10}{1}}{\binom{20}{10}}=\frac{100}{\binom{20}{10}}\approx\frac{100}{2^{20}}\approx 10^{-10}. \tag{6.25}$$

Dies ist also äusserst unwahrscheinlich, wenn die Person wirklich farbenblind war. Es liegt daher nahe, die Hypothese zu verwerfen. Es dies sinnvoll? Dazu muss man sich zunächst fragen, ob dann ein anderer Ausgang des Experiments wesentlich wahrscheinlicher gewesen wäre. Was wäre die Wahrscheinlichkeit sagen wir von 5 roten Kugeln gewesen?

$$\mathbb{P}(Z=5)=\frac{\binom{10}{5}\binom{10}{5}}{\binom{20}{10}}=\frac{100}{\binom{20}{10}}\approx\frac{\sqrt{20}}{10\sqrt{2\pi}}.. \tag{6.26}$$

was um Größenordnungen größer ist.

Beispiel 2.

Eine wichtige Testfrage beim Wetten ist, ob eine Glückspielmaschine faire Resultate liefert. Der einfachste Fall ist etwa, ob in einer Münze Kopf und Zahl tatsächlich mit gleicher Wahrscheinlichkeit vorkommen. Allgemeiner kann man die Frage formulieren, wie man feststellen kann, ob unabhängige Zufallsvariablen X_i wirklich einen versprochenen Mittelwert μ haben. Dazu betrachten wir wieder n unabhängige Realisierungen dieser Zufallsvariablen, $X_1, \ldots, X_n$. Wir wissen bereits, dass ein Schätzer für den Mittelwert durch $\hat{\mu}_n \equiv n^{-1} \sum_{i=1}^n X_i$ gegeben ist. Wenn der beobachtete Schätzer $\hat{\mu}_n$ von μ abweicht, so kann das zwei Ursachen haben: Die Zufallsvariablen haben gar nicht den Mittelwert μ, oder sie haben diesen Mittelwert, und die Abweichung ist rein zufällig. Wir hatten bereits gesehen, dass die Wahrscheinlichkeit der zufälligen Abweichung durch die Varianz kontrolliert ist, und dass die Varianz wiederum durch

$$\hat{\sigma}_n^2 \equiv n^{-1} \sum_{i=1}^n (X_i - \hat{\mu}_n)^2 \tag{6.27}$$

geschätzt werden kann. Man führt nun die Größe

$$T_n \equiv \sqrt{n} \frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n}, \tag{6.28}$$

welche die Abweichung des beobachteten Mittelwerts von dem erwarteten Mittelwert unter der Hypothese, dass dieser μ ist mit der beobachteten Standardabweichung des Mittelwerts vergleicht. Typischerweise sollte dann T von der Ordnung 1 sein, wenn die Hypothese richtig ist. Man kann nun einen Schwellenwert für T festlegen, bis zu dem man die Hypothese akzeptieren will, und oberhalb dessen man sie ablehnt. Man kann die Wahrscheinlichkeitsverteilung von T_n mittels des zentralen Grenzwertsatzes und des Gesetzes der großen Zahlen für große n auch leicht aproximieren: Wenn wir die Modellannahme treffen, dass die X_i unabhängig mit Mittelwert μ und endlicher Varianz σ^2 sind, dann konvergiert nach dem GGZ $\hat{\sigma}_n$ gegen σ und nach dem ZGS T_n gegen eine Gauß-verteilte Zufallsvariable mit Mittelwert 0 und Varianz 1. Also haben wir, dass

$$\mathbb{P}(|T_n| > t) \approx \phi(t) \approx \frac{2}{\sqrt{2\pi t}} \exp(-t^2/2). \tag{6.29}$$

Allgemeine Prinzipien.

Das grundsätzliche Vorgehen beim Testen von Hypothesen können wir wie folgt beschreiben. Beim Testen geht es letztlich immer darum, zu entscheiden, of wir ein bestimmtes statistisches Modell akzeptieren wollen oder nicht. Dazu betrachten wir eine Klasse statistischer Modelle, die durch eine Familie von Parametern, $\theta \in \Theta$ parametrisiert ist. Im vorigen Beispiel wäre θ etwa der Erwartungswert μ. Wir nen-

nen das Wahrscheinlichkeitsmaß des Modells mit Parameter θ $\mathbb{P}^\theta$. Eine Hypothese wird nun dadurch formuliert, dass man den Parameterbereich Θ disjunkt zerlegt in $\Theta = \Theta_0 \cup \Theta_1$. Die Aussage $\theta \in \Theta_0$ nennt man dann Hypothese, bzw. *Nullhypothese*. Die Wahl der Nullhypothese beruht auf a priori Vorwissen bzw. Vorurteilen. Beim Münzwurf wäre dies etwa die Annahme, dass die Münze an sich fair sein sollte. Diese Hypothese wollen wir nur verwerfen, wenn wir eine hinreichend grosse Inkompatibilität mit den Beobachtungen feststellen.

Ein *Test* ist nun (im einfachsten Fall) eine Teilmenge $H \in \mathfrak{F}$ in unserem Raum der möglichen Ausgänge unseres Zufallsexperiments, also ein Ereignis (allgemeiner können auch Tests aus Zufallsvariablem mit Werten in $[0,1]$ konstruiert werden). Beobachten wir nun einen Ausgang, X, unseres Zufallsexperiments, so wollen wir die Entscheidung über die Nullhypothese wie folgt treffen:

- Falls $X \in H$, so akzeptieren wir die Nullhypothese.
- Falls $X \notin H$, so verwerfen wir die Nullhypothese.

Es kann dabei zu zweierlei Fehlentscheidungen kommen:

- Wir können die Hypothese verwerfen, obwohl sie zutrifft. Dies nennt man einen *Fehler erster Art*, oder
- wir können die Hypothese akzeptieren, obwohl sie falsch ist. Dies nennt man einen *Fehler zweiter Art*.

Ein Test sollte idealerweise die Wahrscheinlichkeit für beide Arten von Fehlern klein machen. Aufgrund der Tatsache, dass die Nullhypothese zunächst ja plausibel erscheint, ist es allerdings zunächst wichtig, die Wahrscheinlichkeit eines Fehlers erster Art klein zu machen. Dazu muss man die größtmögliche Wahrscheinlichkeit dafür, dass $X \notin H$ unter der Nullhypothese klein machen.

Definition 6.6. Ein Ereignis $H \in \mathfrak{F}$ heisst Test der Nullhypothese $\theta \in \Theta_0$ *zum Niveau* α, wenn

$$\sup_{\theta \in \Theta_0} \mathbb{P}^\theta(H^c) \leq \alpha. \tag{6.30}$$

Das Niveau des Tests ist also die Wahrscheinlichkeit, einen Fehler erster Art zu machen. Die Wahl des Schwellenwerts α ist natürlich entscheidend und muss sich aus dem Testproblem ergeben. In der Praxis wird gerne $\alpha = 0.05$ gewählt.

Die Wahrscheinlichkeit einen Fehler zweiter Art zu machen, ist komplizierter zu finden. Er tritt ein wenn die Hypothese falsch ist und in dem richtigen Modell"das Ereignis H auftreten kann. Man fragt sich nun, wie wahrscheinlich dies ist. Dazu definiert man die sog. *Gütefunktion* $G_H(\theta) \equiv \mathbb{P}^\theta(H^c)$. Man nennt $G_H(\theta)$ für $\theta \in \Theta_1$ auch die *Macht* oder *Schärfe* des Tests bei θ. In der Tat ist $G_H(\theta)$ die Wahrscheinlichkeit dafür, dass $Y \notin H$, also dass die Nullhypothese verworfen wird, wenn der richtige Parameter θ ist. Umgekehrt ist dann $\beta_H(\theta) = 1 - G_H(\theta)$ die Wahrscheinlichkeit dafür, einen Fehler zweiter Art zu begehen.

Definition 6.7. Ein Test der Nullhypothese $\theta \in \Theta_0$ zum Niveau α heisst *gleichmäßig bester* Test zum Niveau α, falls für alle anderen Tests, H' zum Niveau α gilt, dass

$$G_H(\theta) \geq G_{H'}(\theta), \quad \text{für alle} \theta \in \Theta_1. \tag{6.31}$$

Beispiel: T-Test.

Wir kommen zu unserem Beispiel des Testens der Hypothese $\mathbb{E}X_i = \mu$ zurück. Hier ist unser Parameter $\theta = \mathbb{E}X_i$. Wir betrachten wir zunächst die Nullhypothese mit $\Theta_0 = \{\mu\}$, $\Theta_2 = \mathbb{R}\backslash\{\mu\}$. Als Test bieten sich die Mengen $H_t \equiv \{|T_n| \leq t\}$, für $t > 0$ an. Nach unseren früheren Überlegungen wissen wir, dass

$$\mathbb{P}^\mu(H_t^c) = \mathbb{P}^\mu(|T_n| > t) \approx \frac{2}{\sqrt{2\pi}} \int_t^\infty \mathrm{e}^{-x^2/2} dx = 2\phi(t) \tag{6.32}$$

ist. Damit können wir also leicht durch geeignet Wahl von t einen Test zu jedem gewünschten Niveau konstruieren. Für einen Test auf 5% etwa wählen wir dazu t_c so, dass $2\phi(t_c) = 0.05$, und unser Testereignis $H_{t_c} = \{|T_n| \leq t_c\}$ is dann der gewünschte Test.

Wie sieht es mit der Güte des Tests aus? Falls $\theta = \mu + \delta$, so ist unter $\mathbb{P}^\theta$,

$$T_n = \sqrt{n}\delta/\hat{\sigma}_n + T_n', \tag{6.33}$$

wo T_n' wieder approximativ normal-verteilt ist (unter $\mathbb{P}^\theta$). Damit ist unsere Gütefunktion approximativ (o.b.d. A. sei $\delta > 0$)

$$G_{H_t}(\theta) \approx 1 - \mathbb{P}^\theta(|T_n' - \sqrt{n}\delta/\sigma_n| \leq t) \approx 1 - \frac{1}{\sqrt{2\pi}} \int_{\sqrt{n}\delta/\sigma - t}^{\sqrt{n}\delta/\sigma + t} \mathrm{e}^{-x^2/2} dx. \tag{6.34}$$

Das Integral ist klein, wenn das Integrationsintervall weit von Null weg liegt, es ist aber nahe bei eins, wenn es die Null umfasst. Für den letzteren Fall muss aber $\sqrt{n}\delta/\sigma$ klein sein, d.h. $\delta \sim 1/\sqrt{n}$. Die Wahrscheinlichkeit eines Fehlers zweiter Art wird also nur dann nicht sehr klein, wenn Insbesondere bedeutet dies, dass für große n ein Fehler zweiter Art nur dann mit grosser Wahrscheinlichkeit auftritt, wenn $\delta \leq O(1/\sqrt{n})$ ist. Wir sehen an diesem Beispiel, dass das Testen der Hypothese $\mathbb{E}X_i = \mu$ gegen die Alternative $\mathbb{E}X_i \neq \mu$ problematisch ist, da es sehr wahrscheinlich ist, einen Fehler zweiter Art zu begehen, wenn der wirkliche Erwartungswert sehr nahe bei μ liegt. Die Gütefunktion gibt uns den Hinweis, welche weiteren Werte ausser der Nullhypothese ebenfalls nicht one weiteres verworfen werden sollten.

Die oben betrachtete Hypothese der Form $\Theta_0 = \{\mu\}$ nennt man eine *zweiseitige Hypothese*. Im Gegensatz dazu nennt man Hypothesen der Form $\theta \leq \mu$ oder $\theta \geq \mu$ *einseitige Hypothesen*.

Im obigen Beispiel ist eine einseitige Hypothese etwa $\theta \leq \mu$, wobei θ der Erwartungswert von X_i ist. Unter der Hypothese sollte also das beobachtete $\hat{\mu}_n$ eher kleiner als μ sein. Daher wäre ein Test der Hypothese $\theta \geq \mu$ z.B. das Ereignis $H_t \equiv \{T_n \leq t\}$, wobei T_n wie in (6.28) gegeben ist. Damit H_t die Hypothese zum Niveau α tested, muss dann gelten, dass

$$\max_{\theta \leq \mu} \mathbb{P}^\theta(T_n > t) \leq \alpha. \tag{6.35}$$

Nun wissen wir wieder, dass für grosse n approximativ gilt

$$\mathbb{P}^\theta(T_n > t) = \mathbb{P}(\sqrt{n}((\theta-\mu)/\hat{\sigma}_n + T_n' > t) \approx \phi(t + \sqrt{n}(\mu-\theta)/\hat{\sigma}). \tag{6.36}$$

Dann ist aber

$$\max_{\theta\le\mu} \mathbb{P}^\theta(T_n > t) \approx \max_{\theta\le\mu} \phi(t + \sqrt{n}(\mu-\theta)/\hat{\sigma}) = \phi(t). \tag{6.37}$$

In vielen Situationen, wie etwa der obigen, hat man nicht nur einen Test auf ein bestimmtes Niveau, sondern eine Familie von Tests auf beliebige Niveaus. Einer gegebenen Beobachtung Z kann man dann den sogenannten p-Wert zuordnen, nämlich

$$p(Z) = \min\{\mathbb{P}(H_t^c, t > 0 : Z \in H_t^c\}, \tag{6.38}$$

das heisst, das Niveau des schärfsten Test (Test zum niedrigsten Niveau), bei dem die Nullhypothese verworfen worden wäre. Es ist klar, dass je kleiner der p-Wert ist, um so mehr Vertrauen wird man in eine negatives Testergebnis haben. Der p-Wert wird auch ërreichtes Niveau"(level attained") gennant. Er beschreibt in gewissem Sinn die Unwahrscheinlichkeit der Beobachtung unter der Nullhypothese.

Anmerkung. Zur Notation: Das Niveau zu dem ein Test tested nennt man auch *Signifikanzniveau*. Eine Test zu einem Niveau nennt man auch *Signifikanztest*. Gelegentlich wird ein Test auch *Teststatistik* genannt.

6.4 Stichproben

.

Das vielleicht klassischste Problem der Statistik besteht darin, Aussagen über den Zustand eines Systems zu treffen, das nicht vollständig beobachtet werden kann. Oft geht es dabei darum, die Zusammensetzung einer grossen Gruppe von Individuen, die unterschiedliche Eigenschaften haben, zu schliessen. Dabei möchte man eine kleine Untergruppe (*Stichprobe*) auswählen, deren Zusammensetzung messen, und darauf auf die gesamte Gruppe schliessen. Dies geschieht etwa bei Prognosen zum Wahlausgang bei Wahlen und anderen Meinungsumfragen. Ein anderes Beispiel ist etwa das Problem, die Häufigkeit verschiedenen Spezies von Planzen oder Tieren in einem Biotop zu bestimmen.

Wir betrachten hier die folgende einfache Situation. Gegeben seien N Personen, die entweder die Meinung A oder die Meinung B haben können. Die unbekannt Anzahl der Personen mit Meinung A bezeichnen wir mit N_A. Wir wollen diese Zahl bestimmen, indem wir eine (kleine) Zahl n von Personen zufällig aus den N Personen auswählen und diese Nach Ihrer Meinung fragen. Die Gruppe der n Personen bezeichnen wir als *Stichprobe*. Die Anzahl der Personen in der Stichprobe, die die Meinung A haben bezeichnen wir mit n_A. Wir betrachten Was lernen wir über N_A?

Wir wollen dieses Problem auf zwei Wegen angehen, einmal als Hypothesentest, und einmal als Bayes'schen Schätzproblem.

6.4.1 Stichproben als Hypothesentest

Der Zugang über Hypothesentests bietet sich dann an, wenn wir a priori eine Vermutung über die Zusammensetzung unserer Guppe haben. Angenommen, wir haben die Meinung, dass mindestens pN (für ein $p \in (0,1)$) Personen die Meinung A haben. Dann formulieren wir die
Nullhypothese: $\{N_A \geq pN\}$.

WIr suchen einen Test in der form $\{n_A \geq tn\}$. Wie müssen wir t und n wählen, um die Nullhypothese auf dem Niveau 0.95 zu testen?

Dazu benötigen wir die Wahrscheinlichkeit des Stichprobenausgangs unter der Bedingung, dass $N_A = K$. Nach unserem allgemeinen Prinzip muss gelten, dass

$$\max_{K \geq pN} \mathbb{P}\left(n_A < tn \mid N_A = K\right) \leq 0.05. \tag{6.39}$$

Nehmen wir $N_A = K$ als gegeben an, so ist unsere Stichprobenauswahl ein kontrolliertes Zufallsexperiment, und die Wahrscheinlichkeit den Wert $n_A = k$ zu erhalten, wenn $N_A = K$ gegeben ist, ist hypergeometrisch (siehe Beispiel 1):

$$\mathbb{P}(n_A = k | N_A = K) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}. \tag{6.40}$$

Wir müssen also sicherstellen, dass

$$\sum_{k=0}^{tn-1} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \leq 0.05, \tag{6.41}$$

für alle $K \geq pN$. Um besser zu verstehen, was das uns sagt, nehmen wir an, dass N sehr gross gegen n und auch K gross gegen k ist. Dann können wir die Binomialkeffizienten approximieren als

$$\binom{K}{k} \approx \frac{K^k}{k}, \tag{6.42}$$

und so weiter. (Der Fehler, den wir dabei machen ist ein Faktor $\prod_{i=1}^{K}(1-i/K) \geq (1-k/K)^k \sim \mathrm{e}^{-k^2/K}$, was sehr nahe bei eins liegt, wenn k^2 klein gegen K ist.

$$\mathbb{P}(n_A = k | N_A = K) = \binom{n}{k}\left(\frac{K}{N}\right)^k\left(1-\frac{K}{N}\right)^{n-k}, \tag{6.43}$$

d.h. unsere Wahrscheinlichkeit kann durch eine Binomialverteilung $Bin(n, K/N)$ approximiert werden, was zu der Bedingung

$$\sum_{k=0}^{tn-1}\binom{n}{k}(K/N)^k(1-K/N)^{n-k} \leq 0.05 \tag{6.44}$$

führt. Klarerweise kann dies nur klein sein, wenn $k/n < K/N$ für alle betrachteten Werte von k und K ist. Aufgrund der Monotonieeigenschaften reduzieren sich unsere Bedingungen daher auf

$$\sum_{k=0}^{tn-1} \binom{n}{k} p^k (1-p)^{n-k} \leq 0.05. \tag{6.45}$$

Hier können wir nun aber wieder den Satz von de Moivre-Laplace anwenden. Wir wählen $t = t_c \equiv p - z/\sqrt{n}$. Dann sagt unser Satz, dass der uns sagt, dass für n einigermassen groß, die linke Seite gut durch

$$\frac{1}{\sqrt{2\pi p(1-p)}} \int_{-\infty}^{-z} \mathrm{e}^{-x^2/2p(1-p)} dx = \phi(z/\sigma), \tag{6.46}$$

approximiert wird, wobei wieder $\sigma^2 \equiv p(1-p)$ ist. Also erhalten wir einen Test auf das Niveau 0.95 wenn wir $t = p - 2\sigma/\sqrt{n}$ wählen.

Wie steht es nun mit der Güte dieses Tests? Die Gütefunktion ist in unserem Fall für $K' < pN$

$$G_H(K') = \mathbb{P}(n_A < t_c n | N_A = K') \approx \sum_{k=0}^{t_c n-1} \binom{n}{k} \left(K'/N\right)^k \left(1 - K'/N\right)^{n-k}. \tag{6.47}$$

Wenn wir wieder mit dem zentralen Grenzwertsatz approximieren, erhalten wir

$$\begin{aligned} G_H(K') = \mathbb{P}(n_A < t_c n | N_A = K') &\approx \frac{1}{\sqrt{2\pi p'(1-p')}} \int_{-\infty}^{-2\sigma+\sqrt{n}(p-K'/N)} \mathrm{e}^{-x^2/sp(1-p)} dx \\ &= \phi(2 - \sqrt{n}(p - K'/N)\sigma). \end{aligned} \tag{6.48}$$

Die rechte Seite der Gleichung erreicht den Wert $1/2$ wenn $\sqrt{\sigma}(p - K'/N)\sqrt{n} = 2$, also für $K' = pN - \sigma N/(\sigma\sqrt{n})$. Wir sehen daher, dass Güte des Tests erst dann gegen eins (und damit die Wahrscheinlichkeit einen Fehler zweiter Art zu machen gegen Null geht), wenn der wahre Wert von N_A um einen Betrag von der Ordnung $N/\sqrt{n}$ unter dem der Nullhypothese liegt.

Genauso lassen sich natürlich auch andere Hypothesen über N_A testen.

6.4.2 Stichproben als Bayes'scher Schätzer

In vielen Fällen hat man keine a priori Hypothese sondern möchte aus der Stichprobe den Wert von N_A schätzen. Genau genommen möchte man sogar eine Wahrscheinlichkeitsverteilung für N_A schätzen. Nun ist hier aber eigentlich nichts zufällig (und insbesondere ist nichts wiederholbar), da wir ja nur ein einziges System betrachten und dort N_A eben einen bestimmten Wert hat. Um dennoch probabilistisch argumentieren zu können, nehmen wir den sog. Bayes'schen Standpunkt ein in dem

Wahrscheinlichkeit einfach ein Maß für unser Unwissen ist. Dazu sehen wir auch die gesamte Anzahl der Personen, die die Meinung N_A haben, als Zufallsvariable an. A priori wissen wir nur, dass N_A einen Wert zwischen Null und N annimmt, und nehmen daher als a priori Verteilung $\mathbb{P}(N_A = K) = \frac{1}{N+1}$, für $K \in \{0,\ldots,N\}$. Nun führen wir auf demselben Wahrscheinlichkeitsraum die Zufallsvariable n_A ein, also das Ergebnis unserer Stichprobe. Was uns nun interessiert, ist die Wahrscheinlichkeitsverteilung von N_A, gegeben den beobachteten Wert k der Zufallsvariablen n_A. Nun kennen wir aber die bedingte Verteilung von n_A, gegeben den Wert von N_A, siehe (6.40). Nach Bayes haben wir dann

$$\mathbb{P}(N_A = K|n_A = k) = \mathbb{P}(n_A = k|N_A = K)\frac{\mathbb{P}(N_A = K)}{\mathbb{P}(n_A = k)}. \tag{6.49}$$

$\mathbb{P}(N_A = K)$ kennen wir aus unserer Grundannahme der Gleichverteilung. $\mathbb{P}(n_A = k)$ erhalten daraus (nach Satz 3.2, Punkt (ii)), als

$$\begin{aligned}\mathbb{P}(n_A = k) &= \sum_{K=0}^{N} \mathbb{P}(n_A = k|N_A = K)\mathbb{P}(N_A = K) \\ &= \sum_{K=0}^{N} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}\frac{1}{N+1}.\end{aligned} \tag{6.50}$$

Nun gilt aber

$$\sum_{K=0}^{N} \binom{K}{k}\binom{N-K}{n-k} = \binom{N+1}{n+1}, \tag{6.51}$$

und daher

$$\mathbb{P}(n_A = k) = \frac{\binom{N+1}{n+1}}{(N+1)\binom{N}{n}} = \frac{1}{n+1}, \tag{6.52}$$

d.h. auch die Zufallsvariablen n_A sind a priori gleichverteilt auf Ihrem Wertebereich.

Anmerkung. Die Tatsache, dass, unter der der a priori Annahme, dass wir nichts wissen, auch die Zufallsvariable n_A gleichverteilt sein soll, scheint sehr plausibel und könnte fast als Beweis der Gleichung (6.51) dienen. Es gibt allerdings auch einen rein analytischen Beweis dieser Relation.

Dann erhalten wir für die bedingte Verteilung von K

$$\mathbb{P}(N_A = K|n_A = k) = \frac{n+1}{N+1}\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}. \tag{6.53}$$

Nehmen wir wieder $N \gg n$ etc. an, so können wir wieder approximieren:

$$\begin{aligned}\mathbb{P}(N_A = K | n_A = k) &= \frac{n+1}{N+1}\binom{n}{k}\left(\frac{K}{N}\right)^k\left(1-\frac{K}{N}\right)^{n-k} \\ &= \frac{n+1}{N+1}\binom{n}{k}\exp\left(n\left(\frac{k}{n}\ln(K/N)+(1-\frac{k}{n})\ln(1-K/N)\right)\right) \\ &= \frac{n+1}{N+1}\binom{n}{k}\exp\left(nF(k/n,K/N)\right),\end{aligned} \tag{6.54}$$

Die Funktion $F(\beta,\alpha) = \beta\ln\alpha + (1-\beta)\ln(1-\alpha)$ strebt bei $\alpha = 0$ und 1 gegen $-\infty$. Sie ist als Funktion von α konkav und nimmt ihr einziges Maximum bei $\alpha = \beta$ an, und es gilt $F(\beta,\beta) = \beta\ln\beta + (1-\beta)\ln(1-\beta)$. Es folgt, dass der wahrscheinlichste Wert von K gerade $K = k\frac{N+1}{n+1}$ ist, was wir auch erwarten sollten. Damit haben wir einen Schätzer für K gefunden.

Wir können weiter die Funktion $F(\beta,\alpha)$ in der Nähe ihres Maximums Taylorentwickeln und erhalten

$$F(\beta,\alpha) = F(\beta,\beta) - \frac{1}{2}\frac{1}{\beta(1-\beta)}(\alpha-\beta)^2 + O(|\alpha-\beta|^3). \tag{6.55}$$

Für kleine Werte von $\beta-\alpha$ können wir die Fehlerterme vernachlässigen. Wenn wir $K/N - k/n = x/\sqrt{n}$ setzen (also Werte von K betrachten, die nur um $xN/\sqrt{n}$ vom Schätzer kN/n, so finden wir

$$\mathbb{P}(N_A = K | n_A = k) \approx \frac{n}{N}\binom{n}{k}\exp\left(nF(k/n,k/n) - \frac{x^2}{2k/n(1-k/n)}\right). \tag{6.56}$$

Hier sehen wir schon klar die Gauß-Verteilung im Anzug. In der Tat, wenn wir den Binomialkoeffizienten wir in Gleichung (5.11) approximieren, so erhalten wir

$$\mathbb{P}(N_A = K | n_A = k) \approx \frac{\sqrt{n}}{N}\frac{1}{\sqrt{2\pi k/n(1-k/n)}}\exp\left(-\frac{x^2}{2k/n(1-k/n)}\right). \tag{6.57}$$

Daraus folgt aber (ähnlich wie im Beweis des Satzes von de Moivre-Laplace) folgendes Resultat zeigen. Wir definieren ein Familie von Zufallsvariablen $Z_{N,n} \equiv \sqrt{n}\,(N_A/N - n_A/n)$. Dann gilt:

Satz 6.8. *Für jedes $\beta \in (0,1)$ und $a < b \in \mathbb{R}$ gilt*

$$\lim_{n\uparrow\infty}\lim_{N\uparrow\infty}\mathbb{P}\left(a \le Z_{N,n} \le b \mid n_A = [\beta n]\right) = \frac{1}{\sqrt{2\pi\beta(1-\beta)}}\int_a^b \mathrm{e}^{-\frac{x^2}{2\beta(1-\beta)}}\,dx. \tag{6.58}$$

Die wesentliche Botschaft dieses Resultats ist, dass das Ergebnis einer Stichprobe der Größe n mit k mal der Meinung A uns sagt, dass mit Wahrscheinlichkeit 0.95 in die relative Häufigkeit der Meinung A im Intervall $[k/n - 2\sigma/\sqrt{n}, k/n + 2\sigma/\sqrt{n}]$ liegt, wo $\sigma^2 = \beta(1-\beta)$ ist, wenn $N \gg n$. Die absolute Zahl der Personen mit Meinung A bestimmt unsere Stichprobe dagegen mit einem Fehler der Ordnung $N/\sqrt{n}$. Ist etwa $N = 10^8$ und wir wollen die Häufigkeit der Meinung A auf 1% mit Konfi-

denz 0.95 bestimmen, so müssen wir $n \sim 10000$ wählen. Die Bayes'sche Analyse erlaubt also auf der Basis einer Stichprobe Konfidenzintervalle für den Wert der Anzahl der Personen mit Meinung A in der Gesamtpopulation zu finden. Beachte, dass die Größe des Konfidenzintervalls für die Häufigkeit N_A/N stets von der Ordnung $1/\sqrt{n}$ ist (unabhängig von der Größe von N, vorausgesetzt diese ist groß gegen n.

6.5 χ^2-Anpassungstests

Wir betrachten wieder die Situation eines statistischen Modells wo $X_i, i \in \mathbb{N}$ unabhängige identisch verteilte Zufallsvariablem sind. Das Modell sei parametrisiert durch die Verteilung der Zufallsvariablen X_1, θ. Es sei zunächst der Fall angenommen, wo X_i endlich viele Werte, $\{1,\ldots,\ell\}$ annehmen. Die Hypothesen, die wir betrachten wollen sind $\theta = P$, für eine bestimmte Verteilung P, z. B. die Gleichverteilung. Wir suchen nun nach einem geeigneten Test zu einem Niveau α.

Dazu betrachten wir n Beobachtungen der Zufallsvariablen $X_1,\ldots,X_n$ und die entsprechende empirische Verteilung

$$\nu_n \equiv \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}. \tag{6.59}$$

Wir betrachten den Vektor der Frequenzen

$$L_n \equiv (\nu_n(1),\ldots,\nu_n(\ell)). \tag{6.60}$$

Wir wollen L_n mit dem Wahrscheinlichkeitsvektor

$$(P(1),\ldots,P(\ell)), \tag{6.61}$$

vergleichen. Wir können nun die Likelihood der beobachteten Frequenzen unter der Nullhypothese,

$$\prod_{i=1}^{\ell} P(i)^{n\nu_n(i)}, \tag{6.62}$$

mit der maximalen Likelihood unter der Alternative, $\theta \neq P$, vergleichen. Nun ist aber

$$\sup_{\theta\neq P}\prod_{i=1}^{\ell} \theta(i)^{n\nu_n(i)} = \prod_{i=1}^{\ell} \nu_n(i)^{n\nu_n(i)}. \tag{6.63}$$

Betrachten wir nun das Verhältnis dieser Werte,

$$R_n \equiv \frac{\prod_{i=1}^{\ell} \nu_n(i)^{n\nu_n(i)}}{\prod_{i=1}^{\ell} P(i)^{n\nu_n(i)}}, \tag{6.64}$$

Es folgt dass

$$n^{-1}\ln R_n = \sum_{i=1}^{\ell} \nu_n(i)\ln\left(\frac{\nu_n(i)}{P(i)}\right) \equiv H(\nu_n, P). \tag{6.65}$$

Die Grösse $H(\nu_n, P)$ nennt man die *relative Entropie* der Maße ν_n und P. Sie verschwindet genau dann, wenn $P = \nu_n$. Ihre Grösse ist ein Maß für die Abweichung von ν_n von P. Wir könnten also eine Test von der Form $\{nH(\nu_n, P) \le c\}$ betrachten. Dann müssen wir nur noch wissen, wie c zu wählen ist, um ein gewünschtes Testniveau zu erzielen.

Dazu schreiben wir die relative Entropie geschickt um. Es gilt nämlich

$$\begin{aligned} H(\nu_n, P) &= \sum_{i=1}^{\ell} P(i)\frac{\nu_n(i)}{P(i)}\ln\left(\frac{\nu_n(i)}{P(i)}\right) \\ &= \sum_{i=1}^{\ell} P(i)\left(1 - \frac{\nu_n(i)}{P(i)} + \frac{\nu_n(i)}{P(i)}\ln\left(\frac{\nu_n(i)}{P(i)}\right)\right) \\ &\equiv \sum_{i=1}^{\ell} P(i)\psi\left(\frac{\nu_n(i)}{P(i)}\right), \end{aligned} \tag{6.66}$$

mit $\psi(u) = 1 - u + u\ln u$. Dabei haben wir benutzt, dass offensichtlich

$$\sum_{i=1}^{\ell} P(i)\left(1 - \frac{\nu_n(i)}{P(i)}\right) = \sum_{i=1}^{\ell} P(i) - \sum_{i=1}^{\ell} \nu_n(i) = 1 - 1 = 0 \tag{6.67}$$

ist. Nun ist $\frac{d}{du}\psi(u) = -1 + \ln u + 1 = \ln u$, so dass ψ bei 1 ihr einziges Minimum, 0, annimmt. Weiter ist $\frac{d^2}{du^2}\psi(1) = 1$, so dass die Taylorapproximation um 1 gegeben ist durch $\psi(u) \approx (u-1)^2/2$. Damit erhalten wir für die relative Entropie die Approximation

$$H(\nu_n, P) \approx \frac{1}{2}\sum_{i=1}^{\ell} P(i)\left(\frac{\nu_n(i)}{P(i)} - 1\right)^2 \equiv \frac{1}{2n} D_{n,P}. \tag{6.68}$$

Alternativ schreibt man oft auch um:

$$D_{n,P} = \sum_{i=1}^{\ell} \frac{\left(n\nu_n(i) - nP(i)\right)^2}{nP(i)} \tag{6.69}$$

Das sieht an sich schon ganz gut aus als Maß für die Abweichung von ν_n von P. Der entscheidende Punkt ist aber, dass für die rechte Seite unter der Nullhypothese ein universeller Grenzwertsatz, ähnlich dem zentralen Grenzwertsatz, gilt.

Satz 6.9. *Für jedes ℓ und für jedes P so, dass $P(i) > 0$, für alle $i \in \{1, \ldots, \ell\}$ und für alle $c > 0$ gilt*

$$\lim_{n\uparrow\infty} \mathbb{P}^P\left(D_{n,P} \le c\right) = \chi^2_{\ell-1}(c), \tag{6.70}$$

wobei $\chi^2_{\ell-1}$ die Verteilungsfunktion der χ^2-Verteilung mit $\ell-1$ Freiheitsgraden ist. Dabei ist

$$\chi^2_k(c) = \mathbb{P}\left(\sum_{i=1}^{k} Z_i^2 \leq c\right), \tag{6.71}$$

wo $Z_i, i \in \mathbb{N}$, unabhängige Gauß-verteilte Zufallsvariablen mit Mittelwert 0 und Varianz 1 sind.

Anmerkung. Da in der Formel für $D_{n,P}$ ℓ Quadrate auftreten, könnte man zunächst vermuten, dass in der χ^2-Verteilung ℓ Quadrate von Gauß'schen Zufallsvariablen auftreten sollten. Der Punkt ist aber, dass die Summanden in (6.68) nicht unabhängig sind, da ja stets $\sum_{i=1}^{\ell} \nu_n(i) = 1$, etc.. Daher bleiben im Grenzwert nur $\ell-1$ FFreiheitsgrade"übrig.

Der Beweis dieses Satzes würde hier zu weit führen. Wir beschränken uns auf den Fall $\ell = 1$, der eine direkte Folge des Satzes von de Moivre-Laplace ist.

Beweis. (Im Fall $\ell = 1$). In diesem Fall gilt, dass

$$\begin{aligned} D_{n,P} &= \sum_{i=1}^{2} \frac{1}{nP(i)} \left(n\nu_n(i) - nP(i)\right)^2 \qquad (6.72)\\ &= \left(\frac{1}{nP(1)} + \frac{1}{n(1-P(1))}\right)\left(n\nu_n(1) - nP(1)\right)^2 = \frac{\left(n\nu_n(1) - nP(1)\right)^2}{nP(1)(1-P(1))}. \end{aligned}$$

Nun ist aber

$$\frac{\left(n\nu_n(1) - nP(1)\right)^2}{nP(1)(1-P(1))} = \left(\frac{1}{\sqrt{nP(1)(1-P(1))}} \sum_{k=1}^{n} \left(\mathbb{1}_{X_k=1} - P(1)\right)\right)^2 \equiv Z_n^2. \tag{6.73}$$

Unter der Verteilung $\mathbb{P}^P$ ist $\sum_{k=1}^{n} \mathbb{1}_{X_k=1}$ $\mathrm{Bin}(n, P(1))$-verteilt und daher konvergiert nach dem Satz von de Moivre-Laplace Z_n in Verteilung gegen eine Gauß-verteilte Zufallsvariable Z mit Mittelwert Null und Varianz 1. Daher gilt

$$\begin{aligned} \lim_{n\uparrow\infty} \mathbb{P}^P\left(D_{n,P} \leq c\right) &= \lim_{n\uparrow\infty} \mathbb{P}^P\left(Z_n^2 \leq c\right)\\ &= \lim_{n\uparrow\infty} \mathbb{P}^P\left(|Z_n| \leq \sqrt{c}\right) = \mathbb{P}^P\left(|Z| \leq \sqrt{c}\right)\\ &= \mathbb{P}^P\left(Z^2 \leq c\right), \end{aligned} \tag{6.74}$$

wie behauptet. □

Die Funktionen χ^2_k liegen tabelliert vor. Die zugehörigen Wahrscheinlichkeitsdichten kann man explizit angeben. Insbesondere folgt nun, dass wir, zumindest für grosse n einen Test der Nullhypothese $\theta = P$ zum Niveau α durch

$$\{D_{n,P} \leq \chi^2_{\ell-1,1-\alpha}\} \tag{6.75}$$

erhalten, wo $\chi^2_{k,1-\alpha}$ den Wert bezeichnet so dass $\chi^2_k\left(\chi^2_{k,1-\alpha}\right) = 1-\alpha$. Diesen Test nennt man einen χ^2-Anpassungstest zum Niveau α.

6.6 χ^2-Test für die Normalverteilung

Die Gauß-Verteilung spielt als ein bevorzugtes Modell in der Statistik eine wichtige Rolle. Daher ist es wichtig, auch die Hypothese, dass Zufallsvariablen Gauß-verteilt sind, testen zu können. Die im vorigen Abschnitt verwendete Methode scheint nicht direkt anwendbar, da wir angenommen hatten, dass die Zufallsvariablen nur endlich viele Werte annehmen. Es ist aber leicht, diese Methode an den Fall beliebiger Verteilungen anzupassen. Unser Modell sei also wie oben, aber mit der Nullhypothese, dass die Verteilung der Zufallsvariablen X_i die Gauß-Verteilung mit Mittelwert μ und Varianz σ^2 ist.

Um einen Test zu konstruieren, diskretisieren wir das Problem. Dazu teilen wir den Wertebereich $\mathbb{R}$ in ℓ disjunkte Teilintervalle $I_1,\ldots,I_\ell$ ein, wobei natürlich $\cup_{i=1}^{\ell} = \mathbb{R}$ sein muss. Jede Wahrscheinlichkeitsverteilung $\mathbb{P}$ auf $\mathbb{R}$ induziert dann eine Verteilung $\mathbb{P}^\ell$ auf der endlichen Menge $\{1,\ldots,\ell\}$ durch

$$\mathbb{P}^k(j) = \mathbb{P}(I_j), \quad j \in \{1,\ldots,\ell\}. \tag{6.76}$$

Indem wir nun testen ob die diskretisierte empirische Verteilung nahe bei der diskretisierten Gaußverteilung liegt, erhalten wir jedenfalls eine Test zum gewünschten Niveau für unsere Nullhypothese. Um auch den Fehler zweiter Art einigermassen klein zu machen, muss man die Zerlegung hinreichend fein machen und dann aber auch genügend Beobachtungen haben, damit eine substanziell grosse Zahl der Beobachtungen in jedes Intervall fällt.

6.7 χ^2-Test auf Unabhängigkeit

Eine wichtige Frage, die durch statistische Test geklärt werden soll, ist die nach der Unabhängigkeit verschiedener Eigenschaften. Gerade im medizinischen Bereich wird so etwas sehr häufig untersucht: So fragt man sich, ob etwa der Verlauf einer Krankheit von der Einnahme eines Medikaments abhängig ist oder nicht, oder man möchte wissen, ob der Verzehr gewisser Nahrungsmittel mit dem Gesundheitszustand zusammenhängt, oder nicht. Eine etwas modernere Frage ist es, Zusammenhänge zwischen genetischen Charakteristiken und bestimmten Erkrankungen herauszufinden. Bei all diesen Fragen sucht man im Prinzip nach kausalen Zusammenhängen. Man sollte sich allerdings bewusst sein, dass dies sehr schwierig ist und von statistischen Test selten geleistet werden kann.

Wir wollen die folgende Situation betrachten. Ein Individuum i sei durch zwei Merkmale, X und Y, charakterisiert. (z. B. Form und Farbe), die jeweils Werte in

den endliche Mengen A und B annehmen. Wir bezeichnen den Produktraum mit $E = A \times B$ Unser statistisches Modell besteht in der Grundannahme, dass wir eine beliebige Anzahl von Individuen $i \in \mathbb{N}$ haben, deren Eigenschaften durch unabhängige Zufallsvariablen $(X_i, Y_i), i \in \mathbb{N}$ modelliert sind. Modellparameter ist die gemeinsame Verteilung θ dieses Paars von Zufallsvariablen. Insbesondere wollen wir einen Test auf die Nullhypothese, dass X und Z unabhängige Zufallsvariablen sind, entwerfen.

Wir definieren die sogenannten *Randverteilungen*

$$\theta^A(i) = \sum_{j \in B} \theta(i,j), \quad i \in A, \tag{6.77}$$

$$\theta^B(j) = \sum_{i \in A} \theta(i,j), \quad j \in B. \tag{6.78}$$

Die Nullhypothese lautet dann: $\theta = \theta^A \otimes \theta^B$. Wenn wir mit $\mathscr{M}_A$ und $\mathscr{M}_B$ die Menge der Wahrscheinlichkeitsmaße auf den Mengen A und B bezeichnen, so können wir die Nullhypothese in der expliziteren Form

$$\Theta_0 = \left\{ \theta : \forall_{i \in A, j \in B} \theta(i,j) = \alpha(i)\beta(j), \text{für } \alpha \in \mathscr{M}_A, \beta \in \mathscr{M}_B \right\}, \tag{6.79}$$

schreiben. Um einen Test der Nullhypothese zu entwerfen, wollen wir ganz analog zum Fall des Anpassungstests vorgehen. Dazu betrachten wir die empirische Verteilung

$$\nu_n = \frac{1}{n} \sum_{k=1}^{n} \delta_{(X_k, Y_k)}. \tag{6.80}$$

Die dieser entsprechende Matrix

$$h_n(ij) \equiv \#\{1 \le k \le n : (X_k, Y_k) = (i,j)\}, \quad (i,j) \in A \times B, \tag{6.81}$$

wird in der Statistik *Kontingenztafel* genannt. Wie wollen nun wieder den *Likelihoodkoeffizienten* der Beobachtung bestimmen.

Unter dem Wahrscheinlikeitsmaß θ ist die Likelihood der Beobachtung ν_n gegeben durch

$$\rho_n(\theta; \nu_n) = \prod_{(i,j) \in A \times B} [\theta(i,j)]^{h_n(ij)}. \tag{6.82}$$

Damit ist das Verhältnis der maximalen Likelihood zur maximalen Likelihood unter der Nullhypothese

$$R_n = \frac{\sup_{\theta \in \Theta} \prod_{(i,j) \in A \times B} [\theta(i,j)]^{h_n(ij)}}{\sup_{\alpha \in \mathscr{M}_A, \beta \in \mathscr{M}_B} \prod_{(i,j) \in A \times B} [\alpha(i)\beta(j)]^{h_n(ij)}}. \tag{6.83}$$

Das supremum im Zähler wird wieder von der empirischen Verteilung selbst angenommen, ist also $\prod_{i,j} \nu_n(i,j)^{n\nu_n(i,j)}$. Für die Terme im Nenner gilt

$$\begin{aligned}\prod_{(i,j)\in A\times B}[\alpha(i)\beta(j)]^{h_n(ij)} &= \prod_{(i,j)\in A\times B}[\alpha(i)]^{h_n(ij)} \times \prod_{(i,j)\in A\times B}[\beta(j)]^{h_n(ij)} \\ &= \prod_{i\in A}[\alpha(i)]^{\sum_{j\in B}h_n(ij)} \times \prod_{j\in B}[\beta(j)]^{\sum_{i\in A}h_n(ij)} \\ &= \prod_{i\in A}[\alpha(i)]^{n\nu^A(i)} \times \prod_{j\in B}[\beta(j)]^{n\nu_n^B(j)}, \end{aligned} \tag{6.84}$$

wo wir wieder gesetzt haben $\nu^A(i) = \sum_{j\in B}\nu_n(i,j) = \frac{1}{n}\sum_{j\in B}h_n(ij)$, etc.. Damit ist

$$\begin{aligned}\sup_{\alpha\in\mathscr{M}_A,\beta\in\mathscr{M}_B}\prod_{(i,j)\in A\times B}[\alpha(i)\beta(j)]^{h_n(ij)} &= \left(\sup_{\alpha\in\mathscr{M}_A}\prod_{i\in A}[\alpha(i)]^{n\nu^A(i)}\right)\left(\sup_{\beta\in\mathscr{M}_B}\prod_{j\in B}[\beta(i)]^{n\nu^B(j)}\right) \\ &= \prod_{i\in A}\left[\nu_n^A(i)\right]^{n\nu^A(i)}\prod_{j\in B}\left[\nu_n^B(j)\right]^{n\nu^B(j)} \\ &= \prod_{(i,j)\in A\times B}\left[\nu_n^A(i)\nu_n^B(j)\right]^{n\nu_n(i,j)}. \end{aligned} \tag{6.85}$$

Also erhalten wir

$$R_n = \prod_{(i,j)\in A\times B}\left[\frac{\nu_n(i,j)}{\nu_n^A(i)\nu_n^B(j)}\right]^{n\nu_n(i,j)} = \exp\left(nH(\nu_n,\nu_n^A\otimes\nu_n^B)\right), \tag{6.86}$$

wo H wieder die relative Entropie ist. Unter der Nullhypothese sollte ν_n nahe an einem Produktmaß sein, also H nahe null sein. Genau wie im Fall des χ^2-Anpassungstest kann man die relative Entropie quadratisch approximieren durch

$$H(\nu_n,\nu_n^A\otimes\nu_n^B) \approx \frac{1}{2}\sum_{(i,j)\in A\times B}\nu_n^A(i)\nu_n^B(j)\left(\frac{\nu_n(i,j)}{\nu_n^A(i)\nu_n^B(j)}-1\right)^2 \equiv \frac{1}{2n}\widetilde{D}_n. \tag{6.87}$$

Der Nutzen dieses Verfahren kommt wieder von einem Grenzwertsatz, der diesmal von Pearson bewiesen wurde.

Satz 6.10. *Für jedes Produktmaß $\alpha\otimes\beta$ konvergiert $\widetilde{D}_n$ in Verteilung gegen eine $\chi^2_{(a-1)(b-1)}$-verteilte Zufallsvariable, wo $a = |A|$ und $b = |B|$, d.h., für alle $c > 0$ gilt*

$$\lim_{n\uparrow\infty}\mathbb{P}^{\alpha\otimes\beta}\left(\widetilde{D}_n \le c\right) = \chi^2_{(a-1)(b-1)}([0,c]). \tag{6.88}$$

Der Beweis dieses Satzes würde den Rahmen der Vorlesung sprengen.

Damit habe wir also ein gutes Verfahren zum Testen auf Unabhängigkeit. Aus n Beobachtungen der Merkmale A und B bestimmt man die Kontingenztafel $h_n(ij)$ und somit die empirische Verteilung $\nu_n(i,j) = \frac{1}{n}h_n(ij)$. Weiter bestimmt man die empirischen Randverteilungen der Merkmale A und B, $\nu^A(i), \nu^B(j)$. Diese setzt man in die Formel (6.87) ein und erhält $\widetilde{D}_n$. Will man nun die Nullhypothese, dass die Merkmale A und B unabhängig sind zum Nieveau α testen, so benutzt man als Test

$\{\widetilde{D}_n \leq \chi^2_{(a-1)(b-1),\alpha}\}$, welches man in einer Tabelle nachschlägt. Man verwirft die Nullhypothese also nur, wenn $\widetilde{D}_n > \chi^2_{(a-1)(b-1),\alpha}$ ist.

Kapitel 7
Nochmal: Der zentrale Grenzwertsatz*

Wir wollen zum Abschluß der Vorlesung noch den zentralen Grenzwertsatz 5.5 beweisen. Es gibt eine gane Reihe von Möglichkeiten, dies zu tun. Der hier gegenene Beweis ist dem Buch von Kersting und Wakolbinger [3] entnommen. Er hat den Vorteil, dass er nur die Kenntnis der Taylorformel aus der reellen Analysis voraussetzt.

Gauss'sche Zufallsvariablen haben folgende wichtige Eigenschaft.

Lemma 7.1. *Es seien X_1 und X_2 unabhängige, identisch verteilte Zufallsvariablem mit Mittelwert Null und Varianzen σ_1^2 und σ_2^2. Dann ist $Z \equiv X_1 + X_2$ eine Gauss'sche Zufallsvariable mit Mittelwert Null und Varianz $\sigma_1^2 + \sigma_2^2$.*

Beweis. Man kann dieses Lemma z. B. durch explizite Berechnung der Verteilungsfunktion von Z beweisen. Wir wollen aber ein anderes Argument benutzen, dass ohne Rechnung auskommt. Dazu benutzen wir den Satz von de Moivre Laplace. Es seine $X_i, i \in \mathbb{N}$ unabhängige $Ber(1/2)$ Zufallsvariablen. Dann ist

$$\frac{2}{\sqrt{n}} \sum_{i=1}^{[\sigma_1^2 n]} (X_i - 1/2) = \frac{2\sqrt{[\sigma_1^2 n]}}{\sqrt{n}} \frac{1}{\sqrt{[\sigma_1^2 n]}} \sum_{i=1}^{[\sigma_1^2 n]} (X_i - 1/2). \tag{7.1}$$

Wegen dem Satz von de Moivre Laplace folgt daraus, dass

$$\frac{2}{\sqrt{n}} \sum_{i=1}^{[\sigma_1^2 n]} (X_i - 1/2) \to X_1, \tag{7.2}$$

in Verteilung. Ebenso gilt

$$\frac{2}{\sqrt{n}} \sum_{i=[\sigma_1^2 n]+1}^{[\sigma_1^2 n]+[\sigma_2^2 n]} (X_i - 1/2) \to X_2. \tag{7.3}$$

Weiterhin sind X_1 und X_2 unabhängig. Dazu müssen wir nur zeigen, dass sie unkorreliert sind.

$$\mathbb{E}[X_1X_2] = \lim_{n\uparrow\infty} \mathbb{E}\left(\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{[\sigma_1^2 n]}(X_i - 1/2)\right]\left[\frac{1}{\sqrt{n}}\sum_{i=[\sigma_1^2 n]+1}^{[\sigma_1^2 n]+[\sigma_2^2 n]}(X_i - 1/2)\right]\right) = 0. \quad (7.4)$$

Andererseits ist

$$X_1 + X_2 = \lim_{n\uparrow\infty} \frac{2}{\sqrt{n}} \sum_{i=1}^{[\sigma_1^2 n]+[\sigma_2^2 n]} (X_i - 1/2) = Z. \quad (7.5)$$

Damit ist die Behauptung bewiesen. □

Was wir gerade gezeigt haben, ist, dass die Eigenschaft der Gauss-Verteilung im Lemma 7.1 eine Folge der Tatsache ist, dass die Gauss-Verteilung die Verteilung eines Grenzwertes einer Summe von unabhängigen Zufallsvariablen ist.

Unser Lemma impliziert insbesondere, dass wenn $X_i, i \in \mathbb{N}$ unabhängige, Gauss-verteilte Zufallsvariblen mit Mittelwert Null und Varianz eins sind, dann ist für jedes $n \in \mathbb{N}$,

$$Z_n \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_1 \quad (7.6)$$

ebenfalls eine Gauss-verteilte Zufallsvariablen mit Mittelwert Null und Varianz eins. Wir wollen nun folgenden Satz beweisen.

Lemma 7.2. *Seien $X_i, i \in \mathbb{N}$ wie oben und seien $Y_i, i \in \mathbb{N}$ unabhängige identisch verteilte Zufallsvariablem mit Mittelwert Null und Varianz eins. Es sei $W_n \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$. Sei $f : \mathbb{R} \to \mathbb{R}$ eine dreimal stetig differenzierbare Funktion mit drei beschränkten Ableitungen. Dann gilt*

$$\lim_{n\uparrow\infty} \mathbb{E}\left(f(W_n) - f(Z_n)\right) = 0. \quad (7.7)$$

Beweis. Der Beweis beruht darauf, die Differenz $f(W_n) - f(Z_n)$ in viele kleine Differenzen zu zerlegen. Dazu benutzen wir eine teleskopische Entwicklung":

$$\begin{aligned}
f(W_n) = {} & f\left(n^{-1/2}\sum_{i=1}^{n} Y_i\right) - f\left(n^{-1/2}\left(X_1 + \sum_{i=2}^{n} Y_i\right)\right) \\
& + f\left(n^{-1/2}\left(X_1 + \sum_{i=2}^{n} Y_i\right)\right) - f\left(n^{-1/2}\left(X_1 + X_2 + \sum_{i=3}^{n} Y_i\right)\right) \\
& + f\left(n^{-1/2}\left(X_1 + X_2 + \sum_{i=3}^{n} Y_i\right)\right) - f\left(n^{-1/2}\left(X_1 + X_2 + X_3 + \sum_{i=4}^{n} Y_i\right)\right) \\
& + \ldots \\
& + \ldots \\
& + f\left(n^{-1/2}\left(\sum_{i=1}^{n-1} X_i + Y_n +\right)\right) - f\left(n^{-1/2}\left(\sum_{i=1}^{n} X_i\right)\right) \\
& + f(Z_n). \quad (7.8)
\end{aligned}$$

Daher ist

$$\begin{aligned}\mathbb{E}\left(f(W_n)-f(Z_n)\right) &= \mathbb{E}\left[f\left(n^{-1/2}\sum_{i=1}^{n}Y_i\right)-f\left(n^{-1/2}\left(X_1+\sum_{i=2}^{n}Y_i\right)\right)\right]\\ &+\mathbb{E}\left[f\left(n^{-1/2}\left(X_1+\sum_{i=2}^{n}Y_i\right)\right)-f\left(n^{-1/2}\left(X_1+X_2+\sum_{i=3}^{n}Y_i\right)\right)\right]\\ &+\ldots\\ &+\ldots\\ &+\mathbb{E}\left[f\left(n^{-1/2}\left(\sum_{i=1}^{n-1}X_i+Y_n+\right)\right)-f\left(n^{-1/2}\sum_{i=1}^{n}X_i\right)\right]. \qquad (7.9)\end{aligned}$$

In jedem Summanden unterscheiden sich die Argumente der Funktionen f nun nur in einem Term. Um alles etwas übersichtlicher zu machen, führen wir die Variablen

$$U_k \equiv n^{-1/2}\left(\sum_{i=1}^{k}X_i+\sum_{k+2}^{n}Y_i\right) \qquad (7.10)$$

ein. Die Terme in den einzelnen Zeilen in (7.9) sind dann von der Form $f(U_k+n^{-1/2}Y_{k+1})-f(U_k+n^{-1/2}X_{k+1})$. Die Taylorentwicklung zur Ordnung 2 gibt dann

$$\begin{aligned}&f\left(U_k+n^{-1/2}X_{k+1}\right)-f\left(U_k+n^{-1/2}Y_{k+1}\right)\\ &=0+n^{-1/2}(Y_{k+1}-X_{k+1})f'(U_k)\\ &+n^{-1}\frac{1}{2}(Y_{k+1}^2-X_{k+1}^2)f''(U_k)+R_3, \qquad (7.11)\end{aligned}$$

wo R_3 das Restglied der Ordnung 3 ist, das wir in der Form

$$\begin{aligned}2nR_3 &= Y_{k+1}^2\left(f''\left(U_k+n^{-1/2}\eta Y_{k+1}\right)-f''(U_k)\right) \qquad (7.12)\\ &-X_{k+1}^2\left(f''\left(U_k+n^{-1/2}\eta' X_{k+1}\right)-f''(U_k)\right),\end{aligned}$$

für $\eta,\eta'\in[0,1]$. Den Term, der die Gauss'sche Zufallsvariable X_{k+1} involviert, können wir einfach die Ungleichung

$$\left|f''\left(U_k+n^{-1/2}\eta' X_{k+1}\right)-f''(U_k)\right|\le C|X_{k+1}|, \qquad (7.13)$$

benutzen, da das resultierende $|X_{k+1}|^3$ endliche Erwartung hat. Für den anderen Term geht dies nicht ohne weiteres, da wir nicht voraussetzen, dass $|Y_{k+1}|^3$ endliche Erwartung hat. Wir führen daher eine Konstante $K<\infty$ ein und unterscheiden die Fälle $|Y_{k+1}|\le K$ und $|Y_{k+1}|>K$. Wir erhalten damit

$$\left|f''\left(U_k+n^{-1/2}\eta Y_{k+1}\right)-f''(U_k)\right| \le \mathbb{1}_{|Y_{k+1}|\le K}n^{-1/2}|Y_{k+1}|C+\mathbb{1}_{|Y_{k+1}|>K}C''.$$

Wenn wir diese Abschätzungen in Gleichung (7.11) einsetzen und dann die Erwartung nehmen, erhalten wir

$$\begin{aligned}&\mathbb{E}\left[f\left(U_k+n^{-1/2}Y_{k+1}\right)-f\left(U_k+n^{-1/2}X_{k+1}\right)\right]\\&=0+0+0+n^{-1}\left(C'''n^{-1/2}+K^3n^{-1/2}C+C''\mathbb{E}\left[Y_{k+1}^2\mathbb{1}_{|Y_{k+1}|>K}\right]\right).\end{aligned}\tag{7.14}$$

Da nach Vorraussetzung die Erwartung von Y_{k+1}^2 endlich ist, existiert der Limes $K\uparrow\infty$ von $\mathbb{E}\left[Y_{k+1}^2\mathbb{1}_{|Y_{k+1}|\leq K}\right]$, und somit gilt

$$\lim+K\uparrow\infty E\left[Y_{k+1}^2\mathbb{1}_{|Y_{k+1}|>K}\right]=0.\tag{7.15}$$

Damit folgt aber

$$\lim_{n\uparrow\infty}|\mathbb{E}\left(f(W_n)-f(Z_n)\right)|\leq C'\mathbb{E}\left[Y_1^2\mathbb{1}_{|Y_1|>K}\right],\tag{7.16}$$

für jedes $K<\infty$¡ und daher

$$\lim_{n\uparrow\infty}|\mathbb{E}\left(f(W_n)-f(Z_n)\right)|\leq\varepsilon,\tag{7.17}$$

für jedes $\varepsilon>0$, mithin

$$\lim_{n\uparrow\infty}\mathbb{E}\left(f(W_n)-f(Z_n)\right)=0.\tag{7.18}$$

Was zu beweisen war. □

Aus unserem Lemma folgt aber nun der zentrale Grenzwertsatz sehr leicht.

Beweis (von Satz 5.5). Wenn wir im Lemma 7.7 die Funktion f als die Indikatorfunktion $\mathbb{1}_{[a,b]}$ wählen durften, dann wären wir schon fertig. Die Indikatorfunktion ist aber nicht differenzierbar. Nun kann man sich aber leicht überlegen, dass man eine Stufenfunktion beliebig gut durch glatte Funktionen annähern kann. Insbesondere gibt es für jedes $\delta>0$ zwei Funktionen f_1,f_2, die die Eigenschaften, die im Lemma gefordert werden haben und für die gilt

(i) Für alle $x<a-\delta$, alle $x\in[a,b]$ und alle $x>b+\delta$ ist $f_1(x)=\mathbb{1}_{[a,b]}(x)$;
(ii)Für alle $x<a$, alle $x\in[a+\delta,b-\delta]$ und alle $x>b$ ist $f_2(x)=\mathbb{1}_{[a,b]}(x)$;
(iii)Für alle $x\in\mathbb{R}$ gilt $0\leq f_2(x)\leq\mathbb{1}_{[a,b]}(x)\leq f_1(x)\leq 1$.

Dann gilt zunächst für jedes n,

$$\mathbb{E}\left[f_2(W_n)\right]\leq\mathbb{E}\left[\mathbb{1}_{[a,b]}(W_n)\right]\leq\mathbb{E}\left[f_2(W_n)\right].\tag{7.19}$$

Wegen Lemma 7.7 folgt dann

$$\mathbb{E}\left[f_2(Z)\right]\leq\liminf_{n\uparrow\infty}\mathbb{E}\left[\mathbb{1}_{[a,b]}(W_n)\right]\leq\limsup_{n\uparrow\infty}\mathbb{E}\left[\mathbb{1}_{[a,b]}(W_n)\right]\leq\mathbb{E}\left[f_1(Z)\right],\tag{7.20}$$

wo Z Gauss mit Mittelwert null und Varianz eins ist. Andererseits ist nach Konstruktion der Funktionen f_i,

$$\mathbb{E}[f_1(Z)] \leq \mathbb{P}(a-\delta \leq Z \leq b+\delta), \quad \mathbb{E}[f_2(Z)] \geq \mathbb{P}(a+\delta \leq Z \leq b-\delta). \quad (7.21)$$

Da $\delta > 0$ beliebig kelin sein darf und die Verteilungsfuntion der Gauss-Verteilung stetig ist, können wir die oberen und unteren Schranken in (7.21) beliebig nahe and $\mathbb{P}(a \leq Z \leq b)$ bringen, woraus der zentrale Grenzwertsatz folgt. □

Literaturverzeichnis

1. Hans-Otto Georgii. *Stochastik*. de Gruyter Lehrbuch. Walter de Gruyter & Co., Berlin, 2002.
2. Norbert Henze. *Stochastik für Einsteiger*. Vieweg + Teubner Verlag, 7 edition, 2008.
3. G. Kersting and A. Wakolbinger. *Elementare Stochastik*. Birkhäuser, Basel, Boston, Berlin, 2008.

Sachverzeichnis