Stochastik – ein Fest der Unabhängigkeit

REINHARD WINKLER (TU WIEN)

Die allermeisten Ergebnisse der Stochastik (also der Wahrscheinlichkeitsrechnung und der Statistik) haben als wesentliche Voraussetzung irgendeine Spielart von stochastischer Unabhängigkeit. Dies anhand der wichtigsten Beispiele und dennoch relativ systematisch und umfassend darzustellen ist das Anliegen dieses Artikels. Die Auswahl der einzelnen Inhalte folgt Überlegungen zur Frage: Welche Ideen, Konzepte und Resultate sollten Lehrende an Höheren Schulen unbedingt präsent haben, um Stochastik angemessen unterrichten zu können?

1. Einleitung

Innerhalb der Mathematik an Höheren Schulen scheinen in der Stochastik immer wieder gewisse Unsicherheiten darüber aufzutreten, was denn nun genau zu unterrichten sei. Worum also geht es in der Stochastik überhaupt? Eine sehr knappe Antwort könnte lauten: *Stochastik ist die mathematische Analyse des Zufalls,* oder etwas konkreter, *Maßtheorie plus Unabhängigkeit.*

Erstens ist der grundlegende Begriff der Stochastik, nämlich Wahrscheinlichkeit, so wie sie seit Kolmogorow (vgl. Kolmogorow (1933)) verstanden wird, mathematisch nichts anderes als ein Maß – völlig in Analogie zu Flächen- und anderen Maßen aus der Geometrie. Und zweitens beruhen die Hauptergebnisse der Stochastik, insofern sie über die bloße Maßtheorie hinausgehen, fast durchwegs auf stochastischer Unabhängigkeit als wesentlicher Voraussetzung. Hat man das einmal begriffen, so lösen sich allfällige Mysterien um das Wesen des Zufalls wenigstens hinsichtlich mathematischer Modellbildung von selber auf. Frei nach Carnap (vgl. Carnap (1931)) könnte man von der Überwindung der Metaphysik des Zufalls durch mathematische Analyse der stochastischen Unabhängigkeit sprechen.

Was für die meisten meiner bisherigen Artikel in den Didaktikheften gilt, möchte ich in diesem Fall besonders betonen: Der Text richtet sich nicht direkt an Schüler. Entsprechend ist er nicht als Eins-zu-Eins-Vorlage für den Unterricht gedacht, sondern – bei bewusstem Verzicht auf manch Wichtiges sowie auf strenge Vollständigkeit – als Bündelung jenes Hintergrundverständnisses, das mir für Lehrende wesentlich erscheint, um den Lehrstoff im Schulunterricht sinnvoll zu strukturieren und richtige Schwerpunkte zu setzen. Auch der Zugang zu stochastischer Fachliteratur soll durch den vorliegenden Text erleichtert werden. In der konkreten Unterrichtsgestaltung hingegen sind natürlich immer Vereinfachungen und Anpassungen an die Gegebenheiten in der jeweiligen Klasse, wie sie Lehrenden nie abgenommen werden können, nötig.

Nun zum Inhalt im Einzelnen: Kapitel 2 bringt die konzeptuellen Grundlagen. Und zwar wird in Abschnitt 2.1 erklärt, wie der vage Begriff des Zufalls mathematisch durch jenen der stochastischen Unabhängigkeit modelliert werden kann. Das führt unweigerlich zum Begriff des Wahrscheinlichkeitsmaßes (Abschnitt 2.2) und in weiterer Folge zu Produkträumen und dem Satz von Fubini (Abschnitt 2.3). In Kapitel 3 kommen Zufallsgrößen und deren Unabhängigkeit ins Spiel. Zunächst geht es um den Einfluss von Unabhängigkeit auf Erwartungswert und Varianz (Abschnitte 3.1 und 3.2). Als eine weitere Erscheinungsform von Unabhängigkeit wird in Abschnitt 3.3 die Gedächtnislosigkeit von geometrischer und Exponentialverteilung thematisiert, in Abschnitt 3.4 die Faltung von Maßen. Ein mathematisches Hauptziel stellt Kapitel 4 dar, wo zwei der wichtigsten Resultate der Stochastik besprochen werden: Das Gesetz der großen Zahlen (Abschnitt 4.1, mit Beweis) und der zentrale Grenzwertsatz, der – weil es um einen Grenzwert von Verteilungen, nicht von Zahlen geht – oft auch Grenz*verteilungs*satz genannt wird (Abschnitt 4.2, ohne Beweis aber mit relativ ausführlichen Erläuterungen). Kapitel 5 gibt Einblicke, wie Unabhängigkeit auch für stochastische Prozesse (solche mit unabhängigen Zuwächsen, Abschnitt 5.1) und schließende Statistik (vermittels Unabhängigkeit von Stichproben, Abschnitt 5.2) ganz entscheidend

Ich danke Mathias Beiglböck, Manfred Borovcnik, Martin Goldstern, Gregor Kastner und Manfred Kronfellner für wertvolle Anregungen zur Verbesserung des vorliegenden Artikels.

wirksam wird. Das abschließende Kapitel 6 zieht Bilanz zur Rolle von Maß und Unabhängigkeit in der mathematischen Stochastik und berührt auch philosophische Fragen.

Wer sich für umfassendere Literatur zur Stochastik interessiert, dem sei beispielsweise das allgemein einführende Lehrbuch Georgii (2009) empfohlen. Wer eine systematischere maßtheoretische Basis, die ich hier weitgehend aussparen muss, bevorzugt, wird eher zu Kusolitsch (2011) greifen, zu Bauer (1992) und Bauer (2002) (eher für Fortgeschrittene; Wahrscheinlichkeitstheorie ohne Statistik) oder zu Rudin (2009) (Standardwerk der Höheren Analysis, weniger auf Stochastik bezogen). Für eine wesentlich überschaubarere Darstellung der Grundideen der (Lebesgueschen) Maß- und Integrationstheorie auf ähnlichem Niveau wie der vorliegende Artikel darf ich auf meinen Artikel Winkler (2011) verweisen und für damit verbundene Grundlagenprobleme auf Winkler (2001). Generell finden sich in bisherigen Bänden der Didaktikhefte zahlreiche Artikel über spezielle Themen der Stochastik, sowohl über konkrete Anwendungen als auch über speziellere didaktische Gesichtspunkte.

2. Unabhängigkeit, Wahrscheinlichkeit und Produkträume

2.1. Zufall und stochastische Unabhängigkeit

Kaum etwas eignet sich besser, um die Effizienz mathematischer Begriffe zu illustrieren, als die Frage, was *Zufall* ist. Beginnen wir mit einem Beispiel aus dem Alltag.

Wenn jemand sagt, er habe *zufällig* eine Freundin getroffen, so ist damit meistens gemeint, dass dem Treffen nichts vorausgegangen sei, was dieses Treffen in augenscheinlicher Weise wahrscheinlicher gemacht hätte. Die Zufälligkeit eines Ereignisses bedeutet also eine kausale Unabhängigkeit von anderen Ereignissen, oder genauer: dass gewisse andere Ereignisse nichts an der Wahrscheinlichkeit des beobachteten Ereignisses ändern.

Um das formal zu fassen, schreiben wir $P(A)$ für die Wahrscheinlichkeit (probability) des Ereignisses A und $P(A|B)$ für die Wahrscheinlichkeit von A unter der Hypothese, dass B eintritt. $P(A|B)$ heißt auch die *bedingte Wahrscheinlichkeit* von A unter (der Bedingung) B.[1] Unabhängigkeit von A und B im beschriebenen Sinn bedeutet also $P(A|B) = P(A)$. Um damit mehr anfangen zu können, müssen wir die Größe $P(A|B)$ besser verstehen. Und zwar meinen wir damit Folgendes.

Wissen wir, dass B eintritt, so bedeutet das Eintreten von A, dass sowohl A als auch B eintreten oder, in mengentheoretischer Schreibweise, dass das Ereignis $A \cap B$ eintritt. Als $P(A|B)$ können wir aber nicht einfach $P(A \cap B)$ nehmen. Das würde im Normalfall einen zu niedrigen Wert liefern, denn wir wissen ja immerhin schon, dass wir uns innerhalb von B befinden. Es geht also um die relative Größe von $A \cap B$ innerhalb von B, d.h. um den Wert $P(A|B) = \frac{P(A \cap B)}{P(B)}$. Diese Gleichung verwenden wir als Definition für die *bedingte Wahrscheinlichkeit* $P(A|B)$ (sofern $P(B) \neq 0$). Damit können wir die Bedingung $P(A|B) = P(A)$ für die Unabhängigkeit von A und B umschreiben in die symmetrische Form $P(A \cap B) = P(A) \cdot P(B)$, allgemeiner:

Definition: *Eine beliebige Familie von Ereignissen A_i, $i \in I$, heißt* (stochastisch) unabhängig, *wenn für je endlich viele paarweise verschiedene Indizes $i_1, i_2, \ldots, i_n \in I$*

$$P(A_{i_1} \cap \ldots \cap A_{i_n}) = P(A_{i_1}) \cdot \ldots \cdot P(A_{i_n}).$$

gilt.[2]

[1] Es sei betont, dass zwischen den Ereignissen A und B keinerlei inhaltliche Beziehung wie zeitliche Abfolge vorausgesetzt werden muss, es geht nur um die nun näher zu untersuchenden Wahrscheinlichkeiten.

[2] Man beachte, dass schon bei $I = \{1,2,3\}$ aus der paarweisen Unabhängigkeit mehrerer Ereignisse noch nicht die (gemeinsame) Unabhängigkeit folgt. Als einfachstes Beispiel betrachten wir zwei aufeinanderfolgende Würfe einer fairen Münze, so dass es vier mögliche und gleich wahrscheinliche Ergebnisse (also alle mit $P = \frac{1}{4}$) gibt, die wir als $(0,0), (0,1), (1,0)$ und $(1,1)$ notieren. Die Ereignisse $A_1 = \{(0,0),(1,1)\}$ (*bei beiden Würfen das gleiche Ergebnis*), $A_2 = \{(0,0),(0,1)\}$ (*beim ersten Wurf* 0) und $A_3 = \{(0,0),(1,0)\}$ (*beim zweiten Wurf* 0) sind wegen $P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A_i) \cdot P(A_j)$ für $i, j = 1,2,3$ paarweise unabhängig, nicht aber als dreielementige Familie: $P(A_1 \cap A_2 \cap A_3) = P(\{(0,0)\}) = \frac{1}{4} \neq \frac{1}{8} = \left(\frac{1}{2}\right)^3 = P(A_1) \cdot P(A_2) \cdot P(A_3)$

Allerdings ist eine noch grundlegendere Frage offen: Was ist die Wahrscheinlichkeit P, die wir bisher ganz unbefangen verwendet haben?

2.2. Wahrscheinlichkeit als Maß

Wir abstrahieren von Abschnitt 2.1: Die Wahrscheinlichkeit $P(A)$ eines Ereignisses A ist eine reelle Zahl zwischen 0 und 1. Also ist $P : \mathcal{A} \to [0,1]$ eine Funktion, die auf einem noch näher zu spezifizierenden System $\mathcal{A}$ von Mengen A definiert ist. Die Vereinigung aller $A \in \mathcal{A}$ wollen wir mit Ω bezeichnen. Die Elemente $\omega \in \Omega$ heißen Elementarereignisse, die Mengen $A \in \mathcal{A}$ heißen (messbare) Ereignisse. (Man wünscht sich natürlich die Messbarkeit aller Teilmengen von Ω. Leider lässt sich das in vielen Situationen nicht bewerkstelligen. Diese – rein mathematisch sehr faszinierenden – Schwierigkeiten spielen im vorliegenden Artikel aber nur eine untergeordnete Rolle. Mehr darüber findet sich etwa im bereits erwähnten Artikel Winkler (2001). Im Folgenden möge der Leser sich durch das lediglich der mathematischen Korrektheit geschuldete Auftreten des Mengensystems $\mathcal{A}$ aber möglichst wenig irritieren lassen.)

Weil Ω alle Möglichkeiten umfasst, soll $P(\Omega) = 1$ gelten, und symmetrisch dazu soll die leere Menge $\emptyset$ die Wahrscheinlichkeit $P(\emptyset) = 0$ haben. Implizit haben wir außerdem die Additivität von P verwendet (nämlich in jener Fußnote, wo wir vier gleich wahrscheinlichen, einander ausschließenden und alle Möglichkeiten abdeckenden Ereignissen jeweils die Wahrscheinlichkeit $\frac{1}{4}$ zugewiesen haben): Zwei einander ausschließende Ereignisse A, B (d.h. solche mit $A \cap B = \emptyset$) erfüllen stets $P(A \cup B) = P(A) + P(B)$. Mit diesen Forderungen an P ist ein für die Bedürfnisse des Schulunterrichts schon weitgehend praktikabler Wahrscheinlichkeitsbegriff festgelegt. Trotzdem gibt es gute Gründe, noch etwas weiter zu gehen.

Elementare Darstellungen der Wahrscheinlichkeitstheorie erschöpfen sich oft in rein kombinatorischen Überlegungen nach dem Prinzip: *Wahrscheinlichkeit ist die Anzahl der günstigen dividiert durch die Anzahl der möglichen Fälle*. Tiefer blickt man dagegen, wenn man sich von Analogien zur Geometrie leiten lässt. Es fällt nämlich schnell auf, dass alles, was oben über P gesagt wurde, auch zutrifft, wenn man statt Wahrscheinlichkeiten z.B. Längen von Teilintervallen (oder deren endliche Vereinigungen) des Einheitsintervalls $[0,1]$ misst, Flächen innerhalb des Einheitsquadrats $[0,1]^2$ oder Volumina innerhalb des Einheitswürfels $[0,1]^3$.

Der gemeinsame Begriff im Hintergrund ist der des (Wahrscheinlichkeits-)Maßes. Zu seiner Fixierung sind im Vergleich zum vorangegangenen Absatz nur mehr technische Ergänzungen (Ausdehnung der Additivität von zwei auf abzählbar viele paarweise disjunkte Mengen und die entsprechende Forderung an $\mathcal{A}$, eine σ-Algebra zu sein) nötig, die allerdings beeindruckende Konsequenzen haben.

Definition: *Sei Ω eine Menge*[3]*. Ein System $\mathcal{A}$ bestehend aus Teilmengen von Ω heißt* σ-Algebra *(auf Ω), falls gilt:*

1. $\emptyset \in \mathcal{A}$.
2. *Aus $A \in \mathcal{A}$ folgt $\Omega \setminus A \in \mathcal{A}$.*
3. *Aus $A_0, A_1, A_2, \ldots \in \mathcal{A}$, folgt stets, dass auch $\bigcup_{n\in\mathbb{N}} A_n$ in $\mathcal{A}$ liegt.*

(Wegen $\bigcap_{n\in\mathbb{N}} A_n = \Omega \setminus \bigcup_{n\in\mathbb{N}}(\Omega \setminus A_n)$ folgt aus 2. und 3., dass $\mathcal{A}$ auch gegenüber abzählbaren Durchschnitten abgeschlossen ist.)

Ist $\mathcal{A}$ eine σ-Algebra auf Ω, so heißt eine Funktion $P : \mathcal{A} \to [0,1]$ mit $P(\emptyset) = 0$ und $P(\Omega) = 1$ ein Wahrscheinlichkeitsmaß *(kurz auch W-Maß), falls P* σ-additiv *ist, d.h. falls für alle Folgen von Mengen*

[3] Ω repräsentiert die Menge aller möglichen Elementarereignisse. Wie das genau geschieht, ist oft gar nicht so wichtig. Sehr häufig wählt man daher Ω als Menge von Zahlen oder anderer geeigneter mathematischer Objekte. Im Zusammenhang mit Zufallsgrößen werden wir darauf nochmals zu sprechen kommen.

$A_n \in \mathcal{A}$, $n \in \mathbb{N}$, die $A_i \cap A_j = \emptyset$ für alle $i \neq j$ erfüllen, die Gleichung

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

gilt. Das Tripel $(\Omega, \mathcal{A}, P)$ nennt man dann einen Wahrscheinlichkeitsraum. *Die $\omega \in \Omega$ heißen* Elementarereignisse*, die $A \in \mathcal{A}$ heißen* messbare Mengen *oder schlicht* Ereignisse.

Bei endlichem, n-elementigem $\Omega = \{\omega_1, \ldots, \omega_n\}$ (im Schulunterricht wird man wohl mit diesem Fall und $n = 2$ beginnen) kann als $\mathcal{A}$ einfach die Potenzmenge, das heißt die Menge aller Teilmengen von Ω, genommen werden. Die W-Maße auf diesem Raum erhält man dann, indem man sämtliche n-Tupel $p = (p_1, \ldots, p_n)$ mit $0 \leq p_i \leq 1$ und $\sum_{i=1}^{n} p_i = 1$ betrachtet. Zu jedem solchen p und $A \subseteq \Omega$ hat man als $P(A)$ einfach die Summe aller p_i mit $\omega_i \in A$ zu nehmen. Im Fall $p_i = \frac{1}{n}$ spricht man von *diskreter Gleichverteilung* (hier auf einer n-elementigen Menge).

Entscheidend für die Wirksamkeit der Wahrscheinlichkeitstheorie ist, dass es Maße in Hülle und Fülle auch auf unendlichen Mengen gibt. Als eines der wichtigsten Beispiele greifen wir nochmals den Vergleich mit der Längenmessung von Teilmengen von $\Omega = [0,1]$ mit Hilfe des eindimensionalen Lebesgueschen Maßes $P = \lambda = \lambda_{[0,1]}$ auf. Was genau wir als σ-Algebra $\mathcal{A}$ der messbaren Teilmengen nehmen, soll uns hier nicht so sehr kümmern. Jedenfalls enthält $\mathcal{A}$ alle Teilintervalle von $[0,1]$. Aus den Eigenschaften einer σ-Algebra folgt bereits daraus eine enorme Reichhaltigkeit von $\mathcal{A}$. (Es bedarf einiger Ideen inklusive Verwendung des Auswahlaxioms, um überhaupt Teilmengen von $[0,1]$ zu finden, die nicht in $\mathcal{A}$ liegen.) Über das Maß λ wiederum genügt es hier zu wissen, dass wunschgemäß $\lambda([a,b]) = b - a$ für alle $0 \leq a \leq b \leq 1$ gilt. (Technisch wichtig ist überdies die Vollständigkeit von λ, d.h.: Ist N eine Nullmenge, also $N \in \mathcal{A}$ mit $\lambda(N) = 0$ und $M \subseteq N$, so ist auch M in $\mathcal{A}$ und daher eine Nullmenge.) Der Rest ergibt sich aus der σ-Additivität.

Hier interessiert uns die Interpretation von λ als Wahrscheinlichkeitsmaß: Ist $A \subseteq [0,1]$ eine messbare Teilmenge, so ist die Wahrscheinlichkeit $P(A)$, dass eine beliebig herausgegriffene Zahl $x \in [0,1]$ auch in A liegt, gerade das Maß $\lambda(A)$ von A. Im Gegensatz zur diskreten spricht man von *kontinuierlicher* oder *stetiger Gleichverteilung*, hier auf $[0,1]$.

2.3. Unabhängigkeit und Produkträume

Wir wollen nun die Idee präzisieren, dass Punkte (x,y) aus dem Einheitsquadrat $[0,1]^2$ zufällig ausgewählt werden. Statt des eindimensionalen müssen wir das zweidimensionale Lebesguemaß heranziehen, das zur Unterscheidung mit λ_2 bezeichnet sei. Ist $A \subseteq [0,1]^2$ (man stelle sich A als irgendein flächiges Gebilde innerhalb des Einheitsquadrats vor, z.B. als kleine Kreisscheibe), so entspricht die Wahrscheinlichkeit, dass ein zufälliger Punkt (x,y) in A liegt, gerade dem Flächeninhalt $\lambda_2(A)$.

Spezielle Beispiele sind solche Ereignisse, die nur von x oder nur von y abhängen, am einfachsten etwa die senkrecht bzw. waagrecht streifenförmigen Mengen $A = \{(x,y) \in [0,1]^2 : a \leq x \leq b\}$ mit $0 \leq a \leq b \leq 1$ oder $B = \{(x,y) \in [0,1]^2 : c \leq y \leq d\}$ mit $0 \leq c \leq d \leq 1$. Offenbar ist das Maß ihrer rechteckigen Schnittmenge $\lambda_2(A \cap B) = (b-a)(d-c) = \lambda_2(A) \cdot \lambda_2(B)$, also sind A und B unabhängig (bezüglich λ_2). Mit den Abkürzungen $A' = [a,b]$ und $B' = [c,d]$ gilt aber auch $A \cap B = A' \times B'$ und außerdem $\lambda(A') = \lambda_2(A)$ und $\lambda(B') = \lambda_2(B)$. In $[0,1]^2$ ergibt sich die zweidimensionale Maßzahl für ein kartesisches Produkt daher als Produkt der eindimensionalen: $\lambda_2(A' \times B') = \lambda(A') \cdot \lambda(B')$.

Genau das fordert man von einem *Produktraum* (im maßtheoretischen Sinn) bzw. von einem *Produktmaß*. Dabei kann man die Konstruktion auch auf Produkte mehrerer oder auch unendlich vieler Komponenten ausdehnen, was im Zusammenhang mit Folgen von Zufallsgrößen sehr wichtig ist. Wir dürfen hier auf eine genauere Ausführung der Konstruktionen und Beweise verzichten, weil die Intuition aus dem behandelten Fall des Quadrates (oder des Würfels) als Produkt von zwei (bzw. drei) Intervallen für unsere Zwecke reicht. Als Folgerung gilt der höchst plausible und für das Weitere entscheidende *Satz* *von Fubini*, der die Integration auf Produkträumen beschreibt. Für Funktionen $f : [0,1]^2 \to \mathbb{R}$ (um die genauen technischen Voraussetzungen wollen wir uns hier nicht kümmern) lautet er:

$$\int_{[0,1]^2} f(x,y)\,d\lambda_2(x,y) = \int_{[0,1]} \left(\int_{[0,1]} f(x,y)\,d\lambda(y) \right) d\lambda(x) = \int_{[0,1]} \left(\int_{[0,1]} f(x,y)\,d\lambda(x) \right) d\lambda(y)$$

In Worten und etwas allgemeiner: Das Integral über einen Produktraum lässt sich mittels iterierter Integrale über die beiden ursprünglichen Wahrscheinlichkeitsräume ermitteln[4]. Die Eingangsüberlegungen mit den Mengen A, B, A', B' entsprechen dem Spezialfall, wo f die charakteristische Funktion der Menge $A \cap B$ ist. Wir resümieren:

Stochastische Unabhängigkeit wird durch Produktmaße modelliert, wo höherdimensionale Maßzahlen kartesischer Produkte sich als Produkte der entsprechenden niederdimensionalen Maßzahlen ergeben. Das entspricht geometrischer Flächenberechnung nach der üblichen Formel für Rechtecke. Auf Integrale über zweidimensionale Bereiche überträgt sich das nach dem Satz von Fubini vermittels Iteration eindimensionaler Integrale.

3. Zufallsgrößen und Unabhängigkeit

3.1. Zufallsgrößen und ihr Erwartungswert

In erster Annäherung darf man sich unter einer Zufallsgröße eine numerische Größe X vorstellen, die sich aus zufälligen Ereignissen ω ergibt. Oft, etwa wenn man ein Kontinuum möglicher Werte einer Messgröße im Auge hat, stellt man sich bereits unter dem Ereignis ω selbst eine numerische Größe vor, und die Unterscheidung zwischen ω und $X(\omega)$ erscheint artifiziell. Dennoch hat sich dieser Gesichtspunkt bewährt. Beim Erwartungswert $\mathrm{E}(X)$ von X denke man an das arithmetische Mittel über alle möglichen Werte $X(\omega)$, allerdings gewichtet mit ihrer Wahrscheinlichkeit. Die genaue Definition lautet wie folgt.

[4] Für eine Beweisskizze erinnern wir uns zunächst, dass für eine Funktion $f : [0,1] \to \mathbb{R}$, die auf jedem Teilintervall $I_j = [\frac{j}{n}, \frac{j+1}{n}[$, $j = 0, 1, \ldots, n-1$, jeweils den konstanten Wert $f(\frac{j}{n})$ annimmt, die Beziehung

$$\int_0^1 f(x)\,dx = \int_{[0,1]} f(x)\,d\lambda(x) = \frac{1}{n} \sum_{j=1}^{n} f\left(\frac{j}{n}\right)$$

gilt. Ist f keine so spezielle Treppenfunktion, wie hier vorausgesetzt, so stimmt die zweite Gleichheit natürlich nicht mehr exakt. Immerhin kann aber bei jedem Riemann-integrierbaren und somit erst recht bei jedem stetigen f und zu jedem $\varepsilon > 0$ für hinreichend große n ein Fehler kleiner als ε garantiert werden. Analoges gilt, wenn sich die Integration statt über ein eindimensionales Intervall (hier $[0,1]$) über eine zweidimensionale Fläche, z.B. $f : [0,1]^2 \to \mathbb{R}$, erstreckt. Das Integral $\int_{[0,1]^2} f\,d\lambda_2 = \int_{[0,1]^2} f(x,y)\,d\lambda_2(x,y)$ misst das Volumen $\lambda_3(V)$ des dreidimensionalen Bereichs $V = \{(x,y,z) :\ 0 \le x \le 1, 0 \le y \le 1, 0 \le z \le f(x,y)\}$, wobei der Einfachheit halber $f \ge 0$ vorausgesetzt sei. Stellen wir uns zwecks Veranschaulichung das Einheitsquadrat $[0,1]^2$ in n^2 kleine Quadrate $Q_{i,j}$, $i, j, = 0, 1, 2, \ldots, n-1$, jeweils mit Seitenlänge $\frac{1}{n}$ zerlegt vor. Zunächst sei f auf jedem $Q_{i,j}$ konstant mit Wert $f_{i,j} = f(\frac{i}{n}, \frac{j}{n})$. In Analogie zum eindimensionalen Fall lässt sich das gesuchte Volumen $\lambda_3(V)$ auf mehrere Arten schreiben, z.B.:

$$\lambda_3(V) = \int_{[0,1]^2} f(x,y)\,d\lambda_2(x,y) = \sum_{i,j=0}^{n-1} \int_{Q_{i,j}} f(x,y)\,d\lambda_2(x,y) = \sum_{i,j=0}^{n-1} \frac{1}{n^2} f\left(\frac{i}{n}, \frac{j}{n}\right)$$

Die Summation kann auch als eine iterierte aufgefasst werden, und zwar mittels Vertauschung der Summationsreihenfolge auf zwei Arten, also:

$$\int_{[0,1]^2} f(x,y)\,d\lambda_2(x,y) = \sum_{i=0}^{n-1} \frac{1}{n} \left(\sum_{j=0}^{n-1} \frac{1}{n} f\left(\frac{i}{n}, \frac{j}{n}\right) \right) = \sum_{j=0}^{n-1} \frac{1}{n} \left(\sum_{i=0}^{n-1} \frac{1}{n} f\left(\frac{i}{n}, \frac{j}{n}\right) \right).$$

Die inneren Summen in den großen Klammern stellen jeweils den Wert eines eindimensionalen Integrals von f dar, wobei jeweils eine der Koordinaten festgehalten wird und die andere das Integrationsintervall $[0,1]$ durchläuft. Entsprechend sind die äußeren Summen die Integrale über die Werte dieser inneren Integrale, wenn die im inneren Integral fixierte Variable läuft. Und damit ist der Satz von Fubini für ein $f : [0,1]^2 \to \mathbb{R}$, das auf den Teilbereichen $Q_{i,j}$ konstant ist, bewiesen. Auch Integrale über andere, z.B. auf $[0,1]$ stetige f, lassen sich nach derselben Formel berechnen, weil, genauso wie im Eindimensionalen, beliebig gute Approximationen durch Treppenfunktionen im obigen Sinn möglich sind. Für allgemeinere Bedingungen, unter denen der Satz immer noch gilt, sei etwa auf die ausführlichen Kapitel über Produktmaße in Bauer (1992) oder Rudin (2009) verwiesen.

Ist $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum, so versteht man unter einer *Zufallsgröße* oder auch *Zufallsvariablen* eine Abbildung $X : \Omega \to \mathbb{R}$ (manchmal lässt man als Wertemenge von X auch $\mathbb{R} \cup \{\infty\}$, $\mathbb{R}^n$ oder andere Verallgemeinerungen zu), für die zu jedem reellen Intervall I die Wahrscheinlichkeit, dass $X(\omega) \in I$, definiert ist. Diese Eigenschaft heißt *Messbarkeit der Funktion X* und ist offenbar äquivalent damit, dass die Urbilder $X^{-1}(I) = \{\omega \in \Omega : X(\omega) \in I\}$ von Intervallen I unter X in $\mathcal{A}$ liegen, also messbare Mengen sind. Somit ist durch die Festsetzung $P_X(A) = P(X^{-1}(A))$ ein W-Maß P_X auf $\mathbb{R}$ definiert, die sogenannte (W-)*Verteilung* von X (hier auf $\mathbb{R}$). Wegen ihrer Messbarkeit können Zufallsgrößen (jedenfalls, wenn sie beschränkt sind) sinnvoll integriert werden. Insbesondere ist das (gleich näher zu erläuternde) Integral

$$\mathrm{E}(X) = \int_\Omega X(\omega)\,dP(\omega) = \int_\mathbb{R} x\,dP_X(x),$$

definiert, welches man den *Erwartungswert* von X oder auch von P_X nennt.

Auch ohne die gesamte maßtheoretisch basierte Integrationstheorie aufzurollen, lässt sich leicht verstehen, warum tatsächlich die Interpretation von $\mathrm{E}(X)$ als Mittelwert sinnvoll ist. Nehmen wir dazu zunächst an, dass $\Omega = \{\omega_1, \ldots, \omega_n\}$ eine n-elementige Menge ist, auf der die Wahrscheinlichkeitsverteilung P durch $P(\{\omega_i\}) = p_i$ mit $0 \leq p_i \leq 1$ und $\sum_{i=1}^n p_i = 1$ gegeben sei. Auf dem entsprechenden W-Raum liege die Zufallsgröße $X : \Omega \to \mathbb{R}$, $\omega_i \mapsto x_i$ vor. Definitionsgemäß ist in diesem Fall das Integral für den Erwartungswert gleich

$$\mathrm{E}(X) = \sum_{i=1}^n p_i x_i.$$

Ist P die diskrete Gleichverteilung auf Ω, also $p_i = \frac{1}{n}$, so ist $\mathrm{E}(X)$ also gerade das arithmetische Mittel der Werte $x_1, \ldots, x_n$. Im allgemeineren Fall wird in der Summe jede der möglichen Zahlen x_i mit der Wahrscheinlichkeit p_i gewichtet, mit der sie von X als Wert angenommen wird. Das in Abschnitt 4.1 noch genauer zu behandelnde Gesetz der großen Zahlen besagt, dass die Mittel (bei oft wiederholten unabhängigen Versuchen) gegen den Erwartungswert $\mathrm{E}(X)$ konvergieren. Der Erwartungswert von X lässt sich also als jener Wert deuten, der als langfristiger Mittelwert nach oftmaliger, unabhängiger Wiederholung von X zu erwarten ist.

Zur Absicherung des Verständnisses betrachten wir als Beispiel die stetige Gleichverteilung auf $\Omega = [0,1]$. In diesem Fall wird das Integral in der obigen Definition von $\mathrm{E}(X)$ zu einem gewöhnlichen Integral auf dem Einheitsintervall. Sein Wert entspricht der (vorzeichenbehafteten) Fläche zwischen dem Graphen der Funktion $X : [0,1] \to \mathbb{R}$ und der x-Achse im Bereich $0 \leq x \leq 1$. Der mittlere Wert ergibt sich nach Division durch die Länge des Integrationsintervalls, die hier aber (so wie das Gesamtmaß jedes Wahrscheinlichkeitsraumes) gleich 1 ist. Also können wir $\mathrm{E}(X)$ auch in diesem Fall als Mittelwert von X interpretieren.

3.2. Der Einfluss von Unabhängigkeit auf Erwartungswert und Varianz

Den Erwartungswert $\mathrm{E}(X)$ einer Zufallsgröße X haben wir als Integral $\mathrm{E}(X) = \int_\Omega X(\omega)\,dP(\omega)$ definiert. Somit ist der Erwartungswert linear, insbesondere additiv, d.h. $\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$ für zwei auf demselben W-Raum definierte Zufallsgrößen $X_i : \Omega \to \mathbb{R}$. Für das Produkt dürfen wir keine so einfache Beziehung erwarten, weil das Integral bekanntlich nicht multiplikativ ist: Für $X_1 = X_2 : x \mapsto x$ auf $[0,1]$ bezüglich des Lebesguemaßes $P = \lambda$ z.B. ist $\mathrm{E}(X_1) = \mathrm{E}(X_2) = \frac{1}{2}$, während $\mathrm{E}(X_1 \cdot X_2) = \mathrm{E}(X_1^2) = \frac{1}{3} \neq \frac{1}{4} = \mathrm{E}(X_1) \cdot \mathrm{E}(X_2)$. Sind X_1 und X_2 hingegen unabhängig, so können wir im Sinne von Abschnitt 2.3 den Produktraum $\Omega \times \Omega$ mit Produktmaß P_2 und den Satz von Fubini heranziehen und erhalten:

$$\mathrm{E}(X_1 \cdot X_2) = \int_{\Omega \times \Omega} X_1(\omega_1) \cdot X_2(\omega_2)\,dP_2(\omega_1, \omega_2) = \int_\Omega X_1(\omega_1) \left(\int_\Omega X_2(\omega_2)\,dP(\omega_2) \right) dP(\omega_1)$$

Das innere Integral in der großen Klammer ist die Zahl $\mathrm{E}(X_2)$, die wir herausheben können, und das verbleibende Intergral ist $\mathrm{E}(X_1)$. Also haben wir den äußerst nützlichen Multiplikationssatz für den Erwartungswert unabhängiger Zufallsgrößen X_1, X_2 bewiesen:

$$\mathrm{E}(X_1 \cdot X_2) = \mathrm{E}(X_1) \cdot \mathrm{E}(X_2)$$

Wir wenden uns nun der *Varianz* $\mathrm{Var}(X)$ einer Zufallsgröße X zu. Sie ist eine Messgröße dafür, mit welchen Abweichungen vom Erwartungswert bei X zu rechnen ist. Es erweist sich als zweckmäßig, $\mathrm{Var}(X)$ als den Erwartungswert des Quadrats dieser Abweichung zu definieren, also

$$\mathrm{Var}(X) = \mathrm{E}\left((X - \mathrm{E}(X))^2\right) = \int_\Omega (X - \mathrm{E}(X))^2 \, dP.$$

Ist X nach P_X verteilt und $\mu = \mathrm{E}(X)$ der Erwartungswert, so wollen wir den Wert $\sigma^2 = \mathrm{Var}(X) = \int_\mathbb{R} (x-\mu)^2 \, dP_X(x)$ auch die Varianz von P_X nennen, $\sigma = \sqrt{\sigma^2}$ heißt auch *Streuung*. Klarerweise ist die Varianz quadratisch homogen: $\mathrm{Var}(cX) = c^2 \mathrm{Var}(X)$. Durchwegs elementare Rechnungen, die auf der eben bewiesenen Multplikativität des Erwartungswertes bei Unabhängigkeit beruhen[5], zeigen für unabhängige Zufallsgrößen X_i aber auch die für eine quadratische Größe noch viel bemerkenswertere Additivität:

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i).$$

3.3. Bedingte Unabhängigkeit, Exponentialverteilung und Dichte

Wir kehren nochmals zurück zum wiederholten Münzwurf, d.h. zu unabhängigen, identisch nach der sogenannten *Alternativverteilung* A_p verteilten Zufallsgrößen X_i, die also mit Wahrscheinlichkeit p den Wert 1 annehmen, mit $1-p$ den Wert 0. Aus unseren bisherigen Überlegungen geht hervor, dass die Wahrscheinlichkeit p_k, genau beim k-ten Wurf erstmals eine 1 zu werfen, $p_k = (1-p)^{(k-1)} p$ ist. Dabei handelt es sich um eine diskrete Verteilung auf der Menge $\Omega = \mathbb{N} \setminus \{0\} = \{1, 2, \ldots\}$, die sogenannte *geometrische Verteilung*. Mit der Formel für die geometrische Reihe prüft man leicht $\sum_{k=1}^\infty p_k = 1$ nach. Ob wir beim nächsten Wurf eine 0 oder eine 1 bekommen, wird wegen der Unabhängigkeit der einzelnen Münzwürfe nicht durch die Vorgeschichte beeinflusst. Man spricht von *Gedächtnislosigkeit*.

Die geometrische Verteilung hat ein interessantes stetiges Analogon, mit dem man Ereignisse wie etwa den Zerfall eines radioaktiven Teilchens modelliert. Seine Tendenz zu zerfallen ist nämlich, wie aus der Erfahrung sehr gut abgesichert ist, unabhängig davon, wie lange es schon in nicht zerfallenem Zustand vorliegt, mit anderen Worten: Es hat kein Gedächtnis. Wir wollen die Verteilung der Zufallsgröße $X \geq 0$ ermitteln, welche den Zeitpunkt des Zerfalls beschreibt. Die Gedächtnislosigkeit besagt

$$P(X \geq s) = P(X \geq s+t | X \geq t) = \frac{P(X \geq s+t)}{P(X \geq t)}.$$

Die Funktion $G(x) = P(X \geq x)$ erfüllt also die Funktionalgleichung $G(s+t) = G(s)G(t)$ der Exponentialfunktion (vgl. Winkler (2012)). Weil G außerdem monoton fallend ist, muss es einen Parameter $c > 0$ geben mit $G(x) = e^{-cx}$ für alle $x \geq 0$. Die zugehörige Verteilung P_X hat also die Verteilungsfunktion $F(x) = P(X \leq x) = 1 - e^{-cx}$ und die sogenannte Dichte $f(x) = F'(x) = ce^{-cx}$ und heißt *Exponentialverteilung* zum Parameter c.

Ganz allgemein versteht man unter der *Verteilungsfunktion* F einer W-Verteilung P auf $\mathbb{R}$ die Funktion $F(x) = P(]-\infty, x])$. Eine reelle Funktion f nennt man *Dichte* von P (bezüglich des Lebesguemaßes

[5] Für den Beweis untersuchen wir zwei unabhängige Zufallsgrößen X und Y (der Rest folgt mittels Induktion): $\mathrm{Var}(X+Y) = \mathrm{E}((X+Y) - \mathrm{E}(X+Y))^2 = \mathrm{E}((X - \mathrm{E}(X)) + (Y - \mathrm{E}(Y)))^2 = \mathrm{Var}(X) + 2\mathrm{E}(X - \mathrm{E}(X))(Y - \mathrm{E}(Y)) + \mathrm{Var}(Y)$. Der gemischte Term wird (hier verwenden wir die oben hergeleitete Multiplikativität von E bei Unabhängigkeit) zu $2\mathrm{E}(X - \mathrm{E}(X))\mathrm{E}(Y - \mathrm{E}(Y))$. Die Additivität von E ausnutzend erhalten wir $\mathrm{E}(X - \mathrm{E}(X)) = \mathrm{E}(X) - \mathrm{E}(\mathrm{E}(X)) = 0$. ($\mathrm{E}(X)$ ist eine konstante Zahl, die als Zufallsgröße natürlich gleich ihrem Erwartungswert ist, also $\mathrm{E}(\mathrm{E}(X)) = \mathrm{E}(X)$.) Somit verschwindet im Ausdruck für $\mathrm{Var}(X+Y)$ der gemischte Term und wir haben tatsächlich $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Ergänzende Bemerkung: Für zentrierte Zufallsgrößen X, Y (d.h. $\mathrm{E}(X) = \mathrm{E}(Y) = 0$), lässt sich mit $Z = X + Y$ also schreiben: $\mathrm{E}(X^2) + \mathrm{E}(Y^2) = \mathrm{E}(Z^2)$. Die formale Ähnlichkeit mit dem Satz von Pythagoras, ist durchaus von Interesse. Der Erwartungswert $\mathrm{E}(XY)$ eines Produktes zweier Zufallsgrößen hat nämlich alle Eigenschaften, die man von einem Skalarprodukt erwartet. Gilt $\mathrm{E}(X - \mathrm{E}(X))(Y - \mathrm{E}(Y)) = 0$, so lässt sich das daher als Orthogonalität der Zufallsgrößen $X - \mathrm{E}(X)$ und $Y - \mathrm{E}(Y)$ deuten, die Additivität der Varianzen also tatsächlich als Anwendung des Satzes von Pythagoras. Der Term $\mathrm{E}(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))$ heißt *Kovarianz* von X und Y. Verschwindet sie, so heißen X und Y *unkorreliert*. Aus unabhängig folgt also unkorreliert, die Umkehrung gilt nicht.

λ), wenn $P(A) = \int_A f\,d\lambda$ für alle Intervalle und, als Folgerung aus dem sogenannten Fortsetzungssatz der Maßtheorie, deshalb sogar für alle messbaren Mengen A gilt. Besitzt F eine auf jedem Intervall Riemann-integrierbare Ableitung, so ist nach dem Hauptsatz der Differential- und Integralrechnung $f = F'$ zugleich Dichte von P. Allerdings besitzen zum Beispiel diskrete Verteilungen auf $\mathbb{R}$, die also auf einer endlichen oder höchstens abzählbaren Menge konzentriert sind, sicher keine Dichtefunktionen bezüglich λ.[6]

3.4. Verteilung der Summe unabhängiger Zufallsgrößen und Faltung von Maßen

In allen bisherigen Beispielen haben wir konkrete Wahrscheinlichkeiten unter der Annahme von Gleichverteilung (diskret oder kontinuierlich) bestimmt. Meist hat man es aber mit anderen Verteilungen zu tun, bei denen insbesondere die elementare Wahrscheinlichkeitsdefinition nach dem Muster *günstige durch mögliche Fälle* nicht naiv anwendbar ist. Bei den wichtigsten davon spielt wieder stochastische Unabhängigkeit eine entscheidende Rolle. Einige wenige sollen nun besprochen werden.

Wir setzen beim Beispiel n-maligen Münzwurfs an, wobei wir die Versuchsausgänge $X_1, X_2, \ldots, X_n$ mit 0 oder 1 kodieren und annehmen, dass beide Möglichkeiten jeweils mit Wahrscheinlichkeit $\frac{1}{2}$ auftreten. Wir stellen die Frage, mit welcher Wahrscheinlichkeit $P(S = k)$ die Anzahl $S = S_n = \sum_{i=1}^n X_i$ der i mit $X_i = 1$ nach n Würfen den Wert k annimmt. Bekanntlich lautet die Antwort $P(S = k) = \frac{1}{2^n}\binom{n}{k}$. Denn es gibt genau $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ Möglichkeiten, aus den n Würfen k auszuwählen (für genau die das Ergebnis 1 lautet), und jede davon hat (hier setzen wir wieder Unabhängigkeit voraus) die Wahrscheinlichkeit $\frac{1}{2^n}$. Etwas allgemeiner kann man statt Gleichverteilung auch voraussetzen, dass $P(X_i = 1) = p$ mit einem $p \in [0,1]$, das auch $\neq \frac{1}{2}$ sein darf. In diesem Fall hat man

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

Allerdings ist hier eine Präzisierung angebracht. Mit welchem W-Raum $(\Omega, \mathcal{A}, P)$ haben wir es in diesem Beispiel überhaupt zu tun? Es erweist sich als sinnvoll, $\Omega = \{0,1\}^n$ zu setzen. Denn dann ist jedes $\omega \in \Omega$ von der Form $\omega = (a_1, a_2, \ldots, a_n)$ mit $a_i = 0$ oder $a_i = 1$, je nach Ausgang des i-ten Münzwurfs. Das W-Maß P ist eindeutig bestimmt durch die Forderung $P(\{\omega\}) = P(\{(a_1, a_2, \ldots, a_n)\}) = p^k(1-p)^{n-k}$, wenn k die Anzahl der i mit $a_i = 1$ ist. Die interessierenden Zufallsgrößen sind $X_i : \omega = (a_1, a_2, \ldots, a_n) \mapsto a_i$, $i = 1, \ldots, n$, und $S : \omega = (a_1, a_2, \ldots, a_n) \mapsto \sum_{i=1}^n X_i(\omega) = \sum_{i=1}^n a_i$. Wir sagen, dass alle X_i *identisch verteilt* sind, weil sie alle dieselbe W-Verteilung auf $\Omega_0 = \{0,1\}$ haben, nämlich die Alternativverteilung A_p mit Parameter p (vgl. Abschnitt 3.3). Bei den X_i handelt es sich aber auch um *unabhängige Zufallsgrößen*, weil sich das gemeinsame W-Maß P auf $\Omega = \{0,1\}^n$ als Produktmaß ihrer einzelnen Verteilungen ergibt.

Die Verteilung $\mathrm{B}_{n,p}$ von S dagegen heißt (wegen der auftretenden Binomialkoeffizienten) *Binomialverteilung* auf der Menge $\{0,1,\ldots,n\}$ und erfüllt

$$\mathrm{B}_{n,p}(\{k\}) = P(S = k) = p^k(1-p)^{n-k}\binom{n}{k}.$$

Sie ist ein Beispiel für eine *Faltung*, und zwar hier der n-fachen Faltung von A_p mit sich selbst.

Sind ganz allgemein $X_1, \ldots, X_n$ unabhängige Zufallsgrößen, die nach den W-Maßen $P_{X_1}, \ldots, P_{X_n}$ verteilt sind, so nennt man die (tatsächlich nur von den P_{X_i} abhängige) W-Verteilung P_S ihrer Summe $S = X_1 + \ldots + X_n$ die *Faltung* der Maße P_{X_i} und schreibt $P_S = P_{X_1} * \ldots * P_{X_n}$. Diese Konstruktion lässt sich, wie wir uns nun überlegen wollen, auch auf stetige Verteilungen übertragen.

Seien die unabhängigen Zufallsgrößen X_1 und X_2 nach dem Lebesguemaß $\lambda_{[0,1]}$ auf $[0,1]$, also stetig gleichverteilt. Ihre Summe $S = X_1 + X_2$ lässt sich als Funktion $S : [0,1] \to [0,2]$, $(x_1, x_2) \mapsto x_1 + x_2$ deuten. (In dieser Notation haben wir den hier nicht näher spezifizierten Raum Ω unterdrückt.) Die Verteilung

6 Es gibt noch andere Beispiele, die Schwierigkeiten machen, sogenannte singuläre W-Maße. Das klassische Beispiel ist ein W-Maß, das auf der in $\mathbb{R}$ nirgends dichten Cantormenge konzentriert ist, wo aber jeder einzelne Punkt Maß 0 erhält.

P_S von S lässt sich am besten durch die Dichte $f : [0,2] \to \mathbb{R}$, $f(s) = s$ für $0 \le s \le 1$ und $f(s) = 2-s$ für $1 \le s \le 2$ beschreiben, denn es gilt $P_S(A) = \int_A f(x)\, d\lambda(x)$.

Der Leser ist eingeladen, sich klarzumachen, dass ganz allgemein die Dichte f der Faltung zweier Maße P_1 und P_2 auf $\mathbb{R}$ mit Dichten f_1 und f_2 nach der Formel $f(x) = \int_{\mathbb{R}} f_1(t) f_2(x-t)\, d\lambda(t)$ berechnet werden kann. Der zentrale Grenzwertsatz, dem wir uns noch ausführlich widmen werden, besagt grob gesprochen, dass (unter relativ schwachen technischen Voraussetzungen) die iterierte Faltung einer Folge von W-Verteilungen nach geeigneter Normierung gegen eine Standardnormalverteilung konvergiert.

4. Unabhängigkeit als Schlüssel zu wichtigen Sätzen

4.1. Gesetze der großen Zahlen

Für den Beweis des Gesetzes der großen Zahlen erweist sich die sogenannte *Tschebyscheffsche Ungleichung*

$$P(|Y - \mathrm{E}(Y)| \ge \varepsilon) \le \frac{\mathrm{Var}(Y)}{\varepsilon^2}$$

für eine Zufallsgröße Y als sehr nützlich.[7] Und zwar wenden wir sie auf die gemittelten Summen $Y_n = \frac{1}{n} S_n$, $S_n = \sum_{i=1}^n X_i$, unabhängiger, identisch verteilter Zufallsgrößen X_i mit $\mathrm{E}(X_i) = \mu$ und $\mathrm{Var}(X_i) = \sigma^2$ an. Aus Additivität und quadratischer Homogenität der Varianz folgt $\mathrm{Var}(Y_n) = \frac{1}{n^2}\mathrm{Var}(S_n) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$. Die Tschebyscheffsche Ungleichung angewandt auf Y_n liefert somit $P(|Y_n - \mathrm{E}(Y_n)| \ge \varepsilon) \le \frac{\sigma^2}{n\varepsilon^2}$. Da σ^2 und ε feste positive Zahlen sind, haben wir bewiesen:

Sind für die identisch verteilten, unabhängigen Zufallsgrößen X_i Erwartungswert μ und Varianz σ^2 endlich, so konvergiert für jedes $\varepsilon > 0$ die Wahrscheinlichkeit, dass $\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| > \varepsilon$ gilt, gegen 0 (für $n \to \infty$). (Schwaches Gesetz der großen Zahlen)

Mit etwas mehr Aufwand[8] beweist man unter denselben Voraussetzungen sogar:

Die Mittel $\frac{1}{n}\sum_{i=1}^n X_i$ konvergieren mit Wahrscheinlichkeit 1 gegen $\mu = E(X_i)$. (Starkes Gesetzes der großen Zahlen)

[7] Sie beruht auf einer sehr einfachen unteren Abschätzung für die Varianz $\mathrm{Var}(Y) = \mathrm{E}(Y - \mathrm{E}(Y))^2 = \int_\Omega (Y(\omega) - \mathrm{E}(Y))^2\, dP(\omega)$ einer Zufallsgröße Y: Für ein beliebiges $\varepsilon > 0$ können wir den Integrationsbereich aufteilen in $\Omega = \Omega_1 \cup \Omega_0$ mit $\Omega_1 = \Omega_1(\varepsilon) = \{\omega : |Y(\omega) - \mathrm{E}(Y)| \ge \varepsilon\}$ und $\Omega_0 = \Omega \setminus \Omega_1$. Weil der Integrand überall positiv und auf Ω_1 mindestens ε^2 ist, folgt $\mathrm{Var}(Y) \ge \varepsilon^2 P(\Omega_1)$ und daraus die Tschebyscheffsche Ungleichung.

[8] Um technische Komplikationen, die nur von der Hauptidee ablenken, zu vermeiden, wollen wir annehmen, dass die X_i beschränkt sind, etwa $|X_i(\omega)| \le 1$ für alle $\omega \in \Omega$. (Im unbeschränkten Fall kann man sich helfen, indem man aus der Voraussetzung, dass Erwartungswert und Varianz endlich sind, obere Abschätzungen für die Wahrscheinlichkeit großer Werte der X_i gewinnt, die ebenfalls für den Beweis hinreichen. Es würde sogar genügen, statt Unabhängigkeit Unkorreliertheit vorauszusetzen und, für eine entsprechend modifizierte Fassung, statt identischer Verteilung Beschränktheit der Erwartungswerte und Varianzen.)

Wir betrachten die Ausnahmemengen $A_{\varepsilon,n}$ aller ω, wo das n-te Mittel $Y_n(\omega) = \frac{1}{n}\sum_{i=1}^n X_i(\omega)$ vom Erwartungswert μ mehr als ε abweicht. Bereits weiter oben haben wir (in etwas anderer Notation) aus der Tschebyscheffschen Ungleichung die Beziehung $P(A_{\varepsilon,n}) \le \frac{\sigma^2}{n\varepsilon^2}$ geschlossen. Weil die unendliche Summe über alle $\frac{1}{n^2}$ (Quadrate n^2 zu betrachten, ist der entscheidende Trick; außerdem wird jetzt gleich die σ-Additivität einfließen) konvergiert, ist auch die Wahrscheinlichkeit $P\left(\bigcup_{n=N}^\infty A_{\varepsilon,n^2}\right) \le \frac{\sigma^2}{\varepsilon^2}\sum_{n=N}^\infty \frac{1}{n^2}$ nicht nur endlich, sondern bei hinreichend großem N auch beliebig klein. Dies bedeutet aber: Die Wahrscheinlichkeit $P(A_\varepsilon)$ dafür, dass für unendlich viele Quadratzahlen n^2 der Betrag der Differenz $|Y_{n^2} - \mu|$ größer als ε ist, muss 0 sein. Dies gilt für alle $\varepsilon > 0$. Daher ist (wieder σ-Additivität!) auch $A = \bigcup_{k=1}^\infty A_{\frac{1}{k}}$ eine Nullmenge. Für jedes $\omega \in \Omega \setminus A$ gilt nach Konstruktion $\lim_{n\to\infty} Y_{n^2}(\omega) = \mu$. Um den Beweis zu komplettieren, müssen wir noch auf die Konvergenz $\lim_{n\to\infty} Y_n(\omega) = \mu$ für beliebige n, also nicht nur für Quadratzahlen n^2 schließen. An dieser Stelle ist uns die Beschränktheit der X_i eine angenehme technische Erleichterung. Ist nämlich $n^2 \le i < (n+1)^2$, so folgt $|i - n^2| \le 2n$. In den Mitteln Y_{n^2} und Y_i treten also maximal $2n$ verschiedene Terme auf, je zwei können sich um maximal $2|X_i| \le 2$ unterscheiden, was nach Division durch n^2 bzw. $i \ge n^2$ zu einer maximalen Differenz $|Y_{n^2} - Y_i| \le \frac{2n}{n^2} = \frac{2}{n}$ führt. Da diese Größe gegen 0 konvergiert und die Y_{n^2} gegen μ konvergieren, müssen es also auch die Y_n tun.

Als etwas saloppe, aber einprägsame Formulierung des Gesetzes der großen Zahlen, das man mit gutem Recht den ersten Hauptsatz der Stochastik nennen kann, sollte vom Unterricht wenigstens hängen bleiben: *Arithmetische Mittel unabhängiger Zufallsgrößen konvergieren auf lange Sicht gegen den Erwartungswert.* Die entscheidende Rolle der Unabhängigkeit lässt sich leicht durch plastische Beispiele plausibel machen.[9]

4.2. Zentraler Grenzwertsatz und Normalverteilung

Haben wir das Gesetz der großen Zahlen als den ersten Hauptsatz der Stochastik bezeichnet, so liegt auf der Hand, was der zweite sein soll: Der sogenannte zentrale Grenzwertsatz, der die überragende Rolle der Normalverteilung in der Stochastik zum Ausdruck bringt. Man kann durchaus von einer Verschärfung des Gesetzes der großen Zahlen sprechen. In beiden Fällen geht es nämlich um die Abweichung arithmetischer Mittel von einem Erwartungswert, allerdings mit unterschiedlich feinem Maßstab der Beobachtung.[10]

Zur Illustration: Seien die X_i unabhängig identisch verteilte Zufallsgrößen mit Erwartungswert $\mathrm{E}(X_i) = \mu$ und $\mathrm{Var}(X_i) = \sigma^2$. Dann geht es im (starken) Gesetz der großen Zahlen um die Konvergenzaussage $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu) = 0$. Die Summe der vorzeichenbehafteten Abweichungen vom Erwartungswert wird also klein im Verhältnis zu n. Der zentrale Grenzwertsatz besagt nun, dass diese Summe sogar nur von der Größenordnung $\sqrt{n}$ ist, und noch genauer: Dividieren wir sie durch $\sqrt{n}\sigma$, so erhalten wir eine Zufallsgröße mit einer Verteilung, die sich, bei hinreichend großem n, immer mehr einer festen Verteilung annähert, der sogenannten (Standard-)*Normalverteilung* $\mathrm{N}_{0,1}$ mit Erwartungswert $\mu = 0$ und Varianz $\sigma^2 = 1$. $\mathrm{N}_{0,1}$ hat als Dichtefunktion die berühmte *Gaußsche Glockenkurve*

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

und wird bei allgemeinem μ und σ^2 zur Normalverteilung $\mathrm{N}_{\mu,\sigma^2}$ mit Dichtefunktion $\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.

Allgemein sagt man, die Folge der Zufallsgrößen X_i (die alle endliche Erwartungswerte und Varianzen besitzen mögen), erfüllt den *zentralen Grenzwertsatz*, wenn die W-Verteilungen P_{Y_n} der Zufallsgrößen

$$Y_n = \frac{1}{\sigma_n} \sum_{i=1}^{n} (X_i - \mathrm{E}(X_i)) \quad \text{mit} \quad \sigma_n^2 = \mathrm{Var}(X_1 + \ldots + X_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)$$

gegen die Standardnormalverteilung $\mathrm{N}_{0,1}$ konvergieren. (Das wiederum soll bedeuten, dass für jedes Intervall $I \subseteq \mathbb{R}$ die Konvergenz $\lim_{n\to\infty} P_{Y_n}(I) = \mathrm{N}_{0,1}(I)$ gilt.)

Hinreichend für den zentralen Grenzwertsatz ist beispielsweise, dass die X_i neben unabhängig auch identisch verteilt sind mit endlichem $\mu = \mathrm{E}(X_n)$ und $\sigma^2 = \mathrm{Var}(X_n)$. Die Frage, unter welchen schwächeren Voraussetzungen an die X_i der zentrale Grenzwertsatz immer noch gilt, wird in der Stochastik sehr genau untersucht. Wesentlich (wenn auch im strengen Sinne weder notwendig noch hinreichend) ist wieder die Unabhängigkeit der X_i. Setzt man sie voraus, kann man etwas komplizierte aber sehr präzise Kriterien dafür angeben, dass der zentrale Grenzwertsatz gilt. Und zwar dürfen, grob gesprochen, unter den unendlich vielen $X_1, X_2, \ldots$ keine einzelnen Summanden größenmäßig dominieren. Das ist aber keinesfalls notwendig, denn: Die Summe irgendwelcher unabhängiger und (nicht notwendig standardisierter) normalverteilter Zufallsgrößen ist wieder normalverteilt (wobei sich Erwartungswerte und Varianzen aufsummieren).

Als Folge des zentralen Grenzwertsatzes tritt Normalverteilung typischerweise dann auf, wenn sich eine Größe als Summe sehr vieler kleiner und voneinander unabhängiger Einflüsse ergibt. Es liegt auf der Hand, dass dies wenigstens approximativ in zahllosen Anwendungssituationen zutrifft.

9 Seien etwa die X_i' unabhängige Zufallsgrößen, welche die Werte 0 und 1 jeweils mit Wahrscheinlichkeit $\frac{1}{2}$ annehmen. Von ihnen ausgehend definieren wir die (nicht unabhängigen) Zufallsgrößen X_n durch $X_n = X_k'$ mit $k! \leq n < (k+1)!$. Man überlegt sich leicht, dass die arithmetischen Mittel der X_n mit Wahrscheinlichkeit 1 zwischen 0 und 1 oszillieren und immer wieder nahe an 0 wie auch an 1 herankommen, also sicher nicht gegen $\mathrm{E}(X_n) = \frac{1}{2}$ konvergieren.

10 Auch das Gesetz vom iterierten Logarithmus, das hier nur als Schlagwort erwähnt sei, fügt sich in diesen Kontext.

Man darf aber keine falschen Schlüsse ziehen. Ein Beispiel: Oft wird auch bei Leistungsbeurteilungen wie etwa den Punktezahlen der einzelnen Schüler bei einer Schularbeit Normalverteilung unterstellt. Die Gesamtpunktezahl kommt ja in der Tat als Summe zahlreicher Einzelleistungen zustande. Diese sind erfahrungsgemäß aber keineswegs unabhängig. Beherrscht ein Schüler nämlich ein Stoffgebiet, so wird es dadurch in der Regel wahrscheinlicher, dass er auch ein anderes beherrscht. Deshalb ist es nicht überraschend, wenn immer wieder auch extreme Punktezahlen auftreten, und zwar häufiger als es einer Normalverteilung entspräche. Würde durch die Fragen nicht Verständnis oder Wissen abgeprüft, sondern gälte es zu raten, hätte Normalverteilung weit höhere Plausibilität. Überspitzt formuliert: Bei der Leistungsfeststellung deutet Normalverteilung darauf hin, dass die Ergebnisse eher zufällig zustande gekommen sind und deshalb wenig über die Einzelleistungen aussagen.

Wie steht es mit einem Beweis des zentralen Grenzwertsatzes? Schon die komplizierten technischen Bedingungen, unter denen er gilt, deuten darauf hin, dass weit reichende Theorie nötig ist. Und wirklich übersteigt wohl jeder Beweis auch einfacher Varianten die Möglichkeiten des Schulunterrichts. Um keine Missverständnisse über die strengen Anforderungen an mathematische Beweise aufkommen zu lassen (Heuristiken reichen nicht!), sollte das im Unterricht auch deutlich gemacht werden. Dennoch kann interessierten Schülern Sinnvolles gesagt werden. Es folgen einige Vorschläge.

Es bietet sich an, vom allereinfachsten Fall, der Binomialverteilung aus Abschnitt 3.4 mit $p = \frac{1}{2}$ auszugehen. Zunächst wird man wohl für ein angemessenes n, z.B. $n = 10$, und $k = 0, 1, \ldots, n$, die entsprechenden Werte $2^{-n}\binom{n}{k}$ in eine Graphik eintragen und die Ähnlichkeit mit der Glockenkurve feststellen. Den meisten Schülern wird das vermutlich genügen. Interessiertere werden aber möglicherweise fragen, warum die Normalverteilung gerade durch die spezielle Dichtefunktion $\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ beschrieben wird, insbesondere wie die Konstanten e und π ins Spiel kommen, die doch aus ganz anderen Zusammenhängen vertraut sind. Es liegt nahe, die Binomialkoeffizienten $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, $k = 0, 1, 2, \ldots, n$, und die darin vorkommenden Faktorielle $n!$ für $n \to \infty$ näher zu untersuchen. Wesentliches Hilfsmittel hierfür ist die Stirlingsche Formel:

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

Das Symbol $\sim$ bedeutet asymptotische Äquivalenz, d.h. dass für große n der prozentuelle Unterschied zwischen den Größen links und rechts beliebig klein wird. Glaubt man die Stirlingsche Formel, so sollte es plausibel sein, dass nach geschickter Rechnung (die alleine für den Schulunterricht wahrscheinlich bereits zu langwierig wäre) die gewünschte Beziehung zur Dichte der Normalverteilung hergeleitet werden kann. Doch wie kommen die Zahlen e und π in die Stirlingsche Formel? Was e betrifft, lässt sich das auf Schulniveau erklären, indem man den Logarithmus von $n!$ mittels Ersetzung einer Summe durch ein Integral (*Eulersche Summenformel*) recht grob abschätzt: $\ln n! = \sum_{k=1}^{n} \ln k \approx \int_1^n \ln x\, dx \approx n \ln n - n$ und somit $n! \approx e^{n \ln n - n} = \left(\frac{n}{e}\right)^n$. Damit ist das Auftreten wenigstens der Zahl e erklärt. Eine feinere Anwendung der Eulerschen Summenformel liefert sogar noch den Faktor $\sqrt{n}$. Der verbleibende Term $\sqrt{2\pi}$ ergibt sich schließlich daraus, dass das Integral über die Dichtefunktion eines W-Maßes immer gleich 1 sein muss.[11]

Kraft der sehr allgemeinen Gültigkeit des zentralen Grenzwertsatzes könnte man statt der diskreten Gleichverteilung auf einer zweielementigen Menge, als deren Faltung die Binomialverteilungen ins Spiel gekommen sind, auch von irgendwelchen anderen Verteilungen ausgehen. (Man untersuche beispielsweise die stetige Gleichverteilung $\lambda_{[0,1]}$ auf $[0,1]$.) Bei iterierter Faltung tritt unweigerlich ein glättender Effekt ein, wobei schließlich alle Besonderheiten des ursprünglichen Maßes von der Normalverteilung als Grenzwert nivelliert werden. In Georgii (2009) findet sich ein Beweis des zentralen Grenzwertsatzes für identisch verteilte Zufallsgrößen, der davon Gebrauch macht. Dies und erst recht allgemeinere Versionen, wo etwa die doch etwas restriktive Voraussetzung *identisch verteilt* abgeschwächt wird, können aber bestenfalls in fortgeschrittenen Vorlesungen im Rahmen eines universitären Fachstudiums bewiesen

11 Hinweis: Die Berechnung des Integrals $I = \int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}$ gelingt auf elegante Weise mittels Fubini: $I^2 = \int_{\mathbb{R}} e^{-x^2}\, dx \int_{\mathbb{R}} e^{-y^2}\, dy = \int_{\mathbb{R}^2} e^{-(x^2+y^2)}\, dx dy$, und anschließender Transformation auf Polarkoordinaten $x = r\cos\varphi$, $y = r\sin\varphi$. Auf diesem Wege kommt dann auch die Zahl π ins Spiel.

werden. Die Standardmethode sind Fouriertransformation von Maßen und, darauf beruhend, die Verwendung charakteristischer Funktionen von Zufallsgrößen.[12]

Vom Standpunkt der allgemeinen Theorie ist es auch sehr interessant, dass die Menge der Normalverteilungen bezüglich Faltung abgeschlossen ist gemäß der Formel $N_{\mu_1,\sigma_1^2} * N_{\mu_2,\sigma_2^2} = N_{\mu_1+\mu_2,\sigma_1^2+\sigma_2^2}$. Man spricht, weil ganz generell das Assoziativgesetz $(P_1 * P_2) * P_3 = P_1 * (P_2 * P_3)$ gilt, von einer Faltungshalbgruppe. Die der Nomalverteilungen ist sogar *unbegrenzt teilbar*, weil jedes N_{μ,σ^2} als n-fache Faltungsiterierte von $N_{\frac{\mu}{n},\frac{\sigma^2}{n}}$ auftritt.

Auch wenn die hier angedeuteten Zusammenhänge weit über die Schulmathematik hinausreichen: Mathematisch interessierten jungen Menschen kann damit eine Ahnung vom Charakter der höheren Mathematik vermittelt werden. Vielleicht kommen sie auf den Geschmack und zeigen sogar unverhoffte Begabungen.

5. Unabhängige Zuwächse versus unabhängige Zufallsgrößen

5.1. Stochastische Prozesse mit unabhängigen Zuwächsen

Folgen von Zufallsgrößen oder deren Summen, wie wir sie bisher betrachtet haben, sind einfache Beispiele *stochastischer Prozesse*. Generell geht es dabei um Familien von Zufallsgrößen, die auf einem gemeinsamen W-Raum definiert und durch irgendwelche Gesetzmäßigkeiten miteinander verknüpft sein dürfen. Dabei können sich stochastische mit deterministischen Elementen mischen.

Ein sehr einfaches Beispiel ist die sogenannte Irrfahrt. Dabei geht man von unabhängigen Zufallsgrößen X_i aus, welche mit gleicher Wahrscheinlichkeit $p = \frac{1}{2}$ die beiden Werte 1 bzw. -1 annehmen, und betrachtet die dadurch entstehenden Partialsummen $S_n = \sum_{i=1}^n X_i$. So wird der zufällige Pfad eines Teilchens modelliert, welches sich zu jedem Zeitpunkt $t = 1, 2, \ldots$ auf der reellen Achse eine Einheit entweder nach rechts oder nach links bewegt. Aus dem zentralen Grenzwertsatz folgt, dass die S_n bei großem n nach geeigneter Normierung annähernd normalverteilt sind. Nehmen die X_i statt $1, -1$ als Werte allgemeiner im $\mathbb{R}^s$ die $2s$ achsenparallelen s-dimensionalen Vektoren gleicher Schrittlänge l an, erhält man die s-dimensionale Irrfahrt.

Für $s = 3$ und bei kleinem $l > 0$ ähnelt die Irrfahrt dem vermutlich prominentesten Beispiel eines stochastischen Prozesses, nämlich der aus der Physik bekannten *Brownschen Bewegung*. Im Mikroskop zeigt sich, dass etwa die Luft winzige feste Partikel enthält, die sich auch bei Windstille ständig in erratischer Weise bewegen, weil sie von Gasmolekülen in alle Richtungen angestoßen werden. Diese Bewegungen lassen sich verblüffend gut durch das folgende stochastische Modell beschreiben.

Für jedes $t \geq 0$ sei X_t die Zufallsgröße, welche die Position eines Teilchens zum Zeitpunkt t anzeigt. (Hier nimmt X_t Werte in $\mathbb{R}^3$ an. Der einfacheren Notation halber wollen wir uns aber auf eine Koordinate beschränken.) Für alle Folgen von Zeitpunkten $s_1 \leq t_1 \leq s_2 \leq t_2 \leq \ldots \leq s_n \leq t_n$ sind die Zufallsgrößen $Y_i = X_{t_i} - X_{s_i}$, $i = 1, 2, \ldots, n$, unabhängig und normalverteilt – man spricht von *unabhängigen Zuwächsen*. Herrscht im umgebenden Medium keine Strömung (Drift), so ist $\mathrm{E}(Y_i) = 0$. Die Varianzen sind proportional zur Länge der Zeitintervalle: $\mathrm{Var}(Y_i) = (t_i - s_i)\sigma^2$ mit einem gemeinsamen σ^2, das man sich durch geeignete Wahl der Einheiten auf $\sigma^2 = 1$ normiert denken kann.

[12] Die Fouriertransformierte $\widehat{P}$ eines Maßes P auf $\mathbb{R}$ ist definiert durch $\widehat{P}(t) := \int_{\mathbb{R}} e^{itx}\, dP(x)$. Ist eine Zufallsgröße X nach P verteilt, so heißt diese Funktion $\widehat{P} : \mathbb{R} \to \mathbb{C}$ auch die charakteristische Funktion von X, symbolisch $\widehat{P} = \varphi_X$. Wenn man die involvierten Definitionen in natürlicher Weise auf komplexwertige Funktionen ausdehnt, gilt offenbar $\varphi_X(t) = \mathrm{E}(e^{itX})$. Wegen der Multiplikativität der Exponentialfunktion und, für unabhängige Zufallsgrößen X_1, X_2, des Erwartungswertes folgt: $\varphi_{X_1+X_2} = \mathrm{E}(e^{it(X_1+X_2)}) = \mathrm{E}(e^{itX_1} e^{itX_2}) = \mathrm{E}(e^{itX_1})\mathrm{E}(e^{itX_2}) = \varphi_{X_1}\varphi_{X_2}$. Für die Faltung von Maßen bedeutet das $\widehat{P_1 * P_2} = \widehat{P}_1 \widehat{P}_2$. Der Summation unabhängiger Zufallsgrößen, also der Faltung von Maßen entspricht somit die Multiplikation komplexwertiger Funktionen. In ähnlicher Weise spiegeln sich viele andere interessante Eigenschaften von Maßen wider in der Welt der Fouriertransformierten. Und weil das in einer Weise geschieht, die auch eine Rückübersetzung erlaubt, lässt sich nun erahnen, wie auch sehr allgemeine Versionen des zentralen Grenzwertsatzes bewiesen werden können: Mit Hilfe nützlicher Eigenschaften der Exponentialfunktion zeigt man, dass die Fouriertransformierten der Zufallsvariablen $\frac{1}{\sigma_n} \sum_{i=1}^n (X_i - \mathrm{E}(X_i))$ mit $\sigma_n^2 = \sum_{i=1}^n \mathrm{Var}(X_i)$ gegen die Fouriertransformierte der Standardnormalverteilung $N_{0,1}$ konvergieren.

5.2. Stichproben als unabhängige, identisch verteilte Zufallsgrößen

Bisher haben wir uns vor allem den Grundlagen der Wahrscheinlichkeitstheorie gewidmet. Doch auch in statistischen Anwendungen spielt Unabhängigkeit eine fundamentale Rolle, vor allem für den Begriff einer Stichprobe, welchen man zur Herleitung wesentlicher Eigenschaften von Schätzern und Tests heranzieht. (Für die Regressionsrechnung, auf die ich hier nicht eingehe, finden sich Anknüpfungspunkte in Pauer (2012).) Das wichtigste und typische Grundanliegen der angewandten Statistik besteht darin, aus typischerweise knappen empirischen Daten $x_1,\ldots,x_n$ auf die zugrundeliegende W-Verteilung P zu schließen. Kann man die x_i als Werte von unabhängigen und identisch (nach dem gesuchten P) verteilten Zufallsgrößen $X_1,\ldots,X_n$ interpretieren, so spricht man von einer *Stichprobe*, dem grundlegenden Konzept der beurteilenden Statistik.

Die einfachste Aufgabe besteht darin, den (hoffentlich vorhandenen) Erwartungswert $\mu = \mathrm{E}(X_i)$ der Verteilung P zu schätzen. Klarerweise bietet sich das arithmetische Mittel $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n x_i$ als Schätzung an. Denn zunächst hat $\hat{\mu}_n$, aufgefasst als Wert der Zufallsgröße $Y_n = \frac{1}{n}\sum_{i=1}^n X_i$, selbst den Erwartungswert

$$\mathrm{E}(Y_n) = \frac{1}{n}\sum_{i=1}^n \mathrm{E}(X_i) = \mu.$$

Doch dafür müssten wir nicht alle x_i heranziehen, sondern könnten uns auch mit einem einzigen begnügen. Aber darüber hinaus, so hoffen wir, streut das Mittel $\hat{\mu}_n$ viel weniger als eine einzelne Zufallsgröße X. Unsere Hoffnung lässt sich mathematisch rechtfertigen, wenn wir die X_i als unabhängig und von endlicher Varianz $\mathrm{Var}(X_i) = \sigma^2$ voraussetzen. Denn dann können wir die Additivität der Varianz verwenden und erhalten

$$\mathrm{Var}(Y_n) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{\sigma^2}{n},$$

was mit wachsendem n beliebig klein wird.

In ähnlicher Weise kann man die Varianz σ^2 mit der Größe $\frac{1}{n-1}\sum_{i=1}^n (x_i - \hat{\mu}_n)^2$ schätzen.[13] Wer sich nicht mit Erwartungswert und Varianz zufriedengeben will, kann aufs Ganze gehen und auf die Verteilung P selbst abzielen. Hierzu verwendet man am besten die sogenannte empirische Verteilungsfunktion

$$\hat{F}_n(x) = \frac{1}{n}|\{i:\ x_i \leq x,\, 1 \leq i \leq n\}|$$

als Schätzung für die tatsächliche Verteilungsfunktion F der gesuchten W-Verteilung P. Mit Hilfe des starken Gesetzes der großen Zahlen kann man recht leicht beweisen, dass für $n \to \infty$ die $\hat{F}_n$ mit Wahrscheinlichkeit 1 gleichmäßig gegen F konvergieren (*Satz von Gliwenko und Cantelli* oder auch *Hauptsatz der mathematischen Statistik*). In allen Fällen ist die Unabhängigkeit der Stichprobe die entscheidende Voraussetzung, unter der wünschenswerte Eigenschaften der Schätzer garantiert werden können.

Ganz ähnlich verhält es sich bei statistischen Tests, die allerdings einer anderen Logik folgen. Statt der Schätzung von Parametern ist das Ziel beim Testen die Entscheidung, ob eine gewisse Vermutung über P, die in diesem Zusammenhang auch Nullhypothese genannt wird, verworfen werden soll. Man verwendet für diesen Zweck wieder Testgrößen von der Art obiger Schätzer $\hat{\mu}_n$ etc. und zerlegt die Menge der möglichen Werte in eine Menge des Verwerfens und eine des Nicht-Verwerfens der Nullhypothese. Aus denselben Gründen wie beim Schätzen hängt die Qualität auch statistischer Tests entscheidend von der Unabhängigkeit der verwendeten Stichprobe ab.

[13] Dass im Nenner der Faktor $n-1$ statt n steht, rührt daher, dass man üblicherweise ja nur den Schätzer $\hat{\mu}_n$ und nicht μ selbst zur Verfügung hat, was zu einer systematischen Unterschätzung der Varianz führte, welche aber durch diese Modifikation des Nenners exakt korrigiert werden kann.

6. Resümee

Zum Abschluss sollen einige bemerkenswerte Einsichten hervorgehoben werden, zu denen die hier dargelegten Überlegungen Anlass geben.

Wahrscheinlichkeit als Maß: Für den Laien vielleicht am überraschendsten, wenn nicht sogar paradox mag die Tatsache anmuten, dass der Gegenstand der Stochastik, nämlich das Unsichere, von der Mathematik mit derselben Strenge und dadurch garantierten Sicherheit behandelt werden kann wie Zahlentheorie, Geometrie, Algebra, Analysis und andere klassische Teilgebiete der Mathematik.[14] Der Grund – und damit scheint das Paradoxe noch gesteigert zu werden – liegt nicht in starken Hypothesen bei der Modellbildung, sondern, ganz im Gegenteil, in der besonderen Allgemeinheit des Wahrscheinlichkeitsbegriffs als Maß, d.h. als σ-additive Mengenfunktion. Nicht viele einzelne Zahlen, die als Wahrscheinlichkeiten gedeutet werden können, sind also der Schlüssel zur leistungsfähigen Theorie, sondern eine einfache Eigenschaft des Maßes als Funktion, nämlich die (σ-)Additivität. Sie kann als Strukturverträglichkeit (nämlich zwischen mengentheoretischer Vereinigung und arithmetischer Addition) aufgefasst werden. Das erinnert an Homomorphiebedingungen, Funktionalgleichungen und ähnliche Eigenschaften von Abbildungen, die in mathematischen Strukturtheorien im Mittelpunkt stehen. Mathematik erweist sich also auch in der Stochastik als primär qualitativ und keinesfalls nur quantitativ orientierte Disziplin.

Maßtheorie + Unabhängigkeit = Stochastik: Als wesentliche Ingredienz, welche die Stochastik von der Maßtheorie abhebt, haben wir die stochastische Unabhängigkeit erkannt. Zwar ist sie aufs Engste mit Produktmaßen und somit mit klassischer geometrischer Flächenmessung verwandt (weshalb auch der Satz von Fubini über die Berechnung höherdimensionaler Integrale so wichtig wird). Doch gewinnen Multiplikativität der Erwartungswerte und Additivität der Varianzen unabhängiger Zufallsgrößen erst unter stochastischen Gesichtspunkten ihr volles Profil. Sie sind der technische Kern beim Beweis des Gesetzes der großen Zahlen. Beim zentralen Grenzwertsatz liegen die Dinge zwar deutlich komplizierter, und Beweise sind für den Schulunterricht außer Reichweite. Dass aber auch hier stochastische Unabhängigkeit wesentlich ist, lässt sich auch ohne höhere Analysis plausibel machen. Ähnliches gilt bei wichtigen stochastischen Prozessen wie der Brownschen Bewegung und in der Angewandten Statistik, wenn sie sich auf Stichproben aus unabhängigen Messungen bezieht. Stochastische Unabhängigkeit als roter Faden durch die Stochastik: Gibt es dafür einen tieferen Grund?

Unabhängigkeit und Zufall: In Abschnitt 2.1 haben wir die Definition der stochastischen Unabhängigkeit aus einer Analyse der alltagssprachlichen Verwendung des Wortes *Zufall* gewonnen. Fasst man die Stochastik als die mathematische Modellierung von Zufall auf, so haben wir damit eine durchaus befriedigende Erklärung gefunden, warum gerade die Unabhängigkeit ein für die Stochastik so charakteristischer Begriff ist.

Philosophen werden an dieser Stelle eventuell einwenden, dass das alles aber noch lange nicht erkläre, was Zufall oder auch Wahrscheinlichkeit dem innersten Wesen nach seien. Denn das Funktionieren eines mathematischen Modells muss noch lange nicht jedermann als endgültige Erklärung zufriedenstellen. Und gerade beim vorliegenden Themenkomplex lassen sich Assoziationsketten knüpfen, die ewige Fragen der Philosophie berühren, wie etwa: Zufall - Notwendigkeit - Determinismus - freier Wille - Ethik etc. Manch einer mag es als Schwäche der Mathematik deuten, dass sie zu manchen dieser Themen nichts Eindeutiges zu sagen hat. Man kann aber auch eine Stärke darin sehen, dass die Mathematik ihren Geltungsbereich sehr genau beschreiben und dort Einsichten bieten kann, die in ihrer Klarheit und Si-

[14] In diesem Zusammenhang sei auf ein Phänomen hingewiesen, das sich beispielsweise in der Theorie stochastischer Prozesse, speziell von Martingalen zeigt, und – grob gesprochen – in Folgendem besteht: Rein deterministische Zusammenhänge sind die klassische Domäne der Mathematik. Wie wir gesehen haben, kann die Mathematik auch dort sehr starke Aussagen machen, wo gewissermaßen der reine Zufall herrscht. Schwierig zu verstehen sind dagegen Prozesse, wo sich diese beiden Elemente in komplizierter Weise mischen. Man versucht daher, diese Prozesse in rein deterministische und rein stochastische Anteile zu zerlegen. Gelingt dies, hat man meist schon fast gewonnen.

cherheit unübertroffen sind – sogar wenn, wie im Fall der Stochastik, Wahrscheinlichkeit der Gegenstand der Betrachtung ist.

Literatur

Heinz Bauer. *Maß- und Integrationstheorie.* Walter de Gruyter. Berlin, New York (2., überarbeitete Auflage: 1992).

Heinz Bauer. *Wahrscheinlichkeitstheorie.* Walter de Gruyter. Berlin, New York (5.Auflage: 2002).

Rudolf Carnap. *Überwindung der Metaphysik durch logische Analyse der Sprache.* Erkenntnis 2 (1931/32), 219-241.

Hans-Otto Georgii. *Stochastik – Einführung in die Wahrscheinlichkeitstheorie und Statistik.* Walter de Gruyter. Berlin, New York (4., überarbeitete und erweiterte Auflage: 2009).

Andrei Nikolajewitsch Kolmogorow. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer. Berlin (1933).

Norbert Kusolitsch. *Maß- und Wahrscheinlichkeitstheorie. Eine Einführung.* Springer. Wien, New York (2011).

Franz Pauer. *Interpolation, lineare Gleichungen (mit und ohne Lösungen) und lineare Regression.* DHÖMG[15] 44 (2012), 59-60.

Walter Rudin. *Reelle und Komplexe Analysis* (deutsche Übersetzung des englischsprachigen Originals). Oldenbourg Wissenschaftsverlag (2009, 2.Auflage).

Reinhard Winkler. *Wie macht man 2 aus 1? Das Paradoxon von Banach-Tarski*. DHÖMG[15] 33 (2001), 166-196.

Reinhard Winkler. *Das Maß aller Dinge aus mathematischer Sicht – zu den Grundlagen der Integralrechnung*. DHÖMG[15] 43 (2011), 146-160.

Reinhard Winkler. *Im Anfang war die Exponentialfunktion*. DHÖMG[15] 44 (2012), 98-109.

[15] Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft (ÖMG), bisher: Didaktikhefte der ÖMG, siehe auch `http://www.oemg.ac.at/DK/Didaktikhefte/index.html`.