de Gruyter Lehrbuch

Wengenroth · Wahrscheinlichkeitstheorie

Jochen Wengenroth

Wahrscheinlichkeitstheorie

Walter de Gruyter
Berlin · New York

Prof. Dr. Jochen Wengenroth
Département de Mathématiques
Université de Liège
Grande Traverse 12 (Bât. B37)
4000 Liège
Belgien
E-Mail: J.Wengenroth@ulg.ac.be

♾ Gedruckt auf säurefreiem Papier, das die US-ANSI-Norm über Haltbarkeit erfüllt.

ISBN 978-3-11-020358-5

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Printed in Germany.
Einbandgestaltung: Martin Zech, Bremen.
Druck und Bindung: AZ Druck und Datentechnik GmbH, Kempten.

Für Annette, Hannah und Philipp.

Vorwort

Es gibt keine Wahrscheinlichkeit – so beginnt Bruno de Finetti sein Werk über Wahrscheinlichkeitstheorie. Und auch wenn diese Äußerung mindestens stark übertrieben ist, bietet sie gute Gelegenheit, den Zweck dieser Theorie zu beschreiben. Wahrscheinlichkeitstheorie ist *nicht* die Analyse eines objektiven Sachverhalts, den man Zufall nennen könnte. Ein Verfechter der gegenteiligen Ansicht wird sagen, dass etwa das Ergebnis eines Münzwurfs oder einer Lottoziehung prinzipiell unvorhersehbar sei und durch eben diesen Zufall „gesteuert" werde – was immer das auch sein mag: Karl Popper spricht von *propensities* und stellt sich darunter wohl etwas ähnliches wie magnetische Kräfte vor, von denen einige in die eine Richtung ziehen und manche in die andere, so dass das Ergebnis ein Mittel der wirkenden Kräfte ist. Diesem Verfechter kann man mit dem Argument begegnen, dass *Indeterminiertheit* überhaupt nichts mit der Funktion von Münzwürfen oder Lottoziehungen zu tun hat: Ob die Würfel noch rollen oder aber im Würfelbecher schon gefallen sind, so dass ein Glücksspieler das Ergebnis nicht sieht, spielt für sein Wettverhalten gar keine Rolle – er braucht so oder so ein vernünftiges Modell, in dem er die Situation beschreiben und seine Strategie planen kann.

Ob es nun Indeterminiertheit und Zufall gibt, und allein schon die Frage, ob sich diese Begriffe überhaupt definieren lassen, ist ein alter philosophischer Streit, zu dem wir hier nichts beitragen wollen – die Wahrscheinlichkeitstheorie braucht diese Begriffe nicht, und allein, dass Laplace sowohl Mitbegründer der Wahrscheinlichkeitstheorie als auch mit seinem *Laplaceschen Dämon* der Vertreter des Determinismus schlechthin war, ist ein Indiz dafür.

Der Streit ist für die mathematische Behandlung der Wahrscheinlichkeit deshalb belanglos, weil sie Modelle zur Verfügung stellt, die von nur einem einzigen völlig unstrittigen Aspekt des Wahrscheinlichkeitsbegriffs ausgehen. Jede philosophische Auffassung ist dann auf die logischen Konsequenzen der Wahrscheinlichkeitstheorie verpflichtet.

Trotzdem wollen wir noch kurz die weiteren außermathematischen Aspekte des Begriffs ansprechen. Wenn man wie de Finetti die objektive Existenz von Wahrscheinlichkeit bestreitet, kann man sie als sozusagen psychologisches Phänomen verstehen. Auch falls alle vergangenen und zukünftigen Tatsachen im Prinzip determiniert sind und von einer Intelligenz, die so umfassend ist wie der schon erwähnte Laplacesche Dämon, berechnet werden können, steht außer Frage, dass die menschliche Intelligenz bestenfalls Teilinformationen erkennen und verarbeiten kann. Man muss das Feld dann freilich nicht der Psychologie überlassen, sondern kann Kriterien formulieren, wie mit Teilinformationen *rational* umzugehen ist – wobei rational hier in einem sehr strengen Sinn gemeint ist, nämlich als Vermeidung logischer Widersprüche.

In de Finettis Konzeption wird der rationale Umgang mit unsicheren Informationen durch die Analyse von Wetten auf Ereignisse wie etwa Ergebnisse von Fußballspielen operationalisiert, und er zeigt, dass allein die Vermeidung von Wettsystemen, die zu einem *sicheren* Verlust führen, Bedingungen impliziert, die stringent genug für die mathematische Theorie sind. Wir wollen dies nicht im Detail sondern nur an einem Beispiel ausführen. Bei de Finetti bedeutet, einem Ereignis die Wahrscheinlichkeit von $1/4$ zuzuordnen, die Bereitschaft, einen Wetteinsatz von 0,25 € zu leisten, wenn bei Eintritt des Ereignisses 1 € ausgezahlt wird, und außerdem gegen jeden Einsatz von mehr als 0,25 € bei Eintritt selbst 1 € auszuzahlen. Jemandem der dem gegenteiligen Ereignis eine andere Wahrscheinlichkeit zuordnet als $3/4$ – also zum Beispiel $1/2$ – kann man dann ein System von Wetten vorlegen, bei dem er mit Sicherheit Geld verliert: Er müsste je 1 € auf A und das gegenteilige Ereignis $B = A^c$ setzen und für (etwas mehr) als 2 € Einsatz auf A und 3 € auf B (sozusagen als Bank) Wetten akzeptieren. Im Fall von A bekäme er 4 € ausgezahlt und müsste selbst 8 € zahlen (mit einem Netto-Verlust von 1 €) und im Fall von B bekäme er 2 € und müsste 6 € zahlen (was ebenfalls zu einem Netto-Verlust von 1 € führt).

Diesen Spezialfall der Additivität von Wahrscheinlichkeiten werden wir im 1. Kapitel wiedersehen und im 3. Kapitel in der Linearität von Erwartungswerten. Eine ähnlich Form der Rationalität finden wir auch im 10. Kapitel bei einer finanzmathematischen Anwendung: Der Preis einer „zufälligen" Auszahlungsfunktion wird dadurch bestimmt, wie man an diese Funktion am billigsten mit einer sicheren nachbilden kann.

Der subjektivistische Aspekt, der bei de Finetti im Vordergrund steht, spielt auch im allgemeinen Verständnis von Wahrscheinlichkeit eine wichtige Rolle – selbst wenn man nicht gleich Haus und Hof verwetten will, wenn man eine Aussage über zu erwartende Ereignisse macht. Dennoch ist man seiner Aussage in einem gewissen Sinn verpflichtet (dem Deutschen fehlt leider ein so schöner Begriff wie *commitment*) – ein Wetterexperte, der eben im Fernsehen eine Regenwahrscheinlichkeit von 1 % verkündet hat (ohne übrigens zu sagen, was genau er damit meint), würde sich mit Schirm und Schal sehr suspekt machen.

Daneben gibt es einen „frequentistischen" Aspekt, der darauf beruht, dass sich viele Situationen, in denen von Wahrscheinlichkeit die Rede ist, durch ihre „Wiederholbarkeit" auszeichnen. Es wird oft als Erfahrungstatsache bezeichnet, dass sich die relativen Häufigkeiten von oft ausgeführten Experimenten einem Grenzwert annähern. Richard von Mises hat versucht, den Begriff Wahrscheinlichkeit als diesen Grenzwert zu *definieren*. Und auch wenn dieser Versuch furios gescheitert ist – sowohl auf der philosophischen Seite, weil bei ihm der physikalische Sinn von Ereignis, das insbesondere zu einem bestimmten Zeitpunkt stattzufinden hat, verloren geht, als auch auf der mathematischen, weil seine Definition von zufälligen oder in seiner Terminologie *regellosen* Folgen als haltlos nachgewiesen wurde – spielt dieser Aspekt sowohl in der Alltagssprache als auch im praktischen Umgang etwa bei Versicherungen eine wichtige Rolle. Wir werden ihn in den Grenzwertsätzen der Wahrscheinlichkeitstheorie und insbesondere im Gesetz der großen Zahlen wiederfinden, allerdings nicht im Zusammenhang mit der ganz verschiedenen Definition sondern als eine der beeindruckenden Konsequenzen.

Auch die oben angedeuteten *propensities* von Popper treffen wir in einer gewissen Form im Zentralen Grenzwertsatz im 5. Kapitel wieder: viele kleine unabhängige Einflüsse führen zu einer spezifischen Verteilung von Wahrscheinlichkeiten.

Die mathematisch-axiomatische Theorie der Wahrscheinlichkeit wurde von Andrei N. Kolmogorov begründet: Sie vermeidet alle inhaltlichen und philosophischen Probleme, indem sie nicht definiert, was Wahrscheinlichkeit ist, sondern indem sie die Regeln beschreibt, denen der Begriff genügt.

Kolmogorovs Terminologie war nicht neu sondern zum großen Teil der bereits existierenden Maßtheorie entlehnt. Man betrachtet eine Familie $\mathcal{A}$ von Teilmengen der Menge der möglichen Konstellationen $\omega \in \Omega$ und verlangt von einer *Wahrscheinlichkeitsverteilung* $P : \mathcal{A} \to [0, 1]$ die als *σ-Additivität* bezeichnete Bedingung

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

für alle sich paarweise ausschließenden $A_n \in \mathcal{A}$ sowie $P(\Omega) = 1$. Für $A \in \mathcal{A}$ heißt dann $P(A)$ Wahrscheinlichkeit von A.

Der Begriff *Wahrscheinlichkeit eines Ereignisses* ist also relativ zu der Abbildung P und deshalb nur innerhalb des Modells sinnvoll. Für die Frage nach der *Bedeutung* von Wahrscheinlichkeit mag dieser axiomatische Zugang auf den ersten Blick enttäuschend sein. Auf den zweiten jedoch erahnt man vielleicht auch den Nutzen für die philosophische Frage: Bevor nicht ein Modell, also die Abbildung P, spezifiziert ist, hat der Begriff Wahrscheinlichkeit gar keine feststehende Bedeutung, und aller Streit rührt von unterschiedlichen *impliziten* Annahmen her. Die Angabe von P erfordert, diese Annahmen zu explizieren, und falls Einigkeit über das Modell herrscht, lässt sich der Streit durch Leibniz' berühmte Aufforderung entscheiden: *calculemus!*

Bei allen inhaltlichen Interpretationen ist mindestens die endliche Additivität

$$P(A_1 \cup \cdots \cup A_n) = P(A_1) + \cdots + P(A_n)$$

für disjunkte Ereignisse unstrittig, und obgleich manche Teile der Wahrscheinlichkeitstheorie auch mit dieser schwächeren Eigenschaft auskämen, erlaubt erst die Bedingung für abzählbare Disjunktionen eine elegante mathematische Theorie.

Wir geben an dieser Stelle den heiklen Punkt in Hinblick auf die Anwendungen und Interpretationen zu – mit der Bitte an den Leser, dieses Problem dann wieder zu vergessen. In der Theorie wird stets verlangt, dass der Definitionsbereich $\mathcal{A}$ von P *schnittstabil* ist, das heißt aus $A \in \mathcal{A}$ und $B \in \mathcal{A}$ folgt $A \cap B \in \mathcal{A}$. Während man die Additivität von P mit Rationalitätsforderungen wie zum Beispiel bei de Finetti begründen kann, ist dies für die Schnittstabilität keineswegs so klar. Man kann wohlbegründete Vermutungen über die Erfolgsaussichten des FC A am nächsten Wochenende als auch über die der Borussia B haben und die vielleicht sogar mit Wahrscheinlichkeiten quantifizieren – und sich mit gutem Grund eines Urteils darüber enthalten, dass beide Mannschaften gewinnen, zum Beispiel, weil man nicht weiß, ob sie gegeneinander spielen. Und selbst wenn man das weiß, ist überhaupt nicht evident oder gar durch Rationalitätsforderungen festgelegt, wie das gemeinsame Eintreten zweier Ereignisse zu bewerten ist.

Die Modelle der Wahrscheinlichkeitstheorie sind für jemanden, der nicht einmal weiß, ob A gegen B spielt, also nicht gemacht. Wem dieses Beispiel zu banal ist, denke an einen Arzt, der sich durchaus über zwei mögliche Nebenwirkungen seiner Behandlung im Klaren ist aber nicht darüber, wie die beiden genau zusammenhängen.

Wir werden also stets voraussetzen, dass jemand, der sich zutraut, zwei Ereignisse A und B vernünftig zu bewerten, auch über deren Konjunktion ein rationales Urteil abgeben kann. Und auch wenn wir den Leser eben gebeten haben, diesen heiklen Punkt wieder zu vergessen – man darf durchaus im Gedächtnis behalten, dass die Modellierung ungewisser Situationen eine anspruchsvolle Aufgabe ist.

Soviel beziehungsweise so wenig zur Philosophie der Wahrscheinlichkeit, und nun zur Motivation für dieses Buch.

Bis vor gar nicht so langer Zeit wurden wahrscheinlichkeitstheoretische Modelle vor allem in der Statistik benutzt, um zu entscheiden, welches einer Klasse von Modellen am wenigsten unplausibel ist (die holprige Formulierung ist Poppers *Logik der Forschung* geschuldet, nach der sich eine Theorie oder ein Modell nicht verifizieren sondern bloß als unpassend herausstellen lassen). Die dafür benötigte „klassische“ Wahrscheinlichkeitstheorie ließ sich inklusive der Grundlagen der Maßtheorie in einer an deutschen Universitäten üblichen zweisemestrigen Vorlesung darstellen, so dass womöglich sogar noch etwas Zeit für statistische Anwendungen oder Ausblicke auf die „moderne“ Theorie stochastischer Prozesse blieb. Deren Hauptergebnis (nämlich die Itô-Formel in Kapitel 9), das in seiner Bedeutung mit dem Hauptsatz der Differential- und Integralrechnung der Analysis vergleichbar ist und eine herausragende Rolle bei Anwendungen zum Beispiel in der *Finanzstochastik* spielt, war dann selbst in einfachen Versionen weit jenseits des behandelten Stoffs und blieb als Thema einer spezialisierten Vorlesung über stochastische Prozesse und somit Studierenden mit diesem speziellen Interesse vorbehalten.

Dieses Buch ist ein Versuch, die gesamte für dieses Hauptergebnis benötigte Wahrscheinlichkeitstheorie in einem Umfang darzustellen, der tatsächlich in zwei Semestern zu bewältigen ist, und ohne dabei auf zentrale Ergebnisse der klassischen Theorie zu verzichten. Allerdings ist der vorliegende Text ein *Buch* und kein *Vorlesungsmanuskript*, letzteres würde deutlich mehr Redundanz enthalten und sich auch stilistisch ziemlich unterscheiden.

Die häufig benutzten Mittel, eine elaborierte Theorie kurz darzustellen – nämlich entweder die Grundlagen bloß anzudeuten und für die Details auf die Literatur zu verweisen oder sie gar als Übungsaufgaben getarnt dem Leser zu überlassen – sind in vielerlei Hinsicht unbefriedigend, und man wird feststellen, dass dieses Buch (an manchen Stellen vielleicht sogar etwas übertrieben) eigenständig (oder *self-contained*) ist: Alles was nicht mit Sicherheit im ersten Jahr eines Mathematikstudiums gelehrt wird, findet sich hier mit Beweis. Dadurch werden Überraschungen vermieden wie man sie gelegentlich in dicken Lehrbüchern findet, die – häufig im Zusammenhang mit Anwendungen des dargestellten Materials – in lockerem Ton einige Ergebnisse über das xyz herbei zaubern, die man ja in jedem Buch über die abc-Theorie nachlesen kann.

Was hier im Wesentlichen benötigt wird, sind der Umgang mit Reihen (mit positiven Summanden, was alle Konvergenzfragen auf den Unterschied zwischen beschränkten und unbeschränkten Partialsummenfolgen reduziert), Stetigkeit und partielle Differenzierbarkeit reeller Funktionen, die Hauptachsentransformation der linearen Algebra und die Terminologie metrischer Räume (die findet man zwar im Anhang, der aber nicht als Einführung für jemanden zu verstehen ist, der davon noch nie gehört hat). Vor allem aber ein souveräner Umgang mit den Grundbegriffen der Mengenlehre und insbesondere der Urbildabbildung (einem Leser, der sich seiner selbst vergewissern will, sei die allererste Übungsaufgabe empfohlen) sowie der Wunsch und die Bereitschaft, alle Aussagen im Text zu verifizieren: Ich habe mich bemüht, den Leser dabei nicht allein zu lassen, zum Beispiel, indem ich auf die so beliebte Wendung wie „es ist leicht einzusehen, dass …“ verzichtet und die eingesparte halbe Zeile für einen Hinweis auf das benötigte Argument benutzt habe. Auch wenn Irrtümer natürlich nicht ausgeschlossen sind, halte ich die manchmal nicht ausgeführten Details wie die Verifikation der Voraussetzungen bei der Anwendung eines Satzes in fast allen Fällen für sehr leicht.

Um bei der Behandlung des Themas weder auf tiefer liegende Ergebnisse noch einige Anwendungen zu verzichten, wurden oft Zugänge und Beweise gewählt, die vielleicht nur in einigen Details neu sind aber in ihrer Gesamtheit zu einem deutlichen Unterschied zu der „klassischen Darstellung“ führen.

Zwei stilistische Mittel zur stringenten Darstellung, die dieses Buch deutlich von einem Vorlesungsmanuskript unterscheiden, sind einerseits der Verzicht auf eine starke Untergliederung. Abgesehen von den nummerierten Sätzen, die sozusagen das logische Gerüst der Theorie bilden, finden sich Beispiele und Definitionen im laufenden Text, wobei letztere fett gedruckt und im Index mit Verweis auf die Seitenzahl gesammelt sind. Andererseits erweist sich die etwas ungewöhnlich Benutzung von $a \equiv b$ als definitorische im Unterschied zur behaupteten Gleichheit als sehr nützlich, auch um bloß vorübergehende Bezeichnungen einzuführen, was Wendungen wie „wobei wir a mit b bezeichnen“ erspart.

Die Kürze der Darstellung hat neben ihren Vorzügen – insbesondere ist so viel leichter ein Überblick über die vielen theoretischen Zusammenhänge zu erreichen als bei einer breiteren Darstellung – natürlich ihren Preis: Einerseits verlangt sie vom

Leser ein hohes Maß an Konzentration, und andererseits kommen manche wichtige Themen der Wahrscheinlichkeitstheorie nur am Rande vor. Als größtes Versäumnis betrachte ich dabei eine ziemliche Vernachlässigung zeitdiskreter stochastischer Prozesse und zweitens das Fehlen von Markov-Prozessen, die nur in der Form von Prozessen mit unabhängigen Zuwächsen vorkommen.

Bevor es zur Sache geht, möchte ich den Herren T. Kalmes, H. Luschgy und W. Sendler für eine Reihe kritischer Anmerkungen und hilfreicher Diskussionen danken sowie Ch. Becker und N. Kenessey, die sowohl Teile des Manuskripts gelesen als auch als Hörer meiner Vorlesungen keine Unsauberkeiten haben durchgehen lassen, insbesondere Herr Becker hat seinen Finger in jede wunde Stelle gelegt. Schließlich bedanke ich mich herzlich bei Lisa Schmitt, die das Manuskript in LaTeX umgesetzt hat.

Liège, Mai 2008 Jochen Wengenroth

Inhaltsverzeichnis

Kapitel 1
Ereignisse und Modelle

Wir führen in diesem Kapitel das Vokabular der Wahrscheinlichkeitstheorie ein, das aus suggestiven Sprechweisen für mengentheoretische Zusammenhänge besteht.

Wir betrachten stets eine Menge Ω von **Konstellationen** $\omega \in \Omega$, die zum Beispiel die möglichen Resultate eines Experiments beschreiben. Teilmengen von Ω nennen wir dann **Ereignisse** (von Ω). Ω selbst heißt das **sichere Ereignis**, und die leere Menge heißt **unmögliches Ereignis**. Zum Beispiel kann man mit der Menge $\Omega = \{1, 2, \ldots, 6\}$ die Ergebnisse des Wurfs mit einem Würfel beschreiben, und $A = \{1, 3, 5\}$ ist das Ereignis, dass die gewürfelte Zahl ungerade ist.

Eine Menge $\mathcal{A}$ von Ereignissen von Ω – also eine Teilmenge der **Potenzmenge** $\mathcal{P}(\Omega)$ – heißt **σ-Algebra** (über Ω), falls das sichere Ereignis Element von $\mathcal{A}$ ist, mit jedem Ereignis $A \in \mathcal{A}$ auch das gegenteilige Ereignis $A^c = \Omega \setminus A$ Element von $\mathcal{A}$ ist, und für jede Folge $(A_n)_{n\in\mathbb{N}}$ von Ereignissen $A_n \in \mathcal{A}$ auch die Vereinigung $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$ ist.

Die Ereignisse $A \in \mathcal{A}$ nennen wir **$\mathcal{A}$-zulässige** Ereignisse oder auch **$\mathcal{A}$-messbar** und das Paar $(\Omega, \mathcal{A})$ einen **Messraum**.

Oft ist es suggestiv, eine σ-Algebra $\mathcal{A}$ als ein System von Informationen anzusehen in dem Sinn, dass man von den $\mathcal{A}$-zulässigen Ereignissen weiß, *ob sie eingetreten sind.* Dann kann man die drei Axiome als Rationalitätsforderungen verstehen (wobei die Bedingung für abzählbare Disjunktionen statt für endliche ein Preis für die elegante mathematische Theorie ist). In diesem Sinn beschreiben die minimale σ-Algebra $\{\varnothing, \Omega\}$ und die maximale σ-Algebra $\mathcal{P}(\Omega)$ vollständige Ignoranz beziehungsweise Allwissenheit.

Für eine Menge $\{\mathcal{A}_\alpha : \alpha \in I\}$ von σ-Algebren über Ω ist

$$\bigwedge\{\mathcal{A}_\alpha : \alpha \in I\} \equiv \bigwedge_{\alpha\in I} \mathcal{A}_\alpha \equiv \bigcap_{\alpha\in I} \mathcal{A}_\alpha = \{A \subseteq \Omega : A \in \mathcal{A}_\alpha \text{ für alle } \alpha \in I\}$$

wieder eine σ-Algebra und zwar die (bezüglich der Inklusion) größte, die in allen $\mathcal{A}_\alpha$ enthalten ist (interpretiert man $\mathcal{A}_\alpha$ als Informationssysteme verschiedener Personen, so wäre dieses Minimum $\bigwedge_{\alpha\in I} \mathcal{A}_\alpha$ der „common sense“). Ist $\mathcal{E}$ irgendeine Menge von Ereignissen von Ω, so ist

$$\sigma(\mathcal{E}) \equiv \sigma_\Omega(\mathcal{E}) \equiv \bigwedge\{\mathcal{A} : \mathcal{A}\ \sigma\text{-Algebra mit } \mathcal{E} \subseteq \mathcal{A}\}$$

die minimale σ-Algebra über Ω, die $\mathcal{E}$ umfasst. $\mathcal{E}$ heißt dann ein **Erzeuger** von $\sigma(\mathcal{E})$. Diese **erzeugte** σ-Algebra ist also durch $\mathcal{E}$ eindeutig bestimmt, aber andererseits gibt es in der Regel sehr viele verschiedene Erzeuger.

$\sigma(\mathcal{E})$ lässt sich nur in sehr speziellen Situationen konkret beschreiben. Zum Beispiel ist $\sigma(\{A\}) = \{\emptyset, A, A^c, \Omega\}$ für jedes Ereignis $A \subseteq \Omega$. Das rechte Mengensystem ist nämlich eine σ-Algebra, die $\{A\}$ umfasst, und andererseits enthält jede σ-Algebra mit A auch A^c und sowieso $\emptyset$ und Ω. Ist $\Omega = \{1, \ldots, 6\}$ die Menge der möglichen Resultate eines Würfelwurfs und $A = \{2, 4, 6\}$, so beschreibt $\sigma(\{A\})$ die Information, ob die Augenzahl gerade ist.

Mit dem gleichen Argument wie eben kann man $\sigma(\mathcal{E})$ für eine (höchstens) abzählbare **Zerlegung** $\mathcal{E} = \{A_n : n \in \mathbb{N}\}$ von Ω beschreiben, das heißt falls A_n paarweise disjunkt mit $\Omega = \bigcup_{n\in\mathbb{N}} A_n$ sind: Mit der Bezeichnung $A(J) \equiv \bigcup_{n\in J} A_n$ für $J \subseteq \mathbb{N}$ gilt dann $\sigma(\mathcal{E}) = \{A(J) : J \subseteq \mathbb{N}\}$. Wegen $\Omega = A(\mathbb{N})$, $A(J)^c = A(J^c)$ und $\bigcup_{n\in\mathbb{N}} A(J_n) = A(\bigcup_{n\in\mathbb{N}} J_n)$ ist nämlich das rechte Mengensystem eine σ-Algebra, die alle $A_n = A(\{n\})$ enthält, und jede σ-Algebra enthält mit allen A_n auch die abzählbaren Vereinigungen $A(J)$.

Ist $\mathcal{E} = \{B_1, \ldots, B_n\} \subseteq \mathcal{P}(\Omega)$ endlich und definieren wir für $s \in \{0,1\}^n$ die Mengen $A_s \equiv \bigcap_{s_j=1} B_j \cap \bigcap_{s_j=0} B_j^c$, so ist $\tilde{\mathcal{E}} \equiv \{A_s : s \in \{0,1\}^n\}$ eine Zerlegung von Ω, weil jedes $\omega \in \Omega$ Element genau derjenigen Menge A_s mit $s_j = 1$ falls $\omega \in B_j$ und $s_j = 0$ falls $\omega \notin B_j$ ist. Außerdem gilt $\sigma(\mathcal{E}) = \sigma(\tilde{\mathcal{E}})$, weil B_j die Vereinigung aller A_s mit $s_j = 1$ ist. Also erhalten wir die Darstellung

$$\sigma(\mathcal{E}) = \Big\{ \bigcup_{s\in J} A_s : J \subseteq \{0,1\}^n \Big\}.$$

Selbst für abzählbares $\mathcal{E}$ kann man die erzeugte σ-Algebra im Allgemeinen nur durch einen „transfiniten“ Prozess konstruktiv beschreiben. Wir werden aber eine solche „Konstruktion“ nie benutzen, sondern kommen immer mit der abstrakten aber einfachen Definition als Minimum aller σ-Algebren, die $\mathcal{E}$ enthalten, aus.

Mit Hilfe des Minimums können wir nun auch das Maximum

$$\bigvee_{\alpha\in I} \mathcal{A}_\alpha \equiv \bigvee \{\mathcal{A}_\alpha : \alpha \in I\} \equiv \sigma\Big(\bigcup_{\alpha\in I} \mathcal{A}_\alpha \Big)$$

als die kleinste σ-Algebra, die alle $\mathcal{A}_\alpha$ umfasst, definieren. Schon das Beispiel $\mathcal{A}_n \equiv \sigma_{\mathbb{N}}(\{n\})$ zeigt, dass $\bigcup_{\alpha\in I} \mathcal{A}_\alpha$ selbst im Allgemeinen keine σ-Algebra ist. In der Interpretation als rationale Informationssysteme heißt das, dass die Vereinigung (also zum Beispiel eine Anhäufung von Internetseiten) nicht rational zu sein braucht. Für $I = \{1, \ldots, n\}$ schreiben wir $\bigvee_{\alpha\in I} \mathcal{A}_\alpha = \mathcal{A}_1 \vee \cdots \vee \mathcal{A}_n$ und $\bigwedge_{\alpha\in I} \mathcal{A}_\alpha = \mathcal{A}_1 \wedge \cdots \wedge \mathcal{A}_n$.

Ist Ω mit einer Metrik d versehen (die für uns wichtigen Definitionen und Ergebnisse über metrische Räume findet man im Anhang), so heißt

$$\mathcal{B}(\Omega, d) \equiv \sigma(\{A : A \text{ offen}\})$$

Borel-σ-Algebra über Ω.

Sind speziell $\Omega = \mathbb{R}^n$ und d die euklidische Metrik, so schreiben wir die Borel-σ-Algebra als $\mathbb{B}_n$ und im Fall $n = 1$ als $\mathbb{B} \equiv \mathbb{B}_1$. Die Borel-$\sigma$-Algebra $\mathbb{B}_n$ enthält insbesondere alle Mengen, die sich durch „abzählbare Prozesse" mittels Komplement- und Durchschnittbildung aus offenen Mengen beschreiben lassen, und es ist mühsamer, Ereignisse zu finden, die nicht $\mathbb{B}_n$-zulässig sind, als in den meisten Fällen die Zulässigkeit konkreter Mengen zu zeigen.

Für eine σ-Algebra $\mathcal{A}$ heißt eine Abbildung $\mu : \mathcal{A} \to [0, +\infty]$ ein **Maß** auf $\mathcal{A}$ oder $(\Omega, \mathcal{A})$ und $(\Omega, \mathcal{A}, \mu)$ heißt dann **Maßraum**, falls

$$\mu(\varnothing) = 0 \text{ und } \mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n=1}^{\infty} \mu(A_n)$$

für alle Folgen $(A_n)_{n\in\mathbb{N}}$ *paarweise disjunkter* Ereignisse $A_n \in \mathcal{A}$ (weil alle Summanden positiv sind, steht die Konvergenz der Reihe im Intervall $[0, +\infty]$ nicht in Frage). μ heißt ein **Wahrscheinlichkeitsmaß** oder eine **Verteilung** und $(\Omega, \mathcal{A}, \mu)$ heißt dann **Wahrscheinlichkeitsraum** oder **Modell**, falls $\mu(\Omega) = 1$ gilt. Üblicherweise werden wir ein Wahrscheinlichkeitsmaß mit dem Symbol P bezeichnen. Die Normiertheit impliziert zusammen mit der **σ-Additivität** angewendet auf $A_1 = \Omega$ und $A_n = \varnothing$ für $n \geq 2$ übrigens schon das erste Axiom $P(\varnothing) = 0$.

Wir sehen hier die zweite wichtige Funktion von σ-Algebren für die Theorie, nämlich als Definitionsbereiche von Maßen. Maße μ, die nur die Werte 0 und 1 annehmen, spezifizieren für jedes $\mathcal{A}$-zulässige Ereignis A, *dass* es eingetreten ist, falls $\mu(A) = 1$, beziehungsweise dass es nicht eingetreten ist, falls $\mu(A) = 0$. Das **Dirac-Maß**

$$\delta_a(A) \equiv \begin{cases} 1, & a \in A \\ 0, & a \notin A \end{cases}$$

in einem Punkt $a \in \Omega$ ist das typische Beispiel für diese Situation. Dies ist also der Grenzfall eines allgemeinen Wahrscheinlichkeitsmaßes, das jedem $A \in \mathcal{A}$ die „Eintrittssicherheit" zuordnet. Endliche Additivität wäre dann wiederum eine Rationalitätsforderung und die σ-Additivität ist auch hier der Tribut an die Eleganz der Theorie.

Wir interessieren uns vornehmlich für Verteilungen, allgemeine Maße sind aber oft ein wichtiges Hilfsmittel und als Instrument zur Bestimmung von Längen, Flächen und Volumina auch in anderen Bereichen der Mathematik von zentraler Bedeutung. Bevor wir erste Beispiele für Verteilungen angeben, beweisen wir die grundlegenden Eigenschaften von Maßen. Für eine Folge von Ereignissen A_n schreiben wir dabei $A_n \uparrow A$ und $A_n \downarrow A$ falls $A_n \subseteq A_{n+1}$ und $A = \bigcup_{n\in\mathbb{N}} A_n$ beziehungsweise $A_{n+1} \subseteq A_n$ und $A = \bigcap_{n\in\mathbb{N}} A_n$.

Satz 1.1 (Einmaleins der Maßtheorie)
Für einen Maßraum $(\Omega, \mathcal{A}, \mu)$ und $A, B, A_n \in \mathcal{A}$ gelten folgende Aussagen:

1. *Sind $A_1, \ldots, A_n$ paarweise disjunkt, so gilt* $\mu\Big(\bigcup\limits_{k=1}^{n} A_k\Big) = \sum\limits_{k=1}^{n} \mu(A_k)$.

2. *$\mu(A \cup B) + \mu(A \cap B) = \mu(A) + \mu(B)$.*
3. *Für $A \subseteq B$ ist $\mu(B \backslash A) + \mu(A) = \mu(B)$.*
4. *Für $A \subseteq B$ ist $\mu(A) \leq \mu(B)$.*
5. *$\mu\Big(\bigcup_{n \in \mathbb{N}} A_n\Big) \leq \sum_{n=1}^{\infty} \mu(A_n)$.*
6. *$A_n \uparrow A$ impliziert $\mu(A_n) \to \mu(A)$.*
7. *$A_n \downarrow A$ und $\mu(A_1) < \infty$ implizieren $\mu(A_n) \to \mu(A)$.*

Die Eigenschaften in 1. und 4.–7. heißen **Additivität**, **Monotonie**, **Sub-σ-Additivität** und **Stetigkeit** von unten beziehungsweise oben.

Beweis. 1. folgt aus der σ-Additivität mit $A_k = \emptyset$ für $k > n$. Diese endliche Additivität impliziert die dritte Aussage, da A und $B \setminus A$ disjunkt mit Vereinigung B sind. 4. folgt aus 3. und der Positivität, und wegen $A \cup B = A \cup (B \backslash (A \cap B))$ liefert die Anwendung von 3. auf $A \cap B \subseteq B$ die zweite Aussage.

6. folgt aus der Monotonie, falls $\mu(A_n) = \infty$ für ein $n \in \mathbb{N}$ gilt. Andernfalls sind $B_n \equiv A_n \setminus A_{n-1}$ mit $A_0 \equiv \emptyset$ paarweise disjunkt, und mit 3. folgt

$$\mu(A) = \mu\Big(\bigcup_{n \in \mathbb{N}} B_n\Big) = \sum_{n=1}^{\infty} \mu(B_n) = \lim_{N \to \infty} \sum_{n=1}^{N} \mu(A_n) - \mu(A_{n-1}) = \lim_{N \to \infty} \mu(A_N).$$

2. impliziert induktiv $\mu(\bigcup_{k=1}^{n} A_k) \leq \sum_{k=1}^{n} \mu(A_k)$ und damit folgt 5. aus der Stetigkeit von unten. Wegen $A_1 \backslash A_n \uparrow A_1 \backslash A$ folgt 7. aus 6. und 3. □

Um Verteilungen auf der Potenzmenge $\mathcal{A} = \mathcal{P}(\Omega)$ einer n-elementigen Menge Ω zu definieren, muss man nicht die 2^n Wahrscheinlichkeiten $P(A)$ für $A \subseteq \Omega$ angeben. Wegen der Additivität reicht die Definition von $P(\{\omega\})$ für jedes $\omega \in \Omega$. Mit der Bezeichnung $\sum_{\omega \in A} f(\omega) \equiv \sup\{\sum_{\omega \in E} f(\omega) : E \subseteq A \text{ endlich}\}$ für $f : A \to [0, \infty]$ gilt allgemeiner:

Satz 1.2 (Diskrete Modelle)
Seien $(\Omega, \mathcal{A})$ ein Messraum und $f : \Omega \to [0, +\infty]$ eine Abbildung. Dann ist durch $\mu(A) \equiv \sum_{\omega \in A} f(\omega)$ ein Maß auf $\mathcal{A}$ definiert. Ist Ω höchstens abzählbar, so ist jedes Maß auf $\mathcal{P}(\Omega)$ von dieser Form.

Beweis. $\mu(\emptyset) = 0$ folgt aus der Definition der leeren Summe als 0, und die σ-Additivität folgt aus der unbedingten Konvergenz von Reihen mit positiven Summanden. Für gegebenes Maß μ auf $\mathcal{P}(\Omega)$ liefert bei abzählbarem Ω wegen der σ-Additivität $f(\omega) \equiv \mu(\{\omega\})$ die gewünschte Darstellung. □

In der Situation von Satz 1.2 heißt f eine **Zähldichte** von μ, und μ heißt **diskret**. Das zu $f = 1$ gehörige Maß heißt **Zählmaß** auf Ω. Das Zählmaß auf $\mathbb{N}$ und die Ereignisse $A_n = \{n, n+1, n+2, \ldots\} \downarrow \varnothing$ zeigen, dass Satz 1.1.7 ohne die Voraussetzung $\mu(A_1) < \infty$ im Allgemeinen falsch ist.

Diskrete *Verteilungen* können wir nun durch Zähldichten f mit $\sum_{\omega \in \Omega} f(\omega) = 1$ definieren. Für endliches Ω heißt das Maß mit Zähldichte $f(\omega) \equiv 1/|\Omega|$ **Laplace-Verteilung** auf Ω. Sie ist das paradigmatische Modell für Situationen wie dem Würfeln, in denen alle Konstellationen als gleichmöglich angesehen werden. Speziell auf $\Omega = \{0, 1\}$ heißt die Laplace-Verteilung auch **Bernoulli-Verteilung**.

Die Verteilung auf $\Omega = \{0, \ldots, n\}$ mit Zähldichte $f(k) \equiv \binom{n}{k} p^k (1-p)^{n-k}$ für festes $p \in [0, 1]$ wird mit $B(n, p)$ bezeichnet und heißt **Binomialverteilung** mit Parametern n und p. Wie wir noch begründen werden, ist sie ein angemessenes Modell für die „Verteilung der Treffer mit Wahrscheinlichkeit p in n unabhängigen Experimenten".

Die Verteilung auf $\Omega = \mathbb{N}_0$ mit Zähldichte $f(n) \equiv e^{-\lambda} \lambda^n / n!$ heißt **Poisson-Verteilung** mit Parameter $\lambda \geq 0$ und wird mit $Po(\lambda)$ bezeichnet. Wir werden später sehen, dass sie ein Modell für die „Verteilung seltener Ereignisse" liefert.

Als Modell für „Wartezeiten auf Ereignisse mit Wahrscheinlichkeit p" dient die **geometrische Verteilung** $Ge(p)$ auf $\Omega = \mathbb{N}$ mit Zähldichte $f(n) \equiv (1-p)^{n-1} p$.

Für nicht abzählbares Ω liefert die Definition

$$P(A) \equiv \begin{cases} 0, & \text{falls } A \text{ abzählbar} \\ 1, & \text{sonst} \end{cases}$$

auf $\mathcal{A} \equiv \{A \subseteq \Omega : A \text{ oder } A^c \text{ abzählbar}\}$ eine (nicht diskrete) Verteilung, die gelegentlich als trennendes Beispiel dient ($\mathcal{A}$ ist eine σ-Algebra, weil die abzählbare Vereinigung abzählbarer Mengen wieder abzählbar ist, und P ist eine Verteilung, weil für paarweise disjunkte Ereignisse in $\mathcal{A}$ höchstens eines abzählbares Komplement besitzt).

Schon mit diskreten Verteilungen lassen sich überraschende Beispiele untersuchen. Wir können die j-te Komponente einer Konstellation $\omega = (\omega_1, \ldots, \omega_n) \in \Omega \equiv \{1, \ldots, N\}^n$ im Fall $N = 365$ als Geburtstag des j-ten Kinds einer Klasse mit n Schülern oder im Fall $N = \binom{49}{6}$ als Ergebnis der j-ten von bislang n durchgeführten Lottoziehungen „6 aus 49" interpretieren. Seien P die Laplace-Verteilung auf Ω und $A \equiv \{(\omega_1, \ldots, \omega_n) \in \Omega : \text{es gibt } i \neq j \text{ mit } \omega_i = \omega_j\}$. In der Interpretation als Geburtstage von Kindern einer Schulklasse beschreibt A das Ereignis, dass zwei Kinder am selben Tag Geburtstag feiern. Durch Induktion nach n erhalten wir, dass A^c genau $N(N-1)\cdots(N-n+1)$ Elemente hat, und wegen $|\Omega| = N^n$ folgt

$$P(A) = 1 - P(A^c) = 1 - \frac{N(N-1) - (N-n+1)}{N^n} = 1 - \prod_{k=0}^{n-1} \left(1 - \frac{k}{N}\right).$$

Die einfache Ungleichung $\exp(x) \geq 1 + x$ für $x \in \mathbb{R}$ (die zum Beispiel aus dem Mittelwertsatz folgt) liefert dann

$$P(A) \geq 1 - \exp\Big(-\sum_{k=0}^{n-1} \frac{k}{N}\Big) = 1 - \exp\big(-\frac{n(n-1)}{2N}\big).$$

Weil $P(A) \geq 1/2$ schon für $n(n-1) \geq 2N \log 2$ gilt (also $n \geq 23$ für $N = 365$ und $n \geq 4404$ für $N = \binom{49}{6}$), wird dieses Beispiel manchmal **Kollisionsparadoxon** genannt.

Eine nützliche Verallgemeinerung von Satz 1.1.2 ist:

Satz 1.3 (Siebformel)
Seien $(\Omega, \mathcal{A}, \mu)$ ein Maßraum und $A_1, \ldots, A_n \in \mathcal{A}$ mit $\mu(\bigcup_{j=1}^n A_j) < \infty$. Dann gilt

$$\mu\Big(\bigcup_{j=1}^n A_j\Big) = \sum (-1)^{|S|+1} \mu\Big(\bigcap_{j \in S} A_j\Big),$$

wobei über alle nicht-leeren Teilmengen $S \subseteq \{1, \ldots, n\}$ summiert wird.

In der Siebformel werden also alle $\mu(A_j)$ mit positivem Vorzeichen, alle $\mu(A_j \cap A_k)$ mit negativem Vorzeichen und allgemeiner die Maße aller Schnitte von k verschiedenen Mengen mit dem Vorzeichen $(-1)^{k+1}$ addiert. Weil nach Voraussetzung $\mu(A_j) < \infty$ für alle $j \in \{1, \ldots, n\}$ gilt, ist diese Summe tatsächlich definiert.

Beweis. Für $n = 1$ ist nichts zu beweisen und für $n \geq 2$ folgt aus Satz 1.1.2 und zweimalige Anwendung der Siebformel für $n - 1$ Ereignisse

$$\begin{aligned} \mu\Big(\bigcup_{j=1}^n A_j\Big) &= \mu(A_n) + \mu\Big(\bigcup_{j=1}^{n-1} A_j\Big) - \mu\Big(\bigcup_{j=1}^{n-1} A_j \cap A_n\Big) \\ &= \mu(A_n) + \sum_{\varnothing \neq S \subseteq \{1,\ldots,n-1\}} (-1)^{|S|+1} \mu\Big(\bigcap_{j \in S} A_j\Big) \\ &\quad - \sum_{\varnothing \neq T \subseteq \{1,\ldots,n-1\}} (-1)^{|T|+1} \mu\Big(\bigcap_{j \in T \cup \{n\}} A_j\Big) \end{aligned}$$

und dies stimmt mit der rechten Seite in der Siebformel für $A_1, \ldots, A_n$ überein. □

Eine klassische Anwendung der Siebformel ist das **Rencontre-Problem**. Wir betrachten die Laplace-Verteilung P auf der Menge der Permutationen $\Omega \equiv \{f : \{1, \ldots, n\} \to \{1, \ldots, n\} \text{ bijektiv}\}$ und suchen die Wahrscheinlichkeit der Menge der Permutationen mit mindestens einem Fixpunkt. Die Konstellationen lassen sich etwa als Ergebnisse von Verlosungen auf einer Weihnachtsfeier interpretieren, zu der jeder Gast ein Geschenk mitbringt. Gesucht ist dann die Wahrscheinlichkeit dafür, dass mindestens einem Gast das selbst mitgebrachte Geschenk zugelost wird. Für $A_j \equiv \{f \in \Omega : f(j) = j\}$ und ein k-elementiges $S \subseteq \{1, \dots, n\}$ liefert die Restriktion auf $\{1, \dots, n\} \setminus S$ eine bijektive Abbildung zwischen $\bigcap_{j \in S} A_j$ und der Menge der Permutationen auf S^c. Weil es $\binom{n}{k}$ Teilmengen von $\{1, \dots, n\}$ mit k Elementen gibt, folgt mit der Siebformel

$$P\Big(\bigcup_{j=1}^{n} A_j\Big) = \frac{1}{n!}\sum_{k=1}^{n}(-1)^{k+1}\binom{n}{k}(n-k)! = \sum_{k=1}^{n}(-1)^{k+1}\frac{1}{k!} = 1 - \sum_{k=0}^{n}(-1)^k/k!$$

Weil dies sehr schnell gegen $1 - e^{-1}$ konvergiert, hängt die gesuchte Wahrscheinlichkeit unerwarteter Weise „kaum" von n ab.

Während der Raum der Konstellationen in den bisherigen Beispielen der Situation angepasst war, erscheint die Wahl der Laplace-Verteilung allenfalls plausibel, wenn nicht gar willkürlich. Im Laplace-Modell werden nämlich sehr *spezifische* Annahmen über die Wahrscheinlichkeiten getroffen, die *nicht* – wie manchmal suggeriert wird – daraus abgeleitet werden können, dass man für die zu modellierende Situation keine spezifischen Informationen hat.

Wir werden nun sehen, dass eine Verteilung auf einer σ-Algebra $\mathcal{A}$ schon durch jeden **schnittstabilen** Erzeuger $\mathcal{E}$ (das heißt für alle $A, B \in \mathcal{E}$ gilt $A \cap B \in \mathcal{E}$) eindeutig bestimmt ist, dies setzt der Willkür immerhin eine Grenze.

Dafür benötigen wir ein Hilfsmittel, das wir immer wieder benutzen werden. Ein Mengensystem $\mathcal{D} \subseteq \mathcal{P}(\Omega)$ heißt **Dynkin-System** (über Ω), falls es das sichere Ereignis enthält, mit jedem Ereignis auch das gegenteilige und mit jeder Folge *paarweise disjunkter* Ereignisse auch die Vereinigung. Insbesondere ist also jede σ-Algebra ein Dynkin-System, und wegen

$$\bigcup_{n \in \mathbb{N}} A_n = \bigcup_{n \in \mathbb{N}} A_n \cap A_{n-1}^c \cap \cdots \cap A_1^c$$

sind schnittstabile Dynkin-Systeme schon σ-Algebren. Genau wie im Fall von σ-Algebren ist für $\mathcal{E} \subseteq \mathcal{P}(\Omega)$ durch

$$\delta(\mathcal{E}) \equiv \delta_\Omega(\mathcal{E}) \equiv \bigcap\{\mathcal{D} \text{ Dynkin-System mit } \mathcal{E} \subseteq \mathcal{D}\}$$

das minimale Dynkin-System, das $\mathcal{E}$ umfasst, definiert.

Satz 1.4 (Dynkin-Argument)
Für jedes schnittstabile Mengensystem $\mathcal{E} \subseteq \mathcal{P}(\Omega)$ gilt $\sigma(\mathcal{E}) = \delta(\mathcal{E})$.

Beweis. Weil jede σ-Algebra ein Dynkin-System ist, gilt $\delta(\mathcal{E}) \subseteq \sigma(\mathcal{E})$. Also müssen wir zeigen, dass $\delta(\mathcal{E})$ eine σ-Algebra ist, und wegen des oben gesagten reicht der Nachweis, dass $\delta(\mathcal{E})$ schnittstabil ist.

Für $B \in \delta(\mathcal{E})$ definieren wir $\mathcal{D}_B \equiv \{A \in \delta(\mathcal{E}) : A \cap B \in \delta(\mathcal{E})\}$. Dann ist jedes dieser Mengensysteme ein Dynkin-System, wobei die Stabilität bezüglich Komplementbildung wegen $B \cap A^c = (B^c \cup A)^c = (B^c \cup (A \cap B))^c$ aus der Disjunktheit von B^c und $A \cap B$ folgt, die beide $\delta(\mathcal{E})$ angehören.

Für $B \in \mathcal{E}$ gilt dann $\mathcal{E} \subseteq \mathcal{D}_B$, weil für $A \in \mathcal{E}$ sogar $A \cap B \in \mathcal{E}$ gilt. Also ist auch $\delta(E) \subseteq \mathcal{D}_B$, weil $\mathcal{D}_B$ ein Dynkin-System ist.

Für alle $A \in \delta(\mathcal{E})$ und $B \in \mathcal{E}$ haben wir also $B \cap A \in \delta(\mathcal{E})$ und damit $\mathcal{E} \subseteq \mathcal{D}_A$ gezeigt. Wieder weil $\mathcal{D}_A$ ein Dynkin-System ist, folgt damit $\delta(\mathcal{E}) \subseteq \mathcal{D}_A$ für jedes $A \in \delta(\mathcal{E})$, so dass $\delta(\mathcal{E})$ wie gewünscht schnittstabil ist. □

Für spätere Zwecke beweisen wir die oben erwähnte Eindeutigkeitsaussage für Verteilungen in etwas größerer Allgemeinheit. Wir nennen ein Maß $\mu : \mathcal{A} \to [0, +\infty]$ **σ-endlich** auf $\mathcal{E} \subseteq \mathcal{A}$ (und nur σ-endlich, falls $\mathcal{E} = \mathcal{A}$), falls es Ereignisse $E_n \in \mathcal{E}$ gibt mit $E_n \uparrow \Omega$ und $\mu(E_n) < +\infty$ für jedes $n \in \mathbb{N}$.

Endliche Maße μ, das heißt Maße mit $\mu(\Omega) < \infty$, sind also auf jedem Mengensystem σ-endlich, das eine Folge $E_n \uparrow \Omega$ enthält.

Das Zählmaß auf einer überabzählbaren Menge ist ein Beispiel für ein nicht σ-endliches Maß.

Satz 1.5 (Maßeindeutigkeit)
Seien μ und ν zwei Maße auf $\mathcal{A}$, die auf einem schnittstabilen Erzeuger $\mathcal{E}$ von $\mathcal{A}$ übereinstimmen und σ-endlich auf $\mathcal{E}$ sind. Dann gilt $\mu = \nu$.

Beweis. Sei $(E_n)_{n\in\mathbb{N}}$ eine Folge in $\mathcal{E}$ mit $\mu(E_n) = \nu(E_n) < \infty$ und $E_n \uparrow \Omega$. Für jedes $n \in \mathbb{N}$ ist dann $\mathcal{D}_n \equiv \{A \in \mathcal{A} : \mu(A \cap E_n) = \nu(A \cap E_n)\}$ ein Dynkin-System (die Komplementstabilität folgt aus $\mu(A^c \cap E_n) + \mu(A \cap E_n) = \mu(E_n) < \infty$). Weil μ und ν auf dem schnittstabilen Erzeuger $\mathcal{E}$ übereinstimmen, gilt $\mathcal{E} \subseteq \mathcal{D}_n$, und mit dem Dynkin-Argument folgt $\mathcal{A} = \sigma(\mathcal{E}) = \delta(\mathcal{E}) \subseteq \mathcal{D}_n$ für jedes $n \in \mathbb{N}$. Mit der Stetigkeit von unten erhalten wir für $A \in \mathcal{A}$

$$\mu(A) = \lim_{n\to\infty} \mu(A \cap E_n) = \lim_{n\to\infty} \nu(A \cap E_n) = \nu(A).$$ □

Das Beispiel $\nu = P$ vor dem Kollisionsparadoxon und $\mu = 2P$ zeigt, dass man selbst für endliche Maße auf die Voraussetzung, dass ν auf $\mathcal{E}$ σ-endlich ist, nicht verzichten kann (auf dem Erzeuger $\mathcal{E} \equiv \{A \subseteq \Omega : A \text{ abzählbar}\}$ sind ν und μ beide gleich 0). Sind allerdings ν und μ beide Wahrscheinlichkeitsmaße, so genügt die Übereinstimmung auf einem schnittstabilen Erzeuger $\mathcal{E}$, weil dann auch $\mathcal{E} \cup \{\Omega\}$ schnittstabil ist und ν und μ dort übereinstimmen.

Eine wichtige Anwendung von Satz 1.5 ist, dass Verteilungen P auf der Borel-σ-Algebra $\mathbb{B}$ über $\mathbb{R}$ durch ihre **Verteilungsfunktion** $F : \mathbb{R} \to \mathbb{R}$, $F(x) \equiv P((-\infty, x])$ eindeutig bestimmt sind: Weil sich jedes offene Intervall als

$$(a, b) = \bigcup_{n\in\mathbb{N}} (-\infty, b - 1/n] \setminus (-\infty, a]$$

darstellen lässt und jede offene Menge abzählbare Vereinigung von Intervallen (etwa allen enthaltenen Intervallen mit rationalen Endpunkten) ist, gilt $\{A \subseteq \mathbb{R} \text{ offen}\} \subseteq \sigma(\{(-\infty, x] : x \in \mathbb{R}\})$. Also ist die Menge $\{(-\infty, x] : x \in \mathbb{R}\}$ ein schnittstabiler Erzeuger von $\mathbb{B}$. Genauso zeigt man, dass etwa $\{(-\infty, x) : x \in \mathbb{Q}\}$ ein Erzeuger von $\mathbb{B}$ ist.

Eine weitere Möglichkeit zur Einschränkung der oben monierten Beliebigkeit bei der Wahl von Modellen besteht darin, Verteilungen aus „elementareren" herzuleiten. Zum Beispiel wäre jede Angabe einer Zähldichte für die „Anzahl der Richtigen beim Lotto n aus N" ziemlich willkürlich. Plausibler ist die Annahme der Laplace-Verteilung aller „Ziehungen" in $\Omega \equiv \{Z \subseteq \{1, \ldots, N\} : |Z| = n\}$. Für jeden „Tipp" $T \in \Omega$ und jedes $k \in \{0, \ldots, n\}$ ist durch $Z \mapsto (Z \cap T, Z \cap T^c)$ eine Bijektion zwischen $A_k \equiv \{Z \in \Omega : |Z \cap T| = k\}$, also dem Ereignis „$k$ Richtige", und $\{R \subseteq T : |R| = k\} \times \{F \subseteq T^c : |F| = n - k\}$ definiert. Wegen $|\Omega| = \binom{N}{n}$ folgt

$$P(A_k) = \frac{\binom{n}{k}\binom{N-n}{n-k}}{\binom{N}{n}}.$$

Durch diesen Ausdruck wird die Zähldichte der (speziellen) **hypergeometrischen Verteilung** $H(N, n)$ auf der Potenzmenge von $\{0, \ldots, n\}$ definiert.

Wie in diesem Beispiel lassen sich Ereignisse oft sehr bequem durch Abbildungen auf dem Raum der Konstellationen beschreiben, und in der Wahrscheinlichkeitstheorie ist es üblich, diese Abbildung mit Großbuchstaben vom Ende des lateinischen Alphabets zu bezeichnen (was für Studierende mit Kenntnissen aus der Grundvorlesung über Analysis zunächst gewöhnungsbedürftig ist). Wir hätten etwa das Ereignis $A = \{(\omega_1, \ldots, \omega_n) \in \Omega : \text{ es gibt } i \neq j \text{ mit } \omega_i = \omega_j\}$ aus dem Kollisionsparadoxon auch als $A = \{\omega \in \Omega : X(\omega) < n\}$ mit der Abbildung $X(\omega_1, \ldots, \omega_n) \equiv |\{\omega_1, \ldots, \omega_n\}|$ beschreiben können.

Wir benutzen für eine Abbildung $X : \Omega \to \mathcal{X}$ und ein Ereignis $B \subseteq \mathcal{X}$ außer der üblichen Bezeichnung $X^{-1}(B) \equiv \{\omega \in \Omega : X(\omega) \in B\}$ für das **Urbild** auch die kürzere Schreibweise $\{X \in B\}$ und analog $\{X \in B, Y \in C\} = \{X \in B\} \cap \{Y \in C\}$ oder $\{X = Y\} = \{\omega \in \Omega : X(\omega) = Y(\omega)\}$ für zwei Abbildungen X, Y auf Ω.

Falls $X : \Omega \to \mathcal{X}$ bijektiv ist, stimmt das Urbild mit dem Bild unter der meist ebenfalls mit X^{-1} bezeichneten Umkehrabbildung überein, so dass die leider häufige Verwechslung zwischen *Umkehr-* und *Urbildabbildung* im Fall bijektiver Abbildungen nicht zu Fehlern führt. Verwechslungen von X^{-1} mit dem multiplikativen Inversen $x^{-1} = 1/x$ von reellen Zahlen sind hingegen kaum zu befürchten.

Folgendes Beispiel zeigt die Prägnanz obiger Schreibweisen. Seien $\Omega \equiv \{0, 1\}^n$, $X_j(\omega_1, \ldots, \omega_n) \equiv \omega_j$ und $X(\omega) \equiv \sum_{j=1}^n X_j(\omega)$. Interpretieren wir die j-te Komponente einer Konstellation als Erfolg ($\omega_j = 1$) beziehungsweise Misserfolg ($\omega_j = 0$) in der j-ten Wiederholung eines Versuchs, so beschreibt $X(\omega)$ die Anzahl der Erfolge.

Für die Laplace-Verteilung P auf Ω und $k \in \{0, \dots, n\}$ gilt dann

$$\begin{aligned} P(\{X = k\}) &= P\Big(\bigcup_{|S|=k} \{X_j = 1 \text{ für } j \in S\} \cap \{X_j = 0 \text{ für } j \notin S\}\Big) \\ &= \sum_{|S|=k} 1/2^n = \binom{n}{k} 1/2^n, \end{aligned}$$

weil $\{X_j = 1 \text{ für } j \in S\} \cap \{X_j = 0 \text{ für } j \notin S\}$ einelementig ist. Dies ist die Zähldichte von $B(n, 1/2)$.

Für jede Abbildung $X : \Omega \to \mathcal{X}$ haben wir durch $B \mapsto X^{-1}(B)$ die Urbildabbildung $\mathcal{P}(\mathcal{X}) \to \mathcal{P}(\Omega)$ definiert, und für $\mathcal{B} \subseteq \mathcal{P}(\mathcal{X})$ nennen wir das Bild $X^{-1}(\mathcal{B}) = \{X^{-1}(B) : B \in \mathcal{B}\}$ unter der Urbildabbildung auch kürzer Urbild von $\mathcal{B}$. Mit $\mathcal{B}$ ist auch $X^{-1}(\mathcal{B})$ eine σ-Algebra.

Um wie im Beispiel eben Wahrscheinlichkeiten von Ereignissen $A = X^{-1}(B)$ messen zu können, müssen sie natürlich im Definitionsbereich des Wahrscheinlichkeitsmaßes liegen. Für zwei Messräume $(\Omega, \mathcal{A})$ und $(\mathcal{X}, \mathcal{B})$ nennen wir deshalb eine Abbildung $X : \Omega \to \mathcal{X}$ **messbar** oder genauer $(\mathcal{A}, \mathcal{B})$-messbar, falls $X^{-1}(\mathcal{B}) \subseteq \mathcal{A}$ gilt. Dann schreiben wir auch: $X : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$ ist messbar.

Ist μ ein Maß auf $\mathcal{A}$, so heißt $\mu^X \equiv \mu \circ X^{-1}$ **Bildmaß** von μ unter X. Für $B \in \mathcal{B}$ gilt also

$$\mu^X(B) = \mu(X^{-1}(B)) = \mu(\{X \in B\})$$

und oft schreiben wir dafür auch $\mu(X \in B)$.

Für ein Wahrscheinlichkeitsmaß P nennen wir P^X die **Verteilung** von X (unter P). In dieser Situation heißt X auch $(\mathcal{X}, \mathcal{B})$-wertige **Zufallsgröße** (auf $(\Omega, \mathcal{A}, P)$) und wir schreiben $X \sim P^X$, was insbesondere dann nützlich ist, wenn P^X eine bekannte Verteilung ist. $X \sim B(n, p)$ bedeutet also $P^X = B(n, p)$ und X heißt dann auch $B(n, p)$-verteilt.

$(\mathbb{R}^n, \mathbb{B}_n)$-wertige Zufallsgrößen heißen n-dimensionale **Zufallsvektoren** und im Fall $n = 1$ auch Zufallszahlen oder **Zufallsvariablen**.

Zwei Zufallsgrößen X, Y heißen **identisch verteilt**, falls sie die gleiche Verteilung besitzen (X, Y müssen also gleichen Wertebereich haben, können aber auf verschiedenen Wahrscheinlichkeitsräumen definiert sein). Auch wenn die dadurch definierte Äquivalenzrelation „weit von der Identität entfernt ist", schreiben wir dann $X \stackrel{d}{=} Y$ (das d steht für „distribution"), X und Y sind also **verteilungsgleich**. Ist zum Beispiel $X \sim B(1, 1/2)$, so gilt $X \stackrel{d}{=} X^2 \stackrel{d}{=} 1 - X$. Dieses einfache Beispiel zeigt, dass aus $X \stackrel{d}{=} \tilde{X}$ und $Y \stackrel{d}{=} \tilde{Y}$ *weder* $(X, Y) \stackrel{d}{=} (\tilde{X}, \tilde{Y})$ *noch* $X + Y \stackrel{d}{=} \tilde{X} + \tilde{Y}$ folgt. Dabei sind (X, Y) und $X + Y$ „argumentweise" als $(X, Y)(\omega) = (X(\omega), Y(\omega))$ und $(X + Y)(\omega) = X(\omega) + Y(\omega)$ definiert.

Bei der Notation $X \sim B(n, p)$ werden sowohl der Raum der Konstellationen als auch das Wahrscheinlichkeitsmaß P ignoriert. Der Grund dafür ist (ein metaphysisch zart besaiteter Leser möge die folgenden Zeilen ruhig überspringen, sie sind für die mathematische Theorie belanglos), dass es für die Stochastik in der Regel keine Rolle spielt, wie die Konstellationen zustande kommen, etwa ob die Lottomaschine samstags mit einem Rad oder mittwochs mit einem Gebläse betrieben wird. Auch ist es unerheblich, dass die realen Konstellationen viele weitere Eigenschaften haben können, die im Modell nicht berücksichtigt werden. Zum Beispiel werden beim Lotto die Kugeln nacheinander gezogen, was wir im Beispiel zur hypergeometrischen Verteilung ignoriert haben. Eine sehr vage aber vielleicht nützliche Vorstellung ist manchmal, Ω als sehr groß anzunehmen – etwa als Menge „aller möglichen Weltläufe“ – und P als „den Zufall, der den aktuellen Weltverlauf steuert“. Die Verteilungen P^X von Zufallsvariablen liefern also Kenntnisse über das unzugängliche P. Zurück zur Sache.

Die Messbarkeit einer Abbildung ist eine technisch (und auch bei der Interpretation als Information) wichtige Eigenschaft, die glücklicherweise meistens leicht zu verifizieren ist. Nützlich sind dabei oft folgende Aussagen.

Satz 1.6 (Urbild σ-Algebren)
Seien $X : \Omega \to \mathcal{X}$ eine Abbildung und $\mathcal{E} \subseteq \mathcal{P}(\mathcal{X})$. Dann ist die vom Urbild erzeugte σ-Algebra das Urbild der erzeugten σ-Algebra, also

$$\sigma_\Omega(X^{-1}(\mathcal{E})) = X^{-1}(\sigma_\mathcal{X}(\mathcal{E})).$$

Beweis. $X^{-1}(\sigma_\mathcal{X}(\mathcal{E}))$ ist eine σ-Algebra über Ω, die $X^{-1}(\mathcal{E})$ und daher auch die erzeugte σ-Algebra $\sigma_\Omega(X^{-1}(\mathcal{E}))$ umfasst. Andererseits ist $\mathcal{G} \equiv \{B \subseteq \mathcal{X} : X^{-1}(B) \in \sigma_\Omega(X^{-1}(\mathcal{E}))\}$ eine σ-Algebra über $\mathcal{X}$, die $\mathcal{E}$ und damit auch $\sigma_\mathcal{X}(\mathcal{E})$ umfasst. □

Wegen dieses Satzes muss man für den Nachweis der Messbarkeit einer Abbildung $X : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$ nur $\{X \in B\} \in \mathcal{A}$ für alle Mengen eines Erzeugers von $\mathcal{B}$ überprüfen. Wir haben im Zusammenhang mit der Definition der Verteilungsfunktion nach Satz 1.5 gezeigt, dass $\mathcal{E} \equiv \{(-\infty, x] : x \in \mathbb{R}\}$ ein Erzeuger der Borel-σ-Algebra $\mathbb{B}$ ist. Mit Satz 1.6 erhalten wir also, dass eine Abbildung $X : \Omega \to \mathbb{R}$ genau dann $(\mathcal{A}, \mathbb{B})$-messbar, wenn $\{X \le x\} \in \mathcal{A}$ für alle $x \in \mathbb{R}$ gilt.

Diese Kriterium impliziert zum Beispiel, dass jede monotone Funktion $X : \mathbb{R} \to \mathbb{R}$ bezüglich der Borel-σ-Algebra messbar ist, weil dann Urbilder von Intervallen wieder Intervalle sind.

Satz 1.7 (Komposition und Messbarkeit)
Sind $X : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$ und $Y : (\mathcal{X}, \mathcal{B}) \to (\mathcal{Y}, \mathcal{C})$ beide messbar, so ist auch die Komposition $Y \circ X : (\Omega, \mathcal{A}) \to (\mathcal{Y}, \mathcal{C})$ messbar. Für jedes Maß μ auf $\mathcal{A}$ gilt $\mu^{Y \circ X} = (\mu^X)^Y$.

Beweis. Für $B \in \mathcal{B}$ gilt $(Y \circ X)^{-1}(B) = \{\omega \in \Omega : Y(X(\omega)) \in B\} = \{\omega \in \Omega : X(\omega) \in Y^{-1}(B)\} = X^{-1}(Y^{-1}(B))$. □

Gelegentlich tritt die Situation auf, dass Y bloß auf dem Bild $M \equiv X(\Omega) \subseteq \mathcal{X}$ definiert ist. Wir betrachten dann die **Spur-σ-Algebra** $\mathcal{B} \cap M \equiv \{B \cap M : B \in \mathcal{B}\}$. Fassen wir X als Abbildung $\Omega \to M$ auf, so bleibt wegen $X^{-1}(B \cap M) = X^{-1}(B)$ für alle $B \in \mathcal{B}$ die Messbarkeit erhalten. Ist dann $Y : (M, \mathcal{B} \cap M) \to (\mathcal{Y}, \mathcal{C})$ messbar, so folgt die Messbarkeit der Komposition $Y \circ X$ aus Satz 1.7.

In der Situation von Borel-σ-Algebren erhalten wir als Anwendung der letzten beiden Sätze:

Satz 1.8 (Messbarkeit und Stetigkeit)

1. Sind $\mathcal{A}$ und $\mathcal{B}$ die Borel-σ-Algebren über den metrischen Räumen Ω beziehungsweise $\mathcal{X}$, so ist jede stetige Abbildung von Ω nach $\mathcal{X}$ auch messbar.

2. Seien $\mathcal{X}$ ein metrischer Raum mit Borel-σ-Algebra $\mathcal{B}$ und $X_n : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$ messbar, so dass für jedes $\omega \in \Omega$ der Grenzwert $X(\omega) \equiv \lim_{n\to\infty} X_n(\omega)$ existiert. Dann ist X ebenfalls $(\mathcal{A}, \mathcal{B})$-messbar.

Beweis. 1. Ist $\mathcal{E}$ die Menge der offenen Teilmengen von $\mathcal{X}$, so folgt mit Satz 1.6 die Beziehung $X^{-1}(\mathcal{B}) = X^{-1}(\sigma_{\mathcal{X}}(\mathcal{E})) = \sigma_{\Omega}(X^{-1}(\mathcal{E})) \subseteq \mathcal{A}$, weil stetige Urbilder offener Mengen offen sind.

2. Für jede Teilmenge B von $\mathcal{X}$ ist durch

$$f(x) \equiv \operatorname{dist}(x, B) \equiv \inf\{d(x, b) : b \in B\}$$

eine stetige **Abstandsfunktion** definiert. Sind nämlich $x \in \mathcal{X}$, $\varepsilon > 0$ und $b \in B$ mit $d(x, b) \leq$ dist $(x, B) + \varepsilon$, so folgt für jedes $y \in \mathcal{X}$

$$f(y) - f(x) \leq d(y, b) - d(x, b) + \varepsilon \leq d(y, x) + \varepsilon.$$

Durch Grenzübergang $\varepsilon \to 0$ und Rollentausch erhalten wir, dass f sogar eine **Kontraktion** ist, das heißt $|f(x) - f(y)| \leq d(x, y)$. Ist nun $B \subseteq \mathcal{X}$ abgeschlossen, so gilt

$$\{X \in B\} = \{f \circ X = 0\} = \{\lim_{n\to\infty} f \circ X_n = 0\} = \bigcap_{k\in\mathbb{N}} \bigcup_{n\in\mathbb{N}} \bigcap_{m\geq n} \{f \circ X_m < 1/k\}.$$

Wegen 1. und Satz 1.7 ist daher $\{X \in B\} \in \mathcal{A}$. Weil die Menge der abgeschlossenen Teilmengen die Borel-σ-Algebra erzeugt, folgt aus Satz 1.6 die Messbarkeit von X. □

Messbarkeit ist also nicht nur eine schwächere Eigenschaft als Stetigkeit, sie hat auch viel bessere Permanenzeigenschaften (für die Stetigkeit der Grenzfunktion in der Situation 1.8.2 müsste man etwa gleichmäßige Konvergenz voraussetzen).

Ein weiteres Beispiel für die guten Permanenzeigenschaften ist die **fallweise Beschreibung** von Abbildungen: Sind $X_n : (\Omega, \mathcal{A}) \to (\mathcal{X}, \mathcal{B})$ messbar und $A_n \in \mathcal{A}$ paarweise disjunkt mit $\bigcup_{n\in\mathbb{N}} A_n = \Omega$, so ist durch $Y(\omega) \equiv X_n(\omega)$, falls $\omega \in A_n$,