Institut für Angewandte Mathematik
Universität Heidelberg

Im Neuenheimer Feld 294
69120 Heidelberg

Hermann Rost

Grundvorlesung Wahrscheinlichkeitstheorie

Vorlesungsmanuskript

Sommer 1993
Neufassung vom Sommer 1996

Vorbemerkungen

Die vorliegenden Aufzeichnungen stellen den Inhalt der Grundvorlesung Wahrscheinlichkeitstheorie dar, wie ich sie im SS 1993 und ähnlich im SS 1996 gelesen habe. Zur Stoffauswahl zwei Bemerkungen. Erstens: Fragen der mathematischen Statistik sind vollständig ausgeklammert worden, aus dem einfachen Grund, weil es in Heidelberg eine eigene Grundvorlesung Statistik gibt. Zweitens: Es ist unvermeidlich, daß ein großer Teil des Materials eines solchen Skripts "Standardstoff" sein wird, wie er in vielen Büchern enthalten ist. Ich habe in der Vorlesung aber bewußt einige nicht unbedingt im Durchschnitt aller Einführungstexte enthaltene Gegenstände ausführlicher behandelt, wie Codierung, Große Abweichungen, Perkolation, Poissonprozeß, weil besagter Standardstoff auch anderswo nachgelesen werden kann. Als ein Beispiel für viele sei genannt *U. Krengel: Einführung in die Wahrscheinlichkeitsrechnung und Statistik. Braunschweig: Vieweg 1988*. Die Literaturangaben, die sich am Ende der einzelnen Kapitel finden, beziehen sich entweder auf historisches Quellenmaterial oder weiterführende und vertiefende Texte, aber meist nicht auf "normale" Lehrbücher. Meine Absicht ist es, den Leser dazu zu bringen, auch bei vermeintlich simplen Dingen, die schon in Dutzenden von Büchern aufgeschrieben sind, an einzelnen Stellen einmal gründlicher nachzusehen. Auch der Anfänger in einer Disziplin sollte aufpassen, daß er sich nicht zu früh Scheuklappen an-legt, indem er sich mit *einem* Buch und dem vermeintlich kanonischen Stoff zufrieden gibt.

Heidelberg, 17. März 1996 H.R.

Inhalt

Kapitel Seite

1.Gleichverteilung2
2. Wahrscheinlichkeitsräume 19
3. Die Binomialverteilung35
4. Die Poissonverteilung46
5. Die Shannonsche Entropie54
6. Der zentrale Grenzwertsatz65
7. Der Poissonprozeß76
8. Starkes Gesetz der großen Zahlen84
9.Gerichtete Perkolation96

Die gesammelten Übungsaufgaben und Lösungen werden am Semesterende nachgeliefert.

Kapitel 1. Gleichverteilung

Inhalt: Laplacesche Definition, Kombinatorisches (Kartenspiel, Besetzungszahlen, Zuf. Permutationen), Grundbegriffe (vor allem bedingte Ws.), Buffonsche Nadel.

Solange man nicht weiß, was Wahrscheinlichkeit ist, und vor allem nicht, wie man den Zahlenwert der Ws. in einer konkreten Situation angeben soll, können Symmetrieüberlegungen von großem Nutzen sein. Wenn man etwa ein Experiment mit einer gewissen Menge von möglichen Ausgängen beobachtet, auf welchen aufgrund der Versuchsanordnung eine Symmetriegruppe operiert, die reich genug ist, so kann es sinnvoll sein zu postulieren, daß die Ausgänge gleichwahrscheinlich seien. Man definiert dann die Ws. des interessierenden Er-eignisses nach der klassischen Formel von Laplace ([L], S. XI) als

(1) Wahrscheinlichkeit := (Anzahl der günstigen Fälle)/(Anzahl der möglichen Fälle).

Diese Formel ist in vielen Anwendungssituationen brauchbar, und wir wollen in diesem Kapitel nur Beispiele betrachten, in denen die Ws. nach der Regel (1) gebildet ist. Die didaktische Absicht ist, anhand der Beispiele das grundlegende Vokabular der W.-Theorie einzuführen und gleichzeitig gewisse einfache kombinatorische Argumentationen vorzustellen. Übrigens, aus (1) folgt, daß die Wsen von disjunkten, d.h. sich gegenseitig ausschließenden, Ereignissen sich addieren.

Für den Naturwissenschaftler ist die Anwendung der Laplaceschen Regel alles andere als tri-vial. Er muß sich nämlich Rechenschaft darüber ablegen, ob die ihr zugrundeliegende Sym-metrieannahme gerechtfertigt ist. Das heißt, daß man sich etwas genauer überlegen muß, was ein "möglicher Fall" ist, und ob nicht unser Raster der Daten, die wir in einer konkreten Situ-ation erheben, zu grob ist und ein "Fall" einer möglichen Beobachtung in Wirklichkeit aus vielen "Fällen" zusammengesetzt ist, die man nur bei feinerer Beobachtung wahrnehmen kann, und die mit größerem Recht als gleich wahrscheinlich anzusehen sind. Überlegungen dieser Art werden wir in den folgenden konkreten Beispielen immer wieder anstellen.

Beispiel 1. Ziehen aus einer Urne.

Aus einer Urne mit w weißen und s schwarzen Kugeln wird n-mal gezogen. Nach jeder Zie-

hung werde die gezogene Kugel wieder zurückgelegt. (Ziehen mit Zurücklegen.) Wie ws. ist es, daß genau k-mal eine weiße Kugel erscheint?

Man denkt sich die Kugeln von 1 bis $u := s+w$ durchnumeriert; man tut so, als ob es einem idealen Beobachter möglich sei, die Nummer der jeweils gezogenen Kugel zu registrieren. Die Symmetrieannahme über das Ziehungsergebnis ist, daß alle Sequenzen $(i(1), .., i(n))$, mit $i(m)$ einer Zahl von 1 bis u , gleich ws. sind. Davon gibt es u^n Stück und das ist die Anzahl der möglichen Fälle. Die günstigen Fälle erhält man, indem man zuerst die Zeitpunk-te fixiert, zu denen eine weiße Kugel erscheint. Das kann auf $C(n,k) := n!\,(k!(n-k)!)^{-1}$ Wei-sen geschehen. Zu jeder Wahl der Menge der Zeitpunkte gibt es $w^k\, s^{(n-k)}$ Mglichkeiten, die individuellen Kugeln, die gezogen werden, zu bestimmen. Wir erhalten als Ergebnis

(2) $P(\text{genau k-mal erscheint weiß}) = C(n,k)\, w^k\, s^{(n-k)}\, u^{-n} = C(n,k)\, (w/u)^k\, (s/u)^{n-k}$.

Hieran ist bemerkenswert, daß diese Zahl nur vom Anteil w/u der weißen Kugeln in der Urne abhängt, sich also nicht ändert, wenn man z.B. w und u verdoppelt.

Wenn man den zweiten Satz der Voraussetzung oben ändert in "Eine gezogene Kugel wird nicht mehr in die Urne zurückgelegt" (Ziehen ohne Zurücklegen), ändert sich auch die Antwort auf die Frage. Die Anzahl der möglichen Fälle ist jetzt $u\cdot(u-1)\cdot\ ..\ \cdot(u-n+1)$, die Anzahl der günstigen, ähnlichen dem obigen Argument, $C(n,k)\cdot w\ ..\cdot(w-k+1)\cdot s\ ..\ \cdot(s-n+k+1)$. Somit:

(3) $P(\text{genau k-mal erscheint weiß}) = n!\, w!\, (u-w)!(u-n)!\, /(k!(n-k)!(w-k)!(u-w-n+k)!u!)$.

Dieser Ausdruck ist keine Funktion von w/u allein; er hat aber die schöne Eigenschaft - die man leicht begründen kann - daß er bei festem u und k symmetrisch in w und n ist. Am leichtesten merkt man ihn sich in der Form

(3a) $P(\text{genau k-mal erscheint weiß}) = C(u,n)^{-1}\, C(w,k)\, C(u-w,n-k)$,

mit der Deutung: ein möglicher Fall ist die Bestimmung einer n-elementigen Teilmenge aus der Menge aller Kugeln; er ist "günstig", wenn diese Teilmenge genau k weiße und damit $n-k$ schwarze Kugeln enthält.

Bemerkung. Es könnte jemand auf die Idee kommen, die Gleichverteilung für die Menge aller Zahlen k anzusetzen, d.h. zu sagen, daß jede Zahl von 0 bis n die gleiche Ws. hat, als Anzahl der gezogenen weißen Kugeln bei n-maligem Ziehen zu erscheinen. Wenn man aber etwas genauer hinsieht und danach fragt, wie diese Anzahl zustandekommt, d.h. nach dem Mechanismus des Ziehens, wird man auf eine Antwort ("Modell") vom obigen Typ geführt.

Beispiel 2. Kartenspiel.

Welches ist die Ws., daß der Spieler Ost im Bridge genau 3 Asse erhält?

Wir denken uns je 13 Plätze vor jedem der 4 Spieler aufgebaut, auf welche die Karten zufällig permutiert gelegt werden. Die Symmetrieannahme ist, daß alle 52! Permutationen des Kartenstoßes gleich ws. sind. Es gibt $C(4,3) = 4$ Möglichkeiten zu bestimmen, welche Asse Ost bekommt. Es gibt $13 \cdot 3 = 39$ mögliche Plätze für die 4. As und $13 \cdot 12 \cdot 11$ Möglichkeiten, wo die drei Asse von Ost liegen können. Hat man die Plätze der 4 Asse bestimmt, dann gibt es noch 48! mögliche Permutationen für die restlichen Karten. Somit ist die gesuchte Zahl

(4) P(Ost hat genau 3 Asse) = $4 \cdot 39 \cdot 13 \cdot 12 \cdot 11 \cdot 48!/52! = 39 \cdot 12 \cdot 11 / 51 \cdot 50 \cdot 49 = 0.0412$.

Eine andere Argumentation,welche ein unnötiges Aufbauschen von Zähler und Nenner in (4) vermeidet, ist die folgende: es gibt $C(52,4)$ Möglichkeiten, die Plätze des Tisches auszuwählen, wo Asse liegen. Die Gleichverteilung auf der Menge der Permutationen impliziert die Gleichverteilung auf der Menge dieser $C(52,4)$ Auswahlen. Günstige Auswahlen für unser Ereignis gibt es $C(13,3)$ mal $C(39,1)$; man muß nämlich separat festlegen, wo bei Ost und wo auf den übrigenen Plätzen die Asse placiert werden. Das führt auf die gleiche Zahl:

(4a) P(Ost hat genau 3 Asse) = $13 \cdot 12 \cdot 11 \cdot 39 \cdot 2 \cdot 3 \cdot 4 / 52 \cdot 51 \cdot 50 \cdot 49 \cdot 2 \cdot 3 = 39 \cdot 12 \cdot 11 / 51 \cdot 50 \cdot 49$.

Die zweite Argumentation bedeutet nichts anderes, als Beispiel 2 unter das Schema des Ziehens ohne Zurücklegen zu subsumieren: man zieht 13-mal aus einer Urne mit 52 Kugeln, von denen 4 eine besondere Qualität haben, und fragt nach der Ws. dafür, daß von den besonde-ren Kugeln genau 3 erscheinen.

Beispiel 3. Besetzungszahlen.

Es werden k Kugeln in s Schachteln gelegt. Welches ist die Ws. dafür, daß in die erste Schachtel $k(1)$ Kugeln, in die zweite $k(2)$ Kugeln usw. zu liegen kommen? Die Fragestel-lung ist noch zu präzisieren, da noch nicht klar ist, was als "Fall" im Sinn der Laplaceschen Definition anzusehen ist. Es gibt zwei gängige Präzisierungen, die wir hier vorstellen, sowie eine etwas "abseitigere" dritte Möglichkeit.

a) Die Kugeln sind unterscheidbar.

Man denkt sich die Kugeln von 1 bis k durchnumeriert (die Schachteln sind ohnehin unterscheidbar); ein "Fall" ist, wenn man zu jeder Kugel die Schachtel angibt, in welcher sie liegt. In diesem Sinn gibt es s^k mögliche Fälle. Die günstigen Fälle zählt man, indem man zuerst die Kugeln spezifiziert, welche in der ersten Schachtel liegen; das kann man auf C(k,k(1)) Arten tun. Dann bestimmt man die Kugeln in der zweiten Schachtel; es gibt C(k-k(1),k(2)) Ar-ten, dies zu tun, usw. Das Produkt der Binomialkoeffizienten ergibt schließlich den Wert

(5) $M(k ; k(1),k(2), ..,k(s)) := k!/(k(1)!\, k(2)!\, ..\, k(s)!)$, wenn $k = \sum k(i)$
$= 0$ sonst.

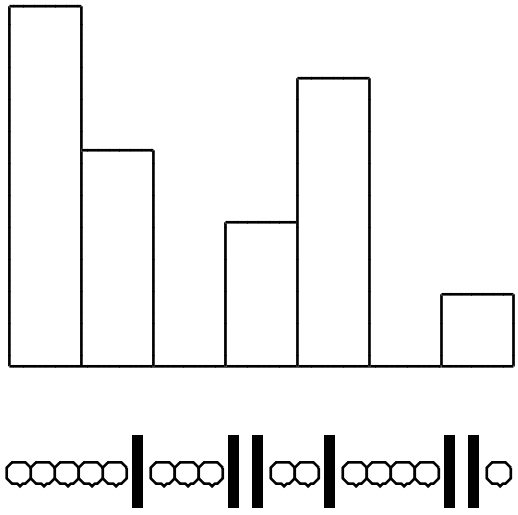
Man nennt M( ) einen Multinomialkoeffizienten. Der Grund liegt in der Identität

(5a) $(u(1) + u(2) .. +u(s))^k = \sum_{k(1), .. ,k(s)} \prod_{1 \le i \le s} M(k;k(1), ..,k(s))u(i)^{k(i)}$,

welche man als Multinomialsatz bezeichnet. (Sie stellt übrigens sicher, daß sich die Gewichte zu 1 summieren: setze alle u(i) = 1 .) Damit ist die gesuchte Ws.

(6) $P(k(i)$ Kugeln in Schachtel $i , 1 \le i \le s) = M(k;k(1), ..,k(s))\, s^{-k}$.

b) Die Kugeln sind ununterscheidbar.



Man postuliert die Gleichverteilung auf der Menge aller s-Tupel (k(1), .. , k(s)) von natürlichen Zahlen mit $\sum k(i) = k$. Die Anzahl der günstigen Fälle ist dann trivialerweise gleich 1 . Die Anzahl der möglichen Fälle bekommt man durch folgendes Bild. Jedes s-Tupel mit Summe k entspricht umkehrbar eindeutig einer Anordnung von k Kugeln und s-1 Trennwänden auf k+s-1 Plätzen, da jede solche Anordnung eindeutig bestimmt ist durch Angabe der Zahl der Kugeln

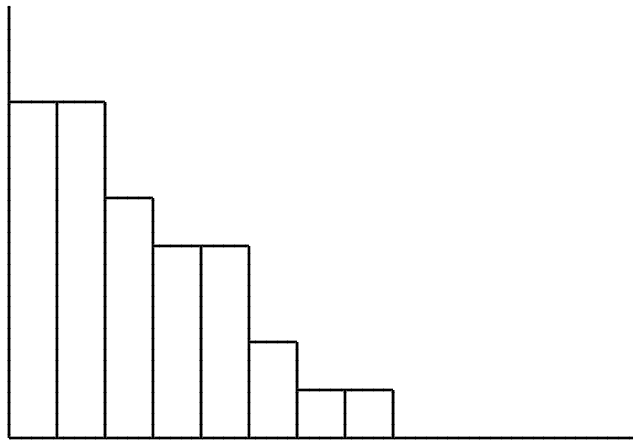
links von der ersten Trennwand, zwischen der ersten und zweiten Wand, .. und rechts von der letzten Wand. (S. obige Zeichnung.) Also gibt es $C(k+s-1 , k)$ Fälle und die gesuchte Ws. ist

(7) $P(k(i)\text{ Kugeln in Schachtel } i , 1 \le i \le s) = 1/C(k+s-1,k) = k!(s-1)!/(k+s-1)!$.

Bemerkung: Die Unterscheidung zwischen den Modellen a) und b) spielt in der Quantenmechanik beim Abzählen der möglichen "Zustände" eines Mehrteilchensystems eine wichtige Rolle. Die jeweiligen Ansätze werden dort unter dem Namen der Maxwell-Boltzmannschen (a) bzw. Bose-Einsteinschen (b) Statistik geführt.

c) Die Schachteln sind ununterscheidbar.

Das klingt im Rahmen der gestellten Frage etwas merkwürdig. Gemeint ist folgendes: ein "Fall" bestehe jetzt in einer noch gröberen Information als in b), nämlich in der Angabe, wieviele Schachteln genau 1 Kugel, 2 Kugeln, usw. enthalten. Oder anders ausgedrückt: man ordnet die Koordinaten im s-Tupel $(k(1), .., k(s))$ so, daß sie eine nichtwachsende Folge bilden und bezeichnet das so erhaltene geordnete Tupel als einen "Fall". Man nennt ein Tupel natürlicher Zahlen (mit endlicher Summe), das in absteigender Reihenfolge angeordnet ist, ein Young-Tableau. Das Problem besteht darin, die Anzahl $N(k,s)$ aller Young-Tableaux mit Summe ("Fläche")gleich k und horizontaler (oder vertikaler) Ausdehnung höchstens s zu



Young-Tableau (Fläche 31, Breite 8, Höhe 7).

bestimmen. Anderes Bild: ein Young-Tableau der Fläche k und Höhe höchstens s ist eine Darstellung der Zahl k als Summe von Zahlen kleinergleich s . Die erzeugende Funktion der $N(k,s)$ ist leicht anzugeben (siehe [PSz], Bd. I, 1. Abschnitt, Aufg. 9):

$$\sum_k N(k,s)\, u^k = \prod_{1 \le i \le s} \{1 - u^i\}^{-1} \quad , \qquad \sum_k N(k,k)\, u^k = \prod_{i \ge 1} \{1 - u^i\}^{-1} \quad .)$$

Weiter in der Entwicklung der allgemeinen Theorie. Wir führen einige Sprechweisen ein:

Def. 1.1. Die Ergebnismenge E ist die Menge der möglichen Fälle (hängt vom gewählten Modell ab); auf E haben wir die Gleichverteilung, welche jedem Ergebnis das Gewicht $1/|E|$ zuordnet; ein Ereignis ist eine Teilmenge der Ergebnismenge; $P(A)$, die Wahrscheinlichkeit des Ereignisses A ist die Summe aller Gewichte seiner Elemente, d.h. $P(A) = |A|/|E|$; eine Zufallsvariable ist eine in E definierte Funktion X mit Werten in einem Raum F ; die Ver-teilung von X ist die Familie $P(X = f) = P(\{e : X(e) = f\}) = |\{e : X(e) = f\}|/|E|$, $f \in F$. Insbesondere ist möglich, daß der Raum F ein Cartesisches Produkt ist, und eine F-wertige Zufallsvariable X sich daher als Paar (oder Tripel usw.) darstellen läßt: $X = (U,V)$. In diesem Fall nennt man die Verteilung von X auch die gemeinsame Verteilung von U und V .

In dieser Sprache kann man den Sachverhalt von Beispiel 1 so ausdrücken: man definiert bei n-maliger Ziehung aus der Urne die Zufallsvariable X als Anzahl der Ziehungen, die eine weiße Kugel erbringen. Dieses X hat im Fall des Ziehens mit Zurückliegen die Verteilung

$$P(X = k) = C(n,k)(w/u)^k (1- w/u)^{n-k} , 0 \le k \le n ;$$

im Fall des Ziehens ohne Zurücklegen hingegen

$$P(X = k) = C(u,n)^{-1} C(w,k) C(u-w,n-k) , 0 \le k \le n.$$

Man nennt die erste Verteilung die Binomialverteilung zu den Parametern n und w/u , die zweite die hypergeometrische Verteilung zu den Parametern u , n und w .

Def. 1.2. Als Erwartung (Erwartungswert, Mittelwert) einer reellen Zufallsvariablen X bezeichnet man die Zahl

$$(8) \qquad EX = \sum_f P(X = f)\cdot f = \sum_e P(e)\cdot X(e) \; (= \sum_e X(e) /|E|) .$$

Aus der zweiten - weniger aus der ersten - Darstellung in (8) folgt sofort

Prop. 1.1. Die Menge der reellen Zufallsvariablen in E bildet einen geordneten Vektorraum; die Erwartung ist eine lineare und monotone Funktion auf diesem Raum. Die Erwartung der konstanten Funktion c (welche ebenfalls eine Zufallsvariable ist) beträgt c .

In Beispiel 1 läßt sich die Erwartung von X , der Anzahl der gezogenen weißen Kugeln, in beiden Varianten leicht ausrechnen. Man hat $X = \sum_{1 \le i \le n} X(i)$, wo $X(i) = 1$, wenn die i-te Ziehung eine weiße Kugel erbringt, und $X(i) = 0$ sonst. Die Erwartung von jedem $X(i)$ ist w/u . (In der ersten Variante ist das klar, in der zweiten zumindest für $i = 1$. Mann überlegt sich aber, daß die Verteilung des Tupels $(X(1), . , X(n))$ auch in der zweiten Variante, der des Ziehens ohne Zurücklegen, symmetrisch bezüglich Permutationen der Indices ist. Also haben alle X(i) die gleiche Erwartung.) Folglich ist $EX = n \cdot w/u$, in jeder Variante.

Beispiel 4. Fixpunkte einer Permutation.

Sei X die Anzahl der Fixpunkte einer zufälligen Permutation von n Elementen. Man be-rechne EX . (Zur Worterklärung: "zufällige Permutation" bedeutet, daß wir die Gleichver-teilung auf der Menge aller Permutationen zugrundelegen. Der Wert von X an der Stelle π ist die Zahl der Fixpunkte der Permutation π .)

Man stellt wieder X als Summe dar, $X = \sum_{1 \le i \le n} X(i)$, wobei man setzt: $X(i)(\pi) = 1$, wenn das i-te Element Fixpunkt von π ist, und $X(i)(\pi) = 0$ sonst. Es gibt $(n-1)!$ Permutationen, welche das i-te Element fest lassen; somit gilt $EX(i) = P(X(i) = 1) = (n-1)!/n! = 1/n$. Damit hat man $EX = 1$, unabhängig von n .

Hier, wie auch schon im Beispiel davor, wäre es übrigens sehr mühsam, alle $n!$ Permutationen einzeln hinzuschreiben, von jeder Permuation die Fixpunktzahl festzustellen und diese Zahlen zu addieren. Der Leser versuche es einmal im Fall $n = 4$, wenn er nichts besseres zu tun hat. - Für spätere Zwecke halten wir den Sprachgebrauch fest:

Def. 1.3. Wenn A ein Ereignis ist, so heißt die Zufallsvariable 1_A , definiert als

(9) $\qquad 1_A(e) := 1$, wenn $e \in A$, und 0 sonst ,

der Indikator von A .

Das folgende Beispiel ist vielleicht etwas künstlich; es wurde aufgenommen, um zusammen mit Beispiel 4 zu unterstreichen, daß sich die Erwartung im Raum der Zufallsvariablen addi-tiv verhält, ohne daß es irgend einer Unabhängigkeitsvorausetzung (s. Kapitel 2) an die Vari-ablen bedürfte.

Beispiel 5. Eine fiktive Lotterie.

Aus einer Urne mit 7 weißen und 5 schwarzen Kugeln wird 10-mal gezogen. Ich darf gegen einen bestimmten Einsatz einen Tippschein abgeben, der für jede Ziehung die Farbe der gezogenen Kugel vorhersagen soll. Der Gewinn, den ich erhalte, ist c mal Anzahl der richtigen Tips. Welches Ausfüllen des Tippscheins ist für mich optimal in dem Sinn, daß der erwartete Gewinn maximal wird?

Antwort: 10-mal w . Begründung: Wenn X(i) der Indikator des Ereignisses "richtige Vorhersage der i-ten Ziehung" ist, so ist der erwartete Gewinn $c\cdot\sum_i EX(i)$. Der i-te Term in dieser Summe ist $c\cdot 7/12$ bzw. $c\cdot 5/12$, je nachdem ob im Tippschein an der i-ten Stelle ein w oder ein s steht. Die Summe wird maximal, wenn jeder Summand den maximal möglichen Wert hat, und das geschieht mit der Tippfolge (w,w, ...,w) .

Bemerkenswert an diesem simplen Beispiel ist, daß dieses Ergebnis sowohl für das Ziehen mit wie für das ohne Zurücklegen richtig ist: es kommt nur auf die Erwartung jedes einzelnen X(i) separat an, nicht auf die gemeinsame Verteilung der X(i) . Es schadet der Optimalität des Tippscheins (w,w, .. , w) überhaupt nicht, daß man beim Ziehen ohne Zurücklegen ga-rantiert mindestens 3 Falsche auf dem Zettel stehen hat.

Wir ändern die Gewinnregel: ich gewinne einen bestimmten Betrag c, wenn ich 10 Richtige habe; wenn nicht, gewinne ich gar nichts. Welcher Tippschein ist jetzt optimal?

Es ist eine Folge x aus Symbolen s und w zu finden, für welche $P(X(i) = x(i),\ 1\le i\le n)$ maximal wird; wenn ich eine solche auf den Tippschein schreibe, habe ich eine Gewinnerwartung von $c\cdot P(x)$. Im Fall des Ziehens mit Zurücklegen ist das die Folge (w, .. , w) : mit der Abkürzung $z(s) = 5$ und $z(w) = 7$ gilt, daß für ein beliebiges $x = (x(1),.. , x(10))$ die An-zahl der auf x führenden Fälle gleich $\prod_i z(x(i))$ ist. Beim Ziehen ohne Zurücklegen wird eine Folge maximaler Ws. anders aussehen: zuerst nur w , soloange bis in der Urne gleich viele Kugeln jeder Farbe sind, dann abwechselnd w und s . Aus Symmetriegründen ist jede durch Permutation daraus hervorgehende Folge ebenso gut. Anwort also im Fall ohne Zurücklegen: jede Folge mit 6-mal w und 4-mal s ist optimal.

Def. 1.4. Wenn B Ereignis ist, $B \neq \varnothing$, dann nennt man die Zuordnung

(10) $$A \to P(A \mid B) := P(A \cap B) / P(B)$$

die bedingte Ws. unter der Bedingung B .

Bemerkung. Man kann streiten, ob es besser ist, die bedingte Ws. unter B als eine Ws. zur Ergebnismenge B aufzufassen - sie ist in diesem Fall natürlich wieder eine Gleichverteilung - oder als eine Ws. zur alten Ergebnismenge E, wobei Ergebnisse, die mit B nicht verträglich sind, das Gewicht null bekommen. Wir empfehlen den zweiten Standpunkt: man kann dann z.B. im Raum der W.-Bewertungen konvexe Mittel bilden, da alle den gleichen Defini-tionsbereich haben. Das schließt nicht aus, daß die andere Deutung mit einer reduzierten Er-gebnismenge in vielen Situationen ebenfalls sehr bequem sein kann.

Beispiel 6. Kartenspiel.

Im Schaffkopf gibt es bekanntlich 4 Spieler und 32 Karten, davon 4 Unter; jeder Spieler erhält 8 Karten. Was ist unter der Bedingung, daß ein bestimmter Spieler 2 Unter hat, die Ws. dafür, daß die restlichen 2 Unter in einer Hand sind?

Am besten nimmt man als Bedingung B das Ereignis, daß der bestimmte Spieler (nennen wir ihn Spieler 1) ein ganz bestimmtes Blatt von 8 Karten hat, unter denen sich auch 2 Unter befin-den. Dann ist B eineindeutig zugeordnet der Menge aller Aufteilungen der restlichen 24 Karten auf die restlichen 3 Spieler. Die Ws., daß Spieler 2 die restlichen 2 Unter hat, ist nach dem Ar-gument von Beispiel 2 gleich $C(24,8)^{-1}\, C(22,6)\, C(2,2) = 7\cdot 8/23\cdot 24$. Die gleiche Ws. gibt es für Spieler 3 und 4 , und da die Ereignisse "Spieler i hat 2 Unter" für $i = 2, 3, 4$ disjunkt sind, ist die gesuchte Ws. die Summe dieser drei Ws.en, also gleich $7/23 = 0.3043$.

Hier haben wir den Standpunkt von Spieler 1 eingenommen, insofern als wir sein Blatt als nicht zufällig angesehen haben und den Zufall nur noch über die Anordnung der restlichen 24 Karten haben verfügen lassen. Denn genau so wird Spieler 1 rechnen, wenn er sich fragt, ob er eine be-stimmte Reizung riskieren soll. In Strenge genommen war jedoch die Bedingung eine anderes, viel größeres Ereignis. Man überlegt sich nun die Formel

Prop. 1.2. ("Satz von der totalen Wahrscheinlichkeit") Sei $E = \cup B(i)$ mit disjunkten $B(i)$; dann gilt für jedes A

(11) $$P(A) = \sum_i P(B(i))\, P(A \mid B(i)) \quad .$$

Beweis : $A = \cup (A \cap B(i))$ (disjunkte Mengen) , $P(A \cap B(i)) = P(A|B(i))\, P(B(i))$.

und modifiziert sie für den Fall, daß $\cup B(i) = B$ nicht notwendig gleich E ist, zu

(11a) $$P(A \mid B) = \sum_i P(B(i) \mid B)\, P(A \mid B(i)) \quad .$$

Man sollte (11) bzw. (11a) so lesen, daß man $P(\cdot)$ bzw. $P(\cdot \mid B)$ als konvexes Mittel der Ws.-Bewertungen $P(\cdot \mid B(i))$ mit Gewichten $P(B(i))$ bzw. $P(B(i) \mid B)$, die sich zu 1 summieren, auffaßt.

Dies wendet man an auf $B = \{$Spieler 1 hat 2 Unter$\}$ und $B(i) = \{$Spieler 1 hat das Blatt i$\}$, wobei durch i eine Liste aller mit B verträglichen Blätter von Spieler 1 indiziert wird. Wir haben oben die bedingte Ws. unter einer beliebigen Bedingung B(i) aus dieser Liste berechnet; ihr Wert war 7/23 . Wenn alle bedingten Ws.en $Ws(\cdot \mid B(i))$ untereinander übereinstimmen, dann stimmen sie auch mit $Ws(\cdot \mid B)$ überein, vermöge (11a).

Andere Argumentation für die Aufgabe von Beispiel 5.
Wir bedingen wieder auf das Blatt, das Spieler 1 in der Hand hält. Spieler 1 weiß, daß 2 Unter bei den Gegnern stehen, sagen wir Herz- und Schellunter. Derjenige Spieler, der den Herzunter hat, wird als Spieler 2 definiert. Jetzt ist noch der Schellunter zu vergeben. Nachdem Spieler 2 bereits eine Karte definitionsgemäß hat, sind bei ihm noch 7 Plätze frei, bei den beiden anderen Spielern aber je 8 . Nach Laplace ist somit 7/23 die Ws. dafür, daß der Schell-unter ebenfalls an Spieler 2 geht.

Beispiel 7. Verbale Fallstricke.

Vor längerer Zeit stand einmal in einem Schulbuch folgende Aufgabe zu lesen: Zwei perfekte Würfel werden gleichzeitig geworfen; man interessiert sich für das Ereignis, daß zwei gleiche Zahlen erscheinen; wie groß ist dessen Ws. unter der Bedingung, daß einer der beiden Würfel die Zahl 6 zeigt?

Die Voraussetzungen sind klar: "perfekt" meint, daß alle 36 möglichen Zahlenpaare (x,y) die gleiche Ws. haben. Die Lösung, welche die Autoren jenes Buchs vorschlugen, lautete: 1/11 . Ihre Begründung war, daß man das bedingende Ereignis "einer der Würfel zeigt die Zahl 6" mathematisch als $\{(x,y) : x = 6 \text{ oder } y = 6\}$ interpretieren müsse; jenes Ereignis bestehe aus 11 Paaren, sein Durchschnitt mit $\{(x,y) : x = y\}$ enthalte genau ein Paar, nämlich (6,6) .

Ein unverbildeter Mensch hingegen gibt spontan die Antwort: 1/6 ; warum sollte die Vorabinformation "ein Sechser" die Chancen für "zwei gleiche" mindern? Er kann aber, auch wenn er sich bei der Lösung der Autoren unwohl fühlt, schlecht gegen ihre Argumente ankommen. Was ist der Grund für das ungute Gefühl, welches die Autoren beim Leser hinterlassen? Nun, wenn man W.-Theorie so versteht, daß in ihr nicht nur logische Spiele getrieben werden, sondern daß hinter ihren Begriffsbildungen und Rechenexempeln noch irgendeine Andeutung von - notwendigerweise idealisierter - Realität steht oder stehen sollte, wird man nach einer Versuchsanord-

nung fragen, die dem Würfelexperiment und dem Bedingen zugrundeliegen könnte. Ein plausibles Szenarium könnte so aussehen, daß die zwei Würfel auf einem Tisch gleichzeitig aus dem Becher gekippt werden und einer von ihnen auf den Boden herunterfällt und unseren Blicken entzogen wird, während der auf dem Tisch gebliebene Würfel die Zahl 6 zeigt. Wir warten auf das Ereignis "zwei gleiche" und fragen bei diesem Wissensstand nach dem Grad unserer Hoffnung, daß der noch unsichtbare Würfel ebenfalls 6 zeigt. In dieser Interpretation ist die Antwort klar: 1/6 . Die Lösung der Autoren hingegen gibt "physika-lisch" keinen rechten Sinn; man kann ihre Bedingung $\{(x,y) : x = 6 \text{ oder } y = 6\}$ nur realisie-ren, indem man einen allwissenden Dämon bemüht, der durch eine willkürliche Vorabinfor-mation die Zahl der möglichen Fälle reduziert. Was lernen wir aus dem mißglückten Rechen-exempel? Ich meine, wir lernen, daß die sprachliche Exegese eines Aufgabentexts unter le-diglich formal-logischen Gesichtspunkten als Übungsmaterial deswegen schlecht ist, weil sie den Gesichtspunkt, daß Mathematik auch eine Realität abbildet (und genau deswegen treibt man Stochastik in der Schule), vollständig ausblendet. Für Schüler, denen man stochastische Denkweisen nahebringen will, ist das Rechenexempel (genauer: seine Lösung) daher sehr un-glücklich gewählt.

Wir erfahren andererseits aus dem mißglückten Beispiel positiv einiges über "das Wesen" der bedingten Wsen. Die Situation von Prop. 1.2 ist typisch für dieses Wesen: man hat es nicht mit dem isolierten Ausrechnen einer W.-Bewertung $P(. \mid B)$ für irgendein B , sondern mit einer Partition der Ergebnismenge $E = \cup B(i)$ zu tun; man soll für jedes mögliche $B(i)$ die bedingte Ws. unter $B(i)$ bereithalten. Dahinter steht die Interpretation, daß $B(i)$ in einem ersten gröberen Versuch ermittelt wird und daß anschließend, darauf aufbauend, die genaue Spezifikation des Elements x der Ergebnismenge in einem zweiten Schritt erfolgt. (Siehe dazu das nächste Kapitel). Zu unserem Beispiel: Selbst wenn wir die Existenz eines allwissenden Dämons akzeptierten, die Vorabinformation, die er uns gab, war willkürlich. Denn welche anderen Dinge hätte er uns sagen können an Stelle von "Ein Würfel zeigt 6"? Wenn etwa das Ergebnis (6,4) ist, hätte er dann gesagt "Einer zeigt 4" oder "Einer zeigt 6"? Und was würde er sagen, wenn das Ergebnis (4,3) ist, etc.? Das heißt, die Aufgabe ist erst dann voll verstanden und daher sinnvoll zu lösen, wenn man die gesamte Strategie des Dämons kennt und daher weiß, welche Ergebnisse mit seinen Worten "Einer zeigt 6" verträglich sind.

Beispiel 8. Zwischenräume.

Wir nehmen ein gut gemischtes Kartenspiel von 52 Karten und legen es ringförmig angeord-net auf den Tisch. Wir zeichnen die 4 Könige aus ; sie bekommen die Nummern 1 , .. , 4 (von Pik bis Kreuz). Durch die 4 Könige wird die Menge der restlichen 48 Karten in 4 Intervalle zerteilt (welche auch leer sein können: wenn zwei Könige nebeneinander liegen). Wir definieren $L(i)$, $1 \leq i \leq 4$, als die Länge desjenigen Intervalls, das sich im Uhrzeigersinn an den i-ten König an-

schließt. Wie ist L(i) verteilt und was ist seine Erwartung?

Die Antwort auf die zweite Frage ist leicht: 12 , aus Symmetriegründen. Die erste ist auch nicht zu schwer: jede Anordnung der 4 Könige ist gleich ws. Davon gibt es 52·51·50·49 mög-liche Fälle. Die für das Ereignis {L(1) = k} günstigen Fälle bestehen aus 52 Möglichkeiten, den Platz von König 1 zu bestimmen, 3 Möglichkeiten, die Nummer des auf ihn folgenden Königs zu wählen, und dann noch (50-k)(49-k) Möglichkeiten, die restlichen 2 Könige auf die verfügbaren Plätze zu setzen. Das heißt

(12) $P(L(1) = k) = (51\cdot 50\cdot 49)^{-1}\cdot 3\cdot(50-k)(49-k) = 17^{-1}\,(1 - k/49)(1 - k/50)$, für $0 \le k \le 48$.

Wir hätten die Antwort auf die erste Frage übrigens auch über Formel (7) finden können: mögli-che Fälle sind alle Anordnungen von 48 ununterscheidbaren Kugeln in 4 Schachteln; günstige Fälle für {L(1) = k} gibt es soviele wie Anordnungen von 48-k Kugeln in 3 Schachteln.

Wir numerieren jetzt die Plätze auf dem Tisch im Uhrzeigersinn. Wir durchwandern die Plät-ze von 1 bis 52 und definieren K(1) als Zahl der besuchten Plätze, bevor wir das erste Mal einen König besuchen, K(2) als Zahl der Plätze zwischen dem ersten und zweiten besuchten König usw., und schließlich K(5) als die Zahl der Plätze nach dem letzten König und vor der Wieder-kehr nach Platz 1. Dann ist die Erwartung eines jeden K(i) gleich 48/5 = 9.6 , aus Symmetrie, und das Tupel L = (L(1), L(2), L(3), L(4)) ist bis auf Permutation identisch mit K := (K(2), K(3), K(4), K(5)+K(1)) . Das Tupel der Erwartungen von L ist (12,12,12, 12), das der Erwar-tungen von K aber (9.6, 9.6, 9.6, 19.2). Wie ist das vereinbar?

Antwort: Das "Aufschneiden" des Rings, d.h. die Wahl einer Schnittstelle zwischen Platz 1 und 52, ist die Ursache dafür, daß K verschieden von L ist. Wenn wir den Ring nach dem Legen der Karten an einer zufälligen (nach der Gleichverteilung gewählten) Stelle aufschnei-den und die K(i) in Bezug auf diese willkürlich gewählte Schnittstelle definieren, d.h. als Platz Nr.1 den ersten auf den Schnitt folgenden Platz bezeichnen, bleibt deren Verteilung die selbe, wie wenn wir die Numerierung der Plätze vor dem Legen der Karten durchführen. Die bedingte Ws., bei gegebenem L , daß die Schnittstelle in das i-te Intervall hineinfällt, ist aber nicht gleich für alle i , sondern proportional zu L(i)+1 . Daher wird dasjenige Intervall, welches die Schnittstelle enthält, gegenüber den anderen "begünstigt". Bei unserem konkrten Problem ergibt sich, daß es im Mittel doppelt so lang ist wie die "nicht begünstigten".

Phänomene wie das hier beschriebene werden als Inspektionsparadoxa bezeichnet. Sie treten auf, wenn die Art der Auswahl (der "Inspektion") eines Intervalls oder eines Blocks in einer Partition die Verteilung der Größe des ausgewählten Objekts beeinflußt. Dazu auch

Beispiel 9. Zyklenlänge in einer zufälligen Permutation.

Eine Menge G der Mächtigkeit n werde zufällig permutiert. Die ausgewählte Permutation π induziert eine Partition von G in Zyklen. (Zwei Elemente x und y von G gehören zum gleichen Zyklus, wenn $\pi^m(x) = y$ für ein natürliches m gilt.) Die Zufallsvariable L sei die Länge eines Zyklus. Wie ist L verteilt?

Die Frage ist in dieser Form unbestimmt, da nicht gesagt wird, welchen Zyklus man bei gege-benem π nehmen soll. Hier sind zwei Vorschläge einer möglichen Präzisierung.
(a) Man wählt aus der Menge der endlich vielen Zyklen zufällig einen aus: Wenn G unter π in m Zyklen zerfällt (diese Zahl ist ebenfalls zufällig), so betätigt man einen neuen Zufalls-generator, der mit Ws. 1/m einen Zyklus herausgreift, und definiert L als dessen Länge.
(b) Man wählt zufällig, nach der Gleichverteilung, ein Element a von G und definiert L als die Länge desjenigen Zyklus, der a enthält.

Die Deutung (b) hat den Vorzug der rechnerischen Einfachheit. Es gilt nämlich

(13) $\qquad P(L = k) = 1/n \quad$ für $1 \le k \le n$.

Ganz gleich, welches a gewählt wird, dieses a gehört genau dann zu einem Zyklus der Län-ge k , wenn $\pi^k(a) = a$ und $\pi^j(a) \neq a$ für $j < k$. Wieviele hierfür günstige Permutationen gibt es? Für die Bahn von a , die Folge $\pi(a), \pi^2(a), .. , \pi^k(a) = a,$ gibt es (n-1) .. (n-k+1) Möglich-keiten, dazu noch (n-k)! für das was außerhalb der Bahn von a geschieht. Man be-kommt insge-samt (n-1)! Möglichkeiten, also ist (n-1)!/n! = 1/n die gesuchte Ws. einer Bahn-länge k .

Die Rechnung zu Deutung (a) ist wesentlich mühsamer und häßlicher. Wir betrachten den Fall n = 4 . Es gibt die folgenden Zyklenlängen (Young-Tableaux) mit respektiven Wsen

(14) (4) (3 , 1) (2 , 2) (2 , 1 , 1) (1 , 1 , 1 , 1)

6/24 8/24 3/24 6/24 1/24 .

Die Ws., daß der zufällig herausgegriffene Zyklus die Länge k hat, ist daher für

(15) k=1 $24^{-1}(8/2 + 6\cdot 2/3 + 1) = 9/24$,

k=2 $24^{-1}(3 + 6/3) = 5/24$,

k=3 $24^{-1} \cdot 8/2 = 4/24$,

k=4 $\quad 24^{-1} \cdot 6 \quad = 6/24$.

Der Vergleich der rechten Spalte in (15) mit der Gleichverteilung (6/24, .. ,6/24) zeigt, daß Deutung b) gegenüber Deutung a) größere Werte von k begünstigt, einmal abgesehen von $k = 4$. Das kann man leicht verstehen: Wir stellen uns vor, daß zuerst die Permutation π und dann das Element a bestimmt wird. Unter der Bedingung, daß π das Young-Tableau (n(1) , .., n(m)) von Zyklenlängen besitzt, ist die Ws. dafür, daß der i-te dieser Zyklen herausge-griffen wird, nicht 1/m wie in Deutung a) , sondern n(i)/n : die zweite W.-Bewertung be-günstigt im Vergleich zur ersten die längeren Zyklen. (Der Fall $L = n (= 4)$ kann nur auftre-ten, wenn es nur einen einzigen Zyklus gibt und damit auch keine Freiheit bei der Auswahl.)

Betrachten wir noch die Mittelwerte von L in beiden Varianten. Im Fall a) hat man für $n = 4$ den speziellen Zahlenwert 55/24 , im Fall b) generell (n+1)/2 , also 60/24 für $n = 4$. Wir definieren die Zufallsvariable M als Zahl der Zyklen der Permutation π . Wir führen noch einige Rechnungen im Modell a) aus. Unter der Bedingung, daß π gegeben ist, hat man

(16a) $\qquad E(L \mid \pi) = n/M(\pi)$, $\qquad$ und daher unbedingt

(16b) $\qquad E\,L = (n!)^{-1} \sum_{\pi} E(L \mid \pi) = n\,E(1/M)$.

Wenn man den Mittelwert von M aus (13) berechnet, erhält man

(17) $\qquad E\,M = 50/24.$

Wir stellen empirisch fest, daß das Produkt der Zahlen in (16b) (zu $n = 4$) und (17), also 55/24 und 50/24 , nicht gleich 4 , sondern größer ist. Allgemein gilt für beliebige positive Zufallsvariable M die Ungleichung $EM \cdot E(1/M) \geq 1$; sie ist ein Spezialfall der sog. Unglei-chung von Jensen (s. Anhang). Wir erwähnen das, um vor leichtfertigem Umgang mit dem Begriff des Mittelwerts zu warnen: Mittelwertbildung ist nicht vertauschbar mit nichtlinearen Operationen im Raum der Zufallsvariablen bzw. der Mittelwerte.

Wir schließen diesen Paragraphen mit zwei klassischen Beispielen, in denen die Gleichverteilung durch eine kontinuierliche Symmetriegruppe festgelegt wird.

Beispiel 10. Buffonsches Nadelproblem.

Wir haben auf einem ebenen Tisch parallele Gerade im Abstand 1 gezeichnet und werfen von weit oben eine Nadel der Länge r auf den Tisch. Mit welcher Ws. wird die Nadel eine der Ge-

raden kreuzen?

Die Lage der Nadel ist durch die Position (x,z) ihres Kopfes und die Richtung φ , in die sie zeigt, bestimmt. Dabei rangieren x und z zunächst von $-\infty$ bis $+\infty$ und φ von 0 bis 2π . Eine Gleichverteilung auf einem unendlichen Raum zu finden ist i.a. unmöglich; für unsere Frage ist aber nur die z-Koordinate (die Geraden mögen in x-Richtung gehen) von Interesse,

und von der brauchen wir nur $y := z \bmod 1$ zu kennen. Wir setzen daher folgende W.-Verteilung an

(18) $$P(Y \in dy, \Phi \in d\varphi) = (2\pi)^{-1} dy\, d\varphi \quad , 0 < y < 1 , 0 < \varphi < 2\pi .$$

(Die Großbuchstaben bezeichnen Zufallsvariablen im Unterschied zu den Elementen der Ergebnismenge.) Die Rechtfertigung für diesen Ansatz ist seine Invarianz unter der Gruppe der euklidischen Abbildungen der Ebene; er ist der einzige, der unter Rotationen und Translationen invariant ist. Die Antwort ist jetzt leicht:

(19) $$P(\text{eine Gerade wird gekreuzt}) = (2\pi)^{-1} \int\int dy\, d\varphi \, (1_{\{y + r\sin\varphi < 0\}} + 1_{\{y + r\sin\varphi > 1\}}) =$$

$$= \pi^{-1} \int\int dy\, d\varphi \, 1_{\{y + r\sin\varphi > 1\}} .$$

Im Spezialfall $r = 1$ ergibt das $\pi^{-1} \int_{0<\varphi<\pi} d\varphi \, \sin\varphi = 2/\pi = 0.63662$. Wenn man will, kann man hierauf eine empirische (stochastische) Messung von π aufbauen; man zählt die relative Häufigkeit des Kreuzens einer Geraden im Lauf einer langen Versuchsserie.

Beispiel 11. Zufällige Gerade.

Gegeben sei ein Kreis. Wie groß ist die Ws. daß eine zufällige Sehne länger ist als die Seite des

dem Kreis einbeschriebenen regulären Dreiecks?

Diese Frage ist klassisch, insofern als sie zeigt, daß verschiedene Begriffe von "zufälliger Sehne" zu verschiedenen Antworten führen. In [P] oder [K] werden einige solcher Begriffe vorgestellt und die Verschiedenheit der Ergebnisse als ein Paradoxon präsentiert. Dessen Auflö-

sung mag in der Geschichte der Stochastik ein ernstes Problem dargestellt haben. Wir wollen uns aber nicht zu sehr mit früheren Irrwegen oder Scheinproblemen aufhalten; daher werden wir von den vorgeschlagenen Lösungen nur die physikalisch plausible vorstellen.

Wir wählen den Mittelpunkt des Kreises als Ursprung des Koordinatensystems und den Kreisradius gleich 1 . Eine Gerade ist beschrieben durch die Gleichung $x\cos\varphi + y\sin\varphi = r$ und daher parametrisierbar durch $r > 0$ und $0 < \varphi < 2\pi$. Das einzige euklidisch invariante Volumenelement ist $dr\,d\varphi$. Sein Gesamtvolumen ist unendlich. Das stört aber nicht, da wir nur den Teil des (r,φ)-Raums betrachten müssen, in welchem die den Kreis schneidenden Geraden liegen, d.h. den Bereich $\{0 < r < 1 , 0 < \varphi < 2\pi\}$. Die Geraden, deren im Kreis verlaufender Abschnitt länger als die Seite des regulären Dreiecks ist, sind offenbar genau die zu einem Parameterwert (r, φ) mit $r < 1/2$ gehörenden. Also ist die gesuchte Ws. gleich 1/2 .

Anhang. Die Ungleichung von Jensen.

Sei A konvexe Teilmenge von R^d und $x(i) \in A$ für $i = 1, .. , n$; seien $p(i)$, $i = 1 , ... , n$, positive Zahlen mit Summe 1 ; sei F eine konvexe Funktion mit Definitionsbereich A . Dann gilt

(J) $$F(\sum p(i)\cdot x(i)) \le \sum p(i)\cdot F(x(i)) .$$

Beweis. Setze $x = \sum p(i)\cdot x(i)$. Es gibt eine affine Funktion G mit den Eigenschaften:

$$G \le F \quad \text{und} \quad G(x) = F(x) .$$

(Das ist eine Folge von Hahn-Banach.) Für die affine Funktion G ist die "Schwerpunktbildung" vertauschbar mit Anwendung von G , d.h. man hat

$$G(x) = G(\sum p(i)\cdot x(i)) = \sum p(i)\cdot G(x(i)) .$$

Hierin ist die linke Seite gleich $F(x)$ und die rechte Seite majorisiert durch $\sum p(i)\cdot F(x(i))$. Das ist genau die Behauptung.

Die oben erwähnte Ungleichung $EM\cdot E(1/M) \ge 1$ für $M > 0$ folgt aus (J), wenn man als A die positive Halbgerade und als F die Funktion $x \to 1/x$ wählt, welche konvex ist. (Und als $p(i)$ die Ws. des Ereignisses $\{M = x(i)\}$, für alle möglichen Werte $x(i)$ von M .)

Literatur.

[K] U. Krengel: Einführung in die Wahrscheinlichkeitstheorie und Statistik. Braunschweig: Vieweg 1988.

[L] P.S. Marquis de Laplace: Théorie analytique des Probabilités. Oeuvres, Tome VII. Gauthier-Villars: Paris 1886. (Nachdruck der 3. Auflage von 1820).

[P] J. Pfanzagl: Elementare Wahrscheinlichkeitsrechnung. Berlin: de Gruyter 1988. (dort vor allem: Abschnitt 5. Geometrische Wahrscheinlichkeiten.)

[PSz] G.Pólya, G. Szegö: Aufgaben und Lehrsätze aus der Analysis. 3. Aufl. Berlin: Springer 1970.

Kapitel 2. Wahrscheinlichkeitsräume

Inhalt: Grundbegriffe in allgemeinerer Form, ohne Gleichverteilungsannahme; eine zentrale Rolle spielen: Unabhängigkeit, bedingte Verteilung und bedingte Ws., letztere auch als Kon-struktionsprinzip. Momente, Varianz, Kovarianz, schwaches Gesetz der großen Zahlen. (Der mathematische Rahmen ist allgemeiner als im letzten Kapitel. Der Hauptunterschied zu Def. 1.1 wird in Def. 2.1 festgehalten; die übrigen Begriffe, wie Verteilung, Zufallsvariable können wörtlich aus Kapitel 1 übernommen werden.)

Def. 2.1. Ein W.-Raum ist eine (zunächst) höchstens abzählbare Menge E , genannt Ergebnismenge, zusammen mit einer W.-Bewertung P , welche jedem $e \in E$ ein Gewicht $P(e)$ zuordnet mit $P(e) \geq 0$ und $\sum_E P(e) = 1$. Ein Ereignis ist eine Teilmenge von E ; die Ws. von A (unter der Bewertung P) ist $\sum_A P(e)$. Ein Träger von P ist jedes $A \subset E$ mit $P(A) = 1$.

Bei allem, was in diesem Paragraphen folgt, wird eine W.-Bewertung P als vorgegeben angenommen; alle Aussagen sind in Bezug auf das jeweilige P zu verstehen.

Def. 2.2. Die bedingte Ws. von A unter der Bedingung B ist $P(A \mid B) := P(A \cap B) / P(B)$. Die bedingte Verteilung einer F-wertigen ZV X unter der Bedingung B ist die Familie $P(X = f \mid B)$, $f \in F$. Wenn F eine Partition von E ist, so ist die bedingte Ws. unter der Bedingung F die Familie aller bedingten Wsen $P(. \mid B)$, wo B die Atome (Blöcke) der Partition F durchläuft.

Zur Illustration des letztgenannten Begriffs erinnern wir an die Ausführungen zu Beispiel 7 ("Verbale Fallstricke") in Kapitel 1. Typischerweise wird man, wenn immer man von einer bedingten Ws. spricht, als mögliche Bedingungen alle Atome einer Partition betrachten.

Def. 2.3. Die ZV X heißt unabhängig von der ZV Y , wenn für alle y die bedingte Verteilung von X unter der Bedingung $\{Y = y\}$ die gleiche ist. (Sie ist dann gleich der unbedingten Verteilung; vgl. Prop. 1.2, Satz von der totalen Ws.)

Bemerkung. Die Familie der Ereignisse $\{Y = y\}$ bildet natürlich eine Partition von E . Man könnte sie die von Y erzeugte Partition nennen. - Weiter, die Unabhängigkeit der ZV X von Y ist äquivalent zu

(1) $P(X = x, Y = y) = P(X = x)\, P(Y = y)$ für alle x , y,

was zeigt, daß Unabhängigkeit eine symmetrische Eigenschaft ist. Wir werden daher einfacher sagen "X und Y sind unabhängig", wenn der Sachverhalt von Def. 2.3 vorliegt.

Def. 2.3a. Eine Familie $X(i)$, $1 \le i \le n$, von ZVen heißt unabhängig, wenn

(2) $P(X(i) = x(i), 1 \le i \le n) = \prod_{1 \le i \le n} P(X(i) = x(i))$ für alle $x(1), .. , x(n)$.

In der Sprache der Partitionen kann man diesen Sachverhalt so fassen: die bedingte Verteilung jedes einzelnen $X(i)$ ist auf allen Atomen der von den übrigen $n-1$ ZVen erzeugten Partition dieselbe.

Eine Standardsituation im stochastischen Modellieren und überaus häufige Vorausetzung in Sätzen der stochastischen Theorie besteht darin, daß man sagt "Die ZVen $X(i)$, $i = 1 .. n$, sind unabhängig; die Verteilung von $X()$ ist $p^{(i)}$ " oder "Die ZVen $X(i)$, $i = 1 .. n$, sind unabhängig identisch verteilt mit Verteilung p ". Vermöge (2) ist klar, daß durch eine solche Angabe die gemeinsame Verteilung der $X(i)$ eindeutig festgelegt und damit alle stochastisch relevante Information über das Modell gegeben ist. Der zweite Fall ist so beliebt, daß er eine eigene Abkürzung bekommen hat: u.i.v. , auf amerikanisch: i.i.d. .

Beispiel 1. Warten auf den ersten Erfolg.

Seien $X(i)$, $i = 1 , .. , n$, unabhhängig und identisch verteilt (u.i.v.) ; es gelte $P(X(i) = 1) = p = 1 - P(X(i) = 0)$, mit $0 < p < 1$. Wir definieren $T := \inf\{t : t \le n \text{ und } X(t) = 1\}$, wenn diese Menge nicht leer ist; andernfalls setzen wir $T = n+1$. Wie ist T verteilt?

Antwort: Offenbar gilt

(3) $P(T = k) = (1 - p)^{k-1} p$ für $1 \le k \le n$; $P(T = n+1) = (1-p)^n$.

Man sieht: wenn n immer größer wird, dann konvergiert die Verteilung in (3) gegen eine Verteilung mit der Eigenschaft

(4) $P(T = k) = (1 - p)^{k-1} p$ für $1 \le k$.

Die Gewichte in (4) addieren sich zu 1 . In diesem Sinn kann man die Voraussetzung eines endlichen n fallenlassen; die Ws., daß T die Begrenzung des Zeithorizonts spürt, d.h. daß die

Menge $\{t : t \le n$ und $X(t) = 1\}$ leer ist, wird beliebig klein für großes n . Man nennt die Verteilung von T in (4) die geometrische Verteilung zum Parameter p . Für die Erwartung von T gilt

$$ET = p \sum_{k\ge 1} (1-p)^{k-1} k = p\, d/dt|_{t=1-p} \sum t^k = p\, d/dt|_{t=1-p} (1 - t)^{-1} = p\, p^{-2} = p^{-1} .$$

Wir machen es nun etwas komplizierter und fragen nach der Wartezeit, bis zwei Einser in Folge kommen, d.h. man betrachtet $S: = \inf \{t : X(t) = X(t-1) = 1\}$. Die Beschränkung auf endliches n haben wir weggelassen, da wir ohnehin wissen, daß S mit Ws. beliebig nahe bei 1 endlich ist; man kann auch so argumentieren, daß S durch das Doppelte einer geometrischen Variablen zum Parameter p^2 majorisiert wird, nämlich $\inf \{t: X(t) = X(t-1) = 1, t \text{ gerade}\}$.
Wir fragen hier nur nach dem Mittelwert m dieser Wartezeit. (Quelle: [E].)

Wir definieren die bedingten Mittelwerte von S , wobei die bedingende Partition die von $X(1)$ erzeugte ist. Sei also

$$m(0): = E(S \mid X(1) = 0) , \qquad m(1):= E(S \mid X(1) = 1) , \qquad \text{sowie } m: = ES .$$

Dann gelten die Beziehungen (mit $q = 1 - p$)

(5) $$m = p\cdot m(1) + q\cdot m(0) ;\ m(1) = 1 + p\cdot 1 + q\cdot(m+1) ;\quad m(0) = 1 + m .$$

Das Gleichungssystem hat die eindeutige Lösung $m(1) = 2p + 2q/p + (q/p)^2$, $m = 2/p + q/p^2$, $m(0) = 1 + 2/p + q/p^2$. Wir haben also gefunden :

(6) $$ES = 2/p + q/p^2 = (1 + p)/p^2 .$$

Die analoge Rechnung für die Wartezeit R bis zum ersten Eintreten des Musters (0,1) ergibt

(7) $$ER = 2 + p/q + q/p .$$

Hieran ist verblüffend, daß im Fall $p = 1/2$ die beiden Erwartungen verschieden sind: $ES = 6$ und $ER = 4$, in scheinbarem Widerspruch dazu, daß beide Muster die gleiche Ws. haben beobachtet zu werden, nämlich $1/4$. Der Leser ist gefragt,wie er sich das Phänomen erklärt.

Beispiel 2. Besetzungszahlen lokal und global.

Es werden n ununterscheidbare Kugeln auf m Schachteln verteilt. Wie ist $X(i)$, die Anzahl der Kugeln in der i-ten Schachtel, verteilt?

Das ist eine unmittelbare Folge von (1.7) . Wir erhalten

(8) $$P(X(1) = k) = C(n+m-1,m-1)^{-1}\, C(n+m-k-2,m-2) =$$

$$((n+m-1)\,..(n+m-k-1))^{-1}\cdot(m-1)\cdot n\,..\,\cdot(n-k+1).$$

Wir interessieren uns für das Grenzverhalten dieses Ausdrucks, wenn n und m nach unendlich gehen, und zwar so, daß n/m , die mittlere Zahl der Kugeln pro Schachtel, gegen einen Limes a mit $0 < a < \infty$ geht. Das Ergebnis ist

(9) $$\lim_{n/m\to a} P(X(1) = k) = a^k\,(1+a)^{-k-1} = (1+a)^{-1}(a/(1+a))^k \quad \text{für } k \geq 0\ .$$

Das bedeutet, daß $X(1)+1$ geometrisch verteilt ist zum Parameter $(1+a)^{-1}$. (Man könnte jetzt lange darüber streiten, ob man die Verteilung von $X(1)$ oder die von $X(1)+1$ als geometrische bezeichnen soll, d.h. ob das untere Ende einer geometrischen Verteilung besser null oder eins ist.) Die Erwartung der Grenzverteilung ist dann $1 + a - 1 = a$, ist also gleich dem Limes der Erwartungen von $X(1)$.

Wir fragen nun danach, wie sich im nämlichen Limes von $n/m\to a$ und $n,m\to\infty$ die gemeinsame Verteilung von $X(1)$ und $X(2)$ verhält. Es gibt den direkten Weg: man hat

(10) $$P(X(1) = j\ ,\ X(2) = k) = C(n+m-1,m-1)^{-1}\, C(n+m-k-j-3,m-3)$$

und führt hierin den Grenzübergang aus. Der andere Lösungsweg ist, zu sagen, daß für jede Wahl von j die bedingte Verteilung von $X(2)$ unter der Bedingung $X(1) = j$ wie in (8) ist, aber mit $n-j$ statt n und $m-1$ statt m . Für jedes j erfüllt die neue Parameterfolge die obigen Voraussetzungen, also konvergieren die bedingten Verteilungen für jedes j gegen die Verteilung (9). Da auch $P(X(1) = j)$ konvergiert, wie bekannt, hat man

(11) $$\lim_{n/m\to a} P(X(1) = k\ ,\ X(2) = j) = (1+a)^{-2}\,(a/(1+a))^{k+j}\ , \qquad \text{für } k\ ,\ j \geq 0\ .$$

Das bedeutet, daß $X(1)$ und $X(2)$ in diesem Limes unabhängig werden. Natürlich ist dasselbe Argument für jede endliche Zahl von Schachteln anwendbar, nicht nur zwei: die Besetzungszahlen von verschiedenen Schachteln werden im Limes unabhängig.

Der Zusammenhang, den die bisherigen Rechnungen zwischen den ununterscheidbaren Kugeln und der geometrischen Verteilung (besser: Klasse der geometrischen Verteilungen) herstellen, soll noch weiter ausgeführt werden. Es gilt nämlich die Aussage:

(12) Seien $X(i)$, $1 \le i \le m$, u.i.v. mit $P(X(i) = k) = p(1-p)^k$ für $k \ge 0$. Dann stimmt die bedingte Verteilung von $X = (X(i))$ unter der Bedingung $\{\sum X(i) = n\}$ überein mit der Bose-Einsteinschen Statistik für die Besetzungszahlen zu n Kugeln und m Plätzen.

Dies folgt sofort aus der Formel $P(X(i) = n(i), 1 \le i \le m) = p^m (1-p)^{\sum n(i)}$, welcher Ausdruck auf der Menge $\{\sum X(i) = n\}$ konstant ist. Die bedingte Verteilung ist daher die Gleichverteilung auf der Menge aller Tupel $(n(i))$ mit Summe n . Man beachte, daß diese bedingte Verteilung für alle Werte des Parameters p dieselbe ist.
Wir haben bisher von bedingter Ws./Verteilung gesprochen als einer von einer gegebenen W.-Bewertung P abgeleiteten Größe. Mindestens genau so wichtig ist aber ein anderer Aspekt: bedingte Verteilungen als eine Möglichkeit, die W.-Bewertung zu definieren. Wir gehen zurück zu Def. 2.1 und Formel (1.11):

(13) $$P(\,.\,) = \sum P(B(i)) \cdot P(\,.\,|\,B(i))\,,$$

wo $(B(i))$ eine Partition der Ergebnismenge ist. Wenn nun die W.-Bewertung P zunächst nur für die Elemente $(B(i))$ einer Partition gegeben ist und weiter für jedes i die bedingte W.-Bewertung $P(\,.\,|\,B(i))$ für alle Teilmengen von E , so läßt sich P vermöge (13) für alle Ereignisse definieren. Die einzigen Restriktionen bei diesem Verfahren sind Gegenstand von

Prop. 2.1. Sei $(B(i))$ eine Partition von E . Zu positiven Gewichten $p(i)$ und einer Familie $(Q(i))$ von W.-Bewertungen auf E definiert man $P = \sum p(i) \cdot Q(i)$. Dieses P ist genau dann eine W.-Bewertung auf E , wenn $\sum p(i) = 1$; es gilt $P(\,.\,|\,B(i)) = Q(i)$ für alle i genau dann, wenn $Q(i)(B(i)) = 1$ für alle i gilt.

Der Beweis ist trivial. Man muß nur etwas aufpassen, wenn $p(i) = 0$ für ein i gilt. In diesem Fall hat man $P(B(i)) = 0$ und die bedingte Ws. zu P unter der Bedingung $B(i)$ ist nicht definiert. Wenn wir aber allgemein die Konvention treffen, daß im Fall $P(B) = 0$ jede W.Bewertung mit Träger B als bedingte Verteilung von P unter B angesehen werden kann, bleibt der Satz von der totalen Ws. gültig, und nur auf den kommt es an. Weiter müssen wir der zweiten Aussage von Prop. 2.1 die Interpretation geben, daß sie nur dann als falsch anzusehen sei, wenn es ein i gibt mit $P(B(i)) > 0$ und $P(\,.\,|\,B(i)) \neq Q(i)$.

Man kann auch durch mehrstufiges Bedingen eine W.-Bewertung aufbauen: auf jeder Stufe werden die Elemente B der Partition, auf welcher die W.-Bewertung bereits erklärt ist, weiter zerlegt, $B = B(j)$; die bedingte W.-Bewertung $P(\,.\,|\,B)$ sagt, in welchem Verhältnis man die Gewichte innerhalb B auf die einzelnen $B(j)$ aufzuteilen hat. Dies führt zu folgender Verallgemeinerung der Proposition:

Prop. 2.2. Sei gegeben ein Baum endlicher Höhe n . Die Menge der Blätter des Baumes heiße E . Für jedes h mit $0 \le h \le n$ sei $F(h)$ diejenige Partition von E , die entsteht, wenn man alle solchen Blätter zusammenfaßt, deren von der Wurzel kommende Wege sich erst in der Höhe h oder darüber trennen. Dann ist $F(0)$ trivial, und wir setzen mit $P(E) = 1$ eine W.-Bewertung auf $F(0)$ fest. An jedem Knoten x , der noch kein Blatt ist, sei eine W.-Bewertung $Q(x, .)$ auf der Menge seiner unmittelbaren Nachfolger gegeben. Setzt man für ein Blatt x in der Höhe $h(x)$

$$P(x) = \prod_{0 \le j < h(x)} Q(x(j) , x(j+1)) , \tag{14}$$

wobei $(x(0), x(1), .., x)$ der Weg von der Wurzel $x(0)$ nach x ist, so definiert P eine W.-Bewertung auf E . Für jedes $h \le n$ ist

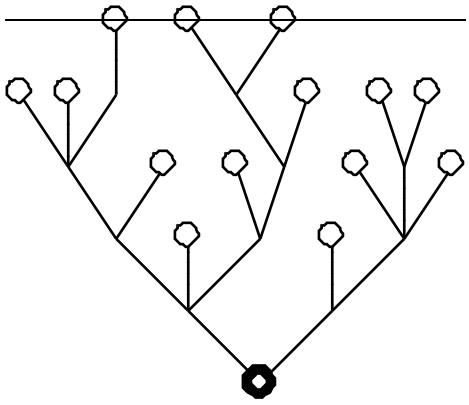
$$P([x]_h) := \prod_{0 \le j < \min(h(x),h)} Q(x(j) , x(j+1)) \tag{15}$$

die Einschränkung von P auf $F(h)$, wobei $[x]_h$ dasjenige Element der Partition $F(h)$ bezeichnet, welches x enthält. Weiter ist für jedes $h \le n$ die bedingte Ws. zu P bedingt auf $F(h)$ gegeben durch

$$P(x \mid B) = \prod_{h \le j < h(x)} Q(x(j) , x(j+1)) \quad (\text{wenn } x \in B \text{ bzw. } 0 \text{ sonst}) . \tag{16}$$

(Das Produkt über die leere Indexmenge ist stets gleich 1 .)

Beweis. Ist trivial, würde aber viele Worte benötigen. Wird daher weggelassen und der Sachverhalt stattdessen durch eine Zeichnung illustriert.



Ein binärer Baum. Dicker Kreis: Wurzel; dünne Kreise: Blätter. Die Höhe ist 5 . Die Partitionen $F(0)$ bis $F(5)$ bestehen aus resp. 1 , 2 , 5 , 9 , 13 , 14 Blöcken.

Beispiel 3. Urnenschema von Pólya.

Sei gegeben eine Urne mit w weißen und s schwarzen Kugeln. Man zieht eine Kugel, legt anschließend die gezogene Kugel zurück und fügt eine neue Kugel der gleichen Farbe hinzu. Dann zieht man wieder und legt eine neue Kugel von der Farbe der eben gezogenen Kugel in die Urne, insgesamt n-mal. Sei $X(i)$ der Indikator des Ereignisses {die i-te gezogene Kugel ist schwarz}, $1 \le i < n$. Wie ist das Tupel $X = (X(i))$ verteilt? (Modell für zwei konkurrierende Species, deren Vermehrungsrate proportional zur jeweiligen Bevölkerungsgröße ist.)

Wir stellen uns die Ergebnismenge, die Menge aller Folgen von Symbolen 0 und 1 der Länge n als Blätter eines vollständigen binären Baums der Höhe n vor. Die Partition $F(h)$ ist erzeugt von derAbbildung, die jeder Folge ihre ersten h Buchstaben zuordnet; ihre Atome entsprechen den Knoten in der Höhe h. Zu jedem Knoten $z = (x(1), .. , x(h))$ gibt es zwei unmittelbare Nachfolger, die durch Anhängen des Symbols 0 oder 1 entstehen; wir schreiben daher einfacher $Q(z,j)$ anstelle von $Q(z , (z,j))$. Diese Zahlen sind gegeben durch

$$Q(z , 1) = (s + \sum_{i \le h} x(i))/(s+w+h) , \; Q(z , 0) = (w + \sum_{i \le h} (1 - x(i)))/(s+w+n) .$$

Das Produkt der $Q(. , .)$ längs eines Wegs von der Wurzel zum Blatt $x = (x(1), .. , x(n))$ ist

$$P(x) = (\prod_{0 \le h < n}(s+w+h))^{-1} \cdot \prod_{0 \le i < m} (s+i) \cdot \prod_{0 \le i < n-m} (w+i) , \tag{17}$$

mit $m = m(x) = \sum_{1 \le i \le n} x(i)$. Wir stellen fest, daß das Tupel der $X(i)$ symmetrisch verteilt ist, d.h. daß seine Verteilung invariant unter Permutation der Indices i ist, denn die r.S. in (17) hängt nur von $m(x)$ ab. Das ist deswegen bemerkenswert, weil in der Vorschrift des Ziehens und Nachlegens auf den ersten Blick keine Invarianz irgendwelcher Art gegenüber einer Permutation der Zeitpunkte des Ziehens zu erkennen ist.

Da die Anzahl aller Folgen x der Länge n mit $\sum x(i) = k$ gleich $C(n,k)$ ist, erhält man

$$P(\sum X(i) = k) = C(s+w+n-1,n)^{-1} \; C(s+k-1,k) \; C(w+n-k-1,n-k) . \tag{18}$$

Im speziellen Fall $s = w = 1$ wird das sehr einfach :

$$P(\sum X(i) = k) = 1/(n+1) \qquad \text{für } 0 \le k \le n.$$

Beispiel 4. Ein Zellvermehrungssprozeß oder ein zufällig ausgedünnter binärer Baum.

Wir denken uns einen (im Prinzip unbeschränkt hohen; wer Angst hat, kann wie in Beispiel 1 einen Plafond n einziehen und, nach gemachter Rechnung, n gegen unendlich streben lassen) binären Baum mit Wurzel. Die Wurzel entspreche der Mutterzelle der 0-ten Generation; sie wird durch die "leere Folge" (Folge der Länge null) repräsentiert. Sie habe 2 mögliche direkte Nachkommen, 1 und 2 . Die Mitglieder der k-ten Generation werden durch Folgen der Länge k von Symbolen 1 und 2 dargestellt; die Regel ist, daß die direkten Nachkommen von x als $x1$ und $x2$ bezeichnet werden. Dies zur Beschriftung des Baums.

Der Zufall kommt dadurch ins Spiel, daß die beiden Angehörigen der Generation 1 unabhängig voneinander nur mit Ws. p ins Leben treten. Weiter, wenn feststeht, daß in der k-ten Generation die Individuen x, y , z ,... ins Leben getreten sind, werden die Lebenden der (k+1)-ten Generation aus der Liste $x1$, $x2$, $y1$, $y2$, $z1$, $z2$ ... unabhängig voneinander mit Ws. p ausgewählt. Um Trivialitäten zu vermeiden, nehmen wir $0 < p < 1$ an.
Einfache Fragen in diesem Modell sind die folgenden: mit welcher Ws. gibt es in der k-ten Generation noch lebende Individuen? Wieviele sind es im Mittel? Allgemein, wie ist ihre Anzahl verteilt?

Wir beginnen mit der ersten Frage. Sei $X(k)$ die Stärke der k-ten Generation und sei $a(k) := P(X(k) = 0)$. Dann hat man die Beziehung

(19) $\quad a(0) = 0 ; \quad a(k+1) = (1 - p)^2 + 2p(1-p)a(k) + p^2a(k)^2 \quad$ für $\; k \geq 0 ,$

wie man durch Bedingen auf die Fälle $X(1) = 0 , 1 , 2$ sieht. Wir setzen $f(u) = (1 + p(u-1))^2$ und schreiben (19) als

(19a) $\quad a(0) = 0 , a(k+1) = f(a(k)) , \quad$ oder $\quad a(k) = f^k(0) , k \geq 0 .$

Die Funktion f bildet das Intervall [0,1] in sich ab; sie ist monoton wachsend und strikt konvex, es gilt $f(0) = (1 - p)^2 > 0$ und $f(1) = 1$. Man kann sich die k-te Iterierte von f so vorstellen, daß man k Stufen auf einer Treppe aus horizontalen und vertikalen Stücken zwischen dem Graphen von f und der Diagonalen $\{(u,u) : 0 \leq u \leq 1\}$ hinaufgeht. Es gibt nun zwei Möglichkeiten: der Graph von f verläuft entweder im ganzen Intervall [0,1) strikt oberhalb der Diagonalen, oder es gibt ein a mit $0 < a < 1$, so daß $f(u) > u$ für $u < a$ und $f(u) <$

Zeichnung. Der Graph von f für $p = 0.6$, $a = 0.444$

u für $u > a$ gilt. Mit anderen Worten: wenn man a als den kleinsten Fixpunkt von f in $[0,1]$ definiert, gilt entweder $a = 1$ oder $0 < a < 1$.

Um zu entscheiden, welche der beiden Möglichkeiten vorliegt, betrachtet man f in einer Umgebung von 1 . Wenn $f'(1) > 1$ gilt, hat man $f(u) < u$ in einer Umgebung von 1 und folglich $a < 1$. Wenn $f'(1) \leq 1$ gilt, dann ist wegen der strikten Konvexität $f'(u) < 1$ für alle $u < 1$ und folglich $f(u) > u$ für alle $u < 1$, also $a = 1$. Nun ist $f'(1) = 2p$; somit haben wir das erste Ergebnis:

(E1) Die Folge der Ws.en, daß der Wachstumsprozeß in der n-ten Generation nicht mehr lebt, ist wachsend in n . Sie strebt gegen 1 genau bei $p \leq 1/2$. (Wir werden später sagen, daß für $p > 1/2$ eine strikt positive Ws. besteht, daß der Prozeß nie ausstirbt.)

Zu den Mittelwerten: ein lebendes Individuum hat zwei potentielle Kinder, die je mit Ws. p ins Leben treten, im Mittel also $2p$ lebende Kinder. Unter der Bedingung $X(k) = m$ gilt also

$$E(X(k+1) \mid X(k) = m) = 2pm .$$

Zeichnung. Der Graph von f für $p = 0.4$, $a = 1$.

Multipliziert man beide Seiten mit $P(X(k) = m)$ und summiert über alle m , liefert Prop 1.2

$$EX(k+1) = 2p\, EX(k) .$$

(E2) Für die Mittelwerte gilt $EX(k) = (2p)^k$ für $k \geq 0$. Die Folge der Mittelwerte ist exponentiell wachsend für $p > 1/2$, sie ist exponentiell fallend oder konstant für $p \leq 1/2$.

Die Bestimmung der gesamten Verteilung von $X(n)$ ist im Prinzip nicht viel schwieriger als die Bestimmung der Aussterbews. Das hierfür benötigte Instrument ist die Methode der erzeugenden Funktionen, das wir jetzt vorstellen.

Einschub. Erzeugende Funktionen und Faltung.

Def. 2.4. Wenn $p(k)$, $k \geq 0$, die Verteilung einer ZV X mit Werten in den natürlichen Zahlen ist, dann heißt erzeugende Funktion dieser Verteilung (oder von X) die Funktion

(20) $$u \to f(u) := \sum_{k \geq o} p(k) \cdot u^k = Eu^X .$$

Als Definitionsbereich von f sei das Intervall $0 \leq u \leq 1$ gewählt. (Auch andere Wahlen sind möglich.) Offenbar bestimmt f die Folge der $p(k)$ eindeutig. Die Angabe der erzeugenden Funktion ist zu verstehen als eine "Codierung" der Zahlenfolge $(p(k))$.

Prop. 2.3. Wenn X und Y unabhängige reelle ZVe sind, hat man

(21) $$E(XY) = (EX)(EY) .$$

Beweis. $\sum_{x,y} P(X=x, Y=y) \cdot x \cdot y = \sum_{x,y} P(X=x) \cdot P(Y=y) \cdot x \cdot y = \{\sum_x P(X=x) \cdot x\} \cdot \{\sum_y P(Y=y) \cdot y\}$.

Prop. 2.4. Wenn X und Y unabhängig sind und die erzeugenden Funktionen f bzw. g haben, dann hat X+Y die erzeugende Funktion f·g .

Beweis. $Eu^{(X+Y)} = Eu^X \cdot u^Y = (Eu^X)\cdot(Eu^Y)$, da auch die ZVen u^X und u^Y unabhängig sind.

Def. 2.5. Wenn drei Verteilungen p , q , r in N mit erzeugenden Funktionen f , g , h in der Relation zueinander stehen, daß $h = f\cdot g$ gilt, nennt man r die Faltung von p und q , geschrieben als $r = p*q$. Die Gewichte der Verteilungen (Koeffizienten der erzeugenden Funktionen) stehen dann in der Beziehung

(22) $$r(n) = \sum_m p(m)\cdot q(n-m) \qquad \text{für alle } n$$

zueinander. In der Sprache der ZVen: $p*q$ ist die Verteilung von X+Y , wenn X bzw. Y nach p bzw. q verteilt und beide unabhängig sind.

‡ Ende des Einschubs

Offenbar ist das oben betrachtete $f(u) = (1 + (p-1)\cdot u)^2$ eine erzeugende Funktion, und zwar der Anzahl der lebenden Nachkommen eines lebenden Individuums. Für jedes $n \geq 0$ definieren wir $g(n, .)$ als die erzeugende Funkton von X(n). Nun gilt für $n \geq 1$ die Zerlegung $X(n) = X(n,1) + X(n,2)$, mit unabhängigen Summanden, wenn man X(n,i) als die Lebenden der n-ten Generation definiert, die in der ersten Generation den Vorfahren i haben. Weiter, unter der Bedingung, daß i ins Leben tritt, ist X(n,i) ebenso verteilt wie X(n-1) unbedingt. Somit bekommt man die erzeugende Funktion von X(n,i) als

$$Eu^{X(n,i)} = 1-p + p\cdot g(n-1,u) , \qquad i = 1 \text{ oder } 2 ,$$

und damit die von von X(n) als

(23) $$g(n,u) = E\, u^{X(n)} = (Eu^{X(n,1)})(Eu^{X(n,2)}) = \{1-p + p\cdot g(n-1,u)\}^2 = f(g(n,u)) .$$

Das bedeutet, daß das Bildungsgesetz der Folge (g(n,u)) für jedes u identisch ist mit dem Bildungsgesetz von (a(n)) ; lediglich die Anfangswerte sind verschieden, da $g(0,u) = u$. Die explizite Lösung der Iterationsvorschrift (23) lautet:

(E3) $$g(n,u) = f^n(u) , \quad 0 \leq u \leq 1 \text{ und } n \geq 0 .$$

Wir sehen, daß (E1) ein Spezialfall von (E3) ist, wegen $g(n,0) = P(X(n) = 0)$. Sei a der kleinste Fixpunkt der Funktion f im Intervall [0,1] . Dann konvergiert für jedes $u < 1$ die Folge (g(n,u)) bei n gegen unendlich gegen a , und allgemein hat man :

(24) $\lim_n g(n,u) =: g(\infty,u)$, wobei $g(\infty,u) = a$ für $u < 1$ und 1 für $u = 1$.

Im Fall $a = 1$ ist $g(\infty, .)$ identisch 1 ; man kann von einer Limesvariablen $X(\infty)$ sprechen, die gleich 0 mit Ws. 1 ist. Im Fall $a < 1$ ist $g(\infty, .)$ unstetig bei 1 , und man wird - mit aller Vorsicht - sagen, $X(\infty)$ sei gleich $+\infty$ mit Ws. $1-a$ und 0 mit Ws. a und die Ws. für alle Werte zwischen 0 und $+\infty$ sei null. Wir stellen also, bei $a < 1$ oder $p > 1/2$, eine Dichotomie im Verhalten von $X(n)$ bei großem n fest: entweder Aussterben oder Wachstum über alle Schranken, aber kein Dazwischenbleiben auf einem endlichen Niveau.

Eine letzte Bemerkung. Unser Vermehrungsmodell ist ein einfaches Beispiel für einen "Phasenübergang", d.h. eine (sprunghafte) qualitative Änderung im asymptotischen Verhalten eines Systems in Abhängigkeit von einem (kontinuierlichen) Parameter. Wir nennen entsprechend $1/2$ den kritischen Wert von p und sprechen von $p > 1/2$ bzw. $p < 1/2$ als dem superkritischen bzw. subkritschen Bereich. Weiter: die "Treppenkonstruktion" der Iterierten von f zeigt, daß $a(n)$ mit exponentieller Geschwindigkeit gegen a strebt - sofern der Winkel zwischen der Diagonalen und dem Graphen von f bei a positiv ist. Das ist im super- und subkritischen Bereich der Fall. Für $p = 1/2$ ist die Konvergenz wesentlich langsamer; man überlegt sich, daß sie die Geschwindigkeit von $1/n$ hat. Eine derart verlangsamte Asymptotik genau an der kritischen Stelle ist typisch für das Szenarium eines Phasenübergangs.

Man nennt die Folge $(X(n))$ einen speziellen Galton-Watson-Prozeß. In allgemeinen GW-Prozessen werden beliebige Verteilungen für die Nachkommenzahl eines Individuums zugelassen, nicht nur $B(2,p)$ wie hier.

Beispiel 5. Faire Glücksspiele.

Das einfachste Modell sieht so aus. Eine Münze wird n-mal unabhängig geworfen; mit Ws. $1/2$ fällt jedesmal "Kopf" oder "Wappen". Wir stellen die Ergebnismenge daher dar als Menge aller Folgen x der Länge n aus zwei Symbolen, die wir $+1$ und -1 nennen wollen (der Grund wird später klar), oder als vollständigen binären Baum der Höhe n . Die W.-Bewertung P legt das Gewicht 2^{-n} auf jede Folge x .

Wir betrachten einen Spieler, der mit der Bank nach folgenden Regeln spielt: er darf vor jedem Wurf einen beliebigen Betrag a auf einen der beiden Ausgänge setzen; wenn der gewettete Ausgang erscheint, bekommt er $2a$ zurück, andernfalls ist der Einsatz verloren. Wir haben das Spiel fair genannt, weil bei einer einmaligen Wette der erwartete Reingewinn (erwarteter Gewinn minus Einsatz) gleich $1/2\cdot(2a + 0) - a$, also gleich null ist. Das Problem besteht darin, eine Aussage über mögliche Gewinne und Verluste während des gesamten Spielverlaufs vom ersten

bis letzten Wurf zu machen, wobei der Spieler bei jedem einzelnen Münzwurf sein Wettverhalten von seinen bisherigen Beobachtungen abhängig machen darf.

Wir müssen zuerst formalisieren, was wir in diesem Modell unter einem möglichen oder erlaubten Wettverhalten (genauer: Strategie oder Handlungsanweisung, an die der Spieler sich oder seinen Agenten für die gesamte Spieldauer bindet) verstehen wollen.

(D1) Eine zulässige Wettstrategie ist eine Folge $a(i)$, $1 \le i \le n$, von reellen Zufallsvaria blen, d.h. Funktionen $a(i,.)$ von x, wobei $a(i, .)$ nur von $x(j)$, $1 \le j < i$, abhängt. (Speziell $a(1, .)$ ist dann eine Konstante.)

Interpretation: $a(i) = a$ mit $a \ge 0$ heißt, daß der Spieler den Betrag a auf den Ausgang +1 im i-ten Wurf setzt; wenn er den Betrag a auf -1 setzt, sagen wir $a(i) = -a$. In jedem Fall gilt dann, daß $G(i)$, definiert als Reingewinn des Spielers im i-ten Wurf, gleich $X(i)\cdot a(i)$ wird, welche Zahl positiv oder negativ sein kann. Die Forderung, daß $a(i ,.)$ nur von $x(j)$, mit $j < i$, abhängen darf, bedeutet, daß wir dem Spieler keine hellseherischen Fähigkeiten zubilligen. Das Hauptresultat bei dieser Formalisierung von zulässigem Wettverhalten ist enthalten in folgendem Satz, der eine mathematische Präzisierung der Alltagsweisheit "In einem fairen Glücksspiel gibt es kein gewinnbringendes Wettsystem" darstellt.
(E1) Unter jeder zulassigen Wettstrategie $a = (a(i))$ gilt für die Erwartung des Gesamtge winns $G = \sum_i G(i)$ die Identität

(25) $$EG = 0 \ .$$

Beweis. Es genügt, wegen der Linearität der Erwartung, $EG(i) = 0$ für jedes i zu zeigen. Nun gilt nach dem Satz von totalen Ws.

(26) $$EG(i) = 2^{1-i} \sum_{x(1),\,..x(i-1)} E(G(i) \mid X(j) = x(j),\ 1 \le j < i) =$$

$$2^{1-i} \sum_{x(1),\,..x(i-1)} E(a(i;x(1), ..,x(i-1))\cdot X(i) \mid X(j) = x(j)\ ,\ 1 \le j < i)\ .$$

Für jeden der 2^{i-1} Summanden gilt: der Faktor $a(i; .)$ in der bedingten Erwartung ist als Konstante anzusehen, da jede einzelne bedingte Verteilung von einer solchen Menge von Folgen x getragen wird, auf welchen $a (i; .)$ konstant ist; wegen der Unabhängigkeit von $X(i)$ und $X(j)$, $j < i$, ist jede bedingte Erwartung von $X(i)$ gleich der unbedingten Erwartung, also gleich 0. Daher ist der Wert jedes Summanden in (26) gleich null, und damit ist $EG(i) = 0$.

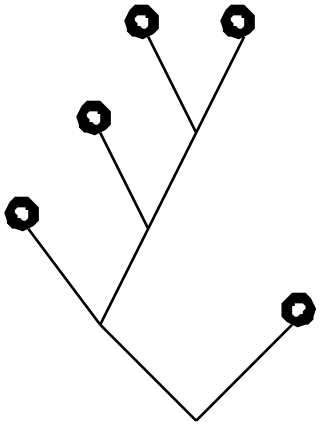
Unser Begriff von zulässiger Wette umfaßt alles, was man sich an Spielsystemen vorstellen kann, darunter auch so extremes Verhalten wie die Petersburger Strategie: man beginnt mit

einem Einsatz a , verdoppelt bei einem Verlust im nächsten Spiel den Einsatz - solange bis man das erste Mal gewinnt; dann hört man zu spielen auf. Man gewinnt bei dieser Strategie mit Ws. $1 - 2^{-n}$ den Betrag a , mit Ws. 2^{-n} hingegen verliert man $(2^n - 1)\cdot a$, die Gewinnerwartung ist in der Tat null. (Das sog. Petersburger Paradoxon kommt durch einen etwas unbedachten Grenzübergang von n gegen unendlich zustande; man gewinnt so angeblich in einem unendlich lange währenden Spiel mit Ws. 1 den Betrag a . Man sollte das Paradox nicht zu ernst nehmen. Was ist null mal unendlich? Man verliert ja auch unendlich viel mit Ws. null!)

Einen Spezialfall einer zulässigen Wettstrategie stellen Strategien dar, bei denen a(i; .) nur der Werte 0 oder 1 fähig ist, mit der Zusatzbedingung, daß aus a(i;x) = 0 folgt, daß auch a(j;x) = 0 für alle j > i . Eine solche Strategie ist eineindeutig einer Abbildung T : E → {0,1,. .. , n} zugeordnet, welche - in der Sprache der binären Bäume wie in Prop. 2 - folgender Bedingung genügt :

(27) Für jedes i ist die Menge {T > i} aus Elementen der Partition *F*(i) zusammengesetzt.

Die Korrespondenz ist klar: T(x) = sup{i : a(i,x) = 1}. Man nennt ein (27) genügendes T eine Stoppzeit, welcher Name sehr anschaulich ist. Das entsprechende Wettverhalten besteht darin, zu allen Zeitpunkten i mit $i \le T$ auf +1 den Betrag 1 zu setzen und nach der Zeit T aufzuhören. Eine Stoppzeit kann man sich veranschaulichen als Möglichkeit, die Zweige des binären Baums der Höhe n zu beschneiden, wobei "beschneiden" bedeutet, daß man sukzessive, von der Wurzel kommend, auf jeder Höhe einzelne Knoten auswählt, von denen man keine Zweige mehr nach oben wachsen läßt.



Darstellung einer Stoppzeit. Ein dicker Knoten in Höhe i entspricht einer Vorgeschichte (x(1), ..,x(i)), bei der zum ersten Mal auf die Teilnahme an der(i+1)-ten Wette verzichtet wird; an den dünnen Knoten heißt die Anweisung"weitermachen".

Bemerkung. Aussage (E1) ist nur mit Vorsicht als Argument gegen die Teilnahme an fairen

Glücksspielen zu betrachten. Wir haben nämlich Spiele am erwarteten Gewinn EG gemessen, einer einzigen reellen Zahl, nicht aber an der gesamtenVerteilung von G. Es ist ja beileibe nicht selten, daß es jemandem Spaß macht, mit der Ws. 10^{-6} eine Million Mark zu gewinnen und mit der Restws. gar nichts, und daß er bereit ist, für diesen Spaß mehr als die erwartete eine DM zu zahlen, sondern etwa das doppelte - das ist das Prinzip des Lotto. Wer diesen Reiz sucht, Riesengewinne, wenn auch mit kleiner Ws., zu machen, verhält sich jedoch nicht unbedingt rational, wenn er sich an einer Wettgemeinschaft beteiligt, mit dem Ergebnis einer n-mal höheren Gewinnchance, bei einem durch n geteilten Gewinn. (Es sei denn, er sucht genau das, kleinere Gewinne, aber höhere Chancen.) Auf jeden Fall idiotisch ist eine Wettgemeinschaft aller Lottospieler: sie garantiert den Verlust des halben Einsatzes und nichts weiter. - Erörterungen des Problems, wie man die Verteilung von G in Mark und Pfennig bewerten soll, hat es unter dem Namen einer Theorie der "utility" oder des subjektiven Nutzens viele gegeben, vor allem in der mathematischen Ökonomie. Er spielt eine Rolle bei der subjektivistischen Begründung des Begriffs der Ws., in der man Axiome für ein unter Unsicherheit "rational" handelndes Individuum aufzustellen trachtet (siehe z.B. [LR]). Der an Analysis und Physik orientierten W.-Theorie sind Fragen dieser Art fremd geblieben.

Def. 2.6. Wenn X reelle Zufallsvariable ist und p natürliche Zahl, dann heißt der Ausdruck

(26) $$EX^p \qquad \text{bzw.} \qquad E|X|^p \qquad \text{bzw. } E(X - EX)^p$$

das p-te Moment bzw. absolute Moment bzw. zentrierte Moment von X. Einen eigenen Namen bekommt das zweite zentrierte Moment: Varianz, geschrieben $\mathrm{var}(X)$. Als Kovarianz von X und Y bezeichnet man die Zahl

(29) $$\mathrm{cov}(X,Y) := E\{(X - EX)(Y - EY)\} \quad (= E(XY) - EX \cdot EY) .$$

Als Kovarianzmatrix eines Tupels $X = (X(i))$ bezeichnet man die Matrix C mit Einträgen

(30) $$C(i,j) = \mathrm{cov}(X(i),X(j)) .$$

Offenbar ist die Kovarianz eine symmetrische Bilinearform im Raum der Zufallsvariablen. Die Kovarianzmatrix C soll man sich vorstellen als die Koeffizientenmatrix der symmetrischen Bilinearform $u \to \mathrm{var}(\sum_i u(i)\, X(i))$, die auf dem Dualraum des Vektorraums V erklärt ist, in welchem X seine Werte annimmt.

Prop. 2.5. Varianz und Kovarianz ändern sich nicht bei der Addition von Konstanten. Die Vari-

anz einer ZV X ist immer nichtnegativ; sie ist null genau dann, wenn X mit Ws. 1 konstant ist. Die Kovarianzmatrix einer V-wertigen ZV X ist stets nichtnegativ. Sie ist ausgear-tet (hat den Eigenwert null) genau dann, wenn es eine niederdimensionale affine Unterman-nigfaltigkeit M von V gibt, so daß X mit Ws. 1 in M liegt.

Prop 2.3. (Andere Fassung) Wenn X und Y unabhängig sind, so ist $\mathrm{cov}(X,Y) = 0$.

Prop 2.6. (Ungleichung von Tschebyschov). Für jede ZV X und jedes $a > 0$ gilt

(31) $$P(|X-EX| \geq a) \leq a^{-2}\, \mathrm{var}(X) .$$

Beweis. Sei $EX = 0$ o.E.d.A. Wir haben die Minorisierung $|X| \geq a \cdot 1_{\{|X|\geq a\}}$. Quadriert man diese Relation und bildet auf beiden Seiten die Erwartung, erhält man (31).

Aus diesen relativ simplen Rechungen bekommt man das nachstehende Ergebnis, das sog. schwache Gesetz der großen Zahlen. Wir führen es als "Satz", weil es für die Theorie der Ws. und ihre Verankerung in der Empirie von unschätzbarer Wichtigkeit ist. Es sagt, daß im Fall der Gültigkeit oder "Wahrheit" eines probabilistischen Modells sämtliche Modellparameter, also die gesamte W.-Bewertung, der empirischen Überprüfung zugänglich ist - sofern es sich um einen beliebig oft unabhängig reproduzierbaren Versuch handelt. (Man muß allerdings hinzufügen, daß die Annahme der Reproduzierbarkeit ihrerseits durch die Beobachtung des Experiments in Serie nicht bestätigt werden kann; die Reproduzierbarkeit muß man a priori postulieren.)

Satz 2.1. (Schwaches Gesetz der großen Zahlen). Sei $X(i)$, $i = 1 .. n$, eine Folge von u.i.v. Zufallsvariablen endlicher Erwartung. Dann gilt für jedes $a > 0$

(32) $$P(|n^{-1} \textstyle\sum_{1\leq i\leq n} X(i) - EX| \geq a) \leq \mathrm{var}(X)\, a^{-2}\, n^{-1}.$$

(33) $$\lim_n P(|n^{-1} \textstyle\sum_{1\leq i\leq n} X(i) - EX| \geq a) = 0.$$

(Die zweite Zeile ist mit etwas Vorsicht zu lesen; strenggenommen müßte man auch an P den Index n anhängen oder sich Gedanken machen, wie unendlich viele u.i.v. ZVe auf ein und demselben W.-Raum erklärt sein können.)

Beweis. Sei $EX = 0$ o.E.d.A. DieVarianz von $n^{-1} \sum_{1\leq i\leq n} X(i)$ ist nach Prop.2.3 gleich $n^{-2} \sum_{1\leq i\leq n} EX(i)^2 = n^{-1}\, \mathrm{var}(X(1))$. Der Rest ist Prop. 2.6.

Man nennt $n^{-1} \sum_{1\leq i\leq n} X(i)$ das Stichprobenmittel (aus den ersten n Beobachtungen). Der Satz

sagt, daß man den (als "unbekannt" anzusehenden) Mittelwert $EX(1)$ bis auf einen beliebig kleinen Fehler a mit beliebig geringer Ws. eines Irrtums - das ist die rechte Seite in (32) - angeben kann, indem man ihn durch das Stichprobenmittel "schätzt", falls n groß genug ist. Wenn man die Erwartungen aller ZVen kennt, dann kennt man auch die Ws.en von Ereignissen, wegen $P(A) = E\, 1_A$. Wir geben dem Sachverhalt in (33) einen eigenen Namen:

Def. 2.7. Eine Folge $(Z(n))$ von Zufallsvariablen konvergiert stochastisch gegen Z , wenn

(34) für alle $a > 0$ gilt $\lim_n P(|Z(n) - Z| > a) = 0$.

Literatur.

[E] A. Engel: Wahrscheinlichkeitsrechnung und Statistik. 2 Bände. Stuttgart: Klett 1973 und 1976.

[LR] R.D. Luce, H. Raiffa: Games and decisions. Introduction and critical survey. New York: Wiley 1957.

Kapitel 3. Die Binomialverteilung

Inhalt: Stirling-Formel. Asymptotik der Binomialgewichte; zentraler Grenzwertsatz von de Moivre-Laplace, lokal und global; Logarithmus der Ws. von "großen Abweichungen".

Einschub: Die Stirling-Formel.

Das wichtigste analytische Hilfsmittel dieses Abschnitts ist die Stirlingsche Formel für die Asymptotik der Gammafunktion oder der Fakultät. Bekanntlich gilt $n! = \Gamma(n+1)$, wobei

(E1) $$\Gamma(n) := \int_{o<x<\infty} x^{n-1} \exp(-x)\, dx .$$

Wir setzen die Formel als bekannt voraus, siehe z.B. [Fo] oder [K]; wer will, findet auf dieser Seite einen etwas anderen Beweis.

Lemma. Sei die Funktion f in R definiert: sie erfülle :

$1 \geq f \geq 0$, $f(0) = 1$, $\sup \{f(x) : |x| \geq r\} < 1$ für alle $r > 0$, f ist integrabel und in einer Umgebung von 0 zweimal differenzierbar, $f''(0) = -1$.

Dann gilt

(E2) $$\int f^n(x)\, dx = (2\pi/n)^{1/2} (1 + o(1)) \text{ bei } n \text{ gegen unendlich.}$$

Beweisskizze. Die Asymptotik des Integrals in (E2) ändert sich nicht, wenn man nur über eine (beliebig klein wählbare) Umgebung von 0 integriert. Entwickelt man dort den Integranden nach Taylor bis zur Ordnung 2 , so sieht man, daß die Asymptotik für alle Funktionen mit den gleichen Ableitungen wie f bis zur Ordnung 2 die nämliche ist; eine spezielle Funktion in dieser Klasse ist $g(x) := \exp(-x^2/2)$. Nun gilt exakt

$$\int g^n(x)\, dx = n^{-1/2} \int g(x)\, dx ,$$

wie man durch die Variablentransformation $x \to x\, n^{-1/2}$ sieht. Die Behauptung folgt aus der Auswertung des Integrals von g , welche den Wert $(2\pi)^{1/2}$ ergibt. (Üblicher Trick mit Integration

von $g(x)\cdot g(y)$ in der Ebene.)

Satz 0 .(Stirling) Für n gegen unendlich gilt die Asymptotik

(E3) $$\Gamma(n+1) = (2\pi n)^{1/2}\,(n/e)^n\,(1 + o(1)) .$$

Beweis. Man setzt $h(x) = c\,x\exp(-x)$ und wählt $c = e$; dann wird das Maximum von h gleich 1 . Der Wert der zweiten Ableitung von h an der Maximalstelle ist -1 . Dann wendet man auf (ein Translat von) h das Lemma an. Andererseits ergibt eine elementare Variablentransformation

$$\int h^n(x)\,dx = c^n \int x^n \exp(-nx)\,dx = c^n\, n^{-(n+1)} \int x^n \exp(-x)\,dx = c^n\, n^{-(n+1)}\,\Gamma(n+1) .$$

Ende des Einschubs

Def. 3.1. Die Binomialverteilung zu den Parametern n und p , geschrieben $B(n,p)$, ist diejenige W.-Verteilung auf den ganzen Zahlen von 0 bis n, welche auf die Zahl k das Gewicht

(1) $$B(n,p;k) := C(n,k)\,p^k\,(1-p)^{n-k}$$

legt. Hierbei ist $C(n,k)$ der Binomialkoeffizient $n!\,(k!\,(n-k)!)^{-1}$.

Die Interpretation ist klar (siehe Kapitel 1): Führt man einen Versuch mit zwei Ausgängen "Erfolg" und "Mißerfolg" , wobei die Ws. für "Erfolg" gleich p sei, n-mal unabhängig durch, so ist $B(n,p;k)$ die Ws. dafür, daß insgesamt k-mal "Erfolg" eintritt. Oder anders, $B(n,p)$ ist die Verteilung einer Zufallsvariablen $\sum_{1\le i\le n} X(i)$ mit u.i.v. Summanden $X(i)$, wobei

$$P(X(i) = 1) = 1 - P(X(i) = 0) = p .$$

Aus der Interpretation folgt sofort

Prop. 3.1. Eine nach $B(n,p)$ verteilte Variable hat Erwartung np und Varianz $np(1-p)$; ihre erzeugende Funktion ist $u \to (1 - p + p\cdot u)^n$.

Aus der Berechnung der Varianz folgt das Gesetz der großen Zahlen für die relative Erfolgshäufigkeit in einer Folge von identischen 0-1-Experimenten. Wir notieren es als eigene Proposition, obwohl es sich nur um einen Spezialfall der Ungleichung von Tschebyschov handelt.

Prop. 3.2. Wenn X nach $B(n,p)$ verteilt ist, so gilt

(2) $P(|X - np| \geq na) = P(|X/n - p| \geq a) \leq a^{-2}(np(1-p))^{-1}$ für jedes $a > 0$.

Im folgenden suchen wir nach schärferen Aussagen über die Gestalt der Binomialverteilung als den in Prop. 3.2 gemachten. Erstens wollen wir von der "Abweichung" $X - np$ die exakte W.-Verteilung beschreiben, wenn man sie im richtigen Maßstab mißt, und nicht nur sagen, daß sie für jedes $a > 0$ mit hoher Ws. absolut kleiner als $a \cdot n$ ist. Zweitens wollen wir genauere Aussagen über die Größe von $P(|X - np| \geq a \cdot n)$ machen . Es wird sich herausstellen, daß sie nicht nur mit der Geschwindigkeit $O(1/n)$, wie in (2) behauptet, sondern exponentiell in n gegen null geht.

Zunächst eine qualitative Aussage.

Prop. 3.3. Die Funktion $k \to b(k) = B(n,p;k)$ ist bei festem p und n nichtfallend bis zu einer Stelle k_o und anschließend nichtwachsend. Die Maximalstelle k_o liegt bis auf einen Fehler von höchstens 1 bei np .

Beweis: Betrachte den Quotienten $b(k)/b(k+1)$. Sein Wert ist $(1-p)\cdot(k+1)/(p\cdot(n-k))$ und daher größer oder kleiner als 1 , je nachdem ob $(k+1)/(n+1)$ kleiner oder größer als p ist.

Die Idee der jetzt folgenden Ausführungen besteht darin, die Funktion $k \to B(n,p:k)$ durch eine analytisch einfache Funktion mit kontinuierlichem Definitionsbereich zu approximieren, d.h. das Histogramm mit Säulenhöhe $B(n,p;k)$ dem Graphen einer glatten Funktion anzunähern. Wir werden dabei zu einem fundamentalen Ergebnis der Analysis kommen, dem Zentralen Grenzwertsatz. In früheren Zeiten war diese Approximation auch numerisch von großer Bedeutung. Das trifft heute nicht mehr im gleichen Maß zu, da die exakten Werte der Binomialgewichte oder ihrer Partialsummen mit der praktisch benötigten Genauigkeit leicht abrufbar sind. Trotzdem ist im Bereich "großer" Stichproben, sagen wir in der Gegend $n \geq 1000$, der zentrale Grenzwertsatz in seiner Einfachheit unschlagbar. Wir betonen aber, daß wir nicht an numerischen Fragen interessiert sind, etwa der Art: "Für welche Werte von n und p ist der zentrale Grenzwertsatz mit welcher Korrektur praktisch zu empfehlen?"

Satz 3.1 (Lokaler Grenzwertsatz). Sei $0 < p < 1$ fest. Es gilt bei n gegen unendlich

(3) $B(n,p:k) = (2\pi p(1-p)n)^{-1/2} \exp(-(k - np)^2/2np(1-p))(1 + o(1))$,

gleichmäßig in den Bereichen $\{k : |k- np| \leq Cn^a\}$ mit $C < \infty$ und $a < 2/3$.

Beweis. Wir setzen $m := n - k$ und $q := 1 - p$. In dem genannten Bereichen strebt sowohl k als auch m gegen unendlich; man kann also die Stirlingformel anwenden. Sie ergibt

(4) $\log B(n,p;k) = -1/2 \log (2\pi km/n) - n\{k/n \log(k/np) + m/n \log(m/nq)\} + o(1)$.

In den genannten Bereichen strebt weiter k/n gegen p und m/n gegen q. Der erste Summand rechts in (4) ist also bereits von der Form $-1/2 \log (2\pi pqn) + o(1)$. Den zweiten Summanden schreiben wir in der Form

(4a) $- n I(r,p)$, mit $r = k/n$ und $I(r,p) := r \log(r/p) + (1-r) \log((1-r)/q)$.

Wir entwickeln $I(\,.\,,p)$ an der Stelle $r = p$ nach Taylor und erhalten

(5) $I(r,p) = (2pq)^{-1} (r - p)^2 + O(|r - p|^3)$.

Multipliziert man (5) mit n, so geht der Fehlerterm rechts nach null, solange $n \cdot |r-p|^3 = n^{-2} \cdot |k - np|^3$ nach null geht, was in den Bereichen $\{k : |k- np| \le Cn^a\}$ der Fall ist. Wenn man also den zweiten Summanden rechts in (4) durch $-n (2pq)^{-1} (r - p)^2 = -(k - np)^2/2np(1-p)$ ersetzt, so begeht man einen beliebig kleinen additiven Fehler, d.h.

$$\log B(n,p;k) = -1/2 \log (2\pi npq) - (k - np)^2/2np(1-p) + o(1) .$$

Übergang zum Exponenten führt auf (3).

Der Satz ist leicht zu merken: die Funktion $x \to (2\pi pqn)^{-1/2} \exp(-(x - np)^2/2npq)$, welche in (3) auftritt, ist bis auf eine affine Verzerrung die standardisierte Gaußische Dichte ("Glockenkurve") $t \to (2\pi)^{-1/2} \exp(-t^2/2)$. Die Verzerrung ist so beschaffen, daß der Mittelwert an die Stelle np rückt, an der auch der Mittelwert der Binomialverteilung sitzt, und daß die Funktion in horizontaler Richtung um den Faktor $(npq)^{1/2}$ auseinandergezogen und vertikal um den gleichen Faktor gestaucht wird. Dieser Faktor ist gleich der Standardabweichung der Binomialverteilung $B(n,p)$. Da die standardisierte Gaußische Dichte die Varianz 1 hat, wie man leicht nachrechnet, stimmen somit die approximierende verzerrte Gaußische Dichte und die Binomialverteilung in Erwartung und Varianz überein. Die kontinierliche Näherungsfunktion hat das Integral 1, ebenso wie die Histogrammfunktion $H(n,p)$ (die dadurch definiert ist, daß sie im Intervall $(k-1/2, k+1/2)$ die konstante Höhe $B(n,p;k)$ hat). Es ist vielleicht günstiger, wenn man das Histogramm affin verzerrt und es mit der standardisierten Glockenkurve vergleicht; der L^1-Abstand von Histogramm und Glockenkurve ist unter der Standardisierung invariant.

Satz 3.2. Für die oben eingeführte Histogrammfunktion $H(n,p)$ gilt

(6) $\lim_n \int |H(n,p;x) - (2\pi npq)^{-1/2} \exp(- (x - np)^2/2npq)| \, dx = 0 .$

Corollar. Wenn X(n) nach B(n,p) verteilt ist, so gilt bei festem p im Limes n gegen unendlich

(7) $P(r < (X(n) - np)(npq)^{-1/2} < s) = (2\pi)^{-1/2} \int_{r<x<s} \exp(- x^2/2) \, dx ,$

gleichmäßig für alle reellen Zahlen $\infty \leq r < s \leq \infty$.

Beweis. Die Aussage (6) folgt unmittelbar aus dem vorigen Satz über gleichmäßige lokale Asymptotik für B(n,p;k) , wenn man bedenkt, daß sie in einem Bereich von x-Werten der Breite n^a mit $a = 2/3 - \epsilon$, ϵ beliebig klein, gilt. Das Integral der Gaußischen Näherungsfunktion über einen solchen Bereich ist nämlich im Limes beliebig nahe bei 1 , da diese Bereiche bei passendem ϵ schneller als $n^{1/2}$ wachsen. Da auch das Integral über H gleich 1 ist und H innerhalb eines solchen Bereichs mit einem beliebig kleinen relativen Fehler durch die Glockenkurve ersetzt werden kann, ändert auch das H-Integral außerhalb dieses Bereichs nur wenig am L^1-Abstand der beiden Funktionen. - Aussage (7) folgt unmittelbar aus (6).

Man nennt das Corollar, welches populärer ist als die schärfere Aussage (6), den Satz von de Moivre-Laplace oder den Zentralen Grenzwertsatz für Binomialverteilungen. Seine Popularität verdankt dieser Satz seiner Universalität, d.h. seiner Gültigkeit weit über die Klasse der Binomialverteilungen hinaus. (Unter allgemeineren Voraussetzungen wird man meist mit einer Aussage vom Typ (7) zufrieden sein und nicht nach einem stärkeren Konvergenzverhalten wie in (6) Ausschau halten.)

Beispiel 1. Meinungsumfrage.

Man zieht aus einer riesigen, mehrere Millionen umfassenden Bevölkerung eine Stichprobe des Umfangs n (in der Größenordnung von 2000) und befragt die ausgewählten Individuen nach ihren Präferenzen als Wähler oder als Verbraucher. Nehmen wir an, es werde eine einfache Frage gestellt, die nur mit Ja oder Nein zu beantworten ist. Der Veranstalter der Umfrage registiriert die Anzahl X der Ja-Antworten. Die Zahl n hat er von Anfang an, beim Aufstellen des Stichprobenplans, festgesetzt; sie wird als nicht zufällig betrachtet. Dann ist ein naheliegendes Modell für diesen Sachverhalt, X als zufällig, und zwar nach B(n,p) verteilt, anzunehmen, wobei p ein unbekanner Parameter ist, nämlich der Anteil der Ja-Stimmen in der Gesamtbevölkerung. Der Sinn des Unternehmens ist es, Aufschluß über p zu bekommen.

Das damit verbundene Problem, wie man die Worte "Aufschluß über p " zu verstehen habe,

gehört zur Statistik; manche Leute würden es als " inverses Problem der Stochastik" bezeichnen. Wir wollen hier auf die spezifisch statistischen Fragen nicht eingehen, sondern nur eine einfache Faustregel für den Umgang mit derartigen Stichprobenzahlen formulieren.

Wenn p der "wahre" Parameter ist, dann ist $n\cdot p$ die Erwartung von X ; die Standardabweichung ist $\sqrt{(n\cdot p(1-p))}$. Es ist daher, wenn wir die Gaußische Näherung zugrundelegen, ebenso ws. , daß X von seinem Mitttelwert um mindestens $c\cdot\sqrt{(n\cdot p(1-p))}$ abweicht, wie das Ereignis, daß eine standardisierte Gaußvariablen einen Betrag größer gleich c hat. Der Zahlenwert für letzteres ist $\sqrt{2/\pi}\cdot\int_{x>c}\exp(-x^2/2)\,dx$; bei $c = 1$ ist das 0.3173 , bei $c = 2$ dagegen 0.0455 . Wenn man von X/n auf p zurückschließen will, muß man daher eine Fehlerbandbreite von $c\cdot\sqrt{p\cdot(1-p)/n}$ nach rechts und links dazu geben, weil in diesem Bereich die "ebenfalls möglichen" Werte von p liegen. Man kennt zwar p nicht, kann aber für die Berechnung der Bandbreite p durch X/n ersetzen. Damit erhält man die Faustregel:

(F) Wenn man $X=k$ beobachtet hat, so soll man den unbekannten Wert von p im Intervall $(k/n - c\cdot\sqrt{k/n\cdot(1-k/n)\cdot 1/n}\ ,\ k/n + c\cdot\sqrt{k/n\cdot(1-k/n)\cdot 1/n})$ vermuten.

Wie man c wählt, ob gleich 2 oder ob man mit einem etwas kleineren Wert zufrieden ist, hängt davon ab, wie sicher man sich sein will, den "wahren" Wert von c in dem in (F) genannten Intervall eingeschlossen zu haben.

Zahlenbeispiel: Sei $n = 2000$, $k = 840$; bei $c = 1.96$ ergibt die Faustregel als Bereich der möglichen p-Werte 0.42 ± 0.0235 oder das Intervall $(0.396\ ,\ 0.444)$, also eine ganz beachtliche Unschärfe. Die Zahl $c = 1.96$ ist so gewählt, daß die "Irrtumsws.", d.h. der Wert des Integrals $\sqrt{2/\pi}\cdot\int_{x>c}\exp(-x^2/2)\,dx$ gleich 0.05 wird.

Beispiel 2. Eine diskretisierte Diffusion (Irrfahrt).

Wir stellen uns ein diffundierendes Teilchen in einem d-dimensionalen Raum derart vor, daß wir den Raum zu einem kubischen Gitter machen und auch die Zeit diskretisieren. (Die Einheiten von Raum und Zeit seien einfachheitshalber als 1 gewählt.) Zu jedem ganzzahligen Zeitpunkt n wird das Teilchen zufällig auf einen anderen Platz gesetzt. Um spezifisch zu sein, lassen wir als Verrückungen oder Zuwächse solche Vektoren zu, die in jeder der d Koordinatenrichtungen den Betrag 1 haben; jede solche Verrückung sei gleich ws., d.h. der Zuwachs zu jedem Zeitpunkt ist gleichverteilt auf der Menge aller 2^d Vektoren mit Koordinaten ± 1 . Außerdem seien die Zuwächse zu verschiedenen Zeitpunkten unabhängig, was wir oben mit dem Wort "zufällig" angedeutet haben. Ein solches Modell für eine zufällige Bewegung nennen wir eine Irrfahrt. (Es hätte auch andere Möglichkeiten der Modellierung gegeben, z.B. die Gleichverteilung auf der Menge

aller 2d Vektoren mit genau einer Koordinate ±1 und den restlichen Koordinaten gleich 0 , die sog. "Nächste-Nachbar-Irrfahrt". Im Fall $d = 2$ sind beide Modelle bis auf eine triviale affine Transformation identisch: der "Stern" der möglichen Zuwächse bildet beidemale ein Quadrat. Das hier gewählte Modell hat für $d>2$ die schöne Eigenschaft, daß die d Koordinaten der Irrfahrt in ihrem gesamten Verlauf voneinander unabhängig sind.)

Wenn die Position des Teilchens zur Zeit 0 der Ursprung war, dann gilt für die Position $S(n) = (S_1(n), .. , S_d(n))$ nach n Schritten: $(S_i(n) + n)/2$ ist nach $B(n, 1/2)$ verteilt, unabhängig für verschiedene Indices i . Im Limes großer Schrittzahlen n ist im "typischen" Bereich von Gitterplätzen (der langsamer wachsen möge als $n^{2/3}$) für jede Koordinate der lokale Grenzwertsatz gültig. Um die Ws. eines d-Tupels x zu bekommen, multipliziert man die Gewichte der einzelnen Koordinaten x_i . Dies führt auf

(8) $$P(S(n) = x) = \begin{cases} (2/n\pi)^{d/2} \exp(-\sum x_i^2/2n)(1 + o(1)), & \text{wenn alle } x_i - n \text{ gerade sind,} \\ 0 & \text{für alle anderen } x . \end{cases}$$

Schaut man in einem räumlich und zeitlich groben Maßstab auf diese Gewichtsverteilung, dann heißt (8) approximativ soviel wie : die W.-Dichte der Position des diffundierenden Teilchens nach der Zeit t ist $f(t,x): = (2t\pi)^{-d/2} \exp(-\sum x_i^2/2t)$. Dieses $f(t, .)$ entsteht, wenn man die mehrdimensionale Gaußische Standarddichte $(2\pi)^{-d/2} \exp(-\sum x_i^2/2)$, welche das Produkt der eindimensionalen Standarddichte an den Stellen x_i ist, um den Faktor $\sqrt{t}$ auseinanderzieht. Man rechnet leicht nach, daß f als Funktion von Raum und Zeit die Gleichung

(9) $$df/dt = 1/2 \cdot \Delta f = 1/2 \cdot \sum_i (d/dx_i)^2 f,$$

erfüllt. Diese Gleichung, genannt Wärmeleitungsgleichung, spielt eine fundamentale Rolle in der Analysis und mathematischen Physik.

Die hier angestellten Überlegungen beziehen sich nur darauf, wie $S(n)$ bei festem n verteilt ist; wir fragen nicht danach, wie die $S(n)$, $n \geq 0$, gemeinsam verteilt sind oder wie ein ganzer Pfad $(0, S(1), S(2) , ... , S(n))$ aussieht, auf wie geraden oder krummen (zufälligen!) Wegen das Teilchen zu seiner Endposition $S(n)$ gelangt. Fragen dieser Art gehören in ein eigenes Kapitel über Irrfahrten.

Beispiel 3. Eine didaktische Bemerkung zum Gesetz der großen Zahlen.

Wenn ein Lehrer seinen Schülern das Gesetz der großen Zahlen erfahrbar machen will, wird er etwa so vorgehen: Er läßt eine Folge $X(i)$, $i = 1 .. n$, von u.i.v. ZVen simulieren, mit n sehr

groß; dann wird er die Partialsummen $S(k) := \sum_{1 \le i \le k} X(i)$ bilden und die Folge $S(k)/k$, $k = 1 .. n$, graphisch darstellen, mit k als Abszisse. Im Normalfall stellt sich anschließend das gewünschte Aha-Erlebnis bei den Schülern ein: die Oszillation im Graphen von $(S(k)/k)$ nimmt ab, je weiter man nach rechts geht; die Folge scheint tatsächlich zu konvergieren. (Aus einem nicht ganz verständlichen Grund wird in vielen Schulbüchern "konvergieren" vermieden und durch "sich stabilisieren" ersetzt.)

Ist dieses Vorgehen korrekt, d.h. liefert das Bild der rechts "sich stabilierenden" Kurve ein wahres, überzeugendes Argument für die behauptete Gesetzmäßigkeit des Zufalls (natürlich modulo der Endlichkeit von n), oder haben wir es mit einem Scheinargument zu tun? Wir wollen die Frage im Spezialfall diskutieren, daß $X(i)$ die Werte 1 (mit Ws. p) und 0 (mit Ws. $1-p$) annimmt, weswegen die Diskussion im Kapitel über Binomialverteilungen ihren Platz findet.

Die Frage stellen heißt sie verneinen. Wir konstruieren Folgen T und U , so daß gilt

(10) $T(k) \le S(k) \le U(k)$ für alle k , $T(n) = S(n) = U(n)$,
$k \to T(k)/k$ ist wachsend und $k \to U(k)/k$ ist fallend.

Das ist leicht zu machen. Man ordnet die $X(i)$ in aufsteigender Reihenfolge, erst alle Nuller und dann die Einser; die k-te Partialsumme der umgeordneten Folge nennt man $T(k)$. Die Folge U wird analog erzeugt, indem man die $X(i)$ absteigend anordnet. Die Folgen $T(k)/k$ und $U(k)/k$ könnte man als "konvergent" ansehen; sie sind monoton und ihre Zuwächse bei n gehen nach null, da sie von der Ordnung $1/n$ sind. In diesem Bild ist auch das zwischen ihnen liegende $S(k)/k$ als "konvergent"anzusehen. Die Pointe dabei ist: um diese "Konvergenz" zu erzielen, haben wir nichts über die $X(i)$ voraussetzen müssen, nur ihre Beschränktheit. Da wir die u.i.v.-Voraussetzung in keiner Weise benutzt haben, können wir auch nichts über das Wesen des Zufalls (oder die Gesetzmäßigkeit des Chaos, oder welch schönen Worte man immer gebrauchen will) einsichtig gemacht haben.

Versuchen wir, aus dieser Aporie herauszukommen. Eine einfache Möglichkeit besteht darin, den in der Simulation benutzen Parameter p den Schülern bekanntzugeben und sie sich davon überzeugen zu lassen, daß $S(n)$ nahe (bis auf $O(\sqrt{1/n})$ bei p liegt. Das wäre in der Tat eine ehrliche empirische Nachprüfung des schwachen Gesetzes; dazu bedürfte es allerdings keiner Graphik und das Wort vom "Stabilisieren" von S wäre überflüssig, da man nur auf den rechten Endpunkt des Graphen schaut. Wir suchen aber lieber nach einem Weg, wie man die gesamte Folge S in die Argumentation einbeziehen kann, nicht nur $S(n)$.

Hier ist ein anderer Vorschlag. Ein Charakteristikum des Gesetzes der großen Zahlen ist seine Reproduzierbarkeit: wenn man zwei Folgen der Länge n bildet, sind beide arithmetischen Mittel

mit hoher Ws. nahe bei p , also auch nahe beieinander. Die letzte Beobachtung ist dann wichtig, wenn p unbekannt ist. Wenn wir also die simulierte Folge X der Länge n in zwei Teilfolgen der Länge $n/2$ zerschneiden, und die arithmetischen Mittel der beiden Teilfolgen betrachten, dann sollten diese beiden Zahlen relativ nahe, bis auf ein $O(1/\sqrt{n})$, beieinander liegen. Ebensogut könnte man sagen, daß $2\cdot S(n/2)/n$ und $S(n)/n$ nahe beieiander liegen sollen. Um sich hiervon zu überzeugen, bedarf es keiner Kenntnis von p . Ein graphische Visualisierung dieses Sachverhalts ist die folgende: man zeichnet nicht die Folge der arithmetischen Mittel, sondern die Folge der Partialsummen $S(k) = \sum_{1 \le i \le k} X(i)$, $k = 0 .. n$, und verbindet die Punkte $(k,S(k))$; den so erhaltenen Polygonzug vergleicht man mit der Geraden $k \to k\cdot S(n)/n$, $k = 0 .. n$. Wenn man den Maßstab so wählt, daß das Zeichenfeld gut ausgefüllt ist, von links unten nach rechts oben, wird man sich (und die Schüler) davon überzeugen, daß Ge-rade und Polygonzug nahe beieinander liegen. "Nahe" heißt, daß der maximale Abstand zwischen Polygonzug und Gerade von der Ordnung $1/\sqrt{n}$ mal der Höhe des Zeichenfelds ist. Will man mit dem Parameter n spielen, so simuliere man verschiedene Graphiken für größenordnungsmäßig verschiedene Werte von n und sehe nach, ob die Übereinstimmung Gerade-Polygonzug mit wachsendem n besser wird.

Wir sind auf dem Umweg über die Didaktik doch noch zu einer Irrfahrtfrage gekommen. Wir stellen uns das Ziel, probabilistisch etwas über den maximalen Abstand zwischen Polygon und Gerade zu sagen. Wir definieren die Zufallsgröße M als maximalen vertikalen Abstand zwischen Gerade und Polygon und fragen, wie M unter der Bedingung, daß $S(n)$ gegeben ist, verteilt ist. Es gibt eine explizite Lösung im Spezialfall eines geraden n und der Bedingung $S(n) = n/2$, welche hier vorgestellt werden soll. Wir lösen das kombinatorische

Problem.

In einer Urne seien m weiße und m schwarze Kugeln. Es wird ohne Zurücklegen $2m$-mal gezogen. Sei $S(k)$ die Anzahl der in den ersten k Ziehungen gezogenen schwarzen Kugeln und $W := \max\{2\cdot S(k) - k , 0 \le k \le 2m\}$. Berechne $P(W \ge c)$ für jedes $c \ge 0$. (Aus Gründen der Ganzzahligkeit ist es bequem, die gesuchte Größe mit 2 zu multiplizieren.)

Lösung. Betrachte die Menge aller Pfade $x = (x(i) , i = 0 .. 2m)$ mit den Eigenschaften: $x(i)$ ist ganz und $x(i) - x(i+1) = \pm 1$ für alle i , $x(0) = 0$. Unter diesen stellen die Pfade mit $x(2m) = 0$ die möglichen Fälle dar (entsprechend der Zusammensetzung der Urne); deren Anzahl ist $C(2m,m)$. Die günstigen Fälle sind die Pfade mit $x(2m) = 0$ und $\max x(i) \ge c$. Wir behaupten, daß die letztere Menge ebenso mächtig ist wie die Menge aller Pfade mit $x(2m) = 2c$. Begründung: wenn ein Pfad einmal das Niveau c erreicht hat, gibt es gleich viele Möglichkeiten für ihn, bis zur Zeit $2m$ nach 0 zurückzukehren als nach $2c$ weiterzusteigen. Die Mächtigkeit der letztgenannten Menge ist aber $C(2m,m+c)$. Damit erhalten wir

(11) $P(W \geq c) = C(2m,m+c)/C(2m,m) = \prod_{0 \leq i < c} (m-i)/(m+c-i)$, gültig für $c \geq 0$.

Der genaue Zahlwert hiervon ist nicht so interessant. Viel wichtiger ist hingegen die Tatsache, daß vermöge des lokalen Grenzwertsastzes (oder der Stirling-Formel) die Asymptotik dieses Ausdrucks im Limes m gegen unendlich und $c = a \cdot \sqrt{2m}$ leicht anzugeben ist:

(12a) $\lim_{m \to \infty} P_{2m}(W \geq a \cdot \sqrt{2m}) = \exp(-2 \cdot a^2)$,

(12b) $\limsup_{m \to \infty} P_{2m}(M \geq b \cdot \sqrt{2m}) \leq 2 \cdot \exp(-8 \cdot b^2)$.

(Die Ungleichung in (12b) rührt daher, daß das links stehende Ereignis die Vereinigung zweier Ereignisse ist, deren Ws. in (12a) berechnet wurde, nämlich $\{\max(S(k) - k/2) \geq b \cdot \sqrt{2m}\}$ und $\{\min(S(k) - k/2) \leq - b \cdot \sqrt{2m}\}$. Diese Ereignisse sind aber nicht disjunkt und daher ist die Summe der Einzelwsen nur eine obere Abschätzung und keine Identität.) Das gefundene Ergebnis ist insofern erstaulich, als es zeigt, daß die maximale Abweichung des zufälligen Polygonzugs von der interpolierenden Geraden von der gleichen Ordnung ist wie die vertikale Abweichung des Endpunkts von seiner Erwartung, nämlich $\sqrt{n}$. Wenn die Höhe des Zeichenfelds von der Ordnung n ist, nimmt man bei großem n den "reskalierten" Pfad wie einen geraden Strich wahr und kann sich so vom Gesetz der großen Zahlen überzeugen. Um auf die Folgen T und U zurückzukommen: bei ihnen ist die maximale Abweichung von der interpolierenden Geraden von der Ordnung n , im oben betrachteten Spezialfall ist sie genau $m/2$ an der Abszisse m . Ihr Graph wird daher makroskopisch nicht als Strich wahrgenommen. - Ohne Beweis sei hinzugefügt, daß die Asymptotik (12) für beliebige "typische" Bedingungen an $S(n)$ gültig ist. Mit dem Herausgreifen des Spezialfalls $S(n) = n/2$ haben wir keine künstliche Asymptotik erzeugt; nur für die exakte Formel (11) braucht man diese spezielle Bedingung. Das zu (11) führende Argument hat den Namen "Spiegelungsprinzip".

Wir wollen, wie oben angekündigt, noch eine andere Folgerung aus der Stirlingschen Asymptotik ziehen. Es geht jetzt nicht mehr darum, die Binomialverteilung $B(n,p)$ im Bereich "mittlerer" Abweichungen, also im Bereich $n \cdot p \pm O(\sqrt{n})$, durch einfachere Ausdrücke zu approximieren, sondern Abschätzungen im Bereich "großer" Abweichungen zu finden, das ist an Stellen $n \cdot (p \pm a)$ mit $a \neq 0$. Da die Wsen im Bereich der großen Abweichungen sehr klein werden, werden wir die Asymptotik ihrer Logarithmen studieren. Es stellt sich heraus, daß die negativen Logarithmen linear in n wachsen und daß die genaue Geschwindigkeit, mit der das geschieht, ausgerechnet werden kann. Wir erinnern an die Abkürzung aus dem Beweis von Satz 3.1:

(13) $I(t,p) := t \log(t/p) + (1-t) \log((1-t)/(1-p))$.

Satz 3.3. (Große Abweichungen) Sei $0 < p < 1$. Für festes t mit $t < p$ (bzw. $t > p$) gilt bei n gegen unendlich

(14) $$\lim_n 1/n \log(\sum_{k<nt} B(n,p;k)) = - I(t,p)$$

(bzw. $$\lim_n 1/n \log(\sum_{k>nt} B(n,p;k)) = - I(t,p)) .$$

.

(Da $I(t,p)$ für $t \neq p$ strikt positiv ist, bedeutet das: die Folge der Wsen von Halbgeraden, die vom Mittelwert $n \cdot p$ einen Abstand $n \cdot a$ mit $a > 0$ besitzen, geht unter $B(n,p)$ exponentiell in n nach null.)

Beweis. Sei o.E.d.A. $t < p$. Zunächst gilt wegen der Monotonie der Gewichte

(15) $$\sum_{k \leq nt} B(n,p;k) \leq (n+1) \cdot t \cdot B(n,p; [nt])$$

(Die eckige Klammer [ ] bezeichnet den ganzzahligen Anteil einer reellen Zahl.) Der Logarithmus der links stehendenSumme unterscheidet sich daher vom Logarithmus ihres größten Summanden höchstens um ein $O(\log n)$, bei Division durch n haben beide Logarithmen das gleiche Limesverhalten. Für die Asymptotik von $\log(B(n,p; [nt]))$ gehen wir zur Formel (4) zurück und machen die gleichen Umformungen wie sie dort folgten, mit dem Unterschied, daß wir $I(t,p)$ nicht nach Taylor entwickeln, sondern unverändert lassen. Man erhält

$$\log B(n,p; [nt]) = -1/2 \log (2\pi t(1-t)n) - n \cdot I(t,p) + o(1) ,$$

somit

$$- n\, I(t,p) - O(\log n) \leq \log B(n,p; [nt]) \leq \log \{\sum_{k \leq nt} B(n,p;k)\} \leq -n \cdot I(t,p) + O(\log n) .$$

Division durch n liefert im Grenzübergang n gegen unendlich die Behauptung.

Wir werden auf Satz 3.3 mehrmals zurückkommen: einmal in Kapitel 5, wenn über die Shannonsche Entropie gesprochen wird, und in allgemeinerer Form in Kapitel 9 bei der sog. Cramérschen oder Bernsteinschen Ungleichung.

Literatur.

[DR] H. Dinges, H. Rost: Prinzipien der Stochastik. Stuttgart: Teubner 1982.

[F] W. Feller: An introduction to probability theory and its applications. Vol. I. 2nd edition. New York: Wiley 1957.

[Fo] O. Forster: Analysis 1, 4. Auflage. Braunschweig: Vieweg 1983.

Kapitel 4. Die Poissonverteilung

Inhalt. Der Poissonische Grenzwertsatz, Beispiele für Poissonische Modellierung. Die Methode der Kopplung von W.-Maßen. Fixpunkte einer Permutation.

Die Sätze des vorigen Paragraphen über das Verhalten der Binomialverteilung bei großem n standen unter der Voraussetzung, daß der Parameter p fest bleibt. Eine andere Asymptotik besteht darin, n und p gleichzeitig so zu verändern, daß n gegen unendlich wächst und $n \cdot p$ konstant bleibt oder einen Limes besitzt. Die elementare Rechnung ergibt:

Sei $\mu > 0$ gegeben. Dann gilt für die Folge der Binomialverteilungen $B(n, \mu/n)$

(1) $$\lim_n B(n, \mu/n ; k) = \exp(-\mu)\, \mu^k/k! \qquad \text{für} \quad k = 0, 1 , 2 \ldots .$$

Def. 4.1. Sei $\mu > 0$. Man nennt die Verteilung mit Gewichten $\exp(-\mu)\, \mu^k/k!$ an den Stellen $k = 0 , 1 , 2 , ..$ die Poissonverteilung zum Parameter μ ; wir werden dafür die Schreibweise $\Pi(\mu)$ verwenden.

Prop. 4.1. Die Verteilung $\Pi(\mu)$ hat den Mittelwert μ und die Varianz μ . Die Familie der Poissonverteilungen ist abgeschlossen unter Faltung, es gilt

(2) $$\Pi(\mu) * \Pi(\lambda) = \Pi(\mu+\lambda) .$$

Beweis. Die erzeugende Funktion von $\Pi(\mu)$ heiße f ; ihre explizite Form ist

(3) $$f(u) = \sum_k u^k \cdot \exp(-\mu) \cdot \mu^k/k! = \exp(-\mu) \cdot \exp(\mu \cdot u) = \exp(\mu \cdot (u-1)) ,$$

womit die zweite Behauptung unmittelbar einsichtig ist. Die erste Behauptung folgt aus der allgemein gültigen Tatsache

(4) wenn f die erzeugende Funktion einer Zufallsvariablen X ist,
dann gilt $EX = f'(1)$ und $EX \cdot (X-1) = f''(1)$

und der Rechnung $var(X) = EX \cdot (X-1) + EX - (EX)^2$.

Eine nach Poisson verteilte Variable zählt nach obiger Herleitung die Gesamtzahl der Erfolge in einem Versuch, der aus vielen unabhängigen identischen Teilen besteht, die jeder eine kleine Erfolgsws. besitzen. Ergebnis (1) ist aber für viele Anwendungssituationen, in denen ein aus unabhängigen Teilen zusammengesetzter Versuch vorliegt, nicht genügend, weil die Voraussetzung der identischen Erfolgsws. zu restriktiv ist. Wir werden unten (Satz 4.1) sehen, daß diese in der Tat fallen gelassen werden kann; es genügt, daß die Teile unabhängig und von kleiner Erfolgsws. sind.

Dazu als Vorbereitung, da wir über Konvergenz von Verteilungen sprechen wollen, etwas über Abstände zwischen W.-Maßen und die Hilfskonstruktion der Kopplung.

Def. 4.2. Seien auf einer endlichen oder abzählbaren Menge I zwei W.-Maße p und q gegeben. Dann bezeichnet man als ihren Variationsabstand die Zahl

(5) $$v(p,q) := \sum_{i\in I} |p(i) - q(i)| .$$

Offenbar hat v die Eigenschaften einer Metrik. *Wegen der Voraussetzung der Abzählbarkeit von* I ist Konvergenz bezüglich v einfach zu charakterisieren.

Prop. 4.2. Einer Folge $(p(n))$ von W.-Maßen konvergiert gegen ein W.-Maß q bezüglich des Variationsabstands genau dann, wenn $(p(n,i))$ gegen $q(i)$ strebt für jedes $i \in I$.

Beweis. Es strebe $(p(n))$ punktweise gegen q . Nun gilt für jedes $J \subset I$

$$\sum_I |p(n,i) - q(i)| \le \sum_J |p(n,i) - q(i)| + \sum_{I\setminus J} q(i) + (1 - \sum_J p(n,i)) \le$$

$$\sum_J |p(n,i) - q(i)| + \sum_{I\setminus J} q(i) + \sum_{I\setminus J} q(i) + \sum_J |p(n,i) - q(i)| = 2\{\sum_J |p(n,i) - q(i)| + \sum_{I\setminus J} q(i)\} .$$

Wählt man hierin J endlich aber groß, wird der zweite Summand ganz rechts beliebig klein; anschließend wählt man n groß genug, um den ersten Summanden klein zu machen. Dies zeigt, daß $(p(n))$ gegen q im v-Abstand strebt. Die Umkehrung ist trivial.

Def. 4.3. Eine Kopplung zweier W.-Maße p und q auf I ist ein W.-Maß r auf $I \times I$, dessen Projektionen oder Marginalmaße mit p bzw. q übereinstimmen, d.h. für welches gilt

$$p(i) = \sum_k r(i,k) \quad \text{und} \quad q(j) = \sum_k r(k,j) \quad \text{für alle } i \text{ und } j .$$

In anderer Sprache : Eine Kopplung von p und q ist eine mögliche gemeinsame Verteilung von zwei Variablen, deren Verteilungen gleich p bzw. q sind. Den Zusammenhang von Kopplung

und Variationsabstand liefert

Prop. 4.3. Für zwei W.-Maße p und q auf einer höchstens abzählbaren Menge I gilt

(6) $$1/2\ v(p,q) = \inf\{1 - \sum_i r(i,i) : r \text{ Kopplung von p und q}\} = \inf\{P(X \neq Y)\},$$

wobei sich das zweite Infimum über alle Räume mit W.-Bewertung P und Variablen X, Y erstreckt, welche unter P die Verteilung p bzw. q haben.

Beweis. Wenn X und Y für sich die Verteilung p bzw. q besitzen, gilt

$$P(X = i) - P(X = Y = i) = \sum_{j \neq i} P(X = i, Y = j),$$
$$P(Y = i) - P(X = Y = i) = \sum_{j \neq i} P(Y = i, X = j).$$

Subtraktion dieser Gleichungen und Summation über i ergibt

$$\sum_i |p(i) - q(i)| = \sum_i |P(X = i) - P(Y = i)| \leq 2 \sum_{(i,j) \text{ mit } i \neq j} P(X = i, Y = j) = 2\{1 - \sum_i r(i,i)\},$$

wobei r die gemeinsame Verteilung von X und Y ist. Das gilt für jede Kopplung r, womit gezeigt ist, daß die linke Seite in (6) kleiner gleich der rechten ist. Für die umgekehrte Ungleichung muß man eine Kopplung r finden, welche Gleichheit in (3) herstellt. Ein mögliches Konstruktionsrezept ist das folgende:

Man definiert X und Y im Einheitsintervall mit Lebesguemaß λ. Zunächst reiht man Teilintervalle der Länge $\min(p(i),q(i))$, $i \in I$, aneinander und setzt auf dem i-ten Teilintervall X und Y beide gleich i. Im verbleibenden Stück der Länge $1 - \sum_i \min(p(i),q(i))$ setzt man X und Y irgendwie fest, so daß am Schluß gilt $\lambda(X = i) = p(i)$ und $\lambda(Y = i) = q(i)$. Das kann man, da man im ersten Schritt für jedes i nur die Masse $\min(p(i),q(i))$ verbraucht hat, welche kleiner als $p(i)$ und $q(i)$ ist. Dann gilt

$$\lambda(X \neq Y) \leq 1 - \sum_i \min(p(i),q(i)) = \sum_i (p(i) - q(i))_+ = 1/2 \sum_i |p(i) - q(i)|.$$

Satz 4.1a. Seien $X(m)$, $1 \leq m \leq n$, unabhängige Zufallsvariable mit Werten 0 oder 1; sei $EX(m) = \mu(m)$ und $\mu = \sum_m \mu(m)$ gesetzt; sei $X = \sum_m X(m)$. Dann gilt

(7) $$\sum_k |P(X = k) - \Pi(\mu,k)| \leq 2 \sum_m \mu(m)^2.$$

Satz 4.1b. Seien für jedes n unabhängige 0-1-Variable $X(m,n)$, $1 \leq m \leq n$, gegeben; sei $\mu(m,n)$

= $EX(m,n)$ und $X(n) = \sum_m X(m,n)$ gesetzt. Unter der Voraussetzung

(8) $\mu := \lim_n \sum_m \mu(m,n)$ existiert und $\lim_n \max_m \mu(m,n) = 0$

konvergiert die Verteilung von $X(n)$ im Variationsabstand gegen $\Pi(\mu)$.

Offenbar ist Version b eine unmittelbare Folge von Version a. Wenn man an quantitativen Aussagen nicht interessiert ist, sondern nur wissen will, ob im Prinzip eine Situation vorliegt, welche eine Poissonische Modellierung rechtfertigt (viele unabhängige Ereignisse mit kleiner Erfolgsws.), genügt es Satz 4.1b im Gedächtnis zu haben. Satz 4.1a in voller Schärfe ist wichtig für manche statistischen Anwendungen, wo die Ersetzung von Binomialvariablen (allgemeiner: Zählvariablen) durch Poissonische Variablen ein beliebter Beweistrick ist; man möchte in Situationen, wo man ihn anwendet, den begangenen Fehler quantitativ kontrollieren.

Beweis von Satz 4.1a. Für den Variationsabstand der Poissonverteilung $\Pi(p)$ und der Verteilung $B(1,p)$, die das Gewicht $1 - p$ auf 0 und p auf 1 legt, gilt die Abschätzung

$$v(\prod(p), B(1,p)) = \{\exp(-p) - (1 - p)\} + \{p - \exp(-p)\cdot p\} + \{1 - \exp(-p)(1 + p)\} =$$

$$= 2\cdot p(1 - \exp(-p)) \le 2\cdot p^2 .$$

Bei geeigneter Kopplung von $X(m)$ mit einer Poissonvariablen $Y(m)$ gleicher Erwartung gilt nach Prop. 4.2 daher $P(X(m) \neq Y(m)) \le \mu(m)^2$. Wenn die Paare $(X(m),Y(m))$, $m = 1 .. n$, unabhängig gewählt werden, so ist $Y = \sum Y(m)$ nach $\Pi(\mu)$ verteilt (Prop. 4.1) und es folgt

$$P(X \neq Y) \le \sum \mu(m)^2 ,$$

und daraus wieder nach Prop. 4.2 die Beziehung (7).

Die Poissonverteilung liefert in vielen Anwendungsbereichen ein brauchbares Modell für eine Zählvariable. Das gilt immer dann, wenn man Grund zur Annahme hat, daß die zu zählenden Ereignisse, deren Gesamtzahl zu modellieren ist, aus einem großen Reservoir von potentiell eintretenden Ereignissen stammen, die jedes für sich aber nur eine kleine Ws. haben, aktuell einzutreten. Dazu einige Beispiele.

In der Demographie. Ein "seltenes" Ereignis ist der Tod eines 100-jährigen Schweizers. Man notiert über mehrere Kalenderjahre hinweg, wieviele Todesfälle 100-jähriger Menschen es in der Schweiz gegeben hat. Jede einzelne Jahreszählung ist als Poissonische Variable anzusehen. Der

Grund: vor 100 Jahren wurden in der Schweiz viele Menschen geboren, in der Größenordnung von 100 000. Jeder einzelne von diesen hat eine sehr geringe Chance, daß er 100 Jahre alt wird, im 101. Lebensjahr stirbt und damit in die heurige Zählung kommt. Ein anderes Beispiel sind die jährlich registrierten tödlichen Unfälle durch Huftritt in der alten preußischen Armee. Auch hier wieder gilt: Begegnungen von Menschen mit Pferden fanden täglich in tausendfacher Zahl statt, pro Jahr vielleicht 1 Million und mehr. Die Ws. daß eine einzelne Begegnung tödlich verläuft, ist äußerst gering, aber im Lauf eines Jahres akkumuliert sich ein gewisses Risiko, das schließlich in einer nicht vernachlässigbaren Zahl von Todesfällen in Erscheinung tritt (siehe nachstehende Tabelle; Quelle [Cz], ursprünglich [B]). Wieder ein anderes Beispiel von diesem Typ ist die Zahl der tödlichen Verkehrsunfälle auf einem bestimmten Straßenabschnitt während eines Jahres.

Tabelle: Todesfälle durch Huftritt in 10 preußischen Korps in den Jahren 1875-94

m	=	0	1	2	3	4	≥ 5
h(m)	=	109	65	22	3	1	0 .

Sei M die Anzahl der tödlichen Unfälle in einem Jahr bei einem Korps. Der Wert von M wurde 200-mal ermittelt (10 Korps über 20 Jahre); die Zahl $h(m)$ zählt, wie oft man dabei $M = m$ gefunden hat. Es gab insgesamt $\sum_m h(m)\cdot m = 122$ Todesfälle bei 200 Beobachtungen. Vergleicht man die Gewichte der Poissonverteilung $\Pi(\rho)$ mit $\rho = 122/200$ mit Zeile 2 der Tabelle (wobei man die drei letzten Einträge jeweils zusammenfaßt), ergibt sich eine sehr gute Übereinstimmung (Chiquadratwert 0.32 bei 2 Freiheitsgraden).

h(m)	=	109	65	22	4	
$200\cdot\Pi(\rho)(m)$	=	108.67	66.29	20.22	4.82	.

(Wir bitten um Nachsicht, daß wir den Begriff des Quiquadratwerts, der die Übereinstimmung von theoretischem Modell und empirischen Daten mißt, hier nicht erklären können.)

In der Unternehmensforschung: Die Zahl der Telefonanrufe, die im Lauf einer Stunde bei einem bestimmten Apparat ankommen. Auch hier gilt wieder: der Betrieb, zu dem der Apparat gehört, hat viele Kunden und damit potentielle Anrufer. Für jeden von diesen besteht eine kleine Ws., daß er während der genannten Stunde mit der an diesem Apparat sitzenden Person etwas besprechen will. Oder: Die Zahl der Schadensfälle, die während eines Jahres von den Versicherten in einer kleinen Stadt ihrer Feuerversicherung gemeldet werden.

In der Physik: Die Anzahl der Impulse, die ein Geigerzähler während eines bestimmten Zeitinter-

valls registriert. Die Probe, die vor dem Geigenzähler liegt, besteht aus sehr vielen Atomen, in der Gegend von 10 hoch 20. Die Ws. daß ein bestimmtes Atom im Zeitintervall der Untersuchung zerfällt und dann im Zählrohr einen Impuls auslöst, ist aber äußerst gering. Wenn die Größe der Probe und das Beobachtungsintervall gut aufeinander abgestimmt sind, kann man mit einer "endlichen" Zahl, sagen wir 5 oder 10, von Impulsen rechnen.

In all diesen Fällen war die Hypyothese der Unabhängigkeit der möglichen Einzelereignisse nicht übermäßig problematisch, obwohl sie in einigen Fällen (Radioaktivität) besser als in anderen (Telefonanrufe) begründbar erscheint. Etwas vorsichtiger wird man bei dem biologischen Beispiel in [F] (S. 153 f.) sein, welches die Anzahlen der Bakterienkulturen in verschiedenen Planquadraten einer Schale, die man einige Zeit hat stehen lassen, als Poissonisch modelliert: es wird schwer zu erklären sein, warum benachbarte Planquadrate dieser Schale sich unabhängig verhalten sollen. Es könnte ja sein, daß die zufällige Verunreinigung, die zum Entstehen einer Bakterienkultur führt, ein "globales" Phänomen ist und alle Planquadrate gleichzeitig betrifft. Wir wollen auf diese Frage noch einmal zurückkommen, wenn wir über die Multinomialverteilung und Poissonmodelle mit räumlicher Struktur sprechen.

Noch eine generelle methodische Bemerkung zur Frage, wann es sich empfiehlt, mit Poisson zu modellieren. Es wird dem Leser aufgefallen sein, daß wir nur solche Beispiele aufgeführt haben, bei denen der Mittelwert der zu beobachtenden (fast-) Poissonischen Zufallsvariablen "nicht zu groß" ist, obwohl eine derartige Restriktion in Satz 4.1a nicht gefordert wird. Wenn man etwa setzt $\mu(i,n) = n^{-a}$, $1 \le i \le n$, mit $1/2 < a < 1$, strebt die Summe der $\mu(i)$ mit wachsendem n nach unendlich, aber die Summe der Quadrate nach null. Das heißt, die Annäherung an eine geeignete Poissonverteilung wird beliebig gut, obwohl der Mittelwert über alle Schranken wächst. Die Poissonische Näherung kann also auch bei großen Zählvariablen sehr gut sein. Was spricht dann dagegen, sie in der Modellierung zu gebrauchen?

Rein mathematisch nichts. Ein praktisches Argument ist jedoch der Umstand, daß es lästig ist, Poissonverteilungen mit großem Parameter zu tabellieren. Und weiter die Tatsache, daß die Asymptotik des lokalen Grenzwertsatzes auch für Poissonverteilungen mit großem Parameter anwendbar ist: man wird in diesen Bereichen dann ohnehin mit der Gaußischen Näherung arbeiten und den spezifisch Poissonischen Charakter der Verteilung ("seltene Ereignisse") nicht mehr wahrnehmen - wenn man nicht eine noch gröbere Optik vorzieht und die Messungen als deterministisch ansieht. (Natürlich gibt es ein Gesetz der großen Zahl für Poissonverteilungen, da die Standardabweichung langsamer wächst als der Mittelwert, $\sqrt{\mu}$ gegen μ.) Letzeres ist der Standpunkt der Versicherer und der Gebrauch, den sie von einer globalen, nicht auf eine kleinere Stadt beschränkten, Schadensstatistik machen. Um das Beispiel der Verkehrstoten zu aufzugreifen: die Zahl der Toten der gesamten BRD während eines Jahres liegt etwa bei 10000 und weist relativ geringe zufallsbedingte Schwankungen auf. Man behandelt diese Zahl praktisch wie eine

deterministische Größe: man diskutiert ihre Zu- oder Abnahme gegenüber dem Vorjahr, welche typischerweise in der Gegend von 3 oder 4 % liegt und versucht, diese mit straßenbaulichen oder gesetzgeberischen oder konjunkturellen Veränderungen zu erklären. Man redet dann aber nicht mehr über den einzelnen Streckenabschnitt, in dem es - zufallsbedingt - in einem Jahr einmal 3 und in einem anderen Jahr 8 Tote gegeben hat.

Wir beenden diesen Paragraphen mit einem Stück kombinatorischer W.-Rechnung, das einfach schön ist. Außerdem stellt es ein Beispiel dafür dar Beispiel, daß eine Summe von 0-1-Variablen (approximativ) Poissonisch sein kann, obwohl die einzelnen Summanden nicht exakt, sondern nur approximativ, unabhängig sind.

Satz 4.2. Es werden n Objekte zufällig permutiert. Es bezeichne X die Zahl der Fixpunkte der Permutation. Dann gilt

(9) $$P(X = k) = (k!)^{-1} \sum_{o \le m \le n-k} (-1)^m/m! \quad \text{für } 0 \le k \le n .$$

Insbesondere strebt im Limes n gegen unendlich die Verteilung von X gegen die Poissonverteilung mit Mittelwert 1 .

Beweis. Wir beweisen (9) für endliches n , womit keinerlei Konvergenzprobleme verknüpft sind. Der Limesaussage folgt dann sofort.

Der erste Schritt ist folgende Beobachtung: seien $Z(i)$, $1 \le i \le n$, irgendwelche 0-1-Variablen und Z ihre Summe; dann gilt

(10) $$E\, Z\cdot(Z - 1)\cdot \ldots \cdot(Z - m+1) = m!\cdot\sum_{i(1)<i(2)\ldots<i(m)} P(Z(i(1)) = Z(i(2) = \ldots = Z(i(m)) = 1) .$$

In der Tat, für jede Realisierung der $Z(i)$ gilt

$$Z(Z-1) \ldots (Z-m+1) = \sum_{i(1), \ldots, i(m)} 1_{\{Z(i(1))= \ldots =Z(i(m)) = 1\}} ,$$

wobei über alle m-Tupel von paarweise verschiedenen Indices summiert wird. Integriert man diese Identität, erhält man (10). (Der Faktor $m!$ rechts in (10) und die Tatsache, daß nur über aufsteigende m-Tupel von Indices summiert wird, kompensieren sich gegenseitig.)

Der zweite Schrit ist elementares Rechnen mit erzeugenden Funktionen, in Verallgemeinerung von (4) : wenn Z eine beliebige Variable mit Werten in den natürlichen Zahlen und wenn G ihre erzeugende Funktion ist, so hat man

(11) $$(d/du)^m\, G(1) = E\, Z\cdot(Z-1)\cdot \ldots \cdot(Z-m+1)\ .$$

Die Ableitung der erzeugenden Funktion von Z ist $\sum_k P(Z=k)\cdot k\cdot(k-1)\cdot \ldots \cdot k(-m+1)\cdot u^{k-m}$; setze hierin $u = 1$. Damit läßt sich durch Taylor-Entwicklung bei $u = 1$ für ein beschränktes Z die erzeugende Funktion mittels der sog. "faktoriellen Momente" ausdrücken :

(12) $$G(u) = \sum_m E\, Z(Z-1)\ldots(Z-m+1)\,(m!)^{-1}\,(u-1)^m\ .$$

Wir kommen jetzt zur speziellen Wahl von Z : sei $Z(i)$ der Indikator des Ereignisses " i ist Fixpunkt" und Z die Anzahl der Fixpunkte. Wir kombinieren (12) mit (10) . Die r.S. in (10) ist im Spezialfall einfach zu berechnen: sie ist 1 für $m \le n$ und 0 sonst. (Man wende die Laplacesche Regel an.) Damit erhalten wir das gewünschte G explizit:

(13) $$G(u) = \sum_{m\le n}(u-1)^m/m!\ .$$

Umsortieren nach Potenzen von u statt $(u-1)$ schließlich ergibt

(12) $$G(u) = \sum_k u^k\,\{\sum_{n\ge m\ge k}(k!(m-k)!)^{-1}(-1)^{m-k}\} = \sum_k u^k\,(k!)^{-1}\{\sum_{j\le n-k}(j!)^{-1}(-1)^j\}$$

und führt nach Koeffizientenvergleich auf die Behauptung (9).

Bemerkung zum Satz. Die Variablen $Z(i)$ sind nicht unabhängig; wie man leicht nachrechnet, gilt $E\,Z(i) = 1/n$ und $E\,Z(i)Z(j) = 1/n(n-1) \ne 1/n^2$ für $i \ne j$.

Literatur.

[B] L. von Bortkiewicz: Das Gesetz der kleinen Zahlen. Leipzig 1898. (Dort S. 22 ff. Bortkiewicz war einer der ersten Autoren, der den Gebrauch einer Poissonvertei lung zur Erklärung empirischer Daten theoretisch begründete.)

[Cz] E. Czuber: Die statistischen Forschungsmethoden. Wien: Seidel 1921. (S. 222)

[F] W. Feller: An Introduction to Probability Theory and its Applications. 2. Auflage. New York: Wiley 1957.

[K] U. Krengel: Einführung in die Wahrscheinlichkeitstheorie und Statistik. Braunsch weig: Vieweg 1988.

Kapitel 5. Die Shannonsche Entropie

Inhalt: Begriff der Entropie, einfache Eigenschaften. Problem des optimalen Codierens, Quellencodierungssatz. Entropie als logarithmische Rate der Anzahl der "wesentlichen" Folgen in einem Produktexperiment. Algorithmus von Huffman. Historisches.

Sei (Ω , A , P) ein diskreter W.-Raum oder, wenn man will, das mathematische Modell für ein Zufallsexperiment. Wir ordnen ihm die Zahl

(1) $$H(P) := H(A , P) := - \sum P(A) \log P(A)$$

zu, wobei die Summation über alle Atome von A geht. (Konvention: $0 \cdot \log 0 = 0$; der Logarithmus sei der zur Basis e , auch wenn für manche Anwendungen die Basis 2 bequemer ist.) Man nennt diese Zahl die Shannonsche Entropie des W.-Raums (Experiments). Da $- \log(.)$ fallend und die Summe über die $P(A)$ für alle Räume gleich 1 ist, wird $H(.)$ zunehmen, wenn man von (Ω , A , P) zu einem "feineren" W.-Raum übergeht, d.h. wenn man Atome von A unterteilt und neue Ereignisse hinzunimmt, wobei man die W.-Bewertung P auf den alten Ereignissen unverändert läßt (siehe (3) unten). Die Entropie ist also ein Maß dafür, wie diffus die Massen $(P(A), A \in A)$ verteilt sind. Sicher könnte man, wenn man nur an einem Diffusivitätsmaß interessiert wäre, die Funktion $\log(.)$ in (1) durch eine beliebige wachsende Funktion ersetzen; der Witz der Definition liegt aber in der (Sub-)Additivität von H bei Produktexperimenten (siehe Prop. 2a/b) und natürlich der Gültigkeit von Satz 1. Zunächst einige elementare Eigenschaften.

Prop. 5.1. Die Shannonsche Entropie genügt der Abschätzung

(2) $$0 \le H(A , P) \le \log n ,$$

wo n die Anzahl der Atome von A ist. Die untere Abschätzung wird zur Gleichheit,wenn $P(A)$ nur die Werte 0 oder 1 annimmt, die obere Abschätzung, wenn $P(A) = 1/n$ für alle $A \in A$ .Wenn $A \subset B$ und $P = Q \mid A$, so gilt

(3) $$H(A , P) \le H(B , Q) .$$

Kein Beweis. Nichttrivial ist allenfalls die obere Abschätzung in (2); sie folgt aus der Jensen-

schen Ungleichung.

Prop. 5.2a. Seien (A , P) und (B , Q) Experimente und sei (C , R) das Produktexperiment, d.h. $C = A \otimes B$ und $R = P \otimes Q$, dann gilt

(4) $$H(C , R) = H(A , P) + H(B , Q) .$$

Beweis. Wenn $A \times B$ ein Atom von C ist, so schreibt man

$$\log R(A \times B) = \log P(A) + \log Q(B) .$$

Diese Identität multipliziert man mit $R(A \times B) = P(A) \cdot Q(B)$ und summiert dann über alle Paare A , B.

Wir fragen, was mit (4) geschieht, wenn man die Beziehung $C = A \otimes B$ beibehält, aber die Forderung der Unabhängigkeit der 1. und 2. Koordinate unter dem Maß R fallen läßt; d.h. man läßt als R ein beliebiges Maß in C zu und wählt P und Q als seine Marginalmaße.

Prop. 5.2b. Unter der obigen Voraussetzung gilt

(5) $$H(C , R) = H(A , P) + \sum_A H(B , R(. \mid A) P(A) .$$

Der zweite Summand rechts erfüllt die Ungleichung

(6) $$\sum_A H(B , R(. \mid A)) P(A) \leq H(B , Q);$$

für die Entropien gilt die Ungleichung

(7) $$H(C , R) \leq H(A , P) + H(B , Q).$$

Man nennt die Größe $\sum_A H(B , R(. \mid A)) P(A)$ die bedingte Entropie von (B , R) unter der Bedingung A ; sie wird manchmal auch geschrieben als $H(B , R \mid A)$. Genauer wäre die Bezeichnung "gemittelte Entropie der bedingten Verteilungen".

Beweis. Aussage (5) folgt aus der Identität $R(A \times B) = R(A) R(B \mid A)$ unter Berücksichtigung von $R(A) = P(A)$. Die Aussagen (6) und (7) sind dann offensichtlich äquivalent. Der Beweis von (6) geht über die Jensensche Ungleichung, wenn man sich klarmacht, daß

$$R = \sum_A R(\,.\,|\,A)\,P(A)$$

als konvexes Mittel der $R(\,.\,|\,A)$ darstellbar und die Funktion $(t_1, \ldots, t_n) \to \sum_i t_i \log t_i$ in R_+^n konvex ist. Das erste ist gerade der "Satz von der totalen Ws.", das zweite ist elementar: in R_+ ist nämlich $t \to t \log t$ konvexe Funktion.

Beispiel 1. Münzwurf.

Der n-fache unabhängige Wurf einer fairen Münze ergibt eine zufällige Folge der Länge n von Symbolen aus einem zweielementigen Alphabet, sagen wir 0 und 1. Jede der 2^n möglichen Folgen hat die gleiche Ws., daher ist die Entropie des Versuchs gleich $n \cdot \log 2$.

Beispiel 2. Ein Ratespiel.

Betrachten wir folgendes Spiel für zwei Personen: die erste denkt sich eine ganze Zahl x mit $0 \le x \le 2^n - 1$, die zweite muß herausbekommen, was x ist. Sie darf dazu Fragen stellen, die sich mit Ja oder Nein beantworten lassen. Wieviele Fragen braucht sie mindestens, um auch im ungünstigsten Fall x eindeutig zu bestimmen?

Es gibt eine Strategie des zweiten Spielers, die mit n Fragen auskommt: Man beginnt mit "Ist x in der oberen Hälfte des Intervalls von 0 bis 2^n-1 ?" und halbiert mit jeder folgenden Frage das mögliche Intervall, in dem sich x aufhalten kann. (Man könnte auch sagen, daß mit jeder Frage eine Ziffer in der Binärentwicklung von x bestimmt wird.) Schwieriger hingegen ist die Frage, ob es eine Strategie gibt, die mit weniger als n Fragen auskommt?

Ein Entropieargument zeigt, daß es eine solche Strategie nicht geben kann. Nehmen wir an, das unbekannte x wäre vom Zufall geliefert worden, nach der Gleichverteilung. Dann ist $n \cdot \log 2$ die Entropie dieses Versuchs. Wenn eine Strategie aus k Fragen gegeben ist, so kann man die Folge der Antworten als Resultat eines Zufallsversuchs (die Antworten erteile ein Schiedsrichter, der x kennt) ansehen. Dieser Zufallsversuch ist ein Produkt aus k Einzelver-suchen mit jeweils 2 Ausgängen. Die Entropie eines Einzelexperiments ist höchstens $\log 2$ (Prop. 1), die Entropie des Gesamtexperiments höchstens $k \cdot \log 2$ (Prop. 2b). Wenn die Ant-worten auf die Fragen x eindeutig festlegen, muß das Gesamtexperiment feiner sein als das Experiment der Wahl von x, seine Entropie also mindestens gleich $n \cdot \log 2$ sein. Folglich ist $k \ge n$. Es gibt keine bessere Strategie.

Beispiel 3. Eine Wägeaufgabe.

Es seien 12 Kugeln gleichen Aussehens gegeben, von denen 11 gleiches Gewicht haben, die zwölfte aber ein davon verschiedenes; sie kann schwerer oder leichter als die übrigen sein. Man hat eine Balkenwaage zur Verfügung, mit welcher man feststellen kann, ob sich in den beiden Waagschalen gleiche Gewichte befinden, oder ob die Ladung in der rechten oder ob die Ladung in der linken Schale schwerer ist. Was ist die Mindestzahl von Wägungen, die man braucht, um zu wissen, welche der Kugeln die von den anderen verschiedene ist und ob sie schwerer oder leichter ist als die anderen Kugeln?

Behauptung: Drei. Obere Abschätzung: es gibt ein Verfahren, das nur 3 Wägungen benötigt. (Finden Sie es!) Untere Abschätzung: ein Entropieargument. Es gibt nämlich 24 mögliche Zustände, in welchen sich das System befinden kann; eine zufällige Wahl des Zustands mit der Gleichverteilung hat die Entropie $\log 24$. Jede Wägung hat 3 mögliche Ausgänge, die Entropie jedes einzelnen Wägeversuchs ist höchstens $\log 3$. Da $2 \cdot \log 3 = \log 9 < \log 24$, kann ein Produktexperiment mit nur 2 Faktoren die Entropie der Zufallswahl nach der Gleichverteilung nicht erreichen.

Wie im Beispiel vorhin gilt: Die Aussage, daß man mit 3 Wägungen tatsächlich auskommt, ist keine Folge eines Satzes über Entropien; man beweist sie konstruktiv, indem man eine Strate-gie angibt. Eine überraschende Erkenntnis gibt es, wenn Sie in Beispiel 3 statt 12 Kugeln de-ren 13 nehmen: Entropiebuchhaltung spricht nicht gegen die Existenz einer Strategie aus 3 Wägungen, wegen $\log 26 \leq 3 \cdot \log 3 = \log 27$; man kann aber "elementar", ohne den Begriff der Entropie zu benutzen, zeigen, daß es eine solche Strategie nicht geben kann.

Beispiel 2 und 3 bilden noch keinen überzeugenden Grund, die Shannonsche Entropie einzuführen: man hätte ebenso gut alles in der Sprache der reinen Kombinatorik (Zählen von möglichen Fällen) formulieren können. Um Shannon voll auszunützen, müssen wir den Bereich der Laplace-Experimente oder Gleichverteilungen verlassen. Das geschieht jetzt, wenn wir Fragen der Codierung behandeln.

Das Problem des optimalen Codierens.

Zunächst ein paar Worte zum Sprachgebrauch. Sei gegeben ein Versuch mit k Ausgängen, welche die Wsen $p(i)$, $1 \leq i \leq k$, tragen. Wir sprechen von einem Code (ausführlicher: binären Code) für diesen Versuch, wenn jedem der n Ausgänge eine endliche 0-1-Folge, ein sog. Codewort, zugeordnet ist, mit der Zusatzbedingung, daß kein Codewort der Anfang eines anderen ist. (Die Zusatzeigenschaft heißt in der Literatur Präfixfreiheit; im vorliegenden Text ist sie *per definitionem* Bestandteil des Begriffs "Code". Sie hat zur Folge, daß die Zuordnung : Aus-

gang → Codewort injektiv ist, was wir sonst eigens hätten fordern müssen. Die Bedeu-tung der Präfixfreiheit ist folgende: man kann in einer ohne Zwischenräume aneinanderge-reihten Sequenz von Codewörtern die Trennungslinien zwischen den einzelnen Wörtern ein-deutig aus der Folge der Symbole 0 und 1 rekonstruieren, sofern man weiß, wo das erste Wort anfängt.) Wenn $x(i)$ das dem i-ten Ausgang zugeordnete Codewort und $n(i)$ seine Länge ist, so heißt $L := \sum p(i)\cdot n(i)$ die mittlere Länge des Codes.

Die Frage, der wir auf den nächsten Seiten nachgehen wollen, lautet: Was ist die minimale mittlere Länge eines Code zu einem gegebenen Experiment und wie findet man einen mög-lichst sparsamen Code?

Satz 5.1. Für die mittlere Länge eines jeden Codes gilt unter den obigen Vorausetzungen

$$L \geq \sum p(i)\ {}^2\log\,(1/p(i)) = (\log 2)^{-1}\, H(p)\ , \tag{8}$$

wobei $H(p)$ die Shannon-Entropie des zugrundeliegenden Versuchs ist. Es gibt einen Code, für welchen gilt

$$L \leq (\log 2)^{-1}\, H(p) + 1\ . \tag{9}$$

Beweis. 1) Wir zeigen zunächst: sei $\{x(1), .. , x(k)\}$ eine beliebige präfixfreie Menge von Binärfolgen und sei $n(i)$ die Länge von $x(i)$, $i = 1 .. k$; dann gilt

$$\sum 2^{-n(i)} \leq 1\ . \tag{10}$$

Sei nämlich $n \geq \max n(i)$. Wir nehmen einen vollständigen binären Baum der Höhe n und repräsentieren in ihm jedes Wort in natürlicher Weise durch einen Knoten, den wir markieren; das Wort 01101 z.B. wird durch den Knoten in Höhe 5 dargestellt, zu dem man von der Wurzel aus gelangt, indem man sukzessive an den Gabelungen links-rechts-rechts-links-rechts geht. Die Präfixfreiheit besagt, daß kein Weg von der Wurzel zu einem markierten Knoten über einen anderen markierten Knoten führt. Wenn also $A(i)$ die Menge aller (von der Wurzel aus aufsteigenden) Wege der Länge n bezeichnet, die durch den i-ten markierten Knoten gehen, $i = 1 .. k$, so sind die $A(i)$ disjunkt. Definiert man das W.-Maß q auf der Menge aller Wege der Länge n als die Gleichverteilung, so folgt

$$\sum q(A(i)) \leq 1\ .$$

Nun ist aber $q(A(i)) = 2^{-n}\, 2^{(n-\, n(i))} = 2^{-n(i)}$, womit die Gültigkeit von (10) bewiesen ist.

2) Für beliebige positive Zahlen $p(i)$, $q(i)$, $1 \le i \le k$, mit $\sum p(i) = 1$ und $\sum q(i) \le 1$ gilt

(11) $$\sum p(i)\cdot\log(p(i) - q(i)) \ge 0 .$$

Durch Übergang von $q(i)$ zu $q(i)/\sum q(j)$ sieht man, daß es genügt den Fall $\sum q(i) = 1$ zu betrachten. Man formt die l.S. in (11) um zu $\sum q(i)\cdot\{p(i)/q(i) \log(p(i)/q(i))\}$, was wegen Jensen größer gleich ist als $t\cdot\log t$ an der Stelle $t = \sum q(i)\{ p(i)/q(i)\} = 1$, d.h. als die Zahl 0 .

3) Sei jetzt ein Code gegeben und $n(i)$ die Länge des i-ten Codeworts. Dann ist (10) erfüllt und man kann (11) anwenden mit $q(i) = 2^{-n(i)}$ und erhält

$$\sum p(i)\cdot\log p(i) + \sum p(i)\cdot n(i)\cdot\log 2 \ge 0 ,$$

was zur Behauptung (8) äquivalent ist.
4) Man beweist (9), indem man die Existenz einer Codierung mit folgender Eigenschaft zeigt:

(E) der i-te Ausgang wird abgebildet auf ein Codewort der Länge $n(i)$,
wobei $n(i)$ definiert ist durch $n(i) \ge {}^2\log(1/p(i)) > n(i) - 1$.

Diese folgt unmittelbar aus der nachstehenden Umkehrung von Beweisteil 1), dem sog.

Lemma von Kraft. Seien $n(i)$, $1 \le i \le k$, natürliche Zahlen, welche der Bedingung (10) genügen. Dann gibt es einen Code, in welchem das i-te Codewort die Länge $n(i)$ hat.

Beweis des Lemmas. Man bringt die $n(i)$ in aufsteigende Reihenfolge. Als erstes Codewort $w(1)$ wählt man das Wort der Länge $n(1)$ das nur aus Nullern besteht. Wenn man die Codewörter $w(j)$, $j < i$, bestimmt hat, definiert man $w(i)$ als das in der lexikographischen Ord-nung (0 kommt vor 1) erste Wort der Länge $n(i)$, welches keine Verlängerung eines bereits vorhandenen $w(j)$ ist. Man muß sich dabei überlegen, ob die Definition realisierbar ist. Das sieht man so ein: Man zählt vor dem i-ten Schritt, wieviele Knoten des binären Baums in der Höhe $n(i)$ ein schon vorhandenes Codewort $w(j)$ verlängern; man kommt auf $2^{n(i)-n(j)}$ Knoten. Nicht verfügbar in Höhe $n(i)$ sind daher $\sum_{j<i} 2^{n(i)-n(j)} = \sum_{j\le i} 2^{n(i)-n(j)} - 1$ Knoten. Diese Zahl ist wegen (10) kleiner oder gleich $2^{n(i)} - 1$. Da es insgesamt $2^{n(i)}$ Knoten in Höhe $n(i)$ gibt, ist mindestens 1 Knoten für den i-ten Schritt verfügbar. (Siehe Zeichnung)

Zeichnung. Konstruktion eines Codes nach dem Lemma von Kraft.
(Gegebene Längen: 2,2,3,3,4,4; gefundene Codewörter : 00 , 01 , 100 , 101, 1100 , 1101.)

Corollar. Jede binäre Codierung eines Experiments, welches aus n unabhängigen Wiederholungen eines einzigen Versuchs mit Ausgangsverteilung (p(1), .., p(k)) besteht, hat eine mittlere Wortlänge pro Wiederholung, die größer gleich $(\log 2)^{-1} H(p)$ ist. Bei hinreichend großem n gibt es Codierungen, welche der Schranke $(\log 2)^{-1} H(p)$ beliebig nahe kommen.

Beweis. Man zerlegt n in Blöcke der Länge M und codiert jeden Block nach Satz 1 derart, daß man die untere Schranke für die mittlere Wortlänge pro Block, $M \cdot (\log 2)^{-1} \cdot H(p)$, um höchstens 1 verfehlt. Der Überschuß über $(\log 2)^{-1} \cdot H(p)$ in der Wortlänge pro Einzelver-such ist dann höchstens 1/M .
Man muß sich in der "Praxis" n als sehr groß vorstellen. Das Modell der unabhängigen Wiederholungen eines Versuchs ist die grobe Beschreibung einer Nachrichtenquelle, welche über einen längeren Zeitraum hinweg sendet. Die Zeit werde so gemessen, daß 1 Versuch pro Zeiteinheit stattfindet. Man fragt, wieviele Buchstaben pro Zeiteinheit man im Mittel mindestens benötigt, um die gesendete Nachricht in ein binäres Alphabet zu übertragen. Satz 1 mit Corollar sagt, daß $(\log 2)^{-1} H(p)$ die minimale Rate (= Anzahl der Buchstaben pro Zeit) ist. Die Wahl der Blocklänge M ist das Ergebnis eines Kompromisses. Wählt man M zu klein, macht sich der bei jeder einzelnen Codierung maximal auftretende Überschuß 1 als 1/M in der Rate störend bemerkbar. Wählt man M zu groß, wird die praktische Berechnung des Codes zu kompliziert, da die Zahl der möglichen Codewörter exponentiell in M anwächst.

Beispiel 4.

Im Fall $(p(1), ...p(k)) = (1/2 , 1/4 , .. 2^{1-k}, 2^{1-k})$ kann man optimal codieren, ohne einen Verlust gegenüber der theoretischen Schranke zu erleiden, da alle Wsen Zweierpotenzen sind. Das "Verfahren von Kraft" (genauer: im Beweis des Lemmas von K.) ergibt die Codewörter

0 , 10 , ... , 11.. .10 ((k-2)-mal Eins) , 11.. .11 ((k-1)-mal Eins),

auf welche man in dieser Reihenfolge die Ergebnisse von 1 bis k abbildet.

Beispiel 5.

Sei (p(1), p(2)) = (2/3 , 1/3) ; als Codewörter bieten sich an 0 und 1 . Die mittlere Wortlänge ist 1.000 , gegenüber $(\log 2)^{-1}$ H(p) = 0.918 . Codiert man einen Block der Länge 3 , so hat man 8 Ausgänge mit Wsen

(8/27, 4/27, 4/27, 4/27, 2/27, 2/27, 2/27, 1/27) .

zu codieren. Eine mögliche Codierung wird durch die Wörter

(00 , 010 , 011, 100 , 101 , 110 , 1110 , 1111)

beschrieben. Die mittlere Länge dieses Codes ist 76/27 , was einer Rate von 76/81 oder 0.938 entspricht. Das bedeutet eine gewisse Verbesserung gegenüber der Codierung zur Blocklänge 1 . Beide Codierungen sind besser als das "Verfahren von Kraft". Dieses hätte im ersten Fall auf (0 , 10) , im zweiten auf (00 , 010 , 011, 100 , 1010 , 1011 , 1110 , 11110) geführt und damit die Raten 4/3 = 1.333 bzw. 81/81 = 1.000 ergeben.

Qualitativ ist klar: wenn man die mittlere Wortlänge minimieren will, muß man seltene Ausgänge auf lange und häufige Ausgänge auf kurze Codewörter abbilden. Die Wahl $^2\log(1/p(i))$ als Länge des Codeworts für einen Ausgang der Ws. p(i) wäre bei einer kontinuierlich veränderlichen Wortlänge optimal; für die Ganzzahligkeit ist ein Preis zu zahlen, der aber für jedes Wort unter 1 gehalten kann. Beispiel 5 zeigt, daß man mit etwas Intelligenz den Fehler in günstigen Fällen wesentlich geringer halten kann. Es gibt sogar einen relativ einfachen Algorithmus, der eine unter der Bedingung der Ganzzahligkeit optimale Lösung liefert, den sog. Huffman-Code. (Siehe [Gu],[Ga] oder unten im Anhang am Ende des Kapitels.) Das MorseAlphabet ist in sehr groben Zügen nach diesem qualitativen Prinzip konstruiert: man codiert die 26 Buchstaben des lateinischen Alphabets in 0-1-Wörter der Länge höchstens 4 , wobei man je nach der Häufigkeit der einzelnen Buchstaben im Englischen längere oder kürzere Codewörter wählt; der Morse-Code ist aber nicht präfixfrei und fällt insofern aus dem Rahmen.

Wir wollen im Rest des Kapitels die Aussage von Satz 5.1 aus einer leicht veränderten Perspektive betrachten. Sei gegeben ein Versuch mit k Ausgängen, deren Wsen p(1) , .. , p(k) seien; die Menge der Ausgänge bezeichnen wir mit E . Der Versuch werde n-mal unabhängig ausgeführt. Dann ist E^n , die Menge aller Folge der Länge n mit Symbolen aus E , die Ergebnismenge des Produktversuchs. Die W.-Bewertung auf E^n werde mit P bezeichnet; sie ist das n-fache Produkt der Bewertungen p = (p(i) , i = 1 .. k) . Uns interessiert die folgende

Frage: Sei $1/2 > \delta > 0$ gegeben; sei $N(n,\delta)$ das Minimum der Zahlen $|A|$, wo A alle Teilmengen von E^n mit $P(A) \geq 1 - \delta$ durchläuft. Wie verhält sich $N(n, \delta)$ für großes n ?

(Mit $|A|$ bezeichnen wir die Kardinalität von A .) In Prosa: Man könnte eine Menge A "wesentlich" nennen, wenn ihr Komplement eine vorgeschrieben kleine Ws. hat; man fragt nach der minimalen Kardinalität einer "wesentlichen" Menge, oder wieviele Elementarereignisse man mindestens braucht, um den W.-Raum (E^n, P) einigermaßen "treu" darzustellen.

Satz 5.2. Es gilt bei fester W.-Bewertung p und $\delta < 1/2$ die Asymptotik

(12) $$\lim_n 1/n \cdot \log N(n,\delta) = H(p) = - \sum p(i) \cdot \log p(i) .$$

Beweis. Wir geben uns Zahlen $r(1) , .. , r(k)$ mit Summe n vor und fragen danach, wieviele Folgen der Länge n es gibt, in denen genau $n(i)$-mal das i-te Element aus E steht. Deren Zahl ist natürlich gleich dem Multinomialkoeffizienten

$$M(n; r(1) , .. , r(k)) = n! \cdot \{\prod r(i)!\}^{-1} .$$

Die Asymptotik dieses Koeffzienten wird nach der Stirlingformel berechnet. Das geht wie bei Formel (4) und (4a) in Kapitel 3; das Ergebnis ist (in für uns ausreichender Feinheit)

(13) $$\log M(n; r(1) , .. , r(k)) = n \cdot H(r(1)/n , .. , r(k)/n) + O(\log n) .$$

Eine leichte Modifikation hiervon erhält man, wenn man über mehrere Häufigkeitsvektoren $(n(i))$ summiert, da ja die Anzahl aller möglichen Häufigkeitsvektoren nur polynomial in n zunimmt:

(13a) $$\log \sum_{r/n \in U} M(n; r(1) , .. , r(k)) \leq n \cdot \sup \{H(s) : s \in U\} + O(\log n)$$

für alle Teilmengen U des Simplex $S := \{ s \in R^k : s(i) \geq 0 , \sum s(i) = 1 \}$. Nun sagt das schwache Gesetz der großen Zahlen, daß bei n-facher unabhängiger Ausführung des Versuchs für jedes einzelne i die relativen Häufigkeiten des i-ten Ausgangs mit hoher Ws. beliebig nahe bei $p(i)$ liegen; daraus folgt, daß mit beliebig hoher Ws. dieses Naheliegen auch simultan für alle i gilt, d.h. daß man hat für das k-Tupel $R = (R(i) , i = 1 .. k)$ der absoluten Häufigkeiten

(14) $$\liminf_n \sum_{r/n \in U} P(R = r) \geq 1 - \delta$$
für jedes $\delta > 0$ und jede Umgebung U von p in S .

Vermöge (14) sieht man, daß eine "wesentliche" Menge von Ergebnissen die Menge aller Folgen ist, bei denen die relative Häufigkeit des i-ten Ausgangs für jedes i nahe bei $p(i)$ liegt; die

Kardinalität dieser Menge ist in (13a) nach oben abgeschätzt; somit hat man

(15) $$\limsup_n 1/n \cdot \log N(n,\delta) \le H(p) .$$

Die untere Abschätzung sieht man so ein: für jedes $x \in E^n$ mit Häufigkeitsvektor r gilt

(16) $$P(x) = \prod_i p(i)^{r(i)} = \exp(n \cdot \sum_i r(i)/n \cdot \log p(i)) .$$

Hierin ist die rechte Seite kleiner gleich $\exp(-n\cdot(H(p) - \epsilon))$, wenn r/n in einer geeigneten Umgebung U von p liegt. In jeder "wesentlichen" Menge von Folgen gibt es aber eine Teilmenge der Ws. mindestens $1 - 2\delta$, so daß alle Folgen dieser Teilmenge eine relative Häufigkeit in U haben. Die Mächtigkeit dieser Teilmenge ist mindestens $(1-2\delta)\cdot \exp(n\cdot(H(p) - \epsilon))$. Daraus folgt die gewünschte Ungleichung, welche den Beweis vollendet:

(17) $$\liminf_n 1/n \cdot \log N(n,\delta) \ge H(p) .$$

Im Licht von Satz 5.2 kann man den Quellencodierungssatz etwa so verstehen: wenn die Quelle mit W.-Bewertung p während der Dauer von n Zeiteinheiten sendet, ist die Zahl der in Betracht zu ziehenden "wesentlichen" Buchstabenfolgen gleich $N(n,\delta)$ oder ungefähr gleich $\exp(n\cdot H(p))$. Um die wesentlichen Folgen eineindeutig in Binärfolgen gleicher Länge L abzubilden, muß L gleich dem Zweierlogarithmus der letzten Zahl sein, also $n\cdot H(p)\cdot(\log 2)^{-1}$. Die Anzahl der benötigten Binärzeichen pro Zeiteinheit ist also gleich $H(p)\cdot(\log 2)^{-1}$.
Historische Bemerkung. Die Asymptotik (14) war schon *Boltzmann* bekannt, siehe [B]. Seine Leistung war es, den Begriff der "Entropie", wie ihn die Physiker seiner Zeit als durch den zweiten Hauptsatz der Thermodynamik definiert gebrauchten, mit einer simplen kombinatorischen oder stochastischen Größe in Verbindung zu bringen. Seine Idee bestand darin, die Entropie eines "Makrozustands" zu definieren - bis auf einen Faktor - als den Logarithmus der Anzahl der "Mikrozustände", aus denen sich der erstere zusammensetzt. Im oben besprochenen diskreten Modell heißt das: die Zahl n spielt die Rolle der "Größe" des Systems oder der Anzahl seiner "Komponenten"; ein Mikrozustand ist ein Element von E^n. Als Makrozustand bezeichnen wir das k-Tupel $r = (r(i) , i = 1 .. k)$ der Häufigkeiten der verschiedenen möglichen Ausprägungen einer Komponente. Der Multinomialkoeffizient $M(n; r(1), .. , r(k))$ zählt die zu einem Makrozustand gehörenden Mikrozustände und sein Logarithmus ist - bis auf einen Faktor - die diesem Makrozustand zugeordnete "Entropie" im Sinn Boltzmanns. Da diese Zahl proportional zu n wächst, bei festen relativen Häufigkeiten r/n , kann man für große Systeme die spezifische Entropie, den Limes des Quotienten Entropie durch Systemgröße definieren, der wegen (13) gleich $H(p)$ ist, wenn r/n gegen p geht. Insofern ist die Wortwahl Shannons in perfekter Übereinstimmung mit dem physikalischen Gebrauch des Wortes Entropie und keine willkürliche Neu-

schöpfung. - Die Mathematisierung von Nach-richtenübertragung und Codierung, der Begriff der Entropie als Maß für die Diffusivität eines Experiments und die damit verbundenen Sätze dieses Kapitels, all das geht auf Arbeiten von *C. Shannon* in den 40er-Jahren zurück, siehe [S].

Anhang: Der Codierungsalgorithmus von Huffman.

Gegeben sei eine W.-Bewertung $p = (p(1), .., p(k))$. Gesucht ist ein binärer Code minimaler mittlerer Länge. Wir skizzieren zunächst den Algorithmus, der von Huffman vorgeschlagen wurde, anhand des nachstehenden Zahlenbeispiels; anschließend argumentieren wir, warum er optimal ist.

Algorithmus. Sortiere die Ausgänge nach wachsender Ws. und etikettiere sie von 1 bis k . Schreibe Wsen und Etiketten auf. Addiere die zwei kleinsten Wsen, fasse diese Ausgänge als 1 Block zusammen, der als neues Etikett beide Etiketten seiner Bestandteile erbt. Sortiere neu. Wiederhole das Verfahren, bis alle Ausgänge in einem einzigen Block verschmolzen sind. -

Zur Tabelle: wir starten mit 6 Zahlen in der ersten Zeile. Wir haben in jeder Zeile das Ergebnis jeweils eines Verschmelzungsschritts festgehalten. Die Einträge sind Etikett und zugehöriger W.-Wert. Um das Codewort für den Ausgang Nr. i zu finden, verfolgt man den Weg der Ziffer i von unten nach oben. Jedesmal wenn der Block, zu dem i gehört, sich teilt, schreibt man 0 , wenn i in den linken, und 1 , wenn i in den rechten Teilblock fällt. Die so erhalte-ne 0-1-Folge ist das Codewort für i . (In jeder Zeile wurde ein Etikett fett gedruckt, und zwar das Etikett des Blocks, der den ursprünglichen Ausgang Nr. 1 enthält.)

1 .05 2 .08 3 .12 4 .19 5 .24 6 .32

3 .12 **12** .13 4 .19 5 .24 6 .32

4 .19 5 .24 **123** .25 6 .32

123 .25 6 .32 45 .43

45 .43 **1236** .57

123456 1.00

Es folgt die Liste sämtlicher Codewörter dieses Beispiels. Die mittlere Codelänge ist $L = 2.38$, die theoretische Schranke ist $(\log 2)^{-1} H(p) = 2.35$.

Ziffer	1	2	3	4	5	6
Codewort	1010	1011	100	00	01	11

Beweisskizze für die Optimalität. Notwendig für die Optimalität ist, daß die Codewörter der beiden seltensten Ereignisse (deren Ws. sei $p(1)$ und $p(2)$) gleich lang sind. Man wählt den optimalen Code daher so, daß sich diese beiden Codewörter nur in der letzen Ziffer unterscheiden. Man faßt diese beiden Ereignisse zusammen und löst das Problem des optimalen Codierens im vergröberten Experiment, das $k-1$ anstelle von k Ausgängen hat; die optimale mittlere Codelänge im ursprünglichen Experiment ist um $p(1)+p(2)$ länger als im vergröberten Experiment. Der Reduktionsschritt wird iteriert; das geschilderte Verfahren macht nichts weiter, als daß es die einzelnen Schritte nachzeichnet.

Literatur.

[B] L. Boltzmann: Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektive den Sätzen über das Wärmegleichgewicht (1877). In: Werke, hrsg. von F. Hasenöhrl, Bd. II, S. 164-223. Leipzig: J.A. Barth 1909.

[Ga] R.G. Gallager: Information and reliable communication. NewYork: Wiley 1968.

[Gu] S. Guiasu: Information Theory with applications. New York. McGraw-Hill 1977.

[S] C.E. Shannon: A mathematical theory of communication. Bell Systems Techn. Journal 27, 379-423 und 623-656. (Nachdruck in Buchform existiert.)

Kapitel 6. Der Zentrale Grenzwertsatz

Inhalt: Der Zentrale Grenzwertwatz für Summen von u.i.v.Variablen; technische Hilfsmittel: schwache Konvergenz von Maßen und Fouriertransformation. Gaußverteilungen. Fixpunktbetrachtungen.

Das erste Ziel ist, den Begriff der schwachen Konvergenz von W.-Maßen (Verteilungen) in R^d einzuführen. Daher ist zunächst zu sagen, was ein W.-Maß in R^d ist. Wir wollen dabei nicht abstrakte Maßtheorie treiben, sondern gewisse analytische Grundlagen der Stochastik wenigstens an einer Stelle sichtbar machen.

Def. 6.1. Ein W.-Maß in R^d ist eine Linearform $\mu : b \to \int b d\mu$ auf den stetigen beschränkten Funktionen auf R^d mit den Eigenschaften

(1) $\int b d\mu \geq 0$ für $b \geq 0$, $\int 1 d\mu = 1$, $\int (\lim_n b_n) d\mu) = \lim_n \int b_n d\mu$
für jede aufsteigende Folge (b_n) mit stetigem Limes.

Die Verteilung der R^d-wertigen ZV X ist das W.-Maß μ mit

(2) $\int b d\mu = E\, b(X)$ für alle b stetig, beschränkt.

(Die zweite Bedingung in (W), die Normierung auf Gesamtmasse 1 , ist der Grund für den Zusatz "Wahrscheinlichkeits-" vor dem Wort "Maß".)

Ohne zusehr ins Detail gehen zu wollen: mit einem W.-Maß im Sinn der Definition ist in kanonischer Weise ein Integral über eine weite Klasse von "anständigen" Funktionen verbunden. Man soll sich ein W.-Maß als eine Massenverteilung in R^d vorstellen. Alle in dieser Vorlesung auftretenden Beispiele von W.-Maßen gehören zu einem der folgenden zwei Typen:

(a) Diskrete W.-Maße. Ein diskretes Maß hat eine höchstens abzählbare Trägermenge $\{x(1) , x(2) , .. \}$. Wenn wir sagen, ein solches μ habe an den Stellen $x(i)$ das Gewicht $p(i)$, oder und dafür in Kurzform schreiben $\mu = \sum p(i) \cdot \delta_{x(i)}$, heißt das formal soviel wie

(3) $\int b d\mu = \sum p(i) \cdot b(x(i))$ für alle b .

Natürlich gilt $\sum p(i) = 1$. Einbegriffen ist der Spezialfall $\mu = \delta_x$ des Deltamaßes bei x .

(b) Absolutstetige W.-Maße. Das sind Maße μ, die sich in der Form

(4) $\int bd\mu = \int b(x)\cdot f(x)\,dx$ für alle b stetig beschränkt

darstellen lassen, wobei $f \geq 0$ und $\int f(x)\,dx = 1$ gilt. Wir sagen im Fall von (4), μ habe die W.-Dichte f und drücken das in Kurzschrift als $d\mu = f(x)\cdot dx$ aus. Die Gaußische Standarddichte ist ein aus Kapitel 3 bekanntes Beispiel einer W.-Dichte. Ein anderes ist die uniforme Dichte im Intervall (a,b) : $f(x) = (b-a)^{-1}\cdot 1_{\{a<x<b\}}$. (Weitere Beispiele folgen im Text.)

Def. 6.2. Das W.-Maß ρ heißt Faltung von μ und π, geschrieben als $\rho = \mu * \pi$, wenn es die Verteilung von $X + Y$ ist, wo X und Y unabhängig sind mit Verteilung μ bzw. π. Wenn μ, π, ρ die Dichten f bzw. g bzw. h haben, ist das gleichbedeutend mit

(5) $h(x) = \int f(x-y)\cdot g(y)\,dy$ (Integration über R^d)

Def. 6.3a. Der Variationsabstand zweier W.-Maße μ und π in R^d ist definiert als

(6) $v(\mu,\pi) = \sup\{\int bd\mu - \int bd\pi : |b| \leq 1\}$

($= \int |f - g|\,dx$ für $d\mu = f\cdot dx$ und $d\pi = g\cdot dx$). Konvergenz im Sinn des Abstands v nennt man starke Konvergenz.

Der Variationsabstand entspricht dem, was wir in Def. 4.2 für einen diskreten Raum definiert haben. Wir haben diesen Begriff hier nur aus Gründen der Vollständigkeit und um des Kontrastes willen erwähnt. Die entscheidende Rolle in diesem Kapitel spielt nämlich die sog. schwache Konvergenz. Für sie ist die Metrik des zugrundeliegenden Raums R^d wesentlich.

Def. 6.3b. Wir sagen, eine Folge von W.-Maßen $(\mu(n))$ in R^d konvergiere schwach gegen ein W.-Maß μ, wenn gilt

(7) $\int bd\mu = \lim_n \int bd\mu(n)$ für alle stetigen beschränkten b .

Bemerkung. Man überlegt sich, wegen (1), daß (7) gleichbedeutend ist mit

(7a) $\int bd\mu = \lim_n \int bd\mu(n)$ für alle stetigen b mit kompaktem Träger.

Auch wenn wir darauf verzichten, schwache Konvergenz zu metrisieren und hier nur von Folgenkonvergenz

sprechen werden, sollte man sich doch ein grobes "metrisches" Bild dieser Konvergenzart machen. Eine Massenverteilung μ ist im schwachen Sinn nahe bei einer Massenverteilung π, wenn die Transportkosten, das Produkt Masse mal Weg, für eine Überführung von μ nach π gering sind. In anderer Sprache: wenn man den Wert der Zufallsvariablen X nur mit einer gewissen, aber kontrollierbaren, Ungenauigkeit feststellen kann und wenn man mit Y das Ergebnis der unscharfen Messung von X bezeichnet, dann ist die Verteilung von Y schwach benachbart zur Verteilung von X. Etwas formeller:

Prop. 6.1. Wenn $X(n)$ stochastisch gegen X strebt, dann strebt die Verteilung von $X(n)$ schwach gegen die Verteilung von X.

Beweis. Sei b stetig mit kompaktem Träger und $0 \le b \le 1$. Dieses b ist gleichmäßig stetig. Es gibt also zu jedem $\epsilon > 0$ ein $\delta > 0$ so daß $P(|b(X) - b(X(n))| > \epsilon) \le P(|X - X(n)| > \delta)$. Wenn die rechte Seite für große n kleiner als ϵ wird, folgt $|Eb(X) - Eb(X(n))| \le 2\epsilon$.

Es ist offensichtlich, daß starke Konvergenz die schwache impliziert; unser Sprachgebrauch hat also das nötige Minimum an innerer Logik. Man kann Prop. 4.1 in der Sprache des jetzigen Kapitels so ausdrücken: "Für W.-Maße, die von den ganzen Zahlen getragen werden, sind schwache und starke Konvergenz dasselbe." Einfache Gegenbeispiele zeigen, daß das falsch wird, wenn man den Bereich diskreter Trägermengen verläßt: wenn z.B. die Punktfolge $(x(n))$ gegen x strebt und alle $x(n)$ verschieden von x sind, dann strebt $(\delta_{x(n)})$ im schwa-chen, aber nicht im starken Sinn, gegen δ_x. - Das Haupthilfsmittel dieses Kapitels ist die Fouriertransformation von W.-Maßen.

Def. 6.4. Wenn μ W.-Maß in R^d ist, dann nennt man die in R^d definierte Funktion

$$t \to \mu\hat{}(t) := \int \exp(it\cdot x)\, \mu(dx) \quad (= E \exp(it\cdot X)) \tag{8}$$

die Fouriertransformierte von μ (oder von X, wenn μ die Verteilung von X ist).

Prop. 6.2. Wenn $\pi = \mu * \nu$, so hat man $\pi\hat{}(t) = \mu\hat{}(t)\cdot\nu\hat{}(t)$ für alle t. In Worten: Der Faltung der Verteilungen entspricht die Multiplikation der Fouriertransformierten.

Beweis. Geht wie bei den erzeugenden Funktionen in Kapitel 2. Beruht auf der Identität

$$\{E \exp(it\cdot X)\}\cdot\{E \exp(it\cdot Y)\} = E \exp(it\cdot(X+Y)) \text{ für } X, Y \text{ unabhängig}. \tag{9}$$

Satz 6.1. Seien $\mu(n)$, $n \ge 1$, und μ W.-Maße in R^d. Dann sind gleichbedeutend

(i) $(\mu(n))$ strebt schwach gegen μ und

(ii) $\lim_n \mu(n)\hat{}(t) = \mu\hat{}(t)$ für alle t .

Beweis (für $d = 1$). Sei f stetig mit kompaktem Träger. Wir setzen für jedes $a > 0$

(10) $$f_a(y) := (a\pi)^{-1} \int f(x)(1 + ((x-y)/a)^2)^{-1}\, dx .$$

Dann gilt für jedes W.-Maß ρ die Identität

(11) $$(2\pi)^{-1}\int\int \rho\hat{}(t) \exp(-itx - a|t|) f(x)\, dt\, dx = \int \rho(dy)\, f_a(y) .$$

In der Tat: die linke Seite wird der Reihe nach umgeformt zu

$$\int\int\int (2\pi)^{-1} \rho(dy) \exp(it(y-x) - a|t|) f(x)\, dt\, dx = (a\pi)^{-1} \int\int \rho(dy) f(x)(1 + ((x-y)/a)^2)^{-1}\, dx =$$

$$\int \rho(dy)\, f_a(y) .$$

Hier haben wir die leicht nachzurechnende Tatsache benutzt, daß

$$a/2 \int \exp(itz)\exp(-a|t|)\, dt = (1 + (z/a)^2)^{-1} .$$

Damit sieht man die Gültigkeit der Implikation (ii) → (i) ein: für jedes a folgt aus (ii) - nach dem Lebesgueschen Satz von der dominierten Konvergenz - die Konvergenz von

$$(2\pi)^{-1}\int\int \mu(n)\hat{}(t) \exp(-itx - a|t|) f(x)\, dt\, dx \quad \text{gegen} \quad (2\pi)^{-1}\int\int \mu\hat{}(t) \exp(-itx - a|t|) f(x)\, dt\, dx$$

und daher wegen (11) von $\int \mu(n)(dy)\, f_a(y)$ gegen $\int \mu(dy)\, f_a(y)$. Da f_a gleichmäßig gegen f strebt bei a gegen 0 , hat man auch

$$\lim_n \int \mu(n)(dy)\, f(y) = \int \mu(dy)\, f(y) ,$$

was wegen obiger Bemerkung die schwache Konvergenz von $\mu(n)$ gegen μ bedeutet. - Die Implikation (i) → (ii) ist trivial.

Corollar. Die Fouriertransformierte bestimmt das Maß eindeutig.

Beweis. Folgt aus (11): wenn die Fouriertransformierten zweier Maße gleich sind, haben diese Maße gleiches Integral auf allen f_a , wo f stetig mit kompaktem Träger ist; damit haben sie gleiches Integral auf diesen f und schließlich auf allen beschränkten stetigen Funktionen.

Beispiele von Fouriertransformierten.

1) Die Exponentialverteilung mit Dichte $f(x) = \exp(-x)\cdot 1_{\{x>0\}}$ hat als FTr die Funktion $t \to (1 - it)^{-1}$.
Die zweiseitige Exponentialverteilung $g(x) = 1/2 \exp(-|x|)$ hat $t \to (1 + t^2)^{-1}$, die im Maßstab veränderte Dichte $a\cdot g(a\cdot x)$ entsprechend die FTr $t \to (1 + (t/a)^2)^{-1}$. Diese Tatsache wurde obigen Beweis bereits benutzt.

2) Die Cauchy-Verteilung mit Dichte $\pi^{-1}(1 + x^2)^{-1}$ hat als FTr $t \to \exp(-|t|)$. (Beweis über den Residuensatz.)

3) Die Gaußsche Standarddichte $\varphi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ hat die FTr $t \to \exp(-t^2/2)$. (Beweis ebenfalls durch Integration im Komplexen.)

4) Die symmetrische uniforme Dichte $1/2a \cdot 1_{\{-a<x<a\}}$ hat die FTr $t \to 1/(at)\cdot\sin(at)$.

Def. 6.5. Die d-dimensionale Gaußverteilung (oder Normalverteilung) zur Erwartung a und Kovarianzmatrix C , wo a ein d-Tupel ist und C eine symmetrische nichtnegative Matrix, ist die Verteilung mit der Fourier-Transformierten

(12) $$t \to \exp(i\textstyle\sum_j a(j)t(j) - 1/2 \sum_{j,k} C(j,k)t(j)t(k)) .$$

Sie wird geschrieben als N(a,C).

Man muß sich überlegen, daß (12) überhaupt als FTr eines W.-Maßes auftritt. Da ein d-Tupel X von unabhängigen Gaußischen Standardvariablen die FTr $\exp(-1/2 \sum t(j)^2)$ hat, sehen wir, daß die Funktion (12) im Fall $a = 0$ und $C = \mathbf{1}$ (Einheitsmatrix in R^d) als FTr auftritt. Wendet man auf dieses X die affine Abbildung $X \to a + AX =: Y$ an, dann hat Y als FTr $t \to \exp(i\sum a(j)t(j) - 1/2 \sum C(j,k)t(j)t(k))$ mit $C = AA^T$. Da sich jede nichtnegative Matrix C in der Form $C = AA^T$ darstellen läßt, ist (12) als FTr realisierbar. Man sieht weiter, daß a die Erwartung und C die Kovarianzmatrix von Y ist; insofern ist die Definition vernünftig. - Unmittelbar aus Def. 6.5, der Substitutionsregel für Integrale und Prop. 6.2 folgt

Prop. 6.3. (a) Wenn C nicht-ausgeartet ist, dann hat N(a,C) eine W.-Dichte, nämlich

(13) $$x \to (2\pi)^{-d/2} (\det C)^{-1/2} \exp\{-1/2 \textstyle\sum B(j,k)(x(j)-a(j))(x(k)-a(k))\}, \text{ mit } B = C^{-1} .$$

(b) Die Faltung zweier Gaußverteilungen (gleicher Dimension) ist wieder Gaußisch. Jedes affine Bild einer Gaußischen ZV ist ebenfalls Gaußisch.

Die zweite Aussage der Proposition macht plausibel, warum das Rechnen mit Gaußverteilungen sehr bequem sein kann. Da die Familie aller Gaußverteilungen in R^d durch Erwartung und Kovarianzmatrix indiziert ist, ist die Idee nicht so fernliegend, im Fall eines relativ geringen Wissens über eine zu modellierende d-dimensionale Variable das Modell in der Klasse der Gaußverteilungen zu suchen, wobei man die unbekannten Parameter mit Hilfe der empirischen (Stichproben-) Größen schätzt (X(j,m) ist die j-te Koordinate der m-ten Beobachtung):

(14) $\underline{X}(j) := 1/n \sum_{m \le n} X(j,m)$ als Schätzer für a(j) und

$\underline{C}(j,k) := 1/n \sum_{m \le n} (X(j,m)-\underline{X}(j))(X(k,m)-\underline{X}(k))$ für C(j,k) .

Das schwache Gesetz der großen Zahlen garantiert die stochastische Konvergenz der unterstrichenen Größen gegen die zu schätzenden "wahren" Parameter. In der älteren Statistik war die Gaußische Modellierung sehr beliebt. Die Kritik liegt auf der Hand: warum um alles in der Welt soll man etwas als Gaußisch annehmen, wenn man keinen Anhaltspunkt dafür hat, daß diese Verteilungsklasse die Daten besser als eine andere beschreibt? Bleibt als Rechtfer-tigung, daß die Algebra der Gaußverteilungen sehr bequem ist, da sie gut an lineare Transfor-mationen der Datensätze angepaßt ist.

Def. 6.6. Eine Familie von W.-Maßen in R (μ_a : a > 0) heißt Skalenfamilie, wenn gilt $\mu_a = T_a\mu_1$; hierbei bezeichnet $T_a\mu$ die Verteilung von a·X , wenn μ die Verteilung von X ist. (Im Fourierbild: $(T_a\mu)\hat{}(t) = \mu\hat{}(at)$ für alle t .)

Wenn man an ein "universelles Fehlergesetz" für die Verteilungen von Meßfehlern glaubt, wäre ein geeigneter Kandidat eine unter Faltung abgeschlossene Skalenfamilie. Denn dann hätte die Summe zweier unabhängiger Meßfehler bis auf einen Maßstabsfaktor die gleiche Verteilung wie ihre Summanden. Mit $N(0,\sigma^2)$, $\sigma > 0$, hat man eine solche Familie gefunden. Ist damit der Beweis erbracht, daß man alle Meßfehler als Gaußisch zu modellieren hätte? Nein. Erstens, warum sollten sich Meßfehler aus unabhängigen Quellen im Gesamtergebnis addieren? Die Fehlerfortpflanzung könnte ja auch so vor sich gehen, daß anstelle der Fehler selbst deren Logarithmen oder Wurzeln addiert werden. Zweitens, es gibt noch andere Skalenfamilien als $N(0,\sigma^2)$, welche unter Faltung abgeschlossen sind, etwa in Beispiel 2 die Cauchyfamilie mit FTr $t \to \exp(-a\cdot|t|)$, $a > 0$. Als schwacher Trost gegen den zweiten Einwand bleibt allerdings das Corollar zu Satz 6.2, siehe unten.

Das Hauptergebnis dieses Kapitels ist der folgende

Satz 6.2. (Zentraler Grenzwertsatz, u.i.v. -Version) Seien $X(i)$, $i \ge 1$, u.i.v. ZVe mit Erwartung 0 und

Varianz c . Sei $Z(n) = n^{-1/2} \sum_{i \le n} X(i)$. Dann strebt die Verteilung von T(n) bei n gegen unendlich schwach gegen N(0,c) .

Andere Formulierung: Sei μ W.-Maß mit Erwartung 0 und Varianz c . Dann strebt die Folge $T_{1/\sqrt{n}}(\mu * \mu * .. * \mu)$ (n-fache Faltung) bei n gegen unendlich schwach gegen N(0,c) .

Beweis. Wenn X(1) die FTr f hat, dann hat Z(n) die FTr $t \to f(t/\sqrt{n})^n =: g_n(t)$. Zu zeigen ist nach Satz 6.1 die punktweise Konvergenz der Folge (g_n) gegen $\exp(-c/2 \cdot t^2)$. Dazu muß f in der Nähe von t = 0 genauer untersucht werden. Das leistet

Prop. 6.4. Wenn eine Verteilung μ Momente bis zur Ordnung p (p ganz) besitzt, dann ist $\mu\hat{}$ p-mal differenzierbar und es gilt

(15) $$(d/dt)^k \, \mu\hat{}(t)|_{t=0} = i^k \int x^k \, \mu(dx) , \qquad k \le p .$$

Beweis. Man differenziert k-mal die Relation $\mu\hat{}(t) = \int \exp(itx) \, \mu(dx)$, wobei man rechts "unter dem Integralzeichen ableitet". Das geht gut für $k \le p$, da für diese k unter dem Integral eine absolut integrable Funktion steht, nämlich $(ix)^k \exp(itx)$. Setzt man t = 0 , folgt (15).

Wegen Prop.6.4 hat man bei t = 0 die Taylor-Entwicklung

(16) $$f(t) = 1 - c \cdot t^2/2 + o(t) \quad \text{oder} \quad f(t) = \exp(-c \cdot t^2/2) + o(t) .$$

Aus dieser folgt für $g_n(t)$ bei festem t und großem n die Asymptotik

(17) $$g_n(t) = \{\exp(-c \cdot t^2/2n) + o(1/n)\}^n = \exp(-c \cdot t^2/2) + o(1)$$

und damit die Behauptpung des Satzes. Wir haben zuletzt den elementaren Hilfssatz benutzt:

Seien a(i) und b(i) , $1 \le i \le n$, komplexe Zahlen vom Betrag höchstens 1 ; dann gilt

(18) $$|\prod a(i) - \prod b(i)| \le \sum |a(i) - b(i)| .$$

Analog zum Poissonschen Grenzwertsatz (Satz 4.1b) gibt es auch zum Zentralen Grenzwertsatz eine Version, die ohne die Voraussetzung identisch verteilter Summanden auskommt. Der Beweis jener allgemeineren Aussage ist nur wenig komplizierter als der eben geführte, er erfordert nur eine etwas genauere Epsilontik. Wir formulieren den Satz, beweisen ihn aber nicht. (Wegen eines Beweises siehe z.B. [Fe], [Fi].)

Satz 6.3. (Zentraler Grenzwertwatz, Version von Lindeberg-Feller). Seien für jedes n gege-ben die Verteilungen $\mu(n,k)$, $1 \le k \le N(n)$. Es gelte für jedes n

(19) $\int x\, d\mu(n,k) = 0$ für alle $k \le N(n)$, $\quad \sum_k \int x^2\, d\mu(n,k) = 1$;

weiter sei erfüllt die sog. Lindeberg-Bedingung

(Li) $\lim_n \sum_k \int x^2\, 1_{\{|x| > a\}}\, d\mu(n,k) = 0$ für alle $a > 0$.

Dann strebt das Faltungsprodukt $\mu(n,1)*..*\mu(n,N(n))$ bei großem n schwach gegen $N(0,1)$.

Bemerkung. Daß die Lindeberg-Bedingung nicht ersatzlos gestrichen werden kann, macht man sich an folgendem Beispiel klar: die $\mu(n,k)$ seien zentrierte Poissonverteilungen mit Parametern $a(n,k)$, die sich in jeder Zeile, d.h. bei festem n über k summiert, zu 1 addieren. Die Faltungsprodukte in jeder Zeile sind dann ebenfalls zentrierte Poissonverteilungen, und zwar zum Parameter 1 , und diese denken nicht daran, gegen $N(0,1)$ zu konvergieren.

Man sollte (Li) so verstehen, daß die Summanden - in der Sprache der ZVen - in jeder Zeile gleichmäßig klein werden, so daß das Nichtberücksichtigen aller Summanden, die ein festes a überschreiten, keinen wesentlichen Einfluß auf die Gesamtvarianz hat. Die Lindeberg-Bedingung entspricht der Voraussetzung " $\max\{ p(n,k) : k \le N(n)\} \to 0$ " in Satz 4.1b.

Corollar zu Satz 6.2. ("Universelles Fehlergesetz") Sei μ ein W.Maß in R mit endlichem 2. Moment. Es gelte

(20) zu $a , b > 0$ gibt es ein $c > 0$ mit $\quad T_a\mu * T_b\mu = T_c\mu$.

Dann ist μ von der Form $N(0,D)$ für ein $D \ge 0$. (In Worten: die zentrierten Gaußverteilungen bilden dieeinzige faltungsabgeschlossene Skalenfamilie mit endlichem 2. Moment.)

Beweis. Wenn $\mu(n)$ die n-fache Faltungspotenz von μ bezeichnet, so gilt: $\mu(n)$ ist von der Form $T_c(\mu)$ als Mitglied der Skalenfamilie. Ein Vergleich der Varianzen zeigt, daß $c = \sqrt{n}$ gelten muß. Man hat daher $T_{\sqrt{n}}\mu = \mu(n)$ oder

(21) $\mu = T_{1/\sqrt{n}}\, \mu(n)$ für alle $n \ge 1$.

Nach Satz 6.2 strebt die rechte Seite in (21) gegen $N(0,D)$, wo D dieVarianz von μ ist.

Dies zur Ehrenrettung der Gaußischen Familie. In der anderen explizit erwähnten faltungsabgeschlossenen Skalenfamilie, der Cauchyfamilie, gibt es nicht einmal ein 1. Moment.

Bemerkung: Übergang zu Dimension $d > 1$.

Wir haben Satz 6.2 nur für eindimensionale Variable formuliert und bewiesen, aus Gründen der Notation. Es ist klar, daß alle Argumente (Taylorentwicklung usw.) sich auf vektorwertige Variable übertragen lassen. Am einfachsten kann man das einsehen, indem man im Fall von R^d -wertigen zentrierten Variablen $X(m)$ für eine beliebige Linearform u auf R^d die reellen Größen $u \cdot X(m)$ betrachtet. Ihre durch $\sqrt{n}$ geteilten Summen erfüllen den eindimensionalen ZGS mit einer Limesvarianz $c(u) = \mathrm{var}(u \cdot X(1)) = \sum C(j,k)u(j)u(k)$, wo C die Kovarianzmatrix von $X(1)$ ist. Auswertung der FTr an der Stelle $t = 1$ ergibt

$$\lim_n E \exp(iu \cdot S(n)/\sqrt{n}) = \exp(-1/2 \sum C(j,k)u(k)u(k)) \qquad \text{für alle } u , \tag{22}$$

und das ist genau die schwache Konvergenz der d-dimensionalen Verteilungen gegen die Gaußverteilung $N(0,C)$. Dieses Argument setzt natürlich voraus, daß Satz 6.1 in der d-dimensionalen Fassung gültig ist. Das ist er in der Tat. Wir formulieren

Satz 6.2 (mehrdimensional). Wenn $X(i)$ u.i.v. R^d-wertige ZVe sind mit Erwartung 0 und Kovarianzmatrix C , dann strebt die Verteilung von $n^{-1/2} \sum_{i \le n} X(i)$ schwach gegen $N(0,C)$.

Noch ein Wort zur mehrdimensionalen Fassung von Satz 6.1. Sein Corollar sagt, daß eine Verteilung μ in R^d durch ihre Fouriertransfromierte eindeutig bestimmt ist. Das hat eine wichtige Konsequenz: eine Verteilung μ ist eindeutig bestimmt, wenn man für jeden eindimensionalen Unterraum des R^d die Projektion von μ auf diesen Unterraum kennt. Mit "Projektion" von μ auf den Unterraum durch e , e ein Einheitsvektor, meinen wir das Bild von μ unter der Abbildung $x \to (e \cdot x)$. Dieses ist definiert als das Maß π in R mit

$$\int b(y)d\pi(y) = \int b((e \cdot x))d\mu \qquad \text{für } b \text{ stetig beschränkt in } R . \tag{23}$$

Die FTr des Bildes von μ unter der Abbildung $x \to (e \cdot x)$ ist, wie leicht zu sehen,

$$R \ni s \to \int \exp(is(e \cdot x))\, \mu(dx) . \tag{24}$$

Kenntnis der rechten Seite in (24) für alle e und alle s ist aber Kenntnis von $\mu\hat{}(t)$, $t \in R^d$.
Diese simple Beobachtung ist die theoretische Grundlage der sog. Tomographie: ein dreidimensionaler Körper K ("Tumor") ist nach obigem Argument eindeutig (bis auf Mengen vom Lebesgue-Maß null) bestimmt, wenn man für jede Ebene das zweidimensionale Maß ihres Schnittes mit K kennt (oder *a fortiori*

für jede Gerade das eindimensionale Maß ihres Schnit-tes mit K). Denn wenn man das Maß der Schnittmenge für alle Ebenen senkrecht zu einer fe-sten Richtung kennt, hat man die Projektion des Maßes $1_K(x)dx$ auf diese Richtung. Beach-te, daß die Projektionen nur auf die Koordinatenachsen i.a. nicht ausreichen, um eine mehrdi-mensionale Verteilung zu identifizieren: eindimensionale Verteilungen lassen sich auf vielfa-che Art zu einer mehrdimensionalen Verteilung "koppeln".

Anhang. Der zentrale Grenzwertsatz als Fixpunktaussage.

Wir wollen im folgenden Abschnitt den ZGS noch aus einer anderen Perspektive sehen ([S], Lektion 15). Wir lassen in der speziellen Version (Satz 6.2) den Index n über die Teilfolge der Zweierpotenzen laufen. Wir nehmen einfachheitshalber an, die Verteilung der X(i) habe eine W.-Dichte; wir bezeichnen mit f_p die W.-Dichte der reskalierten 2^p-ten Partialsumme. Der Übergang von f_p nach f_{p+1} hat eine einfache Form

(25) $$f_{p+1} = T(f_p) \text{ mit } Tf(x) = \sqrt{2}\cdot \int f(x\cdot\sqrt{2} - y)\cdot f(y)\, dy \ .$$

(In Worten läßt sich die Wirkung von T so beschreiben: man faltet f mit sich selbst und kontrahiert anschließend um den Faktor $\sqrt{2}$.) Offenbar ist die Gaußische Standarddichte φ ein Fixpunkt der Abbildung T ; mehr noch: alle Dichten zu einer Verteilung N(0,c) mit beliebigem c sind Fixpunkte von T . Wir stellen uns nun vor, daß die Dichte f in der Nähe der Gaußischen Standarddichte liegt, und setzen sie an in der Form

(26) $$f(x) = \varphi(x)\cdot(1 + \epsilon\cdot h(x)).$$

Wir schreiben Tf in der selben Form, bis auf einen in ϵ quadratischen Term:

(27) $$(Tf)(x) = \varphi(x)\cdot(1 + \epsilon\cdot g(x)) + \epsilon^2\cdot h(x) \ .$$

Das heißt zunächst, man geht von T über zur Abbildung S im Raum $L^1(\varphi\cdot dx)$, die a nach $S(a) := \varphi^{-1}\cdot T(\varphi\cdot a)$ sendet. Wenn wir anschließend S bei a = 1 linearisieren, erhalten wir die lineare Zuordnung $h \to g =: Lh$. Der Zentrale Grenzwertsatz sagt, daß für gewisse a (dieje-nigen, für welche $\varphi\cdot a$ eine W.-Dichte der Varianz 1 ist) die Folge $(S^p(a))$ der Iterierten von S nach 1 konvergiert. Wir wollen untersuchen, wie sich dieser Sachverhalt im Spek-trum von L wiederspiegelt. Wir werden dazu explizit eine aus Eigenvektoren von L beste-hende Basis von $L^2(\varphi\cdot dx)$ (dieser Raum ist schöner als $L^1(\varphi\cdot dx)$) nebst den zugehörigen Eigenwerten angeben. (Das folgende wird nicht bewiesen, ist aber leicht nachzurechnen.)

Zunächst, der Operator L hat die explizite Darstellung

(28) $$Lh(x) = 2\cdot \int h((x + y)/\sqrt{2})\cdot\varphi(y)\, dy \ .$$

Als Eigenvektoren bietet sich das System der Hermite-Polynome an; sie sind definiert durch

(29) $P_n(x) := \varphi(x)^{-1} \cdot (-d/dx)^n \varphi(x))$, $n \geq 0$, oder

(29a) $P_o = 1$, $P_{n+1}(x) = x \cdot P_n(x) - d/dx\, P_n(x)$ für $n \geq 0$.

(Die ersten Polynome sind 1 , x , $x^2 - 1$, $x^3 - 3x$, $x^4 - 3x^2 + 3$, ...)

Behauptung. Es gilt $LP_n = 2 \cdot 2^{-n/2} \cdot P_n$ für $n \geq 0$.

Das heißt, der Operator L wirkt expandierend in zwei Richtungen, welche den Polynomen 1 und x entsprechen; er hat einen Eigenwert 1 in Richtung $x^2 - 1$, und er wirkt kontrahierend in einem Raum der Kodimension 3 . Das Faktum, daß L einen expandierenden Teil hat, wird in den Voraussetzungen des Zentralen Grenzwertsatzes berücksichtigt: da T nur auf W.-Ma-ße angewandt wird, ist L auf den Teilraum orthogonal zu 1 restringiert. Ferner hat die Einschränkung von T auf W.-Maße mit Erwartung 0 zur Folge, daß L auf den Teilraum orthogonal zu x restringiert wird. Die Tatsache schließlich, daß L einen eindimensionalen Unterraum punktweise unverändert läßt, ist eine unmittelbare Konsequenz davon, daß die eindimensionale Mannigfaltigkeit $N(0,c)$, $c > 0$, punktweise invariant unter T ist.

Anstatt eines förmlichen Beweises der Behauptung empfehlen wir der Leserin, sie für kleine n durch direkte Rechnung nachzuprüfen. Als Hilfsmittel empfiehlt sich dabei die Identität:

(30) $\int \varphi(x) \cdot x^{2n} dx = \prod_{0 \leq i < n} (2i+1)$, $\int \varphi(x) \cdot x^{2n+1} dx = 0$ jeweils für $n \geq 0$

(Beweis durch Induktion). Wir lernen noch mehr aus der Spektraldarstellung von L , nämlich daß im kontraktiven Unterraum die Folge $(L^p h , p \geq 0)$ typischerweise mit Geschwindigkeit $2^{-p/2}$ nach 0 geht, da $2^{-1/2}$ der größte Eigenwert von L in diesem Raum ist. Von T ist da-her ebenfalls zu erwarten, daß $(T^p f , p \geq 0)$ für ein f mit Varianz 1 wie $2^{-p/2}$ gegen φ strebt (in einer geeigneten Metrik) und daß die reskalierte n-fache Faltung von f wie $1/\sqrt{n}$ nach φ strebt. Die Vermutung ist richtig; sie wird in der klassisch-analytischen Statistik mit der Methode der sog. Edgeworth-Entwicklung bewiesen. (Siehe [Fe], S. 506 ff.)

Literatur.

[Fe] W. Feller: An introduction to probability theory and its applications. Vol. II. New York: Wiley 1966.

[Fi] M. Fisz: Wahrscheinlichkeitsrechnung und mathematische Statistik. 6. Aufl. Ber lin: VEB Deutscher Verlag der Wissenschaften 1971. (S. 245.)

[S] Ya.G. Sinai: Probability theory: an introductory course. (Übersetzung aus dem Russischen). Berlin: Springer 1992.

Kapitel 7. Der Poissonprozeß

Inhalt: Theorie des Poisson-Punktprozesses und seine Anwendungen in der Modellierung.

Wir greifen in diesem Kapitel die Untersuchungen von Kapitel 4 über den Poissonischen Grenzwertsatz wieder auf. Wir hatten dort gesehen, daß eine Zählvariable approximativ Poissonisch verteilt ist, wenn sie Summe vieler unabhängiger 0-1-Variablen mit kleiner Erwartung ist. Was jetzt neu hinzukommt, ist eine räumliche Struktur: die vielen "kleinen" Variablen in-formieren nicht nur darüber, ob etwas passiert, sondern auch wo. Bereits die Aussage, daß ein entsprechendes Grenzobjekt, nämlich der Poissonprozeß, existiert, ist von Interesse. Dieser Prozeß spielt in der stochastischen Modellierung eine fundamentale Rolle und er ist der Ge-genstand dieses Kapitels.

Def. 7.1. Wenn n natürliche Zahl und $p = (p(i), 1 \le i \le d)$ ein d-Tupel positiver Zahlen mit Summe 1 ist, dann nennt man Multinomialverteilung zu den Parametern p und n, abgekürzt $M(n,p)$, die Verteilung in N^d mit Gewichten

(1) $$M(n,p;n(1), .. ,n(d)) := M(n;n(1), .. , n(d)) \prod_{1 \le i \le d} p(i)^{n(i)} .$$

(Die Zahl $M(n;n(1), .. , n(d)) = n! (\prod n(i)!)^{-1}$ für $\sum n(i) = n$ (und 0 sonst) ist der bereits bekannte Multinomialkoeffizient, vgl. Kapitel 1 oder 5 und die dort gegebene Interpretation.)

Wir wissen aus Kapitel 4, daß Summen von unabhängigen Poissonvariablen wieder Poissonisch sind. Es gibt aber auch umgekehrt eine einfache Konstruktion, Poissonvariable in unabhängige Poissonische Summanden zu zerlegen, die grundlegend für die Mathematik dieses Kapitels ist:

Prop 7.1. (a) Sei $Z = (Z(1) , .. , Z(d))$ ein Tupel unabhängiger Poissonvariablen mit Erwartungen $\mu(1) , .. , \mu(d)$; sei $S = \sum Z(i)$ und $\mu = \sum \mu(i)$. Dann gilt für jedes n : unter der Bedingung $\{S = n\}$ ist Z nach $M(n,p)$ verteilt, wobei $p(i) = \mu(i)/\mu$, $i = 1 .. d$.

(b) Die ZVen S und Z mit Werten in N bzw. N^d mögen folgende Verteilung haben: Gegeben sind Parameter μ und $p = (p(1) , .. , p(d))$; S ist Poissonisch mit Erwartung μ ; für jedes n ist Z unter der Bedingung $\{S = n\}$ nach $M(n,p)$ verteilt. Dann sind die $Z(i)$ unabhängige Poissonvariable zu den Erwartungen $\mu(i) = \mu \cdot p(i)$ für $i = 1 , .. , d$.

Beweis. Zu Behauptung (a). Man hat

(2) $P(Z(i) = n(i),\ 1 \le i \le d \mid S = n) = \{\prod \exp(-\mu(i))\mu(i)^{n(i)}/n(i)!\}\{\exp(-\mu)\ \mu^n/n!\}^{-1}$,

für $\sum n(i) = n$ bzw. 0 sonst. Der rechts stehende Ausdruck ist gleich $M(n,p;n(1), .., n(d))$, wenn man $p(i) = \mu(i)/\mu$ setzt.

Zu Behauptung (b). Die Verteilung von Z ergibt sich nach dem Satz von der totalen Ws. zu

(3) $P(Z(i) = n(i),\ 1 \le i \le d) = \sum_n P(Z(i) = n(i),\ 1 \le i \le d \mid S = n)\ P(S = n) =$

$\sum_n M(n,p;\ n(1), .. , n(d))\ \exp(-\mu)\ \mu^n/n! = \prod \exp(-\mu(i))\mu(i)^{n(i)}/n(i)!$,

wenn man $\mu(i) = \mu \cdot p(i)$ setzt.

Def. 7.2. Sei G ein beschränktes Gebiet in R^d und $\rho > 0$. Ein Poissonprozeß in G zur Dichte ρ ist eine Zufallsvariable Z mit Werten in den endlichen Teilmengen von G , so daß gilt

(P1) $N(A) := |\ Z \cap A\ |$ ist nach $\Pi(\rho \cdot \lambda(A))$ verteilt für alle $A \subset G$;

(P2) für disjunkte $A(1) , .. , A(d)$ sind die $N(A(i))$, $1 \le i \le d$, unabhängig.

($|\ .\ |$ bezeichet die Kardinalität einer Menge und $\lambda(.)$ das Lebesguemaß.) Beachte, die Mächtigkeit der Menge Z und die Lage ihrer Elemente sind beides zufällige Größen.

Wir müssen als erstes nachprüfen, ob die Definition sinnvoll ist, d.h. ob es solche Poissonprozesse überhaupt gibt. Einen ersten positiven Hinweis gibt folgende Überlegung: Wenn wir eine Menge A zerlegen, $A = \cup A(i)$, so ist die Verteilung von N(A) nach (P1) gleich $\Pi(\rho \cdot \lambda(A))$, nach (P2) ist sie wegen $N(A) = \sum N(A(i))$ das Faltungsprodukt der $\Pi(\rho \cdot \lambda(A(i)))$. Vermöge Prop. 4.1. sind aber - glücklicherweise - beide Verteilungen gleich. Ein eigentlicher Existenzbeweis ist das noch nicht; ein solcher ist Gegenstand von

Prop. 7.2. Der folgende Zufallsmechanismus liefert einen Poissonprozeß: Man realisiert zuerst eine ZV Y nach dem Gesetz $\Pi(\rho \cdot \lambda(G))$. Danach, unter jeder möglichen Bedingung $Y = n$, realisiert man n unabhängige in G uniform verteilte ZVe $X(1), .. , X(n)$. Die zu erzeu-gende zufällige Punktmenge wird als $\{X(1), .., X(n)\}$ definiert.

Beweis. Wenn $G = \cup A(i)$ eine Zerlegung von G ist, dann sind unter der Bedingung $\{Y = n\}$ die Zählvariablen $|N(A(i))|$ multinomial nach $M(n, p)$ mit $p(i) = \lambda(A(i))/\lambda(G)$ verteilt. We-gen Prop.1(b) sind die

N(A(i)) unabhängig nach $\Pi(\rho\cdot\lambda(A(i)))$ verteilt und damit ist (P) er-füllt. (Man kann sich in (P2) auf Mengen A(i) mit Vereinigung G beschränken.)
Es folgen einige Beispiele für stochastische Modellierung mit Poissonprozessen. Im Prinzip kann man in fast jedem der in Kapitel 4 genannten Modelle von den Zählvariablen zu einem System von "zufälligen Punkten" im Raum oder in der Zeit übergehen. Insofern sind die nachstehenden Anwendungssituationen nicht wesentlich verschieden von den früher beschriebenen.

Beispiel 1. Bakterienkulturen (s. [F], S. 149 ff.).

Simulation von 100 uniform verteilten Punkten in einem Quadrat

Häufigkeiten der Besetzungszahlen für die 100 Planquadrate:

b	=	0	1	2	3	≥ 4
h	=	39	33	19	7	2
100p	=	36.8	36.8	18.4	6.1	1.9

Hier ist $p = \Pi(1)$. Die Übereinstimmung ist gut, der χ^2-Wert ist 0.65 bei 3 Freiheitsgraden.

In einer geeignet präparierten Petrischale wird nach einiger Zeit nachgesehen, wo sich Bakterienkolonien gebildet haben. Die räumliche Ausdehnung einer Kolonie ist - wenn nicht zuviel Zeit verstrichen ist - praktisch null. Es gibt viele Kolonien in der Schale, und man hat guten Grund zur Annahme, daß die Ausbildung einer Kolonie an einem Ort unabhängig davon ist, was an anderen Orten geschieht. Weiter, wenn man die Schale in räumlich homogener Weise am Anfang präpariert und mit Nährstoffen versieht, darf man erwarten, daß die mittlere Zahl der Kolonien in jedem Teilstück proportional zu dessen Fläche ist. Insofern ist die

Modellie-rung der zufällig auftretenden Punkte durch einen Poissonprozeß gerechtfertigt. (Ein anderes - für uns heutige reichlich makabres - Beispiel derselben mathematischen Struktur, das man ebenfalls in [F] findet, ist ein Stadtplan von London, auf dem die Einschlagstellen von Bomben während des 2. Weltkriegs markiert worden sind.)

Im Beispiel der Petrischale - für die Bomben gilt dieser Einwand in noch viel stärkerem Maße - hat das gewählte Modell die nur schwer oder gar nicht einleuchtende Konsequenz, daß die Gesamtzahl der Kolonien in der Schale Poissonisch verteilt ist. Hingegen haben wir keine Probleme mit der Annahme, daß unter der Bedingung einer bestimmten Zahl von Kolonien deren Positionen sich wie unabhängige uniform verteilte Variable verhalten. Soll man in dieser Lage nur die zweite Eigenschaft bei der Modellierung postulieren und darauf verzichten, die Gesamtzahl als Poissonisch anzusehen? Welche Bedeutung hat überhaupt die zusätzliche Annahme einer Poissonischen Gesamtzahl?

Wir erinnern an die Ausführungen in Kapitel 4: es ist relativ unerheblich, ob wir eine große Zählvariable als deterministisch oder Poissonisch modellieren. Denn selbst wenn sie "in Wahrheit" Poissonisch wäre, so wäre das in der Praxis irrelevant, da der typischerweise unbekannte Parameter der Verteilung erst geschätzt werden müßte. Und eine bessere Schätzung als die tatsächlich beobachtete Zahl wird man schwerlich finden. So sind wir auch im oben simu-lierten Beispiel vorgegangen; wir haben die beobachteten Häufigkeiten mit denen einer Pois-sonverteilung zum Parameter 100 verglichen, weil 100 die Gesamtzahl der Punkte war. Mit welcher anderen Poissonverteilung hätten wir unser Datenmaterial auch vergleichen sollen? Fazit: die Frage, ob man ein "reines" Poissonmodell aufstellen und den unbekannten Parame-ter ρ "anpassen" ("schätzen") soll, oder ob ein Multinomialmodell mit großem deterministi-schen n besser wäre, ist rein akademisch und ohne Einfluß auf den Erklärungswert des Modells für den Effekt der räumlichen Gleichverteilung der einzelnen zufälligen Punkte unter der Bedingung einer bestimmten Gesamtzahl.

Beispiel 2. Ideales Gas.

Wir interpretieren das Gebiet G als Gefäß, in dem sich n Gasmoleküle befinden, und fragen, wie die Positionen der Moleküle im thermischen Gleichgewicht verteilt sind. Das ideale Gas ist dadurch definiert, daß bei gegebener Gesamtzahl n die Positionen der Moleküle u. i.v. sind mit der uniformen Verteilung über G. (Das stochastische Wort der "Unabhängigkeit" paßt gut zum physikalischen Begriff der Idealität, die man als "Abwesenheit einer Wechselwirkung zwischen den Molekülen" charakterisiert.) Wie in Beispiel 1 gilt auch hier, daß man mit gleichem Recht die Positionen der Moleküle durch einen Poissonprozeß hätte beschreiben können. Wenn man sich nur für die "mikroskopischen" Eigenschaften des Gases interessiert - technisch gesprochen: dafür, wie die die Einschränkung des Punktprozesses auf ein Teilgebiet von G aussieht, das im Mittel nur eine beschränkte Zahl von Molekülen enthält - sind beide Modelle bei großem n nicht zu unterscheiden. Im physikalischen Sprachgebrauch nennt man die Modellierung mit festem n kanonisches und die mit Poissonischer Gesamtteilchenzahl großkanonisches Ensemble.

Die Tatsache, daß man die Positionen in Abwesenheit einer Wechselwirkung als unabhängig uniform verteilt annimmt, hat nichts Mysteriöses an sich; sie entspricht der Gleichverteilung im Laplacesche Sinn im hochdimensionalen Konfigurationsraum R^{3n} . Daß die "Gleichverteilung" als Lebesguemaß definiert wird, hängt mit dem invarianten Maß zur Hamiltonschen Dynamik (Satz von Liouville) zusammen. Womit wir bei den Geschwindigkeiten wären.

Natürlich gehört zum idealen Gas auch eine Modellierung der Geschwindigkeiten der Moleküle. Das übliche (Maxwellsche) Modell sieht so aus, daß auch diese u.i.v. sind, mit Gesetz $N(0,c\mathbf{1})$, wo $\mathbf{1}$ die Einheitsmatrix in R^3 ist und c ein positiver Parameter proportional zur Temperatur ist. (Genauer: $c = kT/m$, wo k die Boltzmannkonstante und m die Masse eines Moleküls ist.) Die Verteilung von Ort und Geschwindigkeit simultan in einem Poissonmodell im Sinn von Def. 7.2 auszudrücken, ist mühsamer und wird hier nicht weiter verfolgt; man müßte zu diesem Zweck jene Definition leicht verallgemeinern.

Unabhängig von der Dimension kann man Poissonprozessse auch in Gebieten mit unendlichem Volumen, z.B. in ganz R^d , definieren.

Def. 2a. Der Poissonprozeß zur Dichte ρ in einem beliebigen Gebiet G ist eine ZV Z mit Werten in der Menge der höchstens abzählbaren Teilmengen von G , derart daß die Zählvariablen $N(A) = |A \cap Z|$ für jedes A mit endlichem Volumen nach $\Pi(\rho\cdot\lambda(A))$ verteilt sind, unabhängig für disjunkte A .

Die Frage nach der Existenz ist - auf einem naiven Niveau - trivial: ein "unendlicher" Poissonprozeß ist nichts anderes als eine Sammlung von unabhängigen "endlichen" Poissonprozessen in disjunkten endlichen Gebieten $G(i)$, in welche man sich G zerlegt denkt. - Im Fall der Dimension 1 läßt sich der Poissonprozeß in einer Halbgeraden sehr einfach charakterisieren.

Satz 7.1. Die Zven $X(i)$, $i \geq 1$, seien u.i.v. exponentialverteilt zum Parameter ρ , sei $T(n) := \sum_{1\leq i\leq n} X(i)$ und sei Z als die zufällige Punktmenge $\{T(1), T(2) , ... \}$ definiert. Dann ist Z ein Poissonprozeß der Dichte ρ in der positiven Halbgeraden.

Beweis. Für $t > 0$ setzen wir $N(t) := N((0,t]) = \max\{ n : T(n) \leq t\}$. Dann hat man

(4) $$P(N(t) = n) = P(\sum_{i\leq n}X(i) \leq t < \sum_{i\leq n+1}X(i)) =$$

$$\int dx(1) .. dx(n+1)\ \rho^{n+1} \exp(-\rho\sum_{i\leq n+1}x(i))\cdot 1_{\{t(n) \leq t < t(n+1)\}} ,$$

wo mit $t(n)$ analog zu $T(n)$ die Partialsummen der $x(i)$ bezeichnet seien. Durch Übergang zu den Integrationsvariablen $t(i)$ läßt sich der letzte Ausdruck umformen zu

(5) $$\int dt(1) .. dt(n+1)\ \rho^{n+1} \exp(-\rho t(n+1))\cdot 1_{\{0<t(1)<..t(n)\le t<t(n+1)\}} =$$
$$\int dt(1) .. dt(n)\ \rho^{n} \exp(-\rho t)\cdot 1_{\{0<t(1)< ..<t(n)\le t\}} = \exp(-\rho t)\cdot(\rho t)^n /n! .$$

Damit haben wir gezeigt, daß $N(t)$ Poissonisch verteilt ist mit dem richtigen Mittelwert ρt . Weiter, eine genauere Auswertung von (5) zeigt, daß das Tupel $(T(1), ..., T(n))$ unter der Bedingung $N(t) = n$ folgende Verteilung hat: es ist uniform bezüglich des normierten n-dimensionalen Lebensguemaßes in $(0,t]^n$ verteilt, unter der Bedingung, daß $(t(i))$ eine aufsteigende Folge bildet. Diese Bedingung ist aber unerheblich, wenn man sich nur für die Verteilung der Menge $\{T(1), .., T(n)\}$ interessiert; d.h. $\{T(1), .., T(n)\}$ ist unter $\{N(t) = n\}$ so verteilt, wie wenn die $T(i)$ unabhängige uniforme Variable wären. Die zufällige Menge $Z \cap (0,t]$ erfüllt also die Voraussetzungen von Prop.7.2 und ist ein Poissonprozeß zur Dichte ρ in $(0,t]$. Da dies für beliebiges t richtig ist, stellt Z einen Poissonprozeß zur Dichte ρ in $(0,\infty]$ dar.

Man bekommt nach Satz 1 einen Poissonprozeß in ganz R , indem man zwei Folgen unabhängiger Exponentialvariablen hernimmt und die zugehörgen Partialsummen auf der rechten und auf der linken Halbgeraden aufträgt. In jeder Dimension d hat ein Poissonprozeß in R^d die Eigenschaft der räumlichen Homogenität oder Stationarität: sein Verteilungsgesetz ändert sich nicht, er bleibt Poissonprozeß zur gleichen Dichte, wenn man alle zufälligen Punkte um einen festen Vektor verschiebt; das liegt an der Translationsinvarianz des Lebesguemaßes.

Zurück zur Dimension 1 . Wir fragen, wie die Länge eines Intervalls zwischen zwei aufeinanderfolgenden Punkten in einem Poissonprozeß der Dichte ρ verteilt ist. Diese Frage ist unpräzis, weil sie keine Vorschrift enthält, wie man "ein Intervall" bekommt. Sie führt uns, ähnlich wie in Kapitel 1, auf ein Inspektionsparadoxon. Eine mögliche Antwort besteht darin, in der Konstruktion nach Satz 7.1 ein Intervall $(T(n),T(n+1))$ mit irgendeinem $n \ge 1$ zu nehmen: seine Länge ist $X(n+1)$, und dessen Verteilung ist für jedes feste n exponentiell mit Parameter ρ . Eine andere Antwort besteht darin, auf einen festen Punkt x zu deuten und dasjenige Intervall herzunehmen, welches x enthält. Die Verteilung des so herausgegriffenen Intervalls ist wegen der räumlichen Homogenität des Poissonprozesses für alle x die gleiche, also nehmen wir $x = 0$ an. Nach dem eben Gesagten ist die Länge dieses Intervalls die Summe von zwei unabhängigen Exponentialvariablen zum Parameter ρ , welche den Abstand des ersten Punktes vom Ursprung im rechten bzw. linken Partialsummenprozeß darstellen. Durch die veränderte Inspektionsvorschrift hat sich der Mittelwert des herauszugreifenden Intervalls um einen Faktor 2 vergrößert. In einem gewissen Sinn ist übrigens die erste Anwort die vernünftigere: die mittlere Länge des Intervalls nach der ersten Rechnung ist $1/\rho$, also das Reziproke der Dichte, der mittleren Anzahl der Punkte pro Längeneinheit. Das Gesetz der großen Zahlen bestätigt das: beide Limiten, $T(n)/n$ und $N(t)/t$, existieren im stochastischen Sinn und sind zueinander reziprok, nämlich gleich $1/\rho$ bzw. ρ .

Mehr oder weniger reale Situationen, in denen dieses Paradoxon auftritt, sind die folgenden:
a) Warten auf die Straßenbahn. Wenn man die Abfahrtszeiten der Straßenbahn als Poissonprozeß annimmt

(??) und wenn der mittlere Abstand zwischen zwei Abfahrten 10 Minuten beträgt, so muß ein zufällig ankommender Fahrgast im Mittel nicht etwa 5 Minuten warten (wie es der Fall wäre, wenn die Abfahrten exakt alle 10 Minuten erfolgten), sondern 10 Minuten: so lang ist das Intervall von 0 bis T(1) im Mittel.
b) Wenn man wissen will, wie lange ein Student, der in Heidelberg Examen macht, im Mittel studiert, kann man die Daten auf verschiedene Art erheben. Man kann (1) alle Studenten zu zählen, die 1996 das Examen ablegen und zu registrieren, wie lang sie vorher studiert haben. Oder man zählt (2) alle im SS 1996 eingeschriebenen Studenten und wartet, bis sie ihr Examen ablegen und notiert die entsprechenden Zahlen. Nach dem Gesagten ist klar, daß (2) höhere Mittelwerte produzieren wird als (1): man sagt umgangssprachlich, ein Langzeitstudent habe bei (2) gegenüber (1) eine relativ größere Chance, in die Zählung aufgenommen zu werden, als ein Kurzzeitstudent.

Beispiel 3. Verweildauern.

Wir wollen die rein verbalen Ausführungen des vorigen Absatzes mit einem exakt durchgerechneten Modell absichern. Um bei den Studenten zu bleiben: Zu jedem ganzzahligen Zeitpunkt n (typischerweise im Jahresrhythmus) kommt eine zufällige Zahl $A(n)$ von Studenten in einen Studiengang; die $A(n)$ seien u.i.v. und nach $\Pi(a)$ verteilt, mit einem festen Parameter $a > 0$. Der i-te Student habe eine Verweildauer $T(i)$; die Verweildauern seien ebenfalls u.i.v. und $P(T(i) = k) = p(k)$, $k = 1 , 2 , ..$ (oben offen!) . Wir stellen die folgenden Fragen:

(a) Funktioniert das Modell in zeitlich stationärer Weise? Wenn ja,
(b) wie ist zu einem festen Zeitpunkt die Anzahl der gerade eingeschriebenen Studenten verteilt?
(c) etwas feiner, wieviele Studenten in jeder Altersklasse gibt es zu einem festen Zeitpunkt?

Vermöge Prop.7.1b ist die Rechnung verblüffend einfach. Man betrachte die doppelt indizierte Familie von ZVen $A(n,k)$, $n \in Z$, $k \geq 1$, mit $A(n,k) :=$ Anzahl der zur Zeit n ankommenden Studenten, die k Zeiteinheiten im System bleiben werden. Dann sind die $A(n,k)$ insgesamt unabhängig und Poissonisch mit $EA(n,k) = a \cdot p(k)$. (Man muß sich das bildlich so vorstellen, daß der Strom der ankommenden in unabhängige Schichten zerlegt wird; in der k-ten Schicht sind die Individuen mit Verweilzeit k .) Wenn wir mit $N(n,k)$ die Zahl der zur Zeit n Anwesenden mit Verweildauer k bezeichnen und mit $N(n)$ die Zahl aller zur Zeit n Anwesenden, gilt

$$(6) \qquad N(n,k) = \sum_{o \leq j < k} A(n-j,k) , \qquad N(n) = \sum_k N(n,k) = \sum_{k,j: o \leq j < k} A(n-j,k) .$$

Alle Größen sind Poissonisch als Summen unabhängiger Poissonischer Variablen. Für die Erwartungen gilt

$$(7) \qquad EN(n,k) = a \cdot p(k) \cdot k \qquad \text{und} \qquad EN(n) = a \cdot \sum_k p(k) \cdot k .$$

Damit sind alle gestellten Fragen beantwortet: wenn die mittlere Verweilzeit eines Individuums $\mu := \sum_k$

p(k)·k endlich ist, dann ist die zweite Summe in (7) und damit auch (Beppo Le-vi) in (6) konvergent , die Variable N(n) ist wohl definiert, das System der N(n) , $n \in Z$, wie auch jedes Subsystem N(n,k) , $n \in Z$, ist eine zeitlich stationäre Folge. Das Inspektions-paradoxon kommt ins Spiel, wenn wir nach der Verweildauer V einer zur Zeit n zufällig (genauer: unter der Bedingung {N(n) > 0}) ausgewählten Person fragen: nach Prop.7.1a gilt

(8) $$P(V = m \mid N(n) > 0) = a\cdot p(m)\cdot m/(a\cdot \sum_k p(k)\cdot k) = p(m)\cdot m/\mu \quad .$$

Die Inspektionsvorschrift wirkt sich auf die Ws. von {V = m} gegenüber der entsprechenden Ws. von {T= m} als Multiplikation mit den Faktor m und anschließende Normierung aus.

Beispiel 4. Zerbrechen (Degradation) eines Polymers.

Sei gegeben ein Polymer in Form einer Kette, d.h. ein riesiges Molekül aus linear aneinander gereihten Bausteinen (Atomen oder Gruppen von Atomen). Wir stellen es uns als eine kontinuierliche Strecke der Länge L vor. Die Bindungen zwischen je zwei benachbarten Bausteinen seien einem zeitlich konstanten Risiko der Auflösung (des "Brechens") ausgesetzt, gleichgültig wie groß die Kette gerade ist. Im kontinuierlichen Modell bedeutet das: die Ws. dafür, daß im Zeitintervall dt und im Raumintervall dx ein Bruch erfolgt, ist a·dt·dx - unabhängig davon, ob schon vorher Brüche anderswo erfolgt sind; dabei ist a ein positiver Parameter. Das heißt, wir modellieren die Orte und Zeitpunkte der Brüche durch einen zweidimensionalen Poissonprozeß der Dichte a in einem Gebiet $\{0 < x < L , 0 < t < T\}$. Am Ende der Beobach-tung, zur Zeit T , geben die Projektionen der markierten Raum-Zeit-Punkte auf die Raumkoordinate die Orte der Brüche an. Diese Bruchstellen bilden einen eindimensionalen Poissonprozeß der Dichte aT . Die Verteilung der Länge X(1) des "ersten" Bruchstücks, d.h. des den "Anfang" des Polymers enthaltenden Stücks, und aus Symmetriegründen auch jedes anderen Bruchstücks, ist exponentiell mit Abschneidung oberhalb der Länge L :

(9) $$P(X(1) > x) = \exp(-aTx) \quad \text{für } x \le L \text{ und } 0 \text{ für } x > L .$$

Wenn a·T·L , die erwartete Gesamtzahl der Brüche, hinreichend groß ist, wird es extrem unwahrscheinlich, daß die Abschneidung eine Rolle spielt (das würde sie nur, wenn es überhaupt keinen Bruch gegeben hat). Die (empirische) Längenverteilung ist experimentell meßbar; dadurch ist die Hypothese des konstanten Bruchrisikos überprüfbar.

Literatur.

[F] W. Feller: An introduction to probability theory and its applications. Vol I. 2nd edition.

New York: Wiley 1957.

Kapitel 8. Starkes Gesetz der großen Zahlen

Inhalt. Momenterzeugende Funktion und Entropie, Ungleichung von Bernstein-Cramér. Starkes Gesetz der großen Zahlen. Hilfsmittel der Maximalungleichung.

Das Hauptziel ist der Beweis des in der Überschrift genannten Satzes, zunächst noch unter einer zusätzlichen Voraussetzung, die aber nicht sehr restriktiv ist und die später aufgehoben wird. Die Motivation für das starke Gesetz ist (im Rahmen dieses Kurses) nicht so sehr die gegenüber dem schwachen Gesetz (Satz 2.1) bessere Anwendbarkeit, sondern seine mathematisch einfache, eigentlich: perfekte, Form: es gehört sozusagen zur Allgemeinbildung der Probabilistin. Die Methode der exponentiellen Abschätzungen ("großen Abweichungen") ist hier-bei nicht nur Mittel zum Zweck, sondern von eigenem Interesse; außerdem ist sie bereits in Kapitel 3 oder 5 angelegt. (Siehe dazu z.B. [DS].)

Def. 8.1. Sei X eine beschränkte reelle ZVe. Wir bezeichnen als zugehörige (genauer: zur Verteilung von X gehörige) momenterzeugende Funktion die Zuordnung

(1) $$a \to M(a) := E \exp(aX) , \quad \text{für } a \in R ,$$

als Zustandssumme (oder freie Energie) die Funktion

(2) $$a \to F(a) = \log M(a) ,$$

welche jeweils für $a \in R$ definiert sind. Die Entropiefunktion ist definiert durch

(3) $$I(x) := \sup \{a \cdot x - F(a) : a \in R \} \quad \text{für } x \in R ,$$

wobei der Wert $I(x) = +\infty$ zugelassen wird.

Bemerkung. Die momenterzeugende Funktion ist eine alte Bekannte. Sie ist nichts anderes als die Fouriertransformierte, aber ausgewertet an der imaginären und nicht an der reellen Achse, oder - für ZVe mit Werten in den natürlichen Zahlen - die erzeugende Funkton, wenn man als deren Argument $\exp(a)$ einsetzt. Es folgt unmittelbar, daß bei Faltung die momenterzeugenden Funktionen sich multiplizieren und die

Zustandssummen sich daher addieren. Die Funktion I hat gegenüber der "richtigen" Entropie ihr Vozeichen geändert; wir bitten die Leserin um Entschuldigung, daß wir kein neues Wort für I gefunden haben.

Prop. 8.1. Die Funktion F ist konvex, man hat

(4a) $$F(0) = 0 , \quad F'(0) = EX ,$$

(4b) $$\lim_{a\to-\infty} F'(a) = \text{ess-inf } X , \qquad \lim_{a\to+\infty} F'(a) = \text{ess-sup } X .$$

Die Funktion I ist strikt konvex. Es gilt $I \geq 0$ überall und $I(x) = 0$ genau für $x = EX$.
(Zur Erklärung: ess-sup X ist die kleinste Zahl a, für welche gilt $P(X > a) = 0$.)

Beweis. Sei der Einfachheit halber die Verteilung von X diskret, d.h. vom Typ $\sum p(x)\delta_x$. Dann gilt

(5) $$F'(a) = M(a)^{-1} \sum p(x)\cdot x\cdot\exp(ax) ,$$

$$F''(a) = M(a)^{-2} \{M(a)\cdot\sum p(x)\cdot x^2\cdot\exp(ax) - (\sum p(x)\cdot x\cdot\exp(ax))^2\} .$$

Der erste Ausdruck wird als Mittelwert zur Verteilung $M(a)^{-1} \sum p(x)\cdot\exp(ax)\cdot\delta_x$ erkannt, der zweite als deren Varianz. Aus letzterem folgt die Konvexität von F, aus ersterem die Aussagen in (4b).

Die Konvexität von I folgt unmittelbar aus (3): es ist Supremum von affinen Funktionen. Die strikte Konvexität folgt aus der Differenzierbarkeit von F. (Der Graph von I kann nur dann ein "flaches" Stück aufweisen, wenn mehrere Tangenten an F mit verschiedener Steigung den Graphen von F am selben Punkt berühren, d.h. wenn dieser Graph einen "Knick" hat.) Schließlich ist $I \geq 0$ wegen $F(0) = 0$; die Gleichheit $I(EX) = 0$ gilt, weil die Tangente an F an der Stelle 0 kleiner gleich F ist (F ist konvex); aus der strikten Konvexität von I folgt die Einzigkeit der Minimalstelle.

Prop. 8.2. Es gelten die Ungleichungen

(6) $$P(X > x) \leq \exp(-I(x)) \quad \text{für} \quad x > EX \qquad \text{bzw.}$$

$$P(X < x) \leq \exp(-I(x)) \quad \text{für} \quad x < EX .$$

Beweis. Wir zeigen die erste der Ungleichungen in (6); die andere geht durch Spiegelung daraus hervor. Man hat für beliebiges $a > 0$

$P(X \geq x) = P(\exp(aX) \geq \exp(ax)) \leq E \exp(aX)/\exp(ax) = M(a) \exp(-ax)$, daher

$$-\log P(X \geq x) \geq - F(a) + ax .$$

Da hierin a beliebig war, kann man die r.S. durch $\sup\{ax - F(a) : a > 0\}$ ersetzen. Man muß sich jetzt nur noch überlegen, daß dieser letzte Ausdruck gleich $\sup\{ax - F(a) : a \in R\}$, also gleich $I(x)$ ist, und die Proposition ist bewiesen. Hier wird die Bedingung $x > EX$ benötigt; das Maximierungsproblem, das mit der Definition (3) von I verbunden ist, heißt nämlich geometrisch soviel wie: finde die Stelle a , an der $F'(a) = x$ gilt und nimm die Differenz von ax und $F(a)$. Diese Stelle a liegt aber sicher rechts von 0 , da x als größer als $F'(0) = EX$ vorausgesetzt war. Insofern genügt es, das Supremum über die positiven a zu bilden. (Siehe Zeichnung.) Dieses Argument ist sicher korrekt, wenn $x \leq$ ess-sup X , weil dann die Gleichung $F'(a) = x$ (und sei es mit $a = +\infty$) tatsächlich lösbar ist. Wenn aber $x >$ ess-sup X , ist (6) trivialerweise richtig, da die l.S. dann null wird. Für solche x ist übrigens $I(x) = +\infty$.)

Zeichnung. Funktion F (konvexe Funktion mit $F'(0) = 0$), Tangente an F zur Steigung 1 , Parallele dazu durch den Ursprung. Die Tangente schneidet die y-Achse in der Höhe $- I(1)$

Satz 8.1. (Ungleichung von Bernstein-Cramér) Sei $X(i)$, $1 \leq i \leq n$, eine Familie von beschränkten u.i.v.

Variablen mit Mittelwert μ und I(.) die zugehörige Entropiefunktion. Sei $S(n) = \sum_{1 \le i \le n} X(i)$. Dann gilt

(7) $P(S(n)/n - \mu > x) \le \exp(-n \cdot I(x))$ für $x > \mu$ und

$P(S(n)/n - \mu < x) \le \exp(-n \cdot I(x))$ für $x < \mu$.

Beweis. Ist eine unmittelbare Folge aus Prop. 2. Wenn F die Zustandssumme und I die Entropiefunktion von X(1) ist, dann hat S(n) die Zustandssumme $a \to n \cdot F(a)$ (s. Bemerkung nach Def. 1) und die Entropiefunktion $y \to n \cdot I(y/n)$ (nach (3)). Man setzt $n \cdot x = y$ in (7) und wendet (6) an.

Beispiele der Berechung von Zustandssumme und Entropie.

1) Bernoulli- oder Binomialverteilung.

Betrachte das W.-Maß $p \cdot \delta_1 + q \cdot \delta_o$, wobei $q = 1 - p$ (sog. Bernoulli-Verteilung). Dann ist

(8) $F(a) = \log(q + p \cdot \exp(a))$, $F'(a) = p \cdot \exp(a)[q + p \cdot \exp(a)]^{-1}$,

und man erhält I(x) , für $0 < x < 1$, indem man die Gleichung $F'(a) = x$ löst und, wenn a^* die Lösung dieser Gleichung ist, indem man setzt

(9) $I(x) = a^* x - F(a)$.

Im vorliegenden Fall ergeben sich aus $F'(a^*) = x$ die Beziehungen

$F(a^*) = a^* + \log p - \log x$ und $\exp(a^*) = (q \cdot x)/(p(1-x))$, somit schließlich

(10) $I(x) = x \log(x/p) + (1-x) \log((1-x)/q)$.

Das heißt, wir erhalten als Entropie die schon bekannte Ratenfunktion für die Binomialverteilung aus Kapitel 3, Satz 3, wie eigentlich nicht anders zu erwarten. Der jetzige Satz 1 stellt eine Verbesserung gegenüber jenem früheren Satz dar, da wir jetzt eine finite Abschätzung

(11) $\sum_{k \ge nx} B(n,p;k) \le \exp(-n \cdot I(x))$ für $x > p$

(und analog für $x < p$) ohne irgendwelche zusätzlichen Konstanten erhalten und nicht nur eine Aussage über die logarithmische Wachstumsrate der l.S. in (11).

2) Gaußverteilung.

Sei gegeben die Verteilung N(0,c) . Sie sprengt streng genommen den gesetzten Rahmen, da sie keinen beschränkten Träger hat. Man kann aber alle Argumente übertragen. Es ergibt sich

(12) $M(a) = \exp(1/2 \cdot c \cdot a^2)$, $F(a) = 1/2\ c \cdot a^2$, $I(x) = 1/2 \cdot c^{-1} \cdot x^2$,

die erste Beziehung durch quadratische Ergänzung des Exponenten, der Rest durch direktes Nachrechnen, indem man das Problem (3) löst. Satz 1 sagt dann

(13) $P(S(n)/n > x) \leq \exp(-nI(x)) = \exp(- n \cdot x^2/(2c))$.

Beachte: die r.S. in (13) ist bis auf einen konstanten Faktor die Dichte bei x von $N(0,c/n)$, welches die Verteilung von $S(n)/n$ ist. Insofern haben wir - im vorliegenden Spezialfall - bewiesen, daß die logarithmische Rate, mit welcher die l.S. in (13) bei großem n klein wird, nicht größer sein kann als $1/(2c) \cdot x^2$, was gleich $I(x)$ ist. Diese Eigenschaft ist keineswegs charakteristisch nur für die Gaußverteilungen (siehe auch voriges Beispiel); wir verzichten darauf, die Optimalität der Abschätzung in Satz 8.1 als generelles Resultat zu formulieren.

3) Exponential- und Poissonverteilung.

Diese beiden Verteilungen haben unbeschränkte Träger, sind aber trotzdem behandelbar. Für die Exponentialverteilung zum Parameter 1 , das Maß $\exp(-x) \cdot 1_{\{x>0\}}\, dx$, gilt

(14) $M(a) = 1/(1 - a)$, $F(a) = - \log(1 - a)$, $I(x) = x - 1 - \log x$.

Für die Verteilung $\Pi(t)$ hingegen erhält man (in der offenkundigen Notation von um eine alphabetische Einheit nach vorne verschobenen Buchstaben)

(15) $L_t(a) = \exp(t \cdot (\exp(a) - 1))$, $E_t(a) = t \cdot (\exp(a) - 1)$,

$H_t(x) = t \cdot \{x/t \cdot \log (x/t) - x/t + 1\} = t \cdot H_1(x/t)$.

Die Faltungshalbgruppeneigenschaft der Poissonfamiliedrückt sich darin aus, daß H_t das t-fache einer Funktion von x/t ist. (In beiden Fällen sollte man die übliche Warnung hinzufügen, daß I bzw. H_t nicht in ganz R endlich ist.) Wir hatten in Satz 7.1, beim Zusammenhang zwischen den Partialsummen von Exponentialvariablen $S(n)$, $n \geq 1$, und den Zählvariablen $N(t) := N(0,t]$) des Poissonprozesses gesehen, daß bei geeigneter Darstellung gilt

(16) $\{N(t) < n\} = \{ S(n) > t\}$ für jedes n , t .

Wenn wir $t > n$ wählen, ist sowohl die rechte als auch linke Seite in (16) für man eine exponentielle Abschätzung geeignet, einmal für Poisson- und einmal für Exponentialvariable. Wir wollen nachprüfen, ob die beiden Abschätzungen zum selben Ergebnis kommen. Wir erhalten nach Satz 8.1 und obiger Berechnung der Entropiefunktionen

(17) $P(N(t) < n) \le \exp(- H_t(n)) = \exp(n - t + n\cdot\log(t/n))$,

$P(S(n) > t) \le \exp(-n\cdot I(t)) = \exp(n - t + n\cdot\log(t/n))$.

Wir haben also Glück gehabt. Man erkennt übrigens, daß die rechte Seite in (17), bis auf polynomiale Terme in n , gleich $\exp(-t)\cdot t^n/n!$ ist, also gleich der Ws. dafür, daß eine $\Pi(t)$-Variable gleich n ist. Insofern sieht man wieder, daß die durch Cramér-Bernstein gegebene logarithmische Rate der Abschätzung für beide Verteilungstypen optimal ist.

Wir kommen jetzt zum starken Gesetz der großen Zahlen. Dazu als Vorbereitung

Def. 8.2. Eine Folge $(X(n))$ von ZVen konvergiert fast sicher oder mit Ws. 1 gegen eine ZVe X , wenn gilt

(18) $P(\lim_n X(n) = X) = 1$, ausführlicher

$P(\{\omega : \lim_n X(n)(\omega) = X(\omega)\}) = 1$,

wenn die $X(n)(\omega)$ und $X(\omega)$ Funktionen auf einem W.-Raum mit Elementen ω sind.

Dieser Konvergenzbegriff ist stärker als der der stochastischen Konvergenz. man kann nämlich leicht sehen, daß (14) äquivalent ist zu

(18a) $\lim_n P(\text{es gibt ein } m \ge n \text{ mit } |X(m) - X| > a) = 0$ für alle $a > 0$,

wohingegen stochastische Konvergenz beschrieben wird durch

(St) $\lim_n P(|X(n) - X| > a) = 0$ für alle $a > 0$.

Die Interpretation ist die folgende: man ist für große n nahe am "richtigen" Wert, auch ohne daß man eine genaue Kontrolle über den Wert n hätte; er muß nur größer als eine gewisse Schwelle sein. (In den Anwen-

dungen kann z.B. n Ergebnis eines Zufallsversuchs sein. Technisch gesprochen: die Ausnahmemenge in (St), deren Ws. gegen null geht, kann für jedes n eine andere sein. Im Fall der f.s. Konvergenz (18a) aber ist sie eine monoton fallende Mengenfolge, deren Ws. gegen null strebt.

Prop. 8.3. Eine hinreichende Bedingung für f.s. Konvergenz von (X(n)) gegen 0 ist

(BC) für jedes $a > 0$ ist die Zahlenfolge $P(|X(n)| > a)$, $n \geq 0$, summierbar.

Beweis. Wenn (BC) - was für Borel-Cantelli steht - erfüllt ist, dann ist $Z(a) := \sum_n 1_{\{|X(n)|> a\}}$ mit Ws. 1 endlich (nach Beppo Levi), d.h. mit Ws. 1 sind nur endlich viele der Indikatoren unter dieser Summe ungleich null, oder : $X(n) \leq a$ gilt von einer bestimmten Stelle an. Da dies für jedes $a > 0$ richtig ist, konvergiert die Folge (X(n)) mit Ws. 1 gegen 0 . (Man wählt eine abzählbare Menge von a-Werten, die gegen null gehen; die Ausnahmemenge, wo für ir-gend ein a dieser Menge Z(a) unendlich ist, hat immer noch die Ws. null. Hier wird übrigens die maßtheoretische Fundierung der W.-Theorie benutzt: nämlich, daß eine abzählbare Vereinigung von Nullmengen wieder eine Nullmenge ist.)

Satz 8.2. (Spezialfall des starken Gesetzes der großen Zahlen) Sei X(i), i = 1 , 2 , .. eine u.i. v. Folge von beschränkten ZVen; sei $S(n) = \sum_{1\leq i\leq n} X(i)$ und $\mu = EX(1)$. Dann konvergiert die Folge (S(n)/n) fast sicher gegen die Konstante μ .

Beweis. Vermöge der Ungleichung von Bernstein-Cramér (7) und wegen der strikten Positivität von I(x) für $x \neq \mu$ folgt die Behauptung aus Prop. 3.

An der Formulierung des Satzes stört noch die Voraussetzung der Beschränktheit von X(1). Mit den Methoden dieses Kapitels käme man bei größerer Sorgfalt zwar noch ein Stück wei-ter, aber ohne die Endlichkeit von F (oder M) , zumindest in einer Umgebung von $a = 0$, bricht alles zusammen. Das folgende Beweisprinzip hilft hier weiter:

Lemma. (Maximalungleichung) Sei X(i) , $1 \leq i \leq n$, eine Familie von positiven u.i.v. Variablen; sei $S(k) = \sum_{1\leq i\leq k} X(i)$ und $\mu = EX(1)$. Dann gilt für alle $a > 0$

(19) $$P(\max\{S(k)/k : 1 \leq k \leq n\}> a) \leq \mu/a .$$

Beweis. Für jedes $k \leq n$ und reelle Zahlen s(k) , .. , s(n) gilt: unter der Bedingung $\{S(j) = s(j), k \leq j \leq n\}$ sind die ZVen X(1), .. , X(k) symmetrisch verteilt. Daher hat X(1) hat die be-dingte Erwartung s(k)/k . Oder: wenn f eine Funktion von n-k+1 Variablen ist, hat man

(20) $$E\{X(1)\cdot f(S(k), .. , S(n))\} = k^{-1} E\{S(k)\cdot f(S(k), .. , S(n))\} .$$

Sei nun $A = \{\max \{S(k)/k : 1 \le k \le n\} > a\}$ und $A(k) = \{S(k)/k > a , S(j)/j \le a \text{ für } j > k\}$. Dann sind die $A(k)$ disjunkt und ihre Vereinigung ist A . Wendet man obige Überlegung auf $f = 1_{A(k)}$ an, dann erhält man

(21) $$EX(1)\cdot 1_{A(k)} = k^{-1} ES(k)\cdot 1_{A(k)} \ge a\cdot P(A(k)) ,$$

und nach Addition über k folgt

$$\mu = EX(1) \ge EX(1)\cdot 1_A \ge a\cdot P(A) ,$$

was zu beweisen war.

Verallgemeinerungen des Lemmas: Da im Fall nicht-positiver $X(i)$ stets gilt

$$S(k)^+ \le \sum_{i\le k} X(i)^+ \text{ und } S(k)^- \le \sum_{i\le k} X(i)^- ,$$

kann man die obere und untere Überschreitungsws. separat abschätzen und erhält

(19a) $$P(\max\{|S(k)|/k : 1 \le k \le n\} > a) \le E|X(1)|/a .$$

Da in (19) oder (19a) die r.S. von n nicht abhängt, kann man jeweils auf der linkes Seite n durch unendlich ersetzen. (Das ist übrigens wieder ein maßtheoretisches Argument.)

Satz 8.3. (Starkes Gesetz der großen Zahlen) Sei $X(i)$, $i \ge 1$, eine Folge von u.i.v. Variablen, die ein erstes Moment besitzen; sei $S(n) = \sum_{1\le i\le n} X(i)$. Dann konvergiert die Folge $(S(n)/n)$ fast sicher gegen $EX(1)$.

Beweis. Fixiere a und $\epsilon > 0$. Man approximiert $X(i)$ durch $Z(i) := \max(\min(X(i),c), -c)$, d.h. schneidet von $X(i)$ die oberen und unteren Spitzen ab, wobei c so gewählt wird, daß $Y(i) := |X(i) - Z(i)|$ eine Erwartung höchstens ϵ hat. Dann gilt auch $|EX(i) - EZ(i)| \le \epsilon$. Wegen der Maximalungleichung gilt

(22) $$P(\limsup n^{-1} \sum_{i\le n} Y(i) > a) \le \epsilon/a ,$$

und wegen der schon bewiesenen Version des starken Gesetzes für beschränkte Variable

(23) $$P(\lim n^{-1} \sum_{1\le i\le n} Z(i) = EZ(1)) = 1 .$$

Aus beidem zusammen folgt

(20) $P(\limsup n^{-1} S(n) > EX(1)+\epsilon+a) \le P(\limsup n^{-1} \sum_{1\le i\le n}\{Z(i)+Y(i)\} > EZ(1)+a) \le \epsilon/a$.

Die Ungleichung zwischen den äußeren Ausdrücken gilt für jedes $\epsilon > 0$, woraus folgt

(21) $$P(\limsup n^{-1} S(n)/n > EX(1) + a) = 0 .$$

Das wiederum ist gültig für jedes $a > 0$, daher kann man weiter vereinfachen zu

(22a) $$P(\limsup n^{-1} S(n)/n > EX(1)) = 0 .$$

Aus Symmetriegründen - ersetze $X(i)$ durch $- X(i)$ - hat man die umgekehrte Abschätzung

(22b) $$P(\liminf n^{-1} S(n)/n < EX(1)) = 0 ;$$

aus beidem zusammen folgt die Behauptung des Satzes.

Historische Bemerkung. Das starke Gesetz der großen Zahlen, wie es in Satz 8.3 formuliert worden ist, findet sich in jedem Lehrbuch der W.-Rechnung; es gehört zum "klassischen" Bestand der Stochastik. Es geht auf die Anfangszeit der modernen W.-Theorie, die 20er Jahre, zurück; die Autoren sind Chintschin ([Ch]) und Kolmogorov ([K]). Im Spezialfall, daß $X(i)$ nur der Werte 0 oder 1 fähig ist, hatte schon früher E. Borel ein starkes Gesetz bewiesen ([B]). Der Zusammenhang, in welchem Borel das Problem formulierte, sieht zunächst nicht sehr stochastisch aus; er betrachtete das Einheitsintervall mit dem Lebesguemaß und zeigte, daß in der Dezimalentwicklung einer reellen Zahl die relative Häufigkeit jeder der Ziffern von 0 bis 9 für fast alle Zahlen, im Lebesgueschen Sinn, gleich $1/10$ ist. Man kann aber im Intervall [0,1] mit Lebesguemaß die Zuordnung " $x \to$ i-te Dezimale von x " als ZV auffassen; diese ZVen bilden eine durch i indizierte u.i.v. Familie, und daher bilden für jede Ziffer z die Indikatoren der Ereignisse {die i-te Dezimale ist z} ebenfalls eine u.i.v. Folge.

Die Chintschinsche Formulierung hat den ästhetischen Vorzug, sehr einprägsam zu sein, und den mathematischen Vorzug, daß ihre Voraussetzung (Existenz der Erwartung von $X(i)$) minimal ist. Man kann nämlich relativ leicht zeigen, daß für eine u.i.v. Folge aus der f.s. Konvergenz der arithmetischen Mittel die Existenz der Erwartungen folgt. Wir wollen das hier aber nicht explizit vorrechnen.

Bemerkung zum Sprachbegrauch: die Bezeichnung dieses Satzes mit dem Wort "Gesetz", der wir schon beim schwachen Gesetz der großen Zahlen begegnet sind, legt die Deutung nahe, der Satz sei quasi das mathematische Abbild eines empirischen Sachverhalts (der als Natur-"Gesetz" empfunden wird), nämlich daß die arithmetischen Mittel langer Folgen aus gleichar-tigen zufälligen Zahlen, z.B. die relativen Häu-

figkeiten einer Ziffer im Roulette, im Verlauf langer Beobachtungsdauern einem Grenzwert zustreben. Was immer der Grund für die Wort-wahl sein mag, es verdient festgehalten zu werden, daß nur wenige Sätze der Mathematik (nicht der Physik oder Chemie) es im allgemeinen Sprachgebrauch zum Rang eines "Geset-zes" gebracht haben; die Leserin ist aufgefordert, in ihrem Gedächtnis nach einschlägigen Beispielen aus Topologie, Zahlentheorie, Analysis etc. zu suchen. Dies als Motivation zum

Methodologischen Einschub. Mathematik und Realität.

Nach meiner Ansicht ist es wichtig, daß die (inbesondere die angewandte) Mathematikerin lernt, zwischen mathematischen Theoremen und empirischen Gesetzmäßigkeiten, zwischen Modell und realer Welt, gebührend zu unterscheiden. Ein mathematisches Modell kann bequem, elegant, überzeugend sein und es mag sich in vielen Anwendungssituationen bewähren. Es kann dadurch die Denkgewohnheiten der Mathematikerin so sehr prägen, daß sie die empi-rische Welt nur noch durch die Brille eines bestimmten Gefüges aus mathematischen Begrif-fen, Voraussetzungen und Sätzen zu sehen fähig ist. Wenn dies geschieht, ist Vorsicht gebo-ten; man muß versuchen, wieder Abstand von diesen Denkgewohnheiten zu gewinnen. Ein warnendes Beispiel: das Denken in Begriffen der Kontinuumsanalysis (z.B. Partielle Differentialgleichungen in Strömungsmechanik oder Elastizitätstheorie) hat um die Jahrhundertwende viele Physiker daran gehindert, sich eine atomare Struktur der Materie vorstellen zu können; das vermeintlich empirische, in Wahrheit aber mathematische Prinzip des *"natura non facit saltus"* hat es den Leuten schwer gemacht, das Plancksche Modell der diskreten Energieniveaus im Atom als Erklärung einer Beobachtung zu akzeptieren: man war so sehr in der Welt der stetigen Funktionen heimisch geworden, daß man vergessen hatte, daß sie eigentlich nur mathematische Konstrukte und nicht schlechthin "wirklich" sind. Ein "offenes" Beispiel: ein numerischer Mathematiker hat mir einmal gesagt, daß bei der erfolgreichen Ausübung seines Handwerks drei Dinge zusammenwirkten: die Physik eines Vorgangs, die richtigen Gleichungen und ein guter Algorithmus. Wenn bei ihm ein Algorithmus schlecht laufe, dann könne der Grund hierfür auch darin liegen, daß Analysis und Numerik die Natur nicht gut genug nachgebildet hätten, z.B. weil man beim Modellieren wesentliche Parameter nicht berücksichtig hat. Er sagte auch, eine Kontrolle für gute Numerik sei, daß das gerechnete oder vorausgesagte Ergebnis mit dem Experiment gut übereinstimme. Mein Kommentar: den letzte Satz billige ich, weil er ein legitimes Erfolgskriterium für eine bestimmte Art von angewandter Mathematik nennt. Ich bin jedoch skeptisch, ob "die Physik" erlaubt zu sagen, was gute Algorithmen sind. Ich glaube, hier werden die Ebenen verwischt: für mich ist ein numerischer Algorithmus gut, wenn er der Lösung der analytischen Gleichung, für die er konstruiert wurde, nahekommt; dies zu entscheiden ist ein innermathematisches Problem. - Ende des Einschubs.

Beispiel. Erneuerungsprozesse.

Wir verallgemeinern hier das in Kapitel 7 über Poissonprozesse auf der positiven Halbgeraden Gesagte. Sei

X(i), $i \geq 0$, eine u.i.v. Folge von strikt positiven ZVen; wir setzen nicht voraus, daß sie exponentiell verteilt seien, sondern nur, daß $\mu := EX(i)$ endlich ist. Wir bezeichnen mit $T(n)$ die n-te Partialsumme dieser Folge. (Interpretation: $X(i)$ ist in einem Betrieb die Lebensdauer des i-ten Geräts; wenn das i-te Gerät ausfällt, wird es sofort durch ein neues, das (i+1)-te, ersetzt. Dann ist $T(n)$ der Zeitpunkt der n-ten Erneuerung.) Vermöge Satz 8.3 hat man

(23) $$\lim_n T(n)(\omega)/n = \mu \text{ für P-fast alle } \omega .$$

Wir betrachten jetzt den Punktprozeß $\{T(1), T(2), ..\}$ auf der positiven Halbgeraden und führen wie gehabt die Zählvariablen ein:

(24) $$N(t) = \sup\{ n : T(n) \leq t\} \text{ , für jedes } t \geq 0 .$$

Als Dichte dieses Punktprozesses wird man die Anzahl der Punkte pro Längeneinheit in großen Intervallen bezeichnen, d.h. die Zahl

(25) $$\rho := \lim_t N(t)/t ,$$

falls dieser Limes existiert. (Beachte: wenn X(i) nicht exponentiell ist, kann man nicht damit rechnen, daß die $N(t)$ so etwas wie Summen von u.i.v. Variablen sind, wie im Fall des Poissonprozesses.) Aus (23) folgt durch simples punktweises (für jedes ω) Bilden der Umkehrfunktion zu $T(.)$, welche gleich $N(.)$ ist, die Beziehung

(26) $$\rho = \lim_t N(t)(\omega)/t = 1/\mu \text{ für P-fast alle } \omega .$$

Das hätte man auch nicht anders erwartet. Das starke Gesetz ist bei der hier angestellten Betrachtung dem schwachen Gesetz überlegen, da man einfach punktweise für (fast) jedes ω im Ergebnisraum die Umkehrfunktion bilden darf. Andernfalls hätte man etwas Ärger mit der Ausnahmemenge: die Menge $\{ a < N(t)/t < b\}$ läßt sich nicht einfach als $\{c < T(n)/n < d\}$ für ein festes n darstellen. Man beachte auch, daß ρ nicht von der Realisierung der zufälligen Folge abhängt, sondern mit Ws. 1 konstant ist, d.h. eine das konkrete Modell charakterisierende Zahl.

Der Birkhoffsche Ergodensatz. Für mich ist die beste Art, das starke Gesetz zu "verstehen", es als Spezialfall eines allgemeineren Satzes zu sehen. Die Folge $X = (X(i) , i \geq 1)$ hat die schöne Eigenschaft der Stationarität: ihre Verteilung ändert sich nicht, wenn man alle Indices um 1 erhöht. Man hat im Folgenraum R^N eine Abbildung, nennen wir sie T , welche $(x(1), x(2) , x(3) , ..)$ nach $(x(2) , x(3) , ..)$ sendet; Stationarität besagt, daß TX und X die gleiche Verteilung haben. In der Mechanik weiß man schon lange (Liouville), daß das (später nach Lebesgue benannte) Maß $dp \cdot dq$ im Phasenraum eines zeitlich homogenen Systems sich nicht ändert, wenn man es der Abbildung F der zeitlichen Evolution unterwirft. Diese kommt

dadurch zustande, daß man die das System definierende Differentialgleichung löst und als $F(x)$ den Zustand des Systems zur Zeit 1 nimmt, wenn es zur Zeit 0 im Zustand x war. Sei g eine Observable, d.h. eine reelle Funktion auf dem Phasenraum. Für die Grundlegung der statistischen Mechanik ist es wichtig zu wissen, ob das sog. Zeitmittel von g, d.h.

$$\lim_n 1/n \cdot \sum_{1 \le k \le n} g(F^k(x))$$

(existiert und) gleich dem sog. Raummittel von g ist, d.h. der Größe

$$\int g(x) \cdot dp \cdot dq \cdot (\int 1 \cdot dp \cdot dq)^{-1} .$$

Die Leistung von G.D. Birkhoff (1931) war erstens: zu sehen, daß man die Folge $g(F^k(x))$, $k \geq 1$, als stationäre Folge von Zufallsvariablen auf dem Phasenraum, versehen mit dem normierten Lebesguemaß, ansehen kann; zweitens: zu beweisen, daß für jede stationäre Folge von Zufallsvariablen mit existenter Erwartung die arithmetischen Mittel mit Ws. 1 einen Li-mes besitzen. Ob dieser Limes konstant und damit gleich der Erwartung (= Raummittel) ist, ist eine zweite Frage. Im Fall von u.i.v. Größen gibt viele Möglichkeiten, um sich von der Konstanz des Limes zu überzeugen; das schwache Gesetz der großen Zahlen ist eine davon. (Zur physikalischen Seite siehe [Ch2], oder ein Lehrbuch der Statistischen Mechanik, zur mathematischen Seite etwa Kap. 6 in [D].)

Literatur.

[B] E. Borel: Les probabilités dénombrables et leurs applications arithmétiques. Rend. Circ. Mat. Palermo 27, 247 (1909).

[C] H. Cramér: Sur un nouveau théorème-limite de la théorie des probabilités. Actualités Scientifiques et Industrielles 736, 5-23. Paris: Hermann 1938. (Historisch einer der ersten Belege für exponentielle Abschätzungen.)

[Ch1] A.Ja. Chintschin: Sur la loi des grands nombres. C.R. Acad. Sc. Paris 188, 477-479 (1929).

[Ch2] A.Ja. Chintschin: Mathematische Grundlagen der Statistischen Mechanik. Bibliographisches Institut: Mannheim 1965. (Russisches Original: Moskau 1941).

[D] R. Durrett: Probability: Theory and Examples. Belmont, CA: Wadsworth 1991.

[DS] J.D. Deuschel, D.W. Stroock: Large deviations. Boston: Academic Press 1989.

(Dort S. 9 ff.)

[K] A.N. Kolmogorov: Über die Summen durch den Zufall bestimmter unabhängiger Größen. Math. Annalen 99, 309-319 (1928) und 102, 484-488 (1929).

Kapitel 9. Gerichtete Perkolation

Inhalt: Problem der Perkolation. Beweis der Existenz eines Phasenübergangs.

In diesem Kapitel werden wir eine erste Einführung in ein Teilgebiet der W.-Theorie geben, das sich unter Stochastikern in den letzten Jahrzehnten steigender Beliebtheit erfreut, die Theorie der Perkolation. Sie ist aus mehreren Gründen populär: einmal weil sie geometrische und kombinatorische Strukturen in die sonst vielleicht etwas eintönige analytische Landschaft einführt, zum andern weil sie starke Berührungspunkte zur Physik hat. Man kann sie als "Spielmodell" benutzen, wenn man schwierige Fragen der Feldtheorie auf dem Weg der Analogie zunächst in einem bewußt einfachen Rahmen studieren möchte; die Pointe ist, daß man in ihr Phasenübergänge nachweisen kann, das sind Änderungen des qualitativen Verhaltens, wenn ein Parameter einen bestimmten Schwellenwert überschreitet. Wir sind diesem Phänomen bereits beim Galton-Watson-Prozeß begegnet (Kap. 2); Perkolation hat allerdings eine viel rei-chere Struktur. Wir stellen das vermutlich einfachste nichttriviale Modell vor, die gerichtete Perkolation in Dimension 2; man könnte es auch als Kontakt- (Ansteckungs-) prozeß oder als zufälligen zellularen Automaten bezeichnen. Schon die Namen machen deutlich, daß das Mo-dell für viele Anwender interessant ist, besser: daß es unabhängig von Leuten mit unterschied-lichem Hintergrund entwickelt worden ist.Wir werden die Existenz eines Phasenübergangs in Strenge beweisen.

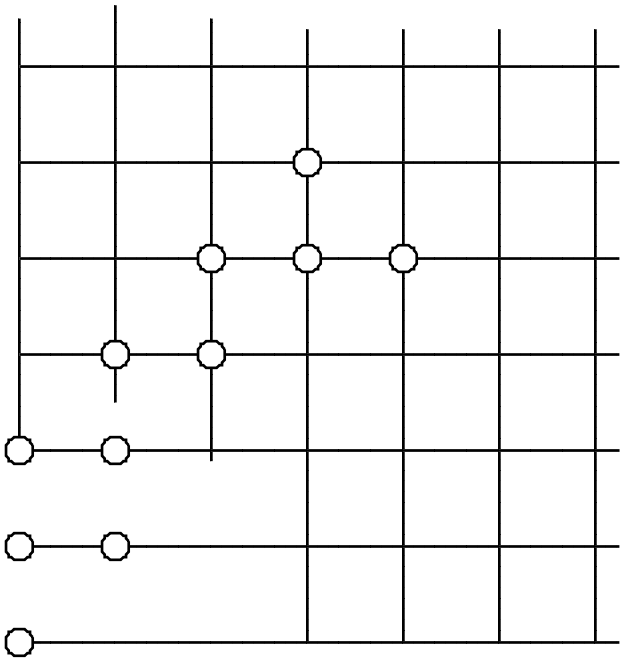
Der zugrundeliegende Raum E ist das ganzzahlige zweidimensionale Gitter. Jeder Punkt x sendet zwei Kanten aus, nach oben und nach rechts, zu den Punkten $x + (0,1)$ und $x + (1,0)$. Wir denken uns diese Kanten als gerichtet, d.h. sie dürfen nur von unten/links nach oben/ rechts durchlaufen werden. Ein Weg kann also keine "Mäander" erst recht keine Schleifen enthalten. Die Menge aller Kanten bezeichnen wir mit K .

Mit jeder Kante k assoziieren wir eine Zufallsvariable $A(k)$, welche die Werte 0 oder 1 annimmt. Wir sagen, k sei offen (passierbar) bzw. blockiert ,wenn $A(k) = 1$ bzw. 0 gilt. Wir setzen voraus:

die $A(k)$ *sind unabhängig identisch verteilt mit* $P(A(k) = 1) = p$.

Dabei ist p ein reeller Parameter zwischen 0 und 1 . Um die Abhängigkeit der W.-Bewertung von diesem Parameter auszudrücken, schreiben wir manchmal P_p statt nur P . Wenn eine Realisierung $a = (a(k))$ des Systems der $A(k)$ gegeben ist, sagen wir, ein Punkt y sei von x aus erreichbar, wenn es einen aus offenen Kanten bestehenden Weg von x nach y gibt. Die Menge aller Gitterplätze, die vom Ursprung (0,0) aus erreichbar sind, bezeichnen wir mit $C(a)$ oder C ("cluster"). Dann ist C vom Zufall abhängige Teilmenge des Gitters; wir fragen nach typischen Eigenschaften von C oder seiner Verteilung, in Abhängigkeit

von p . (Wegen anderer Fassungen des Phänomens "Perkolation", s. Bemerkung 3) am Ende.)



Zeichnung 1. Offene (fett) und blockierte Kanten; das Cluster C(a) (fette Punkte)

Zeichnung 2. Kontaktprozeß.
Fette Punkte: Raum-Zeit-Koordinaten von Infizierten; fette Kanten: Übertragungswege.

Bevor wir die mathematische Argumentation weiter verfolgen, wollen wir den Sachverhalt in eine etwas andere Sprache übersetzen, damit die Worte "Kontaktprozeß" und "zellulärer Automat" verständlich wer-

den. Wir drehen das ganzzahlige Gitter um 45 Grad. Wir interpretieren die vertikale Koordinate als Zeit und die horizontale als Raum. Zu jedem Zeitpunkt haben wir dann ein ganzzahliges eindimensionales Gitter von Raumpunkten, wobei die einzelnen Gitter zu zwei benachbarten Zeitpunkten um eine halbe Gitterkonstante gegeneinander versetzt sind. Sagen wir, die Gitterkonstante sei gleich 2 . Wir setzen zur Zeit 0 Infektionsträger beliebig in den Raum. Jeder von ihnen kann zur Zeit 1 die Infektion auf die beiden nächstgelegenen Plätze bringen, die im Abstand 1 rechts und links von ihm liegen, und zwar mit Ws. p unabhängig auf jeden dieser Plätze. Dieses Spiel setzen wir fort: zur Zeit $n+1$ wird der Gitterpunkt u genau dann infiziert, wenn mindestens einer der zwei Plätze $u \pm 1$ zur Zeit n infiziert ist und die Infektion an u weiterreicht. Das Weitergeben der Infektion von $u \pm 1$ nach u erfolgt unabhängig für verschiedene Plätze u und verschiedene Wahlen des Vorzeichens mit Ws. p . Dies zur Erklärung des Wortes "Kontaktprozeß".

Ein zellulärer Automat besteht aus einem Raum-Zeit-Gitter von Punkten, welche sich in end-lich vielen möglichen "Zuständen" befinden können. Der Zustand aller Punkte $(u,0)$ wird als Anfangswert vorgegeben. Der Zustand des Punktes $(u,n+1)$ bestimmt sich durch eine (translationsinvariante) Regel aus den Zuständen der Punkte $(u-k,n), ...,(u+k,n)$; hier ist k eine den Automaten charakterisierende Konstante, die "Reichweite". Je nachdem, ob die Regel rein de-terministisch ist oder zufällige Elemente enthält, sprechen wir von einem deterministischen oder stochastischen zellulären Automaten. Im hier betrachteten Fall gibt es zwei Zustände an jedem Platz, "infiziert" oder "gesund"; der Automat ist vom stochastischen Typ, $k = 1$.

Die Menge $\{0,1\}^K$ der möglichen Realisierungen der $A(k)$ oder Kantenkonfigurationen werde mit A bezeichnet . Sie ist in natürlicher Weise geordnet; wir sagen, daß $a \le b$ gilt, wenn $a(k) \le b(k)$ für alle $k \in K$ gilt. Wir nennen eine auf A definierte reelle Funktion (bzw. ein Ereignis, d.h. eine Teilmenge von A) monoton, wenn sie (bzw. der Indikator des Ereignisses) monoton bezüglich dieser Ordnung ist. Beispiel für eine monotone Funktion auf A ist die Kardinalität von C ; Beispiele für monotone Ereignisse sind "x ist mit y verbunden" oder "C enthält mehr als 50 Gitterpunkte" oder "C hat unendliche Kardinalität". Konkret interessiert uns in diesem Kapitel die Ws. des letztgenannten Ereignisses als Funktion von p ; genauer: wir wollen wissen, für welche p diese Funktion null ist. Dazu als einfache Vorbe-reitung

Prop. 9.1. Sei f eine monotone positive Funktion auf A . Dann ist die Zuordnung

$$p \to E_p f \tag{1}$$

eine monotone Funktion auf $[0,1]$.

Beweis. Wir konstruieren einen W.-Raum (Ω, F, Q) , auf dem wir die Modelle zu allen p-Werten simultan betrachten können, also eine Kopplung der Modelle zu verschiedenen p . Man wählt als Ω die Menge $(0,1)^K$ und als Q das unabhängige Produkt der Lebesguemaße in den einzelnen Koordinaten. Dann definiert man für jedes p die ZVen

$$A(k,p)(\omega) := 1_{\{\omega(k) < p\}} ,$$

wo $\omega(k)$ die k-te Koordinate von ω ist. Offenbar sind für jedes p die $A(k,p)$ u.i.v. mit Erwartung p. Weiter ist für festes ω und k die Zuordnung $p \to A(k,p;\omega)$ monoton und daher auch $p \to f(a(p;\omega))$ wegen der Monotonie von f. Integration bezüglich Q über ω ergibt die Behauptung.

Corollar. Die Funktion $p \to P_p(|C| = \infty)$ ist monoton im Intervall $[0,1]$. Definiert man

$$p(c) := \inf \{ p : P_p(|C| = \infty) > 0\}, \tag{2}$$

so ist $P_p(|C| = \infty) = 0$ für alle $p < p(c)$ und $P_p(|C| = \infty) > 0$ für alle $p > p(c)$.

Def. 9.1. Die Zahl $p(c)$ heißt kritischer Wert des Parameters p. Wir sagen, Perkolation finde statt, wenn C unendliche Kardinalität hat.

Soweit kommt man mit "weichen" Methoden, wie dem Ausnützen der Monotoniestruktur. Dem Corollar fehlt nur noch eine winzige Kleinigkeit (so scheint es) zur Aussage, daß ein Phasenübergang zwischen Nicht-Perkolation und Perkolation besteht, d.h. daß für gewisse p-Werte der Cluster C fast sicher endlich ist und für andere nicht: wir müssen uns nur noch da-von überzeugen, daß $p(c)$ strikt größer als 0 und strikt kleiner als 1 ist. Diese Kleinigkeit zu ergänzen ist alles andere als trivial: sie ist der eigentliche Inhalt dieses Kapitels. Dabei ist eine untere Abschätzung an $p(c)$ relativ leicht zu bekommen, während sich die obere Ab-schätzung als um einiges mühsamer herausstellt.

Prop. 9.2. (Untere Abschätzung) Es gilt $p(c) \geq 0.5$.

Beweis. Wir gehen ins Bild des Kontaktprozesses mit 1 Infizierten zur Zeit 0. Wir definieren einen zweiten Kontaktprozeß, bei welchem ein Gitterplatz mit mehr als einem Infizierten besetzt sein kann: jeder zur Zeit n an der Stelle u sitzende Infizierte erzeugt mit Ws. p jeweils 1 neuen Infizierten an der Stelle $u + 1$ und unabhängig davon an der Stelle $u - 1$; die verschiedenen Infizierten schaffen sich ihre Nachfolger in unabhängiger Weise. Dann ist offenbar die Zahl der Infizierten an jeder Stelle im zweiten Kontaktprozeß größer als im ersten Kontaktprozeß; man kann vom zweiten zum ersten übergehen, indem man zu jeder Zeit an allen Stellen mit mehr als 1 Infizierten den zweiten, dritten, ..Infizierten (zusammen mit ihren "Nachkommen") aus dem Verkehr zieht. Wenn also für ein p der zweite Kontaktprozeß mit Ws. 1 nach endlicher Zeit ausstirbt, dann auch der erste. Man sieht sofort, daß das zweite Modell ohne seine räumliche Struktur der aus Kapitel 2 bekannte Vermehrungsprozeß (spezi-eller Galton-Watson-Prozeß) ist. Die

erwartete Zahl der Nachkommen eines Individuums in ihm ist 2p . Wie bekannt, stirbt dieser Prozeß für $p \leq 0.5$ aus. Also stirbt für $p \leq 0.5$ auch der uns interessierende erste Kontaktprozeß aus.

Prop. 9.3. (Obere Abschätzung) Es gilt $p(c) \leq 1/6\{15 - \sqrt{93}\} = 0.8927...$.

Beweis. Die Beweisidee besteht grob in folgendem: man zählt alle möglichen endlichen Cluster auf, die den Punkt (0,0) als linke untere Ecke enthalten, und zeigt, daß die Summe der Wsen, daß das zufällige C gleich einem Cluster dieser Liste wird, strikt kleiner als 1 ist - sofern p größer als eine Schranke d ist. Wenn das gelingt, hat man $p(c) \leq d$ gezeigt,weil dann für $p > d$ die Menge C mit positiver Ws. unendliche Mächtigkeit hat.

Wir betrachten die nachstehende Zeichnung und stellen dazu einige geometrische Überlegungen an.

Zeichnung 3. Das Cluster C und sein Rand

Die Zeichnung zeigt ein typisches endliches C (dicke Punkte). Die dicke Linie stellt den vorderen (oder rechten/oberen) Rand des Clusters dar. Sie ist für die Gestalt des Clusters verantwortlich; sie kreuzt nur blockierte Kanten. Denn wenn eine von der dicken Linie gekreuzte Kante nicht blockiert wäre, läge ihr

rechter/oberer Endpunkt ebenfalls im Cluster. Die dünne Linie beschreibt den hinteren (oder linken/unteren) Rand des Clusters. Eine von ihr gekreuzte Kante kann offen oder blockiert sein. (Es ist für die Gestalt des hier gezeichneten Clusters vollkommen egal, ob z.B. die Kante ((4,0) ¬(4,1)) offen oder blockiert ist, da der Punkt (4,0) ohnehin nicht zum Cluster gehört.) Dicke und dünne Segmente zusammen bilden den Rand des Clusters. Dies führt uns auf die folgenden Begriffsbildungen:

Def. 9.2. Eine Kontur ist eine geschlossene Kurve mit folgenden Eigenschaften:

(3a) sie besteht aus Segmenten der Länge eins; jedes Segment hat mit einer Kante des Gra phen (E,K) den Mittelpunkt gemeinsam und verläuft senkrecht zu ihr;
(3b) kein Segment wird zweimal durchlaufen;
(3c) der Punkt (-1/2,-1/2) liegt auf der Kurve; alle anderen Punkte der Kurve liegen rechts/ oberhalb von (-1/2,-1/2) .

Als vorderen (bzw. hinteren) Abschnitt einer Kontur bezeichnet man all die Segmente, in de-nen die Außen-normale des von der Kontur berandeten Gebiets nach rechts oder oben (bzw. nach links oder unten) zeigt. Wir sagen von einer Kontur, sie trete in der Kantenkonfiguration a auf, wenn alle Segmente ihres vorderen Abschnitts eine blockierte Kante kreuzen.

Die folgenden Aussagen sind evident:

(4a) Der vordere und der hintere Abschnitt einer Kontur sind gleich lang.

(4b) Für jede Kantenkonfiguration a ist der Rand von C(a) eine Kontur, die in der Kon figu ration a auftritt.

(4c) Jede Kontur, die in a auftritt, umfaßt den Cluster C(a) in ihrem Inneren. ("Eine auftre tende Kontur wirkt als unübersteigbare Barriere für die Ausbreitung von C .")

(4d) Die Ws. dafür, daß eine bestimmte Kontur der Länge L in der zufälligen Kantenkonfi gurati on auftritt, ist unter dem Parameter p gleich $(1 - p)^{L/2}$.

Angesichts der Aussagen (4) erhalten wir eine obere Abschätzung für die Ws. , daß C endlich ist :

(5) $$P_p(|\,C\,| < \infty) \le \sum_F (1 - p)^{L(F)/2} = \sum_m |\{F : L(F) = 2m\}|\,(1 - p)^m ,$$

wobei die erste Summe über alle Konturen F geht und L(F) die Länge von F bezeichnet. (Es ist klar, daß als Längen nur gerade Zahlen auftreten.)

Es bleibt das kombinatorische Problem zu lösen, die Anzahl der Konturen der Länge $2m$ nach oben abzuschätzen. Eine grobe, für unsere Zwecke genügende, Methode besteht darin, die geschlossene Kurve F gegen den Uhrzeigersinn, beginnend beim Segment $(-1/2,-1/2) \to (1/2,-1/2)$ schrittweise aufzubauen. An jeder Stelle hat man höchstens 3 Möglichkeiten, das nächste Segment zu wählen. Da der erste, zweite und letzte Punkt des Wegs, $(-1/2,-1/2)$, $(1/2,-1/2)$, $(-1/2,1/2)$, schon vorgegeben sind, hat man höchsten 3^{2m-3} Wahlmöglichkeiten, um eine Kontur der Länge $2m$ aufzubauen. Somit gilt

(6) $$P_p(|C| < \infty) \le \sum_{m \ge 2} 3^{2m-3}(1 - p)^m = 1/27\, z^2 (1 - z)^{-1} \quad \text{mit} \quad z = 9(1 - p) .$$

Die r.S. wird kleiner als 1 , sofern $z < 3/2 \{\sqrt{93} - 9\}$ oder $p > 1/6 \{15 - \sqrt{93}\}$. Die numeri-sche Auswertung der unteren Schranke für p ergibt $0.8927..$ Damit ist Prop 3. bewiesen.

Wir fassen die Teilergebnisse zusammen zum

Satz 9.1. Es gibt eine Zahl $p(c)$ mit $0.5 \le p(c) \le 0.893$ derart, daß im Modell der zweidimen-sionalen gerichteten Perkolation zum Parameter p das den Ursprung enthaltende Cluster fast sicher endlich ist, falls $p < p(c)$, und mit positiver Ws. unendlich groß ist, falls $p > p(c)$.

Bemerkungen.

1) Den harten Teil der Aussage des Satzes bildet Prop. 9.3. Die Beweisidee, die wir hier vorgestellt haben, ist unter dem Namen Peierlssche Konturmethode bekannt; sie spielt bei Untersu-chungen zum Ising-Modell eines Ferromagneten eine fundamentale Rolle. Sie kann so zusammengefaßt werden: wenn die Verbindung von $(0,0)$ ins "Unendliche" unterbrochen ist, kann man eine Kontur aus blockierten Kanten zwischen $(0,0)$ und Unendlich legen; das Ereignis, daß eine Kontur der Länge L realisiert wird, hat eine Ws. kleiner als c^L , wobei c vom Modellparameter p abhängt. Wenn $N(L)$, die Zahl der Konturen der Länge L , nicht schneller als exponentiell anwächst, so kann $\sum N(L) \cdot c^L$ und damit die Ws., daß überhaupt eine Kontur auftritt, bei c gegen null beliebig klein gemacht werden. - Diese Überlegungen liefern das

Corollar. $P_p(|C| = \infty)$ strebt gegen 1 bei p gegen 1 .

2) Die beiden Abschätzungen in Prop. 2 und 3 sind mathematisch rigoros, aber numerisch unbefriedigend, weil noch zuviel Luft zwischen ihnen ist. Simulationen legen nahe, an ein $p(c)$ etwa bei 0.6446 zu glauben. Mit den heutigen Rechenmaschinen ist man in der Lage, den gesuchten Zahlenwert auf mehr als 3 Stellen hinter dem Komma mit hinreichender moralischer Gewißheit zu bestimmen. (Ich weiß nicht, zu welchem Kurs $p(c)$ derzeit an der Börse gehandelt wird; es ist für diesen Text auch nicht entscheidend.) Unser Beispiel ist typisch für viele Situationen in der mathematischen Physik, wo man aus Simulationen

sehr genau "weiß", was der kritische Parameterwert ist, wo aber die Mathematiker noch um etliche Zehnerpotenzen dieser Präzision des Wissens oder Glaubens hinterherhinken. Die angesprochene Diskrepanz wird besonders sichtbar bei der hier verwendeten Konturmethode. Man hat zwar neuerdings mit Methoden aus der Vorstellungswelt der zellulären Automaten eine viel bessere obere Ab-schätzung als die hier gebrachten 0.893 erzielt: in [L] wird $p(c) \leq 2/3 = 0.6666..$ bewiesen.
Trotzdem bleibt eine qualitative Diskrepanz zwischen rigoroser Mathematik und Güte der Simulationen, auch wenn sie sich um eine Zehnerpotenz verringert hat.

3) Man kann die qualitative Änderung im Verhalten des Perkolationsprozesses auch an anderen Kriterien sehen als an der Endlichkeit eines festen Clusters. Eine Alternative wäre etwa die folgende (in der Sprache des Kontaktprozesses): Man startet zur Zeit 0 mit $n+1$ Infizierten an den Plätzen $2k$ mit $0 \leq k \leq n$. Man läßt den Prozeß während einer Zeit der Dauer $2n$ laufen und schaut nach, ob es am Ende dieser Zeit noch Infizierte gibt. Wenn ja, sagen wir, der Prozeß habe überlebt. Es stellt sich heraus (was wir hier nicht beweisen), daß für $p < p(c)$ die Überlebensws. mit wachsendem n nach 0 geht, sogar exponentiell, und daß sie für $p > p(c)$ nach 1 strebt, ebenfalls exponentiell. Der Vorteil dieses Kriteriums besteht darin, daß es zeigt, wie sich das makroskopische (mit n als Skaleneinheit) Verhalten des Systems in Raum und Zeit schlagartig an der Grenze $p(c)$ ändert: eine "Störung" am Boden eines großen Rechtecks oder Trapezes bleibt ein "lokales" Ereignis für $p < p(c)$, wird aber plötzlich in ma-kroskopischer Distanz am Deckel des Rechtecks fühlbar, wenn p den kritischen Wert übersteigt. Genau das, die plötzliche Änderung des Makroverhaltens bei kleinen "Drehungen an der Schraube", nennen die Physiker einen Phasenübergang.

Wegen einer genaueren Diskussion des Phänomens Perkolation verweisen wir auf Arbeiten von Durrett, den einführenden Text [D1] und den Übersichtsartikel [D2], sowie die Monographie von Grimmett [G], in der allerdings die gerichtete Perkolation nicht abgehandelt wird.

Literatur.

[D1] R. Durrett: Lecture notes on particle systems and percolation. Monterey: Wadsworth 1988.

[D2] R. Durrett: Oriented percolation in two dimensions. Ann. Probab. 12, 999-1040 (1984).

[G] G. Grimmett: Percolation. Berlin: Springer 1989.

[L] Th.M. Liggett: Survival of discrete time growth models, with applications to oriented percolation. Ann. Appl. Probab. 5, 613-636 (1995).

Auf den folgenden drei Seiten finden sich Bilder von simulierten Kontaktprozessen im sub- , super- und kritischen Regime. Bereits bei 120 bzw. 60 Zeitschritten sind optisch die Unter-schiede zwischen den Regimen deutlich zu erkennen.