Bitte warten Sie, Sie werden verbunden!

ELKE WARMUTH, BERLIN

Zusammenfassung: *Es geht um stochastische Situationen, die man versucht, mit der Binomialverteilung zu modellieren, bei der dieses Modell aber im Allgemeinen nicht adäquat ist. Typischerweise handelt es sich um stochastische Prozesse, bei denen mehrere Komponenten zusammenwirken. Am Beispiel einer solchen Situation soll gezeigt werden, zu welchen anderen Ergebnissen ein sehr einfaches Modell der Bedienungstheorie führt und für welche Parameter sich die Ergebnisse dennoch ähneln und warum sie das tun.*

1 Eine Aufgabe und zwei Lösungen

Die Aufgabe und die Lösungen

Ein Kollege hatte folgende Aufgabe in einer Klausur gestellt:

> *Hotline:* Für mögliche technische Probleme wurde eine Hotline eingerichtet, die mit 5 Serviceberatern besetzt ist. Pro Stunde rufen im Schnitt 40 Kunden an, von denen jeder durchschnittlich 4 Min. 30 Sek. telefoniert.
> Wie groß ist die Wahrscheinlichkeit, dass ein Kunde beim Anruf der Hotline warten muss? Wie viele Serviceberater müssen beschäftigt werden, damit man als Kunde mit mindestens 95 % Sicherheit beim Anrufen sofort einen Serviceberater bekommt?

Im Folgenden betrachten wir nur die erste Frage in dieser Aufgabe.

Lösung des Kollegen: „Natürlich sind die Randbedingungen zu klären (Unabhängigkeit, keine Zeithäufung usw.). Aber ich gehe von dem Modell aus, dass $\frac{180}{300} = p$ die Auslastung beschreibt und sich die Fragestellung mit der Binomialverteilung (X zählt die belegten Leitungen) lösen lässt. Dann ist X binomialverteilt mit $n = 5$ und $p = 0{,}6$ und die gesuchte Wahrscheinlichkeit ist $P(X = 5) = 0{,}6^5$ und beträgt rund 0,08."

Nun kamen aber einige Schüler auf eine andere Idee.

Lösung von Schülern: „X zählt die anrufenden Kunden und ist binomialverteilt mit $n = 40$ und $p = \frac{4{,}5}{60} = 0{,}075$. Ein Anrufer muss warten, wenn schon mindestens fünf Kunden anrufen. Die gesuchte Wahrscheinlichkeit

$$P(X \geq 5) = 1 - \sum_{k=0}^{4} \binom{40}{k} 0{,}075^k \, 0{,}025^{40-k}$$

beträgt rund 0,18."

Gemäß der Maxime „Alle Modelle sind falsch, aber einige sind nützlich." (Sachs 1993; S. 146) wäre zu klären, ob eines der beiden Modelle nützlich ist. Beide können es ja offenbar nicht sein.

Diskussion der ersten Lösung

Interessant ist hier die Klärung der Randbedingungen. Sicherlich wird eine Hotline nicht über den ganzen Tag gleichmäßig nachgefragt sein. In der Aufgabenstellung gibt es dazu keine Aussagen. Das ist schade, denn es sollte doch wohl um Modellierung gehen und nicht um eine eingekleidete Aufgabe zur Binomialverteilung. Nehmen wir also *vereinfachend* an, dass sich das Bedienungssystem in einem sogenannten statistischen Gleichgewicht befindet. Wir werden diesen Begriff später präzisieren und stellen uns vorläufig darunter vor, dass der Kundenandrang im betrachteten Zeitintervall zwar zufälligen Schwankungen unterliegt, die Gesetze dafür aber gleichbleibend sind. Dann verlangen pro Stunde die Kunden im Mittel $40 \cdot 4{,}5 = 180$ Minuten Beratung. Dem stehen $5 \cdot 60 = 300$ Beraterminuten gegenüber. Man könnte $p = \frac{180}{300} = 0{,}6$ als Wahrscheinlichkeit dafür auffassen, dass zu einem zufällig gewählten Zeitpunkt ein bestimmter Berater besetzt ist. In der ersten Lösung werden die fünf Berater als *unabhängig voneinander* agierend und mit *derselben* Besetztwahrscheinlichkeit von 0,6 angenommen. Das ist sicher nicht gerechtfertigt, denn wenn z. B. gerade drei Berater besetzt sind, dann fällt die ganze Last, die ja gemäß Voraussetzung gleichmäßig eintrifft, auf die verbleibenden zwei Berater und schon ändert sich die Besetztwahrscheinlichkeit. Wir sehen, dass die zwei Vorgänge „Ankunft von Anrufen" und „Bedienung von Anrufen" ineinandergreifen und somit sinnvollerweise auch im mathematischen Modell miteinander verbunden werden sollten.

Diskussion der zweiten Lösung

Die Schüler nehmen vereinfachend an, dass *genau* 40 Kunden anrufen. Sie eliminieren also ein Zufallselement aus dem Vorgang „Ankunft von Anrufen". Da jeder Kunde im Mittel 4,5 Minuten im System verbleibt, wird unter der stillschweigenden Annahme, dass ein statistisches Gleichgewicht vorliegt, der Quotient $p = \frac{4{,}5}{60} = 0{,}075$ als Wahrscheinlichkeit

gedeutet, dass ein bestimmter Kunde zu einer zufällig ausgewählten Minute im System ist. Diese Wahrscheinlichkeit ist für jeden Kunden gleich groß. Die Annahme, dass die Kunden unabhängig voneinander anrufen, scheint plausibel zu sein. Damit kommen die Schüler zu dem Schluss, dass die Anzahl X der Kunden, die zu einer zufällig ausgewählten Minute im System sind, binomialverteilt mit den Parametern $n = 40$ und $p = 0{,}075$ ist. Das Problem in dieser Betrachtung stellen die Kunden dar, die alle Berater besetzt vorfinden. Sie bringen das ganze Modell durcheinander, denn es bleibt offen, was mit ihnen passiert. Sie verbleiben länger im System (nämlich um ihre Wartezeit) und somit verändert sich p. Letzten Endes läuft es auf dasselbe hinaus wie bei der ersten Lösung: die beiden zusammenwirkenden Prozesse werden nicht gemeinsam betrachtet.

Man ahnt auch schon, wann die zweite Lösung das Problem in besserer Näherung beschreiben kann. Das wird dann der Fall sein, wenn die Belastung des Systems gering ist und deshalb Wartezeiten kaum auftreten. Jedem ist aus der Erfahrung gut bekannt, dass reale Systeme diese Eigenschaft eher selten besitzen.

Konsequenzen

Im Folgenden soll anhand sehr einfacher Beispiele gezeigt werden, welche mathematischen Werkzeuge geeignet sind, Situationen wie sie in der Aufgabe Hotline beschrieben sind, zu modellieren. Ziel ist es dabei, von dem betrachteten System nicht nur eine Momentaufnahme zu machen, sondern seine zeitliche Entwicklung zu beschreiben. Dies führt auf den mathematischen Begriff des stochastischen (zufälligen) Prozesses und darunter auf die spezielle Klasse der sogenannten Geburts- und Todesprozesse. Solche Prozesse werden wir schließlich zur Modellierung der Hotline verwenden, nicht ohne die damit verbundenen Annahmen kritisch zu diskutieren. Innerhalb des gewählten Modells werden dann Fragen wie die nach der Wahrscheinlichkeit, die Hotline besetzt vorzufinden, beantwortet.

Wir wollen die Leserin und den Leser in die Thematik einführen und Anregungen zum Weiterlesen geben. Im Rahmen eines solchen Aufsatzes ist es selbstverständlich nicht möglich, die gesamte Theorie zu entwickeln. Zur weiteren Vertiefung lohnen sich – wie so oft – Engel (1976) und Krengel (2005). Eine relativ gut zugängliche fachwissenschaftliche Darstellung der Bedienungstheorie bietet Amossowa et al. (1986).

2 Stochastische Prozesse

Ein *stochastischer Prozess* ist eine Schar $(X_t)_{t\in T}$ von Zufallsgrößen X_t, die alle über demselben Wahrscheinlichkeitsraum Ω definiert sind und Werte in einem Zustandsraum E annehmen. Die Variable t steht oft für die Zeit. Für jedes $\omega \in \Omega$ stellt $(X_t(\omega))_{t\in T}$ eine sogenannte Trajektorie dar, d. h. einen möglichen Verlauf des zufälligen Geschehens in der Zeit (vgl. Abb. 1).

Die Zufallsgröße X_t misst den Wert des interessierenden Merkmals zum Zeitpunkt t. Im Hotline-Beispiel wird X_t die Anzahl der Anrufer bei der Hotline zur Zeit t angeben, und zwar einschließlich derjenigen in der Warteschleife. Der Zustandsraum ist in diesem Fall die Menge der natürlichen Zahlen $\{0, 1, 2, \ldots\}$, wobei wir idealisierend annehmen, dass die Warteschleife beliebig viele Anrufer fassen kann.

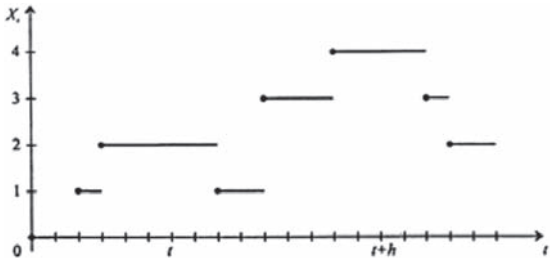

Abb. 1: Trajektorie eines stochastischen Prozesses

Abb. 1 sehr schlecht !!!

Wir betrachten hier nur stochastische Prozesse mit dem Zustandsraum $E = \{0, 1, 2, \ldots\}$ und dem Zeitintervall $T = [0, \infty)$. Aussagen über das Verhalten eines stochastischen Prozesses in der Zeit sind immer Aussagen über das Verhalten der im Allgemeinen unendlich vielen und voneinander abhängigen Zufallsgrößen X_t. Unter plausiblen Annahmen reicht es, für jede beliebige endliche Auswahl von Zeitpunkten $t_1, t_2, \ldots, t_k$ aus T und beliebige natürliche Zahlen $i_1, i_2, \ldots, i_k$ aus E die sogenannten e*ndlichdimensionalen Verteilungen*

$$P(X_{t_1} = i_1, X_{t_2} = i_2, \ldots, X_{t_k} = i_k)$$

zu kennen.

Stationarität, statistisches Gleichgewicht und Ergodizität

Ein stochastischer Prozess heißt *stationär*, wenn sich alle endlichdimensionalen Verteilungen bei einer Verschiebung auf der Zeitachse um einen beliebigen, aber festen Betrag h nicht ändern:

$$P(X_{t_1+h} = i_1, X_{t_2+h} = i_2, \ldots, X_{t_k+h} = i_k) = P(X_{t_1} = i_1, X_{t_2} = i_2, \ldots, X_{t_k} = i_k) \quad \text{(ST)}$$

Die Bedingung (ST) bedeutet nicht, dass sich der Prozess in der Zeit nicht ändert. Sie sagt aber aus, dass das zufällige Verhalten des Prozesses nicht vom Zeitpunkt des Beginns der Beobachtung abhängt. Insbesondere sind bei einem stationären Prozess die eindimensionalen Verteilungen $P(X_t = k)$ unabhängig von t, denn es gilt ja

$$P(X_{t+h} = k) = P(X_t = k).$$

Für die zweidimensionalen Verteilungen $P(X_s = i, X_t = k)$ ist bei einem stationären Prozess nur der Abstand zwischen s und t von Bedeutung, nicht aber ihre Position auf der Zeitachse.

Ein stationärer Prozess ist geeignet, ein *statistisches Gleichgewicht* zu beschreiben. Seine Verteilung nennt man stationäre Verteilung oder Gleichgewichtsverteilung. Viele reale Prozesse haben die Eigenschaft, dass ihr Verhalten nach einer gewissen Zeit (Anfahrphase) durch eine stationäre Verteilung beschrieben werden kann. Man spricht vom *Einschwingen* auf ein statistisches Gleichgewicht.

Wenn ein Einschwingen erfolgt, so können zwei Fälle auftreten:

1. Die Gleichgewichtsverteilung hängt von der Verteilung des Prozesses bei Beginn der Beobachtung ($t = 0$) ab

oder

2. sie ist unabhängig von dieser Anfangsverteilung.

Den zweiten Fall nennt man den *ergodischen*.

3 Markovsche Prozesse

Ein stochastischer Prozess $(X_t)_{t\in T}$ hat die Markov-Eigenschaft, wenn zu jedem Zeitpunkt $t \in T$ die Wahrscheinlichkeitsverteilung für seine weitere Entwicklung nur von seinem Wert X_t zur Zeit t abhängt, nicht aber davon, wie er in diesen Zustand gelangt ist. Genauer, wenn für alle Zeiten $t_1, t_2, \ldots, t_k, t \in T$ mit $t_1 < t_2 < \ldots < t_k < t$ und alle Zustände $i_1, i_2, \ldots, i_k, j \in E$ gilt

$$\begin{aligned} &P(X_t = j \mid X_{t_1} = i_1, X_{t_2} = i_2, \ldots, X_{t_k} = i_k) \\ &\quad = P(X_t = j \mid X_{t_k} = i_k), \end{aligned} \tag{M}$$

dann heißt $(X_t)_{t\in T}$ *Markovscher Prozess*.

Die bedingte Wahrscheinlichkeit des zukünftigen Verhaltens $(X_t = j)$ unter der Bedingung der Vergangenheit $(X_{t_1} = i_1, X_{t_2} = i_2, \ldots, X_{t_k} = i_k)$ hängt bei einem Markovschen Prozess nur vom jüngsten bekannten Zustand $(X_{t_k} = i_k)$ der Vergangenheit ab. Die Information über die gesamte weitere „Vorgeschichte“ $(X_{t_1} = i_1, X_{t_2} = i_2, \ldots, X_{t_{k-1}} = i_{k-1})$ spielt keine Rolle.

Ein sehr einfacher Spezialfall eines Markovschen Prozesses mit $T = \{1, 2, \ldots, n\}$ und $E = \{0, 1\}$ ist eine Bernoulli-Kette, bei der die Zufallsgrößen X_t vollständig unabhängig sind. Bei allgemeinen Markovschen Prozessen gibt es sehr wohl Abhängigkeiten zwischen den Zufallsgrößen X_t, aber sie sind durch die Bedingung (M) stark eingeschränkt.

Übergangswahrscheinlichkeiten

Die endlichdimensionalen Verteilungen eines Markovschen Prozesses sind eindeutig durch die *Anfangsverteilung* $p_i = P(X_0 = i)$, $i \in E$, und die *Übergangswahrscheinlichkeiten* $P(X_t = k \mid X_s = j)$, $s, t \in T$, $j, k \in E$ bestimmt. Wir zeigen das Konstruktionsprinzip am Beispiel der zweidimensionalen Verteilungen:

$$\begin{aligned} P(X_s = j, X_t = k) &= P(X_t = k \mid X_s = j)\, P(X_s = j) \\ &= P(X_t = k \mid X_s = j) \sum_{i\in E} P(X_s = j \mid X_0 = i)\, P(X_0 = i) \end{aligned}$$

Wenn die Übergangswahrscheinlichkeiten $P(X_t = k \mid X_s = j)$ nur von der Zeitdifferenz $t - s$ abhängen, dann heißt der Markovsche Prozess *homogen*. Wir betrachten von nun an nur homogene Markovsche Prozesse und bezeichnen für $t \geq 0$ und $j, k \in E$ mit $p_{jk}(t)$ die Wahrscheinlichkeit, in t Zeiteinheiten vom Zustand j in den Zustand k zu gelangen. Für die eindimensionalen Verteilungen eines homogenen Markovschen Prozesses folgt mit dieser Bezeichnung durch Zerlegung der Ergebnismenge nach dem Anfangszustand j dann:

$$p_k(t) = P(X_t = k) = \sum_{j\in E} p_{jk}(t)\, p_j.$$

4 Geburts- und Todesprozesse

Ein Geburts- und Todesprozess $(X_t)_{t\in T}$ ist ein Markovscher Prozess mit Werten in E, bei dem in kurzen Zeitintervallen im Wesentlichen nur drei Szenarien möglich sind: der Zustand kann sich um 1 erhöhen (Geburt, Anruf kommt an), er kann sich um 1 verringern (Tod, Anruf endet) oder er bleibt gleich.

Um zu präzisieren, was mit „im Wesentlichen“ gemeint ist, verwenden wir die sehr nützliche Symbolik $o(h)$[1]. Eine Funktion $t(h)$ heißt von der Größenordnung $o(h)$ für $h \to 0$, symbolisch $t(h) = o(h)$ für $h \to 0$, falls $\lim_{h\to 0} \frac{t(h)}{h} = 0$. Man sagt auch, $t(h)$ geht

schneller gegen 0 als h. Diese Symbolik ermöglicht es, elegant und zweckgebunden „vernachlässigbar kleine“ Terme zusammenzufassen. Beispielweise gilt

$e^h = 1 + h + o(h)$ für $h \to 0$.

Mit dieser Konvention können wir nun einen Geburts- und Todesprozess durch folgende *infinitesimale Übergangswahrscheinlichkeiten* für $h \to 0$ beschreiben:

$$\begin{aligned} p_{k\,k+1}(h) &= \lambda_k h + o(h), && k \geq 0 \\ p_{k\,k-1}(h) &= \mu_k h + o(h), && k \geq 1 \\ p_{k\,k}(h) &= 1 - \lambda_k h - \mu_k h + o(h), && k \geq 0 \end{aligned} \qquad (\text{Ü})$$

Die *Übergangsintensitäten* λ_k und μ_k sind nichtnegative reelle Zahlen, wobei $\mu_0 = 0$ vereinbart wird. Der Name Intensität rührt daher, dass es sich gerade um die Ableitungen der Übergangswahrscheinlichkeiten in den Nachbarzustand an der Stelle $t = 0$ handelt. Das ist aus den ersten beiden Gleichungen von (Ü) gut ersichtlich, wenn wir beachten, dass $p_{jk}(0) = 0$ für $j \neq k$ ist. Hier bewährt sich auch unter dieser Perspektive das Symbol $o(h)$. Aus (Ü) ergibt sich außerdem, dass die Wahrscheinlichkeit von Änderungen von X_t um mehr als 1 in jedem Intervall der Länge h von der Größenordnung $o(h)$ ist, denn die Summe der drei Übergangswahrscheinlichkeiten beträgt $1 + o(h)$.

Die Änderungen eines Geburts- und Todesprozesses können in einem sogenannten *Übergangsgraphen* veranschaulicht werden:

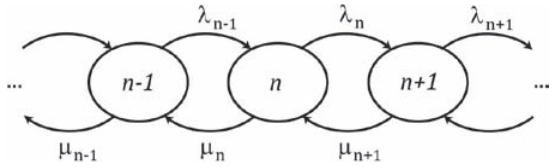

Abb. 2: Übergangsgraph Geburts- und Todesprozess

Differentialgleichungen für Geburts- und Todesprozesse

Mit Hilfe der infinitesimalen Übergangswahrscheinlichkeiten leiten wir nun ein System von Differentialgleichungen für die Zustandswahrscheinlichkeiten $p_k(t) = P(X_t = k)$ her. Aus diesem System von Gleichungen, die die Funktionen $p_k(t)$ samt ihren Ableitungen enthalten, werden wir dann für gewisse Geburts- und Todesprozesse durch Grenzübergang ein lineares Gleichungssystem gewinnen, das es uns ermöglicht, die Gleichgewichtsverteilung zu bestimmen.

Wir teilen das Intervall $[0; t + h]$ in die Teilintervalle $[0; t]$ und $[t; t + h]$ ein. Wenn zum Zeitpunkt $t + h$ der Zustand k vorliegen soll, dann muss gemäß (Ü) zum Zeitpunkt t entweder $k - 1$ oder k oder $k + 1$ vorgelegen haben. Mit der Formel für die totale Wahrscheinlichkeit und (Ü) erhalten wir für $k \geq 1$

$$\begin{aligned} p_k(t+h) &= p_k(t)(1 - \lambda_k h - \mu_k h + o(h)) \\ &\quad + p_{k-1}(t)(\lambda_{k-1} h + o(h)) \\ &\quad + p_{k+1}(t)(\mu_{k+1} h + o(h)). \end{aligned}$$

Wir sortieren die Terme auf der rechten Seite mit dem Ziel, p_k' zu bestimmen, geeignet um:

$$\begin{aligned} p_k(t+h) &= p_k(t) - (\lambda_k + \mu_k)\, p_k(t) h \\ &\quad + \lambda_{k-1}\, p_{k-1}(t) h + \mu_{k+1}\, p_{k+1}(t) h \\ &\quad + o(h)(p_k(t) + p_{k-1}(t) + p_{k+1}(t)). \end{aligned}$$

Für den Differenzenquotienten erhalten wir

$$\begin{aligned} &\frac{p_k(t+h) - p_k(t)}{h} \\ &= -(\lambda_k + \mu_k) p_k(t) + \lambda_{k-1}\, p_{k-1}(t) + \mu_{k+1}\, p_{k+1}(t) \\ &\quad + \frac{o(h)}{h}(p_k(t) + p_{k-1}(t) + p_{k+1}(t)). \end{aligned}$$

Der Grenzübergang für $h \to 0$ liefert mit Hilfe bekannter Grenzwertsätze und der definierenden Eigenschaft von $o(h)$ schließlich

$$p_k'(t) = -(\lambda_k + \mu_k) p_k(t) + \lambda_{k-1}\, p_{k-1}(t) + \mu_{k+1}\, p_{k+1}(t).$$

Im Fall $k = 0$ beachten wir, dass der Zustand 0 nicht durch „Geburt“ entstehen kann und erhalten auf analoge Weise

$$p_0'(t) = -\lambda_0\, p_0(t) + \mu_1\, p_1(t).$$

Das nun vollständige System von Differentialgleichungen (D) ergänzen wir um eine Anfangsverteilung $(p_i)_{i \in E}$.

Gleichgewichtsverteilung für Geburts- und Todesprozesse

Wir interessieren uns nur für den ergodischen Fall, d. h. für eine von der Anfangsverteilung unabhängige stationäre Lösung des Systems (D). Eine solche stationäre Lösung existiert nicht immer. Es könnte z. B. sein, dass sich der Prozess in einem Teil des Zustandsraums „festsetzt“, der vom Anfangszustand abhängt. Wenn wir aber fordern, dass sämtliche Übergangsintensitäten positiv sind, dann folgt daraus, wie

man sich am Übergangsgraphen veranschaulichen kann, dass mit positiver Wahrscheinlichkeit jeder Zustand in E von jedem anderen in einer endlichen Anzahl von Übergängen erreicht werden kann. Solche Geburts- und Todesprozesse nennt man *irreduzibel*. Für irreduzible Geburts- und Todesprozesse gilt folgender

Ergodensatz: Ein irreduzibler Geburts- und Todesprozess ist ergodisch, wenn die Reihe

$$\sum_{k=1}^{\infty} \frac{\mu_1 \mu_2 \ldots \mu_k}{\lambda_1 \lambda_2 \ldots \lambda_k}$$

divergiert und die Reihe

$$\sum_{k=1}^{\infty} \frac{\lambda_0 \lambda_1 \ldots \lambda_{k-1}}{\mu_1 \mu_2 \ldots \mu_k}$$

konvergiert, vgl. z. B. König & Stoyan 1986.

Sei nun der Geburts- und Todesprozess $(X_t)_{t \in T}$ ergodisch. Wir bezeichnen mit $(\pi_k)_{k \in E}$ die eindimensionale stationäre Verteilung, also die Verteilung, die sich durch „Einschwingen“ einstellt:

$$\pi_k = \lim_{t \to \infty} p_k(t).$$

Wenn diese Grenzwerte existieren, dann existieren auch die Grenzwerte der Ableitungen und zwar gilt $\lim_{t \to \infty} p_k'(t) = 0$ für alle k. Das Differentialgleichungssystem (D) geht nach dem Grenzübergang in ein lineares Gleichungssystem über:

$$0 = -\lambda_0 \pi_0 + \mu_1 \pi_1$$
$$0 = -(\lambda_k + \mu_k)\pi_k + \lambda_{k-1} \pi_{k-1}(t) + \mu_{k+1} \pi_{k+1}(t),\ k \geq 1.$$

Die Struktur dieser Gleichungen legt es nahe, folgende äquivalente Umformungen durchzuführen: Man behält die erste Gleichung bei, ersetzt die zweite durch die Summe aus den ersten beiden Gleichungen, die dritte durch die Summe aus den ersten drei usw. Es entsteht das Gleichungssystem

$$0 = -\lambda_0 \pi_0 + \mu_1 \pi_1$$
$$0 = -\lambda_k \pi_k + \mu_{k+1} \pi_{k+1},\ k \geq 1.$$

Für π_k erhalten wir die rekursive Gleichung

$$\pi_{k+1} = \frac{\lambda_k}{\mu_{k+1}} \pi_k,\ k \geq 0 \text{ bzw. } \pi_k = \frac{\lambda_{k-1}}{\mu_k} \pi_{k-1},\ k \geq 1.$$

Die schrittweise Rückführung liefert

$$\pi_k = \frac{\lambda_0 \lambda_1 \ldots \lambda_{k-1}}{\mu_1 \mu_2 \ldots \mu_k} \pi_0,\ k \geq 1.$$

Mit Hilfe der Normierungsbedingung $\sum_{k=0}^{\infty} \pi_k = 1$ bestimmen wir den Anfangswert

$$\pi_0 = \left(\sum_{k=1}^{\infty} \frac{\lambda_0 \lambda_1 \ldots \lambda_{k-1}}{\mu_1 \mu_2 \ldots \mu_k} + 1 \right)^{-1}.$$

Damit haben wir die einzige Gleichgewichtsverteilung im ergodischen Fall gefunden.

5 Modellierung der Hotline

Wir werden nun die Hotline als speziellen Geburts- und Todesprozess modellieren und unsere Ausgangsfrage im Rahmen dieses Modells beantworten. Wie schon in Abschnitt 1 betont, überlagern sich in der Hotline zwei Prozesse: Zu zufälligen Zeitpunkten treffen Anrufe ein, die eine Anrufdauer von zufälliger Länge auslösen. Manche Anrufe geraten darüber hinaus in die Warteschleife und werden vom nächsten freiwerdenden Serviceberater bedient. Der Anruf- und der Bedienprozess greifen ineinander. Der Zustand des Systems Hotline zur Zeit t sei die Anzahl X_t der Anrufer, die sich zum Zeitpunkt t in der Hotline befinden, einschließlich derer in der Warteschleife. Vorerst lassen wir die Anzahl der Serviceberater variabel und bezeichnen sie mit s.

Anrufprozess

Wir nehmen an, dass die Abstände T_k zwischen aufeinanderfolgenden Anrufen unabhängige, mit demselben Parameter λ exponentialverteilte Zufallsgrößen sind. Wir beginnen die Beobachtung zum Zeitpunkt $t = 0$ und bezeichnen für $t > 0$ mit N_t die Anzahl der Anrufe im Intervall $(0, t]$. Die obigen Annahmen kennzeichnen den Anrufprozess (N_t) als Poisson-Prozess. Als interessante Lektüre zum Poisson-Prozess sei Topsøe (1990) empfohlen.

Die Unabhängigkeit erscheint im Hotline-Beispiel plausibel. Der konstante Parameter λ bedeutet eine starke Vereinfachung, wie aus der folgenden Interpretation von λ deutlich wird. Wir deuten den Erwartungswert λ^{-1} von T_k als Vorhersage für den mittleren Abstand zwischen zwei aufeinanderfolgenden Anrufen bei vielen Beobachtungen des Anrufprozesses und gewinnen daraus eine Interpretation von $\lambda = \frac{1}{\lambda^{-1}}$ als mittlere Anzahl von Anrufen pro *Zeiteinheit*. Diese mittlere Anzahl ist also konstant. Im Modell der Hotline wählen wir als Zeiteinheit 1 Minute und setzen gemäß den Daten der Aufgabe $\lambda = \frac{40}{60} = \frac{2}{3}$.

Für eine mit dem Parameter λ exponentialverteilte Zufallsgröße T gilt

$$P(T > h) = e^{-\lambda h} = 1 - \lambda h + o(h),$$
$$P(T \le h) = 1 - e^{-\lambda h} = \lambda h + o(h) \qquad \text{(o)}$$

für $h \to 0+$. Von entscheidender Bedeutung wird die sogenannte *Gedächtnislosigkeit der Exponentialverteilung* sein. Sie besagt im konkreten Fall, dass die Wahrscheinlichkeit dafür, dass in den nächsten v Zeiteinheiten kein Anruf ankommt, wenn seit dem letzten Anruf bereits u Zeiteinheiten vergangen sind, dieselbe ist wie die Wahrscheinlichkeit, dass der Abstand zum nächsten Anruf mehr als v Zeiteinheiten beträgt. In Formeln:

$$P(T > u + v \mid T > u) = P(T > v).$$

Man kann es auch so formulieren: Wenn ich bereits u Zeiteinheiten auf den neuen Anruf vergeblich warte, dann ist der *restliche Abstand* zum neuen Anruf genauso verteilt wie der „frische" Abstand. Dies ist eine weitere sehr einschneidende Eigenschaft. Mit ihrer Hilfe kann gezeigt werden, dass der Prozess $(N_t)_{t\ge 0}$ [2] ein Markovscher Prozess ist.

Bedienprozess

Über die zufälligen Gesprächsdauern der Anrufer nehmen wir an, dass sie voneinander unabhängig und identisch exponentialverteilt mit dem Parameter μ sind. Die Unabhängigkeit ist sicher eine unproblematische Annahme. Wir nehmen zusätzlich vereinfachend an, dass sich die Kunden hinsichtlich der *Verteilung* ihrer Gesprächsdauern nicht unterscheiden. Mit der Annahme der Exponentialverteilung als Gesprächsdauerverteilung hängt wiederum die Markov-Eigenschaft zusammen.

Wenn G die zufällige Gesprächsdauer eines Anrufers bezeichnet, dann gilt analog zu T:

$$P(G > h) = e^{-\mu h} = 1 - \mu h + o(h),$$
$$P(G \le h) = 1 - \mathrm{e}^{-\mu h} = \mu h + o(h).$$

Wir notieren, dass der Erwartungswert von G gleich $\frac{1}{\mu}$ ist, d. h. im Mittel über viele Anrufe dauert ein Gespräch etwa $\frac{1}{\mu}$ Zeiteinheiten. Damit können wir den zweiten Parameter für unser Modell festlegen:

$$\frac{1}{\mu} = 4{,}5.$$

Weiterhin nehmen wir (ohne Bedenken) an, dass die Folge der Zwischenankunftszeiten (T_j) unabhängig ist von der Folge der Gesprächsdauern (G_k). Da sowohl T als auch G stetige Zufallsgrößen sind, ist die Wahrscheinlichkeit dafür, dass *gleichzeitig* ein Anrufer kommt und ein Gespräch endet, gleich Null.

Aus den bisherigen Annahmen lässt sich ableiten, dass der Prozess $(X_t)_{t\ge 0}$, der die Anzahl der Anrufer in der Hotline beschreibt, ein Markovscher Prozess ist. Wir führen hier keinen formalen Beweis, sondern berufen uns inhaltlich auf die Gedächtnislosigkeit der Exponentialverteilung. Wenn wir die Hotline zu einem Zeitpunkt t_k im Zustand i_k vorfinden, dann wissen wir, wie viele Anrufe laufen und wie viele Anrufe gegebenenfalls in der Warteschleife sind. Die Wahrscheinlichkeit für einen Zustand j zu einem zukünftigen Zeitpunkt $t > t_k$ hängt nur von der restlichen Wartezeit auf den nächsten Anruf und den restlichen Gesprächsdauern ab. Wir müssen aber nicht wissen, wann diese Zeiten begonnen haben, d. h. wir müssen keinen Vorgängerzustand von i_k kennen, weil die aufeinanderfolgenden exponentialverteilten Gesprächsdauern ihre Vergangenheit vergessen. Zum Zeitpunkt t_k beginnen sie „von vorn" zu laufen.

Die Konstanz der Parameter λ und μ sichert die Homogenität des Markovschen Prozesses.

Intensitäten λ_k und μ_k

Wir untersuchen nun die Übergangswahrscheinlichkeiten von $(X_t)_{t\ge 0}$ für kurze Zeitintervalle der Länge h und gewinnen daraus die Intensitäten.

Dazu überlegen wir uns zuerst, dass die Wahrscheinlichkeit dafür, dass innerhalb $[t, t + h]$ zwei oder mehr Anrufe ankommen ebenso wie diejenige, dass zwei oder mehr Gespräche enden, von der Größenordnung $o(h)$ ist. Allgemein können wir das so formulieren:

Hilfssatz: Wenn U und V unabhängige exponentialverteilte Zufallsgrößen sind, so gilt: $P(U + V \le h) = o(h)$ für $h \to 0$.

Beweis: Wir zerlegen das Ereignis $\{U + V \le h\}$ so, dass wir neben der Eigenschaft der Exponentialverteilung die Unabhängigkeit ins Spiel bringen können:

$$\begin{aligned} P(U + V \le h) &= P(U + V \le h \mid U \le 0{,}5h)P(U \le 0{,}5) \\ &\quad + P(U + V \le h \mid 0{,}5h < U \le h)P(0{,}5h < U \le h) \\ &\le P(V \le h \mid U \le 0{,}5h)P(U \le 0{,}5h) \\ &\quad + P(V \le 0{,}5h \mid 0{,}5h < U \le h)P(0{,}5h < U \le h) \end{aligned}$$

Wegen der Unabhängigkeit von U und V folgt

$$P(U+V\le h)\le P(V\le h)P(U\le 0{,}5h) + P(V\le 0{,}5h)P(0{,}5h<U\le h)\,.$$

Wir benutzen nun wiederholt die Eigenschaft (o) (vgl. S. 3) und beachten, dass $o(h)\pm o(h)=o(h)$ ist:

$$\begin{aligned}P(U+V\le h) &\le (\lambda h+o(h))(0{,}5\lambda h+o(h))\\ &\quad +(0{,}5\lambda h+o(h))(\lambda h+o(h)-0{,}5\lambda h+o(h))\\ &= 0{,}5\lambda^2h^2+o(h)+0{,}25\lambda^2h^2+o(h)\\ &= o(h)\end{aligned}$$

Damit ist der Hilfssatz bewiesen.

Es ist außerdem deutlich geworden, wie sich die Aussage für mehr als zwei Summanden beweisen lässt.

Bei einem Geburts- und Todesprozess finden in kurzen Zeitintervallen nur Änderungen des Zustands um höchstens 1 statt. Diese Änderungen werden nach der im Hilfssatz formulierten Vorüberlegung dadurch ausgelöst, dass *ein* Gespräch endet und kein Anruf ankommt bzw. *ein* Anruf ankommt und kein Gespräch endet. Alle anderen Szenarien haben die Wahrscheinlichkeit $o(h)$. Im Folgenden bezeichnen wir die restlichen Gesprächsdauern immer mit G und die restlichen Zwischenankunftszeiten mit T.

Sei $X_t=1$, d. h. ein Anrufer wird bedient. Der Systemzustand wechselt auf 0, wenn das Gespräch endet und kein Anruf ankommt. Die übrigen Fälle fassen wir in $o(h)$ zusammen. Wegen

$$\begin{aligned}p_{10}(h) &= P(G\le h,\,T>h)+o(h)\\ &= P(G\le h)P(T>h)+o(h)\\ &= (\mu h+o(h))(1-\lambda h)+o(h)\\ &= \mu h+o(h)\end{aligned}$$

für $h\to 0$ erfolgt der Übergang vom Zustand 1 zum Zustand 0 mit der Intensität $\mu_1=\mu$.

Sei nun $X_t=k$ mit $k\le s$, d. h. k Anrufer werden bedient und kein Anrufer befindet sich in der Warteschlange. Der Systemzustand wechselt genau dann auf $k-1$, wenn eines der Gespräche endet und kein neuer Anruf ankommt. Das endende Gespräch ist dasjenige mit der kürzesten Gesprächsdauer $G=\min(G_1, G_2, \ldots, G_k)$.

Es lohnt sich, die Verteilung von G in einem Hilfssatz voranzustellen:

Hilfssatz: Wenn $G_1, \ldots, G_k$ unabhängige mit dem Parameter μ exponentialverteilte Zufallsgrößen sind, so gilt:

$$P(\min(G_1,\ldots,G_k)\le h)=k\mu h+o(h) \text{ für } h\to 0\,.$$

Beweis: Es ist günstig, zum Gegenereignis überzugehen:

$$\begin{aligned}P(\min(G_1,\ldots,G_k)\le h) &= 1-P(\min(G_1,\ldots,G_k)>h)\\ &= 1-P(G_1>h,\ldots,G_k>h)\\ &= 1-P(G_1>h)\cdot\ldots\cdot P(G_k>h)\\ &= 1-e^{-\mu h}\cdot\ldots\cdot e^{-\mu h}=1-e^{-k\mu h}\\ &= 1-(1-k\mu h+o(h))=k\mu h+o(h)\end{aligned}$$

Beim Übergang von der zweiten zur dritten Zeile haben wir die Unabhängigkeit der Gesprächsdauern benutzt.

Nach dieser Vorarbeit können wir die Übergangswahrscheinlichkeit berechnen:

$$\begin{aligned}p_{k\,k-1}(h) &= P(\min(G_1,G_2,\ldots,G_k)\le h,\,T>h)+o(h)\\ &= P(\min(G_1,G_2,\ldots,G_k)\le h)P(T>h)+o(h)\\ &= (k\mu h+o(h))(1-\lambda h+o(h))+o(h)\\ &= k\mu h+o(h)\,.\end{aligned}$$

Somit haben wir $\mu_k=k\mu$ für $1\le k\le s$ gefunden.

Wenn mehr als s Kunden im System sind, dann „laufen“ s Gespräche und die übrigen $k-s$ Anrufer befinden sich in der Warteschleife und leisten keinen Beitrag zur Übergangsintensität, also gilt $\mu_k=s\mu$ für $k>s$.

Gleichgewichtsbedingung

Wir prüfen, ob für die so definierten Übergangsintensitäten die Bedingungen des Ergodensatzes aus Abschnitt 3 erfüllt sind. Dabei lassen wir die ersten $s-1$ Glieder der Einfachheit halber weg; sie entscheiden ohnehin nicht über Konvergenz oder Divergenz. Somit sollte die Reihe

$$\sum_{k=s}^{\infty}\frac{s!s^{k-s}\mu^k}{\lambda^k}=\frac{s!}{s^s}\sum_{k=s}^{\infty}\left(\frac{s\mu}{\lambda}\right)^k$$

divergieren und die Reihe

$$\sum_{k=s}^{\infty}\frac{\lambda^k}{s!s^{k-s}\mu^k}=\frac{s^s}{s!}\sum_{k=s}^{\infty}\left(\frac{\lambda}{s\mu}\right)^k$$

konvergieren. Die geometrische Reihe $\sum_{k=s}^{\infty}\left(\frac{\lambda}{s\mu}\right)^k$ konvergiert genau dann, wenn der (positive) Quotient $\frac{\lambda}{s\mu}$

kleiner als 1 ist. Dann aber divergiert die Reihe $\sum_{k=s}^{\infty}\left(\frac{s\mu}{\lambda}\right)^k$, da ihre Summanden sämtlich größer als 1 sind. Wir halten als Ergebnis fest:

Satz: Im Modell Hotline existiert eine Gleichgewichtsverteilung, wenn $\frac{\lambda}{s\mu} < 1$ gilt.

Die Bedingung $\frac{\lambda}{s\mu} < 1$ ist überdies sehr plausibel, wenn wir sie in der Form $\lambda < s\mu$ schreiben. Im Mittel kommen λ Anrufe pro Zeiteinheit an. Die s Berater können im Mittel $s\mu$ Anrufer pro Zeiteinheit bedienen. Das System kann sich einschwingen, wenn die mittlere Belastung kleiner als die mittlere Kapazität ist. Der Quotient $\varrho = \frac{\lambda}{\mu}$ heißt in der Bedienungstheorie *Belastungskoeffizient*. Die Ergodizitätsbedingung bedeutet, dass der Belastungskoeffizient pro Serviceberater kleiner als 1 ist. Schon im Fall $\varrho = s$ existiert keine Gleichgewichtsverteilung und die Warteschlange wächst unbegrenzt.

Gleichgewichtsverteilung

Wir spezialisieren nun die im Abschnitt 4 gefundene Gleichgewichtsverteilung für das Hotline-Modell, wobei wir $s = 5$ setzen und ϱ vorerst noch variabel lassen. Wir müssen die Fälle $0 \le k \le 5$ und $k > 5$ unterscheiden. Im ersten Fall erhalten wir

$$\pi_k = \frac{\lambda^k}{\mu \cdot 2\mu \cdot \ldots \cdot k\mu}\,\pi_0 = \frac{\varrho^k}{k!}\,\pi_0,\ 0 \le k \le 5.$$

Im zweiten Fall $k > 5$ sind die Intensitäten konstant und es folgt

$$\pi_k = \frac{\lambda^k}{\mu \cdot 2\mu \cdot \ldots \cdot 5\mu(5\mu)^{k-5}}\,\pi_0 = \frac{\varrho^k}{5!5^{k-5}}\,\pi_0,\ k > 5$$

mit

$$\pi_0^{-1} = \sum_{k=0}^{5}\frac{\varrho^k}{k!} + \frac{5^5}{5!}\sum_{k=6}^{\infty}\left(\frac{\varrho}{5}\right)^k$$

$$= \sum_{k=0}^{5}\frac{\varrho^k}{k!} + \frac{\varrho^6}{5!(5-\varrho)}.$$

Diesen Term vereinfacht ein Computeralgebrasystem zu

$$\frac{\varrho^4 + 8\varrho^3 + 36\varrho^2 + 96\varrho + 120}{24(5-\varrho)}$$

und wir erhalten

$$\pi_0 = \frac{24(5-\varrho)}{\varrho^4 + 8\varrho^3 + 36\varrho^2 + 96\varrho + 120}.$$

Wir merken an, dass die Gleichgewichtsverteilung nicht von λ und μ individuell, sondern nur von deren Quotienten abhängt.

Beispiel: Hotline-Modell für $\varrho = \frac{\lambda}{\mu} = \frac{\frac{2}{3}}{\frac{1}{4{,}5}} = 3$

Die Gleichgewichtsverteilung für $0 \le k \le 11$ lautet auf zwei Nachkommastellen genau

$\pi_0 = 0{,}05$ $\quad \pi_1 = 0{,}14$ $\quad \pi_2 = 0{,}21$

$\pi_3 = 0{,}21$ $\quad \pi_4 = 0{,}16$ $\quad \pi_5 = 0{,}09$

$\pi_6 = 0{,}06$ $\quad \pi_7 = 0{,}03$ $\quad \pi_8 = 0{,}02$

$\pi_9 = 0{,}01$ $\quad \pi_{10} = 0{,}01$ $\quad \pi_{11} = 0{,}00$.

Die Wahrscheinlichkeit, dass ein Anrufer warten muss – die *Besetztwahrscheinlichkeit* – beträgt in diesem Modell

$$p_W = \sum_{k=5}^{\infty}\pi_k = 1 - \sum_{k=0}^{4}\pi_k = 0{,}23.$$

Dieser Wert unterscheidet sich drastisch vom Wert der ersten Lösung und ist auch deutlich verschieden vom Wert der zweiten Lösung. Es ist eine Lösung in einem Modell, das aber den Vorteil hat, die Dynamik der Hotline – wenn auch sehr vereinfacht – zu erfassen. Es ist eine mögliche Lösung der Aufgabe zu den gegebenen Daten.

Hotline unter verschiedener Belastung

Wir wollen die Besetztwahrscheinlichkeit p_W im Markovschen Modell mit der Besetztwahrscheinlichkeit b_W im Binomialmodell mit $n = 40$ und Erfolgswahrscheinlichkeit $p = 0{,}075$ (2. Lösung in Abschnitt 1) in Abhängigkeit von ϱ vergleichen. Da es nur auf das Verhältnis ankommt, lassen wir λ fest und variieren μ. Zwischen p und μ besteht (vgl. Abschnitt 1) der Zusammenhang $p = \frac{\mu^{-1}}{60}$. Ersetzen wir μ^{-1} durch $\frac{3}{2}\varrho$, so folgt $p = \frac{\varrho}{40}$. Das ist wiederum ein plausibles Ergebnis, denn die Belastung verteilt sich gleichmäßig auf die 40 gleichartigen Kunden.

Den Funktionsterm für die Besetztwahrscheinlichkeit p_W in Abhängigkeit von ϱ vereinfacht ein Computeralgebrasystem folgendermaßen:

$$p_W(\varrho) = 1 - \sum_{k=0}^{4}\pi_k(\varrho)$$

$$= \frac{\varrho^5}{\varrho^4 + 8\varrho^3 + 36\varrho^2 + 96\varrho + 120}.$$

Für b_W in Abhängigkeit von ϱ gibt es keine schöne Vereinfachung. Wir haben

$$b_W(\varrho) = 1 - \sum_{k=0}^{4} B\left(k; 40, \frac{\varrho}{40}\right).$$

Wir schauen uns die Graphen der beiden Funktionen an. Sinnvoll sind Argumente im Intervall (0; 5):

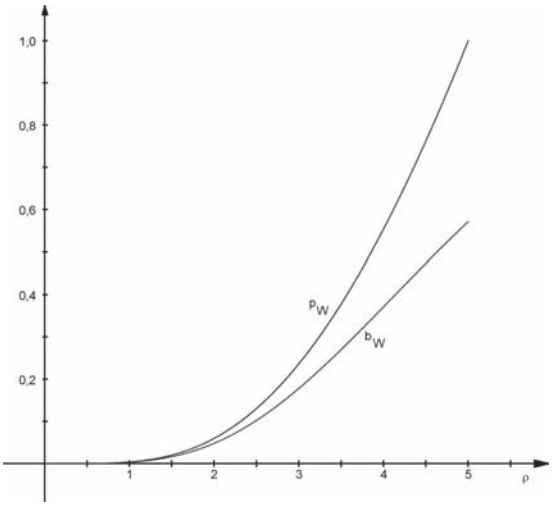

Abb. 3: Vergleich der Besetztwahrscheinlichkeiten

Der obere Graph gehört zu p_W. Nun wird deutlich, dass sich für kleine Belastungskoeffizienten die Funktionen sehr ähneln, während sie mit wachsendem ϱ immer weiter auseinanderdriften. Das war zu erwarten, da die Verknüpfung zwischen Ankunfts- und Bedienprozess umso schwächer ist, je geringer die Belastung ist. Bei sehr geringer Belastung passieren die allermeisten Anrufe die Hotline, ohne warten zu müssen, und folglich ist das Binomialmodell mit den unabhängig und mit derselben Besetztwahrscheinlichkeit agierenden Beratern gar nicht so schlecht. Wenn sich allerdings die Belastung ihrer Grenze $\varrho = 5$ nähert, dann nähert sich der Erwartungswert der Binomialverteilung ebenfalls 5, was bedeutet, dass im Mittel so viele Anrufer wie Berater im System sind, und das führt zu starken Verknüpfungen der beiden Prozesse. Ein Wert von unter 0,6 für die Besetztwahrscheinlichkeit (siehe Abb. 3) ist unglaubwürdig.

6 Ausblick

Wir haben im Modell Hotline lediglich die Besetztwahrscheinlichkeit betrachtet. Man könnte die Erwartungswerte der Anzahl der Kunden im System bzw. der Anzahl der Kunden in der Warteschleife berechnen. Durch Variieren von s könnte man auch die zweite Frage aus der Aufgabenstellung beantworten. Sicher ist dabei der Computereinsatz sinnvoll.

Wir haben die Hotline als Geburts- und Todesprozess modelliert. Sobald man eine der Annahmen über Exponentialverteilung fallen lässt, werden die Verhältnisse schlagartig schwierig. Ebenso verhält es sich mit anderen Warteschlangendisziplinen. Es kommt ja manchmal noch vor, dass vor jedem Schalter eines Amtes eine separate Warteschlange eröffnet wird. Wir haben diese Situation mit Schülerinnen und Schülern in einer Sommerschule modelliert (vgl. Bericht der Sommerschule 2010). Es gelingt nicht, explizite Lösungen für das entstehende Gleichungssystem anzugeben. In solchen Fällen bleibt als Ausweg oft die Simulation. Das ist aber keine Notlösung, sondern eine weit verbreitete und anerkannte Methode, Probleme zu lösen. Sie erfordert ihrerseits Modellbildungskompetenzen, denn simulieren kann man grundsätzlich nur auf der Basis eines Modells.

Anmerkungen

1 Die *o*-Symbolik geht auf Edmund Landau (1877–1938) zurück und wird auch Landau-Symbolik genannt.

2 Wir setzen $N_0 = 0$.

Literatur

Amossowa, N. N.; Gillert, H.; Küchler, U.; Maximow, J. D. (1986): Bedienungstheorie: Eine Einführung. Leipzig: Teubner.

Engel, A. (1976): Wahrscheinlichkeitsrechnung und Statistik, Band 2. Stuttgart: Ernst Klett.

König, D; Stoyan, D. (1986): Methoden der Bedienungstheorie. Leipzig: Teubner.

Krengel, U. (2005): Einführung in die Wahrscheinlichkeitstheorie und Statistik. Wiesbaden: Vieweg.

Sachs, L. (1993): Statistische Methoden – Planung und Auswertung. Berlin/Heidelberg: Springer.

Topsøe, F. (1990): Spontane Phänomene. Stochastische Modelle und ihre Anwendungen. Braunschweig: Vieweg.

Warmuth, E. (1997): Mathematische Modelle diskreter stochastischer Systeme. In: *Lehreinheit Diskrete Simulation*. Hrsg. Land Rheinland-Pfalz: Ministerium für Bildung, Wissenschaft und Weiterbildung.

Bericht der Sommerschule „Lust auf Mathematik“ (2010): http://didaktik1.mathematik.hu-berlin.de/index.php?article_id=373&clang=0. (Zugriff: 01.11.2010)

Anschrift der Verfasserin

Elke Warmuth
Institut für Mathematik
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
warmuth@math.hu-berlin.de