Galton-Watson-Prozesse

Thomas Hotz

FG Wahrscheinlichkeitsrechnung und mathematische Statistik
Institut für Mathematik, TU Ilmenau

28. November 2019

Am Beispiel gewisser Verzweigungsprozesse werden einige stochastische Ansätze zur Analyse von Systemen, die zufälligen Einflüssen unterliegen, vorgestellt. Rot markierte Textstellen erfordern eine mathematische Präzisierung.

Das Modell

Wir betrachten **Galton-Watson-Prozesse**; das sind gewisse **Verzweigungsprozesse**, die ein Bevölkerungswachstum modellieren: Zum Zeitpunkt $t = 0$ (0-te Generation) bestehe die Bevölkerung aus einer vorgegebene Anzahl $X_0 = x_0 \in \mathbf{N}$ von Individuen (Mütter). Wir setzen dann für $t \in \mathbf{N}_0$ rekursiv $X_{t+1} = \sum_{i=1}^{X_t} Z_i^{t+1}$ ($t+1$-te Generation) im Falle $X_t \geq 1$ und $X_{t+1} = 0$ sonst, das heißt im Falle $X_t = 0$; dabei sei Z_i^t, $i, t \in \mathbf{N}$ eine Familie unabhängiger und identisch verteilter (u.i.v.) Zufallsvariablen mit Werten in $\mathbf{N}_0 = \mathbf{N} \cup \{0\}$, wobei das i-te Individuum (Mutter) der t-ten Generation gerade Z_i^{t+1} (direkte) Nachkommen (Töchter) habe. Die Verteilung der Z_i^t heißt entsprechend **Nachkommenverteilung**; sie ist gegeben durch die Angabe der Wahrscheinlichkeiten dafür, dass ein gewisses Individuum (Mutter) $k \in \mathbf{N}_0$ (direkte) Nachkommen (Töchter) hat, $\mathbf{P}(Z_i^t = k) = p_k \in [0, 1]$, wobei $\sum_{k \in \mathbf{N}_0} p_k = 1$ (Gesamtwahrscheinlichkeit 1) erfüllt sein muss.

Man wird sich nun fragen, wie sich die Bevölkerung im Verlauf der Zeit entwickelt, wobei die Schwierigkeit offenbar darin besteht, dass dies von den konkreten **Realisierungen** der Zufallsvariablen Z_i^t abhängen wird, welche *a priori* eben nicht bekannt (da zufällig) sind. Der Fall $p_0 = 1$ ist dabei offenbar uninteressant: Dann würde $\mathbf{P}(X_1 = 0) = 1$ gelten, die Bevölkerung also sofort **fast sicher** (das heißt mit Wahrscheinlichkeit 1) aussterben. Wir gehen daher im Folgenden von $p_0 < 1$ aus.

Man beachte hierbei, dass zwar die Nachkommen einzelner Individuen (stochastisch) unabhängig sind (ihre Anzahlen sich lax gesprochen also nicht gegenseitig beeinflussen), die Gesamtzahlen X_t verschiedener Generationen t jedoch stark voneinander abhängig: X_{t+1} wird (mit großer Wahrscheinlichkeit) größer sein, wenn X_t groß ist. Die Analyse des Verhaltens dieses Prozesses über die Zeit ist daher nicht trivial.

Rechtfertigung eines deterministischen Modells

Wir betrachten die mittlere Zahl an Nachkommen eines gewissen Individuums, $\mu = \mathbf{E}\, Z_i^t = \sum_{k \in \mathbf{N}_0} k\, p_k$. Im Falle $\mu < \infty$, wovon wir ausgehen wollen, sagt man, der **Erwartungswert** der Nachkommenverteilung **existiert** (dies bedeutet, dass wir verlangen, dass die p_k für $k \to \infty$ schnell genug abfallen). Obige Bedingung $p_0 < 1$ impliziert dann gerade $\mu > 0$.

Deterministische Modelle für Bevölkerungsentwicklungen werden typischerweise durch **asymptotische** Betrachtungen für große Bevölkerungszahlen gerechtfertigt. Tatsächlich kann man beweisen:

Gesetz der großen Zahlen: Ist Y_i, $i \in \mathbf{N}$ eine Folge unabhängiger und identisch verteilter, reellwertiger Zufallsvariablen mit $\mathbf{E}\,|Y_i| < \infty$, so gilt $\mathbf{P}\big(\frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow{n \to \infty} \mathbf{E}\, Y_1\big) = 1$.

Existiert darüberhinaus das **zweite Moment** $\mathbf{E}\, Y_1^2 < \infty$, so lässt sich sogar zeigen:

Zentraler Grenzwertsatz: Sei $\sigma^2 = \mathbf{Var}(Y_1) = \mathbf{E}\,Y_1^2 - (\mathbf{E}\,Y_1)^2 > 0$; dann gilt für jedes $c > 0$: $\mathbf{P}\big(\sqrt{n}|\frac{1}{n}\sum_{i=1}^n Y_i - \mathbf{E}\,Y_1| > \sigma c\big) \xrightarrow{n\to\infty} 2\Phi(-c)$, wobei $\Phi : \mathbf{R} \to (0,1)$ bijektiv und monoton wachsend (die Verteilungsfunktion der Standardnormalverteilung) ist, also $\lim_{c\to\infty} \Phi(-c) = 0$ erfüllt.

Für großes $X_0 = x_0$ kann man wegen $X_1 = \sum_{i=1}^{x_0} Z_i^1$ nun unter der Annahme $\mathbf{E}(Z_i^t)^2 = \sum_{k\in\mathbf{N}_0} k^2 p_k < \infty$ (was sicherlich erfüllt ist, da in der Realität $p_k = 0$ für hinreichend großes $k \in \mathbf{N}$ gelten wird) argumentieren, dass für hinreichend großes $c > 0$ mit großer Wahrscheinlichkeit $\sqrt{x_0}|\frac{1}{x_0}X_1 - \mu| \le c$ gelten wird, also $\frac{|X_1 - x_0\mu|}{x_0\mu} \le \frac{c}{\sqrt{x_0}\mu}$. Die (relative) Abweichung zwischen X_1 und $\mu x_0 = \mathbf{E}\,X_1$ ist für großes x_0 also vernachlässigbar.

Iterativ gelangt man nun über die Betrachtung des Erwartungswertes zu einem **deterministischen Modell**: $\mathbf{E}\,X_{t+1} = \mu\,\mathbf{E}\,X_t = \mu^{t+1}x_0$. In diesem sind drei Fälle zu unterscheiden:

$\mu > 1$: exponentielles Wachstum (des Erwartungswertes) von X_t;

$\mu = 0$: (der Erwartungswert von) X_t bleibt konstant;

$\mu < 1$: exponentielle Abnahme des Erwartungswertes von X_t, sodass hier nicht davon ausgegangen werden kann, dass eine rein deterministische Betrachtung des Erwartungswertes für das Verständnis des Langzeitverhaltens genügen wird.

Im letzten Fall gilt irgendwann $0 < \mathbf{E}\,X_t \ll 1$, was $X_t = 0$ mit großer Wahrscheinlichkeit impliziert, sodass der Prozess dann wohl aussterben wird. Wir wollen nun mit Mitteln der Stochastik das Langzeitverhalten dieses Galton-Watson-Prozesses studieren.

Aussterbewahrscheinlichkeit

Galton und Watson selbst hatten dieses Modell betrachtet, um zu berechnen, unter welchen Bedingungen Familiennamen aussterben (X_t war die Anzahl der Männer der t-ten Generation mit diesem Familiennamen). Entsprechend betrachten wir nun die Wahrscheinlichkeit dafür, dass die Bevölkerung ausstirbt, $\mathbf{P}(X_t = 0 \text{ für ein } t \in \mathbf{N})$. Im Falle $x_0 = 1$, wovon wir im Folgenden ausgehen wollen, bezeichnen wir diese **Aussterbewahrscheinlichkeit** mit ρ; andernfalls beträgt sie ρ^{x_0}, da dann x_0 unabhängige Verzweigungsprozesse aussterben müssen.

Natürlich wäre im Falle des Aussterbens auch interessant, wann die Bevölkerung ausstirbt, genauer vielleicht $\mathbf{P}(X_t > 0 \text{ für alle } t \le n)$ für $n \in \mathbf{N}$. Dies werden wir hier aber nicht näher untersuchen.

Während der Erwartungswert also nie 0 wird, ist die Wahrscheinlichkeit dafür, dass die Bevölkerung ausstirbt, im Falle $p_0 = \mathbf{P}(Z_i^t = 0) > 0$ stets positiv (andernfalls wäre es ebenfalls uninteressant) und man wird sich fragen, wie groß sie ist.

Im Falle $\mu < 1$ ist dies leicht zu beantworten: $1 - \rho \le \mathbf{P}(X_t > 0) = \sum_{k\in\mathbf{N}} \mathbf{P}(X_t = k) \le \sum_{k\in\mathbf{N}_0} k\,\mathbf{P}(X_t = k) = \mathbf{E}\,X_t = \mu^t \to 0$ für $t \to \infty$, also gilt dann $\rho = 1$, das heißt die Bevölkerung stirbt in diesem Fall mit Wahrscheinlichkeit 1 aus.

Schwieriger ist der Fall $\mu = 1$ zu behandeln, wobei wir den Trivialfall $p_1 = 1$ (jedes Individuum bekommt genau einen Nachkommen) durch die Annahme $p_0 > 0$ ausgeschlossen haben. Dann hat der Zuwachs $X_{t+1} - X_t = \sum_{i=1}^{X_t}(Z_i^{t+1} - 1)$ gegeben X_t Erwartungswert 0; man sagt, $X_t, t \in \mathbf{N}_0$ sei ein Martingal. Für solche kann man zeigen:

Martingalkonvergenzsatz: Sei Y_t, $t \in \mathbf{N}_0$ ein nichtnegatives Martingal. Dann existiert $Y_\infty \ge 0$ mit $\mathbf{P}(Y_t \xrightarrow{t\to\infty} Y_\infty) = 1$.

In unserem Fall nimmt X_t nur Werte in $\mathbf{N}_0$ an, kann also nur gegen X_∞ konvergieren, wenn X_t schließlich konstant X_∞ wird. $X_\infty > 0$ kommt dann aber nicht in Frage, da $p_0 > 0$ in jedem Zeitschritt zum Aussterben mit Wahrscheinlichkeit $p_0^{X_\infty} > 0$ führen würde, sodass die Wahrscheinlichkeit, dass der Prozess konstant bleibt, höchstens $\lim_{n\to\infty}(1 - p_0^{X_\infty})^n = 0$ ist.

Schließlich betrachten wir noch den Fall $\mu > 1$; in diesem gilt $\rho = \mathbf{P}(X_t = 0 \text{ für ein } t > 0) < 1$, wobei ρ als eindeutig bestimmter Fixpunkt im Intervall $[0,1)$ der **erzeugenden Funktion** der Nachkommenverteilung, $\varphi : [0,1] \to [0,1]$, $\theta \mapsto \sum_{k=0}^{\infty} p_k \theta^k$ für $p_k = \mathbf{P}(Z_i^t = k)$ (mit der Konvention $0^0 = 1$), gegeben ist.

Tatsächlich: Die Ableitungen von φ erfüllen (Differentiation und Summation sind vertauschbar) $\varphi'(\theta) = \sum_{k=1}^{\infty} k p_k \theta^{k-1} \geq 0$ sowie $\varphi''(\theta) = \sum_{k=2}^{\infty} k(k-1) p_k \theta^{k-2} \geq 0$, ja sogar $\varphi''(\theta) > 0$ für $\theta > 0$, da $\mu > 1$ schon $p_k > 0$ für ein $k > 1$ impliziert. φ ist also (streng) monton wachsend und (strikt) konvex mit $\lim_{\theta \uparrow 1} \varphi'(\theta) = \sum_{k=1}^{\infty} k p_k = \sum_{k=0}^{\infty} k p_k = \mu > 1$. Ferner gilt $\varphi(0) = p_0 < 1$ (wegen $\mu > 1$) sowie $\varphi(1) = 1$. Daher existiert ein eindeutig bestimmter Fixpunkt $\rho \in [0,1)$ von φ mit $\varphi(\rho) = \rho$: Auf dem Intervall $[0,\rho]$ gilt dann $\varphi(\theta) \geq \theta$, hingegen $\varphi(\theta) \leq \theta$ auf dem Intervall $[\rho, 1]$.

Wir betrachten nun $\theta_t = \mathbf{P}(X_t = 0)$. Damit dieses Ereignis eintritt, müssen alle Individuen der ersten Generation zum Zeitpunkt t ausgestorben sein; diese bilden X_1 viele u.i.v. Verzweigungsprozesse, nämlich mit Wahrscheinlichkeit p_k gerade k Stück, welche dann nach $t-1$ Zeiteinheiten ausgestorben sein müssen, was (totale Wahrscheinlichkeit!) $\theta_t = \sum_{k=0}^{\infty} p_k \theta_{t-1}^k = \varphi(\theta_{t-1})$ ergibt (wegen $\theta_0 = 0$ und $\theta_1 = p_0$ gilt dies für alle $t \in \mathbf{N}$).

Die Folge θ_t, $t \in \mathbf{N}$ wächst damit schwach monoton und ist (Induktion!) durch ρ beschränkt. Da φ stetig ist, erfüllt ihr Grenzwert $\theta_\infty = \varphi(\theta_\infty)$, er muss also mit ρ übereinstimmen; es gilt aber gerade $\mathbf{P}(X_t > 0 \text{ für alle } t > 0) = \mathbf{P}\big(\cap_{t=1}^{\infty} \{X_t > 0\}\big) = \lim_{t\to\infty} \mathbf{P}(X_t > 0) = \lim_{t\to\infty}(1-\theta_t) = 1 - \rho > 0$.

Beispielsweise sei die Nachkommenverteilung eine Binomialverteilung mit Parametern $n \in \mathbf{N}$ und $q \in (\frac{1}{n}, 1)$, das heißt $p_k = \mathbf{P}(Z_i^t = k) = \binom{n}{k} q^k (1-q)^{n-k}$ für $k \in \{0, \ldots, n\}$, $p_k = 0$ sonst, dann gilt $\mu = nq > 1$ sowie $\varphi(\theta) = \sum_{k=0}^{n} \theta^k \binom{n}{k} q^k (1-q)^{n-k} = (\theta q + 1 - q)^n$, sodass der Fixpunkt ρ die Nullstelle eines Polynoms n-ten Grades ist.

Konkret gilt im Fall $n = 2$ dann $(\rho q + 1 - q)^2 = \rho$, das heißt $q^2\rho^2 + \big(2q(1-q) - 1\big)\rho + (1-q)^2 = 0$; da 1 auch ein Fixpunkt ist, ergibt sich also $\rho = \frac{(1-q)^2}{q^2}$. Bekäme also jeder Vater 2 Kinder, wovon in Deutschland jedes mit Wahrscheinlichkeit ca. $q = 51{,}3\%$ ein Junge wäre (wobei wir unrealistischerweise die Geschlechter der Kinder unabhängig modellieren und davon ausgehen, dass gerade die Jungen ihren Familiennamen behalten und selbst wieder Vater werden), so würde die von ihm gegründete Linie von Familiennamen mit Wahrscheinlichkeit 90% aussterben. Startet man aber mit einer großen Bevölkerung x_0, so ist die Aussterbewahrscheinlichkeit $0{,}9^{x_0}$ zwar nicht null, aber sehr klein.

Epidemien

Als Anwendung betrachten wir die Ausbreitung einer Epidemie in einer Bevölkerung von $n+1$ Individuen, $n \in \mathbf{N}_0$, welche wir durch $B = \{0, \ldots, n\}$ indizieren. Zu Beginn sei nur Individuum 0 infiziert, das heißt wir setzen $I_0 = \{0\}$, wobei I_t die Menge der „Infizierten“ zum Zeitpunkt $t \in \mathbf{N}_0$ bezeichne. Ferner betrachten wir die Menge E_t der aus der Bevölkerung zum Zeitpunkt $t \in \mathbf{N}_0$ „entfernten“ – weil sie isoliert wurden, geheilt und damit resistent oder verstorben sind. Wir nehmen an, dass die Dauer der Krankheit eine Zeiteinheit beträgt, sodass wir mit $E_0 = \emptyset$ beginnen und $E_{t+1} = E_t \cup I_t$ für $t \in \mathbf{N}_0$ erhalten. Die übrigen Individuen $A_t = B \setminus (I_t \cup E_t)$ sind die zur Zeit $t \in \mathbf{N}_0$ für die Krankheit „Anfälligen“. Eine Teilmenge $I_{t+1} \subseteq A_t$ von diesen wird sich infizieren, was $A_{t+1} = A_t \setminus I_{t+1}$, $t \in \mathbf{N}_0$ ergibt. Man wird dann danach fragen, wie viele Personen je erkranken, das heißt die Mächtigkeit von $K = \cup_{t\in\mathbf{N}_0} I_t = \cup_{t\in\mathbf{N}_0} E_t$ ist zu bestimmen.

Wir gehen vom einfachsten denkbaren stochastischen Modell aus und nehmen an, dass für eine Übertragungswahrscheinlichkeit $q \in (0,1)$ unabhängige, **Bernoulli-verteilte** Zufallsvariable $Y_{i,j}$ mit $\mathbf{P}(Y_{i,j} = 1) = 1 - \mathbf{P}(Y_{i,j} = 0) = q$, $i, j \in B$ existieren. Dabei modelliert $Y_{i,j} = 1$, dass Individuum j Individuum i getroffen hat und von diesem infiziert wurde; wir setzen also $I_{t+1} = \{j \in A_t : \text{es gibt ein } i \in I_t \text{ mit } Y_{i,j} = 1\}$ für $t \in \mathbf{N}_0$.

Es ist nahe liegend, I als Verzweigungsprozess anzusehen, wobei die Nachkommenverteilung eine Binomialverteilung mit Parametern n und q ist. Dabei macht man allerdings

drei Fehler: Erstens stehen als Nachkommen eines Individuums nicht alle n übrigen Individuen zur Verfügung, sondern nur noch die Anfälligen; zweitens könnten zwei Infizierte denselben Anfälligen anstecken; und drittens wirken sich diese beiden Fehler auch auf die weiteren Nachfahren aus. Um den Unterschied beschreibbar zu machen, betrachten wir weitere unabhängige, mit Parameter q Bernoulli-verteilte Zufallsvariablen $Z_{i,j}$, $i \in \mathbf{N}_0$, $j \in B$ und setzen $X_0 = 1$ sowie für die den ersten Fehler ausgleichenden „Geburten“ $G_{t+1} = \sum_{i \in I_t} \sum_{j \in B \setminus (A_t \cup \{i\})} Z_{i,j}$, für die den zweiten Fehler ausgleichenden „Mehrfachinfektionen“ $M_{t+1} = \sum_{i \in I_t} \sum_{j \in A_t} Y_{i,j} - |I_{t+1}|$, für die „Folgefehler“ $F_{t+1} = \sum_{i=n+1}^{n+X_t-|I_t|} \sum_{j=1}^{n} Z_{i,j}$ und schließlich $X_{t+1} = \sum_{i \in I_t} \sum_{j \in A_t} Y_{i,j} + G_{t+1} + F_{t+1} \geq |I_{t+1}|$ für $t \in \mathbf{N}_0$. Wie man sich durch Betrachtung der entsprechenden bedingten Verteilungen klar macht, ist X dann ein Verzweigungsprozess für die gewünschte Nachkommenverteilung mit Erwartungswert $\mu = nq$.

Für $\mu > 1$ unterscheiden sich Epidemie und Verzweigungsprozess qualitativ: Die Epidemie endet nämlich, wenn keine Infizierten mehr Anfällige anstecken, spätestens zum Zeitpunkt $n + 1$ mit $E_{n+1} = B$. Für $\mu < 1$ und großes n kann man hingegen zeigen, dass die oben genannten Fehler vernachlässigbar sind, sodass der Fall $\mu = 1$ wieder der kritische ist.

Im Fall $\mu > 1$ würde man in der Epidemiologie von einer **Epidemie** sprechen, im Fall $\mu = 1$ von einer **Endemie**. Ferner bezeichnet man dann μ als **Basisreproduktionsrate**. Wird ein gewisser Anteil $p \in (0, 1)$ der Bevölkerung geimpft, so starten wir mit nur $\tilde{n} = (1 - p)n$ Anfälligen, woraus sich $\tilde{\mu} = q\tilde{n} = q(1 - p)n = (1 - p)\mu$ ergibt. Als **kritische Impfrate** bezeichnet man dann den Anteil $p_c = 1 - \frac{1}{\mu}$, für den sich gerade $\tilde{\mu} = 1$ ergibt; wird ein größerer Anteil der Bevölkerung als dieser geimpft, so stirbt die Krankheit aus!

Beispielsweise beträgt für Pocken $\mu \approx 5$, also $p_c \approx 80\%$, sodass diese Krankheit durch Reihenimpfungen in den 1970er Jahren erfolgreich ausgerottet werden konnte; die Werte für Kinderlähmung sind ähnlich, sodass diese durch eine entsprechende Impfpflicht ebenfalls ausgerottet werden könnte. Für Masern beträgt die Basisreproduktionsrate hingegen $\mu \approx 20$, was $p_c \approx 95\%$ (weltweit!) erfordern würde, was kaum machbar ist (man denke an Impfgegner, aber auch an Personen, die aus gesundheitlichen Gründen nicht geimpft werden können); für Malaria ergibt sich in Teilen Afrikas sogar $\mu \geq 1000$, sodass mehr als 99,9% der Bevölkerung geimpft werden müssten.

Natürlich muss man μ zunächst bestimmen, das heißt aus Daten schätzen, die damit verbundene (stochastische) Unsicherheit quantifizieren, wofür man wiederum das Gesetz der großen Zahlen sowie den zentralen Grenzwertsatz bemühen könnte...

Literatur

Klenke, A. (2006). *Wahrscheinlichkeitstheorie*, 3rd edn, Springer, Berlin.

Kretzschmar, M. and Wallinga, J. (2010). Mathematical models in infectious disease epidemiology, *in* A. Krämer, M. Kretzschmar and K. Krickeberg (eds), *Modern Infectious Disease Epidemiology*, Statistics for Biology and Health, Springer, New York, chapter 12.