Prof. Dr. Matthias Birkner JGU Mainz

Stochastische Modelle der Populationsbiologie: Übungen und Ergänzungen **WS 2022/23**

Blatt 6

Aufgabe 6.1 (Genealogische Variabilität selbst simulieren) Schreiben Sie ein Prgramm in einer Programmiersprache Ihrer Wahl, das einen Kingman-n-Koaleszenten simuliert und das Resultat graphisch ausgibt.

Aufgabe 6.2 (Asymptotische Normalität für Summen von unabhängigen Bernoulli-Variablen nach Le Cam und die Anzahl Typen in einer n-Stichprobe)
(a) Sei $n \in \mathbb{N}$, $p_1, \dots, p_n \in [0,1]$, $\lambda_n := p_1 + \cdots + p_n$. Konstruieren Sie (auf einem geeigneten Wahrscheinlichkeitsraum) Zufallsvariablen $B_1, \dots, B_n$ und X, so dass gilt

(i) $B_1, \dots, B_n$ sind unabhängig, B_j ist Bernoulli(p_j)-verteilt (d.h. $p_j = \mathbb{P}(B_j = 1) = 1 - \mathbb{P}(B_j = 0)$)

(ii) X ist Poisson(λ_n)-verteilt

(iii) $\mathbb{P}(B_1 + \cdots + B_n \neq X) \leqslant \sum_{j=1}^n p_j^2$

(b) Seien $B_1, B_2, \dots$ unabhängig, B_j Bernoulli(p_j)-verteilt mit einem $p_j \in [0,1]$ für $j = 1, 2, \dots$, für $n \in \mathbb{N}$ setzen wir $Z_n := \sum_{j=1}^n B_j$. Es gelte

$$\sum_{j=1}^{\infty} p_j = \infty, \qquad \sum_{j=1}^{\infty} p_j^2 < \infty \tag{1}$$

Dann gilt für $-\infty \leqslant a < b \leqslant \infty$

$$\lim_{n\to\infty} \mathbb{P}\Big(a \leqslant \frac{Z_n - \mathbb{E}[Z_n]}{\sqrt{\mathrm{Var}[Z_n]}} \leqslant b\Big) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

d.h. Z_n ist asymptotisch normalverteilt.

(c) Wir betrachten einen n-Koaleszenten, längs dessen Ästen mit Rate $\theta/2$ Mutationen auftreten. Sei K_n die Anzahl verschiedener Typen an den Blättern, wenn wir das infinitely-many-alleles-Modell zugrunde legen. Zeigen Sie, dass dann K_n für $n \to \infty$ asymptotisch normalverteilt ist.

Hinweise/Diskussion. (a) Sie können beispielsweise unabhängige $X_1, \dots, X_n$ mit $X_j \sim \text{Poisson}(p_j)$ verwenden und damit geeignete Versionen der B_js zu definieren (im Jargon: verwenden Sie eine „Kopplung“). Für $0 < p_j \ll 1$ ist dann $\mathbb{1}(X_j > 0)$ Bernoulli-verteilt mit Parameter $1 - e^{-p_j} = p_j + O(p_j^2)$, was schon „fast“ das Richtige liefert. Um den Erfolgsparameter exakt auf p_j einzustellen, können Sie beispielsweise $\mathbb{1}(X_j > 0) + \mathbb{1}(X_j = 0)\widetilde{B}_j$ betrachten, wobei die $\widetilde{B}_j$ unabhängig sind (und auch unabhängig von den X_j) mit $\widetilde{B}_j \sim \text{Bernoulli}\big(1 - e^{p_j}(1-p_j)\big)$. Wie wahrscheinlich ist das Ereignis $\big\{\mathbb{1}(X_j > 0) + \mathbb{1}(X_j = 0)\widetilde{B}_j \neq X_j\big\}$?

(b) Verwenden Sie (a) und die Tatsache, dass gemäß zentralem Grenzwertsatz für Poisson(λ)-verteilte Zufallsvariablen X_λ gilt, dass $(X_\lambda - \lambda)/\sqrt{\lambda}$ für $\lambda \to \infty$ in Verteilung gegen die Standardnormalverteilung konvergiert.

Beachten Sie: Die Schranke aus (a), Teil *(iii)* ist (nur dann) nützlich, wenn die Summe $\sum p_j^2$ klein ist (bzw. gegen 0 konvergiert, wenn man die Schar in n betrachtet), was nicht wörtlich aus der Annahme (1) folgt. Man kann aber $j_n \in \mathbb{N}$ und eine Nullfolge $\varepsilon_n \to 0$ wählen, so dass

$$\frac{\sum_{j=j_n}^n p_j}{\sum_{j=1}^n p_j} \underset{n\to\infty}{\longrightarrow} 1 \quad \text{und} \quad \sum_{j=j_n}^n p_{n,j}^2 \leqslant \varepsilon_n$$

gilt.

Die Schranke aus (a) stammt aus einer Arbeit von Lucien Le Cam, An approximation theorem for the Poisson binomial distribution, *Pacific J. Math.* 10 (1960), 1181–1197. Man kann darüberhinaus sogar ein X mit den Eigenschaften aus (a) konstruieren, so dass gilt

$$\mathbb{P}(B_1 + \cdots + B_n \neq X) \leqslant \frac{\sum_{j=1}^n p_j^2}{\sum_{j=1}^n p_j} \tag{2}$$

wie Andrew D. Barbour und Peter Hall in dem Artikel On the rate of Poisson convergence, *Math. Proc. Cambridge Philos. Soc.* 95 (1984), 473–480 zeigen. Wenn Sie möchten, verwenden Sie (2) anstelle von Teil (a), das erspart das oben diskutierte „Abschneiden“ der Summanden bis j_n. Andererseits sprengt der Beweis von (2) den Rahmen dieser Übungsaufgabe. (Tatsächlich lässt sich (2) relativ übersichtlich mit der sogenannten Chen-Stein-Methode zeigen, siehe z.B. A.D. Barbour, L. Holst, S. Janson, *Poisson approximation*, Clarendon Press, 1992 oder N. Ross, Fundamentals of Stein's method, *Probab. Surv.* 8 (2011), 210–293, `https://doi.org/10.1214/11-PS182`. Bei Interesse sind Sie natürlich ganz herzlich eingeladen, sich das anzuschauen.)

Aufgabe 6.3 (Eine „Dualitätsformel“ und eine Konstruktion der Wright-Fisher-Diffusion à la Evans) Wir betrachten eine Schar von Cannings-Modellen, im Modell mit Populationsgröße N sei der Nachkommensvektor $(\nu_1^{(N)}, \ldots, \nu_N^{(N)})$. Es gebe 2 Typen, $X_g^{(N)}$ bezeichne die Anzahl Typ-1-Individuen in Generation g. Weiter betrachten wir für eine Stichprobe im N-ten Modell (zur Zeit 0, sagen wir) den Blockzählprozess, d.h. $B_g^{(N)}$ ist die Anzahl Individuen in Generation $-g$, die Vorfahren irgendeines Individuums aus der Stichprobe sind.

(a) Warum gilt für $x_0^{(N)} \in \{0, 1, \ldots, N\}$, $g \in \mathbb{N}_0$ und $n \in \mathbb{N}$ (mit $n \leqslant N$) die folgende Formel?

$$\mathbb{E}_{x_0^{(N)}}\left[\frac{X_g^{(N)}(X_g^{(N)}-1)\cdots(X_g^{(N)}-n+1)}{N(N-1)\cdots(N-n+1)}\right] = \mathbb{E}_n\left[\frac{x_0^{(N)}(x_0^{(N)}-1)\cdots(x_0^{(N)}-B_g^{(N)}+1)}{N(N-1)\cdots(N-B_g^{(N)}+1)}\right]$$

Hierbei bezieht sich der Erwartungswert $\mathbb{E}_{x_0^{(N)}}$ auf der linken Seite auf den Typenzählprozess $(X_g^{(N)})_{g\in\mathbb{N}_0}$ mit Start in $X_0^{(N)} = x_0^{(N)}$, der Erwartungswert $\mathbb{E}_n$ auf der rechten Seite auf den Blockzählprozess $(B_g^{(N)})_{g\in\mathbb{N}_0}$ mit Start in $B_0^{(N)} = n$.

(b) Mit $c_N = \mathbb{E}\big[\nu_1^{(N)}(\nu_1^{(N)}-1)\big]/(N-1)$, $d_N = \mathbb{E}\big[\nu_1^{(N)}(\nu_1^{(N)}-1)(\nu_1^{(N)}-2)\big]/((N-1)(N-2))$ gelte $c_N \to 0$, $d_N/c_N \to 0$ für $N \to \infty$ (so dass, wie wir wissen, der Ahnenprozess einer n-Stichprobe bei geeigneter Zeitskalierung gegen Kingmans n-Koaleszenten konvergiert).

Zeigen Sie: Falls $x_0^{(N)}/N \to z_0$, so gilt für $Z_t := X_{\lfloor t/c_N\rfloor}^{(N)}/N$, $t \geqslant 0$ und $n \in \mathbb{N}$

$$\lim_{N\to\infty} \mathbb{E}_{x_0^{(N)}}\left[Z_t^n\right] = \mathbb{E}_n\left[z_0^{B_t}\right]$$

wobei $(B_t)_{t\geqslant 0}$ der Blockzählprozess des Kingman-Koaleszenten ist.

(c)* Folgern Sie: Für $z \in [0,1]$ und $t \geqslant 0$ gibt es ein eindeutiges Wahrscheinlichkeitsmaß $\kappa_t(z,\cdot)$ auf $[0,1]$ mit der Eigenschaft

$$\int_{[0,1]} x^n \, \kappa_t(z,dx) = \mathbb{E}_n\left[z^{B_t}\right], \quad n \in \mathbb{N}$$

(d)* Die Familie der Wahrscheinlichkeitsmaße $\kappa_t(z,\cdot)$, $z \in [0,1]$ kann als ein stochastischer Kern gewählt werden (d.h. derart, dass $z \mapsto \kappa_t(z,A)$ messbar ist für $A \in \mathcal{B}([0,1])$). Dieses voraussetzend, können Sie den obigen Gedankengang soweit ausbauen, dass er zeigt, dass

$$\int_{[0,1]} \kappa_s(y,A)\, \kappa_t(z,dy) = \kappa_{t+s}(z,A), \quad t,s \geqslant 0,\ z \in [0,1],\ A \in \mathcal{B}([0,1])$$

d.h. die Kerne κ_t, $t \geqslant 0$ bilden eine Markov-Halbgruppe?

(d)** Die Kerne κ_t, $t \geqslant 0$ sind die Übergangskerne der Wright-Fischer-Diffusion $(X_t)_{t\geqslant 0}$, der Lösung der stochastischen Differentialgleichung

$$dX_t = \sqrt{X_t(1-X_t)}\, dB_t$$

(wobei $(B_t)_{t\geqslant 0}$ eine Standard-Brownbewegung ist), d.h. $\kappa_t(z,A) = \mathbb{P}(X_t \in A \,|\, X_0 = z)$.

Bericht. Der Gedankengang dieser Aufgabe entspricht einer stark vereinfachten Version der Resultate von Steven N. Evans, Coalescing Markov labelled partitions and a continuous sites genetics model with infinitely many types, *Ann. Inst. H. Poincaré Probab. Statist.* 33 (1997), 339–358.

Die Aufgabe kann zudem auch als Einladung aufgefasst werden, sich mit Kapitel 1.2.1 des Skripts Stochastische Modelle der Populationsbiologie aus WS 15/16 zu beschäftigen, siehe `https://www.staff.uni-mainz.de/birkner/SMPB1516/smpb1516.pdf`