Stochastische Populationsmodelle

Notizen zu einer Vorlesung an der Johannes-Gutenberg-Universität Mainz, Sommer 2024

Matthias Birkner

Vorläufige Version, 13. Mai 2024

Kommentare, Korrekturvorschläge, Hinweise auf (Tipp-)fehler gerne per Email an `birkner@mathematik.uni-mainz.de` senden

Inhaltsverzeichnis

Kapitel 1

Wright-Fisher-Modell und Kingman-Koaleszent

1.1 Wright-Fisher-Modell: Fundamentalmodell für genetische Drift

Das klassische *Wright-Fisher-Modell*[1,2] ist ein grundlegendes Modell der mathematischen Populationsgenetik zur Beschreibung des Phänomens der sogenannten Gendrift: genetische Typenhäufigkeiten in Populationen verändern sich im Lauf der Zeit aufgrund von Zufälligkeiten im Reproduktionserfolg, auch wenn keine systematischen Unterschiede der Typen oder äußere Einflüsse einwirken. Es ist ein (sehr) idealisiertes Populationsmodell, wir treffen folgende Annahmen:

- feste Populationsgröße: $N \in \mathbb{N}$ Individuen in jeder Generation
- die Population entwickelt sich in diskreten, nicht-überlappenden Generationen
- jedes Individuum hat nur ein „Elter“[3]
- es gibt Zufälligkeit bezüglich der Anzahl der Nachkommen, dabei aber keine systematischen Vorteile einzelner Individuen: die (gemeinsame) Kinderzahlverteilung ist „symmetrisch“
- es gibt verschiedene genetische Typen, die von Elter zu Kind vererbt werden

  Der Einfachheit halber nehmen wir hier (zunächst) an, dass es nur zwei verschiedene Typen, bezeichnet a und A gibt (zudem: keine „Kopierfehler“, sog. Mutationen[4], bei der Vererbung).

[1] Sewall Green Wright, 1889 – 1988, amerikanischer Genetiker

[2] Ronald Aylmer Fisher, 1890 – 1962, britischer Statistiker und Genetiker.

[3] Im Jargon der Genetik sind die Individuen „haploid“ – wörtlich angemessen z.B. für Bakterien, mitochondriale genetische Typen, Y-Chromosom. Viele Spezies sind „diploid“, besitzen also zwei Kopien jedes Chromosoms [ggfs. mit Ausnahme der Geschlechtschromosomen], manche Pflanzen sind „polyploid“. Asymptotisch, mit Ersetzung $N \rightsquigarrow 2N$, ist das Modell aber auch für Gene in diploiden Populationen passend.

[4] Eine sehr schöne Einführung in die Grundlagen der Genetik findet sich beispielsweise auf der Webseite DNA from the beginning `https://www.dnaftb.org/` des Cold Spring Harbor Laboratory.

Nehmen wir an, jedes Individuum einer gegebenen Generation hat (unabhängig) eine zufällige, Poisson-verteilte Anzahl Nachkommen mit Mittelwert 1, sei M_i = Anzahl Nachkommen von Individuum i, $1 \le i \le N$; angesichts der konstanten Gesamtpopulationsgröße müssen wir auf $\{M_1 + M_2 + \cdots + M_N = N\}$ bedingen. Für $m_1, \ldots, m_N \in \mathbb{N}_0$ (mit $m_1 + \cdots + m_N = N$) ist dann (beachte: die Faltungeigenschaft der Poissonverteilung liefert $M_1 + M_2 + \cdots + M_N =^d \mathrm{Pois}(N)$)

$$\mathbb{P}\Big(M_1 = m_1, \ldots, M_N = m_N \,\Big|\, \sum_{i=1}^N M_i = N\Big) = \frac{\prod_{i=1}^N e^{-1}\frac{1^{m_i}}{m_i!}}{e^{-N}\frac{N^N}{N!}} = \frac{N!}{m_1!m_2!\cdots m_N!}\Big(\frac{1}{N}\Big)^N$$

Somit: die gemeinsame Verteilung der Nachkommenszahlen der Individuen einer gegebenen Generation ist $\mathrm{Multinom}(N, \frac{1}{N}, \cdots, \frac{1}{N})$-verteilt. (Wir nehmen zudem Unabhängigkeit über die verschiedenen Generationen an.)

Alternative Interpretation: jedem Individuum in Generation r wird unabhängig ein uniform aus allen Individuen der Generation $r-1$ gewähltes Individuum als Elter zugeordnet.

Wir können daraus die Dynamik des Typenzählprozesses $(X_r^{(N)})_{r \ge r_0}$ ablesen: Sei

$$X_r^{(N)} = \text{Anzahl Typ } A\text{-Individuen in Generation } r$$

(demnach: $N - X_r^{(N)}$ Typ a-Individuen in Generation r).

Für $x, y \in \{0, 1, \ldots, N\}$ gilt somit

$$\mathbb{P}\big(X_{r+1}^{(N)} = y \,\big|\, X_r^{(N)} = x\big) = \binom{N}{y}\big(\frac{x}{N}\big)^y\big(1 - \frac{x}{N}\big)^{N-y},$$

d.h. gegeben $X_r^{(N)} = x$ ist $X_{r+1}^{(N)} \sim \mathrm{Bin}(N, x/N)$. Insbesondere gilt

$$\mathbb{E}\Big[X_{r+1}^{(N)} \,\Big|\, X_r^{(N)} = x\Big] = x, \quad \mathrm{Var}\Big[X_{r+1}^{(N)} \,\Big|\, X_r^{(N)} = x\Big] = N\frac{x}{N}\Big(1 - \frac{x}{N}\Big), \qquad x = 0, 1, \ldots, N \quad (1.1)$$

Abkürzend werden wir die folgende Notation verwenden: $\mathbb{P}_x$ für das W'maß in der Situation, dass $X_0^{(N)} = x$.

Sei $T_{\mathrm{fix}} := \inf\{r \in \mathbb{N}_0 : X_r^{(N)} = 0 \text{ oder } X_r^{(N)} = N\}$ Zeitpunkt, zu dem einer der beiden Typen verloren geht (angesichts $\min_{1 \le x \le N-1} \mathbb{P}_x\big(X_{r+1}^{(N)} \in \{0, N\}\big) > 0$ ist offenbar $\mathbb{P}_x(T_{\mathrm{fix}} < \infty) = 1$ für jedes $x = 0, 1, \ldots, N$).

Wie wahrscheinlich ist es, dass sich Typ A durchsetzt?

$$h(x) := \mathbb{P}_x(X_{T_{\mathrm{fix}}}^{(N)} = N), \quad x = 0, 1, \ldots, N$$

ist die eindeutige Lösung des Gleichungssystems

$$h(x) = \sum_{y=0}^N \binom{N}{y}\Big(\frac{x}{N}\Big)^y\Big(1 - \frac{x}{N}\Big)^{N-y} h(y), \quad x \in \{1, 2, \ldots, N-1\},$$
$$h(0) = 0, \qquad h(N) = 1$$

(Dies ist eine Instanz der allgemeinen Beziehung zwischen Auftreffwahrscheinlichkeiten von Markovketten und diskreten Dirichlet-Problemen, siehe beispielsweise [Geo15, Kap. 6.2], [Bir24, Kap. 7.1].)

Da der Erwartungswert einer $\mathrm{Bin}(N, x/N)$-verteilten Zufallsvariable gerade $N\frac{x}{N} = x$ ist, sieht man, dass der Ansatz $h(x) = x/N$ die eindeutige Lösung liefert, d.h.

$$\mathbb{P}_x(X^{(N)}_{T_{\text{fix}}} = N) = \frac{x}{N}$$

Die Fixationswahrscheinlichkeit entspricht also genau dem Startanteil.

Wie lange wird es typischerweise dauern, bis einer der beiden Typen verschwunden ist? Dazu betrachten wir (zunächst) die erwartete *Stichprobenheterozygotie*

$$\mathbb{E}_x\left[2\frac{X^{(N)}_r}{N}\Big(1-\frac{X^{(N)}_r}{N}\Big)\right]$$

Dies die Wahrscheinlichkeit, in einer zufälligen Stichprobe der Größe zwei (mit Zurücklegen) zwei unterschiedliche genetischen Typen vorzufinden.

Nach (1.1) ist für $x \in \{0, 1, \dots, N\}$

$$\begin{aligned}
&\mathbb{E}\left[2\frac{X^{(N)}_r}{N}\Big(1-\frac{X^{(N)}_r}{N}\Big)\,\middle|\, X^{(N)}_{r-1} = x\right] \\
&\qquad = \frac{2}{N}\mathbb{E}\left[X^{(N)}_r \,\middle|\, X^{(N)}_{r-1} = x\right] - \frac{2}{N^2}\mathbb{E}\left[\big(X^{(N)}_r\big)^2 \,\middle|\, X^{(N)}_{r-1} = x\right] \\
&\qquad = \frac{2x}{N} - \frac{2}{N^2}\left(\mathbb{E}\left[X^{(N)}_r \,\middle|\, X^{(N)}_{r-1} = x\right] + \left(\mathbb{E}\left[X^{(N)}_r \,\middle|\, X^{(N)}_{r-1} = x\right]\right)^2\right) \\
&\qquad = \frac{2x}{N} - \frac{2}{N^2}\left(x\Big(1-\frac{x}{N}\Big) + x^2\right) = \frac{2x}{N}\Big(1-\frac{x}{N}\Big) - \frac{2x}{N^2}\Big(1-\frac{x}{N}\Big) \\
&\qquad = \Big(1-\frac{1}{N}\Big)2\frac{x}{N}\Big(1-\frac{x}{N}\Big)
\end{aligned}$$

somit

$$\begin{aligned}
\mathbb{E}_x\left[2\frac{X^{(N)}_r}{N}\Big(1-\frac{X^{(N)}_r}{N}\Big)\right] &= \mathbb{E}_x\left[\mathbb{E}_x\left[2\frac{X^{(N)}_r}{N}\Big(1-\frac{X^{(N)}_r}{N}\Big)\,\middle|\, X^{(N)}_{r-1}\right]\right] \\
&= \Big(1-\frac{1}{N}\Big)\mathbb{E}_x\left[2\frac{X^{(N)}_{r-1}}{N}\Big(1-\frac{X^{(N)}_{-1}}{N}\Big)\right]
\end{aligned}$$

und iterativ

$$\mathbb{E}_x\left[2\frac{X^{(N)}_r}{N}\Big(1-\frac{X^{(N)}_r}{N}\Big)\right] = \Big(1-\frac{1}{N}\Big)^r 2\frac{x}{N}\Big(1-\frac{x}{N}\Big) \tag{1.2}$$

Wir sehen: um bei großem N eine substantielle Änderung über r Generationen zu sehen, muss $r \propto N$ sein, denn für $r = \lfloor tN \rfloor$ und $x = x^{(N)} = \lfloor yN \rfloor$ mit $t \in (0, \infty)$ und $y \in (0, 1)$ ergibt sich

$$\mathbb{E}_{\lfloor yN \rfloor}\left[2\frac{X^{(N)}_{\lfloor tN \rfloor}}{N}\Big(1-\frac{X^{(N)}_{\lfloor tN \rfloor}}{N}\Big)\right] \underset{N\to\infty}{\longrightarrow} e^{-t}2y(1-y) \tag{1.3}$$

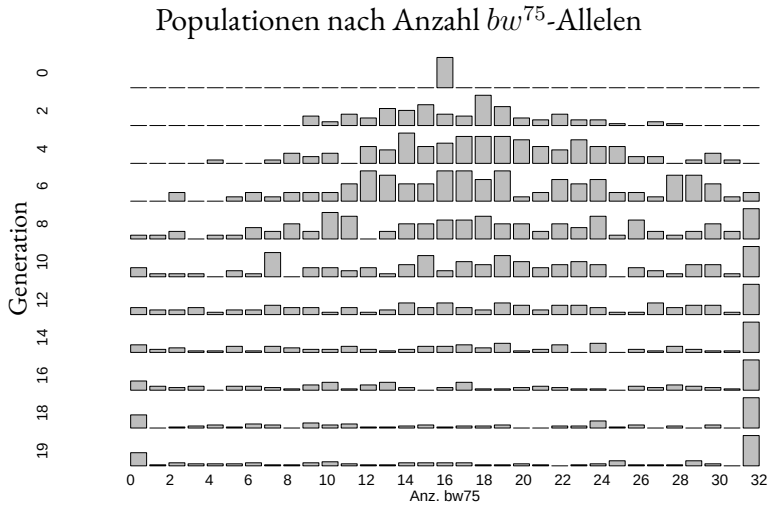


Abbildung 1.1: Beobachtungen aus 105 *Drosophila melanogaster*-Populationsexperimenten (aus P. Buri, *loc. cit.*, Table 14, S. 387). Für die Generationen $0, 2, 4, \ldots, 18, 19$ ist jeweils die empirische Verteilung der bw^{75}-Anteile über die 105 Populationen als Histogramm aufgetragen.

Das Buri-Beispiel

Peter Buri, Gene frequency in small populations of mutant Drosophila, *Evolution* 10, 367–402 (1956) berichtet ein Experiment in „künstlicher Evolution“:

- 105 Populationen von jeweils konstant[5] 16 Taufliegen (8 weibl., 8 männl.) wurden für 19 Generationen (1 Gen. ≈ 14d) unter konstanten Bedingungen gehalten.
- 2 Allele: bw und bw^{75}, die 3 Genotypen bw/bw, bw/bw^{75}, bw^{75}/bw^{75} sind anhand der Augenfarbe unterscheidbar
- Vorexperimente legten nahe, dass diese Genotypen keinen Einfluss auf den erwarteten Reproduktionserfolg haben.
- Die Anzahl bw^{75}-Chromosomen in jeder Population und Generation wurde beobachtet, s.a. Abb. 1.1.

In dieser Laborsituation kann man die Wirkung der Gendrift direkt beobachten, siehe auch Abbildung 1.1. Beispielsweise passt der beobachtete Abfall der (empirischen) Heterozygotie, gemittelt über die 105 Populationen, recht gut zum mittels (1.2) theoretisch vorhergesagten geometrischen Abfallen der erwartete Stichprobenheterozygotie, allerdings muss der „reale“ Populationsgrößenparameter $2N = 32$ durch die „effektive“ Populationsgröße $2N = 23$ ersetzt werden, siehe Abb. 1.2.

[5] Die konstante Populationsgröße wurde jeweils „von Hand“ beim Umsetzen der nächsten Generation in ein neues Glas erzwungen, zwischenzeitlich waren die Populationen natürlich angewachsen.

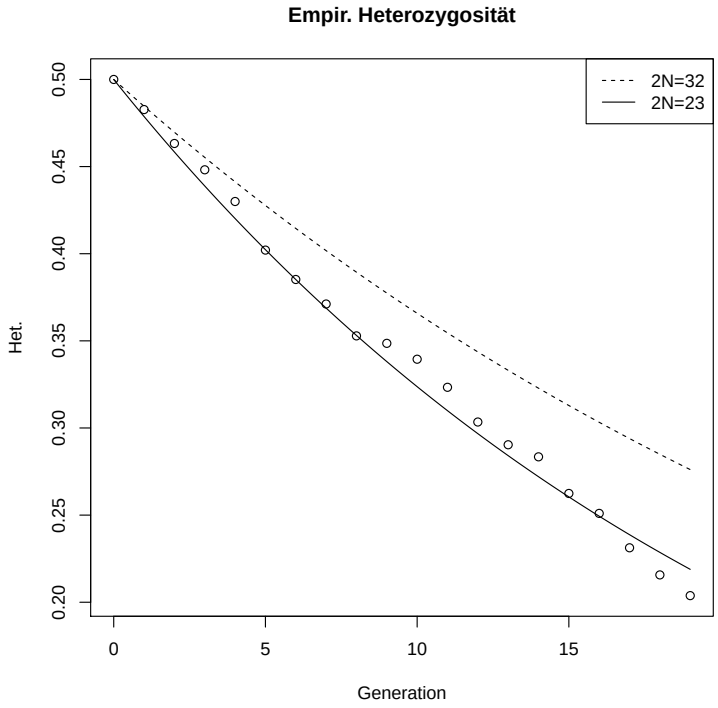


Abbildung 1.2: Beobachtete empirische Heterozygotie und nach (1.2) angepasste Kurven für die (theoretische) erwartete Stichprobenheterozygotie für zwei Parameterwahlen ($2N = 32$ und $2N = 23$)

Erwartete Zeit bis zur Fixierung Anhand der Rechnung rund um die erwartete Stichprobenheterozygotie haben wir in (1.3) bereits gesehen, dass für das Modell mit Populationsgröße N die Zeitskala N relevant ist.

Es gilt

$$\mathbb{E}_x[T_{\text{fix}}] = 1 + \sum_{y=0}^{N} \mathbb{P}_x(X_1 = y)\mathbb{E}_y[T_{\text{fix}}], \quad x = 1, 2, \dots, N-1 \tag{1.4}$$

mit Randwerten $\mathbb{E}_0[T_{\text{fix}}] = \mathbb{E}_N[T_{\text{fix}}] = 0$ (dies verwendet Zerlegung nach dem ersten Schritt und die Markoveigenschaft, siehe z.B. [Bir24, Kap. 7.1]). Allerdings ist das resultierende lineare Gleichungssystem in $N-1$ Unbekannten voll besetzt und nicht explizit lösbar.

Ein prinzipiell gangbarer Weg liegt in der Heuristik, dass $\mathbb{E}_x[T_{\text{fix}}]$ für großes N in „genügend glatter" Weise von $p := \frac{x}{N}$ abhängen sollte, und dann eine Taylorenwicklung von $\mathbb{E}_{pN}[T_{\text{fix}}]$ als Funktion von p anzusetzen.

Satz 1.1. *Sei $x_N \in \mathbb{N}$ mit $x_N/N \to p \in [0,1]$ für $N \to \infty$. Dann gilt*

$$\lim_{N\to\infty} c_N \frac{\mathbb{E}_{x_N}[T_{\text{fix}}^{(N)}]}{2N} = H(p) \tag{1.5}$$

mit $H(p) = -p\log(p) - (1-p)\log(1-p)$.

Wir betrachten hier nur eine Beweisheuristik, siehe Abschnitt A.4 für das volle Argument.

Beweisskizze. Nehmen wir an, es gibt eine genügend glatte Funktion $f_N : [0,1] \to \mathbb{R}_+$ mit

$$f_N\Big(\frac{x}{N}\Big) = \mathbb{E}_x[T_{\text{fix}}^{(N)}],$$

so gilt gemäß (1.4) mit Taylor-Entwicklung von f_N um $\frac{x}{N}$

$$\begin{aligned}
f_N\Big(\frac{x}{N}\Big) &= 1 + \sum_{y=0}^{N} \mathbb{P}_x(X_1 = y) f_N\Big(\frac{y}{N}\Big) \\
&= 1 + \sum_{y=0}^{N} \mathbb{P}_x(X_1 = y)\Big[f_N\Big(\frac{x}{N}\Big) + \Big(\tfrac{y-x}{N}\Big) f_N'\Big(\frac{x}{N}\Big) + \frac{1}{2}\Big(\tfrac{y-x}{N}\Big)^2 f_N''\Big(\frac{x}{N}\Big)\Big] + R(N,x) \\
&= 1 + f_N\Big(\frac{x}{N}\Big) + \frac{1}{N} f_N'\Big(\frac{x}{N}\Big)\mathbb{E}_x[X_1 - x] + \frac{1}{2}\frac{1}{N^2} f_N''\Big(\frac{x}{N}\Big)\mathrm{Var}_x[X_1 - x] + R(N,x) \\
&= 1 + f_N\Big(\frac{x}{N}\Big) + \frac{1}{2N}\frac{x(N-x)}{N^2} f_N''\Big(\frac{x}{N}\Big) + R(N,x)
\end{aligned}$$

mit Restterm

$$R(N,x) = \sum_{y=0}^{N} \mathbb{P}_x(X_1 = y)\frac{1}{6}\Big(\tfrac{y-x}{N}\Big)^3 f_N'''\Big(\zeta\big(\tfrac{x}{N}, \tfrac{y}{N}\big)\Big)$$

($\zeta(\frac{x}{N}, \frac{y}{N})$ ist eine Zahl zwischen $\frac{x}{N}$ und $\frac{y}{N}$).

Nun ist

$$\sum_{y=0}^{N} \mathbb{P}_x(X_1 = y)\Big(\frac{y-x}{N}\Big)^3 = \frac{1}{N^3}\mathbb{E}\Big[\big(Y_{N,x/N} - x\big)^3\Big]$$

mit $Y_{N,x/N} \sim \mathrm{Bin}(N, x/N)$ und es gilt

$$\max_{x=0,1,\dots,N} \mathbb{E}\Big[\big|Y_{N,x/N} - x\big|^3\Big] \le N^{3/2}$$

(siehe z.B. Lemma A.7 in Anhang A.4). Daher ist

$$R(N,x) = o(1/N)$$

zumindest plausibel.

Schreibe $\frac{x}{N} = p$, also erfüllt f_N näherungsweise

$$f_N''(p) = -2N\frac{1}{p(1-p)}, \quad 0 < p < 1 \tag{1.6}$$

mit den Randbedingungen $f_N(0) = f_N(1) = 0$. Man sieht nun leicht, dass eine explizite Lösung von (1.6) für $p \in (0,1)$ gegeben ist durch

$$f_N(p) = -2N\big(p\log(p) + (1-p)\log(1-p)\big),$$

denn

$$f_N'(p) = 2N(\log(p) + 1 - \log(1-p) - 1) = -\frac{2}{c_N}(\log(p) - \log(1-p)),$$

und

$$f_N''(p) = \big(-2N(\log(p) - \log(1-p))\big)' = -2N\Big(\frac{1}{p} + \frac{1}{1-p}\Big).$$

□

Der Satz zeigt insbesondere, dass für $X_0 = N/2$ die erwartete Zeit bis zur Absorption von entweder a oder A gegeben ist durch

$$\mathbb{E}_{N/2}[T_{\text{fix}}] \approx -2N\big(1/2\log(1/2) + 1/2\log(1/2)\big) = 2\log(2)\cdot N \approx 1{,}39\cdot N$$

Generationen.

1.2 Genealogien und Kingmans Koaleszent

Genealogischer Blickpunkt Das Wright-Fisher-Modell genealogisch ausgesprochen: Wir numerieren die Individuen jeder Generation $r \in \mathbb{Z}$ mit $i = 1, 2, \dots, N$ durch. Sei

$$A_{r,i}^{(N)} := \text{Nr. des Vorfahren (in Gen. } r-1\text{) von Ind. Nr. } i \text{ in Generation } r, \tag{1.7}$$

aus der Modellannahme: die $A_{r,i}^{(N)}$, $r \in \mathbb{Z}, i \in [N]$ sind u.i.v. uniform auf $[N] := \{1, 2, \dots, N\}$.

Ahnenverhältnisse Sei $A_{r,i}^{(N)}[k]$ die Nummer des Ahnen vor k Generationen von Individuum Nr. i in Generation r [dieser Ahne lebte in Generation $r-k$], aus (1.7) ist diese

$$\text{rekursiv bestimmt durch } A_{r,i}^{(N)}[1] = A_{r,i}^{(N)} \text{ und } A_{r,i}^{(N)}[k+1] = A^{(N)}_{r-k, A_{r,i}^{(N)}[k]} \text{ für } k \in \mathbb{N}.$$

Wir betrachte eine Stichprobe von n verschiedenen (zufällig gezogenen) Individuen aus Generation $r = 0$, sagen wir die Individuen Nr. $J_1, \dots, J_n$ mit $\mathbb{P}(J_1 = j_1, \dots, J_n = j_n) = 1/(N)_{n\downarrow}$ für paarweise verschiedene $j_1, \dots, j_n \in [N]$. (Wir notieren fallende Faktorielle als $(x)_{k\downarrow} := x(x-1)(x-2)\cdots(x-k+1)$ für $x \in \mathbb{R}$, $k \in \mathbb{N}$ mit Setzung $(x)_{0\downarrow} = 1$.)

Die Verwandtschaftsverhältnisse innerhalb der Stichprobe kodieren wir durch

$$R_k^{(N,n)}, \text{ eine (zufällige) Äquivalenzrelation,}$$

gegeben durch $i \sim_k j$ ($i, j \in [n]$, $k = 0, 1, \dots$), wenn $A_{0,J_i}^{(N)}[k] = A_{0,J_j}^{(N)}[k]$ gilt, d.h. Stichproben i und j haben denselben Ahnen vor k Generationen.

Sei

$$\mathcal{E}_n := \big\{\text{Äquivalenzrelationen auf } [n]\big\}$$

wir notieren $\xi \in \mathcal{E}_n$ etwa durch eine (ungeordnete) Liste der Äquivalenzklassen (z.B. $\xi = \{\{1\}, \{2,3\}\} \in \mathcal{E}_3$ bedeutet $2 \sim_\xi 3$, $1 \not\sim_\xi 2$, $1 \not\sim_\xi 3$).

Wir schreiben $\xi \preceq \eta$, falls

$$i \sim_\xi j \quad \Longrightarrow \quad i \sim_\eta j \qquad \text{gilt,}$$

d.h. η entsteht aus ξ durch Vereinigung einiger Klassen, ggfs. in mehreren Gruppen (beispielsweise ist $\{\{1\}, \{2\}, \{3,4\}, \{5\}, \{6,7\}, \{8\}\} \preceq \{\{1,2,6,7\}, \{3,4,5\}, \{8\}\}$).

Offensichtlich ist $i \sim_0 j \iff i = j$, d.h. $R_0^{(N,n)} = \big\{\{1\}, \{2\}, \dots, \{n\}\big\}$ und es gilt stets $R_k^{(N,n)} \preceq R_{k+1}^{(N,n)}$.

Betrachten wir zunächst den Fall $n = 2$: Für eine Stichprobe der Größe $n = 2$ ist die „korrekte“ Zeitskala der Genealogie [Vielfache von] N, denn die Paarverschmelzungsw'keit ist

$$p^{(N,2)}\big(\{\{1\}, \{2\}\}, \{1,2\}\big) = \frac{1}{N}$$

und somit die Zeit

$$\tau_1^{(N,2)} := \inf\big\{k \in \mathbb{N} : R_k^{(N,n)}\{\{1,2\}\}\big\}$$

bis die Stichprobe ihren ersten gemeinsamen Vorfahren findet (gemessen in Generationen), $\sim \mathrm{geom}(1/N)$, d.h. $\mathbb{P}\big(\tau_1^{(N,2)} > 0\big) = (1 - 1/N)^k$. . Somit für $t \in \mathbb{R}_+$

$$\mathbb{P}\Big(\frac{\tau_1^{(N,2)}}{N} > t\Big) = \Big(1 - \frac{1}{N}\Big)^{\lceil Nt \rceil} \underset{N\to\infty}{\longrightarrow} e^{-t}$$

d.h. für große N ist $\tau_1^{(N,2)}/N$ ungefähr $\mathrm{Exp}(1)$-verteilt. (Das passt auch zur Beobachtung aus (1.3), dass die „relevante“ Zeitskala N ist.)

Für $n = 3$ sieht die Übergangsmatrix $p^{(N,3)}(\cdot,\cdot)$ von $R^{(N,3)}$ folgendermaßen aus:

	$\{\{1\},\{2\},\{3\}\}$	$\{\{1,2\},\{3\}\}$	$\{\{1,3\},\{2\}\}$	$\{\{1\},\{2,3\}\}$	$\{\{1,2,3\}\}$
$\{\{1\},\{2\},\{3\}\}$	$1-3\frac{1}{N}+2\frac{1}{N^2}$	$\frac{1}{N}(1-\frac{1}{N})$	$\frac{1}{N}(1-\frac{1}{N})$	$\frac{1}{N}(1-\frac{1}{N})$	$\frac{1}{N^2}$
$\{\{1,2\},\{3\}\}$	0	$1-\frac{1}{N}$	0	0	$\frac{1}{N}$
$\{\{1,3\},\{2\}\}$	0	0	$1-\frac{1}{N}$	0	$\frac{1}{N}$
$\{\{1\},\{2,3\}\}$	0	0	0	$1-\frac{1}{N}$	$\frac{1}{N}$
$\{\{1,2,3\}\}$	0	0	0	0	1

d.h.

$$p^{(N,3)} = I + \frac{1}{N}\begin{pmatrix} -3 & 1 & 1 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} + \frac{1}{N^2}\begin{pmatrix} 2 & -1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} =: I + \frac{1}{N}Q + \frac{1}{N^2}R.$$

Lemma 1.2. *Für festes $N \geq n$ ist $(R_k^{(N,n)})_{k\in\mathbb{N}_0}$ eine Markovkette mit Werten in $\mathcal{E}_n$. Die Übergangswahrscheinlichkeiten sind gegeben durch*

$$p^{(N,n)}(\xi,\eta) \quad := \quad \mathbb{P}\big(R_{k+1}^{(N,n)} = \eta \,\big|\, R_k^{(N,n)} = \xi\big) = \frac{(N)_{a\downarrow}}{N^b}$$

sofern $\xi \preceq \eta$, wobei η aus $|\eta| = a$ Klassen besteht, ξ aus $b = |\xi| = b_1 + \cdots + b_a$ Klassen besteht und η aus ξ durch Verschmelzen von a Gruppen von Klassen in Gruppengrößen $b_1, \ldots, b_a$ entsteht (d.h. $\eta = \{C_1, \ldots, C_a\}$ und $\xi = \{C_{\alpha\beta} : 1 \leq \alpha \leq a, 1 \leq \beta \leq b_\alpha\}$ mit $C_\alpha = \cup_{\beta=1}^{b_\alpha} C_{\alpha\beta}$ für $\alpha = 1, \ldots, a$).

Beweis. $R_k^{(N,n)} = \xi$ bedeutet, dass es (k Generationen vor der Gegenwart) b verschiedene „aktive Ahnenlinien“ geben muss, d.h. es gibt b paarweise verschiedene Zahlen $i_{\alpha,\beta} \in [N]$, $\beta = 1, \ldots, b_\alpha$, $\alpha = 1, \ldots, a$, so dass

$$A_{0,J_i}^{(N)}[k] = i_{\alpha,\beta} \text{ für } i \in C_{\alpha\beta}, \quad \beta = 1, \ldots, b_\alpha, \ \alpha = 1, \ldots, a$$

gilt. Gegeben dies tritt das Ereignis $\{R_{k+1}^{(N,n)} = \eta\}$ genau dann ein, wenn es paarweise verschiedene $j_1, j_2, \ldots, j_a \in [N]$ gibt mit

$$A_{-k,i_{\alpha,\beta}}^{(N)} = j_\alpha \text{ für } \beta = 1, \ldots, b_\alpha, \ \alpha = 1, \ldots, a. \tag{1.8}$$

Nach Konstruktion des Wright-Fisher-Modells gilt für jede solche Wahl

$$\mathbb{P}\Big(A_{-k,i_{\alpha,\beta}}^{(N)} = j_\alpha \text{ für } \beta = 1, \ldots, b_\alpha, \ \alpha = 1, \ldots, a\Big) = \prod_{\alpha=1}^{a}\prod_{\beta=1}^{b_\alpha}\frac{1}{N} = \frac{1}{N^b}$$

und es gibt $N \cdot (N-1) \cdot (N-2) \cdots (N-a+1) = (N)_{a\downarrow}$ viele mögliche Wahlen. □

Beobachtung 1.3. Für $\xi, \eta \in \mathcal{E}_n$ mit $\xi \preceq \eta$, wobei $|\eta| = a$, $|\xi| = b = b_1 + b_2 + \cdots + b_a$ mit $b_1, \ldots, b_a \geq 1$ zeigt Lemma 1.2

1. $$p^{(N,n)}(\xi,\xi) = \frac{(N)_{b\downarrow}}{N^b} = \prod_{i=0}^{b-1}\Big(1 - \frac{i}{N}\Big) = 1 - \frac{1+2+\cdots+(b-1)}{N} + O\Big(\frac{1}{N^2}\Big)$$
$$= 1 - \frac{1}{N}\binom{b}{2} + O\Big(\frac{1}{N^2}\Big)$$

2. Falls $a = b - 1$, d.h. η entsteht aus ξ durch Verschmelzung genau eines Paars von Klassen,

$$p^{(N,n)}(\xi,\eta) = \frac{(N)_{(b-1)\downarrow}}{N^b} = \frac{1}{N}\prod_{i=0}^{b-2}\Big(1-\frac{i}{N}\Big) = \frac{1}{N} + O\Big(\frac{1}{N^2}\Big)$$

3. Falls $a \le b - 2$, d.h. mehr als zwei Klassen sind an Verschmelzung(en) beteiligt,

$$p^{(N,n)}(\xi,\eta) = O\Big(\frac{1}{N^2}\Big)$$

Dies legt nahe, den zeitreskalierten Prozess der Ahnenverhältnisse $(R^{(N,n)}_{\lfloor Nt\rfloor})_{t\ge 0}$ zu betrachten. Es stellt sich heraus, dass dieser gegen eine zeitkontinuierliche Markovkette konvergiert. Für dazu notwendige Techniken siehe den Exkurs in Abschnitt A.1.

Die allgemeine Struktur des Grenzwerts (für beliebige Stichprobengröße n) ist folgende:

Definition 1.4. Die zeitkontinuierliche Markovkette $(R^{(n)}_t)_{t\ge 0}$ auf $\mathcal{E}_n$ mit Sprungratenmatrix

$$q_{\xi\eta} = \begin{cases} 1 & \text{falls } \eta \text{ aus } \xi \text{ durch Verschmelzung von genau zwei Klassen entsteht,} \\ -\binom{|\xi|}{2} & \text{falls } \eta = \xi, \\ 0 & \text{sonst} \end{cases} \tag{1.9}$$

heißt Kingmans[6] (n-)Koaleszent.

Zumeist betrachten wir den Startzustand $R^{(n)}_0 = \big\{\{1\},\{2\},\dots,\{n\}\big\}$. Wir können den Pfad $(R^{(n)}_t)_{t\ge 0}$ als Baum interpretieren, dessen Blätter mit $1,\dots,n$ markiert sind:
Zu den Zeitpunkten $0 = \tau^{(n)}_n < \tau^{(n)}_{n-1} < \cdots < \tau^{(n)}_2 < \tau^{(n)}_1$, wo

$$\tau^{(n)}_k := \inf\{t \ge 0 : |R^{(n)}_t| \le k\}$$

verschmelzen jeweils zwei Zweige. Für Stichproben $i, j \in [n]$ können wir den genealogischen Abstand von i und j, $\inf\{t \ge 0 : i \sim_{R^{(n)}_t} j\}$ aus dem Baum ablesen.

`vgl. Bild an der Tafel`

Satz 1.5. *Es gilt*

$$(R^{(N,n)}_{\lfloor Nt\rfloor})_{t\ge 0} \longrightarrow \big(R^{(n)}_t\big)_{t\ge 0} \quad \textit{für } N \to \infty.$$

Wir beweisen die in Satz 1.5 formulierte Konvergenz im Sinne der endlich-dimensionalen Verteilungen; tatsächlich gilt auch Konvergenz in Verteilung auf dem Pfadraum $D([0,\infty),\mathcal{E}_n)$.

Beweis von Satz 1.5. Fixiere n. Gemäß Lemma A.2 müssen wir zeigen, dass für $\xi,\eta \in \mathcal{E}_n$ gilt

$$p^{(N,n)}(\xi,\eta) = \delta_{\xi,\eta} + \frac{1}{N}q_{\xi\eta} + o\Big(\frac{1}{N}\Big) \tag{1.10}$$

[da $|\mathcal{E}_n| < \infty$ ist dann der Fehler gleichmäßig klein, d.h. wir zeigen, dass $\lim_{N\to\infty}\max_{\xi,\eta\in\mathcal{E}_n} N\big|p^{(N,n)}(\xi,\eta) - \delta_{\xi,\eta} - (1/N)q_{\xi\eta}\big| = 0$ gilt].

Dies folgt aus Beobachtung 1.3 und der Form der Sprungraten aus (1.9). □

[6] J.F.C. Kingman, The coalescent, Stochastic Process. Appl. 13 (1982), no. 3, 235–248.

Beobachtung 1.6. 1. (Die Zeit bis zum jüngsten gemeinsamen Vorfahren) Aus der Struktur der Sprungratenmatrix (1.9) folgt

$$\tau_1^{(n)} = (\tau_{n-1}^{(n)} - \tau_n^{(n)}) + (\tau_{n-2}^{(n)} - \tau_{n-1}^{(n)}) + \cdots + (\tau_1^{(n)} - \tau_2^{(n)}) \stackrel{d}{=} S_n + S_{n-1} + \cdots + S_2, \qquad n \geq 2,$$

wobei die S_k unabhängige exponentialverteilte Zufallsvariablen mit Parameter $\binom{k}{2}$ sind, somit

$$\mathbb{E}[\tau_1^{(n)}] = \sum_{k=2}^{n} E[S_k] = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n}\Big(\frac{1}{k-1} - \frac{1}{k}\Big) = 2\Big(1 - \frac{1}{n}\Big)$$

und $1 = \mathbb{E}[\tau_1^{(2)}] \leq \mathbb{E}[\tau_1^{(n)}] < \lim_{m\to\infty} \mathbb{E}[\tau_1^{(m)}] = 2$.

Für ein Populationsmodell (aus der von uns betrachteten Schar) mit Populationsgröße N bedeutet dies, das der jüngste gemeinsame Vorfahre der heute lebenden Population im Mittel vor etwa $2N$ Generationen gelebt hat.

Weiter ist

$$\begin{aligned}\operatorname{Var}[\tau_1^{(n)}] &= \sum_{k=2}^{n} \operatorname{Var}[S_k] = \sum_{k=2}^{n} \binom{k}{2}^{-2} = \sum_{k=2}^{n} \frac{4}{k^2(k-1)^2} = \sum_{k=2}^{n} \Big\{4\Big(\frac{1}{k^2} + \frac{1}{(k-1)^2}\Big) + 8\Big(\frac{1}{k} - \frac{1}{k-1}\Big)\Big\} \\ &= \Big\{8\sum_{k=1}^{n-1}\frac{1}{k^2}\Big\} - 4 + \frac{4}{n^2} + \frac{8}{n} - 8 = \Big\{8\sum_{k=1}^{n-1}\frac{1}{k^2}\Big\} - 4\Big(1 - \frac{1}{n}\Big)\Big(3 + \frac{1}{n}\Big),\end{aligned}$$

insbesondere

$$1 = \operatorname{Var}[\tau_1^{(2)}] \leq \operatorname{Var}[\tau_1^{(n)}] < \lim_{n\to\infty} \operatorname{Var}[\tau_1^{(n)}] = 8\frac{\pi^2}{6} - 12 \approx 1.16.$$

Der wesentliche Beitrag zur Gesamtvarianz kommt also von der letzten Verschmelzungszeit S_2.

2. (Teilstichproben-Konsistenz) Sei $\pi_{n,n-1} : \mathcal{E}_n \to \mathcal{E}_{n-1}$ die Einschränkung aller Äquivalenzklassen auf $[n-1]$, so gilt

$$\big(\pi_{n,n-1}(R_t^{(n)})\big)_{t\geq 0} \stackrel{d}{=} \big(R_t^{(n-1)}\big)_{t\geq 0}.$$

Dies folgt aus der Form der Sprungraten oder auch aus der Tatsache, dass für die Approximanten (wie in Satz 1.5) nach Konstruktion $\pi_{n,n-1}(R_k^{(N,n)}) = R_k^{(N,n-1)}$ (realisierungsweise) gilt.

3. (Invarianz der Verteilung unter Permutation der Stichprobennummern) Für eine Permutation σ von $[n]$ und $\xi = \{C_1, \ldots, C_a\} \in \mathcal{E}_n$ sei $\sigma(\xi) = \{\sigma(C_1), \ldots, \sigma(C_a)\}$ die Äquivalenzrelation, die man erhält, indem man die Elemente der Blöcke von ξ gemäß σ umnummeriert. Es gilt

$$\big(\sigma(R_t^{(n)})\big)_{t\geq 0} \stackrel{d}{=} \big(R_t^{(n)}\big)_{t\geq 0}.$$

Dies folgt aus der Symmetrie der Sprungraten oder auch aus der Tatsache, dass für die Approximanten (wie in Satz 1.5) nach Konstruktion $\big(\sigma(R_k^{(N,n)})\big)_{k\in\mathbb{N}_0} \stackrel{d}{=} \big(R_k^{(N,n)}\big)_{k\in\mathbb{N}_0}$ gilt. [Man sagt auch, dass $R_t^{(n)}$ eine austauschbare zufällige Äquivalenzrelation ist.]

Bericht. Mit Beob. 1.6, 2. und Kolmogorovs Erweiterungssatz ist es möglich, den Kingman-Koaleszenten $(R_t)_{t\geq 0}$ mit Stichprobengröße $n = \infty$ als Markovprozess auf $\mathcal{E} := \{\text{Äquivalenzrelationen auf } \mathbb{N}\}$ mit

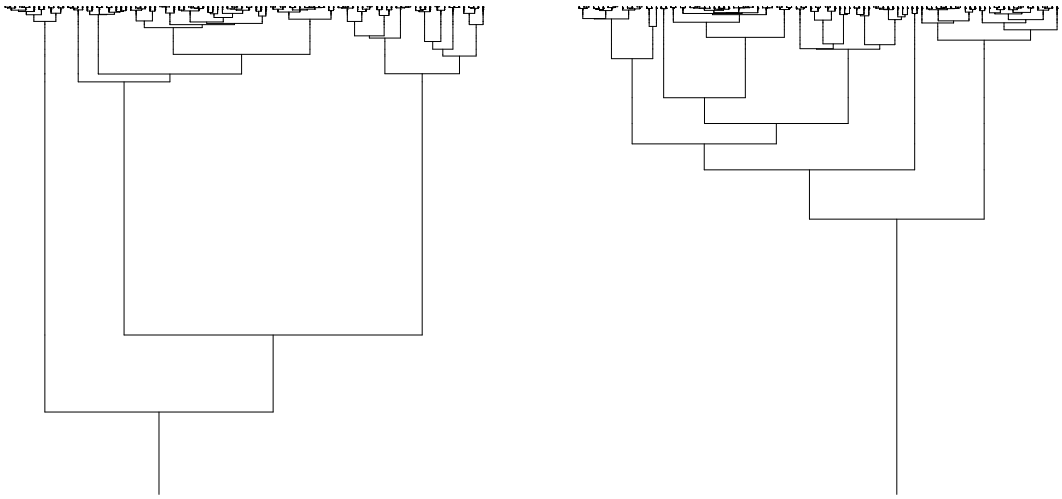


Abbildung 1.3: Zwei Realisierungen des Kingman-100-Koaleszenten (wobei die Blätter jeweils so sortiert wurden, dass der Baum überschneidungsfrei zu zeichnen ist)

Startwert $R_0 = \{\{1\}, \{2\}, \dots\}$ zu definieren mit der Eigenschaft $\big(\pi_{\infty,n}(R_t)\big)_{t\geq 0} \overset{d}{=} \big(\sigma(R_t^{(n)})\big)_{t\geq 0}$ für jedes $n \in \mathbb{N}$.

Siehe auch Konstruktion 1.9 unten für eine explizite Kopplung aller (Kingman-)n-Koaleszenten via "look down", die dies ohne (allzugroßen) Theorie-Aufwand leistet.

Beob. 1.6, 1. zeigt, dass $\mathbb{E}[\tau_1^{(\infty)}] = 2 < \infty$, d.h. auch eine „unendlich große Stichprobe" findet f.s. in endlicher Zeit ihren ersten gemeinsamen Vorfahren. Obwohl $|R_0| = \infty$ ist, gilt $|R_t| < \infty$ für jedes $t > 0$ fast sicher. Man sagt auch, dass der Kingman-Koaleszent „aus dem Unendlichen herabsteigt." Siehe auch die Simulationsbilder in Abbildung 1.3 für einen Eindruck dieses Phänomens.

1.3 Moran-Modell

Das Moran-Modell ist gewissermaßen das „zeitkontinuierliche Analogon" zum Wright-Fisher-Modell, es ist (ebenfalls) eines der fundamentalen Modelle der mathematischen Populationsgenetik.

Definition 1.7 ((Neutrales 2 Typ-)Moran-Modell[7])**.** Man betrachtet eine Population von konstant N (haploiden) Individuen, jedes Individuum besitzt eine unabhängige, $\mathrm{Exp}(1)$-verteilte Lebenszeit und wird am Ende seiner Lebenszeit durch den Nachkommen eines rein zufällig aus der Population gezogenen Individuums ersetzt (es gibt nur ein Elter und, sagen wir, man kann durch sein eigenes Kind ersetzt werden).

Wir nehmen zusätzlich an, dass es zwei Typen A und a gibt, die ohne Mutation vererbt werden. Sei

$$X_t^{(N)} = \text{Anzahl Typ } A\text{-Ind. zur Zeit } t.$$

[7] Nach Patrick Alfred Pierce Moran, 1917–1988 benannt

Angesichts der Gedächtnislosigkeit der Exponentialverteilung ist $(X_t^{(N)})_{t\geq 0}$ eine zeitkontinuierliche Markovkette mit Werten in $\{0, 1, \ldots, N\}$ und Sprungraten

$$q_{i,i+1} = i\frac{(N-i)}{N} = (N-i)\frac{i}{N} = q_{i,i-1}, \quad q_{i,i} = -2\frac{i(N-i)}{N}$$

(die übrigen Einträge der Sprungratenmatrix $Q = (q_{i,j})$ sind $= 0$).

Graphische Konstruktion

Für jedes geordnete Paar (i, j), $i, j \in \{1, \ldots, N\}$, $i \neq j$ sei $(N_t^{(i,j)})_{t\geq 0}$ ein Poissonprozess auf $\mathbb{R}_+$ mit Rate $\frac{1}{N}$, u.a. für verschiedene Paare. Zu den Sprungzeiten von $(N_t^{(i,j)})_{t\geq 0}$ stirbt Individuum j und wird durch einen Nachkommen von Individuum i ersetzt (s.a. Abb. 1.4).

`[Bild an der Tafel]`

Abbildung 1.4: Im Bild: N Kopien der Zeitachse, gerichtete Pfeile zwischen ihnen zu den Sprungzeitpunkten von u.a. Poissonprozessen; das Individuum an der Pfeilspitze stirbt jeweils und wird durch einen Nachkommen des Individuums am Pfeilschaft ersetzt.

Sei

$$X_t(i) = \text{Typ von Individuum } i \text{ zur Zeit } t.$$

Die Dynamik des Prozesses $\big(X_t(1), X_t(2), \ldots, X_t(N)\big)_{t\geq 0}$, der über die Typen der Individuen in der Population Buch führt (und nicht nur über die Anzahlen) ist somit folgende:

Ersetze zu jedem Sprungzeitpunkt t von $N^{(i,j)}$ den Typ $X_{t-}(j)$ durch $X_t(j) = X_{t(-)}(i)$.

Dies ist wohldefiniert, da unabhängige Poissonprozesse f.s. keine gemeinsamen Sprungzeitpunkte besitzen. Diese Konstruktion ist ein Spezialfall einer sogenannten Harris-Konstruktion[8], ein in der Theorie der interagierenden Teilchensysteme übliches (und nützliches) Werkzeug.

1.4 Dualität

Bemerkung 1.8 (Ablesen der Genealogie und der Typen aus der graphischen Konstruktion)**.** Für $t > 0$, $i \in [N]$ sei

$$A_s^{(i,t)} = \text{Nr. des Ahnenindividuums zur Zeit } t-s \text{ von Ind. } i \text{ zur Zeit } t \text{ (für } 0 \leq s \leq t, \text{ Werte in } [N])$$

Zur Konstruktion von $A^{(i,t)} = (A_s^{(i,t)})_{0\leq s\leq t}$ verfolgen wir die derzeitige „Zeitachse“ rückwärts und folgen den Pfeilen jeweils in entgegengesetzter Richtung, vgl. auch Abb. 1.4.

[8] nach Theodore Edward Harris, 1919–2005 benannt

In Formeln können wir den Pfad von $A^{(i,t)}$ beispielsweise folgendermaßen fassen (wir schreiben $N^{(j,i)}([a,b))$ für die Anzahl Sprünge des Poissonprozesses $N^{(j,i)}$ im Zeitintervall $[a,b)$):

Sei $T_0^{(i,t)} := 0$, $\widetilde{A}_0^{(i,t)} := A_0^{(i,t)} := i$, für $k \in \mathbb{N}$ setzen wir

$$T_k^{(i,t)} := \inf\left\{u > T_{k-1}^{(i,t)} : \text{es gibt ein } j \neq i \text{ mit } N^{(j,\widetilde{A}_{k-1}^{(i,t)})}\left([t-u, t-T_{k-1}^{(i,t)})\right) = 1\right\}$$

bzw. $T_k^{(i,t)} := t$, falls es kein solches u gibt. Falls $T_k^{(i,t)} = t$ gilt, so setzen wir $M^{(i,t)} := k$ und wir brechen die Konstruktion hier ab, andernfalls sei

$$\widetilde{A}_k^{(i,t)} \text{ das (f.s.) eindeutig bestimmte } j \text{ mit } N^{(j,\widetilde{A}_{k-1}^{(i,t)})}\left([t-T_k^{(i,t)}, t-T_{k-1}^{(i,t)})\right) = 1$$

und wir setzen die Konstruktion fort. Da die endlich vielen Poissonprozesse $N^{(j,i)}$ f.s. keine Häufungspunkte in $[0,t]$ besitzen, bricht die Konstruktion mit Wahrscheinlichkeit 1 nach endlich vielen Schritten ab und wir setzen dann für $0 < s \leq t$

$$A_s^{(i,t)} := \widetilde{A}_\ell^{(i,t)} \text{ falls } T_\ell^{(i,t)} \leq s < T_{\ell+1}^{(i,t)} \text{ für } 0 \leq \ell < M^{(i,t)}$$

bzw. $A_t^{(i,t)} := \widetilde{A}_{M^{(i,t)}-1}^{(i,t)}$.

$(A_s^{(i,t)})_{0\leq s\leq t}$ ist eine zeitkontinuierliche Markovkette mit (vollkommen symmetrischen) Sprungraten

$$q_{jk} = \begin{cases} \frac{1}{N}, & k \neq j, \\ -\frac{N-1}{N}, & k = j, \end{cases} \tag{1.11}$$

man nennt eine solche Kette auch eine (zeitkontinuierliche) „Irrfahrt auf dem vollständigen Graphen V_N der Ordnung N".

Für $i_1 \neq i_2$ bewegen sich $A^{(i_1,t)}$ und $A^{(i_2,t)}$ unabhängig bis zum „Verschmelzungszeitpunkt"

$$\tau_{i_1,i_2} := \inf\{s \in [0,t] : A_s^{(i_1,t)} = A_s^{(i_2,t)}\},$$

ab dann, d.h. für $u \geq \tau_{i_1,i_2}$, gilt $A_u^{(i_1,t)} = A_u^{(i_2,t)}$.

Für paarweise verschiedene $i_1, i_2, \ldots i_n$ ($\leq N$) bilden

$$A^{(i_1,t)}, \ldots, A^{(i_n,t)} \text{ ein System verschmelzender Irrfahrten auf } V_n$$

und mit

$$k \sim_{s,N} \ell \;:\Longleftrightarrow\; A_s^{(i_k,t)} = A_s^{(i_\ell,t)}, \quad 1 \leq k, \ell \leq n$$

ist

$$\mathcal{R}_s^{(n,N)} := \text{Äquivalenzklassen bezüglich } \sim_{s,N},\ s \in [0,t]$$

ein (zeittransformierter) Kingman-n-Koaleszent. ($(\mathcal{R}_{Ns/2}^{(n,N)})_{s\geq 0}$ wäre wörtlich ein Koaleszent, wenn wir die Zeitachsen in der graphischen Konstruktion „bis $-\infty$ fortsetzten.")

Aus der Konstruktion ergibt sich folgende (realisierungsweise Form) der „Dualität":

$$X_t(i) = X_0(A_t^{(i,t)}) \quad \text{für } 1 \leq i \leq N, t > 0. \tag{1.12}$$

Beweisskizze. Die Tatsache, dass $A^{(i,t)}$ eine zeitkontinuierliche Markovkette ist, folgt anschaulich gesehen aus der Unabhängigkeit der Zuwächse der „treibenden" Poissonprozesse $N^{(j,k)}$, die symmetrische Form der Sprungratenmatrix (1.11) stammt daher, dass alle Poissonprozesse dieselbe Rate $1/N$ haben. Wenn aktuell $A_s^{(i,t)} = j$, so gibt es für $0 < h \ll 1$ und jedes $j' \neq j$ mit Wahrscheinlichkeit $\approx h/N$ einen Sprung von $N^{(j',j)}$ im Zeitintervall $[t-s-h, t-s)$ und dann springt $A^{(i,t)}$ von j nach j'.

Etwas formaler: Sei

$$\mathcal{F}_u^t := \sigma\big(N^{(j,k)}([a,b)) : j \neq k, t-u \leq a < b \leq t\big)$$

die σ-Algebra, die die Informationen über alle Sprünge der $N^{(j,k)}$ zwischen $t-u$ und t enthält. Offenbar kann man $A_s^{(i,t)}$ für $s \leq u$ anhand der Pfade der $N^{(j,k)}$ zwischen $t-u$ und t rekonstruieren (d.h. $A_s^{(i,t)}$ ist $\mathcal{F}_u^t$-messbar für $s \leq u$) und für $s < t$, $j \neq j' \in [N]$ ist auf dem Ereignis $\{A_s^{(i,t)} = j\}$

$$\begin{aligned}
&\frac{1}{h}\mathbb{P}\big(A_{s+h}^{(i,t)} = j' \,|\, \mathcal{F}_s^t\big) \\
&= \frac{1}{h}\mathbb{P}\big(N^{(j',j)}([t-s-h,t-s)) = 1, N^{(j'',j)}([t-s-h,t-s)) = 0 \text{ für } j'' \neq j'\big) + \frac{1}{h}R_h \\
&= \frac{1}{h}\cdot e^{-h/N}\frac{h/N}{1!}\cdot\big(e^{-h/N}\big)^{N-2} + R_h = \frac{1}{N} + o(1)
\end{aligned}$$

für $h \downarrow 0$, wobei der Resterm

$$|R_h| \leq \mathbb{P}\Big(\sum_{k\neq\ell}^N N^{(k,\ell)}([t-s-h,t-s)) \geq 2\Big) = O(h^2)$$

erfüllt. □

Wir können die Formel (1.12) und den dazugehörigen Gedankengang verwenden, um (faktorielle) Momente des Typenanteilsprozesses im Moran-Modell zu berechnen: Betrachten wir ein Moran-Modell mit Populationsgröße N und Startanzahl $X_0^{(N)} = x_0^{(N)}$ von Typ A-Individuen. Für $t \geq 0$ und $n \in \mathbb{N}$ ist

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

die Wahrscheinlichkeit, bei n Zügen ohne Zurücklegen aus der Population zur Zeit t jedesmal Typ A zu ziehen. Andererseits seien $J_1, \ldots, J_n$ mit $\mathbb{P}(J_1 = j_1, \ldots, J_n = j_n) = 1/(N)_{n\downarrow}$ für paarweise verschiedene $j_1, \ldots, j_n \in [N]$ die Nummern der n zur Zeit t gezogenen Individuen. Obige Wahrscheinlichkeit ist

$$\begin{aligned}
&\mathbb{P}_{x_0^{(N)}}\big(X_t(J_1) = X_t(J_2) = \cdots = X_t(J_n) = A\big) \\
&= \mathbb{P}_{x_0^{(N)}}\big(X_0(A_t^{(J_1,t)}) = X_0(A_t^{(J_2,t)}) = \cdots = X_0(A_t^{(J_2,t)}) = A\big) \\
&= \mathbb{E}\Big[\prod_{i=1}^{\#\{A_t^{(J_1,t)},\ldots,A_t^{(J_n,t)}\}} \frac{x_0^{(N)}-i+1}{N-i+1}\Big]
\end{aligned}$$

wobei wir für die erste Gleichung (1.12) verwenden und für die zweite Gleichung beobachten, dass die Nummern $A_t^{(J_1,t)}, \dots, A_t^{(J_n,t)}$ der Ahnenindividuen der Stichprobe (es kann in dieser Liste Mehrfacheinträge geben) gerade $\#\{A_t^{(J_1,t)}, \dots, A_t^{(J_n,t)}\}$ Zügen ohne Zurücklegen aus $[N]$ entsprechen. Da $\#\{A_t^{(J_1,t)}, \dots, A_t^{(J_n,t)}\} =^d \#\mathcal{R}_t^{(n,N)}$ gilt, folgt die „Stichprobendualitätsformel“

$$\mathbb{E}_{x_0^{(N)}}\left[\frac{X_t^{(N)}(X_t^{(N)}-1)\cdots(X_t^{(N)}-n+1)}{N(N-1)\cdots(N-n+1)}\right] = \mathbb{E}\Big[\prod_{i=1}^{\#\mathcal{R}_t^{(n,N)}} \frac{x_0^{(N)}-i+1}{N-i+1}\Big] \tag{1.13}$$

Die Beobachtung, dass $\mathcal{R}_{Nt/2}^{(n,N)} =^d R_t^{(n)}$, der Kingman n-Koaleszent zur Zeit t, liefert damit für $z \in [0,1]$ und $n \in \mathbb{N}_0$

$$\lim_{N\to\infty} \mathbb{E}_{\lfloor Nz\rfloor}\Big[\big(X_{Nt/2}^{(N)}/N\big)^n\Big] = \mathbb{E}_n\big[z^{\#R_t}\big] \tag{1.14}$$

Dies ist zumindest ein Indiz, dass der reskalierte Typenanteilsprozess $\big(X_{Nt/2}^{(N)}/N\big)_{t\geq 0}$ ein Limesobjekt besitzt (die Wright-Fisher-Diffusion).

Bemerkung. Eine analoge Überlegung greift für das Wright-Fisher-Modell aus Abschnitt 1.1 unter Verwendung der Ahneninformationen aus Abschnitt 1.2. Man findet für das Wright-Fisher-Modell

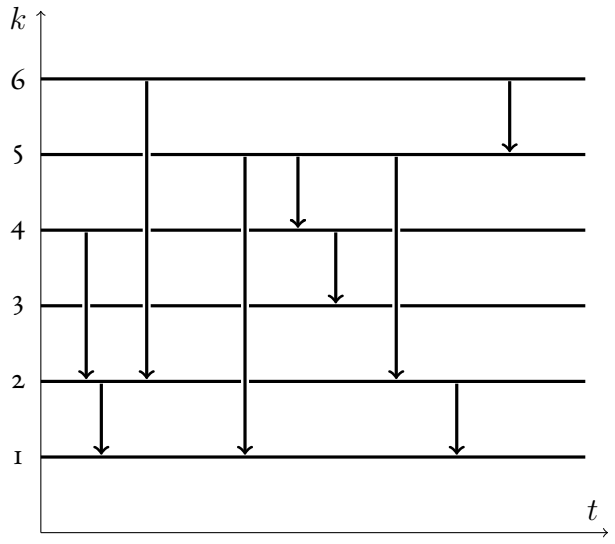
$$\lim_{N\to\infty} \mathbb{E}_{\lfloor Nz\rfloor}\Big[\big(X_{\lfloor Nt\rfloor}^{(N)}/N\big)^n\Big] = \mathbb{E}_n\big[z^{\#R_t}\big]$$

Der Limesprozess ist tatsächlich derselbe wie beim Moran-Modell.

Konstruktion 1.9 (eine explizite Kopplung aller (Kingman-)n-Koaleszenten via “look down”)**.** Für $1 \leq i < j$ seien $(L_{j,i}(t))_{t\geq 0}$ unabhängige Poissonprozesse auf $\mathbb{R}_+$ mit Rate 1. Für $k \in \mathbb{N}$ sei $(A_k(t))_{t\geq 0}$ gegeben als die Lösung von

$$A_k(t) = k - \sum_{1\leq i<j\leq k} \int_0^t (j-i)\mathbf{1}(A_k(s-) = j)\, L_{j,i}(ds), \quad t \geq 0$$

(wobei wir $L_{j,i}$ als die Verteilungsfunktion des – zufälligen – Zählmaßes auffassen, das jeweils an den Sprungstellen von $L_{j,i}$ Atome der Masse 1 besitzt). Zur Veranschaulichung des Systems der Lösungen A_k folgendes Bild: Für jedes $k = 1, 2, \dots$ betrachte eine Kopie der Zeitachse auf Niveau k, für $i < j$ zeichne zu den Sprungzeitpunkten des Prozesses $L_{j,i}$ einen Pfeil von Niveau j nach Niveau i.



Weiter sei $Q = (q_{\ell,m})_{\ell,m\in\mathbb{N}}$ mit

$$q_{\ell,m} = \begin{cases} 1, & 1 \le m < \ell \\ -(\ell-1), & m = \ell, \\ 0, & \text{sonst} \end{cases}$$

Dann gilt:

a) Für $k \in \mathbb{N}$ ist $A_k = (A_k(t))_{t\ge 0}$ Markovkette auf $[k] := \{1,\dots,k\}$, deren Sprungratenmatrix durch die Einschränkung von Q auf $[k]$ gegeben ist.

b) Die Prozesse A_k, $k \in \mathbb{N}$ bilden ein System verschmelzender Markovketten, d.h. für $k \neq \ell$ und $t \ge 0$ gilt

$$A_k(t) = A_\ell(t) \quad \Longrightarrow \quad A_k(t+s) = A_\ell(t+s) \text{ für all } s > 0$$

c) Für $n \in \mathbb{N}$ und $t \ge 0$ definieren wir eine Äquivalenzrelation $R_t^{(n)}$ of $[n]$ via

$$k \sim_{R_t^{(n)}} \ell \quad :\Leftrightarrow \quad A_k(t) = A_\ell(t)$$

$(R_t^{(n)})_{t\ge 0}$ ist verteilt wie Kingmans-n-Koaleszent.

d) $R_t^{(n)}$ entsteht aus $R_t^{(n+1)}$ durch Einschränkung der Klassen von $R_t^{(n+1)}$ auf $[n]$ (eine etwaige leere Klasse $\varnothing = \{n+1\} \cap [n]$ wird stillschweigend entfernt).

Insbesondere ist für $t \ge 0$ die zufällige Äquivalenzrelation R_t of $\mathbb{N}$ via

$$k \sim_{R_t} \ell \quad :\Leftrightarrow \quad k \sim_{R_t^{(n)}} \ell \text{ für } n \ge \max\{k,\ell\}$$

wohldefiniert. Der Prozess $(R_t)_{t\ge 0}$ beschreibt den Kingman-Koaleszenten, der mit unendlich vielen Blättern startet.

1.5 Beispiel: Die Beobachtungen von Dorit et al, 1995

Robert L. Dorit, Hiroshi Akashi und Walter Gilbert berichten in Absence of Polymorphism at the ZFY Locus on the Human Y Chromosome, *Science* 268, 1183–1185 (1995) die Ergebnisse einer genetischen Studie:

- Weltweite[9] Stichprobe von 38 Männern (*homo sapiens*)
- Ein 729 Basenpaare langes, nicht-kodierendes Stück des Y-Chromosoms (das 3. Intron des ZFY-Gens) wurde für jede Stichprobe sequenziert
- Es wurden keinerlei Mutationen gefunden: Alle 38 Stichproben identisch
- Inter-spezies-Vergleich mit Schimpanse, Gorilla, Orang-Utan (und Pavian als "outgroup") zeigt, dass am betrachteten Lokus Mutationen vorkommen können

[9] Loc. cit., S. 1184: "Human DNA samples were obtained from male volunteers who donated hair follicle samples or from cell lines provided by L. L. Cavalli-Sforza and K. K. Kidd. Geographic origins were determined by interview. Whenever possible, geographic origins of parents and grandparents were also ascertained. The samples are grouped by continent of origin, and the number of individuals is given in parentheses. Africa: Nigeria* (1), Ivory Coast (1), Tanzania (1), Southern Africa (2), Algeria (1), Central African Republic* (2), African American (2); Americas: Mexico (2), Guatemala (1), Peru* (1), Argentina (1), Native American (2); Asia: China* (2), Korea (1), Japan* (2), Taiwan (2), Indonesia (1), India (1); Europe/Middle East: Ireland* (1), Belgium (1), Italy* (1), Spain (1), Russia* (2), Poland* (1), Saudi Arabia* (1), Turkey (1); South Pacific: Melanesia (1), New Guinea* (1), Australia* (1). (*) Indicates samples where the 3'-most zinc-finger exon was also sequenced."

- Molekulare Uhr-Annahme und auf Fossilien beruhende Annahmen über die Zeit seit der Aufspaltung von der Vorfahren von Mensch und Schimpanse bzw. Orang-Utan ergeben geschätzte Rate von (fixierten) Mutationen

  $$1{,}35 \times 10^{-3} \text{ Mutationen pro Basenpaar pro Million Jahre}$$

Was können wir angesichts dieser Beobachtungen über die Zeit bis zum jüngsten gemeinsamen Vorfahren der gezogenen 38 Y-Chromosomen (und damit implizit auch über den jgV aller heute lebenden Männer) sagen?

Wir verwenden den Kingman-Koaleszenten als Modell der Genealogie.

A-priori-Verteilung Ohne Berücksichtigung der Beobachtungen würden wir annehmen, dass

$$T_{\text{jgV}} \overset{d}{=} S_{38} + S_{37} + \cdots + S_2$$

wo T_{jgV} die Zeit (in Koaleszenten-Zeiteinheiten) bis zum jüngsten gemeinsamen Vorfahren der 38 gezogenen Männer, die S_k unabhängig mit $S_k \sim \operatorname{Exp}\left(\binom{k}{2}\right)$,

$$1 \text{ Koaleszenten-Zeiteinheit } \mathrel{\widehat{=}}\ N_{\text{eff}} \times g \text{ Jahre}$$

mit N_{eff} … effektive Populationsgröße (für Männer), g … Generationslänge (in Jahren), also

a-priori-Verteilung: $\mathscr{L}(T_{\text{jgV}}) = \overset{38}{\underset{k=2}{\ast}} \operatorname{Exp}\left(\binom{k}{2}\right)$, d.h. mit Lemma A.6 ist die Dichte

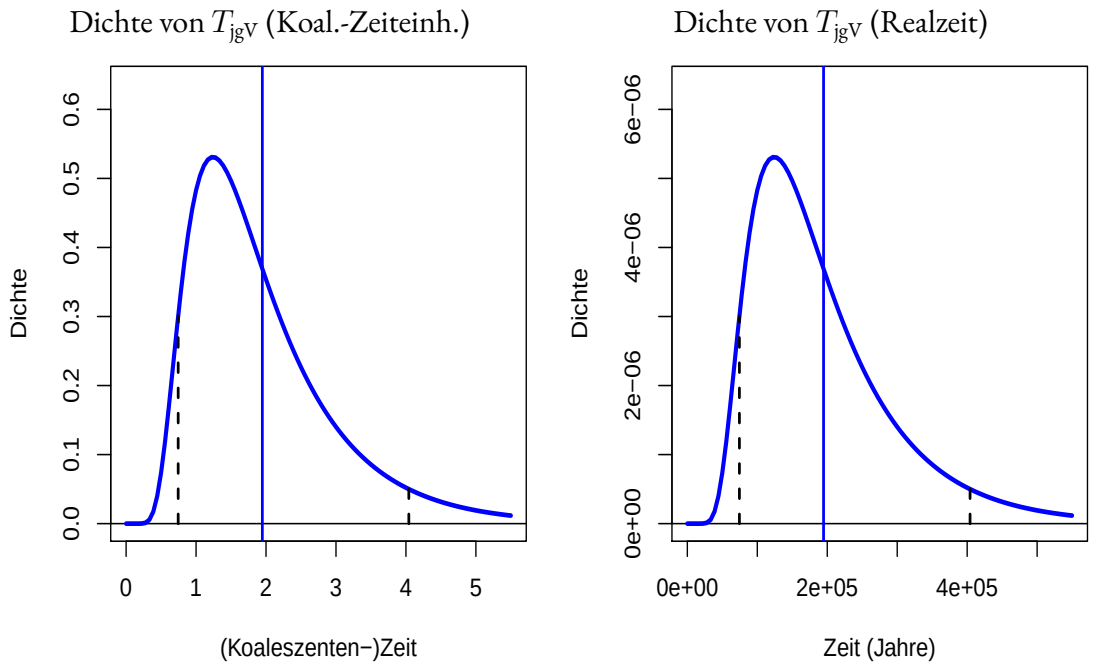
$$f_{\text{a-pri}} = \sum_{i=2}^{38} \binom{i}{2} \exp\left(-\binom{i}{2} t\right) \prod_{j=2, j\neq i}^{38} \frac{\binom{j}{2}}{\binom{j}{2} - \binom{i}{2}}$$

und

$$\mathbb{E}[T_{\text{jgV}}] = \sum_{k=2}^{38} \frac{2}{k(k-1)} = 2\left(1 - \tfrac{1}{38}\right) = \frac{37}{19} \doteq 1{,}947,$$

5%-Quantil von T_{jgV}: $q_{0{,}05} \doteq 0{,}744$, 95%-Quantil: $q_{0{,}95} \doteq 4{,}041$.

Mit Annahmen $N_{\text{eff}} = 5.000$, $g = 20$ Jahre übersetzt sich dies zu MW $\doteq 195.000$ Jahre, $q_{0{,}05} \doteq 74.000$ Jahre, $q_{0{,}95} \doteq 404.000$ Jahre.



A-posteriori-Verteilung Sei S_k die Länge des Zeitintervalls (in Koaleszenten-Zeiteinheiten) während dessen die Genealogie der Stichprobe aus k Linien bestand, M_k die Anzahl Mutationen, die während dieses Intervalls in der Genealogie auftreten.

Gegeben

$$S_k = t \text{ ist } M_k \text{ Poisson-verteilt mit Parameter } tk\frac{\theta}{2}, \text{ d.h.} \tag{1.15}$$

$$\mathbb{P}(M_k = m \,|\, S_k = t) = \exp\Big(-tk\frac{\theta}{2}\Big)\frac{\big(tk\frac{\theta}{2}\big)^m}{m!} \quad \text{wobei} \quad \theta = 2N_{\text{eff}} \times g \times \mu$$

mit N_{eff} effektive Populationsgröße, g Generationslänge (in Jahren), μ Mutationsrate der betrachteten Region im Genom (pro Jahr) (und gegeben $S_2, \ldots, S_n$ sind $M_2, \ldots, M_n$ unabhängig).

Eine heuristische Begründung für (1.15) ist folgende (wir werden dies im weiteren Verlauf der Vorlesung noch genauer betrachten): Angesichts Satz 1.5 entsprechen t Koaleszenten-Zeiteinheiten im Populationsmodell mit Populationsgröße N etwa t/c_N Generationen und somit etwa $tg/c_N = tgN_{\text{eff}}$ Jahren. Wenn wir annehmen, dass pro Jahr (unabhängig von allem anderen) eine Mutation mit der (sehr kleinen) Wahrscheinlichkeit μ auftritt, so ist die Verteilung der Anzahl Mutationen, die wir auf einem Stück der Genealogie dieser Länge sehen,

$$\text{Bin}(tgN_{\text{eff}}, \mu) \approx \text{Poi}(tgN_{\text{eff}}\mu) = \text{Poi}(t\theta/2).$$

Da gegeben $S_k = t$ der Teil der Genealogie, währenddessen k Linien existieren, aus k Stücken von je t Koaleszenten-Zeiteinheiten besteht, ist

$$\mathbb{P}(M_k = m \,|\, S_k = t) = \underbrace{\text{Poi}(t\theta/2) * \cdots * \text{Poi}(t\theta/2)}_{k \text{ mal}} = \text{Poi}(tk\theta/2).$$

(Die Normierung des Mutationsparameters als $\theta/2$ hat historische Gründe und sorgt auch dafür, dass manche Formeln „schöner“ aussehen.)

Frage: Wie ist $S_{38}+\cdots+S_2$ verteilt, gegeben dass $M_{38}+\cdots+M_2 = 0$? $S_k \sim \mathrm{Exp}\big(\binom{k}{2}\big)$, $\mathscr{L}(M_k \mid S_k = t) = \mathrm{Poi}(tk\theta/2)$, dann ist

$$\begin{aligned}\mathbb{P}(M_k = m) &= \int_0^\infty \binom{k}{2} \exp\Big(-\binom{k}{2} t\Big) e^{-tk\theta/2} \frac{(tk\theta/2)^m}{m!}\, dt \\ &= \binom{k}{2} \frac{(k\theta/2)^m}{m!} \int_0^\infty t^m \exp\Big(-\big(\binom{k}{2} + k\theta/2\big)t\Big)\, dt = \frac{k-1}{k-1+\theta}\Big(\frac{\theta}{k-1+\theta}\Big)^m,\end{aligned}$$

(wir substituieren $u = \big(\binom{k}{2} + k\theta/2\big)t$ und nutzen $\int_0^\infty u^m e^{-u}\, du = \Gamma(m+1) = m!$) d.h. $\mathscr{L}(M_k) = \mathrm{Geom}\big(\frac{k-1}{k-1+\theta}\big)$ — man könnte dies alternativ auch über ein „konkurrierende Raten“-Argument einsehen. Weiter ist damit

$$\mathbb{P}(S_k \le t \mid M_k = 0) = \frac{k-1+\theta}{k-1} \int_0^t \binom{k}{2} \exp\Big(-\binom{k}{2} s\Big) e^{-sk\theta/2}\, ds.$$

$$\mathscr{L}(S_k \mid M_k = 0) = \mathrm{Exp}\Big(\frac{k(k-1+\theta)}{2}\Big).$$

Bedingt auf $M_2 = \cdots = M_{38} = 0$ sind $S_2, \ldots, S_{38}$ (weiterhin) unabhängig.

Demnach: Verteilung der Zeit bis zum jüngsten gemeinsamen Vorfahren (in Koaleszenten-Zeiteinheiten), bedingt auf $M := M_2 + \cdots + M_{38} = 0$ ist

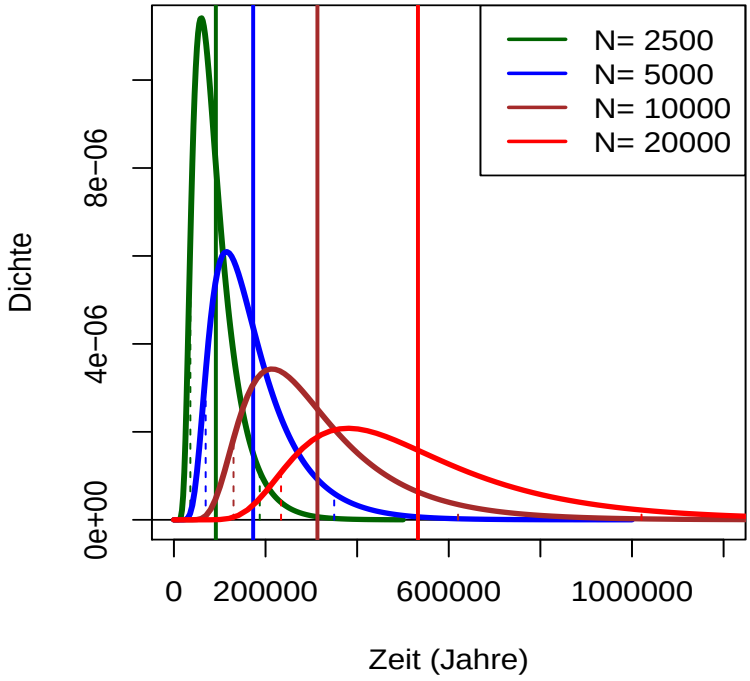
$$T_{\mathrm{jgV}}\big|_{\{M=0\}} \overset{d}{=} S'_{38} + S'_{37} + \cdots + S'_2$$

mit S'_k u.a., $S'_k \sim \mathrm{Exp}\big(\frac{k(k-1+\theta)}{2}\big)$, d.h. $\mathscr{L}(T_{\mathrm{jgV}} \mid M = 0) = \mathop{\ast}_{k=2}^{38} \mathrm{Exp}\big(\frac{k(k-1+\theta)}{2}\big)$. Die Dichte von $\mathscr{L}(T_{\mathrm{jgV}} \mid M = 0)$ (and damit auch den Erwartungswert und die Verteilungsfunktion) können wir wiederum mit Lemma A.6 bestimmen.

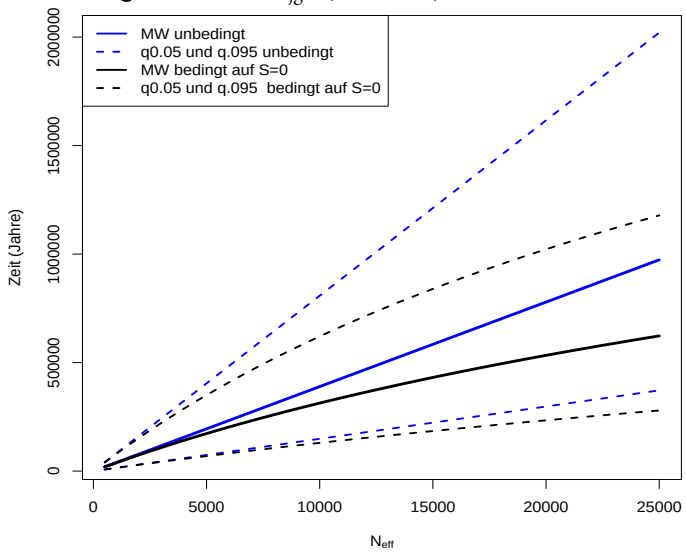
Wir fixieren $g = 20\mathrm{a}$, $\mu = 729 \times 1{,}35 \cdot 10^{-9}\mathrm{a}^{-1} \doteq 0{,}98 \cdot 10^{-6}\mathrm{a}^{-1}$ (aus Dorit et al (1995), diese Werte waren auch in der Literatur unstrittig), so hängt die bedingte Verteilung von T_{jgV} (und nicht nur ihre „Übersetzung in Realzeit“) vom Parameter N_{eff} ab.

N_{eff}	EW	$q_{0,05}$	$q_{0,95}$
2.500	91.519	35.851	187.369
5.000	173.007	69.263	349.909
10.000	313.234	130.095	620.279
20.000	532.785	233.853	1.020.819

Dichte von T_{jgV} (Realzeit)



MW und Quantile von T_{jgV} (Realzeit) als Funktion von N_{eff}



Wir sehen insbesondere: Die Verteilung von T_{jgV} hängt in nicht-linearer Weise von N_{eff} ab.

Literaturbericht Die ursprüngliche Studie erschien in Robert L. Dorit, Hiroshi Akashi und Walter Gilbert, Absence of Polymorphism at the ZFY Locus on the Human Y Chromosome, *Science* 268, 1183–1185 (1995), siehe auch die Diskussionsbeiträge ("technical comments") dazu in *Science* 272, 1356–1362 (1996) von Y.-X. Fu und W.-H. Li, von P. Donnelly, S. Tavaré, D.J. Balding und R.C. Griffiths, von G. Weiss und A. von Haeseler und von J. Rogers, P.B. Samollow und A.G. Comuzzie, die insbesondere eine etwas unpräzise Anwendung der Koaleszenten-Theorie von Dorit et al korrigierten. Die Darstellung fußt in weiten Teilen auf Wakeley [Wak09, Ch. 8.1]

Kapitel 2

Mutationen und der markierte Koaleszent

Wir betrachten in diesem Kapitel die Situation, dass die Individuen der Population verschiedene gentische Typen (abstrakt: aus einer Menge E möglicher Typen) haben und dass – im Gegensatz zur Situation in Kapitel 1 – Kinder mit einer gewissen (typischerweise kleinen) Wahrscheinlichkeit einen anderen Typ als ihr Elter haben (sogenannte „Mutationen“). Konkret stellen wir uns das Wright-Fisher-Modell aus Abschnitt 1.1 mit Populationsgröße N durch folgenden Mechanismus ergänzt vor: Jedes Kind ist unabhängig mit Wahrscheinlichkeit

$$\mu_N = \frac{\theta}{2N} \tag{2.1}$$

ein „mutiertes“ Kind, wobei $\theta \in (0, \infty)$. (Welche Änderung des Typs diese Mutation bewirkt, bleibt noch zu spezifizieren, wir betrachten in den folgenden Abschnitten konkrete Wahlen.)

Annahme (2.1) führt dazu, dass für $t \in (0, \infty)$ auf einer einzelnen Ahnenlinie über $\lfloor Nt \rfloor$ Generationen $\mathrm{Bin}(\lfloor Nt \rfloor, \theta/2N) \approx \mathrm{Pois}(t\theta/2)$ viele Mutationen auftreten. In der Skalierung der Genealogie einer n-Stichprobe wie in Satz 1.5 bedeutet dies: Das Limesobjekt ist ein Kingman-Koaleszent mit n Blättern, längs dessen Ästen Mutationen gemäß einem Poissonprozess mit Rate $\theta/2$ auftreten (vgl. auch die Diskussion in Abschnitt A.1).

Bemerkung. Die Annahme (2.1), die Populationsgröße und Mutationswahrscheinlichkeiten aneinander koppelt, mag auf den ersten Blick unnatürlich erscheinen: Warum sollte die Mutationswahrscheinlichkeit, die sich aus der Effektivität der biochemischen Kopier- und Reparaturmechanismen innerhalb der Zellen eines Individuums, ggfs. im Zusammenspiel mit diversen Umwelteinflüssen, ergibt, irgend etwas mit der Populationsgröße zu tun haben?

Annahme (2.1) bedeutet, dass Mutation und Gendrift „auf derselben Zeitskala“ wirken. Wenn $\mu_N \ll 1/N$, so wirkt die Gendrift (wie wir wissen, über Zeiten der Größenordnung $O(N)$), bevor Mutationen irgendeinen merklichen Einfluss auf die Zusammensetzung der Population haben, wenn $\mu_N \gg 1/N$, so stellt sich „reines Mutationsgleichgewicht“ ein, an dem Gendrift nichts ändert.

Anders gewendet: Für eine gegebene Population (mit endlichem, aber sehr großem N) ist

$$2N\mu_N = \theta$$

der „entscheidende“ Parameter, um die Zeitentwicklung des Typenanteils zu beschreiben.

2.1 Infinitely-many-alleles-Modell (IMA)

Definition 2.1 (Infinitely-many-alleles-Mutationsmechanismus)**.** Wir treffen die Modellannahme, dass jede Mutation einen völlig neuen Typ erzeugt (und die Mutationen sind „neutral“, d.h. sie beeinflussen den Fortpflanzungserfolg nicht). In der Literatur ist auch der Name “infinite alleles model” üblich.

Mathematisch realisieren wir dies durch die Wahl $E = [0,1]$ als Typenmenge, bei jedem Mutationsereignis wird (unabhängig) der neue Typ uniform($[0,1]$)-verteilt gewählt.

Wenn Mutationen sich mit Rate $\theta/2 > 0$ ereignen, ist die Vorwärtsentwicklung des Typs längs einer Abstammungslinie demnach beschrieben durch den Markov-Prozess mit Generator

$$Bf(x) = \frac{\theta}{2} \int_0^1 f(u) - f(x)\, du, \quad x \in [0,1]$$

für $f : [0,1] \to \mathbb{R}$ beschränkt und messbar.

Betrachte Stichprobe der Größe n: Beobachtete genetische Variabilität modelliert durch n-Koaleszent, längs dessen Kanten sich mit Rate $\frac{\theta}{2}$ Mutationen gem. IMA-Modell ereignen.

`[Bild an der Tafel]`

Offenbar ist nur der Teil der Genealogie jeweils „oberhalb“ der jüngsten Mutation relevant, für die Beobachtungen an den Blättern können wir also folgende äquivalente Dynamik betrachten:

Definition 2.2 („Getöteter n-Koaleszent“)**.**

- Beginne mit $\{\{1\},\{2\},\dots,\{n\}\}$ (alle aktiv).
- Jedes Paar von aktiven Klassen verschmilzt mit Rate 1.
- Jede Klasse wird mit Rate $\frac{\theta}{2}$ getötet/inaktiviert: allen Elementen wird derselbe, unif($[0,1]$)-verteilte Typ zugeordnet und die Klasse wird inaktiviert

  (sie hat ihre „definierende Mutation“ getroffen).
- Ende, wenn keine aktiven Klassen mehr übrig.

Analog zu Def. 1.4 (Kingman-Koaleszent) kann man dies formal als zeitkontinuierliche Markovkette auf

$$\widetilde{\mathcal{E}}_n = \big\{\text{Äquivalenzrelationen auf } [n], \text{ deren Klassen als aktiv/inaktiv markiert sind}\big\}$$

ausformulieren (Übung: man stelle die Sprungratenmatrix auf).

Bemerkung 2.3. Angesichts der Symmetrien des Koaleszenten ist die eigentlich relevante Information das *Typenhäufigkeitsspektrum* $(B_1,\dots,B_n)$, wobei

$$B_i = \#\text{Typen, die } i\text{-mal in der Stichprobe vorkommen}, \quad i = 1,\dots,n$$

(offenbar ist $\sum_{i=1}^n i\, B_i = n$).

Gegeben $(B_1, \dots, B_n) = (b_1, \dots, b_n)$ mit $\sum_{i=1}^n i b_i = n$ und

$$\sum_{i=1}^{n} b_i = k$$

sind die beobachteten Typen in der Stichprobe folgendermaßen verteilt: Zerlege $\{1, \dots, n\}$ uniform in k Teilmengen der Größen

$$\underbrace{1, \dots, 1}_{b_1}, \underbrace{2, \dots, 2}_{b_2}, \dots, \underbrace{n}_{b_n},$$

ordne jeder Teilmenge u.a. einen unif$([0,1])$-verteilten Typ zu.

Seien $E_n, \dots, E_1$ ZVn mit Werten in $\{\text{koal}, \text{mut}\}$,

E_k = Typ des Ereignisses, das die Anz. aktiver Klassen von k auf $k-1$ reduziert.

Es gilt

$$\mathbb{P}(E_k = \text{coal}) = \frac{\binom{k}{2}}{\binom{k}{2} + k\frac{\theta}{2}} = \frac{k-1}{k-1+\theta},$$
$$\mathbb{P}(E_k = \text{mut}) = \frac{k\frac{\theta}{2}}{\binom{k}{2} + k\frac{\theta}{2}} = \frac{\theta}{k-1+\theta}$$

und $E_n, \dots, E_1$ sind unabängig (verwende „konkurrierende-Raten“-Argument und Symmetrien der Sprungraten). Also gilt für $e_n, e_{n-1}, \dots, e_1 \in \{\text{coal}, \text{mut}\}$

$$\mathbb{P}(E_n = e_n, \dots, E_1 = e_1) = \frac{\prod_{k=1}^{n} \left(\theta \mathbf{1}(e_k = \texttt{mut}) + (k-1)\mathbf{1}(e_k = \texttt{coal})\right)}{\prod_{k=1}^{n}(k-1+\theta)}. \tag{2.2}$$

Definition 2.4 (Hoppe[1]-Urne)**.** Urne enthält eine schwarze Kugel (“Mutationskugel”) mit Masse θ, farbige Kugeln jew. mit Masse 1.

- Zu Beginn: Urne enthält nur die schwarze Kugel .
- In jedem Schritt: Ziehe eine Kugel mit W’keit proportional zu ihrer Masse.
- Falls farbige Kugel gezogen: Lege zurück zusammen mit einer weiteren Kugel derselben Farbe.
- Falls schwarze Kugel gezogen: Lege zurück zusammen mit einer weiteren Kugel einer völlig neuen Farbe.

Lemma 2.5. *Die von den n nicht-schwarzen Kugeln erzeugte Verteilung der Familiengrößen (Typenhäufigkeitsspektrum) nach n Zügen der Hoppe-Urne entspricht der des getöteten n-Koaleszenten (aus Def. 2.2).*

[1] Fred M. Hoppe, Pólya-like urns and the Ewens’ sampling formula, *J. Math. Biol.* 20, no. 1, 91–94, (1984).

Beweisskizze. Lese (2.2) „rückwärts." □

Sei

$$K_n = \#\text{verschiedene Typen in } n\text{-Stichprobe.}$$

Satz 2.6. *Im IMA-Modell ($\theta > 0$ fest) gilt für $n \to \infty$*

$$\mathbb{E}[K_n] = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1} \sim \theta \log n, \quad \operatorname{Var}[K_n] = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1} \cdot \frac{i-1}{\theta + i - 1} \sim \theta \log n,$$

$$\textit{und} \quad \frac{K_n - \mathbb{E}[K_n]}{\sqrt{\operatorname{Var}(K_n)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Beweis. Verwende Hoppe-Urne, schreibe

$$K_n = A_1 + \cdots + A_n,$$

mit

$$A_i = \mathbf{1}(\text{im } i\text{-ten Zug wurde die schwarze Kugel gezogen}).$$

Nach Konstruktion sind $A_1, \ldots, A_n$ u.a. mit

$$\mathbb{P}(A_i = 1) = \theta/(\theta + i - 1).$$

Für die Asymptotik vergleiche mit Riemann-Integral, für asymptotische Normalität bilde ein (unabängiges, zentriertes, normiertes) Dreiecksschema

$$X_{ni} = \frac{A_i - \frac{\theta}{\theta+i-1}}{\sqrt{\operatorname{Var}[K_n]}},$$

dies erfüllt (trivialerweise) die Lindeberg-Bedingung: Es gilt für

$$S_n = X_{n1} + \cdots + X_{nn}$$

$\operatorname{Var}[S_n] = 1$, für jedes $\varepsilon > 0$ gilt wegen $\operatorname{Var}[K_n] \to \infty$, dass $\mathbb{E}[X_{ni}^2 \mathbf{1}(X_{ni}^2 > \varepsilon)] = 0$ für n genügend groß, insbesondere

$$L_n(\varepsilon) := \sum_{i=1}^{n} \mathbb{E}[X_{ni}^2 \mathbf{1}(X_{ni}^2 > \varepsilon)] \to 0.$$

□

Bemerkung 2.7 (Hoppes Urne und zufällige Permutationen[2])**.** n Züge aus der Hoppe-Urne generieren sukzessive eine zufällige Permutation Π_n von $\{1, \ldots, n\}$ (in Zyklen-Darstellung) folgendermaßen:
Nummeriere die farbigen Kugeln in der Reihenfolge des Erscheinens, wenn im k-ten Zug

[2] Eine Beobachtung aus Paul Joyce, Simon Tavaré, Cycles, permutations and the structure of the Yule process with immigration, *Stochastic Process. Appl.* 25, no. 2, 309–314, (1987).

- schwarze Kugel gezogen : füge neuen Zyklus (k) hinzu,
  (insbesondere: nach dem ersten Zug entsteht die Identität (1))
- farbige Kugel j_k gezogen : füge im Zyklus, der j_k enthält, links von j_k ein.

Für jede Permutation π mit k Zyklen gilt dann

$$\mathbb{P}(\Pi_n = \pi) = \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)},$$

denn π legt die Reihenfolge der Ereignisse i.d. Urne fest, wenn im k-ten Zug schwarze Kugel gezogen: Faktor $\frac{\theta}{\theta+k-1}$, wenn farbige gezogen: Faktor $\frac{1}{\theta+k-1}$.

Man kann dies als eine Version des sogenannten China-Restaurant-Prozesses auffassen, siehe z.B. [Kle20], Kap. 24.3 (und speziell S. 523f dort).

Satz 2.8 (Ewens'sche Stichprobenformel[3])**.** *Seien $b_1, \ldots, b_n \in \mathbb{N}_0$ mit*

$$\sum_{j=1}^{n} b_j = k \leq n \quad \textit{und} \quad \sum_{j=1}^{n} j b_j = n$$

gegeben. Die Wahrscheinlichkeit, in einer n-Stichprobe (im IMA-Modell) jeweils b_j Typen mit genau j Repräsentanten (für $j = 1, \ldots, n$) zu beobachten, ist

$$\frac{n!}{1^{b_1} 2^{b_2} \cdots n^{b_n}} \cdot \frac{1}{b_1! b_2! \cdots b_n!} \cdot \frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)}. \tag{2.3}$$

Man kann (2.3) auch schreiben als

$$C(n,\theta) \times \prod_{j=1}^{n} e^{-\theta/j} \frac{(\theta/j)^{b_j}}{b_j!} \tag{2.4}$$

mit

$$C(n,\theta) = n! \exp\left(\theta \sum_{j=1}^{n} 1/j\right) / (\theta(\theta+1)\cdots(\theta+n-1)),$$

d.h. die Verteilung des Typenhäufigkeitsspektrums $(B_1, \ldots, B_n)$ in einer n-Stichprobe ist $\otimes_{j=1}^{n} \mathrm{Poi}(\theta/j)$, bedingt auf $\sum_{j=1}^{n} j B_j = n$.

Beweis. Man kann dies per Induktion beweisen, indem man nach dem „jüngsten" Ereignis im markierten Koaleszenten zerlegt. Wir betrachten hier ein direktes, kombinatorisches Argument aus dem Artikel Robert C. Griffiths, Sabin Lessard, Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles, *Theoretical Population Biology* 68, no. 3, 167–177, (2005).

Seien $E_n, E_{n-1}, \ldots E_1$ die „Elementarübergänge" in der Historie des getöteten Koaleszenten aus Def. 2.2 (vgl. (2.2) und Diskussion vor der Hoppe-Urne, Def. 2.4),

E_m beschreibt das Ereignis, das die Anzahl aktiver Linien von m auf $m-1$ reduziert.

[3] Warren J. Ewens, The sampling theory of selectively neutral alleles, *Theoretical Population Biology* 3, 87–112, (1972) und S. Karlin, J. McGregor, Addendum to a paper of W. Ewens, *Theoretical Population Biology* 3, 113–116, (1972).

Wir nummerieren die Linien mit $1, \dots, n$ und führen (genauer) Buch, welche Linie von einer Mutation getroffen wird bzw. welche Linie mit welcher Linie verschmilzt, mögliche Werte der Elementarübergänge sind also

$$\texttt{mut}(i), \text{ Linie } i \text{ trifft eine Mutation, und}$$

$$\texttt{koal}(i \to j), \text{ Linie } i \text{ verschmilzt in Linie } j\ (\neq i).$$

Wir denken bei den Verschmelzungereignissen an „gerichtete" Verschmelzungen, d.h. für jedes aktuell noch aktive Paar von Linien i und j

$$\text{verschmilzt Linie } i \text{ in Linie } j \text{ mit Rate } \frac{1}{2}.$$

Eine Liste $e_n, e_{n-1}, \dots, e_1$ solcher möglicher Elementarübergänge nennen wir ein *Feinprotokoll*.

Sei für $m \geq 2$

$$p_m(e_m) = \begin{cases} \dfrac{1/2}{\frac{1}{2}m(m-1) + \frac{\theta}{2}m} = \dfrac{1}{m(m-1+\theta)}, & \text{wenn } e_m \text{ eine Verschmelzung,} \\ \dfrac{\theta/2}{\frac{1}{2}m(m-1) + \frac{\theta}{2}m} = \dfrac{\theta}{m(m-1+\theta)}, & \text{wenn } e_m \text{ eine Mutation,} \end{cases}$$

$p_1(e_1) = 1$, wenn e_1 eine Mutation ist, und $p_1(e_1) = 0$ für eine (dann unmögliche) Verschmelzung.

Für ein gegebenes mögliches Feinprotokoll $e_n, e_{n-1}, \dots, e_2, e_1$ mit k Mutationsereignissen (und somit k Typen) ist

$$\mathbb{P}(E_n = e_n, \dots, E_1 = e_1) = p_n(e_n)p_{n-1}(e_{n-1})\cdots p_1(e_1) = \frac{\theta^k}{\prod_{m=1}^n m(m-1+\theta)} = \frac{\theta^k}{n!(\theta)_{n\uparrow}} \quad (2.5)$$

(Produkt der Übergangswahrscheinlichkeiten der Skelettkette des getöteten Kingman-n-Koaleszenten).

Für $b_1, \dots, b_n \in \mathbb{N}_0$ mit $\sum_{j=1}^n b_j = k$ (und $\sum_{j=1}^n j b_j = n$) gibt es

$$\frac{(n!)^2}{\prod_{j=1}^n (b_j! j^{b_j})} \quad (2.6)$$

mögliche Feinprotokolle, die auf dieses Typenhäufigkeitsspektrum $(b_1, \dots, b_n)$ führen. Das Produkt von (2.5) und (2.6) liefert (2.3).

Zu (2.6): Stellen wir uns für den Moment die k Typen („künstlich") nummeriert vor, mit

$$\text{Typenhäufigkeitsvektor } (n_1, n_2, \dots, n_k),$$

d.h. n_ℓ Stichproben sind vom ℓ-ten Typ, und es gilt

$$\big|\{\ell : n_\ell = j\}\big| = b_j, \quad j = 1, 2, \dots, n.$$

Es gibt

$$n! \times \binom{n}{n_1 \dots n_k} \times (n_1 - 1)! \cdots (n_k - 1)! = \frac{(n!)^2}{\prod_{\ell=1}^k n_\ell} = \frac{(n!)^2}{\prod_{j=1}^n j^{b_j}} \quad (2.7)$$

verschiedene Feinprotokolle mit k nummerierten Typen, die auf diesen Typenhäufigkeitsvektor $(n_1, n_2, \dots, n_k)$ führen:

1. $n!$ Möglichkeiten für die Reihenfolge, in der die Linien inaktiv werden,
2. $\binom{n}{n_1 \dots n_k}$ Möglichkeiten, die n Linien auf die k Typen aufzuteilen
   (n nummerierte Kugeln in k Schachteln legen),
3. für $\ell = 1, \dots, k$ gibt es $(n_\ell - 1)!$ viele Arten, innerhalb von Typ ℓ die „Verschmelzungsziele“ festzulegen.

   (Nehmen wir an, wir haben in Schritt 1 und 2 festgelegt, dass Linien $i_1, i_2, \dots, i_{n_\ell}$ vom Typ ℓ sind und dass diese in der Reihenfolge $i_1, i_2, \dots$ inaktiviert werden. Dann gibt es $n_\ell - 1$ viele Wahlen für das Verschmelzungsziel von i_1, $n_\ell - 2$ viele Wahlen für das Verschmelzungsziel von i_2, u.s.w.)

Schließlich führen

$$b_1! \cdot b_2! \cdot \dots \cdot b_n! \tag{2.8}$$

verschiedene Feinprotokolle mit nummerierten Typen auf dasselbe Feinprotokoll ohne Typennummerierung:

$$\text{Typen } \ell \text{ und } \ell' \text{ können vertauscht werden, sofern } n_\ell = n_{\ell'}. \tag{2.9}$$

Der Quotient von (2.7) und (2.8) liefert (2.6). □

Bemerkung 2.9. Nach Satz 2.6 ist

$$\widehat{\theta}_{\text{naiv}} := \frac{K_n}{\log n}$$

ein (schwach) konsistenter Schätzer für die Mutationsrate θ, d.h. für jedes $\theta \in (0, \infty)$ gilt

$$\widehat{\theta}_{\text{naiv}} \underset{n\to\infty}{\to} \theta \quad \text{stochastisch bzw. in Verteilung.}$$

Satz 2.6 liefert auch asymptotische Normalität von $\widehat{\theta}_{\text{naiv}}$. Allerdings ist

$$\operatorname{Var}_\theta[\widehat{\theta}_{\text{naiv}}] \sim \frac{\theta}{\log n}.$$

(Dies ist allerdings „deprimierend langsam“: z.B. müsste $n = e^{100} \approx 2.7 \cdot 10^{43}$ sein, damit die Streuung $\approx 0.1\sqrt{\theta}$ ist.)

Beobachtung 2.10. Im IMA-Modell ist K_n suffizient für θ, d.h. gegeben $K_n = k$ hängt die Verteilung der beobachteten Typen nicht von θ ab.

Beweis. Sei

$$C_{n,k} := \sum\nolimits'_{(b_1,\dots,b_n)} \frac{n!}{\prod_{i=1}^n i^{b_i}\, b_i!},$$

($\sum'$ bezeichnet die Summe über $(b_1, \dots, b_n) \in \mathbb{N}_0^n$ mit $\sum b_i = k$, $\sum i b_i = n$).

Satz 2.8 (Ewens-Formel) liefert

$$\mathbb{P}_\theta(K_n = k) = C_{k,n}\frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)},$$

für $b_1, \dots, b_n$ mit $b_1 + \cdots + b_n = k$ (und $\sum i b_i = n$) also

$$\mathbb{P}_\theta(B_1 = b_1, \dots, B_n = b_n | K_n = k) = \frac{1}{C_{k,n}}\frac{n!}{1^{b_1}2^{b_2}\cdots n^{b_n}} \cdot \frac{1}{b_1!b_2!\cdots b_n!}.$$

□

Sei $K_n = k$ beobachtet. Der Maximum-Likelihood-Schätzer $\widehat{\theta}_{\mathrm{ML}}$ ist dasjenige $\theta \geq 0$, das die Likelihood

$$L_n(\theta, k) = \mathbb{P}_\theta(K_n = k) = C_{k,n}\frac{\theta^k}{\theta(\theta+1)\cdots(\theta+n-1)}$$

(als Funktion von θ) maximiert.

Es ist

$$\frac{\partial}{\partial\theta}\log L_n(\theta, k) = \frac{\partial}{\partial\theta}\left(k\log\theta - \sum_{i=0}^{n-1}\log(\theta+i)\right) = \frac{k}{\theta} - \sum_{i=0}^{n-1}\frac{1}{\theta+i} = \frac{1}{\theta}\left(k - \sum_{i=0}^{n-1}\frac{\theta}{\theta+i}\right),$$

also ist $\widehat{\theta}_{\mathrm{ML}}$ Lösung von

$$k = \sum_{i=0}^{n-1}\frac{\widehat{\theta}_{\mathrm{ML}}}{\widehat{\theta}_{\mathrm{ML}}+i} \quad \left(\text{„} = \mathbb{E}_{\widehat{\theta}_{\mathrm{ML}}}[K_n]\text{“}\right)$$

(d.h. derjenige θ-Wert, unter dem erwartet=beobachtet).

Die Fisher-Information ist

$$I(\theta) \;=\; \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log L_n(\theta, K_n)\right)^2\right] = \frac{1}{\theta^2}\mathbb{E}_\theta\left[\left(K_n - \sum_{i=0}^{n-1}\frac{\theta}{\theta+i}\right)^2\right] = \frac{1}{\theta^2}\mathrm{Var}_\theta(K_n)$$

($\sim \frac{1}{\theta}\log n$ nach Satz 2.6).

2.1.1 Die GEM-Verteilung

Betrachte die Hoppe-Urne (für festes n alternativ: den Koaleszent mit Mutationen gemäß IMA-Modell), die Typen/Familien seien in Reihenfolge des Erscheinens nummeriert (“age order”)

$$X_k(n) := \text{Größe der } k\text{-ten Familie nach dem } n\text{-ten Zug aus der Hoppe-Urne},$$

(offenbar $X_1(1) = 1$, $X_k(n) = 0$ für $k > n$).

Frage:

$$\left(\frac{1}{n}X_1(n), \frac{1}{n}X_2(n), \frac{1}{n}X_3(n), \dots\right) \underset{n\to\infty}{\longrightarrow} ? \tag{2.10}$$

Beobachtung: $(n+\theta)^{-1}X_1(n)$, $n = 2, 3, \ldots$, ist ein (beschränktes) Martingal:

$$\mathbb{E}\Big[\frac{1}{n+1+\theta}X_1(n+1)\,\Big|\,\mathcal{F}_n\Big] = \frac{X_1(n)+1}{n+1+\theta}\frac{X_1(n)}{n+\theta} + \frac{X_1(n)}{n+1+\theta}\frac{\theta+n-X_1(n)}{n+\theta} = \frac{X_1(n)}{n+\theta}$$

(mit $\mathcal{F}_n = \sigma$(Beobachtungen bis einschließlich n-tem Zug)), konvergiert also f.s. Analog ist für $k \geq 2$ mit α_k = Zeitpunkt des ersten Auftretens von Typ k

$$(n+\alpha_k+\theta)^{-1}X_k(n+\alpha_k), \quad n = 1, 2, \ldots$$

ein (beschr.) Martingal.

Demnach: (2.10) konvergiert f.s. (zumindest koordinaten-weise).

Satz 2.11 (GEM-Verteilung[4])**.** *Seien* $B_1, B_2, \ldots$ *u.i.v.* $\mathrm{Beta}(1,\theta)$*, d.h. sie besitzen die Dichte* $\theta(1-b)^{\theta-1}$ *auf* $[0,1]$*. Die Verteilung des Grenzwerts in (2.10) ist gegeben durch*

$$\Big(B_1, (1-B_1)B_2, (1-B_1)(1-B_2)B_3, (1-B_1)(1-B_2)(1-B_3)B_4, \ldots\Big).$$

Definition 2.12 (Yule[5]-Prozess)**.** Ein zeitkontinuierlicher reiner Geburtsprozess (jedes Individuum erzeugt u.a. mit Rate 1 ein neues Individuum) heißt ein Yule-Prozess.

Es handelt sich also um eine zeitkontinuierliche Markovkette auf $\mathbb{N}$ mit Sprungraten

$$q_{n,n+1} = n = -q_{n,n},\ n \in \mathbb{N} \quad (q_{n,m} = 0 \text{ für } m \neq n, n+1).$$

Es ist [auch] ein Spezialfall eines zeitkont. (Galton-Watson-)Verzweigungsprozesses.

Lemma 2.13. *Sei* $(Y_t)_{t\geq 0}$ *ein Yule-Prozess (mit Geburtsrate* 1 *pro Individuum) und Startwert* $Y_0 = 1$*.*
Es gilt $\mathscr{L}(Y_t) = \mathrm{geom}(e^{-t})$ *und* $(e^{-t}Y_t)_{t\geq 0}$ *ist ein* L^2*-beschränktes Martingal mit*

$$\lim_{t\to\infty} e^{-t}Y_t \overset{d}{=} \mathrm{Exp}(1).$$

Beweis. Sei

$$T_i := |\{t : Y_t = i\}| \text{ die Länge des Zeitintervalls, in dem } i \text{ Ind. leben.}$$

Die Form der Raten zeigt:

$$T_1, T_2, \ldots \text{ sind u.a., } T_i \sim \mathrm{Exp}(i).$$

Somit

$$\mathbb{P}(Y_t > n) \quad = \quad \mathbb{P}(T_1 + \cdots + T_n < t) = \mathbb{P}\Big(\max_{i=1,\ldots,n} \tau_i < t\Big) = (1-e^{-t})^n, \quad n = 0, 1, 2, \ldots$$

mit τ_i u.i.v. $\mathrm{Exp}(1)$, d.h. $\mathscr{L}(Y_t) = \mathrm{Geom}(e^{-t})$.

(Alternativ beachte, dass die Lösung der Vorwärtsgleichung

$$\frac{\partial}{\partial t}\mathbb{P}_1(Y_t = n) = (n-1)\mathbb{P}_1(Y_t = n-1) - n\mathbb{P}_1(Y_t = n),\ \ \mathbb{P}_1(Y_0 = n) = \delta_{1n}$$

[4] Nach Bob Griffiths, Steinar Engen und John William Thomas Mccloskey benannt

[5] nach George Udny Yule, 1871–1951

gegeben ist durch $\mathbb{P}_1(Y_t = n) = e^{-t}(1 - e^{-t})^{n-1}$)

Zusammen mit der Verzweigungseigenschaft

$$\mathscr{L}(Y_t|Y_0 = k + j) = \mathscr{L}(Y_t|Y_0 = k) * \mathscr{L}(Y_t|Y_0 = j)$$

folgt: $\mathbb{E}[Y_{t+h}|Y_t] = e^h Y_t$, d.h. $(e^{-t}Y_t)_{t\geq 0}$ ist Martingal.

Weiter folgt leicht: $\sup_{t\geq 0} \mathbb{E}[(e^{-t}Y_t)^2] < \infty$ und $e^{-t}Y_t \to^d \mathrm{Exp}(1)$. □

Lemma 2.14. *Seien G_1 und G_2 u.a., $G_i \sim \mathrm{Gamma}(\theta_i)$ (d.h. Dichte $(\Gamma(\theta_i))^{-1} g^{\theta_i - 1} e^{-g}$ auf $\mathbb{R}_+$). Dann ist*

$$\mathscr{L}\Big(G_1 + G_2, \frac{G_1}{G_1 + G_2}\Big) = \mathrm{Gamma}(\theta_1 + \theta_2) \otimes \mathrm{Beta}(\theta_1, \theta_2).$$

Beweis. $G := G_1 + G_2$ ($G \sim \mathrm{Gamma}(\theta_1 + \theta_2)$). Die gemeinsame Dichte von (G_1, G) ist

$$f_{(G_1,G)}(g_1, g) = c\mathbf{1}(0 \leq g_1 \leq g) g_1^{\theta_1 - 1} e^{-g_1} (g - g_1)^{\theta_2 - 1} e^{-(g - g_1)} = ce^{-g}\mathbf{1}(0 \leq g_1 \leq g) g_1^{\theta_1 - 1} (g - g_1)^{\theta_2 - 1},$$

demnach ist die bedingte Dichte von G_1, gegeben $G = g$

$$f_{G_1|G=g}(g_1) = c(g)\mathbf{1}(0 \leq g_1 \leq g) g_1^{\theta_1 - 1} (g - g_1)^{\theta_2 - 1},$$

und somit die bedingte Dichte von G_1/G, gegeben $G = g$

$$f_{(G_1/G)|G=g}(b) = \tilde{c}(g)\mathbf{1}(0 \leq b \leq 1) b^{\theta_1 - 1} (1 - b)^{\theta_2 - 1}.$$

Da $\int_0^1 f_{(G_1/G)|G=g}(b)\, db = 1$ gilt, muss

$$\tilde{c}(g) = \Gamma(\theta_1 + \theta_2)/(\Gamma(\theta_1)\Gamma(\theta_2))$$

für jedes $g > 0$ gelten. □

Beweis von Satz 2.11. Darstellung via Yule-Prozess mit Immigration:

Seien $0 < T_1 < T_2 < \cdots$ die Sprungzeitpunkte eines Poissonprozesses auf $[0, \infty)$ mit Rate θ.

Der i-te Immigrant erscheint zum Zeitpunkt T_i, gründet i-te Familie, diese wächst ab dann als Yule-Prozess (vgl. Def. 2.12) unabhängig von allen anderen.

`[Bild an der Tafel]`

Seien

$$\begin{aligned}
Z_i(t) &:= \text{Größe der } i\text{-ten Familie zur Zeit } t \text{ (wir setzen } Z_i(t) = 0 \text{ für } t < T_i,\ Z_i(T_i) = 1),\\
S(t) &:= \sum_{i=1}^{\infty} Z_i(t) \quad \text{die Gesamtgröße der Population zur Zeit } t,\\
\tau_n &:= \min\{t : S(t) = n\} \quad \text{der Zeitpunkt, zu dem die Population auf } n \text{ anwächst.}
\end{aligned}$$

Es gilt

$$\Big(\frac{1}{n}Z_1(\tau_n), \frac{1}{n}Z_2(\tau_n), \frac{1}{n}Z_3(\tau_n), \dots\Big)_{n=1,2,\dots} \stackrel{d}{=} \Big(\frac{1}{n}X_1(n), \frac{1}{n}X_2(n), \frac{1}{n}X_3(n), \dots\Big)_{n=1,2,\dots} \tag{2.11}$$

Dazu Vergleich der Sprungraten: Es gebe aktuell

$$k \text{ Familien d. Größen } j_1, j_2, \dots, j_k \text{ mit } j_1 + \cdots + j_k = n.$$

- $S(t)$ springt nach $n+1$ mit Rate $n+\theta$,
- der Zuwachs
  - betrifft i-te Familie mit W'keit $j_i/(n+\theta)$,
  - erzeugt neue Familie mit W'keit $\theta/(n+\theta)$.

Also: Skelettkette des Yule-Prozesses mit Immigration $\widehat{=}$ Hoppe-Urne.

Lemma 2.13 zeigt:

$$\Big(e^{-t}Z_1(t), e^{-t}Z_2(t), e^{-t}Z_3(t), \dots\Big) \to \Big(e^{-T_1}A_1, e^{-T_2}A_2, e^{-T_3}A_3, \dots\Big) \quad \text{f.s.}, \tag{2.12}$$

(koordinaten-weise) mit $A_1, A_2, \dots$ u.i.v. Exp(1).

Daraus folgt

$$e^{-t}S(t) \to \sum_{n=1}^{\infty} e^{-T_n}A_n \quad \text{f.s.} \tag{2.13}$$

(Zur Rechtfertigung der Grenzwertvertauschung beachte, dass $M_i := \sup_{t\geq 0} e^{-t}Z_i(T_i+t)$ u.i.v. sind mit $\mathbb{E}M_1 < \infty$, also $\limsup M_n/n = 0$ gemäß Borel-Cantelli-Lemma ($\sum_{n=1}^{\infty}\mathbb{P}(M_i > \epsilon n) < \infty$ für jedes $\epsilon > 0$).

Weiterhin gilt $T_n/n \to \theta^{-1}$ f.s. gem. dem starken Gesetz der großen Zahlen, somit ist für $m \geq N_0$

$$\sup_{t\geq 0}\sum_{n=m}^{\infty} e^{-T_n}e^{-(t-T_n)}Z_n(t) \leq \sum_{n=m}^{\infty} e^{-T_n}M_n \leq \sum_{n=m}^{\infty} ne^{-2n/\theta}. \quad)$$

Weiterhin ist

$$\mathscr{L}\Big(\sum_{n=1}^{\infty} e^{-T_n}A_n\Big) = \text{Gamma}(\theta), \tag{2.14}$$

denn $\sum_i \delta_{(A_i,T_i)}$ ist ein Poissonscher Punktprozess auf $\mathbb{R}_+ \times \mathbb{R}_+$ mit Intensitätsmaß $\theta dt \otimes e^{-x}dx$ (und dies ist das Lévy-Maß des Gamma-Prozess/der Gamma-Verteilung, siehe z.B. Klenke [Kle20], Bsp. 16.15)

Für die Form des Intensitätsmaßes: sei $h : (0,\infty) \to \mathbb{R}_+$, sagen wir, stetig mit kompaktem Träger, so ist

$$\int_0^\infty \int_0^\infty h(e^{-t}a)\,\theta dt\, e^{-a}da = \int_0^\infty \int_0^a h(r)\theta\frac{dr}{r}e^{-a}da = \int_0^\infty h(r)\int_r^\infty e^{-a}\,da\,\theta\frac{dr}{r} = \int_0^\infty h(r)\theta e^{-r}\frac{dr}{r}.$$

(Die allgemeine Beobachtung dahinter ist folgende: Wenn $\Pi = \sum\delta_{a_i}$ ein PPP auf E mit Intensitätsmaß ν ist und $f : E \to E'$, dann ist $\tilde{\Pi} = \sum\delta_{f(a_i)}$ ein PPP auf E' mit Intensitätsmaß $\tilde{\nu} = \nu \circ f^{-1}$.)

Das Argument für (2.14) zeigt auch

$$\mathscr{L}(G,T) = \text{Gamma}(1+\theta)\otimes\text{Exp}(1) \;\Rightarrow\; \mathscr{L}(e^{-T/\theta}G) = \text{Gamma}(\theta), \tag{2.15}$$

denn $\sum_{n=1}^{\infty} e^{-T_n}A_n = e^{-T_1}\big(A_1 + \sum_{n=2}^{\infty} e^{-(T_n-T_1)}A_n\big)$.

(Alternativ beachte man, dass $e^{-T/\theta} \sim \text{Beta}(\theta,1)$ gilt und verwende Lemma 2.14.)

Somit gilt

$$\frac{Z_1(t)}{S(t)} = \frac{e^{T_1-t}Z_1(t)}{e^{T_1-t}Z_1(t) + \sum_{i=2}^{\infty} e^{T_1-t}Z_i(t)} \to \frac{A_1}{A_1 + \sum_{i=2}^{\infty} e^{-(T_i-T_1)}A_i} =: B_1 \quad \text{f.s.},$$

und $\mathscr{L}(B_1) = \mathrm{Beta}(1,\theta)$, wobei B_1 und $A_1 + \sum_{i=2}^{\infty} e^{-(T_i - T_1)} A_i$ u.a. (Lemma 2.14).

Sei

$$C_n := A_n + \sum_{i=n+1}^{\infty} e^{-(T_i - T_n)} A_i, \quad B_n := \frac{A_n}{C_n}.$$

Zeige induktiv:

$$\mathscr{L}\Big(C_1, B_1, B_2, \dots, B_n\Big) = \mathrm{Gamma}(1+\theta) \otimes \mathrm{Beta}(1,\theta)^{\otimes n} \quad \text{für } n \in \mathbb{N}. \tag{2.16}$$

Der Fall $n = 1$ stimmt nach obigem.

Für den Schluss von $n \to n+1$: I.V. und Stationarität sowie Unabhängigkeit der Poisson-Zuwächse liefern

$$\mathscr{L}(C_2, B_2, B_3, \dots, B_{n+1}) = \mathrm{Gamma}(1+\theta) \otimes \mathrm{Beta}(1,\theta)^{\otimes n};$$

zudem sind $(C_2, B_2, B_3, \dots, B_{n+1})$ und $(A_1, T_2 - T_1)$ unabhängig.

Nach Def. ist

$$C_1 = A_1 + e^{-(T_2 - T_1)} C_2, \; B_1 = \frac{A_1}{C_1} = \frac{A_1}{A_1 + e^{-(T_2 - T_1)} C_2}.$$

Es ist $e^{-(T_2 - T_1)} C_2 \sim \mathrm{Gamma}(\theta)$ nach (2.15) und $A_1 \sim \mathrm{Exp}(1)$ u.a. von $e^{-(T_2 - T_1)} C_2$, d.h.

$$(C_1, B_1) \sim \mathrm{Gamma}(\theta) \otimes \mathrm{Beta}(1,\theta)$$

nach Lemma 2.14, dies liefert den Induktionsssschluss.

Schließlich gilt

$$\begin{aligned}
\frac{e^{-t} Z_n(t)}{e^{-t} S(t)} \quad \to \quad & \frac{e^{-T_n} A_n}{\sum_{i=1}^{\infty} e^{-T_i} A_i} = \frac{\sum_{i=2}^{\infty} e^{-T_i} A_i}{\sum_{i=1}^{\infty} e^{-T_i} A_i} \times \cdots \times \frac{\sum_{i=n}^{\infty} e^{-T_i} A_i}{\sum_{i=n-1}^{\infty} e^{-T_i} A_i} \times \frac{e^{-T_n} A_n}{\sum_{i=n}^{\infty} e^{-T_i} A_i} \\
= \quad & (1 - B_1) \times \cdots \times (1 - B_n) \times \frac{A_n}{A_n + \sum_{i=n+1}^{\infty} e^{-(T_i - T_n)} A_i} \\
= \quad & (1 - B_1) \times \cdots \times (1 - B_{n-1}) B_n.
\end{aligned}$$

□

Beobachtung 2.15 (Poisson-Dirichlet-Verteilung)**.** Seien $1 \geq V_1 > V_2 > \cdots$ die (der Größe) nach sortierten Einträge (=Typenhäufigkeiten) aus GEM-verteiltem

$$\Big(B_1, (1 - B_1) B_2, (1 - B_1)(1 - B_2) B_3, (1 - B_1)(1 - B_2)(1 - B_3) B_4, \dots\Big).$$

Sei $\Pi = \sum_i \delta_{X_i}$ PPP auf $\mathbb{R}_+$ mit Intensitätsmaß $(\theta/x) e^{-x} dx$ (Π beschreibt die Sprünge eines Standard-Gamma-Subordinators bis zur Zeit θ) und $S := \sum X_i$, seien $X_{[1]} > X_{[2]} > \cdots$ die Ordnungsstatistik der X_is. Dann ist

$$\big(X_{[1]}/S, X_{[2]}/S, \dots\big) \stackrel{d}{=} \big(V_1, V_2, \dots\big).$$

Dies folgt aus dem Beweis von Satz 2.11. Diese Verteilung (ein W'maß auf $\{(x_1, x_2, \dots) \in [0,1]^{\mathbb{N}} : x_1 + x_2 + \cdots = 1\}$), heißt die Poisson-Dirichlet-Verteilung (mit Parameter θ).

Bericht 2.16 (Endlich viele Typen mit elternunabhängiger Mutation)**.** Betrachte Mutationsmodell mit d neutralen Typen (Typenmenge $E = \{1, \dots, d\}$), jede Linie mutiert mit Rate $\theta/2$, Typ nach Mutation ist j mit W'keit π_j (> 0) $((\pi_1, \dots, \pi_d)$ Ws-Gewichte auf $\{1, \dots, d\})$, u.a. vom vorigen Typ.

Die Typenverteilung in einer unendlichen Population im Gleichgewicht ist dann

$$\mathscr{L}\Big(Z_1(\infty), \dots, Z_d(\infty)\Big) \;=\; \mathrm{Dirichlet}(\theta\pi_1, \dots, \theta\pi_d), \tag{2.17}$$

d.h. die Dichte ist

$$\frac{\Gamma(\theta)}{\Gamma(\theta\pi_1)\cdots\Gamma(\theta\pi_d)} x_1^{\theta\pi_1 - 1} \cdots x_k^{\theta\pi_d - 1}$$

bezüglich dem Lebesgue-Maß auf $\{(x_1, \dots, x_d) : 0 \le x_i \le 1, x_1 + \cdots + x_d = 1\}$.

Man kann den allgemeinen Fall aus Beob. 2.15 herleiten: Wenn man jeden Sprung des Gamma-Subordinators unabhängig gemäß π mit einer „Farbe" aus $\{1, \dots, d\}$ einfärbt, so bilden die Sprünge jeder Farbe für sich jeweils unabhängige Gamma-Subordinatoren (mit entsprechend verkleinerter Intensität), somit: $Y_i \sim \mathrm{Gamma}(\theta\pi_i)$ und $Y_1, \dots, Y_d$ unabhängig, so ist

$$\Big(Z_1(\infty), \dots, Z_d(\infty)\Big) \stackrel{d}{=} \left(\frac{Y_1}{Y_1 + \cdots + Y_k}, \dots, \frac{Y_k}{Y_1 + \cdots + Y_k}\right)$$

und die rechte Seite ist $\mathrm{Dirichlet}(\theta\pi_1, \dots, \theta\pi_d)$-verteilt (dies ist eine multivariate Verallgemeinerung von Lemma 2.14, siehe z.B. Ch. 40.5 in Norman L. Johnson, Samuel Kotz, *Distributions in statistics: continuous multivariate distributions*, Wiley, 1972).

2.2 Infinitely-many-sites-Modell (IMS)

Definition 2.17 (Infinitely-many-sites-Modell[6])**.** Man nimmt an, dass jede Mutation eine neue, bisher noch nie mutierte Position am betrachteten Lokus betrifft.

Mathematisch realisiert man dies z.B. folgendermaßen: Die betrachtete Stelle im Genom (eine gewisse Abfolge von Nukleotiden im DNS-Doppelstrang eines Chromosoms) entspricht $[0, 1]$, jede Mutation erhält eine neue, uniform aus $[0, 1]$ gewählte „Position", der Typ eines Individuums ist ein (einfaches) Zählmaß auf $[0, 1]$ (bzw. alternativ eine Teilmenge von $[0, 1]$), der Typ eines Individuums gibt an, wo dieses relativ zu einen „Referenztyp" (oder „Wildtyp") mutiert ist.

Das infinitely-many-sites-Modell (IMS-Modell, in der Literatur auch infinite-sites-Modell genannt) ist für viele praktische Zwecke eine angemessene Approximation für die Beschreibung von Mutationen auf dem Niveau der DNS-Sequenz : Wenn die Mutationsrate pro Basenpaar sehr klein und die betrachtete Stelle im Genom (der sog. Lokus) nicht „zu lang" ist, ist es plausibel, die Möglichkeit der Mehrfachmutation einer Stelle (und andere Effekte, die im IMS-Modell nicht berücksichtigt werden, etwa Rekombination oder Insertionen/Deletionen längerer Stücke im Genom) zu vernachlässigen. (Man denke an eine Abfolge von $L \gg 1$ Basenpaaren, bei Mutation wird eine der zufällig gewählte der L Positionen zufällig modifiziert.)

[6] Eingeführt in G. A. Watterson, On the number of segregating sites in genetical models without recombination, Theoretical Population Biology 7 (2), 256–276, (1975).

Beispiel 2.18. John Parsch, Colin D. Meiklejohn, and Daniel L. Hartl, Patterns of DNA sequence variation suggest the recent action of positive selection in the janus-ocnus region of drosophila simulans, *Genetics* 159:647–657, (2001) berichten (u.a.) genetische Variabilität in einem ca. 1.700 Basenpaare langen Stück des Chromosoms 3 in einer (weltweiten) Stichprobe von 8 *Drosophila simulans* und einer Stichprobe von *Drosophila melanogaster*, zwei verwandten Arten von Taufliegen. An insgesamt 31 Stellen sind Unterschiede zwischen den Individuen sichtbar (siehe Tabelle 2.1, es sind nur die sogenannten variablen oder segregierenden Positionen aufgeführt).

Die Sequenzinformation von *Drosophila melanogaster* — diese Stichprobe bildet bezüglich der 8 Stichproben von *Drosophila simulans* eine „outgroup" — gestattet (zusammen mit den IMS-Modellannahmen) an jeder Position zu entscheiden, welche Base die anzestrale und welche die mutierte ist.

	Position																														
																				1	1	1	1	1	1	1	1	1	1	1	1
				2	2	2	3	5	5	5	6	6	6	6	6	6	6	7	8	1	1	2	3	3	3	4	6	6	6	6	6
	3	8	9	4	9	9	6	3	4	8	4	5	5	5	5	6	6	0	4	3	3	9	0	2	8	5	1	7	8	8	9
	5	3	3	6	1	4	2	8	9	5	4	1	2	7	8	5	7	7	3	2	9	1	0	2	8	2	4	3	1	2	4
s1	c	g	a	t	c	c	a	a	t	a	t	a	a	a	g	c	t	c	g	a	t	a	a	g	c	c	g	a	t	t	c
s2	.	.	c	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	g	.	.	.	.	.
s3	a	c	.	c	a	t	g	c	c	c	g	g	g	g	a	t	c	t	a	t	c	c	t	c	t	g	t	t	g	c	a
s4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
s5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	g	.	.	.	.	.
s6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
s7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	g	.	.	.	.	.
s8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	g	.	.	.	.	.
m1	.	.	.	c	a	t	g	c	c	c	a	g	.	.	.	.	.	t	.	.	c	c	.	.	t	g	.	t	g	c	a

Tabelle 2.1: Beobachtete genetische Variabilität in einer Region in Chromosom 3 aus einer Stichprobe von 8 *Drosophila simulans* (Zeilen s1–s8) und einer Stichprobe von *Drosophila melanogaster* (Zeile m1) aus Parsch et al, *Genetics* 159:647–657, (2001). Siehe Figure 2 dort, wir betrachten hier nur den Teil der Sequenz, der die Gene *janA* und *janB* umfasst.

Zur Beschreibung von beobachteten Sequenzdaten in einer Stichprobe der Größe n im Kontext des IMS-Modells betrachten wir folgende Modellvorstellung: Die n-Stichprobe entsteht aus einem n-Koaleszent, längs dessen Ästen sich mit Rate $\frac{\theta}{2}$ Mutationen ereignen (und jede trifft eine völlig neue Position).

Die Anzahl segregierender Stellen ist

$$S_n = \#\text{ verschiedene Mutationen, die in } n\text{-Stichprobe vorkommen}$$

(im Sinne von: Positionen, an denen sich mindestens zwei Stichproben unterscheiden). Wenn $S_n = s$, so entsprechen die Beobachtungen einer $n \times s$-Datenmatrix $(D_{ik})_{i=1,\dots,n;k=1,\dots,s}$

$$D_{ik} = \mathbf{1}(\text{Stichprobe } i \text{ ist an } k\text{-ter segregierender Stelle mutiert}).$$

Beispielsweise sehen wir in Abbildung 2.1 eine Realisierung mit $S_4 = 3$.

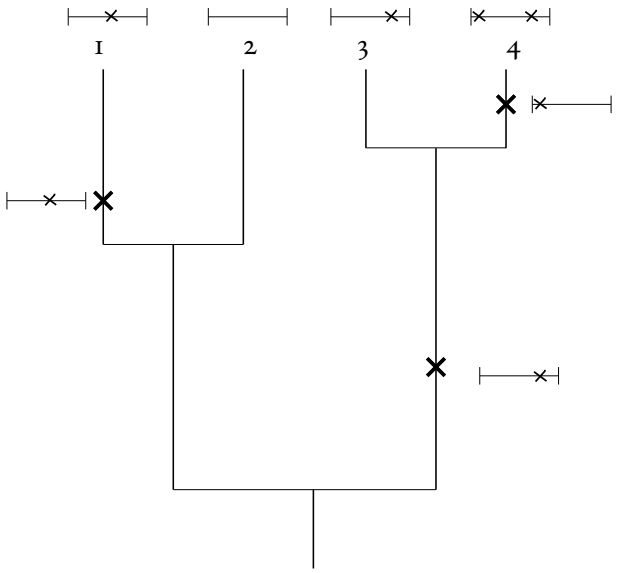


Abbildung 2.1: Ein 4-Koaleszent, längs dessen Kanten Mutationen gemäß IMS-Modell auftreten. Wir registrieren für jede Mutation die mutierte Position (in $[0,1]$) und an den Blättern (den Stichproben) sämtliche Mutationen, die sich auf dem kürzesten Weg vom jeweiligen Blatt zur Wurzel befinden.

Bemerkung 2.19 (Unbekannter Wildtyp)**.** Wenn man „nur" die Stichprobe sieht und keine externen Zusatzinformationen (z.B. eine „outgroup" durch inter-Spezies-Vergleich wie in Bsp. 2.18) besitzt, kann man an den segregierenden Stellen nicht entscheiden, welcher Typ der Wildtyp und welcher die Mutante ist (im Genetik-Jargon: die Mutationen sind „unpolarisiert"). In dieser Situation ist obige Datenmatrix nur bis auf „Umklappen" von Spalten definiert, d.h. die eigentliche Information ist S_n und

$$\Delta_{i,j}(k) = \begin{cases} 1, & \text{Stichproben } i \text{ und } j \text{ an } k\text{-ter segregierender Stelle verschieden,} \\ 0, & \text{sonst} \end{cases}$$

für $k = 1, \dots, S_n$.

Beobachtung 2.20. Es gilt

$$\mathbb{E}_\theta[S_n] = \theta\, h_n, \quad \operatorname{Var}_\theta[S_n] = \theta\, h_n + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2}$$

mit $h_n = \sum_{i=1}^{n-1} \frac{1}{i}$.

$$\widehat{\theta}_W := \frac{S_n}{h_n}$$

ist ein erwartungstreuer Schätzer für θ (der sogenannte Watterson-Schätzer) mit

$$\operatorname{Var}_\theta\left[\widehat{\theta}_W\right] \sim \frac{\theta}{\log n} \quad \text{und} \quad \frac{\widehat{\theta}_W - \theta}{\sqrt{\operatorname{Var}_\theta\left[\widehat{\theta}_W\right]}} \xrightarrow{d} \mathcal{N}(0,1) \qquad \text{für } n \to \infty.$$

Beweis. Schreibe

$$S_n = S_{n,n} + S_{n,n-1} + \cdots + S_{n,2}$$

mit

$$S_{n,j} = \#\text{ Mutationen, während die Genealogie aus } j \text{ Linien besteht.}$$

Wir hatten in der Diskussion in Kapitel 1.5 gesehen, dass $S_{n,j} \sim \mathrm{geom}(\frac{j-1}{\theta+j-1})$ (denn gegeben T_j, die Zeit, während der j Linien in der Genealogie existieren, ist $S_{n,j} \sim \mathrm{Pois}(\frac{\theta}{2} j T_j)$) und $S_{n,n}, \dots, S_{n,2}$ sind unabhängig. Somit

$$\mathbb{E}_\theta[S_{n,j}] = \tfrac{\theta+j-1}{j-1} - 1 = \theta/(j-1), \quad \mathrm{Var}_\theta[S_{n,j}] = \big(\tfrac{\theta}{\theta+j-1}\big)/\big(\tfrac{j-1}{\theta+j-1}\big)^2 = \frac{\theta(\theta+j-1)}{(j-1)^2} = \frac{\theta}{j-1} + \frac{\theta^2}{(j-1)^2},$$

was die Formeln für Erwartungswert und Varianz von S_n beweist.

Zur asymptotischen Normalität von $\widehat{\theta}_W$: Schreibe

$$X_{n,j} := \frac{S_{n,j} - \theta/(j-1)}{\sqrt{\mathrm{Var}_\theta\big[\widehat{\theta}_W\big]}}, \quad j = 2, 3, \dots, n.$$

Die $X_{n,j}$ bilden ein unabhängiges, zentriertes und normiertes Dreiecksschema (für festes n sind $X_{n,2}, \dots, X_{n,n}$ unabhängig mit $\mathbb{E}_\theta[X_{n,j}] = 0$ und $\sum_{j=2}^n \mathrm{Var}_\theta[X_{n,j}] = 1$), für $\varepsilon > 0$ und n so groß, dass $\varepsilon \mathrm{Var}_\theta\big[\widehat{\theta}_W\big] > \theta^2$ gilt, ist

$$\begin{aligned}
\mathbb{E}\big[X_{n,j}^2 \mathbf{1}(X_{n,j}^2 > \varepsilon)\big] &= \frac{1}{\mathrm{Var}_\theta\big[\widehat{\theta}_W\big]} \mathbb{E}\Big[\big(S_{n,j} - \theta/(j-1)\big)^2 \mathbf{1}\big(S_{n,j} > \theta/(j-1) + \sqrt{\varepsilon \mathrm{Var}_\theta\big[\widehat{\theta}_W\big]}\big)\Big] \\
&\le \frac{1}{\mathrm{Var}_\theta\big[\widehat{\theta}_W\big]} \mathbb{E}\Big[S_{n,j}^2 \mathbf{1}\big(S_{n,j} > \theta/(j-1) + \sqrt{\varepsilon \mathrm{Var}_\theta\big[\widehat{\theta}_W\big]}\big)\Big].
\end{aligned}$$

Demnach erfüllt das Schema die Lindeberg-Bedingung

$$\lim_{n\to\infty} \mathbb{E}\big[X_{n,j}^2 \mathbf{1}(X_{n,j}^2 > \varepsilon)\big] = 0 \tag{2.18}$$

und daher ist das renormierte $\widehat{\theta}_W$ asymptotisch normalverteilt (siehe z.B. [Kle20, Satz 15.43]).

(beachte: Für $X \sim \mathrm{geom}(p)$ gilt

$$\begin{aligned}
\mathbb{E}\big[X^2 \mathbf{1}(X \ge m)\big] &= \sum_{k=m}^\infty k^2 p(1-p)^{k-1} \le \sum_{k=m}^\infty (k+1) k p (1-p)^{k-1} = p \sum_{k=m}^\infty \Big[\frac{d^2}{dy^2}(1-y)^{k+1}\Big]_{y=p} \\
&= p\Big[\frac{d^2}{dy^2} \sum_{k=m}^\infty (1-y)^{k+1}\Big]_{y=p} = p\Big[\frac{d^2}{dy^2} \frac{(1-y)^{m+1}}{y}\Big]_{y=p} \\
&= p \frac{(1-p)^{m-1}}{p}\Big((m+1)m + 2(m+1)\frac{1-p}{p} + 2\frac{(1-p)^2}{p^2}\Big) \\
&\le \big((m+1)(m+2) + 2\big)\frac{(1-p)^{m-1}}{p^2},
\end{aligned}$$

demnach für $X = S_{n,j}$ mit $p = p_j = \frac{j-1}{\theta+j-1}$ und z.B. $m = \sqrt{\frac{1}{2}\varepsilon\theta \log n}$ ist

$$\frac{1}{\mathrm{Var}_\theta\big[\widehat{\theta}_W\big]} \mathbb{E}\Big[S_{n,j}^2 \mathbf{1}\big(S_{n,j} > \sqrt{\tfrac{1}{2}\varepsilon\theta \log n}\big)\Big] \le C_\theta \big(\tfrac{\theta}{\theta+j-1}\big)^{\sqrt{\frac{1}{2}\varepsilon\theta \log n}}$$

für ein $C_\theta < \infty$, d.h (2.18) gilt.) □

Bemerkung 2.21 (Alternativer Zugang zu Beob. 2.20)**.** Wir könnten Erwartungswert und Varianz von S_n auch folgendermaßen berechnen: Gegeben die Gesamtlänge L_{ges} des Koaleszenten ist S_n $\text{Poi}\big((\theta/2)L_{\text{ges}}\big)$-verteilt.

$$L_{\text{ges}} \stackrel{d}{=} \sum_{j=2}^{n} jT_j,$$

mit $T_n, T_{n-1}, \ldots, T_2$ u.a., $\mathscr{L}(T_j) = \text{Exp}\big(\binom{j}{2}\big)$, somit

$$\begin{aligned}
\mathbb{E}_\theta[S_n] &= \frac{\theta}{2}\sum_{j=2}^{n} j\binom{j}{2} = \theta\sum_{i=1}^{n-1}\frac{1}{i}, && (2.19)\\
\text{Var}_\theta[S_n] &= \mathbb{E}_\theta\big[\text{Var}_\theta[S_n \mid L_{\text{ges}}]\big] + \text{Var}_\theta\big[\mathbb{E}_\theta[S_n \mid L_{\text{ges}}]\big]\\
&= \mathbb{E}_\theta\big[\frac{\theta}{2}L_{\text{ges}}\big] + \text{Var}_\theta\big[\frac{\theta}{2}L_{\text{ges}}\big] = \theta\sum_{i=1}^{n-1}\frac{1}{i} + \theta^2\sum_{i=1}^{n-1}\frac{1}{i^2}. && (2.20)
\end{aligned}$$

Die Konvergenzordnung $O(1/\log(n))$ der Varianz des Watterson-Schätzers ist zwar von Standpunkt der Statistik gesehen „frustrierend langsam“ (zumal im Vergleich zur klassischen Situation, in der man einen Parameter basierend auf n *unabhängigen* Beobachtungen schätzt, dort hat man typischerweise Abfall der Varianz $O(1/n)$ für plausible Schätzer). Andererseits haben Y. X. Fu and W. H. Li, Maximum Likelihood Estimation of Population Parameters, *Genetics* 134 (4), 1261–1270, (1993) gezeigt, dass es (zumindest asymptotisch) auch nicht besser möglich ist:

Satz 2.22. *Jeder erwartungstreue Schätzer für θ im IMS-Modell*

$$\textit{hat unter } \mathbb{P}_\theta \textit{ mindestens Varianz } \theta\Big/\sum_{k=1}^{n-1}\frac{1}{\theta+k} \quad (\sim \theta/\log n \textit{ für } n\to\infty).$$

Beweis. Nehmen wir (zunächst) an, wir könnten $S_{n,2} = s_{n,2}, S_{n,3} = s_{n,3}, \ldots, S_{n,n} = s_{n,n}$ beobachten (was anhand von Sequenzdaten an den Blättern des Koaleszenten nicht immer möglich ist):

Die Likelihoodfunktion (die Verteilungsgewichte von $(S_{n,n}, \ldots, S_{n,2})$, aufgefasst als Funktion des Parameters θ) ist

$$\begin{aligned}
L_n(s_{n,2}, \ldots, s_{n,n}; \theta) &= \prod_{j=2}^{n}\frac{j-1}{\theta+j-1}\left(\frac{\theta}{\theta+j-1}\right)^{s_{n,j}}\\
&= (n-1)!\theta^{s_n}\prod_{j=2}^{n}\big(\theta+j-1\big)^{-(s_{n,j}+1)}
\end{aligned}$$

mit $s_n = s_{n,2} + \cdots + s_{n,n}$, also

$$\frac{\partial}{\partial\theta}\log L_n(s_{n,2}, \ldots, s_{n,n}; \theta) = \frac{s_n}{\theta} - \sum_{j=2}^{n}\frac{s_{n,j}+1}{\theta+j-1}$$

d.h. $\widehat{\theta}_{\text{ML,hyp}}$, der Maximum-Likelihood-Schätzer für θ basierend auf $(S_{n,n}, \ldots, S_{n,2})$, ist die Lösung (in θ) von $s_n = \theta\sum_{j=2}^{n}\frac{s_{n,j}+1}{\theta+j-1}$.

Weiter ist

$$\frac{\partial^2}{\partial\theta^2}\log L_n(s_{n,2}, \ldots, s_{n,n}; \theta) = -\frac{s_n}{\theta^2} + \sum_{j=2}^{n}\frac{s_{n,j}+1}{(\theta+j-1)^2},$$

die Fisher-Information ist somit

$$\begin{aligned} I(\theta) &= -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log L_n(s_{n,2},\dots,s_{n,n};\theta)\right] \\ &= \mathbb{E}_\theta\left[\frac{S_n}{\theta^2}\right] - \sum_{j=2}^{n}\mathbb{E}_\theta\left[\frac{S_{n,j}+1}{(\theta+j-1)^2}\right] = \frac{\sum_{k=1}^{n-1}1/k}{\theta} - \sum_{j=2}^{n}\frac{\theta+j-1}{(j-1)\,(\theta+j-1)^2} \\ &= \frac{\sum_{k=1}^{n-1}1/k}{\theta} - \sum_{j=2}^{n}\frac{1}{(j-1)(\theta+j-1)} = \frac{1}{\theta}\sum_{k=1}^{n-1}\left(\frac{1}{k}-\frac{\theta}{k(\theta+k)}\right) = \frac{1}{\theta}\sum_{k=1}^{n-1}\frac{1}{\theta+k}. \end{aligned}$$

Gemäß der Cramér-Rao-Ungleichung (siehe z.B. John A. Rice, *Mathematical statistics and data analysis*, Duxbury Press, 1995, Ch. 8.6 oder die knappe Diskussion unten) gilt für jeden erwartungstreuen Schätzer

$$T = T(S_{n,n},\dots,S_{n,2})$$

für θ (d.h. der Schätzer wird durch eine Funktion $T : \mathbb{N}_0^{n-1} \to (0,\infty)$ mit $\mathbb{E}_\theta\big[T(S_{n,n},\dots,S_{n,2})\big] = \theta$ für alle $\theta > 0$ dargestellt)

$$\mathrm{Var}_\theta[T(S_{n,n},\dots,S_{n,2})] \geq \frac{1}{I(\theta)} = \frac{\theta}{\sum_{k=1}^{n-1}\frac{1}{\theta+k}},$$

d.h. die Behauptung.

Nachträge:

1. Heuristik zur Cramér-Rao-Ungleichung:

   Betrachten wir die allgemeine Situation, dass die Beobachtungen X ein Zufallsvektor (mit Werten in einer geeigneten Teilmenge von $\mathbb{R}^d$ für ein d) sind, die Dichte-/Likelihood-Funktion $f(x;\theta)$ (im diskreten Fall: die Gewichte) sei genügend glatt, so dass die folgenden Vertauschungen von Ableitung und Integral gerechtfertigt sind.

   Sei $V(X) := \frac{\partial}{\partial\theta}\log f(X;\theta) = \frac{1}{f(X;\theta)}\frac{\partial}{\partial\theta}f(X;\theta)$ die „Score-Funktion“, es ist $\mathbb{E}_\theta[V(X)] = 0$ stets, denn

   $$\int f(x;\theta)\frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta}f(x;\theta)\,dx = \int\frac{\partial}{\partial\theta}f(x;\theta)\,dx = \frac{\partial}{\partial\theta}\int f(x;\theta)\,dx = \frac{\partial}{\partial\theta}1 = 0.$$

   $$I(\theta) := \mathrm{Var}_\theta[V(X)] = \mathbb{E}_\theta[V(X)^2]$$

   heißt die Fisher-Information, beachte auch

   $$I(\theta) = -\mathbb{E}_\theta[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)]$$

   (d.h. wir können die Fisher-Information als (erwartete) Krümmung der Likelihood-Funktion an den beobachteten Daten interpretieren), denn

   $$\frac{\partial^2}{\partial\theta^2}\log f(x;\theta) = \frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)} - \Big(\frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}\Big)^2 = \frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)} - \Big(\frac{\partial}{\partial\theta}\log f(x;\theta)\Big)^2$$

und

$$\mathbb{E}_\theta[\frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)}] = \int \frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)} f(x;\theta)\,dx = \int \frac{\partial^2}{\partial\theta^2} f(x;\theta)\,dx = \frac{\partial^2}{\partial\theta^2}\int f(x;\theta)\,dx = \frac{\partial^2}{\partial\theta^2} 1 = 0.$$

Sei nun $T(X)$ irgendein Schätzer für θ (d.h. eine Funktion der Beobachtungen X mit Werten in $[0,\infty)$ (und $\mathbb{E}_\theta[(T(X))^2] < \infty$), dann ist

$$\begin{aligned}\mathrm{Cov}_\theta[T(X), V(X)] = \mathbb{E}_\theta[T(X)V(X)] &= \int T(x)\frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta} f(x;\theta)\, f(x;\theta)\,dx \\ &= \int T(x)\frac{\partial}{\partial\theta} f(x;\theta)\,dx = \frac{\partial}{\partial\theta}\mathbb{E}_\theta[T(X)].\end{aligned}$$

Die Cauchy-Schwarz-Ungleichung liefert

$$\sqrt{\mathrm{Var}_\theta[T(X)]\mathrm{Var}_\theta[V(X)]} \geq |\mathrm{Cov}_\theta[T(X), V(X)]| = \left|\frac{\partial}{\partial\theta}\mathbb{E}_\theta[T(X)]\right|$$

und somit gilt

$$\mathrm{Var}_\theta[T(X)] \geq \frac{\left|\frac{\partial}{\partial\theta}\mathbb{E}_\theta[T(X)]\right|^2}{\mathrm{Var}_\theta[V(X)]} = \frac{\left|\frac{\partial}{\partial\theta}\mathbb{E}_\theta[T(X)]\right|^2}{I(\theta)}.$$

falls $T(X)$ ein unverzerrter Schätzer für θ ist, d.h. $\mathbb{E}_\theta[T(X)] = \theta$ für alle θ, so ist natürlich $\frac{\partial}{\partial\theta}\mathbb{E}_\theta[T(X)] = 1$ (und dies ist die Form der Cramér-Rao-Ungleichung, die wir oben verwendet haben).

2. Wir hatten die Schranke unter der Annahme hergeleitet, dass wir $(S_{n,n}, \ldots, S_{n,2})$ tatsächlich beobachten könnten, was anhand der Daten i.A. nicht möglich ist. Intuitiv erscheint es zumindest sehr plausibel, dass jeder „reale“ erwartungstreue Schätzer für θ (d.h. jedes $\widetilde{T} = \widetilde{T}(D)$, das eine Funktion der Datenmatrix D ist, das also weniger Informationen verwenden darf, als wir oben angenommen hatten), ebenfalls mindestens Varianz $1/I(\theta)$ hat.

   Diese Intuition kann man durch den Begriff der (statistischen) Suffizienz folgendermaßen formalisieren: Sei X die „volle“ Information, die die Genealogie der n-Stichprobe und die darauf vorkommenden Mutationen beschreibt, d.h. X enthält die „topologische“ Information, in welcher Reihenfolge die Verschmelzungen der Linien stattfinden, und für jede Kante im Baum die Information, welche Mutationen auf dieser liegen (wir verzichten hier darauf, dies in Formeln zu fassen). Offenbar kann man aus X die Datenmatrix D ablesen und somit kann $\widetilde{T} = \widetilde{T}(D(X))$ als eine Funktion von X interpretiert werden.

   Die entscheidende Beobachtung ist, dass $Y := (S_{n,2}, S_{n,3}, \ldots, S_{n,n})$ *suffizient* für θ ist, d.h. die bedingte Verteilung $\mathscr{L}_\theta(X \mid Y)$ hängt nicht von θ ab — gegeben $S_{n,2} = s_{n,2}, \ldots, S_{n,n} = s_{n,2}$ entsteht X, indem man für $j = n, n-1, \ldots, 2$ auf den j Kanten in „Niveau“ j des Koaleszenten $s_{n,j}$ Mutationen uniform verteilt und unter allen aktuell möglichen Verschmelzungen uniform eine auswählt; demnach enthalten die Gewichte von $\mathscr{L}_\theta(X \mid Y)$ nur kombinatorische Terme, aber keine θ-Abhängigkeit.

   Nun ist (beachte, dass wir den bedingten Erwartungswert bilden können, ohne θ zu kennen)

$$\widehat{T} := \widehat{T}(Y) := \mathbb{E}[\widetilde{T} \mid Y]$$

ebenfalls ein erwartungstreuer Schätzer für θ und nach Konstruktion ist $\widehat{T}$ eine gewisse Funktion von $Y = (S_{n,2}, S_{n,3}, \dots, S_{n,n})$, d.h. nach obigem ist $\mathrm{Var}_\theta[\widehat{T}] \geq 1/I(\theta)$ und folglich auch

$$\mathrm{Var}_\theta[\widetilde{T}] = \mathbb{E}_\theta[\underbrace{\mathrm{Var}_\theta[\widetilde{T}\,|\,Y]}_{\geq 0}] + \mathrm{Var}_\theta[\underbrace{\mathbb{E}_\theta[\widetilde{T}\,|\,Y]}_{=\widehat{T}}] \geq \mathrm{Var}_\theta[\widehat{T}] \geq \frac{1}{I(\theta)}.$$

□

Definition 2.23 (Frequenzspektrum)**.** Sei

$$\xi_i^{(n)} := \#\text{ Mutationen, die in genau } i \text{ der } n \text{ Stichproben vorkommen}, \quad i = 1, \dots, n-1.$$

(Wir nehmen dabei an, dass an jeder Position der anzestrale oder „Wildtyp“ bekannt ist, z.B. durch Interspezies-Vergleich.) Der Vektor

$$\xi^{(n)} = \left(\xi_1^{(n)}, \xi_2^{(n)}, \dots, \xi_{n-1}^{(n)}\right)$$

heißt das Frequenzspektrum (der segregrierenden Stellen).

Wenn der anzestrale Typ nicht bekannt ist, betrachtet man stattdessen das gefaltete Frequenzspektrum $\left(\eta_1^{(n)}, \eta_2^{(n)}, \dots, \eta_{\lfloor n/2 \rfloor}^{(n)}\right)$ mit

$$\eta_i^{(n)} := \xi_i^{(n)} + \xi_{n-i}^{(n)} \mathbf{1}_{i \neq n/2}, \quad 1 \leq i \leq \lfloor n/2 \rfloor.$$

Satz 2.24. *Es gilt*

$$\mathbb{E}_\theta\left[\xi_i^{(n)}\right] = \frac{\theta}{i}, \quad \mathrm{Cov}_\theta\left[\xi_i^{(n)}, \xi_j^{(n)}\right] = \mathbf{1}_{i=j}\frac{\theta}{i} + \theta^2 \sigma_{ij}, \quad 1 \leq i \leq j \leq n$$

mit $h_n := \sum_{i=1}^{n-1} \frac{1}{i}$, $\beta_n(i) := \frac{2n}{(n-i+1)(n-i)}\left(h_{n+1} - h_i\right) - \frac{2}{n-i}$

$$\sigma_{ii} = \begin{cases} \beta_n(i+1), & i < \frac{n}{2}, \\ 2\frac{h_n - h_i}{n-i} - \frac{1}{i^2}, & i = \frac{n}{2}, \\ \beta_n(i) - \frac{1}{i^2}, & i > \frac{n}{2}, \end{cases} \quad \textit{für } i > j \textit{ ist} \quad \sigma_{ij} = \begin{cases} \frac{\beta_n(i+1) - \beta_n(i)}{2}, & i + j < n, \\ \frac{h_n - h_i}{n-i} + \frac{h_n - h_j}{n-j} & \\ \quad - \frac{\beta_n(i) + \beta_n(j+1)}{2} - \frac{1}{ij}, & i + j = n, \\ \frac{\beta_n(j) - \beta_n(j+1)}{2} - \frac{1}{ij}, & i + j > n \end{cases}$$

(und $\sigma_{ij} = \sigma_{ji}$*).*

Die Diagonaleinträge $\sigma_{i,i}$ (d.h. $\mathrm{Var}_\theta\left[\xi_i^{(n)}\right]$) dominieren die Kovarianzmatrix $(\sigma_{i,j})$: Abb. 2.2 zeigt die Diagonaleinträge $\sigma_{i,i}$ und die Antidiagonaleinträge $\sigma_{i,n-i}$ für $n = 25$ und $\theta = 1$, Abb. 2.3 zeigt eine dreidimensionale Darstellung von $(\sigma_{i,j})$ für $n = 25$ und $\theta = 1$, Abb. 2.3 zeigt $(-\sigma_{i,j})$, wobei der besseren Sichtbarkeit wegen die (auch betragsmäßig) deutlich größeren Diagonal- und Antidiagonaleinträge auf 0 gesetzt wurden.

Beweis (der Formel für den Erwartungswert). Wir denken uns die Kanten des n-Koaleszenten auf jedem Niveau (u.a. zufällig) nummeriert.

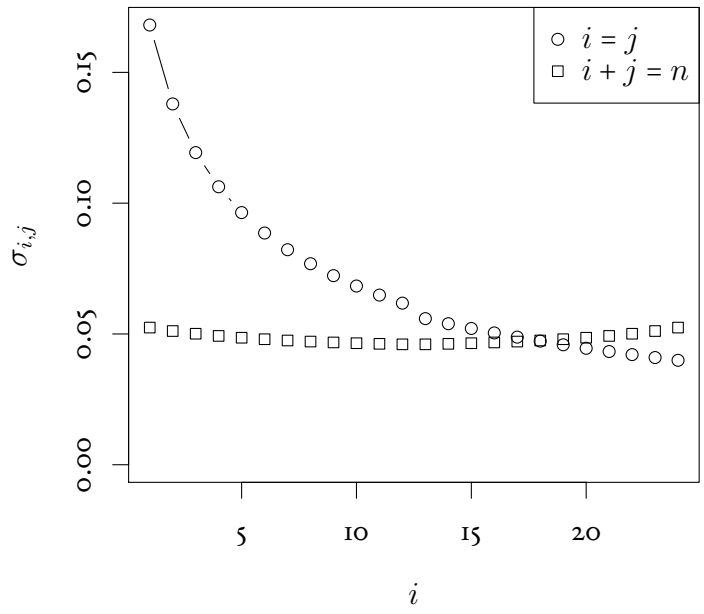


Abbildung 2.2: Diagonaleinträge $\sigma_{i,i}$ und Antidiagonaleinträge $\sigma_{i,n-i}$ der Kovarianzmatrix von $\xi^{(n)}$ für $n = 25$, $\theta = 1$

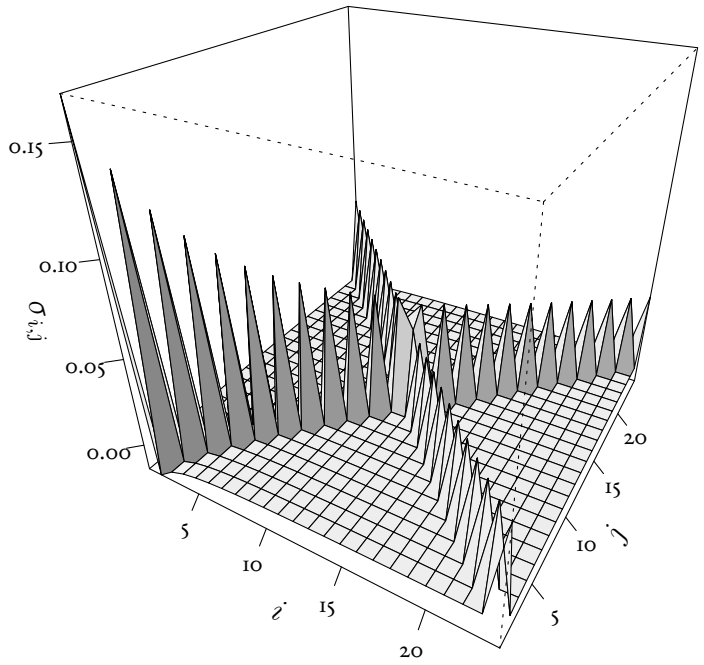


Abbildung 2.3: Kovarianzmatrix von $\xi^{(n)}$ für $n = 25$, $\theta = 1$

Sei

$$\nu_{k,\ell} := \#\,\text{Mutationen auf } \ell\text{-ter Kante auf Niveau } k,$$

$$J_{k,\ell} := \#\,\text{Blätter oberhalb } \ell\text{-ter Kante auf Niveau } k,$$

damit ist

$$\xi_i^{(n)} = \sum_{k=2}^{n-i+1} \sum_{\ell=1}^{k} \nu_{k,\ell} \mathbf{1}(J_{k,\ell} = i). \tag{2.21}$$

Somit gilt

$$\begin{aligned}
\mathbb{E}_\theta\big[\xi_i^{(n)}\big] &= \sum_{k=2}^{n-i+1} \sum_{\ell=1}^{k} \mathbb{E}_\theta\big[\nu_{k,\ell}\mathbf{1}(J_{k,\ell} = i)\big] = \sum_{k=2}^{n-i+1} \sum_{\ell=1}^{k} \mathbb{E}_\theta\big[\nu_{k,\ell}\big]\mathbb{P}_\theta(J_{k,\ell} = i) \\
&= \sum_{k=2}^{n-i+1} \sum_{\ell=1}^{k} \frac{\theta}{k(k-1)} \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} = \sum_{k=2}^{n-i+1} k \frac{\theta}{k(k-1)} \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}} \\
&= \theta \sum_{k=2}^{n-i+1} \frac{1}{k-1} \frac{(n-i-1)!}{(k-2)!(n-i-k+1)!} \frac{(k-1)!(n-k)!}{(n-1)!} \times \frac{i!}{i(i-1)!} \\
&= \frac{\theta}{i} \frac{1}{\binom{n-1}{i}} \sum_{k=2}^{n-i+1} \binom{n-k}{i-1} = \frac{\theta}{i}
\end{aligned}$$

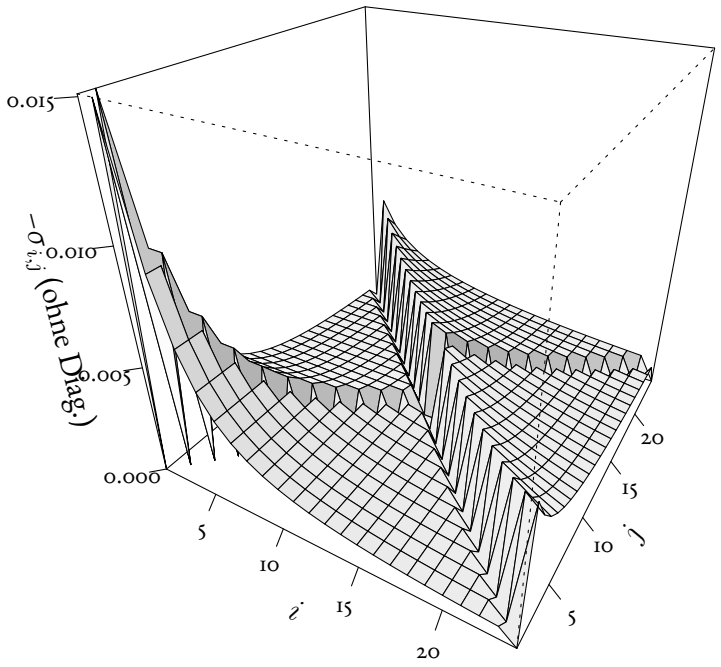


Abbildung 2.4: $(-1)\times$Kovarianzmatrix von $\xi^{(n)}$ für $n = 25$, $\theta = 1$, wobei Diagonal- und Antidiagonaleinträge auf 0 gesetzt wurden

Wir verwenden hierbei in der ersten Zeile, dass

$$\mathbb{E}_\theta\big[\nu_{k,\ell}\big] = \mathbb{E}_\theta\big[\mathbb{E}_\theta\big[\nu_{k,\ell}\,\big|\,T_k\big]\big] = \mathbb{E}_\theta\big[\frac{\theta}{2}T_k\big] = \frac{\theta}{2}\frac{2}{k(k-1)} = \frac{\theta}{k(k-1)}$$

gilt und dass $\nu_{k,\ell}$ (das nur vom Poissonprozess der Mutationen abhängt) und $J_{k,\ell}$ (das nur die Kombinatorik der Abstammungsverhältnisse widerspiegelt) unabhängig sind.

In der zweiten Zeile ersetzen wir

$$\mathbb{P}_\theta(J_{k,\ell} = i) = \frac{\binom{n-i-1}{k-2}}{\binom{n-1}{k-1}},$$

denn mit Korollar A.4 ist die Aufteilung in k (zufällig nummerierte) Familiengrößen uniform auf allen $\big\{(m_1,\dots,m_i) \in \mathbb{N}^i : m_1 + \cdots + m_i = n\big\}$: Es gibt $\binom{n-1}{k-1}$ viele Möglichkeiten, bei $\binom{n-i-1}{k-2}$ davon ist $J_{k,\ell} = i$.

Schließlich beachte in der letzten Zeile $\sum_{k=2}^{n-i+1}\binom{n-k}{i-1} = \binom{n-1}{i}$, denn es gibt $\binom{n-1-(k-1)}{i-1} = \binom{n-k}{i-1}$ viele Teilmengen von $\{1,\dots,n-1\}$ der Größe i, deren kleinstes Element $k-1$ ist, und insgesamt $\binom{n-1}{i}$ Teilmengen von $\{1,\dots,n-1\}$ der Größe i.

Um $\mathbb{E}_\theta[\xi_i^{(n)}\xi_j^{(n)}]$ zu bestimmen kann man die Darstellungen (2.21) für i und für j miteinander multiplizieren und erhält analog zu oben eine Darstellung via eine Doppelsumme über Paare von Kanten im Koaleszenten-Baum. Mittels einer Verfeinerung von Korollar A.4 kann man den kombinatorischen Ausdruck $\mathbb{P}_\theta(J_{k,\ell} = i, J_{k',\ell'} = j)$ bestimmen (man unterscheidet verschiedene Fälle, je nachdem ob die betrachtete Kante (k',ℓ') ein Nachfahre der Kante (k,ℓ) im Baum ist oder nicht) und erhält nach recht umfangreichen Umformungen die oben angegebenen Ausdrücke für $\mathrm{Cov}_\theta\big[\xi_i^{(n)}, \xi_j^{(n)}\big]$, für Details siehe den Artikel von Yun-Xin Fu, Statistical Properties of Segregating Sites, *Theor. Pop. Biol.* 48, 172–197 (1995), in dem dieser Satz bewiesen wurde. □

Tajimas[7] Test

Betrachte eine n-Stichprobe (im IMS-Modell), für $1 \le i < j \le n$ sei

$$\Delta_{i,j} := \text{Anzahl segregierende Stellen, an denen sich Stichproben } i \text{ und } j \text{ unterscheiden.}$$

Die mittlere Anzahl paarweiser Unterschiede,

$$\widehat{\theta}_\pi := \frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n}\Delta_{ij}$$

(„Tajimas $\widehat{\theta}_\pi$“), ist ein (auf den beobachteten Sequenzen basierender) Schätzer für die Mutationsrate θ.

Beobachtung 2.25. Es gebe s segregierende Stellen.

$$\widehat{\theta}_\pi = \frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n}\sum_{m=1}^{s}\mathbf{1}(\text{Stichpr. } i \text{ und } j \text{ unterschiedl. an } m\text{-ter segr. Stelle}) = \frac{1}{\binom{n}{2}}\sum_{k=1}^{n-1}\xi_k^{(n)}\,k(n-k)$$

[7] Fumio Tajima, Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism, *Genetics* 123, 585–595, (1989)

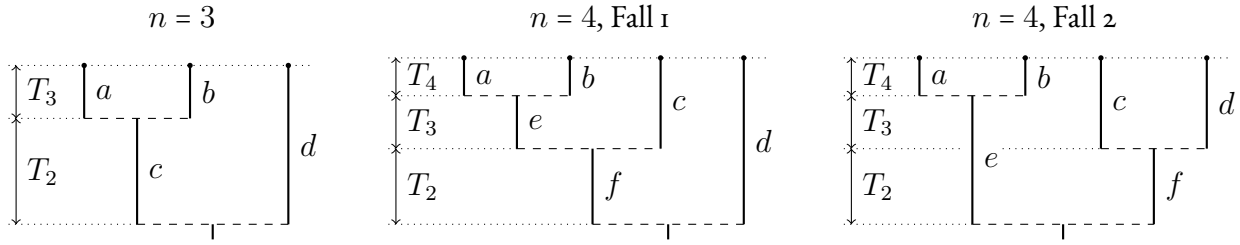


Abbildung 2.5: Formen eines n-Koaleszenten für $n = 3$ (bis auf Umnummerierung der Bläter nur eine Möglichkeit) und $n = 4$ (zwei Möglichkeiten)

(mit $\xi_k^{(n)}$ = # Mut., die in k Stichpr. vorkommen, aus Def. 2.23), d.h. $\widehat{\theta}_\pi$ ist eine (lineare) Funktion des Frequenzspektrums.

(Darüberhinaus kann $\widehat{\theta}_\pi$ ebenso wie S_n und $\widehat{\theta}_W$ als Funktion des gefalteten Frequenzspektrums aufgefasst werden, d.h. wir können dies auch dann bestimmen, wenn wir die anzestralen Typen nicht kennen.)

Proposition 2.26. *Es gilt*

$$\mathbb{E}_\theta\left[\widehat{\theta}_\pi\right] = \theta, \quad \operatorname{Var}_\theta\left(\widehat{\theta}_\pi\right) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2.$$

Insbesondere: $\widehat{\theta}_\pi$ ist erwartungstreuer Schätzer für θ, allerdings ist es nicht konsistent:

$$\lim_{n\to\infty} \operatorname{Var}_\theta\left(\widehat{\theta}_\pi\right) = \tfrac{1}{3}\theta + \tfrac{2}{9}\theta^2 > 0.$$

Bemerkung. $\widehat{\theta}_\pi = \frac{1}{\binom{n}{2}} \sum_{1\le i<j\le n} \Delta_{ij}$ wird in der Literatur auch mit π bezeichnet und die (empirische) „Nukleotid-Diversität“ (engl. “nucleotide diversity”) genannt.

($\mathbb{E}_\theta[\pi] = \theta$ ist einer der Gründe für die Parametrisierung, dass Mutationen mit Rate $\theta/2$ längs der Genealogie erscheinen.)

Beweis. Betrachte zunächst eine Stichprobe der Größe $n = 2$: Es ist

$$\mathbb{E}_\theta[\Delta_{1,2}] = \mathbb{E}_\theta\big[\mathbb{E}_\theta[\Delta_{1,2}\,|\,T_2]\big] = \mathbb{E}_\theta[\theta T_2] = \theta\mathbb{E}_\theta[T_2] = \theta$$

(mit T_2 = Zeit, währenddessen die Genealogie aus 2 Linien besteht = Zeit bis zum jgV der beiden Stichproben, $T_2 \sim \operatorname{Exp}(1)$) und

$$\begin{aligned}\operatorname{Var}_\theta[\Delta_{1,2}] &= \operatorname{Var}_\theta\big[\mathbb{E}_\theta[\Delta_{1,2}\,|\,T_2]\big] + \mathbb{E}_\theta\big[\operatorname{Var}_\theta[\Delta_{1,2}\,|\,T_2]\big] \\ &= \operatorname{Var}_\theta[\theta T_2] + \mathbb{E}_\theta[\theta T_2] = \theta^2\operatorname{Var}_\theta[T_2] + \theta\mathbb{E}_\theta[T_2] = \theta^2 + \theta.\end{aligned}$$

Betrachte nun eine Stichprobe der Größe $n = 3$, es bezeichne η_a die Anzahl Mutationen auf Kante a, etc., siehe Abbildung 2.5. Jede Kante kommt in 2 von 3 paarweisen Vergleichen vor, also ist

$$\widehat{\theta}_{\pi,n=3} = \frac{1}{3}\big(\Delta_{1,2} + \Delta_{1,3} + \Delta_{2,3}\big) = \frac{2}{3}\big(\eta_a + \eta_b + \eta_c + \eta_d\big).$$

Sei T_j die Länge der Zeitspanne, währenddessen der Koaleszent aus j Linien besteht. Nach Definition sind $\eta_a, \eta_b, \eta_c, \eta_d$ unabhängig, gegeben T_3 und T_2, und $\eta_a, \eta_b \sim \mathrm{Poi}(\frac{\theta}{2}T_3), \eta_c \sim \mathrm{Poi}(\frac{\theta}{2}T_2), \eta_d \sim \mathrm{Poi}(\frac{\theta}{2}(T_3+T_2))$, d.h. $\eta_a+\eta_b+\eta_c+\eta_d \sim \mathrm{Pois}(\theta L_3/2)$, wobei $L_3 = 3T_3 + 2T_2$ die Gesamtlänge des Baums ist. Daher ist

$$\begin{aligned}
\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=3}\big] &= \frac{4}{9}\mathrm{Var}_\theta\big[\eta_a+\eta_b+\eta_c+\eta_d\big] \\
&= \frac{4}{9}\mathrm{Var}_\theta\big[\mathbb{E}_\theta[\eta_a+\eta_b+\eta_c+\eta_d \,|\, L_3]\big] + \frac{4}{9}\mathbb{E}_\theta\big[\mathrm{Var}_\theta[\eta_a+\eta_b+\eta_c+\eta_d \,|\, L_3]\big] \\
&= \frac{4}{9}\mathrm{Var}_\theta\Big[\frac{\theta}{2}L_3\Big] + \frac{4}{9}\mathbb{E}_\theta\Big[\frac{\theta}{2}L_3\Big] = \frac{4}{9}\frac{\theta^2}{4}\Big(9\cdot\frac{1}{3^2}+4\cdot 1\Big) + \frac{4}{9}\frac{\theta}{2}\Big(3\cdot\frac{1}{3}+2\cdot 1\Big) = \frac{5}{9}\theta^2+\frac{2}{3}\theta.
\end{aligned}$$

Andererseits ist wegen der Symmetrien der Verteilung des Koaleszenten $\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{1,3}] = \mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{2,3}]$, etc. und somit

$$\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=3}\big] = \frac{1}{9}\cdot 3\cdot\mathrm{Var}_\theta[\Delta_{1,2}] + \frac{1}{9}\cdot 6\cdot\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{1,3}] = \frac{1}{3}(\theta^2+\theta) + \frac{2}{3}\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{1,3}],$$

folglich

$$\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{1,3}] = \frac{3}{2}\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=3}\big] - \frac{1}{2}(\theta^2+\theta) = \frac{1}{3}\theta^2+\frac{1}{2}\theta. \tag{2.22}$$

Betrachte nun eine Stichprobe der Größe $n = 4$: Es gibt 2 mögliche Baumtopologien (siehe Abb. 2.5).

Wir untersuchen zunächst Fall 1 (das mittlere Diagramm in Abb. 2.5). Gegeben T_4, T_3, T_2 sind hier $\eta_a \sim \mathrm{Poi}(\frac{\theta}{2}T_4), \eta_b \sim \mathrm{Poi}(\frac{\theta}{2}T_4), \eta_c \sim \mathrm{Poi}(\frac{\theta}{2}(T_4+T_3)), \eta_d \sim \mathrm{Poi}(\frac{\theta}{2}(T_4+T_3+T_2)), \eta_e \sim \mathrm{Poi}(\frac{\theta}{2}T_3), \eta_f \sim \mathrm{Poi}(\frac{\theta}{2}T_2)$ und unabhängig. Weiter ist in diesem Fall

$$\widehat{\theta}_{\pi,n=4} = \Delta(1) = \frac{1}{\binom{4}{2}}\big(3\eta_a+3\eta_b+3\eta_c+3\eta_d+4\eta_e+3\eta_f\big) =: \frac{1}{6}X_1$$

(beachte: wenn oberhalb einer Kante ℓ Blätter liegen, so tritt sie in $\ell\cdot(n-\ell)$ paarweisen Vergleichen auf), somit ist

$$\begin{aligned}
\mathbb{E}_\theta\big[\widehat{\theta}_{\pi,n=4}\,\big|\,\text{Top.=1}\big] &= \frac{1}{6}\frac{\theta}{2}\mathbb{E}\big[3T_4+3T_4+3(T_4+T_3)+3(T_4+T_3+T_2)+4T_3+3T_2\big] \\
&= \frac{\theta}{12}\Big(\frac{3}{\binom{4}{2}}+\frac{3}{\binom{4}{2}}+3\Big(\frac{1}{\binom{4}{2}}+\frac{1}{\binom{3}{2}}\Big)+\frac{4}{\binom{3}{2}}+3\Big(\frac{1}{\binom{4}{2}}+\frac{1}{\binom{3}{2}}+1\Big)+3\cdot 1\Big) = \frac{17}{18}\theta, \\
\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=4}\,\big|\,\text{Top.=1}\big] &= \frac{1}{36}\mathbb{E}_\theta\big[\mathrm{Var}_\theta[X_1\,|\,T_4,T_3,T_2]\big] + \frac{1}{36}\mathrm{Var}_\theta\big[\mathbb{E}_\theta[X_1\,|\,T_4,T_3,T_2]\big] \\
&= \frac{1}{36}\mathbb{E}_\theta\big[\tfrac{\theta}{2}(9T_4+9T_4+9(T_3+T_4)+9(T_4+T_3+T_2)+16T_3+9T_2)\big] \\
&\quad + \frac{1}{36}\mathrm{Var}_\theta\Big[\frac{\theta}{2}(3T_4+3T_4+3(T_4+T_3)+3(T_4+T_3+T_2)+4T_3+3T_2)\Big] \\
&= \frac{\theta}{72}\Big(\tfrac{9}{6}+\tfrac{9}{6}+9(\tfrac{1}{6}+\tfrac{1}{3})+9(\tfrac{1}{6}+\tfrac{1}{3}+1)+\tfrac{16}{3}+9\cdot 1\Big) \\
&\quad + \frac{\theta^2}{144}\mathrm{Var}_\theta\big[12T_4+10T_3+6T_2\big] \\
&= \frac{53}{108}\theta + \frac{\theta^2}{144}\Big(\tfrac{12^2}{6^2}+\tfrac{10^2}{3^2}+\tfrac{6^2}{1^2}\Big) = \frac{53}{108}\theta+\frac{115}{324}\theta^2.
\end{aligned}$$

Untersuchen wir nun Fall 2 (das rechte Diagramm in Abb. 2.5). Gegeben T_4, T_3, T_2 sind hier $\eta_a \sim \mathrm{Poi}(\frac{\theta}{2}T_4), \eta_b \sim \mathrm{Poi}(\frac{\theta}{2}T_4), \eta_c \sim \mathrm{Poi}(\frac{\theta}{2}(T_4+T_3)), \eta_d \sim \mathrm{Poi}(\frac{\theta}{2}(T_4+T_3)), \eta_e \sim \mathrm{Poi}(\frac{\theta}{2}(T_3+T_2)), \eta_f \sim \mathrm{Poi}(\frac{\theta}{2}T_2)$ und unabhängig, weiter ist in diesem Fall (mit Argumentation analog zu Fall 1)

$$\widehat{\theta}_{\pi,n=4} = \Delta(2) = \frac{1}{\binom{4}{2}}\big(3\eta_a + 3\eta_b + 3\eta_c + 3\eta_d + 4\eta_e + 4\eta_f\big) =: \frac{1}{6}X_2,$$

somit ergibt sich

$$\begin{aligned}
\mathbb{E}_\theta\big[\widehat{\theta}_{\pi,n=4}\,\big|\,\text{Top.=2}\big] &= \frac{1}{6}\frac{\theta}{2}\mathbb{E}\big[3T_4 + 3T_4 + 3(T_4+T_3) + 3(T_4+T_3) + 4(T_3+T_2) + 4T_2\big] \\
&= \frac{\theta}{12}\mathbb{E}\big[12T_4 + 10T_3 + 8T_2\big] = \frac{\theta}{12}\big(\tfrac{12}{6} + \tfrac{10}{3} + 8\big) = \frac{10}{9}\theta, \\
\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=4}\,\big|\,\text{Top.=2}\big] &= \frac{1}{36}\mathbb{E}_\theta\big[\mathrm{Var}_\theta[X_2\,|\,T_4,T_3,T_2]\big] + \frac{1}{36}\mathrm{Var}_\theta\big[\mathbb{E}_\theta[X_2\,|\,T_4,T_3,T_2]\big] \\
&= \frac{1}{36}\mathbb{E}_\theta\big[\frac{\theta}{2}(9T_4 + 9T_4 + 9(T_4+T_3) + 9(T_4+T_3) + 16(T_3+T_2) + 16T_2)\big] \\
&\quad + \frac{1}{36}\mathrm{Var}_\theta\big[\frac{\theta}{2}(3T_4 + 3T_4 + 3(T_4+T_3) + 3(T_4+T_3) + 4(T_3+T_2) + 4T_2)\big] \\
&= \frac{\theta}{72}\mathbb{E}\big[36T_4 + 34T_3 + 32T_2\big] + \frac{\theta^2}{144}\mathrm{Var}_\theta\big[12T_4 + 10T_3 + 8T_2\big] \\
&= \frac{\theta}{72}\big(\tfrac{36}{6} + \tfrac{34}{3} + 32\cdot 1\big) + \frac{\theta^2}{144}\big(\tfrac{12^2}{6^2} + \tfrac{10^2}{3^2} + \tfrac{8^2}{1^2}\big) = \frac{37}{54}\theta + \frac{89}{162}\theta^2.
\end{aligned}$$

Insgesamt ist mit $\mathbb{P}(\text{Top.=1}) = 2/3 = 1 - \mathbb{P}(\text{Top.=2})$ (denn damit der 2. Fall für die Baumtopologie eintritt, muss die zweitjüngste Verschmelzung das Paar von Linien betreffen, das bis dahin noch an keiner Verschmelzung teilgenommen hat, dies ist dann 1 von 3 Möglichkeiten)

$$\mathbb{E}_\theta\big[\widehat{\theta}_{\pi,n=4}\big] = \frac{2}{3}\cdot\frac{17}{18}\theta + \frac{1}{3}\cdot\frac{10}{9}\theta = \theta$$

und

$$\begin{aligned}
\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=4}\big] &= \mathbb{E}_\theta\big[\mathrm{Var}_\theta[\widehat{\theta}_{\pi,n=4}\,|\,\text{Top.}]\big] \;+\; \mathrm{Var}_\theta\big[\mathbb{E}_\theta[\widehat{\theta}_{\pi,n=4}\,|\,\text{Top.}]\big] \\
&= \frac{2}{3}\big(\frac{53}{108}\theta + \frac{115}{324}\theta^2\big) + \frac{1}{3}\big(\frac{37}{54}\theta + \frac{89}{162}\theta^2\big) \;+\; \frac{2}{3}\big(\frac{17}{18}\theta - \theta\big)^2 + \frac{1}{3}\big(\frac{10}{9}\theta - \theta\big)^2 \\
&= \frac{23}{54}\theta^2 + \frac{5}{9}\theta.
\end{aligned}$$

Andererseits ist wie oben wegen der Symmetrien der Verteilung des Koaleszenten

$$\begin{aligned}
\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=4}\big] &= \frac{1}{36}\mathrm{Cov}_\theta\Big[\sum_{1\le i<j\le 4}\Delta_{i,j}\,,\,\sum_{1\le k<\ell\le 4}\Delta_{k,\ell}\Big] \\
&= \frac{1}{36}\Big(6\mathrm{Var}_\theta[\Delta_{1,2}] + 6\cdot 2\cdot 2\,\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{1,3}] + 6\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{3,4}]\Big) \\
&= \frac{1}{6}(\theta^2+\theta) + \frac{2}{3}\big(\frac{1}{3}\theta^2 + \frac{1}{2}\theta\big) + \frac{1}{6}\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{3,4}]
\end{aligned}$$

und somit

$$\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{3,4}] = 6\mathrm{Var}_\theta\big[\widehat{\theta}_{\pi,n=4}\big] - (\theta^2+\theta) - 4\Big(\frac{1}{3}\theta^2+\frac{1}{2}\theta\Big) = \frac{2}{9}\theta^2+\frac{1}{3}\theta. \tag{2.23}$$

Schließlich betrachten wir den allgemeinen Fall einer Stichprobe der Größe n:

$$\begin{aligned}
\mathbb{E}_\theta\big[\widehat{\theta}_\pi\big] &= \frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n}\mathbb{E}_\theta[\Delta_{i,j}] = \frac{1}{\binom{n}{2}}\binom{n}{2}\mathbb{E}_\theta[\Delta_{1,2}] = \theta,\\
\mathrm{Var}_\theta\big[\widehat{\theta}_\pi\big] &= \frac{1}{\binom{\binom{n}{2}}{2}}\mathrm{Cov}_\theta\Big[\sum_{1\le i<j\le n}\Delta_{i,j}\,,\sum_{1\le k<\ell\le n}\Delta_{k,\ell}\Big]\\
&= \frac{1}{\binom{\binom{n}{2}}{2}}\Big(\tbinom{n}{2}\mathrm{Var}_\theta[\Delta_{1,2}] + \tbinom{n}{2}2(n-2)\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{1,3}]\\
&\qquad\qquad + \tbinom{n}{2}\tbinom{n-2}{2}\mathrm{Cov}_\theta[\Delta_{1,2},\Delta_{3,4}]\Big)\\
&= \frac{1}{\binom{n}{2}}\Big(\theta^2+\theta+2(n-2)\Big(\frac{1}{3}\theta^2+\frac{1}{2}\theta\Big)+\tbinom{n-2}{2}\Big(\frac{2}{9}\theta^2+\frac{1}{3}\theta\Big)\Big)\\
&= \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2.
\end{aligned}$$

□

Passen beobachtete Sequenzdaten zum Modell? Unser wahrscheinlichkeitstheoretisches Modell beschreibt die Verteilung von n beobachteten Sequenzen, die wir an den Blättern eines Kingman-n-Koaleszenten ablesen, auf dessen Kanten gemäß einem Poissonprozess mit einer gewissen Rate $\theta/2$ Mutationen liegen, die den Typ jeweils gemäß dem IMS-Modell ändern. Angesichts Satz 1.5 ist die biologische Interpretation, dass wir n Stichproben aus einer „panmiktischen" Population konstanter Größe sehen und dass die genetische Variabilität (am betrachteten Ort im Genom) „neutral" ist (und dass die Annahmen des IMS-Modells wenigstens approximativ zutreffen).

Die Tatsache, dass sowohl $\widehat{\theta}_W$ als auch $\widehat{\theta}_\pi$ in diesem Modell erwartungstreue Schätzer für (das unbekannte) θ sind, gestattet es, für die Nullhypothese „das Modell beschreibt die Daten zutreffend" einen statistischen Test zu formulieren. Wenn das Modell zutrifft, sollte nämlich

$$\widehat{\theta}_\pi - \widehat{\theta}_W \approx 0$$

bis auf „zufällige Fluktuationen" gelten. Diese Idee geht auf F. Tajima zurück, siehe den in Fußnote 7 auf S. 46 zitierten Artikel.

Um einzuschätzen, wie groß die „typischen" Fluktuationen sind, sollten wir (zumindest) $\mathrm{Var}_\theta\left[\widehat{\theta}_\pi - \widehat{\theta}_W\right]$ bestimmen können.

Bericht 2.27. Es gilt $\mathrm{Cov}_\theta\left[S_n, \widehat{\theta}_\pi\right] = \theta + \left(\frac{1}{2}+\frac{1}{n}\right)\theta^2$, also $\mathrm{Cov}_\theta\left[\widehat{\theta}_W, \widehat{\theta}_\pi\right] = \frac{\theta}{h_n} + \left(\frac{1}{2}+\frac{1}{n}\right)\frac{\theta^2}{h_n}$ und somit

$$\mathrm{Var}_\theta\left[\widehat{\theta}_\pi - \widehat{\theta}_W\right] = \left(\tfrac{n+1}{3(n-1)} - \tfrac{1}{h_n}\right)\theta + \left(\tfrac{2(n^2+n+3)}{9n(n-1)} - \tfrac{n+2}{nh_n} + \tfrac{g_n}{h_n^2}\right)\theta^2$$

(mit $h_n = \sum_{i=1}^{n-1} \frac{1}{i}$, $g_n = \sum_{i=1}^{n-1} \frac{1}{i^2}$).

Weiterhin ist

$$\widehat{V} := \alpha_1 S_n + \alpha_2 S_n(S_n - 1) \tag{2.24}$$

mit

$$\alpha_1 = \left(\tfrac{n+1}{3(n-1)} - \tfrac{1}{h_n}\right)/h_n, \quad \alpha_2 = \left(\tfrac{2(n^2+n+3)}{9n(n-1)} - \tfrac{n+2}{nh_n} + \tfrac{g_n}{h_n^2}\right)\Big/(h_n^2 + g_n) \tag{2.25}$$

ein erwartungstreuer Schätzer für $\mathrm{Var}_\theta\left[\widehat{\theta}_\pi - \widehat{\theta}_W\right]$.

Die Formel für $\mathrm{Cov}_\theta[S_n, \widehat{\theta}_\pi]$ kann man mittels einer ähnlichen Zerlegung wie im Beweis von Proposition 2.26 beweisen, siehe F. Tajima, a.a.O. Zusammen mit Beobachtung 2.20 und Proposition 2.26 ergibt sich daraus die Formel für $\mathrm{Var}_\theta[\widehat{\theta}_\pi - \widehat{\theta}_W]$.

Aus Beobachtung 2.20 folgt auch

$$\mathbb{E}_\theta[S_n] = \theta h_n \quad \text{und} \quad \mathbb{E}_\theta[S_n(S_n-1)] = \mathrm{Var}_\theta[S_n] + \big(\mathbb{E}_\theta[S_n]\big)^2 - \mathbb{E}_\theta[S_n] = \theta^2 h_n + \theta^2 g_n,$$

d.h. $\mathbb{E}_\theta[\widehat{V}] = \mathrm{Var}_\theta[\widehat{\theta}_\pi - \widehat{\theta}_W]$.

Definition 2.28 (Tajimas D)**.** $D := \dfrac{\widehat{\theta}_\pi - \widehat{\theta}_W}{\sqrt{\widehat{V}}}$ mit $\widehat{V}$ aus (2.24) heißt Tajimas D.

Die Teststatistik D erfüllt $\mathbb{E}_\theta[D] \approx 0$, $\mathrm{Var}_\theta(D) \approx 1$ (die Erwartung ist nicht exakt $= 0$, da Zähler und Nenner nicht unabhängig sind, die Varianz ist nicht exakt $= 1$, da $\widehat{V}$ nur ein Schätzer für die Varianz des Zählers ist). Die Formulierung ist (beispielsweise) durch den klassischen t-Test inspiriert: Dort normiert man einen empirischen Mittelwert von n Beobachtungswerten mit dem Standardfehler, einem Schätzer für die Streuung.

Um anhand von D einen statistischen Test zu formulieren, benötigen wir (für ein vorgegebenes Signifikanzniveau α) sogenannte kritische Werte, d.h. geeignete Quantile von D unter der Nullhypothese.

Auf dem Ereignis $\{S_n = s\}$ gilt

$$\widehat{\theta}_W = s/h_n, \quad \widehat{V} = \alpha_1 s + \alpha_2 s(s-1),$$

der kleinste möglicher Wert von $\widehat{\theta}_\pi$ ist dann

$$\frac{1}{\binom{n}{2}} s(n-1) = \frac{2s}{n}.$$

Dies geschieht, wenn $\xi_1^{(n)} + \xi_{n-1}^{(n)} = n$, $\xi_i^{(n)} = 0$ für $2 \le i \le n-2$ gilt (insbesondere, wenn alle Mutationen auf sogenannten externen Kanten – die direkt zu einem Blatt führen – liegen). Der kleinste mögliche Wert von D ist dann somit

$$\frac{2s/n - s/h_n}{\sqrt{\alpha_1 s + \alpha_2 s(s-1)}} \xrightarrow[s\to\infty]{} \frac{2/n - 1/h_n}{\sqrt{\alpha_2}} =: d_{\min} \quad \big(= d_{\min}(n)\big).$$

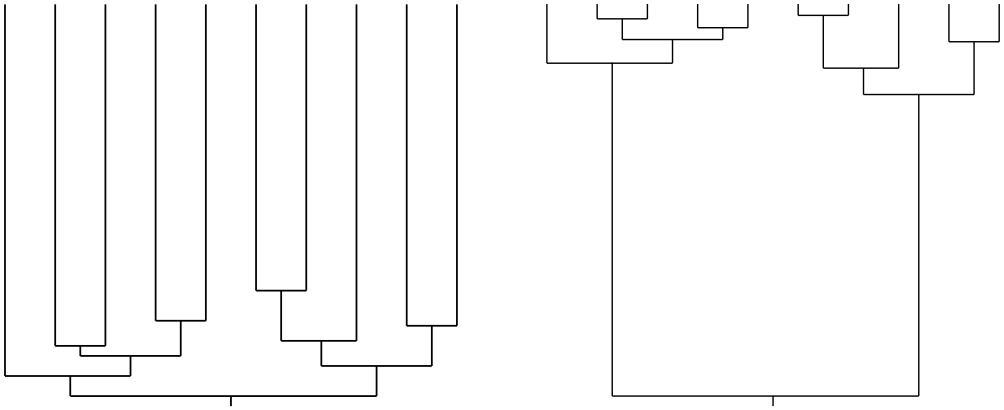


Abbildung 2.6: Ein „sternförmiger“ (links) und ein „Hühnerbein“- (rechts) Koaleszentenbaum

Während ein so kleiner Wert von D unter dem Modell, in dem Mutationen auf den Kingman-Koaleszenten fallen, eher untypisch ist, wäre dies ist in einer „sternförmigen“ Genealogie, in der die externen Äste die Gesamtlänge des Baumes dominieren (siehe Abbildung 2.6), typisch.

Andererseits ist auf $\{S_n = s\}$ der größte mögliche Wert von $\widehat{\theta}_\pi$

$$\frac{1}{\binom{n}{2}} s\lceil n/2\rceil\lfloor n/2\rfloor = 2s\frac{\lceil n/2\rceil\lfloor n/2\rfloor}{n(n-1)}.$$

Dies geschieht, wenn $\xi^{(n)}_{\lceil n/2\rceil} = n$, $\xi^{(n)}_i = 0$ für $i \neq \lceil n/2\rceil$ (d.h. wenn alle Mutationen auf sehr „balanzierten“ Kanten liegen, die die Blätter in genau zwei Hälften teilen). Der größte mögliche Wert von D ist dann

$$\frac{\frac{2s\lceil n/2\rceil\lfloor n/2\rfloor}{n(n-1)} - s/h_n}{\sqrt{\alpha_1 s + \alpha_2 s(s-1)}} \xrightarrow[s\to\infty]{} \frac{\frac{2\lceil n/2\rceil\lfloor n/2\rfloor}{n(n-1)} - 1/h_n}{\sqrt{\alpha_2}} =: d_{\max} \ \big(= d_{\max}(n)\big)$$

Während ein so kleiner Wert von D unter dem Kingman-Koaleszenten untypisch wäre, wäre dies ist in einer (sehr balanzierten) “Hühnerbein-artigen”-Genealogie, in der zwei innere Äste die Gesamtlänge des Baums dominieren (siehe Abbildung 2.6), typisch.

Im Gegensatz etwa zum klassischen t-Test ist die Verteilung von D unter der Nullhypothese

„die Beobachtungen entstehen durch die Typen an den Blättern eines n-Koaleszenten, längs dessen Kanten sich mit Rate $\theta/2$ Mutationen gemäß IMS-Modell ereignen“ (2.26)

nicht explizit bekannt und hängt von dem unbekannten θ ab.

Tajimas pragmatisch-heuristische Lösung: Approximiere die Verteilung von D durch eine skalierte Beta-Verteilung, so dass der Träger = $[d_{\min}, d_{\max}]$, EW= 0 und Var= 1 gilt (was recht plausibel passt, siehe Abbildung 2.7): Verwende die approximative Dichte

$$f_{\text{appr}}(d) = \frac{\Gamma(u+v)(d-d_{\min})^{u-1}(d_{\max}-d)^{v-1}}{\Gamma(u)\Gamma(v)(d_{\max}-d_{\min})^{u+v-1}}, \quad d_{\min} < d < d_{\max} \tag{2.27}$$

mit

$$u = \frac{(1 + d_{\max} d_{\min}) d_{\min}}{d_{\max} - d_{\min}}, \quad v = -\frac{(1 + d_{\max} d_{\min}) d_{\max}}{d_{\max} - d_{\min}}. \tag{2.28}$$

(beachte $d_{\min} < 0 < d_{\max}$).

Diese Formeln entspringen dem Ansatz

$$D \approx (d_{\max} - d_{\min}) B + d_{\min} \quad \text{mit} \quad B \sim \mathrm{Beta}(u, v).$$

$\mathrm{Beta}(u, v)$ hat EW $\frac{u}{u+v}$ und Var $\frac{uv}{(u+v)^2(u+v+1)}$, aus dem Ansatz und den geforderten Normierungen ergibt sich

$$\begin{aligned}
\frac{u}{u+v} &= \frac{-d_{\min}}{d_{\max} - d_{\min}} \qquad \Longrightarrow \quad v = u\frac{d_{\max} - d_{\min}}{-d_{\min}} - u = u\frac{d_{\max}}{-d_{\min}}, \\
\frac{uv}{(u+v)^2(u+v+1)} &= \frac{u^2 \frac{d_{\max}}{-d_{\min}}}{u^2\big(1 + \frac{d_{\max}}{-d_{\min}}\big)^2\big(u(1 + \frac{d_{\max}}{-d_{\min}}) + 1\big)} = \frac{-d_{\max} d_{\min}}{(d_{\max} - d_{\min})^2\big(u(1 + \frac{d_{\max}}{-d_{\min}}) + 1\big)} \\
&= \frac{1}{(d_{\max} - d_{\min})^2} \\
\Longrightarrow \quad u &= \frac{-d_{\max} d_{\min} - 1}{1 - \frac{d_{\max}}{d_{\min}}} = \frac{d_{\min}(1 + d_{\max} d_{\min})}{d_{\max} - d_{\min}}, \quad v = -\frac{d_{\max}(1 + d_{\max} d_{\min})}{d_{\max} - d_{\min}},
\end{aligned}$$

woraus sich (2.28) ergibt.

Definition 2.29 (Tajimas Test)**.** Sei $\alpha \in (0, 1)$, $q_{\mathrm{Beta}(u,v)}(\alpha/2)$, $q_{\mathrm{Beta}(u,v)}(1 - \alpha/2)$ das $\alpha/2$- bzw. $(1 - \alpha/2)$-Quantil der $\mathrm{Beta}(u, v)$-Verteilung mit angepassten Parametern u, v aus (2.28).

Lehne H_0 : (2.26) ab, wenn

$$\begin{aligned}
&D < (d_{\max} - d_{\min}) q_{\mathrm{Beta}(u,v)}(\alpha/2) + d_{\min} \qquad \text{oder} \\
&D > (d_{\max} - d_{\min}) q_{\mathrm{Beta}(u,v)}(1 - \alpha/2) + d_{\min}.
\end{aligned}$$

Dieser Test hält (zumindest approximativ) das Signifikanzniveau α ein.

Beispiel. Für die Daten aus Bsp. 2.18 ergibt sich $n = 8$, $s = 31$, $\xi_1^{(8)} = 13$, $\xi_2^{(8)} = 1$, $\xi_7^{(8)} = 17$, somit $\widehat{\theta}_\pi \doteq 7.93$, $\widehat{\theta}_W \doteq 11.96$, $D \doteq -1.79$

Tajimas Approximation liefert ein 95%-Konfidenzintervall für D unter dem Standard-Kingman-Koaleszenten von $[-1.663, 1.975]$, d.h. die Abweichung von 0 ist auf dem 5%-Niveau signifikant (s. Tajima, a.a.O., Table 2, S 592).

Diskussion. In der biologischen Interpretation nennt man Tajimas Test gelegentlich etwas salopp einen „Test auf Neutralität“, da die Nullhypothese (2.26) aus einem Modell ohne Selektion stammt.

Signifikante Abweichungen von $D \approx 0$ legen Alternativhypothesen nahe, unter denen der Baum, der die Stichproben verbindet, eher nicht wie ein „typischer“ Koaleszent aussieht.

Ein signifikant negatives $D < 0$ passt eher zu einem Baum, in dem externe Äste dominieren (Abbildung 2.6, links). Biologische Szenarien, in denen solche Genealogien typisch sind, wären beispielsweise gerichtete Selektion am betrachteten Ort im Genom (oder in dessen „Nähe“, ein sogenannter selektiver “sweep”) oder eine stark wachsende Population.

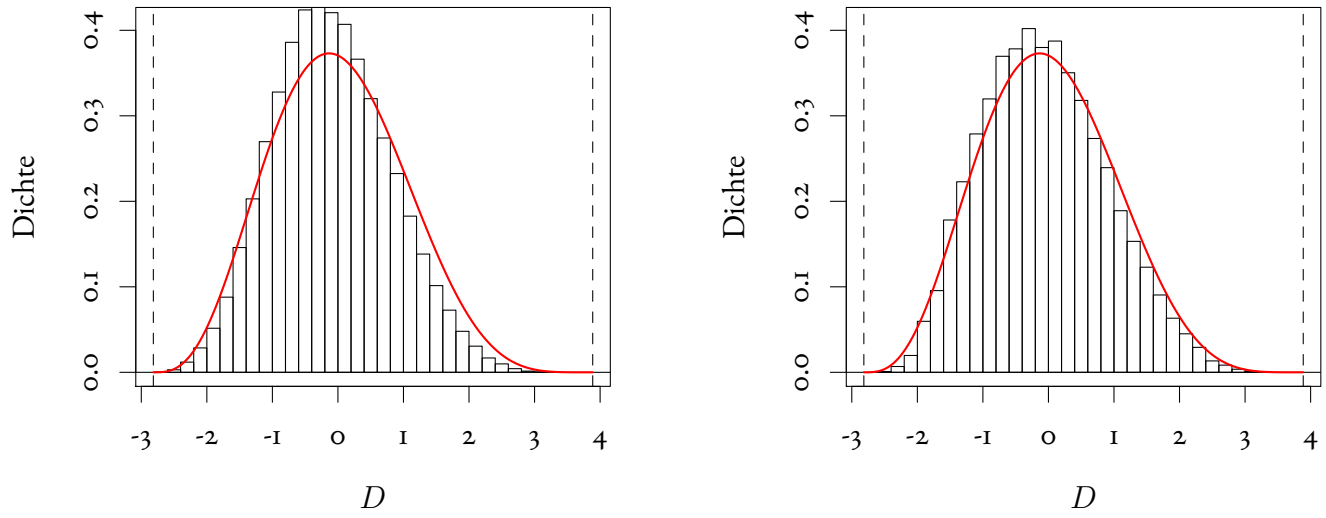


Abbildung 2.7: Simulation der Verteilung von D für $n = 25$ und $\theta = 10$ (links) bzw. $\theta = 2$ (rechts) unter dem Kingman-Koaleszenten mit IMS-Mutationen sowie angepasste skalierte Beta-Dichte aus (2.27). Histogramm jeweils basierend auf 100.000 simulierten Datensätzen.

Ein signifikant positives $D > 0$ passt eher zu einem Baum, in dem wenige interne Äste dominieren (Abbildung 2.6, rechts). Populationsszenarien, in denen solche Genealogien typisch sind, sind beispielsweise (räumlich stark) strukturierte Populationen oder sogenannte balanzierende Selektion (bei der selektive Kräfte gewissermaßen eine genetische Substruktur in der Population aufrecht erhalten).

Eine „exakte“ Version von Tajimas Test

K. L. Simonsen, G. A. Churchill und C. F. Aquadro haben in dem Artikel Properties of statistical tests of neutrality for DNA polymorphism data, *Genetics* 141:413–429, (1995) eine Version von Tajimas Test vorgeschlagen, die ohne die (nicht wörtlich gerechtfertigte) Approximation von D durch eine Beta-Verteilung auskommt[8].

Das unbekannte θ wird dabei als „Störparameter“ (engl. “nuisance parameter”) aufgefasst. Wir wählen $\beta > 0$ (und typischerweise klein) und konstruieren zunächst ein Konfidenzintervall für θ zum Irrtumsniveau β:

Sei für $s \in \mathbb{N}_0$

$$\widehat{\theta}_L(s) = \min\left\{\theta > 0 : \mathbb{P}_\theta(S_n \geq s) > \beta/2\right\}, \quad \widehat{\theta}_R(s) = \max\left\{\theta > 0 : \mathbb{P}_\theta(S_n \leq s) > \beta/2\right\}.$$

[8]Die Konstruktion verwendet ein allgemeines statistisches Prinzip, siehe R. L. Berger und D. D. Boos, P values maximized over a confidence set for the nuisance parameter, *Journal of the American Statistical Association* 89, No. 427, 1012–1016, (1994).

Dies ist mittels der Verteilungsfunktion von S_n unter $\mathbb{P}_\theta$ aus Lemma 2.30 unten zumindest numerisch möglich; da diese als Funktion von θ stetig ist, gilt tatsächlich $\mathbb{P}_{\widehat{\theta}_L(s)}(S_n \geq s) = \beta/2$ und $\mathbb{P}_{\widehat{\theta}_R(s)}(S_n \leq s) = \beta/2$ für $s \in \mathbb{N}_0$.

Da $\theta \mapsto \mathbb{P}_\theta(S_n \geq s)$ monoton wachsend in θ ist, gilt

$$\widehat{\theta}_L(s) > \theta \iff \mathbb{P}_\theta(S_n \geq s) \leq \beta/2 \quad \text{und} \quad \widehat{\theta}_R(s) < \theta \iff \mathbb{P}_\theta(S_n \leq s) \leq \beta/2.$$

Damit gilt

$$\forall\, \theta > 0 \,:\, \mathbb{P}_\theta\big([\widehat{\theta}_L(S_n), \widehat{\theta}_R(S_n)] \ni \theta\big) \geq 1 - \beta,$$

denn für $\theta > 0$ ist

$$\begin{aligned}
&\mathbb{P}_\theta\big([\widehat{\theta}_L(S_n), \widehat{\theta}_R(S_n)] \not\ni \theta\big) = \mathbb{P}_\theta\big(\theta < \widehat{\theta}_L(S_n)\big) + \mathbb{P}_\theta\big(\theta > \widehat{\theta}_R(S_n)\big) \\
&= \mathbb{P}_\theta\big(S_n \in \{s : \theta < \widehat{\theta}_L(s)\}\big) + \mathbb{P}_\theta\big(S_n \in \{s : \theta > \widehat{\theta}_R(s)\}\big) \\
&= \sum_{s\,:\,\mathbb{P}_\theta(S_n \geq s) \leq \beta/2} \mathbb{P}_\theta(S_n = s) + \sum_{s\,:\,\mathbb{P}_\theta(S_n \leq s) \leq \beta/2} \mathbb{P}_\theta(S_n = s) \leq \frac{\beta}{2} + \frac{\beta}{2}.
\end{aligned}$$

Dann bestimmt man bei beobachtetem Wert von S_n (mittels Simulation, für θ-Werte aus einem geeignet feinen Gitter in $[\widehat{\theta}_L(S_n), \widehat{\theta}_R(S_n)]$)

$$\begin{aligned}
D_L^* &= \min\Big\{\frac{\alpha}{2}\text{-Quantil von } \mathscr{L}_\theta(D) \,:\, \theta \in [\widehat{\theta}_L(S_n), \widehat{\theta}_R(S_n)]\Big\}, \\
D_L^* &= \max\Big\{\Big(1 - \frac{\alpha}{2}\Big)\text{-Quantil von } \mathscr{L}_\theta(D) \,:\, \theta \in [\widehat{\theta}_L(S_n), \widehat{\theta}_R(S_n)]\Big\}.
\end{aligned}$$

Somit gilt

$$\forall\, \theta > 0 \,:\, \mathbb{P}_\theta\big(D \notin [D_L^*, D_R^*]\big) \leq \alpha + \beta,$$

d.h. der Test

$$\text{lehne } H_0 : (2.26) \text{ ab, wenn } D < D_L^* \text{ oder } D > D_R^*$$

hält Niveau $\alpha + \beta$ ein (zumindest theoretisch, wenn man die Quantile im 2. Schritt exakt bestimmen könnte).

Beispiel. Für die Daten aus Bsp. 2.18 ($n = 8$, $D \doteq -1.79$) berichten Simonsen, Churchill und Aquadro, a.a.O., Table 3 gemäß diesem Ansatz ein 95%-Konfidenzintervall für D unter dem Standard-Kingman-Koaleszenten von $[-1.80, 1.83]$ (für $n = 10$, $S_n \in [27, 41]$), d.h. die Abweichung ist „gerade so“ nicht signifikant auf dem 5%-Niveau.

Lemma 2.30 (Explizite Verteilung von S_n)**.** *Es gilt*

$$\begin{aligned}
\mathbb{P}_\theta\,(S_n = m) &= \frac{n-1}{\theta} \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-2}{k-1} \Big(\frac{\theta}{\theta + k}\Big)^{m+1}, \quad m \in \mathbb{N}_0, \\
\mathbb{P}_\theta\,(S_n \leq s) &= 1 - \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-1}{k} \Big(\frac{\theta}{\theta + k}\Big)^{s+1}, \quad s \in \mathbb{N}_0.
\end{aligned}$$

Bemerkung. Dies ist eine Version von Lemma A.6 (Dichte der Faltung exponentieller ZVn) für den diskreten Fall (Faltung geometrischer ZVn).

Beweis. $S_n = S_{n,n} + S_{n,n-1} + \cdots + S_{n,2}$ mit $S_{n,j} \sim \operatorname{geom}(\frac{i-1}{\theta+i-1})$ u.a.

Sei $u \in [0,1]$: Es ist

$$\mathbb{E}\left[u^{S_{n,j}}\right] = \sum_{\ell=0}^{\infty} u^\ell \tfrac{j-1}{\theta+j-1}\left(\tfrac{\theta}{\theta+j-1}\right)^\ell = \tfrac{j-1}{\theta+j-1}\frac{1}{1-u\frac{\theta}{\theta+j-1}} = \frac{j-1}{j-1+\theta(1-u)},$$

somit

$$\mathbb{E}\left[u^{S_n}\right] = \prod_{j=2}^{n} \mathbb{E}\left[u^{S_{n,j}}\right] = \prod_{k=1}^{n-1} \frac{k}{k+\theta(1-u)}.$$

Weiter ist

$$\prod_{k=1}^{n-1} \frac{k}{k+z} = \sum_{k=1}^{n-1} \frac{a_{n,k}}{k+z} \quad (z \in \mathbb{C} \setminus -\mathbb{N}) \quad \text{mit } a_{n,k} = \tfrac{(n-1)!}{\prod_{j\neq k}^{n-1}(j-k)} = (-1)^k(n-1)\tbinom{n-2}{k-1},$$

also

$$\mathbb{E}_\theta\left[u^{S_n}\right] = \sum_{m=0}^{\infty} u^m \mathbb{P}_\theta\left(S_n = m\right) = \sum_{k=1}^{n-1} a_{n,k} \sum_{m=0}^{\infty} \left(\frac{\theta}{\theta+k}\right)^m u^m = \sum_{m=0}^{\infty} u^m \sum_{k=1}^{n-1} a_{n,k}\left(\frac{\theta}{\theta+k}\right)^m$$

und

$$\begin{aligned}
\mathbb{P}_\theta\left(S_n \le s\right) &= \sum_{m=0}^{s} \mathbb{P}_\theta\left(S_n = m\right) = \sum_{m=0}^{s} \frac{n-1}{\theta} \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-2}{k-1} \left(\frac{\theta}{\theta+k}\right)^{m+1} \\
&= \frac{n-1}{\theta} \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-2}{k-1} \sum_{m=0}^{s} \left(\frac{\theta}{\theta+k}\right)^{m+1} \\
&= \frac{n-1}{\theta} \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-2}{k-1} \frac{\theta}{\theta+k} \frac{1-\left(\frac{\theta}{\theta+k}\right)^{s+1}}{1-\left(\frac{\theta}{\theta+k}\right)} \\
&= \frac{n-1}{\theta} \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-2}{k-1} \frac{\theta}{k} \left(1-\left(\frac{\theta}{\theta+k}\right)^{s+1}\right) \\
&= \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-1}{k} \left(1-\left(\frac{\theta}{\theta+k}\right)^{s+1}\right) \\
&= 1 - \sum_{k=1}^{n-1} (-1)^{k-1} \binom{n-1}{k} \left(\frac{\theta}{\theta+k}\right)^{s+1}
\end{aligned}$$

(denn $-\sum_{k=1}^{n-1}(-1)^k\binom{n-1}{k} = 1-(1-1)^{n-1} = 1$). □

Literaturverzeichnis

[Bir24] Matthias Birkner, *Einführung in die Stochastik*, `https://www.staff.uni-mainz.de/birkner/GrundlStoch_2324/Stochastik-Einfuehrung_WS23_24.pdf`, 2024, Vorlesungsnotizen, JGU Mainz.

[Geo15] Hans-Otto Georgii, *Stochastik. Einführung in die Wahrscheinlichkeitstheorie und Statistik*, 5th, revised and expanded ed. ed., Berlin: De Gruyter, 2015 (German).

[Kle20] Achim Klenke, *Wahrscheinlichkeitstheorie*, 4th revised and supplemented edition ed., Masterclass, Berlin: Springer Spektrum, 2020 (German).

[KW10] Götz Kersting and Anton Wakolbinger, *Elementare Stochastik.*, 2nd revised ed. ed., Basel: Birkhäuser, 2010 (German).

[Nor97] J. R. Norris, *Markov chains*, Camb. Ser. Stat. Probab. Math., vol. 2, Cambridge: Cambridge University Press, 1997 (English).

[Wak09] John Wakeley, *Coalescent theory: an introduction*, vol. 1, Roberts & Company Publishers Greenwood Village, Colorado, 2009.

Anhang A

Anhang

A.1 Ein Exkurs zum Poissonprozess und zu zeitkontinuierlichen Markovketten

Wir diskutieren hier knapp Poissonprozesse auf $\mathbb{R}_+$ und zeitkontinuierliche Markovketten. Wesentlich mehr dazu findet man z.B. in den Büchern von J. Norris [Nor97] oder von A. Klenke [Kle20, Kap. 17.3].

Sei $c_N \longrightarrow_{N\to\infty} 0$ eine Nullfolge, $\lambda > 0$, $Z_i^{(N)}$, $i \in \mathbb{N}$ u.i.v. $\sim \mathrm{Ber}(\lambda c_N)$ (wir betrachten o.E. nur so große N, dass $\lambda c_N \le 1$), seien

$$T_0^{(N)} := 0, \quad T_\ell^{(N)} := \inf\left\{i > T_{\ell-1}^{(N)} : Z_i^{(N)} = 1\right\},\ \ell \in \mathbb{N}$$

($T_\ell^{(N)}$ ist der Zeitpunkt des ℓ-ten Erfolgs in der Münzwurffolge $(Z_i^{(N)})_{i\in\mathbb{N}}$), dann sind

$$\tau_\ell^{(N)} := T_\ell^{(N)} - T_{\ell-1}^{(N)},\ \ell \in \mathbb{N}$$

u.i.v., $\tau_\ell^{(N)} \sim \mathrm{geom}(\lambda c_N)$, d.h. $\mathbb{P}(\tau_\ell^{(N)} = j) = c_N\lambda(1 - c_N\lambda)^{j-1}$ für $j \in \mathbb{N}$ und für $x \ge 0$ gilt

$$\mathbb{P}(c_N\tau_\ell^{(N)} > x) = \mathbb{P}\big(\tau_\ell^{(N)} > \lfloor\frac{x}{c_N}\rfloor\big) = \big(1 - c_N\lambda\big)^{\lfloor x/c_N\rfloor} \underset{N\to\infty}{\longrightarrow} e^{-\lambda x},$$

d.h. $c_N\tau_\ell^{(N)} \xrightarrow[N\to\infty]{d} \mathrm{Exp}(\lambda)$ (Übung: Beweisen Sie diese Aussagen).

Sei weiter

$$M_k^{(N)} := \big|\{1 \le i \le k : Z_i^{(N)} = 1\}\big| = \max\{\ell \in \mathbb{N}_0 : T_\ell^{(N)} \le k\},$$

offenbar gilt für $0 \le k_0 < k_1 < \cdots < k_m$

$$M_{k_1}^{(N)} - M_{k_0}^{(N)}, M_{k_2}^{(N)} - M_{k_1}^{(N)}, \ldots, M_{k_m}^{(N)} - M_{k_{m-1}}^{(N)} \quad \text{sind unabhängig}$$

und für $0 \le k < k'$ ist $M_{k'}^{(N)} - M_k^{(N)} \sim \mathrm{Bin}(k' - k, c_N\lambda)$, somit gilt für $0 \le t < t'$

$$M_{\lfloor t'/c_N\rfloor}^{(N)} - M_{\lfloor t/c_N\rfloor}^{(N)} \xrightarrow[N\to\infty]{d} \mathrm{Pois}\big(\lambda(t' - t)\big).$$

(Übung: Beweisen Sie diese Aussagen).

Dies lädt ein, folgendes Limesobjekt zu betrachten: Sei $\tau_1, \tau_2, \dots$ u.i.v., $\tau_\ell \sim \mathrm{Exp}(\lambda)$, $T_0 := 0, T_\ell := \tau_1 + \cdots + \tau_\ell$, $\ell \in \mathbb{N}$,

$$M_t := \max\{i \in \mathbb{N}_0 : T_i \le t\}, \quad t \in [0, \infty)$$

der stochastische Prozess $(M_t)_{t\ge 0}$ heißt *Poissonprozess* mit Rate λ. (Beachte: die Definition ist so eingerichtet, dass $t \mapsto M_t$ rechtsstetig ist, man sagt auch: $(M_t)_t$ hat rechtsstetige Pfade.)

Aus obigen Überlegungen folgt für jedes $m \in \mathbb{N}$

$$\big(c_N \tau_1^{(N)}, \dots, c_N \tau_m^{(N)}\big) \xrightarrow[N\to\infty]{d} \big(\tau_1, \dots, \tau_m\big),$$

$$\big(c_N T_1^{(N)}, \dots, c_N T_m^{(N)}\big) \xrightarrow[N\to\infty]{d} \big(T_1, \dots, T_m\big)$$

somit ergibt sich für $t_1 < t_2 < t_m$, $k_1, \dots, k_m \in \mathbb{N}_0$

$$\begin{aligned}
&\mathbb{P}\big(M^{(N)}_{\lfloor t_1/c_N\rfloor} = k_1, \dots, M^{(N)}_{\lfloor t_m/c_N\rfloor} = k_m\big) = \mathbb{P}\big(T^{(N)}_{k_1} \le \lfloor t_1/c_N\rfloor < T^{(N)}_{k_1+1}, \dots, T^{(N)}_{k_m} \le \lfloor t_m/c_N\rfloor < T^{(N)}_{k_m+1}\big) \\
&\xrightarrow[N\to\infty]{d} \mathbb{P}\big(T_{k_1} \le t_1 < T_{k_1+1}, \dots, T_{k_m} \le t_m < T_{k_m+1}\big) = \mathbb{P}\big(M_{t_1} = k_1, \dots, M_{t_m} = k_m\big),
\end{aligned}$$

d.h. die Folge von stochastischen Prozessen $(M^{(N)}_{\lfloor t/c_N\rfloor})_{t\ge 0}$ konvergiert gegen den Prozess $(M_t)_{t\ge 0}$ im Sinne der endlich-dimensionalen Verteilungen.

Aus diesen Beobachtungen folgt

$$\begin{aligned}
&\mathbb{P}\big(M_{t_1} - M_{t_0} = j_1, M_{t_2} - M_{t_1} = j_2 \dots, M_{t_m} - M_{t_{m-1}} = j_m\big) \\
&= \lim_{N\to\infty} \mathbb{P}\big(M^{(N)}_{\lfloor t_1/c_N\rfloor} - M^{(N)}_{\lfloor t_0/c_N\rfloor} = j_1, \dots, M^{(N)}_{\lfloor t_m/c_N\rfloor} - M^{(N)}_{\lfloor t_{m-1}/c_N\rfloor} = j_m\big) \\
&= \lim_{N\to\infty} \mathbb{P}\big(M^{(N)}_{\lfloor t_1/c_N\rfloor} - M^{(N)}_{\lfloor t_0/c_N\rfloor} = j_1\big) \times \cdots \times \mathbb{P}\big(M^{(N)}_{\lfloor t_m/c_N\rfloor} - M^{(N)}_{\lfloor t_{m-1}/c_N\rfloor} = j_m\big) \\
&= \prod_{i=1}^{m} e^{-\lambda(t_i - t_{i-1})} \frac{(\lambda(t_i - t_{i-1}))^{j_i}}{j_i!},
\end{aligned}$$

d.h. die Inkremente eines Poissonprozesses (M_t) sind Poissonverteilt [der Parameter ist $\lambda\times$ die Länge des betrachteten Zeitintervalls] und Inkremente über jeweils disjunkte Zeitintervalle sind unabhängig. Diese beiden Eigenschaften charakterisieren den Poissonprozess [ggfs. mit Forderung der Rechtsstetigkeit].

Der Parameter λ kann als Sprungrate interpretiert werden in dem Sinne, dass für ein (kurzes) Zeitintervall $(t, t+h]$ die Wahrscheinlichkeit, einen Sprung in diesem Zeitintervall zu sehen, $\approx \lambda \times$ Intervalllänge ist, genauer

$$\mathbb{P}\big(M_{t+h} = k+1 \,\big|\, M_t = k\big) = \mathbb{P}\big(M_{t+h} - M_t = 1\big) = e^{-\lambda h}\frac{\lambda h}{1!} = \lambda h + O(h^2) \quad \text{für } h \downarrow 0.$$

Zu allgemeinen zeitkontinuierlichen Markovketten Sei E endliche Menge, $\widehat{p} = (\widehat{p}(x,y))_{x,y\in E}$ stochastische Matrix (d.h. $\widehat{p}(x,y) \ge 0$, $\sum_{y\in E}\widehat{p}(x,y) = 1$ für alle $x \in E$), $\widehat{X} = (\widehat{X}_n)_{n\in\mathbb{N}_0}$ (zeitdiskrete, homogene) $\widehat{p}$-Markovkette (d.h. $\mathbb{P}(\widehat{X}_0 = x_0, \widehat{X}_1 = x_1, \dots, \widehat{X}_n = x_n) = \mathbb{P}(\widehat{X}_0 = x_0)\widehat{p}(x_0,x_1) \times \cdots \times \widehat{p}(x_{n-1}, x_n)$ für $x_0, x_1, \dots, x_n \in E$).

Sei $(M_t)_{t\geq 0}$ Poissonprozess mit Rate $\lambda > 0$, unabhängig von $\widehat{X}$,

$$X_t := \widehat{X}_{M_t}, \quad t \in [0,\infty),$$

so ist [$\widehat{p}^m$ bezeichne die m-te Potenz von $\widehat{p}$, I die $E \times E$-Identitätsmatrix]

$$\begin{aligned} p_t(x,y) &:= \mathbb{P}\big(X_t = y \,\big|\, X_0 = x\big) = \sum_{m=0}^{\infty} \mathbb{P}\big(M_t = m, X_t = y \,\big|\, X_0 = x\big) \\ &= \sum_{m=0}^{\infty} e^{-\lambda t}\frac{(\lambda t)^m}{m!}\widehat{p}^m(x,y) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\frac{(\lambda t)^n}{n!}\frac{(\lambda t)^m}{m!}\big((-I)^n\widehat{p}^m\big)(x,y) \\ &= \sum_{\ell=0}^{\infty}\frac{t^\ell}{\ell!}\lambda^\ell\sum_{m=0}^{\ell}\binom{\ell}{m}\big(\widehat{p}^m(-I)^{\ell-m}\big)(x,y) = \sum_{\ell=0}^{\infty}\frac{t^\ell}{\ell!}\big(\lambda(\widehat{p}-I)\big)^\ell(x,y) = \sum_{\ell=0}^{\infty}\frac{t^\ell}{\ell!}Q^\ell(x,y) = \big(e^{tQ}\big)(x,y) \end{aligned}$$

[wobei die Matrix $Q = (q_{x,y})_{x,y\in E}$ Einträge $q_{x,y} = \lambda(\widehat{p}(x,y) - \delta_{xy})$ besitzt; obige Reihe konvergiert, denn $\max_{x,y\in E}|Q^n_{xy}| \leq (|E|\max_{x,y\in E}|Q_{x,y}|)^n$; beachte auch, dass $\widehat{p}^m$ und I^n kommutieren] und analoge Rechnungen, die die Unabhängigkeit der Zuwächse von $(M_t)_{t\geq 0}$ ausnutzen, zeigen

$$\mathbb{P}\big(X_{t_1} = x_1, \ldots, X_{t_n} = x_n \,\big|\, X_0 = x_0\big) = \prod_{i=1}^{n} p_{t_i - t_{i-i}}(x_{i-1}, x_i)$$

für $0 = t_0 < t_1 < \cdots < t_n$, $x_0, x_1, \ldots, x_n \in E$.

Die Matrix Q heißt die *Sprungratenmatrix* (auch: *Ratenmatrix* oder Q-Matrix) der zeitkontinuierlichen Markovkette X, sie hat die Eigenschaften

$$q_{x,y} \geq 0 \ \text{ für } x \neq y, \qquad \sum_{y\in E, y\neq x} q_{x,y} = -q_{x,x}.$$

Zur Interpretation der Einträge von Q als Sprungraten: Für $x \neq y$ und $h \downarrow 0$ ist

$$\mathbb{P}\big(X_{t+h} = y \,\big|\, X_t = x\big) = \big(e^{hQ}\big)(x,y) = Q^0(x,y) + hQ^1(x,y) + O(h^2) = hq_{x,y} + O(h^2).$$

Kolmogorovs Vorwärts- und Rückwärtsgleichungen für zeitkontinuierliche Markovketten
Für die Sprungratenmatrix $Q = (q_{x,y})_{x,y\in E}$ einer zeitkontinuierlichen Markovkette auf der endlichen Menge E löst $\exp(tQ) = \sum_{n=0}^{\infty}\frac{t^n}{n!}Q^n$ für $t \geq 0$ das System von Differentialgleichungen

$$\frac{\partial}{\partial t}\exp(tQ) = Q\exp(tQ) = \exp(tQ)Q \tag{A.1}$$

demnach für $t \geq 0$ und $P_t = e^{tQ} = (p_t(x,y))_{x,y\in E}$

$$\frac{\partial}{\partial t}P_t = QP_t, \text{ d.h. } \forall\, x,y \in E: \quad \frac{\partial}{\partial t}p_t(x,y) = \sum_z q_{x,z}p_t(z,y) = \sum_z q_{x,z}\big(p_t(z,y) - p_t(x,y)\big), \tag{A.2}$$

$$\frac{\partial}{\partial t}P_t = P_tQ, \text{ d.h. } \forall\, x,y \in E: \quad \frac{\partial}{\partial t}p_t(x,y) = \sum_z p_t(x,z)q_{z,y} = p_t(x,y)q_{y,y} + \sum_{z\neq y} p_t(x,z)q_{z,y} \tag{A.3}$$

(beachte: gliedweise Differentiation der Exponentialreihe ist hier erlaubt).

Die Gleichungen (A.2) und (A.3) haben eine stochastische Interpretation. (A.2) heißt Kolmogorovs *Rückwärtsgleichung*, denn es gilt (wir schreiben $\mathbb{P}_x(\cdot)$ für $\mathbb{P}(\,\cdot\,|X_0 = x)$)

$$\begin{aligned}\frac{p_{t+h}(x,y) - p_t(x,y)}{h} &= \frac{1}{h}\big(\mathbb{P}_x(X_{t+h} = y) - \mathbb{P}_x(X_t = y)\big)\\ &= \frac{1}{h}\Big(\sum\nolimits_z \mathbb{P}_x(X_{t+h} = y|X_h = z)\mathbb{P}_x(X_h = z) - \mathbb{P}_x(X_t = y)\Big)\\ &= \frac{1}{h}\Big(\sum\nolimits_z p_t(z,y)\big(1_{\{x=z\}} + hq_{x,z} + o(h)\big) - p_t(x,y)\Big) = \sum_z q_{x,z}p_t(z,y) + o(1),\end{aligned}$$

man leitet sie also aus her durch „Rückwärtszerlegung“ des Prozesses X im Intervall $[0, t+h]$ gemäß dem Verhalten am Anfang des Intervalls. Analog heißt (A.3) Kolmogorovs *Vorwärtsgleichung*, sie entsteht durch Zerlegung gemäß dem Wert bei t:

$$\begin{aligned}\frac{p_{t+h}(x,y) - p_t(x,y)}{h} &= \frac{1}{h}\Big(\sum\nolimits_z \mathbb{P}_x(X_{t+h} = y|X_t = z)\mathbb{P}_x(X_t = z) - \mathbb{P}_x(X_t = y)\Big)\\ &= \frac{1}{h}\Big(\sum\nolimits_z p_t(x,z)\big(1_{\{z=y\}} + hq_{y,z} + o(h)\big) - p_t(x,y)\Big) = \sum_z q_{x,z}p_t(z,y) + o(1).\end{aligned}$$

Sowohl (A.2) als auch (A.3) sind (im Fall $|E| < \infty$) eindeutig lösbar und beide bestimmen die Halbgruppe von Übergangsmatrizen $(P_t)_{t\geq 0}$ – es sind beides endliche Systeme linearer Differentialgleichungen mit konstanten Koeffizienten.

Beispiel A.1. a) Sei $E = \{0,1\}$, $Q = \left(\begin{smallmatrix} -a & a \\ b & -b \end{smallmatrix}\right)$ mit $a, b > 0$. Für $x, y \in \{0,1\}$ gilt $p_t(x,y) = \delta_{x,y}e^{-(a+b)t} + (1 - e^{-(a+b)t})\mu(y)$ mit $\mu(0) = b/(a+b)$, $\mu(1) = a/(a+b)$.
b) Sei $E = \{0,1,2,3\}$ (oder auch $E = \{\mathsf{A}, \mathsf{G}, \mathsf{C}, \mathsf{T}\}$) und

$$Q = \begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix}.$$

Für $x, y \in E$ gilt $p_t(x,y) = \delta_{x,y}e^{-4t} + \frac{1}{4}(1 - e^{-4t})$.

Lemma A.2. *E endliche Menge, $Q = (q_{xy})_{x,y\in E}$ Sprungratenmatrix, $X^{(N)}$, $N \in \mathbb{N}$ zeitdiskrete E-wertige Markovketten mit Übergangsmatrix*

$$p^{(N)}(x,y) = \delta_{x,y} + c_N q_{xy} + o(c_N), \quad x, y \in E,$$

wo $c_N \to 0$ für $N \to \infty$ und $X_0^{(N)} = x_0 \in E$. Dann konvergieren die (zeitlich reskalierten) Prozesse $(X^{(N)}_{\lfloor t/c_N \rfloor})_{t\geq 0}$ für $N \to \infty$ gegen die zeitkontinuierliche Markovkette X mit Sprungratenmatrix Q (im Sinne der endlich-dimensionalen Verteilungen).

Beweis. Wir schreiben die Übergangsmatrix von $X^{(N)}$ als

$$p^{(N)} = I + c_N Q_N$$

mit $Q_N := c_N^{-1}(p^{(N)} - I)$, somit nach Voraussetzung $Q_N \xrightarrow[N\to\infty]{} Q$ (eintrags-weise).

$$\begin{aligned}(I + c_N Q_N)^{\lfloor c_N^{-1} t\rfloor} &= \sum_{k=0}^{\lfloor c_N^{-1} t\rfloor} \binom{\lfloor c_N^{-1} t\rfloor}{k} c_N^k Q_N^k = \sum_{k=0}^{\lfloor c_N^{-1} t\rfloor} c_N^k \frac{\lfloor c_N^{-1} t\rfloor(\lfloor c_N^{-1} t\rfloor - 1)\cdots(\lfloor c_N^{-1} t\rfloor - k + 1)}{k!} Q_N^k \\ &\to \sum_{k=0}^{\infty} \frac{t^k}{k!} Q^k = e^{tQ} \qquad \text{für } N \to \infty.\end{aligned}$$

Beachte: Für $k \in \mathbb{N}$ gilt

$$c_N^k \frac{\lfloor c_N^{-1} t\rfloor(\lfloor c_N^{-1} t\rfloor - 1)\cdots(\lfloor c_N^{-1} t\rfloor - k + 1)}{k!} Q_N^k \xrightarrow[N\to\infty]{} \frac{t^k}{k!} Q^k$$

(eintrags-weise) und die Beträge der Einträge der Matrix auf der linken Seite sind für genügend großes N

$$\le t^k (2\max\{|Q(x,y)| : x, y \in E\})^k / k!,$$

so dass Grenzwert und Summation vertauscht werden können. Somit

$$\begin{aligned}\mathbb{P}\big(X^{(N)}_{\lfloor c_N^{-1} t_1\rfloor} = x_1, \dots, X^{(N)}_{\lfloor c_N^{-1} t_n\rfloor} = x_n\big) &= \prod_{j=1}^{n} (I + c_N Q_N)^{\lfloor c_N^{-1}(t_j - t_{j-1})\rfloor}(x_{j-1}, x_j) \\ &\xrightarrow[N\to\infty]{} \prod_{j=1}^{n} (e^{(t_j - t_{j-1})Q})(x_{j-1}, x_j) = \mathbb{P}\big(X_{t_1} = x_1, \dots, X_{t_n} = x_n\big).\end{aligned}$$

□

A.2 (Weitere) Eigenschaften des Kingman-Koaleszenten

Sei $\xi_i^{(n)}$, $i = n, n-1, \dots, 1$ der Zustand des n-Koaleszenten zum ersten Zeitpunkt, zu dem i Klassen existieren, d.h. $\xi_i^{(n)} = R^{(n)}_{\tau_i^{(n)}}$ (mit $\tau_i^{(n)} := \inf\{t \ge 0 : |R_t^{(n)}| \le i\}$ wie oben). [$(\xi_i^{(n)})_{i=n,n-1,\dots,1}$ heißt die *Skelettkette* des Kingman-n-Koaleszenten, sie ist (offenbar) eine Markovkette.]

Proposition A.3. *Für $\xi \in \mathcal{E}_n$ mit i Klassen der Größen $\lambda_1, \dots, \lambda_i \in \mathbb{N}$ (mit $\lambda_1 + \cdots + \lambda_i = n$) gilt*

$$\mathbb{P}(\xi_i^{(n)} = \xi) = c_{n,i} w(\xi) \quad \textit{mit} \;\; w(\xi) = \lambda_1! \cdots \lambda_i!, \;\; c_{n,i} = \frac{i!}{n!} \frac{(n-i)!(i-1)!}{(n-1)!}. \tag{A.4}$$

Beispiel. Betrachte $n = 9$, $i = 3$, es ist $c_{9,3} = \frac{3!}{9!}\frac{6!2!}{8!} = 1/1\,693\,440$.

λ_i	3-3-3	4-3-2	5-2-2	4-4-1	5-3-1	6-2-1	7-1-1
w	216	288	480	576	720	1440	5040

Wir sehen: die Verteilung hat mehr Gewicht auf „unbalanzierten Aufteilungen."

Beweis von Prop. A.3. Rückwärtsinduktion über i: Für $i = n$ gilt $\mathbb{P}(\xi_n^{(n)} = \{\{1\}, \dots, \{n\}\}) = 1$ mit $\lambda_1 = \cdots = \lambda_n = 1$, und $c_{n,n} = w(1, \dots, 1) = 1$.

$i \to i-1$: Es ist

$$\mathbb{P}(\xi_{i-1}^{(n)} = \eta \,|\, \xi_i^{(n)} = \xi) = \begin{cases} \dfrac{1}{\binom{i}{2}}, & \text{falls } \eta \text{ aus } \xi \text{ durch Verschmelzung eines Paars von Klassen entsteht,} \\ 0, & \text{sonst.} \end{cases}$$

Sei $\eta \in \mathcal{E}_n$, $|\eta| = i-1$, Klassengrößen $\tilde\lambda_1, \dots, \tilde\lambda_{i-1}$.

$$\begin{aligned}
\mathbb{P}(\xi_{i-1}^{(n)} = \eta) &= \frac{2}{i(i-1)} \sum_{\xi : \xi \prec \eta} \mathbb{P}(\xi_i^{(n)} = \xi) \\
&= \frac{2}{i(i-1)} \sum_{\ell=1}^{i-1} \sum_{m=1}^{\tilde\lambda_\ell - 1} \frac{1}{2} \binom{\tilde\lambda_\ell}{m} c_{n,i} \tilde\lambda_1! \cdots \tilde\lambda_{\ell-1}! m! (\tilde\lambda_\ell - m)! \tilde\lambda_{\ell+1}! \cdots \tilde\lambda_{i-1}! \\
&= \frac{c_{n,i}}{i(i-1)} \sum_{\ell=1}^{i-1} \sum_{m=1}^{\tilde\lambda_\ell - 1} w(\eta) = \frac{c_{n,i} w(\eta)}{i(i-1)} \sum_{\ell=1}^{i-1} \sum_{m=1}^{\tilde\lambda_\ell - 1} 1 \\
&= \frac{c_{n,i} w(\eta)}{i(i-1)} (n - (i-1)) = c_{n,i-1} w(\eta)
\end{aligned}$$

Für das 2. Gleichheitszeichen verwenden wir die Induktionsannahme und zerlegen gemäß der „aufgespaltenen" Klasse: die ℓ-te Klasse hat $\tilde\lambda_\ell$ Elemente, zerlege in 2 Teilmengen d. Größen m und $\tilde\lambda_\ell - m$, es gibt $\frac{1}{2}\binom{\tilde\lambda_\ell}{m}$ mögliche Wahlen; der Faktor $\frac{1}{2}$ entsteht, weil die Klassen in η als ungeordnet aufgefasst werden. □

Korollar A.4. *1. Sei σ eine uniform verteilte Permutation von $\{1, \dots, i\}$, u.a. von $\xi^{(n)}$, $M_i = |C_{i,\sigma(i)}^{(n)}|$ mit $\xi_i^{(n)} = \{C_{i,1}^{(n)}, \dots, C_{i,i}^{(n)}\}$. Dann ist*

$$(M_1, \dots, M_i) \text{ uniform verteilt auf } \big\{(m_1, \dots, m_i) \in \mathbb{N}^i : m_1 + \cdots + m_i = n\big\}.$$

2. Sei $\xi = \{A_1, \dots, A_{i-1}\} \in \mathcal{E}_n$ mit $|A_j| = \lambda_j$. $\mathscr{L}(\xi_i^{(n)} \,|\, \xi_{i-1}^{(n)} = \xi)$ kann folgendermaßen beschrieben werden:

- *wähle A_j mit W'keit $\frac{\lambda_j - 1}{n-i+1}$ ($j \in \{1, \dots, i-1\}$), dann wähle k uniform aus $\{1, \dots, \lambda_j - 1\}$*
- *spalte A_j uniform in zwei Teile der Größen k und $\lambda_j - k$.*

Bemerkung. Für $i = 2$ zeigt Kor. A.4 die „uniforme Aufspaltung" der Stichprobe in zwei älteste Familien.

Beweis von Kor. A.4. 1. Seien $m_1, \dots, m_i$ mit $m_1 + \cdots + m_i = n$ gegeben. Nach Prop. A.3 hat jede Realisierung des (zufällig) geordneten Vektors $(C_{i,\sigma(1)}^{(n)}, \dots, C_{i,\sigma(i)}^{(n)})$, die mit den geforderten Größen m_j verträglich ist, die W'keit $\frac{1}{i!} c_{n,i} m_1! \cdots m_i!$, somit

$$\begin{aligned}
\mathbb{P}\big((M_1, \dots, M_i) &= (m_1, \dots, m_i)\big) = \binom{n}{m_1 \ \dots \ m_i} \frac{1}{i!} c_{n,i} m_1! \cdots m_i! \\
&= \frac{(n-i)!(i-1)!}{(n-1)!} = \frac{1}{\binom{n-1}{i-1}},
\end{aligned}$$

denn es gibt $\binom{n}{m_1 \dots m_i}$ Partitionen, die bezgl. der Größe der Klassen in Frage kommen,

jede hat n. Prop. A.3 dieselbe W'keit $c_{n,i} m_1! \cdots m_i!$,

die W'keit, dass zuf. Perm. σ geg. Ordnung liefert, ergibt nochmals einen Faktor $\frac{1}{i!}$.

(Beachte auch $\#\{\{(m_1, \dots, m_i) \in \mathbb{N}^i : m_1 + \cdots + m_i = n\}\} = \binom{n-1}{i-1}$: n Kugeln in i (nummerierte) Schachteln legen, so dass keine Schachtel leer ist: $n-1$ mögl. Plätze für $i-1$ "Trennwände".)

2. ξ entstehe aus η durch Aufteilen von A_j in 2 Teile der Größen k und $\lambda_j - k$.

$$\begin{aligned}\mathbb{P}(\xi_i^{(n)} = \xi \,|\, \xi_{i-1}^{(n)} = \eta) &= \frac{\mathbb{P}(\xi_{i-1}^{(n)} = \eta \,|\, \xi_i^{(n)} = \xi)\mathbb{P}(\xi_i^{(n)} = \xi)}{\mathbb{P}(\xi_{i-1}^{(n)} = \eta)} \\ &= \frac{\frac{1}{\binom{i}{2}} c_{n,i}\lambda_1!\cdots\lambda_{j-1}!k!(\lambda_j - k)!\lambda_{j+1}!\cdots\lambda_{i-1}!}{c_{n,i-1}\lambda_1!\cdots\lambda_{j-1}!\lambda_j!\lambda_{j+1}!\cdots\lambda_{i-1}!} = \frac{1}{\binom{i}{2}} \cdot \frac{i(i-1)}{n-i+1}\frac{1}{\binom{\lambda_j}{k}} \\ &= \frac{\lambda_j - 1}{n-i+1} \cdot \frac{1}{\lambda_j - 1} \cdot 2\frac{1}{\binom{\lambda_j}{k}}\end{aligned}$$

($\frac{\lambda_j - 1}{n-i+1} \widehat{=}$ Wahl von A_j; $\frac{1}{\lambda_j - 1} \widehat{=}$ Wahl von k; $2\frac{1}{\binom{\lambda_j}{k}} \widehat{=}$ Wahl der Zerlegung von A_j – beachte: Faktor 2, da die Klassen "ungeordnet" angegeben werden) □

Korollar A.5. *Betrachte eine Teilstichprobe der Grösse n in einem Kingman-m-Koaleszenten, mit $m > n$. Dann erfüllt die Wahrscheinlichkeit des Ereignisses $E_{m,n}$, dass der jüngste gemeinsame Vorfahre (jgV) der n-Stichprobe mit der Wurzel des m-Koaleszenten übereinstimmt,*

$$\mathbb{P}(E_{m,n}) \to \frac{n-1}{n+1} \quad \textit{für } m \to \infty.$$

Beweis. Wir betrachten $\xi_2^{(n)}$, die erste Aufspaltung des Koaleszenten von der Wurzel aus betrachtet. Diese resultiert in einer Aufspaltung der trivialen Partition $\{\{1, \ldots, m\}\}$ in eine Äquivalenzrelation mit genau zwei Klassen der Grössen $m - X$ und X, wobei X nach Kor. A.4 auf $[m-1]$ uniform verteilt ist. Falls der jgV der n-Stichprobe nicht mit der Wurzel übereinstimmt, so müssen die Ahnenlinien aller n Individuen der Stichprobe alle entweder in dem Block der Grösse $m - X$ oder in dem Block der Grösse X liegen.

Das erste Ereignis hat Wahrscheinlichkeit $\frac{(m-X)_{n\downarrow}}{(m)_{n\downarrow}}$ und das zweite $\frac{(X)_{n\downarrow}}{(m)_{n\downarrow}}$.

Wir erhalten

$$\begin{aligned}\mathbb{P}(E_{mn,}) &= 1 - \mathbb{P}\big((E_{mn,})^c\big) = 1 - \sum_{k=1}^{m-1}\Big[\frac{(m-k)_{n\downarrow}}{(m)_{n\downarrow}} + \frac{(k)_{n\downarrow}}{(m)_{n\downarrow}}\Big]\underbrace{\mathbb{P}(X=k)}_{\frac{1}{m-1}} \\ &\xrightarrow[m\to\infty]{} 1 - \int_0^1 (x^n + (1-x)^n)\,dx \\ &= 1 - \Big[\frac{1}{n+1}x^{n+1}\Big]_0^1 - \Big[\frac{1}{n+1}(1-x)^{n+1}(-1)\Big]_0^1 = 1 - \frac{2}{n+1}\end{aligned}$$

für $m \to \infty$ durch Konvergenz der Riemann-Summe gegen das Riemann-Integral. □

A.3 Die Verteilung der Summe unabhängiger, exponentialverteilter Zufallsvariablen

Lemma A.6. *Seien $X_1, X_2, \ldots, X_n$ unabhängig, X_i sei exponentialverteilt mit Parameter λ_i und $\lambda_1 < \lambda_2 < \cdots < \lambda_n$. Dann hat $X := X_1 + \cdots + X_n$ die Dichte*

$$f_X(t) = \sum_{j=1}^{n} \lambda_j \exp(-\lambda_j t) \prod_{k=1, k\neq j}^{n} \frac{\lambda_k}{\lambda_k - \lambda_j},$$

insbesondere ist (mit $a_j := \prod_{k=1,k\neq j}^{n} \frac{\lambda_k}{\lambda_k-\lambda_j}$*)*

$$\mathbb{P}(X > t) = \sum_{j=1}^{n} a_j \exp(-\lambda_j t), \quad t \geq 0.$$

Beweisskizze. Die Formel für die Dichte kann man beispielsweise per Induktion durch sukzessive Faltung mit der Exponentialdichte beweisen, für den Induktionsschritt beachten wir

$$\begin{aligned}
&\int_0^t \sum_{j=1}^{n-1} \lambda_j \exp(-\lambda_j s) \prod_{k=1,k\neq j}^{n-1} \frac{\lambda_k}{\lambda_k-\lambda_j} \times \lambda_n \exp(-\lambda_n(t-s))\, ds \\
&= \sum_{j=1}^{n-1} \lambda_j \lambda_n \prod_{k=1,k\neq j}^{n-1} \frac{\lambda_k}{\lambda_k-\lambda_j} \times e^{-\lambda_n t} \int_0^t e^{(\lambda_n-\lambda_j)s}\, ds = \sum_{j=1}^{n-1} \lambda_j \lambda_n \prod_{k=1,k\neq j}^{n-1} \frac{\lambda_k}{\lambda_k-\lambda_j} \times \frac{e^{-\lambda_j t} - e^{-\lambda_n t}}{\lambda_n-\lambda_j} \\
&= \sum_{j=1}^{n-1} \lambda_j e^{-\lambda_j t} \prod_{k=1,k\neq j}^{n} \frac{\lambda_k}{\lambda_k-\lambda_j} - \lambda_n e^{-\lambda_n t} \sum_{j=1}^{n-1} \frac{\lambda_j}{\lambda_n-\lambda_j} \prod_{k=1,k\neq j}^{n-1} \frac{\lambda_k}{\lambda_k-\lambda_j}.
\end{aligned}$$

Dann verwenden wir die Identität

$$\sum_{j=1}^{n-1} \frac{\lambda_j}{\lambda_n-\lambda_j} \prod_{k=1,k\neq j}^{n-1} \frac{\lambda_k}{\lambda_k-\lambda_j} = -\prod_{k=1}^{n-1} \frac{\lambda_k}{\lambda_k-\lambda_n},$$

die (man dividiere beide Seiten durch $\lambda_1\lambda_2\cdots\lambda_{n-1}$ und sortiere Terme) äquivalent ist zu

$$\sum_{j=1}^{n} \prod_{k=1,k\neq j}^{n} \frac{1}{\lambda_k-\lambda_j} = 0. \tag{A.5}$$

Sei $\ell_j(x) := \prod_{k=1,k\neq j}^{n} \frac{x-\lambda_k}{\lambda_j-\lambda_k}$ (das j-te Lagrange-Polynom zu $\lambda_1,\dots,\lambda_n$), $\ell_1(x) + \ell_2(x) + \cdots + \ell_n(x)$ ist ein Polynom in x vom Grad $n-1$, das (mindestens) an den n verschiedenen Stellen $\lambda_1,\dots,\lambda_n$ den Wert 1 annimmt (denn $\ell_j(\lambda_i) = \delta_{ji}$), daher gilt $\ell_1(x) + \ell_2(x) + \cdots + \ell_n(x) \equiv 1$. Die linke Seite von (A.5) ist $(-1)^{n-1}$ mal der Koeffizient von x^{n-1} in diesem Polynom.

Alternativ beachte man, dass für $\zeta \in \mathbb{R}$ ist $\mathbb{E}[e^{i\zeta X_j}] = \int_0^\infty e^{i\zeta x}\lambda_j e^{-\lambda_j x}\, dx = \frac{\lambda_j}{\lambda_j - i\zeta}$, also $\mathbb{E}[e^{i\zeta X}] = \prod_{j=1}^{n} \frac{\lambda_j}{\lambda_j-i\zeta} =: \varphi_1(\zeta)$ gilt, während $\int_0^\infty e^{i\zeta x} \sum_{j=1}^{n} a_j\lambda_j e^{-\lambda_j x}\, dx = \sum_{j=1}^{n} \frac{a_j\lambda_j}{\lambda_j-i\zeta} =: \varphi_2(z)$ und es *ist* $\varphi_2 = \varphi_1$ (φ_2 ist die Partialbruchzerlegungs-Darstellung von φ_1), denn beide sind meromorph auf $\mathbb{C}$ mit jeweils einfachen Polen bei $\zeta = -i\lambda_1,\dots,-i\lambda_n$und $\lim_{|z|\to\infty} \varphi_{1/2}(z) = 0$, $\lim_{z\to -i\lambda_j} \frac{\varphi_1(z)}{\lambda_j - iz} = \lambda_j \prod_{k=1,k\neq j}^{n} \frac{\lambda_k}{\lambda_k-\lambda_j} = \lim_{z\to -i\lambda_j} \frac{\varphi_2(z)}{\lambda_j - iz}$.

Die Formel für den Verteilungsschwanz von X ergibt sich durch entsprechendes Integrieren der Dichte. □

A.4 Erwartete Fixationszeit im Wright-Fisher-Modell: exakte Rechnung

In diesem Kapitel führen wir den vollständigen Beweis von Satz 1.1 aus, d.h. für die Fixationszeit $T_{\text{fix}}^{(N)} = \inf\{r \geq 0 : X_r^{(N)} = 0 \text{ oder } X_r^{(N)} = N\}$ im neutralen 2-Typ-Wright-Fisher-Modell mt

Populationsgröße N gilt

$$\lim_{N\to\infty} c_N \frac{\mathbb{E}_{x_N}[T_{\text{fix}}^{(N)}]}{2N} = H(p) = -p\log(p) - (1-p)\log(1-p) \tag{1.5}$$

falls $x_N/N \to p \in [0,1]$.

Wir schreiben im Folgenden $p_r := X_r^{(N)}/N$ für den Anteil von Typ A-Individuen in der r-ten Generation (und unterdrücken in der Notation, dass dessen Verteilung natürlich auch von N abhängt), sowie $\mathbb{P}_p$ bzw. $\mathbb{E}_p$ für Wahrscheinlichkeiten bzw. Erwartungswerte in der Sitation, dass der Startanteil $p_0 = X_0^{(N)}/N = p$ beträgt.

Wir zeigen zunächst die obere Schranke

$$\mathbb{E}_p\big[T_{\text{fix}}^{(N)}\big] \le 2NH(p) \tag{A.6}$$

Für $p \in \{0,1\}$ gilt die Aussage trivialerweise. Für $p \in \{\frac{1}{N}, \dots \frac{N-1}{N}\}$ liefert die „Ein-Schritt-Analyse“ (1.4)

$$\mathbb{E}_p[T_{\text{fix}}^{(N)}] = 1 + \sum_{q\in\{0,\frac{1}{N},\dots,1\}} \mathbb{P}_p\{p_1 = q\}\mathbb{E}_q[T_{\text{fix}}^{(N)}]. \tag{A.7}$$

Unser Ziel ist nun zu zeigen, dass für alle $p \in I_N = \{0, \frac{1}{N}, \dots, 1\}$ und $N \in \mathbb{N}$ die Funktion

$$f_N(p) := 2NH(p) - \mathbb{E}_p[T_{\text{fix}}^{(N)}],$$

die Ungleichung

$$\sum_{q\in I_N} \mathbb{P}_p(p_1 = q)\big[f_N(q) - f_N(p)\big] \le 0, \tag{A.8}$$

mit $f_N(0) = f_N(1) = 0$ erfüllt. Denn damit ist f_N superharmonisch und $\{f_N(p_r)\}_{r\ge 0}$ ein $\mathcal{F}^N$-Supermartingal [KW10, Kapitel 2.7]. Mit dem Doobschen Stoppsatz [KW10, Satz 1.10] folgt für jedes für $p \in I_N$, dass

$$f_N(p) \ge \mathbb{E}_p\big[f_N(p_{T_{\text{fix}}^{(N)}})\big] = 0,$$

und damit gilt

$$f_N(p) = 2NH(p) - \mathbb{E}_p[T_{\text{fix}}^{(N)}] \ge 0,$$

und dies ist gerade die Aussage (A.6).

Es bleibt also noch (A.8) zu zeigen, was für Randfälle $p \in \{0,1\}$ offensichtlich gilt. Angesichts von (A.7) ist (A.8) für $p \in \{\frac{1}{N}, \dots \frac{N-1}{N}\}$ äquivalent zu folgender Ungleichung:

$$2N \sum_{q\in I_N} \mathbb{P}_p(p_1 = q)\big[H(q) - H(p)\big] \le -1 \tag{A.9}$$

Da die Funktion $H(\cdot)$ symmetrisch bezüglich Spiegelung um $p = 1/2$ ist, d.h. invariant unter der Ersetzung $p \mapsto 1-p$, gilt

$$\begin{aligned}\sum_{q\in I_N} \mathbb{P}_{1-p}(p_1 = q)\big[H(q) - H(1-p)\big] &= \sum_{q\in I_N} \mathbb{P}_{1-p}\{p_1 = 1-q\}\big[H(1-q) - H(1-p)\big] \\ &= \sum_{q\in I_N} \mathbb{P}_p(p_1 = q)\big[H(q) - H(p)\big]\end{aligned}$$

Daher genügt es, den Fall $0 < p \le 1/2$ zu betrachten. H ist im Innern von $[0,1]$ beliebig oft stetig differenzierbar; Taylor-Entwicklung bis zur ersten Ordnung mit Restglied in Integralform liefert

$$\begin{aligned} H(q) &= H(p) + (q-p)H'(p) + \int_p^q (q-t)H''(t)\,dt \\ &= H(p) + (q-p)H'(p) + (q-p)^2 \int_0^1 H''\big(p+u(q-p)\big)(1-u)\,du, \quad q \in I_N \end{aligned}$$

wobei wir in der zweiten Zeile $t = p + u(q-p)$ substituiert haben. Wir definieren ein Wahrscheinlichkeitsmaß $\tilde{\mu}_{N,p}$ auf $(0,1)$ mittels

$$\int_{(0,1)} f(x)\,\tilde{\mu}_{N,p}(dx) := \frac{N}{p(1-p)} \int_0^1 \mathbb{E}_p\Big[f\big(p+u(p_1-p)\big)(p_1-p)^2\Big] 2(1-u)\,du$$

(siehe Lemma A.7 (a)).

Damit schreibt sich

$$\begin{aligned} \sum_{q\in I_N} \mathbb{P}_p(p_1 = q)\big[H(q) - H(p)\big] &= \frac{p(1-p)}{2N} \int_{(0,1)} H''(x)\,\tilde{\mu}_{N,p}(dx) \\ &= -\frac{p(1-p)}{2N} \int_{(0,1)} \varphi(x)\,\tilde{\mu}_{N,p}(dx) \le -\frac{p(1-p)}{2N}\,\varphi\bigg(\int_{(0,1)} x\,\tilde{\mu}_{N,p}(dx)\bigg) \end{aligned}$$

mit $\varphi(x) = \frac{1}{x(1-x)}$, wobei das Ungleichungszeichen aus der Konvexität der Funktion φ auf $(0,1)$ und der Jensenschen Ungleichung folgt. Weiter ist

$$\begin{aligned} \int_{(0,1)} x\,\tilde{\mu}_{N,p}(dx) &= \frac{N}{2p(1-p)} \int_0^1 \mathbb{E}_p\Big[\big(p+u(p_1-p)\big)(p_1-p)^2\Big] 2(1-u)\,du \\ &= \frac{N}{2p(1-p)}\Big(p\mathbb{E}_p\big[(p_1-p)^2\big] + \frac{1}{3}\mathbb{E}_p\big[(p_1-p)^3\big]\Big) \end{aligned}$$

wobei wir den Satz von Fubini sowie $\int_0^1 2u(1-u)\,du = 1/3$ verwenden. Es ist

$$\mathbb{E}_p\big[(p_1-p)^3\big] = \frac{1}{N^3}\mathbb{E}\big[(Y_{N,p} - Np)^3\big] = \frac{1}{N^2}p(1-p)(1-2p)$$

mit $Y_{N,p} \sim \mathrm{Bin}(N,p)$ und $\mathbb{E}_p\big[(p_1-p)^2\big] = p(1-p)/N$ (siehe Lemma A.7 (b)).

Damit finden wir

$$\int_{(0,1)} x\,\tilde{\mu}_{N,p}(dx) = \frac{p}{2} + \frac{1-2p}{6N}$$

und daher

$$\begin{aligned} &\sum_{q\in I_N} \mathbb{P}_p(p_1 = q)\big[H(q) - H(p)\big] \\ &\qquad \le -\frac{p(1-p)}{2N} \frac{1}{\big(p/2 + (1-2p)/(6N)\big)\big(1-p/2-(1-2p)/(6N)\big)} \le -\frac{1}{2N} \end{aligned}$$

(beachte $\frac{p(1-p)}{(p/2+(1-2p)/(6N))(1-p/2-(1-2p)/(6N))} \ge 1$ für $p \le 1/2$). Dies komplettiert den Beweis von (A.9) und damit auch der Aussage (A.6).

Um die Asymptotik von $\mathbb{E}_p[T_{\text{fix}}^{(N)}]$ aus (1.5) zu zeigen, betrachten wir nun die Taylor-Entwicklung von H bis zur zweiten Ordnung. Sei zunächst $\varepsilon \in (0, \frac{1}{2})$ und setze

$$T_\varepsilon^{(N)} := \inf\{r : p_r < \varepsilon \text{ oder } p_r > 1-\varepsilon\}.$$

Für $\zeta \in (\varepsilon/2, 1-\varepsilon/2)$ ist

$$H'''(\zeta) = \frac{1-2\zeta}{\zeta^2(1-\zeta^2)}$$

gleichmäßig beschränkt. Taylor-Entwicklung bis zur zweiten Ordnung mit Restglied in Integralform liefert

$$\begin{aligned} H(q) = H(p) + (q-p)H'(p) + \frac{(q-p)^2}{2}H''(p) \\ + \frac{(q-p)^3}{2}\int_0^1 H'''\big(p+u(q-p)\big)(1-u)^2\,du \end{aligned}$$

Weiter ist

$$\begin{aligned} &\left|H'''\big(p+u(q-p)\big)(1-u)^2\right| \\ &\quad\le \frac{(1-u)^2}{((1-u)p+uq)^2(1-(1-u)p-uq)^2} \le \frac{1}{p^2(1-p)^2} \le \varepsilon^{-4} \end{aligned}$$

Für das zweite Ungleichungszeichen betrachten wir hier eine die Fallunterscheidung: Falls $q < p$, so ist der erste Term in der zweiten Zeile höchstens

$$(1-u)^2/\left(((1-u)p+0)(1-(1-u)p-up)\right)^2 = 1/(p(1-p))^2$$

falls $q \ge p$ gilt, so ist er höchstens $(1-u)^2/\left(p(1-(1-u)p-u)\right)^2 = 1/(p(1-p))^2$.

Somit gilt

$$\begin{aligned} &\sum_{q\in\{0,\frac{1}{N},\dots,1\}} \mathbb{P}_p(p_1 = q)\big[H(q) - H(p)\big] \\ &= \sum_{q\in\{0,\frac{1}{N},\dots,1\}} \mathbb{P}_p(p_1 = q)\Big[(q-p)H'(p) + \frac{1}{2}(q-p)^2H''(p) + \frac{1}{6}(q-p)^3H''(\zeta(q,p))\Big] \\ &= -\frac{1}{2} + \frac{1}{6}\sum_{q\in\{0,\frac{1}{N},\dots,1\}} \mathbb{P}_p(p_1 = q)(q-p)^3H'''(\zeta(q,p)) \\ &= -\frac{1}{2} + R_N(p) \end{aligned}$$

wobei das Restglied die Abschätzung

$$\max_{p\in\{0,\frac{1}{N},\dots,1\}\cap(\varepsilon,1-\varepsilon)} |R_N(p)| \le \varepsilon^{-4} \max_{p\in\{0,\frac{1}{N},\dots,1\}} \mathbb{E}_p\big[|p_1-p|^3\big] \le \frac{1}{\varepsilon^4 N^{3/2}}$$

erfüllt: Für $p \in [0,1]$ ist mit der Jensenschen Ungleichung und Lemma A.7 (b)

$$\mathbb{E}_p\big[|p_1-p|^3\big] \le \big(\mathbb{E}_p\big[|p_1-p|^4\big]\big)^{3/4} \le \big(N^{-2}\big)^{3/4}$$

Die übliche Ein-Schritt-Analyse zusammen mit obigem zeigt, dass

$$f_{N,\varepsilon}(p) := 2NH(p) - \mathbb{E}_p[T_\varepsilon^{(N)}]$$

für $p \in \{0, \frac{1}{N}, \dots, 1\} \cap (\varepsilon, 1-\varepsilon)$ die Gleichung

$$\sum_{q \in \{0, \frac{1}{N}, \dots, 1\}} \mathbb{P}_p(p_1 = q)\big[f_{N,\varepsilon}(q) - f_{N,\varepsilon}(p)\big] = 2NR_N(p)$$

erfüllt. Zudem gilt für $p \in \{0, \frac{1}{N}, \dots, 1\} \cap \big([0,\varepsilon] \cap [1-\varepsilon, 1]\big)$

$$0 \le f_{N,\varepsilon}(p) \le 2N\Big[\varepsilon \log \frac{1}{\varepsilon} + (1-\varepsilon)\log\frac{1}{1-\varepsilon}\Big]$$

Somit erhalten wir `[Referenz für geeigneten Stoppsatz heraussuchen...]`

$$\begin{aligned}\Big|\frac{1}{2N} f_{N,\varepsilon}(p)\Big| &= \Big|\frac{1}{2N}\mathbb{E}_p\Big[f_{N,\varepsilon}(p_{T_\varepsilon^{(N)}}) + \sum_{j=0}^{T_\varepsilon^{(N)}-1} 2NR_N(p_j)\Big]\Big| \\ &\le \Big[\varepsilon\log\frac{1}{\varepsilon} + (1-\varepsilon)\log\frac{1}{1-\varepsilon}\Big] + \frac{1}{\varepsilon^4 N^{3/2}}\mathbb{E}_p\big[T_\varepsilon^{(N)}\big]\end{aligned}$$

Da $\mathbb{E}_p\big[T_\varepsilon^{(N)}\big] \le \mathbb{E}_p\big[T_{\text{fix}}^{(N)}\big] \le 2NH(p)$ gemäß (A.6) gilt, folgt

$$\limsup_{N\to\infty} \max_{p \in \{0, \frac{1}{N}, \dots, 1\}} \Big|\frac{1}{2N} f_{N,\varepsilon}(p)\Big| \le H(\varepsilon)$$

mit $H(\varepsilon) \to 0$ für $\varepsilon \downarrow 0$. Offenbar ist stets $T_{\text{fix}}^{(N)} - T_\varepsilon^{(N)} \ge 0$. Die starke Markov-Eigenschaft (und ein offensichliches Kopplungsargument, sowie Spiegelungssymmetrie um $p = 1/2$) zeigt, dass

$$0 \le \mathbb{E}_p\big[T_{\text{fix}}^{(N)} - T_\varepsilon^{(N)}\big] \le \sup_{0<p'\le\varepsilon} \mathbb{E}_{p'}\big[T_{\text{fix}}^{(N)}\big] \le 2NH(\varepsilon).$$

Insgesamt ergibt sich

$$\begin{aligned}\limsup_{N\to\infty}\Big|\frac{\mathbb{E}_p\big[T_{\text{fix}}^{(N)}\big]}{2N} - H(p)\Big| &\le \limsup_{N\to\infty}\Big|\frac{\mathbb{E}_p\big[T_\varepsilon^{(N)}\big]}{2N} - H(p)\Big| \\ &\quad + \limsup_{N\to\infty}\Big|\frac{\mathbb{E}_p\big[T_{\text{fix}}^{(N)} - T_\varepsilon^{(N)}\big]}{2N}\Big| \le H(\varepsilon) + H(\varepsilon)\end{aligned}$$

und mit $\varepsilon \downarrow 0$ folgt die Behauptung. □

Wir tragen einige kleine Details aus dem Beweis von Satz 1.1 in folgendem Lemma zusammen:

Lemma A.7. *(a) Für $0 < p < 1$ und $N \in \mathbb{N}$ definiert die Formel*

$$\int_{(0,1)} f(x)\,\tilde\mu_{N,p}(dx) := \frac{N}{p(1-p)}\int_0^1 \mathbb{E}_p\Big[f\big(p + u(p_1 - p)\big)(p_1-p)^2\Big]2(1-u)\,du \quad \text{(A.10)}$$

für beschränktes (und messbares) $f : [0,1] \to \mathbb{R}$ ein Wahrscheinlichkeitsmaß $\tilde\mu_{N,p}$ auf $[0,1]$.

(b) Für $Y_{N,p} \sim Bin(N,p)$ gilt

$$\mathbb{E}\big[(Y_{N,p} - Np)^3\big] = Np(1-p)(1-2p),$$
$$\mathbb{E}\big[(Y_{N,p} - Np)^4\big] = Np(1-p)\big(1-3p(1-p)\big) + 3N(N-1)p^2(1-p)^2 \le N^2$$

Beweisskizze. (a) Offensichtlich definiert die rechte Seite von (A.10) ein Maß. Um zu prüfen, dass es sich tatsächlich um ein Wahrscheinlichkeitsmaß handelt, setzen wir $f(x) \equiv 1$ ein: wegen $\mathbb{E}_p\big[(p_1 - p)^2\big] = p(1-p)/N$ und $\int_0^1 2(1-u)\,du = 1$ ist $\tilde{\mu}_{N,p}([0,1]) = 1$.

(b) Man kann $Y_{N,p} - Np = \sum_{i=1}^N (A_i - p)$ mit $A_1, \dots, A_n$ u.i.v. Bernoulli(p) darstellen und dann die Linearität des Erwartungswerts (und geeignete Gedanken zur Kombinatorik) ausnutzen. Alternativ könnte man die momentenerzeugende Funktion viermal ableiten.

Insbesondere liefert die Jensen-Ungleichung ($\mathbb{R}_+ \ni x \mapsto x^{3/4}$ is konkav) dann

$$\mathbb{E}\big[|Y_{N,p} - Np|^3\big] = \mathbb{E}\big[\big(|Y_{N,p} - Np|^4\big)^{3/4}\big] \le \Big(\mathbb{E}\big[|Y_{N,p} - Np|^4\big]\Big)^{3/4} \le \big(N^2\big)^{3/4} = N^{3/2}$$

□

A.5 Ein Steilkurs über Martingale in diskreter Zeit

Dieses Kapitel ist eine Einladung, sich (in sehr knapper Form) mit der Theorie der (zeitdiskreten) Martingale zu beschäftigen. Eine wesentlich gründlichere Behandlung findet sich beispielsweise bei Klenke [Kle20], speziell Kapitel 8–11 (das auch Lesern ans Herz gelegt sei, die an den im Text erwähnten „Übungen“ verzweifeln).

Beispiel A.8. Die symmetrische gewöhnliche Irrfahrt $S_n = X_1 + \cdots + X_n$, X_i u.i.v., $\mathbb{P}(X_i = \pm 1) = 1/2$ ($S_0 := 0$) wird uns hier als „Leib-und-Magen-Beispiel“ dienen.

A.5.1 Filtrationen

Zur Erinnerung: Das „übliche“ Grundobjekt der Wahrscheinlichkeitstheorie ist ein Wahrscheinlichkeitsraum $(\Omega, \mathscr{A}, \mathbb{P})$, bestehend aus Grundraum Ω, σ-Algebra $\mathscr{A}$ auf Ω und einem Wahrscheinlichkeitsmaß $\mathbb{P}$ auf $\mathscr{A}$. Eine *Filtration* $(\mathscr{F}_n)_n$ ist eine aufsteigend geordnete Familie von Teil-σ-Algebren, d.h. $\mathscr{F}_n \subset \mathscr{F}_{n+1} \subset \mathscr{A}$ für $n = 0, 1, 2, \dots$. Eine naheliegende (und nützliche) Interpretation ist, $\mathscr{F}_n$ als die Menge der bis zur Zeit n entschiedenen Ereignisse aufzufassen.

Beispiel A.9. Eine Folge von Zufallsvariablen (X_n) definiert via $\mathscr{F}_n := \sigma(X_0, X_1, \dots, X_n)$ eine Filtration (Übung: Überzeugen Sie sich davon).

A.5.2 Bedingte Erwartung

$\mathscr{G} \subset \mathscr{A}$ eine (Teil-)σ-Algebra. Wenn $\mathscr{G}$ endlich viele Atome $A_1, \dots, A_\ell$ hat, liegt es nahe, die „bedingte Erwartung von X, gegeben die Information aus $\mathscr{G}$“ folgendermaßen zu definieren:

$$\mathbb{E}[X|\mathscr{G}](\omega) = \frac{1}{\mathbb{P}(A_i)}\mathbb{E}[X 1_{A_i}] \quad \text{für } \omega \in A_i. \tag{A.11}$$

Man verallgemeinert (A.11) folgendermaßen: Eine reellwertige ZV Z heißt bedingte Erwartung von X gegeben $\mathscr{G}$ (schreibe $\mathbb{E}[X|\mathscr{G}]$), wenn gilt

1. Z ist $\mathscr{G}$-messbar, d.h. $\{Z \in B\} \in \mathscr{G}$ für jede messbare Teilmenge $B \subset \mathbb{R}$,
2. $\mathbb{E}[HZ] = \mathbb{E}[HX]$ für alle beschränkten $\mathscr{G}$-messbaren ZVn H.

(Übung: Überzeugen Sie sich, dass im Fall $|\Omega| < \infty$ die Version aus (A.11) diese Definition erfüllt.)

Bericht A.10. Für integrierbares X existiert die bedingte Erwartung $\mathbb{E}[X|\mathscr{G}]$ und ist bis auf f.s.-Gleichheit eindeutig bestimmt (Existenz beispielsweise via Projektion auf den Unterraum der (quadratintegrablen) $\mathscr{G}$-messbaren ZVn). Neben den „üblichen" Eigenschaften von Erwartungwerten (Linearität, Positivität) sind zwei wichtige Eigenschaften

1. $\mathbb{E}\big[\mathbb{E}[X|\mathscr{G}]\big|\mathscr{G}'\big] = \mathbb{E}[X|\mathscr{G}']$ wenn $\mathscr{G}' \subset \mathscr{G}$ („Turmeigenschaft"),
2. $\mathbb{E}[YX|\mathscr{G}] = YE[X|\mathscr{G}]$ sofern Y $\mathscr{G}$-messbar ist (und $\mathbb{E}[|XY|] < \infty$).

A.5.3 Martingale

Eine Folge integrierbarer Zufallsvariablen (M_n) (so dass M_n $\mathscr{F}_n$-messbar ist für $n = 0, 1, \dots$) heißt ein *Martingal* (bezüglich der Filtration $(\mathscr{F}_n)$), wenn

$$\mathbb{E}[M_n|\mathscr{F}_{n-1}] = M_{n-1} \text{ (f.s.) für } n = 1, 2, \dots \tag{A.12}$$

gilt. Es gilt dann auch $\mathbb{E}[M_n|\mathscr{F}_m] = M_m$ (f.s.) für $m \le n$ (Übung).

Die symmetrische Irrfahrt (Beispiel A.8) ist ein Martingal (Übung).

Bericht A.11. Wenn in (A.12) das „=" durch „≥" ersetzt wird, spricht man von einem *Submartingal*, wenn es durch „≤" ersetzt wird, von einem *Supermartingal*.

Prävisible Prozesse, „Spielstrategien", Gewinnprozesse als Martingale

$H_1, H_2, \dots$ eine Folge (individuell) beschränkter Zufallsvariablen, so dass H_i $\mathscr{F}_{i-1}$-messbar ist für $i = 1, 2, \dots$ (man nennt dann $(H_i)_{i\ge1}$ auch *prävisibel*), (M_n) ein Martingal. Dann ist auch die Folge (Y_n), definiert durch $Y_0 := 0$,

$$Y_n := \sum_{k=1}^{n} H_k(M_k - M_{k-1}), \quad n = 1, 2, \dots \tag{A.13}$$

ein Martingal (Übung). Wenn man (M_n) als den Gewinnprozess eines Spielers, der in jeder Runde einen „Einheitseinsatz" in einem fairen Spiel wettet, interpretiert, so ergibt dies für (Y_n) folgende Interpretation: Dies ist der Gewinnprozess eines Spielers, der jeweils vor der i-ten Runde den H_i-fachen Einheitseinsatz setzt. Die Bedingung, dass H_i $\mathscr{F}_{i-1}$-messbar sein muss, beschreibt einen Spieler ohne hellseherische Fähigkeiten: die Höhe des Einsatzes muss vor der Kenntnis des Ausgangs der i-ten Runde festgelegt werden.

A.5.4 Stoppzeiten

Eine Zufallsvariable τ mit Werten in $\{0, 1, \dots\}$ mit der Eigenschaft

$$\{\tau \le n\} \in \mathscr{F}_n, \quad n = 0, 1, 2, \dots. \tag{A.14}$$

heißt eine *Stoppzeit* (strenggenommen: $(\mathscr{F}_n)$-Stoppzeit). (A.14) läßt sich folgendermaßen interpretieren: Man kann zu jedem Zeitpunkt n entscheiden, ob τ „bereits eingetreten ist". Äquivalent kann man fordern, dass $\{\tau = n\} \in \mathscr{F}_n$ für alle n (Übung).

Für eine Stoppzeit τ ist die *τ-Vergangenheit* $\mathscr{F}_\tau$ gegeben durch $A \in \mathscr{F}_\tau :\iff A \cap \{\tau \le n\} \in \mathscr{F}_n$ für jedes n ($\mathscr{F}_\tau$ ist eine σ-Algebra, Übung).

Eine wichtige Klasse von Stoppzeiten erhält man mittels $\tau_A := \min\{k \in \mathbb{Z}_+ : X_k \in A\}$, wenn (X_n) eine $(\mathscr{F}_n)$-adaptierte Folge (sagen wir, reellwertiger) Zufallsvariablen und $A \subset \mathbb{R}$ (Überzeugen Sie sich, dass τ_A eine Stoppzeit ist). Sind τ_1, τ_2 Stoppzeiten, so auch $\tau_1 \wedge \tau_2$ und $\tau_1 \vee \tau_2$ (Übung). Warum ist mit τ stets auch $\tau + 5$ eine Stoppzeit, $\tau - 5$ aber im Allgemeinen nicht?

A.5.5 Optionales Stoppen

(M_n) Martingal, τ beschränkte Stoppzeit (d.h. es gibt eine Konstante T mit der Eigenschaft $\mathbb{P}(\tau \le T) = 1$). Dann gilt

Satz A.12 (Optional sampling-Satz, Basisversion)**.** $\mathbb{E}\big[M_\tau\big] = \mathbb{E}\big[M_0\big]$, *allgemeiner* $\mathbb{E}\big[M_\tau\big|\mathscr{F}_n\big] = M_{\tau\wedge n}$ *für* $n = 0, 1, \dots$.

Zum Beweis argumentieren Sie beispielsweise folgendermaßen: Überprüfen Sie zunächst, dass

$$M_\tau = \mathbb{E}\big[M_T\big|\mathscr{F}_\tau\big] \quad \text{fast sicher} \tag{A.15}$$

gilt. Tatsächlich gilt für $A \in \mathscr{F}_\tau$

$$\begin{aligned}\mathbb{E}\big[M_\tau \mathbf{1}_A\big] &= \mathbb{E}\left[\sum\nolimits_{k=0}^{T} M_k \mathbf{1}(\tau = k)\mathbf{1}_A\right] = \sum_{k=0}^{T} \mathbb{E}\big[M_k \mathbf{1}_{A\cap\{\tau=k\}}\big] = \sum_{k=0}^{T} \mathbb{E}\Big[\mathbb{E}\big[M_T\big|\mathscr{F}_k\big]\mathbf{1}_{A\cap\{\tau=k\}}\Big] \\ &= \sum_{k=0}^{T} \mathbb{E}\Big[\mathbb{E}\big[M_T \mathbf{1}_{A\cap\{\tau=k\}}\big|\mathscr{F}_k\big]\Big] = \sum_{k=0}^{T} \mathbb{E}\big[M_T \mathbf{1}_{A\cap\{\tau=k\}}\big] = \mathbb{E}\big[M_T \mathbf{1}_A\big],\end{aligned}$$

wobei an geeigneter Stelle (wo?) die Martingaleigenschaft $M_k = \mathbb{E}[M_T|\mathscr{F}_k]$, die Tatsache $A \cap \{\tau = k\} \in \mathscr{F}_k$ und $\mathbb{E}[\mathbb{E}[\,\cdot\,|\mathscr{F}_k]] = \mathbb{E}[\,\cdot\,]$ ausgenutzt werden. Aus (A.15) ergibt sich sofort die erste Behauptung (warum?).

Für die zweite Behauptung benutzen Sie die Turmeigenschaft der bedingten Erwartung beispielsweise folgendermaßen: $\tau \wedge n$ ist ebenfalls eine Stoppzeit, die offenbar $\tau \wedge n \le \tau$ ($\le T$) erfüllt. Überlegen Sie sich, dass dies $\mathscr{F}_{\tau\wedge n} \subset \mathscr{F}_\tau$ impliziert (ist das anschaulich einsichtig?). Demnach gilt mit (A.15)

$$M_{\tau\wedge n} = \mathbb{E}\big[M_T\big|\mathscr{F}_{\tau\wedge n}\big] = \mathbb{E}\Big[\mathbb{E}\big[M_T\big|\mathscr{F}_\tau\big]\Big|\mathscr{F}_{\tau\wedge n}\Big] = \mathbb{E}\big[M_\tau\big|\mathscr{F}_{\tau\wedge n}\big].$$

Bericht A.13. Man kann in Satz A.12 die Bedingung, dass τ beschränkt sein muss, fallen lassen. Technisch ist dann die entscheidende Bedingung, dass die Familie (M_n) *gleichgradig integrierbar* sein muss (Siehe [Kle20, Abschn. 10.3]). Ganz ohne Bedingungen kann Satz A.12 aber nicht richtig sein, wie die gewöhnliche Irrfahrt (Beispiel A.8) mit $\tau_{\{1\}} := \min\{n : S_n = 1\}$ zeigt: Wegen der Rekurrenz von (S_n) ist $\tau_{\{1\}} < \infty$ f.s., also $S_{\tau_{\{1\}}} = 1$, somit $\mathbb{E}[S_{\tau_{\{1\}}}] = 1 \neq \mathbb{E}[S_0] = 0$. (Für die Glücksspielinterpretation bedeutet dies: Man kann – im Prinzip – aus einem fairen Spiel sicheren Gewinn ziehen, wenn man ggfs. beliebig lange spielen und dabei beliebig hohe Schulden ansammeln darf.)

Bemerkung A.14. Aus Satz A.12 folgt, dass das *gestoppte* Martingal $(M_{\tau\wedge n})_n$ ebenfalls ein Martingal ist, wenn τ eine (beschränkte) Stoppzeit und (M_n) ein Martingal ist.

A.5.6 Konvergenz

Unter („leichten") Bedingungen konvergiert ein Martingal (M_n) fast sicher. Die auf Joseph Doob zurückgehende Beweisidee ist folgende: Wäre dies nicht der Fall, so gäbe es $a < b$, so dass (M_n) unendlich oft zwischen (unterhalb) a und (oberhalb) b oszilliert. Dann könnte man mit folgender Strategie beliebig großen Gewinn erzielen:

- Steige ein, sobald M_n unter a fällt,
- halte, bis M_n über b steigt.
- Erziele mindestens Gewinn $b - a > 0$ aus jeder solchen „Aufkreuzung".

Das widerspricht allerdings den Beobachtungen aus Abschnitt (A.5.3).

Wir wollen diese Idee nun präzisieren. Sei (M_n) ein nach unten beschränktes Martingal, o.E. $M_n \geq 0$ für alle n. (Warum ist die Annahme ≥ 0 keine zusätzliche Einschränkung?)

Seien $0 \leq a < b < \infty$. Setzen Sie $\sigma_0 := 0$,

$$\begin{aligned} \tau_k &:= \inf\{n > \sigma_{k-1} : M_n \leq a\}, \quad k = 1, 2, \ldots, \\ \sigma_k &:= \inf\{n > \tau_k : M_n \geq b\}, \quad k = 1, 2, \ldots. \end{aligned}$$

(Mit Verabredung $\tau_k = \infty$ bzw. $\sigma_k = \infty$, wenn es kein passendes n mehr gibt.)

Überzeugen Sie sich, dass die τ_k und σ_k Stoppzeiten sind. Betrachten Sie beispielsweise eine Skizze, um sich zu vergewissern, dass (M_n)

im Zeitintervall $\{\tau_k, \tau_{k+1}, \ldots, \sigma_k\}$ die k-te Aufkreuzung von (unterhalb) a nach (oberhalb) b ausführt (A.16)

(sofern $\tau_k, \sigma_k < \infty$). Sei

$$U_n^{(a,b)} := \max\{k : \sigma_k \leq n\}$$

die Anzahl abgeschlossener solcher Aufkreuzungen bis zum Zeitpunkt n.

Sei $I_0 := 0$, für $n \geq 1$

$$I_n := \sum_{i=0}^{n-1} \mathbf{1}(\exists\, k : \tau_k \leq i < \sigma_k)(M_{i+1} - M_i),$$

d.h. nur die Inkremente von (M_n) innerhalb der Aufkreuzungsintervalle zählen für (I_n). Verifizieren Sie, dass

$$\mathbb{E}\big[I_n\big|\mathscr{F}_{n-1}\big] = I_{n-1}$$

gilt, d.h. (I_n) ist (ebenfalls) ein Martingal.

Warum gilt

$$I_n \geq (b-a)U_n^{(a,b)} + \left(M_n - M_{\tau_{U_n^{(a,b)}+1}\wedge n}\right) \geq (b-a)U_n^{(a,b)} + (0-a)\,? \tag{A.17}$$

(Hinweis: Für jedes k ist $\sum_{i=\tau_k}^{\sigma_k-1}(M_{i+1}-M_i) = M_{\sigma_l} - M_{\tau_k} \geq (b-a)$.)

Lemma A.15 (Aufkreuzungsungleichung)**.** *Es gilt für jedes* n

$$\mathbb{E}\big[U_n^{(a,b)}\big] \leq \frac{a}{b-a}. \tag{A.18}$$

Offenbar $U_n^{(a,b)} \leq U_{n+1}^{(a,b)}$ für alle n, d.h. die Folge von Zufallsvariablen $(U_n^{(a,b)} : n \in \mathbb{N})$ konvergiert monoton gegen ein $U_\infty^{(a,b)}$, also auch

$$\mathbb{E}\big[U_\infty^{(a,b)}\big] = \lim_{n\to\infty}\mathbb{E}\big[U_n^{(a,b)}\big] \leq \frac{a}{b-a} < \infty$$

(benutzen Sie den Satz von der monotonen Konvergenz für das Gleichheitszeichen und dann (A.18) für die Abschätzung), insbesondere $\mathbb{P}\big(U_\infty^{(a,b)} < \infty\big) = 1$.

Betrachten Sie nun Ereignisse (mit $0 \leq a < b$, $a, b \in \mathbb{Q}$, sagen wir)

$$C^{(a,b)} := \left\{\liminf_{n\to\infty} X_n < a\right\} \cap \left\{\limsup_{n\to\infty} X_n > b\right\}.$$

Argumentieren Sie, dass $C^{(a,b)} \subset \{U_\infty^{(a,b)}\}$, folglich $P(C^{(a,b)}) = 0$ nach obigem, und daher auch

$$\mathbb{P}\left(\bigcup_{\substack{0\leq a<b\\ a,b\in\mathbb{Q}}} C^{(a,b)}\right) = 0 \tag{A.19}$$

gilt. Warum haben Sie damit folgende Version des Martingalkonvergenzsatzes bewiesen?

Satz A.16. *Ein nach unten beschränktes Martingal konvergiert mit Wahrscheinlichkeit* 1.

Bericht A.17. Die Konvergenz $M_n \to M_\infty$ f.s. muss i.A. nicht die Konvergenz der Erwartungwerte implizieren: Betrachten Sie beispielsweise die Irrfahrt aus Beispiel A.8, die beim Auftreffen auf -1 gestoppt wird. Für gleichgradig integrierbare Martingale gilt allerdings auch $\mathbb{E}[M_n] \to \mathbb{E}[M_\infty]$.

A.5.7 Doobsche Ungleichung

Im Allgemeinen ist es sehr schwierig, aus der Verteilung eines stochastischen Prozesses zu festen Zeiten Informationen über das Pfadverhalten wie beispielsweise das laufende Maximum abzuleiten. Im Fall von Martingalen sind die Verhältnisse übersichtlicher:

Satz A.18 (Doobs L^2-Ungleichung)**.** *Sei* (M_n) *Martingal mit* $M_0 \geq 0$ *und* $\mathbb{E}[M_n^2] < \infty$ *für alle* n*,* $M_n^* := \max_{0\leq k\leq n} M_k$*. Dann gilt*

$$\mathbb{E}\big[(M_n^*)^2\big] \leq 4\mathbb{E}\big[M_n^2\big].$$

Für festes $\lambda > 0$ gilt

$$\lambda\mathbb{P}(M_n^* \geq \lambda) \leq \mathbb{E}\big[M_n \mathbf{1}(M_n^* \geq \lambda)\big] \Big(\leq \mathbb{E}\big[|M_n| \mathbf{1}(M_n^* \geq \lambda)\big]\Big). \tag{A.20}$$

Argumentieren Sie beispielsweise folgendermaßen: $\tau := \inf\{k : M_k \geq \lambda\} \wedge n$ ist eine (durch n) beschränkte Stoppzeit, also

$$\begin{aligned}\mathbb{E}[M_n] = \mathbb{E}[M_\tau] &= \mathbb{E}\big[M_\tau \mathbf{1}(M_n^* \geq \lambda)\big] + \mathbb{E}\big[M_\tau \mathbf{1}(M_n^* < \lambda)\big] = \mathbb{E}\big[M_\tau \mathbf{1}(M_n^* \geq \lambda)\big] + \mathbb{E}\big[M_n \mathbf{1}(M_n^* < \lambda)\big] \\ &\geq \lambda\mathbb{P}(M_n^* \geq \lambda) + \mathbb{E}\big[M_n \mathbf{1}(M_n^* < \lambda)\big].\end{aligned}$$

Nun substrahiere $\mathbb{E}\big[M_n \mathbf{1}(M_n^* < \lambda)\big]$ auf beiden Seiten.

Stets gilt

$$(M_n^*)^2 = \int_0^{M_n^*} 2\lambda \, d\lambda,$$

also (wegen $(M_n^*)^2 \leq M_0^2 + M_1^2 + \cdots + M_n^2$ ist der Erwartungswert endlich)

$$\begin{aligned}\mathbb{E}\big[(M_n^*)^2\big] &= \mathbb{E}\left[\int_0^{M_n^*} 2\lambda \, d\lambda\right] = \mathbb{E}\left[\int_0^{\infty} 2\lambda \mathbf{1}(M_n^* \geq \lambda) \, d\lambda\right] = 2\int_0^{\infty} \lambda\mathbb{P}(M_n^* \geq \lambda) \, d\lambda \\ &\leq 2\int_0^{\infty} \mathbb{E}\big[|M_n| \mathbf{1}(M_n^* \geq \lambda)\big] \, d\lambda = 2\mathbb{E}\left[|M_n| \int_0^{M_n^*} d\lambda\right] = 2\mathbb{E}\big[|M_n| M_n^*\big]\end{aligned}$$

Folgern Sie mit der Cauchy-Schwarz-Ungleichung:

$$\mathbb{E}\big[(M_n^*)^2\big] \leq 2\sqrt{E\big[|M_n|^2\big]}\sqrt{\mathbb{E}\big[(M_n^*)^2\big]}.$$

Bericht A.19. Die Ungleichung gilt wörtlich auch für Submartingale. Die Annahme $M_0 \geq 0$ ist eigentlich nicht notwendig (vereinfacht hier nur die Argumentation ein wenig). Es gilt eine analoge Aussage für jedes $p > 1$ statt $p = 2$ (Doobs L^p-Ungleichung).

A.5.8 Die symmetrische gewöhnliche Irrfahrt auf einem Intervall

Seien $a, b \in \mathbb{Z}$, $a < x < b$, $Z^{(x,a,b)}$ die symmetrische gewöhnliche Irrfahrt startend in $Z_0^{(x,a,b)} = x$, gestoppt, sobald $Z_0^{(x,a,b)} \in \{a, b\}$. Prüfen Sie: $((b - Z_n^{(x,a,b)})/(b-a))_n$ und $((b - Z_n^{(x,a,b)})(Z_n^{(x,a,b)} - a) - n)_n$ sind Martingale. Können Sie diese Information benutzen, um die Wahrscheinlichkeit, dass der obere Rand getroffen wird, sowie die erwartete Zeit bis zum Treffen des Rands zu berechnen?