Kapitel II
Das Wright-Fisher-Modell

Das Hardy-Weinberg-Modell aus dem letzten Kapitel basiert im Grunde auf der Annahme einer unendlichen Population, auf deren Genotyp-Frequenzen zufällige Kreuzung wie ein deterministischer Operator agiert mit der Konsequenz, daß sich die Genotyp-Frequenzen schon nach einer Generation stabilisieren. In diesem Abschnitt wollen wir uns dem Effekt einer endlichen Populationsgröße als einem Aspekt des Evolutionsverhaltens zuwenden. Er wurde von Fisher, Haldane und Wright in nicht unbeträchtlichem Umfang untersucht. Diese Endlichkeit impliziert, daß Veränderungen der Genotyp-Frequenzen als Teil eines stochastischen statt deterministischen Prozesses angesehen werden müssen. Um die Bedeutung des stochastischen Faktors theoretisch bewerten zu können, benötigen wir zuerst ein Modell, welches das Verhalten einer Genpopulation im stochastischen Fall geeignet beschreibt. Die Wahl eines solchen Modell ist vielleicht mehr als in irgendeinem anderen Bereich der Theorie ziemlich willkürlich, und wir wollen nicht vorgeben, daß eines der hier gewählten Modelle die Natur tatsächlich in guter Näherung abbildet. Fisher und Wright bedienten sich bei der Analyse ihres Modells, das wir im Anschluß vorstellen wollen, Methoden aus der Theorie der Markov-Ketten sowie einer nahen Verwandten, der Theorie von Diffusionsprozessen, auch wenn ihre Terminologie dies nicht ausdrücklich ausweist.

1. Das Grundmodell

Wir beginnen mit der Beschreibung des einfachsten, *haploiden Modells zufälliger Reproduktion ohne Mutation und Selektion* und betrachten eine endliche Population von N Individuen hinsichtlich eines bestimmten Gens, das wieder entweder vom Typ A_1 oder A_2 ist. Ausgehend von der Elternpopulation, ergibt sich die nächste Generation durch N-maliges Ziehen mit Zurücklegen bei zwei möglichen Ausgängen (Bernoulli-Experimente): Besteht die Elternpopulation aus k Typ-A_1-Allelen und $N-k$ Typ-A_2-Allelen, so ergibt jede Ziehung mit der Wahrscheinlichkeit

$$\alpha_k = \frac{k}{N} \quad \text{bzw.} \quad \beta_k = 1 - \frac{k}{N}$$

ein Typ-A_1 bzw. Typ-A_2-Allel der Tochterpopulation. Die Evolution der Population unter diesem ad infinitum fortgesetzten Mechanismus läßt sich durch

$$M_n \;=\; \sum_{k=1}^{N} X_{n,k}, \quad n \geq 1 \tag{1.1}$$

beschreiben, wobei M_n die Anzahl der Typ-A_1-Allele der n-ten Generation bezeichnet und $X_{n,k}$ das Ergebnis des k-ten Bernoulli-Experiments zur Erzeugung dieser Generation angibt. Die $X_{n,k}, 1 \le k \le N$, sind bedingt unter $M_0, ..., M_{n-1}$ stochastisch unabhängig und identisch $B(1, \alpha_{M_{n-1}})$-verteilt, d.h.

$$P^{(X_{n,1},...,X_{n,N})|M_0=i_0,...,M_{n-1}=i_{n-1}} \;=\; P^{(X_{n,1},...,X_{n,N})|M_{n-1}=i_{n-1}} \;=\; B(1,\alpha_{i_{n-1}})^N. \tag{1.2}$$

Die $X_{n,k}$ hängen also von $M_0, ..., M_{n-1}$ nur über M_{n-1} ab, und vermöge (1.2) folgt weiter

$$P^{M_n|M_0=i_0,...,M_{n-1}=i_{n-1}} \;=\; P^{M_n|M_{n-1}=i_{n-1}} \;=\; B(N,\alpha_{i_{n-1}}). \tag{1.3}$$

$(M_n)_{n\ge 0}$ bildet demnach eine EMK mit Zustandsraum $\{0, ..., N\}$ und Übergangswahrscheinlichkeiten

$$p_{ij} \;=\; \binom{N}{j}\alpha_i^j\beta_i^{N-j}, \quad i,j = 0, ..., N. \tag{1.4}$$

Da hieraus offenkundig

$$E(M_n|M_{n-1}=i) \;=\; N\cdot\frac{i}{N} \;=\; i \tag{1.5}$$

für alle $i = 0, ..., N$ folgt, definiert $(M_n)_{n\ge 0}$ außerdem unter jeder Anfangsverteilung ein (beschränktes) Martingal und konvergiert folglich f.s. und im Mittel gegen einen Limes M_∞. Die mittlere Allelfrequenzen $\frac{E_iM_n}{N}$ für A_1 und $1 - \frac{E_iM_n}{N}$ für A_2 bleiben somit wie im Hardy-Weinberg-Modell über alle Generationen konstant. Aber Vorsicht! Dies sagt nichts über die Fluktuationen aus. Die Zustände 0 und N, in denen die Population nur noch Gene vom Typ A_1 bzw. A_2 enthält, sind absorbierend, und wir werden schon bald sehen, daß – im krassen Gegensatz zum Hardy-Weinberg-Modell – Absorption unter jeder Anfangsverteilung mit Wahrscheinlichkeit 1 irgendwann eintritt (☞ Satz 3.1). Man spricht in diesem Fall von *Fixierung*. Für eine Diskussion über die biologische Rechtfertigung der hier gemachten Voraussetzungen verweisen wir auf Fisher (1930).

2. Das Wright-Fisher-Modell mit Mutationseffekten

Die Berücksichtigung von Mutationseffekten läßt sich etwa durch folgende Modellerweiterung bewerkstelligen: Vor Bildung einer neuen Generation hat jedes Allel die Chance zu mutieren, d.h., sich in ein Allel der anderen Art zu verwandeln. Wir nehmen an, daß eine Mutation $A_1 \to A_2$ mit Wahrscheinlichkeit γ_1 und eine Mutation $A_2 \to A_1$ mit Wahrscheinlichkeit γ_2 geschieht. Es gelten dann weiter (1.1)–(1.4), jedoch mit den neuen Ziehungswahrscheinlichkeiten

$$\begin{aligned}\alpha_k &= \frac{k}{N}(1-\gamma_1) + \left(1-\frac{k}{N}\right)\gamma_2,\\ \beta_k &= \frac{k}{N}\gamma_1 + \left(1-\frac{k}{N}\right)(1-\gamma_2).\end{aligned} \tag{2.1}$$

Zur genaueren Erläuterung schlüsseln wir den Mechanismus weiter auf: Wir nehmen an, daß Mutation der Ziehung nachgeschaltet ist. Sei $Y_{n,k} = 1$ bzw. $= 0$, falls das k-te aus der $(n-1)$-ten Generation selektierte Gen vor Auftreten einer möglichen Mutation vom Typ A_1 bzw. A_2 ist. Die $Y_{n,1}, ..., Y_{n,N}$ erfüllen dann (1.2), sind also bedingt unter M_{n-1} unabhängig und jeweils $B(1, M_{n-1}/N)$-verteilt. Seien weiter $I_{n,k}, J_{n,k}$ unabhängige (auch von M_{n-1} und den $Y_{n,k}$) Bernoulli-Variablen, $I_{n,k} \sim B(1, \gamma_1)$, $J_{n,k} \sim B(1, \gamma_2)$, und setze

$$X_{n,k} \;=\; Y_{n,k}(1 - I_{n,k}) + (1 - Y_{n,k})J_{n,k}. \tag{2.2}$$

$I_{n,k} = 1$ bedeutet demnach eine Mutation $A_1 \to A_2$ des k-ten gezogenen Gens der $(n-1)$-ten Generation und $J_{n,k} = 1$ eine Mutation $A_2 \to A_1$. Wie man sofort einsieht, erfüllen auch hier die $X_{n,k}$ (1.2), jedoch mit den α_k aus (2.1). Es gilt nämlich unter Beachtung der Unabhängigkeitsannahmen

$$\begin{aligned} &P(X_{n,k} = 1 | M_{n-1} = i) \\ &\quad= \; P(I_{n,k} = 0, Y_{n,k} = 1 | M_{n-1} = i) + P(J_{n,k} = 1, Y_{n,k} = 0 | M_{n-1} = i) \\ &\quad= \; P(Y_{n,k} = 1 | M_{n-1} = i) P(I_{n,k} = 0) + P(Y_{n,k} = 0 | M_{n-1} = i) P(J_{n,k} = 1) \\ &\quad= \; \frac{i}{N}(1 - \gamma_1) + \left(1 - \frac{i}{N}\right)\gamma_2 \;=\; \alpha_i \end{aligned}$$

für alle $k = 1, ..., N$.

Sofern $\gamma_1 \gamma_2 > 0$, tritt offenbar in keinem Zustand Fixierung ein. Stattdessen strebt M_n in diesem Fall für $n \to \infty$ in Verteilung gegen einen stationären Limes ξ, den wir als *Genfrequenz im Gleichgewicht* bezeichnen.

3. Das Wright-Fisher-Modell mit Selektionsdruck

Wir kehren zurück zum Grundmodell und wollen für dieses als weitere Variante das Konzept eines Selektionsdrucks zugunsten von, sagen wir, Typ-A_1-Allelen diskutieren. Es sei zunächst bemerkt, daß im Grundmodell (*neutrale Selektion*) die mittleren Reproduktionsraten $r_n = \frac{E(M_n | M_{n-1})}{M_{n-1}}$ und $R_n = \frac{E(N - M_n | M_{n-1})}{N - M_{n-1}}$ für beide Alleltypen stets 1 betragen, wie (1.5) zeigt. Stellen wir uns nun vor, daß der Ziehungsmechanismus Allelen vom Typ A_1 gegenüber denen vom Typ A_2 einen mittleren selektiven Vorteil gewährt, präzisiert durch $r_n = (1+s)R_n$ für alle $n \geq 1$ und ein $s > 0$ (klein). Gesucht sind also Selektionswahrscheinlichkeiten α_k, β_k, die dieses gewährleisten. Da weiterhin $E(M_n | M_{n-1} = i) = N\alpha_i$ für alle $i = 0, ..., N$ gilt, ergeben sich $\alpha_k, \beta_k = 1 - \alpha_k$ vermöge

$$r_n \;=\; \frac{N\alpha_{M_{n-1}}}{M_{n-1}} \;=\; (1+s)\frac{N(1 - \alpha_{M_{n-1}})}{N - M_{n-1}} \;=\; (1+s)R_n, \quad n \geq 1,$$

eindeutig zu

$$\alpha_k = \frac{(1+s)k}{N + sk} \quad \text{und} \quad \beta_k = \frac{N-k}{N+sk}. \tag{3.1}$$

Der Quotient der erwarteten Populationsgrößen von Typ-A_1- und Typ-A_2-Allelen in der n-ten Generation (bedingt unter M_{n-1}) ergibt sich zu

$$\begin{aligned}\frac{E(M_n|M_{n-1})}{E(N-M_n|M_{n-1})} &= \frac{\alpha_{M_{n-1}}}{\beta_{M_{n-1}}} = \frac{(1+s)M_{n-1}}{N-M_{n-1}} \\ &= \left(\frac{1+s}{1}\right)\left(\frac{\text{Anzahl von Typ-}A_1\text{-Genen in der }(n-1)\text{-ten Generation}}{\text{Anzahl von Typ-}A_2\text{-Genen in der }(n-1)\text{-ten Generation}}\right)\end{aligned}$$

und verdeutlicht auf alternative Weise die Bedeutung von Selektion. Beachte, daß Zustände 0 und N auch unter Selektionsdruck absorbierend sind. Eine wichtige Frage lautet demnach auch hier, mit welcher Wahrscheinlichkeit bedingt unter $M_0 = k$ Fixierung eintritt.

4. Die Heterozygosität einer Wright-Fisher-Population

Wir kehren zurück zum Grundmodell und werfen als nächstes einen genaueren Blick auf die Wahrscheinlichkeiten für Fixierung des einen oder anderen Alleltyps, die bezogen auf die EMK $(M_n)_{n\geq 0}$ den Absorptionswahrscheinlichkeiten $q_{i,0}$ und $q_{i,N}$ im Zustand 0 bzw. N unter P_i, $i = 0, ..., N$, entsprechen. Gemäß (1.5) gilt $E(M_n|M_{n-1}) = M_{n-1}$ f.s. für alle $n \geq 1$, d.h. $(M_n)_{n\geq 0}$ bildet (unter jeder Anfangsverteilung) ein beschränktes Martingal und konvergiert folglich f.s. gegen einen Limes M_∞. Mittels dieser Beobachtung und dem Nachweis, daß $P_i(M_\infty = 0 \text{ oder } N) = 1$ für alle $i = 0, ..., N$ gilt, lassen sich q_0 und q_N leicht berechnen.

4.1. Satz. *Unter den gegebenen Annahmen gilt*

$$q_{i,N} = 1 - q_{i,0} = \frac{i}{N} \tag{4.1}$$

für jedes $i = 0, ..., N$.

Beweis: Sei $i \in \{0, ..., N\}$ beliebig vorgegeben. Da alle M_n denselben endlichen Wertebereich besitzen und für $n \to \infty$ P_i-f.s. gegen M_∞ konvergieren, folgt

$$P_i(M_n = M_\infty \text{ für fast alle } n \geq 0) = 1.$$

Dies bedeutet aber, daß die Werte von M_∞ nur absorbierende Zustände sein können, also 0 oder N. Vermöge der Martingaleigenschaft und der trivialerweise vorliegenden gleichgradigen Integrierbarkeit von $(M_n)_{n\geq 0}$ erhalten wir nun

$$Nq_{i,N} = E_i M_\infty = E_i M_0 = i$$

und folglich das Gewünschte. ♢

Im einfachen Wright-Fisher-Modell nimmt also die genetische Variabilität im Zeitablauf ab und verschwindet schließlich ganz. In diesem Zusammenhang ist auch ein Blick auf die

Varianz von M_n interessant. Wir erinnern an die Notation $\alpha_k = \frac{k}{N} = 1 - \beta_k$ für $k = 1, ..., N$ und setzen außerdem $\kappa \stackrel{\text{def}}{=} 1 - \frac{1}{N}$.

4.2. Lemma. *Es gilt unter jeder Anfangsverteilung λ und für alle $n \geq 0$*

$$\mathrm{Var}_\lambda M_n \;=\; (1-\kappa^n) E_\lambda M_0 (N - E_\lambda M_0) \;+\; \kappa^n \mathrm{Var}_\lambda M_0 \tag{4.2}$$

und somit insbesondere

$$\mathrm{Var}_i M_n \;=\; (1-\kappa^n)\, i\, (N-i) \tag{4.3}$$

für jedes $i = 0, ..., N$.

Beweis: Wir benutzen die allgemeine Varianzformel

$$\mathrm{Var}_\lambda M_n \;=\; E_\lambda(\mathrm{Var}(M_n|M_{n-1})) \;+\; \mathrm{Var}_\lambda(E(M_n|M_{n-1}))$$

zur Herleitung einer Rekursionsformel. Unter Beachtung von $E(M_n|M_{n-1}) = M_{n-1}$ f.s. und $P^{M_n|M_{n-1}} = B(N, \alpha_{M_{n-1}})$ ergibt sich

$$\begin{aligned}
\mathrm{Var}_\lambda M_n &= E_\lambda(N\alpha_{M_{n-1}}\beta_{M_{n-1}}) \;+\; \mathrm{Var}_\lambda M_{n-1} \\
&= (1-\kappa) E_\lambda(M_{n-1}(N - M_{n-1}) \;+\; \mathrm{Var}_\lambda M_{n-1} \\
&= (1-\kappa)\Big(N E_\lambda M_{n-1} - \mathrm{Var}_\lambda M_{n-1} - (E_\lambda M_{n-1})^2\Big) \;+\; \mathrm{Var}_\lambda M_{n-1} \\
&= (1-\kappa) E_\lambda M_0 (N - E_\lambda M_0) \;+\; \kappa \mathrm{Var}_\lambda M_{n-1}
\end{aligned}$$

für $n \geq 1$, und daraus folgert man leicht

$$\begin{aligned}
\mathrm{Var}_\lambda M_n &= \kappa^n \mathrm{Var}_\lambda M_0 \;+\; (1-\kappa) \sum_{k=0}^{n-1} \kappa^k E_\lambda M_0 (N - E_\lambda M_0) \\
&= \kappa^n \mathrm{Var}_\lambda M_0 \;+\; (1-\kappa^n) E_\lambda M_0 (N - E_\lambda M_0),
\end{aligned}$$

wie behauptet. ♢

Von nicht unerheblichem Interesse ist die Frage, wie schnell Fixierung in einer Wright-Fisher-Population eintritt. Die Anfangsverteilung λ sei im folgenden beliebig. Eine einfache Rechnung unter Benutzung von (4.2) zeigt

$$\begin{aligned}
E_\lambda M_n(N - M_n) &= E_\lambda M_0 (N - E_\lambda M_0) \;-\; \mathrm{Var}_\lambda M_n \\
&= \kappa^n \Big(E_\lambda M_0 (N - E_\lambda M_0) - \mathrm{Var}_\lambda M_0 \Big) \\
&= \kappa^n E_\lambda M_0 (N - M_0)
\end{aligned} \tag{4.4}$$

für alle $n \geq 0$. Multipliziert man beide Seiten mit $\frac{2}{N^2}$, erhält man offenkundig

$$h_\lambda(n) \;\stackrel{\text{def}}{=}\; E(2\alpha_{M_n}\beta_{M_n}) \;=\; \kappa^n h_\lambda(0) \tag{4.5}$$

für alle $n \geq 0$. $h_\lambda(n)$ wird als *Heterozygosität* der Population in Generation n (unter der Anfangsverteilung λ) bezeichnet und entspricht der Wahrscheinlichkeit, daß zwei zufällig ausgewählte Gene dieser Generation verschiedenen Typs sind. Gemäß (4.5) nimmt die Heterozygosität der Population also ungeachtet der Anfangsverteilung geometrisch schnell ab.

5. Die mittlere Absorptionszeit: Eine Diffusionsapproximation

Wenden wir uns der mittleren Absorptionszeit m_i unter P_i zu, d.h.

$$m_i \stackrel{\text{def}}{=} E_i\tau^A, \quad i = 0, ..., N,$$

mit $\tau^A \stackrel{\text{def}}{=} \inf\{n \geq 0 : M_n = 0 \text{ oder } = N\}$. Offenbar gilt $m_0 = m_N = 0$. Für $1 \leq i \leq N$ erhält man unter Benutzung der Markov-Eigenschaft

$$m_i \;=\; p_{i,0} \cdot 1 \;+\; p_{i,N} \cdot 1 \;+\; \sum_{j=1}^{N-1} p_{i,j}(1 + m_j) \;=\; 1 \;+\; \sum_{j=0}^{N} p_{i,j} m_j. \tag{5.1}$$

Da es unmöglich scheint, die allgemeine Lösung dieses Gleichungssystems explizit zu berechnen, müssen wir nach einer Approximation für große N suchen, was uns in die Welt der Diffusionen führt.

Anstelle der Anzahl von Typ-A_1-Allelen M_n betrachtet man gewöhnlich die zugehörige Allelfrequenz $X_N(n) \stackrel{\text{def}}{=} M_n/N$. Aus Zweckmäßigkeitsgründen erweitern wir den Zeitbereich von $\mathbb{N}_0$ auf $[0, \infty)$, indem wir $X_N(t) \stackrel{\text{def}}{=} X_N(n)$ für $t \in [n, n+1)$ und $n \in \mathbb{N}_0$ definieren. Um zu einem nicht degenerierten Limes zu gelangen, müssen wir sowohl den Raum als auch die Zeit geeignet skalieren, was zu dem Prozeß

$$Y_N(t) \stackrel{\text{def}}{=} X_N(Nt), \quad t \geq 0 \tag{5.2}$$

führt. Die Idee besteht darin, daß $Y_N(\cdot)$ für $N \to \infty$ in Verteilung gegen einen Grenzprozeß $Y(\cdot)$ in stetiger Zeit mit dem stetigen Zustandsraum $[0, 1]$ konvergiert. $Y(\cdot)$ bildet ein Beispiel eines Markov-Prozesses in stetiger Zeit mit stetigen Pfaden, genannt *Diffusionsprozeß* oder einfach *Diffusion*. Zeitskalierungen in Einheiten proportional zur Populationsgröße N sind typisch für populationsgenetische Modelle, und Diffusionstheorie gehört zu den Standardwerkzeugen in der Populationsgenetik. Referenzen??????

Die Eigenschaften einer eindimensionalen quadratisch integrierbaren Diffusion $Y(\cdot)$ sind im wesentlichen durch die infinitesimale Drift und Varianz bestimmt, im zeitlich homogenen Fall gegeben durch

$$\mu(y) \stackrel{\text{def}}{=} \lim_{h \to 0} h^{-1} E(Y(t+h) - Y(t) | Y(t) = y),$$
$$\sigma^2(y) \stackrel{\text{def}}{=} \lim_{h \to \infty} h^{-1} E((Y(t+h) - Y(t))^2 | Y(t) = y).$$

Da im zeitdiskreten Wright-Fisher-Modell M_{n+1} bedingt unter $M_n = i$ eine $B(N, \frac{i}{N})$-Verteilung besitzt, folgt

$$\begin{aligned} E\Big(X_N(n+1) - X_N(n)\Big|X_N(n) = \frac{i}{N}\Big) &= 0, \\ E\Big(\Big(X_N(n+1) - X_N(n)\Big)^2\Big|X_N(n) = \frac{i}{N}\Big) &= \frac{1}{N}\,\frac{i}{N}\Big(1 - \frac{i}{N}\Big), \end{aligned}$$

für alle $n \geq 0$. Setzen wir $h_N \stackrel{\text{def}}{=} \frac{1}{N}$, $n = 0$ und wählen $i = i_N$ zu beliebigem $y \in [0,1]$ derart, daß $y_N \stackrel{\text{def}}{=} \frac{i_N}{N} \to y$, so ergibt sich

$$\mu(y) = \lim_{N\to\infty} h_N^{-1} E(Y_N(h_N) - Y_N(0)|Y_N(0) = y_N) = 0,$$

und

$$\begin{aligned} \sigma^2(y) &= \lim_{N\to\infty} h_N^{-1} E((Y_N(h_N) - Y_N(0))^2|Y_N(0) = y_N) \\ &= \lim_{N\to\infty} h_N^{-1} h_N y_N(1 - y_N) \\ &= y(1-y). \end{aligned}$$

Betrachten wir nun die mittlere Absorptionszeit $m(y)$ in 0 oder 1 für den Diffusionsprozeß $Y(\cdot)$ unter $P(\cdot|Y(0) = y)$, also

$$m(y) \stackrel{\text{def}}{=} E(T^A|Y(0) = y) \quad \text{mit} \quad T^A \stackrel{\text{def}}{=} \inf\{t \geq 0 : Y(t) = 0 \text{ oder } = 1\}.$$

Sofern auch für diese Funktion ein stetiger Übergang vom zeitdiskreten zum zeitstetigen Modell möglich ist, was wir im folgenden unterstellen, läßt sich eine Approximation von $m(y)$ folgendermaßen herleiten. Setzen wir

$$\tau_N^A \stackrel{\text{def}}{=} \inf\{n \geq 0 : X_N(n) = 0 \text{ oder } = 1\}$$

und

$$T_N^A \stackrel{\text{def}}{=} \inf\{t \geq 0 : Y_N(t) = 0 \text{ oder } = 1\},$$

so folgt $m(y) = \lim_{N\to\infty} E(T_N^A|Y(0) = y)$ und

$$\tau_N^A = \inf\{t \geq 0 : X_N(t) = 0 \text{ oder } = 1\} = \frac{T_N^A}{h_N}.$$

Sei $S_N(i)$ im folgenden eine $B(N, \frac{i}{N})$-verteilte Zufallsgröße. Vermöge (5.1) gilt dann für $m_N(y_N) \stackrel{\text{def}}{=} E(T_N^A|Y_N(0) = y_N) = h_N m_{i_N}$

$$m_N(y_N) = h_N + \sum_{j=0}^{N} p_{i_N,j}(N) m_N\Big(\frac{j}{N}\Big) = h_N + E m_N\Big(\frac{S_N(i_N)}{N}\Big).$$

Für unsere weiteren Überlegungen setzen wir voraus, daß $m_N(y_N) = m(y) + o(h_N)$ für $N \to \infty$ gilt und $m(\cdot)$ zweimal stetig differenzierbar ist. Dann ergibt sich mittels der obigen Gleichung

und einer Taylor-Entwicklung 2. Ordnung in y_N

$$\begin{aligned} m(y_N) &= h_N + Em\left(\frac{S_N(i_N)}{N}\right) + o(h_N) \\ &= h_N + m(y_N) + \frac{m''(y_N)}{2}\mathrm{Var}\left(\frac{S_N(i_N)}{N}\right) + o(h_N) \\ &= h_N + m(y_N) + \frac{m''(y_N)}{4N}y_N(1-y_N) + o(h_N) \end{aligned}$$

und daraus weiter nach Multiplikation mit N und anschließendem Grenzübergang $N \to \infty$ die Differentialgleichung

$$m''(y)y(1-y) = -2, \quad 0 < y < 1. \tag{5.3}$$

mit den offensichtlichen Randbedingungen $m(0) = m(1) = 0$. Diese läßt sich mit einer Partialbruchentwicklung und zweimaliger partieller Integration leicht lösen und führt schließich zu folgendem Ergebnis:

5.1. Satz. *Für die mittlere Absorptionszeit $m(y) = E(T^A|Y(0) = y)$ der Wright-Fisher-Diffusion $(Y(t))_{t\geq 0}$ gilt*

$$m(y) = -2(y\log y + (1-y)\log(1-y)), \quad 0 \leq y \leq 1. \tag{5.4}$$

Zurückkehrend zum zeitdiskreten Modell, erhalten wir vermöge der approximativen Beziehung $m_i \approx Nm(\frac{i}{N})$ die folgenden Ergebnisse: Falls $\frac{i}{N} = \frac{1}{2}$, gilt

$$m_N \approx Nm\left(\frac{1}{2}\right) = (2\log 2)N \approx 1.39 \cdot N \text{ Generationen},$$

während $i = 1$

$$m_1 \approx Nm(h_N) = 2(\log N + 1) + o(1) \text{ Generationen}$$

für die erwartete Zeit bis zur Fixierung liefert.

6. Die Genealogie des Wright-Fisher-Modells

Als nächstes betrachten wir das Wright-Fisher-Modell aus der genealogischen Perspektive. In Abwesenheit von Rekombination bildet die DNS-Sequenz, die das betrachtete Gen repräsentiert, eine Kopie einer Sequenz der vorherigen Generation, die ihrerseits die Kopie einer Sequenz der vorherigen Generation darstellt, usw. Jede DNS-Sequenz kann somit als 'Individuum' aufgefaßt werden, das eine 'Mutter' (nämlich die Sequenz, von der sie kopiert wurde) sowie eine zufällige Anzahl von 'Nachkommen' (nämlich die Sequenzen, die von ihr als Kopien in die nachfolgende Generation eingehen) besitzt.

Um diesen Prozeß entweder zeitlich vorwärts oder zeitlich rückwärts zu studieren, markieren wir die Individuen einer gegebenen Generation mit $1, 2, ..., N$ und bezeichnen mit ν_i die

Anzahl der Nachkommen des Individuums i, $1 \leq i \leq N$. Wir nehmen an, daß die ν_i unabhängig und jeweils Poisson-verteilt sind, und betrachten die gemeinsame Verteilung von $(\nu_1, ..., \nu_N)$ unter der Bedingung, daß die Gesamtzahl an Nachkommen $\nu_1 + ... + \nu_N$ wieder N beträgt. Das folgende einfache Lemma zeigt, daß dies zu einer Multinomialverteilung führt, die gar nicht vom Parameter der gegebenen Poisson-Verteilung abhängt.

6.1. Lemma. *Sei $\theta > 0$ beliebig. Gegeben stochastisch unabhängige, jeweils Poi(θ)-verteilte Zufallsgrößen $\nu_1, ..., \nu_N$, gilt*

$$P^{(\nu_1,...,\nu_N)|\nu_1+...+\nu_N=k} = M(k, \tfrac{1}{N}, ..., \tfrac{1}{N}) \tag{6.1}$$

für alle $k \in \mathbb{N}$ und ferner

$$P^{(\nu_{\pi(1)}+...+\nu_{\pi(i)})|\nu_1+...+\nu_N=k} = B(k, \tfrac{i}{N}) \tag{6.2}$$

für jede i-Permutation $(\pi(1), ..., \pi(i))$ von $1, ..., N$. Insbesondere besitzt jedes ν_i bedingt unter $\nu_1 + ... + \nu_N = k$ eine $B(k, \frac{1}{N})$-Verteilung.

Beweis: Für $(m_1, ..., m_N) \in \mathbb{N}_0^N$ mit $m_1 + ... + m_N = k \in \mathbb{N}$ erhalten wir wegen $\nu_1 + ... + \nu_N \stackrel{d}{=} Poi(N\theta)$

$$\begin{aligned} &P(\nu_1 = m_1,..., \nu_N = m_N | \nu_1 + ... + \nu_N = k) \\ &\quad = \frac{P(\nu_1 = m_1, ..., \nu_N = m_N)}{P(\nu_1 + ... + \nu_N = k)} \\ &\quad = \frac{e^{-N\theta} \prod_{j=1}^{N} \frac{\theta^{m_j}}{m_j!}}{e^{-N\theta} \frac{(N\theta)^k}{k!}} \\ &\quad = \frac{k!}{m_1! \cdot ... \cdot m_N!} \left(\frac{1}{N}\right)^k \end{aligned}$$

und folglich (6.1). Die übrigen Behauptungen kann der Leser leicht selbst nachprüfen. ♢

Im folgenden nehmen wir an, daß die Anzahlen der Nachkommen verschiedener Generationen stochastisch unabhängig sind und bedingt unter konstanter Populationsgröße N der Verteilung gemäß (6.1) mit $k = N$ genügen. Der folgende Satz zeigt unter Verwendung eines Galton-Watson-Verzweigungsprozesses (GWP), daß ein derartiges Modell tatsächlich realisierbar ist .

6.2. Satz. *Sei $\{\nu_{n,j} : j, n \in \mathbb{N}\}$ eine Familie stochastisch unabhängiger Poi(θ)-verteilter Zufallsgrößen, $\theta > 0$ beliebig, und $(Z_n)_{n\geq 0}$ der durch $Z_0 = N$ und*

$$Z_n = \sum_{j=1}^{Z_{n-1}} \nu_{n,j}, \quad n \geq 1,$$

spezifizierte GWP. Sei ferner $\mathcal{F}_n \overset{\text{def}}{=} \sigma(\nu_{k,j}, k \leq n, j \geq 1)$. *Dann gilt*

$$P^{(\nu_{n,1},...,\nu_{n,k})|\mathcal{F}_{n-1}, Z_{n-1}=N, Z_n=k} = M(k, \tfrac{1}{N}, ..., \tfrac{1}{N}) \tag{6.3}$$

für alle $k, n \in \mathbb{N}$.

Beweis: Der Nachweis von (6.3) läßt sich unter Hinweis auf Lemma 6.1 leicht führen und bleibt dem Leser überlassen. ♢

Kommen wir zur Verbindung des soeben beschriebenen Populationsmodells mit dem Wright-Fisher-Modell und stellen uns dazu vor, daß jedes Individuum einer gegebenen Generation entweder das Allel A_1 oder A_2 trägt und daß i der Individuen mit A_1 markiert seien, o.B.d.A. die Individuen $1, ..., i$. Da es keine Mutationen gibt, hat ein Individuum vom Typ A_1 nur Nachkommen desselben Typs. Unter Hinweis auf (6.2) ist die Anzahl $\nu_1 + ... + \nu_i$ der Nachkommen vom Typ A_1 folglich $B(N, \frac{i}{N})$-verteilt, was exakt der Konstellation im einfachen Wright-Fisher-Modell entspricht. Zusammenfassend können wir festhalten, daß sich das Wright-Fisher-Modell in ein Galton-Watson-Verzweigungsmodell mit Poissonscher Reproduktionsverteilung 'einbetten' läßt und so eine natürlicherweise eine Genealogie induziert wird. Diese ergibt sich aber auch aus unserer Modellbeschreibung in Abschnitt 1: Zieht man aus einer Urne mit N Kugeln, die die Individuen der gegenwärtigen Generation repräsentieren und mit $1, ..., N$ numeriert sind, nacheinander N-mal eine Kugel mit Zurücklegen und notiert deren Nummern (statt nur dessen Allel-Typ), so erhält man als Ergebnis Zufallsgrößen $Y_1, ..., Y_N$, die jeweils Laplace-verteilt sind auf $\{1, ..., N\}$. Offenbar gibt dann Y_j die Mutter des j-ten Individuums der nächsten Generation an.

Aufgrund dieser Spezifikationen wird deutlich, wie man den Evolutionsprozeß von Eltern zu Kindern zu Enkeln usw. simulieren kann. Bild 6.1 zeigt eine Realisierung eines solchen Prozesses für $N = 9$ Individuen. Eine Prüfung zeigt, daß die Individuen Nr. 3 und 4 der letzten (gezeigten) Generation ihren ersten gemeinsamen Vorfahren (EGV) 3 Generationen vorher besitzen, während die Individuen Nr. 2 und 3 ihren EGV 11 Generationen vorher besitzen. Gegeben eine beliebige Populationsgröße N und eine n-Stichprobe der gegenwärtigen Generation, stellt sich allgemeiner die Frage: Wie sieht die Verwandtschaftsstruktur der Mitglieder in der Stichprobe aus? Die diesbetreffend entscheidende Beobachtung, die sich sofort aus dem o.a. Urnenmodell ergibt, besteht darin, daß Individuen bei Rückwärtsbetrachtung des Prozesses ihre Vorfahren völlig zufällig aus der vorherigen Generation wählen, und daß sukzessive Wahlen unabhängig von Generation zu Generation getroffen werden. Selbstverständlich sind nicht alle Mitglieder vorheriger Generationen Vorfahren von Mitgliedern der aktuellen Stichprobe. In Bild 6.2 sind die Ahnenlinien der Individuen dieser Stichprobe durch gestrichelte Linien hervorgehoben, in Bild 6.3 wurden alle anderen Ahnenlinien entfernt, so daß nur noch die 'erfolgreichen', d.h. zur Stichprobe verwandten Vorfahren gezeigt werden. Die entzerrten Verwandtschaftsbeziehungen aus Bild 6.3 sind schließlich in Bild 6.4 dargestellt und zeigen die Baumstruktur der Genealogie der betrachteten Stichprobe.

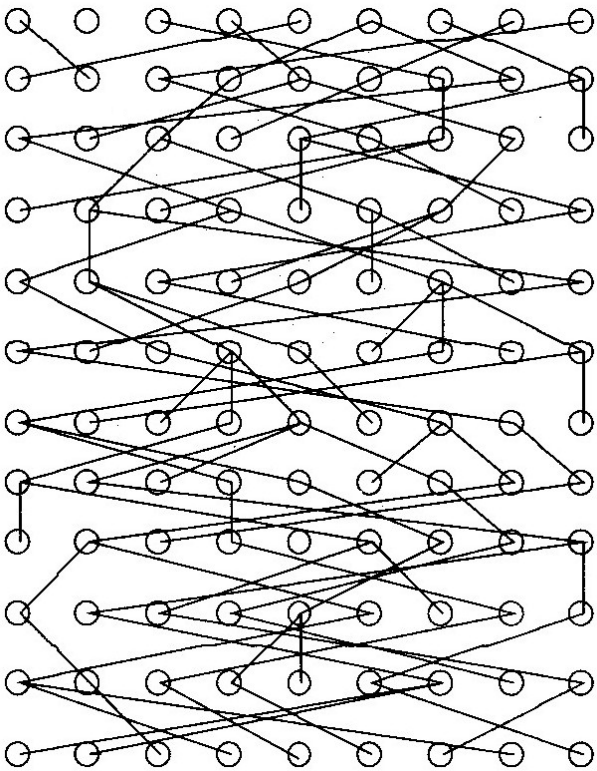

BILD 6.1. Simulation eines Wright-Fisher-Modells mit $N = 9$ Individuen. Generationen evolvieren im Bild abwärts. Die Individuen der letzten Generation sollten mit $1, ..., 9$ durchnumeriert werden. Individuen, die in Eltern-Kind-Beziehung stehen, sind durch eine Linie verbunden. [Quelle: Tavaré (2004)]

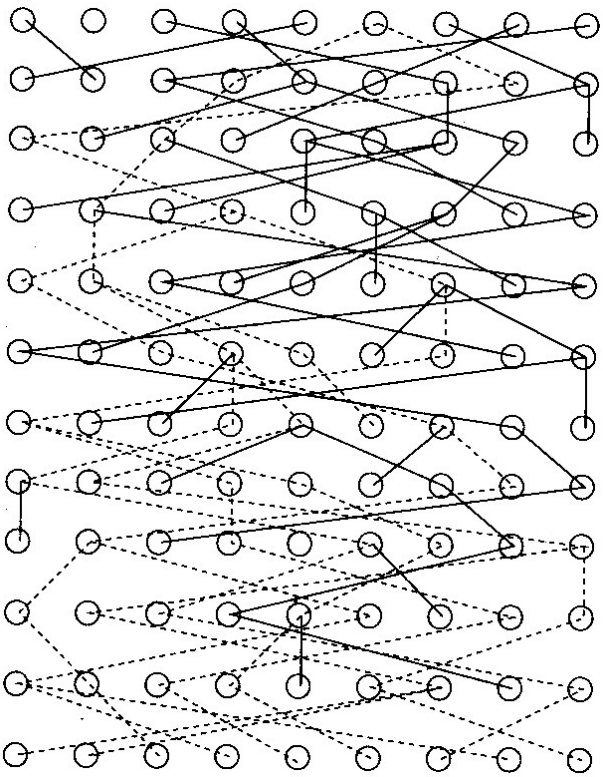

BILD 6.2. Dieselbe Simulation wie in Bild 3.1, wobei die gestrichelten Linien die Ahnenlinien der Individuen der letzten Generation angeben.

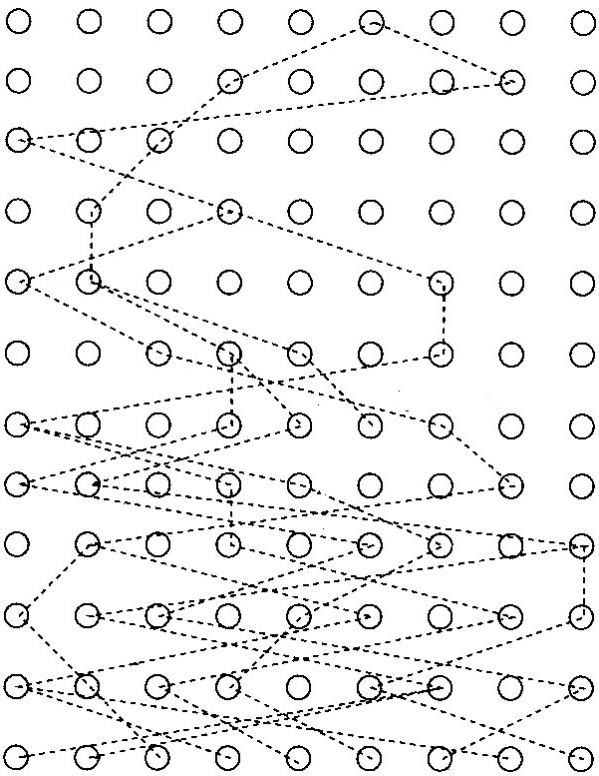

BILD 6.3. Dieselbe Simulation wie in BILD 3.1, aber unter Verzicht auf diejenigen Ahnenlinien, die vor der letzten Generation ausgestorben sind.

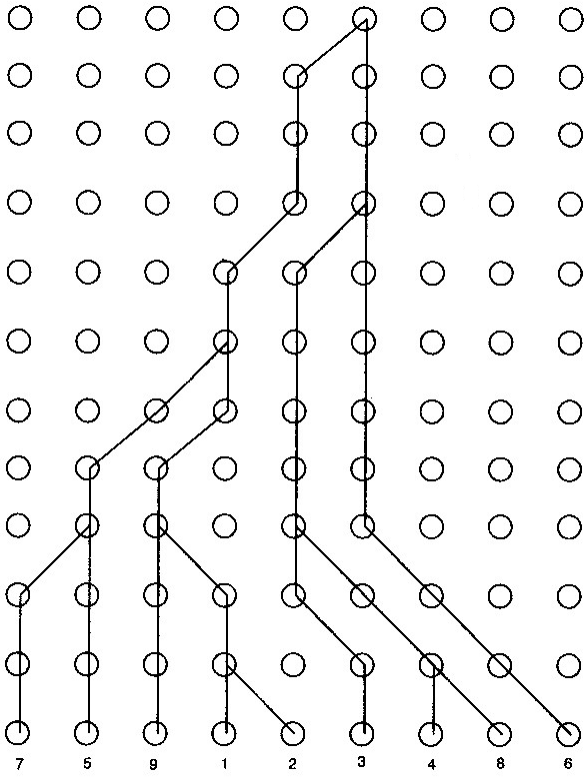

BILD 6.4. Die entzerrte Version von BILD 6.3.

Für ein genetisches Modell ohne Mutation eröffnet das Verständnis des genealogischen Prozesses einen direkten Weg zur Untersuchung von Gen-Frequenzen (☞ Felsenstein (1971)). Zur Demonstration dieser Aussage geben wir eine genealogische Herleitung der Formel (4.5) für die Heterozygosität der gegebenen Wright-Fisher-Population und fragen hierfür, wieviele Generationen man zurückschauen muß, damit zwei zufällig gewählte Gene der gegenwärtigen Generation ihren ersten gemeinsamen Vorfahren besitzen. Da Individuen ihre Eltern unabhängig voneinander und völlig zufällig wählen, gilt

$$P\Big(\text{2 Individuen haben verschiedene Eltern}\Big) \;=\; \kappa \;=\; 1 - \frac{1}{N}.$$

Da aber die Eltern ihrerseits eine Zufallsstichprobe in ihrer Generation bilden, läßt sich das Argument iterieren, und man erhält

$$P\Big(\text{2 Individuen haben ihren EGV vor mehr als } r \text{ Generationen}\Big) \;=\; \kappa^r \tag{6.4}$$

für jedes $r \geq 1$.

Betrachten wir nun die Wahrscheinlichkeit $h_\lambda(r)$, daß unter der Anfangsverteilung λ zwei zufällig mit Zurücklegen gewählte Individuen der r-ten Generation von verschiedenem Genotyp sind. Stimmen beide Individuen überein, was mit Wahrscheinlichkeit $\frac{1}{N}$ geschieht, so ist diese Wahrscheinlichkeit natürlich 0; andernfalls besitzen diese verschiedenen Genotyp dann und nur dann, wenn ihr EGV mehr als r Generationen zurückliegt und ihre Vorfahren in Generation 0 verschiedenen Genotyp besitzen. Die bedingte Wahrscheinlichkeit für das letzte Ereignis (gegeben "EGV $\geq r$") entspricht aber gerade der Chance, daß zwei ohne Zurücklegen gezogene Individuen nicht denselben Genotyp haben und ist deshalb gleich $\frac{2E_\lambda M_0(N-M_0)}{N(N-1)}$. Dies zeigt erneut

$$h_\lambda(r) \;=\; \kappa^r \left(1 - \frac{1}{N}\right) \frac{2E_\lambda M_0(N-M_0)}{N(N-1)} \;=\; \kappa^r h_\lambda(0)$$

für alle $r \geq 0$, d.h. (4.5).

Für große Populationsgrößen N ergibt sich nach Reskalierung der Zeit in Vielfachen von N, daß die Zeit bis zum EGV für eine Stichprobe von 2 Individuen asymptotisch Exp(1)-verteilt ist. Für $t > 0$ und $r = Nt$ ergibt sich nämlich in (6.4)

$$\lim_{N\to\infty} \kappa^{Nt} \;=\; \lim_{N\to\infty} \left(1 - \frac{1}{N}\right)^{Nt} \;=\; e^{-t}.$$

Daß die vorgenommene Reskalierung der Zeit exakt der in Abschnitt 5 entspricht, ist natürlich nicht überraschend, bilden doch Vorwärts- und Rückwärtsbetrachtungen lediglich alternative Perspektiven desselben zugrundeliegenden Modells.

7. Der Ahnenprozeß für große Populationen

Werfen wir als nächstes einen Blick auf die Frage nach der Anzahl von Vorfahren für größere Stichproben. Die Wahrscheinlichkeit, daß die Individuen einer Zufallsstichprobe der

Größe drei verschiedene Eltern haben, beträgt

$$\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right),$$

und das schon früher benutzte Iterationsargument liefert, daß die Stichprobe mit Wahrscheinlichkeit

$$\left(\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\right)^r \;=\; \left(1-\frac{3}{N}+\frac{2}{N^2}\right)^r$$

drei verschiedene Vorfahren über mehr als Generationen r besitzt. Eine erneute Reskalierung der Zeit in Vielfachen von N zeigt, daß diese Wahrscheinlichkeit für große N approximativ e^{-3t} entspricht. Auf der neuen Zeitskala hat die Zeit bis zum EGV einer Stichprobe von drei Genen also eine Exp(3)-Verteilung. Andererseits beträgt die Wahrscheinlichkeit, daß drei Individuen höchstens zwei Eltern besitzen,

$$\frac{3(N-1)}{N^2}+\frac{1}{N^2} \;=\; \frac{3N-2}{N^2}.$$

Da Individuen ihre Eltern zufällig wählen, ergibt sich nämlich die Wahrscheinlichkeit für exakt zwei verschiedene Eltern zu

$$M(3,\tfrac{1}{N},...,\tfrac{1}{N})(E_2) \;=\; |E_2|\,\frac{3!}{2!\,1!\,0!}\,\frac{1}{N^3} \;=\; \frac{3N(N-1)}{N^3} \;=\; \frac{3(N-1)}{N^2},$$

wobei

$$E_2 \;\stackrel{\text{def}}{=}\; \left\{(i_1,...,i_N)\in\mathbb{N}_0^N : \sum_{k=1}^N \mathbf{1}_{\{1\}}(i_k)=\sum_{k=1}^N \mathbf{1}_{\{2\}}(i_k)=1 \text{ und } \sum_{k=1}^N i_k=3\right\}$$

mit der Interpretation von i_k als Häufigkeit, mit der das k-te Individuum der Elterngeneration von einem der Individuen der Stichprobe als Mutter gewählt wird. Daß alle drei Individuen dieselbe Mutter wählen, ergibt sich entsprechend zu

$$M(3,\tfrac{1}{N},...,\tfrac{1}{N})(E_1) \;=\; |E_1|\,\frac{3!}{3!\,0!\,0!}\,\frac{1}{N^3} \;=\; \frac{N}{N^3} \;=\; \frac{1}{N^2},$$

wobei

$$E_1 \;\stackrel{\text{def}}{=}\; \left\{(i_1,...,i_N)\in\mathbb{N}_0^N : \sum_{k=1}^N \mathbf{1}_{\{3\}}(i_k)=1 \text{ und } \sum_{k=1}^N i_k=3\right\}.$$

Als bedingte Wahrscheinlichkeit für exakt zwei Vorfahren in Generation r, gegeben der EGV tritt in der r-ten Generation auf, erhalten wir nun

$$\frac{\left(1-\frac{3}{N}+\frac{2}{N^2}\right)^{r-1}\frac{3N-3}{N^2}}{\left(1-\frac{3}{N}+\frac{2}{N^2}\right)^{r-1}\frac{3N-2}{N^2}} \;=\; \frac{3N-3}{3N-2},$$

und diese strebt für $N\to\infty$ gegen 1. Folglich sinkt die Zahl der gemeinsamen Vorfahren im asymptotischen Modell um genau 1, wenn der EGV gefunden wird.

Wir können die bisherige Diskussion dahingehend zusammenfassen, daß im asymptotischen Modell der EGV einer Stichprobe von drei Genen eine Exp(3)-verteilte Zeit T_3 zurückreicht und dann zwei verschiedene Vorfahren für eine weitere Exp(1)-verteilte Zeit T_2 besitzt. Darüberhinaus sind T_2 und T_3 stochastisch unabhängig.

Betrachten wir nun allgemein die Anzahl verschiedener Eltern einer zufälligen k-Stichprobe, so können wir diese offenbar gleichsetzen mit der Anzahl besetzter Zellen, nachdem k Kugeln in N Zellen zufällig plaziert worden sind. Dies ergibt für die Wahrscheinlichkeit $g_{k,j}$, daß k Individuen j verscheidene Eltern wählen,

$$g_{k,j} \;=\; \frac{N(N-1)\cdot\ldots\cdot(N-j+1)\,\mathcal{S}_k^{(j)}}{N^k}, \quad j=1,2,...,k, \tag{7.1}$$

wobei $\mathcal{S}_k^{(j)}$ eine Stirling-Zahl der zweiten Art ist und angibt, auf wieviele Arten eine k-elementige Menge in j nichtleere Teilmengen partitioniert werden kann. Zur Begründung von (7.1) notieren wir, daß es somit $\mathcal{S}_k^{(j)}$ Möglichkeiten gibt, die k Individuen der Stichprobe in j Gruppen zu unterteilen, und dann $N(N-1)\cdot\ldots\cdot(N-j+1)$ Möglichkeiten, diesen Gruppen j verschiedene Eltern zuzuweisen. Außerdem gibt es N^k Möglichkeiten, k Individuen in beliebiger Weise ihren Eltern zuzuordnen.

Das Verhalten der Ahnenstruktur für festes N zu analysieren, ist schwierig, aber wir werden sehen, daß einfache Approximationen nach Reskalierung der Zeit, wie wir sie für Stichproben der Größe 2 und 3 hergeleitet haben, auch für allgemeine Stichprobengrößen n bestimmt werden können. Als erstes definieren wir den Prozeß

$$A_n^N(t) \;\stackrel{\text{def}}{=}\; \text{Anzahl verschiedener Vorfahren in Generation } t \text{ einer } n\text{-Stichprobe zur Zeit } 0$$

für $t=0,1,2,...$ und $1\le n\le N$. Die vorherigen Überlegungen liefern direkt:

7.1. Lemma. *Der Prozeß $A_n^N(\cdot)$ ist eine DMK mit Zustandsraum $\{1,2,...,n\}$ und Übergangswahrscheinlichkeiten*

$$P(A_n^N(t+1)=j\,|\,A_n^N(t)=k) \;=\; g_{k,j}$$

für alle $j,k\in\{1,2,...,n\}$ mit $j\le k$, wobei die $g_{k,j}$ gemäß (7.1) definiert sind. Im Fall $j>k$ gilt $P(A_n^N(t+1)=j\,|\,A_n^N(t)=k)=0$, und die Übergangsmatrix $G_{N,n}$ hat folglich die Gestalt einer unteren Dreiecksmatrix, nämlich

$$G_{N,n} \;=\; \begin{pmatrix} g_{1,1} & 0 & 0 & \ldots & 0 \\ g_{2,1} & g_{2,2} & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ g_{n,1} & g_{n,2} & g_{n,3} & \ldots & g_{n,n} \end{pmatrix}.$$

Im folgenden machen wir eine Aussage über das Verhalten von $A_n^N(\cdot)$ für festen Stichprobenumfang n und $N\to\infty$. Statt $G_{N,n}$ schreiben wir einfacher G_N. Wir dürfen erwarten,

daß die gewohnte Reskalierung der Zeit $t \mapsto Nt$ im Limes zu einem Markov-Sprungprozeß (MSP) $A_n(\cdot)$ in stetiger Zeit mit Zustandsraum $\{1,2,...,n\}$ führt, genannt *Ahnenprozeß* zum Stichprobenumfang n. Der anschließende Satz bestätigt dies und gibt zudem Auskunft über dessen Übergangsmatrixfunktion. Um $A_n^N(t)$ für alle $t \in [0,\infty)$ definiert zu haben, setzen wir $A_n^N(t) = A_n^N(k)$ für $t \in [k, k+1)$ und $k \in \mathbb{N}_0$.

7.2. Satz. *Sei $n \geq 1$ fest und $\widehat{A}_n^N(t) = A_n^N(Nt)$ für $t \in [0,\infty)$. Dann konvergiert $\widehat{A}_n^N(\cdot)$ in Verteilung gegen einen MSP $A_n(\cdot)$ auf $\{1,2,...,n\}$ mit Anfangswert $A_n(0) = n$ und Q-Matrix $Q_n = (q_{k,j})_{1\leq j,k\leq n}$, definiert durch*

$$q_{k,k} \stackrel{\text{def}}{=} -\binom{k}{2} \quad \textit{und} \quad q_{k.k-1} \stackrel{\text{def}}{=} \binom{k}{2} \quad \textit{für } k = 2,...,n \tag{7.2}$$

sowie $q_{k,j} \stackrel{\text{def}}{=} 0$ sonst.

Beweis: Bezeichnet $(G_n(t))_{t\geq 0}$ die Übergangsmatrixfunktion von $A_n(\cdot)$, also

$$G_n(t) \stackrel{\text{def}}{=} \Big(P_k(A_n(t) = j)\Big)_{1\leq k,j\leq n}$$

für $t \geq 0$, so gilt bekanntlich

$$G_n(t) \;=\; e^{tQ_n} \;\stackrel{\text{def}}{=}\; \sum_{r\geq 0} \frac{t^r Q_n^r}{r!}$$

für alle $t \geq 0$. Wir beschränken uns auf den Nachweis von

$$\lim_{N\to\infty} \Big(P_k(\widehat{A}_n^N(t) = j)\Big)_{1\leq k,j\leq n} \;=\; \lim_{N\to\infty} G_{N,n}^{Nt} \;=\; e^{tQ_n}.$$

Unter Hinweis auf (7.1) erhalten wir wegen $\mathbb{S}_k^{(k-1)} = \binom{k}{2}$

$$g_{k,k-1} \;=\; \frac{N(N-1)\cdot ...\cdot (N-k+2)\,\mathbb{S}_k^{(k-1)}}{N^k} \;=\; \binom{k}{2}\frac{1}{N} \;+\; o\Big(\frac{1}{N^2}\Big)$$

sowie für $1 \leq j < k-1$

$$g_{k,j} \;=\; \frac{N(N-1)\cdot ...\cdot (N-j+1)\,\mathbb{S}_k^{(j)}}{N^k} \;=\; o\Big(\frac{1}{N^2}\Big),$$

falls $N \to \infty$. Dies liefert wegen $g_{k,k} = 1 - \sum_{j=1}^{k-1} g_{k,j}$

$$g_{k,k} \;=\; =\; 1 \;-\; \binom{k}{2}\frac{1}{N} \;+\; o\Big(\frac{1}{N^2}\Big),$$

so daß insgesamt

$$G_{N,n} \;=\; I_n \;+\; N^{-1}Q_n \;+\; o(N^{-2}), \quad N \to \infty,$$

und daraus

$$G_{N,n}^{Nt} \;=\; \Big(I_n \;+\; N^{-1}Q_n \;+\; o(N^{-2})\Big)^{Nt} \;\to\; e^{tQ_n}, \quad N\to\infty,$$

für alle $t \geq 0$ folgt, wobei I_n die n-dimensionale Einheitsmatrix bezeichnet. ♢

Da $A_n(\cdot)$ nur Sprünge nach unten der Höhe 1 macht, handelt es sich um einen reinen Todesprozeß, dessen Verweildauer in einem Zustand k exponentialverteilt ist mit Parameter $\binom{k}{2}$ für $k = 1, ..., n$.

Die Berechnung der Verteilung von $A_n(t)$ ist eher eine Routineangelegenheit und kann etwa mittels Diagonalisierung der Matrix $Q = Q_n$ bewerkstelligt werden. Dazu schreibt man $Q_n = R_n D_n L_n$, wobei D_n die Diagonalmatrix der Eigenwerte $-\binom{k}{2}$, $1 \leq k \leq n$, von Q_n bezeichnet, und R_n, L_n die Matrizen der rechten bzw. linken Eigenvektoren angeben, die so normalisiert sind, daß $R_n L_n = L_n R_n = I_n$ gilt. Bei dieser Vorgehensweise erhalten wir für $k = 1, ..., n$ und $t > 0$

$$P_n(A_n(t) = k) \;=\; \sum_{j=k}^{n} e^{-j(j-1)/2}\frac{(2j-1)(-1)^{j-k}k_{(j-1)}n_{[j]}}{k!(j-k)!n_{(k)}}, \tag{7.3}$$

wobei $a_{(0)} = a_{[0]} \stackrel{\text{def}}{=} 1$ und

$$\begin{aligned} a_{(n)} &\stackrel{\text{def}}{=} a(a+1)\cdot ... \cdot (a+n-1),\\ a_{[n]} &\stackrel{\text{def}}{=} a(a-1)\cdot ... \cdot (a-n+1) \end{aligned}$$

für $n \geq 1$ vereinbart sei. Auch der Erwartungswert läßt sich berechnen:

$$EA_n(t) \;=\; \sum_{k=1}^{n} e^{-k(k-1)t/2}\frac{(2k-1)n_{[k]}}{n_{(k)}}, \tag{7.4}$$

und für die fallenden faktoriellen Momente gilt

$$E(A_n(t))_{[r]} \;=\; \sum_{k=r}^{n} \frac{n_{[k]}}{n_{(k)}} e^{-k(k-1)t/2}(2k-1)\frac{(r+k-2)!}{(r-1)!(k-r)!}, \tag{7.5}$$

falls $r = 2, ..., n$.