

Die approbierte Originalversion dieser Diplom-/Masterarbeit ist an der Hauptbibliothek der Technischen Universität Wien aufgestellt (<http://www.ub.tuwien.ac.at>).

The approved original version of this diploma or master thesis is available at the main library of the Vienna University of Technology (<http://www.ub.tuwien.ac.at/englweb/>).



TECHNISCHE
UNIVERSITÄT
WIEN
Vienna University of Technology

D I P L O M A R B E I T

Zeitdiskrete mathematische Modelle in der Populationsgenetik

Ausgeführt am Institut für
Diskrete Mathematik und Geometrie
der Technischen Universität Wien

unter der Anleitung von
Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Helmut Länger

durch
Gregor Mistelbauer, BSc
3522 Lichtenau im Waldviertel 60

Wien, Februar 2012

Inhaltsverzeichnis

1	Einleitung	2
2	Markov-Ketten	4
3	Mendel'sche Vererbungslehre	9
4	Hardy-Weinberg-Gesetz	11
5	Selektionsmodell	16
5.1	Zeitdiskretes Selektionsmodell	16
5.2	Fundamentalsatz der natürlichen Selektion	17
6	Wright-Fisher-Modell	20
6.1	Neutrales Wright-Fisher-Modell	20
6.2	Wright-Fisher-Modell mit Mutation	26
6.3	Wright-Fisher-Modell mit Selektion und Mutation	34
7	Moran-Modell	36
7.1	Neutrales Moran-Modell	36
7.2	Moran-Modell mit Mutation	39
7.3	Moran-Modell mit Selektion	41
7.4	Moran-Modell für unendlich viele Allele	43
8	Effektive Populationsgrößen	45
8.1	Cannings-Modell	45
8.2	Wright-Fisher-Modell	48
9	Schlussbemerkung	51

Kapitel 1

Einleitung

Eine erste Verbindung zwischen Genetik und Mathematik stellte bereits 1866 der Naturforscher Gregor Johann Mendel her, der sich mit der Vererbung bei Erbsen beschäftigte. Seine allseits bekannten daraus gewonnenen drei Vererbungsregeln bilden eine wichtige Basis für die Gesetze und Modelle, welche in dieser Arbeit beschrieben werden:

Neben den biologischen Grundlagen, welche in Kapitel 3 aufgrund ihrer Bekanntheit nur kurz wiederholt werden, werden in Kapitel 2 die mathematischen Grundlagen erarbeitet. Es handelt sich hierbei um die Theorie der Markov-Ketten, welche nach dem russischen Mathematiker Andrei Markov benannt sind.

Rund 40 Jahre wurde den Ergebnissen von Mendel keine große Aufmerksamkeit geschenkt, bis schließlich der britische Mathematiker Godfrey Hardy und der deutsche Vererbungsforscher Wilhelm Weinberg das nach ihnen benannte Hardy-Weinberg-Gesetz formulierten:

Danach bleiben für eine ideale Population die Allelhäufigkeiten p_i der Allele A_i bereits ab der ersten Generation konstant, während die Genotyphäufigkeiten P_{ij} ab der zweiten Generation konstant bleiben. Die Grundlage für den Beweis, welcher in Kapitel 4 gezeigt wird, bilden die Mendel'schen Vererbungsregeln.

Kurze Zeit danach gewann das Gebiet der Biomathematik an Bedeutung und immer mehr Wissenschaftler beschäftigten sich damit. So auch der amerikanische Genetiker Sewall Wright und der britische Mathematiker Sir Ronald Fisher:

Nach ihnen ist eines der einfachsten mathematischen Modelle der Populationsgenetik benannt, welches in Kapitel 6 diskutiert wird. Die Populationsgröße wird nicht mehr als unbegrenzt angenommen, wodurch anstatt der bisherigen deterministischen Betrachtungsweise ein stochastischer Zugang gewählt werden muss. Dabei entsteht die $(n + 1)$ -te Generation durch zufälliges Ziehen mit Zurücklegen von Genen der n -ten Generation. Zusammen mit der Theorie aus Kapitel 2 lassen sich sogenannte Übergangswahrscheinlichkeiten berechnen, welche dann diverse Eigenschaften des Modells liefern.

Ziel dieser Arbeit ist es, dieses einfache Modell mit einem Allelvorrat von zwei Allelen an einem bestimmten Genort zu verallgemeinern, indem man einerseits ein- beziehungsweise zweiseitige Mutation und/oder Selektion hinzufügt, andererseits den Allelvorrat auf unendlich viele Allele vergrößert.

Das in Kapitel 7 vorgestellte Moran-Modell ist nach dem australischen Statistiker Patrick Moran benannt. Gegenüber dem Wright-Fisher-Modell hat es den Vorteil, dass Ergebnisse

sehr häufig explizit angegeben werden können. Da es pro betrachteten Zeitpunkt im Gegensatz zum Wright-Fisher-Modell nur eine „Geburt“ und einen „Sterbefall“ gibt, gilt für die Übergangswahrscheinlichkeiten $p_{ij} = 0$ für $|i - j| > 1$. Dadurch reduziert sich die Übergangsmatrix zu einer Tridiagonalmatrix, was unter anderem ermöglicht, dass viele Ausdrücke bei diesem Modell explizit bestimmt werden können.

Auch in diesem Kapitel wird zuerst ein Grundmodell mit einem Allelvorrat von zwei Allelen an einem Genort diskutiert, welches dann durch Hinzufügen von Mutation und Selektion oder durch Vergrößern des Allelvorrats verallgemeinert wird.

Als letztes Modell wird das Cannings-Modell, welches vom amerikanischen Mathematiker Chris Cannings entwickelt wurde, kurz vorgestellt. Es handelt sich hierbei um eine Verallgemeinerung des Wright-Fisher-Modells. Das Cannings-Modell dient in dieser Arbeit vor allem der Berechnung der drei in Kapitel 8 definierten effektiven Populationsgrößen.

Betrachtet man Populationen, die bestimmte zusätzliche Eigenschaften, wie zum Beispiel zwei unterschiedliche Geschlechter, aufweisen, so liefern Berechnungen mit der tatsächlichen Populationsgröße keine befriedigenden Ergebnisse. Um die Realität besser annähern zu können, verwendet man je nach Problemstellung die passende effektive Populationsgröße. Dadurch wird eine Basis für Modelle, welche der Realität immer näher kommen, geschaffen.

Kapitel 2

Markov-Ketten

Da der mathematische Teil dieser Arbeit auf der Theorie der Markov-Prozesse¹ aufbaut, wird zu Beginn ein kurzer Überblick darüber gegeben. Neben wichtigen Definitionen enthält dieses Kapitel auch später benötigte Eigenschaften². Da die genannten Sätze und Eigenschaften zwar Basis für viele noch folgende Aussagen sind, jedoch nicht zur Kernaussage dieser Diplomarbeit gehören, werden die Aussagen nur teilweise bewiesen. Für Beweise der restlichen Aussagen sei auf entsprechende Fachliteratur verwiesen.

Definition 2.1 Ein E -wertiger stochastischer Prozess ist eine Familie von Zufallsvariablen $(X_t)_{t \in T}$ mit $X_t : \Omega \rightarrow E$.

Definition 2.2 Ein stochastischer Prozess $(X_t)_{t \in T}$ heißt Markov-Prozess mit Zustandsraum E genau dann, wenn für alle $t_i \in T$ mit $t_0 < t_1 < \dots < t_{n+1}$ und $E_{t_0}, \dots, E_{t_{n+1}} \in E$, $n \in \mathbb{N}$, die Markov-Eigenschaft

$$P(X_{t_{i+1}} = E_{t_{i+1}} \mid X_{t_0} = E_{t_0}, \dots, X_{t_i} = E_{t_i}) = P(X_{t_{i+1}} = E_{t_{i+1}} \mid X_{t_i} = E_{t_i})$$

erfüllt ist.

Definition 2.3 Ein Markov-Prozess mit diskretem Zustandsraum E heißt Markov-Kette.

Die Besonderheit eines Markov-Prozesses und damit insbesondere einer Markov-Kette ist, dass die Wahrscheinlichkeit eines Übergangs zu $X_{t_{i+1}}$ nur vom Zustand in X_{t_i} abhängt. Markov-Ketten sind durch ihre Anfangsverteilung, ihren Zustandsraum und ihre Übergangswahrscheinlichkeiten $p_{ij} = P(X_{n+1} = E_j \mid X_n = E_i)$ eindeutig bestimmt.

Definition 2.4 Ein Zustand $E_i \in E$ heißt:

- i) absorbierend, falls $p_{ii} = 1$ gilt.
- ii) rekurrent, falls $P(X_n = E_i \text{ für ein } n \geq 1 \mid X_0 = E_i) = 1$ gilt.
- iii) transient, falls er nicht rekurrent ist.

Bemerkung. Sobald der absorbierende Zustand E_i erreicht wird, verbleibt der Prozess laut Definition darin, und es kann kein anderer Zustand mehr eintreten.

¹Andrei Andrejewitsch MARKOV, russischer Mathematiker, *14. Juni 1856 in Rjasan, †20. Juli 1922 in Petrograd;

²Die hier erwähnten Aussagen stammen größtenteils aus [1, §2.12].

Definition 2.5 Für die Menge A der absorbierenden Zustände sei die Absorptionszeit \bar{t}_A definiert durch

$$\bar{t}_A := \inf\{n \geq 0 \mid X_n = E_j, E_j \in A\}$$

Definition 2.6 Eine Markov-Kette mit Zustandsraum $E = \{E_0, \dots, E_M\}$ und Übergangsmatrix $P = (p_{ij})$ hat eine stationäre Verteilung $\phi = (\phi_0, \dots, \phi_M)$, falls $\phi = \phi P$ gilt. Für alle $E_j \in E$ ergibt sich daher

$$\phi_j = \sum_{i \in E} \phi_i p_{ij} \quad \text{mit} \quad \sum_{i \in E} \phi_i = 1$$

Bemerkung. Wie bereits in dieser Definition wird auch im Folgenden der Zustand E_i teilweise mit i identifiziert.

Im Anschluss finden sich vier später benötigte Eigenschaften von Markov-Ketten, wobei bis auf Widerruf vorausgesetzt sei, dass E_0 und E_M absorbierende Zustände sind:

i) Die durchschnittliche Absorptionszeit \bar{t}_i bei ursprünglichem Zustand E_i erfüllt

$$\bar{t}_i = \begin{cases} 1 + \sum_{j=0}^M p_{ij} \bar{t}_j & i = 1, \dots, M-1 \\ 0 & i = 0, M \end{cases} \quad (2.1)$$

Bemerkung. Diese Gleichheit ist relativ einfach einzusehen:

Startet man in einem absorbierenden Zustand, so ist die durchschnittliche Absorptionszeit natürlich 0. Daher gilt $\bar{t}_0 = \bar{t}_M = 0$.

Startet man jedoch in einem nichtabsorbierenden Zustand, so gilt: Ist der nach einem Schritt erreichte Zustand absorbierend, so benötigt man keine weiteren Schritte. Ist der nach dem ersten Schritt erreichte Zustand E_j jedoch wiederum nichtabsorbierend, braucht man weitere \bar{t}_j Schritte um einen absorbierenden Zustand zu erreichen. Es ergibt sich daher

$$\begin{aligned} \bar{t}_i &= p_{i0} \cdot 1 + p_{iM} \cdot 1 + \sum_{j=1}^{M-1} p_{ij} (1 + \bar{t}_j) \\ &= \sum_{j=0}^M p_{ij} + \sum_{j=1}^{M-1} p_{ij} \bar{t}_j \\ \bar{t}_0 = \bar{t}_M = 0 & \quad 1 + \sum_{j=0}^M p_{ij} \bar{t}_j \end{aligned}$$

ii) π_i sei die bedingte Wahrscheinlichkeit, dass der absorbierende Zustand E_M unter der Bedingung $X_n = E_i$ letztlich erreicht wird. Es gilt die Rekursion

$$\pi_i = \sum_{j=0}^M p_{ij} \pi_j, \quad \pi_0 = 0, \pi_M = 1 \quad (2.2)$$

Bemerkung. Ausgehend vom Zustand E_i summiert man also über alle möglichen Zustände E_j , $j = 0, \dots, M$, wobei stets mit der Wahrscheinlichkeit π_j , dass vom jeweiligen Zustand E_j der absorbierende Zustand E_M erreicht wird, multipliziert wird. Dies entspricht einfachen Regeln der Wahrscheinlichkeitsrechnung, weswegen auf eine exakte Beweisführung hier verzichtet wird.

Die Bedingung $\pi_0 = 0$ gilt, da E_0 ebenfalls als absorbierend angenommen wird.

- iii) p_{ij}^* sei die bedingte Wahrscheinlichkeit, dass $X_{n+1} = E_j$ unter den Bedingungen, dass $X_n = E_i$ und der absorbierende Zustand E_M letztendlich eintritt. Da es sich bei X um eine Markov-Kette handelt, ergibt sich mit einfachen Regeln der Wahrscheinlichkeitsrechnung die Identität

$$p_{ij}^* = p_{ij} \frac{\pi_j}{\pi_i} \quad (2.3)$$

Beweis.

$$\begin{aligned} p_{ij}^* &= P(X_{t+1} = E_j \mid X_t = E_i, E_M \text{ tritt letztendlich ein}) \\ &= \frac{P(X_{t+1} = E_j \text{ und } E_M \text{ tritt letztendlich ein} \mid X_t = E_i)}{P(E_M \text{ tritt letztendlich ein} \mid X_t = E_i)} \\ &= \frac{p_{ij}\pi_j}{\pi_i} \end{aligned}$$

■

- iv) Startet man im Zustand E_i , so beschreibt \bar{t}_{ij} die durchschnittliche Anzahl des Eintretens von Zustand E_j bis schließlich einer der beiden absorbierenden Zustände E_0 oder E_M eintritt. Es gilt die Rekursion

$$\bar{t}_{ij} = \sum_{k=0}^M p_{ik} \bar{t}_{kj} + \delta_{ij}, \quad \bar{t}_{0j} = \bar{t}_{Mj} = 0. \quad (2.4)$$

Bemerkung. Hier kann wie in der Bemerkung von Punkt ii) argumentiert werden. Falls im Zustand E_j gestartet wird, ist dieser bereits vor dem ersten Übergang einmal eingetreten. Daher ergibt sich das Kronecker³-Symbol δ_{ij} ⁴.

Bemerkung. Mithilfe der \bar{t}_{ij} aus (2.4) kann die durchschnittliche Absorptionszeit \bar{t}_i aus Punkt i) als

$$\bar{t}_i = \sum_{j=0}^M \bar{t}_{ij}$$

dargestellt werden. Summiert man in (2.4) nun auf beiden Seiten über alle $j = 0, \dots, M$, so ergibt sich mit obiger Darstellung auch auf diese Weise die Rekursion aus (2.1):

$$\begin{aligned} \bar{t}_i &= \sum_{j=0}^M \bar{t}_{ij} \\ &= \sum_{j=0}^M \sum_{k=0}^M p_{ik} \bar{t}_{kj} + \sum_{j=0}^M \delta_{ij} \\ &= \sum_{k=0}^M \left(p_{ik} \sum_{j=0}^M \bar{t}_{kj} \right) + 1 \\ &= \sum_{k=0}^M p_{ik} \bar{t}_k + 1 \end{aligned}$$

³Leopold KRONECKER, deutscher Mathematiker, *7. Dezember 1823 in Liegnitz, †29. Dezember 1891 in Berlin;

⁴ $\delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$

Hierbei ist zu erwanen, dass die Bedingungen $\bar{t}_{0j} = \bar{t}_{Mj} = 0$ aus (2.4) die Bedingungen $\bar{t}_0 = \bar{t}_M = 0$ in (2.1) liefern.

Wie bereits in der Einleitung erwahnt wurde, ist fur das spater betrachtete Moran-Modell die Matrix P der ubergangswahrscheinlichkeiten eine Tridiagonalmatrix. Das heit, dass nur die ubergangswahrscheinlichkeiten p_{ii} , $p_{i,i+1}$ sowie $p_{i,i-1}$ ungleich Null sind. In diesem Fall lassen sich viele der oben angefuhrten Ausdrucke explizit bestimmen.

Zur Erleichterung und besseren ubersichtlichkeit fuhren wir die Notationen

$$\begin{aligned}\lambda_i &:= p_{i,i+1} \\ \mu_i &:= p_{i,i-1} \\ \rho_0 &:= 1 \\ \rho_k &:= \frac{\mu_1 \cdots \mu_k}{\lambda_1 \cdots \lambda_k} \quad \text{fur } k = 1, \dots, M-1\end{aligned}$$

ein.

Falls sowohl E_0 als auch E_M absorbierend sind, ergibt sich in diesem Fall mit $p_{ii} = 1 - \mu_i - \lambda_i$ fur (2.2) das System

$$\pi_i = \frac{\mu_i}{\lambda_i + \mu_i} \pi_{i-1} + \frac{\lambda_i}{\lambda_i + \mu_i} \pi_{i+1} \quad i = 1, \dots, M-1$$

welches mit den Bedingungen $\pi_0 = 0$ und $\pi_M = 1$ die Losung

$$\pi_i = \frac{\rho_0 + \dots + \rho_{i-1}}{\rho_0 + \dots + \rho_{M-1}} \quad i = 1, \dots, M-1 \quad (2.5)$$

besitzt.

Die nachsten zwei Vereinfachungen von (2.4) fur \bar{t}_{ij} werden hier ohne Beweis aus [1, S.91-92] zitiert:

Existieren wie bisher zwei absorbierende Zustande E_0 und E_M , so gilt

$$\bar{t}_{ij} = \begin{cases} \frac{(1-\pi_i) \sum_{k=0}^{j-1} \rho_k}{\rho_{j-1} \mu_j} & j = 1, \dots, i \\ \frac{\pi_i \sum_{k=j}^{M-1} \rho_k}{\rho_j \lambda_j} & j = i+1, \dots, M-1. \end{cases} \quad (2.6)$$

Existiert nur der absorbierende Zustand E_0 , so gilt weiterhin $\bar{t}_i = \sum_{j=0}^M \bar{t}_{ij}$. Jedoch gilt

$$\bar{t}_{ij} = \begin{cases} \frac{1}{\mu_j} \left(1 + \sum_{l=1}^{j-1} \left(\prod_{k=1}^l \frac{\lambda_{j-k}}{\mu_{j-k}} \right) \right) & j = 1, \dots, i \\ \bar{t}_{ii} \frac{\lambda_i \cdots \lambda_{j-1}}{\mu_{i+1} \cdots \mu_j} & j = i+1, \dots, M \end{cases} \quad (2.7)$$

Existiert kein absorbierender Zustand, so gibt es eine stationare Verteilung $\phi = (\phi_0, \dots, \phi_M)$, welche durch

$$\phi_i = \phi_0 \frac{\lambda_0 \cdots \lambda_{i-1}}{\mu_1 \cdots \mu_i} \quad (2.8)$$

gegeben ist. Dabei muss ϕ_0 so gewahlt werden, dass $\sum_{i=0}^M \phi_i = 1$ erfullt ist.

Bemerkung. Diese Darstellung ergibt sich aus Definition 2.6:

Da $p_{ij} = 0$ für $|i - j| > 1$, reduziert sich das System $\phi = \phi P$, wobei P die Tridiagonalmatrix der Übergangswahrscheinlichkeiten ist, auf

$$\begin{aligned}(\mu_i + \lambda_i)\phi_i &= \phi_{i+1}\mu_{i+1} + \lambda_{i-1}\phi_{i-1} \\ \Leftrightarrow \mu_{i+1}\phi_{i+1} - \lambda_i\phi_i &= \mu_i\phi_i - \lambda_{i-1}\phi_{i-1}\end{aligned}$$

Definiert man $\psi_i := \mu_i\phi_i - \lambda_{i-1}\phi_{i-1}$, so gilt also

$$\psi_{i+1} = \psi_i$$

Da $\lambda_{-1} = \mu_0 = 0$, folgt für $i = 0$:

$$\begin{aligned}0 &= \mu_1\phi_1 - \lambda_0\phi_0 \\ \Leftrightarrow \phi_1 &= \frac{\lambda_0}{\mu_1}\phi_0\end{aligned}$$

Wegen $\psi_1 = 0$ folgt $\psi_i = 0$ für alle $i = 1, \dots, M$ und daher

$$\phi_{i+1} = \frac{\lambda_i}{\mu_{i+1}}\phi_i, \quad \text{für } i = 0, \dots, M - 1$$

womit sich die gesuchte Behauptung ergibt.

Kapitel 3

Mendel'sche Vererbungslehre

Der Augustinermönch Gregor Mendel⁵ gilt als Pionier der Genetik. Die Ergebnisse seiner Kreuzungsversuche mit Erbsen veröffentlichte er 1866 in der Publikation „Versuche über Pflanzen-Hybriden⁶“:

i) Uniformitätsregel

Kreuzt man homozygote - also reinerbige - Eltern, so sind alle Nachkommen der ersten Tochtergeneration gleich. Mendel kreuzte in seinem Versuch rezessive weißblühende und dominante violettblühende reinerbige Erbsenarten, wobei die Tochtergeneration ausschließlich aus violettblühenden Erbsen bestand. Neben dem Phänotyp⁷ stimmt auch der Genotyp⁸ aller Nachkommen überein:

		AA	
	×	A	A
aa	a	Aa	Aa
	a	Aa	Aa

ii) Spaltungsregel

Kreuzt man diese Tochtergeneration wiederum, so treten im dominant-rezessiven⁹ Erbgang violett- und weißblühende Erbsen im Verhältnis 3:1 auf. Im intermediären¹⁰ Erbgang ist das Verhältnis von violett-, lila- und weißblühenden Erbsen 1:2:1.

		Aa	
	×	A	a
Aa	A	AA	Aa
	a	Aa	aa

iii) Unabhängigkeitsregel

Kreuzt man homozygote Eltern, welche mehrere verschiedene Merkmale aufweisen, so kommt es in der zweiten Tochtergeneration bereits zu neuen Merkmalskombinationen.

⁵Gregor Johann MENDEL, katholischer Priester und Botaniker, *20. Juli 1822 in Heinzendorf, †6. Jänner 1884 in Brünn;

⁶vergleiche [7] und [9]

⁷Als Phänotyp wird das Erscheinungsbild bezeichnet.

⁸Als Genotyp oder Erbbild wird die exakte Genpaarung an einem Ort bezeichnet.

⁹Ein Allel wird dominant genannt, falls es andere Gene überstimmt. Andernfalls handelt es sich um ein rezessives Gen.

¹⁰Hierbei kommt es zu einer Merkmalskoppelung des rezessiven und dominanten Allels.

Gregor Mendel wurde durch das Kreuzen von dominanten glatten (B) gelben (A) Erbsen mit rezessiven runzeligen (b) grünen (a) Erbsen darauf aufmerksam.

×	BA	bA	Ba	ba
BA	BA	BA	BA	BA
bA	BA	bA	BA	bA
Ba	BA	BA	Ba	Ba
ba	BA	bA	Ba	ba

Das Ergebnis waren glatte gelbe, glatte grüne, runzelige gelbe und runzelige grüne Erbsen im Verhältnis 9:3:3:1, welches obiger Tabelle zu entnehmen ist.

Erfolgreich war sein Experiment aufgrund der Homozygotität der verwendeten Elterngene sowie des glücklichen Zufalls, dass bei den untersuchten Merkmalen keine Koppelung der Gene auftrat. Es zeichnete also stets nur ein Gen für ein beobachtetes Merkmal verantwortlich. Treten hingegen Merkmalskoppelungen auf, so kann es zu entsprechenden Abweichungen kommen. Mendel zeigte mit seinen Versuchen, deren Ergebnisse rein auf Gesetzen der Wahrscheinlichkeitsrechnung beruhen, dass Erbmerkmale als diskrete Teilchen anzusehen sind.

Bestätigt wurden die Mendel'schen Regeln erst um 1900 von de Vries¹¹, Correns¹² und Tschermak-Seysenegg¹³.

¹¹Hugo DE VRIES, holländischer Biologe, *16. Februar 1848 in Haarlem, †21. Mai 1935 in Lunteren;

¹²Carl CORRENS, deutscher Botaniker, *19. September 1864 in München, †14. Februar 1933 in Berlin;

¹³Erich TSCHERMAK-SEYSENEGG, österreichischer Botaniker und Genetiker, *15. November 1871 in Wien; †11. Oktober 1962 in Wien;

Kapitel 4

Hardy-Weinberg-Gesetz

In diesem Kapitel betrachten wir eine sehr große Population mit nicht überlappenden Generationen ohne Migration, Selektion und Mutation. Außerdem sollen ausschließlich Zufallspaarungen¹⁴ stattfinden. Populationen, die all diese Eigenschaften besitzen, wobei nicht nur von einer großen, sondern von einer unbegrenzten Population ausgegangen wird, nennt man ideal. Die deterministische Betrachtung ist in diesem Fall wegen der nahezu verschwindenden Varianz aufgrund der großen Individuenanzahl gerechtfertigt:

Mit der Kenntnis der Mendel'schen Regeln konnten Hardy¹⁵ und Weinberg¹⁶ 1908 folgende Gesetzmäßigkeit, welche nun als Hardy-Weinberg-Gesetz bekannt ist, beweisen:

Satz. (*Hardy-Weinberg-Gesetz*¹⁷)

In einer idealen Population gilt:

- a) Die Allelhäufigkeiten¹⁸ p_i , $i = 1, \dots, n$, bleiben von der ersten Generation weg stets unverändert.
- b) Die Genotyphäufigkeiten

$$P_{ij} = \begin{cases} p_i^2 & i = j \\ p_i p_j & i \neq j \end{cases}$$

bleiben für alle $i, j = 1, \dots, n$ ab der zweiten Generation, also der ersten Tochtergeneration, unverändert.

Erfüllt eine Population diese Bedingungen, so befindet sie sich im Hardy-Weinberg-Gleichgewicht.

Beweis. Das Genpaar eines Individuums besteht aus einem Gen des Vaters und einem Gen der Mutter. Es ist allerdings nicht von Bedeutung welches Gen von, wem stammt. Daher wird zwischen den Genotypen $A_i A_j$ und $A_j A_i$ nicht unterschieden, woraus für die Genotyphäufigkeiten sofort $P_{ij} = P_{ji}$ folgt. Die möglichen Genotypen der Population sind daher

- i) $A_i A_i$, $\forall i = 1, \dots, n$, (homozygot) und

¹⁴auch „random mating“ oder Panmixie genannt

¹⁵Godfrey Harold HARDY, britischer Mathematiker, *7. Februar 1877 in Cranleigh, †1. Dezember 1947 in Cambridge;

¹⁶Wilhelm WEINBERG, deutscher Arzt und Vererbungsforscher, *25. Dezember 1862 in Stuttgart, †27. November 1937 in Tübingen;

¹⁷vergleiche [10, §23] und [4, §1.2]

¹⁸Als Allele werden alle möglichen Ausprägungen eines Gens an einem Locus (Genort) bezeichnet.

ii) $A_i A_j$ für $i \neq j$, (heterozygot)

Die Allelhäufigkeit p_i des Allels A_i ist somit durch

$$p_i = \frac{1}{2} \left(2P_{ii} + \sum_{i \neq j} P_{ij} + \sum_{i \neq j} P_{ji} \right) = \sum_{j=1}^n P_{ij} \quad \forall i = 1, \dots, n$$

gegeben. Damit kann man nun auf die Genotyp- und Allelhäufigkeiten der ersten Tochtergeneration (P'_{ij} und p'_i) schließen. Das Punnett-Schema¹⁹ ist ein einfacher Weg, die benötigten bedingten Wahrscheinlichkeiten zu erhalten. Für $k \neq l$ und $i \neq l, k$ ergeben sich folgende Möglichkeiten, aus denen die Genkombination $A_i A_i$ hervorgehen kann:

		$A_i A_i$	
	\times	A_i	A_i
$A_i A_i$	A_i	$A_i A_i$	$A_i A_i$
	A_i	$A_i A_i$	$A_i A_i$

		$A_i A_i$	
	\times	A_i	A_i
$A_i A_k$	A_i	$A_i A_i$	$A_i A_i$
	A_k	$A_k A_i$	$A_k A_i$

		$A_i A_k$	
	\times	A_i	A_k
$A_i A_k$	A_i	$A_i A_i$	$A_i A_k$
	A_k	$A_k A_i$	$A_k A_k$

		$A_i A_l$	
	\times	A_i	A_l
$A_i A_k$	A_i	$A_i A_i$	$A_i A_l$
	A_k	$A_k A_i$	$A_k A_l$

Aus diesen Tabellen kann man die benötigten Wahrscheinlichkeiten für die Berechnung von P'_{ii} für $k \neq l$, $i \neq k, l$ ablesen:

$\alpha \times \beta$	$P(\alpha \times \beta)$	$P(A_i A_i \alpha \times \beta)$
$A_i A_i \times A_i A_i$	P_{ii}^2	1
$A_i A_k \times A_i A_i$	$4P_{ik}P_{ii}$	$\frac{1}{2}$
$A_i A_k \times A_i A_k$	$4P_{ik}^2$	$\frac{1}{4}$
$A_i A_k \times A_i A_l$	$8P_{ik}P_{il}$	$\frac{1}{4}$

Daraus ergibt sich mit dem Satz der totalen Wahrscheinlichkeit

$$\begin{aligned}
 P'_{ii} &= \sum_{\alpha \times \beta} P(A_i A_i | \alpha \times \beta) P(\alpha \times \beta) \\
 &= P_{ii}^2 + 2 \sum_{k \neq i} P_{ik} P_{ii} + \sum_{k \neq i} P_{ik}^2 + 2 \sum_{k \neq i} \sum_{\substack{l < k \\ l \neq i}} P_{ik} P_{il} \\
 &= P_{ii}^2 + 2 \sum_{k \neq i} P_{ik} P_{ii} + \left(\sum_{k \neq i} P_{ik} \right) \left(P_{ik} \sum_{l \neq i, k} P_{il} \right) \\
 &= \left(P_{ii} + \sum_{k \neq i} P_{ik} \right) \left(P_{ii} + \sum_{l \neq i} P_{il} \right) \\
 &= p_i^2
 \end{aligned}$$

¹⁹Reginal PUNNETT, britischer Genetiker, *20. Juni 1875 in Tonbridge, †3. Januar 1967 in Bilbrook;

Zur Berechnung von P'_{ij} mit $i \neq j$ betrachten wir wieder jene Paarungen, aus denen die Kombination $A_i A_j$ hervorgehen kann. Für $k \neq i$, $l \neq j$ und $(k, l) \neq (j, i)$ ergeben sich mit den Punnett-Schemata

		$A_j A_j$	
	\times	A_j	A_j
$A_i A_i$	A_i	$A_i A_j$	$A_i A_j$
	A_i	$A_i A_j$	$A_i A_j$

		$A_j A_j$	
	\times	A_j	A_j
$A_i A_k$	A_i	$A_i A_j$	$A_i A_j$
	A_k	$A_k A_j$	$A_k A_j$

		$A_j A_l$	
	\times	A_j	A_l
$A_i A_i$	A_i	$A_i A_j$	$A_i A_l$
	A_i	$A_i A_j$	$A_i A_l$

		$A_i A_j$	
	\times	A_i	A_j
$A_i A_j$	A_i	$A_i A_i$	$A_i A_j$
	A_j	$A_j A_i$	$A_j A_j$

		$A_j A_l$	
	\times	A_j	A_l
$A_i A_k$	A_i	$A_i A_j$	$A_i A_l$
	A_k	$A_k A_j$	$A_k A_l$

die benötigten Wahrscheinlichkeiten:

$\alpha \times \beta$	$P(\alpha \times \beta)$	$P(A_i A_j \alpha \times \beta)$
$A_i A_i \times A_j A_j$	$2P_{ii}P_{jj}$	1
$A_i A_k \times A_j A_j$	$4P_{ik}P_{jj}$	$\frac{1}{2}$
$A_i A_i \times A_j A_l$	$4P_{ii}P_{jl}$	$\frac{1}{2}$
$A_i A_j \times A_i A_j$	$4P_{ij}^2$	$\frac{1}{2}$
$A_i A_k \times A_j A_l$	$8P_{ik}P_{jl}$	$\frac{1}{4}$

Analog zur obigen Rechnung erhält man

$$\begin{aligned}
P'_{ij} &= \frac{1}{2} \sum_{\alpha \times \beta} P(A_i A_j | \alpha \times \beta) P(\alpha \times \beta) \\
&= P_{ii}P_{jj} + \sum_{k \neq i} P_{ik}P_{jj} + \sum_{l \neq j} P_{ii}P_{jl} + P_{ij}^2 + \sum_{k \neq i} P_{ik} \left(\sum_{\substack{l \neq j \\ (k,l) \neq (j,i)}} P_{jl} \right) \\
&= P_{ii} \left(P_{jj} + \sum_{l \neq j} P_{jl} \right) + \sum_{k \neq i} P_{ik} \left(P_{jj} + \sum_{\substack{l \neq j \\ (k,l) \neq (j,i)}} P_{jl} \right) + P_{ij}^2 \\
&= \left(P_{ii} + \sum_{k \neq i} P_{ik} \right) \left(P_{jj} + \sum_{l \neq j} P_{jl} \right) \\
&= p_i p_j
\end{aligned}$$

und damit die zweite Aussage des Satzes.

Die biologisch wichtige erste Aussage folgt unmittelbar aus der mathematisch einfachen Rechnung

$$p'_i = \sum_{j=1}^n P'_{ij} = \sum_{j \neq i} p_i p_j + p_i^2 = p_i \left(\sum_{j \neq i} p_j + p_i \right) = p_i \sum_{j=1}^n p_j = p_i,$$

womit das Hardy-Weinberg-Gesetz bewiesen ist. ■

Bemerkung. Die Genotyphäufigkeiten der Tochtergeneration berechnen sich aus den Allelhäufigkeiten der Elterngeneration.

Beispiel. Als Beispiel soll der Hardy-Weinberg-Gleichgewichtszustand für zwei Allele bestimmt werden²⁰:

Sei p die relative Häufigkeit des dominanten A_1 -Gens, $q = 1 - p$ die relative Häufigkeit des rezessiven A_2 -Gens. Außerdem seien für die Genotypen A_1A_1 , A_1A_2 und A_2A_2 die relativen Häufigkeiten durch D , H und R mit $D + H + R = 1$ gegeben. Dann ergibt sich der Zusammenhang

$$D + \frac{1}{2}H = p \quad \text{und} \quad R + \frac{1}{2}H = q$$

Laut dem Hardy-Weinberg-Gesetz für $n = 2$ Allele befindet sich die Population im Hardy-Weinberg-Gleichgewicht, falls

$$D = p^2 \qquad H = 2pq \qquad R = q^2$$

gilt.

$H^2 = 4DR$ ist somit eine notwendige Bedingung für das Hardy-Weinberg-Gleichgewicht. Folgende Rechnung zeigt, dass sie auch hinreichend ist:

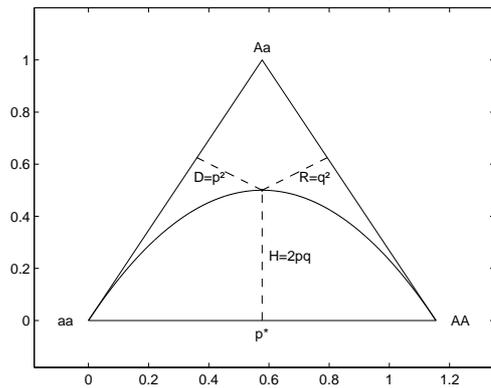
$$\begin{aligned} p^2 &= \left(D + \frac{1}{2}H \right)^2 = D^2 + DH + \frac{1}{4}H^2 = D^2 + DH + DR = D \\ 2pq &= 2 \left(D + \frac{1}{2}H \right) \left(\frac{1}{2}H + R \right) = 2DR + DH + RH + \frac{1}{2}H^2 = DH + RH + H^2 = H \\ q^2 &= \left(\frac{1}{2}H + R \right)^2 = \frac{1}{4}H^2 + HR + R^2 = DR + HR + R^2 = R \end{aligned}$$

Heterozygote Genotypen nehmen ihre maximale Häufigkeit bei $p = q = \frac{1}{2}$ an. Es müssen also zu Beginn gleich viele A_1 -Gene und A_2 -Gene vorhanden sein.

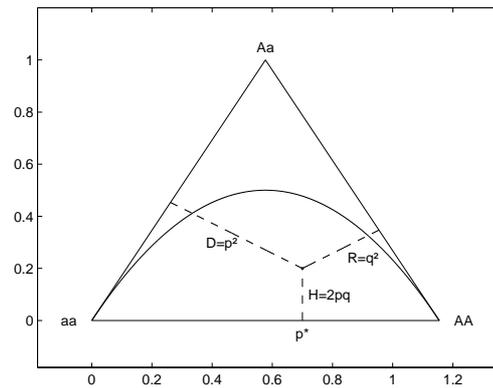
Die Finetti-Diagramme²¹ werden verwendet, um die Genotyphäufigkeiten einer diploiden Population graphisch darzustellen. Hierbei wird die geometrische Eigenschaft, dass für einen beliebigen Punkt in einem gleichseitigen Dreieck die Summe der Normalabstände zu den drei Seiten gleich der Höhe des Dreiecks ist, ausgenutzt. Für die Verwendung in der Populationsgenetik wird die Höhe mit Eins fest gewählt. Liegt ein Punkt auf der Hardy-Weinberg Parabel, so befindet sich die Population im Gleichgewicht. □

²⁰vergleiche [9, §2.23]

²¹Bruno DE FINETTI, italienischer Statistiker, *13. Juni 1906 in Innsbruck, †20. Juli 1985 in Rom;



(a) Population im Hardy-Weinberg-Gleichgewicht



(b) Population nicht im Gleichgewicht

Abbildung 1: de Finetti-Diagramme mit Hardy-Weinberg-Parabel (erstellt in MATLAB)

Bemerkung. Ab der ersten Tochtergeneration bleiben die Genotyphäufigkeiten laut dem Hardy-Weinberg-Gesetz unverändert und die Population befindet sich im Hardy-Weinberg-Gleichgewicht. Geometrisch betrachtet, entspricht der erste Zeitschritt der zur x-Achse orthogonalen Projektion eines beliebigen gegebenen Punktes auf die Hardy-Weinberg-Parabel.

Bemerkung. Da in der Realität Populationen nie unendlich groß sind, betrachtet man häufig Modelle für Populationen mit einem Genpool der Größe $2N$, das heißt diploide Populationen der Größe N beziehungsweise haploide Populationen der Größe $2N$. Dadurch ist eine deterministische Betrachtung nicht mehr gerechtfertigt, und man wählt einen stochastischen Zugang.

Kapitel 5

Selektionsmodell

In diesem Kapitel wird sowohl die zeitdiskrete als auch zeitstetige Version des Fundamentalsatzes der natürlichen Selektion behandelt. Zuvor betrachten wir allerdings das zeitdiskrete Selektionsmodell:

5.1 Zeitdiskretes Selektionsmodell

Im Unterschied zu idealen Populationen tritt in realen Populationen häufig Selektion auf: Dabei unterscheiden sich die Genotypen durch unterschiedliche Sterblichkeit und/oder Fruchtbarkeit. Im Folgenden wird Selektion ausschließlich durch unterschiedliche Überlebenschancen bis ins fortpflanzungsfähige Alter der Individuen modelliert. Wir betrachten also eine bis auf Selektion ideale Population, welche die Bedingungen des Hardy-Weinberg-Gesetzes erfüllt. Wie auch in Kapitel 4 sind die n Allelhäufigkeiten durch p_i , $i = 1, \dots, n$, gegeben. Da ausschließlich Zufallspaarungen stattfinden, ergibt sich für das Genpaar (A_i, A_j) in der Tochtergeneration im Neugeborenenstadium die Häufigkeit $p_i p_j$.

w_{ij} bezeichne die Überlebenschance bis ins fortpflanzungsfähige Alter eines Individuums mit dem Genpaar (A_i, A_j) .

Bemerkung.

- i) Die Konstanten w_{ij} werden Fitnessparameter oder Selektionskoeffizienten genannt.
- ii) Die Fitnessparameter werden in diesem Abschnitt als zeitlich konstant angenommen.
- iii) $w_{ij} \geq 0$ für alle $i, j = 1, \dots, n$.
- iv) Da für die Genpaare (A_i, A_j) und (A_j, A_i) kein Unterschied in den Genotypen $A_i A_j = A_j A_i$ ist, gilt $w_{ij} = w_{ji}$. Damit ist die Selektionsmatrix $W = (w_{ij})_{i,j=1,\dots,n}$ symmetrisch.

Betrachten wir eine Population mit N diploiden Individuen, so besitzen $p_i p_j N$ der Nachkommen das Genpaar (A_i, A_j) , wovon $w_{ij} p_i p_j N$ das fortpflanzungsfähige Alter erreichen. Insgesamt erreichen dieses Alter pro Generation $\sum_{r,s=1}^n w_{rs} p_r p_s N$ Individuen.

Unter allen fortpflanzungsfähigen Nachkommen ergibt sich für die Wahrscheinlichkeit p'_{ij} des Genpaares (A_i, A_j) im fortpflanzungsfähigen Alter beziehungsweise im Neugeborenenstadium der Tochtergeneration aufgrund der Zufallspaarung

$$p'_{ij} = \frac{w_{ij} p_i p_j N}{\sum_{r,s=1}^n w_{rs} p_r p_s N} = \frac{w_{ij} p_i p_j}{\sum_{r,s=1}^n w_{rs} p_r p_s}.$$

p'_i sei die Wahrscheinlichkeit des Allels A_i in der Tochtergeneration. Dann gilt wegen $p'_{ij} = p'_{ji}$

$$p'_i = \frac{1}{2} \sum_{j=1}^n p'_{ij} + \frac{1}{2} \sum_{j=1}^n p'_{ji} = \sum_{j=1}^n p'_{ij} \quad \text{für } i = 1, \dots, n.$$

Insgesamt ergibt sich

$$p'_i = p_i \frac{\sum_{j=1}^n w_{ij} p_j}{\sum_{r,s=1}^n w_{rs} p_r p_s} \quad \text{für } i = 1, \dots, n.$$

Diese Gleichung beschreibt die Wirkung der Selektion auf die Genhäufigkeiten der einzelnen Generationen. Mit einem Vektor $p \in S_n := \{q \in (\mathbb{R}_0^+)^n \mid \sum_{i=1}^n q_i = 1\}$ lässt sich diese Gleichheit auch darstellen als

$$p'_i = p_i \frac{(Wp)_i}{p \cdot Wp}. \quad (5.1)$$

5.2 Fundamentalsatz der natürlichen Selektion

Zeitdiskrete Version

Wir betrachten nun eine diploide Population der Größe N mit den n verschiedenen Allelen A_j , $j = 1, \dots, n$. Der Ausdruck $p \cdot Wp$ aus (5.1) kann als mittlere Fitness der Population, das ist die durchschnittliche Überlebenswahrscheinlichkeit eines zufällig der Population entnommenen Individuums vom Neugeborenenstadium ins fortpflanzungsfähige Alter, interpretiert werden. Diese mittlere Fitness besitzt nun folgende Eigenschaft:

Satz. (*Fundamentalsatz der natürlichen Selektion*)

Für die mittlere Fitness $p \cdot Wp$ einer Generation und die der Tochtergeneration gilt

$$p' \cdot Wp' \geq p \cdot Wp,$$

wobei Gleichheit genau für $p' = p$ gilt.

Bemerkung. Die mittlere Fitness kann von einer Generation zur nächsten also nicht abnehmen, da sie wächst oder zumindest gleich bleibt²².

Zeitstetige Version²³

Wir betrachten wiederum eine diploide Population, wobei nun die Annahme getrennter Generationen nicht mehr berücksichtigt wird. Sei $N_i(t)$ die Anzahl des i -ten Allels in der Gesamtpopulation zum Zeitpunkt t , dann gilt $\sum_{i=1}^n N_i(t) = N(t)$ für alle $t \in \mathbb{R}_0^+$ und damit für die relative Häufigkeit des i -ten Allels $p_i(t) = \frac{N_i(t)}{N(t)}$. Weiters definiere man für den Vektor der Allelhäufigkeiten $p(t) = (p_1(t), \dots, p_n(t))$ die Fitness des Allels A_i durch

$$\Phi_i(p(t)) = \sum_{j=1}^n w_{ij} p_j(t) \quad (5.2)$$

²²Der Beweis ist zum Beispiel in [4, S. 17-19] angeführt.

²³Dieses Kapitel orientiert sich an [10, §24-25]

sowie die mittlere Fitness der gesamten Population durch

$$\Phi(p(t)) = \sum_{i=1}^n p_i(t) \Phi_i(p(t)). \quad (5.3)$$

Mit den Änderungsraten $\dot{\Phi}_i(p(t))$ für $\dot{N}_i(t)$ und $\dot{\Phi}(p(t))$ für $\dot{N}(t)$ gilt trivialerweise

$$\begin{aligned} \dot{N}_i(t) &= \Phi_i(p(t)) N_i(t) \\ \dot{N}(t) &= \Phi(p(t)) N(t) \end{aligned}$$

Mit $\dot{p}_i(t) = \frac{\dot{N}_i(t)N(t) - N_i(t)\dot{N}(t)}{N(t)^2}$ folgt nun direkt die Differentialgleichung

$$\dot{p}_i(t) = p_i(t) (\Phi_i(p(t)) - \Phi(p(t))), \quad i = 1, \dots, n. \quad (5.4)$$

Die Differenz $\Phi_i(p(t)) - \Phi(p(t))$ ist somit die Änderungsrate für $p_i(t)$:

Ist die Fitness eines Allels größer als die mittlere Fitness der Gesamtpopulation, so wächst seine relative Häufigkeit. Ist sie jedoch kleiner, so nimmt die relative Häufigkeit ab.

Satz. (*Fundamentalsatz der natürlichen Selektion*)

Sei $p(t)$ eine Lösung von (5.4) mit $p(0) \in S_n := \{p \in (\mathbb{R}_0^+)^n \mid \sum_{i=1}^n p_i = 1\}$. Dann ist die mittlere Fitness $\Phi(p(t))$ eine monoton wachsende Funktion, die

$$\dot{\Phi}(p) = 2 \sum_{i=1}^n p_i (\Phi_i(p) - \Phi(p))^2, \quad p \in S_n$$

erfüllt. Insbesondere gilt die Äquivalenz

$$\dot{\Phi}(p) = 0 \Leftrightarrow p \text{ ist Gleichgewicht.}$$

Beweis. Sei $p(t)$ eine Lösung von (5.4) mit $p(0) \in S_n$. Wegen der Symmetrie von F gilt:

$$\begin{aligned} \frac{d}{dt} \Phi(p) &\stackrel{(5.3)}{=} \frac{d}{dt} \sum_{i=1}^n p_i \Phi_i(p) \stackrel{(5.2)}{=} \frac{d}{dt} \sum_{i,j=1}^n w_{ij} p_i p_j \\ &= 2 \sum_{i,j=1}^n w_{ij} \dot{p}_i p_j = 2 \sum_{i=1}^n \dot{p}_i \sum_{j=1}^n w_{ij} p_j \\ &\stackrel{(5.2)}{=} 2 \sum_{i=1}^n \Phi_i(p) \dot{p}_i \stackrel{(5.4)}{=} 2 \sum_{i=1}^n p_i \Phi_i(p) (\Phi_i(p) - \Phi(p)) \\ &= 2 \sum_{i=1}^n p_i \Phi_i^2 - 2 \sum_{i=1}^n p_i \Phi_i \Phi - 2 \sum_{i=1}^n p_i \Phi_i \Phi + 2 \sum_{i=1}^n p_i \Phi_i \Phi \\ &\stackrel{\sum_{j=1}^n p_j = 1}{=} 2 \sum_{i=1}^n p_i \Phi_i^2 - 4 \sum_{i=1}^n p_i \Phi_i \Phi + 2 \sum_{i=1}^n p_i \Phi_i \Phi \sum_{j=1}^n p_j \\ &\stackrel{(5.3)}{=} 2 \sum_{i=1}^n p_i \Phi_i^2 - 4 \sum_{i=1}^n p_i \Phi_i \Phi + 2 \sum_{i=1}^n p_i \Phi^2 \\ &= 2 \sum_{i=1}^n p_i (\Phi_i(p) - \Phi(p))^2 \end{aligned}$$

An dieser Formel erkennt man, dass $\dot{\Phi}(p) = 0$ genau dann, wenn für jedes $i = 1, \dots, n$ entweder $p_i = 0$ oder $\Phi_i(p) - \Phi(p) = 0$ gilt. Ist dies der Fall, befindet sich die Population im Gleichgewicht. ■

Es kann gezeigt werden, dass die Lösung $p(t)$ von (5.4) mit $p(0) \in S_n$ für $t \rightarrow \infty$ gegen die Menge der Gleichgewichte konvergiert. Eine Verschärfung dieser Aussage ist, dass die Lösung nicht nur gegen die Menge der Gleichgewichte, sondern gegen ein Gleichgewicht konvergiert²⁴.

²⁴vergleiche [10, §26]

Kapitel 6

Wright-Fisher-Modell

6.1 Neutrales Wright-Fisher-Modell

Das einfachste stochastische Modell²⁵ in der Populationsgenetik ist nach Sewall Wright²⁶ und Sir Ronald A. Fisher²⁷ benannt. Es handelt sich um ein zeitdiskretes Modell für diploide Individuen einer Population, die vorerst folgende vereinfachende Annahmen erfüllt:

- i) diskrete nichtüberlappende Generationen: Dadurch werden die Genpools der einzelnen Generationen getrennt.
- ii) gleiche Fitness der einzelnen Allele verhindert das Auftreten von Selektion: Dadurch hat kein Allel Vor- oder Nachteile bei der Fortpflanzung.
- iii) keine Veränderung der Gene durch Mutation
- iv) keine Ab- oder Zuwanderung von Genen durch Migration
- v) konstante Populationsgröße: Jede Generation verfügt über genau $2N$ Gene.
- vi) „random mating“: Es finden ausschließlich Zufallspaarungen statt.

Aufgrund dieser Eigenschaften werden Modelle solcher Art als neutral bezeichnet. Vorerst sei angenommen, dass an einem Locus lediglich die Allele A_1 und A_2 zur Auswahl stehen.

Aufgrund der Zufallspaarungen ist es möglich, anstatt der N diploiden Individuen $2N$ haploide Individuen zu betrachten. Der Genpool mit $2N$ Genen der $(n + 1)$ -ten Generation entsteht durch zufälliges Ziehen mit Zurücklegen der Gene aus der n -ten Generation. Bezeichne X_n die Anzahl der A_1 -Gene in der n -ten Generation, dann ist X_n eine Zufallsvariable mit Merkmalsraum $\{0, \dots, 2N\}$, die mit einer Bernoulli-Zufallsvariablen

$$Y_{n,i} = \begin{cases} 1 & A_1\text{-Allel wird mit Wahrscheinlichkeit } p \text{ gezogen} \\ 0 & A_2\text{-Allel wird mit Wahrscheinlichkeit } 1 - p \text{ gezogen} \end{cases}$$

als

$$X_n = \sum_{i=1}^{2N} Y_{n,i}$$

²⁵vergleiche [1, §3]

²⁶Sewall WRIGHT, amerikanischer Biologe und Genetiker, *21. Dezember 1889 in Melrose; †3. März 1988 in Madison;

²⁷Sir Ronald Aylmer FISHER, britischer Biologe und Mathematiker, *17. Februar 1890 in London, †29. Juli 1962 in Adelaide;

dargestellt werden kann. X_n ist als Summe von Bernoulli-Verteilungen binomialverteilt mit $\text{Bin}(2N, p)$. Die Wahrscheinlichkeit, dass in der $(n+1)$ -ten Generation j A_1 -Gene auftreten, falls in der n -ten Generation i A_1 -Gene vorhanden waren, liefert die Übergangswahrscheinlichkeiten

$$p_{ij} = P(X_{n+1} = j \mid X_n = i) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j}. \quad (6.1)$$

Mit (6.1) erkennt man, dass es sich bei der Zufallsvariablen X_n um eine Markov-Kette mit Übergangsmatrix $P = (p_{ij})$ handelt. Ebenfalls mit (6.1) erhält man die absorbierenden Zustände 0 und $2N$: Gilt für ein $n_0 \in \mathbb{N}$ $X_{n_0} = 0$ oder $X_{n_0} = 2N$, so gilt das auch für alle $n > n_0$.

Es ergeben sich nun folgende Fragen:

- i) Wie groß ist bei gegebenen $X_n = i$ die durchschnittliche Zeit \bar{t}_i , bis ein absorbierender Zustand erreicht ist?
- ii) Wie oft tritt im Durchschnitt der Zustand j ein, bevor ein absorbierender Zustand erreicht wird?

Diese Fragen können eigentlich mit (2.1) und (2.4) beantwortet werden. Da es aber Schwierigkeiten bereitet, explizite Lösungen zu finden, wollen wir folgende Approximationen betrachten: Laut (2.1) gilt für die durchschnittliche Absorptionszeit \bar{t}_i einer Markov-Kette

$$\bar{t}_i = 1 + \sum_{j=0}^M p_{ij} \bar{t}_j, \quad \bar{t}_0 = \bar{t}_M = 0.$$

Wählt man im betrachteten Modell $M = 2N$ und $\frac{i}{M} = x$ fest, während $\frac{j}{M} = x + \delta x$ mit $\delta(x)$ als Zufallsvariable gewählt wird, so lässt sich obige Gleichung mit $\bar{t}_i = \bar{t}(x)$, wobei $\bar{t} \in C^2(\mathbb{R})$ angenommen sei, schreiben als

$$\begin{aligned} \bar{t}(x) &= 1 + \sum_{\delta x=0, \frac{1}{2N}, \dots, 1} P(x \mapsto x + \delta x) \bar{t}(x + \delta x) \\ &= 1 + \mathbb{E}[\bar{t}(x + \delta x)] \end{aligned} \quad (6.2)$$

Da $\bar{t}(x)$ eine zweimal stetig differenzierbare Funktion der stetigen Variablen x ist, existiert die Taylor²⁸-Reihe:

$$\begin{aligned} \bar{t}(x + \delta x) &= \sum_{n=0}^{\infty} \frac{\bar{t}^{(n)}(x)}{n!} (\delta x)^n \\ &= \bar{t}(x) + (\delta x) \bar{t}'(x) + \frac{(\delta x)^2}{2} \bar{t}''(x) + \dots \end{aligned}$$

womit sich unter Berücksichtigung von Termen bis zur Ordnung 2 die Approximation

$$\bar{t}(x) \approx 1 + \bar{t}(x) + \mathbb{E}[\delta x] \bar{t}'(x) + \frac{\mathbb{E}[(\delta x)^2]}{2} \bar{t}''(x) \quad (6.3)$$

ergibt. Mit den Übergangswahrscheinlichkeiten aus (6.1) folgt

$$\begin{aligned} \mathbb{E}[\delta x] &= \sum_{j=0}^{2N} \left(\frac{j-i}{2N}\right) \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} \\ &= \frac{1}{2N} \left[\sum_{j=0}^{2N} j \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} - \sum_{j=0}^{2N} i \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} \right] \\ &= \frac{1}{2N} (i - i) = 0 \end{aligned}$$

²⁸Brook TAYLOR, britischer Mathematiker, *18. August 1685 in Edmonton, †29. Dezember 1731 in London;

und

$$\begin{aligned}
\mathbb{E}[(\delta x)^2] &= \sum_{j=0}^{2N} \left(\frac{j-i}{2N}\right)^2 \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} \\
&= \frac{1}{4N^2} \sum_{j=0}^{2N} (j^2 - 2ij + i^2) \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} \\
&= \frac{1}{4N^2} \left[\frac{i(2N + 2Ni - i)}{2N} - 2i^2 + i^2 \right] \\
&= \frac{i(2N - i)}{8N^3} \\
&= \frac{x(1-x)}{2N}.
\end{aligned}$$

Damit erhält man die approximierte Differentialgleichung

$$x(1-x)\bar{t}''(x) \approx -4N. \quad (6.4)$$

Durch zweimaliges Integrieren erhält man mit $c, d \in \mathbb{R}$

$$\begin{aligned}
\bar{t}'(x) &\approx -4N \ln x + 4N \ln(1-x) + c \\
\bar{t}(x) &\approx -4N(x \ln x - x) + 4N(1-x - (1-x) \ln(1-x)) + cx + d.
\end{aligned}$$

Mit den Randbedingungen $\bar{t}(0) = \bar{t}(1) = 0$ folgt für die mittlere Absorptionszeit $\bar{t}(x)$ die eindeutige Lösung

$$\bar{t}(x) \approx -4N(x \log x + (1-x) \log(1-x)) \quad (6.5)$$

Beispiel. ²⁹ Betrachtet man für die Anfangshäufigkeiten von A_1 die Werte $\frac{1}{2N}$ beziehungsweise $\frac{1}{2}$, so ergibt sich für die durchschnittliche Absorptionszeit

$$\begin{aligned}
\bar{t}\left(\frac{1}{2N}\right) &\approx 2 + 2 \log(2N) \text{ Generationen} \\
\bar{t}\left(\frac{1}{2}\right) &\approx 2,8N \text{ Generationen}
\end{aligned}$$

Während es bei gleicher Häufigkeit von A_1 - und A_2 -Genen sehr lange dauert bis Absorption eintritt, ist das bei einem sehr kleinen Anteil von A_1 -Genen in der Population eher rasch der Fall. \square

Da in diesem Modell für die Wahrscheinlichkeit π_j wegen (2.2) $\pi_j = \frac{j}{2N}$ gilt, folgt für die Wahrscheinlichkeit p_{ij}^* aus (2.3)

$$p_{ij}^* = \frac{j}{i} \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(\frac{2N-i}{2N}\right)^{2N-j}. \quad (6.6)$$

Aufgrund der Fixation des A_1 -Gens muss pro Generation mindestens ein A_1 -Gen produziert werden. p_{ij}^* ist dann die Wahrscheinlichkeit, dass die restlichen $2N-1$ Gene genau $j-1$ A_1 -Gene erzeugen. Analog zu vorhin erfüllt die bedingte mittlere Absorptionszeit

$$\bar{t}^*(x) \approx 1 + \bar{t}^*(x) + \mathbb{E}[\delta x] \bar{t}^{*'}(x) + \frac{\mathbb{E}[(\delta x)^2]}{2} \bar{t}^{*''}(x), \quad (6.7)$$

²⁹Dieses Beispiel stammt aus [1, S. 93]

Mit (6.6) ergibt sich für die gesuchten Werte

$$\begin{aligned}
\mathbb{E}[\delta x] &= \sum_{j=1}^{2N} \left(\frac{j-i}{2N}\right) \frac{j}{i} \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} \\
&= \frac{1}{2N} \sum_{j=1}^{2N} \left(\frac{j^2}{i} - j\right) \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} \\
&= \frac{1}{2N} \left(\frac{2N + 2Ni - i}{2N} - i\right) \\
&= \frac{2N - i}{4N^2} \\
&= \frac{1 - x}{2N}
\end{aligned}$$

und

$$\begin{aligned}
\mathbb{E}[(\delta x)^2] &= \sum_{j=1}^{2N} \left(\frac{j-i}{2N}\right)^2 \frac{j}{i} \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} \\
&= \frac{1}{4N^2} \sum_{j=1}^{2N} \left(\frac{j^3}{i} - 2j^2 + ij\right) \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} \\
&= \frac{1}{4N^2} \left(\frac{2N^2 + 6N^2i - 3Ni + 2N^2i^2 - 3Ni^2 + i^2}{2N^2} - \frac{i(2N + 2Ni - i)}{N} + i^2\right) \\
&= \frac{1}{8N^4} (2N^2 + 2N^2i - 3Ni - Ni^2 + i^2) \\
&= \frac{N + Ni - i}{4N^3} \frac{2N - i}{2N} \\
&= \frac{x}{2N} \left(\frac{1}{N} \left(\frac{N}{i} + N - 1\right)\right) (1 - x) \\
&\approx \frac{x(1-x)}{2N}.
\end{aligned}$$

Bemerkung. i kann in der letzten Approximation die Werte $1, \dots, 2N$ annehmen. Definiert man die Funktion $B(i) := \frac{1}{N} \left(\frac{N}{i} + N - 1\right)$ für großes, festes N , so zeigt sich, dass für den Großteil der möglichen Werte von i , der Funktionswert sehr nahe an 1 liegt. Daher wird für diese Approximation der Wert 1 gewählt.

Es ergibt sich für die Differentialgleichung (6.7) die Approximation

$$(1-x)t^{*\prime}(x) + \frac{1}{2}x(1-x)t^{*\prime\prime}(x) = -2N$$

Die Lösung dieser Differentialgleichung unter den Bedingungen $\bar{t}^*(1) = 0$, $\lim_{x \rightarrow 0} \bar{t}^*(x) < \infty$ und $x = p$ lässt sich durch Trennen der Variablen sowie Variation der Konstanten berechnen. Für die homogene Lösung ergibt sich

$$\begin{aligned}
2(1-x)t^{*\prime}(x) + x(1-x)t^{*\prime\prime}(x) &= 0 \\
\Leftrightarrow \frac{t^{*\prime\prime}(x)}{t^{*\prime}(x)} &= -\frac{2}{x}
\end{aligned}$$

womit durch zweimaliges Integrieren mit $c, d \in \mathbb{R}$

$$\bar{t}^*(x) = -\frac{c}{x} + d$$

folgt. Der Ansatz $\bar{t}^*(x) = -\frac{c(x)}{x} + d$ liefert

$$\begin{aligned}\bar{t}^{*'}(x) &= -\frac{c'(x)}{x} + \frac{c(x)}{x^2} \\ \bar{t}^{*''}(x) &= -\frac{c''(x)}{x} + \frac{2c'(x)}{x^2} - \frac{2c(x)}{x^3}\end{aligned}$$

Einsetzen in die ursprüngliche Differentialgleichung ergibt

$$\begin{aligned}-2N &= (1-x) \left(-\frac{c'(x)}{x} + \frac{c(x)}{x^2} \right) + \frac{x(1-x)}{2} \left(-\frac{c''(x)}{x} + \frac{2c'(x)}{x^2} - \frac{2c(x)}{x^3} \right) \\ \Leftrightarrow c''(x) &= \frac{4N}{1-x}\end{aligned}$$

und zweimaliges Integrieren folglich mit $c_0, c_1 \in \mathbb{R}$

$$c(x) = 4N(1-x) \log(1-x) - 4N(1-x) + c_0x + c_1.$$

Die allgemeine Lösung ist daher durch

$$\bar{t}^*(x) = -\frac{c(x)}{x} + d = -4N \frac{1-x}{x} \log(1-x) + 4N \frac{1-x}{x} - c_0 - \frac{c_1}{x} + d$$

gegeben. Die gegebene Bedingung $\bar{t}^*(1) = 0$ liefert $c_0 = -c_1$. Mit $\lim_{x \rightarrow 0} \bar{t}^*(x) < \infty$ und $x = p$ erhält man schließlich die gesuchte Lösung

$$\bar{t}^*(p) = -4N \frac{1-p}{p} \log(1-p)$$

Beispiel. ³⁰ Es gelte wieder $p_1 = \frac{1}{2N}$ und $p_2 = \frac{1}{2}$:

$$\begin{aligned}\bar{t}^*(p_1) &\approx 4N - 2 \text{ Generationen} \\ \bar{t}^*(p_2) &\approx 2,8N \text{ Generationen}\end{aligned}$$

Die lange Absorptionszeit für p_2 lässt darauf schließen, dass der größte Eigenwert der Übergangsmatrix von der Größenordnung 1 ist. Der Mathematiker Feller³¹ hat 1951 die Eigenwerte erstmals explizit angegeben³²:

$$\lambda_j = \frac{(2N)(2N-1) \cdots (2N-j+1)}{(2N)^j} \quad j = 1, 2, \dots, 2N \quad (6.8)$$

□

Das neutrale Wright-Fisher-Modell für zwei Allele lässt sich sehr einfach auf k Allele verallgemeinern:

³⁰Dieses Beispiel findet man auch in [1, S.94].

³¹William FELLER, amerikanischer Mathematiker, *7. Juli 1906 in Zagreb, †14. Jänner 1970 in New York City;

³²In Abschnitt 7.1 wird eine allgemeine Form für die Eigenwerte einer Verallgemeinerung des Wright-Fisher-Modells angegeben.

Dazu sei $k > 2$, $k \in \mathbb{N}$, beliebig. Die Zufallsvariablen $X_i(n)$, $i = 1, \dots, k$, geben die Anzahl der A_i -Gene zum Zeitpunkt n an. Daher muss $\sum_{i=1}^k X_i = 2N$ gelten, womit $k - 1$ Elemente des Vektors (X_1, \dots, X_k) unabhängig sind.

Für die Übergangswahrscheinlichkeiten ergibt sich eine Multinomialverteilung

$$P(X_i(n+1) | X_i(n), i = 1, \dots, k) = \frac{(2N)!}{\prod_{i=1}^k (X_i(n+1))!} \prod_{i=1}^k \left(\frac{X_i(n)}{2N} \right)^{X_i(n+1)}. \quad (6.9)$$

Abschließend sollen in diesem Teil noch Wahrscheinlichkeiten für die Fixierung eines Alleltyps angegeben werden. Gesucht sind also Absorptionswahrscheinlichkeiten $q_{i,0}$ im Zustand E_0 beziehungsweise $q_{i,2N}$ im Zustand E_{2N} . Mit (6.1) folgt unmittelbar, dass es sich bei (X_n) - wegen $\mathbb{E}(X_n | X_{n-1} = i) = 2N \frac{i}{2N} = i$ fast sicher für alle $n \geq 1$ - um ein beschränktes Martingal³³ handelt. Daher konvergiert aufgrund der Martingaleigenschaften (X_n) fast sicher gegen einen Grenzwert X_∞ .

Damit kann nun gezeigt werden, dass für ein beschränktes Martingal (X_n) die Absorptionswahrscheinlichkeiten durch

$$q_{i,2N} = 1 - q_{i,0} = \frac{i}{2N}, \quad \forall i \in \{0, \dots, 2N\}.$$

gegeben sind.

Beweis. Sei $i \in \{0, \dots, 2N\}$ beliebig. Weil X_n für alle n denselben Wertebereich hat und X_n für $n \rightarrow \infty$ fast sicher gegen X_∞ konvergiert, folgt $P(X_n = X_\infty \text{ für fast alle } n \geq 0) = 1$. Die Werte von X_∞ können also nur absorbierende Zustände sein. Mit den Martingaleigenschaften und der gleichgradigen Integrierbarkeit von (X_n) gilt dann

$$2Nq_{i,2N} = \mathbb{E}_i X_\infty = \mathbb{E}_i X_0 = i$$

■

Weiters kann die Varianz von (X_n) mit $\alpha_i = 1 - \beta_i = \frac{i}{2N}$ und $\kappa = 1 - \frac{1}{2N}$ für alle $n \geq 0$ als

$$\text{Var}(X_n) = (1 - \kappa^n)(\mathbb{E}X_0)(2N - \mathbb{E}X_0) + \kappa^n \text{Var}(X_0)$$

dargestellt werden:

Beweis. Es gilt mit der bedingten Varianz die Identität

$$\text{Var}(X_n) = \mathbb{E}(\text{Var}(X_n | X_{n-1})) + \text{Var}(\mathbb{E}(X_n | X_{n-1}))$$

Wegen $\mathbb{E}(X_n | X_{n-1}) = X_{n-1}$ fast sicher und $P(X_n | X_{n-1}) \sim \text{Bin}(2N, \alpha_{X_{n-1}})$ gilt mit dem Verschiebungssatz von Steiner³⁴

$$\begin{aligned} \text{Var}(X_n) &= \mathbb{E}(2N\alpha_{X_{n-1}}\beta_{X_{n-1}}) + \text{Var}(X_{n-1}) \\ &= (1 - \kappa)\mathbb{E}(X_{n-1}(2N - X_{n-1})) + \text{Var}(X_{n-1}) \\ &= (1 - \kappa)(2N\mathbb{E}(X_{n-1}) - \text{Var}(X_{n-1}) - (\mathbb{E}(X_{n-1}))^2) + \text{Var}(X_{n-1}) \\ &= (1 - \kappa)(\mathbb{E}X_0)(2N - \mathbb{E}X_0) + \kappa \text{Var}(X_{n-1}) \end{aligned}$$

³³Ein Martingal ist eine Folge (M_0, M_1, \dots) von integrierbaren Zufallsvariablen, welche die Bedingung $\mathbb{E}(M_n | M_0, M_1, \dots, M_{n-1}) = M_{n-1}$ erfüllen.

³⁴Jakob STEINER, schweizer Mathematiker, *18. März 1796 in Utzenstorf, †1. April 1863 in Bern;

Damit folgt

$$\begin{aligned}
\text{Var}(X_n) &= \kappa^n \text{Var}(X_0) + (1 - \kappa) \sum_{j=0}^{n-1} \kappa^j (\mathbb{E}X_0)(2N - \mathbb{E}X_0) \\
&= \kappa^n \text{Var}(X_0) + (\mathbb{E}X_0)(2N - \mathbb{E}X_0)(1 - \kappa)(1 + \kappa + \kappa^2 + \dots + \kappa^{n-1}) \\
&= \kappa^n \text{Var}(X_0) + (1 - \kappa^n)(\mathbb{E}X_0)(2N - \mathbb{E}X_0)
\end{aligned}$$

■

Für das Wright-Fisher-Modell gilt für eine beliebige Anfangsverteilung mit obigen Aussagen für alle $n \geq 0$

$$\begin{aligned}
(\mathbb{E}X_n)(2N - \mathbb{E}X_n) &= \mathbb{E}(X_0(2N - \mathbb{E}X_0)) - \text{Var}(X_n) \\
&= \kappa^n (\mathbb{E}(X_0(2N - \mathbb{E}X_0)) - \text{Var}(X_0)) \\
&= \kappa^n (\mathbb{E}X_0)(2N - X_0) \\
\Leftrightarrow \mathbb{E}(2\alpha_{X_n}\beta_{X_n}) &= \kappa^n \mathbb{E}(2\alpha_{X_0}\beta_{X_0})
\end{aligned}$$

wobei $\mathbb{E}(2\alpha_{X_n}\beta_{X_n})$ der Wahrscheinlichkeit entspricht, dass zwei zufällig gezogene Gene der n -ten Generation vom gleichen Typ sind, und daher die Heterozygotität als

$$h(n) := \mathbb{E}(2\alpha_{X_n}\beta_{X_n})$$

definiert wird.

Bemerkung. Die Heterozygotität nimmt unabhängig von der Anfangsverteilung geometrisch ab.

6.2 Wright-Fisher-Modell mit Mutation

Angenommen im Modell aus Abschnitt 6.1 sei Mutation erlaubt: Ein A_1 -Gen mutiert also mit Wahrscheinlichkeit $u > 0$ zu einem A_2 -Gen, wobei die Mutation von A_2 -Genen zu A_1 -Genen vorerst noch ausgeschlossen sei. Sind in der n -ten Generation also i von insgesamt $2N$ Genen A_1 -Gene, so zieht man in der $(n + 1)$ -ten Generation mit der Wahrscheinlichkeit $\frac{i(1-u)}{2N}$ ein A_1 -Gen. Damit ergeben sich die Übergangswahrscheinlichkeiten

$$p_{ij} = \binom{2N}{j} \left(\frac{i(1-u)}{2N} \right)^j \left(1 - \frac{i(1-u)}{2N} \right)^{2N-j}$$

Von besonderem Interesse ist die durchschnittliche Zeit \bar{t} bis zur Auslöschung aller A_1 -Gene. Analog zur Herleitung von (6.4) benötigt man für die allgemeine approximierte Differentialgleichung (6.3) die beiden Erwartungswerte $\mathbb{E}[\delta x]$ und $\mathbb{E}[(\delta x)^2]$. Mit den direkt darüber angeführten Übergangswahrscheinlichkeiten ergibt sich für festes $x = \frac{i}{2N}$ und $u \approx \frac{1}{N}$

$$\begin{aligned}
\mathbb{E}[\delta x] &= \sum_{j=0}^{2N} \binom{j-i}{2N} \binom{2N}{j} \left(\frac{i(1-u)}{2N} \right)^j \left(1 - \frac{i(1-u)}{2N} \right)^{2N-j} \\
&= \frac{1}{2N} (i(1-u) - i) = -ux
\end{aligned}$$

und

$$\begin{aligned}
\mathbb{E}[(\delta x)^2] &= \sum_{j=0}^{2N} \left(\frac{j-i}{2N}\right)^2 \binom{2N}{j} \left(\frac{i(1-u)}{2N}\right)^j \left(1 - \frac{i(1-u)}{2N}\right)^{2N-j} \\
&= -\frac{i(1-u)}{8N^3} (ui(2N-1) - 2N(1+i) + i) - \frac{i^2(1-u)}{2N^2} + \frac{i^2}{4N^2} \\
&= \frac{x(1-u)}{2N} + \frac{x^2}{2N} (2u - 2Nu^2 - u^2 - 1) \\
&\approx \frac{x(1-x)}{2N}
\end{aligned}$$

Damit ergibt sich für die durchschnittliche Zeit bis zur Auslöschung aller A_1 -Gene die approximierte Differentialgleichung

$$-4Nux\bar{t}'(x) + x(1-x)\bar{t}''(x) = -4N,$$

welche man durch die Substitution $t'(p) = v(p)$, Trennen der Variablen sowie Variation der Konstanten löst. Da dies zu unhandlichen Ausdrücken in der länglichen Rechnung führt, sei hier nur die Lösung unter den Bedingungen $\bar{t}(0) = 0$ und $\lim_{x \rightarrow 1} \bar{t}(x) < \infty$ für $\theta \neq 1$ mit $\theta = 4Nu$ angegeben³⁵:

$$\bar{t}(p) = \int_0^p \frac{4N}{x(1-\theta)} \left((1-x)^{\theta-1} - 1 \right) dx + \int_p^1 \frac{4N}{x(1-\theta)} (1-x)^{\theta-1} \left(1 - (1-p)^{1-\theta} \right) dx \quad (6.10)$$

Für $\theta = 1$ erhält man die Lösung durch Grenzwertbildung.

Beispiel. Wir betrachten einen Genpool der Größe $2N = 10^7$, wobei die Mutationswahrscheinlichkeit u gleich 10^{-6} ist. Dann ergeben sich für die durchschnittliche Dauer bis zur Auslöschung aller A_1 -Gene folgende Werte:

$$\begin{aligned}
\bar{t}\left(\frac{1}{2}\right) &\approx 3,1 \cdot 10^6 \text{ Generationen} \\
\bar{t}\left(\frac{1}{10^6}\right) &\approx 225 \text{ Generationen}
\end{aligned}$$

Mutiert jedoch jedes tausendste A_1 -Gen, das heißt $u = 10^{-3}$, während die anderen Werte unverändert bleiben, so ergibt sich

$$\begin{aligned}
\bar{t}\left(\frac{1}{2}\right) &\approx 9788 \text{ Generationen} \\
\bar{t}\left(\frac{1}{10^6}\right) &\approx 88 \text{ Generationen}
\end{aligned}$$

Diese nur exemplarisch gewählten Beispiele liefern bereits ein Bild, in welchen Zeitdimensionen eine Auslöschung eines einzelnen Gens stattfinden kann: Während in dem Beispiel für $p = 10^{-6}$ bei einem Genpool von 10^7 Genen die Auslöschung in relativ kurzer Zeit erfolgt, dauert es vergleichsweise sehr lange, falls die Hälfte des Genpools aus A_1 -Genen besteht. \square

³⁵vergleiche [1, S. 95-96]

Sei nun auch die Mutation von A_2 -Genen zu A_1 -Genen mit der Mutationsrate v möglich. Da A_1 -Gene weiterhin mit der Mutationsrate u zu A_2 -Genen mutieren, verändern sich die Übergangswahrscheinlichkeiten zu

$$p_{ij} = \binom{2N}{j} \left(\frac{i(1-u) + (2N-i)v}{2N} \right)^j \left(1 - \frac{i(1-u) + (2N-i)v}{2N} \right)^{2N-j}. \quad (6.11)$$

Es existiert eine stationäre Verteilung $\phi = (\phi_0, \dots, \phi_{2N})$, deren exakte Lösung aufgrund der komplexen Struktur eher schwierig ist. Für eine Approximation wird auf [1, §5.6] verwiesen. Der Erwartungswert μ und die Varianz σ^2 der stationären Verteilung können jedoch mit dem bereits Erarbeiteten berechnet werden:

Laut Definition 2.6 gilt $\phi = \phi P$ mit $P = (p_{ij})$ aus (6.11).

Betrachte den Vektor $\tau = (1, 2, \dots, 2N)$. Dann gilt mit der Darstellung für den Erwartungswert

$$\mu = \phi \tau = \phi P \tau.$$

Für die i -te Komponente von $P\tau$ gilt

$$\begin{aligned} (P\tau)_i &= \sum_{j=1}^{2N} j \binom{2N}{j} \left(\frac{i(1-u) + (2N-i)v}{2N} \right)^j \left(1 - \frac{i(1-u) + (2N-i)v}{2N} \right)^{2N-j} \\ &= i(1-u) + (2N-i)v, \end{aligned}$$

wobei die zweite Gleichheit lediglich der Erwartungswert einer Binomialverteilung ist. Damit erhält man

$$\begin{aligned} \mu = \phi P \tau &= \sum_{i=1}^{2N} \phi_i (i(1-u) + (2N-i)v) \\ &= \mu(1-u) + v(2N-\mu) \\ \Leftrightarrow \mu &= 2N \frac{v}{u+v}. \end{aligned} \quad (6.12)$$

Die Varianz kann man natürlich auf analoge Weise berechnen³⁶. Hier soll aber ein anderer Rechenweg aufgezeigt werden³⁷:

x_n bezeichne die relative Häufigkeit des A_1 -Gens in der Population zum Zeitpunkt n . Damit gilt für die Genhäufigkeit zum Zeitpunkt $n+1$

$$x_{n+1} | x_n = x_n(1-u) + (1-x_n)v + e = x_n(1-u-v) + v + e,$$

wobei e etwaige Abweichungen repräsentiert und daher $\mathbb{E}[e] = 0$ gilt. Mit der Notation $y_n = x_n(1-u-v) + v$ ergibt sich mit $x_{n+1} | x_n = y_n + e$

$$\mathbb{E}[x_{n+1}^2 | x_n] = \mathbb{E}[y_n^2 + 2y_n e + e^2 | x_n].$$

Wegen $\mathbb{E}[e] = 0$ folgt

$$\mathbb{E}[e^2] = \mathbb{E}[e^2] - (\mathbb{E}[e])^2 = \text{Var}(e),$$

wobei in diesem Fall

$$\text{Var}(e) = \frac{y_n(1-y_n)}{2N}$$

³⁶siehe zum Beispiel [12, S. 315-316]

³⁷vergleiche [3, S. 307-309]

gelte. Damit ergibt sich

$$\mathbb{E}[x_{n+1}^2 | x_n] = \mathbb{E} \left[y_n^2 + \frac{y_n(1-y_n)}{2N} \right] = \mathbb{E} \left[\frac{y_n}{2N} + y_n^2 \left(1 - \frac{1}{2N} \right) \right].$$

Außerdem ergibt sich mit $\mu' := \mathbb{E}[x_n^2]$ und $\mathbb{E}[y_n] = \mu(1-u-v) + v$

$$\mathbb{E}[y_n^2] = \mu'(1-u-v)^2 + 2\mu v(1-u-v) + v^2.$$

Insgesamt erhält man daher mit $\mu = \frac{u}{u+v}$ - da hier der relative Anteil betrachtet wird, fällt der Faktor $2N$ weg -

$$\begin{aligned} \mu' &= \frac{\mu(1-u-v) + v}{2N} + \left(1 - \frac{1}{2N} \right) (\mu'(1-u-v)^2 + 2\mu v(1-u-v) + v^2) \\ \Leftrightarrow \mu' &= \frac{v}{u+v} + \frac{1 - (2N-1)(2-u-v)v}{2N - (2N-1)(1-u-v)^2}. \end{aligned}$$

Mit dem Verschiebungssatz von Steiner erhält man die Varianz

$$\sigma^2 = \mu' - \mu^2 = \frac{uv}{(u+v)^2} \frac{1}{1 + (4N-2)(u+v) - (2N-1)(u+v)^2}$$

Ohne an Genauigkeit zu verlieren, kann man in obiger Formel $(2N-1)$ durch $2N$ ersetzen. Multipliziert man zusätzlich mit $(2N)^2$, um wieder zu Geburten und Sterbefällen zurückzukehren, so erhält man die in [12] angegebene Form

$$\sigma^2 \approx \frac{4N^2 uv}{(u+v)^2 (4Nu + 4Nv + 1)} \quad (6.13)$$

Mit diesen soeben berechneten Eigenschaften der stationären Verteilung kann die Frage nach der Wahrscheinlichkeit P_2 , dass zwei Gene, welche durch Zufallspaarung vereint werden, vom gleichen Typ sind, beantwortet werden:

Ist die Häufigkeit der A_1 -Gene x , so folgt für die gesuchte Wahrscheinlichkeit $P_2 = x^2 + (1-x)^2$. Der Erwartungswert davon ist $1 - 2\mathbb{E}[x] + 2\mathbb{E}[x^2]$. Für $u = v$ und $\theta = 4Nu$ liefert das mit (6.12) und (6.13) einen approximierten Wert von $\frac{1+\theta}{1+2\theta}$.

Alternativ lässt sich die Wahrscheinlichkeit P_2 , dass zwei zufällig gezogene Gene vom gleichen Alleltyp sind, wie folgt berechnen:

Zwei zufällig gezogene Gene einer beliebigen Generation haben mit Wahrscheinlichkeit $\frac{1}{2N}$ gemeinsame beziehungsweise mit Wahrscheinlichkeit $1 - \frac{1}{2N}$ verschiedene Eltern, welche jedoch mit Wahrscheinlichkeit P_2 vom gleichen Alleltyp sind. Während die Wahrscheinlichkeit, dass keines der Gene oder aber beide Gene Mutanten sind, $(1-u)^2 + u^2$ beträgt, ist die Wahrscheinlichkeit, dass eines der beiden Gene mutiert ist, gleich $2u(1-u)$. Zusammen ergibt das

$$P_2 = \left(\frac{1}{2N} + \left(1 - \frac{1}{2N} \right) P_2 \right) (u^2 + (1-u)^2) + \left(1 - \frac{1}{2N} \right) (1 - P_2) (2u(1-u)) \quad (6.14)$$

Mit $\theta = 4Nu$ folgt nach einigen rein algebraischen Rechenschritten

$$\begin{aligned} P_2 &= \frac{1 + 4Nu - 4Nu^2}{1 + 8Nu - 8Nu^2} \\ &\approx \frac{1 + \theta}{1 + 2\theta} \end{aligned}$$

Angenommen es existieren nun k verschiedene Allele und es tritt ausschließlich symmetrische Mutation zwischen diesen k Allelen auf, dann gibt es eine stationäre Verteilung für die Anzahl der Alleltypen. Der Erwartungswert und die Varianz lassen sich analog zu (6.12) und (6.13) berechnen. Ist die Mutationsrate u , das heißt $P(A_i \text{ mutiert}) = u$ für alle i , so gilt $P(A_i \text{ mutiert zu } A_j) = \frac{1}{k-1}$. Aufgrund der Symmetrie ist die durchschnittliche Anzahl von A_i -Allelen in der stationären Verteilung $\frac{2N}{k}$.

Mit gleichen Argumenten wie oben lässt sich die Wahrscheinlichkeit P_2 wie folgt berechnen:

$$P_2 = \left(\frac{1}{2N} + \left(1 - \frac{1}{2N} \right) P_2 \right) (u^2 + (1-u)^2) + \left(1 - \frac{1}{2N} \right) (1 - P_2) \frac{2u(1-u)}{k-1}.$$

Mit $\theta = 4Nu$ folgt wiederum nach einigen Rechenschritten

$$\begin{aligned} P_2 &= \frac{-k+1+2uk-4Nu-2u^2k+4Nu^2}{-k+1+2uk-4Nuk-2u^2k+2Nu^2} \\ &\approx \frac{k(1-2u)-1+\theta}{k(1-2u)-1+k\theta} \stackrel{k \rightarrow \infty}{\approx} \frac{1-2u}{1-2u+\theta} \approx \frac{1}{1+\theta} \end{aligned} \quad (6.15)$$

Für $k=2$ stimmt die Wahrscheinlichkeit daher mit (6.14) überein, falls man u vernachlässigt. Außerdem erkennt man hier bereits, dass für sehr kleine θ , die Wahrscheinlichkeit ungefähr 1 beträgt. Daher ist es in diesem Fall sehr wahrscheinlich, dass ein einziges Allel sehr häufig vorkommt, während die restlichen nahezu vernachlässigbare Häufigkeit haben.

Existieren unendlich viele verschiedene Allele, so ist Mutation ein wesentlicher Faktor: Vorhandene Allele mutieren mit der Rate u zu Alleltypen, welche bisher noch nie aufgetreten sind, wodurch ständig neue Allele entstehen. Sei für alle $i \in \mathbb{N}$ $X_i(n)$ die Anzahl der Gene vom Alleltyp A_i in der n -ten Generation und u die Mutationsrate. Dann ergibt sich für die Wahrscheinlichkeit, dass in der $(n+1)$ -ten Generation Y_i Gene vom Typ A_i sind und Y_0 verschiedene neue Alleltypen auftreten, die Multinomialverteilung

$$P(Y_0, Y_1, \dots \mid X_1, X_2, \dots) = \frac{(2N)!}{\prod_{i=0}^{\infty} (Y_i!)} u \prod_{i=1}^{\infty} \left(\frac{X_i(1-u)}{2N} \right)^{Y_i} \quad \forall i = 1, 2, \dots \quad (6.16)$$

Bemerkung.

- i) Im zweiten Produkt wurde der Faktor zum Index 0 - gegeben mit u - aus dem Produkt gezogen, da er unabhängig von i ist. Deshalb startet dieses Produkt mit Index 1.
- ii) Aufgrund der Voraussetzung stirbt jeder Alleltyp nach einer gewissen Zeit aus.

Da es nicht von Bedeutung ist, welches Allel genau betrachtet wird, betrachten wir nun eine Population mit $2N$ Genen mit der Struktur $\{a, b, c, \dots\}$. Es existieren also

- a Gene von einem Alleltyp
- b Gene eines anderen Alleltyps
- c Gene eines weiteren Alleltyps
- \vdots

In einer Population der Größe $2N$ gibt es genauso viele unterschiedliche Strukturen, wie es Partitionen der Zahl $2N$ in den natürlichen Zahlen gibt: $\{2N\}$, $\{2N-1, 1\}$, $\{2N-2, 2\}$, $\{2N-$

$2, 1, 1\}, \dots, \{1, 1, \dots, 1\}$. Die Übergangswahrscheinlichkeiten zwischen den einzelnen Strukturen werden durch (6.16) beschrieben. Mithilfe der Markov-Theorie lässt sich zeigen, dass es trotz der hohen Anzahl an möglichen Zuständen eine stationäre Verteilung der Strukturen gibt: Sei P_2^n die Wahrscheinlichkeit, dass zwei zufällig aus der n -ten Generation gezogene Gene vom gleichen Alleltyp sind. Es gilt daher, dass entweder

- i) keines der Gene mutiert ist oder
- ii) beide Gene von einem Vorfahren abstammen oder
- iii) die Gene Vorfahren vom gleichen Alleltyp haben.

Damit gilt

$$P_2^{n+1} = (1 - u)^2 \left(\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) P_2^n \right).$$

Die Population befindet sich im Gleichgewicht, falls $P_2^n = P_2^{n+1} = P_2$ gilt:

$$\begin{aligned} P_2 &= (1 - u)^2 \left(\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) P_2 \right) \\ \Leftrightarrow P_2 &= \frac{1}{1 - 2N + \frac{2N}{(1-u)^2}} \\ &\approx \frac{1}{1 + \theta} \end{aligned}$$

Lässt man im Modell für k Allele k gegen unendlich laufen, so erhält man, wie in (6.15) ersichtlich ist, die gleiche Approximation. Analog sei P_3^n die Wahrscheinlichkeit, dass drei zufällig gezogene Gene vom gleichen Alleltyp sind. Mit Wahrscheinlichkeit $\frac{1}{(2N)^2}$ stammen alle drei Gene von einem, mit Wahrscheinlichkeit $\frac{3(2N-1)}{(2N)^2}$ von zwei und mit Wahrscheinlichkeit $\frac{(2N-1)(2N-2)}{(2N)^2}$ von drei Vorfahren ab. Daraus ergibt sich insgesamt die Formel

$$P_3^{n+1} = \frac{(1 - u)^3}{(2N)^2} (1 + 3(2N - 1)P_2^n + (2N - 1)(2N - 2)P_3^n),$$

womit man im Gleichgewicht nach einigen Umformungsschritten die Approximation

$$P_3 \approx \frac{2}{2 + \theta} P_2 \approx \frac{2!}{(1 + \theta)(2 + \theta)}$$

erhält. Induktives Fortsetzen für wachsende Anzahl an Genen liefert die gesuchte approximierte Wahrscheinlichkeit

$$P_i \approx \frac{(i - 1)!}{(1 + \theta)(2 + \theta) \dots (i - 1 + \theta)},$$

dass i zufällig gezogene Gene vom selben Alleltyp sind.

Beispiel. Die folgende Tabelle soll eine ungefähre Größenordnung dieser Wahrscheinlichkeiten vermitteln. Für dieses Beispiel wurde die Größe des Genpools mit $2N = 10^7$ festgelegt:

	$u = 10^{-6}$	$u = 10^{-3}$
P_2	$4 \cdot 10^{-2}$	$5 \cdot 10^{-5}$
P_{10}	10^{-7}	$7 \cdot 10^{-34}$
P_{100}	$4 \cdot 10^{-23}$	10^{-270}

Bemerkung. Die Wahrscheinlichkeit P_i kann auch als Wahrscheinlichkeit, dass alle i gezogenen Gene vom gleichen Alleltyp sind, interpretiert werden. Die betrachtete Struktur ist dann einfach $\{i\}$.

In der Populationsstruktur $\{i-1, 1\}$ gibt es neben $i-1$ gleichen Genen noch ein dazu verschiedenes Gen. Die Wahrscheinlichkeit, dass i gezogene Gene genau diese Struktur aufweisen, ergibt sich als Differenz der beiden Wahrscheinlichkeiten P_{i-1} und P_i . Es gilt also $P_{i-1,1} = P_{i-1} - P_i$. Die gesuchte Wahrscheinlichkeit $P(\{i-1, 1\})$ ist für $i \geq 3$ durch $i(P_{i-1} - P_i)$ gegeben. Insgesamt gilt die Approximation

$$P(\{i-1, 1\}) \approx \frac{i(i-2)\theta}{(1+\theta)(2+\theta)\dots(i-1+\theta)}.$$

Auf analoge Weise kann man die Strukturwahrscheinlichkeiten für alle möglichen Populationsstrukturen erhalten. Eine Möglichkeit der Darstellung ist die nach Warren J. Ewens benannte „Ewens Sampling Formula“³⁸:

Sei $\bar{A} = (A_1, A_2, \dots, A_n) \in \mathbb{N}^n$ der Vektor, in dem der j -te Eintrag die Anzahl der verschiedenen Gene, welche genau j -mal in der Stichprobe der Größe n vorkommen, angibt. Mit $\bar{a} = (a_1, a_2, \dots, a_n) \in \mathbb{N}^n$ und $S_n(\theta) = \theta(1+\theta)\dots(n-1+\theta)$ gilt

$$P(\bar{A} = \bar{a}) = \frac{n!}{S_n(\theta)} \prod_{j=1}^n \frac{\theta^{a_j}}{(a_j)! j^{a_j}} \quad (6.17)$$

Notwendigerweise gilt $\sum_{j=1}^n jA_j = \sum_{j=1}^n ja_j = n$. Da (6.17) als

$$P(\bar{A} = \bar{a}) = \frac{n! \theta^{\sum_{j=1}^n a_j}}{1^{a_1} \dots n^{a_n} a_1! \dots a_n! S_n(\theta)}$$

notiert werden kann, lässt es sich mit $\sum_{j=1}^n A_j = K$ und $\sum_{j=1}^n a_j = k$ nach Summation schreiben als³⁹

$$P(K = k) = \frac{|S_n^k| \theta^k}{S_n(\theta)}, \quad (6.18)$$

wobei S_n^k der Koeffizient von θ^k in $S_n(\theta)$ ist. Damit folgt für die durchschnittliche Anzahl an Allelen pro Generation⁴⁰

$$\mathbb{E}(K) = 1 + \frac{\theta}{1+\theta} + \frac{\theta}{2+\theta} + \dots + \frac{\theta}{n-1+\theta} \quad (6.19)$$

und weiters

$$Var(K) = \theta \sum_{j=1}^{n-1} \frac{j}{(\theta+j)^2}. \quad (6.20)$$

Wir wollen uns nun der Frage widmen, wie viele verschiedene Alleltypen durchschnittlich zum Zeitpunkt n existieren.

Dafür betrachten wir ein Allel A_k , welches nach einer gewissen Anzahl an Generationen aussterben wird. Vor dem Aussterben existiert das Allel A_k mit der Wahrscheinlichkeit $\frac{1}{2N}$ in

³⁸vergleiche [1, S.114]

³⁹Der Beweis wird zum Beispiel in [5, S.141] ausgeführt.

⁴⁰Für die einzelnen Schritte sei auf [6, §41] verwiesen.

der Population. Damit ist die Häufigkeit von A_k eine Markov'sche Zufallsvariable mit der Übergangsmatrix $P = (p_{ij})$ mit den Einträgen

$$p_{ij} = \binom{2N}{j} \left(\frac{i(1-u)}{2N} \right)^j \left(1 - \frac{i(1-u)}{2N} \right)^{2N-j}$$

Die durchschnittliche Zeit $\mathbb{E}(T)$, welche das Allel A_k in der Population existiert, kann daher angegeben werden. Die durchschnittliche Anzahl an neuen Allelen pro Generation beträgt $2Nu$. Außerdem ist die durchschnittliche Anzahl an verschwundenen Allelen pro Generation gegeben durch den Quotienten $\frac{\mathbb{E}(K)}{\mathbb{E}(T)}$. Damit muss für stationäre Zustände

$$\mathbb{E}(K) = 2Nu\mathbb{E}(T)$$

gelten, da dann auch die durchschnittliche Anzahl an verschwundenen Allelen pro Generation $2Nu$ beträgt. (6.10) mit $p = \frac{1}{2N}$ liefert eine numerische Approximation für $\mathbb{E}(T)$, und damit erhält man schließlich mit numerischen Hilfsmitteln⁴¹ :

$$\mathbb{E}(K) \approx \theta + \int_{\frac{1}{2N}}^1 \frac{\theta}{x} (1-x)^{\theta-1} dx.$$

Eine bessere Approximation erhält man, indem man ein Intervall (x_1, x_2) mit $\frac{1}{2N} \leq x_1 \leq x_2 \leq 1$, in dem die betrachteten relativen Häufigkeiten liegen, wählt:

$$\mathbb{E}(K(x_1, x_2)) \approx \int_{x_1}^{x_2} \frac{\theta}{x} (1-x)^{\theta-1} dx$$

Mit dieser Approximation und der Betrachtung, dass ein Allel mit Häufigkeit x mit der Wahrscheinlichkeit $1 - (1-x)^n$ in einer Stichprobe vom Umfang n enthalten ist, folgt, dass die durchschnittliche Anzahl an verschiedenen Allelen in einer Stichprobe der Größe n gleich

$$\int_0^1 (1 - (1-x)^n) \frac{\theta}{x} (1-x)^{\theta-1} dx$$

ist. Dieser Wert ist gleich $\mathbb{E}(K)$ aus (6.19).

Die Funktion

$$h(x) := \frac{\theta}{x} (1-x)^{\theta-1}$$

wird Häufigkeitsspektrum⁴² genannt. $h(x)\delta x$ gibt die Wahrscheinlichkeit an, dass ein Allel mit einer Häufigkeit, welche im Intervall $(x, x+\delta x)$ liegt, in der Population vorkommt. In Abhängigkeit von θ kann man das Häufigkeitsspektrum folgendermaßen beschreiben:

Während es für $\theta \ll 1$ ein Allel mit hoher und wenige andere mit niedriger Häufigkeit gibt, ist es für $\theta \gg 1$ am wahrscheinlichsten, dass es sehr viele Allele mit niedriger Häufigkeit gibt. Es ist selten der Fall, dass es einige wenige Allele gibt, die sich im mittleren Häufigkeitsbereich befinden.

Bemerkung. Diverse Ergebnisse des Wright-Fisher-Modells für unendlich viele Allele bleiben für kompliziertere Modelle, wie zum Beispiel Modelle mit Berücksichtigung zweier Geschlechter oder geographischer Unterschiede, korrekt, falls $\theta = 4N_e u$ gewählt wird⁴³.

⁴¹Diese Näherungen wurden aus [1, S.115] übernommen, und werden daher hier nicht bewiesen.

⁴²siehe [2, S. 89-92]

⁴³siehe auch Kapitel 8

6.3 Wright-Fisher-Modell mit Selektion und Mutation

Tritt in einer Population neben Mutation auch Selektion auf, so wird diese durch unterschiedliche Fitnesswerte modelliert:

Für zwei Allele können also die drei Genotypen A_1A_1 , A_1A_2 und A_2A_2 mit den zeitlich konstanten Fitnesskoeffizienten w_{11} , w_{12} und w_{22} auftreten. Ist x wieder die Häufigkeit der A_1 -Gene in der Population, so ist die mittlere Fitness \bar{w} definiert durch

$$\bar{w} := x^2w_{11} + 2x(1-x)w_{12} + (1-x)^2w_{22}$$

Die einzelnen Einträge der Übergangsmatrix erfüllen wie bisher

$$p_{ij} = \binom{2N}{j} p_i^j (1-p_i)^{2N-j} \quad (6.21)$$

mit

$$p_i = \frac{1}{\bar{w}} \left([x^2w_{11} + x(1-x)w_{12}] (1-u) + [x(1-x)w_{12} + (1-x)^2w_{22}] v \right),$$

wobei $x = \frac{j}{2N}$ gilt. Anhand dieser Wahrscheinlichkeiten können bereits einige qualitative Aussagen getroffen werden:

- i) Tritt keine Mutation auf, das heißt gilt $u = v = 0$, so können die beiden absorbierenden Zustände $X = E_0$ beziehungsweise $X = E_{2N}$ erreicht werden.
- ii) Für $u > 0$, $v = 0$ kann es zur Auslöschung des A_1 -Genes kommen.
- iii) Analog kann es für $v > 0$ und $u = 0$ zur Auslöschung des A_2 -Genes kommen.
- iv) Gilt hingegen $u > 0$ und $v > 0$, so kann ein stationärer Zustand existieren.

Da für Modelle mit Selektion und Mutation quantitative Aussagen schwer zu treffen sind, wird für entsprechende Approximationen auf [1, §5] verwiesen.

Hier werden nun Approximationen für Modelle mit Selektion, jedoch ohne Mutation, das heißt $u = v = 0$, betrachtet:

Dazu seien, wie in [1, §1.4], $w_{11} = 1 + s$, $w_{12} = 1 + sh$ und $w_{22} = 1$ für ein $s \approx \frac{1}{N}$ und $h \in \mathbb{R}_0^+$. Mit $\alpha = 2Ns$, festem $x = \frac{j}{2N}$ und $x + \delta x = \frac{j}{2N}$, wobei δx wiederum die Zufallsvariable ist, lässt sich Gleichung (2.2) analog zu (6.2) und (6.3) schreiben als

$$\begin{aligned} \pi(x) &= \sum_{\delta x=0, \frac{1}{2N}, \dots, 1} P(x \mapsto x + \delta x) \pi(x + \delta x) \\ &\approx \pi(x) + \mathbb{E}[\delta x] \pi'(x) + \frac{\mathbb{E}[(\delta x)^2]}{2} \pi''(x) \end{aligned}$$

Für $\mathbb{E}[\delta x]$ und $\mathbb{E}[(\delta x)^2]$ ergibt sich nach etwas algebraischer Rechenarbeit mit (6.21), $u = v = 0$, $\alpha = 2Ns$ sowie $i = 2Nx$:

$$\begin{aligned} \mathbb{E}[\delta x] &= \sum_{j=0}^{2N} \frac{j-i}{2N} \binom{2N}{j} \left(\frac{sx(x-h-hx)+x}{sx(x+2h-2hx)+1} \right)^j \left(1 - \frac{sx(x-h-hx)+x}{sx(x+2h-2hx)+1} \right)^{2N-j} \\ &= \frac{\alpha x(1-x)(x+h-2hx)}{2N} + O(N^{-2}) \end{aligned}$$

und

$$\begin{aligned}\mathbb{E}[(\delta x)^2] &= \sum_{j=0}^{2N} \left(\frac{j-i}{2N}\right)^2 \binom{2N}{j} \left(\frac{sx(x-h-hx)+x}{sx(x+2h-2hx)+1}\right)^j \left(1 - \frac{sx(x-h-hx)+x}{sx(x+2h-2hx)+1}\right)^{2N-j} \\ &= \frac{x(1-x)}{2N} + O(N^{-2})\end{aligned}$$

Insgesamt ergibt sich die approximierte Differentialgleichung

$$(2\alpha x + 2\alpha h - 4\alpha hx) \pi'(x) + \pi''(x) = 0.$$

Die Lösung dieser Differentialgleichung ergibt sich leicht aus

$$\frac{\pi''(x)}{\pi'(x)} = -(2\alpha x + 2\alpha h - 4\alpha hx),$$

durch zweimaliges Integrieren mit $c_0, d \in \mathbb{R}$ sowie $c := e^{c_0} \in \mathbb{R}$:

$$\begin{aligned}\log(\pi'(x)) &= -\alpha x^2 - 2\alpha hx + 2\alpha hx^2 + c_0 \\ \Rightarrow \pi'(x) &= c \exp(-\alpha x^2 - 2\alpha hx + 2\alpha hx^2) \\ \Rightarrow \pi(x) &= c \int_0^x \exp(-\alpha z^2 - 2\alpha hz + 2\alpha hz^2) dz + d\end{aligned}$$

Zusammen mit den normierenden Randbedingungen $\pi(0) = 0$ und $\pi(1) = 1$ ergibt sich die Lösung

$$\pi(x) = \frac{\int_0^x \exp(-\alpha z^2 - 2\alpha hz + 2\alpha hz^2) dz}{\int_0^1 \exp(-\alpha z^2 - 2\alpha hz + 2\alpha hz^2) dz}. \quad (6.22)$$

Beispiel. ⁴⁴ Der Einfluss von Selektion auf die Fixationswahrscheinlichkeiten soll hier anhand eines einfachen - jedoch bereits aussagekräftigen - Beispiels aufgezeigt werden:

Dazu sei die Genpoolgröße $2N = 2 \cdot 10^5$, $s = 10^{-4}$ und $h = 0.5$ gewählt. Für diese Wahl von h sind die Fitnessparameter gegeben durch $w_{11} = 1 + s$, $w_{12} = 1 + \frac{s}{2}$ und $w_{22} = 1$. Die Fitness der heterozygoten Genpaarungen liegt also genau in der Mitte zwischen den Fitnesswerten der entsprechenden homozygoten Paarungen. Die Lösung (6.22) vereinfacht sich in diesem Fall zu

$$\pi(x) = \frac{\int_0^x \exp(-\alpha z) dz}{\int_0^1 \exp(-\alpha z) dz} = \frac{1 - \exp(-\alpha x)}{1 - \exp(-\alpha)}.$$

Damit gilt $\pi(0.5) = 0.9999546$. Betrachtet man hingegen das Beispiel ohne Selektion, also $s = 0$, so ergibt sich $\pi(0.5) = 0.5$. Trotz des sehr kleinen Selektionsvorteils pro Generation ergibt sich über die lange Zeit, bis schließlich Fixation eines Gens eintritt, ein erheblicher Vorteil. \square

⁴⁴Dieses Beispiel stammt aus [1, §3.2].

Kapitel 7

Moran-Modell

7.1 Neutrales Moran-Modell

Das neutrale Moran-Modell wurde 1958 von Patrick Moran⁴⁵ entwickelt. Der Vorteil dieses Modells gegenüber dem Wright-Fisher-Modell ist die Anwendung expliziter Ausdrücke für diverse Eingaben. Wir betrachten also zu diskreten Zeitpunkten $n = 1, 2, 3 \dots$ eine haploide Population der Größe $2N$. Pro Zeitpunkt wird ein Individuum zufällig gewählt, welches sich fortpflanzt, und danach eines um zu sterben. Damit bleibt die Populationsgröße immer konstant.

Angenommen wir betrachten $2N$ Gene, welche aus dem Allelvorrat $\{A_1, A_2\}$ gewählt werden können. Dann ergibt sich für die Übergangswahrscheinlichkeiten:

Falls zum Zeitpunkt n die Anzahl der A_1 -Gene gleich i ist, so existieren zum Zeitpunkt $n + 1$ genau

- i) $i - 1$ A_1 -Gene, falls sich ein A_2 -Gen fortpflanzt und ein A_1 -Gen stirbt:

$$p_{i,i-1} = \frac{i}{2N} \frac{2N-i}{2N} = \frac{i(2N-i)}{(2N)^2} \quad (7.1)$$

- ii) $i + 1$ A_1 -Gene, falls sich ein A_1 -Gen fortpflanzt und ein A_2 -Gen stirbt:

$$p_{i,i+1} = \frac{i}{2N} \frac{2N-i}{2N} = \frac{i(2N-i)}{(2N)^2} \quad (7.2)$$

- iii) i A_1 -Gene, falls sich ein A_1 -Gen fortpflanzt und ein A_1 -Gen stirbt oder falls sich ein A_2 -Gen fortpflanzt und ein A_2 -Gen stirbt:

$$p_{ii} = \frac{i}{2N} \frac{i}{2N} + \frac{2N-i}{2N} \frac{2N-i}{2N} = \frac{i^2 + (2N-i)^2}{(2N)^2}$$

Da die restlichen Übergangswahrscheinlichkeiten gleich 0 sind, ist $P = (p_{ij})$ eine Tridiagonalmatrix.

Sind i A_1 -Gene vorhanden, so lässt sich die Fixationswahrscheinlichkeit von A_1 sehr einfach berechnen:

Mit der Notation $\mu_i := p_{i,i-1}$, $\lambda_i := p_{i,i+1}$ folgt aufgrund der Gleichheit von (7.1) und (7.2)

$$\begin{aligned} \rho_0 &= 1 \\ \rho_k &:= \frac{\mu_1 \cdots \mu_k}{\lambda_1 \cdots \lambda_k} = 1 \quad \text{für } k = 1, \dots, 2N-1. \end{aligned}$$

⁴⁵Patrick Alfred Pierce MORAN, australischer Statistiker, *14. Juli 1917 in Sydney, †19. September 1988 in Canberra;

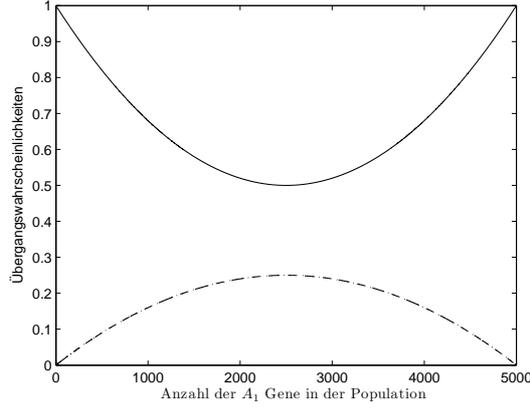


Abbildung 2: Übergangswahrscheinlichkeiten: $p_{i,i-1}=p_{i,i+1}$ ---, p_{ii} — (erstellt in MATLAB)

Da für die gesuchte Wahrscheinlichkeit laut (2.5)

$$\pi_i = \frac{\rho_0 + \dots + \rho_{i-1}}{\rho_0 + \dots + \rho_{2N-1}}$$

gilt, ergibt sich somit der Wert

$$\pi_i = \frac{\sum_{j=0}^{i-1} 1}{\sum_{j=0}^{2N-1} 1} = \frac{i}{2N}.$$

Auch einige weitere Eigenschaften aus Kapitel 2 lassen sich mit diesen Übergangswahrscheinlichkeiten und oben angeführter Notation explizit berechnen:

So tritt beispielsweise ein transientser Zustand E_j t_{ij} -mal ein, bis schließlich ein absorbierender Zustand eintritt, wobei in diesem Fall wegen (2.6)

$$\bar{t}_{ij} = \begin{cases} 0 & j = 0, 2N \\ \frac{2N(2N-i)}{2N-j} & j = 1, \dots, i \\ 2N \frac{i}{j} & j = i+1, \dots, 2N-1 \end{cases}$$

gilt. Damit ergibt sich einerseits sofort mit (2.3) nach Multiplikation mit $\frac{\pi_j}{\pi_i} = \frac{j}{i}$

$$\bar{t}_{ij}^* = \begin{cases} 0 & j = 0, 2N \\ \frac{2N(2N-i)j}{(2N-j)i} & j = 1, \dots, i \\ 2N & j = i+1, \dots, 2N-1 \end{cases}$$

und andererseits folgt für die durchschnittliche Absorptionszeit bei anfänglich transientem Zustand

$$\bar{t}_i = 2N \left(\sum_{j=1}^i \frac{2N-i}{2N-j} + \sum_{j=i+1}^{2N-1} \frac{i}{j} \right) \quad (7.3)$$

und damit wiederum

$$\bar{t}_i^* = 2N \left(\frac{(2N-i)}{i} \left(\sum_{j=1}^i \frac{j}{2N-j} \right) + 2N - i - 1 \right).$$

Beispiel. ⁴⁶ Sei $i = 1$. Es existieren also bis auf ein A_1 -Gen nur A_2 -Gene. Dann gilt für alle $j = 1, \dots, 2N - 1$:

$$\begin{aligned}\bar{t}_{1j}^* &= 2N \\ \bar{t}_1^* &= 2N(2N - 1)\end{aligned}$$

Die durchschnittliche Fixationszeit ist also genau das Produkt der Anzahl der möglichen Geburten und Sterbefälle. \square

Für dieses Modell wollen wir auch die Eigenwerte explizit berechnen:

Um die Eigenwerte angeben zu können, verwenden wir den folgenden Satz, welcher im Jahr 1974 vom amerikanischen Mathematiker Chris Cannings bewiesen wurde.

Bemerkung. Chris Cannings entwickelte ebenfalls ein Populationsmodell, welches einen allgemeineren und realistischeren Zugang als das Wright-Fisher-Modell darstellt. Ohne in dieser Arbeit genauer darauf einzugehen, soll die Idee kurz vorgestellt werden, um in Kapitel 8 und den darunter angeführten Beweis damit arbeiten zu können:

Zu diskreten Zeiten $n = 0, 1, 2, \dots$ sei die Populationsgröße mit $2N$ fest gewählt. Die Vererbung basiert auf folgender Annahme:

Betrachtet man gleichgroße Teilmengen von Nachkommen der Population zum Zeitpunkt n , so haben all diese die gleiche Verteilung zum Zeitpunkt $n + 1$.

Hat das i -te Gen also y_i Nachkommen, so folgt einerseits $y_1 + y_2 + \dots + y_{2N} = 2N$, und andererseits ist die Verteilung von (y_i, y_j, \dots, y_k) unabhängig von i, j, \dots, k .

Diese Annahmen schließen nicht aus, dass ein Gen sowohl zum Zeitpunkt n als auch zum Zeitpunkt $n + 1$ lebt.

Das Wright-Fisher-Modell ist daher nur ein Spezialfall des Cannings-Modells, in dem $(y_1, y_2, \dots, y_{2N})$ - wie in (6.9) zu sehen ist - multinomialverteilt sind.

Satz. Sei $p_{ij} = P(X_{t+1} = E_j \mid X_t = E_i)$ für $i, j = 0, 1, \dots, 2N$. Dann sind die Eigenwerte der Matrix (p_{ij}) gegeben durch

$$\begin{aligned}\lambda_0 &= 1 \\ \lambda_j &= \mathbb{E}(y_1 y_2 \dots y_j), \quad j = 1, 2, \dots, 2N,\end{aligned}\tag{7.4}$$

wobei $y_i, i = 1, \dots, j$, die Anzahl der Nachkommen des i -ten Gens angibt.

Beweis. ⁴⁷Sei $P = (p_{ij})$. Angenommen, es existiert eine nichtsinguläre Matrix Z und eine obere Dreiecksmatrix A , sodass $PZ = ZA$. Aufgrund der Nichtsingularität von Z existiert die Inverse Z^{-1} , und es gilt $P = ZAZ^{-1}$. Die Eigenwerte von P stimmen also mit jenen von A überein. Die besondere Struktur von A liefert sofort die Eigenwerte $\lambda_i = a_{ii}$ von A . Für $Z = (z_{kl})$ sind die Matrixeinträge für $k, l = 0, 1, \dots, 2N$ definiert als $z_{kl} = k^l$. Daraus ergibt sich für das Element am Schnittpunkt der Zeile i mit der Spalte j der Matrizen PZ beziehungsweise ZA

$$(PZ)_{ij} = \sum_{k=0}^{2N} p_{ik} k^j\tag{7.5}$$

$$(ZA)_{ij} = \sum_{k=0}^{2N} a_{kj} i^k = \sum_{k=0}^j a_{kj} i^k\tag{7.6}$$

⁴⁶Dieses Beispiel stammt aus [1, S.105].

⁴⁷Der Beweis orientiert sich an [1, §3.3].

Nun gilt, dass (7.5) als $\mathbb{E}(\{X_{t+1}\}^j | X_t = E_i)$ und (7.6) als $a_{jj}(i(i-1)\dots(i-j+1)) +$ weitere Terme geschrieben werden kann. Die a_{jj} sind also genau dann Eigenwerte von P , falls $\mathbb{E}(\{X_{t+1}\}^j | X_t = i) = a_{jj}(i(i-1)\dots(i-j+1)) +$ weitere Terme gilt.

Im Cannings-Modell gilt ganz allgemein die - hier nicht bewiesene - Gleichheit

$$\begin{aligned} \mathbb{E}(\{X_{t+1}\}^j | X_t = i) &= \mathbb{E}\{y_1 + y_2 + \dots + y_i\}^j & (7.7) \\ &= i\mathbb{E}\{y_1^j\} + \dots + (i(i-1)\dots(i-j+1))\mathbb{E}(y_1 y_2 \dots y_j) \end{aligned}$$

Damit folgt $a_{jj} = \mathbb{E}(y_1 y_2 \dots y_j)$ für alle $j = 0, 1, \dots, 2N$. ■

Wir betrachten nun aus der Gesamtheit von $2N$ Genen j Gene. Dann ist die Wahrscheinlichkeit, dass eines dieser Gene für die Fortpflanzung gewählt wird, gleich der Wahrscheinlichkeit, dass eines dieser Gene stirbt und trägt somit $\frac{j}{2N}$. Da ein Gen in diesem Modell üblicherweise ein Nachkomme seiner selbst ist, kann das Produkt $y_1 y_2 \dots y_j$ nur die Werte 0, 1 und 2 annehmen:

$$y_1 y_2 \dots y_j = \begin{cases} 0, & \text{falls eines der betrachteten } j \text{ Gene stirbt ohne sich fortgepflanzt zu haben.} \\ 2, & \text{falls sich eines der betrachteten } j \text{ Gene fortpflanzt ohne danach zu sterben.} \\ 1, & \text{sonst.} \end{cases}$$

Damit und mit (7.4) folgt für die gesuchten Eigenwerte

$$\begin{aligned} \lambda_0 &= 1 \\ \lambda_j &= 0 \left(\frac{j}{2N} \frac{2N-1}{2N} \right) + 2 \left(\frac{j}{2N} \frac{2N-j}{2N} \right) + 1 \left(1 - \frac{j}{2N} \frac{2N-1}{2N} - \frac{j}{2N} \frac{2N-j}{2N} \right) \\ &= 1 + \frac{2j(2N-j) - j(2N-j) - j(2N-1)}{(2N)^2} \\ &= 1 - \frac{j(j-1)}{(2N)^2}, \quad j = 1, 2, \dots, 2N \end{aligned}$$

Beispiel. Berechnung der Eigenwerte des Wright-Fisher-Modells:

Mit (7.4) sowie dem Erwartungswert einer Multinomialverteilung gilt

$$\begin{aligned} \lambda_j &= \mathbb{E}(y_1 \dots y_j) \\ &= \sum \dots \sum y_1 y_2 \dots y_j \frac{(2N)!}{y_1! \dots y_j! (2N - y_1 - \dots - y_j)!} \left(\frac{j}{2N} \right)^{\sum y_i} \left(1 - \frac{j}{2N} \right)^{2N - \sum y_i} \\ &= \frac{(2N)(2N-1)\dots(2N-j+1)}{(2N)^j} \end{aligned}$$

Die Darstellung stimmt mit den von Feller in (6.8) berechneten Eigenwerten überein. □

7.2 Moran-Modell mit Mutation

Im ersten Schritt trete ausschließlich Mutation von A_1 -Genen zu A_2 -Genen mit der Wahrscheinlichkeit $u > 0$ auf. Da keine Mutation von A_2 nach A_1 stattfindet, sterben die A_1 -Gene eventuell nach einer gewissen Zeit aus. p_{ij} ist wieder die Wahrscheinlichkeit, dass zum Zeitpunkt $n+1$ j A_1 -Gene vorhanden sind, wenn zum Zeitpunkt n die Anzahl der A_1 -Gene i war. Wie in Abschnitt 7.1 gilt $p_{ij} = 0$ für $j \notin \{i-1, i, i+1\}$. Für die restlichen Übergangswahrscheinlichkeiten gilt:

Falls zum Zeitpunkt n die Anzahl der A_1 -Gene gleich i ist, so existieren zum Zeitpunkt $n+1$ genau

- i) $i - 1$ A_1 -Gene, falls sich entweder ein A_2 -Gen fortpflanzt und ein A_1 -Gen stirbt oder ein A_1 -Gen, welches zu einem A_2 -Gen mutiert, sich fortpflanzt und ein A_1 -Gen stirbt:

$$p_{i,i-1} = \frac{i(2N-i)}{(2N)^2} + \frac{ui^2}{(2N)^2}$$

- ii) $i + 1$ A_1 -Gene, falls sich ein A_1 -Gen fortpflanzt und ein A_2 -Gen stirbt, wobei hier lediglich nichtmutierte Gene betrachtet werden:

$$p_{i,i+1} = \frac{i(2N-i)}{(2N)^2} (1-u)$$

- iii) i A_1 -Gene, falls weder $i + 1$ noch $i - 1$ A_1 -Gene existieren:

$$p_{ii} = 1 - p_{i,i-1} - p_{i,i+1}$$

Ist auch Mutation von A_2 -Genen zu A_1 -Genen mit der Wahrscheinlichkeit $v > 0$ möglich, dann ergeben sich analog zu oben die Übergangswahrscheinlichkeiten

$$\begin{aligned} p_{i,i-1} &= \frac{i(2N-i)}{(2N)^2} (1-v) + \frac{ui^2}{(2N)^2} =: \mu_i \\ p_{i,i+1} &= \frac{i(2N-i)}{(2N)^2} (1-u) + \frac{v(2N-i)^2}{(2N)^2} =: \lambda_i \\ p_{ii} &= 1 - p_{i,i-1} - p_{i,i+1} \end{aligned}$$

In diesem Fall existiert für die Anzahl der A_1 -Gene eine stationäre Verteilung ϕ , welche exakt mithilfe von (2.8) angegeben werden kann. Da dies zu eher unhandlichen Ausdrücken und länglichen Rechnungen führt, sei auf [1, S.107] verwiesen.

Beispiel. Dieses Beispiel zeigt die Übergangswahrscheinlichkeiten für das Moran-Modell mit ein- und zweiseitiger Mutation in Abhängigkeit von der relativen Anfangshäufigkeit des A_1 -Gens. Um den Unterschied auch in der Graphik darunter gut zu erkennen, sind die Inputparameter mit $2N = 5000$, $u = \frac{1}{10}$ und $v = \frac{1}{15}$ gewählt.

Während bei einseitiger Mutation die absolute Anzahl der A_1 -Gene natürlich im Laufe der Zeit immer schneller kleiner wird, muss man bei zweiseitiger Mutation in Abhängigkeit von der Anfangshäufigkeit des A_1 -Gens argumentieren:

Es existiert ein im Allgemeinen reelles $c \in [0, 2N]$, das $p_{c,c-1} = p_{c,c+1}$ erfüllt. Ist die Anfangshäufigkeit des A_1 -Gens kleiner als $\frac{c}{2N}$, so nimmt die Anzahl der A_1 -Gene zu, bis schließlich eine stationäre Verteilung eintritt. Ist die Anfangshäufigkeit des A_1 -Gens jedoch größer als $\frac{c}{2N}$, so nimmt die Anzahl der A_1 -Gene ab und die der anderen Gene zu. □

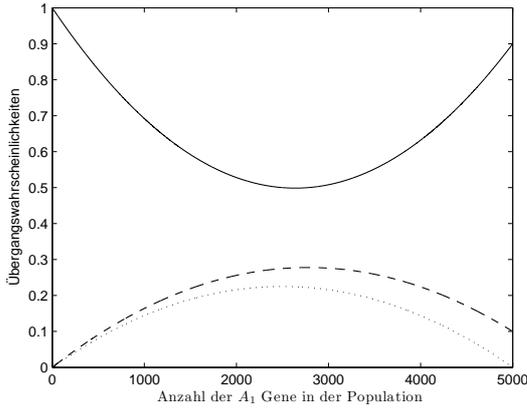
Wir betrachten nun eine Approximation der stationären Verteilung:

Aus (7.3) folgt nach einigen Umformungsschritten und Abschätzungen für den Logarithmus mit $p = \frac{i}{2N}$

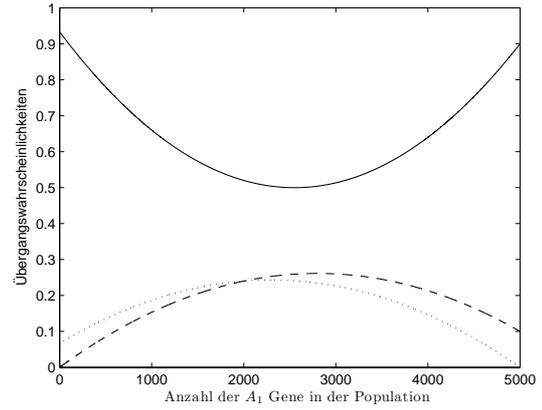
$$\bar{t}(p) \approx -(2N)^2 (p \log p + (1-p) \log(1-p))$$

Dieser Ausdruck ist sehr ähnlich zu (6.5): Der Faktor $2N$ tritt auf, weil einerseits Todesfälle und Geburten betrachtet werden, andererseits aber Generationen. Da die effektive Populationsgröße im Wright-Fisher-Modell die Hälfte der effektiven Populationsgröße⁴⁸ im Moran-Modell

⁴⁸Dafür sei auf Kapitel 8 verwiesen.



(a) Population mit einseitiger Mutation



(b) Population mit zweiseitiger Mutation

Abbildung 3: Übergangswahrscheinlichkeiten: $p_{i,i-1}$ ---, $p_{i,i+1}$ ···, p_{ii} — (erstellt mit MATLAB)

beträgt, tritt noch ein Faktor 2 auf.

Mit (2.7) und den Übergangswahrscheinlichkeiten für Mutation in beide Richtungen folgt mit $\theta = 2Nu$, $p = \frac{i}{2N}$ und $x = \frac{j}{2N}$ schließlich aus den \bar{t}_{ij} eine Approximation für die Anzahl der Geburten und Sterbefälle⁴⁹

$$\bar{t}_i \approx \frac{(2N)^2}{1-\theta} \left(\int_0^p \frac{(1-x)^{\theta-1} - 1}{x} dx + \int_p^1 \frac{(1-x)^{\theta-1}}{x} \left(1 - (1-p)^{1-\theta} \right) dx \right)$$

7.3 Moran-Modell mit Selektion

In diesem Abschnitt betrachten wir das Moran-Modell mit Selektion, welche durch unterschiedliche Sterberaten modelliert wird:

Falls die Anzahl der A_1 -Gene zum Zeitpunkt n gleich i ist, so stirbt als nächstes Gen ein A_1 -Gen mit der Wahrscheinlichkeit $\frac{\eta_1 i}{\eta_1 i + \eta_2 (2N - i)}$ mit $\eta_1, \eta_2 \in \mathbb{R}$.

Für die Übergangswahrscheinlichkeiten ergibt sich damit: Falls zum Zeitpunkt n die Anzahl der A_1 -Gene gleich i ist, so existieren zum Zeitpunkt $n + 1$ genau

- i) $i - 1$ A_1 -Gene, falls sich ein A_2 -Gen fortpflanzt und ein A_1 -Gen stirbt:

$$p_{i,i-1} = \frac{2N - i}{2N} \frac{\eta_1 i}{\eta_1 i + \eta_2 (2N - 1)} \quad (7.8)$$

- ii) $i + 1$ A_1 -Gene, falls sich ein A_1 -Gen fortpflanzt und ein A_2 -Gen stirbt:

$$p_{i,i+1} = \frac{i}{2N} \frac{\eta_2 (2N - i)}{\eta_1 i + \eta_2 (2N - 1)} \quad (7.9)$$

- iii) i A_1 -Gene, falls weder $i + 1$ noch $i - 1$ A_1 -Gene existieren:

$$p_{ii} = 1 - p_{i,i-1} - p_{i,i+1}$$

⁴⁹siehe zum Beispiel [1, S. 108]

Anhand dieser Übergangswahrscheinlichkeiten kann man folgende Fälle unterscheiden:

- i) Gilt $\eta_1 = \eta_2$, so tritt keine Selektion auf, da $p_{i,i-1} = p_{i,i+1}$ gilt.
- ii) Für $\eta_1 < \eta_2$ gilt $p_{i,i-1} < p_{i,i+1}$. Damit hat das A_1 -Gen einen selektiven Vorteil gegenüber dem A_2 -Gen.
- iii) Für $\eta_1 > \eta_2$ gilt $p_{i,i-1} > p_{i,i+1}$. Damit hat das A_1 -Gen einen selektiven Nachteil gegenüber dem A_2 -Gen.

Beispiel. Wir betrachten eine Population mit einem Genpool der Größe $2N = 5000$ und dem Allelvorrat $\{A_1, A_2\}$.

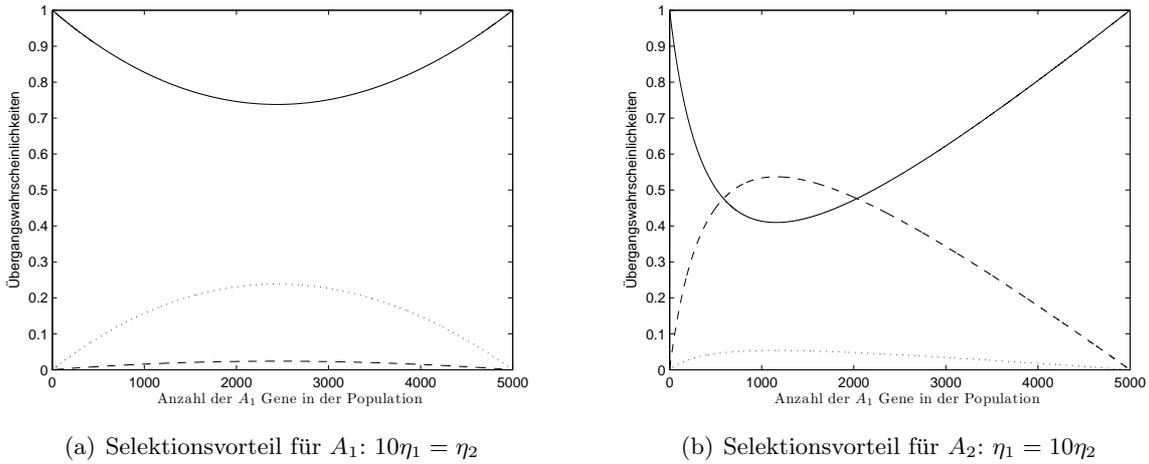


Abbildung 4: Übergangswahrscheinlichkeiten: $p_{i,i-1}$ — — —, $p_{i,i+1}$ ···, p_{ii} — (erstellt mit MATLAB)

□

An den gegebenen Übergangswahrscheinlichkeiten erkennt man, dass es sich bei der Matrix dieser Wahrscheinlichkeiten wiederum um eine Tridiagonalmatrix handelt. Für die Eigenwerte dieser Matrix existiert bis heute keine geschlossene Form. Allerdings lassen sich diverse Ergebnisse aus Kapitel 2 anwenden: So gilt zum Beispiel mit der Notation aus dem eben genannten Kapitel

$$\rho_0 = 1$$

$$\rho_k = \prod_{j=1}^k \frac{\mu_j}{\lambda_j} = \left(\frac{\eta_1}{\eta_2}\right)^k.$$

Die Fixationswahrscheinlichkeit des A_1 -Gens lässt sich nun mit (2.5) bestimmen:

$$\pi_i = \frac{1 - \left(\frac{\eta_1}{\eta_2}\right)^i}{1 - \left(\frac{\eta_1}{\eta_2}\right)^{2N}}$$

Ohne diesen Rechenweg genauer nachzuprüfen, soll hier eine andere recht einfache Berechnungsart⁵⁰ der Fixationswahrscheinlichkeit gezeigt werden:

⁵⁰vergleiche [3, S.155-157] und [8, S.119]

Da für die Übergangswahrscheinlichkeiten

$$p_{i,i+1} + p_{i,i-1} + p_{ii} = 1 \quad (7.10)$$

gilt, folgt die Differenzgleichung

$$\pi_i = \pi_{i+1}p_{i,i+1} + \pi_{i-1}p_{i,i-1} + \pi_i p_{ii}.$$

Aufgrund der Selektion gilt weiters mit

$$p_{i,i+1} = \frac{\eta_2}{\eta_1} p_{i,i-1}$$

und (7.10)

$$\begin{aligned} \frac{\eta_2}{\eta_1} \pi_{i+1} - \left(1 + \frac{\eta_2}{\eta_1}\right) \pi_i + \pi_{i-1} &= 0 \\ \Leftrightarrow \zeta^2 - \left(1 + \frac{\eta_1}{\eta_2}\right) \zeta + \frac{\eta_1}{\eta_2} &= 0 \end{aligned}$$

Daraus ergeben sich für ζ die beiden Werte 1 und $\frac{\eta_1}{\eta_2}$, womit die allgemeine Lösung

$$\pi_i = c_1 + c_2 \left(\frac{\eta_1}{\eta_2}\right)^i$$

mit $c_1, c_2 \in \mathbb{R}$ lautet. Mithilfe der beiden Randbedingungen $\pi_0 = 0$ und $\pi_{2N} = 1$ ergibt sich

$$\pi_i = \frac{1 - \left(\frac{\eta_1}{\eta_2}\right)^i}{1 - \left(\frac{\eta_1}{\eta_2}\right)^{2N}}$$

Bemerkung. Es gibt natürlich auch Ansätze mit unterschiedlichen Geburtenraten und einer Kombination aus unterschiedlichen Geburten- und Sterberaten. Da diese Ansätze sehr ähnlich zum oben angeführten Fall sind, werden sie hier nicht extra diskutiert.

7.4 Moran-Modell für unendlich viele Allele

Wie im Modell für zwei Alleltypen liegt für diese Art des Moran-Modells ein bestimmter Sterbe- und Geburtenprozess vor. Analog zum Wright-Fisher-Modell für unendlich viele verschiedene Alleltypen ist ein Nachkomme mit der Wahrscheinlichkeit u ein neuer noch nie dagewesener Alleltyp. Die Übergangswahrscheinlichkeiten bilden wieder eine Tridiagonalmatrix, deren Einträge - genauso wie bereits im Moran-Modell mit beidseitiger Mutation - durch

$$\begin{aligned} p_{i,i-1} &= \frac{i(2N-i)(1-v) + ui^2}{(2N)^2} \\ p_{i,i+1} &= \frac{i(2N-i)(1-u) + v(2N-i)^2}{(2N)^2} \\ p_{ii} &= 1 - p_{i,i-1} - p_{i,i+1} \end{aligned}$$

gegeben sind. Wählt man ein Allel, so kann keine Aussage über ein stationäres Verhalten gemacht werden. Betrachtet man hingegen eine bestimmte Genstruktur, so existiert wie in den vorherigen Modellen eine stationäre Verteilung. Der Unterschied besteht darin, dass eine exakte

Wahrscheinlichkeit für jede beliebige Struktur angegeben werden kann.

Sei nun β_j die Anzahl von Alleltypen mit genau j Genen in der Population. Dann gilt $\sum_j j\beta_j = 2N$. Die stationäre Verteilung gibt der australische Statistiker Albert Trajstman in [11] so an:

$$P(\beta_1, \dots, \beta_{2N}) = \frac{(2N)! \theta^{\sum_j \beta_j}}{S_{2N}(\theta) \prod_{j=1}^{2N} j^{\beta_j} \beta_j!}$$

mit $S_j(\theta) = \theta(\theta + 1) \dots (\theta + j - 1)$ und $\theta = \frac{2Nu}{1-u}$. Weil in diesem Modell nicht die effektive Populationsgröße verwendet wird, weicht θ von jenem des Wright-Fisher-Modells ab. Diese Gleichung entspricht mit $n = 2N$ der von Ewens entwickelten Formel (6.17), wodurch die gleichen Folgerungen wie in Abschnitt 6.2 gültig sind. Allerdings können im Gegensatz zum Wright-Fisher-Modell - wie bereits ganz zu Beginn des Kapitels erwähnt - explizite Ausdrücke, wie zum Beispiel das Häufigkeitsspektrum, angegeben werden, welche in [1, S.118-119] zu finden sind.

Kapitel 8

Effektive Populationsgrößen

Obwohl das Wright-Fisher-Modell in der bisher behandelten Form für die Beschreibung biologischer Zusammenhänge ungeeigneter als das Cannings-Modell oder das Moran-Modell ist, wird es aufgrund des großen theoretischen Wissens darüber sehr häufig verwendet. Damit diverse Aussagen in vorhergehenden Kapiteln auch für Populationen mit zusätzlichen Eigenschaften, wie zum Beispiel bei Unterscheidung zweier Geschlechter, korrekt bleiben, verwendet man die sogenannten effektiven Populationsgrößen:

Definition 8.1

- i) Die effektive Populationsgröße bezüglich der Eigenwerte ist definiert als

$$N_e^{EW} := \frac{1}{2} \frac{1}{(1 - \lambda_{max})}, \quad (8.1)$$

wobei λ_{max} der größte Eigenwert ungleich Eins der Übergangsmatrix des betrachteten Modells ist.

- ii) Die effektive Populationsgröße bezüglich Inzucht ist definiert als

$$N_e^I := \frac{1}{2} \frac{1}{\pi_2}, \quad (8.2)$$

wobei π_2 die Wahrscheinlichkeit ist, dass zwei zufällig gezogene Gene die gleichen Eltern haben.

- iii) Die effektive Populationsgröße bezüglich der Varianz ist definiert als

$$N_e^{Var} := \frac{1}{2} \frac{x_n(1 - x_n)}{Var(x_{n+1} | x_n)}, \quad (8.3)$$

wobei x_n den relativen Anteil der A_1 -Gene zum Zeitpunkt n angibt.

8.1 Cannings-Modell

Ohne viel über das Cannings-Modell zu wissen, da es nur sehr kurz in der Bemerkung in Abschnitt 7.1 erwähnt wurde, werden in diesem Teil die drei verschiedenen effektiven Populationsgrößen für das Cannings-Modell mit nichtüberlappenden Generationen explizit bestimmt:

Für die Berechnung von N_e^{EW} muss zuerst der größte Eigenwert ungleich Eins bestimmt werden. Laut (7.4) gilt $\lambda_2 = \mathbb{E}(y_1 y_2)$. Da y_i die Anzahl der Nachkommen des i -ten Gens ist,

folgt $\sum_{j=1}^{2N} y_j = 2N$ und damit $Var(\sum_{j=1}^{2N} y_j) = 0$. Wir definieren $Var(y_i) =: \sigma^2$ für alle $i = 1, \dots, 2N$. Mit der Identität

$$Var\left(\sum_{j=1}^{2N} y_j\right) = \sum_{j=1}^{2N} Var(y_j) + 2 \sum_{j=1}^{2N-1} \sum_{k=j+1}^{2N} Cov(y_i, y_j) \quad (8.4)$$

folgt aufgrund der Symmetrie

$$2N\sigma^2 + 2N(2N-1)Cov(y_i, y_j) = 0 \quad (8.5)$$

$$\Leftrightarrow Cov(y_i, y_j) = -\frac{\sigma^2}{2N-1} \quad (8.6)$$

Insgesamt ergibt sich für den Eigenwert mit dem Verschiebungssatz von Steiner und $\mathbb{E}(y_i) = 1$, $i = 1, \dots, 2N$, da jedes Gen im Schnitt genau einen Nachkommen hat,

$$\begin{aligned} \lambda_2 &= \mathbb{E}(y_1 y_2) \\ &= Cov(y_1, y_2) + \mathbb{E}(y_1)\mathbb{E}(y_2) \\ &= 1 - \frac{\sigma^2}{2N-1}. \end{aligned}$$

Die effektive Populationsgröße bezüglich der Eigenwerte ist laut Definition (8.1) daher durch

$$N_e^{EW} = \frac{2N-1}{2\sigma^2}$$

gegeben.

Für die effektive Populationsgröße bezüglich der Varianz muss ein Ausdruck für $Var(X_{n+1} | X_n)$ gefunden werden: Angenommen es existieren i A_1 -Gene zum Zeitpunkt n , das heißt es gilt $X_n = i$. Dann sortiert man die Gene so, dass die ersten i Gene A_1 -Gene sind. Das heißt $y_1 + \dots + y_i$ sind die gesamten Nachkommen aller A_1 -Gene. Dann folgt mit (8.4) und (8.5) - (8.6)

$$\begin{aligned} Var(X_{n+1} | X_n = i) &= Var(y_1 + \dots + y_i) \\ &= i\sigma^2 + i(i-1)Cov(y_1, y_2) \\ &= \frac{i(2N-i)\sigma^2}{2N-1} \end{aligned}$$

Damit gilt mit $x_n = \frac{X_n}{2N}$:

$$Var(x_{n+1} | x_n) = \frac{x_n(1-x_n)\sigma^2}{2N-1}$$

und mit (8.3)

$$N_e^{Var} = \frac{2N-1}{2\sigma^2}.$$

Zur Berechnung der effektiven Populationsgröße bezüglich Inzucht sei wieder um y_i die Anzahl der Nachkommen des i -ten Gens der n -ten Generation. Damit gilt $\sum_{j=1}^{2N} y_j = 2N$, und die

Wahrscheinlichkeit P_2 , dass zwei zufällig gezogene Gene der $(n + 1)$ -ten Generation gleiche Eltern haben bedingt unter gegebenen y_1, \dots, y_{2N} , ist nach einfacher Wahrscheinlichkeitsrechnung gegeben durch

$$P_2 = \sum_{i=1}^{2N} \frac{y_i(y_i - 1)}{2N(2N - 1)}$$

Für den Erwartungswert π_2 dieser Summe gilt wegen $\mathbb{E}(y_i) = 1$, $Var(y_i) = \sigma^2$ und der Symmetrie

$$\begin{aligned} \pi_2 &= \mathbb{E} \left(\sum_{i=1}^{2N} \frac{y_i(y_i - 1)}{2N(2N - 1)} \right) \\ &\stackrel{Symm.}{=} \frac{2N (\mathbb{E}(y_i^2) - \mathbb{E}(y_i))}{2N(2N - 1)} \\ &\stackrel{\mathbb{E}(y_i)=1}{=} \frac{2N (\mathbb{E}(y_i^2) - (\mathbb{E}(y_i))^2)}{2N(2N - 1)} \\ &= \frac{Var(y_i)}{(2N - 1)} \\ &= \frac{\sigma^2}{2N - 1}. \end{aligned}$$

Damit folgt mit (8.2)

$$N_e^I = \frac{2N - 1}{2\sigma^2}.$$

Bemerkung. Für das Cannings-Modell stimmen also alle drei oben betrachteten Arten der effektiven Populationsgröße überein.

Betrachtet man nur den führenden Term, so gilt für Modelle mit nichtüberlappenden Generationen für die effektive Populationsgröße $N_e = \frac{N}{\sigma^2}$. Damit kann nun eine Verbindung zum Wright-Fisher-Modell für unendlich viele Allele hergestellt werden:

Laut der abschließenden Bemerkung von Abschnitt 6.2 bleiben die Aussagen jenes Kapitels richtig, falls $\theta = 4N_e u$ gewählt wird. Setzt man $\theta = 4N_e u = \frac{4Nu}{\sigma^2}$, so bleiben die Aussagen auch für das Cannings-Modell mit nichtüberlappenden Generationen gültig.

Für Modelle mit überlappenden Generationen ist eine sinnvolle Definition für die effektive Populationsgröße $\frac{sN_e}{2N}$, wobei s die Anzahl der Sterbefälle pro Zeiteinheit und N_e eine der drei oben definierten effektiven Populationsgrößen ist. Da für Modelle mit nichtüberlappenden Generationen stets $s = 2N$ gilt, ist diese Definition für beide Arten von Populationen anwendbar. Weiters folgt mit dieser Definition auch sofort die effektive Populationsgröße für das Moran-Modell. In diesem Fall ist $k = 1$, und damit ergibt sich $N_e^{EW} = N_e^I = N_e^{Var} = \frac{1}{2}N$. Die effektive Populationsgröße eines Moran-Modells ist also doppelt so groß wie die effektive Populationsgröße im Wright-Fisher-Modell:

Während im Wright-Fisher-Modell $\sigma^2 \approx 1$ gilt, ist im Moran-Modell $\sigma^2 \approx \frac{1}{N} = \frac{2}{2N}$. Der noch auftretende Faktor $\frac{1}{2N}$ ergibt sich wiederum wegen der Betrachtung von Sterbefällen und Geburten anstatt Generationen.

Abschließend betrachten wir noch die effektive Populationsgröße bezüglich Inzucht $N_e^{I,d}$ für das Cannings-Modell für eine diploide Population. $N_e^{I,d}$ ist definiert als die reziproke Wahrscheinlichkeit, dass zwei zufällig gezogene Gene der $(n+1)$ -ten Generation gleiche Eltern in der n -ten Generation haben. Für das Cannings-Modell heißt das anders formuliert:

Wähle aus der Generation n zwei Gene und schaue, ob zwei zufällig gewählte Gene der $(n+1)$ -ten Generation beide von einem Gen oder je von einem der beiden Gene der Generation n abstammen. Die gesuchte Wahrscheinlichkeit π_2 ist dann der Erwartungswert von

$$\sum_{i=1}^N \frac{(y_i + y_{N+i})(y_i + y_{N+i} - 1)}{2N(2N - 1)}.$$

Sei nun σ_d^2 die Varianz des diploiden Modells, so ergibt sich analog zu obiger Rechnung mit

$$\pi_2 = \mathbb{E} \left(\sum_{i=1}^N \frac{(y_i + y_{N+i})(y_i + y_{N+i} - 1)}{2N(2N - 1)} \right)$$

die gesuchte Populationsgröße

$$N_e^{I,di} = \frac{4N - 2}{\sigma_d^2 + 2}.$$

Das diploide Cannings-Modell ist hierbei definiert als Cannings-Modell, in dem die Austauschbarkeit, welche in Abschnitt 7.1 kurz erwähnt wird, für diploide Individuen gilt. Mit zusätzlichen Annahmen kann σ_d^2 so gewählt werden, dass $N_e^I = N_e^{I,d}$ gilt⁵¹.

8.2 Wright-Fisher-Modell

In diesem Abschnitt betrachten wir nun die effektiven Populationsgrößen für das Wright-Fisher-Modell mit zusätzlichen Eigenschaften:

Existieren in einer Population zwei verschiedene Geschlechter, so seien N^m beziehungsweise N^w die Anzahl diploider männlicher beziehungsweise weiblicher Individuen der Population der Größe N . Da es sich um eine diploide Population handelt, erhält ein Nachkomme je ein Gen vom Vater und eines der Mutter, womit die Anzahl der A_1 -Gene zum Zeitpunkt $n+1$ eine Summe von Binomialverteilungen ist:

$$\begin{aligned} X_{n+1}^m &= \gamma_{n+1}^m + \delta_{n+1}^m \\ X_{n+1}^w &= \gamma_{n+1}^w + \delta_{n+1}^w \end{aligned}$$

mit $\gamma_{n+1}^m \sim \text{Bin}(N^m, \frac{X_n^m}{2N^m})$, $\delta_{n+1}^m \sim \text{Bin}(N^m, \frac{X_n^w}{2N^m})$ und analog $\gamma_{n+1}^w \sim \text{Bin}(N^w, \frac{X_n^m}{2N^w})$, $\delta_{n+1}^w \sim \text{Bin}(N^w, \frac{X_n^w}{2N^w})$.

Damit existiert eine Übergangsmatrix, deren größten Eigenwert ungleich Eins wir nun benötigen, um N_e^{EW} zu berechnen. Dazu suchen wir eine Funktion, welche

$$Y(X^m, X^w) \begin{cases} = 0 & \text{für absorbierende Zustände} \\ > 0 & \text{sonst} \end{cases}$$

⁵¹siehe [1, S.122]

erfüllt. Außerdem soll für $\lambda > 0$

$$\mathbb{E}[Y(X_{n+1}^m, X_{n+1}^w) | X_n^m, X_n^w] = \lambda Y_n$$

gelten. Da solche Funktionen stets existieren, allerdings schwer zu finden sind, sei hier auf [1, §3.7] verwiesen, wo die Lösung

$$Y(X^m, X^w) = \frac{C}{2} \left(\frac{X^m(2N^m - X^m)}{(2N^m)^2} + \frac{X^w(2N^w - X^w)}{(2N^w)^2} \right) + \left(1 - \frac{(X^m - N^m)(X^w - N^w)}{N^m N^w} \right)$$

mit $C = \frac{1}{2} \left(1 - \sqrt{1 - \frac{2}{N^m} - \frac{2}{N^w}} \right)$ angegeben wird. Daraus ergibt sich

$$\begin{aligned} \lambda &= \frac{1}{2} \left(1 - \frac{1}{4N^m} - \frac{1}{4N^w} + \sqrt{1 + \frac{(N^m + N^w)^2}{(4N^m N^w)^2}} \right) \\ \Rightarrow \lambda &\approx 1 - \frac{N^m + N^w}{8N^m N^w} = 1 - \frac{N}{8N^m N^w} \end{aligned}$$

Für die effektive Populationsgröße bezüglich der Eigenwerte gilt also im Wright-Fisher-Modell unter Berücksichtigung zweier Geschlechter mit (8.1)

$$N_e^{EW} \approx \frac{4N^m N^w}{N}$$

Für die Berechnung der effektiven Populationsgröße bezüglich Inzucht benötigen wir die Wahrscheinlichkeit π_2 :

Zwei zufällig gezogene Gene einer Generation haben identische Elterngene, falls beide vom gleichen männlichen oder weiblichen Gen abstammen. π_2 ist daher gegeben durch

$$\frac{1}{2} \frac{N-2}{2N-1} \frac{1}{2N^m} + \frac{1}{2} \frac{N-2}{2N-1} \frac{1}{2N^w}.$$

Mit (8.2) gilt

$$N_e^I = \frac{1}{2\pi_2} \approx \frac{4N^m N^w}{N}$$

Die Berechnung von N_e^{Var} ist für eine zweigeschlechtliche Population mit bisherigen Mitteln nicht möglich, da der Anteil an A_1 -Genen keine Markov'sche Zufallsvariable ist. Einen Weg dieses Defizit der Definition zu umgehen, bieten sogenannte Quasi-Markov'sche Zufallsvariablen. Ohne hier näher darauf einzugehen, sei auf entsprechende Literatur⁵² verwiesen und nur erwähnt, dass

$$N_e^{Var} = \frac{4N^m N^w}{N}$$

gilt. Damit sind auch in dieser Art von Modell alle effektiven Populationsgrößen von gleicher Größenordnung.

Für effektive Populationsgrößen mit anderen Eigenschaften der Population, sei auf [1, Seiten 124-126] und [8, Seiten 172-186] verwiesen.

⁵²siehe zum Beispiel [1, §4]

Bemerkung.

- i) Hier wurde für die Berechnung der effektiven Populationsgrößen immer nur eine zusätzliche Annahme betrachtet. Natürlich ist dies noch immer nicht sehr realitätsnah, da die verschiedensten Populationen eine Vielzahl individueller Eigenschaften aufweisen. Allerdings ist die Verwendung effektiver Populationsgrößen ein Schritt in Richtung besserer Modellierung.
- ii) Obwohl die effektiven Populationsgrößen in den oben betrachteten Fällen von gleicher Größenordnung sind, ist dies nicht immer der Fall:
Betrachtet man zum Beispiel Populationen mit nichtkonstanter Größe in der ein einziger heterozygoter Elternteil eine sehr große Anzahl an Kindern bekommt, sind N_e^{EW} und N_e^{Var} sehr groß während N_e^I ungefähr 1 ist.
Das lässt darauf schließen, dass N_e^{EW} und N_e^{Var} eher durch die zukünftige Evolution der Population geprägt sind, während die vergangene Entwicklung N_e^I beeinflusst.
Je nach Erfordernissen muss daher die passende der drei effektiven Populationsgrößen gewählt werden, um die beste Modellierung der Realität zu bekommen.

Kapitel 9

Schlussbemerkung

Ausgehend von den Mendel'schen Vererbungsregeln lässt sich eine Vielzahl verschiedener Modelle, welche die Evolution von Genen beschreiben, entwickeln:

Deterministische Ansätze, welche bereits vor den stochastischen entwickelt wurden, lassen sich lediglich durch die unbegrenzte Populationsgröße rechtfertigen. Obwohl natürlich diese Modelle eine erste nicht zu verachtende Modellierung der Realität bringen, ist trotz allem die Voraussetzung einer unendlich großen Population nie gegeben.

Daher ist das Wright-Fisher-Modell, bei dem von einer konstanten Größe des Genpools von $2N$ ausgegangen wird, bereits ein erster Schritt in Richtung realitätsnaher Modellierung. Aufgrund der einfachen Struktur dieses Modells ist das Wissen darüber heute schon sehr groß. Dies ist der Grund, warum man dieses Modell realistischeren Modellen in vielen Fällen vorzieht. Natürlich darf man auch die Verallgemeinerungen des Modells durch Mutation und Selektion nicht außer Acht lassen. Denn genau durch diese zusätzlichen Informationen, welche mehr oder weniger gut modellierbar sind, rückt dieses Modell wieder ein Stück näher an die Realität.

Beim Cannings-Modell unterliegt die Evolution keiner Multinomialverteilung, wie es beim Wright-Fisher-Modell der Fall ist. Das alleine ermöglicht natürlich schon eine bessere Modellierung der Realität. Allerdings ist über dieses Modell (noch) nicht so viel bekannt wie über das Wright-Fisher-Modell.

Einen ganz anderen Zugang wählt das Modell von Moran. Dadurch ist es leichter zu handhaben und mit den bis jetzt bekannten mathematischen Mitteln liefert es viele explizite Darstellungen. Die Verallgemeinerungen durch Mutation, Selektion beziehungsweise durch Vergrößern des Allelvorrates ermöglichen wie auch beim Wright-Fisher-Modell eine bessere Modellierung.

Unabhängig vom betrachteten Modell ist die Definition von effektiven Populationsgrößen ein großer Schritt: Durch eine teilweise recht einfache Berechnung der effektiven Populationsgröße kann das Modell erheblich verbessert werden.

Zusammenfassend hat jedes Modell, egal ob deterministisch oder stochastisch, seine Berechtigung. Um die beste Modellvariante für die Lösung zu verwenden, muss man abhängig von der Aufgabenstellung die Vor- und Nachteile jedes einzelnen Modells abwägen.

Weil die Forschung in diesem Gebiet noch relativ jung ist, werden immer wieder neue Ansätze und Modelle entwickelt, wodurch die Realität besser modelliert werden kann.

Literaturverzeichnis

- [1] **Warren J. EWENS:**
Mathematical Population Genetics; I. Theoretical Introduction. Second Edition
Springer, New York 2004
- [2] **Warren J. EWENS:**
The Sampling Theory of Selectively Neutral Alleles
Theoret. Population Biol. 3 (1972), 87-112.
- [3] **J. S. GALE:**
Theoretical Population Genetics
Unwin Hyman, London 1990
- [4] **Josef HOFBAUER/ Karl SIGMUND:**
The Theory of Evolution and Dynamical Systems
Cambridge University Press, Cambridge 1988
- [5] **Fred M. HOPPE:**
The sampling theory of neutral alleles and an urn model in population genetics.
J. Math. Biol. 25 (1987), 123-159.
- [6] **N. L. JOHNSON/ S. KOTZ/ N. BALAKRISHAN:**
Discrete Multivariate Distributions
John Wiley & Sons, New York 1997
- [7] **Gregor MENDEL:**
Versuche über Pflanzen-Hybriden
Gedruckt in den Verhandlungen des naturforschenden Vereins in Brünn; IV. Band; Abhandlungen 1865; Brünn, 1866; Im Verlage des Vereins. Seiten 3-47.
- [8] **Patrick A. P. MORAN:**
The Statistical Processes of Evolutionary Theory
Oxford University Press, New York 1962
- [9] **Wilfried NÖBAUER/ Werner TIMISCHL:**
Mathematische Modelle in der Biologie
Vieweg, Braunschweig-Wiesbaden 1979
- [10] **Jan W. PRÜSZ/ Roland SCHNAUBELT/ Rico ZACHER:**
Mathematische Modelle in der Biologie - Deterministische homogene Systeme.
Birkhäuser, Basel 2007
- [11] **A. C. TRAJSTMAN:**
On a conjecture of G.A. Watterson
Adv. Appl. Prob. 6 (1974), 489-493.

- [12] **Sewall WRIGHT:**
The Distribution of Gene Frequencies in Populations
Proc. National Acad. Sci. USA 23 (1937), 307-320