

Biostatistik, SS 2016

Matthias Birkner

<http://www.staff.uni-mainz.de/birkner/Biostatistik16/>

22.4.2016



JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

Herzlich willkommen zur Vorlesung
Biostatistik im SS 2016!



JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

Inhalt

- 1 Organisatorisches
 - Themenplan
 - Literatur
- 2 Dreisatz, lineare Gleichungen
 - Dreisatz
 - Gleichungen
- 3 Funktionen und Folgen
 - Folgen und Funktionen: mathematische Begriffe und Definitionen
 - Beispiel: Hardy-Weinberg-Gleichgewicht
 - Beispiel: Allelhäufigkeiten unter Selektion

Orte, Zeiten, Übungen, Zulassungskriterien, ...

- Fr. 8–10, Hörsaal 18
- <http://www.staff.uni-mainz.de/birkner/Biostatistik16/>
(dort stehen die Folien, Verweise auf die Übungen, Informationen)
- **Klausur: Mo., 25.7.2016**
(Für B.Sc.-Studenten: mind. 25% der Übungspunkte)
- **Übungen:** elektronisch via *ilias*, von Frederik Klement f.klement@uni-mainz.de betreut. Abgabe jeweils bis montags um 10h (10 Tage Bearbeitungszeit)
- **Tutorien:** 2h/Woche, für Fragen, Besprechung der Übungsaufgaben, etc.
Termine und Informationen zur Anmeldung:
<https://www.stochastik.mathematik.uni-mainz.de/frederik-klement/biostatistik-ss-16/>
- Blatt 0 : wird nächste Woche in den Tutorien besprochen

Voraussichtliche Themen

Mathematik-Wiederholung

- Dreisatz, lineare Gleichungen, Folgen, Funktionen
- Exponential- und Logarithmusfunktion
- Differential- und Integralrechnung

Einführung in die Statistik

- Deskriptive Statistik
- Standardfehler
- Grundbegriffe der Wahrscheinlichkeitstheorie
- t -Test
- Lineare Regression
- Schätzer, Konfidenzintervalle
- Kontingenztafeln und χ^2 -Test
- Rangtests

Meta-Lernziele

- 1 Mathematik und Statistik stellen nützliche Werkzeuge für die Biologie bereit
- 2 Mathematik und Statistik sind (gar) nicht (so) schwer

Literaturhinweise (UB Lehrbuchsammlung)

Es gibt (meines Wissens nach) kein Buch, das exakt nur den Vorlesungsstoff umfasst. Folgende Bücher sind jeweils mehrfach in der UB vorhanden, manche auch als E-Book. Sie überdecken den Stoff der Vorlesung (für den „mathematischen“ Teil 1–3, für den „statistischen“ Teil 4–6) und gehen z.T. deutlich darüber hinaus.

- 1 E. Bohl, *Mathematik in der Biologie*, 4. Aufl., Springer, 2006.
- 2 H. Vogt, *Grundkurs Mathematik für Biologen*, 2. Aufl., Teubner, 1994.
- 3 A. Riede, *Mathematik für Biologen*, Vieweg, 1993.
- 4 F. Bärlocher, *Biostatistik*, Thieme, 1999.
- 5 W. Timischl, *Biostatistik : eine Einführung für Biologen und Mediziner*, 2. Aufl., Springer, 2000.
- 6 W. Köhler, G. Schachtel, P. Voleske, *Biostatistik : eine Einführung für Biologen und Agrarwissenschaftler*, 4. Aufl., Springer, 2007.

Empfehlung: In verschiedene Bücher „hineinschnuppern“
(Frage des persönlichen Stils und Geschmacks)

Quellen und Dank

Diese Vorlesung basiert zu weiten Teilen auf Material von Brooks Ferebee, Gaby Schneider und Anton Wakolbinger (Universität Frankfurt) sowie von Martin Hutzenthaler und Dirk Metzler (LMU München).

Darüberhinaus habe ich Material von Achim Klenke (Mainz) und einige Beispiele von Hans Schuh (Mainz) übernommen.

Ich danke ihnen herzlich für die Bereitstellung des Materials und vieler Beispiele.

Dreisatzrechnung: Anwendbar, wenn zwei Meßgrößen zueinander proportional bzw. indirekt proportional (d.h. die eine proportional zum Kehrwert der anderen) sind.

Beispiel 1. Eine vorgegebene Proteinlösung hat eine Konzentration von $20 \frac{\text{mg}}{\text{ml}}$. Wir möchten $500 \mu\text{l}$ einer Proteinlösung mit einer Konzentration von $0,3 \frac{\text{mg}}{\text{ml}}$ erstellen. Wie viele μl der vorliegenden Proteinlösung (und wieviel zusätzliches reines Lösungsmittel) werden benötigt?

Erinnerung:

$$1\text{l} = 1000\text{ml} = 1\,000\,000\mu\text{l}$$

$$\text{Konzentration (des Proteins in der Lösung)} = \frac{\text{gelöste Proteinmasse}}{\text{Volumen der Lösung}},$$

also sind bei konstanter Konzentration die Masse der gelösten Substanz und das Volumen der Lösung zueinander proportional (d.h. Verdopplung der Masse entspricht Verdopplung des Volumens, etc.).

Beispiel 1. Eine vorgegebene Proteinlösung hat eine Konzentration von $20 \frac{\text{mg}}{\text{ml}}$. Wir möchten $500 \mu\text{l}$ einer Proteinlösung mit einer Konz. von $0,3 \frac{\text{mg}}{\text{ml}}$ erstellen. Wie viele μl der vorliegenden Proteinlösung werden benötigt?

Lösung (Direkter Dreisatz):

(Bei konstanter Konzentration sind Masse (an gelöstem Protein) und Volumen (der Lösung) zueinander proportional.)

1. Schritt : Berechne die in der neuen Lösung gelöste Proteinmasse y

$$\begin{array}{rcl}
 1 \text{ ml} = 1000 \mu\text{l} & \text{enthalten} & 0,3 \text{ mg} \\
 1 \mu\text{l} & \text{enthält} & \frac{0,3}{1000} \text{ mg} \\
 500 \mu\text{l} & \text{enthalten} & y = \frac{0,3 \cdot 500}{1000} \text{ mg} = 0,15 \text{ mg}
 \end{array}$$

2. Schritt : Berechne das benötigte Volumen x der gegebenen Lösung

$$\begin{array}{rcl}
 20 \text{ mg} & \text{in} & 1 \text{ ml} \\
 1 \text{ mg} & \text{in} & \frac{1}{20} \text{ ml} \\
 0,15 \text{ mg} & \text{in} & x = 0,15 \cdot \frac{1}{20} \text{ ml} = 0,0075 \text{ ml} = 7,5 \mu\text{l}
 \end{array}$$

Gewünschtes Gesamtvolumen = $500 \mu\text{l} = 7,5 \mu\text{l} + 492,5 \mu\text{l}$

Demnach : $7,5 \mu\text{l}$ der vorgegebenen Lösung müssen mit $492,5 \mu\text{l}$ reinen Lösungsmittels vermischt werden.

Beispiel 2. 10 μl eines 7 M Harnstoffes wird mit einer Pufferlösung auf 500 μl verdünnt. Was ist die Konzentration des Harnstoffes in der verdünnten Lösung?

Erinnerung (Stoffmengenkonzentration, „Molarität“):

1 mol eines Stoffes sind $6,022 \cdot 10^{23}$ Teilchen (Avogadro-Zahl),
 Masse von 1 mol eines Stoffes
 = (relative) Molekül-/Atommasse des Stoffes in g

Schreibweise:

$M = \frac{\text{mol}}{\text{l}}$ („molar“), 7 M Harnstoff („7 molare(r) Harnstoff(-Lösung)“)
 bezeichnet also eine Lösung, die 7 mol Harnstoff pro Liter enthält.

Lösung (Indirekter Dreisatz):

Konzentration = $\frac{\text{Masse}}{\text{Volumen}}$, also sind bei konstanter Masse der gelösten Substanz die Konzentration und das Volumen der Lösung zueinander indirekt proportional.

In 10 μl gelöster Harnstoff ist 7 M

In 500 μl gelöster Harnstoff ist $x = 7 \cdot \frac{10}{500} \text{ M} = 0,14 \text{ M}$

Demnach : Die verdünnte Harnstofflösung ist 0,14 M.

Bemerkung.

Die chemische Summenformel von Harnstoff ist $\text{CN}_2\text{H}_4\text{O}$, die (relative) Molekülmasse von Harnstoff ist demnach $12 + 2 \cdot 14 + 4 \cdot 1 + 16 = 60$.

Somit: 1 mol Harnstoff entspricht 60 g Harnstoff, die Konzentration der 0,14 M Harnstofflösung aus Beispiel 2, ausgedrückt in mg/ml, ist also

$$0,14 \frac{\text{mol}}{\text{l}} \cdot 60 \frac{\text{g}}{\text{mol}} = 8,4 \frac{\text{g}}{\text{l}} = 8,4 \frac{1000\text{mg}}{1000\text{ml}} = 8,4 \frac{\text{mg}}{\text{ml}}.$$

Ein erneuter Blick auf die Beispielrechnung

Beispiel 1. Eine vorgegebene Proteinlösung hat eine Konzentration von $20 \frac{\text{mg}}{\text{ml}}$. Wir möchten $500 \mu\text{l}$ einer Proteinlösung mit einer Konzentration von $0,3 \frac{\text{mg}}{\text{ml}}$ erstellen. Wie viele μl der vorliegenden Proteinlösung werden benötigt?

Aufstellen einer (linearen) Gleichung:

Sei x = benötigtes Volumen der gegebenen Proteinlösung (in μl).

$$x \mu\text{l} = \frac{x}{1000} \text{ ml}, \quad 500 \mu\text{l} = \frac{500}{1000} \text{ ml} = 0,5 \text{ ml}$$

Die Masse an gelöstem Protein ändert sich beim Verdünnen nicht, also (vorher=nachher)

$$\frac{x}{1000} \cdot 20 \text{ (mg)} = 0,5 \cdot 0,3 \text{ (mg)}$$

$$\iff \frac{x}{50} = 0,15 \quad | \cdot 50$$

$$\iff x = 0,15 \cdot 50 = 7,5,$$

d.h. es werden (wie wir wissen) $7,5 \mu\text{l}$ der gegebenen Lösung benötigt.

Beispiel: Stromrechnung

Ein Verbraucher bezieht seinen Strom zu folgenden Konditionen:
75 € Grundpreis pro Jahr und ein Arbeitspreis von 23 ct/kWh.

- a) Wie hoch ist die Stromrechnung, wenn im Jahr 1550 kWh verbraucht wurden?
- b) Ein anderer Haushalt zahlte bei diesem Tarif im letzten Jahr 827,10 €. Wieviel kWh hat dieser Haushalt verbraucht?
-

Gleichung aufstellen:

Sei p =Stromkosten (in €, für ein Jahr), x = Jahresverbrauch (in kWh)

$$p = 75 + 0,23x$$

a) $x = 1550$, also $p = 75 + 0,23 \cdot 1550 = 431,50$

Die Stromrechnung kostet 431,50 €.

b) $p = 827,10 = 75 + 0,23x \quad | - 75$

$$827,10 - 75 = 0,23x \quad | : 0,23$$

$$\frac{827,10 - 75}{0,23} = 3270 = x$$

Es wurden 3270 kWh verbraucht.

Folgen (abstrakt)

Eine (im Prinzip) unendliche Liste von Zahlen x_1, x_2, x_3, \dots

Beispiele

$$x_n = n : 1, 2, 3, \dots$$

$$x_n = (-1)^n : -1, 1, -1, 1, \dots$$

$$x_n = 1/n : 1, 1/2, 1/3, 1/4, \dots$$

$$x_n = 29 : 29, 29, 29, \dots$$

Rekursive Definitionen

$x_n =$ ein Ausdruck, der x_{n-1} enthält, z.B.

$$x_n = x_{n-1} + 5$$

(also $x_n = x_1 + 5(n-1)$, $n = 2, 3, 4, \dots$) „arithmetische Folge“

$$x_n = ax_{n-1}$$

(also $x_n = x_1 a^{n-1}$, $n = 2, 3, 4, \dots$) „geometrische Folge“

Bem.: Eine „geschlossene“ Form von x_n zu finden kann i.A. schwierig (oder nahezu unmöglich) sein.

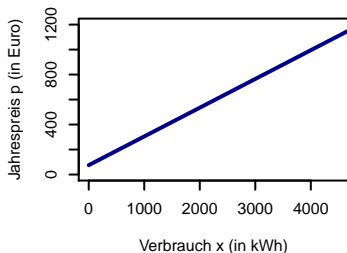
Funktionen (abstrakt)

Sehr allgemein: Eine Zuordnung, die jedem Element des *Definitionsbereichs* genau ein Element des *Wertebereichs* zuweist.

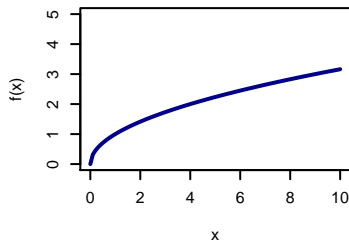
Beispiel: Teilnehmer der Vorlesung \longrightarrow Matrikelnummer
oder Teilnehmer der Vorlesung \longrightarrow Alter in Jahren

Oft werden Funktionen durch algebraische Ausdrücke definiert, z.B.

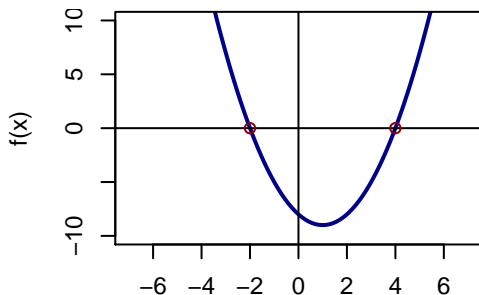
$$p(x) = 75 + 0,23x$$



$$f(x) = \sqrt{x}$$



Funktionen (abstrakt, quadratisch)



In Anwendungen trifft man (gelegentlich) *quadratische* Funktionen, z.B.

$$\begin{aligned} f(x) &= x^2 - 2x - 8 = x^2 - 2 \cdot 1 \cdot x + 1 - 1 - 8 \\ &= (x - 1)^2 - 9 = (x + 2)(x - 4) \end{aligned}$$

Erinnerung („*p-q*-Formel“): Die Gleichung $x^2 + px + q = 0$ hat die (reellwertigen) Lösungen $-\frac{p}{2} - \sqrt{\frac{p^2}{4} - q}$ und $-\frac{p}{2} + \sqrt{\frac{p^2}{4} - q}$ (sofern $\frac{p^2}{4} - q \geq 0$).

Hardy-Weinberg-Gleichgewicht (1908)



Godfrey Harold Hardy
(1877–1947)



Wilhelm Weinberg
(1862–1937)

Idealisierte Population („Mathematisches Modell“): sehr groß, diploid, hermaphroditisch. An einem Genort gebe es zwei verschiedene Allele, A und a .

Annahme: „Neutralität“, d.h. Reproduktionserfolg unabhängig vom Genotyp.

Genotypenhäufigkeiten heute:

Genotyp	AA	Aa	aa	$(x_{AA} + x_{Aa} + x_{aa} = 1)$
Anteil	x_{AA}	x_{Aa}	x_{aa}	

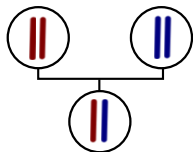
Das bedeutet für die Allelhäufigkeiten

Allel	A	a
Anteil	$p_A = x_{AA} + \frac{1}{2}x_{Aa}$	$p_a = \frac{1}{2}x_{Aa} + x_{aa}$

(offenbar auch $p_A + p_a = 1$)

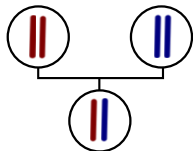
Wir nehmen weiter an:

„rein zufällige Paarungen“,
Mendelsche Segregation



Genotyphäufigkeiten in der nächsten Generation?

Heute:	Genotyp	AA	Aa	aa
	Anteil	x_{AA}	x_{Aa}	x_{aa}
	Allel	A		a
	Anteil	$p_A = x_{AA} + \frac{1}{2}x_{Aa}$		$p_a = \frac{1}{2}x_{Aa} + x_{aa}$



Nächste Generation:

Genotyp	AA	Aa	aa	Allel	A	a
Anteil	x'_{AA}	x'_{Aa}	x'_{aa}	Anteil	p'_A	p'_a

$$x'_{AA} = x_{AA}^2 + 2x_{AA}x_{Aa} \cdot \frac{1}{2} + x_{Aa}^2 \cdot \frac{1}{2} \cdot \frac{1}{2} = (x_{AA} + \frac{1}{2}x_{Aa})^2 = p_A^2$$

Analog:

$$x'_{Aa} = 2(x_{AA} + \frac{1}{2}x_{Aa})(x_{aa} + \frac{1}{2}x_{Aa}) = 2p_A p_a$$

$$x'_{aa} = x_{aa}^2 + 2x_{Aa}x_{aa} \cdot \frac{1}{2} + x_{Aa}^2 \cdot \frac{1}{2} \cdot \frac{1}{2} = (x_{aa} + \frac{1}{2}x_{Aa})^2 = p_a^2$$

Wir sehen:

Anteile heute: x_{AA} , x_{Aa} , x_{aa} (Genotypen),

$$p_A = x_{AA} + \frac{1}{2}x_{Aa}, p_a = \frac{1}{2}x_{Aa} + x_{aa} \text{ (Allele)}$$

Anteile nächste Generation:

$$x'_{AA} = p_A^2, x'_{Aa} = 2p_A p_a, x'_{aa} = p_a^2 \text{ (Genotypen),}$$

$$p'_A = x'_{AA} + \frac{1}{2}x'_{Aa} = p_A^2 + p_A p_a = p_A(p_A + p_a) = p_A,$$

$$p'_a = \frac{1}{2}x'_{Aa} + x'_{aa} = p_A p_a + p_a^2 = p_a(p_A + p_a) = p_a \text{ (Allele)}$$

Also:

- Allelhäufigkeiten sind konstant über die Generationen.
- Unabhängig von den ursprünglichen Genotyphäufigkeiten stellt sich für die Genotypen AA , Aa , aa nach einer Generation das Verhältnis

$$p^2 : 2p(1 - p) : (1 - p)^2$$

ein und ändert sich dann nicht mehr:

Hardy-Weinberg-Gleichgewicht

Eine Anwendung von Wurzeln und quadratischen Gleichungen:

Im HW-Gleichgewicht ist wegen $x_{AA} = p_A^2$ offensichtlich
 $p_A = +\sqrt{x_{AA}}$, analog $p_a = +\sqrt{x_{aa}}$.

Was kann man über die Allelhäufigkeit p_A sagen, wenn man nur x_{Aa} , die Häufigkeit der Heterozygoten, kennt?

$x_{Aa} = 2p_A(1 - p_A) = 2p_A - 2p_A^2$, also

$$p_A^2 - p_A + \frac{1}{2}x_{Aa} = 0.$$

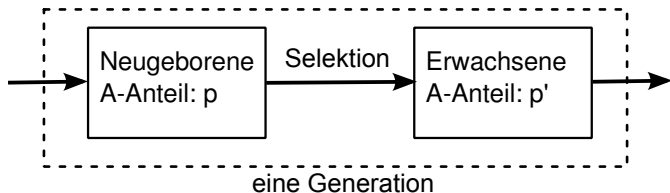
$$p_A = \frac{1}{2} + \sqrt{\frac{(-1)^2}{4} - \frac{1}{2}x_{Aa}} = \frac{1}{2} + \frac{1}{2}\sqrt{1 - 2x_{Aa}} \text{ oder}$$

$$p_A = \frac{1}{2} - \frac{1}{2}\sqrt{1 - 2x_{Aa}} \quad (\text{gemäß „p-q-Formel“})$$

(Zusatzinformation erforderlich, um die Lösung auswählen, z.B. welches der beiden Allele häufiger ist.)

[Auch anschaulich klar: Heterozygote ändern sich nicht bei A-a-Vertauschung.]

Ein Modell für Selektion



Genotyp	AA	Aa	aa
(rel.) Fitness	w_{AA}	w_{Aa}	w_{aa}

Interpretation: Die Chance eines Nachkommen vom Typ AA , bis zum Reproduktionsalter zu überleben, ist proportional zu w_{AA} , etc.

Beispiel 1:

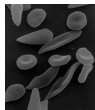
	AA	Aa	aa
	1	0,95	0,9



(nachteilige Flügel­färbung bei *Callimorpha dominula*)

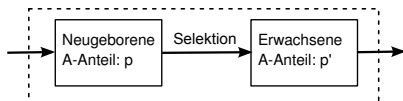
Beispiel 2:

	AA	Aa	aa
	1	1,17	0



(Sichelzellanämie und endemische Malaria)

Ein Modell für Selektion: die nächste Generation



Genotyp	AA	Aa	aa
(rel.) Fitness	w_{AA}	w_{Aa}	w_{aa}
Anteil heute	p^2	$2pq$	q^2
rel. Anteil nächste Gen.	$p^2 w_{AA}$	$2pq w_{Aa}$	$q^2 w_{aa}$
Anteil nächste Gen.	$\frac{p^2 w_{AA}}{w_{ges}}$	$\frac{2pq w_{Aa}}{w_{ges}}$	$\frac{q^2 w_{aa}}{w_{ges}}$

wobei

p ($= p_A$) Anteil des A -Allels,

$q = 1 - p$ ($= p_a$) Anteil des a -Allels,

$$w_{ges} = p^2 w_{AA} + 2pq w_{Aa} + q^2 w_{aa}$$

die „Gesamtfitness“ der (heutigen) Population

Umparametrisierung

Wir schreiben statt w_{AA} , w_{Aa} , w_{aa} lieber

Genotyp	AA	Aa	aa
(rel.) Fitness	1	$1 - hs$	$1 - s$

wobei

s : Selektionskoeffizient, $s \leq 1$,

h : (Koeffizient des) „Heterozygoteneffekt(s)“.

Beispiel 1:

	AA	Aa	aa
	1	0,95	0,9

$s = 0,1$, $h = 0,5$

Beispiel 2:

	AA	Aa	aa
	1	1,17	0

$s = 1$, $h = -0,17$

Verschiedene Szenarien von Selektion

Genotyp	AA	Aa	aa
(rel.) Fitness	1	$1 - hs$	$1 - s$

$s = 0$: neutraler Fall

$s \neq 0$ (wir betrachten den Fall $0 < s \leq 1$, sonst vertausche die Rollen von A und a)

$h = 1$: A rezessiv

$h = 0$: A dominant

$0 < h < 1$: unvollständige Dominanz

$h < 0$: Überdominanz (Aa „am fittesten“)

$h > 1$: Unterdominanz

(Bem.: $h = 1/2$: „additiver Fitnessseffekt“, ist mathematisch besonders angenehm und häufig für $s \approx 0$ gerechtfertigt)

Fragen

Verändern sich die Allelhäufigkeiten im Laufe der Zeit?

Wenn ja, wie?

Wie sieht es nach sehr langer Zeit aus?

Wird sich das A -Allel durchsetzen?

Mathematisch formuliert:

Sei p_n der Anteil A -Allele in der Population in Generation n ,
 $n = 0, 1, 2, \dots$

Wie verhält sich p_n , wenn n nach ∞ strebt?

Die Antwort hängt (hauptsächlich) von h ab ...

Änderung des A -Anteils als Funktion des aktuellen Anteils

Anteil heute: p relative Fitness:

AA	Aa	aa
1	$1 - hs$	$1 - s$

Anteil in der nächsten Generation:

$$N(p) := \frac{p^2 \cdot 1 + \frac{1}{2} \cdot 2pq \cdot (1 - hs)}{p^2 + 2pq(1 - hs) + q^2(1 - s)}$$

(mit $q = 1 - p$)

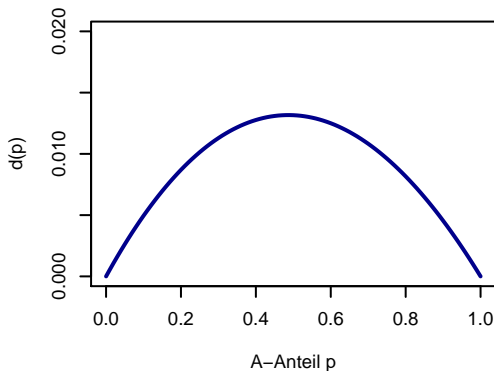
Änderung über eine Generation

$$\begin{aligned} d = N(p) - p &= \frac{p^2 \cdot 1 + \frac{1}{2} \cdot 2pq \cdot (1 - hs)}{p^2 + 2pq(1 - hs) + q^2(1 - s)} - p \\ &= \frac{spq(ph + q(1 - h))}{p^2 + 2pq(1 - hs) + q^2(1 - s)} \end{aligned}$$

$0 \leq h \leq 1$: gerichtete Selektion (Diagramm für $s=0,1$, $h=0,5$)

Änderung des A-Anteils als Funktion von p :

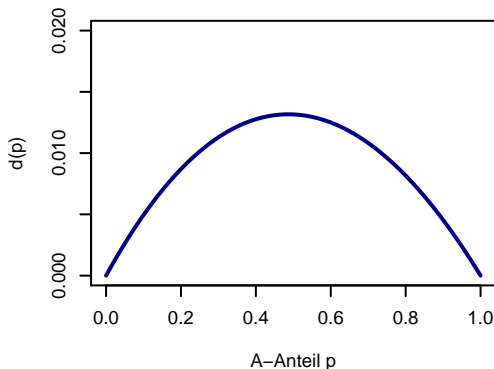
$$d(p) = \frac{sp(1-p)(ph + (1-p)(1-h))}{p^2 + 2p(1-p)(1-hs) + (1-p)^2(1-s)}$$



Wir sehen: $d(p) > 0$ außer für $p \in \{0, 1\}$, d.h. der A-Anteil nimmt stets zu.

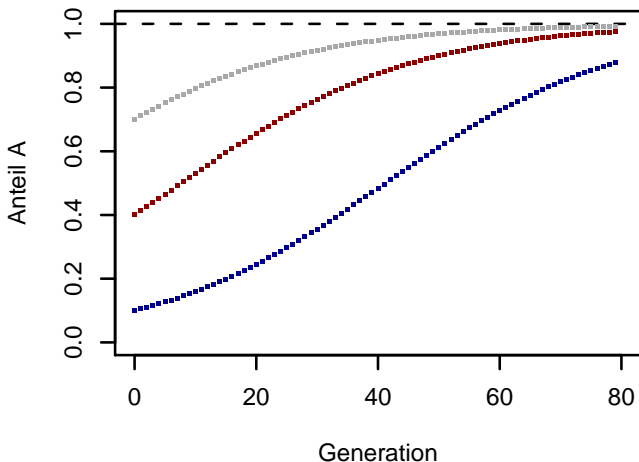
$0 \leq h \leq 1$: gerichtete Selektion (Diagramm für $s=0,1$, $h=0,5$)

Änderung des A -Anteils $d(p)$ als Funktion von p :



$d(p) > 0$, d.h. der A -Anteil nimmt stets zu (außer für $p \in \{0, 1\}$).
Mathematisch ausgedrückt:

$$p_n = N(p_{n-1}) = p_{n-1} + d(p_{n-1}) > p_{n-1}$$

$0 \leq h \leq 1$: gerichtete Selektion, Konvergenz des A-Anteils

Anteil A in Abhängigkeit der Generation für verschiedene Startwerte ($s=0,1$, $h=0,5$)

$0 \leq h \leq 1$: gerichtete Selektion, Konvergenz des A-Anteils

$p_n = p_{n-1} + d(p_{n-1})$, also für $0 < p_1 < 1$:

$p_1 < p_2 < p_3 < \dots \leq 1$, andererseits kann p_n für $\varepsilon > 0$ nicht für alle n unterhalb $1 - \varepsilon$ liegen (denn die Funktion $d(\cdot)$ ist strikt positiv im Intervall $[p_1, 1 - \varepsilon]$), somit haben wir bewiesen

Satz

Im Fall $0 \leq h \leq 1$ gilt für jedes $p_1 \in (0, 1]$: $p_n \rightarrow p^* = 1$ für $n \rightarrow \infty$

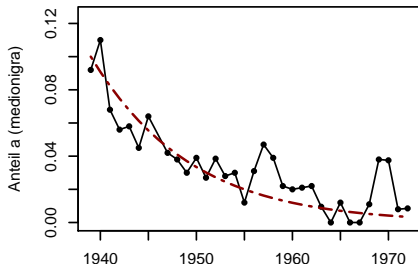
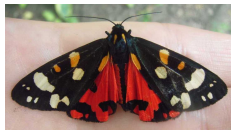
(Man sagt: „Die Folge p_n konvergiert gegen p^* “ und schreibt dies auch als $\lim_{n \rightarrow \infty} p_n = p^*$)

Biologische Interpretation

Im Fall $0 \leq h \leq 1$ setzt sich stets Allel A durch (sofern es zu Beginn vorhanden ist). Die Population wird auf lange Sicht nur aus AA-Individuen bestehen.

Beispiel: Das *medionigra*-Allel (*a*) in einer Population von *Callimorpha dominula* (dt. Schönbär) in Oxford

Beobachteter Anteil *a* 1939–1972 und Modellvorhersage



Nach Kap. 3 in John H. Gillespie, *Population genetics : a concise guide*, Johns Hopkins Univ. Press, 1998, siehe auch Ford, E.B. and P.M. Sheppard, The *medionigra* polymorphism of *Panaxia dominula*. *Heredity* 24:112–134, 1969.

Bemerkung. Die Modellkurve passt recht gut, das beweist allerdings nicht, dass tatsächlich gerichtete Selektion für die beobachteten Änderungen des *a*-Anteils verantwortlich ist — es könnte andere Effekte geben, die z.T. kontrovers in der Literatur diskutiert werden.

$h < 0$: balancierende Selektion

Erinnerung: Rel. Fitness $1 : 1 - hs : 1 - s$, also bedeutet $h < 0$, dass Heterozygote der „fitteste“ Typ sind.

(Erneut:) Änderung als Funktion von p :

$$d(p) = \frac{sp(1-p)(ph + (1-p)(1-h))}{p^2 + 2p(1-p)(1-hs) + (1-p)^2(1-s)}$$

Für welche p ist $d(p) = 0$?

$p = 0$, $p = 1$, oder p Lösung von

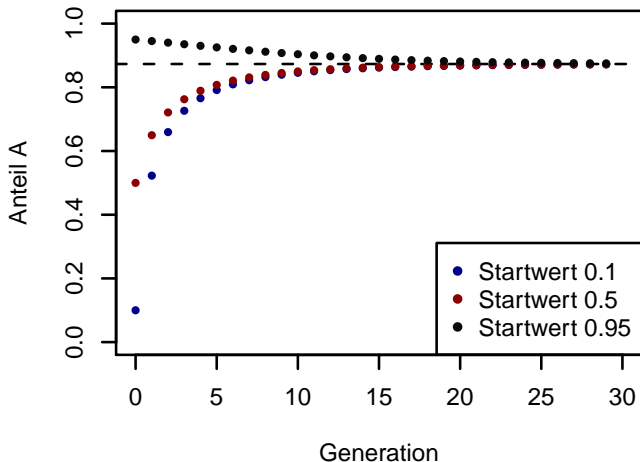
$$ph + (1-p)(1-h) = 0,$$

d.h. $p = \hat{p} = \frac{1-h}{1-2h}$.

Im Beispiel: $h=-0,17$, $\hat{p}=0,87$ (also $1-\hat{p}=0,13$)

$h < 0$: balancierende Selektion

Zeitliche Entwicklung des A-Anteils bei verschiedenen Startwerten ($s=1$, $h=-0,17$)



$h < 0$: balancierende Selektion, Konvergenz

$$p_n = p_{n-1} + d(p_{n-1}), \quad \hat{p} = \frac{1-h}{1-2h}$$

Man kann (leicht) zeigen, dass $p_1 < p_2 < p_3 < \dots \leq \hat{p}$ wenn

$0 < p_1 < \hat{p}$ und

$p_1 > p_2 > p_3 > \dots \geq \hat{p}$ wenn $\hat{p} < p_1 < 1$,

somit

Satz

Im Fall $h < 0$ gilt für jedes $p_1 \in (0, 1)$:

$$\lim_{n \rightarrow \infty} p_n = \hat{p} = \frac{1-h}{1-2h}.$$

(Die Folge p_n konvergiert gegen \hat{p} .)

Biologische Interpretation

Im Fall $h < 0$ bleiben beide Allele in der Population erhalten, das genaue Verhältnis hängt von h ab (das hier die Stärke der Überdominanz misst).



Übrigens: Der Schönbär war Schmetterling des Jahres 2010

www.bund-nrw-naturschutzstiftung.de/schmetterling2010.htm