Einführung in die Mathematische Biologie

Ein Vorlesungsskript von

Ellen Baake und Joachim Hermisson

LMU München
Wintersemester 2004/2005

Inhaltsverzeichnis

Vorbemerkungen

Biologie und Mathematik

Es ist wohl keine Übertreibung zu sagen, dass nicht alle großen Biologen talentierte Mathematiker waren. Charles Darwin schreibt zu diesem Thema in seiner Autobiographie von 1876:

> *During the three years which I spent at Cambridge my time was wasted, as far as the academical studies were concerned, as completely as during the three years which I spent at Edinburgh and at school. I attended mathematics, and even went during the summer of 1828 with a private tutor (a very dull man) to Barmouth, but I got on very slowly. The work was repugnant to me, chiefly from my unbeing able to see any meaning in the early steps in algebra. This impatience was very foolish, and in the after-years I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics for men thus endowed seem to have an extra sense. But I do not believe that I should ever have succeeded beyond a very low grade.*

Dennoch verdankt Darwin den entscheidenden Durchbruch zu seiner Theorie der natürlichen Selektion einer Kombination von Naturbeobachtung und mathematischer Modellierung, wie er selbst ausdrücklich betont:

> *In October 1838, that is, fifteen months after I had begun my systematic enquiry, I happened to read for amusements 'Malthus on Population', and being well prepared to appreciate the struggle for existence which everywhere goes on from long-continued observation of the habits of animals and plants, it at once struck me that under these circumstances favourable variations would tend to be preserved, and unfavourable ones to be destroyed. The result of this would be the formation of new species. Here then I had at last got a theory by which to work.*

Die mathematischen Einsichten von Thomas Malthus (1766-1834), dessen *Essay on the Principle of Population* (1798) Darwin mehr zufällig in die Hände bekam, werden wir im ersten Teil der Vorlesung ausführlich besprechen.
Natürlich hat sich seit den Zeiten Darwins in der Biologie viel getan. Generell wird mathematische Modellierung überall dort wichtig, wo die Wissenschaft über das Sammeln von Daten und Fakten (klassisch: z.B. neue Spezies beschreiben, heute: z.B. Funktion eines Gens beschreiben) hinausgeht und das Gewinnen von neuen Einsichten aus diesen Fakten das Ziel ist. Nach dem Physiker Sir William Bragg ist letzteres die eigentliche Bestimmung jeder Wissenschaft:

> *The important thing in science is not so much to obtain new facts, but to discover new ways to think about them.*

Da die Biologie so komplex und vielfältig ist wie keine andere Naturwissenschaft, steht häufig noch das Sammeln von Daten und Fakten an erster Stelle. Zunehmend können heute aber immer größere Mengen an Daten in kurzer Zeit gewonnen werden (Sequenzdaten, Genexpressionen), sodass nicht das Sammeln selbst, sondern die Gewinnung neuer Erkenntnisse aus diesen Daten zum vorrangigen Ziel wird. Dies erfordert oft komplexe Modelle – eine zunehmende Mathematisierung der Wissenschaft ist die Folge. Natürlich ist diese Entwicklung in den verschiedenen Teilbereichen der Biologie unterschiedlich schnell. In der von Darwin begründeten Evolutionsbiologie ist sie schon so weit vorangeschritten, dass John Maynard Smith in der Einleitung zu seinem Standardlehrbuch *Evolutionary Genetics* trocken bemerkt:

> *If you can't stand algebra, keep out of evolutionary biology.*

Es wird in dieser Vorlesung (zweistündig und ganz am Anfang des Studiums) nicht möglich sein, einen umfassenden Überblick über Modelle und Methoden der mathematischen Biologie zu geben. Das Ziel ist es vielmehr, eine Idee davon zu vermitteln, was es heißt "mathematisch zu modellieren" und mit welchen Methoden man in Modellen zu biologischem Verständnis kommen kann. Dies wird im ersten Teil der Vorlesung beispielhaft anhand einiger Modelle für deterministische Prozesse geschehen. Genauso geht es dann im zweiten Teil um Grundideen der Statistik und der Testtheorie.

Literatur

Zum ersten Teil:

- *+F. R. Adler, Modeling the Dynamics of Life, Brooks/Cole 1997.
- N. F. Britton, Essential Mathematical Biology, Springer 2003.
- *+A. Riede, Mathematik für Biologen, Vieweg 1993.
- +W. Timischl, Biomathematik. Eine Einführung für Biologen und Mediziner, Springer 1995.
- +H. Vogt, Grundkurs Mathematik für Biologen, 2. Auflage, Teubner 1994.

Zum zweiten Teil:

- *+F. R. Adler, Modeling the Dynamics of Life, Brooks/Cole 1997.
- *+G. Gelbrich, Statistik für Anwender, Shaker, 1998
- *+A. Riede, Mathematik für Biologen, Vieweg 1993.
- +R. Sokal, F. J. Rohlf, Biometry, 3. Auflage, Freeman, 1995.
- +H. Vogt, Grundkurs Mathematik für Biologen, 2. Auflage, Teubner 1994.

Die mit * bezeichneten Bücher sind in der Bibliothek des Department Biologie II vorhanden. Die mit + bezeichneten Bücher gibt es in der Lehrbuchsammlung der UB.

Grundlagen

Bevor es mit dem eigentlichen Stoff der Vorlesung losgeht, zunächst ein paar Erläuterungen zu Notationen und ein bisschen Wiederholung aus der Schulmathematik.

Summen und Produkte: Das Summenzeichen ist wie folgt definiert (":=" bedeutet "ist definiert durch"):

$$\sum_{i=k}^{N} x_i := x_k + x_{k+1} + x_{k+2} + \cdots + x_N$$

Die Summationsgrenzen k und N sind beides natürliche Zahlen (mit $N \geq k$). Die Laufvariable i, über die summiert wird, ist hier ein Index einer anderen Variablen x, dies muss aber nicht sein, z.B.:

$$\sum_{\ell=2}^{5} \frac{1}{\ell^2} = \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25}.$$

Rechenregeln: Konstanten, die nicht von der Laufvariablen abhängen, darf man 'ausklammern', also vor die Summe ziehen:

$$\sum_{i=1}^{n} cx_i = cx_1 + cx_2 + \cdots + cx_n = c(x_1 + x_2 + \cdots + x_n) = c\sum_{i=1}^{n} x_i.$$

Außerdem gilt

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$$

Analog zum Summenzeichen ist das Produktzeichen definiert:

$$\prod_{i=k}^{N} y_i := y_k \cdot y_{k+1} \cdot y_{k+2} \cdot \cdots \cdot y_N.$$

Funktionen: Eine wichtige Klasse von Funktionen sind die Polynome. Ein Polynom n-ten Grades ist definiert als:

$$f(x) = \sum_{i=0}^{n} a_i x^i = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

Die a_i sind beliebige Zahlen und heißen Koeffizienten. Polynome ersten Grades entsprechen den Geraden,

$$f(x) = ax + b,$$

mit Steigung a und Schnittpunkt mit der y-Achse bei b. Eine spezielle Gerade ist die Winkelhalbierende $f(x) = x$, die häufig auch mit $f = \mathrm{Id}$ (für *identity*) bezeichnet wird. Polynome zweiten Grades entsprechen den Parabeln,

$$f(x) = ax^2 + bx + c.$$

Die Nullstellen der Parabel lassen sich mit Hilfe der "Mitternachtsformel" berechnen: $f(x_\pm) = 0$ für

$$x_\pm = \frac{1}{2a}\left(-b \pm \sqrt{b^2 - 4ac}\right).$$

Eine weitere wichtige Funktionsklasse sind die Potenzen wie $f(x) = ab^x$. Die Basis b ist irgendeine Zahl. Besonders wichtig ist die Potenz zur Basis $e \approx 2.71828\ldots$, die sogenannte Exponentialfunktion

$$f(x) = e^x = \exp(x).$$

Es sollte bekannt sein, dass

$$\begin{aligned} e^0 &= 1, \\ e^{a+b} &= e^a \cdot e^b, \\ (e^a)^n &= e^{na}. \end{aligned}$$

Die Umkehrfunktion der Exponentialfunktion ist der natürliche Logarithmus $f(x) = \ln x$ (für $x > 0$). Es gilt also:

$$\ln\big(\exp(x)\big) = \exp\big(\ln(x)\big) = x.$$

Analog zu den obigen Rechenregeln für die Exponentialfunktion gilt für den Logarithmus:

$$\begin{aligned} \ln(1) &= 0, \\ \ln(ab) &= \ln a + \ln b, \\ \ln(a^n) &= n \ln a. \end{aligned}$$

Differenzieren: Die Steigung der Tangente an eine Funktion $f(x)$ an der Stelle x ergibt sich aus der Ableitung:

$$\frac{\mathrm{d}}{\mathrm{d}x} f(x) := f'(x) := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Für ein Maximum oder ein Minimum x^* einer differenzierbaren Funktion muss immer gelten, dass die Ableitung an dieser Stelle $f'(x^*) = 0$ ist. Die Ableitungen von einigen wichtigen Funktionen sind:

$$\frac{\mathrm{d}}{\mathrm{d}x} x^n = n x^{n-1}, \qquad \frac{\mathrm{d}}{\mathrm{d}x} e^x = e^x, \qquad \frac{\mathrm{d}}{\mathrm{d}x} \ln x = \frac{1}{x}$$

Rechenregeln für die Ableitung sind die Summenregel,

$$\frac{\mathrm{d}}{\mathrm{d}x}[f(x) + g(x)] = f'(x) + g'(x),$$

die Produktregel,

$$\frac{\mathrm{d}}{\mathrm{d}x}[f(x)g(x)] = f'(x)g(x) + f(x)g'(x),$$

die Quotientenregel,

$$\frac{\mathrm{d}}{\mathrm{d}x} \frac{f(x)}{g(x)} = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2},$$

und die Kettenregel ("äußere mal innere Ableitung"),

$$\frac{\mathrm{d}}{\mathrm{d}x} f\big(g(x)\big) = \frac{\mathrm{d}}{\mathrm{d}g} f\big(g(x)\big) \frac{\mathrm{d}}{\mathrm{d}x} g(x).$$

Integrieren: Die Umkehrung der Differenziation ist die Integration,

$$\int_a^b f(x)\mathrm{d}x = F(x)\Big|_a^b = F(b) - F(a),$$

wobei für die Stammfunktion $F(x)$ gilt: $F'(x) = f(x)$. Die geometrische Bedeutung des Integrals mit den Grenzen a und b ist die Fäche unterhalb des Graphs der Funktion und oberhalb der x-Achse zwischen a und b. Einige wichtige Stammfunktionen sind:

$$\begin{aligned} f(x) = x^n \;&\rightarrow\; F(x) = \frac{x^{n+1}}{n+1} + c, \\ f(x) = \frac{1}{x} \;&\rightarrow\; F(x) = \ln(x) + c, \\ f(x) = e^{ax} \;&\rightarrow\; F(x) = \frac{1}{a}e^{ax} + c, \end{aligned}$$

c ist jeweils eine beliebige Konstante. Das Integral ist das kontinuierliche Gegenstück zur Summe, es gelten deshalb analoge Rechenregeln:

$$\int_a^b cf(x)\mathrm{d}x = c\int_a^b f(x)\mathrm{d}x, \qquad \int_a^b \big(f(x) + g(x)\big)\mathrm{d}x = \int_a^b f(x)\mathrm{d}x + \int_a^b g(x)\mathrm{d}x.$$

Teil 1

Modellierung biologischer Prozesse

Was heißt Modellieren? Ein Modell in der Biologie ist eine Abbildung eines biologischen Systems (eines Teils der Natur) auf einen mathematischen Formalismus. Modelle sind immer Abstraktionen, also vereinfachte Darstellungen der Wirklichkeit. Modellbildung besteht deshalb aus einer Reihe von Entscheidungen oder *Modellannahmen*: Bei einigen Faktoren des biologischen Systems geht man davon aus, dass sie für die vorliegende Fragestellung wesentlich sind, und integriert sie in das Modell, von anderen Aspekten nimmt man an, dass sie keine große Rolle spielen, und man ignoriert sie im Modell (*abstrahiert* von ihnen). Wenn man zum Beispiel die Bevölkerungsentwicklung in Deutschland in den nächsten Jahren modellieren will, ist sicher die gegenwärtige Größe der Bevölkerung und auch ihre Altersstruktur ein wichtiger Faktor, den man in einem Modell berücksichtigen sollte. Auf der anderen Seite kann man von den klimatischen Bedingungen in einem einfachen Modell wahrscheinlich abstrahieren, obwohl genau genommen sicher auch Klimaschwankungen (Hitzewellen und Kälteperioden) eine gewisse Auswirkung auf die jährlichen Geburts- und Todesraten haben.
Allgemein gilt: Je mehr man abstrahiert, oder je stärker (restriktiver) die Modellannahmen sind, desto einfacher, aber auch ungenauer wird das resultierende Modell. Ein zu einfaches Modell hat mit dem System, das es beschreiben soll, nicht mehr viel zu tun. Umgekehrt gilt: Mit zunehmender Komplexität wird das Modell zwar (in der Regel) genauer, dafür aber immer schlechter handhabbar. Aus einem Modell, das genauso komplex ist wie das System, das es beschreiben soll, kann man keine neuen Einsichten gewinnen.
Die Kunst der Modellierung ist es also, die wesentlichen Faktoren in einem biologischen System herauszufinden. Man braucht dafür zweierlei: Einerseits eine gute Kenntnis des biologischen Systems, um eine Vorahnung zu haben, was wichtig sein könnte. Andererseits braucht man auch ein gutes Verständnis des mathematischen Formalismus, um zu wissen, an welchen Stellen im Modell auch Faktoren, die zunächst ganz klein aussehen, plötzlich einen großen Effekt haben (und umgekehrt). Ob man wirklich alle wesentlichen Faktoren berücksichtigt hat, weiß man natürlich immer erst hinterher, wenn sich (idealerweise) Modellvorhersagen durch Messungen bestätigen lassen. Wenn dies gelingt, hat man etwas über die Natur verstanden.

Was ist ein biologischer Prozess? Eine häufige Problemstellung in der Biologie ist die Frage, wie sich eine uns interessierende Größe in einem biologischen System in der Zeit entwickelt. Solche biologischen Prozesse gibt es in ganz verschiedenen Zusammenhängen:

- In der *Ökologie* interessiert man sich zum Beispiel für die Frage, wie sich die Populationsgrößen in einem Ökosystem entwickeln.
- Die *Evolutionsbiologie* beschäftigt sich mit der Entwicklung von Allelfrequenzen in einer Population als Folge von Mutation, Selektion und genetischer Drift.

- Das Forschungsfeld der *Entwicklungsbiologie* definiert sich geradezu über einen biologischen Prozess: Die schrittweise Ausprägung des Phänotyps in der Ontogenese.
- In der *Physiologie* schließlich geht es um Fragen wie die Veränderungen von Enzymkonzentrationen im metabolischen Netzwerk einer Zelle oder die Regulierung des Glukosespiegels im Blut.

Typischerweise möchte man dann zunächst die Gleichgewichtspunkte beschreiben, die sich in einem solchen Prozess einstellen (ökologisches Gleichgewicht, metabolischer Gleichgewichtsfluss, Mutations-Selektions Balance, ...), und fragt zum Beispiel, ob diese Gleichgewichte stabil oder instabil gegenüber einer äußeren Störung sind.
Um die folgenden Fragen wird es im ersten Teil der Vorlesung gehen:

- Wie modelliere ich einen biologischen Prozess?
- Mit was für Methoden kann ich diesen Prozess im Modell charakterisieren?
- Wie kann ich daraus Aussagen über das Verhalten biologischer Systeme gewinnen?

Wir werden uns dabei auf deterministische Prozesse beschränken (im Gegensatz zu stochastischen Prozessen). Das heißt, wir nehmen an, dass sich das Verhalten des Systems in der Zukunft eindeutig aus dem Zustand des Systems in der Gegenwart vorhersagen lässt (Zufall spielt keine Rolle). Als Beispiele werden wir vor allem Modelle aus der Populationsdynamik verwenden. Die Populationsdynamik ist ein Teilbereich der Ökologie und beschreibt die zeitliche Veränderung in der Größe von Populationen.

1.1 Deterministische Prozesse in diskreter Zeit

Viele Prozesse in der Natur werden als in der Zeit kontinuierlich ablaufend betrachtet. Für andere, wie beispielsweise das Wachstum einer Insektenpopulation mit getrennten Generationen oder für Kreuzungsexperimente, ist eine **diskrete Zeitskala** sinnvoller, auf der nur bestimmte Zeitpunkte, z.B. Vielfache der Generationsdauer Δt, angegeben werden. Eine diskrete Betrachtungsweise entspricht oft auch der experimentellen Situation, wo meist nicht kontinuierlich, sondern in regelmäßigen Abständen gemessen wird. Messen wir beispielsweise die Populationsgröße x zu den Zeitpunkten $0, \Delta t, 2\Delta t, 3\Delta t, \ldots$, so bilden die Werte $x_0, x_1, x_2 \ldots$ eine *Folge*, die die zeitliche Veränderung des Systems beschreibt.
Eine **Folge** ist eine Menge von numerierten Zahlen, $x_0, x_1, x_2, \ldots$. Die Nummern sind die natürlichen Zahlen (inkl. 0) und heißen **Indizes**. Die Zahl mit dem Index n heißt das **Folgenglied** x_n.
Eine Folge ist **explizit** gegeben, wenn man eine Formel für beliebige Folgenglieder hat, die ohne Kenntnis vorheriger Folgenglieder (außer dem ersten) auskommt; z.B. die **arithmetische Folge** $x_n = x_0 + cn$ (für $x_0 = 1$ und $c = 2$ ist das gerade die Folge der ungeraden Zahlen). Eine Folge ist **iterativ** (oder **rekursiv**) definiert, wenn eine Vorschrift f gegeben ist, wie man ein Folgenglied aus dem jeweils vorhergehenden gewinnt:

$$x_{n+1} = f(x_n). \tag{1.1}$$

Für die arithmetische Folge lautet diese Vorschrift einfach $x_{n+1} = x_n + c$. Die Funktion f heißt **Iterationsvorschrift** oder **Iterationsfunktion**. In unserem populationsbiologischen Zusammenhang wird die Iterationsfunktion oft auch als **Reproduktionsfunktion** bezeichnet; sie stellt eine **dynamische Regel** dar, die angibt, wie sich die Populationsgröße von einem Jahr

zum nächsten verändert. Die zugehörige *explizite Folge* heißt auch **Zeitverlauf**, **Zeitentwicklung** oder **Populationsentwicklung**.
Der entscheidende Schritt zur mathematische Modellierung eines biologischen Prozesses ist die Konstruktion der Iterationsfunktion aus biologisch motivierten Überlegungen über das System, das man beschreiben möchte. Im folgenden werden wir einige Beispiele hierfür kennenlernen. In einem zweiten Schritt möchte man dann wissen, wie die Reproduktionsfunktion den Zeitverlauf des biologischen Prozesses bestimmt.

1.1.1 Modellbildung I: Geometrisches Wachstum

Betrachten wir, um konkret zu werden, die Vogelpopulation auf einer Insel. Jedes Jahr wird die Populationsgröße bestimmt. Wie kann man die Entwicklung dieser Größe in einem Modell beschreiben? Für das einfachste Modell machen wir die folgenden Annahmen:
Als erstes nehmen wir an, dass die Populationsgröße im Folgejahr, x_{n+1}, nur von der aktuellen Populationsgröße x_n abhängt, und nicht von anderen in der Zeit veränderlichen Größen (wir vernachlässigen also z.B. Klimaschwankungen oder den Einfluss von Populationsgrößen anderer Vogelarten). Außerdem nehmen wir an, dass sich die Populationsgröße zwischen den Jahren nur durch Geburten und Todesfälle ändert. Wir setzen also an: $x_{n+1} = x_n + B_n - D_n$, wobei B_n und D_n die Anzahl der Geburten und Todesfälle in einer Population der Größe x_n angibt. In der einfachsten Näherung sollten diese Zahlen einfach proportional zur ursprünglichen Zahl der Individuen sein, also $B_n = bx_n$ und $D_n = dx_n$. b und d sind Konstanten und geben die *mittlere Zahl der Geburten und Todesfälle* in der Population pro Individuum und pro Jahr an. Mit der Definition des **Wachstumsfaktors** (*growth ratio*) $r := 1 + b - d$ finden wir damit die Reproduktionsfunktion

$$x_{n+1} = f(x_n) = x_n + bx_n - dx_n = rx_n. \tag{1.2}$$

Die Reproduktionsfunktion $f(x) = rx$ ist einfach eine Gerade durch den Ursprung (s. Abb. 1.1). Für jeden **Anfangswert** x_0 können wir durch wiederholtes Anwenden der Iterationsvorschrift (1.2) die Entwicklung der Populationsgröße in aufeinanderfolgenden Jahren voraussagen, d.h:

$$\begin{array}{lllllll} x_1 & = & rx_0 & & & & \\ x_2 & = & rx_1 & = & r^2x_0 & & \\ x_3 & = & rx_2 & = & r^2x_1 & = & r^3x_0 \\ & \vdots & & & & & \end{array}$$

Offensichtlich läßt sich die Folge *explizit* angeben als

$$x_n = r^n x_0. \tag{1.3}$$

Dies ist die sogenannte **geometrische Folge**. Die Populationsentwicklung hängt nur von x_0 und r ab. Diese beiden Werte legen die Entwicklung für alle Zeiten fest. Zwei aufeinanderfolgende Glieder einer geometrischen Folge haben ein konstantes Verhältnis r.

Beispiel: Wachstum einer Vogelpopulation

In einer Vogelpopulation, die eine Insel neu besiedelt hat, werden pro Jahr und Individuum im Mittel $b = 1.1$ neue Vögel geboren und $d = 0.2$ sterben. Das führt zu geometrischem Wachstum mit $r = 1 + 1.1 - 0.2 = 1.9$, also $x_n = (1.9)^n x_0$, mit $x_0 = 10$ in Zahlen:

Zeit n (in Jahren)	0	1	2	3	4	5	6	7	8
Populationsgröße x_n	10	19	36	69	130	248	470	894	1698

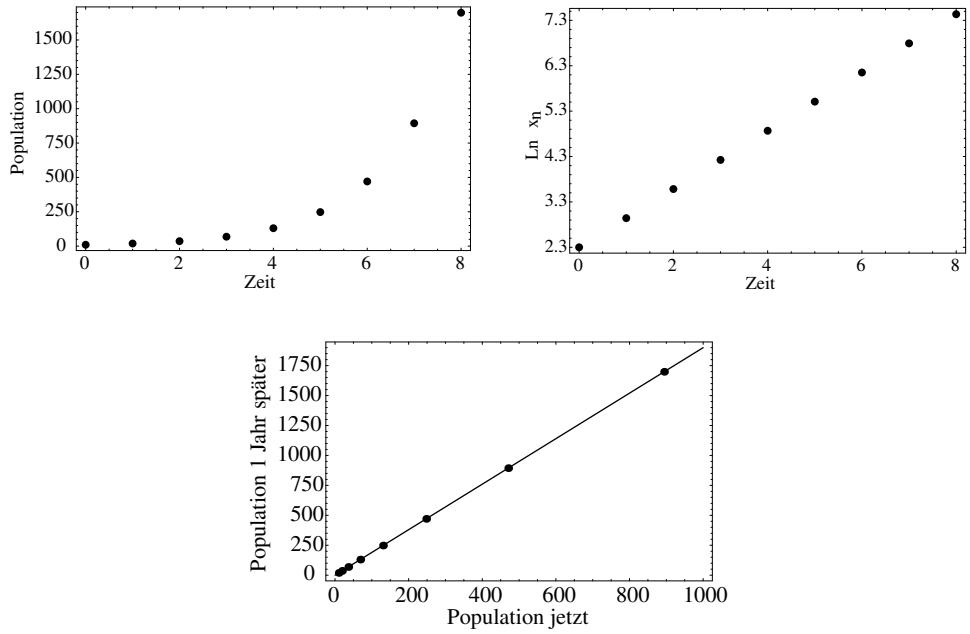


Abbildung 1.1: Geometrisches Wachstum einer Vogelpopulation. Links oben: Populationsgröße x_n als Funktion der Zeit n in Jahren; rechts oben: Logarithmus der Populationsgröße als Funktion der Zeit (sog. halblogarithmische Auftragung); unten: Populationsgröße in der nächsten Generation als Funktion der aktuellen Größe (die Punkte liegen auf dem Graphen der Reproduktionsfunktion).

Wir beobachten noch folgendes: Betrachtet man anstelle der Populationsgröße den (natürlichen) Logarithmus,

$$y_n = \ln(x_n) = \ln(r^n x_0)\,,$$

so liefern die Rechenregeln für Logarithmen

$$y_n = n\ln(r) + \ln(x_0).$$

Wenn x_n geometrisch wächst, wächst also der Logarithmus von x_n linear mit der Zeit (vgl. Abb. 1.1). Die Steigung der zugehörigen Geraden ist $\ln(r)$, der Achsenabschnitt ist $\ln(x_0) \approx 2.3$. An dieser Darstellung läßt sich das geometrische Wachstum besonders leicht erkennen. Man benötigt sie auch, wenn man z.B. voraussagen will, wie lange es dauert, bis sich die Individuenzahl verzehnfacht hat.

Verhalten in Abhängigkeit von r

Allgemein erhalten wir für eine Reproduktionsfunktion der Form $f(x) = rx$ ein qualitativ unterschiedliches Verhalten abhängig vom Wert des Wachstumsfaktors $r \geq 0$:

$r = 0$	sofortiges Aussterben
$0 < r < 1$	geometrisches "Schrumpfen"
$r = 1$	$x_n = x_0$ für alle n
$r > 1$	geometrisches Wachstum

Ein komplexeres Modell

Wir betrachten wieder das Beispiel der Vogelpopulation auf einer Insel. Im letzten Abschnitt hatten wir mit der Iterationsfunktion $f(x) = (1 + b - d)x = rx$ ein sehr einfaches Modell für die zeitliche Dynamik der Populationsgröße konstruiert. Dieses Modell entält zwei Parameter, die Geburtenrate b und die Todesrate d, die wir aus Beobachtungen im Feld schätzen können (die Dynamik hängt allerdings nur von der Differenz $b - d$ ab). Um das Modell des geometrischen Wachstums realistischer zu machen, wollen wir nun einige Aspekte, die wir bisher vernachlässigt hatten, ergänzen:

- Jedes Jahr wandern einige Vögel vom Festland auf die Insel zu. Wir modellieren dies durch einen Beitrag A zu $f(x)$, wobei A einfach eine positive Zahl unabhängig von der Populationsgröße x ist.
- Bisher haben wir eine konstante Geburtenrate b angenommen. Die Zahl der Neugeborenen ist dann einfach proportional zur Populationsgröße $B(x) = bx$. Für sehr kleine Populationen ist das nicht realistisch. Der Grund ist, dass zur Produktion von Nachkommen erst einmal ein Partner gefunden werden muss. Dies gelingt bei kleiner Populationsgröße x nicht immer. Die effektive Anzahl der Geburten ist deshalb zunächst kleiner und erreicht erst für ein hinreichend großes x (wenn die Partnersuche kein Problem mehr ist) die Relation $B(x) = bx$.
- Eine weitere unrealistische Eigenschaft des geometrischen Wachtums-Modells ist es, dass Populationen unendlich groß werden können. In Wirklichkeit ist die Populationsgröße durch den beschränkten Vorrat an Nahrung und Platz immer begrenzt. Wir können dies in unserem Modell durch eine maximale Populationsgröße K (die sogenannte *carrying capacity*) berücksichtigen. Die Iterationsfunktion $f(x)$ kann den Wert K nicht übersteigen, sondern nähert sich ihm für große x asymptotisch an.

Die Änderungen, die sich aus diesen Überlegungen für die Iterationsfunktion $f(x)$ ergeben sind in der Abbildung 1.2 dargestellt.

1.1.2 Modellanalyse I: Cobwebbing

Wir wollen nun eine graphische Methode kennenlernen, die es uns erlaubt, auch für kompliziertere Reproduktionsfunktionen Aussagen über die Entwicklung der Populationsgröße zu treffen. Diese Methode, das sogenannte **Cobwebbing**, besteht aus mehreren Schritten:

1. Zunächst zeichnen wir den Graphen der Reproduktionsfunktion f und den Graphen der Winkelhalbierenden $\mathrm{Id}(x) = x$ und wählen den Anfangswert x_0.
2. Um x_1 aus x_0 zu bestimmen, wenden wir f auf x_0 an, d.h. x_1 ist der Funktionswert von f an der Stelle x_0. Wir zeichnen einen Pfeil von Punkt (x_0, x_0) zum Punkt (x_0, x_1), der auf dem Graphen von f liegt.
3. Um x_2 aus x_1 zu bestimmen, müssen wir zunächst den Wert x_1 von der vertikalen auf die horizontale Achse übertragen. Dies erreichen wir dadurch, dass wir zunächst einen waagerechten Pfeil vom Punkt (x_0, x_1) auf dem Graphen von f zum Punkt (x_1, x_1) auf der Winkelhalbierenden zeichnen. Nun erhält man $x_2 = f(x_1)$ analog zu Schritt 2: durch einen senkrechten Pfeil von (x_1, x_1) zum Punkt (x_1, x_2) auf dem Graphen von f.

Wenn man Schritt 3 von verschiedenen Anfangspunkten x_0 aus wiederholt ausführt, erhält man das in Abb. 1.3 dargestellte Bild.

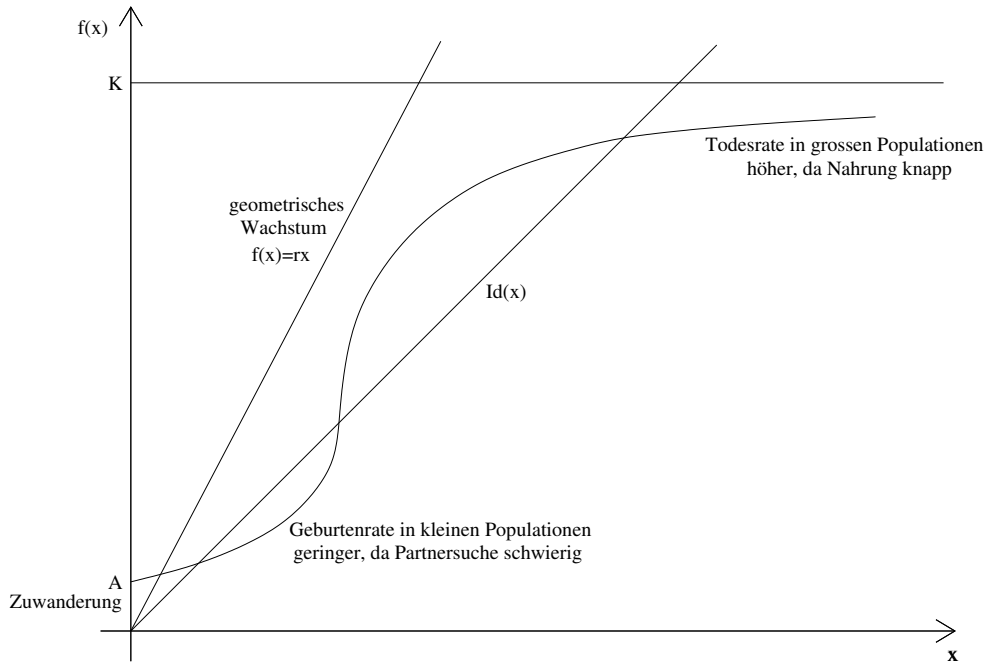


Abbildung 1.2: Vergleich der Iterationsfunktionen für das geometrische und das komplexere Modell.

1.1.3 Gleichgewichtspunkte oder Fixpunkte

Von besonderem Interesse in einem dynamischen Prozess sind die Werte x^*, auf die sich die dynamische Variable (z.B. die Populationsgröße) einstellt wenn man lang genug wartet, also $\lim_{n\to\infty} x_n = x^*$. Der Punkt x^* selbst erfüllt $f(x^*) = x^*$ und heißt **Gleichgewichtspunkt** oder **Fixpunkt**. Etwas präziser:

Definition 1 (Gleichgewichts- oder Fixpunkt) *Sei f eine Iterationsfunktion. Dann heißt eine Lösung x^* der Gleichung*

$$f(x) = x \tag{1.4}$$

Fixpunkt oder Gleichgewichtspunkt. Ist x^* ein Fixpunkt und $x_n = x^*$, dann ist $x_{n+1} = x_n = x^*$.

Graphisch lassen sich die Fixpunkte sehr einfach bestimmen als Schnittpunkte der Reproduktionsfunktion mit der Winkelhalbierenden. Wichtig für die Analyse des dynamischen Prozesses ist nun die Unterscheidung von stabilen und instabilen Fixpunkten:

Definition 2 *Ein Fixpunkt x^* heißt* **global stabil***, wenn die Folge der x_n unabhängig vom Startwert x_0 gegen x^* konvergiert (das heißt der Abstand von x_n zu x^* geht für große n gegen Null).*
Ein Fixpunkt x^* heißt **lokal stabil***, wenn die Folge der x_n für solche x_0, die nah genug an x^* liegen, gegen x^* konvergiert.*
Ein Fixpunkt x^* heißt **instabil***, wenn er nicht lokal stabil ist.*

Stabile Fixpunkte wirken **anziehend**; global stabile ziehen alles an, lokal stabile ziehen an, was in der Nähe ist. Instabile Fixpunkte **stoßen ab**, was in der Nähe ist.

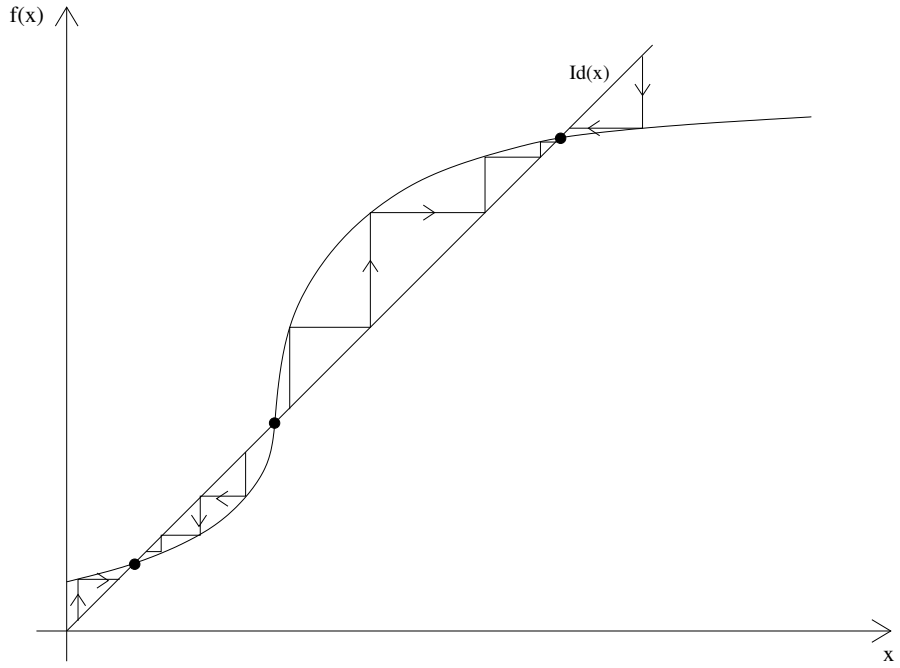


Abbildung 1.3: Iterationsfunktion und Cobwebbing für das Modell einer Vogelpopulation. Beachte, dass die Iterationsfunktion selbst nicht die Populationsgröße als Funktion der Zeit angibt. Die Zeitentwicklung gewinnen wir erst durch das Cobwebbing.

Kriterium für (lokale) Stabilität bzw. Instabilität: Ob ein Fixpunkt stabil oder instabil ist, können wir am konkreten Beispiel meist leicht mit dem Cobwebbing herausfinden. Für die Vogel-Population erhalten wir zum Beispiel zwei stabile und einen instabilen Fixpunkt (siehe Abb. 1.3). Mit Hilfe des graphischen Verfahrens können wir aber auch ein allgemeines Kriterium für die lokale Stabilität eines Fixpunktes ermitteln. Wir betrachten dazu wie sich das Cobweb in Abhängigkeit der Steigung der Iterationsfunktion am Schnittpunkt mit der Winkelhalbierenden ändert. Wir sehen (vgl. Abb. 1.3):

- Ist die Steigung der Iterationsfunktion am Fixpunkt größer als 1 ($f(x)$ schneidet die Winkelhalbierende von unten nach oben), so kommt man auf dem Cobweb immer weiter vom Fixpunkt weg: der Fixpunkt ist instabil.
- Ist die Steigung kleiner als die Steigung der Winkelhalbierenden, aber größer als -1, so nähert man sich auf dem Cobweb dem Fixpunkt an: der Fixpunkt ist (lokal) stabil.
- Ist die Steigung schließlich kleiner als -1 (in biologischen Modellen selten), so ist der Fixpunkt wiederum instabil.

Ist die Steigung genau gleich eins oder gleich -1, so lässt dieses Kriterium keine Aussage zu. Das Kriterium liefert auch keine Information über das *globale* Verhalten!
Mathematisch ausgedrückt lautet das Kriterium: Wir nehmen an, dass x^* ein Fixpunkt der Reproduktionsfunktion f ist. Wenn f eine Funktion ist, die man im Punkt x^* ableiten kann, dann gilt

$$\begin{array}{lcll} |f'(x^*)| < 1 & \Rightarrow & x^* & \text{ist lokal stabil,} \\ |f'(x^*)| > 1 & \Rightarrow & x^* & \text{ist instabil.} \end{array} \tag{1.5}$$

1.2 Nicht-lineare Prozesse

Bei geometrischem Wachtum der Form $f(x) = rx$ können auch anfänglich kleine Populationen innerhalb weniger Generationen zu einer enormen Größe anwachsen wenn der Wachstumsfaktor $r > 1$ ist. Am Beispiel der Vogelpopulation haben wir aber schon im letzten Abschnitt argumentiert, dass dies für große Populationen nicht realistisch ist, da sich das Wachstum bei Überbevölkerung aufgrund von Nahrungsknappheit abschwächen muss. Die mittlere Anzahl der Geburten, b, und die mittlere Anzahl der Tode, d, pro Individuum und Zeiteinheit (und damit $r = 1 + b - d$) hängen selbst von der Populationsgröße ab. Für unser Modell einer Reproduktionsfunktion heißt dies, wir müssen zur allgemeineren Form

$$f(x) = x(1 + b(x) - d(x)) = xr(x)$$

übergehen, in der b, d und r selbst Funktionen von x sind. Beim geometrischen Wachstum war die Reproduktionsfunktion eine Gerade. Man nennt den zugehörigen Prozess deshalb auch einen *linearen Prozess*. Prozesse mit Iterationsfunktionen, die keine Geraden sind, nennt man entsprechend *nicht-lineare Prozesse*. In diesem Abschnitt werden wir ein Modell eines nichtlinearen Prozesses genauer analysieren und auch die biologische Bedeutung des geometrischen Wachstums noch einmal diskutieren.

1.2.1 Modellbildung II: Das Verhulst-Modell

Betrachte einen Histidin-auxotrophen Bakterienstamm, d.h. die Bakterien sind auf Histidinzufuhr angewiesen, um sich zu vermehren. Zu Beginn der Kultivierung einer kleinen Zahl von Bakterien in einem Medium, das eine bestimmte Menge Histidin enthält, werden die Bakterien fast geometrisch wachsen. Wenn die Zahl der Bakterien zunimmt, ohne dass frisches Medium zugeführt wird, beginnt das Histidin knapp zu werden, und die Wachstumsrate wird sich verlangsamen. Um solche und ähnliche Situationen zu beschreiben, macht man oft den folgenden Ansatz:

$$x_{n+1} = x_n r(x_n) = x_n\, r_0\, \frac{c}{c + r_0 x_n} \tag{1.6}$$

mit Konstanten $r_0, c > 0$. Ohne den Bruch wäre das Wachstum geometrisch mit Wachstumsfaktor $r = r_0$. Wenn es unbegrenzte Ressourcen an Histidin gäbe wäre $r_0 x_n$ die Zahl der überlebenden Bakterien in der Generation $n+1$. Das Histidin ist aber beschränkt und wird umso knapper, je mehr Bakterien sich darum 'streiten'. Wir sollten im Modell also berücksichtigen, dass ein Teil der hoffnungsvoll in die nächtes Generation gestarteten Bakterien nicht genügend Nahrung erhält und verhungert. Dies wird durch den Bruch in Gleichung (1.6) erreicht. Dieser Bruch beschreibt den Anteil der nicht verhungerten Bakterien. Er ist für $x_n > 0$ immer kleiner als 1 und wird mit steigendem x_n (d.h. mit knapperen Ressourcen) immer kleiner. Die Gleichung (1.6) ist das sogenannte **Verhulst-Modell**. Die Reproduktionsfunktion lautet

$$f(x) = x\, \frac{r_0 c}{c + r_0 x} = \frac{r_0\, x}{1 + x/c}\,; \tag{1.7}$$

dies ist die Gleichung einer **Hyperbel**. Wir interessieren uns nur für den Bereich $x \geq 0$. Reproduktionsfunktion und Populationsentwicklung sind in Abb. 1.4 gezeigt.

Analyse des Verhulst-Modells

Wir interessieren uns für die Zeitentwicklung der Populationsgröße. Eine vollständige Lösung dieser Frage wäre ihre explizite Angabe, mathematisch ausgedrückt: die explizite Angabe der

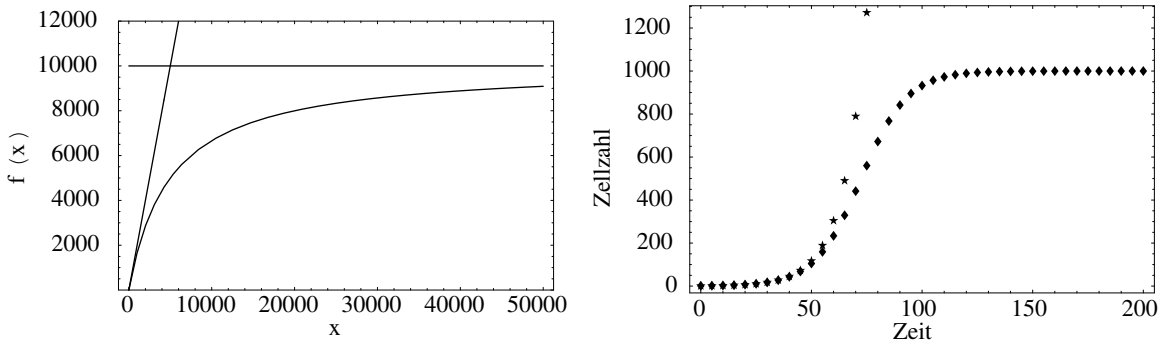


Abbildung 1.4: Reproduktionsfunktionen (links) und Populationsentwicklung (rechts) für das Verhulstmodell mit $r_0 = 1.1$ und $c = 11000$ und für ungebremstes geometrisches Wachstum mit $r = 1.1$. Zeit ist in Generationen gemessen; jede 5. Generation ist gezeigt (Rauten = Verhust, Sterne = geometrisch).

Folge der x_n, die über die Reproduktionsfunktion iterativ definiert ist. Bei einem nicht-linearen Prozess ist dies in der Regel aber nicht möglich. Aus der Analyse der Reproduktionsfunktion können wir aber dennoch sehr viele biologisch wichtige Informationen über den Prozess gewinnen. Dies werden wir nun Schritt für Schritt am Beispiel des Verhulst-Modells diskutieren.

Nullstellen $f(x) = 0$ genau dann, wenn $x = 0$: 'Von nichts kommt nichts'.

Asymptotisches Verhalten Wir untersuchen das Verhalten von $f(x)$, wenn x sehr groß wird. Es gilt:

$$\lim_{x\to\infty} f(x) = \lim_{x\to\infty} \frac{c}{1 + \frac{c}{r_0 x}} = c,$$

die Reproduktionsfunktion hat also eine (horizontale) Asymptote. Das bedeutet: Egal wie viele Individuen man (etwa in ein Habitat) einsetzt, sind bereits in der nächsten Generation – wegen der beschränkten Ressourcen – nur noch höchstens c Individuen vorhanden. Dies entspricht der beschränkten *carrying capacity* im Beispiel der Vogelpopulation im letzten Abschnitt.

Erste Ableitung Die Anwendung der Quotientenregel ergibt

$$f'(x) = \frac{c^2\, r_0}{(r_0 x + c)^2} > 0$$

für alle $x \geq 0$. f ist somit monoton steigend und besitzt keine lokalen Extrema. Das heißt: Je mehr "Eltern" wir haben, desto mehr Nachkommen gibt es in der nächsten Generation.

Zweite Ableitung

$$f''(x) = \frac{-2\, c^2\, r_0}{(r_0 x)^3} < 0$$

für alle $x \geq 0$, $f(x)$ macht also eine Rechtskurve. Das heißt, dass die Steigung von $f(x)$ für große x immer kleiner wird: das Wachstum wird immer stärker gebremst.

Die Fixpunkte des Verhulst-Modells

Die Lösung der Fixpunktgleichung (1.4) für das Verhulst-Modell liefert als Fixpunkte $x_1^* = 0$ und $x_2^* = a - a/r_0$. x_2^* ist nur dann positiv und somit biologisch relevant, wenn $r_0 > 1$ ist. Charakterisierung der Fixpunkte:

Fall 1: $r_0 \leq 1$: In dieser Situation ist 0 der einzige relevante Fixpunkt. Da f rechtsgekrümmt ist und für $x \geq 0$ keinen Schnittpunkt mit der Identitätsabbildung $\mathrm{Id}(x)$ hat, muss die Reproduktionsfunktion unterhalb von $\mathrm{Id}(x)$ liegen, also $f(x) \leq \mathrm{Id}(x)$ für $x \geq 0$. 0 ist global stabiler Fixpunkt, die Population stirbt also allmählich aus, egal, wie groß sie zu Anfang war.

Fall 2: $r_0 > 1$: 0 und $a - a/r_0$ sind die beiden biologisch relevanten Fixpunkte. Da f rechtsgekrümmt ist, liegt der Graph der Reproduktionsfunktion für $0 < x < a - a/r_0$ oberhalb, für $x > a - a/r_0$ dagegen unterhalb der Identitätsabbildung. $x_2^* = a - a/r_0$ ist anziehender, $x_1^* = 0$ ist abstoßender Fixpunkt. Die Population wächst oder schrumpft also so lange, bis sie die Größe x_2^* erreicht, egal, wie groß sie zu Anfang ist (s. Abb. 1.4).

Man beachte die unterschiedliche Bedeutung von c und $x_2^* = a - a/r_0$. Die Asymptote der Reproduktionsfunktion, c, entspricht der maximalen Individuenzahl *nach einer Generation*. Der Fixpunkt x_2^* ist (für $r_0 > 1$) diejenige Individuenzahl, die sich *nach vielen Generationen* schließlich einstellt.

1.2.2 Modellanalyse II: Rückführung auf einen linearen Prozess

Das geometrische Wachstum ist einer der ganz wenigen Prozesse für den man die Zeitentwicklung explizit angeben kann. Man versucht deshalb oft, sich dies auch für kompliziertere Systeme nützlich zu machen: Die Rückführung auf einen linearen Prozess ist eine wichtige Technik zur Analyse nicht-linearer Prozesse.

Lokale Approximation

Eine der wichtigsten Techniken bei der Modellanalyse ist die lokale Approximation: Wenn man einen nicht-linearen Prozess nur für eine beschränkte Anzahl von Generationen bzw. einen kleinen Bereich von Populationsgrößen betrachtet (also nur "lokal"), kann man ihn in der Regel näherungsweise durch einen linearen Prozess (also das geometrische Wachstum) beschreiben. Wir erinnern uns daran, dass die Ableitung einer Funktion f an der Stelle x_0, also $f'(x_0)$, gleich der Steigung der Tangente im Punkt x_0 an den Graphen von f ist. Für kleine Δx darf man daher den Wert der Funktion an der Stelle $x = x_0 + \Delta x$ durch den entsprechenden Wert auf der Tangente annähern (s. Abb. 1.5):

$$f(x_0 + \Delta x) \approx f(x_0) + f'(x_0) \cdot \Delta x = f(x_0) + f'(x_0) \cdot (x - x_0)\,. \tag{1.8}$$

Im letzten Schritt haben wir $\Delta x = x - x_0$ eingesetzt. Da die Tangente eine Gerade ist, spricht man von einer **linearen Näherung**. Wir diskutieren dies an einem Beispiel.

Beispiel: Kleine Bakterienpopulation Solange es genügend Nahrung für alle gibt, sollte das Wachstum einer Population auch nicht gebremst sein. Wir wollen im Verhulst-Modell zeigen, dass die Bakterienpopulation für kleine Populationsgröße x tatsächlich annähernd geometrisch wächst. Uns interessiert also der Prozess in der Nähe von ("lokal um") $x_0 = 0$. Dort erhalten wir für das Verhulst-Modell (Gleichung 1.7)

$$f(x) \approx f(0) + f'(0)(x - 0) = r_0\, x. \tag{1.9}$$

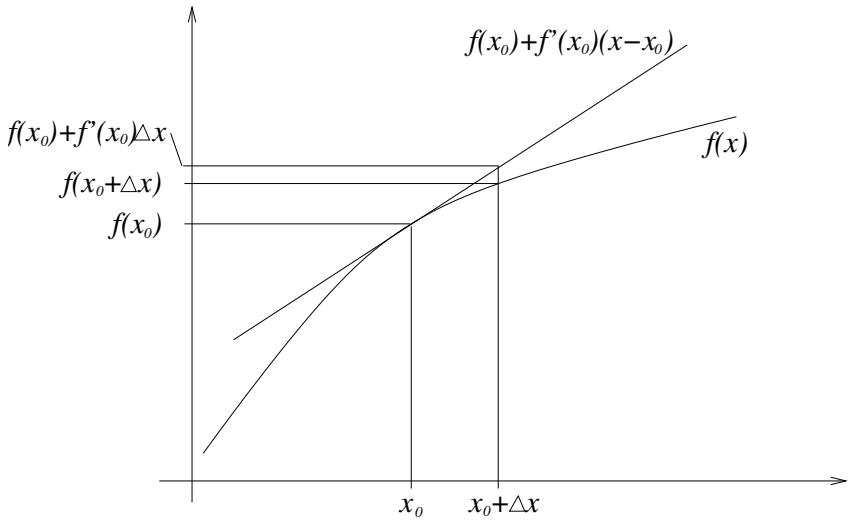


Abbildung 1.5: Lineare Approximation einer Funktion.

Das ist aber gerade die Reproduktionsfunktion für geometrisches Wachstum mit $r = r_0$. Betrachten wir den Fall $r_0 = 1.1$ und $c = 11000$ wie in Abb. 1.4. Wir wollen zeigen, dass die Approximation der Verhulst-Reproduktionsfunktion durch $f(x) = 1.1x$ für kleine x erstaunlich gut ist. Dazu starten wir mit $x_0 = 1$ und berechnen x_n mittels Gleichung (1.7) bzw. mittels der Approximation (1.9):

Zeit	1	2	3	4	5	6	7	8	9	10	...	50
x_n exakt	1.1	1.2	1.3	1.5	1.6	1.8	1.9	2.1	2.4	2.6	...	105
x_n approximativ	1.1	1.2	1.3	1.5	1.6	1.8	1.9	2.1	2.4	2.6	...	117

Wie man auch aus Abb. 1.4 sieht, ist das geometrische Wachstum für die ersten ≈ 50 Generationen eine sehr gute Näherung: so lange ist noch genügend Histidin für alle da. Mit weiter ansteigender Populationsgröße wird die Approximation durch das ungebremste geometrische Wachstum dann aber sehr schlecht.

Transformation auf einen linearen Prozess

Manchmal kann man einen komplizierten biologischen Prozess auch so umformen, dass man am Ende wieder auf das geometrische Wachtum kommt. Wir besprechen dies an einem Beispiel, das deutlich macht, warum das geometrische Wachstum für die Biologie eine so zentrale Bedeutung hat.

Konkurrenz zwischen zwei Populationen Wir betrachten zwei Bakterienstämme, "a-Bakterien" und "b-Bakterien", die wie im obigen Beispiel auf Histidin angewiesen sind. Das Wachstum der einzelnen Populationen erfolgt wieder nach dem Verhulst-Modell. Wir nehmen an, dass die b-Bakterien etwas schneller wachsen als die a-Bakterien und wählen deshalb den anfänglichen Wachstumsfaktor $r_b > r_a$. Die Kapazität c, die durch die Histidinmenge vorgegeben ist, ist in beiden Fällen gleich. Es stellt sich nun die Frage: Was passiert, wenn wir eine Mischung aus a- und b-Bakterien auf das Medium geben? Wir können versuchen, dies in einem gemeinsamen Modell zu beschreiben.
Solange genügend Histidin vorhanden ist (also bei kleiner Größe der Gesamtpopulation), sollte das Wachstum in jedem Bakterienstamm wieder näherungsweise geometrisch sein. Wenn wir

die Populationsgrößen in der n-ten Generation mit a_n und b_n bezeichnen ist für kleine a_n und b_n also $a_{n+1} = r_a a_n$ und $b_{n+1} = r_b b_n$. In der Konkurrenz ums Futter spielen jetzt aber nicht nur die Nachkommen des eigenen Stamms eine Rolle, sondern auch die Nachkommen des jeweils anderen Stamms. Es ist deshalb sinnvoll, den Bruch im Verhulst-Modell (Gleichung 1.6), der den Anteil der Bakterien angibt, die nicht verhungern, von der gesamten konkurrierenden Nachkommenschaft $r_a a_n + r_b b_n$ abhängig zu machen. Damit ist:

$$a_{n+1} = f_1(a_n, b_n) = a_n \, r_a \, \frac{c}{c + r_a a_n + r_b b_n} \tag{1.10}$$

$$b_{n+1} = f_2(a_n, b_n) = b_n \, r_b \, \frac{c}{c + r_a a_n + r_b b_n} \tag{1.11}$$

Diese Gleichungen sind nicht nur nicht-linear, sie hängen auch noch von beiden Populationsgrößen, a_n und b_n, ab. Es handelt sich also um einen mehrdimensionalen Prozess. Solche Prozesse (der Normalfall in der Biologie!) sind in der Regel nur approximativ behandelbar. In unserem Beispiel kann man den Prozess aber stark vereinfachen, wenn man statt nach den einzelnen Populationsgrößen nur nach dem *Verhältnis* $z_n := b_n/a_n$ fragt. Dann erhält man für $z_{n+1} = b_{n+1}/a_{n+1}$

$$z_{n+1} = f_3(z_n) = \frac{r_b}{r_a} \, z_n \, . \tag{1.12}$$

Wir sehen also: Das Verhältnis in der Nachkommengeneration z_{n+1} hängt nur vom Verhältnis in der Elterngeneration z_n ab. Und: Der Prozess ist jetzt linear. Das Wachstum des Verhältnisses der Populationsgrößen ist einfach geometrisch mit $r = r_b/r_a$. Da z_n deshalb *ungebremst* wächst und nach einigen Generationen sehr groß wird heißt dies, dass in der Gesamtpopulation bald (fast) nur noch b-Bakterien zu finden sein werden. Biologisch heißt dies: die b-Bakterien setzen sich durch und die a-Bakterien sterben aus.
Genau diese Überlegung war für Charles Darwin der entscheidende Gedanke für seine Theorie der Evolution durch natürliche Selektion: Wenn eine Art auch nur einen kleinen selektiven Vorteil hat (hier ein etwas schnelleres Wachstum, $r_b > r_a$), dann wird sie sich mit der ganzen Macht des geometrischen Wachstums gegen andere Arten durchsetzen können. Das geometrische Wachstum ist letztlich der Grund dafür, dass natürliche Selektion so wirksam ist und dass es biologische Evolution überhaupt gibt.

1.3 Mehrdimensionale Prozesse

In den letzten Abschnitten haben wir lineare und nicht-lineare Prozesse für eine einzige in der Zeit veränderliche biologische Größe betrachtet. Wenn es zunächst mehrere zeitveränderliche Größen gab, wie die Populationsgrößen a_n und b_n der konkurrierenden Bakterienstämme in Abschnitt 1.2.2, haben wir den Prozess so transformiert, dass wir am Ende nur noch eine solche Größe hatten (bei den Bakterien das Verhältnis der Populationsgrößen $z_n = b_n/a_n$). Einen Prozess mit einer zeitveränderlichen Größe nennt man auch *eindimensional*. In biologischen Prozessen spielen sehr oft viele veränderliche Faktoren eine Rolle, und meistens kann man sie nicht so einfach wegtransformieren. Man hat es dann mit einem *mehrdimensionalen* Prozess zu tun. In diesem Abschnitt wollen wir Techniken zur Beschreibung von mehrdimensionalen Prozessen erarbeiten. Um es nicht gleich zu kompliziert zu machen, werden wir uns auf lineare Prozesse beschränken. Zunächst einmal ein Beispiel:

1.3.1 Modellbildung III: Ein Modell für Populationsstruktur

Wir betrachten ein weiteres Mal die Vogelpopulation aus den Abschnitt 1.1.1. Wir nehmen aber an, dass diese Vögel auf zwei benachbarten Inseln mit unterschiedlichen Bedingungen (z.B. unterschiedlicher Vegetation) leben. In der Ökologie nennt man so etwas eine geographisch strukturierte Population. Die Populationsgrößen $x_1(n)$ und $x_2(n)$ verändern sich von Jahr zu Jahr durch Geburten und Tode und durch Migration zwischen den Inseln. Wir machen also folgenden Ansatz:

$$x_1(n+1) = x_1(n) + B_1(n) - D_1(n) + M_{1\leftarrow 2}(n) - M_{2\leftarrow 1}(n) \tag{1.13}$$
$$x_2(n+1) = x_2(n) + B_2(n) - D_2(n) + M_{2\leftarrow 1}(n) - M_{1\leftarrow 2}(n) \tag{1.14}$$

Beachte, dass der Index hier für die Insel steht. Im Unterschied zu unserer bisherigen Notation schreiben wir die Generationsvariable n jetzt als Argument hinter das x. B und D sind die Geburten und Tode pro Generation, wir nehmen der Einfachheit halber an, dass diese Zahlen auf jeder Insel proportional zur Populationsgröße ist, also $B_1(n) = b_1 x_1(n)$ und $D_1(n) = d_1 x_1(n)$ (entsprechend für die zweite Insel). Wir definieren wieder Reproduktionsparameter $r_1 = 1 + b_1 - d_1$ und $r_2 = 1 + b_2 - d_2$, die dann ebenfalls nicht von der Populationsgröße abhängen. r_1 und r_2 können unterschiedlich groß sein. Wir nehmen an, dass die Zahl der Migranten jeweils proportional zur Populationsgröße auf der Ausgangs-Insel ist, unabhägig von der Situation auf der Ziel-Insel, also $M_{1\leftarrow 2}(n) = m_{1\leftarrow 2} x_2(n)$ und $M_{2\leftarrow 1}(n) = m_{2\leftarrow 1} x_1(n)$. Damit haben wir insgesamt die folgenden Iterationsvorschriften für unser Modell:

$$x_1(n+1) = (r_1 - m_{2\leftarrow 1}) x_1(n) + m_{1\leftarrow 2} x_2(n) \tag{1.15}$$
$$x_2(n+1) = (r_2 - m_{1\leftarrow 2}) x_2(n) + m_{2\leftarrow 1} x_1(n) \tag{1.16}$$

Da auf der rechten Seite keine Produkte, Potenzen oder Quotienten von x_1 und x_2 vorkommen, ist das Modell linear.

Um in einem Modell wie dem obigen rechnen zu können, benötigen wir zunächst einmal einen geeigneten mehrdimensionalen Formalismus. Einen Rahmen hierfür bietet die lineare Algebra, von der wir jetzt einige Grundkonzepte einführen werden.

1.3.2 Vektoren und Matrizen

Vektor: Wir definieren einen m-dimensionalen Vektor $\boldsymbol{a}$ als eine Folge von m in einer Spalte angeordneter Zahlen,

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_m \end{pmatrix} . \tag{1.17}$$

Vektoren werden oft (wie hier) durch fett gedruckte Buchstaben bezeichnet. Die Einträge a_i heißen auch Elemente oder Komponenten des Vektors. Graphisch kann man einen zweidimensionalen Vektor auch als Pfeil in der Ebene darstellen.
Wenn man sich in einem biologischen System für mehrere zeitveränderliche Größen interessiert, fasst man diese oft in einem Vektor zusammen. Zum Beispiel geben in der Populationsbiologie die Komponenten des *Populationsvektors* die Größen verschiedener Teilpopulationen an (Populationen auf verschiedenen Inseln, verschiedene Spezies, Wildtypen und Mutanten, ...). Die Länge eines Vektors (die in Anwendungen in der Physik eine wichtige Rolle spielt) hat in der Biologie meist keine direkte Bedeutung. Die Summe der Vektoreinträge dagegen schon: Bei einem Populationsvektor ist das gerade die Größe der Gesamtpopulation. Die Richtung des Vektors bestimmt das Größenverhältnis der Teilpopulationen.

Matrix: Eine Matrix $\mathcal{A}$ ist ein rechteckiges Feld aus Zeilen und Spalten der Form:

$$\mathcal{A} := \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix} =: (a_{ij}) \tag{1.18}$$

Die a_{ij} heißen Elemente oder Komponenten der Matrix. Der erste Index eines Matrixelementes a_{ij} ist die Zeilennummer, der zweite die Spaltennummer; a_{ij} ist also das Element in der i-ten Zeile und j-ten Spalte. Die **Dimension** der Matrix wird durch die Anzahl der Zeilen und die Anzahl der Spalten festgelegt: Eine $(m \times n)$-Matrix hat m Zeilen und n Spalten. $\mathcal{A}$ aus Gl. (1.18) ist eine (3×4) Matrix. Ein m-dimensionaler Vektor ist nichts anderes als eine Matrix mit der Dimension $(m \times 1)$.

Rechnen mit Vektoren und Matrizen

Addition: Haben zwei Vektoren oder Matrizen dieselbe Dimension, so erfolgen die **Addition** und **Subtraktion** elementweise. D.h. der Vektor $\boldsymbol{c} = \boldsymbol{a} \pm \boldsymbol{b}$ ist definiert durch $c_i = a_i \pm b_i$, die Matrix $\mathcal{C} = \mathcal{A} \pm \mathcal{B}$ entsprechend durch $c_{ij} = a_{ij} \pm b_{ij}$. Für Vektoren kann man (in zwei Dimensionen) Addition und Subtraktion durch geeignetes Aneinanderlegen der Vektorpfeile graphisch darstellen.

Multiplikation: Wir unterscheiden zwei Typen von Multiplikationen. Die einfachste Version ist die Multiplikation einer Matrix oder eines Vektors mit einer reellen Zahl λ. Wir definieren:

$$\lambda\mathcal{A} := (\lambda a_{ij}) .$$

Da wir Vektoren als $(m \times 1)$-Matrizen auffassen können, gilt diese Definition gleichzeitig für das Produkt eines Vektors mit einer Zahl. Es gelten dann die folgenden Rechenregeln:

$$(\lambda_1 \, \lambda_2) \, \mathcal{A} = \lambda_1 \, (\lambda_2 \mathcal{A}) \tag{1.19}$$

$$\lambda(\mathcal{A} + \mathcal{B}) = \lambda\mathcal{A} + \lambda\mathcal{B} \tag{1.20}$$

$$(\lambda_1 + \lambda_2)\mathcal{A} = \lambda_1\mathcal{A} + \lambda_2\mathcal{A}. \tag{1.21}$$

(1.19) ist ein Assoziativgesetz, (1.20) und (1.21) sind Distributivgesetze.
Matrizen können auch miteinander multipliziert werden. Diese **Matrixmultiplikation** ist allerdings etwas komplizierter. Nehmen wir an, $\mathcal{A}$ sei eine $(m \times k)$ Matrix und $\mathcal{B}$ eine $(k \times n)$ Matrix. Das Produkt $\mathcal{C} = \mathcal{A} \cdot \mathcal{B}$ ist dann definiert durch

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{ik}b_{kj} = \sum_{\ell=1}^{k} a_{i\ell}\, b_{\ell j}. \tag{1.22}$$

Um c_{ij} zu bekommen nimmt man also die j-te Spalte der rechten Matrix ($\mathcal{B}$) und die $i$te Zeile der linken Matrix ($\mathcal{A}$), und berechnet dann "erstes Element der $j$ten Spalte mal erstes Element der $i$ten Zeile plus zweites Element der $j$ten Spalte mal zweites Element der $i$ten Zeile, ...". Man beachte, dass die Definition erfordert, dass die Zeilen von $\mathcal{A}$ und die Spalten von $\mathcal{B}$ die gleiche Länge k haben. Die neue Matrix $\mathcal{C}$ hat dann die Dimension $(m \times n)$.
Wir werden im Folgenden nur zwei Fälle brauchen: Die Multiplikation von (2×2) Matrizen und die Multiplikation einer Matrix mit einem Vektor. Dazu jeweils ein Zahlenbeispiel:

$$\begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 5 & 1 \end{pmatrix} \quad ; \quad \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 5 \\ 6 \end{pmatrix} = \begin{pmatrix} 17 \\ 39 \end{pmatrix}$$

Wir sehen insbesondere, dass die Multiplikation einer Matrix mit einem Vektor wieder einen Vektor ergibt. Man kann das auch so sagen: Eine $(m \times m)$ Matrix definiert eine Funktion, die m-dimensionale Vektoren auf m-dimensionale Vektoren abbildet.
Allgemein gelten die folgenden **Rechenregeln** für die Multiplikation von Matrizen:

$$(\mathcal{A}\mathcal{B})\mathcal{C} = \mathcal{A}(\mathcal{B}\mathcal{C}), \quad \text{und} \quad \mathcal{A}(\lambda\mathcal{B}) = (\lambda\mathcal{A})\mathcal{B} = \lambda(\mathcal{A}\mathcal{B}); \tag{1.23}$$

$$(\mathcal{A} + \mathcal{B})\mathcal{C} = \mathcal{A}\mathcal{C} + \mathcal{B}\mathcal{C} \quad \text{und} \quad \mathcal{A}(\mathcal{B} + \mathcal{C}) = \mathcal{A}\mathcal{B} + \mathcal{A}\mathcal{C}\,. \tag{1.24}$$

(1.23) sind wieder Assoziativgesetze, (1.24) Distributivgesetze. **Achtung:** Für die Matrixmultiplikation gilt das Kommutativgesetz nicht ($\mathcal{A}\,\mathcal{B}$ ist normalerweise nicht gleich $\mathcal{B}\,\mathcal{A}$)!

Mehrdimensionale Prozesse und Matrizen

Wir wenden uns nun wieder dem dynamischen Prozess der Vogelpopulationen aus Abschnitt 1.3.1 zu. Mit Hilfe des eben eingeführten Matrixprodukts können wir die Iterationsgleichungen (1.15) folgendermassen zusammenfassen:

$$\begin{pmatrix} x_1(n+1) \\ x_2(n+1) \end{pmatrix} = \begin{pmatrix} r_1 - m_{2\leftarrow 1} & m_{1\leftarrow 2} \\ m_{2\leftarrow 1} & r_2 - m_{1\leftarrow 2} \end{pmatrix} \begin{pmatrix} x_1(n) \\ x_2(n) \end{pmatrix} \tag{1.25}$$

Wir wollen nun einen mehrdimensionalen Prozess ganz allgemein in Vektorschreibweise definieren. Hierzu vergegenwärtigen wir uns zunächst noch einmal die Definition eines solchen Prozesses in einer Dimension:

- In einer Dimension ist ein dynamischer Prozess durch eine Folge von Zahlen $x(n)$ (z.B. Populationsgrößen) gegeben. Diese Folge kann man durch den Startwert $x(0)$ und die Iterationsfunktion definieren. Die Iterationsfunktion f gibt an, wie man von einem Folgeglied zum nächsten kommt, also $x(n+1) = f(x(n))$. Im einfachsten Fall (dem linearen Prozess) ist die Iterationsfunktion einfach die Multiplikation mit einer konstanten (nicht von x abhängigen) Zahl: $x(n+1) = rx(n)$.

Entsprechend kann man einen mehrdimensionalen Prozess durch eine Folge von Vektoren $\boldsymbol{x}(n)$ mit Elementen $x_i(n)$ darstellen. Wie im eindimensionalem Fall kann man die Folge durch einen Startwert $\boldsymbol{x}(0)$ und eine Iterationsfunktion f definieren. Die Iterationsfunktion bildet also Vektoren auf Vektoren ab. Im einfachsten (linearen) Fall ist sie die Multiplikation mit einer konstanten (nicht von den x_i abhängigen) Matrix $\mathcal{R}$:

$$\boldsymbol{x}(n+1) = f(\boldsymbol{x}(n)) = \mathcal{R}\boldsymbol{x}(n). \tag{1.26}$$

Bei einem linearen Prozess in einer Dimension (also dem geometrischen Wachstum) konnten wir den Zeitverlauf der dynamischen Variablen auch explizit angeben: $x(n) = r^n x(0)$. Mit Hilfe der Matrixmultiplikation geht dies für mehrere Dimensionen ganz entsprechend:

$$\boldsymbol{x}(1) = \mathcal{R}\boldsymbol{x}(0) \tag{1.27}$$

$$\boldsymbol{x}(2) = \mathcal{R}\boldsymbol{x}(1) = \mathcal{R}\mathcal{R}\boldsymbol{x}(0) = \mathcal{R}^2\boldsymbol{x}(0) \tag{1.28}$$

und allgemein

$$\boldsymbol{x}(n) = \mathcal{R}^n\boldsymbol{x}(0). \tag{1.29}$$

In einer Dimension hatten wir mit dieser Form schon die vollständige Lösung unseres Problems. In mehreren Dimensionen ist es aber nur ein Zwischenschritt: Wir interessieren uns schließlich für den expliziten Zeitverlauf der einzelnen Vektoreinträge oder ihrer Summe, der Vektor selbst ist dabei nur eine Hilfsgröße. Wie verändert sich ein Vektor, wenn wir ihn wiederholt mit der selben Matrix multiplizieren? Um hierfür ein Gefühl zu bekommen, untersuchen wir zuerst einen ganz einfachen Fall.

Beispiel Wir betrachten wieder die Vogelpopulation auf den zwei Inseln, mehmen aber an, dass es keine Wanderung zwischen den Inseln gibt, die Migration ist also gleich Null. Natürlich haben wir dann eigentlich zwei unabhängige eindimensionale Prozesse und brauchen den mehrdimensionalen Formalismus nicht. Wir wollen das Problem aber trotzdem in diesem Rahmen beschreiben.

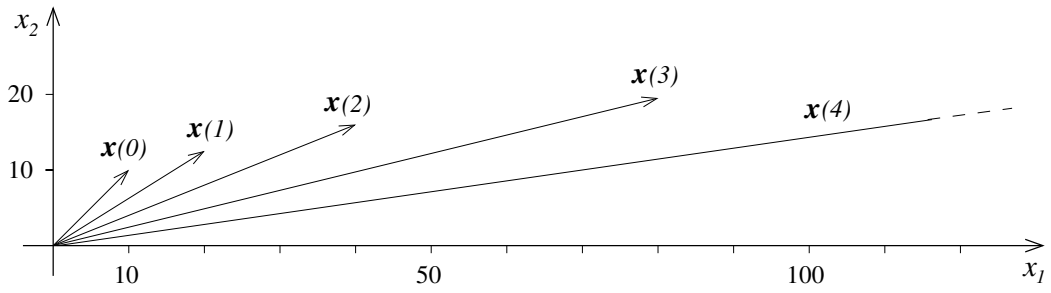


Abbildung 1.6: Entwicklung eines Populationsvektors für das Zwei-Insel Modell ohne Migration. Die Reproduktionsparameter sind $r_1 = 2$ und $r_2 = 1.25$.

In Abbildung 1.6 sehen wir, dass der Populationsvektor durch die Multiplikation mit der Matrix gestreckt und gedreht wird. Es gibt allerdings zwei spezielle Richtungen, in denen ein Vektor nur gestreckt und nicht gedreht wird: Dies ist genau dann der Fall, wenn der Populationsvektor auf der x_1-Achse oder der x_2-Achse liegt, wenn die ganze Population also auf einer Insel konzentriert ist. Vektoren auf der x_1-Achse werden mit r_1 multipliziert, und Vektoren auf der x_2-Achse mit

r_2. Für einen beliebigen Populationsvektor können wir nun wie folgt vorgehen: Offensichtlich können wir jeden beliebigen Vektor $\boldsymbol{x}(0)$ auch als Summe von Vektoren auf der x_1- und x_2-Achse schreiben:

$$\boldsymbol{x}(0) = \boldsymbol{v}^{(1)} + \boldsymbol{v}^{(2)} \tag{1.30}$$

mit

$$\boldsymbol{v}^{(1)} = \begin{pmatrix} x_1(0) \\ 0 \end{pmatrix} \quad \text{und} \quad \boldsymbol{v}^{(2)} = \begin{pmatrix} 0 \\ x_2(0) \end{pmatrix};.$$

Da $\boldsymbol{v}^{(1)}$ und $\boldsymbol{v}^{(2)}$ in jeder Generation einfach mit r_1 bzw. r_2 multipliziert werden, gilt für die Zeitentwicklung des Populationsvektors $\boldsymbol{x}$

$$\boldsymbol{x}(n) = r_1^n \boldsymbol{v}^{(1)} + r_2^n \boldsymbol{v}^{(2)}\,. \tag{1.31}$$

Im Unterschied zu Gleichung (1.27) können wir aus dieser Form sofort die Zeitentwicklung für die Teilpopulationen oder die Gesamtpopulation gewinnen. Wir können zum Beispiel fragen, wie sich der Anteil der Teilpopulationen an der Gesamtpopulation nach langer Zeit einstellt. Wenn (wie in Abb. 1.6) $r_1 > r_2$ ist, ergibt sich für den Anteil der Population auf der ersten Insel:

$$\lim_{n\to\infty} \frac{x_1(n)}{x_1(n)+x_2(n)} = \lim_{n\to\infty} \frac{r_1^n x_1(0)}{r_1^n x_1(0) + r_2^n x_2(0)} = \frac{x_1(0)}{x_1(0) + \lim_{n\to\infty}(r_2/r_1)^n x_2(0)} = 1 \tag{1.32}$$

der Anteil der ersten Teilpopulation wächst also stetig an und geht im Limes vieler Generationen gegen 1. Dies sieht man auch im Beispiel in Abbildung 1.6: Für wachsendes n wird der Winkel zwischen dem Populationsvektor und der x_1-Geraden immer kleiner. Für das Wachstum der Gesamtpopulation nach langer Zeit finden wir entsprechend:

$$\lim_{n\to\infty} \frac{x_1(n+1)+x_2(n+1)}{x_1(n)+x_2(n)} = \lim_{n\to\infty} \frac{r_1^{n+1}x_1(0) + r_2^{n+1}x_2(0)}{r_1^n x_1(0) + r_2^n x_2(0)} \tag{1.33}$$

$$= \lim_{n\to\infty} \frac{r_1 x_1(0) + r_2(r_2/r_1)^n x_2(0)}{x_1(0) + (r_2/r_1)^n x_2(0)} = r_1\,. \tag{1.34}$$

Im Limes spielt also nur noch der größere der beiden Reproduktionsparameter eine Rolle.
Was hilft uns dieses Beispiel nun für den allgemeinen Fall? Wir können uns hierzu folgendes überlegen: Angenommen, es gibt zwei beliebige Richtungen in der Ebene, in denen Vektoren durch die Multiplikation mit der Matrix nicht gedreht, sondern nur gestreckt (mit einer Zahl multipliziert) werden. Dann könnten wir wie im Beispiel den Anfangsvektor als Summe solcher Vektoren darstellen und erhalten dann die Zeitentwicklung analog zu (1.31). Im nächsten Abschnitt wollen wir deshalb die folgende Frage klären: Wann gibt es Vektoren, die durch eine Matrix auf ein Vielfaches von sich selbst abgebildet werden und wie kann man sie gegebenenfalls bestimmen?

1.3.3 Modellanalyse III: Eigenwerte und Eigenvektoren

Für die meisten Vektoren bedeutet die Multiplikation mit einer Matrix eine Streckung und eine Drehung, Vektoren verändern also ihre Länge und ihre Richtung. Wir hatten aber schon im letzten Beispiel gesehen, dass es spezielle Vektoren gibt, die mit einer Matrix nicht gedreht werden, sondern nur gestreckt (oder gestaucht). Die Matrix wirkt auf solche Vektoren also einfach wie ein Faktor. Solche Vektoren nennt man **Eigenvektoren** einer Matrix.

Definition 3 (Eigenvektor, Eigenwert) $\boldsymbol{v}$ *heisst* **Eigenvektor** *zur Matrix* $\mathcal{R}$ *zum* **Eigenwert** λ *wenn gilt*

$$\mathcal{R}\boldsymbol{v} = \lambda\boldsymbol{v} \quad \textit{bzw.} \quad \mathcal{R}\boldsymbol{v} - \lambda\boldsymbol{v} = \boldsymbol{0}. \tag{1.35}$$

0 *ist der Nullvektor (der Vektor mit allen Komponenten gleich 0). Wir setzten in der Definition voraus, dass* $\boldsymbol{v}$ *nicht selbst schon der Nullvektor ist. Der Eigenwert* λ *ist der Streckungs- oder Stauchungsfaktor mit dem* $\boldsymbol{v}$ *multipliziert wird (wenn er negativ ist, dreht sich durch die Multiplikation mit der Matrix die Orientierung des Vektors um). Für einen Eigenvektor spielt nur die Richtung eine Rolle, nicht seine Länge: Insbesondere ist mit* $\boldsymbol{v}$ *auch jeder gestreckte Vektor* $a\boldsymbol{v}$ *selbst wieder Eigenvektor zum gleichen Eigenwert.*

Oben hatten wir schon argumentiert, dass wir eine vollständige Lösung für die Zeitentwicklung eines zweidimensionalen biologischen Prozesses bekommen können, wenn wir zwei solche Eigenvektoren finden können. Dies wollen wir jetzt für eine allgemeine (2×2) Matrix $\mathcal{R}$ versuchen. Wir schreiben hierzu die Gleichung (1.35) noch einmal Zeile für Zeile auf:

$$\begin{aligned}(r_{11} - \lambda)v_1 + r_{12}v_2 &= 0\\ r_{21}v_1 + (r_{22} - \lambda)v_2 &= 0\,.\end{aligned}$$

Dies ist ein lineares Gleichungssystem (LGS). Von den r_{ij} nehmen wir an, dass sie bekannt sind, die Komponenten des Eigenvektors v_i und den Eigenwert λ müssen wir bestimmen. Wir gehen Schrittweise vor. In einem ersten Schritt teilen wir durch v_2. Wir bekommen dann[1]:

$$\begin{aligned}(r_{11} - \lambda)(v_1/v_2) + r_{12} &= 0\\ r_{21}(v_1/v_2) + (r_{22} - \lambda) &= 0,\,.\end{aligned}$$

Die zweite Gleichung nach (v_1/v_2) aufgelöst ergibt:

$$\frac{v_1}{v_2} = \frac{\lambda - r_{22}}{r_{21}}. \tag{1.36}$$

Wenn wir diesen Ausdruck in die erste Gleichung einsetzen, erhalten wir eine quadratische Gleichung für λ,

$$\lambda^2 - \lambda(r_{11} + r_{22}) + r_{11}r_{22} - r_{12}r_{21} = 0\,.$$

Mit der bekannten Formel für quadratische Gleichungen finden wir die zwei Lösungen:

$$\lambda_\pm = \frac{1}{2}\left(r_{11} + r_{22} \pm \sqrt{(r_{11} - r_{22})^2 + 4r_{12}r_{21}}\right)\,. \tag{1.37}$$

λ_+ ist der Eigenwert für das Vorzeichen "+"vor der Wurzel, λ_- der Eigenwert für das Vorzeichen "−". Die Richtung der zugehörigen Eigenvektoren erhalten wir jetzt direkt aus Gleichung (1.36) wenn wir den Eigenwert λ_+ bzw. λ_- einsetzen. Es ergeben sich die folgenden Eigenvektoren $\boldsymbol{v}^+$ und $\boldsymbol{v}^-$:

$$\boldsymbol{v}^\pm = \begin{pmatrix} v_2^\pm \left(r_{11} - r_{22} \pm \sqrt{(r_{11} - r_{22})^2 + 4r_{12}r_{21}}\right)/(2r_{21}) \\ v_2^\pm \end{pmatrix}\,. \tag{1.38}$$

Der Wert $v_2^\pm$ erscheint hier als Faktor in jeder Komponente. Er ändert deshalb nur die Länge des Vektors und nicht seine Richtung. Da für einen Eigenvektor nur die Richtung festgelegt ist, kann $v_2^\pm$ beliebig gewählt werden.

[1]Wir nehmen an, dass $v_2 \neq 0$ ist und man deshalb durch v_2 teilen darf.

Beispiel Wir betrachten das Beispiel der Vogelpopulationen auf zwei benachbarten Inseln aus Abschnitt (1.3.1) mit den Reproduktionsparamtern $r_1 = 1.6$ und $r_2 = 1.2$ und Migrationsparametern $m_{2\leftarrow 1} = 0.2$ und $m_{1\leftarrow 2} = 0.3$. Die Matrix $\mathcal{R}$, die die Iterationsgleichung (1.25) bestimmt hat dann die folgende Form

$$\mathcal{R} = \begin{pmatrix} 1.6 - 0.2 & 0.3 \\ 0.2 & 1.2 - 0.3 \end{pmatrix} = \begin{pmatrix} 1.4 & 0.3 \\ 0.2 & 0.9 \end{pmatrix} . \tag{1.39}$$

Mit diesen Werten erhalten wir aus den Gleichung (1.37) und (1.38) die Eigenwerte

$$\lambda_+ = 1.5 \quad ; \quad \lambda_- = 0.8$$

mit zugehörigen Eigenvektoren

$$\boldsymbol{v}^+ = v_2^+ \begin{pmatrix} 3 \\ 1 \end{pmatrix} \quad ; \quad \boldsymbol{v}^- = v_2^- \begin{pmatrix} -0.5 \\ 1 \end{pmatrix} .$$

Um die explizite Form der Zeitentwicklung für den dynamischen Prozess zu erhalten, müssen wir nun noch den Startvektor als Summe von Eigenvektoren ausdrücken (vgl. Abb. 1.7), also

$$\boldsymbol{x}(0) = \begin{pmatrix} x_1(0) \\ x_2(0) \end{pmatrix} = v_2^+ \begin{pmatrix} 3 \\ 1 \end{pmatrix} + v_2^- \begin{pmatrix} -0.5 \\ 1 \end{pmatrix} = \begin{pmatrix} 3v_2^+ - 0.5v_2^- \\ v_2^+ + v_2^- \end{pmatrix} .$$

Dies ist wieder ein LGS. Nach v_2^+ und v_2^- aufgelöst ergibt

$$v_2^+ = \frac{2x_1(0) + x_2(0)}{7} \quad , \quad v_2^- = \frac{6x_2(0) - 2x_1(0)}{7} ,$$

und wir erhalten die gesuchte explizite Lösung für die Zeitentwicklung analog zu Gleichung (1.31):

$$\boldsymbol{x}(n) = (1.5)^n \, \frac{2x_1(0) + x_2(0)}{7} \begin{pmatrix} 3 \\ 1 \end{pmatrix} + (0.8)^n \, \frac{-2x_1(0) + 6x_2(0)}{7} \begin{pmatrix} -0.5 \\ 1 \end{pmatrix} . \tag{1.40}$$

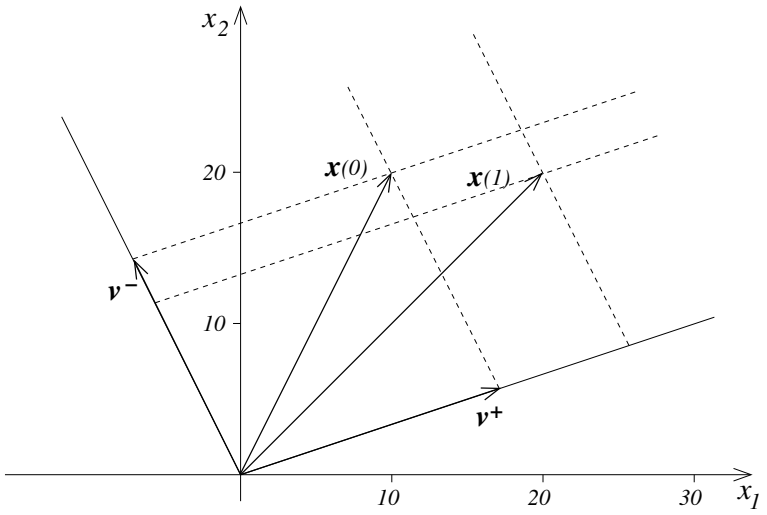


Abbildung 1.7: Eigenvektoren der Reproduktionsmatrix (1.39): Zerlegung des Startvektors (mit $x_1(0) = 10$, $x_2(0) = 20$) und erster Schritt der Populationsentwicklung.

Biologische Bedeutung von Eigenwerten und Eigenvektoren

Im letzten Abschnitt haben wir gesehen, dass eine Zerlegung des Populationsvektors in Eigenvektoren der Reproduktionsmatrix zu einer expliziten Lösung der Zeitentwicklung führt. Eigenwerte und Eigenvektoren haben aber auch eine direkte biologische Bedeutung. Dies sehen wir, wenn wir das Wachstum der Teilpopulationen und ihre relativen Größen nach langer Zeit betrachten. Für das Wachstum der Population auf der ersten Insel berechnen wir z.B.:

$$\begin{aligned}\lim_{n\to\infty}\frac{x_1(n+1)}{x_1(n)} &= \lim_{n\to\infty}\frac{(1.5)^{n+1}[6x_1(0)+3x_2(0)]+(0.8)^{n+1}[x_1(0)-3x_2(0)]}{(1.5)^n[6x_1(0)+3x_2(0)]+(0.8)^n[x_1(0)-3x_2(0)]}\\ &= \lim_{n\to\infty}\frac{1.5[6x_1(0)+3x_2(0)]+0.8(0.8/1.5)^n[x_1(0)-3x_2(0)]}{[6x_1(0)+3x_2(0)]+(0.8/1.5)^n[x_1(0)-3x_2(0)]} = 1.5\end{aligned}$$

Das Wachstum der Population auf der zweiten Insel nach vielen Generationen berechnet sich analog und mit dem selben Ergebnis.

- Allgemein gilt: Der größte Eigenwert der Reproduktionsmatrix, λ^+, ist der Faktor mit dem jede Inselpopulation – und damit auch die Gesamtpopulation – im Limes $n \to \infty$ wächst.

Für die relativen Größen der Inselpopulationen folgt:

$$\lim_{n\to\infty}\frac{x_1(n)}{x_2(n)} = \lim_{n\to\infty}\frac{3[2x_1(0)+x_2(0)](1.5)^n - 0.5[-2x_1(0)+6x_2(0)](0.8)^n}{[2x_1(0)+x_2(0)](1.5)^n + [-2x_1(0)+6x_2(0)](0.8)^n} = \frac{3}{1}$$

- Wir sehen: Die relativen Größen der Teilpopulationen stellen sich im Limes $n \to \infty$ gemäß den Komponenten des Eigenvektors zum größten Eigenwert der Reproduktionsmatrix ein. Der Anteil der beiden Teilpopulationen an der Gesamtpopulation ist dann $x_1(\infty)/(x_1(\infty)+x_2(\infty))$ bzw. $x_2(\infty)/(x_1(\infty)+x_2(\infty))$. Man nennt dies auch die *Limes- oder Grenzverteilung.*

Den Anteil eines Populationsvektors in Richtung des anderen Eigenvektors kann man als Abweichung von der Grenzverteilung auffassen. Der Quotient aus dem kleineren und dem größeren Eigenwert, λ_-/λ_+, ist dann ein Maß dafür, wie schnell das Verhältnis der Populationsgrößen gegen die Grenzverteilung konvergiert (je kleiner der Quotient, desto schneller).
Was wir hier im Beispiel gesehen haben, gilt ganz allgemein für zweidimensionale lineare Prozesse (in entsprechender Verallgemeinerung auch für höherdimensionale Prozesse). Dies ist die Konsequenz eines sehr nützlichen mathematischen Satzes. Auf unsere Situation angepasst lautet er:

Satz [Perron-Frobenius]: *Jede (2×2)-dimensionale Reproduktionsmatrix $\mathcal{R}$ mit positiven Matrixelementen $r_{ij} > 0$ hat zwei reelle, verschieden große Eigenwerte $\lambda_+ > \lambda_-$ mit zugehörigen Eigenvektoren in unterschiedlichen Richtungen. Der explizite Zeitverlauf des dynamischen Prozesses ist dann gegben durch*

$$\boldsymbol{x}(n) = \lambda_+^n \boldsymbol{v}^+ + \lambda_-^n \boldsymbol{v}^-$$

mit Eigenvektoren $\boldsymbol{v}^\pm$ zu $\lambda_\pm$ und $\boldsymbol{x}(0) = \boldsymbol{v}^+ + \boldsymbol{v}^-$. Im Limes $n \to \infty$ stellt sich der Populationsvektor $\boldsymbol{x}(n)$ (für einen beliebigen Startvektor $\boldsymbol{x}(0)$) auf die Richtung von $\boldsymbol{v}^+$ ein. Jede Teilpopulation wächst (oder schrumpft) dann geometrisch mit Faktor λ_+, die relativen Größen der Teilpopulationen bleiben konstant und verhalten sich wie die Komponenten von $\boldsymbol{v}^+$.

Die Bedeutung dieses Satzes besteht darin, dass er uns erlaubt, ein mehrdimensionales biologisches Problem auf ein eindimensionales Problem zu reduzieren. Wenn man zum Beispiel die Rolle der Vogelpopulation im Ökosystem der Inseln studieren will (z.B. in der Konkurrenz zu

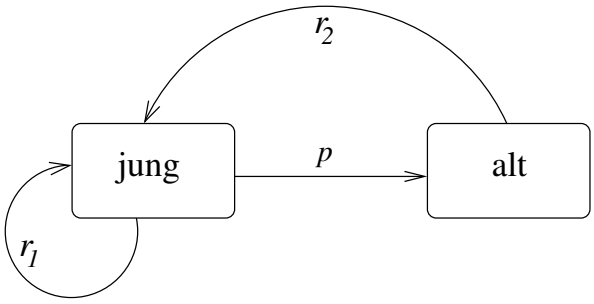


Abbildung 1.8: Modell mit zwei Altersklassen.

anderen Volgelarten) kann man oft einfach die feste Grenzverteilung annehmen und braucht sich um Probleme der Populationsstruktur nicht weiter zu kümmern. In einem übergeordneten biologischen Prozess (z.B. mit den Populationsgrößen meherer Vogelarten) betrachtet man dann nur noch die Gesamtpopulation, die, solange es genügend Nahrung gibt, mit λ_+ als Wachstumsfaktor anwächst.

1.3.4 Anwendung: Populationen mit Altersstruktur

Bislang haben wir immer angenommen, dass die Zahl der produzierten Nachkommen in einer Population unabhängig vom Alter der jeweiligen Individuen ist. Wir wollen nun die realistischere Annahme machen, dass die Zahl der Nachkommen eines Individuums von dessen Alter abhängt. Die Population wird in diskrete Altersklassen eingeteilt, der Einfachheit halber beschränken wir uns auf zwei Klassen, 'jung' und 'alt'. Die Zahl der jungen Individuen zum Zeitpunkt n sei mit $x_1(n)$ bezeichnet, die Zahl der Alten mit $x_2(n)$. Wir betrachten nun das Zusammenspiel zweier Einzelprozesse:
1. *Reproduktion:* Verschieden alte Individuen haben verschiedene Reproduktionsraten. r_1 sei die Nachkommenzahl pro Kopf von Individuen in Altersklasse 1, und entsprechend r_2 diejenige von Individuen in Altersklasse 2. Die Nachkommen sind immer 'jung', unabhängig vom Alter der Eltern, also gibt es im nächsten Jahr

$$x_1(n+1) = r_1 x_1(n) + r_2 x_2(n) \tag{1.41}$$

Individuen in Altersklasse 1.
2. *Das Älterwerden:* Wir wählen die Altersklassen und die Zeiteinheit gerade so, dass nach einem Zeitschritt ein 'junges' Individuum zu einem 'alten' Individuum wird, sofern es nicht zwischendurch gestorben ist. Ein 'altes' Individuum überlebt einen weiteren Zeitschritt in keinem Fall mehr. Wenn p die Überlebenswahrscheinlichkeit von 'jung' nach 'alt' ist, ist damit $px_1(n)$ die Anzahl der Individuen in der zweiten Altersklasse zum Zeitpunkt $n+1$:

$$x_2^{(n+1)} = p x_1^{(n)} \,. \tag{1.42}$$

Die Gleichungen (1.41) und (1.42) können in Matrixschreibweise zusammengefasst werden:

$$\boldsymbol{x}^{(n+1)} = \mathcal{L}\boldsymbol{x}^{(n)} \tag{1.43}$$

mit der sogenannten **Leslie-Matrix**

$$\mathcal{L} = \begin{pmatrix} r_1 & r_2 \\ p & 0 \end{pmatrix}. \tag{1.44}$$

Sie spielt die Rolle der Iterationsfunktion. Wir haben also ein zweidimensionales diskretes dynamisches System. Die Zeitentwicklung der Altersklassen kann man jetzt wie oben beschrieben bestimmen. Nach genügend langer Zeit stellt sich eine sogenannte *stabile Altersverteilung* ein, die sich aus dem Eigenvektor zum größten Eigenwert ergibt. Die angenehme Konsequenz dieser stabilen Verteilung ist wieder, dass man die Altersklassen einer Population in Modellen oft nicht explizit zu modellieren braucht. Man nimmt einfach an, dass sich die stabile Verteilung eingestellt hat und rechnet dann nur noch mit der Größe der Gesamtpopulation und dem größten Eigenwert der Leslie-Matrix als Reproduktionsparameter.

Beispiel: Fibonacci's Kaninchen

Leonardo da Pisa (genannt Fibonacci), 1170–1230, ist der erste bekannte Biomathematiker. Er beschäftigte sich unter anderem mit der Vermehrung von Kaninchen und entwickelte hierfür das folgende Populationsmodell.
Kaninchen werden zwei Jahre alt. Zählen wir Kaninchen in Einheiten von Paaren und nehmen wir an, dass jedes junge Paar und jedes alte Paar jedes Jahr jeweils ein neues (junges) Paar produziert und jedes junge Paar auch das nächste Jahr erlebt. Dann ist $\mathcal{L} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$. Nehmen wir an, zum Zeitpunkt 0 kommt ein junges Paar in eine zuvor nicht von Kaninchen bevölkerte Gegend, also $x_1^{(0)} = 1$, $x_2^{(0)} = 0$. Berechnen wir mal für einige Zeitpunkte die Populationsentwicklung:

$$\begin{aligned}
\boldsymbol{x}^{(0)} &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\
\boldsymbol{x}^{(1)} &= \mathcal{L}\boldsymbol{x}^{(0)} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
\boldsymbol{x}^{(2)} &= \mathcal{L}\boldsymbol{x}^{(1)} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \mathcal{L}^2\boldsymbol{x}^{(0)} \\
\boldsymbol{x}^{(3)} &= \mathcal{L}\boldsymbol{x}^{(2)} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \mathcal{L}^3\boldsymbol{x}^{(0)} \\
\boldsymbol{x}^{(4)} &= \mathcal{L}\boldsymbol{x}^{(3)} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \end{pmatrix} = \mathcal{L}^4\boldsymbol{x}^{(0)} \\
\boldsymbol{x}^{(5)} &= \mathcal{L}\boldsymbol{x}^{(4)} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 5 \\ 3 \end{pmatrix} = \begin{pmatrix} 8 \\ 5 \end{pmatrix} = \mathcal{L}^5\boldsymbol{x}^{(0)}
\end{aligned}$$

Wir fragen zuerst, wie sich die Anzahl der jungen Kaninchen $x_1(n)$ entwickelt. Wir lesen ab, dass dies gerade die Folge $1, 1, 2, 3, 5, 8, \ldots$ ist. Für die Anzahl der alten Kaninchen $x_2(n)$ ergibt sich die Folge $0, 1, 1, 2, 3, 5, \ldots$. Für die Größe der Gesamtpopulation $N(n) := x_1(n) + x_2(n)$ erhalten wir schließlich $1, 2, 3, 5, 8, 13 \ldots$. Dies ist dreimal dieselbe Folge, nur um eine bzw. zwei Folgenglieder verschoben. Sie erfüllt die Iterationsfunktion

$$x(n) = x(n-1) + x(n-2),$$

jedes Folgenglied ist also die Summe der beiden vorausgegangenen. Diese Folge hat viele bemerkenswerte mathematische Eigenschaften und taucht auch in der Natur in zahlreichen Zusammenhängen auf (z.B. beim Wachstum bestimmter Kristalle oder in der Blattstellungslehre der Botanik). Sie wurde von Leonardo anhand des Kaninchenmodells zuerst beschrieben und heißt nach ihrem Entdecker **Fibonacci-Folge**.
Fragen wir als nächstes nach dem Wachstum der Kaninchenpopulation nach langer Zeit und der stabilen Altersverteilung, die sich für $n \to \infty$ einstellt. Hierfür brauchen wir den größten Eigenwert von $\mathcal{L}$ mit zugehörigem Eigenvektor. Wenn wir die Werte der Leslie-Matrix in Gleichung

(1.37) und (1.38) einsetzen, finden wir

$$\lambda_+ = (1+\sqrt{5})/2 \quad ; \quad \boldsymbol{v}^+ = v_2^+ \begin{pmatrix} (1+\sqrt{5})/2 \\ 1 \end{pmatrix}$$

Alte und junge Kaninchen stellen sich also auf das Verhältnis $1 : (1+\sqrt{5})/2$ ein. Für den speziellen Fall von Fibonacci's Kaninchem berechnet man außerdem leicht, dass auch das Verhältnis junge Kaninchen zu Kaninchen insgesamt $1 : (1+\sqrt{5})/2$ ist. Eine Teilung in diesem Verhältnis und dieser besonderen Eigenschaft nennt man auch den *goldenen Schnitt*. Sie spielt auch in der Kunst und der Architektur eine große Rolle.

1.4 Kontinuierliche Entwicklungsprozesse

Bislang haben wir Prozesse in diskreter Zeit untersucht, d.h. wir haben angenommen, dass nur zu festen Zeitpunkten etwas passiert oder dass wir nur in regelmäßigen Abständen messen. Eine solche Annahme ist jedoch nicht immer sinnvoll; oft ist es realistischer, den Zuwachs kontinuierlich in der Zeit zu modellieren. Wir werden jetzt also eine kontinuierliche Zeitvariable t betrachten, (sie durchläuft die positiven reellen Zahlen), und $x(t)$ bezeichnet die Populationsgröße (oder eine andere Größe, die uns interessiert) als Funktion der Zeit. Diese Änderung der Betrachtungsweise führt zur Theorie der gewöhnlichen Differentialgleichungen, in die wir im folgenden einen kurzen Einblick geben werden.

1.4.1 Differentialgleichungen

Betrachten wir eine Bakterienkultur, in der sich jede Stunde 80% der Zellen (synchron) teilen; eine zweite, bei der sich alle 30 min. 40% der Zellen teilen; und eine dritte, bei der sich alle 15 min. 20 % der Zellen teilen. Es handelt sich jeweils um geometrisches Wachstum in diskreter Zeit, aber mit unterschiedlichen Zeitschritten Δt und unterschiedlichem Wachstumsfaktor r. Die Iterationsfunktionen für diese drei Situationen lauten

$$\begin{aligned} x_{n+1} &= x_n + 0.8\, x_n, \quad \Delta t = 1h \\ x_{n+1} &= x_n + 0.4\, x_n, \quad \Delta t = 1/2h \\ \text{und} \quad x_{n+1} &= x_n + 0.2\, x_n, \quad \Delta t = 1/4h\,. \end{aligned} \tag{1.45}$$

Der Zeitverlauf für die ersten 2.5 Stunden ist in Abbildung (1.9) dargestellt. Man sieht, dass die Kurve immer glatter wird, je kleiner das Zeitintervall gewählt wird. Man beobachtet aber auch einen 'Zinseszinseffekt': Je kürzer das Zeitintervall, desto schneller das Wachstum insgesamt, da die neuen Individuen sich ihrerseits schneller wieder teilen und so zum Wachstum beitragen.

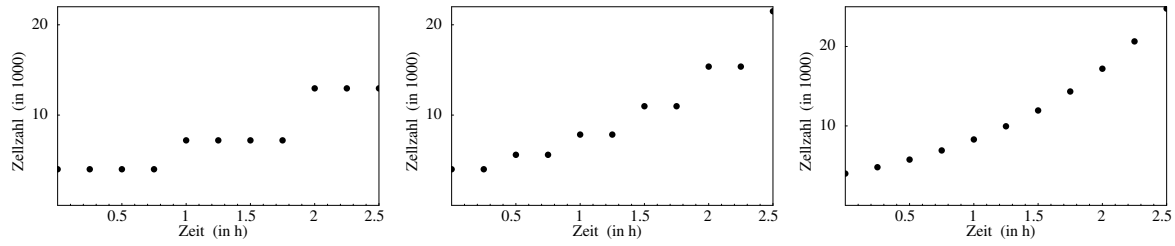


Abbildung 1.9: Populationsentwicklung einer Bakterienkultur, die mit $x_0 = 4000$ Bakterien startet. Links: Jede volle Stunde teilen sich 80% der Zellen sychron. Mitte: Alle 30 min teilen sich 40% der Zellen. Rechts: Alle 15 min teilen sich 20% der Zellen.

Was passiert, wenn wir diesen Prozess fortsetzen? Um das herauszufinden, nehmen wir nun als Zeitvariable $t = n\Delta t$, definieren $x_n = x(n\Delta t) = x(t)$ als Funktion der Zeit, und schreiben, in Verallgemeinerung von (1.45), für beliebig kleine Zeitschritte Δt:

$$x(t + \Delta t) = x(t) + 0.8 \cdot \Delta t \cdot x(t). \tag{1.46}$$

Wenn man das Zeitintervall immer weiter verkürzt, erhält man immer glattere Kurven, die sich immer weniger voneinander unterscheiden (Abb. 1.10). Um diesen Grenzprozess mathematisch zu vollziehen, formen wir (1.46) um in

$$\frac{x(t + \Delta t) - x(t)}{\Delta t} = 0.8 \cdot x(t)\,. \tag{1.47}$$

Der Ausdruck auf der linken Seite ist ein *Differenzenquotient*. Wenn wir nun Δt gegen Null gehen lassen, wird aus dem Differenzenquotienten die Ableitung der Funktion x:

$$\lim_{\Delta t \to 0} \left(\frac{x(t+\Delta t) - x(t)}{\Delta t} \right) = \frac{\mathrm{d}x}{\mathrm{d}t} = 0.8\, x(t) \tag{1.48}$$

oder allgemeiner

$$\dot{x}(t) := \frac{\mathrm{d}x}{\mathrm{d}t} = \lambda\, x(t)\,. \tag{1.49}$$

Die Zeit t ist nun eine *kontinuierliche* Variable. Wir benutzen das Symbol $\dot{x}(t) = \frac{\mathrm{d}}{\mathrm{d}t}x(t)$ als Symbol für die Ableitung nach der Zeit. λ ist die (*instantane)* Wachstumsrate (pro Kopf). Sie hat die Einheit 1/Zeit. Man beachte den Unterschied zwischen r (geometrisches Wachstum) und λ: Während in r alle Nachkommen eingehen (Überlebende eingerechnet), misst λ nur den *Zuwachs* (man könnte auch von einer *Zuwachsrate sprechen*). Eine Population konstanter Größe ist also durch $r = 1$, aber $\lambda = 0$ charakterisiert.

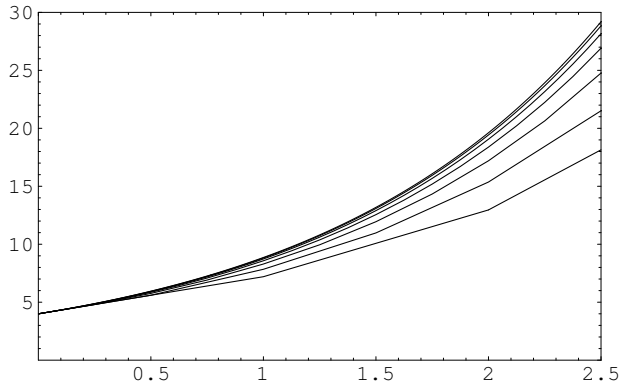


Abbildung 1.10: Populationsentwicklung bei weiterer Unterteilung des Zeitintervalls. Von unten nach oben: $\Delta t = 1,\ 1/2,\ 1/4,\ 1/8,\ 1/16,\ 1/32,\ 1/64$, mit Zuwächsen $r - 1 = 0.8,\ 1/2 \cdot 0.8,\ 1/4 \cdot 0.8, \ldots,\ 1/64 \cdot 0.8$. Der Übersichtlichkeit halber sind die Punkte diesmal durch Linien verbunden. Man beobachtet Konvergenz gegen eine glatte Funktion.

Gleichungen der Form

$$\dot{x}(t) = g\big(x(t)\big) \quad \text{oder kürzer} \quad \dot{x} = g(x) \tag{1.50}$$

bezeichnet man als **(gewöhnliche) Differentialgleichungen**; $g(x)$ heißt **Geschwindigkeitsfunktion** oder **rechte Seite**. Sie gibt an, wie schnell sich die Populationsgröße ändert, wenn sie gerade den Wert x hat. Somit ist eine Differentialgleichung das kontinuierliche Gegenstück zur Iteration in diskreter Zeit. Wird zusätzlich der Wert von x zum Zeitpunkt 0 vorgegeben, also $x(0) = x_0$, so spricht man von einem **Anfangswertproblem** (oder AWP, das 'Problem' besteht darin bei gegebener Differtialgleichung und Anfangswert den Zeitverlauf $x(t)$ zu rekonstruieren). Differentialgleichungen spielen in allen Naturwissenschaften eine wichtige Rolle. Die meisten Naturgesetze sind in Form von Differentialgleichungen formuliert. Sie machen Aussagen über die Änderung eines Systems in Abhängigkeit von seinem augenblicklichen Zustand. Wir werden hier nur den allereinfachsten Fall behandeln: sogenannte *gewöhnliche Differentialgleichungen erster Ordnung*. Das heißt, es kommt nur eine einfache Ableitung nach einer Variablen (der Zeit t) vor. Bei Differentialgleichungen höherer Ordnung kommen auch zweite oder höhere Ableitungen vor. Ein Beispiel sind die Bewegungsgleichungen der Physik, wie etwa die *Schwingungsgleichung*,

$$\ddot{x} = -(D/m)x\,,$$

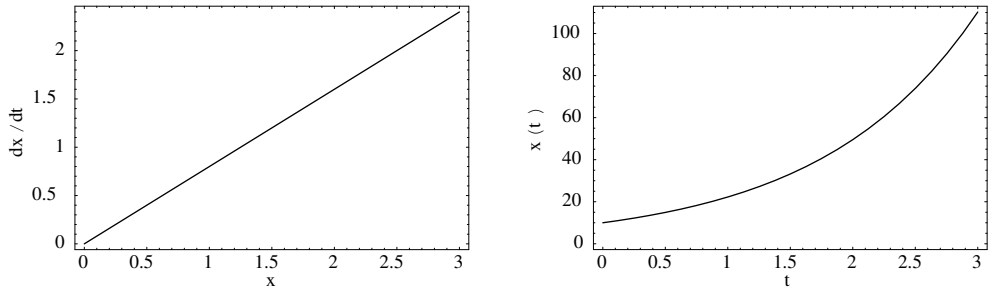


Abbildung 1.11: Die lineare Differentialgleichung $\dot{x} = \lambda x$ (links) und ihre Lösung, $x(t) = x_0 e^{\lambda t}$, für $\lambda = 0.8$ und $x_0 = 10$ (rechts).

wo die Beschleunigung eines Körpers, also die zweite Ableitung nach der Zeit $\ddot{x} := d^2x/dt^2$ als Funktion des Ortes x angegeben wird (D ist die Federkonstante und m die Masse). Die meisten Gesetze sind zudem noch sogenannte *partielle Differentialgleichungen*, das heißt es tauchen Ableitungen nach mehereren Variablen (z.B. Ort und Zeit) auf. Beispiele sind die Schrödingergleichung, die Maxwell-Gleichungen und die Einstein'sche Feldgleichungen aus der Physik, oder auch die Diffusionsgleichung, die in der Biologie viele wichtige Anwendungen hat (im nächsten Abschnitt behandeln wir einen einfachen Spezialfall in dem nur eine gewöhnliche DGL auftritt).

1.4.2 Modellbildung IV: Exponentielles Wachstum

In diskreter Zeit war der lineare Prozess in einer Dimension, also das geometrische Wachstum, der Ausgangspunkt für alle unsere Überlegungen gewesen. Oben haben wir genau für diesen Prozess den Übergang zu kontinuierlicher Zeit durchgeführt und sind so auf Gleichung (1.49) gekommen. In dieser DGL ist die Geschwindigkeitsfunktion linear in x, man nennt sie deshalb die lineare DGL in einer Dimension (genauer: homogen-linear, da die Gerade $g(x) = \lambda x$ durch den Ursprung geht). Sie spielt für die dynamischen Prozesse in kontinuierlicher Zeit die gleiche fundamentale Rolle wie das geometrische Wachstum in diskreter Zeit.
Eine DGL lösen bedeutet, eine Funktion $x(t)$ zu finden, deren Ableitung nach t die geforderte Beziehung (1.50) erfüllt. Wie in diskreter Zeit ist das meistens nicht explizit möglich, aber die lineare DGL ist eine Ausnahme. In diesem Fall können wir die Lösung des Anfangswertproblems raten:

$$x(t) = x_0 e^{\lambda t} . \tag{1.51}$$

Durch Differenzieren ergibt sich sofort, dass $\dot{x}(t) = \lambda e^{\lambda t} x_0 = \lambda x(t)$ und $x(0) = x_0$; Gleichung (1.51) ist also tatsächlich die gesuchte Lösung. Gleichung (1.51) beschreibt das sogenannte **exponentielle Wachstum**. Dieses Modell wurde zuerst von MALTHUS 1798 auf das Wachstum menschlicher Populationen angewandt (und beschreibt die Bevölkerungsexplosion erschreckend gut!). Die instantane Wachstumsrate λ wird deshalb manchmal auch als *Malthus-Parameter* bezeichnet.
Für ein biologisches System gewinnt man das Modell für exponentielles Wachstum unter genau den gleichen Modellannahmen wie das geometrische Wachstum im diskreten Fall (außer, dass man eben kontinuierliche Zeit annimmt). Also: Die Populationsgröße ändert sich nur durch Geburten und Tode und die Zahl der Geburten und Tode ist proportional zur momentanen Populationsgröße. Mit einer Geburtsrate β und Todesrate δ (jeweils pro Kopf) ergibt sich dann Gleichung (1.49) mit $\lambda = \beta - \delta$.

Die enge Beziehung zum geometrischen Wachstum sieht man auch aus Gleichung (1.51) wenn man die Konstante $\exp(\lambda)$ als Wachstumsparameter r wählt. Dann ist

$$x(t) = x_0 e^{\lambda t} = (\exp(\lambda))^t x_0 = r^t x_0 \,.$$

Wenn man nun den in kontinuierlicher Zeit ablaufenden Prozess in diskreten Zeitschritten beschreibt (weil man etwa einmal in der Stunde misst), erhält man wieder ein geometrisches Wachstum. In diesem Sinne sind exponentielles und geometrisches Wachstum "ein und dasselbe". Im Unterschied zur obigen Ableitung des kontinuierlichen Prozesses aus dem diskreten ändern wir hier nichts am Prozess selbst: Geburten und Tode ereignen sich kontinuierlich, auch wenn wir nur in diskreten Zeitschritten messen. Deshalb muss im Wachstumsparameter r auch den Zinseszins-Effekt mit berücksichtigen. Dies ist auch der Fall. Im Vergleich zu oben ist der Wachtumsfaktor $r = \exp(\lambda)$ immer etwas größer als $1 + \lambda$.

Diffusionsprozesse

Da Molekülbewegungen nicht an diskrete Zeitpunkte gebunden sind, wird ein Diffusionsprozess realistischerweise als kontinuierlich in der Zeit beschrieben. Man denke sich also eine Zelle in einem Medium, das einen Farbstoff enthält, der über die Zellmembran diffundieren kann. Die Farbstoffkonzentration im Medium ist c und wird durch den Austausch mit der Zelle nicht nennenswert verändert. Der Farbstoff diffundiert mit (instantaner) Rate $\alpha > 0$ in beiden Richtungen über die Zellmembran. Wir nehmen an, dass die Diffusion proportional zur Farbstoffkonzentration im Medium bzw. in der Zelle ist. Die Konzentration in der Zelle, $x(t)$, folgt dann der Differentialgleichung

$$\dot{x}(t) = \alpha c - \alpha x = \alpha(c - x); \tag{1.52}$$

die Änderung von x ist also proportional zur *Konzentrationsdifferenz*. Gleichung (1.52) ist eine typische *Bilanzgleichung* der Form 'Netto-Zunahme=Influx minus Efflux'. Sie beschreibt den einfachsten Fall einer Diffusion mit nur zwei diskreten Werten für die Konzentration: innerhalb und außerhalb der Zelle. Im allgemeineren Fall ist die Konzentration selbst eine kontinuierliche Funktion des Ortes. Dann wird die Diffusion zu einer partiellen Differentialgleichung und durch die Fick'schen Gesetze beschrieben.
Die Gleichung (1.52) ist immer noch eine lineare Differentialgleichung, allerdings nicht mehr homogen-linear wie beim exponentiellen Wachstum (die Gerade $g(x) = \alpha(c - x)$ geht nicht durch den Ursprung). Man kann eine solche DGL immer noch vollständig und exakt lösen (siehe Übungen). Wir wollen das hier aber nicht tun, sondern nur ihr qualitatives Verhalten diskutieren.

1.4.3 Modellanalyse IV: Phasenliniendiagramm

Analog zum Cobwebbing im diskreten Fall führen wir eine graphische Methode ein, die es uns erlaubt, auch für komplizierte Differentialgleichungen die Zeitentwicklung der Populationsgröße qualitativ zu analysieren. Diese Methode ist das **Phasenliniendiagramms**, wir werden sie am Beispiel des Diffusionsmodells besprechen (siehe Abbildung 1.12).

1. Als erstes zeichnen wir den Graphen der Funktion $g(x)$.

2. Wir denken uns nun $x(t)$ als Bewegung auf der x-Achse, deren Geschwindigkeit $\dot{x}$ an jedem Punkt x durch den Wert $g(x)$ gegeben ist.

3. Wir deuten die Entwicklung der dynamischen Größe in jedem Punkt x durch Pfeile auf der x-Achse an: Wenn $\dot{x} = g(x)$ positiv (negativ) ist, wird x größer (kleiner), also zeichnet

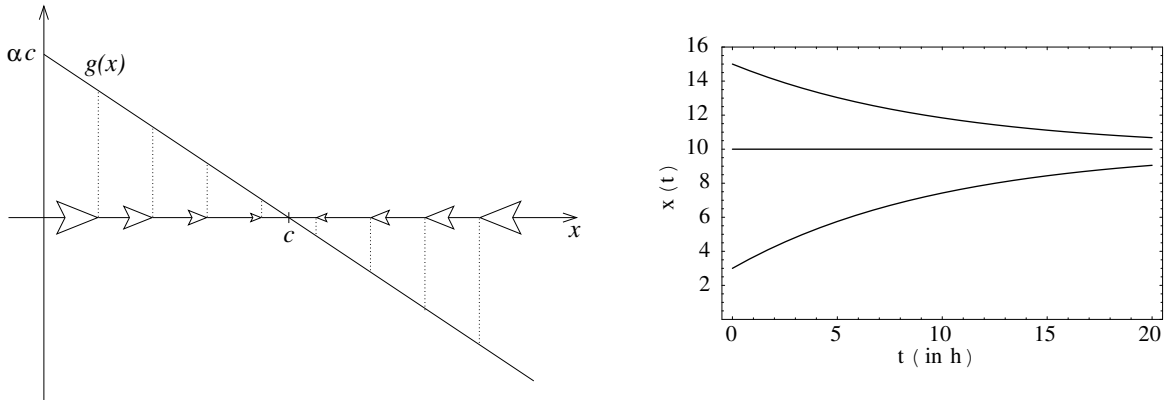


Abbildung 1.12: Stoffaustausch gemäß Gleichung (1.52). Links: Phasenliniendiagramm. Rechts: Die zugehörige Zeitentwicklung $x(t)$ für $c = 10$, $\alpha = 0.1/h$ und zwei verschiedene Anfangswerte $x_0 = 3$ und $x_0 = 15$.

man einen Pfeil nach rechts (links). Die Geschwindigkeit ist umso größer, je größer der Betrag von $g(x)$ ist. Dies deutet man durch verschieden große Pfeile an. Die Bewegung von $x(t)$ folgt dann diesen Pfeilen; daraus läßt sich qualitativ der Zeitverlauf ablesen, auch wenn man die Differentialgleichung nicht explizit lösen kann.

Für das Diffusionsmodell ändert sich x so lange, bis $x = c$ ist, das heißt bis der Konzentrationsausgleich erreicht ist.

1.4.4 Fixpunkte und Stabilität

Wie im diskreten Fall interessieren wir uns für Werte, auf die sich die Größe $x(t)$ nach langer Zeit einstellt. Offensichtlich ändert sich $x(t)$ gerade an den Punkten nicht mehr, an denen die Geschwindigkeitsfunktion $g(x) = 0$ ist. Allgemein nennt man Lösungen x^* der Gleichung $g(x) = 0$ Fixpunkte oder **Gleichgewichtspunkte** der Differentialgleichung $\dot{x} = g(x)$. Geometrisch sind die Gleichgewichtspunkte die Schnittpunkte von $g(x)$ mit der x-Achse.

Die Stabilität der Gleichgewichtspunkte kann man ebenfalls direkt aus dem Phasenliniendiagramm ablesen: Ein Gleichgewichtspunkt x^* ist **lokal stabil** oder **anziehend**, wenn die Pfeile in beiden Richtungen auf ihn zulaufen. Mathematisch ist das offensichtlich genau dann der Fall, wenn die Ableitung $g'(x) = \frac{\mathrm{d}}{\mathrm{d}x} g(x) < 0$ ist. Ist dagegen $g'(x^*) > 0$, so laufen die Pfeile nach links und rechts vom Fixpunkt weg und das Gleichgewicht heißt **instabil** oder **abstoßend**. (Der Fall $g'(x^*) = 0$ spielt selten eine Rolle; wir sparen ihn hier aus.) Im Beispiel des Diffusionsmodells ist $x^* = c$ und $g'(x^*) = -\alpha < 0$, der Gleichgewichtspunkt ist also anziehend, in Übereinstimmung mit Abb. 1.12. **Man beachte:** $\dot{x}(t)$ ist die Ableitung von x *nach der Zeit* t, während $g'(x)$ die Ableitung von g *nach der dynamischen Größe* x ist !!!

Die folgende Tabelle gibt über die Zusammenhänge und Unterschiede zwischen diskreten dynamischen Systemen und Differentialgleichungen Auskuft (der Einfachheit halber nehmen wir wieder die Größe einer Population als die interessierende dynamische Größe).

	diskrete Zeit (n)	kontinuierliche Zeit (t)
dynamisches System	diskrete Iteration	Differentialgleichung
	$x_{n+1} = f(x_n)$	$\dot{x}(t) = g(x)$
charakterisiert durch	Iterationsfunktion $f(x)$	Geschwindigkeitsfunktion $g(x)$
gibt an	Populationsgröße in der nächsten Generation	Zuwachsrate der Population
graphischer Ansatz	Cobwebbing	Phasenliniendiagramm
Gleichgewicht	$f(x^*) = x^*$	$g(x^*) = 0$
Schnittpunkt mit	Winkelhalbierender	x-Achse
lokal stabil wenn	$\lvert f'(x^*)\rvert < 1$	$g'(x^*) < 0$

1.5 Nicht-lineare Prozesse in kontinuierlicher Zeit

Analog zum diskreten Fall kann man auch mit Differentialgleichungen Wachstumsprozesse bei knappen Ressourcen beschreiben. Man wird hierbei automatisch wieder auf nicht-lineare Prozesse geführt, d.h. die Geschwindigkeitsfunktion ist keine Gerade mehr, sondern eine allgemeine Funktion der Populationsgröße x. Im einfachsten nicht-linearen Fall mit einer quadratischen Geschwindigkeitsfunktion erhalten wir eines der wichtigsten Wachstumsmodelle der Biologie, das sogenannte logistische Wachstum.

1.5.1 Modellbildung V: Logistisches Wachstum

Das kontinuierliche logistische Wachstum ist ein naher Verwandter des diskreten Verhulst-Modells (und wird manchmal auch als kontinuierliches Verhulst-Modell bezeichnet). Die Populationsgröße folgt dabei der logistischen Differentialgleichung

$$\dot{x} = g(x) = \frac{\lambda}{K}\, x(K - x) \;=\; \lambda x(1 - x/K)\,, \quad \lambda, K > 0\,. \tag{1.53}$$

λ hat wieder die Bedeutung einer instantanen Wachstumsrate; wenn x sehr viel kleiner ist als K, dann wächst die Population näherungsweise exponentiell nach $\dot{x} = \lambda x$. K ist die maximale Populationsgröße, die das Habitat ernähren kann und wird wie im diskreten Fall als Kapazität oder *carrying capacity* bezeichnet. Die Modellannahme des logistischen Wachstums ist, dass der Zuwachs $\dot{x}$ der Population einerseits proportional zur Zahl der bereits vorhandenen Individuen x und andererseits proportional zur noch nicht ausgenutzten Kapazität (zur noch vorhandenen Nahrung) $K - x$ ist.
Für die Fixpunkte berechnet man: $g(x) = 0$ bei $x_1^* = 0$ und bei $x_2^* = K$. Da $g'(0) = \lambda > 0$ ist, ist der Fixpunkt x_1^* immer instabil; der Fixpunkt x_2^* mit $g'(K) = -\lambda < 0$ ist dagegen stabil. Die gleiche Information kann man auch aus dem Phasenliniendiagramm erhalten (siehe Abb. logifig). Als Zeitverlauf erhält man wie beim diskreten Verhulst-Modell eine S-förmige ("sigmoide") Kurve. Nach genügend langer Zeit wird beim logistischen Wachstum die *carrying capacity* stets voll ausgeschöpft. Man beachte den Unterschied zum diskreten Verhulst-Modell: dort ist die Kapazität eine harte Obergrenze für die Zahl der Nachkommen in *jeder* Generation. Im Gleichgewicht wird sie dort von der Population nicht vollständig ausgeschöpft.

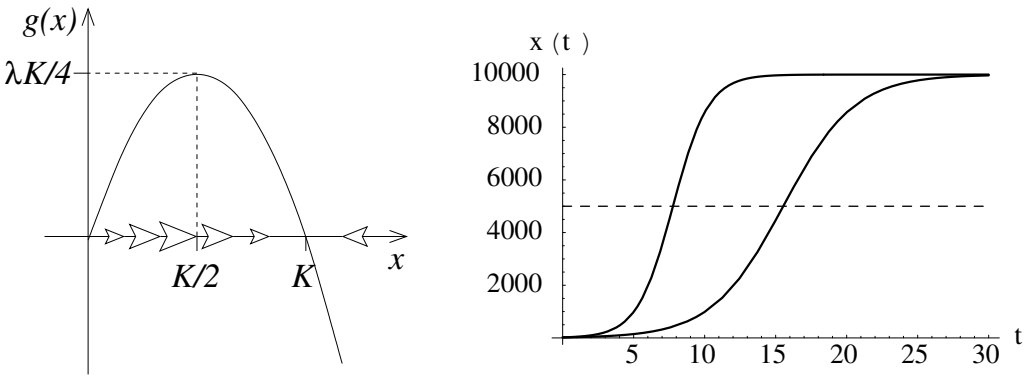


Abbildung 1.13: Phasenliniendiagramm und Zeitverlauf für das logistische Wachstum. Für den Zeitverlauf wurde $K = 10000$ gewählt und zwei verschiedene Werte für λ, 0.4 und 0.8. Für größeres λ wächst die Population anfänglich schneller. Bei $K/2$ (gestrichelte Linie bei 5000) ist das Wachstum maximal: die Kurve $x(t)$ hat hier einen Wendpunkt.

Das logistische Wachstum taucht in der Biologie in vielen Zusammenhängen immer wieder auf. Das Beispiel des Fließgleichgewichts im Chemostat wird in den Übungen besprochen. Im nächsten Abschnitt besprechen wir eine weiteres Modell mit einer interessanten biologischen Anwendung, das letztlich wieder auf die logische Gleichung führt.

1.5.2 Anwendung: Ein epidemiologisches Modell

Im Jahr 1897 entdeckte der englische Arzt Sir Ronald Ross (1857 – 1932, Nobelpreis 1902), dass Malaria von der Anopheles-Mücke übertragen wird. Um die Krankheit auszurotten, schlug er deshalb die Bekämpfung der Mücken vor. Da es aber in vielen Gegenden praktisch unmöglich ist, die Mücke total auszurotten, standen viele Leute solchen Programmen skeptisch gegenüber: der Erreger (Plasmodium) würde in einigen Mücken versteckt die Aktion überdauern und nach Beendigung des Programms (mit hohen Kosten) mit den Mücken rasch wieder zurückkehren. Sir Ronald entwickelte daraufhin ein mathematisches Modell, um seine Zeitgenossen davon zu überzeugen, dass dies ein Fehlschluss ist. Wir besprechen dieses Modell in einer etwas vereinfachten Form.
Im Modell kann Malaria von einem Infizierten mit der Rate α auf einen Nichtinfizierten übertragen werden. Da die Übertragung über die Anopheles-Mücke erfolgt, ist die Übertragungsrate eine Funktion der Mückenzahl, je mehr Mücken, desto größer α. Im Modell wollen wir zunächst die Ausbreitung der Krankheit für ein konstantes α (konstante Mückenzahl) beschreiben. Infizierte können außerdem mit einer Rate μ genesen (wir nehmen vereinfachend an, dass sie weder sterben noch immun werden).

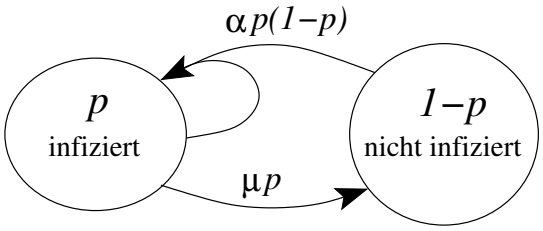


Abbildung 1.14: Ein Modell für eine ansteckende Krankheit.

Sei $p(t)$ der Anteil der Infizierten in einer Population; dann ist $1 - p(t)$ der Anteil der Nichtinfizierten. Da Neuinfektionen das Vorhandensein von Infizierten und Nichtinfizierten voraussetzt, ist der Zuwachs an Infizierten einerseits proportional zu p, andererseits zu $1 - p$, mit Proportionalitätskonstante α. Die Genesung ist dagegen unabhängig von der Zahl der Gesunden, und deshalb nur proportional zu p. Insgesamt ändert sich p dann gemäß

$$\dot{p} = \alpha p(1 - p) - \mu p\,. \tag{1.54}$$

Die Beziehung zum logistischen Wachstum wird klar, wenn man die Gleichung umschreibt:

$$\dot{p} = \frac{\alpha - \mu}{(\alpha - \mu)/\alpha}\left([(\alpha - \mu)/\alpha] - p\right) \tag{1.55}$$

Wenn man nun $\alpha - \mu = \lambda$ setzt und $(\alpha - \mu)/\alpha = K$ bekommt man genau wieder das logistische Wachstum aus Gleichung (1.53). Allerdings kann es hier vorkommen, dass $\mu > \alpha$ ist und wir somit eine *negative* Kapazität K bekämen. Man muss deshalb zwei Fälle, $\alpha > \mu$ und $\alpha < \mu$ unterscheiden. Die Phasenliniendiagramme für beide Fälle sind in Abb. 1.15 gezeigt. Die möglichen Gleichgewichtspunkte sind $p_1^* = 0$ und $p_2^* = 1 - \mu/\alpha$. Ist $\alpha > \mu$, so ist $p_1^* = 0$ instabil

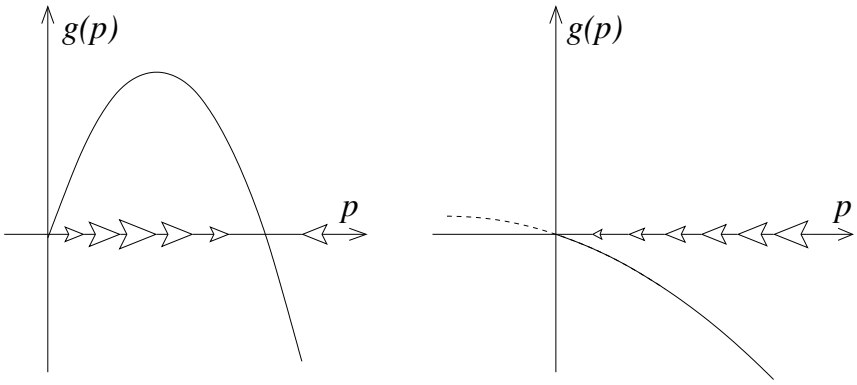


Abbildung 1.15: Phasenliniendiagramm für das Ansteckungsmodell (1.54). Links: $\alpha < \mu$, rechts: $\alpha > \mu$.

und $p_2^* > 0$ und stabil. Es stellt sich also ein Zustand ein, bei dem stets ein Bruchteil $1 - \mu/\alpha$ der Population infiziert ist. Man spricht von einem *endemischen Gleichgewicht*. Ist dagegen $\alpha < \mu$, so ist $p_2^* < 0$. Somit ist $p_1^* = 0$ das einzig relevante Gleichgewicht und stabil; die Krankheit stirbt aus. Insbesondere bedeutet dies, dass man die Übertragungsrate α nicht auf 0 drücken muss, um die Krankheit zu besiegen. α muss nur kleiner als die Genesungsrate μ sein, dann rottet sie sich selbst aus. Für Malaria heißt dies, man muss nicht alle Mücken ausrotten, sondern "nur" die Mückenzahl solange genügend klein halten bis der Erreger verschwunden ist.
Ronald Ross hat seine mathematischen Modelle stets für seinen bedeutendsten wissenschaftlichen Beitrag gehalten (!) – er wurde aber von kaum einem seiner Zeitgenossen verstanden. Erst zwei Jahrzehnte nach seinem Tod wurden mathematische Modelle in der Epidemiologie aufgegriffen und haben dort heute einen festen Platz. Einfache Modelle dienen vor allem dem qualitativen Verständnis der Ausbreitung von Krankheiten. In unserem Beispiel etwa: Abhängig von der Höhe der Übertragungsrate kann sich bei einer Krankheit ein epidemisches Gleichgewicht einstellen (wo immer ein Teil der Bevölkerung krank ist), oder die Krankheit kann sich nicht dauerhaft festsetzen und verschwindet von selbst wieder. Für Malaria sieht man das heute z.B. in vielen Gebieten der USA, wo es Anopheles-Mücken in geringerer Dichte gibt, aber keine Malaria. Für quantitative Vorhersagen, gerade in den komplexen Ökosystemen der tropischen Länder, erwiesen sich einfache Modelle, die man analytisch auswerten kann als nicht ausreichend. Um dort Strategien zur Bekämpfung von Krankheiten zu testen, verwendet man heute sehr viel komplexere Modelle, die mit Computern ausgewertet werden.

1.6 Mehrdimensionale Prozesse: Gekoppelte Differentialgleichungen

Bislang haben wir Differentialgleichungen für eine einzige dynamische Variable (z.B. eine Population) untersucht. Analog zu den diskreten Prozessen wollen wir nun Prozesse in kontinuierlicher Zeit mit mehreren zeitveränderliche Größen untersuchen, zum Beispiel zwei Populationen, die miteinander in Wechselwirkung treten. Wir werden besonders eine offensichtliche Wechselwirkung von einiger biologischer Bedeutung diskutieren, nämlich das Fressen und Gefressenwerden.

1.6.1 Modellbildung VI: Räuber-Beute Prozesse

Wir betrachten die gemeinsame Zeitentwicklung einer Räuber- und einer Beutepopulation (z.B. Bakterien und Amöben, oder Hasen und Füchse). Die Populationsgrößen zur Zeit t seien $x(t)$ (Beute) bzw. $y(t)$ (Räuber). Wie modelliert man ein solches System?
Wenn wir das System mit Differentialgleichungen beschreiben wollen, dann müssen wir den (positiven oder negativen) Zuwachs in den Populationsgrößen, also $\dot{x}$ und $\dot{y}$, in Abhängigkeit von den gegenwärtigen Populationsgrößen (x und y) ausdrücken. Wir suchen also ein System der Form

$$\dot{x} = g(x,y) \tag{1.56}$$
$$\dot{y} = h(x,y)\,. \tag{1.57}$$

Die Geschwindigkeitsfunktionen $g(x,y)$ und $h(x,y)$ geben an, wie sich die Beute- bzw. Räuberpopulation ändert, wenn gerade x Beutetiere und y Räuber da sind. Im Unterschied zu unseren bisherigen, eindimensionalen Modellen hängt der Zuwachs der Beutepopulation x nicht allein von x selbst ab, sondern auch von der Größe der Räuberpopulation y – und umgekehrt. Deshalb sind g und h Funktionen von zwei Variablen, d.h., sie ordnen jedem (x,y)-Zahlenpaar einen Wert $g(x,y)$ bzw. $h(x,y)$ zu. Ein System von Gleichungen der Form (1.56) nennt man ein **System gekoppelter Differentialgleichungen**. Sind zusätzlich Anfangswerte ($x(0) = x_0$, $y(0) = y_0$) vorgegeben, so hat man das zugehörige Anfangswertproblem. Wir werden nun versuchen, biologisch sinnvolle Funktionen g und h zu finden. In einem ersten Schritt betrachten wir jede Population erst einmal in Abwesenheit der anderen, suchen also $g(x,0)$ und $h(0,y)$.
Solange es genügend Raum und Nahrung gibt, wächst die Beutepopulation exponentiell. Mit zunehmender Größe der Population machen die Tiere sich aber gegenseitig Konkurrenz (um Nahrung, Nistplätze,...). Ein Konkurrenzsituation entsteht (mit einer bestimmten Wahrscheinlichkeit) immer dann, wenn zwei Tiere aufeinandertreffen. Die Häufigkeit solcher Treffen nimmt mit der Zahl der möglichen Konkurrenten-Paare in der Population zu. Man modelliert deshalb den Einfluss von Konkurrenz innerhalb einer Art als proportional zum Quadrat der Populationsgröße. Die führt zu folgender Differentialgleichung:

$$g(x,0) = \lambda_x x - \gamma_x x^2, \qquad \text{wobei} \qquad \lambda_x, \gamma_x > 0\,. \tag{1.58}$$

λ_x ist die (instantane) Wachstumsrate der Beute, und γ_x bestimmt die Stärke der innerartlichen Konkurrenz. Wenn wir (1.58) mit Gleichung (1.53) vergleichen sehen wir, dass wir gerade wieder das logistische Wachstum haben; beide Gleichungen stimmen überein, wenn wir $\lambda_x := \lambda$ und $\gamma_x := \frac{\lambda}{K}$ setzen. In der Argumentation, die zu Gleichung (1.53) führte, war die beschränkte Kapazität K der Ausgangspunkt der Überlegungen, hier die Häufigkeit von Konkurrenzsituationen. Eine beschränkte Kapazität $K = \lambda_x/\gamma_x$ ergibt sich als Konsequenz der steigenden Konkurrenz. Die Tatsache, dass man das logistische Wachstum aus unterschiedlichen Argumentationen ableiten kann ist ein Grund dafür, dass es als Modell sehr häufig verwendet wird.

Mit den Räubern verhält es sich ähnlich, allerdings sterben sie aus, wenn keine Beute da ist. Sie schrumpfen also exponentiell mit einer Sterberate λ_y. Darüber hinaus machen sie sich auch noch gegenseitig Konkurrenz (zum Beispiel weil sie Agressionen gegeneinander entwickeln). Wir erhalten die folgende Differentialgleichung,

$$h(0,y) = -\lambda_y y - \gamma_y y^2, \qquad \text{wobei} \qquad \lambda_y, \gamma_y > 0. \tag{1.59}$$

Um nun die Interaktion der beiden Populationen zu beschreiben, nehmen wir an, dass die Zahl der gefressenen Beute einerseits proportional zur Zahl der Beutetiere, andererseits proportional zur Zahl der Räuber ist. Die Beutepopulation erleidet also Verluste um $-p\,xy$ ($p > 0$). Für die Räuberpopulation ist die Möglichkeit, Beute zu machen, natürlich von Vorteil – sie kann dadurch überhaupt erst wachsen, und zwar um $q\,xy$ ($q > 0$). Wenn wir die Gleichungen (1.58) und (1.59) um die entsprechenden Terme ergänzen, erhalten wir das folgende System gekoppelter Differentialgleichungen:

$$\begin{array}{rclcll} \dot{x} & = & g(x,y) & = & \lambda_x x - \gamma_x x^2 - pxy & = & x(\lambda_x - \gamma_x x - py) \\ \dot{y} & = & h(x,y) & = & -\lambda_y y - \gamma_y y^2 + qxy & = & y(-\lambda_y - \gamma_y y + qx)\,. \end{array} \tag{1.60}$$

Wir sehen, dass diese Gleichungen nach einem ganz allgemeinen Prinzip gebildet werden: Terme proportional zu x oder y geben an, wie sich die Populationen entwickeln solange alle Individuen unabhängig von den anderen sind. Terme proportional zu Produkten von Populationsgrößen (x^2, xy und y^2) geben Auskunft darüber was passiert, wenn Tiere der entsprechenden Populationen sich treffen (Konkurrenz, fressen und gefressen werden) und was dies für eine Auswirkung auf den Zuwachs von x oder y hat. Die Parameter λ_x, γ_y, etc. bestimmen wie groß dieser Effekt ist. Für die Parameter in (1.60) werden wir im folgenden häufig die Werte

$$\lambda_x = 1,\ \lambda_y = 0.05,\ \gamma_x = 10^{-4},\ \gamma_y = 10^{-5}\ p = 10^{-4},\ q = 10^{-5} \tag{1.61}$$

verwenden.

Beispiel: Für $(x,y) = (6500, 3000)$ ist $\dot{x} = g(x,y) = 6500 - 10^{-4} \cdot 6500^2 - 10^{-4} \cdot 6500 \cdot 3000 = 325$ und $\dot{y} = h(x,y) = -0.05 \cdot 3000 - 10^{-5} \cdot 3000^2 + 10^{-5} \cdot 6500 \cdot 3000 = -45$. Wenn t in Jahren gemessen wird, bedeutet das: Wenn gerade 6500 Beutetiere und 3000 Räuber da sind, wird die Beutepopulation wachsen (und zwar mit einer momentanen Geschwindigkeit von 325 Individuen pro Jahr), die Räuberpopulation wird schrumpfen (und zwar mit einer momentanen Geschwindigkeit von 45 Individuen pro Jahr). Durch die Änderung der Populationsgrößen ändern sich diese Geschwindigkeiten laufend!

1.6.2 Modellanalyse VI: Phasenebene

Das Differentialgleichungssystem (1.60) ist mehrdimensional *und* nichtlinear. Es gehört also zum schwierigsten Typ – genau dem Typ, zu dem die allermeisten biologisch relevanten Modelle gehören. Eine explizite Lösung solcher Systeme ist nicht möglich. Man ist deshalb wieder auf die qualitative Analyse angewiesen. Zu diesem Zweck stellen wir unser wichtigstes Werkzeug für Systeme zweier gekoppelter Differentialgleichungen vor, die Darstellung in der sogenannten **Phasenebene**, siehe Abbildung (1.16). Die Phasenebene ist die zweidimensionale Entsprechung zum Phasenliniendiagramm. Zu ihrer Konstruktion gehen wir folgendermaßen vor:

1. Wir zeichnen eine x,y-Ebene. Analog zum Phasenliniendiagramm fassen wir jeden Zustand des Systems als Punkt (x,y) in der x,y-Ebene (genauer in deren positivem Quadranten) auf; z.B. entspricht der Punkt $(1000, 10)$ einer Situation mit 1000 Beute- und 10 Räubertieren.

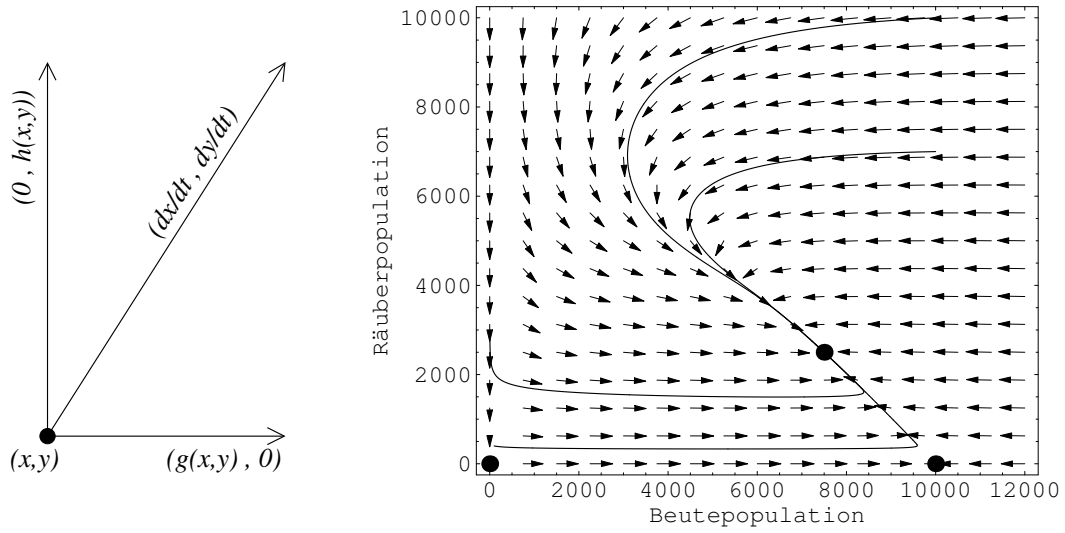


Abbildung 1.16: Differentialgleichungssystem als Vektorfeld. Links: Ein einzelner Punkt (x, y) mit angeheftetem Geschwindigkeitsvektor $(\dot{x}, \dot{y}) = (g(x, y), h(x, y))$. Mitte: Phasenebene für das Räuber-Beute-Modell mit Vektorfeld für die Parameter (1.61). Lösungen für verschiedene Anfangswerte $((10000, 10000); (10000, 7000); (2, 3000); (100, 400)$, durchgezogene Kurven) folgen dem Vektorfeld.

2. Wir stellen uns nun den Zeitverlauf $(x(t), y(t))$ als Kurve in der Ebene vor, wobei jeweils die horizontale Position der Größe der Beutepopulation und die vertikale Position der Größe der Räuberpopulation entspricht.

3. Den Zeitverlauf kennen wir zunächst nicht. Analog zum Phasenliniendiagramm können wir aber *an jedem Punkt* (x, y) einen Vektor anheften mit x-Komponente $g(x, y)$ und y-Komponente $h(x, y)$. Man erhält so ein **Vektorfeld**, s. Abb. 1.16. (Im Unterschied zum Phasenliniendiagramm ist für die Funktionen $g(x, y)$ und $h(x, y)$ selbst im Diagramm kein Platz, sie werden nur über die Pfeile ausgedrückt).

4. Wenn man den Startwert (x_0, y_0) vorgibt, erhält man nun die Lösungskurve (also der Zeitverlauf) indem man den Pfeilen folgt. Die Kurve ist mithin stets tangential zum Vektorfeld. Eine grobe Skizze für die Lösungen $x(t)$ und $y(t)$ als Funktion der Zeit (einzeln) erhält man dann durch Projektion auf die x– bzw. y-Achse.

Abbildung (1.17) zeigt den tatsächlichen Zeitverlauf für das Räuber-Beute Modell und Anfangswert $(x_0, y_0) = (10000, 7000)$.

Qualitatives Verhalten des Räuber-Beute-Modells

Wir wollen erst gar nicht versuchen, explizite Lösungen für gekoppelte Differentialgleichungen zu finden. Stattdessen bleiben wir bei der bewährten Strategie, von den Gleichgewichtspunkten des Systems auf das Langzeitverhalten zu schließen. Die Gleichgewichtspunkte (x^*, y^*) sind solche Wertepaare, an denen sich x und y *beide* nicht ändern, also die Lösungen des **nichtlinearen**

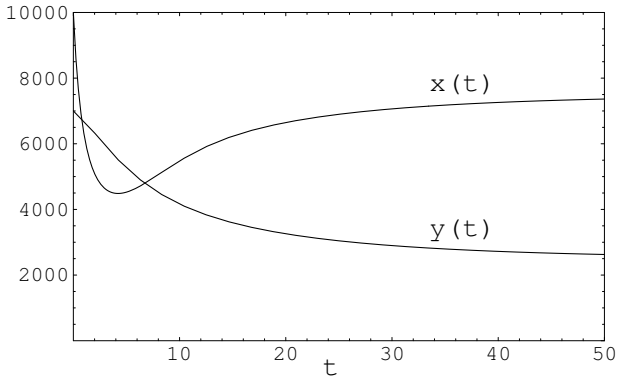


Abbildung 1.17: $x(t)$ und $y(t)$ für die Anfangswerte $(x_0, y_0) = (10000, 7000)$ als Funktionen der Zeit.

Gleichungssystems

$$g(x,y) = 0 \tag{1.62}$$
$$h(x,y) = 0. \tag{1.63}$$

Bevor wir dieses System für unser Räuber-Beute-Modell lösen, überlegen wir uns, was es geometrisch bedeutet. Ganz allgemein kann man die Lösungen der Gleichungen (1.62) und (1.63) als Kurven in der Phasenebene darstellen; man bezeichnet sie als **Nullisoklinen**, s. Abb. 1.18. Die Lösungskurve von $g(x, y) = 0$, also die Menge derjenigen (x, y)-Paare, an denen sich x nicht ändert, ist die **x-Nullisokline**; die Lösungskurve von $h(x, y) = 0$, also die Menge derjenigen Punkte, an denen sich y nicht verändert, ist die **y-Nullisokline**. Ein Schnittpunkt der beiden Nullisoklinen ist dann ein Gleichgewichtspunkt (x^*, y^*) des Systems.
Wir wollen nun die Nullisoklinen und Gleichgewichtspunkte für das Räuber-Beute-Modell berechnen. Dazu gehen wir in drei Schritten vor:

1. Schritt: $g(x, y)$ und $h(x, y)$ faktorisiert schreiben – wie in (1.60):

$$g(x,y) = x(\lambda_x - \gamma_x x - py) \tag{1.64}$$
$$h(x,y) = y(-\lambda_y - \gamma_y y + qx) \tag{1.65}$$

2. Schritt: x- und y-Nullisokline berechnen. Die x-Nullisokline ist das Gebilde, das aus den beiden Geraden

$$x = 0 \quad \text{und} \quad \lambda_x - \gamma_x x - py = 0 \tag{1.66}$$
$$\Leftrightarrow y = \frac{\lambda_x}{p} - \frac{\gamma_x}{p} x \tag{1.67}$$

besteht. Die y-Nullisokline besteht aus den beiden Geraden

$$y = 0 \quad \text{und} \quad -\lambda_y - \gamma_y y + qx = 0 \tag{1.68}$$
$$\Leftrightarrow \quad y = -\frac{\lambda_y}{\gamma_y} + \frac{q}{\gamma_y} x\,. \tag{1.69}$$

Für die Parameterwerte (1.61) sind die beiden Null-Isoklinen in Abb. 1.18 gezeigt.

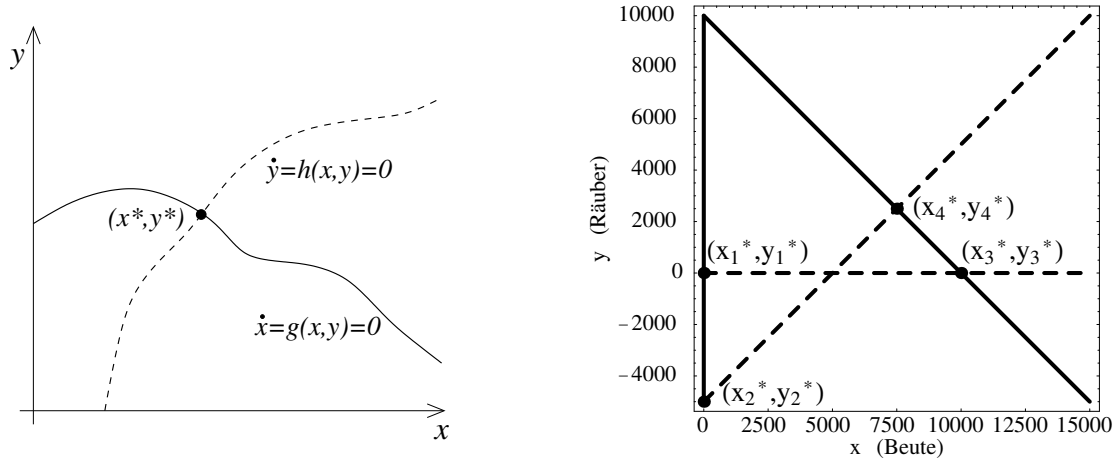


Abbildung 1.18: Links: Nullisoklinen für zwei gekoppelte Differentialgleichungen (allgemeine Situation). Schnittpunkte der x-und y-Nullisoklinen sind Gleichgewichtspunkte des Systems. Rechts: x-Nullisoklinen (fette Linien) und y-Nullisoklinen (gestrichelte Linien) für das Räuber-Beute-Modell (1.60) mit den Parameterwerten (1.61). Wo sich eine fette mit einer gestrichelten Linie schneidet, befindet sich ein Gleichgewichtspunkt.

3. Schritt: Gleichgewichtspunkte als Schnittpunkte der x- und y-Nullisoklinen berechnen. Dazu kombinieren wir die Möglichkeiten in Gl. (1.66) und (1.68) miteinander und erhalten vier Schnittpunkte:

a) $x = 0$ und $y = 0$. $(x_1^*, y_1^*) = (0, 0)$ ist offensichtlich Gleichgewicht des Systems. Sowohl Beute als auch Räuber sind ausgestorben.

b) $x = 0$ und $-\lambda_y - \gamma_y y + qx = 0$. Dies wird gelöst von $(x_2^*, y_2^*) = (0, -\frac{\lambda_y}{\gamma_y})$. Da $y_2^* < 0$, ist dieses Gleichgewicht irrelevant (Räuber können ohne Beute langfristig nicht überleben).

c) $\lambda_x - \gamma_x x - py = 0$ und $y = 0$. Hier erhält man $(x_3^*, y_3^*) = (\frac{\lambda_y}{\gamma_y}, 0)$. Dies ist das Gleichgewicht des logistischen Wachstums von Beute in Abwesenheit von Räubern ($x_3^* > 0$).

d) Einen letzten Gleichgewichtspunkt liefert der Schnittpunkt der beiden Geraden (1.67) und (1.69). Wenn man diese gleichsetzt erhält man

$$\frac{\lambda_x}{p} - \frac{\gamma_x}{p} x = -\frac{\lambda_y}{\gamma_y} + \frac{q}{\gamma_y} x \,. \tag{1.70}$$

Nach x aufgelöst erhält man hieraus x_4^* und wenn man x_4^* in (1.67) oder (1.69) einsetzt auch y_4^* als

$$x_4^* = \frac{\gamma_y \lambda_x + \lambda_y p}{\gamma_x \gamma_y + pq} \qquad \text{und} \qquad y_4^* = \frac{-\gamma_x \lambda_y + \lambda_x q}{\gamma_x \gamma_y + pq} \,. \tag{1.71}$$

Wenn x_4^* und y_4^* beide positiv sind, koexistieren die beiden Spezies.

In unserem Beispiel (1.61) lauten die Gleichgewichtspunkte

$$(x_1^*, y_1^*) = (0, 0)\,, \quad (x_2^*, y_2^*) = (0, -5000), \quad (x_3^*, y_3^*) = (10000, 0)\,, \quad (x_4^*, y_4^*) = (7500, 2500).$$

In diesem Fall ist $(x_4^*, y_4^*) > 0$. Für andere Parameterwerte braucht dies aber nicht der Fall zu sein, wie wir weiter unten noch sehen werden.
Die Berechnung der Stabilitätseigenschaften erfordert Methoden aus der linearen Algebra, die wir in dieser Vorlesung nicht mehr behandeln werden. Wir wollen uns auf graphische Methoden konzentrieren. Wenn man sich die Mühe macht, ein Vektorfeld zu zeichnen (oder ein Computerprogramm hat, das dies kann), kann man die Stabilitätseigenschaften oft einfach ablesen: Aus Bild 1.16 sieht man z.B., dass das Räuber-Beute-Gleichgewicht (x_4^*, y_4^*) **(lokal) stabil** ist, d.h. Pfeile laufen *aus allen Richtungen* darauf *zu*; die anderen beiden Gleichgewichtspunkte sind dagegen **instabil**, d.h. **mindestens in einer Richtung** laufen die Pfeile *davon weg*. Als nächstes wollen wir überlegen, wie man die Stabilität von Gleichgewichtspunkten herausfinden kann, ohne gleich ein ganzes Vektorfeld zu zeichnen.
Wählen wir dazu $\lambda_y = 0.2$, ansonsten dieselben Parameter wie in (1.61). In diesem Fall sind die Nullisoklinen

$$x\text{-Nullisokline:} \quad x = 0 \qquad \text{und} \qquad y = 10000 - x\,; \tag{1.72}$$

$$y\text{-Nullisokline:} \quad y = 0 \qquad \text{und} \qquad y = -20000 + x\,. \tag{1.73}$$

Sie sind in Abb. 1.19 gezeigt. Diesmal gibt es kein relevantes Räuber-Beute-Gleichgewicht, da $y_4^* < 0$. Wir überlegen nun für $x, y > 0$ die grobe Richtung des Vektorfeldes. Dazu reicht es, sich über die Vorzeichen von $\dot{x}$ und $\dot{y}$ in den verschiedenen Segmenten der Phasenebene Klarheit zu verschaffen (s. Abb. 1.19 links). Wir wissen, dass $\dot{x} > 0$ ($\dot{x} < 0$), wenn $g(x, y) > 0$ ($g(x, y) < 0$). Für positive x und y ist das aber gerade dann der Fall, wenn $y < 10000 - x$ (bzw. $y > 10000 - x$) ist, also für Punkte unterhalb (oberhalb) der x-Nullisoklinen; vgl. Gl. (1.72). Andererseits ist $\dot{y} > 0$ ($\dot{y} < 0$), wenn $h(x, y) > 0$ ($h(x, y) < 0$). Für positive x und y ist das der Fall, wenn $y < -2000 + x$ (bzw. $y > -2000 + x$) ist, also für Punkte unterhalb (oberhalb) der y-Nullisoklinen; vgl. Gl. (1.73). Weiterhin ist entlang der positiven y-Achse stets $\dot{x} = g(0, y) = 0$ und $\dot{y} = h(0, y) < 0$; entlang der positiven x-Achse ist $\dot{y} = h(x, 0) = 0$ und $\dot{x} = g(x, 0) > 0$ ($\dot{x} = g(x, 0) < 0$) wenn $x < 10000$ ($x > 10000$).
Aus diesen Überlegungen ergeben sich die Richtungsvektoren in Abb. 1.19 rechts. Man sieht, dass die Lösungen aus allen Richtungen auf das Gleichgewicht 'Beute ohne Räuber', (x_3^*, y_3^*), zulaufen. Die Räuber sterben also aus. Es ist zwar Beute vorhanden, aber die Räuber können nicht genug Beutetiere erlegen, um ihren Nahrungsbedarf zu decken.

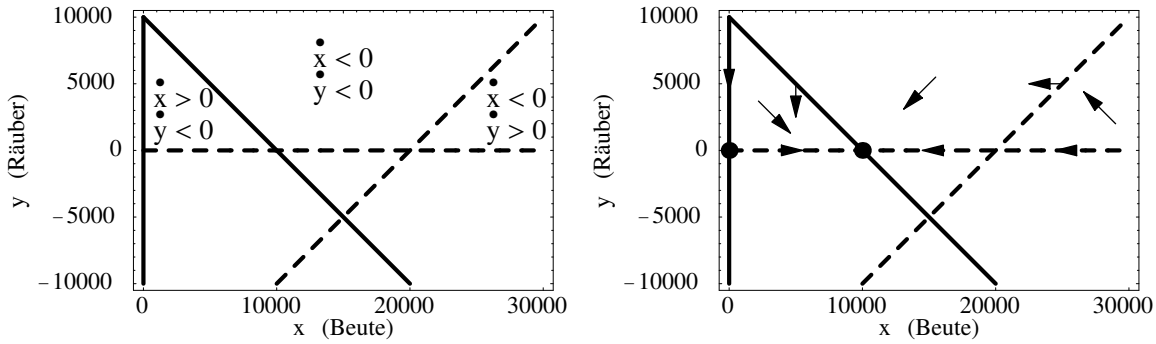


Abbildung 1.19: Nullisoklinen für das Räuber-Beute-Modell mit $\lambda_y = 0.2$. Links: Die verschiedenen Segmente der Phasenebene mit den entsprechenden Vorzeichen von $\dot{x}$ und $\dot{y}$. Rechts: Die zugehörigen Richtungspfeile.

1.6.3 Anwendung: Das Konkurrenzmodell der Ökologie

Zum Abschluss des ersten Teils der Vorlesung wollen wir noch einen weiteren wichtigen Fall eines Prozesses mit zwei Populationen diskutieren, der sich mit den selben Methoden behandeln lässt. Das Konkurrenzmodell der Ökologie beschreibt zwei Populationen, die um dieselbe Ressource konkurrieren. Die wesentliche Frage, die wir stellen wollen ist: Unter welchen Bedingungen ist Koexistenz von Arten möglich? Das Differentialgleichungssystem des Konkurenzmodells lautet

$$\begin{aligned} \dot{x} &= \lambda_x x - \gamma_x x^2 - \gamma_{xy} xy \\ \dot{y} &= \lambda_y y - \gamma_y y^2 - \gamma_{yx} xy \end{aligned} \tag{1.74}$$

Der Einfachheit halber nehmen wir an, dass beide Populationen ohne Konkurrenz gleich schnell wachsen und setzen die Wachstumsrate auf $\lambda_x = \lambda_y = 1$. Die γ-Parameter messen die Stärke der Konkurrenz; sie sind stets positiv. Es ist sinnvoll, ihre Bedeutung einzeln zu betrachten: die Koeffizienten γ_x und γ_y messen, wie stark das Wachstum der Spezies x und y jeweils durch innerartliche Konkurrenz beeinträchtigt wird. γ_{xy} misst den Konkurrenzeffekt der Spezies y auf x, und umgekehrt γ_{yx} die Stärke der Konkurrenz, die y durch x spürt.
Zur Behandlung des Modells berechnen wir nun zunächst die Nullisoklinen:

$$\begin{aligned} g(x,y) = x(1-\gamma_x x - \gamma_{xy} y) = 0 \quad &\text{für} \quad x = 0 \quad \text{und} \quad y = \frac{1-\gamma_x x}{\gamma_{xy}} \\ h(x,y) = y(1-\gamma_y y - \gamma_{yx} x) = 0 \quad &\text{für} \quad y = 0 \quad \text{und} \quad y = \frac{1-\gamma_{yx} x}{\gamma_y} \end{aligned} \tag{1.75}$$

Daraus ergeben sich die folgenden Gleichgewichtspunkte als Schnittpunkte der Isoklinen:

$$(x_1^*, y_1^*) = (0,0) \quad ; \quad (x_2^*, y_2^*) = (0, 1/\gamma_y) \quad ; \quad (x_3^*, y_3^*) = (1/\gamma_x, 0)$$

$$(x_4^*, y_4^*) = \Big(\frac{\gamma_y - \gamma_{xy}}{\gamma_x\gamma_y - \gamma_{xy}\gamma_{yx}}, \frac{\gamma_x - \gamma_{yx}}{\gamma_x\gamma_y - \gamma_{xy}\gamma_{yx}} \Big)$$

Der vierte Fixpunkt ist nur dann biologisch relevant wenn x_4^* und y_4^* beide positiv sind. Nur wenn (x_4^*, y_4^*) im positiven Quadranten liegt *und* stabil ist wird es Koexistenz geben. Je nachdem, ob γ_x größer oder kleiner als γ_{yx} ist, und ob γ_y größer oder kleiner als γ_{xy} ist, gibt es vier Fälle, die man unterscheiden muss.

1. Nehmen wir zuerst an, dass γ_x kleiner ist als γ_{yx}, aber $\gamma_y > \gamma_{xy}$. Biologisch heißt das, dass die Spezies x sich selbst weniger "schadet" (Konkurrenz macht) als der Spezies y. Spezies y dagegen macht sich selbst stärker Konkurrenz als der Spezies x. Spezies y ist also durchweg weniger kompetitiv als Spezies x. Im Diagramm der Phasenebene (Abb. 1.20 oben) sieht man, dass der Fixpunkt (x_4^*, y_4^*) nicht im positiven Quadranten liegt, es gibt keine Koexistenz. Der einzige stabile Fixpunkt ist $(x_2^*, y_2^*) = (0, 1/\gamma_y)$: y stirbt aus und nur x überlebt.

2. Im entsprechenden Fall mit x und y vertauscht (also $\gamma_x > \gamma_{yx}$ und $\gamma_y < \gamma_{xy}$) sind die Rollen vertauscht: nur die kompetitivere Spezies y überlebt.

3. Nehmen wir jetzt an, dass gilt $\gamma_x < \gamma_{yx}$ und $\gamma_y < \gamma_{xy}$. Dann schaden beide Spezies der jeweils anderen mehr als sich selbst. In der Phasenebene (Abb. 1.20 Mitte) sehen wir, dass wir jetzt einen Gleichgewichtspunkt im positiven Quadranten haben. Dieser Gleichgewichtspunkt ist aber *instabil*, er wird also in der Realität nicht erreicht werden. Es gibt *zwei* stabile Gleichgewichtspunkte, $(x_2^*, y_2^*) = (0, 1/\gamma_y)$ und $(x_3^*, y_3^*) = (1/\gamma_x, 0)$. Je nach

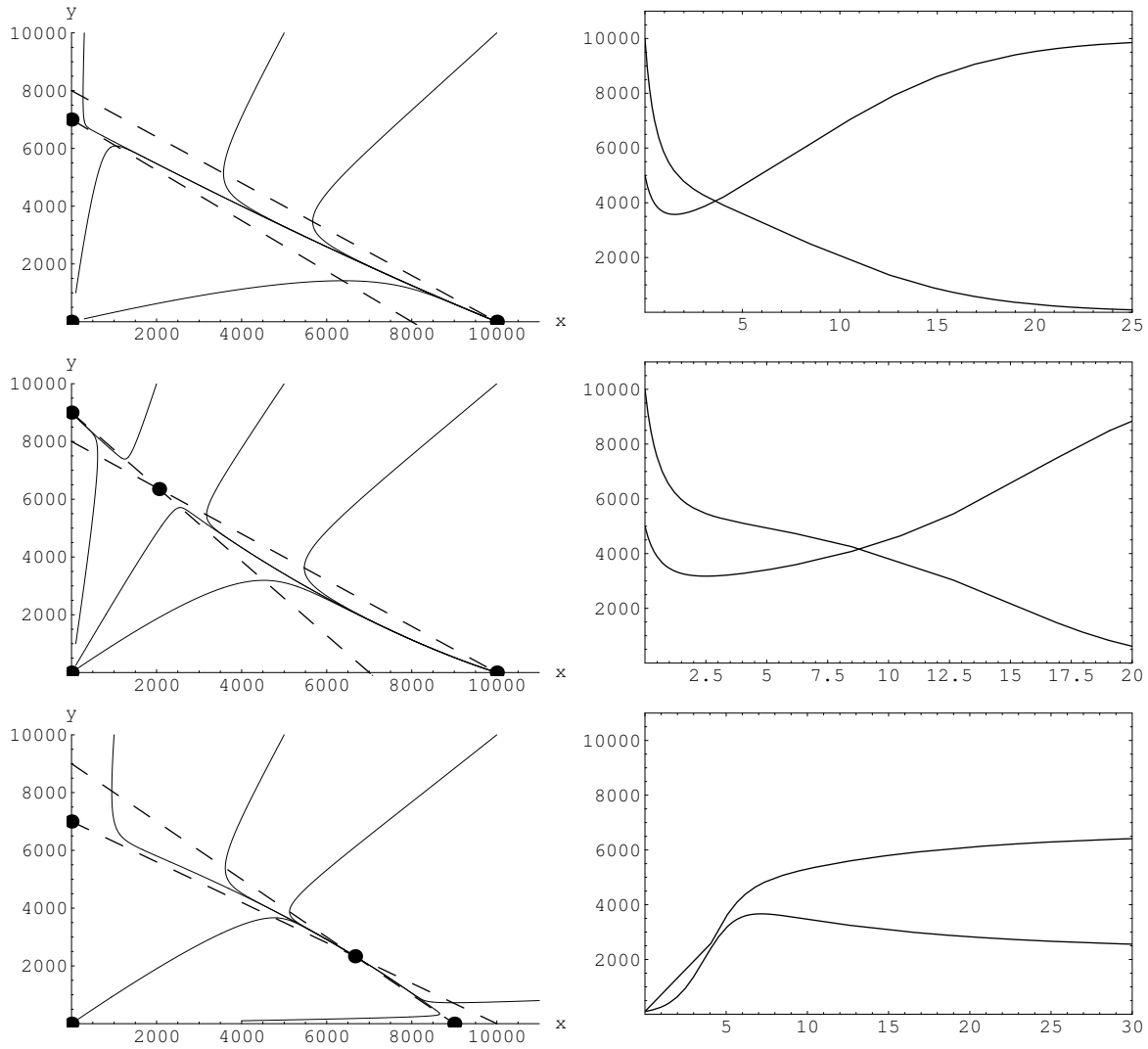


Abbildung 1.20: Darstellung der drei Fälle der Konkurrenz zwischen zwei Arten in der Phasenebene und im Zeitverlauf. Oben: Spezies x ist der überlegene Kompetitor und verdrängt Spezies y. Mitte: Spezies x und y können sich wechselseitig verdrängen, je nach den Anfangsbedingungen. Unten: Koexistenz.

dem *Anfangswert* des Prozesses wird entweder der eine oder der andere Punkt erreicht. Je häufiger eine Art am Anfang ist, desto kompetitiver ist sie. Insbesondere haben einzelne "Eindringlinge" einer Spezies auf dem Gebiet der anderen Spezies keine Chance, sich zu vermehren.

4. Nehmen wir als letzten Fall an, dass sich beide Spezies selbst mehr Konkurrenz machen als der jeweils anderen. Wir haben also $\gamma_x > \gamma_{yx}$ und $\gamma_y > \gamma_{xy}$, die innerartliche Konkurrenz ist stets größer als die zwischenartliche Konkurrenz. Im Diagramm der Phasenebene (Abb. 1.20 unten) sehen wir, dass wir in diesem Fall tatsächlich einen stabilen Gleichgewichtspunkt für Koexistenz bekommen.

Wir erhalten über das Modell also eine Antwort auf die Frage, wann Koexistenz von Arten möglich ist: jede Art muss sich jeweils selbst stärker Konkurenz machen als der anderen. Biolo-

gisch ist das nur dann möglich, wenn die Arten gegenseitig nicht um *jede* Ressource konkurrieren, sondern jeweils noch ihre speziell eigenen haben. Man spricht dann von einer Differenzierung der *ökologischen Nischen* der beiden Arten und sagt auch: "koexistierende Arten dürfen sich nicht zu ähnlich sein".

Teil 2

Wahrscheinlichkeitsrechnung und Statistik

Im ersten Teil der Vorlesung haben wir deterministische Prozesse betrachtet, also solche, bei denen der Zufall keine Rolle spielte. In der Biologie ist der Zufall aber häufig ein wesentliches Element – sowohl in Abläufen in der Natur, als auch bei der experimentellen Erhebung von Daten. Der enge Zusammenhang zwischen Biologie und Wahrscheinlichkeitstheorie zeigt sich auch darin, dass wesentliche Teile der mathematischen Statistik Anfang des letzten Jahrhunderts ganz explizit zur Behandlung von Problemen aus der Populationsgenetik entwickelt wurden. Der mathematische Umgang mit Zufall und Wahrscheinlichkeiten ist deshalb das Thema des zweiten Teils der Vorlesung.

2.1 Grundbegriffe und Definitionen

Bevor wir zu den grundlegenden Definitionen kommen, starten wir mit einem Beispiel.

Die Gen-Lotterie

Vererbungslehre ist angewandte Wahrscheinlichkeitstheorie: Kinder erben zufällige Kombinationen der Gene ihrer Eltern. Betrachten wir im einfachsten Fall einen einzigen Gen-Locus mit zwei Allelen A und a. A könnte zum Beispiel der Wildtyp sein und a ein Krankheitsgen. Wir nehmen an, dass beide Eltern heterozygot sind, also den Genotyp Aa haben. Kinder erben je ein Allel von beiden Eltern, es gibt also drei Möglichkeiten für den Genotyp des Kindes: Homozygot für eines der beiden Allele, AA oder aa, oder heterozygot, Aa. Wir fragen nach der Wahrscheinlichkeit dieser Ereignisse. Da jedes Elternteil in der Mitose genausoviele Gameten mit dem A und dem a Allel produziert, erbt ein Kind von Mutter und Vater jedes der beiden Allele mit Wahrscheinlichkeit von je $1/2$. Außerdem hängt die Wahrscheinlichkeit, A oder a vom Vater zu erben, nicht davon ab, welches Allel von der Mutter vererbt wurde. Daraus ergeben sich direkt die Wahrscheinlichkeiten für den Genotyp des Kindes:

- Für den homozygoten Typ AA muss das Kind das A-Allel von Mutter und Vater erben. Jedes dieser Ereignisse einzeln tritt in der Hälfte aller Fälle ein. Da die Ereignisse unabhängig sind, finden wir den Genotyp AA in der Hälfte der Hälfte aller Fälle, also mit Wahrscheinlichkeit $Pr(AA) = (1/2) \cdot (1/2) = 1/4$. Genauso gilt $Pr(aa) = 1/4$.
- Für den heterozygoten Typ Aa gibt es zwei Möglichkeiten. Entweder erbt das Kind Allel A von der Mutter und Allel a vom Vater oder umgekehrt a von der Mutter und A vom Vater.

Jedes dieser Ereignisse tritt mit Wahrscheinlichkeit 1/4 ein. Da beide Ereignisse nicht gleichzeitig eintreten können, finden wir den heterozygoten Typ einfach in der Summe aller Fälle dieser beiden Ereignisse, also mit Wahrscheinlichkeit $Pr(Aa) = (1/4) + (1/4) = 1/2$.

2.1.1 Ereignis und Zufallsvariable

Allgemein nennt man die Menge $\Omega = \{\omega_i\}$ aller möglichen Ausgänge ω_i eines Zufallsexperimentes die **Ergebnismenge**. Jede Teilmenge $A \subseteq \Omega$ heißt **Ereignis** des Zufallsexperimentes. Ein Ereignis mit nur einem Element heißt **Elementarereignis**. Jedem Ereignis A ordnet man eine Wahrscheinlichkeit $Pr(A)$ zu. Empirisch entspricht $Pr(A)$ der **relativen Häufigkeit** des Ereignisses A wenn man das Experiment sehr oft (bzw. im Grenzfall unendlich oft) durchführt,

$$Pr(A) = \lim_{n\to\infty} \frac{n_A}{n}\,, \tag{2.76}$$

wobei n die Anzahl der Experimente insgesamt ist, und n_A die Anzahl der Fälle ist, in denen das Ereignis A eintritt.[2] Es ist immer $0 \leq Pr(A) \leq 1$. Das **Gegenereignis** zu einem Ereignis A ist $\bar{A} := \Omega \setminus A$. Seine Wahrscheinlichkeit – also die Wahrscheinlichkeit, dass A *nicht* eintritt – ist $Pr(\bar{A}) = 1 - Pr(A)$. Wenn zwei Ereignisse A und B sich gegenseitig *ausschließen*, also nie gemeinsam auftreten können, dann ist die Wahrscheinlichkeit für das zusammengesetzte Ereignis, dass A *oder* B eintritt durch die Summe der Einzelwahrscheinlichkeiten gegben,

$$Pr(A \,\text{oder}\, B) = Pr(A) + Pr(B). \tag{2.77}$$

Im oben besprochenen Zufalls'experiment' der Vererbung besteht die Ergebnismenge aus den vier geordneten Allel-Paaren $\{(A,A),(A,a),(a,A),(a,a)\}$, wobei das erste Allel in jedem Paar das von der Mutter vererbte ist. Die Wahrscheinlichkeit für jedes der vier Elementarereignisse ist 1/4. Die Wahrscheinlichkeit für das Ereignis "heterozygot" ergibt sich als Summe der Wahrscheinlichkeiten der sich ausschließenden Ereignisse (A,a) und (a,A)), $Pr(Aa) = Pr((A,a) \,\text{oder}\, (a,A)) = 1/4 + 1/4 = 1/2$.

Sehr häufig besteht ein Zufallsexperiment aus der Kombination mehrerer Teilexperimente. Es ist dabei nicht unbedingt notwendig, dass die Teilexperimente zeitlich hintereinander ausgeführt werden, man muss sie nur gedanklich voneinander trennen können. Ein Standardbeispiel ist das Würfeln mit zwei Würfeln. Auch die Vererbung ist ein Beispiel für ein kombiniertes Experiment: Die Teilexperimente bestehen aus der Vererbung eines Allels von der Mutter und der Vererbung eines Allels vom Vater.

Beispiel: Messung mehrerer phänotypischer Merkmale

Wir betrachten als weiteres Beispiel eine typische Situation der organismischen Biologie: Tiere messen und wiegen. Gemessen werden sollen Größe und Gewicht von Individuen aus einer Population. Aus Erfahrung von vorangegangenen Messungen wissen wir, dass 10% aller Tiere in der Population größer als 30 cm sind und 20% schwerer als 1kg. Wir fragen nach der Wahrscheinlichkeit, dass ein Tier größer als 30 cm *und* schwerer als 1kg ist. Nachdem die Größenmessung bereits durchgeführt worden ist, wissen wir von einem bestimmten Individuum bereits, dass es größer als 30 cm ist. Wir fragen danach, was jetzt die Wahrscheinlichkeit dafür ist, dass es schwerer als 1kg ist.

Diese Fragen führen auf zwei zentrale Begriffe der Wahrscheinlichkeitstheorie, die auch für die

[2] "Streng mathematisch" betrachtet verwenden wir hier das Gesetz der großen Zahlen.

Planung von Experimenten und die Auswertung von Daten in der Biologie von großer Bedeutung sind, die **bedingte Wahrscheinlichkeit** und die statistische **Abhängigkeit** bzw. **Unabhängigkeit** von Ereignissen. Seien allgemein A und B Ereignisse des ersten und des zweiten Experiments mit Wahrscheinlichkeiten $Pr(A)$ und $Pr(B)$. $Pr(A, B)$ ist die Wahrscheinlichkeit, dass beide Ereignisse eintreten. Dann ist die bedingte Wahrscheinlichkeit $Pr(B|A)$ von B bei gegebenem A definiert als

$$Pr(B|A) = \frac{Pr(A, B)}{Pr(A)}. \tag{2.78}$$

Wenn $n_{A,B}$ die Zahl aller Fälle ist, in denen A und B eintritt finden wir mit Gleichung (2.76)

$$Pr(B|A) = \frac{Pr(A, B)}{Pr(A)} = \lim_{n\to\infty} \frac{n_{A,B}/n}{n_A/n} = \lim_{n\to\infty} \frac{n_{A,B}}{n_A}. \tag{2.79}$$

Empirisch ist $Pr(B|A)$ deshalb die relative Häufigkeit des Ereignisses B unter allen Fällen, in denen Ereignis A eintritt wenn man das Experiment sehr oft wiederholt. Im Beispiel: Die relative Häufigkeit von Tieren die schwerer als 1kg sind, wenn man nur Tiere betrachtet, die größer als 30 cm sind. Im allgemeinen sind $Pr(B|A)$ und $Pr(B)$ verschiedene Größen: Unter den großen Tieren ist es in aller Regel wahrscheinlicher, dass man ein schweres Tier findet, als unter allen Tieren, also $Pr(\text{“schwer”}|\text{“groß”}) > Pr(\text{“schwer”})$. Man sagt, Größe und Gewicht eines Individuums sind stochastisch abhängige Messgrößen. Der gegenteilige und besonders wichtige Fall ist die Unabhängigkeit von Ereignissen.

Definition 4 (Stochastische Unabhängigkeit) *Zwei Ereignisse A und B heißen (stochastisch)* unabhängig*, wenn die Information, ob das eine Ereignis eintritt oder nicht, nichts an der Wahrscheinlichkeit für das andere Ereignis ändert. Dies ist genau dann der Fall wenn die Wahrscheinlichkeit, dass* beide *Ereignisse eintreten durch das Produkt der Einzelwahrscheinlichkeiten gegeben ist,*

$$Pr(A, B) = Pr(A)Pr(B). \tag{2.80}$$

Aus Gleichung (2.78) folgt dann (falls $Pr(A) > 0$) $Pr(B|A) = Pr(B)$. Analog folgt $Pr(A|B) = Pr(A)$.

Unabhängigkeit bedeutet, dass zwei Ereignisse im wahrscheinlichkeitstheoretischen Sinn keinen Einfluß aufeinander ausüben. Dies ist nicht ganz dasselbe wie die Abwesenheit realer Beeinflussung; wesentlich ist nur, dass das Eintreten des einen Ereignisses keinen Einfluss auf die Wahrscheinlichkeit des anderen Ereignisses hat (siehe dazu die Aufgaben). Das Vererbungsexperiment ist ein Beispiel, in dem die beiden Teilexperimente (Vererbung von der Mutter und vom Vater) unabhängig sind.
Die bisher eingeführten Grundbegriffe lassen sich gut im Bild des Wahrscheinlichkeitsbaums darstellen, siehe Abbildung (2.21). Man beachte, dass die Reihenfolge der Teilexperimente im Baum in Prinzip beliebig ist. In der Praxis hängt sie nur davon ab, ob man zur bedingten Wahrscheinlichkeit vom Typ $Pr(A|B)$ oder zu $Pr(B|A)$ den direkteren Zugang hat.

Häufig interessiert man sich bei einem Zufallsexperiment nicht direkt für die Vielfalt der möglichen Ereignisse selbst, sondern für zähl- oder messbare Größen, die sich aus den Ereignissen ableiten. Ein einfaches Beispiel ist die Augensumme beim zweimaligen Werfen eines Würfels. In der Genetik interessiert man sich für den Effekt des Genotyps auf ein phänotypisches Merkmal, wie die Körpergröße oder die Fitness (die mittlere Anzahl an Nachkommen pro Generation). Dies führt zur Definition der **Zufallsvariable**.

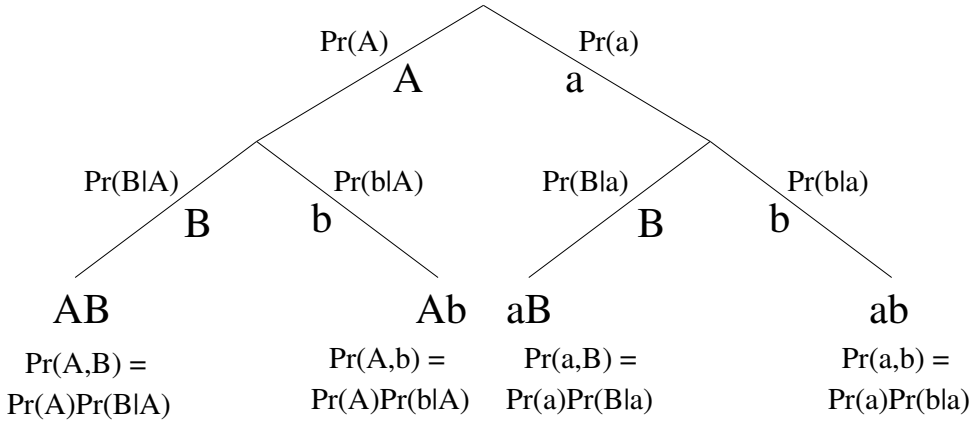


Abbildung 2.21: Wahrscheinlichkeitsbaum für ein zweistufiges Zufallsexperiment. Jede Verzweigung an einer Astgabel entspricht dem Elementarereignis eines Teilexperiments, A und $a(= \bar{A})$ für das erste Teilexperiment, B und b für das zweite. Jeder Ast des Baumes (vom Stamm bis zum Blatt) entspricht einem Elementarereignis des Gesamtexperiments, AB, Ab, aB und ab. An jeder Verzeigung sind die Wahrscheinlichkeiten für die Ereignisse der Teilexperimente angegeben. Für das erste Teilexperiment an der Verzweigung am Stamm die Gesamtwahrscheinlichkeiten $Pr(A)$ und $Pr(a)$, an den Unterverzweigungen in den Ästen jeweils die bedingten Wahrscheinlichkeiten $Pr(B|A)$ etc. Die Wahrscheinlichkeit eines Elementarereignisses des gesamten Experiments erhält man, indem man die Wahrscheinlichkeiten entlang des entsprechenden Astes multipliziert. Unabhängigkeit der beiden Teilexperimente besteht genau dann, wenn die bedingten Wahrscheinlichkeiten in allen Ästen gleich sind: $Pr(B|A) = Pr(B|a) = Pr(B)$ und $Pr(b|A) = Pr(b|a) = Pr(b)$. Ereignisse in verschiedenen Ästen des Baumes schließen sich gegenseitig aus. Wahrscheinlichkeiten für zusammengesetzte Ereignisse (z.B. Ab oder aB) erhält man, indem man über die Wahrscheinlichkeiten der entsprechenden Elementarereignisse summiert.

Definition 5 (Zufallsvariable, ZV) *Wenn Ω die Ereignismenge eines Zufallsexperiments ist, dann heißt eine Abbildung $X : \Omega \to \mathbb{R}$ (reelle) Zufallsvariable (ZV). Die Werte, die eine Zufallsvariable annehmen kann, heißen ihre* **Realisierungen**.

Wir beschränken uns zunächst auf den Fall einer **diskreten ZV** X, die nur endlich viele Werte $x_1, x_2, \ldots, x_n$ annehmen kann. Auch Größen, die eigentlich kontinuierlich veränderbar sind, wie Längen und Gewichte, stellt man oft effektiv durch eine diskrete ZV dar indem man die Werte in Klassen einteilt. Dies geschieht in der Praxis schon dadurch, dass man nicht beliebig genau messen kann, sondern z.B. auf ganze Millimeter rundet. Eine Zufallsvariable (X) bezieht sich auf die Situation vor dem Experiment, dessen Ausgang noch ungewiss ist. Hat man das Experiment durchgeführt, so hat man eine konkrete Realisierung (x), deren Wert nun feststeht. Im folgenden werden Zufallsvariablen stets mit Großbuchstaben, die zugehörigen Realisierungen mit Kleinbuchstaben bezeichnet.

Nehmen wir an, dass im obigen Beispiel die Fitness der Genotypen $w_{AA} = 1$, $w_{Aa} = 1 - hs$ und $w_{aa} = 1 - s$ ist. Dann ist die Fitness W des Kindes eine Zufallsvariable mit diesen Werten als Realisierungen. Der Parameter h ist der sogenannte Dominanzparameter. Insbesondere bedeutet $h = 0$ volle Dominanz des A-Allels. In diesem speziellen Fall hat die Zufallsvariable W nur zwei Realisierungen, $w_1 = 1$ und $w_2 = 1 - s$.

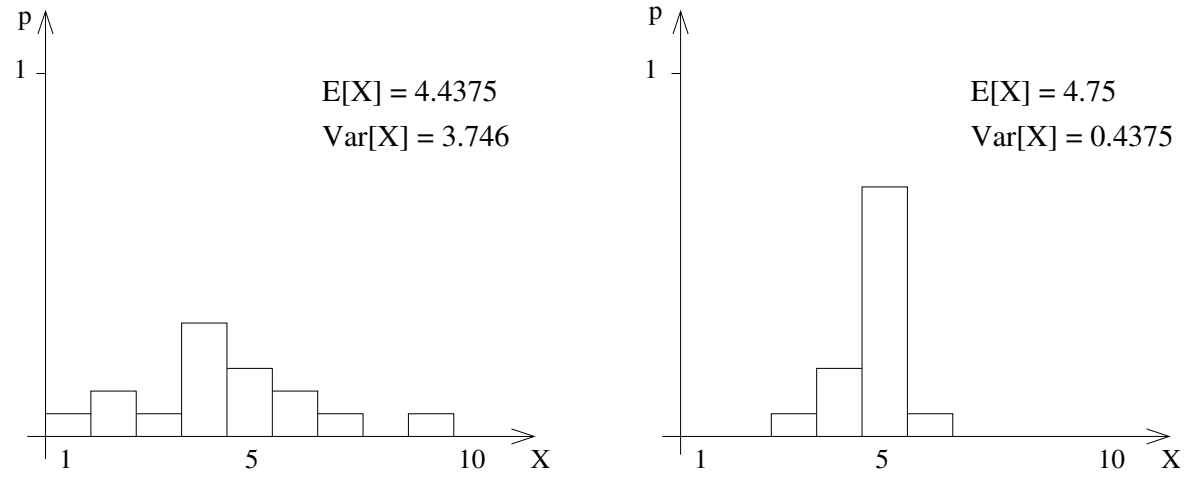


Abbildung 2.22: Darstellung der Wahrscheinlichkeitsverteilung einer Zufallsvariablen im Säulendiagramm. Links Verteilung mit großer, rechts mit kleiner Varianz.

Jedem Wert x_i einer diskreten ZV X kann man nun eine Wahrscheinlichkeit zuordnen, indem man die Wahrscheinlichkeiten aller Elementarereignisse, die zum Wert x_i führen, addiert:

$$p_i := Pr(X = x_i) = \sum_{\omega : X(\omega) = x_i} Pr(\omega). \tag{2.81}$$

Die Zuordnung der p_i zu den x_i nennt man **Wahrscheinlichkeitsverteilung** (W-Verteilung) von X. Anschaulich kann man eine W-Verteilung als Säulendiagramm darstellen (Abbildung 2.22). Wir wollen nun versuchen, W-Verteilungen zu charakterisieren. Zunächst interessiert uns, welchen Wert unsere Zufallsgröße *im Mittel* annimmt, wenn man das Experiment unendlich oft wiederholt. Das ist der Erwartungswert:

Definition 6 (Erwartungswert) *Ist X eine diskrete ZV mit Realisierungen $x_1, \ldots, x_n$, dann heißt*

$$\mu := \mathbf{E}[X] := \sum_{i=1}^{n} p_i \, x_i$$

der Erwartungswert von X.

Die mittlere Fitness für das Kind in unserem Beispiel ist $\mathbf{E}[W] = (1/4) \cdot 1 + (1/2) \cdot (1 - hs) + (1/4) \cdot (1 - s) = 1 - (hs/2) - (s/4)$.
Der Erwartungswert gibt den 'Durchschnittswert' einer Zufallsvariablen an, sagt aber nichts über die 'Gestalt' der Verteilung aus. Eine wichtige Kenngröße ist die Streuung der Verteilung um den Erwartungswert. Man misst sie anhand der *mittleren quadratischen Abweichung vom Erwartungswert*, der sogenannten **Varianz**:

Definition 7 (Varianz, Standardabweichung) *Die Größe*

$$\sigma^2 := \mathrm{Var}[X] = \mathbf{E}\left[\big(X - \mu\big)^2\right] = \sum_i (x_i - \mu)^2 \, p_i = \big(\sum_i x_i^2 \, p_i\big) - \mu^2$$

heißt Varianz von X. Die Größe

$$\sigma = \sqrt{\mathrm{Var}[X]}$$

heißt Standardabweichung.

2.1.2 Mehrere Zufallsvariablen

Oft interessiert man sich in einem System für die Beziehungen zwischen mehreren ZVn (wie im obigen Beispiel zwischen Größe und Gewicht). Man definiert dann eine gemeinsame W-Verteilung, $p_{ij} = Pr(X = x_i, Y = y_j)$, analog zu (2.81) als Summe über die Wahrscheinlichkeiten aller Elementarereignisse ω, für die $X(\omega) = x_i$ und $Y(\omega) = y_i$ ist. Zwei ZVn X und Y heißen unabhängig, wenn alle ihre Realisierungen x_i und y_j paarweise unabhängige Ereignisse sind. Dies ist genau dann der Fall, wenn die gemeinsame W-Verteilung faktorisiert:

$$p_{ij} = Pr(X = x_i, Y = y_j) = Pr(X = x_i)Pr(Y = y_j) = p_{X,i}\, p_{Y,j}, \tag{2.82}$$

wobei $p_{X,i}$ die Wahrscheinlichkeiten ist, dass die ZV X den Wert x_i annimmt und $p_{Y,j}$ die Wahrscheinlichkeit, dass Y den Wert y_j annimmt. Als Gegenstück zur Varianz für eine ZV ist für zwei ZVn ein Maß für ihre Korrelation definiert, die sogenannte **Kovarianz**:

Definition 8 (Kovarianz, Korrelationskoeffizient) *Die Größe*

$$\mathrm{Cov}[X, Y] = \mathbf{E}\left[(X - \mu_X)(Y - \mu_Y)\right] = \sum_{i,j}(x_i - \mu_X)(y_j - \mu_Y)\, p_{ij} = \big(\sum_{i,j} x_i y_j\, p_{ij}\big) - \mu_X\mu_Y$$

heißt Kovarianz von X und Y. Offensichtlich ist $\mathrm{Cov}[X, X] = \mathrm{Var}[X]$. *Die Größe*

$$\rho_{XY} = \frac{\mathrm{Cov}[X, Y]}{\sigma_X \sigma_Y}$$

heißt Korrelationskoeffizient von X und Y.

Positive Korrelation bedeutet, dass eine Tendenz besteht, nach der $X(\omega)$ für diejenigen ω die größeren Werte annimmt, für die auch $Y(\omega)$ die größeren Werte annimmt. Dann werden nämlich $(x_i - \mu_X)$ und $(y_j - \mu_Y)$ oft das gleiche Vorzeichen haben und das Produkt $(x_i - \mu_X)(y_j - \mu_Y)$ ist positiv. Negative Kovarianz deutet auf die umgekehrte Tendenz hin. Die Kovarianz selbst ist aber noch kein gutes Maß für die Stärke dieser Tendenz. Um dies zu sehen, stellen wir uns vor, dass X die Körpergröße ist, x_i also eine Länge. Wenn man diese in Metern statt in Zentimetern angibt, ändern sich alle Größen x_i, μ_X und auch $\mathrm{Cov}[X, Y]$ um einen Faktor hundert. Man kann also nicht sagen, ob eine Kovarianz von 10.3 'groß' oder 'klein' ist. Man verwendet deshalb den Korrelationskoeffizienten als normiertes Maß, das von der gewählten Skala unabhängig ist. Für den Korrelationskoeffizienten gilt $-1 \leq \rho_{XY} \leq 1$. Die Extremwerte 1 und -1 bedeuten eine maximale positive oder negative Korrelation. Sie treten auf, wenn die ZVn mit einer Beziehung $Y = aX + b$ (a und b Konstanten) strikt voneinander abhängen.
Für *unabhängige* ZVn X und Y gilt stets, dass sie auch unkorreliert sind, denn dann ist

$$\sum_{i,j} x_i y_j\, p_{i,j} = \sum_{i,j} x_i y_j\, p_{X,i}\, p_{Y,j} = \big(\sum_i x_i\, p_{X,i}\big)\big(\sum_j y_j\, p_{Y,j}\big) = \mu_X\mu_Y$$

und damit $\mathrm{Cov}[X, Y] = 0$. Die Umkehrung gilt aber nicht: Die Kovarianz zwischen zwei ZVn kann auch verschwinden, wenn sie stochastisch abhängig sind.

2.1.3 Rechenregeln

Für Erwartungswerte, Varianzen und Kovarianzen gelten die folgenden Rechenregeln (a, b, c und d sind reelle Konstanten):

$$\begin{aligned}
\mathbf{E}[aX + c] &= a\mathbf{E}[X] + c \\
\mathbf{E}[X + Y] &= \mathbf{E}[X] + \mathbf{E}[Y] \\
\mathrm{Var}[aX + c] &= a^2 \,\mathrm{Var}[X] \\
\mathrm{Cov}[aX + c, bY + d] &= ab \,\mathrm{Cov}[X, Y] \\
\mathrm{Var}[X + Y] &= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X, Y] \\
\mathrm{Cov}[X + Y, Z] &= \mathrm{Cov}[X, Z] + \mathrm{Cov}[Y, Z]
\end{aligned}$$

Aus der vorletzten Zeile folgt insbesondere, dass sich die Varianz einer Summe von unkorrelierten ZVn gleich der Summe der Varianzen der einzelnen ZVn ist.

2.2 Diskrete Verteilungen

In den nächsten beiden Abschnitten werden wir einige häufig verwendete Verteilungen von ZVn genauer studieren. Dabei geht es jeweils darum den Realisierungen der ZV Wahrscheinlichkeiten zuzuordnen. Um die Wahrscheinlichkeit eines Ereignisses zu bestimmen, versucht man häufig zuerst, das Ereignis als Summe von Elementarereignissen auszudrücken, die alle die gleiche (bekannte) Wahrscheinlichkeit p_0 haben. Dann bekommt man die Wahrscheinlichkeit des Ereignisses einfach als p_0 mal der Zahl dieser Elementarereignisse.
Um zum Beispiel die Wahrscheinlichkeit zu bestimmen, mit zwei Würfeln die Augensumme 10 zu erzielen, macht man sich zuerst klar, dass es 36 mögliche Ausgänge des Zufallsexperiments gibt, die alle gleich wahrscheinlich sind, also Wahrscheinlichkeit $p_0 = 1/36$ haben. Von diesen Möglichkeiten gibt es genau drei, in denen das gewünschte Ereignis eintritt (nämlich (5,5), (6,4) und (4,6)). Die Wahrscheinlichkeit für die Augensumme 10 ist deshalb $p(10) = 3/36 = 1/12$.
Wahrscheinlichkeit hat also viel mit Zählen zu tun. Wir beginnen deshalb mit einem kurzen Abschnitt über Kombinatorik.

2.2.1 Elementare Kombinatorik

Kombinatorik ist die Kunst des Abzählens. Sie beginnt mit der Zahl der Möglichkeiten, n *unterscheidbare* Objekte *anzuordnen* (der Zahl der **Permutationen**): Diese ist gleich $n!$. Dabei bezeichnet $n!$ die **Fakultätsfunktion**:

$$\begin{aligned} n! &:= 1 \cdot 2 \cdot \ldots \cdot n \quad \text{für } n \geq 1, \\ 0! &:= 1\,. \end{aligned} \tag{2.83}$$

Gleichung (2.83) läßt sich folgendermaßen verstehen: Für den ersten Platz in der Reihe hat man noch alle n Objekte zur Auswahl; für den zweiten die jetzt noch übrigen $n-1$; usw., und für den letzten Platz gibt es nur noch eine einzige Möglichkeit. So erhält man den Ausdruck für $n!$ (von rechts nach links gelesen).
Mit derselben Überlegung findet man die Zahl der Möglichkeiten, aus den n Objekten k ($\leq n$) *auszuwählen und anzuordnen* als

$$n(n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}\,. \tag{2.84}$$

Wir suchen nun noch die Zahl der Möglichkeiten, k der n Objekte *ohne Berücksichtigung der Reihenfolge auszuwählen* (die Zahl der **Kombinationen**). Dazu überlegen wir zuerst, dass die (uns aus Gl. (2.84) schon bekannte) Zahl der Möglichkeiten, k aus n Objekten auszuwählen und anzuordnen, gleich der Zahl der Kombinationen (von k aus n) mal der Zahl der Permutationen von k Objekten ist; also ist die gesuchte Zahl der Kombinationen gleich

$$\frac{n!}{(n-k)!k!} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{1 \cdot 2 \cdot \ldots \cdot k} =: \binom{n}{k}. \tag{2.85}$$

$\binom{n}{k}$ heißt **Binomialkoeffizient**.

Beispiele:

- Es gibt $6! = 720$ Möglichkeiten, 6 Personen in eine Schlange zu stellen, und $6 \cdot 5 \cdot 4 \cdot 3 = \frac{6!}{(6-4)!} = 360$ Möglichkeiten, eine *Schlange* aus 4 Personen zu bilden, wenn man insgesamt 6 Personen zur Auswahl hat. Dagegen gibt es nur $\binom{6}{4} = \frac{6!}{4!\cdot 2!} = 15$ Möglichkeiten, eine *Gruppe* aus 4 Personen zu bilden, wenn man 6 Personen zur Auswahl hat. (Man beachte: Schlangen sind geordnet, Gruppen nicht!)

- Wenn es in einer Population an einem Gen-Locus n verschiedene Allele gibt, dann gibt es n verschiedene homozygote Genotypen und $\binom{n}{2} = n(n-1)/2$ Möglichkeiten für einen heterozygoten Genotyp (also $(n^2+n)/2$ verschiedene Genotypen überhaupt).

2.2.2 Die Binomialverteilung

Beispiel: Wir betrachten wie im letzten Abschnitt zwei für einen Gen-Locus heterozygote Eltern mit Genotyp Aa und nehmen an, dass Allel A dominant ist. Zum Beispiel könnte a ein Krankheitsgen einer rezessiv vererbten Krankheit sein, die nur im Homozygoten aa auftritt. Wir wissen schon, dass die Wahrscheinlichkeit für das Auftreten der Krankheit bei jedem Kind dann $1/4$ ist. Wir können nun weiter nach der Wahrscheinlichkeit fragen, dass zwei von vier Kindern – oder allgemein k von n Kindern – krank sind.

Fragen dieser Art führen auf die sogenannte Binomialverteilung. Betrachten wir allgemein ein Zufallsxperiment mit nur zwei möglichen Ausgängen, das Ereignis A und sein Gegenereignis $\bar{A}$. (Das heißt, uns interessieren nur ob A zutrifft oder nicht: z.B krank oder nicht krank.) Ein solches Experiment heißt **Bernoulliexperiment**. Wir definieren $p = Pr(A)$ (und damit $1-p = Pr(\bar{A})$). Wenn wir dieses Experiment mit gleichbleibender Wahrscheinlichkeit p für A n-mal wiederholen, können wir eine Zufallsvariable X definieren, die die Anzahl des Eintretens von A beschreibt. Dies bezeichnet man als **Binomialexperiment**. Uns interessiert die Verteilung von X, also die Wahrscheinlichkeit, dass das Ereignis $X = k$ eintritt.
Man kann sich dies in zwei Schritten überlegen. Zunächst wählt man aus der Menge von n Ergebnissen k aus, bei denen das Ereignis eingetreten sein soll, bei den anderen nicht. Wie wir oben gesehen haben gibt es $\binom{n}{k}$ Möglichkeiten, eine solche Auswahl zu treffen. Da die einzelnen Teilexperimente des Binomialexperiments unabhängig sind, ergibt sich die Wahrscheinlichkeit für jede dieser Möglichkeiten aus dem Produkt der Wahrscheinlichkeiten für die Teilexperimente: dabei erhalten wir k mal den Faktor p für das Eintreten von A auf und $n-k$ mal der Faktor $1-p$ für das Nicht-Eintreten. Die Wahrscheinlichkeit für $X = k$, also dass A genau k mal vorkommt, ist damit

$$Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}\,, \quad k = 0, \ldots, n\,. \tag{2.86}$$

Die durch (2.86) definierte Wahrscheinlichkeitsverteilung heißt **Binomialverteilung** mit Parametern n und p; s. Abb. 2.23. Man sagt, X ist $B(n,p)$ verteilt. Erwartungswert und Varianz der Binomialverteilung sind

$$\mu = np\,, \quad \sigma^2 = np(1-p)\,. \tag{2.87}$$

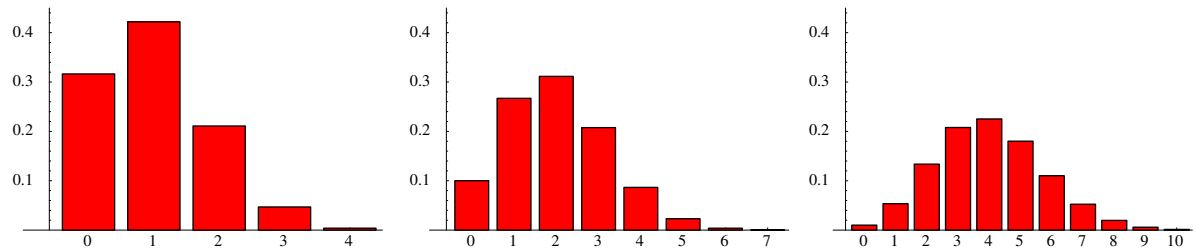


Abbildung 2.23: Binomialverteilung für $p = 1/4$ und $n = 4, 8$ und 16 (von links nach rechts).

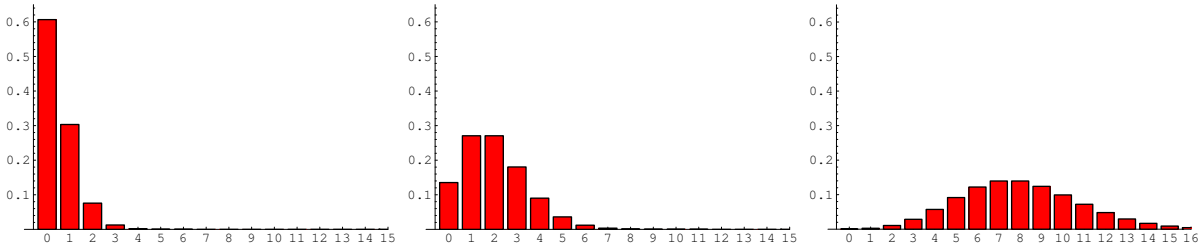


Abbildung 2.24: Poisson-Verteilung für Parameter $\lambda = 0.5$, $\lambda = 2$ und $\lambda = 8$ (entsprechend einer Zeit zum letzten gemeinsamen Vorfahren von $2,5 \cdot 10^5$, 10^6 und $4 \cdot 10^6$ Generationen bei einer Mutationsrate von $p = 10^{-6}$ für den untersuchten DNA Abschnitt).

2.2.3 Die Poisson-Verteilung

Eine häufige Situation bei einem Binomialexperiment ist der Fall mit 'riesigem' n und 'winzigem' p. In diesem Fall ist das Hantieren mit der Binomialverteilung unbequem, und man verwendet stattdessen die **Poissonverteilung** mit **Intensitätsparameter** $\lambda := np$. Man erhält sie aus der Binomialverteilung im Limes

$$n \to \infty, \quad np \equiv \lambda > 0\,. \tag{2.88}$$

Die Wahrscheinlichkeit, dass das Ereignis A genau $X = k$ mal eintritt, lautet dann (ohne Herleitung):

$$Pr(X = k) = \frac{\lambda^k}{k\,!}\, e^{-\lambda}, \quad k = 0, 1, 2, \ldots\,. \tag{2.89}$$

Man beachte, dass die Poissonverteilung – im Gegensatz zur Binomialverteilung – nur einen Parameter hat. Man kann sie deshalb auch dann benutzen, wenn man n und p nicht einzeln angeben kann, sondern nur ihr Produkt, λ – vorausgesetzt, man darf annehmen, dass n sehr groß und p sehr klein ist. Eine Faustregel ist, dass die Poissonverteilung $B(n, p)$ hinreichend gut approximiert, wenn sowohl $p \leq 0.1$ als auch $n \geq 30$ ist. Eine charakteristische Eigenschaft der Poisson-Verteilung ist, dass Erwartungswert und Varianz gleich sind:

$$\mu = \lambda\,, \quad \sigma^2 = \lambda\,. \tag{2.90}$$

Beispiel: Die molekulare Uhr

Um die phylogenetische Verwandschaft zwischen zwei Spezies zu klären, interessiert man sich für die Zeit, die seit dem letzten gemeinsamen Vorfahren beider Spezies verstrichen ist. In der molekularen Phylogenie untersucht man hierfür homologe Sequenzabschnitte nicht-kodierender DNA in beiden Spezies und zählt die abweichenden Basenpaare (die Anzahl der Substitutionen). Die Mutationswahrscheinlichkeit pro Basenpaar und Generation ist sehr klein ($\approx 10^{-9}$), für einen Abschnitt von z.B. 1000 Basenpaaren ist sie von der Größenordnung $p = 10^{-6}$. Wenn seit dem letzten gemeinsamen Vorfahren n Generationen verstrichen sind, dann liegen zwischen den beiden Spezies heute $2n$ Generationen. Wir nehmen an, dass die Mutationsrate über die gesamte Zeit konstant geblieben ist und dass die Wahrscheinlichkeit vernachlässigbar ist, dass eine Base zweimal mutiert. Dann ist die Anzahl der Substitutionen auf dem DNA Abschnitt poissonverteilt mit Parameter $\lambda = 2np$.

2.2.4 Die hypergeometrische Verteilung

Angenommen, in einer Population mit $N = 1000$ Individuen gibt es $K = 30$ mit einer bestimmten Mutation und $L = N - K = 970$ Wildtypen. Wenn man nun willkürlich eine Stichprobe der Größe n aus der Population herausgreift, was ist dann die Wahrscheinlichkeit, $X = k$ Mutanten (und $\ell = n - k$ Wildtypen) zu erwischen?
Dies ist die Frage nach der Häufigkeit eines Ereignisses in einer sogenannten 'Stichprobe ohne Zurücklegen'. Wenn wir uns vorstellen, dass wir die Individuen der Stichprobe der Reihe nach aus der Population auswählen, ist der wesentliche Unterschied zum Binomialexperiment, dass die Wahrscheinlichkeit in jedem Teilexperiment vom Ausgang der vorherigen Teilexperimente abhängt. Mit anderen Worten: Die Wahrscheinlichkeiten sind bedingt, die Teilexperimente nicht unabhängig. Man löst das Problem deshalb auf anderem Wege und allein durch Abzählen: Es gibt insgesamt $\binom{N}{n}$ gleichwahrscheinliche Möglichkeiten, n aus N Individuen auszuwählen. Ebenso gibt es $\binom{K}{k}$ Möglichkeiten, k aus den K Mutanten auszuwählen; für *jede* dieser Möglichkeiten gibt es wiederum $\binom{L}{\ell}$ Möglichkeiten, ℓ aus den L Wildtypen auszuwählen. Von den $\binom{N}{n}$ Möglichkeiten, die es insgesamt gibt, führen also gerade $\binom{K}{k}\binom{L}{\ell}$ zum Ergebnis k. Die gesuchte Wahrscheinlichkeit ist dann 'die Zahl der günstigen durch die Zahl der möglichen Fälle', also

$$Pr(X = k) = \frac{\binom{K}{k}\binom{L}{\ell}}{\binom{N}{n}}\,. \tag{2.91}$$

X heißt *hypergeometrisch* oder $H(n, K, N)-$verteilt. Erwartungswert und Varianz sind

$$\mu = \frac{n\,K}{N}\,, \quad \sigma^2 = n\,\frac{KL(N-n)}{N^2(N-1)}\,. \tag{2.92}$$

Beispiel: Da die hypergeometrische Verteilung die Verteilung in einer Stichprobe angibt, ist sie in der biologischen Anwendungen sehr häufig. Das Paradebeispiel ist aber die Gewinnchance beim Zahlenlotto. Bei '6 aus 49' sind von insgesamt $N = 49$ Kugeln sind $K = 6$ Kugeln markiert (das Ergebnis der Ziehung). $n = 6$ werden geraten (auf dem Lottoschein). Die Wahrscheinlichkeit, $k = 4$ richtige zu tippen, ist

$$Pr(X = 4) = \frac{\binom{6}{4}\binom{43}{2}}{\binom{49}{6}} = \frac{15 \cdot 903}{13983816} \simeq 0.000097\,.$$

2.2.5 Die geometrische Verteilung

Wir betrachten nun eine etwas andere Situation. Ein Bernoulli-Experiment mit Wahrscheinlichkeit p für das Ereignis A wird so lange wiederholt, bis A (zum ersten Mal) eintritt. Die Zufallsgröße X ist jetzt die *Anzahl der Experimente*. Die Wahrscheinlichkeit, dass man gerade i Versuche braucht, bis man das gewünschte Ereignis erhält, ist

$$Pr(X = i) = p\,(1-p)^{i-1}\,, \quad i = 1, 2, \ldots\,. \tag{2.93}$$

Gleichung (2.93) ist die sogenannte **geometrische Verteilung** mit Parameter p. Sie hat Erwartungswert und Varianz

$$\mu = \frac{1}{p}, \quad \sigma^2 = \frac{1-p}{p^2}\,. \tag{2.94}$$

Beispiel: Es wird mit einem idealen Würfel gewürfelt, bis eine 6 erscheint. Hier ist $p = 1/6$, also $\mu = 6$ und $\sigma^2 = 30$.

2.3 Kontinuierliche Verteilungen

Viele im täglichen Leben vorkommende Größen (z.B. die Körpergröße) können Werte aus einem Intervall auf der reellen Zahlengeraden (im Extremfall sogar alle reellen Zahlen) annehmen und sind somit kontinuierlich verteilt. Die Wahrscheinlichkeit, einen Wert *exakt* zu treffen, ist stets null: $Pr(X = x) = 0$. Die Definition der zu X gehörigen Verteilung ist daher nicht mehr über die Zuordnung $p_i = Pr(X = x_i)$ möglich. Wir haben uns bisher damit beholfen, dass wir die möglichen Werte in Klassen eingeteilt haben. Dies ist aber immer etwas ungenau und vor allem bei sehr vielen Klassen unhandlich. Besser ist die Beschreibung mittels einer sogenannten **Dichtefunktion**.

2.3.1 Dichtefunktionen

Wir betrachten die Wahrscheinlichkeitsverteilung einer kontinuierlichen ZV. Zur Darstellung in einem Säulendiagramm teilen wir die Realisierungen der ZV in Klassen ein. Die Wahrscheinlichkeit entspricht der Höhe der Säule (s. Abb. 2.25, links). Eine andere Möglichkeit ist es, die Wahrscheinlichkeit über den *Flächeninhalt* der Säule zu bestimmen (bei einer Breite der Säule von 1 ist beides dasselbe). In einem weiteren Schritt machen wir die Klasseneinteilung jetzt immer feiner und definieren die Wahrscheinlichkeit weiter über den Flächeninhalt der immer dünner werdenden Säulen. Im Grenzfall unendlich feiner Klasseneinteilung bekommt man dann die Dichtefunktion $f(x)$ als Hüllkurve $f(x)$ (s. Abb. 2.25).

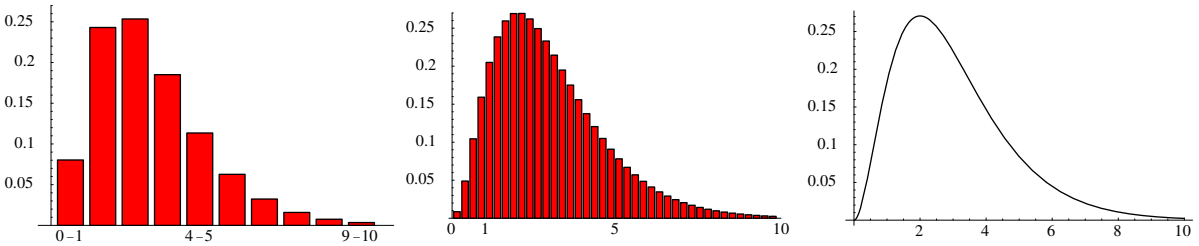


Abbildung 2.25: Klasseneinteilung für stetige ZV und Dichtefunktion, die man im Grenzfall unendlich feiner Klasseneinteilung erhält wenn man die Wahrscheinlichkeit als Flächeninhalt der Säule darstellt.

Dichtefunktionen haben nichtnegative Werte (also $f(x) \geq 0$ für alle $x \in \mathbb{R}$), und die Fläche unter der Kurve ist auf 1 normiert:

$$\int_{-\infty}^{\infty} f(x)dx = 1. \tag{2.95}$$

Die Wahrscheinlichkeit, dass X einen Wert zwischen a und b annimmt, ist gegeben durch

$$Pr(a \leq X \leq b) = Pr(X \in [a, b]) = \int_a^b f(x)dx; \tag{2.96}$$

s. Abb. 2.26. Insbesondere ist die Wahrscheinlichkeit, dass X nicht größer ist als ein Wert x,

$$F(x) := Pr(X \leq x) = \int_{-\infty}^{x} f(y)\, dy. \tag{2.97}$$

$F(x)$ heißt **Verteilungsfunktion** und gibt die Fläche unter der Dichtefunktion links von x an. Die Verteilungsfunktion ist eine Stammfunktion der Dichte $f(x)$, umgekehrt ist die Dichtefunktion deshalb die Ableitung der Verteilungsfunktion: $f(x) = F'(x)$. Wir werden häufig benutzen, dass

$$Pr(X \in [a, b]) = F(b) - F(a)\,. \tag{2.98}$$

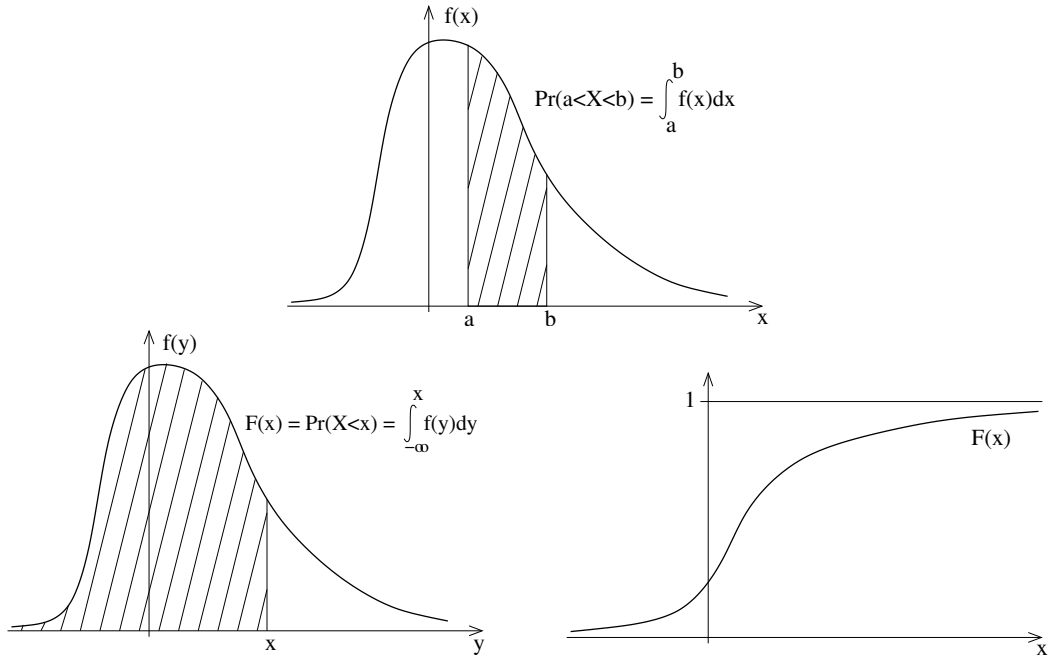


Abbildung 2.26: Dichtefunktion, Wahrscheinlichkeit und Verteilungsfunktion. Oben: Die Wahrscheinlichkeit, dass sich X zwischen a und b aufhält, ist gleich dem Flächeninhalt unter der Dichtefunktion $f(x)$ im Intervall $[a, b]$, s. Gl. (2.96). Unten: Die Verteilungsfunktion $F(x)$ gibt die Wahrscheinlichkeit an, dass X nicht größer ist als x. $F(x)$ ist Stammfunktion zu $f(x)$.

Zur Charakterisierung von Dichtefunktionen definiert man analog zum diskreten Fall den Erwartungswert, die Varianz und die Kovarianz. Statt der Summen treten nun Integrale auf. Der Erwartungswert einer kontinuierlich verteilten ZV X mit Dichte f ist

$$\mu = \mathbf{E}(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx. \tag{2.99}$$

Die Varianz ist

$$\sigma^2 = \mathrm{Var}[X] = \int_{-\infty}^{\infty} (x-\mu)^2\, f(x)\, dx \tag{2.100}$$

Für zwei kontinuierlich verteilte Zufallsvariablen X, Y mit gemeinsamer Dichtefunktion $f(x, y)$ ist

$$Pr(X \in [a,b], Y \in [c,d]) = \int_c^d \Big(\int_a^b f(x,y)dx \Big) dy \tag{2.101}$$

und die Kovarianz ist

$$\mathrm{Cov}[X,Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f(x,y) dx\, dy\,. \tag{2.102}$$

Für unabhängige ZVn faktorisiert die gemeinsame Dichte

$$f(x,y) = f(x)f(y) \tag{2.103}$$

und die Kovarianz ist Null. Alle in Abschnitt 2.1.3 angegebenen Rechenregeln gelten auch im kontinuierlichen Fall.

2.3.2 Gleichverteilung

Die einfachste kontinuierliche Verteilung ist die *Gleichverteilung*, bei der alle Werte in einem Intervall $[\alpha, \beta]$ gleich wahrscheinlich sind. Mit $\alpha = 0$ und $\beta = 20$ beschreibt sie z.B. die Wartezeit im S-Bahnhof, wenn die Züge pünktlich alle 20 min fahren und man zu einer zufälligen Zeit auf den Bahnsteig kommt. Die Dichtefunktion ist

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha}, & \alpha < x \leq \beta \\ 0, & \text{sonst.} \end{cases} \tag{2.104}$$

Die zugehörige Verteilungsfunktion lautet:

$$F(x) = \begin{cases} 0, & x \leq \alpha \\ \frac{x-\alpha}{\beta-\alpha}, & \alpha < x \leq \beta \\ 1, & x > \beta\,. \end{cases} \tag{2.105}$$

Der Erwartungswert ist

$$\mu = \frac{1}{\beta-\alpha}\int_\alpha^\beta x\,dx = \frac{1}{2}\frac{1}{\beta-\alpha}[x^2]_\alpha^\beta = \frac{1}{2}\frac{1}{\beta-\alpha}(\beta^2-\alpha^2) = \frac{1}{2}(\alpha+\beta) \tag{2.106}$$

und die Varianz

$$\begin{aligned} \sigma^2 &= \frac{1}{\beta-\alpha}\int_\alpha^\beta x^2 dx - \mu^2 = \frac{1}{3}\frac{\beta^3-\alpha^3}{\beta-\alpha} - \frac{1}{4}(\alpha+\beta)^2 \\ &= \frac{1}{12(\beta-\alpha)}(4\beta^3 - 4\alpha^3 - 3\alpha^2\beta + 3\alpha^3 - 6\alpha\beta^2 + 6\alpha^2\beta - 3\beta^3 + 3\alpha\beta^2) \\ &= \frac{1}{12(\beta-\alpha)}(\beta^3 - \alpha^3 + 3\alpha^2\beta - 3\alpha\beta^2) = \frac{1}{12(\beta-\alpha)}(\beta-\alpha)^3 = \frac{1}{12}(\beta-\alpha)^2. \end{aligned} \tag{2.107}$$

2.3.3 Normalverteilung

Die bei weitem wichtigste kontinuierliche Verteilung ist die **Normalverteilung** (oder Gauß-Verteilung). Eine stetige Zufallsgröße X heißt normalverteilt mit Erwartungswert μ und Varianz σ^2 (kurz: $N(\mu, \sigma^2)$), wenn sie durch die Dichtefunktion

$$f(x) = \frac{1}{\sqrt{2\,\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{2.108}$$

beschrieben wird. Die überwiegende Mehrheit biologischer Messgrößen, wie Körperlänge, Gewicht, Gerbstoffgehalt, Enzymaktivität usw., ist in guter Näherung normalverteilt.
Wir betrachten zunächst eine spezielle Wahl der Parameter, nämlich $\mu = 0$ und $\sigma^2 = 1$. Die zugehörige Verteilung, $N(0, 1)$, heißt **Standardnormalverteilung**. Eine standardnormalverteilte Zufallsvariable Z hat Verteilungsfunktion

$$\Phi(z) := Pr(Z \leq z) = \frac{1}{\sqrt{2\,\pi}}\int_{-\infty}^{z}\exp\left(-\frac{t^2}{2}\right)dt; \tag{2.109}$$

die Wahrscheinlichkeit, dass $Z \in [a, b]$ ist, ist dann $\Phi(b) - \Phi(a)$. Für das Integral auf der rechten Seite von (2.109) gibt es aber keine geschlossene Formel; man muss die Werte für Φ deshalb in einer Tabelle nachschauen oder vom Computer ausrechenen lassen. Die Normalverteilungstabelle auf Seite 90 gibt zu vorgebenem z die Fläche der Standardnormalverteilung links von z an, d.h.

$\Phi(z) = Pr(Z \leq z)$. In den Tabellen sind üblicherweise nur Werte $z \geq 0$ erfasst. Für $z < 0$ ergibt die Symmetrie der Verteilung

$$\Phi(-z) = 1 - \Phi(z). \tag{2.110}$$

Wir lösen uns nun von der Einschränkung auf die Standardnormalverteilung und betrachten eine Zufallsgröße $X \sim N(\mu, \sigma^2)$ mit beliebigem μ und σ^2. Für die zugehörige Verteilungsfunktion gibt es keine Tabellen; man muss die Situation daher auf die Standardnormalverteilung zurückführen. Das gelingt mit Hilfe der sogenannten **Standardisierung**: Man ersetzt X durch $Z := \frac{X-\mu}{\sigma}$, und entsprechend x durch $z := \frac{x-\mu}{\sigma}$. Z ist nun standardnormalverteilt, und $Pr(X \leq x) = Pr(Z \leq z) = \Phi(z)$. Daraus ergibt sich

$$Pr(a < X \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \tag{2.111}$$

Beispiel: Die Länge X (in μm) von Pantoffeltierchen sei $N(200, 100)$ verteilt. Die Wahrscheinlichkeit, dass ein Tierchen zwischen 190 und 210 μm lang ist, ist dann

$$\begin{aligned} Pr(X \in [190, 210]) &= \Phi\Big(\frac{210-200}{10}\Big) - \Phi\Big(\frac{190-200}{10}\Big) \\ &= \Phi(1) - \Phi(-1) = 2\,\Phi(1) - 1 = 2 \cdot 0.8413 - 1 \simeq 0.68. \end{aligned}$$

mit dem Tabellenwert $\Phi(1) = 0.8413$ (s. S. 90).

Eine wichtige Eigenschaft der Normalverteilung ist, dass die Summe von unabhängigen, normalverteilten Zufallsvariavblen wieder normalverteilt ist.

Additionssatz der Normalverteilung: *Sind X_i, $i = 1, \ldots, n$, unabhängige, normalverteilte ZV, wobei X_i nach $N\big(\mu_i, \sigma_i^2\big)$ verteilt ist, so ist ihre Summe normalverteilt, mit Erwartungswert $\mu = \sum_i \mu_i$ und Varianz $\sigma^2 = \sum_i \sigma_i^2$,*

$$X = \sum_i X_i \sim N\Big(\textstyle\sum_i \mu_i, \sum_i \sigma_i^2\Big). \tag{2.112}$$

2.3.4 Der Zentrale Grenzwertsatz (ZGS)

Die fundamentale Bedeutung der Normalverteilung ist die Folge des *zentralen Grenzwertsatzes*. Seine Aussage erweitert den Additionssatz auf die Summe von unabhängigen ZV mit ganz beliebigen Verteilungen.

Zentraler Grenzwertsatz: *Seien X_i, $i = 1, \ldots, n$, unabhängige ZV mit Erwartungswerten μ_i und Varianzen σ_i^2. Dann ist ihre Summe $X = \sum_i X_i$ für großes n näherungsweise normalverteilt mit Erwartungswert $\mu = \sum_i \mu_i$ und Varianz $\sigma^2 = \sum_i \sigma_i^2$,*

$$X = \sum_{i=1}^{n} X_i \simeq N\Big(\textstyle\sum_i \mu_i, \sum_i \sigma_i^2\Big). \tag{2.113}$$

Die meisten Größen, die man in der Natur beobachtet, werden von einer großen Zahl zufälliger Faktoren beeinflusst. Phänotypische Merkmale z.B. sind meistens das Resultat zahlreicher Umweltfaktoren und der Beiträge vieler Gene. Solange nicht einige wenige Beiträge dominieren, kann

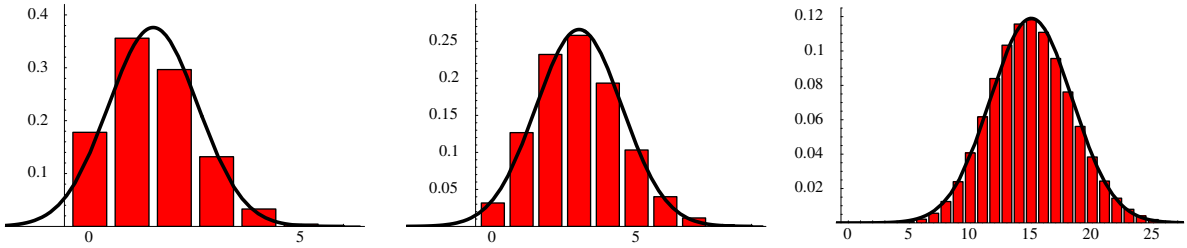


Abbildung 2.27: Approximation der Binomialverteilung durch die Normalverteilung. Exakte Verteilung (dargestellt als Histogramm) für $p = 1/4$ und $n = 6, 12, 60$ (von links nach rechts), und deren Approximation durch die Normalverteilung $N(np, np(1-p))$.

man in guter Näherung annehmen, dass die Faktoren additiv sind. Dann ist das phänotypische Merkmal X

$$X = \sum_i X_i \tag{2.114}$$

wobei die Zufallsvariablen X_i die Beiträge der einzelnen Faktoren darstellen. Die Verteilungen der X_i sind normalerweise unbekannt. Wichtig ist aber nur, dass die Beiträge unabhängig sind. Der ZGS sorgt dann dafür, dass die Summe (das phänotypische Merkmal X) näherungsweise normalverteilt ist. Der ZGS erlaubt auch die Approximation anderer Verteilungen durch die Normalverteilung, und zwar immer dann, wenn die Ereignisse aus hinreichend vielen unabhängigen Einzelereignissen zusammengesetzt sind. Darüber gibt die folgende Liste Auskunft.

Beispiele und 'Faustregeln' für die Normalapproximation:

1. Die Binomialverteilung $B(n, p)$ ist die Verteilung der Summe n unabhängiger Bernoulliexperimente mit Erfolgswahrscheinlichkeit μ. Für hinreichend große n kann man die Binomialverteilung durch die Normalverteilung approximieren. Als Faustregel gilt: Ist X binomialverteilt mit $\sigma^2 = np(1-p) \geq 9$, so ist X näherungsweise normalverteilt nach $N(np, np(1-p))$. Eine Illustration findet sich in Abb. 2.27.

2. Auch die Poissonverteilung geht für hinreichend großem Parameter λ in die Normalverteilung über: Ist X Poisson-verteilt mit $\lambda = \mu = \sigma^2 \geq 10$, so ist X näherungsweise normalverteilt nach $N(\lambda, \lambda)$.

3. Bei der hypergeometrischen Verteilung $H(n, K, N)$ sind die Ziehungen der einzelnen Elemente einer Stichprobe *nicht* unabhängig. Sie geht deshalb bei einer großen Stichprobengröße n i.A. *nicht* in die Normalverteilung über. Nur wenn die Gesamtmenge N (z.B. Populationsgröße) sehr viel größer ist als die Stichprobe sind die Ziehungen "fast" unabhängig. Dann geht die hypergeometrische Verteilung in die Binomialverteilung über und kann (wenn die Stichprobe groß genug ist) wieder durch die Normalverteilung approximiert werden: Ist X hypergeometrisch verteilt mit $\frac{n}{N} \leq 0.05$ und $n\frac{K}{N}(\frac{N-K}{N}) \geq 9$, so ist X näherungsweise normalverteilt nach $N(\mu, \sigma^2)$ mit $\mu = n(K/N)$ und $\sigma^2 = n\frac{K}{N}(\frac{N-K}{N})$.

2.4 Grundlagen der Statistik

In den letzten Abschnitten haben wir Zufallsvariablen (Messgrößen) mit festen, als bekannt angenommenen Wahrscheinlichkeitsverteilungen oder Dichtefunktionen betrachtet. Aus diesen haben wir Wahrscheinlichkeiten für abgeleitete Ereignisse berechnet, wie zum Beispiel das k-fache Auftreten eines Ereignisses in einer Stichprobe. Ein Gebiet der Biologie, für das diese Betrachtungsweise relevant ist, ist die Mendel'sche Verebungslehre. In den allermeisten Fällen ist die Zielrichtung aber andersherum: Man kennt die Verteilung der Zufallsvariable nicht, oder zumindest nicht vollständig, hat aber eine Stichprobe und möchte daraus Aussagen über die unbekannte Verteilung gewinnen. Dies ist Gegenstand der mathematischen Statistik – und des letzten Teils dieser Vorlesung, der jetzt in Angriff genommen wird.

2.4.1 Messwerte und ihre Darstellung

Wir betrachten ein Merkmal (eine beliebige Messgröße) X als die Zufallsvariable, deren Verteilung wir bestimmen wollen. Empirisch tut man dies, indem man Wahrscheinlichkeiten durch relative Häufigkeiten annähert. Hierfür führt man die Messung von X (das Zufallsexperiment) wiederholt aus. Wesentlich ist dabei, dass die Messungen *unabhängig* sind. Man erhält eine Menge von n Messwerten, $x_1, x_2, \ldots, x_n$, die man auch als *Stichprobe* vom Umfang n bezeichnet. Die x_i sind Realisierungen der Zufallsvariablen X.
Man unterscheidet verschiedene Typen von Messwerten. Hierfür werden Merkmale in sogenannte *Skalenniveaus* eingeteilt, die wir hier in der Reihenfolge aufsteigender *Datengüte* aufführen. Vom Skalenniveau hängt ab, welche Werkzeuge später für die Behandlung der Daten eingesetzt werden können.

1. **Nominalskala:** Die Realisierungen nominaler Merkmale besitzen keine Rangordnung; es geht um qualtitative Eigenschaften (z.B. Farbe, Geschlecht).

2. **Ordinalskala:** Die Realisierungen haben eine Rangordnung; man kann für je zwei Werte sagen, welcher der größere ist, kann aber die Größe des Unterschieds nicht genau quantifizieren. So ist klar, dass die Ordnungsrelation *kleines Bier* < *mittleres Bier* < *großes Bier* besteht. Die genaue Mengendifferenz ist durch diese Ordnungsrelation jedoch nicht erfasst. Der Mittelwert hat für ordinale Daten keine Bedeutung.

3. **Intervallskala:** Hier hat man klare quantitative Angaben; man kann Differenzen bilden, und gleiche Differenzen bedeuten gleiche Abstände (z.B. Temperatur; Kontostand). Bei intervallskalierten Daten hat der Mittelwert eine Bedeutung.

Hat man seine Daten erhoben, so kann man sie in Tabellenform oder graphisch darstellen.

1. **Urliste:** Ungeordnete Form von Messdaten $(x_1, x_2, \ldots, x_n)$, z.B. in der Reihenfolge, in der sie gemessen wurden.

2. **Häufigkeitstabelle:** Die Messwerte werden geordnet, und gleiche Werte werden zusammengefasst ($\xi_1 < \xi_2 \ldots < \xi_r$; $r \leq n$). In einer Tabelle wird notiert, wie häufig die jeweiligen Werte beobachtet wurden. Die Anzahl n_j des Auftretens von ξ_j heißt *absolute Häufigkeit*, während $h_j := \frac{n_j}{n}$ dessen *relative Häufigkeit* ist. Natürlich ist $\sum_{j=1}^{r} n_j = n$ und $\sum_{j=1}^{r} h_j = 1$.

3. **Säulendiagramm:** graphische Darstellung von n_j (oder h_j) versus ξ_j. Oft ist es sinnvoll, Messwerte zu größeren Klassen zusammenzufassen. Werden alle Intervalle gleich breit gewählt, so spricht man von einem *Histogramm*. Histogramme dienen der diskreten Approximation der Wahrscheinlichkeitsverteilung durch empirische Daten.

Beispiel: Es wird die Zahl der BOHNEN PRO SCHOTE bei $n = 50$ Bohnenhülsen ermittelt.

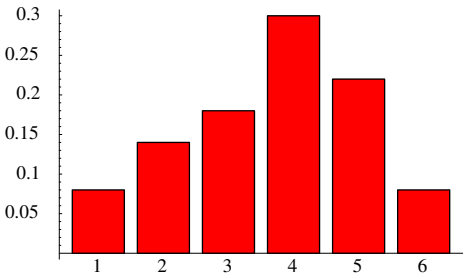
Urliste:

1, 2, 1, 4, 5, 3, 5, 4, 5, 5, 4, 4, 4, 4, 4, 3, 6, 2, 2, 6, 3, 6, 4, 5, 1,
3, 4, 4, 5, 5, 2, 4, 2, 5, 3, 1, 3, 4, 5, 5, 4, 3, 3, 4, 4, 5, 6, 2, 2, 3

Häufigkeitstabelle:

ξ_j	1	2	3	4	5	6
n_j	4	7	9	15	11	4
h_j	0.08	0.14	0.18	0.3	0.22	0.08

Histogramm:



In einer "unendlich großen" Stichprobe kann man alle Aspekte der Wahrscheinlichkeitsverteilung der zugrundeliegenden ZV über die relativen Häufigkeiten beliebig genau abschätzen. In der Realität ist der Stichprobenumfang aber (oft sehr) begrenzt und man muss sich auf die Schätzung einiger zentraler Eigenschaften der Verteilung konzentrieren. Für die Wahrscheinlichkeitsverteilung eines intervallskalierten Merkmals sind der Erwartungswert μ und die Varianz σ^2 die wichtigsten Bestimmungsgrößen. Insbesondere ist Normalverteilung durch μ und σ^2 vollständig festgelegt. Wir beginnen deshalb mit der Schätzung dieser beiden Größen aus einer Stichprobe.

2.4.2 Schätzung des Erwartungswertes

Gegeben sei eine Stichprobe $(x_1, x_2, \ldots, x_n)$ unabhängiger intervallskalierter Messdaten. Die Verteilung der zugrundeliegenden Zufallsvariablen X und insbesondere ihren Erwartungswert μ kennen wir nicht, er soll aus der Stichprobe geschätzt werden.
Der Schätzwert für μ, der sich automatisch nahelegt ist der **Mittelwert** der Stichprobe

$$\overline{x} = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{j=1}^{r} \xi_j h_j. \tag{2.115}$$

In der Sprache der Statistik spricht man von einem **Schätzwert** oder **Schätzer** des Erwartungswertes (der Hut $\hat{}$ ist zu lesen als 'Schätzer für'). Beim BOHNENBEISPIEL (s.o.) ist der Mittelwert $\overline{x} = 3.68$ (Bohnen pro Schote).
Im Gegensatz zu μ hängt $\hat{\mu}$ von der zufällig gewählten Stichprobe ab. Wir können uns vorstellen, dass wir wiederholt Stichproben vom Umfang n ziehen und jeweils den Mittelwert bilden. Dann ist der Mittelwert selbst eine Zufallsvariable, deren Verteilung wir analysieren wollen. Hierfür betrachten wir die unabhängigen Messungen nicht als n Realisierungen derselben ZV X, sondern unterscheiden n ZVn $X_1, \ldots, X_n$ für die Ergebnisse der n Teilexperimente. Die X_i sind *unabhängig* und *identisch verteilt* (abgekürzt **i.i.d.** für 'independent and identically distributed'). 'Identisch verteilt' heißt, sie haben alle dieselbe Wahrscheinlichkeitsverteilung, nämlich die von X. Mit dieser Interpretation ist $\bar{x}$ eine Realisierung der Zufallsvariablen

$$\bar{X} := \frac{1}{n}\sum_{i} X_i. \tag{2.116}$$

Ihr Erwartungswert und ihre Varianz sind

$$\mathbf{E}(\bar{X}) = \mathbf{E}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu\,, \tag{2.117}$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\Big(\frac{1}{n}\sum_{i} X_i\Big) = \frac{1}{n^2}\mathrm{Var}\Big(\sum_{i} X_i\Big) = \frac{1}{n^2}\sum_{i}\mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i}\sigma^2 = \frac{1}{n}\,\sigma^2\,, \tag{2.118}$$

wobei wir die Rechenregeln für Erwartungswerte und Varianzen aus 2.1.3 (insbesondere die Additivität von Varianzen unabhängiger ZVn) und die identische Verteilung der X_i verwendet haben. Wir machen die folgenden Beobachtungen:

- Der Erwartungswert von $\bar{X}$ stimmt mit dem Erartungswert von X überein. Das heißt, wenn wir "unendlich viele" Mittelwerte von Stichproben vom Umfang n bestimmen, erhalten wir im Mittel den 'wahren' Erwartungswert μ der zugrundeliegenden Verteilung. Man bezeichnet den Stichprobenmittelwert als Schätzer des Erwartungswerts deshalb als **erwartungstreu** oder **unverzerrt** (*unbiased*).
- Die Varianz von $\bar{X}$ wird mit wachsendem Stichprobenumfang immer kleiner. Gemeinsam mit der Erwartungstreue besagt dies, dass die Schätzung mit zunehmendem Stichprobenumfang immer 'besser' wird in dem Sinne, dass sie den wahren Wert immer besser trifft. Man sagt, der Mittelwert ist ein **konsistenter** Schätzer für den Erwartungswert.

Für große n können wir sogar eine Aussage über die gesamte Verteilung des Mittelwerts machen. Da $\bar{X}$ als Summe unabhängiger ZV definiert ist, ist dies eine Konsequenz des zentralen Grenzwertsatzes:

Zentraler Grenzwertsatz der Statistik *Seien $X_1, \ldots, X_n$ unabhängig und identisch verteilte ZVn mit Erwatungswert μ und Varianz σ^2. Dann ist $\bar{X} = \frac{1}{n}\sum_i X_i$ für großes n näherungsweise normalverteilt mit Erwatungswert μ und Varianz σ^2/n, oder:*

$$\lim_{n\to\infty} Pr\left(\frac{(\bar{X}-\mu)\sqrt{n}}{\sigma} \le z\right) = \Phi(z). \tag{2.119}$$

Unabhängig von der Verteilung der zugrundeliegenden ZV X sind die Mittelwerte von Stichproben bei hinreichend großem Umfang n stets normalverteilt.

2.4.3 Schätzung der Varianz

Der gängige Schätzer für die Varianz σ^2 der Verteilung von X ist die sogenannte **empirische Varianz**

$$s^2 = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\left(\Big(\sum_{i=1}^{n} x_i^2\Big) - n\,\overline{x}^2\right); \tag{2.120}$$

$s := \sqrt{s^2}$ ist die zugehörige empirische Standardabweichung. Im obigen Beipiel der Bohnen ist die empirische Varianz $s^2 = 1.936$ (Bohnen pro Schote)^2, und die Standardabweichung $s = 1.392$ (Bohnen pro Schote).
Hierbei fällt auf, dass durch $n-1$ und nicht durch n geteilt wird. Der Grund ist, dass man σ^2 erwartungstreu schätzen möchte. Analog zur Vorgehensweise beim Mittelwert können wir s^2 als Realisierung der Zufallsgröße $S^2 := (1/(n-1))\sum_i (X_i - \bar{X})^2$ auffassen. Für sie gilt (ohne Rechnung) $\mathbf{E}(S^2) = \sigma^2$, und $\lim_{n\to\infty}\mathrm{Var}(S^2) = 0$. Die empirische Varianz ist somit eine erwartungstreue und konsistente Schätzung der Varianz σ^2 der der Stichprobe zugrundeliegenden

Verteilung – also wieder eine Schätzung mit 'guten' Eigenschaften. Die naheliegende Wahl, den Schätzer als Realisierung von $\tilde{S}^2 := (1/n)\sum_i (X_i - \bar{X})^2$ zu definieren, führt dagegen auf eine im Mittel zu kleine Schätzung, denn (ohne Rechnung)

$$\mathbf{E}(\tilde{S}^2) = \mathrm{Var}(X) - \mathrm{Var}(\bar{X}) = \big(1 - (1/n)\big)\sigma^2 \,.$$

Die Ursache ist, dass $\tilde{S}^2$ die Schwankung um den *Stichprobenmittelwert* misst, während eigentlich die Schwankung um den wahren (aber unbekannten) Erwartungswert μ geschätzt werden soll. Da der Mittelwert aus der Stichprobe selber ermittelt wurde, fällt die Schwankung um ihn etwas geringer aus als die um μ. $\tilde{S}^2$ liefert zwar auch einen Schätzer für die Varianz, dieser ist aber nicht erwartungstreu.

2.4.4 Schätzung weiterer Parameter der Verteilung

Erwartungswert und Varianz sind die beiden wichtigsten Kenngrößen einer Verteilung. Man bezeichnet sie auch als **Momente**; Mittelwert und empirische Varianz werden entsprechend auch **Momentenschätzer** genannt. Momentenschätzer kann man auch benutzen, um weitere unbekannte Parameter von Verteilungen zu schätzen, etwa das p eines Bernoulliexperiments, oder den Intensitätsparameter der Poisson-Verteilung. Zu diesem Zweck werden die Beziehungen zwischen dem zu schätzenden Parameter und den Momenten der Verteilung benutzt, um den Parameter als Funktion der Momente auszudrücken. Die Momente werden dann durch die Momenten*schätzungen* ersetzt; so ergibt sich ein Schätzwert für den Parameter.

Beispiele:

1. Ist X binomialverteilt mit bekanntem n und unbekanntem p (Wahrscheinlichkeit für das Eintreten von A). Dann ist p der Erwartungswert der identisch verteilten Zufallsvariablen X_i, die den Wert 1 annehmen, wenn A im $i$ten Teilexperiment eintritt, und 0 sonst. Wenn n_A die Zahl der Experimente mit Ergebnis A ist, dann ist

   $$\hat{p} = \frac{n_A}{n}, \tag{2.121}$$

   eine Realisierung der ZV $(1/n)\sum_i X_i$ und ein erwartungstreuer und konsistenter Schätzer für p.

2. Sei X eine Poisson-verteilte ZV und $(x_1, \ldots, x_n)$ eine Stichprobe, wobei wir den Parameter λ nicht kennen. Dann ist

   $$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{2.122}$$

   erwartungstreuer, konsistenter Schätzer des Parameters λ. Achtung: Es ist nicht zu empfehlen, anstelle des Mittelwerts die empirische Varianz zur Schätzung von λ zu verwenden, denn der resultierende Schätzer hat eine höhere Varianz als der Schätzer aus (2.122).

2.4.5 Zwei Zufallsvariablen und Kovarianzschätzung

Häufig ist man nicht allein an der Verteilung einer Messgröße interessiert, sondern am Zusammenhängen zwischen verschiedenen Größen X und Y. Gemessen wurde z.B. für $n = 18$ Personen die Körpergröße x_i und das Körpergewicht y_i:

x_i/y_i : 187/79, 184/81, 166/53, 182/64, 168/52, 170/55, 165/60, 166/51, 184/79, 176/60, 170/56, 175/70, 190/85, 185/80, 171/53, 185/74, 170/66, 187/84.

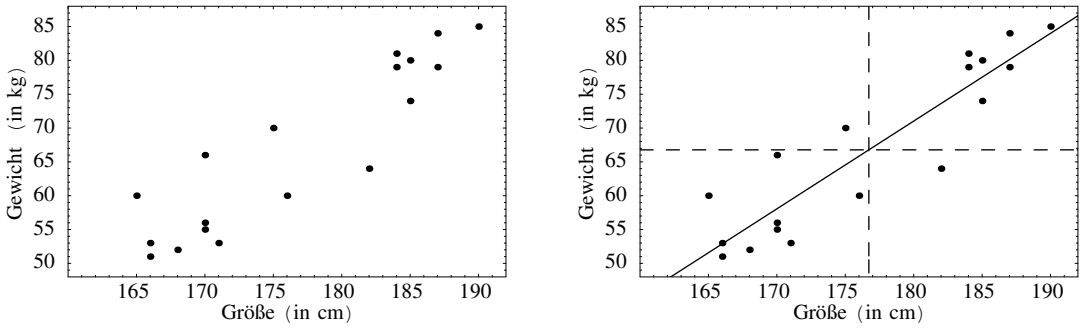


Abbildung 2.28: Links: Scatterplot für das Größe–Gewicht Beispiel. Rechts: Scatterplot mit Mittelwerten (gestrichelt) und Regressionsgeraden.

Solche Daten lassen sich übersichtlich in einem *Scatterplot* darstellen, in dem jedes (x_i, y_i)-Paar als Punkt in der x, y-Ebene eingetragen wird, s. Abb. 2.28.
Als Maß für den Zusammenhang zwischen zwei ZV hatten wir in Abschnitt 2.1.3 die Kovarianz und den Korrelationskoeffizienten definiert. Die zugehörigen erwartungstreuen und konsistenten Schätzer sind die *empirische Kovarianz*,

$$s_{xy} := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\Big((\sum_{i=1}^{n} x_i y_i) - n\bar{x}\bar{y}\Big), \tag{2.123}$$

und der *empirische Korrelationskoeffizient* $\hat{\rho}_{xy}$,

$$\hat{\rho}_{xy} := \frac{s_{xy}}{s_x \cdot s_y}, \tag{2.124}$$

wobei $\bar{x}$, $\bar{y}$ die Mittelwerte und s_x, s_y die empirischen Standardabweichungen von X und Y sind. Allgemein charakterisiert das Vorzeichen von s_{xy} bzw. $\hat{\rho}_{xy}$ die Form der Punktwolke im Scatterplot (Abb. 2.29).

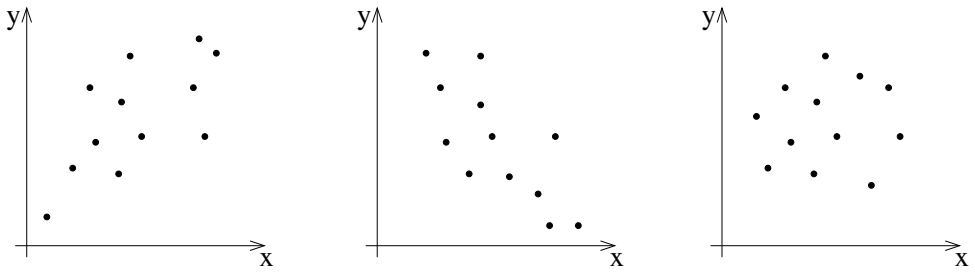


Abbildung 2.29: Daten mit positiver (links), negativer (Mitte) und verschwindender (rechts) Kovarianz bzw. Korrelation.

Eine wesentliche Eigenschaft der "wahren" Kovarianz war insbesondere, dass sie für unabhängige ZV gleich null ist. Für die Schätzwerte s_{xy} (und $\hat{\rho}_{xy}$) gilt deshalb, dass sie für unabhängige ZV in einer hinreichend großen Stichprobe nahe 0 sein sollten, s. Abb. 2.29. Besteht dagegen ein strikt linearer Zusammenhang zwischen den beiden ZV, also Korrelation $\rho_{XY} = 1$ (oder $\rho_{XY} = -1$), dann liegen auch in jeder Stichprobe alle Punkte im *Scatterplot* auf einer Geraden mit positiver (oder negativer) Steigung und die empirische Kovarianz ist $\hat{\rho}_{xy} = 1$ (bzw. $\hat{\rho}_{xy} = -1$).

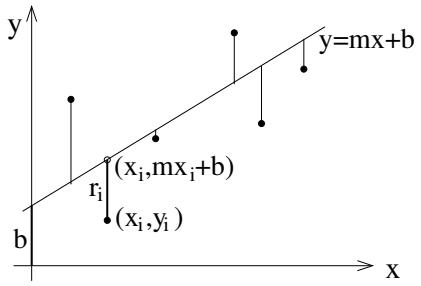


Abbildung 2.30: Methode der kleinsten Quadrate zur Ermittlung der Regressionsgeraden. Die vertikalen Linien geben den Abstand r_i der Messpunkte (x_i, y_i) von der Geraden an.

Im KÖRPERGRÖSSE/KÖRPERGEWICHT-BEISPIEL ist $\bar{x} = 176.72$, $\bar{y} = 66.78$, $s_x^2 = 74.88$ und $s_y^2 = 151.11$. Man erhält $s_{xy} = 97.17$ und $\rho_{xy} = \frac{97.17}{\sqrt{74.88 \cdot 151.11}} = 0.913$, Größe und Gewicht sind also in der Sichprobe stark positiv korreliert. Dies ist ein Hinweis darauf, dass zwischen X und Y auch "in Wahrheit" (in einer unendlich großen Stichprobe) ein positiver Zusammenhang besteht.

2.4.6 Lineare Regression

Auch wenn die Korrelation zwischen zwei intervallskalierten ZV X und Y nicht total ist ($\rho_{xy} = \pm 1$), die Punkte im Scatterplot also nicht genau auf einer Geraden liegen, möchte man einen linearen Zusammenhang zwischen X und Y durch eine Geradengleichung beschreiben. Anschaulich möchte man in der (x, y)-Ebene eine Gerade so gut wie möglich an die n Messpunkte anpassen.
Gegeben sind n Paare von Messdaten $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Gesucht sind die Parameter m (Steigung) und b ($y-$Achsenabschnitt) der Geradengleichung

$$y = f_{m,b}(x) = mx + b. \tag{2.125}$$

m und b sollen so bestimmt werden, dass die Messpunkte möglichst wenig von der Geraden abweichen. Die Abstände der Messpunkte von der Geraden sind die sogenannten *Residuen*:

$$r_i := y_i - f_{m,b}(x_i)\,. \tag{2.126}$$

siehe Abb. 2.30. In der **Methode der kleinsten Quadrate** sucht man diejenigen Werte für m und b, für die die Summe der Abweichungs*quadrate* am kleinsten wird. Zu diesem Zweck minimiert man die Funktion

$$\Delta(m, b) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \big(y_i - f_{m,b}(x_i)\big)^2\,. \tag{2.127}$$

$\Delta(m, b)$ ist minimal für

$$b = \overline{y} - m\overline{x} \quad \text{und} \quad m = \frac{s_{xy}}{s_x^2} = \rho_{xy}\frac{s_y}{s_x}\,. \tag{2.128}$$

Die ermittelte Gerade heißt *Regressionsgerade*. Sie ist steigend (fallend), wenn $\rho_{xy} > 0$ ($\rho_{xy} < 0$). m nennt man auch den *Regressionskoeffizienten*. Aus der Gleichung für b liest man ab, dass der Punkt $(\overline{x}, \overline{y})$ immer auf der Regressionsgeraden liegt. Für das KÖRPERGRÖSSE/KÖRPERGEWICHT-BEISPIEL erhält man $m = 1.297$ und $b = -162.4$; s. auch Abb. 2.28 (rechts).

2.5 Konfidenzintervalle

Im letzten Abschnitt haben wir für unbekannte Parameter einer Verteilung, wie den Erwartungswert, einzelne Schätzwerte ermittelt. Um beurteilen zu können, wie gut eine solche **Punktschätzung** den wahren Parameterwert trifft, würde man gerne ein Intervall angeben, in dem dieser Wert mit einer gewissen Sicherheit liegt. Dies führt auf die Definition von sogenannten **Konfidenzintervallen**. Wir benötigen hierfür einige Vorbereitung und gehen in mehreren Schritten vor.

2.5.1 Ableitung

1. Schritt: Wir betrachten zunächst die Wahrscheinlichkeit, eine standardnormalverteilte Zufallsvariable Z in einem *vorgegebenen symmetrischen Intervall* $[-c, c]$ anzutreffen. Das ist

$$\Pr(Z \in [-c, c]) = 2\Phi(c) - 1\,. \tag{2.129}$$

2. Schritt: Es soll nun umgekehrt ein symmetrisches Intervall so bestimmt werden, dass die Werte von Z mit *vorgegebener Wahrscheinlichkeit* $1-\alpha$ hineinfallen (bzw. mit Wahrscheinlichkeit α herausfallen). Dazu muss die Gleichung

$$2\Phi(c) - 1 = 1 - \alpha \;\Leftrightarrow\; \Phi(c) = 1 - \frac{\alpha}{2} \tag{2.130}$$

nach c aufgelöst werden. Da es keine explizite Formel für Φ gibt, muss der Wert der Normalverteilungstabelle entnommen werden. Allgemein bezeichnet man denjenigen Wert z_q, für den

$$\Phi(z_q) = q \tag{2.131}$$

gilt, als q-**Quantil** der Standardnormalverteilung. Damit ergibt sich

$$\Pr(Z \in [-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}]) = 1 - \alpha. \tag{2.132}$$

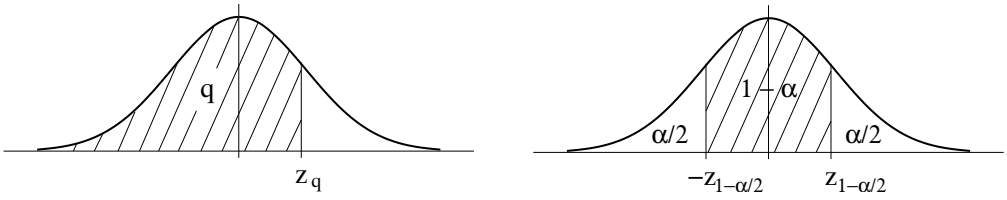


Abbildung 2.31: q-Quantil der Standardnormalverteilung und Intervall in dem die Werte von Z mit Wahrscheinlichkeit $1 - \alpha$ liegen .

Beispiel: Für $\alpha = 0.05$ ist nach $\Phi(z_{0.975}) = 0.975$ zu suchen. Aus der Normalverteilungstabelle ergibt sich $z_{0.975} = 1.96$. Das gesuchte Intervall in dem 95% der Werte von Z liegen ist deshalb $[-1.96, 1.96]$.

ACHTUNG! In manchen Büchern, und auch **in dieser Vorlesung bis WS 98/99**, werden Quantile genau andersherum definiert, nämlich über $\Phi(z_q) = 1 - q$. Das beachte man beim Umgang mit alten Skripten, der Klausurensammlung und den zugehörigen Lösungen. Definition (2.131) ist aber in der Literatur sehr viel üblicher!

3. Schritt: Für eine beliebige normalverteilte ZV $X \sim N(\mu, \sigma^2)$ berechnen wir das Intervall über die 'Umkehrung der Standardisierung': Wir schreiben X als $X = \mu + \sigma Z$. Dann ist Z standardnormalverteilt ($Z \sim N(0,1)$), und wir erhalten

$$\Pr(X \in [\mu \pm z_{1-\frac{\alpha}{2}}\sigma]) = 1 - \alpha\,. \tag{2.133}$$

Der Effekt dieser Prozedur ist, dass das Intervall nun symmetrisch um den vorgegebenen Erwartungswert μ herum liegt und mit der Standardabweichung σ gestreckt wird.

4. Schritt: Im nächsten Schritt wenden wir dieses Resultat auf die Verteilung der Mittelwertsmessung an. Im letzten Abschnitt hatten wir den Mittelwert $\bar{x}$ einer Stichprobe als Realisierung der ZV $\bar{X} = \frac{1}{n}\sum_i X_i$ interpretiert, wobei die X_i unabhängige und identisch verteilte ZV mit $\mathbf{E}(X_i) = \mu$ und $\mathrm{Var}(X_i) = \sigma^2$ sind. Wir nehmen an, dass n hinreichend groß ist, dass $\bar{X}$ in guter Näherung normalverteilt ist nach $N(\mu, \sigma^2/n)$. (Wenn die X_i selbst normalverteilt sind, gilt dies aufgrund des Additionssatzes der Normalverteilung sogar *exakt*, auch für kleine n.) Wir fragen nach dem symmetrischen Intervall um μ, in dem $\bar{x}$ mit Wahrscheinlichkeit $1-\alpha$ liegt. Aus Gl. (2.133) folgt

$$\Pr\big(\bar{X} \in [\mu \pm z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]\big) \simeq 1 - \alpha\,. \tag{2.134}$$

Das Intervall wird offensichtlich immer kleiner, je größer n ist. Wir können auch sagen, wie groß n sein muss, damit es eine vorgegebene Breite nicht überschreitet. Dazu setzen wir die Intervallgrenzen $[\mu \pm \delta]$ und lösen (2.134) nach n auf:

$$z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} = \delta \quad \Rightarrow \quad n = \left(\frac{z_{1-\frac{\alpha}{2}} \cdot \sigma}{\delta}\right)^2. \tag{2.135}$$

n ist die Zahl der Versuche, die mindestens ausgeführt werden muss, damit die Abweichung des Mittelwerts vom Erwartungswert mit Wahrscheinlichkeit $1-\alpha$ nicht größer ist als δ.

BEISPIEL: Wie oft muss man (mit einem idealen Würfel) mindestens würfeln, damit der Mittelwert der Augenzahl mit 95% Wahrscheinlichkeit zwischen 3.4 und 3.6 liegt? Erwartungswert und Varianz der Augenzahl beim einmaligen Würfeln sind $\mu = 3.5$ und $\sigma^2 = 2.917$ (siehe Übungen). Es ist $\delta = 0.1$ und $z_{0.975} = 1.96$. Einsetzen in (2.135) ergibt $n = \frac{1.96^2 \cdot 2.917}{0.1^2} \simeq 1121$.

5. Schritt: Bisher haben wir Intervalle um den wahren Erwartungswert herum definiert, in dem Stichprobenmittelwerte mit einer vorgegebenen Wahrscheinlichkeit liegen. Die Ausgangsfragestellung war aber, ein Intervall um den gemessenen Stichprobenmittelwert herum zu definieren, in dem der wahre Mittelwert mit einer gewissen Sicherheit liegt. Wir halten hierfür fest, dass die Aussage '$\bar{X} \in [\mu \pm \delta]$' gleichbedeutend ist mit '$\mu \in [\bar{X} \pm \delta]$' – denn beides besagt, dass $\bar{X}$ und μ nicht weiter als δ voneinander entfernt sind. Wir können (2.134) daher auch lesen als

$$\Pr\big(\mu \in [\bar{X} \pm z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]\big) \simeq 1 - \alpha\,. \tag{2.136}$$

$[\bar{X} \pm z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$ heißt **Konfidenzintervall** (oder **Vertrauensintervall**) für den Erwartungswert μ zum Konfindenzniveau $1-\alpha$. Die Zufallsvariable in (2.136) ist nicht der wahre Erwartungswert μ, sondern der Mittelpunkt des Konfidenzintervalls $\bar{X}$. Die Wahrscheinlichkeit ist also eine Aussage über die Klasse von zufälligen Intervallen, die man erhält, wenn man wiederholt Stichproben vom Umfang n zieht: In $((1-\alpha)\cdot 100)\%$ aller dieser Intervalle wird μ liegen. Für ein einzelnes, beobachtetes Konfidenzintervall $[c_1, c_2]$ heißt dies: Die Aussage "$\mu \in [c_1, c_2]$" ist in $((1-\alpha)\cdot 100)\%$ aller Fälle richtig.

Wir können nun Konfidenzintervalle für verschiedene Fälle konkret angeben.

2.5.2 Konfidenzintervall für den Erwartungswert der Normalverteilung

Gegeben sei eine Stichprobe vom Umfang n mit empirischem Mittelwert $\bar{x}$. Es wird vorausgesetzt, dass die zugrundeliegende Zufallsgröße normalverteilt ist, oder dass n so groß ist, dass die Normalapproximation anwendbar ist.

1. Wenn die **Varianz σ^2 von X bekannt** ist, folgt das symmetrische $1-\alpha$-Konfidenzintervall für den Erwartungswert μ von X direkt aus Gl. (2.136),

$$\left[\bar{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\,,\ \bar{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]. \tag{2.137}$$

2. In den meisten Fällen ist aber die wahre **Varianz σ^2 unbekannt** und man muss mit der **empirischen Varianz** s^2 arbeiten. Zu diesem Zweck betrachtet man an Stelle der standardnormalverteilten ZV Z die ZV

$$T := \frac{\bar{X} - \mu}{\sqrt{S^2/n}}\,. \tag{2.138}$$

Diese Größe kombiniert mit dem Mittelwert $\bar{X}$ und der empirischen Varianz S^2 (vgl. Abschnitt 2.4.3) zwei Zufallsvariablen, deren Wert aus der Stichprobe bestimmt werden. Die Dichte von T kann man mit einigem Aufwand bestimmen, die Verteilungsfunktion ist

$$Pr(T \le t) = c\int_{-\infty}^{t}\left(1 + \frac{\tau^2}{n-1}\right)^{-n/2} d\tau\,, \tag{2.139}$$

wobei c eine Normierungskonstante ist. Dies ist die sogenannte **Studentsche t-Verteilung** mit $n-1$ Freiheitsgraden (kurz: T ist t_{n-1}-verteilt). Wie bei der Normalverteilung kennt man keinen geschlossenen Ausdruck für das Integral und muss die Werte in Tabellen nachschlagen. Das symmetrische $1-\alpha$-Konfidenzintervall für den Erwartungswert μ ist dann

$$\left[\bar{x} - t_{n-1,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\,,\ \bar{x} + t_{n-1,1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right], \tag{2.140}$$

wobei $t_{n-1,q}$ das q-Quantil der t_{n-1}-Verteilung ist, also derjenige Wert, der

$$Pr(T \le t_{n-1,q}) = q \tag{2.141}$$

erfüllt. Quantilwerte können der Tabelle der Student-t-Verteilung im Anhang auf Seite 91 entnommen werden. Da die t-Verteilung zwei Zufallsgrößen kombiniert, hat sie eine größere Varianz als die Standardnormalverteilung. Deshalb ist das Konfidenzintervall mit dem t-Quantil $t_{n-1,q}$ besonders für kleine n breiter als das entsprechende Intervall mit dem Quantil z_q der Normalverteilung (z.B. ist $t_{2,0.975} = 4.3 > 1.96 = z_{0.975}$). Es gilt jedoch $\lim_{n\to\infty} t_{n-1,q} = z_q$; für große n ist daher die Approximation durch die Normalverteilung zulässig (siehe Tabelle S. 91).

2.5.3 Konfidenzintervall für den Parameter p eines Bernoulli-Experiments

Bei n Wiederholungen eines Bernoulli-Experiments wurde n_A mal das Ereignis A beobachtet, woraus sich $\hat{p} = n_A/n$ als Schätzer für p, die Wahrscheinlichkeit von A, ergibt. Für $n\hat{p}(1-\hat{p}) \ge 9$ ist die Normalapproximation zulässig und wir können nach obigem Muster vorgehen. Wenn wir

annehmen, dass n sehr groß ist, können wir die Varianz des (einfachen) Bernoulli-Experiments mit $s^2 = \frac{n}{n-1}\hat{p}(1-\hat{p}) \simeq \hat{p}(1-\hat{p})$ abschätzen und erhalten mit der Quantile der Normalverteilung

$$\left[\hat{p} - z_{1-\frac{\alpha}{2}}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\ ,\ \hat{p} + z_{1-\frac{\alpha}{2}}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right] \tag{2.142}$$

für das symmetrische $1-\alpha$-Konfidenzintervall von p. (Falls (2.142) Intervallgrenzen < 0 bzw. > 1 liefert, müssen diese natürlich durch 0 bzw. 1 ersetzt werden.)

Beispiel:

Die Geburtsstatistik der Schweiz weist zwischen 1950 und 1970 $n = 1944700$ Geburten aus, davon $n_J = 977600$ Jungengeburten. Es sollen das 99% und das 99.9% Konfidenzintervall für die Wahrscheinlichkeit p einer Jungengeburt ermittelt werden. Man erhält $\hat{p} = n_J/n = 0.512984$. Da n riesig ist, ist die Normalapproximation anwendbar und die Varianz kann mit $\sigma^2 \simeq s^2 \simeq 0.249831$ abgeschätzt werden. Die relevanten Quantilwerte sind $z_{0.995} = 2.575$ und $z_{0.9995} = 3.300$. Damit ergeben sich die 99% und 99.9% Konfidenzintervalle als $[0.51206, 0.513907]$ bzw. $[0.511904, 0.514164]$. Keines schließt den Wert 0.5 ein, Jungengeburten sind also mit 99.9% iger Sicherheit häufiger als Mädchengeburten.

Bemerkungen

- Die Breite des Konfidenzintervalls einer Mittelwertschätzung hängt von drei Faktoren ab.
  1. Varianz σ^2: Eine Größe, die weniger schwankt, kann man enger eingrenzen.
  2. Stichprobenumfang n: größere Stichprobe ergibt sicherere Aussage.
  3. Niveau $1-\alpha$: Mehr Sicherheit (kleineres α) erfordert breiteres Intervall, das mehr Werte umfasst.
- Übliche Konfidenzniveaus sind 95%, 99% und 99.9%. Sie entsprechen den gängigen Signifikanzniveaus für statistische Tests (siehe nächster Abschnitt).
- Alternativ zu den Intervallen auf diesen Niveaus gibt man in der graphischen Darstellung von Daten oft auch sogenannte **Fehlerbalken** (*error bars*) von der Breite der empirischen Standardabweichung an. Für die Schätzung eines Mittelwerts μ durch eine Stichprobe vom Umfang n ist der Fehlerbalken durch das Intervall $[\bar{x} - s/\sqrt{n}\,,\ \bar{x} + s/\sqrt{n}]$ gegeben. Dahinter steckt (wie oben) die Annahme, dass n hinreichend groß ist, dass die Werte um μ herum normalverteilt sind und dass s als gute Schätzung von σ betrachtet werden kann. Dann ist der Fehlerbalken ein Konfidenzintervall zum Niveau $2\Phi(1) - 1 \simeq 0.68$ (für kleine n ist das Niveau etwas niedriger). Als Faustregel kann man sich außerdem merken, dass man ein 95%-Konfidenzintervall ungefähr durch Verdopplung der Länge des Fehlerbalkens erhält.

2.6 Das Testen von Hypothesen

Experimentelle Daten dienen letzten Endes immer dem Zweck, bestimmte Vermutungen zu bestätigen oder zu widerlegen. Bei zufälligen Größen kann man solche Entscheidungen nie mit letzter Sicherheit treffen; man benötigt hier das Konzept des **statistischen Tests**.

2.6.1 Das Testprinzip

Ausgangspunkt für den Test ist eine **Hypothese.** Eine Hypothese ist in diesem Zusammenhang ganz allgemein eine Annahme über eine Zufallsvariable. Ein **Test** einer Hypothese ist ein Prüfverfahren, das man anwendet, um festzustellen, ob die Hypothese abgelehnt werden soll oder nicht. Das Konzept des statistischen Tests soll nun anhand eines Beispiels ausführlich beprochen werden.
Wir betrachten die schon im letzten Abschnitt erwähnte Schweizer Geburtenstatistik. Zwischen 1950 und 1970 gab es in der Schweiz $n = 1944700$ Geburten insgesamt und davon $n_J = 977600$ Jungengeburten. Anhand dieser Stichprobe soll getestet werden, ob Jungen- und Mädchengeburten in der Schweiz gleich wahrscheinlich sind. Ausgangapunkt für einen Test ist stets die präzise Formulierung der **Nullhypothese** H_0. Die Nullhypothese ist ein statistisches Modell der Wirklichkeit. Ein Modell besteht immer aus einer Reihe von Annahmen. Im Beispiel der Geburtenstatistik können die Annahmen von H_0 z.B. wie folgt aussehen:

1. Wir nehmen an, dass die Resultate einzelner Geburten (Junge oder Mädchen) voneinander unabhängig sind.

2. Wir nehmen an, dass die Wahrscheinlichkeit für eine Jungengeburt bei einer Geburt im Beobachtungszeitraum einen konstanten Wert p hatte.

3. Wir nehmen an, dass dieser Wert $p = p_0 = 0.5$ ist.

Wenn wir für die $i$te Geburt eine ZV X_i definieren mit $X_i = 1$ für einen Jungen und $X_i = 0$ für ein Mädchen, so ist jede einzelne Geburt unter der Nullhypothese ein Beroulliexperiment mit Mittelwert $p_0 = 0.5$ und Varianz $\sigma^2 = p_0(1 - p_0) = 0.25$. Aus den ersten beiden Annahmen folgt, dass die X_i unabhängig und identisch verteilt sind. Da n sehr groß ist, können wir für die Mittelwertsmessung deshalb den Zentralen Grenzwertsatz anwenden und erhalten die Verteilung des Mittelwerts $\bar{X} = (1/n)\sum_i X_i$ unter H_0,

$$H_0 : \qquad \bar{X} \sim N(p_0, p_0(1-p_0)/n) = N(0.5, 0.25/1944700)\,. \tag{2.143}$$

Wir wollen nun die Nullhypothese testen. Dies geschieht, indem man den gemessenen Mittelwert mit der Vorhersage der Nullhypothese vergleicht. Es gibt zwei mögliche Resultate: Wir lehnen die Nullhypothese ab (verwerfen H_0), wenn der beobachtete Stichprobenmittelwert $\bar{x}$ unter H_0 'zu unwahrscheinlich' ist, das heißt, er weicht zu weit vom angenommenen Mittelwert $p_0 = 0.5$ ab, als dass wir ihn als zufällige Schwankung akzeptieren wollen. Anderenfalls verwerfen wir H_0 nicht. Um genau festzulegen, was 'zu unwahrscheinlich' sein soll, wählt man vor (!) der Durchführung des Tests eine kleine Zahl α, das sogenannte **Signifikanzniveau** des Tests oder die **Irrtumswahrscheinlichkeit**. H_0 wird dann abgelehnt, wenn $\bar{x}$ in den **Ablehnungsbereich** $\{\bar{x} : |\bar{x} - p_0| > \delta\}$ fällt. Dabei wird δ so bestimmt, dass

$$Pr(|\bar{X} - p_0)| > \delta) = \alpha. \tag{2.144}$$

Nach Gl. (2.135) muss δ dann gerade als

$$\delta = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \tag{2.145}$$

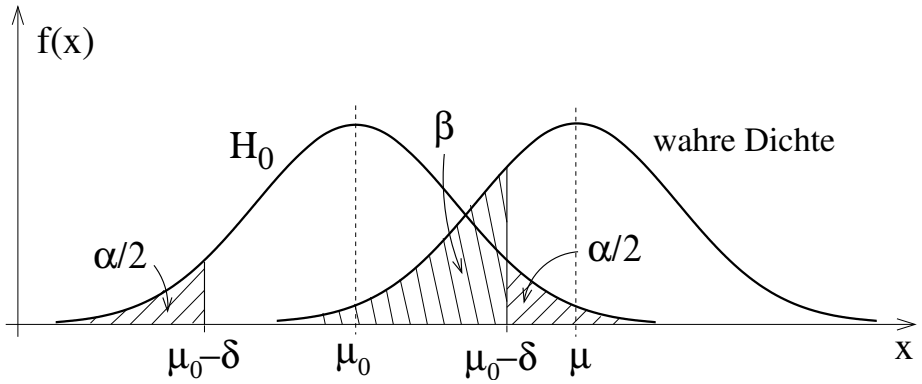


Abbildung 2.32: Gauß-Tests mit Fehler erster und zweiter Art. μ_0 ist der Erwartungswert unter H_0, μ ist der wahre (aber unbekannte) Erwartungswert. Die Grenzen $\mu_0 \pm \delta$ hängt von der Wahl des Signifikanzniveaus α ab.

gewählt werden. H_0 wird also verworfen, falls

$$\frac{|\bar{x} - p_0|}{\sigma}\sqrt{n} > z_{1-\frac{\alpha}{2}}\,. \tag{2.146}$$

Andernfalls behält man die Nullhypothese bei. Man kann daraus allerdings nicht schließen, dass sie richtig ist. Man kann lediglich folgern, dass man aufgrund der Stichprobe keine signifikante Abweichung zum Signifikanzniveau α feststellen kann. Wenn wir für das Geburtenbeispiel z.B. $\alpha = 0.001$ wählen, ist

$$\frac{|\bar{x} - p_0|}{\sigma}\sqrt{n} = 36.20 > z_{0.9995} = 3.3, \tag{2.147}$$

H_0 wird also abgelehnt. Aus der Ablehnung der Nullhypothese folgt, dass zumindest eine der Annahmen, die man unter H_0 gemacht hat, wahrscheinlich nicht zutrifft. Im Geburtenbeispiel gibt es zwei Möglichkeiten: Entweder die Stichprobenmittelwerte sind nicht normalverteilt. Dies wäre der Fall, wenn die einzelnen Geburten nicht als unabhängige Bernoulliexperimente verstanden werden können (Annahmen 1 und 2). Oder die Annahme für Erwartungswert und Varianz unter H_0 sind nicht zutreffend. Da beide aus dem Parameter p für die Wahrscheinlichkeit der Jungengeburt folgen, führt dies zu dem Schluss, dass die Annahme 3, $p = p_0 = 0.5$ bzw. gleiche Wahrscheinlichkeiten für Jungen- und Mädchengeburten, falsch ist. Man kann nun weitere Argumente dafür suchen, dass es keinen Grund gibt, an Annahmen 1 und 2 zu zweifeln (es sind hierfür auch weitere Tests möglich) und kommt dann endlich zu dem Schluss, dass Jungengeburten in der Schweiz in der Tat signifikant häufiger sind (zum Niveau $\alpha = 0.001$) als Mädchengeburten.

Fehler 1. und 2. Art

Da mit zufälligen Daten keine absolute Sicherheit zu erreichen ist, können bei der Ablehnung oder Beibehaltung von H_0 Fehler nie ganz vermieden werden. Man unterscheidet zwei Typen von Fehlern (vgl. Abb. 2.32):

- **Fehler 1. Art**: Die Nullhypothese wird abgelehnt, obwohl sie richtig ist. Diesen Fehler hat man bei einem Test unter Kontrolle: Der Test ist gerade so konstruiert, dass der Fehler erster Art mit der vorgegebenen Irrtumswahrscheinlichkeit α geschieht.
- **Fehler 2. Art**: Die Nullhypothese wird beibehalten, obwohl sie falsch ist. Die Wahrscheinlichkeit β, mit der dies geschieht, hängt von der wahren (aber unbekannten) Verteilung

ab und kann deshalb nicht sicher kontrolliert werden. Man kann aber sagen: β wird umso größer, je kleiner α ist.

In Tabellenform zusammengefasst:

	Das Testergebnis lautet	
	H_0 ablehnen	H_0 beibehalten
H_0 ist richtig	Fehler 1. Art α	kein Fehler $1-\alpha$
H_0 ist falsch	kein Fehler $1-\beta$	Fehler 2. Art β

Bemerkungen

- Man wählt eine Nullhypothese meist 'konservativ': Sie verneint die Unterschiede, Veränderungen oder Zusammenhänge, die man nachweisen möchte. Das "erwünschte" Ergebnis eines Tests ist dann die Ablehnung der Nullhypothese. Den Fehler, den man dabei macht, hat man durch die Wahl von α unter Kontrolle.

- Je stärker man eine Nullhypothese wählt, das heißt, je mehr Annahmen man in sie aufnimmt, desto einfacher ist sie statistisch signifikant abzulehnen. Aber: Je stärker H_0, desto schwächer ist auch das Resultat bei einer Ablehnung. Schließlich weiß man am Ende nur, dass *irgendeine* Annahme unter H_0 wahrscheinlich falsch war. Man versucht deshalb, H_0 immer so stark wie möglich zu machen, aber außer der Annahme, die man wirklich testen will, möglichst keine weiteren unabgesicherten Annahmen aufzunehmen.

- Die Wahl des Signifikanzniveaus α ist im Prinzip willkürlich. Als Regel kann man nur sagen, dass α klein sein sollte, wenn man mit der Ablehnung von H_0 einen neuen Zusammenhang zeigen will, damit man hinreichend sicher sein kann. In der Praxis werden allerdings fast ausschließlich drei Niveaus verwendet, und zwar 0.05, 0.01 und 0.001.

- Comuterprogramme für statistische Tests berechnen in der Regel sogenannte p-Werte auf Grund derer H_0 abgelehnt oder beibehalten wird. Der p-Wert gibt das kleinste Signifikanzniveau an, unter dem die Nullhypothese noch abgelehnt werden würde. Der Test ist dann signifikant, wenn p kleiner als das *vorher festgelegte* α ist. In der Literatur wird der genaue p-Wert in diesem Fall mit dem Resultat angegeben.

2.6.2 Der Zoo der Erwartungswerttests

Nachdem wir nun das Prinzip des statistischen Tests kennengelernt haben, soll jetzt eine Übersicht über einige Tests für den Erwartungswert einer Verteilung gegeben werden (sogenannte Erwartungswerttests). Es gibt eine große Anzahl verschiedener Erwartungswerttests, die je nach Situation zur Anwendung kommen. Der Einfachheit halber nehmen wir hier generell an, dass von einer Normalverteilung der Mittelwertschätzung ausgegangen werden kann (also Messungen unabhängig und Stichprobe hinreichend groß, bzw. die interessierende Größe ist selbst schon normalverteilt). Um dann den jeweils 'richtigen' Test zu finden, geht man anhand folgender Kriterien vor:

1. Vergleicht man seine Beobachtungen mit einem vorgegebenen 'Idealwert' μ_0 (wie im Geburtenbeispiel), oder vergleicht man zwei Zufallsgrößen X und Y miteinander (oft sind das Daten aus 'Experiment' und 'Kontrolle', z.B. das Wachstum von Pflanzen mit und ohne Wachstumshormon). Im ersten Fall ist die zu testende Annahme unter H_0 $\mu = \mu_0$ und

man macht einen **Ein-Stichproben-Test** (manchmal auch **einfacher Test** genannt). Im zweiten Fall testet man $H_0 : \mu_X = \mu_Y$ und macht einen **Zwei-Stichproben-Test** (oder **doppelten Test**).

2. Ist die Varianz der Zufallsgröße (unter der Nullhypothese) bekannt, oder muss sie aus der Stichprobe geschätzt werden? Ist die Varianz bekannt, so verwendet man **Gauß-Tests**, ansonsten **t-Tests**.

3. Ist die Fragestellung **einseitig** oder **zweiseitig**? Wenn man einfach nach der Abweichung zweier Mittelwerte fragt (wie beim Geburtenbeispiel), dann ist die Fragestellung zweiseitig. Die Nullhypothese lautet dann $H_0 : \mu = \mu_0$ und große Abweichungen der Prüfgröße nach oben *oder* nach unten (bezüglich μ_0) führen gleichermaßen zur Ablehnung. Oft weiß man aber aus theoretischen oder praktischen Überlegungen schon vor (!) Durchführung des Experiments, dass $\mu_X \geq \mu_0$ (oder $\mu_X \leq \mu_0$) sein sollte (diese Vermutung darf sich aber nicht erst aus der Stichprobe selbst ergeben, die man für den Test verwendet). Wenn wir zum Beispiel testen, ob die Reduktion von Futter bei Ratten einen signifikanten Effekt auf das Gewicht von Neugeborenen hat, können wir u.U. von vorneherein davon ausgehen, dass sich das Gewicht jedenfalls nicht erhöhen wird. In diesem Fall lehnt man die Nullhypothese nur bei einer großen Abweichung in einer Richtung ab (Abweichung nach unten für die Ratten), die Fragestellung ist einseitig und der Ablehnungsbereich besteht dann aus den $\alpha \cdot 100\%$ *größten (kleinsten)* Werten; vgl. Abb. 2.33. Ein einseitiger Test ist auch dann angebracht, wenn nur Abweichungen von μ_0 in einer Richtung relevant sind, wie bei der Überschreitung (Unterschreitung) eines Grenzwertes. Die Nullhypothese lautet dann $H_0 : \mu \leq \mu_0$ (bzw. $H_0 : \mu \geq \mu_0$).

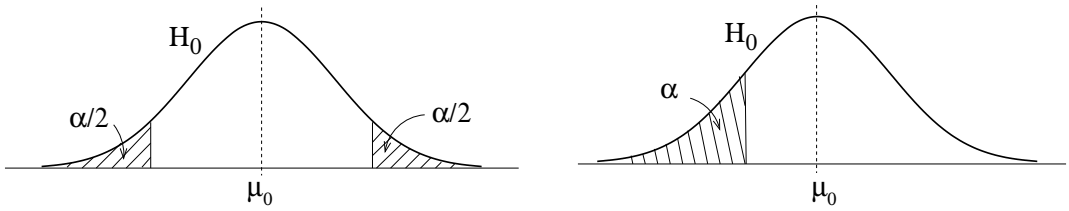


Abbildung 2.33: Links: Zweiseitiger Test. Rechts: Einseitiger Test wenn $\mu \leq \mu_0$ bekannt oder nur dieser Fall relevant ist (Fall $\mu \geq \mu_0$ spiegelbildlich).

In den nächsten beiden Abschnitten folgt eine Übersicht über wichtige Erwartungswerttests.

2.6.3 Ein-Stichproben-Tests

Die Frage, ob der unbekannte Erwartungswert μ einer Zufallsvariablen X vom vermuteten Wert μ_0 abweicht, soll zum Signifikanzniveau α beantwortet werden. Gegeben sei eine Stichprobe vom Umfang n mit Mittelwert $\bar{x}$ und empirischer Varianz s^2.

1. **Einfacher Gauß-Test:** Die Varianz σ^2 unter H_0 wird als bekannt vorausgesetzt. Der Test beruht darauf, dass die ZV

$$Z := \frac{\bar{X} - \mu_0}{\sigma}\sqrt{n} \tag{2.148}$$

unter H_0 standardnormalverteilt ist.

(a) Zweiseitiger Test: Verwerfe H_0, falls

$$\frac{|\bar{x} - \mu_0|}{\sigma}\sqrt{n} > z_{1-\frac{\alpha}{2}}\,, \tag{2.149}$$

wobei $z_{1-\frac{\alpha}{2}}$ das $1-\frac{\alpha}{2}$-Quantil der Standardnormalverteilung bezeichnet.

(b) Einseitiger Test: $\mu \geq \mu_0$ von vornreherein bekannt oder nur Abweichung $\mu > \mu_0$ relevant. Verwerfe H_0, falls

$$\frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} > z_{1-\alpha}\,. \tag{2.150}$$

(c) Einseitiger Test: $\mu \leq \mu_0$ bekannt oder nur Abweichungen $\mu < \mu_0$ relevant. Verwerfe H_0, falls

$$\frac{\bar{x} - \mu_0}{\sigma}\sqrt{n} < -z_{1-\alpha}\,. \tag{2.151}$$

Bei einem signifikanten Ergebnis eines einfachen Gauß-Tests und Ablehnung der Nullhypothese gibt es drei Möglichkeiten:

I Der Erwartungswert μ weicht tatsächlich von μ_0 ab.

II Die unter H_0 angenommene Varianz trifft nicht zu.

III Die Mittelwertschätzung $\bar{X}$ ist nicht (näherungsweise) normalverteilt.

Um auf I schließen zu können (was meistens das Ziel ist), muss man II und III ausschliessen können. III kann ausgeschlossen werden, wenn entweder die ZV X selbst normalverteilt ist oder die Stichprobe genügend groß gewählt wurde und die Messungen unabhängig waren (ZGS!). Für den Ausschluss von II braucht man eine Kenntnis der Varianz unabhängig von der Stichprobe (oder man weiss, wie im Geburtenbeispiel, dass wenn II zutrifft, auch I zutreffen muss, da Erwartungswert und Varianz gekoppelt sind).

2. **Einfacher t-Test:** Die Varianz σ^2 unter H_0 ist *unbekannt*; daher muss die *empirische Varianz* s^2 der Stichprobe herangezogen werden. Der Test beruht darauf, dass die ZV

$$T := \frac{\bar{X} - \mu_0}{S}\sqrt{n} \tag{2.152}$$

unter H_0 Student-t-verteilt ist mit $k = n - 1$ Freiheitsgraden (dies ist ganz analog zur Überlegung zu Konfidenzintervallen bei unbekannter Varianz).

(a) Zweiseitiger Test: H_0 ist abzulehnen, falls

$$\frac{|\bar{x} - \mu_0|}{s}\sqrt{n} > t_{n-1,1-\frac{\alpha}{2}}\,, \tag{2.153}$$

wobei $t_{n-1,1-\frac{\alpha}{2}}$ das $1-\frac{\alpha}{2}$-Quantil der Student-t-Verteilung mit $k = n - 1$ Freiheitsgraden bezeichnet.

(b) Einseitige Tests: Es wird als bekannt vorausgesetzt, dass $\mu \geq \mu_0$ gilt (bzw. $\mu \leq \mu_0$). Dann ist H_0 ist abzulehnen, falls

$$\frac{\bar{x} - \mu_0}{s}\sqrt{n} > t_{n-1,1-\alpha} \quad (\text{bzw. } < -t_{n-1,1-\alpha})\,. \tag{2.154}$$

Bei einem signifikanten Ergebnis eines einfachen t-Tests und Ablehnung von H_0 gibt es zwei Möglichkeiten: Entweder μ weicht tatsächlich von μ_0 ab oder die Mittelwertschätzung X ist nicht normalverteilt. Für eine große Stichprobe, oder wenn X selbst bekanntermaßen normalverteilt ist, kann Letzteres ausgeschlossen werden.

Beispiel Ist der mittlere Cholesterinwert μ von Vegetariern signifikant ($\alpha = 0.01$) geringer als der Normalwert ($\mu_0 = 180$ mg/100ml)? Wir gehen davon aus, dass er jedenfalls nicht erhöht sein sollte, setzen also voraus, dass $\mu \leq \mu_0$ gilt und testen einseitig. Bei einer Gruppe von 9 Vegetariern werden folgende Werte gemessen: 154, 119, 177, 150, 138, 185, 167, 158, 174. Man erhält $\bar{x} = 158$, $s = 20.7$, und wegen $\frac{158-180}{20.7}\sqrt{9} = -3.19 < -2.90 = -t_{8,0.99}$ wird H_0 abgelehnt. Wenn wir davon ausgehen können, dass Cholesterinwerte normalverteilt sind, ist der Cholesterinwert von Vegetariern damit signifikant niedriger als der Normalwert.

2.6.4 Zwei-Stichproben-Tests

Der Vergleich zweier Stichproben (z. B. aus Experiment und Kontrolle) ist eine Standardsituation in der Biologie. Wir besprechen Tests für zwei verschiedene Versuchspläne (*experimental designs*). Bei **gepaarten Stichproben** macht man an n Objekten (bzw. Individuen) je zwei Messungen. Zum Beispiel kann man dasselbe Merkmal unter verschiedenen Bedingungen (in verschiedenen *treatments*) messen: Blutdruck mit und ohne Belastung, Genexpressionen (der gleichen Gene) in männlichen und weiblichen Drosophilen, etc. Die einzelnen Messungen in einer Stichprobe sollen wieder unabhängig sein, die beiden Stichproben untereinander sind es aber nicht. Man kann je eine Messung der einen Stichprobe einer Messung der anderen Stichprobe zuordnen – man spricht auch von *verbundenen* Stichproben. Ein alternatives Design verwendet **unabängige Stichproben**. Hier ist es wesentlich, dass die Messungen in beiden Stichproben sich gegenseitig nicht statistisch beeinflussen (zum Beispiel Messung eines Merkmals an Individuen aus zwei verschiedenen Populationen).

Gepaarte Stichproben: t-Differenzentest

An n Objekten (Individuen) werden jeweils die Merkmale X und Y gemessen, bzw. dasselbe Merkmal unter verschiedenen Bedingungen. Die Messungen beider Stichproben sind paarweise verbunden, X und Y sind *nicht* unabhängig. Die Erwartungswerte $\mu_X := \mathbf{E}(X)$ und $\mu_Y := \mathbf{E}(Y)$ sind unbekannt, ebenso die Varianzen. Man testet auf Gleichheit der Erwartungswerte, also $H_0 : \mu_X - \mu_Y = 0$. Der Test lässt sich auf einen einfachen t-Test für die Differenzen $D := X - Y$ zurückführen. Man nimmt unter H_0 an, dass die Zufallsvariable $\bar{D} = \bar{X} - \bar{Y}$ normalverteilt ist, mit Erwartungswert $\mu_D = 0$ und unbekannter Varianz. Dann ist unter H_0

$$\frac{\bar{D}}{S_D}\sqrt{n} \sim t_{n-1}\,, \tag{2.155}$$

wobei S_D^2 die empirische Varianz von D als Zufallsvariable bezeichnet. Man berechnet also aus der Stichprobe die Differenzen

$$d_i = x_i - y_i\,, \quad i = 1, \ldots, n\,, \tag{2.156}$$

und deren Mittelwert und empirische Varianz

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \bar{x} - \bar{y} \quad , \quad s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2\,. \tag{2.157}$$

H_0 wird abgelehnt, falls

$$\frac{|\bar{d}|}{s_d}\sqrt{n} > t_{n-1,1-\frac{\alpha}{2}}\,. \tag{2.158}$$

Ganz analog werden entsprechende einseitige Tests durchgeführt.

Beispiel: Der Ertrag X (in Zentnern) von 8 Kirschbäumen im Jahr 1990 wird mit dem Ertrag Y *derselben* 8 Bäume im Jahr 1991 verglichen. Es soll zum Signifikanzniveau 0.05 getestet werden, ob der Ertrag vom Jahr abhängt.

Baum-Nr. i	1	2	3	4	5	6	7	8	
x_i	3.6	3.1	3.4	3.2	3.5	3.1	3.2	3.5	$\bar{x} = 3.33$
y_i	3.5	3.5	3.4	3.6	4.0	3.5	3.3	3.1	$\bar{y} = 3.48$
d_i	-0.1	-0.4	0.	-0.4	-0.5	-0.4	-0.1	0.4	$\bar{d} = -0.15$

Es ergibt sich $s_d^2 = 0.104$ und somit $\frac{|\bar{d}|}{s_d}\sqrt{n} = 1.32 < 2.365 = t_{7,0.975}$, also kein signifikanter Unterschied.

Unabhängige Stichproben: Doppelter t-Test

X und Y seien *unabhängige* ZV mit unbekannten Erwartungswerten $\mu_X := \mathbf{E}(X)$, $\mu_Y := \mathbf{E}(Y)$ und Varianzen $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$. Gegeben sind Stichproben von X und Y vom Umfang n_x bzw. n_y. Ihre Mittelwerte sind $\bar{x}$ bzw. $\bar{y}$, und ihre empirischen Varianzen s_x^2 bzw. s_y^2. Es soll auf Gleichheit der Erwartungswerte getestet werden. Die Differenz der *Stichprobenmittel*,

$$\bar{D} := \bar{X} - \bar{Y}\,, \tag{2.159}$$

ist unter H_0 eine normalverteilte ZV mit $\mathbf{E}(\bar{D}) = 0$ und

$$\sigma_{\bar{D}}^2 := \mathrm{Var}(\bar{D}) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{1}{n_x}\mathrm{Var}(X) + \frac{1}{n_y}\mathrm{Var}(Y). \tag{2.160}$$

Eine weitere wichtige Annahme der Nullhypothese ist nun, dass die Varianzen von X und Y *gleich* sind, $\mathrm{Var}(X) = \mathrm{Var}(Y) = \sigma^2$ (sog. Annahme der *Varianzhomogenität*). Dann ist

$$\sigma_{\bar{D}}^2 = \Big(\frac{1}{n_x} + \frac{1}{n_y}\Big)\sigma^2 = \frac{n_x + n_y}{n_x n_y}\,\sigma^2\,. \tag{2.161}$$

Das unbekannte σ^2 wird nun aus der Stichprobe abgeschätzt, und zwar als *gewichtetes Mittel* von s_x^2 und s_y^2 mit Gewichtsfaktoren $n_x - 1$ und $n_y - 1$:

$$\hat{\sigma}^2 = s^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}\,. \tag{2.162}$$

Somit ergibt sich als Schätzwert für die Varianz von $\bar{D}$:

$$\hat{\sigma}_{\bar{D}}^2 = \frac{n_x + n_y}{n_x n_y}\,s^2\,. \tag{2.163}$$

Mit diesen Schätzwerten kann man zeigen, dass $\frac{\bar{D}}{S}\sqrt{\frac{n_x n_y}{n_x+n_y}}$ Student-t-verteilt ist mit $k = n_x + n_y - 2$ Freiheitsgraden. H_0 ist deshalb zu verwerfen, falls

$$\frac{|\bar{d}|}{s}\sqrt{\frac{n_x n_y}{n_x + n_y}} > t_{n_x+n_y-2,1-\frac{\alpha}{2}} \quad \text{wobei} \quad \bar{d} := \bar{x} - \bar{y}. \tag{2.164}$$

Analog konstruiert man entsprechende einseitige Tests.

Beispiel: In einer Kölner Klinik wurden im Jahr 1985 $n_x = 269$ Mädchen und $n_y = 288$ Jungen geboren, mit Durchschnittsgewicht $\bar{x} = 3050$g bzw. $\bar{y} = 3300$g. Die zugehörigen empirischen Standardabweichungen waren $s_x = 460$g und $s_y = 470$g. Es soll zum Signifikanzniveau $\alpha = 0.01$ getestet werden, ob Jungen und Mädchen das gleiche erwartete Geburtsgewicht haben. Es ist $|\bar{d}| = |\bar{x} - \bar{y}| = |3050 - 3300| = 250$, $k = n_x + n_y - 2 = 555$ und $s^2 = \frac{268 \cdot 460^2 + 287 \cdot 470^2}{555} = 216409.2$. Bei einem doppelten t-Test ist deshalb $\frac{\bar{x}-\bar{y}}{s}\sqrt{\frac{n_x n_y}{n_x+n_y}} = 6.34 > 2.576 = z_{0.995} \simeq t_{555,0.995}$ und die Nullhypothese wird abgelehnt. Unter der (sehr plausiblen) Annahme der Normalverteilung und der (ggf. mit einem eignenen, vorbereitenden Test zu begründenden) Annahme der Varianzhomogenität kann man zum Niveau $\alpha = 0.05$ schließen, dass Jungen im Mittel schwerer sind als Mädchen.

2.6.5 Der Chi-Quadrat-Anpassungstest

Die bisher behandelten statistischen Tests betrafen Erwartungswerte von Verteilungen. Manchmal interessiert man sich aber nicht so sehr für Unterschiede im Erwartungswert, sondern für andere Kenngrößen der Verteilung (wie die Varianz) oder man will wissen, ob es überhaupt einen Unterschied in den Verteilungen gibt. In vielen Fällen, wenn die Daten nicht metrisch sind, sondern nur ordinal oder nominal, ist der Erwartungswert (und die Varianz) gar nicht definiert. Um einen Unterschied in der Verteilung von zwei Stichproben feststellen zu können, braucht man hier also andere Methoden. Eine Methode, die nicht spezielle Kenngrößen wie den Erwartungswert untersucht, sondern die gesamte Verteilung ist der sogenannte **Anpassungstest**.

Beispiel

Wir betrachten eine Pflanzenart mit den Genotypen aa (1; weiße Blüte), Aa (2; rosa Blüte) und AA (3; rote Blüte) und kreuzen Aa mit Aa. Falls der intermediäre Mendelsche Erbgang die Situation perfekt beschreibt, erwartet man in der Nachkommenschaft aa, Aa und AA im Verhältnis $p_1 = 1/4$, $p_2 = 1/2$, und $p_3 = 1/4$. Anhand einer Stichprobe der Größe $n = 400$ soll überprüft werden, ob die Verteilung in der Nachkommengeneration mit den Mendelschen Verhältnissen kompatibel ist. Die Stichprobe liefert die $n_1 = 115$ weiße Blüten, $n_2 = 171$ rosa Blüten und $n_3 = 114$ rote Blüten. Kann aufgrund dieser Beobachtung die Nullhypothese "Mendelscher intermediärer Erbgang" abgelehnt werden?

Das Prinzip des Chi-Qudrat-Tests

Wir betrachten nur den Fall diskreter Zufallsvariablen. Sei Y eine diskrete Zufallsvariable, die r verschiedene Werte $y_1, y_2, \ldots, y_r$ annehmen kann. Im Beispiel sind das die Bütenfarben weiß, rosa und rot. Die Nullhypothese behauptet, dass die Verteilung von Y durch vorgegebene Wahrscheinlichkeiten $p_1, p_2, \ldots, p_r$ beschrieben wird:

$$H_0: \text{ Es gilt } \quad Pr(Y = y_i) = p_i \quad \text{für alle } \quad i = 1, \ldots, r. \tag{2.165}$$

Im Beispiel sind die Wahrscheinlichkeiten für die einzelnen Blütenfarben unter H_0 wie oben angegeben durch die Mendelschen Gesetze definiert. H_0 soll nun anhand einer Stichprobe überprüft werden. Wir nehmen an, dass in einer Stichprobe vom Umfang n der Wert y_i gerade n_i-mal vorkommt ($n_1 + \ldots + n_r = n$). Diese n_i sind die absoluten Häufigkeiten, wie man sie auch in Histogrammen verwendet. Im Kontext des Anpassungstests nennt man sie auch **beobachtete Häufigkeiten**. Nehmen wir nun an, dass die Nullhypothese wahr ist, so finden wir mit Wahrscheinlichkeit p_i beim zufälligen Herausgreifen eines Objekts aus der Grundgesamtheit den Wert y_i. Der Erwartungswert für N_i (die zu n_i gehörige Zufallsvariable) ist somit gleich $e_i = np_i$. Diese

Zahlen e_i nennen wir die **erwarteten Häufigkeiten** für die Werte y_i. Im obigen Beispiel sind die erwarteten Häufigkeiten $e_1 = 400 \cdot 1/4 = 100$, $e_2 = 400 \cdot 1/2 = 200$ und $e_3 = 400 \cdot 1/4 = 100$. Als Maß für die Abweichung der Stichprobe von der hypothetischen Verteilung dient nun die Summe der *relativen quadratischen Abweichungen* der beobachteten von den erwarteten Häufigkeiten:

$$\chi^2 := \sum_{i=1}^{r} \frac{(n_i - e_i)^2}{e_i} \,. \tag{2.166}$$

Man kann zeigen, dass die zugehörige Zufallsvariable,

$$X^2 := \sum_{i=1}^{r} \frac{(N_i - e_i)^2}{e_i} \,, \tag{2.167}$$

für große n näherungsweise χ^2-verteilt ist mit $k = r - 1$ Freiheitsgraden, d.h.

$$Pr(X^2 < x^2) = c \int_0^{x^2} u^{k/2-1} e^{-u/2} du \,. \tag{2.168}$$

Dabei ist c eine Normierungskonstante. Diese Dichte verschwindet für negative Werte von X^2 – schließlich können Abweichungs*quadrate* nie negativ sein. Der Test ist also immer *einseitig*.
Die Nullhypothese ist abzulehnen, falls für eine vorgegebene Irrtumswahrscheinlichkeit α

$$\chi^2 > \chi^2_{k,1-\alpha} \tag{2.169}$$

ist, wobei $\chi^2_{k,q}$ das q-Quantil der χ^2-Verteilung mit k Freiheitsgraden bezeichnet. Das heißt, $\chi^2_{k,q}$ ist derjenige Wert, der

$$Pr(X^2 \leq \chi^2_{k,q}) = q \tag{2.170}$$

erfüllt. Eine Quantiltabelle findet sich im Anhang auf Seite 92.
Im Beispiel erhalten wir $\chi^2 = (115 - 100)^2/100 + (171 - 200)^2/200 + (114 - 100)^2/100 = 8.46 > 5.99 = \chi^2_{2,0.95}$, H_0 wird also abgelehnt. (Nun darf man nach den biologischen Ursachen für die Abweichung von den Mendelschen Verhältnissen forschen: z.B. könnte Selektion gegen Heterozygote im Spiel sein, oder Wechselwirkung des Gens für die Blütenfarbe mit anderen Teilen des Genoms.)

Bemerkungen

- Die Symbole X^2, χ^2 und x^2 haben sich eingebürgert obwohl sie leicht missverständlich sind. Der Exponent "hoch 2" ist als Teil des Symbols zu lesen. X^2, χ^2 und x^2 werden also wie gewöhnliche Variablen behandelt. Insbesondere ist X^2 nicht mit dem Quadrat einer Zufallsvariablen X zu verwechseln!

- Es ist zu beachten, dass die χ^2-Verteilung eine schlechte Näherung ist, wenn die Stichprobe zu klein ist. Als Faustregel gilt: Der Test darf im Fall $r \leq 8$ nur benutzt werden, wenn alle $e_i \geq 5$ sind, und im Fall $r > 8$, wenn alle $e_i \geq 1$. Wenn die Daten (mindestens) ordinalskaliert sind und n groß genug ist, lassen sich die Faustregeln erfüllen, indem man benachbarte Werte zu größeren Klassen zusammenfasst.

- Bisher haben wir nur die Situation betrachtet, dass wir die Verteilung unter H_0 vollständig kennen, ohne dass irgendwelche Parameter spezifiziert werden mussten. Oft steht man aber vor der Situation, dass man einen (oder mehrere) Parameter der hypothetischen Verteilung erst aus der Stichprobe schätzen muss. Dann verringert sich die Zahl der Freiheitsgrade um a, die Zahl der geschätzten Parameter, also $k = r - a - 1$. Dazu geben wir noch ein Beispiel an.

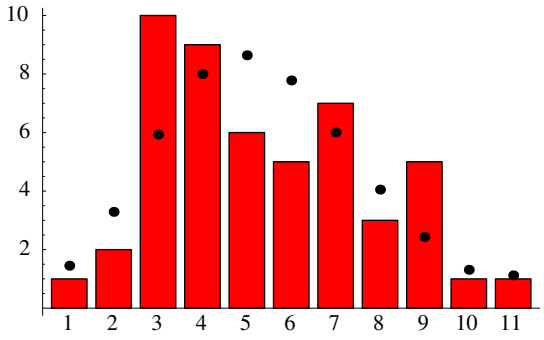


Abbildung 2.34: Beobachtete (Histogramm) und erwartete (Punkte) Häufigkeiten beim radioaktiven Zerfall. Für den Test werden die Randklassen zusammengefasst, damit e_i für alle Klassen ≥ 1 ist.

Beispiel: radioaktiver Zerfall

Mit einem Geiger-Müller-Zählrohr werden während einer gesamten Messdauer von 50 Sekunden die radioaktiven Emissionen pro Sekunde gemessen. Es ergibt sich folgende Urliste:

y_i : 7, 4, 3, 6, 4, 4, 5, 3, 5, 3, 5, 5, 3, 2, 5, 4, 3, 3, 7, 6, 6, 4, 3, 11, 9,
6, 7, 4, 5, 4, 7, 3, 2, 8, 6, 7, 4, 1, 9, 8, 4, 8, 9, 3, 9, 7, 7, 9, 3, 10.

Es soll überprüft werden, ob die Beobachtungen mit einer Poisson-Verteilung verträglich sind ($\alpha = 0.05$). Dazu muss zunächst der Parameter der Poisson-Verteilung aus der Stichprobe geschätzt werden: $\hat{\lambda} = \bar{y} = 5.4$. Mit Hilfe dieser Schätzung werden die erwarteten Häufigkeiten ermittelt: $e_i = n \cdot Pr(Y = i) = \frac{\hat{\lambda}^i}{i!} e^{-\hat{\lambda}}$. Die Beobachtungen werden so zu Klassen zusammengefasst, dass alle e_i größer 1 sind (zusammenfassen von Randklassen). Die beobachteten Häufigkeiten ergeben sich direkt aus der Urliste.

# Zerfälle	≤ 1	2	3	4	5	6	7	8	9	10	≥ 11
beobachtet	1	2	10	9	6	5	7	3	5	1	1
erwartet	0.23+1.22 = 1.45	3.29	5.93	8.00	8.64	7.78	6.00	4.05	2.43	1.31	0.64+0.29+... = 1.12

Es sind $r = 11$ Klassen, die Zahl der Freiheitsgrade beträgt $r - 1 - 1 = 9$, und $\chi^2 = 8.61 < \chi^2_{9,0.95} = 16.92$. Es ist also keine signifikante Abweichung von der Poissonverteilung festzustellen.

Anhang

Normalverteilungstabelle

$\Phi(z)$: Verteilungsfunktion der Standardnormalverteilung.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.7	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999
3.8	.9998	.9999	.9999	.9999	.9999	.9999	.9999	1.0000	1.0000	1.0000

Quantilentabelle der t-Verteilung

Quantilen $t_{k,q}$ der Studentschen t-Verteilung. Freiheitsgrade k in Zeilen, q-Werte in Spalten. Die letzte Zeile (∞) führt das entsprechende q-Quantil z_q der Standardnormalverteilung auf.

$k \diagdown q$	0.75	0.875	0.95	0.975	0.9825	0.99	0.995	0.999	0.9995
1	1.000	2.414	6.314	12.706	25.452	31.821	63.656	318.289	636.578
2	0.816	1.604	2.920	4.303	6.205	6.965	9.925	22.328	31.600
3	0.765	1.423	2.353	3.182	4.177	4.541	5.841	10.214	12.924
4	0.741	1.344	2.132	2.776	3.495	3.747	4.604	7.173	8.610
5	0.727	1.301	2.015	2.571	3.163	3.365	4.032	5.894	6.869
6	0.718	1.273	1.943	2.447	2.969	3.143	3.707	5.208	5.959
7	0.711	1.254	1.895	2.365	2.841	2.998	3.499	4.785	5.408
8	0.706	1.240	1.860	2.306	2.752	2.896	3.355	4.501	5.041
9	0.703	1.230	1.833	2.262	2.685	2.821	3.250	4.297	4.781
10	0.700	1.221	1.812	2.228	2.634	2.764	3.169	4.144	4.587
11	0.697	1.214	1.796	2.201	2.593	2.718	3.106	4.025	4.437
12	0.695	1.209	1.782	2.179	2.560	2.681	3.055	3.930	4.318
13	0.694	1.204	1.771	2.160	2.533	2.650	3.012	3.852	4.221
14	0.692	1.200	1.761	2.145	2.510	2.624	2.977	3.787	4.140
15	0.691	1.197	1.753	2.131	2.490	2.602	2.947	3.733	4.073
16	0.690	1.194	1.746	2.120	2.473	2.583	2.921	3.686	4.015
17	0.689	1.191	1.740	2.110	2.458	2.567	2.898	3.646	3.965
18	0.688	1.189	1.734	2.101	2.445	2.552	2.878	3.610	3.922
19	0.688	1.187	1.729	2.093	2.433	2.539	2.861	3.579	3.883
20	0.687	1.185	1.725	2.086	2.423	2.528	2.845	3.552	3.850
21	0.686	1.183	1.721	2.080	2.414	2.518	2.831	3.527	3.819
22	0.686	1.182	1.717	2.074	2.405	2.508	2.819	3.505	3.792
23	0.685	1.180	1.714	2.069	2.398	2.500	2.807	3.485	3.768
24	0.685	1.179	1.711	2.064	2.391	2.492	2.797	3.467	3.745
25	0.684	1.178	1.708	2.060	2.385	2.485	2.787	3.450	3.725
26	0.684	1.177	1.706	2.056	2.379	2.479	2.779	3.435	3.707
27	0.684	1.176	1.703	2.052	2.373	2.473	2.771	3.421	3.689
28	0.683	1.175	1.701	2.048	2.368	2.467	2.763	3.408	3.674
29	0.683	1.174	1.699	2.045	2.364	2.462	2.756	3.396	3.660
30	0.683	1.173	1.697	2.042	2.360	2.457	2.750	3.385	3.646
32	0.682	1.172	1.694	2.037	2.352	2.449	2.738	3.365	3.622
34	0.682	1.170	1.691	2.032	2.345	2.441	2.728	3.348	3.601
36	0.681	1.169	1.688	2.028	2.339	2.434	2.719	3.333	3.582
38	0.681	1.168	1.686	2.024	2.334	2.429	2.712	3.319	3.566
40	0.681	1.167	1.684	2.021	2.329	2.423	2.704	3.307	3.551
50	0.679	1.164	1.676	2.009	2.311	2.403	2.678	3.261	3.496
60	0.679	1.162	1.671	2.000	2.299	2.390	2.660	3.232	3.460
70	0.678	1.160	1.667	1.994	2.291	2.381	2.648	3.211	3.435
80	0.678	1.159	1.664	1.990	2.284	2.374	2.639	3.195	3.416
100	0.677	1.157	1.660	1.984	2.276	2.364	2.626	3.174	3.390
150	0.676	1.155	1.655	1.976	2.264	2.351	2.609	3.145	3.357
200	0.676	1.154	1.653	1.972	2.258	2.345	2.601	3.131	3.340
∞	0.674	1.150	1.645	1.960	2.240	2.326	2.576	3.090	3.291

Quantilentabelle der χ^2-Verteilung

Quantilen $\chi^2_{k,q}$ der χ^2-Verteilung. $k = n - a - 1$ ist die Zahl der Freiheitsgrade (in Zeilen), q-Werte in Spalten.

$k \backslash q$	0.7	0.75	0.8	0.85	0.9	0.95	0.975	0.99	0.995
1	1.07	1.32	1.64	2.07	2.71	3.84	5.02	6.63	7.88
2	2.41	2.77	3.22	3.79	4.61	5.99	7.38	9.21	10.60
3	3.66	4.11	4.64	5.32	6.25	7.81	9.35	11.34	12.84
4	4.88	5.39	5.99	6.74	7.78	9.49	11.14	13.28	14.86
5	6.06	6.63	7.29	8.12	9.24	11.07	12.83	15.09	16.75
6	7.23	7.84	8.56	9.45	10.64	12.59	14.45	16.81	18.55
7	8.38	9.04	9.80	10.75	12.02	14.07	16.01	18.48	20.28
8	9.52	10.22	11.03	12.03	13.36	15.51	17.53	20.09	21.95
9	10.66	11.39	12.24	13.29	14.68	16.92	19.02	21.67	23.59
10	11.78	12.55	13.44	14.53	15.99	18.31	20.48	23.21	25.19
11	12.90	13.70	14.63	15.77	17.28	19.68	21.92	24.73	26.76
12	14.01	14.85	15.81	16.99	18.55	21.03	23.34	26.22	28.30
13	15.12	15.98	16.98	18.20	19.81	22.36	24.74	27.69	29.82
14	16.22	17.12	18.15	19.41	21.06	23.68	26.12	29.14	31.32
15	17.32	18.25	19.31	20.60	22.31	25.00	27.49	30.58	32.80
16	18.42	19.37	20.47	21.79	23.54	26.30	28.85	32.00	34.27
17	19.51	20.49	21.61	22.98	24.77	27.59	30.19	33.41	35.72
18	20.60	21.60	22.76	24.16	25.99	28.87	31.53	34.81	37.16
19	21.69	22.72	23.90	25.33	27.20	30.14	32.85	36.19	38.58
20	22.77	23.83	25.04	26.50	28.41	31.41	34.17	37.57	40.00
21	23.86	24.93	26.17	27.66	29.62	32.67	35.48	38.93	41.40
22	24.94	26.04	27.30	28.82	30.81	33.92	36.78	40.29	42.80
23	26.02	27.14	28.43	29.98	32.01	35.17	38.08	41.64	44.18
24	27.10	28.24	29.55	31.13	33.20	36.42	39.36	42.98	45.56
25	28.17	29.34	30.68	32.28	34.38	37.65	40.65	44.31	46.93
30	33.53	34.80	36.25	37.99	40.26	43.77	46.98	50.89	53.67
40	44.16	45.62	47.27	49.24	51.81	55.76	59.34	63.69	66.77
50	54.72	56.33	58.16	60.35	63.17	67.50	71.42	76.15	79.49
60	65.23	66.98	68.97	71.34	74.40	79.08	83.30	88.38	91.95
70	75.69	77.58	79.71	82.26	85.53	90.53	95.02	100.43	104.21
80	86.12	88.13	90.41	93.11	96.58	101.88	106.63	112.33	116.32
90	96.52	98.65	101.05	103.90	107.57	113.15	118.14	124.12	128.30
100	106.91	109.14	111.67	114.66	118.50	124.34	129.56	135.81	140.17
150	158.58	161.29	164.35	167.96	172.58	179.58	185.80	193.21	198.36
200	209.99	213.10	216.61	220.74	226.02	233.99	241.06	249.45	255.26
500	516.09	520.95	526.40	532.80	540.93	553.13	563.85	576.49	585.21